Scientific Paper / Artículo Científico

UNIVERSIDAD POLITÉCNICA SALESIANA ECUADOR
INGENIUS
Revista de Ciencia y Tecnología
https://doi.org/10.17163/ings.n25.2021.02
pISSN: 1390-650X / eISSN: 1390-860X

Impact of oversampling algorithms in the classification of Guillain-Barré syndrome main subtypes

Impacto de los algoritmos de sobremuestreo en la clasificación de subtipos principales del síndrome de Guillain-Barré

Manuel Torres-Vásquez[1,2], José Hernández-Torruco[1],
Betania Hernández-Ocaña[1], Oscar Chávez-Bosquez[1*]

Received: 15-05-2020, Reviewed: 22-07-2020, Accepted after review: 25-09-2020

Resumen

Guillain-Barré Syndrome (GBS) is a neurological disorder where the body's immune system attacks the peripheral nervous system. This disease evolves rapidly and is the most frequent cause of paralysis of the body. There are four variants of GBS: Acute Inflammatory Demyelinating Polyneuropathy, Acute Motor Axonal Neuropathy, Acute Sensory Axial Neuropathy, and Miller-Fisher Syndrome. Identifying the GBS subtype that the patient has is decisive because the treatment is different for each subtype. The objective of this study was to determine which oversampling algorithm improves classifier performance. In addition, to determine whether balancing the data improves the performance of the predictive models. Three oversampling methods (ROS, SMOTE, and ADASYN) were applied to the minority class. Three classifiers (C4.5, SVM and JRip) were used.

Abstract

El síndrome de Guillain-Barré es un trastorno neurológico donde el sistema inmune del cuerpo ataca al sistema nervioso periférico. Esta enfermedad es de rápida evolución y es la causa más frecuente de parálisis del cuerpo. Existen cuatro variantes de SGB: polineuropatía desmielinizante inflamatoria aguda, neuropatía axonal motora aguda, neuropatía axonal sensorial aguda y síndrome de Miller-Fisher. Identificar el subtipo de SGB que el paciente contrajo es determinante debido a que el tratamiento es diferente para cada subtipo. El objetivo de este estudio fue determinar cuál algoritmo de sobremuestreo mejora el rendimiento de los clasificadores. Además, determinar si balancear los datos mejoran el rendimiento de los modelos predictivos. Aplicamos tres métodos de sobremuestro (ROS, SMOTE y ADASYN) a la clase minoritaria, utilizamos tres clasificadores (C4.5, SVM y JRip).

[1,*] División Académica de Ciencias y Tecnologías de la Información, Universidad Juárez Autónoma de Tabasco, Cunduacán, Tabasco, México. Corresponding author ✉: oscar.chavez@ujat.mx.
https://orcid.org/0000-0001-8475-0914 https://orcid.org/0000-0003-3146-9349
https://orcid.org/0000-0001-5700-7615 https://orcid.org/0000-0002-0324-9886

[2] Tecnológico Nacional de México campus Centla, División Sistemas Computacionales, Frontera, Centla, Tabasco, México.

Suggested citation: Torres-Vásquez, M.; Hernández-Torruco, J.; Hernández-Ocaña, B. and Chávez-Bosquez, O. (2021). «Impact of oversampling algorithms in the classification of Guillain-Barré syndrome main subtypes». Ingenius. N.° 25, (january-june). pp. 20-31. doi: https://doi.org/10.17163/ings.n25.2021.02.

The performance of the models was obtained using the ROC curve. Results show that balancing the dataset improves the performance of the predictive models. The SMOTE Algorithm was the best balancing method, in combination with the classifier JRip for OVO and the classifier C4.5 for OVA.

Keywords: ADASYN, Classifiers, Unbalance, ROS, SMOTE, Wilcoxon.

El rendimiento de los modelos se obtuvo mediante la curva ROC. Los resultados muestran que balancear el *dataset* mejora el rendimiento de los modelos predictivos. El algoritmo SMOTE fue el mejor método de balanceo en combinación con el clasificador JRip para OVO y el clasificador C4.5 para OVA.

Palabras clave: ADASYN, clasificadores, desbalanceo, ROS, SMOTE, Wilcoxon.

1. Introduction

The Guillain-Barré Syndrome (GBS) is defined as an autoimmune polyradiculoneuropathy and is the most frequent cause of acute generalized paralysis [1]. The GBS occurs when the immune system attacks part of the peripheral nervous system. This disease evolves rapidly and is characterized by weakness of the legs which further advances to the arms, "ascending paralysis". The initial symptoms are muscle weakness and tingling in the extremities. The severe cases require mechanical ventilation. The cause is unknown, but two thirds of the cases precede to a respiratory infection or acute gastroenteritis. It has been recently associated to the Zika virus. The GBS affects between 0.4 and 2.4 cases per 100,000 inhabitants/year. It appears at any age, but it often shows a higher frequency in people between 50 and 80 years old. It is slightly more frequent in men than in women. It has a mortality rate between 2% and 8%. Most people eventually recover completely when the disease is mild or moderate, and in other cases there may remain harms in the nervous system for long time or even permanently [2]. Electrophysiological and nerve conduction studies determine the tests for diagnosing GBS. There are four main subtypes of GBS:

- Acute Inflammatory Demyelinating Polyneuropathy **(AIDP)**.
- Acute Motor Axonal Neuropathy **(AMAN)**.
- Acute Sensory Axial Neuropathy a **(AMSAN)**.
- Miller-Fisher Syndrome **(MF)**.

The recovery of the patient largely depends on the prompt identification of the subtype of GBS. Each subtype should be treated in a different manner, and the treatment and costs vary according to the subtype developed by the patient. In severe cases that generate temporary or permanent immobility, the rehabilitation therapies are often long and costly generating psychological and economic implications to the sick person and to the relatives.

Machine Learning is a branch of Artificial Intelligence that uses different mathematical, statistical and optimization techniques, with the purpose of developing information analysis tools so that computers «learn» through examples [3]. At present, disciplines such as finance, oil, marketing, sales and health utilize automatic learning as technological tool to make predictions. Specifically, in the health area, an increasing number of models are being developed for diagnosing diseases such as cancer [4], [5], diabetes [6], [7], Parkinson [8] and Alzheimer [9], with excellent results.

Classification algorithms are in charge of analyzing the data provided and determining the patients that are healthy and the ones that are sick. However, one of the most common problems in medical diagnosis is the disproportionality of cases. In real life there are more healthy patients than sick patients. For example, if it is desired to diagnose patients with diabetes, it will be found that a larger number of people are healthy and a smaller number are sick with diabetes. This disproportionality in the data is known as data unbalancing. There are two types of unbalancing: binary and multiclass unbalancing. Binary unbalancing occurs when in a dataset of two classes, one of the classes has a larger number of data (majority class) with respect to the other class (minority class). On the other hand, multiclass unbalancing occurs when the dataset comprises more than two classes, and the data distribution is unequal for each of the classes [10].

Data unbalancing may affect the result of the classifiers since it tends to bias the results towards the majority class (healthy patients). The standard classification algorithms are built for balanced data, i.e., the same number of healthy and sick cases. For example, for the case of patients with diabetes, the classifier will ignore the patients with diabetes and will only take into account healthy patients. The problem is that it is desired to determine sick patients and not the healthy ones. For this reason, it is necessary to use techniques that help balancing the data.

In the specialized literature there are three techniques most commonly used to overcome the problem of data unbalancing [11].

- **At the data level.** This technique adds or eliminates data to the class, until balancing the dataset. This technique is also known as sampling and is divided in three groups:
  - Oversampling: consists in adding data to the minority class until reaching balance with the majority class.
  - Downsampling: consists in eliminating data from the majority class until reaching equilibrium with the minority class.
  - Hybrid: this technique combines oversampling and downsampling simultaneously, to reach a better balance between classes.
- **At the algorithm level.** They adapt or create classification algorithms to reinforce the prediction of the class.
- **Sensitive cost.** considers the costs associated with the incorrect classification of the samples. It uses different cost matrices that describe the costs of incorrectly classifying any particular data example.

The technique at the data level is one of the most popular because it is independent of the classifier used, and besides the data are treated before being used by the classifier. The oversampling technique is the most commonly used since it adds data to the minority class. There are different oversampling techniques that generate data, yielding good results with respect to downsampling which may eliminate important data and affect the result of the classifier [12].

On the other hand, besides data unbalancing, the distribution of the instances affects the results of the classifiers [13]. There are techniques that add synthetic data to the minority class and locate them in strategic places to resolve the unbalancing problem and the position of the instances.

The objective of this study was twofold. The first was identifying which of the three oversampling algorithms used to balance the original GBS dataset improves the results of the classification algorithm. The second objective was to establish if balancing the data improves the performance of the predictive models created with balanced data, with respect to models created with unbalanced data. For this purpose, Wilcoxon statistical test is utilized to know if there is a statistically significant difference between such models. At present, there are no studies in the specialized literature to identify the main subtypes of the GBS using Automatic Learning algorithms. In previous studies [14], [15], predictive models were created using the original unbalanced dataset. In this experimental study, the training subsets are balanced using three oversampling techniques (ROS, SMOTE and ADASYN). Results demonstrate that balancing the data improves the performance of the predictive models. A performance of 90% was achieved in some cases.

For this study, two binarization techniques (OVO and OVA) were first used to create 10 binary subsets. Then, the subsets were divided in training sets with 66% of the data, and test sets with 33% of the data. Once the training data were obtained, three balancing methods (ROS, SMOTE and ADASYN) were applied to oversample the minority class and balance it with the majority class. Once the data were balanced, three classification algorithms were applied with different approaches: C4.5 (decision tree), SVM (Support Vector Machine), JRip (Ripper). The performance of the predictive models was determined using the Area Under the Curve (AUC) of the ROC Curve. The results of the predictive models are the average of the AUC for 60 runs. At last, Wilcoxon test is applied to the models created with balanced data that outperformed the models created with unbalanced data, to know if there is a statistically significant difference between such models.

2. Materials and Methods

2.1. Dataset

The dataset used in this study is a collection of 129 patients diagnosed with GBS. One of the 4 main subtypes of GBS was identified to each of these patients. Table 1 shows the main features of the dataset.

Table 1. Features of the dataset

Characteristic	**Value**
Number of classes	4
Number of instances	129
Number of attributes	16
Class 1 Instances (AIDP)	20
Class 2 Instances (AMAN)	37
Class 3 Instances (AMSAN)	59
Class 4 Instances (MF)	13

This information was obtained through the National Institute of Neurology and Neurosurgery of the City of Mexico (Instituto Nacional de Neurología y Neurocirugía de la Ciudad de México). The original dataset comprises 356 variables. In a previous paper 16 variables were identified as the most relevant ones [16]. The first 4 variables were clinical, and the following 14 belong to the nerve conduction test. The variables used in the experiments are shown in the following:

v22: Symmetry (in weakness)
v29: Affectation of extraocular muscles
v30: Ptosis
v31: Cerebellar implication
v63: Amplitude of the left median motor nerve
v106: Area under the curve of the left ulnar motor nerve
v120: Area under the curve of the right ulnar motor nerve
v130: Amplitude of the left tibial motor nerve
v141: Amplitude of the right tibial motor nerve
v161: Area under the curve of the right peroneal motor nerve
v172: Amplitude of the left median sensory nerve
v177: Amplitude of the right median sensory nerve
v178: Area under the curve of the right median sensory nerve
v186: Latency of the right ulnar sensory nerve
v187: Amplitude of the right ulnar sensory nerve
v198: Area under the curve of the right sural sensory nerve

2.2. Automatic learning algorithms

2.2.1. Oversampling algorithms

Oversampling algorithms are a technique at the data level that add data to the minority class, with the purpose of balancing the unbalanced dataset. There are diverse algorithms to oversample the classes. Three techniques that generate instances with different approaches were used for this study:

1. The Random Oversampling (ROS) Algorithm obtains a random sample from instances of the minority class and makes a copy of them. The duplicated instances are placed randomly in the dataset. ROS is a non-heuristic method whose objective is to balance the minority class with the majority class [17].

2. The Synthetic Minority Oversampling Technique (SMOTE) oversamples the minority class generating synthetic instances with the purpose of balancing the minority class with the majority class [18]. The new synthetic instances are generated through interpolation between various instances of minority classes, based on the nearest neighbor rule. SMOTE performs this procedure in the «feature space». The procedure to generate synthetic data is the following: (a) Determine the oversampling percentage necessary to be generated. (b) In order to generate the synthetic objects, carry out the following procedure: (b1) Randomly select an instance of a minority class. (b2) Randomly choose its k-nearest neighbors according to the Euclidean distance. (b3) Take the difference between the feature vector and each of the selected neighbors. (b4) Multiply the difference times a random number between 0 and 1. (b5) Add this last value to the original value of the sample. (b6) Return the synthetic sample. (c) The new synthetic sample will be placed between the instance originally selected and each of the k-nearest neighbors.

   The main difference between SMOTE and ROS is that ROS duplicates data from the minority class and adds them randomly. SMOTE generates synthetic data and places them in a neighborhood of the minority class.

3. The Adaptive Synthetic Sampling Approach for Imbalanced Learning (ADASYN) is an extension of SMOTE. ADASYN has two objectives: the first is creating synthetic instances through linear interpolation between the instances of the minority class, to reduce the imbalance of the minority class with the majority class of the dataset. The second objective that makes ADASYN different with respect to SMOTE is that the data generated adaptively changes the decision boundary adding data in the zone of the minority class difficult to learn compared to the data of the minority class which are easy to learn, through a density distribution. ADASYN seeks to give more weight to the data of the minority class which are difficult to learn [19].

2.2.2. Classification algorithms

Three classification algorithms that determine their results through different approaches were utilized. The objective is contrasting the results of each of them:

1. Decision tree (C4.5): Is a supervised learning algorithm in which each branch node represents a choice among various options, and each leaf node represents a decision. The classification technique is performed by means of division criteria, with a structure of inverted tree, similar to a flowchart. It handles continuous and discrete characteristics. It has high precision, stability, is fast, easy to interpret and robust in the presence of noise. C4.5 bases its results in a hierarchical and inductive learning manner, i.e., in the discovery of patterns from examples [20].

2. Support Vector Machine (SVM): Is a supervised learning algorithm which is employed for binary classification. It belongs to the family of linear classifiers, i.e., the original data are resized by means of a mathematical function to search for a linear separability between them. SVM is based on the concept of constructing an optimal hyperplane, i.e., it creates a straight line that separates the classes. The objective is to find the best hyperplane that divides the dataset and maximizes the margin between the classes [21].

3. Ripper (JRip): Is one of the most popular algorithms for classification problems, with a rule-based approach. The classes are examined in increasing size, and an initial set of rules is generated for the class using the reduced incremental error JRip (RIPPER). It proceeds treating all examples of a particular sense in the training data as one class, and finding a set of rules to cover all members in that class. Afterwards, it passes to the following class and does the same, repeating this procedure until all classes are covered [22].

2.3. Performance Measure

The performance of the classification algorithms is evaluated using the graph or curve of the Receiver Operating Characteristics (ROC), and the Area Under the Curve (AUC). The ROC curve measures how well are the predictions classified, as well as the quality of the model predictions [23]. The ROC curve is defined as the sensitivity, which is the rate of true positives shown in Equation 1. The 1-specificity is the rate of false positives, shown in Equation 2. For this experiment, it is used to identify among one of the GBS subtypes.

$$sensitivity = \frac{VP}{VP + FN} \tag{1}$$

$$1 - specificity = \frac{FP}{VN + FP} \qquad (2)$$

The Area Under the Curve (AUC) enables identifying a class. For example, recognizing if a patient suffers a particular disease or is healthy. In this performance measure, the values $\geq .900$ are considered excellent models. The values $\geq .700$ indicate that they are good models. However, values $\leq .500$ are considered bad models.

2.4. Binarization techniques

In classification problems it is common to find datasets that are constituted by more than two classes, which are known as multiclass datasets. Some classification algorithms are only capable of discriminating between two classes. For this reason, it is common to transform a multiclass problem in binary subproblems. Two binarization techniques are found in the literature; One-Vs-One (OVO) and One-Vs-All (OVA) [24].

The OVO technique divides a problem of n classes into $n(n-1)/2$ binary subproblems, forming all possible pairs of classes. The OVA technique takes a class as the minority class, and the remaining classes are combined to form the majority class. This procedure is performed n times according to the number of classes that constitute the dataset. The OVO and OVA binarization techniques are used to discriminate one class from the others. In medical diagnosis problems, they are used to identify a sick patient from other healthy patients. Figures 1 and 2 show the 4 subsets obtained with the OVA technique and the 6 subsets obtained with the OVO technique, from the original GBS dataset.

2.5. Validation

For each classification, the model is validated using the train-test evaluation. The dataset is divided in two subsets of data. The first is the training data, which are used to build the model. The second are called test data, which are maintained apart and are used to evaluate the model. It is employed $\frac{2}{3}$ of the dataset for training and $\frac{1}{3}$ for testing the model.

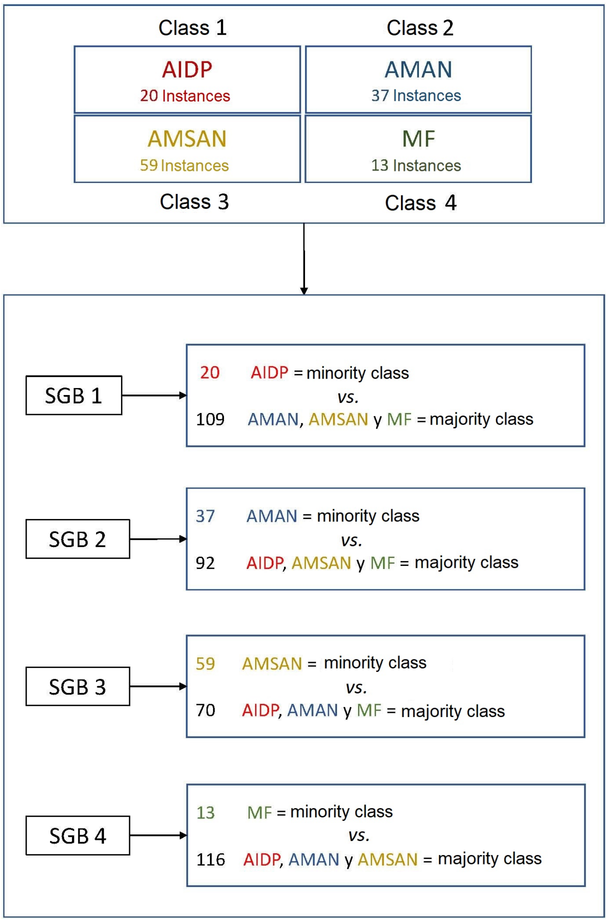

Figure 1. Binarization One-Vs-All (OVA)

3. Experimental procedure

As first step, the original unbalanced multiclass dataset was converted in two binary subproblems. Two different binarization techniques (OVO and OVA) are utilized. The difference between the two binarization techniques is the following: the OVO technique creates all possible combinations that can be formed with the n classes that constitute a dataset; on the other hand, OVA takes one class and converts it in minority class and the remaining classes are combined to form the majority class. OVA creates subsets depending on the total number of classes in the original dataset. The objective of creating binary subsets is that the balancing methods used in this study identify only two classes, the minority class which is oversampled until it is balanced with the majority class.

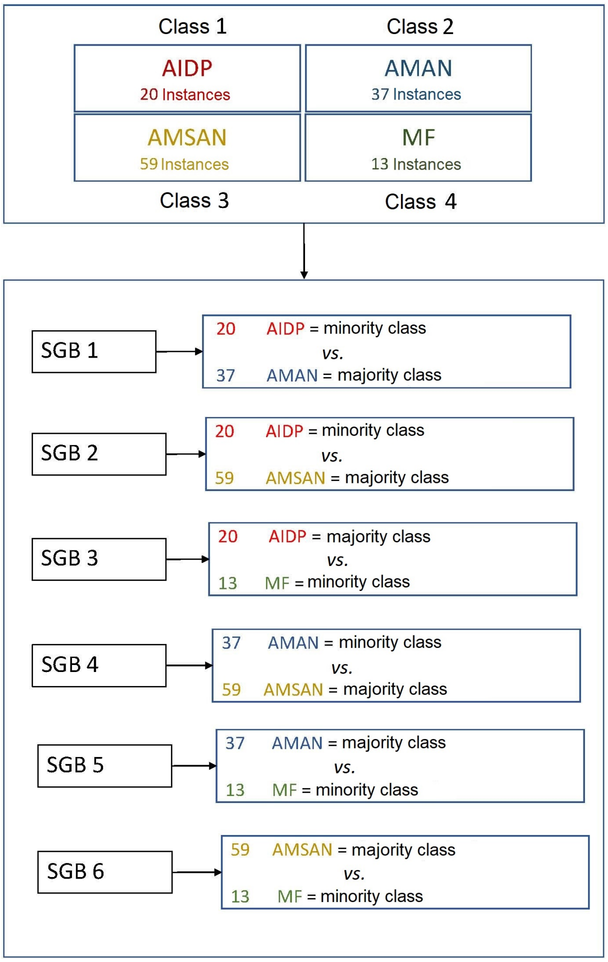

Figure 2. Binarization One-Vs-One (OVO)

Table 2. Subsets obtained with the OVA technique

Subset	Minority class	Majority class
SGB1	20	109
SGB2	37	92
SGB3	59	70
SGB4	13	116

A total of 10 binary datasets were obtained applying the two binarization techniques. Table 2 shows the 4 datasets created with the OVA technique, and Table Tabla 3 shows the 6 binary datasets created with the OVO technique. The first column shows the subsets obtained with binarization technique. The second column contains the number of instances that constitute the minority class. The third column shows the number of instances in the majority class. It may be observed that the OVA technique has a greater data unbalance between the minority class and the majority class, with respect to the OVO technique.

Aplicando las dos técnicas de binarización obtuvimos un total de 10 *datasets* binarios. En la Tabla 2 se muestran los 4 *dataset* creados con la técnica OVA y en la Tabla 3 se muestran los 6 *dataset* binarios creados con la técnica OVO. En la primera columna se muestran los subsets obtenidos con la técnica de binarización. En la segunda columna se observa el número de instancias que forman la clase minoritaria. La tercera columna muestra el número de instancias que integran la clase mayoritaria. Podemos observar que la técnica OVA tiene un mayor desbalanceo de datos entre la clase minoritaria y la clase mayoritaria respecto a la técnica OVO.

Table 3. Subsets obtained with the OVO technique

Subset	Minority class	Majority class
SGB1	20	37
SGB2	20	59
SGB3	13	20
SGB4	37	59
SGB5	13	37
SGB6	13	59

As a second step, each of the 10 subsets were divided. The data were split in 2/3 for training and the remaining 1/3 for testing. In the following, three oversampling algorithms (ROS, SMOTE and ADASYN) were applied to the minority class in the training data, until balancing it with the majority class. The testing data were used to measure the performance of the models obtained.

Table 4. Results of the balanced subsets applying oversampling methods to the minority class for OVA

Subset	Data A	Data B	Class a	Class b
SGB1	14	59	73	73
SGB2	25	37	61	62
SGB3	40	7	47	47
SGB4	9	69	78	78

Datos A: Unbalanced training data.
Datos B: Data generated with SMOTE, ROS and ADASYN.
Clase a: Balanced minority class.
Clase b: Original majority class.

Table 4 shows the 4 balanced subsets for the OVA technique. Table 5 shows the 6 balanced subsets for the OVO technique. The first column shows the subsets of the binarization technique. The second column shows the minority class with the number of instances that constitute it. The third column shows the number of instances that were generated for each oversampling algorithm. Columns 4 and 5 show the balanced minority class and majority class, respectively.

The next step was obtaining the predictive models applying three classification algorithms (C4.5, SVM and JRip) to the 10 balanced subsets. It were conducted 60 independent runs calculating the Area Under the Curve (AUC) for the 10 subsets. The predictive models are the result of the average of the AUCs for the 60 runs. On the other hand, the same procedure was carried out using the unbalanced subsets to obtain predictive models with unbalanced data.

Table 5. Results of the balanced subsets applying oversampling methods to the minority class for OVO

Subset	Data A	Data B	Class a	Class b
SGB1	14	9	23	23
SGB2	14	26	40	40
SGB3	9	5	14	14
SGB4	25	15	40	40
SGB5	9	16	25	25
SGB6	9	31	40	40

Datos A: Unbalanced training data.
Datos B: Data generated with SMOTE, ROS and ADASYN.
Clase a: Balanced minority class.
Clase b: Original majority class.

The last step was to compare the performance of the models obtained with balanced data, with the models obtained with unbalanced data. The Wilcoxon statistical test was used to know if there is a statistically significant difference between the models, provided that balanced models have outperformed unbalanced models. A significance value 0.05 was utilized.

The experiments were conducted in the R software, designed for statistical analysis. The R Studio version 1.2.1335 was utilized as integrated development environment. The packages used for balancing the data were: unbalanced for the ROS algorithm [25], DMwR for the SMOTE algorithm [26] and UBL for the ADASYN algorithm [27]. RWeka 0.4-39 [28] was used for the classification algorithms C4.5 and JRip, and 071 1.7-0 [29] was used for the SVM classifier.

The linear SVM classifier was optimized through the tune function, assigning values of 0.001, 0.01, 0.1, 1, 10, 50, 80, 100 for the parameter C. The JRip and C45 classifiers do not require optimization of hyperparameters.

4. Results and Discussion

Tables 6 and 9 show the results of the predictive models obtained. Three balancing methods (ROS, SMOTE and ADASYN) were applied. Six unbalanced subsets obtained were oversampled with the OVO binarization technique, and four subsets with the OVA binarization technique. Each value is the average of the results obtained through 60 runs. The classifiers C4.5, SVM and JRip were applied once the training set was balanced. The models were evaluated using the ROC metrics. The Wilcoxon statistical test was applied to the balanced models against the unbalanced models, when the balanced models outperformed the unbalanced models, with the objective of knowing if the performance of the balanced models obtained a statistically significant difference.

The structure of the Tables is the following: the first column shows the subsets obtained by means of the OVO and OVA binarization techniques, the GBS subtypes that constitute it, as well as the number of instances for each subtype. The second column shows the three classifiers used to obtain the predictive models for each subset. The third column shows the results of the predictive models using unbalanced data. Columns 4, 5 and 6 show the models obtained using balanced data applying three oversampling techniques (ROS, SMOTE and ADASYN). It is also observed that the values in bold letter are the predictive models which, besides outperforming unbalanced models, obtained a statistically significant difference. Table 6 shows the results of the 72 predictive models obtained using the OVO binarization technique.

Of these models, 18 were created with unbalanced data and 54 were obtained using balanced data applying the three oversampling methods. It was found that 32 balanced models could not outperform the unbalanced models. Other 15 balanced models outperformed the unbalanced models, but no statistically significant difference was found. On the other hand, 7 balanced models outperformed the unbalanced models, and in addition they had a statistically significant difference.

The best results were achieved with the subset GBS6, obtaining 3 models with statistically significant difference. On the other hand, the subsets GBS2 and GBS4 had 2 models each with statistically significant difference. The subsets GBS1, GBS3 and GBS5 exhibited the worst performance with respect to the unbalanced models, since a statistically significant difference was not found in any of them.

With respect to the balancing methods, Table 7 shows the results of the ranking obtained for each method. These results were obtained assigning a position to each method depending on its performance with each subset. For every row, a value is assigned to each oversampling method. In the first row, a value of 1 is assigned to SMOTE, since it obtained the best performance. A value of 2 was assigned to ROS since it obtained the second-best performance, and finally the value of 3 is assigned to ADASYN because it was the method with the worst performance. This operation is performed for every row. Subsequently, all the values for each method are added and divided by the number of rows to obtain the average. For example, SMOTE obtained the first place 5 times, the second place 6 times, the third place 5 times and the fourth place 2 times. The sum of these values is 40, which is divided by the number of rows in the table, 18 for this case. The result is 2.222, which holds number 1 in the ranking [30] because it is the lowest average.

For OVO, the SMOTE algorithm was the balancing method with the best performance, with an average

Table 6. Table of results of the predictive models applying ROS, SMOTE and ADASYN to oversample the minority class.

Subset	**Classifier**	**Unbalanced data**	**Balancing applying ROS**	**Balancing applying SMOTE**	**Balancing applying ADASYN**
GBS1	C4.5	0.9604	0.9514	0.9576	0.9292
AIDP-AMAN	SVM	0.9576	0.9465	0.9618	0.9486
20-37	JRip	0.9563	0.9507	0.9403	0.9396
GBS2	C4.5	0.8585	0.8160	0.8551	0.8529
AIDP-AMSAN	SVM	0.8472	0.8306	0.8333	0.8484
20-59	JRip	0.8260	0.8178	**0.8549***	**0.8545***
GBS3	C4.5	0.8132	0.8111	0.7965	0.7854
AIDP-MF	SVM	0.6556	0.6340	0.6535	0.6792
20-13	JRip	0.8556	0.8493	0.7382	0.8396
GBS4	C4.5	0.9258	0.9093	0.9093	0.8897
AMAN-AMSAN	SVM	0.8760	0.8692	0.8827	0.8845
37-59	JRip	0.8782	**0.9059***	**0.9065***	0.8877
GBS5	C4.5	0.8736	0.8826	0.8868	0.8486
AMAN-MF	SVM	0.8806	0.8729	0.8847	0.8910
37-13	JRip	0.8854	0.8958	0.8889	0.8833
GBS6	C4.5	0.8007	**0.8411***	0.7839	0.8209
AMSAN-MF	SVM	0.7089	**0.7600***	0.7534	**0.7746***
59-13	JRip	0.8580	0.8561	0.8720	0.8264

The values are the average of 60 runs of the ROC curves using OVO.

score of 2.2222. The algorithms ADASYN and ROS held the second place, because both obtained the same average score of 2.7222. With respect to the classifiers, Table 8 shows that the JRip classifier obtained the best performance with an average score of 1.6667. The C4.5 classifier obtained the second place with an average score of 1.8333. At last, the SVM classifier obtained the worst performance with an average score of 2.500.

Table 7. Results of the ranking by balancing method for OVO

Method	**Ranking**	**Average score**
SMOTE	1	2.2222
ADASYN	2	2.7222
ROS	2	2.7222

Table 9 shows the results of 48 predictive models, obtained using the OVA binarization technique. Among these, 12 models were created with unbalanced data and 36 were obtained using balanced data applying three oversampling methods. It was found that 15 balanced models could not outperform the unbalanced models; 9 balanced models outperformed the unbalanced models, but no statistically significant difference was found. On the other hand, 12 balanced models outperformed the unbalanced ones, and besides had a statistically significant difference.

Table 8. Results of the ranking by classifier for OVO

Classifier	***Ranking***	**Average score**
JRip	1	1.6667
C4.5	2	1.8333
SVM	3	2.5000

The best performances were obtained with the subsets GBS1 and GBS4. In the subset GBS1, 8 balanced models improved the unbalanced models, of which 5 models obtained a statistically significant difference. In the GBS4 subset, 6 balanced models outperformed the unbalanced models, of which 5 models obtained a statistically significant difference. With the subset GBS2, 5 balanced models outperformed the unbalanced models, but only 2 models obtained a statistically significant difference. In the SGB3 subset it performed the worst. Only 3 balanced models exceeded the unbalanced data, without finding a statistically significant difference.

Table 10 shows the results of the ranking for the balancing methods applying the OVA binarization technique. The SMOTE algorithm obtained the best performance with an average score of 1.9167, holding the first place. The ADASYN algorithm obtained the second place, with an average score of 2.1667. At

Table 9. Table of results of the predictive models applying ROS, SMOTE and ADASYN to oversample the minority class

Subset	**Classifier**	**Unbalanced data**	**Balancing applying ROS**	**Balancing applying SMOTE**	**Balancing applying ADASYN**
GBS1	C4.5	0.7894	0.7873	0.8042	**0.8162***
AIDP-ALL	SVM	0.7162	0.7262	**0.7750***	**0.7722***
20-109	JRip	0.7826	0.7921	**0.8102***	**0.8215***
GBS2	C4.5	0.8729	0.8653	0.8900	0.8949
AMAN-ALL	SVM	0.8564	0.8489	0.8490	**0.8871***
37-92	JRip	0.8608	0.8513	0.8699	**0.8949***
GBS3	C4.5	0.8723	0.8455	0.8795	0.8493
AMSAN-ALL	SVM	0.7948	0.7982	0.7881	0.7827
59-70	JRip	0.8470	0.8358	0.8442	0.8536
GBS4	C4.5	0.7808	0.7806	**0.8951***	0.7331
MF-ALL	SVM	0.6464	**0.7590***	**0.7516***	**0.6991***
13-116	JRip	0.8319	0.8440	**0.8826***	0.7882

The values are the average of 60 runs of the ROC curves using OVA.

last, ROS was the balancing algorithm with the worst performance, holding the third place with an average score of 3.0833.

With respect to the classifiers, the results of the ranking are observed in Table 11. The C4.5 classifier obtained the first place with an average score of 1.2500. The JRip classifier ended up in the second place, with an average score of 1.500. The third place was obtained by the SVM classifier, with an average score of 2.7500.

Table 10. Results of the ranking by balancing method for OVA

Method	***Ranking***	**Average score**
SMOTE	1	1.9167
ADASYN	2	2.1667
ROS	3	3.0833

The OVA binarization technique obtained the best results. A total of 36 predictive models were obtained with balanced data. Among these, 12 predictive models obtained a statistically significant difference. The SMOTE algorithm was the balancing method with the best results. The JRip classifier was the best algorithm, according to the ranking.

Table 11. Results of the ranking by classifier for OVA

Method	***Ranking***	**Average score**
C4.5	1	1.2500
JRip	2	1.5000
SVM	3	2.7500

The OVA binarization technique obtained the worst performance. A total of 54 predictive models were obtained with balanced data, of which 7 predictive models achieved a statistically significant difference. The ADASYN algorithm obtained the best performance as oversampling method. The C4.5 classifier obtained the best performance, since it obtained the lowest average score.

5. Conclusions

In this research work, three oversampling algorithms (ROS, SMOTE and ADASYN) were explored, with the objective of knowing which obtains the best performance; in addition, to know if balancing the original dataset improves the performance of the predictive models obtained with unbalanced data. These experiments were conducted with a real dataset of patients diagnosed with some subtype of GBS. Initially, binary subsets were created applying two techniques (OVO and OVA) to the original dataset. It was obtained 10 subsets divided in: 6 subsets with the OVO technique, and 4 subsets with the OVA technique. Each subset was split into 66% of the data for training and 34% of the data for testing. The minority classes of the training subsets were oversampled applying ROS, SMOTE and ADASYN, with the purpose of balancing the minority class with the majority class. Once the subsets were balanced, three classifiers were applied: C4.5, JRip and SVM. The results are the average of 60 runs of the ROC curve. Wilcoxon test was applied to the predictive models obtained with balanced data that outperformed the models with unbalanced data, to know if there is a statistically significant difference between them.

The OVA binarization technique obtained the best result compared to the OVO technique. Applying the OVA technique, 36 predictive models were obtained with balanced data, of which 12 had a statistically significant difference. The best algorithm for balancing the data was SMOTE with respect to ROS and ADASYN. The SMOTE algorithm improved the performance of the predictive models according to their oversampling features. SMOTE adds instances to the minority class, extrapolating new instances instead of duplicating them, as it is done by the ROS algorithm. The ROS algorithm copies instances of the minority class and adds them randomly, duplicating information that may confuse the classifiers. On the other hand, ADASYN is a variant of SMOTE which adds instances to the minority class that are difficult to learn, specially the ones located in the decision boundary; this approach may not be enough information for the classifier to identify the classes and improve the result. The C4.5 classifier obtained the best performance according to the average score for OVO.

The results demonstrate that balancing the data improves the performance of the predictive models obtained with unbalanced data. On the other hand, using automatic learning algorithms in disease diagnosis problems is feasible, and may contribute to the identification of the GBS subtype that a patient gets. As future works, hybrid oversampling and downsampling techniques will be explored, in addition to the use of other classifiers.

References

[1] P. A. van Doorn, "Guillain-Barré syndrome," in *Dysimmune Neuropathies.* Elsevier, 2020, pp. 5–29. [Online]. Available: https://doi.org/10.1016/B978-0-12-814572-2.00002-9

[2] A. Tellería-Díaz and D. Calzada-Sierra, "Síndrome de Guillain-Barré," *Revista de Neurología*, vol. 34, no. 10, pp. 966–976, 2002. [Online]. Available: https://doi.org/10.33588/rn.3410.2001280

[3] E. Alpaydin, *Introduction to Machine Learning.* MIT press, 2020. [Online]. Available: https://bit.ly/2HvdROG

[4] J. A. Cruz and D. S. Wishart, "Applications of Machine Learning in cancer prediction and prognosis," *Cancer Informatics*, vol. 2, p. 117693510600200, jan 2006. [Online]. Available: https://doi.org/10.1177/117693510600200030

[5] A. R. Vaka, B. Soni, and S. R. K., "Breast cancer detection by leveraging Machine Learning," *ICT Express*, may 2020. [Online]. Available: https://doi.org/10.1016/j.icte.2020.04.009

[6] H. Kaur and V. Kumari, "Predictive modelling and analytics for diabetes using a machine learning approach," *Applied Computing and Informatics*, dec 2018. [Online]. Available: https://doi.org/10.1016/j.aci.2018.12.004

[7] N. P. Tigga and S. Garg, "Prediction of Type 2 Diabetes using Machine Learning classification methods," *Procedia Computer Science*, vol. 167, pp. 706–716, 2020. [Online]. Available: https://doi.org/10.1016/j.procs.2020.03.336

[8] Z. K. Senturk, "Early diagnosis of parkinson's disease using machine learning algorithms," *Medical Hypotheses*, vol. 138, p. 109603, may 2020. [Online]. Available: https://doi.org/10.1016/j.mehy.2020.109603

[9] A. Khan and S. Zubair, "An improved multimodal based Machine Learning approach for the prognosis of Alzheimer's disease," *Journal of King Saud University - Computer and Information Sciences*, apr 2020. [Online]. Available: https://doi.org/10.1016/j.jksuci.2020.04.004

[10] A. Fernández, S. García, M. Galar, R. C. Prati, B. Krawczyk, and F. Herrera, *Learning from Imbalanced Data Sets.* Springer International Publishing, 2018. [Online]. Available: https://doi.org/10.1007/978-3-319-98074-4

[11] G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, and G. Bing, "Learning from class-imbalanced data: Review of methods and applications," *Expert Systems with Applications*, vol. 73, pp. 220–239, may 2017. [Online]. Available: https://doi.org/10.1016/j.eswa.2016.12.035

[12] A. Fernández, S. García, F. Herrera, and N. V. Chawla, "SMOTE for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary," *Journal of Artificial Intelligence Research*, vol. 61, pp. 863–905, apr 2018. [Online]. Available: https://doi.org/10.1613/jair.1.11192

[13] K. Napierala and J. Stefanowski, "Types of minority class examples and their influence on learning classifiers from imbalanced data," *Journal of Intelligent Information Systems*, vol. 46, no. 3, pp. 563–597, jul 2015. [Online]. Available: https://doi.org/10.1007/s10844-015-0368-1

[14] J. Canul-Reich, J. Frausto-Solís, and J. Hernández-Torruco, "A predictive model for Guillain-Barré syndrome based on single learning algorithms," *Computational and Mathematical Methods in Medicine*, vol. 2017, pp. 1–9, 2017. [Online]. Available: https://doi.org/10.1155/2017/8424198

[15] J. Canul-Reich, J. Hernández-Torruco, O. Chávez-Bosquez, and B. Hernández-Ocaña, "A predictive model for Guillain-Barré syndrome based on ensemble methods," *Computational Intelligence and Neuroscience*, vol. 2018, pp. 1–10, 2018. [Online]. Available: https://doi.org/10.1155/2018/1576927

[16] J. Hernández-Torruco, J. Canul-Reich, J. Frausto-Solís, and J. J. Méndez-Castillo, "Feature selection for better identification of subtypes of Guillain-Barré syndrome," *Computational and Mathematical Methods in Medicine*, vol. 2014, pp. 1–9, 2014. [Online]. Available: https://doi.org/10.1155/2014/432109

[17] A. Fernández, S. del Río, N. V. Chawla, and F. Herrera, "An insight into imbalanced big data classification: Outcomes and challenges," *Complex & Intelligent Systems*, vol. 3, no. 2, pp. 105–120, 2017. [Online]. Available: https://doi.org/10.1007/s40747-017-0037-9

[18] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, jun 2002. [Online]. Available: https://doi.org/10.1613/jair.953

[19] H. He, Y. Bai, E. A. García, and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," in *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*. IEEE, jun 2008. [Online]. Available: https://doi.org/10.1109/IJCNN.2008.4633969

[20] S. Ruggieri, "Efficient C4.5 [classification algorithm]," *IEEE Transactions on Knowledge and Data Engineering*, vol. 14, no. 2, pp. 438–444, 2002. [Online]. Available: https://doi.org/10.1109/69.991727

[21] T. S. Furey, N. Cristianini, N. Duffy, D. W. Bednarski, M. Schummer, and D. Haussler, "Support Vector Machine classification and validation of cancer tissue samples using microarray expression data," *Bioinformatics*, vol. 16, no. 10, pp. 906–914, 2000. [Online]. Available: https://doi.org/10.1093/bioinformatics/16.10.906

[22] A. Rajput, R. P. Aharwal, M. Dubey, S. Saxena, and M. Raghuvanshi, "J48 and JRip rules for e-governance data," *International Journal of Computer Science and Security (IJCSS)*, vol. 5, no. 2, p. 201, 2011. [Online]. Available: https://bit.ly/3jt2jrY

[23] R. Kannan and V. Vasanthi, "Machine learning algorithms with ROC curve for predicting and diagnosing the heart disease," in *Soft Computing and Medical Bioinformatics*. Springer Singapore, jun 2018, pp. 63–72. [Online]. Available: https://doi.org/10.1007/978-981-13-0059-2_8

[24] A. Fernández, V. López, M. Galar, M. J. del Jesús, and F. Herrera, "Analysing the classification of imbalanced data-sets with multiple classes: Binarization techniques and ad-hoc approaches," *Knowledge-Based Systems*, vol. 42, pp. 97–110, apr 2013. [Online]. Available: https://doi.org/10.1016/j.knosys.2013.01.018

[25] A. D. Pozzolo, O. Caelen, and G. Bontempi, *unbalanced: Racing for Unbalanced Methods Selection*, 2015, R package version 2.0. [Online]. Available: https://doi.org/10.1007/978-3-642-41278-3_4

[26] L. Torgo, *Data Mining with R, learning with case studies*. Chapman and Hall/CRC, 2010. [Online]. Available: https://bit.ly/3jtkeyV

[27] P. Branco, R. P. Ribeiro, and L. Torgo, "UBL: an R package for utility-based learning," *CoRR*, vol. abs/1604.08079, 2016. [Online]. Available: https://bit.ly/35yeFtU

[28] I. H. Witten, E. Frank, M. A. Hall, and C. Pañ, *Data Mining, Practical Machine Learning Tools and Techniques*, Elsevier, Ed. Morgan Kaufmann, 2017. [Online]. Available: https://doi.org/10.1145/507338.507355

[29] D. Meyer, E. Dimitriadou, K. Hornik, A. Weingessel, and F. Leisch, *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*, 2018, R package version 1.7-0. [Online]. Available: https://bit.ly/3mm1d3s

[30] A. S. Hussein, T. Li, W. Y. Chubato, and K. Bashir, "A-SMOTE: A new preprocessing approach for highly imbalanced datasets by improving SMOTE," *International Journal of Computational Intelligence Systems*, 2019. [Online]. Available: https://bit.ly/3mhotiT