西南交通大学学报

第 58 卷 第 6 期
2023 年 12 月

JOURNAL OF SOUTHWEST JIAOTONG UNIVERSITY

Vol. 58 No. 6
Dec. 2023

ISSN: 0258-2724 DOI：10.35741/issn.0258-2724.58.6.21

Research article

Computer and Information Science

ENHANCING IMBALANCED DATA CLASSIFICATION: A CASE STUDY OF PORTUGUESE BANK MARKETING

加强不平衡数据分类：葡萄牙银行营销案例研究

Mahmoud Rajallah Asassfeh [a], Mohammad Rasmi [a], Abdullah Alqammaz [a,*], Ahmad Bany Doumi [b], Khaled Al-Qawasmi [a], Ala'a Al-Shaikh [a]

[a] Department of Cyber Security, Faculty of Information Technology, Zarqa University
Zarqa, Jordan, masassfeh@zu.edu.jo, mmousa@zu.edu.jo, a.qammaz@zu.edu.jo, kqaqasmi@zu.edu.jo, ashaikh@zu.edu.jo

[b] Department of Computer Science, Faculty of Computer Science and Information Technology, Jerash University
Jerash, Jordan, a.banydoumi@jpu.edu.jo

Received: October 23, 2023 ▪ Review: November 27, 2023
▪ Accepted: December 17, 2023 ▪ Published: December 29, 2023

This article is an open-access article distributed under the terms and conditions of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0)

Abstract

In classification tasks, it is presumed that the number of classes of observations is balanced. While classification models usually give a heavily biased weight to the class that has higher occurrence, building an efficient classification model is likely a challenge when feeding it with imbalanced dataset observations or samples of data. This study introduces a new method for addressing imbalanced datasets in classification tasks, particularly focusing on predicting long-term deposits in banking institutions. The method involves systematic evaluation and comparison of random oversampling (ROS) and synthetic minority over-sampling technique (SMOTE) while employing meticulous feature selection to optimize classification precision. This new methodology showcased competitive performance, notably achieving an accuracy of 89.1% and a G-mean of 0.61 with SMOTE at a 500% ratio encompassing all features in Experiment 2 and an accuracy of 87.2% and a G-mean of 0.677 with ROS at a 500% ratio using the top 15 features in Experiment 3.

Keywords: Synthetic Minority Oversampling Technique, Random Oversampling, Imbalanced Data, Feature Selection, Random Forest

摘要 在分类任务中，假设观察类别的数量是平衡的。虽然分类模型通常会给出现率较高的类别赋予严重偏差的权重，但在向其提供不平衡的数据集观察或数据样本时，构建有效的分类模型可能是一个挑战。本研究引入了一种解决分类任务中不平衡数据集的新方法，特别关注预测银行机构的长期存款。该方法涉及随机过采样（活性氧）和合成少数过采样技术（斯莫特）的系统评估和比较，同时采用细致的特征选择来优化分类精度。这种新方法展示了具有竞争力的性能，特别是

在包含实验2中所有特征的500%比率下，斯莫特实现了89.1%的准确度和0.61的G均值，在活性氧为使用实验3中前15个特征的500%比率。

关键词: 合成少数过采样技术、随机过采样、不平衡数据、特征选择、随机森林

I. INTRODUCTION

Marketing selling campaigns are a standard method to enhance business. These campaigns target a segment of customers by contacting them through various channels such as fixed phone, mobile phone, or email. The dataset used in this paper was collected from bank marketing campaigns of Portuguese banking institutions by contacting their clients to determine if their product (i.e., bank term deposit) would be (yes) or (no). Due to the high volume of bank marketing data and imbalanced data problem, it is necessary to investigate an efficient and accurate classification model to support the data analysis process effectively. Consequently, this study proposed a random forest-based classification model to predict the anticipation in long-term deposits and handle the imbalanced data problem. In essence, the contribution of this study lies in a multifaceted investigation. First, this study delves into the appraisal of the random forest classifier (RFC) as a formidable tool for feature and predicting bank term deposits. Second, our inquiry extends to the assessment of oversampling techniques and their influence on elevating both the stability and performance of the classifier within the imbalance problem. Imbalanced data relate to categorization difficulties in which one class significantly surpasses the other. Binary classification is more prone to imbalance than multi-level classification [13]. Extreme imbalance data, for example, can be found in banking or financial data, such as our bank marketing dataset. An algorithm cannot obtain the information required to make an accurate prediction about the minority class from an imbalanced dataset. As a result, it is advised to employ a balanced classification dataset; most real-world classification issues exhibit some level of imbalance in classes, which occurs when each class does not include an equal amount of dataset. It is critical to correctly alter a developed approach to match the tasks' objectives. Otherwise, an unrobust outcome may be delivered [11].

To be clearer, suppose a dataset has two classes, A and B. The occurrence of class A is approximately 90% of the dataset, whereas the occurrence of class B is approximately 10%. However, the target prediction is mainly interested in detecting class B. In this case, when a model is trained, high accuracy results may be obtained for class A but not for class B. Instead, a correctly calibrated approach would have a lower accuracy but a significantly greater true positive rate (or recall), which indicates a robust classification performance. Such cases often occur, especially due to harmful content on the Internet and/or within complex data collection procedures, e.g., medical data. At this point, there are a few strategies that deal with class imbalance where some strategies are appropriate for most classification issues, while others may be better suited to certain levels of imbalanced data case. For the purposes of this study, these are described within the terms of binary classification. In addition, identifying the minority class is also presumed.

The primary objectives of this study revolve around mitigating the challenges posed by imbalanced datasets in the context of predicting long-term deposits within banking institutions. Specifically, this study aims to systematically evaluate and compare the effectiveness of two prominent re-sampling techniques, random oversampling (ROS) and synthetic minority over-sampling technique (SMOTE), in enhancing classification model performance. Moreover, this study seeks to conduct meticulous feature selection alongside various re-sampling ratios to discern the optimal combination that maximizes classification precision. Indeed, the selection of the prediction of long-term deposits in banking institutions as the focal point of this study stems from the prevalence of imbalanced datasets in this domain, presenting practical implications. The findings from this research bear significance in real-world applications within banking and finance, offering potential strategies to refine classification models and inform decision-making processes, particularly where imbalanced datasets are pervasive.

II. METHODOLOGY

The methodology of this study is based on five major stages: data understanding and preprocessing, feature selection, modeling, evaluation, and analysis. Figure 1 illustrates the methodology and demonstrates the sequencing of the main stages.

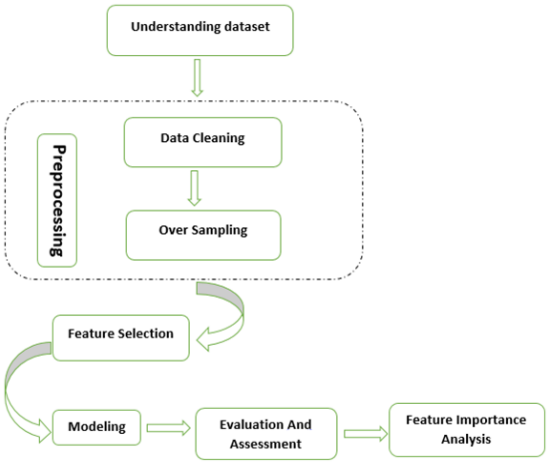

Figure 1. Proposed methodology (The authors)

The following subsections describe the main stages of this methodology:

A. Understanding the Dataset

This study used a freely available dataset, namely the Portuguese Bank Marketing Dataset (PBMD), which resulted from a bank marketing campaign conducted by a Portuguese banking institution. The data is collected based on phone calls where each client is contacted more than once to determine whether the product (bank term deposit) is (yes) or (no). The target of the data is to predict whether the client would anticipate or not (yes, no) a term deposit [1], [3].

This dataset is the result of Portuguese financial institutions' phone marketing operations; they contact the client more than once to determine whether the product (bank term deposit) is (yes) or (no). more details on the dataset are described in the following section.

B. Data Prepossessing

1) Data Cleaning

Data cleaning is an essential step that handles unknown values and incorrect inputs. The classification model is adversely affected by error input and/or an incomplete record of the data, which affect its performance. This study performed the following data cleaning process:

- Eliminating not available data samples in each attribute;
- As recommended by the author of the dataset, the "feature duration" attribute is discarded from the dataset.

C. Data Re-sampling

For data resampling, this study uses random oversampling and SMOTE techniques. Random oversampling (ROSE) is used to produce synthetic data based on the bootstrap approach. It handles both categorical and continuous data by producing artificial examples from conditional density estimates of the two classes.

In brief, the ROSE package has different types of sampling methods: oversampling, undersampling, and bothsampling. The purpose of the oversampling technique (Figure 2) is to oversample the minority class. The purpose of the undersampling technique (Figure 3) is to undersample the majority class. The bothsampling technique combines oversampling and undersampling techniques. The majority class is undersampled without replacement, whereas the minority class is oversampled with replacement. Indeed, ROS sampling generates data artificially and delivers a more accurate estimate of the original data. Consequently, to balance the dataset in our experiment, we used the oversampling method [6].

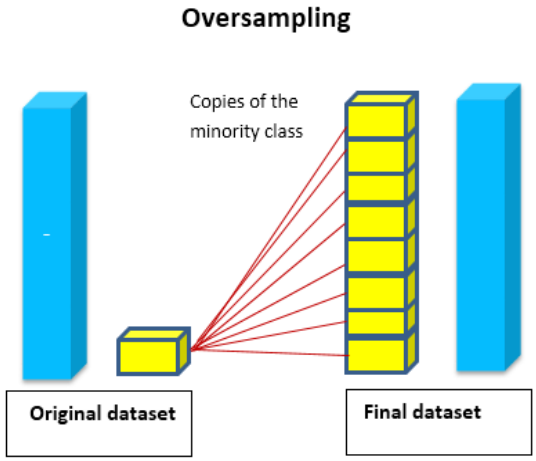

Figure 2. Oversampling method (The authors)

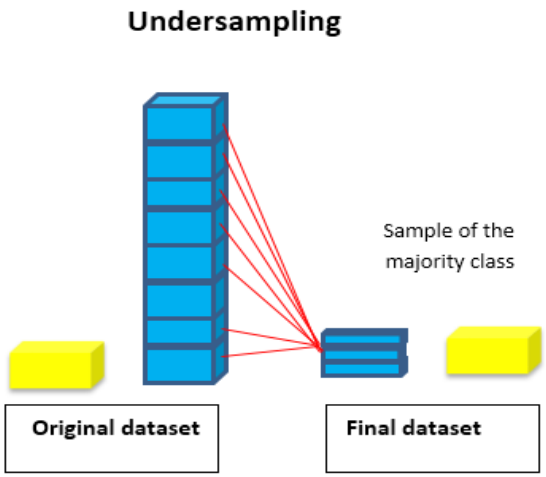

Figure 3. Undersampling method (The authors)

On the other hand, when adding precise clones of minority instances to the main dataset, the SMOTE sampling technique is used to avoid overfitting. SMOTE generates new artificially similar instances by creating convex combinations of surrounding instances, which generates new instances of the minority class. As

illustrated in Figure 4, it creates lines in the feature space between minority points and performs sampling along these lines. This provides a balanced dataset with less overfitting to some extent [2].

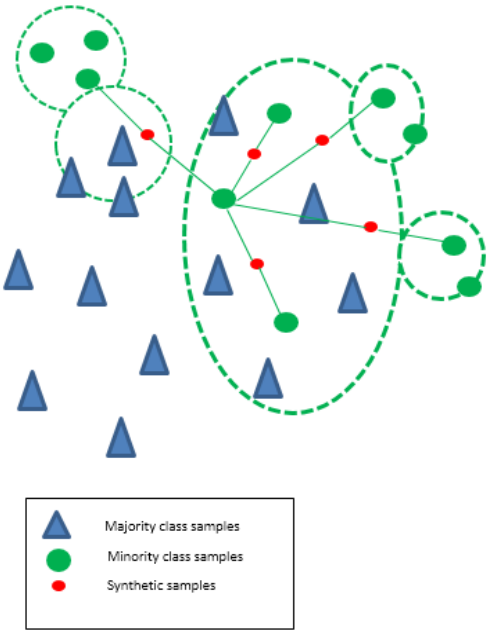

Figure 4. SMOTE method (The authors)

D. Feature Selection

Feature selection is a crucial step in data mining because it helps eliminate irrelevant inputs [16]. In addition, it has several advantages, including simplifying model interpretation, mitigating overfitting, and reducing computational costs and training time [10], [16], [18]. When it comes to assessing feature relevance, one of the most renowned machine learning algorithms is the random forest [7], [17]. The random forest algorithm employs two key techniques: mean decrease accuracy and mean decrease impurity [12].

In the mean decrease impurity approach, random forest comprises numerous decision trees, each focusing on a single feature to divide the dataset into two sets where responses with the same values end up together. The criterion for choosing which feature to divide by is known as impurity. During the tree training process, the algorithm evaluates how each feature contributes to the reduction of impurities within a tree. This process is averaged, and the features are ranked accordingly.

E. Modeling

This study leverages the RFC, which is a well-established and widely adopted tool in the realm of predictive analytics. The selection of the RFC is underpinned by its notable prevalence in prior literature, which confirms its efficacy and utility for solving similar prediction tasks. The allure of this classifier lies in its inherent simplicity and user-friendliness, making it an accessible choice for researchers and practitioners alike. The primary objective of this modeling phase is to discern the optimal prediction methodology that excels in terms of generalization performance [14]. By deploying the RFC, this study aims to harness its powerful ensemble of decision trees to construct a robust predictive model for anticipating long-term deposits, thus contributing to the advancement of predictive modeling techniques in this domain.

F. Evaluation

This study employs the most prevalent assessment metric in the literature to assess the performance of the constructed classification model, including classification accuracy, sensitivity, specificity, and G-means metrics. Moreover, the confusion matrix serves as a foundation for these measurements. The purpose of this stage is to identify which classification model performs better in predicting tasks on the dataset used in this study. The following are the evaluation metrics equations [9], [15]:

$$\text{Accuracy} = (TP+TN=TP+FP+FN+TN)\_100\% \quad (1)$$

$$\text{Sensitivity} = (TP=TP+FN)\_100\% \quad (2)$$

$$\text{Specificity} = (TN=FP+TN)\_100\% \quad (3)$$

$$\text{Precision} = (TP=TP+FP)\_100\% \quad (4)$$

In Equations 1-4, TP represents the number of true positives, FN represents the number of false negatives, TN represents the number of true negatives, and FP represents the number of false positives. As illustrated in Figure 5, they are defined in the confusion matrix. The G-mean is another relevant number for measuring the classification accuracy of imbalanced positive and negative samples.

$$G-mean = \frac{\sqrt{\text{sensitivity}\times\text{specifity}}}{1} \quad (5)$$

		Actual Class	
		Negative	**Positive**
Predicted Class	**Negative**	True Negative (TN)	False Negative (FN)
	Positive	False Positive (FP)	True Positive (TP)

Figure 5. Confusion matrix (The authors)

Obviously, the value of the G-mean indicates the performance of the classifier, especially in the imbalanced dataset. The larger the G-mean value, the better the classifier performance.

G. Experimental Results

This section goes over the experiments conducted using the methodology stated in Figure 1 and bank marketing datasets. In the experiments, this study considers the following scenarios; this study explores three distinct experimental scenarios to assess the impact of oversampling and feature selection on the RFC applied for predictive modeling:

- *Scenario 1:* The RFC classifier is initially applied without oversampling, using the input variables, as delineated in Table 1. Subsequently, a feature selection algorithm is employed. This scenario aims to discern the effect of feature selection on the RFC performance.
- *Scenario 2:* In this scenario, SMOTE is employed as the first step to balance class distributions. Following oversampling, a feature selection algorithm is applied to order the features based on their importance. Finally, the RFC is executed. This scenario was designed to evaluate the potential enhancements achieved by incorporating SMOTE oversampling in the model.
- *Scenario 3:* Similarly to Scenario 2, this scenario employs the random over-sampling (ROS) technique as the initial step, followed by feature selection and ultimately, the RFC. The specific aim of this scenario is to determine the effectiveness of ROS oversampling in improving model outcomes.

Table 1.
Attributes of the bank marketing dataset [4], [8]

	Attribute name	Description	Type
1	Age	Age of the client	N
2	Job	Type of the client's job	C
3	Marital	The client's status.	C
4	Education	What is the highest level of education attained by the client?	C
5	Default	Does the client have credit?	C
6	Housing	Is the client in possession of a housing loan?	C
7	Loan	Is the client in possession of a personal loan?	C
8	Contact	What is the client contact type?	C
9	Month	What is the last month of the year of the contract with the client?	C
10	Day of the Week	What is the last day of the week of the contract with the client?	C
11	Duration	How long does it communicate with the client?	N
12	Campaign	Count of contacts made during this campaign and for this client	N
13	Pday	The number of days since the client was last reached by a previous campaign.	N
14	Previous	Number of contacts made prior to this campaign and for this client	N
15	Poutcome	Result of the preceding marketing campaign	C
16	Emp.var.rate	Employment variation rate	N
17	Cos.price.idx	Consumer price index	N
18	Cons.conf.idx	Consumer confidence index	N
19	Euribor3m	Euribor 3 month rate	N
20	Nr.employed	Number of employees	N
21	Label	Has the client made a term deposit?	C

Notes: N - numeric, C - categorical

H. Environmental Settings

All experiments were performed in a Windows operating system using the R programing language and an Intel R Core TM i5-4200U CPU@ 1.6 GHz 2.30 GHz processor. Each data mining (DM) model underwent ten iterations for robustness and consistency. The R package, a comprehensive suite of software tools for data management, computation, and graphical representation, plays a pivotal role in this analysis. This multifaceted package offers the following key attributes [5]:

- *Efficiency in data handling and storage:* The R package excels in proficiently managing and storing data, ensuring streamlined data manipulation.
- *Array operations:* This provides a set of operators tailored for array operations, with a particular focus on matrices, facilitating complex calculations.
- *Rich toolkit for data analysis:* This substantial collection of intermediary tools caters to diverse data analysis requirements, enhancing the versatility of the R package.
- *Graphical data analysis and visualization:* This package boasts robust capabilities for graphical data analysis and visualization, rendering results both on the computer screen and in print.
- *Interactive data analysis:* R predominantly serves as a platform for crafting novel methods of interactive data analysis. Its

rapid growth is complemented by an array of supplementary packages. Many R programs are designed for transient use and are tailored to specific data analysis tasks.

I. Experimental Setup

In all experiments, 70% of the data are used for training purposes, and the remaining 30% are used for testing purposes and model validation. The process is run ten times with the classifier being tested on the test data each time, and all testing results are reported at the end of the process.

Therefore, three main experiments were performed.

J. Experiment I: Classification without Re-sampling

In this experiment, the RFC was applied to the dataset without incorporating any oversampling techniques. The primary objective at this stage was to assess the significance of the dataset's features. Subsequently, the RFC classifier was employed to analyze the classifier's performance when using the top 20 features, all features, the top 15, top 10, and top 5 features. This comprehensive analysis aimed to discern the classifier's behavior under varying feature subsets while avoiding resampling techniques.

The results of this trial are presented in Table 2, with a visual representation provided in Figure 6 for enhanced clarity. A close examination of the findings reveals that the RFC achieved notably high classification accuracy rates, with the top 15 features exhibiting the highest accuracy, followed by the full set of top 20, top 10, and top 5 features. However, it is crucial to note that for datasets characterized by imbalanced class distributions, an evaluation based solely on accuracy may not suffice.

Table 2.
Result of the classifier with and without SMOTE oversampling (The authors)

		Sensitivity		Specificity		Precision yes		Precision no		Accuracy		G-MAIN	
Without sampling		**AVG**	**Std**	**AVG**	**Std**	**Avg**	**Std**	**Avg**	**Std**	**Avg**	**Std**	**Avg**	**Std**
	All features	0.294	0.014	0.974	0.0023	0.606	0.010	0.91	0.0015	0.893	0.00032	0.535	0.013
	Top 15	0.277	0.00083	0.977	.00036	0.62	.0031	.909	6.42E-05	.894	.00022	.52	.00068
	Top 10	.304	.0011	.968	9.72E-05	.565	.0020	.912	.00012	.889	4.95E-05	.543	.00095
	Top 5	.257	.0015	.967	9.72E-05	.509	.0020	.905	.00018	.882	.00025	.498	.0015
Smote 100	**All features**	.359	.0018	.959	.0019	.544	.0108	.917	.00019	.888	.0015	.587	.0013
	Top 15	.329	.00072	.965	.00015	.560	.0008	.914	7.68E-05	.889	9.89E-05	.564	.00059
	Top 10	.264	.00083	.970	.000202	.539	.00091	.906	7.83E-05	.885	8.57E-05	.506	.00075
	Top 5	.191	.00041	.984	.00011	.621	.0019	.9	5.21E-05	.890	.00013	.434	.00048
Smote 200	**All features**	.382	.0012	.956	.00019	.540	.00056	.92	.00014	.888	8.57E-05	.605	.000093
	Top 15	.342	.0033	.966	.00039	.581	.0011	.916	.00036	.892	.00015	.575	.0027
	Top 10	.274	.00042	.969	.00019	.546	.0019	.908	6.28E-05	.886	.00023	.515	.00044
	Top 5	.187	.00083	.983	5.61E-05	.593	.00096	.899	9E-05	.888	8.57E-05	.429	.00095
Smote 300	**All features**	.392	.0014	.953	.00017	.529	.0013	.921	.00017	.886	.00023	.611	.00112
	Top 15	.355	.00083	.963	.00015	.563	.00104	.917	9.66E-05	.890	.00015	.585	.000676
	Top 10	.288	.00042	.968	5.61E-05	.553	.00069	.910	5.05E-05	.888	8.57E-05	.529	.000388
	Top 5	.207	.0025	.981	5.61E-05	.602	.0030	907	.00028	.889	.000301	.451	.002752
Smote 400	**All features**	.382	.0011	.958	.00015	.553	.00080	.920	.00013	.890	.000131	.605	.000848
	Top 15	.333	.0011	.966	.000202	.572	.0022	.915	.00014	.889	.000297	.567	.00098
	Top 10	.272	.00072	.972	9.72E-05	.568	.0015	.908	9.09E-05	.889	.000171	.514	.000706
	Top 5	.219	.0015	.980	9.72E-05	.594	.0022	.903	.000171	.886	.000216	.463	.00159
Smote 500	**All features**	.391	.0021	.953	.000149	.529	.0014	.921`	.000248	.891	.000257	.610	.00161
	Top 15	.343	.00042	.965	.000225	.572	.00146	.916	4.27E-05	.891	.000178	.575	.000320
	Top 10	.266	.00072	.979	.5.61E-05	.642	.00122	.908	8.6E-05	.895	.000131	.511	.000703
	Top 5	.238	.00083	.978	.000112	.599	.00182	.905	9.93E-05	.890	.000171	.483	.000856
Smote 600	**All features**	.397	.00083	.953	.000257	.534	.00129	.921	9.54E-05	.887	.000216	.615	.000621
	Top 15	.348	0	.967	.000149	.587	.00109	.916	1.18E-05	.893	.000131	.580	4.45E-05
	Top 10	.276	.00041	.977	9.72E-05	.619	.00071	.909	4.06E-05	.893	4.95E-05	.519	.000368
	Top 5	.238	0	.977	9.72E-05	.584	.00103	.905	8.58E-06	.889	8.57E-05	.483	2.40E-05
Smote 700	**All features**	.383	.00041	.954	.00015	.531	.00090	.920	5.32E-05	.886	.000148	.605	.000333
	Top 15	.308	.0025	.968	.00025	.564	.00332	.912	.000304	.889	.000445	.546	.00227
	Top 10	.245	.0011	.977	.000112	.592	.0008	.906	.000117	.890	8.57E-05	.489	.00107
	Top 5	.178	.00072	.983	.000112	.580	.00087	.898	7.11E-05	.887	4.95E-05	.418	.00082

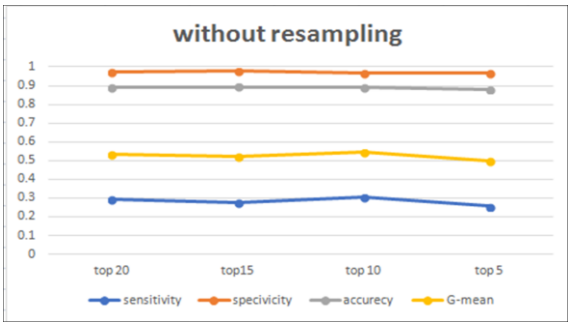

Figure 6. Results of the classifier without sampling (The authors)

To gain a more comprehensive perspective, it is imperative to consider sensitivity, specificity, and the geometric mean (G-mean). In particular, focusing on sensitivity ratios, which are pivotal for recall of the positive class in this context, the top 15 features demonstrated the most favorable performance with a 62% ratio, closely followed by the top 20 features, top 10 features, and top 5 features, registering ratios of 60.6%, 56.5%, and 50.9%, respectively. These findings shed valuable insights into the classifier's behavior across different feature subsets under the constraints of an imbalanced class distribution.

K. Experiment II: Classification with Oversampling (SMOTE)

In this study, we employ SMOTE to address class label imbalance within the dataset. Subsequently, we apply the RFC to assess the significance of the features. The RFC is then used on varying feature subsets, including the top 20 features (all features), top 15, top 10, and top 5 features, to evaluate its performance following the oversampling procedure. Importantly, the oversampling process is exclusively applied to the training data to ensure fairness in testing using unrepresented or modified data. Different oversampling ratios are explored to identify the optimal resampling ratio. In total, the RFC is trained on different feature subsets (20, 15, 10, 5), resulting in 32 unique outcomes.

The evaluation results of this experiment are presented in Table 2, and Figures 7–10 visually demonstrate the impact of oversampling ratios on classifier performance. Notably, the use of SMOTE oversampling has a limited influence on classification accuracy, as indicated in Figure 7. However, a closer look at the sensitivity values (Figure 8), which are particularly relevant for assessing minority class recall, reveals a steady improvement in classifier sensitivity with oversampling.

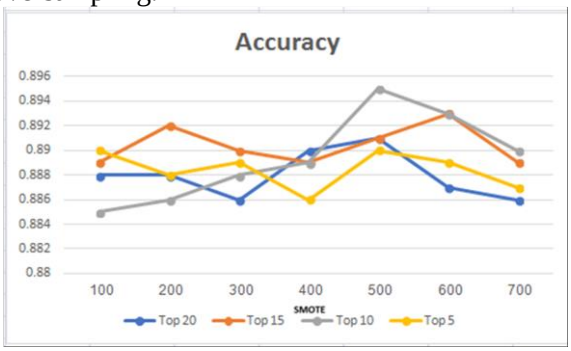

Figure 7. Accuracy (The authors)

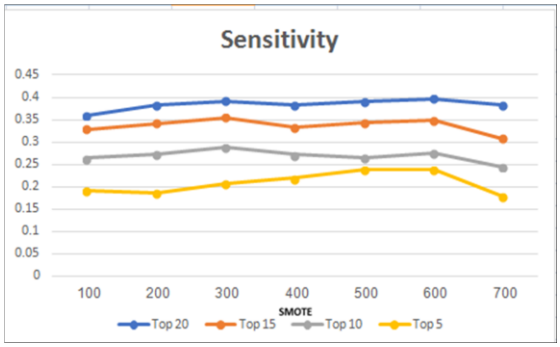

Figure 8. Sensitivity (The authors)

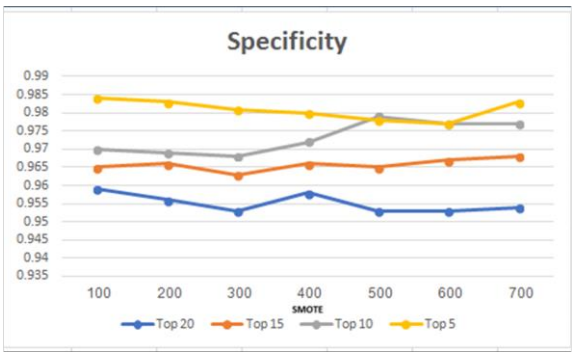

Figure 9. Specificity (The authors)

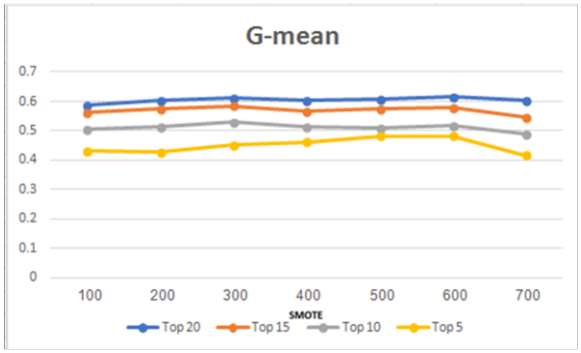

Figure 10. G-mean (The authors)

Furthermore, examining the G-mean results in Figures 10 and 11, it becomes evident that the classifier experiences significant enhancements, particularly at the 600% oversampling ratio, with respect to sensitivity and G-mean. Feature selection was conducted after each resampling ratio. The results illustrate feature importance in Table 3. Subsequently, the RFC is applied to the top 20, top 15, top 10, and top 5 features, with the outcomes documented in Table 2. Notably, a slight variation in accuracy was observed above and below the results obtained when all features were used. However, a reduction in the number of selected features leads to a decrease in the sensitivity and G-mean.

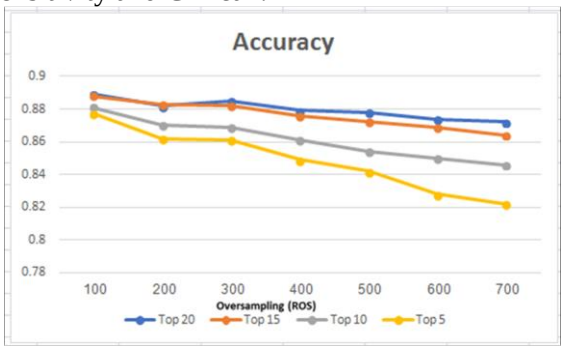

Figure 11. Accuracy (The authors)

Table 3.
Importance of features after SMOTE re-sampling (The authors)

	Data without SMOTE	SMOTE200	SMOTE300	SMOTE400	SMOTE500	SMOTE600	SMOTE700
1	Euribor3m	Euribor3m	Nr.emplpyed	Previous	Previous	Previous	Previous

Continuation of Table 3

2	Age	Nr.emplpyed	Euribor3m	Euribor3m	Euribor3m	Nr.emplpyed	Cons.conf.idx
3	Job	Cons.conf.idx	Cons.conf.idx	Nr.emplpyed	Cons.price.idx	Euribor3m	Cons.price.idx
4	Education	Cons.price.idx	Cons.price.idx	Cons.price.idx	Nr.emplpyed	Cons.conf.idx	Euribor3m
5	Nr.emplpyed	Previous	Previous	Cons.conf.idx	Cons.conf.idx	Cons.price.idx	Nr.emplpyed
6	Day-of-week	Emp.var.rate	Emp.var.rate	Emp.var.rate	Emp.var.rate	Emp.var.rate	Emp.var.rate
7	Campaign	Age	Age	Month	Month	Month	Month
8	Pdays	Job	Month	Age	Age	Age	Age
9	Month	Month	Job	Job	Job	Campaign	Pdays
10	Marital	Education	Campaign	Campaign	Campaign	job	Campaign
11	Poutcome	Campaign	Education	Pdays	Pdays	Pdays	job
12	Cons.conf.idx	Day-of-week	Pdays	Education	Education	Poutcome	Poutcome
13	Housing	Pdays	Day-of-week	Day-of-week	Day-of-week	Education	Loan
14	Emp.var.rate	Marital	Loan	Loan	Loan	Loan	Education
15	Cons.price.idx	Loan	Poutcome	Poutcome	Poutcome	Day-of-week	Day-of-week
16	Loan	Poutcome	Marital	Marital	Marital	Contact	Contact
17	Previous	Housing	Contact	Contact	Contact	Marital	Marital
18	Contact	Contact	Housing	Housing	Housing	Housing	Housing
19	Default	Default	Default	Default	Default	Default	Default
	Data without SMOTE	**ROS-200**	**ROS-300**	**ROS-400**	**ROS-500**	**ROS-600**	**ROS-700**
1	Euribor3m	Euribor3m	Euribor3m	Euribor3m	Euribor3m	Euribor3m	Euribor3m
2	Age	Age	Age	Age	Age	Age	Age
3	Job	Job	Job	Job	Job	Job	Nr.emplpyed
4	Education	Nr.emplpyed	Education	Nr.emplpyed	Nr.emplpyed	Nr.emplpyed	Job
5	Nr.emplpyed	Education	Nr.emplpyed	Education	Education	Education	Education
6	Day-of-week	Campaign	Day_of_week	Emp.var.rate	Emp.var.rate	Campaign	Campaign
7	Campaign	Day-of-week	Campaign	Campaign	Campaign	Emp.var.rate	Day_of_week
8	Pdays	Emp.var.rate	Pdays	Day_of_week	Day_of_week	Day_of_week	Emp.var.rate
9	Month	Month	Month	Cons.conf.idx	Cons.conf.idx	Month	Month
10	Marital	Cons.conf.idx	Poutcome	Month	Month	Marital	Marital
11	Poutcome	Pdays	Marital	Marital	Marital	Cons.conf.idx	Cons.conf.idx
12	Cons.conf.idx	Poutcome	Cons.conf.idx	Poutcome	Poutcome	Pdays	Pdays
13	Housing	Marital	Housing	Pdays	Pdays	Poutcome	Housing
14	Emp.var.rate	Housing	Cons.price.idx	Housing	Housing	Cons.price.idx	Cons.price.idx
15	Cons.price.idx	Cons.price.idx	Emp.var.rate	Cons.price.idx	Cons.price.idx	Housing	Poutcome
16	Loan	Loan	Loan	Loan	Loan	Loan	Loan
17	Previous	Contact	Previous	Default	Default	Default	Previous
18	Contact	Previous	Contact	Contact	Contact	Contact	Default
19	Default	Default	Default	Previous	Previous	Previous	Contact

L. Experiment III: Classification with Random Oversampling (ROS)

In this study, we tackle the challenge of class label imbalance by implementing the random oversampling (ROS) technique on the dataset. Following this, we use the RFC to discern the importance of various features. The RFC is subsequently applied to different feature subsets, specifically the top 20 (comprising all features), top 15, top 10, and top 5 features. The classifier's performance is evaluated after resampling, incorporating varying oversampling ratios to identify the most effective model.

To evaluate the classifier's performance, we train it on data oversampled with ROS at various ratios, ranging from 100% to 700%, with a 100% increment at each step. This ratio represents the proportion of instances created from the rare class through ROS. Furthermore, the RFC is trained on different feature subsets at each resampling ratio (20, 15, 10, 5), yielding 32 unique outcomes. The results of this experiment are presented in Table 4, while Figures 11-14 vividly illustrate the impact of the oversampling ratio on the classifier's performance.

Table 4.
Result of the classifier with and without oversampling (ROS) (The authors)

		Sensitivity		**Specificity**		**Precision yes**		**Precision no**		**Accuracy**		**G-MAIN**	
Without sampling		**AVG**	**Std**	**AVG**	**Std**	**Avg**	**Std**	**Avg**	**Std**	**Avg**	**Std**	**Avg**	**Std**
	All features	.294	.014	.974	.0023	.606	.010	.91	.0015	.893	.00032	.535	.013
	Top 15	.277	.00083	.977	.00036	.62	.0031	.909	6.42E-05	.894	.00022	.52	.00068
	Top 10	.304	.0011	.968	9.72E-05	.565	.0020	.912	.00012	.889	4.95E-05	.543	.00095
	Top 5	.257	.0015	.967	9.72E-05	.509	.0020	.905	.00018	.882	.00025	.498	.0015
ROS 100	**All features**	.351	.0014	.962	.00024	.555	.00059	.917	.00015	.889	4.95E-05	.581	.0011
	Top 15	.348	.0012	.961	.00011	.547	.00081	.916	.00014	.888	.00013	.579	.00101
	Top 10	.340	.0011	.954	.00015	.501	.0015	.915	.00014	.881	.00025	.569	.00095

	Continuation of Table 4											
	Top 5	.344	.0028	.949	9.72E-05	.475	.0013	.915	.00067	.877	.00066	.571	.00235
ROS 200	**All features**	.416	.0012	.944	9.61E-05	.503	.00066	.923	.00015	.882	.00013	.627	.00093
	Top 15	.437	.00093	.943	.000312	.509	.00128	.925	.00012	.883	.00022	.641	.00064
	Top 10	.421	.00181	.931	.000168	.451	.00163	.923	.00023	.870	.00035	.626	.0014
	Top 5	.430	.00166	.921	.000488	.423	.00233	.923	.00024	.862	.00060	.629	.0014
ROS 300	**All features**	.420	.00666	.947	.000149	.510	.00448	.925	.00080	.885	.00087	.630	.0050
	Top 15	.447	.00705	.940	.001898	.494	.00621	.927	.00077	.882	.00134	.648	.0048
	Top 10	.425	.00524	.927	.001457	.436	.00674	.924	.00088	.869	.00179	.628	.0042
	Top 5	.433	.00631	.917	.000342	.409	.00254	.925	.00075	.861	.00044	.630	.0045
ROS 400	**All features**	.440	.00149	.938	5.61E-05	.489	.00071	.925	.00018	.879	.00014	.642	.0011
	Top 15	.450	.00109	.934	.000148	.477	5.26E-05	.926	.00013	.876	0	.648	.00074
	Top 10	.458	.0014	.915	.000112	.421	.000831	.926	.000182	.861	.000198	.647	.00101
	Top 5	.464	.0011	.901	.001158	.386	.000321	.926	.000746	.849	.00142	.646	.00049
ROS 500	**All features**	.456	.0015	.934	.000423	.484	.001499	.927	.000176	.878	.00034	.653	.00102
	Top 15	.497	.00083	.922	.000148	.463	.000686	.931	.000108	.872	.00017	.677	.00057
	Top 10	.481	.00109	.905	.000202	.405	.001018	.928	.000154	.854	.00029	.659	.00082
	Top 5	.486	.00083	.890	.000112	.375	.000348	.928	.000104	.842	9.89E-05	.648	.00054
ROS 600	**All features**	.478	.00109	.928	.000158	.471	.000402	.929	.000167	.874	.000125	.666	.00071
	Top 15	.504	.00071	.919	.000393	.455	.001249	.932	9.58E-05	.869	.000357	.680	.00050
	Top 10	.481	.00181	.900	.000341	.393	9.10E-05	.928	.000208	.850	8.57E-05	.658	.00111
	Top 5	.497	.00143	.873	.000202	.346	.000324	.928	.000177	.828	4.95E-05	.659	.00088
ROS 700	**All features**	.479	.00143	.926	.00028	.465	.0004759	.929	.000164	.872	.000131	.664	.00091
	Top 15	.503	.00041	.912	.000312	.437	.001077	.931	.7.50E-05	.864	.000324	.677	.00039
	Top 10	.483	.00124	.896	.000148	.384	.000892	.928	.0001704	.846	.000262	.658	.00089
	Top 5	.516	.00109	.864	.000244	.338	.000415	.930	.000139	.822	.000178	.667	.00066

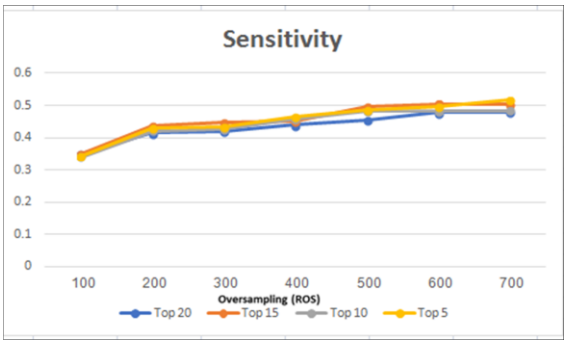

Figure 12. Sensitivity (The authors)

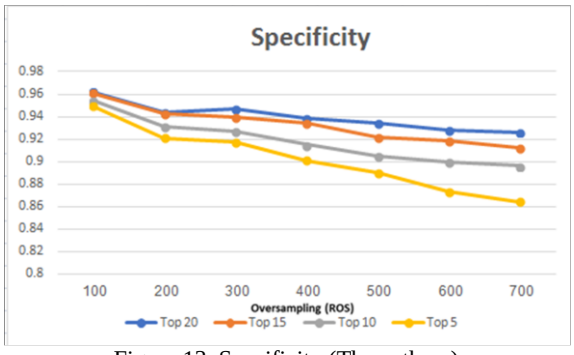

Figure 13. Specificity (The authors)

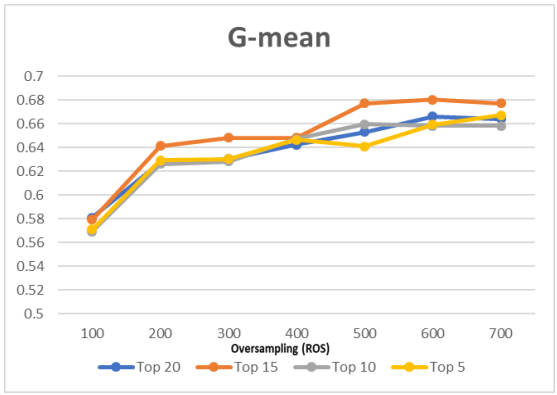

Figure 14. G-mean (The authors)

Notably, Figure 11 demonstrates that ROS oversampling leads to a marginal reduction in accuracy as the resampling ratio increases. Focusing on sensitivity values (Figure 12), a key metric for assessing minority class recall, we observed a consistent enhancement in the classifier's sensitivity, outperforming the SMOTE technique. Turning our attention to the G-mean results in Figures 14 and 15, we note a slight but consistent improvement in classifier performance, with ROS surpassing SMOTE, and the optimal result achieved at the 600% oversampling ratio for G-mean.

Feature selection is systematically applied after each resampling ratio. The feature importance results are presented in Table 5. The RFC is then applied to the top 20, top 15, top 10, and top 5 features, and the results are documented in Table 4. The analysis reveals a decrease in accuracy when the number of selected features is reduced, which is consistent with the behavior observed for the G-mean.

Comparing the evaluation results in Tables 2 and 4, we observe that feature selection after SMOTE is more reliable than that after ROS. This distinction arises from ROS dependence on a bootstrap approach, which may generate a surplus of replica data.

In this context, the best model for Experiment 2 was achieved with a 500% ROS resampling rate and all features selected, resulting in an accuracy of 89.1% and a G-mean of 0.61. For Experiment 3, the optimal model arises when ROS resampling is set at 500%, with 15 features selected, yielding an accuracy of 87.2% and a G-mean of 0.677.

III. Conclusion and Future Work

This study addresses imbalanced datasets in the context of predicting long-term deposits

within banking institutions. Through the use of random oversampling (ROS) and SMOTE, the impact of these techniques on enhancing classification model performance and stability was investigated. Experiment 2 showcased SMOTE re-sampling at a 500% ratio, encompassing all features, resulting in a robust model with 89.1% accuracy and a G-mean of 0.61. Conversely, Experiment 3 demonstrated the efficacy of ROS re-sampling at a 500% ratio with the top 15 features, achieving 87.2% accuracy and an impressive G-mean of 0.677.

Comparison with existing literature underscores the effectiveness of ROS and SMOTE methodologies in mitigating imbalanced dataset challenges, showcasing competitive performance. In addition, deliberate feature selection in conjunction with various re-sampling ratios offers nuanced insights into model behavior and sensitivity to feature subsets.

The outcomes have significant practical implications, particularly in domains such as banking and finance, where imbalanced datasets are prevalent. The demonstrated efficacy of ROS and SMOTE techniques, coupled with insights from feature selection, highlights their potential adoption in real-world scenarios. Emphasis on metrics beyond accuracy, including sensitivity, specificity, and G-mean, underscores the importance of a comprehensive evaluation framework in imbalanced classification tasks.

Recommendations include exploring hybrid approaches, integrating multiple re-sampling techniques, and employing advanced ensemble methods for improved predictive performance. Further research avenues include exploring diverse classifiers and additional feature engineering techniques to enhance model robustness against class imbalance. This study's comprehensive evaluation lays the groundwork for future research, facilitating the development of reliable predictive models in the presence of imbalanced datasets.

ACKNOWLEDGMENT

This research was funded by the Deanship of Research and Graduate Studies at Zarqa University, Jordan.

REFERENCES

[1] MORO, S., CORTEZ, P., and RITA, P. (2014) A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, 62, pp. 22-31.

[2] SONI, D. (2018) Dealing with imbalanced classes in machine learning.

[3] VAIDEHI, R. (2016) Predictive modeling to improve success rate of bank direct marketing campaign. *International Journal of Management and Business Studies*, 6 (1), pp. 22-24.

[4] VAJIRAMEDHIN, C. and SUEBSING, A. (2014) Feature selection with data balancing for prediction of bank telemarketing. *Applied Mathematical Sciences*, 8 (114), pp. 5667-5672.

[5] VENABLES, W.N., SMITH, D.M., and R DEVELOPMENT CORE TEAM (2009) An introduction to R.

[6] LUNARDON, N., MENARDI, G., and TORELLI, N. (2013) R Package 'ROSE': Random Over-Sampling Examples.

[7] MORO, S., LAUREANO, R., and CORTEZ, P. (2011) Using data mining for bank direct marketing: An application of the CRISP-DM methodology. In: *Proceedings of the European Simulation and Modelling Conference, Guimaraes, October 2011*. Guimaraes: University of Minho, pp. 117-121.

[8] KAGGLE (2018) *Portuguese Bank Marketing Data Set.* [Online] Available from: https://www.kaggle.com/datasets/yufengsui/portuguese-bank-marketing-data-set/ [Accessed 08/12/23].

[9] ALAGHA, A.S., FARIS, H., HAMMO, B.H., and ALA'M, A.Z. (2018) Identifying β-thalassemia carriers using a data mining approach: The case of the Gaza Strip, Palestine. *Artificial Intelligence in Medicine*, 88, pp. 70-83.

[10] CASTILLO, P.A., MORA, A.M., FARIS, H., MERELO, J.J., GARCÍA-SÁNCHEZ, P., FERNÁNDEZ-ARES, A.J., DE LAS CUEVAS, P., and GARCÍA-ARENAS, M.I. (2017) Applying computational intelligence methods for predicting the sales of newly published books in a real editorial business management environment. *Knowledge-Based Systems*, 115, pp. 133-151.

[11] ISSA, G. (2021) A new two-step ensemble learning model for improving stress prediction of automobile drivers. *The International Arab Journal of Information Technology*, 18 (16), pp. 819-829.

[12] CHEN, R.C., DEWI, C., HUANG, S.W., and CARAKA, R.E. (2020) Selecting critical features for data classification based on machine learning methods. *Journal of Big Data*, 7 (1), 52.
[13] BAROT, P. and JETHVA, H. (2022) MiNB: Minority Sensitive Naïve Bayesian Algorithm for Multi-Class Classification of Unbalanced Data. *The International Arab Journal of Information Technology*, 19 (4), pp. 609-616.
[14] AL-SHERIDEH, A.S., MAABREH, K., MAABREH, M., AL MOUSA, M.R., and ASASSFEH, M. (2023) Assessing the Impact and Effectiveness of Cybersecurity Measures in e-Learning on Students and Educators: A Case Study. *International Journal of Advanced Computer Science and Applications*, 14 (5), pp. 158-164.
[15] AL-KHATEEB, M., AL-MOUSA, M., AL-SHERIDEH, A., ALMAJALI, D., ASASSFEHA, M., and KHAFAJEH, H. (2023) Awareness model for minimizing the effects of social engineering attacks in web applications. *International Journal of Data and Network Science*, 7 (2), pp. 791-800.
[16] AL-QAMMAZ, A.Y., AHMAD, F.K., and YUSOF, Y. (2018) Social spider optimisation algorithm for dimension reduction of electroencephalogram signals in human emotion recognition. *International Journal of Engineering & Technology*, 7 (2.15), pp. 146-149.
[17] JUNOH, A.K., ALZOUBI, W.A., ALAZAIDAH, R., and AL-LUWAICI, W. (2020) New features selection method for multi-label classification based on the positive dependencies among labels. *Solid State Technology*, 63 (2s). Available from https://www.solidstatetechnology.us/index.php/JSST/article/view/6464.
[18] KHALIFEH, A.F., ALQAMMAZ, A.Y., ABUALIGAH, L., KHASAWNEH, A.M., and DARABKH, K.A. (2022) A machine learning-based weather prediction model and its application on smart irrigation. *Journal of Intelligent & Fuzzy Systems*, 43 (2), pp. 1835-1842.

参考文:

[1] MORO, S.、CORTEZ, P. 和 RITA, P. (2014)
预测银行电话营销成功与否的数据驱动方法。决策支持系统，62，第 22-31 页。
[2] SONI, D. (2018)
处理机器学习中的不平衡类别。
[3] VAIDEHI, R. (2016)
提高银行直销活动成功率的预测建模。《国际管理与商业研究杂志》，6 (1)，第 22-24 页。
[4] VAJIRAMEDHIN, C. 和 SUEBSING, A. (2014)
用于预测银行电话营销的数据平衡特征选择。应用数学科学，8 (114)，第 5667-5672 页。
[5] VENABLES, W.N.、SMITH, D.M. 和 R 开发核心团队 (2009) 右简介。
[6] LUNARDON, N.、MENARDI, G. 和 TORELLI, N. (2013)
右包“玫瑰”：随机过采样示例。
[7] MORO, S.、LAUREANO, R. 和 CORTEZ, P. (2011)
使用数据挖掘进行银行直接营销：CRISP-DM方法的应用。见：欧洲模拟和建模会议记录，吉马良斯，2011年10月。吉马良斯：米尼奥大学，第 117-121 页。
[8] 卡格尔 (2018)
葡萄牙银行营销数据集。[在线]可从：https://www.kaggle.com/datasets/yufengsui/portuguese-bank-marketing-data-set/ [访问日期：23年8月12日]。
[9] ALAGHA, A.S.、FARIS, H.、HAMMO, B.H. 和 ALA’M, A.Z. (2018)
使用数据挖掘方法识别β-地中海贫血携带者：巴勒斯坦加沙地带的案例。医学中的人工智能，88，第 70-83 页。
[10] CASTILLO, P.A.、MORA, A.M.、FARIS, H.、MERELO, J.J.、GARCÍA-SÁNCHEZ, P.、FERNÁNDEZ-ARES, A.J.、DE LAS CUEVAS, P. 和 GARCÍA-ARENAS, M.I. (2017)
应用计算智能方法在真实的编辑业务管理环境中预测新出版图书的销售情况。基于知识的系统，115，第 133-151 页。
[11] ISSA, G. (2021)

一种新的两步集成学习模型，用于改进汽车驾驶员的压力预测。《国际阿拉伯信息技术杂志》，18 (16)，第 819-829 页。

[12] 陈 R.C.、DEWI, C.、黄 S.W. 和 CARAKA, R.E. (2020) 基于机器学习方法选择数据分类的关键特征。大数据杂志，7（1），52。

[13] BAROT, P. 和 JETHVA, H. (2022) 米诺布：用于不平衡数据多类分类的少数敏感朴素贝叶斯算法。《国际阿拉伯信息技术杂志》，19 (4)，第 609-616 页。

[14] AL-SHERIDEH, A.S.、MAABREH, K.、MAABREH, M.、AL MOUSA, M.R. 和 ASASSFEH, M. (2023) 评估电子学习中网络安全措施对学生和教育工作者的影响和有效性：A案例分析。国际高级计算机科学与应用杂志，14 (5)，第 158-164 页。

[15] AL-KHATEEB, M.、AL-MOUSA, M.、AL-SHERIDEH, A.、ALMAJALI, D.、ASASSFEHA, M. 和 KHAFAJEH, H. (2023) 最小化社会影响的意识模型网络应用程序中的工程攻击。国际数据与网络科学杂志，7 (2)，第 791-800 页。

[16] AL-QAMMAZ, A.Y.、AHMAD, F.K. 和 YUSOF, Y. (2018) 用于人类情感识别中脑电图信号降维的社交蜘蛛优化算法。国际工程与技术杂志，7 (2.15)，第 146-149 页。

[17] JUNOH, A.K., ALZOUBI, W.A., ALAZAIDAH, R., 和 AL-LUWAICI, W. (2020) 基于标签间正相关性的多标签分类新特征选择方法。固态技术，63（2s）。可从https://www.solidstatetechnology.us/index.php/JSST/article/view/6464 获取。

[18] KHALIFEH, A.F.、ALQAMAZ, A.Y.、ABUALIGAH, L.、KHASAWNEH, A.M. 和 DARABKH, K.A. (2022) 基于机器学习的天气预报模型及其在智能灌溉中的应用。智能与模糊系统杂志，43 (2)，第 1835-1842 页。