KNN Algorithm for Predicting MTsN Padang Panjang Students' to High School Placement

1st Risna Zahira
School of Computing
Telkom University
Bandung, Indonesia
risnazahir@student.telkomuniversity.ac.id

2nd Putu Harry Gunawan
School of Computing
Telkom University
Bandung, Indonesia
phgunawan@telkomuniversity.ac.id

3nd Indwiarti Indwiarti
School of Computing
Telkom University
Bandung, Indonesia
indwiarti@telkomuniversity.ac.id

***Abstract*—Education is a top priority in Indonesia's development, with every region committed to enhancing its quality. To support individual skill development, high-quality advanced education is essential. Therefore, developing a predictive model can help students and schools prepare for the admission process to top-tier institutions. This study predicts the admission outcomes of MTsN Padang Panjang students to prestigious high schools, including MAN Insan Cendekia, SMAN 1 Sumatera Barat, SMAN 2 Padang Panjang, and SMAN 1 Padang Panjang. The model employs the K-Nearest Neighbors (KNN) algorithm utilizing historical data such as academic scores, achievement records, preferred school, and alumni distribution. The KNN algorithm was selected for its proven effectiveness in various prediction and classification tasks. Pre-processing involves managing data imbalance through SMOTE oversampling, handling outliers using the IQR method, standardizing data, and selecting relevant features. The model is evaluated using Precision, Recall, and F1-Score metrics, achieving an overall accuracy of 92%, demonstrating its effectiveness in classifying data. Unlike previous studies that focused on predicting outcomes for a single school, this study introduces a novel approach by applying a multiclass prediction model across several prestigious schools, incorporating feature selection and parameter tuning. These findings highlight the model's potential to improve student placement predictions and aid in postsecondary school selection.**

***Index Terms*—K-Nearest Neighbors (KNN), prediction model, student graduation.**

I. INTRODUCTION

Education is at the core of Indonesia's development agenda, with the third-largest education system in Asia and the fourth-largest in the world, encompassing over 50 million students and 2.6 million teachers across more than 250,000 schools. Two ministries oversee the management of the education system: the Ministry of National Education (*Kementerian Pendidikan Nasional, Kemendiknas*), which manages 84% of schools, and the Ministry of Religious Affairs (*Kementerian Agama, Kemenag*), responsible for 16%. Private schools also play a significant role [1].

Every region in Indonesia strives to improve the quality of education, including West Sumatra. According to Governor Regulation (*Peraturan Gubernur, Pergub*) Number 12 of 2021 on New Student Admissions for Public Senior High Schools, West Sumatra provides four admission pathways for high schools: zoning, affirmation, parental/guardian assignment, and achievement [2].

Several prestigious schools, such as MAN Insan Cendekia, SMAN 1 Sumatera Barat, SMAN 2 Padang Panjang, and SMAN 1 Padang Panjang, are the primary targets for students aspiring to continue their education at the high school level. The competition to gain admission to these schools is highly competitive due to limited enrollment quotas. To minimize failure in the admission selection process, machine learning can be utilized to develop predictive models that provide valuable insights for students and schools in preparing for the selection process to these top high schools [3] [4].

Previous studies have indicated that various machine learning algorithms are effective in predicting student performance. Research conducted by Supriadi et al. [5] examined the effectiveness and accuracy of the K-Nearest Neighbor (KNN) method in classifying student graduation levels using historical student data, such as exam scores and report card grades, as predictive features, achieving high accuracy and stable performance.

Research by Wiyono et al. [6] compared three machine learning algorithms Support Vector Machine (SVM), KNN, and Decision Tree, to identify the most effective model for predicting student performance, particularly in distinguishing between active and inactive students. The overall accuracy of each algorithm was SVM at 95%, KNN at 92%, and Decision Tree at 93%. Although SVM achieved the highest accuracy, the KNN algorithm remains advantageous due to its ease of implementation, flexibility for multi-class data, and suitability for small datasets with good performance and minimal risk of overfitting.

Previous studies using the KNN algorithm focused on predicting student graduation at one school or college and were less than optimal in selecting features or algorithm parameters, so that prediction accuracy was limited. This study introduces a novel approach by developing a multi-class prediction model for four prestigious high schools: MAN Insan Cendekia, SMAN 1 Sumatera Barat, SMAN 1 Padang Panjang, and SMAN 2 Padang Panjang. The model utilizes historical data from MTsN Padang Panjang, the State Islamic Junior High School *(Madrasah Tsanawiyah Negeri, MTsN)* of Padang Panjang, including academic grades and achievement records as input variables. MTsN Padang Panjang data was chosen because it is one of the best and most favorite junior high schools in West Sumatra, with a track record of superior academic achievement [7]. The selection of this school is expected to represent the pattern of student admissions in high schools, so that the prediction results obtained can be applied more widely.

II. METHODS

A. Research Design

The research began with a literature review of previous studies to gain background knowledge and ideas for solutions

to the identified problem. The study continued with the collection of academic grades and achievement data from MTsN Padang Panjang, which was then analyzed to gain an initial understanding of its basic structure. The data was subsequently cleaned and checked for balance. If the dataset was found to be imbalanced, data balancing techniques were applied. Next, data visualization was performed to examine the relationships between features and the target variable. The data was then transformed into a format compatible with processing. Feature selection was conducted by analyzing the correlation between features and the target variable. Following these steps, the KNN model was applied to the data to evaluate its performance, and the research concluded with an analysis of the results. Fig.1 presents the flowchart used in this research to facilitate understanding of the research process.

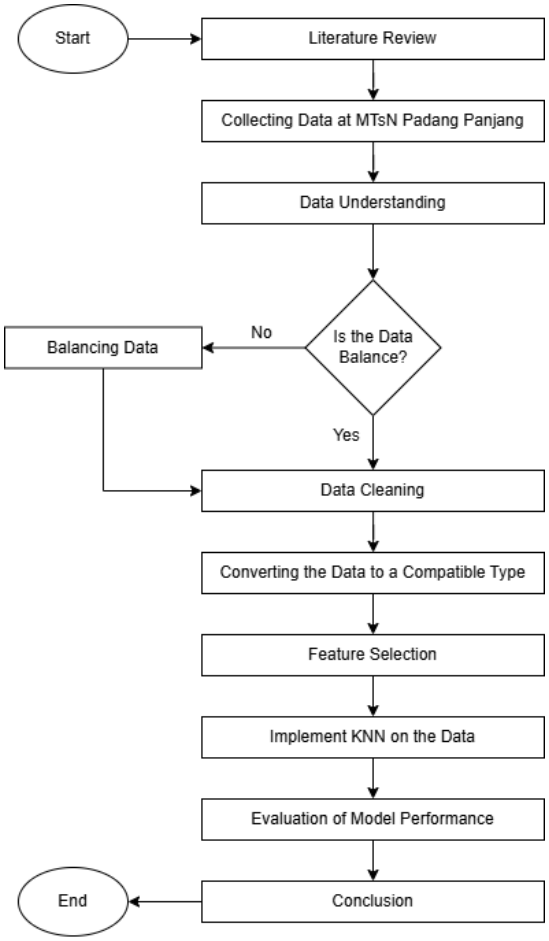

Fig. 1. Flowchart of the Research Process

B. K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) is a classification algorithm commonly utilized for predicting performance across diverse domains. The performance of KNN depends heavily on high quality and accurate training data [8]. The algorithm determines the class of test data by evaluating the distance between test data points and training data, assigning the class based on the majority category among its closest neighbors. Although straightforward in concept, KNN faces challenges in selecting the optimal value for K and choosing an appropriate distance metric [9]. One commonly used metric is Euclidean Distance, which measures how close or far the test data point is from training data points by considering the absolute differences in attributes. The formula is:

$$euc(X,Y) = \sqrt{\sum_{i=1}^{n}(X_i - Y_i)^2} \tag{1}$$

Where X and Y are two data vectors where n represents the number of features in the dataset, and X_i and Y_i are the values of the i-th feature in each vector. By calculating this distance, KNN determines the relationship between test and training data, enabling the prediction of the most likely patterns. The smaller the distance, the greater the similarity between the data points [10].

The parameter K defines the number of nearest neighbors used to decide the class of test data. Choosing the right K is critical to avoid overfitting or underfitting. Cross-validation, a technique for validating model performance by splitting the dataset into subsets, is used to optimize K and reduce the risk of overfitting. Grid Search is also employed to find the best K value by testing various values within a specified range. These processes enhance the overall performance of the KNN algorithm. KNN model performance can be evaluated using a Confusion Matrix, which provides a clear breakdown of correct or incorrect predictions, thus enabling a deeper understanding of the model's behavior [11]. In KNN, the confusion matrix helps identify classification errors and evaluate metrics such as accuracy, precision, recall, and F1-Score [12]. It generates four key values: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). These values are then used to compute essential evaluation metrics as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{4}$$

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5}$$

True Positive (TP) indicates the count of positive samples accurately classified as positive, whereas True Negative (TN) represents the count of negative samples correctly identified as negative. Conversely, False Positive (FP) refers to the number of negative samples incorrectly classified as positive, while False Negative (FN) represents the number of positive samples mistakenly classified as negative.

C. Exploratory Data Analysis

The data collection method was conducted quantitatively through direct communication with the school to obtain relevant data. The dataset includes the recapitulation of report card grades, student achievement records, data on preferred schools, and alumni distribution from six academic years, from 2019 to 2024. The dataset consists of twenty-three columns and 1.496 rows. The information includes the Student Identification Number (*Nomor Induk Siswa, NIS*), National Student Identification Number (*Nomor Induk Siswa Nasional, NISN*), student name, gender, grades for 15 subjects

(as detailed in Table I), the final average score, and certificates reflecting student achievements.

TABLE I
FIFTEEN SUBJECTS AND THEIR DESCRIPTIONS

Subject	Description
QH (Quran Hadist)	Quranic Studies and Hadith
AA (Akidah Akhlaq)	Islamic Morality
FIK (Fikih)	Islamic Jurisprudence
SKI (Sejarah Islam)	Islamic History
PPKn (Kewarganegaraan)	Citizenship
BIND (Bahasa Indonesia)	Indonesian
BAR (Bahasa Arab)	Arabic
MTK (Matematika)	Mathematics
IPA (Science)	Science
IPS (Sosial)	Social
BING (Bahasa Inggris)	English
SB (Seni Budaya)	Arts and Culture
PJOK (Olahraga)	Physical Education
PRKTI (Praktik)	Practical Skills
MULOK (Tahfiz)	Quran Memorization

Additionally, the dataset records the Preferred School, indicating the schools targeted by students for further education, and the Advanced School, which represents the alumni distribution and the schools where students were ultimately accepted and enrolled.

The input variables in this prediction model are the grades of 15 subjects, the final average score, and the Preferred School. The Advanced School column in the dataset serves as the target variable and contains four primary categories: MAN Insan Cendekia, SMAN 1 Sumatera Barat (SMAN 1 SUMBAR), SMAN 2 Padang Panjang (SMAN 2 PP), and SMAN 1 Padang Panjang (SMANSA).

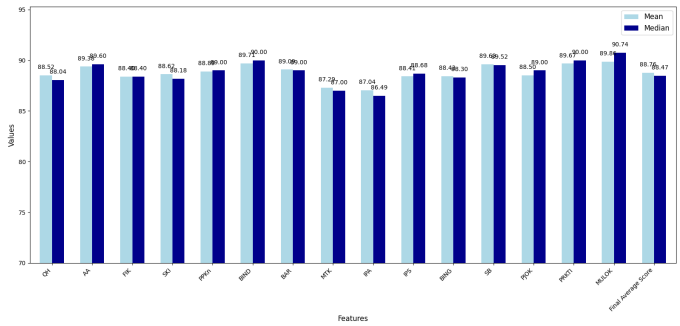

Fig. 2. Mean and Median of Numeric Features

Fig.2 shows the mean and median values of various numeric features in the dataset. The chart shows that most features, such as QH, AA, FIK, SKI, and Final Average Score, have mean and median values that are almost aligned, indicating symmetrical distributions with minimal skewness. Features like BAR and MULOK show mean values slightly higher than the median, suggesting the presence of some higher values affecting the distribution. MTK exhibits a broader range with a slight difference between mean and median, reflecting the presence of lower values. Overall, academic performance is consistent across features, with mean and median values generally falling within the range of 85 to 90.

Fig.3 shows the distribution of certificates and students' school preferences, with 41.4% of students holding a certificate (Y) and 58.6% not holding a certificate (N). SMANSA is the most preferred school, chosen by 34.5% of students,

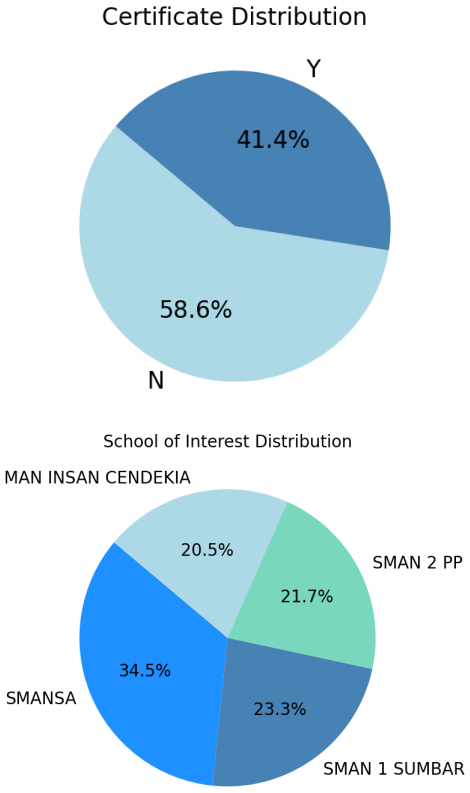

Fig. 3. Certificate and Preferred School Distribution

followed by SMAN 1 SUMBAR at 23.3%, SMAN 2 PP at 21.7%, and MAN Insan Cendekia at 20.5%.

D. Preprocessing

After the data collection process and initial understanding of the data in the Exploratory Data Analysis stage. The initial target distribution revealed an imbalance in the data, as shown in Table II. This imbalance can adversely affect the model's performance, especially in predicting minority classes. To address this issue, additional data was collected to narrow the value range between classes, resulting in the final value distribution shown in the "Final Value" column. Furthermore, oversampling was applied to the training data using the Synthetic Minority Over-sampling Technique (SMOTE) to mitigate data imbalance [13].

TABLE II
TARGET DATA DISTRIBUTION

Advanced School	Initial Value	Final Value
MAN Insan Cendekia	132	277
SMAN 1 SUMBAR	200	344
SMAN 2 PP	383	387
SMANSA	432	488

Despite using SMOTE, the target distribution remained slightly imbalanced due to the need for a higher volume of synthetic data for minority classes [14]. Excessive oversampling can compromise data quality. Additionally, oversampling was applied only to the training data, leaving the test data with its original imbalanced distribution, which influenced model evaluation. However, this slight imbalance had some benefits, such as maintaining dataset realism, enabling the model to capture relevant patterns, improving generalization to real-world data, and preventing overfitting to minority classes. This approach preserved data variation

and ensured the model focused on dominant patterns from the majority class, resulting in more stable predictions.

Next, the data cleaning process involved removing irrelevant and unnecessary columns, such as NIS, NISN, Student Name, and Gender. Any null values and duplicates present in the data were eliminated to ensure the dataset was of high quality and accuracy. For categorical columns, such as Certificate, Preferred School, and Advanced School, encoding was performed to convert them into numerical formats [15]. The Certificate column was encoded as ”Y” = 1 and ”N” = 0, while the Preferred School and Advanced School columns were encoded as follows: MAN Insan Cendekia = 0, SMAN 1 SUMBAR = 1, SMAN 2 PP = 2, and SMANSA = 3.

The Interquartile Range (IQR) method was employed to address outliers. This approach involves calculating the range between the first quartile (Q1) and the third quartile (Q3). Data points falling outside the normal range were identified as outliers and removed. This step ensured that the data was more representative and less influenced by extreme values that could disrupt the analysis or modeling process [16].

After cleaning and encoding, numerical data was normalized using the StandardScaler from the sklearn.preprocessing library. Standardization was performed to scale numerical values to a uniform range, a critical step for ensuring the KNN algorithm operates optimally on datasets with consistent feature scales [17].

Two types of features were analyzed: numerical features and categorical features. Fig.4 presents the distribution of grades for several subject features and the final average score concerning the target Advanced School. The distribution of the final average score is relatively stable with a narrow range across all target categories, with only a few outliers, indicating data consistency. The distributions of the subject QH, AA, and FIK features are similar to the final average score, showing small variations and minimal outliers. In contrast, MTK and PJOK features exhibit numerous outliers with extreme values.

Fig.5 illustrates the distribution of categorical features, namely Certificate and Preferred School, concerning the target Advanced School. For the Certificate feature, Y indicates having a certificate of achievement in competitions, while N indicates otherwise. Each target has a distribution for Y and N. A majority of students accepted into MAN Insan Cendekia hold achievement certificates (Y), highlighting this feature's importance in increasing the likelihood of being admitted to schools with stringent selection processes. Additionally, most students succeeded in enrolling in advanced schools that matched their Preferred School. The high correlation between Preferred School and Advanced School suggests that Preferred School is a significant predictor for Advanced School.

E. Feature Selection and Data Splitting.

The feature selection process is performed using boxplot visualization analysis and involved statistical tests such as Analysis of Variance (ANOVA), which was used to analyze the differences in mean values across features within target categories, and Mutual Information, which measured the relevance of features to the target variable. Feature selection is essential for enhancing the accuracy of machine learning models across various applications. [18] [19] [20].

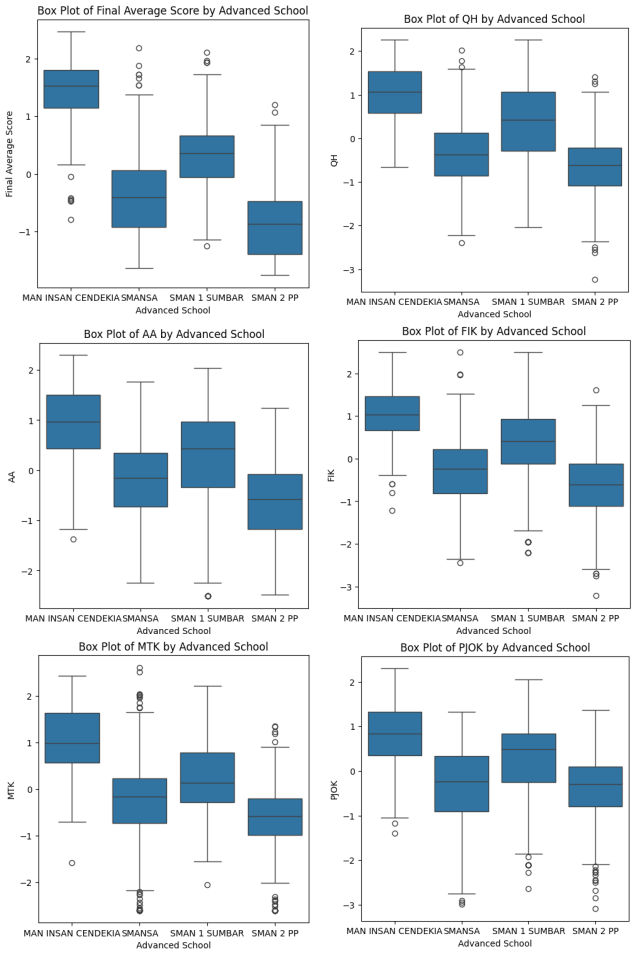

Fig. 4. Numerical Feature Distribution by Target Category

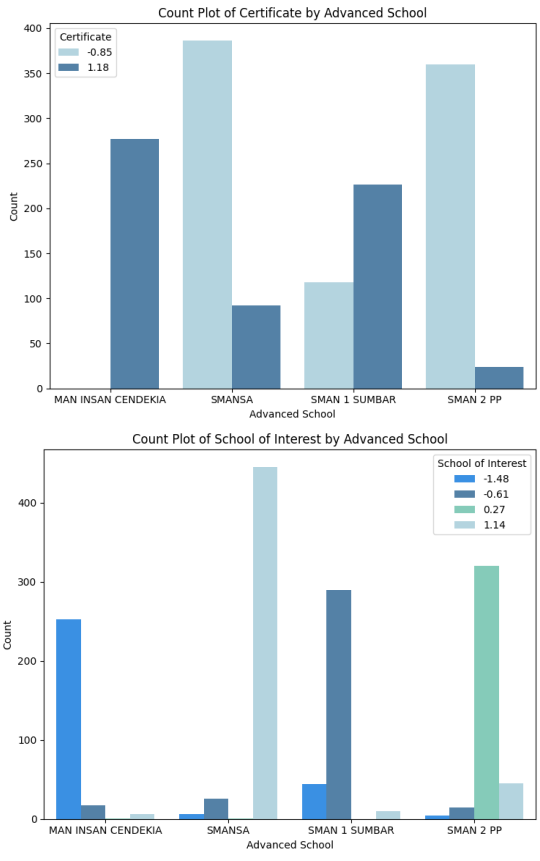

Fig. 5. Categorical Feature Distribution by Target Category

In this study, the relationship between features and the target variable was assessed using mutual_info_classif() from the sklearn.feature_selection library, which calculates the degree of relevance a feature has to the target. A high mutual information score indicates that the feature significantly influences the target variable. Additionally, the ANOVA test was used to evaluate the correlation between features and the target by measuring the F-Statistic, which quantifies the variability between groups within target categories, and calculating the P-Value to determine the significance of the correlation between features and the target. Features with low P-Values were considered significant and relevant to the target.

The best features selected based on the ANOVA test results were the Final Average Score, Certificate, Preferred School, and the subjects Natural Sciences (IPA) and Islamic Cultural History (SKI).

Meanwhile, based on the analysis of numerical and categorical features, the best features for the prediction model are Final Average Score, Certificate, Preferred School, and subjects QH, AA, and FIK. These features demonstrate stable distributions, strong relevance to the target variable, and potential to improve the model's predictive performance.

Next, the processed dataset is divided into two parts: 80% for training data to train the model and 20% for testing data to evaluate model performance.

F. K-Nearest Neighbor Implementation

This study employed the K-Nearest Neighbors (KNN) algorithm for classification, implemented using the Scikit-learn library in Python. Initially, the parameter n_neighbors = 23 was used to build the first model, indicating that the model considered 23 nearest neighbors to determine the class of a data point. To improve accuracy, hyperparameter tuning was conducted to identify the optimal value for n_neighbors. Using the Grid Search technique with 10-fold cross-validation, various values of n_neighbors ranging from 1 to 30 were tested. The results indicated that the best-performing model was achieved with n_neighbors = 3.

III. RESULTS AND DISCUSSIONS

A. Model with Raw Data

TABLE III
COMPARISON OF PRECISION, RECALL, AND F1-SCORE USING RAW DATA

Target	Precision	Recall	F1-Score
MAN Insan Cendekia	60%	52%	56%
SMAN 1 SUMBAR	44%	43%	44%
SMAN 2 PP	66%	81%	73%
SMANSA	57%	48%	52%
Overall Accuracy	**61%**		

TABLE IV
CONFUSION MATRIX USING RAW DATA

	Class 1	Class 2	Class 3	class 4
Class 1	15	9	2	3
Class 2	6	16	1	14
Class 3	0	1	65	14
Class 4	4	10	30	41

Table III presents the model evaluation results using raw data. The best model accuracy achieved was 61%. The highest precision was recorded for the SMAN 2 PP class 66% while the lowest was for the SMAN 1 SUMBAR class 44%, indicating a high number of incorrect predictions for certain classes. Recall was also better for majority classes, such as SMAN 2 PP 81%, compared to minority classes, such as MAN Insan Cendekia 56%.

The confusion matrix in Table IV shows imbalanced predictions across classes. Majority classes like SMAN 2 PP had 65 correct predictions, while MAN Insan Cendekia only had 15, with many misclassifications into other classes. The model tends to prioritize dominant classes, resulting in poor predictions for minority classes.

Without applying SMOTE to address data imbalance and selecting relevant features, the model's performance becomes inconsistent, with only a few classes dominating the predictions.

B. Model with Processed Data

This study evaluated the performance of the K-Nearest Neighbors (KNN) algorithm using two variations of feature selection.

- Features based on boxplot visualization: Final Average Score, Certificate, Preferred School, and the subjects QH (Qur'an and Hadith), AA (Islamic Morals), and FIK (Islamic Jurisprudence).
- Features based on ANOVA test: Final Average Score, Certificate, Preferred School, and the subjects IPA (Natural Sciences) and SKI (Islamic Cultural History).

These two feature variations were compared to determine which set yielded the best accuracy and performance for the model.

The evaluation results showed that the feature variation based on boxplot visualization performed best, achieving an overall accuracy of 92%, while the feature variation based on the ANOVA test achieved an overall accuracy of 89%. In addition to accuracy, other evaluation metrics such as precision, recall, and F1-Score were also analyzed for each class.

TABLE V
COMPARISON OF PRECISION, RECALL, AND F1-SCORE USING BOXPLOT

Target	Precision	Recall	F1-Score
MAN Insan Cendekia	95%	91%	93%
SMAN 1 SUMBAR	84%	92%	88%
SMAN 2 PP	95%	92%	94%
SMANSA	94%	93%	94%
Overall Accuracy	**92%**		

TABLE VI
CONFUSION MATRIX USING BOXPLOT

	Class 1	Class 2	Class 3	class 4
Class 1	52	4	0	1
Class 2	2	56	2	1
Class 3	1	1	68	3
Class 4	0	6	2	98

Based on Table V, the boxplot visualization-based feature achieves the best balance between precision, recall, and F1-Score. This model excels in predicting the MAN Insan Scholar class with 95% precision, although it is a minority class, while SMAN 1 SUMBAR shows lower precision, indicating a tendency for misclassification. The confusion matrix values shown in Table VI show that the model has

successfully classified class 1 (MAN Insan Scholar) correctly with 52 values, 56 values for class 2 (SMAN 1 SUMBAR), 68 values for class 3 (SMAN 2 PP). And class 4 (SMANSA) with the best performance with 98 values successfully classified correctly. Although there are some misclassifications, the model still shows the best performance.

TABLE VII
COMPARISON OF PRECISION, RECALL, AND F1-SCORE USING ANOVA FEATURES

Target	Precision	Recall	F1-Score
MAN Insan Cendekia	92%	95%	93%
SMAN 1 SUMBAR	81%	92%	86%
SMAN 2 PP	90%	86%	88%
SMANSA	93%	87%	90%
Overall Accuracy	**89%**		

TABLE VIII
CONFUSION MATRIX USING ANOVA FEATURES

	Class 1	Class 2	Class 3	class 4
Class 1	54	2	0	1
Class 2	3	56	1	1
Class 3	2	3	63	5
Class 4	0	8	6	92

In Table VII, the features based on the ANOVA test showed better recall for the MAN Insan Cendekia class, achieving 95%, indicating the model's strong ability to recognize data from this class almost entirely. However, the precision for the SMAN 1 SUMBAR class remained the lowest at 81%, which affected the F1-Score for this class. Based on the confusion matrix in Table VIII, class 4 (SMANSA) has the best performance with the number of correctly classified values of 92 values and only a few misclassifications. Class 1 (MAN Insan Cendekia) and class 2 (SMAN 1 SUMBAR) show a relatively large number of errors. Class 3 (SMAN 2 PP) experiences more misclassifications than other classes.

ANOVA-based features are less optimal as they only consider the individual relationship between features and the target, without accounting for interactions between features, sensitivity to data imbalance, and distribution information such as outliers. In contrast, features identified through boxplot provide a more comprehensive understanding of data patterns, ultimately leading to better model performance.

TABLE IX
MODEL PREDICTION RESULT

Advanced School	Predicted_Class
MAN INSAN CENDEKIA	MAN INSAN CENDEKIA
MAN INSAN CENDEKIA	MAN INSAN CENDEKIA
SMANSA	MAN INSAN CENDEKIA
MAN INSAN CENDEKIA	SMAN 1 SUMBAR
MAN INSAN CENDEKIA	SMAN 1 SUMBAR
SMAN 1 SUMBAR	MAN INSAN CENDEKIA
SMAN 1 SUMBAR	MAN INSAN CENDEKIA
SMANSA	SMANSA
SMAN 2 PP	SMAN 2 PP
SMAN 1 SUMBAR	SMAN 1 SUMBAR
SMANSA	SMAN 1 SUMBAR
SMANSA	SMANSA
SMAN 1 SUMBAR	SMAN 1 SUMBAR
SMAN 2 PP	SMAN 2 PP
SMANSA	SMANSA
SMANSA	SMAN 1 SUMBAR
SMANSA	SMANSA

Table. IX shows some of the model's prediction results, where the "Advanced School" column represents the actual data, and the "Predicted Class" column represents the model's predictions, with the majority of predictions being accurate. Misclassifications in these classes are caused by overlapping input data characteristics or high similarities between classes, such as between MAN Insan Cendekia and SMAN 1 SUMBAR or SMAN 1 SUMBAR and SMANSA, making it challenging for the model to distinguish classes accurately. However, the number of correctly classified instances and the percentage of evaluation metrics indicate that the model has excellent overall performance.

The slightly imbalanced class distribution did not negatively affect the model's performance for minority classes like MAN Insan Cendekia. This was due to the combination of oversampling using SMOTE and relevant feature selection, which allowed the model to learn important patterns from minority classes without overfitting. In other words, despite the smaller amount of data for the MAN Insan Cendekia class compared to majority classes, the model was still able to deliver the best precision, recall, and F1-Score for this class.

Unbalanced test data enhances model generalization by reflecting real-world conditions where class distributions are often unequal. This ensures the model is evaluated in realistic scenarios, testing its ability to recognize both majority and minority classes effectively. By exposing the model to unbalanced data, it avoids overfitting to artificially balanced distributions and better handles class disparities. Additionally, this approach provides a comprehensive evaluation of metrics like Precision, Recall, and F1-Score across all classes, ensuring the model's robustness and reliability in diverse, real-world applications.

Overall, the features based on boxplot visualization proved superior in providing balanced precision and recall, with an overall accuracy of 92%. These results indicate that despite the class imbalance, the approach used in this study successfully addressed the challenge and delivered optimal performance across all classes, including minority classes.

IV. CONCLUSION

This study successfully developed a predictive model for student admission to prestigious high schools based on junior high school historical data using the K-Nearest Neighbors (KNN) algorithm. The model demonstrated excellent performance with an overall accuracy of 92%. The study highlights that the use of SMOTE and optimal feature selection significantly impacts model performance. Based on the evaluation metrics analysis, the Precision, Recall, and F1-Score values showed highly satisfactory results for most classes. The MAN Insan Cendekia, SMANSA, and SMAN 2 PP classes achieved an F1-Score of 93%, while the SMAN 1 SUMBAR class scored slightly lower at 88% due to its lower precision. The best features identified through box plot visualization analysis include the Final Average Score, Achievement Certificate, Preferred School, and key subjects such as QH (Qur'an and Hadith), AA (Islamic Morals), and FIK (Islamic Jurisprudence). Despite the slightly imbalanced class distribution, the combination of SMOTE and effective feature selection allowed the model to maintain good predictive performance across all classes, including minority categories. These findings indicate that using SMOTE and

optimal feature selection significantly enhances the quality of the dataset and the model's performance.

The data from MTsN Padang Panjang was chosen due to its high quality and excellent track record of academic achievements. This school is expected to represent the student admission patterns of high schools. However, the scalability of the model to other contexts or datasets has some limitations, as student admission patterns in other schools may vary due to differences in curriculum, selection standards, and data distribution. Future research can be expanded by involving data from various other schools to improve the generalizability of the predictive model. Additionally, different algorithms could be explored to handle larger or more complex datasets.

REFERENCES

[1] World Bank, "World Bank and education in Indonesia," World Bank, [Online]. Available: https://www.worldbank.org/en/country/indonesia/brief/world-bank-and-education-in-indonesia.

[2] Government of West Sumatra Province, Regulation of the Governor of West Sumatera Number 12 of 2021 concerning Admission of New Learners at Senior High Schools, Vocational High Schools, and Boarding Schools, Padang: West Sumatera Provincial Government, 2021.

[3] A. Maulana, T. R. Noviandy, N. R. Sasmita, M. Paristiowati, R. Suhendra, E. Yandri, J. Satrio, and R. Idroes, "Optimizing university admissions: A machine learning perspective," Journal of Educational Management and Learning, vol. 1, no. 1, pp. 1, 2023, doi: 10.60084/jeml.v1i1.46. Available online: www.heca-analitika.com/jeml.

[4] B. Assiri, M. Bashraheel, and A. Alsuri, "Improve the accuracy of students admission at universities using machine learning techniques," Proc. 7th Int. Conf. Data Science and Machine Learning Applications (CDMA), 2022, pp. 1–6, doi: 10.1109/CDMA54072.2022.00026.

[5] D. Supriadi, Y. Prana, and E. Suryana, "Application of K-Nearest Neighbor for classification of graduation rates in students of SMAN 11 Bengkulu City," Media Infota, vol. 19, no. 2, pp. 1-6, 2023, doi: https://doi.org/10.47738/ijiis.v1i1.17

[6] S. Wiyono and T. Abidin, "Comparative study of machine learning KNN, SVM, and Decision Tree algorithm to predict student's performance," Int. J. Res. - Granthaalayah, vol. 7, no. 12, pp. 129-135, 2019, doi: https://doi.org/10.29121/granthaalayah.v7.i1.2019.1048.

[7] MTsN Padang Panjang. Retrieved December 24, 2024, from: https://mtsnpadangpanjang.sch.id/

[8] A. M. Habibi and R. R. Santika, "Implementation of the K-Nearest Neighbor Algorithm in Determining Majors Using the Web-Based Euclidean Distance Method at SMP Setia Gama," vol. 3, pp. 7-14.

[9] E. Purwaningsih and E. Nurelasari, "Application of K-Nearest Neighbor for Graduation Level Classification on Students," Jurnal Informatika, vol. 10, no. 1, pp. 46–56, 2021.

[10] W. Wahyono, I. N. P. Trisna, S. L. Sariwening, M. Fajar, and D. Wijayanto, "Comparison of distance measurement on k-nearest neighbour in textual data classification," Jurnal Teknologi dan Sistem Komputer, vol. 8, no. 1, pp. 54–58, Jan. 2020, doi: 10.14710/jtsiskom.8.1.2020.54-58.

[11] M. Fahmy, "Confusion Matrix in Binary Classification Problems: A Step-by-Step Tutorial," ERJ, 2020.

[12] N. Azizah and Y. Santoso, "Application of K-Nearest Neighbor Algorithm To Predict Potential Dropout Students"

[13] G. A. Pradipta, R. Wardoyo, A. Musdholifah, I. N. H. Sanjaya, and M. Ismail, "SMOTE for handling imbalanced data problem: A review," in Proc. 2021 Sixth Int. Conf. Informatics and Computing (ICIC), Jakarta, Indonesia, 2021, pp. 1–6, doi: 10.1109/ICIC54025.2021.9632912.

[14] Nasim Matar, Bilal Sowan, and Amneh Al-Jaber, "Evaluating Models Performance for Credit Risk Detection for Imbalanced Data," 2024 2nd International Conference on Cyber Resilience (ICCR), 2024.

[15] N. Sameera and M. Shashi, "Encoding approach for intrusion detection using PCA and KNN classifier," in Proc. Third Int. Conf. Computational Intelligence and Informatics, K. Raju, A. Govardhan, B. Rani, R. Sridevi, and M. Murty, Eds., Advances in Intelligent Systems and Computing, vol. 1090, Springer, Singapore, 2020, pp. 187–199, doi: 10.1007/978-981-15-1480-7_15.

[16] A. A. Alabrah, "An Improved CCF Detector to Handle the Problem of Class Imbalance with Outlier Normalization Using IQR Method," Italian National Conference on Sensors, 2023. doi: 10.3390/s23094406.

[17] A. Berdaly and Z. Abdiakhmetova, "Predicting heart disease using machine learning algorithms," J. Math. Mech. Comput. Sci., vol. 11, no. 3, 2022. doi: 10.26577/jmmcs.2022.v115.i3.10.

[18] T.-N. Tai, H. Tseng, and Y.-T. Sung, "Impact of Feature Selection Algorithms on Readability Model," in Proc. Taiwan Conf. Comput. Linguistics and Speech Processing, 2023.

[19] M. H. M. Zaki, M. A. A. Aziz, S. Sulaiman, and N. Hambali, "Feature Selection Methods Application Towards a New Dataset based on Online Student Activities," J. Electr. Electron. Syst. Res., vol. 23, no. 1, 2023. doi: 10.24191/jeesr.v23i1.004.

[20] N. Nurrahma and R. Yusuf, "Comparing Different Supervised Machine Learning Accuracy on Analyzing COVID-19 Data using ANOVA Test," in Proc. 6th Int. Conf. Interactive Digital Media (ICIDM), 2020, doi: 10.1109/ICIDM51048.2020.9339676.