The Adaboost Integration to the C4.5 Algorithm in Improving Study Interest Classification Accuracy

S Fahriah[*1], N D Kamarudin[2], L Triyono[3], A Rizaldy[4]

[1]Technical Information, Politeknik Negeri Semarang

[2]Faculty of Science and Defence Technology, National Defence University of Malaysia

[3]Computer Engineering Technology, Politeknik Negeri Semarang

[4]Information System Department, Universitas Islam Negeri Alauddin Makassar

E-mail: sirlifahriah@polines.ac.id[1], nursyafiqah@siswa.ukm.edu.my[2], liliek.triyono@polines.ac.id[3], adhy.rizaldy@uin-alauddin.ac.id[4]

Abstract. Specialization is one of the important things in focusing on the student's field of study. At one university with a computer science faculty, there is an Informatics Engineering undergraduate study program divided into two specializations, namely intelligent systems and software engineering design. At one university with a computer science faculty, there is a bachelor's program in Informatics Engineering divided into two specializations, namely intelligent systems and software engineering design. Students need help to choose one of the specializations in the informatics engineering study program. To overcome these problems, the authors provide solutions in the form of ideas that can help students determine specialization. In this problem, the algorithm will be used is the C4.5 algorithm based on forward selection plus adaboost. The results of the specialization classification use the selected attributes and iterate over the cross-validation to produce the proper accuracy. Testing the C4.5 algorithm produces an accuracy of 93.89%, and the C4.5 algorithm based on forward selection produces an accuracy of 94.44%, while using the C4.5 algorithm based on forward selection with the addition of adaboost produces an accuracy of 94.63%. These tests prove that there is an increase in accuracy by adding selection and adaboost features to the C4.5 algorithm.

Keywords: Classification, Algorithm C4.5; feature selection, adaboost, specialization in the field of study.

1. Introduction

The field of interest is the most critical factor in education. The field of interest in an education or university is one of the determinants of the future for all students—the main area of interest, especially for students who want to work in the field of interest. Many students need help understanding their talents, so it is challenging to determine areas of interest that match them. According to some experts, talent or area of interest is a condition in which a person is able to achieve a particular skill, knowledge, and skill

through special training [1]. Talent is a condition or quality possessed by each individual, which will likely develop in the future [2]. Some students need help determining their area of interest. As a result of choosing an area of interest, there will be several impacts. One of the impacts that arise from selecting an inappropriate field of interest is the need for more interest in student learning. The choice of the field of interest that is not right for students is different from talent so that student interest in learning decreases. Many cases are found due to choosing the wrong field of interest that is different from talent resulting in student interest in learning to fall. Decreased student interest in learning because of choosing the wrong field of interest can affect student performance in attending lectures. This results in a decrease in the quality of students' academic scores. Of all the impacts that may occur, another impact is due to inappropriately choosing the field of interest, namely students are reluctant to study the field of study that has already been chosen, resulting in a decrease in academic value which has the impact of not graduating on time.

In most educational psychology research, students' interests and talents are pretty related to academic achievement [3]. In terms of areas of interest, students of the informatics engineering study program at one university are divided into two areas of interest, namely intelligent systems and software design engineering. To get a pattern from the data to be analyzed so that it becomes a decision result, several processes and stages are needed such as the selection of attributes or parameters from the data that has been processed. In this interest field classification study, researchers select attributes based on several considerations that greatly affect the results of specialization decisions.

Based on the previous review, here are some factors that influence a research including an algorithm that is not suitable based on the dataset, attributes that are not suitable to be used as a parameter, the distribution of the data set between training and testing. Based on several studies that have been done previously, it is possible to use various data mining algorithms to solve similar problems regarding the selection of majors. These algorithms include decision trees, naive bayes, clustering and so on. Based on several studies that have been carried out, the researcher sees that the performance of the C4.5 algorithm is suitable for determining the area of interest [4]. C4.5 selects attributes according to the gain ratio. This algorithm overcomes the weakness of leaning towards attributes that have more value when using gain information. Some datasets also need help with data quality issues or data imbalance. Class imbalances occur in several datasets, and dataset imbalances will directly affect classification accuracy [4]. The decision tree method has a good performance in handling on time or late classification but the decision tree needs to improve in the high level of class imbalance. To overcome these problems can be done with methods that can balance the class and increase accuracy.

Adaboost is a boosting method that can balance classes by giving weights to the level of misclassification that can change the distribution of data [15]. This problem approach is planned to use the adaboost implementation on the feature selection-based C4.5 algorithm. The application of adaboost to the feature selection-based C4.5 algorithm will later be applied to the accuracy of selecting student interest areas. By using this approach is expected to increase the value of accuracy and produce accurate decisions.

2. Literature Review

The following are some of the studies that examine the selection of students' areas of interest:

Al-malaise et al [5] have conducted research on student performance in 2014. This study classified student performance which tends to be low. In this study, the C4.5 algorithm was added with the samme adaboost technique. With the addition of the samme technique, the results of the C4.5 algorithm are better with an accuracy rate of 78% while 60% without samme adaboost.

Prafulla Kalapatapu et al [6] investigated the classification of music in India in 2016. The results show that using feature selection only sometimes increases accuracy, but can still increase accuracy in some situations depending on the right method. The selection feature selection based on gain information

from neural networks and SVM is the best for the Indian song data set. In another study, Al-Sarem [7] researched in 2015. This research is for academic purposes based on the C4.5 algorithm. The study results show that the C4.5 algorithm can be applied in this case and provides accuracy in classification.

In 2016, Wiwit Supriyanti et al [8] researched on the Comparison of the Performance of the C4.5 and Naïve Bayes Algorithms for the Accuracy of Student Concentration Selection. The trial was carried out using data obtained from the University of Muhammadiyah Surakarta. The results of the comparison between C4.5 and Naïve Bayes combined with forward selection obtained the highest accuracy result, namely C4.5 at 84.98%.

On another occasion, Holisatul Munawaroh et al [9] investigated the comparison of ID3 and C5.0 algorithms in identifying majors of high school students. The data used in this study is the data of class X students from the class of 2011/2012 obtained from SMAN 2 Bangkalan as much as 200 data. This study resulted in an accuracy of C5.0 95% better than ID3 by 93%.

Another study by Yusuf S Nugroho et al [10] compared the nave Bayes algorithm, C4.5 and K-means for the selection of majors at SMA N 1 Boyolali. The data used consisted of gender, interests, average science scores, average social studies scores, science psychotest, social studies psychotest, school origin and majors. Based on the accuracy and recall values, the decision tree method is better than other methods with an accuracy value of 79.14% and a recall value of 90.80%. Based on the precision value, the Naive Bayes method is better than the other methods with a value of 77.51%.

3. Research Method

3.1. Research Design

To support research in the field of data mining, the application of the CRISP-DM (Cross Industry Standard Process for Data Mining) method will be carried out. CRISP-DM is a model of the data mining process to describe the sequence in data mining, providing a standard data mining process that is used as a strategy in solving general problems. While the research in this study used experimental research. Experimental research involves investigating the treatment on variable parameters and using tests determined by the researcher. The research method carried out several steps such as data set collection, data set processing, proposed method, evaluation and validation of results.

3.2. Collection of Data Sets

Data collectors in this study took data from Informatics Engineering students from a university. The data taken in this study is the data of students of the Informatics Engineering study program of approximately 950 records. The data will be divided into training data to get the data model and testing data to test the data model. The distribution of training and testing data is divided into a ratio of 9 to 1. The data includes student ID, student status, social studies, courses that have been taken along with their grades and areas of interest taken. Not all attributes are used based on the data obtained, but through preprocessing the data to obtain attributes that may affect the selection of the student's field of interest. The requirements for the specialization that have been determined are some of the courses that have been written in the academic guidebook obtained at the time of entering the lecture. At this stage the data attributes used include: Course Values and areas of interest.

3.3. Data Set Processing

The initial data processing process is needed to prepare data that is truly valid before being processed. In this study, it was done by cleaning the data or cleaning the blanks. At this stage, the initial processing of the data set is carried out to simplify the data so that the data can be recognized and used as a model of the proposed algorithm. The data will be processed with data mining tools. If there is inconsistent data or missing values, selecting data with data filters is necessary. It is done for data that has missing values by deleting records on data with missing values. Filtered data is data that has complete attributes as specified.

Researchers only use complete data with selected attributes. After the data is selected and cleaned the variables or attributes that have an effect are obtained so that the variables become 12 regular attributes and one label attribute and can be directly processed.

3.4. Proposed method

The proposed method for classifying study interests is the application of adaboost on the C4.5 algorithm based on Forward Selection. The dataset will be selected using the Forward Selection method as a feature selection method for attributes that are less influential or irrelevant to the dataset with the aim of increasing accuracy. Furthermore, classification is carried out with the application of adaboost on the C4.5 algorithm as an algorithm that classifies the selection of fields of interest for informatics engineering students. As for the validation using 10-fold cross validation. The stages and methods are described as below:

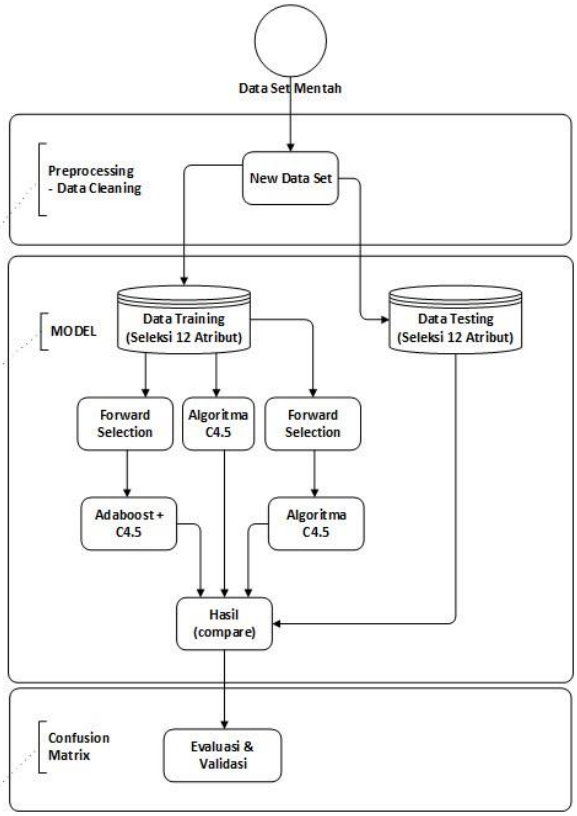

Figure 1. Proposed method

3.5. Validation and Evaluation of Results

After going through several stages and processes, the result of the research is a model and the accuracy of the data that has been processed. The model or pattern that can be used as a classification for the selection of the field of interest of the informatics engineering student is therefore necessary for evaluation and validation. Furthermore, in this stage, an evaluation of the model that was obtained in the previous phase will be carried out. The evaluation aims to adjust the model obtained so that it is right on target according to the target to be achieved. At this stage the model formed from the application of adaboost on the C4.5 algorithm based on Forward Selection will be tested using a Confusion Matrix diagram to determine the

accuracy value. From the results of the Confusion Matrix diagram, it will be known the results of positive predictions that are true and false, as well as the results of negative predictions that are true and false. Then the validation process is carried out using the k-fold cross validation method to determine the average success of a system, the method is to iterate by randomizing the input attributes so that the system is tested for several random input attributes [22]. The results of the classification are expected that by using the adaboost method and the selection of forward selection features can optimize accuracy and good performance against the C4.5 algorithm.

4. Results

After going through several stages and processes, the result of the research is a model and the accuracy of the data that has been processed and tested. The model or pattern can be used as a classification for the selection of the field of interest of the informatics engineering student. The results of the research are expected to produce an appropriate and accurate classification of the selection of fields of study interest.

4.1. Experiment with C4.5. Algorithm

At this stage the data will be tested only using the C4.5 algorithm. Testing by classifying the data so as to produce a model or a certain rule (rule) for the classification of the selection of the field of interest for informatics engineering students. Testing data as many as 300 records consisting of 12 attributes and 1 label with reference to the calculation of entropy and gain information. From the calculation results will produce a decision tree which is a new rule of data to make it easier to understand.

4.2. Testing on C4.5. Algorithm

To find out how the performance of the C4.5 algorithm in classifying the selection of students' fields of interest, a confusion matrix was tested. From the dataset, which amounted to 540 students of informatics engineering, there were 327 areas of interest for SC and 213 areas of interest for RPLD. Of the 327 data on the field of interest for SC students classified as field of interest in SC is correct (TP), a total of 314 data and students classified as field of interest are incorrect (TN) as many as 13 data. Meanwhile, from the 213 data on the RPLD interest field, students classified the RPLD interest area as correct (FP) with as many as 193 data, and students classified the RPLD interest area as incorrect (FN) with 20 data.

4.3. Experimental C4.5 Algorithm based on Forward Selection

The use of forward selection in this study is to select attributes that have no effect or are less relevant to the problem and are expected to increase the accuracy of the C4.5 algorithm. The input variables for the specified course values include Calculus I Values, Basic Programming Values, Object-Based Programming Values, Software Engineering Values, Database Values, Informatics Logic Values, Programming Algorithms Values, Discrete Mathematics Values, Data Structure Values, Matrix Values & Vector Space, Value Calculus II, Probability Value and Statistics. From the data, initialization of the attributes in the data will be carried out.

4.4. Testing with the C4.5 Algorithm based on Forward Selection

To find out how the performance of the forward selection-based C4.5 algorithm in classifying the selection of students' fields of interest, a confusion matrix was tested. The above experiment uses attributes that have passed the selection process. The attributes used in the above test are attributes that are considered influential for the attribute classification process including Calculus I Value, Basic Programming Values, Software Engineering Values, Database Values, Informatics Logic Values, Programming Algorithms Values, Matrix Values & Vector Spaces, Values Calculus II, Value Probability and Statistics. The results of the experiment with the forward selection based C4.5 algorithm prove that

this method can increase the accuracy of the C4.5 algorithm without forward selection from 93.89% to 94.44%.

4.5. Experiment using Adaboot the C4.5 algorithm based on Forward Selection

This research uses the adaboost method on the C4.5 algorithm based on forward selection to balance unbalanced data sets. although it has been tested on each algorithm, namely C4.5, C4.5 based on forward selection to find the best value. However, testing is still carried out using the adaboost method on the forward selection-based C4.5 algorithm. In this test, 540 data were used with 327 SC class data and 213 RPLD class data. The results of testing the adaboost application formula model on the C4.5 algorithm based on forward selection based on data mining tools is to make ten iterations and produce 10 C4.5 algorithm architecture based on forward selection. The decision tree generated by the application of adaboost on the C4.5 algorithm based on forward selection is 7 models and produces a different root for each.

4.6. Testing using Adaboot the C4.5 algorithm based on Forward Selection

In the third or last test, the researcher tested the performance of adaboost by applying it to the c4.5 algorithm based on forward selection. In this study, although it has been tested on each algorithm, namely C4.5 and C4.5 based on forward selection, testing is still carried out by applying adaboost to the C4.5 algorithm based on forward selection and then comparing the three methods. As in the previous test, to find out how the performance of adaboost applied to the forward selection-based C4.5 algorithm in classifying the selection of students' fields of interest, a confusion matrix was tested. Experiments using attributes that have passed the feature selection process in testing the C4.5 algorithm based on forward selection and adaboost will be applied. The attributes used in the above test are attributes that are considered influential for the attribute classification process including Calculus I Value, Basic Programming Values, Software Engineering Values, Database Values, Informatics Logic Values, Programming Algorithms Values, Matrix Values & Vector Spaces, Values Calculus II, Value Probability and Statistics. The results of the experiment with the application of adaboost on the forward selection-based C4.5 algorithm prove that this method can improve the accuracy that has increased to more. It is evident from the C4.5 algorithm without forward selection which was originally 93.89% increased to 94.44% with the forward selection-based C4.5 algorithm and further increased by the application of adaboost on C4.5 based on forward selection with an accuracy of 94.63%.

5. Evaluation and Validation

At this stage, evaluation and validation are carried out to test the algorithm that has been applied and selected with the k-fold cross validation evaluation method, namely by forming k subsets of the existing data set. The method begins by dividing the data a number of k-fold according to the researcher's determination. In the process the data will use 600 data with the application of the adaboost method on the forward selection based C4.5 algorithm. At this stage, the data will be divided into ten iterations of the same size. Furthermore, the training and testing process is carried out ten times with a comparison of 9 to 1 on the training and testing data. The 10-fold cross-validation method will work by carrying out tests that are repeated 10 times, and the measurement results in the form of an average accuracy of 10 times of testing.

5.1. Testing using K-Fold cross-validation

The initial stage of the 10-fold cross-validation method is to divide the 600 data sets into ten subsets (sections). The results of the division of 600 are each subset of 60 data. The first fold consists of a combination of 9 different subsets of data, which are combined into one and used as training data. While the remaining subset acts as testing data. Then the training and testing process is carried out until it folds

to 10. The table for the distribution of training data and testing data for the 10-fold cross-validation method is as shown in the table below:

Table 1. 10-Fold Cross Validation

	Data Training		**Data Testing**		
Fold	Subset	Data	Subset	Data	Accuracy
1	S2, S3, S4, S5, S6, S7, S8, S9, S10	540	S1	60	96.48%
2	S1, S3, S4, S5, S6, S7, S8, S9, S10	540	S2	60	93.70%
3	S1, S2, S4, S5, S6, S7, S8, S9, S10	540	S3	60	94.07%
4	S1, S2, S3, S5, S6, S7, S8, S9, S10	540	S4	60	95.56%
5	S1, S2, S3, S4, S6, S7, S8, S9, S10	540	S5	60	97.22%
6	S1, S2, S3, S4, S5, S7, S8, S9, S10	540	S6	60	96.30%
7	S1, S2, S3, S4, S5, S6, S8, S9, S10	540	S7	60	93.89%
8	S1, S2, S3, S4, S5, S6, S7, S9, S10	540	S8	60	94.44%
9	S1, S2, S3, S4, S5, S6, S7, S8, S10	540	S9	60	95.19%
10	S1, S2, S3, S4, S5, S6, S7, S8, S9	540	S10	60	95.93%
Average					95.28%

In the table above, it can be seen that the accuracy value obtained for each fold is stable. The lowest accuracy value is in the 2nd fold, which is 93.70%, and the highest accuracy value is in the 5th fold of 97.22%. The accuracy obtained is based on the results of calculations based on the confusion matrix. Below is an example of a confusion matrix of the highest accuracy.

Table 2. 5th Confusion Matrix Fold

Table	**true SC**	**true RPLD**	**Class Precision**
Pred. SC	315	9	97.22%
Pred. RPLD	6	210	98.22%
Class Recall	98.13%	95.89%	

Accuracy = (315 + 210) / (315+210+6+9) * 100%

= (525 / 540) * 100%

= 95.28%

From the accuracy results, the performance of the adaboost on the forward selection-based C4.5 algorithm for the classification process for the selection of fields of interest for informatics engineering students can be analyzed. The applied algorithm is good by producing an average accuracy value of 95.28%, which was tested using a 10-fold cross-validation test on 540 training data and 60 testing data.

6. Discussion

In this study, the classification of the selection of fields of interest for Informatics Engineering students using the C4.5 algorithm, C4.5 based on forward selection and the application of adaboost on C4.5 based on forward selection. By using the same data and attributes, namely 540 as training data, the best method in this experiment was to apply the adaboost method to the forward selection based C4.5 algorithm. The result of applying the adaboost method on C4.5 based on forward selection produces the greatest accuracy with increasing accuracy. The purpose of this study is to increase the accuracy value generated by the C4.5 classification algorithm by reducing some attributes in the dataset using forward selection and balancing the data using adaboost. The dataset is processed using a comparison of the results obtained between the C4.5 algorithm, C4.5 based on forward selection and the application of adaboost on C4.5 based on forward selection. the comparison of the three tests can be seen in the Table 3:

Table 3. Comparison of Accuracy

Method	Accuracy
C4.5 Algorithm	93.89%
C4.5 + FS	94.44%
C4.5 + FS + Adaboost	94.63%

It can be seen in the three experiments that have been carried out above by comparing methods with the same data. The experiments using the adaboost method on C4.5 based on forward selection with 540 training data resulted in the greatest accuracy of 94.63%. By using the evaluation method 10-fold cross validation which is used to determine the average success of the accuracy values obtained is valid or not, as well as measuring the performance of the algorithm, the average accuracy of 10 test iterations is 95.28%.

7. Conclusion

The conclusions obtained are based on the results of the application of the C4.5 algorithm, C4.5 based on forward selection and the application of adaboost on C4.5 based on forward selection for the classification of the selection of fields of interest for informatics engineering students as follows:

1. The use of adaboost and forward selection as a feature selection boosting method is proven to increase accuracy where the attributes that were originally involved in the dataset are reduced by selecting only the influential attributes and the unbalanced data becomes balanced.
2. The test results of the performance of adaboost implementation on C4.5 based on forward selection with 540 data produced the highest accuracy with 94.63% accuracy compared to the previous C4.5 algorithm 93.89% and C4.5 based on forward selection 94.44%.
3. The results of the application of the adaboost method on the forward selection-based C4.5 algorithm are said to be good and can be applied to this case with data from informatics engineering students.

References

[1] M. Musrofi, Melesatkan Prestasi Akademik Siswa, Yogyakarta: Pedagogja, 2010.

[2] S. Liliana, "Penerapan Algoritma C4.5 untuk penentuan Informatika dan Komputer (STMIK) Indonesia," *Jurnal GEMA AKTUALITA*, vol. 2, no. 1, Juni 2013.

[3] F. A. Prastya, "Penerapan Algoritma C4.5 untuk Prediksi Jurusan Siswa SMAN 3 Rembang," *Skripsi Teknik Informatika Universitas Dian Nuswantoro Semarang*, 2015.

[4] Z. Quan, G. Lin-gang, W. Chong-jun, W.-j. and C. Shi-fu, "Using An Improved C4.5 for Imbalanced Dataset of Intrusion," *National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210093.*

[5] D. A. AL-Malaise, D. A. Malibari and M. Alkhozae, "Students' Performance Prediction System Using Multi Agent Data Mining Technique," *International Journal of Data Mining & Knowledge Management Process (IJDKP),* vol. 4, no. 5, 2014.

[6] P. Kalapatapu, S. Goli, P. Arthum and A. Malapati, "A Study On Feature Selection and Classification Techniques Indian Music," *Procedia Computer Science,* no. 98, p. 125 – 131, 2016.

[7] M. Al-Sarem, "Building a Decision Tree Model for Academic Advising Affairs Based on the Algorithm C4. 5," *International Journal of Computer Science Issues*, vol. 12, no. 5, pp. 1694-0814, 2015.

[8] W. Supriyatin , Kusrini and A. Amborowati, "Perbandingan Kinerja Algoritma C4.5 dan Naive Bayes Untuk Ketepatan Pemilihan Konsentrasi Mahasiswa," *INFORMA Politeknik Indonusa Surakarta*, vol. 1, pp. 2442-7942, 2016.

[9] H. Munawaroh, "Perbandingan Algoritma ID3 dan C5.0 dalam Identifikasi Penjurusan Siswa SMA," *Proceding Program Studi Teknik Informatika, Universitas Trunojoyo.*

[10] Y. S. Nugroho and I. Graha, "Data MIning Jasa Pengiriman Titipan Kilat di PT CItra Van dengan Metode Decision Tree," *Naskah Publikasi Program Studi Teknik Informatika Fakultas Komunikasi dan Informatika Universitas Muhammadiyah Surakarta*, Juni 2014.

[11] W. Wiwid, " Optimasi Klasifikasi Status Gizi Balita Berdasarkan Indeks Antropometri Menggunakan Algoritma C4.5 Adaboost Classification," *Jurnal Ilmiah Komputerisasi Akuntansi, Vol.12, No.2, , pp. 45-51, Desember 2019*.

[12] F. A. Hermawati, Data Mining Bab I Teknik Data Mining, Yogyakarta: CV ANDI OFFSET, 2013.

[13] Kusirni and t. l. Emha, " Algoritma Data Mining," Yogyakarta, *Andi offset*, 2009.

[14] A. Sijabat, "Penerapan Data Mining Untuk Pengolahan Data Siswa dengan Metode Decision Tree (Studi kasus: Yayasan Kristen Andreas)," *Informasi dan Teknologi Ilmiah (INTI),* vol. V, no. 3, pp. 2339-210X, Pebruari 2015.

[15] H. A. Yoga, Kusrini and T. L. Emha, " Klasifikasi Penerimaan Mahasiswa Baru Menggunakan Algortima C4.5 Dan Adaboost (Studi Kasus : STMIK XYZ)," *CSRID Journal, Vol.9, no.1, Hal. 1-11, Februari 2017.*

[16] D. H. Kamagi and S. Hansun, Implementasi Data Mining dengan Algoritma C4.5 untuk Memprediksi Tingkat Kelulusan Mahasiswa, Tangerang, 2014, pp. 2085-4552.

[17] R. A. Saputra, "Komparasi Algoritma Klasifikasi Data MIning Untuk Memprediksi Penyakit Tuberculosis (TB): Studi Kasus Puskesmas Karawang Sukabumi," 2014.

[18] P. G. S. Cipta Nugraha and G. S. Mahendra, " Explorasi Algoritma C4.5 Dan Forward Feature Selection Untuk Menentukan Debitur Baik Dan Debitur Bermasalah Pada Produk Kredit Tanpa Agunan (KTA)," Jurnal Sains Teknologi, Vol. 9, no. 1, 2020.

[19] D. S. M. S. N, "Intrusion Detection using Naive Bayes Classifier with Feature Reduction," *Procedia Technology Elsevier, no. 4, pp,* pp. 119-128, 2012.

[20] P. Yang, "Ensemble-Based Wrapper Methods for Feature Selection and Class Imbalance Learning," *Springer, pp,* pp. 544-555, 2013.

[21] Z. Imaduddin and H. A. Tawakal, "Deteksi Dan Klasifikasi Daun Menggunakan Metode Adaboost dan SVM," *Seminar Nasional Teknologi Informasi dan Multimedia 2015,* pp. 2302-3805, 2015.

[22] A. Syamsudin, "Kebutuhan Penelitian di Bidang Ilmu Pendidikan.," 2004.