Project:
Prediction of Stroke

By Samd Guizani (Cohort A23 – SPOC)

CONTENTS

EXECUTIVE SUMMARY

Objective:

The primary aim of this study is to develop a model that predicts the likelihood of a patient having a stroke using a set of descriptors such as age, gender, and existing health conditions.

Dataset:

- Total Observations: 5111
- Variables: 12 (including unique patient ID, demographic information, health indicators, and stroke occurrence)

Key Variables:

- Target: Stroke occurrence (binary: 1 if stroke, 0 if no stroke)
- Predictors: age, gender, hypertension, heart disease, marital status, work type, residence type, average glucose level, BMI, and smoking status.

Exploratory Data Analysis:

1. **Missing Values:** BMI has 201 missing values, addressed through multivariate imputation.
2. **Imbalance Issues:**
    - Stroke occurrence is rare (249 out of 5111 cases).
    - Hypertension and heart disease are also underrepresented.
3. **Categorical Variables:**
    - work_type: "Never_worked" category was merged with "children" due to age correlation.
    - gender: "Other" gender category was excluded due to rarity.
    - smoking_satus: often “unknown”, potentially affecting model performance.
4. **Continuous Variables:**
    - Age is approximately evenly distributed, with a concentration between 40-60 years.
    - Avg_glucose_level shows bimodal distribution.
    - BMI has outliers (BMI > 65), which were excluded.

Correlations:

- Positive correlation found among age, BMI, and average glucose level.
- Stroke likelihood increases with age, higher glucose levels, hypertension, heart disease, being married, self-employment, and smoking status.
- Gender and residence type showed little correlation with stroke occurrence.

Feature Engineering:

- Binary categorical variables were encoded into binary formats.
- One-hot encoding was applied to multilevel categorical variables.
- Missing BMI values were imputed using correlations with other variables.

Modeling Approach:

- Scaling: Variables were standardized.
- Model Selection and Evaluation: Models tested include Logistic Regression, SVM, Decision Tree, Random Forest, Gradient Boosting, Histogram-based Gradient Boosting, and Neural Network.
- Hyperparameter Tuning: Utilized Grid Search with cross-validation to optimize models.
- Performance Metrics: Emphasis on recall score to minimize false negatives, despite reduced precision score.

Results:

- Best Performing Model: Logistic Regression with a recall of 0.9 but low precision of 0.13.
- Other Models:
    - SVM and Random Forest showed high recall but similar precision to Logistic Regression.
    - Gradient Boosting and Neural Network performed poorly on recall.

Conclusion:

The study successfully developed a classification model to predict stroke risk, with Logistic Regression emerging as the best model. However, the model faces challenges due to:

- Imbalance in the dataset (low stroke occurrence).
- Limited discriminative predictors.

Future improvements could involve collecting more detailed biological indicators to enhance model accuracy.

PROBLEM STATEMENT

The objective of the study is to propose a model that predicts whether a patient has stroke or not, based on descriptors (such as age, gender, existing disease, etc.)

The dataset contains 5111 observations with the following 12 variables:

5. **id**: unique patient identifier
6. **gender**: “Male”, “Female” or “Other”
7. **age**: age of the patient
8. **hypertension**: 0 (if the patient doesn’t have hypertension) or 1 (if the patient has hypertension)
9. **heart_disease**: 0 (if the patient doesn’t have a heart disease) or 1 (if the patient has a heart disease)
10. **ever_married**: “No” or “Yes”
11. **work_type**: “children”, “Govt_job”, “Never_worked”, “Private” or “Self-employed”
12. **Residence_type**: “Rural” or “Urban”
13. **avg_glucose_level**: average glucose level in the blood
14. **bmi**: body mass index
15. **smoking_status**: “formerly smoked”, “never smoked”, “smokes” or “Unknown” (in this case the information for the patient is not available)
16. **stroke**: 1 (if the patient had a stroke) or 0 (if the patient didn’t have a stroke)

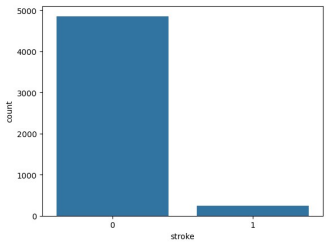
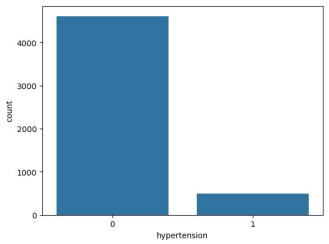
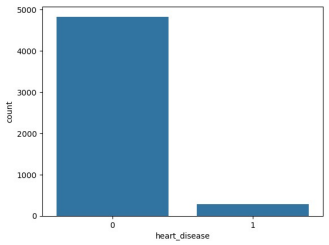
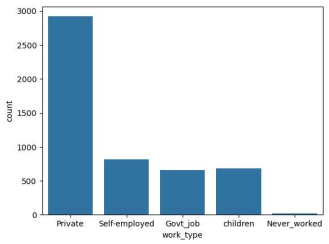
stroke is the target variable to be predicted. It is a binary value (0 or 1), hence the model to develop is a ***classification model***.

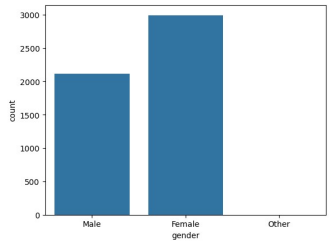
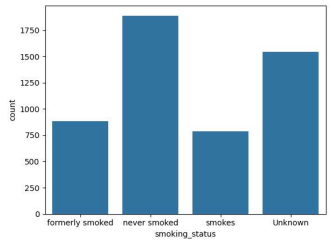
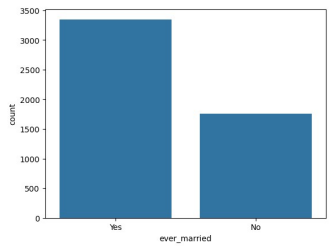
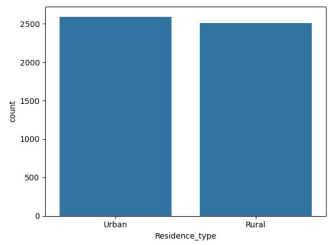
The rest of the variables, except id (used as an index), are the features or predictors, among which 3 are continuous (age, bmi and avg_glucose_level) and 8 are categorical variables (binary or multilevel).

EXPLORATORY DATA ANALYSIS

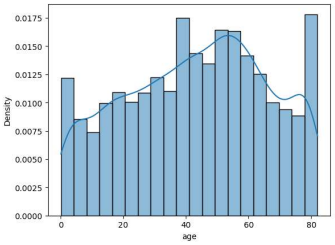
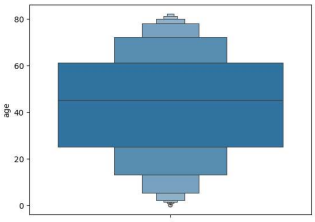
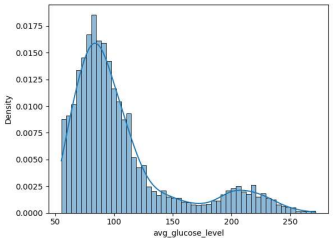
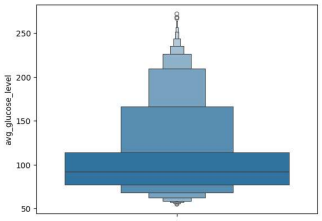
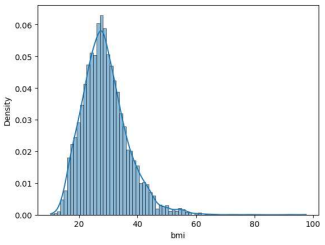
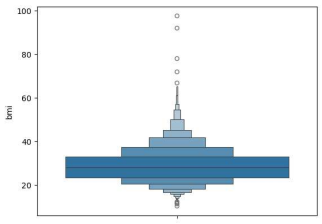
It is to be noted that bmi variable has 201 missing values and will be dealt with in section Features engineering.

EXPLORING UNIVARIATE CATEGORICAL VARIABLES

stroke variable shows a very high unbalance. Patients with stroke represent only 249 among the 5000+ observations. Unbalancing of target variable may be a problem to achieve good classification performance.	
hypertension and heart disease variables are unbalanced. Patients with hypertension or heart disease are underrepresented.	 
Over-representation of “Private” work_type is observed. Also, Never_worked category is highly under-represented. Exploring further “Never_worked” patient category shows the age of these 22 patients is ranging between 13 and 23 and none had a stroke. Therefore, decision was made to group “Never_worked” under “children”	

For gender, a reasonable balance between “Male” and “Female” categories is observed. However, “Other” category is very rare. Indeed, only 1 observation falls in this category, and it is a patient with no stroke). Hence, it was excluded from the dataset.	
smoking_status counts show that the status of a large proportion of patients is unknown. It could be detrimental to the model performance, since smoking is a risk factor for stroke.	
Most of the patients have been married.	
residence_type is equally distributed between “Urban” and “Rural” categories.	

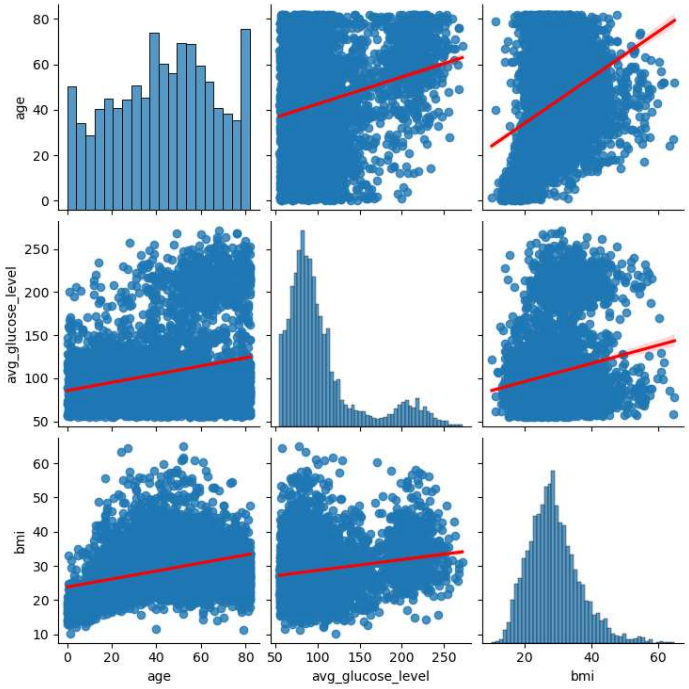
EXPLORING UNIVARIATE CONTINUOUS VARIABLES

age distribution is rather symmetric and close to uniform (although a higher representation of age range between 40 and 60 years can be seen on the histogram). No obvious outlying values are observed.	 
Avg_glucose_level shows a bimodal distribution. No obvious outlying values are observed.	 
bmi has an approximate symmetric distribution. However, 5 outliers are observed (bmi > 65, which is very uncommon). Further exploration of these 5 observations revealed the patients did not have stroke. So, decision was made to exclude them.	 

EXPLORING CORRELATIONS

Continuous variables - Pairwise correlations

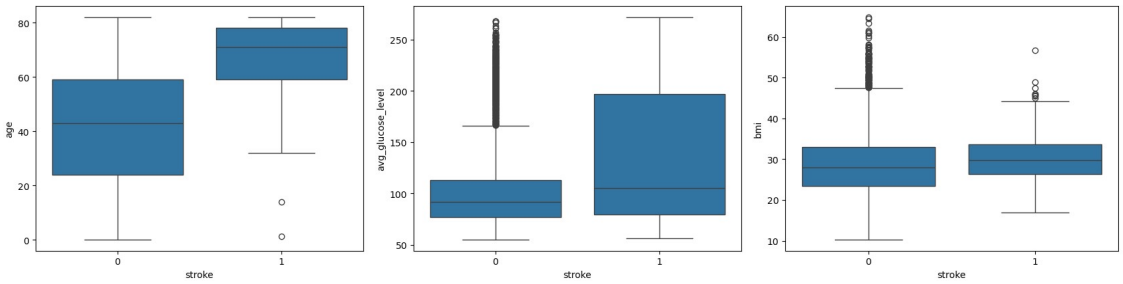
Pairwise plots of the 3 continuous variables reveal a positive correlation between bmi, age and avg_glucose_level. These correlations can help with estimating the missing bmi values.



Continuous variables – Correlation vs. target

Using box plots by stroke, it can be noticed that patients with stroke are older and have higher and (more dispersed) avg_glucose_level. These 2 features are likely to be interesting predictors in the classification model.

On the other hand, bmi correlation with stroke seems weaker. It is observed that higher bmi relates to stroke, but the difference in median or distribution between the 2 groups is small.

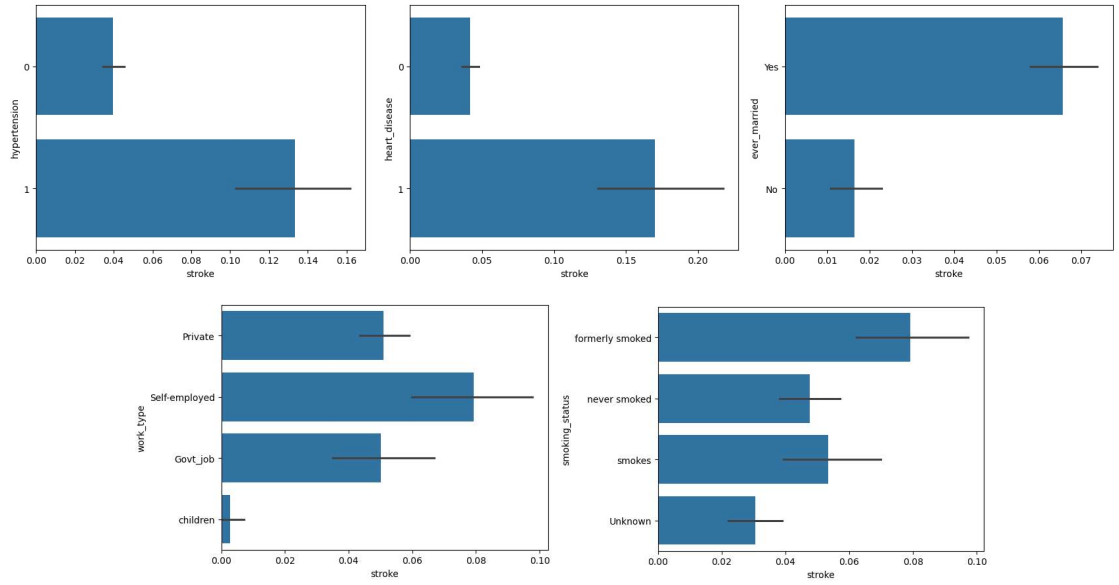


Categorical variables vs. target correlations

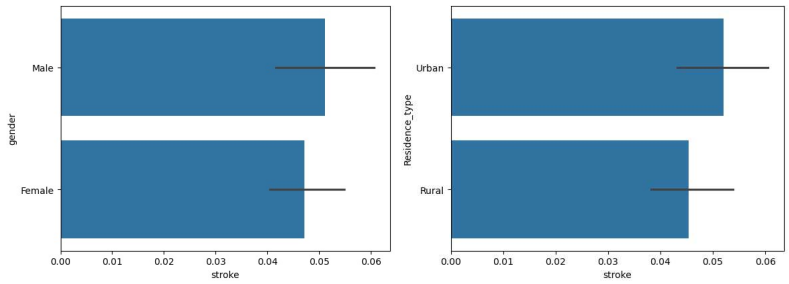
Based on contingency tables and bar charts, it can be confirmed a strong link of stroke vs.:

- hypertension and heart disease: both conditions yield a much higher risk of developing stroke.
- ever_married: being or having been married is correlated with stroke, but the link could be coincidentally related to age range (young people, such as children, are not married and as presented previously, age is a factor increasing stroke risk)
- work_type: self-employment increases risk of stroke
- smoking_status: former smokers and smokers are at high risk of stroke.

These 5 variables are therefore considered interesting predictors to include in the model.

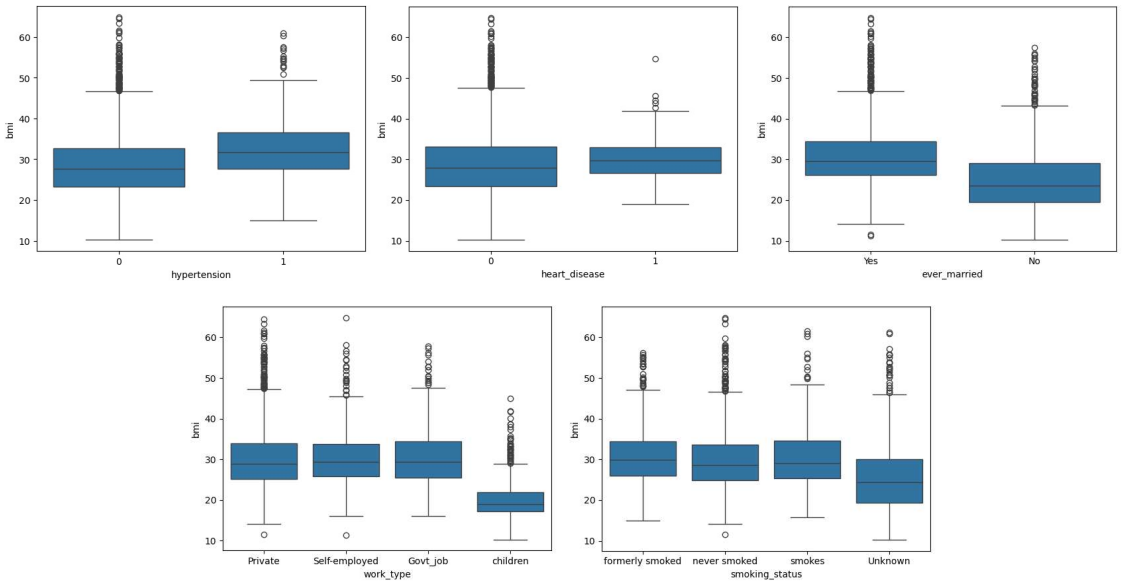


Gender and residence_type are not well correlated to developing stroke.



bmi – Correlations vs. categorical variables

A specific focus was put to find correlation between bmi and other variables. The purpose is to confirm relationships that could help estimating the missing bmi values. It was found slight dependency of bmi with hypertension, heart disease, ever_married, work_type and smoking_status.



FEATURES ENGINEERING

ENCODING CATEGORICAL VARIABLES

gender (after discarding “Other” category), ever_married and residence_type are binary variables for which the levels are expressed in text values. Therefore, they were converted to Boolean variables.

work_type and smoking_status are multilevel categorical variables. Therefore, they have been converted into binary variables using one-hot encoding.

IMPUTING MISSING BMI VALUES

bmi variable shows 201 missing values. As noted previously, bmi was found to be correlated with other variables (noticeably age, avg_glucose_level, hypertension, heart disease, ever_married, work_type and smoking_status). These correlations will be used to impute the missing values using a multi-variate method[1].

[1] 6.4. Imputation of missing values — scikit-learn 1.5.1 documentation

MODELLING

PREDICTING FEATURES SCALING

Some of the variables in the dataset are continuous (age, avg_glucose_level and bmi) while others are binary (post-feature engineering). So, in order to convert them on a comparable scale, a standard scaling was applied[2]. Each standardized variable is transformed by subtracting its mean and dividing by its standard deviation.

MODEL DEVELOPMENT AND EVALUATION, FINE-TUNING

The dataset was split into 2 groups: a train set and a test set (20% of the data). The split was stratified on stroke variable, to ensure similar proportions of patients with and without stroke in the 2 subsets. Obviously, the model is trained only using the train set. The test set is used to evaluate the fitted model performance and compare different solutions.

Several algorithms are available for classification. For this study, the list of models tested is:

- Logistic regression
- Support vector machine
- Decision tree
- Random forest
- Gradient Boosting
- Histogram-based gradient boosting
- Multi-layer perceptron (Neural Net)

Moreover, the performance of a classification algorithm highly depends on the chosen hyper-parameters. Therefore, it is recommended to assess several combinations and, using cross-validation method, choose the set of hyperparameters that yields the best model performance. One way of executing such search is referred to as Grid Search[3].

---

[2] StandardScaler — scikit-learn 1.5.1 documentation
[3] 3.2. Tuning the hyper-parameters of an estimator — scikit-learn 1.5.1 documentation

In this study, to facilitate evaluation of solutions based on various algorithms, a class called "**Classifier**" has been defined. Each object is instantiated by passing a base model to optimize (from the list above), a train and a test set, as well as the settings to apply for the hyperparameter Grid Search (notably, the space of hyperparameters and the scoring metric to use during the search: e.g. recall, precision or f1). Within the class Classifier, Methods are defined to train the model, get predictions from the test set and compute model performance metrics (such as precision and recall scores, confusion matrix and classification report). Once the model tuning is completed, the Classifier object stores:

- the base model (not tuned)
- the fitted model (which includes the best performing hyperparameter set)
- the training and testing data sets as well as the prediction for the test set
- the settings used for the hyperparameter Grid Search (such as the searched parameter space)

Finally, applying print() function to a Classifier object, prints a summary which includes:

- the type of classification algorithm and the set of hyperparameters yielding the best model performance
- the full set of parameters (to be used to setup an identical model)
- the performance metrics observed on test set (confusion matrix and classification report which includes important score such as recall, precision and f1)

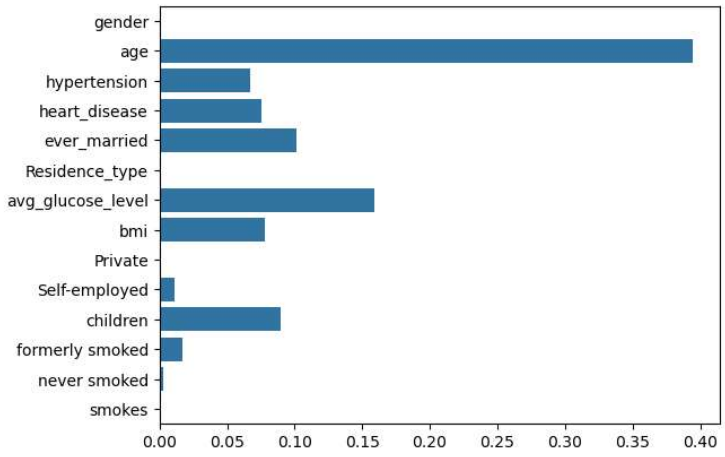
In this study, recall was chosen as the scoring method during the hyperparameters Grid Search. Indeed, the classifier should identify the patients with a risk of stroke and avoid miss-classify them as safe. In other words, reducing the rate of false negatives is highly important. Nevertheless, improving recall comes at the cost of reducing precision. It is therefore anticipated that the number of false positives can be high[4].

The results of developed and optimized models are summarized in the following table. It turns out that the best performing model is the Logistic Regression. Alternatively, Support Vector Machine and Random Forest are also high on recall score, with a slight advantage for Support Vector Machine (since it has a slightly better precision score, potentially reducing the rate of false positives).

[4] N.B.: To better balance the recall vs. precision scores, using F1 (instead of recall) scoring for the Grid Search was also tried. As anticipated, this yields slightly lower recall and slightly higher precision. But it is preferable in this case to keep the focus on recall, to avoid missing patients with a death threatening stroke condition.

Model	Best hyperparameters	Recall (for stroke=1)	Precision (for stroke=1)	F1 score (for stroke=1)
Logistic Regression	C=0.001, class_weight='balanced', l1_ratio=0.5, penalty='elasticnet', solver='saga'	**0.9**	0.13	0.22
Support Vector Machine	C=0.1, class_weight='balanced', gamma='auto', kernel='sigmoid'	**0.88**	0.13	0.23
Decision Tree	class_weight='balanced', min_samples_leaf=33	0.76	0.13	0.22
Random Forest	class_weight='balanced', max_depth=2, min_samples_leaf=27, min_samples_split=3, n_estimators=200	**0.88**	0.12	0.22
Gradient Boosting	max_leaf_nodes=40, min_samples_leaf=15	0.02	0.33	0.04
Histogram-based Gradient Boosting	class_weight='balanced', l2_regularization=100, max_leaf_nodes=50	0.76	0.17	0.27
Multi-Layer Perceptron	hidden_layer_sizes=(150, 75)	0.10	0.13	0.11

Additional valuable information can also be found thanks to the tree- and ensemble-based models: the importance of the features in the classification. As an example, for the Random Forest model, the importance of the predicting variables is presented on the following bar chart:



This graph confirms, at a great extent, the observations previously made in Exploratory data analysis (EDA) section: age, avg_glucose_level as well as categorical variables such as hypertension, heart_disease,

ever_married and children (i.e. young population, not having worked) are strong predictors to determine if stroke is a risk for the patient.

Noticeably, bmi turns out to be an important variable although the EDA did not show a strong correlation to stroke risk.

smoking_status, on the other hand, seems to be a moderately important variable to predict stroke. Being a former smoker seems to be indicative of a stroke risk. However, whether being a current smoker or not doesn't influence the classification. It is also likely that the smoking_status variable influence is undermined by the fact that a large number of patients are reported with unknown smoking_status.

CONCLUSION

In this study, the objective was to develop a classification model that would allow to predict the risk of stroke given a set of features about patients.

The exploratory data analysis revealed that stroke was correlated with:

- age: older patients are more at risk of stroke
- avg_glucose_level: higher level is associated with higher risk
- underlying disease: hypertension or heart_disease are risk factors
- lifestyle and habits: work_type and smoking_status are influential factor as well as marital status (but this variable may be coincidentally correlated with other variables such as age)

Multiple classification algorithms have been tested and the best performance was obtained with Logistic Reagression. Nevertheless, it must be noted that reaching a good recall score (0.9), i.e. reducing the risk of false negatives, comes at the price of a poor precision score (meaning a high rate of false positive stroke patients).

Achieving good performing classification models is difficult with this dataset for multiple reasons:

- the stroke cases are limited and represent less than 5% of the total number of observations
- the number and nature of predictors truly discriminative of a stroke risk, such as age, hypertension and heart disease, are limited. An improvement to the data could be the collection of biological indicators (such as blood tests) which may carry information on the risk of stroke.