ISTANBUL TECHNICAL UNIVERSITY ★ GRADUATE SCHOOL

NEW PROPOSED METHODS FOR SYNTHETIC MINORITY OVER-SAMPLING TECHNIQUE

M.SC. THESIS

Hakan KORUL

Department of Data Engineering and Business Analytics

Big Data and Business Analytics Programme

AUGUST 2024

ISTANBUL TECHNICAL UNIVERSITY ★ GRADUATE SCHOOL

NEW PROPOSED METHODS FOR SYNTHETIC MINORITY OVER-SAMPLING TECHNIQUE

M.SC. THESIS

Hakan KORUL
(528211075)

Department of Data Engineering and Business Analytics

Big Data and Business Analytics Programme

Thesis Advisor: Assist. Prof. Dr. Mehmet Ali ERGÜN

AUGUST 2024

İSTANBUL TEKNİK ÜNİVERSİTESİ ★ LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

SENTETİK AZINLIK AŞIRI ÖRNEKLEME TEKNİĞİ İÇİN YENİ ÖNERİLEN YÖNTEMLER

YÜKSEK LİSANS TEZİ

Hakan KORUL
(528211075)

Veri Mühendisliği ve İş Analitiği Anabilim Dalı

Büyük Veri ve İş Analitiği Programı

Tez Danışmanı: Dr. Öğr. Üyesi Mehmet Ali ERGÜN

AĞUSTOS 2024

Hakan KORUL, a M.Sc. student of ITU Graduate School of Science Engineering and Technology student ID 528211075 successfully defended the thesis entitled "NEW PROPOSED METHODS FOR SYNTHETIC MINORITY OVER-SAMPLING TECHNIQUE", which he prepared after fulfilling the requirements specified in the associated legislations, before the jury whose signatures are below.

Thesis Advisor :	**Assist. Prof. Dr. Mehmet Ali ERGÜN** Istanbul Technical University	**………..**
Jury Members :	**Assoc. Prof. Dr. Ömer Faruk BEYCA** Istanbul Technical University	**………..**
	Prof. Dr. Selim ZAİM Ibn Haldun University	**………..**

Date of Submission : **21 August 2024**
Date of Defense : **21 August 2024**

To my family,

FOREWORD

Dealing with imbalanced datasets and successfully predicting the minority class has always been a challenging problem. I have always encountered this difficulty in my own experience. Although the performance improves with existing synthetic data generation methods, my motivation is to develop a better method than these methods.

I am immensely grateful to my thesis advisor, Assistant Professor Doctor Mehmet Ali ERGUN, for his unwavering patience and belief in both myself and the challenges I faced throughout my extensive master's degree journey. Additionally, I extend my thanks to him for guiding me in discovering my interests and providing invaluable instruction on conducting thorough research.

August 2024

Hakan Korul
Data Scientist

TABLE OF CONTENTS

ABBREVIATIONS

SMOTE **:** Synthetic Minority Oversampling Technique

LIST OF TABLES

Page

LIST OF FIGURES

Page

NEW PROPOSED METHODS FOR SYNTHETIC MINORITY OVER-SAMPLING TECHNIQUE

SUMMARY

Machine Learning (ML), artificial intelligence (AI), aims to mimic human learning processes using data and algorithms. Machine learning algorithms are trained on data, which can be labeled or unlabeled, with the aim of finding patterns by considering each data variable and its corresponding label. Algorithms trained on data predict future data labels based on patterns learned from previously observed data variables. In machine learning algorithms, there is an error function that evaluates the algorithm's performance on the dataset, and to improve performance, the algorithm tries to find the one with a lower error value than the error function obtained previously.

The proliferation of ML algorithms and their utilization across various fields in the literature has led to the resolution of many challenging problems through machine learning. These problems can generally be categorized as supervised and unsupervised learning. Supervised learning, where data labels are known, can be further classified into regression and classification. Regression algorithms are used for problems such as predicting stock prices, predicting house prices, and forecasting weather conditions, while classification algorithms are employed for tasks like customer churn prediction, email spam detection, and image recognition. Problems where labels are not available in the dataset fall under unsupervised learning, which typically includes tasks like customer segmentation.

Ensuring machines are trained properly and achieve high performance relies heavily on the quality of the data. The data used for machine learning ideally should be clean, balanced, and representative of real-world conditions. Additionally, the size of the dataset is crucial, as larger datasets usually lead to better results. Alongside these data characteristics, having balanced classes in classification problems is vital for classification machine learning algorithms.

When it comes to classification problems, having balanced class distributions in the data is crucial for classification machine learning algorithms. If the class distribution is imbalanced, it leads to what is known as an imbalanced dataset. Training a machine learning model on a problem with balanced classes as opposed to imbalanced classes can result in noticeable differences in performance. For instance, in a problem with balanced classes, where 4 out of 10 customers churn, meaning almost a 50% probability, even a model with random guessing can make successful predictions. However, if the problem is such that only 1 out of 100 customers churn, random guessing becomes insufficient, posing an extremely challenging problem for machine learning models.

To address this issue, various methods have been developed. In the example mentioned above, where 99 customers do not churn and 1 customer churns, these problems are typically categorized into majority and minority classes. These different methods can essentially be classified into three categories: increasing the number of minority class samples to approach the number of majority class samples (oversampling), decreasing the number of majority class samples to approach the number of minority class samples (undersampling), or using a combination of both methods (hybrid sampling).

Directly comparing oversampling and undersampling methods can be quite challenging, as these methods may perform differently on different datasets. However, certain characteristics of the dataset can help expedite the selection of the appropriate method. For example, if the dataset is not very large, undersampling may not be suitable because it would eliminate majority class samples, resulting in a significantly smaller dataset. In such cases, oversampling methods that increase the overall dataset size should be preferred.

Oversampling methods can be further divided into random oversampling and SMOTE. The most significant difference between the two methods lies in how they handle the generation of synthetic data. While random oversampling duplicates existing minority class samples in the dataset, SMOTE generates synthetic data points close to existing minority class samples using various techniques. Since random oversampling does not introduce new examples to the machine learning algorithm and can lead to overfitting, SMOTE methods generally yield better performance improvements.

In the data generation stage with the SMOTE method, first, the number of k nearest neighbors is determined for each sample. To generate synthetic data, two examples are selected: the first is the minority sample itself, and the second is a randomly chosen minority class example among its k nearest neighbors. Then, the difference between the neighbor example and the minority example is calculated. Subsequently, this difference is multiplied by a random number between 0 and 1 and added back to the minority sample. This process ensures that the newly generated synthetic data is placed between the minority sample and one of its k nearest neighbors randomly selected.

Following the emergence of the SMOTE method in the literature, efforts have been focused on developing this method from various perspectives. Borderline SMOTE and K-Means SMOTE are just some of the newly proposed methods in this context. In the Borderline SMOTE method, a border between the minority and majority classes is defined, and synthetic data generation is focused only on the minority samples on this borderline. In the K-Means SMOTE method, the K-means technique is utilized, where the training set is divided into k clusters, and then decisions are made on which clusters to use for oversampling, followed by applying SMOTE steps within these clusters.

In my research, I developed three different new SMOTE methods. The first one, Genetic SMOTE, focuses on variable alteration using the crossover logic commonly used in genetic algorithms between a sample and its neighbor. Randomly selected features of the sample are chosen, then these features are taken from the neighbor, while the remaining features are directly taken from the sample, resulting in the generation of new synthetic data of the same size. The second one, Dual Borderline SMOTE, is inspired by the Borderline SMOTE method. Instead of considering the ratio of minority samples within the total number of neighbors applied for Borderline, which ranges between 0.5 and 1, I adjusted it to range between 0 and 1. The significant difference between this method and Borderline SMOTE is that, while Borderline SMOTE does not consider whether the neighbors of the minority sample are on the border after finding a minority sample on the border, in this method, both the sample and its neighbor must be on the newly defined border. The last method developed is Genetic Dual Borderline SMOTE, which combines these two methods. This method applies the Dual Borderline SMOTE rules for selecting minority samples and their close minority sample neighbors, meaning that both the selected sample and its neighbor must lie on the defined borderline. When generating synthetic data from the selected two examples, instead of applying SMOTE as in the Dual Borderline SMOTE method, it focuses on variable alteration similar to the Genetic SMOTE steps.

In the final section, to compare the performance of the three developed methods with the other three methods used, 8 datasets and 4 different machine learning algorithms were employed. For each dataset and algorithm, a total of 6 different SMOTE methods were evaluated for performance. During performance comparison, various parameter combinations of machine learning algorithms and SMOTE methods were tested for each dataset, and the parameters yielding the best results were selected for each dataset, model, and SMOTE method. The success of the SMOTE methods was compared using the parameters that provided the best F1 score. In the performance measurement, the F1 score of the minority class was preferred over metrics like accuracy, which are not suitable for imbalanced datasets. To reduce randomness and ensure a more reliable model, a 5-fold cross-validation score was used to calculate the F1 score. A total of 32 F1 scores (8 datasets, 4 machine learning algorithms) were obtained for each SMOTE method, and the F1 scores were illustrated in figures either overall or based on the model. Additionally, each dataset and machine learning algorithm were ranked from 1 to 6 for SMOTE methods, and the rankings were compared in the study. It has been observed that the newly developed methods outperform the existing ones. In model-based analyses, while there is not a significant difference in the performance of SMOTE methods in linear models, especially in ensemble methods, the Genetic Dual Borderline SMOTE method stands out noticeably from the others.

SENTETİK AZINLIK AŞIRI ÖRNEKLEME TEKNİĞİ İÇİN YENİ ÖNERİLEN YÖNTEMLER

ÖZET

Makine öğrenimi (ML), yapay zeka (AI) veri ve algoritmaların kullanılarak insan öğrenme süreçlerinin taklit edilmesidir. Makine öğrenme algoritmaları, veriyle eğitilir; bu veriler etiketli veya etiketsiz olabilir ve algoritma, her bir veri değişkeni ve etiketi göz önünde bulundurarak verideki desenleri belirlemeyi hedefler. Eğitilen algoritmalar, gelecekteki henüz bilinmeyen veri etiketlerini önceden öğrendikleri desenlerle tahmin ederler. Bu algoritmalarda hata fonksiyonu bulunur ve algoritma bu hata üzerinden veri seti üzerindeki başarısını değerlendirir; performansı iyileştirmek için ise daha önce elde edilen hata fonksiyonundan daha düşük bir hata değeri elde etmeye çalışır.

ML algoritmalarının geniş kullanımı ve farklı alanlarda literatürdeki artan önemi, birçok endüstride birçok zorlu problemi çözmeye yönelik makine öğrenimi uygulamalarının artmasına neden olmuştur. Bu problemler genellikle denetimli ve denetimsiz öğrenme olmak üzere iki ana kategoriye ayrılır. Veri etiketlerinin bilindiği denetimli öğrenme, regresyon ve sınıflandırma olarak daha da sınıflandırılabilir. Regresyon algoritmaları, stok fiyatları, ev fiyatları ve hava durumu gibi tahmin problemleri için kullanılırken, sınıflandırma algoritmaları müşteri ayrılma tahmini, e-posta spam tespiti ve görüntü tanıma gibi sınıflandırma görevleri için kullanılır. Etiketlerin bilinmediği problemler, denetimsiz öğrenme altında incelenir ve genellikle müşteri segmentasyonu gibi kümeleme problemlerini içerir.

Makinelerin doğru bir şekilde eğitilmesi ve yüksek performans sağlaması, kullanılan veri kalitesine büyük ölçüde bağlıdır. Makine öğrenimi için kullanılan verilerin ideal olarak temiz, dengeli ve gerçek dünya koşullarını temsil edici olması gerekir. Ayrıca, veri setinin boyutu da önemlidir; genellikle daha büyük veri setleri daha iyi sonuçlar verir. Bu veri özelliklerinin yanı sıra, sınıflandırma problemleri için dengeli sınıflara sahip olmak, sınıflandırma makine öğrenme algoritmaları için kritik öneme sahiptir. Sınıflandırma problemleri için dengeli sınıf dağılımına sahip olmak, makine öğrenme algoritmaları için önemlidir. Sınıf dağılımı dengesiz ise, bu durum dengesiz bir veri seti olarak bilinir. Dengeli sınıflara sahip bir problemde model eğitmek, dengesiz sınıflara sahip bir problemde model eğitmekten farklılık gösterir ve farklı performans sonuçlarına neden olabilir. Örneğin, dengeli sınıflara sahip bir problemde, 10 müşteriden 4'ü ayrılıyorsa, neredeyse %50 oranında rasgele başarılı tahminler yapılabilir. Ancak, eğer karşılaşılan problemde 100 müşteriden sadece 1'i ayrılıyorsa ve ayrılan müşteriyi tahmin etmek istiyorsanız, rasgele tahmin yetersiz kalır ve makine öğrenme modelleri için son derece zorlayıcı bir problem oluşur.

Bu sorunu çözmek için çeşitli yöntemler geliştirilmiştir. Yukarıda bahsedilen örnekte, 99 müşterinin ayrılmadığı ve 1 müşterinin ayrıldığı durumlar genellikle dengesiz sınıf problemi olarak adlandırılır ve bu tür veri setleri çoğunluk ve azınlık sınıflarını içerir. Bu problemi için geliştirilen farklı yöntemler genellikle azınlık sınıfı örneklerinin sayısını artırmayı, çoğunluk sınıfı örneklerinin sayısını azaltmayı veya her iki yöntemin kombinasyonunu kullanmayı amaçlar. Bu yöntemleri doğrudan

karşılaştırmak zor olabilir, çünkü bu yöntemler farklı veri kümelerinde farklı performanslar gösterebilir. Ancak, veri setinin belirli özellikleri, uygun yöntemin seçimini kolaylaştırabilir. Örneğin, veri kümesi çok büyük değilse, undersampling uygun olmayabilir çünkü çoğunluk sınıfı örneklerini ortadan kaldırarak önemli ölçüde daha küçük bir veri kümesine neden olur. Bu durumlarda, genel veri kümesi boyutunu artıran oversampling yöntemleri tercih edilmelidir.

Oversampling yöntemleri daha da genişletilebilir ve rasgele oversampling ve SMOTE olmak üzere ikiye ayrılabilir. İki yöntem arasındaki en önemli fark, sentetik veri üretme şekillerinden kaynaklanmaktadır. Rasgele oversampling, veri kümesindeki mevcut azınlık sınıfı örneklerini birebir kopyalar; SMOTE ise farklı teknikler kullanarak mevcut azınlık sınıfı örneklerine yakın sentetik veri noktaları oluşturur. Rasgele oversampling, makine öğrenme algoritmasına yeni örnekler tanıtmadığı ve aşırı öğrenmeye yol açabileceği için genellikle SMOTE yöntemleri rastgele oversampling yönteminden daha iyi performans gösterir. SMOTE yöntemiyle veri üretme aşamasında öncelikle k adet yakın komşunun sayısı belirlenir. Sentetik veri üretmek için iki örnek seçilir: ilki azınlık örneği kendisi, ikincisi ise k adet en yakın komşuları arasından rasgele seçilen komşu azınlık sınıfı örneğidir. Daha sonra, komşu örneği ile azınlık örneği arasındaki fark hesaplanır. Ardından, bu fark 0 ile 1 arasında rasgele bir sayıyla çarpılır ve azınlık örneğine geri eklenir. Bu süreç, yeni üretilen sentetik verinin azınlık örneği ile komşusunun arasında bulunmasını sağlar.

SMOTE yönteminin literatürde ortaya çıkmasının ardından, bu yöntemi farklı bakış açılarıyla geliştirmeye yönelik çabalar yoğunlaşmıştır. Borderline SMOTE ve K-Means SMOTE, bu bağlamda önerilen yeni yöntemlerden sadece birkaçıdır. Borderline SMOTE yönteminde, azınlık ve çoğunluk sınıfları arasında bir sınır tanımlanır ve sentetik veri oluşturma işlemi yalnızca bu sınırlarda bulunan azınlık örneklerine odaklanır ve bu sınır üzerinde yer almayan azınlık sınıfına ait örneklerden ekstra sentetik veri üretiminin öğrenmeye bir katkısı olmayacağı belirtilir. K-Means SMOTE yönteminde ise, eğitim seti k adet kümeye ayrılır ve ardından hangi kümelerin oversampling için kullanılacağına karar verilir. Bu karar aşamasında, cluster içerisindeki azınlık ve çoğunluk sınıfı oranları dikkate alınır. Bu kararların ardından bu kümeler içinde bağımsız şekilde ve ağırlıklı bir biçimde SMOTE adımları uygulanır.

Bu çalışmada, daha başarılı sonuçlar elde etmek için üç farklı yeni SMOTE yöntemi geliştirilmiştir. İlk yöntem olan Genetik SMOTE, genetik algoritmalarda sıkça kullanılan çaprazlama mantığına dayanarak değişkenler arasındaki değişime odaklanır. Bu yöntemde, öncelikle 1 ile veri kümesindeki örneklerin değişken uzunluğu arasında rastgele sayılar belirlenir. Bu rastgele sayılar, hangi değişkenlerin değiştirileceğinin indeksini temsil eder. Sentetik veri üretiminde, önce örnek kopyalanır, sonra rastgele seçilen indeksteki değişkenler, ilgili örneğin seçilen komşusundaki değişkenlerle güncellenir. Böylece, verideki örneklerle aynı uzunlukta sentetik veri üretilmiş olur. Bu yöntemde, literatürdeki diğer SMOTE yöntemlerinden farklı olarak, yeni üretilecek sentetik verinin, örnek ve komşusu arasında olacak şekilde yeni bir değişken değeri üretmeyerek, veride bulunan mevcut değişkenleri kullanarak sentetik veri üretilir.

İkinci yöntem olan Dual Borderline SMOTE, Borderline SMOTE yönteminden esinlenmiştir. Bu yöntemin Borderline SMOTE yönteminden önemli bir farkı, Borderline SMOTE ile sentetik veri üretiminde azınlık örneğinin sınırda bulunması yeterlidir. Azınlık örneğinin komşularının da tanımlanan sınırlarda olup olmadığını göz önünde bulundurmamasıdır. Ancak bu yeni geliştirilen yöntemde, hem örnek hem de komşusunun yeni tanımlanan sınırlar üzerinde olması gerekmektedir. Ayrıca diğer

bir fark ise, Borderline SMOTE için sınır tanımlamasında, azınlık örneğinin toplam komşu sayısı içerisinde azınlık komşu örneklerinin oranının 0.5 ile 1 arasında olması gerekliyken, yeni yöntemde sınır tanımı için bu oran 0 ve 1 arası şeklinde düzenlenmiştir. Burada bu oran 0 ise, azınlık örneğinin toplam k adet komşusu içerisinde azınlık örneği bulunmadığını ve bu örnekten üretilecek sentetik veri üretiminin öğrenmeyi yanıltacağı düşünülürken, bu oranın 1 olması ise bütün komşuların azınlık sınıfına ait olduğunu gösterir ve zaten yoğun bir azınlık bölgesinden tekrardan yeni bir sentetik veri üretmenin öğrenme için faydalı olmadığı düşünülmüştür.
Son geliştirilen yöntem ise, ilk iki yöntemin beraber uygulandığı Genetic Dual Borderline SMOTE' tur. Bu yöntem, sentetik veri üretmek için gereken örnek ve komşu seçimi için Dual Borderline SMOTE kurallarını uygular. Yani seçilen örnek ve komşusunun, tanımlanan sınırlar üzerinde olması gerekmektedir. Seçilen iki örneğe dayanarak sentetik veri oluşturulurken, Genetik SMOTE adımlarında olduğu gibi yeni bir değişken üretmek yerine, genetik algoritmalarında kullanılan örnek ve komşusu arasında değişken değişimine odaklanır.
Son bölümde, geliştirilen üç yöntemin diğer mevcut üç yöntemle karşılaştırılması için 8 veri seti ve 4 farklı makine öğrenme algoritması kullanıldı. Her veri seti ve algoritma için toplamda 6 farklı SMOTE yöntemi performans açısından değerlendirildi. Performans karşılaştırması sırasında, her veri seti için makine öğrenme algoritmaları ve SMOTE yöntemlerinin çeşitli parametre kombinasyonları test edildi ve her veri seti, model ve SMOTE yöntemi için en iyi sonuçları sağlayan parametreler seçildi. Böylece SMOTE yöntemlerinin başarısı, en iyi F1 skorunu sağlayan parametreler kullanılarak karşılaştırıldı. Performans ölçümünde, dengesiz veri kümeleri için uygun olmayan doğruluk gibi metrikler yerine dengesiz veri setinin performansı için daha uygun bir metrik olan azınlık sınıfının F1 skoru metriği tercih edildi. Daha güvenilir bir başarı metriği oluşturmak için 5 kat çapraz doğrulama skoru kullanılarak F1 skoru hesaplandı. Her bir SMOTE yöntemi için toplamda 32 F1 skoru (8 veri seti, 4 makine öğrenme algoritması) elde edildi ve SMOTE yöntemlerinin F1 skoru, figürler ile genel veya model bazlı şeklinde grafiksel olarak gösterildi. Ayrıca, her veri seti ve makine öğrenme algoritması için SMOTE yöntemleri 1 ile 6 arasında başarısına göre sıralandı ve bu sıralamalar yine grafikler ile karşılaştırıldı. Yeni geliştirilen yöntemlerin mevcut yöntemleri geride bıraktığı gözlemlenmiştir. Model tabanlı analizlerde, doğrusal modellerde SMOTE yöntemlerinin performansında belirgin bir fark olmasa da, özellikle ensemble yöntemlerinde, Genetik Dual Borderline SMOTE yöntemi diğerlerinden belirgin şekilde ayrışmaktadır.

1. INTRODUCTION

In today's world, where machine learning is becoming increasingly prevalent, various industries are endeavoring to solve challenging problems by training machines. Imbalanced classes pose a significant challenge, which is commonly encountered in real-world applications. Imbalanced class distributions are frequently observed in scenarios such as automated inspection systems overseeing products on a manufacturing assembly line, where the number of defective products is significantly lower than that of non-defective ones. Similarly, in credit card fraud detection, legitimate transactions far outnumber fraudulent ones. In both cases, there is an uneven distribution of instances across different classes (KrishnaVeni and Sobha, 2011).

Issue of class imbalance in machine learning pertains to classification tasks with unequal distribution among data classes (Last et al., 2017). In such problems, a class is considered rare when the number of samples is significantly lower compared to the other class, making the problem more challenging. Recognizing infrequent occurrences poses a predictive challenge in machine learning, and due to their rare nature in everyday observations, predicting such instances encounters difficulties (Haixiang et al., 2017).

The difficulty of these problems arises primarily because the machine encounters the majority class frequently, allowing it to learn that class effectively, while struggling to learn the minority class due to its rare occurrence. Performance evaluation is crucial in addressing such problems. Accuracy, one of the most popular metrics, is not suitable and can be misleading. A simple prediction in a dataset with 98% normal pixels and 2% abnormal pixels, such as the mammography dataset, would result in an accuracy value of around 0.98. However, accuracy is not suitable for such scenarios, and more weight should be given to correctly predicting the minority class (abnormal pixels) (Chawla et al., 2002).

Most traditional classification algorithms aim to minimize the error rate, representing the percentage of incorrectly predicted class labels. These algorithms often overlook the distinctions between various types of misclassification errors, assuming that all errors carry equal costs. However, in many real-world scenarios, this assumption does not hold true. The differences between different misclassification errors can be significant. For instance, in the medical diagnosis of a specific cancer, the model's inability to predict a cancer patient can jeopardize the patient's life; therefore, this error is highly significant (Ganganwar, 2012).

In the second part of this study, the definition of imbalanced datasets is provided, followed by a discussion of popular methods including SMOTE, Borderline SMOTE, and K-Means SMOTE along with their details. The third part of the study elaborates on the newly developed methods. In the fourth section, a detailed comparison of the performance of the newly developed methods with the existing three SMOTE methods is presented. Lastly, in the conclusion part, the findings of the study are summarized, and suggestions for future research on further enhancing this proposed method are provided.

2. RELATED STUDIES

Approaches to handling imbalanced datasets can be categorized into solutions at the data level and the algorithmic level. Techniques at the data level alter the distribution of imbalanced datasets, resulting in balanced datasets presented to the learner to enhance the detection rate of the minority class. On the other hand, methods at the algorithmic level either adapt existing data mining algorithms or propose new ones to address the challenge of imbalance (Han et al., 2005).

Data-level methods focus on resolving class imbalance in the training set by modifying it. They eliminate the imbalance by reducing the number of the majority class (undersampling) through selecting random samples from the majority class. For the minority group, they attempt to reduce class imbalance by reintroducing the same samples to the training set or by generating synthetic minority samples suitable for the dataset (oversampling), taking into account the samples in the minority class. SMOTE methods are a type of oversampling method that produces new minority samples and are very successful in preventing overfitting compared to random oversampling methods, which duplicate the same data (Last et al., 2017).

Resampling methods aim to facilitate the learning process by bringing the distribution between classes in the unbalanced class problem closer to non-skewed. Resampling methods are divided into three categories: oversampling, undersampling, and the hybrid method, which is a combination of these two approaches. 156 papers on resampling techniques were examined. Among these papers, over-sampling methods were selected 84 times, under-sampling methods were chosen 39 times, and hybrid-sampling methods were selected 33 times (Haixiang, 2017).

SMOTE (Synthetic Minority Over-sampling Technique) method, which generates new synthetic data by considering the minority samples in the training cluster. Essentially, this algorithm creates new synthetic data points between the minority samples in the training set and their nearest neighbors. When producing new variable values between an existing sample and its neighbor, the algorithm performs a vector subtraction between the sample and its neighbor. It then multiplies the result by a random number between 0 and 1 and adds the vector back to the sample. This ensures that the variables of the newly generated data point fall between the sample and its neighbor (Chawla et al., 2002). Figure 2.1 mathematically illustrates how the SMOTE method is applied between two samples.

Consider a sample (6,4) and let (4,3) be its nearest neighbor.
(6,4) is the sample for which k-nearest neighbors are being identified.
(4,3) is one of its k-nearest neighbors.
Let:
f1_1 = 6 f2_1 = 4 f2_1 - f1_1 = -2
f1_2 = 4 f2_2 = 3 f2_2 - f1_2 = -1
The new samples will be generated as
(f1',f2') = (6,4) + rand(0-1) * (-2,-1)
rand(0-1) generates a random number between 0 and 1.

Figure 2.1 : Example of generation of synthetic examples (SMOTE) (Chawla et al., 2002).

Figure 2.2 visualizes how the SMOTE method positions the synthetic data using linear interpolation with a random number between 0 and 1.

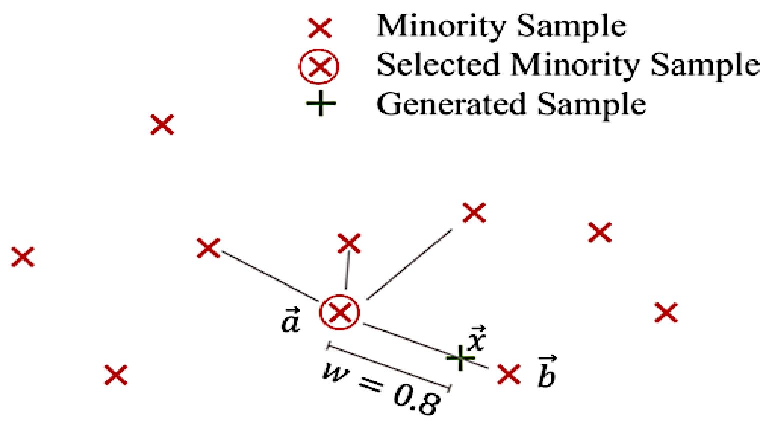

Figure 2.2 : SMOTE linearly interpolates a randomly selected minority sample and one of its k = 4 nearest neighbors (Last et al., 2017).

Generating synthetic data for minority samples located far from the border region will not significantly contribute to the classification process. As a result, they proposed a new algorithm called Borderline SMOTE, which considers all neighbors of minority samples in the training set and specifically targets minority samples in the border region, while considering the majority count among all neighbors (Han et al., 2005).

In their article, the algorithm is explained as follows: For each minority sample, a certain number of neighbors in the entire training set are considered, and among these neighbors, there may be samples belonging to the minority or majority class. Upon examining the distribution, if all neighbors belong to the majority class, this point is classified as noise. If the number of majority neighbors in the total number of neighbors is equal to or greater than half of the total number of neighbors but less than the total number of neighbors, this point is defined as 'DANGER,' while the other case is defined as safe. Only for minority samples in the 'DANGER' zone, the nearest minority sample neighbors are calculated. Then, synthetic data is generated between this minority sample and its closest neighbors, similar to the SMOTE method (Han et al., 2005).

In Figure 2.3, the visualization of the dataset highlights how the Borderline SMOTE algorithm defines borderline samples during the synthetic data generation process and how the generated synthetic data is positioned between the two classes.

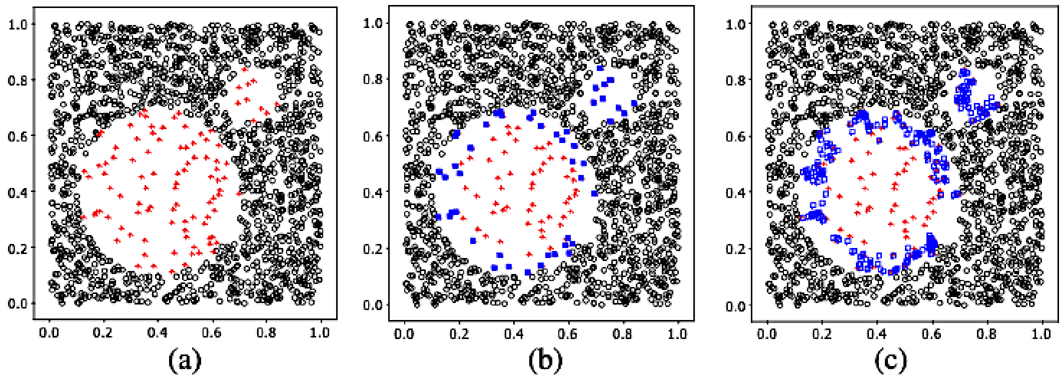

Figure 2.3 : (a) The initial distribution of the dataset. (b) The minority samples located at the borderline (represented by solid squares). (c) The synthetic minority samples generated at the borderline (depicted by hollow squares) (Han et al., 2005).

K-means SMOTE method, a new synthetic data generation algorithm that combines K-means and SMOTE methods. The algorithm consists of three stages: clustering, which involves dividing the training cluster into groups using K-means; filtering, which examines the minority and majority numbers in the cluster to determine the clusters selected for oversampling; and finally, oversampling, which produces synthetic data using the SMOTE method for the clusters selected in the filtering stage (Last et. al., 2017). Figure 2.4 illustrates the stages of K-means SMOTE. Initially, the training set is grouped into three clusters using the k-means algorithm. Subsequently, for each cluster, the imbalance ratio is calculated by dividing the minority sample count within the cluster by the majority count. For example, in the top-left cluster, the total number of minority class samples is 5, while the majority class sample count is 2, resulting in an imbalance ratio of 5/2. This process is repeated for other clusters, and clusters with an imbalance ratio greater than 1, indicating that the minority class count is greater than the majority class count, are filtered. In the next oversampling stage, only the selected clusters are used. During the oversampling stage, for each selected cluster's minority samples, k nearest minority neighbors are found within the same cluster, and new synthetic data is generated using the SMOTE method between the sample and its neighbor in the cluster.

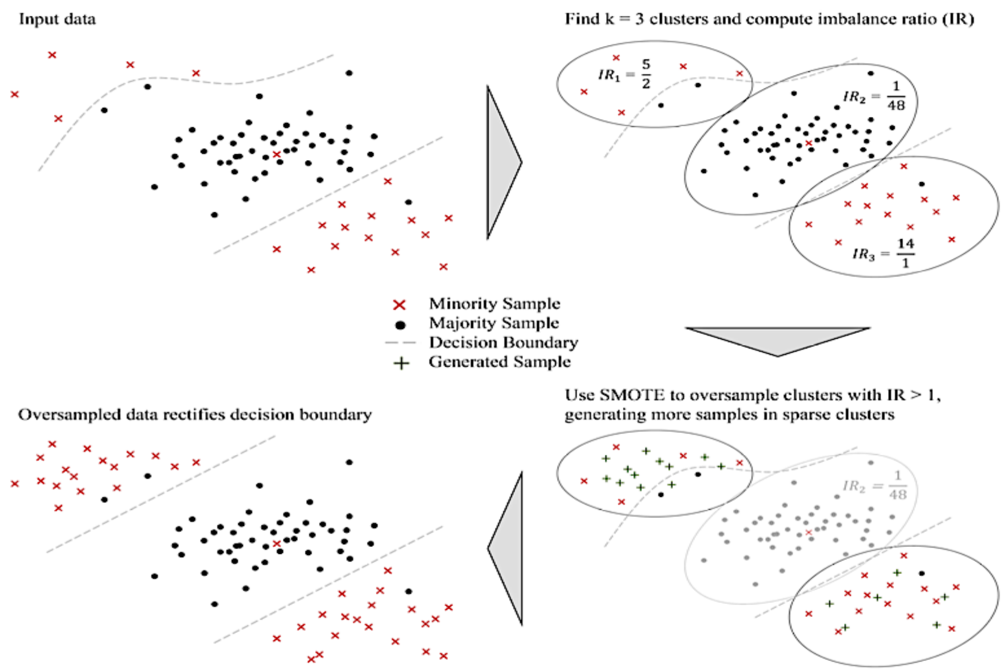

Figure 2.4 : K-Means SMOTE oversamples safe areas and combats within-class imbalance (Last et al., 2017).

3. PROPOSED METHODS

3.1 Genetic SMOTE

The first method, Genetic SMOTE, focuses on the variation of features between samples and their neighbors, rather than generating synthetic data with a random specific alpha coefficient commonly used in many methods in the literature. It gets its name from the crossover logic found in genetic algorithms.With this method, instead of creating an entirely new synthetic data with different variables from scratch, a random selection is made between 1 and the number of features to determine which features will be altered. Subsequently, synthetic data is copied from the sample, and then the selected variables to be changed are taken from the neighbor. This way, a new synthetic data is generated. In Figure 3.1, an attempt is made to visualize how new synthetic data is generated using this method when the randomly selected index numbers are 1, 3, and 4.

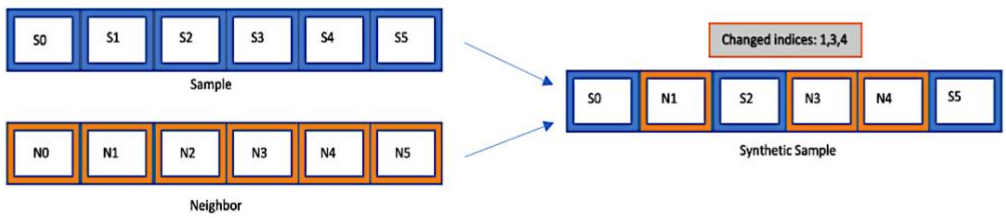

Figure 3.1 : Crossover between minority sample and its neighbor.

The algorithm of the method is as follows: Suppose that the minority class is *MN*, and

$$MN = \{mn_1, mn_2, mn_3, mn_4, \ldots, mn_n\} \quad (3.1)$$

where n is the total number of minority samples in *MN*.

Step 1. For each mn_i (i = 1,2,..., n) in the minority class *MN*, find its *k* nearest neighbors from the entire minority class *MN*, and

$$NN_i = \{nn_{i1}, nn_{i2}, \ldots, nn_{ik}\} \quad (3.2)$$

where NN_i is the set of nearest neighbors of the minority sample mn_i. Here, k is the selected total number of nearest neighbors and all samples in NN_i are also part of the *MN*.

Step 2. For each mn_i (i = 1,2,..., n) and its nearest neighbors in NN_i, select *T* random numbers *(RN)*, where *T* must be between 1 and *L*-1, and *L* is the total length of samples.

$$RN = \{rn_1, rn_2, \ldots, rn_T\} \tag{3.3}$$

Create a synthetic sample *S* by copying mn_i .

$$S = [mn_{i1}, mn_{i2}, \ldots mn_{iL}] \tag{3.4}$$

Here, mn_{i1} represents the first element of the vector mn_i, and mn_{iL} represents the last element of the vector mn_i. Then for each random number in *RN*, change the value of *S*, by replacing *RNth* position of element in *S* vector, which is mn_{iRN}, with NN_{i1RN} (*RNth* element of first neighbor of mn_i). As seen in equation 3.5, the modified version of *S* can be defined as follows: for every position *(rn)* in *RN*, change the value of the position *rn* in *S* vector by assigning the value of the position *rn* in nearest neighbor.

$$S[rn] = NN_{irn} \tag{3.5}$$

Repeat these steps until the desired number of synthetic data points is achieved.

For example, assume that the first element of the minority class (mn_1) is [6, 12, 5, 14, 24], and among the 5 (k) nearest neighbors, suppose that nn_{12}, the second nearest neighbor of mn_1, is [3, 2, 13, 1, 33]. In the next step, choose random numbers between 1 and 5 (L, the length of samples). Suppose the chosen random numbers are RN is {2, 4}. First, create a synthetic sample S by copying all elements of mn_1. S becomes [6, 12, 5, 14, 24]. Then, for each element in RN, change the value of S so that the second element of S will be the second element of nn_{12} (which is 2), and the fourth element of S will be the fourth element of nn_{12} (which is 1). After these steps, the final version of S will become [6, 2, 5, 1, 24]. This process is performed for all neighbors of mn_1, generating a total of k synthetic data points. Repeat this step for all minority class examples until the desired total number of synthetic data points is reached.

3.2 Dual Borderline SMOTE

The second method, Dual Borderline SMOTE, is inspired by the Borderline SMOTE method. In this method, the ratio of how many neighbors of the minority sample are belong to minority class is calculated from the total number of chosen neighbors. If this minority ratio value is greater than 0 and less than 1, the minority sample is considered suitable for synthetic data generation. If the minority ratio value is 0, it indicates that all neighbors are minority, suggesting that the sample is in a safe or minority-dense area. In such a case, generating synthetic data from this sample may not be beneficial for the model. If the ratio value is 1, the minority sample is considered dangerous or an outlier. In this case, generating synthetic data may lead to mislabeling issues, so it is not preferred.

While these methods resemble the Borderline SMOTE method, the factor that distinguishes them from the algorithm is the importance of both the minority sample and its neighbor's minority ratio values. In the Borderline method, only the sample being on the border is sufficient, and it does not consider whether its neighbors meet the same condition when looking at them. In the Dual Borderline SMOTE method, both the minority sample and its close minority neighbor must have a minority ratio value between 0 and 1. If this condition is met, a synthetic data point is generated between these two examples.

The algorithm of the method is as follows: Suppose that the all training set is A, the minority class is N and the majority class is J and

$$N = \{n_1, n_2, n_3, \ldots., n_{nnum}\},\ J = \{j_1, j_2, j_3, \ldots., j_{jnum}\} \tag{3.6}$$

where $nnum$ and $jnum$ are the total number of minority and majority samples.

Step 1. For each n_i (i = 1,2,..., $nnum$) in the minority class N, find its k nearest neighbors from the entire training set A, and

$$NN_i = \{nn_{i1}, nn_{i2}, \ldots., nn_{ik}\} \tag{3.7}$$

where NN_i is the nearest neighbors of the minority sample n_i, k is the selected total number of nearest neighbors and all samples in NN_i are also part of the A.

Step 2. The examples in the NN_i set can belong to either the minority or majority class. For instance, nn_{i1}, the nearest neighbor of the minority sample n_i, may belong to the majority class, while nn_{i2}, the second nearest neighbor, may belong to the minority class. Let's denote the total number of minority examples among the k nearest neighbors as k_n for each sample. For each minority example, the minority neighbor ratio can be calculated as k_n / k. If this ratio is 0, it is defined as a safe zone*;* if it is 1, it is considered a risky zone; and if it is between 0 and 1, it is identified as an suitable zone. The minority samples located in the suitable zone are considered for synthetic data generation.

Step 3. The examples in SUITABLE are also examples in minority class N.

$$\text{SUITABLE} = \{ n'_1, n'_2, n'_3, \ldots, n'_{snum} \},\ \ 0 \leq snum \leq nnum \qquad (3.8)$$

For each n'_i, this time find k nearest neighbors only in minority class N and take only neighbors if it is example of SUITABLE. Suppose that from k nearest neighbors only k' *neighbors in* SUITABLE.

Step 4. Calculate differences like dif_j $(j=1,2,..,k')$ by assessing the disparities between n'_i and its k' closest neighbors located in SUITABLE zone from N. Subsequently, we multiply dif_j by a randomly generated factor r_j (j = 1,2,..., k'), where r_j falls within the range of 0 to 1. The number of k' new synthetic examples is produced as shown in equation 3.9.

$$s_j = n'_i + r_j \times dif_j, \qquad j = 1,2,\ldots,k' \qquad (3.9)$$

These steps are repeated until the desired number of synthetic data points is achieved.

In the SMOTE method, the problem of generating synthetic data from noisy minority group samples is addressed by defining algorithmic noise, and synthetic data generation from these samples is avoided. Additionally, the issue of generating more synthetic data from dense regions in the SMOTE method is similarly resolved by not generating synthetic data from these samples. Unlike Borderline SMOTE, since it performs a two-way control, it is more focused on border points, which increases the likelihood of producing more beneficial synthetic data for the machine learning model.

3.3 Genetic Dual Borderline SMOTE

The third method is a combination of the two methods mentioned above. In the Genetic SMOTE method, instead of generating features for each feature between the sample and its neighbor, as many SMOTE methods do, it changes the features randomly. On the other hand, Dual Borderline SMOTE examines the minority neighbor ratio within the total of k neighbors for both the sample and its neighbor to generate synthetic data between them. If these two instances are suitable, it produces synthetic data between the sample and its neighbor.

By combining these two methods, the new proposed method selects only the minority sample and its neighbor with minority neighbor ratio values between 0 and 1 to produce synthetic data (SUITABLE zone explained in step 1,2,3 of Dual Borderline SMOTE, see pages 9-10 for a detailed explanation). After determining these two suitable points, the variables to be changed are randomly selected, and these variables are replaced in the sample and its neighbor (random feature crossover explained in step 2 of Genetic SMOTE, see page 8 for a detailed explanation). Thus, the new synthetic data is produced by combining the random feature crossover from Genetic SMOTE with the suitable point selection from Dual Borderline SMOTE.

4. RESEARCH METHODOLOGY

4.1 Metrics

There are several methods to evaluate model performance in binary classification problems, with accuracy, precision, recall, F1 score, and AUC scores being among the most popular. However, not every performance metric is suitable for use in imbalanced datasets. Some metrics have been developed specifically to accurately evaluate success in imbalanced datasets and are commonly applied in such scenarios.

Although accuracy is perhaps the most popular and easily understood metric, it is misleading when used for imbalanced datasets. The primary goal in imbalanced datasets is typically to correctly predict the minority class. Since the accuracy metric incorporates information about how many of the majority group predictions were correct, the resulting score is often high and misleading. Equation 4.1 shows the mathematical formula for calculating the accuracy metric in binary classification models. Here, the denominator includes all predictions, while the numerator consists only of the correct predictions, i.e., the true positives and true negatives.

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN) \tag{4.1}$$

On the other hand, the F1 score is a more reliable metric, especially in imbalanced datasets, unlike the accuracy metric. The F1 score represents the harmonic mean of precision and recall values. Equation 4.2 shows the mathematical formula for calculating the F1 score. While precision measures the proportion of predicted minority class samples that actually belong to the minority class, recall evaluates how many of the samples that truly belong to the minority class are correctly predicted as such.

$$F1 = (2 * \text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall}) \tag{4.2}$$

These metrics can be better understood by examining the confusion matrix in Figure 4.1 below. In essence, TP (True Positives) and TN (True Negatives) represent correctly predicted instances, while FP (False Positives) and FN (False Negatives) denote incorrectly predicted examples.

	P Positives	N Negatives	
PP Predicted Positives	TP True Positives	FP False Positives	Precision $\frac{TP}{PP}$
PN Predicted Negatives	FN False Negatives	TN True Negatives	
	Sensitivity / Recall $\frac{\mathrm{TP}}{P}$	Specificity $\frac{TN}{N}$	

Figure 4.1 : Confusion matrix (Last et al., 2017).

4.2 Oversamplers

In this study, six different oversampling methods were employed. The SMOTE, Borderline SMOTE, and K-means SMOTE methods were utilized by importing them from the imbalance-learn library in Python. Additionally, three distinct new algorithms, namely Genetic SMOTE, Dual Borderline SMOTE, and Genetic Dual Borderline SMOTE, were developed and employed.

These oversampling methods involve different parameters. Although there is no general rule for the minority sample ratio in these methods, we considered a scenario where the number of minority samples equals the number of majority samples to compare different SMOTE methods. In all datasets, synthetic data was generated using the SMOTE method in such a way that the number of minority class samples in the training set equaled the number of majority class samples. No sampling was applied to the test dataset to reflect real-world conditions.

Another crucial parameter for SMOTE methods is the number of neighbors. In this study, the same number of neighbors was tested for each SMOTE method. However, different cluster balance threshold values were experimented with for the K-means SMOTE method to ensure proper functionality.

- SMOTE, Borderline SMOTE
  - K-Nearest Neighbors $\in$ {2, 3, 4, 5, 6}
- K-means SMOTE
  - K-Nearest Neighbors $\in$ {2, 3, 4, 5, 6}
  - Cluster Balance Threshold $\in$ {'auto', 0.1, 0.2, 0.4, 0.6}
- Genetic SMOTE, Dual Borderline SMOTE, Genetic Dual Borderline SMOTE (Proposed Methods)
  - K-Nearest Neighbors $\in$ {2, 3, 4, 5, 6}

4.3 Classifiers

After applying SMOTE methods to each dataset, the performance of the test data was assessed using different models. Four distinct models were employed: Random Forest, Linear SVM, Logistic Regression, and Light GBM.

Various hyperparameter combinations for these models were executed with different combinations of SMOTE methods in the preceding step. Consequently, the best validation cross-validation (CV) F1 score for the minority class was determined for each dataset, model, and SMOTE method.

The application of the models and SMOTE method follows this process:

1. Each dataset is partitioned into five separate folds for training and validation, ensuring the minority-to-majority class ratio remains consistent across both sets.
2. The SMOTE method is applied to the training set in each fold.
3. Subsequently, the training set is fitted with the model, and the F1 score is computed by predicting the test set.
4. The validation CV score for this SMOTE method and model is determined by averaging the F1 scores across the five separate validation sets.

For each dataset, all model and SMOTE method combinations are tested using the above procedure, and the maximum score for the relevant dataset, model, and SMOTE method is identified.

- Random Forest Classifier
  - Number of Estimators $\in$ {2, 4, 8}
  - Min samples Leaf $\in$ {4, 8, 16}
  - Max Depth $\in$ {None, 2}
- Logistic Regression
  - Penalty $\in$ {'l2', None}
  - C $\in$ {0.01, 0.1, 1, 5}
- Linear SVC
  - Penalty $\in$ {'l2', 'l1'}
  - C $\in$ {0.01, 0.1, 1, 3}
- Light GBM
  - Max Depth $\in$ {-1, 2, 4}
  - Number of Estimators $\in$ {20, 40}
  - Minumum Child Samples $\in$ {40, 60}

4.4 Datasets

Eight binary classification datasets were utilized to compare the performance of SMOTE methods. Two of these datasets were sourced from Kaggle, while the remaining six were obtained from the Keel dataset repository. Information about these datasets, including size, number of variables, and the imbalance ratio (i.e., the ratio of the majority class size to the minority class size), can be found in Table 4.1.

Table 4.1 : Summary information of the used datasets.

data_name	data_size	feature_size	imbalance_ratio
healthcare_stroke_kaggle	4,909	17	5.66
it_churn_kaggle	7,032	31	2.76
page_blocks0_keel	5,472	11	8.79
yeast-2_vs_4_keel	514	9	9.08
yeast-0-5-6-7-9_vs_4_keel	528	9	9.35
vowel0_keel	988	14	9.98
shuttle-c0-vs-c4_keel	1,829	10	13.87
yeast-1_vs_7_keel	459	8	14.30

4.5 Results

In the experimental results section, eight different datasets, four different models, and six different SMOTE methods were tested, and the cross-validation F1 score for class 1 (the minority class) was obtained. The model and SMOTE parameters that provided the best cross-validation score for each dataset, model, and SMOTE method were utilized. When comparing the scores, the optimal situations for each SMOTE method were considered.

For each SMOTE method, a score calculation was performed. In this calculation, the F1 value was not directly used. Instead, for each dataset and model, the best F1 value obtained when using the respective SMOTE method was subtracted from the best F1 score obtained when no SMOTE method was used for the same dataset and model. This difference is then referred to as the 'score.'

In Table 4.2, the score values calculated for the relevant model and SMOTE method on the "yeast-1_vs_7_keel" dataset can be seen.

Table 4.2 : Scores of SMOTE methods on yeast-1_vs_7_keel dataset.

smote_method	lgbm	linearsvc	logisticregression	randomforest
Borderline SMOTE	0.159	0.134	0.135	0.195
Dual Borderline SMOTE	0.205	0.138	0.139	0.232
Genetic Dual Borderline SMOTE	0.250	0.143	0.144	0.195
Genetic SMOTE	0.153	0.157	0.158	0.152
K-Means SMOTE	0.031	0.013	0.015	0.033
SMOTE	0.179	0.129	0.130	0.186

In Figure 4.2, the average scores of each the SMOTE method are calculated without any dataset and model breakdown. As we can see in Figure 4.2, newly developed SMOTE methods has demonstrated better performance compared to other SMOTE methods.

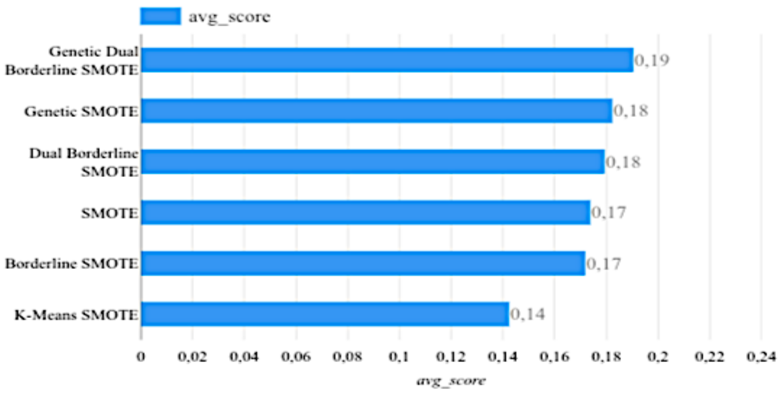

Figure 4.2 : Average scores of SMOTE methods.

In Figure 4.3, the score metric is analyzed on a model basis. There wasn't much of a difference between the SMOTE methods for linear models. However, when it came to ensemble models, the new SMOTE algorithms performed better than the others.

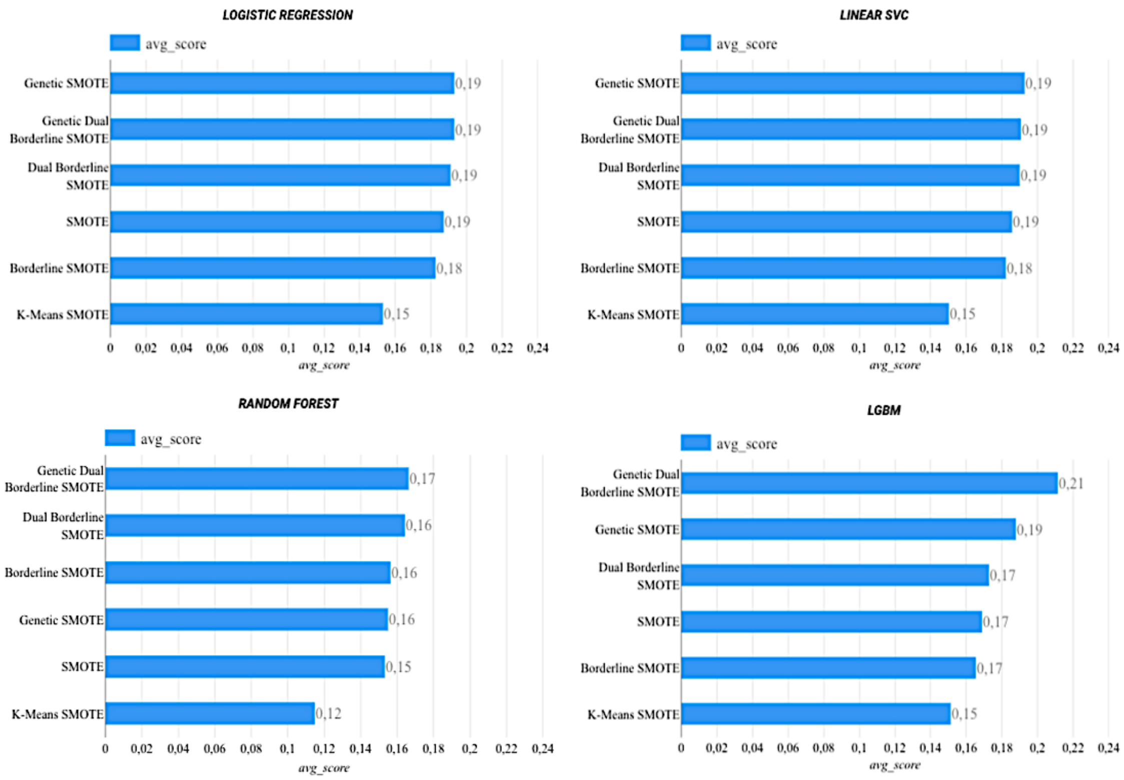

Figure 4.3 : Model based average scores of SMOTE methods.

With this score metric, ranks were also found based on dataset and model. For instance, in the "yeast-1_vs_7_keel" dataset in Table 4.3, the Genetic Dual Borderline SMOTE method achieved the highest score in the LGBM model with a value of 0.25. Therefore, the rank of this method is designated as 1. Similar rank values are assigned to other SMOTE methods.

Table 4.3 : Ranks of SMOTE methods on yeast-1_vs_7_keel dataset.

smote_method	lgbm	linearsvc	logisticregression	randomforest
Borderline SMOTE	4	4	4	2
Dual Borderline SMOTE	2	3	3	1
Genetic Dual Borderline SMOTE	1	2	2	3
Genetic SMOTE	5	1	1	5
K-Means SMOTE	6	6	6	6
SMOTE	3	5	5	4

In Figure 4.4, the average rank values for each SMOTE method are calculated without any dataset and model breakdown. It is clearly evident that the newly developed SMOTE methods outperform the existing ones in terms of average rank values.

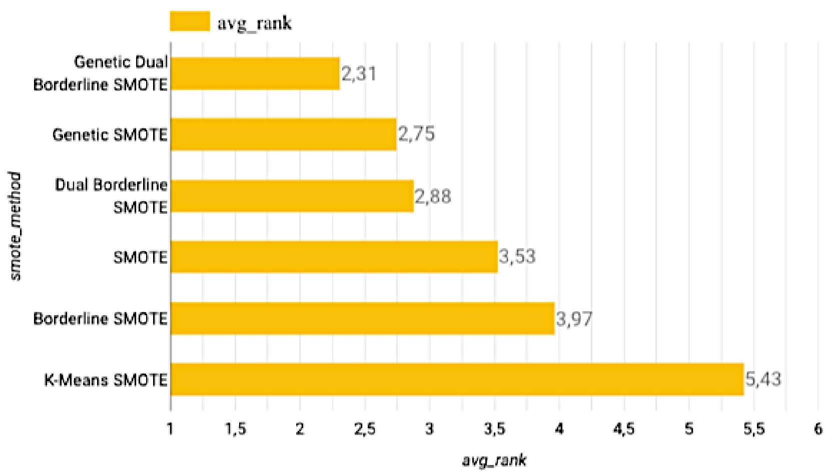

Figure 4.4 : Average rank values of SMOTE methods.

In Figure 4.5, the average rank values of SMOTE methods are calculated based on the model. While the rank values are close in linear models, in ensemble algorithms, especially in the light GBM algorithm, the developed Genetic Dual Borderline SMOTE method is noticeably superior to the others.

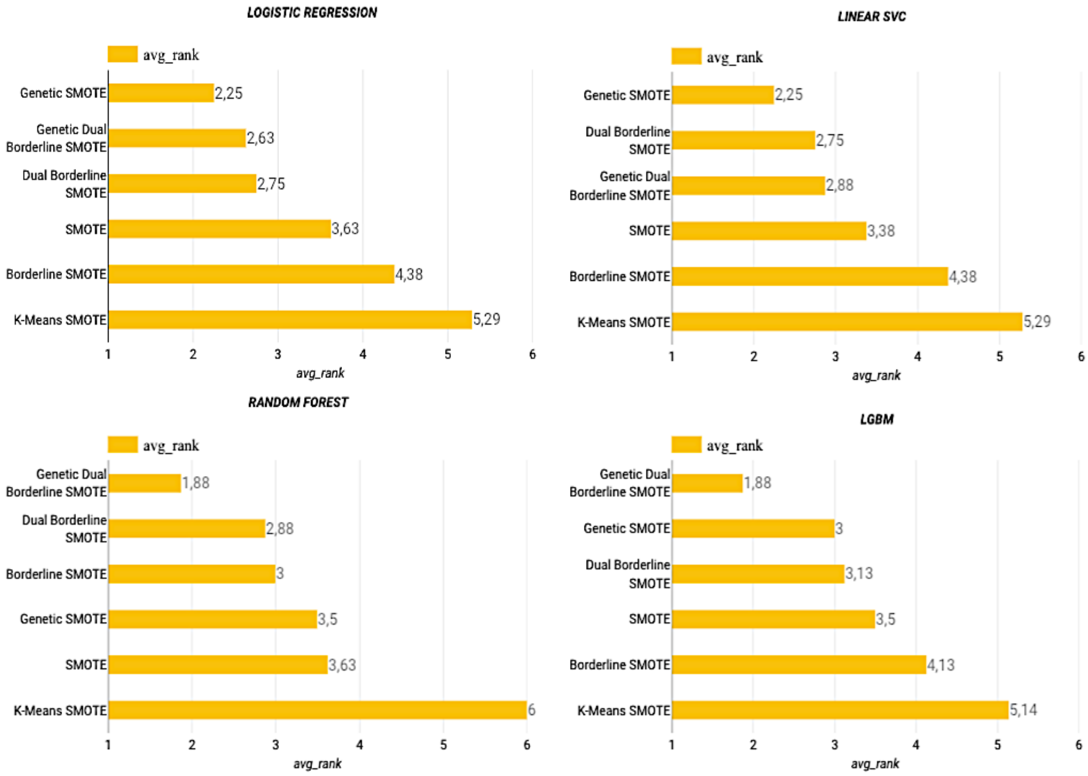

Figure 4.5 : Model based average rank values of SMOTE methods.

5. CONCLUSION

Imbalanced datasets can cause serious problems in classification problems. Having a small number of samples in the train set makes it difficult for the model to accurately predict this class in the test set, so many different methods have been developed to solve this problem.

Synthetic data generation is definitely the most popular method. The most popular methods are SMOTE, Borderline SMOTE, K-means SMOTE methods that can meet this need, but synthetic data production can actually be a very risky process. For example, in the smote method, synthetic data produced regardless of the majority class may actually cause noise.

In this study, three different synthetic sample generation methods were tested with various datasets and models. The Genetic SMOTE method is inspired by genetic algorithms. Unlike other SMOTE methods in the literature that generate synthetic data with new variable values between the sample and its neighbor, it generates synthetic data by crossing over the random feature values of two existing minority samples. Another method, Dual Borderline SMOTE, decides whether to generate synthetic data for a given sample and its neighbor by examining the ratio of minority class samples to their neighbors within the number of nearest neighbors in the training set. Here, the expectation is that the ratio for each point falls between 0 and 1. If the ratio for any of them is 0, it is assumed that the sample is in a dense minority area, and the new variable to be generated will not contribute. If it is 1, it is considered as noise, as all neighbors are expected to belong to the majority class, and generating variables from here is considered risky. The last proposed method, Genetic Dual Borderline SMOTE, as the name suggests, is a combination of two methods, that is, in the selection of two points for synthetic data production, Dual Borderline SMOTE rules apply, while in the production of synthetic data between these two points, variable change is taken as basis, as it is in Genetic SMOTE.

The study involved eight different imbalanced datasets, four different models, and six different SMOTE methods, including three proposed methods. Parameter combinations yielding the best F1 scores for SMOTE methods were preferred. During score calculation, the F1 score of SMOTE methods was subtracted from the F1 score each dataset and model pair would have achieved if no SMOTE method was applied. Averages were computed in this scoring, revealing that the proposed methods obtained a higher average score than the others.

Alongside the scores, the success of SMOTE methods for each dataset and model was ranked, and it was determined that the proposed methods outperformed other methods. Among the proposed methods, the Genetic Dual Borderline SMOTE method performed better than the Genetic SMOTE method, and the Genetic SMOTE method performed better than the Dual Borderline SMOTE method.

In future research, instead of selecting crossover variables entirely at random, different methods can be utilized to determine the importance of each variable in the training set. Only these significant variables can then be randomly selected for crossover between samples and their neighbors. Conducting variable exchange solely among the important variables may provide additional contributions to the model's learning process. In addition to this, in the method applied for selecting the minority sample and its neighbor, the defined SUITABLE zone can be regenerated with different ratios of the total number of minority examples among the k nearest neighbors, and their performances can be compared.

REFERENCES

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002) SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, 321-357.

Ganganwar, V. (2012) An overview of classification algorithms for imbalanced datasets. *International Journal of Emerging Technology and Advanced Engineering*, *2*(4), 42-47.

Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., & Bing, G. (2017) Learning from class-imbalanced data: Review of methods and applications. *Expert systems with applications*, *73*, 220-239.

Han, H., Wang, W. Y., & Mao, B. H. (2005) Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. In *International conference on intelligent computing* (pp. 878-887). Berlin, Heidelberg: Springer Berlin Heidelberg.

KrishnaVeni, C. V., & Sobha Rani, T. (2011) On the classification of imbalanced datasets. *IJCST*, *2*(SP1), 145-148.

Last, F., Douzas, G., & Bacao, F. (2017) Oversampling for Imbalanced Learning Based on K-Means and SMOTE.

CURRICULUM VITAE

Ad-Soyad : Hakan Korul

EDUCATIONAL STATUS:

- **Bachelor's Degree** : 2020, Istanbul Technical University, Management Faculty, Industrial Engineering

JOB EXPERIENCE:

- Yapı Kredi Bank / Senior Data Scientist (2023 – Current)
- Bitaksi / Data Scientist (2022 – 2023)
- Hepsiburada / Growth Data Analyst (2021 - 2022)
- LC Waikiki / International Store Merchandiser (2020 – 2021)