Educational Data Mining Techniques Approach to Predict Student's Performance

Annisa Uswatun Khasanah and Harwati

***Abstract*—Predicting student's performance is one way that can be conducted by university to monitor their student to prevent student failed. Student final GPA is one parameter that must be full fill by student to graduate from university and it can be used to measure student's performance. Educational Data Mining is popular techniques to predict student's performance. This study tried to implement two popular data mining clustering and classification analysis to predict student's performance. K-means algorithm is used since it is very popular and easy to be implemented clustering algorithm. Linear Regression and Support Vector Machine (SVM) then used to predict the final GPA since the attributes used in this study is numerical data. The clustered data and non-clustered data were evaluated in the classification analysis and the MSE was compared. The result showed that clustered data had smaller RMSE and Linear Regression was better than SVM.**

***Index Terms*—Student, performance prediction, educational data mining.**

I. Introduction

Data mining is the process of automatically discovering useful information in large data repositories [1]. Data mining have been applied in many different scopes such as engineering, medical, marketing and also education. Educational Data Mining Techniques is the implementation of data mining techniques in education domain. There have been increasing number of researches interest in educational data mining especially to predict student performance [2]–[5].

Student performance plays important rule to measure the quality of the students, moreover in this current condition where universities operate in high competitive environment. In Indonesia, there was a rapid increase in the number of college since 2005 [6]. While in Yogyakarta, which is popular as a student city, in 2015 there were 130 colleges including academy, polytechnic, institute, university and others [7]. Predicting the student performance is one way that can be conducted by university to monitor their student to prevent student failed.

Student's performance prediction can be done by implementing popular data mining techniques such as classification method. Classification can be defined as the task of assigning objects to one of several predefined categories [1]. Lot of scholars have already applied classification method in educational data mining. Mueen *et al*. [8] used three different data mining classification algorithm (Naïve Bayes, Neural Network using Multilayer Perception with back-propagation type supervised-learning algorithm, and Decision Tree) to predict course final exam of the students, and Naïve Bayes was outperforming the other algorithm. Kabakchieva [9] used Decision tree classifier, Bayes classifiers and a Nearest Neighbour classifier to predict student's performance in Bulgarian University based on personal and pre-university characteristics. Ahmed and Elaraby [10] used decision tree (using ID3 algorithm) to predict the final grade mark of students in Information System department. Yadav and Pal [4] used C4.5, ID3 and CART decision tree algorithms to predict student's performance in the final exam. The result showed that C4.5 technique has highest accuracy.

This study will implement two data mining techniques, clustering analysis and classification analysis. The student data from Industrial Department, Universitas Islam Indonesia was firstly clustered. K-means algorithm was used since K-means is popular and easy to be implemented algorithm [11]. After the optimum number of cluster has be defined, the classification will be applied. Linear Regression and Support Vector Machine (SVM) were used since all of the attributes used in this study were numerical data. The clustered and non-clustered data will be evaluated in classification step, then the results based on Root Mean Squared Error (RMSE) will be compared. The rest of this paper is organized as follows, Section II present the literature review, Section III presents the research methods related to this study, Section IV shows the results and discussion. The concluding is finally made in Section V.

Manuscript received May 20, 2018; revised September 4, 2018. This work was supported in part by the Department of Industrial Engineering Universitas Islam Indonesia.

The authors are with the Department of Industrial Engineering Universitas Islam Indonesia, Jalan Kaliurang km. 14,5 Yogyakarta, Indonesia (e-mail: annisa.uswatun@uii.ac.id, harwati@uii.ac.id).

II. Literature Review

A. K-Means Algorithm

K-means [12] is one of the most popular and widely used clustering technique [11], [13] that firstly proposed by MacQueen in 1967 [14]. K-means is easy to be implemented and it also well known for its computational efficiency in large data set [11], [13]. It is also mentioned by Verma *et al*. [15], K-Means algorithm is faster than other clustering algorithm and also produces quality clusters when using huge dataset.

Described by Tan *et al*. [1], the K-means algorithm starts with K cluster centroids, which are initially randomly selected or derived from some a priori information. Each point in data set is then assigned to the closest centroid, and each collection of points assigned to a centroid is a cluster. The centroid of each cluster is then updated based on the point assigned to each cluster. This process is repeated until no point change clusters, or equivalently, until the centroids

doi: 10.18178/ijiet.2019.9.2.1184

remain the same. K-means algorithm can be further discussed as follows:

1) determine the number of cluster, K,
2) generate K cluster centroids randomly,
3) calculate the Euclidean distance using this following equation,

$$d(Xi, Zj) = \sqrt{(Xi - Zj)^2} \tag{1}$$

where Xi is the coordinate of the data and Zj is the coordinate of the cluster centroid.

4) assign each data point to the cluster that the distance between cluster centroid and the data point is the smallest,
5) recalculate the new centroids,
6) repeat until the centroids do not change.

There have been lot of scholars who implemented K-means algorithm such as Chandhok *et al*. [16], Yao *et al*. [17] and Moftah *et al*. [18] who successfully implemented K-means for image segmentation. Oyelade *et al*. [19] implemented K-means algorithm for analyzing students data of a private Institution in Nigeria, while Jaganathan and Jaiganesh [20] used K-means algorithm in web document clustering.

B. Linear Regression

Linear regression is an approach for modeling the relationship between dependent variable Y and one or more explanatory variables denoted X. This method has been also widely used as prediction method, such as Nguyen *et al*. who implemented Linear Regression to predict author age. Naseem *et al*. [21] implemented Linear Regression for face recognition, Antoch *et al*. [22] implemented Linear Regression to predict electrical consumption in Sardinia, Jung *et al*. [23] used Linear Regression to predict cancer incident and mortality in Korea, Melo-Espinosa *et al*. [24] implemented Linear Regression in surface tension prediction to know the relationship between surface tension of different vegetable oils and their fatty acid composition, while Chen [25] used Linear Regression to predict concrete compressive strength of Electric Arc Furnace Oxidizing slag.

C. Support Vector Machine

Support Vector Machine (SVM) is one of machine learning technique that can solve big data classification problem [26]. SVM works very well with high dimensional data and avoid the curse of dimensionality problem [1]. SVM are a set of maximum-margin classifier that minimize the classification error and maximize the geometric margin [27]. SVM has been implemented by scholars in many different scopes. Bauer *et al*. [28] implemented SVM for brain tissue segmentation to predict brain tumor. Wang *et al*. [29] proposed a color image segmentation using pixel wise SVM. Hu *et al*. [30] classifying species of fish in China based on color and texture features and using a multi-class support vector machine (MSVM). Zhang and Wu [31] used SVM to classified fruits using computer vision. Zhang *et al*. [32] used Kernel Support Vector Machine Decision Tree to distinguish among elderly subjects with Alzheimer's disease (AD), mild cognitive impairment (MCI), and normal controls (NC) based on Structural Magnetic Resonance Imaging. While, Shi *et al*. [33] proposed algorithms to forecast power output of zPhotovoltaic systems in China based upon weather classification and SVM

III. RESEARCH METHOD

The research flow chart for this study followed the Cross Industry Standard Process for Data Mining (CRISP-DM) framework. This framework was firstly developed in late 1996 [34]. The CRISP-DM methodology is described in terms of a hierarchical process model, consisting of sets of tasks. CRISP-DM provides a non-proprietary and freely available standard process for fitting data mining into the general problem-solving strategy of a business or research unit. There are six phases in CRISP-DM, including: Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation, and Deployment as shown in Fig 1.

The business process describes the background and objectives of this study as discussed in introduction in Section I. Data understanding explain the initial data. Data in this research were collected from student data base that can be accessed from Universitas Islam Indonesia's information system (UNISYS). There are four attributes used in this study as follows,

TABLE I: ATTRIBUTES

Senior high school grade	SHSG	(in scale (0-10)
Attendance in first semester	ATT	(in %)
GPA in first semester	GPA1	(in scale 0-4)
Final GPA	FGPA	(in scale 0-4)

The FGPA is the final GPA when the students graduated or drop out. It was used as the parameter for the student's performance. SHSG, ATT, and GPA1 used to estimate the new student performance (FGPA) based in their first semester. After the data were collected, the data was cleaned and transformed in Data Preparation Step. The initial data consisted of 178 data of student in academic year 2007 and after cleaning the data, it is only 104 data set available.

The next step is Modelling step. The student data with three attributes exclude the FGPA were clustered with K-means algorithm. To estimate the student performance, the clustered data and non-clustered data were applied in the classification (estimation) model using Linear Regression (LR) and Support Vector Machine (SVM), then the result were compared. This result that will be evaluated whether the model in fact achieves the objectives set or not. The Modelling and Evaluations step will further discuss in Section IV. The results of the study will be summarized in the Section V. Software Rapid Miner was used to apply the clustering and classification (estimation) model.

IV. RESULT AND DISCUSSION

The correlation matrix is presented in the Table II before conducting further analysis to explain the correlation between attributes SHSG, ATT, GPA1 and FGPA.

Bold value represent that those values are smaller that alpha (0.050) which indicates a probably significant difference between the actual mean value. It means that both of the attributes have correlation. It can be concluded from the correlation matrix that ATT and GPA1 has high positive correlation with FGPA.

TABLE II: CORRELATION MATRIX

Attributes	SHSG	ATT	GPA1	FGPA
SHSG	1	0.188	0.297	0.287
STT	0.188	1	0.691	0.838
GPA1	0.297	0.691	1	0.851
FGPA	0.287	0.838	0.851	1

Fig. 2 shows correlation between ATT, GPA1 and FGPA. It can be seen that the higher the student's attendance, they tend to have higher GPA1. Students who are diligent and have a good GPA in the first semester, they tend to have good final GPA.

A. Clustering Using K-Means

In this step, students were segmented based on three attributes (SHSG, ATT, GPA1) using K-means algorithm. The number of K used in this study are 3, 4, and 5. The cluster performance evaluated based on Davies Bouldin Index. DBI is one of measurement that can be used to find K-optimum in clustering algorithm such as K-means [35]–[37]. Lower DBI indicates better result.

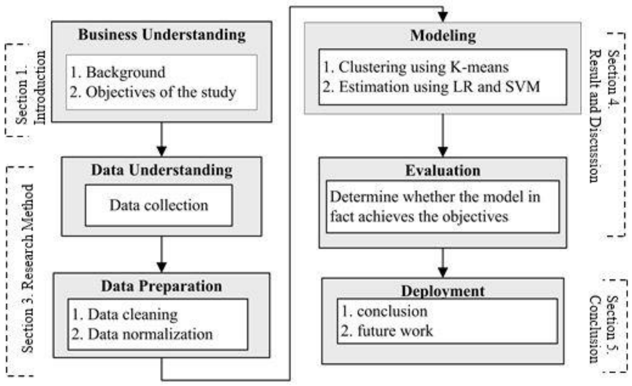

Fig. 1. Research framework.

TABLE III: DAVIES BOULDIN INDEX FOR CLUSTERING RESULT EVALUATION

Number of cluster	DBI
3	0.948
4	0.928
5	1.128
6	1.269

Based on DBI that shown in Table III, it can be evaluated that the K-optimum is 4 represented with the smallest number of DBI. The clustering analysis results with 4 cluster including the cluster profile are represented in Table IV.

TABLE IV: CLUSTERING RESULT

Cluster	SHSG	ATT	GPA1	Cluster member	Cluster Profile
1	8.18	0.87	2.35	23	smart enough and quite diligent
2	8.49	0.93	3.19	40	smart and diligent
3	7.35	0.50	1.60	11	not smart and not diligent
4	7.04	0.91	2.68	30	smart enough but diligent

The profile of each cluster is represented with the average value (mean) of each attributes. It can be concluded that the biggest cluster is cluster 2 which is dominated with smart and diligent students. These results then perform in estimation step.

B. Estimation Using Linear Regression and Support Vector Machine

In this study, the estimation step was performed by implementing LR and SVM algorithm. The purpose of this step is to compare the estimation accuracy between non clustered data from the original data and clustered data from clustering analysis. 10-fold cross validation was used to perform both of the classification algorithm and RMSE was used to evaluate the estimation results. Table V shows the RMSE from both algorithms for each non clustered and clustered data.

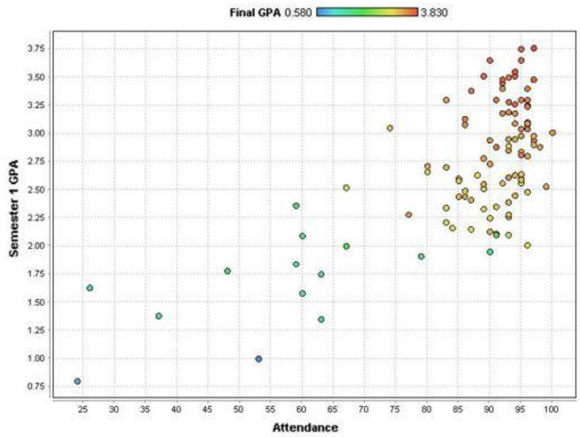

Fig. 2. Correlation graph for ATT GPA1 AND FGPA.

TABLE V: RMSE FOR CLASSIFICATION RESULT

Data	LR	SVM
no cluster	0.084	2.182
4 cluster	0.076	0.271

It can be seen that the classification results from clustered data is better than non-clustered data. It also can be concluded that LR is outperform SVM.

V. CONCLUSION

From the study that have conducted, it can be concluded that all the algorithm used can achieve the objective of the study. The optimum number of cluster can be defined after implementing K-means and it was 4 clusters. And it also can be concluded that cluster the data first before doing the classification analysis can minimize the RMSE. The results also showed that Linear Regression is better in predicting student's final GPA than SVM.

ACKNOWLEDGMENT

The author would like to thanks for Universitas Islam Indonesia who provide the data and for the financial support.

REFERENCES

[1] P.N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*, MA: Pearson Education, 2006.

[2] B. K. Baradwaj and S. Pal, “Mining educational data to analyze students’ performance,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 2, no. 6, pp. 63–69, 2011.

[3] B. K. Bhardwaj and S. Pal, “Data mining : A prediction for performance improvement using classification,” *Int. J. Comput. Sci. Inf. Secur.*, vol. 9, no. 4, 2011.

[4] S. K. Yadav and S. Pal, “Data mining : A prediction for performance improvement of engineering students using classification,” *World Comput. Sci. Inf. Technol. J. WCSIT*, vol. 2, no. 2, pp. 51–56, 2012.

[5] M. M. A. Tair and A. M. El-halees, “Mining educational data to improve students’ performance : A case study mining educational data to improve students’ performance : A case study,” *International Journal of Information and Communication Technology Research*, vol. 2, no. 2, 2012.

[6] Tempo, Tiap Dua Hari, Satu Perguruan Tinggi Muncul di Indonesia. (2015). [Online]. Available: http://nasional.tempo.co/read/news/2015/06/04/079672015/tiap-dua-hari-satu-perguruan-tinggi-muncul-di-indonesia

[7] Kementerian Riset Teknologi Dan Pendidikan Tinggi, Grafik Jumlah Perguruan Tinggi. (2016). [Online]. Available: http://forlap.dikti.go.id/perguruantinggi/homegraphpt

[8] A. Mueen, B. Zafar, and U. Manzoor, “Modeling and predicting students’ academic performance using data mining techniques,” *I.J. Modern Education and Computer Science*, vol. 11, pp. 36-42, 2016.

[9] D. Kabakchieva, “Predicting student performance by using data mining methods for classification,” *Cybern. Inf. Technol.*, vol. 13, no. 1, pp. 61–72, 2013.

[10] A. B. E. D. Ahmed and I. S. Elaraby, “Data mining: A prediction for student’s performance using classification method,” *World J. Comput. Appl. Technol.*, vol. 2, no. 2, pp. 43–47, 2014.

[11] W. Kwedlo, “A clustering method combining differential evolution with the K -means algorithm,” *Pattern Recognit. Lett.*, vol. 32, no. 12, pp. 1613–1621, 2011.

[12] J. MacQueen, “Some methods for classification and analysis of multivariate observations,” *Some Methods for Classification and Analysis of Multivariate Observations*, vol. 1, pp. 281-297, 1967.

[13] N. Dhanachandra, K. Manglem, and Y. J. Chanu, “Image segmentation using k-means clustering algorithm and subtractive image segmentation using k-means clustering algorithm and subtractive clustering algorithm,” *Procedia - Procedia Comput. Sci.*, vol. 54, pp. 764–771, 2015.

[14] A. Shafeeq, “Dynamic clustering of data with modified k-means algorithm dynamic clustering of data with modified k-means algorithm,” in *Proc. International Conference on Information and Computer Networks*, vol. 27, no. 1–6, 2014.

[15] M. Verma, M. Srivastava, N. Chack, A. K. Diswar, and N. Gupta, “A comparative study of various clustering algorithms in data mining Manish Verma , Mauly Srivastava , Neha Chack, Atul Kumar Diswar, Nidhi Gupta,” *International Journal of Engineering Research and Applications (IJERA)*, vol. 2, no. 3, pp. 1379–1384, 2012.

[16] M. C. Chandhok, S. Chaturvedi, and A. A. Khurshid, “An approach to image segmentation using k-means clustering algorithm,” *International Journal of Information Technology (IJIT)*, vol. 1, no. 1, pp. 11–17, 2012.

[17] H. Yao, Q. Duan, D. Li, and J. Wang, “An improved K-means clustering algorithm for fish,” *Math. Comput. Model.*, vol. 58, no. 3–4, pp. 790–798, 2013.

[18] H. M. Moftah, A. T. Azar, E. T. Al-Shammari, N. I. Ghali, A. E. Hassanien, and M. Shoman, “Adaptive k-means clustering algorithm for MR breast image segmentation,” *Neural Comput. Appl.*, vol. 24, no. 7–8, pp.1917–1928, 2014.

[19] O. J. Oyelade, O. O. Oladipupo, and I. C. Obagbuwa, “Application of k means clustering algorithm for prediction of students academic performance,” *Int. J. Comput. Sci. Inf. Secur. IJCSIS*, vol. 1, no. 1, pp. 292–295, 2010.

[20] P. Jaganathan and S. Jaiganesh, “An improved K-means algorithm combined with Particle Swarm Optimization approach for efficient web document clustering,” *Green Computing, Communication and Conservation of Energy (ICGCE)*, 2013.

[21] I. Naseem, R. Togneri, and M. Bennamoun, “Linear regression for face recognition,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 11, pp. 2106–2112, 2010.

[22] J. Antoch, L. Prchal, M. Rosaria De Rosa, and P. Sarda, “Electricity consumption prediction with functional linear regression using spline estimators,” *J. Appl. Stat.*, vol. 37, no. 12, pp. 2027–2041, 2010.

[23] K.W. Jung, Jung, S. Park, Y.J. Won, H. J. Kong, J. Y. Lee, H. G. Seo, and J.-S. Lee, “Prediction of cancer incidence and mortality in Korea, 2012,” *Cancer Res. Treat.*, vol. 44, no. 1, pp. 25–31, 2012.

[24] E. A. Melo-espinosa, Y. Sánchez-borroto, and M. Errasti, “ISES solar world congress surface tension prediction of vegetable oils using artificial neural networks and multiple linear regression,” *Energy Procedia*, vol. 57, pp. 886–895, 2014.

[25] L. Chen, “A Multiple linear regression prediction of concrete compressive strength based on physical properties of electric arc furnace oxidizing slag,” *International Journal of Applied Science and Engineering*, vol. 7, no. 2, pp. 153–158, 2010.

[26] S. Suthaharan, *Support Vector Machine*, Boston, MA: Springer, 2016.

[27] H. Hernault, H. Prendinger, D. A. du Verle, and M. Ishizuka, “HILDA : A discourse parser using support vector machine classification,” *Dialogue and Discourse*, vol. 1, no. 3, pp. 1–33, 2010.

[28] S. Bauer, L. P. Nolte, and M. Reyes, “Fully automatic segmentation of brain tumor images using support vector machine classification in combination with hierarchical conditional random field regularization,” *Int. Conf. Med. Image Comput. Comput. Interv.*, vol. 6893, no. 354-361, 2011.

[29] X. Wang, T. Wang, and J. Bu, “Color image segmentation using pixel wise support vector machine classification,” *Pattern Recognit.*, vol. 44, no. 4, pp. 777–787, 2011.

[30] J. Hu, D. Li, Q. Duan, Y. Han, G. Chen, and X. Si, “Fish species classification by color , texture and multi-class support vector machine using computer vision,” *Comput. Electron. Agric.*, vol. 88, pp. 133–140, 2012.

[31] Y. Zhang and L. Wu, Classification of Fruits Using Computer Vision and a Multiclass Support Vector Machine, *Sensors*, vol. 12, pp. 12489–12505, 2012.

[32] Y. Zhang, S. Wang, and Z. Dong, “Classification of alzheimer disease based on structural magnetic resonance imaging by kernel support vector machine decision tree,” *Progress In Electromagnetics Research*, vol.144, pp.171–184, 2014.

[33] J. Shi, W. J. Lee, Y. Liu, Y. Yang, and P. Wang, “Forecasting power output of photovoltaic systems based on weather classification and support vector machines,” *IEEE Trans. Ind. Appl.*, vol. 48, no. 3, pp. 1064–1069, 2012.

[34] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, and C. Shearer, R. Wirth, *CRISPDM 1.0 Step-by-Step Data Mining Guide*, Technical report: CRISP-DM, 2000.

[35] E. Rendón, I. Abundez, A. Arizmendi, and E. M. Quiroz, “Internal versus External cluster validation indexes,” *Int. J. Comput. Commun.*, vol. 5, pp. 27-34, 2011.

[36] S. M. S. Hosseini, A. Maleki, and M. R. Gholamian, “Cluster analysis using data mining approach to develop CRM methodology to assess the customer loyalty,” *Expert Syst. Appl.*, vol. 37, no. 7, 2010.

[37] S. Petrovic, “A comparison between the silhouette index and the davies-bouldin index in labelling ids clusters,” *Tthe 11th Nord. Work. Secur. IT-Systems, Nord.*, pp. 53–64, 2006.

Annisa Uswatun Khasanah was born on June 21, 1990 in Bantul (Indonesia). In 2008 she was graduated from SMA N 1 Yogyakarta. In July 2012 she was graduated from Industrial Engineering Universitas Gadjah Mada (ST), Yogyakarta. At the 7th semester she took fast tract program to continue master degree at the same university while finishing bachelor degree with scholarship from Ministry of Education. In September 2012 she took double degree for his mater program in Industrial Management, National Taiwan University of Science and Technology, Taipei Taiwan and graduated in July 2013 (MBA). After finishing the MBA she came back to UGM to finish her M.Sc. She worked on Data Mining Application for his master and her research in the last four years. Now she works as a lecturer in Universitas Islam Indonesia.

Harwati was born on May 22, 1982 in Medan (Indonesia). In 2000 she has graduated from high school no 1 of Medan North Sumatera. In 2005 has graduated with cum laude from Gadjah Mada University on specially modelling in Industrial Engineering. From 2009 to 2011 she was studied in Bandung Institute of Technology on specialty in supply chain management. In 2011 she defended the thesis “Benefit Sharing Trough Supply Chain Integration in Vendor–Managed Inventory”. Now she works in Universitas Islam Indonesia as lecturer and junior researcher.