Project: Stroke Risk Prediction with Machine Learning Techniques

Cavallini Sara*[1] and Eusebio Alberto†[2]

[1, 2] *Politecnico di Milano*

Abstract

A stroke occurs when the blood flow to a specific part of the brain is suddenly interrupted. This interruption leads to the gradual death of brain cells, resulting in disabilities that vary based on the affected brain region. Early identification of stroke symptoms is crucial for effective prediction and fostering a healthy lifestyle. This research projects is aimed at evaluating the work of Dritsas and Trigka,[1] employing machine learning (ML) techniques to develop and evaluate multiple models, aiming to create a robust framework for long-term stroke risk prediction. The above mentioned study's main contribution is the development of a stacking method that demonstrates high performance, validated by various metrics including AUC, precision, recall, F-measure, and accuracy. However, our evaluation of the paper's results revealed some inconsistencies that warrant further investigation.

1 Introduction

In this section, we present the project scope, our methodology, and how our work differs from the original paper by Dritsas and Trigka. Our primary goal was to replicate and evaluate their study on stroke prediction using machine learning techniques. We began by preprocessing the data, following similar steps as outlined in the paper, including handling missing values, filtering out patients and augmenting the dataset.

We conducted an extensive Exploratory Data Analysis (EDA) to ensure the integrity and quality of the data. Our EDA confirmed that the data distribution and percentages matched those reported in the original study, indicating accurate preprocessing.

Furthermore, we addressed the class imbalance in the dataset using the SMOTE oversampling technique, unlike the original study, which only augmented the training dataset. This adjustment was made to ensure a balanced representation of classes in the entire dataset.

In terms of feature ranking, we applied the random forest classifier and information gain ranking, which corroborated the original study's findings on the importance of age, BMI, and glucose level in stroke prediction.

Finally, we trained multiple models, tuning hyperparameters using GridSearchCV to identify the best-performing models. We implemented and evaluated both stacking and voting methods as described in the original study, comparing the results against the reported metrics. Our evaluation revealed some inconsistencies in the reported performance metrics, which we discuss in the following sections.

To address some of the discrepancies between the paper methods and the recommended best practices, we have developed two version of the notebook that we will refer to as version A and B. The first tries to replicate the paper steps, while the latter introduces variations and best practices to address some of the problems that we have found while analyzing the paper.

*sara.cavallini@mail.polimi.it

†alberto.eusebio@mail.polimi.it

2 Exploratory Data Analysis

In this section, we present the methods used to replicate the paper and additional steps taken to ensure a fair analysis. Our Exploratory Data Analysis (EDA) includes data preparation, cleaning, handling outliers, categorical encoding, and addressing data imbalances.

2.1 Data preparation

The paper experiment were performed over a publicly available dataset,[2] that was available on the Kaggle platform. To prepare the data for model training, we followed a systematic approach to clean and preprocess the dataset. This involved several key steps to ensure data quality and consistency.

2.1.1 Data cleaning

We identified and handled missing or null values in the dataset. Specifically, the BMI column contained some N/A values, which we removed in version A to maintain data integrity. In line with the original paper, we only considered patients above 18 years old in both versions, removing those younger to ensure consistency in the age demographic. Additionally, in version A we filtered out patients with unknown gender and smoking status to reduce uncertainty in the data.

However, as it appears from our analysis, the amount of null values are all concentrated in the BMI class and they account for up to 16.06% of the stroke values, so dropping them will consist in a significant information loss. In version B of the notebook, we have decided to impute these values with the mean of the category BMI. Similarly, the single 'Other' label in the gender category was set to the most frequent label in the class: 'female'. Handling the imputation of the 'Unknown' label in the smoking status feature is equally important, as these samples account for 13.45% of the stroke class samples. Since the 'Unknown' label is the second largest one for the class in the dataset, simply imputing it to the majority class would not capture all the variance of the dataset, so we have decided to train a Random Forest Classifier to reclassify this label. It is important to note that we did not use the stroke column when training the classifier, not to introduce bias in the dataset itself.

2.1.2 Outliers

In the original study, the authors removed the single lowest BMI value, which we also replicated to maintain consistency. We conducted further analysis to identify and address any other significant outliers that could skew the model results, ensuring a robust dataset for training. In version A we have removed the single lowest value of the BMI, while in version B we have adopted the Inter Quartile Method to classify and remove outliers both in the BMI and average glucose level categories.

2.1.3 Categorical encoding

Categorical features in the dataset were encoded using one-hot encoding for multi-class features. This approach was chosen to facilitate the effective use of categorical data in machine learning models. One-hot encoding was used to prevent any ordinal relationship assumptions. Furthermore, the numerical features such as BMI, age and average glucose level were normalized.

2.1.4 Addressing imbalances

The original paper highlighted a class imbalance between smokers and non-smokers, which we confirmed during our analysis. To address this, we used the SMOTE (Synthetic Minority Over-sampling Technique) method to balance the dataset by oversampling the minority class. In the original study, the resampling was performed on the entire dataset, so we have replicated this procedure in notebook A. However, since is common practice to resample only the training dataset, in notebook B we have done so. This choice was motivated by the need for a clean test dataset, to evaluate the performances of the algorithms on.

We also compared SMOTE with SMOTENC (SMOTE for Nominal and Continuous features), a variation that considers categorical features while generating synthetic samples. In fact, applying SMOTE on the dataset corrupts the one-hot encoded features that must be then rounded down to avoid multi-class definition errors that arise when resapling the dataset. In fact, consider the touple (0, 0.32, 1) obtained through resampling and belonging to the smoking-status category, it makes no sense to say that the patient belongs to the 'smoking' category and also partially to the 'never smoked' one.

2.1.5 Data splitting

The dataset was randomly split into two parts, for training and test. The training subset included 80% of the data in the dataset, while the testing dataset accounted for 20%. As stated before, while in notebook A we have performad resampling on the whole dataset, in notebook B we have decided to oversample only the training subset. This led to significantly different performance results in the two dataset.

2.2 Data description

In this section, we analyze the correlation between different features and present various graphs to illustrate these relationships. Each graph is accompanied by an explanation of the results obtained and their implications for stroke prediction.

2.3 Graphs

2.3.1 Imbalanced dataset

In this first part, we are going to present some of the graphs belonging to the imbalanced dataset. 1 shows clearly the underrepresentation of the stroke class.

We start by plotting the distribution of the numerical features inside the dataset. 2a, 2b, 2c are the plots of the distributions for the BMI, the average glucose level and age categories. Interestingly, the average glucose level category has two peaks.

Following in the analysis, 3a. 3c and 3b plot the correlation between the numerical features inside the dataset. We can clearly see no linear dependencies in these plots, but it is evident a correlation

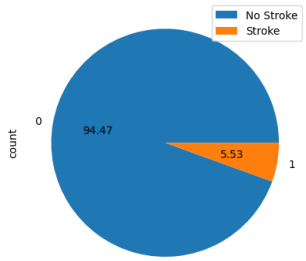


Figure 1: Imbalance distribution

between age and the incidence of higher values of blood sugar, that also seems to lead to a higher incidence of stroke. Interestingly, the 25-40 bmi values(overweight, obese I) seems to be the range in which the higher glucose levels are concentrated. However, this is also the range in which we have most of the samples.

2.3.2 Balanced dataset

After oversampling the dataset, we will show the plots of some of the features of the dataset, to understand better which one are more influential than others in predicting stroke.

The age group graph (Figure 4) shows the percentage of stroke occurrences across different age buckets. We observe that stroke incidence increases significantly with age, particularly in the 75+ age group, which has the highest stroke rate. This indicates that age is a critical factor in stroke risk.

The BMI category graph (Figure 5) illustrates the distribution of stroke occurrences across different BMI categories. Overweight and Obese I categories show a higher percentage of stroke patients compared to other BMI groups.

The gender graph (Figure 6) presents the percentage of stroke occurrences among males and females. The data shows a slightly higher stroke rate in females compared to males, indicating potential gender-related risk factors for stroke.

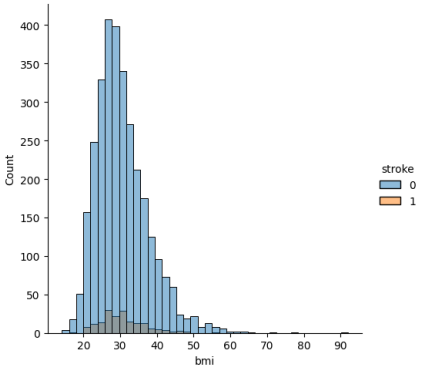
The heart disease graph (Figure 7) shows the percentage of stroke occurrences among patients with and without heart disease. Patients with heart disease have a higher stroke rate, however it is also shown that most of the patients suffering from stroke, did not seem to suffe from heart disease. This highlights the importance of monitoring cardiovascular health to prevent strokes.

A similar argument can be drawn from 8 that apparently does not contribute so much to the stroke class, but this can be explained as a lack of screening.

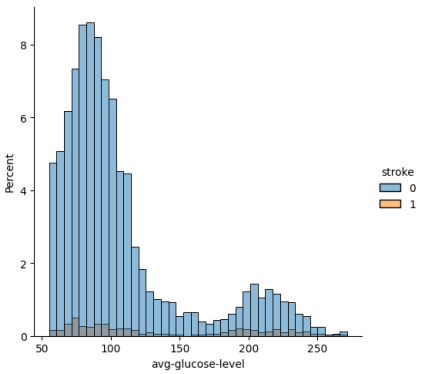
The plot 9 shows no significant stroke imbalance between living in a urban area or in the country side. So we are expecting this feature to have less importance in the model training.

The smoking status graph (Figure 10) displays the stroke occurrences among patients with different smoking habits. Surprisingly, former smokers and never smokers show higher stroke rates compared to current smoker. Nonetheless, this trend may also be biased by the uneven distribution of the data.

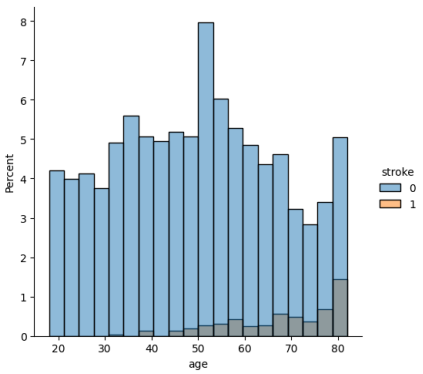
The work type graph (Figure 11) shows the distribution of stroke occurrences across different employment sectors. The private sector has the highest stroke rate, potentially due to higher representation in the dataset. Self-employed individuals also show a notable stroke rate.



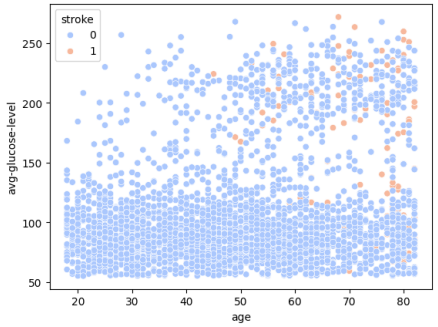
(a) BMI distribution



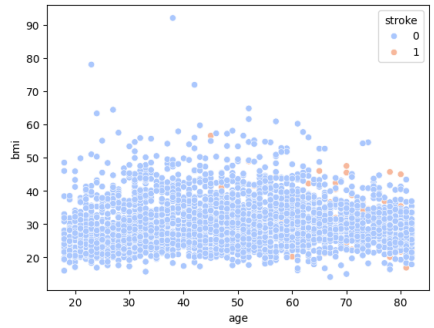
(b) Average Glucose level distribution



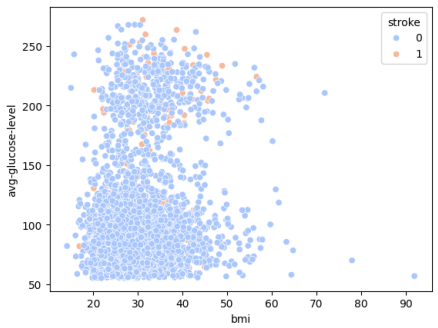
(c) Age distribution



(a) Age-Glucose correlation plot



(b) Age-BMI correlation plot



(c) BMI-Glucose correlation plot

2.4 Feature Ranking

In this section, we show the graphs of features obtained both for the paper replica and the best practice analysis. The results are very similar both for notebook A and B, so we will represent only the results for A. As also shown in the original paper, the Age feature is the most influential one, both for the information gain and Random Forest methods.

12a and 12b depict the importance of each feature, computed using both a Random Forest Classifier and for the Information Gain method. Both methods indicate that age is the most relevant feature, other features such as bmi and average glucose level have an high rank in both methods. The information gain method was applied fitting a decistion tree, using 'entropy as criterion'. Out of curiosity also the mutual information score for the features were calculated. This technique computes the statistical correlation between the feature and the output labels. The result of the latter is shown in 12c which also includes age and bmi among the most important features.

3 Model training and evaluation

In this section, we detail the models employed in our study, the methods used for hyperparameter tuning, and the evaluation metrics used to assess model performance.

3.1 Models training

We employed several machine learning models to predict stroke risk, following the methodology outlined in the original paper. The models used in our study include:

- Logistic Regression
- Decision Tree Classifier
- Random Forest Classifier
- STochastic Gradient Descent (SDG) Classifier

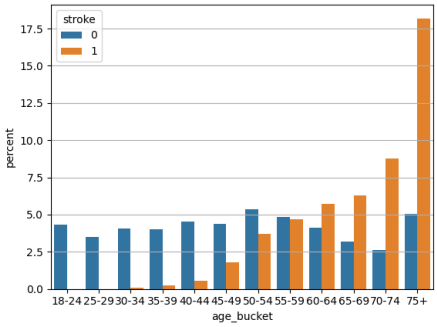


Figure 4: Age graph balanced dataset

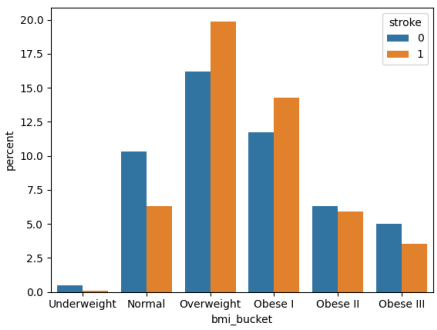


Figure 5: BMI graph balanced

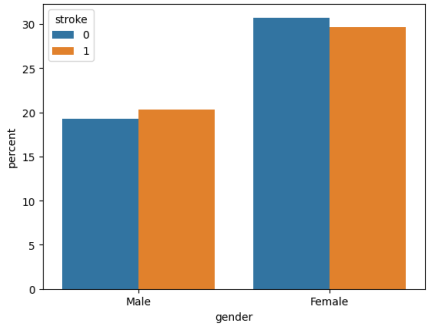


Figure 6: Gender influence

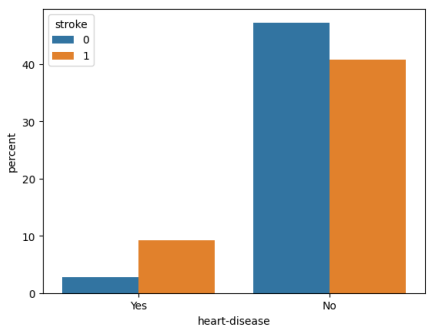


Figure 7: Heart disease incidence

- k-Nearest Neighbors (k-NN) Classifier
- Gaussian Naive Bayes
- Multi Layer Perceptron (MLP) Classifier
- Stacking Classifier
- Majority Voting Classifier

Each model was trained on the preprocessed dataset using all the features. The stacking classifier, which combines a Gaussian Naive Bayes Classifier, a Random Forest Classifier and two decision trees imblementing both the Gini Criterion and Entropy criterion to split nodes in the classifier.

3.2 Hyperparameter tuning

To optimize the performance of each model, we employed GridSearchCV for hyperparameter tuning. This method involves an exhaustive search over a specified parameter grid, evaluating each combination of parameters using cross-validation to identify the best-performing set.

We used GridSearchCV with 10-fold cross-validation to systematically explore the hyperparameter space for each model. This approach ensures that the selected hyperparameters provide the best generalization performance on unseen data.

The hyperparameter grids for each model were defined based on common practices and literature, ensuring a comprehensive exploration of the parameter space. The optimal hyperparameters identified through this process were then used to train the final models.

3.3 Evaluation Metrics

To evaluate the performance of the trained models, we used a variety of metrics, including:

- **Accuracy**: The ratio of correctly predicted instances to the total instances.
- **Precision**: The ratio of true positive predictions to the sum of true positive and false positive predictions.
- **Recall**: The ratio of true positive predictions to the sum of true positive and false negative predictions.
- **F1-Score**: The harmonic mean of precision and recall, providing a single metric that balances both concerns.
- **Area Under the Curve (AUC)**: The area under the Receiver Operating Characteristic (ROC) curve, representing the model's ability to distinguish between classes.

These metrics provide a comprehensive evaluation of model performance, highlighting both the accuracy and the ability to correctly identify stroke occurrences.

3.4 Results

In this section we will present the results obtained from the model training. For simplicity, we will present only two matrices, each one refering to one notebook, that display the results obtained for each metric.

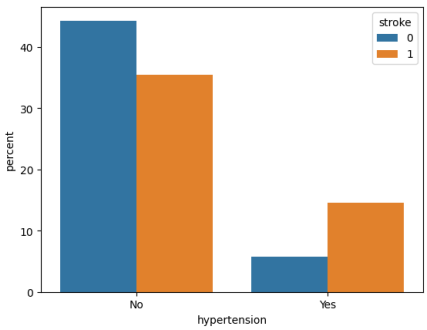


Figure 8: Hypertension incidence

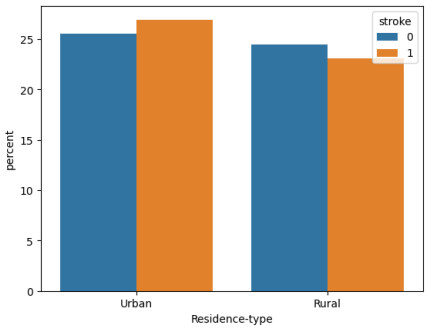


Figure 9: Residence type incidence

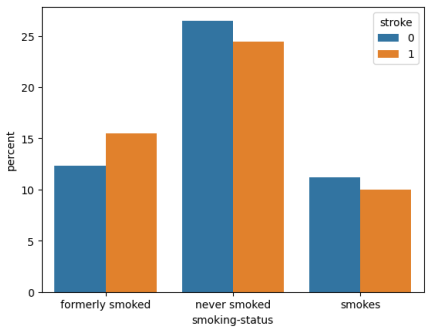


Figure 10: Smoking incidence

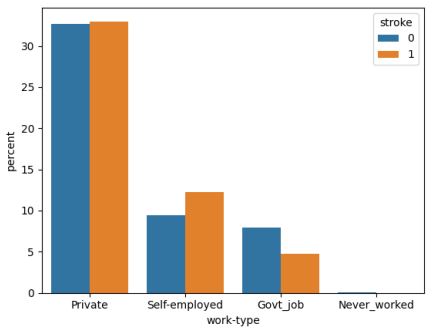


Figure 11: Work incidence

3.4.1 Evalutation methodology

As already declared, we have developed two version of the notebook: version A tries to replicate the methodologies stated in the paper, while version B implements best practices. An important difference between version A and B is the dataset on which the models are tested on. In version A, since the whole dataset was augmented using the SMOTE technique, not only the train dataset but also the test one is balanced, this 'drugs' the metrics results, loosing the real world distribution of data. In version B we have decided to oversample only the train dataset, leaving the test dataset untouched, resulting in a more fair evalutaion of the model performance.

3.4.2 notebook A

As we can clearly see from the matrix reported in 13a, the results of each of the trained models seem to indicate a good model training. The models that performed overall best are the stacking, the random forest classifier and the majority voting ones with ROC AUC score of 96%, 94% and 93%. For these models, also the recall metric seems to be pretty high. Having a high recall is foundamental when treating medical data, as it is better to have false positive than false negative in these cases and let other specialists double-check the data.

3.4.3 notebook B

13b is showing another figure of the data instead. When testing the algorithms on data coming from real world, we can clearly see that the performance in precision and recall drop significantly, showing that other models such as K-Nearest Neighbours and Logistic regression, may achieve better results in unseen data.

4 Conclusions

4.1 Paper Replication

In the first part of our project, we meticulously followed the procedures and algorithms outlined in the referenced paper. This replication allowed us to validate the findings and understand the underlying assumptions and methodologies employed by the original authors. The trends in our findings differ only slightly from those showed in the paper, and are mainly imputable to some hyperparameters used in the SMOTE function that we were not able to replicate.

This part confirms the Stacking method as the most performing one, followed by Random Forest and Majority Voting.

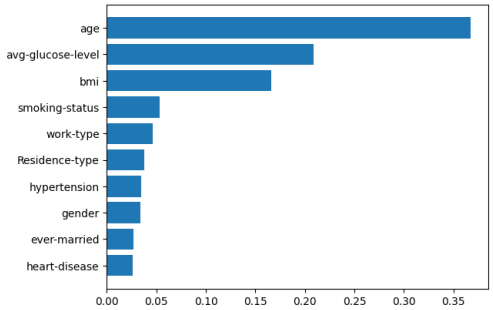
4.2 Best Practices Implementation

The second part of the project revolved around the implementation of best practices to handle outliers and implement a fairer evaluation of our trained modes. In this section we have registered severe underperformance of the models on the real data distribution, rising awareness. In particular, it was concerning the low recall score of the models on the unseen data.

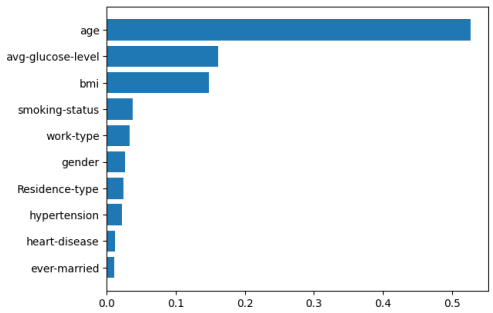
This part shows new emerging best predictions such as KNN, Logistic Regression and Naive Bayes Classifiers, that are chosen based on their high AUC and recall scores.

References

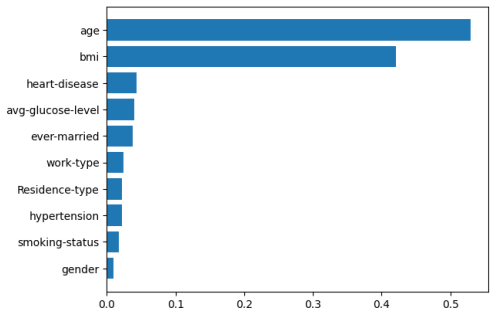
[1] Dritsas E, Trigka M. Stroke Risk Prediction with Machine Learning Techniques. Sensors. 2022;22(13). Available from: https://www.mdpi.com/1424-8220/22/13/4670.



(a) Random forest feature ranking

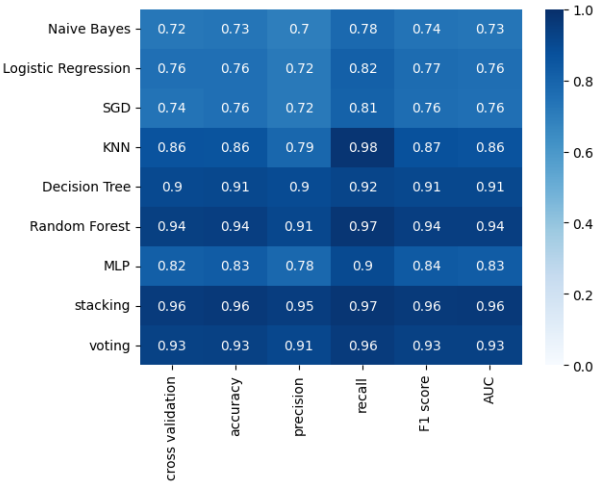


(b) Information gain feature ranking

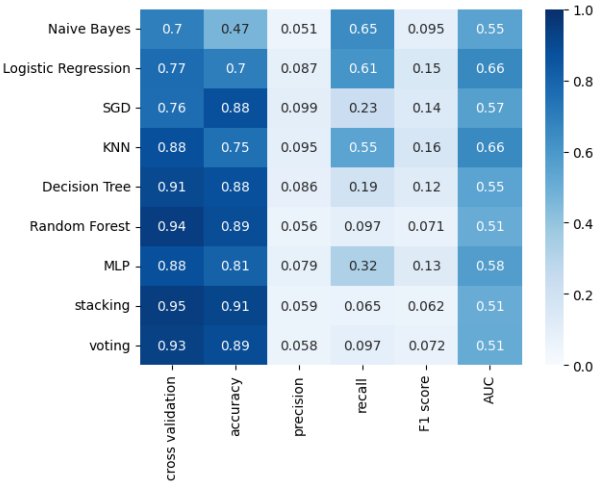


(c) Mutual information feature ranking

[2] Palacios FS. Stroke Prediction Dataset;. Available from: `https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset`.



(a) Models performance A



(b) Models performance B