提升不平衡資料集之分類效能－應用不同採樣技術

沈宜君[1]，翁政雄[2]
[1]國立彰化師範大學資訊管理系
E-mail: sisisibibi@gmail.com
[2]國立彰化師範大學資訊管理系
E-mail: peteweng@gmail.com

摘要

本研究探討不同採樣技術對機器學習模型處理不平衡資料集的影響。實驗比較了 ROS、RUS、SMOTE、ADASYN 和伽瑪分佈等採樣方法，並結合 LR、RF、SVM 和 KNN 等分類模型，使用四個不平衡比例介於 10:1 至 15:1 的資料集進行測試。研究採用 Minority Recall、Majority Recall 及 F1-Score 評估模型在少數與多數類別的分類性能。實驗結果顯示：採樣方法皆能顯著改善模型對少數類別的識別能力，其中 RUS 在 Minority Recall 表現最佳，但會影響 Majority Recall；SMOTE 和 ADASYN 表現最穩定，既提升 Minority Recall 又維持理想的 Majority Recall。伽瑪分佈採樣則提供了介於兩者之間的平衡解決方案。在模型性能方面，RF 不僅在原始資料上表現較佳，結合採樣技術後更展現優異的穩定性。除此之外，本研究發現：沒有一種方法在所有情況下都表現最佳。在實際應用中，建議根據實際需求選擇合適的採樣方法和模型組合。

關鍵字：不平衡資料集;採樣方法;機器學習;羅吉斯回歸;支援向量機

Abstract

This study investigates the impact of various sampling techniques on machine learning models' performance in handling imbalanced datasets. The experiment compared sampling methods including ROS, RUS, SMOTE, ADASYN, and Gamma Distribution, combined with classification models such as LR, RF, SVM, and KNN. Four datasets with imbalance ratios between 10:1 and 15:1 were used for testing. The study employed Minority Recall, Majority Recall, and F1-Score to evaluate model performance on both minority and majority classes.

Experimental results show that all sampling methods significantly improved models' ability to identify minority classes. RUS achieved the best Minority Recall performance but affected Majority Recall; SMOTE and ADASYN demonstrated the most stable performance, enhancing Minority Recall while maintaining satisfactory Majority Recall. Gamma Distribution sampling provided a balanced solution between these approaches. Among the models, RF not only performed better on original data but also showed excellent stability when combined with sampling techniques. We found that no single method performed best in all scenarios. For practical applications, it is recommended to select appropriate sampling methods and model combinations based on specific requirements.

Keywords *：Imbalanced data, Sampling, Machine Learning, Logistic Regression, SVM*

1. 簡介

機器學習和資料分析領域中，資料不平衡 (Imbalanced data) 是某些類別的樣本數量顯著多於其他類別，此情況會對模型的學習效果和判斷造成顯著影響，為常見且具挑戰性的問題，如醫學診斷[17]、詐欺檢測[12]等，都涉及資料不平衡。在這些領域中，少數類別通常是我們最關心的，但傳統的機器學習方法往往在處理這類資料時表現不佳。

本研究旨在比較不同採樣技術對不平衡資料集的機器學習模型性能的影響。另外，為更具針對性的評估模型表現，引入 Majority Recall 及 Minority Recall 作為新的評估指標。

2. 文獻探討

2.1 不平衡資料集

處理不平衡資料時需面臨各種挑戰，包括少數類別的重要性、資料預處理、算法調整以及混合方法的應用。處理代表性不足的資料和類別分布時，傳統機器學習需要進行適當的修改和調整，以確保在實際應用中能有效處理不平衡問題來提高模型的整體準確性[8]。

2.2 採樣方法

傳統的分類器往往偏向於預測多數類別，雖然可以提升整體準確率，但在實際應用中可能難以有效處理少數類別。為了解決這個問題，可以透過採樣方法來平衡資料類別，從而提升分類器的性能。針對資料不平衡問題，採樣方法被廣泛研究和應用。

比較過採樣 (Oversampling) 和欠採樣 (Undersampling)的優劣，相關研究顯示：增加少數類別樣本比減少多數類別樣本更有效[13]。使用 SMOTE 透過在少數類別樣本間進行插值來生成新樣本處理不平衡資料，能顯著改善少數類別的預測能力，尤其是 RF 結合 SVM SMOTE [2]。此外，基於伽瑪分佈的過採樣方法，在多數資料集表現優異，尤其在 F1-Score 及 AUC 指標上[11]。

然而，採樣的時間點也是影響模型效能的關鍵。不恰當的時間點可能導致資料外洩 (data leakage)，這可能導致過於樂觀的結果[16]，通常為在資料預處理階段或模型訓練過程中不當的處理資料。例如，在分割資料之前進行縮放或正規化，可能將測試集中的資訊外洩至訓練集中，因此需注意此風險，以確保模型的準確性和可靠性。

2.3 評估指標

準確率 (Accuracy) 是常用的衡量標準，但對於不平衡資料集，準確率不再是合適的衡量標準，因為與多數類別相比，少數類別對準確率的影響很小[15]。因此，本研究採用 F1-Score、Recall 等作為分類效能評估指標。

3. 模型及方法

3.1 機器學習模型

羅吉斯迴歸 (Logistic Regression, LR) 用於預測二元結果的機率。它將資料擬合至一個邏輯函數，可以預測諸如成功/失敗、是/否、1/0 等事件發生的機率。LR 廣泛應用於醫療、商業、教育等領域[10][14]。隨機森林 (Random Forest, RF) 是一種集成學習方法，常用於分類和迴歸任務。RF 由多棵決策樹 (decision tree) 組成[4]，其原理是隨機採樣資料集中的多個樣本來訓練多個獨立的決策樹。每個決策樹對新資料進行預測，最終通過對所有決策樹的預測結果進行多數表決來得出預測結果，從而提升了分類器的準確性和穩定性。

支持向量機 (Support Vector Machine, SVM) 的目標是尋找一個超平面來劃分資料中不同類別的資料[3]。本研究使用線性可分 SVM，旨在找到一個能完全分離兩個類別並使間隔最大化的超平面。K 近鄰 (K Nearest Neighbor, KNN) 是一種基本且廣泛使用的監督學習方法[6]。在 KNN 演算法中，物件的分類是根據其最近鄰的 K 個樣本的類別來決定，類似「物以類聚」的原理。其運作方式是計算未知資料點與已知資料點的距離，找出最近的 K 個鄰居，並根據這 K 個鄰居的多數類別來決定新點的類別。

3.2 採樣方法

3.2.1 隨機過採樣 (Random Oversampling, ROS)

透過隨機重複少數類別的樣本來平衡資料集，從而達到與多數類別相似的樣本量，優點為操作簡單，缺點為可能導致模型過擬合[2]。

3.2.2 隨機欠採樣 (Random Undersampling, RUS)

與 ROS 相反。此方法透過隨機刪除減少多數類別樣本來平衡資料集，其優點為操作簡單且速度提升，缺點為可能丟失重要資訊，樣本數的減少也可能導致模型的泛化能力不足[2]。

3.2.3 SMOTE (Synthetic Minority Over-sampling Technique, SMOTE)

SMOTE 透過在少數類樣本進行線性插值來生成新樣本[5]。這種方法旨在增加少數類樣本的數量，從而達到類別平衡。核心步驟如下：

- 選擇一少數類別樣本 X_{chosen}。
- 找到該樣本在少數類別中的 k 個最近鄰樣本。$X_{nearest}$ 為 k 個最近鄰中的一個。
- 生成新樣本：透過在 X_{chosen} 和 $X_{nearest}$ 之間線性插值來生成新樣本。公式如下：

$$X_{new} = X_{chosen} + (X_{nearest} - X_{chosen}) * \delta;\ \delta \in [0.1] \quad (1)$$

其優勢在於能透過創建合成樣本來生成新資料點，有助於緩解模型過度擬合的問題。

3.2.4 ADASYN (Adaptive Synthetic Sampling Approach, ADASYN)

基於 SMOTE 的改進方法[9]。生成新樣本時考慮了樣本的分佈密度，尤其針對與多數類樣本重疊較多的樣本來生成更多的合成樣本，且可以自適應地決定每個少數類樣本需要生成的新樣本數量。

3.2.5 伽瑪分佈 (Gamma Distribution Sampling)

伽瑪分佈採樣是統計學和各種科學領域中廣泛使用的連續概率分布，亦可作為採樣方法來處理不平衡資料集，研究證明其優於其他方法[11]。優點為合成新樣本時能夠引入更多的變異，從而提高模型的泛化能力，且可根據不同的需求調整標準差和均值，以模擬不同的資料分布特徵。然而，參數選擇不當可能導致生成的資料與實際分布差異較大，導致模型性能下降。核心步驟如下：

$$f(x;\alpha,\theta) = \frac{x^{\alpha-1}+e^{-\frac{x}{\theta}}}{\Gamma(\alpha)\theta^{\alpha}} \quad (2)$$

- 選擇 α 、θ 參數。
- 計算分佈最大值的座標

$$X = \theta(\alpha - 1) \quad (3)$$

- 選擇 n 個少數點

(n =多數類別數量-少數類別數量)

- 對於每個被選到的鄰居點 p：

從k個鄰居中選一點p'，並定義向量：

$$v = p' - p \quad (4)$$

- 使用伽瑪分佈生成值t，定義新少數類別：

$$q = p + (t - n) - v \quad (5)$$

4. 實驗

4.1 資料集說明

本研究使用 Imblearn.dataset 提供的不平衡資料集，其資料集也被其他學者用於研究資料不平衡問題[11]。如表 1 所示，本研究擷取其中 4 項來自不同領域的資料集 (包含汽車、犯罪率、手勢、醫療) 作為實驗資料，其不平衡程度落在 10:1 至 15:1。欄位 Ratio 為不平衡程度、#S 為資料筆數、#F 為特徵欄位數。

表 1. 資料集項目集合

ID	Dataset	Repository	Ratio	#S	#F
1	car_eval_34	UCI	12:1	1,728	21
2	us_crime	UCI	12:1	1,994	100
3	libras_move	UCI	14:1	360	90
4	thyroid_sick	UCI	15:1	3,772	53

4.2 評估指標

我們使用 F1-Score、Minority Recall、Majority Recall 作為效能評估指標。F1-Score 為 Precision 和 Recall 的加權調和平均，用以評估模型的綜合性能表現。

$$Precision = \frac{TP}{TP+FP} \quad (6)$$

$$Recall = \frac{TP}{TP+FN} \quad (7)$$

$$F1-Score = \frac{2*Precision*Recall}{Precision+Recall} \quad (8)$$

其中，

TP: True Positive; FP: False Positive;

TN: True Negative; FN: False Negative.

Minority Recall 與 Majority Recall 是本研究專用的評估指標，源自傳統的 Recall 概念。不同之處為 Minority Recall 專注於評估模型對少數類別的識別能力。它計算在所有實際屬於少數類別的樣本中，被正確預測為少數類別的比例。此指標更針對性的反映模型在不平衡資料集中對少數類別的分類性能；Majority Recall 反之。

$$Minority\ Recall = Minority(\frac{TP}{TP+FN}) \quad (9)$$

$$Majority\ Recall = Majority(\frac{TP}{TP+FN}) \quad (10)$$

4.3 實驗流程

圖 1 為本研究實驗流程。首先進行資料清洗(data cleaning)，移除不正確、不完整及重複的資料。接著進行採樣 (Sampling) 用以平衡資料類別。最後採用 MinMaxScaler 進行標準化，此法不僅實現了特徵縮放也具有正規化 (Normalization) 的效果。

在此使用 K-折交叉驗證 (K-Fold Cross Validation) 評估模型性能。我們對每種模型和採樣方法的組合，重複上述過程並記錄評估指標。此方法確保了結果的可靠性和穩定性。此處採用 K=5。

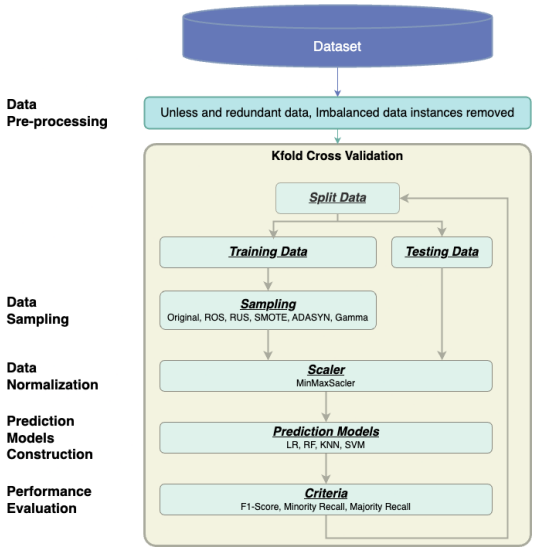

圖 1 實驗流程圖

4.4 分類性能分析

由表 2 至表 4 的實驗結果顯示：未經採樣處理前，各模型均呈現出對多數類別的偏好傾向，其 Majority Recall 大多維持在 0.98 以上。然而，這也反映了模型在不平衡資料集上的偏差問題。某些極端案例甚至出現完全無法識別少數類別的情況。此現象凸顯了在不平衡資料集上，模型容易被多數類別主導，導致分類決策偏向樣本量較大的類別。

使用採樣技術後，在類別平衡性有顯著改善。特別是 SMOTE 和 ADASYN 這類過採樣方法，展現出最穩定的效果，能在提升 Minority Recall 至 0.8 以上的同時，仍維持較佳的 Majority Recall，也發現 SMOTE、ADASYN 分數大致相同。相較之下，RUS 雖然能有效提升 Minority Recall，但往往導致 Majority Recall 顯著下降，進而影響 F1-Score。

表 2. 各採樣方法 F1-Score 分類性能

Performance based on F1-Score

		Sampling					
Dataset	Model	Non	SMOTE	ADASYN	RUS	ROS	Gamma
1	LR	**0.855**	0.835	0.827	0.638	0.827	0.603
	RF	0.880	0.910	0.911	0.738	**0.914**	0.893
	SVM	0.871	**0.913**	**0.913**	0.678	**0.913**	0.602
	KNN	0.097	0.527	0.546	0.537	0.097	**0.612**
2	LR	**0.494**	0.480	0.459	0.455	0.479	0.456
	RF	0.464	0.484	0.489	0.447	0.458	**0.495**
	SVM	**0.510**	0.471	0.450	0.438	0.445	0.449
	KNN	0.416	**0.440**	0.430	0.405	0.423	0.402
3	LR	0.553	**0.745**	0.728	0.400	0.729	0.720
	RF	0.567	**0.839**	0.823	0.409	0.683	0.790
	SVM	**0.813**	0.753	0.717	0.489	0.697	0.648
	KNN	**0.813**	0.760	0.760	0.491	0.758	0.694
4	LR	0.154	**0.451**	0.449	0.327	0.446	0.323
	RF	0.838	0.868	**0.875**	0.710	**0.875**	0.805
	SVM	0.000	**0.486**	0.481	0.342	0.483	0.330
	KNN	0.500	**0.529**	0.516	0.415	0.522	0.440

表 3. 各採樣方法 Minority Recall 分類性能

Performance based on Minority Recall

		Sampling					
Dataset	Model	Non	SMOTE	ADASYN	RUS	ROS	Gamma
1	LR	0.777	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
	RF	0.843	0.903	0.918	**1.000**	0.940	0.895
	SVM	0.851	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
	KNN	0.052	0.888	0.873	**0.926**	0.052	0.820
2	LR	0.393	0.780	0.793	**0.873**	0.800	0.820
	RF	0.360	0.493	0.527	**0.867**	0.380	0.580
	SVM	0.393	0.767	0.780	**0.860**	0.760	0.800
	KNN	0.327	0.713	0.707	**0.820**	0.473	0.707
3	LR	0.410	0.840	0.840	0.910	0.880	**0.920**
	RF	0.420	0.760	0.760	**0.880**	0.550	0.750
	SVM	0.710	0.840	0.840	**0.960**	0.840	0.880
	KNN	0.710	0.870	0.870	**1.000**	0.750	0.790
4	LR	0.087	0.883	0.879	0.853	0.879	**0.909**
	RF	0.776	0.840	0.844	**0.948**	0.844	0.749
	SVM	0.000	0.874	0.861	0.879	0.874	**0.918**
	KNN	0.381	0.602	0.606	**0.866**	0.580	0.490

表 4. 各採樣方法 Majority Recall 分類性能

Performance based on Majority Recall

		Sampling					
Dataset	Model	Non	SMOTE	ADASYN	RUS	ROS	Gamma
1	LR	**0.997**	0.967	0.965	0.904	0.965	0.889
	RF	**0.994**	0.993	0.992	0.940	0.990	0.991
	SVM	**0.991**	0.984	0.984	0.919	0.984	0.888
	KNN	**1.000**	0.875	0.888	0.871	**1.000**	0.925
2	LR	**0.984**	0.879	0.863	0.838	0.872	0.853
	RF	**0.985**	0.957	0.951	0.834	0.979	0.941
	SVM	**0.988**	0.878	0.861	0.831	0.863	0.855
	KNN	**0.980**	0.875	0.870	0.817	0.939	0.854
3	LR	**1.000**	0.967	0.958	0.806	0.952	0.949
	RF	**1.000**	0.997	0.994	0.818	0.997	0.988
	SVM	**1.000**	0.967	0.961	0.848	0.952	0.934
	KNN	**1.000**	0.967	0.967	0.827	0.982	0.964
4	LR	**0.998**	0.867	0.867	0.780	0.865	0.757
	RF	**0.996**	0.994	0.994	0.953	0.994	0.993
	SVM	**1.000**	0.887	0.888	0.785	0.886	0.762
	KNN	**0.991**	0.957	0.952	0.848	0.958	0.952

在模型選擇上，RF 即使在未採樣的情況下也能維持相對較高的 Minority Recall，而在搭配採樣技術後，其性能提升更為顯著。此結果顯示：處理不平衡資料時，採樣技術的選擇與模型的選用同等重要，兩者的適當配合能有效改善分類器在各類別上的平衡表現。

4.5 可視化分析

圖 2 展示 Minority Recall 和 F1-Score 的中位數表現。在 Minority Recall 效能中，可以觀察到所有模型在未採樣的情況下性能較差。經過 SMOTE 和 ADASYN 採樣後，各模型的性能都有明顯提升，維持在 0.75 至 0.85 之間。當使用 RUS 時，所有模型都達到最高的 Minority Recall，約在 0.90 左右。然而，在 ROS 中，除了 LR 和 SVM 能維持高 Recall 外，RF 和 KNN 的性能顯著下降。在 F1-Score 效能中，RF 整體性能最為出色，特別是在 SMOTE 和 ADASYN 採樣下，F1-Score 達到約 0.85 的最高點。值得注意的是，雖然 RUS 能提高 Minority Recall，但從 F1-Score 分數來看，所有模型的整體性能反而下降，這表示在提升少數類別辨識能力的同時，可能犧牲了對多數類別的準確度。KNN 在各種採樣方法下的 F1-Score 相對較低，顯示其在處理不平衡資料集時可能不是最佳選擇。

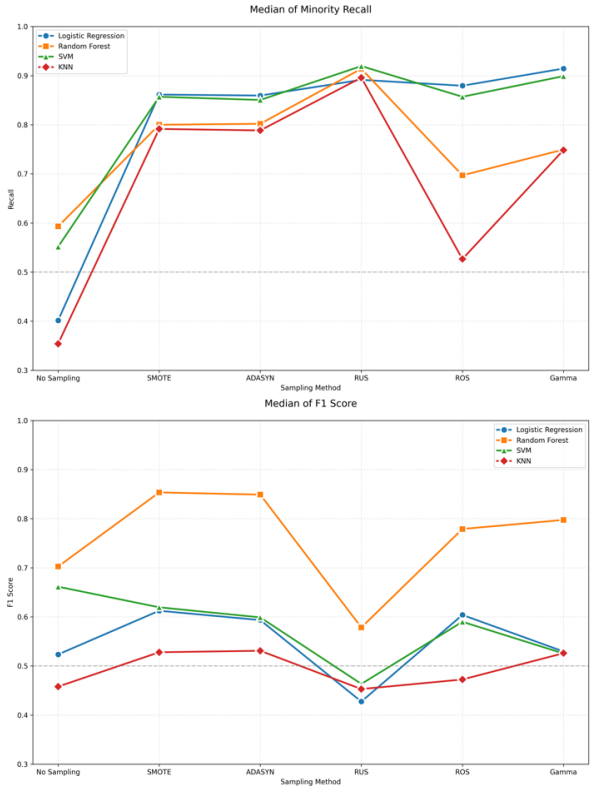

圖 2 不同採樣方法與模型的折線圖

5. 結論

本研究探討了不同採樣方法對處理不平衡資料集的影響，並評估了這些方法與多種機器學習模型的組合效果，我們得出：

1. 採樣技術的影響

   所有的採樣方法均顯著的提升了分類性能，尤其是在少數類別的分類能力。

2. RUS 的欠採樣挑戰

   雖然 RUS 在提升 Minority Recall 方面表現突出，但會犧牲 Majority Recall ，可能導致整體 F1-Score 下降。因此，在使用 RUS 時需謹慎考量其可能造成的資訊流失，RUS 是此研究中唯一的欠採樣方法，本研究認為，大量降低多數類別的特徵，會導致辨識多數類別的重要特徵丟失，導致其在 Majority Recall 表現不佳。

3. 模型選擇的重要性

   RF 在所有採樣方法下均展現出穩定的性能，特別是在未經採樣的情況下仍能維持較高的 Minority Recall。相對而言，KNN 在處理不平衡資料集時效果不佳，無論採用何種採樣方法，其 F1-Score 均顯著低於其他模型。

不同的採樣方法和模型組合在各種情況下性能各異，沒有一種方法在所有情況下都表現最佳。在實際應用中，需要根據具體需求和資料集性質來選擇適當的方法和模型組合。未來研究可透過不同資料不平衡比探索不同的採樣及模型分類效能。

參考文獻

[1] 傅國瑋(2020)。SMOTE 演算法和集成學習在不平衡資料預測之應用。〔碩士論文。國立臺北大學〕**臺灣博碩士論文知識加值系統**。

[2] Batista, G. E. A. P. A., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. SIGKDD Explor. Newsl., 6(1), 20–29.

[3] Boser, B., Guyon, I., & Vapnik, V. (1996). A Training Algorithm for Optimal Margin Classifier. Proceedings of the Fifth Annual ACM Workshop on Computational Learning Theory, 5.

[4] Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5–32.

[5] 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.

[6] Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. IEEE Transactions on Information Theory, 13(1), 21–27.

[7] Cox, D. R. (1958). The Regression Analysis of Binary Sequences. Journal of the Royal Statistical Society. Series B (Methodological), 20(2), 215–242.

[8] Haibo He, & Garcia, E. A. (2009). Learning from Imbalanced Data. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263–1284.

[9] He, H., Bai, Y., Garcia, E. A., & Li, S. (2008). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), 1322–1328.

[10] Jain, H., Khunteta, A., & Srivastava, S. (2020). Churn Prediction in Telecommunication using Logistic Regression and Logit Boost. Procedia Computer Science, 167, 101–112.

[11] Kamalov, F., & Denisov, D. (2020). Gamma distribution-based sampling for imbalanced data. Knowledge-Based Systems, 207, 106368.

[12] Lopo, J. A., & Hartomo, K. D. (2023). Evaluating Sampling Techniques for Healthcare Insurance Fraud Detection in Imbalanced Dataset. Jurnal Ilmiah Teknik Elektro Komputer Dan Informatika (JITEKI), 9(2), Article 2.

[13] Mohammed, R., Rawashdeh, J., & Abdullah, M. (2020). Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results. 2020 11th International Conference on Information and Communication Systems (ICICS), 243–248.

[14] Singh, H. P., & Alhulail, H. N. (2022). Predicting Student-Teachers Dropout Risk and Early Identification: A Four-Step Logistic Regression Approach. IEEE Access, 10, 6470–6482. IEEE Access.

[15] Sun, Y., Wong, A. K. C., & Kamel, M. S. (2009). CLASSIFICATION OF IMBALANCED DATA: A REVIEW. International Journal of Pattern Recognition and Artificial Intelligence, 23(04), 687–719.

[16] Zheng, A., & Casari, A. (2018). Feature Engineering for Machine Learning (1st ed.). O'Reilly Media, Inc.

[17] Zhu, C., Idemudia, C. U., & Feng, W. (2019). Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques. Informatics in Medicine Unlocked, 17, 100179.