International Journal of Computer and Information System (IJCIS)
Peer Reviewed – International Journal
Vol : Vol. 01, Issue 01, May 2020
e-ISSN : 2745-9659
https://ijcis.net/index.php/ijcis/index

Estimation System For Late Payment Of School Tuition Fees

1st Muqorobin, 2nd Kusrini, 3rd Siti Rokhmah, 4th Isnawati Muslihah
[1,3,4]Institut Teknologi Bisnis AAS Indonesia Surakarta
[2]Universitas Amikom Yogyakarta.
[1,3,4]Jl. Slamet Riyadi No. 361 Windan, Makamhaji, Kartasura, Sukoharjo, Indonesia
[2]Jl. Ring Road Utara Condong Catur, Sleman, Yogyakarta, Indonesia
[1]robbyaullah@gmail.com, [2]kusrini@amikom.ac.id [3]elfathiey@gmail.com, [4]isnawatimuslihah12345@gmail.com

Abstract*—The Surakarta Al-Islam Vocational School is a private educational institution that requires all students to pay school tuition fees. Education is an obligation for all Indonesian citizens. The cost of education is one of the most important input components in implementing education. Because cost is the main requirement in achieving educational goals. SPP School is a routine school fee that is carried out every month. Based on last year's School Admin report, many students were late in paying school tuition fees, around 60%. This is a very big problem because the income of school funds comes from school tuition. The purpose of this research is that the researcher will build a prediction system using the best classification method, which is to compare the accuracy level of the Naïve Bayes method with the K-K-Nearest Neighbor method. Because both methods can make class classifications right or late, in paying school fees. processing using dapodic data for 2017/2018 as many as 236 data. In improving accuracy, the researcher also applies feature selection with Information Gain, which is useful for selecting optimal parameters. System testing is carried out using the Confusion Matrix method. The final results of this study indicate that the Naïve Bayes Method + Information Gain Method produces the highest accuracy, namely 95% compared to the Naïve Bayes method alone, namely 85% and the K-NN method, namely 81%.

Keywords : Comparison, Naive Bayes, K-NN, Prediction, Tuition Fee

I. INTRODUCTION

Attending the basic education is an obligation for every Indonesian citizen. This is stated in a statutory regulation in article 31 paragraph (1) of the 1945 Constitution and Minister of Education and Culture Regulation No. 19 of 2016 about the Smart Indonesia Program[1]. Meanwhile, the financial cost is one of the important components in the process of implementing education[2]. One of the financing sources comefrom the Education Development Donation or simply called tuition fee[3]. SMK Al-Islam Surakarta is one of the private educational institutions of the Al-Islam Surakarta Foundation which focuses on teaching Information Technology and Islamic Sciences. In financing their operational of the school, they mostly chargedit to students, especially in school tuition payments that has to paid monthly.

The problem that often arises in SMK Al-Islam Surakarta is when many students are late in paying school tuition fees. This is a serious problem because school tuition fees are one of the main sources of funds in improving the quality of school education. A tuition fee is used to cover operational costs including the salaries of teachers and employees. Based on data from the financial administration section, that in the 2017/2018 school year there were 60% of students who were late in paying school tuition. To overcome the problem of delay, it is necessary to predict students who have the potential to be late in making payments so that the school can take anticipatory action.

There are several predictive algorithms that can be used including Naïve Bayes and K-Nearest Neighbor (K-NN). Both algorithms are included in the Top 10 algorithms in data mining[4]. Research on the use of the Naive Bayes algorithm has been done to predict the level of smooth credit card payment and the student graduation rate on time[5][6]. K-NN has been used in many classification processes including the classification process of skin conditions, emotion recognition from multichannel EEG signals and Gene expression cancer classification[7]. Algorithm comparison has also been done in several studies to compare Naïve Bayes, K-NN with other algorithms[8].

This study will compare the accuracy of the Naïve Bayes algorithm and the K-NN algorithm in predicting the accuracy of tuition fee payments[9]. After each algorithm is ran by using all of available parameters, the experiment also compared how if the algorithm ran with a specific selected parameter usinga feature selection technique namely Information Gain[10]. Research that has utilized information gain to select features includes the research of classifying analytical sentiment documents process[11]. The best algorithm from this study will be applied in the system of late payment prediction of school tuition[12].

II. RESEARCH METHOD

The data used in this study are sourced from official school Education Based Dataof the 2017/2018 school year and school tuition fee payment transaction reports.The attributes of the data to be used as determinant variables are Parent Income, Family Dependent, Father's Education, Father's Age, Mother's Education and Mother's Age[13].

The total amounts of data record taken were 236. The data records obtained arethen splattedintoTraining Data and TestingData. The training data is used to create theprediction knowledge model with the observed methods Naive Bayes and K-NN, while testing data is used to determine the level of accuracy of the model. From the total amounts of available data, the proportion between training data and testing data will be 75% and 25% respectivelyaccording to the proportion reference in data usage. In other words, from 236 total data, the number of training data and testing data are177 and 59 respectively[14].

The study also combines the Naive Bayes and K-NN methods with Information Gain and compare them alltogether to get the best result. The combinations will be

using the Naïve Bayes algorithm with all variables, using Naïve Bayes algorithm with selected variables from Information Gain, using K-NN algorithm with all variables, and using K-NN algorithm with selected variables from Information Gain.

The models created by each method are then tested with 4-fold cross validation that is the testing in each method is carried out 4 times with a combination of different training data and testing data as shown in Table 1.

Table 1. The Plotting of Training and Testing Data in Each Fold

Fold	Training Data Rec No.	TestingData Rec No.
1	178-236	1-59
2	1-59 and 119-236	60-117
3	1-118 and178-238	119-177
4	1-177	178-236

The prediction performance of testing data in each fold will be calculated using Confusion matrix. The matrix produces the value of accuracy, precision and recall. Each of these values will be averaged over the entire fold. The average results of accuracy, precision, recall and F-Measure of each method were compared to determine the best method in predicting data of overdue payment of school tuition.

2.1 *Naïve Bayes* Method

Naïve Bayes algorithm is one of popular classification methods using probability and statistics. The basic form of Bayes methods can be shown as Equation 1.

$$P(A|B) = \frac{P(B|A)*P(A)}{P(B)} \quad \text{.....(1)}$$

With:
P(A|B) : The Posterior value of AwhenB is occurred
P(B|A) : The Likelihood value of B when A is occurred
P(A) : The Prior value of class A
P(B) : The Evidence value of a class

The probability A as B, is obtained from probability B when A multiplied by probability of A and divided by probability of B. The use of Naive Bayes on a data with more than one feature / attribute, causing Equation 1 to be more complex as shown in Equation 2

$$P(A|B_1 \dots B_n) = \frac{\mathrm{P(A)P}(B_1 \dots B_n|\mathrm{A})}{\mathrm{P}(B_1 \dots B_n)} \quad \text{.....(2)}$$

The value of P $(B_1 \dots B_n)$ is constant for each experiment so that the maximum value of a class is determined by the maximum value between P (A) P $(B_1 \dots B_n|$ A)[15]. The equation function formed becomes a maximum multiplication for the prior value and the likelihood function, the function is shown in Equation 3.

$$f_c(F) = {}^{\arg}{}_{c \in F}^{max} P(A)(\prod_{i=1}^{n} P(B_i|A)) \quad \text{.....(3)}$$

2.2 K-NN (*Nearest Neighbor*) Method

K-Nearest Neighbor (K-NN) algorithm is a well known algorithm in machine learning area, because its easy and simple process. In the K-NN algorithm all available data must have a label to identify the closest data in the comparison process.The working principle of K-NN is to find the closest distance between the evaluated data with the nearest K-Neighbor in the training data[16]. The steps of the K-NN algorithm are:

a. Determine parameter K
b. Calculate the distance between testing data and training data. If the data is numeric, then we use Euclidean distance as shown in Equation 4.

$$D(x_i, y_i) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad \text{.....(4)}$$

With:
Xi = training data
Yi = testing data
D (xi, yi) = distance
i = variable data
n = dimension data

c. Sort all the distances descending.
d. Select the closest distance to parameter k
e. Select the highest amount of class then classify

2.3 Variables Selection with Information Gain

Information Gain is the simplest feature selection method by making attributes ranking and widely used in text categorization applications, microarray data analysis and image data analysis[17]. Information Gain can help reduce noise caused by irrelevant features. Information Gain detects features that have the most information based on a particular class[18]. Determining the best attribute is startedby calculating the entropy value. Entropy is a measure of class uncertainty using the probability of particular events or attributes[19]. The formula for calculating entropy is shown in Equation 5. After the entropy value obtained, the Information Gain calculation can be done using Equation 6.

$$Entropy(S) = \sum_{i=1}^{c} -P_i log_2(P_i) \quad \text{.....(5)}$$

With c is the number of values in the classification class and Pi is the number of samples for class *i*.

$$Gain(S, A) = Entropy(S) - \sum values(A) \frac{|S_v|}{|S|} Entropy(S_v) \quad \text{.....(6)}$$

2.4 Confusion Matrix Testing

Accuracy calculations in data mining can be done by entering a set of testing data into the data mining model and comparing the classification values produced by the model as a prediction to the actual value in the test data. Simple classification for a prediction usually consists of two classes, which indicate that the main observation target or event is *occurred*(Positive) or *not-occurred* (Negative)[20]. Confusion matrix is a method used to calculate accuracy as above. Accuracy measurement is done by confusion matrix testing as shown in Table 2.

Table 2. Confusion Matrix Model

Real Data	Predicted Classification	
	Positive	Negative
Positive	TP	FN
Negative	FP	TN

with:
TP is True Positive, that is, real data is positive and correctly classified (positive) by the system.
TN is True Negative, that is, the real data is negative and correctly classified (negative) by the system.
FN is False Negative, that is, the real data is negative but is wrongly classified (positive) by the system.
FP is False Positive, that is, real data is positive but is classified as wrong (negative) by the system.

From the training data set that is applied to the model, the confusion matrix produces several values, namely Accuracy, Precision, Recall, and F-Measure. Accuracy is the percentage of cases that have predictions and real values are both positive (TP) orboth negative (TN) compared to the total number of cases [15]. Precision or confidence is the ratio between cases that have predictions and real values are both positive (TP) compared to the overall positive predicted cases (TP + FP). Recall or sensitivity is the ratio of cases that have predictions and real values that are equally positive (TP) compared to cases having positive real data (TP + FN) [20]. The Confusion Matrix from table 2 is used in the calculation of accuracy as shown in equation 7, the precision in equation 8, Recall in equation 9 and F-Measure in equation 10.

$$Accuracy = \frac{\sum_{i=1}^{l} \frac{TP_i + TN_i}{TP_i + TN_i + FP_i + FN_i}}{l} * 100\%$$

$$Precision = \frac{\sum_{i=1}^{l} TP_i}{\sum_{i=1}^{l} FP_i + TP_i} * 100\%$$

$$Recall = \frac{\sum_{i=1}^{l} TP_i}{\sum_{i=1}^{l} TP_i + FN_i} * 100\%$$

$$F\text{–}Measure = \frac{2 * Precision * Recall}{Precision + Recall} * 100\% \quad \text{.....(7)}$$

III. RESULT AND ANALYSIS

The research process starts from retrieving data from the school database and transforming the data into a dataset as in Table 3.

Table 3. Source Dataset

Rec	Parents Income	Family Dependent	Father Education	Father Age	Mother Education	Mother Age	RSClass
1	2 - 4 M	Moderate	Elementary	Early elderly	Junior High School	Early adult	Ontime
2	< 1 M	Many	Elementary	Early elderly	Junior High School	Early elderly	Ov
3	1 - 2 M	Few	Bachelor	Early elderly	Diploma 3	Early elderly	On
4	< 1 M	Moderate	Junior High School	Late elderly	Elementary	Late elderly	Ov
..	..	…	..	…	..	…	
236	< 1 M	Few	Elementary	Early elderly	High School	Late adult	Ov

3.1 Information Gain Variable Selection

Information gain is used to select the optimal attributes in making predictions. Calculations are based on the dataset in Table 3. The steps in calculating the information Gain method are as follows :

1. Determining TotalEntropy

To calculate total entropy as in Equation 5.From Tabel 3we get the number of total data 236 with two classes that is "Ontime" with104 data and "Overdue" with 132 data. The total entropy computed as:

$$\text{Total Entropy} = \left(-\frac{104}{236} * \log_2\left(\frac{104}{236}\right)\right) + \left(-\frac{132}{236} * \log_2\left(\frac{132}{236}\right)\right) = 0{,}99$$

2. Calculating the Entropy of the Attribute Values
Calculating the values' entropy needs the total number of records that the value occured, the number of records that have class "Ontime" in the occurence of the value, and the number of records that have class "Overdue" in the occurence of the value.The Entropy valuesprocessedwith Equation 5 and the results are shown in Table 4.

Table 4. Entropy of Attribute Values

Attribute	Value	Ontime	Overdue	Total	Entropy
Parents Income	< 1 M	29	41	70	0,98
	1 - 2 M	26	30	56	1,00
	2 - 4 M	23	38	61	0,96
	> 4 M	26	23	49	1,00
Dependent	Many	11	21	32	0,93
	Fair	50	51	101	1,00
	Few	43	60	(7)103	0,98
	Junior High			(8)	
	School	22	22	44	1,00
…	…	…	…	(9)…	…
Mother Age	Late Adult	46	60	106	0,99
	Early Adult	5	1	6	0,65
	Late Elderly	7	12	19	0,95
	Early Elderly	45	58	103	0,99
	Elderly	1	1	2	1,00

3. Calculating Gain Value
Using Equation 6, the gain value of the attributes in dataset is calculated. The sorted results are put alltogether in Table 5.

Table 5. Attribute Gain Ranking

No	Attribute's Parameter	Gain Value
1	Father Education	0,066
2	Mother Education	0,056
3	Father Age	0,016
4	Mother Age	0,014
5	Parents Income	0,009
6	Family Dependent	0,008

We take 4 attributes that have the highest Gain value, that is Father's Education, Mother's Education, Father's Age and Mother's Age, and expect this attribute selection can improve the models.

3.2 Naive Bayes Calculation

The 4-fold cross validation is carried out to dataset in Table2. Below is the calculation for Fold 4 using all determinant variable:

1. Determine the value for each class of 177 training data
   K1 (Class "Ontime") = 73
   K2 (Class "Overdue") = 104
2. Determine the Probability of Each Attribute Value from 177 training data
   "Ontime" probability is calculated from the number of data in the "Ontime" class on an attribute value divided by the number of class data "Ontime" (K1).The "Overdue" probability is calculated from the number of " Overdue " class on an attribute value divided by thenumber ofclass data " Overdue " (K2). The complete calculation results are shown in Table 6.

Table 6. Calculation of Attribute Value Probabilities

Attribute	Vaue	Count		Probability	
		Ontime	Overdue	Ontime	Overdue
Parents Income	< 1 M	21	31	0,29	0,30
	1 - 2 M	18	23	0,25	0,22
	2 - 4 M	18	32	0,25	0,31
	> 4 M	16	18	0,22	0,17
Dependent	Many	8	18	0,11	0,17
	Fair	38	41	0,52	0,39
	Few	27	45	0,37	0,43
Father Education	Diploma 3	4	4	0,05	0,04
	Bachelor	7	14	0,10	0,13
	Elementary	13	31	0,18	0,30
	High School	33	36	0,45	0,35
	Junior High School	16	19	0,22	0,18
FatherAge	Late Adult	19	32	0,26	0,31
	Early Adult	4	1	0,05	0,01
	Late Elderly	10	19	0,14	0,18
	Early Elderly	37	48	0,51	0,46
	Elderly	4	5	0,05	0,05
Mother Education	Diploma 3	7	5	0,10	0,05
	Bachelor	19	46	0,26	0,44
	Elementary	25	30	0,34	0,29
	High School	18	18	0,25	0,17
	Junior High School	32	49	0,44	0,47
Mother Age	Late Adult	4	1	0,05	0,01
	Early Adult	6	8	0,08	0,08
	Late Elderly	30	45	0,41	0,43

3. Examining the Testing Data
   Testing Data is taken from the dataset in Table 3 which is not used as training data, that is 59 data as shown this.

Table 7. Testing Data

Record	Father Education	Father Age	Mother Education	Mother Age	RS
178	Elementary	Late Elderly	Elementary	Late Adult	Ov
179	High School	Early Elderly	High School	Early Elderly	On
180	Elementary	Early Elderly	High School	Early Elderly	Ov
181	High School	Early Elderly	Junior High School	Early Elderly	On
..	..	…	..	…	..
236	Elementary	Early Elderly	High School	Late Adult	Ov

Each data in Table 7 is predicted to find out whether it is overdue or on-time payment byusing Equation 3. The results are shown in Table 8.

Table 8. Data Testing Calculation Results

Record	Ori. Result	Predict. Result
178	Overdue	Overdue
179	Ontime	Ontime
180	Overdue	Ontime
181	Ontime	Ontime
..	..	…..
236	Overdue	Overdue

4. Test Result
   Based on the results of testing data of 59 records in Table 8, testing can be done using the Confusion Matrix method that can be obtained as shown in Table 9.

Table 9. Testing Results 4-Fold Cross Validation with the Naïve Bayes Algorithm

K fold	Accuracy	Presisi	Recall	F-Measure
K1	56%	40%	48%	43%
K2	56%	50%	50%	50%
K3	56%	59%	43%	50%
K4	86%	97%	81%	88%
Average	64%	61%	56%	58%

3.3 Implementation of the Naive Bayes Method with Information Gain

The calculation is done in the same way as the calculation using the Bayes method, but only uses the 4 best parameters.The results of system testing on the Naïve Bayes method using selected variables from the Information Gain calculation are shown in Table 10.

Table 10. Results of Testing the Naïve Bayes + Information Gain Method

K fold	Accuracy	Precision	Recall	F-Measure
K1	56%	40%	48%	43%
K2	61%	58%	56%	57%
K3	59%	59%	46%	52%
K4	92%	100%	86%	93%
Average	67%	64%	59%	61%

3.4 K-NN Calculation

Predictions using the K-NN method can be done by calculating the proximity value between training data and data testing with the Euclidean Distance equation [21]. To calculate the proximity distance between training data and testing data, the data set in Table 3 needs to be converted into numbers. The first step in calculating the proximity of training data and data testing requires a K value to produce the best accuracy value. The researchers try to predict the late payment with a K value of 1,3,5,7,9,11,13,15. After determining the value of K, the data in Table 2 that has been converted will be broken down into training data and testing data. For the 4th Fold testing, the training data is shown in Table 11 and the testing data is in Table 12.

Table 11. K-NN Training Data

Record	Parents Income	Family Dependent	Father Education	Father Age	Mother Education	Mother Age	Result
1	30	15	5	15	15	5	Ontime
2	10	20	5	15	15	15	Overdue
3	20	10	30	15	25	15	Ontime
4	10	15	15	20	5	20	Overdue
..	..	…	..	…	..	…	..
177	40	15	5	20	5	15	Overdue

Table 12. K-NN Testing Data

Record	Parents Income	Family Dependent	Father Education	Father Age	Mother Education	Mother Age	Ori Result	Predict Result
178	10	15	5	20	5	10	Overdue	Overdue
179	20	10	20	15	20	15	Ontime	Overdue
180	30	15	5	15	20	15	Ontime	Ontime
181	40	20	20	15	15	15	Ontime	Ontime
..	..	…	..	…	..	…	..	--
236	40	15	5	20	5	15	Overdue	Overdue

$$D(a,b)=\sqrt{\sum_{k=1}^{d}(a_k - b_k)^2} \quad (11)$$

The confusion matrix calculation using 4-Fold Cross Validation is shown in Table 13.

Table 13. Testing Result of K-NN Algorithm

K (on K-NN)	Accuracy	Precision	Recall	F-Measure
1	58%	37%	53%	43%
3	64%	51%	62%	55%
5	61%	50%	57%	53%
7	62%	57%	57%	56%
9	64%	59%	58%	58%
11	66%	62%	60%	61%
13	64%	63%	60%	61%
15	64%	61%	61%	60%
Average	63%	55%	59%	56%

3.5 Calculation of K-NN + Information Gain

With Information Gain feature selection added to to K-NN method, the confusion matrix calculations using 4-Fold Cross Validation are shown in Table 14.

Table 14. Testing Result of K-NN+IGAlgorithm

K (onK-NN)	Accuracy	Precision	Recall	F-Measure
1	58%	42%	51%	46%
3	61%	48%	56%	51%
5	61%	49%	58%	52%
7	64%	52%	60%	55%
9	61%	44%	58%	48%
11	60%	46%	57%	48%
13	59%	46%	54%	48%
15	60%	46%	56%	49%
Average	60%	46%	56%	50%

3.6 Method Comparison

The results of each methods obeservation before to be put alltogether in a table as shown in Table 15.The table shows that the Naïve Bayes + Information Gain algorithm has the best performance with an accuracy value of 67%.

Table 15. Algorithm Comparison

Algorithm	Accuracy	Precision	Recall	F-Measure
Naïve Bayes	64%	61%	56%	58%
NB + IG	67%	64%	59%	61%
K-NN	63%	55%	59%	56%
K-NN + IG	60%	46%	56%	50%

IV. CONCLUSION

This research has produced a data mining model that is able to carry out the Ontime or Overdue class classification based on attributes: parent income, family dependence, father's education, father's age, mother's education and mother's age. To make the best method selection, the researchers made a comparison of four methods, namely the Naïve Bayes method, Naïve Bayes + Information Gain, K-NN, and K-NN + Information Gain. The best method is obtained from the combination of Naïve Bayes algorithm with information gain feature selection which produces an accuracy value = 67%, precision = 64%, recall = 59% and f-measure = 61%.

REFERENCES

[1] S. P. Rochmiyati, "Kebijakan pendidikan bahasa Indonesia dalam perspektif pendidikan nasional," *Caraka*, vol. 1, no. 2, pp. 3–14, 2015.

[2] Menteri Pendidikan dan Kebudayaan, "Peraturan Menteri Pendidikan Dan Kebudayaan Republik Indonesia Nomor 19 Tahun 2016 Tentang Program Indonesia Pintar," pp. 1–9, 2016.

[3] I. Sanjiwani, "Analisis Biaya Pendidikan Dan Dampaknya Terhadap Kualitas Proses Pembelajaran Dan Aspirasi Pendidikan Siswa (Studi Tentang Persepsi Para Siswa Sma Dwijendra Denpasar Tahun Pelajaran 2011/2012)," *J. Adm. Pendidik.*, vol. 3, no. 2, 2012.

[4] H. Asri, H. Mousannif, H. Al Moatassime, and T. Noel, "Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis," *Procedia Comput. Sci.*, vol. 83, no. Fams, pp. 1064–1069, 2016.

[5] M. Hasan, "Menggunakan Algoritma Naive Bayes Berbasis," vol. 9, pp. 317–324, 2017.

[6] W. Gata *et al.*, "Algorithm Implementations Naïve Bayes, Random Forest. C4.5 on Online Gaming for Learning Achievement Predictions," vol. 258, no. Icream 2018, 2019.

[7] Y. A. Gerhana, W. B. Zulfikar, A. H. Ramdani, and M. A. Ramdhani, "Implementation of Nearest Neighbor using HSV to Identify Skin Disease," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 288, no. 1, 2018.

[8] M. Tayyib *et al.*, "Accelerated sparsity based reconstruction of compressively sensed multichannel EEG signals," *PLoS One*, vol. 15, no. 1, p. e0225397, 2020.

[9] N. Y. Moteghaed, K. Maghooli, and M. Garshasbi, "Improving Classification of Cancer and Mining Biomarkers from Gene Expression Profiles Using Hybrid Optimization Algorithms and Fuzzy Support Vector Machine," *J. Med. Signals Sens.*, vol. 8, no. 1, pp. 1–11, 2018.

[10] R. A. Saputra, "Komparasi Algoritma Klasifikasi Data Mining Untuk Memprediksi Penyakit Tuberculosis (Tb): Studi Kasus Puskesmas Karawang," *Semin. Nas. Inov. dan Tren*, no. April, pp. 1–8, 2014.

[11] C. Darujati, "PERBANDINGAN KLASIFIKASI DOKUMEN TEKS MENGGUNAKAN METODE NAÏVE BAYES DENGAN K-NEAREST NEIGHBOR Abstrak," *Univ. Stuttgart*, vol. 13, no. 1, pp. 1–9, 2010.

[12] A. D. Rachid, A. Abdellah, B. Belaid, and L. Rachid, "Clustering prediction techniques in defining and predicting customers defection: The case of e-commerce context," *Int. J. Electr. Comput. Eng.*, vol. 8, no. 4, pp. 2367–2383, 2018.

[13] M. Sadikin and F. Alfiandi, "Comparative study of classification method on customer candidate data to predict its potential risk," *Int. J. Electr. Comput. Eng.*, vol. 8, no. 6, pp. 4763–4771, 2018.

[14] M. Wang, Z. H. Ning, C. Xiao, and T. Li, "Sentiment classification based on information geometry and deep belief networks," *IEEE Access*, vol. 6, pp. 35206–35213, 2018.

[15] S. Saha and D. Nandi, "Data Classification based on Decision Tree, Rule Generation, Bayes and Statistical Methods: An Empirical Comparison," *Int. J. Comput. Appl.*, vol. 129, no. 7, pp. 36–41, 2015.

[16] Y. F. Safri, R. Arifudin, and M. A. Muslim, "K-Nearest Neighbor and Naive Bayes Classifier Algorithm in Determining The Classification of Healthy Card Indonesia Giving to The Poor," *Sci. J. Informatics*, vol. 5, no. 1, p. 18, 2018.

[17] A. Dey, "Machine Learning Algorithms: A Review," *Int. J. Comput. Sci. Inf. Technol.*, vol. 7, no. 3, pp. 1174–1179, 2016.

[18] S. Chormunge and S. Jena, "Efficient feature subset selection algorithm for high dimensional data," *Int. J. Electr. Comput. Eng.*, vol. 6, no. 4, pp. 1880–1888, 2016.

[19] G. C. Sutradana and M. D. R. Wahyudi, "Penerapan Data Mining Untuk Analisis Pengaruh Lama Studi Mahasiswa Teknik Informatika Uin Sunan," *Penerapan Data Min. Untuk Anal. Pengaruh Lama Stud. Mhs. Tek. Inform. Uin Sunan Kalijaga Yogyakarta Menggunakan Metod. Apriori*, vol. 1, no. 3, pp. 153–162, 2017.

[20] I. Handayani, "Application of K-Nearest Neighbor Algorithm on Classification of Disk Hernia and Spondylolisthesis in Vertebral Column," *Indones. J. Inf. Syst.*, vol. 2, no. 1, p. 57, 2019.

[21] Abdullah, Robi W., et al. "Keamanan Basis Data pada Perancangan Sistem Kepakaran Prestasi Sman Dikota Surakarta." Creative Communication and Innovative Technology Journal, vol. 12, no. 1, 2019, pp. 13-21.

[22] Muqorobin, M., Apriliyani, A., & Kusrini, K. (2019). Sistem Pendukung Keputusan Penerimaan Beasiswa dengan Metode SAW. Respati, 14(1).

[23] Muqorobin, M., Hisyam, Z., Mashuri, M., Hanafi, H., & Setiyantara, Y. (2019). Implementasi Network Intrusion Detection System (NIDS) Dalam Sistem Keamanan Open Cloud Computing. Majalah Ilmiah Bahari Jogja, 17(2), 1-9.

[24] K. Kusrini, E. T. Luthfi, M. Muqorobin and R. W. Abdullah, "Comparison of Naive Bayes and K-NN Method on Tuition Fee Payment Overdue Prediction," 2019 4th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE), Yogyakarta, Indonesia, 2019, pp. 125-130, doi: 10.1109/ICITISEE48480.2019.9003782.