Determining student's single tuition fee category using correlation based feature selection and support vector machine

by Wiyli Yustanti

Determining Student's Single Tuition Fee Category Using Correlation Based Feature Selection and Support Vector Machine

W Yustanti
Department of Informatics Engineering
Universitas Negeri Surabaya
Surabaya, Indonesia
wiyliyustanti@unesa.ac.id

Y Anistyasari
Department of Informatics Engineering
Universitas Negeri Surabaya
Surabaya, Indonesia
yenian@unesa.ac.id

Elly Matul Imah
Mathematics Department
Universitas Negeri Surabaya
Surabaya, Indonesia
ellymatul@unesa.ac.id

***Abstract*— the government has issued the regulation about the enactment of a single tuition fee based on the socio-economic conditions of each student since 2013. All public universities are required to implement this policy. Therefore, each university needs to create a formulation that can be used to categorize a student into which cost group. The results of the data collection found that the parameters used to determine the classification of tuition fees between one universities with another are different. In this research, taken a sampling of student data at one public university database. Before classifying, the attribute of dataset was selected using correlation based feature selection (CFS). The classifier hath has been used in this study is Support Vector Machine (SVM).**

Keywords—single tuition fee; confirmatory factor analysis; ordinal data; structural equation modelling, sosio economy status

I. INTRODUCTION

Single Tuition Fee (STF) is a policy of Indonesian government which is enforced since 2013. This policy is aimed at helping and easing the cost of student education with cross-subsidy system through the category of STF adjusted to the **student's socio**-economic status (SES). Single tuition fee makes it easy to predict student tuition expenses for each semester and there will be no additional fees. In the regulation of the ministry of higher education number 39 of 2017 mentioned that students can apply for payment waivers of STF if the university decision is not in accordance with the conditions of their parents. This policy raises the consequence that universities should be able to make an accurate formulation regarding the determination of STF categories on each student based on their economic capabilities. The false prediction in the classification of STF resulted in complaints or even the failure of students to re-register as a freshman at the university. Based that reason, the main goal of this research is to develop the model of classification algorithm in predicting STF based on university database in order to assist the authority makes the decision in **classifying the student's STF.**

Some studies have been conducted related to determine the single tuition. Ariady et al has done a project about determining STF using Fuzzy C Means approach by using 7 variables predictors. This paper did not describe clearly the accuracy of prediction based the method because the focus of that research is just applying Fuzzy C Means in decision support system (DSS) [1]. The second paper is presented by Gumelia et al in national conference which had studied about prediction the STF of level 1 based on Additive Weighting (SAW) method with 10 variables. This study explained that there is a specific constraint for determining first level of STF, it must meet at least minimum 5 percent from the total of new students. [5]. Previous researchers are not focus to model the algorithm but just implement the method in a sample data. Considering that there is still no research related to the classification modeling for the determination of STF category, then in this study will be tested to see how the best model of the determination process of STF based on existing data in the university database. The formal definition of STF is the cost of each student based on his economic ability [8]. The STF consists of several groups determined on the basis of economic capacity of student, parents of students or others who finance it. Based on the definition of STF, the term of socio-economic is also very important. Socioeconomic status (SES) is one of the most widely studied constructs in the social sciences. Most of them describe that SES is constructed from three kinds of capitals. These are financial capital (material resources), human capital (non-material resources such as education), and social capital (resources achieved through social connections). In other paper mentioned that SES is about family background it is said that in determining family background mostly used parental income, education and occupations and also home resource. [9].

The classification of SFS in this study uses correlation based feature selection (CFs) and Backpropagation Neural Network. CFs are commonly used algorithms for feature selection as performed by Elen et al, she use CFS for Feature Selection Methods in the Analysis of a Population Survey Dataset[11]. Machine learning is also often used for SES classification, like researches that conducted by Michael[12], Zhang[13], Victor Soto[14]. Based on the literatures, this study used CFS as feature selection and SVM as classifier for **Determining Student's** Single Tuition Fee Category.

978-1-5386-3172-0/17/$31.00© 2017 IEEE

II. DATA PREPROCESSING

A. Single Tution Fee (STF)

Paragraph 5 of article 1 of the ministerial regulation No. 55 of 2017 on a single college tuition at the state university defines STF as the cost of each student based on his economic ability. The STF consists of several groups determined on the basis of economic capacity of student, parents of students or others who finance it. Each public university proposes a STF grouping model to the finance minister in order to be established formally. The regulation also states that university leaders can provide STF relief and / or re-enforce STF to students if there are [2]:

- The discrepancies in the economic capacity of the student submitted by the student, the student's parent, or any other party financing it; And / or
- Changes in the economic capability of students, parents, or others who finance it.

Besides, it is also stipulated that public university is prohibited to collect base money and / or other levies other than STF from new students of diploma and undergraduate program for the benefit of direct learning service. The university does not bear student fees consisting of personal cost, the cost of community service program, dormitory fees and learning activities and research carried out independently.

B. Socioeconomic Status (SES)

There are several definition of socioeconomic status based on some reviews of papers. Socioeconomic status (SES) is one of the most widely studied constructs in the social sciences. Most of them describe that SES is constructed from three kinds of capitals. These are financial capital (material resources), human capital (nonmaterial resources such as education), and social capital (resources achieved through social connections). SES is about family background It is said that in determining family background mostly used parental income, education and occupations, and also home resource [9]. An expanded SES measure could include measures of additional household, neighborhood, and school resources.

This study will rely on the definitions and measures as described by a recommendation of panel discussion of experts result about socioeconomic status as a construct [3]. SES can be defined wid**ely as one's access to financial, social, cultural, and** human capital resources. The path diagram of this concept can be explain in Fig. 1. Fig. 1 also is known as path analysis, which is a development technique of multiple linear regression to test the contribution shown by path coefficient on each path diagram of the causal relationship between variables X_1 X_2 and X_3 to Y and their impact on Z. This analyzes the causal relationships that occur in multiple regression if the independent variables affect the dependent variable not only directly but also indirectly. Based on fig.1, SES is constructed from 4 principle factor (financial, human, social and culture) which are called unobserved variable (oval symbol). This is also called as latent variable. Latent variable cannot directly measured but it can be measured using other variable which are called as indicator variable (square symbol)

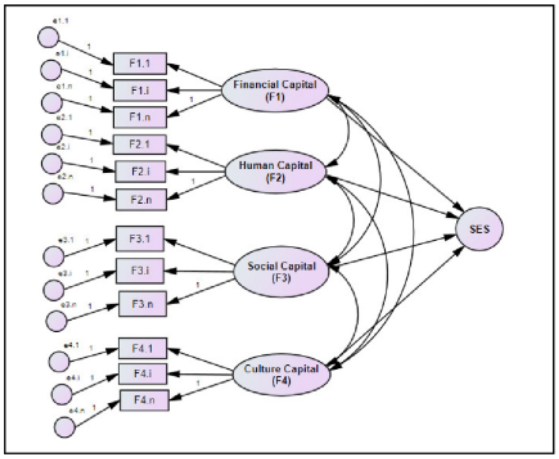

Fig. 1. Path diagram of SES Construct and Measures

C. Correlation Based Feature Selection (CFS)

Representing set of feature for build a classification model for a particular task is an important phase in machine learning. Correlation-based Feature Selection (CFS) is an algorithm that able to handling evaluation formula with an appropriate correlation measure and a heuristic search strategy[14]. CFS is a simple filter algorithm that ranks subsets according to a correlation based heuristic evaluation function. The bias of the evaluation function is toward subset that contain features that are highly correlated with the class and uncorrelated with each **other. CFS's feature subset evaluation function can be seen on** Eq. 1.

$$M_S = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}} \tag{1}$$

Where M_S is the heuristic of feature subset S containing k feature, $\overline{r_{cf}}$ is the mean feature-class correlation$(f \in S)$, and $\overline{r_{ff}}$ is the average feature-feature inter-correlation. Eq. 1 can be thought of as providing an indication of how predictive of the class a set of feature are; the denominator of how much redundancy there is among the features. CFS assumes that features are conditionally independent given the class. CFS treats feature uniformly by discretizing all continuous feature in the training data at outset.

III. CLASSIFICATION

Support Vector machine (SVM) is classification algorithm with a great mathematically concept. SVM is proposed by Vapnick using basic concept Maximal Margin Classifier[15]. It is simple to understand the basic ideas behind more sophisticated SVMs. Consider a linearly separable dataset $\{(X_i, d_i)\}$, where X_i is the input pattern for the i:th example and d_i is the corresponding desired output {-1, 1}. The assumption, *the dataset is linearly separable*, means that there exists a hyper plane working as the decision surface. We can write:

978-1-5386-3172-0/17/$31.00© 2017 IEEE

$$W^TX_i + b \geq 0, then d_i = +1$$
$$W^TX_i + b \leq 0, then d_i = -1 \qquad (2)$$

where $W^TX_i + b$, is the output function. The distance from the hyper plane to the closest point is called the geometric margin. The idea is, to have a good machine, so the geometric margin needs to be maximized. First, we introduce the marginal function $W^TX_i + b$ because the dataset is linearly separable we can rewrite as (3), as follow:

$$W^TX_i + b = +1$$
$$W^TX_i + b = -1 \qquad (3)$$

where $X^+(X^-)$ is the closest data point on the positive (negative) side of the hyperplane. Now it is straight forward to compute the geometric margin.

$$\gamma = \frac{1}{2}\left(\frac{W^TX^+ + b}{|w|} - \frac{W^TX^- + b}{|w|}\right)$$
$$= \frac{1}{2|w|})W^TX^+ + b - W^TX^- - b)$$
$$= \frac{1}{2|w|}(1-(-1)) = \frac{1}{|w|} \qquad (4)$$

Hence, equivalent to maximize the geometric margin is fixing the functional margin to one and minimizing the norm of the weight vector |w|. This can be formulated as a quadratic problem with inequality constraints $d(w^Tx_i + b) \geq 1$.

$$min: \frac{1}{2}W^TW \ (quadratic - problem) \qquad (5)$$

subject to: $d(w^Tx_i + b) \geq 1$

By the use of Lagrange multipliers$\alpha_i \geq 0$ the original problem is transformed into the dual problem. From the Kuhan–Tuker theory we have the following condition:

$$\alpha_i[d_i(W^Tx_i + b) - 1] = 0 \qquad (6)$$

It means that only the points with functional margin unity contribute to the output function. These points are called the Support Vectors, which support the separating hyper plane. In non-linear classification problem, SVM was developing by using Mercer theorem that commonly knowns as Kernel Trick.

IV. RESULT AND DISCUSS

A. DATASET

The data is collected since 2016 to 2017 in registration process of new student in a public university. The student must fill the form provided in registration online system. In order to understand the data, the Table I below will describe the table fields.

TABLE I. LATEN AND INDICATOR VARIABLES IN STUDENT DATABASE

Latent Variable		Indicator Variable		Measurement Scale
ξ_1	Financial Capital Resources (FCR)	x_1	**Mother's** employement	Ordinal
		x_2	**Father's employement**	Ordinal
		x_3	**Mother's salary**	Ordinal
		x_4	**Father's salary**	Ordinal
		x_5	**Mother's other income**	Ordinal
		x_6	**Father's other income**	Ordinal
		x_7	Number of dependent	Ordinal
		x_8	House tenure	Ordinal
		x_9	Electricity Power	Ordinal
		x_{10}	Land Size	Ordinal
		x_{11}	House Size	Ordinal
		x_{12}	Landhouse Tax Value	Ordinal
		x_{13}	Roof Material	Nominal
		x_{14}	Floor Material	Nominal
		x_{15}	Wall Material	Nominal
		x_{16}	Wall Condition	Ordinal
		x_{17}	Livingroom Condition	Ordinal
		x_{18}	Roof Condition	Ordinal
		x_{19}	Bathroom Condition	Ordinal
		x_{20}	Kithcen Condition	Ordinal
		x_{21}	Guestroom Condition	Ordinal
		x_{22}	Family room Condition	Ordinal
		x_{23}	Bedroom Condition	Ordinal
		x_{24}	Balcony Condition	Ordinal
		x_{25}	Has Bathroom	Nominal
		x_{26}	Has Washing Area	Nominal
		x_{27}	Has Toilet	Nominal
		x_{28}	Water Bill	Ordinal
		x_{29}	Electricity Bill	Ordinal
		x_{30}	Phone Bill	Ordinal
		x_{31}	Internet Bill	Ordinal
		x_{23}	Number of People at Home	Scale
		x_{33}	Motor Tenure	Ordinal
		x_{34}	Car Tenure	Ordinal
		x_{35}	Chilren are Schooling	Scale
ξ_2	Human Capital Resources (HCR)	x_{36}	**Mother's education**	Ordinal
		x_{37}	**Father's education**	Ordinal
ξ_3	Social Capital Resource (SCR)	x_{38}	Is Father Alive	Nominal
		x_{39}	**Father's** Relationship	Nominal
		x_{40}	Is Mother Alive	Nominal
ξ_4	Culture Capital Resource (CCR)	x_{41}	Distance from City	Ordinal
		x_{42}	Source of Water	Nominal
		x_{43}	Source of Electricity	Nominal

B. EXPERIMENTAL RESULT

Experiment was run using WEKA. Dataset consist of 44 feature that classify into 6 classes, K1, K2, K3, K4, K5, and K6. Using CFS 43 feature have been selecting, the selected feature is **father's** employment (x_2), **mother's** salary (x_3), **father's** salary(x_4), house tenure (x_8), electricity power (x_9), house size(x_{11}), land house tax value (x_{12}), kitchen condition (x_{20}), electricity bill (x_{29}), internet bill(x_{30}), motor tenure (x_{33}), car tenure (x_{34}), **mother's** education (x_{36}). Selected feature will be compare to full feature for classifying **Student's** Single Tuition Fee Category. Dataset is imbalanced data with ratio of imbalanced class 1:50. The classification result can be seen on Table 2.

978-1-5386-3172-0/17/$31.00© 2017 IEEE

TABLE II. CONFUSSION MATRIX OF SINGLE TUITION FEE CATEGORY USING SVM BEFORE CFS

confusion matrix		Classified as					
		a	b	c	d	e	f
real classes	a = K3	565	225	77	0	4	0
	b = K4	271	2124	6	186	1	0
	c = K2	251	10	138	1	42	0
	d = K5	0	255	0	1734	0	0
	e = K1	6	1	46	0	63	0
	f = K6	0	0	0	50	0	0

TABLE III. CONFUSSION MATRIX OF SINGLE TUITION FEE CATEGORY USING SVM AFTER CFS

confusion matrix		Classified as					
		a	b	c	d	e	f
real classes	a = K3	562	209	99	0	1	0
	b = K4	161	2281	0	146	0	0
	c = K2	184	11	232	0	15	0
	d = K5	0	180	0	1802	0	7
	e = K1	4	1	58	0	53	0
	f = K6	0	0	0	27	0	23

Base on Table 2, we can see after feature selection, the minor classes still able to classify with a good performance than before feature selection processing. The accuracy of classification before selecting the feature using CFS is 66.52%, and the accuracy of classification after selecting the feature using CFS is 81.78%. Detail of performance classification student's single tuition fee category as preliminary for develop automation system for determining student's single tuition fee can be seen on Table 4.

TABLE IV. EVALUATION MEASURE OF CLASSIFICATION USING CFS AND SVM

Class	Recall	F-Measure	ROC Area
K3	0.645	0.631	0.904
K4	0.881	0.866	0.903
K2	0.525	0.558	0.947
K5	0.906	0.909	0.963
K1	0.457	0.573	0.985
K6	0.46	0.575	0.984
Weighted Avg.	0.818	0.816	0.928

ROC of classification in table 4 shown good performances, but global accuracy need to be increased. Class K2, K3, and K4 is non-linear separable, and very difficult to classify see on Table 3. This is why the global accuracy needs to be improved. Scatter plot of mother salary and father salary related to category of student's single tuition fee and be seen on Fig.2.

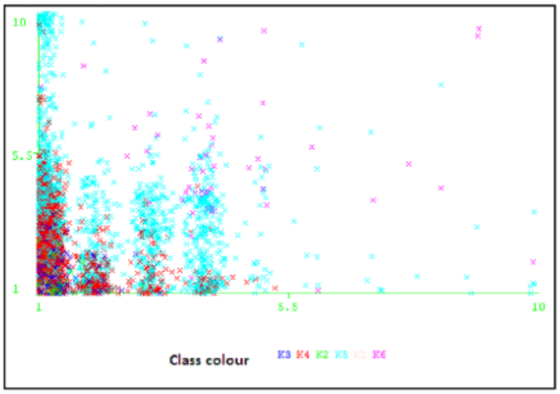

Fig. 2. Scatter plot of mother salary and father salary

V. CONCLUTION

The results of this study have shown that the use of Correlation Based Feature Selection (CFS) for selecting the best feature has been able to improve classification performances. The accuracy of classification before using CFS is 66.52%, then after selecting feature using CFS increase to 81.78%. ROC Area of classification using combination of CFS and SVM is very good, the ROC area of classification is 92.8%, and this value is equally in every class. Class K2, K3, and K4 is very difficult to separating, so to handling this problems we need to improve our methods.

REFERENCES

[1] Ariyady Kurniawan Muchsin, Made Sudarma," Penerapan Fuzzy C-Means Untuk Penentuan Besar Uang Kuliah Tunggal Mahasiswa Baru", Bali, Jurnal Lontar Komputer, Vol. 6, no.3, Desember, ISSN: 2088-1541, DOI: 10.24843/LKJITI.6.3.16975, 2015

[2] Cavdar, S. C., & Aydin, A. D.,"An Experimantal Study on Relationship between Student Socio-Economic Profile, Financial Literacy, Student Satisfaction and Innovation within the Framework of TQM",Procedia-Social and Behavioral Sciences, 195, 739-748,2015

[3] Ensminger ME, Forrest CB, Riley AW, Kang M, Green BF, Starfield B, Ryan SA,"The validity of measures of socioeconomic status of adolescents", Journal of Adolescent Research, May;15(3):392-419, 2000

[4] Gegel, L., Lebedeva, I., & Frolova, Y,"Social Inequality in Modern Higher Education. Procedia-Social and Behavioral Sciences" 214, 368-374,2015

[5] Gusmelia Testiana, Rachmansyah," Pemanfaatan Metode Simple Additive Weighting (SAW) untuk Penentuan Penerima UKT Kelompok 1", Seminar Nasional Teknologi Informasi, Komunikasi dan Industri (SNTIKI) 9, Pekanbaru, 18-19 Mei, ISSN (Online) : 2579-5406,2017

[6] Jerrim, J., Chmielewski, A. K., & Parker, P. ,"Socioeconomic inequality in access to high-status colleges: A cross-country comparison",Research in Social Stratification and Mobility, 42, 20-32,2015

[7] Leigh, A., Jencks, C., & Smeeding, T. M. " Health and economic inequality", The Oxford Handbook of Economic Inequality, Oxford University Press, Oxford, 384-405,2009

[8] Regulation of the minister of research, technology and high education of the republic of Indonesia number 39 year 2017 about the single tuition fee for public universities,Jakarta, May 2017.

[9] U.S. Department of Education, Improving the Measurement of Socioeconomic Status for the National Assessment of Educational Progress:A Theoritical Foundation , Recommendations to the National Center for Education Statistics, Institute of Education Sciences, National Center for Education Statistics, Novermber 2012, retrieved from https://nces.ed.gov/nationsreportcard/pdf/researchcenter/Socioeconomic_Factors.pdf

[10] White, K. R.," The relation between socioeconomic status and academic achievement", Psychological bulletin, 91(3), 461,1982

[11] E. Pitt and R. Nayaki, "The Use of Various Data Mining and Feature Selection Methods in the Analysis of a Population Survey Dataset," in *Proceedings 2nd International Workshop on Integrating Artificial Intelligence and Data Mining (AIDM 2007) CRPIT*, 2007, pp. 87–97.

[12] M. S. C. Thomas, N. a Forrester, and A. Ronald, "Modeling socioeconomic status effects on language development.," *Dev. Psychol.*, vol. 49, no. 12, pp. 2325–43, 2013.

[13] X. Zhang, K. Tocque, J. Boothby, P. Cook, and M. Li, "Exploration of Relationship between Social Status and Mortality Rates in England," *Neural Networks*, 2008.

[14] M. Hall, "Correlation-based Feature Selection for Machine Learning," *Methodology*, vol. 21i195-i20, no. April, pp. 1–5, 1999.

[15] E. M. Imah, F. Al Afif, M. Ivan Fanany, W. Jatmiko, and T. Basaruddin, "A comparative study on Daubechies Wavelet Transformation, Kernel PCA and PCA as feature extractors for arrhythmia detection using SVM," in *IEEE Region 10 Annual International Conference, Proceedings/TENCON*, 2011, pp. 5–9.

978-1-5386-3172-0/17/$31.00© 2017 IEEE

Determining student's single tuition fee category using correlation based feature selection and support vector machine

ORIGINALITY REPORT

19%	17%	14%	8%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Hamid Khorrami, Majid Moavenian. "A comparative study of DWT, CWT and DCT transformations in ECG arrhythmias classification", Expert Systems with Applications, 2010 Publication	3%
2	Khorrami, H.. "A comparative study of DWT, CWT and DCT transformations in ECG arrhythmias classification", Expert Systems With Applications, 201008 Publication	3%
3	www.researchgate.net Internet Source	3%
4	Submitted to School of Business and Management ITB Student Paper	3%
5	Malika-Djahida Kedir-Talha, Najissa Sadi-Ahmed, Mohamed Amine Ait Amer. "The lifted wavelet transform for encephalic signal	2%

diagnostic", 2014 6th International Conference of Soft Computing and Pattern Recognition (SoCPaR), 2014
Publication

6 H Syahputra, Sutrisno, S Gultom. "Decision Support System For Determining The Single Tuition Group (UKT) In State University Of Medan Using Fuzzy C-Means", Journal of Physics: Conference Series, 2020
Publication
2%

7 Bing Niu, Yuhuan Jin, WenCong Lu, GuoZheng Li. "Predicting toxic action mechanisms of phenols using AdaBoost Learner", Chemometrics and Intelligent Laboratory Systems, 2009
Publication
2%

8 www.scms.waikato.ac.nz
Internet Source
2%

9 Submitted to Higher Education Commission Pakistan
Student Paper
2%

Exclude quotes	Off	Exclude matches	< 2%
Exclude bibliography	On		