Optimization of Feature Selection Using Genetic Algorithms to Increase Payment Delay Prediction Results (Subang Polytechnic State Case Study)

Tri Herdiawan Apandi[1], Dwi Vernanda[2], Rian Piarna[3]
{h.apandi@gmail.com[1], yogurt.nda@gmail.com[2], piarnarian@gmail.com[3]}

Politeknik Negeri Subang[123]

Abstract. The Indonesian government established a policy on tuition fees in higher education, namely the Single Tuition System (UKT), including the Subang State Polytechnic (POLSUB) using this system. Currently in POLSUB there are frequent delays in UKT payments, the number of delays is not much but continues to increase every semester. The purpose of this study is to classify students who are late and timely in making UKT payments using the C4.5 algorithm. Based on the comparison of evaluation and validation results from the four partitions shows that the data results from partitions 1 to 4 producing the best accuracy is on the 2nd partition has an accuracy rate of 75%. While the results using genetic algorithms improve accuracy results to 83.64%. In selecting the initial data collection feature, there are 13 features and after testing there are only 7 features that are most influential.

Keywords: C4.5 algorithm, cross validation, confusion matrix, polsub, UKT.

1 Introduction

The Indonesian Government sets a policy regarding the cost of education in higher education, namely in the form of a Single Tuition System (UKT), each student can have a different amount of UKT this is adjusted to the circumstances economics and study programs at selected State Universities. The procedure for the use and nominal stipulation of UKT has been regulated by Law No. 12 of 2011 concerning Higher Education was assembled by State University Operational Assistance, Single Tuition and Single Tuition [1]. The UKT system was chosen by the government to respond to the problem of the cost of education in universities which continued to increase from year to year, so that it had a negative impact on state universities that seemed exclusive [2].

Subang State Polytechnic (POLSUB) is a state university that implements payment of tuition fees with the UKT system. New students who have been declared accepted or for old students must register and make UKT payments at the beginning of the semester. UKT payments are made by students one week before the lecture begins, but for students who have not paid at the deadline, they can do a payment suspension.

Currently in POLSUB there are frequent delays in UKT payments, although the number of delays is not much but continues to increase every semester. Late payments are influenced by various factors, one of which is the economic condition of the student's family. Delays in UKT payments have an impact on POLSUB's operational activities, especially in academic activities,

ICCSET 2018, October 25-26, Kudus, Indonesia

DOI 10.4108/eai.24-10-2018.2280507

where students are not allowed to attend the lecture process as long as they have not fulfilled their obligations. This also has an impact on POLSUB's financial reporting.

Late payment of UKT can be overcome by predicting the possibility of late payment, so that the POLSUB Treasurer can find out students who are likely to be late in UKT payments. Prediction is done by classifying data usingTechnique Data Mining andalgorithm Decision Tree C4.5 [3]. Data Mining is defined as a process for finding a pattern in the data [4]. One technique in data mining is classification, classification is a process to find a model or function that describes the difference in data class [5]. Decision Tree C4.5 is an algorithm for classification techniques. In general, classification usingalgorithms decision tree has a good level of accuracy because the process can be done quickly and simply [6]. The Decision Tree is considered a feature that is not interconnected for that needs an optimization. Optimization is done to choose a feature that is very influential, the algorithm that will be suitable for this problem is the genetic algorithm. The Genetic Algorithm works by selecting the most influential individuals / features [7]. The contribution to this research is the best way to select features with genetic algorithms.

2 Previous Research

There have been a number of previous studies conducted by many other researchers relating to service satisfaction, as will be explained below:

Much Aziz Muslim in his article entitled "Improving Algorithm Accuracy Using Split Models for Credit Card Risks" The result is the first dataset is selected 16 features, second dataset 12 features, third dataset 8 features, fourth dataset 4 features. Each dataset that has been divided based on its features will be calculated using the C4.5 Algorithm. From the results of this accuracy calculation, the best accuracy obtained in the third dataset is 75.1%. This proves that the selection of features is very influential on the results of accuracy. [8]

Rezha, Rochmah and Siswidiyanto with the title "Analysis of the Influence of Public Service Quality on Community Satisfaction (Study of the Recording Service of Electronic Identity Cards (E-KTP) in Depok City)". Selection of the right features will affect the results of the accuracy that will be obtained. There are feature categories that will be connected to each other [9].

Hamta in his research entitled "Analysis of the Application of Data Mining in Measuring the Level of Community Satisfaction in Batam Samsat Services". In this paper explained the effect of variables or features greatly affects the level of accuracy [10].

The research conducted by Maris with the title "Customer Satisfaction Analysis Using Algorithm C4.5". Themethod C4.5 implementation ofusing customer data can be used to determine customer satisfaction. The ratio of training data used affects the value of accuracy in each experiment [11].

Universal in his research "Naïve Bayes Algorithm Optimization Using Genetic Algorithm for the Prediction of Fertility (Fertility)"Naïve Bayes plus optimization using Genetic Algorithms will improve the accuracy of the results [12].

Wahyuni et al with the title of the research "Prediction of the results of the Jakarta dki legislative election using naïve bayes with genetic algorithms as a selection feature". While predictions using naïve bayes and AG as a selection feature will increase the level of predictive accuracy [13].

Oman Somantri and M. Khambali (2017) in his research entitled "Feature Selection Classification of Short Story Categories Using Naïve Bayes and Genetic Algorithms" results

showed that the Naive Bayes algorithm performed feature selection using Genetic Algorithms experienced an accuracy increase of 6% [14].

Based on the results of the literature study will be compared the results of accuracy using the distribution of training data and data testing will be compared the results of accuracy by using optimization using genetic algorithms. The result is that children get features that will affect and features that improve accuracy.

3 Research Methodology

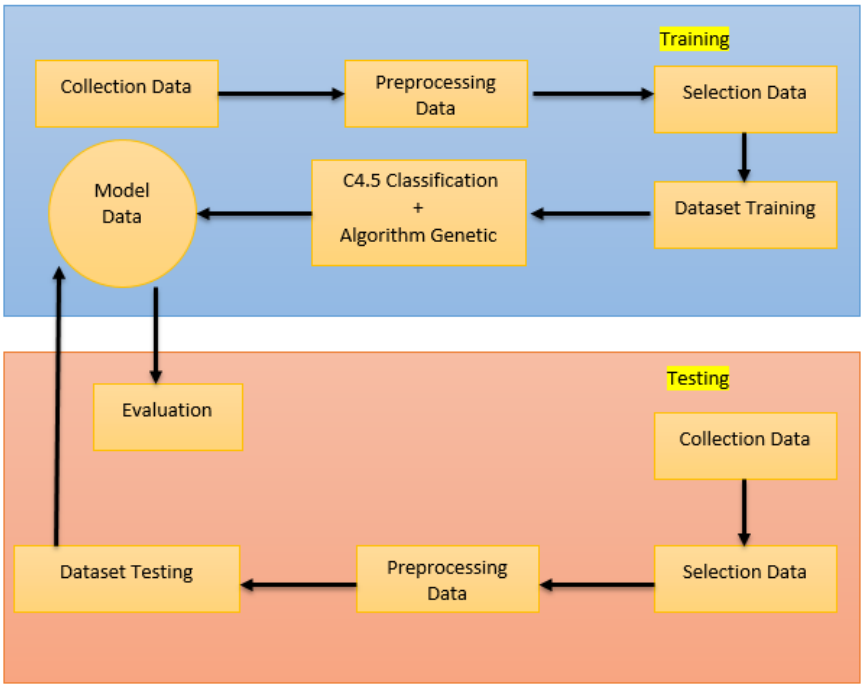

Fig. 1. Research methodology.

This research contained 2 major processes that were carried out, namely the training process and theprocess testing. The training process is a process that runs in order to form a data model that will be used for the next stage. The next stage is the stage of theprocess testing to see the model that has been formed in thephase training has good results or not.process Training and testing there are several processes. Theprocess training includes:

a. Colecction data is collecting data from existing students at Subang State Polytechnic. Collecting data using questionnaires, questionnaire instruments can be seen in appendix 3. These questionnaires are variables that will affect student UKT payments, these variables are obtained from the results of interviews with financial staff obtained variable delay in UKT payments. This dataset has 13 selected features, this feature is: father's salary, mother's salary, dependents, expenses, transportation, snacks, place to

live, family, activities, year of entry, campus facilities, academic services, and financial services.

b. Preprocessing data is converting numerical data into nominal data so that it can be processed in the next stage.
c. Selection of data is the selection of data to be used for the next stage. This selection will discard incomplete data, the incompleteness of the data is caused when filling out questionnaires that are not in accordance with the filling rules.
d. Training dataset is data that is ready to be processed using the C4.5 algorithm. After the data is ready to be processed, the data classification process will be carried out.
e. C4.5 Clasification and Genetic Algorithm is the process of data classification using the C4.5 algorithm to produce a decision tree (decision tree) data that will be used for theprocess testing plus optimization to improve accuracy.
f. Data Model results obtained after classifying by Algorithm C4.5 in the form of a decision tree (decision tree).
g. Evaluation done by analyzing the results of the classification. Data measurement is done by confusion matrix to evaluate the results of thealgorithm decision tree (C4.5). Confusion matrix is a table consisting of the number of lines of test data that are predicted to be true and incorrect by the classification model.

In the testing process, it is more or less the same as theprocess, *training* but there is a small difference in the process that is carried out, namely not doing classification. Data that is ready to be processed will be tested using a model produced bydata *training*. The model produced by thedata *training* is a decision tree (*decision tree*) [15]. After testing the model (decision tree) will be evaluated with a *confusion matrix* in the form of accuracy, *precedence*, and *recall*. These results will be the basis that the decision tree generated from training data has a good level of accuracy in predicting delays in student UKT payments.

4 Results

4.1 Testing results

The data used in this study amounted to 102 data obtained from questionnaires distributed to students of academic year 2017/2018 Information Management Polytechnic of Subang. From the total sample used then divided into 4 partitions, partition 1 for 90% training data and 10% testing data, partition 2 for 80% training data and 20% testing data, and partition 3 for 70% training data and 30% testing data, and partition 4 for 60% training data and 40% data testing.

Based on the sample data tested usingmethod decision tree withdata testing of 20% obtained an accuracy rate of 50%. As shown in table 1 below.

Table 1. Performance vector data partition 2.

Accuracy: 50%			
	True on time	True late	Class precision
Pred. on time	9	6	60.00%
Pred. late	4	1	20.00%
class recall	69.23%	14.29%	

Based on the sample data that has been tested using themethod decision tree (C4.5)then it will be validated to get the best results so that it can be applied to predict the possibility of UKT payment delays at the Subang State Polytechnic. The results of the validationmodel decision tree of partition 2 datausing cross validation obtained an accuracy rate of 75%. As listed in table 2 below.

Table 2. Performance vector data partition 2.

Accuracy: 75%			
	True on time	True late	Class precision
Pred. on time	12	4	75.00%
Pred. late	1	3	75.00%
class recall	92.31%	42.86%	

Based on sample data that has been tested using themethod *decision tree* (C4.5)and optimized using genetic algorithms, the results will be listed in table 3 below.

Table 3. Performance vector uses optimization.

Accuracy: 83.64%			
	True on time	True late	Class precision
Pred. timely	20	3	86.96%
Pred. late	14	64	82.28%
class recall	58.87%	95.59%	

4.2 Comparison of testing results

Based on the overall data that has been tested, the authors get a comparison of the results between partition 1, partition 2, partition 3, and partition 4 to analyze the possibility of delays in UKT payments at Subang State Polytechnic with the following results: Accurate comparison between Partition 1, Partition 2, Partition 3, and Partition 4. Comparison of accuracy can be seen in Figure 2. Comparison of Data Accuracy.

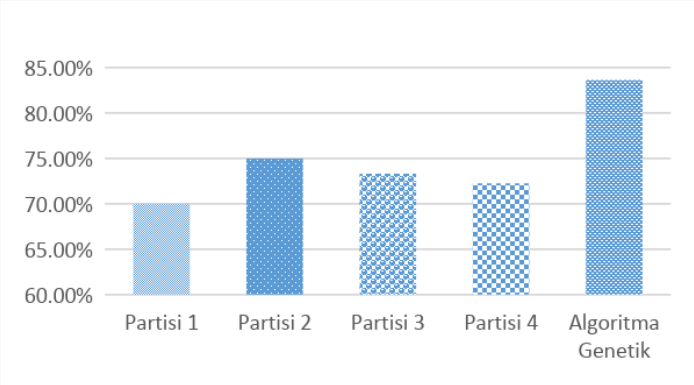

Fig. 2. Comparison of data accuracy.

From Figure 2 above shows that partition 2 data testing has a greater accuracy than other partition data, which is 75%. Partitions 1 through 4 show poor results because the features of the initial results collected by 13 features are many features that are not interconnected. Using genetic algorithms produces excellent accuracy of 83.64% and obtained features that should be used in subsequent studies are father's salary, maternal salary, dependents, activities, pocket money, transportation and campus facilities.

5 Conclusion

Based on the results of the research that has been done, some conclusions can be drawn, including:

a. From the results of the data set partition test found the highest accuracy rate on partition 2 data is 80% training data and results in an accuracy of 75.00%. Thus the model obtained from partition 2 can be used as training data to predict new data.
b. Attributes of academic services have a considerable influence in all partitions tested, after an experiment using RapidMiner that the attributes of academic services are at thenode top.
c. Genetic Algorithms affect the accuracy of the results to 83.64% and the most influential feature is the father's salary, maternal salary, dependents, activities, pocket money, transportation and campus facilities.

Reference

[1] Muchsin and Sudarma, "Application of Fuzzy C-Means for the Determination of the New Student College Single Money," *Lontar Komput*, vol. 6, no. 3, 2015.
[2] "R. Indonesia, Law," no. 12, 2012.
[3] I. Print, I. Online, H. Sulastri, and A. Gufroni, "Application Of Data Mining In Patient," *GROUPSJ. Teknol. Sist. Inf*, vol. 2, pp. 299–305, 2017.
[4] I. H. Witten and E. Frank, *Data mining Practical Machine Learning Tools and Techniques*, 2nd ed. 2005.
[5] W. G. W. X, Z. X,., "Data mining with big data," *EEE Trans. Knowl Data Eng*, vol. 26, no. 1, pp. 97–107, 2014.
[6] Han and M. Kamber, *Data mining: Concepts and Techniques. 3rd Edition*, 1st ed. San Fr. Morgan Kaufmann Publ, 2011.
[7] A. Zagorecki, "Feature Selection for Naive Bayesian Network Ensemble using Evolutionary Algorithms." pp. 381–385, 2014.
[8] M. A. Muslim, A. Nurzahputra, and B. Prasetiyo, "Improving Accuracy of C4. 5 Algorithm Using Split Feature Reduction Model and Bagging Ensemble for Credit Card Risk Prediction," 2018, pp. 141–145.
[9] F. Rezha, S. Rochmah, and Siswidiyanto, "Analysis of the Influence of Public Service Quality on Community Satisfaction (Study of Electronic Identity Card Recording Services (e-KTP) in Depok City)." p. 10, 2016.
[10] F. Hamta, "Analysis of the Application of Data Mining in Measuring the Level of Community Satisfaction in Batam Samsat Services." pp. 1–17, 2017.
[11] R. Maris, "Customer Satisfaction Analysis Using C4 Algorithms. 5." pp. 1–14, 2016.
[12] D. Buani, "Optimization of the Naïve Bayes Algorithm by Using Genetic Algorithms for Prediction of Fertility," vol. 4, no. 1, pp. 54–63, 2016.
[13] D. Wahyuni, T. Sutojo, and A. Luthfiarta, "Prediction Of DKI Jakarta Legislative Election Result

Using Naïve Bayes With Genetic Algorithm As Selection Features." pp. 1–14, 2014.
[14] O. Somantri and M. Khambali, "Feature Selection Classification of Short Story Categories Using Naïve Bayes and Genetic Algorithms," *J. Nas. Tech. Electr. Eng. Inf.*, vol. 6, no. 3, pp. 301–306, 2017.
[15] K. Pliakos, "Mining features for biomedical data using clustering tree ensembles," Biomed. Informatics," vol. 85, pp. 40–48, 2018.