Available online at https://journal.rescollacomm.com/index.php/ijqrm/index

International Journal of Quantitative Research and Modeling

e-ISSN 2721-477X
p-ISSN 2722-5046

Vol. 1, No. 3, pp. 123-134, 2020

Data Mining Implementation Using Naïve Bayes Algorithm and Decision Tree J48 In Determining Concentration Selection

Budiman[1*], Reni Nursyanti[2], R Yadi Rakhman Alamsyah[3], Imannudin Akbar[4]

[1,2,3,4]*Department of Technology and Informatics, Universitas Informatika dan Bisnis Indonesia*.

** Corresponding author email: budiman1982@gmail.com*

Abstract

Computerization of society has substantially improved the ability to generate and collect data from a variety of sources. A large amount of data has flooded almost every aspect of people's lives. AMIK HASS Bandung has an Informatic Management Study Program consisting of three areas of concentration that can be selected by students in the fourth semester including Computerized Accounting, Computer Administration, and Multimedia. The determination of concentration selection should be precise based on past data, so the academic section must have a pattern or rule to predict concentration selection. In this work, the data mining techniques were using Naive Bayes and Decision Tree J48 using WEKA tools. The data set used in this study was 111 with a split test percentage mode of 75% used as training data as the model formation and 25% as test data to be tested against both models that had been established. The highest accuracy result obtained on Naive Bayes which is obtaining a 71.4% score consisting of 20 instances that were properly clarified from 28 training data. While Decision Tree J48 has a lower accuracy of 64.3% consisting of 18 instances that are properly clarified from 28 training data. In Decision Tree J48 there are 4 patterns or rules formed to determine concentration selection so that the academic section can assist students in determining concentration selection.

Keywords: Concentration, Classification, Naïve Bayes, Decision Tree J48

1. Introduction

The rapid development, of information technology, is undisputed. Along with these developments, all transaction data has been evolved by applying information technology. Thus the computerization of the community has substantially improved the ability to generate and collect data from various sources. Vast amounts of data have flooded almost every aspect of people's lives.

The growth of explosive data has been stored, while data has generated an urgent need for new techniques and automated tools that can help intelligently turn large amounts of data into useful information and knowledge. This led to the development of computer science called data mining with its various applications. More popular data mining referred to as Knowledge Data Discovery or KDD is automatic or practical pattern extraction that represents knowledge implicitly stored or captured in large databases, data warehouses, web, other large information repositories, or data streams (Larose, 2015).

AMIK HASS Bandung is a private college in Bandung. In carrying out the learning process AMIK HASS Bandung has an Informatics Management Study Program consisting of three areas of concentration that can be selected by students in the fourth semester including Computerized Accounting, Computer Administration, and Multimedia. Concentration selection is an effort to determine interest in improving the field of science and competency that will be chosen by students based on the results of consultation with their respective guardian lecturers. Also, the academic section will evaluate student data in the form of gender, GPA, and Class. This activity takes a long time because the determination of concentration selection must be precise based on past data, so the academic section must have a pattern or rule to predict concentration selection. To solve the problem several methods can be applied in concentration selection at AMIK HASS Bandung. In this work, the data mining techniques used are Naïve Bayes and Decision Tree J48 using WEKA tools. Based on the background above, the purpose of this work is how to determine the pattern or rules of concentration selection and how much accuracy the application of Naïve Bayes data mining algorithms and Decision Tree J48 in concentration selection predictions.

Previous relevant work has been done by Nematzadeh (2012), researchers try to classify researchers as "Expert" and "Novice" based on cognitive factors to get the best possible answers. The domain of this work is based on the academic environment. An important point of this work is to classify researchers based on the Naive Bayes technique and Decision Tree J48 ultimately choosing the best method based on the highest accuracy of each method to help researchers get the best feedback based on their demands in the digital library. Based on the best accuracy, it can be concluded that web developers can use Naïve Bayes or Naïve Bayes update techniques compared to Decision Tree J48 to classify researchers and help them to get the best feedback based on their demands in the digital world of libraries (Nematzadeh, 2012).

Further work was carried out by George Dimitoglou et al. (2012), who examined the accuracy of data mining and machine learning with Naïve Bayes and J48 algorithms to predict the survival of lung cancer patients. The study showed an accuracy rate of about 90% on one of Naïve Bayes and J48's algorithms. The results of such a treating doctor can theoretically collect some medical measurements such as tumor size and location, treatment options, and others to predict with a fairly high degree of accuracy whether the patient is likely to live for five years or more. Given the high mortality rate (> 90%) patients in the study can be utilized to examine the survivability of patients over a shorter period, between 12 and 18 months (Dimitoglou et al, 2012).

In addition, much work has been done in data mining techniques in the field of education in various cases including Merceron, A et al. has a case study on mining education data sets to identify the behavior of failing students and to warn students about the risks before the final exam (Merceron and Yacef, 2005). Al-Radaideh (2006) applied the decision tree to predict the final grades of students studying C++ Courses at Yarmouk University. Jordan. Romero et al. (2008). have done work in the application of data mining techniques for moodle course management and

data mining techniques that have been widely used for e-learning data mining. In addition, educational data mining work was carried out by Minaei-Bidgoli et al. (2003). Beikzadeh et al (2005) does work using educational data mining to identify and improve. It has been observed that there has been an improvement in the decision-making process. Waiyamai et al. (2003) in his work used data mining to help develop a new curriculum, and to help students choose the appropriate courses. Rao et al. (2016) work on learning models to predict student performance using classification techniques. It also shows comparative performance analysis of J48, naïve Bayesian classifier, and random forest algorithms.

Comparing data mining classification techniques is Algorithm C4.5, AODE, Naive Bayesian, K-Nearest Neighbor to analyze and predict student performance aimed at improving skills in achieving the final goal of the semester (Mayilvaganan and Kalpanadevi, 2014).

The study aims to determine hidden knowledge and patterns about student performance by applying two classification algorithms, KNN and Naive Bayes to the secondary school education data set at the Gaza Strip Environment Ministry in 2015. The main purpose of classification can be to help the ministry of education to improve the performance and initial prediction of student performance. Teachers can also take appropriate evaluations to improve student learning. Experiment results showed that Naïve Bayes was better than K-Nearest Neighbor by receiving the highest accuracy score of 93.6% (Amra and Maghari, 2017).

Further work was carried out by Devasia et al. (2016), the work aimed at developing a web-based application to utilize Naive Bayes techniques in retrieving information contained in the Higher Education database. The increase in the number of students who did not continue studying affects the reputation of educational institutions. The experiment was conducted on 700 students consisting of 19 attributes. Results prove that the Naive Bayes algorithm provides higher accuracy compared to other methods such as Regression, Decision Tree, Neural Network, and others.

The current work is different from the previous work, which determines the comparison and prediction of the selection of student concentration in the fourth semester using gender attributes, GPA, and class to help students in determining the concentration that should be taken.

2. Data Mining

Data mining is a process in analyzing the data of various perspective data and summarizing it to produce useful information. Technically the process of data mining is to find patterns and relationships in a large relational database. Data sources can include databases, data warehouses, the web, other repositories of information, or data that dynamically flows into the system. In large-scale information technology can develop transaction and analytical systems separately, in data mining provides a relationship between the two. Data mining can find new relationships and patterns in data. It is found in the areas of statistics, machine learning, artificial intelligence, and neural networks (Rao et al., 2016; Han et al., 2012).

3. Naïve Bayes

Naïve Bayes is a classification with probability and statistical methods put forward by British scientist Thomas Bayes (Han et al., 2012). This algorithm uses the Bayes theorem and assumes that all independent variables are class variable values. This method only requires the amount of training

data to determine the approximate parameters required in the Process classification. NBC often works much better in the most complex real-world situations than expected. Bayes theorem is a mathematical formula that used to determine conditional probability in equation 1 (Saritas and Yasar., 2019).

$$P(C_i \mid X) = \frac{P(X \mid C_i)P(C_i)}{P(X)} \quad (1)$$

Description:
P(Ci|X) = Probability of Ci hypothesis if given fact or record X (Posterior probability)
P(X|Ci) = look for parameter values that give the most likelihood
P(Ci) = Prior probability from X (Prior probability)
P(X) = Number of probability tuple that appears

4. Decision Tree J48

Decision Tree is one of the most intuitive and popular data mining methods, especially in providing explicit rules for proper classification and handling of heterogeneous data. The Decision Tree is on the line between predictive and descriptive methods.

The Decision Tree technique is used in classification to detect individual division criteria from population into specified classes (many cases n = 2) starting with selecting variables that based on the category to provide the best separation of individuals in each class, thus providing sub-populations called nodes, each containing the largest proportion of individuals in a single class. Then the same operation will be repeated on each newly acquired node until there is no further separation from the individual that may or is desired according to the criteria depending on the tree type.

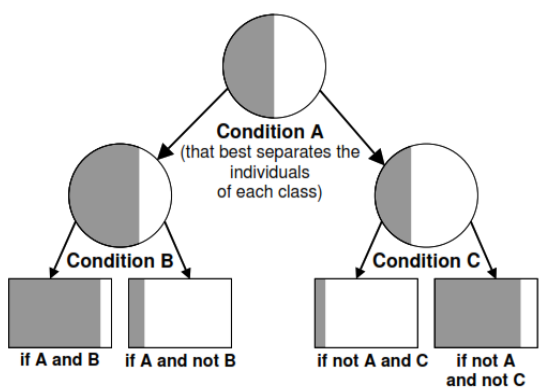

Figure 1: Decision Tree

Figure 1 is a Decision Tree that shows the induction of the decision tree building a flow chart-like structure in which each internal node (non-leaf) shows a test on the attribute, each branch

corresponds to the test result, and each external node (leaf) indicates the predicted class. On each node, the algorithm selects the "best" attribute to partition data into each class (Ye N, 2013). Decision Tree J48 is an implementation of the C4.5 algorithm developed by the WEKA project team.

5. Research Methods

This work is a test of student data that chooses the concentration in the fourth semester taken from the Academic Section which is poured in the form of a table. The data will be done twice using Naïve Bayes and Decision Tree J48 with the machine learning tool "WEKA".

The prediction made in this work is to determine the concentration chosen by a student who will take the study in the fourth semester with the following conditions:

a. Gender: whether male or female.
b. Class: whether the class is regular or non-regular
c. GPA: what is the GPA in the third semester with a range of <3 or >=3

These three conditions will predict students who will choose concentration as an interest namely Computer Administration, Computerized Accounting, or Multimedia by studying past events with various conditions.

The data in this work is 111 data sets with a 75% split test percentage mode used as training data as a model and 25% as test data to be tested against established models. The following is the concentration selection data that will be performed on each test, namely:

Table 1. Data Set

Number	Name	Gender	Class	GPA	Concentration
1	Aep Sofyan	Male	Regular	>= 3	Multimedia
2	Agus Aswandi	Male	Regular	< 3	Multimedia
3	Agus Kurnia	Male	Regular	>= 3	Administration Computer
4	Agustiana	Female	Regular	>= 3	Administration Computer
5	Ajeng Asrining Puri	Female	Regular	< 3	Computerized Accounting
6	Ali Akbar Rausyan Fikri	Male	Non-Regular	>= 3	Administration Computer
7	Amelia Widhiayuni Safitri	Female	Regular	>= 3	Multimedia
8	Andi Supriyatna	Male	Regular	>= 3	Administration Computer
9	Anggy Sulastiani	Female	Non-Regular	>= 3	Multimedia
10	Anisafitri	Female	Non-Regular	>= 3	Computerized Accounting
11	Anissa Anggraeni	Female	Regular	< 3	Multimedia

12	Apep Bayu Gunawan	Male	Regular	>= 3	Computerized Accounting
13	Aprilianti Karim	Female	Regular	>= 3	Computerized Accounting
14	Arief Kusnandar	Male	Regular	>= 3	Computerized Accounting
…	…	…	…	…	…
109	Zaeni Wahab	Male	Regular	< 3	Multimedia
110	Zahra Ghaisani Arifah	Female	Regular	>= 3	Computerized Accounting
111	Zella Adiga Pertiwi	Female	Regular	>= 3	Computerized Accounting

Table 1 is referred as a data set consisting of 4 attributes and 75% of the 111 records to be tested namely Gender {Male, Female}, Class {Regular, Non-Regular}, GPA {<3, >=3} and Concentration {Computer Administration, Computerized Accounting, Multimedia}. A total of 25% of the 111 records will be predicted and compare to the accuracy of the two experiments conducted.

6. Results And Discussions

Prediction testing was conducted using two classification techniques namely Naïve Bayes and Decision Tree J48. Here are the test results against the training set.

6.1. Naïve Bayes Classification

Figure 2 is the result of Naïve Bayes classification testing of training sets, testing is done by the same method on Decision Tree J48.

```
Correctly Classified Instances          20               71.4286 %
Incorrectly Classified Instances         8               28.5714 %
Kappa statistic                          0.5077
Mean absolute error                      0.3264
Root mean squared error                  0.387
Relative absolute error                 75.5144 %
Root relative squared error             82.4332 %
Total Number of Instances               28

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.769    0.267    0.714      0.769    0.741      0.501    0.810     0.774     Multimedia
                 0.000    0.042    0.000      0.000    0.000      -0.079   0.792     0.302     Administration Computer
                 0.909    0.176    0.769      0.909    0.833      0.717    0.882     0.747     Computerized Accounting
Weighted Avg.    0.714    0.199    0.634      0.714    0.671      0.503    0.836     0.696
```

Figure 2: Naïve Bayes Test Results

In Figure 2 the test results using the Naïve Bayes classification have an accuracy rate of 71.4% which states the correct prediction ratio with the overall testing set tested, while the absolute error means 0.3264. The proximity of the data in the Multimedia class of 71.4% shows that the correct percentage of students choosing the multimedia concentration of the entire student predicted chose multimedia concentration. The Computerized Accounting Class has data proximity of 76.9% indicating that the correct percentage of students choosing the concentration of Computerized Accounting from the entire student predicted chose the concentration of Computerized Accounting.

A return score for multimedia classes of 76.9% indicates that the percentage of students predicted chose multimedia concentration over students who chose multimedia concentration. While the return of grades for Computerized Accounting class of 90.9% indicates that the percentage of students who are predicted to choose the concentration of Computerized Accounting versus the overall student who chose the concentration of Computerized Accounting.

```
=== Confusion Matrix ===

  a  b  c   <-- classified as
 10  1  2 |  a = Multimedia
  3  0  1 |  b = Administration Computer
  1  0 10 |  c = Computerized Accounting
```

Figure 3: Confusion Matrix Naïve Bayes

Figure 3 is confusion matrix Naïve Bayes, the first line there is "10 1 2" indicating that there are multimedia class instances in the testing set of 10 correctly predicted as Multimedia, 1 is incorrectly classified as Computer Administration and 2 are incorrectly classified as Computer Administration. In the second line, there is a "3 0 1" indicating that there are instances of the Computer Administration class in the testing set of 3 incorrectly classified as Multimedia and 1 classified as Computerized Accounting. In the third line, there is a "1 0 10" indicating that there is an instance of the Computerized Accounting class in the testing set and 1 is incorrectly classified as Multimedia, and 10 is correctly predicted as Computerized Accounting. Figure 4 shows the predicted results using the Naïve Bayes classification.

No.	1: Gender Nominal	2: Class Nominal	3: GPA Nominal	4: prediction margin Numeric	5: predicted Concentration Nominal	6: **Concentration** Nominal
1	Male	Regular	)= 3	0.006545	Multimedia	Multimedia
2	Female	Regular	)= 3	0.585995	Computerized Accounting	Computerized Accounting
3	Female	Regular	)= 3	0.585995	Computerized Accounting	Computerized Accounting
4	Male	Regular	)= 3	0.006545	Multimedia	Multimedia
5	Female	Non-Regular	)= 3	0.44851	Computerized Accounting	Computerized Accounting
6	Male	Regular	)= 3	-0.006545	Multimedia	Computerized Accounting
7	Female	Regular	)= 3	0.585995	Computerized Accounting	Computerized Accounting
8	Male	Regular	)= 3	-0.038152	Multimedia	Administration Computer
9	Male	Non-Regular	(3	0.520129	Multimedia	Multimedia
10	Male	Regular	(3	0.554137	Multimedia	Multimedia
11	Female	Regular	)= 3	0.585995	Computerized Accounting	Computerized Accounting
12	Female	Non-Regular	)= 3	-0.44851	Computerized Accounting	Administration Computer
13	Female	Regular	)= 3	0.585995	Computerized Accounting	Computerized Accounting
14	Female	Regular	)= 3	-0.585995	Computerized Accounting	Multimedia
15	Female	Non-Regular	)= 3	0.44851	Computerized Accounting	Computerized Accounting
16	Female	Regular	)= 3	0.585995	Computerized Accounting	Computerized Accounting
17	Male	Regular	(3	0.554137	Multimedia	Multimedia
18	Female	Regular	(3	0.097174	Computerized Accounting	Computerized Accounting
19	Male	Regular	)= 3	0.006545	Multimedia	Multimedia
20	Female	Regular	)= 3	-0.585995	Computerized Accounting	Multimedia
21	Male	Non-Regular	)= 3	-0.164267	Administration Computer	Multimedia
22	Male	Regular	)= 3	0.006545	Multimedia	Multimedia
23	Male	Regular	(3	0.554137	Multimedia	Multimedia
24	Male	Regular	)= 3	-0.038152	Multimedia	Administration Computer
25	Female	Regular	)= 3	0.585995	Computerized Accounting	Computerized Accounting
26	Male	Regular	)= 3	0.006545	Multimedia	Multimedia
27	Male	Regular	)= 3	-0.038152	Multimedia	Administration Computer
28	Male	Non-Regular	(3	0.520129	Multimedia	Multimedia

Figure 4: Results Prediction on arffview for Naïve Bayes

6.2. Decision Tree J48

Figure 6 is the result of Decision Tree J48 testing against the testing set. Testing was conducted in the same method against Naive Bayes.

```
=== Evaluation on test split ===

Time taken to test model on test split: 0 seconds

=== Summary ===

Correctly Classified Instances          18               64.2857 %
Incorrectly Classified Instances        10               35.7143 %
Kappa statistic                          0.4776
Mean absolute error                      0.3257
Root mean squared error                  0.3878
Relative absolute error                 75.3568 %
Root relative squared error             82.5961 %
Total Number of Instances               28

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.385    0.000    1.000      0.385    0.556      0.501    0.841     0.795     Multimedia
                 0.750    0.292    0.300      0.750    0.429      0.335    0.755     0.268     Administration Computer
                 0.909    0.176    0.769      0.909    0.833      0.717    0.853     0.735     Computerized Accounting
Weighted Avg.    0.643    0.111    0.809      0.643    0.647      0.562    0.833     0.696
```

Figure 5: Decision Tree J48 Test Results

Figure 5 shows the test results using the Decision Tree J48 classification having an accuracy rate of 64.3% with an absolute error of 0.3257 stating the correct prediction ratio with the overall testing set tested.

The proximity of the data in the Multimedia class is 100% indicating that the correct percentage of students choosing the multimedia concentration of the entire student is predicted to choose the multimedia concentration. The Computerized Accounting Class has data proximity of 76.9% indicating that the correct percentage of students choosing the concentration of Computerized Accounting from the entire student predicted chose the concentration of Computerized Accounting. While the Computer Administration Class has data proximity of 30% indicates that the correct percentage of students choosing computer administration concentrations from all students is predicted to choose the concentration of Computer Administration.

```
=== Confusion Matrix ===

  a  b  c   <-- classified as
  5  6  2 |  a = Multimedia
  0  3  1 |  b = Administration Computer
```

Figure 6: Confusion Matrix Decision Tree J48

The returned score for multimedia classes of 38.5% indicates that the percentage of students predicted to choose multimedia concentration over actual students chose multimedia concentration. The grade returned for the Computer Administration class of 75% indicates that the percentage of students who are predicted to choose the concentration of computer administration rather than the actual student chooses the multimedia concentration. While the returned grades for computerized accounting classes of 90.9% show that the percentage of students who predicted choosing computerized accounting concentration versus students as a whole chose computerized accounting concentration. Figure 6 shows the predicted results using the Decision Tree J48 classification.

Based on the data in figure 6 obtained in the confusion matrix for decision tree J48 classification, the first line there is "5 6 2" indicating that there are multimedia class instances in the testing set among them 5 correctly predicted as Multimedia, 6 are incorrectly classified as Computer Administration and 2 are incorrectly classified as Computer Administration. In the second line, there is a "0 3 1" indicating that there is an instance of the computer administration class in the test set 3 correctly predicted as computer administration and 1 is incorrectly classified as computer administration. In the third line, there is a "0 1 10" indicating that there is an instance of computerized accounting class in the test set 1 incorrectly classified as computerized accounting and 10 is correctly predicted as computerized accounting. Figure 7 is the predicted result of 28 data tests using the Decision Tree J48 classification.

No.	1: Gender Nominal	2: Class Nominal	3: GPA Nominal	4: prediction margin Numeric	5: predicted Concentration Nominal	6: **Concentration** Nominal
1	Male	Regular	)= 3	-0.051282	Administration Computer	Multimedia
2	Female	Regular	)= 3	0.552632	Computerized Accounting	Computerized Accounting
3	Female	Regular	)= 3	0.552632	Computerized Accounting	Computerized Accounting
4	Male	Regular	)= 3	-0.051282	Administration Computer	Multimedia
5	Female	Non-Regular	)= 3	0.552632	Computerized Accounting	Computerized Accounting
6	Male	Regular	)= 3	-0.102564	Administration Computer	Computerized Accounting
7	Female	Regular	)= 3	0.552632	Computerized Accounting	Computerized Accounting
8	Male	Regular	)= 3	0.051282	Administration Computer	Administration Computer
9	Male	Non-Regular	(3	0.333333	Multimedia	Multimedia
10	Male	Regular	(3	0.333333	Multimedia	Multimedia
11	Female	Regular	)= 3	0.552632	Computerized Accounting	Computerized Accounting
12	Female	Non-Regular	)= 3	-0.578947	Computerized Accounting	Administration Computer
13	Female	Regular	)= 3	0.552632	Computerized Accounting	Computerized Accounting
14	Female	Regular	)= 3	-0.552632	Computerized Accounting	Multimedia
15	Female	Non-Regular	)= 3	0.552632	Computerized Accounting	Computerized Accounting
16	Female	Regular	)= 3	0.552632	Computerized Accounting	Computerized Accounting
17	Male	Regular	(3	0.333333	Multimedia	Multimedia
18	Female	Regular	(3	0.552632	Computerized Accounting	Computerized Accounting
19	Male	Regular	)= 3	-0.051282	Administration Computer	Multimedia
20	Female	Regular	)= 3	-0.552632	Computerized Accounting	Multimedia
21	Male	Non-Regular	)= 3	-0.051282	Administration Computer	Multimedia
22	Male	Regular	)= 3	-0.051282	Administration Computer	Multimedia
23	Male	Regular	(3	0.333333	Multimedia	Multimedia
24	Male	Regular	)= 3	0.051282	Administration Computer	Administration Computer
25	Female	Regular	)= 3	0.552632	Computerized Accounting	Computerized Accounting
26	Male	Regular	)= 3	-0.051282	Administration Computer	Multimedia
27	Male	Regular	)= 3	0.051282	Administration Computer	Administration Computer
28	Male	Non-Regular	(3	0.333333	Multimedia	Multimedia

Figure 7: Predicted Results on arffview for Decision Tree J48

Figure 8 shows the visualization of the tree formed from the Decision Tree J48 classification model.

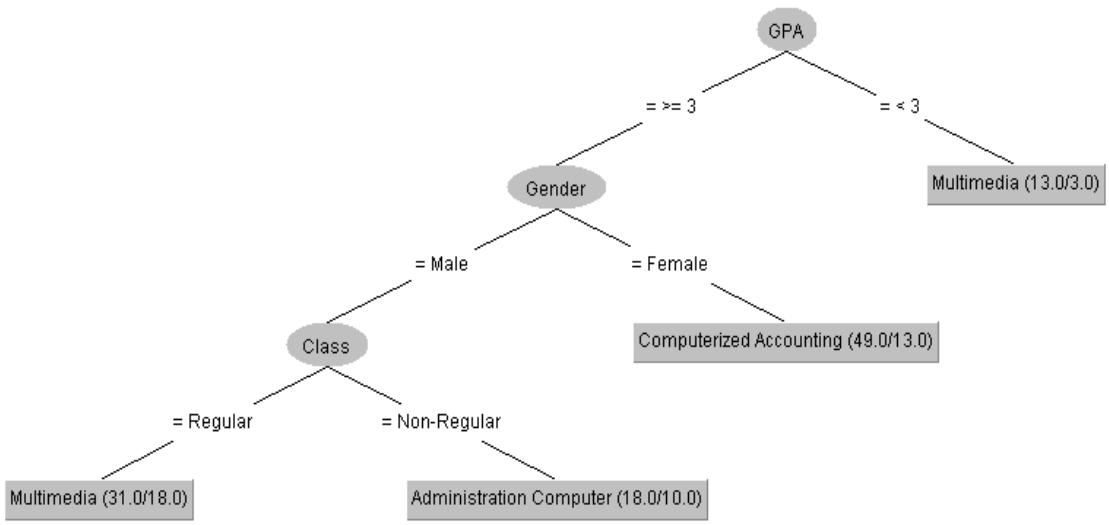

Figure 8: Tree Visualization Results

Based on the results of the visualization of the tree in figure 9 then the pattern or rules formed in the Decision Tree classification are as follows:

- IF "GPA >=3" AND "Gender = Male" AND "Class = Regular" THEN "Multimedia"
- IF "GPA >=3" AND "Gender = Male" AND "Class = Non-Regular" THEN "Computer Administration"
- IF "GPA>=3" AND "Gender = Female" THEN "Computerized Accounting"
- IF "GPA < 3" THEN "Multimedia"

7. Model Evaluation

After analyzing the results, table 2 shows the difference between the two algorithms in the test against the data set.

Table 2. Evaluation of Two Models

Criteria	Naïve Bayes	Decision Tree J48	Results
Correctly Classified Instances	71.42 %	64.28 %	Naïve Bayes
Incorrectly Classified Instances	28.57 %	35.71 %	Naïve Bayes
Accuracy	0.714	0.643	Naïve Bayes
Precision for Multimedia	0.714	1.000	Decision Tree J48
Precision for Computer Administration	0.000	0.300	Decision Tree J48
Precision for Computerized Accounting	0.769	0.769	-

Table 2 above shows that Naïve Bayes is a good model for Correctly Classified Instances, Incorrectly Classified Instances, and Accuracy criteria. While for precision criteria for Multimedia, and Precision for Computer Administration model Decision Tree J48 shows better.

Table 3 shows the difference in the average proximity of the data, the average return of the value, and the length of time in which it is required in the classification process.

Table 3. Differences In Average Precision, Average Recall Time Taken

Classifier	Average Precision	Average Recall	Time Taken (s)
Naïve Bayes	0.634	0.714	0
Decision Tree J48	0.809	0.643	0.02

Table 3 shows the average proximity of data generated by Naïve Bayes by 63.4% and Decision Tree J48 is higher at 81% which states that the average percentage of students choosing the concentration of all students predicted. While the average return of grades produced on Naïve Bayes was 71.4% and Decision Tree J48 was lower which was 64.3% stating that the average percentage of students predicted in the selection of a concentration compared to the overall students who chose that concentration. While the time it takes to build a model on Naïve Bayes takes 0 seconds and Decision Tree J48 takes 0.02 seconds.

8. Conclusion

Based on test results using Naïve Bayes and Decision Tree J48 with split percentage mode in the same data set, some conclusions can be drawn as follows:

1. There are 4 patterns or rules formed to determine the selection of concentration so that the academic section can assist students in determining concentration selection.
2. While the Decision Tree J48 classification has a lower accuracy of 64.3% consists of 18 instances that are clarified correctly from 28 training data. While the mean absolute error value in the Decision Tree J48 classification has a lower value. The smaller the absolute error mean value, the better the classification model.

References

Al-Radaideh, Q. A., Al-Shawakfa, E. M., and Al-Najjar, M. I. (2006, December). Mining student data using decision trees. In *International Arab Conference on Information Technology (ACIT'2006), Yarmouk University, Jordan*.

Amra, I. A. A., and Maghari, A. Y. (2017, May). Students performance prediction using KNN and Naïve Bayesian. In *2017 8th International Conference on Information Technology (ICIT)* (pp. 909-913). IEEE.

Beikzadeh, M., & Delavari, N. (2005). A new analysis model for data mining processes in higher educational systems. *On the proceedings of the 6th Information Technology Based Higher Education and Training*, 7-9.

Devasia, T., Vinushree, T. P., and Hegde, V. (2016, March). Prediction of students performance using Educational Data Mining. In *2016 International Conference on Data Mining and Advanced Computing (SAPIENCE)* (pp. 91-95). IEEE.

Dimitoglou, G., Adams, J. A., and Jim, C. M. (2012). Comparison of the C4. 5 and a Naïve Bayes classifier for the prediction of lung cancer survivability. *arXiv preprint*, *4*(8), 1-9.

Han, J., Kamber, M., and Pei, J. (2012). Data mining: concepts and techniques. *Morgan Kaufman Publishers*, *10*, 978-981.

Larose, D. T. (2015). *Data mining and predictive analytics*. New York: John Wiley & Sons.

Mayilvaganan, M., and Kalpanadevi, D. (2014, December). Comparison of classification techniques for predicting the performance of student academic environment. In *2014 International Conference on Communication and Network Technologies* (pp. 113-118). IEEE.

Merceron, A., & Yacef, K. (2005, May). Educational Data Mining: a Case Study. In *AIED* (pp. 467-474).

Minaei-Bidgoli, B., Kashy, D. A., Kortemeyer, G., & Punch, W. F. (2003, November). Predicting student performance: an application of data mining methods with an educational web-based system. In *33rd Annual Frontiers in Education, 2003. FIE 2003.* (Vol. 1, pp. T2A-13). IEEE.

Nematzadeh, B. Z. (2012). Comparison Of Decision Tree And Naive Bayes Methods In Classification Of Researcher's Cognitive Styles, *Academic Environment*, *3*(2), 23-34.

Rao, K. P., Rao, M. C., and Ramesh, B. (2016). Predicting learning behavior of students using classification techniques. *International Journal of Computer Applications*, *139*(7), 15-19.

Romero, C., Ventura, S., and García, E. (2008). Data mining in course management systems: Moodle case study and tutorial. *Computers & Education*, *51*(1), 368-384.

Saritas, M. M., and Yasar, A. (2019). Performance analysis of ANN and Naive Bayes classification algorithm for data classification. *International Journal of Intelligent Systems and Applications in Engineering*, *7*(2), 88-91.

Waiyamai, K. (2003). Improving quality of graduate students by data mining. *Department of Computer Engineering, Faculty of Engineering, Kasetsart University, Bangkok.*

WEKA. [Online]. [cited 2020 August 14. Available from: https://waikato.github.io/weka-wiki/not_so_faq/j48_numbers/.

Ye, N. (2013). *Data mining: theories, algorithms, and examples.* New York: CRC press.