Computer-based Techniques for Predicting the Failure of Student Studies Using the Decision Tree method

D Arifin¹, A Hadiana²

¹Departemen Magister Sistem Informasi, Universitas Komputer Indonesia, Bandung, Indonesia.
²Lembaga Ilmu Pengetahuan Indonesia, Bandung, Indonesia.

Email: 75117003@mahasiswa.unikom.ac.id

Abstract. The purpose of this study is to predict students who have the potential to drop out of college so that the selection process for prospective students is more effective. Based on the problems that have been raised. The research method used was the forecasting method proposed to predict prospective students who drop out before entering college using the Decision Tree C4.5 method and Forward Selection. The tool used in this study used rapidminer 9.2. Based on the research results obtained, using 90% training data and 10% testing data resulted in an accuracy of 82.52% and obtained attribute models that affect the classification of student graduation, namely the Study Program and Age attributes.

1. Introduction

One of the factors that determine the quality of higher education is the percentage of students who can complete the study period. The high number of students who cannot complete their studies will make various problems involved in student studies such as the number of students who drop out of school in the college environment. The increase in education at the tertiary level has increased the number of applicants each year. This is inversely proportional to the number of graduates issued each year due to various kinds of reasons and problems experienced by each student, this becomes a problem that must be resolved by the college. Drop out is the process of withdrawing student status from students, which is caused by certain things that have been determined by the university in question [1].

The drop-out problem attracted many researchers to conduct research on student failure and school failure. In 2016, research was conducted on decision tree C4.5 method using 546 customer data resulting in an accuracy of 89.69% [2]. In 2009 research was conducted on the comparability of Decision Tree algorithms, Bayesian classifiers, logistic models, rule-based learners, and random forests by using 648 student data sets to make drop-out predictions, and in this study, the decision tree showed the highest accuracy [3]. According to Adhatrao in the study "Predicting Students Performance using ID3 and C4.5 Classification Algorithms" explained that the results of the prediction model C4.5 algorithm based on Decision Tree gave an accuracy for a total of 182 students, the average percentage of accuracy achieved in Bulk and Singular Evaluations is approximately 75.275, that the Decision tree is a fairly good model [4]. According to Sivakumar et al., that the accuracy of the classification tree decision algorithm that is fixed in the dataset has a greater effect [5-8]. according to Karthick et al., in his research, the decision tree method has the highest value of accuracy compared to the method of naive Bayes [9]. The decision tree method has the highest value of accuracy compared to the method of Support Vector Machine (SVM)[10]. in a study conducted uses the decision tree method for providing a ranking for the hospitals
based on the facilities available in the hospital produce Accuracy is nearly 80% and the class precision values are more than 60%[11]. Based on the study it has been proven that the decision tree method is suitable for large datasets[12]. In this study, besides getting good accuracy values, it was also aimed at obtaining influential attribute models by applying Feature Selection. Feature Selection is one way to determine the most influential attributes in the dataset. Feature Selection plays the role of choosing the right subset of the original feature set because not all features/attributes are relevant to the problem [13]. In order for universities to provide scholarships to students, an accurate study failure prediction technique is needed based on student data, so that prospective student who has the potential to fail the study are not selected to get a scholarship [14].

The purpose of this study is to predict students who have the potential to drop out of college so that the prospective student selection process is more effective and the method used is the Decision Tree C4.5 method and forward selection.

2. Method

2.1. Algorithm Analysis
Algorithms are effective methods expressed as limited circuits, algorithms are also collections of commands to solve a problem. Since the analysis of algorithms is independent of the computer or programming language used, algorithms are given in pseudo-code [15].

In scientific disciplines, the algorithm is studied abstractly. One of the topics of the algorithm is the application of decision tree predictions or what is often called the Decision Tree.

2.2. Decision Tree
The decision tree is a very strong and well-known method of classification and prediction. The decision tree method converts very large facts into decision trees that represent rules. Rules can be easily understood with natural language.

The process in the decision tree is to change the form of data (tables) into a tree model, change the tree model into a rule, and simplify the rule. A decision tree is (usually) a binary tree, with an inequality test on one of the attributes at each internal node, and a target attribute value associated with each leaf[16].

The decision tree is a classification method that is popular and widely used practically. Data Mining is the art and science of discovering useful innovative patterns from data[17]. When compiling a decision tree the first thing to do is to determine which attribute will be the root node and which attribute will be the next node. Good attribute selection is an attribute that allows getting the smallest decision tree of its size. Or attributes that can separate objects according to their class. Heuristically the attribute chosen is the attribute that produces the most "purest" node (the cleanest). The purity measure is expressed by the level of impurity, and to calculate it, can be done using the Entropy concept.

Entropy is diversity or diversity. Entropy expresses the impurity of a collection of objects. If given a set of objects with label/output y which consists of objects labeled 1, 2 to n, Entropy of objects with this n class can be calculated by the following formula.

\[ \text{Entropy} (S) = \sum_{i=1}^{n} - p_i \log_2 p_i \]  

Information:

\( S = \text{Set of cases} \)
\( n = \text{Number of S partitions} \)
\( p_i = \text{Si's proportion to S} \)

Then after calculating Entropy, calculate Information Gain:

\[ \text{Gain} (S, A) = \text{entropy} (S) - \sum_{i=1}^{n} \frac{|S_i|}{S} \times \text{Entropy} (S_i) \]  

Information:

\( S = \text{Case Set} \)
\( A = \text{Attribute} \)
Information gain is the most popular criteria for attribute selection. Information gain can be calculated from output data or dependent variable $y$ grouped by attribute $A$, denoted by the gain $(y, A)$. Information gain, gain $(y, A)$, from attribute $A$ relative to output data $y$ is: Where value $(A)$ is all possible values of attribute $A$ and is a subset of $y$ where $A$ has value $c$.

According to in terms of information gain is information acquisition, information gain will experience problems for attributes that have very varied values. The process in the decision tree is to change the form of data (tables) into a tree model, change the tree model into a rule, and simplify the rule.

C4.5 algorithm is the development of the ID3 algorithm, where development is carried out in terms of ability to overcome missing data, can handle continuous data, pruning. In general, the steps of the C4.5 algorithm for building decision trees are as follows:

1. Select the attribute as root.
2. Create a branch for each value.
3. Share cases in branches.
4. Repeat the process for each branch until all cases in the branch have the same class.

2.3. Data Mining Process

The terms data mining and knowledge discovery databases (KDD) are often used interchangeably to explain the process of extracting information in a database. Broadly speaking, the KDD process can be explained as follows:

1. Data Selection

   Select data from a set of operational data before the information excavation stage is carried out. The selection data that will be used is stored separately from the operational database (See Table 1).
Table 1. Attributes

| No | Attribute            | Value                                                                 |
|----|----------------------|----------------------------------------------------------------------|
| 1  | Study Programs       | 1. Diploma Three of Midwifery                                       |
|    |                      | 2. Diploma Three of Nursing                                        |
|    |                      | 3. Diploma Three of Optician Refraction                             |
|    |                      | 4. Bachelor of Nursing                                              |
|    |                      | 5. Bachelor of Public Health                                        |
|    |                      | 6. Profession of Nursing                                            |
| 2  | Class                | 1. Regular                                                          |
|    |                      | 2. Extension                                                        |
| 3  | Sex                  | 1. Female                                                           |
|    |                      | 2. Male                                                             |
| 4  | Place of Birth       | 1. Bandung                                                          |
|    |                      | 2. Outside Bandung                                                 |
| 5  | Origin of School     | 1. Jawabarat                                                        |
|    |                      | 2. Outside Jawabarat                                                |
| 6  | Age                  | 1. <20                                                              |
|    |                      | 2. 20-29                                                            |
|    |                      | 3. 30-39                                                            |
|    |                      | 4. >=40                                                             |
| 7  | IQ                   | 1. <90                                                              |
|    |                      | 2. 90-109                                                           |
|    |                      | 3. 110-199                                                          |
|    |                      | 4. >=120                                                            |
| 8  | Grade Point Semester | 1. <2.75                                                            |
|    |                      | 2. 2.75 - 2.99                                                      |
|    |                      | 3. 3 - 3.49                                                         |
|    |                      | 4. >3.5                                                             |
| 9  | Grade Point Semester | 1. <2.75                                                            |
|    |                      | 2. 2.75 - 2.99                                                      |
|    |                      | 3. 3 - 3.49                                                         |
|    |                      | 4. >3.5                                                             |
| 10 | Grade Point Semester | 1. <2.75                                                            |
|    |                      | 2. 2.75 - 2.99                                                      |
|    |                      | 3. 3 - 3.49                                                         |
|    |                      | 4. >3.5                                                             |
| 11 | Grade Point Semester | 1. <2.75                                                            |
|    |                      | 2. 2.75 - 2.99                                                      |
|    |                      | 3. 3 - 3.49                                                         |
|    |                      | 4. >3.5                                                             |
| 12 | Study                | 1. Failed                                                           |
|    |                      | 2. Succeed                                                          |

2. Cleaning
This cleaning process includes removing data duplication, checking inconsistent data, and correcting errors in data.

3. Transformation
Some data mining techniques require special data formats before they are applied because they are transformed so that the data is suitable for the data mining process. This transformation and selection of data also determine the quality of the results of data mining.

4. Data Mining
By using certain techniques or methods that are applied to selected data to look for patterns or information from the data.
5. Evaluation / Interpretation
The pattern or information generated from the data mining process is presented in an easily understood form. This stage includes examining whether the pattern or information found is contrary to the facts or hypotheses that existed before.

2.4. Confusion Matrix
A Confusion matrix is a tool used to evaluate classification models to estimate objects that are right or wrong. A matrix of predictions that will be compared with the original class of input or in other words contains information on actual values and predictions on classification.

In measuring performance using confusion matrix, there are 4 (four) terms as a representation of the results of the classification process. The four terms are True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN). True Negative Value (TN) is the number of negative data detected correctly, while False Positive (FP) is negative data but detected as positive data. Meanwhile, True Positive (TP) is positive data that is detected correctly. False Negative (FN) is the opposite of True Positive, so data is positive, but it is detected as negative data (See Table 2).

| Classification | Predicted class |
|----------------|-----------------|
| Class = Yes    | Class = No      |
| Class = Yes    | a (true positive-TP) b (false negative-FN) |
| Class = No     | c (false positive-FP) d (true negative-TN) |

The formula for calculating the accuracy of the matrix is:

\[
\text{Accuracy} = \frac{TP + TN}{TP + FN + FP + TN} \times 100\% = \frac{A + D}{A + B + C + D} \times 100\% \quad (2.3)
\]

\[
\text{Precision} = \frac{TP}{TP + FP} \times 100\% = \frac{A}{A + C} \times 100\% \quad (2.4)
\]

\[
\text{Recall} = \frac{TP}{TP + FN} \times 100\% = \frac{A}{A + B} \times 100\% \quad (2.5)
\]

Information
- TP is True Positive, which is the amount of positive data that is correctly classified.
- TN is True Negative, which is the amount of negative data that is correctly classified.
- FN is False Negative, which is the amount of negative data but incorrectly classified.
- FP is False Positive, which is the number of positive data but incorrectly classified.

Precision is the level of accuracy between the information requested by the user and the answer given by the system. Whereas recall is the success rate of the system in rediscovering information.

2.5. Split Validation
Split Validation is a validation technique that divides data into two parts randomly, partly as training data and others as test data. By using Split Validation, a training experiment will be conducted based on a predetermined split ratio, then the remainder of the split ratio training data will be considered as test data. Training data is data that will be used in conducting learning while the test data is data that has
never been used as learning and will function as test data for the truth or accuracy of learning outcomes (See Figure 1).

3. Results and Discussions

3.1. Data collection
This study uses data on students at the Dharma Husada Bandung School of Health Sciences as many as 1436 students from 2009 to 2014. The data is divided into training data and test data. Training data is used to find patterns of existing data, while test data is used to measure the accuracy of the data classification.

3.2. Validation and Evaluation
The main objective of this study is to analyze the predictions of potentially non-active students by applying data mining classification techniques with decision tree algorithms. In the testing phase of this model, the data used has passed the preprocessing stage. Next is the design of the model that will be used (See Figure 2).
Information:
1. Read Excel
   This operator is used to import the data set to be used, in this study the data is imported from the excel file.
2. Validation
   The validation method used in the study is Split Validation, this validation only divides the total from the entire dataset into training data and test data based on a predetermined ratio.
3. Decision Tree
   The classification method used in this study
4. Apply Model
   Implement the model that has been created
5. Performance
   Operators used to measure the accuracy of the model's performance

3.3. Experiment Results
In this study tested the accuracy of the student's graduation classification using the Decision Tree method. The highest accuracy results in experiment 5 in the table using training data 50%, 50% test data resulted in the accuracy of 80.78% which was supported by the precision value of 84.14% and the recall value of 92.14% (See Table 3).

| Experiment | 1   | 2   | 3   | 4   | 5   | 6   | 7   | 8   | 9   |
|------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Training Data | 90% | 80% | 70% | 60% | 50% | 40% | 30% | 20% | 10% |
| Testing Data  | 10% | 20% | 30% | 40% | 50% | 60% | 70% | 80% | 90% |
| Accuracy     | 78.32% | 77.40% | 76.10% | 79.62% | 80.78% | 77.15% | 79.50% | 77.00% | 77.03% |
| Precision    | 81.45% | 82.28% | 76.10% | 83.61% | 84.14% | 82.28% | 82.31% | 83.01% | 79.28% |
| Recall       | 92.66% | 89.04% | 100% | 91.08% | 92.14% | 89.18% | 93.07% | 78.76% | 94.51% |
Next test the accuracy of student graduation classifications using the Decision Tree and Forward Selection methods. The highest accuracy results in experiment 1 produced an accuracy of 82.52% which was supported by the precision value of 82.31% and the recall value of 98.17% (See Table 4).

Table 4. The results of the accuracy of the Decision Tree and Forward Selection methods

| Experiment | 1   | 2   | 3   | 4   | 5   | 6   | 7   | 8   | 9   |
|------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Training Data | 90% | 80% | 70% | 60% | 50% | 40% | 30% | 20% | 10% |
| Testing Data  | 10% | 20% | 30% | 40% | 50% | 60% | 70% | 80% | 90% |
| Accuracy     | 82.52% | 79.17% | 82.13% | 77.87% | 79.11% | 80.05% | 80.30% | 80.05% | 78.34% |
| Precision    | 82.31% | 81.42% | 84.38% | 80.27% | 81.36% | 82.18% | 82.55% | 83.21% | 77.85% |
| Recall       | 98.17% | 94.06% | 93.90% | 94.05% | 94.15% | 94.21% | 93.99% | 92.45% | 100.00% |

3.4. Discussion
The purpose of this study is to obtain the relevant model features/attributes and the highest value of accuracy. Experiments of the Decision Tree method obtained the highest accuracy of 80.78%. Subsequent experiments with the Decision Tree and Forward Selections methods as a selection feature obtained the highest accuracy results of 82.52% and obtained attribute models that influence student graduation classifications, namely the attributes of the Study Programs and Age.

4. Conclusion
The results of the discussion can be concluded that Forward Selection methods can help to improve the accuracy of the decision tree classification method. In this case, the decision tree method utilizes the Forward Selection feature selection function for selecting data attributes with the characteristics of the data itself and increasing the accuracy of the prediction accuracy. The Decision Tree and Forward Selection method are more accurate and effective in classifying student graduation status with an accuracy of 82.52%. By obtaining the most influential attributes namely: Study Programs, Age.

References
[1] Nurhayati, S., and Luthfi, E. T. (2015). Prediksi Mahasiswa Drop Out Menggunakan Metode Support Vector Machine. Sisfotenika, 5(1), 82-93.
[2] Sudrajat, R., Irianingsih, I., and Krisnawan, D. (2017, January). Analysis of data mining classification by comparison of C4.5 and ID algorithms. In IOP Conference Series: Materials Science and Engineering (Vol. 166, No. 1, p. 012031).
[3] Dekker, G. W., Pechenizkiy, M., and Vleeshouwers, J. M. (2009). Predicting Students Drop Out: A Case Study. International Working Group on Educational Data Mining, vol.2, no 1,pp. 33-36.
[4] K. Adhatrao, A. Gaykar, A. Dhawan, R. Jha and V. Honrao, 2013, “Predicting Students Performance Using ID3 and C4.5 Classification Algorithms”, International Journal of Data Mining and Knowledge Management Process (JDKP), vol.3. no 4 pp. 45-56.
[5] S. Sivakumar, S. Venkataraman, and R. Selvaraj, 2016, "Predictive Modeling of Student Dropout Indicators in Educational Data Mining using Improved Decision Tree", Indian Journal of Science and Technology, vol. 9. No. 4 pp 115-125.
[6] Anuradha, C., and Velmurugan, T. (2015). A comparative analysis on the evaluation of classification algorithms in the prediction of Students performance. *Indian Journal of Science and Technology*, 8(15), 1-12.

[7] Chitra, R., and Seenivasagam, V. (2013). Review of heart disease prediction system using data mining and hybrid intelligent techniques. *ICTACT Journal on Soft Computing*, 3(04), 605-609.

[8] Kumar, M. (2016, October). Predicting Students’ Performance Using Classification Techniques in Data Mining. In *International Journal of Technology and Computing (IJITC)* (Vol. 2, No. 10 (October, 2016)). Techlive Solutions.

[9] Karthick, S., Malathi, D., and Yashwant, K. (2018). Analysing the causes of cyclonic severity in bay of bengal and its effects in biotic species using XGBoost algorithm. *Indian Journal of Public Health Research and Development*, 9(11), 2241-2245.

[10] Mardiani, G. T. (2018, August). Construction industry project planning information system. In *IOP Conference Series: Materials Science and Engineering* (Vol. 407, No. 1, p. 012093). IOP Publishing.

[11] Garwe, T., Stewart, K., Newgard, C. D., Stoner, J., Sacra, J. C., Cody, P., and Albrecht, R. M. (2019). Survival Benefit of Treatment at or Transfer to a Tertiary Trauma Center among Injured Older Adults. *Prehospital Emergency Care*, (just-accepted), vol 14 no.21 pp.1-18.

[12] Hassan, F., Farhan, S., Fahiem, M. A., and Tauseef, H. (2018). A Review on Machine Learning Techniques for Software Defect Prediction. *Technical Journal*, 23(02), 63-71.

[13] Kurniawan, B. (2018, August). Integrated Information System for Radio Frequency Identification Based Administration and Academic Activities on Higher Education. In *IOP Conference Series: Materials Science and Engineering* (Vol. 407, No. 1, p. 012097). IOP Publishing

[14] Rahman, M. A. (2017). Algoritma C45 Untuk Menentukan Mahasiswa Penerima Beasiswa (Studi Kasus: PPS IAIN Raden Intan Bandar Lampung). *Jurnal Teknologi Informasi Magister*, I(02), 118-128.

[15] Soegoto, E. S., and Jayaswara, M. R. (2018, August). Web and Android Programming Course Information System. In *IOP Conference Series: Materials Science and Engineering* (Vol. 407, No. 1, p. 012063).

[16] Hamoud, A., Hashim, A. S., and Awadh, W. A. (2018). Predicting Student Performance in Higher Education Institutions Using Decision Tree Analysis. *Hamoud, AK, AS Hashim, and WA Awadh, Predicting Student Performance in Higher Education Institutions Using Decision Tree Analysis. International Journal of Interactive Multimedia and Artificial Intelligence*, 5(2), 26-31.

[17] Gil, N. R. D. A. (2018). *Marketing preditivo: exemplo de uma campanha de crédito pessoal* (Doctoral dissertation) 14(5), 104-115.