ISSN 2278-3091
Volume 10, No.1, January - February 2021
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse521012021.pdf
https://doi.org/10.30534/ijatcse/2021/521012021

Educational Data Mining in Predicting Student Final Grades

William Willibrordus Damopolii[1], Nathan Priyasadie[2], Amalia Zahra[3]
Computer Science Department, BINUS Graduate Program - Master of Computer Science
Bina Nusantara University, Jakarta, Indonesia 11480
[1]william.damopolii@binus.ac.id, [2]nathan.priyasadie@binus.ac.id, [3]amalia.zahra@binus.edu

ABSTRACT

Educational data mining is a field of science that extracts knowledge from educational data. One of its implementations is to predict student performance, it helps teachers to identify students that need more support. This can potentially increase learning effectiveness and elevate overall student's grades. There are various algorithms and optimization solutions to predict student's performance. In this paper, we use real data from one of Indonesia's public junior high schools to compare naive bayes, decision tree, and k-nearest neighbor algorithms and implement feature selection and parameter optimization to identify which combination of algorithm and optimization can achieve the highest accuracy in predicting student grades, i.e. 7-grade classification. The results show that k-NN achieves the highest accuracy with 77.36%, where both feature selection and parameter optimization are applied.

Key words: Educational Data Mining, Knowledge Discovery, Student Academic Performance

1. INTRODUCTION

Nowadays, people are more connected to the Internet than ever. As a result of this technological revolution, the tremendous amount of information transactions over the Internet generates a tremendous amount of data. These data can be considered meaningful if we are able to extract the relevant information correctly, especially with the help of data mining.

Data mining is a method to find patterns and knowledge from a large amount of data. The process includes data preprocessing, data mining, pattern evaluation, and knowledge presentation. Data mining can be applied in various fields, such as business intelligence, health informatics, finance, and many others [1]. In this study, we will focus on the implementation of data mining in the field of education.

Educational data mining is an emerging interdisciplinary research field dedicated to researching and exploring methods to extract meaningful information from the massive data in educational environment [2]. Since information technology plays a big role in supporting the education field in the past decade, almost every institution stores their information inside a student information system [2]. This information includes student demographic, parent information, scores etc.

Applying data mining techniques to education processes will be meaningful in obtaining relevant trends, performance summaries and insights, which in turn might help students identify what aspect they need to improve. The aspect can be their academic performance, life cycle management, courses selection, measuring their retention rate, and the grant fund management of an institution [3]. One of the implementations of educational mining is to predict student's grades. Grades are essential components in every education field as they are calibers to reflect capability and performance of students in that educational institution. Predicting student final performance can also motivates the institution to create more effective teaching methods and a more conducive environment for the students [2]. Giving more support to students, who have been predicted earlier to have lower grades, can potentially increase learning effectiveness and elevate the overall student grade. In the end, having a good grade gives the student a bigger opportunity to get accepted in better higher education.

The objective of prediction is to estimate the unknown value of a variable that describes the student [2]. There are 2 ways to predict student performance, regression and classification. In this paper, we will only be focusing on classification. This study will be using real data from one of the public junior high schools in Indonesia SMPN 124 Jakarta, with a student's historical grade combined with student's sociodemographic variables to predict the student's final grade. There are many classification algorithms in data mining and this study will be comparing 3 different algorithms, Naive Bayes, Decision Tree, and K-Nearest Neighbors, with 2 data mining optimization methods, parameter optimization and feature selection, to find the best combination based on accuracy performance.

2. RELATED WORKS

Yadav et al [4]. conducted a study to compare multiple decision tree algorithm such as ID3, C4.5, and CART using 90 engineering student's data obtained from VBS Purvanchal University (Uttar Pradesh) on the sampling method for Institute Technology for session 2010 show that C4.5 algorithm can best classify student performance with 78.6% accuracy.

Chaudhari et al [5]. conducted a study to predict student performance acquired from SSBT College of Engineering and Technology, Jalgaon. Using various student behaviour variables such as number of library visits, hours spent on study, ability to time management etc. Naive bayes algorithm shows the best performance with 96% accuracy, compared to C-means algorithm at 95% and K-means algorithm 94%

Amrieh et al [6]. conducted a study to predict their academic achievement based on Kalboard 360 learning management system data, implementing a decision tree, artificial neural network, and naive bayes algorithm and using ensemble methods improve the performance of the classifier up to 22.1% and by utilizing student behaviour features increase the classifier accuracy up to 25.8% resulting in 82% accuracy in decision tree, 80% accuracy in artificial neural network, and 80% accuracy in naive bayes.

Hussain et al [7]. conducted a study to identify academically weak students using data from Digboi College, Duliajan College, and Doomdooma College. The study implemented deep learning using the sequential neural model with the adam optimization method. The study then also compared other classification methods such as the artificial immune recognition system v2.0 and adaboost. The highest accuracy achieved was 95.34% produced by deep learning technology.

Ahmad et al [8]. conducted a study to predict the student's academic performance of a first year bachelor student in computer science course. The data is collected from eight-year period intakes from July 2006/2007 until July 2013/2014 that contain student's demographic, previous academic record and family background information. Best prediction result was achieved by implementing the rule based algorithm with 71.3% accuracy followed by the decision tree with 68.8% and naive bayes with 67%.

Almarabeh [9] conducted a study to analyze student performance using classification techniques. The author used various data such as midterm score, student attendance, laboratory experiment, workshop and other factors to predict student final score. With Train-Test 80:20, Rule Based algorithms showed the best performance with 71.3% accuracy.

In 2019 Saa et al [10]. conducted a study to predict student academic performance using educational data mining. The author used student demographics, course instructor information, student general information, and student previous performance information from a private university in United Emirates Arab. Random Forest algorithm outperformed the other classifiers with 75.52% accuracy followed by Logistic Regression algorithm.

Yao et al [11]. conduct a study to predict secondary school students' final score using their personal data. The dataset consists of several variables such as parent information, student health condition, financial condition, attendance etc. With feature selection, the J48 algorithm showed the best result with 84.39% accuracy, while without feature selection OneR algorithm showed the best performance with 84.19% accuracy.

Rifat et al [12]. conducted a study to predict students' performance using student transcript data from a renowned university in Bangladesh. To predict students' final score, the authors used six state-of-the-art classification algorithms. The results showed that the Random Forest algorithm gave the best performance with 94.1% accuracy, followed by the Tree Ensemble algorithm.

3. METHODOLOGY

This study is conducted using naive bayes, decision tree, and k-nearest neighbor algorithms with RapidMiner software due to its extensive set of classification and optimization algorithm [13], the work focuses on comparing different algorithm's performance combined with feature selection and parameter optimization to find the best combination based on accuracy performance.

3.1 Data Collection

The first step in this research is data collection. We need to find and gather the right data for the algorithm, data may be scattered in different spreadsheets, databases, or websites. We used SMPN 124 Jakarta class of 2020 and 2019 with a total of 432 number of initial student data with 30 variables, which are entrance grades, gender, religion, type of living, transportation method, parent's education, parent's occupation, parent's income, and students' grades in 1st and 2nd semester. The grades in each semester include religion education, civic, bahasa indonesia, english, mathematics, natural science, social science, art and culture, and sports subject. After we gather all data, we join them into a single dataset.

3.2 Data Preprocessing

A. Remove Duplicates and Missing Values

After gathering all the data, we removed duplicates and missing value attributes from the data, resulting in 264 data left.

B. Convert Data

The next step is to convert numerical data into categorical data with Table 1 mapping.

Table 1: Mapping Rule

Numerical values	Categorical Values
>=95	A
90-94	B
85-89	C
80-84	D
75-79	E
70-74	F
<70	G

Which resulting in Table 2 student's attribute.

Table 2: Variables of Dataset

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
Religion	Student's Religion	Islam, Catholic, Christian
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 Degree, Master 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
1st and 2nd 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

3.3 Data Splitting

The data is split into 2, training and testing, with 80:20 ratio. The training data set is 80%, while the testing data set is 20% of total data. Training dataset will be used to train the algorithm in order to classify student data, while the testing data will be used to test the performance of the model trained.

3.4 Classification

Classification refers to a method of grouping of items based on qualitative information about one or more characteristics of the item, and grouping items according to a set of previously labelled items.

Classification aims to identify characteristics that indicate the group to which each case belongs. This pattern can be used to understand existing data and predict behavior of new data. Data mining creates classification models by examining already classified data (cases) and finding predictive pattern inductively. These existing cases may come from historical databases [17].

The data classified using following classification algorithms:

A. Naïve Bayes

Naive Bayes is one of the simplest and the most commonly used classifiers [18]. It assumes the conditional independence of a class, that is, given the class label of a tuple, it assumes that the values of the variables are conditionally independent of the others [1].

It is based on the application of Bayes' theorem to handle simple probabilistic classification. It assumes that the existence of a particular characteristic of a class is unrelated to the existence of any other characteristic, Even if these characteristics depend on the existence of another characteristic, the naive bayes classifier will treat all these variables as independently contributing to the possibility of classifying them into a specific class [19].

B. Decision Tree

Decision tree is a divide-and-conquer classification method. One of its advantages lies in interpretability of the constructed model. With this interpretability, information related to the identification of important features and relationships between classes can be used to support the design of future experiments and data analysis [20].

C. K-Nearest Neighbor

K- Nearest Neighbor is a technique for classifying elements by evaluating the k number of closest neighbors. An object is classified according to the majority votes of its neighbors, and the object is assigned to the most common category among its k nearest neighbors.

K-Nearest Neighbor has several main advantages such as simplicity, effectiveness, intuitiveness and competitive classification performance in many fields [17].

3.5 Feature Selection

Feature selection is one of the data preprocessing techniques in data mining to increase the data quality by minimizing the number of variables that need to be processed while maintaining the most relevant variable. It enhances classification accuracy, and learning runtime required. There are many feature selection algorithms, one of them is forward selection [14].

Forward selection is a method of adding variables to the model one at a time [15]. It starts with an empty selection of variables, and then adds every unused variable of the given data in each round. For each added variable, then calculate the performance. Only the variable with the highest performance improvement are added to the selection. Then start a new round with the modified selection [16].

3.6 Parameter Optimization

Parameter Optimization is one technique to increase the accuracy of the data mining algorithm by tuning its parameters based on testing dataset performance. It runs the algorithm with all the predefined parameter configuration possibilities, finding the best possible parameter configuration based on the algorithm's model accuracy.

The parameters that will be tuned in the decision tree are the criterion on which attribute will be selected for splitting, tree's maximal depth, pruning and minimal gain value. K-NN parameters that will be tuned are k value, measure type and weighted vote. The parameters that will be tuned in naive bayes are the laplace correction techniques.

3.7 Evaluation

Each algorithm will be evaluated based on accuracy measurement (1). Accuracy is the percentage of the correctly identified label, We used 4 variables True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) as presented in Figure 1. To calculate it, we first add the numbers of data that are correctly classified by classifier, divided by the total number of data classified, as illustrated in Equation (1).

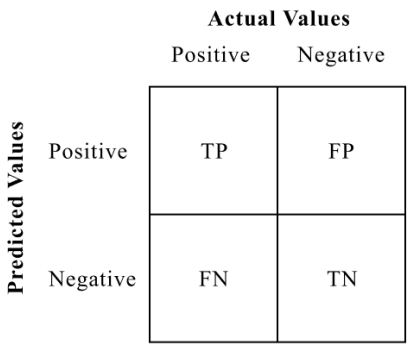

Figure 1 : Confusion Matrix

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

4. RESULT AND ANALYSIS

Table 3 summarizes the algorithm performance for naive bayes, decision tree, and k-nearest neighbor (K-NN), with 4 results of accuracy for each algorithm, without any optimization, using only features selection (FS), using only parameter optimization (PO), and with both optimization (FS+PO).

Table 3: Accuracy Result

Algorithm	Without Optimization	FS	PO	FS + PO
Naive Bayes	69.81%	75.47 %	73.47 %	76.92%
Decision Tree	62.26%	77.36 %	64.15 %	73.58%
K-NN	67.92%	69.81 %	69.81 %	**77.36%**

Naive bayes' best performance at 76.92% is achieved by applying parameter optimization by setting the laplace correction parameter to false and implementing feature selection which results in only using Bahasa Indonesia 2nd semester grade, natural science 2nd semester grade, english 2nd semester grade, gender and type of living variables.

Decision tree best performance at 77.36% is achieved by implementing parameter optimization, tuning the criterion based on gain ratio, set the maximal depth to 10, applying pruning and with 0.01 minimal gain.

K-NN best performance at 77.36% is achieved by using parameter optimization by tuning the k value to 100, set the KNN to nominal measure and weighted vote to false and applying feature selection that results in only using Bahasa Indonesia 1st semester grade, english 2nd semester grade, gender and mother's education variable.

In general, the experiment shows that feature selection and parameter optimization improve the accuracy of the classifier algorithm. However, when combined with optimization, it does not always result in better accuracy, such as that in the Decision Tree experimental result. Feature selection works better compared to parameter optimization in improving the accuracy due to the large amount of data variables. It also shows that various algorithms show different accuracy results, K-Nearest Neighbor with feature selection and parameter optimization shows the same result as decision tree with features selection display best accuracy value of 77.36%.

5. CONCLUSION

This study is designed to compare various classification algorithms and optimization to predict student's grade, we applied Naive Bayes, Decision Tree, and K-Nearest Neighbor with feature selection and parameter optimization. Among these algorithms, the best accuracy is achieved by K-Nearest Neighbor with feature selection and parameter optimization that show the same accuracy result with Decision Tree with feature selection at 77.36% accuracy. There are limitations on this study such as lack of data varieties, and the number of data processed due the original dataset containing a lot of missing values.

REFERENCES

1. J. Han, J. Pei and M. Kamber, ***DATA MINING: Concepts and Techniques***, Morgan Kaufmann, 2011.
2. 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), 40(6), pp. 601-618, 2010.
3. M. Goyal and R. Vohra, **Applications of Data Mining in Higher Education**, International Journal of Computer Science Issues (IJCSI) 9, no. 2, pp. 113, 2012.
4. S. K. Yadav and S. Pal, **Data Mining: A Prediction for Performance Improvement of Engineering Students using Classification**, World of Computer Science and Information Technology Journal, vol. 2, no. 2. pp. 51-56, 2012.
5. K. P. Chaudhari, R. A. Sharma, S. S. Jha and R. J. Bari, **Student Performance Prediction System using Data Mining Approach**, Int J Adv Res Comput Commun Eng 6, no. 3, pp. 833-839, 2017.
6. E. A. Amrieh, T. Hamtini and I. Aljarah, **Mining Educational Data to Predict Student's Academic Performance using Ensemble Methods**, International Journal of Database Theory and Application 9, no. 8, pp. 119-136, 2016.
7. S. Hussain, Z. F. Muhsion, Y. K. Salal, P. Theodoru, F. Kurtoğlu and G. C. Hazarika, **Prediction Model on Student Performance based on Internal Assessment using Deep Learning**, International Journal of Emerging Technologies in Learning (iJET) 14, no. 08, pp. 4-22, 2019.
8. F. Ahmad, N. H. Ismail and A. A. Aziz, **The Prediction of Students' Academic Performance using Classification Data Mining Techniques**, Applied Mathematical Sciences 9, no. 129, pp. 6415-6426, 2015.
9. H. Almarabeh, **Analysis of Students' Performance by using Different Data Mining Classifiers**, International Journal of Modern Education and Computer Science 9, no. 8, pp. 9, 2017.
10. A. A. Saa, M. Al-Emran and K. Shaalan, **Mining Student Information System Records to Predict Students' Academic Performance**, In International conference on advanced machine learning technologies and applications, pp. 229-239, 2019.
11. Y. Yao, Z. Chen, S. Byun and Y. Liu, **Using Data Mining Classifiers to Predict Academic Performance of High School Students**, Scientific Cyber Security Association (SCSA), pp. 18-35, 2019.

12. M. R. I. Rifat, A. Al Imran and A. S. M. Badrudduza, **Educational Performance Analytics of Undergraduate Business Students**, International Journal of Modern Education and Computer Science 11, no. 7, pp. 44, 2019.
13. S. Slater, S. Joksimović, V. Kovanovic, R. S. Baker and D. Gasevic, **Tools for Educational Data Mining: A Review,** Journal of Educational and Behavioral Statistics 42, no. 1, pp. 85-106, 2017.
14. A. G. Karegowda, A. S. Manjunath and M. A. Jayaram, **Comparative Study of Attribute Selection using Gain Ratio and Correlation Based Feature Selection**, International Journal of Information Technology and Knowledge Management 2, no. 2, pp. 271-277, 2010.
15. J. M. Sutter and J. H. Kalivas, **Comparison of Forward Selection, Backward Elimination, and Generalized Simulated Annealing for Variable Selection**, Microchemical journal 47, no. 1-2, pp. 60-66, 1993.
16. "Forward Selection (RapidMiner Studio Core)" [Online].Available: https://docs.rapidminer.com/latest/studio/operators/modeling/optimization/feature_selection/optimize_selection_forward.html [Accessed 20 January 2021]
17. S. B. Imandoust and M. Bolandraftar, **Application of K-Nearest Neighbor (KNN) Approach for Predicting Economic Events: Theoretical Background**, International Journal of Engineering Research and Applications 3, no. 5, pp. 605-610, 2013.
18. C. C. Aggarwal and C. X. Zhai, **A Survey of Text Classification Algorithms**, Mining Text Data, pp. 163-222, 2012.
19. S. A. Pattekari and A. Parveen, **Prediction system for Heart Disease using Naïve Bayes**, International Journal of Advanced Computer and Mathematical Sciences 3, no. 3, pp. 290-294, 2012.
20. A. J. Myles, R. N. Feudale, Y. Liu, N. A. Woody and S. D. Brown, **An Introduction to Decision Tree Modeling**, Journal of Chemometrics: A Journal of the Chemometrics Society 18, no. 6, pp. 275-285, 2004.