Supervised Learning Project

The second leading cause of death worldwide is stroke according to the world health organization. It is a medical condition induced by an insufficient blood supply to the brain that can have a more or less serious impact depending on how quickly it is treated. Whenever it does not lead to death often leaves the survivor with some severe disabilities which might concern both the physical and psychological spheres. It is a very widespread pathology indeed among people over the age of 25 about 25% will experience stroke soon or later. Not only the studies about the biological and medical aspects of the disease furnish the necessary knowledge to fight it. Also the statistical analysis of inherent data can provide some answers useful for preventing it. In this paper a logistic regression has been applied to determine a model able to classify those patients who are more likely to have a brain attack, grounding the analysis on the value of several factors of risk. Subsequently a penalized logistic regression is run to perform subset selection and to determine the most relevant predictors.

I. Data Set

The dataset collects patients' information useful to predict the likelihood of having a stroke. For each of the 5100 observations 12 attributes are reported:

- **id**: unique number to identify the patient.
- **gender**: "Male", "Female" or "Other".
- **age**: age of the patient.
- **hypertension**: 0 if the patient doesn't have hypertension, 1 if the patient has hypertension. It concerns high blood pressure.
- **heart_disease**: 0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease. The scope is broad, it includes diseases such as coronary artery disease, arrythmia or atherosclerosis.
- **ever_married**: "No" or "Yes".
- **work_type**: "Children", "Govt_job", "Never_worked", "Private" or "Self-employed". The data includes information on children.
- **residence_type**: "Rural" or "Urban".
- **avg_glucose_level**: average glucose level in blood in mg/dl.
- **bmi**: body mass index. It is a measure of body fat based on weight and height.
- **smoking_status**: "formerly smoked", "never smoked", "smokes" or "Unknown".
- **stroke**: 1 if the patient had a stroke or 0 if not. It is a brain damage due to insufficient blood supply.

It collects measures of different types, nominal, continuous, discrete and dichotomous as shown below in table (1). Also, there are several missing values of different types, some blank, some N/As and some misleading information that can be considered as such.

The goal of the analysis is to detect a statistical learning model able to predict the risk of having a stroke with the best accuracy as possible as well as to determine the most relevant factors of risk.

Table (1). Original data set.

Id	*Gen.*	*Age*	*Hyp.*	*He_di.*	*Ev_m.*	*Wo_ty.*	*Re_ty*	*Avg_gl.*	*Bmi*	*Sm_sta.*	*Stro.*
9046	Male	67	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
51676	Female	61	0	0	Yes	Self-employed	Rural	202.21	N/A	never smoked	1
31112	Male	80	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1
60182	Female	49	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
1665	Female	79	1	0	Yes	Self-employed	Rural	174.12	24	never smoked	1
56669	Male	81	0	0	Yes	Private	Urban	186.21	29	formerly smoked	1
53882	Male	74	1	1	Yes	Private	Rural	70.09	27.4	never smoked	1
10434	Female	69	0	0	No	Private	Urban	94.39	22.8	never smoked	1

The dataset is available on Kaggle only for educational purposes.

II. Analysis

The dataset includes different types of missing values. In order to make the analysis more accurate these must be dropped or eventually transformed. The variables involved are *gender*, *bmi* and *smoking status.* The first, the easiest to treat, counts only one meaningless observation where sex is identified as "other". Since it is just one it can be dropped. Differently, the body mass index variable counts 201 N/As while the smoking status variable includes 1544 'unknown' records. The observations involved are too many to be eliminated because the resulting loss of information would be too high. Therefore they both have been substituted by using the K-nearest neighbour algorithm[1] based on the Gower distance[2]. Moreover individuals underage, as well as people who never worked, have been removed from the analysis. The former reported misleading information, e.g. 8 years old smoker, while the latter represented an insignificant fraction of the sample (0.43%), therefore it was dropped to simplify the analysis. The sample is now left with 4248 observations. The following step was about removing the column *id*, useless for our purpose and converting the type of some variables in order to make the data suitable for further processing.

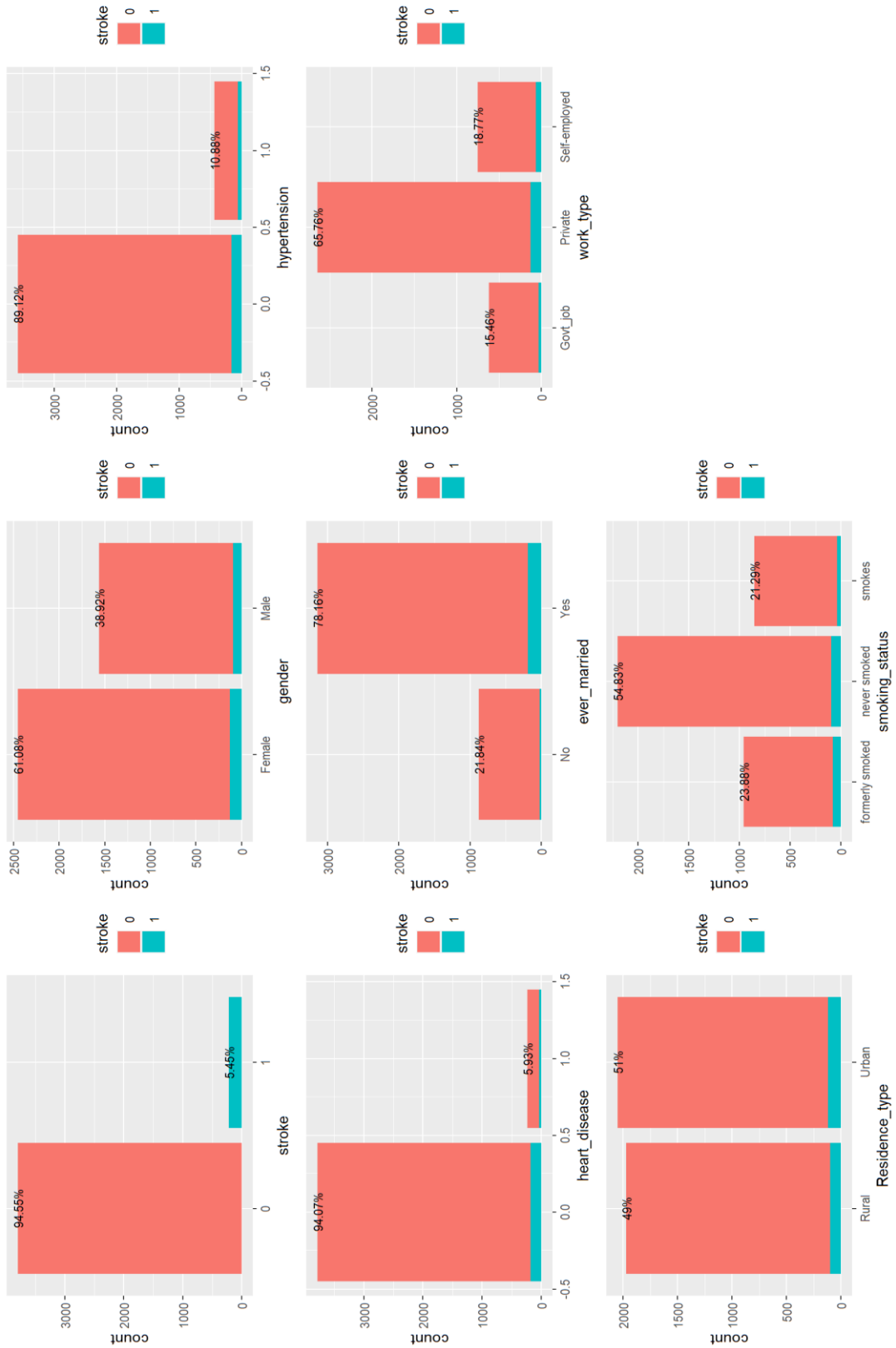
After the data explorative phase comes a quick descriptive data analysis. Here below in figure (1) are reported some histograms concerning all the categorical and dichotomous variables, visualizing the absolute frequency for each level as well as the fraction of patients who experienced a stroke[3]. The factors of risk that most showed a higher relative frequency of stroke cases are *hypertension*, *heart disease*, *smoking status* and surprisingly *ever married*. All the others did not show any suspicious and evident correlation with the number of stroke cases, at least according with such a superficial descriptive analysis.

[1] It assigns a value to the missing record according to the class majority -for categorical- or to the mean -for numerical- considering the values of the k closest observations.

[2] The distance between two different observations x_i and x_j is $s_{iJ} = \sum_{k=1}^{p} s_{iJk}\delta_{iJk} / \sum_{k=1}^{p} \delta_{iJk}$. Quantitative predictors: $s_{ijk} = 1 - |x_{ik} - x_{Jk}|/R_k$ where R_k is the range. Qualitative predictors: $s_{iJk} = 1\{x_{ik} = x_{Jk}\}, 0\{x_{ik} \neq x_{Jk}\}$. Where $0 \leq s_{ij} \leq 1$ and $\delta_{iJk} = 1$ if the values can be compared, 0 otherwise. Maximum similarity when s_{ij} is equal to 1.

[3] Coloured in teal.

Figure (1). Descriptive analysis of the categorical variables.



Another relevant finding that has to be taken into consideration is the fact that the dataset is very unbalanced in the response, as it is often the case with health data. Indeed, about 95% of the patients did not experience stroke while only 5% did so. The balancing of the dataset is required to make correct and meaningful predictions. If this fact is ignored also a naive classifier would obtain a high accuracy just by predicting the most common outcome, in this case, not having stroke. However such a classifier would make errors of only one type and would never be able to foresee a potential risk of stroke, which is actually the ultimate goal of the analysis. In order to settle this, right after the split of the data set into training and test, with a 3 to 1 ratio, the SMOTE[4] algorithm is applied to the training set. In this case the minority class is increased to match the size of the majority one by creating new synthetic data points via K-Nearest Neighbour algorithm. It is a more accurate oversampling technique with respect a random data augmentation based on the simple insertion of random copies of observations of interest since it mitigates better the risk of overfitting.

The last three features, *bmi*, *age* and *average glucose level* are numerical and their graphs report on the y axis the number of cases and on the x axis the values of the variable. Meanwhile they differentiate among stroke and not stroke cases, as reported in figure (2). While the number of cases who did not experience stroke begin to decline for ages above approximately 55, the number of cases experiencing stroke goes in the opposite direction. Such a tendency to increase seems to suggest some correlation among the age of the patient and the probability of having a stroke. However the other two variables did not show a similar trend given that the two lines, red and teal, seems to follow the same path on a different scale. Further investigation is required. To conclude the descriptive part of the analysis the *glucose level* is weakly correlated with both, *bmi* and *age* (16% and 22.8%), while only the former two present several outliers as reported in figure (3).

Figure (2). Numerical variables.

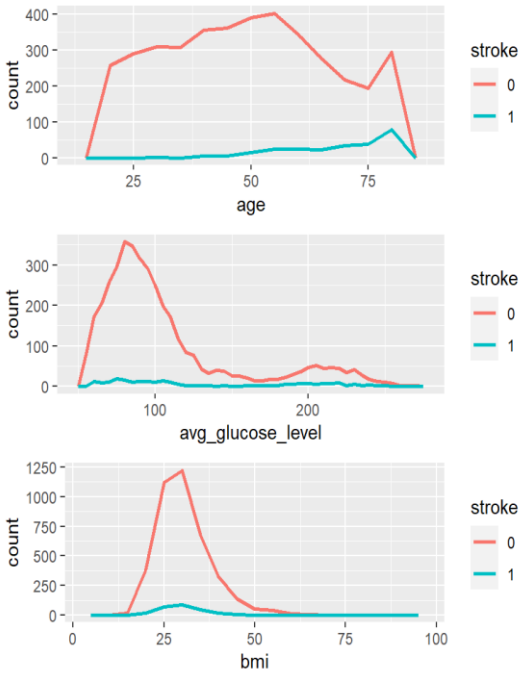
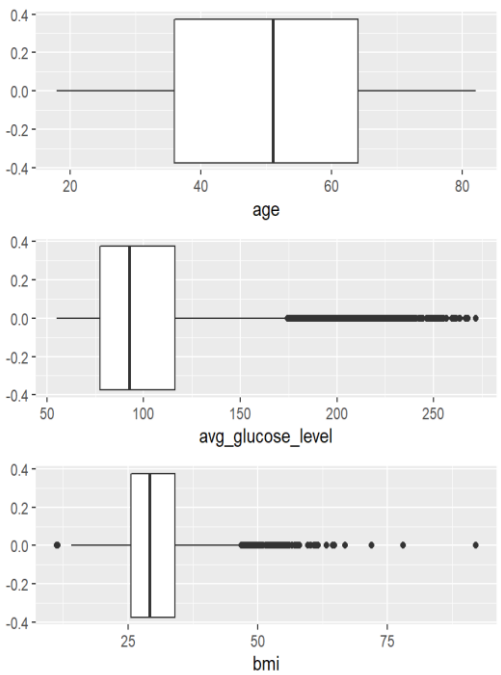


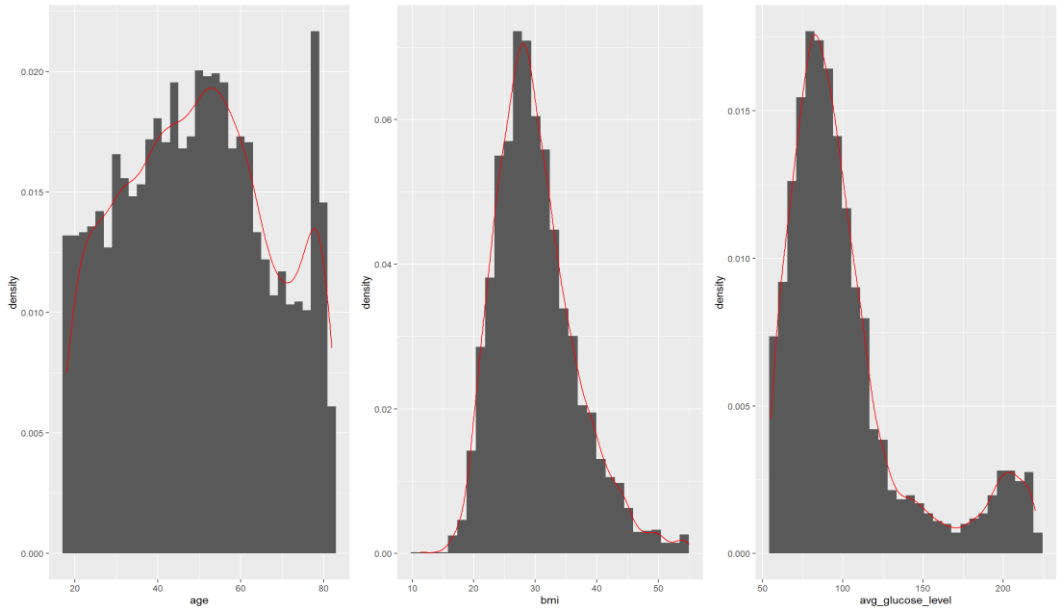
Figure (3). Box plots.



[4] Synthetic Minority Sampling Technique. First it selects a random observation of a positive class and then it creates a new synthetic data point from the observations selected via K-NN.

Before modelling the data the outliers should be managed since they might be too influential. The highest bound for *bmi* is 97.6, which is a feasible value, indeed even higher values have been recorded in history. However values above 55 are extremely rare. In fact 55 is the body max index of a person 1.75 meters tall weighting about 165 kilograms. In the data set considered there are 33 values scoring a bmi over 55 which are all dropped from the analysis. The same reasoning is applied to the observations scoring an *average glucose level* over 220 mg/dl, which are 205. The data set is finally left with 4016 individuals. Here below the density distributions of the aforementioned variables are shown.

Figure (4). Probability distributions of respectively age, bmi and average glucose level.



Next, the balanced training set is fit to the multiple logistic regression model. The model is of the kind:

$$\log\left(\frac{P(x)}{1-P(x)}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p \text{ [5] } (1)$$

The term on the left is called the log of the odds. The odds is a ratio between the probability of occurrence of the event of interest, in our case experiencing a stroke, and the probability of its complementary event, i.e. the probability of not having a stroke. For example an odds ratio of 3 means that the event of interest is 3 times more likely to happen than its counterpart. Given than the outcome is binary, and one of the two event has to take place, an odd of 3 corresponds to a probability of 75% . Finally the logarithm is taken to simplify the function on the right, linearizing it, where the betas are the coefficients to be estimated while x_{ip} is the score of feature p of observation i. Once the betas are estimated via the maximum likelihood[6] estimator, function (1) can be used to

[5] Derived from: $P(x) = \frac{e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}}{1+e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}}$. Where the term on the right is the logistic function.

[6] $L(y_1, y_2 \ldots y_n; \theta) = \prod_{i=1}^{n} f(y_i; \theta) \rightarrow \hat{\theta} = \max_{\theta} L(Y; \theta)$. For the binary logistic model, based on the Binomial distribution, the likelihood to be maximized is: $L(p) = \prod_{i=1}^{n} p^{y_i}(1-p)^{1-y_i} \rightarrow \ln(L(p) = \sum_{i=1}^{n}\left[\ln(1-p) + y_i \ln\left(\frac{p}{1-p}\right)\right] = \sum_{i=1}^{n}\left\{y_i\left(\beta_0 + \sum_{J=1}^{p} x_{iJ}\beta_J\right) - \ln\left(1 + e^{\beta_0+\sum_{J=1}^{p} x_{iJ}\beta_J}\right)\right\}$

classify a patient on the base of its factors of risk. However, since *P(x)* does not have a linear relationship with x_i the coefficients' interpretability is not straightforward, indeed it depends on the level of x_i. Despite this their sign is very meaningful in the sense that a positive one signals an increase in probability while a negative a decrease, for any level. The results of the multiple logistic regression are reported in table (2), here below.

Table (2). Logistic regression coefficients. In black the statistical significant coefficients at a 5% level.

Term	*Estimate*	*Std.error*	*Statistic*	*P.value*
(Intercept)	-5.0493018979753	0.29763694820505	-16.96463402284	1.50031492699e-64
age	0.08740162772391	0.00280008440436	31.213926118733	6.8959796481e-214
hypertension	0.34267222644341	0.104274823601886	3.2862412479518	0.00101534009815
heart_disease	0.41581760247275	0.129016596044696	3.2229776262946	0.00126865462649
avg_glucose_level	0.00610874435843	0.000746126572647	8.1872762375460	2.67204395627e-16
bmi	-0.0007318862475	0.00625642421635	-0.116981557238	0.90687466831826
gender_Male	0.21159077993491	0.07656292949857	2.7636191734129	0.00571642133760
ever_married_Yes	-0.6277141069125	0.114504069645392	-5.482024428097	4.20486153786e-08
work_type_Private	0.0557569229192	0.105306576834918	0.529472371005	0.59647780581037
work_type_Self.employed	-0.0142470381435	0.124001846669405	-0.114893757844	0.908529332848259
Residence_type_Urban	-0.0572776862731	0.0751185968349017	-0.762496754286	0.445763579568069
smoking_status_never.smoked	-0.9836897722777	0.0871147947494628	-11.29187958379	1.43925501476e-29
smoking_status_smokes	-0.2615433430941	0.106479006399636	-2.456290229761	0.014037973238570

The variables which showed a statistical significance at a 5% level are *age*, *hypertension*, *heart disease*, *average glucose level*, *gender*, *ever married* and *smoking status.* The first five have a positive impact on the probability of having stroke while the last two a negative one. As for *age*, *hypertension*, *heart disease*, *average glucose levels* and *gender* results are coherent with the scientific literature regarding this topic. On the other hand *ever married* and *smoking status* report curious results. According with the coefficients people who have being married have less chances to have a stroke with respect those who have never been. Furthermore the coefficients concerning smoke habits are meaningless. Having never smoked or currently smoking reduce the risk of getting a stroke with respect having formerly smoked. Clearly such a result is questionable and should be investigated more since the coefficients should have shown opposite signs. Once the test set is fit to the model the confusion matrix reports the following results:

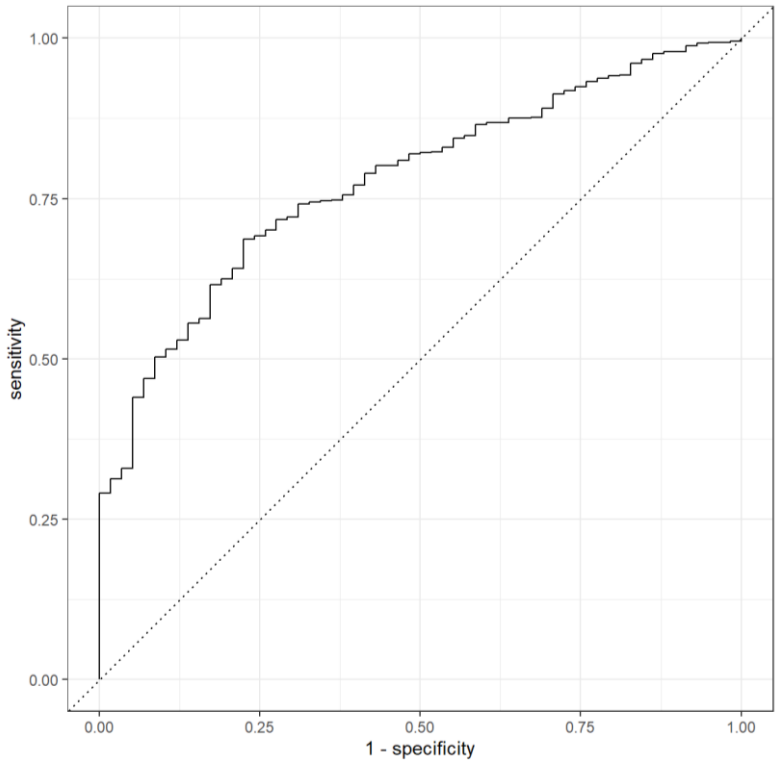
- 685 out of 946 true negatives have been detected.
- 40 out of 58 true positives have been detected.

The overall prediction accuracy of the model is 72.2%[7]. This means that the model predicts the true class in about 72% of the cases. Here below the ROC curve is displayed, a graph reporting the false positive rate on the x axis against the true positive rate on the y axis, for different classification thresholds[8].

[7] $\frac{\text{true negatives+true positives}}{\text{test set size}}$

[8] Represents the probability above which an observation is classified within a class.

Figure (5). Roc curve.



$$Sensisitivity = True\ positives/(true\ positives + false\ negatives)$$
$$1 - Specificity = False\ positives/(false\ positives + true\ negatives)$$[9]

A measure of performance for the logistic model is the area under the curve (AUC). Closer it is to 1 better is the model. In this case the AUC scores 0.777. However avoiding to drop the outliers improves the logistic model performances. Indeed the AUC increases of about 2%, up to 0.797, ceteris paribus. Outliers in this case make the algorithm more accurate. Instead, as intuited, avoiding to apply the SMOTE algorithm, leaving the dataset unbalanced, leads to high but misleading performances in terms of predictability. Indeed about 95% of observations are classified correctly, unfortunately all in the most common class, that is, not having stroke. On the other hand standardization[10] has controversial results. If the model without outliers is implemented, the accuracy is modestly improved by standardizing the numerical features[11]. Such an improve does not occur with the model including extreme observations[12]. Indeed by standardizing the variables the model becomes a little less accurate. Finally, with great amazement, the model having the highest prediction accuracy is the one trained on the full dataset, including all the outliers (children as well) without standardizing[13]. The percentage of observations correctly classified is 77.6% and the AUC is 0.846.

[9] Sensitivity represents the portion of positives correctly classified. Similarly 1-specificity represents the portion negatives misclassified. By decreasing the threshold the two quantities increase up to 1 since more positives are predicted.

[10] $x_i \to \frac{x_i - u_x}{\sigma_x}, \forall x_i \in X$. Data are centred and scaled.

[11] AUC=0.815.

[12] AUC=0.791.

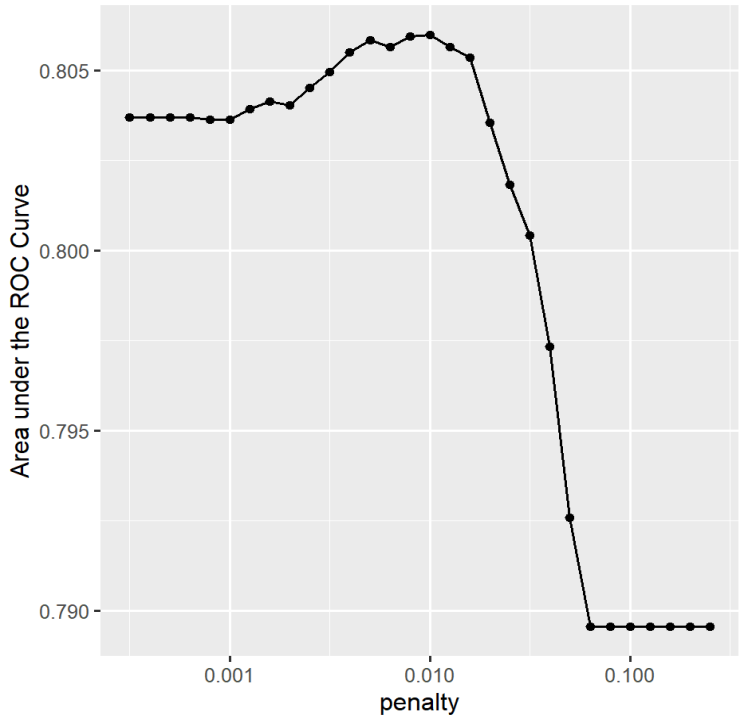
[13] Not required for classic logistic regression.

The next step consists in applying the LASSO[14] technique to the logistic model in order to reduce the variance, enhance interpretability and perform features selection. Such a method implies the maximization of the following log-likelihood function:

$$\sum_{i=1}^{n}\left\{y_i\left(\beta_0+\sum_{J=1}^{p}x_{iJ}\beta_J\right)-\ln\left(1+e^{\beta_0+\sum_{J=1}^{p}x_{iJ}\beta_J}\right)\right\}-\lambda\sum_{J=1}^{\rho}|\beta_J|$$

The last term of the equation represents the LASSO penalty responsible for constraining the magnitude of the coefficients. The *l1-norm*, the penalty associated with the absolute value, differently from the *l2-norm* of the ridge regression[15], assigns to some of the coefficients a 0 value. In this sense it performs features selection, by discarding some of the less significant variables out of the model. The tuning parameter λ determines the severity of the penalty, higher it is, more coefficients are driven to 0. However since the penalty is sensible to the magnitude of the coefficients the data must be scaled. In this analysis all the numerical predictors are standardized, including the dummies. The lambda parameter is not defined a priori indeed first it must be tuned to select the most optimal value. The tuning is performed by subsetting the training set into two different data sets, one used for training the algorithm and the other one for validating it. In this way models with different penalty levels, in this case 30 models, could be compared. The matric used for assessing the best model is once again the value of the AUC. The results are shown graphically in figure (6).

Figure (6). AUC associated with 30 models, each with a different lambda.



By increasing furtherly the value of the penalty the curve shown above decreases monotonically towards 0.5. The first 17 models score a similar AUC, in the range 0.804-

[14] Least absolute shrinkage and selection operator. It can be applied to both, regression and classification problems.

[15] $\lambda\sum_{J=1}^{p}\beta_J^2$.

0.806. The value picked for the penalty is 0.0126, the furthest to the right before the graph begins to decline. The following steps consist in fitting a logistic regression with the selected LASSO penalty to determine the relevant features. Coefficients are reported in table (3).

Table 3. Logistic regression with LASSO coefficients.

term	***estimate***	***penalty***
(Intercept)	-1.01029432408413	0.0126
age	1.51568526742402	0.0126
hypertension	0.00587379491528013	0.0126
heart_disease	0.0202337575418476	0.0126
avg_glucose_level	0.162205772172752	0.0126
bmi	0	0.0126
gender_Male	0	0.0126
ever_married	0	0.0126
work_type_Private	0	0.0126
work_type_Self.employed	0	0.0126
Residence_type_Urban	0	0.0126
smoking_status_never.smoked	-0.330	0.0126
smoking_status_smokes	-0.0237	0.0126

The variables *bmi*, *gender*, *ever married*, *work type* and *smoking status* have been discarded by the LASSO from the previous model. Those features are those having the weakest correlation with the response.

III. Conclusions

Assessing whether a patient has high or low probability of experiencing a stroke is of undoubtful importance in order to prevent the occurrence of such a dangerous medical condition. Statistical learning methods can provide a powerful tool to support difficult decisions concerning individuals' health. With the aim to build a predictive model able to determine the likelihood for an individual of experiencing a stroke a dataset reporting several factors of risk has been analysed.

Given the classification nature of the problem a logistic regression has been applied to model the probability of having stroke. The model with the best prediction accuracy is the one trained on the whole dataset, including outliers, classifying correctly about 77% of the observations of the validation set.

Furthermore, by applying a penalized logistic regression via LASSO, the most important features for predicting the response have been identified. Those with a positive impact on the probability of exhibit the symptoms are *age*, *hypertension*, *heart disease*, *average glucose level* and *smoking status*, coherently with the scientific literature concerning this topic. However the last variable was associated with deceptive coefficients. Those who never smoked, as well as those who are smoking, have less probability of experiencing a brain attack with respect those who used to smoke. Both dummies scored negative coefficients while one should expect the coefficients to go in opposite direction. This might be due to the bad quality of the data. Indeed during the analysis several incoherences concerning this variable have been encountered.