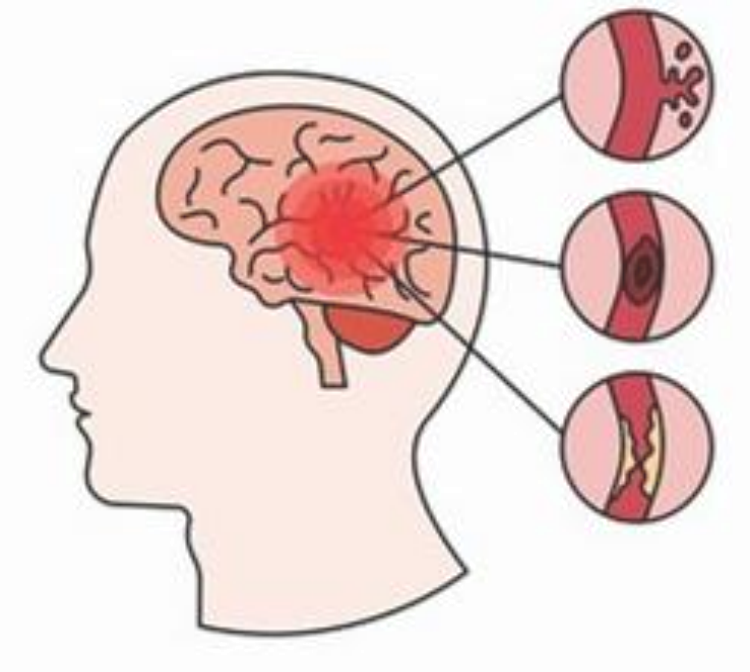
Nova IMS Information Management School

Big Data Analytics

2025



Stroke Prediction

Ana Escarei, 20241641

Maria Inês Lopes, 20240340

José Nobre, 20241614

Rúben Carvalho, 20241390

1. Introduction

A stroke occurs when the blood supply to a part of the brain is interrupted, either due to a blockage or a rupture of a blood vessel (Associação Nacional AVC , 2025).

There are two main types of strokes: Ischemic and Hemorrhagic. An Ischemic stroke occurs when a blood clot blocks an artery, preventing blood flow to the brain (Associação Nacional AVC , 2025). This is the most common type. The main causes of this kind of stroke are atherosclerosis and plaque buildup in the vessel walls, which can lead to cerebral thrombosis or cerebral embolism (American Stroke Association, 2025) A Hemorrhagic stroke, also known as brain hemorrhage, occurs when a blood vessel bursts, causing bleeding in the brain, and it also interrupts blood supply (Associação Nacional AVC , 2025). It is the less common type of stroke. This interruption in blood flow may be caused by an intracerebral hemorrhage or a subarachnoid hemorrhage (American Stroke Association, 2025). Finally, there is the Transient Ischemic Attack, also known as a Mini-Stroke, which is a “temporary blockage of blood flow to the brain. The clot usually dissolves on its own or gets dislodged, and the symptoms usually last less than five minutes.” (American Stroke Association, 2025)

The main goal of this project, using this dataset, is to understand the likelihood of someone having a stroke based on certain risk factors. These factors include age, gender, high blood pressure, pre-existing heart disease, marital status, work, type of residence, glucose level, body mass index (BMI), and smoking status.

When searching for a dataset, our criteria were that it should cover a current topic in the healthcare field and have a data size and number of variables that would allow us to apply, within a Databricks project, the theoretical concepts learned in class.

2. Data Exploration

- **Data Preprocessing**

In this project, we used a dataset with the goal of predicting the probability of a patient suffering from a Stroke. Each record in the dataset corresponds to a

patient and contains relevant information about their clinical and demographic profile.

This dataset includes a total of 5010 rows and 12 variables, of which 6 are numerical (age, hypertension, heart disease, glucose level, body mass index (BMI), and 5 are categorical (gender, marital status, work type, residence type, and smoking habits).

In the initial data processing, the category “other” was identified in the gender variable. Therefore, a filter was applied to keep only the rows that did not contain the value “other”. In the age variable, 115 decimal values were found, presumably corresponding to children. Thus, we chose to retain only rows where age is an integer, discarding decimal values.

Regarding BMI, 201 records had missing values (N/A), which were replaced with null values and subsequently removed from the dataset. During this phase, outliers were also detected and removed, specifically cases where the BMI exceeded 60 or the blood glucose level surpassed 260 mg/dL. To support this analysis, an illustrative chart of these outliers is provided in Annex 1.

To prepare the dataset for modeling, a pipeline was developed to perform the necessary transformations on the variables. This ensures that the dataset is ready for training and validating predictive models.

To ensure data consistency and enable efficient operations, it was necessary to convert the data types of different *DataFrame columns*. Categorical variables were converted to *String Type*, while numerical variables were converted to *Integer Type*. Additionally, the identification column (*id*) was removed, as it did not contain any relevant information for the analysis.

- **Feature Selection**

To identify which variables, have the highest impact on stroke prediction, the Random Forest algorithm was used. After training the model, it was possible to extract the relative importance of each variable using the *feature Importance* attribute. These values were converted into an ordered list, which allowed us to highlight the features that contributed most to the model’s decisions. Based on this analysis, the 10 most relevant variables were selected, considering their weight in reducing uncertainty throughout the decision trees.

3. Models

Two distinct models, Logistic Regression and Random Forest, were then evaluated to compare their performance and determine which one best fits the characteristics of the dataset. Both models were trained and adjusted using details such as AUC, precision, recall and F1 score, allowing the optimization of the predictive capacity for stroke risk.

- **Hyperparameter tuning with Wight Col on Logistic Regression**

First, we used the Logistic Regression model. We chose this model because it is a classic statistical model, widely used in binary classification problems, where we can understand and interpret the influence of each variable on the probability of a stroke occurring or not.

We tested our model using several features and the ones that gave us the best results were those with 5, 10 and 12 features. Of these three, the one that presented the best results was the one with 10 features.

- **Hyperparameter tuning with Wight Col on Random Forest**

Next, we used the Random Forest model. We applied this model because it could connect non-linear relationships and complex interactions between variables, based on multiple decision trees. We tested this model because it is useful in problems where the data has high dimensionality and possible correlations between variables. We also tested different features, and the ones that gave us the best results were those with 8 and 10 features, with the latter presenting the best results.

- **Plotting of AUC and Confusion Matrix for the best model**

According to the results we obtained, the best model was Random Forest with 10 features. We chose this model after testing several models with different combinations of variables and concluded that it performed best. Our choice considered the analysis of some metrics, with special emphasis on AUC, F1 score, precision and recall.

Our model had an AUC of 0.839, which indicates a good ability to distinguish between patients with and without stroke. The F1 score of 0.2078 was the highest

among all the configurations tested, thus reflecting a balance between precision (0.117) and recall (0.902). The recall value, in a clinical context, is quite important because a high recall value means a greater ability to correctly identify positive cases, that is, people who have suffered a stroke. Another important point is false negatives, because in a health context, having a false negative represents a real risk of stroke that passed and was not properly identified, leading to serious consequences for the patient.

The ROC curve (appendix 4) was generated by this model and the shape of the curve, which approaches the upper left corner of the graph, indicates the high rate of true positives with a low rate of false positives.

The selection of 10 features allows maintaining interpretability and avoiding overfitting, and this model is the most suitable to maximize predictive performance without compromising the practical viability of the solution.

Considering the model used, a confusion matrix was also generated (appendix 5), which provided us with important information about the classifier's performance regarding the hits and errors made in each class. In relation to true negatives, i.e. cases correctly classified as "Without Stroke", we had a value of 595; in relation to false positives, cases classified as "With Stroke" although they were "Without Stroke", we had a value of 278. In relation to false negatives, cases of stroke that the model was unable to identify, we had 4; and finally, true positives, cases correctly identified, were 37. The matrix is an effective model in identifying real cases of stroke due to its high sensitivity; however, its low high precision can lead to false alarms and conclusions that are not so positive for the patient.

The variables with the greatest impact on stroke prediction were: ever_married_index, smoking_status_ohe, gender_index, work_type_ohe, scaled_numeric_vector and residence_type_index. Although most features are aligned with risk factors, some require more critical interpretation.

The ever_married_index feature stood out as the most prevalent in the model. Although, at first glance, it may suggest a direct relationship between marital status and stroke risk, it is important to consider that this variable may indirectly reflect the "age" factor. Older individuals are more likely to have been married, and age is recognized as one of the main risk factors for stroke.

The importance of smoking_status_ohe is in line with the fact that smoking is a risk factor for cardiovascular and neurological diseases, including stroke. Similarly, the gender_index variable may also reflect known variations between genders regarding the incidence and evolution of strokes. In turn, work_type_ohe may relate to the level of stress in the work environment. The scaled_numeric_vector variable, which represents the combination of attributes such as BMI and blood glucose, proved to be relevant because it directly aggregates physiological factors associated with the occurrence of stroke. Residence_type_index may reflect inequalities in access to health care between urban and rural areas, as well as differences in lifestyles.

4. Conclusion

In carrying out this project, the main difficulty results from a limitation of the Databricks Community, which does not allow the sharing of Notebooks, preventing the production of code by the various elements of the group in a single Notebook, forcing them to work separately, sometimes in duplicating and exporting and importing Notebooks for comparison and sharing of code.

Regarding the Dataset, the fact that it was not balanced created some problems that were overcome with the inclusion of adjusted parameters in the models.

In short, the project allowed testing and training several prediction models, in search of the most final solution to the problem under analysis.

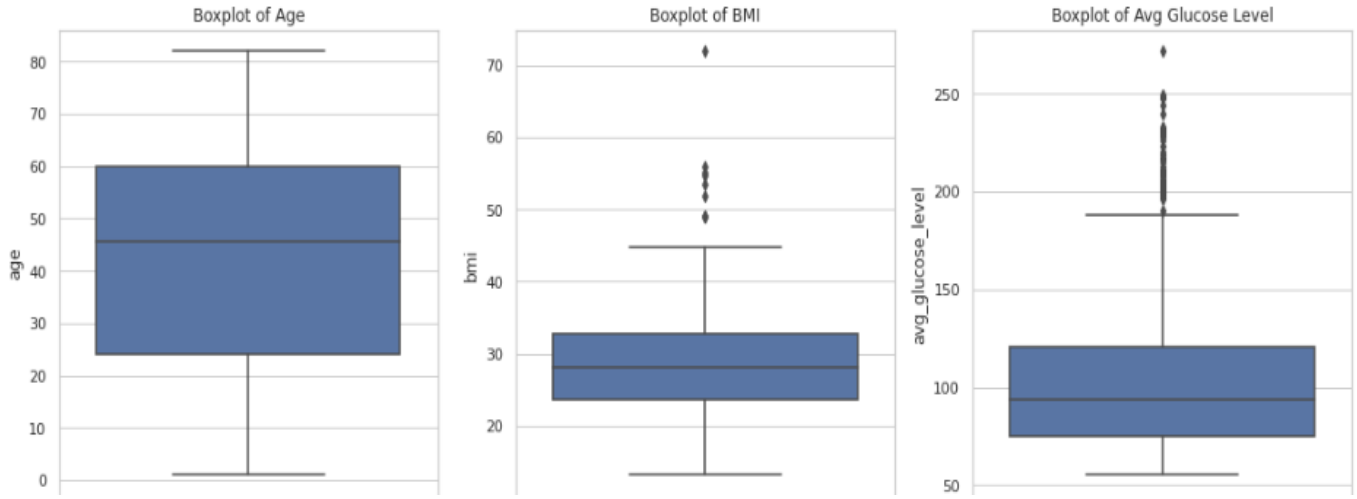
In the end, we obtained a model with good predictive capacity, since the number of false negatives is low, although it presents a high number of false positives. In healthcare, it is preferable to err on the side of excess in disease prevention measures. It is suggested that the model be further improved so that more assertive diagnoses promote the effectiveness of resource management.

5. Bibliograph

1. American Stroke Association. (2025). Retrieved from https://www.stroke.org
2. Associação Nacional AVC . (2025). Retrieved from https://www.anavc.pt
3. Instituto Regional de Estatística. (2025). *Menos Mortes por Doenças do Aparelho Circulatório*.
4. *Kaggle*. (2025). Retrieved from https://www.kaggle.com/
5. Stroke Association. (2025). Retrieved from https://www.stroke.org.uk

6. Appendix

Appendix 1 – Chart of the detected outliers.



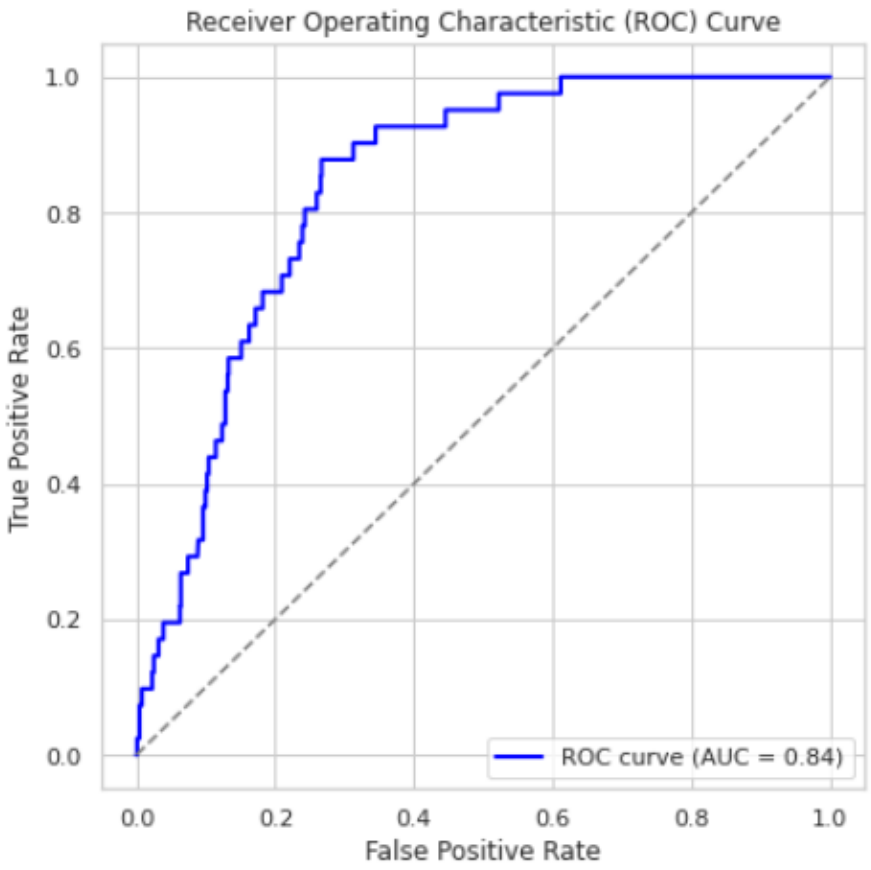
Appendix 2 – Logistic Regression Table

	5 Features	10 Features	12 Features
Best regParam	0.01	0.2	0.1
Best elasticNetParam	0.0	0.25	0.5
Best maxIter	10	10	10
Best model AUC	0.827843	0.862319	0.833069
Accuracy	0.711160	0.705689	0.671772
True Positives (TP)	33	36	28
False Positives (FP)	256	264	298
False Negatives (FN)	8	5	2
True Negatives (TN)	617	609	586
Precision	0.114187	0.120000	0.085890
Recall	0.804878	0.878049	0.933333
F1 Score	0.200000	0.211144	0.157303

Appendix 3 – Random Forest table

	8 Features	10 Features
Best numTrees	150	150
Best maxDepth	4	4
Best maxBins	64	64
Best model AUC	0.812880	0.839145
Accuracy	0.683807	0.691466
True Positives (TP)	29	37
False Positives (FP)	282	278
False Negatives (FN)	7	4
True Negatives (TN)	596	595
Precision	0.093248	0.117460
Recall	0.805556	0.902439
F1 Score	0.167147	0.207865

Appendix 4 – Graphic Receiver Operating Characteristics Curve



Appendix 5 – Confusion Matrix