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

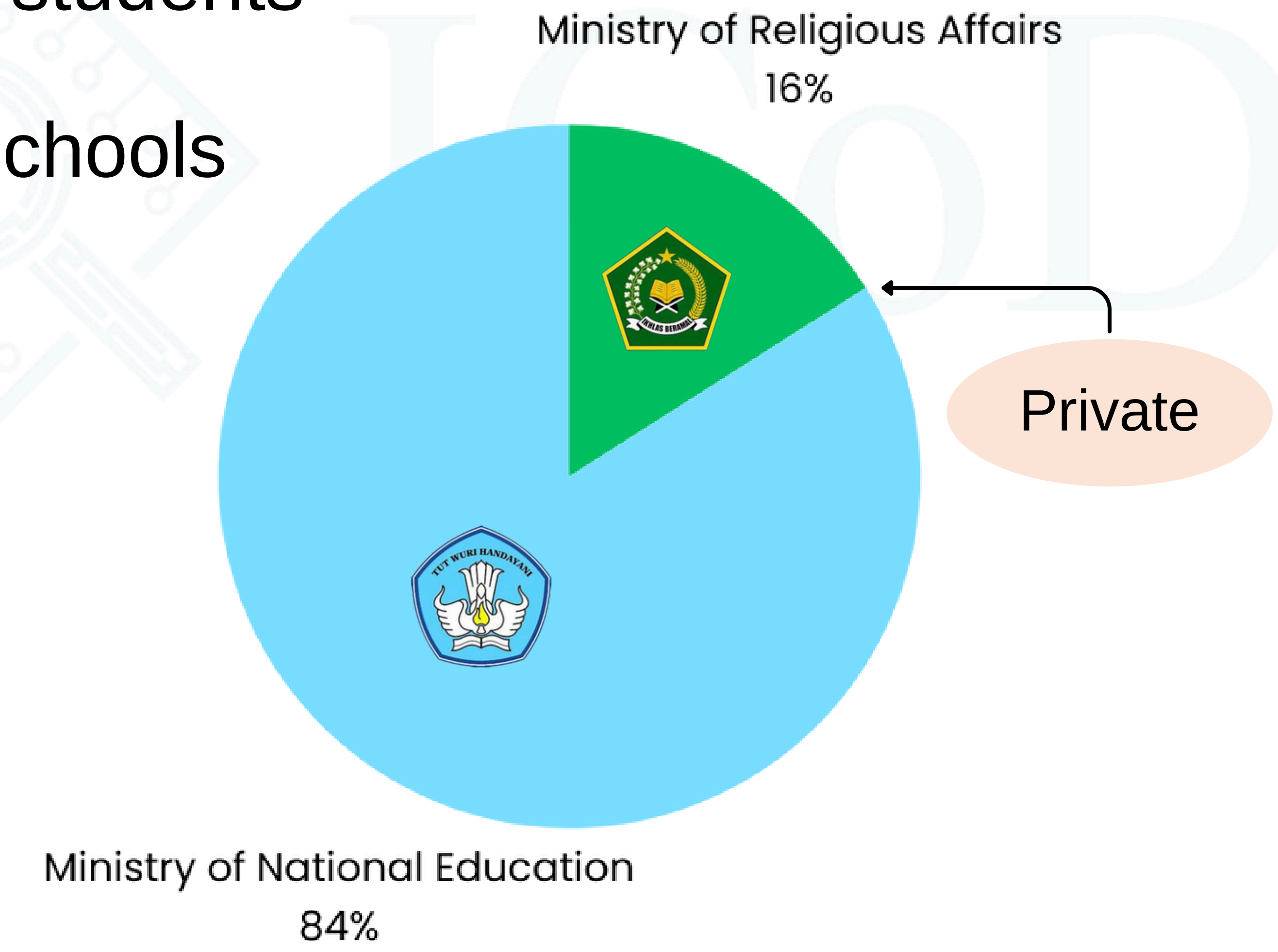
Risna Zahira; Putu Harry Gunawan

Telkom University
Bandung, Indonesia

# BACKGROUND

Based on the 2014 **World Bank report on Education in Indonesia**, 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.

- More than 50 million students
- 2.6 million teachers
- More than 250.000 schools



Each region is committed to improving the quality of education, particularly in West Sumatra.

The high competition for admission to prestigious schools.

The challenges of the admission process require a predictive model to support the readiness of students and schools.

# RELATED WORKS

Supriadi et al. tested 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 cards as predictive features, achieving high accuracy and stable performance.

The study by Wiyono et al. compared three machine learning algorithms Support Vector Machine (SVM), KNN, and Decision Tree to determine the best 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%.

# NOVELTY

**Previous Research**

The focus was on predicting student graduation at a single school or university (single class/target) and was less optimal in selecting features or algorithm parameters, resulting in limited prediction accuracy.

>

Developed a multi-class prediction model using historical data from MTsN Padang Panjang, combining feature selection and algorithm parameter tuning.

The aim of this study is to develop a predictive model for the graduation 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, using the KNN algorithm.

# METHODOLOGY

## WHY KNN?

- Easy to implement
- Flexible for multi-class data
- Good performance
- Low risk of overfitting

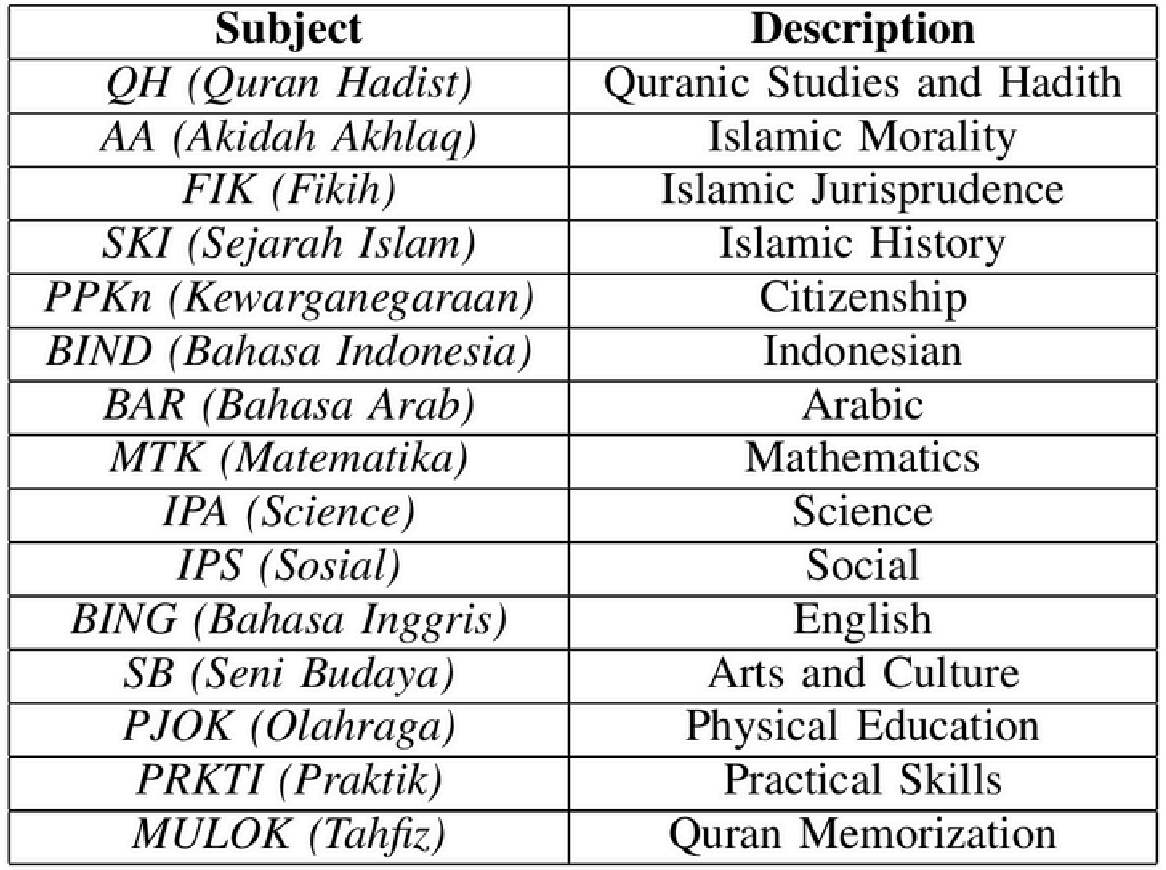
By considering the Euclidean Distance value, the K parameter, and proper feature selection.

## DATASET

The dataset consists of 23 columns and 1.147 rows, collected from 2019 to 2024. The information in the dataset includes:

- Student Identification Number (NIS)
- National Student Identification Number (NISN)
- Student Name
- Gender
- Scores for 15 Subjects
- Final Average Score
- Certificate
- Preferred School
- Advanced School

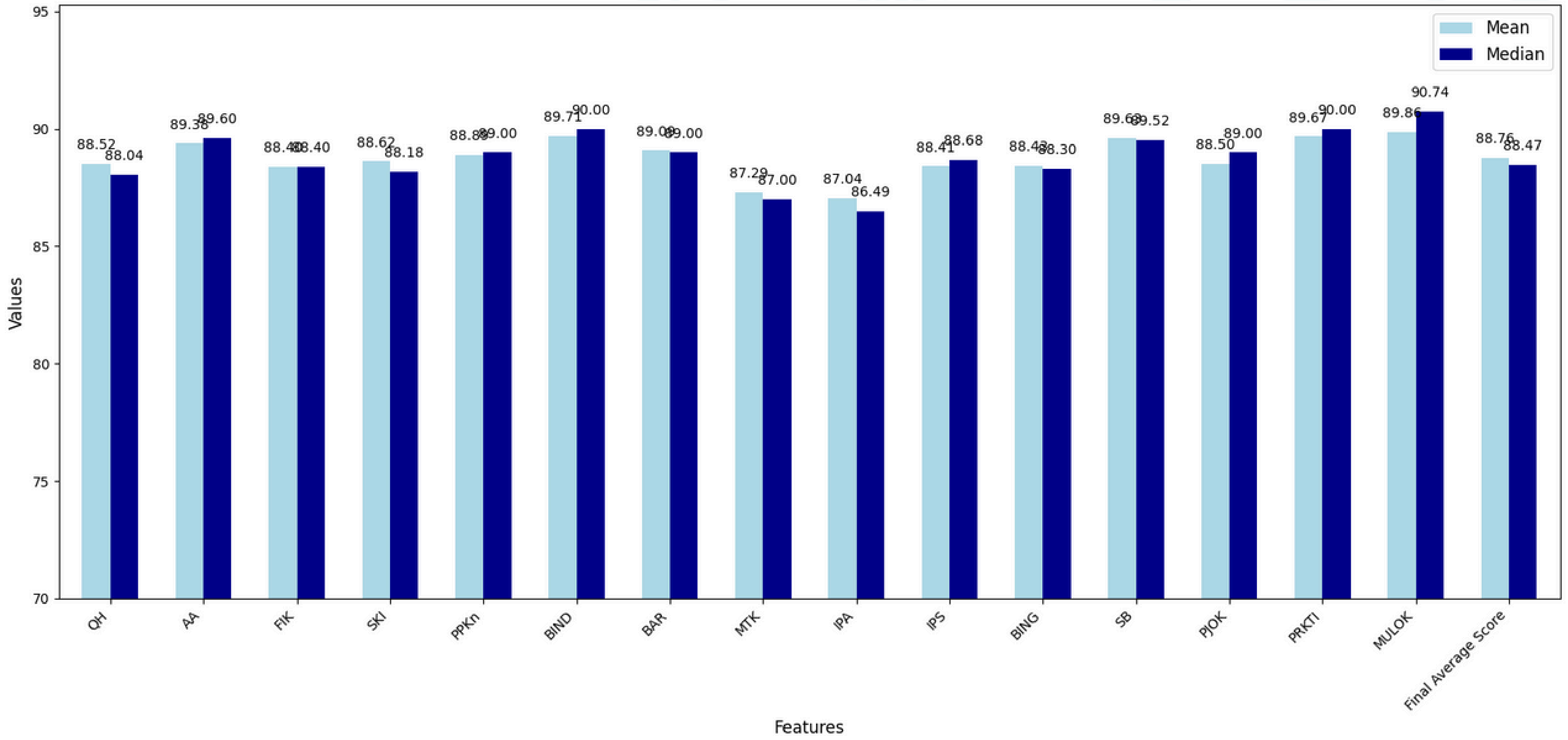
TABLE I



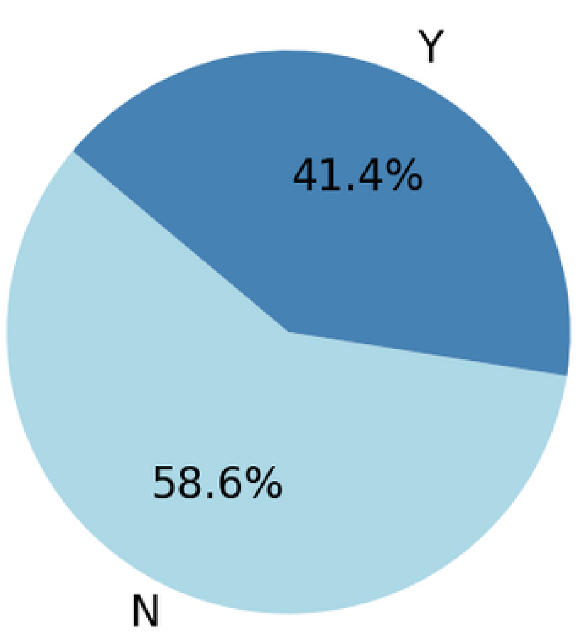
| Subject | Description |
|---|---|
| *QH (Quran Hadist)* | Quranic Studies and Hadith |
| *AA (Akidah Akhlaq)* | Islamic Morality |
| *FIK (Fikih)* | Islamic Jurisprudence |
| *SKI (Sejarah Islam)* | Islamic History |
| *PPKn (Kewarganegaraan)* | Citizenship |
| *BIND (Bahasa Indonesia)* | Indonesian |
| *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 |

TABLE II

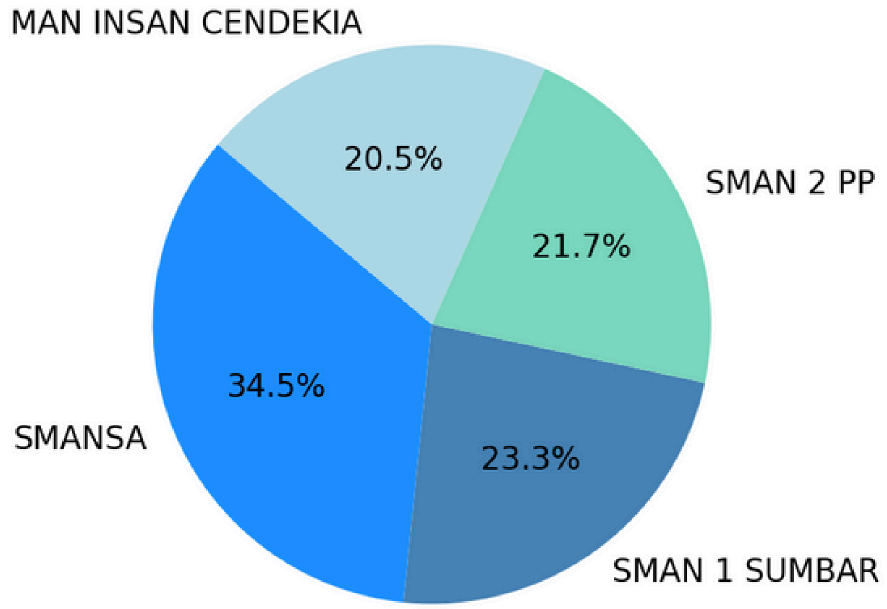
| Advanced School | Initial Value | Final Value |
|---|---|---|
| MAN Insan Cendekia | 132 | 277 |
| SMAN 1 Sumatera Barat (SMAN 1 SUMBAR) | 200 | 344 |
| SMAN 2 Padang Panjang (SMAN 2 PP) | 383 | 387 |
| SMAN 1 Padang Panjang (SMANSA) | 432 | 488 |



Certificate Distribution



School of Interest Distribution



## Data Cleaning

- Removed Irrelevant Columns: NIS (Student Identification Number), NISN (National Student Identification Number), Student Name, and Gender
- Eliminated Null Values and Duplicates

## Encoding

The Certificate column was encoded as:

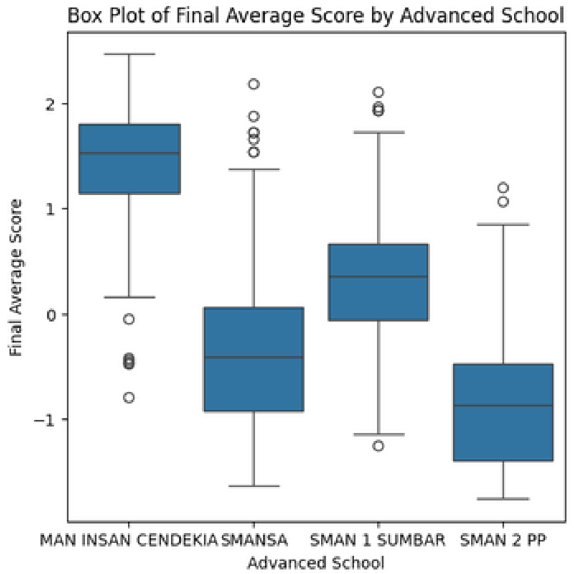
- ”Y” = 1
- ”N” = 0

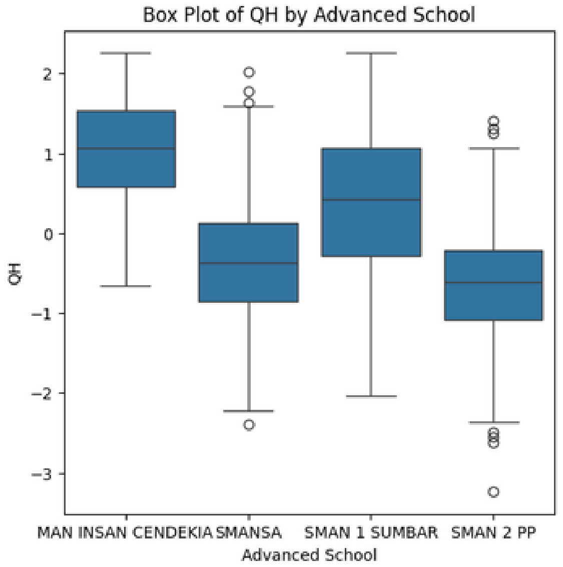
The Preferred School and Advanced School columns were encoded as:

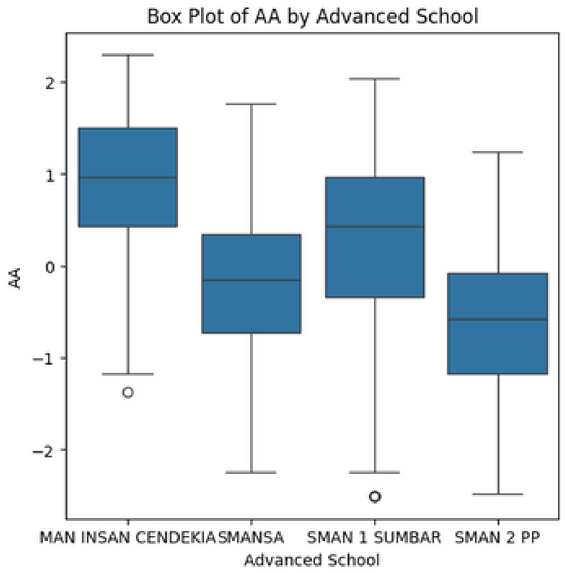
- MAN Insan Cendekia = 1
- SMAN 1 SUMBAR = 2
- SMAN 2 PP = 3
- SMANSA = 4

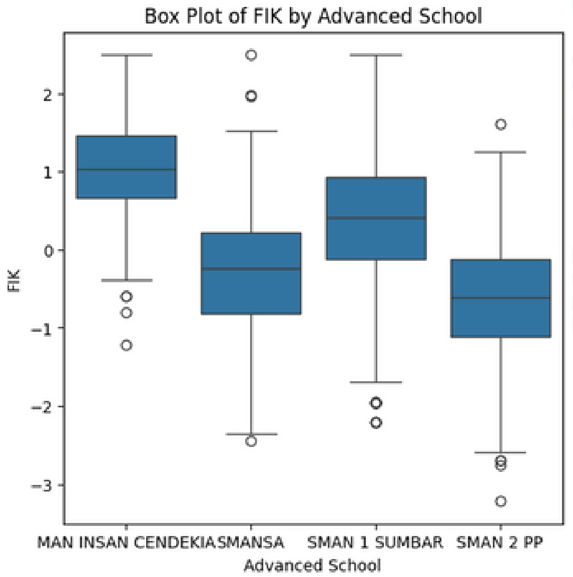
## Handle Outliers and Numerical Data Normalization

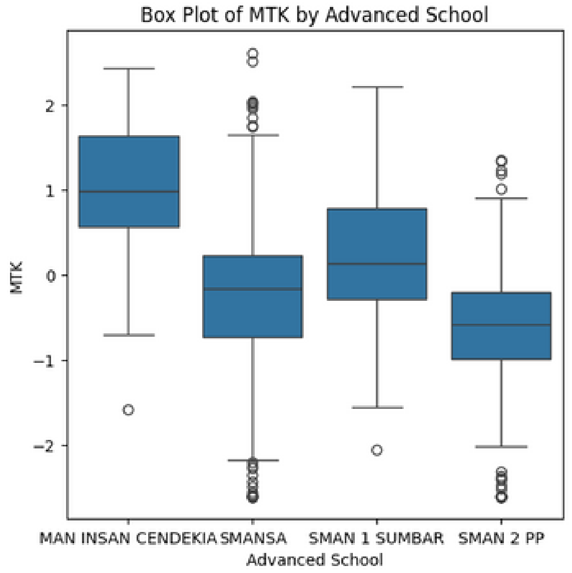
Applying the Interquartile Range (IQR) method to identify and eliminate outliers, ensuring more representative and consistent data, and normalizing numeric data using StandardScaler from sklearn.preprocessing to ensure uniform feature scales, so that the KNN algorithm can work optimally.

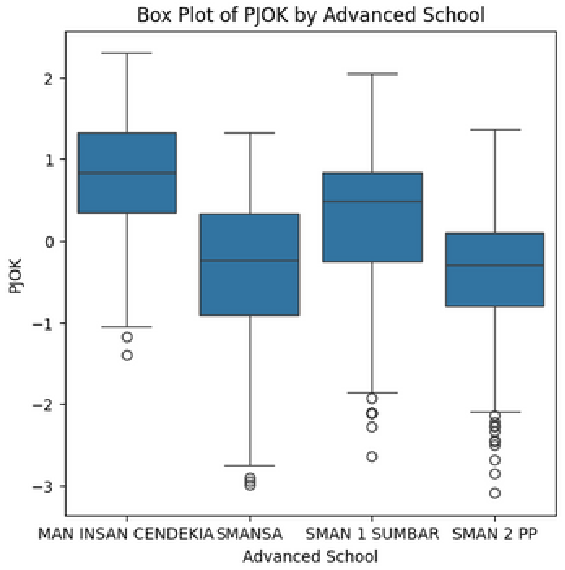

Box Plot of Final Average Score by Advanced School

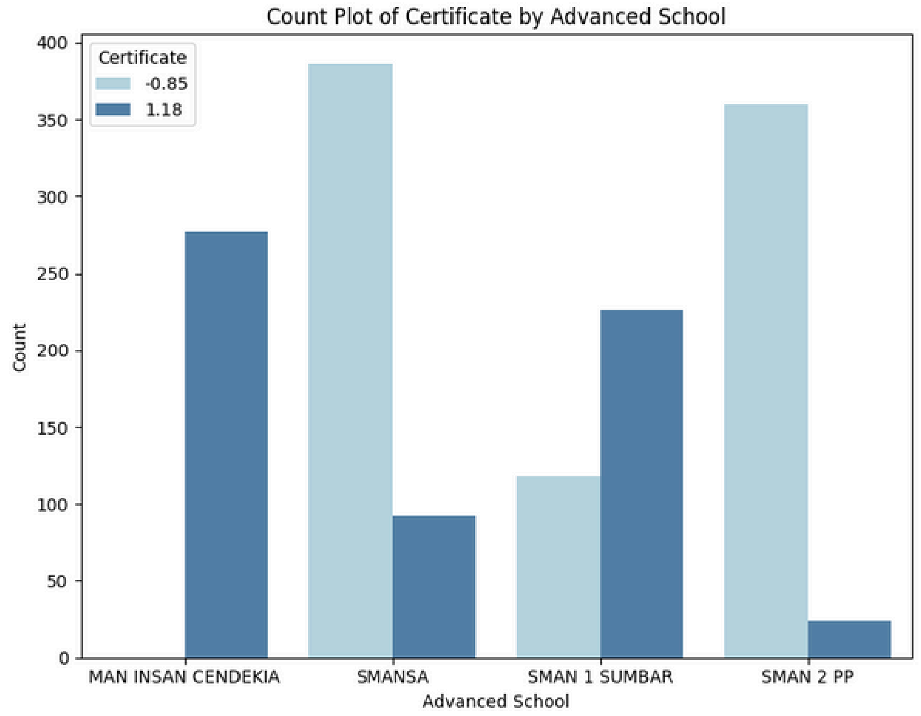

Box Plot of QH by Advanced School

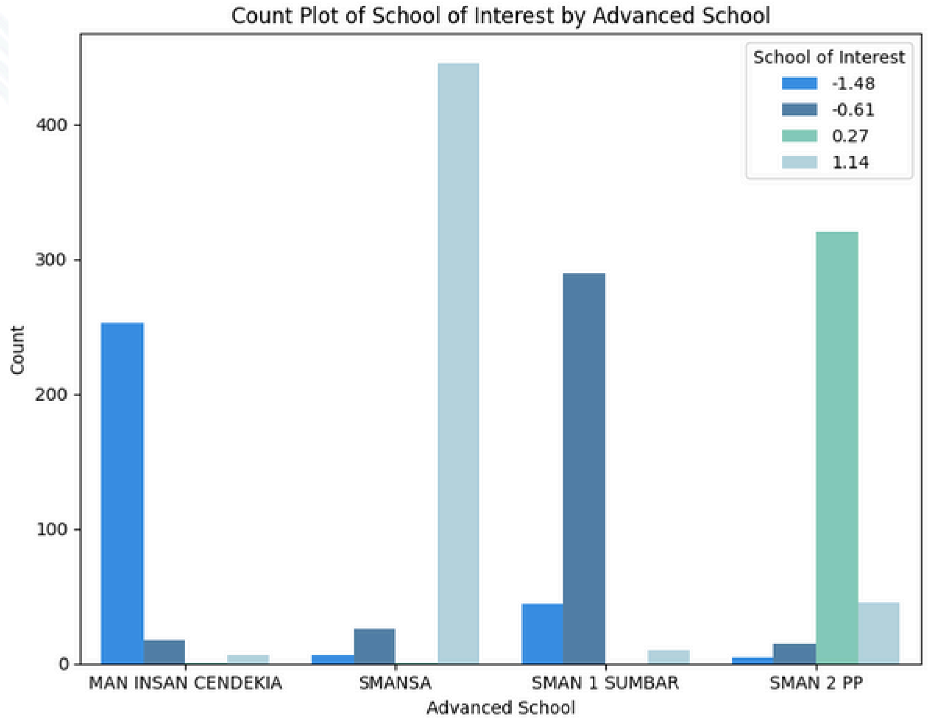

Box Plot of AA by Advanced School


Box Plot of FIK by Advanced School


Box Plot of MTK by Advanced School


Box Plot of PJOK by Advanced School


Count Plot of Certificate by Advanced School


Count Plot of School of Interest by Advanced School

# FEATURES SELECTION

Based on boxplot visualization:

- Final Average Score
- Certificate
- Preferred School
- and the subjects QH, AA, and FIK

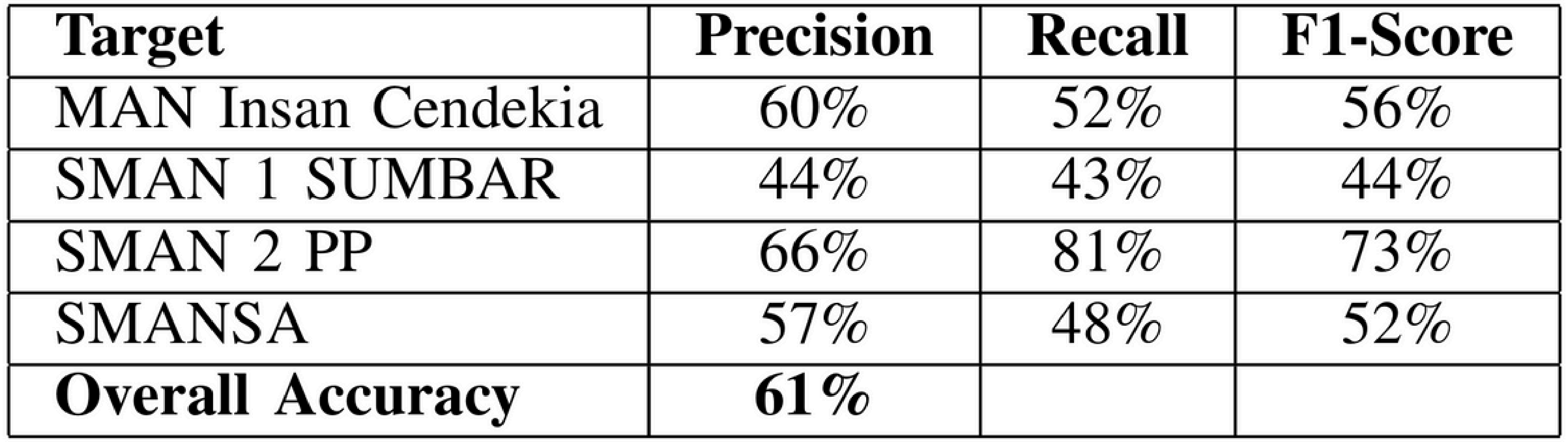
Based on ANOVA test:

- Final Average Score
- Certificate
- Preferred School
- and the subjects IPA and SKI

# RESULT AND DISCUSSION

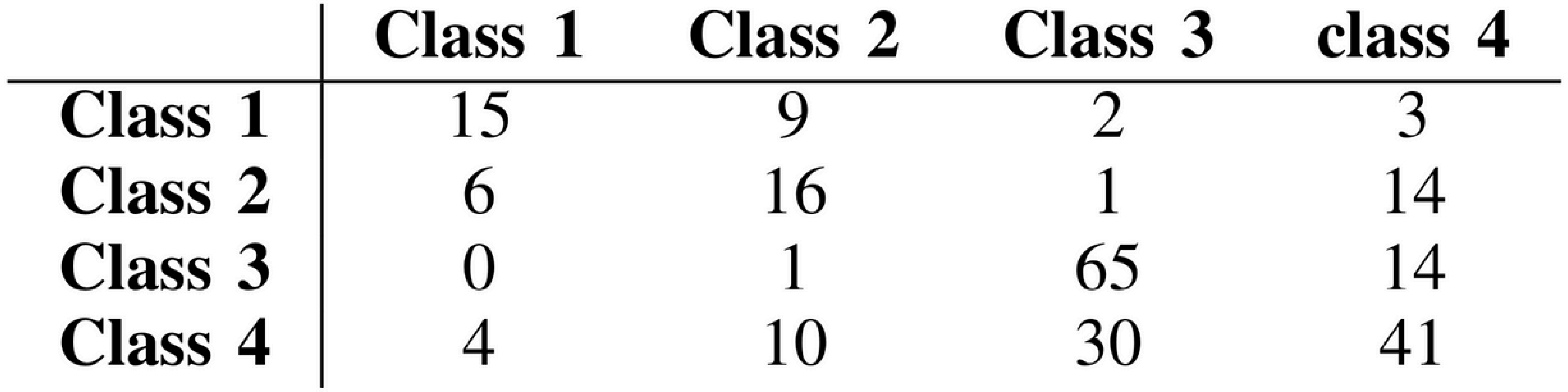
## A. Model with Raw Data

TABLE III



| 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

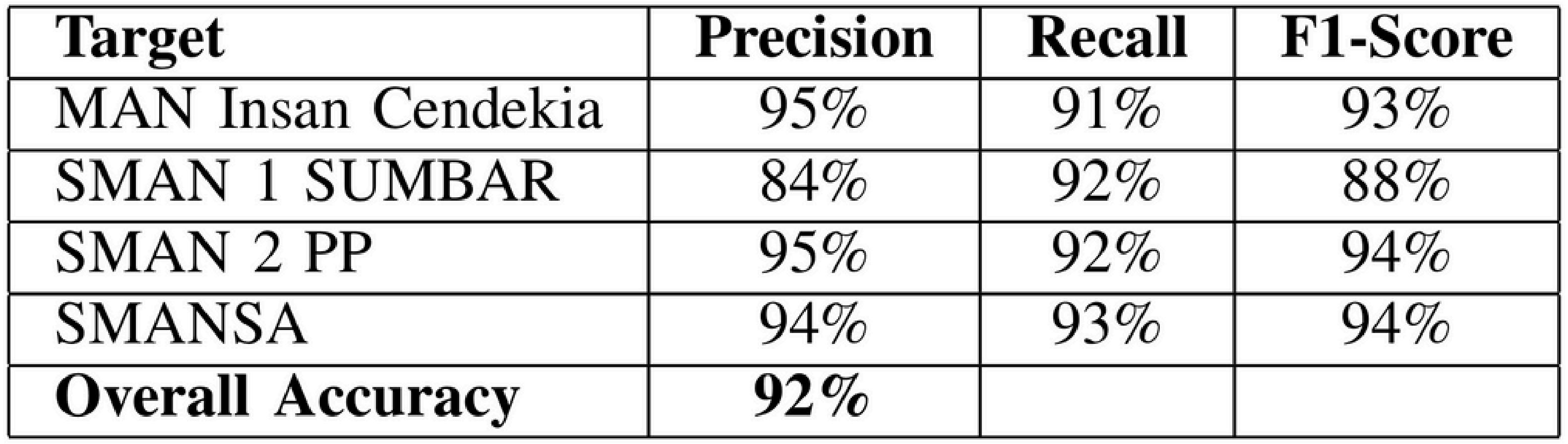


| | Class 1 | Class 2 | Class 3 | class 4 |
|---|---|---|---|---|
| **Class 1** | 15 | 9 | 2 | 3 |
| **Class 2** | 6 | 16 | 1 | 14 |
| **Class 3** | 0 | 1 | 65 | 14 |
| **Class 4** | 4 | 10 | 30 | 41 |

## B. Model with Processed Data

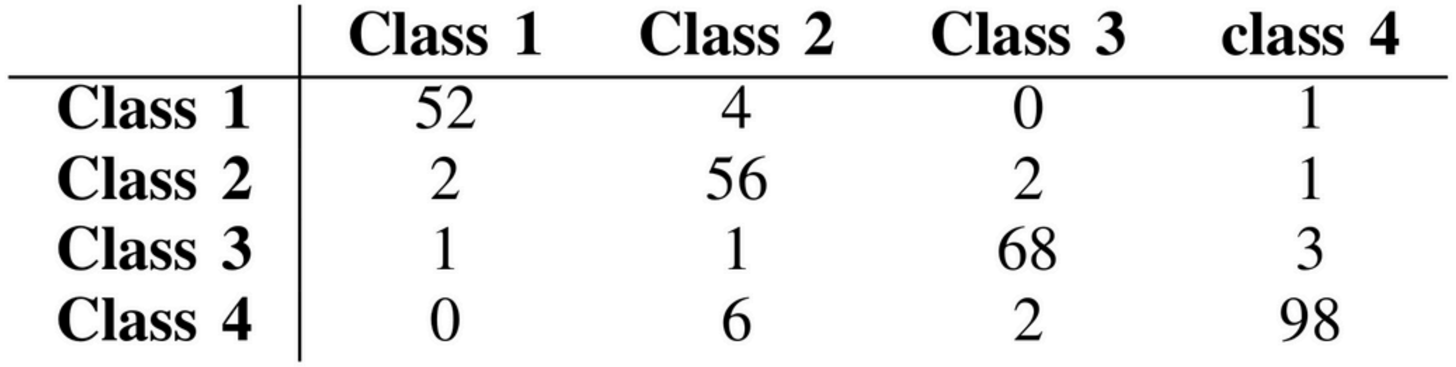
### Using Boxplot

TABLE V



| 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%** | | |

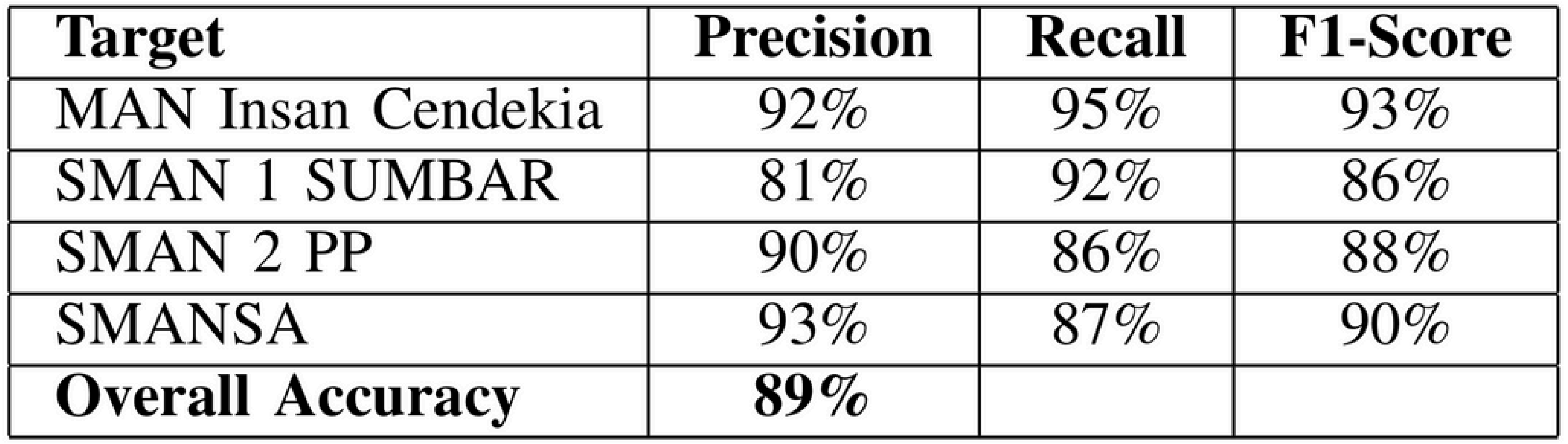
TABLE VI



| | **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 |

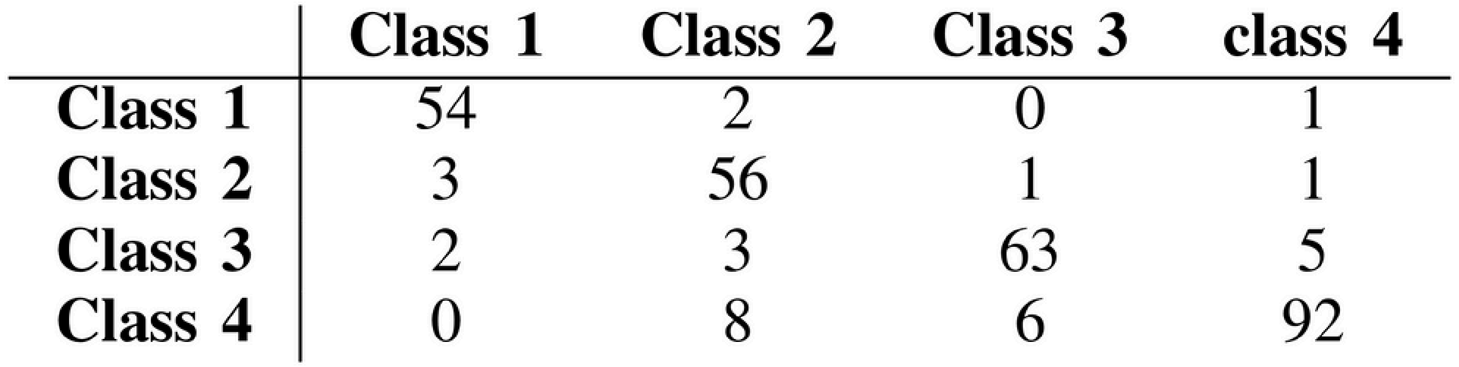
### Using ANOVA Features

TABLE VII



| 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%** | | |

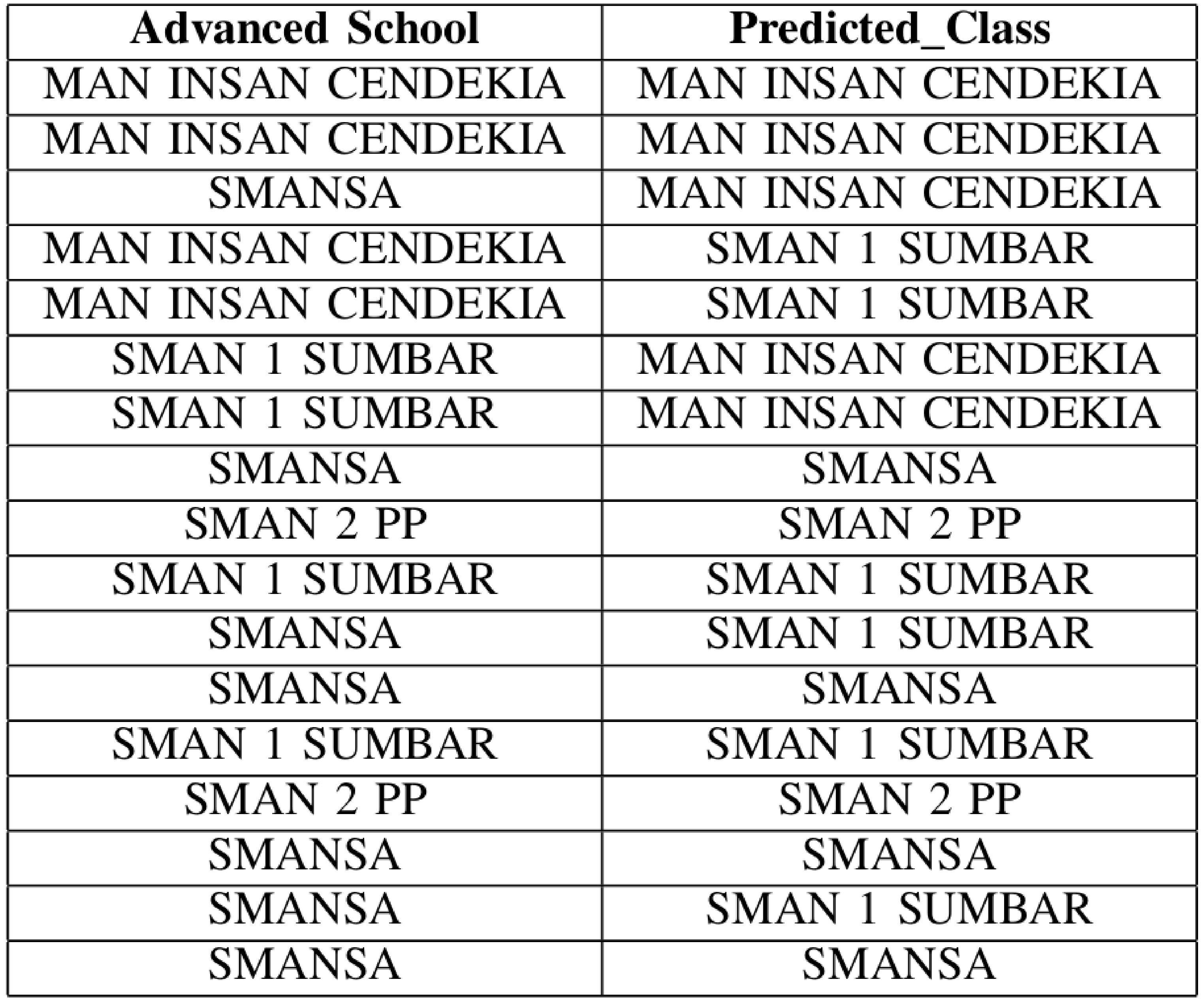
TABLE VIII



| | **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 |

# Model Prediction Result

TABLE IX



| 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 |

# CONCLUSION

This study successfully developed a predictive model using the K-Nearest Neighbors (KNN) algorithm to predict student admission to prestigious schools based on historical data from MTsN Padang Panjang. The model demonstrated excellent performance with an accuracy of 92%. The combination of SMOTE and optimal feature selection (Final Average Score, Achievement Certificates, Preferred School, and and several subjects, namely QH, AA, and FIK) significantly enhanced the dataset quality and model performance, including for minority categories.

## FUTURE WORK

Model Generalizability

Exploration of Other Algorithms

# THANK YOU