Comparison of Data Mining Algorithm Performance on Student Savings Dataset

Y D P Negara¹, A F Doni²

¹,²Lecturer, Dept. of Informatics Engineering. University of Trunojoyo Madura
Bangkalan, Indonesia

Yudha.putra@trunojoyo.ac.id, doni.fatah@trunojoyo.ac.id

Abstract. Sabilillah Educational Cash Unit is a unit within the sabilillah educational foundation which is engaged in education. The stored cash processing data will be utilized using data mining so that it can be used as a decision support for finding information that is useful in evaluating the data used. Various methods contained in the data mining, the authors will make a comparison of the method techniques from the data mining. The use of the decision tree and K-means method is implemented using the Rapid Miner application, which will later be analysed of each of these methods to determine the strategy to look for students who have the potential to save the hajj savings. This research was conducted with a group of data to determine the percentage value of precision, recall and accuracy. The results of this study that the C.4.5 method has a better value than other methods on the recall and accuracy, while the K-Means method has a precision value better than the other methods.

1. Introduction

Sabilillah Education Cash Unit is a unit within the sabilillah education foundation located on Jl. Rajawali III Sampang and is engaged in education. The Sabilillah Education Cash Unit must meet the needs of students every day in managing cash and are required to be able to take the right decisions in determining cash management strategies. To be able to do this, the Unit needs an adequate source of information for further analysis. In the Cash Unit of Education, a lot of transactions are recorded. The availability of very large sales data tends only to be used as archives and reports and has not been optimally utilized because there is no decision support system and methods that can be used to design a business strategy to increase the amount of cash. By conducting data mining, there are several methods that can be used to make comparisons in the above cases, including using the id3 algorithm, and K-Means so that we know what techniques are suitable to be applied so that the level of accuracy of students is diligently saving high.

2. Literature Review
2.1. Related Works
This study refers to relevant previous research. There are some results from previous studies as a reference basis for this related research. Researchers refer to several studies that have topics related to the author's research. Previous studies discussing the Comparative Analysis of Decision Tree Algorithms focus on comparison of different decision tree algorithms for data analysis[6]. The results of this study, Random Forest provides better predictive results. Other research conducted is Predicting Student Performance in Higher Education Institutions Using Decision Tree Analysis. The results show that the use of decision tree analysis can improve academic performance and improve institutional success[7]. Other research that is carried out presents the model based on the decision tree algorithm and suggests the best algorithm based on performance. Three classifiers were built (J48, Random Tree and REPTree). The result is J48 algorithm is considered as the best algorithm based on its performance compared to Random Tree and RepTree algorithms[8]. Other studies compare between K-means Algorithms and EM Clustering using improved elementary school student attendance performance[9]. Other studies discuss classifying student data into clusters based on similarity of data using the K-Means Clustering algorithm, the results obtained can contribute to improving data mining significantly to support strategic promotion in getting new prospective students[10].

3. Methods
Research methodology begins with observation. The flow of research begins with observation and data retrieval (raw data), processing raw data into data sets, cross validation testing of each model, evaluating and comparing performance.

The method used in this study is C4.5 and K-means. Some of the most popular algorithms today are K-Means and C.45. K-Means Clustering is a popular clustering algorithm with local optimization[11][12], while the popular C.45 algorithm used in several other studies provides a high degree of accuracy[13][14].

![Research Methodology](image)

Figure 1. Research Methodology[15]

4. Result
The purpose of this study is to compare methods in data mining for Sampang Sabilillah cash data units using predetermined variables. Data analysis for the selection of attributes needed in conducting research is explained in table 1.

| Table 1. Data Atribut |
Atribut | Variable
---|---
Hajj savings program | Y
Magnitude of Savings | X1
Total Savings | X2
Id_member | X3

4.1. Results of Decision Tree Implementation on Rapidminer

In this data mining implementation to analyze the determination of the Sampang sabilillah cash unit data classification using the Rapid Miner 5 application. In the analysis of the decision tree (C4.5). The design process using the Decision Tree method using the RapidMiner 5 application is shown in Figure 2. The results of Figure 1 are then executed and viewed using Meta Data seen in Figure 3.

![Decision Tree Design](image)

**Figure 2.** Decision Tree Design.

![RapidMiner results on the Decision Tree method](image)

**Figure 3.** RapidMiner results on the Decision Tree method

4.2. Results of K-Means Implementation on Rapidminer
In the implementation of data mining to analyze the determination of the Sabilillah Sampang cash unit data classification using the Rapid Miner 5. application in the K-Means analysis. The design process with K-Means using the RapidMiner 5 application is shown in Figure 4. The results of Figure 1 are then executed and viewed using the Scatter Plot seen in Figure 5.

Figure 4. RapidMiner results on the K-means method

Figure 5. RapidMiner results on the K-means method

Table 2. Comparison Data

| Method   | Precision | Recall | Accuracy |
|----------|-----------|--------|----------|
| C 4.5    | 0 %       | 100%   | 46.22%   |
| K-Means  | 40.30%    | 37.51% | 39.40%   |

Table 2 shows that C.4.5 is a method that has a higher level of Recall and accuracy than the other methods, then the Precision K-Means has a higher value than the other methods.

5. Conclusion

Based on the results of the study, there are several conclusions are Based on the results of a comparison of the two methods, namely the Decision Tree and K-means, the most influential variable on the priority outcomes of the Hajj program is the Savings Number variable. Based on the Recall and accuracy values, the C.4.5 method has a higher value than other algorithms with a precision value of 100% and an
accuracy value of 46.22%. Based on the value of precision, K-Means has a higher value compared to other methods with a recall value of 39.40%

References
[1] D. M. B. Tarigan, D. Palupi Rini, and Sukemi, “Particle swarm optimization-based on decision tree of c4.5 algorithm for upper respiratory tract infections (URTI) prediction,” J. Phys. Conf. Ser., vol. 1196, no. 1, pp. 0–7, 2019.
[2] A. Abu, “Educational Data Mining &amp; Students’ Performance Prediction,” Int. J. Adv. Comput. Sci. Appl., vol. 7, no. 5, pp. 212–220, 2016.
[3] R. A. Rustia, M. M. A. Cruz, M. A. P. Burac, and T. D. Palaoag, “Predicting Student’s Board Examination Performance using Classification Algorithms,” pp. 233–237, 2018.
[4] M. H. I. Shovon, “Prediction of Student Academic Performance by an Application of K-Means Clustering Algorithm,” Int. J. Adv. Res. Comput. Sci. Softw. Eng. Res., vol. 2, no. 7, pp. 846–849, 2012.
[5] A. Dutt, “Clustering Algorithms Applied in Educational Data Mining,” Int. J. Inf. Electron. Eng., no. April, 2014.
[6] S. Sathyadevan and R. R. Nair, “Comparative Analysis of Decision Tree Algorithms: ID3, C4.5 and Random Forest,” vol. 410, pp. 549–562, 2016.
[7] A. K. Hamoud, A. S. Hashim, and W. A. Awadh, “Predicting Student Performance in Higher Education Institutions Using Decision Tree Analysis,” Int. J. Interact. Multimed. Artif. Intell., vol. 5, no. 2, p. 26, 2018.
[8] V. Mhetre and M. Nagar, “Classification based data mining algorithms to predict slow, average and fast learners in educational system using WEKA,” Proc. Int. Conf. Comput. Methodol. Commun. ICCMC 2017, vol. 2018-Janua, no. Iccmc, pp. 475–479, 2018.
[9] M. Singh, H. Nagar, and A. Sant, “K-mean and EM Clustering algorithm using attendance performance improvement Primary school Student,” Jisrset, vol. 2, no. 1, pp. 131–133, 2016.
[10] S. Abadi et al., “Application model of k-means clustering: insights into promotion strategy of vocational high school,” Int. J. Eng. Technol., vol. 7, no. 2.27, p. 182, 2018.
[11] E. Jukes, “Encyclopedia of Machine Learning and Data Mining (2nd edition),” Ref. Rev., vol. 32, no. 7/8, pp. 3–4, 2018.
[12] N. Dhanachandra, K. Manglem, and Y. J. Chanu, “Image Segmentation Using K-means Clustering Algorithm and Subtractive Clustering Algorithm,” Procedia Comput. Sci., vol. 54, pp. 764–771, 2015.
[13] P. Sharma, K. Saxena, and R. Sharma, “Computational Intelligence in Data Mining—Volume 1,” vol. 410, no. Cvd, pp. 285–294, 2016.
[14] C. Anuradha and T. Velmurugan, “Performance Evaluation of Feature Selection Algorithms in Educational Data Mining,” Int. J. Data Min. Tech. Appl., vol. 5, no. 2, pp. 131–139, 2017.
[15] R. Sumitha, E. S. Vinothkumar, and P. Scholar, “Prediction of Students Outcome Using Data Mining Techniques,” Int. J. Sci. Eng. Appl. Sci., vol. 2, no. 6, pp. 132–139, 2016.

1. The first section in your paper
The first paragraph after a heading is not indented (Bodytext style).
Other paragraphs are indented (BodytextIndented style).

2. Another section of your paper
The first paragraph after a heading is not indented (Bodytext style).
Other paragraphs are indented (BodytextIndented style).
2.1. A subsection
Some text.

2.1.1. A subsubsection. The paragraph text follows on from the subsubsection heading but should not be in italic.

References
[1] A reference
    This reference has two entries but the second one is not numbered (it uses the ‘Reference (no number)’ style.
[2] Another reference
[3] More references