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

Implementation Internet of Things for Feeding Catfish Water Quality Analysis Using Linear Regression and K-Nearest Neighbor

Yaddarabullah[1], Egie Hermawan[2]
[1,2]*Department of Informatics Engineering, Universitas Trilogi*
Email : yaddarabullah@trilogi.ac.id, egihermawan@trilogi.ac.id

Abstract— Cultivation of catfish *(Clarias Gariepinus)* is a promising business field and also a very productive activity because public interest in catfish is high. This factor is observed by market demand for catfish which is increasing from year to year. In catfish farming, you must pay attention to the acidity of the water (pH), temperature, and oxygen levels, which can change if too much feed is given. This can cause catfish seedlings to die and affect the catfish harvest. Catfish farmers often provide excessive food which causes many catfish seeds to die. This research will conduct a study on an Internet Of Things technology that can be used to monitor the acidity level in water pH, temperature, and oxygen levels as well as feed fish. The Internet of Things is very influential for monitoring the quality of catfish ponds by distributing information data resulting from sensor monitoring. The data obtained will be predicted for water quality in the pond by implementing a Linear Regression method. Furthermore, the acquisition of data from the predictions that have been carried out will be processed again to go to the next phase, namely classifying with the K-Nearest Neighbor algorithm method to carry out the identification phase of water types based on the nearest neighbors. This prediction is used to anticipate and notify catfish farmers through applications if there is a water acidity level (pH), temperature, and oxygen and feed levels that have run out.

Keywords— *Water Quality Analysis, Internet of things, K-Nearest Neighbor*

I. INTRODUCTION

Cultivation of catfish (*Glarias Gariepinus)* is a very productive activity because public interest in catfish is high. In helping the success in the cultivation of catfish farm, good, management is needed. Various attempts have been made to develop [1]. One of the main factors involved in cultivating catfish is feeding regularly, because it greatly affects the growth of the catfish itself. Excess feed (overfeeding) can lead to various adverse effects including faster growth of catfish seedlings at the beginning of cultivation, reduced capacity and quality in soil and water, as well as disease infection and can cause many catfish seedlings to die. While the lack of feed (underfeeding) low can result in disruption of growth in catfish seedlings. The two factors above can affect the yield of the catfish, therefore technological developments are very important and have a very good impact on these problems. Technology comes from a manual system which then operates and develops rapidly and increases towards an automated and integrated system [2]. The catfish pond in Mundusari Village, Subang Regency, West Java has a length of 3 meters and a width of 2 meters, and has a water depth of about 50 cm. The quality of pond water for catfish farming can be reviewed in various sources such as river water, then lakes, the effect of feed given to fish and wells. In catfish pond cultivation in Mundusari Village, Subang Regency, West Java, two types of feed are given to the catfish, namely pellets as the main feed and tiren chicken when the fish farmers are experiencing a decline in economic conditions. In addition to the various kinds of water, rainwater can also be used in catfish farming, but due to the high acidity and cold temperatures, separate treatment must be obtained before being used for cultivation. Water for the development of catfish should not be contaminated by various kinds of waste, because it will have an impact and endanger the life of the catfish. The quality of the water in the catfish pond must be considered and the qualifications for catfish cultivation include clear color and not polluted. The balance of pH levels in water is the most important tip to become a good predictor of catfish cultivation benchmarks. It is very mandatory in catfish cultivation to maintain the stability of the pH of the water with levels of 5.5 - 7.5. The condition of the pH of the water can be said to be neutral if it is in the pH level of 6-7, if the condition of the water is in the pH range above 7 then the water is in the alkaline category, if the pH level of the water is below 6 then the water is in the acidic category. In acidic conditions, fungi and bacteria will grow and multiply, catfish have resistance at pH level conditions of 5.5 – 7.5. Water with an environment that does not meet the stated criteria can create a source of disease which in the future can endanger the growth of the catfish [3]. These factors can be implemented by monitoring the condition of the pH level of the fish pond using the use of an internet of things technology. In taking the data it is necessary to obtain an information on the condition of the pH

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

level in the catfish pond. The Internet of Things is a series of objects in the form of hardware that can exchange information, both service operators and other devices that have been connected to a system until finally they can provide greater impact. The results from these sensors will be used to predict the quality of pool water within the next 14 days, in predicting the quality of the pool water, a method is needed, while the method used is Linear Regression [4]. The application of the system to the K-Nearest Neighbor method is by comparing the data that has been tested and the data that has been trained on the database [5]. Based on these problems, a study was conducted to create a feeding system for catfish and analyze water quality based on temperature, pH, and also ammonia in the water by applying a Linear Regression and K-Nearest Neighbor method based on Smart Mobile Internet Of Things. The results of this study are expected to help farmers in preventing the death of catfish seedlings in the context of cultivating catfish.

II. RESEACRH METHODOLOGY

The methodology used in this research consists of several stages. This stage will predict and also detect water quality in the catfish pond based on the pH level of the water, temperature, and ammonia.

A. Data Acquisition

The data acquisition process is collecting data values from the device to display the data information obtained. The data acquisition system in this study obtains data from IoT devices by using temperature sensors, pH sensors and ammonia sensors. The location used for data collection is Pamanukan Catfish Farm, Subang Regency. The working process of the data collection system on the IoT weather station tool is shown in Figure 2 below :

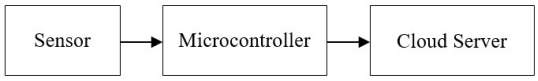

Fig. 1. Work Process.

In Figure 1 above is the working process of the data acquisition system on the IoT water acidity (pH) sensor tool which is used to obtain data results as information. Each sensor device takes data consisting of 3 objects, namely temperature, pH and ammonia. The data obtained will go to the Rasberry Pi to be sent to the Cloud Server as a data storage container. Data consisting of temperature, air humidity and soil moisture obtained through sensors will be displayed on the website, then the website display will be changed to mobile. In this mechanism, data can enter the stage of the forecasting process.

B. Forecasting

Linear regression is also a statistical method whose function is to test the degree between the causal factor (*x*) and the outcome variable. The causal factor is usually represented by *X*, and the outcome variable is represented by *Y*. Simple linear regression or commonly abbreviated as SLR (Simple Linear Regression) is also one of the statistical methods used to predict or predict the quality and quantity characteristics in production. The general equation for the simple linear regression method in this study can be seen in the equation 1 model below [6]:

$$\boldsymbol{Y = a + bX} \tag{1}$$

The coefficients of equations *a* and *b* can be determined using the least squares method. The least squares method is a method for determining the coefficients of an equation. Is the smallest value obtained from the sum of the squares between the points with the smallest regression line that can be searched. Therefore, the application of the model in the above equation is carried out to determine the value of an effect variable (*y*). After that, to find a coefficient value (*b*), a formulation model is needed which can be seen in the model equation 3.2, which is as follows:

$$\boldsymbol{b = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2}} \tag{2}$$

The application of the model in the above equation is carried out to obtain a coefficient value (*b*), namely *n* = total data (sum *x* multiplied by number of variables *y*) - (sum *x*) multiplied (number of variables *y*) / overall data (*n*) multiplied by (sum *x* times the number of *x*) – (sum of *x*)2 [7]. In the next step to find the value for constant (a) a formula is needed in the model equation 3.3 as follows:

$$\boldsymbol{a = \frac{\sum y}{n} - \frac{b\sum x}{n}} \tag{3}$$

The application of the model in the above equation is carried out to obtain a constant value (*a*), which is the total number of variables y – the result of the coefficient value (*b*) multiplied by the total number of variables x and divided by the number of data (*n*). The steps for carrying out calculations on a simple linear regression are as follows:

1) Determine a causal factor variable (*x*) and an effect variable (*y*) on a data.
2) Compile the data that will be used as the causal factor variable (*x*) and the effect variable (*x*).
3) Accumulating related data.
4) Performs a calculation to find a value on the variables *x*2, *y*2, *xy* value, and the whole of each.
5) Look for the value of the coefficient (*b*) and the value of the constant (*a*) based on the formula model equation 3.1 and model equation 3.2 above.

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

6) Create an equation model for simple linear regression by looking for MSE (Mean Squared Error).

7) Carry out a prediction implementation with the error percentage obtained on the causal factor variable or the effect variable [8].

Based on the description above, a flow diagram is obtained to carry out the process of forecasting stages, which aims to determine the predicted value for the next 14 days by collecting variable data obtained from Internet of Things tools. The next step is to find a value of x2 which is designed to find the value of the independent variable and find the value of the dependent variable of x and y. After looking for this value, the next step is to find the value of the constant (a) and the value of the coefficient (b). If the value has been obtained, a search for the predicted value and MSE (Mean Squared Error) value will be carried out. After that, the next process is to find the predicted value for the next 14 days. The data obtained from the Internet of Things tool can be seen in Table 1, as below:

TABLE 1. WEATHER TEMPERATURE DATA

Day (x)	Temp (y)	X2	xy
1	27.3	1	27.3
2	27.4	4	54.8
3	27.9	9	83.7
4	28.6	16	114.4
5	28.8	25	144
6	28.7	36	172.2
7	28.3	49	198.1
8	28.8	64	230.4
9	29	81	261
10	29.2	100	292
11	28.9	121	317.9
12	28.9	144	346.8
13	27.6	169	358.8
14	21.6	196	302.4
105	391	1015	2903.8

Table 1 shows that the temperature data obtained will be used as simulation data for linear regression analysis calculations, and will be predicted in the next 14 days or for two weeks based on the amount of data obtained. It can be seen at the end of table x that there is a total of 105, and in the table of temperature (y) there is a total of 391. This number is not the number of days or the temperature that is totaled but a total number of variables x and variable y used for linear regression calculations. The total data is obtained based on the formula for the linear regression equation model, not the total number of day data (x) or temperature data (y). The next step is to find a value of x2 that will be used to determine a value of b, which is a coefficient variable (x). Regarding the calculation applied, the value of an x will be multiplied by the number itself or the multiplication of the square, then the total number for the x2 variable is 1015. In the next step, doing a calculation to find a value for the xy variable in the temperature data, the total total is obtained. at xy value is 2903.8. The next step is to search for a variable coefficient value at x (b) based on the temperature variable using a formula in the following equation 2 model:

$$b = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2} \quad (4)$$

The next step is to search for a constant variable (a) by implementing a formula in equation 3 as shown below:

$$a = \frac{\sum y}{n} - \frac{b\sum x}{n} \quad (5)$$

The next step is to do a search on the linear regression equation based on the temperature variable by implementing a formula listed in the equation 1 model as follows:

$$Y = a + bX \quad (6)$$

After getting the results of the linear regression equation, the next step is to perform calculations to find a prediction. The method used is Root Mean Square Error (RMSE). MSE estimates a prediction method by squaring the error which is then executed by an addition based on the number of observations [9]. This MSE can minimize the biggest prediction error. The formula for the RMSE can be seen in the equation 4 model as follows:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(yi-\hat{y})^2}{N}} \quad (7)$$

In the prediction search by looking for the average error based on the equation model above, namely by reducing the predicted value that has been obtained with the actual value [8].

$$RMSE = \sqrt{\frac{\sum_{i=1}^{14}(28.74857143 - 27.3)^2}{14}}$$

Based on the acquisition of the MSE value, a data prediction for the 1st variable (y) is obtained which can be seen in Table 3 as follows:. The results obtained for 2 weeks are calculated by applying the formula contained in the equation 1 model as follows:

$$Y = a + bX \quad (7)$$

In the prediction search above, multiplying the day (x) that will be predicted, then the predicted temperature on the 15th day is 26.98. The temperature prediction obtained can be seen in Table 2 as follows:

TABLE 2. RESULT OF PREDICTION EXPERIMENTS

Day	Temperature Forecast
15	26.98
16	26.85
17	26.73
18	26.60
19	26.47

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

Day	Temperature Forecast
20	26.35
21	26.22
22	26.09
23	25.97
24	25.84
25	25.72
26	25.59
27	25.46
28	25.34

Table 2 is the result of the forecasting stage that will be used for the next 2 weeks based on data on temperature. Obtaining the highest forecasting results is on the 15th day with the acquisition value of 26.98 degrees Celsius.

C. K-Nearest Neighbor (KNN) Classification

K-Nearest Neighbor (K-NN) is a collection of instance-based learning. This algorithm is also an instance-based learning technique. KNN is solved by finding the group of k objects that are closest (similar) to the objects in the new data or test data in the training data. K-Nearest Neighbor can provide benefits and provide good accuracy, so that the recommendations generated according to the needs or interests of users[10]. The K-Nearest Neighbor (KNN) algorithm is a data collection classification method based on previously classified learning data. Included in supervised learning, where the results of new query instances are classified according to the most common categories in KNN. The algorithm determines the KNN based on the shortest distance from the test sample to the training sample. After taking the KNN, most of the KNN will be used as a test sample prediction. Usually, the nearest neighbor or far neighbor is calculated based on the Euclidean distance*Case folding*: a stage to uniform a sentence or text into a standard. Process case folding used so that the system does not process a word that has the same meaning even though the writing pattern is different [11]. The steps for using the KNN method are described as follows:

1) Determine the parameter k.
2) Calculates the distance between the data to be evaluated with all training data.
3) Line up the distance formed (ascending order).
4) Determine the closest order k.
5) Match the appropriate class.
6) Find the number of classes from the nearest neighbor, and set this class as the data class to evaluate [12].

The search for the value of k is determined by the equation model (3.5) as follows:

$$k = \sqrt{n} \tag{8}$$

The application of the above formula equation model is used to determine the value where n is the number of sample data. The value of k is basically the total of odd numbers in the selection process which serves to prevent the same distance value from occurring. The next stage is data normalization, the Min-Max Normalization formula. Min-max normalization is one of the most common ways to normalize data. For each feature, the minimum value of the feature is changed to 0, the maximum value is changed to 1, and every other value is converted to a decimal between 0 and 1. The application of this method can be seen in the model equation 7 as follows [13]:

$$x* = \frac{x-min(x)}{max(x)-\min(x)} \tag{9}$$

Based on the description above, a flow diagram is obtained to perform a classification. The data will be obtained using the K-Nearest Neighbor algorithm method in the classification process to determine the k parameters of the predicted data. After finishing determining the parameters at k and before calculating the distance, what must be done is to find the normalization value of each training data and test data. Preprocessing is an early stage that must be done in data mining. Preprocessing goals in data mining is to prepare data raw before any other process. Preprocess data is done by eliminating data that do not match or change the data into a form that easier to process by the system [14]. The application of the search for normalization values can be done by implementing an equation model (6):

$$x* = \frac{x-min(x)}{max(x)-\min(x)} \tag{10}$$

The calculation above is to find the normalization of temperature, water pH, and ammonia in the first training data. The formula is applied according to the total number of training data, which is 14 data. The thing that must be met to find a distance is to find the entire normalization value first. A series of test data obtained through the forecasting stage using the Linear Regression method for the next 14 days. When the search process has been completed, the search results will find out most of the objects obtained by relying on training data and testing data [15]. This series of processes will produce classification results to determine the water quality suitable for catfish farming based on weather predictions and future water quality.

D. Testing

In this process, a system test is carried out by implementing a blackbox method. The application of this method is carried out to ensure that all requirements have been implemented as well as identify deficiencies in the system and provide output that meets the expectations of the examiner by conducting a system accuracy test.

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

III. RESULT AND DISCUSSION

After analyzing and simulating the design contained in the previous chapter, the results and discussion of the system built were obtained. This is based on the method used in this study, namely Linear Regression and the K-Nearest Neighbor algorithm. The process of sending data contained in Internet of Things devices must comply with a data communication architecture protocol that has been created, namely on an IoT gateway as a message receive from a sensor and continue sending the data to the server. In the next stage, implement the forecasting interface which is a menu display to see the prediction results obtained by displaying the results of the graph of the prediction processing results for each data such as temperature, pH, and ammonia. The graph displays original data and predictive data. There are 2 plots in displaying graphs consisting of forecasting graphs from 1 week data that have been taken from sensor devices. The first plot displays the original data graph and the second plot displays the predicted data graph on the data for the next 1 week. The forecasting mechanism also displays the results of the Mean Squared Error (MSE) value obtained based on each forecasting process from each data. The display of the forecasting tempeature interface can be seen in Figure 2 as follows:

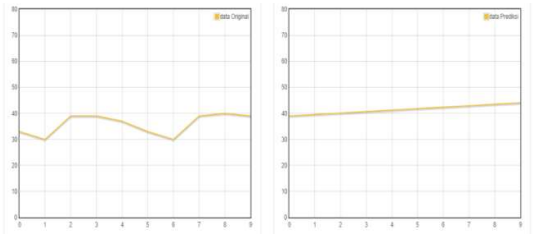

Fig. 2. Linear Regression Temperature Forecasting

Figure 2 shows a forecasting process carried out by a linear regression algorithm, the forecasting was carried out for monitoring weather conditions around the pond and in order to prevent and carry out maintenance on the catfish pond, the forecasting was carried out for 10 days. At the left image show temperature was recorded in 10 days and the right image show trend of forecasting for next 10 days. These means that the average temperature in the next 10 days is between 40 to 50. In the next stage, the forecasting process is carried out using linear regression on the pH of the water, as shown in Figure 3as follow:

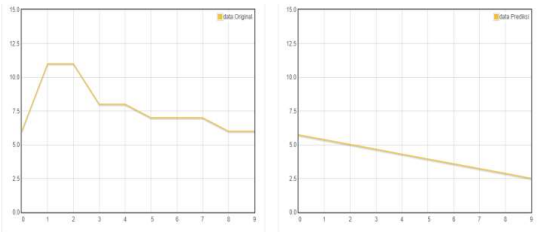

Fig. 3. Linear Regression pH Forecasting

Figure 3 shows a forecasting process carried out by a linear regression algorithm, the forecasting was carried out for monitoring the pH conditions of the water contained in the pond and performing maintenance on the catfish pond, the forecasting was carried out for 10 days. In Figure 3, there is a fairly high increase in data in the original data, namely on day 2 and day 3 on the graph, and experienced a very drastic decrease in the forecasting mechanism. At the left image show pH was recorded in 10 days and the right image show trend of forecasting for next 10 days. These means that average trend of pH decreasing in the next 10 days. In the next stage, the forecasting process is carried out using linear regression on the ammonia gas, as shown in Figure 4 as follow:

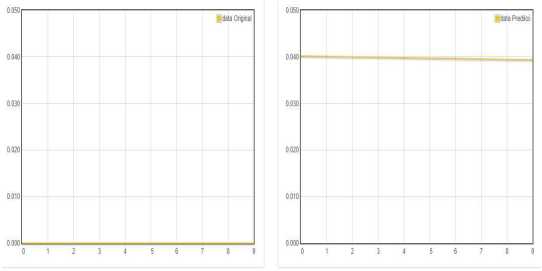

Fig. 4. Linear Regression Ammonia Forecasting

Figure 4 shows a forecasting process carried out by a linear regression algorithm, the forecasting was carried out for monitoring ammonia gas around the water in the pond and performing maintenance on the catfish pond, the forecasting was carried out for 10 days. Ammonia gas usually arises from the remnants of food contained in the pond and also fish waste that is trapped in the pond. The last stage is a classification process that functions as a recommendation for suitable care for catfish using a K-Nearest Neighbor algorithm method. This classification process relies on factual data that has been obtained from the research location. In the next stage, making a classification interface is a display of the results of the classification of recommendations for the type of treatment that is suitable for the fish pond according to the conditions of temperature and humidity, pH, and also levels of ammonia gas. The results of the classification will be sorted so that the earliest treatment is the one that is highly recommended to be carried out in the condition of the catfish pond. The display in the classification process will display each Euclidean Distance obtained. The appearance of the classification interface by implementing the K-Nearest Neighbor Algorithm method can be seen in Table 3 as follow:

TABLE 3. CLASSIFICATION USING THE K-NEAREST NEIGHBOR

No	T	pH	Ammonia	Treatment Type	K-NN results
1	35	11.5	0.012	Waste 50% Water and Replace	1.6143297322556016

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

				50% With New Water	
2	40	7.15	0.022	Waste 50% Water and Replace 50% With New Water	3.1715 916852 23332
3	33	7.15	0.2	Waste 50% Water and Replace 50% With New Water	3.1715 916852 23332
4	35	7.15	0.23	No Treatment	4.0090 805867 56908
5	33	11.7	0.02	No Treatment	4.1959 576361 59408
6	35	6.15	0.012	No Treatment	4.5977 596520 23774
7	41	8.65	1.02	Sprinkle PK medicine and combine it with salt water then dissolve it, and pour it into the fish pond in question	5.8485 947444 18905

In table 3, the system of the K-Nearest Neighbor Algorithm provides a recommended action according to the previous data collection stage, classification is carried out for the next 10 days based on forecasting data performed by the previous Linear Regression Algorithm.

IV. CONCLUSIONS

Linear Regression and K-Nearest Neighbor based on the Internet of Things can be implemented on a website system for water quality analysis and automatic feeding of catfish. The system can retrieve online data that is accessed using the API and displays forecasting graphically from original data and predictive data and presents recommendations for the type of treatment that is suitable for catfish ponds found in Mundusari Village, Subang Regency, West Java. Internet of Things devices can generate data from individual sensors such as DHT11 for temperature, pH sensors, and MQ-2 for ammonia. The data collection process was carried out for 14 days, from 21 July 2021 to 03 August 2021.

REFERENCES

[1] Al-Dosary, N. M. N., Al-Hamed, S. A., & Aboukarima, A. M. (2019). K-NEAREST NEIGHBORS METHOD FOR PREDICTION OF FUEL CONSUMPTION IN TRACTOR-CHISEL PLOW SYSTEMS. *Engenharia Agrícola*, *39*(6), 729–736.

[2] Priyanka, G., Prakash, N., Kishore, A., & S, N. K. (2020). *Smart IoT Based Fish Pond Monitoring System to Enhance Fish Cultivation*. *29*(5), 6408–6417.

[3] Guo, Y., Han, S., Li, Y., Zhang, C., & Bai, Y. (2018). K-Nearest Neighbor combined with guided filter for hyperspectral image classification. *Procedia Computer Science*, *129*, 159–165. https://doi.org/10.1016/j.procs.2018.03.066

[4] Hu, G., Yang, Z., Zhu, M., Huang, L., & Xiong, N. (2018). *Automatic classification of insulator by combining k-nearest neighbor algorithm with multi-type feature for the Internet of Things*. 1–10.

[5] Fitriani, R., & Vitriani, Y. (2018). The Comparison of Linear Regression Method and K-Nearest Neighbors in Scholarship Recipient. *2018 19th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD)*, 194–199.

[6] Arafat, A. I., Akter, T., Ahammed, F., Ali, Y., & Nahid, A. (2020). A dataset for internet of things based fish farm monitoring and notification system. *Data in Brief*, *33*, 106457. https://doi.org/10.1016/j.dib.2020.106457

[7] Wahla, A. H., Chen, L., Wang, Y., Chen, R., & Wu, F. (2019). Automatic wireless signal classification in multimedia Internet of Things: An adaptive boosting enabled approach. *IEEE Access*, *7*, 160334–160344.

[8] Wang, W., & Lu, Y. (2018). Analysis of the Mean Absolute Error (MAE) and the Root Mean Square Error (RMSE) in Assessing Rounding Model. *IOP Conference Series: Materials Science and Engineering*, *324*(1). https://doi.org/10.1088/1757-899X/324/1/012049

[9] Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)?–Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, *7*(3), 1247–1250.

[10] Kang, S. (2021). K-nearest neighbor learning with graph neural networks. *Mathematics*, *9*(8). https://doi.org/10.3390/math9080830

[11] Guo, G., Wang, H., Bell, D., Bi, Y., & Greer, K. (2003). KNN model-based approach in classification. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *2888*(August), 986–996. https://doi.org/10.1007/978-3-540-39964-3_62

[12] Kurniadi, D., Abdurachman, E., Warnars, H. L. H. S., & Suparta, W. (2018). The prediction of scholarship recipients in higher education using k-Nearest neighbor algorithm. *IOP Conference Series: Materials Science and Engineering*, *434*(1). https://doi.org/10.1088/1757-899X/434/1/012039

[13] Nababan, A. A., Sitompul, O. S., & Tulus. (2018). Attribute Weighting Based K-Nearest Neighbor Using Gain Ratio. *Journal of Physics: Conference Series*, *1007*(1). https://doi.org/10.1088/1742-6596/1007/1/012007

[14] Prasath, V. B. S., Alfeilat, H. A. A., Hassanat, A. B. A., Lasassmeh, O., Tarawneh, A. S., Alhasanat, M. B., & Salman, H. S. E. (2017). *Distance and Similarity Measures Effect on the Performance of K-Nearest Neighbor Classifier -- A Review*. 1–39. https://doi.org/10.1089/big.2018.0175

[15] Taunk, K., De, S., Verma, S., & Swetapadma, A. (2019). A brief review of nearest neighbor algorithm for learning and classification. *2019 International Conference on Intelligent Computing and Control Systems (ICCS)*, 1255–1260.