Educational Data Mining in Predicting Student Final Grades on Standardized Indonesia Data Pokok Pendidikan Data Set

Nathan Priyasadie, Sani Muhammad Isa
BINUS Graduate Program – Master of Computer Science
Bina Nusantara University, Jakarta
Indonesia, 11480

***Abstract*—Educational Data Mining has been implemented in predicting student final grade in Indonesia. It can be used to improve learning efficiency by paying more attention to students who are predicted to have low scores, but in practice it shows that each algorithm has a different performance depending on the attributes and data set used. This study uses Indonesian standardized students' data named Data Pokok Pendidikan to predict the grades of junior high school students. Several prediction techniques of K-Nearest Neighbor, Naive Bayes, Decision Tree and Support Vector Machine are compared with implementation of parameter optimization and feature selection on each algorithm. Based on accuracy, precision, recall and F1-Score shows that various algorithm performs differently based on the high school data set, but in general Decision Tree with parameter optimization and feature selection outperform other classification algorithm with peak F1-Score at 61.48% and the most significant attribute in are First Semester Natural Science and First Semester Social Science score on predicting student final score.**

Keywords*—Educational data mining; student performance; classification models; feature selection; parameter optimization

I. Introduction

Educational data mining is a rapidly growing multidisciplinary area of study devoted to studying and developing techniques for extracting useful information from enormous amounts of data generated in educational settings [1]. Because information technology has been significant in improving the area of education over the last decade, nearly every institution now maintains a student information system [1]. This information includes student demographic, parent information, scores etc. Applying data mining techniques to educational processes can be beneficial in identifying important trends, performance summaries, and insights, which will assist students in identifying areas for improvement. An institution's academic performance, life cycle management, course selection, retention rate measurement, and grant money management may all be considered [2].

Predicting student grades is one of educational mining's applications. Grades are critical components of education since they act as a barometer of a student's competency and performance within that institution. Predicting a student's final grade might also encourage a school to improve its teaching techniques and create a more pleasant learning environment [1]. By providing additional support to students who were previously projected to have lower grades, it is possible to enhance learning efficiency and the overall student grade [3]. Finally, a high score improves a student's chances of admission to a more prestigious higher education program.

Data from Organization for Economic Co-operation and Development (OECD) shows that Indonesian student ranked 72 out of 77 countries on Programme for International Student Assessment (PISA) report in 2018, and this rank tends to stagnate for the last 10-15 years. It can be concluded that education in Indonesia is still lagging compared to other countries.

The purpose of prediction is to determine the value of an unknown variable that correspond to the student [1]. In Indonesia there have been several researches that investigate student performance prediction using Naive Bayes (NB) [4], Decision Tree (DT) [5], Support Vector Machine (SVM) [6], K-Nearest Neighbor (K-NN) [7] and Regression Analysis [8] but there is no research that predicts student grades by using standardized national socio-demographic aspects of students such as the Data Pokok Pendidikan (DAPODIK).

The purpose of this study is to find the best algorithm to predict student final score using standardized DAPODIK data combined with student historical grade from three public junior high schools in Indonesia. With standardized data, schools throughout Indonesia can determine the best method to predict student grades in their schools. It can be used to improve learning efficiency by providing additional support to students who were previously projected to have lower grades. This study will compare four different algorithms, Naive Bayes, Decision Tree, K-Nearest Neighbors and Support Vector Machine with two data mining optimization methods, parameter optimization (PO) and feature selection (FS).

II. Related Work

Mengash [9] in their study found that using Artificial Neural Networks to predict student performance of 2039 Computer Science students at a Saudi Public University from 2016 to 2019 had an accuracy rate of greater than 79%, outperforming other classification techniques such as Decision Trees, Support Vector Machines, and Naive Bayes. It compares various pre-admission criteria (high school grade average, Scholastic Achievement Admission Test score, and General

Aptitude Test score). The findings indicate that the Scholastic Achievement Admission Test score is the most reliable predictor of future student performance of any pre-admission criteria. As a result, admissions systems should give this score a higher weight.

Rifat et al. [10] perform research to predict students' performance using transcript data from a Bangladeshi institution. The authors utilized six cutting-edge classification algorithms (Gradient Boosted Tree, Random Forest, Tree Ensemble, Decision Tree, Support Vector Machines and K-Nearest Neighbor) to forecast students' final grades. The findings indicated that the Random Forest algorithm performed the best, with an accuracy of 94.1%, followed by the Tree Ensemble method.

Yao et al. [11] perform research to determine the final score of secondary school students utilizing their personal data. The data set contains a variety of factors, including parent information, student health status, financial status and. attendance etc. With feature selection, the J48 algorithm achieved the highest accuracy of 84.39%, whereas without feature selection, the OneR algorithm achieved the highest accuracy of 84.19%.

Saa et al. [12] gathered data on student demographics, course teacher information, student general information, and prior performance from a private institution in the United Arab Emirates using various algorithms (Decision Tree, Random Forest, Gradient Boosted Trees, Deep Learning, Naive Bayes, Logistic Regression and Generalized Linear Model). With 75.52% accuracy, the Random Forest method topped the other classifiers, followed by the Logistic Regression technique.

Fairos et al. [13] conduct research to predict student performance using Universiti Technology Cawangan Kelantan and Universiti Teknologi MARA Cawangan Negeri Sembilan student data with total 631 transcript from 2013 to 2016, with various attributes such as gender, all the course enrolled by student including the course grade. They develop a model to predict student performance using K-Nearest Neighbor, Naive Bayes, Decision Tree and Logistic Regression Model. It shows that Naive Bayes outperform other classification algorithm with 89.26% accuracy.

Based on previous study shows that data set plays a big role in determining which algorithm is the best for predicting student final score. On this research a standardized data set is used to determine which algorithm is best to be applied throughout high school in Indonesia.

III. Methodology

This research uses The Cross Industry Standard Process for Data Mining (CRISP-DM) [14]. CRISP-DM is the most used methodology for developing Data Mining projects; it consists of six steps as visualized in Fig. 1. The first step is business understanding where the purpose is to provide context for the objectives and data. The second step is data understanding where its purpose is to determine what can be expected and accomplished from the data. The third step is data preparation where it involves cleaning, integrating, and formatting the data [15]. The fourth step is modelling where the Naive Bayes, Decision Tree, Support Vector Machine and K-Nearest Neighbor algorithm are used then optimized using feature selection and parameter optimization method to produce the best prediction model. The last step is Evaluation of each model based on accuracy, precision, recall and F1-Score.

A. Data Understanding

The first step is to collect data from various sources, locate and gather data for training and validate the algorithm, which may be spread over many spreadsheets, databases, or webpages. This research uses data from 3 high schools from the Jakarta class of 2020 and 2019. With a total of 926 student data each with 33 variables, in xlsx format with Table I attributes.

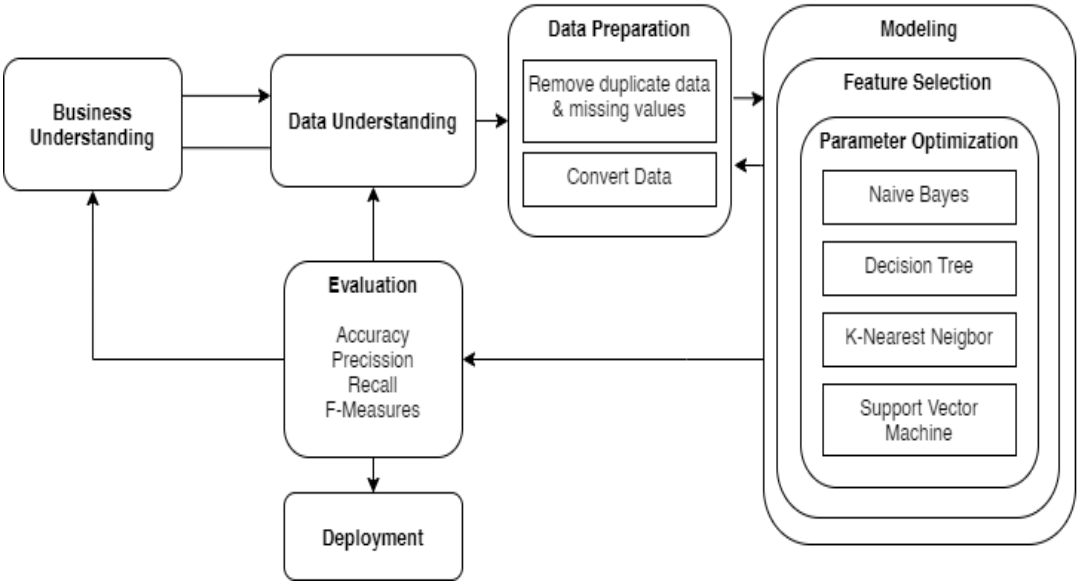

Fig. 1. CRISP-DM Methodology.

TABLE I. DATA ATTRIBUTES

Variables	***Description***	***Possible Value***
Final Grades	Average of Student's Final Grade	A, B, C, D, E, F, G
Entrance Grades	Student's Final Grade in Primary School	A, B, C, D, E, F, G
Gender	Student's Gender	Male, Female
Type of living	Student's living types	Living with Parents, Boarding House, Living with Guardian Others
Transportation Method	Student's Transportation Method to School	Car, Motorcycle, Bicycle, Public Transportation, Taxibike, On Foot, Others
Father's Education	Father's latest Education	None, Primary School, Junior High school, Senior High school, Diploma, Bachelor's Degree, Master's Degree
Father's Occupation	Father's latest Occupation	General employees, Entrepreneur, Merchant, Deceased, Laborer, Government Employees/Soldiers/Police, Others
Father's Income	Father's Monthly Income	No Income, <Rp 500.000, Rp 500.000 - Rp 999.999, Rp 1.000.000 - Rp 1.999.999, Rp 2.000.000 - Rp 4.999.999, Rp 5.000.000 - Rp 20.000.000, >Rp 20.000.000
Mother's Education	Mother's latest Education	None, Primary School, Junior Highschool, Senior Highschool, Diploma, Bachelor's degree, Master's degree
Mother's Occupation	Mother's latest Occupation	General employees, Entrepreneur, Merchant, Deceased, Laborer, Government Employees, Soldiers/Police, Others
Mother's Income	Mother's Monthly Income	No Income, <Rp 500.000, Rp 500.000 - Rp 999.999, Rp 1.000.000 - Rp 1.999.999, Rp 2.000.000 - Rp 4.999.999, Rp 5.000.000 - Rp 20.000.000, >Rp 20.000.000
First and Second Semester's Grades	Grades in Religion, Civic, Bahasa Indonesia, English, Mathematics, Natural Science, Social Science, Art and Culture, Sports	A, B, C, D, E, F, G

After being collected, the data is combined into a single data set.

B. Data Preparation

This step is to delete duplicate data or have empty attributes leaving 759 student data remaining. Then, numerical attributes are converted into categorical attributes based on Table II rules.

TABLE II. MAPPING RULES

Numerical Values	**Categorical Values**
Above 95	A
90-94	B
85-89	C
80-84	D
75-79	E
70-74	F
Below 70	G

Then the data is divided into two categories: training and testing, The training data set comprises 80% of the total data will be used to train the algorithm for classifying student data, whereas the testing data set comprises 20% will be used to evaluate the trained model's performance.

C. Modelling

Classification is a concept that refers to the act of classifying things according to information about one or more of its attributes, as well as categorizing them according to a collection of already classified items [16]. This research uses RapidMiner software which has a large collection of classification and optimization methods [17].

Naive Bayes is one of the simplest and most frequently used classification methods [18]. This method is based on Bayesian theory of probability, which assumes that a class is independent of each other [19]. With a simple concept, Naive Bayes uses a conditional probability model with P(six) as the probability of the class and assumes that the value of a predictor (x) in a particular class (c) does not depend on the value of other predictors. Naive Bayes can be described in the following equation 2.2.

$$ax(t) = Ax(t) + Bu(t) + B1w(t) \quad (1)$$

Naive Bayes has the advantages of being fast and efficient in using memory, able to handle quantitative data and discrete data, resilient to noise and only requires a small amount of data for classification and can handle missing values by ignoring values during probability calculations [20].

Decision Tree is a method for classifying by looking for differences between classes and dividing them using attributes by making a diagram in the form of a tree. This method uses a divide-and-conquer approach, one of the advantages is the ease of reading the model that has been made, with this convenience, information related to the identification of important attributes and relationships between classes can be used for analysis and research in the future [21]. By splitting to determine branches, there are several ways that can be used, such as Gini Impurity which looks for branches that have the most homogeneous results, which means that the results of the division have similar characteristics.

K-Nearest Neighbor performs classification by comparing input with training data like it, each data consists of n-attributes represented by a point on the n-dimensional graph, if given a data whose class is not known then K-Nearest Neighbor will look for several k training data closest to the location with the data [19] . After knowing the number of closest samples, the algorithm can estimate the class of the data based on the number of closest samples, the distance can be measured using several formulas such as Euclidean Distance. The advantage of K-Nearest Neighbor is that it can group a lot of data efficiently and in a fast time [20], but it has the disadvantage that it can become significantly slower with an increasing amount of data.

Support Vector Machine is a method that can be used to classify linear and non-linear data [22]. The way it works is by doing non-linear mapping to change the training data to a higher dimension, in this dimension he looks for the most optimal linear hyperlane separator. With enough nonlinear mappings that have high dimensions, data from the two classes can always be separated by hyperlane. Support Vector Machines finds hyperlane using support vectors and margins [19]. The advantages of Support Vector Machines are that it works well if there is a clear distance between class differences, effective for cases where the number of dimensions is more than the number of sample data, but the disadvantages are that it is not suitable for large data sets and does not perform well for data sets that have a lot of noise.

Feature selection reduces the number of dimensions of the data set thereby reducing processor and memory usage [23]. With this feature selection removes irrelevant attributes from the data set and improves the accuracy of the algorithm. For this study forward selection is used where it starts with an empty attribute set and adds attributes in it until the stopping criterion is met [24]. This method allows avoiding the use of additional memory and processor and improves the accuracy of the algorithm by removing irrelevant attributes from the data set.

Parameter optimization is a technique used to find the best combination of parameters to get the optimum performance of each algorithm. In the approach there are several ways such as through the grid, evolutionary and quadratic. By running iterations according to the provisions, then trying to calculate new parameters that may be between the previous parameters, and after that compare the results of the accuracy of the initial parameters and the parameters of the calculation results. The grid search is originally an exhaustive search based on defined subset of the hyper-parameter space. The hyper-parameters are specified using minimal value (lower bound), maximal value (upper bound) and number of steps [25]. In this case the grid search is used since those best ranges and dependencies are known.

D. Evaluation

In this research method the evaluation will be carried out using multi-class confusion matrix to evaluate each model accuracy, precision, recall and F1-Score.

E. Result and Analysis

Tables below summarizes the algorithm performance for Naive Bayes, Decision Tree, Support Vector Machines and K-Nearest Neighbor without any optimization, and with both feature selection and parameter optimization on different high school testing data set. For more concise tabulation the methods Naïve Bayes, Decision Tree, Support Vector Machines, K-Nearest Neighbor, combined feature selection and parameter optimization are abbreviated to NB, DT, SVM, K-NN and FS+PO, respectively.

On Table III shows result for high school A data set, Decision Tree with optimization shows to be the best overall algorithm and with best F1-Score at 54.01% and the most significant attribute on that algorithm is *First Semester Social Science*, *Second Semester English* and *Gender*, while K-Nearest Neighbor with optimization achieved best accuracy score at 77.36% while Naive Bayes with optimization achieved best recall score at 52.41%.

On Table IV shows result for high school B data set, Decision Tree with optimization shows to be the best overall algorithm and with best accuracy at 85.71% and precision at 64.40% and the most significant attribute is First Semester Religion score, First Semester Natural Science score, First Semester Sports score, First Semester Arts score, First Semester Social Science score and Second Semester Arts score while Naïve Bayes without optimization achieved best recall score at 61.11%.

On Table V shows result for high school C data set, Naive Bayes with optimization shows to be the best overall algorithm on all measurements and the most significant attributes are First Semester Natural Science score, First Semester Social Science score and Gender.

TABLE III. PERFORMANCE ON HIGH SCHOOL A TESTING DATA SET

Algorithm	Optimization Method	Accuracy	Precision	Recall	F1-Score
NB	-	69.81%	60.60%	48.72%	54.01%
	PO + FS	76.92%	55.47%	52.41%	53.89%
DT	-	62.26%	55.90%	39.12%	46.02%
	PO + FS	73.58%	62.46%	50.63%	55.92%
SVM	-	66.04%	32.72%	36.22%	34.38%
	PO + FS	71.70%	60.58%	44.37%	51.46%
K-NN	-	67.92%	58.58%	42.29%	49.11%
	PO + FS	77.36%	38.54%	44.57%	41.33%

TABLE IV. PERFORMANCE ON HIGH SCHOOL B TESTING DATA SET

Algorithm	Optimization Method	Accuracy	Precision	Recall	F1-Score
NB	-	76.79%	53.12%	61.11%	56.70%
	PO + FS	83.93%	60.60%	61.01%	60.80%
DT	-	66.07%	43.75%	42.56%	43.14%
	PO + FS	85.71%	64.40%	58.83%	61.48%
SVM	-	66.04%	32.72%	36.22%	34.38%
	PO + FS	82.14%	60.91%	57.54%	59.17%
K-NN	-	69.64%	47.62%	44.15%	45.81%
	PO + FS	78.57%	53.08%	57.00%	54.97%

TABLE V. PERFORMANCE ON HIGH SCHOOL C TESTING DATA SET

Algorithm	Optimization Method	Accuracy	Precision	Recall	F1-Score
NB	-	48.78%	42.39%	51.62%	46.55%
	PO + FS	79.41%	66.62%	66.62%	66.62%
DT	-	63.44%	50.91%	50.86%	58.88%
	PO + FS	70.73%	57.30%	64.93%	60.87%
SVM	-	60.98%	42.00%	36.70%	39.17%
	PO + FS	68.75%	41.86%	45.56%	43.63%
K-NN	-	58.54%	39.78%	38.17%	38.95%
	PO + FS	75.61%	45.65%	51.49%	48.39%

In general, the experiment shows that feature selection and parameter optimization improve the accuracy of the classifier algorithm up to 62.79%. However, it also shows that various algorithms show different accuracy results with different high school data set. Decision Tree with optimization shows to be the best overall combination to predict student performance on A and B high school data set with peak F1-Score at 61.48%, meanwhile Naive Bayes with optimization shows to be the best combination on high school C data set with 66.62% F1-Score And in almost every data set shows that the most significant attributes are First Semester Natural Science, First Semester Social Science score on predicting student final score.

IV. CONCLUSION

The result of this study found that: (a) Overall best F1-Score is achieved by Decision Tree with feature selection and parameter optimization. (b) In general parameter optimization and feature selection show to improve algorithm performance. (c) The most significant attributes in predicting student score are First Semester Natural Science score and First Semester Social Science score. (d) Even with the same attributes from different schools' data set each algorithm performs differently. With these results it can be concluded that the research has achieved its objectives. But there is a room of improvement on this research since there are lack of data varieties because we're only using data from single province in Indonesia.

REFERENCES

[1] C. Romero and S. Ventura, Educational Data Mining: A Review of the State of the Art, IEEE Transactions on Systems Man and Cybernetics Part C (Applications and Reviews), 2010.

[2] M. Goyal and R. Vohra, "Application of Data Mining in Higher Education," International Journal of Computer Science Issues (IJCSI), 2012.

[3] R. Asif, A. Merceron, S. A. Ali and N. G. Haider, "Analyzing Undergraduate Student's Performance," Computers & Education, 2017.

[4] M. Jannah, "Penerapan Data Mining Prediksi Nilai Ujian Nasional (UN) Siswa SMP Menggunakan Metode Naive Bayes," Jurnal Informatika, Manajemen dan Komputer, 2020.

[5] P. Mayadewi and E. Rosely, "Prediksi Nilai Proyek Akhir Mahasiswa Menggunakan Algoritma Klasifikasi Data Mining," Seminar Nasional Sistem Informasi Indonesia, 2015.

[6] R. Thaniket, Kusrini and E. T. L., "Prediksi Kelulusan Mahasiswa Tepat Waktu Menggunakan Algoritma Support Vector Machine," Jurnal Teknologi dan Rekayasa, 2020.

[7] S. R. Rani, S. R. Andani and D. Suhendro, "Penerapan Algoritma K-Nearest Neighbor untuk Prediksi Kelulusan Siswa pada SMK Anak Bangsa," Prosiding Seminar Nasional Riset Information, 2019.

[8] H. Susanto and Sudiyatno, "Data Mining Untuk Memprediksi Prestasi Siswa Berdasarkan Sosial Ekonomi, Motivasi, Kedisiplinan Dan Prestasi Masa Lalu," Jurnal Pendidikan Vokasi, 2014.

[9] H. A. Mengash, "Using Data Mining Techniques to Predict Student Performance to Support Decision Making in University Admission Systems," Institute of Electrical and Electronics Engineers, 2020.

[10] M. R. I. Rifat, A. A. Imran and A. S. M. Badrudduza, "Educational Performance Analytics of Undergraduate Business Students," I.J. Modern Education and Computer Science, 2019.

[11] Y. Yao, Z. Chen, S. Byun and Y. Liu, "Using Data Mining Classifiers to Predict Academic Performance of High School Students," Scientific and Practical Cyber Security Journal, 2019.

[12] A. A. Saa, M. Al-Emran and K. Shaalan, "Mining Student Information System Records to Predict Students' Academic Performance," Advances in Intelligent Systems and Computing, 2019.

[13] W. Fairos, W. F. W. Yaacob, S. Azlin, S. Nasir, W. Faizah, N. M. Sobri and C. Mara, "Supervised data mining approach for predicting student performance," Indonesian Journal of Electrical Engineering and Computer Science, 2019.

[14] O. Marbán, G. Mariscal and J. Segovia, Data Mining and Knowledge Discovery in Real Life Applications, 2009.

[15] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer and R. Wirth, "CRISP-DM 1.0 - Step-by-step data mining guide," 2000.

[16] S. B. Imandoust and M. Bilandraftar, "Application of K-Nearest Neighbor (KNN) Approach for Predicting Economic Events: Theoretical Background," International Journal of Engineering Research and Application, pp. 605-610, 2013.

[17] S. Slater, S. Joksimovic, V. Kovanovic and R. Baker, "Tools for Educational Data Mining: A Review," Journal of Educational and Behavioral Statistics, pp. 85-106, 2017.

[18] C. C. Aggarwal and C. Zhai, "A Survey of Text Classification," 2012.

[19] J. Han, M. Kamber and J. Pei, Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers, 2012.

[20] Defiyanti, J. Sofi and Mohamad, "Integrasi Metode Klasifikasi Dan Clustering dalam Data Mining," Konferensi Nasional Informatika (KNIF), 2015.

[21] A. J. Myles, R. N. Feudale, Y. Liu, N. A. Woody and S. D. Brown, "An introduction to decision tree modeling," Journal of Chemometrics, 2004.

[22] Boser, G. Bernhard E, V. Isabelle M and V. N, "A Training Algorithm for Optimal Margin Classifiers," Proceedings of the Fifth Annual Workshop on Computational Learning Theory, 1992.

[23] S. A. Ö. T. İ. Esra Mahsereci Karabulut, "A comparative study on the effect of feature selection on classification accuracy," Procedia Technology, 2012.

[24] G. Borboudakis and I. Tsamardinos, "Forward-Backward Selection with Early Dropping," Journal of Machine Learning Research 20, 2019.

[25] I. Syarif, A. Prugel-Bennett and G. Wills, "SVM Parameter Optimization Using Grid Search and Genetic Algorithm to Improve Classification Performance," Telecommunication Computing Electronics and Control, 2016.