IJOBIT
International Journal of Business and Information Technology

Copyright © The Author(s)
Vol. 2, No. 2, Desember 2020

IMPLEMENTATION OF THE DECISION TREE MODEL ON MACHINE LEARNING TO PREDICT POTENTIAL NEW STUDENTS

Ade Onny Siagian[1], Haudi[2]
[1] Universitas Bina Sarana Informatika
[2]STAB Dharma Widya
[1]ade.aoy@bsi.ac.id, [2]haudi@stabdharmawidya.ac.id

Article Information
Received: Oktober 2, 2021
Revised: November 6, 2021
Online: Desember 25, 2021

Keywords
Prediction, Machine learning, Python Programming, Decision Tree

ABSTRACT

According to a previous study, “Implementation of Naïve Bayes Classifier-based Machine Learning to Predict and Classify New Students at Matana University” has an accuracy of 0.73 or 73%. This is not optimal, the accuracy needs to be improved. In this research, to increase accuracy by using a different model, namely the Decision Tree model. The reason for choosing the Decision Tree is that there are not many predictors used (4 predictors) and can be used for classification or prediction. The 4 predictors are frequency, position, majors of students in SMA/K, and research programs of interest. The target is the entry status of prospective students. The research procedures that were tried were information gathering, pre-processing, machine learning processes with the Decision Tree model and visualization of the results. The programming language used is Python. The results of this Decision Tree show changes, through 10 executions the average accuracy of the ratio of training information and test information is 70: 30 of 0.727 or 72.7% (lowest accuracy is 47% and highest is 87%), for a ratio 80: 20 of 0, 73 or 73% (the lowest accuracy is 60% and the highest is 90%). Thus, the results of the Decision Tree model on average have not exceeded the results of the Naïve Bayes Classifier model. Further research, increase the amount and alteration of information, reduce the difference in results each time the model is executed, try other models, and improve the application ready to use.

INTRODUCTION

According to Gartner research, by 2022 it is estimated that artificial intelligence (AI) has the potential to be worth $3.9 trillion in business. (Gartner, 2018). This makes artificial intelligence, including machine learning, placed in a more strategic position in the business. Marketing is a business unit that is affected by the advancement of artificial intelligence (Sterne, 2017). The application of artificial intelligence and machine learning in the business world is certainly not arbitrary. One of the most important things in this application is accuracy. The higher the accuracy, the more reliable it is. This is very supportive in making business decisions.

In a previous study, entitled "Implementation of Naïve Bayes Classifier-based Machine Learning to Predict and Classify New Students at Matana University", obtained an accuracy of 0.73 or 73%

(Simon, 2021). The accuracy is quite high, it is possible to improve the accuracy. The way to improve the accuracy is by changing the model used. In this study, try to apply the Decision Tree model.

Decision tree is a form of decision making whose hierarchy imitates a tree, has roots, branches and leaves. It is also used as a learning technique in artificial intelligence with the term Decision Tree Learning (DTL). According to Suyanto, DTL is a "machine learning technique that builds a hierarchical sequential structure classification rule representation by recursively partitioning the training data set" (Suyanto, 2018). Classification and regression problems can be solved by Decision Tree. This study chose Decision Tree because the predictors used were not many (four predictors), namely frequency, JaBoDeTaBek, major and study program and can be used for classification or prediction. The implementation of the Decision Tree is done through the creation of a program. Through this Decision Tree-based program, it is hoped that it can simplify simulation, calculation and visualization, speed up the acquisition of results, reduce calculation errors, compared to without the program.

Python is a popular programming language today, especially for the development of machine learning, NLP and neural networks (Analytics Insight, 2021). In addition, Python is also widely used in data science. Python has many libraries, such as Numpy, Pandas, Sklearn, Matplotlib and many more. This is the reason Python was chosen to predict potential new students with Decision Tree.

METHODS

This research method consists of four stages, namely data collection, pre-processing, machine learning process with Decision Tree and visualization of results, Figure 1.

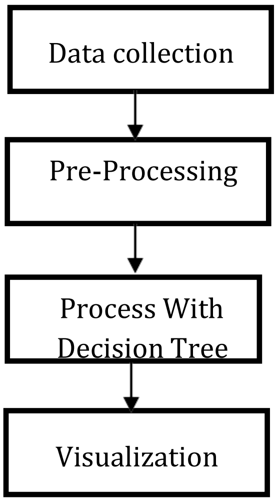

Figure 1. Research Method

Research data obtained from the results of previous studies, in the form of a table with comma separated values (csv) format. The file consists of four columns for a predictor and a target. Columns as predictors are frequency, JaBoDeTaBek, majors and study programs. Column as target is login status. The format of the dataset can be seen in Table 1. The encoding for the study program can be seen in Table 2.

Table 1. Data Set Format

Frequency	JaBoDeTaBek	Major	Study Program	login status

Table 2. Coding Of Study Programs

Code	Study Program Name
1	Architecture
2	Accountancy
3	Visual communication design
4	Physics
5	Hospitality and Tourism
6	Management
7	Information Systems
8	Computer system
9	Statistics
10	Technical Information

Pre-processing is done through the program using the Python programming language. The libraries used are Numpy, Pandas and Sklearn (Scikit-Learn). Coding is done for the frequency predictor variable, JaBoDeTaBek, majors and study programs. Duplicate data (duplicate data) is eliminated.

At the stage of the process is done by applying the Decision Tree. Comparison of training data and test data in the process is 70:30 and 80:20. In Python, the Sklearn library uses the Classification And Regression Tree (CART) algorithm to train the Decision Tree model (Géron, 2019). Criterion used is entropy (a measure of how random a group of data is).

Visualization is the last stage in this research method. At this stage the results or outputs of the process are visualized in the form of graphs so as to facilitate understanding and decision making.

RESULTS AND DISCUSSION

The application of the Decision Tree model in machine learning to predict potential new students has been carried out in a program with Python as the programming language. The program carried out ten executions of the Decision Tree model with two different ratios of training data and test data, namely 70:30 and 80:20. The results of the ten executions can be seen in Table 3.

Table 3. Results of the Decision Tree process

No	70:30	80:20
1	0,73	0,8
2	0,8	0,7
3	0,87	0,7
4	0,8	0,9
5	0,67	0,6
6	0,87	0,9
7	0,87	0,6
8	0,6	0,6
9	0,47	0,7
10	0,6	0,8
x̄	**0,727**	**0,73**

From the results of Table 3, it can be seen that the averages are not much different between the ratios of 70:30 and 80:20, namely 0.727 and 0.73. However, if you look at the top and bottom accuracy levels, it can be seen that the 70:30 ratio has the lowest accuracy of 0.47 and the top 0.87, while the 80:20 ratio has the lowest accuracy of 0.6 and the top accuracy of 0.9.

The visualization of these results can be seen in Figure 2. From the visualization It can be seen that the eighth executions have the same results, while the first, fifth and sixth each ratio is not much different. As for the far different results, it can be seen in the seventh and ninth executions.

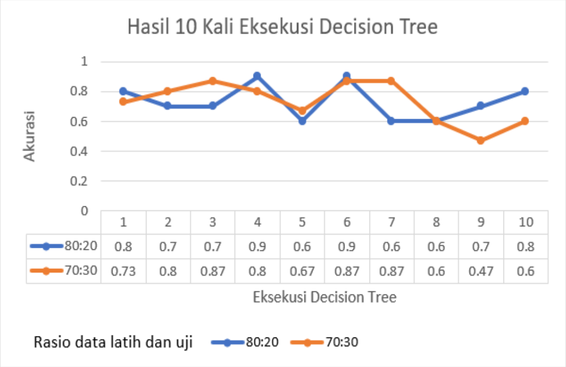

Figure 2. Visualization of results

CONCLUSION

In previous studies, the Naïve Bayes Classifier model with an accuracy of 0.73 or 73%, is not much different from the average obtained through the Decision Tree model, both ratios of 70:30 and 80:20. Here it can be seen that the Decision Tree model is not stable where the training data changes, the results or outputs can change. Although the accuracy has been up to 0.9 or 90% (80:20 ratio), but it has also been up to 0.47 or 47% (70:30 ratio), with an average of 0.727 or 72.7% (70:30 ratio) and 0.73 or 73% (80:20 ratio). Thus, the overall accuracy of the Decision Tree model is not stable and on average it has not exceeded the Naïve Bayes Classifier model. Further research, increasing the amount and variety of data, finding solutions to reduce the difference in results each time the model is executed, using other models, and develop ready-to-use applications.

ACKNOWLEDGMENT

The author would like to thank the relevant parties who have provided material and moral support in writing this article. We also thank the IJOBIT Journal for providing the opportunity to share knowledge that will be useful in the future.

REFERENCES

Analytics Insight. (2021). What Are The Best Programming Languages For Artificial Intelligence, https://www.analyticsinsight.net/what-are-the-best-programming-languages-for-artificial-Intelligence/

Barus, S.P. (2021). Implementation of Naïve Bayes Classifier-based Machine Learning to Predict and Classify New Students at Matana University. Journal of Physics: Conference Series 1842 (1),012008

Gartner. (2018). Gartner Says Global Artificial Intelligence Business Value to Reach $1.2 Trillion in 2018. https://www.gartner.com/en/newsroom/press-releases/2018-04-25-gartner-says-global-artificial- intelligence-business-value-to- reach-1-point-2-trillion-in-2018

Géron, A. (2019). Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow. O'Reilly Media

Siagian, AO, Gunartin, K Nufus, HS Nur'aini Yusuf, A Maddinsyah, A Muchtar. Journal of A Systematic Literature, Review of Education Financing Model in Indonesian School. 2020.

Siagian, AO, Contribution of Inventory Accounting Systems in Improving Inventory Internal Control, Journal of Social Science, 1 (2), 2020 pp. 1-6. http://jsss.co.id/index.php/jsss/article/view/12

Sterne, J. (2017). Artificial Intelligence for Marketing, Practical Applications. John Wiley & Sons

Suyanto. (2018). Machine Learning: Basic and Advanced Levels. Bandung: Informatics.

Sugiyono. Quantitative Research Methods, Qualitative, and R & D." Bandung: Alfabeta, 2017.