2017 4th International Conference on New Media Studies
Yogyakarta, Indonesia, November 08-10, 2017

Determining Single Tuition Fee Of Higher Education In Indonesia : A Comparative Analysis Of Data Mining Classification Algorithms

Muhammad Nur Yasir Utomo[1], Adhistya Erna Permanasari[2], Eddy Tungadi[3], Irfan Syamsuddin[4]
[1,2]Department of Electrical Engineering and Information Technology
Universitas Gadjah Mada
Yogyakarta, Indonesia
[3,4]Computer and Network Engineering Study Program
Department of Electrical Engineering
Politeknik Negeri Ujung Pandang
Makassar, Indonesia
[1]m.nur.yasir@mail.ugm.ac.id, [2]adhistya@ugm.ac.id, [3]eddy.tungadi@poliupg.ac.id, [4]irfans@poliupg.ac.id

***Abstract*—Student's Single Tuition Fee or *Uang Kuliah Tunggal* (UKT) is a subsidy policy in higher education by the Indonesian government. This policy regulates the tuition fees incurred by each student at each semester in every higher education institutions. Since the cost of UKT expenses is influenced by the financial ability of each student, therefore the cost of education among students must be grouped into several classes. Until recently, there has been no standard to make such classification whereas such determination is an important task to solve by every higher institution in Indonesia. This study aims to compare five data mining classification algorithms (Gaussian Naïve Bayes, Multinomial Naïve Bayes, Bernoulli Naïve Bayes, Decision Tree and SVM) to find the best algorithm for the case of determining the UKT classes. The experiment is conducted using 230 training data and 10-fold cross-validation evaluation. Based on the result, Decision Tree managed to obtain average accuracy value of 0.814 or 81.4%. Finally, Decision Tree is used to classify the UKT classes of 3258 data of students.**

Keywords—classification algoritm; data mining; supervised learning; machine learning; ukt; student financial;

I. INTRODUCTION

The Indonesian government is aware that the plurality of societies' economic condition is very diverse. Therefore, policies related to public spending for services should pay attention to justice aspects. These policies are to ensure every citizen can access the services provided by the state, including education service by implementing Student's Single Tuition Fee or *Uang Kuliah Tunggal* (UKT) which has been set on Permenristekdikti Number 22 Year 2015 [1].

UKT is a policy that regulates the tuition fees incurred by each student in each semester at the College. UKT costs are divided into various classes. The determination of UKT cost class for a student is influenced by his financial capability. In its implementation, universities should analyze student data to know their financial capability to determine the appropriate UKT classes for them.

The main criteria used to determine the UKT classes are the parent income criteria [1]. However, the financial capability of the students cannot be seen only from the income of their parents. Other influential criteria such as the number of sibling, local origin, and the study program field also have an impact on the students' financial capability.

With the number of criteria and data to be analyzed, the determination of the UKT classes for each student takes a lot of time, effort and cost if it is done manually [2]. To address this problem, a method is needed to help decision-making in determining the appropriate UKT classes for each student based on their financial capabilities.

In this research, data mining classification algorithms are explored and compared to one and the other in the case of determining the student's UKT classes. We used supervised learning techniques to test the classification algorithms. The classification algorithms that we explore in this research are Decision Tree, Naïve Bayes, and Support Vector Machine (SVM).

II. LITERATURE REVIEW

Student's Single Tuition Fee or *Uang Kuliah Tunggal* (UKT) is an Indonesian government policy to guarantee the right of every citizen to get the proper education. UKT is designed as an educational financing solution at the university level by applying the concept of cross-subsidy. Amount of expenses on UKT, as referred to Article 2 paragraph (2) Permenristekdikti Number 22 Year 2015 [1], consists of several classes of cost, ranging from low-cost to normal-cost class.

The determination of a student's UKT class is a classification problem that analyzes the student's financial capability based on criteria that can affect their financial condition. For classification problems, data mining is commonly used techniques [3].

TABLE I. NEW STUDENT REGISTRATION DATABASE OF STATE POLYTECHNIC OF UJUNG PANDANG

No.	Participant No.	Local Origin	Father's Income	Mother's Income	Num. of Sibling	Prodi Field
1	10082	Pangkajene Kepulauan Regency	Rp. 1.000.000 - Rp. 2.000.000	No income	4	Engineering
2	10031	Pangkajene Kepulauan Regency	Rp. 2.000.000 - Rp. 3.000.000	No income	3	Engineering
3	20063	Polewali Regency	Rp. 2.000.000 - Rp. 3.000.000	No income	2	Commerce
…	…	…	…	…	…	…
3528	11849	Makasar City	Rp. 2.000.000 - Rp. 3.000.000	Rp. 1.000.000 - Rp. 2.000.000	5	Engineering

Data mining is a series of processes to extract new information from a set of data [4]. Data mining is widely used for the purposes of classification, prediction, and clustering. For the case of determining the UKT class, data mining can be used to classify financial capability. Researchers have shown that various purposes of financial capability analysis can be done with data mining [4] such as determination of college students that in financial hardship [3], selecting which student get scholarship grant [5], and the classification for financial fraud detection [6].

Research [3] conduct an analysis to find and determine students with difficult economic conditions based on their family's capability in paying tuition fees. This study is conducted because the economic conditions that burden the students can negatively affect mental health, academic performance, student behavior, and their social life [3]. Finally, this research explores data mining with a multi-label learning problem. The students' habits on campus are explored from various perspectives such as student card activity, internet usage and places visited by the student on campus. This study compares several methods such as Support Vector Machine (SVM) and Multiple Kernel Learning (MKL) with a self-developed method called Dis-Hard. The proposed method by research [3] gets better performance results from existing methods. This study shows, the economic capability of a student can be extracted with data mining.

Data mining and machine learning as techniques to determine the eligibility of students in receiving educational scholarships is conducted on research [5]. This research takes some students characters such as moral, intellectual, to health as variables to determine the eligibility of the students to get scholarship. Using the Naïve Bayes method (NB), research [5] manages to find which student deserve or not deserve to get a scholarship. This research shows the potential of Naive Bayes method can be used in the case of UKT.

The use of data mining classification techniques for financial needs is also conducted by research [6]. The aim of this research is to use data mining classification method as a detector in financial fraud. The study uses three different methods, namely SVM, Naïve Bayes and K-Nearest Neighbors (KNN). Research [6] has finally found that SVM delivers better result compared to NB and KNN.

Other data mining classification method is Decision Tree which is work by mapping training data into a tree or hierarchical structures [7], [8]. Research [9] shows that extracting information also can be done using the Decision Tree method. The research utilizes the decision tree method to classify bad debts. The analysis that performed on the financial market object finds that the decision tree gives high result of accuracy to classify bad debts, it is even can provide an explanation of the factors that affect each classification result.

Based on the literature on the ability of data mining classification algorithm in the field of education and finance above, the research eventually explored and compared data mining classification algorithms to find the best algorithm for determining the student's UKT classes. Some of the classification algorithms used are Naïve Bayes, SVM, and Decision Tree.

III. RESEARCH METHODOLOGY

There are three major stages in this research, namely defining variables, data normalization, and the training-test process of the algorithms with supervised learning. These stages are described as follows:

A. Data and Reasearch Variables

The source of data used in this research is data from the new student enrollment database of State Polytechnic of Ujung Pandang (PNUP) in 2014. In this database, the data of students that required for determining their UKT classes are available. The amount of available data reached 3528 students data as shown in Table 1. This research uses the data to get variables that can affect student financial capability, including:

- Father and Mother Income

  Parental income variable is chosen because it can represent economic income of the student's family. Since one kind job can result different income, the value of the parent's income gives clearer representation of the student economic condition.

- Local Origin

  The local origin variable is associated with an additional expenditure that students should incur if they are coming from different regions with the region of the college. For example, additional expenditure for boarding and transportation. These additional expenditures have an impact on the economic capability of the students.

- Number of Sibling

  The number of children in a family affects the economic expenditure. The more children, the more likely the expenditure is. So the variable number of siblings of the students should be taken into account as variables that may affect their financial capability.

- Prodi Field

  The Prodi Filed variable is considered as a variable that can affect economic expenditure because the regulation on Permenristekdikti Number 22 Year 2015 shows that the tuition fee for Engineering field is greater than Commerce field.

Other data that needed is the data of UKT expense classes. This data are obtained from the attachment of Permenristekdikti Number 22 Year 2015 [1] for State Polytechnic of Ujung Pandang as shown in Table 2. These data is used as the target label.

TABLE II. UKT CLASSES ACCORDING TO GOVERNMENT RULE

Single Tuition Fee (Per Semester) for PNUP in Rupiah				
Class 1 (UKT1)	**Class 2 (UKT2)**	**Class 3 (UKT3)**	**Class 4 (UKT4)**	**Class 5 (UKT5)**
500.000	1.000.000	1.750.000	3.000.000	4.000.000

B. Data Normalization

The data obtained from PNUP contain two type of data, which are integer and string. Data normalization is conducted on each of the influential features that are still in string format to make it easier to process. These features are Parents Income, Local Origin and Study Program Field. We convert these string data into integer data using some rules. The rules can be explained as follows:

- Normalization features of father's income and mother's income

  Student data obtained is still in the form of raw data, where the field of father and mother income are still in string format, these data do not match the feature format that can be processed in machine learning. So the data is converted to an integer using the rules shown in Table 3.

TABLE III. RULE FOR INCOME NORMALIZATION

Income Variations	Normalization
No income	0
Rp. 1.000.000 - Rp. 2.000.000	1
Rp. 2.000.000 - Rp. 3.000.000	2
Rp. 3.000.000 - Rp. 4.000.000	3
Rp. 4.000.000 - Rp. 5.000.000	4
Over Rp. 5.000.000	5

- Normalization feature of local origin

  Local origin normalization is conducted by the rule that if the student origin is not from Makassar City then it is considered as 1, and if the student is from Makassar City is considered as 0:

TABLE IV. RULE FOR LOCAL ORIGIN NORMALIZATION

Local Origin	Normalization
Makassar City	0
Not Makassar City	1

- Normalization feature of study program field

  In this study, we used the field of the student's study program that is engineering field or commerce field. The data containing the 'engineering' string is converted into integer '2' and 'commerce' string into integer '1', the normalization rule for this feature shown in Table 5.

TABLE V. RULE FOR PRODI FIELD NORMALIZATION

Prodi Field	Normalization
Commerce	1
Engineering	2

Finally, all the data in Table 1 are normalized so that it has the same data type that is the integer shown in Table 6 below:

TABLE VI. FINAL RESULT OF DATA NORMALIZATION

No.	Participant No.	Local Origin	Father Income	Mother Income	Num. of Sibling	Prodi Field
1	10082	1	1	0	4	2
2	10031	1	2	0	3	2
3	20063	1	2	0	2	1
…	…	…	…	…	…	…
3528	11849	0	2	1	5	2

Furthermore, all variables will be used as a single feature to determine the UKT classes of each student.

C. Training dan Testing

In this section, some algorithms are tested to find the best algorithm in determine UKT classes of the students. Some of the algorithms tested in this research are Gaussian NB, Multinomial NB, Bernoulli NB, Decision Tree and SVM. The supervised learning technique is used to train the classifier.

The supervised learning technique is used because its capability that can make predictions based on the labeled training data used [10]. This capability makes the supervised learning technique can replicate the humans decision if it

trained using labeled data from human assessment. The process from training to obtaining predictions is illustrated in Figure 1.

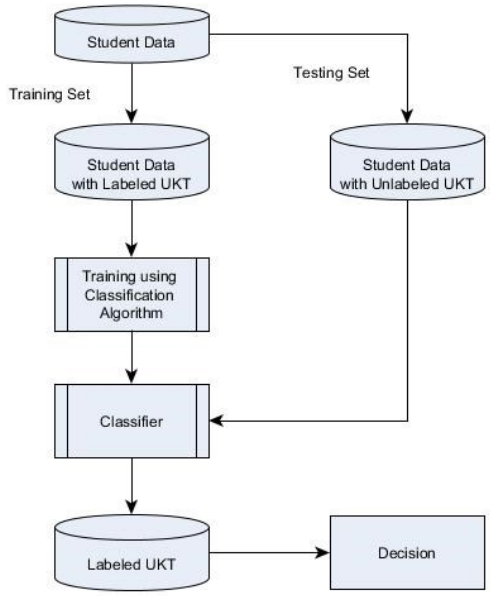

Fig. 1. Classification Process using Supervised Learning

For the experiment, only 230 data from 3528 unlabeled data of the student data are used. The 230 data are then labeled manually based on human assessment as UKT1 (64), UKT2 (43), UKT3 (51), UKT4 (35) and UKT5 (37).

Normalization of influential variables that have been made before then used as features. For each data in 230 data, every features are combined to forms a numerical patterns such as [2,0,3,0,1], [3,2,1,1,2] etc. The numerical pattern of each data is then used to train the classifier.

Algorithm 1: Classification Process for Each Algorithm

Labelled = $\{x_i y_i\}_{i=1}^{l}$

Unlabeled = $\{x_n\}_{n=1}^{u}$

TrainClassifier = classificationAlgo($\{x_i y_i\}_{i=1}^{l}$)

for all Unlabeled **do** n interation

- classify $f(x)$ using TrainClassifier
- save the result

Endfor

Fig. 2. Classification Process for Each Algorithm

The training-prediction process is illustrated in Figure 2. The experiment also performs 10-fold cross-validation to ensure the accuracy level of each algorithm. All of these processes are conducted using Python programming language. The consideration of the use of Python because it can process data quickly, besides, there is Scikit-Learn package support that has the classification function of Decision Tree, Support Vector Machine, Naïve Bayes and others.

IV. Result and Discussions

The evaluation result of the training and prediction process can be described as follows:

A. Results of Each Algorithm

The result of the each algorithms that is trained using supervised learning shows the following results:

TABLE VII. Accuracy Results of Classification Algorithms

	Gaussian NB	Multinomial NB	Bernoulli NB	Decision Tree	SVM
1	0.577	0.615	0.231	0.731	0.654
2	0.800	0.640	0.480	0.840	0.840
3	0.640	0.480	0.400	0.760	0.680
4	0.792	0.708	0.250	0.917	0.958
5	0.870	0.609	0.217	0.739	0.783
6	0.727	0.636	0.500	0.818	0.909
7	0.864	0.682	0.364	0.864	0.818
8	0.857	0.476	0.429	0.857	0.857
9	0.762	0.524	0.286	0.762	0.714
10	0.810	0.714	0.381	0.857	0.857
Max	0.870	0.714	0.500	0.917	**0.958**
Mean	0.770	0.608	0.354	**0.814**	0.807

Table 7 shows that this research uses 10-fold cross-validation on each of the classification algorithms tested. Decision Tree and SVM algorithms are two algorithms that always get adjacent results on each iteration, even getting the same accuracy value at the 2nd, 8th and 10th iterations.

Table 7 also shows that SVM get the highest maximum accuracy value of 0.958. However, Decision Tree algorithm yields as the best algorithm with highest average accuracy value of 0.814.

B. Selection of The Best Algorithm

Decision Tree and SVM are the two suitable algorithms for determining UKT classes. Based on Table 7, Decision Tree is the algorithm with the best average accuracy value. Decision Tree prediction results are strongly influenced by the datasets and variables used, the more obvious features of training data the more Decision Tree can clearly build the tree to predict the class [11]. A good and stable prediction result by Decision Tree also shows that the training set and the variables used are appropriate for the case [12]. On the other hand, SVM is the algorithm that get the highest maximum accuracy value. However, SVM prediction results fluctuate due to its originally design for binary classification, while there are 5 classes in the case of determining the UKT class.

Decision Tree finally chosen as the best algorithm to determine UKT class. The task requires an algorithm that works stably, the decision tree has this capability. Table 7 shows that Decision Tree has better average of accuracy value compared to other algorithms, including SVM.

C. Implementing Decision Tree Algorithm

Finally, with the decision tree algorithm and 230 data for training set, 3528 student data is processed to determine their UKT classes. The result is shown in Figure 3.

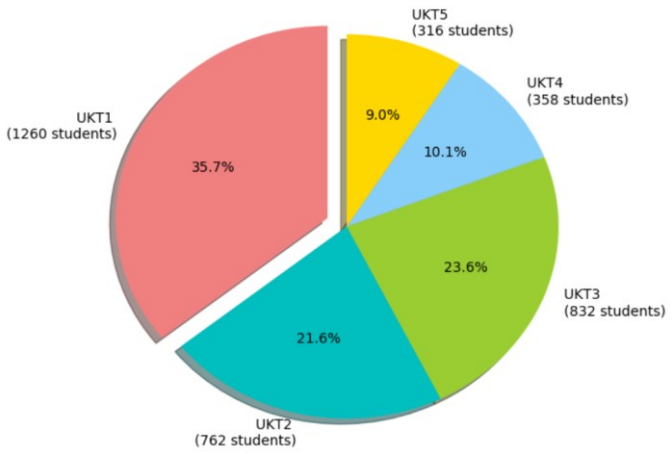

Fig. 3. Result of determining UKT class of 3528 student data

Figure 3 shows the distribution of predicted UKT classes of 3528 student data. Of all available data, 1260 data or 35.7% of data are UKT1, 762 data or 21.6% of data are UKT2, 832 data or 23.6% of data are UKT3, 358 data or 10.1% of data are UKT4 and 316 or 9.0% of data are UKT 5.

V. Conclusion

Data mining classification algorithm is suitable to be applied in case of determination of student's UKT classes. Based on the results of experiments conducted with five algorithms using 230 training data and 10-fold cross-validation, Decision Tree is found as the most suitable algorithms for determining UKT classes. Decision Tree prediction results yield average accuracy value of 0.814 or 81.4%. Finally, the Decision Tree algorithm is argued as the best algorithm to be used in determining the student's UKT classes.

In the future, this study might be extended in several ways, such as using the case as a new problem based learning to enhance teaching materials [13] and extending the experiment by using more training data and variables to get firm difference of accuracy between the algorithms [14].

Acknowledgment

This work is partially supported by the Indonesia Endowment Fund of Education (LPDP), Ministry of Finance, Indonesia and State Polytechnic of Ujung Pandang (PNUP), Makassar, Indonesia.

Reference

[1] Ministry of Research, Technology and Higher Education of the Republic of Indonesia (Kemenristekdikti RI), "Peraturan Menteri Riset, Teknologi, dan Pendidikan Tinggi Republik Indonesia Nomor 22 Tahun 2015," *Kemenristekdikti RI*, 2015.

[2] A. Dutt, M. A. Ismail, and T. Herawan, "A Systematic Review on Educational Data Mining," *IEEE Access*, Vol. 5, pp. 15991–16005, 2017.

[3] C. Guan, X. Lu, X. Li, E. Chen, W. Zhou, and H. Xiong, "Discovery of College Students in Financial Hardship," *Proceedings - IEEE International Conference on Data Mining, ICDM*, pp. 141–150, 2016.

[4] B. T. Xiaoyu Liu, Zhe Huang, "Review on the data mining technology and the applications on financial analysis area," *International Conference on Communication and Electronics Systems (ICCES)*, pp. 1–7, 2016.

[5] W. Wei, J. Han, J. Kong, and H. Xia, "Prediction of The Scholarship Using Comprehensive Development," *International Conference on Enterprise Systems*, pp. 183–188, 2016.

[6] S. O. Moepya, "Applying Cost-Sensitive Classification for Financial Fraud Detection under High Class-Imbalance," *2014 IEEE International Conference on Data Mining Workshop*, vol. 202, no. January, pp. 183–192, 2014.

[7] J. Han and M. Kamber, "Data Mining: Concepts and Techniques," *Annals of Physics.*, vol. 54, p. 770, 2006.

[8] E. Drabikov and E. Feckov, "Decision Trees - a Powerful Tool in Mathematical and Economic Modeling," *Internasional Carpathian Control Conference (ICCC)*, pp. 34–39, 2017.

[9] A. T. Khadija Alaoui Hamidi, Abdelaziz Berrado, Loubna Benabbou, "A Classification Based Framework for Credit Risk Assessment in the Moroccan Financial Market," *International Conference on Intelligent Systems: Theories and Applications (SITA)*, pp. 5–10, 2016.

[10] P. R. D. Roshani Ade, "Classification of Students by Using an Incremental Ensemble of Classifiers," *International Conference on Reliability, Infocom Technologies and Optimization (ICRITO)*, 2014.

[11] T. Hasbun, A. Araya, and J. Villalon, "Extracurricular Activities as Dropout Prediction Factors in Higher Education Using Decision Trees," *International Conference on Advanced Learning Technologies (ICALT)*, pp. 242–244, 2016.

[12] H. Gulati, "Predictive Analytics Using Data Mining Technique," *International Conference on Computing for Sustainable Global Development*, pp. 713–716, 2015.

[13] I. Syamsuddin, "Problem Based Learning on Cloud Economics Analysis Using Open Source Simulation," *International Journal of Online Engineering*, vol. 12, no. 6, pp. 4–9, 2016.

[14] R. Jindal, R. and M.D.Borah, "A survey on educational data mining and research trends," *International Journal of Database Management Systems*, vol. 5, no. 3, pp. 53–73, 2013.