Dominant Criteria and Its Factor Affecting Student Achievement Based on Rough-Regression Model

Riswan Efendi
Faculty of Science and Technology, Universitas Islam Negeri Sultan Syarif Kasim Riau,
28293 Pekanbaru, Indonesia
riswan.efendi@uin-suska.ac.id

Novi Yanti
Faculty of Science and Technology, Universitas Islam Negeri Sultan Syarif Kasim Riau,
28293 Pekanbaru, Indonesia
novi_yanti@uin-suska.ac.id

Alex Wenda
Faculty of Economics and Sosial, Universitas Islam Negeri Sultan Syarif Kasim Riau,
28293 Pekanbaru, Indonesia
alexwenda@uin-suska.ac.id

Susnaningsih Mu'at
Faculty of Economics and Sosial, Universitas Islam Negeri Sultan Syarif Kasim Riau,
28293 Pekanbaru, Indonesia
susnaningsih@uin-suska.ac.id

Noor Azah Samsudin
Faculty of Science Computer and Information Technolgy, Universiti Tun Hussein Onn Malaysia
86400 Batu Pahat, Malaysia
azah@uthm.edu.my

Mustafa Mat Deris
Faculty of Science Computer and Information Technolgy, Universiti Tun Hussein Onn Malaysia
86400 Batu Pahat, Malaysia
mmustafa@uthm.edu.my

***Abstract*— the ordinary least square model has been widely considered to estimate the significant factors which influence the student achievement. Some factor is qualitative type and measured using criteria or categories. However, the decisive criteria for each factor which affect to the cumulative grade point average of student cannot be determined by this model. In this paper, we are interested to build a new procedure using rough-regression model in determining the dominant criteria from each factor based on generalization of dependency attribute. Based on result, the proposed procedure is capable to investigate the dominant criteria and factors affecting student achievement, such as, language spoken with dominant criteria is "many-many", FB friend with dominant criteria is "many" and fast food with dominant criteria is "never". This proposed procedure is very appropriate to implement for handling categorical data.**

Keywords—rough-regression, dominant criteria, FB friend, fast food, language spoken, CGPA.

I. INTRODUCTION

In education and psychology applications, regression models have been widely applied to investigate the inter-relationships between explanatory and response factors (attributes) which influence the academic performance of the universities students. For example, there was a negative association between times spent on Facebook and GPA (grade point average) for freshmen, sophomores, and juniors [3, 4]. Moreover, the increasing of cell phones use was associated with decreasing of academic performance for college students in U.S [17]. However, the implementation of rough sets into regression model is still limited to investigate the education and psychology attributes. Additionally, the attributes from both domains are categorical values.

Moreover, there was an association between language proficiency and multilingualism to academic performance, especially international students by using statistical approach [1, 2]. Based on [5-7], the high academic performance have been achieved significantly when the student never consumed fast food using adjusted odds ratio. Motivated by three different factors, namely, number of FB friend, number of language spoken, and fast food per week, we are interested to determine the dominant criteria from each factor which affect to academic performance of university students using rough sets and regression models.

Recently, rough-regression model has been introduced in medical applications, especially in prediction cholesterol level, flu diagnosis and cancer diagnosis [13-15]. However, the discussion between rough set and regression models has been presented separately in the previous studies for education domains. In this paper, we discuss the generalization of dependency attribute of rough sets in determining dominant criteria for each significant factor in regression model. Additionally, the proposed procedure will be implemented to calculate and determine the decisive criteria for student CGPA.

II. ROUGH SET AND REGRESSION MODELS

A. Rough Sets and Applications

Pawlak has been initiated in 1982 the rough sets theory for uncertainty and categorical attributes analysis [8], there are some components in this theory related with information system $S = (U, \Omega, V_q, f_q)$ [8-10].

A students, patients and observations are called as objects in rough sets. While, the factors, variables, and characteristic information are denoted as attributes. A decision table is addressed for information systems which organized using the row and column correspond to objects and attributes, respectively [8-10]. Based on [10], there are some parameter should be determined in the rough sets model, such as, the indiscernibility, lower-upper approximations, and boundary region as expressed in Equations (1) – (5) accordingly.

$$IND(M) = \{(x, y) \in U^2 : \forall a \in M, a(x) = a(y)\} \quad (1)$$

$$\underline{M}(X) = \left\{\{x \in U | [x]_B \subseteq X\}\right\}, \quad (2)$$

$$\bar{M}(X) = \{x \in U | [x]_B \cap X \neq \emptyset\}, \quad (3)$$

$$\mathrm{BND}(X) = \underline{M}(X) - \bar{M}(X), \quad (4)$$

$$b(X) = \frac{\underline{M}(X)}{\bar{M}(X)}, \quad (5)$$

while, the association value among variables or factors is written as [11]:

$$l = \frac{\sum_{x \in U/D} |\underline{C}(X)|}{|U|}; C, D \subseteq A \wedge C \cap D = \emptyset. \quad (6)$$

978-1-5386-7440-6/18/$31.00 ©2018 IEEE

The highest value of l is called as a decisive factor. Majority, rough sets theories can be implemented to investigate the dependency between conditional attributes and decision attribute, rules development for decision making for various domains, such as, cognitive sciences and medical data.

B. Regression Models and Its Applications

Based on [15], a simple ordinary least square model is written as:

$$Z = a + bQ + e, \tag{7}$$

where Z is a response factor, Q is an explanatory factor, a and b are constant and coefficient regression, while e is residual. The steps in constructing Eq. (7) are detailed as follows [12]:

- Evaluate correlation value between Z and F.
- Calculate constant and coefficient.
- Evaluate the validity of F and Z using R-square value and ANOVA and F-test.
- Evaluate the significant constant and coefficient model using t-test.
- Evaluate the residual model using normal test.

In applications, regression models can be used to measure the causal effect between independent and dependent factors in various domains, such as, sciences, medical, education, economics, others.

III. PROPOSED DOMINANT CRITERIA USING ROUGH-REGRESSION MODEL

Based on Eq. (7), the regression model for investigating factors which affect to the CGPA student can be generalized as follows [12]:

$$CGPA(Y) = k_0 + k_1P_1 + k_2P_2 + \cdots + k_nP_n + \tag{8}$$

From Eq. (8), $P_1, \ldots, P_n$ are explanatory factors which affect to response factor, Y. On the other hands, $k_0, \ldots, k_n$ are coefficients. In this section, our goal is to estimate the coefficients and the significant explanatory factors based on steps given in Section IIB. In this case, we assumed that all factors affect to the student CGPA. By using generalization of dependency attribute in Eq. (6), the decisive criteria for each factor can be derived as follows:

Step 1: Transform numerical values of explanatory and response factors into criteria as presented in Table I.

TABLE I. CATEGORICAL VALUE OF FACTORS

ID	P_1	P_2	⋯	P_n	$CGPA$
S_1	Very Low	Very Low	⋯	Very small	SC Lower
S_2	Very Low	Average	⋯	Very small	First class
S_3	Very low	Very Low	⋯	Moderate	SC Upper
S_4	Many	Very low	⋯	High	First class
⋯	⋯	⋯	⋯	⋯	⋯
S_n	Many many	Never	⋯	Very high	First class

Step 2: Organize each factor by following criteria into sets as follows:

For P_1,

$$U/P_1(\text{Criteria}_1) = \{\forall R_i \text{ with the first criteria}\},$$
$$\ldots$$
$$U/P_1(\text{Criteria}_p) = \{\forall R_i \text{ with the } p \text{ criteria}\},$$

For P_2,

$$U/P_2(\text{Criteria}_1) = \{\forall R_i \text{ with the first criteria}\},$$
$$\ldots$$
$$U/P_2(\text{Criteria}_p) = \{\forall R_i \text{ with the } p \text{ criteria}\},$$
$$\ldots$$

For P_n,

$$U/P_n(\text{Criteria}_1) = \{\forall R_i \text{ with the first criteria}\},$$
$$\ldots$$
$$U/P_n(\text{Criteria}_p) = \{\forall R_i \text{ with the } p \text{ criteria}\},$$

Step 3: Generalize decisive criteria (DC) by using Eq. (6) for each factor as follows:

For P_1,

$$\text{Decisive Criteria}_1 \rightarrow k_1^Y = \frac{U/X_1(\text{Criteria}_1)}{U/Y},$$
$$\ldots$$
$$\text{Decisive Criteria}_p \rightarrow k_p^Y = \frac{U/X_1(\text{Criteria}_p)}{U/Y}.$$

For P_2,

$$\text{Decisive Criteria}_1 \rightarrow k_1^Y = \frac{U/X_2(\text{Criteria}_1)}{U/Y},$$
$$\ldots$$
$$\text{Decisive Criteria}_p \rightarrow k_p^Y = \frac{U/X_2(\text{Criteria}_p)}{U/Y}.$$

$$\ldots$$

For P_n,

$$\text{Decisive Criteria}_1 \rightarrow k_1^Y = \frac{U/X_n(\text{Criteria}_1)}{U/Y},$$
$$\ldots$$
$$\text{Decisive Criteria}_p \rightarrow k_p^Y = \frac{U/X_n(\text{Criteria}_p)}{U/Y}.$$

Step 4: Determine maximum value from each factor based on Step 3.

IV. IMPLEMENTATION

We collected row data from 56 students Multimedia Department, UTHM. The data collection included number of language spoken (P_1), number of Facebook friend (P_2), number of fast food a week (P_3) and cumulative grade point average (CGPA) (Y). By using this data set, we obtained the mathematical model as follows:

$$CGPA = 3.51 + 0.0879LS - 0.00071FB - 0.0515Food \tag{9}$$

Based on coefficient values in Eq. (9), the coefficient of language spoken is highest among the factors above, so that this factor gave the significant impact to CGPA if compared with Facebook friend and fast food factors. On the other hands, we want to find out which criteria (condition) from language spoken, Facebook friend and fast food affect to the student CGP decisively. By applying proposed steps previous section, in the dominant criteria are derived as follows:

Step 1: Transform numerical values of explanatory and response factors into criteria as presented in Table II.

TABLE II. CATEGORICAL VARIABLES

ID	P_1	P_2	P_3	$CGPA\ (Y)$
s_1	Low	Very low	Very rarely	First class
s_2	Moderate	Moderate	Very rarely	First class
s_3	Low	Very low	Rarely	SC upper
s_4	Low	Low	Rarely	SC upper
…	…	…	…	…
s_{56}	Moderate	Very low	Very rarely	First class

Step 2: Define sets of CGPA based on its criteria as follows:

U/Y = {First class, Second class upper, Second class lower},

$$U/Y = \{\{s_{1,}s_{2,}s_{5,\ldots,}s_{56}\},\ \{s_{3,}s_{4,}s_{6,\ldots,}s_{49}\},\ \{s_{11,}s_{40}\}\} \quad (10)$$

Step 3: Define sets of independent factors and its criteria as follows:

For P_1,

$U/(\text{Very low}) = \{s_{21}, s_{29}, s_{47}\}$,
$U/(\text{Low}) = \{s_1, s_3, s_4, \ldots, s_{51}\}$,
$U/(\text{Moderate}) = \{s_2, s_5, s_6, \ldots, s_{56}\}$,

…

$U/(\text{MMMM}) = \{s_{10}\}$.

For P_2,

$U/(\text{Very low}) = \{s_1, s_3, s_6, \ldots, s_{56}\}$,
$U/(\text{Low}) = \{s_4, s_5, s_7, \ldots, s_{54}\}$,
$U/(\text{Moderate}) = \{s_2, s_8, s_{43}\}$,
…
$U/(\text{Too many}) = \{s_{20,}s_{28}, s_{36}, s_{40}\}$.

For P_3,

$U/(\text{Very low}) = \{s_1, s_3, s_6, \ldots, s_{56}\}$,
$U/(\text{Low}) = \{s_4, s_5, s_7, \ldots, s_{54}\}$,
$U/(\text{Moderate}) = \{s_2, s_8, s_{43}\}$,
…
$U/(\text{Too many}) = \{s_{20,}s_{28}, s_{36}, s_{40}\}$.

Step 4: By using Eq. (6) and Section III, compute the dominant criteria for each factor:

For P_1,

$\text{Very low} \rightarrow k_1^Y = 0$, $\text{Low} \rightarrow k_2^Y = 0$,
$\text{Moderate} \rightarrow k_3^Y = 0$, $\text{Many} \rightarrow k_4^Y = 0$,
$\text{Many many} \rightarrow k_5^Y = \frac{3}{56}$, $\text{Many many many} \rightarrow k_6^Y = \frac{1}{56}$,
$\text{Many many many many} \rightarrow k_7^Y = \frac{1}{56}$.

For P_2,

$\text{Very low} \rightarrow k_1^Y = 0$, $\text{Low} \rightarrow k_2^Y = 0$,
$\text{Moderate} \rightarrow k_3^Y = 0$, $\text{Many} \rightarrow k_4^Y = \frac{4}{56}$,
$\text{Too many} \rightarrow k_5^Y = 0$.

For P_3,

$\text{Never} \rightarrow k_1^Y = \frac{2}{56}$, $\text{Very rarely} \rightarrow k_2^Y = 0$,
$\text{Rarely} \rightarrow k_3^Y = 0$, $\text{Sometimes} \rightarrow k_4^Y = 0$,
$\text{Often} \rightarrow k_5^Y = \frac{1}{56}$, $\text{Very Often} \rightarrow k_5^Y = 0$,
$\text{Very Very Often} \rightarrow k_6^Y = 0$.

Step 5: Determine the highest value of decisive criteria from each factor as follows:

For X_1,

Dominant criteria = $\max\left\{0,0,0,0,\frac{3}{56},\frac{1}{56},\frac{1}{56}\right\} = \frac{3}{56}$.

For X_2,

Dominant criteria = $\max\left\{0,0,0,\frac{4}{56},0\right\} = \frac{4}{56}$

For X_3,

Dominant criteria = $\max\left\{\frac{2}{56},0,0,0,\frac{1}{56},0,0\right\} = \frac{2}{56}$.

Step 6: Evaluate the decisive category and its factors which give impact to the student achievement (CGPA) as presented in Table III.

TABLE III. DEPENDENCY CRITERIA AND VARIABLES

Variables/Criteria	First class	Second class upper	Second class lower
P_1/Very low	0	0	0
P_1/Low	0	0	0
P_1/Moderate	0	0	0
P_1/Many	0	0	0
P_1/MM	**3/56**	0	0
P_1/MMM	1/56	0	0
P_1/MMMM	1/56	0	0
P_2/Very low	0	0	0
P_2/Low	0	0	0
P_2/Moderate	0	0	0
P_2/Many	0	**4/56**	0
P_2/Too many	0	0	0
P_3/Never	**2/56**	0	0
P_3/Very rarely	0	0	0
P_3/Rarely	0	0	0
P_3/Sometimes	0	0	0
P_3/Often	0	1/56	0
P_3/Very Often	0	0	0
P_3/Very very often	0	0	0

Based on Table III, the dominant criteria of P_1 to CGPA is "many many (MM = 3/56)". It can be interpreted that first class student who able to speak in 5 different languages has highest CGPA if compared with other. This condition is also supported by previous study [1, 2]. On the other words, the student's CGPA is very depended on language spoken with "MM" criteria. For Facebook friend, the student who have a number of friends with the criteria "many (M = 4/56)" obtain a second class upper of CGPA. Sound similar with discussion in [3, 4] While, students who "never (N = 2/56)" eat fast food every week have a high CGPA if compared with student who eat fast food frequently. Thus, the fast food consumption is very significant factor to the academic performance as mentioned in [5-7]. From all dominant criteria, two criteria (P_1 /MM) and P_3 /Never) contribute to student CGPA (first class level). Other criteria

(P_2/Many) contributes to the second class upper. In this study, we applied upper approximation to calculate the dominant criteria of each factor. So that, this the main different with previous study [16].

V. CONCLUSION

In this study, a new procedure was proposed to determine the decisive criteria for each explanatory factor which contribute to the response factor. The proposed procedure has been implemented in handling the explanatory factors and its decisive category which gave impact to the academic achievement. There are three significant factors which affected to the academic performance (CGPA) for the Multimedia students, namely, number of language spoken, number of FB friend and fast food with specific criteria. These factors are also supported by previous studies [1-7].

In the previous papers, the decisive categories from each factor are not yet explained in detail, especially conventional regression models. Interestingly, the proposed procedure is generalized from dependency attribute and upper approximation of rough sets. Based on our perspective, proposed procedure is very appropriate to analyze the categorical data or variables which fully uncertainty. The merit of this proposed procedure is not very strict with statistical assumptions.

ACKNOWLEDGMENT

This paper is supported by Lembaga Pengabdian Masyarakat (LPM), Universitas Islam Negeri Sultan Syarif Kasim, Riau, Indonesia.

REFERENCES

[1] L. Amy and S. Crist, "Why bilingual boys get better grades in English-only America? The impact gender, language and family interaction on academic achievement of Latino/ a children of immigrants," Ethnic and Racial Studies, vol. 32, pp. 346-368, 2009.

[2] N. M. Martirosyan, E. Hwang and R. Wanjohi, "Impact of english profiency on academic performance of International students," Journal of International Students, vol. 5, pp. 60-75, 2015.

[3] Thuseethan and Kuhanesan, "Influence Facebook in academic perfromance of Sri Lankan university students," Global Journal of Comp. Science and Tech., Network, Web and Security, vol. 4, pp. 29-35, 2014.

[4] R. Junco "Student class standing, facebook use, and academic performance," Journal of Applied Developmental Psychology, vol. 36, pp. 18-29, 2015.

[5] F. Kobayashi, "Academic achievement, BMI, and fast food intake of American and Japanese college students," Nutrition and Food Science, vol. 39, pp. 555-566, 2009.

[6] A. Ross,"Nutrition and its effects on academic performance How can our schools improve?," Master Thesis of Art Education, Northen Michigan University, pp. 1-58, 2010.

[7] R. Aresegowda, N. A. Rani, P. Mukherjee, A. Nusrath, "Assessment of dietry trends and its impact on academic performance among young adult medical students of a tertirary care teaching hospital. Int. Journal of Medical Science, vol.5, pp. 2346-2350, 2016

[8] Z. Pawlak, "Rough sets," International Journal of Computer and Information Science, vol.11, pp. 341-356, 1982.

[9] Z. Pawlak, "Rough sets: Theoretical aspects of reasoning about data," Kluwer Academic Pubsliher, Dordrecht, 1991.

[10] F. E. H. Tay and L. Shen, Economic and financial using rough sets model," European Journal of Operation Research, vol. 141, pp. 641-659, 2002.

[11] T. Herawan, M. M. Deris, and J. H. Abawajy,"A rough set approach for selecting clustering attribute," Knowledge-Based Systems, vol. 23, pp. 220-231, 2010.

[12] J. M. Wooldridge, "Introductory Econometrics A Modern Approach," Third edition, Thomson, South Western, USA, 2006.

[13] Efendi, R., Deris, M.M.: Decision Support Model in Determining Factors and Its Dominant Criteria Affecting Cholesterol Level Based on Rough-Regression. Recent Advances on Soft Computing and Data Mining, pp. 243-251, 2018.

[14] Efendi, R., Samsudin, N.A., Deris, M.M, and Ting YG: Flu Diagnosis System Using Jaccard Index and Rough Set Approaches, Journal of Physic, Conference Series, 2018

[15] Efendi, R., Samsudin, N.A., Deris, M.M.: Medipre: Medical diagnosis prediction using rough-regression approximation. ACM Proceeding on High Compilation, Computing and Communications, pp. 35-39, 2018.

[16] Rissino, S., Torres, G. L. Rough set theory-fundamental concepts, principals, data extraction, and applications, Julio Ponce and Adam Karahoca (Ed), Data Mining and Knowledge Discovery in Real Life App. Inform, pp. 35-58, 2009.

[17] A. Lepp, J. E. Barkley, A. C. Karpinski, "The relationship between cell phone use, academic performance, anxiety, and satisfaction with life in college students," Computer in Human Behaviour, vol. 31, pp. 343-350, 2014.