e-ISSN: 2614-8404
p-ISSN:2776-3234

Comparison of ANN Backpropagation Algorithm and Random Forest Regression in Predicting the Number of New Students

Padmavati Darma Putri Tanuwijaya[1], Jhonatan Laurensius Tjahjadi[2], Yosefina Finsensia Riti[3*)]
[1]Program Studi Ilmu Informatika, Fakultas Teknik, Universitas Katolik Darma Cendika Surabaya
[2]Program Studi Ilmu Informatika, Fakultas Teknik, Universitas Katolik Darma Cendika Surabaya
[3]Program Studi Ilmu Informatika, Fakultas Teknik, Universitas Katolik Darma Cendika Surabaya
Email: [1]padmavati.tanuwijaya@student.ukdc.ac.id, [2]jhonatan.tjahjadi@student.ukdc.ac.id, [3*]yosefina.riti@ukdc.ac.id

Abstract **–** Higher education institutions are educational units located at a higher level after high school or vocational school. Catholic University Darma Cendika Surabaya (UKDC) faces challenges in managing the admission of new students due to variations in the number of prospective students applying to each department, which is also influenced by changing trends in interests and job needs in Indonesia. The use of Artificial Neural Network with Backpropagation and Random Forest Regression algorithms for comparing the prediction of new student admissions in the following year will be beneficial for the administration of Catholic University Darma Cendika Surabaya (UKDC) to gain a clearer understanding of the dynamics of admissions and to support decision making in the future development of the university. The predicted number of students joining Catholic University Darma Cendika Surabaya (UKDC) in the 2024 period using Artificial Neural Network is 219 students with a Mean Squared Error (MSE) of 0,1046 and Root Mean Square Error (RMSE) of 0,32.

Keywords – Artificial Neural Network, Backpropagation, Random Forest Regression , MSE, RMSE.

I. INTRODUCTION

In a university there are three important roles in its implementation, namely: carry out education, carry out research, and carry out community service [1]. The new student admission process is the starting point of the University in forming a diverse academic community and has the potential to develop their interests and ambitions, this is a key aspect of higher education administration that will affect the continuation and academic development of the University [2]. Darma Cendika Catholic University (UKDC) has a variety of variations in new student admissions each year such as the number of prospective new students who register in each department, this must be considered so that new student admissions can increase the following year. Darma Cendika Catholic University can prepare by predicting the number of students who will register in the following year by predicting it. In this study, researchers predicted new students in the following year using the Artificial Neural Network and Random Forest Regression algorithms using historical data on student admissions from 2017-2023.

Artificial Neural Network (ANN) is an information processing system that can produce predictive values from data with high accuracy using artificial neural networks that model complex relationships between inputs and outputs so that they resemble the characteristics of human nerves. ANN is also able to identify patterns in data when processing large datasets [3]. Meanwhile, Random Forest Regression is a regression learning algorithm used to produce predictive values more strongly and accurately than a single model. Random Forest Regression is suitable for solving various types of problems [4].

Previous research has illustrated the progress of the use of Artificial Neural Network (ANN) application in designing optimization of rice production prediction using multilayer [5], application in predicting nickel ore production [6], use in predicting rainfall [7], use in predicting student graduation predicate [3], use in predicting stock prices [8] and Random Forest Regression use in predicting the number of class participants for schedule planning [4], use in predicting house prices [9], use in predicting coffee quality [10], use in predicting Ciliwung river water quality [11], use in predicting cell phone prices [12] for different algorithm applications. However, each researcher has various tests, at this time there is still no one who uses Artificial Neural Network and Random Forest Regression models as a comparison to predict the number of new students at Darma Cendika Catholic University Surabaya which is a research factor with a different context.

In this research, we aim to achieve a more accurate and relevant prediction comparison with our university situation. The purpose of this research is to provide predictions using Artificial Neural Network and Random Forest Regression that are useful for the administration of Darma Cendika Catholic University Surabaya to provide a clearer view of the dynamics of enrollment in support of decision making on the future development of the university.

II. RESEARCH METHODOLOGY

In this research, the process began with the collection of a dataset. For gathering data on new students, we requested information from the admission office of Darma Cendika Catholic University in Surabaya. Subsequently, we

e-ISSN: 2614-8404
p-ISSN:2776-3234

conducted data cleaning to remove unnecessary information, ensuring that the resulting data would yield optimal results. After the data cleaning process, we proceeded to implement the data into the prediction program we had developed. The process flow is depicted in Figure 1.

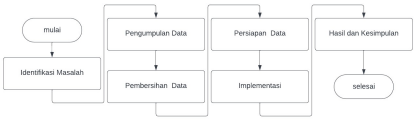


Fig 1. Flowchart of the Research Process.

1. Problem Identification
   Problem identification is crucial for formulating the research focus. The identified issues include:

- Accessibility of data collection through the administration of New Student Admissions.
- Quality of the provided data.
- Variables influencing the data results.

2. Data Collection
   Data collection involved obtaining the dataset for new student admissions from 2017 to 2023 through the UKDC administration.

Table 1. Raw Data for New Student Admissions in 2017

No.	No. Pendaftaran	NPM	Nama	Fakultas
1	004336	17126002	Melanny Natalia	Ekonomi
2	004321	17120003	Hanna Kristina	Ekonomi
3	004342	17120005	Bernadeta Dete	Ekonomi
4	004341	17126003	K. Mega Lestari	Ekonomi
5	004339	17120004	Trisha Claudia Prinzessa F	Ekonomi

Kelas	Gel	Jalur	Alamat
Sore	Gel 1	Umum	Perum. Taman Pondok Jati H/17
Pagi	Gel 1	JMLA	Jl. Mojoarum IX/15
Pagi	Gel 1	Umum	Waru Gunung RT. 05, RW. 03, Karang Pilang, Surabaya
Sore	Gel 1	JMLA	Jl. Kapas Baru 2 / 108, Surabaya
Pagi	Gel 1	JMLA	Deltasari Indah Blok AV-36, Sidoarjo

Kota	Propinsi	Pulau	Asal Sekolah
Sidoarjo	Jawa Timur	Jawa	SMAK ST. Yusuf Karang Pilang
Surabaya	Jawa Timur	Jawa	SMAK Kr. YBPK 1, Surabaya
Surabaya	Jawa Timur	Jawa	SMAK ST. Yusuf Karang Pilang
Surabaya	Jawa Timur	Jawa	SMK Adhikawacana, Surabaya
Sidoarjo	Jawa Timur	Jawa	SMAK Untung Suropati, Sidoarjo

3. Data Cleaning
   Data cleaning involves making changes to the dataset, addressing issues such as handling missing data, ensuring data consistency, and other necessary adjustments. This step is crucial during the data preparation process, allowing the transformation of data into a model dataset ready for analysis and comprehensible to the researcher [6].
4. Data Preparation
   In this study, a division of 80-20 is employed, with 80% as the training dataset and 20% as the test data. Subsequently, data normalization and scaling are performed to ensure that the data has a suitable range and characteristics, aiming to enhance accuracy and reduce the potential for data leakage [13].
5. Student Prediction Process
   The prediction of the number of new students involves the application of Artificial Neural Network and Random Forest Regression through several stages of the process.

Fig. 2 Flowchart of Student Prediction Process.

6. Prediction
   Prediction is an effort or action in which someone anticipates or estimates something in the future by utilizing relevant information from previous periods [4].
7. Artificial Neural Network
   Artificial Neural Network is a data classification model that shares a concept almost similar to the human brain's neural system. The goal of an Artificial Neural Network is to enable a computer to have cognitive thinking abilities, mimicking the way the human brain operates to solve problems [3][13].
8. Backpropagation

e-ISSN: 2614-8404
p-ISSN:2776-3234

Backpropagation, developed by Rumelhart, Hinton, and Williams around 1986, resulted in an iterative algorithm that is considered simple and easy to use [14][15]. Backpropagation is a method in artificial neural networks that uses supervised learning algorithms. It involves multiple perceptrons with many layers used to adjust weight values in hidden layers [15].

9. Random Forest Regression
   Random Forest Regression is a machine learning technique that combines predictions from various algorithms to generate a random forest regression. It produces an average result from hundreds to thousands of decision trees consisting of decision nodes and leaf nodes used for sample evaluation through its test function [4].
10. MSE and RMSE
    Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) are methods for measuring how far the predictions of a model are from the actual data. They provide an indication of the standard deviation of the difference between predicted and observed values, yielding optimal predictive recommendations [16][17].

$$\square\square\square = \frac{1}{\square^{\square\square\square\square}} \square_{\square}^{\square^{\square\square\square\square}} (\square_{\square} - \square_{\square})^2 \quad (1)$$

$$\square\square\square\square = \sqrt{\frac{\square|\square_{\square} - \square_{\square}|^2}{\square}} \quad (2)$$

where :

t_k = The actual data.

y_k = The predicted data.

n = Total number of observations.

11. Implementation

    This research utilizes Google Colab as the text editor provided by Google. Google Colab is based on the Python programming language and is designed to be accessible online, allowing researchers to develop and build Artificial Neural Network (ANN) models for predictions more conveniently and efficiently.

    The researcher implemented the Artificial Neural Network (ANN) model using the KERAS library from TensorFlow. The model consists of an input layer, 14 hidden layers, and an output layer. For the Random Forest Regressor model, the researcher used the Scikit-Learn library, which employs a tree-based algorithm suitable for regression data.

12. Measurement of Error Rate

    After implementation, the process involves measuring the average error rate using Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) [18]. These metrics provide insights into the accuracy of the prediction models, with lower values indicating better performance by minimizing the differences between predicted and actual values.

III. RESULTS AND DISCUSSION

3.1 Data Cleaning

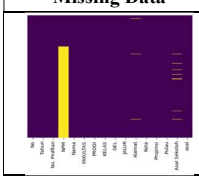
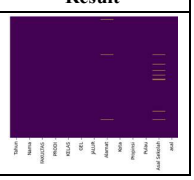
The following Tables 2, 3, 4 present the results after the data cleaning process and grouping of the initial dataset.

Table 2. Total data of new student admissions

Prodi	2017	2018	2019	2020	2021	2022	2023
Total	173	200	282	204	173	226	211

In the results of the missing data check, as indicated in Table 6's missing data column, it is evident that the NPM column has a high number of missing values. The researcher performed data cleaning by removing several columns, such as NMP, and irrelevant tables, such as No and No. Pndftan. This process resulted in the cleaned data image.

Table 3. Missing Data from Dataset

Missing Data	Result
	

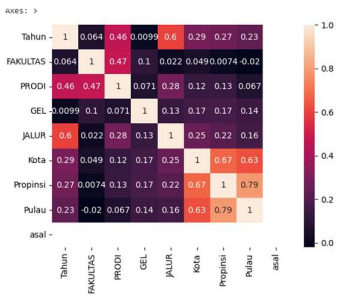
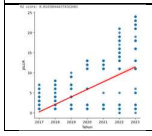
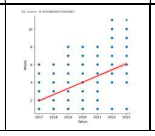
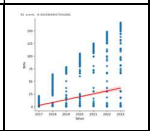


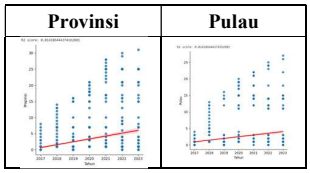
Fig 3. Correlation Matrix among Data.

Based on the correlation matrix above, linear regression is determined for the top 5 data, namely pathway, program, city, province, and island. The table provides a visualization of the linear regression for each correlation.

Table 4. Missing Data from Dataset

Jalur	Prodi	Kota
		

e-ISSN: 2614-8404
p-ISSN:2776-3234



3.2 Results of CSV Data Transfer

In "Figure 4," the results of transferring data from an Excel file to a CSV format are presented.

```
tahun;jumlah
2017;173
2018;200
2019;282
2020;204
2021;173
2022;226
2023;211
```

Figure. 4 Result Transferring Data to a CSV file.

3.3 Implementation

Conducting a test by setting the parameter for 2023 as the prediction target involves creating a new CSV file by excluding data from the year 2023.

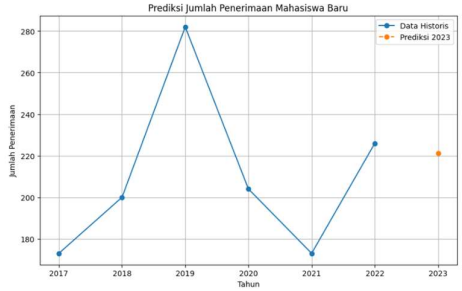


Fig 5. Prediction of New Students in 2023 using ANN.

Analysis of the predicted number of new student admissions for the year 2023 using ANN:

1. In the prediction section, the estimated number of new student admissions expected to join Darma Cendika Catholic University (UKDC) in 2023 is approximately 221.31 or 221 students.
2. The model produces an MSE value of around 0.116, indicating a low level of prediction error in the Artificial Neural Network model.
3. The RMSE value is approximately 0.34, signifying that the predictions are close to the actual data.

Comparing the predicted results with the actual data, there is a difference of around 4.74%, with the predicted value being slightly higher than the actual data (211 to 221 in the prediction results).

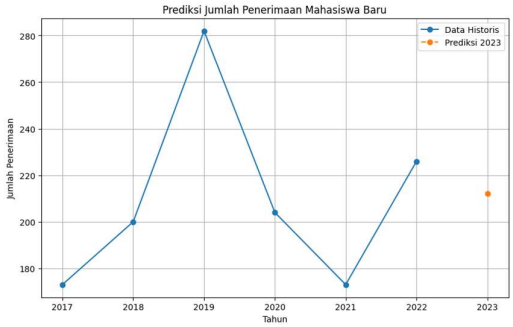


Fig 6. Prediction of New Students in 2023 using Random Forest Regressor.

Analysis Results of Predicting the Number of New Students in 2023 using Random Forest Regressor:

In the prediction section, the estimated number of new student admissions expected to join Darma Cendika Catholic University (UKDC) in 2023 is approximately 212.21 or 212 students.

1. The model produces an MSE value of around 0.3651, indicating a low level of prediction error in the Random Forest Regressor model.
2. The RMSE value is approximately 0.60, signifying that the predictions are close to the actual data.
3. Comparing the predicted results with the actual data, there is a difference of around 0.47%, with the predicted value being slightly higher than the actual data (211 to 212 in the prediction results).

Further testing will be conducted with datasets from 2017 to 2023, followed by predictions for new student admissions in 2024. Figure 4 represents the graphical results of predicting new student admissions in 2024.

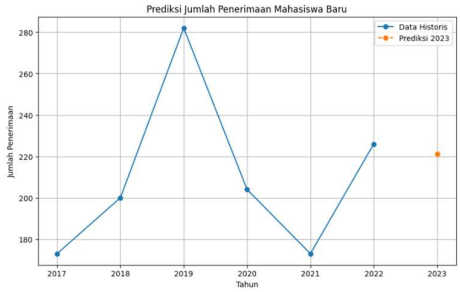


Fig 7 Prediction of New Students in 2024 using ANN.

Analysis Results of Predicting the Number of New Students in 2024 using ANN:

1. In the prediction section, the estimated number of new student admissions expected to join Darma Cendika Catholic University (UKDC) in 2024 is approximately 219.26 or 219 students.
2. The model produces an MSE value of around 0.1046, indicating a low level of prediction error in the Artificial Neural Network model.
3. The RMSE value is approximately 0.32, signifying that the predictions are close to the actual data.

e-ISSN: 2614-8404
p-ISSN:2776-3234

These results suggest that the Artificial Neural Network model provides accurate predictions for the number of new student admissions in 2024.

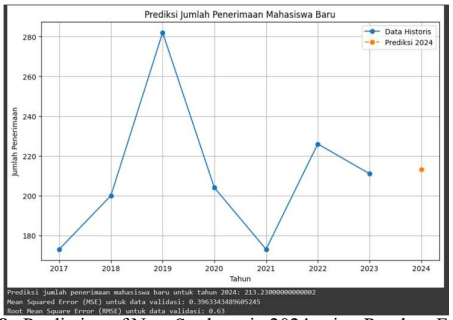


Fig 8. Prediction of New Students in 2024 using Random Forest Regressor.

Analysis Results of Predicting the Number of New Students in 2024 using Random Forest Regressor:

1. In the prediction section, the estimated number of new student admissions expected to join Darma Cendika Catholic University (UKDC) in 2024 is approximately 213.23 or 213 students.
2. The model produces an MSE value of around 0.3963, indicating a low level of prediction error in the Random Forest Regressor model.
3. The RMSE value is approximately 0.63, signifying that the predictions are close to the actual data.

These results suggest that the Random Forest Regressor model provides accurate predictions for the number of new student admissions in 2024.

IV. CONCLUSION

From the analysis results, it can be concluded that the Artificial Neural Network (ANN) model provides a more accurate prediction of the number of new student admissions for the year 2024, estimating 219 students compared to the Random Forest Regressor's estimate of 213 students. The ANN model exhibits a lower prediction error rate with an MSE value of approximately 0.1046 and better prediction accuracy with an RMSE value of around 0.32, compared to the Random Forest Regressor model. However, the difference in error values being higher in the ANN model compared to the Random Forest Regressor with a lower error rate can be attributed to several factors:

1. The dataset has limited data that influences the model's performance, especially the ANN model, which requires a substantial amount of data for effective training.

2. The ANN model is more complex than the Random Forest Regressor, potentially leading to overfitting, whereas the Random Forest Regressor demonstrates robustness due to its use of multiple trees from random subsets.

In this study, it can be concluded that the Artificial Neural Network (ANN) model provides a more accurate prediction of the number of new student admissions for 2024, estimating around 219 students, compared to the Random Forest Regressor model, which predicts approximately 213 students. The ANN model has a lower prediction error rate, as indicated by an MSE value of around 0.1046, and better prediction accuracy, with an RMSE of around 0.32, compared to the Random Forest Regressor model. However, despite the ANN model providing more accurate predictions and having a lower error rate, the possibility of errors cannot be ruled out due to the relatively small amount of data. Thus, the results of this prediction study can provide support for university administration and development in the future.

REFERENCES

[1] R. N. Sedyati, “Perguruan tinggi sebagai agen pendidikan dan agen pertumbuhan ekonomi,” vol. 16, pp. 155–160, 2022, doi: 10.19184/jpe.v16i1.27957.

[2] M. Metode, M. Carlo, K. K. Simulasi, and M. Carlo, “Simulasi prediksi jumlah mahasiswa baru universitas dehasen bengkulu menggunakan metode monte carlo,” vol. VII, 2020, doi:10.33369/pseudocode.7.1.8-16.

[3] A. R. Suleman and I. Palupi, “Penerapan Artificial Neural Network (ANN) untuk Prediksi Prestasi Akhir Mahasiswa Melalui Nilai Mata Kuliah Dasar Tingkat 1,” vol. 10, no. 2, pp. 1849–1859, 2023.

[4] R. Septiriana, A. Perwitasari, and R. Septiriana, “Jurnal Mantik Prediction of the Number of Course Participants Using Random Forest Regression Algorithm,” vol. 6, no. 3, 2022, doi:10.35335/mantik.v6i3.3175.

[5] H. Putra and N. Ulfa, “Jurnal Nasional Teknologi dan Sistem Informasi Penerapan Prediksi Produksi Padi Menggunakan Artificial Neural Network Algoritma Backpropagation,” vol. 02, pp. 100–107, 2020, doi:10.25077/TEKNOSI.v6i2.2020.100-107.

[6] M. E. Rosadi, D. Agustini, M. Farida, and D. D. Anjani, “Analisis Penerapan Neural Network dalam Memprediksi Produksi Bijih Nikel di Indonesia,” vol. 4, no. 1, pp. 40–50, 2022, doi:doi.org/10.30645/brahmana.v4i1.108.

e-ISSN: 2614-8404
p-ISSN:2776-3234

[7] M. R. M, R. D. Putri, M. R. N. R, S. Amin, and M. Akli, "Pengaplikasian Artificial Neural Network (ANN) dalam Memprediksi Curah Hujan Menggunakan Python," pp. 369–373.

[8] I. I. Ridho, C. F. Ramadhani, and A. P. Windarto, "Penerapan Artificial Neural Network dengan Metode Backpropagation Dalam Memprediksi Harga Saham (Kasus : PT . Bank BCA , Tbk) diantaranya oleh A . Santoso dan S . Hansun [1] dalam riset mereka yang sebenarnya . Dalam kesimpulannya , penerapan NN dalam memprediksi harga," vol. 8, pp. 295–303, 2023.

[9] E. Fitri, "JOURNAL OF APPLIED COMPUTER SCIENCE AND TECHNOLOGY (JACOST) Analisis Perbandingan Metode Regresi Linier , Random Forest Regression dan Gradient Boosted Trees Regression Method untuk Prediksi Harga Rumah," vol. 4, no. 1, pp. 58–64, 2023, doi:10.52158/jacost.v4i1.491.

[10] K. Ciptady, M. Harahap, and Y. Ndruru, "Prediksi Kualitas Kopi Dengan Algoritma Random Forest Melalui Pendekatan Data Science," vol. 2, no. 1, 2022, doi:10.47709/dsi.v2i1.1708.

[11] S. Kasus, S. Ciliwung, C. River, and C. Study, "Jurnal Teknologi Lingkungan Prediksi Kualitas Air Sungai Menggunakan Metode Pembelajaran Mesin : Studi Kasus Sungai Ciliwung Prediction of River Water Quality Using Machine Learning Methods : Ciliwung River Case Study," vol. 24, no. 2, pp. 273–282, 2023.

[12] C. Science, V. W. Siburian, and I. E. Mulyana, "Prediksi Harga Ponsel Menggunakan Metode Random Forest," vol. 4, no. 1, pp. 978–979, 2018.

[13] N. A. Riani, R. Andreswari, R. Fauzi, and S. Informasi, "IMPLEMENTASI ALGORITMA ARTIFICIAL NEURAL NETWORK," vol. 4307, no. 3, pp. 241–247, 2021.

[14] M. Adi, P. Hutabarat, M. Julham, A. Wanto, P. Algoritma, and M. Produksi, "Meychael Adi Putra Hutabarat* 1 , Muhammad Julham 2 , Anjar Wanto 3," vol. 4, no. 1, pp. 77–86, 2018.

[15] J. R. Saragih, M. Billy, S. Saragih, and A. Wanto, "ANALISIS ALGORITMA BACKPROPAGATION DALAM PREDIKSI NILAI EKSPOR (JUTA USD)," vol. 15, no. 2, pp. 254–264, 2018, doi:10.23887/jptk-undiksha.v15i2.14362.

[16] R. H. Dananjaya, "PENERAPAN ARTIFICIAL NEURAL NETWORK (ANN) DALAM MEMPREDIKSI," vol. 10, no. 4, pp. 419–426, 2022, doi:10.20961/mateksi.v10i4.65034.

[17] R. Sistem, D. Kartini, F. Abadi, and T. H. Saragih, "Prediksi Tinggi Permukaan Air Waduk Menggunakan Artificial Neural," vol. 1, no. 10, pp. 39–44, 2021, doi:10.29207/resti.v5i1.2602.

[18] G. I. Marthasari, S. A. Astiti, and Y. Azhar, "Prediksi Data Time-series menggunakan Jaringan Syaraf Tiruan Algoritma Backpropagation Pada Kasus Prediksi Permintaan Beras," vol. 6, no. 3, pp. 187–193, 2021, doi:https://dx.doi.org/10.30591/jpit.v6i3.2627.

[19] C. E. Larsen, R. Trip and C. R. Johnson, "Methods for procedures related to the electrophysiology of the heart", U.S. Patent 5,529,067, (1995) June 25.