Comparison of Data Mining Classification Algorithms for Student Performance

Emny Harna Yossy
Computer Science Department, BINUS Online Learning, Bina Nusantara University, Indonesia 11480
Computer Science Department, BINUS Graduate Program - Doctor of Computer Science, Bina Nusantara University, Indonesia 11480
emny.yossy@binus.ac.id

Yaya Heryadi
Computer Science Department, BINUS Graduate Program - Doctor of Computer Science, Bina Nusantara University, Indonesia 11480
yayaheryadi@binus.edu

Lukas
Computer Science Department, BINUS Graduate Program - Doctor of Computer Science, Bina Nusantara University, Indonesia 11480
lookus@gmail.com

***Abstract*—Student performance has an important role to measure student quality. Student quality can be measured through predictions of student performance. Prediction can be done using data mining techniques. One technique that can be used is the classification method. The research aims to find out which classification model has the best performance related to student performance data. The data used is taken from UCI Machine Learning, namely student performance. The study used seven methods, namely K-nearest neighbor, classification and regression trees, naïve bayes, adaboost, extratree, bernaoulli naïve bayes, random forest. The technology used to compare the seven methods uses Python programming. Testing the performance of methods using cross validation. The results of this study are the comparison of student performance classification algorithms on student math, namely K-Nearest Neighboring of 86.52%, classification and regression tests of 86.08%, naïve bayes of 84.78%, adaboost of 88.04%, extratree of 81.30%, bernaoulli naïve bayes of 79.34%, random forest E of 87.82%, random forest G of 89.78%. Based on these results we know that the best classification method is the random forest G of 89.78%.**

Keywords—data mining, student performance, classification, algorithms.

I. INTRODUCTION

Along with the development of technology, especially the development of greater data storage. Data is a repository of information that can be used to analyze organizational needs. One organization that has a large data store is an educational organization. Educational organizations use data to get information, especially information about students. Student data has many attributes so that we can make predictions such as student performance predictions. Prediction can be done using data mining techniques.

Student performance has an important role to measure student quality. Student quality can be measured through predictions of student performance. Prediction can be done using data mining techniques. One technique that can be used is the classification method. Classification is grouping data based on characteristics that have similarities and differences[1].

Data mining in the field of education is not like data mining in general because the data hierarchy is different from other fields[2]. Data mining methods in education are classified into five dimensions, one of which is predictions such as predicting output values based on input data [3]. Predict data mining there are several data mining techniques using algorithms such as naïve bayes, decision trees, K-nearest neighbors, neural networks [4].

This study aims to determine the accuracy of the most accurate classification algorithm to measure student performance predictions. Predictions were made on mathematical student data in Portugal that researchers obtained from UCI Machine Learning datasets. Based on the description above, a comparison of classification data mining algorithms is performed on student performance.

II. LITERATURE REVIEW

A. Related Works

Cortez researches related to predicting secondary school student performance. The prediction is done using the Decision Trees algorithm, Random Forest, Neural Networks and Support Vector Machines. The result is the accuracy of student performance will be good if the first and second values are met[5]. Saa, et al, examined related to the identification of factors that influence student performance. The results of the study are the most common actors grouped in four main categories, namely student grades and prior student performance, student e-Learning activities, student demographics, and student social. The best algorithm prediction is to use decision tree algorithms, naïve bayes classifiers, and artificial neural networks[6].

Abazeed et al. Conducted a study of the predictions and evaluations of students who needed attention and corrective actions and looked for deviations before they occurred and which led to reduced performance and reduced failure rates. Prediction using random tree and the a priori algorithms. The results of the study are prediction and evaluation models[7]. Haris et al. Examined the area in which students were able to enrolment subjects. Prediction using Regression and Classification algorithms (Neural Networks, Bayesian Networks, Decision Trees, Support Vector Machines and Instance Based). The result is giving recommendations for the subjects to be chosen [8].

Edin et al. Conducted a study related to data mining on student performance using the decision tree algorithm and naïve bayes. The result is a demographic variable that influences student graduation[9]. Ahmed, et al, predict student performance using the decision tree algorithm to find out students who need special attention in order to successfully graduate. Prediction results are lists of students who are predicted to need special attention [10].

Khasanah, et al, examined student performance at the university level. This study aims to monitor student performance and prevent failure. The prediction is done using a linear regression algorithm and support vector machine. The results of this study are linear regression algorithms better than support vector machines[11].
Mueen et al. Examined student performance using the naïve bayes algorithm, neural network, and decision tree to predict final exam scores [12]. Kabakchieva, examines student performance related to Bulgarian University using the Decision Tree to predict the final grades of students majoring in information systems[13]. Yadav, et al., Predicted student performance for final exam scores using decision tree algorithms [14].
Based on the related works, then data mining classification algorithm in this study uses seven models including:

1. K-Nearest Neighbour
2. Classification and Regression Trees (CART)
3. Naïve Bayes
4. AdaBoost
5. ExtraTree
6. Bernaulli Naïve Bayes
7. RandomForest

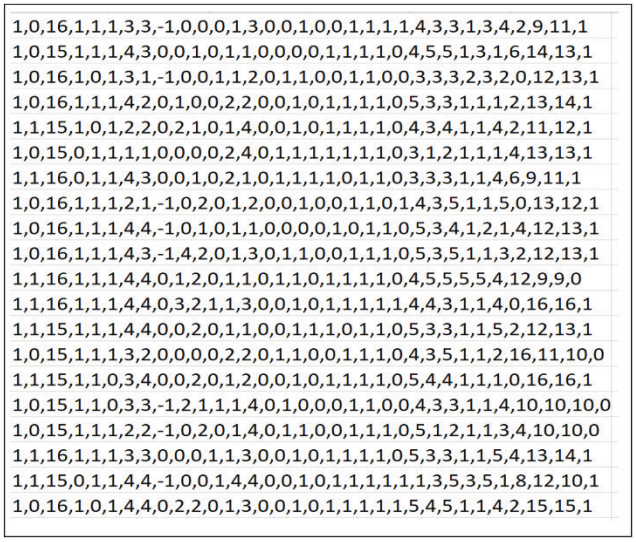
B. *Python*

In this research, technology is using Python language [15]. Python is using Anaconda-Spyder as a platform [16]. Program supported by libraries, among others: pandas, numpy, from sklearn import metrics, model_selection, classification_report, accuracy_score KNeighborsClassifier, DecisionTreeClassifier, GaussianNB, BernoulliNB, auc, precision_recall_curve, roc_curve, RFE, LinearDiscriminantAnalysis, from sklearn.feature_selection import, confusion_matrix, RandomForestClassifier, AdaBoostClassifier, ExtraTreesClassifier, PCA, dan from matplotlib import pyplot[17].

III. RESEARCH METHOD

A. *Data*

Data retrieved from:
https://archive.ics.uci.edu/ml/datasets/student+performance.
Data sources come from: http://www3.dsi.uminho.pt/pcortez. Paulo Cortez, University of Minho, GuimarÃ£es, Portugal[5].
The data taken is related to the achievement of secondary education of two Portuguese schools. The data attributes include student grades, demographic, social and school features. In this study the data used are student math data. The data consisted of 394 students. Student math data provided is data in the form of text and numerics. To facilitate the data mining process, researchers convert data into numerics, as follows:



```
1,0,16,1,1,1,3,3,-1,0,0,0,1,3,0,0,1,0,0,1,1,1,1,4,3,3,1,3,4,2,9,11,1
1,0,15,1,1,1,4,3,0,0,1,0,1,1,0,0,0,0,1,1,1,1,0,4,5,5,1,3,1,6,14,13,1
1,0,16,1,0,1,3,1,-1,0,0,1,1,2,0,1,1,0,0,1,1,0,0,3,3,3,2,3,2,0,12,13,1
1,0,16,1,1,1,4,2,0,1,0,0,2,2,0,0,1,0,1,1,1,1,0,5,3,3,1,1,1,2,13,14,1
1,1,15,1,0,1,2,2,0,2,1,0,1,4,0,0,1,0,1,1,1,1,0,4,3,4,1,1,4,2,11,12,1
1,0,15,0,1,1,1,1,0,0,0,0,2,4,0,1,1,1,1,1,1,1,0,3,1,2,1,1,1,4,13,13,1
1,1,16,0,1,1,4,3,0,0,1,0,2,1,0,1,1,1,1,0,1,1,0,3,3,3,1,1,4,6,9,11,1
1,0,16,1,1,1,2,1,-1,0,2,0,1,2,0,0,1,0,0,1,1,0,1,4,3,5,1,1,5,0,13,12,1
1,0,16,1,1,1,4,4,-1,0,1,0,1,1,0,0,0,0,1,0,1,1,0,5,3,4,1,2,1,4,12,13,1
1,0,16,1,1,1,4,3,-1,4,2,0,1,3,0,1,1,0,0,1,1,1,0,5,3,5,1,1,3,2,12,13,1
1,1,16,1,1,1,4,4,0,1,2,0,1,1,0,1,1,0,1,1,1,1,1,0,4,5,5,5,5,4,12,9,9,0
1,1,16,1,1,1,4,4,0,3,2,1,1,3,0,0,1,0,1,1,1,1,1,4,4,3,1,1,4,0,16,16,1
1,1,15,1,1,1,4,4,0,0,2,0,1,1,0,0,1,1,1,0,1,1,0,5,3,3,1,1,5,2,12,13,1
1,0,15,1,1,1,3,2,0,0,0,0,2,2,0,1,1,0,0,1,1,1,0,4,3,5,1,1,2,16,11,10,0
1,1,15,1,1,0,3,4,0,0,2,0,1,2,0,0,1,0,1,1,1,1,0,5,4,4,1,1,1,0,16,16,1
1,0,15,1,1,0,3,3,-1,2,1,1,1,4,0,1,0,0,0,1,1,0,0,4,3,3,1,1,4,10,10,10,0
1,0,15,1,1,1,2,2,-1,0,2,0,1,4,0,1,1,0,0,1,1,1,0,5,1,2,1,1,3,4,10,10,0
1,1,16,1,1,1,3,3,0,0,0,1,1,3,0,0,1,0,1,1,1,1,0,5,3,3,1,1,5,4,13,14,1
1,1,15,0,1,1,4,4,-1,0,0,1,4,4,0,0,1,0,1,1,1,1,1,1,3,5,3,5,1,8,12,10,1
1,0,16,1,0,1,4,4,0,2,2,0,1,3,0,0,1,0,1,1,1,1,1,5,4,5,1,1,4,2,15,15,1
```

Fig. 1. Data of Student Math.

B. *Methods*

In this study, researchers used seven methods to model comparations classifications of data. The model used is K-Nearest Neighbor, CART, Naïve Bayes, AdaBoost, ExtraTree, Bernaulli Naïve Bayes, RandomForest consists of Random forest E and random forest G. The technology used to compare the models is Python. The tools used are Spyder. The results of the comparison are described in the form of area under curve graphic. The research stages are as follows:

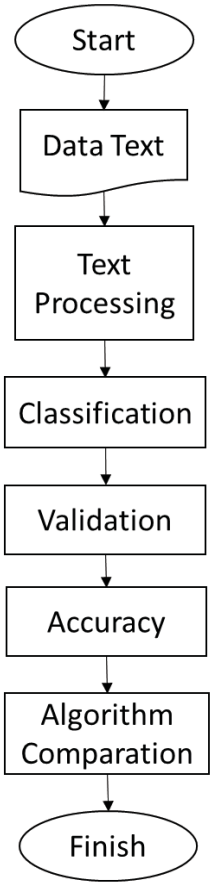


Fig.2. Research stages.

Research stages include:

1. Data Text
   Data is taken from UCI Machine Learning dataset. Data on student performance. The data processed is student math data consisting of 33 attributes. Data in the form of text and numeric.
2. Text Processing
   After the data is obtained, text processing begins by converting the data into numeric form and saving it in format.csv.
3. Classification
   After the data is processed, the researchers conducted the classification using seven classification methods.
4. Validation
   Then, researchers validate by applying cross validation.
5. Accuracy
   Next, the researcher made an accuracy calculation in the form of precision and recall.
6. Algorithm comparation
   Next the researchers performed the algorithm compation model classification described in the plot.

IV. Result and Discussion

A. Implementation

The implementation starts from text processing. Text converted to numeric. Then the data is made into an array of 33 values according to the many attributes of the data. Arrays are taken in two dimensions, the values of X and Y.
Next, the system validates in a way to apply Cross Validation with Model Comparations Classifications, as follows:

	school	sex	age	...	G1	G2	G3
count	460.000000	460.000000	460.000000	...	460.000000	460.000000	460.000000
mean	0.723913	0.395652	16.789130	...	11.528261	11.702174	0.713043
std	0.447547	0.489523	1.179477	...	2.578034	2.740083	0.452834
min	0.000000	0.000000	15.000000	...	5.000000	0.000000	0.000000
25%	0.000000	0.000000	16.000000	...	10.000000	10.000000	0.000000
50%	1.000000	0.000000	17.000000	...	11.000000	11.000000	1.000000
75%	1.000000	1.000000	18.000000	...	13.000000	13.000000	1.000000
max	1.000000	1.000000	22.000000	...	18.000000	19.000000	1.000000

[8 rows x 33 columns]
KNN: 0.865217
CART: 0.860870
NB: 0.847826
AdaBoost: 0.880435
ExtraTree: 0.813043
BernoulliNB: 0.793478
RandomForestE: 0.878261
RandomForestG: 0.897826
AUC CART=0.942

Fig. 3. Apply Cross Validation with Model Comparations Classifications

Validation is done by calculating the amount of data, the mean, standard deviation, minimum, 25%, 50%, 75% and maximum. From these values a classification comparison calculation is performed so that the accuracy of the algorithm is obtained.
It can be seen from the pictures above that the results of comparison of student performance classification algorithms on student math are K-Nearest Neighboring of 86.52%, classification and regression tests of 86.08%, naïve Bayes of 84.78%, adaboost of 88.04 %, extratree of 81.30%, bernaoulli naïve bayes of 79.34%, random forest E of 87.82%, random forest G of 89.78%. Based on these results we know that the best classification algorithm is the random forest G of 89.78%. Grafic comparations classifications models as follows:

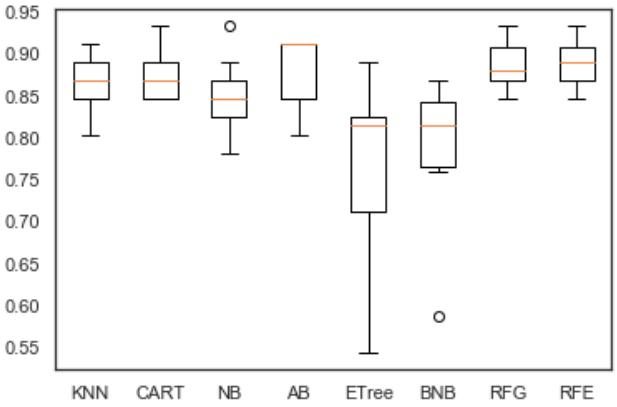


Fig.4. Algorithms Comparison.

The next stage is scoring in the under curve area from the comparison algorithm classification results. Scoring is done to get the value of precision, recall and f-measure and support. This is following scoring with Area Under Curve (AUC) from classifications models:

	precision	recall	f1-score	support
0	0.74	0.80	0.77	132
1	0.92	0.89	0.90	328
micro avg	0.86	0.86	0.86	460
macro avg	0.83	0.84	0.84	460
weighted avg	0.87	0.86	0.86	460

Fig. 5. Scoring with AUC from Classifications Models.

From the results of the scoring can be described Area Under Curve Graphic from classification methods as follows:

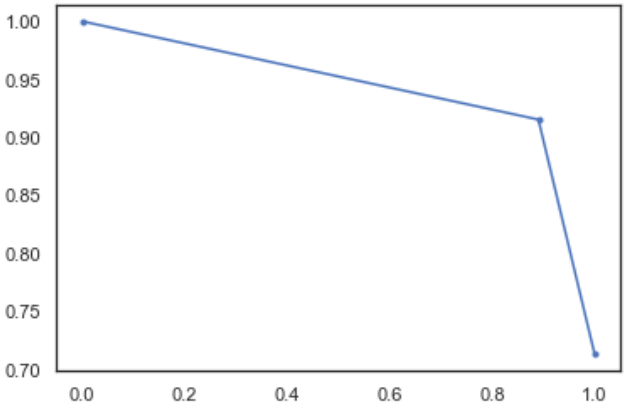


Fig.6. Graphic from AUC.

V. CONCULSIONS

The result of the study shows the best classification algorithm for student performance with the student math data is the random forest G of 89.78%. In addition, three algorithm which have the best performance are random forest, adaboost dan K-Nearest Neighboring. In the future, this research can improve to research more than seven algorithms, can research about the most influential features for student performance.

ACKNOWLEDGEMENT

This paper is used as part of Soft Computing subjects, Doctoral Computer Science, Bina Nusantara University. Thank you to Bina Nusantara University for funding from this paper.

REFERENCES

[1] T. Pang-Ning, M. Steinbach, and V. Kumar, "Introduction to data mining: Solution Manual," *Libr. Congr.*, p. 796, 2006.

[2] R. S. J. d. Baker, "International Encyclopedia of Education (3rd edition)," in *Data Mining for Education*, In McGaw, B., Peterson, P., 2010, pp. 112–118.

[3] S. N. U Manzoor, "An agent based system for activity monitoring on network-ABSAMN," *Expert Syst. Appl.*, vol. 8, no. 36, pp. 10987–10994, 2009.

[4] J. Han, *Data Mining: Concepts and Techniques (The Morgan Kaufmann Series in Data Management Systems)*. 2011.

[5] P. Cortez, "Using Data Mining to Predict Secondary School Student Performance," *University of Minho*. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/student+performance.

[6] A. Abu Saa, M. Al-Emran, and K. Shaalan, *Factors Affecting Students' Performance in Higher Education: A Systematic Review of Predictive Data Mining Techniques*, no. 0123456789. Springer Netherlands, 2019.

[7] A. Abazeed and M. Khder, "A Classification and Prediction Model for Student ' s Performance in University Level," no. 1993, 2017.

[8] N. A. Haris and F. A. Rahman, "A Study on Students Enrollment Prediction using Data Mining," no. c, 2016.

[9] E. Osmanbegovic and M. Suljic, "Data Mining Approach for Predicting Student Performance," – *J. Econ. Bus.*, 2012.

[10] A. Badr, E. Din, and I. S. Elaraby, "Data Mining : A prediction for Student ' s Performance Using Classification Method," vol. 2, no. 2, pp. 43–47, 2014.

[11] A. U. Khasanah, "Educational Data Mining Techniques Approach to Predict Student ' s Performance," vol. 9, no. 2, pp. 115–118, 2019.

[12] A. Mueen, B. Zafar, and U. Manzoor, "Modeling and Predicting Students' Academic Performance Using Data Mining Techniques," *Int. J. Mod. Educ. Comput. Sci.*, vol. 8, no. 11, pp. 36–42, 2016.

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

[14] L. X. Wang, "The MW Method Completed: A Flexible System Approach to Data Mining," *IEEE Trans. Fuzzy Syst.*, vol. 11, no. 6, pp. 68–782, 2003.

[15] "Python." [Online]. Available: https://www.python.org/.

[16] "Anaconda." [Online]. Available: https://www.anaconda.com/.

[17] "Scikit-Learn." [Online]. Available: https://scikit-learn.org/.