Logistic Regression Model for Predicting SNBP Admission Based on Academic Data

Daffa Naufal Rahimi
Faculty of Science and Technology
Universitas Islam Negeri Sumatera Utara Medan
Medan, Indonesia
dapa.nopal001@gmail.com

Yusuf Ramadhan Nasution
Faculty of Science and Technology
Universitas Islam Negeri Sumatera Utara Medan
Medan, Indonesia
ramadhannst@uinsu.ac.id

Abstract— The National Selection Based on Achievement (SNBP) is a crucial pathway for prospective students to access higher education; however, the uncertainty surrounding admission outcomes often causes anxiety among prospective students. This study aims to develop an SNBP admission prediction model based on logistic regression using academic data from students at the Darul Arafah Raya Islamic Boarding School. The method used is binary logistic regression with parameter estimation via the Newton-Raphson method. The research data consists of 261 academic records of students from 2022 to 2025, divided into training and testing datasets. Model evaluation was conducted using accuracy, precision, recall, F1 score, and AUC-ROC metrics. The results show that the model achieved convergence at the seventh iteration with an accuracy rate of 81.25 percent. The precision and recall values were 82.35 percent, respectively, while the AUC-ROC value was 0.9049, which falls into the "good classification" category. It can be concluded that the logistic regression model is effective for predicting SNBP graduation based on average report card scores and is suitable for implementation as a decision support system for students in estimating their admission chances.

Keywords— Academic Data, Graduation Prediction, Logistic Regression, Newton-Raphson, SNBP

Article info: Date Submitted: 2026-03-20 | Date Revised: 2026-04-06 | Date Accepted: 2025-04-07

I. INTRODUCTION

The contemporary adoption of technology integrated seamlessly within the structural framework of the current educational system, specifically facilitated through the active use of specialized learning software applications, consistently generates a highly comprehensive digital track record consisting entirely of academic activity data. This resulting comprehensive dataset, which encapsulates student academic activity, can subsequently be analyzed in a rigorous manner by utilizing the specific framework known as the Educational Data Mining approach, widely referred to by its formal acronym, EDM [1]. This EDM approach deliberately applies sophisticated data mining techniques to identify completely hidden patterns residing within the data, accurately project the future academic achievements of the students, and optimize the overall operational effectiveness of ongoing teaching and learning activities [2]. Nevertheless, despite these proven capabilities, a significantly large number of educational institutions have not yet optimally utilized this incredibly valuable data. Consequently, this unutilized data merely ends up functioning as static administrative reports that ultimately serve only to continuously burden the database. This inefficient situation directly triggers the recognized phenomenon described as "data rich but information poor," namely a condition where the incredibly abundant volume of available educational data is simply not directly proportional to the actual quality of meaningful information that can be successfully extracted [3].

The National Selection Based on Achievement, formally designated by its acronym SNBP, distinctly represents an inherently highly competitive state university admission pathway strictly based on the comprehensive achievement track records accumulated by applying students [4]. Considering the challenging reality that the specific graduation indicators utilized for this pathway are not definitively defined [5], the crucial process of

DOI: https://doi.org/10.37396/jsc.v9i1.649

selecting the appropriately eligible candidates, as conducted by the respective schools, is very often based merely on the highly subjective intuition of human selectors alongside basic manual estimation techniques. This strictly conventional, non-automated approach to candidate selection inherently risks triggering severe inaccuracies in the high-stakes process of determining truly eligible candidates. Ultimately, these triggered inaccuracies caused by subjective intuition and manual estimation can fundamentally harm and significantly diminish the graduation chances of those specific students who actually possess genuinely capable academic capacities that rightfully deserve formal recognition.

In a structured effort to actively overcome this prevalent, damaging uncertainty continuously surrounding the entire administrative selection process, it is absolutely imperative and vitally important that modern educational institutions decisively transition to adopt a firmly data-driven decision-making paradigm. Within the boundaries of this specific data-driven paradigm, the machine learning approach clearly emerges as a highly relevant, scientifically sound, and perfectly suited solution explicitly designed to successfully extract those hidden patterns directly from the documented academic track records belonging to previous alumni. Through the careful, precise, and highly controlled implementation of appropriate computational algorithms, the thorough, exhaustive analysis of this extensive historical data is fundamentally able to greatly facilitate high schools in their crucial administrative task of determining properly eligible student candidates. Furthermore, this advanced computational facilitation firmly guarantees that the final determination of eligible student candidates is consistently executed precisely, fairly, and entirely objectively.

The specific machine learning algorithm deemed most exceptionally relevant for achieving this analytical purpose is Logistic Regression, explicitly defined as a methodical statistical approach utilized for accurately modeling the underlying relationship between specific independent predictor variables and corresponding categorical response variables [6]. To properly accommodate entirely different data structures, this versatile algorithm is distinctly divided into three main operational variants: the binary logistic regression variant designed specifically for handling dichotomous responses [7], the multinomial regression variant utilized exclusively for processing nominal responses containing more than two distinct categories [8], and the ordinal regression variant specifically applied to structured, tiered responses [9]. Aligning perfectly with these established variants, this current research specifically focuses on and implements a binary logistic regression model. The core intent of this model implementation is to simultaneously analyze and accurately predict the students' ultimate graduation status within the highly competitive SNBP selection pathway, categorized strictly into two dichotomous outcomes, Passed and Not Passed, by directly utilizing the average academic grades of previous alumni as the predictor variable.

Previous research by Huriyah et al. [10] confirmed the remarkable effectiveness of the binary logistic regression model in accurately predicting the numerical chances of passing the SNMPTN pathway within the Statistics study program. By involving a comprehensive series of designated predictor variables—which include the ranked order of study program preferences, the history of national and provincial achievements, and the calculated average grades across six distinct academic subjects—against a strictly dichotomous graduation status utilized as the designated response variable, the implemented modeling achieved a validated accuracy rate measuring exactly 84.95%, accompanied by a prediction error of 15.05%. This exceptionally high accuracy rate strongly indicates that the binary logistic regression algorithm is highly reliable to be replicated for predicting students' probability of passing the SNBP selection pathway by systematically utilizing the historical data consisting of alumni's average academic grades.

Furthermore, reinforcing this algorithmic choice, a comparative study executed by Wardana et al. [11] systematically evaluated the predictive performance of three distinct machine learning algorithms, namely the Decision Tree, Naïve Bayes, and Logistic Regression methodologies, in strictly predicting state university entrance selections. The

DOI: https://doi.org/10.37396/jsc.v9i1.649

conclusive research results absolutely confirmed that Logistic Regression is unequivocally the most superior predictive model with a documented 85% accuracy rate, successfully surpassing the Decision Tree at 78% and Naïve Bayes at 71%. Consequently, this specific research aims to meticulously implement a fully functional binary logistic regression model to precisely predict the statistical probability of student eligibility in the SNBP selection pathway. The application of this structured modeling is deliberately designed to directly facilitate educational institutions, specifically guidance and counseling teachers, in determining eligible candidates precisely and objectively. The intentional selection of this algorithm relies completely on undeniable empirical evidence from various previous studies confirming the reliability level and superior performance of logistic regression in accurately predicting university entrance selection results.

II. METHOD/MATERIAL

A. Logistic Regression

Logistic regression is a statistical model used to analyze the relationship between one or more independent variables and a single categorical dependent variable. This model aims to test the significance of the simultaneous and partial effects of independent variables on the dependent variable, as well as to predict the probability of an event occurring based on given values of the independent variables [12]. If the dependent variable has two categories, the technique used is called binary logistic regression. Meanwhile, if the dependent variable has more than two categories, the method used is multinomial logistic regression [13]. This study uses a binary logistic regression model. The mathematical equation for the binary logistic regression model is expressed as follows [14].

$$g(x) = \beta_0 + \beta_1 X_n \qquad (1)$$

B. Newton-Raphson Method

The Newton-Raphson method is a numerical method used to find the roots of nonlinear equations. This method starts with an initial guess and approximates the root of the equation by using the derivative at that point. This process is performed iteratively to estimate the root of the equation [15]. The Newton-Raphson algorithm involves the following steps [16]:

1. Determine the initial value x_0.
2. Calculating the first derivative $f'(x)$.
3. Applying the iterative formula $x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$.
4. Repeat the process until convergence is achieved.

C. Python

Python is a programming language that uses an interpreter to execute its code. The interpreter can translate the code directly, and Python can run on various platforms, such as Windows, Linux, and others [17]. Using Python offers advantages in terms of data processing speed, flexibility, analysis, and accurate visualization capabilities. Tools such as pandas, numpy, statsmodels, and matplotlib are used to perform data cleaning, statistical calculations, logistic regression modeling, and the presentation of analysis results [18].

D. Research Framework

DOI: https://doi.org/10.37396/jsc.v9i1.649

Figure 1. Research Framework

1) Data Collection

Data collection was conducted among students at the Darul Arafah Raya Islamic Boarding School who had registered for the SNBP from 2022 to 2025. The academic data collected included the students' names, average report card grades, SNBP admission status, and year of registration. However, the variables used in this study were average report card grades and SNBP admission status, with average report card grades as the independent variable and SNBP admission status as the dependent variable. Based on the data collection process, a total of 261 student academic records were obtained.

2) Preprocessing

The collected data then underwent preprocessing to improve data quality and generate an accurate predictive model. At this stage, all data types in the raw data were converted to numerical values to facilitate the analysis process, both during training and testing. Next, data rows containing missing values were removed so that only data with complete values for both variables were used. Variables not required for modeling—namely, student names and SNBP registration years—were removed to streamline the dataset. The entire preprocessing process aimed to ensure the quality of the data used in building the predictive model.

3) Data Splitting

After undergoing the preprocessing stage, the data was divided into two groups: training data and testing data. This division aims to build an accurate predictive model and test the model's performance on previously unseen data. The training data was drawn from the academic records of students at the Darul Arafah Raya Islamic Boarding School for the period from 2022 to 2024, while the testing data was drawn from the year 2025. The data splitting process resulted in a composition of 197 data points for training and 64 data points for testing.

4) Model Training

The training data obtained from data splitting is used during the model training phase to estimate the coefficients of the logistic regression model. These coefficients form the logistic regression equation, which is applied to the testing data to predict SNBP admission. The estimation process is performed iteratively using the Newton-Raphson method on the training data to obtain optimal values for the intercept, coefficients, and odds ratio, thereby producing an accurate predictive model.

5) Evaluation

After the logistic regression model coefficients were obtained and the model was established, a testing and evaluation process was conducted to measure the model's reliability and performance in predicting the SNBP admission results for students at the Darul Arafah Raya Islamic Boarding School. The testing involved predicting SNBP admission outcomes on the test data using the logistic regression model developed during the training phase. The prediction results were then evaluated using a confusion matrix and several evaluation metrics, namely accuracy, precision, recall, F1-score, and AUC-ROC.

DOI: https://doi.org/10.37396/jsc.v9i1.649

III. RESULT AND DISCUSSION

This study utilized academic data from 261 students at the Darul Arafah Raya Islamic Boarding School who had registered for the SNBP from 2022 to 2025. The data underwent a preprocessing stage, during which unnecessary variables—such as student names and registration years—were removed, leaving only the independent variable (average report card scores) and the dependent variable (SNBP graduation status). Additionally, all missing values in both variables were removed to improve data quality. Subsequently, the data was divided into two parts: training data and testing data. The training data covers students' academic records from 2022 to 2024, while the testing data covers academic records from 2025. The training and testing data samples are presented in Tables 1 and 2 below.

Table 1. Training Data Sample

Average Score	Graduation Status
92.11	0
89.73	0
90.32	0
93.69	0
91.20	0
…	…
97.95	1
94.06	1
93.05	0
98.39	1
92.12	0

Table 2. Testing Data Sample

Average Score	Graduation Status
92.79	0
90.99	0
91.55	0
93.02	1
96.14	0
…	…
98.15	1
99.00	1
96.34	1
94.47	1
95.51	1

After the training and testing data were split, the next step was to estimate the parameters of the logistic regression model using the Newton-Raphson method. This method was applied iteratively to the training data to obtain the optimal intercept and regression coefficients. The algorithm was implemented using the Python programming language. Based on the computational results, model convergence was achieved at the 7th iteration. Details of the intercept, coefficients, log-likelihood, and changes in the intercept value during the iteration process are presented in Table 3.

Table 3. The Newton-Raphson Convergence Process

Iteration	β_0	β_1	Log-Likelihood	$\lvert\Delta\beta_0\rvert$
1	-52.133914	0.549999	-136.549995	52.133914
2	-83.483624	0.879932	-71.171380	31.349710
3	-104.477355	1.100737	-60.041160	20.993731

DOI: https://doi.org/10.37396/jsc.v9i1.649

4	-111.637127	1.176032	-57.692441	7.159772
5	-112.257371	1.182554	-57.466593	0.620244
6	-112.261505	1.182598	-57.458040	0.004134
7	-112.261954	1.182603	-57.457988	0.000000

Based on Table 3, it can be seen that the intercept and regression coefficients reached an optimal point at the 7th iteration. This is indicated by the value of $|\Delta\beta_0|$ approaching zero at that iteration. In addition, the log-likelihood value also showed stability from the 4th to the 7th iteration. Based on these results, the intercept, coefficients, and odds ratios presented in Table 4 were obtained. Based on these parameters, the logistic regression model equation is obtained as follows.

$$z = -112{,}261954 + 1{,}182603 \times X \quad (2)$$

$$p = \frac{1}{1+e^{-z}} \quad (3)$$

Table 4. Logistic Regression Coefficient

Logistic Regression Coefficient	Score	Description
β_0 (Intercept)	-112.261954	Konstanta model
β_1 (Slope)	1.182603	Koefisien nilai rata-rata
Odds Ratio	3.262855	e^{β_1}

After the logistic regression coefficients and model equation were successfully obtained through the model training phase, the next step was to implement the equation to predict SNBP pass rates. The implementation was carried out using a threshold value of 0.5, where data with a probability (p) value less than 0.5 was predicted to fail, while data with a p value greater than or equal to 0.5 was predicted to pass. A sample of the prediction results along with the z and p values for the test data is presented in Table 5.

Table 5. Sample of Predicted Data Testing Results

Average Score (X)	z	p	Predicted Results
92.79	-2.5282	0.0739	0
90.99	-4.6569	0.0094	0
91.55	-3.9946	0.0181	0
93.02	-2.2562	0.0948	0
96.14	1.4335	0.8074	1
…	…	…	…
98.15	3.8105	0.9783	1
99.00	4.8157	0.9920	1
96.34	1.6700	0.8416	1
94.47	-0.5414	0.3679	0
95.51	0.6885	0.6656	1

After implementing the logistic regression model on the test data, the next step is to evaluate the model's reliability and performance in predicting the SNBP admission results of students at the Darul Arafah Raya Islamic Boarding School. The evaluation was conducted using a confusion matrix and several metrics, namely accuracy, precision, recall, F1-score, and AUC-ROC. The results of the confusion matrix are presented in Figure 2.

DOI: https://doi.org/10.37396/jsc.v9i1.649

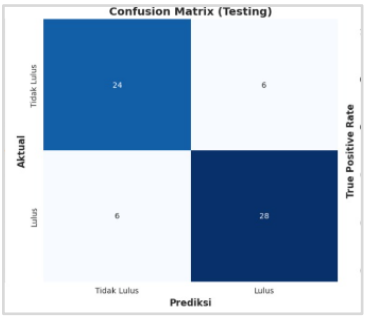

Figure 2. Confusion Matrix Results

Based on the confusion matrix in Figure 2, the results show 28 true positives, 24 true negatives, 6 false positives, and 6 false negatives. Based on these values, other evaluation metrics were calculated and are presented in Figure 3.

METRIK EVALUASI:

Metrik	Training	Testing
Accuracy	0.8731	0.8125
Precision	0.8481	0.8235
Recall	0.8375	0.8235
F1-Score	0.8428	0.8235
Log Loss	0.2917	0.3853
AUC-ROC	0.9470	0.9049

Figure 3. Model Evaluation Results

Based on the results of the classification model evaluation using a confusion matrix on the test data, an accuracy of 81.25% was obtained. This value indicates that the model has a good level of accuracy in classifying the data overall. The precision and recall values, at 82.35% each, indicate that the model is capable of making positive predictions with a relatively low error rate and has a good ability to accurately identify all positive data. Additionally, an F1-score of 82.35% demonstrates a good balance between precision and recall, meaning the model is not biased toward either type of error (false positives or false negatives). An AUC-ROC value of 0.9049 indicates that the model has a strong ability to distinguish between positive and negative classes. This score falls within the "good classification" category, meaning the model is capable of providing accurate prediction probabilities in distinguishing between 'Pass' and "Fail" data. Overall, the developed model performs well and is suitable for use as a graduation prediction tool, although there is still room for improvement to minimize classification errors.

IV. Conclusion

Based on the research findings and discussion, it can be concluded that the binary logistic regression model is effective for predicting success in the National Selection Based on Achievement (SNBP) using the average report card scores of students at the Darul Arafah Raya Islamic Boarding School. The parameter estimation process using the Newton-Raphson method converged at the 7th iteration, yielding an optimal regression equation. Model evaluation on the testing data showed good performance with an accuracy rate of 81.25%, precision and recall values of 82.35% respectively, and an AUC-ROC value of 0.9049, which falls into the "good classification" category.

Thus, this model is suitable for implementation as a decision support system to help students estimate their chances of passing the SNBP before official registration takes place. However, this study still has limitations because it only utilizes academic variables (report card grades). For future research, it is recommended to integrate non-academic variables

DOI: https://doi.org/10.37396/jsc.v9i1.649

such as competition achievements or school accreditation, as well as to compare it with other machine learning algorithms to improve the robustness and accuracy of the predictions.

ACKNOWLEDGEMENT

The author would like to express gratitude to the State Islamic University of North Sumatra (UINSU) in Medan for providing facilities and academic support throughout this research process. Gratitude is also extended to the Darul Arafah Raya Islamic Boarding School for granting permission and access to academic data on students from 2022 to 2025, enabling the completion of this research. The author also appreciates the contributions of colleagues and journal reviewers who provided valuable feedback to enhance the quality of this article.

REFERENCES

[1] S. A. A. Kharis and A. H. A. Zili, "Learning Analytics dan Educational Data Mining pada Data Pendidikan," *Jurnal Riset Pembelajaran Matematika Sekolah*, vol. 6, no. 1, pp. 12–20, Mar. 2022, doi: 10.21009/jrpms.061.02.

[2] M. Yağcı, "Educational data mining: prediction of students' academic performance using machine learning algorithms," *Smart Learning Environments*, vol. 9, no. 1, p. 11, Dec. 2022, doi: 10.1186/s40561-022-00192-z.

[3] C. Chandra, I. Fenriana, and R. Rimbawan, "Implementasi Data Mining Untuk Mengetahui Pola Pembelian Pelanggan Pada Produk Vin's Cafe Dengan Algoritma Apriori Dan Pengujian Kualitas Melalui Metode ISO 9126," *ALGOR*, vol. 4, no. 1, pp. 11–20, Sep. 2022, doi: 10.31253/algor.v4i1.1538.

[4] Supangat and Rafif Giovanni, "Evaluasi Tingkat Persaingan Siswa Dalam Seleksi Nasional Masuk Perguruan Tinggi Negeri Menggunakan Algoritma Naive Bayes," *Journal of Scientech Research and Development*, vol. 6, no. 1, pp. 1055–1068, Jul. 2024, doi: 10.56670/jsrd.v6i1.337.

[5] I. Hanum, Y. Sholva, H. Sastypratiwi, and F. Asrin, "Model Prediksi Keketatan Lolos SNMPTN Menggunakan Algoritma K-Nearest Neighbor," *Jurnal Ilmiah ILKOMINFO - Ilmu Komputer & Informatika*, vol. 6, no. 2, pp. 179–190, Jul. 2023, doi: 10.47324/ilkominfo.v6i2.205.

[6] A. Tripena, R. Maharsi, Y. Lianawati, and A. A. Setyawan, "Analisis Faktor-Faktor Yang Mempengaruhi Kemiskinan Rumah Tangga Di Desa Kotayasa Melalui Pendekatan Regresi Logistik Biner," *Jurnal Elektro Luceat*, vol. 9, no. 2, 2023.

[7] N. K. Hasibuan, S. Dur, and I. Husein, "Faktor Penyebab Penyakit Diabetes Melitus dengan Metode Regresi Logistik," *G-Tech: Jurnal Teknologi Terapan*, vol. 6, no. 2, pp. 257–264, Sep. 2022, doi: 10.33379/gtech.v6i2.1696.

[8] R. Prabowo, H. Sujaini, and T. Rismawan, "Analisis Sentimen Pengguna Twitter Terhadap Kasus COVID-19 di Indonesia Menggunakan Metode Regresi Logistik Multinomial," *Jurnal Sistem dan Teknologi Informasi (JustIN)*, vol. 11, no. 2, p. 366, Jul. 2023, doi: 10.26418/justin.v11i2.57449.

[9] M. D. Purnama and A. Sofro, "Pemodelan Faktor-Faktor Yang Mempengaruhi Indeks Pembangunan Manusia Jawa Timur Dengan Regresi Logistik Ordinal," *MATHunesa: Jurnal Ilmiah Matematika*, vol. 12, no. 3, pp. 654–661, Jul. 2024, doi: 10.26740/mathunesa.v12n3.p654-661.

[10] N. Satyahadewi, R. Tamtama, H. Perdana, and S. K. Huriyah, "Binary Logistics Regression To Predict The Opportunity Of SNMPTN Graduation In Statistics Study Program Of Tanjungpura University," *Mathline : Jurnal Matematika dan Pendidikan Matematika*, vol. 8, no. 1, pp. 17–28, Feb. 2023, doi: 10.31943/mathline.v8i1.269.

[11] O. Y. Wardana, M. Ayub, and A. Widjaja, "Perbandingan Akurasi Model Pembelajaran Mesin untuk Prediksi Seleksi Masuk Perguruan Tinggi Negeri," *Jurnal Teknik Informatika dan Sistem Informasi*, vol. 9, no. 1, Apr. 2023, doi: 10.28932/jutisi.v9i1.6126.

[12] E. Roflin, F. Riana, E. Munarsih, Pariana, and I. A. Liberty, *Regresi Logistik Biner dan Multinomial* . Pekalongan: PT. Nasya Expending Management, 2023.

[13] A. Ermillian and K. Nugroho, "Perancangan Model Deteksi Potensi Siswa Putus Sekolah Menggunakan Metode Logistic Regression Dan Decision Tree," *Jurnal Informatika: Jurnal Pengembangan IT*, vol. 9, no. 3, pp. 281–295, Dec. 2024, doi: 10.30591/jpit.v9i3.8007.

[14] R. herina Situngkir and P. Sembiring, "Analisis Regresi Logistik untuk Menentukan Faktor-Faktor yang Mempengaruhi Kesejahteraan Masyarakat Kabupaten/Kota di Pulau Nias," *FARABI: Jurnal Matematika dan Pendidikan Matematika*, vol. 6, no. 1, pp. 25–31, May 2023, doi: 10.47662/farabi.v6i1.432.

[15] L. A. Mukaromah and M. R. Atsani, "Penerapan Metode Bisection Dan Newton-Raphson Untuk Penyelesaian Akar Persamaan Non-Linier Menggunakan Matlab," *Jurnal Teknik Informatika dan Sistem Informasi*, vol. 4, no. 2, 2024.

[16] N. Rohmawati, I. S. N. Inayah, and A. Wibowo, "Perbandingan Penggunaan Pythondan Exceldalam Menyelesaikan Persamaan Tak Linier Metode Newton Raphson," *Numerical: Jurnal Matematika dan Pendidikan Matematika*, vol. 9, no. 1, 2025.

[17] S. Rahman *et al.*, *Python: Dasar dan Pemrograman Berorientasi Objek*. Tahta Media Group, 2023.

[18] H. Anugrah, V. T. Pasau, Jufri, and Asran, "Analisis Faktor Yang Mempengaruhi Pencapaian Akreditasi Unggul Pada Program Studi Teknik Informatika Menggunakan Regresi Logistik," *Jurnal DIPAKOMTI*, vol. 17, no. 2, 2026.

DOI: https://doi.org/10.37396/jsc.v9i1.649