JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. VIII No 3, Agustus 2022, hlm. 295 - 304 ISSN 2550-0201 (Online)
DOI: https://doi.org/10.33330/jurteksi.v8i3.1680
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

STUDENTS GRADUATION PREDICTION BASED ON ACADEMIC DATA RECORD USING THE DECISION TREE ALGORITHM C4.5 METHOD

Narita Ayu Prahastiwi[1*], Rachmadita Andreswari[1], Rokhman Fauzi[1]
[1]Sistem Informasi, Telkom University
email: *naritaayup@student.telkomuniversity.ac.id

Abstract: An application can assist organizations in achieving the goals to be achieved by facilitating ongoing work processes. This happened in the Information Systems Study Program at one of the best private universities, namely Telkom University, where the SI Study Program has a website called PIPE and has one feature to be able to predict student graduation. However, this feature is currently being developed with an easy flow, so it requires development in the implementation of graduation achievements. Researchers solve these problems by building an assessment model based on academic data on the effect of choosing a specialization. Data mining is needed in this study to form predictive patterns, then one of the data mining groups is based on classification and using machine learning to perform automated assessments so that they can be sustainably performed. In determining the time and delay, using the decision tree method based on the C4.5 algorithm. The accuracy results obtained using the C4.5 algorithm are 94.11%, then the factor that becomes the root node is Jumlah SKS Lulus and the results have an influence on the selection of specialization. So that the results of this graduation model can be applied to the PIPE application.

Keyword: C4.5 algorithm; classification; decision tree; graduation prediction

Abstrak: Sebuah aplikasi dapat membantu organisasi dalam mencapai tujuan yang ingin dicapai dengan memudahkan proses kerja yang sedang berlangsung. Seperti yang terjadi pada Prodi Sistem Informasi yang ada pada salah satu Perguruan Tinggi Swasta terbaik yaitu Universitas Telkom, dimana pada Prodi SI memiliki website bernama PIPE dan memiliki salah satu fitur untuk dapat melakukan prediksi kelulusan mahasiswa. Namun fitur tersebut saat ini dikembangkan dengan alur penentuan sederhana, sehingga memerlukan pengembangan dalam hal implementasi algoritma prediksi kelulusan. Peneliti melakukan penyelesaian masalah tersebut dengan membangun model prediksi kelulusan berdasarkan rekam data akademik terhadap pengaruh pemilihan peminatan. *Data mining* dibutuhkan dalam penelitian ini untuk membentuk pola penyelesaian prediksi, kemudian salah satu pengelompokan *data mining* berdasarkan tugasnya adalah klasifikasi dan menggunakan *machine learning* untuk melakukan prediksi kelulusan secara otomatis terhadap data baru agar dapat dilakukan secara berkelanjutan. Dalam melakukan klasifikasi prediksi kelulusan tepat waktu dan terlambat, menggunakan metode *decision tree* berdasarkan algoritma C4.5. Hasil akurasi yang didapat dengan menggunakan algoritma C4.5 adalah sebesar 94,11%, kemudian faktor yang menjadi *root node* adalah Jumlah SKS Lulus dan hasil memiliki pengaruh terhadap pemilihan peminatan. Sehingga hasil model prediksi kelulusan ini dapat diterapkan pada aplikasi PIPE.

Kata kunci: algoritma C4.5; decision tree; klasifikasi; prediksi kelulusan.

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. VIII No 3, Agustus 2022, hlm. 295 - 304 ISSN 2550-0201 (Online)
DOI: https://doi.org/10.33330/jurteksi.v8i3.1680
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

INTRODUCTION

Telkom University is the best Private Universities (PTS) No. 1 from the Ministry of Education and Culture of the Republic of Indonesia for two consecutive years in 2019 and 2020 [1]. Each new academic year the SI Study Program will accept new students, where students must complete the IS scientific level based on the curriculum that has been set for eight semesters or four years of normal study, seven semesters of three and a half years of study and a maximum of fourteen semesters. From 2016-2020 the SI Study Program succeeded in achieving the graduation target that had been set based on the FRI baseline target, it is hoped that the SI Study Program can continue to maintain the graduation target, by continuously monitoring the progress of SI Study Program students based on academic data records. The SI study program has a website called the Select Specialization (PIPE) application, in the PIPE application there is a feature to predict student graduation, but this feature is currently being developed with a simple determination flow, so it requires improvements in the implementation of the graduation prediction algorithm, and has other features as well for the selection of specializations from the seven specializations in the SI Study Program. Therefore, a prediction model for student graduation is needed with a path of determination and a supporting theoretical basis so that it can be applied to the PIPE application.

To achieve this goal in predicting student graduation, optimal data utilization is needed on large amounts of data and the data collection is stored in the i-Gracias system of Telkom University. Utilizing the data, one of which is by using Knowledge Discovery in Database (KDD) [2]. KDD is the process of extracting and exploring large amounts of previously unknown data to obtain knowledge and meaningful data by using certain techniques or methods [3], one of the stages in the KDD process is data mining. Data mining is used as a process of finding patterns from large amounts of data, coming from data sources such as databases [4]. One of the data mining groupings based on the task is classification and data mining inherits many aspects and techniques from the scientific field, one of which is machine learning [5]. Classification is used to carry out the training process for target functions that group each attribute into available label classes [6]. While machine learning is used as machine learning on new data based on the training process to classify objects based on certain characteristics [7].

Well-known algorithms for predicting on-time graduation rates are ID3, CART, Naïve Bayes, fuzzy AHP (FAHP) and C4.5 algorithms [8]. The C4.5 algorithm has the highest performance value and level of accuracy compared to ID3 and CART in classifying [9] and against Naïve Bayes the C4.5 algorithm also has higher accuracy in determining the time of graduation on time with quite a lot of variables [10]. Therefore, the C4.5 algorithm can help determine graduation predictions on time and is also one of the algorithms used to form a decision tree [11]. Decision Tree is a data mining technique in the form of a decision tree that is used to predict object membership based on different classes or labels [6]. In determining student graduation predictions, using class labels based on academic data records, including having an influence on the selection of student

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. VIII No 3, Agustus 2022, hlm. 295 - 304 ISSN 2550-0201 (Online)
DOI: https://doi.org/10.33330/jurteksi.v8i3.1680
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

specialization by predicting student graduation. The dataset used comes from data from students of the SI Study Program at Telkom University in 2016-2017. This study builds a predictive model by applying several methods and theoretical foundations in order to produce output in accordance with the goals and needs.

METHOD

The systematic research uses several stages, namely, the preparation stage, data processing and evaluation. The stages of the research are shown in Figure 1, and then explain each of the stages carried out.

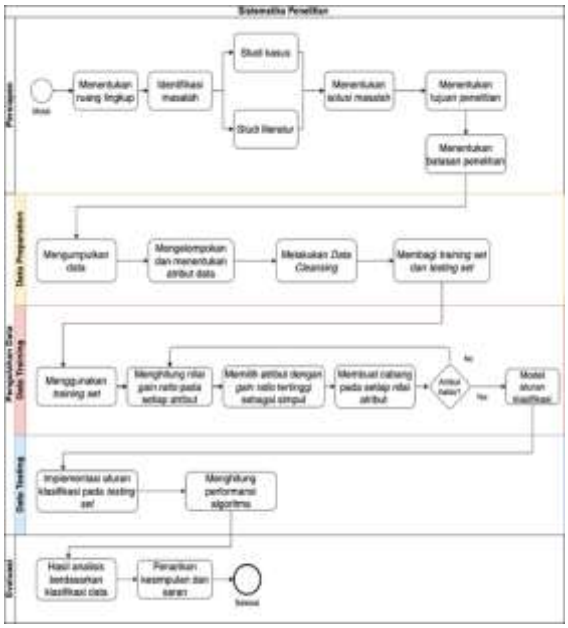

Figure 1. The Systematic Research

The first stage is the preparation stage, starting with determining the scope of the problems that exist in the SI study program at Telkom University. Then identify the problem based on the information that has been collected, then the process of determining the problem is determined by two things, namely through existing case studies and literature studies by looking at existing references or best practices. The literature study is also used as a basis for reference in providing appropriate problem solutions. The literature study that was collected was about data classification, decision trees and the application of the C4.5 Algorithm for predicting student academic performance. Processing and generating input data from the extraction of large amounts of data using KDD.

Knowledge Discovery in Database (KDD) is the process of identifying new patterns in valid, potentially useful data and ultimately the patterns can be understood in the data. [12]. The process contained in the KDD is iterative and interactive from the steps that can belustrated as shown in Figure 2 [12].

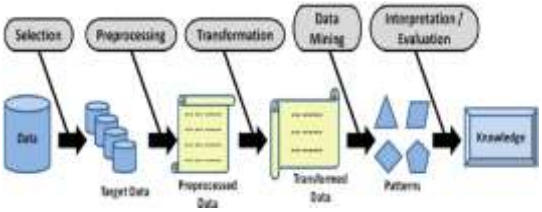

Figure 2. Process of KDD [12]

1. Selection
   Data selection selection will be used from a set of operational data and aims to create a target set from the original data or select attributes or data samples.
2. Pre-processing
   Pre-processing or cleaning aims to clean data, remove duplicated data, determine the right strategy for missing data fields and correct data errors, such as typographical errors.
3. Transformation
   The transformation process in charge of reducing and projecting data.

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. VIII No 3, Agustus 2022, hlm. 295 - 304 ISSN 2550-0201 (Online)
DOI: https://doi.org/10.33330/jurteksi.v8i3.1680
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

4. Data Mining
   The process of finding patterns that attract large amounts of data using the method.
5. Interpretation / Evaluation
   The aim is that the information or patterns that have been obtained can be visualized and easily understood by interested parties and can be seen again and again from the previous process steps.

Data mining is also one of the steps in the KDD process, with data mining being able to determine and perform classification patterns, predict, estimate and obtain information using certain techniques or methods, where the data sources can be obtained from databases, data warehouses, or data streamed. dynamically [7]. One of the scientific fields in data mining is machine learning. Machine Learning (ML) can be defined as a computer application that is adopted by means of self-learning based on existing training data to produce predictions in the future [13]. Therefore, the larning machine can predict graduation automatically for new data so that it can be used sustainably performed.

The second step is the data processing stage, in this process it is divided into three phases, namely the data preparation phase, training data and data testing. Classification is a grouping of data mining used based on the task and is also one of the main applications in ML. Classification has two processes, namely the process of building a model based on the existing training data (training set) and the classification process using the model to classify the new data (testing set) using certain techniques or methods. Classification can be defined as the process of training the target function and grouping each attribute into the available label classes [6]. The classification process is based on four components [6].

1. Class
   It is a categorical dependent attribute that represents a "label" on an object.
2. Predictor
   Is an independent attribute represented by the characteristics of data attributes.
3. Training Set
   Is a dataset or data set that contains the values of the two components above which are used to determine the suitable class based on the predictor.
4. Testing Set
   Contains new data to be classified by the model that has been created and evaluates classification accuracy.

One method of classification algorithm is Decision Tree (DT). DT is a data mining technique in the form of a decision tree that is used to predict object membership based on different classes, taking into account the values according to their attributes [6]. Decision Tree also a flow chart with the concept of a tree structure, which consists of nodes as attribute tests, each branch as a result of the test, and each leaf node as a class or decision [6]. Using the C4.5 Algorithm in determining the results of the DT classification in the prediction model of student graduation. Algorithm C4.5 or Classification version 4.5 is one of the algorithms to form a decision tree based on training data, can be used in classification methods, class predictions and seen based on gain ratio [14]. The C4.5 algorithm is the development of the ID3 algorithm, the ID3 algorithm is the most basic decision tree learning algorithm developed by J. Ross Quinlan where the ID3 algorithm can be

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. VIII No 3, Agustus 2022, hlm. 295 - 304 ISSN 2550-0201 (Online)
DOI: https://doi.org/10.33330/jurteksi.v8i3.1680
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

implemented using a recursive function [14].

The C4.5 algorithm is one of the supervised learning-based algorithms which is also a development of the ID3 algorithm, using the C4.5 algorithm, DT can be built from a set of training data with information entropy, this is included in the statistical classifier [14]. The following are the stages of the C4.5 . algorithm process [15]:

1. Prepare training data. Training data is usually obtained from historical data that has happened before and has been grouped into certain classes.
2. Counting and determining the root of the tree (root). The root will be selected by calculating the gain value of each attribute, the highest gain ratio value will be chosen as the first root, before getting the gain ratio requires the results of entropy, gain and also split info in accordance with equations (1) to (4) [15]:

$$Entropy\ (\mathrm{S}) = \sum_{i=1}^{n} -pi * \log 2\ pi \quad (1)$$

Description :
S : case set
n : number of partitions S
pi : the proportion of Si to S

$$Gain(\mathrm{S},\mathrm{A}) = Entropy(\mathrm{S}) - \sum_{i=1}^{n} \frac{|Si|}{|S|} * Entropy(\mathrm{Si}) \quad (2)$$

Description :
S : case set
A : attribute
n : number of attribute partitions A
|Si| : number of cases on partition i
|S| : number of cases in S

$$Gain\ Ratio(\mathrm{S},\mathrm{A}) = \frac{Gain(\mathrm{S},\mathrm{A})}{SplitInfo(\mathrm{S},\mathrm{A})} \quad (3)$$

Description :
S : sample space (data) used for training
A : attribute
Gain(S,A) : information gain atribut A
SplitInfo(S,A) : split information atribut A

$$SplitInfo(\mathrm{S},\mathrm{A}) = -\sum_{i=1}^{n} \frac{Si}{S} \log 2 \frac{Si}{S} \quad (4)$$

Description :
S : sample space (data) used for training
A : attribute
Si : number of samples for attribute i

3. Create a branch for each record.
4. Divide cases into branches.
5. Repeat the process until all cases in the branch result in the appropriate decision or have the same class.
6. The decision tree partitioning process will stop when [16] :
   a. All records in node N get the same class.
   b. There are no attributes or variables in the partitioned record anymore.
   c. There are no records in the empty branch.

The last stage of evaluation, which is to analyze the results obtained in the

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. VIII No 3, Agustus 2022, hlm. 295 - 304 ISSN 2550-0201 (Online)
DOI: https://doi.org/10.33330/jurteksi.v8i3.1680
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

previous stage which can be used to predict student graduation targets.

RESULT AND DISCUSSION

Data preprocessing is a process carried out to remove unnecessary data and improve the consistency of the content in the data, integrate data from data sources (databases) and be carried out before data mining to improve the overall quality of the pattern to be searched [7]. Using raw data from student academic data records, making it as input data that contains data that is used as a predictive factor, so this preprocessing is done to make it easier to transform and clean data.

The graduation prediction model is obtained from the classification results of student academic data records which will be used as input data and produce output in the form of a model that can predict student graduation. Obtained from the SISFO section of Telkom University in the form of Comma Separated Values (CSV) format in the form of student data for 2016 totaling 339 records. Then the input data will be used as a training set as training data in the formation of the prediction model. The DT classification is also used as a testing set as an implementation of the prediction model results that have been carried out. The data used as a training set of the overall data is 80% and 20% for the testing set. After that, enter 2017 data that has been adjusted with a total of 368 records as the implementation of the model. Determination of independent attributes is based on references [8] and selected three attributes, namely; TAK, Parents' Income and School Origin, then use several other attributes because they are in accordance with the research objectives. Based on the influence of the selection of student specialization on the prediction of graduation and also based on the student's academic value, the following is an explanation of the attributes used as predictors in the input data :

1. Number of Passed Credits, containing the number of Semester Credit Units (SKS) that passed in semester 1 to semester 6, a total of 118 credits. It is possible to find out how far the number of student credits that have been fulfilled in six semesters, to be able to complete credits in normal study programs, namely 141 Credits.
2. Total MKPP Values, containing the total scores in the prerequisite courses (MK) for student specialization with the highest total score of 72. Served as an attribute because the tendency of students to enter specialization is assessed based on the value of the MK Prerequisite for Specialization.
3. Specialization, contains information about each student's specialization.
4. TAK, contains the number of points or scores from the Student Activity Transcript (TAK). With less and enough categories, less if points < 60 and enough if points > 59.
5. EPRT, containing points or scores from language test results, EPrT is also used as one of the requirements to take part in the PA/TA/Thesis trial [17]. With less and enough categories, less if points < 450 and enough if points > 449.
6. Origin of School, contains the type of origin of the student's school during high school education. With the categories of Private MA, State MA, State High School, Private High School, State Vocational School and

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. VIII No 3, Agustus 2022, hlm. 295 - 304 ISSN 2550-0201 (Online)
DOI: https://doi.org/10.33330/jurteksi.v8i3.1680
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

Private Vocational School.

7. Parent's Income, containing information on the income of parents (guardians) of students based on Gross Domestic Product (GDP) per capita, Parent's income is also used as a graduation prediction factor in the study [18].

The modeling stage is by building a predictive model in accordance with the research, there are two interrelated processes, namely the training process and the testing process. Figure 3 depicts an outline flowchart for the training and testing process in building the C4.5 Algorithm decision tree.

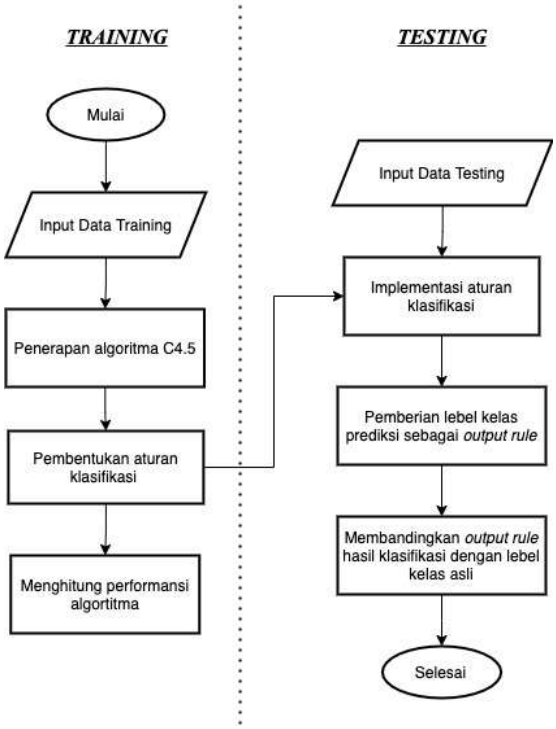

Figure 3. Flowchart of Training Set & Testing Set

Implementation is done by entering the input data into the training set and testing set and the results of the classification performance evaluation are based on the confusion matrix. The confusion matrix is an indicator of the overall accuracy as well as the sensitivity and accuracy of each class [19].

Table 1. Confusion Matrix Performance Evaluation

	On Time Prediction	Late prediction	Class Recall
Actual On Time	59 (TP)	3 (FN)	95%
Late Actual	1 (FP)	5 (TN)	83%
Class Precision	98%	62%	

From the results of the evaluation of the implementation, the results obtained accuracy of 94.11% and is a comparison of the accuracy of the prediction results with the actual data. Figure 4 depicts the results obtained in the prediction classification model using a decision tree state that the attribute that has the highest importance level is the number of credits passed, with the highest gain ratio and total entropy information as the root node.

$$Entropy(Total) = \left(\left(-\frac{235}{271}\right) * \log 2\left(\frac{235}{271}\right) + \left(-\frac{36}{271}\right) * \log 2\left(\frac{36}{271}\right)\right) = 0{,}565$$

$$Gain\ Ratio(Total, Jumlah\ SKS\ Lulus) = \frac{0{,}3269177}{0{,}191077773} = 1{,}710914$$

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. VIII No 3, Agustus 2022, hlm. 295 - 304 ISSN 2550-0201 (Online)
DOI: https://doi.org/10.33330/jurteksi.v8i3.1680
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

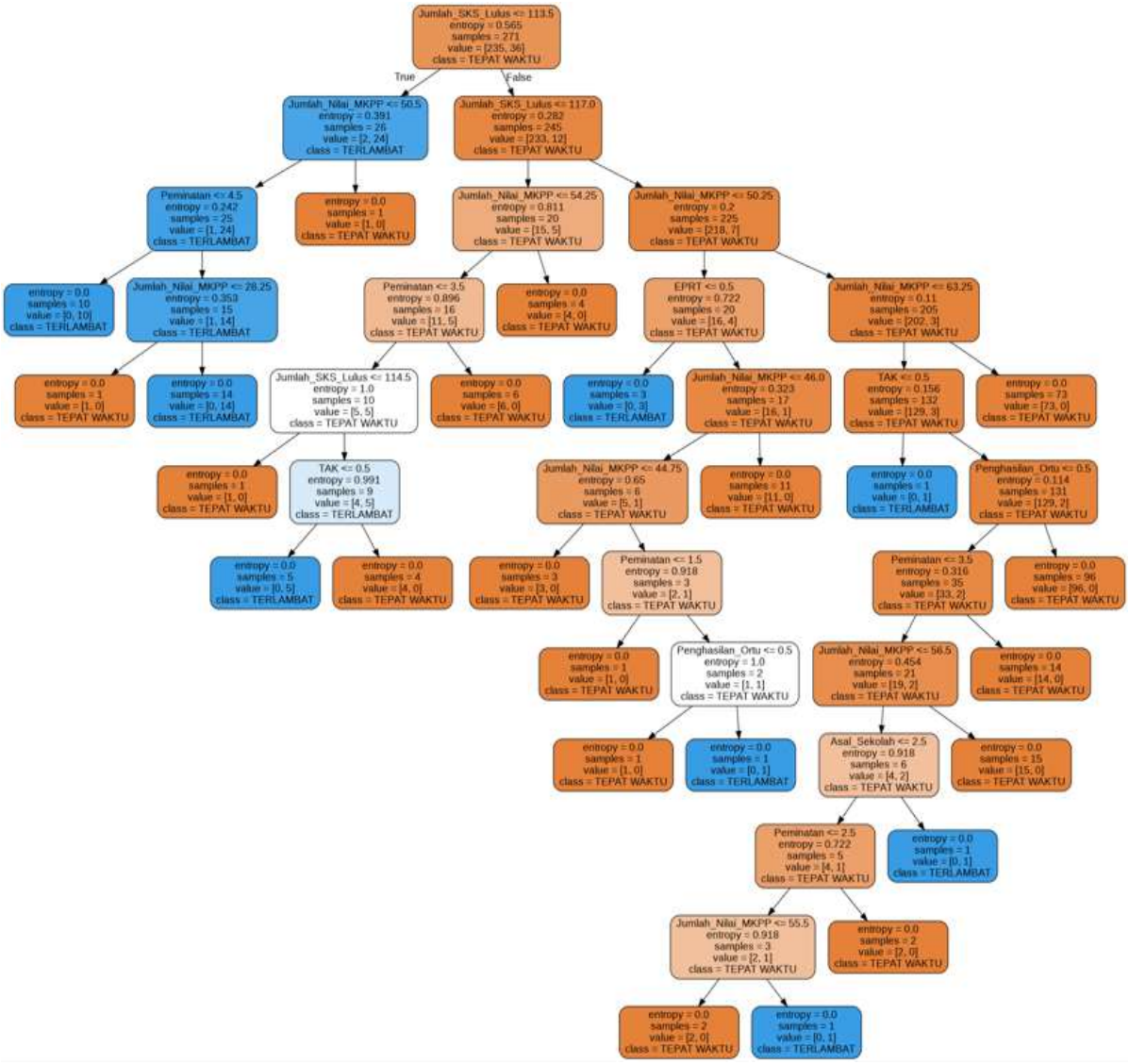

Figure 4. Decision Tree Classification Results

CONCLUSION

The results of the evaluation of the classification performance on the graduation prediction model for the 2016 IS study student academic data records were 339 records, resulting in an accuracy of 94.11%, precision results for the class on time of 98% and recall of 97%. All attributes that are predictors have an influence on the prediction of graduation.

The rule obtained from the prediction model is that if a student has a total of 118 credits passed, the total value of MKPP is above 64 and the results of specialization are all specializations, then the results can be predicted that students will graduate on time, but if the number of credits passed is less than 114, the total value of MKPP below 29 and the

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. VIII No 3, Agustus 2022, hlm. 295 - 304 ISSN 2550-0201 (Online)
DOI: https://doi.org/10.33330/jurteksi.v8i3.1680
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

results of specialization are all specializations, then the results can be predicted that students graduate late. It is also seen that there is an effect of choosing a specialization on the prediction of graduation. Thus, making predictions on the effect of selecting students' specializations can be done using the decision tree algorithm C4.5 and also the results of the research have been carried out with a more systematic path of determination and models so that they can be applied to PIPE applications on an ongoing basis.

BIBLIOGRAPHY

[1] U. Telkom, "Keunggulan Telkom University," *telkomuniversity.ac.id*, 2020. [Online]. Available: https://telkomuniversity.ac.id/keunggulan/.

[2] P. Chapman *et al.*, *CRISP-DM*. 2000.

[3] I. G. A. S. Melati, Linawati, and I. A. D. Giriantari, "Knowledge Discovery Data Akademik Untuk Prediksi Pengunduran Diri Calon Mahasiswa," *Maj. Ilm. Teknol. Elektro*, vol. 17, no. 3, 2018.

[4] N. Tanjung, D. Irmayani, and V. Sihombing, "Implementation of C5 . 0 Algorithm for Prediction of Student Learning Graduation in Computer System Architecture Subjects," *SinkrOn*, vol. 7, no. 1, pp. 274–280, 2022.

[5] M. Bramer, *Principles of Data Mining*, no. February. London: Springer, 2007.

[6] F. Gorunescu, *Data Mining Concepts, Models and Techniques*, Volume 12. Berlin: Springer, 2011.

[7] J. Han, M. Kamber, and J. Pei, *Data Mining Concepts and Techniques*. USA: Elsevier, 2012.

[8] R. Andreswari, M. A. Hasibuan, D. Y. Putri, and Q. Setyani, "Exploration Analysis of Data Mining Algorithm to Predict Student Graduation Target," *Atl. Highlights Eng. IEEE*, vol. Vol 2, no. July, 2019.

[9] M. Yusa, E. Utami, and E. T. Luthfi, "EVALUASI PERFORMA ALGORITMA KLASIFIKASI DECISION TREE ID3 ," *InfoSys J.*, vol. Vol 4, pp. 23–34, 2016.

[10] Risqiati and B. Ismanto, "Analisis Komparasi Algoritma Naive Bayes Dan C4-5 Untuk Waktu Kelulusan Mahasiswa," *IC- Tech*, vol. XII, no. 1, pp. 33–38, 2017.

[11] C. N. Dengen, Kusrini, and E. T. Luthfi, "Implementasi Decision Tree Untuk Prediksi Kelulusan Mahasiswa Tepat Waktu," *ResearchGate*, vol. 10, no. 1, pp. 1–11, 2020.

[12] F. Gullo, "From Patterns in Data to Knowledge Discovery : What Data Mining Can Do," *Elsevier*, vol. 62, pp. 18–22, 2015.

[13] D. E. Goldberg and J. H. Holland, "Genetic Algorithms and Machine Learning," in *Machine Learning 3*, Kluwer Academic, 1988, pp. 95–99.

[14] A. Verma, "STUDY AND EVALUATION OF CLASSIFICATION ALGORITHMS IN DATA MINING," *Int. Res. J. Eng. Technol.*, vol. Vol 5, pp. 1297–1307, 2018.

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. VIII No 3, Agustus 2022, hlm. 295 - 304 ISSN 2550-0201 (Online)
DOI: https://doi.org/10.33330/jurteksi.v8i3.1680
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

[15] Y. Kustiyahningsih and E. Rahmanita, "Aplikasi sistem pendukung keputusan menggunkaan algoritma c4.5. untuk penjurusan sma," *SimanteC*, vol. 5, pp. 101–108, 2016.

[16] Rismayanti, "Implementasi Algoritma C4.5 Untuk Menentukan Penerima Beasiswa di STT Harapan Medan," *J. Media Infotama*, vol. 12, no. 2, pp. 116–120, 2016.

[17] U. Telkom, "EPRT Kemahasiswaan," *studentstelkomuniversity.com*, 2021. [Online]. Available: https://studentstelkomuniversity.com/?s=eprt.

[18] K. D. Kolo, S. A. Adepoju, and J. K. Alhassan, "A Decision Tree Approach for Predicting Students Academic Performance," *Int. J. Educ. Manag. Eng.*, no. October, pp. 12–19, 2015.

[19] M. Wibowo, F. Noviyanto, S. Sulaiman, and S. M. Shamsuddin, "Machine Learning Technique For Enhancing Classification Performance In Data Summarization Using Rough Set And Genetic Algorithm," *Int. J. Sci. Technol. Res.*, vol. 8, pp. 1108–1117, 2019.