Jurnal Mantik, 5 (4) (2022) 2688-2693

Published by:Institute of Computer Science (IOCS)

Jurnal Mantik

Journal homepage: www.iocscience.org/ejournal/index.php/mantik/index

Decision Tree Model For Predicting Work Schedules Using Scikit-Learn

Siti Hapsoh Lubis[1], Muhammad Iqbal[2]

[1,2]Computer System, Sains and Teknologi, Universitas Pembangunan Panca Budi Medan,
Jl. Gatot Subroto KM. 4,5 Medan, 20122, Indonesia

Email : hapsohs307@agmail.com, muhammadiqbal@dosen.pancabudi.ac.id

A R T I C L E I N F O

Article history:

Received: Jan 09, 2022
Revised: Jan 14, 2022
Accepted: Feb 16, 2022

Keywords:

Decision Tree,
Classification,
Scikit-Learn,
Python

A B S T R A C T

Predicting category and numerical data, such as working schedule data, is difficult since it necessitates a specific process. A decision tree is one of many categorization methods that can handle both category and numerical input. Scikit learn, a python library that may be used for decision trees, is one example. Although Scikit-optimized learn's CART algorithm could only handle numerical data, it did provide certain features to deal with categorical data. To forecast working schedules, this study used scikit-learn to create a decision tree model. There are 54 variables, three of which are category and one of which is numerical. A 6-depth decision tree model was created as a result of the implementation. The evaluation yielded a positive outcome, with accuracy and precision above 0.7 and 0.9, respectively. The optimal division of data is 30% validation and 70% training. In comparison to KNN, the decision tree model has higher accuracy, with decision tree accuracy exceeding 0.8 while KNN accuracy is below.

1. Introduction

Data and information cannot be separated from multimedia in today's technological advancements. Data and information are no longer only presented as text, but also as images, audio, and video. The image is one of the components of multimedia that is critical as a source of information. RGB is one type of image that contains elements of the Red, Green, and Blue colors.

To detect color, an object must first be recognized. The HSV (Hue Saturation Value) image format is one of the image formats used in color segmentation. HSV has a three-part scope: Hue, which represents color, Saturation, which represents color dominance, and Value, which represents brightness. Segmentation is a technique that separates portions of an image in order to obtain accurate object recognition results. As a result, color segmentation using the HSV method will typically detect colors with varying degrees of dominance and brightness depending on the detected object (Putranto, Yoga, Hapsari, & Wijana, 2010).

In digital image processing, a threshold (threshold value) is a limit value. It is a technique for segmenting images with a significant difference in image intensity between the background and the main object. The presence of a value barrier between an object and its background is referred to as the threshold (Ardhianto, Hadikurniawati, & Budiarso, 2013).

Matlab is a programming language with advantages for technical computing needs and programming interfaces that make data analysis, algorithm development, and graph calculations easier (Firmansyah, 2007). Matlab is used to track the color of an object in a video. Matlab R2016b was used in this paper because it is more capable of managing videos than the previous version of matlab.

The study (Achamad & Slamat, 2012) used the C.45 decision tree algorithm to develop a model that predicts employee job performance. Age, gender type, education level, religion, level, length of service, and state of health are some of the variables that are used. The type of data in each variable, on the other hand, is either numerical or categorical data. Schedule data is a type of data that can be used as a class label. There are two types of schedule data: A and B. This study makes an application that can determine employee work schedules automatically by inputting the value of the required features by inputting the value of the required

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

features by inputting the value of The accuracy of the C4.5 decision tree method that was used is around 87 percent.

2. Method

This study's data contains four variables in the form of numeric and categorical data. Gender, age, job class, and employment status are the factors considered. The class label is determined by the work schedule, which is divided into morning and evening shifts. The data used consisted of 54 records from a single company in Jakarta.

Gender, job class, and employment status are the three categorical variables. Because the decision tree in scikit-learn can only analyse numerical data, categorical data must first be preprocessed. Before they may be used in a decision tree model, these variables must be preprocessed. The LabelEncoder package in scikit-learn was used for preprocessing. This module assigns labels to categorical data classes ranging from 0 to n. Table 1 shows the results of tagging with LabelEncoder (Vaish, 2017).

TABLE 1.
CATEGORY VARIABLE PREPROCESSING RESULTS

Variabel	Before	After
Jenis Kelamin	Pria	0
	Wanita	1
Golongan	1	0
Pekerjaan	2	1
	4	2
	5	3
Status	Kontrak	0
Kepegawaian	Tetap	1

This study's data contains four variables in the form of numeric and categorical data. Gender, age, job class, and employment status are the factors considered. The class label is determined by the work schedule, which is divided into morning and evening shifts. The data used consisted of 54 records from a single company in Jakarta.

Gender, job class, and employment status are the three categorical variables. Because the decision tree in scikit-learn can only analyse numerical data, categorical data must first be preprocessed. Before they may be used in a decision tree model, these variables must be preprocessed. The LabelEncoder package in scikit-learn was used for preprocessing. This module assigns labels to categorical data classes ranging from 0 to n. Table 1 shows the results of tagging with LabelEncoder (Vaish, 2017).

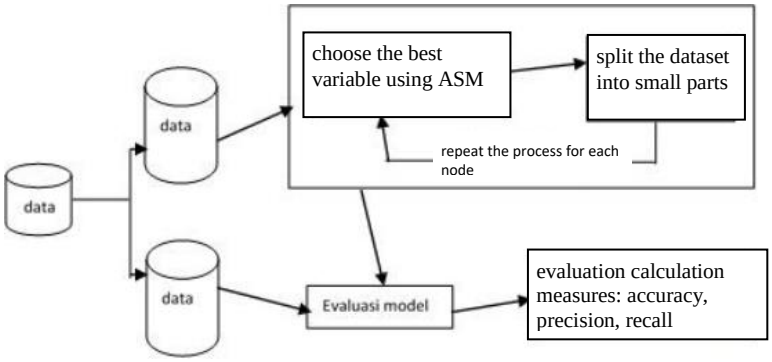

Figure 1. Decision Tree Steps (Navlani, 2018)

The term ASM, which stands for Attribute Selection Measure, is used in Figure 1 to describe one method of selecting separator criteria to group data as efficiently as feasible (Navlani, 2018). The Gini Index is the ASM utilized in the CART algorithm, which is the method used in the scikit-learn decision tree. This instrument determines the purity of each node, with a value greater than zero indicating the presence of

samples from other classes (Ceballos, 2019). Equation 1 shows the general formula for the Gini index (Navlani, 2018).

$$Gini\ (D) = 1 - \sum_{i=1}^{m} p_i^2 \quad \text{.....(1)}$$

pi is the chance that a pair of data in D belongs to class Ci. The next step is to analyze the decision tree model that has been created. The initial step is to experiment with different percentages of training and test data to see which one is optimal for developing a decision tree model. The data was randomly divided into training and test data from the 54 available data.

Accuracy, precision, recall, and F1 measure are determined for each % of training and test data. The comparison of classification results that match all classification results is known as accuracy. The algorithm's recall is how well it recognizes a class, while precision is how accurate the classification results are across all data, and the F1-measure is a mix of recall and precision that measures the method's overall performance (Ridok & Latifah, 2015). The macroaverage value will be utilized in this test for precision, recall, and the F1 measure.

The second assessment compares the decision tree to KNN, which is one of the most effective classification algorithms. The value of accuracy is the evaluation that is being compared. The KNN parameter setting was not done in this assessment, hence the default setting of the KNN module in the scikit-learn package was used.

3. Result and Discussion

Tests with varying percentages of training data and test data were carried out using ten variations, namely utilizing a percentage of test data of 5%, 10%, 15%, and so on, up to 50%. The maximum depth value and the minimum sample leaf value were not established in this test. Figure 2 shows the outcomes of the tests. The chart shows that all evaluation values are more than 0.7, indicating that the decision tree method's performance with the best percentage of test data is 30% of the data. test and 70% of training data. Although there isn't much of a variation in the evaluation values for each % of test data. With 30 percent test data, the accuracy, precision, recall, and F1-measure scores are 0.892, 0.93, 0.82, and 0.85, respectively. The evaluation values are consistent, implying that the decision tree performed admirably.

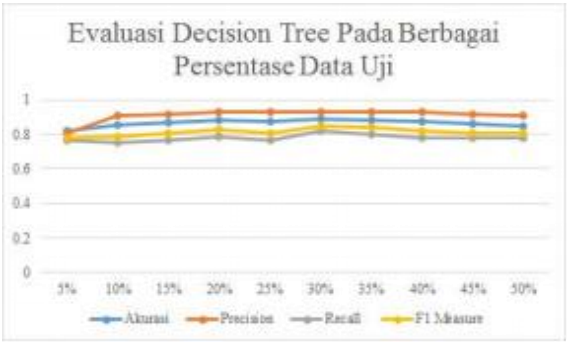

Figure 2. Evaluation of Decision Trees on Various Percentages of Test Data

Tests with varying percentages of training data and test data were carried out using ten variations, namely utilizing a percentage of test data of 5%, 10%, 15%, and so on, up to 50%. The maximum depth value and the minimum sample leaf value were not established in this test. Figure 2 shows the outcomes of the tests. The chart shows that all evaluation values are more than 0.7, indicating that the decision tree method's performance with the best percentage of test data is 30% of the data. test and 70% of training data. Although there isn't much of a variation in the evaluation values for each % of test data. With 30 percent test data, the accuracy, precision, recall, and F1-measure scores are 0.892, 0.93, 0.82, and 0.85, respectively. The evaluation values are consistent, implying that the decision tree performed admirably.

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

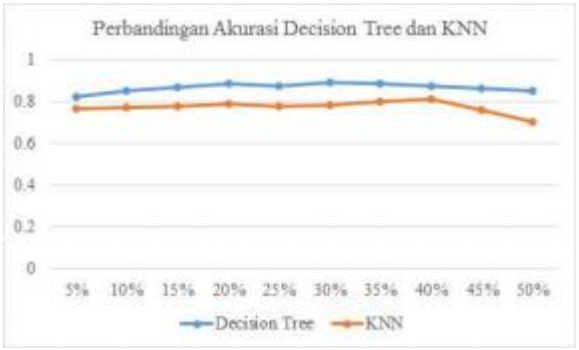

Figure 3. Comparison of Decision Tree Accuracy and KNN

Figure 3 shows the decision tree prediction model with 30 percent test data and no extra parameters. The graphviz tool produced the following image (Ellson, Gansner, Koutsofios, North, & Woodhull, 2001). The tree has a depth of 6, with the root at the 0th level and the lowest two leaves at the 6th level. Each node has a minimum sample size of two.

The selected class is class 1, which is the morning schedule, as seen in the root. There are samples that correspond to class 2, especially the night schedule, with a Gini index root of 0.418. A total of 37 samples were employed in the root, including 26 examples from class 1 and 11 samples from class 2. The age variable 22.5 was chosen because it has the lowest purity value after separation.

The Gini index value is 0 at the 1st depth, in the true branch, indicating that all eight samples belong to the same class, namely 1. The remainder can be found in the false branch. The job class variable is what is used to separate the nodes. The labor class variable is divided down into two branches at the second depth, namely gender and age 29.5. The data is split into branches indefinitely until all of it is used.

Nodes of a categorical decision tree are typically phrased as "whether gender is male." Meanwhile, in the scikit-learn decision tree model, category is treated as numerical, resulting in a node that reads "is the gender 0.5." This is because categorical data is converted to numerical and handled as continuous data in a decision tree during the preprocessing process using the LabelEncoder. Because encoder labels convert category data into ordered integers, they aren't suitable for use, especially if the data isn't ordinal. Because the training and test data are encoded at the same time, the accuracy and precision values of the decision tree evaluation are quite good, particularly over 0.7 and above 0.9. The decision tree will have trouble anticipating data that is entered individually and is still categorical in nature.

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

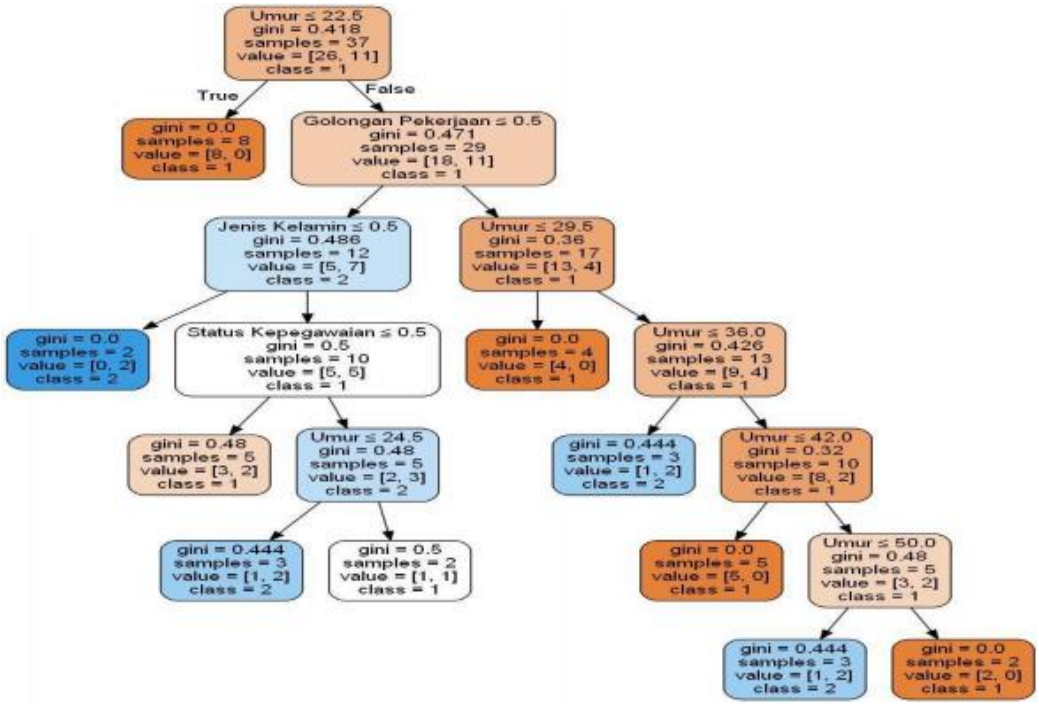

Figure 4. Decision Tree Model For Predicting Work Schedules

4. Conclusion

To forecast work schedules, a decision tree model was created using 54 work schedule data with numerical and categorical characteristics. Because the model is built with the scikit-learn toolkit, categorical data must be preprocessed using the Label Encoder module. The decision tree model has a high evaluation value, with accuracy and precision exceeding 0.7 and 0.9, respectively. The test findings with varied percentages of test data revealed that the evaluation results were not significantly different, but that the 30% test data had the highest assessment value. The decision tree outperforms the KNN in terms of accuracy, with the decision tree scoring more than 0.8 and the KNN scoring lower. Because categorical data is regarded an ordered number, the translation of categorical data to numerical data by scikit-learn is not appropriate.

5. References

[1] Achamad, B. D. M., & Slamat, F. (2012). Klasifikasi Data Karyawan Untuk Menentukan Jadwal Kerja Menggunakan Metode Decision Tree. Jurnal IPTEK, 16(1), 17–23.

[2] Aguilar-chinea, R. M., Castilla, I., Expósito, C., Aguilar-chinea, R. M., Castilla, I., Moreno-vega, J. M., & Moreno-vega, J. M. (2019). Using a decision tree algorithm to predict the robustness of a transshipment schedule. Procedia Computer Science, 149, 529–536. https://doi.org/10.1016/j.procs.2019.01.172

[3] Ceballos, F. (2019). Scikit-Learn Decision Trees Explained - Training, Visualizing, and Making Predictions with Decision Trees. Retrieved from https://towardsdatascience.com/scikit-learn-decision-trees-explained-803f3812290d

[4] Ellson, J., Gansner, E., Koutsofios, L., North, S., & Woodhull, G. (2001). Graphviz - Open Source Graph Drawing Tools. In Lecture Notes in Computer Science (pp. 483–484). Springer-Verlag. Retrieved from https://graphviz.gitlab.io/

[5] Floares, A. G., Calin, G. A., & Manolache, F. B. (2016). Bigger Data is Better for Molecular Diagnosis Tests Based on Decision Trees. In In: Tan Y., Shi Y. (eds) Data Mining and Big Data. DMBD 2016. Lecture Notes in Computer Science, vol 9714 (pp. 288–295). Springer, Cham. https://doi.org/https://doi.org/10.1007/978-3-319-40973-3_29

[6] Loh, W. (2011). Classification and Regression Trees. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, (January 2011). https://doi.org/10.1002/widm.8

[7] Navlani, A. (2018). Decision Tree Classification in Python. Retrieved from https://www.datacamp.com/community/tutorials/decision-tree-classification-python

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

[8] Ochiai, Y., Masuma, Y., & Tomii, N. (2019). Improvement of timetable robustness by analysis of drivers' operation based on decision trees. Journal of Rail Transport Planning & Management, 9(March), 57–65. https://doi.org/10.1016/j.jrtpm.2019.03.001
[9] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, E. (2011). Scikit-learn: Machine Learning in {P}ython. Journal of Machine Learning Research, 12, 2825--2830.
[10] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, E. (2019). Decision Tree. Retrieved from https://scikit-learn.org/stable/modules/tree.html
[11] Ridok, A., & Latifah, R. (2015). Klasifikasi Teks Bahasa Indonesia Pada Corpus Tak Seimbang Menggunakan NWKNN. In Konferensi Nasional Sistem & Informatika 2015 (pp. 9–10).
[12] Topîrceanu, A., & Grosseck, G. (2017). Decision tree learning used for the classification of student archetypes in online courses. Procedia Computer Science, 112, 51–60. https://doi.org/10.1016/j.procs.2017.08.021
[13] Vaish, P. (2017). Decision Trees in scikit-learn. Retrieved August 14, 2019, from https://adataanalyst.com/scikit-learn/decision-trees-scikit-learn/