Aptisi Transactions on Technopreneurship (ATT)
Vol. 4 No. 3 November 2022

P-ISSN: 2655-8807
E-ISSN: 2656-8888

Application of Data Mining for Slot Time Prediction at International Airports in Indonesia: J48 Algorithm

Renddy Wandhana Suryaman[1], Gunawan Wang[2], Viany Utami Tjhin[3]
Magister Manajemen Sistem Informasi, Bina Nusantara University, Jakarta, Indonesia[1,2,3]
Email coresponding: renddyws@gmail.com[1]

Author Notification
29 August 2022
Final Revised
15 September 2022
Published
30 September 2022

Renddy Wandhana Suryaman, Gunawan Wang, & Viany Utami Tjhin. (2022). Application of Data Mining for Slot Time Prediction at International Airports in Indonesia: J48 Algorithm . Aptisi Transactions on Technopreneurship (ATT), 4(3), 215–225.
DOI: https://doi.org/10.34306/att.v4i3.263

Abstract

In aviation, the safety and smooth flow of flight traffic is a business core, where every flight traffic service is expected to avoid delays caused by aircraft movement either in the air or on land. Therefore, the time slot at the airport is essential for the accuracy of the movement of aircraft, both Departure and Arrival, and this is intended to avoid delays caused by the accumulation of queues of planes that will depart and planes that will land, with a large number of aircraft movements at the International Airport. Soekarno Hatta requires analysis with data mining techniques such as the J48 algorithm and Decision Tree, and Naïve Bayes.

Keywords: *Data mining, J48, Decision Tree, Naïve Bayes, Slot Time.*

1. Introduction

An aviation traffic service is said to be good if all systems in the aviation traffic service are integrated and run smoothly so that the process of air traffic services as the company's core business can run well. AirNav Indonesia is an aviation navigation service company that prioritizes flight service and safety [1]. Therefore, if there is a long queue for each plane that will land or will fly at all airports in Indonesia, especially Soekarno Hatta international airport, this will significantly affect the safety. of aviation and business process companies [2] [3]. Thus, the time slot at each airport in Indonesia is a serious concern, especially during the golden time when every airline is fighting for the slot time at that hour, if not appropriately managed, then chaos can occur, and the impact on flight safety can result in incidents or aircraft accidents.

In aviation traffic services, the flight plan is the primary data, which contains all information about the aircraft's departure plan to its destination and other important information [4]. The flight plan data is also data that will then be the basis for billing for flight traffic services provided by AirNav Indonesia to airlines. The flight plan data will be sent to the Air Traffic Service System (ATS System) and activated by Departure news when the plane has flown, then deactivated with Arrival news after the plane lands safely [5].

The process of assigning time slots is still manual based on several parameters: requests from airlines, availability of aircraft parking at airports, willingness to accept airports, and availability of time slots at both departure and destination airports [6]. However, suppose the company has a good analysis of the time slot by managing every aircraft movement, whether it is on schedule or delayed or cancelled flights due to operational or technical problems, which are not caused by queues [7]. In that case, this will be a very effective and efficient service that can have a direct impact on every stakeholder at the airport, both the company itself and the airport and airline, which leads to excellent and timely service for aircraft passengers as the main customers in the air transportation service business process [8].

P-ISSN: 2655-8807
E-ISSN: 2656-8888

Soekarno Hatta International Airport has an average take-off and landing movement of between 800 to 1000 daily and reaches 1200 per day at certain times. Soekarno Hatta international airport can accommodate 86 aircraft movements per hour using two runways. Increasing airport capacity by increasing the number of runways or moving some flights to another nearest airport is not the right solution [9]. Due to very detailed and strict flight regulations, it requires a lot of money and very long preparation. This is due to safety issues. Is the business core in the air transportation business.

Over time and economic development, this will encourage the development of air transportation routes which will increase aircraft movements, so there is a need for mitigation by analyzing aircraft movements to increase the occupancy of the available time slots [10]. Each element that is part of data mining for a time slot will be explained further.

Therefore, a data mining technique solution with classification modelling is one solution to predict the availability of time slots, especially at Soekarno Hatta international airport, whether it will be available or not. Because every day, there will be delays or cancellations of flights and extra flights that can change the arrangement of time slots at each hour, especially at the golden time to increase occupancy, which has an impact on the smooth flow of flight traffic, where each stakeholder receives the final effect [11]. It is improving service to consumers, the efficiency of expenditure (avtur aircraft, aircraft parking, etc.), and increasing the company's income. Data mining is a process that applies various techniques such as statistics, math, artificial intelligence, and machine learning. These techniques are used to identify and extract information to increase knowledge about sources of large databases and data warehouses [12]. Predictive modelling is a process that uses various combinations of data mining and data analysis techniques to produce the desired forecast information.

2. Research Method

Data mining comes from the word mining, which means mining. It can be used to explore the data owned if it is developed. Data mining is an integrated information investigation consisting of a series of exercises based on the meaning of the target to be broken down, examining information for understanding and evaluating the results. Information mining collection not only collects information but includes the investigation and expectation of the data to be displayed. The information collected is stored in a data set and then handled so that it tends to be used for decision-making.

To dig further data-based information, usually using data mining and Knowledge Discovery in Databases (KDD), where the implementation is done alternately. Data mining and KDD are used because the information to be processed is usually in an extensive database, but the two are still interrelated. The following is an overview of the KDD process:

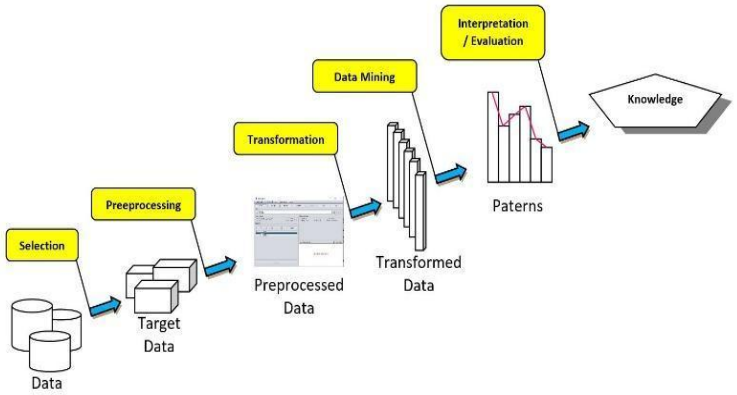

Figure 1. KDD Schematic

The stages of the KDD process in data mining, the steps are as follows:

1. Data Selection, namely the process of collecting data at related locations according to the relevant analysis.
2. Data Preprocessing/Cleaning, which is the stage of the data cleaning process, as well as checking if there is a lack of data, data that is still empty, duplication, and data that is less relevant.
3. Transformation is the stage of selecting data from the results of data preprocessing/cleaning that aims to obtain data as expected in data mining.
4. Data mining is a way to find the desired pattern through the application of a method from the data to be displayed.
5. Interpretation/Evaluation (Interpretation or Evaluation) is the translation stage of a technique or pattern from data mining results to facilitate understanding of the information generated.

2.1 Decision Tree

One of the information commonly used to mine methods is the Decision Tree. The adaptability of this procedure makes it very appealing, especially as it presents the positive side of a direct perception in which the parts of the tree encapsulate the characterization. The choice tree has three classical styles, namely:

1. The classification Tree is applied when the estimated result is class membership, for example, the decision tree algorithm.
2. Regression Tree is applied when the estimated result is a natural number, for example, fuel prices and building values.
3. CART (C&RT) is a denomination and a regression tree. Decision trees have been developed, but ID3 and decision trees are two of the favourite techniques for research analysis. These two techniques have the same principle because the decision tree algorithm was developed through ID3. However, there are different concepts between these two techniques, namely:
   - Choice trees can handle consistent and discrete quality and the preparation of information with missing quality or invalid input.
   - It will manage the results obtained from the choice tree calculation after the properties' determination is carried out by utilizing the Gain Ratio.
   - The calculation of the choice tree is an improvement from ID3 using the Gain Ratio. Advantages to refreshing the data acquisition using the equation:

$$GainRatio(S.A) = \frac{Gain(S.A)}{SplitInfo(S.A)}$$

Where:
S : Space/Sample data is used for data training
A : Attribute Gain (S,A) = information gain for attribute A Split Info (S,A) = split information for attribute A

The property with the highest Gain Ratio is chosen as the test quality for the hub. This approach applies standardization to data acquisition by utilizing what is called parted data, with the equation:

$$SplitInfo(S.A) = -\sum_{i=1}^{i} \frac{S_i}{S} log_2 \frac{S_i}{S}$$

Where :
S : The sample (data) space used for training.
A : Attributes.
Si : Number of samples on attribute

When constructing a selection tree, there may be confusion or blank information in the preparation information. Tree pruning should be possible to recognize and remove these branches so that the tree is simpler and clearer for better arrangement. There are two strategies for pruning the tree of choice, namely:

- Through Pre Pruning for example, stop building from the beginning of the subtree so that it does not go further in pruning the preparation information. Pre Pruning equation

$$e = \frac{r + \frac{Z^2}{2n} + Z\sqrt{\frac{r}{n} - \frac{r^2}{n} + \frac{Z^2}{4n^2}}}{1 + \frac{Z^2}{n}}$$

Where
r : error rate comparison value
n : total sample
z : Φ-1(c)
c : confidence level

- Using Post Pruning for example working on a tree by removing some of the branches of a subtree each time it is created. This technique is a standard part of the choice tree calculation.

2.2 Naive Bayes

Naive Bayes is a characterization strategy often used to measure values whose sign is unclear. Using the Naïve Bayes strategy requires a little preparation of the exact information called preparing the information, which is used to determine the assessed limits needed for the order cycle. The Naïve Bayes strategy is also a technique that estimates the probability of one class from each current set of properties and determines which type is the most ideal. The Naïve Bayes calculation has cycle phases that must be completed, to be more specific, namely:

1. Count the number of classes / labels.
2. Count Number of Cases Per Class
3. Multiply All Class Variables
4. Result Analysis Per Class
5.

Below is the equation formula for Bayes' theorem:

$$P(H \mid X) = (P(X \mid H).P(H))/(P(X))$$

Where :
X : Data with unknown class
H : Hypothesis data is class specific
P(H| X) : Probability of hypothesis H based on condition X (posteriori probability)
P(H) : Probability of hypothesis H (prior probabilities)
P(X|H) : Probability of X according to condition on hypothesis H

To understand the Naive Bayes technique, it must be noted that the interaction characterization requires various guidelines to find out what class makes sense for the test being investigated. In this way, the above Naive Bayes technique is modified as follows:

$$P(C|F1 \ldots Fn) = \frac{P(C)P(F1 \ldots Fn|C)}{P(F1 \ldots Fn)}$$

Where Variable C discusses class, while variable F1 ... Fn discusses the quality of direction expected to play the set. Then the equation makes sense that the probability of the inclusion of an example of a certain quality in class C (Posterior) is the probability of the presence of class C (before the passage of the model, commonly referred to earlier), multiplied by the probability of occurrence of the test attribute in class C (also called probability), separated by the probability of development attribute test worldwide (also called proof). In this way, the above equation can also be arranged equation, namely:

$$Posterior = \frac{prior\ x\ likelihood}{evidence}$$

$$P(C|F_1, \dots, F_n = P(C)P(F_1, \dots, F_n|C)$$
$$= P(C)P(F_1|C)P(F_2, \dots, F_n|C, F_1)$$
$$= P(C)P(F_1|C)P(F_2|C, F_1)P(F_3, \dots, F_n|C, F_1, F_2$$
$$= (C)P(F_1|C)P(F_2|C, F_1)P(F_3|C, F_1, F_2)P(F_4, \dots, F_n|C, F_1, F_2, F_3)$$
$$= P(C)P(F_1|C)P(F_2|C, F_1)P(F_3|C, F_1, F_2) \dots P(F_n|C, F_1, F_2, F_3, \dots, F_{n-1})$$

The value of evidence (evidence) is generally assigned to each class in one example. The back-end values will be contrasted, and the back-end values will be reversed from the different classes to decide which class to classify as an example. Further elaboration of the Bayesian equation is solved by elaborating (|1,…,) using the accompanying duplication rule: It is quite possible to see that the consequences of elaboration lead to more complex elements affecting the probability price, which are very difficult to investigate individually. Therefore, estimation becomes challenging to do. This is where the very high presumption of freedom (innocentness) is used, that each clue (F1, F2...Fn) is autonomous from one another. With these assumptions, the accompanying parable is:

$$P(F_i|F_j) = \frac{P(F_i \cap F_j)}{P(F_j)} = \frac{P(F_i)P(F_j)}{P(F_j)} = P(F_i)$$

For i≠j , so

$$P(F_i|C, F_j) = P(F_i|C)$$

Above condition is a model of the Naive Bayes hypothesis which will then be used in the grouping system. For characterization with persistent information, the Gauss Density equation is used:

$$P(X_i = x_i|Y = y_j) = \frac{1}{\sqrt{2\pi\sigma_{ij}}} e^{-\frac{(x_i - \mu_{ij})^2}{2\sigma^2 ij}}$$

Where:
P : Probability
Xi : Attribute i
xi : Value of attribute i
Y : The class to be searched yi : Sub class Y to search
μ : mean, which contains the average of all attributes
σ : Standard Deviation, which contains the variance of all attributes.

3. Findings

Data were analyzed using the classification method to predict the allocation of time slots at Soekarno Hatta airport.

P-ISSN: 2655-8807
E-ISSN: 2656-8888

3.1 Data Preparation

The data collection here comes from a web based flight plan system and appraisal value data will be obtained from the appraisal system and exported to excel according to the excel template to be processed.

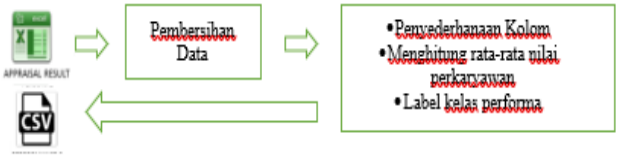

Figure 2. Export Dataset

3.2 Initial Data Processing

Next is the cleaning of information through flight information obtained through data collection, especially by removing unclear notes and repeating these records. it can also take redundant characteristics such as feature numbers, aircraft registrations, and flight numbers. This is done because this characteristic significantly affects the handling of information in the system. The following information is considered used:

C1 : Scheduled Flight
C2 : Un-schedule Flight
C3 : Domestic Flight
C4 : International Flight
C5 : PBN (performance based navigation)
C6 : Winter Season
C7 : Summer Season
C8 : Night Flight
C9 : Day Flight

And Each of the parameters has the following values:
1 : Not recommended
2 : Less Recommended
3 : Operational Considerations
4 : Recommended
5 : Highly Recommended

Data sets that have been exported from the web based flight plan system are as shown below:

P-ISSN: 2655-8807
E-ISSN: 2656-8888

FPL	C1	C2	C3	C4	C5	C6	C7	C8	C9	Hasil
1	1	5	5	5	5	5	5	5	5	Sangar Disarankan
2	1	1	1	2	2	2	3	2	2	Tidak Disarankan
3	4	3	5	5	5	5	5	3	5	Disarankan
4	5	5	5	5	5	5	5	5	5	Disarankan
5	4	4	5	5	5	5	4	4	4	Disarankan
6	5	5	5	5	3	4	4	5	5	Disarankan
7	4	4	4	5	5	5	4	5	5	Disarankan
8	4	5	4	5	4	5	4	5	4	Disarankan
9	5	5	5	5	5	5	4	4	4	Disarankan
10	5	5	5	5	4	4	5	5	4	Disarankan
11	2	1	3	2	2	2	2	3	1	Tidak Disarankan
12	4	4	4	4	4	4	5	5	5	Disarankan
13	5	5	5	5	5	5	5	5	5	Sangat Disarankan
14	4	5	4	5	4	5	4	3	5	Disarankan
15	3	5	5	5	5	5	4	4	5	Disarankan
16	5	5	5	5	4	4	4	4	4	Disarankan
17	5	4	5	4	5	4	5	5	4	Disarankan
18	4	4	5	5	4	4	5	5	5	Disarankan
19	4	5	4	5	4	5	5	4	4	Disarankan
20	5	3	3	5	5	5	5	5	5	Disarankan
21	5	4	5	5	4	3	3	5	5	Disarankan
22	2	5	5	5	5	5	5	5	5	Disarankan
23	5	3	5	5	3	5	5	5	5	Disarankan
24	4	5	5	4	5	4	5	4	4	Disarankan
25	4	4	3	3	3	3	3	4	4	Disarankan
26	4	3	3	3	3	3	3	4	4	Disarankan
27	3	2	3	3	3	3	3	3	3	Kurang Disarankan

Figure 3. Dataset

- Mining Models

Various algorithms and techniques of Classification, Clustering, Regression, Artificial Intelligence (AI), Neural Networks, Naive Bayes, Decision Trees, Genetic Algorithms, K-Neighbors etc. Classification is one of the most frequently studied problems by data mining and machine learning (ML) researchers, which consists of predicting the value of an attribute (category/class) based on the importance of other attributes (attribute prediction). There are different classification methods. In this research, we use the Decision Tree and Naive Bayes algorithm.

- Application of Decision Tree on Performance of Employee

A Decision Tree is an additional method of supporting decisions by applying a graph as a tree or decision model that will show or explain the likelihood that a child will occur. Decision trees are commonly used in research operations to make decisions and identify the best method for achieving specific goals. In this study, data processing will use Weka. For the first step, we choose an excel file to be processed using Weka. The first step will be like the picture below.

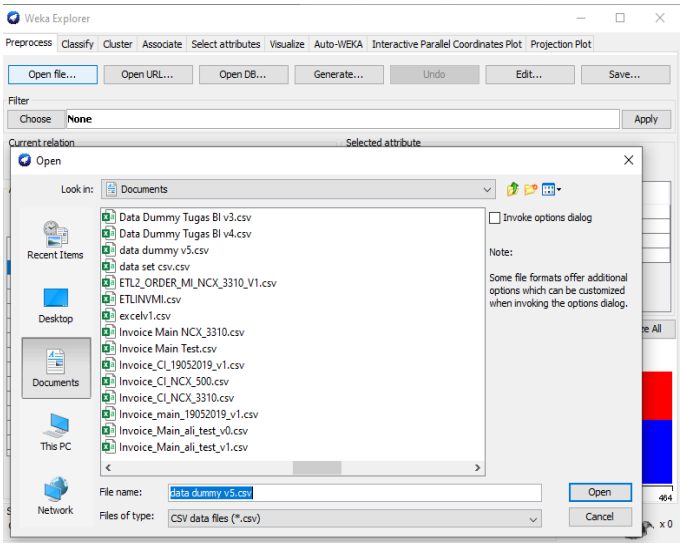

Figure 4. Import Weka Data

Then after importing data into Weka then select the algorithm, namely J48 on Weka as shown below:

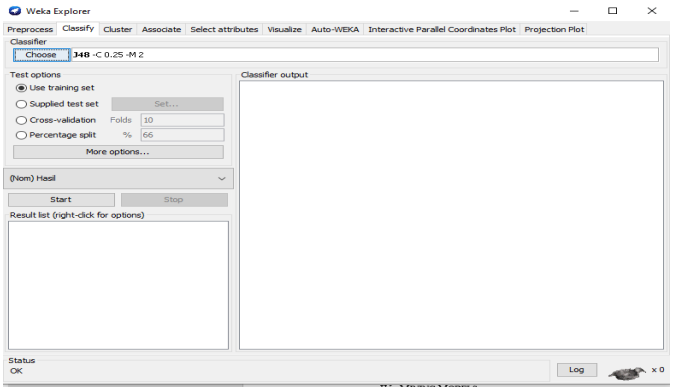

Figure 4. J48 Algorithm

For testing here use 2 options, namely use training set, cross validation folds 10 %,. For data processing use training set as shown below:

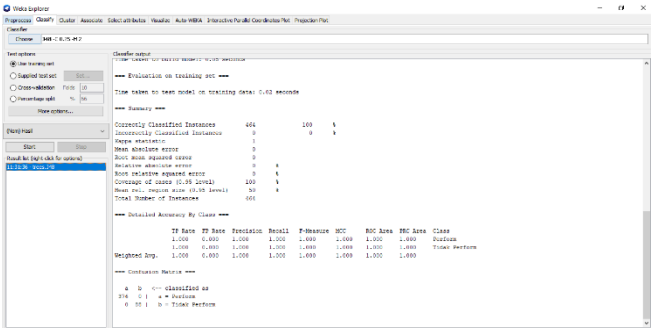

Figure 5. J48 Algorithm Use Training Set

and Visualization with Tree as shown below:

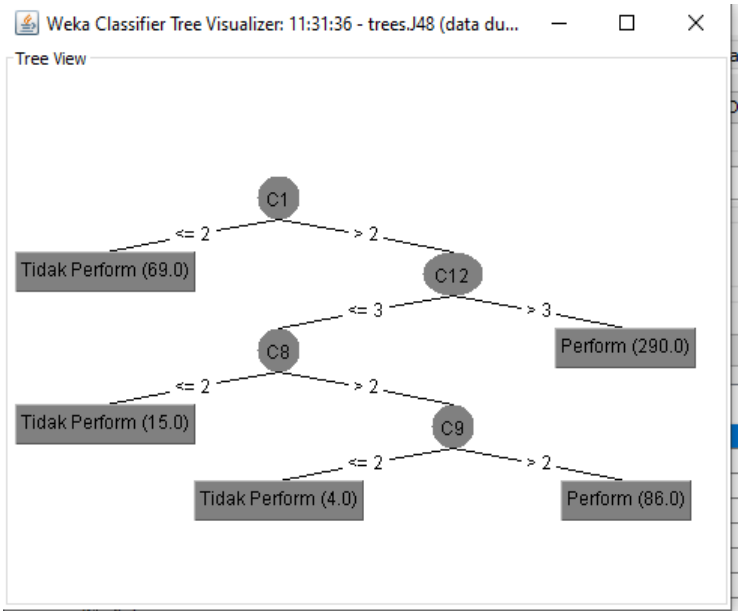

Figure 6 . Visualization of the J48 Algorithm Tree Use Training Set

For validation using cross validation folds 10% as shown below:

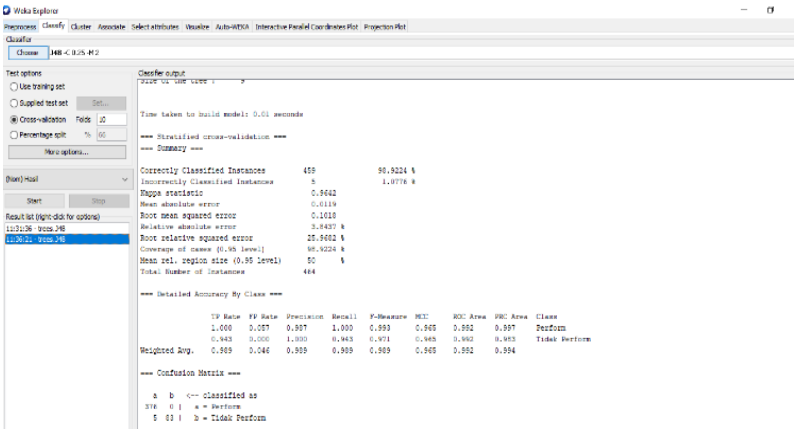

Figure 7. J48 Cross Validation Algorithm

The picture above is data processing using 10% cross validation. The advantages of using cross validation where the data to be displayed will be more accurate. The following is a visualization of the tree cross validation folds 10%

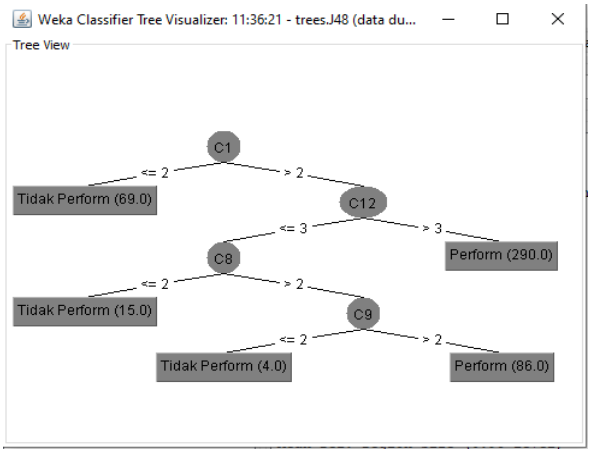

Figure 8. Visualization of the J48 Algorithm Tree Use Cross Validation

- Application of Naïve Bayes on Performance of Employee

For Naïve Bayes processing, it is almost the same as using the decision tree algorithm on Weka where you have to choose a predetermined excel file. Then choose an algorithm that is nave Bayes. For data processing itself, it is the same as using the training set and cross validation folds 10%. The following uses the use training set

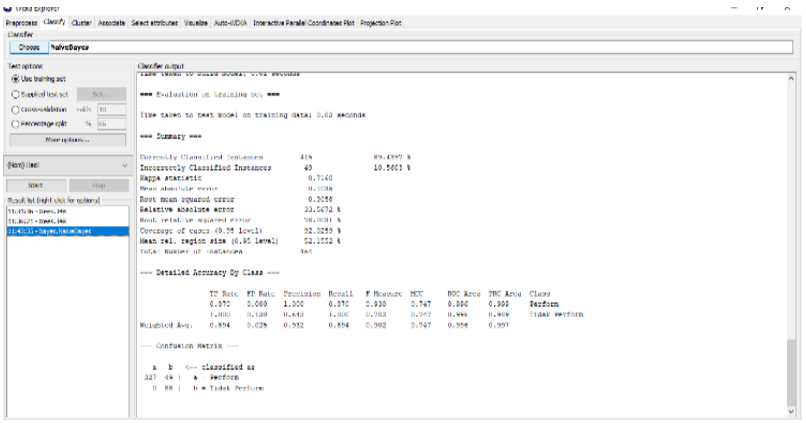

Figure 8. Data Processing Using Naïve Bayes Use Training Set

And images are carried using cross validation

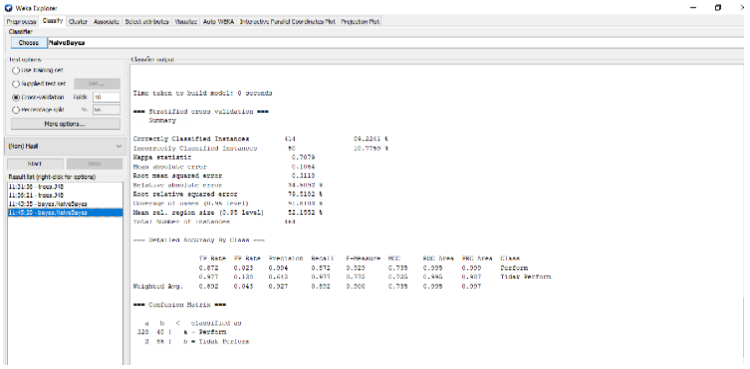

Figure 8. Data processing using Naive Bayes Cross Validation

- Model evaluation

For application to the decision tree algorithm and with the number of attributes after processing the source, there are nine attributes with a sum of the weight of 484 and instances of 464. From the results of the decision tree algorithm, it has correctly classified instances of around 98, 9224% and incorrectly classified instances of approximately 1.0076%. By using cross-validation and using cross-validation, we get 100% correct instances and 0% incorrect instances. . ROC Area performs 0.997. It does not perform around 0.983 using cross-validation, and the following rules are obtained:

R1: If the flight type is a scheduled flight and has performance-based navigation (PBN), it is a highly recommended aircraft.
R2: If the international flight is less than three and the unscheduled flight is more significant than two, and it has a recommended result
R3: If C1 is less than two, then it can be an operational consideration for giving the slot time
R4: If PBN is less than equal to 2, then the result is less recommended
R5: If C1 and C3 are less than two, then no, the result is not recommended

The results of the Naive Bayes algorithm have correctly classified instances at 87.931% and incorrectly classified models at 12.096%. ROC Area level performs 0.997 for and does not perform using cross-validation 10%. Utilising the use training set, the correct instance is 89.4937%, the incorrect model is 10.5063%, the roc area to perform is 0.996, and the roc area is 0.996.

4. Conclusion

For the application of the decision tree algorithm and with the number of attributes after processing the source, there are 15 attributes with a sum of the weight of 484 and 464 instances. Cross-validation and using cross-validation obtained 100% correct instances and 0% incorrect instances. ROC Area performs 0.997 and does not perform around 0.983. With a minor error, it can conclude that the decision tree algorithm is more accurate in predicting and seeing the available time slots as a whole and the space occupancy.

References

[1] I. Artamonov, N. Danilochkina, I. Pocebneva, and K. Karmokova, "Using data integrity models for aviation industry business process quality management," *Transp. Res. Procedia*, vol. 63, pp. 1668–1673, 2022.

[2] S. Bhargav and N. Mehra, "Study of employee attrition in business process outsourcing companies in India," *Int. J. Res. Soc. Sci.*, vol. 8, no. 9, pp. 348–358, 2018.

[3] A. Bitkowsk, "The relationship between Business Process Management and Knowledge Management-selected aspects from a study of companies in Poland," *J. Entrep. Manag. Innov.*, vol. 16, no. 1, pp. 169–193, 2020.

[4] P. Balakrishna, R. Ganesan, and L. Sherry, "Accuracy of reinforcement learning algorithms for predicting aircraft taxi-out times: A case-study of Tampa Bay departures," *Transp. Res. Part C Emerg. Technol.*, vol. 18, no. 6, pp. 950–962, 2010.

[5] G. Praetorius, F. van Westrenen, D. L. Mitchell, and E. Hollnagel, "Learning lessons in resilient traffic management: a cross-domain study of vessel traffic service and air traffic control," in *HFES Europe Chapter Conference Toulouse 2012*, 2012, pp. 277–287.

[6] S. Hamzah and S. A. Adisasmita, "Aircraft parking stands: proposed model for Indonesian airports," *Procedia Environ. Sci.*, vol. 28, pp. 324–329, 2015.

[7] R. Kurniawan, A. Sutawan, and R. Amalia, "Information System Ordering Online Restaurant Menu At Hover Cafe," *Aptisi Trans. Manag.*, vol. 4, no. 1, pp. 32–40, 2020.

[8] E. Chowns, "Is community management an efficient and effective model of public service delivery? Lessons from the rural water supply sector in Malawi," *Public Adm. Dev.*, vol. 35, no. 4, pp. 263–276, 2015.

[9] A. Carlin and R. E. Park, "Marginal cost pricing of airport runway capacity," *Am. Econ. Rev.*, vol. 60, no. 3, pp. 310–319, 1970.

[10] M. Doepke and M. Tertilt, "Does female empowerment promote economic development?," *J. Econ. Growth*, vol. 24, no. 4, pp. 309–343, 2019.

[11] A. Alwiyah, S. Sayyida, P. A. Sunarya, and D. Apriliasari, "Inovasi Manajemen Pengajuan Judul Kuliah Kerja Praktek (KKP) berbasis Laravel Framework," *Technomedia J.*, vol. 7, no. 2, pp. 168–180, 2022.

[12] I. Amsyar, E. Cristhopher, U. Rahardja, N. Lutfiani, and A. Rizky, "Application of Building Workers Services in Facing Industrial Revolution 4.0," *Aptisi Trans. Technopreneursh.*, vol. 3, no. 1, pp. 32–41, 2021