E-ISSN : 2614-8404
P-ISSN : 2776-3234

JISA

(JURNAL INFORMATIKA dan SAINS)

VOL 4 No 1
June

Published by :
Program Studi Teknik Informatika
UNIVERSITAS TRILOGI

2021

JISA
(Jurnal Informatika dan Sains)

Volume 4, Edition 1, June 2021

Published by:
Program Studi Teknik Informatika
Universitas Trilogi

JISA	Vol : 4	Ed :1	Page : 01-95	Jakarta, Jun 2021	e-ISSN: 2614-8404	p-ISSN: 2776-3234

JISA
(Jurnal Informatika dan Sains)

Volume 4, Edition 1, June 2021

Editorial Address
Ruang Dosen Fakultas Industri Kreatif dan Telematika Lantai 3
Jalan Taman Makam Pahlawan No. 1, Kalibata, Pancoran, RT.4/RW.4, Duren Tiga, Pancoran, Kota Jakarta Selatan, Daerah Khusus Ibukota Jakarta 12760Telp :(021) 798001

Published by:
Program Studi Teknik Informatika
Universitas Trilogi

JISA	Vol : 4	Ed :1	Page : 01-95	Jakarta, Jun 2021	e-ISSN: 2614-8404	p-ISSN: 2776-3234

e-ISSN: 2614-8404
p-ISSN: 2776-3234

Table of Content

e-ISSN: 2614-8404
p-ISSN: 2776-3234

Implementation of K-Means Clustering Algorithm in Mapping the Groups of Graduated or Dropped-out Students in the Management Department of the National University

Muhammad Darwis [1], Liyando Hermawan Hasibuan [2], Mochammad Firmansyah [3], Nur Ahady[4], Rizka Tiaharyadini[5]
[1]Program Studi Magister Ilmu Komputer, Fakultas Teknologi Informasi, Universitas Budi Luhur
[2]Program Studi Magister Ilmu Komputer, Fakultas Teknologi Informasi, Universitas Budi Luhur
[3]Program Studi Magister Ilmu Komputer, Fakultas Teknologi Informasi, Universitas Budi Luhur
[4]Program Studi Magister Ilmu Komputer, Fakultas Teknologi Informasi, Universitas Budi Luhur
[5]Program Studi Magister Ilmu Komputer, Fakultas Teknologi Informasi, Universitas Budi Luhur
email: [1]darwis.mawardin@gmail.com, [2]liyandohermawan@gmail.com, [3]bgolem@gmail.com [4]nahady18@gmail.com, [4]rizkatiaharyadini27@gmail.com

Abstract* - The dropout rate at the National University is still high. National Universities must make efforts to anticipate this rate and increase the number of graduates. This study aims to determine the characteristics of students who are likely to graduate or drop out (DO) in the management department of the National University, Jakarta. The study was conducted by implementing the K-Means algorithm, where each data is grouped according to the closest distance to the centroid. Determination of Cluster C1 graduate or C2 drop out is based on the attributes of status of students (active, leave, out and non-active), educational status (graduated or DO), GPA, total credits taken and length of study. To facilitate the clustering process, Orange tools are used that provide K-Means algorithm features. The total data input in this study were 1988 students from various classes. As a result, a pattern or mapping of graduated or DO students was found based on the attributes mentioned earlier. Testing the results of this cluster with the silhouette method, by measuring the distance between cluster members, both C1 and C2, showed good Silhouetter value, reaching 85% which indicates that this clustering method can be applied as an effort to overcome the high dropout rate. The management department, National University can use the results of this study to predict the graduation of their students.

Keywords - data mining; clustering; K-mean algorithm; orange; graduation; mapping;

I. INTRODUCTION

The National University management department is part of the faculty of economics which was founded in 1964, then its status was registered and recognized in 1985 by the Ministry of Education and Culture. Since then, the department has successfully graduated thousands of graduates, who have worked in various sectors in companies, government agencies, universities, and also entrepreneurs. The National University has been recognized as one of the higher education institutions in Indonesia [1].

However, along the way, not a few students have dropped out (DO) in the management department of the National University. This has an impact on the ups and downs of credibility and public trust of the university. To overcome this and maximize the learning process, it is better for the National University to know the patterns and mapping of student groups who have passed or dropped out, based on their academic data. Thus, the National University can make preventive efforts to prevent dropouts in students and maintain their credibility and reputation.

Data mining can provide solutions to the National University through the clustering method. One of the features that can be used in this case is the K-Means Clustering algorithm. With this technique, a pattern or mapping will be produced that can be used by the National University to determine the characteristics of graduated or dropped-out students based on their academic data so far.

The K-Means clustering algorithm has been widely used by various researchers to group objects based on their conditions and characteristics. As long as data is available, this method can be applied in various fields, such as health, education, disaster, military and so on [2], [3], [4], [5], [6], [7], [8], [9], [10], [11]. In the field of education, especially the clustering method at universities, researchers usually try to make patterns or mappings that are related to students, such as interests, academic conditions and students' graduation [12], [13], [14], [15], [16], [17], [18].

The National University has documented the student data well in their system. However, the data is still in raw form so it is difficult to read. To better utilize and maximize the existence of this data, the National University should implement data mining in it. Besides being very helpful, data mining methods are trusted and have been widely used

e-ISSN: 2614-8404
p-ISSN: 2776-3234

by various other institutions to extract more value from their data.

Therefore, in this study, the data processing and grouping of students in the Management Department of the National University were carried out using the clustering method. Just as other researchers have done, this study will utilize the K-Means clustering method to do this. The number of datasets processed in this study were 1,988 records. With this method, a pattern and mapping will be generated that can be used to see and predict the likelihood that students will graduate or drop out.

II. RESEARCH METHOD

The research methodology in the clustering process using the K-Means method in mapping the groups of the graduated or dropped-out students at the Management Department of the National University includes several stages: literature review, determination of the data sets and data pre-processing—which consists of data validation, data transformation and data reduction as well as data exploration using Orange application. To test the results of this study, silhouette was applied. The results will show whether or not this clustering method can be used as an effort to overcome the high dropout rate at the National University. The description of the flow and stages of the research that the authors did for this study is as shown in Figure 1.

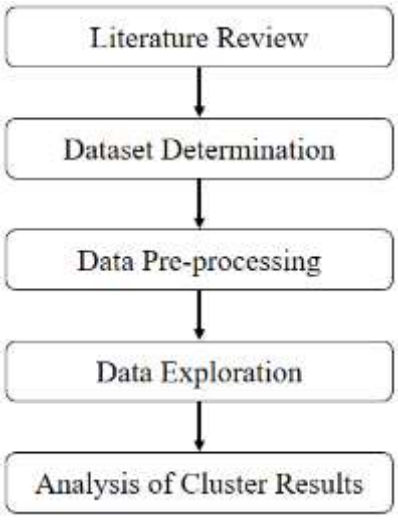


Figure 1. Flow and research stages

A. Literature Review

As literature to support this research, the authors use several trusted scientific journals, proceedings, e-books and websites. These papers contain about clustering, such as the Application of the K-Means Method for Student Clustering Based on Academic Values with a Weka Interface Case Study at the Department of Informatics, UMM Magelang [12], Prediction of Students Academic Execution Using K-Means and K-Medoids Clustering Technique [12] 18], K-Means Cluster Analysis in the Grouping of Students' Capabilities [17] and the Implementation of the K-Means Method in Mapping Student Groups through Lecture Activity Data [15].

Based on the literature review, it is known that the K-means clustering method in education has been widely used, especially in grouping students, both learning activities and predicting academic scores. The percentage of grouping shows good results, so that it can be applied to solve problems at the university. The focus of this research is to cluster students who are likely to graduate or drop out (DO) using the K-Means method.

B. Dataset Determination

The data set in this study was obtained from the Department of Management, the National University, Jakarta. Furthermore, the data is processed to the data pre-processing stage. The number of datasets processed in this study were 1,988 records or rows of data containing detailed information about the academic data of students majoring in Management at the National University of various generations, from 1990 to 2016. In detail, from the number of student data collected in this study, 56% were male and 44% were female. The number of student data from the class of 1990-1995 was 48% or 960 students, the class of 1996-2000 was 17% or 343 students, the class of 2001-2005 was 20% or 390 students, the class of 2006-2010 was 3% or 65 students, the class of 2010-2016 was 12% or 230 students. For this study, several attributes were used as consideration for creating student clusters, whether they pass or drop out. These attributes are the GPA score, the number of credits taken, length of study, initial status and final status of students. From these attributes, it is hoped that a cluster will be formed that can help the National University to make predictions for students' graduation or drop-out.

C. Data Pre-processing

This stage includes 3 main parts: data validation, data transformation, and data reduction. For information, to facilitate the implementation of this stage, several features in the Orange application are used. The details of the three stages are:

1. Data validation

The data validation stage in this study was carried out to ensure that the training data set was in good condition and there were no more missing values in it. The missing value in question is due to incomplete data, outliers (abnormal data) or data with inconsistent values. From the data validation stage, data will be generated where there are no missing values or normal conditions for the next stage.

2. Data transformation

The next stage is data transformation, which is performed with the aim of maintaining and ensuring the accuracy of the training data set. For this stage, the outlier technique contained in the Orange application is also applied, namely the Outlier Detection Method feature. From this data transformation process, more accurate data will be generated for the next stage.

3. Data reduction

Data reduction is done to take data sampling from the existing training set. The goal is that the selected data sets are really ready to be classified. Of the total 1,988 data records, then filtering was carried out until the remaining 60% of the data or 1,173 records were left. So, from this

e-ISSN: 2614-8404
p-ISSN: 2776-3234

stage, a sample data will be generated that is ready to be processed to the next stage.

D. Data Exploration

This stage is one of the important stages in the clustering process, because this is where the data will actually be processed: the value is calculated and the results are analyzed. In this study, data processing and calculation methods were carried out using the features in the Orange application. There are several features that will be used in it, such as the outlier feature, K-Means, select rows and a scatter plot to see the clustering results.

E. Clustering Result Analysis

At this stage, the results of the classification or grouping that have been obtained from calculations using the Orange application are analyzed and reviewed. From this stage, patterns might be as the results of K-Means clustering.

III. RESULTS AND DISCUSSION

Research for the clustering of students majoring in Management at the National University starts from collecting training data sets, designing and implementing data mining. Later, a certain pattern will be generated that can be used to predict the possibility of students graduating or dropping out. These stages include identification of training sets, data preparation, data exploration using Orange and analysis of the results.

A. Data Set

The first stage carried out in this study was the collection and definition of the dataset as shown in Figure 2. In this data, which contains data from students of the Management Department of the National University of various generations from 1990 to 2016, there are 1,988 data records available. The dataset is then uploaded to Orange to facilitate data processing, including data preparation. In this set, it was seen that there were still missing values of 0.3% or around 34 rows or data records.

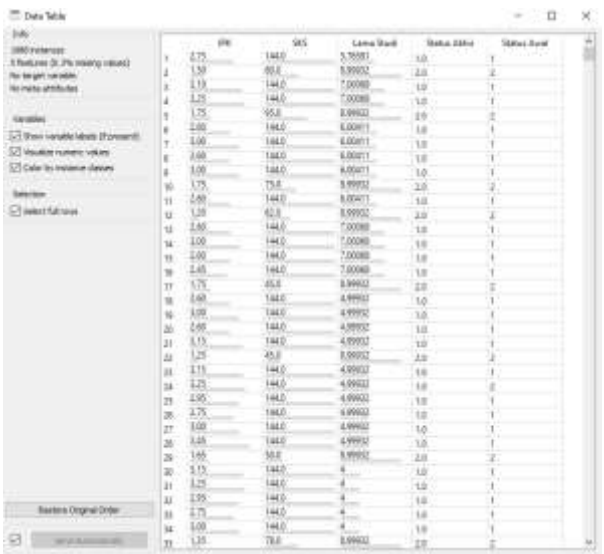


Figure 2. Dataset of Management Department students, the National University, Jakarta

From the data table, it can be seen that there are at least 5 attributes that have been selected from the data set and then processed to form a cluster. These attributes include:

a. GPA, contains GPA data records for each student and is a numerical attribute. The value is a range of GPA numbers from 0-4 which have been classified based on the Regulation of the Minister of Education and Culture of the Republic of Indonesia Number 49 of 2014 concerning National Higher Education Standards regulating the assessment and cumulative achievement index contained in articles 23 and 24. Assessment reports are in the form of students' success qualifications in taking a course stated in the range of:
   - GPA <= 4, in good category.
   - GPA <= 3, in sufficient category.
   - GPA <= 2, in the poor category.
   - GPA <= 1, in the very poor category.

b. SKS, is a numeric attribute that contains data on the number of credits that have been taken by students. The values are in the range of 1 to 144.

c. Duration of study, namely data on the length of study that students have taken until they graduate or drop out and is a numerical attribute.

d. Initial Status, is student status data during the lecture process. The authors has classified this data and made it a numerical attribute. The value is 1 for "active" status, 2 for "inactive" status, 3 for "leave" status and 4 for "out" status.

e. Final Status is data on the final academic status of students, graduated or dropped out (DO). The authors made it as a numerical attribute with a value of 1 for "pass" status and 2 for "Drop Out" status.

B. Data Pre-processing

This section contains the data preparation stage, which consists of data validation, data transformation and data reduction. The authors use the Orange application in all stages of this preparation to make the data processing easy and effective.

a. Data validation

As previously explained, there are missing values in the dataset used in this study, so it is necessary to pre-process the data first. This is also to improve the validation of existing data. For this reason, the Preprocess feature contained in Orange is used to do this. Furthermore, there is also a Data Table feature that can be used to view data, including data from the validation results. As a result, the missing values in the dataset were eliminated so that the number of data became 1,954 records and could be continued to the next process. These results are as shown in Figure 3.

e-ISSN: 2614-8404
p-ISSN: 2776-3234

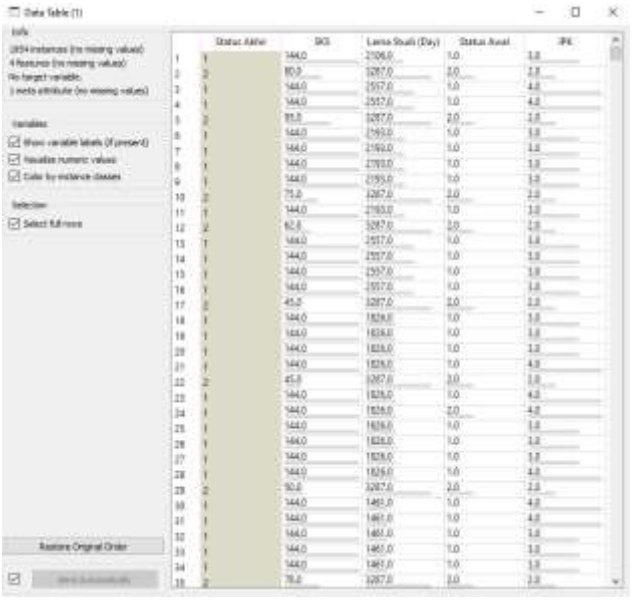


Figure 3. Results of data validation

b. Data transformation

To perform the data transformation process, the Outlier feature in Orange is used. Outlier means there is an unnatural (anomalous) data set in the data set, so that it must be corrected before the next stage. After that, the results are displayed in the form of Data Inliers, which means data that is not normal or all data other than normal data. The description of the Outlier feature as a transformation process in Orange is as shown in Figure 4.

Figure 4. Outlier process in Orange

As additional information, before the dataset is entered into the Orange application, data creations have been carried out. The data creations are the addition of several attributes that are deemed necessary and changing some attributes that contain categorical data into numeric data. This aims to facilitate the classification process with the K-Means feature. These attributes include initial status, final status and length of study as described in the dataset section of this paper.

c. Data reduction

This data reduction is basically a sampling of existing data, especially because there are quite a lot of rows. In addition to streamlining the algorithm, it is also less heavy and doesn't take long to process data. Although several rows of data are reduced at this stage, the quality of the resulting data remains the same, so that it still meets the research requirements.

In the Orange application, to perform this stage, the "Data Sampler" feature can be used which is linked from the Preprocess feature. As previously explained, the data used at this stage is about 60% of the total data or as many as 1,954 records. The technique used is random sampling where the data will be randomly selected. From the "Data Sampler" output, the Outlier feature is then used, so that later, inlier data is obtained, namely data other than data from outlier data. The data is then processed further so that a total of 592 records are obtained from all existing data. Inliers data obtained as output is as shown in Figure 5. The inliers data will be processed further at a later stage. At this stage, the data is really ready to be processed and explored further.

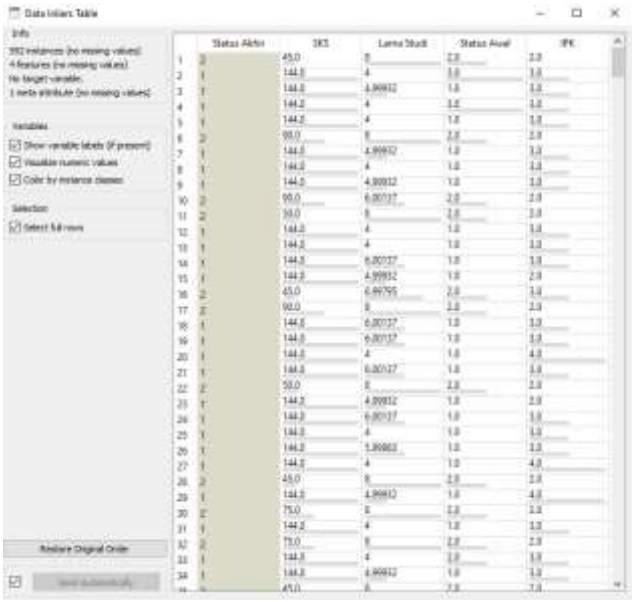


Figure 5. Data sampling inliers

d. Data exploration

In the processing and exploring of the dataset using Orange, the features are emphasized. In addition, the process of identifying the intensity of the relationship between attributes is carried out. At this stage also, a trial with the Unvariate Analysis technique was carried out. Actually, at this stage, testing can also be done with the Bivariate Analysis and Multivariate Analysis techniques. However, due to limitations, only Unvariate Analysis is used where the properties of each attribute are investigated. The technique is done by applying the "Feature Statistics" feature provided by Orange. The picture is as in Figure 6.

Figure 6. Results of "feature statistics" on Orange

From Figure 6, the results of data processing can generally be read. The data center for each attribute is SKS = 113.43, GPA = 2.87, length of study = 5.75 years, initial status = 1.57 and final status = 1.35. Next is data processing using the K-Means method with the aim of grouping students' data into 2 clusters. To do clustering in Orange, use the "K-Means" feature and then select the attribute row that becomes the center of the cluster with the "Select Row" feature. Figure 7 illustrates the results of clustering against the dataset. Cluster C1 shows groups of students who have

e-ISSN: 2614-8404
p-ISSN: 2776-3234

successfully passed while cluster C2 shows students who drop out.

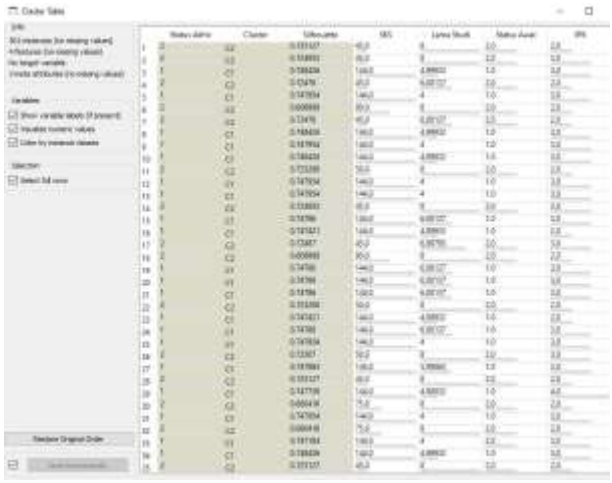


Figure 7. The results of clustering with the K-Means feature

In theory, according to [15], the following are the calculation steps using the K-Means method:

a) Determine the number of clusters.
b) Allocate data into clusters randomly.
c) Calculate the centroid / average of the data in each cluster
d) Allocate each data to the nearest centroid / average.
e) Return to Step c), if there is still data that has moved to different clusters or if the change in the centroid value is above the specified threshold value or if the value change in the objective function used is above the specified threshold value

One alternative to the application of K-Means with several related calculation theories development is Eucledian Distance (L2-Norm). The distance between two points is formulated as follows:

$$d(x,y) = \|x,y\|^2 = \sum_{i-1}^{n}(x_i - y_i)^2$$

Information:
d = determinant (Euclidean Distance)
x = the center of the cluster
y = data
n = amount of data
i = data to-

Furthermore, [15] describes the shortest distance between the centroid and the document to determine the cluster position of a document. For example, document A has the shortest distance to centroid 1 compared to the others, then document A is included in group 1. Recalculate the position of the new centroid for each centroid (Ci.j) by taking the average of the documents that enter the initial cluster (Gi..j). Iteration is carried out continuously until the group position does not change. The following is the formula for determining the centroid:

$$C_i = \frac{x_1 + x_2 + x_{..} + x_{...}}{\sum x}$$

Information:
x1 = the value of the 1st data record
x2 = 2nd record data value
Σx = number of data records

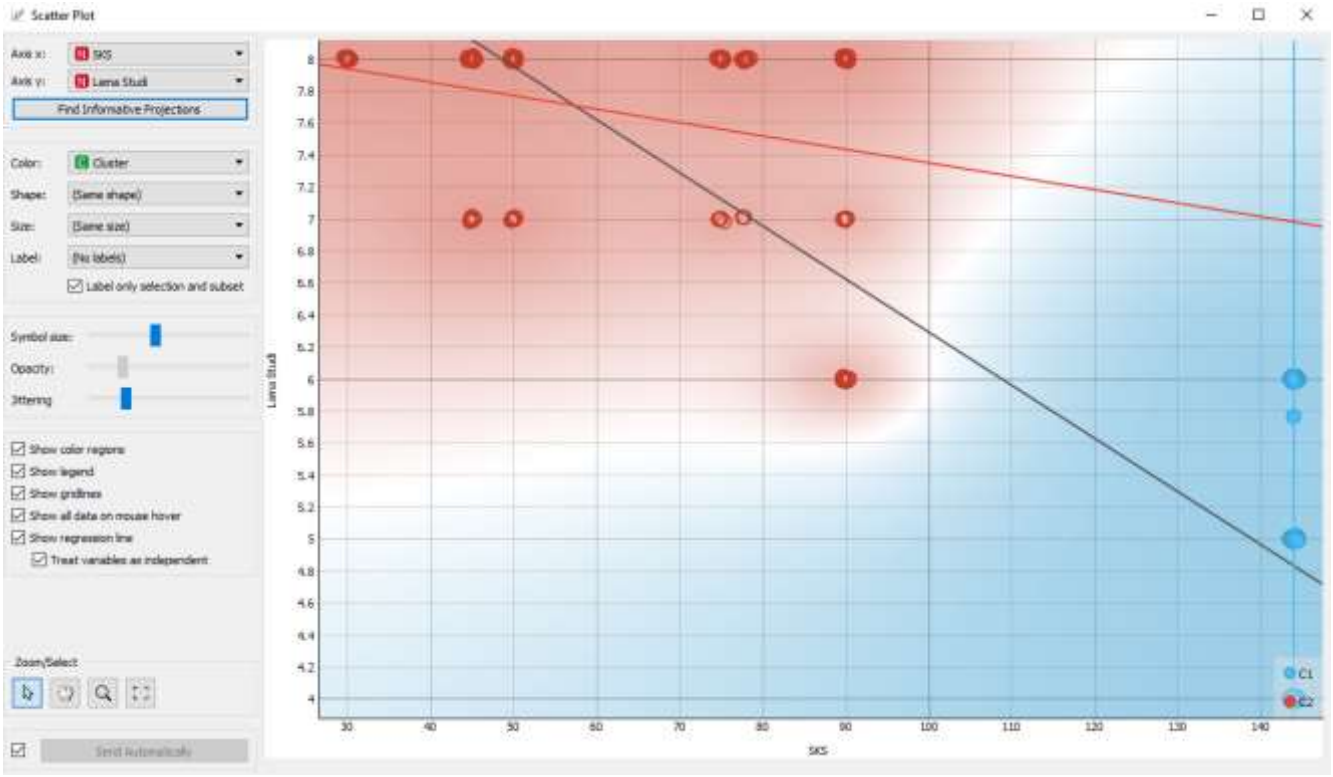


Figure 8. Clustering graphs of graduating and dropout students with the Scatter Plot feature based on credits and length of study

e-ISSN: 2614-8404
p-ISSN: 2776-3234

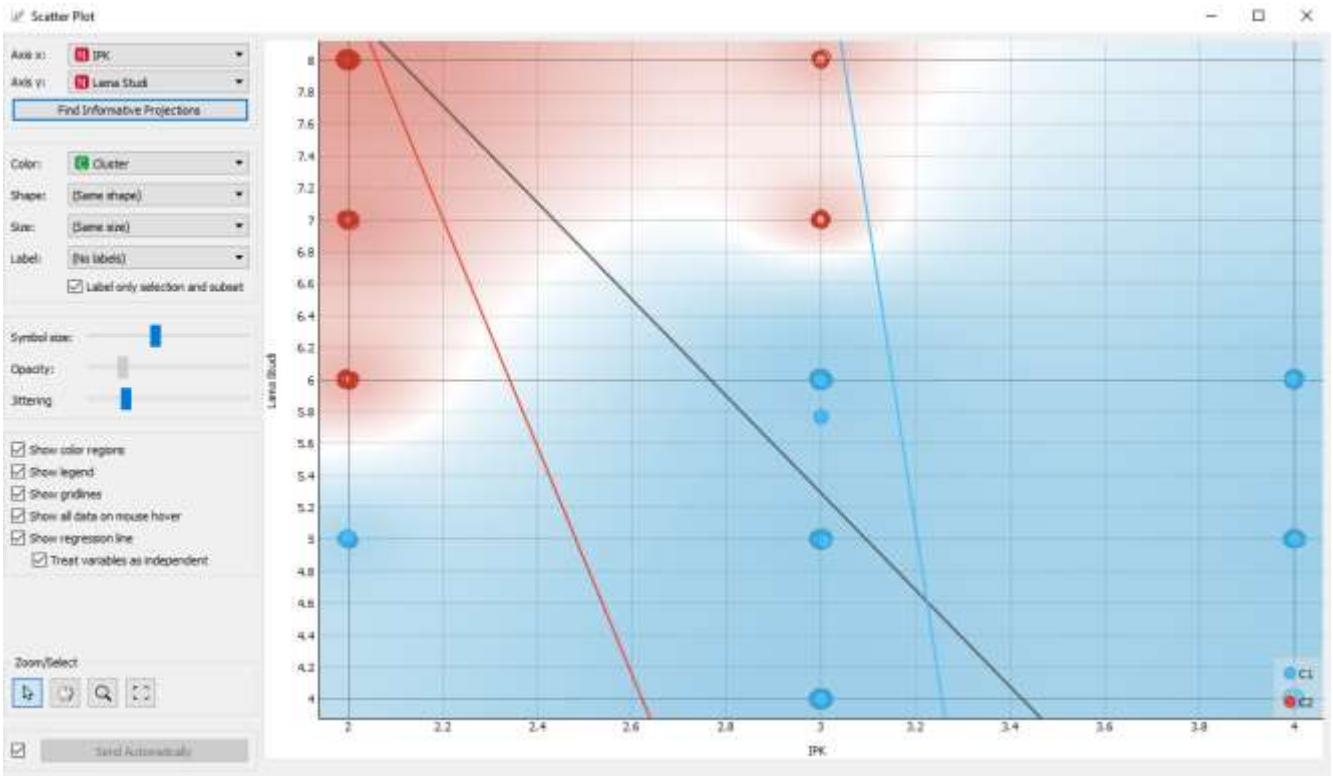


Figure 9. Clustering graphs of graduating and dropout students with the Scatter Plot feature based on GPA and length of study

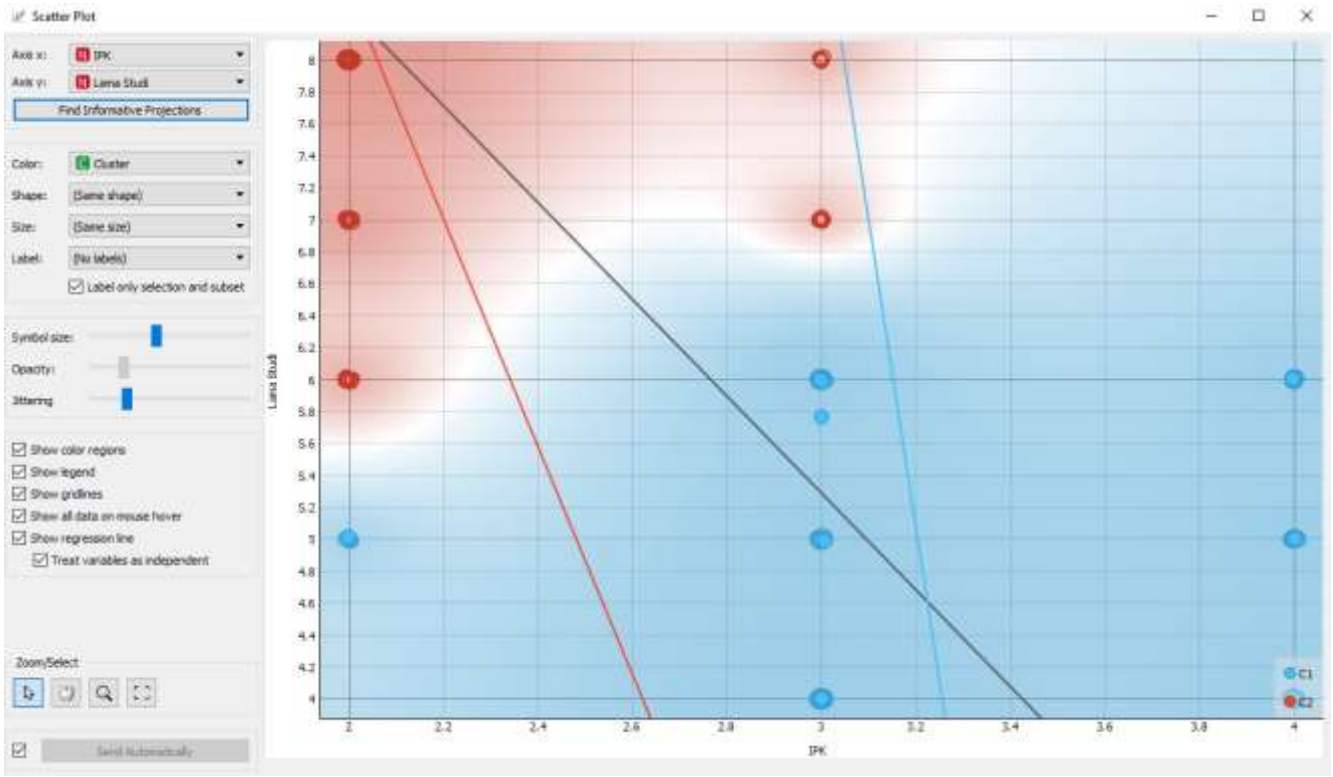


Figure 10. Clustering graphs of of graduating and dropout students with the Scatter Plot feature based on their final status and length of study

By processing data using Orange, it is quite easy to visualize the results of the cluster, namely the "Scatter Plot" feature. The cluster graph was then analyzed and tested. For more details, one of the cluster graphs is shown in Figure 8.

e-ISSN: 2614-8404
p-ISSN: 2776-3234

C. Analysis of Classification Results

The cluster graph shown in Figure 8, illustrates the grouping of students in the Department of Management of the National University who have successfully passed or dropped out based on the number of credits taken and the length of study. Cluster C1, which is marked in blue contains a group of students who have successfully passed. Meanwhile, cluster C2 is a group of students who drop out, which is marked in red. In the C2 cluster group, it can be seen that the length of study for DO students is over 6 years with the number of credits taken below 100. Meanwhile, in the C1 cluster, it can be seen that the group of students who successfully passed education for 6 years and under with the full number of credits is 144. This graph shows the irregularity of the academic conditions of dropping out students where they have been recorded and studied for a long time, but still attend lectures with a small number of credits. This is an anomalous condition where the length of the study is inversely proportional to the number of credits taken.

Figure 9 shows the clustering of students who graduated or dropped out (DO) based on the GPA obtained and the length of study. Cluster C2 shows groups of students who drop out and is marked in red, while cluster C1 is a group of students who have successfully completed their studies, which is marked in blue. From this graph, it can be seen that on average students with cluster C1, graduating and completing the study, have a GPA> = 3 with a length of study <= 6 years. Meanwhile, the group of students who dropped out, cluster C2, had a GPA> = 3 with a length of study> = 6 years. This is certainly quite reasonable because the academic conditions of students who drop out usually have a GPA below the average, especially if they have taken courses for more than 4 years.

The cluster graphic in Figure 10 emphasizes the relationship between the number of credits a student has taken or the GPA a student has with the length of study he/she has taken. The graph depicts students whose status has finally passed, cluster C1 is marked in blue and groups of students who drop out are marked in red. In line with the pattern in Figure 8 and Figure 9, in this graph it can be seen that the group of students who successfully graduated, cluster C1, has a length of study period under 6 years. Meanwhile, cluster C2, students who drop out (DO), have a study period of more than 6 years.

This study shows that students who have taken education above the normal time, which is 4 years or more but are still recorded taking a small number of credits, it is necessary to take preventive measures so they will not drop out. The university should have taken anticipatory steps, so that the students concerned can complete their studies. Furthermore, students who have a GPA below 3 and have taken education above the normal time, 4 years or below, need to be given special attention as a preventive measure.

There are several other insights that can be taken from research that applies clustering with the K-Means method based on Figure 7, including:

a) Students who have taken SKS> = 90 but have GPA> = 2, need to be given special attention. These students need to be supported in order to increase their achievement index, especially if the students have taken education above the normal time, 4 years.

Figure 11. Cluster category

b) Students whose status was initially inactive and had just taken the number of SKS <= 90, need special attention. Do not let it drag on and become a drop out (DO) in the final status, especially if the students have taken education above the normal time, 4 years. The university needs to be pro-active in communicating with these students.

c) Students whose GPA> = 3, and have taken SKS> = 90, but the length of study is above normal, 4 years, need to be given special attention, especially if they have been recorded as inactive or on leave.

To test the results of the clusters produced in this study, the Silhouette Plot feature which is also included in the Orange application is used. According to [19], the Silhouette Coefficient is used to see the quality and strength of the cluster, how well an object is placed in a cluster. This method is a combination of the cohesion and separation method. The stages of calculating the Silhouette Coefficient are as follows:

a) Calculate the average distance from a document for example i with all other documents that are in the same

e-ISSN: 2614-8404
p-ISSN: 2776-3234

cluster:

$$a(i) = \frac{1}{|A|-1} \sum j \in A, j \neq i \; d(i,j)$$

where j is another document in one cluster A and d(i, j) is the distance between document i and j.

b) Calculate the average distance from document i to all documents in other clusters, and take the smallest value.

$$d(i,C) = \frac{1}{|A|} \sum j \in C \; d(i,j)$$

where d (i, C) is the average distance of document i with all objects in other clusters C where A ≠ C.

c) The Silhouette Coefficient value is:

$$b(i) = \min C \neq A \, d(i,j)$$

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

The results of the Silhoutte Coefficient shown by the application of Orange in this study are as shown in Figure 11. The silhouette method average value approach is used to estimate the quality of the clusters formed. The higher the average value is, the better it is. Based on the graph in Figure 11, it can be seen that the optimal cluster formed at C1, k = 1, with an average Silhoutte value of more than 0.85 or 85%. All graduate student members, cluster C1, showed average grades. In cluster C2, k = 2, the Silhouette values vary between 0.5 - 0.8. However, the value is quite good. This shows that the clusters formed in this study have been tested and can be used by the Management Department of the National University to predict the likelihood that students will graduate well or experience drop out (DO).

rsitas

IV. CONCLUSION

The K-means algorithm can be used to cluster and classify students who graduate or drop out (DO). The results of the clustering are used as predictions to determine the possibility of students' academic conditions that can pass or experience dropouts. Insights obtained from the results of this calculation can be used by the Department of Management of the National University to take preventive steps and ensure that students do not drop out. This K-means method is easily implemented in Orange because it has proven features, so it is quite fast and accurate.

The results of the silhouette test in this study reached 85%, which indicates that the clustering method to determine the characteristics of students who are likely to graduate or drop out (DO) in the management department of the National University, Jakarta, can be applied as an effort to overcome the high dropout rate. With the results of clustering like this, for example, it makes it easier for universities to control and observe the academic conditions of their students. Finally, the campus can improve academic services and graduate students with good academic conditions.

REFERENCES

[1] Universitas Nasional, "Sejarah Perkembangan Fakultas Ekonomi Universitas Nasional," *Universitas Nasional*, 2020. [Online]. Available: http://fe.unas.ac.id/profil-fakultas/sejarah-perkembangan-fakultas/.

[2] W. M. P. Dhuhita, "Clustering Menggunakan Metode K-Means Untuk Menentukan Status Gizi Balita," vol. 15, no. 2, 2015.

[3] M. H. Adiya and Y. Desnelita, "Jurnal Nasional Teknologi dan Sistem Informasi Penerapan Algoritma K-Means Untuk Clustering Data Obat-Obatan Pada RSUD Pekanbaru," vol. 01, pp. 17–24, 2019.

[4] A. V. Vidhyapeetham, "Crime Analysis and Prediction using Optimized K-Means Algorithm 1 1," no. Iccmc, pp. 915–918, 2020.

[5] P. Manivannan, "Dengue Fever Prediction Using K-Means Clustering Algorithm," pp. 1–5, 2017.

[6] S. Kaur, "Disease Prediction using Hybrid K-means and Support Vector Machine," 2016.

[7] X. Chen and P. Miao, "Image Segmentation Algorithm Based on Particle Swarm Optimization with K-means Optimization," *2019 IEEE Int. Conf. Power, Intell. Comput. Syst.*, pp. 156–159, 2019.

[8] R. Alzu, "Medical Image Segmentation via Optimized K-Means," pp. 959–962, 2017.

[9] F. Natalia, R. I. Desanti, and F. V. Ferdinand, "Prediction and Visualization of Flood Occurences in Tangerang using K-Medoids , DBScan , and X-Means Clustering Algorithms," pp. 43–47, 2019.

[10] A. Suresh, "Prediction of major crop yields of Tamilnadu using K-means and Modified KNN," *2018 3rd Int. Conf. Commun. Electron. Syst.*, no. Icces, pp. 88–93, 2018.

[11] E. Hot and V. Popovi, "Soil data clustering by using K-means and fuzzy K-means algorithm," pp. 890–893, 2015.

[12] ASRONI and R. ADRIAN, "Penerapan Metode K-Means Untuk Clustering Mahasiswa Berdasarkan Nilai Akademik Dengan Weka Interface Studi Kasus Pada Jurusan Teknik Informatika UMM Magelang," *J. Ilm. SEMESTA Tek.*, vol. 18, no. 1, pp. 76–82, 2015.

[13] F. Nur, R. Fauzan, J. Aziz, B. D. Setiawan, and I. Arwani, "Implementasi Algoritma K-Means untuk Klasterisasi Kinerja Akademik Mahasiswa," vol. 2, no. 6, pp. 2243–2251, 2018.

[14] H. Priyatman, F. Sajid, and D. Haldivany, "Klasterisasi Menggunakan Algoritma K-Means Clustering untuk Memprediksi Waktu Kelulusan," vol. 5, no. 1, pp. 62–66, 2019.

[15] A. Fadlil, M. Teknik, I. Universitas, A. Dahlan, T. Elektro, and U. Ahmad, "Implementasi Metode K-Means Dalam Pemetaan Kelompok Mahasiswa Melalui Data Aktivitas Kuliah," vol. 3, no. 1, pp. 22–31, 2018.

[16] D. W. Widodo, "Implementasi Algoritma K-Means Clustering Untuk Mengetahui Bidang Skripsi Mahasiswa Multimedia Pendidikan Teknik Informatika Dan Komputer Universitas Negeri Jakarta," *J. Pinter*, vol. 1, no. 2, pp. 157–166, 2017.

[17] B. Poerwanto and R. . Fa'rifah, "Analisis cluster k-means dalam pengelompokan kemampuan mahasiswa," *J. Sci. Pinisi*, vol. 2, no. 2, pp. 92–96,

e-ISSN: 2614-8404
p-ISSN: 2776-3234

2015.

[18] S. Agrawal, "Prediction of Students Academic Execution Using K-Means and K-Medoids Clustering Technique," *2018 2nd Int. Conf. Trends Electron. Informatics*, no. Icoei, pp. 1308–1315, 2018.

[19] F. M. IZZADIN, "Optimasi Jumlah Cluster K-Means Dengan Metode Elbow dan Silhouette Untuk Pengelompokan Luas Panen Palawija Kabupaten Magelang Pada Tahun 2017," *Medium The Startup*, 2019. [Online]. Available: https://medium.com/@16611050/optimasi-jumlah-cluster-k-means-dengan-metode-elbow-dan-silhouette-untuk-pengelompokan-luas-panen-200131515c4f.

e-ISSN:2614-8404
p-ISSN:2776-3234

System Development for Learning Process Monitoring in Private Lesson Institution Using Codeigniter Framework

Arief Herdiansah
Program Studi Teknik Informatika, Fakultas Teknik, Universitas Muhammadiyah Tangerang
Email : arief_herdiansah@umt.ac.id

***Abstract* – The monitoring information system is one of the main functions in managing student data in Generasi Cerdas private lesson. Currently, the method of delivering information on the results of the learning process to the parents of students is still carried out with a conventional process, private lesson generation smarts make reports on the results of the learning process and reports on student attendance using Microsoft Excel, then print them and give them to parents through their children. The problems that occur are often students forget to provide the report to parents and sometimes even documents are lost so parents have to ask to return to the private lesson. This study provides the results of an analysis and design of a monitoring information system that can be used to facilitate services to students in monitoring including grades and student absence. This application developed using UML Design and PHP Programming with CodeIgniter Framework and use MySQL Database. CodeIgniter is an open source framework in the form of a PHP framework with an MVC model (Model, View, Controller) for building dynamic websites. The resulting application has been able to provide information of teacher, student, student value and student absence data to assist in processing student and teacher data in the Generasi Cerdas private lesson, so that the institution has a faster, more efficient monitoring information system and easier to use.**

Keywords – Learning, Monitoring, Private Lesson, CodeIgniter

I. INTRODUCTION

An information system is a system that involves information technology, including computers, software, databases, communication systems, the Internet, mobile devices and others that work specifically on tasks, interacting by informing various actors in different organizational or social contexts [1]. Application of information systems in educational institutions can make the work carried out more neatly, quickly and accurately [2]. A computerized system will make it easier and help companies make decisions [3] [4].

Private lesson is one of the activities carried out to provide assistance to students in order to get more optimal achievements or learning outcomes at the institutions where they study. Tutoring aims to make students able to adjust to the current educational situation. By following this, students will get many benefits, which are the students' understanding of subjects that have been considered difficult, developing the ability to socialize, and also improving the achievements of the students themselves.

Generasi Cerdas is one of the private lesson institution in Tangerang City. At present, the private lesson institution has problems in data management, especially in monitoring teaching and learning process, processing and checking student achievement data are still using a manual system where the data storage is still scattered in each teacher so that it has not been well documented.

The above problems can be solved by designing an information system monitoring teaching and learning process on that course based information technology. Identification of problems that exist include:

1. The process of recording data on teaching and learning process information is not efficient because it is done manually.
2. There is still a duplication of student data, especially teacher data relating to learning and teaching hours.
3. Parents have difficulty in monitoring the process of teaching and learning carried out by their children.

Based on the above, the researcher made a study to develop a web-based information system that is easy to use by teachers, students and parents, but this research is limited by the scope of limited with data management restrictions as follows:

1. Student data, subject data, lesson schedule data, student grade data and student attendance data in the course.
2. Data from teaching hours of teaching in the course. Reports containing information relating to the learning and teaching process in the course.

The learning process that is currently running increasingly depends on the use of technology and the process of integrating new technological trends into the education system is continuously being improved for better results for students wherever they are [5]. The use of technology in the learning process has the main objective of facilitating interaction between teachers, students and parents so that the objectives of the teaching and learning process can be achieved [6] [7]. The system is developed to have a dashboard, which is a menu that displays intelligent business data that is useful for the analysis process [8].

PHP is a server-side scripting language that integrates with HTML to create dynamic web pages. The purpose of server-side scripting is syntax and the commands, which are given, will be fully executed in the system [9] [10].

There are several PHP programming language

e-ISSN:2614-8404
p-ISSN:2776-3234

frameworks, including: CakePHP, Laravel, CodeIgniter which are widely used by information system developers, each of which has advantages.

CodeIgniter is the simplest PHP framework with the smallest number of files compared to other PHP frameworks and the CodeIgniter framework has the best performance for Complex Data, CRUD Operations, and Image Upload tasks [11].

As a framework, CodeIgniter has advantages in terms of fairly complete libraries and packages, making it easier for developers to design a website. The developer doesn't need to code everything from scratch, just use the library provided

Research on "Making SMS Gateway Application for Academic Information at Be Excellent course, Pacitan". The SMS gateway service for academic information can be used by the Be Excellent course, Pacitan to disseminate information to students and parents. The Be Excellent course can provide academic information which can be accessed by students or parents by auto response or by broadcast. With the SMS gateway service, academic information from Be Excellent can be received directly by students 'or parents' cellphones via short messages so that information can be conveyed more quickly and on target [12].

Research on "Design and Development of Information Systems for Web Based Subjects and SMS Gateway". In the previous system, parents did not get grades directly from the school except at the end of each semester, so they had difficulty knowing the development of their children's grades while at school. Therefore, the application of the SMS Gateway is expected to make it easier for parents and students to find out the value of the Enrichment Exams, Midterm Exams, and End of Semester Exams via SMS. After testing the system, it is concluded that this system can be applied in elementary schools and junior high schools [13].

The two studies above show the importance of using IT-based information systems to help the learning process and inform parents about the results, so that teachers and parents can work together to support children's education. The research carried out is applied research where the system development process uses the PHP programming language CodeIgniter framework which will make the resulting application powerful and easy to use by the user.

II. RESEARCH METHODOLOGY

2.1. Research Methods

This study applied a mixed research method (quantitative & qualitative), where data collection would be carried out by means of a survey method using a statement/ questioner and conducted direct interviews with relevant parties. In this study, researchers took data from 65 respondents consisting of 30 students, 30 parents and 5 teachers.

2.2. Sample Selections

The sampling method used was purposive sampling. Sample taking with purposive sampling is a sampling technique by taking respondents selected by researchers according to the specific characteristics of the sample.

The criteria for the people chosen as respondents in this study are:

1. Knowing the role of Bimbel Generasi Cerdas as an institution engaged in education.
2. Knowing the importance of data entry activities of teaching and learning.
3. Recognizing the importance of data and the importance of having a backup of documents digitally to protect/ to reduce losses if something unexpected happens/ forces majeure.
4. Knowing the importance of data management as a stakeholder in decision making and being useful in determining development.

Recognizing that the implementation of the information system application monitoring of teaching and learning process will have a positive effect on all parties involved in the Bimbel Generasi Cerdas.

Data collection methods used in this study are:

1. Interview Methods.
   Researchers have prepared a list of questions relating to information gathering, to ask: teachers, guardians of students and management of the Bimbel Generasi Cerdas.
2. Observation Methods.
   Observation is an activity of direct observation of the profile of the organization and the object of research. The observation process was carried out to study data from the results of activities carried out and organizational archive documents, especially those managed by the Bimbel Generasi Cerdas, organizational goals and structure, business processes, availability of technological infrastructure, and information technology policies that exist in the Bimbel Generasi Cerdas.
3. Literature Study Methods.
   Researchers collect data by studying, researching, and reading books, journals, theses,which are related to the development of monitoring information systems that will be developed.

3.4. Technical Analysis Data and Systems

In the analysis process, the analysis techniques used are:

1. Analysis techniques approach to Object Oriented Analysis (OOA) or object-oriented analysis with UML. UML is a modeling language in development, in the field of software engineering to visualize system designs [14]. The system analysis process that will be developed is carried out on the results of the stages of data collection obtained, namely from the results of interviews, surveys, direct observations and literature studies conducted by researchers to obtain specifications of the system requirements which will be developed.
2. Analysis of Functional, Non-Functional Needs of Users

Table 1. Functional and Non Functional Table

Fuctional	
Needs Analysis	
No	**Management of Bimbel Generasi Cerdas wants this system to be able to:**
1	Display the login menu by entering the user name and password.

e-ISSN:2614-8404
p-ISSN:2776-3234

2	Display the main display menu as Admin.
3	Display the student data menu as Admin.
4	Display the teacher data menu as Admin.
5	Display the lesson data menu as Admin.
6	Display the schedule data menu as Admin.
7	Display the student grades menu as Admin.
8	Display the student attendance menu as Admin.
9	Display the admin's account settings menu.
10	Display the teaching schedule for Teacher.
11	Display the main display menu as a Teacher.
12	Display the input students' attendance menu as Teacher.
13	Display the input students' grades menu as Teacher.
14	Display the teacher's account setting menu.
15	Display the main display menu as Students.
16	Display the Students' grades view menu.
17	Display the Students' attendance view menu.
18	Display the Students' account settings menu.
Non Functional	
No	**Management of Bimbel Generasi Cerdas wants this system to be able to:**
1	Have an attractive application framework.
2	Have a user friendly application display.
3	Be a web based.

3.5.Research Steps

In this study, the authors conducted the stages of research using the waterfall system development model.

The waterfall method is often called the classic life cycle, where it illustrates a systematic and sequential approach to software development, starting with the specification of user needs and then continuing through the stages of planning, modeling, construction, as well as the delivery of the system to the customers/ users (deployment), which ends with support for the complete software produced [15].

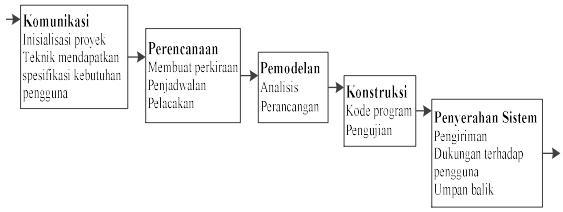


Figure 1. Model waterfall system development [15]

The research steps will go through several stages of the work process, namely:

1. Requirement Stage, the stage where the researcher determines the object of research and conducts research analysis.
2. System Design Stage, the stage where researchers conduct system design using UML, including use case diagrams, activity diagrams, sequence diagrams, state chart diagrams, class diagrams.
3. Implementation Stage, which is the stage where the researcher implements the system first developed in a small program, which is integrated in the next stage, each unit is developed and tested for its function called unit testing.
4. Verification Stage, which is the stage of the researcher verifying all the units making this system that was developed in the implementation stage, integrated into the system after testing each unit, to find out any failures or errors.

Maintenance Stage, namely the stage that the researcher carries out maintenance, and corrects errors or failures that have not been found in the previous step.

III. RESULTS AND DISCUSSION

3.1. Use Case Diagram

The use case diagram illustrates the workflow of the system in a very simple way, the main function of the system and the various types of users who will interact with the system, as in picture 2-4.

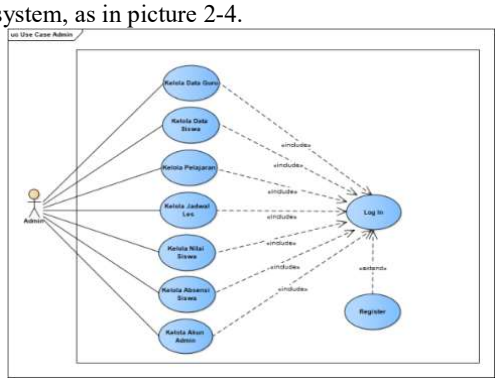


Figure 2. Admin use case diagram

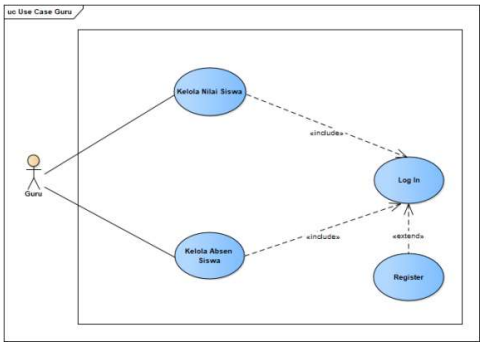


Figure 3. Teacher use case diagram

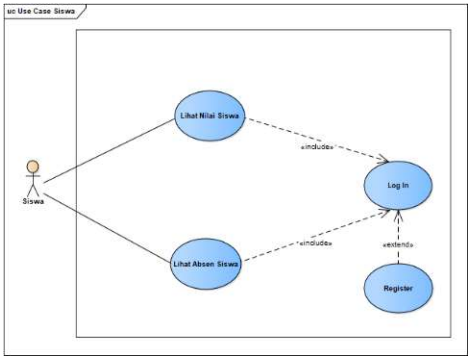


Figure 4. Student and Parent use case diagram

3.2.Sequence Diagram

Sequence diagrams are diagrams, that illustrate objects, participate in use cases and messages or information on activities carried out between them from

e-ISSN:2614-8404
p-ISSN:2776-3234

time to time for a use case. Figure 5 illustrates the sequence diagram of managing student grades by the admin actor.

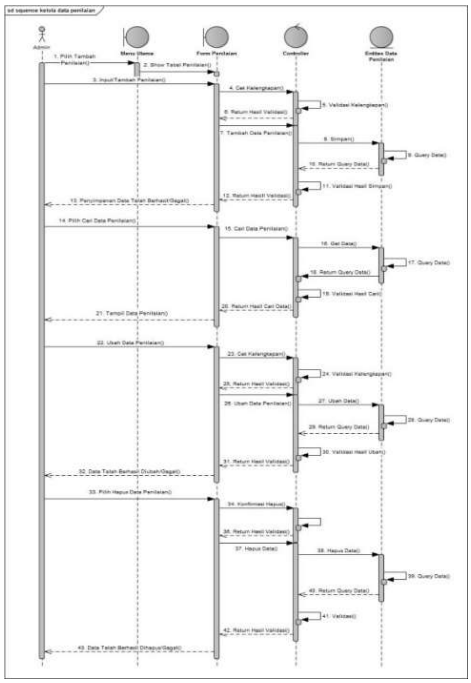


Figure 5. Sequence diagram system

3.3. Application Development Using Codeigniter

CodeIgniter's work process is very simple, if the user wants to access the application via a browser, then the steps are:

a) Each user requests the application, it will be directed to the index.php page.
b) Routing will determine the flow of requests from users. If the requested page is cached, the routing will perform step 3.
c) If the routing leads to caching, then the page displayed is the cached page.
d) If the routing points to security, then all data from the user will be filtered to increase security before being directed to the controller.
e) The controller will call the model, library, helper, and other tools needed for the application page requested by the user.
f) User requests will be displayed on the screen.

3.4. GUI Design

3.4.1. Log in Menu

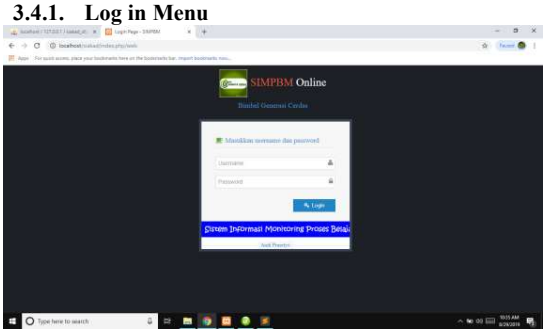


Figure 6. Log in Menu

This menu is the first menu, where the user must enter the appropriate user name and password that has been registered by the system administrator.

3.4.2. Dashboard Menu

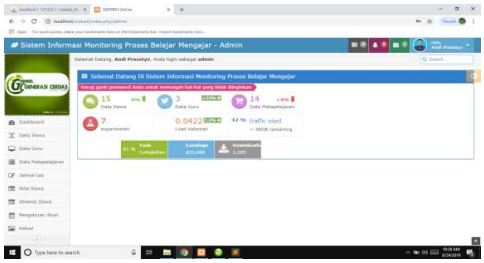


Figure 7. Dashboard Menu

This menu is what will view after the user has successfully logged into the system. In this menu, users can see the student data dashboard, subject teacher data, experimental data, the number of assignment pages and user traffic.

3.4.3. Student Data Menu

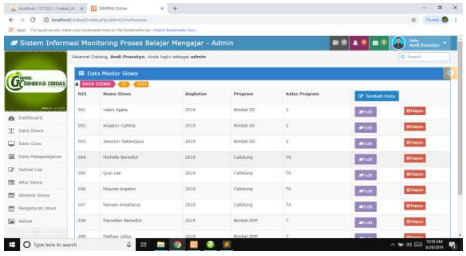


Figure 8. Student Data Menu

This menu is a menu that displays student data, data that can be seen by the user depends on the access rights the user has, if the user is a teacher, he will be able to see all the students he teaches, but if the user is a student or parent, then only data students or their child's data that can be seen by the user.

3.4.4. Managing Students Grades Menu

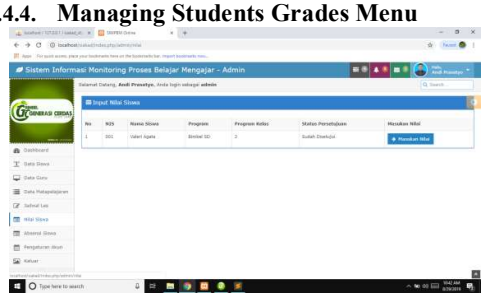


Figure 9. Managing Students grades Menu

This menu is a menu that can be used by the teacher to manage data on the students he teaches. Teachers who can view and manage student data are only teachers who teach the class of subjects they teach.

3.4.5. Managing Students Attendance Menu

e-ISSN:2614-8404
p-ISSN:2776-3234

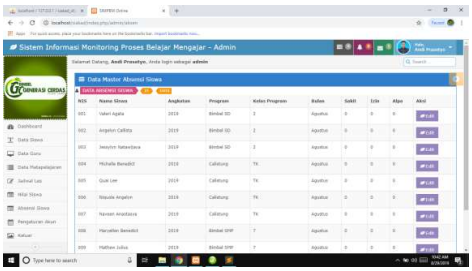


Figure 10. Managing Student Attendance Menu

This menu is a menu that displays student attendance data, data that can be seen by the user depending on the access rights the user has, if the user is a teacher, he will be able to see all the student absences he teaches, but if the user is a user. are students or parents, only the student or child absences data can be seen by the user.

3.5. Black Box Testing

System testing aims to assess the quality of the system based on black box testing with 17 test cases.

Table 2. Black Box testing

Application Name : System Information Learning Process Monitoring at Private Lessons Institution				Testing date: 10-01-2021	
				Tester: Andi	
	Page Tested	Actor Act	System React		Result
			True	False	
A.	**Admin Page System**				
1.	Admin Home page	Select Log in menu	Get the admin log in page	Failed to get admin log in page	valid
2.	Log in Admin	Input username and password	Get the admin main page	Failed to get admin main page	valid
3.	Log out	Select log out menu	Logged out and display the homepage	Failed to log out	valid
4.	Teachers' data	Add teacher data	Teacher data added	Display an error message if there is data that is not filled in	valid
		Change the teacher data according to the desired changes	The latest teacher data will be successfully saved	Display an error message if there is data that is not filled in	valid
		Delete the teacher data	Teacher data will be deleted	Teacher data will not be deleted	valid
5.	Students' data	Add student data	Student data added	Display an error message if there is data that is not filled in	valid
		Change student data	The latest student data will be successfully saved	Display an error message if there is data that is not filled in	valid
		Delete student data	Student data will be deleted	Student data will not be deleted	valid
6.	Subject data	Add subject data by entering all data	Subject data added	Display an error message if there is data that is not filled in	valid
		Change the subject data	The latest subject data will be successfully saved	Display an error message if there is data that is not filled in	valid
		Delete subject data	Subject data will be deleted	Subject data will not be deleted	valid
7.	Tutoring Schedule data	Add tutoring schedule data	Tutoring schedule data added	Display an error message if there is data that is not filled in	valid
		Change tutoring schedule data	The latest tutoring schedule data will be successfully saved	Display an error message if there is data that is not filled in	valid
		Delete tutoring schedule data	Tutoring schedule data will be deleted	Tutoring schedule data will not be deleted	valid
8.	Students' grades data	Add student grade data	Student grade added	Display an error message if there is data that is not filled in	valid
		Change student grade data	The latest student grade data will be successfully saved	Display an error message if there is data that is not filled in	valid
		Delete student grade data	Student grade data will be deleted	Student grade data will not be deleted	valid
9.	Students' Absence	Add student attendance data	Student attendance added	Display an error message if there is data that is not filled in	valid
		Change student attendance data	The latest student attendance data will be successfully saved	Display an error message if there is data that is not filled in	valid
		Delete student attendance	Student attendance will be deleted	Student attendance will not be deleted	valid
10.	Changing password	Change admin password	Password will be successfully changed	Password will not be changed	valid
B.	**Teacher Page System**				
1.	Teachers' Data	Add teacher data	Teacher data added	Display an error message if there is data	valid

e-ISSN:2614-8404
p-ISSN:2776-3234

				that is not filled in	
		Change the teacher data	The latest teacher data will be successfully saved	Display an error message if there is data that is not filled in	valid
		Delete teacher data	Teacher data will be deleted	Teacher data will not be deleted	valid
2.	Students grade data	Add student grade data	Student grade data added	Display an error message if there is data that is not filled in	valid
		Change student grade data	The latest student grade data will be successfully saved	Display an error message if there is data that is not filled in	valid
		Delete student grade data	Student grade data will be deleted	Student grade data will not be deleted	valid
3.	Students' attendance data	Add student attendance data	Student attendance data added	Display an error message if there is data that is not filled in	valid
		Change student attendance data	The latest student attendance data will be successfully saved	Display an error message if there is data that is not filled in	valid
		Delete student attendance data	Student attendance data will be deleted	Student attendance data will not be deleted	valid
4.	Changing password	Change teacher password	Password will be successfully changed	Password will not be changed	valid
C.	**Student Page System**				
1.	Students' grade data	See student grade data	Display student grade data	Student grade data will not be displayed	valid
2.	Students' attendance data	See student attendance data	Display student attendance data	Student attendance data will not be displayed	valid
3.	Changing password	Change student password	Password will be successfully changed	Password will not be changed	valid

IV. CONCLUSION

The design of the monitoring information system of student learning outcomes in the Bimbel Generasi Cerdas is the development of an ongoing system. Various problems that have arisen have been attempted to be handled with this proposed new system. The conclusions that can be drawn from the development of this information system include:

1. Information system monitoring teaching and learning process that was developed has been able to process data and present it into a useful information for students, teachers and parents of students.
2. The developed teaching and learning monitoring information system can process student grades and absences data quickly, precisely and accurately.

REFERENCES

[1] S. K. Boell and D. Cecez-Kecmanovic, "What is an Information System?," *IEEE 2015 48th Hawaii Int. Conf. Syst. Sci.*, pp. 4959–4968, 2015, doi: 10.1109/HICSS.2015.587.

[2] N. Fitriawati, A. Herdiansah, and A. Gunawan, "Sistem Informasi Program Keluarga Harapan Studi Kasus Kecamatan Kosambi Tangerang," *J. Tek. Inform. Univ. Muhammadiyah Tangerang*, vol. 3, no. 2, pp. 21–26, 2019.

[3] A. Herdiansah, N. Handayani, and A. Kurniawan, "Development of Decision Support Systems Selection of Employee Acceptance Using Weighted Product Method," *J. Inf. Syst. Informatics*, vol. 1, no. 2, pp. 87–97, 2019.

[4] A. Herdiansah, "Sistem Pendukung Keputusan Referensi Pemilihan Tujuan Jurusan Teknik di Perguruan Tinggi Bagi Siswa Kelas XII IPA Menggunakan Metode AHP," *J. MATRIK*, vol. 19, no. 2, pp. 223–234, 2020, doi: https://doi.org/10.30812/matrik.v19i2.579.

[5] A. Elsaadany and K. Abbas, "Development and Implementation of E-Learning System in Smart Educational Environment," *IEEE 39th Int. Conv. Inf. Commun. Technol. Electron. Microelectron.*, pp. 1004–1009, 2016, doi: 10.1109/MIPRO.2016.7522286.

[6] N. Komalasari, D. F. Murad, D. Agustine, M. Irsan, J. Budiman, and E. Fernando, "Effect of education, performance, position and information technology competency of information systems to performance of information system," *2018 Int. Semin. Res. Inf. Technol. Intell. Syst. ISRITI 2018*, pp. 221–226, 2018, doi: 10.1109/ISRITI.2018.8864437.

[7] K. Kavran and N. Majstorović, "Custom applications for information system for higher education in Croatia," *Proc. Elmar - Int. Symp. Electron. Mar.*, vol. 2015-Novem, no. September, pp. 238–241, 2015, doi: 10.1109/ELMAR.2015.7334537.

[8] M. Gounder, V. Iyer, and A. Al-Mazyad, "A Survey on Business Intelligence tools for University Dashboard Development," *IEEE - 3rd MEC Int. Conf. Big Data Smart City*, pp. 1–7, 2016, doi: 10.1109/ICBDSC.2016.7460347.

[9] A. Rudyanto, *Pemrograman Web Dinamis Menggunakan PHP dan MySQL*, Ed.1. Yogyakarta: Andi Offset, 2011.

[10] A. Kadir, *Buku Pintar Programmer Pemula PHP*, Ed.1. Yogyakarta: Mediakom, 2013.

[11] X. Li, S. Kaman, and J. Chishti, "An empirical study of three PHP frameworks," *IEEE 4th Int. Conf. Syst. Informatics*, pp. 1636–1640, 2017, doi: 10.1109/ICSAI.2017.8248546.

[12] R. Liatmaja and I. Wardati, "Sistem Informasi Akademik Berbasis Web pada Lembaga Bimbingan Belajar Be Excellent Pacitan," *Indones.*

e-ISSN:2614-8404
p-ISSN:2776-3234

J. Netw. Secur., vol. 2, no. 2, pp. 58–63, 2013.

[13] A. Rifai and H. Mustafidah, "Rancang Bangun Sistem Informasi Nilai Mata Pelajaran Berbasis Web dan SMS Gateway (Build-Up Lesson Score Information System Based on Web and SMS Gateway)," *J. JUITA*, vol. II, no. 4, pp. 239–248, 2013.

[14] M. . Dennis, A., Wixson, Haley, B., Roth, *System Analyst and Design*, 5th edd. USA: Don Fowley Publisher, 2012.

[15] R. Pressman, *Software Engineering: A Practitioner's Approach*, 7th editio. New York: The McGraw-Hill Company, 2015.

e-ISSN: 2614-8404
p-ISSN: 2776-3234

K-Means Cluster Analysis of Sex, Age, and Comorbidities in the Mortalities of Covid-19 Patients of Indonesian Navy Personnel

Bambang Suharjo[1*)], Muhammad Satria Yuda Utama[2]
[1]Magister Ilmu Komputer, Fakultas Teknologi Informasi, Universitas Budi Luhur
[2]Master of Data Science and Business Analytic, Faculty of Computing, Asia Pacific University of Technology and Innovation
Email: [1]bambang_suharjo@tnial.mil.id, [2]satria@gmail.com

Abstract – **Covid-19 disease is still ongoing. It is necessary to do intensive research related to age, sex and congenital diseases so that management can be better planned. The research was conducted using data from Indonesian Navy personnel and their families, retired Indonesian Navy and their families. This study used k-means clustering for data grouping of Indonesian Navy personnel based on age, sex and congenital disease characteristics. The results of the k-means cluster clustering show that the k = 2 cluster has not been able to provide an explanation of the relationship between age, sex and comorbidity with the risk of death due to Covid-19. However, in the cluster with k = 3, it turns out that deaths due to Covid-19 are related to old age, men, even though there is no congenital disease. Meanwhile, using the k = 4 cluster, it is increasingly clear that deaths due to Covid-19 are closely related to old age, both men and women, with comorbidities.**

Keywords – comorbidity, Mortality, Covid-19, K-Means Cluster

I. INTRODUCTION

History records that pandemics have occurred many times and claimed millions of lives. There have been many significant pandemics and have caused enormous negative impacts in various fields including health, economy and even national security in the world [1]. Coronavirus disease 2019 (corona virus disease / Covid-19) is a new name given by the World Health Organization (WHO) for patients with the 2019 corona virus infection. The disease caused by the corona 19 virus was first reported from the city of Wuhan, China by the end of 2019. This positive single-strain RNA virus occurs by infecting the human respiratory tract. This virus is sensitive to heat and can be effectively inactivated by disinfectants containing chlorine. The source of the Covid-19 virus is thought to have come from animals, especially bats, and other vectors such as bamboo mice, camels and weasels. Common symptoms due to exposure to the Covid-19 virus include fever, cough and difficulty breathing. Clinical syndromes that appear after exposure to the Covid-19 virus can be grouped into uncomplicated, mild pneumonia and severe pneumonia [2], [3]. The spread is rapid throughout the world and the threat of a new pandemic until early 2021 has yet to be contained [4].

In Indonesia, exposure to the covid-19 virus was discovered in early March 2020. The development of sufferers due to this virus continues to increase until the beginning of 2021, reaching more than 1 million sufferers. Even though vaccines are starting to be found and used against this virus, various research is still needed to address the threat of this new virus pandemic as a whole.

Various studies that have been conducted in Indonesia and other countries show a link between age, sex and comorbidity to the dangers of the covid-19 virus on safety and recovery [2], [5]. Comorbidity is a patient congenital disease before exposure to a disease. In exposure to the disease caused by the Covid-19 virus, based on various studies, many comorbidities are believed to be dangerous for patient safety [6], [7], [8], [9], [10].

Comorbidity puts Covid-19 patients into vicious cycle of life and is strongly associated with significant morbidity and mortality. Comorbid individuals must adopt vigilant precautions and require careful management [6]. In the study, patients with the characteristics of old age, male, and critical illness were at increased risk of death compared to patients with other conditions at younger age, women and had no comorbidities Covid-19 patients with diabetes, chronic lung disease, cardiovascular disease, hypertension, HIV and other comorbidities may develop life-threatening situations [6], [7], [8], [9], [10].

This study will reveal the grouping of patients due to exposure to the Covid-19 virus in the Indonesian Navy with the characteristics of age, sex and comorbidity. Clustering can be used to partition data into groups, or clusters. A cluster can be described as a group of data objects that are more similar to other objects in their cluster than to data objects in other clusters [11], [12], [13].

Some previous research shown researches about clustering to analyze covid-19, age, gender, and comorbidities, such as:

The research aims to assess the relationship between sex, age, and comorbidity and mortality in Covid-2019 patients using clustering. The conclusion obtained is that gender, age, and comorbidities are partially related to the risk of death from Covid -19 [14]. Next research on the use of the k-means clustering of covid data obtained from Kaggle resulted in clusters with different levels of sufferers and deaths. With these results, it is recommended to carry

e-ISSN: 2614-8404
p-ISSN: 2776-3234

out different treatments in areas in different clusters [15]. Another study conducted grouping of districts and cities in Central Java, Indonesia based on Covid -19 cases using k-means clustering. The results show that two of the 3 clusters are areas that must be considered by the government because they are areas with a high number of active cases and high cases of Covid -19 deaths [16].

In another study, Artificial Neural Networks and k-means cluster were used on data on the current situation of the spread of Covid-19 in Indonesia for clustering. The resulting clusters consist of many provincial clusters in Indonesia with groupings on the characteristics of positive growth, recovery and death [17], [18], [19].

Based on these previous studies, it appears that clustering of Covid-19 sufferers is important. Thus, it needs to be made more comprehensive by adding patient comorbidities to the analysis. The results of these research are expected to be able to reveal more detailed characteristics about sufferers of exposure to the covid-19 virus. This is important to do to support the current pandemic policy model and its future anticipation.

II. RESEARCH METHODOLOGY

A. Data

The data used in this study were Covid-19 patient data, including: Indonesian Navy personnel and their families, retired members of the Indonesian Navy and their families. Data collected from March 2020 to December 2020. Data was collected using the census method, covering all of these data on those who suffer from Covid-19 which were obtained from the Indonesian Navy Health Service database.

B. Research Steps

The research steps were carried out as follows:

1. Data is collected from the Indonesian Navy covid-19 database
2. Cleaning data by eliminating the variables of the patient's name, unit origin, and the relationship with members of the Indonesian Navy if the family.
3. Data transformation
4. Create a descriptive analysis
5. Choose the best number of clusters using the elbow method.
6. Perform clustering calculations using the k-means method for the best clusters according to the elbow method
7. Analysis of the characteristics of the resulting clusters
8. Perform a comparison analysis of the characteristics of the clustering analysis results.

C. Flowchart Diagram

Based on the research steps above, the research flowchart was compiled as follows.

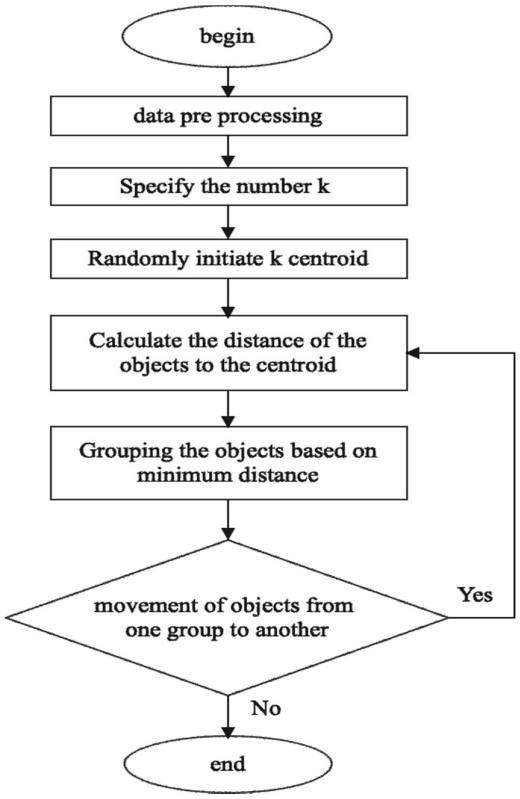


Figure 1. Flowchart diagram of the research

III. RESULTS AND DISCUSSION

A. Descriptive of Indonesian Navy Covid-19 Patients

Data on patients with Covid-19 as a whole can be described starting from gender (male and female), results of hospital treatment or independent isolation (recovering and dying), comorbidity (heart, lung and diabetes), presented in Figure 2, as following.

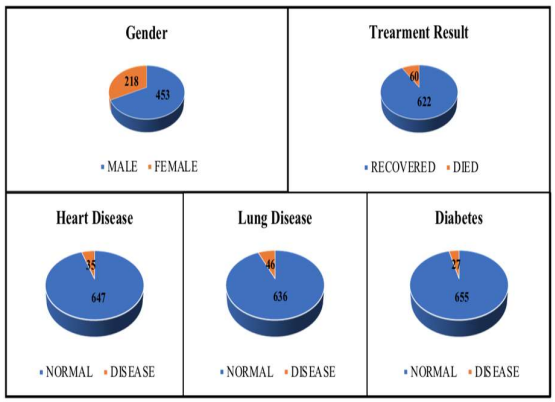


Figure 2. Distribution of Indonesian Navy Covid-19 sufferers based on gender (male and female), treatment results, and comorbidities.

Based on the pie chart above, it appears that the number of Indonesian Navy sufferers of Covid-19 from

e-ISSN: 2614-8404
p-ISSN: 2776-3234

March 2020 to December 2020 reached 682 people, with the breakdown of men = 453 people, 218 women. Meanwhile, 622 people were recovered from the treatment and 60 people died. The co-morbidities of the sufferers included: 35 heart disease, 46 lung disease and 27 diabetes. In addition, the age of the patients was between 4 years old until 88 years old and the mean was 36.8 years.

B. Cluster Analysis using k-mean

To get the size of the number of good clusters, it is necessary to select k. One way of selecting k can be done with the elbow method. The pseudocode for selecting k with Python language is as follows.

```
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
from kneed import KneeLocator
from sklearn.cluster import KMeans
sse = []
for i in range(1, 11):
    kmeans = KMeans(n_clusters = i, init = 'k-means++', random_state = 42)
    kmeans.fit(TRANSF_DATA)
    sse.append(kmeans.inertia_)
plt.plot(range(1, 11), sse, marker="o", color="red")
plt.title('Elbow METHOD')
plt.xlabel('Number of clusters')
plt.ylabel('SSE')
plt.show()
kl = KneeLocator(range(1, 11), sse, curve="convex", direction="decreasing")
print()
print('BEST K USING ELBOW METHOD IS','"', kl.elbow, '"')
```

Figure 3. Pseudocode of elbow method

The output results from the pseudocode above, then the elbow plot can be presented in accordance with the figure 4, as follow

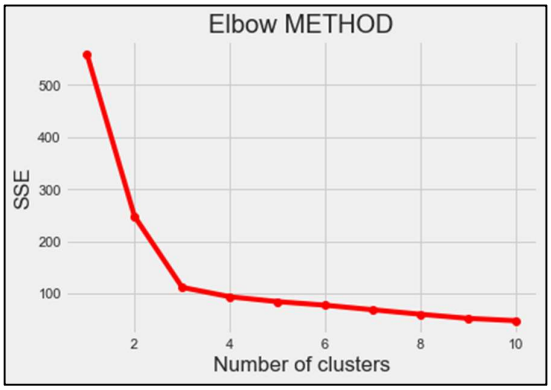


Figure 4. Elbow plot of best number of clusters

From the elbow plot, it can be seen that the elbows are at k = 2 to 4. Thus, the three options (k = 2, k = 3, and k = 4) are a priority for clustering calculations. Furthermore, the pseudocode used to carry out the clustering process using the k-means cluster method and plot the clustering results on the number of clusters 2, 3 and 4 which are carried out using Python, as follows

```
X = x_scaled
kmeans = KMeans(n_clusters = 2, init = 'k-means++', random_state = 42)
y_kmeans = kmeans.fit_predict(X)
print(y_kmeans)
plt.scatter(X[y_kmeans == 0, 0], X[y_kmeans == 0, 1], s = 100, c = 'red', label = 'Cluster 1')
plt.scatter(X[y_kmeans == 1, 0], X[y_kmeans == 1, 1], s = 100, c = 'blue', label = 'Cluster 2')
plt.scatter(kmeans.cluster_centers_[:, 0], kmeans.cluster_centers_[:, 1], s = 300, c = 'black',
            label = 'Centroids')
plt.title('CLUSTERS (K=2) OF INDONESIAN NAVY COVID 19')
plt.xlabel('CRITERIA')
plt.ylabel('RESULT')
plt.legend()
plt.show()
```

Figure 5. Pseudocode of clustering plots

The output result of pseudocode in Figure 6 can be presented as follows.

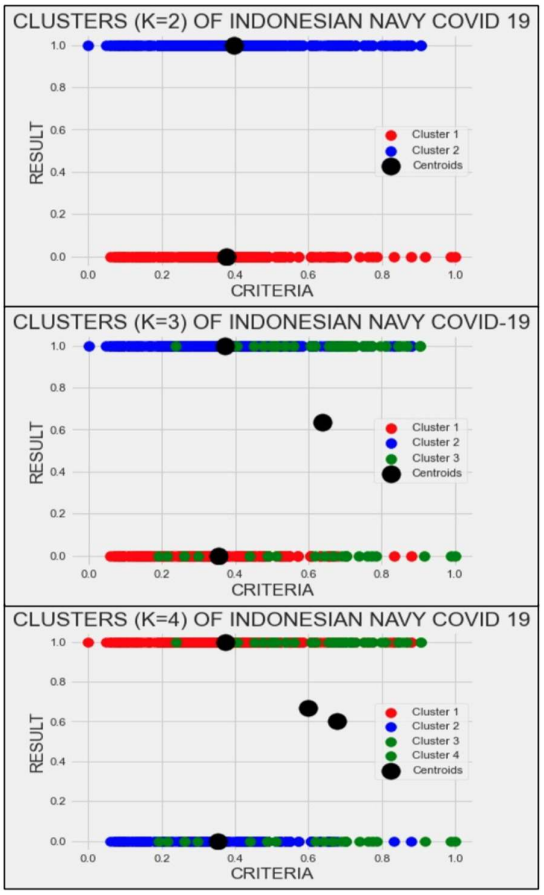


Figure 6. Plot of Objects clustering

Using centroid of every cluster, we can conclude that the characteristic of every kind of clustering, can be shown in table 1,2, and 3 below.

Table 1. Characteristic of Covid-19 of Indonesian navy personnel using 2 clusters

	Characteristic	**Percentage of member**
Cluster 1	Younger, dominated by female, dominated by normal heart, dominated by normal lung, and dominated by normal diabetes, and dominated by recovered.	32.7%
Cluster 2	Older, dominated by male, dominated by normal heart, dominated by normal lung, and dominated by normal diabetes, and dominated by recovered.	67.3%

e-ISSN: 2614-8404
p-ISSN: 2776-3234

Based on the description of the characteristics of Indonesia n Navy personnel exposed to covid-19, young or old, most recovered and a small proportion died. However, it cannot be concluded that comorbid disorders are associated with mortality. So that clustering with cluster= 2 has not been a ble to explain in detail the comorbidity and mortality due t o covid-19.

Table 2. Characteristic of Covid-19 of Indonesian navy personnel using 3 clusters

	Characteristic	**Percentage of member**
Cluster 1	Young, dominated by women, dominated by normal heart, dominated by normal lungs, normal and dominated by normal diabetes, and recovered	29.4%
Cluster 2	Older, male, dominated by normal heart, dominated by normal lungs, normal and dominated by normal diabetes, and recovered	61.6%
Cluster 3	Old, dominated by men, dominated by normal heart, dominated by normal lungs, donimated by normal diabetes and died	8.8%

The results of the clustering according to table 2 show that a high risk of death is closely related to old age and men, with or without comorbidities. The conclusion from clustering into 3 clusters has not shown a detailed conclusion. Thus, it needs to be developed into 4 clusters.

Table 3. Characteristic of Covid-19 of Indonesian navy personnel using 4 clusters

	Characteristic	**Percentage of member**
Cluster 1	Young, male, dominated by no congenital disease, recovered	61.7%
Cluster 2	Young, female, dominated by no congenital disease, recovered	29.5%
Cluster 3	Elderly, male-dominated, some have cardiac comorbidities, mostly pulmonary comorbidities, no diabetic comorbidities, died	4.4%
Cluster 4	Elderly, predominantly male, some have cardiac comorbidities, no pulmonary comorbids, some have diabetes comorbidities, died	4.4%

Based on the clustering in table 3 which contains 4 clusters, it appears that the risk of death from covid-19 will be high if the elderly, male or female, with cardiovascular, lung and diabetes comorbidities. Therefore, it is necessary to be more careful, more intensive treatment in patients with these characteristics.

IV. CONCLUSION

Based on the results of the k-means cluster clustering, it appears that the k = 2 cluster has not been able to provide an explanation of the relationship between age, sex and comorbidity with the risk of death due to covid-19. However, in clusters with k = 3, it appears that deaths from covid-19 are related to older age, men, even though there is no congenital disease. Meanwhile, using the k = 4 cluster, it is increasingly clear that deaths from covid-19 are closely related to older age, both men and women, with comorbidities.

REFERENCES

[1] Qiu, W., Rutherford, S., Mao, A., & Chu, C. (2017). The Pandemic and its Impacts. *Health, Culture and Society*, *9*, 1–11. https://doi.org/10.5195/hcs.2017.221

[2] Yuliana. (2020). Corona virus diseases (Covid -19); Sebuah tinjauan literatur. *Wellness and Healthy Magazine*, *2*(1), 187–192.

[3] Anjorin, A. A. (2020). The coronavirus disease 2019 (COVID-19) pandemic : A review and an update on cases in Africa The coronavirus disease 2019 (COVID-19) pandemic : A review and an update on cases in Africa. *Asian Pacific Journal of Tropical Medicine*, *13*(April). https://doi.org/10.4103/1995-7645.281612

[4] Handayani, D., Hadi, D. R., Isbaniah, F., Burhan, E., & Agustin, H. (2020). Penyakit Virus Corona 2019. *Jurnal Respirologi Indonesia*, *40*(2), 119–129.

[5] Putri, R. N. (2020). Indonesia dalam Menghadapi Pandemi Covid-19. *Jurnal Ilmiah Universitas Batanghari Jambi*, *20*(2), 705. https://doi.org/10.33087/jiubj.v20i2.1010

[6] Filardo, T. D., Khan, M. R., Krawczyk, N., Galitzer, H., Karmen-Tuohy, S., Coffee, M., Schaye, V. E., Eckhardt, B. J., & Cohen, G. M. (2020). Comorbidity and clinical factors associated with COVID-19 critical illness and mortality at a large public hospital in New York City in the early phase of the pandemic (March-April 2020). *PLoS ONE*, *15*(11 November), 1–16. https://doi.org/10.1371/journal.pone.0242760

[7] Ejaz, H., Alsrhani, A., Zafar, A., Javed, H., Junaid, K., Abdalla, A. E., Abosalif, K. O. A., Ahmed, Z., &

e-ISSN: 2614-8404
p-ISSN: 2776-3234

Younas, S. (2020). COVID-19 and comorbidities: Deleterious impact on infected patients. *Journal of Infection and Public Health*, *13*(12), 1833–1839. https://doi.org/10.1016/j.jiph.2020.07.014

[8] Gold, M. S., Sehayek, D., Gabrielli, S., Zhang, X., McCusker, C., & Ben-Shoshan, M. (2020). COVID-19 and comorbidities: a systematic review and meta-analysis. *Postgraduate Medicine*, *132*(8), 1–7. https://doi.org/10.1080/00325481.2020.1786964

[9] Kun'ain, U. I. A., Rahardjo, S. S., & Tamtomo, D. G. (2020). Meta-Analysis: The Effect of Diabetes Mellitus Comorbidity on the Risk of Death in Covid-19 Patients. *Indonesian Journal of Medicine*, *5*(4), 368–377. https://doi.org/10.26911/theijmed.2020.05.04.12.

[10] Yang, J., Zheng, Y., Gou, X., Pu, K., Chen, Z., Guo, Q., Ji, R., Wang, H., Wang, Y., & Zhou, Y. (2020). Prevalence of comorbidities and its effects in coronavirus disease 2019 patients: A systematic review and meta-analysis. *International Journal of Infectious Diseases*, *94*, 91–95. https://doi.org/10.1016/j.ijid.2020.03.017

[11] Frigui, H. (2008). Cluster Analysis: Basic Concepts and Algorithms. In *2008 1st International Workshops on Image Processing Theory, Tools and Applications, IPTA 2008*. https://doi.org/10.1109/IPTA.2008.4743793

[12] Kumar, M., & Verma, A. (2018). Clustering Techniques - A Review. *International Journal of Computer Sciences and Engineering*, *6*(6), 1091–1099. https://doi.org/10.26438/ijcse/v6i6.10911099

[13] Thrun, M. (2018). Approaches to Cluster Analysis. In *Projection-Based Clustering through Self-Organization and Swarm Intelligence* (pp. 21–31). https://doi.org/10.1007/978-3-658-20540-9_3

[14] Biswas, M., Rahaman, S., Biswas, T. K., Haque, Z., & Ibrahim, B. (2021). Association of Sex, Age, and Comorbidities with Mortality in COVID-19 Patients: A Systematic Review and Meta-Analysis. *Intervirology*, *64*(1), 36–47. https://doi.org/10.1159/000512592

[15] Indraputra, R. A., & Fitriana, R. (2020). K-Means Clustering Data COVID-19. *Jurnal Teknik Industri*, *10*(3), 275–282.

[16] Mahmudan, A. (2020). Clustering of District or City in Central Java Based COVID-19 Case Using K-Means Clustering. *Jurnal Matematika, Statistika Dan Komputasi*, *17*(1), 1–13. https://doi.org/10.20956/jmsk.v17i1.10727

[17] Khotimah, T., & Darsin. (2020). Clustering Perkembangan Kasus Covid-19 di Indonesia Menggunakan Self Organizing Map. *Jurnal Dialektika Informatika (Detika)*, *1*(1), 23–26.

[18] R, R. P., & E, Y. A. (2020). *Analisis Cluster Virus Corona (COVID-19) di Indonesia pada 2 Maret 2020 – 12 April 2020 dengan Metode K-Means Clustering*. *May*, 1–6.

[19] Virgantari, F., & Faridhan, Y. E. (2020). K-Means Clustering of COVID-19 Cases in Indonesia s Provinces. *Proceedings of the International Conference on Global Optimization and Its Applications*.

e-ISSN: 2614-8404
p-ISSN: 2776-3234

Optimization of Support Vector Machine Method Using Feature Selection to Improve Classification Results

Saikin *), Sofiansyah Fadli[2], Maulana Ashari[3]
[1,3]Program Studi Sistem Informasi, STMIK Lombok
[2]Program Studi Teknik Informatika, STMIK Lombok
Email: [1]eken.apache@gmail.com, [2]sofiansyah182@gmail.com, [3]aarydarkmaul@gmail.com

Abstract – **The performance of the organizations or companiesare based on the qualities possessed by their employee. Both of good or bad employee performance will have an impact on productivity and the impact of profits obtained by the company. Support Vector Machine (SVM) is a machine learning method based on statistical learning theory and can solve high non-linearity, regression, etc. In machine learning, the optimization model is a part for improving the accuracy of the model for data learning. Several techniques are used, one of which is feature selection, namely reducing data dimensions so that it can reduce computation in data modeling. This study aims to apply the method of machine learning to the employee data of the Bank Rakyat Indonesia (BRI) company, so that it can improve the performance of the classification algorithm by removing some features that have no correlation to the objective label and have a significant effect on the classification results. The method used is SVM method by increasing the accuracy of learning data by using a feature selection technique using a wrapper algorithm. From the results of the classification test, the average accuracy obtained is 72 percent with a precision value of 71 and the recall value is rounded off to 72 percent, with a combination of SVM and cross-validation. Data obtained from Kaggle data, which consists of training data and testing data. each consisting of 30 columns and 22005 rows in the training data and testing data consisting of 29 col-umns and 6000 rows. The results of this study get a classification score of 82 percent. The precision value obtained is rounded off to 82 percent, a recall of 86 percent and an f1-score of 81 percent**.

Keywords – K-Fold Algorithm;SVM Method; Classification; Machine Learning

I. INTRODUCTION

The performance of the organizations or companies are based on the qualities possessed by their employee. both of good or bad employee performance will have an impact on productivity and the impact of the benefits that are obtained by the company. Machine learning enables companies to measure employee performance based on Key Performance Indicator (KPI), by applying learning to historical data, making it easier for companies to make decisions. Machine learning algorithms are often used to be solutions to solve problems related to data learning. Algorithms used are classification, clustering and regression algorithms.

Support Vector Machine (SVM) is a machine learning method based on statistical learning theory and can solve high nonlinear, regression, etc. in the sample space and can also be used as a predictive system identification tool [1]. This algorithm is also flexible, it can be applied to the field of data modeling where data classification and data analysis are regression in nature. SVM is an algorithm for making predictions, both predictions in regression and classification cases. which is how it works to get the optimal separator function (hyperplane) to separate observations that have different target variable values, from the concept promoted by this SVM algorithm which makes SVM work well on high-dimensional data sets, even SVM also uses kernel techniques to map the original data from the original dimension to another dimension which is relatively higher [2]. Prediction using SVM is very sensitive to the value of the parameters, being the soft-marginal value constant C of various kernel parameters [3].

In machine learing, model optimization is part of improving the results of model accuracy for data learning. Several techniques are used, one of which is feature selection, namely reducing data dimensions so that it can reduce computation in data modeling. The main purpose of feature selection is to reduce the number of features used in the classification while maintaining an acceptable classification accuracy [4]. Feature selection can have a big impact on the effectiveness of the resulting classification algorithm [5], in some cases, as a result of feature selection, the accuracy of future classification can be improved [6].

One of the algorithms used for feature selection is the Wrapper method. The wrapper method is a method of selecting features as a blackbox to find the best sub-set of attributes. in a previous study [7] in the form of "Combination of the Correlated Naive Bayes Method and the Wrapper Feature Selection Method for Health Data Classification". The aim of this study was to combine the Correlated Naive Bayes method and Wrapper-based feature selection for health data classification. The stages of this research consisted of several stages, namely (1) collecting the Pima Indian Diabetes and Thyroid dataset from the UCI Machine Learning Repository, (2) pre-processing data such as transformation, scaling, and Wrapper-based feature selection, (3) classification using Correlated Naive Bayes and Wrapper Feature Selection Method, and (4) performance testing based on its accuracy using 10-fold cross validation method. Based on the results of the tests

e-ISSN: 2614-8404
p-ISSN: 2776-3234

that have been carried out, the combination of the Correlated Naive Bayes method with Wrapper-based feature selection gets the best accuracy of both the data used. For the Pima Indian Diabetes dataset, the accuracy is 71.4% and the Thyroid dataset accuracy is 79.38%. Thus, the combination of the Correlated Naive Bayes method and Wrapper-based feature selection resulted in better accuracy without feature selection with an increase of 4.1% for the Pima Indian Diabetes dataset and 0.48% for the Thyroid dataset.

Based on previous research, this study aims to apply the machine learning method to the employee data of the Bank Rakyat Indonesia (BRI) company [19]. The method used is the Support Vector Machine method by increasing the accuracy of data learning with feature selection techniques using the wrapper algorithm. Data obtained from Kaggle data, which consists of training data and testing data. each of which consists of 30 columns and 22005 rows of training data and data testing consisting of 29 columns and 6000 rows. modeling results will predict employees into the best performance class and not.

II. RESEARCH METHODOLOGY

A. Literature Review

Research conducted by [8] with the research title "Seleksi Fitur Warna Citra digital Biji Kopi MenggunaanMetode Principal Component Analysis". The results of this study show that the average training process on feature data after the feature selection process has increased compared to without feature selection. This can be seen from the 5 times the training process with feature selection, the best accuracy value is 90.8%, while without feature selection the best accuracy is 89.6%. in a study conducted by [9] entitled " Principal Component Analysis Support Vector Machine (PCA-SVM) untukklasifikasi Kesejahtraaan Rumah Tanggak di Kabupaten Brebes" The results of household poverty classification in Brebes Regency use PCA-SVM with the RBF kernel approach to training data. obtained 667 respondents who were classified appropriately. There are 93 households classified as not poor and 574 households classified as poor.

In the testing data, there were 285 respondents who were classified appropriately. There are 35 households classified as not poor and 250 households classified as poor. in a study conducted [9] entitled " Seleksi Fitur Dan Optimasi Parameter K-NN BerbasisAlgoritmaGenitika Pada Dataset Medis". This study concludes that a genetic algorithm is applied to select features and optimize k parameters for k's closest neighbors to improve the accuracy of the five medical data sets used as benchmarks. The proposed method proved to be effective in increasing accuracy, and the difference in test results between the five datasets resulted in a significant difference.

Support Vector Machine (SVM) was developed by Vapnik in 1992 together with Bernhard Boser and Isabelle Guyon [10]. SVM is a machine learning method that performs a technique to find a classifier function that can split data into two different classes. 11]. The strategy used is to minimize errors in training data and the Vapnik-Chervokinensis (VC) dimension called Structural Risk Minimization (SRM). The aim of SVM is to get the best hyper-plane separating the two classes [12]. Getting the best hyperplane is the same as maximizing the distance between the hyperplane with the closest pattern from each class (margin). The advantage of the SVM method is its generalizability, which is the ability to classify other data that is not included in the data used in machine learning [13]. Feature selection is a process that involves a subset of feature sets that produce output such as the entire feature set. Feature selection is usually used to select optimal features, reduce dimensions, increase accuracy of classification algorithms, and remove irrelevant features [14]. The feature selection technique is divided into 3 groups namely Filter, Wrapper, and Embedded [15]. Filter-based feature selection research was carried out by [16].

B. Method

The stages of this research were carried out from the stage of collecting the dataset, processing the classification data to testing the classification results. The details of the research stages are drawn as shown in Figure 1.

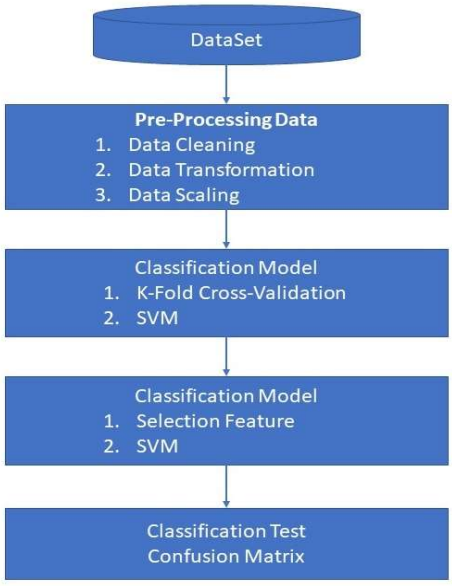


Figure 1. Research Flow

C. Data Retrieval

Data and research variables are secondary data taken from the Kaggle website, in the form of BRI bank employee data consisting of training data and testing data. training data consists of 28 columns and 22005 rows, and testing data consists of 28 columns and 6000 rows. The following displays the features for training data.

Table 1. Dataset Features

	Variabel
1	job_level
2	job_duration_in_current_job_level
3	person_level
4	job_duration_in_current_person_level
5	job_duration_in_current_branch
6	Employee_type
7	Employee_status
8	Gender
9	Age
10	marital_status_maried(Y/N)

e-ISSN: 2614-8404
p-ISSN: 2776-3234

11	number_of_dependences
12	number_of_dependences (male)
13	number_of_dependences (female)
14	Education_level
15	GPA
16	year_graduated
17	job_duration_as_permanent_worker
18	job_duration_from_training
19	branch_rotation
20	job_rotation
21	assign_of_otherposition
22	annual leave
23	sick_leaves
24	Best Performance
25	Avg_achievement_%
26	Last_achievement_%
27	Achievement_above_100%_during3quartal
28	achievement_target_1
29	achievement_target_2
30	achievement_target_3

D. Pre-Processing Data

Data cleaning is the process of data cleaning from several inconsistent data values. Its main purpose is to eliminate misinformation related to data. Data cleaning is carried out on data with high outlier values and data on empty and nan value data. empty data is used for the following, which is a display of missing data based on the highest order:

	Mising Values	% of Total Values
achievement_target_3	6727	30.570325
achievement_target_2	6727	30.570325
achievement_target_1	6727	30.570325
Achievement_above_100%_during3quartal	6302	28.638946
Last_achievement_%	6302	28.638946
Avg_achievement_%	6289	28.579868
Education_level	3608	16.396274
GPA	3503	15.919109
year_graduated	3503	15.919109
job_duration_as_permanent_worker	2055	9.338787
Employee_type	12	0.054533

Figure 2. Data Missing Values

E. Data Transformation

The classification method works with numeric data types, data transformation is used to change the form of data types other than Number into number data types. In the data set used in this study, there are several features that are object data types or categorical data types.

F. Scaling

The use of scaling to reduce the dominance of features whose range values are higher than features with smaller ranges. The use of scaling will affect the classification results because there will be no features that have a dominant value in the classification results, all will be calculated based on a maximum range value of 1 and a minimum. In the dataset, there are several features whose range values are higher than other features.

G. K-Fold Cross-Validation

Determination of training data and test data using the K-fold cross-validation technique where the amount of K is used, namely 4. where using 4 K, will divide the data into 3 subsamples into training and 1 sub sample into testing data. The following illustrates the distribution of training data and testing data using 4K.

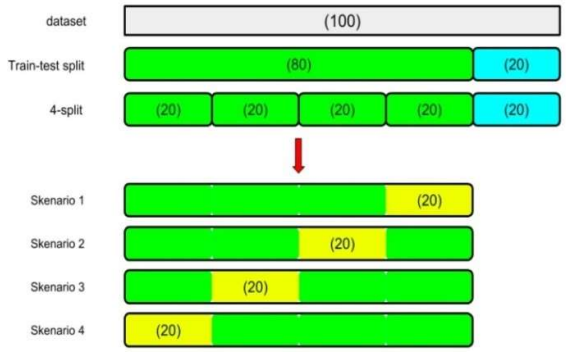


Figure 3. K-Fold Cross-Validation Scenario

H. Feature Selection

Feature selection is used here by looking for correlations that have a high impact on the classification results. features that do not have a significant effect on the classification result will be discarded. The method used is the analysis of variance (ANOVA), which is a statistical method that functions to test the significant effect on the average between groups of variables. The results of data visualization, there are several features that do not have a significant effect on the classification results such as Achievement_above_100% _during3quartal, Best Performance, sick_leaves, GPA, Education_level and gender.

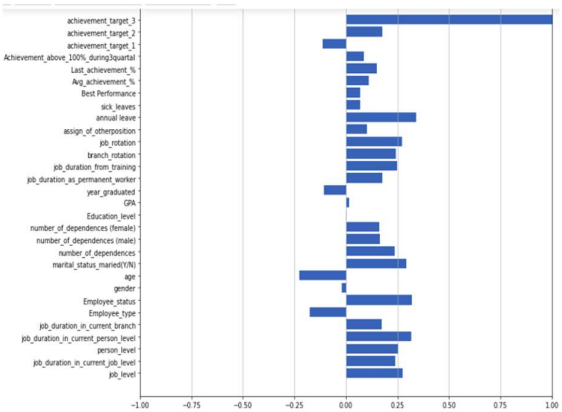


Figure 4. Features and Targets Correlation

I. Support Vector Machine

The fourth step is classification using SVM. The SVM algorithm has the ability to analyze data and perform pattern recognition [12]. SVM does its job by searching for the best hyperline, where what is meant by hyperline is the

e-ISSN: 2614-8404
p-ISSN: 2776-3234

dividing line between two classes [1]. The bigger the margin or dividing line, the smaller the level of misclassification that occurs [24]. SVM is also known as using the parameter penaltykernel method, the kernel contained in SVM such as kernellinear, radial Basis Function and Polynomial. In this study, kernelliniear will be used and the number of parameter penalties of C = 1.0.

III. RESULTS AND DISCUSSION

A. Classification Results Testing

Testing the results of classification using confusionmatrix is to look for true positive, true negative, false positive and false negative values. By applying two stages, namely the first stage by testing the classification results of the combination of the SVM algorithm with cross-validation. The second stage is to test the SVM classification results from feature selection.

B. SVM and K-Fold Cross-Validation Testing

In the classification using the SVM linear kernel and the parameter value C = 1.0 and using 4-fold cross-validation, the resulting average accuracy value is 72 percent.

```
=======================Score SVM dengan 4-fold cross-validation===
[0.72952012 0.72370334 0.73291323 0.72678788]
==========================================================



==================================Rata-rata Score===============
0.7282311432306585
==========================================================
```

Figure 5. SVM Classification with 4-Fold Cross-Validation

The results of the classification test using confusion matrix for classification with 4-fold cross-validation show that the predicted true positive value (true as best performance) is 1939, while the true negative value (correct prediction is no best performance) is 2105. While the false negative value (which is wrongly predicted as (best performance) is 831 and false positive (wrongly predicted as no best performance) is 673.

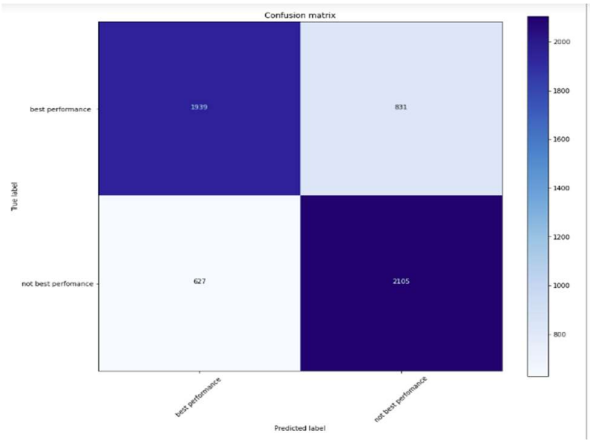


Figure 6. Confusion Matrix for SVM 4-Fold Cross-validation classification

C. SVM Testing With Feature Selection

In the SVM test with a 4-fold cross-validation score, an average of 72 percent was obtained, while using a combination of SVM with feature selection resulted in 82 percent, an increase of about 10 percent. The precision value obtained is rounded off to 82 percent, 86 percent recall and 81 percent f1-score.

```
==========================================================
Hasil klasifikasi SVM dengan Seleksi fitur = 0.8167605889838211
==========================================================
Nilai accuracy        = 0.8158851326790258
Nilai Precision Score = 0.8194404055703874
Nilai Recall          = 0.8667642752562226
nilai F1 Score        = 0.8154561244293783
==========================================================
```

Figure 7. SVM Score ValueandFeature Selection

The test results use confusion matrix, the predicted true positive value is 2121 best performance, and the true negative value is 2368. Meanwhile, the true positive value is 268 while the false negative value is 649.

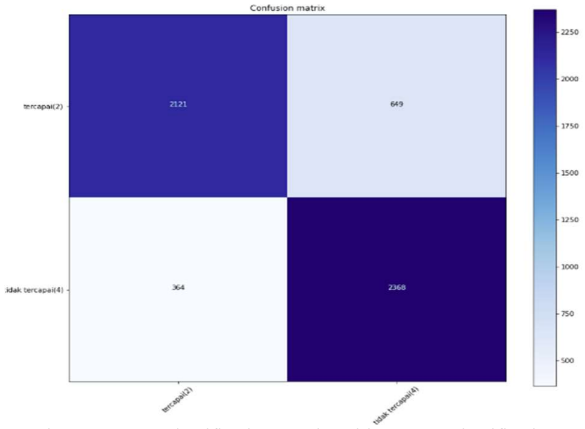


Figure 8. SVM Classification Matrix With Feature Classification

D. Discussion

The results of the confusion matrix measurement show a false positive value and false shows a high enough value. Testing the results of classification using confusion matrix is to find true positive, true negative, false positive and false negative values. By expecting two stages, namely the first stage by testing the classification results of the combination of the SVM algorithm with cross-validation. The second stage is to test the SVM classification results from feature selection.

IV. CONCLUSION

From the results of the classification test, the average accuracy obtained is 72 percent with a precision value of 71 and the recall value is rounded off to 72 percent, with a combination of SVM and cross-validation. To improve the accuracy, a feature selection experiment was carried out and searched for several features by looking for high correlation values and removing some features with low correlation values to the target data. The results obtained from the SVM modeling trials with feature selection obtained accuracy rounded to 82 percent. The accuracy value is 81 percent, the precision value is 82, the recall value is 86 percent and the fi-score is 81 percent. Confusion matrix testing results in true positive 2121 best performance, true negative 2368, true positive 268 while false negative is 649.By applying feature selection to the SVM algorithm, the accuracy value increased from 72 percent to 82 percent, an average increase of 10 percent.

e-ISSN: 2614-8404
p-ISSN: 2776-3234

REFERENCES

[1] Agustian Noor. (2018). *Perbandingan algoritma Support Vector machine biasa dengan support vector mechine berbasis particale swarm optimization untuk prediksi gempa bumi.* Jurnal Humaniora dan Teknologi.DOI:10.34128/jht.v4i1.37.

[2] Prasetyo. (2014). *Data Mining Mengolah data menjadi informasi*. Andi. Yogyakarta

[3] Ultach Enri. (2018). *Optimasi Parameter Support Vector Machine Untuk Prediksi Nilai Tukar Rupiah Terhadap Dolar Amerika.* Jurnal Gerbang.Vol 8 No 1.

[4] Raymer, M. L. Punch, W. F., Goodman, E. D., Kuhn, L. A., & Jain, A. K. (2000). *Dimensionality reduction using genetic algorithms.* IEEE Transactionson Evolutionary Computation, 4(2), 164-171.

[5] Jain, A., & Zongker, D. (1997). *Feature Selection: Evaluation, Application and Small Sample Performance.* IEEE Transactions on Pattern Analysis andMachine Intelligence. 19(2). 153-158.

[6] Maimon, O., & Rokach, L. (2010). *Data Mining and Knowledge Discovery Handbook (Second Edition ed.).* New York: Springer.

[7] Hairani., Muhammad Innuddin. (2019).*Kombinasi Metode Correlated Naive Bayes dan Metode Seleksi Fitur Wrapper untuk Klasifikasi Data Kesehatan.* Jurnal Teknik Elektro Vol. 11 No. 2

[8] Rizki Tri Prasetio. (2020). *Seleksi Fitur Dan Optmiasi Parameter K-NN Berbasis Algoritma Genitika Pada Dataset Medis.*Jurnal Renponsif. Vol. 2. No. 2.

[9] Diani. (2017). *Analisis Pengaruh Kernel Support Vector Machine (SVM) pada Klasifikasi Data Microarray untuk Deteksi Kanker,* Indonesian Jurnal Of Computing. Bandung.Vol. 2 No. 1. DOI: 10.21108/INDOJC.2017.2.1.169.

[10] Han, J., Kamber, M., & Pei, J. (2012). *Data Mining Concepts and Techniques 3rd Edition.* USA: Morgan Kaufmann.

[11] Vapnik, V. N. (2002). *The Nature of Statistical Learning Theory 2nd Edition.* New York: Springer-Verlag

[12] Gunn, S. (1998). *Support Vector Machines for Classification and Regression.* Southampton: University of Southampton.

[13] J. C. Ang, A. Mirzal, H. Haron, and H. N. A. Hamed, " Supervised, unsupervised, and semi-supervised feature selection: A review on gene selection," IEEE/ACM Trans. Comput. Biol. Bioinforma., vol. 13, no. 5, pp. 971–989, 2016, DOI: 10.1109/TCBB.2015.2478454.

[14] E. Hancer, B. Xue, and M. Zhang, " Differential evolution for filter feature selection based on information theory and feature ranking," Knowledge-Based Syst., vol. 140, pp. 103–119, 2018, DOI: 10.1016/j.knosys.2017.10.028.

[15] M. Alirezanejad, R. Enayatifar, H. Motameni, and H. Nematzadeh, "Heuristic filter feature selection methods for medical datasets," Genomics, vol. 112, no. 2, pp. 1173–1181, 2020, DOI: 10.1016/j.ygeno.2019.07.002.

[16] Abdillah, Abdul, Azis., Prianto, Budi. (2019). *Pembelajaran Mesin Menggunakan Principal Component Analysis dan Support Vector Machines untuk Mendeteksi Diabetes.* J. Matem. Sains. DOI Number: 10.5614/jms.2019.24.1.2.

[17] Buntoro, G.A. (2017). *Analisis Sentimen Calon Gubernur DKI Jakarta 2017 Di Twitter.* INTEGER: Journal of Information Technology, Vol. 2, Ed. 1. DOI: 10.31284/j.integer.2017.v2i1.95

[18] Hikmawan, S., Pardamean, A., Khasanah, S.N., (2020).*Sentimen Analisis Publik Terhadap Joko Widodo Terhadap Wabah Covid-19 Menggunakan Metode Machine Learning.* Jurnal Kajian Ilmiah, Vol. 20. Ed. 2. DOI:10.33633/tc.v19i4.4044

[19] Sari, Erna DH., Irhamah. (2019). *Analisis Sentimen Nasabah Pada Layanan Perbankan Menggunakan Metode Regresi Logistik Biner, Naïve Bayes Classifier (NBC), dan Support Vector Machine*

e-ISSN: 2614-8404
p-ISSN: 2776-3234

(SVM). Jurnal Sains dan Seni. Vol. 8. No. 2. ISSN.
2337-3520. Institut Teknologi Sepuluh Nopember.

e-ISSN: 2614-8404
p-ISSN: 2776-3234

Web-Based Scheduling Application and Motion Sensor Using Arduino Mega

Januardi Nasir
[1]Program Studi Sistem Informasi Fakultas Teknik Universitas Nahdhatul Ulama Sumatera Barat
Email:januardinasir@gmail.com

Abstract – **Waste will appear in a building including lights, air conditioners or fans that no one is using. One reason for this is due to a lack of discipline culture. Culture turns off electrical appliances when they are not used that are less. Use of PHP language and MySQL database. Motion sensors that use infrared passively or better known as PIR (Passive Infra Red) can be used for security. The microcontroller itself is a chip or IC (Integrated circuit) that can be programmed using a computer. The purpose of this research is to find out how to make a web application that can control electronic equipment in buildings, to find out how to make a motion sensor circuit with Arduino Mega so that electronic devices can be active. or off, to find out which is more efficient between using web applications and motion sensors in buildings. The results of this research are that in making a web walker application that can control electronic equipment in buildings requires: web server (hosting), internet connection, ethernet shield, arduino mega, module relay and the use of motion sensors with Arduino Mega, the sensitivity level and time delay can be adjusted. giving a signal when there is movement of a human object.**

Keywords – Web-Based Applications, Motion Sensor, Microcontroller

I. INTRODUCTION

The development of technology today is very helpful for humans, one of which is helping to control the use of electric power at the Nadhatul Ulama University, West Sumatra, where buildings that have lots of space will cause many problems. One of them is related to the efficiency of the power usage monitoring system. This waste is caused by lights, air conditioners, or fans that don't have any users. This problem occurs when someone is not disciplined to turn off the lights when they are not needed. Real examples are lights, air conditioners, and fans in university buildings where the process of turning them on and off is manually controlled by someone. Employees sometimes forget to turn it off when the room is not used for teaching and learning. Forgetting to turn will also be a problem. Lecturers who are already on the fourth floor have to come down to make sure the controllers turn on the lights. This will affect the harmonious relationship between lecturers and campus employees..

The web application can be a solution to the above problems. Besides being able to run on the network the installation process is also easy. The use of language PHP and MySQL database can be utilized. Where the benefits can run on various operating systems if it has been placed on the server and can be accessed anywhere as long as having a network connection. Human error will be minimized. Wasteful use of electric power will also be suppressed. Motion sensors that use passive infrared better known as PIR (Passive Infra-Red) can be used for security. This tool will detect infrared waves generated by living creatures within its range and will issue an output that can be utilized. In this research will be investigated how the coverage detection sensor, living creatures, and anything that can be detected by these sensors. This plan will be implemented by researchers in the classroom. It can also be installed in the workspace at the company. This sensor will detect the presence or absence of people. If there is someone then the light will stay alive. However, if no light will die.

Arduino is an electronic device or an electronic circuit board open-source in which there are main components of a chip microcontroller with the type of firm Atmel AVR. The microcontroller itself is a chip or IC (integrated circuit) that can be programmed using a computer. The purpose of embedding the program in the microcontroller is that electronic circuits can read input, process the input, and then generate the desired output. So microcontroller serves as the brain that controls the input, process, and output of an electronic circuit. (Christopher, 2015).

A. Web-Based Applications

Client-server is a paradigm in information technology that refers to a way to distribute applications into two sides: the client and the server. In the client/server model, an application is divided into two separate parts, but it still a unity there are client component and a server component. The client component is also often referred to as the front-end, while the server component is referred to as the back-end. The client component and application is running in a workstation and receive input data and users[1]. The client component will prepare the data entered by the user by using specific processing technology and send it to the server components that run on the server machine. Generally in the form of a request to several services that are owned by the server. The server component will accept the request and the client as well as the direct and reversed the results of such processing to the client. Clients also receive information on the results of the data process performed by the server and display it to the user to use

e-ISSN: 2614-8404
p-ISSN: 2776-3234

applications that interact with the user[2]. An example of s simple client/server applications is web designed by using Active Server Pages (ASP), or PHP.PHP or ASP scripts will be executed on the webserver (Apache or Internet Information Services), while the scripts that run on the client-side will be executed by the web browser on the client computer. Client-server resolves the issue on which the software uses the database so that each computer does not need to be installed database. The method of the client database server can be installed on a computer as a server and the application is installed on the client[3]. If it is explained briefly it can be argued that the server (Apache Web server) task is to serve a client request. Client (e.g. Firefox browser) the duty is to ask service on the server. An explanation of the relationship of client/server is mainly in the process that occurs in an HTML-based website, PHP, and MySQL, which are further explained with pictures and explanations as follows[4].

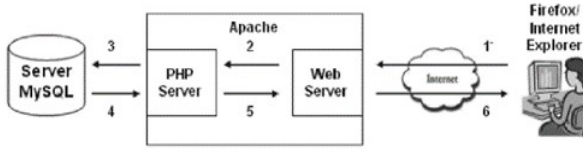


Figure 1. Dynamic web structure with PHP

B. Motion sensor

Infrared is electromagnetic radiation of a wavelength longer than visible light, but shorter than radio wave radiation. It means below (infra = under, in Latin) red, red is the color of visible light with the longest wavelengths. Infrared radiation has a range of three orders and has a wavelength between 700 nm and 1 mm. Sensors PIR Passive Infrared often called Pyroelectric or IR motion sensors[5]. PIR sensor can detect motion, especially coming from the man while in range (range) sensor. This sensor has advantages such as its shape is small, inexpensive, low power consumption, easy to use, and durable. For this reason, these sensors are widely used in applications in the home and for business purposes[6]. PIR (Passive Infra-Red) can be used for security. This tool will detect infrared waves generated by living creatures within its range and will issue an output that can be utilized. In this research will be investigated how the coverage detection sensors and living beings or anything that can be detected by these sensors[7]. PIR (Passive Infrared Receiver) is an electronic component in the form of an infrared-based sensor. PIR sensor is not like the IR sensor that has an IR LED and phototransistor. PLR, unlike IR LEDs that emit anything. PIR sensor only responds to energy and passive IR rays possessed by each object detected by it. Objects that can be detected by these sensors typically are living beings, like humans, cats, dogs, and something that has large enough volume. In the PIR sensor, some parts have their respective functions, namely Fresnel lens, IR Filter, Pyroelectric sensor, Amplifier, and Comparator[8]. The PIR sensor works by capturing the heat energy generated from the passive infrared light beam which every object is above zero temperature. For example, the human body has a body temperature of around 36 ° C. The radiation of infrared light is captured by a pyroelectric sensor which is the core of this PIR sensor. Infrared cause Pyroelectric sensor consisting of gallium nitride, cesium nitrate, and lithium tantalate generate electric current. Infra-red can generate electric current causes bring heat energy. The process is almost the same as the electric current formed when sunlight on the solar cell. Distance range PIR sensor itself can be regulated according to needs. The maximum distance is approximately ± 10 meters and a minimum of ± 30 cm[9].

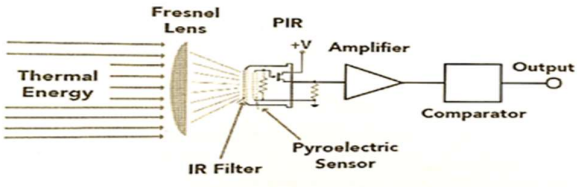


Figure 2. Passive Infrared Receiver

C. Arduino Mega

Arduino is an electronic device or an electronic circuit board that has open-source key components such as a chip microcontroller with the type of AVR from Atmel firm. The microcontroller itself is a chip or IC (integrated circuit) that can be programmed using a computer. The purpose of embedding the program in the microcontroller is that electronic circuits can read input, process the input, and produce output as desired. So microcontroller serves as the brain that controls the input, process, and output of an electronic circuit[10]. One type of Arduino is the Arduino Mega 2560. Arduino Mega 2560 is a microcontroller-Atmega based 2560 with 16 MHz Clock Speed and flash Memory 256KB. Can run on power 7-12V. Have 54 pins digital input/output pin 22-53 and added to the PWM pin, 14 pins PWM on pins 0-13, 16 pin analog inputs on pins A0-A15, USB connection, an auxiliary power supply connection, and the reset button[11].

II. RESEARCH METHODOLOGY

. A. Design And Implementation

Here is a drawing block diagram generally try to combine web-based applications, motion sensor and Arduino mega as well as the device to be in control.

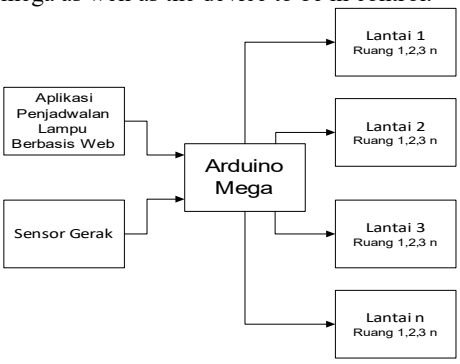


Figure 3. *Web-based applications, motion sensor, and Arduino mega*

e-ISSN: 2614-8404
p-ISSN: 2776-3234

The working principle of the image above as follows: the microcontroller will get input from a light scheduling application web-based and motion sensors. A light scheduling application web-based will contain a view which is used to turn on or turn off an electronic device with a predefined schedule. This application can be used to control remotely, as long as there is an internet connection

The motion sensor or Passive Infra-Red (PIR) was placed in a room that will be created automatically. If anyone entered the room, the lights, air conditioning or fan will on. Meanwhile, if the person left the room, the lights, air conditioning, or fan would off. In brief, this sensor will be a trigger or being input to Arduino Mega.

Arduino Mega is needed because of the number of input and output, or I/O more. Another Arduino has fewer I/O. This microcontroller will act as the brains of the system. It will accept input from a web-based scheduling application and input from motion sensors or Passive Infra-Red (PIR).

The output of the Arduino Mega is an out signal that would drive the relay. This relay will drive the contactor for lights, air conditioning, or fan. Not only the lights, air conditioning, or fan but another electronic device can also be controlled. As motorcycle safety in machinery production, heater scheduling or heater, turn on the TV, radio, DVD player, or other electronic devices.

Use case researchers use to communicate at a high level what the system needs to do, and each of the UML diagram techniques for building a program presents functions in different ways, each view has a different purpose. The following is the use case of the system that will be made in theory.

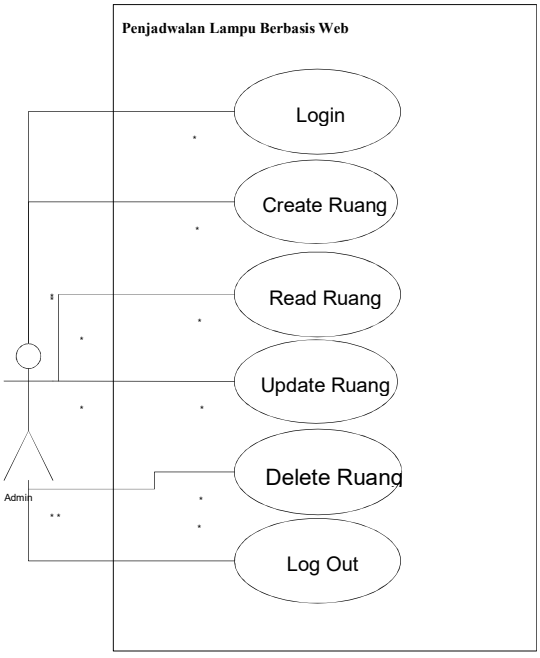


Figure 4. *Use case Application web*

ERD is useful for modeling systems that will later develop the database. This model also helps system researchers when conducting database analysis and design because this model can show the kinds of data needed and the correlation between the data therein.

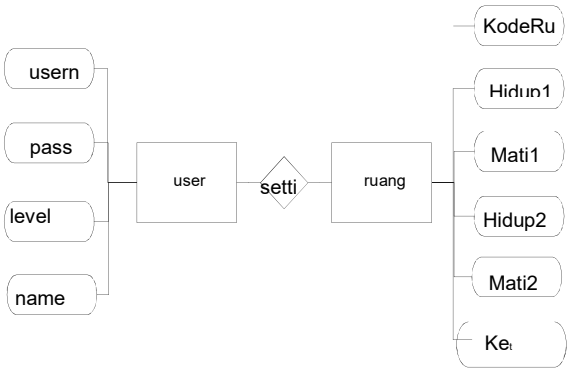


Figure 5. *ERD*

This file design is used as a basis for creating a database. We use a mysql database called "iot". Here are the details of the file design that will be applied.

Table 1. Desain User

No	Field Name	Type	Size	Description
1	id_user	Integer	2	Primary Key
2	username	Varchar	50	
3	password	Varchar	200	
4	level	Varchar	20	
5	name	Varchar	40	

Tabel 2. Desain Room

No	Field name	Type	Size	Description
1	KodeRuang	Varchar	5	Primary Key
2	Hidup1	Time		
3	Mati1	Time		
4	Hidup2	Time		
5	Mati2	Time		
6	Ket	Varchar	20	

B. Microcontroller circuit to Electronic Devices

Here is an overview of its electronic devices. To make this circuit it requires some electronic components such as the Arduino Mega Shield Ethernet, 8-channel Relay Module, PIR Sensor, Adaptor 5-12Volt DC, jumper cables, and boxes and terminal.

Ethernet shield will be used by researchers to receive data from the hosting server that contains 0 (zero) and 1 (one) as the trigger input to the Arduino Mega. There is a LAN port that will get DHCP IP by way of the access point. The access point used by the researchers is TP-Link MR-3220. The access point is connected with an internet connection using a USB modem of 3/tree operator.

e-ISSN: 2614-8404
p-ISSN: 2776-3234

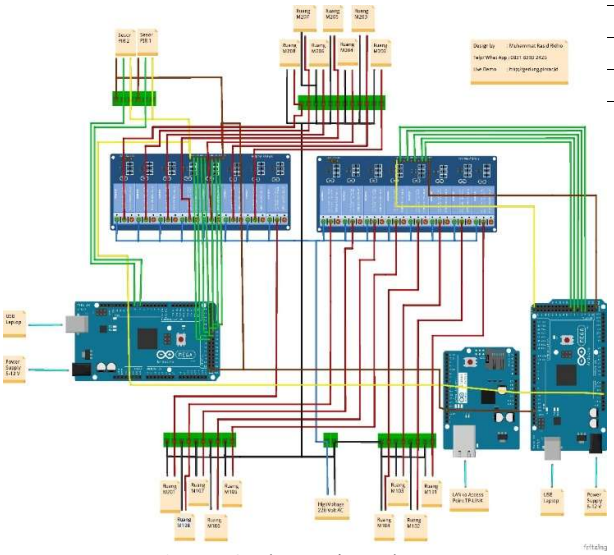


Figure 6. Electronics scheme

The amount of data received zero or one based on MySQL database queries contained in server hosting. Where storage scheduling can be set according to the needs of the building owner. Arduino Mega receives digital signals coming from the Ethernet Shield. This signal is used as an input signal following space or electronic devices. There are fifty-three inputs and outputs on this mega Arduino. So that the amount can be maximized. While analog input with a capacity used only seven pieces. Due to the limitations of the material, the researchers only use output number 41 to number 48. This output Terminat will issue a voltage of about five volts if it gets command 1 (one). The digital voltage will be used to trigger a relay module with eight channels.

Arduino mega needs a power supply of only five to twelve volt DC (direct current). So researchers simply use the adapter from 220 VAC to 9 Volt DC. The relay module used by researchers is 8 units in each module. Relay obtains input from Arduino mega. The relay itself is used as a switch if it gets a signal.

The relay that researchers spend is Active Low, which means that when it gets a signal from Arduino mega relay will off. And conversely, if it does not get a signal from Arduino mega relay will active. Researchers have to change the coding program, reverse output results. Zero means on and one means off.

III. RESULTS AND DISCUSSION

D. Testing of Scheduling Application Web-Based

Testing is done by way of running web applications and try with various scenarios of load control (lighting, air conditioning, etc.). The first scenario by connecting a local network with a hub device or Switch

dengan perangkat Hub atau Swicth.

Tabel 3. Device Aplikasi Web

No	Device to Access http://gedung.pintar.id	True/ False
1	*Personal Computer*	True
2	Laptop	True
3	Handphone Android	True
4	Handphone IOS	True
5	Tablet Samsung	True

Second, by using the Access Point. The next scenario uses the internet. Applications deployed in hosting that has been made by researchers at the domain http://gedung.pintar.id.This address can be accessed by any device as long as there is an internet connection. Researchers have applied it on PC, laptops, and mobile devices such as mobile phones and tablets. All run well.

Figure 7. Results of running the application in the browser

E. Motion Sensor Testing

The sensor used by the researchers is a motion sensor or Passive Infrared (PIR). Testing was conducted to determine how sensitive sensors detect people coming into the room. To be set when the lights will be off or when the light will on. Researchers encountered some problems such as the sensitivity of the sensor and the duration of the sensor's signal to the microcontroller.

Tabel 4. Distance Sensor

No	Distance Sensor PIR ke objek	True / False
1	< 6 Meter	True
2	6,5 Meter	True
3	7 Meter	True
4	7,5 Meter	True
5	8 Meter	False

The problem that appears when it is tested is about the distance between the PIR sensors with the arrival of the object in this case humans. After being tested by the researcher's farthest sensitivity within 6.5 meters to 7.5 meters. It is set on a variable resistor using a screwdriver. If space is greater, it can be used for more than two sensors. Researchers also should pay attention to the angle of its sensitivity. At the moment, the test sensitivity of angle reaches 105 degrees.

The second problem is about the length of time the sensor provides a signal to the Arduino mega. Having tested, there is a setting to adjust the delay time for the first time that needs the objects to move continuously. If it stops moving, the lights will die instantly. by setting the potential meter the delay time can be more than eight minutes. With this, an object or a human does not need to move

e-ISSN: 2614-8404
p-ISSN: 2776-3234

continuously for powering electronic devices such as lights, tv, and others.

Figure 8. Physical hardware

.

IV. CONCLUSION

The conclusion of this study is to know how to make a web application that can control electronic equipment in buildings, to find out how to make a motion sensor circuit with Arduino Mega so that electronic devices can be active or off, to find out which one is more efficient between the use of web applications and motion sensors in buildings. . The results of this research are that in making a web walker application that can control electronic equipment in buildings requires: web server (hosting), internet connection, ethernet shield, arduino mega, module relay and the use of motion sensors with Arduino Mega, the sensitivity level and time delay can be adjusted. Giving a signal when there is movement of a human object.

REFERENCES

[1] B. Festus, F. R. Amodu, and K. W. Thomas, "Development of a microcontroller based automatic night lightning system using motion detector," Int J Biosen Bioelectron, vol. 4, no. 6, pp. 267–270, 2018, doi: 10.15406/ijbsbe.2018.04.00138.

[2] Y. R, B. D, and H. V. | K. S. | P. Roy, "IoT based Classroom Automation using Arduino," Int. J. Trend Sci. Res. Dev., vol. Volume-2, no. Issue-2, pp. 306–313, 2018, doi: 10.31142/ijtsrd9404.

[3] H. R. Hatem, J. N. Shehab, and I. Abdul-Rahman, "ARDUINO Microcontroller Based Building Security System," Eng. Technol. J., vol. 35, no. 5, pp. 532–536, 2017.

[4] A. N. Vaghela, B. D. Gajjar, and S. J. Patel, "Automatic Switch using PIR Sensor," Int. J. Eng. Dev. Res., vol. 5, no. 1, pp. 2321–9939, 2017, [Online]. Available: https://www.ijedr.org/papers/IJEDR1701109.pdf.

[5] Y. Hashim and M. N. Shakib, "Automatic control system of highway lights," TELKOMNIKA (Telecommunication Comput. Electron. Control., vol. 18, no. 6, p. 3123, 2020, doi: 10.12928/telkomnika.v18i6.16497.

[6] C. S. Swetha, "Intruder detection security system," pp. 1170–1174, 2020.

[7] P. Oyekola, T. Oyewo, A. Oyekola, and A. Mohamed, "Arduino based smart home security system," Int. J. Innov. Technol. Explor. Eng., vol. 8, no. 12, pp. 2880–2884, 2019, doi: 10.35940/ijitee.L3052.1081219.

[8] M. T. A. Zaen and A. Yunandy, "Pengendali Lampu Monitoring Rumah Dengan Short Message Service (Sms) Berbasis Arduino Uno," J. Inform. dan Rekayasa Elektron., vol. 1, no. 2, p. 47, 2018, doi: 10.36595/jire.v1i2.59.

[9] H. M. Fadhil, A. Kadhum, and R. Abdulkadhum, "Multi-effectiveness Smart Home Monitoring System Based Artificial Intelligence through Arduino," J. Softw., vol. 12, no. 7, pp. 546–558, 2017, doi: 10.17706/jsw.12.7.546-558.

[10] A. D. Achmad, Z. Zainuddin, J. Toding, and R. Kalau, "Sistem keamanan perumahan berbasis mikrokontroler arduino uno," J. Ilm. Techno Entrep. Acta, vol. 1, no. 1, pp. 1–8, 2016.

[11] K. Srividhyasaradha, I. Joe Louis Paul, and S. Sasirekha, "RFID and PIR motion sensor based automated attendance system for educational institutions," Int. J. Recent Technol. Eng., vol. 8, no. 2 Special Issue 8, pp. 1275–1279, 2019, doi: 10.35940/ijrte.B1052.0882S819.

e-ISSN: 2614-8404
p-ISSN: 2776-3234

COMPATIBILITY OF SELECTION OF STUDENT DEPARTMENTS USINGk-NEAREST NEIGHBOR AND NAÏVE BAYES CLASSIFIER IN INFORMATICS PRIVATE VOCATIONAL SCHOOL, SERANG CITY

Budi Pangestu
Master Program in Computer Science, Faculty of Information Technology, Budi Luhur University Jl. Raya Ciledug,Petukangan Utara, Kebayoran Lama, South Jakarta 12260 Tel. (021) 5853753, Fax. (021) 5869225
E-mail: pangesturj45@gmail.com

***Abstract*- Selection of majors by prospective students when registering at a school, especially a Vocational High School, is very vulnerable because prospective students usually choose a major not because of their individual wishes. And because of the increasing emergence of new schools in cities and districts in each province in Indonesia, especially in the province of Banten. Problems experienced by prospective students when choosing the wrong department or not because of their desire, so that it has an unsatisfactory value or value in each semester fluctuates, especially in their Productive Lessons or Competencies. To provide a solution, a departmental suitability system is needed that can provide recommendations for specialization or major suitability based on students' abilities through attributes that can later assist students in the suitability of majors. The process of classifying the suitability of majors in data mining uses the k-Nearest Neighbor and Naive Bayes Classifier methods by entering 16 (sixteen) criteria or attributes which can later provide an assessment of students through this test when determining the majors for themselves, and there is no interference from people. another when choosing a major later. Research that has been carried out successfully using the k-Nearest Neighbors method has a higher recall of 99%, 81% accuracy and 82% precision compared to the Naïve Bayes Classifier whose recall only yields 98% while the accuracy and precision is the same as the k- Nearest Neighbors.**

Keywords: Data Mining, Department Suitability, K-NN, NBC, Classification

I. INTRODUCTION

Selection of appropriate majors will increase interest and provide comfort for someone in learning. On the basis of the same abilities, it is expected that learning activities can run smoothly and do not experience difficulties and can increase students' interest and learning achievement. Conversely, a lack of interest in learning is due to mistakes in choosing a major. The algorithms that will be used in the classification of majors at the SMKS Informatics at Serang City are k-Naerest Neighbors (K-NN) and Naive Bayes Classifier (NBC). Because this study uses a classification which will compare the level of accuracy, precision and recall of the attributes that are owned, namely the attributes that will be used to classify the suitability of student majors, namely, Academic Values for Semester I and II which consist of the values of Religious Education and Character, Pancasila and Citizenship Education, Indonesian, English, Mathematics, Basic Expertise, Expertise Program Basics, Interview Results, Al-Quran Reading and Writing Results, Interests, Counseling Guidance Teacher Recommendations. Where Precision is the level of accuracy between the information requested by the user with the answer given by the system. Meanwhile, Recall is the success rate of the system in recovering information. Accuracy is defined as the level of closeness between the predicted value and the actual value. The following illustration illustrates the difference between accuracy and precision. As for determining the suitability, the results of Semester 1 and Semester 2 scores will be seen. If there is a drop in semester 2 scores, there is a possibility that the student feels that the department he chose is not in accordance with what he wants and is expected at the beginning of registration. The current condition, which is faced at the location of the researcher, every time he enters semester 3 (three) or 4 (four) there are students whose grades fluctuate and cause these students to tend to reflect and feel that the student chose the wrong major when registering first so that the score does not match desire and learning or how to learn does not focus on the material being taught. The school has tried various methods when learning and specifically for these students, it is often activated in class, but because the students feel they are not in accordance with their wishes or have a wrong direction which results in not being enthusiastic about studying subjects from their majors which results in their grades fluctuating every semester or every there are assignments from the subject teacher.

Research that has been conducted by [1] conducted research on the Implementation of the Fuzzy K-Nearest Neighbor (FK-NN) Classification Method for Fingerprint Access Points in Indoor Positioning, which uses Fuzzy K-

e-ISSN: 2614-8404
p-ISSN: 2776-3234

Nearest Neighbor (FK-NN) and K -Nearest Neighbor (K-NN) with the results of research carried out from the initial stage to testing, and the results of testing the accuracy of client positions that have been carried out from the K-NN and FK-NN methods, resulting in a percentage index (%) on the K-NN method for k = 1 the value reaches 96%, k = 2 to k = 7 the value reaches 76%, and k = 8 to k = 10 the value reaches 73%. Meanwhile, the FK-NN method for k = 1 and k = 2 the value reaches 96%, k = 3 to k = 8 the value reaches 76%, k = 9 the value reaches 73%, and k = 10 the value reaches 76%. Based on these results, it shows that the system is running well and the accuracy results from the implementation of the FK-NN classification method for Fingerprint Access points in Indoor Positioning have a fairly good level of accuracy than the KNN method.

Subsequent research has also been carried out by [2] who conducted research on the Comparison of the Performance of the Naive Bayes and K-Nearest Neighbor Methods for Indonesian Language Article Classification, which uses Naive Bayes and K-Nearest Neighbor with the results of the Naive Bayes method having better performance good with an accuracy rate of 70%, while the K-Nearest Neighbor method has a fairly low level of accuracy, namely 40%.

Subsequent research has also been carried out by [3], who conducted Determination Analysis of High School Departments based on the Fuzzy Tsukamoto Method and the K-Nearest Neighbor (K-NN) Algorithm, using the Logical Fuzzy method, Tsukamoto Method, Data Maining, K-Nearest. National Algorithm with the results of the research that has been done The Tsukamotodan K-NN method can be used as decision support for majoring high school students based on the students' abilities, interests and talents. The Tsukamoto method performs majors calculating the percentage of department recommendations based on Fuzzy logic. The K-NN method determines the pointing by calculating the distance between the data stored as training data and new data as testing data

Subsequent research has also been carried out by [4], who applied the Naïve Bayes Classifier Method to Determine Final Project Topics on the STIKOM Binaniaga Library Website. Using the Naïve Bayesian Classification (NBC) with the research results that have been described, conclusions can be drawn: 1. The Naive Classifier method can be used to determine and display the topic of the IT department in the proposed title. 2. The accuracy and finding in determining the topic in the data of the new IT department thesis title is influenced by the learning data or training data in each category. This training data contains words that often appear in each category or words that can represent certain categories. Further research has also been carried out by Mafakhir [5] who conducted research on the Application of the Naïve Bayes Classifier Method for Student Designation at Madrasah Aliyah Al-Falah Jakarta, with the Naïve Bayes Classifier method, with the results of testing and testing of the methods and prototypes that have been developed, it is concluded that the Naïve Bayes method can be used to classify majors. students. However, the process of converting numeric values into categorical values results in low accuracy, precision, and recall results. Simplification causes detailed information to be lost.

Further research has also been carried out by [6],[7] who conducted research on the comparative analysis of the naïve Bayes Classifier and K-Nearest Neighbor methods against data classification, with the Naive Bayes method, the k-nn algorithm, confusion matrix, with the results of the comparison of the two methods. It can be concluded that the k-NN method has better accuracy than the NBC method. This is evidenced by an accuracy rate of 80% for the k-NN method and 73% for the nbc calculated using the Confusion matrix method.

This study will compare 2 methods using the k-Nearest Neighbor[8] and Naïve Bayes Classifier methods[9],[10],[11],[12] to see the suitability of the Department to students at the Informatics SMKS Serang City which uses the attributes that are owned, namely Semester I and II Academic Values which consist of the value of Religious Education and Character, Pancasila and Citizenship Education, Indonesian, English, Mathematics, Basic Expertise, Expertise Program Basics, Interview Results, Al-Quran Reading and Writing Results, Interests, Counseling Guidance Teacher Recommendations.

II. RESEARCH METHODOLOGY

The object of research is sourced from Basic Education Data (DAPODIK) and Data from Semester 1 and 2 e-report cards for SMKS Informatics at Serang City from 2014-2019, so it can be seen what kind of students study the subject in their majors. Where in the development of this research using software with the PHP programming language (Hypertext Processor), so it takes some software that supports it, namely as follows:

a. Web Browser
b. Web Server (XAMPP)
c. Editor (Sublime / Notepad ++ / Visual

e-ISSN: 2614-8404
p-ISSN: 2776-3234

Studio)

d. Output (Microsoft Office Word and Excel)

A. Technique of Analysis

The analysis technique that will be used is data mining which compares the k-Nearest Neighbor method with the Naïve Bayes Classifier using simple nonprobablity sampling. Which is used to process the data that will be made for Testing data and Test data on the suitability of the Department can be seen in Figure .

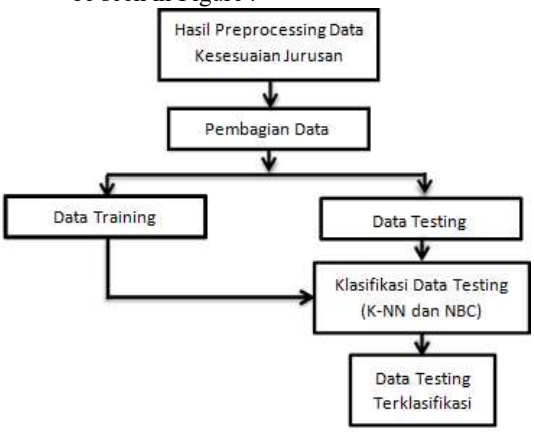


Figure 1. Data Analysis Techniques

B. Design

The system that will be created is a system that will compare the k-Nearest Neighbor[13] and Naïve Bayes Classifier methods[14],[15] to determine the suitability of the department in the selection of the department later.

The process to be designed in making a prototype or system is:

1) Import excel data

Import data is done to see the data that has been processed which will later process the data in the system with the data format in the form of .cvs or .xls

2) Preprocessing

Data that has been imported will be checked for eligibility in the data by the system, and later processed by the system.

3) Comparison results

After the data is in accordance with the needs, it will be processed directly by the system with a comparison stage in the research analysis process. The results of the comparison will be in the form of a description of Appropriate or Unsuitable, which will see the results of data processing in the selection of majors by students.

The steps in calculating the NBC value in the system can be described in the NBC Flowchart which can be seen in Figure 2.

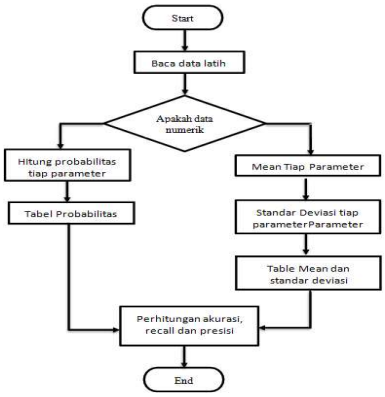


Figure 2. NBC Flowchart

III. RESULT AND DISCUSSION

The data analysis activity in this study aims to provide a detailed picture of what is examined in this study. In analyzing the data in this study the researcher used the method of observation on the attributes used in the calculation based on 997 data taken from 2014-2019 shown in table There are 4 (four) majors that will be of interest, namely Accounting, Office Administration, Software Engineering and Computer and Network Engineering.

Table 1 Total Data Amount

Id	Minat	Smstr Ganjil	Smstr Genap	Hsl Intrview	Hsl BTQ	Rek BK	Hasil
1	AK	83	82	80	90	AK	SESUAI
2	AK	82	86	90	80	AK	SESUAI
3	AK	78	74	80	90	AK	SESUAI
4	AK	83	85	80	80	AK	SESUAI
5	AK	82	86	80	70	AK	SESUAI
6	AK	79	84	90	80	AK	SESUAI
7	AK	77	82	80	80	APK	TIDAKSESUAI
8	AK	81	80	70	90	AK	SESUAI
9	AK	83	89	50	70	AK	SESUAI
....	.....	.....	.....	.....	.....	.....	....
990	APK	84	83	80	70	APK	SESUAI
991	APK	80	78	70	80	APK	SESUAI
992	APK	84	80	50	80	APK	TIDAKSESUAI

Table 2 Test Data

Id	Minat	Smstr Ganjil	Smstr Genap	Hsl Intrview	Hsl BTQ	Rek BK	Hasil
1	AK	83	82	80	90	AK	SESUAI
2	AK	82	86	90	80	AK	SESUAI
3	AK	78	74	80	90	AK	SESUAI
4	AK	83	85	80	80	AK	SESUAI
5	AK	82	86	80	70	AK	SESUAI
6	AK	79	84	90	80	AK	SESUAI
7	AK	77	82	80	80	APK	TIDAKSESUAI
8	AK	81	80	70	90	AK	SESUAI
9	AK	83	89	50	70	AK	SESUAI
....	.....	.....	.....	.....	.....	.....	....
990	APK	84	83	80	70	APK	SESUAI
991	APK	80	78	70	80	APK	SESUAI
992	APK	84	80	50	80	APK	TIDAKSESUAI

A. Research Results

In the results of this study, it is explained

e-ISSN: 2614-8404
p-ISSN: 2776-3234

where to start taking the training data and test data that is owned and then processing the data by entering the data in the calculation of the k-NN and NBC methods which will later calculate in each method and confusion matrix will be carried out. The results of the k-NN and NBC calculations will produce the suitability of the majors that students choose. For the results of the research and steps in the system to determine the suitability of the Koran itself, it will be discussed in prototype testing to show that the results of the application made are as expected. The flowchart for comparison of the k-Nearest Neighbor and Naïve Bayes Classifier methods is shown in Figure 3.

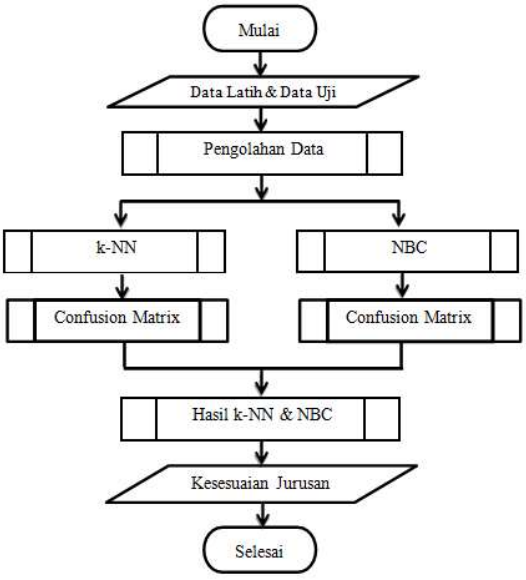


Figure 3. Classification Process Flow

B. System Design

System design determines how the system will meet the objectives of this study in the form of hardware, software, interface displays, databases and files that will be needed in designing this system.

In this design, it consists of a Class Diagram which will display several descriptions of the system, namely the attributes and operations in a class. The class diagram itself can be seen in Figure 4B

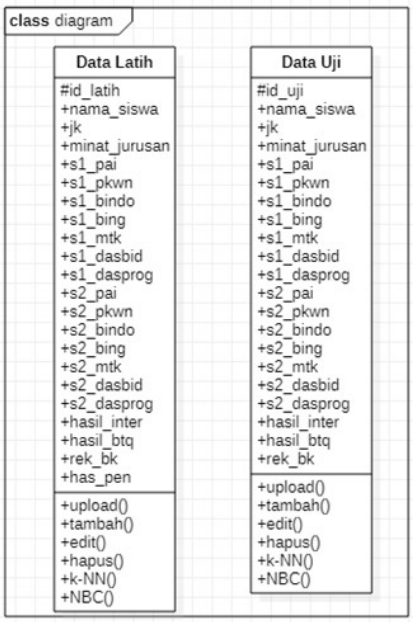


Figure 4. Class Diagram

Consists of 2 tables in the database for the suitability of this direction, and can be seen in Figure 5.

Figure 5. Database Design

C. Display Design

The Min-Max value for odd semesters is taken from data obtained in research conducted by researchers which can be seen in Figure 6.

e-ISSN: 2614-8404
p-ISSN: 2776-3234

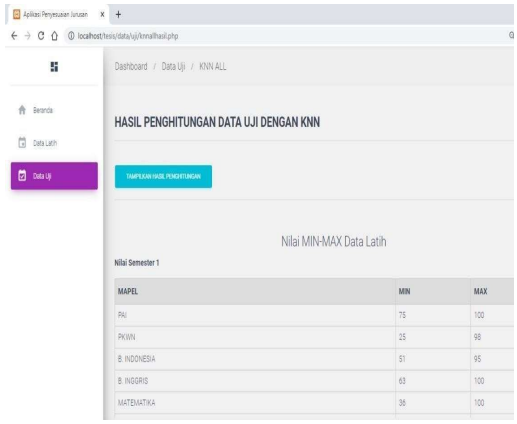


Figure 6. Display of Odd Semester Min Max Value

The Min-Max value for the even semester is not much different from the odd semester value of data taken from the research place which can be seen in Figure 7

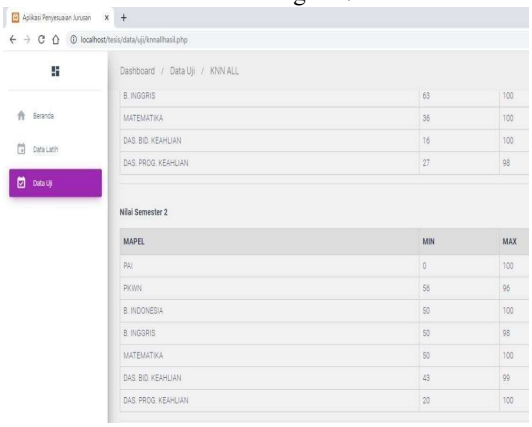


Figure 7. Display of Min Max Even Semester Value

The calculation of the Confusion Matrix on the test data can be seen in Figure 8.

Figure 8. Display Confusion Matrix Value

D. White Box

The tester in white box testing is knowledgeable about coding and writing test cases with the appropriate parameters. This mainly concerns the control flow and data flow of a program. White Box itself has several techniques in testing, such as: Data Flow Testing, Control Flow Testing, Basic Path / Path Testing, and Loop Testing can be seen in Figure 9

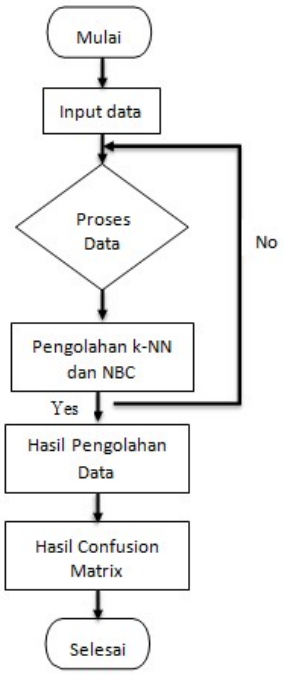


Figure 9. White Box Testing

E. Implementation Plan

1. Implementation of Development
   In the system development stage itself, the system will be refined to be more tailored to the needs, with an estimated time of around 2 months to adjust and be able to complete future developments.
2. System Socialization
   After the development of the system that has been socialized at the beginning of the meal, it will be socialized again for the system with the latest developments to the SMKS Informatics of Serang City.
3. Application of the system
   For the implementation of the system itself, it will be carried out with existing standards in the location, by preparing the necessary equipment, starting from hardware and software, which will be applied to the system, with a series ranging from 3 weeks - 1 month.
4. System Trial
   After it is finished, it will be tested on the whole system and look for whether there are still errors or deficiencies in the system.
5. Evaluation of System Trials
   If there are still errors or shortcomings that

e-ISSN: 2614-8404
p-ISSN: 2776-3234

result, the system will be refined and repaired again according to the needs of the Serang City Informatics SMKS.

6. Refinement of Systems and Procedures
   If the evaluation of the trial has been carried out and no more errors appear in the system, improvements will be made and included in standard operating procedures at the Serang City Informatics SMKS.

IV. CONCLUSION

This conclusion is drawn from the results of research that has been carried out starting from, problems, hypotheses and discussions that have been tested, the following conclusions are obtained:

a. Based on the results of the accuracy, recall and precision test in both methods, it was found that k-Nearest Neighbor has a higher recall of 99%, 81% accuracy and 82% precision compared to the Naïve Bayes Classifier whose recall only produces 98% while for accuracy and precision the same as k. -Nearest Neighbors.
b. So by using the k-Nearest Neighbor method it can be used for calculations in the Adjustment of Majors to Students at Informatics SMKS Serang City

After the results are obtained from this study, it is advisable to carry out further research with the following additions:

a. The data obtained from the research site must be better and try to include all data, not just certain items,
b. Research is carried out annually, in order to see and take into account the suitability of the majors in students later.
c. The stages of implementation for students and the committee for admitting new students are very much needed in the first stage of selecting a department directed by the committee to prospective new students.

REFERENCES

[1] Billyan, B. F., Bhawiyuga, A., & Primananda, R. (2017). Implementasi Metode Klasifikasi Fuzzy K-Nearest Neighbor (FK-NN) Untuk Fingerprint Access Point Pada Indoor Positioning. *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer (JPTIK)*, *1*(11).

[2] Devita, R. N., Herwanto, H. W., & Wibawa, A. P. (2018). Perbandingan Kinerja Metode Naive Bayes dan K-Nearest Neighbor untuk Klasifikasi Artikel Berbahasa indonesia. *Jurnal Teknologi Informasi Dan Ilmu Komputer*, *5*(4), 427. https://doi.org/10.25126/jtiik.201854773

[3] Ariani, F., & Endra, R. Y. (2013). Implementation of Fuzzy Inference System with Tsukamoto Method for Study Progamme Selection. *2nd International Conference on Engineering and Technology Development (ICETD)*, *Icetd*, 189–200.

[4] Ghaniy, R., & Sihotang, K. (2019). Penerapan Metode Naïve Bayes Classifier Untuk Penentuan Topik Tugas Akhir. *Teknois : Jurnal Ilmiah Teknologi Informasi Dan Sains*, *9*(1), 63–72. https://doi.org/10.36350/jbs.v9i1.7

[5]. Mafakhir, A. Z., & Solichin, A. (2020). Penerapan Metode Naïve Bayes Classifier Untuk Penjurusan Siswa Pada Madrasah Aliyah Al-Falah Jakarta. *Fountain of Informatics Journal*, *5*(1), 21. https://doi.org/10.21111/fij.v5i1.4007

[6]. Handayani, I., & Ikrimach, I. (2020). Accuracy Analysis of K-Nearest Neighbor and Naïve Bayes Algorithm in the Diagnosis of Breast Cancer. *Jurnal Infotel*, *12*(4), 151–159. https://doi.org/10.20895/infotel.v12i4.547

[7]. Antaristi, M., & Kurniawan, Y. I. (2017). Aplikasi Klasifikasi Penentuan Pengajuan Kartu Kredit Menggunakan Metode Naive Bayes di Bank BNI Syariah Surabaya. *Jurnal Teknik Elektro*, *9*(2). https://doi.org/10.15294/jte.v9i2.12496

[8]. Indrayuni, E. (2017). Text Mining dalam Analisis Sentimen Review Restoran Menggunakan Algoritma K-Nearest-Neighbor (KNN). *JURNAL TEKNIK INFORMATIKA STMIK ANTAR BANGSA*, *3*(2).

[9] Sipayung, E. M., Maharani, H., & Zefanya, I. (2016). Perancangan Sistem Analisis Sentimen Komentar Pelanggan Menggunakan Metode Naive Bayes Classifier. *Jurnal Sistem Informasi*.

[10] Zuhri, F. N., & Alamsyah, A. (2017). Analisis sentimen masyarakat terhadap brand smartfren menggunakan naive bayes classifier di forum kaskus. *E-Proceeding of Management*.

[11] Indriani, A. (2014). Klasifikasi Data Forum dengan menggunakan Metode Naïve Bayes Classifier. *Seminar Nasional Aplikasi Teknologi Informasi (SNATI) Yogyakarta*.

[12] Oktasari, L., Chrisnanto, Y. H., & Yuniarti, R. (2016). Text Mining Dalam Analisis Sentimen Asuransi Menggunakan Metode Niave Bayes Classifier. *Prosiding SNST*.

[13] Rivki, M., & Bachtiar, A. M. (2017). IMPLEMENTASI ALGORITMA K-

e-ISSN: 2614-8404
p-ISSN: 2776-3234

NEAREST NEIGHBOR DALAM PENGKLASIFIKASIAN FOLLOWER TWITTER YANG MENGGUNAKAN BAHASA INDONESIA. *Jurnal Sistem Informasi*. https://doi.org/10.21609/jsi.v13i1.500

[14] Rinawati, R. (2017). Penentuan Penilaian Kredit Menggunakan Metode Naive Bayes Berbasis Particle Swarm Optimization. *J-SAKTI (Jurnal Sains Komputer Dan Informatika)*. https://doi.org/10.30645/j-sakti.v1i1.28

[15] Saleh, A. (2015). Implementasi Metode Klasifikasi Naïve Bayes Dalam Memprediksi Besarnya Penggunaan Listrik Rumah Tangga. *Creative Information Technology Journal*.

e-ISSN: 2614-8404
p-ISSN: 2776-3234

Analysis of Factors that Affect the Success of E-Learning Implementation of STMIK BI Balikpapan

Surmiati[1], Elvin Leander Hadisaputro[2*)], Joy Nashar Utamajaya[3]
[1,2,3]Program Studi Sistem Informasi, STMIK Borneo Internasional
Email: [1]surmiati@stmik-borneo.ac.id, [2]elvin.leander@stmik-borneo.ac.id , [3]joy.nashar@stmik-borneo.ac.id

Abstract – **E-learning in higher education is a technique to improve learning and teaching experience, and as a tool to educate students through digital media, with or without the guidance of their instructors. STMIK BI Balikpapan has been using it since 2015, but its implementation has not been as optimal as expected. The research aims to identify the factors that influence the success of the application of e-learning in STMIK BI Balikpapan by referring to the model adopted from TAM (Technology Acceptance Model) and TOE (technological, organizational and environment). The research respondents were 94 people. Data were collected through questionnaires and analyzed using the Structural Equation Model (SEM) through the Smart PLS program. The results showed that of the four hypotheses tested, one hypotheses had significant influence (habits) and the other three hypotheses were not significant (connections, motivation and facility).**

Keywords – The success of the application of e-learning, Structural equation model (SEM), TAM, TOE, and Smart PLS

I. INTRODUCTION

Currently the use of IT has penetrated into various fields of life, including in the world of education, starting from the level of basic education to higher education. An example is the use of E-learning which involves the use of information and communication technology (ICT) to convey teaching and learning. E-learning includes the use of many ICT technologies and has been defined as teaching and learning activities that are facilitated online through network technology.

Likewise, STMIK Borneo International Balikpapan, which also concentrates on increasing the use of online e-learning applications by using the internet to improve the quality of education in these institutions. With easy access to E-learning without being bound by time and place, it is hoped that it can increase student motivation and learning achievement on campus.

Learning to use E-Learning at STMIK BI Balikpapan is one example of the use of existing technology since 2015. E-learning that is implemented is Moodle which provides several features including: teaching materials from lecturers, student grades, assignments, attendance, and online Discussion forum. Of course, the use of this technology can make it easier for students and lecturers to interact. However, the implementation is not the case. Utilization of e-learning technology has not been maximized. Some of the problems found are: The intensity of the use of E-learning by lecturers is still very low. There is still a lack of interaction between educators (lecturers) and students / students, and between students. E-learning requires educators (lecturers) to change the role of educators from what originally mastered conventional learning techniques, to learning techniques using ICT. Allocation of time needed to complete assignments is because most students come from the working class. Facilities and equipment needed in learning activities are inadequate. From these conditions it makes the writer feel interested in analyzing the factors that influence the successful implementation of E-Learning in STMIK Borneo Internasional Balikpapan.

This study uses a model adopted from the TAM (Technology Acceptance Model) and TOE (technology, organization and environment) models. The TAM model was first introduced by Davis in 1986, on the adaptation of TRA (Theory of Reasoned Action) (Fishbein & Ajzen, 1975) to explain technology adoption behavior. The TAM model is one of the models used to measure user acceptance on an information system. This model provides a theoretical basis for understanding the factors that influence the acceptance of an information system in an organization [1]. The TOE (Technology-Organization-Environment) model was first introduced by Rocco DePietro, Edith Wiarda and Mitchell Fleischer (1990). The TOE model was further developed by [2]. The TOE model describes a process by which companies use and implement technological innovations that are influenced by the technological context, organizational context, and environmental context. The TOE model uses three main variables namely Technological context, Organizational context, and Environmental context.

Through this case study, the author will examine several things related to the successful application of e-learning, namely: accessibility factors, habits or habits, student motivation, and available facilities. The author wants to know whether the four variables have a significant influence on the successful implementation of E-learning in STMIK Borneo Internasional Balikpapan. To conduct this study, the author collected research data from 7 permanent lecturers and 188 students at STMIK Borneo Internasional.

The research model used in this study are based on the result of previous studies. In [3] the factors used are (1) understanding and management of e-learning programs conducted by teachers, (2) understanding of e-learning based learning owned by students, (3) Availability of facilities and infrastructure contained in SMKN 2 Pengasih in implementing e-learning based learning. In [4] the research aims to look for factors that influence online discussion participation in SCeLE MTI UI. The results showed that in general the factors influencing SCeLE online discussion participation were extrinsic motivation,

e-ISSN: 2614-8404
p-ISSN: 2776-3234

habits, information quality, performance expectancy, social influence, system quality, and service quality. In [5] TAM is use to find the factors that has the significant effect towards the intention to use of mobile banking system.

This research raises the formulation of the problem namely:

1. Does accessibility have an influence on the successful implementation of E-Learning in STMIK Borneo Internasional Balikpapan?
2. Does the habit factor of habit effecton the successful implementation of E-Learning in STMIK Borneo Internasional Balikpapan?
3. Does the student motivation factor have a significant influence on the successful implementation of E-Learning in STMIK Borneo Internasional Balikpapan?
4. Does the factor of facilities owned like cell phones affect the successful implementation of E-Learning in STMIK Borneo Internasional Balikpapan?

II. RESEARCH METHODOLOGY

A. Research Hypothesis

The following is the elaboration of each hypothesis tested in this study, including:

H1: Accessibility factors have a significant positive effect on the successful implementation of the Balikpapan International STMIK Borneo E-learning

H2: The habit factor has a significant positive effect on the successful implementation of the STMIK Borneo International Balikpapan E-learning

H3: Learning motivation factors have a significant positive effect on the successful implementation of the STMIK Borneo International Balikpapan E-learning

H4: Facilities factor has a significant positive effect on the successful implementation of the STMIK Borneo International Balikpapan E-learning

B. Research Variable

Research variables are attributes or properties or values of people, objects or activities that have certain variations determined by researchers to be studied and then drawn conclusions[6]. The following is the division of research variables is:

1. Exogenous Variables (Independent Variable)

Exogenous variables are variables that are not influenced by previous variables (antecedents). The independent variables in this study are:

1) The accessibility factor is symbolized by X1
2) The habit factor is symbolized by X2
3) The learning motivation factor is symbolized by X3
4) The facility factor is symbolized by X4

2. Endogenous Variables (Dependent Variable)

Endogenous variables are variables that are influenced by previous variables. The dependent variable in this study, namely:

1) The successful application of E-Learning is symbolized by Y

The indicator of the variable in this study are:

(X1) Accessibility, Accessibility is how something is easy to reach,to enter, use, see,etc. The indicators are based on the research of [3], which are:

- I can access the internet at any time
- I can access the internet anywhere
- I can see the picture display is good and the sound is clear
- I can only access the internet in certain places
- I can only access the internet at certain times
- I access the internet with poor sound and picture quality

(X2) Habit, habit is defined as something that is usually done. Learning habits are a way of acting that is obtained through a learning process that is carried out repeatedly. The indicators are based on the research of [7], indicators are:

- I use internet services any time
- I use internet services at certain hours
- I use internet services for work purposes
- I use internet services as a learning resource
- I am using the internet service for a very limited time

(X3) Motivation. Motivation is an impulse that arises in a person consciously or unconsciously to take an action with a specific purpose. The indicator are based on the research of [4]. The indicators are:

- I often search and find other sources related to lecture materials
- I am able to complete assignments on time
- I want to make friends via the internet
- I can add knowledge about general knowledge via the internet
- I use the internet to fill my spare / free time

(X4) Facilities. Facilities is the extent to which an individual believes that the technical and organizational infrastructure is available to support the use of the system / technology [7]. Facilities influence on behavioral intention and acceptance of E-learning. The indicator are based on the research of [8] and [9].The indicators are:

- I only use a mobile phone with limited specifications
- I am using a mobile phone with advanced specifications
- I only use a laptop
- I use my laptop and mobile phone to get internet service
- I use facilities that are still very limited

C. Population

The population of this study is 7 permanent lecturers, which is 3 lecturers from manajemen informatika program and 4 lecturers from the information system program, and 188 active students, with 8 students from Manajemen Informatika Program and 180 students from the Information System Program, at STMIK Borneo Internasional.

D. Data Analysis Method

Data analysis was performed using Partial Least Square (PLS) analysis. One of the multivariate statistical research techniques that tries to make comparisons between several dependent variables and multiple independent variables,

e-ISSN: 2614-8404
p-ISSN: 2776-3234

PLS is part of the SEM statistical method to solve multiple regression [10]. Following are the steps in this research:

1) First step: designing a measurement model (outer model). In this stage, the researcher defines and specifies the relationship between the latent construct and the indicator whether it is reflective or formulative. Tests conducted on the outer model [11]:
   - Convergent Validity. The convergent validity value is the factor loading value on the latent variable with its indicators. Expected value> 0.7.
   - Discriminant Validity. This value is a cross loading factor value which is useful for knowing whether the construct has adequate discriminant, that is by comparing the loading value of the intended construct must be greater than the value of loading with other constructs.
   - Composite Reliability. Data that has composite reliability> 0.7 has a high reliability
   - Average Variance Extracted (AVE). Expected AVE value> 0.5.
   - Cronbach Alpha. Reliability tests were strengthened with Cronbach Alpha. Expected value> 0.6 for all constructs.
2) Second step: designing a structural model (inner model). In this stage, the researcher formulates the relationship model between constructs. Evaluation of the inner model uses Coefficient of determination (R^2), Predictive Relevance (Q^2) and Goodness of Fit Index (GoF). Hypothesis testing is done by looking at the probability value and t-statistics. For probability values, the p-value with an alpha of 5% is less than 0.05. The t-table value for alpha 5% is 1.96. So the hypothesis acceptance criteria is when t-statistics> t-table.
3) Third step: construct a path diagram. The main function of the path diagram is to visualize the relationship between indicators and their constructs and between constructs that will make it easier for researchers to see the model as a whole.
4) Hypothesis Test, In general the explanatory research method is a method approach that uses PLS. This is because in this method there is hypothesis testing. Testing the hypothesis can be seen from the t-statistic value and the probability value. For testing hypotheses using statistical values, for alpha 5% the t-statistic value used was 1.96. So the acceptance / rejection criteria Hypothesis is Ha accepted and H0 rejected when t-statistics> 1.96. To reject / accept the hypothesis using probability, Ha is accepted if the value of p <0.05.

III. RESULTS AND DISCUSSION

A. Result

The number of questionnaires distributed was 108 questionnaires, the number of questionnaires that were filled in completely and returned were 94 questionnaires, the number of questionnaires returned but not filled in was 9 questionnaires, and the number of questionnaires that were not returned was 5 questionnaires.

Outer Model

All loading factors for this research have values above 0.70, as shown in table 1, so that no constructs for all variables are eliminated.

Table 1 Loading Factor

	Facilities	Habit	Accessibility	Learning Motivation	Application of E-Learning
X1.1			0,966		
X1.2			0,96		
X1.3			0,969		
X1.4			0,883		
X2.1		0,888			
X2.2		0,846			
X2.3		0,876			
X2.4		0,759			
X3.1				0,847	
X3.2				0,838	
X3.3				0,834	
X3.4				0,802	
X4.1	0,863				
X4.2	0,902				
X4.3	0,912				
X4.4	0,769				
Y1.1					0,964
Y1.2					0,97
Y1.3					0,95
Y1.4					0,901

Discriminant validity is carried out to ensure that each concept of each latent variable is different from the other variables. The model has good discriminant validity if each loading value of each indicator of a latent variable has the greatest loading value with another loading value of another latent variable. The discriminant validity test results are obtained as in table 2.

Table 2 Cross Loading Value

	Facilities	Habit	Accessibility	Learning Motivation	Application of E-Learning
X1.1	0.119	0.349	0.966	0.434	0.189
X1.2	0.132	0.302	0.960	0.399	0.145
X1.3	0.105	0.295	0.969	0.414	0.119
X1.4	0.132	0.340	0.883	0.417	0.081
X2.1	0.286	0.888	0.175	0.615	0.473
X2.2	0.323	0.846	0.182	0.530	0.326
X2.3	0.261	0.876	0.444	0.610	0.383
X2.4	0.451	0.759	0.362	0.504	0.360
X3.1	0.237	0.561	0.292	0.847	0.407
X3.2	0.197	0.594	0.346	0.838	0.288
X3.3	0.202	0.535	0.526	0.834	0.237
X3.4	0.289	0.545	0.353	0.802	0.381
X4.1	0.863	0.257	0.111	0.213	0.182
X4.2	0.902	0.420	0.139	0.268	0.350
X4.3	0.912	0.330	0.060	0.238	0.331
X4.4	0.769	0.273	0.156	0.281	0.168
Y1.1	0.324	0.417	0.098	0.377	0.964

e-ISSN: 2614-8404
p-ISSN: 2776-3234

Y1.2	0.359	0.435	0.165	0.415	0.970
Y1.3	0.336	0.435	0.130	0.351	0.950
Y1.4	0.207	0.473	0.181	0.420	0.901

From table 2, it can be seen that each latent variable has a good discriminant validity because each loading value of each indicator of a latent variable has the greatest loading value with other loading values to other latent variables.

Table 3 presents the Composite Reliability and AVE values for all variables. It can be concluded that all constructs meet reliable criteria. This is indicated by the value of composite reliability above 0.70 and AVE above 0.50 as recommended criteria.

Table 3 Composite Reliability and Average Variance Extracted

	Cronbach's Alpha	rho_A	Composite Reliability	Average Variance Extracted (AVE)
Facilities	0.889	0.964	0.921	0.745
Habit	0.864	0.883	0.908	0.712
Accessibility	0.961	1.041	0.971	0.893
Learning Motivation	0.852	0.873	0.899	0.689
Application of E-Learning	0.961	0.962	0.972	0.896
AVERAGE				0.787

Inner Model

Testing the inner model or structural model is done to see the relationship between the construct, the significance value and the R-square of the research model. The structural model is evaluated using R-square for the dependent construct of the t test as well as the significance of the coefficient of structural path parameters. The structural model is shown in figure 1.

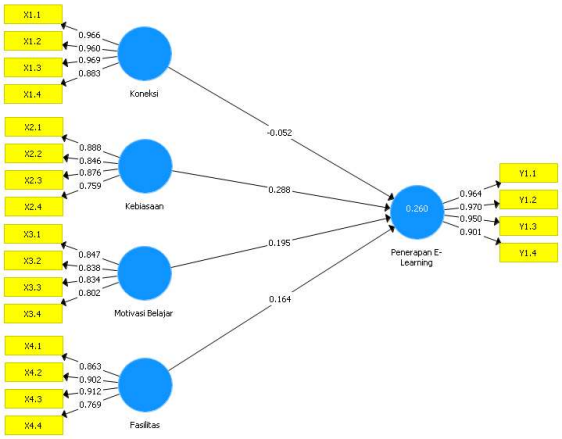


Figure 1 Structural Model

Information:

: Endogen Variable

: Moderation Variable Indicator

Table 4 is the result of R-square estimation using SmartPLS.

Table 4 R-Square Value

	R Square	R Square Adjusted
E-Learning Implementation	0.260	0.226

Table 4 shows the *R-square* value for the E-Learning Implementation variable obtained by 0.260. Furthermore, testing the inner model can be done by looking at the value of *Q-square* (predictive relevance). To assess the Q-square can be done using the equation:

$$Q^2 = 1 - (1 - R^2) \quad \text{.....(1)}$$

The Q-square result using equation (1) is 0,260 or 26%.

The next step of testing the model structure is to assess the *Goodness of Fit (GoF)*. The GoF value can be found using the equation:

$$GoF = \sqrt{AVERAGE\ x\ R^2} \quad \text{.....(2)}$$

Using the *AVERAGE* value from table 3 and the R^2 value from table 4, then the result for *Goodness of Fit* is 0,452, which categories large category according to [12].

Hypothesis Testing Result

The significance of the estimated parameters provides very useful information about the relationship between the research variables. The basis used in testing the hypothesis is the value contained in the output result for inner weight. Table 5 provides the estimated output for structural model testing.

Table 5 Result For Inner Weight

	Original Sample (O)	Sample Mean (M)	Standard Deviation (STDEV)	T Statistics (|O/STDEV|)	P Values
Facilities -> Application of E-Learning	0.164	0.180	0.095	1.722	**0.086**
Habit -> Application of E-Learning	0.288	0.284	0.127	2.270	**0.024**
Accessibility -> Application of E-Learning	-0.052	-0.036	0.125	0.414	**0.679**
Learning Motivation -> Application of E-Learning	0.195	0.211	0.116	1.676	**0.094**

In PLS statistical testing every hypothesized relationship is carried out using simulations. In this case the bootstrap method is performed on the sample. Bootstrap testing is also intended to minimize the problem of research data abnormalities. The bootstrapping test results of the PLS analysis are as follows:

1. Effect of Accessibility on E-Learning Implementation

The test results indicate that the effect of accessibility variables with the application of e-learning shows the path coefficient of -0.052 with a t value of 0.414. This value is smaller than t table (1.960). This result means that accessibility does not have a positive and significant effect on the application of e-learning. This also shows that the indicators of accessibility consisting of can be accessed

e-ISSN: 2614-8404
p-ISSN: 2776-3234

at any time, can be accessed anywhere, display good pictures and clear sound, can only be accessed in certain places, certain times with poor sound and picture quality does not affect the indicators - e-learning application indicators consisting of E-learning can be accessed easily, E-learning makes it easy for students to learn from various sources, E-learning helps students interact with lecturers and friends, e-learning motivates students to complete assignments on time.

2. The influence of habits on the implementation of e-learning

The test results show that the influence of the habit variable with the application of e-learning shows the path coefficient of 0.288 with a t value of 2.270. This value is greater than t table (1.960). This result means that habits have a positive and significant effect on the application of e-learning. It also shows that indicators of habits which consist of using internet services at any time, using internet services at certain hours, using internet services for work purposes, using internet services with very limited time affect indicators of the application of e-learning that consists of E-learning that can be accessed easily, E-learning makes it easy for students to learn from various sources, E-learning helps students interact with lecturers and friends, e-learning motivates students to complete assignments on time.

3. The Effect of Learning Motivation on the Implementation of E-Learning

The test results show that the influence of learning motivation variables with the application of e-learning shows the path coefficient of 0.195 with a t value of 1.676. This value is smaller than t table (1.960). This result means that learning motivation has no positive and significant effect on the application of e-learning. It also shows that indicators of learning motivation consisting of searching for and finding other sources related to lecture material, looking for friends, adding insight into general knowledge, filling in spare time / leisure do not affect the indicators of the application of e-learning consisting of E-learning can be accessed easily, E-learning makes it easy for students to learn from various sources, E-learning helps students interact with lecturers and friends, e-learning motivates students to complete assignments on time.

4. Effect of Facilities on the Implementation of E-Learning

The test results show that the effect of facility variables with the application of e-learning shows a path coefficient of 0.164 with a t value of 1.722. This value is smaller than t table (1.960). This result means that the facility has no positive and significant effect on the application of e-learning. This also shows that the indicators of the facilities which consist of only using mobile phones with limited specifications, using mobile phones with advanced specifications, only using laptops, using laptops and mobile phones to obtain internet services do not affect the indicators of e-implementation learning that consists of E-learning can be accessed easily, E-learning makes it easy for students to learn from various sources, E-learning helps students interact with lecturers and friends, e-learning motivates students to complete assignments on time.

B. Discussion

Based on the results of statistical calculations, it can be concluded that the construct of accessibility has no effect on the construct of implementing e-learning directly. This can be seen from the t-statistic value smaller than 1.96 which is equal to 0.414. This shows that accessibility does not have a direct influence on the application of e-learning. results of this study contradict the research of [4] that found that there are several factors that influence online discussion participation, namely: extrinsic motivation, habits, information quality, performance expectancy, social influence, system quality, and service quality. Factors that greatly influence behavioral intention are performance expectancy, habits, extrinsic motivation, and social influence. Meanwhile, factors that greatly affect user satisfaction are information quality, system quality and service quality.

With reference to the results of statistical calculations, it can be concluded that the construct of habits influences the construct of the application of e-learning directly. This can be seen from the t-statistic value greater than 1.96 which is equal to 2.270. This shows that habits affect the application of e-learning. The results of this study contradict the research of [4] that found that there are several factors that influence online discussion participation, namely: extrinsic motivation, habits, information quality, performance expectancy, social influence, system quality, and service quality. Factors that greatly influence behavioral intention are performance expectancy, habits, extrinsic motivation, and social influence. Meanwhile, factors that greatly affect user satisfaction are information quality, system quality and service quality.

By using the results of statistical calculations, it can be concluded that the construct of learning motivation does not have a significant positive effect on the construct of the application of e-learning directly. This can be seen from the t-statistic value which is smaller than 1.96 which is equal to 1.676. This shows that learning motivation does not have a significant direct effect on the application of e-learning. These results are different from previous studies, namely [13], that found that the effectiveness of e-Learning is influenced by the ease of use factor, the wealth of media interaction used (media richness), and external motivation (extrinsic motivation).

Referring to the results of statistical calculations, it can be concluded that the facility construction has no significant positive effect on the construct of e-learning implementation directly. This can be seen from the t-statistic value smaller than 1.96 which is 1.722. This shows that facilities do not have a significant direct effect on the application of e-learning. The results of this study differ from the study of [14] that gender factors influence students in using the SCeLe Online Discussion System at the University of Indonesia Information Technology Masters Program. In male sex, the influences are e-learning motivation, social influence and facilitating condition, while for female gender the factors that have significant influence are social influence, teacher's roles and facilitating condition.

e-ISSN: 2614-8404
p-ISSN: 2776-3234

IV. CONCLUSION

Based on the results of the analysis that has been done, the factors that has significant effect on the application of learning at STMIK Borneo Internasional is Habits, while Accessibility, Learning motivation and Facilities does not have a significant effect on the application of e-elearning. For further researchers, it is expected to expand the research area other than at the STMIK Borneo International Balikpapan to have a bigger view of the application of e-learning.

REFERENCES

[1] F. D. Davis, "Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology," *MIS Q.*, vol. 13, no. 3, pp. 319–340, 1989.

[2] L. G. Tornatzky and M. Fleischer, "The processes of technological innovation," *J. Technol. Transf.*, vol. 16, no. 1, pp. 45–46, 1990.

[3] A. Ahmadi and M. Wakid, "Evaluasi Pelaksanaan E-learning Pembelajaran Sistem Kelistrikan Siswa Kelas X Teknik Otomotif SMK N 2 Pengasih," *J. Pendidik. Tek. Otomotif*, vol. 14, no. 2, pp. 23–40, 2016.

[4] M. Sihotang and W. S. Nugroho, "Investigasi Faktor-Faktor Yang Mempengaruhi Partisipasi Diskusi Online: Studi Kasus Student-Centered E-Learning Environment Magister Teknologi Informasi Universitas Indonesia," *J. Sist. Inf.*, vol. 11, no. 1, pp. 1–12, 2015.

[5] S. Sopyah, "Analisis Penerimaan Mobile Banking Menggunakan Teori TAM," Universitas Bina Nusantara, 2018.

[6] Sugiyono, *Metode Penelitian Kuantitatif, Kualitatif dan R&D*. Bandung: Alfabet, 2019.

[7] V. Venkatesh, J. Y. L. Thong, and X. Xu, "Consumer Acceptance and Use of Information Technology : Extending The Unified Theory of Acceptance and Use of Technology," *MIS Q.*, vol. 36, no. 1, pp. 157–178, 2012.

[8] M. Kocaleva, I. Stojanovic, and Z. Zdravev, "Model of e-Learning Acceptance and Use for Teaching Staff in Higher Education Institutions," *Int. J. Mod. Educ. Comput. Sci.*, vol. 7, no. 4, pp. 23–31, 2015.

[9] T. Handayani and S. Sudiana, "Analisis Penerapan Model Utaut (Unified Theory of Acceptance and Use of Technology) Terhadap Perilaku Pengguna Sistem Informasi (Studi Kasus: Sistem Informasi Akademik Pada Sttnas Yogyakarta)," *Angkasa J. Ilm. Bid. Teknol.*, vol. 7, no. 2, p. 165, 2017.

[10] I. Ghozali and H. Latan, *Partial Least Square Konsep, Teknik, dan Aplikasi Menggunakan Program Smart PLS 3.0 untuk Penelitian Empiris*. Universitas Diponegoro, 2017.

[11] A. S. Hussein, *Modul Ajar Penelitian Bisnis dan Manajemen Menggunakan Partial Least Square (PLS) dengan Smart PLS 3.0*. Malang: Jurusan Manajemen Fakultas Ekonomi dan Bisnis UNIVERSITAS BRAWIJAYA, 2015.

[12] M. Tenenhaus, S. Amato, and V. E. Vinzi, "A global Goodness – of – Fit index for PLS structural," in *XLII SIS scientific meeting*, 2004, pp. 739–742.

[13] W. Wu and L. Hwang, "The Effectiveness of e-Learning for Blended Courses in Colleges: a Multi-level Empirical Study," *Int. J. Electron. Bus. Manag.*, vol. 8, no. 4, pp. 312–322, 2010.

[14] A. T. Winoto, "Analisis Faktor- Faktor yang Mempengaruhi Mahasiswa dalam Menggunakan Sistem Diskusi Online SCeLe di Program Magister Teknologi Informasi Universitas Indonesia," Universitas Indonesia, 2013.

e-ISSN: 2614-8404
p-ISSN: 2776-3234

A Study of V2V Communication on VANET: Characteristic, Challenges and Research Trends

Ketut Bayu Yogha
[1]Program Studi Teknik Informatika, Fakultas Industri Kreatif dan Telematika, Universitas Trilogi
Email: ketutbayu@trilogi.ac.id

Abstract – Vehicle to Vehicle (V2V) communication is a specific type of communication on Vehicular Ad Hoc Network (VANET) that attracts the great interest of researchers, industries, and government attention in due to its essential application to improve safety driving purposes for the next generation of vehicles. Our paper is a systematic study of V2V communication in VANET that cover the particular research issue, and trends from the recent works of literature. The paper is essential to give reader a brief description about recent V2V Communication studies especially focus on characteristic, challenges and future research trends. We begin the article with a brief V2V communication concept and the V2V application to safety purposes and non-safety purposes; then, we analyze several problems of V2V communication for VANET related to safety issues and non-safety issues. Next, we provide the trends of the V2V communication application for VANET. Finally, provide SWOT analysis as a discussion to identify opportunities and challenges of V2V communication for VANET in the future. The paper does not include a technical explanation. Still, the article describes the general perspective of VANET to the reader, especially for the beginner reader, who intends to learn about the topic.

Keywords – V2V communication, VANET, vehicular communication

I INTRODUCTION

An autonomous vehicle positioned as an interactive robot or agent system that capable of autonomously overcome the various situation in the driving task. An autonomous vehicle is limited in perception, calculation, and decision-making processes from multiple interactive robot system addressed to overcome safety internallyAmessages can be forward based on the highest pheromone intensity from source to destination. Some example of the trajectory-based routing algorithm in unicast communication protocol are Greedy Perimeter Stateless Routing (GPSR) [13][38], ACO-based Routing [39][40], and PSO-Based Routing [41][42][43]

II. V2V COMMUNICATION IN VANET

V2V communication utilizes Wi-Fi technology known as Dedicated Short Range Communication (DSRC) between each vehicle and GPS technology that offers a detail positioning view through the communication exchanges with similarly equipped vehicles. DSRC is a special purposes communication designed for the vehicular vehicle to provide a short-range communication with a neighbor vehicle or with the environment to gain a cooperative driving situation [10]. The DSRC uses 75 MHz spectra for vehicular communication and utilizes the radio technology based on IEEE 802.11p with 3 to 27 Mbps of bandwidth[8]. Several components are required to provide the V2V communication. These components integrated and mediated communication protocol explained in Figure 3.

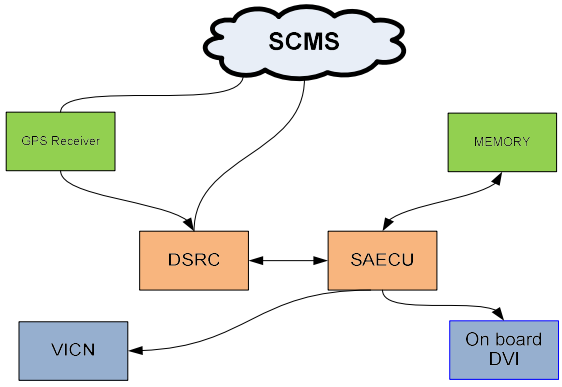


Figure 3. V2V Communication Components

DSRC is a dedicated radio unit that works as the data receiver and transmitter. The GPS receiver responsible for providing vehicle position and time; the data will become an input to DSRC while Memory capable as a storage to store the information from the Safety Application Electronic Control Unit (SAECU). SAECU capable of enhancing safety by calculating input from DSRC and memory and Vehicle's Internal Communication Networks(VICN). SCMS is the facility to ensure security certificates while the communication occurs among the vehicles.

A. V2V Communication Applications

Based on works of literature, we present various applications of V2V communication, categorize for two broad purposes: safety purposes and non-safety purposes
Safety Purposes. The primary goals in this category are to minimize the safety issues by providing guidance or other information for the driver to prevent or anticipate road accidents such as pre-crash or post-crash situations, blind spot anticipation, intersection assistance, etc. The V2V

e-ISSN: 2614-8404
p-ISSN: 2776-3234

communication designed for safety-critical, which means that it requires strict allowable latency in the count of milliseconds and maximum communication range in meters[44]. V2V safety purpose given in Table 2.

Table 2. V2V Safety Purposes

No	Safety purposes	Services
1	Collision Warning System	a. Pre-crash avoidance warning b. Post-crash avoidance warning c. Blind-spot warning d. Blind-intersection warning
2	Cooperative Driving System	a. Lane-change warning b. Safety Platooning formation

Consider an example shown in Fig 4. A post-crash avoidance warning involving three vehicles where the blue vehicle obtaining a broadcast warning messages from the another vehicles that encounter collision around the blue vehicle by means that the vehicles involved in the accident give the warning signals to another vehicles around the communication range to help driver to react properly to avoid the accident or to slow down the vehicle to prevent another crash,. This ability utilize ad-hoc communication among vehicles.

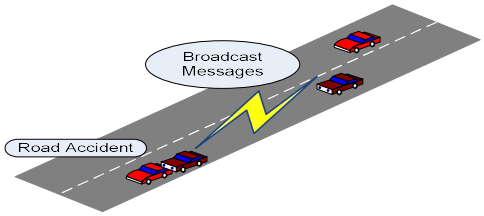


Figure 4. Post-Crash Avoidance Warning

Another high-risk collision spot generally is in the intersection. Figure 5 illustrated blind side possibility while the vehicles crossed the intersection. Each car in the DSRC range will receive a broadcast message from another vehicle and vice-versa that inform the position and direction of each other vehicles movement in a specific area that potentially risks both of the cars in dangerous situation. The ad-hoc communication essential to coordinating the move in a formation/platoon situation.

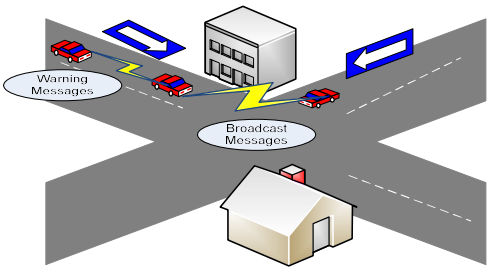


Figure 5. Blind-intersection warning

The vehicle's platoon should be able to follow the leader by continuously maintain and anticipate dynamic situations along the road—the illustration of the vehicles' detachment shown in Figure 6.

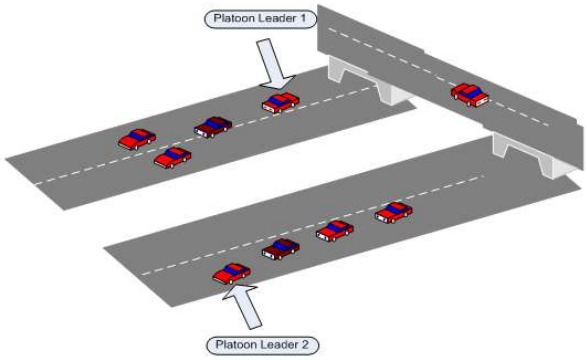


Figure 6. Platooning of Vehicles

The message exchangeability among the vehicles intends to minimize potential accidents and enhance the safe driving within the car with driving assistance features for both autonomous vehicles and non-autonomous vehicles [45]. Furthermore, V2V communication will strengthen the safety support in five-level autonomy in autonomous vehicles where the combination between AI, Vehicle technology, IoT, and communication ability will accelerate the massive use of autonomous vehicles in the future[8][6].

In general, the non-safety purposes is apart from the safety aspect of driving, many other services such as mobility, environment, infotainment, etc. are provided inside the vehicles [44] that require a various point of technology to build it. In this study, we focus on services that make or have a correlation with ad-hoc message exchanges, especially using the V2V communication platform. The detail Non-Safety facilities shown in Table3.

Table 3. Non-safety Purposes

No	Non-Safety purposes	Services
1	Advance Navigation	a. Traffic flow improvement, b. Dynamic eco-routing, c. Congestion Warning.
2	Eco-fuel consumption	

V2V communication enables advance navigation by utilizing information exchanging by other vehicles to provide information about traffic conditions, congestion warnings, or road accident information by combining with

e-ISSN: 2614-8404
p-ISSN: 2776-3234

GPS and IoT technology to guarantee traffic fluidity and circulation[11]. The system combines GPS information and a realtime information from the surrounded vehicle in communication ranges then the navigation module calculates both data to decide the alternative routes available to avoid the congestion before the car passes the congestion spots.

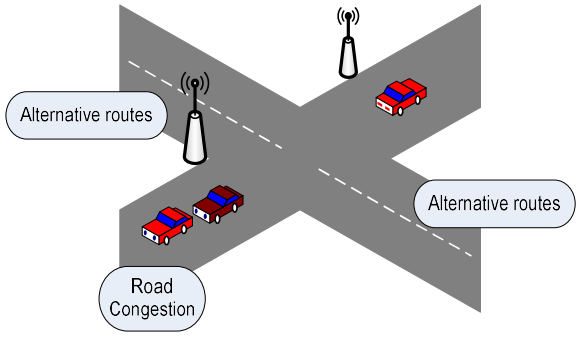


Figure 6. Congestion Warning Services

Figure 6. Describe the illustration of congestion warning services in V2V communication. The traffic information is essential to maintain vehicle movement on various traffic conditions, which have implications for efficient fuel consumption, optimize travel time, and obtaining driving comfort [46].

III. CHALLENGES OF V2V AS A DRIVING COMMUNICATION IN VANET: A BRIEF REVIEW

VANET is a geographical routing protocol that contains unique network characteristics due to particular purposes implementation in vehicular communication that requires a different approach from the global networks[47]. VANET is specifically designed as the next level of driver assistance system to enhance the safety issues in the automotive industry shortly. Moreover, It requires particular purposes of communication protocol to support mobility, quality of services, safety, etc. in the driver assistance system both implemented in autonomous vehicles or non-autonomous vehicles.

A frequent movement of large scale vehicles has implications for several consequences and challenges that must be solved to maintain the functionality of the networks. In VANET, several problems are emerging and become a fundamental issue in VANET[27].

A.Dynamic Topology

A constant movement in a highly active network situation, making the topology of the network continually changing over time, hence establishing and maintain communication is difficult; this situation is called dynamic topology issues. It needs a proper approach to keep the communication process below maximum latency time that implies to quality of services of the network

The establishment of communications between vehicles requires the knowledge of the node positions and their movements, which are very difficult to predict due to the mobility pattern of each car over a dynamic network connection and topology[6].

B.Dynamic Network Connection

Since the vehicle is moving in a highly active network situation over time, it will be resulting in the "ON/OFF" connection along the way. Moreover, the condition can be worst in the presence of radio obstacles that potentially interfere with the communication channel. When the congestion occurs, a path between two nodes wishing to communicate and ensure continuous communication in well-state services, but on the other hand, in the case of low vehicle density, it will inflict frequent disconnection that possible to produce high-rate of failure connection. Both situations require a robust routing protocol to recognize the situation and provide an alternative link rapidly to ensure the quality of communication.

C.Real-time Constraint

The messages exchanged in the VANET network mostly do not cost high resources and high data rates. Unfortunately, the issues are to keep end-to-end delay stays minimum is essential to maintain excellent quality services. For example, sending the warning message broadcast must have a minimum delay to keep the real-time services, or as a consequence, the warning message will no longer be helpful to anticipate the accident or avoid a collision.

D.Quality of Services (QoS)

QoS defined as a standard requirement that needs to meet while establishing end-to-end connections to maintain data exchanges[47]. Various factors and constraints should consider earlier to keep good QoS as each application has its own QoS standard. A right routing approach is essential to efficiently set up new routes when the other one is no longer available due to the changes of various variables such as vehicle velocity, position, topology, distance, etc.

E.Data Security and Privacy issues

The implementation of multiple intelligent on-board potentially stores a large amount of personal information that record individual activities and habit besides the vehicular trajectory data itself, this issues possibly affect the public acceptance issue for VANET system beside Dependability, and vulnerability of leakage of personal information issues. Moreover, another threat could emerge from manipulating the messages or recording the trajectory of the vehicle remotely[6].

IV. THE TRENDS OF V2V COMMUNICATION FOR VANET

As one type of communication in VANET, V2V communication developed along with the development of VANET research itself. Many key important topics in vehicular communication are currently under intensive study and discussion to be enhanced or modify the potential advantages of V2V communication in VANET[27]. V2V ensure the safety and non-safety applications[44], Eco-driving and reduce carbon emission[48], traffic congestion control[19], advance cooperative driving[18][2][20], crash avoidance system[15], *etc*. to offer the next level of safety and driving comfort.

e-ISSN: 2614-8404
p-ISSN: 2776-3234

A. Intelligent Transportation System (ITS)

The application of ITS is essential for modern urban traffic operation. The ITS combines several technology such as Information technology, sensing anf electronic technology, GPS, computational ability, angineering and even the geographical information system. Several urban city or country have been implement the ITS such as Macao intelligent system[49], ITS for good transport and public transport in Italy[50], multi modal ITS in Russia[51], ITS for railway operation in St. Petersburg[52], Central Infrastructur of ITS in Taiwan[53], Central Infrastructur of ITS in Bangkok[54], ITS for freight expedition in Zimbabwe[55], ITS in UK [56], and ITS in China[57]

The previous research proposes by Molosaine. N.R. *et al.* [12] said that it would become an alternative solution in the modern transportation system to reduce road accidents. The idea emerges based on the fast improvement of wireless technology that enables interconnection between Vehicle to Vehicle(V2V) and Vehicle to Infrastructure (V2I), the interconnection is more efficient than just isolated systems to obtain the transportation solution shortly. For example, to minimize road accidents, driving efficiency by avoiding road congestion. Wireless communication is the technology that enables to interconnect various components of the transportation communication system such as sensors, vehicles, and road infrastructures. The wireless technology is the most vital technology that improves vehicular communication, especially V2V communication in VANET. Where VANET will play a significant role in the development of ITS[12].

Daniel. A. *et al.* [11] said that V2V communication for VANET would automate interaction among vehicles and infrastructure to provide a higher level of safety, comfort, and competence in vehicular communication as the architecture model shown in Fig7. In the illustration, the base station responsible for gathering and analyzing the data from the actual information in map databases and CCTV cameras pointed at the specific road, including the traffic signal data and other potential traffic congestion.

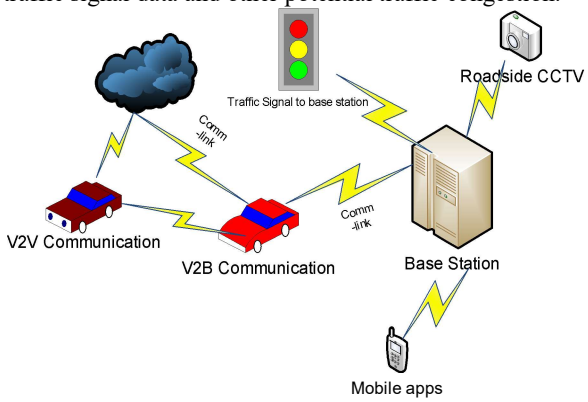


Figure 8. System Model for cooperative communication in ITS

The base station in Figure 8 is essential to provide information as an input for intelligent navigation assistants inside the vehicles in a collaborative environment or the smartphone through a specific application installed on the phone. After receiving the information, the vehicle can utilize it and makes an optimal decision regarding the available path selection at that time. The involvement of Big data and deep learning also accelerates the development of ITS. Zhu. L. *et al.* [58] proposed the architecture model of conducting Big Data analytic in ITS and detail explanation about each aspect component in the architecture shown in Figure 9.

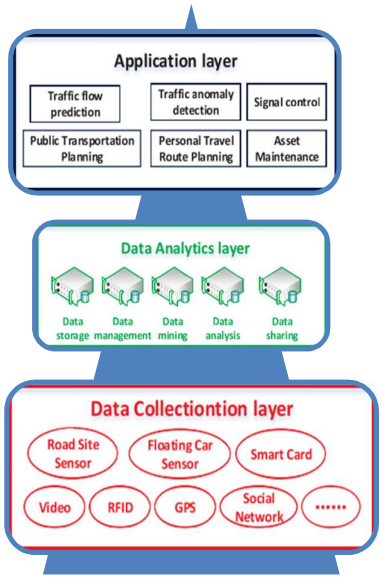


Figure 9. Big Data Analytics Architecture in ITS

The architectures consist of three levels of layers which are Data Connection Layer (DCL), Data Analytics Layer (DAL), and Application Layer (AL). DCL gathering and provides the necessary data for the upper layer. A large amount of information collected from the various sensors, videos, GPS, social networks, etc. DAL will receive data from the DCL and then utilize multiple data analytics methods before sending it to the data storage. DAL also manages the data, analysis, mining, and sharing of the data to the upper layer. DAL is a central point of the architecture. Application layer extracts and utilize the process result from the previous sheet and apply it to different transportation condition such as traffic prediction, traffic guidance, emergency, etc.

Several method and approach proposed to enhance the ITS implementation such as Big Data Analytics for ITS using hadoop[59], Cooperative sensing and mining system for ITS[60], an agent-based approach for ITS[61], then deep-learning approach for ITS [62], a data-driven based of ITS[63], GIS-based ITS[63], GIS-based ITS[64], and the implementation of LiDAR technology to support ITS[65][66]

B.Crash Avoidance System(CAS)

Safety issues are critical in the development of vehicle technology, and all stakeholders must guarantee to enhance safety as one of the top priorities to create safer transport. Ghatwai. et al. (2017) said that a 95% fatal accident caused by human error, the victim number possibly reduce by designing a precision driving assistance system. One of the essential driving assistance features is a crash avoidance system[15]. Several projects developed to ensure the

e-ISSN: 2614-8404
p-ISSN: 2776-3234

system effectively supports driving assistance in the last ten years.

One of the significant projects in the crash avoidance system is PreVent Project. PreVent is an integrated safety platform by advancing current sensor technology with communication technology to speed up the market introduction and penetrated the advance vehicle safety system. PreVent emerged because of the slow commercialization of vehicle safety platforms due to a lack of sensor performance and high production costs that impact low public trust and awareness of these potential vehicle safety systems[16]. The project divided into two main activities, Vertical Function Fields (VFF) and horizontal movements. The VFF focus to make electronic safety zone inside and outside (environment) the vehicle to support the driver while driving on the road and in an accident situation. The VFF function includes several features: Safe speed and following secure functions, Lateral control support functions, Intersection safety functions, and Collision mitigation functions

The horizontal activities focus on safety functions integration and evaluate legal, safety impact by different stakeholders and create dissemination strategies for the expert in the fields, regulator, and public. PreVent is a well-established prototype for the future development of vehicle safety technology both for the autonomous and non-autonomous vehicle by combining various technology that intends to support the vehicle safety and creates, promotes, disseminates, and executes the development of vehicle security.

Connected Vehicle Crash Avoidance (COVCRAF)[14] emerges as an alternative from a previous collision avoidance system that depends on a sensor-based mechanism to detect and avoid the road hazard condition. In a sensor-based order, the information gathered from the environment depends on the internal sensor capability without receiving information from other vehicles around the vehicles at various road conditions. As a result, we can not synchronize the movement simultaneously to respond to the actual situation as a group of cars. COVCRAF is a Cooperative onboard Road Hazard Signaling (RHS) that utilized V2x technology to enhanced awareness of the driving conditions. COVCRAF enables direct interaction among the nearby vehicles by continuously send information about the presence of hazardous driving situations by using wireless communication to enhance safety driving on the road. The V2V platform used when a group of vehicles makes direct communication to create interaction as a response to road hazard situations.

Another potential communication strategy potentially used in the crash avoidance system is a Long-Term Evolution Vehicle (LTE-V)[67].]. LTE-V is a scheduling strategy that utilized radio resources to shared information among a group of users (vehicles) efficiently to share the radio resources; it requires two types of scheduling strategy, A Dynamic Scheduling(DS) strategy and Semipersistent Scheduling (SPS).

DS can use in various services with a control signal overhead; it is the most flexible scheduling strategies in LTE –V. The second strategy is SPS that design to support Voice-over-Internet Protocol (VOIP), SPS allocates a traffic channel periodically without interferences mechanism from control messages during traffic conditions [67]. LTE-V potentially used in the crash avoidance system due to its performance that reduces the minimum delay transmission during messages interchanges among vehicles that effectively used in the crash avoidance system in the future[68]**.**]. Future research in LTE-V continues to grow to analyze any aspects of LTE-V performance in various cases in the crash avoidance system and any other condition in VANET.

C.Advances Cooperative Driving Ability (ACDA)

Cooperative Driving Ability (CDA) [20]] mostly used to support safety and non-safety services that consist of several features described in Table 2 and Table 3. In the last decade, the development of GPS technology, IoT, and social media has triggered enormous research and project in VANET communication technology, especially in integrating VANET technology with those emerging environments. The challenges that emerge in the integration process are exponentially proportional to the benefits and opportunities of VANET research in the future[18].

We explore several opportunities related to the implementation of CDA for VANET both to support safety and non-safety application. First, traffic congestion avoidance, Congestion problem is not directly related to safety issues. Still, it is potentially associated with another question, such as time efficiency, high fuel consumption, environmental problem, or even reduce national productivity, especially in a densely populated city. The congestion can be in the local area, but the impact is significant in the broader field of the country, especially related to the economy and transportation efficiency[17]. In conclusion, it is vital to develop traffic congestion methods as an alternative solution besides exists approach that utilizes the GPS-based system and social media as a source of traffic information.

Brennand. C. et al. (2017) proposed Fast-offset Xpath Service (FOXS)**.** The service reduces the possibility of being in a traffic jam by classified and suggests various alternative routes to vehicles, FOXS developed using fog computing paradigm. As a simulation result, FOXS reduce 70% of the stop time, which is estimated to decrease 29% CO2 release by the vehicles in the air. FOXS also 11.5 % reduce the packet collision metric and reduce 30% messages delay in communication evaluation.

Hsu. C. et al. (2017) examined several potential congestion avoidance procedures in V2V communication to evaluate and validate the congestion avoidance procedure [19]. Hsu. C. et al. (2017) emulate 80 Onboard equipment or reference unit that transmitting signal in 10 Hz-800Hz in the simulation. The test procedure based on the simulation result of busy channel percentages. Each of the congestion algorithms tested using the 80 reference unit. As a threshold, if the simulation detects three or four reference unit in busy states that indicate the potential of road congestion, the percentages is between 50% to 80%. The optimal congestion algorithm should be below the situation by creating alternative routes or channels. The study validates the result by using GPS data-generation

e-ISSN: 2614-8404
p-ISSN: 2776-3234

from the virtual vehicles to measure the performances of the algorithm[19].

D.Cooperative Car Parking System(CCPS)

The enormous number of vehicles that stream to a car park during the rush hour or holiday season will deliver to the congestion inside the car parking area to circulates the movement of the vehicle in a limited parking space. The condition impacts the significant time and fuel consumed for each car as they search for the parking space. Aliedani. A and Loke. S (2017), The average elapsed time for the vehicles to find the available parking space is 20 minutes due to the limited parking space that does not support the adequate information for the driver in a contention level to find the parking spot[21].

In general, the sensing technology used to arranges the occupancy situation and reservation mechanisms development to reduces the level of competitiveness among the vehicles to find the spot. Several car parking approaches proposed to support that conventional way to reduce the average elapsed time for the car to find the parking space in a specific area. The Co-Park (Aliedani proposes a cooperative Car Parking Algorithm. An et., al[21]. Using the multi-agent-based approach and utilize the DSRC protocol, CoPark project simulates the cooperation among the vehicle by exploiting autonomous software agent that enables to do V2V communication in DSRC ranges. An initial belief function works as a heuristic to be communicated with other cars to provide the parking space information from in DSRC ranges. The idea potentially reduces the level of competitiveness among the vehicles and, at the same time, offers realtime local information that essential to minimize the time elapse and decision to make for the driver.

Adewumi. O, et al. (2014)) proposed a heuristic-based algorithm to optimize the parking space allocation—the project tested inside the university parking area. The Pattern search algorithm and particle swarm pattern search investigated to answer two main problems: minimizing the conflict related to available reserved spaces to the user and determines the number of authorized parking spaces to be issued for an unreserved parking slot[69]. As a result, Adewumi. O, et al. (2014) build the hybrid algorithm called Particle Swarm Pattern Search (PSPS) to prevail better performance than the separate algorithm.

Another research proposed by Correa. A. *et al.* (2017) suggests an infrastructure-based network around the parking area to support cooperative vehicle communication. The study aimed at a parking system that possible to accommodate both ordinary vehicles and an autonomous car equipped with vehicle communication technology utilize a road-side facility around the parking area using the V2I concept. The Idea is to overcome the GPS limitation in a positioning accuracy by creating a tree-based searching to select the parking space based on historical data and distribute the information through the vehicles using available infrastructure in the parking area[68]. The simulation shows that the system is near effective performance by considering various communication ranges and autonomous car penetration rates [69].

Another cooperative car system such as Development of agent-based CPS for smart parking system[70], smart indoor parking system based on collaborative palnning of parking space[71], and the implementation of Markov chain as a model baed prediction for parking avaibility[72].

E.Platoon/Formation

Besides Congestion control and parking services, The V2V communication also has a potential implementation as a formation or platoon control[18]. According to Abunei. An *et al.* (2019), 1.35 million people die yearly in various road accidents worldwide. Platooning potentially reduces the road accident or vehicle collision; it also potentially reduces fuel consumption and enhances safety and driving comfort while running as a platoon[73]. Jia proposed the study of platoon-based cooperative driving. D, and Ngudoy. D (2016). They explored and study the relationship between V2V communication and platoon based cooperative driving by designing a consensus-based controller to optimize the movement of platoon vehicles. To build the design, Jia. D, and Ngudoy. D (2016) focuses on the microscopic traffic level by providing a theoretical foundation about V2Vcommunication with cooperative driving behavior to create and maintain formation while moving in the traffic[18].

Several parameters must be meet the requirement such as speed synchronization, space arrangement, type of platoon formation, platoon size, inter-vehicle communication strategy, and time headway [74]. So, it is essential to investigate potential disturbance and stability issues to maintain the formation, which is indirectly related to the problem of driving safety when the configuration of a group of vehicles is running. Studli. S. et al. (2017) investigate several potential control issues such as disturbances amplification, stability, sensitivity analysis of platoon vehicles, as they are performing a cyclic formation or bi-directional formation[74]. Bian. Y., et al. (2019) explores predecessor factors of following strategy to reduce time headway as an essential requirement while performing a stable string the formation. A new method proposed called Constant Time Headway (CTH).With a lower time headway and sufficient information topology matrix obtained better performance of consistent platoon performances[23].

Abunei. A. et al. (2019) introduce a customizable and low-cost V2V platform various VANET standards in 5.9 GHz and 700 MHz bands called Velocity-based Vehicles Platooning (VVP), it is one of the most significant improvements in vehicles communication technology nowadays. VVP enables a group of vehicles to maintain the distance among them in a high-speed situation, and it synchronized any movement in a small group of cars. VVP based on leader-follower synchronization that harmonizes leader and follower movement in various control variables: velocity, steering angle, inertia, and vehicle position using the V2V communication in DSRC range. The system is 10% more efficient in reducing fuel consumption and reduced gas emission[73].

The simulation process simulates the 5.9 GHz and 700 MHz bands to examine the performance in an emergency. As a result, the 700 MHz has better performances in harsh

e-ISSN: 2614-8404
p-ISSN: 2776-3234

situation compare with the other one. The result is essential as an alternative solution for the future development of the On-Board Unit (OBU) that currently used a 5.9 GHz band as they based communication frequency. Another research proposed by Kim. J and Han. Y (2019) focuses on onV2V communication in the group cast platooning scheme by formulating Markov Decision Process is used to optimizing join retransmission control to maximize the time headway in a single-line vehicle platooning[75]. Single-line vehicle platooning can be used in various types of land transportation such as car[22], bus platooning[76], truck platoon[10].

Platoon based cooperative driving ability attracts the great interest of researchers and industries attention in due to its essential application to improve safety driving, security, or even fuel efficiency by reducing air drag and maintaining constant speed in various road conditions. Several projects developed worldwide to extend the platoon ability in multiple types of vehicles, such as semi-autonomous truck expedition and heavy-duty truck [77][78][79] and a platoon of a vehicle for passenger car[80]. The research and development for the formation of the truck are higher than for passenger cars due to the financial ability and less risk of safety issues for the passenger.

F.Cooperative Lane Changes Protocol (CLCP)

The increasing number of vehicles that are running every day in every city worldwide is an important aspect to support the economic movement in every country. Unfortunately, this routine and essential activity also potentially caused many traffic accidents worldwide with a lot of the number of casualties which indirectly impact the economic losses and images of the country[81]. The active safety driving assistant is essential to enhance the driving safety and comfort by minimizing driver error or anticipating various harsh conditions the road[24][26]. Recently, several car manufacturers and research institutes around the world struggle to provide and improves an optimal driving assistant, especially in developing the lane change assistant system for autonomous or non-autonomous vehicles that provide proper decision to make the lane change in safe and efficient ways[82].

Ruina. D. et al. (2014) define that lane change warning system must guarantee two significant aspects, the car following scheme and collision avoidance scheme to anticipate the emergency, in examples, direct collision with another vehicle or other transportation modes. V2V communication is significant to enhance the lane change warning system by providing realtime information from the local environment around the car. In the last ten years, the safety system usually depends on sensor-based in OBU without collecting data from the surrounding environment such as vehicle or road infrastructure in an emergency caused by vehicle movement on the road. Ruina. D. et al. (2014) proposed a central lane change logic system in the following Figure 10.

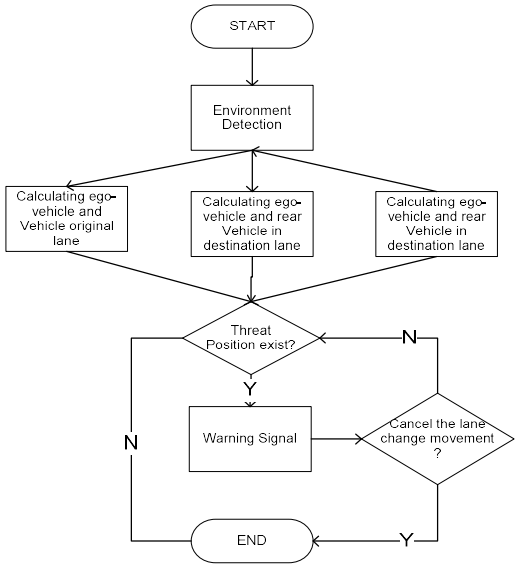


Figure 9. The lane change logic system

Figure 10 describes if there is a potential collision threat that exists in the current situation. The system will generate the warning system; otherwise, the warning signal will be inactive, and the lane change aborted. The V2V communication work as support features to gather data from another vehicle as an input to be calculated and the proposed decision whether the system will give an alert or aborted the warning regarding anticipate the situation. To maintain continuous communication, so the system can make proper lane changing decision, a reliable and intensive connection is obligatory. Wang. L. et al. (2018) proposed a simple communication scheme using ACK messages to ensure constant communication. Based on the simulation result, Wang. L. et al. (2018) claim that communication capability not only elevating the driving safety but also increasing the efficiency of road traffic[82].

Calculating the data and create proper decision is another important aspect of the lane- change control logic system. Recently, the development of a machine learning algorithm has been rapidly growing research in the field, especially in the lane change system. Liu. X. et al. (2019) used a deep learning method to improve the decision-making process while the lane change is needed. The proposed model uses a historical driver experiences from the driving log and the V2V memory effect as a situational maneuvers assessment. Liu. X. et al. (2019) claim that the Deep Learning Networks (DNN) model achieves higher identification accuracy not only in a lane-changing decision but also the reason to keep the lane maneuver compare with the conventional machine learning[83]. Sakr. A. et al. (2018) analyzing the performance of three supervised-learning techniques, the forest random, vector machine, and decision tree with gradient boosting. The proposed research is essential as guidance for future research to choose the most efficient among the three supervised-learning techniques[84].

Cui Y. et al. (2020) proposes a LiDAR-based system as an innovation in V2V communications technology. Light Detection and Ranging (LiDAR) used to identify and

e-ISSN: 2614-8404
p-ISSN: 2776-3234

predict the lane change situation. The system accommodates a real-life situation where not all vehicles can be connected, and there is a conventional car that is not supported by V2V technology, it is more than 60% is not equipped by vehicle communication technology. LiDAR provides an alternative solution to accommodate those unconnected vehicles by use LiDAR as a roadside guiding facility for the mixed-traffic condition. The LiDAR system work as a data collector to record real-time data as training input to the back-end system. The back-end system will identify and predict vehicle movement in the future based on previous condition records by LiDAR[25].

VI. DISCUSSION

To describe the trends, we provided a chart designated the sixty works of literature related to represents the distribution of V2V communication research trends in VANET application in figure 11.

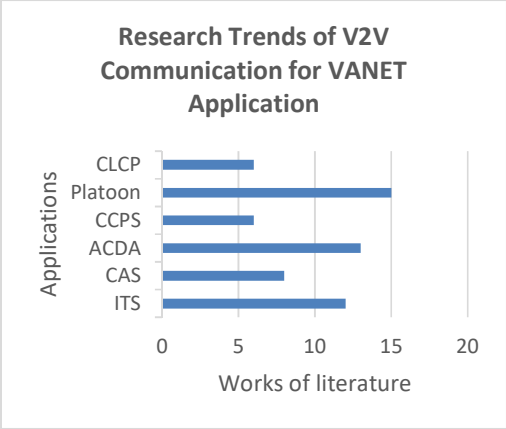


Figure 10. Research Trends Chart of V2V Communication for VANET Application

Figure 10 shows that the implementation of V2V communication for platoon has the highest research interest(25%), followed by ACDA research(21%), and ITS(20%). The lowest research interest is CLCP (10%) and CCPS (10%). The chart cannot be generalized as a big picture of overall research trends around the world, but we hope to give a brief overview of V2V implementation in a smaller scope of application on VANET

We define several aspects which can accelerate and obstruct the development of VANET to support the vehicular technology in near future from various point of views such as economy, legal, infrastructure, public, community acceptance, and the related technology itself. We used A SWOT matrix to analyze the correlation between strength, weakness, opportunity, and threat define in Table 4.

Table 4. SWOT Analysis

V2V	Strength(S)	Weakness(W)
Opportunity(O)	- Wireless based network development, - IoT & Big Data - Progressive AI development - Growing Autonomous Vehicles industry, - Government support, - Adolescent Technology - Global trends in mobile vehicular environment	- High cost in R&D, - More expensive product, - Government Infrastructure support, - Different safety Standard Regulation, - Algorithm performance and validation
Threats(T)	- world economy fluctuation, - Public Trust in a safety issue - Legal issue - Unclear market segmentation	- High-cost R&D, - Expensive product, - conflict of interest (regulator developer)

Table 4 shows that the expansion utility of wireless technology combined with the IoT and Big Data Analysis directly impacts vehicular research to exploit and integrated those technologies to create a more intelligent vehicle. On the other hand, artificial intelligence and growing autonomous vehicles boost the VANET research rapidly in some regions such as North America, Europe, and East Asia, which well-known as a center of car production and development. Unfortunately, the opportunity development in the vehicular industry is also facing several weaknesses related to high-cost R&D, which impacts the price rising for the consumers. Different safety standards and regulations related to the safety issue in different regions retard the industry to exploit the opportunities in any aspect of technology development. Although the existing algorithms are robust and reliable, there is still challenging to validate and examine their performances in various conditions, regulation, and different safety standards.

This emerging industry also facing severe threats and weakness, especially in fundamental economic aspects such as world economic fluctuation that impact the R&D and makes the limitation in market segmentation. The low Public trust to use the vehicles equipped with V2V communication to enhance safe driving requires the car producer to continuously improving the safety technology and disseminating it to a vast community that will raise the cost without unpredictable profit in short and middle terms.

Although VANET is a progressive innovation in the vehicular industry soon, The development of VANET is also inseparable from the challenges to resolve in hierarchical stages for a long time. The situation opens up research opportunities in various fields and development opportunities for university and research institutions. Still, it will require a lot of resources in the development and dissemination of the technology to gain the public trust to buy and use the vehicles equipped with this VANET technology. As a growing field of research, these challenges will slow down VANET implementation, especially to boost the application of autonomous vehicles shortly.

V. CONCLUSION

The study explores and provides review literature of fundamental V2V communication in VANET. Several challenges and trends are elaborated. The V2V communication in VANET potentially improves the future transportation that provides a higher standard of safety driving, enhances the driving comfort, and supports the big

e-ISSN: 2614-8404
p-ISSN: 2776-3234

idea of ITS both to answer the safety issue or non-safety issues in the future. Unfortunately, the study also found that VANET technology still used in limited segmentation and needs further research to implement VANET in a vehicle effectively. A SWOT matrix describes several fundamental factors that need to solved by all stakeholders, such as a legal problem, global economic fluctuation, low public trust, the multi-standard problem in a different region, the regulations, infrastructures problem, and high-cost of R&D

REFERENCES

[1] Li, D., Liu, M., Zhao, F., & Liu, Y. (2019). Challenges and countermeasures of interaction in autonomous vehicles. *Science China Information Sciences*, *62*(5), 3–5. https://doi.org/10.1007/s11432-018-9766-3

[2] Wang, N., Wang, X., Palacharla, P., & Ikeuchi, T. (2018). Cooperative autonomous driving for traffic congestion avoidance through vehicle-to-vehicle communications. *IEEE Vehicular Networking Conference, VNC*, *2018-Janua*, 327–330. https://doi.org/10.1109/VNC.2017.8275620

[3] Wang, Y., Yao, J., & Chen, G. (2018). An evolving super-network model with inter-vehicle communications. *Journal of the Franklin Institute*. https://doi.org/10.1016/j.jfranklin.2018.07.036

[4] Mchergui, A., Moulahi, T., Alaya, B., & Nasri, S. (2017). A survey and comparative study of QoS aware broadcasting techniques in VANET. *Telecommunication Systems*, *66*(2), 253–281. https://doi.org/10.1007/s11235-017-0280-9

[5] Caballero-gil, C., Caballero-gil, P., & Molina-gil, J. (2015). *Self-Organized Clustering Architecture for Vehicular Ad Hoc Networks*. *2015*. https://doi.org/10.1155/2015/384869

[6] Liang, W., Li, Z., Zhang, H., Sun, Y., & Bie, R. (2014). Vehicular ad hoc networks: Architectures, Research issues, Challenges and trends. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *8491*, 102–113. https://doi.org/10.1155/2015/745303

[7] Demba, A., & Moller, D. P. F. (2018). Vehicle-to-Vehicle Communication Technology. *IEEE International Conference on Electro Information Technology*, *2018-May*(1), 459–464. https://doi.org/10.1109/EIT.2018.8500189

[8] Özdemir, Ö., Kılıç, İ., Yazıcı, A., & Özkan, K. (2016). A V2V System Module for Inter Vehicle Communication. *Applied Mechanics and Materials*, *850*, 16–22. https://doi.org/10.4028/www.scientific.net/amm.850.16

[9] Shaikh, S. N., & Patil, S. R. (2016). A robust broadcast scheme for vehicle to vehicle communication system. *Conference on Advances in Signal Processing, CASP 2016*, 301–305. https://doi.org/10.1109/CASP.2016.7746184

[10] Gao, S., Lim, A., & Bevly, D. (2016). An empirical study of DSRC V2V performance in truck platooning scenarios. *Digital Communications and Networks*, *2*(4), 233–244. https://doi.org/10.1016/j.dcan.2016.10.003

[11] Daniel, A., Paul, A., Ahmad, A., & Rho, S. (2016). Cooperative Intelligence of Vehicles for Intelligent Transportation Systems (ITS). *Wireless Personal Communications*, *87*(2), 461–484. https://doi.org/10.1007/s11277-015-3078-7

[12] Moloisane, N. R., Malekian, R., & Capeska Bogatinoska, D. (2017). Wireless machine-to-machine communication for intelligent transportation systems: Internet of vehicles and vehicle to grid. *2017 40th International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2017 - Proceedings*, 411–415. https://doi.org/10.23919/MIPRO.2017.7973459

[13] Moussaoui, B., Fouchal, H., Ayaida, M., & Mermiz, S. (2016). Unicast routing on VANETs. *Proceedings of the 2016 Federated Conference on Computer Science and Information Systems, FedCSIS 2016*, *8*, 1089–1092. https://doi.org/10.15439/2016F262

[14] Outay, F., Bargaoui, H., Chemek, A., Kamoun, F., & Yasar, A. (2019). The covcrav project: Architecture and design of a cooperative v2v crash avoidance system. *Procedia Computer Science*, *160*, 473–478. https://doi.org/10.1016/j.procs.2019.11.062

[15] Ghatwai, N. G., Harpale, V. K., & Kale, M. (2017). Vehicle To vehicle communication for crash avoidance system. *Proceedings - 2nd International Conference on Computing, Communication, Control and Automation, ICCUBEA 2016*, 1–3. https://doi.org/10.1109/ICCUBEA.2016.7860118

[16] Matthias Schulze, Tapani Mäkinen, Joachim Irion, Maxime Flament, T. K. (2008). *Preventive and Active Safety Applications Integrated Project*. 198.

[17] Brennand, C. A. R. L., Filho, G. P. R., Maia, G., Cunha, F., Guidoni, D. L., & Villas, L. A. (2019). Towards a Fog-Enabled Intelligent Transportation System to Reduce Traffic Jam. *Sensors (Basel, Switzerland)*, *19*(18), 1–30. https://doi.org/10.3390/s19183916

[18] Jia, D., & Ngoduy, D. (2016). Platoon based cooperative driving model with consideration of realistic inter-vehicle communication. *Transportation Research Part C: Emerging Technologies*, *68*, 245–264. https://doi.org/10.1016/j.trc.2016.04.008

[19] Hsu, C., Fikentscher, J., & Kreeb, R. (2017). Development of potential methods for testing congestion control algorithm implemented in vehicle-to-vehicle communications. *Traffic Injury Prevention*, *18*(S1), 51–57.

[20] Desai, P., Loke, S. W., & Desai, A. (2017).

e-ISSN: 2614-8404
p-ISSN: 2776-3234

Cooperative vehicles for robust traffic congestion reduction: An analysis based on algorithmic, environmental and agent behavioral factors. *PLoS ONE*, *12*(8), 1–20. https://doi.org/10.1371/journal.pone.0182621

[21] Aliedani, A., & Loke, S. W. (2018). Cooperative car parking using vehicle-to-vehicle communication: An agent-based analysis. *Computers, Environment and Urban Systems*, *October 2017*, 101256. https://doi.org/10.1016/j.compenvurbsys.2018.06.002

[22] Hasrouny, H., Samhat, A. E., Bassil, C., & Laouiti, A. (2018). Trust model for secure group leader-based communications in VANET. *Wireless Networks*, *6*, 1–23. https://doi.org/10.1007/s11276-018-1756-6

[23] Bian, Y., Zheng, Y., Ren, W., Li, S. E., Wang, J., & Li, K. (2019). Reducing time headway for platooning of connected vehicles via V2V communication. *Transportation Research Part C: Emerging Technologies*, *102*(March), 87–105. https://doi.org/10.1016/j.trc.2019.03.002

[24] Liu, Z. Q., Zhang, T., & Wang, Y. F. (2019). Research on Local Dynamic Path Planning Method for Intelligent Vehicle Lane-Changing. *Journal of Advanced Transportation*, *2019*. https://doi.org/10.1155/2019/4762658

[25] Cui, Y., Wu, J., Xu, H., & Wang, A. (2020). Lane change identification and prediction with roadside LiDAR data. *Optics and Laser Technology*, *123*(September 2019), 105934. https://doi.org/10.1016/j.optlastec.2019.105934

[26] Peng, T., Su, L., Zhang, R., Guan, Z., Zhao, H., Qiu, Z., Zong, C., & Xu, H. (2020). A new safe lane-change trajectory model and collision avoidance control method for automatic driving vehicles. *Expert Systems with Applications*, *141*. https://doi.org/10.1016/j.eswa.2019.112953

[27] Eze, E. C., Zhang, S. J., Liu, E. J., & Eze, J. C. (2016). Advances in vehicular ad-hoc networks (VANETs): Challenges and road-map for future development. *International Journal of Automation and Computing*, *13*(1), 1–18. https://doi.org/10.1007/s11633-015-0913-y

[28] Tilahun, S. L., & Tawhid, M. A. (2018). Swarm hyperheuristic framework. *Journal of Heuristics*, *25*(4), 809–836. https://doi.org/10.1007/s10732-018-9397-6

[29] Chiu, K. L., & Hwang, R. H. (2012). Communication framework for vehicle ad hoc network on freeways. *Telecommunication Systems*, *50*(4), 243–256. https://doi.org/10.1007/s11235-010-9401-4

[30] Zeadally, S., Hunt, R., Chen, Y. S., Irwin, A., & Hassan, A. (2012). Vehicular ad hoc networks (VANETS): Status, results, and challenges. *Telecommunication Systems*, *50*(4), 217–241. https://doi.org/10.1007/s11235-010-9400-5

[31] Cheng, J., Cheng, J., Zhou, M., Liu, F., Gao, S., & Liu, C. (2015). Routing in internet of vehicles: A review. *IEEE Transactions on Intelligent Transportation Systems*, *16*(5), 2339–2352. https://doi.org/10.1109/TITS.2015.2423667

[32] Ferreiro-Lage, J. A., Gestoso, C. P., Rubiños, O., & Agelet, F. A. (2009). Analysis of unicast routing protocols for VANETs. *Proceedings of the 5th International Conference on Networking and Services, ICNS 2009*, 518–521. https://doi.org/10.1109/ICNS.2009.96

[33] Bilal, S. M., Bernardos, C. J., & Guerrero, C. (2013). Position-based routing in vehicular networks: A survey. *Journal of Network and Computer Applications*, *36*(2), 685–697. https://doi.org/10.1016/j.jnca.2012.12.023

[34] Saleh, A. I., Gamel, S. A., & Abo-Al-Ez, K. M. (2017). A Reliable Routing Protocol for Vehicular Ad hoc Networks. *Computers and Electrical Engineering*, *64*, 473–495. https://doi.org/10.1016/j.compeleceng.2016.11.011

[35] Zhou, Q., Fan, Y., & Wei, C. (2012). Heuristic routing protocol research on opportunistic networks. *Proceedings of the 14th IEEE International Conference on High Performance Computing and Communications, HPCC-2012 - 9th IEEE International Conference on Embedded Software and Systems, ICESS-2012*, 1704–1707. https://doi.org/10.1109/HPCC.2012.254

[36] Liu, Z. Y., Zhou, J. G., Zhao, T., & Yan, W. (2009). An opportunistic approach to enhance the geographical source routing protocol for vehicular ad hoc networks. *IEEE Vehicular Technology Conference*, 1–5. https://doi.org/10.1109/VETECF.2009.5378797

[37] Suhendra, T., & Priyambodo, T. K. (2017). Analisis Perbandingan Algoritma Perencanaan Jalur Robot Bergerak Pada Lingkungan Dinamis. *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, *11*(1), 21. https://doi.org/10.22146/ijccs.15743

[38] Hu, L., Ding, Z., & Shi, H. (2012). An improved GPSR routing strategy in VANET. *2012 International Conference on Wireless Communications, Networking and Mobile Computing, WiCOM 2012*, 1–4. https://doi.org/10.1109/WiCOM.2012.6478416

[39] Silva, R., Lopes, H. S., & Godoy, W. (2013). A heuristic algorithm based on ant colony optimization for multi-objective routing in vehicle Ad Hoc networks. *Proceedings - 1st BRICS Countries Congress on Computational Intelligence, BRICS-CCI 2013*, 435–440. https://doi.org/10.1109/BRICS-CCI-CBIC.2013.78

[40] Rajesh Kumar, M., & Routray, S. K. (2017). Ant Colony based Dynamic source routing for VANET. *Proceedings of the 2016 2nd International Conference on Applied and Theoretical Computing*

e-ISSN: 2614-8404
p-ISSN: 2776-3234

and Communication Technology, ICATccT 2016, 279–282. https://doi.org/10.1109/ICATCCT.2016.7912008

[41] Koulinas, G., Kotsikas, L., & Anagnostopoulos, K. (2014). A particle swarm optimization based hyper-heuristic algorithm for the classic resource constrained project scheduling problem. *Information Sciences*, *277*, 680–693. https://doi.org/10.1016/j.ins.2014.02.155

[42] Abba, S., & Lee, J. A. (2017). Bio-inspired self-aware fault-tolerant routing protocol for network-on-chip architectures using Particle Swarm Optimization. *Microprocessors and Microsystems*, *51*, 1339–1351. https://doi.org/10.1016/j.micpro.2017.04.003

[43] Okulewicz, M., & Mańdziuk, J. (2017). The impact of particular components of the PSO-based algorithm solving the Dynamic Vehicle Routing Problem. *Applied Soft Computing Journal*, *58*, 586–604. https://doi.org/10.1016/j.asoc.2017.04.070

[44] Singh, P. K., Nandi, S. K., & Nandi, S. (2019). A tutorial survey on vehicular communication state of the art, and future research directions. *Vehicular Communications*, *18*, 100164. https://doi.org/10.1016/j.vehcom.2019.100164

[45] Balzano, W., Murano, A., & Vitale, F. (2016). V2V-EN - Vehicle-2-Vehicle Elastic Network. *Procedia Computer Science*, *58*, 497–502. https://doi.org/10.1016/j.procs.2016.09.084

[46] Cherkaoui, B., Beni-Hssane, A., Fissaoui, M. El, & Erritali, M. (2019). Road traffic congestion detection in VANET networks. *Procedia Computer Science*, *151*, 1158–1163. https://doi.org/10.1016/j.procs.2019.04.165

[47] Boussoufa-Lahlah, S., Semchedine, F., & Bouallouche-Medjkoune, L. (2018). Geographic routing protocols for Vehicular Ad hoc NETworks (VANETs): A survey. *Vehicular Communications*, *11*, 20–31. https://doi.org/10.1016/j.vehcom.2018.01.006

[48] Darwish, T., Abu Bakar, K., & Hashim, A. (2017). Green geographical routing in vehicular ad hoc networks: Advances and challenges. *Computers and Electrical Engineering*, *64*, 436–449. https://doi.org/10.1016/j.compeleceng.2016.09.030

[49] Li, D., Deng, L., Cai, Z., Franks, B., & Yao, X. (2018). Intelligent Transportation System in Macao Based on Deep Self-Coding Learning. *IEEE Transactions on Industrial Informatics*, *14*(7), 3253–3260. https://doi.org/10.1109/TII.2018.2810291

[50] Benza, M., Bersani, C., D'Incà, M., Roncoli, C., Sacile, R., Trotta, A., Pizzorni, D., Briata, S., & Ridolfi, R. (2012). Intelligent Transport Systems (ITS) applications on dangerous good transport on road in Italy. *Proceedings - 2012 7th International Conference on System of Systems Engineering, SoSE 2012*, 223–228. https://doi.org/10.1109/SYSoSE.2012.6384180

[51] Malygin, I., Komashinsky, V., & Tsyganov, V. V. (2017). International experience and multimodal intelligent transportation system of Russia. *Proceedings of 2017 10th International Conference Management of Large-Scale System Development, MLSD 2017*, 1–5. https://doi.org/10.1109/MLSD.2017.8109658

[52] Seliverstov, Y. A., Malygin, I. G., Komashinskiy, V. I., Tarantsev, A. A., Shatalova, N. V., & Petrova, V. A. (2017). The St. Petersburg transport system simulation before opening new subway stations. *Proceedings of 2017 20th IEEE International Conference on Soft Computing and Measurements, SCM 2017*, 284–287. https://doi.org/10.1109/SCM.2017.7970562

[53] Lin, L. T., Huang, H. J., Lin, J. M., & Young, F. F. (2009). A new intelligent traffic control system for Taiwan. *2009 9th International Conference on Intelligent Transport Systems Telecommunications, ITST 2009*, 138–142. https://doi.org/10.1109/ITST.2009.5399369

[54] Poolsawat, A., Ayutaya, K. S. N., & Pattara-Atikom, W. (2009). Impact of intelligent traffic information system on congestion saving in Bangkok. *2009 9th International Conference on Intelligent Transport Systems Telecommunications, ITST 2009*, 153–156. https://doi.org/10.1109/ITST.2009.5399364

[55] Muchaendepi, W., Mbohwa, C., & Kanyepe, J. (2019). Intelligent Transport Systems and its Impact on Performance of Road Freight Transport in Zimbabwe. *IEEE International Conference on Industrial Engineering and Engineering Management*, *2019-December*, 80–83. https://doi.org/10.1109/IEEM.2018.8607409

[56] Nkoro, A. B., & Vershinin, Y. A. (2014). Current and future trends in applications of Intelligent Transport Systems on cars and infrastructure. *2014 17th IEEE International Conference on Intelligent Transportation Systems, ITSC 2014*, 514–519. https://doi.org/10.1109/ITSC.2014.6957741

[57] Zhu, T., & Liu, Z. (2015). Intelligent Transport Systems in China: Past, Present and Future. *Proceedings - 2015 7th International Conference on Measuring Technology and Mechatronics Automation, ICMTMA 2015*, 581–584. https://doi.org/10.1109/ICMTMA.2015.146

[58] Zhu, L., Yu, F. R., Wang, Y., Ning, B., & Tang, T. (2019). Big Data Analytics in Intelligent Transportation Systems: A Survey. *IEEE Transactions on Intelligent Transportation Systems*, *20*(1), 383–398. https://doi.org/10.1109/TITS.2018.2815678

[59] Vidya, V. M., & Deepa, N. (2019). Big data analytics in intelligent transportation systems using hadoop. *International Journal of Recent Technology and Engineering*, *7*(6), 75–80.

e-ISSN: 2614-8404
p-ISSN: 2776-3234

[60] Wu, F. J., Zhang, X., & Lim, H. B. (2014). A cooperative sensing and mining system for transportation activity survey. In *IEEE Wireless Communications and Networking Conference, WCNC* (pp. 3284–3289). https://doi.org/10.1109/WCNC.2014.6953075

[61] Chen, B., & Cheng, H. H. (2010). A review of the applications of agent technology in traffic and transportation systems. *IEEE Transactions on Intelligent Transportation Systems*, *11*(2), 485–497. https://doi.org/10.1109/TITS.2010.2048313

[62] Veres, M., & Moussa, M. (2019). Deep Learning for Intelligent Transportation Systems: A Survey of Emerging Trends. *IEEE Transactions on Intelligent Transportation Systems*, 1–17. https://doi.org/10.1109/tits.2019.2929020

[63] Zhang, J., A, A, A, A, A, A, A, & A. (2011). Data-driven intelligent transportation systems: A survey. In *IEEE Transactions on Intelligent Transportation Systems* (Vol. 12, Issue 4, pp. 1624–1639).

[64] Wang, X., & Yan, S. (2011). Design and implementation of intelligent public transport system based on GIS. *2011 International Conference on Electric Information and Control Engineering, ICEICE 2011 - Proceedings*, 4868–4871. https://doi.org/10.1109/ICEICE.2011.5777492

[65] Anand, B., Barsaiyan, V., Senapati, M., & Rajalakshmi, P. (2019). Real time LiDAR point cloud compression and transmission for intelligent transportation system. *IEEE Vehicular Technology Conference*, *2019-April*, 1–5. https://doi.org/10.1109/VTCSpring.2019.8746417

[66] Eckelmann, S., Trautmann, T., Ußler, H., Reichelt, B., & Michler, O. (2017). V2V-Communication, LiDAR System and Positioning Sensors for Future Fusion Algorithms in Connected Vehicles. *Transportation Research Procedia*, *27*, 69–76. https://doi.org/10.1016/j.trpro.2017.12.032

[67] Li, W., Ma, X., Wu, J., Trivedi, K. S., Huang, X. L., & Liu, Q. (2017). Analytical Model and Performance Evaluation of Long-Term Evolution for Vehicle Safety Services. *IEEE Transactions on Vehicular Technology*, *66*(3), 1926–1939. https://doi.org/10.1109/TVT.2016.2580571

[68] Li, J., Zhang, Y., Shi, M., Liu, Q., & Chen, Y. (2020). Collision avoidance strategy supported by LTE-V-based vehicle automation and communication systems for car following. *Tsinghua Science and Technology*, *25*(1), 127–139. https://doi.org/10.26599/TST.2018.9010143

[69] Correa, A., Boquet, G., Morell, A., & Vicario, J. L. (2017). Autonomous car parking system through a cooperative vehicular positioning network. *Sensors (Switzerland)*, *17*(4). https://doi.org/10.3390/s17040848

[70] Sakurada, L., Barbosa, J., Leitao, P., Alves, G., Borges, A. P., & Botelho, P. (2019). Development of Agent-Based CPS for Smart Parking Systems. *IECON Proceedings (Industrial Electronics Conference)*, *2019-October*, 2964–2969. https://doi.org/10.1109/IECON.2019.8926653

[71] Shi, Y., Pan, Y., Sun, X., Xie, R., Chen, W., & Shen, S. (2018). Collaborative Planning of Parking Spaces and AGVs Path for Smart Indoor Parking System. *Proceedings of the 2018 IEEE 22nd International Conference on Computer Supported Cooperative Work in Design, CSCWD 2018*, 496–500. https://doi.org/10.1109/CSCWD.2018.8465323

[72] Tilahun, S. L., & Di Marzo Serugendo, G. (2017). Cooperative multiagent system for parking availability prediction based on time varying dynamic markov chains. *Journal of Advanced Transportation*, *2017*. https://doi.org/10.1155/2017/1760842

[73] Abunei, A., Comsa, C. R., Caruntu, C. F., & Bogdan, I. (2019). Redundancy based V2V communication platform for vehicle platooning. *ISSCS 2019 - International Symposium on Signals, Circuits and Systems*, 9–12. https://doi.org/10.1109/ISSCS.2019.8801781

[74] Stüdli, S., Seron, M. M., & Middleton, R. H. (2017). Vehicular Platoons in cyclic interconnections with constant inter-vehicle spacing. *IFAC-PapersOnLine*, *50*(1), 2511–2516. https://doi.org/10.1016/j.ifacol.2017.08.449

[75] Kim, J., Han, Y., & Kim, I. (2019). Efficient Groupcast Schemes for Vehicle Platooning in V2V Network. *IEEE Access*, *7*, 171333–171345. https://doi.org/10.1109/ACCESS.2019.2955791

[76] Tripathy, R., Harmalkar, J., & Kumar, A. (2019). A functionally safe dual-bus platoon architecture for future smart cities. *Proceedings of the International Conference on Trends in Electronics and Informatics, ICOEI 2019*, *2019-April*(Icoei), 682–686. https://doi.org/10.1109/icoei.2019.8862618

[77] Bhoopalam, A. K., Agatz, N., & Zuidwijk, R. (2018). Planning of truck platoons: A literature review and directions for future research. *Transportation Research Part B: Methodological*, *107*, 212–228. https://doi.org/10.1016/j.trb.2017.10.016

[78] Larson, J., Liang, K. Y., & Johansson, K. H. (2015). A distributed framework for coordinated heavy-duty vehicle platooning. *IEEE Transactions on Intelligent Transportation Systems*, *16*(1), 419–429. https://doi.org/10.1109/TITS.2014.2320133

[79] Kokkinogenis, Z., Teixeira, M., D'Orey, P. M., & Rossetti, R. J. F. (2019). Tactical level decision-making for platoons of autonomous vehicles using auction mechanisms. *IEEE Intelligent Vehicles Symposium, Proceedings*, *2019-June*(Iv), 1632–1638. https://doi.org/10.1109/IVS.2019.8814122

[80] Maiti, S., Winter, S., & Kulik, L. (2017). A

e-ISSN: 2614-8404
p-ISSN: 2776-3234

conceptualization of vehicle platoons and platoon operations. *Transportation Research Part C: Emerging Technologies*, *80*, 1–19. https://doi.org/10.1016/j.trc.2017.04.005

[81] Dang, R., Ding, J., Su, B., Yao, Q., Tian, Y., & Li, K. (2014). A lane change warning system based on V2V communication. *2014 17th IEEE International Conference on Intelligent Transportation Systems, ITSC 2014*, 1923–1928. https://doi.org/10.1109/ITSC.2014.6957987

[82] Wang, L., Iida, R. F., & Wyglinski, A. M. (2018). Coordinated Lane Changing Using V2V Communications. *IEEE Vehicular Technology Conference*, *2018-Augus*, 1–5. https://doi.org/10.1109/VTCFall.2018.8690643

[83] Liu, X., Liang, J., & Xu, B. (2019). A Deep Learning Method for Lane Changing Situation Assessment and Decision Making. *IEEE Access*, *7*, 133749–133759. https://doi.org/10.1109/ACCESS.2019.2940853

[84] Sakr, A. H., Bansal, G., Vladimerou, V., & Johnson, M. (2018). Lane Change Detection Using V2V Safety Messages. *IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC*, *2018-Novem*, 3967–3973. https://doi.org/10.1109/ITSC.2018.8569690

e-ISSN: 2614-8404
p-ISSN: 2776-3234

Development of Student Associations Information System at Universitas Pembangunan Nasional Veteran Jakarta

Muhammad Adrezo[1], Rio Wirawan[2]
[1]Program Studi Informatika, Fakultas Ilmu Komputer, UPN Veteran Jakarta
[2]Program Studi Sistem Informasi, Fakultas Ilmu Komputer, UPN Veteran Jakarta
Email: [1]muhammad.adrezo@upnvj.ac.id, [2]rio.wirawan@upnvj.ac.id

Abstract – **Universitas Pembangunan Nasional Veteran Jakarta (UPN Veteran Jakarta) is one of the public universities which views student associations to play an important role in student self-development. Student's self-development can be realized if students participate in every activity. but a lot of problems that occur because the process related to student association is still done manually without using an information system, where students have to come to campus to take care of all the needs to hold an activity. So that we need a system that aims to improve services to student associations as well as facilitate the management of existing student associations data and can increase the credibility of UPN Veteran Jakarta itself. It is called SIWA. It is expected to minimize errors that occur and manage business processes that exist in each student association. So that the benefits of the system are that information on Real Estimate of Cost, submission of activity proposals, accountability reports and annual reports can be managed properly, minimize errors that occur and manage business processes that exist in each student association. Besides that, it can also support a paperless culture in the college environment. This information system is built based on a web-based system and its development uses the waterfall method.**

Keywords – Information System, Waterfall, Student Associations.

I. INTRODUCTION

In the development of technology era where technology can be found in all areas. Most organizations need to digitize business processes in their organization. Especially with the Covid-19 pandemic, organizations are necessary to optimize service from offline to online service. UPN Veteran Jakarta as a higher education institution recognized student associations have an important role as a medium for student's self-development. Student's self-development can be realized if students participate in every activity. On the other hand, student association have some problems in reporting activities in UPN Veteran Jakarta, especially in this pandemic where students must comply with existing health protocols and sometimes activity data is not stored properly so it is troublesome when they want to be reviewed. This happens because the process related to student association is still done manually without using an information system, where students have to come to campus to take care of all the needs to hold an activity.

Several studies have been conducted regarding student associations, Kurniawati, Hari and Darmanto [1], conducted research on the information system for the administration of student association activities (SIPAWA) at Widya Kartika University, Surabaya. This application is built using the Waterfall method. This system is used as a student association administration management information system (SIPAWA), Real Estimate of Cost information system, submission of student activity proposals, accountability reports and so that annual reports can be managed properly.

Research in this scope is also done by Ardian, Suryawan and Hartono [2], they make a system to manage the administration of student associations to help the institutions carry out supervision and guidance of student organizations at STIMIK STIKOM Indonesia. Meanwhile, the analysis and design used in this study is Structured Analysis and Structured Design method. The information system development uses Statement of Purpose (SOP), Event List, Context Diagram, Data Flow Diagram (DFD). Database design is done using the Entity Relationship Diagram (ERD). This system will provide information about the condition of student associations. Assessment of the condition of student associations is based on the activeness of student association member, the number of activities, student participation in student association activities, and discipline in student associations in terms of administration. The assessment was carried out using the Simple Additive Weighting (SAW) method. Furthermore, the development of the archive digitization application for the secretariat of student association in the STIMIK STIKOM BALI was carried out by Yuningsih [3], the Laravel framework was used in application development. In addition, research conducted by Pertiwi [4] and Pratiwi [5], shows that the relationship between student organizations and students is important in creating leadership and learning motivation on student achievement.

In this pandemic, a student associations information system is needed so that students don't need to come to campus to take care of documents related to student associations activities. This information system is used for Real Estimate of Cost, activity reports, evaluation of activities, etc. In order that all business processes in the student association can run even better in this era, especially during the Covid-19 pandemic, universities are required to provide online-based services supported by the readiness of technological devices at the university. With this system, UPN Veteran Jakarta can provide optimal

e-ISSN: 2614-8404
p-ISSN: 2776-3234

service to student association and can increase the credibility of UPN Veteran Jakarta itself. In addition, it is expected to minimize errors that occur and manage business processes that exist in each student association. So that the benefits generated later, it is hoped that information on Real Estimate of Cost, submission of activity proposals, accountability reports and annual reports can be managed properly. Besides that, it can also support a paperless culture in the college environment.

Based on that explanations, it can be said that this system is needed to improve the quality of service to student associations so as to facilitate data collection and submission of activities to be carried out by student organizations at UPN Veteran Jakarta. In its development, this system will be developed using the waterfall method. Waterfall has been used by many researchers in the system development process as has been done in research [6], [7], [8], [9] and [10].

II. RESEARCH METHODOLOGY

A. Data Collection

Data collection is the process of systematically collecting and confirming information about variables of interest. Where someone can answer questions about system requirements and discussion about the desired system. The following is an overview of the activities carried out at Universitas Pembangunan Nasional Veteran Jakarta in the data collection process for Student Association Information System development.

a) Observation
In this process, we find out the requirements for student association information system at Universitas Pembangunan Nasional Veteran Jakarta.

b) Interview
In this process, interview carried out to the parties concerned, namely student associations, student association supervisor, vice rector, Financial Division, Public Relation Division and the Division of Academic, Student Affairs, Planning and collaboration. To get information, data, and find out the process flow of the system. Table 1 shows number of correspondence.

Table 1. Correspondence

No	*Test Case*	*Number of Correspondence*
1	student associations	30
2	student association supervisor	5
3	vice rector	1
4	Financial Division	2
5	Public Relation Division	2
6	Division of Academic, Student Affairs, Planning and collaboration	2
Total		**42**

B. System Design

The research method used is the waterfall model methodology which explains the systematic stages because the process flows from beginning to end. Among them are system analysis, system design, implementation, testing and maintenance as shown in Figure 1. The model encompasses the following activities:

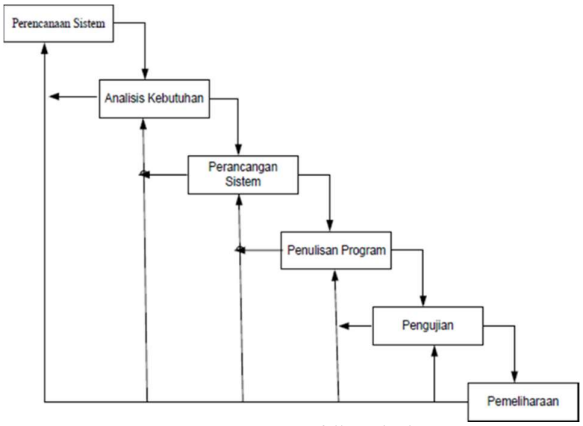


Figure 1. Waterfall Method

a) System Planning
In order to produce quality software, careful planning is needed by conducting a feasibility study. Feasibility studies include: economic, operational, and technical.

b) Requirement analysis
The purpose of system analysis is to determine problems in order to improve the system. So that it is hoped that it can work by analyzing the situation, then the existing problems will be resolved.

c) System Design
The design outlines screen layouts, business rules, process diagrams and other documentation. The results of this stage will describe the new system as a collection of modules or subsystems.

d) Coding/Implementation
In this stage, the implementation of designs and designs that have been carried out. So that at this stage it produces an information system (software).

e) Testing
After the software is built, testing is carried out to test the reliability of the software that has been built. This is done to ensure software reliability.

f) Maintenance
This stage aims to deal with the finished software so that it can function properly and avoid disturbances that cause damage. At this stage, updates can also be made to improve existing software.

C. Use Case

To describe the actors who interact with the system, use case diagrams can be used. The use case diagram is made in accordance with the business processes that have been identified in the system analysis. Functional and operational systems by defining usage scenarios agreed

e-ISSN: 2614-8404
p-ISSN: 2776-3234

upon between the user and the developer. The following is the entire SIWA use case at UPN Veteran Jakarta in Figure 2.

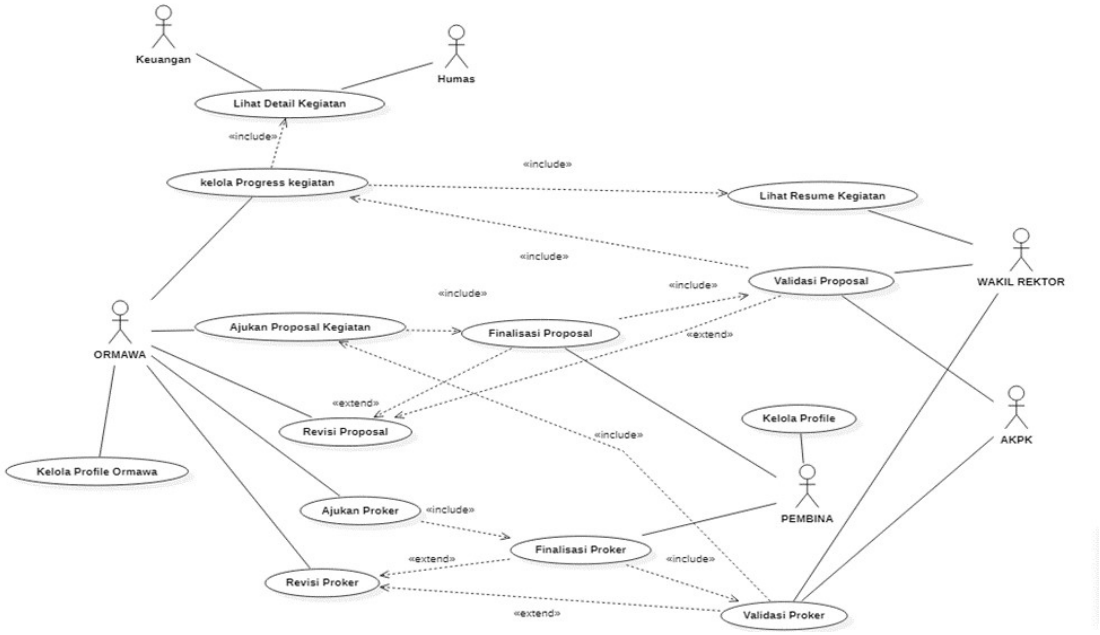


Figure 2. Use Case Diagram

III. RESULTS AND DISCUSSION

The Student Association Information System (SIWA) at UPN Veteran Jakarta has been made and taste as follows:

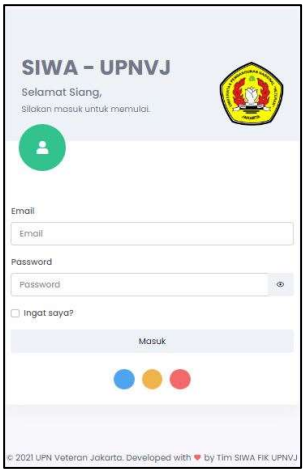


Figure 3. Login

Login page shows in Figure 3. When a user accesses the system, all users will go to the login page and be asked to enter their username and password. After that the user will automatically enter the dashboard page according to the user level.

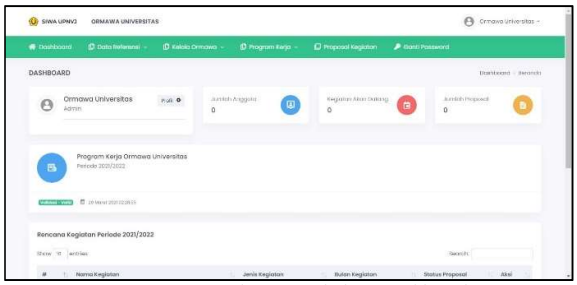


Figure 4. Student Associations Dashboard

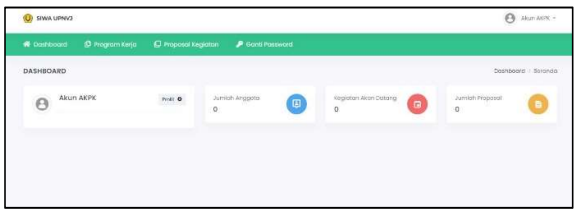


Figure 5. Verifier Dashboard

Figure 4 and Figure 5 show the dashboards for the student association level and the Verifier level. On the student associations dashboard, there are several menus including the Work plan menu, activity proposals and activity reports. Furthermore, there is a verifier dashboard which consists of two levels, supervisor level for each student association and the highest level verifier (vice rector and the Division of Academic, Student Affairs, Planning and collaboration) to verify work plan proposal and activity proposal. In Figure 6, Figure 7 and Figure 8 show work plan page, activity proposal page and activity planning list.

e-ISSN: 2614-8404
p-ISSN: 2776-3234

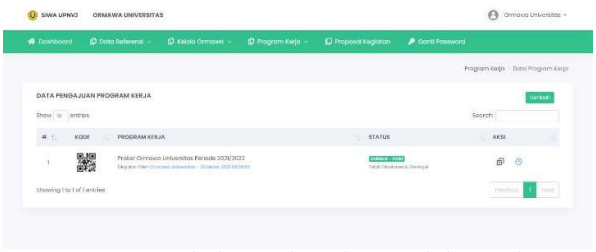


Figure 6. Work plan Page in Student Association Level

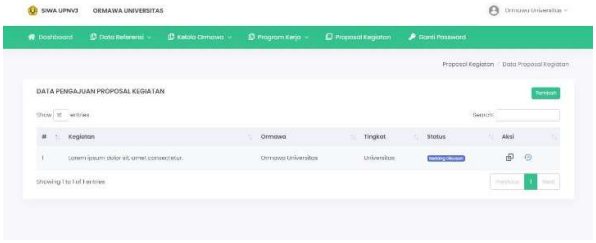


Figure 7. Activity Proposal Page in Student Association Level

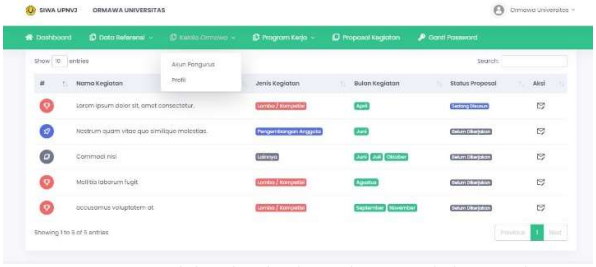


Figure 8. Activity Planning in Student Association Level

On the work plan page there is a list of work plans and we can add, edit and delete work plans. Furthermore, on the activity proposal page we can see a list of activity proposals that have been made and see the status of the proposal whether it has been accepted or not. And we can edit the proposal if there is a revision of verifier. On the dashboard, we can also see a list of activities that will be carried out throughout the annual work plan that has been made.

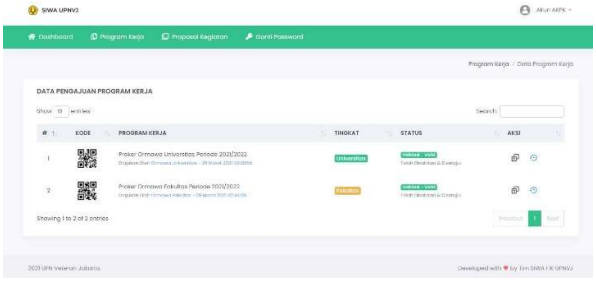


Figure 9. List of Work Plan Proposal Page in Verifier Level

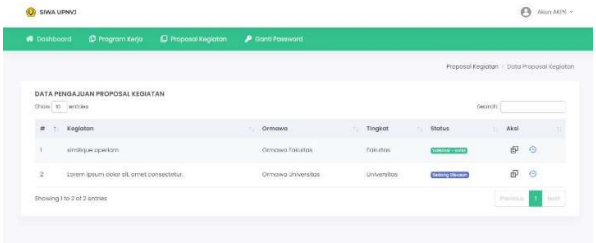


Figure 10. List of Activity Proposal Page in Verifier Level

Figure 9 and Figure 10 show work plan menu and activity proposal menu in verifier level. On the work plan menu and activity proposal in verifier level, Verifiers can view proposals and make notes for each part of the proposal. In addition, the verifier can provide a conclusion whether the proposal is accepted, revised or rejected.

In the development of this system, black box testing is carried out. Black box testing has been used by many researchers and developers to test a system. The following [11], [12], [13] and [14] are some studies that have been carried out using black box testing. Table 2 shows Black Box testing.

Table 2. Black Box Testing

No	*Test Case*	*Expected Output*	*Actual Output*	*Status*
1	Check the program of activities recorded in the work planning on the dashboard page	A list of activity programs that correspond to the initial work planning is displayed	A list of activity programs that correspond to the initial work planning is displayed	Pass
2	Click the proposal details in the action column	The system redirects to the proposal detail page	The system redirects to the proposal detail page	Pass
3	Check input with proposal data	Data can be saved by clicking save	Data can be saved by clicking save	Pass
4	Check the input form then click to another menu, then continue input the data	Data is not saved if you do not save first	Data is not saved if you do not save first	Pass
5	Check the input form then click save	The button will only appear if there is already saved data	The button will only appear if there is already saved data	Pass
6	check work planning proposal of student association	Can check the proposal	Can check the proposal	pass
7	Give feedback on the proposal	Can provide feedback on proposals	Can provide feedback on proposals	Pass
8	Click the approve proposal button (finalization)	Can provide proposal finalization status	Can provide proposal finalization status	Pass
9	Click the validation proposal button (validated)	Can provide validation	Can provide validation	pass

IV. CONCLUSION

Based on the stages taken in the development of the student association information system (SIWA), where the development uses the waterfall method and this system built based on web-based system, it can be concluded that the existence of this information system is able to assist in

e-ISSN: 2614-8404
p-ISSN: 2776-3234

organizing documents both proposals, reports, etc. so that it is easily accessible. it can also support a paperless culture in the college environment. In the other hand, Students no longer need to meet directly with their supervisors, especially during the Covid-19 pandemic.

REFERENCES

[1] Kurniati, F., Hari, Y., Darmanto. (2019). Pengembangan Sistem Informasi Pengelolaan Administrasi Kegiatan Organisasi Kemahasiswaan (SIPAWA) di Universitas Widya Kartika Surabaya. In *Prosiding SNST Fakultas Teknik.* 1(1), 107-112.

[2] Ardian, D. P. Y., Suryawan, I. W. D., Hartono, E. (2018). Sistem Informasi Pengelolaan Administrasi Organisasi Kemahasiswaan di STIMIK STIKOM Indonesia. Jurnal Teknologi Informasi dan Komputer 4(2), 156-165. DOI: 10.36002/jutik.v4i2.548.

[3] Yuningsih, L. (2017). Implementasi Framework Laravel pada Aplikasi Digitalisasi Arsip Sekretariat Organisasi Mahasiswa STIMIK STIKOM Bali. In *E-Proceeding KNS&I STIKOM Bali*, 379-383.

[4] Pertiwi, M.C., Sulistiyawan, A., Rahmawati, I., Kaltsum, H.U. (2015). Hubungan Organisasi dengan Mahasiswa dalam Menciptakan Leadership. In *Prosiding Seminar Nasional dan Call for Papers "Aktualisasi Bimbingan Konseling pada Pendidikan Dasar Menuju Peserta Didik yang Berkarakter"*, 227-234. ISBN: 978-602-70471-1-2

[5] Pratiwi, S.S. (2017). Pengaruh Keaktifan Mahasiswa dalam Organisasi dan Motivasi Belajar Terhadap Prestasi Belajar Mahasiswa Fakultas Ekonomi Universitas Negeri Yogyakarta. Jurnal Pendidikan dan Ekonomi 6(1), 54-64.

[6] Amrin, A., Larasati, M.D., Satriadi, I. (2020). Model Waterfall untuk Pengembangan Sistem Informasi Pengolahan Nilai pada SMP Kartika XI-3 Jakarta Timur. Jurnal Teknik Komputer 6(1), 135-140. DOI: 10.31294/jtk.v6i1.6884.

[7] Wiro, G. Samito. (2017). Penerapan Metode Waterfall pada Desain Sistem Informasi Geografis Industri Kabupaten Tegal. Jurnal Informatika: Jurnal Pengembangan IT (JPIT) 2(1), 6-12.

[8] Tabrani, M., Pudjiarti, E. (2017). Penerapan Metode Waterfall Pada Sistem Informasi Inventori PT. Pangan Sehat Sejahtera. Jurnal Inkofar 1(2), 30-40.

[9] Nere, M., Buani, D.W.P. (2018). Penerapan Metode Waterfall pada Sistem Informasi Jasa Laudry (SIJALY) JENSCHAX Laundy Bekas. Jurnal TECHNO Nusa Mandiri 15(2), 69-76.

[10] Purnia, D. S., Rifai, A., Rahmatullah, S. (2019). Penerapan Metode Waterfall dalam Perencanaan Sistem Informasi Aplikasi Bantuan Sosial Berbasis Anroid. In *Prosiding Seminar Nasional Sains dan Teknologi, 1-7.*

[11] Hanifah, U., Alit, R., Sugiarto. (2016). Penggunaan Metode Black Box pada Pengujian Sistem Informasi Surat Keluar Masuk. SCAN-Jurnal Teknologi Informasi dan Komunikasi 11(2), 33-40.

[12] Jaya, T.S. (2018). Pengujian Aplikasi dengan Metode Blackbox Testing Boundary Value Analysis (Studi Kasus: Kantor Digital Politeknik Negeri Lampung). Jurnal Informatika: Jurnal Pengembangan IT (JPIT) 3(2), 45-48.

[13] Cholifah, W. N., Yulianingsih, Sagita, S. M. (2018). Pengujian Black Box Testing pada Aplikasi Action & Strategy Berbasis Android dengan Teknologi PHONEGAP. Jurnal String, 3(2), 206-210.

[14] Ningrum F. C., Suherman, D., Aryanti, S., Prasetya, H. A., Saifudin, A. (2019). Pengujian Black Box pada Aplikasi Sistem Seleksi Sales Terbaik Menggunakan Teknik Equivalence Partitions. Jurnal Informatika Universitas Pamulang, 4(4), 125-130.

[15] Sutabri. (2005). *Sistem Informasi Manajemen.* Yogyakarta: Andi, 21.

[16] O'Brien, J. A. (2009). *Enterprise Business Management Information System*. McGraw-Hill/Irwin, 304.

e-ISSN: 2614-8404
p-ISSN: 2776-3234

Implementation of Social Network Analysis in the Spread of Natuna Issues on Twitter

Ashif Dzilfiqar Thayyibi[1*)], Juliana Mansur[2]
[1]Program Studi Magister Ilmu Komputer, Fakultas Teknologi Informasi, Universitas Budi Luhur
[2]Program Studi Magister Ilmu Komputer, Fakultas Teknologi Informasi, Universitas Budi Luhur
email: [1]ashifdzilfiqar@mail.com, [2]julianbisnis@gmail.com

Abstract* – *Currently, the growth of internet users has been accompanied by the development of applications that support interaction among users, which is called social media. One of the popular social media in society today is twitter. Data on Twitter can be presented in a graph structure visualization in nodes that represent actors and edges that represent relationships between actors. In an effort to find the most influential actors and actors who interact the most in spreading the Natuna topic on social media twitter, an analysis will be carried out using the Social Network Analysis method using the Degree Centrality approach. The data used in this study were taken from December 20, 2019 at 00.00 WIB to January 7, 2020 at 10.00 WIB consisting of 71,477 nodes and 147066 edges. The results of this study can be concluded that the @susipudjiastuti account is the most influential actor and plays an important role in social networking because the @susipudjiastuti account is the most linked account with 29755 links. Meanwhile, the @ shaktia704 account was the most active account during the data collection period, which reached 259 links.

***Keywords* – social network analysis; directed graph; degree centrality; twitter;**

I. INTRODUCTION

Currently, the growth of internet users has been accompanied by the development of applications that support interaction among users, which is called social media. One of the popular social media in society today is twitter. According to a survey conducted by eMarketer, there were 22.8 million active users of social media twitter in Indonesia in 2019 [1].Based on the survey conducted by eMarketer, it can be concluded that Twitter is a valuable source of social media data for analyzing opinions or opinions [2] and is an option in conducting research for social network analysis (social network analysis).

In Twitter, data can be visualized in a graphical structure formed by nodes that represent actors and edges that represent relationships between actors. The use of a graphical structure in the Twitter social media network makes it easy to visualize the nodes that have relations with other nodes, whether few or many relationships within the network or outside the group's network. The role of nodes in a central position in the group has an important function as control and stability in the group [3], which in turn shows the influence of these actors in the group. Edge and node relationships in twitter are referred to as follower (number of other actors who follow) and following (number of actors followed) between actors also affect how much popularity of actors in a group.

This research took the topic of Social Network Analysis on the Natuna Case on the Twitter Social Network. Natuna is one of the waters in the Riau Islands, which is one of the outer areas of the Unitary State of the Republic of Indonesia, which is an issue that is contested between Indonesia and neighboring countries. Having an area with abundant resources is one of the reasons Natuna is being contested. Each state claims that these waters belong to them with proof of their claim. In this study, data obtained from social media twitter was used on December 20, 2019 at 00.00 WIB to January 7, 2020 at 10.00 WIB.

Social Network Analysis is a study that studies human relations by utilizing graph theory. Looking at the problems above, the application of Social Network Analysis in an application that is able to describe the relationships between individuals by visualizing in the form of graphs is likely to help the process of solving existing problems. In addition, a calculation process will be carried out on each relationship between individuals to find the centrality of a social network based on the position of each related individual in the network structure [4].

Network according to Kadushin is defined as a collection of objects or nodes and a mapping or description of the relationships between these nodes. The relationship that occurs between one node and another is an edge or link [5].

According to Marin and Wellman, a social network is a group of nodes connected relevantly by one or more relationships. Nodes or network members are units that are connected by relations in the patterns studied [6]. Wasserman and Faust also expressed their opinion about social networks as a social structure consisting of individuals or organizations called "nodes", which are connected by one or more specific types of interdependence, such as friendship, kinship, common interest, exchange of money, dislike , knowledge or prestige [7].

This study aims to find centrality based on the degree centrality approach of a social network based on the position of each individual linked in the network structure. In addition, this study also conducted graphical visualization of existing relationships using the Gephi version 0.92 application to find the most influential actors and those with the most interaction.

II. RESEARCH METHODOLOGY

The interactions that occur in the “Natuna” case on the social media network twitter form an information

e-ISSN: 2614-8404
p-ISSN: 2776-3234

dissemination, then visualized and analyzed will produce information that is useful for improving the process of information dissemination by the government. Social network analysis itself is a science that studies the relationship between one entity unit and other entity units with the help of graph theory [8].

The SNA method and technique were chosen because this method can provide an overview or visualization down to the smallest relationship that occurs only in one individual to another in the network, this SNA method can also be used to find the nodes, communities, and informal hierarchies that have the most influence. in the network [9].

There are several concepts in the social network analysis approach, apart from describing the patterns formed from the relationships between nodes or actors, SNA is more often used to determine the central node in a network by calculating several centrality values, among which the commonly calculated ones are:

A. *Degree centrality* [10] calculates the number of interactions a node has. To calculate the degree centrality value of this node, it can be done using the following formula:

$$CD(ni) = d(ni) \quad (1)$$

Information:

d(ni) = the number of interactions this node has with other nodes on the network.

B. *Betweenness centrality* [11] calculate how often a node is passed by other nodes to go to a particular node in the network. This value serves to determine the role of the actor who is the bridge connecting interactions in the network. To calculate the degree centrality value of a node, it can be done using the following formula:

$$CB(ni) = \Sigma gjk(ni)/gjk \quad (2)$$

Information:

gjk (ni) = the number of shortest paths from node js at node k passing through node i.
gjk = the number of shortest paths between 2 nodes in the network.

C. Closeness centrality [12], calculating the average distance between a node and all other nodes in the network or in other words measuring the closeness of a node to other nodes. In a network with g nodes, the closeness centrality of these nodes is as follows:

$$CC(ni) = [N\text{-}1 / \Sigma\, d(ni, nj)] \quad (3)$$

Information:

N = the number of nodes in the network d (ni, nj) = the number of shortest paths connecting nodes ni and nj.

D. Eigenvector centrality [13], performs measurements that give higher weight to nodes connected to other nodes that also have high centrality values. To calculate the eigenvector centrality value of a node, it can be done using the following formula:

$$Ci(\beta) = \Sigma(\alpha + \beta cj)\, Aji \quad (4)$$
$$C(\beta) = \alpha(I\text{-}\beta A)\text{-}1A1$$

Information:

α = normalization constant (scale vector).
B = symbolizes how much a node has a centrality weight in a node that also has a high centrality value.

Where A is the adjacency matrix, I is the identity matrix and 1 is the matrix. The amount of β is the radius of power of a node. If β is positive, then it has high centrality bonds and is connected with people who are central. Meanwhile, if β is negative, then it has high centrality bonds but is connected to people who are not central. If β = 0, you will get degree centrality.

The method in this research is carried out in several stages as shown in the figure 1 [14].

e-ISSN: 2614-8404
p-ISSN: 2776-3234

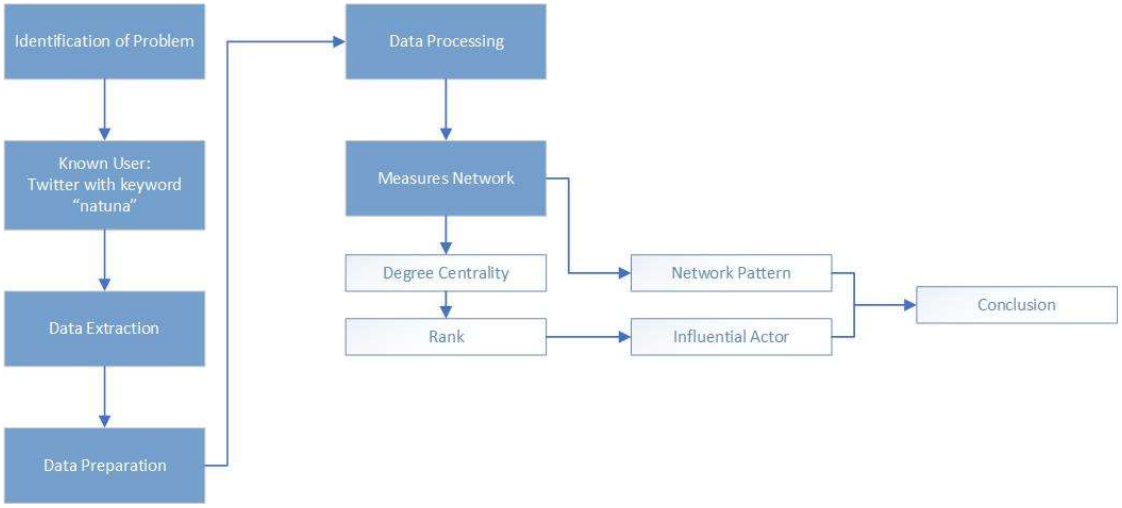


Figure 1. Research Methods

1. Identification of problems
   At this stage, the identification of the research problem is carried out. The process of identifying this problem is carried out by observing the phenomena behind the research.

2. *Known user*
   At this stage, determine the object of research. The object of research is the interaction carried out by accounts on the Twitter social network, which interact on the topic Natuna and with the keyword Natuna.

3. Data *extraction*
   At this stage, the data collection process is carried out on the twitter social network. The data taken contains the keyword natuna.

4. Data Processing and Measures Networks
   At this stage, network interaction data processing is carried out. The pattern used for graph visualization interaction uses directed type.

5. Centrality Value Calculation
   At this stage, the centrality value is calculated, namely the degree centrality node or actor to identify influential actors with a high number of interactions.

6. *Rank*
   This stage is the stage of ranking the calculated centrality values of actors in the network.

7. Draw conclusions
   In this stage, conclusions are drawn from the SNA analysis using the centrality approach.

Data processing in this study was carried out using Gephi software version 0.9.2. Gephi software is an open source application for network exploration and manipulation. A network module to be developed can be imported, visualized, mapped, filtered, manipulated and exported in the Gephi software [15]. Data processing in the software is carried out in the following steps:

(1) Import network data sets that have been previously created using the help of spreadsheets in Microsoft Excel. The data set that can be used is only the dataset with a text-based .csv extension. The dataset is separated into two parts, first import data set nodes and second step import data set edges. (The nodes dataset contains a list of actor names on the network, and the edges dataset contains data on relationships or interactions that occur between nodes on the network.

(2) Choose which visualization algorithm to use. This algorithm functions to determine the layout of the nodes to be visualized in the sociogram. In addition, the selection of the algorithm also affects the form of network visualization that will be produced. In this study, the algorithm used is the Fruchterman Reingold algorithm [15].

(3) Set the algorithm configuration by changing the attribute column such as area, gravity and speed available in the properties window according to the desired configuration, then click the run button.

(4) Personalize the visualized network. In this process, you will adjust the appearance of the color, shape, labeling the nodes in the network and you can also adjust the thickness and thickness of a line edge between the nodes and give a name to the edges.

(5) Calculates network property values. In this study, the value of network property attributes was calculated in the

e-ISSN: 2614-8404
p-ISSN: 2776-3234

form of the value of Total Node, Total Edges, Average Degree, Average Weighted Degree, Network Diameter, and Number of Communities. All these attributes can be calculated by clicking one button at a time in the setting column of the statistics window.

(6) Displays the ranking of nodes that have the highest influence or interaction value on the network. This step can be done in two ways, first by looking directly at the data table window, and the second is by configuring the display by changing the size of the node or node label in the network visualization image in accordance with the order of values owned by the nodes (the larger value owned by the node, the greater the appearance of the node in the network visualization image). The second way can be done by changing the configuration in the appearance window.

(7) Export the visualization image with the .pdf, .png and .svg file extensions. The results of data calculation can also be exported by accessing the menu in the data table window. The results of data processing will be a file with the extension .csv format.

III. RESULTS AND DISCUSSION

The research was conducted by visualizing the interaction data of Natuna's information dissemination on Twitter using Gephi software version 0.9.2. Twitter data with the keyword natuna taken from December 20, 2019 at 00.00 WIB to January 7, 2020 at 10.00, consists of 71,477 actors and 147,066 interactions.

The visualization in Figure 2 below shows the formation of several groups that have been differentiated based on the color of the network or so-called modularity class . From the visualization, it can be seen that there are 3 large groups that have the most interactions on the topic of Natuna, which are shown in purple, green, and blue. The formation of these groups uses a modularity approach in the Gephi application. This degree centrality approach calculates the number of interactions a node has (indegree and outdegree).

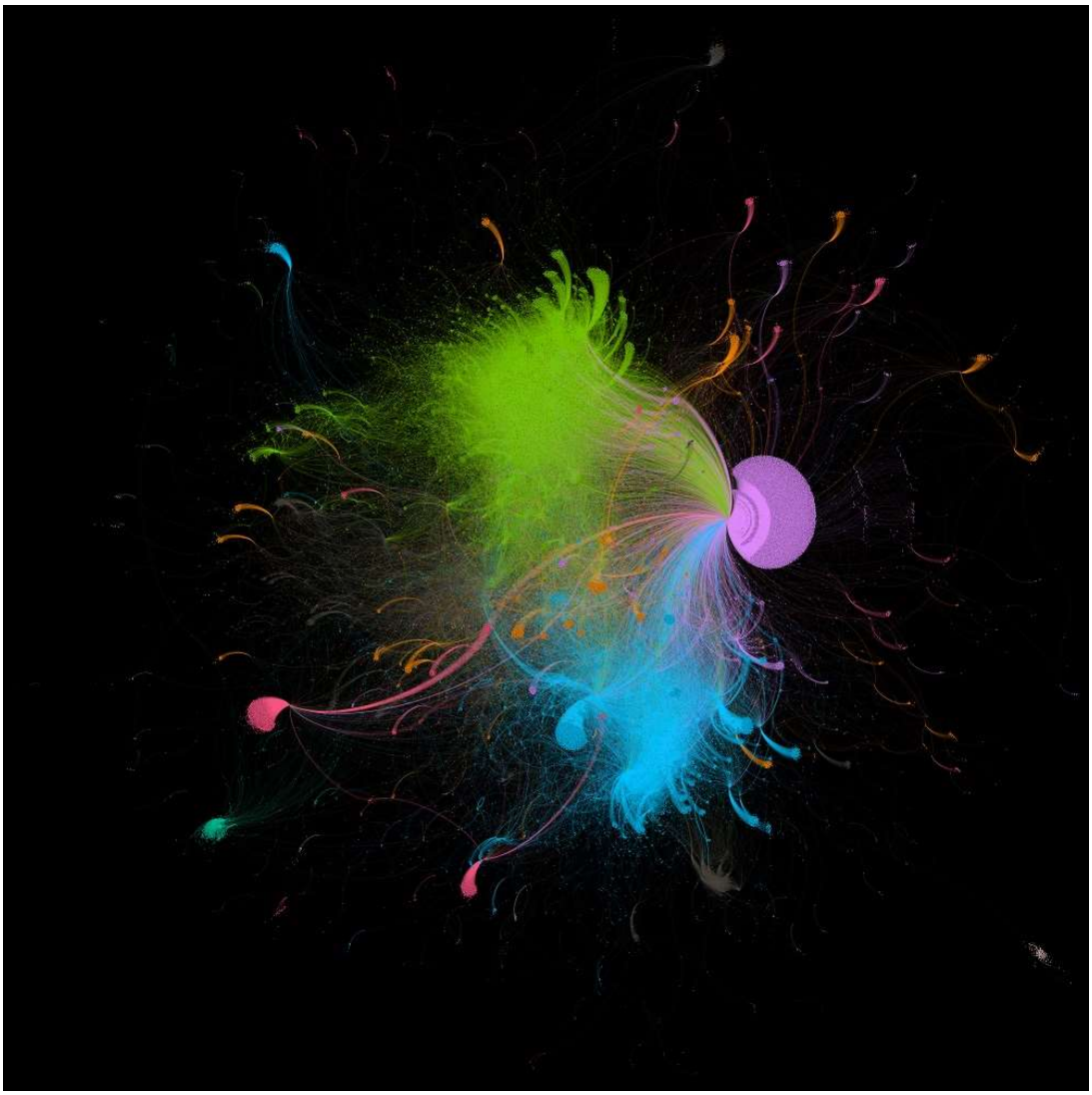


Figure 2. Visualization of the formation of issue groups

e-ISSN: 2614-8404
p-ISSN: 2776-3234

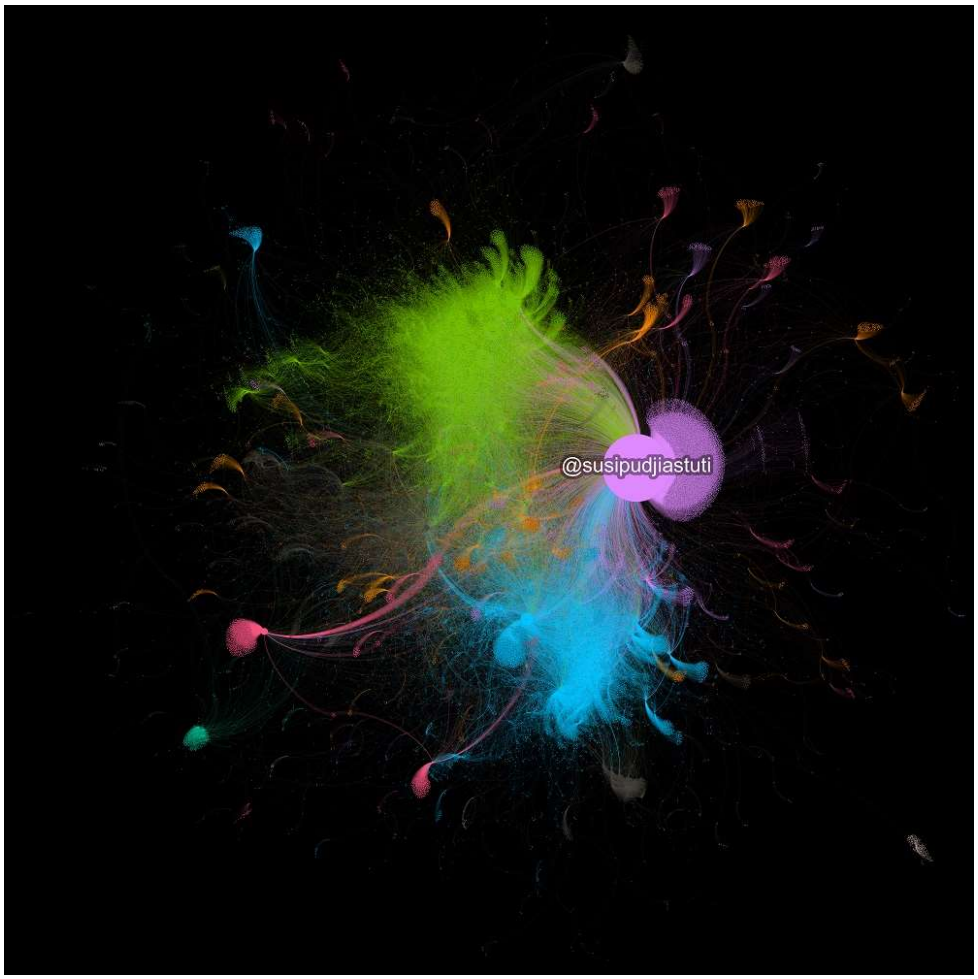


Figure3. Indegree Visualization

So for indegree is the total interaction received by a node or can be called to find which influencer / account has an important role in the network or the account that is most linked by other accounts, visualization can be independent can be seen in Figure 3. While outdegree is the total interaction which is made by a node / account which interacts most actively, outdegree visualization can be seen in Figure 4.

From Figure 3, namely indegree visualization, it can be seen that the @susipudjiastuti account is the actor who plays the most role in the spread of the Natuna topic because the @susipudjiastuti account is the most targeted. Notice table I below shows the top 10 influencers, a huge difference from the top 10 most linked accounts. The @susipudjiastuti account is in the first place linked by 29755 accounts. The big difference is that the visualization only shows @susipudjiastuti, while other accounts are so small that they are not visible on the visualization.

Tabel 1. Indegree value top 10 accounts

No	Label	Node_type	In-Degree
1	@susipudjiastuti	TW_Account	29755
2	@jokowi	TW_Account	5648
3	@D4tuk_T4mburin	TW_Account	3519
4	@JeromePolin	TW_Account	2817
5	@ustadtengkuzul	TW_Account	2381
6	#natuna	Hashtag	2191
7	@geloraco	TW_Account	2165
8	@hnurwahid	TW_Account	1837
9	@do_ra_dong	TW_Account	1621
10	@_TNIAU	TW_Account	1553

Figure 4 shows the outdegree visualization, which is the account that is most actively interacting, namely the @ shaktia704 account, in this visualization only the @ shaktia704 account is visible because of the difference in the value of interactions on that account. See table 1 below for a breakdown of the top 10 account outdegree values.

Tabel 2. Outdegree value top 10 accounts

No	Label	Node_type	Out-Degree
1	@shaktia704	TW_Account	259
2	@GunGunG49169853	TW_Account	124
3	@ANDRE_OCTA_	TW_Account	90
4	@AlvaroDeBazan2	TW_Account	88
5	@AliDavala99	TW_Account	86
6	@zenoldgonzalez	TW_Account	73
7	@margaretaputr13	TW_Account	62
8	@stevansixciokre	TW_Account	60
9	@AriestaRiico	TW_Account	59
10	@MaskuMasku1	TW_Account	59

e-ISSN: 2614-8404
p-ISSN: 2776-3234

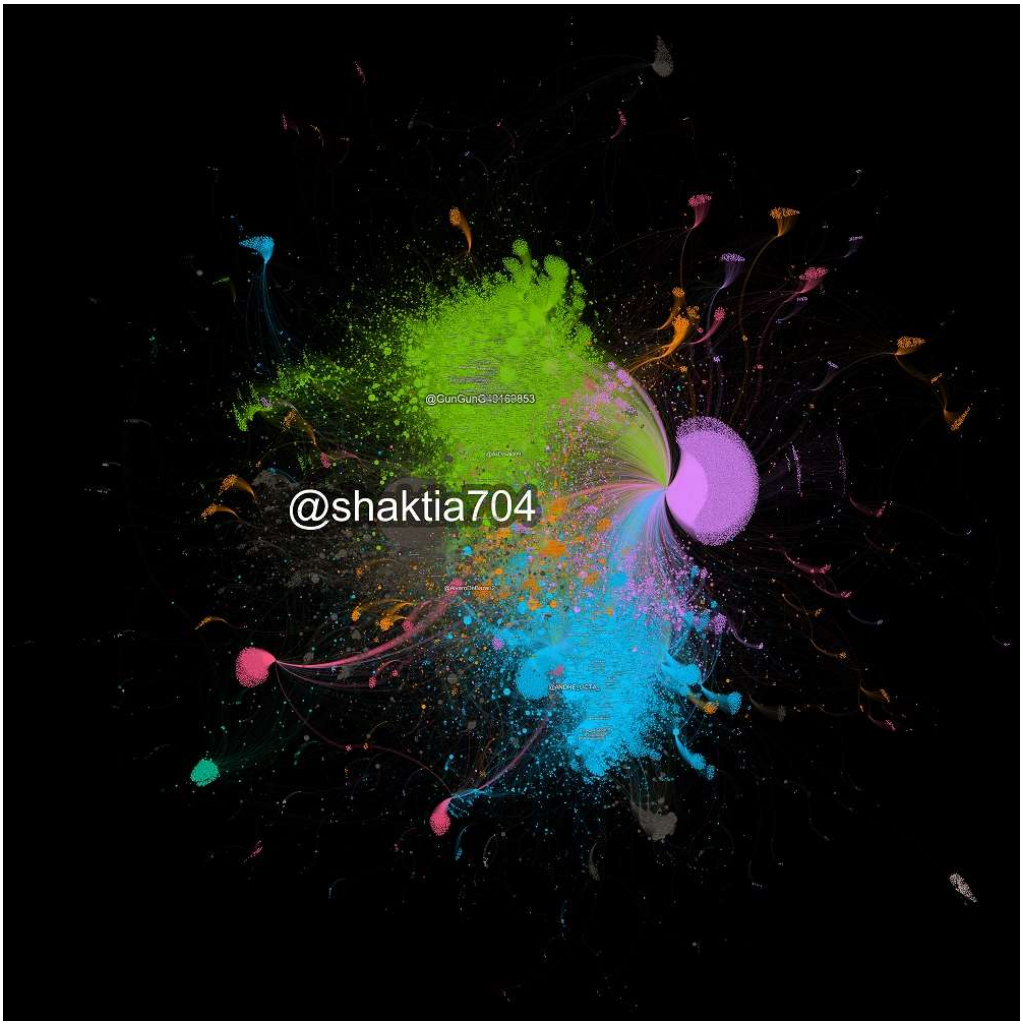


Figure 4. Outegree Visualization

In table 1, it can be seen from the results of in-degree calculations that the @susipudjiastuti account has the highest value, this is because the @susipudjiastuti account is the main actor who plays a role in the topic of Natuna. Then in table 2, it can be seen that the @ shaktia704 account has the highest out-degree value because the @ shaktia704 account is the most actively interacting.

Table 3. Betweenness Centrality scores top 10 accounts

No	Label	Node_type	Betweenness
1	@Aryprasetyo85	TW_Account	6.387338011628019E7
2	@susipudjiastuti	TW_Account	6.2843006970944755E7
3	@D4tuk_T4mburin	TW_Account	1.3921774279551372E7
4	@Na_T1N4	TW_Account	1.3921774279551372E7
5	@DonAdam68	TW_Account	1.3457628354689362E7
6	@Zahrah40291660	TW_Account	1.2423828393732902E7
7	@jr_kw19	TW_Account	8955050.328199675
8	@7intaPutih	TW_Account	6906324.800256199
9	@johhhnygudhel	TW_Account	6619723.706064895
10	@Luana0119411577	TW_Account	5196291.787102646

Table 3 shows the top 10 rankings betweenness centrality calculations obtained from Gephi. From these calculations it can be seen that there are 2 accounts that choose the highest value, namely the @Aryprasetyo85 and @susipudjiastuti accounts. This shows that the two accounts become a bridge or meeting point for netizens in their discussion of Natuna on social media twitter.

e-ISSN: 2614-8404
p-ISSN: 2776-3234

Figure 5. Betweenness Visualization

Figure 5 visualizes the results of betweenness centrality, which shows that the @ Aryprasetyo85 and @susipudjiastuti accounts are meeting points for netizens on issues about Natuna on Twitter.

Table 4. Eigenvector Centrality scores top 10 accounts

No	Label	Node_type	Eigenvector
1	@susipudjiastuti	TW_Account	1.0
2	@jokowi	TW_Account	0.19216753986353677
3	@D4tuk_T4mburin	TW_Account	0.11837080026996827
4	@JeromePolin	TW_Account	0.09457074633210014
5	@ustadtengkuzul	TW_Account	0.08003242969610756
6	#natuna	Hashtag	0.07804578730451751
7	@hnurwahid	TW_Account	0.061940705758999355
8	@do_ra_dong	TW_Account	0.055108563151993695
9	@_TNIAU	TW_Account	0.052416368798993716
10	#jokowikawalnatuna	Hashtag	0.049589729591855675

Table 4 shows the top 10 ranking results of the Eigenvector centrality calculation obtained from Gephi. From these calculations it can be seen that the account with the highest value is the @susipudjiastuti account. This shows that the @susipudjiastuti account is the center of conversation in this case is the Natuna case, because all influencers are related to the @susipudjiastuti account.

The results of eigenvetor centrality can be seen in Figure 6 which visualizes that the @susipudjiastuti account is the center of conversation among netizens in the Natuna case on Twitter.

e-ISSN: 2614-8404
p-ISSN: 2776-3234

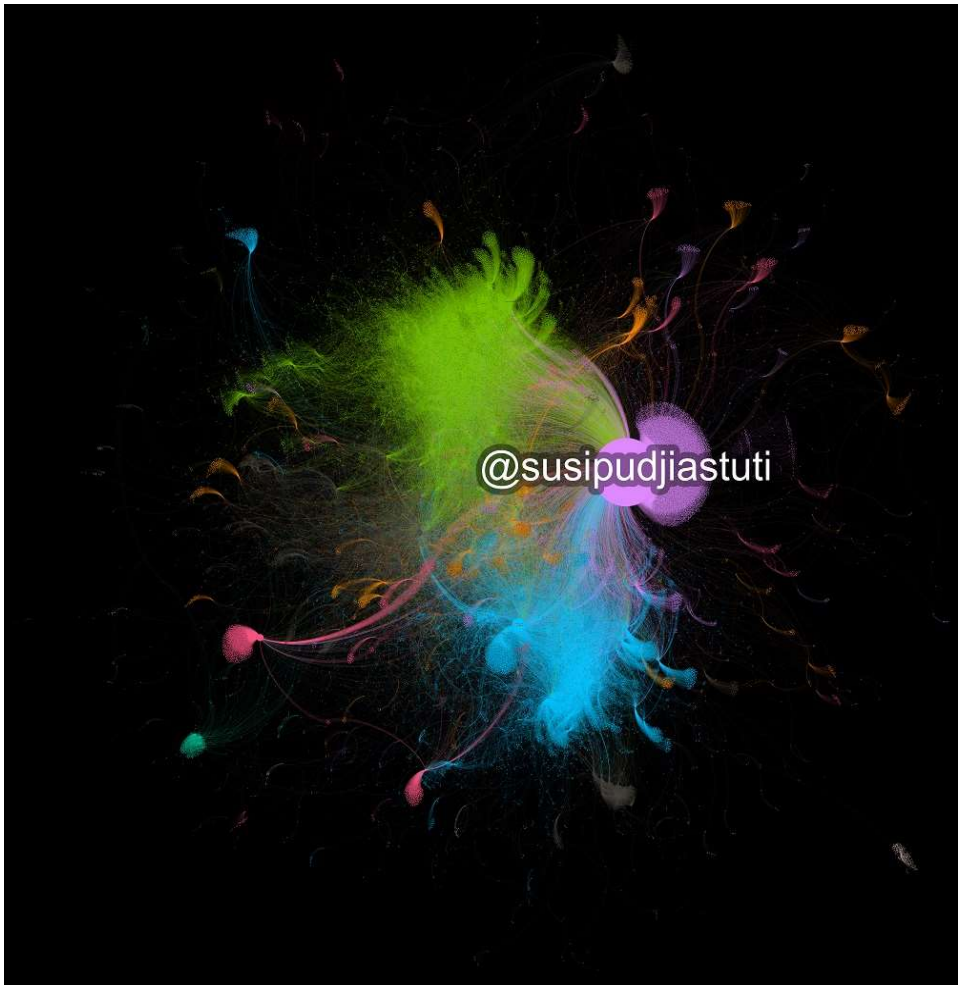


Figure 6. Eigenvector Visualization

IV. CONCLUSION

In the discussion of the results of the research that has been carried out, in order to answer the research questions that have been formulated, it can be concluded as follows:

1. The @susipudjiastuti account has an important role in the network because the account is often linked, reaching 29755.
2. The @shaktia704 account is the most active in the period of data taken, which reached 259.
3. The @susipudjiastuti and @ Aryprasetyo85 accounts become a bridge or meeting point for netizens in discussions about Natuna.
4. The @susipudjiastuti account became the center of discussion in the Natuna case.

REFERENCES

[1] R. Nugraha, "Trending Topik Twitter sebagai Sarana Penyebaran Isu Kontroversial," *www.kompasiana.com*, 2020. https://www.kompasiana.com/nugraha88/5e1087aed541df79920f6b22/trending-topik-twitter-sebagai-sarana-penyebaran-isu-kontroversial?page=all (accessed Jan. 17, 2021).

[2] L. Zhang, R. Ghosh, M. Dekhil, M. Hsu, and B. Liu, "Combining lexicon-based and learning-based methods for twitter sentiment analysis," *HP Lab. Tech. Rep.*, no. 89, 2011.

[3] S. Fortunato, "Community detection in graphs," *Phys. Rep.*, vol. 486, no. 3–5, pp. 75–174, 2010, doi: 10.1016/j.physrep.2009.11.002.

[4] B. Susanto, H. Lina, and A. R. Chrismanto, "Penerapan Social Network Analysis dalam Penentuan Centrality Studi Kasus Social Network Twitter," *J. Inform.*, vol. 8, no. 1, 2012, doi: 10.21460/inf.2012.81.111.

[5] C. Kadushin, "Kadushin , Charles . Understanding social networks : theories , Kadushin , Charles . Understanding social New York , NY : Oxford

e-ISSN: 2614-8404
p-ISSN: 2776-3234

University Press , Keen On Mass Comms ?," vol. 18, no. April, pp. 2013–2015, 2012.

[6] A. Marin and B. Wellman, *Handbook of Social Network Analysis*. 2009.

[7] S. Wasserman and K. Faust, *Social Network Analysis: Methods and applications*. Cambridge University Press, 1994.

[8] M. Tsvetovat and A. Kouznetsov, *Social Network Analysis for Startups*. 2011.

[9] A. Bohn, I. Feinerer, K. Hornik, and P. Mair, "Content-based social network analysis of mailing lists," *R J.*, vol. 3, no. 1, pp. 11–18, 2011, doi: 10.32614/rj-2011-003.

[10] F. Bloch and M. O. Jackson, "Centrality Measures in Networks," *SSRN Electron. J.*, no. January, 2016, doi: 10.2139/ssrn.2749124.

[11] F. Y. Pratama, "Simulasi Jejaring Jalan Kota Pontianak Dengan Betweeness Centrality dan Degree Centrality," *Progr. Stud. Tek. Ind. Fak. Tek. Univ. Tanjungpura*, pp. 1–6, 2018.

[12] J. Zhang and Y. Luo, "Degree Centrality, Betweenness Centrality, and Closeness Centrality in Social Network," vol. 132, no. Msam, pp. 300–303, 2017, doi: 10.2991/msam-17.2017.68.

[13] G. Lohmann *et al.*, "Eigenvector centrality mapping for analyzing connectivity patterns in fMRI data of the human brain," *PLoS One*, vol. 5, no. 4, 2010, doi: 10.1371/journal.pone.0010232.

[14] M. S. Setatama and D. Tricahyono, Ir., M.M., Ph.D., "Implementasi Social Network Analysis pada Penyebaran Country Branding 'Wonderful Indonesia,'" *Indones. J. Comput.*, vol. 2, no. 2, p. 91, 2017, doi: 10.21108/indojc.2017.2.2.183.

[15] T. M. J. Fruchterman and E. M. Reingold, "Graph Drawing by Force-directed Placement," *Softw. Pract. Exp.*, vol. 21, pp. 1129–1164, 1991.

e-ISSN: 2614-8404
p-ISSN: 2776-3234

Online System on Monitoring and Feedback for Education

Sudirman[1*)]
[1]Program Studi Pendidikan Matematika, Fakultas Keguruan dan Ilmu Pendidikan,
Universitas Qamarul Huda Badaruddin Bagu
Email: sudirman@uniqhba.ac.id

Abstract – This study reported the staging of process on developing a mobile application for real-time data management information system on monitoring and feedback in early childhood education, it can help tracking child development and education and assist teacher in monitoring and feedback on child services. A formative study was carried out to gather necessary information through data mapping, in-depth interviews with key stakeholders, document reviews, application development, deployment the application, field assessments, usability testing and integrated analytics that involving 253 respondents. To obtain a full picture on early childhood education, data on child growth and education shall be mapped and linked in one mobile application. The monitoring and feedback of mobile app were conducted by tracking of performance using Meraki software include connectivity, battery charge, disc usage, last online time, and location. Tracking of data entry onto the application by the users through reporting dashboard installed. Analysis from the daily tracking data presented to the users through coaching activity to monitor their performance indicators are on-time submission was 89.38%, completeness percentage data quality were 93.38%. Using a tablet PC or mobile phone, data could be easily entered at any time by the person. Due to still poor infrastructure at the grass root level, the system also allows a safety store offline that could automatically link to server when network connection is available. Based on the study this application is applicable for online monitoring and feedback on early child development and education. Results suggest the use of real time rapid analysis of these routine assessments of provider performance and the application usability enables a dynamic process of continuous quality improvement.

Keywords – Online System, Monitoring, Feedback, Education

I. INTRODUCTION

Many field workers who rely on paper record-keeping complain that manually compiling monthly and yearly reports for their supervisors takes more time than it should. Moreover, supervisors complain that reports they receive from workers are incomplete or poorly compiled [1]. One of the benefits of switching to a paperless record-keeping system is the ability to automate and standardize reporting at all levels in the field system [2]. Data entered could be automatically synced with the reporting module, so workers can access and compile their reports at any time [17]. They can easily track their progress during the month or year without having to manually compile data. Supervisors and reporting authorities can rest assured that the data being reported is accurate and reflects real service provision and health events on the ground. They can easily detect anomalies with digitized reporting and significantly reduce the time to respond to an emergency, such as a disease outbreak, when it occurs [2].

In rural areas, or anywhere field workers might be spread out and hard to reach, having an online web portal and dashboard for daily monitoring is an efficient and smart way to ensure workers are regularly providing timely care to their clinets [3]. The smart registry web portal allows end user login for monitoring clinet data and printing paper reports of their data if required for submission [18]. Supervisors at higher levels can login to monitor their education workers and view their service provision in real time along with aggregate data across all workers at a particular field level. The web portal can also archive data, in case a education worker needs to review older records which are no longer stored on the application [4].

Paper registers present strategic challenges for tabulation and access to real-time data for decision-making, monitoring Frontline Education Workers (FEW) performance at district or national level, and providing a reasonable level of accountability for authentic and complete individual data records. The burden of paper-based reporting takes valuable time away from service provision, often results in the duplication of information across multiple registers, and requires manual tabulation of the data for summary reporting [20]. Paper-based data also does not facilitate continuity of care between visits or across providers. Tracking services provided to a client requires scrutiny of multiple documents, which are often not organized by client name. Consequently, clients who have missed services or appointments are not identified in a timely fashion, leading to a missed window of opportunity for intervention. Developing aggregate monthly or quarterly reports from this paper-based data is error-prone and time consuming, a task that is repeated at several levels of supervision before a compiled report reaches the senior management layer. By the time data reaches the decision maker, the opportunity to use the data for real-time strategic decisions has passed[20].

The ubiquity of mobile technology, such as phones and tablets, and their increasing penetration among even the most remote and marginalized populations has provided a platform for innovators to creatively target pervasive education system challenges. Even in the absence of structured mobile interventions, education system actors have used mobile devices to improve communications across the developing world, and as such provide an opportunity to strengthen education system [21].

Monitoring following an application notably to oversee the compliance of the application usage by the front

e-ISSN: 2614-8404
p-ISSN: 2776-3234

education workers. This will also allow us to understand how each user utilizes the application in daily basis. Furthermore, any issue either regarding the application/device or the utilization itself can be early detected through this research.

There is a clear and urgent need for an integrated education information system to generate quality data, reduce the workload of frontline education workers, and provide data in realtime for supervisor and policy makers to guide strategy and improve education outcomes.

.

II. RESEARCH METHODOLOGY

A formative study was carried out to gather necessary information through data mapping, in-depth interviews with key stakeholders, document reviews, application development, deployment the application, field assessments, usability testing and integrated analytics that involving 253 respondents.

The methodology used in developing the application as follow:

2.1 Literature Study

Collecting data from books, literatures, or objects related to the topic of Child care including module of Community Development Workers was published by World Bank and Partners in 2013. Anthropometry calculation algorithm with Z Score. for the classification, we using the table calculation from "Anthropometry book 2010" from Indonesia Ministry of Health and for weight indicator, we using standard of red borderline[19].

2.2 Hardware and Software Requirements

Child care application consists of two application system that is Application Server which keeps overall data from Client Application which only contains data based on coverage area of community development in field. The server is responsible for requests for both existing and new data.

Tools Selection: tools used to build information systems based on Android mobile is Java and MySQL programming.

IDE: Android Studio is the official Integrated Development Environment (IDE) for Android app development, based on IntelliJ IDEA.

2.3 Hardware Specification

a. Server

Child care and education application was required server specifications as below: Product/Service Virtual Private Server - VPS Hazelnut Active, Backup Quota 10 x GB, OS Template is debian-8.0-x86_64, Platform Linux x86_64, OS Package Debian GNU/Linux 8.0 (for AMD64/Intel EM64T) OS EZ template, CPU Cores 4, CPU Limit 1600 MHz, Memory 4.00 GB, HDD 80.0 B.

b. Client

The specification of client as follow : Screen size 5”, Brand Samsum Galaxy, OS Android 6.0, RAM 2 GB, Storage 16B.

c. Laptop

Laptop Specifications are Brand is Lenovo, Type Legion Y520-N21D, Processor is Intel Core i7-7700HQ, RAM 16 GB DDR4; 2 x SODIMM Slot, HDD 1TB SATA SSD + 256GB, VGA using NVIDIA®GeForce®GTX 1050 Ti, DVD Writer, Screen 15.6", DOS

a. Webserver: Nginx

1. Programming Language Backend: PHP version 7 with Framework used is Laravel version 5.0.
2. Database: MySQL has several features among others
3. Portability: Supports various operating systems like Windows, Linux, FreeBSD, Mac OS X Server, Solaris, Amiga, and more.
4. Open Source : MySQL is distributed open source, under the GPL license so it can be used free of charge.
5. Multiuser: MySQL can be used by multiple users at the same time without any problems or conflicts.
6. Performance Tuning: MySQL has an amazing speed in handling simple queries, in other words it can process more SQL per unit of time.
7. Column Type: MySQL has a very complex column type, such as signed / unsigned integer, float, double, char, text, date, timestamp, and others.
8. Commands and Functions: MySQL has full operators and functions that support SELECT and WHERE commands in the query.
9. Security: MySQL has several layers of securities such as sub netmask level, hostname, and user permissions with a detailed licensing system as well as encrypted passwords.
10. Scalability and Restrictions: MySQL is capable of handling large-scale databases, with more than 50 million records and 60 thousand tables and 5 billion rows. In addition, the index limit that can fit up to 32 indexes in each table.

b. Procedures of Application Design

The design of the application was done by designing UI (User Interface) design, database like variables/fields, values, label, logical check, range check, calculation and function design in the application.

a. Logic

The use case staff described the interaction between the teacher and the system. Tutor/teacher are required to login in order to ensure access to data. Login using username and Password for each teacher. Username and Password will be sent to the server for validation. The server that receives data from the application detects every request based on the URL address it receives. Data from this URL address will determine the type of request the server must perform at the same time respond to requests with the appropriate data. If there is conformity with existing staff

e-ISSN: 2614-8404
p-ISSN: 2776-3234

data it will be replies in the form of basic data from Staff and data in the form of access rights code. The basic data that is sent back to the client is: Full name, NIK, Locations covering the names of sub-districts, districts and provinces After successfully logging in, the first thing to do is fill in the local database by requesting the server to send data based on the location of the work area of the staff officer. After the initial data is filled then the activity can be done to recording the data of visits to the Child care service. Data from this Server as the basic data used for Mother and Child Identity in Child care applications.

b. Activity
Activity describe activity in the system has been built, how each flow begins, the decisions that may occur, and how they end. Activity diagrams can also describe parallel processes that may occur in some executions. Activity diagram consists of staff login process, master data management process, process.

c. Class
Class described the state (attribute / property) of a system, all at once offer services to manipulate the situation (method / function). Class is a specification which, if instantiated, will produce an object and is the core of object-oriented development and design.

d. Database SQLite
The database design was a translation of the class diagram in the form of tables containing field names, field types, key types, and field actions. Database for data storage in This application, which is useful to accommodate the required data.

e. UIX
Based on Use case and flowcharts. Each element break into deliverable and lay down a strategy to go ahead with. Design part of (UI/UX) and prepare a design that delivers the best user experience. The UI prototype tested on different devices. We make sure that smooth navigation on the Mobile App that already created.

c. Coding
Perform database creation and mobile application development. This project used Model View Presenter (MVP) is the latest and greatest Android architecture pattern. This decouple business logic (Model) from view logic (Activity / Fragment) by utilizing an intermediate step called the Presenter.

a. **View**
The view was extremely limited in MVP, it's only works on display data and navigate to a new screen when the presenter tells it to. The View has no visibility of the Model, except for the POJOs / Entities. In regards to Android specifically, this would include my Activities, Fragments, Recycler View Adapters, and anything that extends the Android View class. My personal preference is to only let the Activities & Fragments talk to the Presenters and leave the Views & Adapters to only display data and delegate events (On Click) back to the Activity or Fragment.

b. **Presenter**
The presenter lays between the View and the Model, and it acts to events passed from the view. For instance, when the Finish Button (to save data) is clicked inside the view, it would call presenter. Save (). Once this occurs, the presenter utilizes the Model to determine if all criteria are met (I.e. a valid email address) and, if so, we can safely save the data. The presenter would then either notify the view to display an error message, or notify the view to navigate.

c. **Model**
The Model includes business logic that was entirely decoupled from the UI / Platform specific logic. This encompasses my Entities, backend services / helpers (web, database, etc), and business logic. It will used wrapper class (either called Model or Interactor) that will talk directly to the backend services and hold business logic.

d. Z Score calculation

a. Data Source
we used growth chart indicator on daily basis that can be downloaded at http://who.int/childgrowth/standards/ on file that contains the percentiles of z score. The file contains the coefficients of L, M and S that can be used to calculate Z Score.

b. Calculation
Flow process of calculation as follow, app would prepare all the indicator into 2-dimensional array with 4 column each, contains age, L, M and S value, user input the child visit date, weight and length/height data into app, app calculate the children's current age on day unit, by find the range between visit date and the child birth date. by using the age (calculation result on step b) as the index, app would get L, M and S data on that index and calculate Z Score using LMS formula below:

$$Z_{ind} = \frac{\left[{}^{y}/_{M(t)}\right]^{1^{L(t)}-1}}{S(t)L(t)} \quad (1)$$

Anthropometry has a same calculation algorithm with Z Score. for the classification, we using the table calculation from anthropometry book from Indonesia Ministry of Health.

e. App design flow and Testing
This stage was done to test the functionality of applications that have been built **(**Figure. 1**).**

e-ISSN: 2614-8404
p-ISSN: 2776-3234

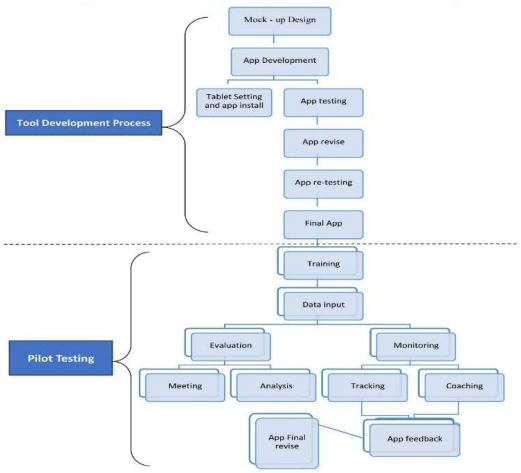


Figure 1. Design flow of application development

III. RESULTS AND DISCUSSION

To obtain a full picture on childhood development and education in early education, data on child growth and education shall be mapped and linked in one application. We introduced a mobile app to systematically compile the individual as well as group data (i.e. school profiles) across different aspects of child life, ranging from Child care and education [5]. Using a tablet PC or mobile phone, data could be easily entered at any time (real-time data) by the person (Figure 2).

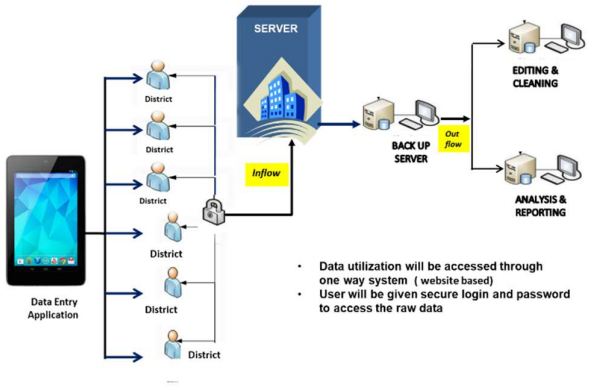


Figure 2. Data Management System

The monitoring and feedback were basically conducted every day through following several ways: Tracking of tablet or smartphone performance was using tracking software, Meraki (Figure 3.) installed in the device prior to deployment. The tablet or smartphone performance indicators include connectivity, battery charge, disc usage, last online time, and location.

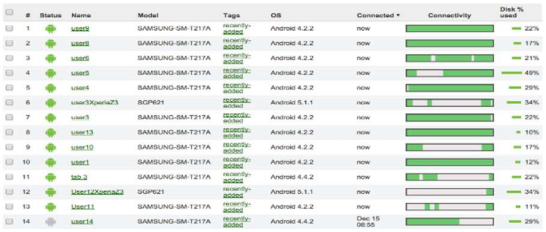


Figure 3. Tracking result by Meraki Software

Tracking of data entry onto the application by the users through a Reporting Dashboard built for monitoring and evalution use (Figure 4) and (Figure 5).

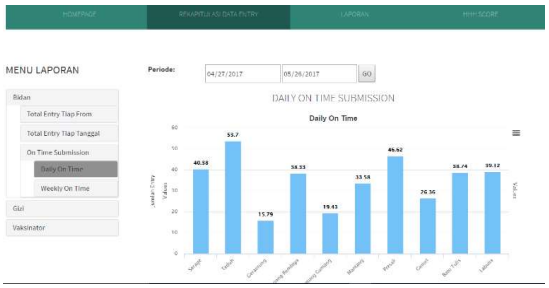


Figure 4. Reporting Dashboard

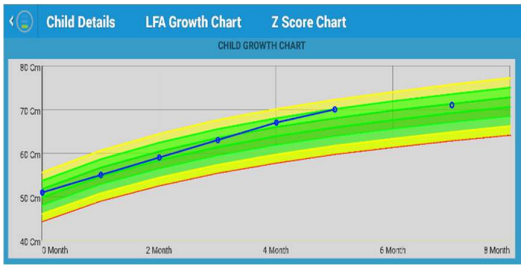


Figure 5. Z score report on child growth

All data log that were generated from the tracking device and Reporting Dashboard are collected and accumulated for later being analyzed. This analysis is used to understand the usage pattern as well as evaluate the users' performance.

Analysis from the daily tracking data was also presented to the users through coaching activity. The purpose is for the education workers to also be able to monitor their own performance. Other than above analysis, during coaching, it was also provided the users with analysis of other performance indicators driven from data entry, such as on-time submission (Figure 6.) and data quality (Figure 7). Through this data-driven coaching, we also encouraged and supported the education workers for a high compliance of the application usage.

e-ISSN: 2614-8404
p-ISSN: 2776-3234

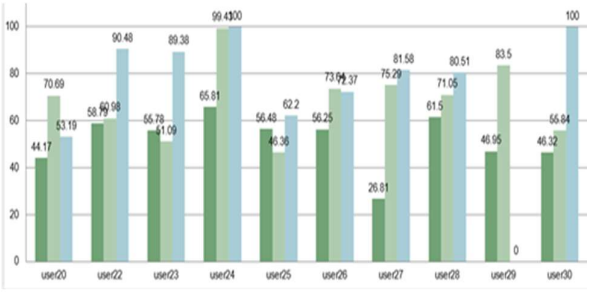


Figure 6. Ontime submission data on child development

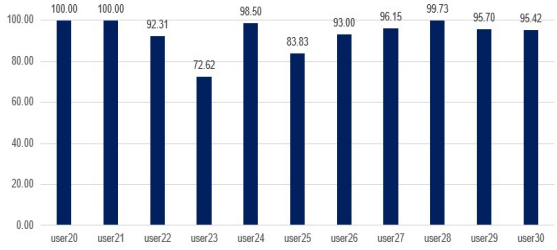


Figure 7. Data quality and completeness of child development

Due to still poor infrastructure at the grass root level, the system also allows a safety stored offline that could automatically link to server when network connection is available. The immediate data entry was provided real-time data report that could be accessed by any relevant stakeholders at any levels to response accordingly. However, to avoid misuse of data, the access has also been restricted with a secured login system [6].

Based on this study, these forms were launched simply within the smart register screens at the tap of a button, and allow offline data entry where network connection is not always available. Data has been safely stored offline until the device has a network connection again and the data is then submitted to the secured server [7]. There was been a backup server provided to keep the data updated if the main server gets into trouble. With this application, users can easily jump between questions, answering them in whichever order best matches their workflow. This application allows projects to include data entry validations and mandatory questions in their forms [8]. In addition, the application offers advanced features such as data entry calculations and cascade selects, which are useful in forms where the user must select their location from a long, expandable list. Smart registers make these once time-consuming tasks easy to accomplish. Smart register has a customizable array of sort and filter options to rearrange and filter down the list of clients to a new list that matches the user's immediate work needs. Each smart register is equipped with a smart search feature, obviating the need to scroll and scroll through the lists when trying to search for a single respondents. The search results were instant, meaning the results start appearing as soon as we start typing. The search feature was also customizable to whatever search term is needed, whether a name or an ID number [9]. The application allows data entry directly in the interface. Data was collected on the app with smart paper forms, which are built to resemble paper, but supports advanced skip/form logic including constraint checks. To reduce typing errors, the packages used a built-in data check algorithm to check the consistency and validity of each entry. If there is error or inconsistency found then it will be fixed directly [10]. After all the data was entered into the server, then they have to be edited and cleaned before being analyzed.

Data was entered automatically synced with the reporting module, so teacher or child workers can access their reports at any time. They can easily track their progress without having to manually compile data each time. Supervisors and reporting authorities can rest assured that the data being reported is accurate and reflects real service provision and health events on the ground[11]. In rural areas, or anywhere health/community development workers might be spread out and hard to reach, having an online web portal and dashboard for daily monitoring is an efficient and smart way to ensure workers are regularly providing timely care. The smart registry web portal allows end user login for monitoring their own data and printing paper reports of their data if required for submission [16]. Supervisors at higher levels can login to monitor their child workers and view their service provision in real time along with aggregate data across all workers. The web portal can also store archived data, in case a child worker needs to review older records which are no longer stored on the app. Currently, it has comprises of a server backend and Android based mobile phone client [12].

The servers has kept in a high dedicated connectivity location, an undergraduate of computer was provided by university is responsible to maintain and daily backup the data in those servers [13]. The operator who has been selected have a capability and skills to monitor, manage and maintain the server or server management. That person has to monitor all the other data and then coach the couple and what to do. All the primary data source inflow and outflow should be from the bottom of the page and the other user access should be on the top. Data Utilization for all stakeholders or others can be accessed through one gate (website based) Through a secure login, users were able to access the display data for analysis and reporting [14]. Data in a database or in a statistical package has been restricted to those who have a password for access. In any reports or publications the confidentiality of all were retained. Data collected during project was a real-time data processing and directly transfer into the server. Only limited personnel have an access to the data concerned [15]. Data has been accessed by certain personnel under the study for purpose of data analyses and reporting process. checking the data that has been collected for validity and internal consistency by automated data processing scripts customized to the needs of the project data. The scripts is flag in real time inconsistencies and alert a supervisor of potential problems requiring correction. checking the validity and internal consistency check for all data that goes into the server database on daily basis.

IV. CONCLUSION

Based on the study, this application is easily applicable for real-time monitoring and evaluation on early childhood education. Results suggest the use of real time rapid analysis of these routine assessments of provider

e-ISSN: 2614-8404
p-ISSN: 2776-3234

performance and the application usability enables a dynamic process of continuous quality improvement. It has also been shown to increase frontline education workers performance and responsiveness to the uptake of application and identified key implementation challenges early on. Developing both rapid analysis processes and evaluation techniques that utilize the real time data made available by application is crucial for continuous quality improvement and sustainability of online education approaches

REFERENCES

[1] Abubakar A, Holding P, Van Baar A, Newon CRJC and Van de Vijver FJR (2008). Monitoring psychomotor development in a resource limited setting: and evaluation of the Kilifi Developmental Inventory. *Annals of Tropical Pediatrics: International Child Health*, 28 (3), 217-226.

[2] Abubakar A, Holding P, van de Vijver F J, Bomu G, Van Baar A. (2010) Developmental monitoring using caregiver reports in a resource-limited settting: the case of Kilifi, Kenya. *Acta Pediatr* 99:29.

[3] Achenbach, T M and Edelbrock CS (2001). *Child behavior checklist*. Burlington, VT

[4] Bayley, N (2006). *Bayleys Scales of Infant and Toddler Development 3rd Edition (Bayleys III).* The Psychological Corporation, San Antonio, TX

[5] Carter AS, Briggs-Gowan MJ, Jones SM and Little TD (2003). The infant-toddler social and emotional assessment (ITSEA): Factor structure, reliability, and validity. *Journal of abnormal child psychology*. 31(5). 495-514.

[6] Gartstein MA and Rothbart MK (2003). Studying infant temperament via the revised infant behavior questionnaire. *Infant Behavior and Development*, 26(1): 64-68.

[7] Jaramillo, A. and A. Mingat. (2003). *Early Childhood Care and Education in Sub-Saharan Africa: What would it take to meet the Dakar goal*? The World Bank: Africa Region.

[8] Kochanska G, Murray KT and Harlan, ET (2000). Effortful control in early childhood: Continuity and change, antecedents, and implications for social development. *Developmental psychology*, 36(2), 220-232.

[9] Newnham CA, Milgrom J, Skouteris H. Effectiveness of a modified Mother-Infant Transaction Program on outcomes for preterm infants from 3 to 24 months of age. *Infant Behav Dev*. 2009 Jan;32(1):17-26.

[10] Peacock S, Konrad S, Watson E, Nickel D, Muhajarine N (2013) Effectiveness of home visiting programs on child outcomes: a systematic review, *BMC Public Health* 13:1728

[11] Prado EL, Alcock KJ, Muadz H, Ullman MT, Shankar AH, for the SUMMIT Study Group. (2012) Maternal multiple micronutrient supplements and child cognition: a randomized trial in Indonesia. *Pediatrics* 130:e536–e546.

[12] Prado EL, Ullman MT, Muadz H, Alcock KJ, Shankar AH, for the SUMMIT Study Group. (2012). The effect of maternal multiple micronutrient supplementation on cognition and mood during pregnancy and postpartum in Indonesia: a randomized trial. *PLoS One* 7:e32519.

[13] Shonkoff, J.P., & Phillips, D. (Eds.) (2000). *From neurons to neighborhoods: The science of early childhood development*. Committee on Integrating the Science of Early Childhood Development. Washington, DC: National Academy Press.

[14] Stipek, D.(2004). *The early childhood classroom observation measure*. Unpublished manuscript, Stanford University.

[15] Smith LB and Thelen E (2003). *Development as a dynamic system*. Trends in cognitive sciences, 7(8), 343-348.

[16] Squires J and Bricker D (2009). *Ages and Stages Questionnaires*, Third Edition (ASQ-3). Baltimore, MD: Brookes Publishing

[17] Walker SP, Wachs TD, Grantham-McGregor S, Black MM, Nelson CA, Huffman SL,Baker-Henningham H, Chang SM, Hamadani JD, Lozoff B , Meeks-Gardner JM, Powell CA, Rahman A, Richter L (2011)Inequality in early childhood: risk and protective factors for early child development. *Lancet*. 378: 1325–38.

[18] Shankar AH, Jahari AB, Sebayang SK, Aditiawarman, Apriatni M, Harefa B, Muadz H, Soesbandoro SD, Tjiong R, Fachry A, Shankar AV, Atmarita, Prihatini S, Sofia G. (2008) Effect of maternal multiple micronutrient supplementation on fetal loss and infant death in Indonesia:

e-ISSN: 2614-8404
p-ISSN: 2776-3234

a double-blind cluster-randomised trial. *Lancet*. 371:215-27.

[19] Shankar AV, Zaitu A, Kadha JK, Sebayang SK, Apriatni M, Sulastri A, Sunarsih E, Shankar AH. (2009) Programmatic effects of a large scale multiple micronutrient supplementation trial in Indonesia: using community facilitators as intermediaries for behavior change. *Food Nutr Bull*. 30:S207-S214.

[20] Zurovac D, Otieno G, Kigen S, Mbithi AM, Muturi A, Snow RW, et al. Ownership and use of mobile phones among health workers, caregivers of sick children and adult patients in Kenya: cross-sectional national survey. *Global. Health*. 2013;9(20). 2.

[21] Labrique A. Where there is no "mHealth": *Mobile Phone Ownership and Use in Rural Bangladesh*. mHealth Summit 2012 [Internet]. Washington DC Area; 2012. Available from: http://www.mhealthsummit.org/sites/default/files/Research - Maternal and Child Health.pdf

e-ISSN: 2614-8404
p-ISSN: 2776-3234

PREDICTION OF INCOMING ORDERS USING THE LONG SHORT-TERM MEMORY METHOD AT PT. XYZ

Lukman Irawan[1], Fauzi[2], Denny Andwiyan[3]
[1] Master Program in Computer Science, Faculty of Information and Technology, Budi Luhur University
[2] Master Program in Computer Science, Faculty of Information and Technology, Budi Luhur University
[3]Science and Technology Faculty, Information System Department, Raharja University
email: [1]lukman.irawan26@gmail.com, [2]fauzi.said25@gmail.com, [3]andwiyan@raharja.info

***Abstract* – Currently the need for domestic packaging paper continues to increase, driven by the level of consumer awareness about sustainable packaging. PT XYZ is a local company engaged in the Corrugated Cardboard Box (KKG) industry. So far, the problems in fulfilling incoming orders every month are not optimal with an average of about 30% inaccuracy. This is because the orders that enter cannot be predicted. As an effort to win market competition in packaging paper, PT. XYZ must improve the fulfillment of incoming orders by predicting incoming orders using the Long Short-Term Memory (LSTM) method. The aim of this research is to provide a predictive model for incoming orders in accordance with the needs of order fulfillment to be applied to production planning. So that order fulfillment can be on time. The method used in predicting incoming orders is the Long Short-Term Memory (LSTM) method using weighting evaluations with the lowest Root Mean Squared Error (RMSE) and Augmented Dickey-Fuller test (ADF). The test results of the LSTM method with parameter sizes of Batch: 1 Epochs: 5000 Neurons: 1 show that the RMSE for MDM products is 8.767582 and 0.287924, LNR products are 10.623984 and 0.466621, WTP products are 1.636849 and 0.361515 lower than the size of the fit parameters for other LSTM models, and the ADF Statistic value for MDM products -6.137597, LNR -6.753697, WTP -4.872927.**

***Keywords* – Prediction, Planning, LSTM, *Time Series*.**

I. INTRODUCTION

Paper Packaging was the first flexible packaging before the invention of plastic and aluminum foil. Currently, paper packaging is still widely used and is able to compete with other packaging such as plastic and metal because it is cheap, easy to obtain and widely used. The weakness of packaging paper for packaging food ingredients is that it is sensitive to water and is easily affected by environmental humidity.

Currently, the demand for domestic packaging paper continues to increase driven by the increasing level of consumer awareness, about sustainable packaging, together with strict regulations imposed by various environmental protection agencies, regarding the use of environmentally friendly packaging products, which is driving an increasing market for packaging based paper.

PT XYZ has been established since 1993 and is a local company engaged in the corrugated cardboard box (KKG) packaging industry in Indonesia, with an area of + 2.0 hectares in the western area of Purwakarta, West Java. and has a production capacity of +500 tons per month.

PT. XYZ produces using the Make to Order (MTO) system, where several production activities such as final assembly and component manufacturing wait until there is an incoming order from the customer. However, some activities, such as providing production capacity, are carried out on the basis of forecasting or predicting incoming orders, where prediction of incoming orders is an activity to estimate the size of incoming orders for certain goods in a certain period and marketing area. So predictive figures can be made for a monthly period. In the hierarchy of predictions, different models can be made.

So far, the problem of fulfilling incoming orders every month is considered not optimal, as described in Table 1.1 which explains the status of the accuracy of the fulfillment of incoming orders which is divided into 2 (two) namely "Right" and "Incorrect", then percentage in weight (Tons) :

Table 1. Percentage of Accuracy of Distribution of Packaging Paper Requests

Status Ketepatan Pemenuhan	Jumlah Permintaan	Jumlah Weight (Ton)	Presentase Jlm Weight (Ton)
TEPAT	1,221	25,215	70%
TIDAK TEPAT	566	10,842	30%
Grand Total	**1,787**	**36,057**	**100%**

From Table 1 above, it can be seen that the percentage of accuracy in fulfilling incoming orders during 2019 was only around 70% and the inaccuracy in fulfilling paper requests was around 30%. Then in Table 2 explains the average percentage of fulfillment accuracy based on the type of customer (Large Customers, Medium Customers, and Small Customers) :

Table 2. Average Percentage of Demand Distribution Accuracy

Formula		A	B	C = A + B	D = A / C	E = B / C
No	Type Pelanggan	Status (Weight - Ton)		Total Ketepatan (Weight - Ton)	Presentase Status	
		Tepat	Tidak Tepat		Tepat	Tidak Tepat
1	Pelanggan Besar	22,265	2,820	25,084	89%	11%
2	Pelanggan Sedang	2,066	6,657	8,723	24%	76%
3	Pelanggan Kecil	884	1,365	2,250	39%	61%
Grand Total		**25,215**	**10,842**	**36,057**	**70%**	**30%**

e-ISSN: 2614-8404
p-ISSN: 2776-3234

In Table 2. above, it can be seen that the average percentage of the accuracy of fulfillment of incoming orders for large customers with an average accuracy of fulfillment of about 89% and inaccuracy of fulfillment of about 11%, medium customers with an average fulfillment of about 24% and inaccuracy of fulfillment of about 76% while for small customers the average accuracy of fulfillment is around 39% and the inaccuracy of fulfillment is around 61%.

This is because the ups and downs of incoming orders cannot be predicted properly. So that the production team has difficulty in planning the management of maximum production capacity to fulfill incoming orders. Figure 1 shows the trend of packaging paper demand during 2019 :

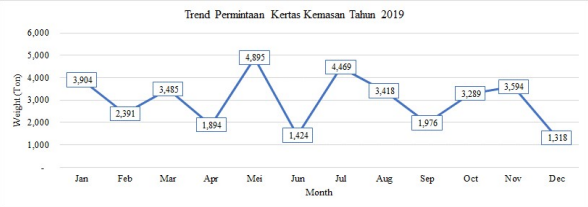


Figure 1. Incoming Order Trend in 2019 Tahun.

As a result of the problems caused above can give a bad predicate for the level of service provided by PT. XYZ on customers and also has the potential to reduce the company's profits.

Time series prediction methods such as Support Vector Machine (SVM) [2] [3], Recurrent Neural Network (RNN) [4] and LSTM [1] are proposed by many researchers to predict order.

RNN [4] has advantages in predicting sequential data, strong in each processing, RNN will store the internal state, namely St, which is given from one time step to the next time step. This is the *"Memory"* of the RNN, but has the disadvantage that the number of layers in the RNN itself can be very long, as long as the number of rows in the input, and this poses a problem in itself because the RNN often has to store dependencies from inputs that are located quite far apart which are commonly called *"Vanishing Gradient"*.

LSTM [1] can solve the RNN problem, namely Vanishing Gradient, because it has a different processing with ordinary RNN modules. Another difference is that additional signals are given from one time step to the next, namely the cell state and memory cell, represented by the symbol Ct. which is appropriate for the characteristics of the demand for packaging paper in this study.

In this study, LSTM will be applied to select the appropriate and optimal cell state and memory cells, so that the results of the prediction model for packaging paper demand are more accurate according to the company's needs. The expected result is that the demand prediction model for packaging paper using the Long Short-Term Memory (LSTM) method can help improve the management of packaging paper demand according to company needs and can also help improve PT. XYZ to customers with timely fulfillment and according to incoming orders.

Based on the background of the problem in this study, there is no good prediction of incoming orders. With the aim of being able to apply an incoming order prediction model with the Long Short-Term Memory (LSTM) Method.

II. RESEARCH METHOD

A. Prediction Theory

To solve problems in the future that cannot be ascertained, people always try to solve them with models of approaches that are in accordance with the actual behavior of the data, as well as in making predictions [13].

Predicting (forecasting) demand for products and services in the future and its parts is very important in planning and monitoring production [14]. A prediction has many meanings, so the prediction needs to be planned and scheduled so that it will take a period of time at least in the period of time needed to make a policy and determine several things that affect the policy.

Predictions are needed in addition to estimating what will happen in the future, decision makers also need to make plans.

B. Definition of Prediction

Prediction is an estimate of the expected level of demand for a product or several products in a certain period of time in the future. Therefore, Prediction is basically an estimate, but using certain methods Prediction can be more than just one estimate. It can be said that Prediction is a scientific estimate although there will be some errors due to the limitations of human abilities.

Before describing this Prediction method, it is first described about the definition of Prediction itself. Prediction is the activity of estimating the expected level of product demand for a product or several products in a certain period of time in the future [5].

According to Buffa: "Prediction or forecasting is defined as the use of statistical techniques in the form of a future picture based on the processing of historical figures" [6].

According to Makridakis: "Prediction is an integral part of management decision-making activities" [7].

Organizations always set goals and objectives, try to estimate environmental factors, and then choose actions that are expected to result in the achievement of these goals and objectives. The need for prediction increases in line with management's efforts to reduce its dependence on things that are uncertain. Prediction becomes more scientific in nature in the face of management environment. Because every organization is related to each other, good or bad forecasts can affect all parts of the organization [7].

C. Prediction Technique

Prediction techniques, in general, the time series method can be grouped into:

1. Averaging Method

Used for conditions where each data at different times has the same weight so that random fluctuations in data can be soaked with the average, usually used for short-term predictions [8]. The methods included in it, among others:

- Simple Average

$$F_{T+n} = \overline{X} = \sum_{i=n}^{T+(n-1)} \frac{X_i}{T}$$

Formula description:

e-ISSN: 2614-8404
p-ISSN: 2776-3234

X = F = Prediction results
T = Period
Xi = Demand in period t

- Simple Moving Average
  If stationary data is obtained, this method is good enough to predict the situation. Formula used :

$$F_{T+n} = \overline{X} = \frac{X_1 + X_2 + \ldots\ldots + X_n}{T}$$

Formula description:
X = F = Prediction results
T = Period
Xi = Demand in period t

- Double Moving Average
  Jika data tidak stasioner serta mengandung pole trend, maka dilakukan moving average terhadap hasil single moving average. Rumus yang digunakan :

$$S'_t = \frac{X_t + X_{t-1} + \ldots + X_{t-1}}{N}$$

2. Metode *Smoothing* (Pemulusan)
   Used in conditions where the weight of the data in one period is different from the data in the previous period and forms an Exponential function which is commonly called Exponential smoothing [9]. The methods included in it, among others:

- *Single Exponential Smoothing*
  This method greatly reduces the problem of data distortion because there is no need to store historical data anymore. The effect of the size of a is in the opposite direction to the effect of entering the number of observations. This method always follows any trend in the actual data because all it can do is set future forecasts with a percentage of the last error. To determine a close to optimal requires several trials. Formula used :

$$F_{t+1} = F_t + \alpha \times (X_t - F_t)$$

Where : Ft+1 = Prediction result t + 1
a = Smoothing constant
Xt = Demand in period t
Ft = Previous period

- *Double Exponential Smoothing* one parameter of *Browns*.
  The rationale for Browns linear exponential smoothing is similar to that of a linear moving average, because both single and multiple smoothing values lag behind the actual data if there is an element of trend. The equation used in this method is as follows:

$$S'_t = aX_t + (1-a)S'_{t-1}$$

Formula description :
Xt = Demand in period t
S't = Smoothing value I period t
S"t = Smoothing value II period t
S't-1 = Previous first smoothing value (t-1)
S"t-1 = Previous second smoothing value (t-1)
a = Smoothing constant
at = Interception in period t
bt = Period trend value t
Ft+1 = Prediction Results for the period t+1
m = Number of forecasted future time periods

- *Regresi Linier*
  Linear regression is used for prediction if the existing data set is linear, meaning that the relationship between the time variable and demand is in the form of a line (linear). The linear regression method is based on the calculation of the least square error, namely by calculating the smallest distance to a point in the data to draw a line. As for the linear regression prediction equation, three constants are used, namely a, b and Y [10]. With each formulation is as follows:

$$b = \frac{n\sum X_i Y_i - \sum X_i \sum Y_i}{n\sum X_i^2 - (\sum X_i)^2}$$

Formula description :
y = Predicted Variables
a,b = Prediction Parameters
t = Independent variable

D. Research design
In this study, several steps will be taken to achieve the research objectives. These steps can be illustrated through the flow chart in Figure 2. This research starts from collecting data from raw data. The next stage is the determination of the network architecture design by determining the input and output patterns for training and testing purposes on an artificial neural network (ANN). This stage is then followed by the determination of the training algorithm.

Next is the training stage for the data that has been normalized and the architecture determined, the training is carried out first for the standard backpropagation algorithm, after that the training is carried out again by adding the learning rate and momentum coefficient to the weight update. The purpose of the training was to determine the value of the Root Mean Squared Error (RMSE) [11] and the Augmented Dickey–Fuller test (ADF) [12]. Carry out the testing phase of the test data, with the aim of knowing the level of validation of the results.

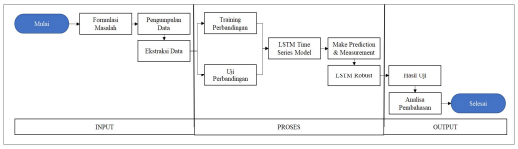


Figure 2. Research Design

The model experiment phase begins with model making and LSTM Network training. The training results are in the form of an LSTM model in the form of a CSV file. The model obtained is then used in the prediction process by loading the model file. The final stage of the experiment is the process of denormalizing the test data to get the predicted value and the evaluation value of the model's performance results. The benefit of the training and prediction processes being done separately is when running the prediction process if there is no new data. There is no need to model the training data again, but directly load it from the file. This will speed up the prediction process because without having to retrain the same data.

e-ISSN: 2614-8404
p-ISSN: 2776-3234

E. Data collection

The data used in this study is data on the realization of the demand for packaging paper at PT. XYZ from 2016 to 2019 in the form of secondary data that is quantitative. Consists of 8 attributes and 166.899 rows. The description of the realization of data on the fulfillment of packaging paper demand at PT. XYZ can be seen in table 3.

Table 3. Overview of Incoming Order Data

Order Date	Req. Ship. Date	Mills	Order Number	Cust. ID	Mat. ID	Prod. Group	SO Weight (KG)
3/18/2016	3/30/2019	NBL	2611004828	Cust1	MT001	LNR	19.80
3/18/2016	3/30/2019	NBL	2611004828	Cust3	MT002	LNR	19.20
3/18/2016	3/30/2019	NBL	2611004828	Cust1	MT003	LNR	19.60
5/15/2016	6/15/2019	NBL	2611004829	Cust4	MT002	LNR	24.00
5/15/2016	6/15/2019	NBL	2611004830	Cust2	MT003	LNR	23.80
1/13/2019	1/30/2017	NBL	2611006651	Cust39	MT096	LNR	30.11
1/6/2019	1/30/2017	NBL	2611006637	Cust19	MT053	MDM	20.47
1/6/2019	1/30/2017	NBL	2611006637	Cust19	MT021	MDM	13.10
…	…	…	…	…	…	…	…
12/3/2018	12/9/2020	NBL	2611003948	Cust94	MT259	LNR	0.11
12/4/2018	12/9/2020	NBL	2611003952	Cust76	MT424	WTP	1.00

F. Data processing

Before the implementation stage is implemented, the preprocessing stage is first carried out. The number of initial data that can be obtained from data collection is 1,335,192 records, but not all data is used and not all attributes are used because the data must go through the initial data processing stage or is called data preparation..

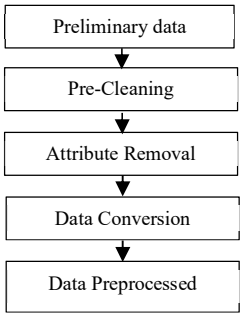


Figure 3. Data Processing

G. Proposed method

Long Short-Term Memory (LSTM) [1] was first mentioned in 1997 by Hochreiter and Schmidhuber. LSTM is also known as a neural network with an adaptable architecture, so its shape can be adjusted depending on the application. Below Figure 4 shows a model diagram for the LSTM method.

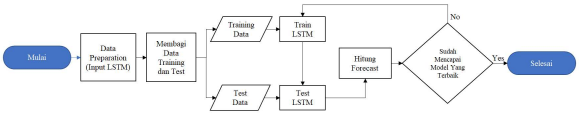


Figure 4. LSTM Method Model Diagram

Long Short-Term Memory is a derivative of the RNN (Recurrent Neural Network) method. RNN is an iterative neural network which is specially designed to handle sequential data. However, RNN has a vanishing and exploding gradient problem, namely if there is a change in the range of values from one layer to another.

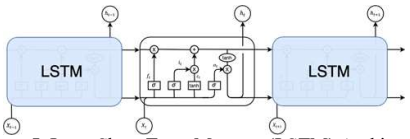


Figure 5. Long Short Term Memory (LSTM) Architecture

The hidden layer consists of memory cells, one memory cell has three gates, namely input gate, forget gate, and output gate. The input gate controls how much information should be stored in the cell state. This prevents the cell from storing unnecessary data. Forget gate functions to control the extent to which the value remains in the memory cell. Output Gate serves to decide how much content or value is in a memory cell, it is used to calculate output.

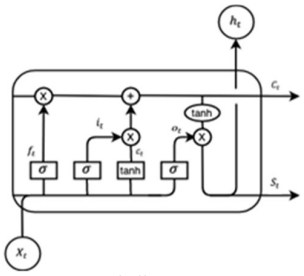


Figure 6. Shell Memory LSTM

Next on an architecture. The LSTM is built and designed to overcome the problem of gradient vanishing from RNNs when dealing with vanishing and exploding gradients. The LSTM architecture consists of an input layer, an output layer, and a hidden layer which is presented in Figure 5.

Figure 6 presents the contents of the hidden layer of LSTM namely memory cells. A memory cell in the LSTM stores a value or state (cell state), either for a long or short period of time. The explanation for the gates in one Long Short Term Memory (LSTM) memory cell is as follows:

1. Input Gate (i_t)

   The input gate takes the role of taking the previous output and the new input and passing them through the sigmoid layer. This gate returns a value of 0 or 1. The formula for i_t is:

$$i_t = \sigma(W_i S_t - 1 + W_i X_t)$$

   Formula description :
   $W i$ = Weight of Input Gate
   $S t - 1$ = Previous state at time t-1
   Xt = Input on time t
   σ = Sigmoid activation function

   The input gate value is multiplied by the output of the candidate layer ($\tilde{C}$). Formula for ($\tilde{C}$) is :

$$\tilde{C} = tanh(W_c S_t - 1 + W_c X_t)$$
$$ct = (it * \tilde{C}_t + f_t * c_t - 1)$$

   Formula description,
   $\tilde{C}$ = Intermediate cell state.
   $W c$ = Weight of cell state.
   $S t - 1$ = Previous state at time t– 1.
   $X t$ = Input on time t.

   The previous state is multiplied by the forget gate and then added to the new candidate function allowed by

e-ISSN: 2614-8404
p-ISSN: 2776-3234

the output gate.

2. Forget Gate (f_t)
Forget gate is a sigmoid layer that takes the output at time t – 1 and input at time t and combines them and applies the sigmoid activation function. Since it is sigmoid, the output of this gate is 0 or 1. If $f_t = 0$ then the previous state will be forgotten, while if $f_t = 1$ the previous state has not changed. Formula of f_t is :

$$f_t = \sigma(W_f S_t{-}1 + W_f X_t)$$

Formula description,
Wf = Weight of forget gate.
$St-1$ = Previous state at time t - 1.
Xt = Input on time t.
σ = Sigmoid activation function

This layer applies a hyperbolic tangent to the previous mix of input and output. Returns the candidate vector to be added to the state.

3. Output Gate (O_t)
The output gate controls how many states pass to the output and works in the same way as any other gate. And finally generate a new cell state (ht). Formula of ot and ht is :

$$ot = \sigma(W_o S_t{-}1 + W_o X_t)$$
$$ht = ot * tanh(ct)$$

Formula description,
Wo = Weight of output gate.
$St-1$ = Previous state or current state t - 1.
Xt = Input on time t.
σ = Sigmoid activation function

The prediction accuracy is obtained from the data that has been trained and the key to the success of both is the number of hidden layers. There are two attributes used in this study, namely the date of sale and the value or value of daily income from orders entered by PT. XYZ. The LSTM Method Training Model diagram is shown in Figure 7 below 7:

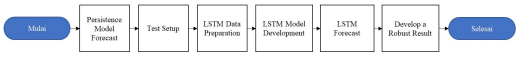


Figure 7. LSTM Method Training Model Diagram

H. Result Evaluation

For testing the performance of the model using the Root Mean Square Error (RMSE). Root Mean Square Error (RMSE) is an alternative method for evaluating forecasting techniques used to measure the accuracy of the forecast results of a model. The resulting value RMSE is the average value of the square of the number of errors in the prediction model. Root Mean Square Error (RMSE) is a technique that is easy to implement and has been frequently used in various studies related to RMSE forecasting which is expressed by the following formula :

Root Mean Square Error (RMSE)

$$\sqrt{\frac{1}{n}\sum_{i}^{n}(y_i - y_i)^2}$$

Formula description:
$\tilde{y}i$ = Forecasting value
yi = Actual value
n = Amount of data

I. Unit Root Test

Stationarity is one of the important prerequisites in time series data models. Stationary data is data that shows the mean, variance and auto variance (on the lag variation) remains the same at any time the data is formed or used, meaning that with stationary data the time series model can be said to be more stable. If the data used in the model is not stationary, then the data is reconsidered for its validity and stability.

One of the formal concepts used to determine the stationarity of data is through a unit root test. This test is a popular test, developed by David Dickey and Wayne Fuller (1979) as the Augmented Dickey-Fuller (ADF) Test [15]. If a time series data is not stationary at zero order, I(0), then the data stationarity can be searched through the next order so that the stationarity level is obtained on the nth order (first difference or I(1), or second difference or I(2).), etc. Several models can be selected to perform the ADF Test:

$\Delta Yt = \delta Yt\text{-}1 + ut$ (without intercept)
$\Delta Yt = \beta + \delta Yt\text{-}1 + ut$ (with intercept)
$\Delta Yt = \beta 1 + \beta 2t + \delta Yt\text{-}1 + ut$ (intercept with time trend)
Δ = first difference of the variables used
t = variabel trend

The hypothesis for this test is :
H0 : $\delta = 0$ (there is a unit root, not stationary)
H1 : $\delta \neq 0$ (no unit root, stationary)

III. RESULT AND DISCUSSION

A. Data preparation

The data collected consists of 8 columns and 166.899 rows. The data is historical data on demand transactions with a time span from January 2016 to December 2019. The data is in the form of an excel file with 8 variables and 1,335,192 records, hereinafter referred to as paper-sales.

The variables in the transaction data include: Order Date, Req. Ships. Date, Factory, Order Number, Customer ID, Sales Person, Material ID, and SO Weight (MT).

Table 4. Display of Incoming Order Data

Order Date	Req. Ship. Date	Mills	Order Number	Cust. ID	Mat. ID	Prod. Group	SO Weight (KG)
3/18/2016	3/30/2019	NBL	2611004828	Cust1	MT001	LNR	19.80
3/18/2016	3/30/2019	NBL	2611004828	Cust3	MT002	LNR	19.20
3/18/2016	3/30/2019	NBL	2611004828	Cust1	MT003	LNR	19.60
5/15/2016	6/15/2019	NBL	2611004829	Cust4	MT002	LNR	24.00
5/15/2016	6/15/2019	NBL	2611004830	Cust2	MT003	LNR	23.80
1/13/2019	1/30/2017	NBL	2611006651	Cust39	MT096	LNR	30.11
1/6/2019	1/30/2017	NBL	2611006637	Cust19	MT053	MDM	20.47
1/6/2019	1/30/2017	NBL	2611006637	Cust19	MT021	MDM	13.10
...	...	...	...	...	...	...	...
12/3/2018	12/9/2020	NBL	2611003948	Cust94	MT259	LNR	0.11
12/4/2018	12/9/2020	NBL	2611003952	Cust76	MT424	WTP	1.00

Display This data set describes the number of monthly incoming orders for a period of 4 years, before testing the data shown in the figure:

e-ISSN: 2614-8404
p-ISSN: 2776-3234

```
Month
2016-01-01    36.8
2016-02-01    27.6
2016-03-01    29.3
2016-04-01    36.0
2016-05-01    37.4
Name: Sales, dtype: float64
```

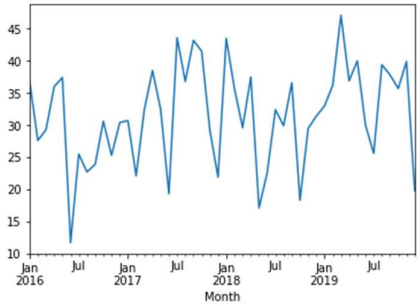


Figure 8. Display of Incoming Order Data Collection for 4 Years

B. Dataset Test Setting

The paper-sales dataset will be divided into 2 (two) parts, namely a training set and a test set. The first 2 (two) years of data will be taken for the training data set and the other 2 (two) years of data will be used for the test set.

The process of dividing the dataset or what is called the Train-Test Split is as follows :

```
Observations: 48
Training Observations: 36
Testing Observations: 12
```

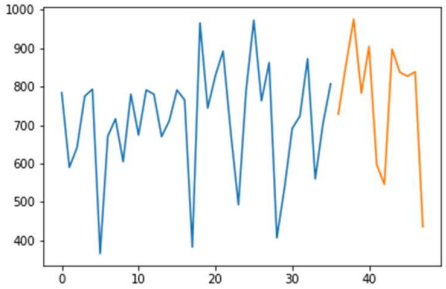


Figure 9. Display of Test Data Test

Furthermore, the model will be developed using the training dataset and will make predictions on the test dataset.

A rolling forecast scenario will be used, also called Walk-Fordward model validation, where each step of the test dataset will be executed individually, then the actual expected values from the test dataset will be fetched and made available to the forecast model at the next time step. This mimics a field scenario where research on sales of new packaging paper will be made available each month and used in the forecast for the following month.

From all estimates on the dataset will be collected error scores are calculated to summarize the skills of the model. The root mean squared error (RMSE) will be used because it can produce a score that is in the same unit as the estimated data, namely monthly sales of packaged paper.

C. Estimated Persistence Model

A good basic approximation for a time series with a linear upward trend is the persistence estimate. Persistence estimates where observations from the previous time step (t-1) are used to predict the observations in the current time step (t). We can implement this by taking the last observations from the training and history data accumulated with walk-forward validation and using them to predict the current time step.

The following is the persistence estimation model on the paper-sales-l dataset as follows::

```
RMSE: 9.054
```

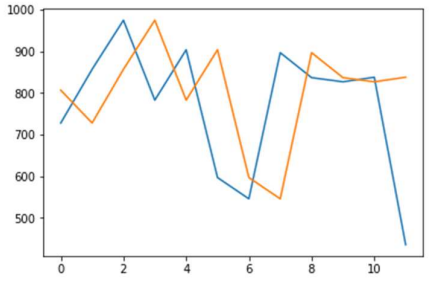


Figure 10. Display of the Persistence Model

In the persistence estimate model above, an RMSE value of around 198.7 monthly packaging paper sales is generated for the estimate on the test data set set.

D. Turning Datasets into Supervised Learning

The LSTM model in Keras assumes that the data will be divided into input (X) and output (y) components. For Time Series, it can be achieved by using the last time step observation (t-1) as input and the current time step observation (t) as output.

Then to combine these two series together to create a Data Frame that can be used in Supervised Learning. The series that is pushed down will have a new position at the top with no value. The NaN value (no numbers) will be used in this position which replaces this NaN value with a value of 0, which the LSTM model must learn as the start of the circuit. Here are the results:

```
      0     0
0   0.0  36.8
1  36.8  27.6
2  27.6  29.3
3  29.3  36.0
4  36.0  37.4
```

E. Converting Dataset to stationary data

The Packaging Paper Sales dataset is classified as non-stationary where there is a structure in the data that depends on time. Specifically, there is an increasing trend in the data. Stationary data are easier to model and are more likely to produce more skilled estimates perkiraan.

The standard way to clear a trend is to differentiate the data. That is, the previous research time step (t-1) is subtracted from the current study (t). This removes the trend and we are left with a series of differences, or changes in the study from one time step to the next. Here are the results:

```
Month
```


e-ISSN: 2614-8404
p-ISSN: 2776-3234

```
2016-01-01    36.8
2016-02-01    27.6
2016-03-01    29.3
2016-04-01    36.0
2016-05-01    37.4
Name: Sales, dtype: float64
0    -9.2
1     1.7
2     6.7
3     1.4
4   -25.7
dtype: float64
0    27.6
1    29.3
2    36.0
3    37.4
4    11.7
dtype: float64
```

From the results above, the first 5 rows of loaded data are printed, then the first 5 rows are from different series, then finally the first 5 rows with the difference operation being reversed. For the first study in the original dataset was removed from the data the inverse difference. Moreover, the last data set matches the first as expected.

F. Splitting the Dataset to Scale

Like other Neural Networks (NN), LSTM expects data to be within the scale of the activation function used by the network. The default activation function for the LSTM is the hyperbolic tangent (tanh), which returns a value between -1 and 1. This is the range of interest for time series data.

For a balanced experiment, the values of the scaling coefficients (min and max) must be calculated on the training dataset and applied to scale any test and forecast datasets. This avoids contaminating the experiment with knowledge from the test data set, which may give the model a small advantage.

Next convert the dataset to the range [-1, 1] using the MinMaxScaler class. Like other scikit-learn transform classes, it requires data to be provided in a matrix format with rows and columns. Therefore, we must reshape the NumPy array before performing the transformation.

```
Month
2016-01-01    36.8
2016-02-01    27.6
2016-03-01    29.3
2016-04-01    36.0
2016-05-01    37.4
Name: Sales, dtype: float64
0    0.418079
1   -0.101695
2   -0.005650
3    0.372881
4    0.451977
dtype: float64
0    36.8
1    27.6
2    29.3
3    36.0
4    37.4
dtype: float64
```

G. Estimate using the Long-Short Term Memory (LSTM) Model

Once the LSTM model matches the training data, the model can be used to make estimates. In this study it has some flexibility which can decide to adjust the model once on all training data, then predict each new time step one by one from the test data (fixed approach), or can adjust the model or update the model each time step from the test data as new observations from test data available (dynamic approach).

To make an estimate, we can call the predict() function on the model. This takes the NumPy 3D array input as an argument. In that case, it would be a layer with one value, research on the previous time step.

To run this model, you can call a function called forecast(). With model fit, the batch size used when adjusting the model (for example 1), and a row of test data, the function will separate the input data from the test row, reshape it, and return the prediction as a single floating point value.

During training, the internal state is reset after each epoch. When estimating in this study did not reset the internal state among estimates. In fact, the model builds state as we estimate each time step in the test data set set.

The scaling and reverse-scaling behavior has been moved to the scale() and invert_scale() functions for simplicity. The test data is scaled using a fit of scaler on the training data, as needed to ensure the min/max values of the test data do not affect the model. The sequence of data transformations is adjusted for convenience, first creating stationary data, then supervised learning problems, then scaling. Distinctions are made on the entire data set before being broken down into training and test sets for convenience. In this case, it is easy to collect observations during validation going forward and differentiate them as we proceed where it is not decided not to read them further. Below is the result :

```
Month=1, Predicted=25.133321, Expected=33.000000
Month=2, Predicted=26.136585, Expected=36.200000
Month=3, Predicted=28.210680, Expected=47.100000
Month=4, Predicted=38.214326, Expected=36.900000
Month=5, Predicted=32.778191, Expected=40.000000
Month=6, Predicted=34.051595, Expected=30.000000
Month=7, Predicted=33.395497, Expected=25.600000
Month=8, Predicted=33.054117, Expected=39.400000
Month=9, Predicted=34.184841, Expected=37.800000
Month=10, Predicted=33.495201, Expected=35.700000
Month=11, Predicted=33.384978, Expected=39.900000
Month=12, Predicted=34.004823, Expected=19.800000
Test RMSE: 8.914
```

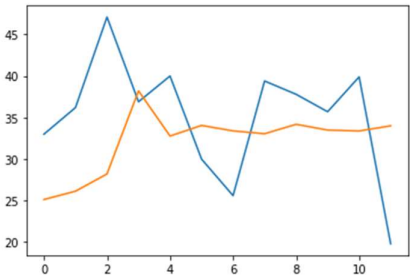


Figure 11. LSTM Model Display (1, 500, 1)

e-ISSN: 2614-8404
p-ISSN: 2776-3234

```
Month=1, Predicted=31.046228, Expected=33.000000
Month=2, Predicted=32.255115, Expected=36.200000
Month=3, Predicted=33.345093, Expected=47.100000
Month=4, Predicted=39.694085, Expected=36.900000
Month=5, Predicted=38.614510, Expected=40.000000
Month=6, Predicted=38.535921, Expected=30.000000
Month=7, Predicted=39.055871, Expected=25.600000
Month=8, Predicted=36.120883, Expected=39.400000
Month=9, Predicted=35.680632, Expected=37.800000
Month=10, Predicted=39.032655, Expected=35.700000
Month=11, Predicted=39.866720, Expected=39.900000
Month=12, Predicted=40.316062, Expected=19.800000
Test RMSE: 8.754
```

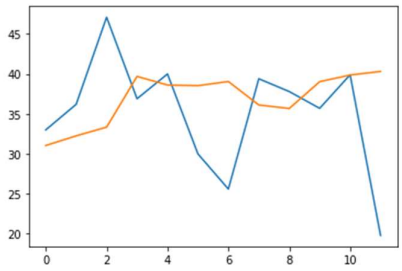


Figure 12. LSTM Model Display (1, 1500, 1)

```
Month=1, Predicted=26.131302, Expected=33.000000
Month=2, Predicted=27.513795, Expected=36.200000
Month=3, Predicted=30.229067, Expected=47.100000
Month=4, Predicted=40.482170, Expected=36.900000
Month=5, Predicted=38.248608, Expected=40.000000
Month=6, Predicted=35.598107, Expected=30.000000
Month=7, Predicted=33.753025, Expected=25.600000
Month=8, Predicted=30.770659, Expected=39.400000
Month=9, Predicted=33.095257, Expected=37.800000
Month=10, Predicted=33.409154, Expected=35.700000
Month=11, Predicted=32.137640, Expected=39.900000
Month=12, Predicted=33.928015, Expected=19.800000
Test RMSE: 8.576
```

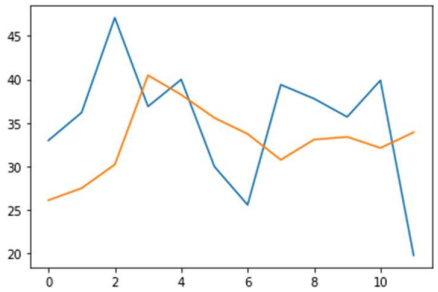


Figure 13. LSTM Model Display (1, 5000, 1)

H. Developing Robust Results

The problem with Neural Network (NN) is that this model gives different results to the initial conditions. One approach might be to improve the Random Number Seed used by Keras to ensure the results are reproducible. Another approach would be to control for random initial conditions using different experimental settings.

Next, the walk-forward model and validation will be placed in a loop with a fixed number of repetitions. Each run iteration of the RMSE can be recorded. We can then summarize the distribution of the RMSE scores. Below is the result:

```
1) Test RMSE: 8.735
2) Test RMSE: 9.242
3) Test RMSE: 8.620
4) Test RMSE: 8.716
5) Test RMSE: 8.544
6) Test RMSE: 8.818
7) Test RMSE: 8.509
8) Test RMSE: 8.628
9) Test RMSE: 8.590
10) Test RMSE: 8.544
...
29) Test RMSE: 8.659
30) Test RMSE: 10.027

rmse
count  30.000000
mean    8.767582
std     0.287924
min     8.508875
25%     8.621584
50%     8.711215
75%     8.782992
max    10.026662
```

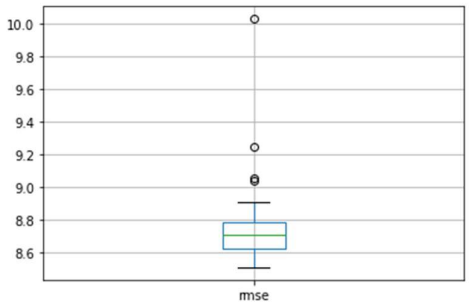


Figure 14. Robust Result Display (1, 5000, 1)

From the results above, it can be seen that the average value and standard deviation of the RMSE are 8.767582 and 0.287924 monthly incoming orders. A box and whisker plot is created from the distribution shown below. It captures the middle of the data as well as the extent and outliers. Then for measurement.

I. Augmented Dickey-Fuller test (ADF)

Statistical tests make strong assumptions about the dataset used. ADF to inform the extent to which the null hypothesis can be rejected or failed to be rejected. The results must be interpreted in order for certain problems to be meaningful.

The null hypothesis of this test is that the time series can be represented by a unit root, which is not stationary (has some time-dependent structure). The alternative hypothesis (rejecting the null hypothesis) is that the time series does not move.

- Hypothesis Zero (H0): If it fails to be rejected, this indicates that the time series has a unit root, meaning it is not stationary. It has some time-dependent structure.
- Alternative Hypothesis (H1): The null hypothesis is rejected; it shows the time series has no unit root, which means it doesn't move. It has no time-dependent structure.

The result uses the p-value of the test. A p-value below the threshold (such as 5% or 1%) indicates we rejected the null hypothesis (stationary), otherwise, a p-value above the threshold indicates we failed to reject the null hypothesis (non-stationary)..

e-ISSN: 2614-8404
p-ISSN: 2776-3234

- p-value> 0.05: Failed to reject the null hypothesis (H0), the data has a unit root and is non-stationary.
- p-value <= 0.05: Reject the null hypothesis (H0), the data has no unit root and is stationary.

Below is an example of calculating the Augmented Dickey-Fuller test on the Incoming Orders dataset for each product :

Table 5. ADF Recap for MDM, LNR and WTP Products

Measurement	MDM	LNR	WTP
ADF Statistic	-6.137597	-6.753697	-4.872927
P-value	0.000000	0.000000	0.000000
1%	-3.578	-3.578	-3.578
5%	-2.925	-2.925	-2.925
10%	-2.601	-2.601	-2.601

J. Comparison of LSTM with DES and Linear Regression

Before conducting research using the LSTM method, a comparison has been made with the Double Exponential Smoothing (DES) method [9] and Linear Regression [10] using the Minitab software. This is done to take the lowest error, of the three methods. Below table 6 shows the comparison:

Table 6. Comparison of LSTM, DES, Linear Regression Errors

Measurement	LSTM	DES	Regresi Linier
MSE	8.576	10.689	9.3279

From table 6 it can be seen that the lowest error using the Long Short-Term Memory (LSTM) method is 8.576.

K. *Fit* comparison of LSTM Models

In this study, the network parameters will not be adjusted, instead will use a comparison of the 3 configurations below, the RMSE and ADF Test will be measured:

Table 7. Comparison results of LSTM.

Measurement *fit* Model LSTM	***Batch:* 1** ***Epochs:* 500** ***Neuron:* 1**	***Batch:* 1** ***Epochs:* 1500** ***Neuron:* 1**	***Batch:* 1** ***Epochs:* 5000** ***Neuron:* 1**
RMSE	8.914	8.754	8.576
ADF Test	ADF Statistic: -6.137597 p-value: 0.000000 1%: -3.578 5%: -2.925 10%: -2.601		

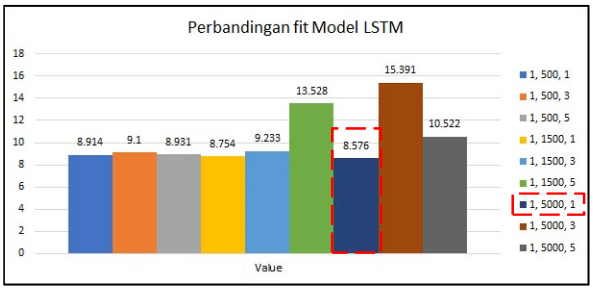


Figure 15. Comparison graph of LSTM . Model fit

From the results of the comparison of the fit of the LSTM model above, it can be seen that the parameter size of Batch: 1 Epochs: 5000 Neuron: 1 shows RMSE 8,576 lower than the size of the other LSTM model fit parameters, and the ADF Statistic value is -6.137597 The more negative this statistic, the more likely we are to rejecting the null hypothesis means that the dataset is stationary and the p-value 0.000000 means that the data does not have a unit root and is stationary.

IV. CONCLUSION

Based on the discussion that has been carried out, the conclusions are :

1. From the analysis of the trials that have been carried out, it can be concluded that the LSTM model is able to produce time series data predictions in this case, namely incoming orders with a small and accurate error rate. The average value and standard deviation of RMSE for MDM products are 8.767582 and 0.287924, RMSE for LNR products are 10.623984 and 0.466621, RMSE for WTP products are 1.636849 and 0.361515 monthly incoming orders.
2. The test results show that the number of 1 LSTM batch, epoch = 1500, and LSTM unit = 1 is the most optimal model parameter. The composition of the data sharing affects the prediction accuracy results. The best composition is obtained at 80% train data and 20% test data.
3. Predictions for the next 12 months are likely to remain stable. The future prediction data generated is between the range of initial orders for MDM products from 25.986923 to the last data, which is 34.050922, for LNR products from 39.865912 to the last data, which is 27.63460, for WTP products from 3.758679 until the last data is 2.678200.

REFERENCES

[1] Hochreiter, S., & Schmidhuber, J. (1997). LSTM 1997. *Neural Computation*.

[2] Ou, Y., Qian, H., & Xu, Y. (2009). Support vector machine based approach for abstracting human control strategy in controlling dynamically stable robots. Journal of Intelligent and Robotic Systems: Theory and Applications. https://doi.org/10.1007/s10846-008-9292-8

[3] Qian, M., Hongquan, W., Yongsheng, W., & Yan, Z. (2008). SVM based prediction of Spontaneous Combustion in Coal Seam. Proceedings of the 2008 International Symposium on Computational Intelligence and Design, ISCID 2008. https://doi.org/10.1109/iscid.2008.193

[4] Tseng, Y. C., Chen, C. C., Lee, C., & Huang, Y. K. (2007). Incremental in-network RNN search in wireless sensor networks. Proceedings of the International Conference on Parallel Processing Workshops. https://doi.org/10.1109/ICPPW.2007.47

[5] Biegel, J. E. (1961). Statistics in Forecasting. Management International, 162-181.

e-ISSN: 2614-8404
p-ISSN: 2776-3234

[6] Buffa, Elwood S., dan Sarin, R. K. (1996). Manajemen Operasi dan Produksi Modern. Edisi 8. Jakarta: Binarupa Aksara.
[7] Wheelwright, S., Makridakis, S., Makridakis, S., Wheelwright, S., Gross, C. W., Peterson, R. T., … O'Neill, W. J. (1978). Forecasting Methods for Management. Journal of Marketing Research. https://doi.org/10.2307/3150640
[8] Mitropolsky, I. A. (1967). Averaging method in non-linear mechanics. International Journal of Non-Linear Mechanics, 2(1), 69-96.
[9] Kitagawa, G. (1991). A nonlinear smoothing method for time series analysis. Statistica Sinica, 371-388.
[10] Andrews, D. F. (1974). A robust method for multiple linear regression. Technometrics, 16(4), 523-531.
[11] Willmott, C. J. (1981). On the validation of models. Physical geography, 2(2), 184-194.
[12] Cheung, Y. W., & Lai, K. S. (1995). Lag order and critical values of the augmented Dickey–Fuller test. Journal of Business & Economic Statistics, 13(3), 277-280.
[13] Rescher, N. (1998). Predicting the future: An introduction to the theory of forecasting. SUNY press.
[14] Syntetos, A. A., Babai, Z., Boylan, J. E., Kolassa, S., & Nikolopoulos, K. (2016). Supply chain forecasting: Theory, practice, their gap and the future. European Journal of Operational Research, 252(1), 1-26.
[15] Dickey, David A dan Wayne A. Fuller, (1979), Distribusi of Estimators for Autoregressive Time Series With a Unit Root, Journal of the American Statistical Association, Vol. 74, No. 366

e-ISSN: 2614-8404
p-ISSN: 2776-3234

Prediction of Electrical Energy Consumption Using LSTM Algorithm with Teacher Forcing Technique

Sasmitoh Rahmad Riady[1*)], Tjong Wan Sen[2]
[1,2]Information Technology, Faculty of Computing, President University
Email: [1]sasmitohrr@student.president.ac.id, [2]wansen@president.ac.id

Abstract **– Electrical energy is an important foundation in world economic growth, therefore it requires an accurate prediction in predicting energy consumption in the future. The methods that are often used in previous research are the Time Series and Machine Learning methods, but recently there has been a new method that can predict energy consumption using the Deep Learning Method which can process data quickly for training and testing. In this research, the researcher proposes a model and algorithm which contained in Deep Learning, that is Multivariate Time Series Model with LSTM Algorithm and using Teacher Forcing Technique for predicting electrical energy consumption in the future. Because Multivariate Time Series Model and LSTM Algorithm can receive input with various conditions or seasons of electrical energy consumption. Teacher Forcing Technique is able lighten up the computation so that it can training and testing data quickly. The method used in this study is to compare Teacher Forcing LSTM with Non-Teacher Forcing LSTM in Multivariate Time Series model using several activation functions that produce significant differences. TF value of RMSE 0.006, MAE 0.070 and Non-TF has RMSE and MAE values of 0.117 and 0.246. The value of the two models is obtained from Sigmoid Activation and the worst value of the two models is in the Softmax activation function, with TF values is RMSE 0.423, MAE 0.485 and Non-TF RMSE 0.520, MAE 0.519.**

Keywords – Multivariate Time Series, Deep Learning, Teacher Forcing, LSTM.

I. INTRODUCTION

Energy is an important foundation for economic development in a country [1] and electricity is one of the main energy sources [2]. Therefore, energy policy for a country is very important, because it not only helps the development of the country but also affects the environment both in the field of industrial operations and in the realm of low-income to elite residential areas. Due to the large amount of capital investment and the length of time it took to expand the capacity of electrical energy projects. Therefore, a good estimate or prediction is one of the requirements in the development of a more effective energy policy. because it can reduce the possibility of errors in electrical system planning. Therefore, producing an accurate forecast of electricity consumption is very important [3].

In recent years, many studies have used techniques in predicting electrical energy consumption, either using machine learning [4] or deep learning [5]. In a study conducted by Karimbatar et al regarding data mining for energy consumption by comparing several algorithms in machine learning with Regression models [6], Neural Network [7], and SVM [8] and taking the best model obtained by the Regression model with a relative error of 0.9%, the output of this prediction shows that the average electricity consumption rate increases by about 3.2% per year and will reach 7,076,796 MW in 2020 from a population growth of 22.28% [9]. In another study for the realm of machine learning conducted by Choi regarding the energy consumption analysis for homes using the K-means clustering algorithm from the data generated for K-7 with a silhouette score of 0.799 [10] Another study was also conducted by Nallathambin et al to predict consumption. The electrical energy that will be applied to the USA using the Decision Tree and Random Forest algorithms and experimental solutions states that the RF model provides better accuracy, namely 95.78% than the DT model with an accuracy of 91.6% and each model has an error rate 0.197 and 0.906 [11].

But at this time there is a new method, namely the Deep Learning method, which can process training quickly, at this time there are also many researchers conducting research on predictions using the method as reflected by Kim et al regarding the prediction of household electricity consumption using CNN-LSTM Hybrid Network [12], the proposed method can be quickly and accurately in predicting irregular energy consumption trends in the dataset of household power consumption. However, because the proposed method was processed earlier by the sliding window algorithm [13], this caused a prediction delay in the actual data [14], in other studies carried out by Young-Jun in electrical energy forecasting By comparing the models contained in the Deep Learning [15] including the LSTM, Gru, and SEQ2SEQ models with the results of the LSTM experiment get the best results with RMSE 0.96 [16], but this value is not good enough to use the actual data to use seasonal data features and Long term in forecasting more accurate electrical energy. Therefore, the researcher will propose a multivariate time series model [17] using the LSTM algorithm as a model and electrical energy prediction algorithm and Teacher Forcing Technique [18] to help in long-term predictions using public consumption datasets taken from the Smart Meters in London Some conditions or seasons. Because the Multivariate Time Series LSTM model algorithm can combine several input to training and testing and produce an output, therefore from various conditions or seasons for the consumption of electrical energy in the dataset will be used as input and will produce an output, namely predictions in the future Come accurately and precisely [19] But the algorithm has weaknesses in the long-term prediction because of the high computing side, then the teacher forcing will help in the long-term predictions because the algorithm can train repetitive networks quickly and efficiently due to output from Repeated LSTM will be used as a subsequent input so

e-ISSN: 2614-8404
p-ISSN: 2776-3234

that it will produce low computing using the basic truth from the previous time step as input [20]. With a dataset that the multivariate time series model and the algorithm can reduce RMSE [21] and can predict the consumption of electricity in the long term.

From the results of the training data, a comparison will be made with several activation functions such as ReLu, Softmax [22], Sigmoid [23] and the newest activation function found in RNN, namely Swish [24].

II. RESEARCH METHODOLOGY

At this research stage, there are steps taken by researchers including using literature studies, data analysis using data visualization techniques [25], then using preprocessing data [26] where the data that has been collected and processed is returned to be entered into the model architecture then the result data will be trained to get the results of a prediction of electrical energy consumption, as shown in Figure 1 below.

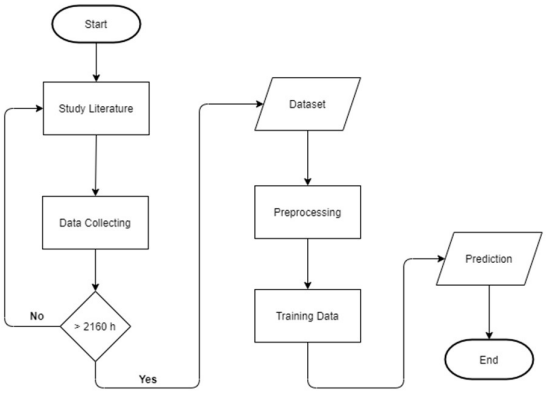


Figure 1. Flow Research Stages

A. Data Collecting

At this stage are the steps in data collection to be used as a dataset of several seasons that are correlated with data on electrical energy consumption for each housing block. The following is the flow of the data collection stages as in Figure 2 below.

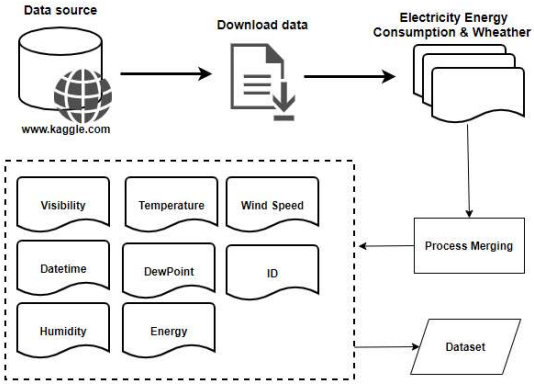


Figure 2. Data Collecting

As in Figure 2, where electricity consumption data is taken from public data, namely www.kaggle.com about smart meter data in London, data is downloaded by 2 GB and analyzed, the data contains several folders and files. DataSet represents the reading of household electrical energy consumption in London in Kilowatt Hour (KWH) from November 2011 to February 2014. The file series is categorized into 112 blocks (block 0 to block 111) in the original dataset. In addition to electrical energy consumption data there are several files, including Acorn Details, Information Household, UK Holiday Bank, Weather Daily and Weather per hour. For this study we combine electrical energy consumption data with weather per hour for one housing block. The dataset will be trained is 2903 with variables datetime, visibility, temperature, dew point, pressure, windspeed, humidity, ID, and energy.

B. Training Model

The model that was built using the LSTM Multivariate Time Series Model with Teacher Forcing Technique, before going to the process, this study will propose to do some comparisons including comparing the LSTM algorithm with Non-Teacher Forcing and Teacher Forcing [27], to find which performance loss is better, then the results of the comparison will be continued using the Multivariate Time Series model and comparisons of several activation functions [28] contained in the Deep Learning method, the following is the flow of the training model that will be used in this study, as shown in Figure 3.

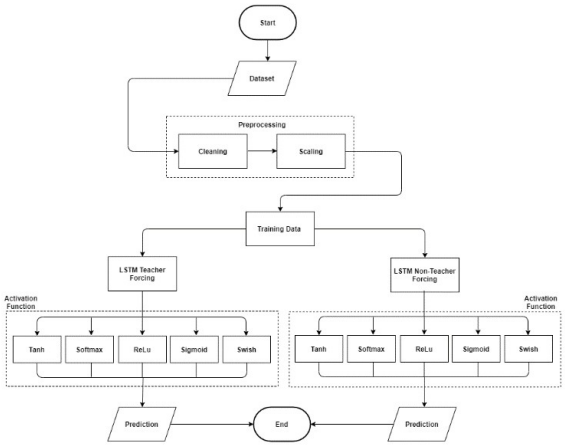
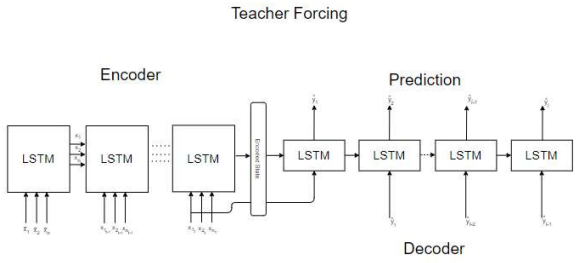


Figure 3. Training Model

C. Architecture Model

The following is an architectural model using a comparison between LSTM Non-Teacher Forcing and LSTM Teacher Forcing in order to find the best performance loss of the two models, then the model will be entered into the Multivariate Time Series model, then it will be compared with several activations to find RMSE results. better. As in Figure 4 and Figure 5.



e-ISSN: 2614-8404
p-ISSN: 2776-3234

Figure 4. LSTM Teacher Forcing

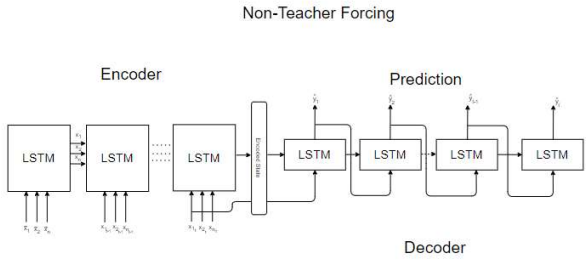


Figure 5. LSTM Non-Teacher Forcing

D. Activation Function

From the results of the eating model architecture, several activations will be compared, including sigmoid activation, ReLu, Sigmoid and Custom Activation. Development of the Sigmoid Activation, this activation functions to make the Neural Network non-linear [29]. Sigmoid will accept a single number and convert the x value into a value that has a range from 0 to 1, which has the following formula.

$$S(x) = \frac{1}{1 + e^{-x}} \quad (1)$$

And next is the ReLU or Rectified Linear Unit activation which is a pretty good activation function because ReLU greatly accelerates the convergence process carried out with stochastic gradient descent when compared to sigmoid/tanh with the following formula.

$$f(x) = \max(0, x) \quad (2)$$

Because ReLU basically only creates a delimiter on the number zero, meaning that if x ≤ 0 then x = 0 and if x > 0 then x = x. Their experiment shows that Swish [30] tends to perform better than ReLu on deeper models across a number of challenging data sets with the following formula.

$$f(x) = x.sigmoid(x) \quad (3)$$

E. Measurement Step

experiment carried out which performance is better. Measurement using RMSE and MAE. Root Mean Square Error (RMSE) is the sum of the squared error or the difference between the true value and the predetermined predictive value. With the RMSE.

$$RMSE = \sqrt{\sum \frac{(Y' - Y)^2}{n}} \quad (4)$$

Mean Absolute Error (MAE) shows the mean error value which is the error of the true value with the predicted value. MAE itself is generally used for measuring error prediction in time series analysis. The formula for MAE itself is defined as follows:

$$MAE = \frac{|Y' - Y|}{n} \quad (5)$$

hat formula (5) shows that Y' is Prediction Value, Y is Actual Value, and n is Total of Data.

III. RESULTS AND DISCUSSION

A. Preprocessing

At this stage the data will be merged to form multivariate time series data so that data from several files such as weather and energy can be combined into a data file. The following is an example of a dataset that has been merged.

Table 1. Data Merging

visibility	temperature	datetime	dewpoint	pressure	Windspeed	humidity	id	energy
14.31	6.86	2013-10-29T23:00:00Z	4.61	1017.98	3.29	0.86	MAC000002	0.457
14.31	6.1	2013-10-30T00:00:00Z	4.09	1018.38	2.95	0.87	MAC000002	0.414
13.45	5.94	2013-10-30T01:00:00Z	4.17	1018.89	3.14	0.88	MAC000002	0.408
13.23	5.54	2013-10-30T02:00:00Z	3.92	1019.25	3.02	0.89	MAC000002	0.352
14.31	5.06	2013-10-30T03:00:00Z	3.28	1019.57	2.74	0.88	MAC000002	0.25
14.31	5.04	2013-10-30T04:00:00Z	3.24	1019.84	2.77	0.88	MAC000002	0.217
14.31	4.81	2013-10-30T05:00:00Z	3.29	1020.25	2.16	0.9	MAC000002	0.211
…	…	…	…	…	…	…	…	…
12.23	8.69	2014-02-27T17:00:00Z	1.55	1006.43	4.11	0.61	MAC000002	0.667
12.41	7.22	2014-02-27T18:00:00Z	2.16	1006.68	3.43	0.7	MAC000002	1.7
12.52	6.22	2014-02-27T19:00:00Z	2.07	1006.88	2.99	0.75	MAC000002	2.259
14.03	5.94	2014-02-27T20:00:00Z	2.07	1006.74	3.25	0.76	MAC000002	1
16.09	5.03	2014-02-27T21:00:00Z	1.67	1006.36	3.06	0.79	MAC000002	1.766
14	4.1	2014-02-27T22:00:00Z	1.64	1005.67	3.02	0.84	MAC000002	2.465

e-ISSN: 2614-8404
p-ISSN: 2776-3234

The following is a description of the attributes resulting from the merging process, namely Time, Visibility, Temperature, Dew Point, Pressure, Wind Speed, Humidity, ID and Energy. Produces 2904 raw data or 2904 hours of data. Then proceed with the indexing, cleaning, and scaling data stages. Table 2 is an example of the final prepocesing stage.

B. Teacher Forcing and Non-Teacher Forcing with Univariate Time Series Model

Figure 4 and Figure 5 have discussed the model architecture for teacher forcing and non-teacher forcing, the data that will be used first is the univariate time series data where the attributes to be learned are the data frame index and energy which have been changed from per hour to per minute, at this stage the data will be learned using an ADAM optimization of 15 epochs and a batch size of 100, here are the results of the model that has been trained.

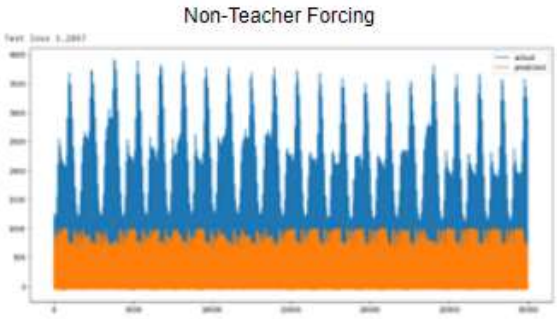


Figure 6. Teacher Forcing Univariate Model

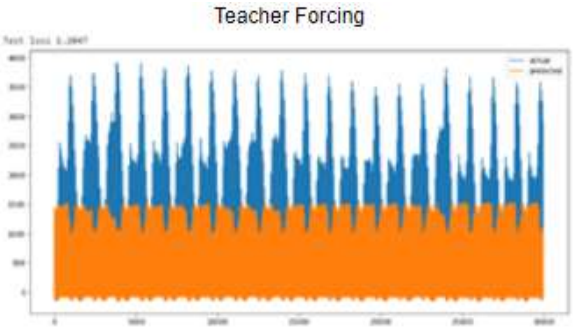


Figure 7. Non-Teacher Forcing Univariate Model

From the results of training and testing, the LSTM model using the Teacher Forcing technique is better for performance loss and prediction on the univariate model, next is training on electrical energy consumption data with the Multivariate Time Series model using Teacher Forcing and Non-Teacher Forcing Techniques

A. Teacher Forcing and Non-Teacher Forcing with Multivariate Time Series Model

At this stage is to conduct training on each attribute that will be used as input to be trained on the model that has been made, the following is a multivariate time series variable model that will be trained on the LSTM model using TF and Non-TF.

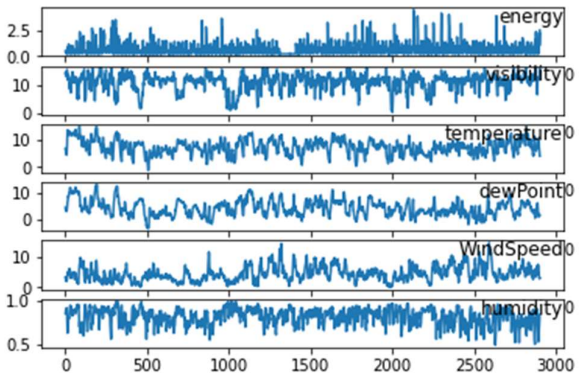


Figure 8. Multivariate Time Series Electricity Consumption

In the previous training, univariate time series data were trained without using an activation function, now multivariate data will be trained using Tanh, Softmax, ReLu, Sigmoid, and Swish activations. The image below is the result of the training.

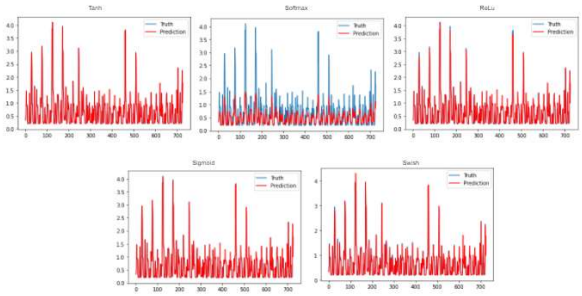


Figure 9. Prediction Result LSTM Teacher Forcing using Activation.

And next is to measure the prediction results for the Non-Teacher Forcing LSTM model as follows.

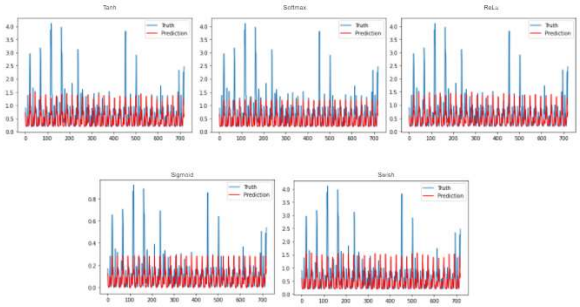


Figure 10. Prediction Result LSTM Non-Teacher Forcing Activation

Table 2. Activation Deferent

Activation	Teacher Forcing		Non-Teacher Forcing	
	RMSE	MAE	RMSE	MAE
Tanh	0.009	0.084	0.510	0.513
Softmax	0.423	0.485	0.520	0.519
ReLu	0.020	0.104	0.508	0.514
Sigmoid	0.006	0.070	0.117	0.246
Swish	0.025	0.117	0.505	0.515

e-ISSN: 2614-8404
p-ISSN: 2776-3234

IV. CONCLUSION

The training model in this study uses several activation functions including tanh, Softmax, ReLu, Sigmoid, and swish activation. The results obtained show that Teacher Forcing LSTM is better than Non-Teacher Forcing LSTM in terms of performance loss and prediction which results in quite a significant difference. The TF value is RMSE 0.006, MAE 0.070 and Non-TF itself has RMSE and MAE values of 0.117 and 0.246. The value of the two models is obtained from sigmoid activation and the worst value of the two models is in the Softmax activation function, with TF values namely RMSE 0.423, MAE 0.485 and Non-TF RMSE 0.520, MAE 0.519.

REFERENCES

[1] T. Ula, "Dampak Konsumsi Energi Terbarukan Terhadap Pertumbuhan Ekonomi: Studi di Asia Tenggara," *J. Econ. Sci.*, vol. 5, no. 2, pp. 26–34, 2019.

[2] dkk Ahmad Wahid, "Analisis Kapasitas Dan Kebutuhan Daya Listrik Untuk Menghemat Penggunaan Energi Listrik Di Fakultas Teknik Universitas Tanjungpura," *J. Tek. Elektro UNTAN*, vol. 2, no. 1, p. 10, 2014.

[3] O. Somantri, "Prediksi Kebutuhan Permintaan Energi Listrik Menggunakan Neural Network Berbasiskan Algoritma Genetika," no. September, pp. 1–135, 2015.

[4] S. Seyedzadeh, F. P. Rahimian, I. Glesk, and M. Roper, "Machine learning for estimation of building energy consumption and performance: a review," *Vis. Eng.*, vol. 6, no. 1, 2018, doi: 10.1186/s40327-018-0064-7.

[5] M. Alanbar, A. Alfarraj, and M. Alghieth, "Energy consumption prediction using deep learning technique case study of computer college," *Int. J. Interact. Mob. Technol.*, vol. 14, no. 10, pp. 166–177, 2020, doi: 10.3991/ijim.v14i10.14383.

[6] B. Shyti and D. Valera, "The Regression Model for the Statistical Analysis of Albanian Economy," *Int. J. Math. Trends Technol.*, vol. 62, no. 2, pp. 90–96, 2018, doi: 10.14445/22315373/ijmtt-v62p513.

[7] B. Cetişli and A. Barkana, "Speeding up the scaled conjugate gradient algorithm and its application in neuro-fuzzy classifier training," *Soft Comput.*, vol. 14, no. 4, pp. 365–378, 2010, doi: 10.1007/s00500-009-0410-8.

[8] A. Abubakar *et al.*, "A support vector machine classification of computational capabilities of 3D map on mobile device for navigation aid," *Int. J. Interact. Mob. Technol.*, vol. 10, no. 3, pp. 4–10, 2016, doi: 10.3991/ijim.v10i3.5056.

[9] N. Karimtabar, S. Pasban, and S. Alipour, "Analysis and predicting electricity energy consumption using data mining techniques - A case study I.R. Iran - Mazandaran province," *2015 2nd Int. Conf. Pattern Recognit. Image Anal. IPRIA 2015*, no. Ipria, pp. 0–5, 2015, doi: 10.1109/PRIA.2015.7161634.

[10] Y. J. Lee and H. J. Choi, "Forecasting building electricity power consumption using deep learning approach," *Proc. - 2020 IEEE Int. Conf. Big Data Smart Comput. BigComp 2020*, pp. 542–544, 2020, doi: 10.1109/BigComp48618.2020.000-8.

[11] S. Nallathambi and K. Ramasamy, "Prediction of electricity consumption based on DT and RF: An application on USA country power consumption," *Proc. - 2017 IEEE Int. Conf. Electr. Instrum. Commun. Eng. ICEICE 2017*, vol. 2017-Decem, no. 1, pp. 1–7, 2017, doi: 10.1109/ICEICE.2017.8191939.

[12] S. Chan, I. Oktavianti, and V. Puspita, "A Deep Learning CNN and AI-Tuned SVM for Electricity Consumption Forecasting: Multivariate Time Series Data," *2019 IEEE 10th Annu. Inf. Technol. Electron. Mob. Commun. Conf. IEMCON 2019*, pp. 488–494, 2019, doi: 10.1109/IEMCON.2019.8936260.

[13] H. Lim, Y. Kim, C. Y. Lee, and K. Cheun, "An Efficient Sliding Window Algorithm Using Adaptive-Length Guard Window for Turbo Decoders," *J. Commun. Networks*, vol. 14, no. 2, pp. 195–198, 2012, doi: 10.1109/JCN.2012.6253068.

[14] T. Y. Kim and S. B. Cho, *Predicting the Household Power Consumption Using CNN-LSTM Hybrid Networks*, vol. 11314 LNCS. Springer International Publishing, 2018.

[15] A. Sherstinsky, "Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) network," *Phys. D Nonlinear Phenom.*, vol. 404, no. March, pp. 1–43, 2020, doi: 10.1016/j.physd.2019.132306.

[16] S. Goodman and N. Ding, "TeaForN : Teacher-Forcing with N-grams," pp. 8704–8717, 2020.

[17] F. Karim, S. Majumdar, H. Darabi, and S. Harford, "Multivariate LSTM-FCNs for time series classification," *Neural Networks*, vol. 116, pp. 237–245, 2019, doi: 10.1016/j.neunet.2019.04.014.

[18] K. Drossos, S. Gharib, P. Magron, and T. Virtanen, "Language modelling for sound event detection with teacher forcing and scheduled sampling," *arXiv*, no. October, 2019, doi: 10.33682/1dze-8739.

[19] A. A. Ismail, T. Wood, and H. C. Bravo, "Improving long-horizon forecasts with expectation-biased LSTM networks," *arXiv*, 2018.

[20] A. Goyal, A. Lamb, Y. Zhang, S. Zhang, A. Courville, and Y. Bengio, "Professor forcing: A new algorithm for training recurrent networks," *Adv. Neural Inf. Process. Syst.*, no. Nips 2016, pp. 4608–4616, 2016.

[21] A. S. Wardana and M. I. Ananta Timur, "Collaborative Filtering Recommender System pada Virtual 3D Kelas Cendekia," *IJEIS (Indonesian J. Electron. Instrum. Syst.*, vol. 8, no. 1, p. 73, 2018, doi: 10.22146/ijeis.28729.

[22] C. E. Nwankpa, W. Ijomah, A. Gachagan, and S. Marshall, "Activation functions: Comparison of trends in practice and research for deep learning," *arXiv*, pp. 1–20, 2018.

[23] Julpan, E. B. Nababan, and M. Zarlis, "Analisis

e-ISSN: 2614-8404
p-ISSN: 2776-3234

Fungsi Aktivasi Sigmoid Biner Dan Sigmoid Bipolar Dalam Algoritma Backpropagation Pada Prediksi Kemampuan Siswa," *J. Teknovasi*, vol. 02, no. 1, pp. 103–116, 2015.

[24] M. Çelebi and M. Ceylan, "The New Activation Function for Complex Valued Neural Networks: Complex Swish Function," no. July 2019, pp. 169–173, 2019, doi: 10.36287/setsci.4.6.050.

[25] N. Ahmad Syaripul and A. Mukharil Bachtiar, "Visualisasi Data Interaktif Data Terbuka Pemerintah Provinsi Dki Jakarta: Topik Ekonomi Dan Keuangan Daerah," *J. Sist. Inf.*, vol. 12, pp. 15–29, 2016.

[26] C. V. Gonzalez Zelaya, "Towards explaining the effects of data preprocessing on machine learning," *Proc. - Int. Conf. Data Eng.*, vol. 2019-April, pp. 2086–2090, 2019, doi: 10.1109/ICDE.2019.00245.

[27] M. Sangiorgio and F. Dercole, "Robustness of LSTM neural networks for multi-step forecasting of chaotic time series," *Chaos, Solitons and Fractals*, vol. 139, p. 110045, 2020, doi: 10.1016/j.chaos.2020.110045.

[28] A. Yadav, C. K. Jha, and A. Sharan, "Optimizing LSTM for time series prediction in Indian stock market," *Procedia Comput. Sci.*, vol. 167, no. 2019, pp. 2091–2100, 2020, doi: 10.1016/j.procs.2020.03.257.

[29] O. Contribution, "Nonlinear Neural Networks: Principles, Mechanisms, and Architectures," *Pattern Recognit. by Self-Organizing Neural Networks*, vol. 1, 2020, doi: 10.7551/mitpress/5271.003.0004.

[30] H. H. Chieng, N. Wahid, O. Pauline, and S. R. Kishore Perla, "Flatten-T Swish: A thresholded ReLU-Swish-like activation function for deep learning," *arXiv*, vol. 4, no. 2, pp. 76–86, 2018.