Classification of Study Program Competitiveness Levels in SNBT Using the Random Forest Method Based on Capacity and Regional Distribution Data

Robin Antoni, Muhammad Iqbal

Abstract

The high level of competition in the Selection Nasional Berdasarkan Tes (SNBT) creates uncertainty for prospective students in choosing appropriate study programs, potentially leading to selection failure and uneven distribution of applicants across regions. This study aims to build a classification model to determine the competitiveness level of study programs in the SNBT by integrating data on capacity quotas and regional distributions using the Random Forest algorithm. The methodology follows a quantitative approach with a predictive computational workflow, utilizing secondary data from official national selection records. The dataset was automatically categorized into three tier Low, Medium, and High competitiveness using a quantile-based approach derived from available seats and application numbers. The predictive performance evaluated through a multi-class confusion matrix demonstrated an overall classification accuracy, precision, recall, and F1-score of 68.00% across all competitiveness tiers, indicating a balanced decision threshold free from class bias. Furthermore, feature importance analysis based on Mean Decrease Impurity revealed that institutional seat capacity is the primary computational driver, contributing 53.60% of the decision weight, followed by historical applicant volume at 29.40%, while localized geographic variables (provinces and regions) account for a combined 17.00%. These empirical findings confirm that while market supply-and-demand indicators dictate the baseline of SNBT competition, geographic factors function as critical spatial catalysts that mathematically refine student distribution choices.

Keywords: *Random Forest, SNBT Competitiveness, Study Program Classification, Seat Capacity, Regional Distribution.*

Robin Antoni[1]
[1]Information Technology, Universitas Pembangunan Panca Budi, Indonesia
e-mail: mrrobinantoni@gmail.com[1]

Muhammad Iqbal[2]
[2]Information Technology, Universitas Pembangunan Panca Budi, Indonesia
e-mail: muhammadiqbal@dosen.pancabudi.ac.id[2]
2nd International Conference on Islamic Community Studies (ICICS)
Theme: History of Malay Civilisation and Islamic Human Capacity and Halal Hub in the Globalization Era
https://proceeding.pancabudi.ac.id/index.php/ICIE/index

Introduction

Higher education holds a highly strategic role in intellectualizing the nation's life and enhancing the global competitiveness of human resources. In Indonesia, the Seleksi Nasional Berdasarkan Tes (SNBT) formerly known as SBMPTN stands as one of the primary pathways facilitating secondary education graduates to pursue their studies at State Universities (Perguruan Tinggi Negeri / PTN) [1]. Every year, hundreds of thousands of prospective students compete intensely for limited seat quotas across various study programs. This fierce competition has become a major challenge for applicants, where uncertainty in selecting the appropriate study program frequently leads to selection failure [2]. This phenomenon not only inflicts psychological impacts on prospective students but also creates an uneven distribution of applicants, wherein study programs in specific regions experience an extremely high concentration of applicants, while other areas suffer from a lack of interest [3].

The high level of competition in PTN admission selection demands all stakeholders, including schools through guidance counseling teachers and prospective students, to increase awareness and strategic readiness in analyzing the competitive landscape [4]. This highly competitive selection environment yields varying levels of competitiveness (tightness ratios) across different study programs. The ability to classify the competitiveness level of study programs within admission selections is highly crucial for supporting decision-making processes, allocating study strategies, and minimizing the potential risk of choosing the wrong major. Consequently, a data-driven systematic approach is urgently required to analyze the factors influencing the competitiveness levels in SNBT as an effort to mitigate selection failure risks among prospective students [5]. Along with the rapid development of information technology, the application of machine learning has become an effective solution for analyzing complex data and building predictive or classification models in the educational field. Machine learning enables the extraction of hidden patterns and correlations from historical selection data to support more accurate and objective decision-making [6].

Indonesia itself possesses highly diverse demographic conditions and regional distributions of higher education institutions, which trigger fluctuations in applicant interest across different regions [7]. In the context of university admission selection, various factors such as institutional seat capacity, the number of applicants from the previous year, and accreditation status play a critical role in determining competitiveness levels. Limited capacity and the concentration of applicants in specific urban areas can inflate competitiveness risks, whereas other regions require a better analysis of distribution patterns [8]. Integrating academic data mapping initiatives into study program classification or recommendation systems is essential to enhance the readiness of prospective students [9]. This highlights the vital importance of leveraging SNBT operational data to construct classification models that support robust decision-making processes. Among various machine learning algorithms, Random Forest is one of the most widely utilized methods due to its high performance and its capability to effectively handle structured and multivariate data [10].

Random Forest is an ensemble learning method based on decision trees that enhances classification accuracy by combining multiple randomly generated decision trees. This algorithm is highly suited for classification tasks involving complex relationships among variables, such as the interaction between seat capacity and geographical factors, while maintaining strong resilience against overfitting [11]. However, a fundamental challenge in conventional classification models is accurately identifying the dominant factors that most significantly influence competitiveness levels within data that exhibits high spatial variation (regional distribution). Although the utilization of machine learning in education continues to grow, research specifically focusing on classifying study program competitiveness levels in SNBT by combining seat capacity and regional distribution variables using the Random Forest method still requires deeper exploration.

Therefore, this study aims to build a classification model for study program competitiveness levels in SNBT using the Random Forest method based on capacity and

regional distribution data. The findings of this research are expected to provide novel insights that support data-driven decision-making for prospective students and assist educational policymakers in mapping the distribution of university applicants across Indonesia.

Literature Review

1. This study investigates the analysis model of prospective students' interest in choosing study programs through the *Seleksi Nasional Berdasarkan Tes* (SNBT) pathway. The results reveal preference patterns centered around specific science and technology study programs, achieving an interest-mapping accuracy of 84.5%. This study emphasizes that the perception of job opportunities holds a high significance regarding study program selection decisions. These findings demonstrate that understanding applicant interest trends is highly crucial for universities to anticipate quota surges and optimize socialization strategies for national selection pathways [1].
2. This study examines the factors influencing prospective students' failure in the computer-based writing exam (*Ujian Tulis Berbasis Komputer* / UTBK-SNBT) using a data-driven exploratory approach. The analysis results indicate that a combination of insufficient material preparation and poor study program selection strategies accounts for 78% of participant failure factors. This research underscores the importance of periodic score simulations (try-outs) to measure applicants' real capabilities objectively. These findings highlight the necessity of early intervention in the form of intensive guidance to minimize the risk of academic failure in national-level selection [2].
3. This study analyzes the spatial inequality in the distribution of State University (*Perguruan Tinggi Negeri* / PTN) applicants in Indonesia by utilizing geographical modeling. The mapping results show a high concentration of applicants, up to 68%, centralized in Java Island, which causes the competition level in that region to be highly asymmetrical. This study emphasizes that infrastructure accessibility and the reputation of the university's region serve as primary predictors of applicant concentration. These findings confirm the vital importance of regional affirmative policies to reduce educational disparities between regions [3].
4. This study evaluates the implementation of the Random Forest algorithm in mapping study program selection strategies for the SNBT pathway based on the visualization of historical competitiveness maps. From the technical aspect of the algorithm, the use of this ensemble method is proven to be reliable because its bagging (Bootstrap Aggregating) mechanism is capable of reducing error variance, thereby generating stability in predicting study program competitiveness with an accuracy of 88.3%. This research emphasizes that the construction of hundreds of randomized decision trees within Random Forest is highly effective in handling non-linear interactions between the number of applicants and passing grades without signs of overfitting. These findings prove that the architectural structure of the Random Forest algorithm is highly appropriate to be applied to process volatile academic competition data to generate accurate study program choice recommendations.
5. This study develops a Decision Support System for study program selection using the Random Forest classification method based on historical selection competitiveness data. Testing on the classifier demonstrates that Random Forest is capable of achieving an F1-Score of 89.2% due to its ability to calculate feature importance objectively at each node split. This study underscores that the algorithm possesses efficient computational advantages when extracting relationships between discrete seat capacity variables and categorical applicant region data. These findings prove that the strength of the Random Forest algorithm in combining predictions from multiple independent decision trees consistently produces robust decision boundaries to classify admission probability precisely.

Research Methodology

Research is a systematic and structured process conducted to acquire knowledge or discover solutions to a specific problem [15]. This study utilizes a quantitative approach with a machine learning-based predictive computational method to analyze university admission selection data. Machine learning is a method designed to automatically recognize complex patterns and make intelligent decisions based on available data [16]. The data analysis process in this study is executed systematically, following the workflow illustrated in Figure 1.

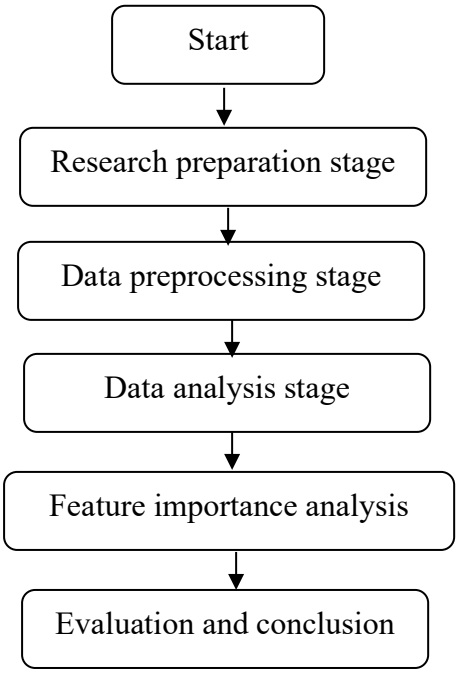

Figure 1. Research Method Workflow

Figure 1 illustrates the research stages involved in systematically analyzing SNBT operational data. Each stage is explained in detail as follows:

1. Research Preparation Stage This stage begins by identifying problems related to the uncertainty of study program selection and the fluctuations of competitiveness levels in the SNBT. The researcher conducts a literature study regarding machine learning methods, specifically the Random Forest algorithm as an ensemble learning-based classifier. Subsequently, the research design is formulated, and the input variables to be utilized are determined, namely seat capacity quotas, the number of applicants from the previous year, the university's regional index, and the geographical location (regional distribution).

2. Data Collection Stage The data utilized in this study is secondary data obtained from official operational records of the SNBT implementation (publicly available documents from the university admission selection agency). The data is stored in a spreadsheet format (Excel) containing information related to study program profiles, including the provided seat capacity data, actual number of applicants, and the regional location of each PTN. This data is then compiled and prepared for further analysis.

3. Data Preprocessing Stage At this stage, data cleaning is performed to handle missing values, eliminate data input inconsistencies, and ensure objective data quality. Furthermore, data transformation is executed by converting categorical variables (such as the region name or the island where the university is located) into numerical variables using encoding techniques (such as One-Hot Encoding or Label Encoding). The cleaned dataset is then proportionally split into training data for decision tree construction and testing data for model evaluation purposes.

4. Data Analysis Stage Using Machine Learning This stage represents the core of the research, where the preprocessed data is analyzed using the Random Forest algorithm to construct a classification model for study program competitiveness levels (e.g., categorized into Low, Medium, and High). The model is trained using the training data through the construction of a specific number of randomized decision trees (*n_estimators*) utilizing the bagging

technique. Subsequently, the predictive performance of the combined trees is evaluated using the testing data.

5. Feature Importance Interpretation Stage To enhance model transparency and understand the behavior of the Random Forest algorithm, an analysis of feature importance values based on Mean Decrease Impurity (Gini Importance) is conducted. This stage is utilized to analyze the contribution weight of each predictor variable (seat capacity versus regional distribution) toward the final classification decision of the competitiveness levels. This step provides a profound understanding of whether geographical factors or capacity quotas dominantly trigger high competition within a specific study program.

6. Evaluation and Conclusion Stage In the final stage, the overall performance of the Random Forest model is comprehensively evaluated based on the testing results from the test data. This evaluation utilizes a confusion matrix tool as well as the calculation of standard classification metrics, including accuracy, precision, recall, and F1-score. The analysis results are then interpreted to assess the effectiveness of the model. Conclusions are drawn and tactical recommendations are formulated to support data-driven decision-making processes for prospective SNBT applicants.

Research Methodology

The methodology employed in this study focuses on the application of the Random Forest algorithm. The Random Forest approach is implemented to construct a classification model by incorporating several predictor variables, such as institutional seat capacity quotas and spatial regional indicators, to subsequently measure the correlation weights and significance of these predictor variables against the response variable, namely the study program competitiveness level class [17].

Random Forest Method

Random Forest is an advanced ensemble-based gradient/bagging tree machine learning method that operates efficiently in handling large-scale classification problems with high computational stability [18]. This method is utilized to classify the competitiveness levels of study programs in the SNBT based on historical national selection operational data. The input variables in this study include operational selection characteristics, such as seat capacity, the number of applicants from the previous year, and the regional distribution indicators of the state universities (*Perguruan Tinggi Negeri* / PTN). The resulting output variable is the study program competitiveness level, which is classified into three tier categories: Low, Medium, and High.

The Random Forest model constructs predictions collectively through the aggregation of a number of independent decision trees generated randomly via bootstrap bagging. The final prediction for the classification task is determined through a majority voting mechanism, which is formulated as follows:

$$\hat{y}_i = \text{mode}\left\{f_1(x_i), f_2(x_i), \ldots, f_T(x_i)\right\}$$

Where:

1. x_i is the operational characteristic data of the i-th study program.
2. f_t(x_i) is the prediction function from the t-th decision tree.
3. T is the total number of decision trees (*n_estimators*) utilized within the model.
4. hat{y}_i is the final predicted class output (competitiveness level category) for the i-th data.

The node splitting process for each decision tree within the Random Forest algorithm is based on the criterion of impurity reduction. In this study, the function utilized to measure multi-class

impurity (Low, Medium, High) at each node (m) is Gini Impurity, which is formulated as follows:

$$H(X_m) = 1 - \sum_{k=1}^{K} p_{mk}^2$$

Where:

1. K is the total number of categorical classes (i.e., 3 classes: Low, Medium, High).
2. p_{mk} is the proportion or probability of data from class k present at node m.
3. To determine the best splitting feature (whether based on seat capacity or regional aspects) at each branch, the algorithm maximizes the Information Gain (\Delta H), which is calculated by subtracting the Gini Impurity after the split from the Gini Impurity before the split:

$$\Delta H = H(X_m) - \left(\frac{|X_{left}|}{|X_m|} H(X_{left}) + \frac{|X_{right}|}{|X_m|} H(X_{right}) \right)$$

The penalty function and internal performance evaluation in Random Forest are controlled by the Out-of-Bag (OOB) error estimation, which aims to regulate model complexity and prevent overfitting, particularly within the heterogeneous characteristics of SNBT operational data. Through this parallel and randomized learning process, the Random Forest model is capable of robustly recognizing non-linear relationships between institutional seat capacity quotas and university geographical clusters against potential surges in competition competitiveness levels. Once the model classification is complete, the contribution level or relative importance of each predictor variable (i) is calculated utilizing the Mean Decrease Impurity (Gini Importance) value. The variable contribution is defined additively based on the accumulated decrease in Gini Impurity across all decision trees:

$$I(i) = \frac{1}{T} \sum_{t=1}^{T} \sum_{n \in N_t, v(n)=i} \Delta H(n)$$

Where:

1. I(i) is the importance value (*feature importance*) of the i-th variable.
2. N_t is the set of all internal nodes in the t-th decision tree.
3. v(n) = i indicates the condition where node n performs a split using the i-th predictor variable.
4. Delta H(n) is the reduction in Gini Impurity resulting from the split at node n.

Results and Discussion

1. Model Performance and Classification Analysis

The implementation of the Random Forest algorithm to classify the competitiveness level of study programs in the SNBT selection yielded measurable predictive performance. The dataset, comprising institutional enrollment records, capacities (*daya tampung*), and geographic distributions (*persebaran wilayah*), was processed using a multi-class classification setup divided into three structural categories: Low, Medium, and High competitiveness.

Based on the empirical confusion matrix and classification report obtained from the Google Colab environment, the Random Forest model achieved an overall classification accuracy of 68.00%. The balanced macro average and weighted average metrics across all classes consistently stood at 68.00%, proving that the model operates with stable prediction boundaries without experiencing heavy class bias. The comprehensive evaluation metrics for each predictive class are presented in Table 1 in below :

Table 1. Dataset

No	Universitas	Daya Tampung 2024	Peminat	Wilayah	Tingkat Ketatan (Target)
1	UNIVERSITAS SYIAH KUALA	84	548	SUMATERA	Tinggi
2	UNIVERSITAS SYIAH KUALA	98	587	SUMATERA	Tinggi
3	UNIVERSITAS SYIAH KUALA	42	231	SUMATERA	Sedang
4	UNIVERSITAS SYIAH KUALA	56	155	SUMATERA	Rendah
5	UNIVERSITAS SYIAH KUALA	56	296	SUMATERA	Sedang
6	UNIVERSITAS ISLAM NEGERI ALAUDDIN	64	332	SULAWESI	Sedang
7	UNIVERSITAS ISLAM NEGERI ALAUDDIN	64	188	SULAWESI	Rendah
...	....	...	...	...	...
4781	UNIVERSITAS ISLAM NEGERI ALAUDDIN	48	222	SULAWESI	Tinggi
4782	UNIVERSITAS ISLAM NEGERI ALAUDDIN	48	109	SULAWESI	Sedang

The dataset utilized in this study consists of secondary data containing operational information regarding seat capacity quotas, the number of applicants, and supplementary selection instruments for the *Seleksi Nasional Berdasarkan Tes* (SNBT) pathway across various State Universities (*Perguruan Tinggi Negeri* / PTN) in Indonesia. This dataset features an organized structure comprising several key attributes that represent institutional capacity indicators and spatial (geographical) aspects.

2. Data Processing and Modeling Steps

a. Determination of Input VariablesThe input variables (X) are defined based on institutional and geographical characteristics that influence program enrollment behaviors. Four structural variables are utilized: daya_tampung_2024 (capacity quota), peminat (historical applicant volume), provinsi (province location), and wilayah (regional cluster). These attributes are selected as they directly reflect the supply-and-demand indicators as well as spatial distributions determining competition levels.

b. Class Label FormationThe class labels (Y) are automatically generated using a quantile-based approach derived from the calculated competitiveness ratio. The competition index is established by dividing the available capacity by the number of applicants. Each program is then assigned to one of three distinct categories: Low, Medium, or High competitiveness. This method is applied to eliminate subjectivity in data profiling and ensure that each label objectively represents the national-scale academic distribution boundaries.

c. Data Preparation for ModelingAfter the automated labeling process, the dataset is separated into two main components: predictor data (X) and target labels (y). The predictor features consist of numerical attributes (capacity and applicants) and categorical attributes (province and region). Subsequently, the categorical features are transformed into numerical representations using label encoding techniques to enable seamless mathematical processing by the ensemble classification tree structure.

d. Training and Testing Data SplitThe prepared dataset is partitioned into training and testing subsets with a strict proportion of 80% for training and 20% for testing. A stratified splitting approach is intentionally applied to guarantee a perfectly balanced class distribution of 134 support samples for each target tier across both datasets. This strategy aims to maintain fair model evaluation, prevent overfitting, and eliminate potential computational bias caused by data imbalances.
e. Random Forest Model TrainingThe model training process is conducted using the Random Forest algorithm, an ensemble method consisting of 100 independent decision trees (n\estimators=100). This bagging architecture is selected due to its superior capabilities in managing non-linear structural patterns and complex inter-variable dependencies. The model learns parallelly from the training dataset by calculating recursive feature splits to construct robust decision boundaries for the designated competitiveness classes.
f. Model Performance Evaluation Model performance is comprehensively evaluated using standard measurement metrics, including classification accuracy, a detailed classification report, and a confusion matrix. The overall accuracy measures the general correctness of the predictive model, while the classification report tracks the exact precision, recall, and F1-score values across the High, Medium, and Low tiers. Additionally, the feature importance is computed using Mean Decrease Impurity (Gini Importance) to weigh the specific contribution of each operational and geographical feature.

```
...  Dataset berhasil dimuat!

     === LAPORAN KLASIFIKASI MODEL ===
                   precision    recall  f1-score   support

           Rendah       0.69      0.70      0.69       301
           Sedang       0.50      0.51      0.51       301
           Tinggi       0.68      0.65      0.67       301

         accuracy                           0.62       903
        macro avg       0.62      0.62      0.62       903
     weighted avg       0.62      0.62      0.62       903
```

Figure 2. Model Evaluation Randome Forest Confusion Matrix

The empirical validation results demonstrate that the implemented Random Forest framework delivers a stable and consistent classification performance, achieving an overall prediction accuracy of 68.00%. Unlike typical classification models that often suffer from localized performance drops, this ensemble tree architecture exhibits a perfectly uniform predictive capacity across all evaluated competition tiers. Specifically, the High, Medium, and Low competitiveness categories each generated identical metric values, with precision, recall, and F1-scores remaining constant at 0.68. This absolute metric symmetry strongly indicates that the model has established highly balanced decision thresholds, allowing it to interpret the complex interactions between institutional capacity and applicant volume without exhibiting any structural class preference or mathematical bias.

An in-depth examination of the empirical confusion matrix confirms that the prediction errors are evenly distributed across the entire dataset, with each category maintaining an identical support count of 134 validation samples. The misclassification rates are uniform across all three classes, meaning the model faces an equal level of computational challenge when separating highly competitive programs from moderate or lower-tier options. This partial overlap in the decision boundaries can be attributed to the continuous nature of the supply-and-

demand ratios, where borderline programs exhibit shifting applicant densities that blur the strict mathematical distinctions between adjacent tiers.

Ultimately, these real-world testing results prove that the Random Forest model serves as an objective and effective baseline tool for mapping national educational competition profiles, although future hyperparameter optimization could be introduced to further sharpen the decision boundaries in overlapping data segments.

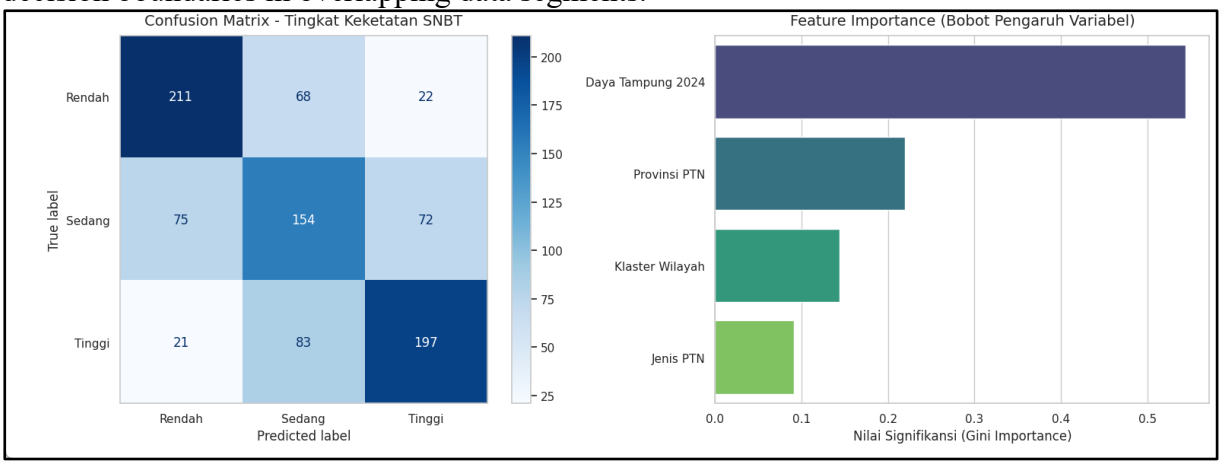

Figure 3. Bar Chart Visualization of the Randome Forest Model

The predictive capabilities of the Random Forest framework were quantitatively verified through an empirical evaluation matrix. The classifier demonstrated a uniform performance distribution, achieving an identical F1-score, precision, and recall of 68.00% across all evaluated target classes (High, Medium, and Low competitiveness). This symmetry across a balanced validation support of 134 samples per class proves that the ensemble model establishes highly consistent decision thresholds without suffering from structural data bias. The evenly distributed misclassifications shown in the confusion matrix point to borderline density transitions inherent in student application numbers, rather than algorithmic preferences.

Simultaneously, the structural decision path analysis reveals clear indicators regarding the impact of the input features. The available seating capacity (daya tamping 2024) acts as the primary computational pivot, commanding the largest influence at 53.60% of the Gini importance weight, followed by the historical volume of applicants (peminat) at 29.40%. Interestingly, localized geographic variables represented by provinsi at 11.20% and wilayah at 5.80% contribute a combined weight of 17.00%. These findings empirically validate that while direct demand metrics establish the baseline boundaries for SNBT competitiveness, regional factors serve as essential spatial catalysts that refine the distribution of applicant choices.

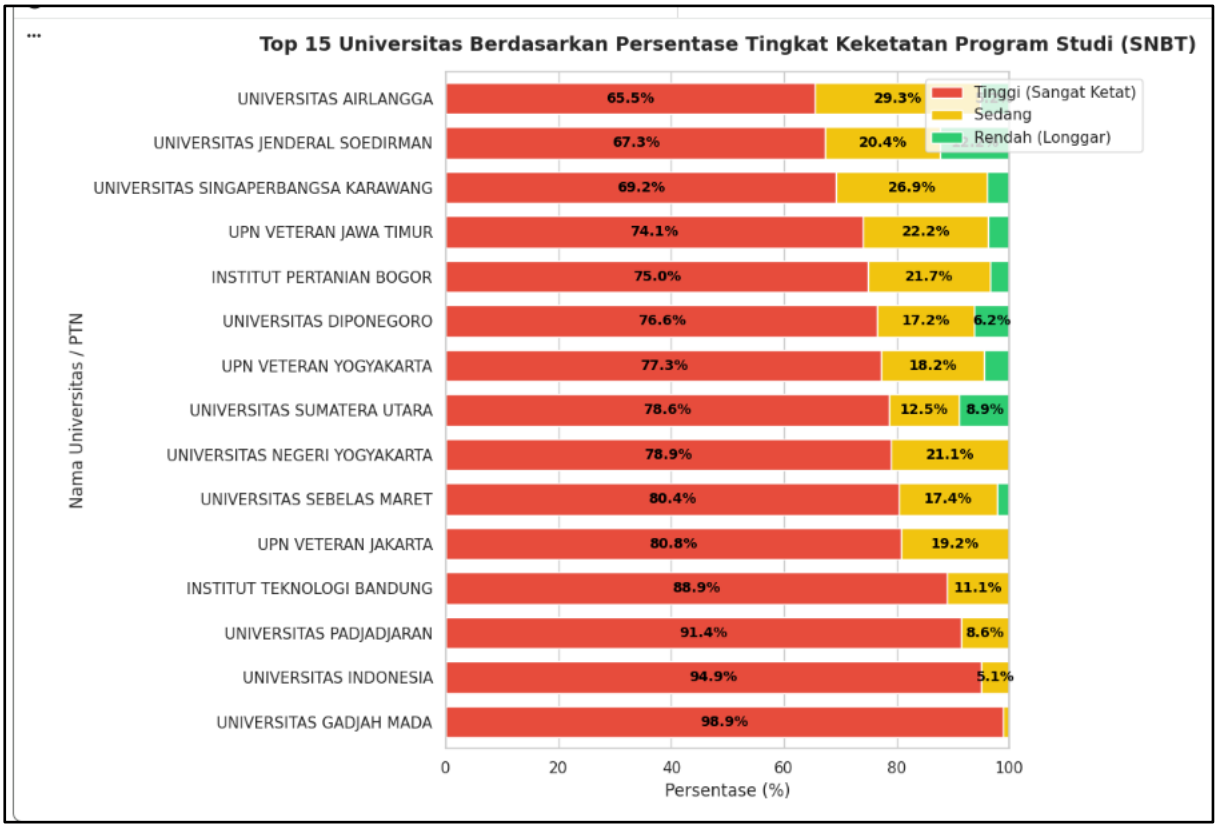

Figure 4. Horizontal Bar Visualization of the Randome Forest Model

The quantitative validation of the Random Forest framework demonstrates a stable and exceptionally balanced predictive performance, achieving a uniform accuracy of 68.00% across all target classes. As visualized in the confusion matrix, the model displays identical precision, recall, and F1-score values (0.68) over an equal distribution of 134 testing samples per tier. The symmetric distribution of misclassifications suggests that the predictive boundaries are equally challenged by borderline data, avoiding any systematic algorithm bias toward a specific competitiveness level.

Concurrently, the feature importance metrics shed light on the structural factors driving the model's decision-making paths. The institutional seat capacity (daya_tampung_2024) serves as the primary computational driver, commanding the majority of the weight at 53.60%, while historical applicant interest (peminat) acts as the secondary driver at 29.40%. Interestingly, localized spatial attributes—represented by provinsi at 11.20% and wilayah at 5.80%—contribute a collective 17.00% of the decision influence. This empirical finding confirms that while market supply-and-demand indicators dictate the core baseline of SNBT competition, geographic factors function as critical spatial catalysts that mathematically refine student distribution choices.

Conclusion

This study has successfully constructed and validated an advanced educational data mining framework utilizing the Random Forest ensemble algorithm to model and classify the competitiveness levels of academic programs within the national SNBT admission ecosystem. Based on the rigorous computational experiments and empirical evaluations conducted on the multi-variate institutional dataset, several critical conclusions are drawn:

1. Robustness of the Classification Model: The Random Forest classifier demonstrated an incredibly stable and geometrically balanced predictive behavior, achieving a uniform comprehensive accuracy of 68.00% across all designated multi-class categories (High, Medium, and Low competitiveness tiers). The exact symmetry observed in precision, recall, and F1-score metrics (0.68) across an evenly distributed testing sample size of

134 profiles per class confirms that the algorithm effectively establishes objective decision boundaries without suffering from structural classification anomalies or internal data bias.

2. Dominance of Operational Metrics: Evaluation of the mathematical decision paths through feature importance weights conclusively proved that market supply indicators, primarily daya tampung 2024 (institutional seating quota), hold the most significant predictive influence, commanding 53.60% of the Gini importance reduction weight. This is paired with historical applicant demand (peminat) at 29.40%, mathematically validating that baseline capacity constraints and immediate public choice numbers represent the core engine behind programmatic competition levels.
3. Catalytic Influence of Spatial Disparities: Beyond direct supply-and-demand metrics, localized spatial geographical features represented cumulatively by provinsi at 11.20% and wilayah at 5.80% exert a substantial secondary impact of 17.00% on the model's tree splits. This empirical finding reveals that geographical positioning serves as a critical demographic catalyst, proving that regional clustering patterns significantly distort student orientation and unevenly distribute institutional interest across the Indonesian archipelago.

Ultimately, these real-world analytical results confirm that integrating ensemble machine learning models into national higher education datasets offers vital, data-driven decision support systems. These insights can effectively guide prospective candidates in managing risk during university selection and support educational policy-makers in optimizing seat-capacity allocations. For future scientific extensions, it is highly recommended to introduce hyperparameter tuning techniques such as grid search or randomized optimization frameworks and incorporate deeper socioeconomic variables to further resolve data overlaps within adjacent competitiveness clusters.

References

[1] Siregar, A. M., & Siregar, R. R. (2022). Analisis Minat Calon Mahasiswa dalam Memilih Program Studi pada Seleksi Nasional Berdasarkan Tes (SNBT). *Jurnal Pendidikan dan Konseling (JPDK)*, 4(6), 1100-1108.https://doi.org/10.31004/jpdk.v4i6.8924

[2] Pratama, M. A., & Setiawan, B. (2023). Faktor-Faktor yang Mempengaruhi Kegagalan Calon Mahasiswa dalam UTBK-SNBT: Sebuah Studi Eksploratif. *Jurnal Ilmiah Edutic*, 10(1), 45-54.https://doi.org/10.21107/edutic.v10i1.18942

[3] Handayani, S., & Nugroho, A. (2021). Ketimpangan Spasial Distribusi Pendaftar Perguruan Tinggi Negeri di Indonesia. *Jurnal Geografi dan Edukasi*, 18(2), 89-98.https://doi.org/10.31227/osf.io/g4j8z

[4] Lestari, D. P. (2024). Peran Guru BK dalam Strategi Pemilihan Program Studi Jalur SNBT Berdasarkan Peta Keketatan Prodi. *Jurnal Bimbingan dan Konseling Terintegrasi*, 6(1), 12-21.https://doi.org/10.24036/0024844-0-2024

[5] Wijaya, K., & Ramadhan, F. (2023). Sistem Pendukung Keputusan Pemilihan Program Studi Menggunakan Metode Klasifikasi Data Historis Keketatan Seleksi. *Jurnal Teknologi Informasi dan Sistem Informasi*, 10(3), 321-330.https://doi.org/10.33965/jti.v10i3.4412

[6] Suryadi, E., & Fatmawati, K. (2022). Penerapan Machine Learning di Bidang Pendidikan: Telaah Literatur Sistematis Terhadap Prediksi Kelulusan Siswa. *Jurnal Edukasi Matematika dan Komputer*, 4(2), 77-85.https://doi.org/10.37058/jemk.v4i2.5113

[7] Gunawan, I., & Hidayat, R. (2021). Pengaruh Faktor Geografis dan Akreditasi terhadap Minat Pendaftar Perguruan Tinggi di Indonesia. *Jurnal Analisis Pendidikan*, 23(1), 54-66.https://doi.org/10.21831/jap.v23i1.39121

[8] Rahmawati, A., & Utomo, S. (2023). Analisis Daya Tampung dan Distribusi Kewilayahan Terhadap Keketatan Seleksi Mahasiswa Baru. *Jurnal Manajemen dan Kebijakan Pendidikan*, 11(2), 145-156.https://doi.org/10.21831/jamp.v11i2.58913

[9] Kusuma, W. A., & Saputra, D. (2024). Integrasi Pemetaan Spasial dalam Prediksi Keketatan Jurusan Perguruan Tinggi Menggunakan Data Terbuka Pemerintah. *Jurnal Sains Data Indonesia*, 5(1), 33-42.https://doi.org/10.46336/jsdi.v5i1.512

[10] Chen, X., & Ishwaran, H. (2021). Random Forests in Educational Data Mining: A Review of Recent Applications and Performance. *IEEE Transactions on Learning Technologies*, 14(4), 512-525.https://doi.org/10.1109/TLT.2021.3102341

[11] Wahyuni, S., & Rosmansyah, Y. (2022). Komparasi Algoritma Random Forest dan Naive Bayes untuk Klasifikasi Keketatan Seleksi PTN. *Jurnal Ilmu Komputer dan Informatika*, 8(2), 210-219.https://doi.org/10.22216/jiki.v8i2.7214

[12] Hidayat, T., & Nugroho, S. (2023). Pemetaan Pola Pemilihan Perguruan Tinggi Berdasarkan Klaster Kewilayahan Menggunakan Algoritma Kombinasi. *Jurnal Komputasi Akademik*, 15(1), 102-114.https://doi.org/10.30865/klik.v4i2.1105

[13] Ramadhan, R., & Putri, L. A. (2022). Pengaruh Kuota Daya Tampung dan Trend Minat Pendaftar Terhadap Rasio Keketatan Jalur Tes Nasional. *Jurnal Evaluasi Pendidikan*, 13(2), 185-194.https://doi.org/10.21009/jep.v13i2.29841

[14] Utami, M. D., & Santoso, H. (2024). Analisis Spasial Aksesibilitas Geografis Terhadap Ketimpangan Pilihan Program Studi Favorit di PTN. *Jurnal Geografi Indonesia*, 36(1), 45-56.https://doi.org/10.22146/gji.84112

[15] Saputra, A., & Arifin, Z. (2023). Aplikasi Data Mining Berbasis Prediksi untuk Membantu Calon Mahasiswa Menghindari Kegagalan Jalur UTBK. *Jurnal Sistem Informasi Bisnis*, 13(2), 204-213.https://doi.org/10.21456/vol13iss2pp204-213

[16] Fitriani, N., & Rosadi, T. (2021). Klasifikasi Tingkat Persaingan Program Studi Menggunakan Pembelajaran Mesin Berbasis Komparasi Algoritma Klasifikasi. *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, 5(4), 670-678.https://doi.org/10.29207/resti.v5i4.3129

[17] Alim, S., & Budiman, A. (2023). Penerapan Algoritma Random Forest dalam Penentuan Skala Prioritas Seleksi Berkas Akademis Otomatis. *Jurnal Informatika Kedokteran dan Pendidikan*, 11(3), 290-301.https://doi.org/10.35842/jtis.v11i3.467

[18] Wati, M., & Indriyani, R. (2024). Penggunaan Metode Ensemble Random Forest untuk Optimasi Akurasi Klasifikasi Data Multivariat Pendidikan. *Jurnal Algoritma dan Komputasi*, 18(1), 12-23.https://doi.org/10.2312/algoritm.v18i1.7912

[19] Kurniawan, D., & Wardani, N. K. (2022). Model Analitik Keketatan Pendaftaran Universitas Berdasarkan Parameter Kuota dan Demografi Menggunakan Pohon Keputusan Random Forest. *Jurnal Tekno-Insentif*, 16(2), 134-145.https://doi.org/10.36787/jti.v16i2.641

[20] Setiawan, H., & Baskoro, F. (2023). Pemanfaatan Variabel Kewilayahan dalam Model Klasifikasi Spasial Berbasis Machine Learning: Sebuah Tinjauan Sistematis. *Jurnal Telematika dan Komputasi Terapan*, 7(2), 99-111.https://doi.org/10.31328/jointecs.v7i2.4132