Available at https://journalenrichment.com/index.php/jr/

Enrichment: Journal of Multidisciplinary Research and Development

Predict GO-JEK Driver Income Level in the Bali Region using Decision Tree

I Dewa Gede Ngurah Bramasta Darmawan[1*], Dian Eka Ratnawati [2]

[1,2]Universitas Brawijaya Malang, Yogyakarta, Indonesia

[*]Email: : bramasta@ub.ac.id, dian_ilkom@ub.ac.id

ARTICLE INFO

Keywords:
GO-JEK, J48 algorithm, information technology, confusion matrix, ROC/AUC curve, revenue

ABSTRACT

The work in the transportation industry is one of several that is impacted by information technology, which is now developing quickly. The emergence of information technology in the transportation industry led to the creation of GO-JEK. GO-JEK is growing in popularity with the general public because, in addition to offering a wide range of services, it also creates new jobs with more flexible working hours for the community. Of course, each driver's income from working as a GO-JEK driver is unique. However, a number of things can have an impact on this income. With a focus on the Bali region, this study tries to forecast income levels for drivers of GO-JEK. J48 algorithm is used to process the data, producing a decision tree with 19 rule models. An accuracy value of 77.61%, a precision value of 78.95%, a value recall of 81.08%, and an AUC value of 0.808 are obtained from the rule model test utilizing the confusion matrix and ROC/AUC curve. These results demonstrate how effective the decision tree and rule model are.

INTRODUCTION

Initially, information technology could only be used to make phone calls, but it has since advanced rapidly. Today's information technology can be used for more than just phone calls, it can also be used for teaching and learning, as well as shopping (Joianus 2020). Many current jobs have also been influenced by information technology developments, one of which is the emergence of online motorcycle taxis as a result of the incorporation of information technology developments into the traditional motorcycle taxi transportation industry. Online motorcycle taxis currently come in a wide variety, with GO-JEK being one of the most well-known (Purwanto, Luthfi, and Arsal 2018), (Aziah and Adawia 2018).

Nadiem Makarim founded the company in 2009, and it is a well-known online motorcycle taxi service. GO-JEK is growing in popularity because the services it offers its customers are so diverse, ranging from picking up passengers to picking up goods (Aziah and Adawia 2018). People who want to work as GO-JEK driver partners can also find work with the company. This job is appealing because, in addition to flexible working hours, profit sharing between drivers and GO-JEK is also a profitable distribution for drivers, with 80% profit sharing for drivers and 20% profit sharing for GO-JEK parties (Giri and Dewi 2017). The number of drivers in the Bali region reached 14,500 in 2017, as did the number of downloads of the Gojek Driver application on the Google Play Store, which reached five million downloads and 500 thousand reviews (Giri and Dewi 2017; Google Playstore n.d.). GO-JEK drivers can earn up to 500,000 rupiah per month, but their earnings may decrease due to competition from Grab, Maxim, and others (Giri and Dewi 2017). Each driver's income will also differ due to a variety of influencing factors, one of which is the difference in the number of orders and the length of time worked, which can be a differentiator in generating income for each GO-JEK driver. This study was carried out to assist in predicting the level of income of GO-JEK drivers in the Bali region, by using attributes that can affect GO-JEK driver income and the J48 algorithm. The research will result in decision trees and rule models, which will then be tested using the confusion matrix and ROC/AUC curves to determine how effective and accurate they are. This section will describe similar types of research and materials used to support research.

A. Related Works

This subsection will explain the type of research that was used to assist with this research.

1) Rizky Evita Putri (2021) : The purpose of this study was to forecast the efficiency of an injection machine in a factory. It used the C4.5 algorithm to classify a variety of data within the factory in order to forecast engine performance effectiveness. The resulting decision tree and model rules were then tested using the matrix confusion method and the ROC/AUC curve to determine the classification model's accuracy level as a result of C45 data processing. The C4.5 algorithm was used in this study, which was similar to the J48 algorithm. When compared to the C4.5 algorithm, the J48 algorithm is more accurate.
2) Abdul Rohman and Anief Rufiyanto (2019) : The C4.5 algorithm is also used in this study to forecast the number of students graduating from an Indonesian university. The confusion matrix and ROC/AUC curves are also used in tests on the resulting decision trees and rule models. The difference between this study and previous research is that in this study, the data is processed with an additional k-fold cross validation. The purpose of using k-fold cross validation is to make the process of sharing data between test and training data easier.

B. Materials

This subsection will explain the type of research that was used to assist with this research

1) Decision Tree : Decision Tree is one of the most commonly used methods in machine learning, image processing, and determining a pattern of information. The classification algorithm includes a Decision Tree, which can organize a large amount of data. Using training datasets as a basis, decision trees can be used to classify knowledge and information (Charbuty and Abdulazeez 2021; Wahyuni 2018).
2) *J48 Algorithm* : This algorithm can generate binary trees, which are built during the classification process, and each node is stored in a database. When creating a decision tree, this algorithm will ignore any missing values or attributes (Kaunang 2019). The J48 algorithm is used because it can process different types of data, both discrete and continuous data, and it can generate simple results rules (Wahyuni and Anggraini 2022). The decision tree that is generated has nodes and leaves, with each node representing the value of an attribute and each leaf representing a class (Azlan et al. 2016). The nodes and leaves of the decision tree are formed using the J48 algorithm's two formulas (Pakpahan 2021). First, the formula for calculating the Gain value (1), which is used to determine the first node (the root) and the next node.

$$\text{Gain S, A} = \text{Entropy S} - \sum_{i=1}^{n} \frac{|Si|}{|S|} \times \text{Entropy S} \quad (1)$$

Second, the formula for calculating entropy values (2) to aid in the production of leaves from a decision tree.

$$\text{Entropy S} = \sum_{i=1}^{n} (- p \times \log_2 pi) \quad (2)$$

3) GO-JEK : GO-JEK is an online transportation company founded in 2009 by Nadiem Makarim. GO-JEK not only provides passenger pick-up services, but also a slew of other options, such as ordering food delivery and picking up goods. GO-JEK began with only 20 driver partners and has since grown to include over two million driver partners throughout Indonesia. GO-JEK now has its own e-wallet service, making transactions between customers and drivers easier (Aziah and Adawia 2018; PT GoTo Gojek Tokopedia Tbk n.d.).
4) Confusion Matrix : Confusion Matrix is a test method used to evaluate or assess the classification method's performance. The confusion matrix compares the obtained classification results to the expected results (Caesaria, Astiningrum, and Syulistyo 2020). This method of testing includes four assessments: the first is an accuracy assessment, the second is a precision assessment, the third is a recall assessment, and the fourth is an error assessment. The following table shows how testing is carried out using four attributes that state the classification of the number of correct test data and the number of incorrect test data (Pratiwi, Handayani, and Sarjana 2021) (**Error! Reference source not found.**).

Table I. Confusion Matrix

Observed	**Predicted Class**	
	True	False
True	True Positive (TP)	False Negative (FN)
False	False Positve (FN)	True Negative (TN)

Using the formula below, these four attributes will be used to calculate accuracy (3), precision (4), and recall (5) (Pakpahan 2021).

a. Accuracy

$$\text{Accuracy} = \frac{TP+TN}{TN+TP+FP+FN} \times 100\% \quad (3)$$

b. Precision

$$\text{Precision} = \frac{TP}{TP+FP} \times 100\% \quad (4)$$

c. Recall

$$\text{Recall} = \frac{TP}{TP+FN} \times 100\% \quad (5)$$

5) *ROC/AUC Curve* : The Receiver Operating Characteristic (ROC) is a graphical representation used to assess the accuracy of a diagnostic. The higher the curve is above the diagonal line, the more accurate the predicted value (Nengsih, Zein, and Hayati 2021). The AUC (Area Under Curve) method is required to calculate the value of the area under the ROC curve. AUC is a performance measurement method used to determine whether or not there is a difference in specific conditions. The higher the AUC value, the more accurate the classification, conversely, the lower the value, the less accurate the classification (Hoo, Candlish, and Teare 2017),(Rohman and Rufiyanto 2019),(Putri 2021). The classification of AUC values is shown in the **Error! Reference source not found.**.

Table 2. Auc Classification

No.	Value Range	Classification Group
1	0.5 – 0.6	Wrong Classification
2	0.61 – 0.7	Bad Classification
3	0.71 – 0.8	Enough Classification
4	0.81 – 0.9	Good Classification
5	0.91 – 1.0	Perfect Classification

6) *RapidMiner Studio* : RapidMiner Studio is a data mining algorithm-processing software. With the help of the provided *operators*, RapidMiner can make it easier for users to process data, whether it's data calculations or other large amounts of data. This operator includes data-modifying functions. Data is simply connected to the operator's nodes; if you want to see the results of data processing, connect the operator node to the result node, and the results will be displayed. RapidMiner results can be displayed graphically (Rahmat et al. 2017).

METHOD

Before implementing the J48 algorithm, interviews were conducted to determine the factors that influence GO-JEK drivers' earnings. The interviews were informal so that the researcher could get closer to the participants' drivers and extract information from them The interview produced several attributes that could be included in the questionnaire questions and included in the dataset. These attributes include, "City of origin", "Days worked", "Time worked", "Hours", "Age", "Number of orders", "Type of orders", "Do you have another job?", "Total bonus" , "Monthly income", and "Is the income enough?".

Data will also be collected using a questionnaire, because a questionnaire allows one to reach out to GO-JEK drivers in any location, ensuring that the data collected is in accordance with predetermined targets. The questions compiled into the questionnaire are questions compiled from the results of interviews with drivers about the factors that affect their income, after which the questionnaire will be distributed to drivers to fill out. Apart from distributing questionnaires, we also visited several GO-JEK drivers' bases directly. The questionnaire, which was distributed to drivers, received responses from 334 people. The next 334 people will become data, which will be processed in the next stage.

The following step is to perform data preprocessing and data transformation on the data collected via a questionnaire. The preprocessing stage's goal is to remove attributes that aren't required for the next stage, resulting in fewer attributes used but accurate results. The transformation stage is used to simplify attribute values. For example, the age attribute, which has four value ranges, is reduced to only two value ranges. The process aims to simplify manual calculations as well as to simplify and facilitate computer calculations while still providing accurate results. The transformation stage is also used to add classes to the dataset so that the data can be classified as ">= 3 Million" or "< 3 Million". The way to determine this class is from “Monthly income” attribute.

The data mining process will be carried out in the following stage. The RapidMiner Studio 9.10.011 application is used to perform the data mining process. The data that has been preprocessed and transformed will be entered into RapidMiner Studio and divided into two datasets: one for test data (20% of total data) and one for training data (80% of total data). The J48 algorithm will then be used to generate decision trees and rule models from the training data. These results will be tested later to determine how accurate the classification results are in predicting the income level of GO-JEK drivers in real terms.

The evaluation process will then be carried out using the confusion matrix and ROC/AUC curve testing methods. Testing is carried out with predetermined test data that accounts for 20% of the total data. The test results will show the level of accuracy of the resulting rule model as well as the quality of the classification results. These values will later serve as a reference for using the rule model in real life, and will determine whether or not the rule model predicts correctly.

RESULTS AND DISCUSSION

This section will explain the results of data processing using RapidMiner Studio to implement the J48 algorithm, as well as the results of testing using the confusion matrix and ROC/AUC curve on the J48 algorithm results. These are the results obtained from data based on the results of interviews and the distribution of questionnaires, which have been preprocessed and transformed.

A. Decision Tree

Fig. 1 is the result of the decision tree, after processing data with the J48 algorithm (these results are a re-drawing of the actual results in RapidMiner Studio).

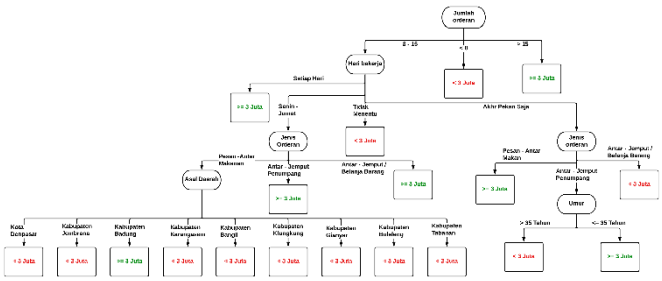

Fig. 1. Decision tree results from the implementation of the J48 algorithm

The results of the decision tree in Fig. 1 can also be made in the form of a rule model to make it easier to understand. The rule model that is formed from the decision tree is 19 rules, namely:

1. If "Number of orders" per day is 8 – 15, and "Working days" every day, then "Monthly income" >= 3 Million
2. If "Number of orders" per day is 8-15, "Working days" Monday - Friday, "Order type" - food delivery and "Regional origin" Denpasar City, then "Monthly income" < 3 Million
3. If "Number of orders" per day is 8-15, "Working days" Monday - Friday, "Type of orders" - food delivery and "Regional origin" Jembrana Regency, then "Monthly income" < 3 Million
4. If "Number of orders" per day is 8-15, "Working days" Monday - Friday, "Type of orders" - food delivery and "Regional origin" Badung Regency, then "Monthly income" >= 3 Million
5. If "Number of orders" per day is 8-15, "Working days" Monday - Friday, "Type of orders" - food delivery and "Regional origin" Karangasem Regency, then "Monthly income" < 3 Million
6. If "Number of orders" per day is 8-15, "Working days" Monday - Friday, "Type of orders" - food delivery and "Regional origin" Bangli Regency, then "Monthly income" < 3 million
7. If "Number of orders" per day is 8-15, "Working days" Monday - Friday, "Type of orders" - food delivery and "Regional origin" Klungkung Regency, then "Monthly income" < 3 Million
8. If "Number of orders" per day is 8-15, "Working days" Monday - Friday, "Type of orders" - food delivery and "Regional origin" of Gianyar Regency, then "Monthly income" < 3 Million
9. If the "Number of orders" per day is 8 - 15, "Working days" Monday - Friday, "Type of orders" - food delivery and "Regional origin" of Buleleng Regency then "Monthly income" < 3 Million
10. If "Number of orders" per day is 8-15, "Working days" Monday - Friday, "Type of orders" - food delivery and "Regional origin" of Tabanan Regency, then "Monthly income" < 3 Million
11. If "Number of orders" per day is 8 - 15, "Working days" Monday - Friday, and "Order type" picks up passengers, then "Monthly income" >= 3 Million

12. If "Number of orders" per day is 8 - 15, "Working days" Monday - Friday, and "Type of orders" pick up / shop for goods, then "Monthly income" >= 3 Million
13. If the “Number of orders” per day is 8 – 15, and “Working days” are uncertain, then “Monthly Revenue” < 3 Million
14. If the "Number of orders" per day is 8-15, "Working days" weekends only, and "Order type" orders - food delivery, then "Monthly income" >= 3 Million
15. If “Number of orders” per day is 8 – 15, “Working days” only weekends, “Order type” pick-up and pick-up passengers and “age” > 35 years, then “Monthly income” < 3 Million
16. If “Number of orders” per day is 8 – 15, “Working days” weekends only, “Order type” pick-up and pick-up passengers and “age” <= 35 years, then “Monthly income” >= 3 Million
17. If "Number of orders" per day 8-15, "Working days" only weekends and "Type of orders" pick-up / shopping for goods, then "Monthly income" < 3 Million
18. If "Number of orders" per day < 8, then "Revenue per month" < 3 million
19. If "Number of orders" per day > 15, then "Revenue per month" >= 3 Million

This rule demonstrates that not all of the attributes in the dataset are used in the decision tree formation..

B. Confusion Matrix

The confusion matrix is one of the testing methods used in this study. The confusion matrix is used to assess how good/accurate the decision tree's predictions are. TABLE III shows the results of the confusion matrix, which yield three values: accuracy, precision, and recall.

Table 3. Confusion Matrix Results

Evaluation	Value
Accuracy	77.61%
Precision	78.95%
Recall	81.08 %

The results in TABLE III show that the decision tree and the resulting rule model are fairly good or accurate, which means that the rule model almost always produces predictions that are correct. The obtained accuracy, precision, and recall results indicate that the resulting rule model is already very good. Accuracy is defined as how precise the prediction results are generated from existing data, and it receives an accuracy of 77.61%. Precision refers to how precise the data can be for positive/">= 3 million" results from all predicted data, with a precision of 78.95%. Recall is intended to be how precise the predictions that produce a ">= 3 million" class of all data that has a ">= 3 million" class, and for the results obtained from recall of 81.08%.

C. ROC/AUC Curve

The ROC/AUC curve is used to test how good the resulting classification model is. It can be said that the ROC/AUC curve will not always be directly proportional to the results of the confusion matrix. Confusion matrix that gets good results, not necessarily the ROC/AUC curve also gets good results. The results in Fig. 2 show that the results of the ROC/AUC curve can be categorized as a result of enough classification based on TABLE II, because it gets an AUC value of 0.808. This result means that the resulting classification model, namely the decision tree and the rule model, can differentiate well between GO-JEK drivers who get more than three million or those who get less than three million or in other words, the predictions produced can be said to be enough accurate.

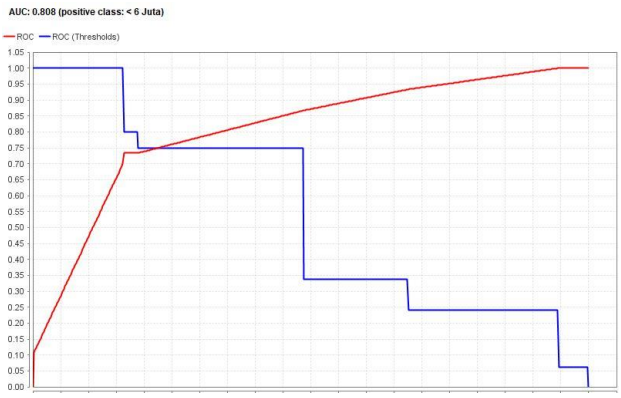

Fig. 2. ROC/AUC Curve Result

CONCLUSION

This research aims to help people in the Bali region who want to become GO-JEK drivers or those who have become GO-JEK drivers, to be able to predict or estimate income and levels income each of them based on several factors, namely, "Monthly income", "Number of orders", "Type of orders", "Working days", "Age", "Amount of bonuses", "Working hours", "Working time", and several other factors. These factors were obtained from the results of interviews with 25 GO-JEK drivers in the region of Bali. These factors were then processed into questionnaire questions and distributed back to the drivers. The results of the questionnaire respondents will then be used as the main dataset, to be processed with the J48 algorithm. The dataset is processed first with preprocessing to remove attributes that are not really needed, then processed with transformation to simplify the value of each attribute and add classes. The ">= 3 million" and "< 3 million" classes were determined "Montly income" attribute. The data that has gone through the transformation will then be processed with the J48 algorithm and with the help of the RapidMiner Studio application. The dataset will be divided into two, namely training data of 80% of all data and test data of 20% of all data. Processing training data with the J48 algorithm, produces a decision tree and a rule model of 19 rules.

These results will then be tested using the test data that has been prepared and with the help of the confusion matrix method and the ROC/AUC curve. The results of the test are as follows, the accuracy value obtained is 77.61%, the precision value obtained is 78.95%, the recall value obtained is 81.08% and the AUC value obtained is 0.808. The test results can be said to be enough good results, because it can be said that it has been able to predict the income of each GO-JEK driver accurately and well, even though the prediction results cannot be said to be perfect. These results can of course be different if the number of attributes removed is fewer or greater in the preprocessing process, as well as the simplified value transformation process is different, so the results from the decision tree to the test results will also experience differences, which can be better or worse. The amount of data for each class also affects the results obtained, if the factors/attributes that determine the class in the dataset are increased or decreased, then the results obtained will also be different.

Suggestions for the next similar or similar research, to use a comparison of the number of different test data and training data, as well as the addition of the number of factors that influence income GO-JEK drivers. Another suggestion, to determine the class in the dataset, you can use different methods, such as C4.5 algorithm, so that the amount of data for each class will also be different from this research, and can produce different results. The next suggestion is that the next research is expected to use a different classification method, so the results obtained will also be different, and can be compared with the results of this study to see which method has better results.

REFERENCES

Aziah, Ayu, and Popon Rabia Adawia. 2018. "Analisis Perkembangan Industri Transportasi Online Di Era Inovasi Disruptif (Studi Kasus PT Gojek Indonesia)." *Cakrawala* 18(2):149–56. doi: 10.31294/jc.v18i2.

Azlan, Wan Amirah W., Siaw Hong Liew, Yun Huoy Choo, Hazli Zakaria, and Yin Fen Low. 2016. "Wavelet Feature Extraction and J48 Decision Tree Classification of Auditory Late Response (ALR) Elicited by Transcranial Magnetic Stimulation." *ARPN Journal of Engineering and Applied Sciences* 11(10):6319–23.

Caesaria, Aura Kanza, Mungki Astiningrum, and Arie Rachmad Syulistyo. 2020. "Identifikasi Komponen Gui Pada Prototipe Aplikasi Mobile." *Jurnal Informatika Polinema* 6(2):51–56. doi: 10.33795/jip.v6i2.321.

Charbuty, Bahzad, and Adnan Abdulazeez. 2021. "Classification Based on Decision Tree Algorithm for Machine Learning." *Journal of Applied Science and Technology Trends* 2(01):20–28. doi: 10.38094/jastt20165.

Giri, Putu Citrayani, and Made Heny Urmila Dewi. 2017. "ANALISIS FAKTOR-FAKTOR YANG MEMPENGARUHI PENDAPATAN DRIVER GO-JEK DI KOTA DENPASAR, BALI."

Google Playstore. n.d. "Gojek Driver - Aplikasi Di Google Play."

Hoo, Zhe Hui, Jane Candlish, and Dawn Teare. 2017. "What Is an ROC Curve?" *Emergency Medicine Journal* 34(6):357–59. doi: 10.1136/emermed-2017-206735.

Joianus. 2020. *Pengaruh Perkembangan Teknologi Informasi Bagi Gaya Hidup Gereja*.

Kaunang, Fergie Joanda. 2019. "Penerapan Algoritma J48 Decision Tree Untuk Analisis Tingkat Kemiskinan Di Indonesia." *CogITo Smart Journal* 4(2):348–57. doi: 10.31154/cogito.v4i2.141.348-357.

Nengsih, Warnia, M. Mahrus Zein, and Nazifa Hayati. 2021. "Coarse-Grained Sentiment Analysis Berbasis Natural Language Processing – Ulasan Hotel." *Jurnal Nasional Teknik Elektro Dan Teknologi Informasi* 10(1):41–48. doi: 10.22146/jnteti.v10i1.548.

Pakpahan, N. S. 2021. "Implementasi Data Mining Menggunakan Algoritma J48 Dalam Menentukan Pola Itemset Belanja Pembeli (Study Kasus: Swalayan Brastagi Medan)." *Journal of Computing and Informatics Research*

1(1):7–13.
Pratiwi, Banu Putri, Ade Silvia Handayani, and Sarjana. 2021. "Pengukuran Kinerja Sistem Kualitas Udara Dengan Teknologi Wsn Menggunakan Confusion Matrix." *Jurnal Informatika Upgris* 6(2):66–75. doi: 10.26877/jiu.v6i2.6552.
PT GoTo Gojek Tokopedia Tbk. n.d. "Gojek Super App: Ojek Online, Taksi Online, Pesan Makan, Kirim Barang, Pembayaran."
Purwanto, Andhika Cahya, Asma Luthfi, and Thriwarty Arsal. 2018. *Eksistensi Ojek Pangkalan Didalam Perkembangan Transportasi Berbasis Informasi Dan Teknologi*. Vol. 7.
Putri, Rizky Evita. 2021. "IMPLEMENTASI DATA MINING UNTUK PREDIKSI EFEKTIVITAS PADA MESIN INJECTION MENGGUNAKAN ALGORITMA C4.5 (Studi Kasus : PT. Tridaya Artaguna Santara)." 1–177.
Rahmat, Brilian, Agum Agidtama Gafar, Nurul Fajriani, Umar Ramdani, Fitria Rihin Uyun, Yuwanda Purnamasari P., and Natalis Ransi. 2017. "Implemetasi K-Means Clustering Pada Rapidminer Untuk Analisis Daerah Rawan Kecelakaan." *Seminar Nasional Riset Kuantitatif Terapan 2017* (April):58–60.
Rohman, Abdul, and Anief Rufiyanto. 2019. "Implementasi Data Mining Dengan Algoritma Decision Tree C4 . 5 Untuk Prediksi Kelulusan Mahasiswa Di Universitas Pandaran." *Proceeding SINTAK 2019* 134–39.
Wahyuni, Ana, and Sri Anggraini. 2022. "Implementasi Algoritma J48 Data Mining Untuk Inovasi Bisinis Perhotelan Di Masa Pandemi Covid- 19 (Studi Kasus Hotel SNS Semarang)." XIII(1):182–92.
Wahyuni, Sri. 2018. "Implementation of Data Mining to Analyze Drug Cases Using C4.5 Decision Tree." *Journal of Physics: Conference Series* 970(1). doi: 10.1088/1742-6596/970/1/012030.