Jurnal Teknik Informatika (JUTIF)
Vol. 4, No. 3, June 2023, pp. 565-572

DOI: https://doi.org/10.52436/1.jutif.2023.4.3.920
p-ISSN: 2723-3863
e-ISSN: 2723-3871

IMPLEMENTATION OF THE RANDOM FOREST ALGORITHM IN CLASSIFYING THE ACCURACY OF GRADUATION TIME FOR COMPUTER ENGINEERING STUDENTS AT DIAN NUSWANTORO UNIVERSITY

Devi Ayu Rahmawati[1], Nitho Alif Ibadurrahman[*2], Junta Zeniarja[3], Novi Hendriyanto[4]

[1,3,4]Faculty of Computer Science, Informatics Engineering, Dian Nuswantoro University, Indonesia
[2]Faculty of Computer Science, Informatics Engineering, Technische Universität Berlin, Germany
Email: [1]111201912334@mhs.dinus.ac.id, [2]ibadurrahman@campus.tu-berlin.de, [3]junta@dsn.dinus.ac.id, [4]nvhendriyanto@dsn.dinus.ac.id

(Article received: February 15, 2023; Revision: March 21, 2023; published: June 26, 2023)

Abstract

To ensure the existence of a university remains intact, one way that can be done is by optimizing the performance of the students so that they can graduate on time. A high percentage of on-time graduation will result in a good assessment of the accreditation of the department in the university. However, there are many factors that affect the graduation rate, such as the student's academic performance, extracurricular activities, and other factors. The data of graduation of students in the Computer Science program at the Faculty of Computer Science, Dian Nuswantoro University, for the academic years 2008-2017 is the object of this study. The objective of this research is to create the best classification model using the Random Forest algorithm to predict the accuracy of the graduation time of students, which will be useful for policy making in the future. The results of the classification using this algorithm received an accuracy of 93% for the training data and 91% for the test data.

Keywords: *Classification, Graduation, Random Forest, Student.*

1. INTRODUCTION

Maintaining an existence in a university can be done in various ways, one of which is by optimizing the performance of students so that they can graduate from the university within the specified time frame, in other words, students are required to be able to graduate on time. The higher the percentage of students who graduate on time, the better the assessment of the department's accreditation in a university. There are several aspects that affect students in completing their studies, so there needs to be an analysis to produce new knowledge that can be beneficial for the university. Many aspects affect the graduation rate, such as grade point average, student activities, and other aspects [1].

Based on data obtained from the Data and Information Center (DINUS Hospitality) of Dian Nuswantoro University, the Computer Science S1 program has a comparison of the number of students who graduate on time and late from the 2008 to 2017 cohort, with details as shown in the table 1.

Based on Table 1, it is known that the number of students who graduate late is still quite high. Therefore, this issue requires more attention from the academic community to increase the quantity of students who graduate on time in order to maintain the existence of the campus. To maintain or increase the number of students who can graduate on time, a classification system for student graduation using data mining is needed to create an early warning system.

Table 1. Comparison of graduation labels for Computer Science students

No	Year	Number of Students	
		On Time Graduates	Late Graduates
1	2008	-	18
2	2009	-	34
3	2010	-	229
4	2011	228	237
5	2012	196	248
6	2013	190	194
7	2014	235	192
8	2015	249	80
9	2016	219	179
10	2017	299	-
Total		**1616**	**1411**

Source : Data and Information Center, Dian Nuswantoro University

Data mining is a series of processes for extracting and uncovering important patterns in order to obtain new knowledge from a large data warehouse, thus assisting in decision-making. Therefore, data mining is also known as Knowledge Discovery in Database (KDD) [2]. Data mining uses experiences or even mistakes from the past to improve the quality of a model that will be used or analyzed [3]. Data mining is often used to identify and extract useful data and insights from large databases using mathematical, statistical, machine learning, or artificial intelligence methods [4]. There

are various methods in data mining for processing various raw data to obtain knowledge, including Estimation, Prediction, Classification, Clustering, and Association.

Classification is a learning process on training data in order to produce a rule (model). The result of the learning process will be used to classify input attributes to class attributes in test data [5]. The classification method has various algorithms that can be used to help solve a problem. One example of a classification algorithm is Random Forest.

Random Forest is one of the supervised learning algorithms. This algorithm is an implementation of homogeneous Ensemble Learning [6] by combining the output of several decision trees [7] to form a result. Random Forest uses bagging [8] or random feature methods to create uncorrelated forests of Decision Trees. If the Decision Tree considers all possible feature separations, Random Forest only selects a subset of those features randomly each time. The Random Forest algorithm has three main hyperparameters that must be set before testing, which are the number of trees, node size, and the number of features sampled. Therefore, Random Forest is a flexible and easy-to-use algorithm for solving classification and regression problems [9].

A research on classification using the Random Forest algorithm has been conducted by Junta Zeniarja, Abu Salam, and Farda Alan Ma'ruf in 2022. The reputation of a university is influenced, among other things, by the number of students who can graduate on time. In this research, the researchers compared several classification algorithms, such as Decision Tree, Random Forest, K-NN, Naive Bayes, and SVM, to see the effectiveness of these algorithms in creating a predictive model for student graduation. Based on the test results, it was found that the highest accuracy value was obtained by the Random Forest algorithm, which was 77.35% [10].

A research on classification using the Random Forest algorithm was also conducted by Christian Cahyaningtyas, Danny Manongga, and Irwan Sembiring in 2022. The research was conducted by comparing three classification algorithms, namely Naive Bayes, K-NN, and Random Forest, to find the best-performing algorithm and comparing three feature selection methods to find the best method to improve algorithm performance. The results obtained from this research are that the Random Forest algorithm is the best algorithm for classifying harvest data, with an accuracy value of 89.14% using the Backward Elimination method, which can increase accuracy to 96.67% [11].

Based on the research that has been conducted and the description above, research is carried out using the Random Forest algorithm. The purpose of this research is to determine the knowledge generated from the analysis of graduation data for students in the Informatics Engineering program at Dian Nuswantoro University and to determine how to implement the Random Forest algorithm in solving problems related to improving student graduation rates on time.

2. RESEARCH METHOD

The object of this research is the students of the Informatics Engineering program at Dian Nuswantoro University, who graduated between 2008-2017, with a total of 3027 data records. The data was obtained from the university's data and information center, and the researcher aims to classify the time of graduation of the students using data mining techniques. This research is conducted because it is a topic of interest among the academic community and to maintain the accreditation of the university. For this research, the researcher chose the Cross Industry Standard Process for Data Mining (CRISP-DM) as the data mining model [10].

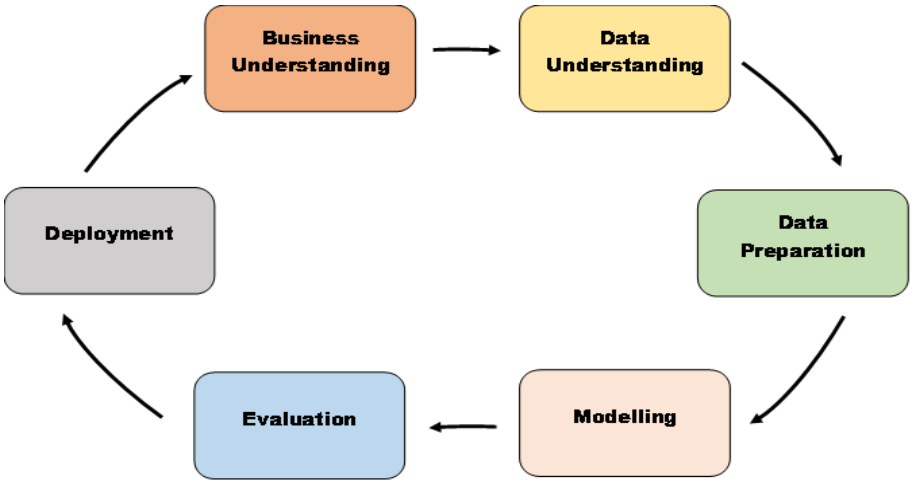


Figure 1. The research flow using CRISP-DM (Cross-Industry Standard Process for Data Mining)

Figure 1 illustrates the research workflow, starting from the business understanding phase to the deployment phase, using the Cross-Industry Standard Process for Data Mining (CRISP-DM) method. By studying previous research related to the classification of student graduation time using the

Random Forest algorithm, the following steps were taken in this study:

1. Business Understanding

Business Understanding The research process begins with business understanding, starting from defining the problem, goals, and solution requirements from a business perspective. The understanding of the research subject is crucial in achieving the desired results. To comprehend the survey object, the researcher gathers information from the data source and delves into it.

2. Data Understanding

Data understanding focuses on the process of understanding data and determining whether the stored data quality is adequate for modeling or not. This phase begins with selecting and identifying datasets to meet business goals. Next, the researcher can identify data quality by performing data cleaning and handling missing values, if any, to discover initial insights from the data.

3. Data Preparation

This phase involves all activities to prepare the final dataset. Data is prepared before implementing it into the model that will be built. Data processing is commonly referred to as the pre-processing stage.

4. Modeling

In this stage, the researcher proposes using the Random Forest algorithm as the modeling technique to solve the existing problem.

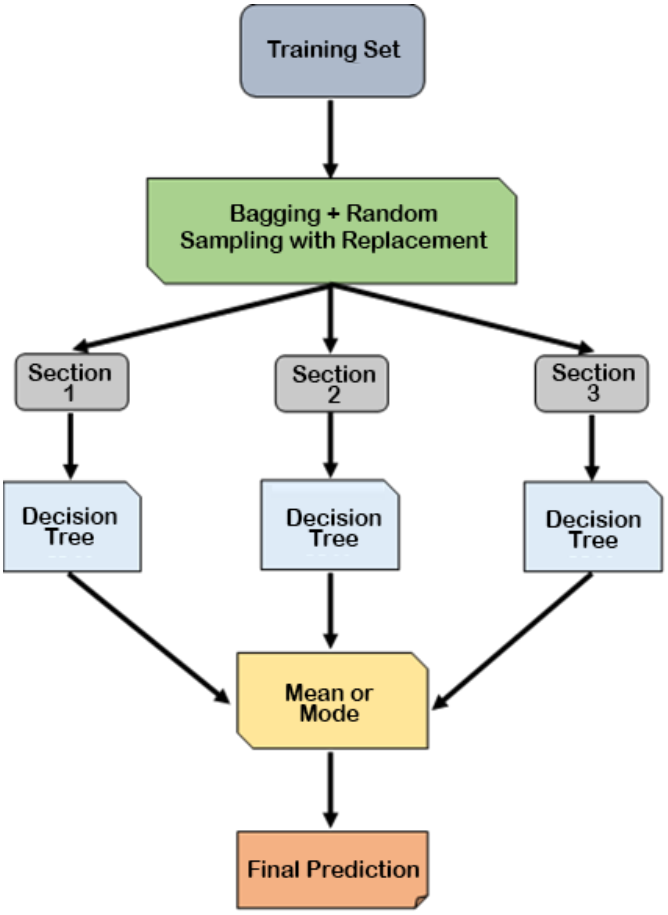


Figure 2. Flowchart of the Random Forest algorithm

Based on Figure 2, the explanation of the steps of the Random Forest algorithm are as follows:

a) Prepare the training data that will be used in the classification process.
b) The existing training data will undergo Bagging and Random Sampling processes to produce new training data that is collected in the bag.
c) For each bag that is generated, a number of randomly selected features will be adopted from its source training set.
d) Each decision tree model will be trained by each different bag, resulting in diverse trained models.
e) Each trained model that is generated will be used to predict a set of new feature values.
f) The predictions of each model that is generated will be combined through a majority voting process to produce the final prediction value.

5. Evaluation

This step involves evaluating the results of the modelling conducted by the researcher using the Confusion Matrix to analyze the performance of the modelling process. The Confusion Matrix is a parameter for measuring performance in classification problems where the output generated can be in the form of two or more classes [12].

6. Deployment

The results of this research provide an analysis that leads to decision-making and knowledge about patterns of students in classifying graduation time using the Random Forest algorithm. After obtaining a well-accurate modelling result with optimal support, it is expected that this can solve the existing problems and can be applied to the academic system in the future.

3. RESULTS AND DISCUSSION

3.1. Business Understanding

The business understanding in this research focuses on the process of classifying students' graduation time as on-time or delayed. The classification criteria is based on the duration of study, which should not exceed 4 years or 8 semesters. The purpose of data mining or the aim of this research is to extract knowledge from the available data in order to understand the patterns of student study that lead to on-time or delayed graduation. The researcher will build an appropriate model to address this issue, which will help academic staff in taking preventive measures in the future through an early warning system.

3.2. Data Understanding

The process carried out in this phase is the acquisition of raw data which is carried out in accordance with the research theme with the required attributes. Figure 3 shows the data for students in the Informatics Engineering program at Dian Nuswantoro University that the researcher obtained from the university's data and information center. The raw data obtained by the researcher consists of 3027 data records for students from the 2008-2017 academic years.

3.3. *Data Preparation*

The data preparation process was carried out using two tools, namely Microsoft Excel and Google Colab/Jupyter Notebook. The researcher used Microsoft Excel to create a labeling of graduation time using the IF ELSE formula, resulting in a data group with labels of on-time, late, and not fulfilled. The data to be used for the next process are only the on-time and late labeled data.

Next, the researcher used Google Colab/Jupyter Notebook to perform further data cleaning. The obtained data has several issues, especially with incomplete data. Therefore, data cleaning is necessary to fix problematic data to avoid interfering with the subsequent research process.

id	gender	date_of_birth	age	pathways	entry_year	ips1	ips2	ips3	ips4	ips
3930	L	09/09/1990	18	Reguler	01/01/2008	2,45	2,21	2,75	2,43	3,(
3974	L	27/03/1990	18	Reguler	01/01/2008	0,47	2,11	1,70	2,39	1,5
4008	L	01/03/1988	20	Reguler	01/01/2008	2,09	2,59	3,00	2,87	3,2
4050	L	14/11/1988	20	Reguler	01/01/2008	2,16	2,37	2,42	2,75	2,(
4058	L	06/05/1989	19	Reguler	01/01/2008	2,24	3,00	3,13	3,05	2,(
4139	L	26/12/1989	19	Reguler	01/01/2008	0,63	2,11	2,42	2,25	1,2
4261	L	27/12/1989	19	Reguler	01/01/2008	3,50	3,13	3,29	2,17	2,(
4283	L	07/08/1990	18	Reguler	01/01/2008	1,64	1,61	2,11	1,55	2,1
4295	L	04/08/1990	18	Reguler	01/01/2008	1,11	2,37	2,56	2,21	2,4
4355	L	28/07/1990	18	Reguler	01/01/2008	0,17	0,83	0,94	1,50	1,7
4367	L	26/09/1990	18	Reguler	01/01/2008	2,91	2,92	1,86	3,33	2,3
4413	L	17/03/1989	19	Reguler	01/01/2008	0,62	0,50	0,00	1,23	1,9
4441	L	15/02/1990	18	Reguler	01/01/2008	1,89	2,55	2,76	2,41	2,8

Figure 3. Example of raw student data table (truncated attributes)

3.4. Modelling

1. Dataset Preparation

Dataset Preparation The dataset consists of 3027 data with 15 independent variables (X) and 1 dependent variable (y) as the class label, which is the graduation label attribute. The label consists of 2 classes, namely on-time (0) and late (1). Each class obtained a percentage of 53.39% for the on-time class label and 46.61% for the late class label. Then, the data is randomly split into two with 85% for training data and 15% for testing data. From this split, it obtained 2572 training data with a percentage of 1373 (53.38%) on-time label data and 1199 (46.62%) late label data. Meanwhile, the testing data consists of 455 data with a percentage of 243 (53.41%) on-time label data and 212 (46.59%) late label data.

2. Data Sampling

The training data appears to have imbalanced classes, so it needs to undergo data sampling. In this study, the researcher performed data sampling by combining two methods, namely SMOTE and TomekLinks. The parameter used by the SMOTETomek method is the majority on the sampling_strategy.

3. Classification with Random

Forest The application of the SMOTETomek method and Random Forest algorithm will be linked through the pipeline. The parameters used in the Random Forest algorithm are max_depth and random_state.

3.5. Evaluation

Based on the experimental results that the researcher conducted using the Random Forest classification algorithm, it was found that the model can differentiate data into classification classes. Overall, the model evaluation results are already very good based on the high F1 Score value in each class, indicating that the classification results are balanced between the two classes.

Accuracy Random Forest: 0.9310202395704255

	precision	recall	f1-score	support
LATE	0.94	0.91	0.93	1129
ON TIME	0.93	0.95	0.94	1292
accuracy			0.93	2421
macro avg	0.93	0.93	0.93	2421
weighted avg	0.93	0.93	0.93	2421

Figure 4. The result of the classification_report on the training set

Figure 4 shows that in the training data, a precision value of 93% for the on-time class and 94% for the late class, a recall value of 95% for the on-time class and 91% for the late class, as well as an f-1 score value of 94% for the on-time class and 93% for the late class. Therefore, the final accuracy result obtained was 93% with detailed confusion matrix values shown in Figure 5.

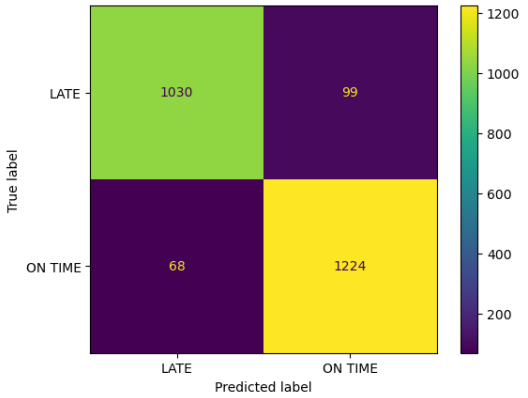


Figure 5. The result of the confusion_matrix on the training set

Accuracy Random Forest: 0.9174917491749175

	precision	recall	f1-score	support
LATE	0.92	0.90	0.91	282
ON TIME	0.92	0.93	0.92	324
accuracy			0.92	606
macro avg	0.92	0.92	0.92	606
weighted avg	0.92	0.92	0.92	606

Figure 6. The result of the classification_report on the testing set

Figure 6 shows that in the testing data, a precision value of 92% was obtained for both class, a recall value of 93% for the on-time class and 90% for the late class, as well as an f-1 score value of 92% for the on-time class and 91% for the late class. Therefore, the final accuracy result obtained was 92% with detailed confusion matrix values shown in Figure 7.

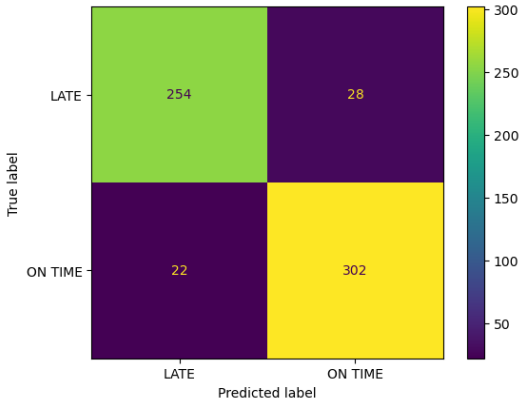
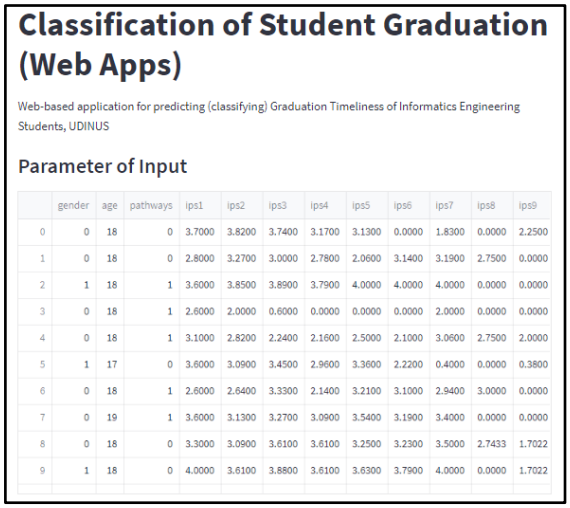


Figure 7. The result of the confusion_matrix on the testing set.

3.6. Deployment

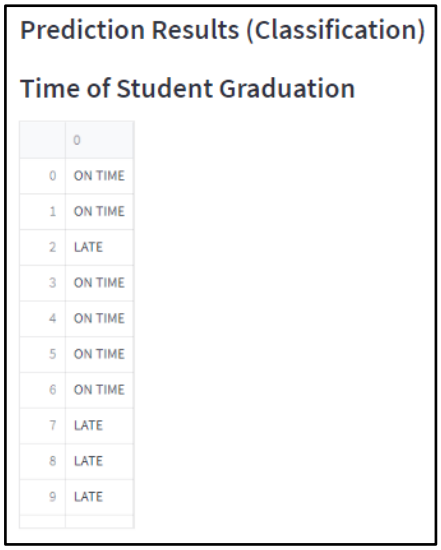
After obtaining the best modeling results with optimal accuracy and supporting values, the next step is to apply the model to a web-based application using the Streamlit framework. Streamlit is a Graphical User Interface (GUI) that can be used to deploy a model. Streamlit is a web framework used to easily distribute models and visualizations using the Python language [13]. The website interface can be seen in Figure 8, Figure 9, and Figure 10.

Figure 8. Display of input form on the Streamlit website



	gender	age	pathways	ips1	ips2	ips3	ips4	ips5	ips6	ips7	ips8	ips9
0	0	18	0	3.7000	3.8200	3.7400	3.1700	3.1300	0.0000	1.8300	0.0000	2.2500
1	0	18	0	2.8000	3.2700	3.0000	2.7800	2.0600	3.1400	3.1900	2.7500	0.0000
2	1	18	1	3.6000	3.8500	3.8900	3.7900	4.0000	4.0000	4.0000	0.0000	0.0000
3	0	18	1	2.6000	2.0000	0.6000	0.0000	0.0000	0.0000	2.0000	0.0000	0.0000
4	0	18	1	3.1000	2.8200	2.2400	2.1600	2.5000	2.1000	3.0600	2.7500	2.0000
5	1	17	0	3.6000	3.0900	3.4500	2.9600	3.3600	2.2200	0.4000	0.0000	0.3800
6	0	18	1	2.6000	2.6400	3.3300	2.1400	3.2100	3.1000	2.9400	3.0000	0.0000
7	0	19	1	3.6000	3.1300	3.2700	3.0900	3.5400	3.1900	3.4000	0.0000	0.0000
8	0	18	0	3.3000	3.0900	3.6100	3.6100	3.2500	3.2300	3.5000	2.7433	1.7022
9	1	18	0	4.0000	3.6100	3.8800	3.6100	3.6300	3.7900	4.0000	0.0000	1.7022

Figure 9. Display of input parameters



	0
0	ON TIME
1	ON TIME
2	LATE
3	ON TIME
4	ON TIME
5	ON TIME
6	ON TIME
7	LATE
8	LATE
9	LATE

Figure 10. Display of prediction results.

4. DISCUSSION

In this study, to broaden the perspective and to become a comparative factor, the researcher used several journals that discussed similar research. The selected journals had a correlation with the Random Forest algorithm. The researcher compared the evaluation stage in several previous studies to determine the comparison of the accuracy of the Random Forest algorithm with the benchmark algorithm.

The first study is a journal written by [14] entitled "Comparison of Machine Learning Methods in Diagnosing Depression Mental Disorders." Depression is one of the fatal diseases that can cause the risk of suicide. The purpose of this research is to assist in diagnosing symptoms of depression so that treatment can be given as soon as possible and to prevent the risk of suicide. In this study, the researcher compared several Machine Learning algorithms, including K-NN, Naïve Bayes, Random Forest, and Decision Tree to create an independent predictive model for diagnosing depression. The results of testing the four algorithms showed that the Random Forest algorithm showed the best accuracy results of 80.02%, which can still be improved by utilizing optimization of the algorithm.

The second study is a journal written by [15] entitled "Prediction of Student Academic Performance using Random Forest and C4.5 Algorithms." Student academic performance is one of the benchmarks for the success of a study program at a university. Therefore, this research was conducted to predict academic performance of students. The researcher compared several classification algorithms, namely Random Forest and C4.5, to see their effectiveness. The results of the study showed that the model using the Random Forest algorithm had better performance than the C4.5 algorithm with an accuracy value of 92.4%, precision of 91.4%, and recall of 92.4%.

In this research, in order to broaden the perspective and to become a comparative factor, the researcher used several literature in the form of journals that discuss similar research. The selected journals are those related to the Random Forest algorithm. The researcher compared the evaluation stages in several previous studies to determine the comparison of the accuracy of the Random Forest algorithm with the comparison algorithm.

The first study is a journal by [14] titled "Comparison of Machine Learning Methods in Diagnosing Depressive Disorders". Depression is a fatal disease that can lead to suicidal risk. The purpose of this research is to help diagnose depression symptoms so that treatment can be done as early as possible and to prevent the risk of suicide. In this study, the researcher compared several Machine Learning algorithms, including K-NN, Naïve Bayes, Random Forest, and Decision Tree to create an independent prediction model for diagnosing depression. The results of testing these four algorithms showed that the Random Forest algorithm showed the best accuracy of 80.02%, which can still be improved by optimizing the algorithm.

The second study is a journal by [15] titled "Prediction of Student Academic Performance using Random Forest and C4.5 Algorithm". The academic achievement of students is one of the benchmarks for the success of running a study program at a university. Therefore, this research was conducted to predict the academic performance of students. The researcher compared several classification algorithms, namely Random Forest and C4.5, to see their effectiveness. The results showed that the model using the Random Forest algorithm performed better than the C4.5 algorithm with an accuracy value of 92.4%, precision of 91.4%, and recall of 92.4%.

The third study is a journal by [16] titled "Comparison of Classification Models for Student Academic Performance Evaluation". In order to achieve quality education, universities need to evaluate both new student admissions and student

graduation. In this study, the researcher conducted research to determine the appropriate classification model for predicting student graduation by comparing several algorithms, including Decision Tree, Random Forest, Naïve Bayes, Support Vector Machine, Neural Network, Gradient Boosted Trees, Logistic Regression, K-Nearest Neighbor, and Auto Multilayer Perceptron. Based on the experiments conducted, the Random Forest algorithm had the best performance compared to other algorithms with an average value of 83.8%.

The fourth study is a journal by [2] titled "Comparison of Performance of C4.5, Random Forest, and Gradient Boosting Algorithms for Commodity Classification". Commodities in a region have their own advantages, one of which is in Riau Province with commodities in the plantation sector which is very promising. This commodity data can be processed to obtain classification patterns. In this study, the researcher compared the effectiveness of algorithms. The comparison results showed that the Random Forest algorithm using shuffle sampling technique had the best performance with an accuracy value of 97.04%.

The fifth study is a journal article by [12] titled "Comparison of K-Nearest Neighbor and Random Forest Methods in Predicting the Accuracy of Classification of Wart Disease Treatment". Warts are skin diseases caused by the Human papillomavirus (HPV). Researchers compared the performance of the K-NN and Random Forest classification algorithms in predicting and diagnosing wart disease. From the test results, it was found that the best accuracy rate was achieved by the Random Forest algorithm, which was 86.56%.

The sixth study is a journal article by [8] titled "Comparison of Classification Methods and Analysis of Academic Factors for Student Graduation Patterns in Higher Education". In this study, the researchers compared the Decision Tree and Random Forest algorithms to determine the factors that influence graduation and analyze student academic patterns. Based on the test results, the Random Forest algorithm with cross-validation and hyperparameter tuning (Grid Search CV) had better accuracy results than the Decision Tree algorithm with an accuracy rate of 96.1%.

The seventh study is a journal article by [10] titled "Feature Selection and Comparison of Classification Algorithms for Predicting Student Graduation". The reputation of a university is influenced by, among other things, the number of students who can graduate on time. In this study, the researchers compared several classification algorithms, such as Decision Tree, Random Forest, K-NN, Naïve Bayes, and SVM to see the effectiveness of these algorithms in creating a model for predicting student graduation. Based on the test results, the Random Forest algorithm had the highest accuracy rate of 77.35%.

The eighth study is a journal article by [17] titled "Comparison of Decision Tree, Random Forest, SVM, and K-NN Algorithms in the Classification of Airline Passenger Satisfaction". Competition among airlines is increasing. Customer satisfaction can be used as an indicator to evaluate the quality of airline services. In this study, the researchers compared the Decision Tree, Random Forest, SVM, and K-NN algorithms to determine the patterns of customer satisfaction in using an airline. From this study, it was found that the highest accuracy rate was achieved by the Random Forest algorithm, which was 96%

The latest study is a journal article by [11] titled "Comparison Analysis of Algorithms and Feature Selection for Broiler Chicken Harvest Classification". This study compared three classification algorithms, namely Naïve Bayes, K-NN, and Random Forest, to find the best-performing algorithm and compared three feature selection methods to find the best method to improve algorithm performance. The result of this study showed that the Random Forest algorithm was the best algorithm for classifying broiler chicken harvest data, with an accuracy rate of 89.14% using the Backward Elimination method, which could increase the accuracy rate to 96.67%.

5. CONCLUSION AND SUGGESTIONS

Based on the research conducted by applying the Random Forest algorithm for the classification of on-time graduation of students, both in training and testing data, it can be concluded that the Random Forest algorithm is good enough for classification in the case of on-time graduation of students. This is evidenced by the testing conducted on 2,572 training data and 455 testing data with an accuracy rate of 93% in training data and 91% in testing data, which proves that the Random Forest algorithm can be used to build a good model that can solve existing problems. Therefore, further research can be developed by using other machine learning algorithms or deep learning algorithms and optimization to improve the accuracy of the model to be built.

REFERENCES

[1] E. Purnamasari, D. P. Rini, and Sukemi, "The Combination of Naive Bayes and Particle Swarm Optimization Methods of Student's Graduation Prediction," *Jurnal Ilmiah Teknik Elektro Komputer dan Informatika*, vol. 5, no. 2, pp. 112–119, Feb. 2019, doi: 10.26555/jiteki.v5i2.15317.

[2] E. Ismanto and M. Novalia, "Komparasi Kinerja Algoritma C4.5, Random Forest, dan Gradient Boosting untuk Klasifikasi Komoditas," 2021.

[3] E. S. Susanto and H. Al Fatta, "Prediksi Kelulusan Mahasiswa Magister Teknik Informatika Universitas Amikom Yogyakarta Menggunakan Metode K-Nearest Neighbor,"

2018.

[4] L. O. M. Zulfiqar, N. Renaningtias, and M. Y. Fathoni, "Educational Data Mining in Graduation Rate and Grade Predictions Utilizing Hybrid Decision Tree and Naïve Bayes Classifier," Jul. 2020, pp. 151–157. doi: 10.5220/0009907101510157.

[5] B. A. Arifiyani and R. S. Samosir, "Sistem Simulasi Prediksi Profil Kelulusan Mahasiswa Dengan Decison Tree," *Kalbiscentia*, vol. 5, no. 2, 2018.

[6] I. D. Mienye, Y. Sun, and Z. Wang, "An improved ensemble learning approach for the prediction of heart disease risk," *Inform Med Unlocked*, vol. 20, Jan. 2020, doi: 10.1016/j.imu.2020.100402.

[7] A. Yanuar, "Random Forest," *Universitas Gadjah Mada Menara Ilmu Machine Learning*, Jul. 28, 2018. https://machinelearning.mipa.ugm.ac.id/2018/07/28/random-forest/ (accessed Oct. 23, 2022).

[8] R. Aprillya, Perbandingan, M. Klasifikasi, R. A. Putri, and N. S. Fatonah, "Perbandingan Metode Klasifikasi serta Analisis Faktor Akademis Pola Kelulusan Mahasiswa di Perguruan Tinggi," vol. 7, no. 2, 2022.

[9] P. Probst, M. N. Wright, and A. L. Boulesteix, "Hyperparameters and tuning strategies for random forest," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 9, no. 3. Wiley-Blackwell, May 01, 2019. doi: 10.1002/widm.1301.

[10] J. Zeniarja, A. Salam, and F. A. Ma'ruf, "Seleksi Fitur dan Perbandingan Algoritma Klasifikasi untuk Prediksi Kelulusan Mahasiswa," *Jurnal Rekayasa Elektrika*, vol. 18, no. 2, Jul. 2022, doi: 10.17529/jre.v18i2.24047.

[11] C. Cahyaningtyas, D. Manongga, and I. Sembiring, "Algorithm Comparison And Feature Selection for Classification of Broiler Chicken Harvest," *Jurnal Teknik Informatika (Jutif)*, vol. 3, no. 6, pp. 1717–1727, Dec. 2022, doi: 10.20884/1.jutif.2022.3.6.493.

[12] U. Erdiansyah, A. Irmansyah Lubis, and K. Erwansyah, "Komparasi Metode K-Nearest Neighbor dan Random Forest Dalam Prediksi Akurasi Klasifikasi Pengobatan Penyakit Kutil," *Jurnal Media Informatika Budidarma*, vol. 6, no. 1, p. 208, Jan. 2022, doi: 10.30865/mib.v6i1.3373.

[13] W. Hastomo, N. Aini, A. Satyo, B. Karno, and L. M. R. Rere, "Metode Pembelajaran Mesin untuk Memprediksi Emisi Manure Management," 2022.

[14] R. Wajhillah, S. Bahri, and A. Wibowo, "Komparasi Metode Machine Learning pada Diagnosa Gangguan Kejiwaan Depresi," 2020. [Online]. Available: https://www.kaggle.com/everseek/depression

[15] S. Linawati, Siti Nurdiani, Kartika Handayani, and Latifah, "Prediksi Prestasi Akademik Mahasiswa Menggunakan Algoritma Random Forest Dan C4.5," *Jurnal Khatulistiwa Informatika*, vol. VIII, no. 1, 2020, [Online]. Available: www.bsi.ac.id

[16] R. Rachmatika and Achmad Bisri, "Perbandingan Model Klasifikasi untuk Evaluasi Kinerja Akademik Mahasiswa," *JEPIN (Jurnal Edukasi dan Penelitian Informatika)* , vol. 6, no. 3, 2020.

[17] M. H. Setiono, "Komparasi Algoritma Decision Tree, Random Forest, Svm Dan K-Nn Dalam Klasifikasi Kepuasan Penumpang Maskapai Penerbangan," *INTI Nusa Mandiri*, vol. 17, no. 1, pp. 32–39, Oct. 2022, doi: 10.33480/inti.v17i1.3420.