Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 2, April 2024
DOI : https://doi.org/10.33395/sinkron.v8i2.13658

e-ISSN : 2541-2019
p-ISSN : 2541-044X

Implementation of Exploratory Data Analysis and Artificial Neural Networks to Predict Student Graduation on-Time

Sonia Sri Muliani[1)*], Volvo Sihombing[2)], Ibnu Rasyid Munthe[3)]
[1,2,3)]Universitas Labuhanbatu, Indonesia
[1)]soniasrimuliani969@gmail.com, [2)]volvolumbantoruan@gmail.com, [3)]ibnurasyidmunthe@gmail.com

Abstract: Almost all universities in Indonesia face the problem of a low number of students graduating on time. This will affect higher education accreditation. For this reason, universities must pay attention to the timely graduation of their students. The way that can be taken is to predict students' graduation on time. This research aims to predict students' timely graduations using a combination of exploratory data analysis and artificial neural networks. Exploratory data analysis is used to study the relationship between features that influence students' on-time graduation, while artificial neural networks are used to predict on-time graduation. This research goes through method stages, starting with determining the dataset, exploratory data analysis, data preprocessing, dividing training and test data, and applying artificial neural networks. From the research, it was found that Work features and GPS features greatly influence graduation on time. Students who study while working are less likely to graduate on time compared to students who do not work. Students who have an average GPS above 3.00 for eight consecutive semesters will find it easier to graduate on time. Meanwhile, Age and Gender features have no effect on graduating on time. With a percentage of 50% training data and 50% test data, epoch 100, and learning rate 0.001, the best network model was obtained to predict graduation on time with an accuracy rate of 69.84%. The research results also show that the amount of test data and the learning rate can influence the level of accuracy. Meanwhile, the number of epochs does not affect the level of accuracy.

Keywords: Artificial Neural Network; Exploratory Data Analysis; Graduation; Machine Learning; Prediction.

INTRODUCTION

The low number of students graduating on time is a problem faced by almost all universities in Indonesia (Harun et al., 2022). This can be seen from the unequal percentage of students graduating from year to year or the comparison of the number of students who have graduated and who have not graduated in several classes (Priyatman, Sajid, & Haldivany, 2019). With longer graduation times, students face greater financial responsibilities, while universities have limited resources (Liao et al., 2019). Universities must pay attention to the graduation rate of their students in an effort to improve their reputation in the eyes of society and increase their accreditation in the eyes of the government. In an effort to maintain or improve higher education accreditation, one of the assessment indicators used is the student graduation rate (Qisthiano, Kurniawan, Negara, & Akbar, 2021). This assessment is contained in the BAN-PT Study Program Accreditation Instrument 4.0 document on the Tridharma Outcomes and Achievements element (Tinggi, 2019). In this document, there is an assessment matrix for the self-evaluation report and study program performance report, which takes into account the average length of study and students' on-time graduation.

Increasing the quantity of students graduating on time is important for the sustainability of a university because a high graduation rate will determine the level of the university in the world of education (Suhaimi, Abdul-Rahman, Mutalib, Hamid, & Malik, 2019). For this reason, efforts to minimize the low graduation rate need to be made by universities; one way is by predicting student graduation (Apridiansyah, Veronika, & Putra, 2021; Dengen, Kusrini, & Luthfi, 2020; Masrizal & Hadiansa, 2017; Putri & Waspada, 2018). Predicting graduation allows students to make informed decisions about their academic and career paths. For higher education institutions, these efforts are useful for identifying students who may not graduate and providing appropriate support to ensure their success (Pelima, Sukmana, & Rosmansyah, 2024). The benefit of graduation prediction is that the results can be used by universities to find useful patterns from large graduation data (Kartarina, Sriwinarti, & Juniarti, 2021; Qisthiano et al., 2021), which are used to predict future graduation outcomes (Basheer, Mutalib, Hamid, Abdul-

* Corresponding author

Rahman, & Malik, 2019). By making predictions, the imbalance between student acceptance and graduation can be overcome, so that students who may not graduate on time can be identified. The results of the predictions can be used by universities to create and implement appropriate policies for remediation and retention (Lagman et al., 2020).

In the field of education, machine learning techniques have been widely implemented in assessing student academic performance (Alyahyan & Düştegör, 2020; Pelima et al., 2024), both for classification and prediction purposes (Figueroa-Cañas & Sancho-Vinuesa, 2020). Machine learning techniques help collect important data, which makes student performance prediction models significant (Tampakas, Livieris, Pintelas, Karacapilidis, & Pintelas, 2019). Meanwhile, to predict student graduation, researchers have also applied various machine learning algorithms previously. By applying the Naïve Bayes algorithm, which uses a dataset of alumni from several universities in Palembang as test data, the results obtained show that the accuracy level of the algorithm is 81% (Qisthiano et al., 2021). Other algorithms, such as support vector machines, were also applied, and the results obtained were 90% accuracy (Bangun, Mawengkang, & Efendi, 2022). Application of the K-Nearest Neighbors algorithm with the highest accuracy rate of 98.5% (Muliono, Lubis, & Khairina, 2020). Not only that, a comparative analysis between algorithms for predicting student graduation was also carried out. By comparing the Naïve Bayes and K-Nearest Neighbors algorithms, this research shows that the accuracy level of the two algorithms is the same, namely 90%. However, the K-Nearest Neighbors algorithm is considered better because it requires fewer processes than Naïve Bayes (Gunawan, Zarlis, & Roslina, 2021). A comparison between Naïve Bayes and artificial neural networks in predicting the study period of undergraduate students was also carried out. According to the results of this research, artificial neural networks are superior in prediction accuracy, namely 78.58% (Azahari, Yulindawati, Rosita, & Mallala, 2020). Other research has compared and evaluated three algorithms, namely, decision trees, artificial neural networks, and support vector machines. The results of this research show that the accuracy level of the three algorithms is more than 80% (Riyanto, Hamid, & Ridwansyah, 2019). The results of this research show that the artificial neural network algorithm is a very popular model because it can be implemented using non-linear data and can accommodate large amounts of data. This is in line with what was produced by research (Rodríguez-Hernández, Musso, Kyndt, & Cascallar, 2021): the ability of artificial neural networks to produce predictive models can utilize the interaction of all predictor variables to estimate better outcome variables. Artificial neural networks have the ability to handle nonlinear relationships between independent and dependent variables in very large data analyses. In addition, artificial neural networks are able to produce target output that is more similar to the actual output (Bukhari et al., 2022).

Based on the advantages of artificial neural networks as previously described, several research results are also described that apply the artificial neural network algorithm to predicting student graduation. Research conducted by (Yaqin, Laksito, & Fatonah, 2021) in predicting early student graduation resulted in an accuracy of 77% at a learning rate of 0.01. This research only used the grade point average (GPA) feature from semester 1 to semester 4. In research conducted by (Ridwan, Lubis, & Kustanto, 2020) and (Fiqha, Yandris, & Nasution, 2022), the same features were used, namely: student identification number, grade point from semester 1 to semester 4, semester 1 to 4 course credits, average GPA, and total course credits. The same results were obtained, namely, a prediction accuracy of 98.27%. However, what differentiates the two studies lies in the parameter settings for the number of epochs, learning rate, and momentum.

The studies described previously did not explain how the features used in the dataset were obtained. It does not explain how the relationship between features affects students' on-time graduation. Meanwhile, the benefit of graduation prediction is to find patterns in large amounts of graduation data (Kartarina et al., 2021; Qisthiano et al., 2021). For this reason, before carrying out the prediction process, further analysis is needed on how the correlation between the features used affects students' on-time graduation. For this reason, this research is different from previous studies, where exploratory data analysis was used to obtain useful information by finding relationships between features in the dataset (Aldera, Emam, Al-Qurishi, Alrubaian, & Alothaim, 2021; Indrakumari, Poongodi, & Jena, 2020). Exploratory data analysis is used to get a better picture of a data set, find anomalies and outliers, and test basic assumptions (Kulkarni & Shivananda, 2019). Exploratory data analysis is a method of summarizing data by taking its main characteristics and displaying them with appropriate representations. Exploratory data analysis concentrates more on examining the assumptions required for hypothesis testing and model fitting, as well as dealing with missing values and changing features as needed (Sahoo, Samal, Pramanik, & Pani, 2019). From this perspective, exploratory data analysis is very useful for studying hidden structures, covering important data attributes sequentially, extracting patterns, and testing hypotheses (Abukmeil, Ferrari, Genovese, Piuri, & Scotti, 2021).

Based on the explanation above, this research aims to apply exploratory data analysis and artificial neural networks to predict students' timely graduation. An exploratory data analysis approach is used to look at the features in the dataset that influence students' timely graduation before the prediction process is carried out. Highly correlated features will be retained, while uncorrelated features will be discarded. The results of selecting highly correlated features will be used in the prediction process. Meanwhile, neural networks are used to predict students'

* Corresponding author

timely graduation by applying various models of different test data variations, epochs, and learning rates. The results of predictions are measured by the level of accuracy. So you can see which model has a high level of accuracy. It is hoped that this research can contribute to overcoming the problem of a lack of students graduating on time, so that it becomes a solution for university administrators to consider in order to reduce the number of students who graduate late.

METHOD

This section will outline the proposed research methodology. This chapter provides an explanation of the dataset used for research and its sources. This chapter also discusses the process of exploratory data analysis and data pre-processing methods, which help clean and prepare data for analysis. This chapter also discusses the division of training data and test data and the implementation of artificial neural networks. Overall, the methodology chapter provides a clear and thorough overview of the research design and steps, allowing the research to be repeatable and valid. Python is the programming language used in this research, starting from the exploratory data analysis stage to the application of artificial neural networks. Google Colab is a text editor used to write and run Python program code. The following are the research stages of the process of predicting students' on-time graduation, which are shown in Figure 1.

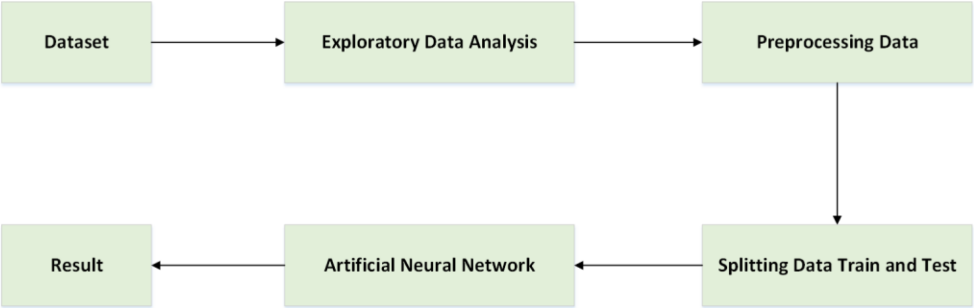

Dataset
Exploratory Data Analysis
Preprocessing Data
Splitting Data Train and Test
Artificial Neural Network
Result

Fig 1. The Research Stages

The dataset used in the research is public secondary data. The dataset was sourced from Kaggle (Athallah, 2023). This dataset is used to predict the possibility of whether a student will graduate on time or late. In this research, the exploratory data analysis process is divided into three structured parts: data identification, univariate analysis, bivariate analysis, and multivariate analysis. The data identification process in this research aims to: understand the data features used; see the data type for each feature; and observe the number of missing values for each feature. The univariate analysis process aims to analyze the distribution of data for each feature used so that the distribution and statistical summary can be understood. Meanwhile, the bivariate and multivariate analysis processes aim to see the relationship between two features (bivariate) and three or more features (multivariate). Each part of this exploratory data analysis helps to gain an in-depth understanding of the characteristics of students' exact graduation data, which is an important basis for the next prediction step that applies artificial neural networks.

Data preprocessing in this research includes the processes of handling missing values, feature coding, data normalization, and feature selection. In the process of handling missing values, missing values will be filled with the value 0 or values that appear frequently. The feature coding process uses two methods, namely, one-hot encoding and label encoding. One-hot encoding is used for categorical data types that only consist of two options. Meanwhile, label encoding is used for categorical data types that consist of three or more choices. The data normalization process in this research was carried out by changing the data into a range of 0 and 1. This needs to be done because there are significant differences in the value range between different features. Feature selection is carried out to select features that have a high correlation with the target variable, in this case, the graduation feature. Features that have a high correlation with the target will be retained. Meanwhile, features with low correlation values will be discarded.

In this research, five scenarios for dividing training data and test data were implemented, namely: 90% training data and 10% test data; 80% training data and 20% test data; 70% training data and 30% test data; 60% training data and 40% test data; and 50% training data and 50% test data. This was done to see the extent of the influence of data sharing on network accuracy. The artificial neural network implementation model that is built consists of 8 input layers, 10 hidden layers, 10 output layers, and 2 classification nodes for prediction results. The number of epochs consists of 100 and 500 iterations, and the learning rate consists of 0.001, 0.01, 0.1, and 0.5. The test results will show a comparison of the accuracy levels of five scenarios for dividing training data and test data by applying various numbers of epochs and learning rates. Next, we will see which artificial neural network model has a higher level of accuracy.

* Corresponding author

RESULT

This section presents the results of applying exploratory data analysis and artificial neural networks to predicting students' on-time graduation. The research results are described systematically based on the research steps described in the previous methodology section.

Table 1. The Graduation Dataset

id	gender	work	age	gps1	gps2	gps3	gps4	gps5	gps6	gps7	gps8	gpa	graduation
1	Female	Working	25	2,76	2,8	3,2	3,17	2,98	3	3,03	0	3,07	Late
2	Female	not-Working	24	3	3,3	3,14	3,14	2,84	3,13	3,25	0	3,17	Late
3	Female	Working	26	3,5	3,3	3,7	3,29	3,53	3,72	3,73	0	3,54	Late
4	Female	not-Working	23	3,17	3,41	3,61	3,36	3,48	3,63	3,46	0	3,41	Late
5	Female	Working	26	2,9	2,89	3,3	2,85	2,98	3	3,08	0	3,09	Late
6	Male	Working	23	2,95	2,82	3,09	3,1	2,78	3,16	3,23	0	3,07	Late
7	Female	not-Working	22	2,76	3,14	2,6	2,95	3,23	3,33	3,3	3,3	3,06	on-Time
8	Female	not-Working	23	2,62	2,89	2,32	2,5	2,5	2,86	3,05	2,5	2,91	on-Time
9	Female	Working	22	3,6	3,54	3,52	3,39	3,52	3,68	3,15	0	3,4	Late
10	Female	Working	24	2,71	2,55	1,77	2,11	1,93	2,13	1,78	0,2	2,2	Late

The dataset in Table 1 is a sample dataset consisting of 379 rows and 14 features. These features are gender, work, age, GPS 1 to GPS 8, GPA, and graduation. ID is a unique identifier for each student. Gender consists of two categories, namely "male" and "female.". Work is a category of students who study while "working" or "not working." Age is the age of each student. GPS 1 to 8 is the student's semester grade point, starting from semester 1 to semester 8. GPA is the average grade point the student obtained after graduating. Graduation is a category of whether students graduate "on time" or "late".

Table 2. Data Identification

Features	Data Type	Type of Data	Total of missing	Percentage of missing
id	int64	Numeric	0	0,0%
gender	object	Categorical	0	0,0%
work	object	Categorical	0	0,0%
age	int64	Numeric	0	0,0%
gps 1	float64	Numeric	0	0,0%
gps 2	float64	Numeric	0	0,0%
gps 3	float64	Numeric	0	0,0%
gps 4	float64	Numeric	0	0,0%
gps 5	float64	Numeric	0	0,0%
gps 6	float64	Numeric	0	0,0%
gps 7	float64	Numeric	0	0,0%
gps 8	float64	Numeric	7	1,85%
gpa	float64	Numeric	3	0,79%
graduation	object	Categorical	0	0,0%

Table 2 shows the results of the data identification process. Data identification is the initial process of exploratory data analysis. This table shows that there are 14 data features consisting of two types of data, namely, category data and numerical data. Category data consists of: gender, work, and graduation, which are object data types. Meanwhile, numeric data consists of: id and age, which are of the int64 data type; as well as gps 1, gps 2, gps 3, gps 4, gps 5, gps 6, gps 7, gps 8, and gpa, which are of the of the float 64 data type. From this table, also shown is the total missing data, namely, 7 data points for the GPS 8 feature with a percentage of 1.85% and 3 missing data points for the GPA feature with a percentage of 0.79%. Missing data is data that contains nothing at all.

* Corresponding author

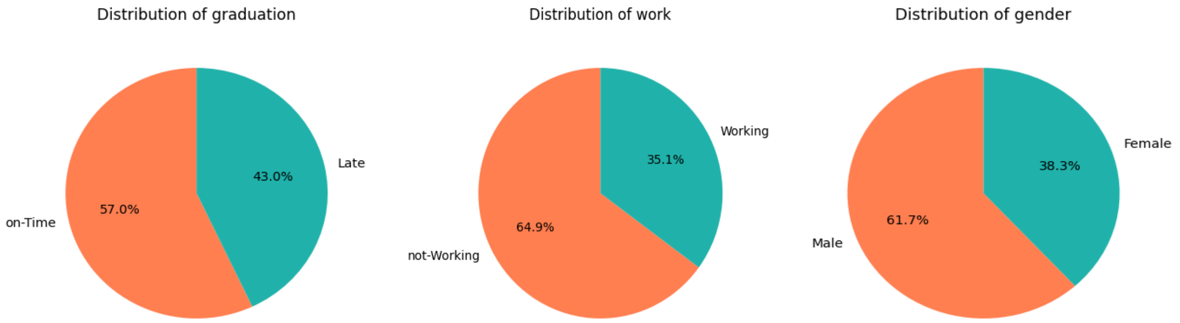
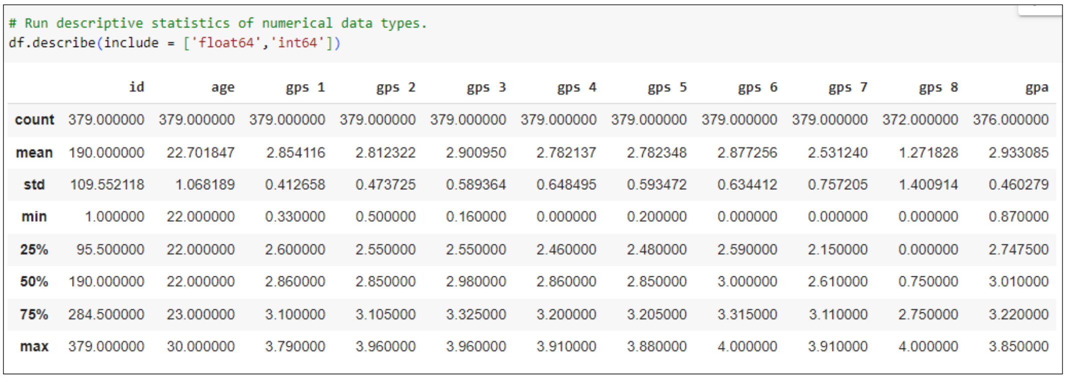


Fig 2. Univariate Analysis

Figure 2 shows the results of the univariate analysis on the distribution of graduation, work, and gender feature data. Data distribution for the Graduation feature is a dependent attribute. The number of students who graduated on time was 57.0%, and 43% were late. This shows that more students graduate on time than late. Meanwhile, in terms of work distribution, the number of students who do not work is 64.9%, and those who do work are 35.1%. Meanwhile, in terms of gender distribution, 61.7% of students are male and 38.3% are female.



```
# Run descriptive statistics of numerical data types.
df.describe(include = ['float64','int64'])
```

	id	age	gps 1	gps 2	gps 3	gps 4	gps 5	gps 6	gps 7	gps 8	gpa
count	379.000000	379.000000	379.000000	379.000000	379.000000	379.000000	379.000000	379.000000	379.000000	372.000000	376.000000
mean	190.000000	22.701847	2.854116	2.812322	2.900950	2.782137	2.782348	2.877256	2.531240	1.271828	2.933085
std	109.552118	1.068189	0.412658	0.473725	0.589364	0.648495	0.593472	0.634412	0.757205	1.400914	0.460279
min	1.000000	22.000000	0.330000	0.500000	0.160000	0.000000	0.200000	0.000000	0.000000	0.000000	0.870000
25%	95.500000	22.000000	2.600000	2.550000	2.550000	2.460000	2.480000	2.590000	2.150000	0.000000	2.747500
50%	190.000000	22.000000	2.860000	2.850000	2.980000	2.860000	2.850000	3.000000	2.610000	0.750000	3.010000
75%	284.500000	23.000000	3.100000	3.105000	3.325000	3.200000	3.205000	3.315000	3.110000	2.750000	3.220000
max	379.000000	30.000000	3.790000	3.960000	3.960000	3.910000	3.880000	4.000000	3.910000	4.000000	3.850000

Fig 3. Descriptive Statistics of Numerical Data Types

The describe() function in Figure 3 displays descriptive statistics from a data frame or series. A summary of the central tendency and distribution of the dataset is displayed by this function, which helps us get a quick overview of the dataset. Count is the number of rows of dataset data, mean is the average, std is the standard deviation, min is the minimum value, and max is the maximum value. The average student GPA is 2.93, with a minimum GPA of 0.87 and a maximum GPA of 3.85.

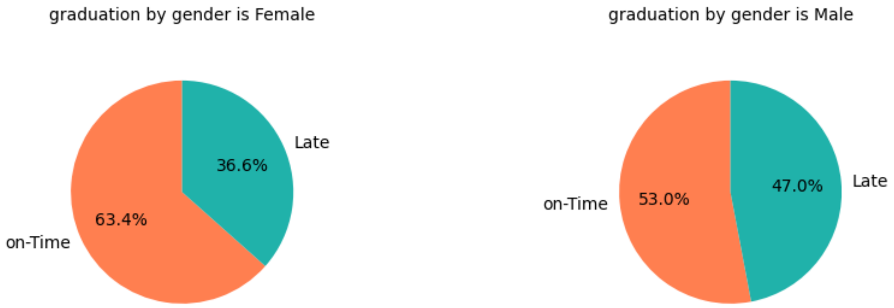


Fig 4. Graduation Correlation by Gender

Figure 4 shows the results of the bivariate and multivariate analyses. The pie chart shows the relationship between graduation rates and gender. The pie chart shows that female students are more likely to complete their studies on time with a percentage of 63.4% compared to male students with a percentage of 53%.

* Corresponding author

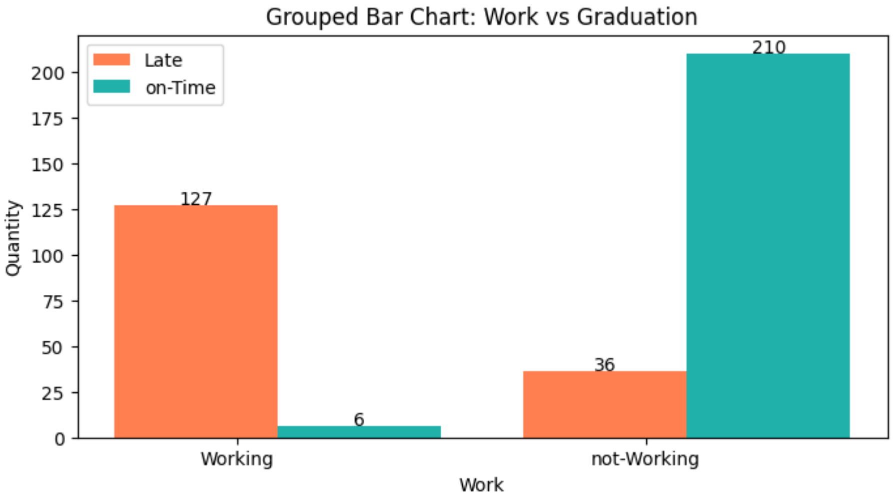


Fig 5. Correlation of Graduation by Work

The bar chart in Figure 5 shows the relationship between graduation rates and student status. From the graph, it appears that students who study while working are less likely to graduate on time compared to students who do not work.

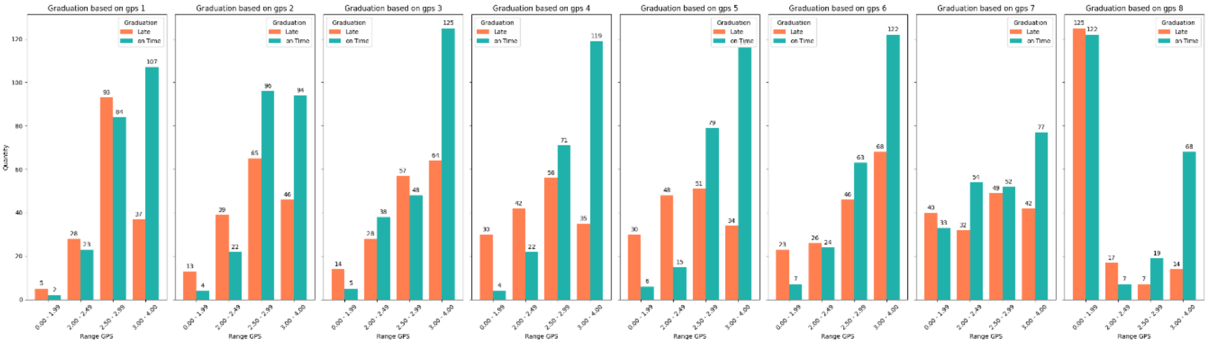
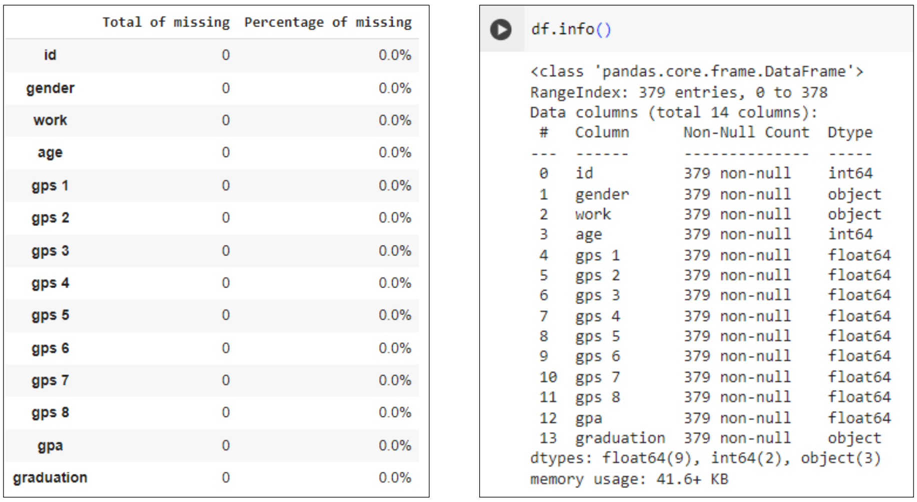


Fig 6. Graduation based on GPS range

Figure 6 shows the student graduation rate based on the GPS range each semester. The graph shows that many students who have a GPS average above 3.00 for eight consecutive semesters graduate on time.



	Total of missing	Percentage of missing
id	0	0.0%
gender	0	0.0%
work	0	0.0%
age	0	0.0%
gps 1	0	0.0%
gps 2	0	0.0%
gps 3	0	0.0%
gps 4	0	0.0%
gps 5	0	0.0%
gps 6	0	0.0%
gps 7	0	0.0%
gps 8	0	0.0%
gpa	0	0.0%
graduation	0	0.0%

```
df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 379 entries, 0 to 378
Data columns (total 14 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   id          379 non-null    int64
 1   gender      379 non-null    object
 2   work        379 non-null    object
 3   age         379 non-null    int64
 4   gps 1       379 non-null    float64
 5   gps 2       379 non-null    float64
 6   gps 3       379 non-null    float64
 7   gps 4       379 non-null    float64
 8   gps 5       379 non-null    float64
 9   gps 6       379 non-null    float64
 10  gps 7       379 non-null    float64
 11  gps 8       379 non-null    float64
 12  gpa         379 non-null    float64
 13  graduation  379 non-null    object
dtypes: float64(9), int64(2), object(3)
memory usage: 41.6+ KB
```

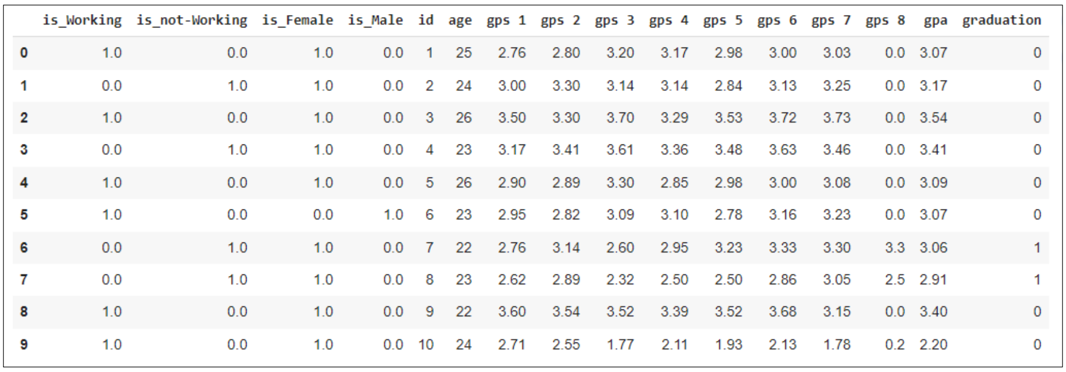
(a) (b)

Fig 7. Results of Handling Missing Values

* Corresponding author

In this section, the results of data preprocessing will be displayed, which include the results of handling missing values, feature encoding, data normalization, and getting highly correlated features. From table 2, it has been shown that the missing data is numeric type data, with a total of 7 data points for the GPS feature 8 with a percentage of 1.85% and 3 data points for the GPA feature with a percentage of 0.79%.

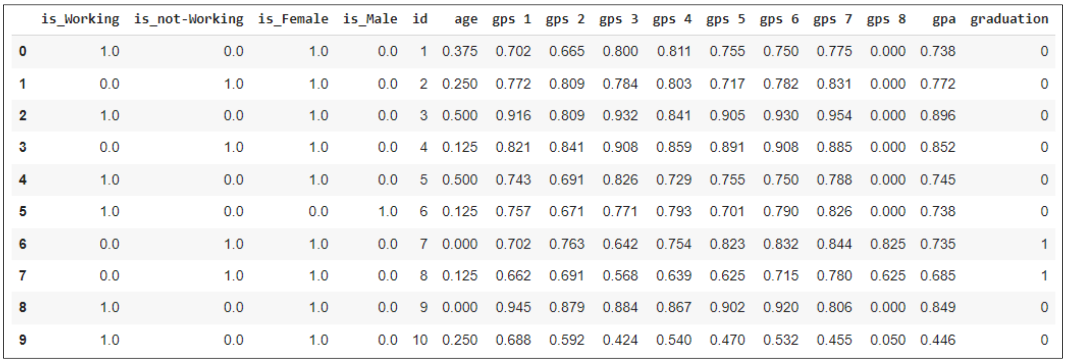
Figure 7 shows the results of handling missing data. In part (a), it can be seen that the GPS line 8 and GPA values have changed to 0. This means that the missing values for these features have been handled. The process of handling missing data in the GPS 8 feature is done by filling in the missing values in each column with a value of 0. Meanwhile, handling missing data in the GPA feature is done by filling in the missing values in the column with a median GPA value of 3.01. From the data info in part (b) it can be seen that all features are complete with a total of 379 rows. Next, the results of feature encoding are shown in Figure 8.



	is_Working	is_not-Working	is_Female	is_Male	id	age	gps 1	gps 2	gps 3	gps 4	gps 5	gps 6	gps 7	gps 8	gpa	graduation
0	1.0	0.0	1.0	0.0	1	25	2.76	2.80	3.20	3.17	2.98	3.00	3.03	0.0	3.07	0
1	0.0	1.0	1.0	0.0	2	24	3.00	3.30	3.14	3.14	2.84	3.13	3.25	0.0	3.17	0
2	1.0	0.0	1.0	0.0	3	26	3.50	3.30	3.70	3.29	3.53	3.72	3.73	0.0	3.54	0
3	0.0	1.0	1.0	0.0	4	23	3.17	3.41	3.61	3.36	3.48	3.63	3.46	0.0	3.41	0
4	1.0	0.0	1.0	0.0	5	26	2.90	2.89	3.30	2.85	2.98	3.00	3.08	0.0	3.09	0
5	1.0	0.0	0.0	1.0	6	23	2.95	2.82	3.09	3.10	2.78	3.16	3.23	0.0	3.07	0
6	0.0	1.0	1.0	0.0	7	22	2.76	3.14	2.60	2.95	3.23	3.33	3.30	3.3	3.06	1
7	0.0	1.0	1.0	0.0	8	23	2.62	2.89	2.32	2.50	2.50	2.86	3.05	2.5	2.91	1
8	1.0	0.0	1.0	0.0	9	22	3.60	3.54	3.52	3.39	3.52	3.68	3.15	0.0	3.40	0
9	1.0	0.0	1.0	0.0	10	24	2.71	2.55	1.77	2.11	1.93	2.13	1.78	0.2	2.20	0

Fig 8. Results of Feature Encoding

The feature coding process is carried out by changing the object data type to float and integer data types. In this research, gender and work features are converted into float data types. Meanwhile, the graduation feature has been changed to an integer data type. After that, a new feature is produced to replace the old feature. In Figure 8, you can see the addition of four new features, namely: is_working, is_not-working, is_female, and is_male. The is_working and is_not-working features are two new features that replace the work feature. Meanwhile, the is_female and is_male features are two new features that replace the gender feature. By using the one-hot encoding method, the value is filled with "0" if it does not match the original value, and the value "1" will be filled in if it matches the original value. In the graduation feature, the label encoding method is used. In this feature, the "On Time" value is replaced with "1", while the "Late" value is replaced with "0." Next, the results of the data normalization are shown.

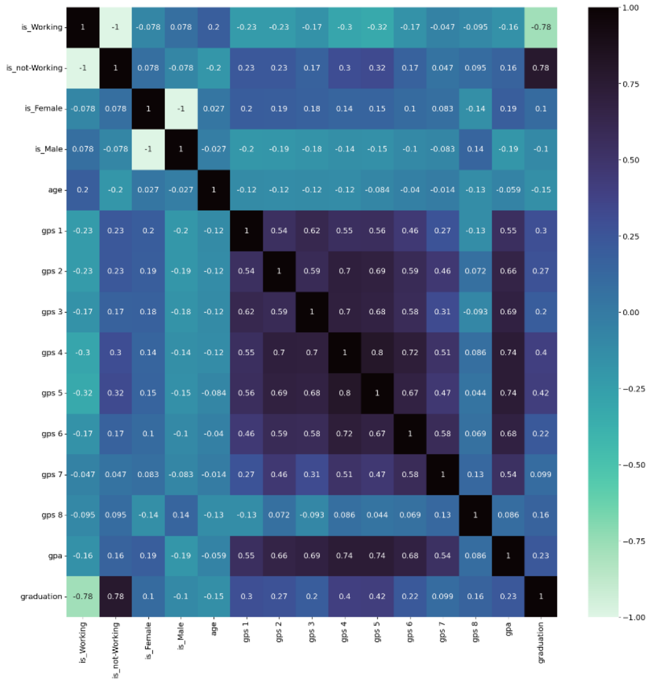


	is_Working	is_not-Working	is_Female	is_Male	id	age	gps 1	gps 2	gps 3	gps 4	gps 5	gps 6	gps 7	gps 8	gpa	graduation
0	1.0	0.0	1.0	0.0	1	0.375	0.702	0.665	0.800	0.811	0.755	0.750	0.775	0.000	0.738	0
1	0.0	1.0	1.0	0.0	2	0.250	0.772	0.809	0.784	0.803	0.717	0.782	0.831	0.000	0.772	0
2	1.0	0.0	1.0	0.0	3	0.500	0.916	0.809	0.932	0.841	0.905	0.930	0.954	0.000	0.896	0
3	0.0	1.0	1.0	0.0	4	0.125	0.821	0.841	0.908	0.859	0.891	0.908	0.885	0.000	0.852	0
4	1.0	0.0	1.0	0.0	5	0.500	0.743	0.691	0.826	0.729	0.755	0.750	0.788	0.000	0.745	0
5	1.0	0.0	0.0	1.0	6	0.125	0.757	0.671	0.771	0.793	0.701	0.790	0.826	0.000	0.738	0
6	0.0	1.0	1.0	0.0	7	0.000	0.702	0.763	0.642	0.754	0.823	0.832	0.844	0.825	0.735	1
7	0.0	1.0	1.0	0.0	8	0.125	0.662	0.691	0.568	0.639	0.625	0.715	0.780	0.625	0.685	1
8	1.0	0.0	1.0	0.0	9	0.000	0.945	0.879	0.884	0.867	0.902	0.920	0.806	0.000	0.849	0
9	1.0	0.0	1.0	0.0	10	0.250	0.688	0.592	0.424	0.540	0.470	0.532	0.455	0.050	0.446	0

Fig 9. Results of Data Normalization

In Figure 9, it can be seen that the values for the age, GPS, and GPA features have changed to values in the range 0 to 1. Then, the normalized data is saved in the form of a CSV file, to be used again in the feature selection process.

* Corresponding author



	is_Working	is_not-Working	is_Female	is_Male	age	gps 1	gps 2	gps 3	gps 4	gps 5	gps 6	gps 7	gps 8	gpa	graduation
is_Working	1	-1	-0.078	0.078	0.2	-0.23	-0.23	-0.17	-0.3	-0.32	-0.17	-0.047	-0.095	-0.16	-0.78
is_not-Working	-1	1	0.078	-0.078	-0.2	0.23	0.23	0.17	0.3	0.32	0.17	0.047	0.095	0.16	0.78
is_Female	-0.078	0.078	1	-1	0.027	0.2	0.19	0.18	0.14	0.15	0.1	0.083	-0.14	0.19	0.1
is_Male	0.078	-0.078	-1	1	-0.027	-0.2	-0.19	-0.18	-0.14	-0.15	-0.1	-0.083	0.14	-0.19	-0.1
age	0.2	-0.2	0.027	-0.027	1	-0.12	-0.12	-0.12	-0.12	-0.084	-0.04	-0.014	-0.13	-0.059	-0.15
gps 1	-0.23	0.23	0.2	-0.2	-0.12	1	0.54	0.62	0.55	0.56	0.46	0.27	-0.13	0.55	0.3
gps 2	-0.23	0.23	0.19	-0.19	-0.12	0.54	1	0.59	0.7	0.69	0.59	0.46	0.072	0.66	0.27
gps 3	-0.17	0.17	0.18	-0.18	-0.12	0.62	0.59	1	0.7	0.68	0.58	0.31	-0.093	0.69	0.2
gps 4	-0.3	0.3	0.14	-0.14	-0.12	0.55	0.7	0.7	1	0.8	0.72	0.51	0.086	0.74	0.4
gps 5	-0.32	0.32	0.15	-0.15	-0.084	0.56	0.69	0.68	0.8	1	0.67	0.47	0.044	0.74	0.42
gps 6	-0.17	0.17	0.1	-0.1	-0.04	0.46	0.59	0.58	0.72	0.67	1	0.58	0.069	0.68	0.22
gps 7	-0.047	0.047	0.083	-0.083	-0.014	0.27	0.46	0.31	0.51	0.47	0.58	1	0.13	0.54	0.099
gps 8	-0.095	0.095	-0.14	0.14	-0.13	-0.13	0.072	-0.093	0.086	0.044	0.069	0.13	1	0.086	0.16
gpa	-0.16	0.16	0.19	-0.19	-0.059	0.55	0.66	0.69	0.74	0.74	0.68	0.54	0.086	1	0.23
graduation	-0.78	0.78	0.1	-0.1	-0.15	0.3	0.27	0.2	0.4	0.42	0.22	0.099	0.16	0.23	1

Fig 10. Results of Correlated Features

The feature selection process uses the Seaborn library in Python, using the heatmap() function. In this study, a threshold of 0.2 was used. This means that if the correlation between existing features and the graduation feature is smaller than 0.2, then the feature will be discarded. The darker the color of the correlation, the closer the value is to 1. Meanwhile, the lighter the color, the closer the closer the value is to 0. Figure 10 shows the correlation between features based on their numerical values. From the correlation hetmap results between features, there are eight features whose values are above 0.2, namely: is_working, is_not-working, gps 1, gps 2, gps 4, gps 5, gps 6, and gpa. These eight features will still be maintained. Meanwhile, features such as is_female, is_male, age, gps 3, gps 7, and gps 8 were removed because their correlation values were below 0.2. At this stage, data preprocessing has been completed. Eight features have been produced that will be used as input variables in the artificial neural network, with the target variable being graduation, which has on-time and late classifications.

As explained in the methodology section, this research applies five scenarios for dividing training data and test data on artificial neural networks and uses various numbers of epochs and learning rates as well. So, the test results are shown in the following tables:

Table 3. Test Results on 10% Test Data

Epoch	Learning Rate	Accuracy
100	0.001	68.33%
	0.01	57.48%
	0.1	57.48%
	0.5	57.48%
500	0.001	68.33%
	0.01	57.48%
	0.1	57.48%
	0.5	57.48%

From Table 3, it can be seen that the highest accuracy value from artificial neural network testing with 10% test data was obtained from a network model with a number of epochs of 100 and 500 at a learning rate of 0.001 with an accuracy of 68.33%.

* Corresponding author

Table 4. Test Results on 20% Test Data

Epoch	Learning Rate	Accucary
100	0.001	64.36%
	0.01	59.08%
	0.1	59.08%
	0.5	59.08%
500	0.001	64.36%
	0.01	59.08%
	0.1	59.08%
	0.5	59.08%

From Table 4, it can be seen that the highest accuracy value from artificial neural network testing with 20% test data was obtained from a network model with a number of epochs of 100 and 500 at a learning rate of 0.001 with an accuracy of 64.36%.

Table 5. Test Results on 30% Test Data

Epoch	Learning Rate	Accucary
100	0.001	63.77%
	0.01	59.62%
	0.1	59.62%
	0.5	59.62%
500	0.001	63.77%
	0.01	59.62%
	0.1	59.62%
	0.5	59.62%

From Table 5, it can be seen that the highest accuracy value from artificial neural network testing with 30% test data was obtained from a network model with a number of epochs of 100 and 500 at a learning rate of 0.001 with an accuracy of 63.77%.

Table 6. Test Results on 40% Test Data

Epoch	Learning Rate	Accucary
100	0.001	68.72%
	0.01	57.27%
	0.1	57.27%
	0.5	57.27%
500	0.001	68.72%
	0.01	57.27%
	0.1	57.27%
	0.5	57.27%

From Table 6, it can be seen that the highest accuracy value from artificial neural network testing with 40% test data was obtained from a network model with a number of epochs of 100 and 500 at a learning rate of 0.001 with an accuracy of 68.72%.

Table 7. Test Results on 50% Test Data

Epoch	Learning Rate	Accucary
100	0.001	69.84%
	0.01	56.61%
	0.1	56.61%
	0.5	56.61%
500	0.001	69.84%
	0.01	56.61%
	0.1	56.61%
	0.5	56.61%

From Table 7, it can be seen that the highest accuracy value from artificial neural network testing with 50% test data was obtained from a network model with a number of epochs of 100 and 500 at a learning rate of 0.001 with an accuracy of 69.84%.

DISCUSSIONS

From this research, it can be explained that the application of exploratory data analysis is able to provide a strong basis for exploring further the influence of features on the research dataset. With exploratory data analysis, it can be proven that students who study while working are less likely to graduate on time than students who do not work. Students with a GPS average above 3.00 for eight consecutive semesters will graduate on time. However, it is important to remember that maintaining a GPS average above 3.00 does not automatically guarantee that a student will graduate on time. Many factors influence this phenomenon, but it is important to conduct thorough research on the various other features that influence students' on-time graduation. The feature selection process has proven that there is no significant influence of gender and age features on timely graduation.

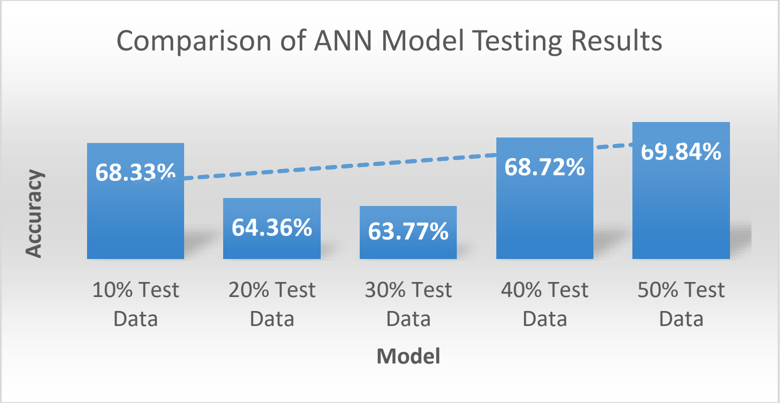


Fig 11. Comparison of Artificial Neural Network Model Results

From Figure 11, it can be seen that the best artificial neural network model is a network model with a total of 50% test data and an accuracy result of 69.84%. From the graph, you can also see a trend of decreasing accuracy at the amount of test data of 20% to 30% and increasing again at the amount of test data of 40% to 50%. These results prove that there is an influence of the amount of test data and learning rate on the level of accuracy. This research proves that there is an increase in accuracy caused by the learning rate. A hyperparameter known as the learning rate regulates how much the model must change in response to error estimates that change each time the model weights are updated. A learning rate that is too small does produce good accuracy, but it can result in a long training process. Meanwhile, a learning rate value that is too large can cause weights that are not ideal or an unstable training process. Meanwhile, applying a different number of epochs does not affect the accuracy results at all.

This research certainly has limitations in terms of the number of datasets, the number of features, the number of training and testing schemes, and the algorithm applied. What are the limitations that have an impact on the accuracy results, which only produce a figure of 69.84%, not above 90%? Previous research has proven that the exploratory data analysis method has succeeded in increasing the accuracy of the predictive model compared to the basic model (Rahmat et al., 2020). However, in this study, it has not been possible to prove the effect of exploratory data analysis on the level of prediction accuracy. For further research, a comparative study can be carried out using exploratory data analysis and without exploratory data analysis on cases of predicting students' on-time graduation. So we can see the effect of exploratory data analysis on the level of accuracy in predicting students' on-time graduation. In future research, a larger number of datasets with a much more complex number of features can also be applied. In order to gain a more comprehensive understanding of the influence of features on students' timely graduation, For a number of training and testing schemes, various numbers of hidden layers can be applied to the artificial neural network, and various types of learning rates can be applied. Meanwhile, regarding the type of algorithm, it is possible to compare several types of machine learning algorithms for predicting students' graduation on time.

CONCLUSION

This research has succeeded in applying exploratory data analysis and artificial neural network methods to predict students' on-time graduation. The results of applying exploratory data analysis have proven the strong influence of the work feature and GPS feature on the student's on-time graduation dataset. The influence of the work feature proves that students who study while working are less likely to graduate on time compared to students

* Corresponding author

who do not work. Meanwhile, the influence of the GPS feature proves that students who have an average GPS above 3.00 for eight consecutive semesters find it easier to graduate on time. From this research, it is also shown that the work features, gps 1, gps 2, gps 4, gps 5, gps 6, and gpa, are highly correlated with students graduating on time. So these features are maintained. Meanwhile, the features of gender, age, gps 3, gps 7, and gpss 8 do not have a significant correlation with students graduating on time. So these features were removed. In the application of artificial neural networks, a model that has a high level of accuracy is a network model with a test data percentage of 50%, a number of epochs of 100, and a learning rate of 0.001, which produces an accuracy of 69.84%.

REFERENCES

Abukmeil, M., Ferrari, S., Genovese, A., Piuri, V., & Scotti, F. (2021). A Survey of Unsupervised Generative Models for Exploratory Data Analysis and Representation Learning. *ACM Comput. Surv.*, *54*(5). https://doi.org/10.1145/3450963

Aldera, S., Emam, A., Al-Qurishi, M., Alrubaian, M., & Alothaim, A. (2021). Exploratory Data Analysis and Classification of a New Arabic Online Extremism Dataset. *IEEE Access*, *9*, 161613–161626. https://doi.org/10.1109/ACCESS.2021.3132651

Alyahyan, E., & Düştegör, D. (2020). Predicting academic success in higher education: literature review and best practices. *International Journal of Educational Technology in Higher Education*, *17*(1). https://doi.org/10.1186/s41239-020-0177-7

Apridiansyah, Y., Veronika, N. D. M., & Putra, E. D. (2021). Prediksi Kelulusan Mahasiswa Fakultas Teknik Informatika Universitas Muhammadiyah Bengkulu Menggunakan Metode Naive Bayes. *JSAI : Journal Scientific and Applied Informatics*, *4*(2), 236–247. https://doi.org/10.36085/jsai.v4i2.1701

Athallah, H. (2023). Kelulusan Mahasiswa. Retrieved March 20, 2024, from kaggle website: https://www.kaggle.com/datasets/hafizhathallah/kelulusan-mahasiswa

Azahari, A., Yulindawati, Y., Rosita, D., & Mallala, S. (2020). Komparasi Data Mining Naive Bayes dan Neural Network memprediksi Masa Studi Mahasiswa S1. *Jurnal Teknologi Informasi Dan Ilmu Komputer (JTIIK)*, *7*(3), 443–452. https://doi.org/10.25126/jtiik.2020732093

Bangun, O., Mawengkang, H., & Efendi, S. (2022). Metode Algoritma Support Vector Machine (SVM) Linier Dalam Memprediksi Kelulusan Mahasiswa. *Jurnal Media Informatika Budidarma*, *6*(4), 2006. https://doi.org/10.30865/mib.v6i4.4572

Basheer, M. Y. I., Mutalib, S., Hamid, N. H. A., Abdul-Rahman, S., & Malik, A. M. A. (2019). Predictive analytics of university student intake using supervised methods. *IAES International Journal of Artificial Intelligence (IJ-AI)*, *8*(4), 367–374. https://doi.org/10.11591/ijai.v8.i4.pp367-374

Bukhari, M. M., Ullah, S. S., Uddin, M., Hussain, S., Abdelhaq, M., & Alsaqour, R. (2022). An Intelligent Model for Predicting the Students' Performance with Backpropagation Neural Network Algorithm Using Regularization Approach. *Human-Centric Computing and Information Sciences*, *12*. https://doi.org/10.22967/HCIS.2022.12.044

Dengen, C. N., Kusrini, K., & Luthfi, E. T. (2020). Implementasi Decision Tree Untuk Prediksi Kelulusan Mahasiswa Tepat Waktu. *SISFOTENIKA*, *10*(1), 1. https://doi.org/10.30700/jst.v10i1.484

Figueroa-Cañas, J., & Sancho-Vinuesa, T. (2020). Early Prediction of Dropout and Final Exam Performance in an Online Statistics Course. *IEEE Revista Iberoamericana de Tecnologias Del Aprendizaje*, *15*(2), 86–94. https://doi.org/10.1109/RITA.2020.2987727

Fiqha, I., Yandris, G. J., & Nasution, F. A. (2022). Implementation of Neural Network Algorithms in predicting student graduation rates. *Sinkron*, *6*(1), 248–255. https://doi.org/10.33395/sinkron.v7i1.11254

Gunawan, M., Zarlis, M., & Roslina, R. (2021). Analisis Komparasi Algoritma Naïve Bayes dan K-Nearest Neighbor Untuk Memprediksi Kelulusan Mahasiswa Tepat Waktu. *Jurnal Media Informatika Budidarma*, *5*(2), 513. https://doi.org/10.30865/mib.v5i2.2925

Harun, R. R., Septyanun, N., Yuliani, T., AM, J., Hamdi, & Rejeki, S. (2022). Lulus Tepat Waktu: Sebuah Motivasi Dan Kode Etik Belajar Bagi Mahasiswa di Perguruan Tinggi. *JCES (Journal of Character Education Society)*, *5*(3), 773–779. https://doi.org/10.31764/jces.v3i1.9118

Indrakumari, R., Poongodi, T., & Jena, S. R. (2020). Heart Disease Prediction using Exploratory Data Analysis. *Procedia Computer Science*, *173*, 130–139. https://doi.org/https://doi.org/10.1016/j.procs.2020.06.017

Kartarina, K., Sriwinarti, N. K., & Juniarti, N. luh P. (2021). Analisis Metode K-Nearest Neighbors (K-NN) Dan Naive Bayes Dalam Memprediksi Kelulusan Mahasiswa. *JTIM : Jurnal Teknologi Informasi Dan Multimedia*, *3*(2), 107–113. https://doi.org/10.35746/jtim.v3i2.159

Kulkarni, A., & Shivananda, A. (2019). *Exploring and Processing Text Data BT - Natural Language Processing Recipes: Unlocking Text Data with Machine Learning and Deep Learning using Python* (A. Kulkarni & A. Shivananda, Eds.). Berkeley, CA: Apress. https://doi.org/10.1007/978-1-4842-4267-4_2

Lagman, A. C., Alfonso, L. P., Goh, M. L. I., Lalata, J. A. P., Magcuyao, J. P. H., & Vicente, H. N. (2020).

* Corresponding author

Classification Algorithm Accuracy Improvement for Student Graduation Prediction Using Ensemble Model. *International Journal of Information and Education Technology*, *10*(10), 723–727. https://doi.org/10.18178/ijiet.2020.10.10.1449

Liao, S. N., Zingaro, D., Thai, K., Alvarado, C., Griswold, W. G., & Porter, L. (2019). A Robust Machine Learning Technique to Predict Low-performing Students. *ACM Transactions on Computing Education*, *19*(3), 1–19. https://doi.org/10.1145/3277569

Masrizal, M., & Hadiansa, A. (2017). Prediksi Jumlah Lulusan Mahasiswa STMIK Dumai Menggunakan Jaringan Syaraf Tiruan. *INFORMATIKA*, *9*(2), 9–14. https://doi.org/10.36723/juri.v9i2.98

Muliono, R., Lubis, J. H., & Khairina, N. (2020). Analysis K-Nearest Neighbor Algorithm for Improving Prediction Student Graduation Time. *SinkrOn*, *4*(2), 42. https://doi.org/10.33395/sinkron.v4i2.10480

Pelima, L. R., Sukmana, Y., & Rosmansyah, Y. (2024). Predicting University Student Graduation Using Academic Performance and Machine Learning: A Systematic Literature Review. *IEEE Access*, *12*(February), 23451–23465. https://doi.org/10.1109/ACCESS.2024.3361479

Priyatman, H., Sajid, F., & Haldivany, D. (2019). Klasterisasi Menggunakan Algoritma K-Means Clustering untuk Memprediksi Waktu Kelulusan Mahasiswa. *Jurnal Edukasi Dan Penelitian Informatika (JEPIN)*, *5*(1), 62. https://doi.org/10.26418/jp.v5i1.29611

Putri, R. P. S., & Waspada, I. (2018). Penerapan Algoritma C4.5 pada Aplikasi Prediksi Kelulusan Mahasiswa Prodi Informatika. *Khazanah Informatika : Jurnal Ilmu Komputer Dan Informatika*, *4*(1), 1–7. https://doi.org/10.23917/khif.v4i1.5975

Qisthiano, M. R., Kurniawan, T. B., Negara, E. S., & Akbar, M. (2021). Pengembangan Model Untuk Prediksi Tingkat Kelulusan Mahasiswa Tepat Waktu dengan Metode Naïve Bayes. *Jurnal Media Informatika Budidarma*, *5*(3), 987. https://doi.org/10.30865/mib.v5i3.3030

Rahmat, F., Zulkafli, Z., Juraiza Ishak, A., Mohd Noor, S. B., Yahaya, H., & Masrani, A. (2020). Exploratory Data Analysis and Artificial Neural Network for Prediction of Leptospirosis Occurrence in Seremban, Malaysia Based on Meteorological Data. *Frontiers in Earth Science*, *8*(November), 1–14. https://doi.org/10.3389/feart.2020.00377

Ridwan, R., Lubis, H., & Kustanto, P. (2020). Implementasi Algoritma Neural Network dalam Memprediksi Tingkat Kelulusan Mahasiswa. *Jurnal Media Informatika Budidarma*, *4*(2), 286. https://doi.org/10.30865/mib.v4i2.2035

Riyanto, V., Hamid, A., & Ridwansyah, R. (2019). Prediction of Student Graduation Time Using The Best Algorithm. *Indonesian Journal of Artificial Intelligence and Data Mining (IJAIDM)*, *2*(1), 1–9. https://doi.org/10.24014/ijaidm.v2i1.6424

Rodríguez-Hernández, C. F., Musso, M., Kyndt, E., & Cascallar, E. (2021). Artificial neural networks in academic performance prediction: Systematic implementation and predictor evaluation. *Computers and Education: Artificial Intelligence*, *2*, 100018. https://doi.org/https://doi.org/10.1016/j.caeai.2021.100018

Sahoo, K., Samal, A. K., Pramanik, J., & Pani, S. K. (2019). Exploratory Data Analysis using Python. *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, *8*(12), 4727–4735. https://doi.org/10.35940/ijitee.L3591.1081219

Suhaimi, N. M., Abdul-Rahman, S., Mutalib, S., Hamid, N. H. A., & Malik, A. M. A. (2019). Review on Predicting Students' Graduation Time Using Machine Learning Algorithms. *International Journal of Modern Education and Computer Science*, *11*(7), 1–13. https://doi.org/10.5815/ijmecs.2019.07.01

Tampakas, V., Livieris, I. E., Pintelas, E., Karacapilidis, N., & Pintelas, P. (2019). Prediction of Students' Graduation Time Using a Two-Level Classification Algorithm. In M. Tsitouridou, J. A. Diniz, & T. A. Mikropoulos (Eds.), *International Conference on Technology and Innovation in Learning, Teaching and Education* (pp. 553–565). Cham: Springer International Publishing.

Tinggi, B. A. N. P. *Matriks Penilaian LED Dan LKPS Program Sarjana*. , Pub. L. No. Peraturan BAN-PT Nomor 5 tahun 2019, 1 (2019). Indonesia.

Yaqin, A., Laksito, A. D., & Fatonah, S. (2021). Evaluation of Backpropagation Neural Network Models for Early Prediction of Student's Graduation in XYZ University. *International Journal on Advanced Science, Engineering and Information Technology*, *11*(2), 610–617. https://doi.org/10.18517/ijaseit.11.2.11152