The Role of SMOTE in Enhancing Naive Bayes Classification for Major Choice Prediction

Elvi Rahmi[1,a)], Eva Yumami[1,b)]

[1]Department of Software Engineering Engineering, Politeknik Negeri Bengkalis, Bengkalis, Indonesia

[a)]Corresponding author: elvirahmi@polbeng.ac.id
[b)]evayumami@polbeng.ac.id

Abstract. This study examines the application of the Synthetic Minority Oversampling Technique (SMOTE) to address class imbalance within a dataset used for predicting high school major selection. The dataset comprises 468 training instances, including 306 labeled as 'IPA' and 162 labeled as 'IPS'. Despite the implementation of SMOTE, the results reveal no significant enhancement in the predictive performance of the models, as both the SMOTE and non-SMOTE models achieved an accuracy of 100%, an F1-score of 100%, and a recall of 100%. This finding suggests that other factors, such as the selection of relevant features, hyperparameter tuning, and model complexity, may have a more substantial impact on prediction performance. Additionally, the study proposes several recommendations for future research, including conducting a more in-depth feature analysis, exploring alternative classification algorithms with advanced class imbalance handling mechanisms, and performing meticulous hyperparameter optimization to improve overall model performance.

Keywords: SMOTE, class imbalance, naïve bayes classifier, predictive modeling

INTRODUCTION

The choice of academic major is a pivotal decision in a student's life, influenced by a complex interplay of factors such as personal interests, aptitudes, and external influences. In Indonesia, secondary school students typically choose between the natural sciences (IPA) and social sciences (IPS) tracks, a decision that significantly impacts their future academic and career paths.

Previous research has explored the prediction of students' major choices using various machine learning algorithms, with Naive Bayes being a popular choice. While some studies have reported exceptionally high accuracy rates, these results often warrant careful interpretation due to potential overfitting issues. Overfitting occurs when a model becomes too closely tailored to the training data, leading to poor generalization performance on unseen data. Moreover, the prevalence of class imbalance, where one class (e.g., IPA) is significantly more represented than the other (e.g., IPS), can further exacerbate the challenges of building accurate prediction models.

To address these limitations, this study focuses on evaluating the effectiveness of various techniques for handling class imbalance in improving the performance of Naive Bayes models for predicting high school students' major choices. By mitigating the effects of overfitting and class imbalance, this research aims to develop more robust and reliable prediction models that can provide valuable insights for students, educators, and policymakers

METHODS

This research aims to evaluate the effectiveness of implementing class imbalance handling techniques in improving the performance of the Naive Bayes model in predicting high school students' major choices and to identify the contribution of these techniques to the accuracy and reliability of predicting high school students' major choices. To achieve these objectives, this research is divided into five stages.

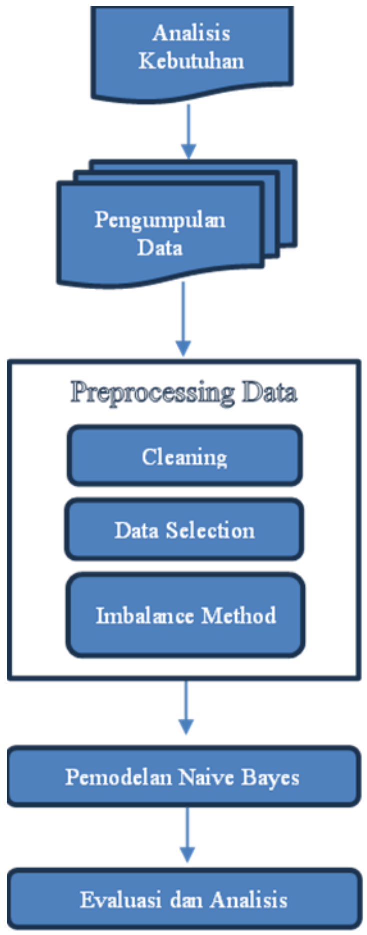


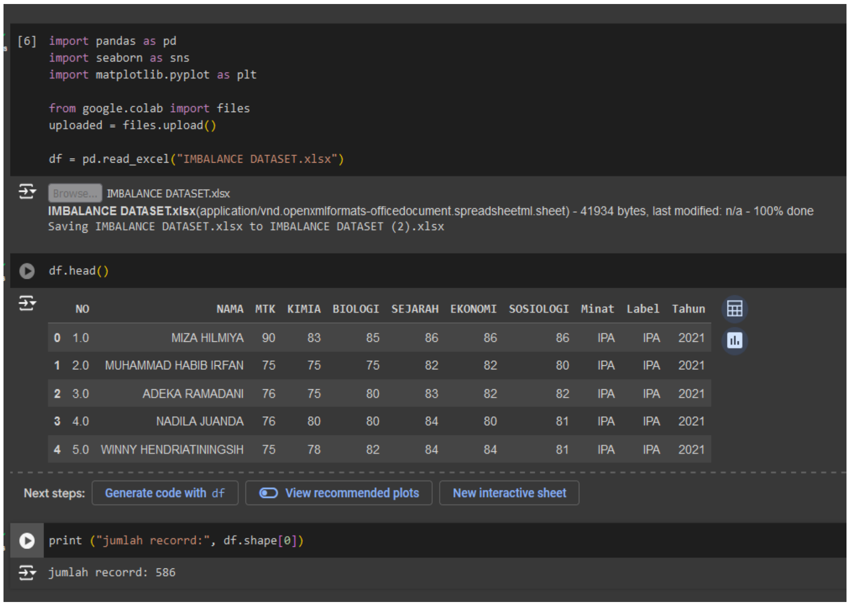
Figure 1. Researh Methodology

Needs Assessment

The research commenced with a needs analysis to identify the specific objectives of the study. The primary focus of this research is to enhance the accuracy of predicting high school students' major choices and to comprehend how class imbalance handling techniques can improve the performance of the Naive Bayes model in the context of student major selection.

Data Collection

The research utilized data from MAN 1 Bengkalis, comprising academic grades in Mathematics, Chemistry, Biology, History, Economics, and Sociology, as well as students' declared major choices between Science (IPA) and Social Sciences (IPS). This data was employed by the school to determine the appropriate placement of students into either the Science or Social Sciences track. The dataset encompassed a three-year period, consisting of 586 instances and 7 attributes.



```
[6] import pandas as pd
    import seaborn as sns
    import matplotlib.pyplot as plt

    from google.colab import files
    uploaded = files.upload()

    df = pd.read_excel("IMBALANCE DATASET.xlsx")
```

Browse... IMBALANCE DATASET.xlsx
IMBALANCE DATASET.xlsx(application/vnd.openxmlformats-officedocument.spreadsheetml.sheet) - 41934 bytes, last modified: n/a - 100% done
Saving IMBALANCE DATASET.xlsx to IMBALANCE DATASET (2).xlsx

```
df.head()
```

	NO	NAMA	MTK	KIMIA	BIOLOGI	SEJARAH	EKONOMI	SOSIOLOGI	Minat	Label	Tahun
0	1.0	MIZA HILMIYA	90	83	85	86	86	86	IPA	IPA	2021
1	2.0	MUHAMMAD HABIB IRFAN	75	75	75	82	82	80	IPA	IPA	2021
2	3.0	ADEKA RAMADANI	76	75	80	83	82	82	IPA	IPA	2021
3	4.0	NADILA JUANDA	76	80	80	84	80	81	IPA	IPA	2021
4	5.0	WINNY HENDRIATININGSIH	75	78	82	84	84	81	IPA	IPA	2021

Next steps: Generate code with df | View recommended plots | New interactive sheet

```
print ("jumlah recorrd:", df.shape[0])
jumlah recorrd: 586
```

Figure 2. Researh Dataset

The dataset was divided into training and testing sets. The training set included data from the 2021-2022 and 2022-2023 academic years, while the testing set contained data from the 2023-2024 academic year. Detailed information on the dataset, including the number of instances, attributes, and sample data, is provided in Tables

Table 1. Data on Student Preferences for Science and Social Studies Majors, 2021-2023

No	Name	Count
1	IPA 2021-2022	186
2	IPS 2021-2022	58
3	IPA 2022-2023	99
4	IPS 2022-2023	40
5	IPA 2023-2024	65
6	IPS 2023-2024	138
Total		**586**

Table 2. Description of Dataset Attributes

No	Attribute	Description	Label
1	Math	Previous semester’s mathematics grades	X1
2	Chemistry	Previous semester’s chemistry grades	X2
3	Biology	Previous semester’s biology grades	X3
4	History	Previous semester’s history grades	X4
5	Economic	Previous semester’s economics grades	X5

6	Sosiology	Previous semester's sosiology grades	X6
7	Preference	Student's Major Preferences Natural Sciences (IPA) = 0 Social Sciences (IPS) = 1	X7
8	Major	IPA = 0 IPS = 1	Y

Table 3 Dataset Sampel

No	X1	X2	X3	X4	X5	X6	X7	Y
1	90	83	85	86	86	86	0	0
2	75	75	75	82	82	80	0	0
3	76	75	80	83	82	82	0	0
4	76	80	80	84	80	81	0	0
5	75	78	82	84	84	81	0	0
6	75	76	79	84	82	81	0	0
7	75	77	77	86	86	80	0	0
8	76	79	80	85	82	80	0	0
9	76	77	78	83	80	80	0	0
10	76	80	82	83	85	80	0	0
11	76	77	81	85	82	82	0	0
12	80	76	79	85	86	85	0	0
13	74	78	81	84	82	82	0	0
14	74	75	76	84	80	79	0	0
15	75	76	75	85	79	81	0	0
...	...	...	...	...	...	...	...	...
572	80	73	80	82	81	80	1	1
573	78	75	76	80	75	80	1	1
574	80	78	84	83	83	81	1	1
575	74	74	75	82	81	82	1	1
576	80	75	78	80	85	82	1	1
577	86	88	86	87	87	87	1	1
578	80	77	82	82	81	80	1	1
579	79	76	78	82	81	80	1	1
580	81	77	80	82	81	82	1	1
581	80	75	78	83	82	80	1	1
582	80	73	82	83	80	80	1	1
583	80	77	82	80	80	80	1	1
584	81	78	78	83	81	80	1	1
585	80	78	80	82	83	81	1	1
586	78	74	79	83	81	80	1	1

Preprocessing Data

. Data pre-processing is a crucial stage in data mining aimed at enhancing dataset quality. This study focuses on addressing class imbalance within the dataset. The dataset used consists of 350 instances of the Natural Sciences track (IPA) and 236 instances of the Social Sciences track (IPS).

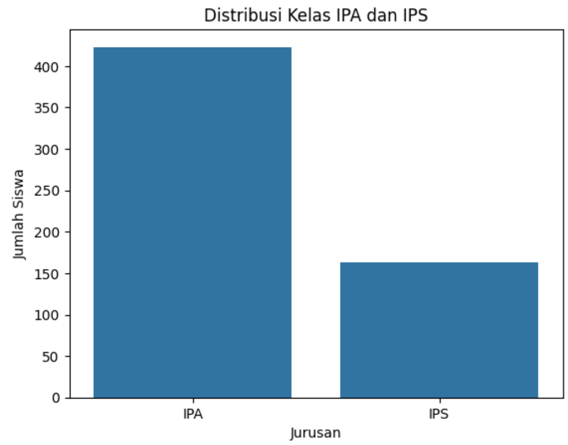


Figure 3 Distribution of Science (IPA) and Social Sciences (IPS) Classes

Figure 4 Subject Grade Distribution and Class Imbalance

The algorithm implemented to address data imbalance in this dataset is SMOTE (Synthetic Minority Oversampling Technique). Before implementing this technique, the first step in preparing the dataset involved checking for any missing data. This process is essential to ensure that data quality is not compromised by gaps that could impact the analysis and final outcomes. The results of this check showed no missing values in any dataset column. Subsequently, categorical feature encoding was performed. At this stage, categorical features in the dataset needed to be converted into numerical format to be usable in the analysis and modeling processes. Encoding is a technique used to transform categorical data into numerical data that can be processed by machine learning algorithms.

The next step involved applying SMOTE to the training data. This technique generates synthetic samples for the minority class by creating interpolations between existing data points. This process increases the number of instances in the minority class, helping the model learn from a more balanced dataset. New data points in the minority class are generated using the following equation. The following figure illustrates the comparison of instance counts between classes in the training data before and after the application of the SMOTE technique.

$$Y' = Y^i + (Y^j - Y^i) * \gamma \quad (1)$$

The following figure illustrates the comparison of instance counts between classes in the training data before and after the application of the SMOTE technique.

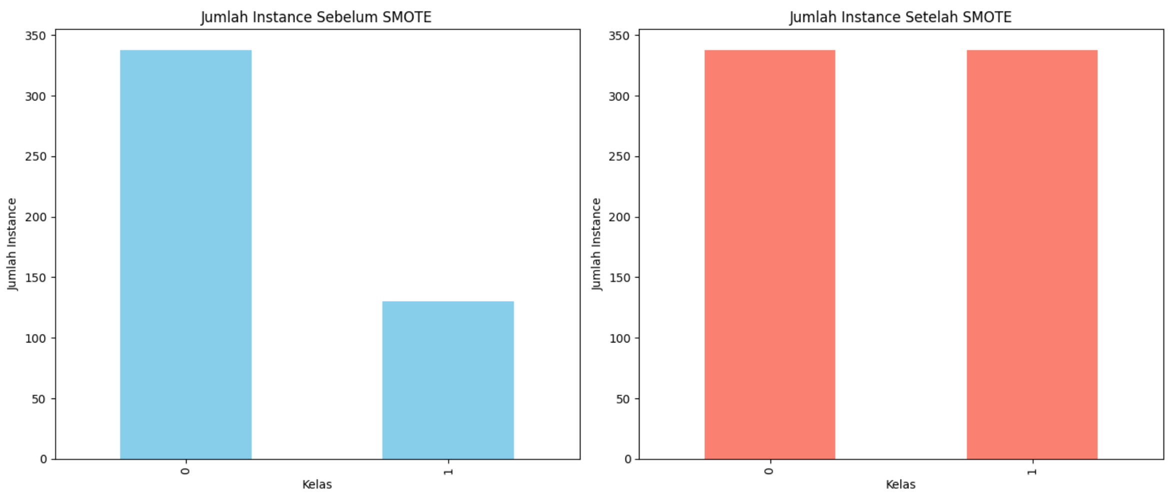


Figure 5 Application of the SMOTE Technique

Naïve Bayes Modelling

After implementing the SMOTE technique, the next step is the development of the Naive Bayes model. Naive Bayes is a probabilistic classification method based on the application of Bayes' theorem, with the assumption of feature independence given the class. The following are the calculation steps to illustrate Naive Bayes modeling in predicting high school major selection.

1. Calculating Prior Probability
   Out of a total of 468 training data instances, there are 306 instances labeled as IPA and 162 instances labeled as IPS. Table 3.4 presents the probability values for the training data.

Table 4. Class Label Probability

No	Label	Probability Value
1	IPA	0.653846153846154
2	IPS	0.346153846153846
Total		**1**

2. Calculating Conditional Probability
   To calculate this conditional probability, it is necessary to determine the mean (μ) and standard deviation (σ) for each feature based on the class.

Table 5 Mean Calculation Values

Label Kelas	MTK	KIMIA	BIOLOGI	SEJARAH	EKONOMI	SOSIOLOGI
IPA	77,35	77,54	79,71	83,57	81,56	81,19
IPS	79,23	78,16	81,30	83,40	82,11	81,25

Table 6 Standard Deviation Calculation Values

Label Kelas	MTK	KIMIA	BIOLOGI	SEJARAH	EKONOMI	SOSIOLOGI
IPA	3,899	2,505	2,478	2,007	2,722	2,359
IPS	3,436	3,264	3,343	2,829	2,362	2,352

With the previously calculated mean (μ) and standard deviation (σ) values, the next step is to calculate the conditional probability for Test Data Case 1 for each feature to determine its class. The calculations are performed using the following equation.

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma_{ik}} e^{\frac{(x_k-\mu_{ik})^2}{2\sigma^2}} \quad (2)$$

The results of the conditional probability calculations for Test Data Case 1 are presented in Table 7 below.

Label Kelas	MTK	KIMIA	BIOLOGI	SEJARAH	EKONOMI	SOSIOLOGI	MINAT	LABEL
	87	81	83	85	83	83	IPA	?
IPA	0,0095	0,0972	0,1055	0,2186	0,2103	0,1937	0,997	5,66E-07
IPS	0,0167	0,1515	0,1919	0,2023	0,2418	0,1973	0,006	9,81E-09

3. Based on the calculations above, the probability value for the interest in 'IPA' for test data case 1 is 5.7×10−75.7×10−7, while the probability value for the interest in 'IPS' is 9.8×10−99.8×10−9. With the higher probability value for 'IPA' compared to 'IPS', it can be predicted that the student with these attribute values is likely to belong to the 'IPA' label.

RESULTS AND DISCUSSION

The results of the study indicate that there is no significant difference in the performance of the model using SMOTE compared to the model not using SMOTE. Both the SMOTE model and the non-SMOTE model achieved an accuracy of 100%, an F1-score of 100%, and a recall of 100%. This suggests that the SMOTE technique does not provide a significant performance improvement in this case.

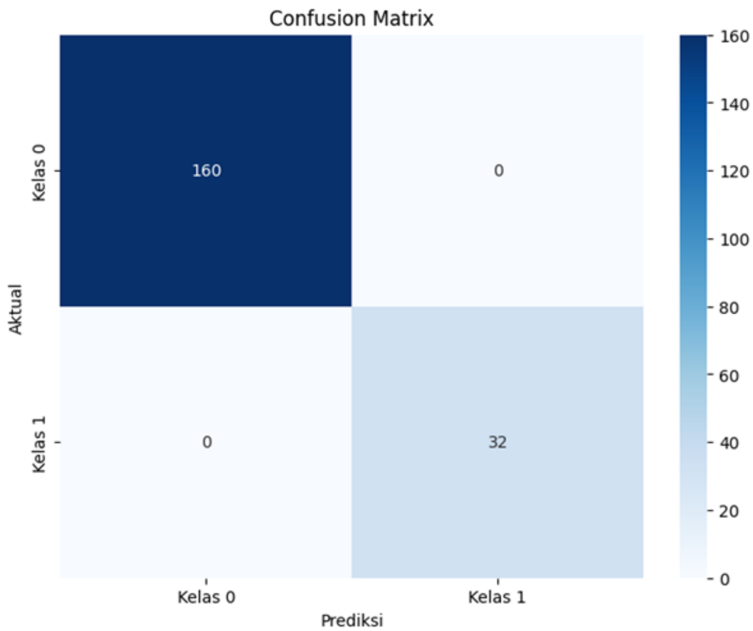


Figure 6 Confusion Matrix for Major Prediction

CONCLUSIONS

Based on the analysis conducted, it can be concluded that the application of the SMOTE oversampling technique on the studied dataset does not result in a significant improvement in the performance of the major choice prediction model. This finding indicates that other factors, such as the selection of relevant features, hyperparameter tuning, and model complexity, have a more dominant influence on prediction performance.

Several recommendations can be made for future research. First, a more in-depth feature analysis is necessary to identify the features that contribute most significantly to the predictions. Second, exploration of other classification algorithms with more advanced mechanisms for handling class imbalance could be undertaken. Third, meticulous hyperparameter optimization of the model may enhance overall model performance.

REFERENCES

[1] Ahmed. Tarek, "*Reservoir Engineering Handbook*", 3rd Edition, Elsevier Inc, 2006.
[2] Amyx. J.W, D.M, Bass. Jr, R.L. Whiting, *"Petroleum Reservoir Engineering-Physicsl Properties*", McGraw-Hill Book Company, New York-Toronto-London, t960.
[3] Clark. N.J, "*Element of Petroleum Reservoir*", Revision Edition, American Institute of Mining,.