

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



# JISA

(JURNAL INFORMATIKA  
dan SAINS)



**VOL 5 No 1**  
**June**

Published by :

**Program Studi Teknik Informatika**  
**UNIVERSITAS TRILOGI**

**2022**

# **JISA**

## **(Jurnal Informatika dan Sains)**

Volume 5, Edition 1, June 2022

---

### **Implementation of the AHP-SMARTER Method in the Decision Support System for Giving Sanctions for Violation of Student Disciplines**

*Sofiansyah Fadli, Maulana Ashari, Hasyim Asyari, Ahmad Susan Pardiansyah*

### **Implementation of Data Mining on Tourist Visits Patterns on Lombok Island Tourism Objects**

*Saikin, SofiansyahFadli, Maulana Ashari*

### **Website System Design Using Agile Kanban Based On QR Code**

*Alton Gunawan Purwanto, Ricky Yohanes Wijaya, Timotius, Indra Budi Trisno*

### **Grouping of Village Status in West Java Province Using the Manhattan, Euclidean and Chebyshev Methods on the K-Mean Algorithm**

*Gatot Tri Pranoto, Wahyu Hadikristanto, Yoga Religia*

### **Application of Data Mining to Determine Promotion Strategy Using Algorithm Clustering at SMK Yadika 1**

*Jerry Watulangkouw*

### **Online Monitoring and Analysis of Lube Oil Degradation for Gas Turbine Engine using Recurrent Neural Network (RNN)**

*Febrianto Nugroho, Rusdianto Roestam*

### **Development of a Village Information System for Acceleration of Village Services in Desa Tegal Kecamatan Kemang Bogor**

*Deden Ardiansyah, Prihastuti Harsani, Eneng Tita Tosida, Abimanyu Oki Saputera, Andhika Bhayangkari*

### **Design and Development E-Mading System for Information Students**

*Geovanne Farell, Igor Novid, Sandi Rahmadika*

### **Naive Bayes and Support Vector Machine Algorithm for Sentiment Analysis Opensea Mobile Application Users in Indonesia**

*Laurenzius Julio Anreaja, Norma Nobuala Harefa, Julius Galih Prima Negara, Venantius Nathan Hermanu Pribyantara, Agung Budi Prasetyo.*

### **Music Genre Recommendations Based on Spectrogram Analysis Using Convolutional Neural Network Algorithm with RESNET-50 and VGG-16 Architecture**

*I Nyoman Purnama*

### **The Classification of Anxiety, Depression, and Stress on Facebook Users Using the Support Vector Machine**

*Tsania Maulidia Wijiasih, Rona Nisa Sofia Amriza, Dedy Agung Prabowo*

**Decision Support System Scholarship Selection Using Simple Additive Weighting (SAW) Method**

*Budi Arifitama*

**Enterprise Content Management (ECM) System Architecture for Capital Project at Oil and Gas Company**

*Arief Herdiansah*

---

Published by:  
**Program Studi Teknik Informatika**  
**Universitas Trilogi**

<b>JISA</b>	<b>Vol : 5</b>	<b>Ed :1</b>	<b>Page : 001-089</b>	<b>Jakarta, Jun 2022</b>	<b>e-ISSN: 2614-8404</b>	<b>p-ISSN: 2776-3234</b>
-------------	----------------	--------------	---------------------------	------------------------------	------------------------------	------------------------------

# JISA

## (Jurnal Informatika dan Sains)

Volume 5, Edition 1, December 2022

---

### Advisor

Yodfiatfinda.,P.hD

### Editor in Chief

Budi Arifitama, S.T., MMSI

### Editorial Board

Ade Syahputra, S.T., M.Inf.Comm.Tech.Mgmt.

: Universitas Trilogi

Yaddarabullah, S.Kom., M.Kom.

: Universitas Trilogi

Maya Cendana, S.T., M.Cs.

: Universitas Bunda Mulia

Silvester Dian Handy Permana, S.T., M.T.I.

: Universitas Trilogi

Ketut Bayu Yogha. B, S.Kom., M.Cs

: Universitas Trilogi

Ninuk Wiliani.,S.Si.,M.Kom

: Institut Teknologi dan Bisnis BRI

Dwi Pebrianti,Ph.D

: Universiti Malaysia Pahang, Malaysia

Dr.Wahyu Caesarendra

: Universiti Brunei Darussalam

### Reviewers

Prof. Ir. Suyoto, M.Sc. Ph.D

: Universitas Atma Jaya Yogyakarta

Dr. Ir. Albertus Joko Santoso, M.T.

: Universitas Atma Jaya Yogyakarta

Setiawan Assegaff, ST, MMSI, Ph.D

: STIKOM Dinamika Bangsa, Jambi

Michael Marchenko, Ph.D

: Universitas Trilogi, Jakarta

Dwi Pebrianti,Ph.D

: Universiti Malaysia Pahang, Malaysia

Prof.Dr.Hoga Saragih.,ST.,MT

: Universitas Bakrie

Isham Shah Hassan.,Ph.D

: Port Dickson Polytechnic Malaysia

Prof.Dr Abdul Talib Bon

: Universiti Tun Hussein Onn, Malaysia

Wiwin Armoldo Oktaviani, S.T, M.Sc

: Universitas Muhammadiyah Palembang,

Yosi Apriani, S.T, M.T

: Universitas Muhammadiyah Palembang,

Dr. Gandung Triyono.,M.Kom

: Universitas Budi Luhur

Ir. Lukito Edi Nugroho, M.Sc., Ph.D

: Universitas Gadjah Mada

Dr. Soetam Rizky Wicaksono

: Universitas Ma chung,

### Secretariat

Asih Wulandini

### 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 : 5	Ed :1	Page : 001-089	Jakarta, Jun 2022	e-ISSN: 2614-8404	p-ISSN: 2776-3234
------	---------	-------	-------------------	----------------------	----------------------	----------------------

## Table of Content

<b>Implementation of the AHP-SMARTER Method in the Decision Support System for Giving Sanctions for Violation of Student Disciplines</b> <i>Sofiansyah Fadli, Maulana Ashari, Hasyim Asyari, Ahmad Susan Pardiansyah</i>	1-11
<b>Implementation of Data Mining on Tourist Visits Patterns on Lombok Island Tourism Objects</b> <i>Saikin, SofiansyahFadli, Maulana Ashari</i>	12-18
<b>Website System Design Using Agile Kanban Based On QR Code</b> <i>Alton Gunawan Purwanto, Ricky Yohanes Wijaya, Timotius, Indra Budi Trisno</i>	19-27
<b>Grouping of Village Status in West Java Province Using the Manhattan, Euclidean and Chebyshev Methods on the K-Mean Algorithm</b> <i>Gatot Tri Pranoto, Wahyu Hadikristanto , Yoga Religia</i>	28-34
<b>Application of Data Mining to Determine Promotion Strategy Using Algorithm Clustering at SMK Yadika 1</b> <i>Jerry Watulangkouw</i>	35-49
<b>Online Monitoring and Analysis of Lube Oil Degradation for Gas Turbine Engine using Recurrent Neural Network (RNN)</b> <i>Febrianto Nugroho, Rusdianto Roestam</i>	50-53
<b>Development of a Village Information System for Acceleration of Village Services in Desa Tegal Kecamatan Kemang Bogor</b> <i>Deden Ardiansyah, Prihastuti Harsani, Eneng Tita Tosida. Abimanyu Oki Saputera, Andhika Bhayangkari</i>	54-57
<b>Design and Development E-Mading System for Information Students</b> <i>Geovanne Farell, Igor Novid , Sandi Rahmadika</i>	58-61
<b>Naive Bayes and Support Vector Machine Algorithm for Sentiment Analysis Opensea Mobile Application Users in Indonesia</b> <i>Laurenzius Julio Anreaja, Norma Nobuala Harefa, Julius Galih Prima Negara, Venantius Nathan Hermanu Pribyantara, Agung Budi Prasetyo.</i>	62-68
<b>Music Genre Recommendations Based on Spectrogram Analysis Using Convolutional Neural Network Algorithm with RESNET-50 and VGG-16 Architecture</b> <i>I Nyoman Purnama</i>	69-74
<b>The Classification of Anxiety, Depression, and Stress on Facebook Users Using the Support Vector Machine</b> <i>Tsania Maulidia Wijiasih, Rona Nisa Sofia Amriza, Dedy Agung Prabowo</i>	75-79
<b>Decision Support System Scholarship Selection Using Simple Additive Weighting (SAW) Method</b> <i>Budi Arifitama</i>	80-84

<b>JISA</b>	<b>Vol : 5</b>	<b>Ed.1</b>	<b>Page : 001-089</b>	<b>Jakarta, Jun 2022</b>	<b>e-ISSN: 2614-8404</b>	<b>p-ISSN: 2776-3234</b>
-------------	----------------	-------------	---------------------------	------------------------------	------------------------------	------------------------------

**Enterprise Content Management (ECM) System Architecture for Capital Project at  
Oil and Gas Company**  
*Arief Herdiansah*

**85-89**

<b>JISA</b>	<b>Vol : 5</b>	<b>Ed.1</b>	<b>Page : 001-089</b>	<b>Jakarta, Jun 2022</b>	<b>e-ISSN: 2614-8404</b>	<b>p-ISSN: 2776-3234</b>
-------------	----------------	-------------	---------------------------	------------------------------	------------------------------	------------------------------

# Implementation of the AHP-SMARTER Method in the Decision Support System for Giving Sanctions for Violation of Student Disciplines

Sofiansyah Fadli<sup>1\*</sup>, Maulana Ashari<sup>2</sup>, Hasyim Asyari<sup>3</sup>, Ahmad Susan Pardiansyah<sup>4</sup>

<sup>1,2,3</sup> Program Studi Teknik Informatika, STMIK Lombok

<sup>4</sup> Program Studi Sistem Informasi, STMIK Lombok

Email: <sup>1</sup>[sofiansyah182@gmail.com](mailto:sofiansyah182@gmail.com), <sup>2</sup>[aarydarkmaul@gmail.com](mailto:aarydarkmaul@gmail.com), <sup>3</sup>[hasyimasyari25@gmail.com](mailto:hasyimasyari25@gmail.com),  
<sup>4</sup>[ahmad.pardiansyah84@gmail.com](mailto:ahmad.pardiansyah84@gmail.com)

**Abstract** – Violations of school rules are often carried out by students, including lack of respect for teachers, students who are not on time, often late for class, skipping classes, jumping fences, smoking and not paying attention to the rules and other regulations in school. This study aims to build a decision support system for violations of student discipline that has the ability to analyze each of the criteria and sub-criteria that have been determined by the school. In this case, students who violate school rules will be punished and given sanctions so as to provide an output value of priority intensity which results in a system that provides an assessment of violations against students. The method used in building this decision support system is by combining the Analytical Hierarchy Process (AHP) method and the Simple Multi Attribute Rating Technique Exploiting Rank (SMARTER) method. Weighting criteria using the AHP method and for ranking using the SMARTER method. The system created can be used to assist in processing data on violations of school rules. With this decision support system, it is hoped that policy makers will have no difficulty in determining what types of actions and sanctions will be given to students who violate school rules.

**Keywords** – Decision Support System, AHP Method, SMARTER Method, School Rules

## I. INTRODUCTION

Each school has its own policy in determining the level of student discipline. The Integrated Islamic Vocational High School (SMK) of Generasi Muslim Cendikia (GMC) still uses a system of calculating points for violations and determining the sanctions for violations that are still manual, namely by recording all events or student problems into a book. The decision support system suggested by the counseling guidance teacher is a system that makes it easier to evaluate the level of student discipline and sanctions for violations effectively and efficiently. Giving sanctions by teachers in the teaching process is influenced by several factors, namely the seriousness factor in learning, consequences, delinquency at the school level, and family stability factors.[2]. Education in Indonesia not only prioritizes the development of cognitive aspects or knowledge of students, but also pays attention to individual development as a whole person[4].

SMK-IT GMC is a vocational school that has quite a lot of students. Every school must have rules and regulations that must be obeyed and followed by every student but not infrequently these rules and regulations are violated, the violations that often occur are students who are not on time, often late for class, skipping class time, jumping fences, smoking and so on.

According to[5]The system of sanctions for violations of the rules in some schools is still in the form of warning letters and direct reprimands to students. Along with presents development of technology and communication a new challenge that can make guidance and counseling more practical. One of them is a Decision Support System which

is an approach to decision making[6]. The method that can support solving this problem is by combining the Analytical Hierarchy Process (AHP) method and the Simple Multi Attribute Rating Technique Exploiting Rank (SMARTER) method.

The system built can be used to assist in processing data on violations of school rules, especially student violations[3]. Although basically there are rules and sanctions that have been implemented in schools, the sanctions are still handled in the usual way without clear differences between the violations committed and the sanctions given (different violations the sanctions are almost the same).

Therefore, researchers want to design a decision support system for sanctions for violating student rules. Every student who violates the rules will be given sanctions so that it can provide a deterrent effect and increase the values of decency and order in the school environment. This is useful to facilitate decision making related to disciplinary issues.

## II. RESEARCH METHODOLOGY

### A. Decision Support System

Decision Support Systems (DSS) are usually built to support a solution to a problem or to an opportunity. Decision Support System (DSS) applications are used in decision making[7]. Decision Support System (DSS) application uses a flexible, interactive and adaptable CBIS (Computer Based Information System), which was developed to support solutions to unstructured specific management problems[8].

### B. AHP (Analytical Hierarchy Process)



This method was first developed by Saaty (Saaty, 1980)[9]. The hierarchical model stated by Saaty is a functional hierarchical model with the main input being human perception.

In general, the steps in using the AHP method for solving a problem are as follows[10]:

- Defining the problem and determining the desired solution.
- Determining the priority of elements
- Synthesis

The things to do in this step are:

- Sum the values of each column in the K matrix.
- Divide each value from the column by the corresponding column total to obtain a normalized matrix.
- Sum the values of each row and divide by the number of elements to get the priority weight value.

- Measuring Consistency

The things that are done in this step are as follows:

- Each value in the first column is multiplied by the priority weight of the first element, then each value in the second column is multiplied by the priority weight of the second element and so on.
- Sum each row ( $\sum$  row).
- The result of the sum of the rows is divided by the priority element in question so that it gets Lambda.

$$\lambda = \frac{\sum \text{row}}{\text{priority}} \quad (1)$$

- Sum Lambda ( $\lambda$ ) and the result is divided by the number of elements present, the result is called  $\lambda$  max.

$$\lambda_{\max} = \frac{\sum \lambda}{n} \quad (2)$$

- Calculate Consistency Index (CI) with formula:

$$CI = \frac{(\lambda_{\max} - n)}{n - 1} \quad (3)$$

- Compare Consistency Ratio (CR) with formula:

$$CR = CI/RC \quad (4)$$

Table1. Random Consistency Value (RC)

N	1,2	3	4	5	6	7	8
Rin	0.00	0.58	0.90	1.12	1.24	1.32	1.41

- Checking hierarchy consistency

### C. Simple Multi Attribute Rating Technique Exploiting Rank (SMARTER)

According to[2]states that SMARTER is a multi-criteria decision-making technique based on the theory that each alternative consists of a number of criteria that have values and each criterion has a weight that describes its importance when compared to other criteria. This weighting is used to assess each alternative in order to obtain the best alternative. SMARTER uses a linear additive model to predict the value of each alternative. The analysis involved is transparent so this method provides a high level of understanding of the problem and can be accepted by decision makers[1].

The model used in SMART is shown in the equation:

$$U(ai) = \sum_{j=1}^k Wj Ui(ai) \quad (5)$$

Information :

$Wj$  = The weighting value of the J-th criterion of the k criteria.

$U(ai)$  = The utility value of the I-th criterion for the I-th criterion

Where  $I = 1, 2, \dots, m$

The steps of the SMARTER method are as follows[3]:

- Determine the number of criteria for the decision to be taken.
- Giving weight to each criterion by using an interval of 1-100 for each criterion with the most important priority.
- Calculating the normalization of each criterion by comparing the value of the weight of the criteria with the number of weights of the criteria, using the formula:

$$NWj = \frac{Wj}{\sum_{n=1}^k Wn} \quad (6)$$

Information :

$NWj$  = Normalization of J-th criterion weights

$Wj$  =J-th criterion weight

$k$  = Number of criteria

$Wn$  = The weight of the N-th criterion.

- Provide a criterion value for each alternative
- Calculates final grades and performs rankings using the SMARTER model.

### D. Research Stages

To assist in the preparation of this research, it is necessary to have a clear framework for the stages[11]. This framework is the steps that will be taken in solving the problems that will be discussed.



Figure 1. Research Stages





- a. Identification of problems that occur in SMK-IT Generasi Muslim Cendikia is the current system that is still not standardized in this case different violations (mild and severe) but the handling is the same and the sanctions given are sometimes the same as other violations. In giving sanctions, there is only a warning and a statement letter, so there are several procedures that are not in accordance with the procedures that should have been applied to students.
- b. In this study, data collection was done by interview, observation and literature study. At this stage, it is done to find out, get data and information that will later support this research[12].

**Observation Method**

Observations were carried out directly at SMK-IT GMC by looking at the daily lives of students and teachers as well as existing problems to find out the types of violations and sanctions that students received if they violated the rules and regulations.

**Interview Method**

Interviews were conducted by asking directly to the Guidance Counseling teacher who directly handles problematic students at SMK-IT GMC.

**Library Study Method**

Literature study is done by reading various kinds of information related to the research title. Researchers took reference sources from national scientific journals and books from the internet.

- c. The problem analysis step is needed to determine recommendations for sanctions for violations of school rules committed by students. With this data analysis, a clear picture of the problems discussed will be obtained[7].
- d. Decision Support System Design, this stage is the activity carried out to make the formulation of the model, the selection of what criteria are taken into consideration for decision makers to decide the best alternative, measure and predict the results that occur.[4].
- e. In this study, the authors implement the AHP-SMARTER method so that they are able to provide recommendations for sanctions for violations of school regulations committed by students. This phase translates the design results into software.
- f. The process of testing the application using blackbox. Testing is done by testing all existing navigation, this test ensures that the processes carried out produce output that is in accordance with the design that has been made[13].
- g. Conclusions are drawn after the design, implementation, and testing stages have been completed[14]. This stage discusses the results of the final goal to be achieved, namely the creation of a decision support system application that can later benefit schools related to the provision of appropriate sanctions in accordance with existing standard procedures.[13].

**E. Research Material**

The research material used to make a decision support system for the awarding of sanctions for student discipline violations is by using the AHP-SMARTER method. With

the object of research SMK-IT Generasi Muslim Cendikia.

**F. Design Model**

Research with the application of the SMARTER Method in determining the sanctions for violations which will be combined with the AHP method, will use linear sequential in the design model. The activities in linear sequential are:

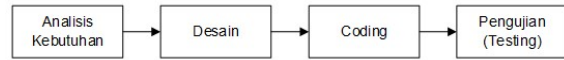


Figure 2. Linear Sequential Model[15]

- a. Requirements analysis is the stage of analyzing the needs needed in making software
- b. The design stage is the translation stage of the analyzed data into a form that is easily understood by users.
- c. Coding is the stage of translating data that has been designed using a particular programming language.
- d. Testing is the stage of testing the software that has been made.

**III. RESULTS AND DISCUSSION**

The implementation of this system is carried out using two process methods, namely weighting criteria using the AHP method and ranking using the SMARTER method.

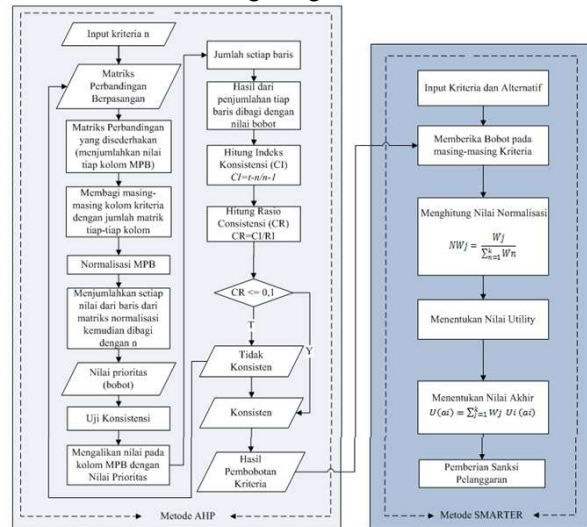


Figure 3. AHP-SMARTER Method Completion Flowchart

The dotted line indicates the transfer of the calculation process from the AHP method to the SMARTER method, indicating the separation between the AHP process and the SMARTER process. In the AHP method, after the weights are obtained, it will be continued by testing the consistency. The goal is whether the weights are consistent or not. If the weights are consistent, it will result in weighting, and if not, it will return to the pairwise comparison matrix. After the weight is obtained, it will be continued with ranking using the SMARTER method[1]. The weights obtained in the AHP method will be used as weights on the criteria.

**Decision Support System Criteria and Alternatives**

The criteria used in this decision support system are as follows:



Table 2. Violation Criteria

No	Criteria	Type of Violation	Point
1	C1	<b>Attendance at school</b>	
		a Absence without explanation 1-3 times	5
		b Absence without explanation 4-6 times	10
		c Absence without explanation 7-10 times	15
	d Absence without explanation more than 10 times	20	
2	C2	<b>School Uniform</b>	
		a Uniforms not in accordance with the terms of the day of use	5
		b Not wearing shoes at school	5
		c Wearing a hat in class or hijab is not uniform	5
	d Incomplete attribute	5	
3	C3	<b>Leaving School</b>	
		a In effective hours without explanation	10
	b Permission to leave and not return to school is not in the school's interest	15	
4	C4	<b>Courtesy of Association</b>	
		a Jump over the fence	15
		b Dating in the school environment	20
		c Mocking/ threatening/ hitting teachers/ employees	50
	d Caught pregnant, pregnant, married	10	
5	C5	<b>Discipline</b>	
		a Male student wear earrings, bracelets, necklaces, tattoos	10
		b Male student with long hair, dyeing hair other than black	20
		c Bringing books, magazines, tapes, VCDs is prohibited	25
		d Smoking or carrying a smoking device in the school environment	30
		e Smoking outside the school environment wears school attributes	30
		f Bring a cellphone and use it during class hours	30
		g Getting into fights or molesting fellow students	50
		h Carrying and using illegal drugs and beverages	75
		i Arrested for a crime and proven	10
		j Carrying sharp weapons & firearms, thereby harming and threatening the safety of others	10

Table 3. School Action

No	Action Code	Point Range	School Action
1	T0	0.1 – 0.9	Verbal Reprimand
2	T1	1 – 10	Held coaching by Guidance teachers and homeroom teachers
3	T2	11 – 25	Parents are called to school, Coaching is held by Guidance Counseling teachers and homeroom teachers, Make guidance statements
4	T3	26 – 40	Parents are called to school, Guidance is held by the Guidance Counseling teachers and homeroom teacher, Makes a guidance statement and gives the 1st Warning Letter to parents/guardians
5	T4	41 – 55	Parents are called to school, Guidance is held by the Guidance Counseling teachers and homeroom teacher, Makes a guidance statement and gives

No	Action Code	Point Range	School Action
6	T5	56– 75	a 2nd warning letter to parents/guardians Parents are called to school, Guidance from the principal is witnessed by the homeroom teacher, Counseling Guidance teacher and students, Makes a statement letter stamped 6000 about willingness to be issued if the score is above 75 and does not go up class
7	T6	76 – 100	Parents are called to school, students are returned to parents

Table 4. Type of Sanction

No	Sanction Code	Point Range	Type of Sanction
1	S0	0.1 – 0.9	Doing Cleaning
2	S1	1 – 10	Not allowed to follow class hours until the change of lessons
3	S2	11 – 25	Make a statement known to the homeroom teacher and parents/guardians
4	S3	26 – 40	1st Warning Letter and 2 day suspension
5	S4	41 – 55	2nd Warning Letter and 5 day suspension
6	S5	56– 75	Stay in class
7	S6	75 – 100	Expelled from school

The alternatives used in this decision support system are as follows:

This alternative set is the students of SMK-IT GMC, as a sample taken as many as 5 students, so that if there are 5 alternative decisions, then these alternatives can be written as  $A = \{A_i | i = 1, 2, 3, 4, 5\}$  with:

- A1: Student 1
- A2: Student 2
- A3: Student 3
- A4: Student 4
- A5: Student 5

#### Calculation Using AHP Method

The next stage is to determine the priority of the elements by compiling criteria and sub-criteria in the form of a pairwise comparison matrix[8]. To find out the results of the weighting of the criteria used in calculating the priority of criteria and sub-criteria with the AHP method, it is necessary to search for values. How to get a value that can be with a certainty value or by conducting a survey through several respondents using a questionnaire sheet[11]. The value of certainty is a value that is directly given for certain criteria, while the value of the questionnaire is the value obtained from the assessment given by the respondent where each respondent gives a different preference value using a scale of 1-9 [8].

Determining the priority of elements by compiling these criteria in the form of a pairwise comparison matrix[9].

Table 5. Pairwise comparison matrix

	C1	C2	C3	C4	C5
C1	1.000	0.500	0.500	0.500	0.500
C2	2.000	1.000	0.500	0.500	0.333
C3	2.000	2.000	1.000	0.500	0.500
C4	2.000	2.000	2.000	1.000	0.500
C5	2.000	3.000	2.000	2.000	1.000
<b>Total</b>	9.000	8.500	6.000	4.500	2.833



Next is to calculate the value of the criteria column elements, where each criterion column element is divided by the number of matrices for each column in table 5, then add up the row matrix of the values of each element.

Table 6. Normalization Matrix of Criteria Element Values

	C1	C2	C3	C4	C5	Total
<b>C1</b>	0.111	0.059	0.083	0.111	0.176	0.541
<b>C2</b>	0.222	0.118	0.083	0.111	0.118	0.652
<b>C3</b>	0.222	0.235	0.167	0.111	0.176	0.912
<b>C4</b>	0.222	0.235	0.333	0.222	0.176	1.190
<b>C5</b>	0.222	0.353	0.333	0.444	0.353	1.706
<b>Total</b>	1.000	1.000	1.000	1.000	1.000	5.000

After determining the number of criteria columns, the next step is to calculate the priority value of the criteria or create a criteria consistency matrix with the formula for the number of criteria elements divided by the number of criteria in this case 5.

Table 7. Average matrix of criteria consistency normalization

	C1	C2	C3	C4	C5	Priority
<b>C1</b>	0.111	0.059	0.083	0.111	0.176	0.108
<b>C2</b>	0.222	0.118	0.083	0.111	0.118	0.130
<b>C3</b>	0.222	0.235	0.167	0.111	0.176	0.182
<b>C4</b>	0.222	0.235	0.333	0.222	0.176	0.238
<b>C5</b>	0.222	0.353	0.333	0.444	0.353	0.341
<b>Total</b>	1.000	1.000	1.000	1.000	1.000	1.000

The next stage is to multiply the elements in the pairwise comparison matrix column multiplied by the priority value results in Table 7, the multiplication results are then added up per each row.

Table 8. The summation matrix of each row

	C1	C2	C3	C4	C5	Quantity Per Line
<b>C1</b>	0.108	0.065	0.091	0.119	0.171	0.554
<b>C2</b>	0.216	0.130	0.091	0.119	0.114	0.671
<b>C3</b>	0.216	0.261	0.182	0.119	0.171	0.949
<b>C4</b>	0.216	0.261	0.365	0.238	0.171	1.250
<b>C5</b>	0.216	0.391	0.365	0.476	0.341	1.789

The next step is to add up the matrix of the sum of each row in Table 6 with the result of the "priority" value in Table 8.

Table 9. The sum of the number of elements per line with the priority value

	Quantity Per Line	Priority	Result
<b>C1</b>	0.554	0.108	0.662
<b>C2</b>	0.671	0.130	0.801
<b>C3</b>	0.949	0.182	1.131
<b>C4</b>	1.250	0.238	1.488
<b>C5</b>	1.789	0.341	2.130
		<b>t =</b>	5.194
		<b>CI =</b>	0.048
		<b>IR =</b>	1.12
		<b>CR =</b>	0.043
		<b>Consistence</b>	

From table 8, the following values are obtained:  
 $t = (1/5) * ((0.554/0.108) + (0.671/0.130) + (0.949/0.182) + (1.250/0.238) + (1.789/0.341)) = 5.194$   
 For  $n = 5$  obtained  $RI_6 = 1.12$  so that:  
 $CI = (5.194-5) / (5-1) = 0.048$   
 $RI_6 = 1.12$   
 $CR = (CI/RI_6) = -0,048/ 1.2 = -0.043$   
 Therefore  $CR \leq 0,1$  then the consistency ratio of the calculation is acceptable (consistent).

From the results of the calculations in the table above, the value of the preference weights can show that the most important weight order criteria with a weight of 34.1%. Next are the criteria for Politeness in Association with a weighted value of 23.8%, the criteria for leaving school with a value of 18.2%, the criteria for school uniforms 13.0% and the criteria for school attendance with a weighting value of 10.8%.

Table 10. Criteria Weight Preference

No	Criteria	(%) Weight	Weight (Wj)
<b>1</b>	Attendance at school	10.8 %	0.108
<b>2</b>	School uniform	13.0 %	0.130
<b>3</b>	Leaving school	18.2 %	0.182
<b>4</b>	Courtesy of association	23.8 %	0.238
<b>5</b>	Discipline	34.1 %	0.341
<b>Total</b>		<b>100%</b>	<b>1</b>

Calculation Using the SMARTER Method

Weighting on SMART uses a scale between 0 and 1, making it easier to calculate and compare values for each alternative[1]. The model used in SMART is shown in

Table 11. Criteria Weight Preference

No	Criteria	Type of Violation	Point	Weight Wj
<b>1</b>	<b>C1</b>	<b>Attendance at school</b>		
	a	Absence without explanation 1-3 times	5	10.8 %
	b	Absence without explanation 4-6 times	10	
	c	Absence without explanation 7-10 times	15	
	d	Absence without explanation more than 10 times	20	
<b>2</b>	<b>C2</b>	<b>School Uniform</b>		
	a	Uniforms not in accordance with the terms of the day of use	5	13.0 %
	b	Not wearing shoes at school	5	
	c	Wearing a hat in class or hijab is not uniform	5	
	d	Incomplete attribute	5	
<b>3</b>	<b>C3</b>	<b>Leaving School</b>		
	a	In effective hours without explanation	10	18.2 %
	b	Permission to leave and not return to school is not in the school's interest	15	
<b>4</b>	<b>C4</b>	<b>Courtesy of Association</b>		
	a	Jump over the fence	15	23.8 %
	b	Dating in the school environment	20	
	c	Mocking/ threatening/ hitting teachers/ employees	50	
<b>5</b>	<b>C5</b>	<b>Discipline</b>		
	a	Male student wear earrings, bracelets, necklaces, tattoos	10	34.1 %
	b	Male student with long hair, dyeing hair other than black	20	
	c	Bringing books, magazines, tapes, VCDs is prohibited	25	
	d	Smoking or carrying a smoking device in the school environment	30	
	e	Smoking outside the school environment wears school attributes	30	
	f	Bring a cellphone and use it during class hours	30	
	g	Getting into fights or molesting fellow students	50	



h	Carrying and using illegal drugs and beverages	75
i	Caught pregnant, pregnant, married	100
j	Arrested for a crime and proven	100
k	Carrying sharp weapons & firearms, thereby harming and threatening the safety of others	100

Sample Calculation Using the SMARTER Method  
NIS : 1719

Name : Supriadi

Type of Violation :

1. Do not enter without information 4 days a week
2. Incomplete attribute
3. Permission to leave and not return to school and not in the interest of the school
4. Jump over the fence
5. Bring cellphones to school and use them during class hours

Calculations using the SMARTER method are as follows :

a. Finding the utility value is as follows:

Utility value formula :

$$U_i(a_i) = 100 \frac{(C_{max} - C_{min})}{(C_{max} - C_{min})} \% \quad (7)$$

Information:

$U_i(a_i)$  = the utility value of the 1st criterion for the i-th criterion

$C_{max}$  = maximum criterion value

$C_{min}$  = minimum criterion value

How to get the utility value as follows:

1. School Attendance Criteria

$$U_i(a_i) = 100 \frac{(10 - 5)}{(20 - 5)} \%$$

$$U_i(a_i) = 100 \frac{(5)}{(15)} \%$$

$$U_i(a_i) = 100 \cdot 0.3333 \%$$

$$U_i(a_i) = 33.33$$

2. School Uniform Criteria

$$U_i(a_i) = 100 \frac{(5 - 5)}{(5 - 5)} \%$$

$$U_i(a_i) = 100 \frac{(0)}{(0)} \%$$

$$U_i(a_i) = 100 \cdot 0 \%$$

$$U_i(a_i) = 0$$

3. Criteria for Leaving School

$$U_i(a_i) = 100 \frac{(15 - 10)}{(15 - 10)} \%$$

$$U_i(a_i) = 100 \frac{(5)}{(5)} \%$$

$$U_i(a_i) = 100 \cdot 1 \%$$

$$U_i(a_i) = 100$$

4. Criteria for Courtesy of Association

$$U_i(a_i) = 100 \frac{(15 - 15)}{(50 - 15)} \%$$

$$U_i(a_i) = 100 \frac{(0)}{(35)} \%$$

$$U_i(a_i) = 100 \cdot 0 \%$$

$$U_i(a_i) = 100$$

5. Order Criteria

$$U_i(a_i) = 100 \frac{(30 - 10)}{(100 - 10)} \%$$

$$U_i(a_i) = 100 \frac{(20)}{(90)} \%$$

$$U_i(a_i) = 100 \cdot 0.2222 \%$$

$$U_i(a_i) = 22.22$$

b. The result value is obtained from:

Formula = Value of utility x normalization

1. School Attendance Criteria

$$\text{Result} = 33.33 \times 0.108 = 3.60$$

2. School Uniform Criteria

$$\text{Result} = 0 \times 0.13 = 0$$

3. Criteria for Leaving School

$$\text{Result} = 100 \times 0.182 = 18.2$$

4. Criteria for Courtesy of Association

$$\text{Result} = 0 \times 0.238 = 0$$

5. Order Criteria

$$\text{Result} = 22.22 \times 0.341 = 7.58$$

c. Looking for the Final Result of SMARTER Calculation

$$= U(a_i) \sum_{j=1}^m NW_j U_i(a_i) \quad (8)$$

$$\text{Result} = 3.60 + 0 + 18.2 + 0 + 7.58$$

$$= 29.38$$

NIS : 3454

Name : Muhamad Sunardi

Type of Violation :

1. Did not enter / did not attend without explanation / alpha more than 3 times
2. Hijab is not uniform
3. Uniforms not in accordance with the terms of the day of use

Calculation using the SMARTER method

a. Finding the utility value is as follows:

1. School Attendance Criteria

$$U_i(a_i) = 100 \frac{(5 - 5)}{(20 - 5)} \%$$

$$U_i(a_i) = 100 \frac{(0)}{(15)} \%$$

$$U_i(a_i) = 100 \cdot 0 \%$$

$$U_i(a_i) = 0$$

2. - School Uniform Criteria

$$U_i(a_i) = 100 \frac{(5 - 5)}{(5 - 5)} \%$$

$$U_i(a_i) = 100 \frac{(0)}{(0)} \%$$

$$U_i(a_i) = 100 \cdot 0 \%$$

$$U_i(a_i) = 0$$

- School Uniform Criteria

$$U_i(a_i) = 100 \frac{(5 - 5)}{(5 - 5)} \%$$

$$U_i(a_i) = 100 \frac{(0)}{(0)} \%$$



$$U_i(ai) = 100 \cdot 0\%$$

$$U_i(ai) = 0$$

5. Order Criteria

$$U_i(ai) = 100 \frac{(50 - 10)}{(100 - 10)}\%$$

$$U_i(ai) = 100 \frac{(40)}{(90)}\%$$

$$U_i(ai) = 100 \cdot 44.44\%$$

$$U_i(ai) = 44.44$$

b. The result value is obtained from:

Formula = Value of utility x normalization

1. School Attendance Criteria

$$\text{Result} = 0 \times 0.108 = 0$$

2. -School Uniform Criteria

$$\text{Result} = 0 \times 0.13 = 0$$

- School Uniform Criteria

$$\text{Result} = 0 \times 0.13 = 0$$

5. Order Criteria

$$\text{Result} = 44.44 \times 0.341 = 15.15$$

c. Finding the Final Result of SMARTER Calculation

$$= U(ai) \sum_{j=1}^m NW_j U_i(ai)$$

$$\text{Result} = 0 + 0 + 0 + 15.15 = 15.15$$

Table 12. SMARTER Calculation Result

No	Student Name	Criteria	Point	Normalization
1	Supriadi	C1.b	10	0.108
		C2.d	5	0.130
		C3.b	15	0.182
		C4.a	15	0.238
		C5.f	30	0.341
2	Muhamad Sunardi	C1.a	5	0.108
		C2.a	5	0.130
		C2.c	5	0.130
		C5.g	50	0.341
		C3.a	15	0.182
3	Lalu Akbar Hasibuan	C4.a	15	0.238
		C5.d	30	0.341
		C5.f	30	0.341
		C1.c	15	0.108
		C2.d	5	0.130
4	Roy Ardianto Putra	C5.f	30	0.341
		C1.a	5	0.108
		C3.b	15	0.182
		C4.a	15	0.238
		C5.b	20	0.341
5	Rumlan Hasanudin	C5.e	30	0.341
		C1.c	15	0.108
		C2.c	5	0.130
		C2.d	5	0.130
		C3.a	10	0.182
6	Maulana Gilang Apriano	C1.d	20	0.108
		C2.a	5	0.130
		C3.b	15	0.182
		C5.f	30	0.341
		C1.c	15	0.108
7	Wahyuni Sawitri	C2.c	5	0.130
		C2.d	5	0.130
		C3.a	10	0.182
		C1.d	20	0.108
		C2.a	5	0.130
8	Marhan Ristu	C3.b	15	0.182
		C5.f	30	0.341
		C1.d	20	0.108
		C2.d	5	0.130
		C4.a	15	0.238
9	Lalu Fikto Alanda Sofia	C4.b	20	0.238
		C1.d	20	0.108
		C2.d	5	0.130
		C4.a	15	0.238
		C4.b	20	0.238

Table 13. Advanced SMARTER Calculation Results

Utility Value	Final Result	Action	Type of Sanction		
33.33	0	100	29.38	T3	S3
0	0	0	0	0	0
22.22	0	0	15.15	T2	S2
0	0	0	0	0	0
44.44	0	100	33.36	T3	S3
0	0	0	0	0	0
22.22	0	0	14.78	T2	S2
22.22	0	0	0	0	0
66.67	0	100	29.57	T3	S3
0	0	0	0	0	0
22.22	0	0	14.78	T2	S2
66.67	0	0	0	0	0
0	0	0	25.4	T3	S3
0	0	0	0	0	0
100	0	100	29	T3	S3
100	0	0	7.58	T1	S1
22.22	0	0	0	0	0
100	0	0	14.13	T2	S2
0	0	0	0	0	0
14	0	0	0	0	0

Table 14. Value Range

No	Value Range	Information
1	1 – 10	Normal
2	11 – 25	Slight/Light
3	26 – 40	Medium
4	41 – 55	Heavy Enough
5	56 – 74	Heavy
6	75 – 100	Very Heavy

Use Case Diagram

- In the Use Case Diagram below, there are 4 actors who play a role in the running of the program. The first actor is the BK teacher, the BK teacher can do the login process, manage data such as student data, violation data, witness data, action data, summons, and change passwords.
- The second actor is students, in this system students can log in and view their own data.
- The third actor is the principal, in this system the principal can log in and see all the existing data. The principal also received a report
- The fourth actor is the student's guardian, in this system the student's guardian can log in and view the data on rules, violations, sanctions and student/children's own data. Guardians of students can also receive summons.





NIS	Nama Siswa	Kelas	Jurusan	JK	Tempat Lahir	Tanggal Lahir	Alamat	Nama Ayah	Nama Ibu
1777	Nurul Hidayah	XII	AP2	IP	Pengapung	2002-11-20	KIJIG	H Hidayah	Hj Miasari
1779	Piangga Cahaya Widaya	XII	AP2	L	Mong 1	2001-02-28	Mong 1	Salim	Minatip
1780	RANI SEPTIA KINGSURUM	XII	AP2	IP	KUKUN	2001-12-17	KUKUN	AMAZ RANI	MIATRE

Figure 8. Student Data Form

### Display of Violation Data Form

On the violation data page, each violation is directly inputted by selecting the criteria for the violation, the type of violation, the point of violation and the percentage of weight that has been normalized into decimal form.

Kode Pelanggaran	Jenis Pelanggaran	Point	Persentase	Aksi
C1 - Kehadiran di Sekolah	a. Tidak masuk / tidak hadir tanpa keterangan' alpha 3-9 kali	5	0.108	Ubat / Hapus
C1 - Kehadiran di Sekolah	b. Tidak masuk / tidak hadir tanpa keterangan' alpha 4-6 kali	10	0.108	Ubat / Hapus
C1 - Kehadiran di Sekolah	c. Tidak masuk / tidak hadir tanpa keterangan' alpha 7-10 kali	15	0.108	Ubat / Hapus
C1 - Kehadiran di Sekolah	d. Tidak masuk / tidak hadir tanpa keterangan' alpha lebih dari 10 kali	20	0.108	Ubat / Hapus
C2 - Seragam Sekolah	a. Seragam tidak sesuai dengan ketentuan hari pengembangannya	5	0.130	Ubat / Hapus
C2 - Seragam Sekolah	b. Tidak berpakaian memakai sandal selama di sekolah	5	0.130	Ubat / Hapus
C2 - Seragam Sekolah	c. Memakai topi dalam kelas / tidak memakai seragam	5	0.130	Ubat / Hapus
C2 - Seragam Sekolah	d. Akibat tidak lengkap	5	0.130	Ubat / Hapus
C3 - Meninggalkan Sekolah	a. Pada jam absen tanpa keterangan	10	0.182	Ubat / Hapus
C3 - Meninggalkan Sekolah	b. Ditun keluar dan tidak kembali lagi ke sekolah / bujukan kesengajaan sekolah	15	0.182	Ubat / Hapus

Figure 9. Violation Data Form

### Display of Rules of Conduct

The fields on the code of conduct data page are the code of conduct and the name of the code of conduct.

Kode Tata Tertib	Nama Tata Tertib	Aksi
A - Kewajiban Siswa	1. Siswa masuk sebelum jam pelajaran dimulai jam 07.35. Wktu dan pulang jam 14.15, kecuali hari libur/masuk, pulang jam 11.15 Wktu.	Ubat / Hapus
A - Kewajiban Siswa	2. Menggunakan bahasa seragam: a. Serin dan Serasa (Pilih-Abu dan Topi b. Ratu dan Karim; (Pakaian seragam: lumJCM" (Pakaian Intaj) d. Sabtu; (Pakaian Pramuka	Ubat / Hapus
A - Kewajiban Siswa	3. Mengikuti kegiatan sekolah seperti: Upacara, Intaj dan kegiatan kegiatan lain yang ditentukan sekolah.	Ubat / Hapus
A - Kewajiban Siswa	4. Menjaga kebersihan ruang kelas dan lingkungan sekolah.	Ubat / Hapus
A - Kewajiban Siswa	5. Menjaga ketertarikan sarana dan prasana sekolah.	Ubat / Hapus
A - Kewajiban Siswa	6. Menggunakan bahasa yang santun terutama bahasa Indonesia yang baik dan benar.	Ubat / Hapus
A - Kewajiban Siswa	7. Meminta (dan terima) beasiswa, beasiswa atau beasiswa yang sangat urgen dan dituntut pada buku gribit.	Ubat / Hapus
A - Kewajiban Siswa	8. Hadir dalam setiap upacara/mas (SPS) minimal 90%.	Ubat / Hapus

Figure10. Rules of Conduct Form

### Sanction Form Display

The sanction data page is filled with inputting the sanction code, point range and type of sanction.

Kode Sanksi	Rentang Point	Jenis Sanksi	Aksi
S0	0.1-0.9	Teguran Lisan	Ubat / Hapus
S1	1-10	Tidak dibenarkan mengikuti jam pelajaran setelah penggantian pelajaran	Ubat / Hapus
S2	11-25	Membuat pernyataan di sekolah oleh wali kelas/orang tua/wali murid	Ubat / Hapus
S3	26-40	SP 1 dan dikorsing 2 hari	Ubat / Hapus
S4	41-55	SP 2 dan dikorsing 5 hari	Ubat / Hapus
S5	56-75	Tinggal kelas	Ubat / Hapus

Figure 11. Sanction Form

### Action Data Form Display

The action data page is filled with inputting the action code, point range and type of action.

Kode Tindakan	Rentang Point	Jenis Tindakan	Aksi
T0	0.1-0.9	Diberikan Teguran oleh guru BK	Ubat / Hapus
T1	1-10	Diberikan pembinaan oleh guru BK dan wali kelas	Ubat / Hapus
T2	11-25	Orang tua dipanggil ke sekolah, Diberikan pembinaan oleh guru BK dan wali kelas, Membuat pernyataan di sekolah.	Ubat / Hapus

Figure12. Action Form

### Violator Data Input Form Display

On the violator's data input page, all student data already exists so that if there are students who violate the admin immediately look for the student's name and click the violating button.

NIS	Nama Siswa	TTL	Kelas	Jurusan	Aksi
1719	SUPIRI	Batang, 2002-02-22	XII	IPA	Pelanggaran
1720	SURBATI	LOKAI, 2002-05-26	XII	IPA	Pelanggaran
1721	TAURMAN ADZHAR	HAWUN, 2002-06-02	XII	IPA	Pelanggaran
1722	Abdul Rahman	BERAH, 2001-06-21	XII	AP1	Pelanggaran
1723	ANUS MARA	Gumung Batu, 2001-12-31	XII	AP1	Pelanggaran
1724	AYUN ARFIN	Sengkot, 2003-02-17	XII	AP1	Pelanggaran
1725	BAIQ NEDA YASNITA	BUN GUNBUK, 2002-12-23	XII	AP1	Pelanggaran
1726	BAIQ SOVIANTI	BUNURU, 2002-09-24	XII	AP1	Pelanggaran
1727	Dani Triana Ramdani	Lemah, 2002-11-18	XII	AP1	Pelanggaran

Figure 13. Violator Data Input Form

After selecting the violating student, it will be processed by selecting the criteria for the violation and the type of violation then the violation process.



**Proses Pelanggaran Siswa**

Tanggal Kejadian: 16/04/2021

NIS: 1361

Nama Siswa: Rang Widawati

Kelas: XI

Jurusan: SMA

**Kriteria Pelanggaran**

No	Kriteria Pelanggaran	Jenis Pelanggaran	Poin	Aksi
1	C1 - Kehadiran di Sekolah	a. Tidak masuk / tidak hadir tanpa keterangan/ alpha 1-3 kali	5	<input type="checkbox"/>
2	C1 - Kehadiran di Sekolah	b. Tidak masuk/ tidak hadir tanpa keterangan/ alpha 4-6 kali	10	<input type="checkbox"/>
3	C1 - Kehadiran di Sekolah	c. Tidak masuk/ tidak hadir tanpa keterangan/ alpha 7-10 kali	15	<input type="checkbox"/>
4	C1 - Kehadiran di Sekolah	d. Tidak masuk/ tidak hadir tanpa keterangan/ alpha lebih dari 10 kali	20	<input type="checkbox"/>

Figure 14. Student Violation Filling Page

After the violation process will be summed up all types of violations committed then will be shown the type of sanctions that will be given and the actions that will be taken by the Counseling Guidance teacher.

#### Display of Violation Point Calculation Result Form

The results obtained after all violations are processed are the display of the number of points, the sanctions obtained and the actions to be taken by the Counseling Guidance teacher. After the calculation results appear, the Admin can print the results by clicking the print results button.

**Proses Pelanggaran Oleh Siswa**

Proses pelanggaran berhasil dilakukan sebanyak 8 pelanggaran

**Hasil perhitungan point pelanggaran**

**37.75**

Sanksi yang didapatkan berdasarkan point pelanggaran: (S3) - SP 1 dan skorsing 2 hari

Tindakan yang dilakukan berdasarkan point pelanggaran: (T3) - Orang tua dipanggil ke sekolah. Diadakan pertemuan oleh guru BK dan wali kelas. Membuat pernyataan bimbingan dan membuat surat pernyataan 1 untuk orang tua/wali murid.

No	Kriteria Pelanggaran	Jenis Pelanggaran	Poin	Persentase
1	C1 - Kehadiran di Sekolah	b. Tidak masuk/ tidak hadir tanpa keterangan/ alpha 4-6 kali	10	0.108
2	C1 - Kehadiran di Sekolah	b. Tidak masuk/ tidak hadir tanpa keterangan/ alpha 7-10 kali	15	0.108
3	C2 - Seragam Sekolah	c. Memakai topi dalam kelas/ jilbab tidak seragam	5	0.130
4	C2 - Seragam Sekolah	d. Absen/ tidak lengkap	5	0.130

Figure 15. Violation Point Calculation Results Page

After the violation committed by the student is processed, the admin can print the violation card.

**DINAS PENDIDIKAN DAN KEBUDAYAAN**  
**YAYASAN GENERASI MUSLIM CENDEKIA**  
**SMK-IT GMC**

**KARTU PELANGGARAN SISWA**  
TANGGAL 2021-09-06

**Data Siswa**

Nama: SALLY KARLINA  
Tempat Lahir: Kuta, 2001-08-01  
Kelas: XI  
Jurusan: API  
Jenis Kelamin: P  
Alamat: Kuta II  
Nama Ayah: MURNAN  
Nama Ibu: Ganing

**Daftar Pelanggaran**

No	Kriteria Pelanggaran	Jenis Pelanggaran	Poin	Persentase
1	C1 - Kehadiran di Sekolah	b. Tidak masuk/ tidak hadir tanpa keterangan/ alpha 4-6 kali	10	0.108
2	C1 - Kehadiran di Sekolah	c. Tidak masuk/ tidak hadir tanpa keterangan/ alpha 7-10 kali	15	0.108
3	C2 - Seragam Sekolah	c. Memakai topi dalam kelas/ jilbab tidak seragam	5	0.130
4	C2 - Seragam Sekolah	d. Absen/ tidak lengkap	5	0.130
5	C3 - Masing-masing Sekolah	b. Bawa keluar dan tidak kembali lagi ke sekolah/ bukan kepentingan sekolah	15	0.182
6	C4 - Sopan Santun Pergaulan	a. Melompat pagar	15	0.25
7	C5 - Ketertiban	d. Merokok/ membawa alat untuk merokok di lingkungan sekolah	30	0.25
8	C5 - Ketertiban	e. Memakai di luar lingkungan sekolah/ memakai atribut sekolah	30	0.25

**Hasil perhitungan point pelanggaran**

**37.75**

Sanksi yang didapatkan berdasarkan point pelanggaran: (S3) - SP 1 dan skorsing 2 hari

Tindakan yang dilakukan berdasarkan point pelanggaran: (T3) - Orang tua dipanggil ke sekolah. Diadakan pertemuan oleh guru BK dan wali kelas. Membuat pernyataan bimbingan dan membuat surat pernyataan 1 untuk orang tua/wali murid.

Pangung, 2021-09-06  
Mengetahui Guru BK/Wali Kelas

Figure 16. Violation Result Print Form page

## IV. CONCLUSION

Based on the research carried out up to the stage of designing, implementing, and testing the software, it can be concluded that from testing the process of calculating student discipline violations with the AHP-SMARTER method, it can be used and is able to provide the right solution in making decisions about giving sanctions to participants. students who violate school rules. From the results of this study, the 5 highest violations committed by students were taken by looking at the first violation point 78.5 sanctions given S6 and actions taken by T6, the two students with 46.5 violation points with S4 sanctions and T4 sanctions, the third students with 31.25 violation points with S3 sanctions and T3 sanctions, the four students with 21.5 violation points with S2 sanctions and T2 actions and the five students with violation points 15.75 with a S2 sanction and T2 action. The decisions taken by the Counseling Guidance Teachers, homeroom teachers and principals can be accounted for with the support of model calculations in the decision support system.

## ACKNOWLEDGMENTS

The author would like to thank Ristekbrin who has fully funded this research through the Penelitian Dosen Pemula (PDP) Grant for the 2020 Proposed Year and the 2021 implementation year, all agencies and individuals who have provided support and assistance during the implementation of the research.

## REFERENCES

- [1] Riski, Dahana Erpa, Tommy, Hendrawan Aloysius, and Anardani Sri, "Rancang Bangun Sistem Pendukung Keputusan Pemberian Sanksi Pelanggaran Siswa Menggunakan Metode SMARTER," *Seminar Nasional Teknologi Informasi dan Komunikasi*, pp. 286-292, 2018.
- [2] Mesliani, Solikhun, and Fauzan M, "Sistem Pendukung Keputusan Dalam Penentuan Sanksi Pelanggaran Peraturan Sekolah Bagi Siswa Sekolah Dasar Negeri 098023Kecamatan Bosar Maligas," in *Seminar Nasional Matematika dan Terapan*, Medan, 2019, pp. 538-544.





- [3] Taufan, Asri, Zaen Mohammad, Daniatan, Janiah Baiq, and Fadli Sofiansyah, "Penerapan Metode SMART Dalam Sistem Pendukung Keputusan Pemberian Sanksi Pelanggaran Tata Tertib Siswa (Studi Kasus: SMK Negeri 1 Pujut)," *MISI (Jurnal Manajemen informatika & Sistem Informasi)*, vol. 4, no. 1, pp. 63-72, 2021.
- [4] Siregar Juarni, "Sistem Pendukung Keputusan Penentuan Proritas Konseling Siswa, Menggunakan Pendekatan AHP-TOPSIS," *Jurnal Sistem Informasi Smik Antar Bangsa*, vol. 6, no. 2, pp. 107-122, 2017.
- [5] Deswari Alan, "Sistem Pengambilan Keputusan Pemberian Sanksi Pelanggaran Kedisiplinan Pada SMP 1 Muhammadiyah Talang Padang," *Proceding KMSI (Konferensi Mahasiswa Sistem Informasi)*, pp. 39-51, 2014.
- [6] Delli, Wihartiko Fajar, Tita, Tosida Eneng, and Jaman, Sentosa Lola, "Sistem Penunjang Keputusan Strategi Tindakan Atas Pelanggaran Siswa Dengan Metode Analytical Network Process," *Komputasi: Jurnal Ilmiah Ilmu Komputer dan Matematika*, vol. 15, no. 1, pp. 102-110, 2018.
- [7] Perdana Adidtya and Budiman Arief, "Analysis of Multi-attribute Utility Theory for College Ranking Decision Making," *Sinkron : Jurnal dan Penelitian Teknik Informatika*, vol. 4, no. 2, pp. 19-26, 2020.
- [8] Kusumadewi Sri, Hartati Sri, Harjoko Agus, and Wardoyo Retantyo, *Fuzzy Multi-Attribute Decision Making (Fuzzy)*. Yogyakarta: Graha Ilmu, 2006.
- [9] Kusrini, *Konsep dan Aplikasi Sistem Pendukung Keputusan*. Yogyakarta: Andi, 2007.
- [10] Fadli Sofiansyah, Imtihan Khairul, and Fahmi Hairul, *Mengenal dan Memahami Sistem Pendukung Keputusan*. Jawa Tengah: CV. Amerta Media, 2020.
- [11] Sa'adati Yuan, Fadli Sofiansyah, and Imtihan Khairul, "Analisis Penggunaan Metode AHP dan MOORA untuk Menentukan Guru Berprestasi sebagai Ajang Promosi Jabatan," *SINKRON (Jurnal & Penelitian Teknik Informatika)*, vol. 3, no. 1, pp. 82-90, 2018.
- [12] Ipinuwati Sri, "Sistem Pendukung Keputusan Pemberian Sanksi Pelanggaran Kedisiplinan Siswa Pada Smk PGRI I Kedondong," *Jurnal Informatika*, vol. 14, no. 2, pp. 153-168, 2014.
- [13] Fadli Sofiansyah and Imtihan Khairul, "Penerapan multi-Objective Optimization On The Basis Of Ratio Analysis (MOORA) Method Dalam Mengevaluasi Kinerja Guru Honorer," *JIRE (Jurnal Informatika & Rekayasa Elektronika)*, vol. 2, no. 2, pp. 10-19, 2019.
- [14] Maryaningsih and Suranti Dewi, "Penerapan Metode Simple Multi Atributte Rating Technique Dalam Pemilihan Dosen Terbaik," *JIKO (Jurnal Informatika dan Ilmu Komputer)*, vol. 4, no. 1, pp. 8-15, 2021.
- [15] S, Pressman Roger, *Software Engineering: A Practitioner's Approach*. New York: Mc Graw Hill, 2009.
- [16] Dwi, Lestari Yuyun and Mardiana, "Decision Support System For Determining the Best College High Private Using Topsis Method," *Sinkron : Jurnal dan Penelitian Teknik Informatika*, vol. 4, no. 2, pp. 27-33, 2020.
- [17] Cahyo, Buono Lintang, Pandiangan Nurlala, and Zubaedah Reza, "Implementation of the Simple Multi Attribute Ranking Technique Method as a Model for Decision Making in Determining the Talents and Interests of Children in Continuing Education," in *Journal of Physics: Conference Series*, Indonesia, 2021, pp. 1-6.
- [18] Syahrian, Harahap Ahmad and Firman, "Sistem Pengaduan Layanan Gangguan Pelanggan Speedy Rantauprapat Berbasis WEB," *U-NET : Jurnal Teknik Informatika*, pp. 1-5, 2017.
- [19] Mariskhana Kartika, Dewi, Sintawati Ita, Widiarina, and Rusdiansyah, "Decision Support System for increasing position of Office at PT. Gramedia Asri Media using Profile Matching Method," *Kartika, Mariskhana; Ita, Dewi, Sintawati; Widiarina; Rusdiansyah*, vol. 5, no. 2, pp. 221-228, 2021.
- [20] Destiana Henny, Sudradjat Adjat, and Amira, Sefenizka Aprilah, "Decision Support System for Determining Exemplary Students Using SAW Method," *Sinkron : Jurnal dan Penelitian Teknik Informatika*, vol. 5, no. 1, pp. 138-145, 2020.



# Implementation of Data Mining on Tourist Visits Patterns on Lombok Island Tourism Objects

Saikin<sup>1\*</sup>, Sofiansyah Fadli<sup>2</sup>, Maulana Ashari<sup>3</sup>

<sup>1,3</sup>Program Studi Sistem Informasi, STMIK Lombok

<sup>2</sup>Program Studi Teknik Informatika, STMIK Lombok

Email: [1eken.apache@gmail.com](mailto:1eken.apache@gmail.com), [2sofiansyah182@gmail.com](mailto:2sofiansyah182@gmail.com), [3arydarkmaul@gmail.com](mailto:3arydarkmaul@gmail.com)

**Abstract** – Foreign tourists entering Indonesia in 2017 and 2018 have increased. From the data obtained on the website of the Ministry of Tourism (Kemenpar) the number of foreign tourists in 2017 was 14,039,799, while in 2018 there were 15,806.1, with a comparison of the number of tourists from the two years, the percentage increase in tourists was 12.58%. The data analysis approach using a classification model is a data analysis approach by studying the data and making predictions with the new data. in the classification model, there are many algorithms that can be applied in data analysis, one of which is the Decision Tree algorithm. This study aims to analyze the pattern of tourist visits based on the objects visited by the number of tourists visiting certain tourist objects. From the modeling using the Decision Tree C4.5 Algorithm and the scenario of splitting the data into three parts, the highest accuracy value was obtained for splitting data of 80:20 for train and testing data and max depth 7, which obtained an accuracy of 94% for train data and 92% for data. testing. Modeling with the Bootstrap Aggregating Method, the accuracy score obtained on training data is 93% and testing data is 92. percent. 3 accuracy results from using bagging reduce the accuracy of the C4.5 algorithm on the data training side from 94% to 93 percent, while the accuracy of testing data is still the same, namely 92%.

**Keywords** – C4.5 Algorithm, Data Mining, Decision Tree, Bootstrap Aggregating Method.

## I. INTRODUCTION

Foreign tourists entering Indonesia in 2017 and 2018 have increased. From the data obtained on the website of the Ministry of Tourism (Kemenpar) the number of foreign tourists in 2017 was 14,039,799, while in 2018 there were 15,806.1, with a comparison of the number of tourists from the two years, the percentage increase in tourists was 12.58%. . The number of incoming tourists is calculated from various entrances scattered in several areas that have tourist attractions, including the Lombok area in the Province of West Nusa Tenggara (NTB). Although from statistical data obtained from the tourism service website in the area in the range of 2017 to 2018, there was a decrease in the number of tourists, but in that year the number of tourists entered was more than two million tourists.

The number of tourist visits certainly increases the activity of the economic industry in various sectors, such as hospitality, culinary, transportation services and also tour and travel services. Companies engaged in these fields provide travel services for tourists, who will visit certain tourist objects. In providing services the company will adjust to the needs and interests of tourists, but the obstacle is the difficulty of predicting the needs or interests of tourists, an obstacle faced by companies engaged in tour and travel services. Understanding the needs and interests of tourists in choosing a tourist attraction to be visited by using existing data in the past to predict the interest of tourists in visiting the selected tourist attraction. Analysis of past data to predict future needs will assist management in making decisions, especially those closely related to tourist visiting patterns.

The data analysis approach using a classification model is a data analysis approach by studying the data and making predictions with new data. in the classification model, there are many algorithms that can be applied in data analysis,

one of which is the Decision Tree algorithm. Decision trees are one of the most popular classification methods because they are easy to interpret by humans. The concept of a decision tree is to convert data into a decision tree and decision rules, [1]. Decision trees are widely used to solve decision-making cases such as medicine (diagnosing patient disease), computer science (data structure), psychology (decision-making theory) and so on [2].

Research conducted by [3] entitled Predicting the Number of Foreign Tourist Visits to Bali Using Support Vector Regression with Genetic Algorithms. This study aims to predict tourist visits, the results of this study. Based on the MAPE value test, the obtained value is 2.513% with the best parameters, namely the lambda range 1 - 10, the complexity range 1 - 100, the epsilon range 0.00001 - 0.001, the gamma range 0.00001 - 0.001, sigma range 0.01 - 3.5, SVR 1250, generation GA 90, population 70, crossover rate 0.6, mutation rate 0.4, number of features 2 and number of prediction period 1 month. And succeeded in modeling the data on foreign tourist visits to Bali according to short-term predictions. while the research conducted by [4] Decision Tree in Analyzing Lake Poso Tourist Visitor Data for Decision Making. The results obtained from this study are the number of visitors more than 28,984 having the statement "Many" which is dominated by local tourists while the value with the description "Less" is for foreign tourists. This is one of the important points in determining the right strategy to develop tourism in Poso Lake.

Lombok Island is one of the areas that is a favorite of tourists visiting both local and foreign tourists. To find out the interests of tourists in choosing tourist objects to be visited, it is necessary to analyze tourist visit data to make it easier to provide services for tourists. Based on previous research that predicts tourist visits with the Support Vector



Machine (SVM) method and analyzes the number of tourist visits using the decision tree method. This study tries to analyze the pattern of tourist visits based on the objects visited by the number of certain tourist objects. The method used in this research is the decision tree C4.5 method and the bootstrap aggregating (bagging) method. Tools used to process jupyter notebook traveler data. The results of this study will also provide recommendations to the company to provide services or offer tourist objects provided to tourists.

## II. RESEARCH METHODOLOGY

### A. Literature Review

Modeling the number of foreign tourist arrivals in Batam using arima and time series regression. [5] The purpose of this study is to model with arima and time series regression and use it on serial data and trend and seasonal data. 1. The best suitable ARIMA model is ARIMA (0,1,1) (0,1,1)<sub>12</sub>, after testing the significance and residual assumptions. 2. The appropriate time series regression model is one involving the time variable (t), the month dummy variable (January to December) and observations to (t-2) and (t-3). Tests of residual assumptions can also be met. 3. After calculating RMSE and MAPE, the best model is given by Time Series Regression.

Prediction of the Number of Foreign Tourist Visits to Bali Using Support Vector Regression with Genetic Algorithms. [3] The purpose of this study is to predict the number of foreign tourist visits to Bali using SVR with Genetic Algorithm optimization. The test results show the MAPE value obtained is 2.513% with the best parameters, namely lambda range 1-10, complexity range 1-100, epsilon range 0.00001 - 0.001, gamma range 0.00001 - 0.001, sigma range 0.01 - 3, 5, SVR 1250, generation GA 90, population 70, crossover rate 0.6, mutation rate 0.4, number of features 2 and number of prediction period 1 month. Based on the test results, the GA-SVR method on data on foreign tourist visits to Bali is suitable for short-term predictions.

Decision Tree in Analyzing Lake Poso Tourist Visitor Data for Decision Making. [4] The purpose of this study is to analyze data on tourist visits to Poso Lake by using the Decision Tree algorithm for decision making. The results of this study, visitors to Lake Poso tourism with the number of visitors more than 28,984 have a lot of information and are dominated by local tourists, while foreign tourists with more than 417 visitors and less than 1,874 still have less information.

[6] Village Classification in Gianyar Regency: Extraction and Classification of Village Potential. The purpose of this study is to identify potential natural attractions, socio-cultural potential of the community, as well as supporting facilities and infrastructure that are useful in building or developing villages in Gianyar Regency as DTW. The results of potential identification are then used as information in clustering villages as tourist attractions. The results of potential identification are then used as information in clustering villages as tourist attractions. The conclusion of this study is that there are 6 potential attributes of the village that can be used to develop the villages of Gianyar Regency as a tourist attraction. These six attributes are: (a) Village atmosphere; (b)

Uniqueness of flora & fauna; (c) village community arts; (d) The existence of temples as attractions; (e) Accessibility; and (f) Travel comfort; Three village clusters were formed based on their potential, namely a village cluster that has developed as a DTW consisting of 13 villages (Cluster I), a developing village cluster of 24 villages (Cluster II), and an undeveloped village cluster consisting of 33 villages (Cluster III); Cluster I has advantages in the uniqueness of flora & fauna, comfort, and artistic potential of the village community. Cluster II excels in the attributes of village atmosphere and temples as an attraction, while cluster III excels in accessibility to and between villages within the same cluster.

[7] Tourism Market Segmentation in Yogyakarta: Classification of Lifestyles of Domestic Tourists, the purpose of this research, is to identify the underlying dimensions of the lifestyles of local tourists to classify tourists who visit according to their typology of lifestyle, and to illustrate how to understand the various segments of local tourists. The results of this research are Metropolis Culture Aspiring Domestic tourists in Jogja in this cluster like shopping activities and usually buy something to take home when visiting Jogja. Happy to spend time off to do recreational activities and interested in visiting Jogja because of getting information from the media. Self-quality Explore Where domestic tourists in Jogja in this cluster have the opinion that if it is very thick with its customs and local wisdom, its culture must be preserved. In addition, tourists are also very optimistic about the future of tourism conditions in Jogja and believe that there will be more other tourists visiting the Jogja Aspiring Vacationer This cluster represents a segment of domestic tourists who have the opinion that education is very important and educational background does not affect one's income.

### B. Theoretical Foundation

#### 1. Data Mining

Data mining is a discipline that studies methods and extracts knowledge or finds patterns from data. Data itself is a recorded fact and has no meaning. And knowledge is a pattern, rule or model that emerges from the data, so data mining is called Knowledge Discovery in Database (KDD). [8].

Data mining is the analysis of the collection review is the analysis of the review of the data set to find unexpected relationships and summarize the data in a different way than before, which is understandable and useful for the data owner. [1] Data mining is an interactive and iterative process that involves several processes used, including knowing the type of data application, data selection, data cleaning, data integration, data reduction and transformation, data mining algorithms in selecting results interpretation development techniques and using the knowledge of the resulting process determined. [9].



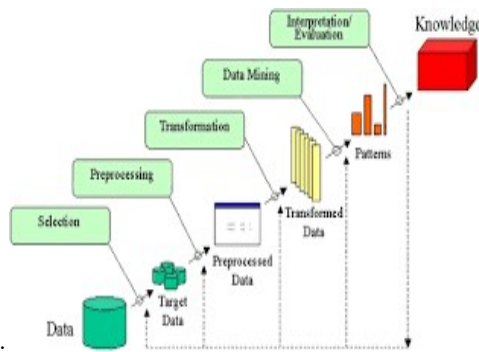


Figure 1. Knowledge Data Discovery (KDD) [10]

In the KDD process can be divided into the following [10]:

- a) Selection
- b) Preprocessing
- c) Transformation
- d) Data mining
- e) Interpretation / evaluation.

In data mining, before extracting data, the dataset used must first enter the pre-processing phase. In the pre-processing phase, the data is first normalized so that later the implementation of the algorithm can run well. Then the dataset used must be large or large so that the level of the resulting pattern is getting better. Pre-processing is very useful for analyzing multi-variate datasets, the target is determined and then cleaned first. Data cleaning removes noise and missing data [11].

## 2. Decision Tree

A decision tree is a predictive model using a tree structure or hierarchical structure. Apart from being relatively fast in construction, the results of the models built are also easy to understand, so this Decision Tree is the most popular classification method used. A decision tree is a flow-chart like tree structure, where each internal node shows a test on an attribute, each branch shows the results of the test and the leaf node shows the classes or class distribution. [12].

$$Entropy(S) = \sum_{i=1}^n - p_i * \log_2 p_i \quad (1)$$

$$Gain(S, A) = Entropy(S) - \sum_{i=1}^n \frac{|S_i|}{|S|} * Entropy(S_i) \quad (2)$$

$$SplitEntropy_A(S) = - \sum_{i=1}^n \frac{|S_i|}{|S|} * \log_2 \frac{|S_i|}{|S|} \quad (3)$$

$$GainRatio(A) = \frac{Gain(A)}{SplitEntropy(A)} \quad (4)$$

where:

- S = Case set
- A = Attribute
- n = Number of partitions on attribute A
- |S<sub>i</sub>| = Number of cases in the i-th partition
- |S| = Number of cases in S
- Ph = Proportion from S<sub>i</sub> to against S

## 3. C4.5 Algorithm

The C4.5 algorithm is a group of Decision Tree algorithms. This algorithm has input in the form of training samples and samples. Training samples in the form of sample data that will be used to build a tree that has been tested for truth. While samples are data fields that will later be used as parameters in classifying data [12].

At the learning stage, the C4.5 algorithm has 2 working principles, [12]:

- a. Decision tree creation. The purpose of the decision tree induction algorithm is to construct a tree data structure that can be used to predict the class of a new case or record that does not yet have a class. C4.5 constructs a decision tree using the divide and conquer method. At first, only root nodes were created by applying the divide and conquer algorithm. This algorithm chooses the best case solution by calculating and comparing the gain ratio, then the nodes formed at the next level, divide and conquer algorithm will be applied again until the leaves are formed.
- b. Making rules (rule set). The rules formed from the decision tree will form a condition in the form of if-then. These rules are obtained by tracing the decision tree from root to leaf. Each node and branching conditions will form a condition or an if, while for the values contained in the leaf will form a result or a then.

## 4. Bagging

Bagging is a voting method in which base-learners are differentiated from their training through a slightly different training set. Generating a sample L that is slightly different from the given sample is done by bootstrapping, when given a training set X of size N, then N random samples of X are shown with replacement [13]. Bagging



was invented by [14] which stands for “bootstrap aggregating”. [15] in his book states that bagging is a technique of the ensemble method by manipulating training data, training data is duplicated d times with sampling with replacement, which will produce d new training data, then from d training data The classifiers will be built called bagged classifiers [15].

According to [14] the stages of bagging can be considered as follows:

1. Bootstrap stages
  - a. Take n samples randomly from the training data.
  - b. After the sample is taken, arrange the best tree based on the training data.
  - c. Repeat steps a-b b times to obtain B classification trees.
2. Aggregating Stages

The aggregating stage is identical to the majority vote, namely making predictions / guesses from a combination of B fruit classification trees.

### C. Methodology

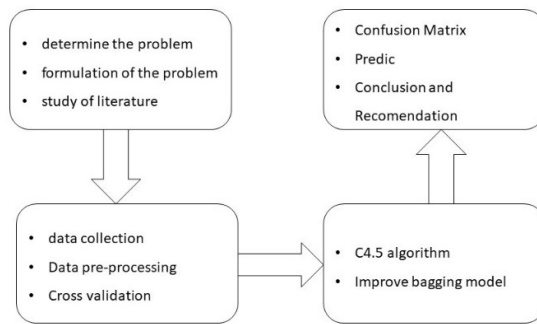


Figure 2. Research Flow

### D. Data Collection

The data used in this study is secondary data, obtained from the office of the Lombok Ceria Holiday tour and travel company. The data obtained is in the form of company daily activity data, namely data on tourist attraction visits, from 2014 to 2015. The data consists of 11 (eleven) columns namely date, name, participants (number of participants), hotels, programs (tourism objects to be visited), restaurants, transportation, guides (tour guides), remarks, expansions and invoices. Feature deletion is also carried out on features that have no effect on data modeling, such as feature names, feature deletion will be carried out.

### E. Data Pre-processing

Data cleaning aims to clean up the missing (Nan value) and or empty (null values). In the data obtained, there are several data features that are missing or null values. Several techniques are used to handle missing data, the first method is to fill in the blank or missing data with the highest value (max) of the feature, and the second method is done to the missing data by deleting features whose missing values are too high. Removed features such as restaurants, remarks, expansions and invoices. The figure table below shows the percentage of missing data for each feature.

	TOTAL DATA MISSING	persen_missing
tanggal	0	0.000000
nama	0	0.000000
Peserta	11	6.043956
Hotel	11	6.043956
Program	0	0.000000
Resto	95	52.197802
transport	31	17.032967
Guide	58	31.868132
Remark	150	82.417582
Expenses	176	96.703297
Invoice	135	74.175824

Figure 3. The percentage of missing data on each feature.

### F. Feature Breakdown

After doing data cleaning, the next step is to transport data. Before carrying out data transportation, data is separated on the date feature, in that feature there are three values namely year, month and date, so the separation is done. Data transformation is not carried out on all data features, but feature transformation is carried out on only a few features such as hotel features, program features, transport features. display data transformation as Figure below:

peserta	tanggal	bulan	tahun	hotel	program	Transport
rendah	akhir	triwulan3	2014	0.166667	0.714286	0.166667
rendah	akhir	triwulan3	2014	0.166667	0.714286	0.166667
rendah	akhir	triwulan3	2014	0.166667	0.285714	0.166667
rendah	akhir	triwulan3	2014	0.166667	0.285714	0.166667
rendah	akhir	triwulan3	2014	0.333333	0.714286	0.500000

Figure 4. Data Transformation.

### G. Data Binning

Data binning aims to group data features based on certain criteria into smaller ones. In the processed data, the features of the binned data are the features of the participants grouped into low, medium and high. The date features are grouped into early, mid and late groups, while the number features are grouped into quarter1, quarter2, quarter3 and quarter4 groups.

peserta	tanggal	bulan
rendah	akhir	trwiulan4
	awal	triwulan1
	pertengahan	triwulan3
	akhir	triwulan3
sedang	pertengahan	triwulan2
	akhir	trwiulan4
rendah	awal	triwulan2
tinggi	akhir	trwiulan4
sedang	awal	triwulan1

Figure 5. Data Binning



H. Target Labels

Determination of the target label from the classification is determined by conducting data clusters, where the target label is divided into two classes, namely the high level 0 visitor class marked with the number 0 (zero), and the 1 or low level class marked with the number 1 (one).

kelas_kunjungan	
1	139
0	126

Figure 6. Classification target label

III. RESULTS AND DISCUSSION

Classification of data is divided in two ways, namely by classification with the decision tree algorithm C4.5, the second way by using bootstrap aggregating (Bagging). data classification with the C4.5 algorithm uses three data splitting scenarios, namely :

- 1) Data training and testing 50%.
- 2) Data training 70% and data testing 30%.
- 3) Data training 80% and data testing 20%.

A. Classification With C4.5 Algorithm.

In the classification with the C4.5 algorithm, three scenarios and one to seven max depths were tried, from the results obtained, the highest value and range of accuracy values in the training data and testing data were in the 80% vs 20% splitting scenario, and in the max depth experiment. 7. The accuracy value obtained is 0.94 for training data and 0.92 for testing data.

Table 1. C4.5 Classification accuracy value

kenario splitting		Max Depth						
		1	2	3	4	5	6	7
50% :	Train	0.89	0.90	0.93	0.94	0.95	0.96	0.96
	Test	0.86	0.86	0.88	0.87	0.86	0.86	0.86
70 % :	Train	0.88	0.88	0.91	0.93	0.94	0.94	0.95
	Test	0.86	0.86	0.89	0.88	0.89	0.89	0.89
80 % :	Train	0.88	0.88	0.91	0.91	0.93	0.93	0.94
	Test	0.86	0.86	0.88	0.88	0.92	0.90	0.92

B. Confusion Matrix Algoritma C4.5

The results of the classification with the C4.5 algorithm were tested using a confusion matrix, on training data and testing data. in the training data the values that are correctly predicted (True Positive) are 94, the values that are correctly predicted are sala (True Negative) are 104, while the false positive and True negative data are each 7. While the confusion matrix on the data testing data that is predicted to be true positive is 22 and predicted is 27, while on true negative 1 and false positive is 3.

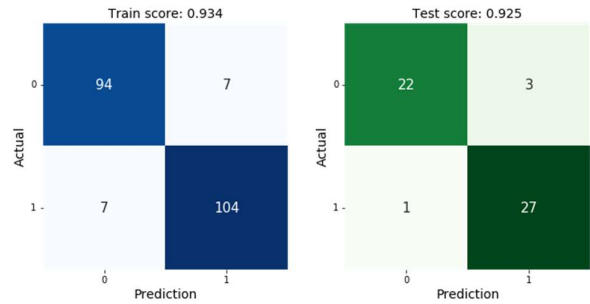


Figure 7. Confusion matrix classification

The test results using the confusion matrix, the accuracy value is 0.92 percent and the precision value is 0.96 percent and the recall value is 0.9 percent and the f1-score value is 0.93 percent.

```

=====
accuracy_score of C4.5 : 0.9245283018867925

precision_score of C4.5 : 0.9642857142857143

recall_score of C4.5 : 0.9

F1-score of C4.5 : 0.9310344827586207
=====
    
```

Figure 8. Classification accuracy value

C. Bootstrap Aggregating (Bagging)

Testing the model using the bagging method, the value of n\_estimator or the number of decision trees built seven times. Then in aggregating so that the accuracy value obtained on the training data and testing data is 0.93% and testing is 0.92%.

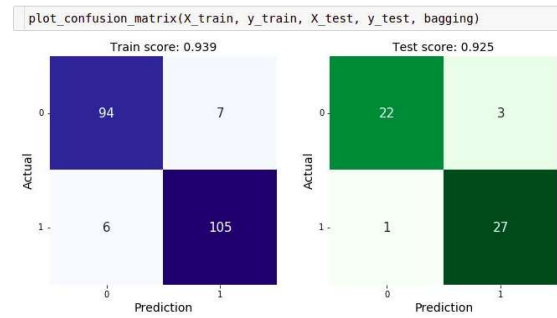


Figure 9. Confusion Matrix Model Bagging



#### D. Modeling Results

Judging from the modeling results on the month of visit data which is divided into four quarters. In class 0 visits, the highest number of visits occurred in the 1st quarter, and in class 1 visits, the highest number of visits occurred in January.

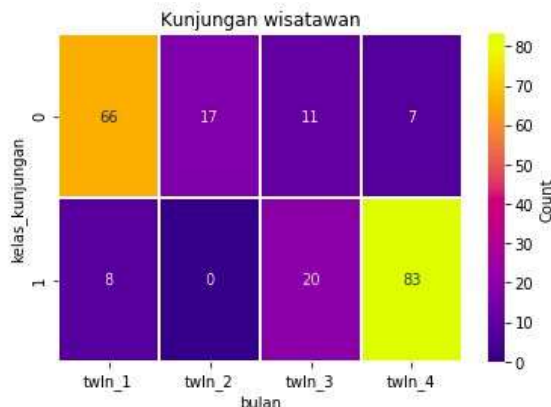


Figure 10. Quarterly visiting class

In the class of visits to tourist objects, the modeling results show that the highest number of level 0 visitors tested Gili Nanggu attractions, and the highest number of level 1 visitors visited Mandalika Kuta, Gili Nanggu and Pink Beach attractions.

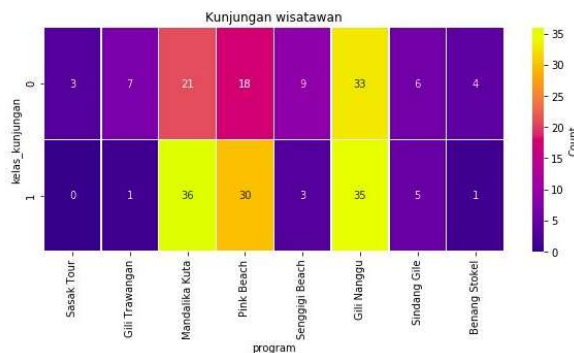


Figure 11. Tourist Attraction Visiting Class.

#### IV. CONCLUSION

##### A. Conclusion

1. From the modeling using the Decision Tree C4.5 algorithm and the scenario of splitting the data into three parts, the highest accuracy value obtained is 80:20 data splitting for train and testing data and max depth 7, which gets 94% accuracy for train data and 92% for data testing.
2. Modeling with the Bootstrap Aggregating Method, the accuracy score obtained on training data is 93% and testing data is 92. percent.
3. The accuracy results from using bagging reduce the accuracy of the C4.5 algorithm on the data training side from 94% to 93 percent, while the accuracy of testing data is still the same at 92%.
4. The prediction results for the highest class of visits are at level 0 in the first quarter where in the first quarter it is in January, February and March. while at the level

occurred in the 4th quarter, namely October, November and December.

5. The attractions with the most visitors at level 0 of the visitor class are Gili Nanggu attractions, while at level 1 tourists tend to choose Mandalika Kuta, Gili Nanggu and Pink Beach attractions.

##### B. Recommendation

In this study using data from 2014 to 2015 for further research it is recommended to use more recent data and use the ARIMA method due to the nature of the data obtained, namely time series data.

#### REFERENCES

- [1] Larose, Daniel T, "Discovering Knowledge in Data : An Introduction to Data Mining", John Willey & Sons, Inc. 2005.
- [2] Prasetyo, Eko, "Data Mining", Yogyakarta: Andi Offset. 2014.
- [3] Listiya Surtiningsih, Muhammad Tanzil Furqon, Sigit Adinugroho, "Prediksi Jumlah Kunjungan Wisatawan Mancanegara Ke Bali Menggunakan Support Vector Regression dengan Algoritma Genetika" Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer, Vol 2 No 8, 2018.
- [4] Fredryc Joshua Pa'o, Hendry Hendry, "Decision Tree dalam Menganalisis Data Pengunjung Wisata Danau Poso untuk Pengambilan Keputusan", Jurnal Sistem Komputer dan Informatika (JSON), Vol 2, No 3, 2021.
- [5] Ely Kurniawati, One Yantri, "Pemodelan Jumlah Kunjungan Wisatawan Mancanegara Di Batam Dengan Menggunakan Arima Dan Regresi Time Series", Jurnal Dimensi, Vol 7, No 3, 2018.
- [6] Mirah P Handayani, Putu Suciptawati, Trisna Darmayanti, Eka N Kencana, "Klasifikasi Desa/Kelurahan di Kabupaten Gianyar: Ekstraksi dan Klasifikasi Potensi Wisata", Jurnal Master Pariwisata (JUMPA) Volume 07, Nomor 02, 2021.
- [7] Agnessia Mega Cahyani Andri Saputri, Devilia Sari, "Segmentasi Pasar Turisme Di Yogyakarta: Klasifikasi Gaya Hidup Wisatawan Domestik", eProceedings of Management, Vol 6, No 2, 2019
- [8] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth, "From Data Mining to Knowledge Discovery in Databases", AI Magazine Volume 17 Number 3. 1996
- [9] Liao, "Recent Advances in Data Mining of Enterprise Data: Algorithms and Application", Singapore: World Scientific Publishing, 2007.
- [10] Yuli Mardi, "Data Mining: Klasifikasi Menggunakan Algoritma C4.5", Jurnal Edik Informatika Penelitian Bidang Komputer Sains dan Pendidikan Informatika, V2.i2 (213-219).
- [11] Ahmad Rofiqul Muslikh, Heru Agus Santoso, Aris Marjuni, "Klasifikasi Data Time Series Arus Lalu Lintas Jangka Pendek Menggunakan Algoritma Adaboost Dengan Random Forest", Vol. 14, No. 1, 2018.
- [12] Sunjana, "Klasifikasi Data Nasabah Sebuah Asuransi Menggunakan Algoritma C4.5, Seminar



- Nasional Aplikasi Teknologi Informasi (SNATI)", Yogyakarta, 2010.
- [13] Clancey, W.J, "Communication, Simulation, and Intelligent Agents: Implications of Personal Intelligent Machines for Medical Education", In Proceedings of the Eighth International Joint Conference on Artificial Intelligence, 556-560. Menlo Park, Calif.: International Joint Conferences on Artificial Intelligence, 1983.
- [14] Leo Beiman, "Machine Learning, Statistics Department", University of Californiaa. Berkeley, CA 94720, 1996.
- [15] Amin, N.A.S., Istadi, I, "Different Tools on Multiobjective Optimization of a Hybrid Artificial Neural Network – Genetic Algorithm for Plasma Chemical Reactor Modelling", In Olympia Roeva (Editor) Real-World Applications of Genetic Algorithms. Croatia: InTech Publisher, 2012.
- [16] Saikin Saikin, Sofiansyah Fadli, Maulana Ashari, "Optimization of Support Vector Machine Method Using Feature Selection to Improve Classification Results," *JISA (Jurnal Informatika dan Sains)*, Vol 4, No 1, 2021.





# Website System Design Using Agile Kanban Based On QR Code

Alton Gunawan Purwanto<sup>1</sup>, Ricky Yohanes Wijaya<sup>2</sup>, Timotius<sup>3</sup>, Indra Budi Trisno<sup>4</sup>

<sup>1,2,3,4</sup> Program Studi Teknik Informatika, Universitas Widya Kartika

Email: <sup>1</sup>[altongunawan27@gmail.com](mailto:altongunawan27@gmail.com), <sup>2</sup>[rickyoyohanes98@gmail.com](mailto:rickyoyohanes98@gmail.com), <sup>3</sup>[timoti.tt15@gmail.com](mailto:timoti.tt15@gmail.com), <sup>4</sup>[indrabt@gmail.com](mailto:indrabt@gmail.com)

**Abstract** – Business developments that occur today cannot be separated from the role of technology and information systems. The system that will be discussed in this journal is a Web-based Restaurant System that uses a QR code. This system will handle at least 5 types of users, namely admin, waiter, chef, cleaning service, and customers where the five kind of users have different access rights. Restaurant information systems using this QR Code can speed up performance in ordering menus, serving ordered dishes, paying bills to customers and cleaning tables after use. The use of the QR code itself is very suitable in the restaurant business and marketing practices because of its advantages, namely being mobile and easy to get information. This will improve the quality of the restaurant in terms of service and time. This Web-Based Restaurant Menu Ordering Application is designed using web programming languages, namely PHP. For the database, it will use MySQL and the system development method will use the Agile Kanban method which can help an organization produce software that has been tested and is ready to use. The results of the system design will be in the form of a comparison table for the old and new systems depicted by the PIECES table, then UML for system modeling, and website interface. The system created will help in reducing the amount of paper used for order receipts. In addition, users can also order menus during the COVID-19 pandemic with restaurants that use this system.

**Keywords** – QR Code, Restaurant, PHP, Agile Kanban, PIECES

## I. INTRODUCTION

Information technology, including computers, software, databases, internet, and other tools works specifically by giving a lot of information to various actors in various contexts. A computerized system will facilitate and assist companies in making decisions, especially being able to help various business fields in the midst of this covid-19 pandemic. The covid-19 pandemic has paralyzed all activities in various circles [1]. Covid-19, which has had many variants and has spread to almost all countries, requires changing almost all of our daily activities. The government has suggested many precautions for covid-19 to stop the spread of this virus, including social distancing, wearing masks, and avoiding crowds [2]. However, basic human needs such as food and drink cannot be replaced. The covid-19 pandemic has made people who want to visit restaurants to be more careful because it is considered as a place which has high percentage risk of the covid-19 virus can be transmitted [3]. In addition, the restaurant system which is still manual by means of the waiter approaching the visitor then writing a message on a piece of paper and delivering the customer's order to the chef makes the service take a long time, miscommunication between the buyer and the waiter and can cause crowds of visitors.

Based on these problems, to improve efficiency in the restaurant system, menu serving management, and resources are very important matters. How to increase efficiency in order to provide services to customers, so that they can produce good restaurant services [4]. Problem identification includes:

1. There are government precautions to avoid crowds which can facilitate the spread of the covid-19 virus.
2. The recording process is still manual which causes time efficiency.
3. There are still orders errors due to miscommunication between waiters and visitors

To overcome this problem, a web-based restaurant system was created using a QR code [5]. This research is limited by the limitations of the scope and management of data boundaries such as menu data, visitor data, payment data, ordering data, menu data, payment data, and table status data.

QR code is a two-dimensional barcode that contains written and printed data in a more concise medium. QR codes were first introduced by the Japanese company Denso-Wave in 1994 [6]. QR stands for Quick Response as it aims to allow its contents to be encoded at high speed. QR code is capable of storing all types of data, such as numeric, alphanumeric, binary, and kanji/kana [7]. The way the QR code works itself is first the pattern of the QR code is taken using a cellphone camera (HP) or another scanner that is able to translate the QR code. Then the pattern on the QR code is decoded using special software that can read the information stored in the QR code pattern. The QR code system consists of an encoder and a decoder. The encoder is responsible for encoding the data and generating the QR code, while the decoder decodes the data from the QR code [8]. The QR code will be integrated with the table and directly connected to the ordering system.

The program planning that is made will be displayed in a UML diagram and the framework used is the kanban method. UML is a predefined modeling language for



software development. The UML approach uses an object-oriented approach. UML consists of 14 diagrams, which are used in this system, including Class Diagrams, Use Case Diagrams, and Sequence Diagrams [9].

Web based means the program to be made based on the website using the PHP programming languages as the main components. Which the main component is PHP as a Server-Side programming language. PHP is an acronym for Hypertext Preprocessor, an open source server-side scripting based programming language. Then the script from PHP will be processed on the server-side [10]. The database used is MySQL which is the most frequently used query programming language. MySQL is one type of database that is open source and is widely used to build web-based applications as a source of data processing. SQL (Structured Query Language) is the standard language used to access database servers. In the 70s this language was developed by IBM, which was then followed by Oracle, Informix, and Sybase. By using SQL, the process of accessing the database becomes easy [11]. By implementing the two things above, it is expected to create an orderly, neat, and smooth flow/process of work and the programs made can be more neat and structured. And finally, a program cannot be considered feasible if it has not received testing. The testing method used is black box testing.

Black Box is a testing method used to test an application/software without having to pay attention to the details of the application/software. Black Box testing only checks the output value based on the input value. Black Box testing process is carried out by trying the application/software and entering data on each form which aims to find out whether the application/software is running as desired [12].

In developing this system software, Agile Kanban methodology is used. This method is part of the SDLC development that can track the progress of each task being done and can limit the amount of work being done to provide a time limit for completing a task. Components in Kanban use boards and cards, the board is an environment of kanban containing various cards with tasks that need to be completed [13].

The web program is considered very suitable to be applied in the current era of technology, and also because of the restrictions due to the increasingly widespread Covid-19 pandemic. The development of this application will certainly improve the services provided by restaurants because currently each restaurant continues to strive to improve quality by providing the best service and one of them is by utilizing information technology to support the restaurant business [14].

The research topic written can be concluded that the purpose of this research by developing a Web-Based Restaurant System program using a QR Code is to improve restaurant service for customers, facilitate ordering by customers, reduce the spread of covid-19, and facilitate the management of restaurant activities. The results of the research are expected to be implemented in a restaurant so that the restaurant will benefit from the system built, namely the convenience between customers and employees.

## II. RESEARCH METHODOLOGY

In developing a restaurant system website, the PIECES analysis method is used, which is a technique to identify and solve problems that occur in information systems. PIECES consists of performance, information analysis, economic analysis, security analysis, efficiency analysis, and service [15]. To produce quality software in a short time and at a low cost using a development process called the System Development Life Cycle (SDLC). The development stage in the SDLC can help an organization to be able to produce software that has been tested and is ready for use. The SDLC design method on this system uses the Agile method with the kanban framework. This Agile Kanban method is used to track progress and visualized using a kanban board, resulting in transparency on each member's work. Agile Kanban is more effective in managing software development. Compared to agile scrum, agile kanban does not have to wait as long as a sprint when a task has been completed. Kanban board has many forms of stages but the main ones are backlog, in-progress, and complete. Kanban implementation can create a more structured workflow [13] [16].

There are 6 stages used in this project management, and these stages are described as follows,

### 1. Backlog

In the backlog, all tasks will be broken down into more detailed tasks, which will be performed and sorted according to different priority levels of work. All tasks related to this system will be collected in the backlog.

### 2. To Do

After all the ideas are collected in the backlog, the tasks will be moved to the to do section. In this section, it is determined that all tasks are of high quality and will be carried out during development.

### 3. In Progress

If there is a task to be done, then the task is moved from the to do stage to the in progress stage. At this stage, the work on the task is being carried out in order of priority. To increase efficiency and streamline workflows due to many unfinished tasks, it is necessary to set boundaries for work in progress, to focus the team on completing tasks systematically.

### 4. Testing

If there is a task that needs to be tested, the task will be moved from the In progress stage to the Testing stage. This stage is optional, testing at this stage is carried out by the user and the team is also responsible for monitoring the progress of the test and following up on any problems that arise.

### 5. Feedback

At this Feedback stage, the user provides input on deficiencies or problems with the tasks that have been tested by the user during the Testing stage.

### 6. Done

A task is considered complete and will be entered into the Done stage if the results of the task have met the requirements and development standards,



the functional test has passed, and there are no errors in the task.

### III. RESULTS AND DISCUSSION

The following are the results of the analysis based on the research method we used,

#### 1. PIECES Table

The comparison between the old system and the system made is shown in table 1 using the PIECES analysis indicator table.

Table 1. PIECES Analysis System Comparison

Component Analysis	Old System	New System
Performance	Menu ordering is still done manually	Menu ordering has been done online through the restaurant application
Information	Menu information for visitors can only be obtained if the waiter approaches the visitor	Menu information, and menu availability can be seen through the restaurant application.
Economy	The use of paper for receipts from ordering menus by visitors can be a waste of money.	The cost to buy paper is no longer needed because now visitors can view the order history independently through the restaurant application.
Control	In the process of ordering the menu, it is the waiter who is in charge of passing it on to the chef.	Menus that have been ordered by customers will automatically appear in the chef's application so that it can save time.
Efficiency	The resources needed are very large because they are still doing manual data collection, wasting time, and having many waiters to serve many tables	All activities carried out by customers are carried out independently, and can reduce unnecessary resources.
Service	In terms of service, visitors have to wait a while for the order to be forwarded to the chef.	In terms of service, visitors make menu selections and confirm orders, the orders will be directly forwarded to the chef.

#### 2. Use Case Diagram

This section will discuss the interactions between actors such as customers, waiters, admins, chefs, and cleaning services with the system described by the use case diagram in Figure 1.

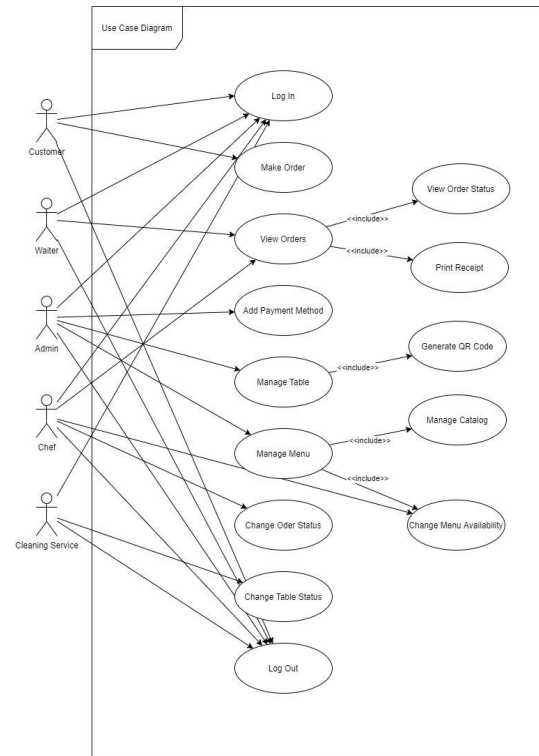


Figure. 1 Restaurant System Use Case Diagram

#### 3. Sequence Diagram

This section will discuss the sequence diagram used by this system, there are 9 sequence diagrams shown in Figures 2-10, namely logging in, logging out, viewing orders, making orders, managing tables, managing menus, managing payment methods, managing table statuses, and change the order status.

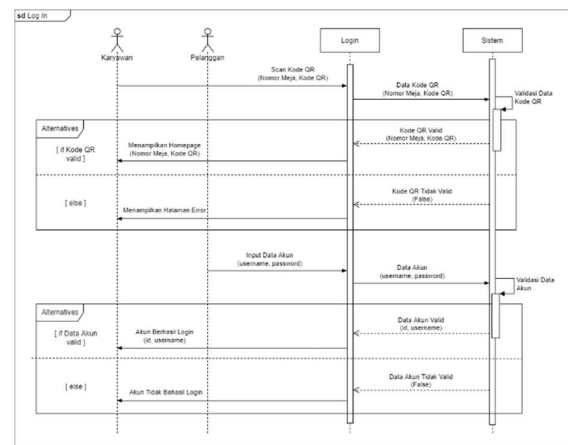


Figure. 2 Sequence Diagram Log In



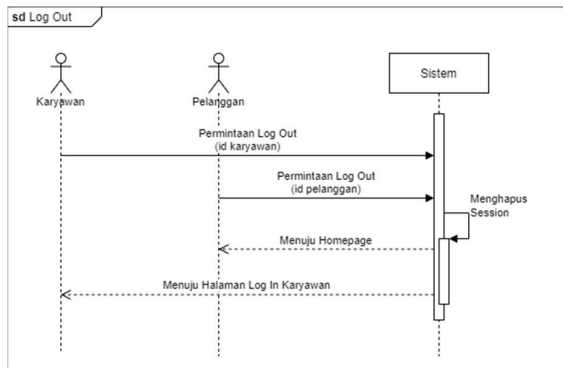


Figure. 3 Sequence Diagram Log Out

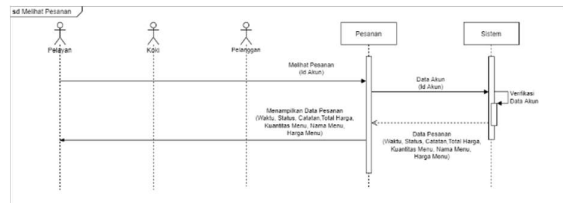


Figure. 4 Sequence Diagram View Orders

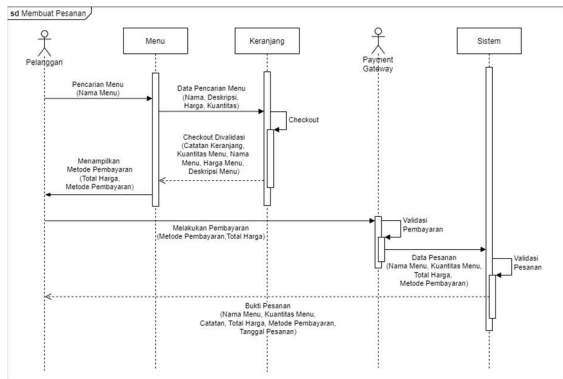


Figure. 5 Sequence Diagram Make an Order

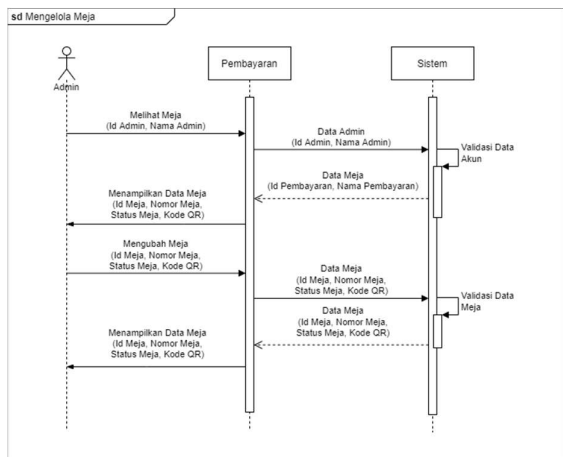


Figure. 6 Sequence Diagram Managing Table

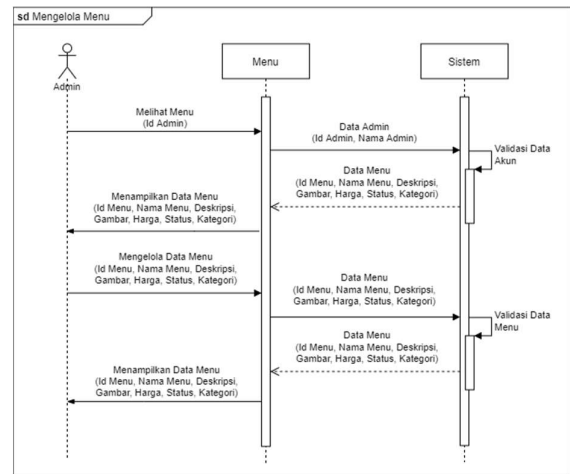


Figure. 7 Sequence Diagram Manage Menu

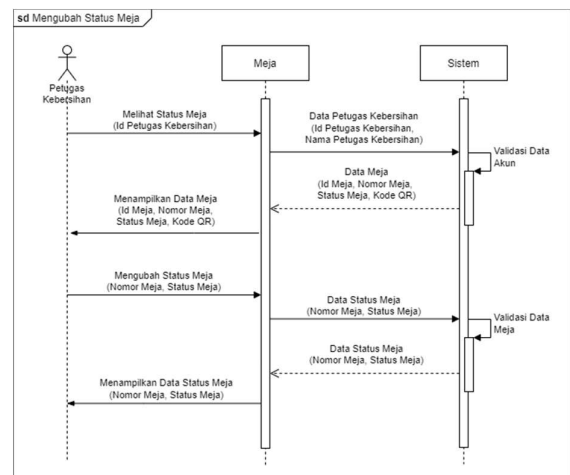


Figure. 8 Sequence Diagram Manage Payment Methods

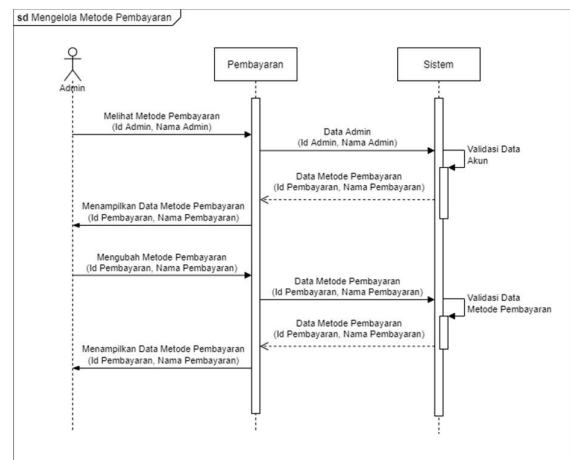


Figure. 9 Sequence Diagram Managing Table Status

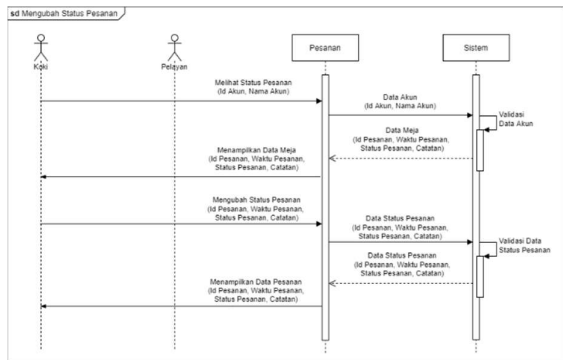


Figure. 10 Sequence Diagram Change Order Status

#### 4. Class Diagram

There are 12 classes in this system, namely Payment, Chef, Waiter, Customer, Basket, Basket Items, Orders, Order Details, Menu, Admin, Cleaning, and Table classes. The interactions between classes are shown in Figure 11 while the properties of the classes are shown in Figure 12.

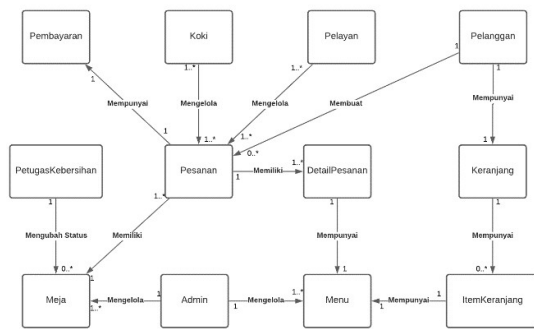


Figure. 11 Restaurant System Class Diagram



Figure. 12 Properties of Restaurant System Class Diagram

#### 5. Website Interface

In the website interface, it will display the appearance of the website that has been designed according to the UML Diagram created. The



following website interface has been included below,

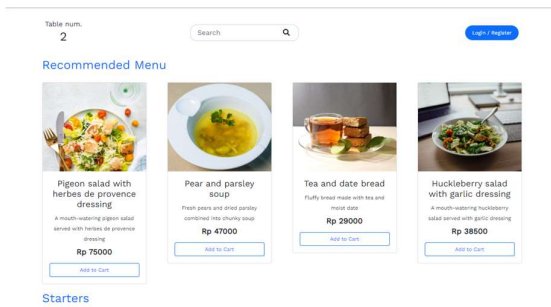


Figure. 13 Customer Homepage

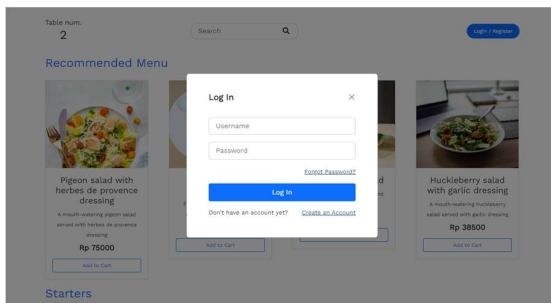


Figure. 14 Customer Log In

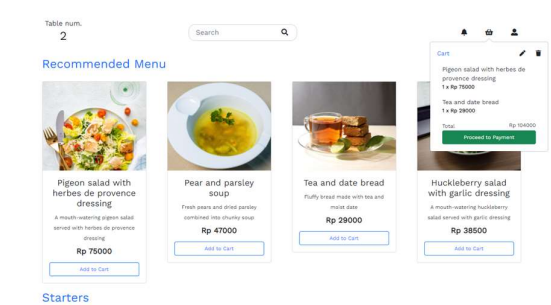


Figure. 15 Customer Cart

Figure 13 is the initial menu display, where customers can register an account and enter an existing account as shown in Figure 14 to be able to place an order by entering the menu in the restaurant into the basket in Figure 15.

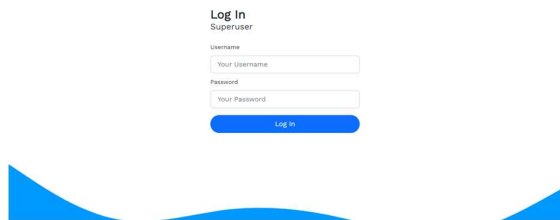


Figure. 16 Superuser Log In

Figure 16 is the initial menu for administrators, waiters, chefs, and cleaning services to be able to enter the system according to the username and password that has been registered by the administrator.

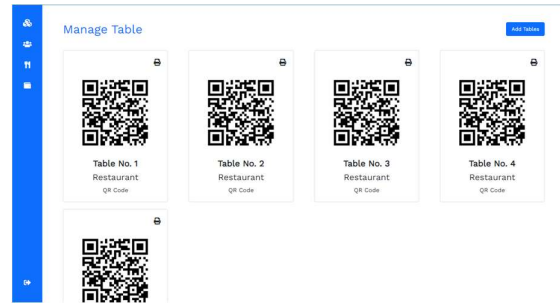


Figure. 17 Admin Menu - Manage Table

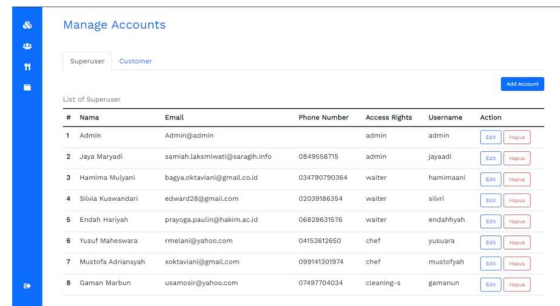


Figure. 18 Admin Menu - Manage Accounts

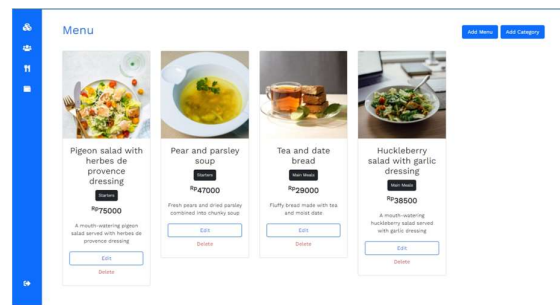


Figure. 19 Admin Menu - Manage Menu

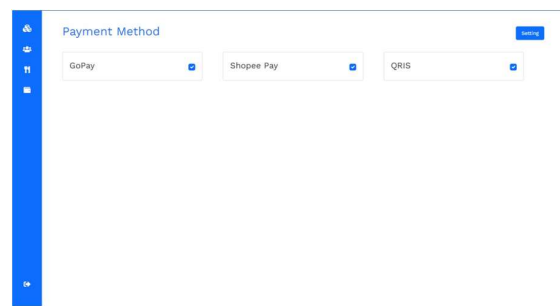


Figure. 20 Admin Menu - Manage Payment Method

Figures 17-20 is the menu that will be displayed after successfully logging in for an account with administrator privileges. In this menu, administrators can manage all accounts, manage menus, add payment methods, manage categories, and manage tables.



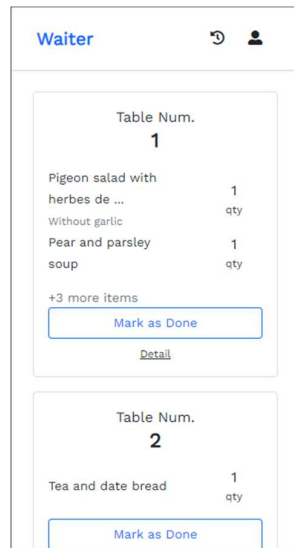


Figure. 21 Waiter Homepage

Figure 21 is a menu that will be displayed after successfully logging in for an account with waiter privileges. In this menu, the waiter can view order details and change the order status.

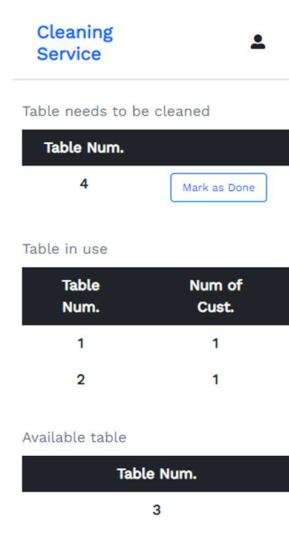


Figure. 23 Cleaning Service Dashboard

Figure 23 is a menu that will be displayed after successfully logging in for an account with access rights as a cleaning service. In this menu, the cleaning service can only change the usage status on each table.

To be able to test the results of this website system, a black box test was carried out whose results are shown in table 2 below. The black box test results on 4 system pages for customers, admins, waiters, chefs, and cleaning services get results that pass for each feature tested. Based on the research conducted, the PIECES analysis table, UML, and website interface have been designed.

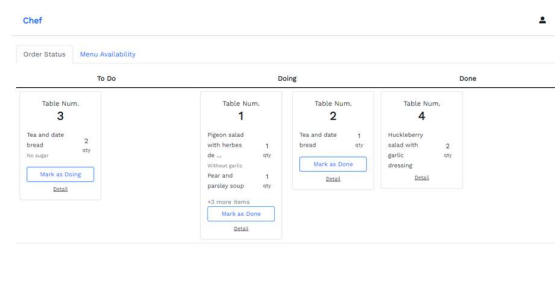


Figure. 22 Chef Homepage

Figure 22 is the menu that will be displayed after successfully logging in for an account with chef access rights. In this menu, chefs can view orders, adjust menu availability, change menu status, and change order status.

Table 2. Black Box Testing Result

No	Test Case	Expected Output	Actual Output	Status
<b>A Customer Page System</b>				
1	Create Account	Customer can create accounts	Customer can successfully create accounts	pass
2	Scan QR Code	Customer can scan QR Code to go to customer homepage	Customer can successfully scan QR Code to go to customer homepage	pass
3	Log In	Customer can log in according to account data	Customer success log in according to account data	pass
4	Search Menu	Customers can search the menu in the search input field accordingly	Customers successfully search the menu in the search input field accordingly	pass
5	Add Menu to Cart	Customers can choose the desired menu and collect it to cart	Customers can successfully choose the desired menu and collect it to cart	pass
6	Make Order	Customers can choose and make orders according to the menu presented on the customer homepage	Customers can successfully choose and make orders according to the menu presented on the customer homepage	pass
7	Receive Payment	Customer gets a notification	Customer successfully gets a	pass



	Notifications	regarding the success or failure of the payment made	notification regarding the success or failure of the payment made	
8	Log Out	Customer can logged out from system	Customer successfully logged out from system	pass
9	View Order History	Customers can see details and history of transactions that have been made	Customers can see details and history of transactions that have been made	pass
<b>B Admin Page System</b>				
1	Log In	Admin can log in according to account data	Admin success log in according to account data	pass
2	Manage Payment Method	Admin can manage payment method	Admin can successfully manage payment method	pass
3	Manage Table	Admin can manage table	Admin can successfully manage table	pass
4	Manage Menu	Admin can manage menu and catalog	Admin can successfully manage menu and catalog	pass
5	Log Out	Admin can logged out from system	Admin successfully logged out from system	pass
6	Print QR Code	Admin can print QR code for each table	Admin can successfully print QR code for each table	pass
<b>C Waiter Page System</b>				
1	Log In	Waiter success log in according to account data	Waiter success log in according to account data	pass
2	View Order	Waiter can view and change the order status	Waiter can view and change the order status	pass
3	Log Out	Waiter can logged out from system	Waiter successfully logged out from system	pass
<b>D Chef Page System</b>				
1	Log In	Chef success log in according to account data	Chef success log in according to account data	pass
2	View Order	Chef can view the order status	Chef can successfully view the order status	pass
3	Change Order status	Chef can change order status	Chef can successfully change order status	pass
4	Adjust menu availability	Chef can adjust menu availability	Chef can successfully adjust menu availability	pass
5	Log Out	Chef can logged out from system	Chef successfully logged out from system	pass
<b>E Cleaning Service Page System</b>				
1	Log In	Cleaning Service can log in according to account data	Cleaning Service success log in according to account data	pass

2	View Table Status	Cleaning service can see table usage status	Cleaning service can successfully see table usage status	pass
3	Change Table Status	Cleaning Service can change the usage status on each table	Cleaning Service can change the usage status on each table	pass
4	Log Out	Cleaning Service can logged out from system	Cleaning Service successfully logged out from system	pass

#### IV. CONCLUSION

Based on the results of the development of a restaurant system that was made according to the needs of the community during the covid-19 pandemic as well as the results and discussions above, it can be concluded that the system that has been created has been in accordance with the research objectives, and with the creation of this system can make the restaurant system more structured and organized. The use of this web-based system can also help reduce paper usage. On the other hand, customers can be more comfortable in ordering menus during the covid-19 pandemic with restaurants that use this system. The author's suggestions for this system to be able to help other developers maximize the performance of this system, by making a financial reporting system to be able to describe the financial condition of the restaurant.

#### REFERENCES

- [1] H. N. Saturwa, S. Suharno, and A. A. Ahmad, "The impact of Covid-19 pandemic on MSMEs," *J. Ekon. dan Bisnis*, vol. 24, no. 1, pp. 65–82, Mar. 2021, doi: 10.24914/JEB.V24I1.3905.
- [2] W. Zentrato, "Gerakan Mencegah Daripada Mengobati Terhadap Pandemi Covid-19," *J. Educ. Dev.*, vol. 8, no. 2, pp. 242–248, May 2020, Accessed: Nov. 13, 2021. [Online]. Available: <http://journal.ipts.ac.id/index.php/ED/article/view/1689>
- [3] The Lancet Respiratory Medicine, "COVID-19 transmission—up in the air," *Lancet Respir. Med.*, vol. 8, no. 12, p. 1159, Dec. 2020, doi: 10.1016/S2213-2600(20)30514-2.
- [4] J. Caroline El Fiorenza, A. Chakraborty, R. Rishi, and K. Baghel, "Smart Menu Card System," *Proc. 3rd Int. Conf. Commun. Electron. Syst. ICCES 2018*, pp. 847–849, Oct. 2018, doi: 10.1109/CESYS.2018.8724045.
- [5] F. FAISAL and M. A. F. ANAS, "PEMANFAATAN KODE QR PADA PENINGKATAN PELAYANAN DAN KEPUASAN PELANGGAN PADA RESTORAN," *J. INSTEK (Informatika Sains dan Teknol.*, vol. 5, no. 1, pp. 111–120, Mar. 2020, doi: 10.24252/INSTEK.V5I1.14504.
- [6] A. Priyambodo, K. Usman, and L. Novamizanti, "Implementation of Android-Based Qr Code in the Presence System," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 7, no. 5, pp. 1011–1020, 2020, doi: 10.25126/jtiik.202072337.





- [7] E. F. Nurdiansyah and I. Afrianto, "IMPLEMENTASI QR CODE SEBAGAI TIKET MASUK EVENT DENGAN MEMPERHITUNGGAN TINGKAT KOREKSI KESALAHAN," *J. Teknol. dan Inf.*, vol. 7, no. 2, pp. 25–44, Sep. 2017, doi: 10.34010/JATILV7I2.491.
- [8] S. Tiwari, "An Introduction to QR Code Technology," pp. 39–44, Jul. 2017, doi: 10.1109/ICIT.2016.021.
- [9] T. A. Kurniawan, "Pemodelan Use Case (UML): Evaluasi Terhadap beberapa Kesalahan dalam Praktik," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 5, no. 1, pp. 77–86, Mar. 2018, doi: 10.25126/jtiik.201851610.
- [10] A. O. Ameen, M. A. Aremu, M. Olagunji, J. B. Awotunde, and O. T. Olowe, "Design and Development of a Safe and Effective Online Marketplace For Nigeria," *GOUni J. Manag. Soc. Sci.*, vol. 5, no. 1, pp. 80–89, 2017.
- [11] D. S. R. M. Ninik Sri Lestari1, "Perancangan Aplikasi Pembuatan Kartu Keluarga Berbasis Web Menggunakan Php Dan Mysql," *Isu Teknol. Sst Mandala*, vol. 15, no. 2, pp. 1–13, 2020.
- [12] F. C. Ningrum, D. Suherman, S. Aryanti, H. A. Prasetya, and A. Saifudin, "Pengujian Black Box pada Aplikasi Sistem Seleksi Sales Terbaik Menggunakan Teknik Equivalence Partitions," *J. Inform. Univ. Pamulang*, vol. 4, no. 4, pp. 125–130, Dec. 2019.
- [13] A. S. Vidianto and W. H. Haji, "Sistem Informasi Manajemen Proyek Berbasis Kanban (Studi Kasus: PT. XYZ)," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 7, no. 2, pp. 283–292, 2020, doi: 10.25126/jtiik.2020701676.
- [14] F. Nurlaila, "Aplikasi Pemesanan Makanan pada Restoran 1953 Indonesia Berbasis Web," *J. Inform. Univ. Pamulang*, vol. 4, no. 1, p. 16, Mar. 2019, doi: 10.32493/INFORMATIKA.V4I1.2585.
- [15] N. L. Kakihary, "Pieces Framework for Analysis of User Saticfaction Internet of Things-Based Devices," *J. Inf. Syst. Informatics*, vol. 3, no. 2, pp. 243–252, Jun. 2021, doi: 10.33557/JOURNALISI.V3I2.119.
- [16] J. Petersen, "MEAN Web Application Development with Agile Kanban," *West. Oregon Univ. Digit. Commons*, 2016.



# Grouping of Village Status in West Java Province Using the Manhattan, Euclidean and Chebyshev Methods on the K-Mean Algorithm

Gatot Tri Pranoto<sup>1\*)</sup>, Wahyu Hadikristanto<sup>2</sup>, Yoga Religia<sup>3</sup>

<sup>1</sup>Informatics Engineering Study Program, Faculty of Engineering, Pelita Bangsa University

<sup>2</sup>Informatics Engineering Study Program, Faculty of Engineering, Pelita Bangsa University

<sup>3</sup>Informatics Engineering Study Program, Faculty of Engineering, Pelita Bangsa University

email: <sup>1</sup>[gatot.pranoto@pelitabangsa.co.id](mailto:gatot.pranoto@pelitabangsa.co.id), <sup>3</sup>[wahyu.hadikristanto@pelitabangsa.ac.id](mailto:wahyu.hadikristanto@pelitabangsa.ac.id), <sup>2</sup>[yoga.religia@pelitabangsa.ac.id](mailto:yoga.religia@pelitabangsa.ac.id)

**Abstract** – Village Potential Data in 2014 (Podes 2014) in West Java Province is data released by the Central Statistics Agency in collaboration with the Ministry of Villages PDPT in unsupervised form and consists of 5319 village data. The 2014 Podes data in West Java Province is based on the level of village development (village specific) in Indonesia, by using the village as the unit of analysis. However, village funds have not been distributed effectively and accurately in accordance with the conditions and potential of the village due to the lack of clear information about the status of the village. So that there are no priority villages that should receive greater funds and attention from the government. One of the algorithms that can be used for the clustering process is the k-means algorithm. Grouping data using k-means is done by calculating the shortest distance from a data point to a centroid. In this study, a comparison of the distance calculation methods on k-means between Manhattan, Euclidean and Chebyshev will be carried out. Tests will be performed using execution time and davies bouldin index. From this test, the 2014 Village Potential data in West Java Province has been grouped into 5 village statuses by obtaining the number of villages for each cluster, namely cluster as many as 694 villages, cluster as many as 567 villages, cluster as many as 1440 villages, cluster is 1557 villages and cluster is 1061 villages. For distance calculation, Chebyshev has the most efficient accumulation of time compared to Manhattan and Euclidean. Meanwhile, the Euclidean method has the Davies Index compared to the Manhattan and Chebyshev methods.

**Keywords** – Village Development, k-means, Manhattan, Euclidean, Chebyshev, Davies Bouldin index.

## I. INTRODUCTION

The Republic of Indonesia is one of the countries in Southeast Asia which is located between two continents (Asia and Australia) and two oceans (India and the Pacific) which makes Indonesia also known as the Archipelago (Intermediate Archipelago). Indonesia is the largest archipelagic country in the world consisting of 17,504 islands. With a population of 222 million in 2006, Indonesia is the country with the fourth largest population in the world. Indonesia consists of various ethnic groups, languages and different religions. The total area of Indonesia is 1,913,578 km<sup>2</sup>, and has the second largest biodiversity in the world [10].

The large number of people and the rapid development in cities have an impact on the Indonesian economy. So that the development makes people come to the city to get a job and settle down. This phenomenon is also known as urbanization. Urbanization in Indonesia causes many problems, the impact on the village is the reduction in human resources because the population moves to cities which causes villages in Indonesia to not experience real development. The existence of urbanization is triggered by development facilities between rural and urban areas [14].

The Ministry of Villages, Development of Disadvantaged Regions and Transmigration of the Republic of Indonesia (Kemendes) is a ministry within the Government of Indonesia in charge of developing rural areas and rural areas, empowering rural communities, accelerating development of disadvantaged areas, and

transmigration. The Ministry of Villages is under and responsible to the President. Based on Presidential Regulation Number 18 of 2020 concerning the 2020-2024 National Medium-Term Development Plan (RPJMN), the 2020-2024 RPJMN is a strategic document that contains development plans that must be carried out by the government for the next five years. The 2020-2024 RPJMN is used as an official reference for local governments and other stakeholders in implementing development. The RPJMN 2020-2024 was raised with the aim of mapping rural conditions in Indonesia based on their level of development, setting development targets/targets in the next five years that must be achieved by village development actors, and photographing the performance of village development that has been carried out. In the 2020-2024 RPJMN for the area and spatial planning of the sub-sector of village development and rural areas, it contains village development targets that must be achieved in the next five years, namely reducing the number of underdeveloped villages to 5,000 villages and increasing the number of independent villages to at least 2,000 villages in 2024. Realizing this requires mapping the status of the village to be built.

The Ministry of Village based on the Regulation of the Minister of Villages, Development of Disadvantaged Regions, and Transmigration of the Republic of Indonesia number 2 of 2016 concerning the index of developing villages, states that the status of villages is categorized into 5 categories, namely independent villages, developed villages, developing villages, underdeveloped villages and

very underdeveloped villages with the explanation as follows: following:

1) Independent Village or the so-called Sembada Village is an Advanced Village that has the ability to carry out village development to improve the quality of life and life as much as possible for the welfare of the Village community with social resilience, economic resilience, and ecological resilience in a sustainable manner. Independent Village or Madya Village is a Village that has a Developing Village Index greater ( $>$ ) than 0.8155.

2) Advanced Village or the so-called Pre-Sufficient Village is a Village that has the potential of social, economic and ecological resources, as well as the ability to manage them to improve the welfare of the Village community, quality of human life, and alleviating poverty. Advanced Village or Pre-Madya Village is a Village that has a Developing Village Index less and equal to ( $\leq$ ) 0.8155 and greater ( $>$ ) than 0.7072.

3) Developing Village or what is called Madya Village is a potential village to become an advanced village, which has the potential of social, economic and ecological resources but has not managed them optimally for improving the welfare of the village community, quality of human life and alleviating poverty. Developing Village or Middle Village is a Village that has a Developing Village Index less and equal to ( $\leq$ ) 0.7072 and greater ( $>$ ) than 0.5989.

4) Disadvantaged Villages or so-called Pre-Madya Villages are villages that have potential social, economic, and ecological resources but have not, or have not managed them in an effort to improve the welfare of the Village community, the quality of human life and experience poverty in its various forms. Disadvantaged Villages or Pre-Madya Villages are Villages that have a Developing Village Index less and equal to ( $\leq$ ) 0.5989 and greater ( $>$ ) than 0.4907.

5) Very Disadvantaged Villages or so-called Pratama Villages are villages that experience vulnerability due to natural disasters, economic shocks, and social conflicts so that they are not able to manage the potential of social, economic and ecological resources, and experience poverty in various forms. Very Disadvantaged Villages or Pratama Villages are Villages that have a Developing Village Index less and smaller ( $\leq$ ) than 0.4907.

Village status is in fact inseparable from village development that uses funds from the government. In the Law of the Republic of Indonesia number 6 of 2015 concerning villages, article 86 paragraph 2 states that the Government and Regional Governments are obliged to develop a village information system and development of rural areas. The Ministry of Villages in collaboration with the National Development Planning Agency and the Central Statistics Agency issued data on village potential (Podes) in West Java Province which has 5,319 instances and consists of 42 dependent attributes without labels. Podes data is a measurement method that is compiled based on the level of village development in Indonesia by making the village the unit of analysis. Podes data is compiled with reference to Law Number 6 of 2014 concerning Villages,

which is intended to capture the level of village development in West Java Province and can be used as a reference for policy planning and village development supervision [19].

In information technology, data is an important part that cannot be separated from information retrieval. Data mining is a series of activities used to find new, hidden or unexpected patterns contained in the data. The term data mining is often considered as a synonym for knowledge discovery from data (KDD), namely the discovery of knowledge from data that focuses on the purpose of the mining process [21]. Data mining can be used to perform clustering, classification and association. Clustering is a grouping process that is carried out by finding similar characteristics between data according to certain class groups [8]. In simple terms, clustering can be used to analyze a set of data and generate a set of clustering rules that can be used to group future data.

In the real world sometimes data is not only grouped into status binary (binary class), but needs also to be grouped into the multi-status (multi-class). In the case of multi-class data-sets, grouping will be more difficult than in the case of classes binary. K-mean is an interactive clustering algorithm that partitions the data-set into number of K clusters a predetermined. Ever done a comparative study of clustering, the partition-based clustering hierarchical based and clustering. density-based. In this study, it was revealed that k-mean which is a partition-based algorithm provides better performance and k-mean is superior for large/lots of data compared to clustering hierarchical and-based density [6]. In addition, several other studies also mention that clustering using the k-means algorithm is faster than algorithms clustering other and also produces quality clusters when using large datasets [7] [20] [12].

The k-mean algorithm has been used for multi-class grouping which shows effective results and is able to improve performance classification [1]. In another case, two algorithms were compared, namely k-mean and Density-Based Spatial Clustering of Applications with Noise (DBSCAN) which said that the two algorithms are algorithms clustering the most commonly used for grouping data with different criteria, and the test results show that k-mean algorithm is better than DBSCAN in time efficiency analysis [5]. In addition, in research on the comparison of algorithms clustering data mining comparing k-mean, Density based clustering and Hierarchical Clusterings concluded that the k-mean algorithm is superior in time efficiency compared to the other two algorithms and also k-mean is able to distribute cluster instances quite accurately [17]. When the k-mean algorithm performs data grouping, the k-mean algorithm will calculate the closest distance between a data set to appoint centroid. The calculation of the distance on the k-mean algorithm can use Manhattan, Euclidean and Chebyshev. Each method of calculating distance is superior to one another depending on the dataset used [3] [18] [15].

The calculation of the distance in the k-means algorithm can use Manhattan, Euclidean and Chebyshev. Awasthi in his research on the comparison of the Manhattan and



I10	Availability and ease of access to midwives' practices		
I11	Availability and convenience to access to poskesdes or polindes		
I12	Availability and ease of access to pharmacies		
I13	Availability of shops, minimarkets, or grocery stores	Economic	Infrastructure Condition of Infrastructure
I14	Availability of markets		
I15	Availability of restaurants, restaurants or food stalls/shops		
I16	Availability of hotel accommodation or lodging		
I17	Availability bank aan		
I18	Electrification	of Energy Infrastructure	
I19	Lighting conditions on main roads		
I20	Fuel for cooking		
I21	Sources of drinking water	Health and Sanitation	Infrastructure
I22	Sources of water for bathing/washing		
I23	Defecation facilities		
I24	Availability and quality of cellular communication facilities	Communication and Information	Infrastructure
I25	Availability of internet facilities and postal or goods delivery		
I26	Traffic and road quality	Means of Transportation	Accessibility / Transportation
I27	Road accessibility		
I28	Availability of public transport		
I29	transport operations		
I30	Travel time per kilometer of transportation to the sub-district office	Transportation accessibility	
I31	Cost per kilometer of transportation to the sub-district office		
I32	Travel time per kilometer of transportation to the regent		
I33	Cost per kilometer of transportation to the regent		
I34	Handling of extraordinary events	Public Health Public	Services
I35	Handling of malnutrition		

I36	Availability of sports facilities	Sports	
I37	Existence of k groups Sports		
I38	Completeness of village government	Independence	of Government Administration
I39	Village autonomy		
I40	assets/wealth		
I41	Quality of human resources of village head	Quality of Human Resources	
I42	Quality of human resources of village secretary		

### C. Sample Selection Methods

Data samples random taken from the original data is data 2014 Village Potential in West Java province as many as 15 villages and 42 indicators initialized as I1 to I42 to be grouped using the k-means algorithm, the data obtained are shown in Table 3.

Table 3. Table of Indicators for Compiling Podes 2014

x	Nama Desa	I1	I2	I3	I4	I5	...	...	I4 1	I4 2
1	CARIU	3.0	3.0	4.0	4.0	3.0			4.0	3.0
		0.0	0.0	0.0	0.0	0.0			0.0	0.0
2	JABON MEKAR	3.0	3.0	4.0	4.0	3.0			5.0	5.0
		0.0	0.0	0.0	0.0	0.0			0.0	0.0
3	CIHIDEUNG HILIR	2.0	4.0	4.0	4.0	3.0			4.0	3.0
		0.0	0.0	0.0	0.0	0.0			0.0	0.0
4	CISOLOK	2.0	3.0	4.0	4.0	3.0			4.0	3.0
		0.0	0.0	0.0	0.0	0.0			0.0	0.0
5	KADUPANDAK	3.0	4.0	5.0	5.0	0.0			4.0	5.0
		0.0	0.0	0.0	0.0	0.0			0.0	0.0
6	BANJARAN KULON	4.0	4.0	4.0	4.0	3.0			5.0	3.0
		0.0	0.0	0.0	0.0	0.0			0.0	0.0
7	JAGABAYA	4.0	3.0	4.0	3.0	3.0			4.0	3.0
		0.0	0.0	0.0	0.0	0.0			0.0	0.0
8	DEPOK	3.0	4.0	4.0	2.0	3.0			0.0	3.0
		0.0	0.0	0.0	0.0	0.0			0.0	0.0
9	LINGGAJAYA	2.0	3.0	3.0	4.0	1.0			4.0	3.0
		0.0	0.0	0.0	0.0	0.0			0.0	0.0
10	KRAMATMULYA	3.0	3.0	3.0	3.0	3.0			4.0	3.0
		0.0	0.0	0.0	0.0	0.0			0.0	0.0
11	PABUARAN WETAN	3.0	3.0	4.0	3.0	3.0			4.0	5.0
		0.0	0.0	0.0	0.0	0.0			0.0	0.0
12	KADIPATEN	3.0	4.0	4.0	3.0	3.0			4.0	0.0
		0.0	0.0	0.0	0.0	0.0			0.0	0.0
13	JATIBARANG	4.0	4.0	5.0	4.0	3.0			4.0	5.0
		0.0	0.0	0.0	0.0	0.0			0.0	0.0
14	KASOMALANG WETAN	4.0	4.0	5.0	5.0	3.0			4.0	5.0
		0.0	0.0	0.0	0.0	0.0			0.0	0.0
15	DAWUAN TIMUR	4.0	3.0	2.0	3.0	3.0			4.0	3.0
		0.0	0.0	0.0	0.0	0.0			0.0	0.0

From table 3 will be used for grouping into 5 clusters initialized as C1 SD C5. The grouping will be done using the k-means algorithm with three different distance calculation methods, namely Manhattan, Euclidean and Chebyshev.

### D. Evaluation

In this study, the evaluation will be carried out by

grouping the village potential data. The grouping will use the k-mean algorithm where the distance calculation will use Manhattan, Euclidean and Chebyshev. The results of the grouping obtained are the grouping of village potential data into 5 clusters, namely cluster 0, cluster 1, cluster 2, cluster 3 and cluster 4. Until this stage it is not a known cluster which can be called a cluster independent village, developed village, developing village, underdeveloped villages and very underdeveloped villages.

In the village potential data, each attribute/indicator has a value of 0 to 5, where a value of 0 is the lowest value while a value of 5 is the highest value. From each cluster obtained has value centroid, where the centroid is the "midpoint" of the cluster. So to determine the status of the village, we can calculate the number of centroids for each cluster, which can be written with the equation:

$$Status\ desa = \sum_{i=1}^8 CI_{i1}, CI_{i2}, \dots, CI_{i42}$$

From equation 8, CI is the centroid of each indicator and each cluster has 42 indicators. Determination of village status will be sorted based on the sum of the values centroid of each indicator in each cluster, where the lowest sum value will be initialized as very underdeveloped village status and the highest sum value will be initialized as independent village status.

#### E. Validation

In this study, validation will be carried out to test the distance calculation method on which k-mean algorithm is most effectively used for grouping village potential data. The test will be carried out using tools RapidMinerto obtain the accumulated time and the value of the Bouldin Index for each distance calculation method used. The best time efficiency is the one that has the minimum accumulated time. Meanwhile, by using the Davies Bouldin Index, a cluster will be considered to have scheme clustering an optimal which has a Davies Bouldin Index minimal.

### III. RESULTS AND DISCUSSION

#### A. Testing the proposed

Method Testing the Manhattan, Euclidean and Chebyshev distance calculation methods on the k-means algorithm used to group the 2014 Village Potential data in West Java Province will be carried out using latest model clustering that is validated using execution time and Davies Bouldin Index.

##### 1. Manhattan Distance

From the use of the k-means algorithm with the Manhattan Calculation method distance to group the 2014 Podes data in West Java Province which amounted to 5319 villages, the number of villages from each obtained cluster was as follows:

- Cluster 0: 1,005 villages
- Cluster 1: 1,189 villages
- Cluster 2: 1,084 villages
- Cluster 3: 447 villages
- Cluster 4: 1,594 villages

When viewed from the number of centroids calculated by the equation above and the number of villages in each cluster, the status of the village can be obtained from the k-

means grouping using the Manhattan Calculation method distance as shown in Table 4.

Table 4. Status and Number of Villages Using Manhattan

Cluster	Jumlah Centroid	Status Desa	Jumlah Desa
Cluster 0	3,12	Tertinggal	1.005 desa
Cluster 1	3,35	Maju	1.189 desa
Cluster 2	3,73	Mandiri	1.084 desa
Cluster 3	2,75	Sangat Tertinggal	447 desa
Cluster 4	3,34	Berkembang	1.594 desa

##### 2. Euclidean Distance

From the use of the k-means algorithm with the Euclidean Calculation method distance to group the 2014 Podes data in West Java Province which amounted to 5,319 villages, the number of villages from each obtained cluster was as follows:

- Cluster 0: 1,061 villages
- Cluster 1: 567 villages
- Cluster 2: 1,557 villages
- Cluster 3: 1,440 villages
- Cluster 4: 694 villages

When viewed from the number of centroids calculated by the equation above and the number of villages in each cluster, the status of the village can be obtained from the k-means grouping using the Euclidean Calculation method distance as shown in Table 5.

Table 5. Status and Number of Villages Using Euclidean

Cluster	Jumlah Centroid	Status Desa	Jumlah Desa
Cluster 0	3,71	Mandiri	1.061 desa
Cluster 1	3,19	Tertinggal	567 desa
Cluster 2	3,42	Maju	1.557 desa
Cluster 3	3,24	Berkembang	1.440 desa
Cluster 4	2,87	Sangat Tertinggal	694 desa

##### 3. Chebyshev Distance

From the use of the k-means algorithm with the Chebyshev Calculation method distance to group the 2014 Podes data in West Java Province, amounting to 5319 villages, the number of villages from each obtained cluster is as follows:

- Cluster 0: 1,366 villages
- Cluster 1: 1,004 villages
- Cluster 2: 306 villages
- Cluster 3: 1,214 villages
- Cluster 4: 1,429 villages

When viewed from the number of centroids calculated by the equation above and the number of villages in each cluster, the status of the village can be obtained from the k-means grouping using the Chebyshev Calculation method distance as shown in Table 6.

Table 6. Status and Number of Villages Using Chebyshev

Cluster	Jumlah Centroid	Status Desa	Jumlah Desa
Cluster 0	3,53	Mandiri	1.366 desa
Cluster 1	3,04	Sangat Tertinggal	1.004 desa
Cluster 2	3,42	Berkembang	306 desa
Cluster 3	3,45	Maju	1.214 desa
Cluster 4	3,23	Tertinggal	1.429 desa

### B. Testing the proposed

Accumulation of time is carried out by executing 5 times for each distance calculation method used. The 5 executions will then be averaged to obtain the most efficient execution time for each distance calculation method. From the tests that have been carried out, it is obtained that the length of execution time is different, as for the length of execution time from the test of the Manhattan, Euclidean and Chebyshev distance calculation methods that have been carried out, it can be seen in Figure 1.

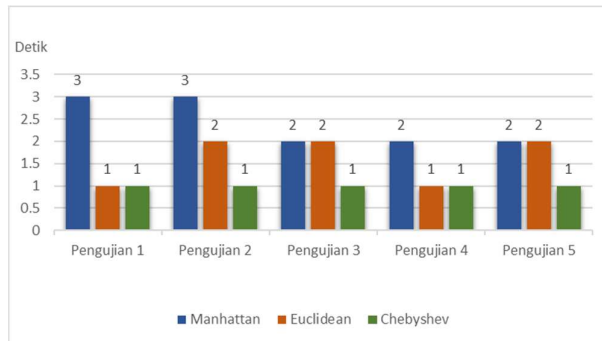


Figure 1. Execution Time

In Figure 1. It can be seen that the execution time of the method Manhattan distance to test 1 to test 5 in a row is 3 seconds, 3 seconds, 2 seconds, 2 seconds and 2 seconds, so when taken on average execution time of a Manhattan distance is 2.4 seconds. Meanwhile, the execution time of the Euclidean Method distance for testing 1 to 5, respectively, is 1 second, 2 seconds, 2 seconds, 1 second and 2 seconds, so that the average execution time of the Euclidean distance is 1.6 seconds. Then the execution time of the Chebyshev Method distance for testing 1 to 5, respectively, namely 1 second, 1 second, 1 second, 1 second and 1 second, so that when taken the average execution time of Chebyshev distance is 1 second. The more easily the execution time required for the Manhattan, Euclidean and Chebyshev methods can be seen in Table 7.

Table 7. Old Time Execution

Pengujian	Waktu Eksekusi		
	Manhatta n	Euclidean	Chebyshev
1	3 detik	1 detik	1 detik
2	3 detik	2 detik	1 detik
3	2 detik	2 detik	1 detik
4	2 detik	1 detik	1 detik
5	2 detik	2 detik	1 detik
Rata-rata	<b>2,4 detik</b>	<b>1,6 detik</b>	<b>1 detik</b>

### C. Testing the Davies Bouldin Index

In this study, the Davies Bouldin Index (DBI) was used to validate the data in each cluster. Measurement using DBI aims to maximize the distance inter-cluster. By using DBI A cluster will be considered to have scheme clustering an optimal if it has a Davies Index minimal. As for the tests that have been carried out, the values obtained Davies Index from the Manhattan, Euclidean and Chebyshev methods are shown in Figure 2.

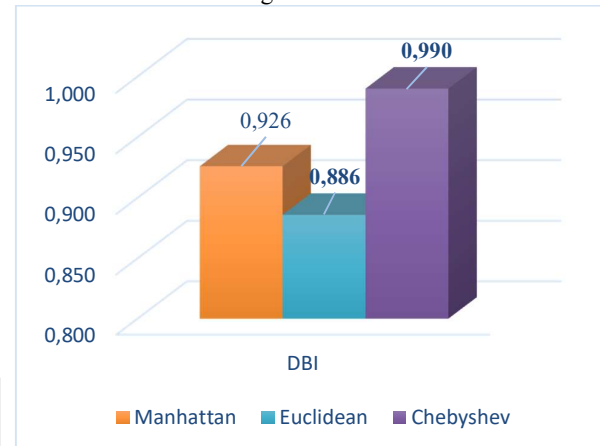


Figure 2. Davies Index of Manhattan, Euclidean and Chebyshev methods

From Figure 2, it can be seen that the value *Davies Index* from the Manhattan method is 0.926, the value *Davies Index* from the Euclidean method is 0.886 and the value *Davies Index* from the Chebyshev method is 0.990. As for the easier value of the *Davies Index* from the Manhattan, Euclidean and Chebyshev methods, it can be seen in Table 8.

Table 8. Status and Number of Villages Using Chebyshev

Davies Bouldin Index		
Manhattan	Euclidean	Chebyshev
0,926	0,886	0,990

From Table 8 it can be seen that the most optimal value of the Manhattan, Euclidean and Chebyshev methods is the Euclidean Method distance with the value Davies Index of 0.886.

### D. Analysis of Test Results

From testing the 2014 Village Potential data grouping method in West Java Province using the k-means algorithm with the distance calculation methods Manhattan, Euclidean and Chebyshev that have been carried out, the results are:

1. The test model used can run well and shows the results in the form of values centroid for each cluster from the methods Manhattan, Euclidean and Chebyshev, so that the status of the village can be determined from the number of centroids in each cluster.
2. The use of the distance calculation method used affects the amount of data in each cluster.
3. The accumulated time obtained from the tests that have been carried out shows that the distance calculation method Chebyshev has the most efficient execution time with an average accumulation time of 1 second.

4. By using the test, the Davies Bouldin Index shows that the distance calculation method Euclidean has the Davies Index most optimal value with a value of 0.886.

From the tests that have been carried out, it can be seen that the 2014 Village Potential data grouping in West Java Province using the k-means algorithm with the distance calculation method Chebyshev has the most efficient accumulation of time compared to Manhattan and Euclidean, while the Euclidean method has the value Davies Index most optimal compared to the method. Manhattan and Chebyshev. So when viewed from the quality of the cluster based on the Davies Index, the obtained cluster status of the village is from the k-means algorithm with the distance calculation method Euclidean as follows:

- Cluster Very Disadvantaged Village As many as 694 villages
- Cluster Disadvantaged Village As many as 567 villages
- Cluster Developing Village Of 1,440 villages
- Cluster Advanced Village as many as 1,557 villages
- Cluster Independent Village of 1,061 villages

#### IV. CONCLUSION

From the discussion and evaluation in the previous chapters, the 2014 Village Potential data grouping in West Java Province into 5 groups using the k-means algorithm with the Manhattan, Euclidean and Chebyshev distance calculation methods, the conclusions are:

1. The 2014 Village Potential data in West Java Province has been grouped into 5 village statuses with the obtained number of villages for each cluster, namely cluster as many as 694 villages, cluster as many as 567 villages, cluster as many as 1440 villages, cluster as many as 1557 villages and cluster as many as 1061 villages.
2. The grouping of Village Potential data in 2014 in West Java Province into 5 village statuses using the k-means algorithm with the Chebyshev distance calculation method has the most efficient accumulation of time compared to Manhattan and Euclidean, while the Euclidean method has the Davies Index most optimal

#### REFERENCES

- [1] Al-robby, M. F., & El-halees, A. M. (2013). Classifying Multi-Class Imbalance Data. 37(5), 74–81.
- [2] Amandeep Kaur Mann, N. K. (2013). Review Paper on Clustering Techniques. Global Journal of Computer Science and Technology.
- [3] Awasthi, R., Tiwari, A. K., & Pathak, S. (2013). Empirical Evaluation On K Means Clustering With Effect Of Distance Functions For Bank Dataset. International Journal of Innovative Technology and Research, 1(3), 233–235.
- [4] Mishra, B. K., Rath, A., Nayak, N. R., & Swain, S. (2012). Far efficient K-means clustering algorithm. ACM International Conference Proceeding Series. <https://doi.org/10.1145/2345396.2345414>
- [5] Chakraborty, S., Nagwani, N. K., & Dey, L. (2011). Performance Comparison of Incremental K-means and Incremental DBSCAN Algorithms. International Journal of Computer Applications. <https://doi.org/10.5120/3346-4611>
- [6] Chaudhari, B., & Parikh, M. (2012). A Comparative Study of Clustering Algorithms using Weka Tools. International Journal of Application or Innovation in Engineering and Management (IJAIEM).
- [7] Claypo, N., & Jaiyen, S. (2015). Opinion mining for Thai restaurant reviews using K-Means clustering and MRF feature selection. 2015 7th International Conference on Knowledge and Smart Technology (KST), 105–108. <https://doi.org/10.1109/KST.2015.7051469>
- [8] Deepa, V. K., Remy, J., & Geetha, R. (2013). Rapid development of applications in data mining. 2013 International Conference on Green High Performance Computing, ICGHPC 2013. <https://doi.org/10.1109/ICGHPC.2013.6533916>
- [9] Ding, S., Wu, F., Qian, J., Jia, H., & Jin, F. (2015). Research on data stream clustering algorithms. Artificial Intelligence Review. <https://doi.org/10.1007/s10462-013-9398-7>
- [10] Direktorat Jenderal Pemerintahan Umum, K. D. N. (n.d.). <https://www.bps.go.id/statictable/2014/09/05/1366/luas-daerah-dan-jumlah-pulau-menurut-provinsi-2002-2016.html>.
- [11] Gandhi, G., & Srivastava, R. (2014). Review Paper: A Comparative Study on Partitioning Techniques of Clustering Algorithms. International Journal of Computer Applications, 87(9), 10–13. <https://doi.org/10.5120/15235-3770>
- [12] Ghosh, S., & Kumar, S. (2013). Comparative Analysis of K-Means and Fuzzy C-Means Algorithms. International Journal of Advanced Computer Science and Applications. <https://doi.org/10.14569/ijaacs.2013.040406>
- [13] Grabusts, P. (2011). The choice of metrics for clustering algorithms. Vide. Tehnologija. Resursi - Environment, Technology, Resources. <https://doi.org/10.17770/etr2011vol2.973>
- [14] Harahap, F. R. (2013). Dampak Urbanisasi Bagi Perkembangan Kota Di Indonesia. Society, 1(1), 35–45. <https://doi.org/10.33019/society.v1i1.40>
- [15] Kouser, K., & Sunita, S. (2013). A comparative study of K Means Algorithm by Different Distance Measures. International Journal of Innovative Research in Computer and Communication Engineering.
- [16] KumarSagar, H., & Sharma, V. (2014). Error Evaluation on K-Means and Hierarchical Clustering with Effect of Distance Functions for Iris Dataset. International Journal of Computer Applications. <https://doi.org/10.5120/15066-3429>
- [17] Pratap, S., Kushwah, S., Rawat, K., & Gupta, P. (2012). Analysis and Comparison of Efficient Techniques of Clustering Algorithms in Data Mining. 3, 109–113.
- [18] Singh, A., Yadav, A., & Rana, A. (2013). K-means with Three different Distance Metrics. International Journal of Computer Applications. <https://doi.org/10.5120/11430-6785>
- [19] Soleh, A. (2017). Strategi Pengembangan Potensi Desa. Jurnal Sungkai, 5(1), 35–52.
- [20] Verma, M., Srivastava, M., Chack, N., Diswar, A. K., & Gupta, N. (2012). A Comparative Study of Various Clustering Algorithms in Data Mining. International Journal of Engineering Research and Applications Wwww.ljera.Com.
- [21] Xu, L., Jiang, C., Wang, J., Yuan, J., & Ren, Y. (2014). Information security in big data: Privacy and data mining. IEEE Access. <https://doi.org/10.1109/ACCESS.2014.2362522>



# Application of Data Mining to Determine Promotion Strategy Using Algorithm Clustering at SMK Yadika 1

Jerry Watulangkouw<sup>1\*)</sup>

<sup>1</sup>Program Studi Magister Ilmu Komputer, Fakultas Teknologi Informasi, Universitas Budi Luhur  
email: [jerrywatulangkouw@gmail.com](mailto:jerrywatulangkouw@gmail.com)

**Abstract** – The Promotion Strategy is very important to achieve the desired target, in determining the School Promotion Strategy for the results of new student admissions and in recommending the right promotion that can be used to overcome the problems faced by SMK Yadika which experienced a decrease in the number of new students from the 2017/2018 class entering 267 new student, then experiencing difficulties in determining promotion strategies, and promotion decisions taken by the school are sometimes not right on target, even though the position of SMK Yadika has a very strategic environment or place that can produce and get a lot of students. This study aims to apply the K-Means algorithm in the Promotion Strategy grouping which produces seven clusters based on the K Optimal Davies Bouldin Index so that it can be used to determine the right promotion strategy and develop an information system prototype to assist schools in compiling and deciding the right promotion. The results of this research, schools can carry out promotions based on the origin of the student's school, promotions based on the field of study of interest, promotions based on the study program expertise, promotions based on competency skills, and promotions based on the district where the student lives or domicile. With the results of clustering using the K-Means methodology, Cluster 1 (17.71%), cluster 2 (32.67%), cluster 3 (10.43%), cluster 4 (5.7%), cluster 5 (4.55 %), cluster 6 (3.34%), and cluster 7 (25.78%).

**Keywords** – Data Mining, Promotion Strategy, Clustering, Davies Bouldin Index, CRIPS-DM

## I. INTRODUCTION

Determining the right Promotion Strategy at this time can outperform the competition between schools, so promotional strategy efforts for an educational institution are needed to be able to get new students and students. Schools as educational service providers need to improve themselves and learn to have the initiative to increase customer satisfaction, in this case prospective students and students. Therefore, a promotion strategy, especially in the field of education, is needed to win a competition between schools so as to increase the interest of prospective new students or students to see the strategic position of the school that can produce a lot of students and students, and also to increase the acceleration of quality improvement and school management professionalism. Competition between vocational education institutions actually provides benefits for customers in this case prospective students and students, this is because customers have many choices in deciding which schools are suitable for their prospective students and students.

In Previous Research Asril et al (2015) with research on graduate data analysis with data mining to support the Promotion strategy of Lancang Kuningan University, the problem obtained The number of private universities and high schools that have been established requires Lancang Kuning University to carry out a series of various active promotions, so that no less competitive with other universities, in this study using the K-Means Clustering Algorithm Method with the attributes of Address, Name of

Region, and GPA while the results obtained are four clusters are formed. Rony (2016) with research on the Application of Data Mining to Use the Clustering Algorithm to Determine New Student Promotion Strategies at the LP3I Jakarta Polytechnic the problems obtained With the very large amount of data the LP3I Polytechnic has difficulty getting identification of prospective students who register at the LP3I Polytechnic Jakarta, In This study uses the K-Means Clustering Algorithm Method, with the attribute of residence, while the results obtained form 4 clusters

Promotional strategies currently running in schools tend to pick up or come to nearby junior high schools by distributing brochures, presenting school profiles, putting up banners, organizing events or competitions in junior high schools, and finally promoting through social media. Then every year Yadika 1 Vocational School has a diverse number of students, such as in the 2017/2018 semester academic year Odd there are 267 new students, in 2018/2019 Odd the number of new students is 263 students, while in 2019 semester / 2020 Odd The number of new students is 231 students. By looking at the decreasing number of students per semester, a strategy is needed to promote Yadika 1 Vocational School.

## II. RESEARCH METHODOLOGY

"Data mining is the analysis of data to find clear relationships and conclude that they were not previously known in a way that is currently understood and useful for the owner of the data".

From some of the opinions above, it can be concluded that Data Mining is a technique used to explore and find a previously unknown relationship from a large and large number of data so that information is needed and can be used later.

The terms data mining and knowledge discovery in databases (KDD) are often used interchangeably to describe the process of extracting hidden information in a large database. And the stages in this data mining are like the KDD process which can be broadly explained as follows han jiawei in Sumangkut et al [1]:

#### 1. Data Selection

Selection of new data sets of operational data needs to be done before the stage of extracting information in data mining begins. The selected data that will be used for the data mining process is stored in a file, separate from the operational database.

#### 2. Pre-processing/Cleaning

Before the data mining process can be carried out, it is necessary to carry out a cleaning process on the data that is the focus of KDD. The cleaning process includes, among others, removing duplicate data, checking for inconsistent data, and correcting errors in data, such as typographical errors.

#### 3. Transformation

Coding is a transformation process on the data that has been selected, so that the data is suitable for the data mining process.

#### 4. Data Mining

Data mining is the process of looking for interesting patterns or information in selected data using certain techniques or methods.

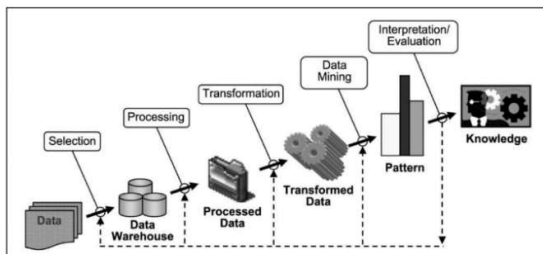
#### 5. Interpretation/Evaluation

The pattern of information generated from the data mining process needs to be displayed in a form that is easily understood by interested parties.

#### 6. Knowledge

*The last stage of the data mining process is how to formulate decisions or actions from the results of the analysis obtained*

Furthermore, Budiman [2] describes the Knowledge Discovery in Database Stages as shown in Figure 2.



According to Badrul [3] data mining is divided into several groups based on the tasks that can be done, which are as follows:

#### 1. Description

Sometimes researchers and analysts simply want to try to find ways to describe the patterns and trends contained in the data. For example, polling officers may not be able to find information or facts that those who are not professional enough will have little support in the presidential election. Descriptions of patterns and trends often provide possible explanations for a pattern or trend.

#### 2. Estimation

Estimation is almost the same as classification, except that the estimation target variable is more numerical than categorical. The model is built using a complete record that provides the value of the target variable as the predicted value. Furthermore, in the next review, the estimated value of the target variable is made based on the value of the predictive variable.

#### 3. Prediction

Prediction is almost the same as classification and estimation, except that in predicting the value of the results will be in the future. Some of the methods and techniques used in classification and estimation can also be used (for appropriate circumstances) for prediction.

#### 4. Classification

In classification, there are categorical variable targets. For example, income classification can be separated into three categories, namely high income, medium income, and low income.

#### 5. Clustering

Clustering is grouping records, observing, or paying attention and forming classes of objects that have similarities. Cluster is a collection of records that have similarities with one another and have dissimilarities with records in other clusters.

#### 6. Association

The task of association in data mining is to find attributes that appear at one time. In the business world it is more commonly called shopping cart analysis.

Several previous studies that have the same object of study regarding educational assistance are summarized as a study review in this paper. The research in question is:

1. This research was conducted by Asril, Wiza and Yunefri (2015) to determine the promotion strategy of Lancang Kuningan University. The full title is Graduate Data Analyst with data mining to support the promotion strategy of Lancang Kuningan University. The problem faced in this research is the number of private universities and high schools that have been established requiring Lancang Kuning University to

- carry out a series of various active promotions, so as not to lose to compete with other universities. The right promotion strategy uses the K-means Clustering Algorithm [4]
2. Research conducted by Priambudi (2015) to determine product sales strategy. The full title of this research is PT Mayora's Product Sales Strategy using the Apriori method and Data Mining Implementation. The problem in this research is how to make an application that can assist the development of schools in determining marketing strategies by utilizing transaction data. The process of implementing the right sales strategy is carried out through the Apriori Method and the implementation of data mining [5].
  3. Research conducted by Rony (2016) to determine new student promotion strategies. The full title of the research is Application of Data Mining to Use Clustering Algorithm to Determine New Student Promotion Strategy at LP3I Polytechnic Jakarta. The process of applying the right promotion strategy is done through the K-Means Clustering algorithm, starting with a random selection which is the number of clusters that you want to form. Then set the K values randomly, temporarily the value becomes the center of the cluster or commonly called the centroid [6].
  4. This research was conducted by Suprawoto (2016) to support the selection of marketing strategies. Title Classification of Student Data Using the K-Means Method to support the selection of marketing strategies. The problem faced is the amount of data that has accumulated from year to year needs to be analyzed to be able to open up opportunities to produce useful information in making alternative decisions for higher education management. The process of implementing the right promotion strategy is done through the K-Means Clustering algorithm [7].
  5. This research was conducted by Wirta and Erlin (2016) to choose a promotion strategy for new student admissions. The full title is Implementation of the K-Means Cluster Analysis Method to choose a new student admission promotion strategy. The process of applying the right promotion strategy is done through the K-Means Clustering algorithm starting with a random selection which is the number of clusters that you want to form [8].
  6. This research was conducted by Mochammad C. A. (2018) to determine the promotion strategy at the Baitussalam Intensive Vocational School Tanjunganom Nganjuk. The full title of this research is Data Mining using the K-Means algorithm to determine the promotion strategy at Baitussalam Intensive Vocational School Tanjunganom Nganjuk. The process of applying the right promotion strategy is done through the K-Means Clustering algorithm, starting with a random selection which is the number of clusters that you want to form. Then assign K values randomly [9].
  7. This research was conducted by Achyani (2018) for the optimization of direct marketing predictions. The full title is There are classification and regression problems with linear or nonlinear kernels which can be an ability of classification learning algorithms. There are classification and regression problems with linear or nonlinear kernels which can be an ability of the classification learning algorithm. The process of applying the Particle Swarm Optimization method for direct marketing prediction optimization [10].
  8. This research was conducted by Kusumo, Sedyono and Marwata (2019) to support higher education promotion strategies. The full title is Apriori Algorithm Analysis to Support Higher Education Promotion Strategy. The problem faced is that universities are currently required to have a competitive advantage by utilizing all available resources (Azimah & Sucahyo, 2007). The high level of competition between educational institutions has resulted in each institution having to be able to manage its institution professionally. The process of implementing the right promotional strategy information support is carried out through the Apriori Algorithm Method [11].
  9. This research was conducted by Jaini (2019) for grouping sales and product marketing strategies. The problems faced in conducting promotions are based on products that are experiencing a trend in society. In addition, new products that are approaching the expiration date will be promoted. It affects the effectiveness of marketing. The process of implementing the right sales support and product marketing strategy is carried out through the Fuzzy C-Means and K-Medoids Algorithm Methods [12].
  10. This research was conducted by Takdirillah (2020) on transaction data to support sales strategy information. Full Title Implementation of Data Mining Using Apriori Algorithm Against Transaction Data to Support Sales Strategy Information. Problems regarding stockpiling that can harm retail store entrepreneurs are quite common. The process of implementing the right sales strategy information support is carried out through the Apriori Algorithm Method [13].

### III. RESEARCH METHODOLOGY

#### A. Research Methods

In this study, the methodology used is CRISP-DM to analyze and process data. The research stages are divided into several stages as shown in Figure 1.



separation between clusters, which can be seen in equation 3.1.

$$inter = \min \{ \|m_k - m_{kk}\| \} \quad \forall k = 1, 2, \dots, K - 1 \text{ dan } k = k + 1, \dots, K \dots \dots \dots \quad (1)$$

Then the term intra is used to measure the cohesiveness of a group. The standard deviation is used to check the proximity of the data points of each cluster which can be seen in equation 3.2.

$$\sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - X_m)^2} \quad (2)$$

The cluster that will be analyzed produces a pattern of adjacent item sets from the results of student data analysis which will be used as item recommendations for the promotion strategy of Yadika 1 Vocational High School, West Jakarta City.

b. Design Technique

After conducting the analysis, then proceed with system design based on the problem analysis that has been carried out, namely:

- 1) The first stage is analyzing business understanding and preprocessing data using the CRISP DM method, clustering student data using the K-Means method, calculating the K-Means method testing by calculating the Centroid value and Davies Bouldin Index to get product results for strategy recommendations promotion.
- 2) The second stage is to design system interactions with actors using Use Case Diagrams and Use Case Diagrams.
- 3) The third stage is to design a database using streamlit as a library and also to upload the dataset and to download the results.
- 4) The fourth stage is designing the system interface (user interface).
- 5) The fifth stage is designing applications using the Python programming language.

c. Testing technique

After doing the design, then proceed with the system testing technique based on the research design that has been done, namely :

- 1) The technique of testing the results of clustering is by looking at the results of the Davies Bouldin Index, if the purity results are close to 1, it indicates the better the cluster formed..
- 2) The testing technique in the development of this information system is using black boc testing is a test that is carried out only observing the results of execution through test data and checking the functionality of the software..

C. Research Steps

Based on the research method, sample selection method, data collection method, analysis technique, design and testing of food data, research steps were formed which can be seen in Figure 2..

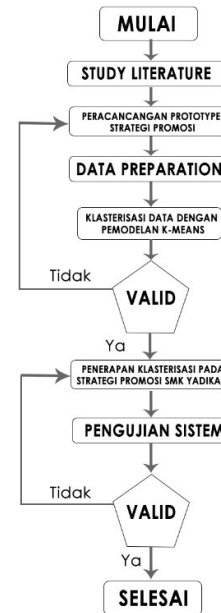


Figure 2. Research Stages

Description of Research Steps:

1. Start
2. Study Literature  
This stage is a series of activities related to the method of collecting library data, reading and recording and managing research materials
3. This stage is the process of designing a display system design. Promotional strategies are made as examples to develop products as an illustration for direct userslangung
4. Data Preparation  
This stage includes all activities to build the final dataset (data to be processed at the modeling stage) from raw data. This stage can be repeated several times. This stage also includes the selection of tables, records, and data attributes, including the process of cleaning and transforming data to be used as input in the modeling stage.
5. Data Clustering with K-Means Modeling  
At this stage, the selection and application of various modeling techniques will be carried out and several parameters will be adjusted to obtain optimal values.

In particular, there are several different techniques that can be applied to the same data mining problem. On the other hand, there are modeling techniques that require special data formats. So at this stage it is still possible to return to the previous stage.

6. Valid

If at this stage it is appropriate, it will proceed to the Clustering Application stage. If at this stage it is not appropriate or fails, it will return to the prototype design stage.

7. Implementation of Clustering on Promotion Strategy for Yadika 1 Vocational High School 1

At this stage of implementing clustering, it is one of the tools in data mining that aims to group objects into clusters in the promotion strategy at SMK Yadika 1.

8. System Testing

In the testing phase of this system based on the research design that has been carried out, namely the technique of testing the results of clustering by looking at the results of the Davies Bouldin Index and the testing technique in developing this information system using black box testing.

9. Valid

If at this stage it is appropriate, then everything is finished and running smoothly. If at this stage it is not appropriate or fails, it will return to the stage of implementing clustering in the promotion strategy of Yadika Vocational High School 1.

10. Finish

#### IV. RESULTS AND DISCUSSION

##### A. CRISP-DM Method

###### 1. Business Understanding

There are management difficulties in determining the Promotion Strategy of the Yadika 1 Vocational High School in West Jakarta. So in this study, the application of data mining with the K-Means method will be carried out to cluster the promotion strategy into seven clusters according to the Optimal K obtained when searching through the Davies Bouldin Index. Of the seven clusters, Tigas will be a promotional recommendation for promotion strategies at SMK Yadika 1, West Jakarta City.

###### 2. Data Understanding

In this study, the attributes used in this study refer to the criteria approach that has the most influence on the promotion strategy. Based on the results of observations that have been made, the authors obtained student data at SMK Yadika 1. The data obtained in this study was obtained through the Administrative section of SMK Yadika, the data obtained were student data from 2017 to 2019 The number of students in this study amounted to 1034 students with each department. From 761 student data that has been obtained, then the data is in Cleaning (Data Cleaning) then obtained a total of students who can be used in this study amounted to 508 students, Attributes before being selected amounted to 21 attributes.

###### 3. Data Preparation

The data obtained for this study were 508 student data from 1036 student data obtained at SMK Yadika 1

West Jakarta. To get quality data, the author uses the initial data technique, data selection, data labeling, and conversion of data labeling results. This stage performs some preparation of data processing. Preparation of the data process, namely: Data Cleaning.

###### 4. Data Integration and Data Reduction

This data selection is a data selection process by focusing on data that can be used to determine the promotion strategy of the Yadika 1 Vocational School. After getting the selection results, the attributes used are 11 attributes, including name, gender, city of birth, religion, sub-district, batch, year of the last report card, field of study of expertise, program of study of expertise, competence of expertise, and the origin of the student's school. The results of the attributes after being selected can be seen in table 2

Table 2. Attributes After Selection

No	Nama Atribut	Tipe Data	Keterangan
1.	Name	String	Nama Siswa
2.	Gender	String	Jenis kelamin siswa
3.	City of Birth	string	Kota Lahir
4.	Religion	string	Agama siswa
5.	Districts	string	Kecamatan siswa
6.	Force	int	Angkatan masuk sekolah
7.	Report Three Year	int	Tahun 3 Raport siswa
8.	Field of study Expertise	String	Bidang Studi Keahlian siswa
9.	Skill Study program	String	Program Studi Keahlian
10.	Skill Competency	String	Kompetensi keahlian
11.	Student's School Origin	String	Asal SMP Siswa

The next stage is data integration. Data integration is the process of converting or merging data into a format suitable for processing in data mining. Often the data that will be used in the data mining process has a format that cannot be directly used. Therefore, the format needs to be changed. In this study, nominal data is initialized in the form of numbers so that it can be processed using the K-means Clustering algorithm.

At this stage, data labeling is carried out on the data that has been selected. The data can be seen in Table 4.5 :

Table 3. Selected data

No	Attribute Name	Data Type
1.	Name	String
2.	Gender	String
3.	City of Birth	String
4.	Religion	String
5.	Districts	String
6.	Force	Int
7.	Report Three Year	Int
8.	Field of study Expertise	String
9.	Skill Study program	String
10.	Skill Competency	String
11.	Student's School Origin	String

The results of the data labeling are then transferred to notepad++ with .arff format. The form of the data is shown in Figure 3 :



From the results of the above calculations, it is found that the distance between student 1 and the center of cluster 4 is 105 .

$$DM5 = \sqrt{\begin{matrix} (1-2)^2 + (18-18)^2 + (4-4)^2 + \\ (16-16)^2 + (2-2)^2 + (2-2)^2 + \\ (1-1)^2(1-1)^2 + (1-1)^2(113-41)^2 \end{matrix}} = 5185$$

From the results of the above calculations, it is found that the distance between student 1 and the center of cluster 5 is 5185

$$DM6 = \sqrt{\begin{matrix} (1-2)^2 + (18-18)^2 + (4-1)^2 + \\ (16-10)^2 + (2-2)^2 + (2-2)^2 + \\ (1-1)^2(1-2)^2 + (1-2)^2(113-75)^2 \end{matrix}} = 1492$$

From the results of the above calculations, it is found that the distance between student 1 and the center of cluster 6 is 1492 .

$$DM7 = \sqrt{\begin{matrix} (1-2)^2 + (18-18)^2 + (4-2)^2 + \\ (16-10)^2 + (2-2)^2 + (2-2)^2 + \\ (1-1)^2(1-2)^2 + (1-2)^2(113-72)^2 \end{matrix}} = 1724$$

From the results of the above calculations, it is found that the distance between student 1 and the center of cluster 7 is 1724 .

In this case  $d(m_i, m_j)$  represents the Euclidean distance from  $m$  to  $m_j$ . Calculating WCV That is by choosing the smallest distance between the data and the centroid in each cluster. The following is the closest distance (iteration 1) in the form of a table which can be seen in table 4:

Table 4. Nearest Distance (Iteration 1)

Jarak Terdekat
21,84
15,00
777,14
1048,89
131,19
1472,20
26,22
12770,47
12772,18
12338,11
10419,81
1698,11
2603,75
2604,18
12553,49
8714,78
12771,11
12770,90
487,82

909,49
12771,75
12782,61
3292,78
9803,11
21,84

WCV = 17707.41

So that the ratio =  $BCV/WCV = 3583.43 / 17707.41 = 0,20236895$

Because this step is the first iteration then proceed to the next step.

Based on the calculations that have been carried out, the results for each cluster are obtained, for Cluster 1 as many as 111 with a percentage of 21.85%, Cluster 2 as many as 32 with a percentage of 6.3%, Cluster 3 as many as 30 with a percentage of 5.9%, Cluster 4 as many as 83 with a percentage of 16.33%, Cluster 5 as many as 99 with a percentage of 19.48%, Cluster 6 as many as 100 with a percentage of 19.68% and Cluster 7 as many as 53 with a percentage of 10.43%. The following is a graph of the results of manual calculations obtained as shown in Figure 4:

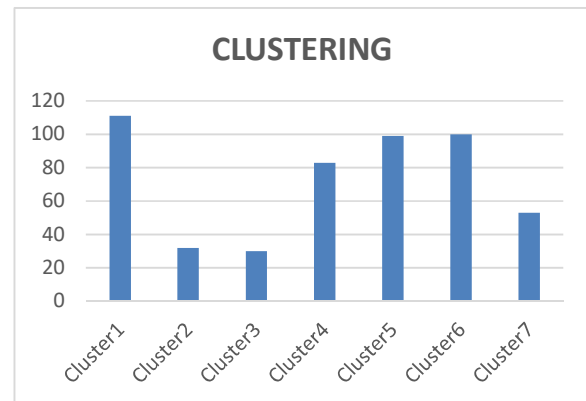


Figure 4. Manual Calculation

### C. Results of Data Visualization of Each Attribute Using WEKA

#### 1. Results of Visualization of Religious Attributes

The following is a form of attribute visualization that is used as a reference for promotional strategies using WEKA tools, namely:

##### a. Gender Attribute Visualization

Visualization of the Address Attribute. It is known that from 508 data in the column selected attribute there is no missing data as much as 0 or 0%. The minimum statistic has a value of 0, the maximum statistic has a value of 2, the statistical mean (average) has a value of 1.283, the statistical standard deviation has a value of 0.455, Gender Attribute Visualization can be seen in Figure 5.



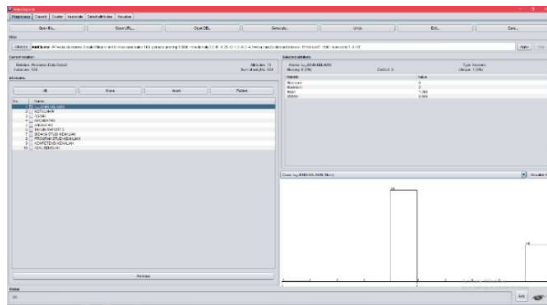


Figure 5 Visualization of Gender Attributes

b. Visualisasi b. City of Birth Attribute Visualization

Visualization of the City of Birth Attributes. It is known that from 508 data in the column selected attribute there is no missing data as much as 0 or 0%. The minimum statistic has a value of 0, the maximum statistic has a value of 72, the statistical mean (average) has a value of 23,337, the statistical standard deviation has a value of 13,848, Visualization of the City of Birth Attributes can be seen in Figure 6.

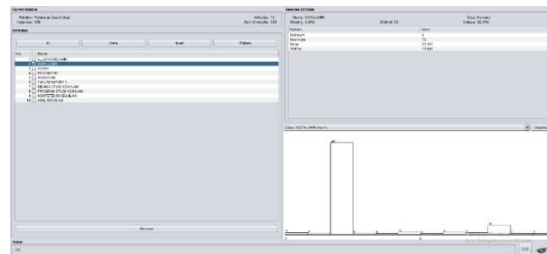


Figure 6. Visualization of Birth City Attributes

c. Religious Attribute Visualization

Visualization of Major Attributes. It is known that from 508 data in the column selected attribute there is no missing data as much as 0 or 0%. The minimum statistic has a value of 1, the maximum statistic has a value of 5, the statistical mean (average) has a value of 1.758, the statistical standard deviation has a value of 1.058, the Visualization of Religious Attributes can be seen in Figure 7..

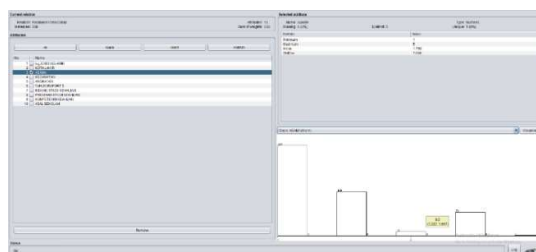


Figure 7. Visualization of Religious Attributes

d. District Attribute Visualization

Visualization of General Subject Value Attributes. It is known that from 508 data in the column selected attribute there is no missing data as much as 0 or 0%. The

minimum statistic has a value of 0, the maximum statistic has a value of 16, the statistical mean (average) has a value of 11,352, the standard deviation statistic has a value of 3.334, Visualization of District Attributes can be seen in Figure 8.

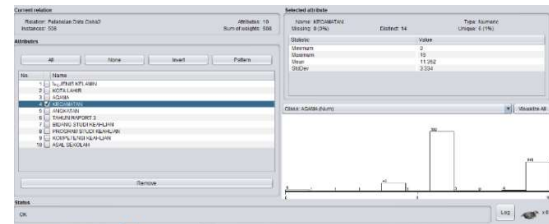


Figure 8. Visualization of District Attributes

e. VForce Attribute Visual

Visualization of School Origin Attributes. It is known that from 508 data in the column selected attribute there is no missing data as much as 0 or 0%. The minimum statistic has a value of 0, the maximum statistic has a value of 3, the statistical mean (average) has a value of 2.413, the statistical standard deviation has a value of 0.571, Visualization of Force Attributes can be seen in Figure 9.

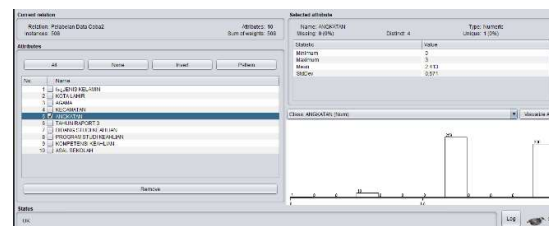


Figure 9 Visualization of Force Attributes

f. Last Report Year Attribute Visualization

Visualization of Religious Attributes. It is known that from 508 data there are 0 or 0% missing data. The minimum statistic has a value of 0, the maximum statistic has a value of 3, the statistical mean (average) has a value of 2.419, the statistical standard deviation has a value of 0.561, Visualization of the Attributes of the Year Reports can be seen in Figure 10.



Figure 10 Visualization of the Last Report Year Attributes

g. Visualization of Field of Study Attributes of Expertise Visualisasi dari Atribut Bidang Studi Keahlian. Diketahui bahwa dari 508 data terdapat missing data sebanyak 10 atau 1%. Pada statistik minimum terdapat nilai 1, statistik maximum terdapat nilai 2, statistik mean (rata-rata) terdapat nilai 1,495, statistik standard deviasi terdapat nilai 0,5, Visualisasi Atribut Bidang Studi keahlian dapat dilihat

di Gambar 11.

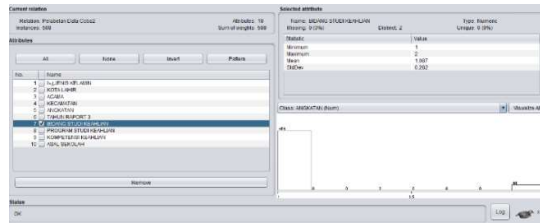


Figure 11 Visualization of the Attributes of the Field of Expertise

h. Visualisasi h. Visualization of Skills Study Program Attributes

visualization of Kelurahan Attributes. It is known that from 508 data there are 0 or 0% missing data. The minimum statistic has a value of 1, the maximum statistic has a value of 3, the statistical mean (average) has a value of 1.591, the statistical standard deviation has a value of 0.645, Visualization of the Attributes of the Skills Study Program can be seen in Figure 11.

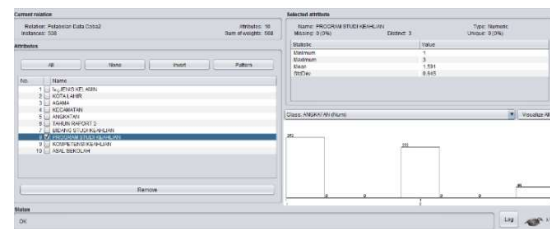


Figure 11. Visualization of Skills Study Program Attributes

i. Visualization of Skill Competency Attributes

It is known that from 508 data there are 0 or 0% missing data. The minimum statistic has a value of 1, the maximum statistic has a value of 3, the statistical mean (average) has a value of 1.591, the statistical standard deviation has a value of 0.645. Visualization of competency attributes can be seen in Figure 12.

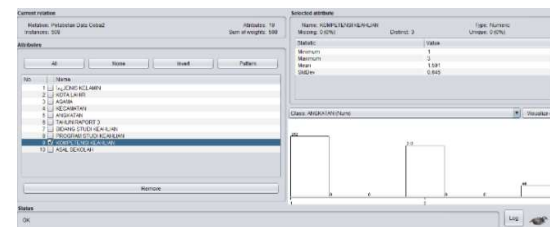


Figure 12 Visualization of Skill Competency Attributes

2. Visualization of School Origin Attributes

Visualization of the Residence Type Attribute. It is known that from 508 data there are 0 or 0% missing data. The minimum statistic has a value of 1, the maximum statistic has a value of 3, the statistical mean (average) has a value of 1.247, the statistical standard deviation has a value of 0.645, Visualization of School Origin Attributes can be seen in Figure 13..

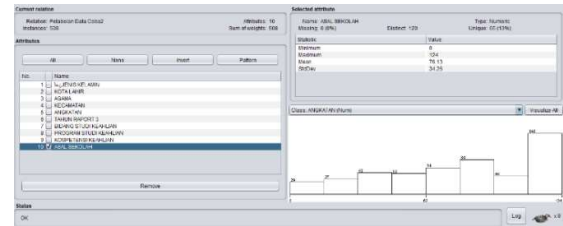


Figure 13 Visualization of School Origin Attributes

2. Cluster Analysis Results With Weka Tools

The results of the cluster analysis, where there are 7 predetermined clusters, the calculation is continued until all data is counted and produces groups into clusters with minimal distances. The iteration was stopped because there were the same cluster center numbers in the 8th iteration. The results of the cluster formed after the 8th iteration did not change, so the iteration was stopped. Clusters were chosen randomly after that the closest distance to the clusters was found in Cluster 4, Cluster 2, Cluster 3, Cluster 6, Cluster 5 and Cluster 7. From the results of cluster 1 it can be seen that the number of students was 57 students (11%). cluster 2 shows that the number of students is 75 students (15%), Then from the results of cluster 3 it can be seen that the number of students is 70 students (14%), then from the results of cluster 4 it can be seen that the number of students is 182 (36%), then from the results cluster 5 shows that the number of students is 44 (9%), then from the results of cluster 6 it can be seen that the number of students is 58 (11%), then from the results of cluster 7 it can be seen that the number of students is 22 (4%). The results of the cluster analysis with the Weka tools can be seen in Figure 14:.

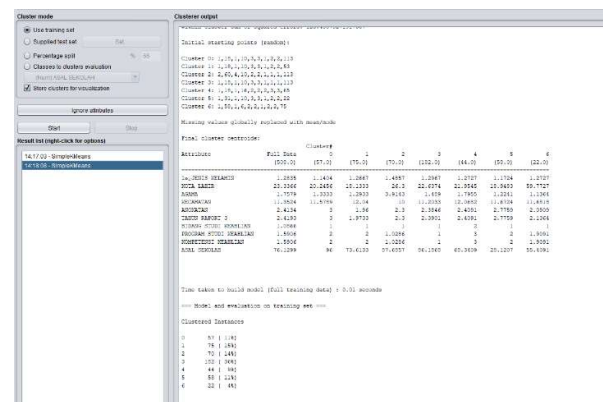


Figure 14 Cluster Analysis

3. Clustering Results Using Weka

From the data from the results of the clustering that has been carried out, it can be determined several promotional strategies that can be carried out by SMK Yadika 1 in conducting promotions based on the Marketing Department which are divided into 7 groups / clusters. The results of Clustering using Weka can be seen in table 5.

Table 5. Clustering Results Using Weka

Cluste r 1	Cluste r 2	Cluste r 3	Cluste r 4	Cluste r 5	Cluste r 6	Clus ter 7
1	1	2	1	1	1	1
18	18	60	18	18	31	58
1	1	4	1	1	1	1
10	10	10	10	16	10	6
3	3	2	3	2	3	2
3	3	2	3	2	3	2
1	1	1	1	2	1	1
2	2	1	1	3	2	2
2	2	1	1	3	2	2
113	53	113	113	65	75	75

#### 4. Cluster Analysis Results Using Python

##### a. Selection of K values using the Davies Bouldin Index (DBI) method

After clustering, the next step is to analyze the results of the clustering using Python to get the "Davies Bouldin index" value. From the analysis carried out, for the lowest DBI with a DBI value of 0.552 and the optimal K is 7, and this value is far below 1.0 or it can be concluded that the grouping process is quite good according to Figure 15 below.

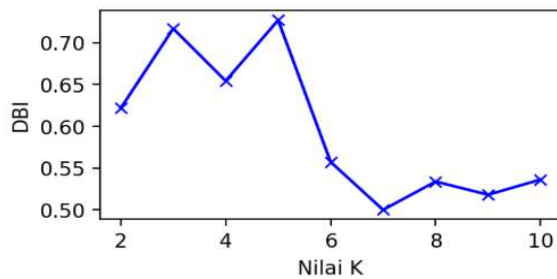


Figure 15 Davies Bouldin Index and K Optimal Values

#### D. K-means Modeling

After seeing the results of the Davies Bouldin Index (DBI) Method, for K-Means Modeling using Python, we can see in Figure 16 below :

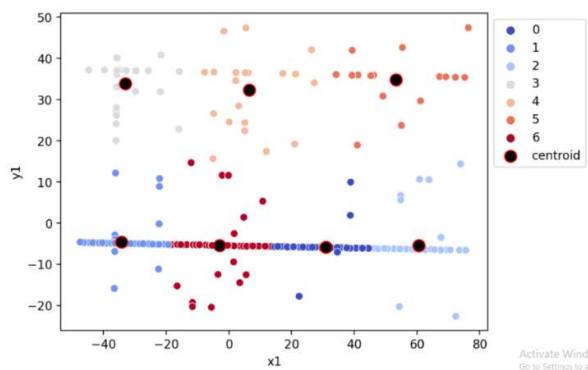
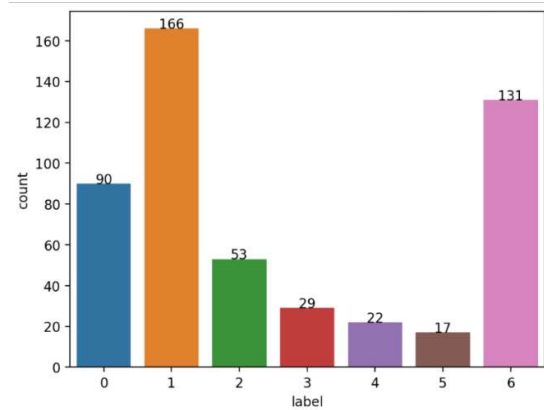


Figure 16 K-Means Modeling

##### 1. Clustering Results

The results of the cluster analysis, where there are 7 clusters that have been determined according to the optimal K, the calculation results in groups into clusters

with a minimum distance. Because there are the same cluster center numbers in the 5th iteration. that the number of students is 90 students (17.71%), Then from the results of cluster 2 it can be seen that the number of students is 166 students (32.67%), Then from the results of cluster 3 it can be seen that the number of students is 53 students (10.43%), then from the results of cluster 4 it can be seen that the number of students is 29 (5.7%), then from the results of cluster 5 it can be seen that the number of students is 22 (4.33%), then from the results of cluster 6 it can be seen that the number of students is 17 (3.34%), then from the results of cluster 7 it can be seen that the number of students as many as 131 (25.78%). with the data population according to Figure 17.



```
[7]:
```

K	dbi
0	2 0.635089
1	3 0.741507
2	4 0.684332
3	5 0.769454
4	6 0.599900
5	7 0.552802
6	8 0.615288
7	9 0.632581
8	10 0.653840

Figure 17 Clustering Results

##### 2. Business Use Case

The following design is a business use case for the system built on the K-Means Calculation application design to determine the best promotion strategy according to Figure 19 below :

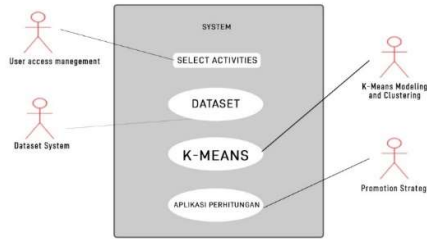
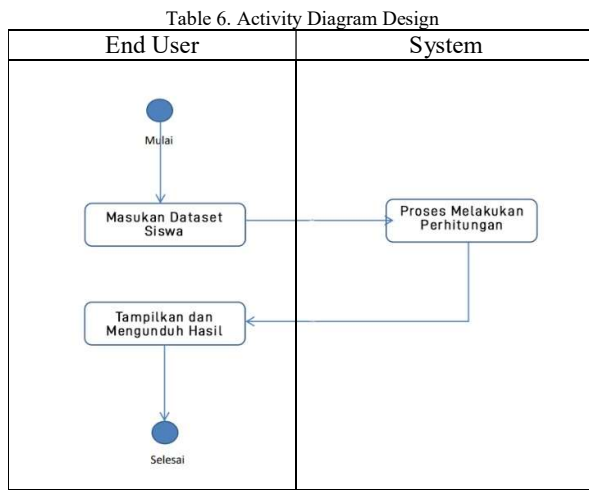


Figure 19. Business Use Case

### 3. Activity Diagram

The following is an activity diagram design that was built, the stages of the process of using the K-Means calculation application to determine the promotion strategy in accordance with the system design according to Table 6 below :



### 4. Prototype Implementation

At the implementation stage of this prototype, the prototype was built using the Python programming language with streamlit as the interface and uploading the dataset and downloading the results, the following is the initial display of the prototype calculation application according to Figure 20 below :

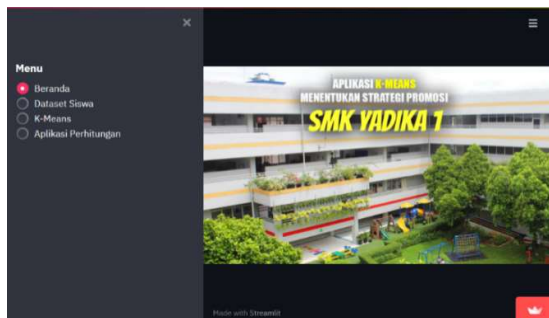


Figure 20 Initial Screen of Applications

On the first page of the prototype there are several menus including the homepage, student datasets, K-Means modeling, and Calculation Applications as the K-Means

calculation process to determine the best promotion strategy. On the student dataset page to see the attributes that will be used in the K-Means search process according to Figure 21 below.

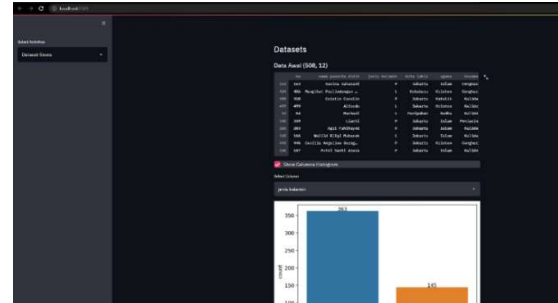


Figure 21 Student Dataset Pages

On the K-Means page there are the results of selecting the K value using the Davies Bouldin Index (DBI) Index, the K-Means Model Simulation with the K value can be changed and the clustering results according to Figure 22 below

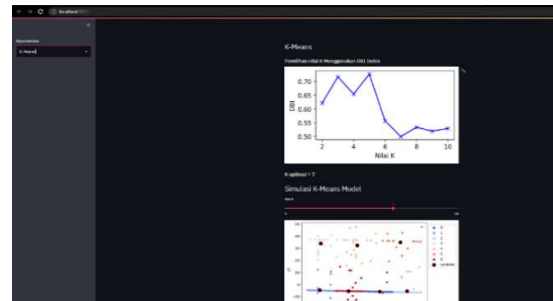



Figure 22 K-Means . Prototype Pages

### 5. Black Box Testing

#### a. Black Box Testing on the Calculation Application Page

In the Black Box test on the calculation application page to see whether the expected results are appropriate or not, it can be seen in table 7.

Tabel 7 Hasil Aplikasi Perhitungan K-Means

No	Testing Scenario	Expected results	Conclusion
1.	Upload a Dataset whose data is incorrect and does not match the format that has been set in the calculation application, then let it process the inputted data, Testing:	The system will display errors or process data that does not match the format set in the calculation application. Test result : 	Already appropriate
2.	Upload a Dataset whose data is	The system will show the process that took place until	Already

	incorrect True according to the format that has been set in the calculation application, then let it process the inputted data, Testing:	the result and can be downloaded. Test result :	appropriate
--	--	---	-------------




b. *Black Box Testing on Dataset Pages*

Pada pengujian *Black Box* pada halaman dataset perhitungan untuk melihat hasil yang diharapkan apakah sesuai atau tidak bisa dilihat pada tabel 8.

Table 8 Student Dataset Results

No	Testing Scenario	Expected results	Conclusion
1.	The system will display the dataset menu if you click the dataset menu button.	The system displays the dataset menu, the test results are:	Already appropriate
2.	In the Dataset, the system will display a menu for each Gender attribute if you click the button according to the attribute that will be displayed.	The system displays the Gender attribute on the dataset menu, the test results are:	Already appropriate
3.	In the dataset, the system will display a menu for each attribute City of birth if you click the button according to the attribute that will be displayed.	The system displays the City of birth attribute on the dataset menu. The test results are:	Already appropriate
4.	In the dataset, the system will display the Religion attribute menu if you click the button according to the attribute to be displayed.	The system displays the Religion attribute on the dataset menu, the test results are:	Already appropriate
5.	In the dataset, the system will display the sub-	The system displays the sub-	Already appropriate


	district attribute menu if you click the button according to the attribute that will be displayed.	on the dataset menu, the test results are:	
6.	In the Dataset, the system will display the Force attribute menu if you click the button according to the attribute to be displayed.	The system displays the Force attribute on the dataset menu, the test results are:	Already appropriate
7.	In the dataset, the system will display the last Report Year attribute menu for the last report card if you click the button according to the attribute to be displayed.	The system displays the last Report Year attribute in the dataset menu, the test results are:	Already appropriate
8.	In the dataset, the system will display the attributes of the field of expertise if you click the button according to the attribute that will be displayed.	The system displays the attributes of the field of expertise on the dataset menu, the test results are:	Already appropriate
9.	In the dataset, the system will display an attribute menu for the study program of expertise if you click the button according to the attributes that will be displayed.	The system displays the attributes of the skill study program on the dataset menu, the test results are:	Sudah sesuai
10.	In the dataset, the system will display the skill competency attributes on the dataset menu, the test results are:	The system displays the skill competency attributes on the dataset menu, the test results are:	Already appropriate

	the attribute to be displayed.		
11	In the dataset, the system will display the menu for each school attribute in detail if you click the button according to the attribute to be displayed.	The system displays the school's attributes on the dataset menu. The test results are: 	Already appropriate

c. Black Box Testing on K-Means Pages

In the Black Box test on the calculation application page to see whether the expected results are appropriate or not, it can be seen in table 9.

**Table 9 Results of K-Means**

No	Testing Scenario	Expected results	Conclusion
1.	The system will display the K-Means menu, if you click the dataset menu button, the results of the K-Means Model will be displayed.	The system displays the K-Means menu, the test results are: 	Already appropriate
2.	The K-Means Menu will display a K-Means Model Simulation whose K value can be changed if you click the simulation button with a K value from numbers 2 to 10 then it will be displayed according to your choice.	The system displays the K-Means Model Simulation by selecting the value of K 10 on the K-Means menu. The test results are:	Already appropriate
3.	The K-Means menu will display Cluster Options from Cluster 1 to Cluster 10, according to the Optimal K we want.	The system displays a choice of clusters that match our choice, the test results are:	Already appropriate

**V. CONCLUSION**

A. Conclusion

Based on the results of observations and research that has been done at SMK Yadika 1, it can be concluded as follows::

1. Applying the K-Means Method to determine the best promotion strategy in accordance with SMK Yadika 1.
  - a. The application of the K-Means Clustering Algorithm resulted in 7 clusters where Cluster 1 90 data (17.71%), cluster 2 166 data (32.67%), cluster 3 53 data (10.43%), cluster 4 29 data (5.7%), cluster 5 22 data (4.55%), cluster 6 17 data (3.34%), and cluster 7 131 data (25.78%) and the Davies Boulding Index (DBI) with a value of 0.552 and an optimal K of 7.
  - b. Based on the calculations that have been made, the Promotion Strategy that can be carried out by the school so that the promotion is carried out more effectively and efficiently are:
    - 1) Promotion based on Student's School Origin.
    - 2) Promotion based on the most sought after Field of Study of Expertise.
    - 3) Promotion based on Skills Study Program
    - 4) Promotion based on Skill Competence
    - 5) Promotion Student
2. Develop a data mining prototype to determine the best strategy according to the needs of SMK Yadika 1.

B. SUGGESTION

From the conclusions mentioned above, the authors provide suggestions for further development of the application of the K-means Clustering Method to determine the best promotion strategy, namely by developing a system for language that is easier to understand so as to assist schools in receiving the information generated and this research can be developed with adding the amount of new data in the attributes of the data mining prototype.

**REFERENCES**

- [1] Sumangkut, K., Lumenta, A. S. M. and Tulenan, V. 2016. Analysis of Daily Mart supermarket shopping patterns to determine the layout of goods using the FP-Growth Algorithm. *Journal of Informatics Engineering*, 8(1).
- [2] Budiman, I. 2015. Application of Classification Data Mining Functions for Prediction of Timely Student Study Period in Higher Education Academic Information Systems. *Journal of Jupiter*, 7(1), pp. 39–50.
- [3] Badrul, M. 2016. Association Algorithm With Apriori Algorithm for Sales Data Analysis. *Journal of Pilar Nusa Mandiri*, 12(2), pp. 121–129.
- [4] Asril, E., Wiza, F. and Yunefri, Y. 2015. Analysis of Graduate Data with Data Mining to Support the Promotion Strategy of Lancang Kuning University. *Journal of Information & Communication Technology Digital Zone*, 6(2), pp. 24-32.
- [5] Priambudi, D. S. 2015. Pt.Mayora's Product Sales Strategy Using Apriori Methods And Data Mining Implementation. *Thesis Article*, pp. 1–9.
- [6] Rony, S. 2016. Data Mining Application Using K-Means Clustering Algorithm to Determine New Student Promotion Strategy (Case Study: Polytechnic Lp3i Jakarta). *Journal of Lanterns ICT*, 3(1), pp. 76–92.

- [7] Suprawoto, T. 2016. Classification of Student Data Using the K-Means Method to Support the Selection of Marketing Strategies. *Journal of Informatics and Computers*, 1(1), pp. 12–18.
- [8] Wirta, A. and Erlin. 2016. Implementation of the K-Means Cluster Analysis Method for Selecting New Student Admission Promotion Strategies. *National Seminar on computer science*, pp. 9–15.
- [9] Mochammad C. A. 2018. Data Mining Uses the K-Means Algorithm to determine promotion strategies in intentional vocational schools. *Thesis Article*, pp. 1–12.
- [10] Achyani, Y. E. 2018. Application of Particle Swarm Optimization Method in Optimizing Direct Marketing Predictions. *Journal of Informatics*, 5(1), pp. 1–11.
- [11] Kusumo, H., Sedyono, E. and Marwata, M. 2019. Analysis of Apriori Algorithms to Support Higher Education Promotion Strategies. *Walisongo Journal of Information Technology*, 1(1), p. 49.
- [12] Jaini, A. 2019. Application of the Fuzzy C-Means Algorithm and K-Medoids for Grouping Sales and Product Marketing Strategies. *State Islamic University of Sultan Syarif Kasim Riau*.
- [13] Takdirillah, R. 2020. Application of Data Mining Using Apriori Algorithm Against Transaction Data to Support Sales Strategy Information. *Journal of Informatics Education*, 4(1), pp. 37–46.

# Online Monitoring and Analysis of Lube Oil Degradation for Gas Turbine Engine using Recurrent Neural Network (RNN)

Febrianto Nugroho<sup>1\*)</sup>, Rusdianto Roestam<sup>2</sup>

<sup>1,2</sup>Information Technology Program, Faculty of Computing, President University  
Email: <sup>1</sup>febrianto.nugroho@student.president.ac.id, <sup>2</sup> rusdianto@president.ac.id

**Abstract** – Lubrication is one of the important aspects of the engine that will impact the overall performance of the gas turbine engine. Degradation of oil is usually known by offline analysis that use oil sample to check some properties and contaminant. The offline analysis will take a longer time, as needed to collect the sample, send it to the laboratory, analyze the sample and create the report. The purpose of this research is to analyze oil parameters in real-time so can predict oil degradation. Sensors and transducers installed on the lube oil system can read some parameters of the oil then transmit easily to the server. The method that will use in this paper is Recurrent Neural Network (RNN) with multi-step Long Short Term Memory (LSTM). The result of this paper will predict oil degradation on the future operation of gas turbine engine.

**Keywords** – lubrication, oil, gas turbine, IoT, condition monitoring

## I. INTRODUCTION

A gas turbine engine is equipment that uses gas as fluid to rotate the turbine with internal combustion. It converts kinetic energy to mechanical energy [1]. This engine consists of two main components, there are stator and rotor. The rotor supports by some bearings when it is rotating. The bearings and other parts such as gears must be lubricated prior, during, and post-operation to reduce the heat produced by friction to keep bearing at design working temperature [2].

Healthiness lubrication systems need to monitor continuously to ensure the safe operation of the turbine and prevent catastrophic failure [3]. Lubrication systems contain some equipment, piping, and oil itself. This paper focuses on oil that flows through the system. New oil is used for the first time of engine operation. During normal operation, the quality of oil will degrade over time. Degradation of the oil will affect the performance of lubrication itself. Level of degradation does not easy and fast to know as need to perform offline lube analysis that take several weeks from taking sampling, shipping to laboratory, reading each properties, analysis and generate report.

There were several previous papers related lube oil system analysis that can divide by three topics, first is fault finding [4], second is condition-based maintenance [5] and third is online machine health monitoring [6]. Lube oil parameters used to find fault equipment on the lube oil system [4], while other research [5], it used for predicting maintenance plan of equipment. Measuring multiple oil properties as online health monitoring tried to replace some of offline sampling parameters [6].

In this research will focus to analyze the historical parameters of the oil to predict oil degradation on the future operation of gas turbine engine. The method that will be used is Recurrent Neural Network to solve the problem that uses time-series data [7] while other research uses a combination of Rough Set and Feed-forward Neural

Network for fault diagnostic and condition-based maintenance of lubrication system.

## II. RESEARCH METHODOLOGY

This research will be performed using the historical log of the gas turbine engine from Company XYZ. Historical log contains all parameters of gas turbine engine includes lube oil system which recorded every hour from November 2019 to November 2020. Based on the type of dataset that will be used, RNN is one of the best deep learning methods to predict future state of time series data. RNN use their internal state memory for processing sequences. So, it usually used for time series forecasting, audio analysis, handwriting recognition and other application [8]. However, there is limitation of memory for simple RNN, so LSTM variant used to handle this issue as it can save information for longer time than simple RNN [9].

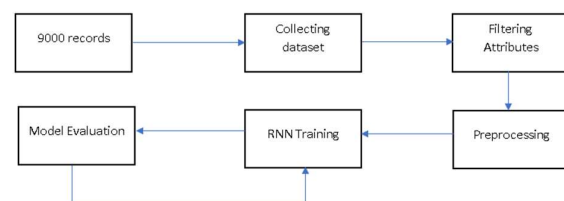


Figure 1 Oil degradation prediction step

### A. Dataset collecting

Research starts from collecting a dataset of the engine from server. Raw data set will filter for lubrication system attributes and engine load then export to csv file. There are seven attributes that will be used for training. Parameters that will be used in this research consist of time, engine running hour, lube oil header pressure and temperature, bearing drain temperature, delta bearing drain temperature, and total kW (kilowatt).

### B. Preprocessing step

Preprocessing step is used for preparing the dataset so it can be ready to process to the next step which is the training



step. This step consists of data split and normalization. The data split that is used for training is 70 % of first 1000 data.

Table 1 Dataset sample from engine historical log

TIME	RH	LO_PRESS	LO_TEMP	BRG_DRN	DELTA_BRG_DRN	kW
11/6/2019 0:00	10740	2.82	60.94	105.78	44.83	8322
11/6/2019 1:00	10741	2.81	62.11	106.44	44.33	8645.5
11/6/2019 2:00	10742	2.81	61.5	105.58	44.08	8238
11/6/2019 3:00	10743	2.81	60.17	104.5	44.33	7801.5
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
11/21/2020 8:00	20210	2.95	61.39	90.22	28.83	2023.96
11/21/2020 9:00	20211	2.96	63.33	91.5	28.17	1998.18
11/21/2020 10:00	20212	2.96	64.56	92.5	27.94	2027.13

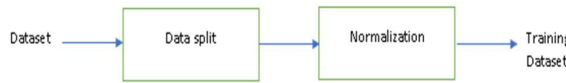
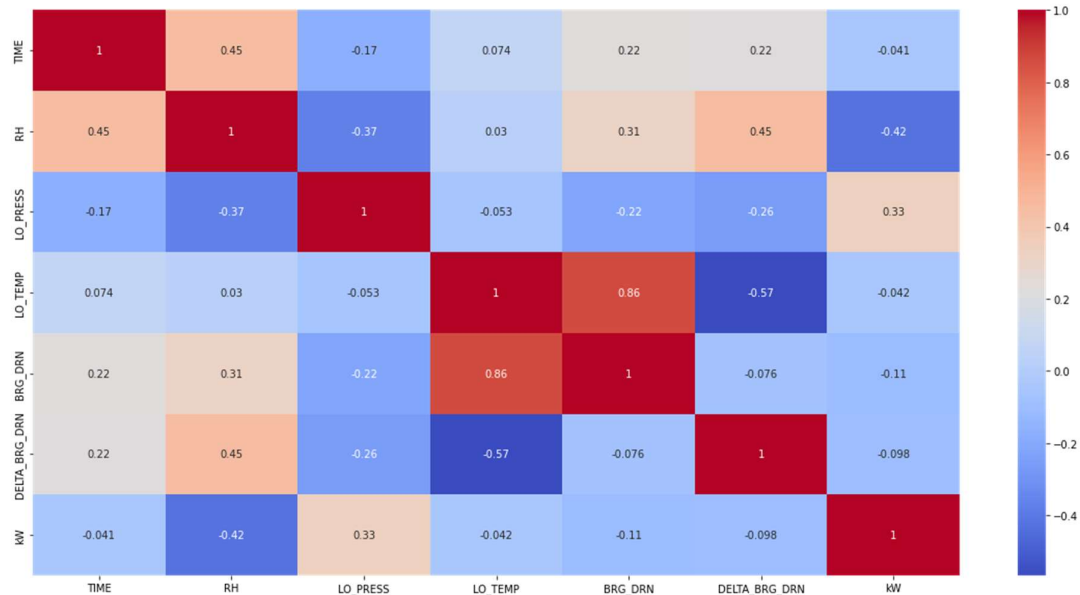


Figure 2 Pre-processing step

Normalization will transform the numeric value become 0 to 1 value.

Table 2 Correlation Matrix



There are many factors that affecting of increasing delta bearing drain temperature, based on correlation matrix on table 2, it shows that BRG\_DRN, LO\_TEMP and RH (running hour) are top three of importance label that have more correlation with delta bearing drain temperature.

**B. Insight of data**

Comparison of parameters with many combinations can get some insight of dataset. Load expected of the engine is above 9000 kW. Filtering applied for kW that only show above 8000 kw data as this is minimum load requirement. Optimum delta bearing temperature at load above 9000 kW is between 42-48 deg C as shown on Figure 3. Engine load is decreasing by increasing running hour as delta bearing drain temperature increase, shown on Figure 4. Based on Figure 5 delta bearing temperature increasing linear around

**C. Training step**

In this training step, data set will split in two parts which consist of train data and test data. Base on empirical studies, it showed that using 70-80% train data and 20-30% test data will get best result [10]. In this research will use 80:20 ratio between train data and test data. Training method that will be used is one of best accuracy methods, so it will compare all methods to get the best one.

**III. RESULTS AND DISCUSSION**

**A. Correlation Matrix**

All labels on dataset will analyze to get correlation between them [11]. In this paper will focus on DELTA\_BRG\_DRN (delta between lube oil temperature header) and LO\_TEMP (lube oil bearing drain temperature) feature to see degradation of lube oil. Increasing delta temperature means performance of absorbing heat from bearing by lube oil is decreasing. Thus, will decreasing the load (kW) of gas turbine engine to maintain bearing drain temperature below alarm set point.

8 deg C for 4,000 running hours period. After replacing the lube oil, delta drain bearing temperature back to normal then slightly increasing linear with running hour.

**C. Prediction**

Using RNN-LSTM, data set will train and create the model, from the model it can predict future data using Tensorflow module [12] with multi-step prediction [13]. Based on prediction, control system can send alarm at 24 hours prior to reach high set point so engine operator can act to prevent unplanned shutdown. Figure 7 shows prediction for next hour based on 24 hours last reading.

For predict to next 24-hour show on Figure 8. MAE used for evaluation as it is better than MSE (Mean Squared Error) [14]. It got MAE (Mean absolute error) higher than one step prediction.



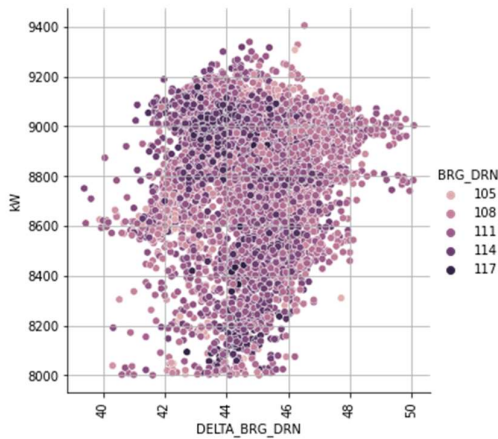


Figure 3 Delta bearing drain vs. kW with bearing drain

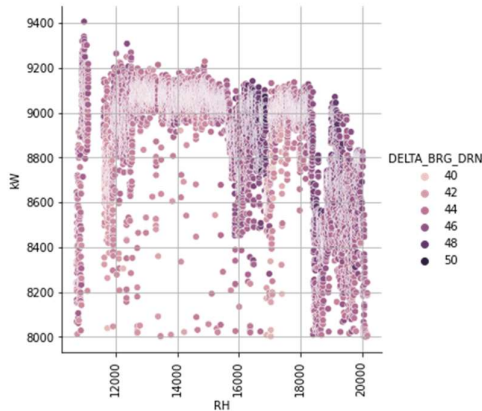


Figure 4 RH and kW with Delta bearing drain

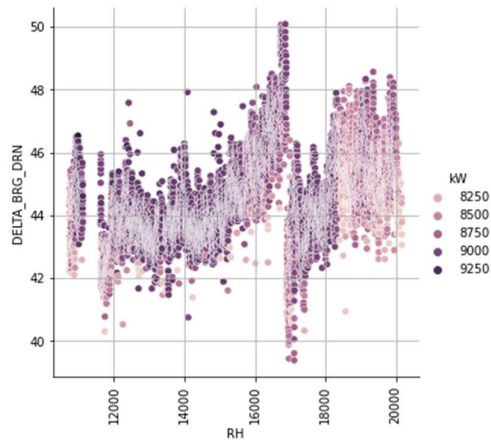


Figure 5 RH vs. delta bearing drain with kW

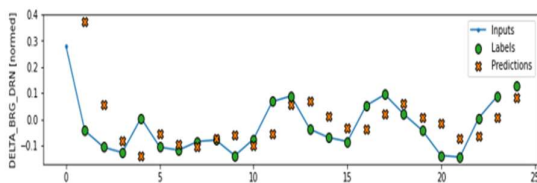


Figure 6 Prediction single-step next an hour

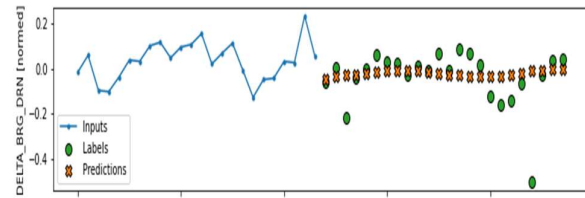


Figure 7 Multi-step model using RNN-LSTM to predict next 24h

Table 3 Model performance

Step	Loss	MAE (Mean Absolute Error)
Single	0.086	0.1292
Multi	0.7654	0.4701

#### IV. CONCLUSION

Future prediction of lube oil degradation for next 24 hours give useful insight that will be used as decision to treat and or to replace the lube oil. This project is initial concept to add analysis feature on the HMI and or Server. Based on results from prediction, it needs more work to get better method than current research by combination of method and or modify layer of the method. Beside on the lube oil temperature prediction, next research should consider also to add other oil properties measurement system such as LOAC (Lab-On-A-Chip) [15] for more accurate analysis. In the future, prediction system will be embedded in to Turbine control and monitoring system as added value of artificial neural network instead only automatic control system.

#### REFERENCES

- [1] Kurz, R., Meher-Homji, C., Fellow, B., Manager, J. M., & Gonzalez, F., "GAS TURBINE PERFORMANCE AND MAINTENANCE", Proceedings of the Forty-Second Turbomachinery Symposium, 2013, Turbomachinery Laboratory, Texas A&M Engineering Experiment Station.
- [2] Nehal S. Ahmed and Amal M. Nassar (May 22nd, 2013). Lubrication and Lubricants, Tribology - Fundamentals and Advancements, Jürgen Gegner, IntechOpen, DOI: 10.5772/56043.
- [3] Benbouzid, M., Berghout, T., Sarma, N., Djurović, S., Wu, Y., & Ma, X. (2021). Intelligent condition monitoring of wind power systems: State of the art review. In *Energies* (Vol. 14, Issue 18). MDPI. <https://doi.org/10.3390/en14185967>
- [4] S. Yi, N. Zhao, S. Li and Z. Xu, "A study on fault diagnostic method for the lube oil system of gas turbine based on rough sets theory," 2014 11th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), 2014, pp. 42-48, doi: 10.1109/FSKD.2014.6980804.
- [5] L. Ciani, G. Guidi and G. Patrizi, "Condition-based Maintenance for Oil&Gas system basing on Failure Modes and Effects Analysis," 2019 IEEE 5th International forum on Research and Technology for Society and Industry (RTSI), 2019, pp. 85-90, doi: 10.1109/RTSI.2019.8895587.



- [6] Xiaoliang Zhu, Chong Zhong, Jiang Zhe, Lubricating oil conditioning sensors for online machine health monitoring – A review, *Tribology International*, Volume 109, 2017, Pages 473-484, ISSN 0301-679X, <https://doi.org/10.1016/j.triboint.2017.01.015>.
- [7] Petneházi, G. (2018). *Recurrent Neural Networks for Time Series Forecasting*. <http://arxiv.org/abs/1901.00069>
- [8] Chatterjee Chandra, (15 December 2021), URL: <https://towardsdatascience.com/implementation-of-rnn-lstm-and-gru-a4250bf6c090>
- [9] Manaswi, N. K. (2018). *Deep Learning with Applications Using Python*. Apress.
- [10] Gholamy, A., Kreinovich, V., & Kosheleva, O. (2018). *A Pedagogical Explanation A Pedagogical Explanation Part of the Computer Sciences Commons*. [https://scholarworks.utep.edu/cs\\_techrep/](https://scholarworks.utep.edu/cs_techrep/)[https://scholarworks.utep.edu/cs\\_techrep/1209](https://scholarworks.utep.edu/cs_techrep/1209)
- [11] Aditya, (15-December-2021). URL: <https://www.kaggle.com/adityadesai13/used-car-dataset-ford-and-mercedes/code>
- [12] USENIX Association., ACM SIGMOBILE., ACM Special Interest Group in Operating Systems., & ACM Digital Library. (2005). *Papers presented at the Workshop on Wireless Traffic Measurements and Modeling: June 5, 2005, Seattle, WA, USA*. USENIX Association. [https://www.tensorflow.org/tutorials/structured\\_data/time\\_series](https://www.tensorflow.org/tutorials/structured_data/time_series) . <https://doi.org/10.5281/zenodo.4724125>
- [13] Chandra, R., Goyal, S., & Gupta, R. (2021). Evaluation of Deep Learning Models for Multi-Step Ahead Time Series Prediction. *IEEE Access*, 9, 83105–83123. <https://doi.org/10.1109/ACCESS.2021.3085085>
- [14] Qi, J., Du, J., Siniscalchi, S. M., Ma, X., & Lee, C.-H. (2020). *On Mean Absolute Error for Deep Neural Network Based Vector-to-Vector Regression*. <https://doi.org/10.1109/LSP.2020.3016837>
- [15] Alabi, O., Wilson, R., Adegbotolu, U., & Kudehinbu, S. (2019). *SPE-195708-MS Realtime Lubricating Oil Analysis to Predict Equipment Failure*.



# Development of a Village Information System for Acceleration of Village Services in Desa Tegal Kecamatan Kemang Bogor

Deden Ardiansyah<sup>1\*</sup>, Prihastuti Harsani<sup>2</sup>, Eneng Tita Tosida<sup>3</sup> Abimanyu Oki Saputera<sup>4</sup>, Andhika Bhayangkari<sup>5</sup>

<sup>1</sup> Computer Engineering Study Program, Vocational School, Pakuan University

<sup>2,3,4,5</sup> Computer Science Study Program, Faculty of Mathematics and Natural Sciences, Pakuan University

Jl. Pakuan, Tegallega. Central Bogor District, Bogor City. West Java 16143

[ardiansyahdeden@unpak.ac.id](mailto:ardiansyahdeden@unpak.ac.id)

## Abstract

*The Village Information System (SID) is an information system that changes raw data into ready-to-use information. In addition, SID will provide convenience to village officials in providing services to the community. The development of this SID is expected to be able to provide acceleration and improve the performance of village officials in terms of service quality to the community, productivity, responsiveness, responsibility and productivity. The development of a village information system in service activities in Tegal village is a transformation from manual to computerized, so systematic efforts are needed in the preparation involving subjects, objects and methods related to the transformation process. The development of the village information system uses the software development life cycle (SDLC). Efforts to control the quality of the Tegal Village Information System use four characteristics of ISO 9126, to know that the parts in the application system have correctly displayed error messages if an error occurs in inputting data. The result of this service activity is that every Village Apparatus can understand the material that has been submitted and can practice the results of the village administration work in a computerized manner based on the Village Information System.*

**Keywords – Village; Village Information System; Village Services; ISO9126, SDLC, Black Box.**

## I. INTRODUCTION

The village is a legal community unit that has territorial boundaries that are authorized to regulate and manage government affairs, the interests of the local community based on community initiatives, origin rights, and/or traditional rights that are recognized and respected in the government system of the Unitary State of the Republic of Indonesia [1]. Village Administration is the administration of government affairs and the interests of the local community in the government system of the Unitary State of the Republic of Indonesia. The main function of the village government is to serve the village community[2]. The village as the smallest administrative government in Indonesia which is tasked with carrying out services to the community is part of the implementation of e-government in Indonesia, it is required to be able to follow the development of information and communication technology in managing the village population administration data [3].

Tegal Village is a village located in Kemang District, Bogor Regency. The condition of Tegal village, Kemang sub-district, is currently in more dynamic village development. The Tegal village community has dynamic demands and always wants fast service to become a new problem faced by the Tegal village apparatus. Currently, the Tegal village apparatus still uses conventional services in direct contact with the community, especially in public services. Another problem faced by the Tegal village apparatus is communication skills using technology[4].

Communication is an intermediary for presenting information to the public [5]. Good communication is needed by the Tegal village apparatus so that any information can be conveyed to the Tegal village community. The need for communication media is one of the tools to make it easier for information to be conveyed properly to the public [6]. The communication media that will be built in Tegal village are the website media and the Village Information System (SID). Media websites and SID are media that can provide convenience for village officials in providing information on village performance to the community[7].

There are still many village population administration service systems that are conventional, resulting in village officials and villagers, wherein the power management process errors often occur caused by humans, wasting time and costs[8]. The Village Information System (SID) is an information system that changes raw data into ready-to-use information. In addition, SID will provide convenience to village officials in providing services to the community[9]. The development of this SID is expected to be able to provide acceleration for village government offices, especially Tegal village, Bogor district. so as to improve the performance of village officials in terms of service quality to the community, productivity, responsiveness, responsibility and productivity.

SID Tegal Village is built based on user needs. These needs are obtained by means of needs analysis. Analysis of application requirements used with systematic survey in accordance with the SDLC (System Development Cycle) method to all staff and community components in Tegal

Village. SDLC is a good method for a dynamic village information system[10], [11].

The development of the Tegal village SID was carried out in two stages of testing, namely testing using ISO 9126 and blackbox testing. Each test has a different function. Testing with ISO 9126 was conducted to determine Functionality, Reliability, Usability, and Efficiency. While testing black box serves to find out the results of input and output from the Tegal Village SID[12].

## II. RESEARCH METHOD

The development of a village information system in service activities in Tegal village is a transformation from manual to computerized, so in the preparation, systematic efforts are needed regarding the subjects, objects and methods associated with the transformation process. The development of the village information system uses the software development life cycle (SDLC) method [13] with the stages of Analysis, Design, Coding, Testing and Implementation.

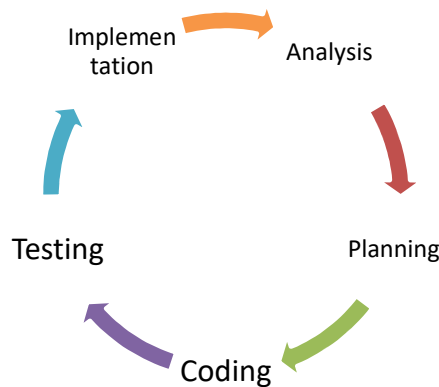


Figure 1 SDLC

### Analysis

The analysis phase of the development of the Village Information System with the survey method resulted in two criteria, namely the benefit criteria and the problem criteria. The criteria for benefits are in accordance with the objectives in realizing the ideals of getting used to processing and reading documents in digital form to accelerate village services. There are several benefits offered by the system, including 1) Time efficiency, 2) Better Documentation Management, 3) Better work comfort, 4) Support for better decisions, 5) More controllable management, 6) Improved organizational image.

The problem analysis criteria resulted in several sources of problems in Tegal village, Kemang sub-district, Bogor district, namely 1) Village documents have not been systematically documented, 2) The community is still difficult to access services and information about the village, 3) There is no Village Information system, 4) Lack of skills village apparatus in managing systematic data based on village information systems.

### Design

The design stage is carried out after generating a benefit analysis and problem analysis carried out in the previous stage. The design stage is described in detail as in Figure 2. Analysis of the problem after the survey is collected and the best solution is designed for each root cause with the aim of producing the desired benefits. The output of the system design to be built is an Integrated Village Information System and an increase in the quality of the village apparatus' ability to serve the community.

### Coding

The coding stage is the implementation stage of the design as expected. This stage is often referred to as design development. The designs developed at this stage are database system design, data flow diagram design, user experience design and user interface design using HTML, PHP and MySQL program code.

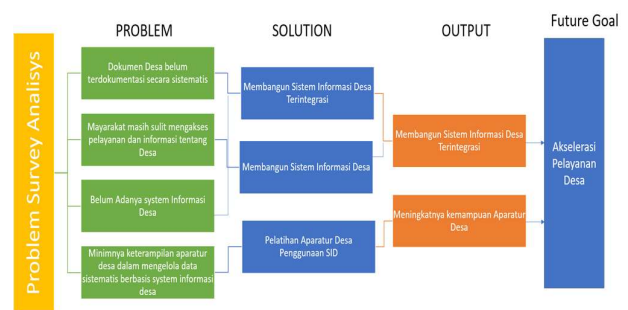


Figure 2 System Design Analysis

### Testing

The testing phase is carried out using Quality Control of the Village Information System using four characteristics of ISO 9126, namely functionality, reliability, usability and efficiency. The results of data analysis obtained from the questionnaire, there are quality control results using QC ISO 9126. Software Testing styles that are carried out are Blackbox Testing and Community Service Satisfaction Surveys.

### Implementation

Implementation of the last stage in the development of the village information system that is currently being built. This stage can be an early stage in further development and be the last stage if the system is not continued in the next development. Therefore, this stage is a very important stage to determine the efficiency generated after the system is run.

## III. Results and Discussion

The results of this service activity are in accordance with the desired output, namely an integrated village information system and increasing the ability of the village apparatus.



### Village Information System

The resulting village information system is a system built using HTML, PHP and MySql programming. Tegal Village Information System Display As shown in Figure 3

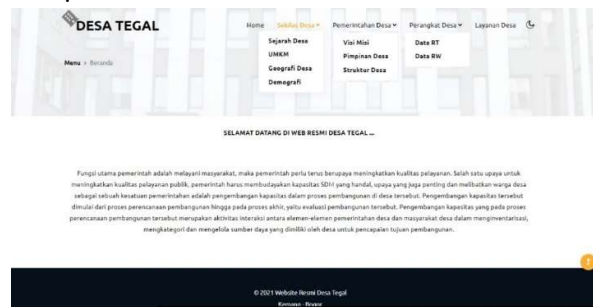


Figure 3 Tegal Village SID Display

Efforts to control the quality of the Tegal Village Information System use four characteristics of ISO 9126, namely functionality, reliability, usability and efficiency. The results of data analysis obtained from the questionnaire, there are quality control results using QC ISO 9162.

Tabel 1 Pengujian ISO 9126

Aspect	Real Score	Ideal Score	% Real Score	Criteria
Functionality	812	900	91,22%	Very Well
Reliability	416	500	85,2%	Very Well
Usability	735	800	93,87%	Very Well
Efficiency	238	300	77,33%	Good
<b>Total</b>	<b>2201</b>	<b>2500</b>	<b>89,04%</b>	Very Well

Based on table 1, it can be concluded that the quality of the Financial Transaction Information System for web-based SMEs is very good. the percentage is 89.04%. The highest quality aspect is based on the Usability aspect with a percentage of 93.87%, followed by the Functionality aspect at 91.22%. Reliability with a percentage of 85.2%, while the lowest quality aspect is the aspect of efficiency with a percentage of 77.33%. The conclusion from these results is that the system can run very well.

Efforts to control the quality of the Tegal village information system use the black box testing method with the aim of knowing that the parts in the application system have correctly displayed error messages if an error occurs in data input [14]. Black Box Testing is carried out to observe the results of execution through test data with the aim of checking the functionality of the software being built [15].

The tests used are 1) Equivalence Partitioning, which is entering data that does not match the data type or entering random data 2) Comparison Testing is seeing the system interface display on different web browsers, 3) Behavior Testing, is creating new data repeatedly to avoid the data stack and the system can accept data with a number of more than 50 4) Performance Testing is evaluating the program's ability to operate correctly in terms of memory consumption flow, data flow and execution speed. The memory usage test was carried out on a similar web browser. The results of the test are as in table 2.

Table 2 Blackbox System Results

Method	Input	Observation result	Output	Conclusion
Equivalence Partitioning	desategal.id	The inputted data has been successfully stored and can be displayed on the front end	Data appears on the front end	Succeed
Comparison Testing	desategal.id	Chrome = No Problem Edge = No Obstacle Mozilla = No Problem	SID Looks perfect on every web browser platform	Succeed
Behavior Test	desategal.id	The addition of user data and information data is successful.	Data appears on the front end	Succeed
Performance Test	desategal.id	The data that was input successfully performed very well	Data appears on the front end	Succeed

### Capacity Building for Village Apparatus

The result of this service activity is that every Village Apparatus can understand the material that has been submitted and can practice the results of village administration work in a computerized manner based on the Village Information System.

## IV. Conclusion

Tegal Village is a village located in Kemang District, Bogor Regency. The condition of Tegal village, Kemang sub-district, is currently in more dynamic village development. The Tegal village community has dynamic demands and always wants fast service to become a new problem faced by the Tegal village apparatus. Currently, the Tegal village apparatus still uses conventional services in direct contact with the community, especially in public services. Another problem faced by the Tegal village apparatus is communication skills using technology.

The development of this SID is expected to be able to provide acceleration for village government offices, especially Tegal village, Bogor district. so as to improve the performance of village officials in terms of service quality to the community, productivity, responsiveness, responsibility and productivity. The development of a village information system in service activities in Tegal village is a transformation from manual to computerized, so systematic efforts are needed in the preparation involving subjects, objects and methods related to the transformation process. The development of the village information system uses the software development life cycle (SDLC) method with the stages of Analysis, Design, Coding, Testing and Implementation.

The result of this service activity is that every Village Apparatus can understand the material that has been submitted and can practice the results of



computerized Village administration work based on the Village Information System.

### Acknowledgments

The author would like to thank the Directorate General of Higher Education in 2021 for the Independent Learning Campus Free Research program (MBKM) and Community Service based on research results and the PTS Prototype Cooperation of Pakuan University with the Directorate General of Higher Education in 2021

### References

- [1] D. Bender, "DESA - Optimization of variable structure Modelica models using custom annotations," *ACM Int. Conf. Proceeding Ser.*, vol. 18-April-2, no. 1, pp. 45–54, 2016, doi: 10.1145/2904081.2904088.
- [2] M. Praseptiawan, E. D. Nugroho, and A. Iqbal, "Pelatihan Sistem Informasi Desa untuk Meningkatkan Kemampuan Literasi Digital Perangkat Desa Taman Sari," *ABDIMAS J. Pengabd. Masy.*, vol. 4, no. 1, pp. 521–528, 2021, doi: 10.35568/abdimas.v4i1.1206.
- [3] R. Fitri, A. N. Asyikin, and A. S. B. Nugroho, "Pengembangan Sistem Informasi Desa Untuk Menuju Tata Kelola Desa Yang Baik (Good Governance) Berbasis Tik," *POSITIF J. Sist. dan Teknol. Inf.*, vol. 3, no. 2, pp. 99–105, 2017, doi: 10.31961/positif.v3i2.429.
- [4] T. M. Sahri and M. Paramita, "Pemberdayaan Masyarakat Melalui Zakat Infaq Shadaqoh Wakaf (Ziswaf) Dalam Meningkatkan Ekonomi Masyarakat Community Empowerment Through Zakat Infaq Shadaqoh Wakaf (Ziswaf) in Improving Community Economy," *J. Qardhul Hasan; Media Pengabd. Kpd. Masy. p-ISSN 2442-3726 e-ISSN 2550-1143*, vol. 6, pp. 121–126, 2020.
- [5] F. Rozi, T. Listiawan, and Y. Hasyim, "Pengembangan Website Dan Sistem Informasi Desa Di Kabupaten Tulungagung," *JUPI (Jurnal Ilm. Penelit. dan Pembelajaran Inform.)*, vol. 2, no. 2, pp. 107–112, 2017, doi: 10.29100/jupi.v2i2.366.
- [6] D. Ardiansyah, E. T. Tosida, and A. D. Waluyo, "Optimization of accounting information system reinforcing of tourism based small and medium enterprises (Smes)," *Int. J. Sci. Technol. Res.*, vol. 9, no. 3, pp. 1282–1286, Mar. 2020.
- [7] E. T. Tosida, A. D. Sanurbi, and A. WArtini, "Optimization of Indonesian Telematic SMEs Cluster: Industry 4 . 0 Challenge," in *IOP Conference Series*, 2017, p. 166, [Online]. Available: <http://iopscience.iop.org/article/10.1088/1757-899X/166/1/012017/meta>.
- [8] F. Matatula and Rosmiati, "PENGEMBANGAN WEBSITE DAN SISTEM INFORMASI DESA DI DESA PANGKAN KECAMATAN PAKU KABUPATEN BARITO TIMUR," *J. Sains Komput. dan Teknol. Inf.*, vol. 4, pp. 45–49, 2021.
- [9] A. Mardiasuti, "Evaluasi Terhadap Kualitas Pelayanan Publik Melalui Kajian Indeks Kepuasan Masyarakat (IKM) pada Unit Referensi Perpustakaan Universitas Gadjah Mada," *Berk. Ilmu Perpust. dan Inf.*, 2016, doi: 10.22146/bip.8835.
- [10] I. Journal and S. Engineering, "system development life cycle," vol. 2, no. 1, pp. 15–24, 2016.
- [11] P. Tooptompong and K. Piromsopa, "Using factor analysis techniques to identify SME information systems' functionality requirements," *International Journal of Interdisciplinary Organizational Studies*. 2018, doi: 10.18848/2324-7649/CGP/v12i03/13-30.
- [12] K. Moumane, A. Idri, and A. Abran, "Usability evaluation of mobile applications using ISO 9241 and ISO 25062 standards," *Springerplus*, 2016, doi: 10.1186/s40064-016-2171-z.
- [13] A. R. Otero and A. R. Otero, "System Development Life Cycle," in *Information Technology Control and Audit*, 2018.
- [14] U. Hanifah, R. Alit, and S. Sugiarto, "Penggunaan Metode Black Box Pada Pengujian Sistem Informasi Surat Keluar Masuk," *SCAN - J. Teknol. Inf. dan Komun.*, vol. 11, no. 2, pp. 33–40, 2016, [Online]. Available: <http://ejournal.upnjatim.ac.id/index.php/scan/article/view/643>.
- [15] B. A. Priyaungga, D. B. Aji, M. Syahroni, N. T. S. Aji, and A. Saifudin, "Pengujian Black Box pada Aplikasi Perpustakaan Menggunakan Teknik Equivalence Partitions," *J. Teknol. Sist. Inf. dan Apl.*, vol. 3, no. 3, p. 150, 2020, doi: 10.32493/jtsi.v3i3.5343.



# Design and Development E-Mading System for Information Students

Geovanne Farell<sup>1\*)</sup>, Igor Novid<sup>2</sup>, Sandi Rahmadika<sup>3</sup>

<sup>1</sup>Program Studi Teknik Informatika, Fakultas Teknik, Universitas Negeri Padang

<sup>2</sup>Program Studi Teknik Informatika, Fakultas Teknik, Universitas Negeri Padang

<sup>3</sup>Department of IT Convergence and Application Engineering, Pukyong National University

Email: <sup>1</sup>[geovannefarell@ft.unp.ac.id](mailto:geovannefarell@ft.unp.ac.id), <sup>2</sup>[igornovid@ft.unp.ac.id](mailto:igornovid@ft.unp.ac.id), <sup>3</sup>[sandika@pukyong.ac.kr](mailto:sandika@pukyong.ac.kr)

**Abstract** – This study aims to facilitate the provision of information and reduce the amount of paper used in the Department of Electrical Engineering, Faculty of Engineering, State University of Padang. With the large number of papers used to provide information on conventional e-mading, it causes too much paper to be wasted and is not very effective in providing information. With a lot of wasted paper also causes inefficient time. Because there is too much information and announcements that are reported to the Department of Electronic Engineering, the admin in the Department of Electrical Engineering, Faculty of Engineering, State University of Padang cannot manage. Therefore, for now it is necessary to have digital madding to manage information and reduce the amount of paper use. There are 2 methods used in this research, namely Data Collection Methods and System Development Methods. This data collection method uses interview and observation techniques, while the system development method uses USDP. In this study using the USDP method, where this method can perform object-oriented software development, design and analysis of software design using the UML (Unified Modeling Language) approach. The results of this study are a E-Mading Information System that can be used by students, lecturers and staff of the Department of Electronics Engineering.

**Keywords** – Web Based; Information System; Digital Madding

## I. INTRODUCTION

E-Mading is a type of information media that is the simplest and easiest to use to be a means of communication such as mass media. Every piece of information provided can be a good and appropriate marketing tool for information. Not only as a marketing tool, e-mading can also be used as a communication tool for students. This is an activity of a department or organization that can show dynamics[1]. Therefore, e-mading management must be done properly and correctly so that the development process can always be followed[2].

With the many developments in information technology making one of the main factors in the provision and dissemination of information so that the academic community including lecturers, engineering faculty staff and students are encouraged to use such technology. The development of information technology can also reach a process within the scope of the engineering faculty to improve the effectiveness and efficiency of performance[2][3].

One example is the efficient performance of the archives department because it still employs a conventional e-mading system. In a data collection method, several methods can be used, namely the method of observation, interviews and library research. In a USDP system development can be used as a reference object-oriented software development[4]. This UML was chosen as one of the tools to analyze the software requirements to suit what is desired. The design of a digital e-mading system is expected to be able to operate and maximize the system that has been running so that the information is right on target and can facilitate the management or access of information. Digital madding can also reduce the amount of excessive paper usage so that it can also support a government program in reducing the amount of paper usage[5].

The system is a unity, both real or abstract objects consisting of various components or elements that are

interrelated, interdependent, mutually supportive, and as a whole unit in one entity to achieve certain goals effectively and efficiently. A system consists of various interconnected subsystems, to get the output of a process into information, then input in the form of data is needed. This definition is a collection of several interconnected components in achieving goals[6]. This system itself consists of several groups in its definition is the emphasis on components and emphasis on procedures. In this case the system can also be defined to be an interconnected whole. Data in the form of raw materials if not processed will not be useful. The data can produce information if it is processed through a model. The model that processes the data is called the data processing cycle[7].

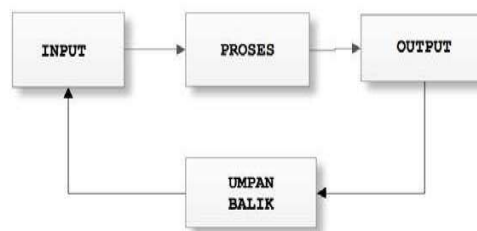


Figure. 1 Data Processing Cycle

Fig. 1 describes an event that defines reality through the input element. After that, the data is processed into an output or information that will be used. An information is received by the user, then there is a response in the form of evaluation and the results of the response will be data that will be entered into input again, and so on. The quality of information depends on three things: accurate, timely and relevant[8]. A simple communication medium on campus is a e-mading. E-mading has many functions as the role of one of the student activity facilities, including



communicative, informative, creative and creative[9]. E-mading plays a role in the formation and formation of students, both in the aspects of knowledge skills/abilities, interests and talents and attitudes. Most of the writers use e-mading as a place to practice. Start from the habit of writing simple things, and in the end the insight will be open to having an interest in developing the writing in-depth. The database is an important part of an information system because it can make a basis in providing information for its users[10]. Databases that determine a quality of information generated by a system, because a system will change or make the database as a center for holding all data has been arranged so that users can add, change and reduce the data needed in accordance with the wishes or goals in making digital madding[11]. The purpose of establishing a database in a company is to facilitate the retrieval of data as shown in Fig. 2 that the database can replace and modify the file cabinet save lots of documents.

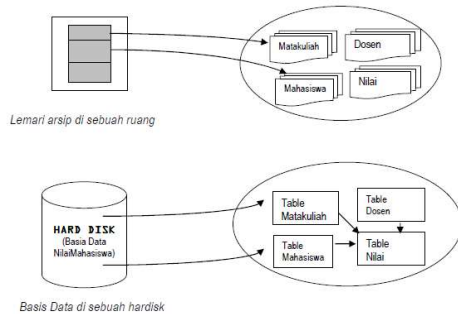


Figure. 2 Archive and database

Some modeling of a system that can be used, one of which is a modeling device for the Unified Modeling Language that is used for systems that are the object-oriented paradigm[12]. According to Jesse (2003: 16) To improve the results of development and productivity, can use UML. UML was created for object-oriented system design. Basically, modeling is done to simplify complex problems so that it can be more easily learned and understood. To understand UML, it takes a form of concept from a modeling language, and learn the three main elements of the UML, namely building blocks, rules that state how a building block can be put together and some common mechanisms[13].

## II. RESEARCH METHODOLOGY

In this study using the type of research development. Development research is a type of research that has a goal in developing and producing a product. The products that can be produced in this research are digital madding applications. Application development design uses the main tool in designing software applications, namely UML (Unified Modeling Language)[14]. UML has the following stages:

### 1. Use Case Diagram

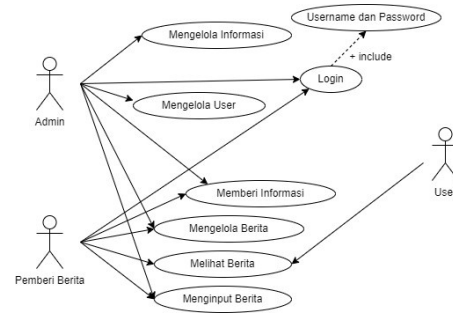


Figure. 3 The step of Use Case Diagram

In the Use Case diagram above there is an admin who can manage the input and output processes of data and processes on the system. Announcers can provide information to the admin so that information can be disseminated through digital e-mading[15]. While the User can see the information available on the digital e-mading whose data has been managed by the admin.

### 2. Class Diagram

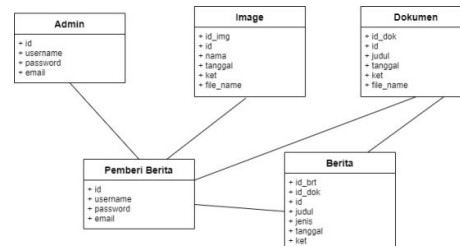


Figure. 4 Example of Class Diagram

In the Class diagram above are the details of the class-class relationships found in digital applications available at the Faculty of Engineering (FT) of the Universitas Negeri Padang (UNP).

## III. RESULTS AND DISCUSSION

### 1) Main Page

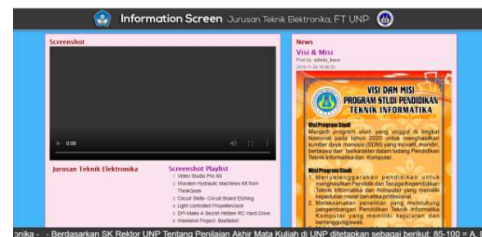


Figure. 5 Main Page

The above display is the main page that will appear on digital madding. This digital madding has a header, body, and footer. In the body there are 2 pages:

- Screenshot Page
- News Page

The News page contains the order of information or the latest announcements being displayed. While the Screenshot Page is a video display that will be displayed.



2) Admin Page

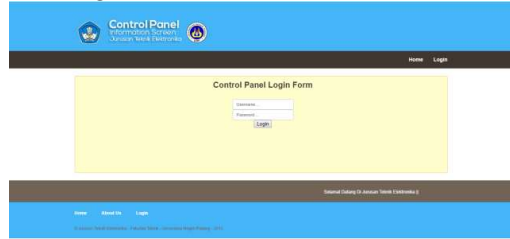


Figure. 6 Admin Page

The view above is the admin page. This page can only be accessed by admin which contains all data that can only be managed by the system.

3) Menu Home

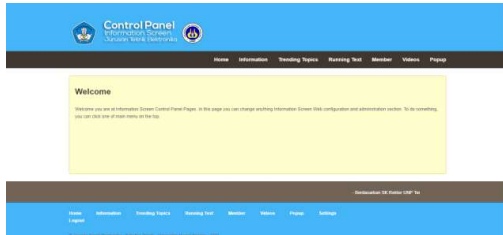


Figure. 7 Home Menu

This is the Home Menu display on the Admin Page. This display will appear when the admin is successful in logging in by inputting the username and password.

4) Menu Information

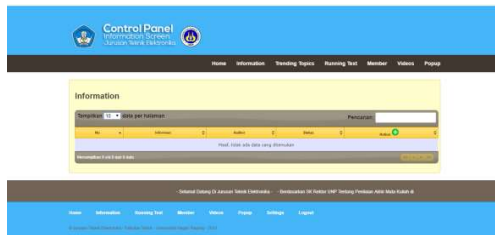


Figure. 8 Information Menu

This display is a display of input information that will be displayed on the start page of e-mading. Admin can add the latest information that will be displayed on digital madding through this Information Menu.

5) Menu Trending Topics



Figure. 9 Trending Topic Menu

This display is the input news display that will be

displayed on the e-mading home page. Admin can add the latest news that will be displayed on digital madding through this Trending Topic Menu.

6) Menu Running Text



Figure. 10 Running Text Menu

The running text in the footer located on the Home page can be added and changed by admin in this menu.

7) Menu Member



Figure. 11 Member Menu

In this menu, the admin can add and even edit the list of digital madding members.

8) Menu Videos



Figure. 12 Videos Menu

In this menu the admin can add a list of videos that will available on digital e-mading.

9) Menu Popup



Figure. 13 Popup Menu

Admin can add a pop-up image on this menu.



#### IV. CONCLUSION

Based on the results and discussion obtained, it can be concluded :

1. This research has succeeded in achieving the goal of making digital e-mading that can be used by information users to be able to access information contained on the website.
2. With the digital bulletin board, the problem of delivering information through a e-mading (bulletin board) namely the limited space of information that can be loaded can be resolved.

#### REFERENCES

- [1] Indera and H. Ramasudha, "Sistem Informasi Elektronik Mading (E-Mading) Ukm Dan Fakultas Ilmu Komputer Pada Ibi Darmajaya Berbasis Android," *Tenika*, vol. 12, no. x, pp. 57–63, 2017.
- [2] M. J. Kim, J. S. Park, and H. K. Chin, "An influence of the digital native characteristics to acceptance of digital signage engagement and media attitude," *Indian J. Sci. Technol.*, vol. 8, no. 23, 2015, doi: 10.17485/ijst/2015/v8i23/79206.
- [3] S. Handoko, "Penerapan Media Sosial Pada Papan Informasi Digital Interaktif," pp. 544–551, 2018.
- [4] P. Thanthirige *et al.*, "No 主観的健康感を中心とした在宅高齢者における健康関連指標に関する共分散構造分析Title," vol. 13, no. August, pp. 50–60, 2016.
- [5] Z. Hajah, D. Darlis, and D. A. Nurmantris, "IMPLEMENTASI MADING ONLINE BERBASIS WEB MENGGUNAKAN FRAMEWORK LARAVEL DI SDN 05 SURABAYO Implementation Online Mading Based on Web Using Laravel Framework in SDN 05 Surabaya," vol. 7, no. 6, pp. 3180–3189, 2021.
- [6] A. Voutama and E. Novalia, "Perancangan Aplikasi M-Magazine Berbasis Android Sebagai Sarana Mading Sekolah Menengah Atas," *J. Tekno Kompak*, vol. 15, no. 1, p. 104, 2021, doi: 10.33365/jtk.v15i1.920.
- [7] R. E. G. Rahayu and Z. K. Pujaeri, "Rancang Bangun Sistem Informasi Absensi Fingerprint, Agenda, Mading Digital di SMK Wikrama 1 Garut Berbasis Web," *J. Algoritm.*, vol. 17, no. 2, pp. 561–568, 2021, doi: 10.33364/algoritma/v.17-2.561.
- [8] T. W. P. T, U. H. Medan, J. H. M. Joni, and C. No, "Dan Komputer Universitas Harapan Medan," pp. 160–169, 2020.
- [9] R. B. Aminullah, D. Darlis, and D. A. Nurmantis, "E-Mading Berbasis Website Menggunakan Raspberry Pi E-Mading Based Website Use Raspberry Pi," *e-Proceeding Appl. Sci.*, vol. 6, no. 2, pp. 2294–2300, 2020, [Online]. Available: <https://openlibrarypublications.telkomuniversity.ac.id/index.php/appliedscience/article/view/13414>.
- [10] H. Ramasudha, B. E. Mahasiswa, M. P. Mahasiswa, and H. Mahasiswa, "Sistem Informasi Elektronik Mading ( E-Mading ) UKM dan Fakultas Ilmu Komputer IIB Darmajaya," vol. 12, no. x, pp. 1–7, 2018.
- [11] T. I. Sri Haryanti, Bambang Eka Purnama, "Rancang Bangun Sistem Informasi E-Mading Manajemen Masjid," *Vol 2 No 2*, pp. 630–638, 2012.
- [12] M. Kegiatan, B. Mengajar, P. Sma, and C. Islami, "Kata Kunci :," vol. 3, no. 1, pp. 1–5, 2015.
- [13] P. Pelaku, U. Mikro, K. D. Menengah, R. Jumardi, and A. Nugroho, "Aplikasi Mading Digital Sebagai Media Promosi," vol. 4, no. 6, pp. 501–508, 2021.
- [14] D. Komalasari and I. Solikin, "Desain Aplikasi E-Mading pada Sekolah MA Miftahul Huda Tugu Agung," *Pros. Semin. Nas. Darmajaya*, vol. 1, no. 1, pp. 27–34, 2018, [Online]. Available: <https://jurnal.darmajaya.ac.id/index.php/PSND/article/view/1226>.
- [15] Lila Setiyani, J. A. Haris, and E. Tjandra, "Rancang Bangun Papan Informasi Digital (Digital Signage) Berbasis Web Menggunakan Sistem Operasi Linux dengan Server NGINX pada STMIK Rosma Karawang," *Metik J.*, vol. 4, no. 2, pp. 83–91, 2020, doi: 10.47002/metik.v4i2.185.



# Naive Bayes and Support Vector Machine Algorithm for Sentiment Analysis Opensea Mobile Application Users in Indonesia

Laurenzius Julio Anreaja<sup>1</sup>, Norma Nobuala Harefa<sup>2</sup>, Julius Galih Prima Negara<sup>3</sup>, Venantius Nathan Hermanu Priyantara<sup>4</sup>, Agung Budi Prasetyo<sup>5</sup>.

<sup>1</sup>Information Systems Study Program, Faculty of Industrial Technology, Atma Jaya University Yogyakarta

<sup>2</sup>Information Systems Study Program, Faculty of Industrial Technology, Atma Jaya University Yogyakarta

<sup>3</sup>Departement of Informatics, Atma Jaya University Yogyakarta

<sup>4</sup>Information Systems Study Program, Faculty of Industrial Technology, Atma Jaya University Yogyakarta

<sup>5</sup>Information Systems Study Program, Faculty of Industrial Technology, Atma Jaya University Yogyakarta

Email: <sup>1</sup>laurenziusjulio11@gmail.com, <sup>2</sup>normanharefa@gmail.com, <sup>3</sup>julius.galih@uajy.ac.id,

<sup>4</sup>venantiusnathan@gmail.com, <sup>5</sup>Agungpraset54@gmail.com

**Abstract** –Opensea is an NFT buying and selling application-based platform that is booming in the community. One way to find out the public's perception of the Opensea application is by sentiment analysis, as done in this study. Data that is used is user review data for the Opensea application in the Indonesian play store. The sentiment analysis technique used is the Naïve Bayes Classifier and the Support Vector Machine (SVM) method. Both are used to compare public responses from sentiment analysis of reviewed data labeled as positive, negative, and neutral. Based on this study, it was found that the Naive Bayes algorithm gives the results that class precision is 87.31%, class recall is 71.02%, and accuracy is 89.81%. While the SVM algorithm gives the results that class precision is 94.23%, class recall 71.96%, and Accuracy 90.78%. It is concluded that the SVM algorithm has a better performance than the Naive Bayes algorithm.

**Keywords** – Opensea, NFT, Sentiment Analysis, Google play store, SVM, Naive Bayes

## I. INTRODUCTION

Opensea application users began to boom in Indonesia in early 2021 because a student from Semarang City, Sultan Gustaf Al Ghozali, got 1.5 billion from the sale of selfie photos in the form of NFT entitled Ghozali Everyday on the OpenSea platform [1]. The OpenSea Marketplace is another place where collectibles sold for Decentraland can be found. The Decentraland collection is linked to NFT. An NFT is a digital token that functions as a digital certificate of ownership for a digital asset such as an artist's digital collection or image. [2].

OpenSea has recorded \$20.37 billion in sales and has over 1.2 million active traders in its network [3].

We opted to collect sales data from OpenSea, the largest active NFT marketplace. Statistics show that OpenSea has accumulated over \$20 billion in trade volume, boasting more than 1.2 million traders. The data set was built using the provided OpenSea API, where we made our queries against the Events endpoint [4].

Play Store is a digital content provider service owned by Google that provides various online product stores such as applications, games, movies or music, and books of various categories. Google Play Store can be accessed through the website, android application, and Google TV. In the Google Play Store application, there are several features, one of which is the rating and review feature from users of available applications or services. A review or review is a text or sentence that contains an assessment or comment on a person's work. The importance of these reviews is often

used as a benchmark for an application, whether it is recommended or not for new users [5].

Sentiment analysis or opinion mining is the process of understanding, extracting, and processing textual data automatically to obtain sentiment information contained in an opinion sentence. Sentiment analysis is carried out to see opinions or opinion tendencies towards a problem or object by someone, whether they tend to have negative or positive views or opinions [6].

In this study, sentiment analysis was carried out to see reviews from users of the OpenSea application. These reviews could be put into three categories, namely positive, neutral and negative.

Many studies have used machine learning algorithms with support vector machines (SVM) and Naïve Bayes (NB) being the most commonly used. Naïve Bayes (NB) is a technique based on Bayes' theorem. The Naive Bayes algorithm assumes that the presence of certain features in a class does not correlate with the presence of other features. This model is easy to build and very useful for very large data sets. Despite its simplicity, Naive Bayes is known to outperform even the most complex classification methods [7].

Support Vector Machine (SVM) is a classification and regression method commonly used for linear and non-linear problems. It has the advantage of applying linear splits to high-dimensional non-linear input data, and this is achieved by using the required kernel functions. The effectiveness of the Support Vector Machine is strongly influenced by the type of kernel function selected and



applied based on the characteristics of the data. Many studies have reported that the Support Vector Machine is the most accurate method for text classification [8].

In previous research regarding the analysis of sentiment on E-Wallet Review (OVO). This study uses 500 positive reviews and 500 negative reviews as training data. The results of this study indicate that the use of the Naive Bayes algorithm produces an accuracy value of 93.10 percent. In comparison, the research results from the SVM algorithm are 91.30 percent. Based on these results, the accuracy value generated by the Naive Bayes algorithm and SVM was found that SVM is the best algorithm for classifying [9]. Also, previous research regarding the sentiment analysis of the Indonesian Police Mobile Brigade Corps based on Twitter posts using the SVM and NB methods resulted in an accuracy value of 86.96% with the SVM approach, 86.96% precision value, and 86.96% recall value [10].

The purpose of this study is to predict sentiment labels on reviews from users of the OpenSea application on the Google Play Store using the Naive Bayes method and Support Vector Machine as a classification model.

## II. RESEARCH METHODOLOGY

The object of this research is the Indonesian people's tweets against the metaverse on Twitter social media. In this study, there are several steps taken in analyzing the sentiments of the Indonesian people towards metaverse technology. The steps taken in this research can be seen in figure 1.

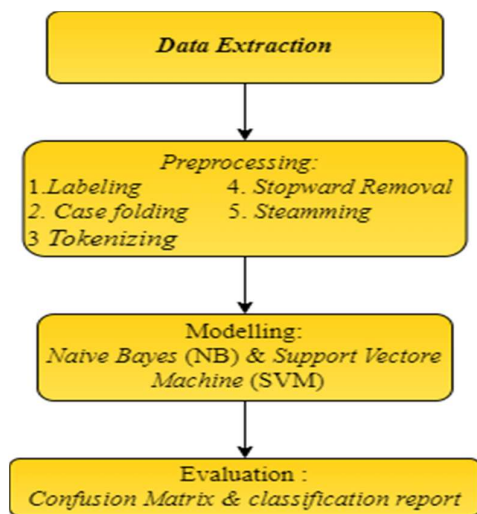


Figure 1. Research Process

### 2.1 Data Extraction

Collecting data in this study was obtained from reviews of users of the OpenSea application on the Google Play Store, using the web scraping technique, namely the technique which is used to extract data in large quantities large than the website where the data already extracted is saved in CSV (Comma Separated Value) format [11]. The web scraper process in the google play store uses the

Google-play-scraper. Google-play-scraper is Node js module to scrape application data from the Google Play store [12]. The data is then processed using the python language to go to the next stage, which is preprocessing.

### 2.2 Preprocessing Data

Pre-processing is a stage that is carried out to process and improve data so that it can be processed after the data to be analyzed has been obtained [13]. The following are:

1) Labeling the data in this study will be carried out by two people. The first person is tasked with manually classifying positive, negative, and neutral sentiments, while the second person re-examines the correctness of the classification results that have been carried out by the first person.

2) Case Folding is the stage to change sentences that have uppercase (capital letters) into lowercase (lowercase). This is done in order to obtain structured and consistent data in the use of capital letters.

3) Tokenizing is the stage to separate sentences into several pieces of words called tokens. Separate words using space punctuation restrictions. The following is an example of tokenizing in the table.

4) Stop word removal is a step to get rid of various useless words in a sentence with the help of the Sastrawi library. Sastrawi library is a library that can also be used to perform stopword removal with unimportant words in Indonesian [14].

5) Stemming is a step taken by researchers to remove prefixes and suffixes for each token with the help of Sastrawi stemmer. Sastrawi stemmer is a stemmer library that is used to overcome the problem of changing words with words into basic words in the Indonesian language [15].

### 2.3 Modeling

Modeling is a method in which a model represents correlation relationships between one set of data and the other set of data [16]. The first step in starting data modeling is to partition or divide the data into training data and testing. The data modeling process for the case of sentiment analysis is carried out using several classification methods, including supervised learning, such as Support Vector Machine (SVM) and Naive Bayes. This algorithm was chosen because it is a commonly used method in sentiment analysis.

Naive Bayes Classifier is the probability used to determine text document class and can process large amounts of data with high accuracy results [17]. SVM (Support Vector Machine) is a machine learning algorithm that is used to divide each data class to find the most optimal hyperplane [18]. The SVM algorithm tries to find a hyperplane to maximize the distance between classes. In this way, SVM can guarantee the ability of high



generalization for data that will be predicted [19].

### 2.4 Evaluation

The evaluation process in this research uses a confusion matrix and classification report. The confusion matrix is a table that is used to describe the performance of a classification algorithm. A confusion matrix visualizes and summarizes the classification algorithm's performance in label comparison and machine learning prediction results [20]. Classification reports are used to measure the predictive quality of the classification of a particular algorithm so as to show the precision, recall, and accuracy of an application of the model algorithm [21]. The aim is to see and compare the accuracy, precision, and recall of SVM and Naive Bayes models in analyzing sentiment.

## III. RESULTS AND DISCUSSION (10 pt, Capital, Bold)

### 3.1 Data Extraction

The dataset that will be used in the research is taken from user reviews of the Opensea application on the play store by doing web scrapping via google-play-scrapper and python language. The dataset collected reviews in Indonesian as many as 1028 reviews.

### 3.2 Preprocessing

#### 3.2.1 Labelling

The dataset obtained is then carried out manually, labeling the sentiment by two people. Where the first person gives the label and the second person checks the correctness of the labeling. The results of the labeling obtained 731 positive sentiments, 231 negative sentiments, and 86 neutral sentiments. The results of the labeling stage can be seen in table 1.

Table 1. Labeling

Review	Sentimen
Mantap gw Dapet 3 ETH Berkat aplikasi ini	Positive
KOK NFT SAYA HILANG ATAU BERKURANG.. TIDAK ADA PENJELASAN DARI PIHAK OPENSEA	Negative

#### 3.2.2 Case Folding

For the dataset that has gone through the labeling process then, every uppercase letter in the comments column will be changed to lowercase, and the number will

be removed. The results of the case folding process can be seen in Table 2.

Table 2. Case Folding

Review	Sentimen
mantap gw dapet eth berkat aplikasi ini	Positive
kok nft saya hilang atau berkurang.. tidak ada penjelasan dari pihak opensea	Negative

#### 3.2.3 Tokenizing

At this stage, the sentence is broken down into words with punctuation and whitespace boundaries. The results of the tokenizing process can be seen in Table 3.

Table 3. Tokenizing

Review	Sentimen
mantap gw dapet eth berkat aplikasi ini	Positive
kok nft saya hilang atau berkurang tidak ada penjelasan dari pihak opensea	Negative

#### 3.2.4 Stopword Removal

After the dataset goes through the tokenizing process, the next step is to delete words that are not important and interfere with the sentiment analysis process through the Stopword Removal stage. The results of this stage can be seen in table 4.

Table 4. Stopword Removal

Review	Sentimen
mantap gw dapet eth berkat aplikasi	Positive
kok nft hilang berkurang penjelasan pihak opensea	Negative

#### 3.2.5 Stemming

The data preprocessing process is then ended by removing the affixes for each word so that the resulting words are



only the basic words. The results of this stage can be seen in Table 5.

Table 5. Stemming

Review	Sentimen
mantap gw dapet eth berkat aplikasi	Positive
kok nft hilang kurang penjelasan dari pihak opensea	Negative

### 3.3 Data Modelling

The training and testing data used in this data modeling is 80%: 20%. This means that from 1028 the training data collection owned is 822 records while the testing data owned is 206 records. Based on the results of the tests conducted on the Opensea application, user comment test data, which consists of 3 labels, namely positive, negative, and neutral, using the Naive Bayes classifier obtained a match accuracy with the train data of 89.81%. Meanwhile, using the Support Vector Machine algorithm, the accuracy was 90.78%. This means that the Naive Bayes model is more accurate than SVM in this study.

The visualization of the bar chart of the number of positive, negative, and neutral sentiments from the Support Vector Machine can be seen in table 8, and the distribution of the most dominant words in the positive, neutral, and negative labels are presented in the form of a word cloud. The word cloud in the positive class is shown in figure 9, while the negative class word cloud is shown in figure 10.

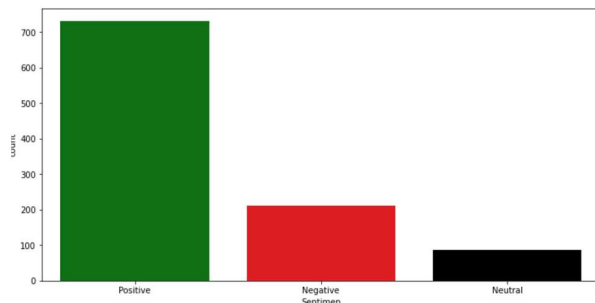


Figure 2. Distribution of the results of the analysis using the Support Vector Machine



Figure 3. Positive words

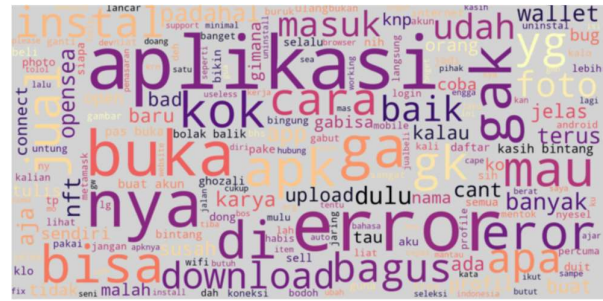


Figure 4. Negative Words

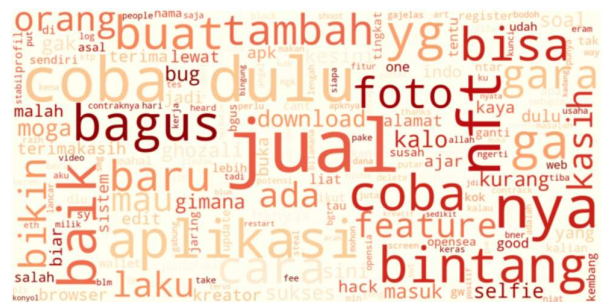


Figure 5. Neutral Words

### 3.4 Evaluation

After the model is created, it needs to be evaluated using a confusion matrix. Evaluation is done using confusion matrix so we can know the exact result of true positive, true negative, true neutral false positive, false negative, and false neutral. True positive, true negative, true neutral false positive, false negative, and false neutral. True positive is the successful positive class classified as the positive class, the true negative is the successful negative class classified as the negative class, and true neutral is the successful neutral class classified as the positive class. false positive is a negative class, and neutral class is classified as a positive class, false negative is a positive class, and neutral class is classified as a negative class. False neutral is a negative class, and a positive class is classified as a neutral class. The classification report is used to determine the class recall and class precision on a model that is being run.

In the evaluation of the naive bayes model with the confusion matrix, the results obtained the results of the true Positive = 143, false positive = 15, true negative = 37, false negative = 5, true neutral = 5, and false neutral = 1. The results of the confusion matrix can be seen in Figure 6.



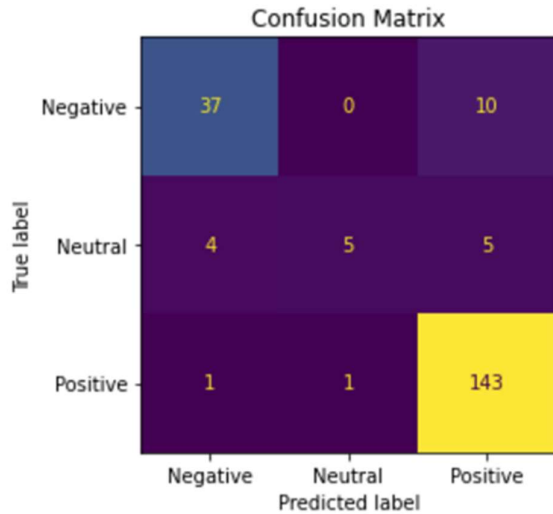


Figure 6. Confusion Matrix Naive Bayes

While the classification report in Naive Bayes shows class precision as negative, neutral, and positive is 88.10%, 83.33%, and 90.51%, while class recall is negative, positive, neutral that is 78.72%, 35.71%, 98.62%, so the results obtained are an average class precision of 87.31%, an average class recall of 71.02%, an accuracy of 89.81%. The results of the Classification report can be seen in figure 7.

	precision	recall	f1-score	support
Negative	0.8810	0.7872	0.8315	47
Neutral	0.8333	0.3571	0.5000	14
Positive	0.9051	0.9862	0.9439	145
accuracy			0.8981	206
macro avg	0.8731	0.7102	0.7585	206
weighted avg	0.8947	0.8981	0.8881	206

Figure 7. Classification Report Naive Bayes

Table 6 below is a combination of the results of the confusion matrix with the classification report on the Naive Bayes evaluation and is shown in tabular form so that it is easy to see the correlation.

Table 6. Summary of Confusion Matrix and Classification Report Naive Bayes

	True Negative	True Neutral	True Positive	Precision
Pred Negative	37	4	1	88.10%
Pred Neutral	0	5	1	83.33%
Pred Positive	10	5	143	90.51%

Recall	78.72%	35.71%	98.62%	
--------	--------	--------	--------	--

In the evaluation of the SVM model with the confusion matrix, the results of the true positive = 144, false positive = 16, true negative = 38, false negative = 3, true neutral = 5, and false neutral = 0. The results of the confusion matrix can be seen in figure 8.

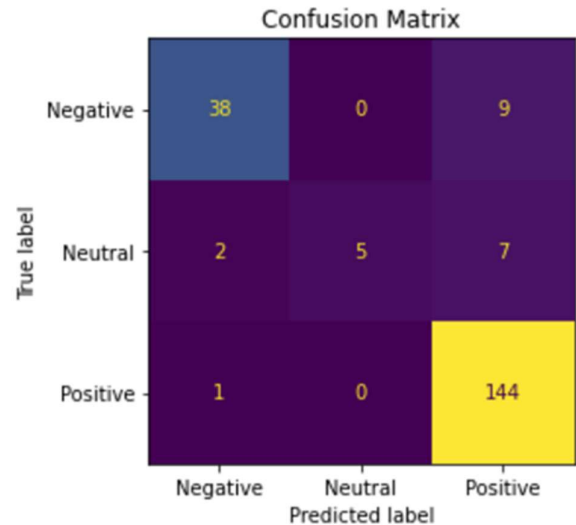


Figure 8. Confusion Matrix SVM

While the classification report in SVM shows Class precision as negative, neutral, and positive that is 92.68%, 100.00%, and 90.00% while Class recall is negative, neutral, positive is 80.85%, 35.71%, 99.31%, so the results obtained are an average class precision of 94.23%, an average class recall of 71.96%, and an accuracy of 90.78%. The results of the Classification report can be seen in figure 9.

	precision	recall	f1-score	support
Negative	0.9268	0.8085	0.8636	47
Neutral	1.0000	0.3571	0.5263	14
Positive	0.9000	0.9931	0.9443	145
accuracy			0.9078	206
macro avg	0.9423	0.7196	0.7781	206
weighted avg	0.9129	0.9078	0.8975	206

Figure 9. Classification Report SVM

The combination of the results of the confusion matrix with the classification report on the Support Vector Machine evaluation is shown in tabular form so that it is easy to see the correlation. The result can be seen in table 7.

Table 7. Summary of Confusion Matrix and Classification Report SVM

	True Negative	True Neutral	True Positive	Precision
--	---------------	--------------	---------------	-----------





<b>Pred Negative</b>	38	2	1	92.68%
<b>Pred Neutral</b>	0	5	0	100.00%
<b>Pred Positive</b>	7	9	144	90.00%
<b>Recall</b>	80.85%	35.71%	99.31%	

Based on the results of the comparison of the SVM and Naive Bayes algorithms, In table 8, the Naive Bayes algorithm gives the results that class precision is 87.311%, class recall is 71.02%, and accuracy is 89.81%. While the SVM algorithm gives the results that class precision is 94.23%, class recall 71.96%%, and accuracy 90.78%.The result can be seen in table 8.

Table 8. Performance comparison between Naive Bayes and SVM

	<b>Accuracy</b>	<b>Precision</b>	<b>Recall</b>
<b>Naïve Bayes</b>	89.81%	87.31%	71.02%
<b>SVM</b>	90.78%	94.23%	71.96%

#### IV. CONCLUSION

Based on the results of the sentiment analysis in this study, it can be seen that the Opensea Application Review dataset predicted using the Naïve Bayes algorithm and SVM showed significant results. The Naive Bayes algorithm gives the results that class precision is 87.31%, class recall is 71.02%, and accuracy is 89.81%. While the SVM algorithm gives the results that class precision is 94.23%, class recall 71.96%%, and accuracy 90.78%. It is concluded that the SVM algorithm has a better performance than the Naive Bayes algorithm. This research has not compared the performance with other machine learning algorithms besides Naive Bayes and SVM, so it is necessary to make a comparison with other classification machine learning algorithm models. Such as lexicon, linear regression, and random forest so that later it can improve the accuracy of sentiment classification in similar research.

#### REFERENCES

[1] Natasya Salim, "Minat Terhadap NFT Bertambah

Sejak Nama Ghazali Viral, Pakar Serukan Adanya Regulasi," 2022.

<https://www.abc.net.au/indonesian/2022-01-19/minat-nft-di-indonesia-meningkat-tapi-waspada-risiko-kejahatan/100765112>.

[2] Mohamed-amine et all, "How should metaverse augment humans with disabilities?," in *13th Augmented Human International Conference Proceedings*, 2022, p. 9, [Online]. Available: <https://archive-ouverte.unige.ch/unige:160466>.

[3] DeepRadar, "NFT MarketPlace Ranking," 2022. <https://dappradar.com/nft/marketplaces> (accessed Jul. 11, 2022).

[4] B. White, *Characterizing the OpenSea NFT Marketplace*, vol. 1, no. 1. Association for Computing Machinery, 2021.

[5] S. A. Aaputra, "Sentiment Analysis Analisis Sentimen E-Wallet Pada Google Play Menggunakan Algoritma Naive Bayes Berbasis Particle Swarm Optimization," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 3, no. 3, pp. 377–382, 2019.

[6] I. Rozi, S. Pramono, and E. Dahlan, "Implementasi Opinion Mining (Analisis Sentimen) Untuk Ekstraksi Data Opini Publik Pada Perguruan Tinggi," *J. EECCIS*, vol. 6, no. 1, pp. 37–43, 2012.

[7] D. D. Tran, T. T. S. Nguyen, and T. H. C. Dao, "Sentiment Analysis of Movie Reviews Using Machine Learning Techniques," *Lect. Notes Networks Syst.*, vol. 235, no. December 2017, pp. 361–369, 2022, doi: 10.1007/978-981-16-2377-6\_34.

[8] I. Santoso, Windu Gata, and Atik Budi Paryanti, "Penggunaan Feature Selection di Algoritma Support Vector Machine untuk Sentimen Analisis Komisi Pemilihan Umum," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 3, no. 3, pp. 364–370, 2019, doi: 10.29207/resti.v3i3.1084.

[9] I. Ajzen, "The theory of planned behavior," *Organ. Behav. Hum. Decis. Process.*, vol. 50, no. 2, pp.



- 179–211, Dec. 1991, doi: 10.1016/0749-5978(91)90020-T.
- [10] Sularso et al, “Sentiment Analysis of the Indonesian Police Mobile Brigade Corps Based on Twitter Posts Using the SVM And NB Methods,” *J. Phys. Conf. Ser.*, vol. 1201, 2019, [Online]. Available: [https://www.researchgate.net/publication/333585160\\_Sentiment\\_Analysis\\_of\\_the\\_Indonesian\\_Police\\_Mobile\\_Brigade\\_Corps\\_Based\\_on\\_Twitter\\_Posts\\_Using\\_the\\_SVM\\_And\\_NB\\_Methods/](https://www.researchgate.net/publication/333585160_Sentiment_Analysis_of_the_Indonesian_Police_Mobile_Brigade_Corps_Based_on_Twitter_Posts_Using_the_SVM_And_NB_Methods/).
- [11] R. Hanifah and I. S. Nurhasanah, “Implementasi Web Crawling Untuk Mengumpulkan Web Crawling Implementation for Collecting,” *J. Teknol. Inf. dan Ilmu Komput.*, vol. 5, no. 5, pp. 531–536, 2018, doi: 10.25126/jtiik20185842.
- [12] Latif et al, “Data Scraping from Google Play Store and Visualization of its Content for Analytics,” 2019, [Online]. Available: [https://www.researchgate.net/publication/342636207\\_Data\\_Scraping\\_from\\_Google\\_Play\\_Store\\_and\\_Visualization\\_of\\_its\\_Content\\_for\\_Analytics](https://www.researchgate.net/publication/342636207_Data_Scraping_from_Google_Play_Store_and_Visualization_of_its_Content_for_Analytics).
- [13] D. P. Sari, “Pemanfaatan NFT Sebagai Peluang Bisnis Pada Era Metaverse,” *J. Akrab Juara*, vol. 7, no. 1, pp. 237–245, 2022, [Online]. Available: <https://dspace.uin.ac.id/handle/123456789/29069>.
- [14] B. Siswanto, “Sentiment Analysis in Indonesian on Jakarta Culinary as A Recommender System,” 2021, [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/9702772>.
- [15] M. A. Rosid, A. S. Fitriani, I. R. I. Astutik, N. I. Mulloh, and H. A. Gozali, “Improving Text Preprocessing for Student Complaint Document Classification Using Sastrawi,” *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 874, no. 1, 2020, doi: 10.1088/1757-899X/874/1/012017.
- [16] Will Koehrsen, “Modeling: Teaching a Machine Learning Algorithm to Deliver Business Value,” 2018. <https://towardsdatascience.com/modeling-teaching-a-machine-learning-algorithm-to-deliver-business-value-ad0205ca4c86> (accessed May 11, 2022).
- [17] Pristiyono, M. Ritonga, M. A. Al Ihsan, A. Anjar, and F. H. Rambe, “Sentiment analysis of COVID-19 vaccine in Indonesia using Naïve Bayes Algorithm,” *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 1088, no. 1, p. 012045, 2021, doi: 10.1088/1757-899x/1088/1/012045.
- [18] M. Bloodgood, “Support Vector Machine Active Learning Algorithms with Query-by-Committee Versus Closest-to-Hyperplane Selection,” *Proc. - 12th IEEE Int. Conf. Semant. Comput. ICSC 2018*, vol. 2018-Janua, no. 2, pp. 148–155, 2018, doi: 10.1109/ICSC.2018.00029.
- [19] S. Fransiska and A. Irham Gufroni, “Sentiment Analysis Provider by.U on Google Play Store Reviews with TF-IDF and Support Vector Machine (SVM) Method,” *Sci. J. Informatics*, vol. 7, no. 2, pp. 2407–7658, 2020, [Online]. Available: <http://journal.unnes.ac.id/nju/index.php/sji>.
- [20] A. Kulkarni, D. Chong, and F. A. Batarseh, “Foundations of data imbalance and solutions for a data democracy,” *Data Democr. Nexus Artif. Intell. Softw. Dev. Knowl. Eng.*, pp. 83–106, Jan. 2020, doi: 10.1016/B978-0-12-818366-3.00005-8.
- [21] Muthukrisman, “Understanding the Classification report through sklearn,” 2018. <https://muthu.co/understanding-the-classification-report-in-sklearn/>.



# Music Genre Recommendations Based on Spectrogram Analysis Using Convolutional Neural Network Algorithm with RESNET-50 and VGG-16 Architecture

I Nyoman Purnama<sup>1</sup>

<sup>1</sup>Sistem Informasi, STMIK PRIMAKARA

Email: [purnama@primakara.ac.id](mailto:purnama@primakara.ac.id)

**Abstract** – Recommendations are a very useful tool in many industries. Recommendations provide the best selection of what the user wants and provide satisfaction compared to ordinary searches. In the music industry, recommendations are used to provide songs that have similarities in terms of genre or theme. There are various kinds of genres in the world of music, including pop, classic, reggae and others. With genre, the difference between one song and another can be heard clearly. This genre can be analyzed by spectrogram analysis. Convolutional Neural Network(CNN) is a neural network algorithm that is commonly used to recognize and classify image data. In this study, an image spectrogram analysis was developed which will be the input feature for the Convolutional Neural Network. CNN will classify and provide song recommendations according to what the user wants. In addition, testing was carried out with two different architectures from CCN, namely VGG-16 and RESNET-50. From the results of the study obtained, the best accuracy results were obtained by the VGG-16 model with 20 epochs with accuracy 60%, compared to the RESNET-50 model with more than 20 epochs. The results of the recommendations generated on the test data obtained a good similarity value for VGG-16 compared to RESNET-50.

**Keywords** – Recommendation, VGG16, Resnet50, CNN, Spectrogram, Music

## I. INTRODUCTION

Music is an inseparable part of people's lives. Music often accompanies someone in their activities. Sometimes listening to music can also affect the mood of the listeners. Usually someone listens to music according to his feelings at that time. So that the role of music becomes important in managing people psychology[1].

Correct song listened by someone can affect listener's feelings. Due to the large amount of music available either through the internet or other music service applications, it will be difficult for people to choose the songs they want. Music is also distinguished by a variety of genres, speed, tempo and themes that vary and vary widely[2]. For western songs, the genres are distinguished by Hip-Hop, International, Electronic, Folk, Experimental, Rock, Pop, and Instrumental. This makes it difficult for music lovers to choose the right song.

Music lovers usually choose songs using manual method in finding the desired music. Like asking for recommendations from friends or listening to music shows to choose music[3]. Often the song that is listened to, does not match his mood or is not a fan of the genre of the song. Recommendations are implemented in various music player platforms on the internet, to provide more experience in listening to the music. The recommendation system is able to predict the favorite music desired by the user. Besides for users, recommendations are also useful for music service providers, because they can increase user satisfaction for using the music service.

Deep learning is a part of Artificial Neural Network-based Machine Learning. With deep learning, a computer

can classify and recommend data in the form of images or sounds[4]. One of the methods commonly used for the classification and recommendation process is the Convolutional Neural Network (CNN). CNN is an extension of Multilayer Perceptron. CNN is able to learn from an image by using supervised learning techniques. This technique will provide a target for output by comparing past learning experiences. There are several architectures that can be used to optimize CNN so that they can have optimal classification results. There are the VGG architecture, mobileNet, ResNet etc. ResNet, short for Residual Networks is a classic neural network[5]. This model is also the winner of the ImageNet challenge in 2015. This model is also easier to optimize, and can get accuracy from great depths increases. ResNet 50 is the best CNN architecture, it is proved on the research by Talo was to conduct research on the classification of brain diseases with MRI images[6]. The spectrogram is a visual representation of the frequency spectrum of the signal. The spectrogram is formed using the Fourier transform. Making a spectrogram with FFT (Fast Fourier Transform) is done by first taking the data in the time domain, and breaking the data into several parts, and doing a Fourier Transform to calculate the magnitude of the frequency spectrum for each part.

The spectrogram is very useful for analyzing sound, where the spectrogram forms a ratio of magnitude to frequency at a given time. Music recommendations can also be made based on the mood of the user. Where in the research that has been done by Amala George et al. The system developed is able to analyze the mood of the user based on his face, then analyze it using the CNN algorithm[2]. From the results of this mood classification, recommendations are then given using the recommendation



module. From the research that has been done, the accuracy is 98%.

Research of music types Classification using the CNN algorithm has also been carried out by analyzing spectrogram images. The spectrogram image that has been generated from the music, then deep learning process will be carried out by using the CNN model. Based on the research that has been done, it is found that the use of 35 epochs has an optimal accuracy of 81.33%. When compared with the KNN method, CNN produces a better level of accuracy[1]. Other research on spectrogram analysis for music genre classification using CNN and Mel-spectrograms has been carried out and the test results depend on the number of datasets, training iterations and computer specifications greatly affect the level of accuracy and duration of modeling. The resulting accuracy is very optimal in classifying music genres, which is 99% for the RELU activation function and 95% for ELU[7].

Music recommendations based on genre have also been carried out using the Convolutional Recurrent Neural Network. Where in this study also uses a spectrogram and analyzes it using CRNN. This study also compared the use of CRNN and CNN methods to classify music genres. From the research results, it is found that CRNN which takes into account the frequency and time sequence features has better performance than CNN[8]. Research on next-song recommendation has also been carried out, where Neural network has performed well in all types of tests. In this study it was concluded that the NN-based next-song recommenders, CNN-rec, NN-rec and Word2Vec, outperform the non-NN based ones[9]. In this research demonstrate that the NN-based next-song recommenders, which combine users' general preference and sequential listening patterns, have the highest performance.

Music recommendation using deep content also done by Aaron van den Oord [10]. In their research showed that recent advances in deep learning translate very well to the music recommendation setting in combination with approach used in this study, with deep convolutional neural networks significantly outperforming a more traditional approach using bag-of-words representations of audio signals. Also other research on music recommendation done by using user behaviour [11]. The approach considered genre, recording year, freshness, favor and time pattern as factors to recommend songs. The evaluation results demonstrate that the approach is effective.

Research on music recommendations by genre is carried out by comparing several machine learning algorithms such as KNN, RF, NB, DT dan SVM[12]. According to the results summarized in this research, SVM achieved better classification results than other methods. In addition, changing the window size and window type caused very small performance changes. Research about music recommendation using similarity between using decided genre value and using feature vector distance also have been done by Jonseol Dee et al. In their paper, proposed a recommendation system based on a preference classification using real-time user brainwaves and genre feature classification. Proposed user's preference classifier achieved an overall accuracy of 81.07%[13].

Based on the research that has been done previously, this study will carry out a music genre recommendation process using the GZTAN dataset which is composed of 10 types of genres, where the music data is first processed using a spectrogram. The results will be classified using the CNN algorithm with RESNET50 and VGG16 architecture. The results of the recommendations generated will be tested whether they are in accordance with the song desired by the user.

## II. RESEARCH METHODOLOGY

The method used in this research is the dataset preparation process, pre-processing, spectrogram, classification process and calculating similarity using cosine similarity.

### A. Dataset

This research uses a dataset in the form of spectrogram images taken from the GTZAN dataset. To simplify the classification of music data using a neural network, it is necessary to change the music data into a mel-spectrogram to be processed by the Neural Network. GTZAN consists of music data and Mel spectrogram results from that music file. Where this dataset is a public dataset that is widely used for evaluating the introduction of music genres (Music Genre Recognition / MGR). GTZAN is a collection of music collected from 2000-2001, which comes from various sources such as CDs, radio and microphone recordings. This dataset consists of 10 genres, namely blues, classical, country, disco, hip-hop, jazz, metal, pop, reggae and rock. The duration of each of these music is 30 seconds. Each genre contains 100 music files. The number of datasets used in this study is divided into 3 parts : training data, validation data and test data. With details in each section as follows:

Table 1. Number dataset used on each class

Genre	Training data	Validation data	Testing data
Blues	80	10	10
Classical	80	10	10
Country	80	10	10
Disco	80	10	10
Hiphop	80	10	10
Jazz	80	10	10
Metal	80	10	10
Pop	80	10	10
Reggae	80	10	10
Rock	80	10	10

### B. Spectrogram

The spectrogram is a visual representation of the frequency spectrum of the signal[14]. In the GTZAN dataset, spectrograms have been generated and stored in their respective classes. Before being entered into the CNN network, this data is further divided into training data, validation data and test data. Each of these spectrogram images will be included in the array, then labeled according to their respective index folders. Then after being given a label, the data will be appended into an array to make it easier to pass the data. From the spectrogram image there are many values and features of the music file that can be displayed. The following is an example of an illustration of the spectrogram of each class in this study.



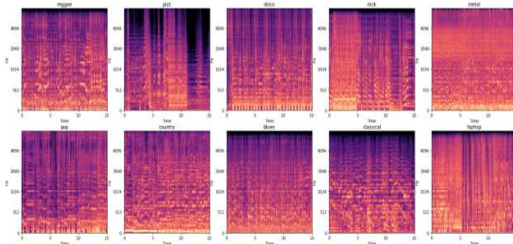


Figure 2 Example of Spectrogram on each classes

Based on the picture above, it can be seen that there are differences in the spectrograms of each genre. The image on the right shows a spectrogram for the hip hop and rock genres, here you can see the frequency density compared to the spectrogram on the left. The spectrogram image is a wave generated as an audio representation in the time, frequency and magnitude domains. To generate spectrograms from each music genre and use it on an artificial neural network, in this study, the Librosa library was used. With librosa, we can retrieve important features in a music file, such as tempo, chroma and spectrogram.

### C. Convolutional Neural Network

Convolutional Neural Network is an artificial neural network that is widely used in the field of image classification. In this study, the audio/music signal is represented as a spectrogram which has a 2D image. CNN is used to classify music genres with the help of spectrograms. Based on the spectrogram images of each music genre, the pattern of the audio signal can be seen. So that each of these genres can be input of the CNN artificial neural network.

In this study, two CNN architectures were used, namely Resnet and VGG16. Resnet is a residual network, which is in charge of image recognition. RESNET-50 is an improved version of VGG-16. Where the last number of this architecture represents the number of layers in the network. RESNET stands for Residual Network which is an artificial neural network innovation that won the 2015 ILSVRC classification competition with an error rate of only 3.15%[15]. While VGG-16 stands for Visual Geometry Group and 16 is the number of layers. The VGG-16 is also a well-known model that participated in the 2014 ILSVRC and obtained an accuracy rate of 92.7%. VGG-16 is also used in image classification and is very popular because of its ease of implementation. The following in Figure 3 is a comparison of the RESNET and VGG architectures.

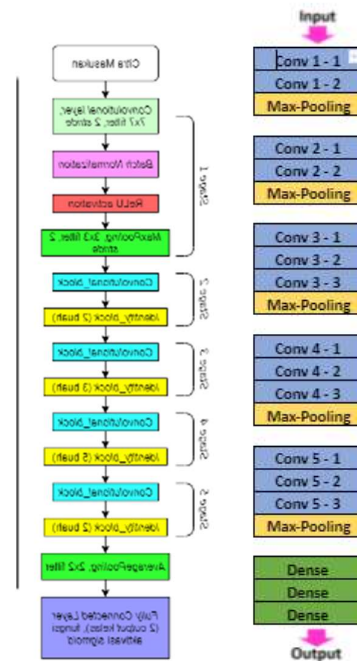


Figure 3. Comparison of Resnet-50 and VGG-16 . architectures

### D. Research flow

The research flow used in this study can be described in outline as follows:

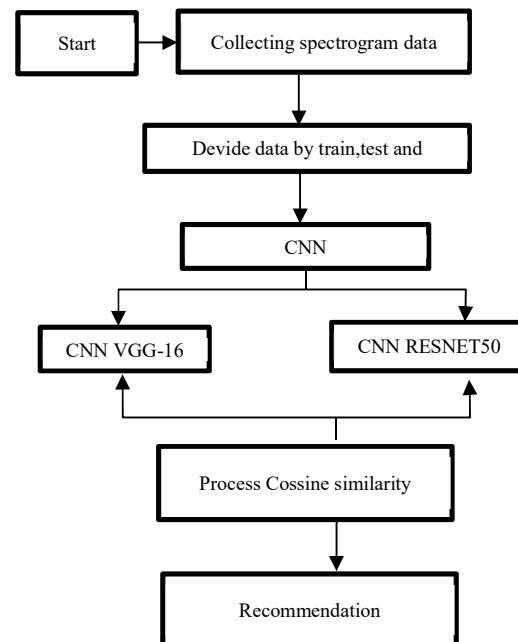


Figure 4. Research flow

The first step in this research is to collect the dataset, in the form of a spectrogram image from the GTZAN dataset. After that by using the required libraries such as *imageDataGenerator* in the Keras library to manage training, test and validation data. After that the MFCC image data from 10 music genres have been grouped by category. The next step is to build a CNN model with Keras.



There are 2 different processes will be carried out using the VGG-16 and Resnet 50 architectures.

After the model is obtained from the training process using two different architectures. Then the process of finding similarities between feature vectors is carried out using cosine similarity. The application will display a recommendation of 5 songs that match those in the validation data. Where the recommendation process is carried out by calculating the value of the similarity of features between one music and another. The first process is to choose music from each genre that will be used as the basis for the recommendation system. Then the forecast from the music base is calculated based on an artificial neural network. The cosine similarity value is calculated from the 2 featured vector being compared. To calculate the similarity of 2 pieces of music with the number of features N, where the first music has a feature vector  $x=[x_1,x_2,x_3,\dots,x_n]$  and the second music has a feature vector  $y=[y_1,y_2,y_3,\dots,y_n]$  then the formula which is used as follows:

$$\cos \theta = \frac{\sum_{i=1}^n x_i y_i}{\sqrt{\sum_{i=1}^n x_i^2} \sqrt{\sum_{i=1}^n y_i^2}}$$

Figure 5. Cossine similarity formula

### III. RESULTS AND DISCUSSION

System implementation is done using Google Colabs. The libraries used in the making of this research are numpy, pandas, librosa, Keras and Scikit learns. This research uses a spectrogram image dataset obtained from the GTZAN dataset with a total of 1000 music data and is divided into 10 categories namely Blues, Classical, Country, Disco, Hip Hop, Jazz, Metal, Pop, Reggae, Rock. This spectrogram image will be the input for the Convolutional Neural Network. Where the image in the form of a mel spectrogram representation of this audio file is saved in “.jpg” format.

To build the image dataset in this study, the *ImageDataGenerator* library was used. As a parameter of this library, we must divide the GTZAN dataset into 3 folder namely training, test and validation data. The number of classes used is 10 classes which are divided into their respective folders. Some parameters that must be initialized are `batch_size=64` and the number of initial epochs used is 20. An important parameter that we need to initialize is `input_shape` for all images, in this study we set the input shape at (224,224,3) / RGB channel and normalize image with a scale of 1./255. To implement CNN in this research, using Python programming language with Keras library and tensorflow. CNN modeling is done by initializing the CNN network layer parameters, namely the number of conv2d layers in the model=2, the number of conv2d layers=32, filter size=(3,3), initializer=glorot\_uniform, activation function=relu, layer dropout=0.2 and the optimized optimizer. "Adam" is used. The CNN model makes several layers, namely convolution layers, pooling layers, dropout layers, flatten layers, dense layers and the RELU activation function. The result of the convolution process is a feature map that is used in the

convolution process repeatedly. The resulting model output is shown as shown below.

Layer (type)	Output Shape	Param #
input_1 (InputLayer)	[None, 224, 224, 3]	0
block1_conv1 (Conv2D)	(None, 224, 224, 64)	1792
block1_conv2 (Conv2D)	(None, 224, 224, 64)	36928
block1_pool (MaxPooling2D)	(None, 112, 112, 64)	0
block2_conv1 (Conv2D)	(None, 112, 112, 128)	73856
block2_conv2 (Conv2D)	(None, 112, 112, 128)	147584
block2_pool (MaxPooling2D)	(None, 56, 56, 128)	0
block3_conv1 (Conv2D)	(None, 56, 56, 256)	295168
block3_conv2 (Conv2D)	(None, 56, 56, 256)	590080
block3_conv3 (Conv2D)	(None, 56, 56, 256)	590080

Figure 6. CNN output model

The next process is to carry out the transfer learning process with 2 different architectures, namely VGG16 and RESNET50. Transfer learning is the process of using an existing model for different problems. By using transfer learning, it is hoped that the results of the training will be better. The parameters needed in this transfer learning process are *lastfourtrainable*, if the value of this parameter is false then the last fully connected layer will be trained. If true then the last 4 layer models that have parameters will be trained. For these two architectural models, "adam" optimization is used. The training process was carried out on each model of 20 epochs. This training will produce a model that will be used in the testing process.

#### A. Result analysis

After the training process was carried out, the precision, recall, and f1-score values were obtained from each music class. Following are the values of Precision, recall, f1-score and accuracy on the VGG16 model.

	precision	recall	f1-score	support
blues	0.56	0.50	0.53	10
classical	0.91	1.00	0.95	10
country	0.67	0.40	0.50	10
disco	0.60	0.30	0.40	10
hiphop	0.60	0.60	0.60	10
jazz	0.53	0.80	0.64	10
metal	0.55	0.60	0.57	10
pop	0.67	0.60	0.63	10
reggae	0.70	0.70	0.70	10
rock	0.36	0.50	0.42	10
accuracy			0.60	100
macro avg	0.61	0.60	0.59	100
weighted avg	0.61	0.60	0.59	100

Figure 7. VGG16 Accuracy value

The results of the confusion matrix for the CNN-VGG16 model are shown in Fig. where the results obtained are quite good in classifying the class of the music dataset used. The best classification process was obtained from classical, jazz and reggae classes. While



the lowest classification was obtained in the disco class.

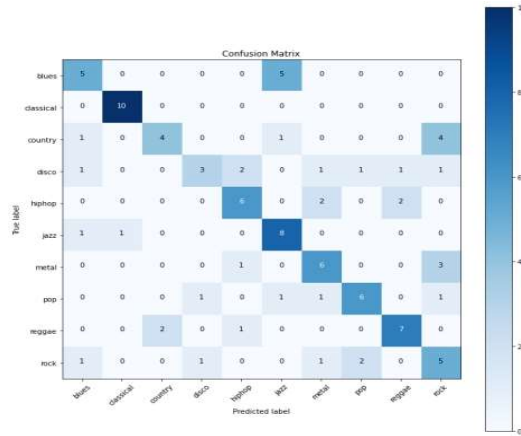


Figure 8. VGG16 Confusion matrix

After the process of model formation with transfer learning VGG16 and RESNET50. Then proceed with making feature extraction. In the VGG16 model, for example, in this study, we will take a model that has been previously stored in the training process. After that, the output weight will be obtained before the classification layer of this model. From this model we will derive the feature vectors for the training and validation datasets. The result of this feature vector is then searched for its similarity with cosine similarity.

In Figure 9, 5 music recommendations are obtained based on the spectrogram image of the music file desired by the user. Where "test image" is the music spectrogram testing data. As Seen with VGG16, the recommended spectrogram has almost the same shape as the test spectrogram. With the test image from the Blues class, the recommendation results are also obtained from the Blues class with a similarity level of 1.

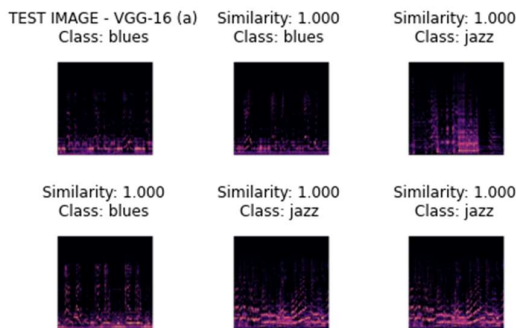


Figure 9. Recommendation output for VGG16 model

While on RESNET50, it has the same testing process as VGG16. After the experiment, the training process with RESNET50 requires a larger epoch to get better accuracy results. In this study, quite good accuracy results were obtained at epochs of 30 for RESNET50. The picture below shows the calculation results of Precision, recall, f1-score and accuracy on the RESNET50 model. The resulting Accuracy value is slightly lower than the VGG16 model with a larger number of epochs. The results of the confusion matrix for the CNN-RESNET50 model are shown in Fig.

Where the results obtained are still lower than the VGG16 model in classifying classes from the music dataset used.

	precision	recall	f1-score	support
blues	0.75	0.30	0.43	10
classical	1.00	0.70	0.82	10
country	0.00	0.00	0.00	10
disco	0.26	0.60	0.36	10
hiphop	0.67	0.20	0.31	10
jazz	0.50	0.50	0.50	10
metal	0.00	0.00	0.00	10
pop	0.57	0.80	0.67	10
reggae	0.23	0.90	0.37	10
rock	0.00	0.00	0.00	10
accuracy			0.40	100
macro avg	0.40	0.40	0.35	100
weighted avg	0.40	0.40	0.35	100

Figure 10. RESNET50 Accuracy value

The results of the confusion matrix for the CNN-RESNET50 model are shown in Fig. Where the results obtained are still lower than the VGG16 model in classifying classes from the music dataset used. The best classification is obtained from the reggae and pop classes. In rock class, the RESNET50 model is not able to give good classification results.

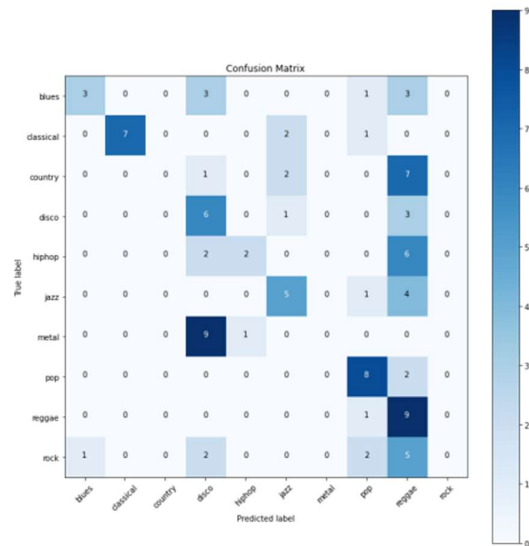


Figure 11 RESNET50 Confusion matrix

For the testing process, the steps taken are the same as the process in the VGG16 model, that is looking for feature extraction from the test image and looking for its cosine similarity with feature extraction from the training dataset. So that the results of the music spectrogram recommendations are obtained in accordance with the testing dataset used. The following is in Figure 12 the results of 5 image similarities from the tested test data. It can be seen that the results of the spectrogram recommendation are quite good, only the level of similarity is lower than the VGG16 model.



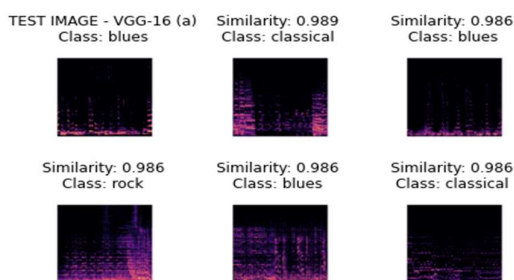


Figure 12. Recommendation output for RESNET50 model

#### IV. CONCLUSION

In this research it is implemented using Python, Google Colab, and TensorFlow and hard libraries. Input shape on CNN model in This research is 224x224 pixels, the filter size is 3x3, the number of epochs is 20 and 30, and the training data is 799 and the validation is 100 data. CNN is the most widely used method in image data. For research with audio data, this data is first processed by spectrogram analysis in the form of Cartesian coordinates with the amplitude of the music as the y-axis. In this study, the spectrogram results become input for CNN with VGG16 and RESNET50 architectures.

Based on the results of the design of prediction system using the CNN method, the accuracy value for the VGG16 training data model is 0.8408, the training data loss is 0.4827, the test data accuracy is 0.6094 and the test data loss is 1.2762. Meanwhile, for the RESNET50 model, the training data accuracy value is 0.6286, the training data loss is 1.0383, the test data accuracy is 0.3438 and the test data loss is 1.8529. So, from these results it can be concluded that the results in both the data is still underfitting. This is because there are still many datasets that are more numerous in number and variants that have characteristics that are similar to each class.

The best accuracy results were obtained by the VGG16 model with 20 epochs compared to the RESNET50 model with more than 20 epochs. The results of the recommendations generated on the test data obtained a good similarity value for VGG16 compared to RESNET50. The suggestion for this research is that in the future it can increase the dataset so that the accuracy obtained is even better, because in this study the songs in the dataset do not have clear boundaries between one genre and another. In addition, the epoch value during the training process is also further improved so that the accuracy level is even better for each CNN model.

#### REFERENCES

[1] Y. M. G. Costa, L. S. Oliveira, A. L. Koerich, and F. Gouyon, "Music genre recognition using spectrograms," *Int. Conf. Syst. Signals, Image Process.*, pp. 151–154, 2011.

[2] A. George, S. Suneesh, S. Sreelakshmi, and T. E. Paul, "Music Recommendation System Using CNN," vol. 9, no. 6, pp. 4197–4200, 2020.

[3] C. R. Wairata, E. R. Swedia, and M. Cahyanti, "Pengklasifikasian Genre Musik Indonesia Menggunakan Convolutional Neural Network," *Sebatik*, vol. 25, no. 1, pp. 255–261, 2021, doi: 10.46984/sebatik.v25i1.1286.

[4] J. Dias, "Music genre classification from Spectrogram using CNN," [Online]. Available: [http://cs230.stanford.edu/files\\_winter\\_2018/projects/6936608.pdf](http://cs230.stanford.edu/files_winter_2018/projects/6936608.pdf).

[5] Faiz Nashrullah, Suryo Adhi Wibowo, and Gelar Budiman, "The Investigation of Epoch Parameters in ResNet-50 Architecture for Pornographic Classification," *J. Comput. Electron. Telecommun.*, vol. 1, no. 1, pp. 1–8, 2020, doi: 10.52435/complete.v1i1.51.

[6] M. Talo, O. Yildirim, U. B. Baloglu, G. Aydin, and U. R. Acharya, "Convolutional neural networks for multi-class brain disease detection using MRI images," *Comput. Med. Imaging Graph.*, vol. 78, p. 101673, Dec. 2019, doi: 10.1016/J.COMPMEIMAG.2019.101673.

[7] D. Lionel, R. Adipranata, and E. Setyati, "Klasifikasi Genre Musik Menggunakan Metode Deep Learning Convolutional Neural Network dan Mel- Spektrogram," *J. Infra Petra*, vol. 7, no. 1, pp. 51–55, 2019, [Online]. Available: <http://publication.petra.ac.id/index.php/teknik-informatika/article/view/8044>.

[8] Adiyansjah, A. A. S. Gunawan, and D. Suhartono, "Music recommender system based on genre using convolutional recurrent neural networks," *Procedia Comput. Sci.*, vol. 157, pp. 99–109, 2019, doi: 10.1016/j.procs.2019.08.146.

[9] K.-C. Hsu, S.-Y. Chou, Y.-H. Yang, and T.-S. Chi, "Neural Network Based Next-Song Recommendation," 2016, [Online]. Available: <http://arxiv.org/abs/1606.07722>.

[10] D. G. W. Ingram *et al.*, "Computer Aided Design, International Conference, University of Southampton, Engl, Apr 24-28 1972.," *Inst Electr Eng, Conf Publ*, no. 86. IEE, pp. 1–9, 1972.

[11] Y. Hu, "12th International Society for Music Information Retrieval Conference ( ISMIR 2011 ) NEXTONE PLAYER: A MUSIC RECOMMENDATION SYSTEM BASED ON USER BEHAVIOR," no. Ismir, pp. 103–108, 2011.

[12] A. Elbir, H. Bilal Çam, M. Emre Iyican, B. Öztürk, and N. Aydin, "Music Genre Classification and Recommendation by Using Machine Learning Techniques," *Proc. - 2018 Innov. Intell. Syst. Appl. Conf. ASYU 2018*, no. October 2018, 2018, doi: 10.1109/ASYU.2018.8554016.

[13] J. Lee, K. Yoon, D. Jang, S. J. Jang, S. Shin, and J. H. Kim, "Music recommendation system based on genre distance and user preference classification," *J. Theor. Appl. Inf. Technol.*, vol. 96, no. 5, pp. 1285–1292, 2018.

[14] M. H. Ashshiddieqy, Jondri, and A. Rizal, "Klasifikasi Suara Paru Dengan Convolutional Neural Network (CNN)," *eProceedings Eng.*, vol. 07, no. 02, pp. 8506–8512, 2020.

[15] W. Setiawan, "Perbandingan Arsitektur Convolutional Neural Network Untuk Klasifikasi Fundus," *J. Simantec*, vol. 7, no. 2, pp. 48–53, 2020, doi: 10.21107/simantec.v7i2.6551.



# The Classification of Anxiety, Depression, and Stress on Facebook Users Using the Support Vector Machine

Tsania Maulidia Wijiasih<sup>1\*</sup>, Rona Nisa Sofia Amriza<sup>2</sup>, Dedy Agung Prabowo<sup>3</sup>

<sup>1</sup>Information Systems Study Program, Faculty of Informatics, Institut Teknologi Telkom Purwokerto

<sup>2</sup>Information Systems Study Program, Faculty of Informatics, Institut Teknologi Telkom Purwokerto

<sup>3</sup>Informatics Engineering Study Program, Faculty of Informatics, Institut Teknologi Telkom Purwokerto

Email: <sup>1</sup>18103123@ittelkom-pwt.ac.id, <sup>2</sup>rona@ittelkom-pwt.ac.id, <sup>3</sup>dedy@ittelkom-pwt.ac.id

**Abstract**– Social media remains an essential platform for connecting people with friends, family, and the world around them. However, when events spread on social media are primarily negative, it will cause depression, anxiety, and stress that tend to increase. This study aims to classify depression, anxiety, and stress using the Support Vector Machine. The data in this study were obtained from active Facebook users using the Depression Anxiety Stress Scale (DASS 21) questionnaire. This study adopted the Knowledge Discover Database process. The result of this study is an evaluation of the performance of the Support Vector Machine classification of depression, anxiety, and stress. The accuracy of the Support Vector Machine in this study is 98.96%.

**Keywords** – Support Vector Machine, DASS 21, Depression, Anxiety, Stress, Facebook

## I. INTRODUCTION

Social media is a computer technology that facilitates sharing of ideas, thoughts, and information through the internet network [1]. Data reported by Internet World Stats states that Indonesia ranks third in the world's most prominent use of social media, Facebook, reaching 176.5 million users in June 2021, which is equivalent to 63.9% of the total population of Indonesia [2]. The 2014 Indonesia Family Life Survey (IFLS) surveyed 22,423 individuals in Indonesia; the survey found that one standard deviation of social media use was associated with a 9% increase in CES-D scores (Center For Epidemiological Studies Depression Scale) [3]. It proves that social media has a negative impact on mental health [3]. Social media itself is seen as social support among users. Still, it can harm mental health, specifically, those who already have a significant degree of depression, anxiety, and stress [4].

Furthermore, Tang, Wang, and Norman (2013) found that activities on social media such as sharing, liking, messaging, and other activities increased stress. Moreover, excessive use of social media Facebook has become a severe source of stress because people often share all kinds of feeds, stories, and comments, from economics, politics, and social issues to personal problems [5]. Another thing is the desire to upload the best photos of yourself to get compliments or likes, and the pressure of bringing out the best of yourself can make the Facebook user feel anxious. In addition to anxiety, friends' achievements on Facebook are one of the factors which affect a person's mental health condition [6]. From the problems above, a classification is needed to classify active Facebook users affected by depression, anxiety, or stress to achieve a good life balance. Positive mental health can help individuals work productively and reach one's full potential.

The initial stage of this research is to collect data. In collecting data, this research was conducted by distributing questionnaires. The questionnaire itself is a research

instrument consisting of a series of questions or other types of instructions that aim to collect information from a respondent. Several studies have used previous questionnaires to assess levels of depression, anxiety, and stress, such as the Perceived Stress Scale (PSS-10) [7], Subjective Units of Distress Scale (SUDS)[8], The Hamilton Rating Scale for Depression (HAM-D) [9], Hamilton Anxiety Rating Scale (HAM-A) and Depression, Anxiety, and Stress Scale (DASS 21) [10]. DASS-21 used in this study is because it has been used in several studies and has high consistency [11].

Several previous researchers used the Support Vector Machine to classify depression, anxiety, and stress in conducting the classification. Research by Zhang et al. predicts Social Anxiety Disorder using the Support Vector Machine, and the results of this study show an accuracy of 76.25%. It shows that the Support Vector Machine makes a good diagnosis of the potential for Social Anxiety Disorder [12]. Subsequent research was conducted by Frick et al. using the Support Vector Machine to classify Social Anxiety Disorder, and the result is an accuracy of 72.6% [13]. Another study conducted by Pantazatos et al. used the Support Vector Machine and got high accuracy results of 89% [14]. Therefore, this study identifies the classification of depression, anxiety, and stress on social media Facebook using a Support Vector Machine; this model determines the distance using a support vector, so the computing process becomes faster and produces high accuracy in classification.

## II. RESEARCH METHODOLOGY

The object of this research is the social media Facebook, and the subject of this research is active Facebook users. The respondents of this study were obtained by distributing Google Forms via social media such as Facebook and Twitter. The questionnaire contains



the questions from Depression Anxiety Stress Scale 21 (DASS21).

Figure 1 shows the research stages, starting with the study of literature and data collection to achieve data results that can be processed in the Knowledge Discovery Database and then evaluate the performance of the Support Vector Machine.

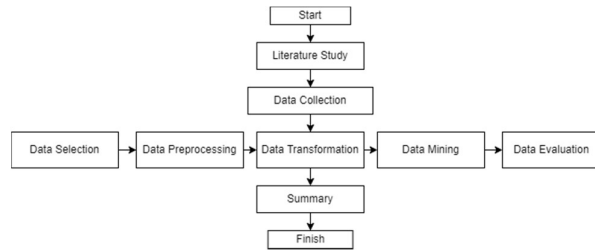


Figure 1. Research Stages

#### A. Literature Study

Literature studies are carried out by reading scientific sources such as books and journals related to the research topic or research question. This stage aims to find how this research relates to existing knowledge.

#### B. Data collection

Data collection in this study was carried out by distributing a Google Form containing a Depression Anxiety and Stress Scale (DASS 21) questionnaire to active Facebook users using Convenience Sampling. The questionnaire was shared on several social media platforms, such as Facebook and Twitter.

#### C.KDD

The Knowledge Discovery Database in this research transforms data into valuable knowledge. The context of this research is the classification of depression, anxiety, and stress in Facebook users. The stages in the KDD process are explained below.

##### (a). Data Selection

The researcher selects the data for the classification process at this data selection stage. The data used comes from Depression Anxiety Stress Scale 21 questionnaire. However, this data is not in accordance with the classification process, so the researcher needs to select the appropriate data.

##### (b). Data Preprocessing

In this data selection stage, noise or irrelevant data is removed from the previous data collection. This stage is necessary so that there is no duplication of data, inconsistent data, or correcting errors in the data. The results of the DASS 21 questionnaire are then labeled using the formulation below:

$$\text{Total} = (\sum \text{sub item}) \times 2$$

The total value of each item calculates by performing an addition to all of the sub-items and then multiple by two. After the total value of each item is obtained, the next step is to compare each item's value, and the highest value of the item is chosen to be a label.

##### (c). Data Transformation

After cleaning the data, then continued with the Data Transformation stage. In this stage, we change the data format, structure, or value into the form required in the data mining process.

##### (d). Data Mining

After the data transformation has been carried out in the previous process, it continues with data mining, extracting potentially valuable patterns. At this stage, the Support Vector Machine is applied.

##### (e). Data Evaluation

At this stage, the researcher evaluates the performance from the classification result. The evaluation in this research uses a confusion matrix. The output of this stage is accuracy, precision, f1, and recall.

### III. RESULTS AND DISCUSSION

The total respondents to the DASS 21 questionnaire were 193 respondents with, 67 male and 126 female. It can be seen in the image below:

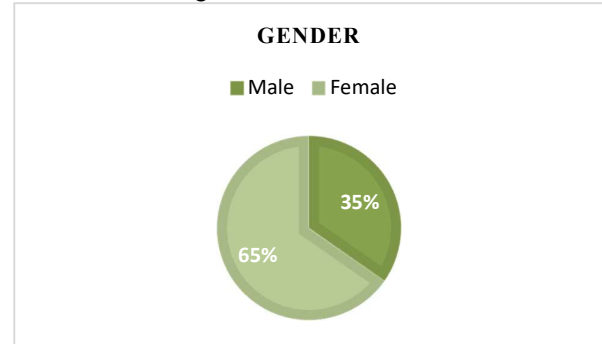


Figure 2. Gender

Most respondents belonged to the age group of 20-30 years, 131 people, followed by 13-20 years, 41 people, 40-50 years, 11 people, then 50 years, 11 people, and 30-40 years total of six people. It can be seen in the image below:

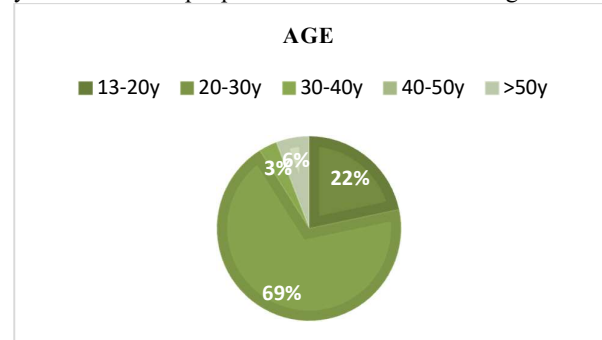


Figure 3. Age

Respondents with jobs as students or college students dominate in this study. It can be seen in the image below:



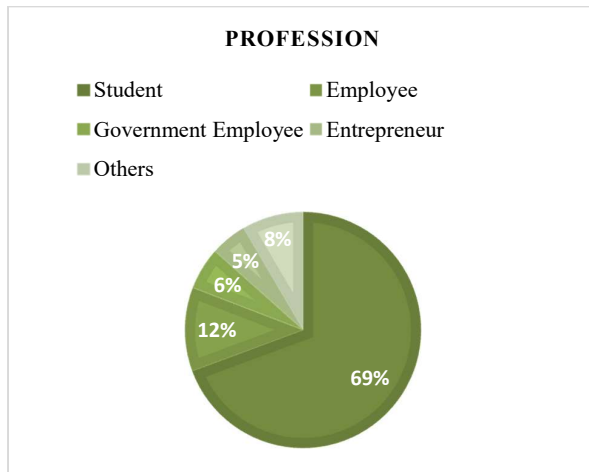


Figure 4. Profession

The domicile of the respondents is very diverse, from the islands of Java, Sumatra, and Kalimantan to Sulawesi.

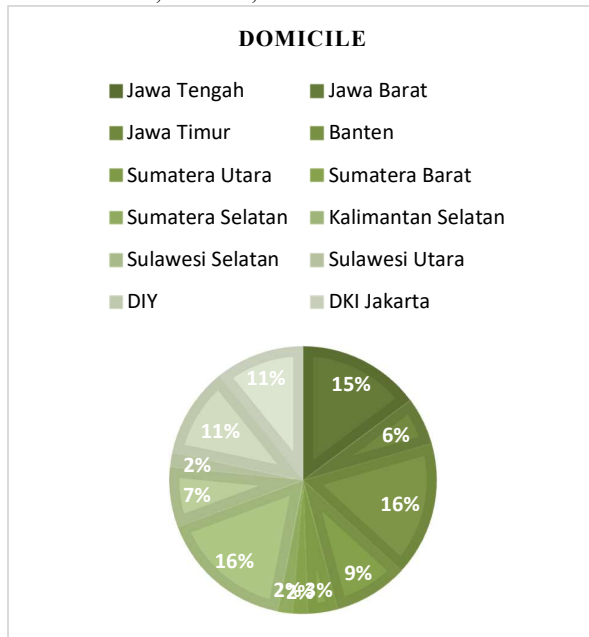


Figure 5. Domicile

The data used is a DASS 21-question instrument. There are 21 questions, with 7 for depression questions, 7 for anxiety questions, and 7 for stress questions. The total number of respondents is 261, but only 193 can be categorized as users who experience depression, anxiety, and stress

The following process is labeling. Labeling is done on the respondent's data obtained from the previous process. Labeling was performed using a DASS score of 21.

Table1. Sample of Respondent Data

Mental Illness	Items	Respondent 1	Respondent 2	Respondent 3
Depression	D1	2	1	2
	D2	2	1	1
	D3	2	0	1

	D4	2	0	1
	D5	2	1	1
	D6	2	1	1
	D7	2	0	1
Anxiety	A1	2	1	0
	A2	2	0	0
	A3	3	1	1
	A4	2	1	1
	A5	2	1	1
	A6	3	1	1
	A7	2	0	1
Stress	S1	2	1	0
	S2	2	1	1
	S3	2	1	1
	S4	3	0	1
	S5	3	0	1
	S6	3	0	1
	S7	2	0	1
CLASS		1	2	3

$$\text{Total depression [1]} = (\sum \text{sub item depresi}) \times 2 \\ (D1 + D2 + D3 + D4 + D5 + D6 + D7) \times 2 \\ = 28$$

$$\text{Total Anxiety [1]} = (\sum \text{sub item kecemasan}) \times 2 \\ (A1 + A2 + A3 + A4 + A5 + A6 + A7) \times 2 \\ = 32$$

$$\text{Total Stress [1]} = (\sum \text{sub item stres}) \times 2 \\ (S1 + S2 + S3 + S4 + S5 + S6 + S7) \times 2 \\ = 34$$

Label[1] = Stress

$$\text{Total depression [2]} = (\sum \text{sub item depresi}) \times 2 \\ (D1 + D2 + D3 + D4 + D5 + D6 + D7) \times 2 \\ = 8$$

$$\text{Total Anxiety [2]} = (\sum \text{sub item kecemasan}) \times 2 \\ (A1 + A2 + A3 + A4 + A5 + A6 + A7) \times 2 \\ = 10$$

$$\text{Total Stress [2]} = (\sum \text{sub item stres}) \times 2 \\ (S1 + S2 + S3 + S4 + S5 + S6 + S7) \times 2 \\ = 6$$

Label[2] = Anxiety

$$\text{Total depression [3]} = (\sum \text{sub item depresi}) \times 2 \\ (D1 + D2 + D3 + D4 + D5 + D6 + D7) \times 2 \\ = 16$$

$$\text{Total Anxiety [3]} = (\sum \text{sub item kecemasan}) \times 2 \\ (A1 + A2 + A3 + A4 + A5 + A6 + A7) \times 2 \\ = 10$$

$$\text{Total Stress [3]} = (\sum \text{sub item stres}) \times 2 \\ (S1 + S2 + S3 + S4 + S5 + S6 + S7) \times 2 \\ = 12$$

Label[3] = Depression

Calculation of each depression, anxiety, and stress item was carried out using a DASS score of 21. Furthermore, each total mental illness was multiplied by two, and then compared the results of the calculation of each item. Respondent 1 was labeled depression because the calculated DASS 21 score for depression was more significant than the calculated DASS 21 score for anxiety and stress. Respondent 2 was labeled anxiety, and respondent three was labeled depression. The calculation was carried out on all 193-respondent data.

After the data of 193 respondents were labeled depression, anxiety, and stress, it was continued by selecting the data to be used for processing in data mining. Name, email address, gender, age, occupation, and domicile data on the questionnaire results were deleted.

The label data is then transformed from numeric to categorical. Furthermore, the data is processed using the Support Vector Machine to produce a classification of depression, anxiety, and stress. Based on data of 193 respondents who have been tested, results of the calculation of f1, precision, recall, and accuracy are obtained. The Support Vector Machine model produces an accuracy of 98.96%, F1 is 95.75%, precision is 99.15%, and recall is 97.26% (Table 2).

Table2. Result

Classification	Mental illness	Accuracy	F1	Precision	Recall
SVM	Depression Anxiety Stress	98.96%	95.75%	99.15%	97.26%

The result of each item was evaluated using a confusion matrix. The mathematic formulation for accuracy, precision, recall, and f1 for each item score is defined below:

$$\begin{aligned} \text{Depression Accuracy} &= \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \\ &= \frac{118 + 73}{118 + 1 + 1 + 73} \times 100\% \\ &= 98.96\% \end{aligned}$$

$$\begin{aligned} \text{Anxiety Accuracy} &= \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \\ &= \frac{17 + 173}{17 + 173 + 2 + 1} \times 100\% \\ &= 98.44\% \end{aligned}$$

$$\begin{aligned} \text{Stress Accuracy} &= \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \\ &= \frac{55 + 137}{55 + 137 + 0 + 1} \times 100\% \\ &= 99.48\% \end{aligned}$$

$$\begin{aligned} \text{Depression Precision} &= \frac{TP}{TP + FP} \times 100\% \\ &= \frac{118}{118 + 1} \times 100\% = 99.15\% \end{aligned}$$

$$\begin{aligned} \text{Anxiety Precision} &= \frac{TP}{TP + FP} \times 100\% \\ &= \frac{117}{117 + 2} \times 100\% = 98,31\% \end{aligned}$$

$$\begin{aligned} \text{Stress Precision} &= \frac{TP}{TP + FP} \times 100\% \\ &= \frac{55}{55 + 0} \times 100\% = 100\% \end{aligned}$$

$$\begin{aligned} \text{Depression Recall} &= \frac{TP}{TP + FN} \times 100\% \\ &= \frac{118}{118 + 1} \times 100\% = 99.15\% \end{aligned}$$

$$\begin{aligned} \text{Anxiety Recall} &= \frac{TP}{TP + FN} \times 100\% = \frac{17}{17 + 1} \times 100\% \\ &= 94.44\% \end{aligned}$$

$$\begin{aligned} \text{Stress Recall} &= \frac{TP}{TP + FN} \times 100\% = \frac{55}{55 + 1} \times 100\% \\ &= 98.21\% \end{aligned}$$

$$\begin{aligned} \text{Depression F1} &= 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \\ &= 2 \times \frac{99.15 \times 99.15}{99.15 + 99.15} = 99.15\% \end{aligned}$$

$$\begin{aligned} \text{Anxiety F1} &= 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \\ &= 2 \times \frac{89.47 \times 97.44}{89.47 + 97.44} = 93.28\% \end{aligned}$$

$$\begin{aligned} \text{Stress F1} &= 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \\ &= 2 \times \frac{91.66 \times 98.21}{91.66 + 98.21} = 94.82\% \end{aligned}$$

Where TN (True Negative) is a negative data with true value, FP (False Positive) is negative data that identifies as positive data. For the more TP (True Positive) have actual value, and FN (False Negative) is a positive data identified as a negative data.

Based on the Support Vector Machine classification method, the result shows that depression accuracy is 98.96%, which explains that 118 instances labeled as depression have a correct value. Based on the Support Vector Machine classification method, the result shows that depression accuracy is 98.44%, which explains that 17 instances labeled as depression have a correct value. Based on the Support Vector Machine classification method, the result shows that depression accuracy is 99.48%, it explains that 55 of instances labeled as depression have a correct value.

The second confusion matrix score is precision. The one instance cannot be labeled as depression; it shows that the precision of depression is 99.15%. Then, two instances cannot be labeled as anxiety; it indicates that anxiety's precision is 89.47%. All of the instances labeled as stress show that the precision of stress is 100%.

The recall value for depression is 99.15% because one instance didn't classify as depression. Next, recall's anxiety is 94.44% because one instance didn't classify as anxiety. Finally, the recall value for stress was 98.21% because one instance wasn't classified as stress.

The last confusion matrix is F1. 99.15% for depression, 93.28% for anxiety and 94.82% for stress.

#### IV. CONCLUSION

From the test results, the Support Vector Machine model has high accuracy because it has advantages such as determining the distance using a support vector to make the



computing process faster [15]. The Support Vector Machine creates a decision function or hyperplane that can differentiate between categories. The resulting decision function or hyperplane will be used to predict a predetermined class, so the classification accuracy is high [16]. The Support Vector Machine method is the best method for classifying depression, anxiety, and stress in Facebook users. It was shown with 98.96% accuracy. Furthermore, this research can be developed by adding other data mining methods such as naïve Bayes, random forest, and decision tree to see the comparative performance of the model. Furthermore, other social media use can be subject to future research, for example, Twitter users, Instagram users, and YouTube users.

#### REFERENCES

- [1] M. Dollarhide, "Social Media Definition," *Investopedia*, 2021.
- [2] Internet World Stats, "Internet World Stats," *Internet World Stats*, 2015.
- [3] S. Sujarwoto, G. Tampubolon, and A. C. Pierewan, "A Tool to Help or Harm? Online Social Media Use and Adult Mental Health in Indonesia," *Int. J. Ment. Health Addict.*, vol. 17, no. 4, pp. 1076–1093, 2019, doi: 10.1007/s11469-019-00069-2.
- [4] P. D. Lauren Reining, M.A., Michelle Drouin, "College Students in Distress: Can Social Media be a Source of Social Support?," *Park. Heal. Res. Repos.*, 2018.
- [5] L. Weng and F. Menczer, "Topicality and impact in social media: Diverse messages, focused messengers," *PLoS One*, vol. 10, no. 2, pp. 1–17, 2015, doi: 10.1371/journal.pone.0118410.
- [6] S. Budury, A. Fitriyani, and K. -, "Penggunaan Media Sosial Terhadap Kejadian Depresi, Kecemasan Dan Stres Pada Mahasiswa," 2019. doi: 10.36376/bmj.v6i2.87.
- [7] E. H. Lee, "Review of the psychometric evidence of the perceived stress scale," *Asian Nurs. Res. (Korean. Soc. Nurs. Sci.)*, vol. 6, no. 4, pp. 121–127, 2012, doi: 10.1016/j.anr.2012.08.004.
- [8] L. Horwitz, "Book Reviews: The Practice of Supportive Psychotherapy," *J. Am. Psychoanal. Assoc.*, vol. 36, no. 1, pp. 197–199, 1988, doi: 10.1177/000306518803600115.
- [9] D. Carrozzino, C. Patierno, G. A. Fava, and J. Guidi, "The hamilton rating scales for depression: A critical review of clinimetric properties of different versions," *Psychother. Psychosom.*, vol. 89, no. 3, pp. 133–150, 2020, doi: 10.1159/000506879.
- [10] M. HAMILTON, "the Assessment of Anxiety States By Rating," *Br. J. Med. Psychol.*, vol. 32, no. 1, pp. 50–55, 1959, doi: 10.1111/j.2044-8341.1959.tb00467.x.
- [11] S. Verma and A. Mishra, "Depression, anxiety, and stress and socio-demographic correlates among general Indian public during COVID-19," *Int. J. Soc. Psychiatry*, vol. 66, no. 8, pp. 756–762, 2020, doi: 10.1177/0020764020934508.
- [12] W. Zhang *et al.*, "Diagnostic prediction for social anxiety disorder via multivariate pattern analysis of the regional homogeneity," *Biomed Res. Int.*, vol. 2015, 2015, doi: 10.1155/2015/763965.
- [13] A. Frick *et al.*, "Classifying social anxiety disorder using multivoxel pattern analyses of brain function and structure," *Behav. Brain Res.*, vol. 259, pp. 330–335, 2014, doi: 10.1016/j.bbr.2013.11.003.
- [14] S. P. Pantazatos, A. Talati, F. R. Schneier, and J. Hirsch, "Reduced anterior temporal and hippocampal functional connectivity during face processing discriminates individuals with social anxiety disorder from healthy controls and panic disorder, and increases following treatment," *Neuropsychopharmacology*, vol. 39, no. 2, pp. 425–434, 2014, doi: 10.1038/npp.2013.211.
- [15] V. Kecman, "Support Vector Machines – An Introduction 1 Basics of Learning from Data," *StudFuzz*, vol. 177, pp. 1–47, 2005.
- [16] V. N. Vapnik, *Statistics for Engineering and Information Science Springer Science+Business Media, LLC*. 2000.



# Decision Support System Scholarship Selection Using Simple Additive Weighting (SAW) Method

Budi Arifitama<sup>1\*)</sup>

<sup>1</sup>Program Studi Teknik Informatika, Universitas Trilogi

Email: [budiarif@trilogi.ac.id](mailto:budiarif@trilogi.ac.id)

**Abstract** – Scholarships are given to students to motivate students and compete with each other in pursuit of the best grades and achievements during their studies. As the name implies, factors such as GPA, competition participation, lecturer recommendations, and organizational participation are the criteria that will be considered for the selection process. In addition, parental income will also be an additional criterion. To minimize errors and reduce bias in the selection process, students who are eligible for scholarships will be assisted by using a Decision Support System (DSS). DSS will support decision making in selecting outstanding scholarship recipients from a pool of alternatives, namely students who register for the outstanding scholarship program by calculating student eligibility based on consideration of the criteria that students have in accordance with predetermined criteria, the alternatives in this research are student from the university. This calculation is carried out using the Simple Additive Weighting (SAW) method which is suitable for use in Multiple Attribute Decision Making (MADM) problems. As a result, each student will get an eligibility score which will influence the final decision. After the ranking of students who are most entitled to a scholarship according to the system calculations are obtained, the final decision will still be taken by the university.

**Keywords** – *Scholarship Awardee, Decision Support System, Simple Additive Weighting (SAW)*

## I. INTRODUCTION

Achievement can be likened to a measure of the success of a student during his academic journey, both at school and college. Scholarships are the provision of financial assistance to students for the continuation of their education and can be one of the things that can motivate students in pursuing achievements[1]. Providing scholarships to outstanding students is not only a reward or a gift, but also triggers academic competition between fellow students. The existence of competition in a learning ecosystem is very important to build the character of students who are persistent and exemplary, so that they can adapt more quickly to the world of work later. In addition, scholarships also provide an alternative to students with financial deficiencies. Scholarships are a cutting-edge solution for universities to provide assistance and awards as well as learning motivation to students which are useful in helping to improve the accreditation of study programs and the reputation of the university[2].

Traditionally, scholarships are awarded after an intensive selection process carried out by the responsible department of the university. This process usually does not take place quickly because many factors can cause the selection of scholarships to be not targeted quickly, for example, such as human error [3]. There are many factors or criteria that must be considered carefully in making decisions, namely in the form of things that directly impact student achievement such as current IPS/GPA scores, increase in GPA compared to last semester's GPA, and competition participation. Not only that, other things such as the family's financial condition that does not meet can also be taken into consideration [4]. One way that can be used to make this selection process faster is by utilizing a Decision Support System.

Decision Support System (DSS) is a system developed with a specific purpose, namely to assist an organization or individual in making a decision. DSS has been widely

applied to help solve problems in the world of education and in general the way it works is to calculate how well an alternative choice is based on the preferences or criteria you want to consider[5]. There are various calculation methods used, such as Simple Additive Weighting (SAW), Weighted Product (WP), and Analytical Hierarchy Process (AHP). Each method has a different way of calculating to get the final result. The SAW method itself can be used to solve decision-making problems that have many attributes (Multiple Attribute Decision Making) whose way of working is by calculating the total weight of each criterion owned by the alternative[6],[7], [8]. Simple Additive Weighting Method is used based on previous research stated that it is mostly suitable to compute decision support system based problem, especially for scholarship awardee.

This study will use the Simple Additive Weighting method to design a Decision Support System that aims to assist the university in selecting students who deserve scholarships. The next section will describe the theoretical basis that will be used in this study, discussing the mathematical formulas of calculations used in the SAW method. In Chapter 3, it is shown the application of the SAW formula in the selection process for scholarship recipients.

## II. RESEARCH METHODOLOGY

Today, every field can take advantage of technology to help facilitate the work carried out in that field. For the case of selecting scholarship recipients, it is also possible to apply a system that can make it easier, namely the decision support system (SPK). This system provides an alternative that can be an aid to decision makers. It can also be said that this system converts existing data into information for decision making from semi-structured problems [9],[10]. The decision support system is intended to facilitate decision making by providing alternatives that can be chosen [11].



Simple additive weight (SAW) is a method that is often used in a decision support system, this method is also known as the weighted addition method. SAW is a method that looks for the weighted sum of the rating criteria on the alternatives for each criterion [12].

The calculation steps using the Simple Additive Weighting (SAW) method:

1. Determining Alternative (Ai)
2. Determine the criteria to be used as a reference in decision making (Cj)
3. Determine the preference weight or level of importance (W) for each criterion
4. Determine the Match Value of each criterion
5. Make a decision matrix (x) obtained from the suitability rating for each alternative (Ai) with each criterion (Cj).
6. Perform the normalization step of the decision matrix (x) by calculating the value of the normalized performance rating (Rij) from the alternative (Ai) on the criteria (Cj) with the formula:

$$R_{ij} = \left\{ \frac{x_{ij}}{\text{Max}\{x_{ij}\}} \right\}$$

If j is an attribute of benefit (benefit)

$$R_{ij} = \left\{ \frac{\text{Min}\{x_{ij}\}}{x_{ij}} \right\}$$

If j is an attribute of cost (cost)

7. The result of normalization (Rij) forms a normalized matrix (R)

$$R = \begin{bmatrix} R_{11} & \dots & R_{1j} \\ \vdots & \ddots & \vdots \\ R_{i1} & \dots & R_{ij} \end{bmatrix}$$

8. The final result of the preference value (Vi) is obtained from the sum of the normalized matrix row elements (R) with the preference weights (W) corresponding to the matrix column elements (W).

With:

- = rank for each alternative
- = weighted value of each criterion
- = normalized performance rating value.

### III. RESULTS AND DISCUSSION

The results of the research on the calculation of decision support systems to determine scholarships with the SAW method can be seen as follows:

#### a. Alternatives

Determine alternatives for the selection of scholarship recipients, namely 23 students who register for the outstanding scholarship program.

Table 1. ALTERNATIVE DECISION

Code	Description
A1	Ade Budiyanto
A2	Dewi Kuswandari
A3	Elvina Usamah
A4	Gabriella Mandasari
A5	Genta Safitri
A6	Hamima Kuswandari
A7	Hasan Hutagalung
A8	Ibrahim Firgantoro
A9	Jarwadi Prasasta
A10	Keisha Puspasari
A11	Mahmud Firmansyah
A12	Nasrullah Wijaya
A13	Nilam Widiastuti
A14	Okta Gunarto
A15	Prima Simanjuntak
A16	Agus Juliansyah
A17	Radit Hutasoit
A18	Saadat Wacana
A19	Sabrina Yuniar
A20	Tedi Hutasoit
A21	Unggul Natsir
A22	Vanya Andriani
A23	Vicky Nurdianti

#### b. Criteria

criteria that become the reference for consideration for the selection of scholarship recipients in the form of academic and non-academic factors are in the following table 2.

Table 2. DECISION CRITERIA

Code	Description
C1	GPA This Semester
C2	Percentage increase in GPA
C3	Organizational participation
C4	Participation in competitions
C5	Lecturer recommendation
C6	Parent Income

#### c. Attribute Criteria



Gives attributes to each criterion that has been determined. There are 2 types of attributes that can be assigned to each criterion, namely *benefits* and *costs*.

- a) *Benefit*, given to criteria that are beneficial or *beneficial*.
- b) *Cost*, given to the criteria that are giving a loss or *cost*.

Table 3. ATTRIBUTE CRITERIA

Code	Description	Attribute
C1	GPA this semester	Benefit
C2	Percentage increase in GPA	Benefit
C3	Organizational participation	Benefit
C4	Participation in competitions	Benefit
C5	Lecturer recommendation	Benefit
C6	Parents income	Cost

c) *Criteria Weight*

Giving weight to each predetermined criterion. The nominal number of weights corresponds to how important the criteria are related to the selection process.

Table 4. WEIGHT CRITERIA

Code	Description	Weight
C1	GPA this semester	5.0
C2	Percentage increase in GPA	1.5
C3	Participation in organizations	1.0

C4	Participation in competitions and the like	2.0
C5	Lecturer recommendations	1.0
C6	Parents income	1.5

d) *Alternative Values on Each Criterion*

Give a weighted value for each alternative to each criterion according to the suitability of the alternative to each of the relevant criteria. attribute type *benefit*, the higher the alternative weight value means the higher the possibility of the alternative being the best choice at the time of calculation, on the contrary, *cost* means the lower the probability.

Table 5. CRITERIA WEIGHTING

Alt.	Criteria					
	C1	C2	C3	C4	C5	C6
A1	3.8	1.0	1.0	2.0	10.0	8.0
A2	3.8	2.0	3.0	3.0	8.0	12.0
A3	3.0	5.0	4.0	1.0	2.0	6.0
A4	2.9	1.0	10.0	1.0	1.0	6.0
A5	3.0	2.0	5.0	1.0	1.0	7.0
A6	3.1	1.0	6.0	4.0	3.0	15.0
A7	3.7	3.0	1.0	10.0	5.0	20.0
A8	3.3	4.0	7.0	2.0	4.0	11.0
A9	3.4	3.0	3.0	8.0	6.0	8.0
A10	2.9	4.0	4.0	4.0	1.0	9.5
A11	3.0	4.0	6.0	1.0	1.0	9.0
A12	2.2	4.0	5.0	1.0	1.0	6.0
A13	3.8	3.0	9.0	4.0	2.0	7.0
A14	3.1	2.0	8.0	3.0	4.0	11.0
A15	3.4	1.0	2.0	2.0	1.0	20.0
A16	2.9	3.0	1.0	1.0	3.0	7.6
A17	1.9	2.0	5.0	1.0	1.0	6.6
A18	2.8	4.0	4.0	4.0	1.0	13.0
A19	3.7	2.0	6.0	7.0	7.0	5.0
A20	2.8	5.0	10.0	2.0	1.0	9.0
A21	2.9	2.0	8.0	8.0	6.0	5.0
A22	3.9	1.0	4.0	1.0	4.0	8.2
A23	3.7	3.0	3.0	2.0	1.0	14.0

e) *Value Normalization*





Normalize the weight value of each alternative to simplify the calculation process. criteria *benefit* use the formula:

$$R_{ij} = \left\{ \frac{x_{ij}}{\text{Max}\{x_{ij}\}} \right\}$$

An example for the first alternative (A<sub>1</sub>):

$$R_{11} = \left\{ \frac{x_{11}}{\text{Max}\{x_{11}\}} \right\} = \left\{ \frac{3.8}{3.8} \right\} = 1.00$$

$$R_{12} = \left\{ \frac{x_{12}}{\text{Max}\{x_{12}\}} \right\} = \left\{ \frac{1.0}{5.0} \right\} = 0.20$$

$$R_{13} = \left\{ \frac{x_{13}}{\text{Max}\{x_{13}\}} \right\} = \left\{ \frac{1.0}{10.0} \right\} = 0.10$$

$$R_{14} = \left\{ \frac{x_{14}}{\text{Max}\{x_{14}\}} \right\} = \left\{ \frac{2.0}{10.0} \right\} = 0.20$$

$$R_{15} = \left\{ \frac{x_{15}}{\text{Max}\{x_{15}\}} \right\} = \left\{ \frac{10.0}{10.0} \right\} = 1.00$$

criteria *benefit* use the formula:

$$R_{ij} = \left\{ \frac{\text{Min}\{x_{ij}\}}{x_{ij}} \right\}$$

An example for the first alternative (A<sub>1</sub>):

$$R_{16} = \left\{ \frac{\text{Min}\{x_{16}\}}{x_{16}} \right\} = \left\{ \frac{8.0}{20.0} \right\} = 0.75$$

So that the normalization value is obtained as shown in the following table.

Table 6. NORMALIZATION VALUE

Alt.	Criteria					
	C1	C2	C3	C4	C5	C6
A1	1.00	0.20	0.10	0.20	1.00	0.75
A2	0.99	0.40	0.30	0.30	0.80	0.50
A3	0.80	1.00	0.40	0.10	0.20	1.00
A4	0.76	0.20	1.00	0.10	0.10	1.00
A5	0.77	0.40	0.50	0.10	0.10	0.86
A6	0.82	0.20	0.60	0.40	0.30	0.40
A7	0.96	0.60	0.10	1.00	0.50	0.30
A8	0.87	0.80	0.70	0.20	0.40	0.55
A9	0.90	0.60	0.30	0.80	0.60	0.75
A10	0.75	0.80	0.40	0.40	0.10	0.63
A11	0.79	0.80	0.60	0.10	0.10	0.67
A12	0.58	0.80	0.50	0.10	0.10	1.00
A13	0.99	0.60	0.90	0.40	0.20	0.86

A14	0.82	0.40	0.80	0.30	0.40	0.55
A15	0.89	0.20	0.20	0.20	0.10	0.30
A16	0.76	0.60	0.10	0.10	0.30	0.79
A17	0.50	0.40	0.50	0.10	0.10	0.91
A18	0.72	0.80	0.40	0.40	0.10	0.46
A19	0.96	0.40	0.60	0.70	0.70	1.20
A20	0.73	1.00	1.00	0.20	0.10	0.67
A21	0.76	0.40	0.80	0.80	0.60	1.20
A22	1.02	0.20	0.40	0.10	0.40	0.73
A23	0.98	0.60	0.30	0.20	0.10	0.43

#### f) Final Results

The last stage is to calculate the final value by finding the total *sum* of the results of the multiplication of the alternative normalization values with the appropriate weighting criteria preferences.

$$V_i = \sum_{j=1}^n W_j R_{ij}$$

For example for the first three alternatives (A<sub>1</sub>, A<sub>2</sub>, A<sub>3</sub>):

$$V_1 = (1.00 \times 5.0) + (0.20 \times 1.5) + (0.10 \times 1.0) + (0.20 \times 2.0) + (1.00 \times 1.0) + (0.75 \times 1.5) = 7.925$$

$$V_2 = (0.99 \times 5.0) + (0.40 \times 1.5) + (0.30 \times 1.0) + (0.30 \times 2.0) + (0.80 \times 1.0) + (0.50 \times 1.5) = 8.011$$

$$V_3 = (0.80 \times 5.0) + (0.40 \times 1.5) + (0.30 \times 1.0) + (0.30 \times 2.0) + (0.80 \times 1.0) + (0.50 \times 1.5) = 7.790$$

After all the final alternative values are calculated, then they are sorted so that a ranking list is obtained as shown in the following table.

Table 7. FINAL RESULTS RANKING

Kode	Alternatif	Nilai
A19	Sabrina Yuniar	9.916
A21	Unggul Natsir	9.219
A13	Nilam Widiastuti	9.046
A9	Jarwadi Prasasta	9.039
A7	Hasan Hutagalung	8.766
A2	Dewi Kuswandari	8.011
A1	Ade Budiyo	7.925
A8	Ibrahim Figrantoro	7.875
A3	Elvina Usamah	7.790
A20	Tedi Hutasoit	7.675
A22	Vanya Andriani	7.489



A14	Okta Gunarto	7.313
A23	Vicky Nurdityanti	7.225
A10	Keisha Puspari	7.188
A11	Mahmud Firmansyah	7.037
A4	Gabriella Mandasari	6.906
A18	Saadat Wacana	6.814
A6	Hamima Kuswandari	6.694
A5	Genta Safitri	6.557
A16	Prima Simanjuntak	6.490
A12	Nasrullah Wijaya	6.413
A15	Prima Simanjuntak	5.912
A17	Radit Hutasoit	5.257

#### IV. CONCLUSION

The Decision Support System made using the SAW method has succeeded in helping make decisions to determine which students are most worthy of receiving outstanding scholarships. The SAW method has been successfully used to help solve problems that are Multiple Attribute Decision Making (MADM) or decision-making problems with many attributes. By using the SAW method, the results obtained in the form of a list of alternative rankings that are considered the most suitable for receiving scholarships, so that they can help facilitate the decision-making process.

#### REFERENCES

- [1] V. Tasril, "Sistem Pendukung Keputusan Pemilihan Penerimaan Beasiswa Berprestasi Menggunakan Metode Elimination Et Choix Traduisant La Realite," *INTECOMS J. Inf. Technol. Comput. Sci.*, 2018, doi: 10.31539/intecomsv1i1.163.
- [2] R. Wati, S. A. Winanda, H. Margahana, and E. Dwiyani, "Sistem Pendukung Keputusan Penerimaan Pegawai Dengan Metode Weighted Product Berbasis Web," *SPEKTRUM J. Pendidik. Luar Sekol.*, 2020.
- [3] V. V. Wang, A. S. Sukamto, and E. E. Pratama, "Sistem Pendukung Keputusan Seleksi Mahasiswa Penerima Beasiswa BBP-PPA dengan Metode TOPSIS pada Fakultas Teknik UNTAN," *J. Sist. dan Teknol. Inf.*, 2019, doi: 10.26418/justin.v7i2.29656.
- [4] J. Fitriana, E. F. Ripanti, and T. Tursina, "Sistem Pendukung Keputusan Pemilihan Mahasiswa Berprestasi dengan Metode Profile Matching," *J. Sist. dan Teknol. Inf.*, 2018, doi: 10.26418/justin.v6i4.27113.
- [5] O. Veza and N. Y. Arifin, "SISTEM PENDUKUNG KEPUTUSAN CALON MAHASISWA NON AKTIF DENGAN METODE SIMPLE ADDITIVE WEIGHTING," *J. Ind. Kreat.*, 2020, doi: 10.36352/jik.v3i02.29.
- [6] R. T. Aprilia Triase; Sriani, Sriani, "Penentuan Tempat Menginap Dengan Menggunakan Fuzzy

- Multiple Attribute Decision Making," *Algoritm. J. Ilmu Komput. Dan Inform.*, 2017.
- [7] I. Mulyadin and D. S. Winarso, "Sistem Pendukung Keputusan Pemilihan Smartphone Menggunakan Metode Simple Additive Weighting," *CAHAYATECH*, 2019, doi: 10.47047/ct.v7i2.13.
- [8] R. Fauzan, Y. Indrasary, and N. Muthia, "Sistem Pendukung Keputusan Penerimaan Beasiswa Bidik Misi di POLIBAN dengan Metode SAW Berbasis Web," *J. Online Inform.*, vol. 2, no. 2, p. 79, 2018, doi: 10.15575/join.v2i2.101.
- [9] R. T. W. Nugraha, B. Arifitama, and Y. Yaddarabullah, "Decision Support System for Rewarding Courier Employees in North Jakarta Using Profile Matching," *J. Integr.*, 2021, doi: 10.30871/ji.v13i1.2535.
- [10] E. Zuraidah and L. Marlinda, "System Penunjang Keputusan Pemilihan Tempat Wisata Lombok Menggunakan Metode Preference Ranking Organization For Enrichman Evaluation (PROMETHEE)," *J. Tek. Komput.*, 2018.
- [11] T. Susilowati, E. Y. Anggraeni, Fauzi, W. Andewi, Y. Handayani, and A. Maselena, "Using Profile Matching Method to Employee Position Movement," *Int. J. Pure Appl. Math.*, vol. 118, no. 7, 2018.
- [12] G. E. Rinaldhi, "Penerapan Metode Simple Additive Weighting ( SAW ) Untuk Sistem Pendukung Keputusan Penentuan Penerimaan Beasiswa Bantuan Siswa Miskin ( Bsm ) Pada Sma Negeri 1 Subah Kab . Batang," *Jur. Tek. Inform. Fak. Ilmu Komput. Univ. Dian Nuswantoro Semarang*, pp. 1–9, 2011.



# Enterprise Content Management (ECM) System Architecture for Capital Project at Oil and Gas Company

**Arief Herdiansah**

Informatics Department, Engineering Faculty, Muhammadiyah Tangerang University  
Email: arief\_herdiansah@umt.ac.id

**Abstract** – The Enterprise Content Management System (ECM) is an improvement over the Document Management System (DMS) solution. In a DMS solution, corporate documents will be managed at the document level (access rights and document distribution processes) while in an ECM solution, documents can be managed down to the level of content contained in the document. Companies who are engaged in the oil and gas industries, are facing intense business competition and to be able to win the competition, companies must be able to increase their HSE (Health, Safety, and Environmental) supervision. Document content management systems are needed by companies that have capital projects, including oil and gas companies. Before a company builds and implements an ECM solution to manage project capital documents, it is necessary to build a system design architecture so that the built ECM solution can meet user needs in managing documents down to the content level. ECM is one of the enterprise solutions, so an Enterprise Architecture (EA) scale system design is needed with a focus on Digital Enterprise Architecture (DEA). This research produces a reference in the process of making ECM architecture in order to manage project capital documents in oil and gas companies.

**Keywords** – ECM, DMS, Digital Document, Capital Project, Oil and Gas Company

## I. INTRODUCTION

Oil and gas companies have projects that are included in the capital project category which is included in the category of capital projects which are long-term capital-intensive investment projects with a purpose to build upon, add to, or improve a capital asset [1], [2]. Capital projects in the process industries involve the construction of physical plant facilities and materials processing equipment to produce a new product for expected profit or alternatively to maintain or develop operating-level capabilities [3], [4]. Capital projects are defined by their large scale and large cost relative to other investments that involve less planning and resources [5], [6]. As one of capital project, the oil and gas company needs a system to support the business activities of the oil industry, cater multiple phases from detail design, execution, and operations, and also support document management for corporate. Organizations that run capital projects face several critical information problems related to managing information contained in project documents, it likes to lose information lost between project phases so that project completion is delayed and increases Project costs, due to lack of integration between systems and workflow processes that use project document [7], [8]. For this reason, a document management system is needed that is more than just digital archive management but also a solution for managing content contained in company documents [9], [10]. Therefore, it is needed to implement Enterprise Content Management (ECM) solution to meet and accommodate the need for digital document management and document content management owned by the company and perform document control process for both capital projects (feed, EPCI, and operation) with considering costs effectiveness in implementing ECM solutions.

In the process of implementing the ECM solution, it is necessary to design an ECM system architecture so that it can describe the details of the ECM solution design which consists of functional architecture, infrastructure and information architecture, in order to describe the ECM solution design to align with the improvement of digital corporate document management including business process enhancement and could be a reference technical design document for ECM enhancement in a future. The ecm system is a development of a DMS (Document Management System) solution [11], [12]. DMS is a softcopy document management system responsible for the efficient and systematic control of the creation, receipt, maintenance, use and disposition of records, including processes for capturing and maintaining evidence of and information about business activities and transactions in the form of records [13], [14].

This research is different from previous research on ECM, where in this study the researcher specializes in the process of designing an ECM system for managing capital project documents in oil and gas companies.

## II. RESEARCH METHODOLOGY

The type of research used in this research is descriptive quantitative research method by finding and collecting information about the implementation of the ECM system in oil and gas companies. The information is then clearly defined the goals to be achieved, plans the approach, collects data as material for making a planning report on the design of the ECM system capital project. Data collected by conducting a direct discussion process with related parties where the ECM design was developed. In addition, researchers also use input from information obtained from books and previous research [15]–[17].



The two things that were first formulated from the results of data collection were the ECM Business Use Case for capital projects and the Enterprise Architecture Principles.

### 2.1. ECM Business Use Case

ECM allows companies to have a single platform in managing project capital documents digitally. The benefits of using a single ECM Platform enable you to collaboratively create, manage, deliver, and archive the content that drives business operations. The ECM Platform makes it possible to distribute all of this content across internal and external systems, applications, and End-User communities. The ECM system has document management features that can be applied to business use cases, project document workflows including the use of standard features of a DMS solution as shown in table 1 below:

Table 1. ECM business use case

No	Use Case Name	Remarks
<b>1</b>	<b>General System Functionality</b>	
1.1	Login to ECM	DMS & ECM standard feature
1.2	Create New Project/Domain	DMS & ECM standard feature
1.3	Add Members to Project/Domain Group	DMS & ECM standard feature
1.4	Document Number Generation from Creating/Importing Document	DMS & ECM standard feature
<b>2</b>	<b>As-built/Project Data Management</b>	Integrated with asset operation application
<b>3</b>	<b>Document Management</b>	
3.1	Search Document	DMS & ECM standard feature
3.2	Check Out and Check In Document	DMS & ECM standard feature
3.3	View Revisions of Document	DMS & ECM standard feature
3.4	View Versions of Document	DMS & ECM standard feature
<b>4</b>	<b>Transmittal Management</b>	
4.1	Distribution Matrices	DMS & ECM standard feature
4.2	Project Document (External Document Flow)	DMS & ECM standard feature
4.3	Corporate Document (Internal Document Flow)	DMS & ECM standard feature
<b>5</b>	<b>Integration with design/engineering tools</b>	
5.1	Integrate Other Engineering Tools with ECM (ECM as a document repository)	Need to create an integration script
<b>6</b>	<b>Intelligent Document</b>	
6.1	Search Content of the Document	ECM can search the content of the document
<b>7</b>	<b>Tag Management</b>	ECM has relationship function between documents. So, one document can have relationship with another document.
7.1	Add Relationship	DMS & ECM standard feature
7.2	View Relationship	DMS & ECM standard feature
3.5	View Relationships of Document	DMS & ECM standard feature
3.6	View Locations of Document	DMS & ECM standard feature
3.7	Export Document	DMS & ECM standard feature
3.8	Subscribe Document	DMS & ECM standard feature
3.9	Add Relationship	DMS & ECM standard feature
3.10	View Comparison	DMS & ECM standard feature



## 2.2. Enterprise Architecture Principles

Enterprise Architecture (EA) and Digital Enterprise Architecture (DEA) are different information technology architecture designs. EA focuses on structuring the company based on the main framework of IT tools, while DEA focuses on document management and lifecycle so that documents can be easily accessed, modified, and managed at any time following company developments [18]. EA is increasingly being developed to manage information systems for business purposes [19]. EA has now inspired IT architects and technology innovators to design and deliver new operating models, resulting in businesses that are now the center stage [20].

The Architecture Principle of EA consists of several layers, namely: Business Principles, Data Principles, Application Principles and Technology Principles [18], [21]. ECM architecture is closely related to Data Principles (Data is an Asset, Data is Shared, Data is Accessible, Data Trustee, Common Vocabulary and Data Definitions, Data Security) and Technology Principles (Requirements-Based Change, Responsive Change Management, Control Technical Diversity, interoperability) [18].

## III. RESULTS AND DISCUSSION

### 3.1. ECM Architecture for Capital Project

The ECM architecture involves a multi-tiered Client/Server where the software system will serve as a platform for the development of business-specific solutions: enabling users to customize with in order to define specific data objects; create forms and reports; define user workflows with face application programming (API) that enables system automation [22], [23].

ECM does not refer to a single technology, but ECM is a system that combines methods, tools, and strategies in order to support the process of storing and retrieving digital documents, managing content, and sending information throughout the life cycle of a document [24], [25].

ECM system for management capital project document includes these key capabilities:

- a) Project information governance through project role-based security
- b) Simple engineering document lifecycles,
- c) Revision Codes
- d) Access control based on discipline and document classification types
- e) Easy-to-use configuration options that give document controllers control over their project content and collaboration between disciplinary engineers
- f) Automatic numbering and properties population, inherited from the project work breakdown folder structure hierarchy and document templates
- g) Folder templates that can encapsulate project work breakdown folder structure hierarchy configuration rules (permissions, permissible documents, Property Inheritance, etc).

The application of ECM for the management of

capital project documents essentials define the solution's key features, which are managed with a common configuration model that provides a consistent and integrated system for engineering information management.

Figure 1 is a high-level view of the current (and planned) ECM Solution capabilities. The ECM Essentials Solution represented is typically the first phase for most Organisations.

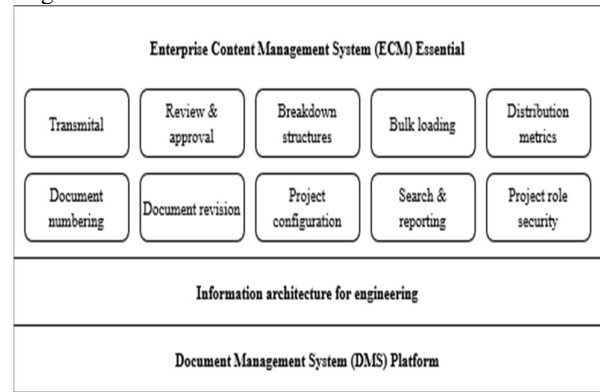


Figure 1. ECM High-Level Solution Architecture

ECM for capital project essentials provide the standard basic feature as shown on the image above. If the customization required, ECM provide the SDK file which can be used by the ECM administrators/developers to start customizing the ECM system. Also for integration part, ECM provide some extensions with the 3rd party application which can be used based on customer's requirement.

### 3.2. ECM System Architecture for Capital Project at Oil and Gas Company

The first thing to create is the infrastructure architecture as shown in Figure 2. Technically, users can use web service application as a front-end server to serve site requests directly. However, in a production environment, user may want to use some web servers as front-end to route the requests to the web service application. Using a web server to handle the requests gives performance and security benefits. If user using web service application HTTP as a front-end web server, then must consider securing that as well. Web service application HTTP must be publish to HTTPS and use different URL if there is an external user want to access the URL. Web service application HTTP must be publish to HTTPS and use different URL if there is an external user want to access the URL

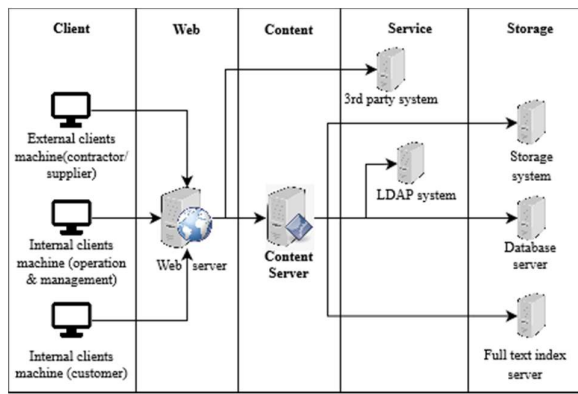


Figure 2. ECM for Capital Project Infrastructure Architecture

The implementation of ECM in oil and gas companies will involve the functions of the company's Document Management, Project manager, Controller, Engineering Design (FEED), Procurement and Construction (EPC) and EPC Suppliers/Contractors.

The next stage is to create a functional architecture as shown in figure 3 the functional architecture shows that SCM Solution allows an Organization and its contractors to easily interact on digital document management solutions and manage them. The new ECM shall be accessed by Intranet User, Extranet User, and Internet User. For Internet and Extranet User, will be using port to get the access of it.

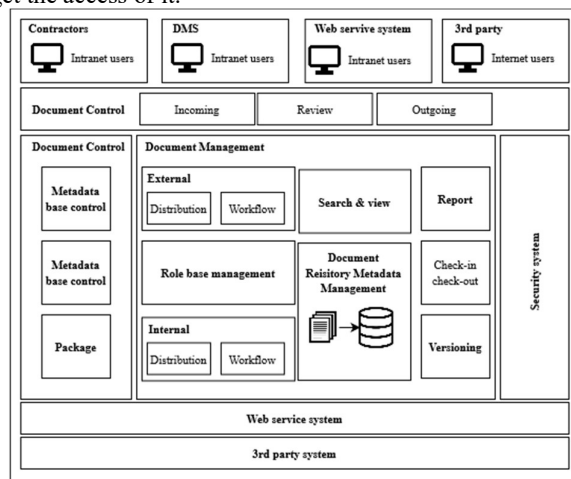


Figure 3. ECM system functional architecture

. Based on the figure above, the two big process of ECM functionality, can be divided into document control and document management.

- 1) In document control process, the documents can be controlled through validation process against the document loading file when perform bulk loading, updating metadata based on right permission, and transmittal packages by creating Incoming, Internal (Review/Approval), and Outgoing Transmittal Types, with their status being easily tracked and references between Transmittals and Content.
- 2) In document management process, internal and external project have the transmittal process

(distribution and workflow). The document controller will typically manage the process of a transmittal process from start to finish and they are the context throughout the process. For the user who involved in transmittal workflow (user who becomes reviewer and responsible person) can be configured through role base management. Role base management uses content file and metadata from Document Repository.

- 3) Role base management can also be used to manage the permission, so that user with right permission can search or view document, view the report, check in and check out the document, and also do versioning against the document.
- 4) The documents or files have metadata that is stored in the repository of ECM.
- 5) ECM can be integrated with other system (third party application) by utilizing web service interface (example integration with SharePoint (SharePoint Connectors) and PEGA (PEGA Page Connector). EDMS exposes its entire set of services through standard-based web services and java-based API. Customizations and integrations are achieved through these basic set of classes and interfaces.
- 6) Through the web service systems, third party application can upload, view, and edit the document using the applied URL web services from EDMS

#### IV. CONCLUSION

Based on the information and descriptions generated from research conducted related to Enterprise Content Management (ECM) system architecture for capital projects at oil and gas companies, it can be concluded:

- 1) Capital project run by an oil and gas company is a large project that involves a team with a variety of functions and documents produced from these functions and it requires an ECM system that is designed according to the flowflow needs of each user.
- 2) The design of the ECM system must refer to the business use cases of the users in each function
- 3) The design of the ECM system must be able to accommodate the involvement of contractors, internal teams, project teams and 3rd parties with strong data access security.
- 4) The research conducted has not discussed and provided details of the digital document storage capacity needed to support the ECM system capital project in oil and gas companies.

#### REFERENCES

- [1] S. Yun, J. Hoi, D. P. de Oliveira, and S. P. Mulva, "Development of performance metrics for phase-based capital project benchmarking," *International Journal of Project Management*, vol. 34, no. 3, pp. 389–402, 2016.



- [2] B. Sanchez and C. Haas, "Capital project planning for a circular economy.," *Construction Management and Economics*, vol. 36, no. 6, pp. 303–312, 2018.
- [3] M. Boukar and I. Muslu, "Administration And Academic Staff Performance Management System Using Content Management System (Cms) Technologies," *IEEE - International Conference on Electronics, Computer and Computation (ICECCO)*, pp. 151–154, 2013, doi: 10.1109/ICECCO.2013.6718251.
- [4] C. S. Young and D. Samson, "Project success and project team management: Evidence from capital projects in the process industries," *Journal of Operations Management*, vol. 26, pp. 749–766, 2008.
- [5] H. L. Chen, "Performance measurement and the prediction of capital project failure," *International Journal of Project Management*, vol. 33, no. 6, pp. 1393–1404, 2015.
- [6] E. C. Ness and D. A. Tandberg, "Use of quality function deployment in civil engineering capital project planning," *The determinants of state spending on higher education: How capital project funding differs from general fund appropriations*, vol. 84, no. 3, pp. 329–362, 2013.
- [7] A. Herdiansah, D. Nurnaningsih, Y. Sugiyani, and T. Handayani, "Implementation transmittal solutions at capital project to increase the effectiveness," in *ICSTEIR 2020 IOP Conf. Series: Materials Science and Engineering*, 2020, pp. 1–11.
- [8] P. J. Laudon, C.K., Laudon, *Management Information System: Managing The Digital Firm*. USA: Pearson Education, Inc, 2012.
- [9] S. Huang and X. Meng, "Research and Application of Integration Solution for Enterprise- Level Heterogeneous Document Management System," *Journal of Physics: Conference Series*, vol. 1621, no. 1, pp. 1–6, 2020.
- [10] M. A. K. Nagar, L. A. Rahoo, H. A. Rehman, and S. Arshad, "Education Management Information Systems in the Primary Schools of Sindh a case study of Hyderabad Division," *2018 IEEE 5th International Conference on Engineering Technologies and Applied Sciences, ICETAS 2018*, pp. 1–5, 2019, doi: 10.1109/ICETAS.2018.8629249.
- [11] V. L. Orlov and E. A. Kurako, "Electronic document management systems and distributed large-scale systems," in *2017 Tenth International Conference Management of Large-Scale System Development (MLSD)*, 2017, pp. 1–5.
- [12] M. Başbüyük and A. Ergüzen, "Electronic Document Management System for Kırıkkale University," *Unified Journal of Computer Science Research*, vol. 1, no. 2, pp. 8–15, 2015.
- [13] H. T. Pho and T. Tambo, "Integrated management systems and workflow-based electronic document management: An empirical study," *Journal of Industrial Engineering and Management (JIEM)*, vol. 7, no. 1, pp. 194–217, 2014.
- [14] S. J. Mary and Usha, "Web Based Document Management Systems in Life Science Organization," in *2015 Online International Conference on Green Engineering and Technologies (IC-GET 2015)*, 2016, pp. 1–6.
- [15] Sudaryono, *Metodologi Riset di Bidang IT: Panduan Praktis, Teori dan Contoh Kasus*, Ed.1. Yogyakarta: Andi Offset, 2015.
- [16] Tohirin, *Metode Penelitian Kualitatif dalam bimbingan dan konseling*. Jakarta: Raja Grafindo Persada, 2011.
- [17] Sugiyono, *Metode Penelitian Kombinasi (mixed Methodes)*, 1st ed. Bandung: Alfabeta - Bandung, 2013.
- [18] I. S. Rozas, K. Khalid, N. Yalina, N. Wahyudi, and D. Rolliawati, "Digital Enterprise Architecture for Green SPBE in Indonesia," *CCIT (Creative Communication and Innovative Technology) Journal*, vol. 15, no. 1, pp. 26–24, 2020.
- [19] F. Gampfer, A. Jürgens, M. Müller, and R. Buchkremer, "Past, current and future trends in enterprise architecture—A view beyond the horizon," *Computers in Industry*, vol. 100, pp. 70–84, 2018.
- [20] K. Costello, "The Evolution of Enterprise Architecture - Smarter With Gartner," IASA An Assosiation for All IT Architects," <https://www.gartner.com/smarterwithgartner/the-evolution-of-enterprise-architecture>, 2020.
- [21] The Open Group, "The TOGAF Standard, Version 9.2 Part III: ADM Guidelines and Techniques," <https://pubs.opengroup.org/architecture/togaf9-doc/arch/chap20.html>, 2018.
- [22] K. Normantas and V. Gediminas, "Extracting Business Rules from Existing Enterprise Software System," in *Deriving business rules from the models of existing information systems*, 2011, pp. 1–15.
- [23] C. Maican and R. Lixandriou, "A system architecture based on open source enterprise content management systems for supporting educational institutions," *International Journal of Information Management*, vol. 36, no. 2, pp. 207–214, 2016.
- [24] E. Mixon, "What is enterprise content management? Guide to ECM," <https://www.techtarget.com/searchcontentmanagement/definition/enterprise-content-management-ECM>, 2020.
- [25] M. Rahimi and M. Rosman, "Reviewing the Concept of Enterprise Content Management (ECM)," *Journal of Digital Information Management*, vol. 18, no. 4, pp. 125–138, 2020.

