2020 the third International Conference on Vocational Education and Electrical Engineering (ICVEE)

NO. 45512/UN38.1/DL.01.02/2020

UNESA

CERTIFICATE

TO:

Yuni Yamasari

as Presenter

Combining the Unsupervised Discretization Method and the Statistical Machine Learning on the Students' Performance

in 2020 the third International Conference on Vocational Education and Electrical Engineering (ICVEE). With the theme "Strengthening the Framework of Society 5.0 through Innovations in Education, Electrical Engineering, and Informatics Engineering".

Surabaya, October 3-4, 2020

Chairman of ICVEE

Prof. Dr. Bambang Suprianto, M.T.

Vice Rector I of UNESA

Prof. Dr. Bambang Yulianto, M.Pd.

IEEE Scopus ATLANTIS PRESS WEB OF SCIENCE

PROCEEDING

UNESA

IEEE
INDONESIA SECTION

ICVEE

2020 the third International Conference on Vocational Education and Electrical Engineering (ICVEE)

Strengthening the framework of Society 5.0 through Innovations in Education, Electrical, Engineering and Informatics Engineering

IEEE Catalog Number : CFP20X27-ART
ISBN : 978-1-7281-7434-1

3-4 OCTOBER 2020
UNIVERSITAS NEGERI SURABAYA

VITUAL EVENT
SURABAYA - INDONESIA

http://icvee.conference.unesa.ac.id

2020 the third International Conference on Vocational Education and Electrical Engineering (ICVEE)

Proceeding

2020 the third International Conference on Vocational Education and Elecrical Engineering (ICVEE) on October 3-4, 2020 in the virtual event, Surabaya, Indonesia.

IEEE catalog number : CFP20X27-ART
ISBN : 978-1-7281-7434-1

Conference Record Number 50212

Message from the General Chair

It gives me great pleasure to all of the keynote/invite speakers, distinguished guests, and ICVEE participants, welcome to 2020 the third International Conference on Vocational Education and Electrical Engineering (ICVEE). Due to the COVID-19 ICVEE conference which is organized by the Department of Electrical Engineering and Departement of Informatics, Universitas Negeri Surabaya and technical sponsorship IEEE Indonesia section hold the conference in the virtual event. The theme of our conference is " Strengthening the framework of Society 5.0 through Innovations in Education, Electrical Engineering, and Informatics Engineering".

This year we receive 134 articles and resulted in 71 articles that have been presented at this conference and will be submitted to the IEEE explorer. The article comes from some domestics and international universities. The International author and co-author come from Brazil, Jerman, Philippines, Japan, Taiwan, Singapore, Malaysia, Thailand, Saudi Arabia, and Australia. We would like to appreciate all of the keynotes and invite speakers, reviewers, committees, and participants for all the support and participation. We would like to give gratitude to the Universitas Negeri Surabaya as the organizer and IEEE Indonesian Section as a technical Co-sponsorship.

Finally, I wish all participants always successful and enjoy this conference. I hope this program will be interesting and useful for all the ICVEE participants.

Prof. Dr. Bambang Suprianto., MT

General Chair

Organizer and committee

2020 the third International Conference on Vocational Education and Electrical Engineering (ICVEE)

Organized by

Technical Co-Sponsorship

Committee:

General Chair

- Prof. Dr. Bambang Suprianto, M.T (Universitas Negeri Surabaya)

General Co-Chair

- Dr. Nurhayati, S.T., M.T. (Universitas Negeri Surabaya)

TPC

- Dr. Elly Matul Immah, M.Kom. (Universitas Negeri Surabaya)
- Setya Chendra Wibawa, M.Kom. (Universitas Negeri Surabaya)

Publication Chair

- Salamun Rohman Nudin, S.Kom., M.Kom. (Universitas Negeri Surabaya)

Advisory Board

- Prof. Takeshi Fukusako, Kumamoto University (Japan)
- Prof. João Fransisco Justo (Polytechnic School of the University of São Paulo, São Paulo, Brasil)
- Prof. Nobuo Funabiki (Okayama University, Japan)
- Prof. Poki chen (National Taiwan University, Taiwan)
- Prof. Gamantyo Hendrantoro (Institut Teknologi Sepuluh Nopember, Indonesia)
- Prof. Alexandre Maniçoba De Oliveira (IFSP, Brazil)
- Prof. Sven Shculte (Technische Universität Dortmund, German)
- Prof. Dr. H. Bambang Yulianto., M.Pd (Universitas Negeri Surabaya, Indonesia)
- Prof. Dr. Supari Muslim , M.Pd (Universitas Negeri Surabaya, Indonesia)
- Prof . Dr. Ismet Basuki, M.Pd (Universitas Negeri Surabaya, Indonesia)
- Prof. Munoto, M.Pd (Universitas Negeri Surabaya, Indonesia)
- Prof. Ekohariadi, M.Pd (Universitas Negeri Surabaya, Indonesia)
- Spits Warnars Harco, S.Kom., M.T.I., Ph.D. (Binus University, Indonesia)

Section Chair

- Dr. Kurnianingsih (Vice-Chair of the IEEE Indonesia Section)

Technical Program Committee/reviewer

- Eko Setijadi, MT.,Ph.D – Institut Teknologi Sepuluh Nopember, Indonesia
- Shintami Hidajati, PhD – Institut Teknologi Sepuluh Nopember, Indonesia
- Jiapan Guo., MSc.Ph.D – University of Groningen, Netherlands
- Prof. Alexandre Maniçoba De Oliveira – Instituto Federal de São Paulo, IFSP, Brazil
- Warangkana Chaihongsa, D.Eng– King Mongkut's Institute of Technology Langkrabang, Thailand
- Dr. Umi Laili Yuhana., M.Kom – Institut Teknologi Sepuluh Nopember, Indonesia
- Dr. Tri Budi Santoso., ST., MT. – Politeknik Elektronika Negeri Surabaya, Indonesia
- Dr. Rr. Ani Dijah Rahajoe., ST.,M.Cs – Universitas Bhayangkara Surabaya
- Dr. Amirullah., ST., MT. – Universitas Bhayangkara Surabaya
- Ir. Wijono, Ph.D – Universitas Brawijaya, Indonesia
- Dr. Eng. Adi Wibowo, S.Si., M.Kom – Universitas Diponegoro, Indonesia
- Krismadinata, Ph. D – Universitas Negeri Padang, Indonesia
- Dr. Diana Purwitasari – Institut Teknologi Sepuluh Nopember, Indonesia
- Royyana Muslim Ijtihadie., Ph.D– Institut Teknologi Sepuluh Nopember, Indonesia
- Dr. Eng. Asep Bayu Dani Nandiyanto – Universitas Pendidikan Indonesia, Indonesia
- Didin Wahyudin, Ph.D – Universitas Pendidikan Indonesia, Indonesia
- Didik Nurhadi., PhD – Universitas Negeri Malang, Indonesia
- Iwan Kustiawan, Ph. D – Universitas Pendidikan Indonesia, Indonesia
- Prof. Moh. Khairuddin, Ph.D – Universitas Negeri Yogjakarta, Indonesia
- Dr. Ade Gafar Abdullah – Universitas Pendidikan Indonesia, Indonesia
- Dr. Nyoman Gunantara., MT– Universitas Udayana, Indonesia
- Dr. Umaisaroh., MT – Universitas Mercu Buana, Jakarta, Indonesia
- Dr. Indah Kurniawati., MT– Universitas Muhammadiyah Surabaya, Indonesia
- Dr. Farah Afianti., M. Kom– Telkom University, Indonesia
- Dr. Mike Yuliana – Politeknik Elektronika Negeri Surabaya, Indonesia
- Dr. Syahfrizal tahcfullah., MT– Universitas Borneo Tarakan, Indonesia
- Teuku Muhammad Roffi., Ph.D – Universitas Pertamina, Indonesia
- Dr. Hakkun Elmunsyah– Universitas Negeri Malang, Indonesia
- Dr. Verry Ronny Palinglingan– Universitas Negeri Manado, Indonesia
- Dr. Ing. Parabelem Tinno dolf Rompas., M.Eng – Universitas Negeri Manado, Indonesia
- Dr. Lala Septem Riza., MT – Universitas Pendidikan Indonesia, Indonesia
- Dr. Endi Suhendi., M.Si. – Universitas Pendidikan Indonesia, Indonesia
- Ilmi Jazuli Ichsan., M.Pd – Universitas Negeri Jakarta, Indonesia
- Eppy Yundra, Ph. D – Universitas Negeri Surabaya, Indonesia
- Dr. Lilik Anifah, MT – Universitas Negeri Surabaya, Indonesia
- Dr. IGP Asto Buditjahjanto, MT – Universitas Negeri Surabaya, Indonesia
- Dr. Nurhayati, ST., MT. – Universitas Negeri Surabaya, Indonesia
- Dr. Lusia Rakhmawati, M.T. – Universitas Negeri Surabaya, Indonesia
- Dr. Yuni Yamasari., M.Kom– Universitas Negeri Surabaya, Indonesia
- Unit Three Karini., Ph.D– Universitas Negeri Surabaya, Indonesia
- Hapsari Peni Tjahyaningtijas., MT– Universitas Negeri Surabaya, Indonesia
- Utama Alan Deta., M.Pd– Universitas Negeri Surabaya, Indonesia
- Naim Rochmawati., M.Kom – Universitas Negeri Surabaya, Indonesia
- Dr. Ricky Eka Putra, S.Kom., M.Kom– Universitas Negeri Surabaya, Indonesia

Organizing Committee

- Dr. Maspiyah, M.Kes.
- Dr. Edy Sulistyo., M.Pd
- Dr. Agus Wiyono.
- Dedy Rahman Prehanto, S.Kom., M.Kom
- Prof. Dr. Munoto, M.Pd.

- Ir. Achmad Imam Agung, M.Pd.
- I Kadek Dwi Nuryana, S.T.,M.Kom.
- Prof. Dr. Bambang Suprianto, M.T.
- Dr. Nurhayati, S.T., M.T.
- Aries Dwi Indriyanti, S.Kom., M.Kom
- Dr. Euis Ismayati.,M.Pd.
- Yeni Anistyasari, S.Pd.,M.Kom
- Dr. Ricky Eka Putra, S.Kom., M.Kom.
- I Gusti Lanang Eka Putra., S.Kom., M.Kom.
- Nur Kholis, S.T., M.T.
- Eppy Yundra, P.hD
- Unit Three Kartini., Ph.D
- Aditya Chandra H., ST., MT.
- Farid Baskoro, ST., MT.
- Dr. I G.P. Asto B., M.T.
- Dr. Lilik Anifah.,S.T.,M.T
- Dr. Meini Sondang, M.Pd.
- Dr. Lusia Rakhmawati, MT
- Hapsari Peni Tjahjaningtijas.,MT
- Setya Chendra Wibawa., S.Pd., M.Kom
- Dodik Arwin Darmawan., S.T., M.T.
- Rifqi Firmansyah, S.T., M.T.
- Arif Widodo, S.T., M.Sc.
- Naim Rohmawati, M.Kom
- Yuli Sutoto, S.Pd, M.Pd.
- Salamun Rohman Nudin, S.Kom.,M.Kom.
- Rahardian Bisma, S.Kom.,M.Kom
- Mahendra Widyartono, ST., MT
- Widi Ariwibowo, S.T., M.T.
- Dr. Joko, S.Pd., M.Pd.
- Drs. Bambang Sujatmiko.,M.T.
- Reza Rahmadian, S.ST., MengSc
- Miftahur Rohman, S.T.,M.T
- Fendi Ahmad, S.Pd.,M.Pd.
- L. Endah Cahya Ningrum, S.Pd.,M.Pd.
- Rindu P, S.Kom., M.Kom.
- Martini Dwi Endah Susanti, S.Kom., M.Kom
- Paramitha Nerisafitra, S.ST., M.Kom.
- Syarifuddin Zuhri., S.Pd.,M.T.
- Subuh Isnur Haryudo, S.T., M.T.
- Marisa, S.E

Table of Contents

Combining the Unsupervised Discretization Method and the Statistical Machine Learning on the Students' Performance

Yuni Yamasari
Department of Informatics
Universitas Negeri Surabaya
Surabaya, Indonesia
yuniyamasari@unesa.ac.id

Anita Qoiriah
Department of Informatics
Universitas Negeri Surabaya
Surabaya, Indonesia
anitaqoiriah@unesa.ac.id

Naim Rochmawati
Department of Informatics
Universitas Negeri Surabaya
Surabaya, Indonesia
naimrochmawati@unesa.ac.id

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

Hapsari P. A. Tjahyaningtijas
Department of Electrical Engineering
Universitas Negeri Surabaya
Surabaya, Indonesia
hapsaripeni@unesa.ac.id

Puput W. Rusimamto
Department of Electrical Engineering
Universitas Negeri Surabaya
Surabaya, Indonesia
puputwanarti@unesa.ac.id

***Abstract*— the suitability of the data with the method in the process of data mining is very important to increase the process performance. However, In Educational Data Mining (EDM), not much research has focused on this field. Therefore, this study proposes to combine an unsupervised discretization method called "equal width interval" and logistic regression as statistical machine learning to improve the performance of the model relating to students' performance. The discretization method is performed on student data with several intervals, namely: 3-interval, 4-interval, and 5-interval. Then, these intervals are combined with logistic regression in two regularizations, namely: lasso and ridge. Evaluation is carried out on all combinations. The experimental results indicate that the highest performance, in terms of the accuracy level, is achieved by the model combining a 3-interval and logistic regression on all regularization. This combination can increase the model performance based on the average accuracy level of about 4.08-8.49 on the ridge regularization and about 4.28-8.6 on the lasso regularization.**

Keywords—students' performance, data mining, machine learning, logistic regression, discretization

I. INTRODUCTION

Nowadays, almost all education institutions explore information and communication technology to enhance their process, for example learning process [1][2][3], evaluation process [4][5] despite its security [6]. This situation generates massive data pushing research about Educational Data Mining (EDM) [7]. The one of popular tasks in EDM is the classification of students' performance[8]. Here, student data is mined to get information about the students' performance. Before the mining process is done, methods sometimes are applied in the pre-processing stage [9]. They are the normalization, feature extraction, feature selection, the discretization method, etc. One of the objectives is to improve the performance of the system built [10].

Relating to the discretization method, it is used to transform the numerical method to a categorical method. Also, this method changes the non-standard probability distribution to the standard probability distribution. This method is applied with many reasons as follows: algorithms of machine learning requiring categorical or ordinal variables [11], the non-standard probability distributions causing the performance degradation of machine learning, the result of mapping smoothing out the relationships between observations because of providing a high-order ranking of values [12]. There are two mainstreams of the discretization methods, namely: the supervised methods and the unsupervised methods. The methods included in the unsupervised discretization methods are equal width interval, equal frequency interval, etc. For the supervised method, they are adaptive quantizes, chi merge, predictive value max, etc. [13]

The logistic regression is one of the methods in machine learning requiring the categorical variables. This method is grouped in the statistical machine learning [14]. As we have known, statistics have a very important role in the development of other sciences to draw conclusions, test hypotheses or theories, understand phenomena, analyze experiments, determine decisions, and so forth. Meanwhile, Machine learning, which is one branch of artificial intelligence (artificial intelligence) is currently continuing to experience growth and increasingly popular. The development of statistics and machine learning is of course because it can not be separated from the main factor, namely data. Machine learning has at least two main objectives, namely: solving problems in predicting the future (unobserved event) and gaining knowledge (knowledge discovery). Statistical machine learning refers to techniques for predicting the future and getting knowledge from data rationally. To be able to get these goals, statistical machine learning can be the right tool or method. Statistics acts as a learning base that utilizes statistical theory to inference and interprets the models, while machine learning focuses on the use of models to predict new data. Statistics and machine learning form a concept called Statistical machine learning using logistic regression models. Logistic regression models are among the models that are often used by machine learning practitioners [15].

Our paper focuses on the exploration of one of the discretization method called "equal width interval" and the statistical machine learning, namely: logistic regression on the students' performance domain.

978-1-7281-7434-1/20/$31.00©2020 IEEE

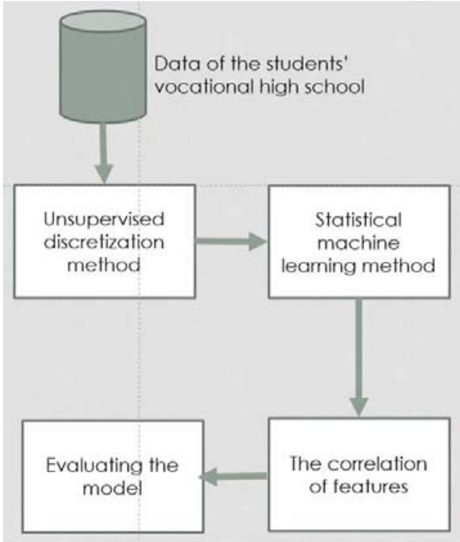

Fig. 1. The proposed architecture

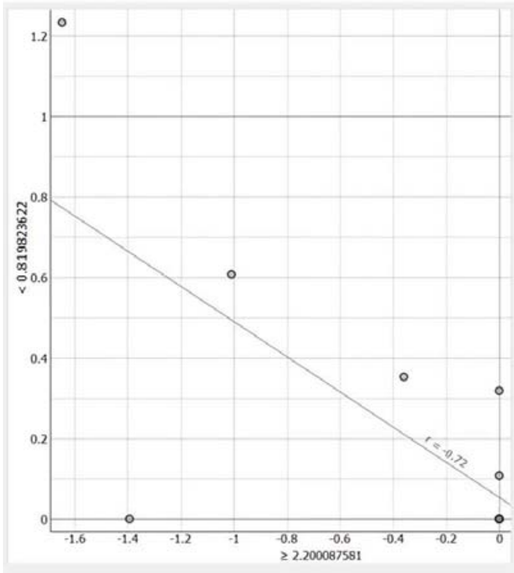

Fig. 2. Regression line with the highest correlation on discretization-3-interval-lasso

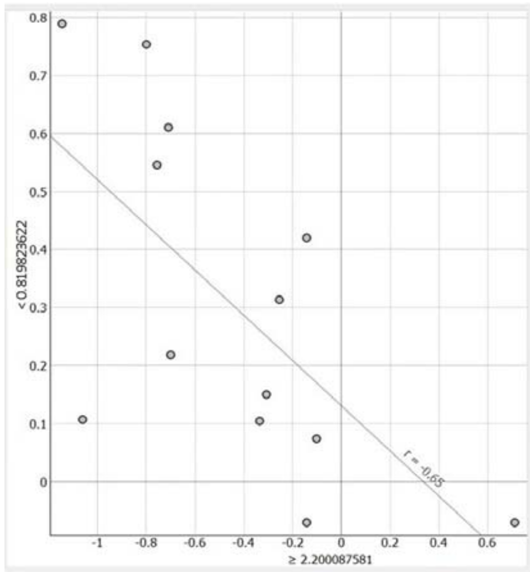

Fig. 3. Regression line with the highest correlation on discretization-3-interval-ridge

II. Method

In this section, this proposed method is depicted in Fig.1. Here, this architecture consists of many stages. They are as follows:

Stage 1: data of students' vocational high school

This research mine the student data of the previous work [16]. The student data is collected when the students join the evaluation process in the e-learning system. The data consists of 5 features that are extracted from 101 features of the raw dataset. They are as follows: Done, PercentTrue, Time, Hint, and Score which having the numeric data type.

Stage 2: the exploration of the unsupervised discretization method

In this stage, we apply the unsupervised discretization method called "equal width interval" with formula as follows:

$$w = \frac{(value_{max} - value_{min})}{k} \quad (1)$$

Where, w = *width* of an interval, k = the number of intervals which can be determined manually. For the range threshold on $value_{min} + iw$, where, $i = 1, ..., k-1$, each range can be defined as follows:

$$value_{min} + w, value_{min} + 2w, ..., value_{min} + (k-1)w \quad (2)$$

We divide student data into some intervals, namely: 3, 4, and 5. Then we definite as 3-interval, 4-interval, and 5-interval. This stage is done to know how many of the best intervals are implemented on our student data. This information is very important to the labeling process relating to the students' performance evel.

Stage 3: logistic regression

We combine the result of the previous stage with logistic regression in this stage. Also, there are two regularizations in the logistic regression, namely: Lasso [17] and Ridge[18][19]. We explore all regularizations to be experimented in our research to reach the optimum result.

Stage 4: analyzing the correlation of features

In this stage, we analyze the correlation on all features. We do one by one of the features to observe the correlation between the one features with the others. This stage also can be used to evaluate the relevant features to determine the target. Further, we also do the visualization for this step.

Stage 5: evaluating the model

The last stage evaluates the model built by the combination of the discretization method and logistic regression using the accuracy level metric. The stage observes the highest performance of the model.

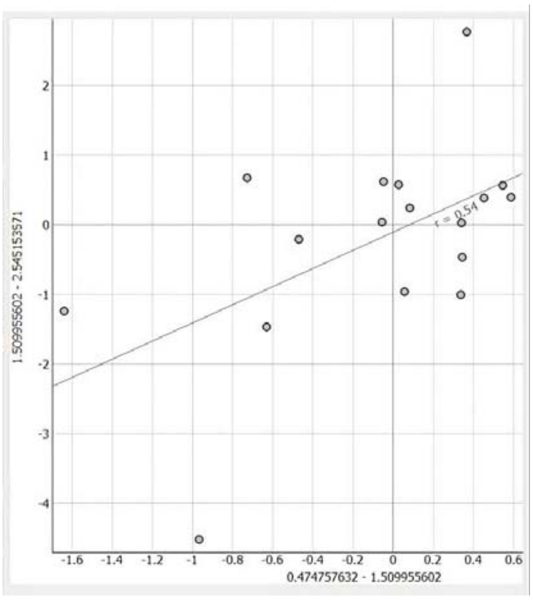

Fig. 4. Regression line with the highest correlation on discretization-4-interval-lasso

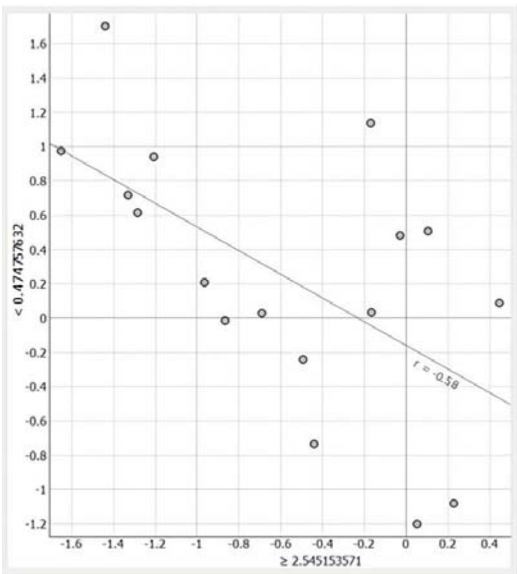

Fig. 5. Regression line with the highest correlation on discretization-4-interval-ridge

III. RESULT AND DISCUSSION

In this section, the proposed architecture is executed and then the result is analyzed. There are 2 sub-section explained, namely: the features correlation and the performance measurement. The performance of the model is evaluated using the accuracy level.

A. The correlation of features

The first session describes the correlation of features. After the discretization method is applied, we do combine the results of the discretization method using the logistic regression on all regulations.

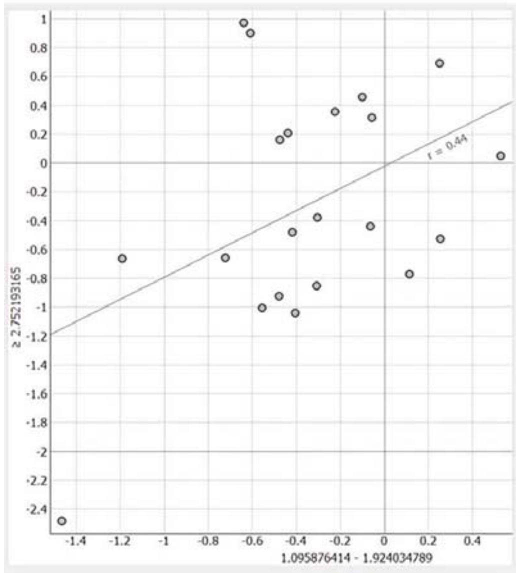

Fig. 6. Regression line with the highest correlation on discretization-5-interval-lasso

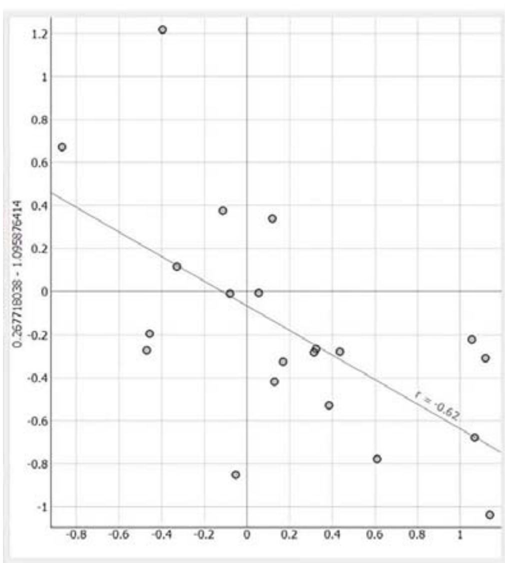

Fig. 7. Regression line with the highest correlation on discretization-5-interval-ridge

For that, there are 6 combinations, namely: 3-interval-lasso, 3-interval-ridge, 4-interval-lasso, 4-interval-ridge, 5-interval-lasso and 5-interval-ridge.

Then, we compute the correlation of features on each combination which is stated with the r symbol. In here, we only visualize the highest results on the all combinations depicted in the Fig.2-Fig.7. The general, the combinations of the discretization method, and the logistic regression on the ridge regulation correlate higher than others in all cases. In detail, the highest correlation is reached by 3-interval-lasso, namely: around r = -0.72. Contrarily, the lowest correlation of about r= 0.44 is achieved by 5-interval-lasso.

TABLE I. THE ACCURACY LEVEL OF THE COMBINATION OF LOGISTIC REGRESSION AND DISCRETIZATION METHOD WITH 3-INTERVAL

	3-interval-lasso					**3-interval-ridge**				
	Repeat train									
Training set size	2	3	5	10	20	2	3	5	10	20
10%	**85.3**	**85.3**	**85.3**	**85.3**	**85.3**	84.8	85	84.9	84.7	84.5
20%	84.6	84.6	84.6	84.6	84.6	84.6	84.6	84.6	84.4	84.4
30%	85	85	85	85	85	85	85	84.8	84.9	84.9
40%	**85.3**	**85.3**	**85.3**	**85.3**	**85.3**	**85.3**	**85.3**	**85.3**	85.1	85.2
50%	**84.2**	**84.2**	**84.2**	**84.2**	**84.2**	**84.2**	**83**	**85.3**	83.9	83.9
60%	84.8	84.8	84.8	84.8	84.8	84.8	84.8	84.8	84.8	84.8

TABLE II. THE ACCURACY LEVEL OF THE COMBINATION OF LOGISTIC REGRESSION AND DISCRETIZATION METHOD WITH 4-INTERVALS

	4-interval-lasso					**4-interval-ridge**				
	Repeat train									
Training set size	2	3	5	10	20	2	3	5	10	20
10%	79.4	80.1	80.6	81	81.2	79.9	80.4	80.8	80.5	80.3
20%	80.2	80.6	80.9	81.1	81.1	80.8	81	80.7	81	80.9
30%	**78.7**	79.6	80.2	80.7	80.9	80.6	80.4	80.5	80.9	80.7
40%	79.4	79.9	79.7	80	80.4	**79.4**	**79.4**	80	80.1	80.5
50%	78.9	79.5	80	80.4	80.3	80.7	80.7	80.7	80.5	80.4
60%	**82.6**	**82.6**	**82.6**	**82.6**	**82.6**	80.4	81.2	81.7	**82**	**82**

TABLE III. THE ACCURACY LEVEL OF THE COMBINATION OF LOGISTIC REGRESSION AND DISCRETIZATION METHOD WITH 5-INTERVALS

	5-interval-lasso					**5-interval-ridge**				
	Repeat train									
Training set size	2	3	5	10	20	2	3	5	10	20
10%	**77.5**	**77.5**	77.3	**77.5**	77.3	76	76.8	76.5	76.7	76.5
20%	75.3	**73.6**	74.9	75.9	76.3	75.8	**75.1**	75.8	76.3	76.5
30%	76.9	76.7	76	75.7	75.7	**76.9**	76.7	76.7	76.5	76.4
40%	76.5	76.5	75.6	76	76.2	76.5	76.5	75.6	75.9	75.9
50%	77.2	76	76.5	76.3	76.8	76.3	76	76.1	76.1	76.7
60%	76.1	76.1	76.1	76.1	76.1	76.1	76.1	76.1	75.9	75.9

Further, the 3-interval is illustrated in Fig.2-3. The correlation exploration of all features on the 3-interval is found that the highest correlation on logistic regression with lasso regulation. It is around r = -0.72. This condition occurs on the value of axis-x >= 2.200087581 and the value of axis-y <0.819823672. For logistic regression with ridge regulation, the highest correlation is about r = -0.65 with the value of axis-x >=2.200087581and on the value of axis-y < 0.8198234672.

The 4-interval is illustrated in Fig.4-5. The correlation exploration of all features on the 4-interval is found that the highest correlation on logistic regression with lasso regulation. It is around r = 0.54. This condition occurs on the value of axis-x 0.474757632-1.509955602 and the value of axis-y = 1.509955602-2.545153571. For logistic regression with ridge regulation, the highest correlation is about r = -0.58 with the value of axis-x >=2.545153571 and axis-y < 0.474757632.

The 5-interval is illustrated in Fig.6-7. The correlation exploration of all features on the 4-interval is found that the highest correlation on logistic regression with lasso regulation. It is around r = 0.44. This condition occurs on the value of axis-x 0.474757632-1.509955602 and the value of axis-y = 1.509955602-2.545153571. For logistic regression with ridge regulation, the highest correlation is about r = -0.62 with the value of axis-x >=2.545153571 and axis-y < 0.474757632.

B. The performance measurement

The next discussion is about the model performance, in terms of level accuracy. Discretization methods at all intervals and all regulations are presented in TABLE I-III. This step is carried out to analyze which intervals and regulations have the best performance in logistic regression. Trials are conducted on a set of training sizes multiples of 10 from 10% -60% and repetition of train 2, 3, 5, 10, and 20.

The trial results of a combination of 3-interval discretization and logistic regression regulation of the lasso and ridge are presented in TABLE I. The table shows the highest accuracy level of 85.3% occurred in the trial scenario training size of 10% and 40% with all the repeat train settings for lasso regulation. Whereas in ridge regulation, the test scenario is the training set size of 40% with repeat train 2, 3, 5, and the training set size 50% with repeat train 5. Conversely, in discretization 3 intervals, logistic regression with lasso regulation experiences the lowest accuracy level of 84.2 % occurs in the trial scenario the training set size is 50% for all repeat trains. Whereas logistic regression with ridge regulation experienced the lowest accuracy level of 83% in the training set size of 60% and repeat train 3.

The experimental result of a combination of 4-interval and logistic regression with the regulation on the lasso and ridge is presented in TABLE II. The table shows the highest level of accuracy in both regulations that occurred in the training set= 60%. For lasso, the highest accuracy level is 82.6% on all repeat trains. The ridge reaches the highest accuracy level of

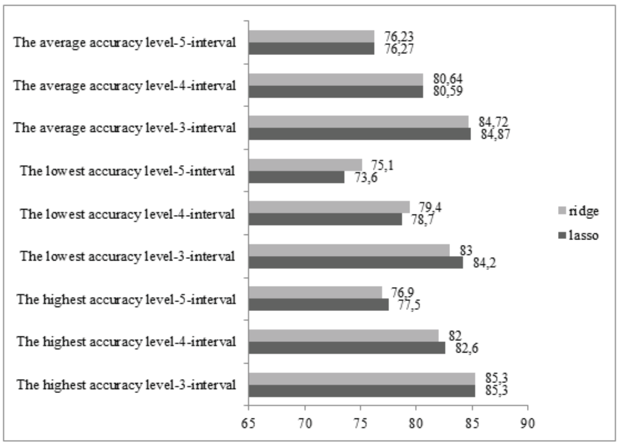

Fig. 8. The comparison of accuracy level on all intervals

82% on repeat trains 10 and 20. Meanwhile, the lowest accuracy level of 78.7% is experienced by Lasso regulations in the training set size of 30% with repeat train 2. For ridge, the lowest level of accuracy 79.4% is experienced oh the training set size of 40% with repeat train 2 and 3.

The experiment result of a combination of 5-interval and logistic regression with the regulation on lasso and ridge is represented in TABLE III. The table shows the highest accuracy level of 77.5% occurring on the training set size of 10% in almost all repeat train settings for lasso regulation. Meanwhile, in ridge regulation, the highest accuracy level of 76.9% is achieved with a training set size of 30% and repeat train 2. Conversely, logistic regression with the Lasso regulation experience the lowest accuracy level of 73.6% occurred in the training set size of 20% and repeat train 3. While in logistic regression with ridge regulation, the lowest accuracy level is 75.1% in the training set size of 20% and repeat train 2.

The overall results of the experiment presented in TABLE I-III show special conditions, namely: achieving quite high accuracy with a small training set size (10%). This is possible because the data of students selected in the construction of the model are sufficiently representative. Moreover, the data selected for this training set is most likely to have the most influence on the target. Thus, this data is sufficient to represent the data as a whole although only slightly in size. To further clarify the results obtained, the average accuracy level, lowest accuracy level, and highest accuracy level are presented in Fig.8. The discretization method for 3-interval dominates the highest results among the others, and then followed by the discretization method for 4-interval and finally 5-interval in all regulations of logistic regression

IV. CONCLUSION

The Equal width interval can be combined with logistic regression with lasso and ridge regulation in the students' performance data. Among the intervals that have been carried out, the combination of discretization of 3 intervals in this realm and logistic regression of all regulations has been proven to achieve the best results, in terms of the highest level of accuracy.

ACKNOWLEDGMENT

The authors wish to thank you for the support of the Department of Informatics, Universitas Negeri Surabaya, Indonesia on this research.

REFERENCES

[1] R. Yunis and K. Telaumbanua, "Pengembangan E-Learning Berbasiskan LMS untuk Sekolah, Studi Kasus SMA/SMK di Sumatera Utara," *J. Nas. Tek. Elektro dan Teknol. Inf.*, vol. 6, no. 1, Feb. 2017.

[2] P. Wanarti, E. Ismayanti, H. Peni, and Y. Yamasari, "The Enhancement of Teaching-Learning Process Effectiveness through The Development of Instructional Media Based on E-learning of Surabaya's Vocational Student," in *Proceedings of the 6th International Conference on Educational, Management, Administration and Leadership*, 2016, pp. 342–346.

[3] U. Aich and S. Banerjee, "Application of teaching learning based optimization procedure for the development of SVM learned EDM process and its pseudo Pareto optimization," *Appl. Soft Comput.*, vol. 39, pp. 64–83, Feb. 2016.

[4] Y. Yamasari, S. M. S. Nugroho, K. Yoshimoto, H. Takahashi, and M. H. Purnomo, "Identifying Dominant Characteristics of Students' Cognitive Domain on Clustering-based Classification," *Int. J. Intell. Eng. Syst.*, vol. 13, no. 1, 2020.

[5] M. W. Rodrigues, L. E. Zárate, and S. Isotani, "Educational Data Mining: A review of evaluation process in the e-learning," *Telemat. Informatics*, May 2018.

[6] P. Maniriho and T. Ahmad, "Information hiding scheme for digital images using difference expansion and modulus function," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 31, no. 3, pp. 335–347, Jul. 2019.

[7] L. C. Liñán, Á. Alejandro, and J. Pérez, "Educational Data Mining and Learning Analytics: differences, similarities, and time evolution Learning Analytics: Intelligent Decision Support Systems for Learning Environments," *RUSC. Univ. Knowl. Soc. J.*, vol. 12, no. 3, pp. 98–112, 2015.

[8] A. Peña-Ayala, *Educational data mining: Applications and trends*. 2014.

[9] N. T. Pang, M. Steinbach, and V. Kumar, *Introduction to Data mining*. 2006.

[10] Y. Yamasari, S. M. S. Nugroho, R. Harimurti, and M. H. Purnomo, "Improving the cluster validity on student's psychomotor domain using feature selection," in *2018 International Conference on Information and Communications Technology (ICOIACT)*, 2018, pp. 460–465.

[11] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.

[12] M. Kuhn and K. Johnson, *Feature Engineering and Selection : a Practical Approach for Predictive Models.* CRC Press LLC, 2019.

[13] J. Dougherty, R. Kohavi, and M. Sahami, "Supervised and Unsupervised Discretization of Continuous Features," Morgan Kaufmann Publishers, 1995.

[14] M. Sugiyama, *Introduction to Statistical Machine Learning*. Elsevier Inc., 2015.

[15] A. Genkin, D. D. Lewis, and D. Madigan, "Large-Scale Bayesian Logistic Regression for Text Categorization," *Technometrics*, vol. 49, no. 3, pp. 291–304, Aug. 2007.

[16] Y. Yamasari, S. Mardi, S. Nugroho, K. Yoshimoto, H. Takahashi, and M. H. Purnomo, "Expanding Tree-Based Classifiers Using Meta-Algorithm Approach: An Application for Identifying Students' Cognitive Level," *Int. J. Innov. Comput.*, vol. 15, no. 6, pp. 2085–2107, 2019.

[17] R. Tibshirani, "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 58. WileyRoyal Statistical Society, pp. 267–288, 1996.

[18] B. F. Swindel, "Geometry of Ridge Regression Illustrated," *Am. Stat.*,

vol. 35, no. 1, p. 12, Feb. 1981.

[19] N. R. Draper and R. C. van Nostrand, "Ridge Regression and James-Stein Estimation: Review and Comments," *Technometrics*, vol. 21, no. 4, p. 451, Nov. 1979.