





The Use of Data Mining Techniques in the Diagnosis and Prevention of Cerebrovascular Accident (CVA)

Maria Adriana Ferreira da Silva   [Universidade Federal Rural do Semi-Árido | maria.silva78326@alunos.ufersa.edu.br]

Angélica Félix de Castro  [Universidade Federal Rural do Semi-Árido | angelica@ufersa.edu.br]

Isaac de Lima Oliveira Filho  [Universidade do Estado do Rio Grande do Norte | isaacol-iveira@uern.br]

Marcelino Pereira dos Santos Silva  [Universidade do Estado do Rio Grande do Norte | prof.marcelino@gmail.com]

 Universidade Federal Rural do Semi-Árido, Rua Francisco Mota Bairro, 572 - Pres. Costa e Silva, Mossoró - RN, 59625-900, Brazil.

Received: 13 December 2022 • Published: 12 November 2024

Abstract Over the years, there has been a rise in the occurrence of Cerebrovascular Accident (CVA) cases, due to the increase in the elderly population. Current data indicate that stroke is one of the leading causes of death and disability worldwide, affecting millions of people and leaving survivors with numerous sequelae, whether they are physical or mental. Many factors such as diabetes, smoking, high blood pressure, and others, favor the onset of stroke, which increases mortality rates, making it necessary to know these factors in order to contribute to early preventive measures. In this sense, the purpose of this article is to use six data mining algorithms with the objective of helping to identify and diagnose people prone to having a stroke based on risk factors and indicative signs. The algorithms used were: Decision Tree, K-Nearest Neighbors (K-NN), Multilayer Perceptron Neural Network (MLP), Support Vector Machine (SVM), Naive Bayes, and the Apriori algorithm. The results showed that the MLP and decision tree algorithms obtained the best results, indicating their use in intelligent solutions for this area.

Keywords: Stroke, Risk factors, Indicative signs, Data mining

1 Introduction

Stroke can be defined as an alteration of the blood flow in the brain and can be classified into three types: Ischemic vascular accident, Hemorrhagic vascular accident, and Transient ischemic attack, the first one being the most common type. Data indicate that stroke is one of the leading causes of death and disability worldwide [Campbell and Khatri, 2020].

In its initial phase, stroke can manifest indicative signs even before its occurrence. The signs that occur most frequently are: numbness on only one side of the body, mental confusion, difficulty speaking or understanding and others. Early identification of these signs is of utmost importance, as a very serious injury can cause the patient's instant death or leave disabilities that can significantly affect their life [Brasil, 2013].

Knowing the risk factors for stroke is also essential to prevent its occurrence and consequently reduce the costs that may come with rehabilitation and hospitalization. The main risk factors for stroke are modifiable and nonmodifiable. Modifiable risk factors are those that the person can control, such as: hypertension, dyslipidemia, diabetes, smoking, sedentary lifestyle, among others. Nonmodifiable risk factors are factors that cannot be controlled, such as: family history of stroke, age, race, sex, and others [Meschia *et al.*, 2014].

According to data from the World Stroke Organization, one in six people in the world will have a stroke during their

lifetime. Such data show the need to develop solutions aimed at improving the health and quality of life of these people, both in terms of rehabilitation and prevention and diagnosis of diseases and health promotion [Brasil, 2013].

In Brazil, although there has been a drop in mortality rates, stroke still represents the leading cause of death and disability in the country. In 2013, the annual incidence rate was 108 cases per 100,000 inhabitants with fatality within 30 days of 18.5% and at 12 months of 30.9%, with a recurrence rate of 15.9% [Brasil, 2013].

Therefore, in the present work, computational data mining techniques were used to assist in the identification and diagnosis of individuals prone to suffering a stroke using an international database. The results of this study can help in the early recognition of indicative signs and risk factors associated with stroke, as well as facilitate the initial care for the preservation of complications that may occur [Brasil, 2013].

In this study, the importance of addressing an increasingly challenging issue represented by the rising number of stroke cases, one of the leading causes of death and disability worldwide, is discussed. The analysis of risk factors and indicative signs underscores the urgency for preventive strategies. In this context, the study provides a significant contribution by applying six data mining algorithms to identify and assist in diagnosing individuals with a higher propensity for stroke. The promising results, particularly from the MLP and Decision Tree algorithms, suggest their potential for innovative solutions in this critical area of healthcare.

This article is organized as follows: In Section 2 the relevant concepts for this research are described; in Section 3 related works are reported; in Section 4 the materials and methods that involve the description of the methodology used are presented; Section 5 presents the results and discussions; and, finally, Section 6 presents the final considerations.

2 Data Mining Fundamentals

This Section is intended to present the theoretical foundation on the concepts and themes relevant to this work. Initially, the knowledge discovery process in a database is presented, followed by the presentation of the main algorithms aimed at the data mining process.

2.1 The Knowledge Discovery in Databases (KDD) Process

KDD is a process characterized by several steps, intending to identify understandable, valid, new, and potentially useful patterns from large data sets. The KDD process includes the following steps: pre-processing, data selection, data mining, evaluation of identified patterns, and visual representation of analysis results [Fayyad *et al.*, 1996]. According to Ferrari and Castro [2016] these five steps of the KDD process can be simplified into four main parts, namely:

- **Data management:** consists of organizing data, that is, quantitative or qualitative values referring to a set of items, which allows an efficient recovery of data.
- **Data preparation or pre-processing:** these are steps prior to mining that aim to prepare data for efficient and effective analysis. It includes processes aimed at cleaning, integrating, selecting or reducing and transforming data.
- **Data mining:** this process step consists of applying algorithms capable of extracting patterns from pre-processed data.
- **Knowledge assessment or validation:** is related to the evaluation of mining results aiming at identifying important knowledge that has been extracted and can be interpreted.

2.2 Algorithms for data mining

Data mining is an area characterized by the junction of different areas of knowledge, such as: database, statistics, image and signal processing, spatial data analysis, artificial intelligence, among others [Géron, 2019]. Several algorithms aimed at solving different real-world problems can be found in the literature. Among these algorithms, those based on supervised and unsupervised learning are the most common.

Supervised learning is based on a set of input data that includes the desired outputs, also called labels. In unsupervised learning, there is no presence of labeled data because its objective is to find similar groups and identify existing relationships between them Géron [2019]. In this work, five supervised learning algorithms will be used along with one for finding associations among the data and explaining such associations:

- **Apriori:** it is an algorithm used for mining association rules, based on the principle that any subset of frequent itemsets must be a frequent itemset. The algorithm uses the breadth-first search strategy, taking into account the frequent itemsets generated in the previous level. After its generation, a frequency test of these itemsets is performed, traversing the database again [Faceli *et al.*, 2011].
- **Decision Tree:** its structure is organized in the form of a tree, in which the highest node of the tree is known as the root node and the internal nodes correspond to the leaves. The path from the root to a leaf node corresponds to a classification rule, where each branch represents a test result and the leaf nodes represent classes or class distributions [Ferrari and Castro, 2016].
- **K-Nearest Neighbors (K-NN):** the nearest neighbors algorithm has variations defined by the number of neighbors to be considered. The k-nearest neighbors technique considers the k neighbors of the training set closest to the point looking for the nearest neighbors of the object to determine which class it belongs to [Faceli *et al.*, 2011].
- **Naive Bayes (NB):** it is a statistical algorithm based on Bayes' Theorem and on the assumption that attribute values are conditionally independent of their class [Faceli *et al.*, 2011].
- **Support Vector Machine (SVM):** it is an algorithm aimed at performing linear or non-linear classifications [Géron, 2019].
- **Multilayer Perceptron (MLP):** artificial Neural Networks are computer systems distributed in interconnected processing units. These units are known as artificial neurons. To solve nonlinearly separable problems using ANNs, the most used alternative is to add one or more intermediate layers. Multilayer perceptron (MLP) networks present this characteristic, as they consist of one or more intermediate layers of neurons and an output layer [Faceli *et al.*, 2011].

3 Related Works

In the literature, several solutions have been developed with the objective of detecting and predicting the diagnosis of stroke, using concepts of Artificial Intelligence, with emphasis on machine learning and data mining techniques. Some of these solutions are presented below.

Govindarajan *et al.* [2020] used several algorithms to classify stroke traits based on the patients' symptoms. Some of the algorithms used were: artificial neural networks, support vector machines, random forests, and others. Data were extracted from the patient registration forms - in total, information was collected from 507 patients. The results showed that among the tested algorithms, artificial neural networks trained with a gradient descent algorithm showed better classification accuracy, greater than 95% when compared to other tested algorithms.

Singh and Choudhary [2017] propose a stroke prediction approach using data from the Cardiovascular Health Study (CHS), which consists of 5888 samples. The C4.5 decision

tree algorithm was used for the feature selection process. The principal component analysis (PCA) algorithm was used for dimensionality reduction and the Back Propagation Neural Network algorithm was used for classification. The results showed that the method performed well.

Yu *et al.* [2019] described the design and implementation of a system for early detection of stroke in Koreans over 65 years old. It was carried out based on the National Institutes of Health (NIH). Data were collected between 2015 and 2017. With the results, it was possible to identify that the C4.5 decision tree algorithm performed well when compared to other machine learning algorithms.

Zhang *et al.* [2018] propose a stroke risk detection model, using machine learning and optimization methods. The study was based on a dataset from the biomedical tests and on the basic demographic characteristics of each patient. For the proposed model, 792 patients from a community hospital in Beijing were observed. Support vector machines (SVM) were combined with the Glow-Worm Swarm Optimization (GSO) algorithm based on the standard deviation of the resources. The dataset was divided into a test set and a validation set using the 10-Fold cross-validation strategy. The results showed that the model performed better in detecting stroke risk.

In a different perspective from the cited studies, this work aims to apply six data mining algorithms in an international dataset in a way that makes it possible to assist in the identification and diagnosis of stroke based on indicative signs and patient risk factors.

4 Methodology

The methodology used in this work is based on the stages of the KDD process proposed by Ferrari and Castro [2016] and was subdivided into five stages, namely: I) analysis of the information from the original database; II) pre-processing of data; III) existing relationship between attributes; IV) definition of experiments; and V) validation of the algorithms. Each of these steps is described below.

4.1 Data management

A database from The data set from the International Stroke Trial database [Sandercock *et al.*, 2011] will be used for this work, an international study that aimed to make individual patient data from the International Stroke Trial (IST) available for public use, one of the largest randomized trials ever performed on acute stroke to facilitate the planning of future studies and to allow additional secondary analyses. The International Stroke Trial (IST) includes 19,435 instances and 112 attributes.

In summary, the information present in the database consists of identifying whether the early administration of aspirin, heparin, both, or neither influenced the clinical course of a stroke. In addition to these, there is an attribute that allows identifying the presence or absence of the stroke. For this reason and also for having a large number of attributes with missing data, it was necessary to perform the pre-processing of the data. Thus, to carry out the pre-processing

and application of the algorithms, the Python programming language was used, making use of the scikit-learn, pandas, seaborn and matplotlib libraries, and the Collaboratory or Colab tool from Google as the Integrated Development Environment (IDE).

4.2 Preparation or pre-processing of data

Initially, the hospital code was removed since it was not relevant to the study. Then, the existence of duplicate values was verified by applying the duplicated method, so that no patient with duplicate information was identified, and it was not necessary to treat these cases. With the verification of missing values through the isnull function, attributes with more than 80% of missing values were identified. Among them, the existence of 36 attributes with a large portion of missing values (NaN) was observed. In this case, these attributes were excluded taking into account the large amount of missing information. After that, the remaining 76 attributes were analyzed in order to use only the attributes related to indicative signs and risk factors. The others were not used as they were not relevant to the present work. At the end of the analysis, 15 attributes were pre-selected.

Even after the removal of attributes with a large amount of missing data and keeping only the attributes that had relevance to the search, some data was still missing, such as the attributes RATRIAL and DNOSTRK. The RATRIAL attribute was related to atrial fibrillation and contained 984 empty instances and the target attribute of the study, and the "DNOSTRK" attribute was the classifying attribute that identified or not the presence of stroke and contained 26 empty instances. Such instances had to be excluded, given the importance of these values for stroke identification.

After selecting these attributes, two continuous attributes were identified (age (AGE) and systolic blood pressure (RSBP)). For these attributes, the values-counts function was used to verify the existence of non-standard values and none of them were identified. In addition, it was decided to categorize these attributes.

Assuming that the tendency for stroke in patients of advanced age is higher, it was decided to categorize age into 4 groups. To make such classification, art. 2 of the Child and Adolescent Statute, which considers a child to be a person under twelve years of age [Brasil, 1990]. The groups of young people, adults, and elderly people were based on the guidelines presented by Cerqueira e Francisco [2022] which consider young people, people between 12 and 19 years old, adults, people between 20 and 59 years old and elderly people, the ones who are 60 years of age or older.

For blood pressure (RSBP), according to de Saúde [2021] there are six groups to specify the degree of hypertension. Therefore, blood pressure was categorized based on these 6 predefined groups. The categories are presented below:

Age categorization:

- Children - From 0 to 11 years old
- Young people - From 12 to 19 years old
- Adults - From 20 to 59 years old
- Elderly people - Over 60 years old

Blood Pressure Categorization:

- Optimal - From 0 to 119 mmHg
- Normal - From 120 to 129 mmHg
- Pre-Hypertension - From 130 to 139 mmHg
- Stage 1 Hypertension - From 140 to 159 mmHg
- Stage 2 Hypertension - From 160 to 179 mmHg
- Stage 3 Hypertension - From 180 to 300 mmHg

At the end of these steps, all attributes of the dataset became categorical. Thus, to help in the application of algorithms where values are treated internally as real values, making it impossible to process text values, the categories were transformed into numerical values, using the replace function. After carrying out all the procedures mentioned, a dataset was obtained containing 18,425 instances and a total of 15 fully valued attributes, about 94.8% of the original base.

Finally, the presence of imbalanced data in the classification variable “DNOSTRK” was identified. According to Azank and Gurgel [2020], this imbalance can be defined by the small incidence of a category within a dataset (minority class) compared to the other categories (majority classes). In most cases, this results in much information about the most incident categories and little information about the minority ones, which can, in many cases, interfere in the results of the algorithms.

One way to remove the bias caused by the difference in the proportions of the categories is to manipulate the amount of data that are actually used, trying to equal the number of observations between the classes [Azank and Gurgel, 2020]. There are several ways to solve the problem of imbalanced data, among them: the NearMiss method, Undersampling, Smote and Oversampling. The NearMiss method is an algorithm that consists of reducing, at random, the examples of the majority class, selecting the examples based on the distance. Undersampling is characterized by randomly reducing the majority class examples. The Smote method consists of generating synthetic (non-duplicate) data of the minority class from its neighbors. And Oversampling is a method that consists of replicating random data from the minority class [Santana, 2020].

Therefore, these four methods were applied in order to identify which one presented the best result for the imbalance problem of the classifying variable in the data set. To identify which one presented the best results, the Random Forest algorithm was applied to validate each one of the methods, taking into account the parameters of accuracy, precision, recall, and F1-score. Thus, the NearMiss and Smote methods were the ones that presented the best results. However, we chose to use the NearMiss method, since the Smote method can harm the performance of the algorithms for minority classes, as it creates synthetic data for the minority class and can cause the problem of data overfitting. For the NearMiss method, the NearMiss variant is used to address class imbalance. When applying NearMiss with the fit_resample() method, a balanced dataset is obtained, where instances of minority classes are preserved while instances of majority classes are reduced. Table 1 presents a description of the resulting attributes after the execution of the pre-processing.

Table 1. Resulting attributes after preprocessing

Attribute	Description
SEX	Patient's sex (M = male; F = female)
AGE	Age group (Child, Youth, Adult and Elderly)
RSLEEP	Symptoms observed upon awakening (Y = Yes; N = No)
RATRIAL	Atrial fibrillation (Y = Yes; N = No)
RVISINF	Visible infarction on CT (Y = Yes; N = No)
RSBP	Systolic blood pressure (mmHg) (Optimal, Normal, Prehypertension, Stage 1 Hypertension, Stage 2 Hypertension, and Stage 3 Hypertension)
RDEF1	Facial deficit (Y = Yes; N = No; C = Not informed)
RDEF2	Arm/hand deficit (Y = Yes; N = No; C = Not informed)
RDEF3	Leg/foot deficit (Y = Yes; N = No; C = Not informed)
RDEF4	Dysphasia (difficulty speaking) (Y = Yes; N = No; C = Not informed)
RDEF5	Hemianopia (difficulty seeing) (Y = Yes; N = No; C = Not informed)
RDEF6	isuospatial disorder (Y = Yes; N = No; C = Not informed)
RDEF7	Brainstem pain (Y = Yes; N = No; C = Not informed)
RDEF8	Other deficit (Y = Yes; N = No; C = Not informed)
DNOSTRK	Patient classification without stroke (Y = Yes; N = No)

4.3 Relationship between attributes

Looking to better understand the relationship between the data for classification, a data analysis was performed. Initially, an analysis was carried out to identify whether there was any relationship between sex and age and the occurrence of stroke. With this, it was possible to identify that there are higher incidences of stroke in men of advanced age, when compared to women of the same age (Figure 1).

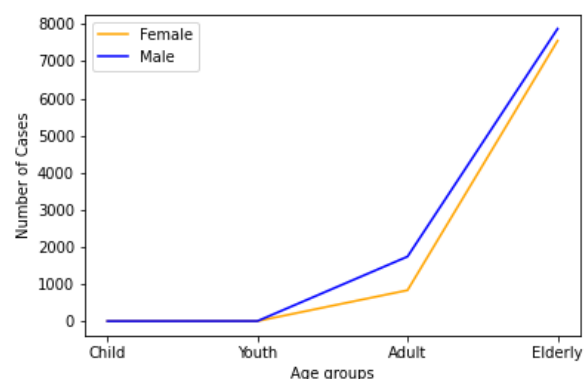


Figure 1. Relationship of stroke occurrence with sex and age.

After that, an analysis was performed to identify whether blood pressure would also be related to stroke occurrence. In this case, it was observed that even in people with “Optimal” and “Normal” pressure, stroke occurs. From people with pre-hypertension to people with Stage 3 hypertension, the number of people with stroke was higher, and more prevalent in men with Stage 1 hypertension (Figure 2).

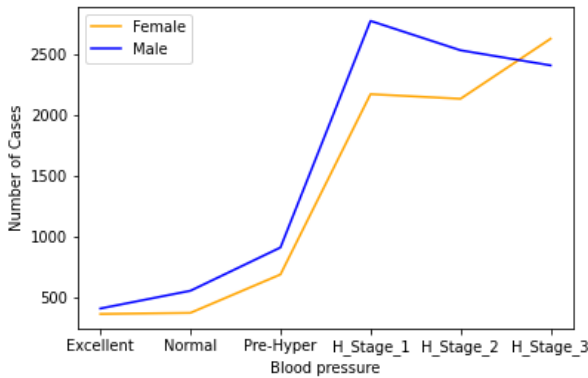


Figure 2. Relationship between stroke occurrence and blood pressure.

The relationship between stroke occurrence and the indicative signs presented in the data set was also verified, such as: facial deficit (RDEF1), deficit in the arm or hand (RDEF2), deficit in the leg or foot (RDEF3), difficulty speaking (RDEF4), difficulty seeing (RDEF5), visuospatial disorder (RDEF6), brainstem pain (RDEF7), and other deficits (RDEF8). With this, it was possible to perceive that the symptoms of facial deficit (RDEF1), deficit in the arm or hand (RDEF2), deficit in the leg or foot (RDEF3), and difficulty speaking (RDEF4) were the symptoms that were more related to stroke occurrence (Figure 3).

On the other hand, symptoms of difficulty seeing (RDEF5) and visuospatial disorder (RDEF6) showed almost linear values, while symptoms of brainstem pain (RDEF7) and other deficits (RDEF8) were not as closely related to stroke with very low values. Furthermore, the occurrence of these symptoms was also found to be more prevalent in men (Figure 3).

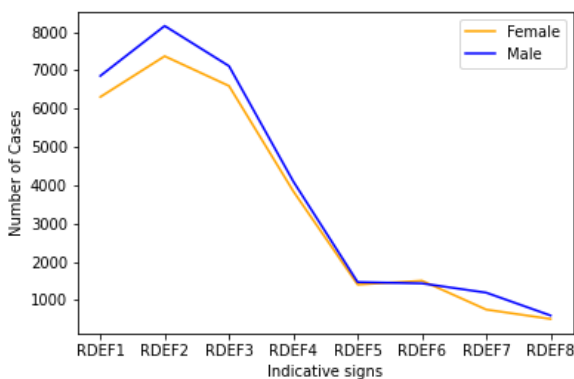


Figure 3. Relationship between stroke occurrence and indicative signs.

4.4 Definition of experiments

After performing the data analysis to identify existing relationships, the experiment was defined. To carry out the experiment, six data mining algorithms were selected to be evaluated and to find the best result for the problem addressed. The selected algorithms were: Decision Tree, K-Nearest Neighbors (K-NN), Naive-Bayes, Support Vector Machine (SVM), Multilayer Perceptron (MLP), and the Apriori algorithm. They were chosen due to the wide use of these algorithms in classification problems and the use of Apriori to analyze what are the possible association rules existing in the data set and to help in the identification of individuals most susceptible to having a stroke. Table 2, the hyperparameters defined for each algorithm are presented.

Table 2. Hyperparameters of the algorithms.

Algorithms	Hiperparâmetros
Decision tree	criterion='entropy', random-state=0
SVM	kernel='linear', C=1
MLP	random-state=1, max-iter=300
K-NN	n-neighbors=11, weights='distance'
Naive Bayes	Default setting

In SVM, a linear kernel was used to maintain a simple and interpretable model, while the regularization parameter C was adjusted to balance between avoiding overfitting and maintaining model accuracy. In K-NN, a sufficient number of nearest neighbors was used, and greater weight was assigned to the nearest neighbors to handle unequal distributions of instances per class. In the decision tree, the quality of the splits was measured based on the entropy of the classes to ensure reproducibility of the results. In MLP, the maximum number of iterations during training was limited for result reproducibility and to attempt to avoid overfitting. These choices were made considering the specific characteristics of each algorithm.

4.5 Validation of algorithms

To test and validate the algorithms, the training and testing method based on k-fold cross-validation was used, with $k = 10$. Thus, in order to obtain the best performance among the techniques for the work context, it was considering the average accuracy for all iterations and the standard deviation, given by the square root of the variance of the data, in which the closer to zero, the better the homogeneity of the data set. In addition, accuracy, recall and F1-score were also taken into account. The results obtained will be discussed in Section 5.

5 Results and discussions

The methodology used in this work was based on the KDD process, proposed by Ferrari and Castro [2016] and consisted of the following steps: analysis of information from the original database; data pre-processing; existing relationship between attributes; definition of experiments; and validation of the algorithms. Among the preprocessing activities, the most

important ones were the removal of irrelevant attributes, the verification and removal of missing data, and the techniques to handle imbalanced data. These activities were essential to improve the data quality and, consequently, the performance of the classification models used. In the experiment, six data mining algorithms were used: Decision Tree, K-NN, Naive Bayes, SVM, and MLP in order to identify which of these techniques presented the best results for classification, and the use of the algorithm Apriori to analyze the possible association rules existing in the dataset.

After carrying out the experiments, the results of the algorithms were obtained in the database, by evaluating the performance of each one. Given an analysis of the results obtained, the decision tree algorithms and the MLP were the ones that presented the best results. The SVM presented an average accuracy of 86.4% and a standard deviation of 0.036, being the highest deviation when compared to the other algorithms. The K-NN obtained an average accuracy of 83.0% and a standard deviation of 0.022. The Naive Bayes algorithm showed the lowest average accuracy among all the algorithms and the second highest standard deviation of the experiment, around 0.035. Table 3 presents the results of the techniques in terms of standard deviation and average accuracy, respectively.

Table 3. Standard deviation and average accuracy of the algorithms.

Algorithms	Standard Deviation	Average Accuracy
Decision tree	0.023	93.2%
SVM	0.036	86.4%
MLP	0.017	93.3%
K-NN	0.022	83.0%
Naive Bayes	0.035	80.3%

When analyzing the metrics shown in Table IV, it is inferred that the decision tree algorithm presented the best percentage to classify cases with stroke. The results of the MLP algorithm were very close, with an accuracy of 99%, a recall of 86%, and an F1-score of 93%. The SVM algorithm presented the lowest precision for classifying stroke cases with an accuracy of 92%. K-NN had a precision of 99%, a recall of 65%, being the lowest result when compared to the other algorithms, and an F1-score of 85%. Naive Bayes was the one that presented the lowest percentages when compared to the other algorithms. This can be explained due to the characteristics of the dataset when compared to the characteristics of the most appropriate algorithms for a certain type of values, for example, the decision tree that works well with categories.

Table 4. Result of algorithm metrics.

Algorithms	Precision	Recall	F1-Score
Decision tree	99%	88%	93%
SVM	92%	88%	90%
MLP	99%	86%	92%
K-NN	99%	65%	78%
Naive Bayes	99%	58%	

The next step consisted of applying the association algo-

rithm. The data was imported into Weka¹, and the association technique was applied using the Apriori algorithm. The analyzed attributes were: "SEX", "AGE", "RSLEEP", "RATRIAL", "RVISINF", "RSBP", "RDEF1", "RDEF2", "RDEF3", "RDEF4", "RDEF5", "RDEF6", "RDEF7", "RDEF8", "DNOSTRK" and "DPE". The configuration of the Apriori algorithm was defined as follows: "Apriori -N 10 -T0 -C0.9 -D0.05 -U 1.0 -M0.1 -S-1.0 -c-1". According to Weka [2022] the algorithm standardization is described in Table 5.

Table 5. Apriori algorithm configuration.

Algorithms	Description
N	Represents the maximum number of rules that will be shown in the output of the algorithm.
T	Represents the metric defined to rank the rules. The metric to be defined can be: confidence, lift, leverage, conviction.
C	Represents the minimum score for the defined rule.
D	Represents the delta and consists of reducing support iteratively, starting from the upper limit until the lower limit is reached.
U	Represents the upper limit for minimum support.
M	Represents the lower limit for minimum support.
S	Represents the significance level. If used, rules are tested for significance at a given level.
c	Represents the class index.

Table 6 presents the association rules with the application of the Apriori algorithm from 18,425 instances.

Table 6. Association rules generated by the Apriori algorithm.

Number of rules	Rules found
1	RATRIAL = N DNOSTRK = N 14890 == DPE = N 14799
2	RDEF7 = N 14871 == DPE = N 14780
3	RATRIAL = N 15260 == DPE = N 15164
4	RDEF8 = N DNOSTRK = N 15710 == DPE = N 15610
5	DNOSTRK = N 18006 == DPE = N 17890
6	RDEF8 = N 16061 == DPE = N 15956
7	AGE = IDOSO2 DNOSTRK = N 15432 == DPE = N 15325
8	AGE = IDOSO2 15755 == DPE = N 15643
9	RDEF2 = Y DNOSTRK = N 15535 == DPE = N 15424
10	RDEF2 = Y 15809 == DPE = N 15693

¹Software tool composed of a collection of machine learning algorithms for solving data mining problems Weka [2022]. Available for download at: <https://sourceforge.net/projects/weka/>

Data analysis pointed to 10 rules that presented the patterns found, in which there was a relational crossing between the attributes, based on the existence of stroke, age, symptoms observed upon awakening, and the presence or absence of pulmonary embolism. The interpretation of the obtained results demonstrates that in a hypothetical situation, a cause arising from a disease does not necessarily have an expected consequence. The simulation of real user interactions could be inferred the following assertions:

- If a patient does not have atrial fibrillation and has a stroke, then that patient does not have a pulmonary embolism.
- If a patient does not have atrial fibrillation, then that patient does not have a pulmonary embolism.
- If a patient has any other deficit and has a stroke, then that patient does not have a pulmonary embolism.
- If a patient has a stroke, then that patient does not have a pulmonary embolism.
- If a patient is over 60 years of age and has a stroke, then that patient does not have a pulmonary embolism.
- If a patient has an arm deficit and has a stroke, then that patient does not have a pulmonary embolism.

In this case, the decision tree technique and the MLP obtained the best results, with the MLP being considered the best since it presented the lowest standard deviation, followed by the decision tree, SVM, K-NN, and Naive Bayes. From the data generated by the Apriori algorithm, it was perceived that the occurrence of pulmonary embolism was not related to any of the cases, it was also identified which age may be related to the presence of stroke and the deficits that may occur.

6 Final considerations

The main objective of this work consisted in the use of six data mining algorithms in order to assist in the identification and diagnosis of individuals prone to having a stroke based on indicative signs (numbness on only one side of the body, mental confusion, difficulty speaking or understanding and others) and risk factors. In summary, it was possible to verify that with the analysis of the relationship between the data and the application of the algorithms it is possible to help in the prediction of stroke occurrence. Through this analysis, it was possible to observe that stroke occurs more frequently in men with advanced age and blood pressure, as well as indicative signs that can influence the occurrence of stroke.

From the results generated by the algorithms, it was found that the MLP algorithm presented the best results for the classification, followed by the decision tree, SVM, K-NN, and Naive Bayes, and, based on the rules generated by the Apriori algorithm, it was found that there is no relationship between pulmonary embolism and stroke in any of the cases tested. In further works, we intend to improve the pre-processing of missing values and imbalanced data, make use of algorithm committees and improve the parameters of the Apriori algorithm, to generate better results. In addition, the algorithm that presented the best results in this work will be considered

for implementations of computational solutions aimed at predicting data concerning stroke.

References

- Azank, F. and Gurgel, G. K. (2020). Dados desbalanceados — o que são e como lidar com eles. <https://medium.com/turing-talks/dados-desbalanceados-o-que-o-que-são-e-como-evitá-los-43df4f9732b>.
- Brasil (1990). Lei 8.069, de 13 de julho de 1990. dispõe sobre o estatuto da criança e do adolescente e dá outras providências. *Diário Oficial [da] República Federativa do Brasil*.
- Brasil, M. d. S. (2013). Diretrizes de atenção à reabilitação da pessoa com acidente vascular cerebral.
- Campbell, B. C. V. and Khatri, P. (2020). Stroke. *The Lancet*, 396(10244):129–142. DOI: 10.1016/s0140-6736(20)31179-x.
- Cerqueira e Francisco, W. d. (2022). Faixa etária da população brasileira. <https://educador.brasilecola.uol.com.br/estrategias-ensino/faixa-etaria-populacao-brasileira.htm>.
- de Saúde, S. M. (2021). *Hipertensão Arterial: Manejo clínico na Atenção Primária à Saúde*, volume 1. Patrícia Aparecida Piva — Gerência técnica de Doenças Renocardiovasculares e Diabetes, 1 edition. Anotação.
- Faceli, K., Lorena, A., Gama, J., and Carvalho, A. (2011). *Inteligência Artificial—uma abordagem de aprendizado de máquina*. Rio de Janeiro: LTC.
- Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., et al. (1996). Knowledge discovery and data mining: Towards a unifying framework. In *KDD*, volume 96, pages 82–88.
- Ferrari, D. G. and Castro, L. N. d. (2016). *Introdução a mineração de dados*. Saraiva Educação SA.
- Géron, A. (2019). *Mãos à Obra: Aprendizado de Máquina com Scikit-Learn & TensorFlow*. Alta Books.
- Govindarajan, P., Soundarapandian, R. K., Gandomi, A. H., Patan, R., Jayaraman, P., and Manikandan, R. (2020). Classification of stroke disease using machine learning algorithms. *Neural Computing and Applications*, 32(3):817–828. DOI: 10.1007/s00521-019-04041-y.
- Meschia, J. F., Bushnell, C., Boden-Albala, B., Braun, L. T., Bravata, D. M., Chaturvedi, S., Creager, M. A., Eckel, R. H., Elkind, M. S., Fornage, M., et al. (2014). Guidelines for the primary prevention of stroke: a statement for healthcare professionals from the american heart association/american stroke association. *Stroke*, 45(12):3754–3832.
- Sandercock, P. A., Niewada, M., and Członkowska, A. (2011). The international stroke trial database. *Trials*, 12(1):1–7.
- Santana, R. (2020). Lidando com classes desbalanceadas — machine learning. <https://minerandodados.com.br/lidando-com-classes-desbalanceadas-machine-learning/>.
- Singh, M. S. and Choudhary, P. (2017). Stroke prediction using artificial intelligence. In *2017 8th Annual Industrial Automation and Electromechanical Engineering Confer-*

ence (*IEMECON*), pages 158–161. DOI: 10.1109/IEMECON.2017.8079581.

Weka (2022). Class apriori.

<https://weka.sourceforge.io/doc.dev/weka/associations/Apriori.html>.

Yu, J., Kim, D., Park, H., Chon, S.-c., Cho, K. H., Kim, S.-J., Yu, S., Park, S., and Hong, S. (2019). Semantic analysis of nih stroke scale using machine learning techniques. In *2019 International Conference on Platform Technology and Service (PlatCon)*, pages 1–5. DOI: 10.1109/PlatCon.2019.8668961.

Zhang, Y., Song, W., Li, S., Fu, L., and Li, S. (2018). Risk detection of stroke using a feature selection and classification method. *IEEE Access*, 6:31899–31907. DOI: 10.1109/ACCESS.2018.2833442.