Published *online* on the journal's web page : http://innovatics.unsil.ac.id

Innovation in Research of Informatics (INNOVATICS)

| ISSN (Online) 2656-8993 |

Implementation of Data Mining at Laboratory Vocational High School Using the C4.5 Algorithm to Predict Students Major Preferences

Nurisya Rahma Suherman[1], Ruuhwan[2], Aso Sudiarjo[3]

[1,2,3]Department of Informatics, Perjuangan University, Tasikmalaya 46115, Indonesia
[1]1903010085@unper.ac.id, [2]ruuhwan@unper.ac.id, [3]asosudiarjo@unper.ac.id

ARTICLE INFORMATION

Article History:
Received: October 30, 2023
Last Revision: November 13, 2023
Published *Online*: November 15, 2023

KEYWORDS

Accuracy,
C.45 Algorithm,
Data Mining,
Decision Tree,
RapidMiner

CORRESPONDENCE

Phone: +6285861425367

E-mail: *1903010085*@unper.ac.id

ABSTRACT

Education or the learning process is the primary thing for human life. Therefore, a place for acquiring knowledge is established, which is called a school. Schools have their own levels, ranging from early childhood education to higher education institutions. When students enter high school, they are required to make decisions in choosing their majors. Accompanied by technological advancements, the issues in high school major selection can be effectively and efficiently addressed using data mining. Common issues that usually arise include lack of accuracy, precision, and requiring a significant amount of time. Hence, the issues within major selection necessitate the use of data mining, employing the C4.5 algorithm method, to determine the accuracy and precision of large datasets. This research achieved with RapidMiner the result is accuracy score of 94.44%, precision of 81.37%, and sensitivity of 74.00%. Additionally, it also generated a decision tree and with Python has an accuracy of 93% because it automatically rounds the values, so there is no significant difference between the two tools. This proves that the C4.5 algorithm produces fairly accurate performance.

1. INTRODUCTION

Education or the learning process is fundamental to human life, as the education system provides extensive knowledge and intellectual enlightenment to the nation's life [1]. Moreover, through education, individuals can develop their potentials with the knowledge and experiences gained in their daily lives [2]. Hence, a place for acquiring knowledge is established, which is called a school, as the creation of schools is highly necessary for the students' future development. Schools have their own levels, ranging from early childhood education to higher education institutions. When students reach high school, they are required to make decisions in choosing their majors [3].

Major selection is a mandatory choice for students who intend to enter the realm of high school [4]. This process assesses and directs students' abilities in developing, enhancing, and deepening the necessary skills and academic values. Usually, students choose their majors based on their interests, academic capabilities, or a sense of curiosity to explore new territories previously unknown to them. Some students might also find themselves uncertain about which major to pursue. Thus, the school should provide guidance and motivation to ensure that students do not make misguided choices. Major selection is highly beneficial in honing students' skills for their future, whether they wish to pursue higher education or embark on careers in their desired fields.

Students who encounter difficulties in selecting a major might be struggling due to lack of self-confidence and poor academic performance [5]. Accompanied by technological advancements, the issues in high school major selection can be effectively and efficiently addressed by utilizing available technology. Common issues that usually arise include lack of accuracy, precision, and requiring a significant amount of time. Therefore, the challenges within major selection necessitate the use of data mining with classification techniques using the algorithm to determine the accuracy and precision of large datasets [6].

In this research, data mining use as process of extracting information or knowledge from large datasets of

student's major preferences, which will later prove invaluable for analyzing or predicting outcomes in specific future situations with implementing C.45 algorithm [7]. Data mining involves extracting information from a dataset by searching for patterns or specific rules, often requiring a substantial amount of data. It encompasses various techniques, methods, and algorithms, one of which is classification techniques, with decision tree methods commonly known as decision trees.

2. RELATED WORK

In a previous study conducted by [8] describe the prediction of interest in major selection of high school students using Naïve Bayes algorithm and the results showed that with the use of the Naïve Bayes algorithm on student data to predict major preferences, an accuracy rate of 93.75%, precision rate of 83.33%, and recall rate of 100% were achieved [9].

The majors of senior high school students are determined based on test scores, academic score, readability, and talent. This research [10] focuses on compiling academic scores for science and social knowledge. Different algorithms like C4.5, Naive Bayes, K-NN, Rule Induction, and others can help with classification. Cross validation and T-Test were used to compare algorithms. Naïve Bayes was found to be the best algorithm with 79.51% accuracy and AUC value of 0.861.

In research [11] explain the utilization of the Naïve Bayes and K-Nearest Neighbor Algorithms for Class XI Student Major Classification. The algorithm applied was Naïve Bayes and K-Nearest Neighbor to compare the highest accuracy values. The data used are 277 records and 4 attributes. The results with the Naïve Bayes Algorithm produced an accuracy of 81.82%, and 55 data out of 277 data were used. Meanwhile the K-Nearest Neighbor algorithm obtained an accuracy of 92.73% with the same amount of data. The results of the best overall algorithm with 1st place are K-Nearest Neighbor and 2nd is Naïve Bayes.

Majoring is an effort to help and guide students in choosing a specialization or major at school with special studies that will be of interest to the student. Based on a survey at SMA Negeri 15 Pekanbaru [12], the process of majoring students had problems such as difficulties experienced by the school in analyzing and evaluating manually when determining student majors one by one. This will of course take up a lot of time and energy. Data Mining there is a classification technique that is used to classify data so that it makes it easier to classify student majors [13]. Classification was carried out on the student data using the K-Nearest Neighbor (KNN) algorithm with RapidMiner tools. The classification modeling results obtained were then compared with the simulation parameters with the maximum accuracy results at the value k=3 with optimal accuracy results of 93.52%, average precision of 88.14%, and average recall of 100.00%.

So, this research will apply the C4.5 algorithm to predict majors in high school students, according to their interests and abilities. With the existing problems, it is hoped that the C4.5 algorithm will be able to become a tool in majoring high school students, of course also using RapidMiner as a tool that supports the classification of students' majors.

3. METHODOLOGY

This research has flow stages, as shown in Figure 1 below for this research.

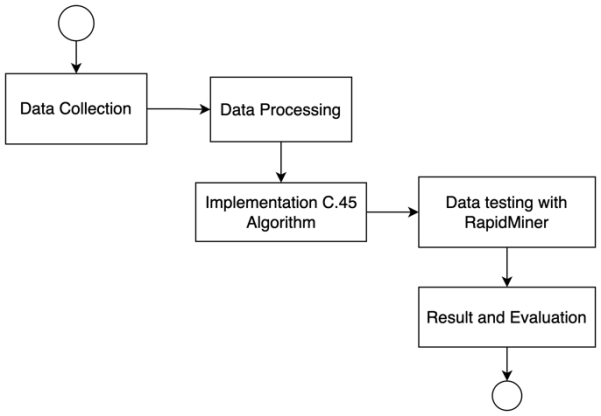

Figure 1. Methodology

3.1 Data Collection

The data used were primary data obtained directly from SMK Laboratory Jakarta, resulting in a total of 544 student data with 7 attributes.

3.2 Data Processing

The training dataset is a set of data used to train or build a model. The validation dataset is used to optimize the model during training. The model is trained using the training dataset, and its performance is tested using the validation dataset. The testing dataset is used to evaluate the model after the training process is complete. This data is unseen, meaning both the model and humans should not have seen these samples during training. It's important that the training, validation, and testing datasets are representative samples for the given problem.

3.3 Implementation of C.45 Algorithm

The C4.5 Algorithm is one of the classification algorithms. It's used to determine the accuracy level of predicting large datasets. The algorithm required to build a decision tree is the C4.5 algorithm, as it can generate rules and a decision tree to improve the accuracy of predictions. Some developments of the C4.5 algorithm include handling missing values, dealing with continuous data, and pruning.

3.4 Data Testing with Rapidminer

RapidMiner is a data mining software used for modeling that generates rules or patterns. It can extract and identify information within large datasets. For example, it can aid the process of student major selection using the C4.5 Algorithm.

3.5 Result and Evaluation

The results obtained from this data processing show that there are influencing factors after using RapidMiner. These factors can serve as an evaluation for school management. Its advanced analytics capabilities, serves as a catalyst for data-informed decision-making, ushering in a new era of efficiency and effectiveness in school management practices.

4. RESULT AND DISCUSSION

After conducting interviews with the school authorities, a dataset of prospective students along with their scores was obtained, which constitutes essential data for this research. The following is the dataset of prospective students obtained for the data mining process in data collection.

4.1 Data Collection

Student data.

TABLE 1. DATA OF STUDENTS

Students Name	Eng	Math	Indo	IPA	Result
A Gani Seya	78	81	85	85	Accepted
Muhammad Iqbal	75	74	82	86	Accepted
Fajri Abdi	84	91	83	84	Accepted
Abdillah	88	71	86	87	Accepted
Abdul Ahmad	78	97	87	88	Accepted
…	…	…	…	…	…
Ade Syaira	72	60	75	65	No Accepted
Ade Sufakni	70	66	76	64	No Accepted
Yogi Akhmad	91	95	85	83	Accepted
Yossa Faras	78	95	84	86	Accepted
Yossy Prili	95	95	85	88	Accepted
Yosua Mariga	86	97	88	88	Accepted
Yuda Addy	93	96	87	88	Accepted
Yusril Muzack	71	73	62	81	No Accepted
Zakaria Lordi	93	92	61	85	Accepted
Zaky Alatief	72	74	84	81	Accepted
Zalfar Azri	76	73	82	60	No Accepted
Zildan	97	63	90	80	Accepted
Zulkifli Farihan	94	95	88	84	Accepted

4.2 Data Processing

Data selection from prospective students at Laboratory Vocational Schools used in this research were 544 datasets. The student data table shows the variables contained in the dataset, including academic scores, interview tests and health tests. This major prediction will be useful for schools in determining the appropriate major for prospective students. Once the data has been collected completely, the next step is to ensure that there are no copies of the data and ensure that the data contains values or is not empty so as not to hinder the classification process. The data transformation carried out is in the form of an excel file so that it can be operated directly into the RapidMiner application.

4.3 Implementation of C.45 Algorithm

Data testing was carried out with the RapidMiner application. The following steps are carried out during

TABLE 1. DATA OF ENTROPY AND GAIN

Subject	Accepted	No Accepted	Entropy	Gain
English >70	492	31	0.3245539	0.0416916
English <70	10	11	0.9983636	
Math >70	499	27	0.2647693	0.0869414
Math <70	15	15	1	
Indo >70	489	28	0.3038064	0.0539455
Indo <70	13	14	0.9990102	
IPA >70	476	25	0.2859728	0.0523604
IPA <70	26	17	0.9681647	

4.4 Data Testing with RapidMiner

After carrying out the calculations manually, the next stage is testing the data in the RapidMiner software using the same dataset as the manual calculations. The following is how to apply student data to the RapidMiner application.

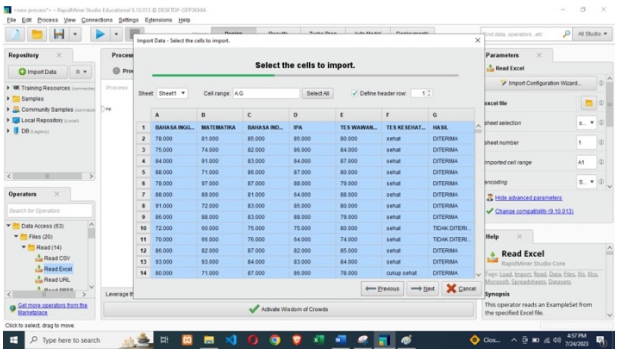

FIGURE 2. CHOOSING COLUMNS AND ROW

The next stage, drag the Decision Tree operator and apply the model to the Main Process. This Apply model is used as a prediction on training data and as a link between the Decision tree operator and Performance.

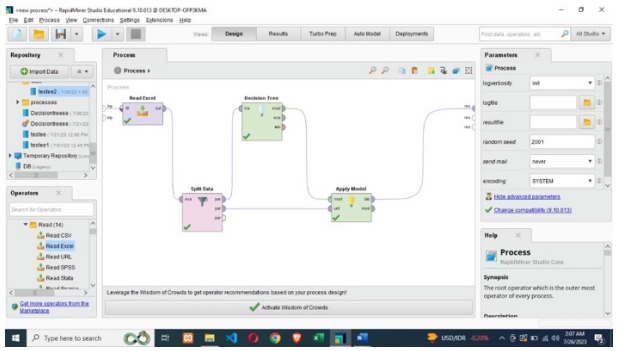

FIGURE 3. VIEW OPERATOR DECISION TREE

Before adding the Performance operator. Click run to see the prediction results. In the following image you can see the difference between "Results" based on original data and based on predictions that have been split into the data.

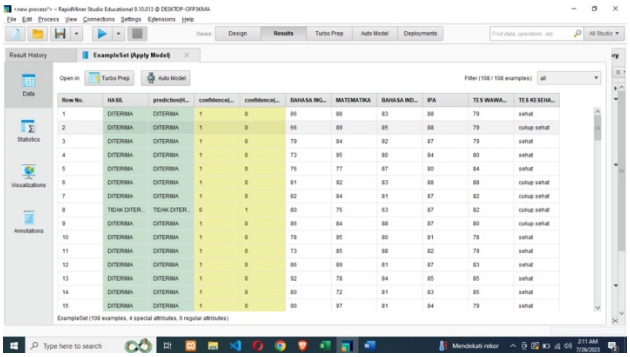

FIGURE 4. DISPLAY DATA THAT SPLIT

To test the accuracy of predictions, you need the Performance operator and select accuracy which is useful as a counter to the accuracy of the data.

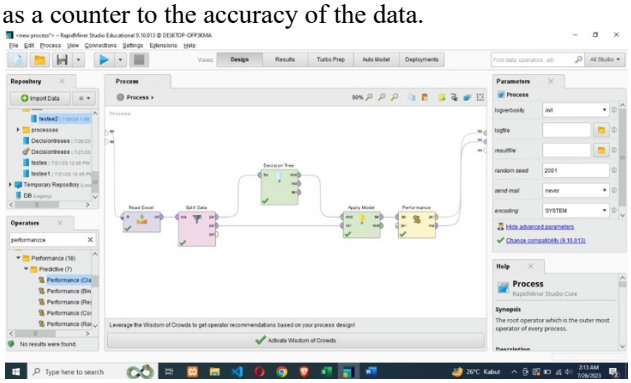

FIGURE 5. DISPLAY ACCURACY AND PERFORMANCE

Decision tree results and calculation of prediction accuracy on data.

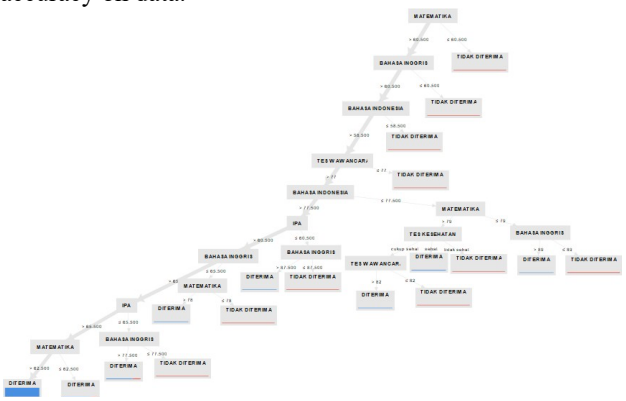

FIGURE 6. TREE RESULT DISPLAY

accuracy: 94.44%

	true DITERIMA	true TIDAK DITERIMA	class precision
pred. DITERIMA	98	4	96.08%
pred. TIDAK DITERIMA	2	4	66.67%
class recall	98.00%	50.00%	

FIGURE 7. DETAILS PARAMETER DISPLAY

4.5 Result and Evaluation

Based on the results of applying the data set using the C4.5 Algorithm in the RapidMiner application, it was found that the "Number of values" is the root of the decision tree. After calculating all the variables, Mathematics is found which is the root tree (node). Then roles are formed in the decision tree. The list of rules is as follows:

Tree

```
MATEMATIKA > 60.500
|   BAHASA INDONESIA > 58.500
|   |   TES WAWANCARA > 76
|   |   |   IPA > 60.500
|   |   |   |   BAHASA INDONESIA > 77.500
|   |   |   |   |   BAHASA INGGRIS > 65.500
|   |   |   |   |   |   IPA > 65.500: DITERIMA {DITERIMA=372, TIDAK DITERIMA=0}
|   |   |   |   |   |   IPA ≤ 65.500
|   |   |   |   |   |   |   BAHASA INGGRIS > 77.500
|   |   |   |   |   |   |   |   TES WAWANCARA > 79.500: DITERIMA {DITERIMA=9, TIDAK DITERIMA=2}
|   |   |   |   |   |   |   |   TES WAWANCARA ≤ 79.500: TIDAK DITERIMA {DITERIMA=0, TIDAK DITERIMA=1}
|   |   |   |   |   |   |   BAHASA INGGRIS ≤ 77.500: TIDAK DITERIMA {DITERIMA=0, TIDAK DITERIMA=3}
|   |   |   |   |   BAHASA INGGRIS ≤ 65.500
|   |   |   |   |   |   MATEMATIKA > 78: DITERIMA {DITERIMA=4, TIDAK DITERIMA=0}
|   |   |   |   |   |   MATEMATIKA ≤ 78: TIDAK DITERIMA {DITERIMA=0, TIDAK DITERIMA=4}
|   |   |   |   BAHASA INDONESIA ≤ 77.500
|   |   |   |   |   MATEMATIKA > 80
|   |   |   |   |   |   TES KESEHATAN = cukup sehat
|   |   |   |   |   |   |   MATEMATIKA > 84.500: DITERIMA {DITERIMA=3, TIDAK DITERIMA=0}
|   |   |   |   |   |   |   MATEMATIKA ≤ 84.500: TIDAK DITERIMA {DITERIMA=0, TIDAK DITERIMA=1}
|   |   |   |   |   |   TES KESEHATAN = sehat: DITERIMA {DITERIMA=11, TIDAK DITERIMA=0}
|   |   |   |   |   |   TES KESEHATAN = tidak sehat: TIDAK DITERIMA {DITERIMA=0, TIDAK DITERIMA=1}
|   |   |   |   |   MATEMATIKA ≤ 80
|   |   |   |   |   |   BAHASA INGGRIS > 89: DITERIMA {DITERIMA=1, TIDAK DITERIMA=0}
|   |   |   |   |   |   BAHASA INGGRIS ≤ 89: TIDAK DITERIMA {DITERIMA=0, TIDAK DITERIMA=9}
|   |   |   IPA ≤ 60.500
|   |   |   |   BAHASA INGGRIS > 89: DITERIMA {DITERIMA=2, TIDAK DITERIMA=0}
|   |   |   |   BAHASA INGGRIS ≤ 89: TIDAK DITERIMA {DITERIMA=0, TIDAK DITERIMA=4}
|   |   TES WAWANCARA ≤ 76: TIDAK DITERIMA {DITERIMA=0, TIDAK DITERIMA=1}
|   BAHASA INDONESIA ≤ 58.500: TIDAK DITERIMA {DITERIMA=0, TIDAK DITERIMA=2}
MATEMATIKA ≤ 60.500: TIDAK DITERIMA {DITERIMA=0, TIDAK DITERIMA=6}
```

FIGURE 8. TREE RESULT DISPLAY

Next, accuracy, precision and recall results are needed to test the data prediction results. Following are the test results:

a. Accuracy, the resulting predictions are TPN (Positive and Negative) Based on the overall data, "What percentage of prospective students have been predicted to be accepted or not accepted from the total?" Accuracy results have been obtained, namely 94.44%.

accuracy: 94.44%

	true DITERIMA	true TIDAK DITERIMA	class precision
pred. DITERIMA	98	4	96.08%
pred. TIDAK DITERIMA	2	4	66.67%
class recall	98.00%	50.00%	

FIGURE 9. ACCURACY RESULT

b. Precision, the resulting prediction is Positive "What percentage of prospective students were correctly not accepted out of the total predicted to be accepted and not accepted?" The precision result is 81.37%.

weighted_mean_precision: 81.37%, weights: 1, 1

	true DITERIMA	true TIDAK DITERIMA	class precision
pred. DITERIMA	98	4	96.08%
pred. TIDAK DITERIMA	2	4	66.67%
class recall	98.00%	50.00%	

FIGURE 10. ACCURACY RESULT

c. Recall, the resulting prediction is Positive "What percentage of prospective students are predicted to be accepted and not accepted compared to the total number of prospective students who are not accepted?" The recall result is 74.00%.

weighted_mean_recall: 74.00%, weights: 1, 1

	true DITERIMA	true TIDAK DITERIMA	class precision
pred. DITERIMA	98	4	96.08%
pred. TIDAK DITERIMA	2	4	66.67%
class recall	98.00%	50.00%	

FIGURE 11. ACCURACY RESULT

Next, you can see the data display because of the comparison between the prediction results and the data set below.

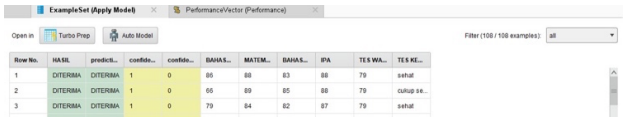

FIGURE 12. OVERALL RESULTS OF SPLIT DATA

The following are prediction results that are in accordance or correct with the data set.

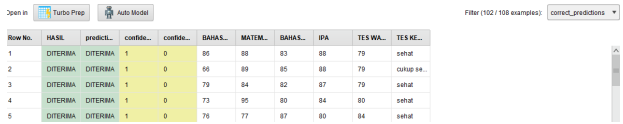

FIGURE 13. APPROPRIATE DATA RESULTS

The following are prediction results that do not match or are wrong with the data set.

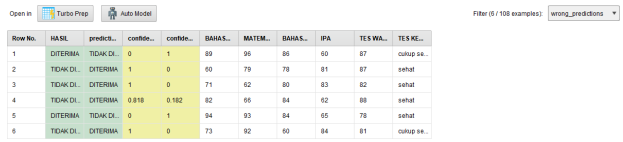

FIGURE 14. INAPPROPRIATE DATA RESULTS

4.5.1 Decision Tree

The following is the result of a decision tree using Python with the first node having a gain value of 0.132, a samples value of 380 and a velocity value of 353.27. This analysis sets the stage for further exploration, refinement, and interpretation, empowering data scientists to harness the power of decision trees for predictive modeling and actionable insights.

The results of the research that has been implemented and described in applying the C4.5 Algorithm to determine the prediction of majoring patterns at the Jakarta Laboratory Vocational School, the researcher can conclude that by applying the RapidMiner and Python applications, the RapidMiner Software produces an accuracy of 94.44%, with a precision of 81.37% and a recall of 74.00% and Python has an accuracy of 93% because Python automatically rounds the values so there is no significant difference between the two tools, so this proves that the C4.5 Algorithm produces quite accurate performance.

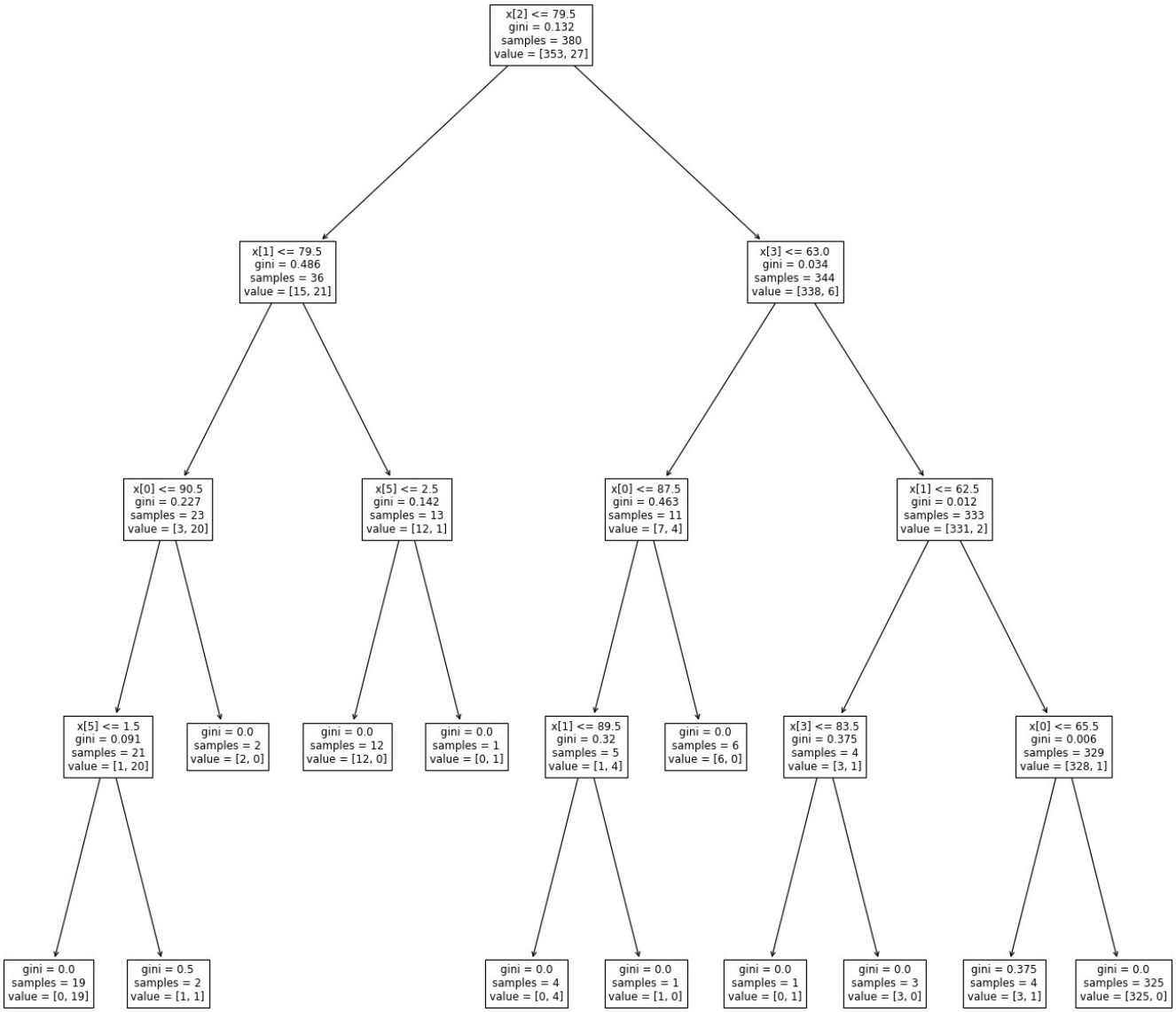

FIGURE 15. DECISION TREE USING PHYTON

5. CONCLUSIONS

Based on the results of the research that has been carried out, it can be concluded as follows. Using Rapidminer software and using the k-means algorithm with sales data for 11 months with calculations carried out to produce 5 clusters. Based on the comparison results of 3 K-means algorithms with different K values, namely 3, 4, 5, the result from Davies Bouldin with a value close to 0 is the value with K 5, with the result from Davies Bouldin being -0.912.

REFERENCES

[1] A. Zuhdi, F. Firman, and R. Ahmad, "The importance of education for humans," *SCHOULID: Indonesian Journal of School Counseling*, vol. 6, no. 1, p. 22, Feb. 2021. DOI: 10.23916/08742011

[2] A. Haleem, M. Javaid, M. A. Qadri, and R. Suman, "Understanding the role of digital technologies in education: A review," *Sustainable Operaationsa and Computers*, vol. 3, pp. 275–285, 2022. DOI: 10.1016/j.susoc.2022.05.004

[3] R. Cheng, J. Gao, and H. Zhang, "Identify the Reasons of Students' Choices for Majors and Courses," in *Proceedings of the 2022 International Conference on Science Education and Art Appreciation (SEAA 2022)*, Paris: Atlantis Press SARL, 2023, pp. 156–166. DOI: 10.2991/978-2-494069-05-3_20

[4] C. Coman, L. G. Țîru, L. Meseșan-Schmitz, C. Stanciu, and M. C. Bularca, "Online Teaching and Learning in Higher Education during the Coronavirus Pandemic: Students' Perspective," *Sustainability*, vol. 12, no. 24, p. 10367, Dec. 2020. DOI: 10.3390/su122410367

[5] M. Tahir and C. Anwar Korompot, "The Impact Of Self-Confidence On Students Public Speaking Ability," 2023.

[6] B. Çığşar and D. Ünal, "Comparison of Data Mining Classification Algorithms Determining the Default Risk," *Scientific Programming*, vol. 2019, pp. 1–8, Feb. 2019. DOI: 10.1155/2019/8706505

[7] G. A. Putri, D. Maryono, and F. Liantoni, "Implementation of the C4.5 Algorithm to Predict Student Achievement at SMK Negeri 6 Surakarta," *IJIE (Indonesian Journal of Informatics Education)*, vol. 4, no. 2, p. 51, Dec. 2020. DOI: 10.20961/ijie.v4i2.47124

[8] D. Putra and A. Wibowo, "Prediksi Keputusan Minat Penjurusan Siswa SMA Yadika 5 Menggunakan Algoritma Naïve Bayes," *Prosiding Seminar Nasional Riset Dan Information Science (SENARIS)*, vol. 2, pp. 84–92, 2020.

[9] M. Awaludin, V. Yasin, and M. Wahyuningsih, "Optimization of Naïve Bayes Algorithm Parameters for Student Graduation Prediction at Universitas Dirgantara Marsekal Suryadarma," *Journal of Information System, Informatics and Computing Issue Period*, vol. 6, no. 1, pp. 91–106, 2022. DOI: 10.52362/jisicom.v6i1.785

[10] A. R. Kadafi, "Perbandingan Algoritma Klasifikasi Untuk Penjurusan Siswa SMA," *Jurnal ELTIKOM*, vol. 2, no. 2, pp. 67–77, Dec. 2018. DOI: 10.31961/eltikom.v2i2.86

[11] M. Yudhi Putra and D. Ismiyana Putri, "Pemanfaatan Algoritma Naïve Bayes dan K-Nearest Neighbor Untuk Klasifikasi Jurusan Siswa Kelas XI," *Jurnal Tekno Kompak*, vol. 16, no. 2, pp. 176–187, 2022.

[12] Q. A. A'yuniyah and M. Reza, "Penerapan Algoritma K-Nearest Neighbor Untuk Klasifikasi Jurusan Siswa Di Sma Negeri 15 Pekanbaru," *Indonesian Journal of Informatic Research and Software Engineering (IJIRSE)*, vol. 3, no. 1, pp. 39–45, Mar. 2023. DOI: 10.57152/ijirse.v3i1.484

[13] A. Y. Kuntoro, H. Hermanto, T. Asra, F. Syukmana, and H. Wahono, "Classification of Student Majors with C4.5 and Naive Bayes Algorithms (Case Study: SMAN 2 Bekasi City)," *Semesta Teknika*, vol. 23, no. 1, 2020. DOI: 10.18196/st.231251

AUTHORS

Nurisya Rahma Suherman

Graduated from the Department of Informatics, Faculty of Engineering, Perjuangan University of Tasikmalaya, Indonesia.

Ruuhwan, M.Kom.

Lecturer in Department of Informatics, Faculty of Engineering, Perjuangan University of Tasikmalaya, Indonesia.

Aso Sudiarjo, M.Kom.

Lecturer in Department of Informatics, Faculty of Engineering, Perjuangan University of Tasikmalaya, Indonesia.