INTERNATIONAL JOURNAL OF SOCIAL HEALTH
https://ijsh.ph/index.php/rp
Volume 2 Number 2 February 2023 E-ISSN: 2984-7079

ALGORITHM C4.5 APPLICATION OF INTEREST AND TALENT DATA MINING AT SMK NEGRI 1 BONGAS

Anissa Ocktoviani, Arif Rinaldi Dikananda
Sekolah Tinggi Manajemen Informatika dan Komputer IKMI Cirebon, Indonesia
anissaocktoviani02@gmail.com

KEYWORDS

data mining, decision tree, interests and talents, method c4.

ABSTRACT

The purpose of this study is to classify student majors to simplify and speed up the determination of the selection of majors so that the process resulting from this selection is more accurate and objective. The design method that will be applied to data mining to determine interests and talents is the C4.5 Decision Tree Algorithm Method. The data used in this research are 331 datasets. The data is classified using the C4.5 Decision Tree Algorithm method. The research results show that the classification of interests and talents can be classified to determine student majors using the C4.5 Algorithm. 2) Of the 331 data divided by 80% training data and 20% testing data, the level of accuracy in the C4.5 Decision Tree Algorithm is 89.39%. So, it can be concluded that the accuracy results are lower than previous studies, which produced an accuracy rate of 100%. This means that by analyzing student interest and aptitude data using C4.5, schools can build a classification model that can assist students in choosing a major that suits their interests and talents thereby increasing the accuracy and efficiency of major selection

DOI: 10.58860/ijsh.v2i2.26

Corresponding Author: Anissa Octoviani
E-mail: anissaocktoviani02@gmail.com

INTRODUCTION

Interest is an ongoing attitude that can make a person's attention pattern so that he becomes selective toward the object of his interest. Talent is an innate ability that is innate to nature and intelligence (Merawati & Rino, 2019). Ignorance of students' abilities, talents, or interests will impact the development of potential and even careers for the future (Narulita et al., 2021). The SMK majors have four majors for students, namely Computer and Network Engineering (TKJ), Motorcycle Engineering and Business (TBSM), Automotive Light Vehicle Engineering (TKRO), and Hospitality (PH). The main factors used to determine interest in talent in choosing majors include scores from PAI, PPKN, Mathematics, Indonesian, Science, Social Studies, English, student interests, and scores (Kuniasari & Fatmawati, nd)

Implementing interest and aptitude tests carried out by a person forms individual data. The data from the test will continue to grow over time and will only be used as an archive or just a report by the administering agency. From these data sets, new information can be extracted through data mining which can be useful for certain institutions. By digging into this data set, new hidden knowledge can be obtained, which has benefits (Irawan, 2019). Based on this data, determine the level of accuracy of students' interests and talents using the C4.5 Decision Tree Algorithm Method. Algorithm C4.5 is a decision tree classification algorithm that can produce decision trees that are easy to interpret, have an acceptable level of accuracy, and can handle discrete and numeric type attributes (Dharshinni, 2021).

Based on research results (Swastina, 2013), the Decision Tree Algorithm C4.5 is more accurate than the Naïve Bayes Algorithm in determining the majors' suitability and recommending student

majors. Thus, the Decision Tree Algorithm C4.5 is accurately applied to determine student suitability with an accuracy rate of 93.31% and an accuracy of major recommendations of 82.84% (Rahayu, 2014).

Classification is an algorithm in data mining that groups data into certain criteria or categories by reading previously existing data (Wanto et al., 2020). The final classification task used is the Decision Tree, and the Algorithm used is Method C4.5. A Decision Tree is a machine learning algorithm that uses a set of rules to make decisions with a tree-like structure that models possible outcomes, resource costs, utilities, and possible consequences or risks (Widiastuti et al., 2023).

Table 1. Interest and Talent Data Sample

Name	Option 1	Option 2	IPA	IPS	PPKN
Galang Ardiansyah	TKRO	TKJ	82	79	78
Strong Maulana	TKRO	TKJ	80	80	80
Sigit Arimukti	TKJ	TKRO	77	71	82
Muhammad Dayu Aji	TKRO	TKJ	72	76	78
Augustine's Daughter	TKJ	TKRO	77	72	80

Based on Table 1, the sample data comes from the talent interest dataset, where each data consists of the attributes Name, Choices 1 & 2, Science Value, IPS, and PPKN. The attribute name states the student's name, the attribute choices 1 and 2 state the student's choice of majors, and the attribute value of Science, Social Sciences, PPKN states the student's report card value.

There is another problem that arises is that not everyone knows their talent. Finding out the potential of someone's talent is not easy because there is a fundamental difference between talent and interest. Interest is often interpreted as talent, even though interest is not talent. Interest is strongly influenced by the environment around family, friends, and society, while talent is a person's ability acquired from birth (Syamsu et al., 2019).

The purpose of this final project is to classify student majors to simplify and speed up the determination of the selection of majors so that the process resulting from this selection is more accurate and objective. The data used in this research are 331 datasets. The data is classified using the C4.5 Decision Tree Algorithm method. This final project is supported by the Knowledge Discovery In Databases (KDD) method, a non-trivial process for finding and identifying patterns (patterns) in data, where the patterns found are valid, new, useful, and understandable. So that with this final assignment project, it can help schools find out the accuracy of students in interest talent data and decision tree results from that data.

Based on the background above, the purpose of this study is to find out and analyze the C4.5 algorithm for the application of data mining interests and talents at SMK negri 1 Bongas. The application of C4.5 in the analysis of interests and talents in SMK can provide several benefits, including: Increasing the effectiveness of major selection. By analyzing student interest and aptitude data using C4.5, schools can build a classification model that can assist students in choosing a major that suits their interests and talents. That way, students have a better chance of pursuing a career that matches their interests and talents. And improve the quality of teaching. By understanding students' interests and talents, teachers can choose teaching methods that are more effective and can help students learn better. This can improve the quality of teaching and help students to better understand the lesson.

METHODS

The primary data source in this final assignment was obtained from the administration teacher. This data is taken from the school directly during the internship. This data is secondary data that is recapitulated from the school. This dataset contains history, organizational structure, data on teachers, students, and school staff, as well as school facilities and infrastructure.

Data obtained from SMK Negeri 1 Bongas Jl. Raya Margamulya NO. 276 B, Kec. Bongas, Indramayu Regency, West Java 45255. On January 4, data were collected at the school. The data collection technique was used directly through the school's computer, transferring files, and storing the data on a flash drive. The document was obtained directly from the school through the Administrative Teacher.

Data analysis techniques use Decision Trees or decision trees and C4.5 Algorithms. Decision Tree or decision tree is a classification method that uses a tree structure, where each node represents an attribute, and its branches represent attribute values. In contrast, the leaves are used to represent classes. The tree is used as a reasoning procedure to get answers to the problems entered. The tree formed is not always a binary tree. If all the features in the set use two kinds of categorical values, the resulting tree will be a binary tree. If the feature contains more than two kinds of categorical values or uses a numeric type, the resulting tree is usually not a binary tree. Each internal node represents a variable, and the leaf nodes represent a class (Pasaribu, 2021) . Flexibility makes this Method attractive, especially because it provides the advantage of visualizing suggestions (Decision Tree form suggestions) that make the prediction procedure observable (Prasetyo et al., 2013) . Several researchers have presented that the decision tree method is widely used to solve decision-making cases, such as in the fields of medicine (diagnosis of patient disease), computer science (data structures), psychology (decision-making theory), and so on.

Algorithm C4.5 is a classification method that involves constructing decision trees (Yulia et al., 2022). Where each branch then leads to another node or final decision Algorithm, C4.5 is the Algorithm used to construct a decision tree (Prasetio et al., 2021). Decision trees are a familiar method of classification and prediction. Decision trees are useful for exploring data and finding hidden relationships between several candidate input variables and a target variable. Many algorithms can form decision trees, including ID3, CART, and C4.5 (Achmad et al., 2012). Algorithm C4.5 is the development of the ID3 Algorithm. The process in the decision tree is to change the form of data in tabular form to a tree model, to change the tree model to a rule, and to simplify the rule.

RESULTS AND DISCUSSION

Application of the Decision Tree Algorithm Model C4.5

The application of the C4.5 Decision Tree Algorithm Model was designed using Rapidminer tools. In this final project, several operators are used, including the following:

1. Retrieve Data

   This operator loads datasets imported into the local repository into the process. Using the Retrieve Data operator, a recap dataset of talent interest students who are accepted and not accepted in choosing a major will be loaded into the process column and produce an Example set of 331 data.

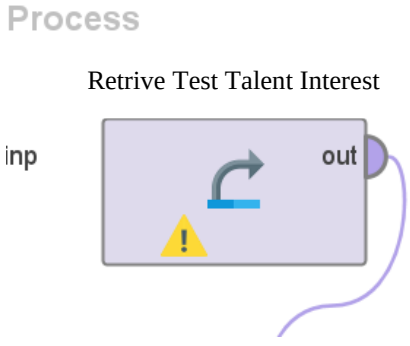
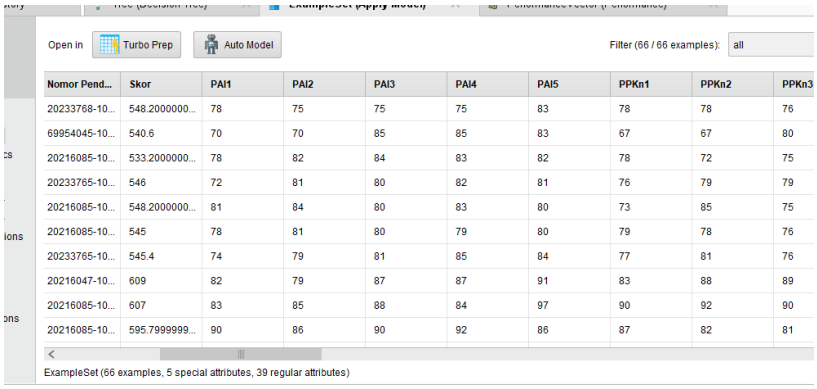


Figure 1. Data Retrieve Operators



Open in Turbo Prep Auto Model Filter (66 / 66 examples): all

Nomor Pend...	Skor	PAI1	PAI2	PAI3	PAI4	PAI5	PPKn1	PPKn2	PPKn3
20233768-10...	548.2000000...	78	75	75	75	83	78	78	76
69954045-10...	540.6	70	70	85	85	83	67	67	80
20216085-10...	533.2000000...	78	82	84	83	82	78	72	75
20233765-10...	546	72	81	80	82	81	76	79	79
20216085-10...	548.2000000...	81	84	80	83	80	73	85	75
20216085-10...	545	78	81	80	79	80	79	78	76
20233765-10...	545.4	74	79	81	85	84	77	81	76
20216047-10...	609	82	79	87	87	91	83	88	89
20216085-10...	607	83	85	88	84	97	90	92	90
20216085-10...	595.7999999...	90	86	90	92	86	87	82	81

ExampleSet (66 examples, 5 special attributes, 39 regular attributes)

Figure 2. Data Retrieve Results

2. Set Roles

   Set Role distinguishes coordinate attributes and position predictions that will be included in the 'label ' category. So that when categorizing data, 'labels' are not included in the calculation, and change the results with Status Attribute as Label and NISN as ID.

NISN	Status
0079593287	Tidak Diterima
0073553042	Tidak Diterima
0073608924	Tidak Diterima
0075089646	Tidak Diterima
0076453203	Tidak Diterima
0076032195	Tidak Diterima
0074064192	Tidak Diterima
0064524642	Diterima
0139636068	Diterima
0077799672	Diterima

Figure 3. Role Set Results

3. Select Attributes

   Select Attribute is selecting a dataset that will be selected and not selected in the final project rapid miner testing the C4.5 Decision Tree Algorithm.

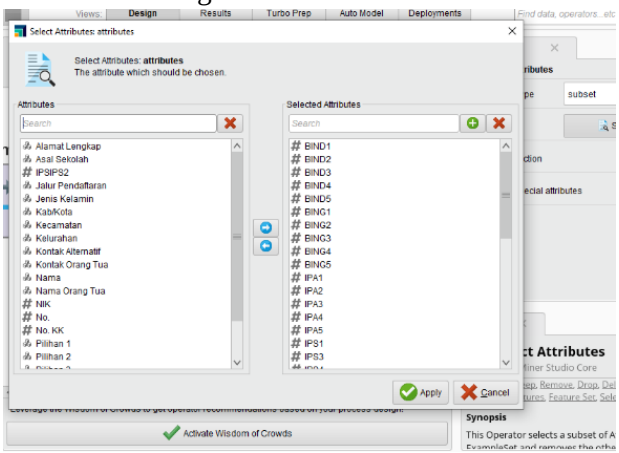


Figure 4. Select Attribute results

4. Split Data

   The distribution ratio for this project is 80% for training data and 20% for testing data, with a total of 331 datasets. There are 265 Training Data and 66 Testing Data.

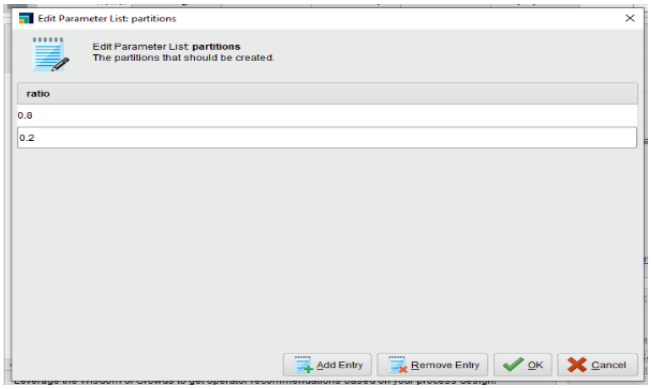


Figure 5. Split Data Process

Table 2. Split Data Results

Ratio	Data Type	Amount of data
80%	Training Data	265
20%	Data Testing	66

5. DecisionTree

   A decision tree is a decision-making method that organizes options into a branching form.

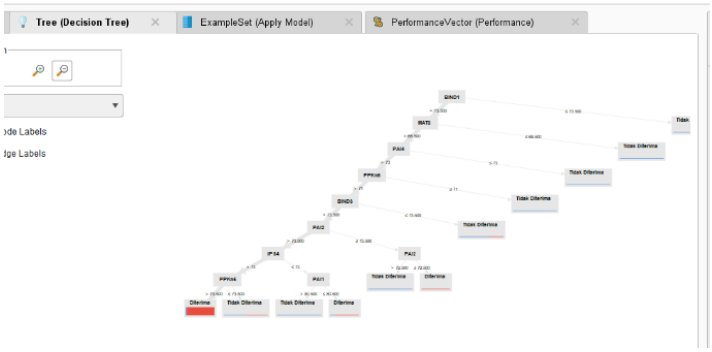


Figure 6. Results of the Apply Model

Results of the Accuracy Level of the Algorithm Method C4.5

performance

The performance operator evaluates a performance model that automatically provides a list of performance modification values according to a given task. The level of accuracy in performance is 89.39% which is a better result than the previous level of accuracy.

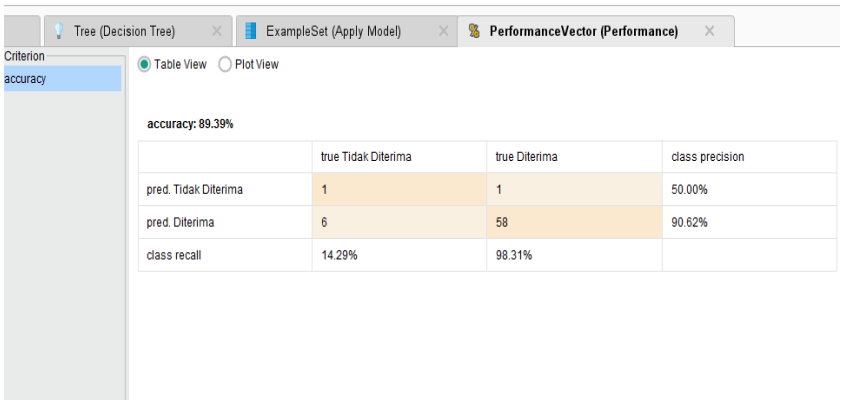


Figure 7. Accuracy Results

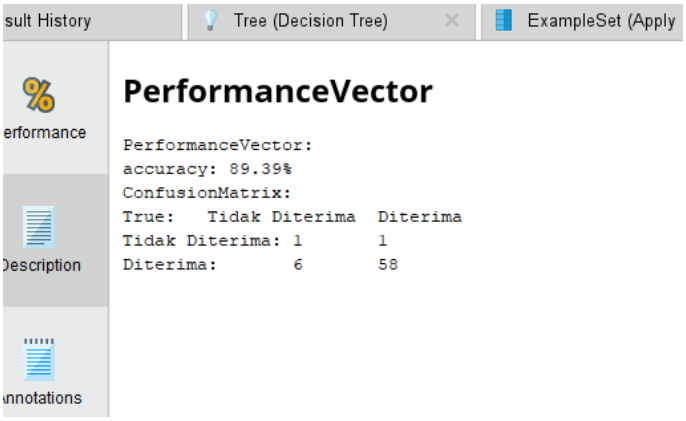


Figure 8. Performance Description

CONCLUSION

Based on the final project regarding students' interests and talents, several conclusions can be drawn: 1) Classification of interests and talents can be classified to determine student majors using the C4.5 Algorithm 2) Of the 331 data divided by 80% training data and 20% testing data, it shows the level the accuracy of the C4.5 Decision Tree Algorithm is 89.39%. The accuracy results are lower than previous studies, which produced an accuracy rate of 100%. This is because the data used is different from before.

REFERENCES

Achmad, B. D. M., Slamat, F., & ITATS, F. T. I. (2012). Klasifikasi data karyawan untuk menentukan jadwal kerja menggunakan metode decision tree. *Jurnal Iptek*, *16* (1).

Dharshinni, N. P. (2021). Classification Of Major Selection Based On Student's Expertise Using C4. 5 Algorithm. *INFOKUM*, *9* (2, June), 412–418.

Irawan, Y. (2019). Penerapan Data Mining Untuk Evaluasi Data Penjualan Menggunakan Metode Clustering Dan Algoritma Hirarki Divisive Di Perusahaan Media World Pekanbaru. *Jurnal Teknologi Informasi Universitas Lambung Mangkurat (JTIULM)*, *4* (1), 13–20.

Kuniasari, R., & Fatmawati, A. (n.d.). *Penerapan Data Mining dengan Algoritma C4. 5 untuk Penentuan Jurusan Sekolah Menengah Atas Implementation of Data Mining with C4. 5 Algorithm for Determining Senior High School*.

Merawati, D., & Rino, R. (2019). Penerapan Data Mining Penentu Minat Dan Bakat Siswa Smk Dengan Metode C4. 5. *ALGOR*, *1*(1), 28–37.

Narulita, S., Oktaga, A. T., & Susanti, I. (2021). Pengujian Akurasi Model Prediksi Menggunakan Metode Data Mining Classification Decision Tree Algoritma C4. 5 untuk Penentuan Peminatan Peserta Didik. *Media Aplikom*, *13*(2), 65–79.

Pasaribu, A. F. O. (2021). Analisis Pola Menggunakan Metode C4. 5 Untuk Peminatan Jurusan Siswa Berdasarkan Kurikulum (Studi Kasus: Sman 1 Natar). *Jurnal Teknologi Dan Sistem Informasi*, *2*(1), 80–85.

Prasetio, A., Hasibuan, M. H., & Sitompul, P. (2021). Simulasi Penerapan Metode Decision Tree (C4. 5) Pada Penentuan Status Gizi Balita. *Jurnal Nasional Komputasi Dan Teknologi Informasi*, *4*(3).

Prasetyo, E., Rahajoe, R. A. D., Agustin, S., & Arizal, A. (2013). Uji Kinerja Dan Analisis K-Support Vector Nearest Neighbor Terhadap Decision Tree dan Naive Bayes. *Jurnal Eksplora Informatika*, *3*(1), 1–6.

Rahayu, E. B. (2014). Algoritma C4. 5 Untuk Penjurusan Siswa SMA NEGERI 3 PATI. *Progr. Stud. Tek. Inform. Fak. Ilmu Komput*, 3–6.

Swastina, L. (2013). *Penerapan Algoritma C4. 5 Untuk Penentuan Jurusan Mahasiswa*.

Syamsu, S., Muhajirin, M., & Wijaya, N. S. (2019). Rules Generation Untuk Klasifikasi Data Bakat dan Minat Berdasarkan Rumpun Ilmu Dengan Decision Tree. *Inspiration: Jurnal Teknologi Informasi Dan Komunikasi*, *9*(1), 40–51.

Wanto, A., Siregar, M. N. H., Windarto, A. P., Hartama, D., Ginantra, N. L. W. S. R., Napitupulu, D., Negara, E. S., Lubis, M. R., Dewi, S. V., & Prianto, C. (2020). *Data Mining: Algoritma dan Implementasi*. Yayasan kita menulis.

Widiastuti, N., Hermawan, A., & Avianto, D. (2023). Komparasi Algoritma Klasifikasi Datamining Untuk Prediksi Minat Pencari Kerja. *Jurnal Teknoinfo*, *17*(1), 219–227.

Yulia, D., Kusuma, A. P., & Permadi, D. F. H. (2022). Penerapan Algoritma C4. 5 Untuk Prediksi Minat Penjurusan Siswa Di Smkn 1 Kademangan. *JATI (Jurnal Mahasiswa Teknik Informatika)*, *6*(2), 893–900.