Journal of Informatics and Web Engineering

Vol. 3 No. 2 (June 2024) eISSN: 2821-370X

Ensemble-SMOTE: Mitigating Class Imbalance in Graduate on Time Detection

Theng-Jia Law[1*], Choo-Yee Ting[1], Hu Ng[1], Hui-Ngo Goh[1], Albert Quek[1]

[1]Multimedia University, 63000 Cyberjaya, Selangor, Malaysia

**corresponding author: (1181103129@student.mmu.edu.my, ORCiD: 0009-0001-2361-3614)*

Abstract – In education, detecting students graduating on time is difficult due to high data complexity. Researchers have employed various approaches in identifying on-time graduation with Machine Learning, but it remains a challenging task due to the class imbalance in the dataset. This study has aimed to (i) compare various class imbalance treatment methods with different sampling ratios, (ii) propose an ensemble class imbalance treatment method in mitigating the problem of class imbalance, and (iii) develop and evaluate predictive models in identifying the likelihood of students graduating on time during their studies in university. The dataset is collected from 4007 graduates of a university from year 2021 and 2022 with 41 variables. After feature selection, various class imbalance treatment methods were compared with different sampling ratios ranging from 50% to 90%. Moreover, Ensemble-SMOTE is proposed to aggregate the dataset generated by Synthetic Minority Oversampling Technique variants in mitigating the problem of class imbalance effectively. The dataset generated by class imbalance treatment methods were used as the input of the predictive models in detecting on-time graduation. The predictive models were evaluated based on accuracy, precision, recall, F0.5-score, F1-score, F2-score, Area under the Curve, and Area Under the Precision-Recall Curve. Based on the findings, Logistic Regression with Ensemble-SMOTE outperformed other predictive models, and class imbalance treatment methods by achieving the highest average accuracy (87.24), recall (92.50%), F1-score (91.30%), and F2-score (92.02%) from 6th until 10th trimester. To assess the effectiveness of class imbalance treatment methods, Friedman test is performed to determine on significant difference between the models after applying Shapiro-Wilk test in normality test. Consequently, Ensemble-SMOTE is ranked as the top-performers by achieving the lowest value in the average rank based on the performance metrics. Additional research could incorporate and examine more complicated approaches in mitigating class imbalance when the dataset is highly imbalanced.

Keywords—Graduate on Time, In-University, Class Imbalance, Artificial Intelligence, Machine Learning

Received: 10 April 2024; Accepted: 09 May 2024; Published: 16 June 2024

I. INTRODUCTION

Students who graduate on time (GOT) are those who completed their studies timely within the time frame specified by the university [1, 2]. Its significance extends beyond individual achievement, serving as a metric for evaluating institutional quality and performance [1, 3]. However, the journey towards GOT is often beset by challenges when the academic success is multifaceted [4], with students grappling to maintain academic momentum and overcome obstacles that may impede timely completion. These challenges manifest in various forms, including the accumulation of failed courses over semesters [3, 5]. While certain measures such as adjusting passing thresholds or attendance tracking might seem to bolster graduation rates, concerns linger regarding their impact on the overall quality of graduates [6]. Hence, the ability to identify students at risk of delayed graduation becomes imperative, facilitating proactive interventions to support them and uphold the quality of graduates [7, 8].

MULTIMEDIA UNIVERSITY

Journal of Informatics and Web Engineering
https://doi.org/10.33093/jiwe.2024.3.2.17

Published by MMU Press. URL: https://journals.mmupress.com/jiwe

In response to the need for targeted support, researchers have increasingly turned to Machine Learning (ML) integration to identify at-risk students during their university tenure. Nonetheless, detecting GOT proves intricate due to the intricate nature of educational data, characterized by high dimensionality and class imbalance. The latter poses a significant hurdle, stemming from uneven distributions of graduation rates across student cohorts. Whether due to varying speeds of progress or other factors [3, 9], this imbalance complicates accurate identification of at-risk students and hampers the implementation of tailored interventions. Without effective strategies to address class imbalance, educational institutions risk inefficiencies in resource allocation, potentially yielding suboptimal outcomes for both students and institutions alike. Thus, rectifying class imbalance becomes paramount to enhancing the efficacy of timely graduation interventions.

Researchers have thus explored diverse class imbalance treatment methods to tackle this issue, aiming to extract insights from skewed data distributions. While class imbalance treatment with data-level methods is commonly employed, scant attention has been paid to investigating the impact of different sampling ratios in mitigating class imbalance. Moreover, the potential benefits of aggregating oversampling, undersampling, or hybrid methods through ensemble techniques remain largely unexplored.

To address these challenges, the objectives of this study were:

1. To compare various class imbalance treatment methods with different sampling ratios in mitigating the problem of class imbalance.
2. To propose an ensemble class imbalance treatment method in mitigating the problem of class imbalance.
3. To develop and evaluate predictive models in identifying the likelihood of students graduating on time during their studies in university.

II. NUMBER OF MINORITY CLASS IN CLASS IMBALANCE

Table 1. Number of Minority Class in Dataset used by Researchers in Mitigating Class Imbalance

Author	Less than 10%	Less than 20%	Less than 30%	Less than 40%	Less than 50%
[10]		✓	✓		
[11]					✓
[12]	✓	✓	✓		
[13]	✓				
[14]			✓		
[15]	✓				
[16]			✓		
[17]	✓				
[18]				✓	
[19]					✓
[20]	✓				
[21]	✓				
[22]	✓				
[23]	✓				
[24]			✓		
[25]	✓				
[26]	✓		✓	✓	
[27]	✓				
[28]		✓			
[29]				✓	
[30]	✓				
[31]				✓	

Table 1 illustrates the prevalence of minority classes within datasets utilized by researchers to tackle the challenge of class imbalance in identifying GOT. This imbalance arises when the dominance of the majority class eclipses the presence of minority classes, resulting in biased predictive models that yield unpredictable outcomes [12, 15, 16, 27, 28, 30]. Addressing the representation of minority classes becomes imperative within this imbalance paradigm.

In binary scenarios, researchers meticulously define the thresholds for high class imbalance, typically when the minority class constitutes less than 8% of the dataset [17, 20, 21, 23, 25] while acknowledging imbalance when the minority class falls below 35% [18, 31]. For example, in a recent study, a random split allocated 85% of the dataset to training data, revealing a distribution of 53.38% for on-time graduations and 46.62% for late graduations [11]. The delineation between moderate and extreme class imbalance is finely drawn, with percentages such as 24.82% indicating moderate imbalance and 12.41% signifying extreme imbalance [10]. Furthermore, class imbalance treatments become imperative, as evidenced by efforts to address the scenario where 30.27% of students did not graduate on time in a dataset from the Academic Administration Bureau of Universitas Advent Indonesia (UNAI) [14]. Al-Shabandar et al. [16] addressed the class imbalance when the Harvard and Open University Learning Analytics Dataset (OULAD) dataset collected contains 78% failing students and 22% students succeeding. Importantly, challenges arise when the distribution of the minority class falls below 10% of the dataset [22, 27], exacerbating the influence of the majority class. These observations underscore the critical importance of tailored interventions to rebalance datasets and mitigate the adverse effects of class imbalance.

Conversely, in multi-class scenarios, researchers confront a different set of challenges surrounding class imbalance. Instances where only a fraction of training data, such as 9.92%, 11.41%, or 26.49%, represents the minority class highlight the complexity of addressing imbalance in a multi-class context, as observed in studies focusing on student performance at MARA Technological University (UiTM) [12]. Moreover, highly imbalanced datasets, where only 9% and 6% of samples belong to the Excellent class in datasets from Iran and Portugal respectively, underscore the need for nuanced approaches to handle disparities across multiple classes [13]. Furthermore, the identification of multi-class imbalance, as noted by Said et al. [15], where approximately 7% of the dataset represents minority classes such as students who did not graduate on time, emphasizes the ongoing challenge of equitable representation across diverse class categories. Thus, the exploration of class imbalance in multi-class scenarios necessitates tailored methodologies and interventions to ensure fair and accurate predictive modeling.

The exploration of class imbalance within datasets utilized for detecting GOT reveals a nuanced landscape where minority classes often face underrepresentation. Researchers navigate various degrees of imbalance, ranging from moderate to extreme, and grapple with the challenges posed by highly skewed distributions. By addressing these disparities and implementing appropriate class imbalance treatments, such as data preprocessing techniques and model adjustments, researchers endeavor to mitigate bias and enhance the robustness of predictive models. Moving forward, a concerted effort to acknowledge and rectify class imbalances will be essential in ensuring the reliability and fairness of predictive analytics in the domain of GOT detection.

III. CLASS IMBALANCE TREATMENT TECHNIQUES IN MITIGATING CLASS IMBALANCE

Table 2. Class Imbalance Treatment Techniques used by Researchers in Mitigating Class Imbalance of Graduate on Time

Author	Random Oversampling	Synthetic Minority Over-sampling Technique	Adaptive Synthetic Sampling	Borderline Synthetic Minority Over-sampling Technique	Support Vector Machine Synthetic Minority Over-sampling Technique	Random Undersampling	NearMiss	One-sided Selection	Tomek Links	Edited Nearest Neighbors	Synthetic Minority Over-sampling Technique and Tomek Links	Synthetic Minority Over-sampling Technique and Edited Nearest Neighbors	Synthetic Minority Over-sampling Technique and Nominal
[10]	✓	✓				✓							
[12]		✓	✓	✓									

[13]	✓	✓		✓	✓				✓	✓	✓	✓	
[14]		✓											
[15]		✓											
[16]		✓											
[17]	✓	✓				✓					✓	✓	
[18]		✓	✓	✓	✓								
[19]		✓											
[20]		✓											
[21]	✓	✓	✓	✓	✓	✓					✓	✓	✓
[23]	✓	✓				✓						✓	
[24]		✓	✓	✓			✓	✓	✓	✓	✓		
[25]		✓	✓	✓									
[26]		✓											
[27]	✓	✓	✓							✓		✓	
[28]		✓		✓	✓				✓	✓	✓	✓	
[29]			✓										
[30]	✓	✓										✓	
[31]		✓											
[32]		✓											
[33]		✓											
[34]		✓									✓		

Once class imbalance is identified, various techniques are implemented to mitigate this issue, with a notable approach being data-level techniques, as depicted in Table 2. These techniques aim to balance the data before model development, thereby improving model performance [17]. Among these techniques, class imbalance treatment was commonly utilized, including oversampling [10, 12 - 21, 23 - 34], undersampling [10, 13, 17, 21, 23, 24, 27, 28], and hybrid techniques [13, 17, 21, 23, 24, 27, 28, 30, 34].

Among the widely used oversampling techniques, Synthetic Minority Oversampling Techniques (SMOTE) stands out for its effectiveness in mitigating class imbalance in the context of GOT. Researchers implementing SMOTE have successfully overcome class imbalance and improved prediction accuracy by 1 – 2% accuracy across the entire dataset [14]. However, in highly imbalanced scenarios with less than 5% minority class, SMOTE may not achieve substantial improvements in balancing True Positive Rate (TPR) and True Negative Rate (TNR) compared to hybrid methods [17]. Nevertheless, SMOTE has shown promising results in terms of Area Under the Curve (AUC), with predictive models achieving the highest AUC of 69% compared to other class imbalance techniques when dealing with minority class instances as low as 4.7% [23]. Additionally, predictive models enhanced with SMOTE have demonstrated the highest average F1-score of 76.20% after implementing feature selection with Pearson correlation [25]. Moreover, in their study, researchers observed an improvement in average recall from 75.20% to 88.20% by increasing the number of minority class instances through SMOTE. However, the performance may degrade when the classification becomes more challenging due to increased noise in the minority class and overlapping regions between majority and minority classes [25].

To address the challenges introduced by oversampling methods, particularly SMOTE, hybrid methods combining oversampling and undersampling of class distribution have been proposed. Among these, SMOTE and Tomek Links (SMOTE-Tomek) [11, 34], and SMOTE and Edited Nearest Neighbors (SMOTE-ENN) [17, 21] have demonstrated superior performance in dealing with high class imbalance. In the study of Mduma [17], SMOTE-ENN outperformed other class imbalance treatment methods by eliminating misclassified samples using its nearest neighbors after applying SMOTE. Comparative analysis across different sampling ratios, ranging from 10% to 100%, revealed the superiority of SMOTE-ENN, achieving the highest Geometric Mean (G-Mean) of 92.70% with a standard deviation of 0.005 in multi-class scenarios when the sampling ratio was 100% [21]. These findings highlighted the effectiveness of hybrid methods in addressing class imbalance and improving the performance of predictive models in challenging scenarios.

A notable gap in existing studies on mitigating class imbalance problems to the limited exploration and comparison of different sampling ratios. While various class imbalance treatment methods have been implemented to address class imbalance, there is a lack of comprehensive studies that systematically compare the effectiveness of different sampling ratios across diverse class imbalance treatment methods. Understanding how different sampling ratios impact the performance and generalizability of predictive models is essential to select the most suitable approach in handling class imbalance. Additionally, there is a dearth of studies that investigate the aggregation of different class imbalance treatments. While individual techniques such as oversampling, undersampling, and hybrid methods have been extensively studied, little attention has been given to the potential benefits and challenges of combining these techniques in an integrated data-level framework. Investigating the synergistic effects of aggregating multiple class imbalance treatments could provide valuable insights into optimizing predictive model performance and enhancing the robustness of class imbalance mitigation strategies.

IV. MACHINE LEARNING TECHNIQUES IN DETECTING GRADUATE ON TIME

Table 3. Machine Learning Techniques used by Researchers in Detecting Graduate on Time

Author	Logistic Regression	Naïve Bayes	Decision Tree	Support Vector Machine	K-Nearest Neighbors	Random Forest	Adaptive Boosting	Extreme Gradient Boosting	Light Gradient Boosting Machine	Categorical Boosting
[10]						✓	✓			
[11]						✓				
[12]	✓	✓	✓	✓	✓					
[13]	✓	✓	✓	✓	✓	✓		✓		
[15]	✓	✓	✓	✓	✓	✓				
[16]	✓					✓				
[17]	✓					✓				
[18]	✓		✓	✓	✓					
[19]						✓				
[20]	✓	✓	✓	✓						
[21]		✓		✓	✓	✓				
[22]		✓	✓	✓	✓	✓	✓	✓		
[23]	✓			✓						
[25]	✓		✓	✓		✓			✓	
[26]					✓					
[27]	✓		✓	✓	✓	✓				
[28]	✓			✓	✓	✓				
[29]	✓	✓	✓	✓	✓					
[30]								✓		✓
[31]			✓	✓		✓				
[32]		✓				✓				
[33]		✓		✓						
[34]				✓		✓				

To detect the students who graduate on time, researchers have developed various ML techniques after the class imbalance treatment as illustrated in Table 3. In addressing the challenges of detecting GOT and the class imbalance, ML techniques such as Random Forest (RF) have gained prominence. Researchers recommended RF due to its efficient implementation with ensemble of randomized Decision Trees (DT) and its capability to mitigate overfitting post class imbalance treatment [11, 17, 19]. For instance, in a study considering SMOTE-applied clustered training datasets created by latent class analysis (LCA) based on student information, RF emerged as the top performer, showcasing improvements ranging from 6% to 10% across all classes compared to individually trained models [19]. Other than that, the researchers achieved 91.75% accuracy, 92.52% precision, 94.74% recall, and 96.58% F1-score after applying SMOTE-Tomek [11]. However, the researchers reported the superiority of Logistic Regression (LR) over RF in terms of G-Mean and F1-score over RF in correctly classifying student dropouts and minimizing misclassifications [17].

The superior performance of LR is widely acknowledged in detecting GOT. Researchers have demonstrated the use of ensemble models in which three top-performers, LR were integrated with the bootstrap aggregation, achieving the highest accuracy of 95.45%, as well as the highest precision and recall with balanced dataset generated by SMOTE [18]. In the study by Buniyamin et al. [12], LR and Support Vector Machine (SVM) showcased better performance than K-Nearest Neighbors (KNN) when class imbalance treatment method, Borderline SMOTE was used. Nevertheless, SVM outperformed LR by achieving higher AUC (69%) when SMOTE is applied to oversample the minority class in the dataset [23]. These findings highlight the varied effectiveness of different class imbalance treatment methods depending on the characteristics and predictive models, as emphasized by Buniyamin et al. [12]. To address this challenge, researchers recommend exploring alternative class treatment methods and ensemble methods such as Extreme Gradient Boosting (XGBoost) to enhance the performance on the imbalanced datasets.

The exploration of ensemble models delves into boosting algorithms like XGBoost and Categorical Boosting (CatBoost), which have demonstrated enhanced performance in predicting on-time graduation amidst class imbalance challenges. XGBoost, for instance, stands out for its ability to significantly improve various performance metrics such as accuracy, recall, F1-score, and AUC. This improvement stems from XGBoost's adeptness at leveraging misclassified data during training to iteratively generate additional Decision Trees, thereby refining the model's predictive capabilities [28]. Additionally, researchers have observed notable advancements by combining XGBoost with SMOTE and CatBoost with RandomOverSamplerSMOTEENN through risk priority rules [30]. By leveraging these two boosting algorithms, the researchers have achieved notable improvements in the precision and recall of predictions, particularly in identifying students at risk of dropping out. These findings underscore the potential of ensemble approaches, particularly boosting algorithms like XGBoost and CatBoost, in mitigating the challenges posed by class imbalance and improving the accuracy and reliability of predictive models in educational settings.

Despite the implementation of ML techniques for detecting on-time graduation within imbalanced datasets, challenges persist due to the intricate and varied nature of educational data. Researchers stress the complexity inherent in educational datasets, encompassing factors ranging from academic performance to socio-economic backgrounds and individual circumstances. Moreover, the dynamic nature of educational environments, including changes in curriculum, teaching methodologies, and student demographics, further complicates accurate prediction of graduation outcomes. Researchers also emphasize the inherent difficulties in effectively implementing ML techniques within educational contexts, noting that the effectiveness of ML algorithms can vary significantly based on data characteristics such as dataset size, class distribution imbalance, and the presence of noise or outliers [12, 35, 36]. Thus, identifying optimal ML techniques and tailored class imbalance treatment strategies for specific datasets remains a significant challenge in achieving effective detection of on-time graduation.

V. METHODS

This section explains the methods used to (i) compare various class imbalance treatment methods in mitigating the problem of class imbalance with different sampling ratios, (ii) propose an ensemble class imbalance treatment method in mitigating the problem of class imbalance, and (iii) develop and evaluate predictive models in identifying the likelihood of students graduating on time during their studies in university. The approaches used in this work are illustrated in Figure 1. Following the completion of data preprocessing and feature selection, various class imbalance treatment methods are developed and compared with different sampling ratios. Furthermore, an ensemble algorithm aggregating the SMOTE variants is implemented to detect GOT effectively in context of class imbalance. The dataset generated by the class imbalance treatment methods are used as the input for the predictive models with the important variables identified. The predictive models are evaluated based on performance metrics. To further explain the optimal class imbalance treatment method, statistical test is performed. The lowest value in average rank is awarded to the top-performing class imbalance treatment methods for mitigating class imbalance in detecting GOT.

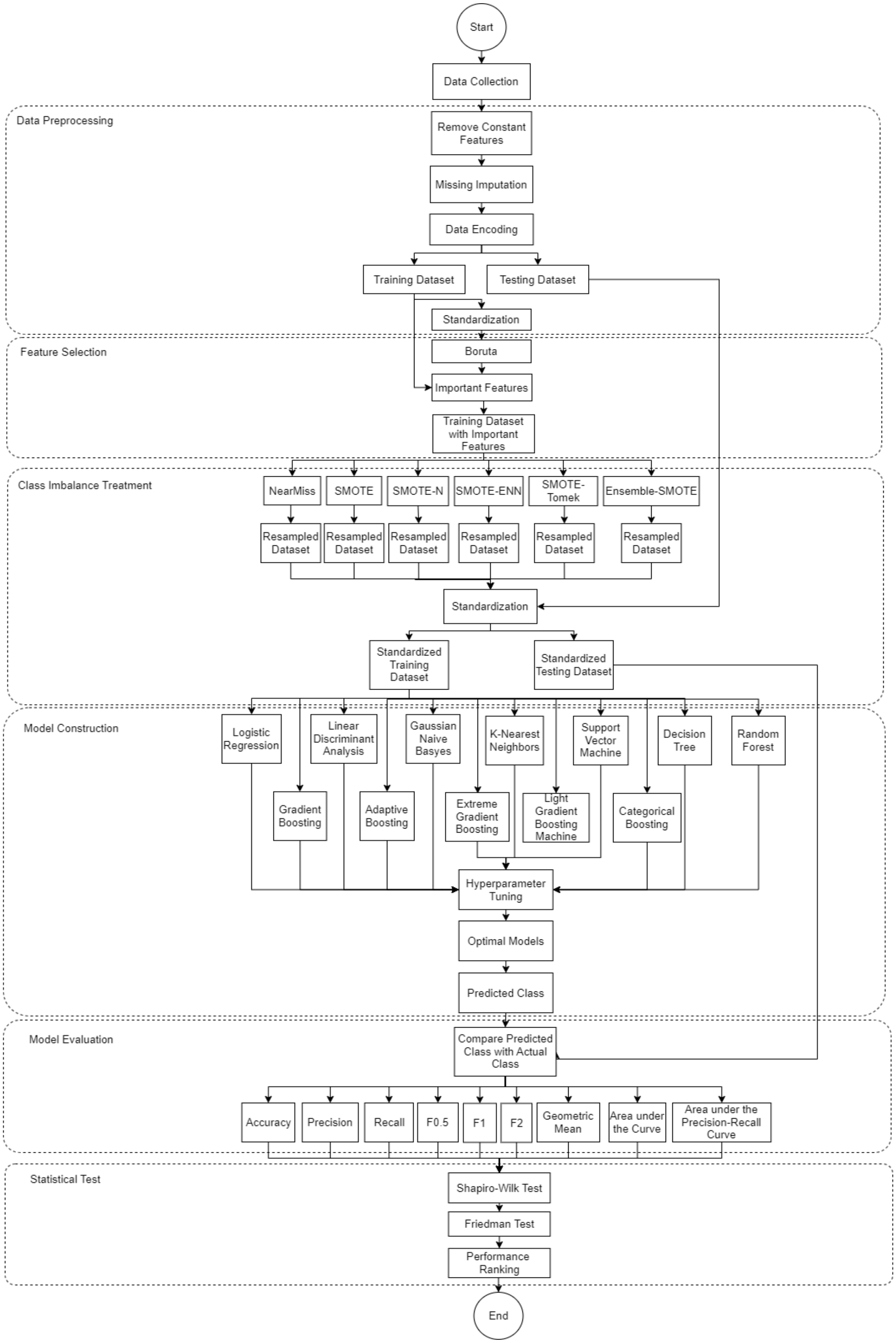


Figure. 1 Flowchart of Methods

A. Data Source

Table 4. Features in Dataset Collected

Data	Feature
Student Profile	• Home state • Home district • Race • Gender • Disability • Marital status • Date of birth
Registration Details	• Campus • Program description • Faculty domain • Faculty description • Admit term begin date • Admit term • Expected graduate term • Expected end of study date
Grades of Sijil Pelajaran Malaysia	• Malay Language • English Language • Mathematics • History • Additional Mathematics • Physics • Chemistry • Biology • Moral Education • Chinese Language • Principles of Accounting • Science
Grades of English Test	• Malaysian University English Test (MUET) • Test of English as a Foreign Language Exam (TOEFL) International English Language Testing System (IELTS)
GPA of Trimester	GPA of each trimester from 1st trimester until 11th trimester

Before students further their studies to universities, Sijil Pelajaran Malaysia (SPM) or Malaysian Certificate of Education is one of the commonly taken public examination in Malaysia [37]. This examination contains certain mandatory subjects such as Malay Language, English Language, Moral Education, History, Mathematics, and Science. This study draws its data from graduates of a university from year 2021 and 2022, comprising a comprehensive dataset that encompasses student profiles, registration details, and the grades in the SPM and English Test as well as the GPA of each trimester from 1st trimester until 11th trimester, as outlined in Table 4. The dataset encapsulates information from 4007 Malaysian students, covering 41 variables, excluding the crucial GOT indicator. In this work, GOT indicator is used as the target variable namely *Yes*, and *No*. Within this dataset, 73.55% of students successfully graduated on time, while 26.45% of students did not meet the stipulated time frame for completion.

B. Data Preprocessing

Table 5. Distribution of Missing Values in Training Data

Feature	Number of Missing Records	Percentage of Missing Records (%)
TOFEL	3577	99.20
IELTS	3529	97.86

Science	2754	76.37
Principles of Accounting	2742	76.04
GPA of 11th trimester	2569	71.24
Chinese Language	2496	69.22
Biology	2430	67.39
GPA of 10th trimester	2363	65.53
Moral Education	2304	63.89
Chemistry	2132	59.12
Physics	2125	58.93
GPA of 9th trimester	1950	54.08
Additional Mathematics	1734	48.09
GPA of 8th trimester	1347	37.35
History	1268	35.16
Mathematics	1264	35.05
GPA of 7th trimester	1184	32.83
GPA of 6th trimester	1084	30.06
GPA of 5th trimester	811	22.49
GPA of 3rd trimester	762	21.13
GPA of 1st trimester	661	18.33
GPA of 2nd trimester	629	17.44
GPA of 4th trimester	625	17.33

Before applying feature selection methods, the dataset undergoes a rigorous preprocessing phase to ensure its cleanliness and quality. This phase involves removing constant features and addressing missing values to provide a solid foundation for subsequent analyses. Following the removal of constant features and extraction of the year of birth, the dataset is divided into 90% training data and 10% testing data to facilitate model training and evaluation. Table 5 provides insight into the extent of missing values present in each feature within the training data. Features with more than 75% missing values, such as TOEFL, IELTS, Science, and Principles of Accounting, are deemed unsuitable for analysis and are consequently excluded from the dataset. Conversely, features with less than 76% missing values undergo an imputation process to salvage valuable data. For categorical features, missing values are replaced with *no_data*, while non-categorical features undergo imputation using two methods: median and KNN. The choice of imputation method for non-categorical features is based on achieving an imputed distribution closest to the original distribution's average difference between the mean and standard deviation. This meticulous approach ensures that missing values are handled effectively, preserving the integrity of the dataset and enabling robust analysis and modelling.

Table 6. Value Encoded for Grades of MUET and SPM

Feature	Original Value	Value Encoded
MUET	Band 6	6
	Band 5	5
	Band 4	4
	Band 3	3
	Band 2	2
	Band 1	1
	Irrelevant	0
SPM	A	4
	B	3
	C	2
	D	1
	Irrelevant	0

Following missing value imputation, the dataset undergoes encoding to prepare it for feature selection methods. Within the dataset, binary values *Yes* and *No* are encoded as 1 and 0, respectively. Additionally, for SPM and

English Test grades, values are encoded based on their ordinal values as delineated in Table 6. Subsequently, the encoded data undergoes further scaling using the Standard Scaler. This process is vital as it aims to eliminate the mean, scale to unit variance, and standardize the features. Such scaling is imperative to mitigate potential issues with feature selection and predictive models, particularly when certain features exhibit higher variance compared to others. By standardizing the scale of features, a uniform representation is maintained, ensuring that variables with disproportionately high variances do not exert undue influence. This optimization enhances the efficacy of both feature selection and predictive modeling techniques, thus contributing to improved model performance and interpretability.

C. Feature Selection

To streamline data dimensionality and enhance model efficiency, Boruta is employed in conjunction with Random Forest, leveraging balanced class weight and a maximum depth of 5. Boruta aids in identifying significant features, which are subsequently utilized as input variables for predictive modeling. This rigorous approach ensures the selection of influential in-university variables, thereby facilitating accurate prediction of students likely to achieve timely graduation. By focusing on key features identified by Boruta, predictive models are equipped to provide insightful assessments, thereby optimizing interventions and support strategies for at-risk students.

D. Class Imbalance Treatment

```
Algorithm 1 Ensemble-SMOTE
Input:
    • Real data, $R = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$ where $y_i \in [0, 1, \dots, k]$ and $k$ = number of classes
    • Sampling ratio, $r$
Output: Data generated by Ensemble-SMOTE, $F$
Algorithm:
1.  $y_{min}$ ← minority class
2.  $B$ ← {SMOTE, SMOTE-N, SMOTE-ENN, SMOTE-Tomek} with $r$
3.  $portion \leftarrow \left\lfloor \frac{r \cdot n_{max} - n_{min}}{n_B} \right\rfloor$
4.  $F$ ← copy of $R$
5.  for each $B_i \in B$ do
6.      $R_{resampled}$ ← dataset generated by $B_i$ with $R$
7.      $f \leftarrow R \cup R_{resampled}$
8.      $f$ ← drop all duplicated records of $f$
9.      $f$ ← first $portion$ of $y_{min}$ in $f$
10.     $F \leftarrow F \cup f$
11. end for
12. $F$ ← drop duplicates of $F$ except for the first occurrence
13. return $F$
```

The training dataset is composed of 73.66% of students who graduated on time and 26.34% who did not. However, such imbalanced class distributions raise concerns about potential bias towards the majority class in predictive models, risking misclassification of minority classes. In response, this study employs various class imbalance treatment techniques, including NearMiss, SMOTE, SMOTE and Nominal (SMOTE-N), SMOTE-ENN, and SMOTE-Tomek. Consequently, the training data undergoes resampling for each technique, starting from an initial sampling ratio of 50% up to 90%, with subsequent scaling using Standard Scaler. Each model is then evaluated using the resampled training data, incorporating the significant features identified by the feature selection method.

Additionally, an ensemble algorithm, Ensemble-SMOTE is implemented to explore the effectiveness of combining existing SMOTE variants (SMOTE, SMOTE-N, SMOTE-ENN, and SMOTE-Tomek) depicted in Algorithm 1. This algorithm aggregates datasets generated by these techniques, with sampling ratio, r and training data, R comprising input variables, x_i and target variables, y_i. Let B represent the set of different variants of SMOTE for the ensemble algorithm to resample the dataset with sampling ratio, r. To ensure equal resampling across each B_i with r, $portion$ is determined by dividing the number of aggregated techniques, n_B with the remaining number of minority class instances, n_{min} (Algorithm 1, line 3). The floor of this division yields the largest integer less than or equal to the

$portion$, ensuring each B_i contributes equally while reducing noise from other techniques. For each B_i, a resampled dataset, $R_{resampled}$ is generated using R and r (Algorithm 1, line 6). Instances other than R (f) are extracted by eliminating duplicate instances from both R and $R_{resampled}$ (Algorithm 1, line 8). After appending f to a copy of R, the ensemble algorithm returns the dataset, dropping duplicate instances except for the first occurrence.

In this work, to simplify and enhance the performance of the Ensemble-SMOTE, a 50% sampling ratio, r is employed. This aims to mitigate excessive noise generation from each B_i, which could otherwise increase bias and misclassification in predictive models. The dataset generated by each class imbalance treatment method is used as the input for the predictive models with the important features identified after standardization using Standard Scaler.

E. Model Construction

Table 7. Hyperparameter Values Search

Model	Hyperparameter	Hyperparameter Description	Search Value
Logistic Regression	C	Controls the inverse of the regularization strength, preventing overfitting.	0.0001, 0.01, 1, 10, 100
Linear Discriminant Analysis	solver	Specifies the algorithm to use in the optimization problem.	svd, lsqr, eigen
Gaussian Naïve Bayes	var_smoothing	Adds a specified value to the diagonal of the covariance matrix of attributes.	1.0, 0.1, 0.01, 0.001, 0.0001, 1 x 10^{-9}
K-Nearest Neighbors	n_estimators	Specifies the number of neighbors.	3, 5, 7, 9, 11
Support Vector Machine	C	Controls the tradeoff between smooth decision boundaries and classifying training points correctly.	0.001, 0.1, 1
Decision Tree	max_depth	Limits the number of nodes in the tree, preventing overfitting.	None, 3, 5, 7
Random Forest	max_depth	Specifies the maximum depth of the individual trees.	None, 3, 5, 7
	n_estimators	Specifies the number of trees.	100, 200
	min_samples_leaf	Controls the minimum size of samples required to split a node further.	1, 2, 3, 4
Gradient Boosting	max_depth	Specifies the maximum depth of the individual estimators.	3, 5, 7
Adaptive Boosting	learning_rate	Shrinks the contribution of each base learner.	0.001, 0.1, 1
	n_estimators	Specifies the number of base learners.	50, 100, 200
Extreme Gradient Boosting	max_depth	Specifies the maximum depth of a tree.	None, 3, 5, 7, 9, 11
	min_child_weight	Specifies the minimum sum of instance weight needed in a child.	None, 1, 2, 3
	colsample_bytree	Specifies the subsample ratio of columns when constructing each tree.	None, 0.5, 0.6, 0.7, 0.8
Light Gradient Boosting Machine	max_depth	Specifies the maximum depth of the tree.	-1, 3, 5, 7
	n_estimators	Specifies the number of estimators.	100, 200
	min_child_weight	Specifies the minimum sum of instance weight needed in a child.	0.001, 0.1, 1
Categorical Boosting	depth	Specifies the maximum tree depth.	3, 5, 6, 7

	min_data_in_leaf	Specifies the minimum number of data points in a leaf.	1, 5, 10

To detect whether a student graduating on time, predictive models are constructed such as Logistic Regression (LR), Linear Discriminant Analysis (LDA), Gaussian Naïve Bayes (GNB), K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), Gradient Boosting (GB), Adaptive Boosting (AdaBoost), Extreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LGBM), and Categorical Boosting (CatBoost). To further improve the model performance, hyperparameter tuning is performed by using GridSearchCV with the hyperparameter values to be searched as indicated in Table 7.

F. Model Evaluation

Table 8. Performance Metric

Metric	Formula	Description
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$	Evaluates the number of correct predictions from a model.
Precision	$\frac{TP}{TP+FP}$	Evaluates the percentage of positive predicted cases that are positive.
Recall	$\frac{TP}{TP+FN}$	Measures the total number of the positive cases that are captured by the positive predictions.
F0.5-score	$1.25 \cdot \frac{Precision \cdot Recall}{0.25 \cdot Precision + Recall}$	Represents the weighted mean of precision and recall, assigning more weights to precision than recall.
F1-score	$2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$	Represents the harmonic mean of precision and recall.
F2-score	$5 \cdot \frac{Precision \cdot Recall}{4 \cdot Precision + Recall}$	Represents the weighted mean of precision and recall, assigning more weights to recall than precision.
Geometric Mean	$\sqrt{Precision \cdot \frac{TN}{TN+FP}}$	Evaluates a model's ability to accurately detect instances from both classes.

After constructing the predictive models, the performance is evaluated based on several metrics outlined in Table 7. These metrics include accuracy, precision, recall, F0.5-score, F1-score, F2-score, Geometric Mean (G-Mean). Additionally, the Area under the Curve (AUC) is computed to assess the trade-off between correctly predicted positive classes and incorrectly predicted negative classes. Area under the Precision-Recall Curve (PR-AUC) is computed to assess the model ability in balancing between precision and recall in the class imbalance context. In the context of evaluation metrics, four crucial measurements are utilized: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). TP represents the total number of positive classes correctly identified, while TN indicates the accurate identification of negative classes. Conversely, if the model incorrectly predicts the negative class as positive, it results in FP. FN denotes the total instances where positive classes are mistakenly predicted as negative. These metrics offer a comprehensive understanding of the models' performance, providing insights into their ability to correctly identify both positive and negative classes, thus informing the overall efficacy of the predictive models.

G. Statistical Test

The effectiveness of class imbalance treatment methods in mitigating class imbalance in GOT is assessed through tests of statistical significance. To begin, the Shapiro-Wilk test is conducted to ascertain the normal distribution of model performance with class imbalance treatment methods before proceeding with further statistical tests.

H_0 : The model performances with class imbalance treatment methods are normally distributed.

H_a : The model performances with class imbalance treatment methods are not normally distributed.

The hypothesis of Shapiro-Wilk test is formed as above, including null hypothesis, H_0 and alternative hypothesis, H_a. If the p-value of the Shapiro-Wilk test is less than the significant level ($\alpha = 0.05$), then H_0 is rejected, suggesting that the model performances with class imbalance treatment methods are not normally distributed. Subsequently, to compare and assess effectiveness, the Friedman test is employed if the model performances are non-normally distributed; otherwise, Analysis of Variance (ANOVA) is utilized.

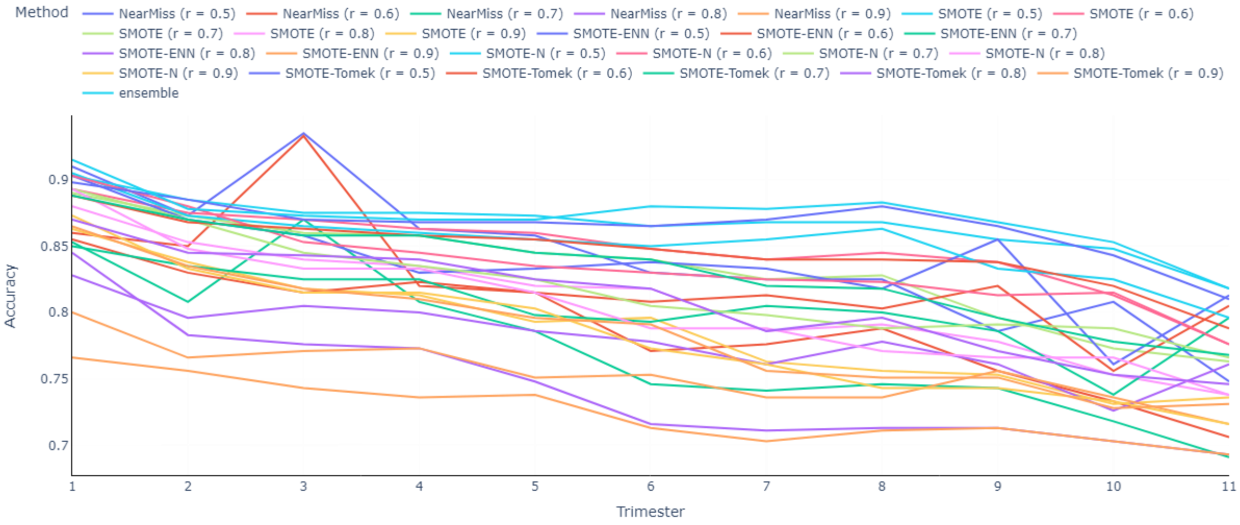
H_0 : There is no significant difference between the model performances with class imbalance treatment methods.

H_a : There is significant difference between the model performances with class imbalance treatment methods.

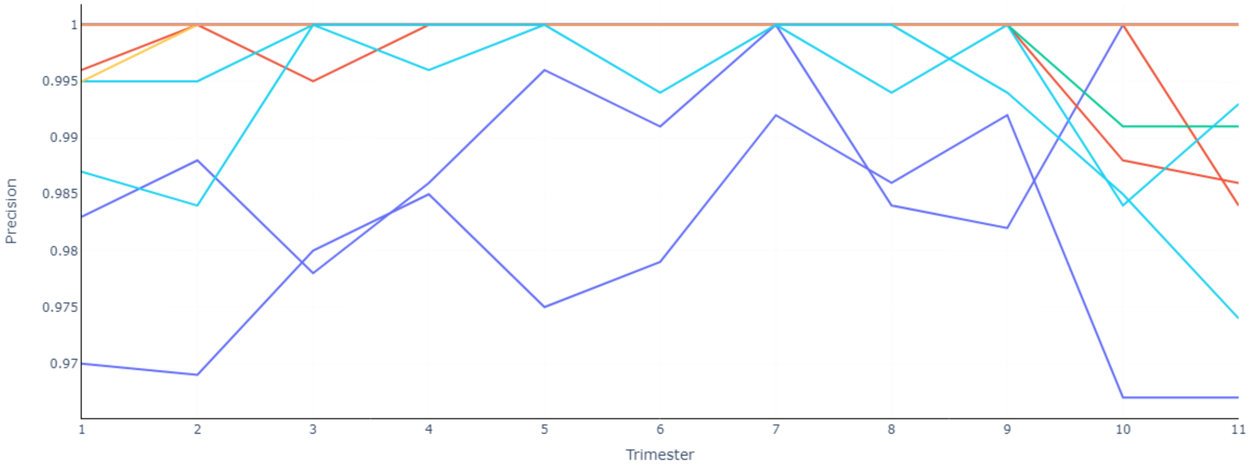
In the comparison of model performances, the hypotheses (H_0 and H_a) are formulated as above for either ANOVA or the Friedman test. If the p-value of the test is less than α, H_0 is rejected, indicating that the class imbalance treatment methods employed yield significant differences in model performances. Leveraging performance metrics, the relative efficacy of class imbalance treatment methods in addressing imbalanced data issues can be evaluated and assigned ranks. Consequently, the most effective class imbalance treatment method is determined by the highest average rank across the performance metrics.

VI. FINDINGS

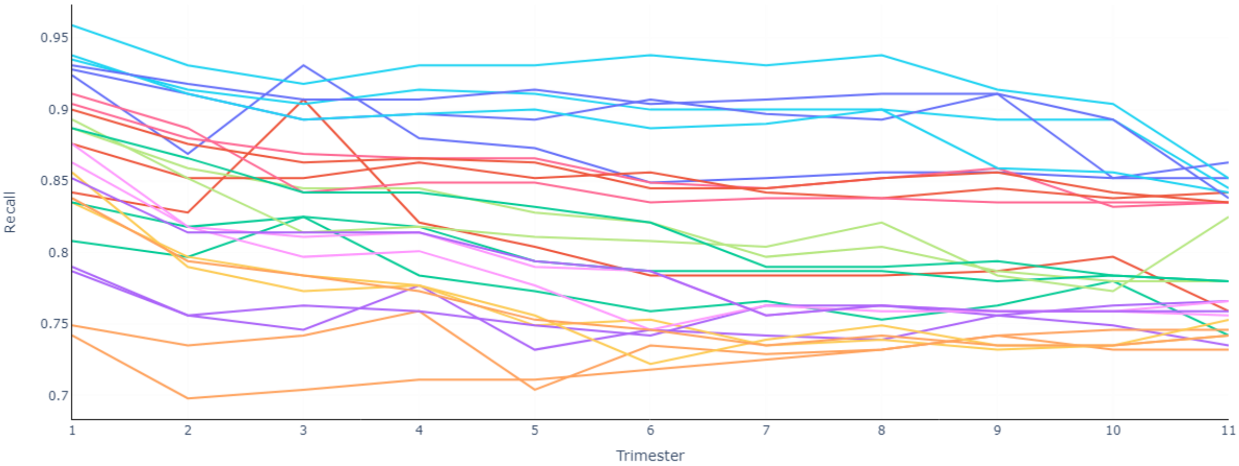
A. Model Performance with Class Imbalance Treatment Methods



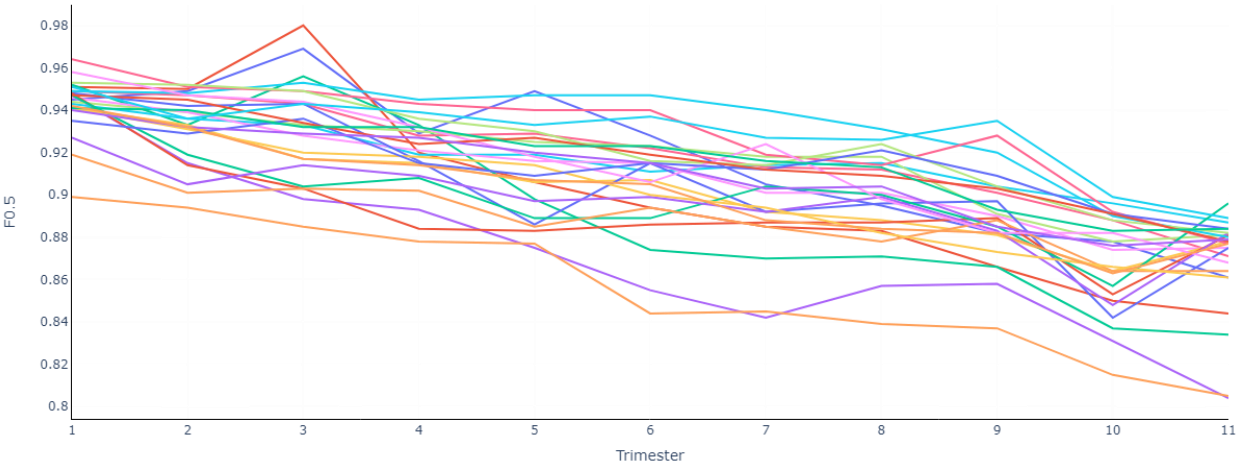
(a)


Precision
1
0.995
0.99
0.985
0.98
0.975
0.97
1
2
3
4
5
6
7
8
9
10
11
Trimester

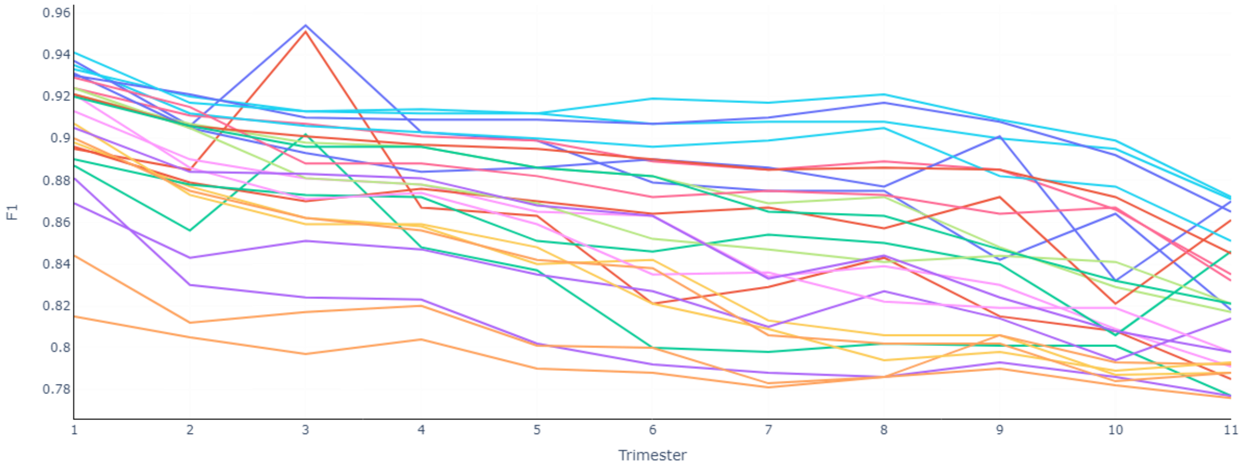
(b)


Recall
0.95
0.9
0.85
0.8
0.75
0.7
1
2
3
4
5
6
7
8
9
10
11
Trimester

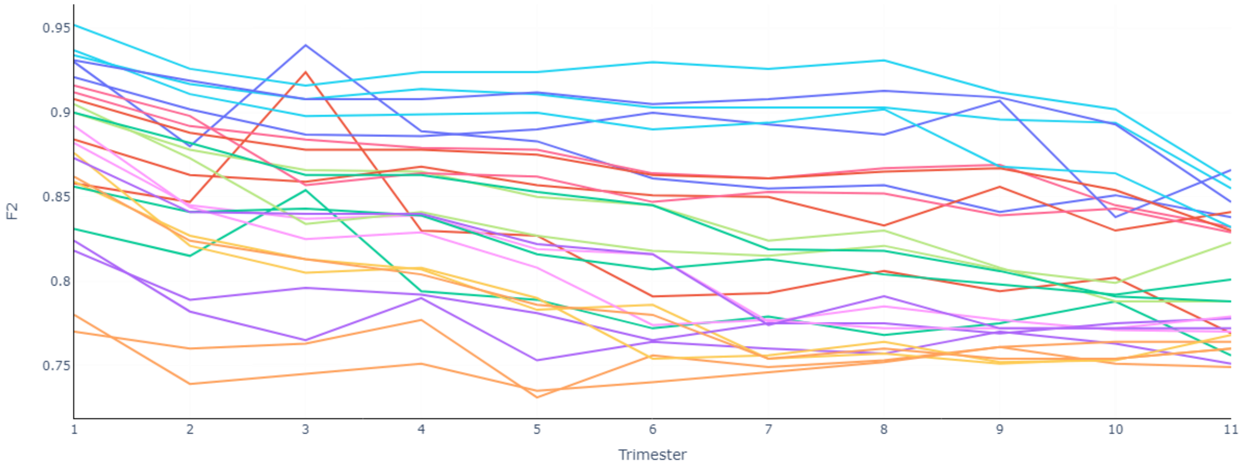
(c)


F0.5
0.98
0.96
0.94
0.92
0.9
0.88
0.86
0.84
0.82
0.8
1
2
3
4
5
6
7
8
9
10
11
Trimester

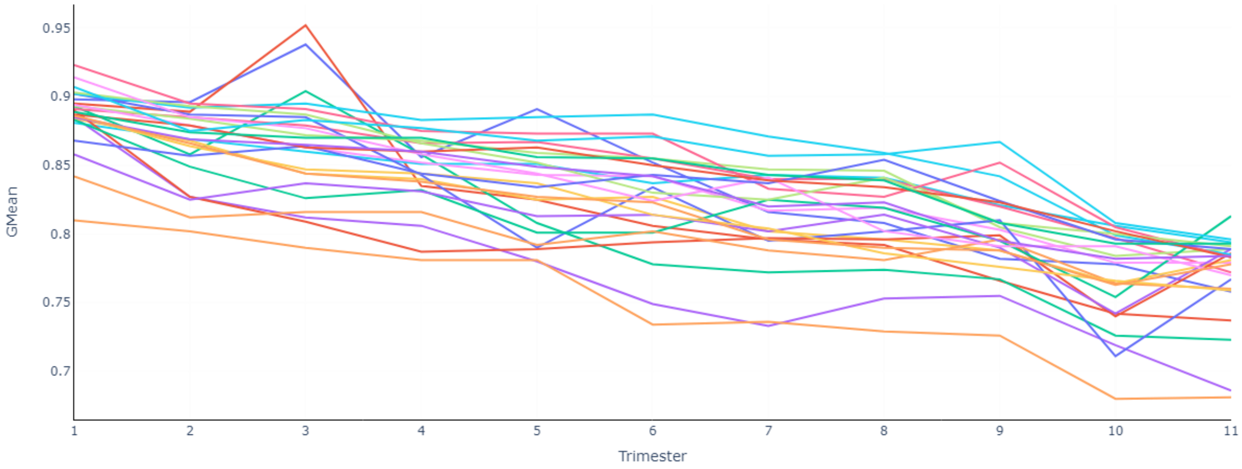
(d)


F1
0.96
0.94
0.92
0.9
0.88
0.86
0.84
0.82
0.8
0.78
1
2
3
4
5
6
7
8
9
10
11
Trimester

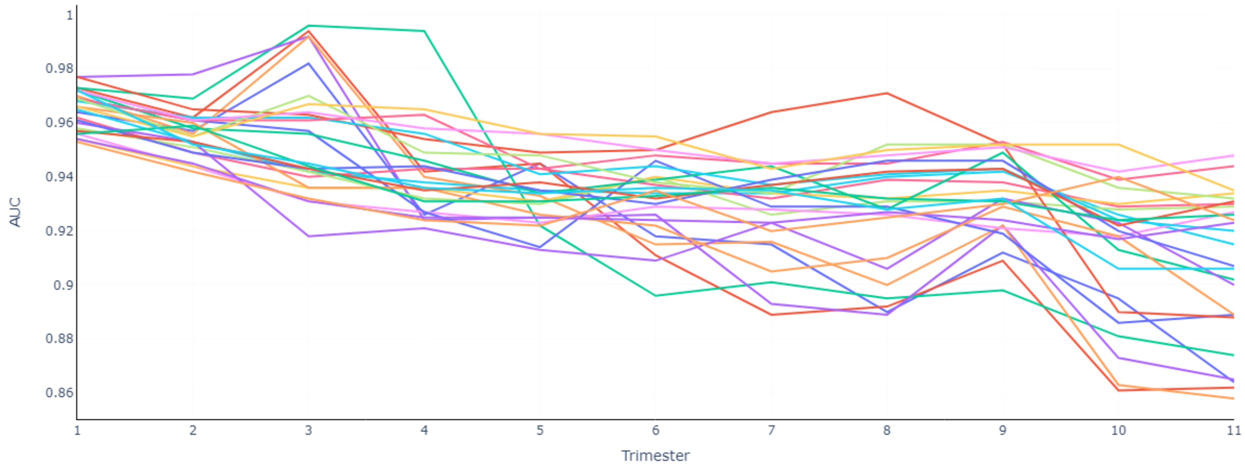
(e)


F2
0.95
0.9
0.85
0.8
0.75
1
2
3
4
5
6
7
8
9
10
11
Trimester

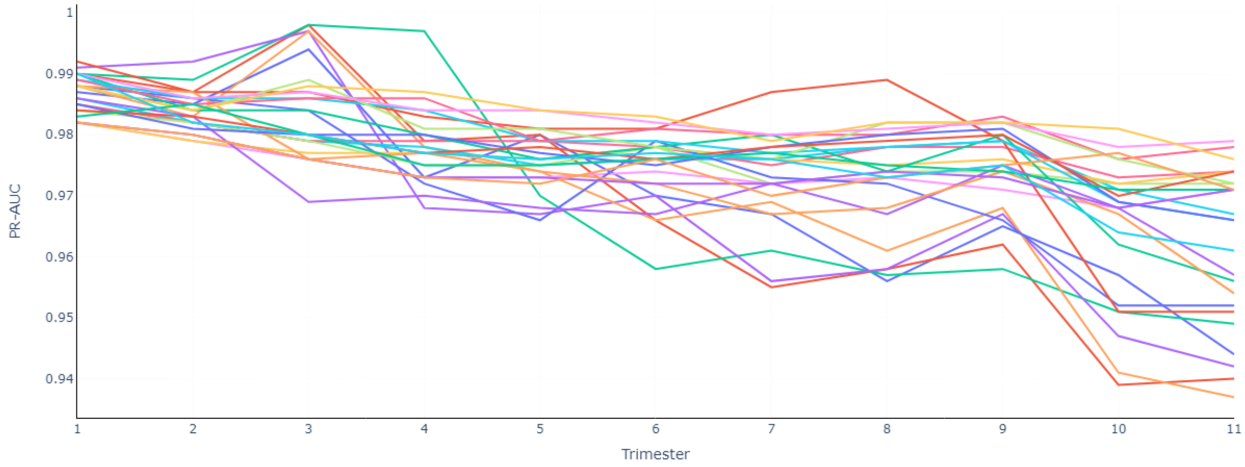
(f)


GMean
0.95
0.9
0.85
0.8
0.75
0.7
1
2
3
4
5
6
7
8
9
10
11
Trimester

(g)



(h)



(i)

Figure. 2 Highest Model Performance for each Class Imbalance Treatment Methods based on (a) accuracy, (b) precision, (c) recall, (d) F0.5-score, (e) F1-score, (f) F2-score, (g) G-Mean, (h) AUC, and (i) PR-AUC.

Figure 2 indicates the highest model performance for each class imbalance treatment methods based on the performance metrics after class imbalance treatment, hyperparameter tuning and feature selection. Notably, NearMiss exhibited exceptional precision rates of 100% across all trimesters when the sampling ratio ranged from 50% to 90%. Precision refers to the ratio of true positive predictions to the total number of positive predictions made by the predictive models. Achieving 100% precision indicates that all instances predicted as positive were indeed true positives. However, despite achieving prefect precision, the predictive models utilizing NearMiss suffered in terms of recall when compared to other class imbalance treatment methods. Recall, also known as sensitivity, measures the ratio of true positive predictions to the total number of actual positive instances in the dataset. A lower recall suggests that the model fails to correctly identify a significant portion of positive instances, leading to missed opportunities of correct predictions. Consequently, the lower recall resulted in diminished overall performance across various metrics such as accuracy, recall, F0.5-score, F1-score, F2-score, G-Mean, AUC, and PR-AUC. These metrics collectively evaluate different aspects of model performance, including its ability to balance between precision and recall, its robustness to class imbalances, and its overall predictive power. Thus, despite NearMiss's high precision rates, its lower recall adversely affected the overall performance of the predictive models in terms of these evaluation metrics.

Based on G-Mean, LGBM with SMOTE-N achieved more than 85.90% when the sampling ratio of 50% is used. G-Mean is used to assess the performance of binary classification models, particularly in imbalanced datasets. It considers both the recall and specificity of the model, providing a balanced evaluation of its performance. In this context, achieving a G-Mean score exceeding 85.90% suggests that the combination of LGBM with SMOTE-N, when using a sampling ratio of 50%, effectively balances between recall and specificity, resulting in strong overall performance in classification tasks. This indicates that the model is adept at correctly identifying both positive and negative instances, even in the presence of class imbalances.

Other than that, CatBoost with NearMiss, SMOTE-N, and SMOTE-ENN achieved the highest AUC (94.80% - 99.60%) and PR-AUC (97.90% - 99.80%) across all trimesters. AUC is used to evaluate the performance of binary classification models, representing the area under the Receiver Operating Characteristic (ROC) curve. A higher AUC value indicates better discrimination between positive and negative classes. Similarly, PR-AUC measures the area under the Precision-Recall curve, providing insight into the model's ability to identify positive instances while minimizing false positives. In this context, CatBoost, in conjunction with NearMiss, SMOTE-N, and SMOTE-ENN, effectively balances between precision and recall, achieving high accuracy in identifying positive instances while maintaining low false positive rates.

Despite the formidable challenges posed by class imbalance within the GOT dataset, LR coupled with Ensemble-SMOTE emerges as a standout performer, showcasing unparalleled effectiveness in addressing these issues. Across a pivotal period spanning the 6th to the 10th trimesters, LR with Ensemble-SMOTE consistently excels, boasting the highest levels of accuracy, recall, F1-score, and F2-score among all class imbalance treatment methodologies. This remarkable achievement signifies LR with Ensemble-SMOTE as a formidable solution for mitigating the inherent biases present in the GOT dataset. Its ability to maintain accuracy levels ranging from 85.30% to 88.30% underscores its robustness in correctly identifying both positive and negative instances, even amidst significant class imbalances. Moreover, with recall rates reaching impressive heights of 90.40% to 93.80%, LR with Ensemble-SMOTE demonstrates a keen aptitude for capturing a substantial portion of the actual positive instances, minimizing the risk of overlooking critical data points.

The F1-score and F2-score, metrics that encapsulate the harmonic mean of precision and recall, further emphasize the reliability and effectiveness of LR with Ensemble-SMOTE. Scoring between 89.90% and 92.10% for the F1-score, and between 90.20% and 93.10% for the F2-score, LR with Ensemble-SMOTE showcases its ability to strike a balance between precision and recall, crucial for achieving high-performance classification in imbalanced datasets. In essence, LR with Ensemble-SMOTE stands as a beacon of excellence in the realm of class imbalance treatment methods within the GOT dataset, offering unparalleled accuracy, recall, and balance between precision and recall. Its consistent and superior performance from the 6th to the 10th trimesters underscores its pivotal role in enhancing the reliability and effectiveness of predictive modelling efforts within the domain of GOT analysis.

B. Statistical Analysis of Model Performance with Class Imbalance Treatment Methods

A statistical examination was conducted to assess potential variations in performance among predictive models utilizing different class imbalance treatment methods, focusing on metrics such as accuracy, precision, recall, F0.5-score, F1-score, F2-score, AUC, and PR-AUC. The Shapiro-Wilk test was employed to ascertain the normal distribution of performance metrics, guiding the selection of the appropriate statistical test. Should the performance metrics exhibit normal distribution, ANOVA would be applied; however, if non-normality is observed, the Friedman test would be utilized.

Table 9. Shapiro-Wilk test

	Accuracy	Precision	Recall	F0.5-score	F1-score	F2-score	G-Mean	AUC	PR-AUC
Statistics	0.9465	0.808	0.9187	0.8487	0.8799	0.9054	0.959	0.9239	0.9304
p-value	1.31 x 10^{-33}	0	1.40 x 10^{-39}	0	1.40 x 10^{-45}	7.43 x 10^{-42}	4.27 x 10^{-30}	1.32 x 10^{-38}	2.67 x 10^{-37}
Reject H_0	Yes								

Table 9 reveals that when the p-value of the Shapiro-Wilk test falls below 0.05, the null hypothesis is rejected. Consequently, ANOVA is deemed inappropriate, and the Friedman test is chosen as the alternative method. Table 10 presents the outcomes of the Friedman test, wherein rejection of the null hypothesis occurs when the p-value is less than 0.05. This outcome indicates a significant disparity in model performances among different class imbalance treatment methods, implying that these methods may have distinct impacts on the overall performance of predictive models.

Table 10. Friedman test

	Accuracy	Precision	Recall	F0.5-score	F1-score	F2-score	G-Mean	AUC	PR-AUC
Statistics	2627.125	3032.907	2900.078	2199.667	2731.776	2863.038	2111.794	3004.658	2992.446
p-value	0								
Reject H_0	Yes								

Table 11. Average Performance Rank

Class Imbalance Method	Accuracy	Precision	Recall	F0.5-score	F1-score	F2-score	G-Mean	AUC	PR-AUC	Mean Rank
NearMiss (r = 0.5)	7.46	22.29	6.49	9.91	7.05	6.59	10.97	15.10	15.60	11.27
NearMiss (r = 0.6)	13.02	16.80	12.82	12.96	12.93	12.81	13.46	15.53	16.22	14.06
NearMiss (r = 0.7)	16.42	10.98	17.03	15.22	16.73	17.01	14.92	14.78	15.37	15.38
NearMiss (r = 0.8)	20.01	9.04	20.45	18.14	20.31	20.44	17.16	14.18	15.02	17.19
NearMiss (r = 0.9)	23.01	7.84	23.59	20.82	23.39	23.55	19.49	15.44	15.86	19.22
SMOTE (r = 0.5)	6.71	16.25	6.67	8.81	6.49	6.56	10.44	9.39	9.12	8.94
SMOTE (r = 0.6)	9.36	13.33	10.22	9.31	9.64	10.07	10.18	9.93	10.20	10.25
SMOTE (r = 0.7)	15.86	12.07	16.36	15.02	16.07	16.33	14.81	15.84	16.22	15.40
SMOTE (r = 0.8)	18.37	8.65	18.84	16.51	18.61	18.84	15.17	15.17	15.26	16.16
SMOTE (r = 0.9)	20.67	7.74	21.08	18.57	20.93	21.03	16.72	16.08	15.98	17.64
SMOTE-N (r = 0.5)	4.56	20.59	4.02	7.32	4.19	4.01	8.60	8.95	8.73	7.89
SMOTE-N (r = 0.6)	6.47	16.26	6.91	7.02	6.42	6.87	**8.09**	**7.99**	**7.74**	8.20
SMOTE-N (r = 0.7)	9.42	14.28	10.12	8.48	9.60	10.09	8.61	10.16	10.21	10.11
SMOTE-N (r = 0.8)	11.79	12.17	12.22	10.17	12.04	12.19	9.39	10.74	10.44	11.24
SMOTE-N (r = 0.9)	14.58	9.33	14.94	12.53	14.85	14.90	11.09	11.92	10.79	12.77
SMOTE-ENN (r = 0.5)	7.23	24.59	3.64	12.30	5.94	4.06	14.06	15.44	15.36	11.40
SMOTE-ENN (r = 0.6)	12.26	18.15	10.65	14.89	11.48	10.84	15.97	13.78	13.71	13.53
SMOTE-	15.02	14.23	14.86	15.23	15.06	14.97	15.53	14.72	14.77	14.93

ENN (r = 0.7)										
SMOTE-ENN (r = 0.8)	19.69	9.05	19.92	18.30	19.84	19.89	17.47	18.31	18.22	17.85
SMOTE-ENN (r = 0.9)	22.94	**6.91**	23.39	21.11	23.19	23.38	19.45	18.07	18.34	19.64
SMOTE-Tomek (r = 0.5)	7.18	16.35	6.93	9.43	6.98	6.89	11.05	9.81	9.24	9.32
SMOTE-Tomek (r = 0.6)	10.07	13.61	10.55	10.18	10.14	10.44	11.29	10.56	10.24	10.79
SMOTE-Tomek (r = 0.7)	16.26	10.96	16.56	15.53	16.42	16.57	15.15	16.23	16.16	15.54
SMOTE-Tomek (r = 0.8)	18.94	9.01	19.33	17.37	19.16	19.28	15.95	17.15	17.06	17.03
SMOTE-Tomek (r = 0.9)	21.23	7.73	21.69	18.90	21.54	21.65	17.22	17.52	17.06	18.28
Ensemble-SMOTE	**2.46**	22.78	**1.71**	**6.95**	**1.99**	**1.72**	8.75	8.22	8.09	**6.96**

Table 11 indicates the average performance ranking of each class imbalance treatment method based on accuracy, precision, recall, F0.5-score, F1-score, F2-score, AUC, and PR-AUC. A lower value in the average rank signifies superior performance across predictive models throughout the trimesters. Based on the findings, the superior performance of Ensemble-SMOTE over other class imbalance treatment methods can be explained by analyzing its average rank across various performance metrics. Ensemble-SMOTE exhibits a notably low average rank of 6.96, indicating its consistent effectiveness in enhancing model performance across different evaluation criteria.

Specifically, Ensemble-SMOTE achieves competitive rankings in critical metrics such as accuracy, recall, F0.5-score, F1-score, and F2-score, which are pivotal for assessing the overall effectiveness of predictive models. Its average rank of 2.46 for accuracy, 1.71 for recall, 6.95 for F0.5-score, 1.99 for F1-score, and 1.72 for F2-score demonstrates its proficiency in correctly identifying positive instances while maintaining a balance between precision and recall, ultimately leading to high-quality predictions. Although precision ranks slightly higher at 22.78, its overall performance across multiple metrics remains impressive, contributing to Ensemble-SMOTE's overall effectiveness. Additionally, the ensemble nature of Ensemble-SMOTE likely contributes to its superior performance. By combining multiple variants of SMOTE, it leverages the strengths of each technique while mitigating their individual weaknesses. This ensemble approach allows Ensemble-SMOTE to effectively address class imbalances and improve model generalization across different trimesters, resulting in consistently superior performance compared to other class imbalance treatment methods.

In summary, the statistics provided highlight Ensemble-SMOTE's remarkable ability to enhance model performance across various evaluation metrics, positioning it as a superior choice for mitigating class imbalance issues in predictive modelling tasks within the context of the given dataset.

V. CONCLUSION

In conclusion, this work has aimed to (i) compare various class imbalance treatment methods in mitigating the problem of class imbalance with different sampling ratios, (ii) propose an ensemble class imbalance treatment method in mitigating the problem of class imbalance, and (iii) develop and evaluate predictive models in identifying the likelihood of students graduating on time during their studies in university. After feature selection, NearMiss, SMOTE, SMOTE-N, SMOTE-ENN, and SMOTE-Tomek were compared with different sampling ratios ranging from 50% to 90%. Moreover, an ensemble algorithm, Ensemble-SMOTE is developed to aggregate the dataset generated by the SMOTE variants such as SMOTE, SMOTE-N, SMOTE-ENN, SMOTE-Tomek in mitigating the problem of class imbalance effectively. The dataset generated by class imbalance treatment methods were used as the input of the predictive models, namely LR, LDA, GNB, KNN, SVM, DT, RF, GB, AdaBoost, XGBoost, LGBM, and CatBoost in detecting GOT. The predictive models were evaluated based on accuracy, precision, recall, F0.5-score, F1-score, F2-score, AUC, and PR-AUC. Based on the findings, LR with Ensemble-SMOTE outperformed other predictive models and class imbalance treatment methods by achieving the highest accuracy (85.30% - 88.30%), recall (90.40% - 93.80%), and F1-score (90.20% - 93.10%) from 6th until 10th trimester. By performing statistical test analysis, Ensemble-SMOTE is ranked as the top-performers by achieving the lowest value in the average rank based on the performance metrics. This suggests that Ensemble-SMOTE could generate useful synthetic samples with less noise, improving the model performance in detecting GOT. In the future, additional research could incorporate and examine more complicated approaches in mitigating class imbalance when the dataset is highly imbalanced.

ACKNOWLEDGEMENT

The researchers sincerely appreciate and express gratitude for financial support from TM Research & Development Grant (TM R&D), MMUE/220028.

AUTHOR CONTRIBUTIONS

Theng-Jia Law - Conceptualization, Methodology, Evaluation, Writing - Original Draft Preparation and Editing;
Choo-Yee Ting - Supervision, Writing – Review;
Hu Ng - Project Leader, Supervision, Writing – Review;
Hui-Ngo Goh - Supervision, Writing – Review;
Quek Albert - Data Providers, Writing – Review

CONFLICT OF INTERESTS

No conflict of interests were disclosed.

REFERENCES

[1] K. Anwar, H. Hanafiah, and A. Ebun, "Predicting Student Graduation Using Artificial Neural Network: A Preliminary study of Diploma In Accountancy Program at UiTM Sabah," 2020.

[2] G. Sidhu, S. Kannan, A. S. Samsul Kamil, and R. Du, "Sustaining Students' Quality Learning Environment by Reviewing Factors to Graduate-on-Time: A case study," *Environment-Behaviour Proceedings Journal*, vol. 8, pp. 127–133, 2023, doi: 10.21834/ebpj.v8i24.4649.

[3] A. Anggrawan, H. Hairani, and C. Satria, "Improving SVM Classification Performance on Unbalanced Student Graduation Time Data Using SMOTE," *International Journal of Information and Education Technology*, vol. 13, no. 2, pp. 289–295, 2023.

[4] R. Garc\'\ia-Ros, F. Pérez-González, F. Cavas-Mart\'\inez, and J. M. Tomás, "Effects of pre-college variables and first-year engineering students' experiences on academic achievement and retention: a structural model," *International Journal of Technology and Design Education*, vol. 29, pp. 915–928, 2019.

[5] N. Mohammad Suhaimi, S. Abdul-Rahman, S. Mutalib, N. H. Abdul Hamid, and A. Md Ab Malik, "Predictive model of graduate-on-time using machine learning algorithms," in *Soft Computing in Data Science: 5th International Conference, SCDS 2019, Iizuka, Japan, August 28–29, 2019, Proceedings 5*, 2019, pp. 130–141.

[6] K. T. Chui, D. C. L. Fung, M. D. Lytras, and T. M. Lam, "Predicting at-risk university students in a virtual learning environment via a machine learning algorithm," *Computers in Human Behavior*, vol. 107, p. 105584, 2020, doi: https://doi.org/10.1016/j.chb.2018.06.032.

[7] F. T. Anggraeny, A. K. Darmawan, A. Anekawati, I. Yudhisari, and others, "Early Prediction for Graduation of Private High School Students with Machine Learning Approach," 2023.

[8] A. Desfiandi and B. Soewito, “Student Graduation Time Prediction using Logistic Regression, Decision Tree, Support Vector Machine, and AdaBoost Ensemble Learning,” *International Journal of Information System and Computer Science*, vol. 7, no. 3, pp. 195–199, 2023.

[9] J. M. Aiken, R. De Bin, M. Hjorth-Jensen, and M. D. Caballero, “Predicting time to graduation at a large enrollment American university,” *Public Library of Science One*, vol. 15, no. 11, p. e0242334, 2020.

[10] T. Wongvorachan, S. He, and O. Bulut, “A Comparison of Undersampling, Oversampling, and SMOTE Methods for Dealing with Imbalanced Classification in Educational Data Mining,” *Information*, vol. 14, no. 1, 2023, doi: 10.3390/info14010054.

[11] D. A. Rachmawati, N. A. Ibadurrahman, J. Zeniarja, and N. Hendriyanto, “Implementation of the Random Forest Algorithm in Classifying the Accuracy of Graduation Time for Computer Engineering Students at Dian Nuswantoro University,” *Jurnal Teknik Informatika (Jutif)*, vol. 4, no. 3, pp. 565–572, 2023, doi: 10.52436/1.jutif.2023.4.3.920.

[12] N. Buniyamin and others, “Mitigating imbalanced classification problems in academic performance with resampling methods/A’zraa Afhzan Ab Rahim and Norlida Buniyamin,” *Journal of Electrical and Electronic Systems Research (JEESR)*, vol. 23, no. 1, pp. 45–56, 2023.

[13] R. Ghorbani and R. Ghousi, “Comparing Different Resampling Methods in Predicting Students’ Performance Using Machine Learning Techniques,” *Institute of Electrical and Electronics Engineers Access*, vol. 8, pp. 67899–67911, 2020, doi: 10.1109/ACCESS.2020.2986809.

[14] Y. T. Samuel, J. J. Hutapea, and B. Jonathan, “Predicting the Timeliness of Student Graduation Using Decision Tree C4.5 Algorithm in Universitas Advent Indonesia,” in *2019 12th International Conference on Information & Communication Technology and System (ICTS)*, 2019, pp. 276–280. doi: 10.1109/ICTS.2019.8850948.

[15] M. Ben Said, Y. Hadj Kacem, A. Algarni, and A. Masmoudi, “Early prediction of Student academic performance based on Machine Learning algorithms: A case study of bachelor’s degree students in KSA,” *Education and Information Technologies (Dordr)*, pp. 1–24, 2023.

[16] R. Al-Shabandar, A. J. Hussain, P. Liatsis, and R. Keight, “Detecting At-Risk Students With Early Interventions Using Machine Learning Techniques,” *Institute of Electrical and Electronics Engineers Access*, vol. 7, pp. 149464–149478, 2019, doi: 10.1109/ACCESS.2019.2943351.

[17] N. Mduma, “Data Balancing Techniques for Predicting Student Dropout Using Machine Learning,” *Data (Basel)*, vol. 8, no. 3, 2023, doi: 10.3390/data8030049.

[18] S. Verma, R. K. Yadav, and K. Kholiya, “A scalable machine learning-based ensemble approach to enhance the prediction accuracy for identifying students at-risk,” *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 8, 2022.

[19] S. Hutt, M. Gardner, A. L. Duckworth, and S. K. D’Mello, “Evaluating fairness and generalizability in models predicting on-time graduation from college applications.,” *International Educational Data Mining Society*, 2019.

[20] E. P. Jiang, “Applying a Hybrid Sampling and Boosting Approach to Predict Student Retention,” *International Journal of Machine Learning and Computing*, vol. 12, no. 5, 2022.

[21] Y. Alshamaila *et al.*, “An automatic prediction of students’ performance to support the university education system: a deep learning approach,” *Multimedia Tools and Applications*, pp. 1–28, 2024.

[22] A. Gonzalez-Nucamendi, J. Noguez, L. Neri, V. Robledo-Rella, and R. M. G. García-Castelán, “Predictive analytics study to determine undergraduate students at risk of dropout,” *Frontiers in Education (Lausanne)*, vol. 8, 2023, doi: 10.3389/feduc.2023.1244686.

[23] S. W. Masood and S. A. Begum, “Comparison of Resampling Techniques for Imbalanced Datasets in Student Dropout Prediction,” in *2022 Institute of Electrical and Electronics Engineers Silchar Subsection Conference*, 2022, pp. 1–7. doi: 10.1109/SILCON55242.2022.10028915.

[24] L. Sha, M. Raković, A. Das, D. Gašević, and G. Chen, “Leveraging Class Balancing Techniques to Alleviate Algorithmic Bias for Predictive Tasks in Education,” *Institute of Electrical and Electronics Engineers Transactions on Learning Technologies*, vol. 15, no. 4, pp. 481–492, 2022, doi: 10.1109/TLT.2022.3196278.

[25] C. H. Cho, Y. W. Yu, and H. G. Kim, “A Study on Dropout Prediction for University Students Using Machine Learning,” *Applied Sciences*, vol. 13, no. 21, 2023, doi: 10.3390/app132112004.

[26] Y. ÜNAL, A. SAĞLAM, and O. KAYHAN, "Improving classification performance for an imbalanced educational dataset example using SMOTE," *Avrupa Bilim ve Teknoloji Dergisi*, pp. 485–489, 2019, doi: 10.31590/ejosat.638608.

[27] A. Nabil, M. Seyam, and A. Abou-Elfetouh, "Prediction of Students' Academic Performance Based on Courses' Grades Using Deep Neural Networks," *Institute of Electrical and Electronics Engineers Access*, vol. 9, pp. 140731–140746, 2021, doi: 10.1109/ACCESS.2021.3119596.

[28] H. Sahlaoui, E. A. A. Alaoui, S. Agoujil, and A. Nayyar, "An empirical assessment of smote variants techniques and interpretation methods in improving the accuracy and the interpretability of student performance models," *Education and Information Technologies (Dordr)*, pp. 1–37, 2023.

[29] A. F. U. , G. B. S. , & G. M. Bako H. S., "Predicting Timely Graduation of Postgraduate Students using Random Forests Ensemble Method," vol. 7, pp. 177–185, 2023, doi: 10.33003/fjs-2023-0703-1773.

[30] S. Kim, E. Choi, Y.-K. Jun, and S. Lee, "Student Dropout Prediction for University with High Precision and Recall," *Applied Sciences*, vol. 13, no. 10, 2023, doi: 10.3390/app13106275.

[31] D. K. Dake, C. Buabeng-Andoh, and others, "Using machine learning techniques to predict learner drop-out rate in higher educational institutions," *Mobile Information Systems*, vol. 2022, 2022.

[32] V. Flores, S. Heras, and V. Julian, "Comparison of Predictive Models with Balanced Classes Using the SMOTE Method for the Forecast of Student Dropout in Higher Education," *Electronics (Basel)*, vol. 11, no. 3, 2022, doi: 10.3390/electronics11030457.

[33] G. Pratape, K. Rao Meesala, S. Panda, and P. Goyal, "Predicting Graduation and Dropout Rates : A Machine Learning Approach," in *2023 International Conference on Advanced Computing & Communication Technologies (ICACCTech)*, 2023, pp. 603–609. doi: 10.1109/ICACCTech61146.2023.00103.

[34] S. Alwarthan, N. Aslam, and I. U. Khan, "An Explainable Model for Identifying At-Risk Student at Higher Education," *Institute of Electrical and Electronics Engineers Access*, vol. 10, pp. 107649–107668, 2022, doi: 10.1109/ACCESS.2022.3211070.

[35] T. A. Khan, R. Sadiq, Z. Shahid, M. M. Alam, and M. B. M. Su'ud, "Sentiment Analysis using Support Vector Machine and Random Forest," *Journal of Informatics and Web Engineering*, vol. 3, no. 1, pp. 67–75, 2024.

[36] M. M. Hussain, S. Akbar, S. A. Hassan, M. W. Aziz, and F. Urooj, "Prediction of Student's Academic Performance through Data Mining Approach," *Journal of Informatics and Web Engineering*, vol. 3, no. 1, pp. 241–251, 2024.

[37] Ministry of Education Malaysia (2024) MOE. [Online]. Available: https://www.moe.gov.my/