








Article

OUCH: Oversampling and Undersampling Cannot Help Improve Accuracy in Our Bayesian Classifiers That Predict Preeclampsia

Franklin Parrales-Bravo ^{1,*}, Rosangela Caicedo-Quiroz ^{2,*}, Elena Tolozano-Benitez ³,
Víctor Gómez-Rodríguez ⁴, Lorenzo Cevallos-Torres ¹, Jorge Charco-Aguirre ¹ and Leonel Vasquez-Cevallos ⁵

¹ Grupo de Investigación en Inteligencia Artificial, Facultad de Ciencias Matemáticas y Físicas, Universidad de Guayaquil, Guayaquil 090514, Ecuador; lorenzo.cevallost@ug.edu.ec (L.C.-T.); jorge.charcoa@ug.edu.ec (J.C.-A.)

² Centro de Estudios para el Cuidado Integral y la Promoción de la Salud, Universidad Bolivariana del Ecuador, km 5 ½ vía Durán—Yaguachi, Durán 092405, Ecuador

³ Centro de Estudios en Tecnologías Aplicadas, Universidad Bolivariana del Ecuador, km 5 ½ vía Durán—Yaguachi, Durán 092405, Ecuador; etolozanob@ube.edu.ec

⁴ Instituto Superior Tecnológico Urdesa (ITSU), Av. Pdte. Carlos Julio Arosemena Tola km 2 ½, Guayaquil 090615, Ecuador; vgoomez@itsu.edu.ec

⁵ SIMUEES Simulation Clinic, Universidad Espíritu Santo, Samborondón 092301, Ecuador; leonelvasquez@uees.edu.ec

* Correspondence: franklin.parralbs@ug.edu.ec (F.P.-B.); rcaicedo@ube.edu.ec (R.C.-Q.)

Abstract: Unbalanced data can have an impact on the machine learning (ML) algorithms that build predictive models. This manuscript studies the influence of oversampling and undersampling strategies on the learning of the Bayesian classification models that predict the risk of suffering preeclampsia. Given the properties of our dataset, only the oversampling and undersampling methods that operate with numerical and categorical attributes will be taken into consideration. In particular, synthetic minority oversampling techniques for nominal and continuous data (SMOTE-NC), SMOTE—Encoded Nominal and Continuous (SMOTE-ENC), random oversampling examples (ROSE), random undersampling examples (UNDER), and random oversampling techniques (OVER) are considered. According to the results, when balancing the class in the training dataset, the accuracy percentages do not improve. However, in the test dataset, both positive and negative cases of preeclampsia were accurately classified by the models, which were built on a balanced training dataset. In contrast, models built on the imbalanced training dataset were not good at detecting positive cases of preeclampsia. We can conclude that while imbalanced training datasets can be addressed by using oversampling and undersampling techniques before building prediction models, an improvement in model accuracy is not always guaranteed. Despite this, the sensitivity and specificity percentages improve in binary classification problems in most cases, such as the one we are dealing with in this manuscript.

Keywords: preeclampsia; bayesian network classifiers; class imbalance; oversampling; undersampling; SMOTE-NC; ROSE; SMOTE-ENC

MSC: 68T37



Citation: Parrales-Bravo, F.; Caicedo-Quiroz, R.; Tolozano-Benitez, E.; Gómez-Rodríguez, V.; Cevallos-Torres, L.; Charco-Aguirre, J.; Vasquez-Cevallos, L. OUCH: Oversampling and Undersampling Cannot Help Improve Accuracy in Our Bayesian Classifiers That Predict Preeclampsia. *Mathematics* **2024**, *12*, 3351. <https://doi.org/10.3390/math12213351>

Academic Editors: Raydonal Ospina, Victor Leiva and Cecilia Castro

Received: 24 September 2024

Revised: 22 October 2024

Accepted: 23 October 2024

Published: 25 October 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Nowadays, Electronic Medical Records (EMRs) are being used to collect clinical data for training prediction models in a variety of medical fields [1]. Modern hospitals provide a wide range of monitoring and data collection tools that make it inexpensive to gather and store data for intra- and interhospital information systems [2]. It is well acknowledged, nonetheless, that data preprocessing is required prior to training predictive models [3,4].

Regarding preeclampsia, it is important to mention that, according to the Ministry of Health's 2020 Report, hypertension resulting from severe preeclampsia [5] is the primary

cause of maternal death in the Ecuadorian province of Guayas. Moreover, ten to fifteen percent of all maternal fatalities globally are presently attributed to preeclampsia, a placenta-related illness [6]. Thus, early identification and treatment of this illness can greatly enhance maternal and perinatal outcomes [7]. As in other fields [8,9], we can make use of information and communication technologies to improve the situation. Using machine learning (ML) techniques might aid in the creation of early preeclampsia prediction models [3].

One important problem in machine learning to be mindful of while training classification models is class imbalance, which occurs when a class label occurs significantly more frequently than another [10]. Clinical data frequently show this circumstance [11]. Traditional classification methods may perform badly on imbalanced datasets [12] because they are meant to learn models on balanced datasets.

Our earlier research [3] had the purpose of building a Bayesian Network Classifier (BNC) for preeclampsia diagnosis, which also allows us to understand the fundamental factors that give rise to this condition. Regarding the data, information from the medical records of patients who were treated from 2017 to 2023 at the “IESS Los Ceibos” hospital in Guayaquil, Ecuador, was gathered retrospectively. The model trained with the TAN_{cl} algorithm and the feature subset selection (FSS) task was the best among them, achieving accuracy close to 90%. In addition, the medical interpretation of the causal influence relationships in the classifying models was performed, which usually agreed with what the medical literature has mentioned to date. Since the study addressed the use of an unbalanced dataset, we leave as future work the consideration of some data augmentation techniques such as the Synthetic Minority Oversampling Technique (SMOTE) to improve the robustness and accuracy of the model. Consequently, we think it is appropriate to examine, in the present manuscript, how undersampling and oversampling can improve the performance of BNCs.

It has been widely mentioned in the literature that class balancing produces improvements in the accuracy of classification models. For example, consider the studies presented in [13–15]. Therefore, our motivation is to review whether that improvement can always occur or, on the contrary, if this study can serve as an example demonstrating how accuracy is not always improved when class balancing is considered.

Our contributions are listed as follows:

- The present study addresses the influence of oversampling and undersampling methods on Bayesian classifiers that predict preeclampsia.
- Contrary to what is widely mentioned in various works, our results demonstrate that an improvement in model accuracy is not always guaranteed.
- Despite the previous point, a balanced performance has been obtained in the classification of positive and negative cases of preeclampsia.
- This study can serve as an example of the following conclusion: while using oversampling and undersampling techniques to address imbalanced training datasets can help to balance the detection of positive and negative cases in binary classification problems like the one under review in this manuscript, an improvement in the models' accuracy is not always guaranteed.

The paper's remaining sections are arranged as follows: Previous studies that are pertinent to this inquiry are presented in Section 2. A concise synopsis of the approach employed to examine the performance of Bayesian models trained on augmented data with ROSE, SMOTE-NC, SMOTE-ENC, OVER, and UNDER is given in Section 3. Section 4 goes into depth on the findings and the analysis of the findings. Finally, concluding thoughts are included in Section 5.

2. Related Work

Medical datasets are usually imbalanced [16]. The dataset used in our previous work is no exception to that [3]. In this section we describe the use of class balancing techniques on clinical datasets. In addition, we review our previous work, its findings, and characteristics of the dataset, with an emphasis on the uneven distribution of its class labels. Then, we

explore techniques that allow class balancing on our dataset, which has continuous and nominal variables. Finally, we indicate those class balancing techniques that will be used in the present study.

2.1. Class Balancing in Clinical Datasets

The use of class balancing techniques on clinical datasets has been carried out in several studies. For example:

- In [16] undersampling and oversampling methods such as SMOTE have been used for class balancing in two lung cancer datasets. According to their results, undersampling techniques have the highest standard deviation (SD), and oversampling techniques have the lowest SD. They conclude that oversampling is a stable method to balance the class because their lowest SD achieved when training classification models on balanced training sets.
- In [17], authors presented an empirical performance evaluation of classification models for five imbalanced clinical datasets, Breast Cancer Disease, Coronary Heart Disease, Indian Liver Patient, Pima Indians Diabetes Database, and Coronary Kidney Disease. They evaluated the performance of classification models built on training sets that considered the following techniques: Undersampling, Random oversampling, SMOTE, ADASYN, SVM-SMOTE, SMOTEEN, and SMOTETOMEK. According to the results, SMOTEEN performed better than the other six data balancing techniques for all five clinical datasets.
- In [18], classification algorithms, combined with resampling strategies and dimensionality reduction methods, were investigated to find a prediction model to correctly identify between growth-hormone treated and non-treated animals. According to their results, SMOTE helped to improve the performance of their classification models.

All in all, within the medical field, many works have benefited from the use of class balancing techniques to improve the performance of their classifier models.

2.2. Class Imbalance Problem of Our Previous Work

In our earlier research [3], the Naïve Bayes, Semi Naïve Bayes (FSSJ), and Chow-Liu Tree Augmented Naïve Bayes (TAN_{cl}) algorithms were used to create explanatory classification models to predict the risk of preeclampsia. A Non-Redundant Feature Selection Technique (NoReFS) was considered to carry out the feature selection procedure. The best model among them was the one that was trained using the TAN_{cl} and features selected by NoReFS. According to the most accurate model, individuals who have a serious vaginal infection are often older than 35. According to the best model, the following factors increase the risk of preeclampsia: tobacco use, rural residency, severe vaginal infections, age over 35, family history of diabetes, and personal experience with hypertension.

In terms of the data gathered, the hospital “IESS Los Ceibos” provided 1467 EMRs. Sixty-four fundamental numerical and category features were gathered. These included demographics, age, cultural traits, and medical history, either personal or familial. “Disease1” was the attribute that needed to be classified (or predicted). Its two category values were positive and negative; positive represents patients who have the condition, while negative represents the opposite circumstance. The distribution of positive and negative values was found using the medical records, as Table 1 illustrates. The baseline accuracy is around 76% when all records are labeled as negative cases.

Table 1. Distribution of positive–negative cases of preeclampsia.

Case	Number of Records
Positive	351
Negative	1116

Table 1 indicates that 76% of patients do not exhibit preeclampsia, indicating an imbalance in our data. Therefore, in our prior work [3], we employed the F1-score measure rather than accuracy to compare the model performances on a class imbalance dataset, saving the usage of undersampling and oversampling strategies for further research. As a result, the effects produced when balancing our training dataset that contains both numerical and categorical variables will be analyzed in this work.

2.3. Class Balancing Techniques That Handles Nominal and Continuous Variables

Both oversampling and undersampling are common methods used for balancing the class. Oversampling is frequently utilized when there is an unequal distribution of data, with one class being considerably underrepresented (as in our dataset). However, until the majority class becomes as popular as the minority class, undersampling is frequently utilized to exclude records that include it. Oversampling could be a more workable method if the dataset's distribution is imbalanced, per [19]. Important data might be lost as a result of undersampling, which might not be able to balance the class distribution [20].

The Synthetic Minority Oversampling Technique (SMOTE) [21] is a widely used oversampling method. In it, the closest minority class neighbors in the sample are used to randomly generate new instances of the minority class; the nearest neighbors are identified by calculating the Euclidean distance between data points in the feature space. SMOTE is a popular technique that outperforms basic random oversampling (OVER). It cannot be used, nevertheless, when the datasets have both nominal and continuous characteristics. In [21], SMOTE-NC (SMOTE-Nominal Continuous) was also proposed as a solution to this issue. Through the preservation of the original labels of categorical features in the resampled data, it handles nominal and continuous characteristics differently.

Another oversampling-based technique called Random Oversampling Examples (ROSE) has been proposed in [22]. By producing synthetic data points that, with regard to a probability distribution centered on the chosen sample, are as close as feasible to the genuine ones, it adds fresh examples of the minority class. ROSE employs the imbalance ratio (IR) to quantify the unequal distribution of data across classes. When the IR value is equal to 1, the dataset is in appropriate balance. Greater IR values correspond to a greater variation in class sizes. ROSE employs an estimated conditional kernel density of the two classes [23] to generate synthetic data points. ROSE is made to handle both continuous and categorical variables, just like SMOTE-NC.

Recently, in [24], a novel minority oversampling technique called SMOTE-ENC (SMOTE—Encoded Nominal and Continuous) is proposed. In it, nominal features are encoded as numeric values, and the difference between two numerical values representing those features indicates the degree of change in affiliation with the minority class. SMOTE-ENC can handle datasets that only contain nominal features, which is not possible in the SMOTE-NC method. Moreover, according to their authors, this method overcomes the problem of SMOTE-NC, which fails to interpret the difference in association between each label and the minority class target in the case of multi-label nominal features.

All in all, we will examine the performance of the Bayesian classification models that are trained on augmented training data when considering the following methods: Random Undersampling (UNDER), Random Oversampling (OVER), ROSE, SMOTE-NC, and SMOTE-ENC. This is because we have a class imbalanced dataset combining categorical and numerical variables.

3. Methodology

This section describes the steps used to assess how well the BNCs, which were built on a balanced training dataset when using OVER, UNDER, ROSE, SMOTE-NC, and SMOTE-ENC, performed in predicting the risk of preeclampsia. The process we followed to achieve our objective is depicted in Figure 1. Details for each of the stages outlined therein are provided below.

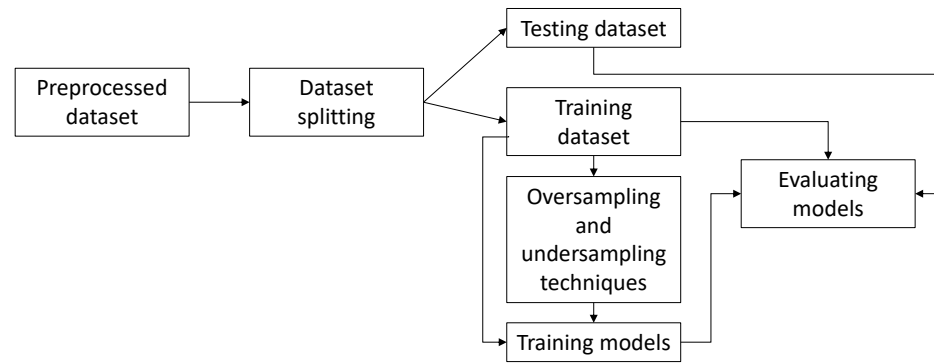


Figure 1. Workflow of the methodology applied to evaluate the performance of BNCs that predict the risk of preeclampsia.

3.1. Preprocessed Dataset

This stage finds out if there are duplicate columns in the dataset. We remove duplicate columns. Furthermore, columns that are the same in every EMR are eliminated. We took into account both methods in accordance with the recommendations made in [1,2].

Moreover, we will apply the Correlation-based feature selection (CFS) technique to reduce the number of features in our dataset. CFS calculates the correlation between each pair of features and then selects the subset of features that have the highest correlation with the target variable and the lowest correlation with each other [25]. All in all, CFS is a powerful technique that can help identify the most relevant variables, even improving the performance of classification models as in [1,2,26,27].

3.2. Dataset Splitting: Training and Testing Datasets

Our collected preeclampsia dataset is split 70:30 between training and test datasets, following the methodology outlined in our earlier work [3]. The training dataset in that study included the first 70% of the data, with the remaining 30% designated for the test dataset. To evenly distribute the data between both datasets, we first shuffle the records into a random order before splitting them. Furthermore, to ensure that the dataset is split while preserving class proportions, we apply a stratified split.

3.3. Oversampling and Undersampling Techniques

Since we want to balance the classes before the model training starts, we only apply oversampling and undersampling to the training dataset. This way, we do not affect the evaluation of the performance of models on the test dataset.

As aforementioned, because our dataset includes both numerical and categorical variables, we will take into account the following techniques: ROSE, SMOTE-NC, Random Oversampling (OVER), Random Undersampling (UNDER), and ROSE. We will use the R statistical program (version 4.3.2) and the functions and packages listed in Table 2 to do this. We used the default settings for all functions.

Table 2. Packages used to deal with class imbalance in R.

Technique	Package	Version	Function
UNDER	caret	6.0–94	downSample
OVER	caret	6.0–94	upSample
ROSE	ROSE	0.0–4	ROSE
SMOTE-NC	RSBID	0.0.2.0000	SMOTE_NC

Regarding the SMOTE-ENC method, we will use it from the code available in [28]. With Python code, we will generate the balanced training dataset under this technique, and then we will load it into the R code.

3.4. Training Models

As in our earlier work [3], the Chow-Liu approach of “Tree Augmented Naïve Bayes” (TAN_{cl}), adapted from [29], “Naïve Bayes” (NB), and a Semi Naïve Bayes with a forward approach to constructive induction called “Forward Sequential Selection and Joining” (FSSJ), as defined [30], will be the Bayesian classification algorithms selected to perform the model training.

We will incorporate in this work the use of the “Hill-climbing Super-Parent Tree Augmented Naïve Bayes” (TAN_{hcspt}) because it allows an efficient interactive exploration of the space of augmented Bayesian classifiers, generally choosing a different set of augmenting arcs to add that the considered by the TAN_{cl} method, producing more accurate results [31]. We will also consider other techniques implemented in the `bnclassify` package (version 0.4.5.9999) for the R statistical software (version 4.3.2). These are the following: “Averaged One-Dependence Estimators” (AODE) [32,33], “Model Averaged Naïve Bayes” (MANB) [34], “Attribute-Weighted Naïve Bayes” (AWNB) [35], and the “Hill-climbing k -dependence Bayesian classifier” (kDB) [36]. Finally, a novel technique called “Correlation-Based Feature Weighting Filter for Naïve Bayes” (CBFW) method [37] will be considered, whose implementation in Python is found in [38].

3.5. Evaluating Models

The preeclampsia BNC models are assessed to see which one performs better after training with each of the strategies mentioned in the preceding stage. We will take the accuracy measure into consideration to evaluate model performance because the dataset produced by the oversampling and undersampling strategies is balanced.

A 10-fold stratified cross-validation is taken into consideration on the training dataset to provide an honest estimation of the accuracy of the trained models. It learns the models from the training subsamples by repeating the learning procedures used to obtain the structure and parameters of the BNC.

Additionally, the test dataset is used to compare the goodness of those models when using them to classify unseen data (data not used to train the models). Additional metrics such as sensitivity, specificity, precision, recall, F1-score, and AUC-ROC will be taken into consideration in this step. These metrics are particularly important in the context of imbalanced datasets, like the one in our study, where the goal is not only to improve overall accuracy but to ensure that the model effectively identifies minority class cases—in this case, patients with preeclampsia.

4. Results and Analysis

4.1. Feature Selection

Table 3 presents the clinical features that were selected using the CFS technique, which were the same as those selected in the prior study [3].

Table 3. Selected features when applying the NoReFS approach.

Feature	Description	Labels
“Hypertensionpersonalhistory”	Hypertension personal history	yes/no
“Parity”	The number of times the fetus has reached a viable gestational age	1/2/3/4/5/6/7 or more
“Gravidity”	The number of times the woman has been pregnant	1/2/3/4/5/6/7 or more
“Fetalstatus”	Previous fetal status at birth	born alive/stillborn/NA
“Tobaccouse”	Tobacco use	yes/no
“Diabetesfamilyhistory”	Existence of relatives with diabetes	yes/no
“Nupucells1”	Patient vaginal infection	mild/moderate/severe
“Maternalage-categorized”	Maternal age by ranges	State0: <35/ State1: ≥35
“Education_Level”	Education level	primary/secondary/tertiary
“Specificplacearealivedincountyof”	Area where the patient resides	urban/rural

4.2. Oversampling and Undersampling on the Training Data

Class imbalance approaches were used on the training dataset (1027 EMRs), which makes up 70% of the original dataset (1467 EMRs), as previously reported. The distribution of positive and negative instances of preeclampsia acquired using various approaches is displayed in Table 4. The distribution of the training dataset without the use of any oversampling or undersampling techniques is referred to as the “Baseline” in this table and throughout the remainder of the paper.

Table 4. Training dataset characteristics.

Technique	EMRs	Positive Cases	Negative Cases
Baseline	1027	24.15%	75.85%
UNDER	498	50%	50%
OVER	1564	50%	50%
ROSE	1027	49.18%	50.82%
SMOTE-NC	1564	50%	50%
SMOTE-ENC	1564	50%	50%

According to [39], ROSE can produce out-of-range values, including negative area sizes, that are not feasible in the actual world. However, we do not observe any unrealistic values provided by any of the approaches, notably in numerical labels, when we examine our training dataset augmented with ROSE.

4.3. Honest Estimation of the Accuracy of Models

Table 5 displays the distribution of accuracy values achieved by the BNC models while employing 10-fold cross-validation. We built BNC models when applying Baseline, OVER, UNDER, ROSE, SMOTE-NC, and SMOTE-ENC on the training dataset. Furthermore, Figure 2 presents a box plot that allows the user to visually evaluate the distribution of accuracy values presented in Table 5.

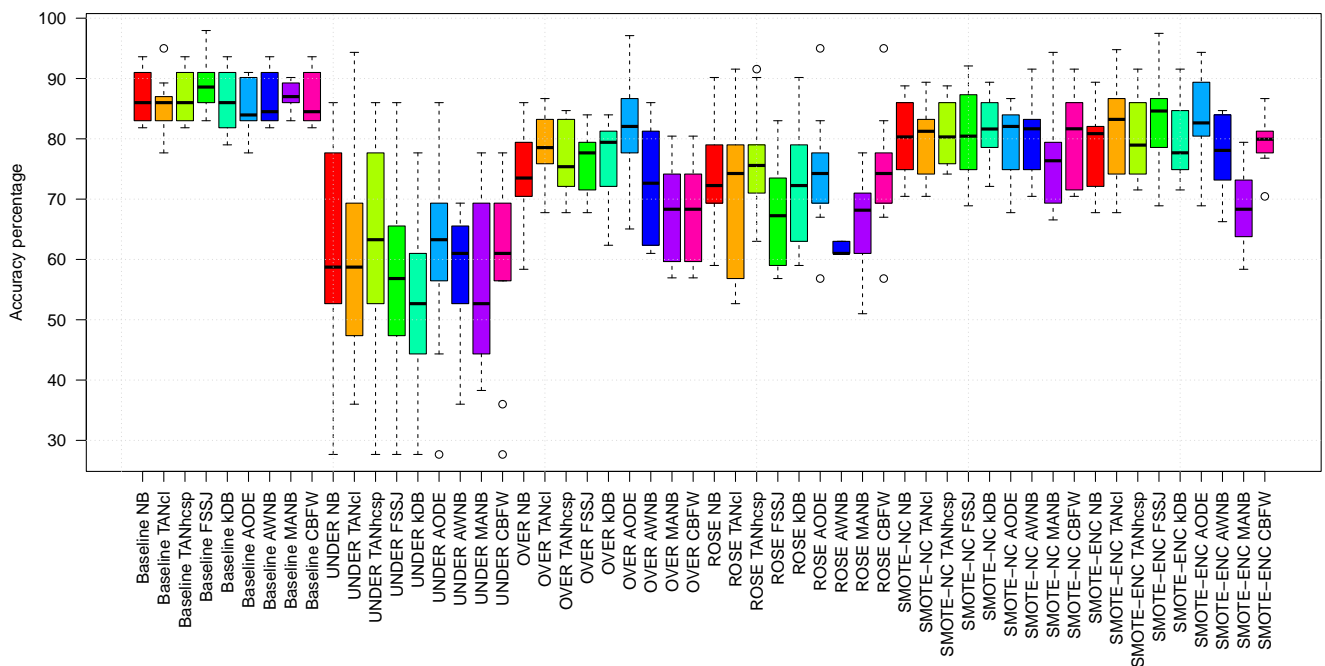


Figure 2. Box plot on the accuracy values of BNC models when using a 10-fold cross-validation. White circles indicate outliers. Identical colors indicate that the same algorithm was used to train the BNC, although they used different techniques for balancing the training set.

Table 5. Descriptive statistics of accuracy values of Bayesian models when using a 10-fold cross-validation.

Technique	Model	Mean	Std.Dev.	Max.	Min.	Range
Baseline	NB	86.42%	4.21%	93.60%	81.83%	11.77%
	TAN _{cl}	85.57%	4.67%	95.00%	77.66%	17.33%
	TAN _{hcsp}	86.42%	4.21%	93.60%	81.83%	11.77%
	FSSJ	88.92%	4.90%	97.95%	83.00%	14.95%
	kDB	86.02%	4.73%	93.60%	79.00%	14.60%
	AODE	85.15%	4.41%	91.00%	77.66%	13.33%
	AWNB	86.44%	4.53%	93.60%	81.83%	11.77%
	MANB	86.85%	2.55%	90.16%	83.00%	7.16%
	CBFW	86.44%	4.53%	93.60%	81.83%	11.77%
UNDER	NB	60.16%	17.45%	86.00%	27.66%	58.33%
	TAN _{cl}	60.01%	18.39%	94.33%	36.00%	58.33%
	TAN _{hcsp}	61.83%	17.45%	86.00%	27.66%	58.33%
	FSSJ	57.59%	18.90%	86.00%	27.66%	58.33%
	kDB	54.03%	15.98%	77.66%	27.66%	50.00%
	AODE	61.83%	16.54%	86.00%	27.66%	58.33%
	AWNB	58.57%	11.02%	69.33%	36.00%	33.33%
	MANB	56.53%	13.67%	77.66%	38.27%	39.39%
	CBFW	58.42%	16.11%	77.66%	27.66%	50.00%
OVER	NB	72.95%	8.76%	86.00%	58.36%	27.63%
	TAN _{cl}	78.04%	5.79%	86.67%	67.75%	18.91%
	TAN _{hcsp}	76.91%	6.12%	84.68%	67.75%	16.92%
	FSSJ	76.67%	5.17%	83.97%	67.75%	16.21%
	kDB	76.37%	6.86%	83.97%	62.35%	21.62%
	AODE	82.08%	8.93%	97.11%	65.05%	32.05%
	AWNB	72.69%	9.34%	86.00%	61.00%	25.00%
	MANB	67.78%	7.49%	80.44%	56.94%	23.49%
	CBFW	67.78%	7.49%	80.44%	56.94%	23.49%
ROSE	NB	73.78%	8.74%	90.16%	59.00%	31.16%
	TAN _{cl}	71.97%	14.04%	91.55%	52.66%	38.88%
	TAN _{hcsp}	76.38%	8.97%	91.55%	63.00%	26.55%
	FSSJ	67.96%	8.57%	83.00%	56.83%	26.16%
	kDB	72.96%	10.43%	90.16%	59.00%	31.16%
	AODE	74.85%	10.04%	95.00%	56.83%	38.16%
	AWNB	61.80%	1.03%	63.00%	61.00%	2.00%
	MANB	66.80%	8.15%	77.66%	51.00%	26.66%
	CBFW	74.85%	10.04%	95.00%	56.83%	38.16%
SMOTE-NC	NB	80.48%	6.11%	88.77%	70.45%	18.31%
	TAN _{cl}	79.97%	6.29%	89.37%	70.45%	18.91%
	TAN _{hcsp}	81.02%	5.32%	88.77%	74.15%	14.61%
	FSSJ	80.70%	7.62%	92.08%	68.89%	23.18%
	kDB	81.81%	5.68%	89.37%	72.11%	17.26%
	AODE	79.93%	6.39%	86.67%	67.75%	18.91%
	AWNB	80.22%	6.57%	91.55%	70.45%	21.09%
	MANB	76.41%	8.48%	94.33%	66.55%	27.77%
	CBFW	79.69%	7.47%	91.55%	70.45%	21.09%
SMOTE-ENC	NB	78.59%	6.91%	89.37%	67.75%	21.62%
	TAN _{cl}	81.05%	8.47%	94.78%	67.75%	27.02%
	TAN _{hcsp}	79.97%	6.96%	91.55%	71.52%	20.02%
	FSSJ	83.73%	8.15%	97.48%	68.89%	28.59%
	kDB	79.41%	6.42%	91.55%	71.52%	20.02%
	AODE	83.50%	7.01%	94.33%	68.89%	25.43%
	AWNB	77.52%	6.79%	84.68%	66.26%	18.42%
	MANB	68.83%	6.40%	79.42%	58.36%	21.05%
	CBFW	79.94%	4.61%	86.67%	70.45%	16.21%

It is evident from Table 5 and Figure 2 that the model trained using FSSJ on the initial imbalanced training dataset (Baseline FSSJ) achieved the best mean accuracy percentage, which was 88.92%. However, they do not differ much from the mean accuracy percentages obtained by other baseline models, with mean accuracies around 86%. Furthermore, the standard deviation obtained by the Baseline FSSJ model (4.90%) is slightly higher than that obtained by the other models trained with the baseline dataset. Among them, the model that achieves good performance with the lowest standard deviation is the Baseline MANB model, with a mean accuracy of 86.85% and a standard deviation of 2.55%. Examining the maximum values, we see that the BNC models built on the baseline training dataset have obtained maximum accuracy between 90% and 98%. All in all, models trained under the original (Baseline) training dataset obtain an accuracy close to 85%, with range values close to 10%, which indicates a small dispersion in the accuracies obtained from the cross-validation.

Regarding models built on training datasets that handled class balancing, the model trained with FSSJ when applying SMOTE-ENC (SMOTE-ENC FSSJ) to perform class balancing yielded the best mean accuracy percentage of 83.73%, followed closely by AODE with a mean accuracy of 83.50%. Moreover, its mean accuracy percentage is slightly close to that achieved by the model trained using FSSJ on the initial imbalanced training dataset (Baseline FSSJ), which was 88.92%. However, Table 5 indicates that these mean accuracy percentages obtained with Baseline FSSJ are less dispersed than those produced with SMOTE-ENC FSSJ. In fact, the achieved range value (14.95%) is lower than the value obtained with SMOTE-ENC FSSJ (28.59%). The TAN_{cl} , FSSJ, AODE, and CBFW models trained under the training dataset augmented with SMOTE-ENC obtained slightly better mean accuracy than those trained using the training dataset augmented with SMOTE-NC. The opposite case occurs when training the NB, TAN_{hcspr} , kDB, AWINB, and MANB models with the training dataset augmented with SMOTE-NC. In other words, in some cases, SMOTE-ENC helps models achieve better results, while in other cases, SMOTE-NC does so.

Regarding models built when undersampling the training dataset (UNDER), we see that the accuracies obtained by cross validation are quite dispersed, sometimes obtaining a performance of 86.33% and other times below 28%, having mean accuracies close to 60%. Moreover, the lowest outcomes—a mean accuracy percentage range of 54.03%—came from balancing the class with UNDER kDB. The low performance of models built from a training dataset balanced with UNDER can be explained because, as mentioned above (Section 2), the dataset balanced by UNDER may have lost important information, which could lead to misclassification of instances that would have followed the patterns learned from the removed information.

Overall, the findings indicate that, with a mean accuracy percentage of 88.92%, the model trained with FSSJ on the baseline training dataset (Baseline FSSJ) has the best average accuracy of the results. Moreover, the SMOTE-ENC FSSJ model was the best among those built under balanced training data, achieving an average accuracy of 83.50%. Thus far, we may infer that while we can employ any oversampling or undersampling strategy to deal with imbalanced training datasets, we cannot always guarantee an improvement in the models' accuracy as in the present work.

4.4. Performance of Models on the Testing Dataset

Table 6 displays the outcomes of the models' performance on the testing dataset. It illustrates that the Bayesian model trained using FSSJ and MANB on the original training dataset (Baseline) yielded the best accuracy results (88.64%).

Table 6. Performances of BNC models on the testing dataset.

Technique	Model	Accuracy	Sensitivity	Specificity	Precision	Recall	F1-Score	AUC
Baseline	NB	86.69%	39.16%	98.32%	60.00%	39.16%	47.39%	0.419884
	TAN _{cl}	81.84%	35.00%	96.05%	45.29%	35.00%	39.48%	0.429377
	TAN _{hcsp}	86.69%	39.16%	98.32%	60.00%	39.16%	47.39%	0.419884
	FSSJ	88.64%	35.00%	99.71%	70.00%	35.00%	46.66%	0.498206
	kDB	86.69%	39.16%	98.32%	60.00%	39.16%	47.39%	0.419884
	AODE	84.75%	35.00%	99.87%	52.85%	35.00%	42.11%	0.433069
	AWNB	85.72%	18.33%	99.92%	50.00%	18.33%	26.82%	0.428850
	MANB	88.64%	35.00%	99.71%	70.00%	35.00%	46.66%	0.434651
CBFW	86.69%	39.16%	98.32%	60.00%	39.16%	47.39%	0.419884	
UNDER	NB	77.96%	64.16%	82.15%	47.14%	64.16%	54.35%	0.573787
	TAN _{cl}	64.36%	72.50%	61.89%	38.30%	72.50%	50.12%	0.559547
	TAN _{hcsp}	67.28%	60.00%	69.49%	37.27%	60.00%	45.98%	0.516403
	FSSJ	63.39%	72.50%	60.63%	37.77%	72.50%	49.67%	0.496888
	kDB	71.16%	64.16%	73.29%	40.95%	64.16%	49.99%	0.577479
	AODE	68.25%	68.33%	68.22%	39.78%	68.33%	50.29%	0.433597
	AWNB	76.01%	51.66%	83.41%	42.25%	51.66%	46.49%	0.446782
	MANB	88.64%	35.16%	91.42%	68.33%	39.16%	49.79%	0.425158
CBFW	68.25%	68.33%	68.22%	39.78%	68.33%	50.29%	0.433597	
OVER	NB	72.13%	64.16%	74.55%	41.70%	64.16%	50.55%	0.588028
	TAN _{cl}	73.10%	60.00%	77.08%	41.57%	60.00%	49.11%	0.540032
	TAN _{hcsp}	76.01%	76.66%	75.82%	47.20%	76.66%	58.43%	0.585918
	FSSJ	68.25%	68.33%	68.22%	39.78%	68.33%	50.29%	0.517879
	kDB	74.07%	76.66%	73.29%	45.55%	76.66%	57.15%	0.581699
	AODE	72.13%	68.33%	73.29%	42.55%	68.33%	52.45%	0.576425
	AWNB	73.10%	76.66%	72.02%	44.78%	76.66%	56.53%	0.600686
	MANB	88.64%	35.16%	91.42%	70.00%	35.00%	46.66%	0.547943
CBFW	73.10%	60.00%	77.08%	41.57%	60.00%	49.11%	0.540032	
ROSE	NB	55.63%	56.83%	51.66%	29.23%	51.66%	37.33%	0.443090
	TAN _{cl}	52.71%	55.83%	51.77%	29.29%	51.77%	38.43%	0.481065
	TAN _{hcsp}	51.74%	51.66%	51.77%	27.85%	51.66%	36.19%	0.512711
	FSSJ	57.57%	43.33%	61.89%	27.39%	43.33%	33.56%	0.488976
	kDB	56.60%	55.83%	56.83%	30.75%	55.83%	39.66%	0.491613
	AODE	51.74%	55.83%	50.50%	28.96%	55.83%	38.14%	0.437816
	AWNB	56.60%	51.66%	58.10%	29.60%	51.66%	37.64%	0.442035
	MANB	31.35%	85.00%	15.06%	29.35%	85.00%	43.63%	0.426213
CBFW	63.25%	59.16%	64.49%	33.88%	59.16%	43.09%	0.565243	
SMOTE-NC	NB	70.19%	47.50%	77.08%	35.71%	47.50%	40.77%	0.558597
	TAN _{cl}	69.22%	47.50%	75.82%	35.00%	47.50%	40.30%	0.550158
	TAN _{hcsp}	68.25%	51.66%	73.29%	35.64%	51.66%	42.18%	0.531170
	FSSJ	59.51%	64.16%	58.10%	34.07%	64.16%	44.51%	0.569936
	kDB	69.75%	55.83%	73.29%	37.50%	55.83%	44.86%	0.559651
	AODE	69.75%	55.83%	73.29%	37.50%	55.83%	44.86%	0.534862
	AWNB	56.60%	60.00%	55.56%	31.81%	60.00%	41.58%	0.561761
	MANB	50.77%	64.16%	40.37%	34.65%	85.00%	49.23%	0.541719
CBFW	63.39%	72.50%	60.63%	37.77%	72.50%	49.67%	0.496888	
SMOTE-ENC	NB	79.45%	84.59%	74.32%	77.64%	84.59%	80.97%	0.674054
	TAN _{cl}	83.10%	91.48%	74.72%	79.07%	91.48%	84.82%	0.811358
	TAN _{hcsp}	84.72%	91.48%	77.97%	81.19%	91.48%	86.03%	0.824361
	FSSJ	84.72%	91.48%	77.97%	81.19%	91.48%	86.03%	0.824361
	kDB	79.59%	85.54%	73.64%	76.98%	85.54%	81.03%	0.772367
	AODE	63.25%	59.16%	64.49%	33.88%	59.16%	43.09%	0.565243
	AWNB	74.72%	92.56%	56.89%	69.54%	92.56%	79.41%	0.723558
	MANB	65.00%	97.31%	28.78%	60.79%	97.31%	74.83%	0.649694
CBFW	64.36%	72.50%	61.89%	38.30%	72.50%	50.12%	0.559547	

In Table 6 we observe that the best accuracy among the models built on balanced training datasets were those that applied OVER and UNDER with MANB (88.64%). Surprisingly, models trained on the reduced training dataset (UNDER) performed poorly in the training phase (Table 5), but performed well when training models with MANB, with accuracy above 88.64%, obtaining a performance like the models trained under the original dataset (Baseline). This implies that the patterns found in the records of the test set have been represented within the models learned under the balanced training set. Furthermore, models trained on the balanced training set with SMOTE-ENC have also achieved good performance, with those trained with FSSJ and TAN_{hcsp} standing out (84.72%).

In Table 6, we have included the percentages of sensitivity and specificity to analyze the goodness of the model when identifying positive and negative cases of preeclampsia. Specificity measures the proportion of negative instances labeled as negative, whereas sensitivity measures the percentage of positive cases classified as positive. Furthermore, sensitivity and specificity are represented in Figure 3 as the abscissas and ordinates, respectively. As can be observed, the Baseline kDB, SMOTE-ENC FSSJ, SMOTE-ENC TAN_{hcsp} , and SMOTE-ENC TAN_{cl} models make up the Pareto optimum front (non-dominated solutions). Among them, when the sensitivity and specificity percentages were concurrently maximized, the SMOTE-ENC TAN_{hcsp} and SMOTE-ENC FSSJ models, which obtained the same values of sensitivity and specificity, produced the best trade-off. In the figure, a red circle encircles it.

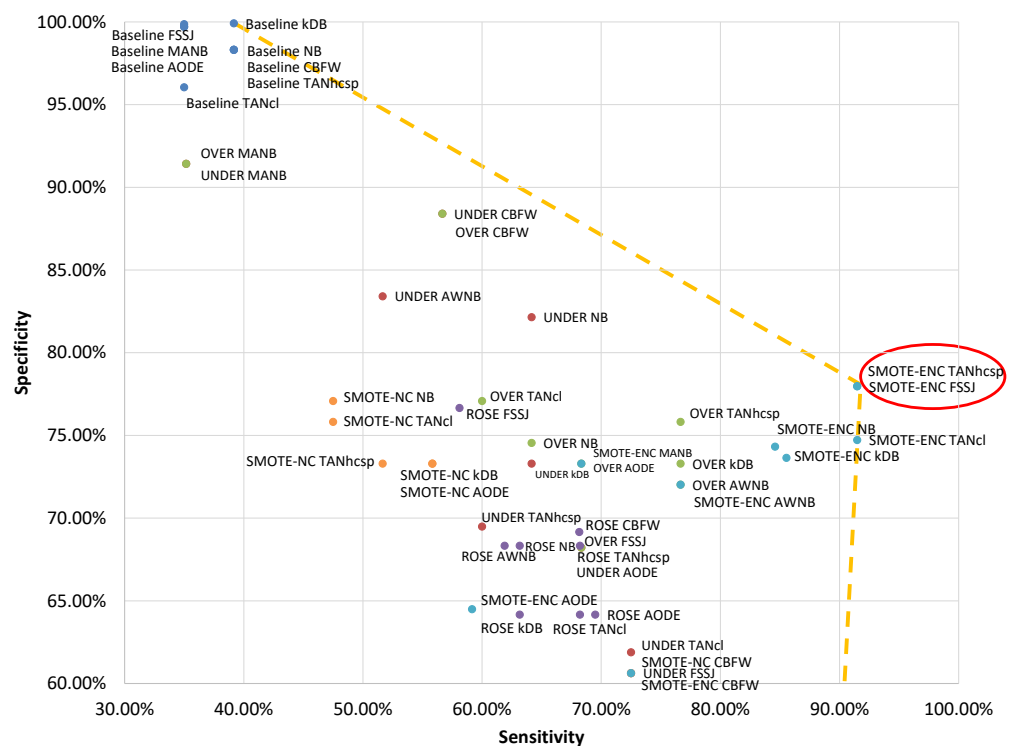


Figure 3. Sensitivity vs. specificity percentages achieved by the BNC models. The dashed yellow line marks the Pareto optimal front. The red circle encloses the best tradeoff solution.

Table 6 and Figure 3 demonstrate that, while the Baseline FSSJ model achieved the highest percentages of accuracy, its percentages of sensitivity and specificity are uneven. In contrast, although with slightly lower accuracy than that achieved by Baseline FSSJ, the SMOTE-ENC TAN_{hcsp} and SMOTE-ENC FSSJ models are able to better detect both positive and negative cases of preeclampsia. This implies that the algorithms used to train BNC models were better able to capture the patterns that define positive cases of preeclampsia due to the balanced training dataset. With a sensitivity and specificity of 91.48% and 77.97%, respectively, the SMOTE-ENC TAN_{hcsp} and SMOTE-ENC FSSJ models achieved

an accuracy of 84.72% without sacrificing the accurate categorization of both positive and negative cases of preeclampsia.

Overall, the performance of BNC models on the test dataset allows us to draw the conclusion that, while models built on balanced training datasets do not always result in significantly improved accuracy, they do achieve a better balance between sensitivity and specificity percentages in a binary classification problem like the one this manuscript addresses.

5. Conclusions

This study looked at how training the Bayesian classification models that predict the likelihood of developing preeclampsia is affected by oversampling and undersampling strategies. The oversampling and undersampling techniques that deal with numerical and categorical variables were taken into consideration since we found both types in our dataset. These techniques included UNDER, OVER, ROSE, SMOTE-NC, and SMOTE-ENC.

The results indicate that there is no improvement in accuracy percentages while class imbalances on the training dataset were handled. But when it came to correctly classifying positive and negative instances of preeclampsia in the test dataset, the models that were built on the balanced training datasets yielded a more balanced outcome in terms of sensitivity and specificity. We can draw the conclusion that, while using oversampling and undersampling techniques to address imbalanced training datasets can help balance the sensitivity and specificity percentages in binary classification problems, like the one under review in this manuscript, an improvement in the models' accuracy is not always guaranteed. This finding may vary across different datasets and applications.

As future work, and with the aim of offering a direction for practical application of our conclusion, we plan to expand the study of the performance of classifier models, which were trained on balanced data with oversampling and undersampling, to different domains and applications.

Author Contributions: All the authors have contributed to the work presented in this paper. Conceptualization, F.P.-B.; methodology, F.P.-B.; validation, F.P.-B., R.C.-Q., E.T.-B., V.G.-R., L.C.-T., J.C.-A. and L.V.-C.; investigation, F.P.-B. and J.C.-A.; writing—original draft preparation, F.P.-B.; writing—review and editing, F.P.-B., R.C.-Q., E.T.-B., V.G.-R., L.C.-T., J.C.-A. and L.V.-C. All authors have read and agreed to the published version of the manuscript.

Funding: This research has been supported by “Universidad de Guayaquil” under project FCI-008-2021. Also, it was co-financed by “Universidad Bolivariana del Ecuador” under project PROY-INV-UBE-013-2022.

Data Availability Statement: The datasets presented in this article are not readily available because the data are part of an ongoing study. Requests to access the datasets should be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Bravo, F.P.; García, A.A.D.B.; Veiga, A.B.G.; De La Sacristana, M.M.G.; Piñero, M.R.; Peral, A.G.; Džeroski, S.; Ayala, J.L. SMURF: Systematic Methodology for Unveiling Relevant Factors in retrospective data on chronic disease treatments. *IEEE Access* **2019**, *7*, 92598–92614. [CrossRef]
2. Bravo, F.P.; García, A.A.; Russo, L.; Ayala, J.L. SOFIA: Selection of Medical Features by Induced Alterations in Numeric Labels. *Electronics* **2020**, *9*, 1492. [CrossRef]
3. Parrales-Bravo, F.; Caicedo-Quiroz, R.; Rodríguez-Larraburu, E.; Barzola-Monteses, J. ACME: A Classification Model for Explaining the Risk of Preeclampsia Based on Bayesian Network Classifiers and a Non-Redundant Feature Selection Approach. *Informatics* **2024**, *11*, 31. [CrossRef]
4. Parrales-Bravo, F.; Torres-Urresto, J.; Avila-Maldonado, D.; Barzola-Monteses, J. Relevant and Non-Redundant Feature Subset Selection Applied to the Detection of Malware in a Network. In Proceedings of the 2021 IEEE Fifth Ecuador Technical Chapters Meeting (ETCM), Cuenca, Ecuador, 12–15 October 2021; pp. 1–6.
5. Ministerio de Salud Pública del Ecuador. Gaceta de Muerte Materna SE14. 2020. Available online: <https://bit.ly/3Poz790> (accessed on 28 March 2022).

6. Parrales-Bravo, F.; Saltos-Cedeño, J.; Tomalá-Esparza, J.; Barzola-Monteses, J. Clustering-based Approach for Characterization of Patients with Preeclampsia using a Non-Redundant Feature Selection. In Proceedings of the 2023 3rd International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME), Tenerife, Spain, 19–21 July 2023; pp. 1–6.
7. De Kat, A.C.; Hirst, J.; Woodward, M.; Kennedy, S.; Peters, S.A. Prediction models for preeclampsia: A systematic review. *Pregnancy Hypertens* **2019**, *16*, 48–66. [[CrossRef](#)] [[PubMed](#)]
8. Parrales-Bravo, F.; Caicedo-Quiroz, R.; Barzola-Monteses, J.; Guillén-Mirabá, J.; Guzmán-Bedor, O. CSM: A Chatbot Solution to Manage Student Questions About payments and Enrollment in University. *IEEE Access* **2024**, *12*, 74669–74680. [[CrossRef](#)]
9. Barzola-Monteses, J.; Guerrero, M.; Parrales-Bravo, F.; Espinoza-Andaluz, M. Forecasting energy consumption in residential department using convolutional neural networks. In Proceedings of the Conference on Information and Communication Technologies of Ecuador, Guayaquil, Ecuador, 24–26 November 2021; Springer: Berlin/Heidelberg, Germany, 2021; pp. 18–30.
10. Liu, Y.; Li, B.; Yang, S.; Li, Z. Handling missing values and imbalanced classes in machine learning to predict consumer preference: Demonstrations and comparisons to prominent methods. *Expert Syst. Appl.* **2024**, *237*, 121694. [[CrossRef](#)]
11. Roy, D.; Roy, A.; Roy, U. Learning from Imbalanced Data in Healthcare: State-of-the-Art and Research Challenges. In *Computational Intelligence in Healthcare Informatics*; Springer: Berlin/Heidelberg, Germany, 2024; pp. 19–32.
12. Demir, S.; Şahin, E.K. Evaluation of oversampling methods (OVER, SMOTE, and ROSE) in classifying soil liquefaction dataset based on SVM, RF, and Naïve Bayes. *Avrupa Bilim ve Teknoloji Dergisi* **2022**, *34*, 142–147. [[CrossRef](#)]
13. Moreno-Barea, F.J.; Jerez, J.M.; Franco, L. Improving classification accuracy using data augmentation on small data sets. *Expert Syst. Appl.* **2020**, *161*, 113696. [[CrossRef](#)]
14. El Gannour, O.; Hamida, S.; Lamalem, Y.; Mahjoubi, M.A.; Cherradi, B.; Raihani, A. Improving skin diseases prediction through data balancing via classes weighting and transfer learning. *Bull. Electr. Eng. Inform.* **2024**, *13*, 628–637. [[CrossRef](#)]
15. Eid, A.M.; Soudan, B.; Nassif, A.B.; Injadat, M. Comparative study of ML models for IIoT intrusion detection: impact of data preprocessing and balancing. *Neural Comput. Appl.* **2024**, *36*, 6955–6972. [[CrossRef](#)]
16. Khushi, M.; Shaukat, K.; Alam, T.M.; Hameed, I.A.; Uddin, S.; Luo, S.; Yang, X.; Reyes, M.C. A comparative performance analysis of data resampling methods on imbalance medical data. *IEEE Access* **2021**, *9*, 109960–109975. [[CrossRef](#)]
17. Kumar, V.; Lalotra, G.S.; Sasikala, P.; Rajput, D.S.; Kaluri, R.; Lakshmana, K.; Shorfuzzaman, M.; Alsufyani, A.; Uddin, M. Addressing binary classification over class imbalanced clinical datasets using computationally intelligent techniques. *Healthcare* **2022**, *10*, 1293. [[CrossRef](#)] [[PubMed](#)]
18. Mooijman, P.; Catal, C.; Tekinerdogan, B.; Lommen, A.; Blokland, M. The effects of data balancing approaches: A case study. *Appl. Soft Comput.* **2023**, *132*, 109853. [[CrossRef](#)]
19. Mohammed, R.; Rawashdeh, J.; Abdullah, M. Machine learning with oversampling and undersampling techniques: Overview study and experimental results. In Proceedings of the 2020 11th International Conference on Information and Communication Systems (ICICS), Irbid, Jordan, 7–9 April 2020; pp. 243–248.
20. Kubus, M. Evaluation of resampling methods in the class unbalance problem. *Econometrics* **2020**, *24*, 39–50. [[CrossRef](#)]
21. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [[CrossRef](#)]
22. Menardi, G.; Torelli, N. Training and assessing classification rules with imbalanced data. *Data Min. Knowl. Discov.* **2014**, *28*, 92–122. [[CrossRef](#)]
23. Tsou, C.S.; Liou, C.; Cheng, L.; Zhou, H. Quality prediction through machine learning for the inspection and manufacturing process of blood glucose test strips. *Cogent Eng.* **2022**, *9*, 2083475. [[CrossRef](#)]
24. Mukherjee, M.; Khushi, M. SMOTE-ENC: A novel SMOTE-based method to generate synthetic data for nominal and continuous features. *Appl. Syst. Innov.* **2021**, *4*, 18. [[CrossRef](#)]
25. Hall, M.A. Correlation-Based Feature Selection for Machine Learning. Ph.D. Thesis, The University of Waikato, Hamilton, New Zealand, 1999.
26. Akande, K.O.; Owolabi, T.O.; Olatunji, S.O. Investigating the effect of correlation-based feature selection on the performance of neural network in reservoir characterization. *J. Nat. Gas Sci. Eng.* **2015**, *27*, 98–108.
27. Bravo, F.P.; García, A.A.D.B.; Gallego, M.M.; Veiga, A.B.G.; Ruiz, M.; Peral, A.G.; Ayala, J.L. Prediction of patient’s response to OnabotulinumtoxinA treatment for migraine. *Heliyon* **2019**, *5*, e01043. [[CrossRef](#)]
28. Mukherjee, M.; Khushi, M. GitHub—Mimimkh/SMOTE-ENC-Code: A New Smote Method for Dataset with Continuous and Multi-Level Categorical Features—github.com. 2021. Available online: <https://github.com/Mimimkh/SMOTE-ENC-code> (accessed on 20 September 2024).
29. Friedman, N.; Geiger, D.; Goldszmidt, M. Bayesian network classifiers. *Mach. Learn.* **1997**, *29*, 131–163. [[CrossRef](#)]
30. Pazzani, M.J. Constructive induction of Cartesian product attributes. In *Feature Extraction, Construction and Selection: A Data Mining Perspective*; Springer: Berlin/Heidelberg, Germany, 1998; pp. 341–354.
31. Keogh, E.J.; Pazzani, M.J. Learning the structure of augmented Bayesian classifiers. *Int. J. Artif. Intell. Tools* **2002**, *11*, 587–601. [[CrossRef](#)]
32. Webb, G.I.; Boughton, J.R.; Wang, Z. Averaged One-Dependence Estimators: Preliminary Results. In Proceedings of the AusDM, Canberra, Australia, 2 December 2002; pp. 65–74.
33. Webb, G.I.; Boughton, J.R.; Wang, Z. Not so naive Bayes: Aggregating one-dependence estimators. *Mach. Learn.* **2005**, *58*, 5–24. [[CrossRef](#)]

34. Dash, D.; Cooper, G.F. Exact model averaging with naive Bayesian classifiers. In Proceedings of the ICML, San Francisco, CA, USA, 8–12 July 2002; pp. 91–98.
35. Hall, M. A decision tree-based attribute weighting filter for naive Bayes. In Proceedings of the International Conference on Innovative Techniques and Applications of Artificial Intelligence, Cambridge, UK, 11–13 December 2006; Springer: Berlin/Heidelberg, Germany, 2006; pp. 59–70.
36. Sahami, M. Learning Limited Dependence Bayesian Classifiers. In Proceedings of the KDD, Portland, OR, USA, 2–4 August 1996; pp. 335–338.
37. Jiang, L.; Zhang, L.; Li, C.; Wu, J. A correlation-based feature weighting filter for naive Bayes. *IEEE Trans. Knowl. Data Eng.* **2018**, *31*, 201–213. [[CrossRef](#)]
38. Kirsten, L. GitHub—CBFW_Naive_Bayes: Python implementation of “A Correlation-Based Feature Weighting Filter for Naive Bayes”—github.com. 2022. Available online: https://github.com/LucasKirsten/CBFW_Naive_Bayes/tree/master (accessed on 10 September 2024).
39. CrossValidated. ROSE and SMOTE Oversampling Methods. 2019. Available online: <https://shorturl.at/dIEQW> (accessed on 8 May 2024).

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.