UKT (Single Tuition) Classification Prediction uses MKNN (K-Nearest Neighbor Modification) Algorithm

1st Dziki Adli
Department of Information Technology
Politeknik Caltex Riau
Pekanbaru, Indonesia
dziki20s2tk@mahasiswa.pcr.ac.id

2nd Dadang Syarif Sihabudin Sahid
Department of Information Technology
Politeknik Caltex Riau
Pekanbaru, Indonesia
dadang@pcr.ac.id

***Abstract*—Islamic University of Sultan Syarif Kasim (*UIN SUSKA*) Riau Province has used an information system, namely the *Sistem Registrasi* (SIREG) to facilitate the registration of prospective students and also SIREG also provides a decision on determining the *UKT* of students who have been declared graduated. But there has never been an evaluation of SIREG's accuracy in determining the *UKT*. From these problems, a model is needed to be implemented to facilitate *SIREG* officers in conducting classifications to establish *UKT* new students. Using the MKNN method and supported by the K-Fold Cross Validation validation method, the classification results get an accuracy value of 71%.**

Keywords—Classification, KNN, K-Fold Cross Validation

I. INTRODUCTION

One of the public universities in Indonesia and has used the UKT system is UIN Suska Riau, the UKT payment system has been implemented in UIN Suska Riau since the 2014/2015 school year which has been implemented in the SIREG application. The process for determining the UKT group requires rigor and time, as student data will be compared to UKT criteria one by one. The decision-making system in determining existing UKT groups uses 5 criteria in assessing the ability of parents of students.

MKNN is an algorithm developed from the KNN algorithm, the MKNN algorithm adds a new process to perform classification i.e. calculation of validity values to consider validity between training data and weighted voting calculations to calculate the weight of each nearby. The addition of 2 new processes in the MKNN is expected to correct any errors in the KNN process. Based on the research will be carried out algorithm classification prediction using K-NN Modification on UKT UIN Suska data because it has been done in previous research for UKT predictions at Riau University using KNN. It is expected that by using the MKNN algorithm, better accuracy results are obtained.

II. RELATED WORK

Sukamto, 2020 [1], the research conducted resulted in an accuracy value from classification using KNN algorithm obtained by 84.21% to predict the UKT that will be paid by prospective students, especially the S1 study program of FMIPA Information System, Riau University. The criteria used are gross income, tuition insurers, the number of dependents listed in the electricity card, the status of residence, the state of the walls of the residence, the state of the roof of the residence, the total area of land ownership and the cost of electricity usage a month. The UKT groups are UKT1, UKT2, UKT3, UKT4, UKT5 and UKT6. The data used is students of S1 FMIPA Universitas Riau in the class of 2016, 2017, and 2018. The ratio for training data and test data is 90%: 10%.

Okfalisa, 2018 [2], research was conducted using the Dataset of the Hope Family Program Implementation Unit using 7,395 records of data has been compared KNN with MKNN using 3 types of K-Fold Cross Validation method and k = 10. Confusion Matrix calculations in Cross 2 get the highest average accuracy value yield of 93.945%. While in the accuracy comparison between KNN and MKKN it was found that KNN has the highest accuracy value of 94.95% and the average accuracy is 93.94%. Using MKKN obtained the highest accuracy of 99.51% and an average accuracy of 99.20%. So it can be said that using MKNN is better 5-7% than KNN.

Lestari, 2017 [3], this research was conducted for the classification of Acceptance of Toyota Astra scholarships by applying knn and naïve bayes algorithms. Then the values of the two algorithms will then be compared. KNN has a higher accuracy value than the accuracy value of the naïve bayes algorithm.

Parvin, 2008 [4], this study compared two algorithms that are one clump namely KNN and MKNN. The development that occurs is performance improvement that is a kind of preprocessing on data training. Add a new value called "validity" to train the sample. Validity takes into account the

stability and reliability value of each data train against its neighbors' data.

Pisarenko, 2021 [5], the proposed method mknn is applied to analyze seismic intensity in two seisgomenic regions. The graph of increased seismic activity can be identified by the MKNN and some of the quantitative statistical characteristics of this graph will be determined and discussed.

Masoodi, 2018 [6], mknn algorithm applied to tracking systems. Research was also conducted on determining the value of K. Simulation of the application of algorithms in case studies has errors in Euclideans below 1 meter for an area of 100 square meters or about 1%. With the same problem, other models are also applied but have a complicated system. Another advantage of MKNN is simplicity.

Of the overall related work that has been described in detail, the author again made sure to apply MKNN to the UKT classification case study. Given the validity of the new value enhanced in MKNN [4] the study applied validation with K-Fold Cross Validation.

III. THEORETICAL FOUNDATION

A. UKT

The Ministry of Education and Culture of the Republic of Indonesia stipulates the Regulation of the Minister on Single Tuition (UKT) which began to be implemented in the academic year 2013/2014. Single tuition is a single tuition fee that is borne by each Prospective New Student based on his or her economic ability. Each State University has a different UKT rate, this is influenced by the regional tentacle and its study program (Permendikbud No.55, 2013).

B. Data Normalization

Data normalization is required in data preprocessing tools used in data mining systems and calculations to narrow the range of data training and distribution of data evenly, in normalization there are several normalization techniques such as min-max normalization, z-score normalization, decimal scalling and sigmoidal normalization. While in this study the normalization of data used is min-max normalization. Min-max normalization is a transformation of the value of an existing data with the smallest range value (min) of 0 and the largest value (max) of 1, in the equation of Min-Max Normalization shown in Equation (1). [7]

$$V' = \frac{V - minA}{maxA - minA} \cdot (new_{maxA} - new_{minA}) + new_{minA} \quad (1)$$

V' = The value of the new data results from normalization
V = Values of data before normalizing newmax
A = Newmin's latest maximum value limit
A = Latest minimum value limit
Max A = maximum value in the column
Min A = minimum value in the column

C. K-Nearest Neighbor Algorithm

The K-NN method is one of the simplest and most intuitive algorithmic techniques in the field of statistical discrimination. The process of calculating euclidean distance in this algorithm is to first create a dataset from training data that the class has known first, the next step is that the dataset of the test data will be classified based on the closest distance of each training data and depending on the k value used. Euclidean equation to perform calculations of the distance between the test data point and the training data point where i is a representation of the attribute value and n dimensions of the attribute shown in the equation (2) [8].

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

d(x,y) = Euclidean distance between training data point x and test data y
x_i = Sample training data
y_i = Test data
n = Attribute dimensions

D. Modified K-Nearest Neighbor Algorithm

This MKNN method is the result of a modification with the goal of optimizing the KNN method, in the MKNN method is divided into two processes, namely, first is the validation of training data and then second is the KNN weighting process or weight voting. Classification is performed on test data based on the highest-grade weight in class K training data that has been validated at the closest distance. Unlike KNN, the KNN method does not go through the process of validation of training data. This method of training data validation can maximize training data with high validity and has a close proximity to test data.[4].

E. Validity of data

The process of data validity must be passed in the MKNN process, the validity of each point is calculated according to its nearest neighbor based on k. the validity of this training data is to function to know the number of points with the same table for all the data in the training data. In this process all function values in s depend on the nearest neighbor is equal value or not on training data. 12 Equation of data validity shown in equation (3)[4].

$$validitas(x) = \frac{1}{k}\sum_{i=1}^{k} S(label(x))(Ni(x))) \quad (3)$$

K = number of closest points
Label(x) = class label x
Ni(x) = nearest point class label x
S = 1 when class is the same or worth 0 when the class is not the same

F. Weight voting

In MKNN each data is calculated in weight. Weight voting is useful on training data with high validity and the closest distance to the test data. The first work of weight voting is to calculate the weight of each neighbor, both the validity of the sample of training data multiplied by the Euclidean weight. The multiplication process can reduce weaknesses in each data that has distance problems with weight in the outlier. Weight voting equation shown in equation (4)[4]

$$W(i) = validitas(i) * \frac{1}{d_e + a} \quad (4)$$

W(i) = calculation of weight voting to i
Validitas(i) = validity value to i
d_e = Euclidean distance training data and test data
α = alpha value

G. K-Fold Cross Validation

Cross-validation or rotation estimation is a model validation technique to assess how statistical analysis results will generalize independent data sets. This technique is primarily used to predict models and estimate how accurate a predictive model is when executed in practice. One technique of cross validation is k-fold cross validation, which breaks down data into k parts of the same size. The use of k-fold cross validation to eliminate biases in data. Training and testing are done k times. In the first experiment, the S1 subset was treated as testing data and the other subset was treated as training data, in the second trial the S1 subset, S3,... Sk becomes training data and S2 becomes testing data, and as accurate as it is[9].

H. Confusion Matrix

Confusion matrix is a method that is usually used to perform accuracy calculations on the concept of data mining. Measurement of the performance of a classification system is important. The performance of a classification system describes how well the system classifies data.

Basically, the confusion matrix contains information that compares the results of classifications carried out by the system with the results of classifications that should be.

IV. RESEARCH METHODS

The stages for completing the research are shown in Figure 1.

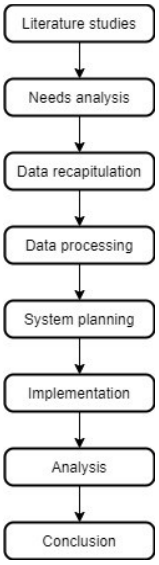

Fig. 1. Methodology flow diagram

In this research begins by collecting then studying the study of the literature used, the next step is to analyze the needs needed, collect then process the data used, design the system to be created, implement the design that has been made, test the system that has been made, analyze the results obtained based on the results of the test, the last step is to make conclusions from the research that has been done.

A. Data sources and research variable

Data is sourced from the database system SIREG UIN SUSKA Riau in 2020 with the amount of data 6000 student data. Variables used in existing weighting are:

- Amount of parents' gross income
- Number of unmarried children
- Number of vehicles
- Number of houses
- Amount of land

While additional variables that will be predicted using MKNN plus 3 variables are:

- Monthly expenses
- Electricity costs per month
- Size of electricity meter

B. Process Flow Diagram

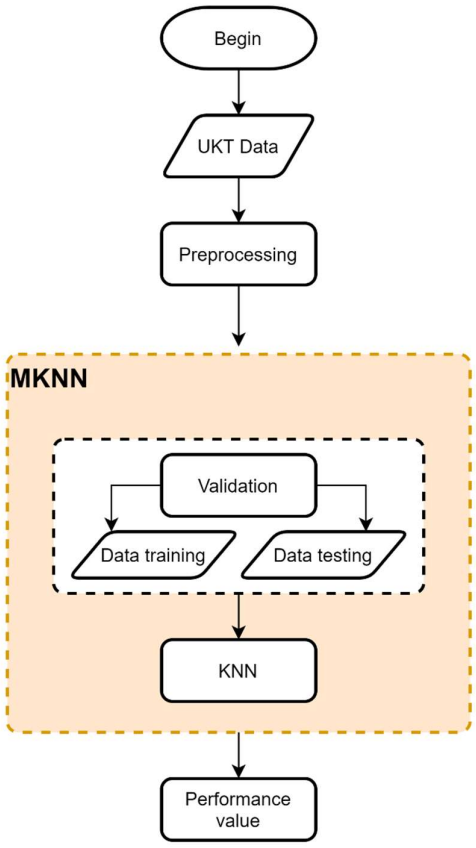

Fig. 2. System process flow diagram

Figure 2. Explain about the stages of the system that is done, ranging from preprocessing, validation, process with the main algorithm, namely MKNN and accuracy assessment.

V. RESULTS AND DISCUSSIONS

1) Preprocessing process

At the preprocessing stage, data transformation is carried out, namely:

- values in parent earnings column.
- value in the monthly electricity cost column.
- values in the field of land
- value on wattage column of electricity meter

2) Validation with Cross Validation, K=4 and with dataset sharing 80:20

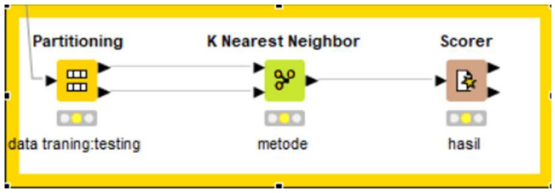

Fig. 3. Validation process using cross validation

3) Validation with K-Fold Cross Validation, K=4 and with dataset sharing 80:20

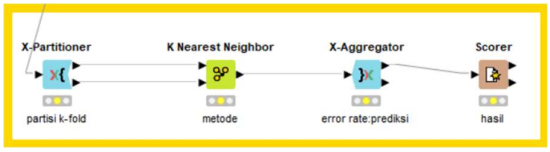

Fig. 4. Validation process using K-Fold cross validation

Figures 3 and 4 are stages of MKNN because they have been modified with the addition of validation methods in the process of separating data into training and testing.

4) Confusion matrix results

a) With Cross Validation

Confusion Matrix - 0:34 - Scorer (hasil)

File Hilite

ukt_rekom...	3	2	4	5	1
3	26	10	12	0	0
2	10	33	3	0	2
4	8	1	86	3	0
5	0	0	10	33	0
1	2	13	1	0	0
6	0	0	1	2	0
7	0	0	0	0	0

Correct classified: 190 Wrong classified: 81

Accuracy: 70.111 % Error: 29.889 %

Cohen's kappa (κ) 0.604

Fig. 5. Accuracy uses confusion matrix for cross validation

b) With K-Fold Cross Validation

Confusion Matrix - 0:39 - Scorer (hasil)

File Hilite

ukt_rekom...	3	2	4	5	1
3	277	84	118	3	7
2	67	443	32	1	10
4	87	24	806	41	5
5	7	0	119	256	0
1	14	76	9	0	44
6	3	0	3	29	0
7	0	0	0	0	0

Correct classified: 1,938 Wrong classified: 767

Accuracy: 71.645 % Error: 28.355 %

Cohen's kappa (κ) 0.626

Fig. 6. Accuracy uses confusion matrix for cross validation

Figures 5 and 6, describe the results of the performance of systems designed using different validation processes. K-Fold and Cross Validation have the advantages of each performance, but are varied from K more K-Fold Cross Validation.[10]. MKNN is an optimization algorithm from conventional KNN algorithms by adding validation methods successfully get maximum performance values. MKNN with cross validation managed to get a value of 70.11% and MKNN with K-Fold Cross Validation managed to increase the value to 71.64%..

VI. CONCLUSION

The system that has been built to determine the cost of each student's UKT, SIREG, still needs to be done quality analysis gradually and in-depth. Because based on the analysis using the 2020 tuition fee dataset, the analysis model applied can measure quality based on the classification generated by the SIREG application. By applying the MKNN algorithm, the accuracy value of the decisions generated by SIREG earns 71.64% points. This value is relatively good considering that the dataset used only the dataset in the last 1 year.

REFERENCES

[1] S. Sukamto, Y. Adriyani, and R. Aulia, "Prediksi Kelompok UKT Mahasiswa Menggunakan Algoritma K-Nearest Neighbor," *JUITA J. Inform.*, vol. 8, no. 1, p. 121, 2020, doi: 10.30595/juita.v8i1.6267.

[2] Okfalisa, I. Gazalba, Mustakim, and N. G. I. Reza, "Comparative analysis of k-nearest neighbor and modified k-nearest neighbor algorithm for data classification," *Proc. - 2017 2nd Int. Conf. Inf. Technol. Inf. Syst. Electr. Eng. ICITISEE 2017*, vol. 2018-Janua, pp. 294–298, 2018, doi: 10.1109/ICITISEE.2017.8285514.

[3] D. R. D. Lestari, "Perbandingan Klasifikasi Beasiswa Toyota Astra Menggunakan K-Nearest Neighbor Classifier dan Nive Bayes Sebagai Penentu Metode Klasifikasi Pada Sistem Pendukung Keputusan Penerimaan Beasiswa Toyota Astra (Studi : Institut Teknologi Sepuluh Nopember)," *Inst. Teknol. Sepuluh Nop.*, pp. 9–10, 2017, [Online]. Available: http://repository.its.ac.id/id/eprint/42490.

[4] H. Parvin, H. Alizadeh, and B. Minaei-bidgoli, "MKNN : Modified K-Nearest Neighbor," *Proc. World Congr. Eng. Comput. Sci. WCECS*, pp. 22–25, 2008.

[5] V. F. Pisarenko and D. V. Pisarenko, "A Modified k-Nearest-Neighbors Method and Its Application to Estimation of Seismic Intensity," *Pure Appl. Geophys.*, Apr. 2021, doi: 10.1007/s00024-021-02717-y.

[6] M. Masoodi, E. A. Sekehravani, and M. Maesoumi, "Rssi-Based Modified K-Nearest Neighbors Algorithm for Indoor Target Tracking," *Far East J. Electron. Commun.*, vol. 18, no. 2, pp. 345–356, 2018, doi: 10.17654/ec018020345.

[7] S. Jain and K. Asawa, "Modeling of emotion elicitation conditions for a cognitive-emotive architecture," *Cogn. Syst. Res.*, vol. 55, no. June, pp. 60–76, 2019, doi: 10.1016/j.cogsys.2018.12.012.

[8] K. Schliep, K. Hechenbichler, and A. Lizee, "Weighted k-Nearest Neighbors," *2016*, vol. 399, p. 15, 2016, [Online]. Available: https://cran.r-project.org/web/packages/kknn/kknn.pdf.

[9] M. Bramer, *Principles of Data Mining*, no. January 2007. 2007.

[10] U. B. Hatta, "JURNAL IPTEKS TERAPAN Research of Applied Science and Education V15.i1 (34-47)," vol. 1, no. March, pp. 34–47, 2021.