Stroke Prediction

Developing a predictive model for stroke events

Statistics for Data Science

Group 2

Don Bresee

Heather Hainsworth

Maggie Lau

Melina Raptis

Fan Ye

Echo Zhang

April 18, 2022

Table of Contents

Objective

According to the Ontario Stroke Network, Strokes are the third highest cause of death and the number one leading source of disability in Canada. Strokes are caused by an interruption in blood flow to the brain (or part of), which happens when the blood vessels get blocked or rupture. When a stroke occurs, the brain cells in the affected area die due to the lack of glucose and oxygen. The longer the obstruction, the more likely the brain will become permanently damaged. Almost 50,000 Canadians will be affected by a stroke each year and 14,000 Canadians will die from strokes annually. Strokes also have an estimated $3.6 billion dollar cost per year to the nation in medical costs and other incidental costs [1]. In short, strokes are a common and costly healthcare issue; therefore it is important to understand the factors that might cause a stroke in order to reduce the likelihood of a stroke.

The Ontario Stroke Network further identifies that there are several risk factors that contribute to whether a patient will experience a stroke. Age is a major factor as the risk substantially increases as a patient ages. In fact, the risk of stroke doubles every decade after the age 55. Gender, ethnicity, blood pressure, cholesterol levels, weight, smoking and diet are just a few other examples of other risk factors for stroke [1]. This study will also take into account other available attributes from patient data that are generally not identified as typical risk factors for stroke such as marital status and area of residence.

The goal of this Data Analysis study is to predict the likelihood of a patient getting a stroke based on given parameters from a dataset of patients. Machine learning models were used to predict the probability of someone having a stroke. In order to determine which patients are more at risk. Ensuring that the proper medical resources are allocated to these groups to mitigate / reduce future strokes in high risk patients.

Data Preparation

The data set used is sourced from the link below:
https://www.kaggle.com/fedesoriano/stroke-prediction-dataset?select=healthcare-dataset-stroke-data.csv

The dataset consists of 5,110 observations with 12 attributes. These 12 attributes are listed below[2]:

1. ID - an anonymous identifier for each individual patient
2. Gender: describing female, male and other
3. Age
4. Hypertension
5. Heart Disease
6. Ever_maried - whether the patient has ever been married
7. Work_type - is the patient employed, self-employed, unemployed, etc.
8. Residence_type - If the patient lives in an urban area or rural area
9. Avg_glucose_level - Average glucose level measured in patient's blood
10. BMI - body mass index

11. Smoking_status - If the patient never smoked, smoked, smoked in the past
12. Stroke - If the patient experienced a stroke

Data Cleanup

The data quality in the dataset used was good however a few attributes required further cleaning to improve its usefulness.

ID - This column was dropped in our analysis as it was simply an identifier for each patient.

BMI - This column had 201 instances of missing data. The boxplot (Figure 1) shows that BMI follows a normal distribution except for some outliers. Therefore, the decision was made to fill missing data with the median value of the BMI from the dataset which was 28.1.

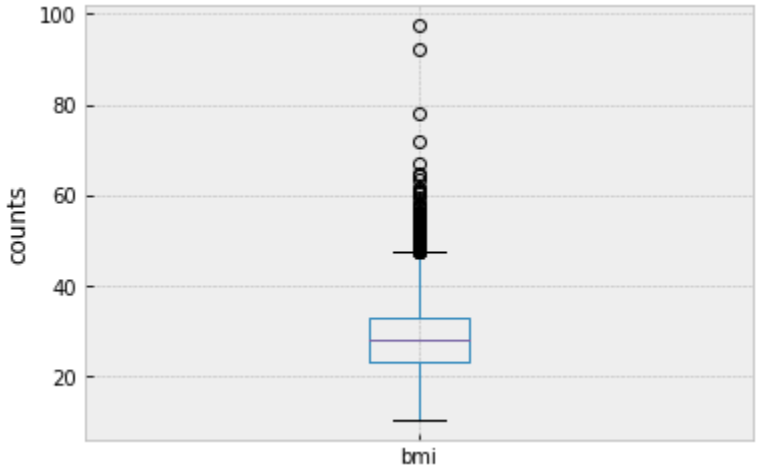


Figure 1. Boxplot of BMI in original database

Gender - There was 1 instance of “Other” identified in the gender column. The decision was made to drop this row from the dataset.

Smoking_Status - Within this column, there were 1,544 instances of “Unknown” which was 30.2% of the dataset. As this was a substantial proportion of the dataset, it would not be ideal to drop the missing data. Further action was required to fill the missing values based on observations from the other available attributes.

The crosstab function was used to compare smoking_status with various other attributes. It was found that residence type and hypertension had little relation with smoking status. However it was found that children or people who never worked were more likely to have never smoked. People who work in private companies have a high chance of smoking. It was also observed that people who are employed and have no heart disease, likely never smoked. People who were working and do have heart disease, if they were never married, they were likely to still be smoking.

Taking these observations into account, the following decisions were made to substitute the missing values in the smoking_status column:

- If the patient was a child or had never worked, then smoking status was set to ‘never smoked’
- If the patient did not have heart disease, then smoking status was set to ‘never smoked’
- If the patient did have heart disease, and was never married, then smoking status was set to ‘smokes’
- If the patient had heart disease and was married, then smoking status was set to ‘formerly smoked’

With the above rules, all rows of smoking_status were filled to either “formerly smoked”, “never smoked”, or“smokes” to complete the dataset.

Preparation for Analysis

Since a “stroke” event is represented as a binary variable (yes or no) the logistic regression model was selected as our primary classification algorithm to train and test the data model. In order to do so, all categorical data was expressed as dummy variables. The binary variables ‘gender’ (Male = 1; Female =0), ‘Ever_married’ (Yes = 1; No=0) , and ‘residence_type’ (Rural = 1; Urban = 0) were converted to nominal variables with only two values of 0 and 1. For the variables ‘work_type’ and ‘smoking_status’ that had more than two levels, one hot encoding was used to create 5 and 3 dummy variables respectively for use in the model. After cleaning and conversion, the dataset had 5,109 observations and 17 attributes including “gender”, “age”, “hypertension”, “heart_disease”, “ever_married”, “Residence_type”, “avg_glucose_level”, “bmi”, “stroke”, “Govt_job”, “Never_worked”, “Private”, “Self-employed”, “children”, “formerly smoked”, “never smoked” and “smokes”.

Model Development and Prediction

Model Building

To begin building the model, the dataset was split into two parts. 70% of randomly selected instances were used to train the model while the other 30% was used to test the model. The training dataset had 3,576 instances with 3,416 of no stroke and 160 of stroke, suggesting the dataset was not-balanced. To balance two classes in “stroke”, the Synthetic Minority Oversampling Technique (SMOTE) was applied to inflate the undersampled class and resample the dataset. The final training dataset had 6,832 instances with 50% of stroke and 50% of no stroke.

The approach that was taken to build the model was to create an initial model using the training set with 70:30 split with the logistic regression model from the Statsmodels library. Then the logistic regression model would be repeated using the resampled (balanced) data for comparison. Other common classification algorithm models were also developed. The best model can be obtained through cross validation by checking predictions against the test dataset. The results of these models are detailed in the following section.

Analysis of Models

Initially the logistic regression model was developed based on the original non-balanced data. The summary for this initial model can be viewed in the Appendix - Analysis of Models - Figure 1. The logistic regression model was repeated again using the balanced data. The summary and analysis for this subsequent model can be viewed in the Appendix - Analysis of Models - Figures 2 to 4. The model parameters were optimized using the Newton Method. All the attributes in the dataset were selected at the beginning. The model evaluation parameters such as Pseudo R-squared and P>|Z| were further improved by tuning the predictors. It was found that the model was optimized when removing the predictors of “Never worked” and “Smokes”. The two models were analyzed for accuracy, Matthews Correlation Coefficient (MCC), precision, recall, specificity, F1 score and AUROC (Area Under the Receiver Operating Characteristic). The results of the analysis are compared as shown.

	Type	Non Balanced	Balanced (threshold=0.5)	Balanced (threshold=0.4)
0	Accuracy	0.941944	0.869537	0.846706
1	MCC	0.068306	0.142420	0.167205
2	Precision	0.500000	0.155280	0.155660
3	Recall	0.011236	0.280899	0.370787
4	Specificity	0.999307	0.905817	0.876039
5	F1-Score	0.021978	0.200000	0.219269
6	AUROC	0.809697	0.772071	0.772071

Table 1. Comparison of Logistic Regression Model, Non-balanced vs. Balanced

The prediction based on model training of unbalanced data has a high true negative rate but a very low true positive rate. As a result, the accuracy, specificity and precision score are higher, while the MCC, recall and F1-Score are much lower than that of the balanced data. Additionally, the discrimination threshold of the prediction was tuned to optimize the recall, MCC and F1-score. It was found that the optimal prediction was achieved when the threshold value was set as 0.4. Overall the model built on the balanced data is an improved model.

The next approach taken was to develop several common models from the sci-kit learn library, such as Decision Tree Classifier, Random Forest Classifier, AdaBoost Regression, K-nearest Neighbors (KNN) and Support Vector Machines. The models were subsequently analyzed and compared against each other to determine the optimal model for this dataset. Below is the comparison summary chart for the models. More detailed results for each model can be viewed in the Appendix - Other Classification Models.

Type	Accuracy	MCC	Precision	Recall	Specificity	F1-Score	AUROC
Non Balanced	0.941944	0.068306	0.500000	0.011236	0.999307	0.021978	0.809697
Balanced (threshold=0.5)	0.869537	0.142420	0.155280	0.280899	0.905817	0.200000	0.772071
Balanced (threshold=0.4)	0.846706	0.167205	0.155660	0.370787	0.876039	0.219269	0.772071
Balanced (SKLearn)	0.869537	0.142420	0.155280	0.280899	0.905817	0.200000	0.593358
Decision Tree Classfier	0.895629	0.111649	0.155340	0.179775	0.939751	0.166667	0.559763
Random Forest Classifier	0.923679	0.103489	0.195652	0.101124	0.974377	0.133333	0.537750
Support Vector Machine	0.712329	0.222402	0.134855	0.730337	0.711219	0.227671	0.720778
Ada Boost Classifier	0.867580	0.139694	0.152439	0.280899	0.903740	0.197628	0.592319
KNeighbors	0.800391	0.183718	0.144262	0.494382	0.819252	0.223350	0.656817

Table 2. Comparison Summary of Models

The results were also plotted visually to compare each characteristic for the models. Below are charts comparing accuracy, specificity, precision and area under ROC (AUROC). The best performance for these characteristics was the logistic regression model from non-balanced original data.

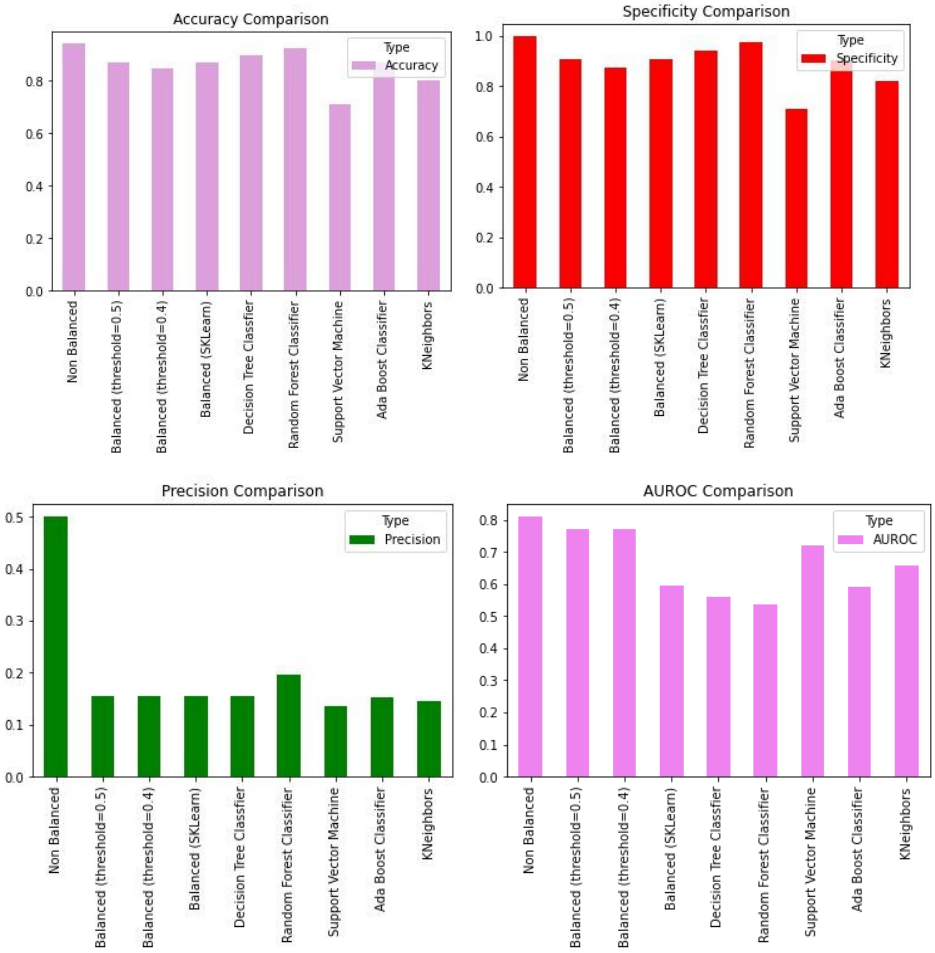


Figure 2, 3, 4, 5 (top left to bottom right) - Accuracy, Specificity, Precision, AUROC comparison

However the MCC comparison shows that in actuality the model built from non-balanced data is not a good predictor due to the lowest MCC result.

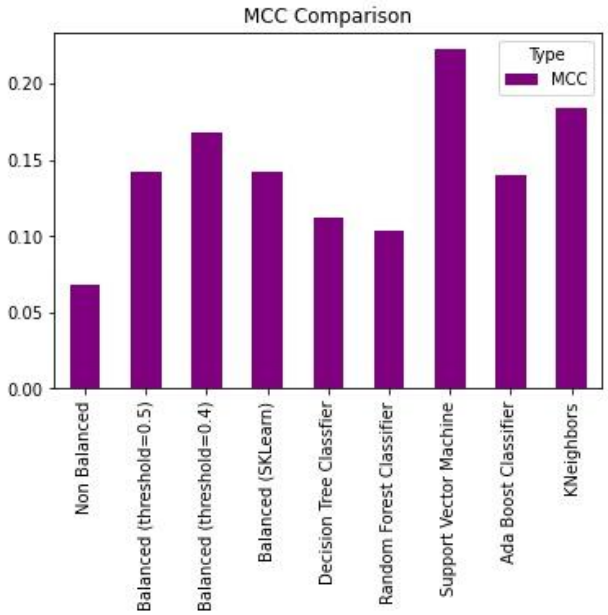


Figure 6 - MCC Comparison

The chart below for recall shows an advantage with the balanced logistic regression model over the non-balanced. However other models such as Support Vector Machine and KNN fare better.

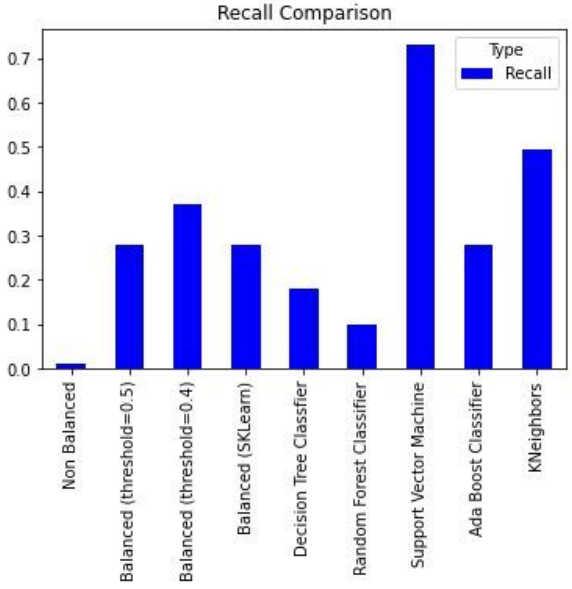


Figure 7 - Recall Comparison

The F1-score is used when the False Negatives and False Positives are crucial and it is a better metric when there are non-balanced classes. In the context of stroke prediction, the F1 score shows the logistic regression model from the balanced data is the better model compared to

non-balanced logistic regression, though again the Support Vector Machine and KNN models scored high as well.

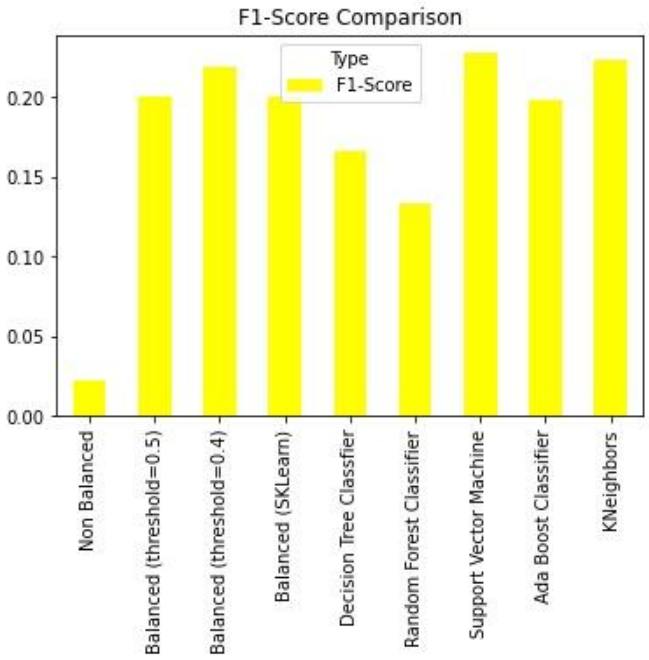


Figure 8 - F1-Score Comparison

Stroke Risk Factors based on the logistic regression model using the balanced dataset

The Logistic Regression (Balanced data) - Model Summary (see figure 2 in appendix) indicates that all selected indicators are statistically significant since the p-value is less than 0.05. Because of this, the coefficient of each input variable can be used to identify the stroke risk factors.

For quantitative variables, a positive coefficient indicates an increased risk of stroke. From the model, we can tell that by increasing the age, the bmi and the avg_glucose_level the likelihood of a stroke increases.

For the gender variable (recalling Male :1 and Female:0), we can say that $e^{-0.63}$ =0.53 is the odds ratio that associates being a male with the risk of stroke. This means being a male is associated with 47% (1-0.53=0.47) reduction in relative risk of stroke.

Similarly, for the ever_married (recalling Yes = 1; No=0) and residence_type (Rural = 1; Urban = 0) variables, we have negative coefficients, indicating being married and living in a rural residence also reduces the relative risk of stroke.

To avoid multicollinearity in the logistic regression model, one reference variable was assigned to each of the categorical variables. The reference variable for the work_type is ‘Never_worked’ and the reference variable for the smoking status is ‘smokes’. Since all the odds ratio are negatives, we can tell that ‘Never_worked’ and ‘smokes’ also increase the risk of stroke.

Conclusions

Based on the stroke prediction dataset and the analysis performed using the logistic regression model it was concluded that predictors can be used to identify patients at high risk of having a stroke. To determine this the models were optimized by balancing the data and tuning the predictor features and discrimination threshold. The balanced data resulted in improved MCC, recall, and F1 score over the non-balanced data leading to a better model.

The analysis of other models within the SKLearn library showed various models performing better in various characteristics. The Support Vector Machine and K Nearest Neighbours both out-performed other models in many characteristics like MCC, recall, and F1-score.

These models could be used to predict the likelihood of whether a patient is at risk of stroke. The model validated that increasing Age, BMI and Avg. glucose levels increase the probability of a stroke. Additionally, being a male is associated with a 52% reduction in the relative risk of a stroke, being married, and living in a rural residence also reduced the relative risk of a stroke. With this information the patient could be better informed on their overall health risks and the patient's future risk of stroke could be reduced by implementing preventative health measures (diet, exercise etc.) and allocating additional medical resources (more health monitoring etc.).

Appendix

Analysis of Models

Figure 1: Logistic Regression (Original Non-balanced data) - Model Summary

```
                           Results: Logit
================================================================
Model:              Logit            Pseudo R-squared: 0.184
Dependent Variable: stroke           AIC:              1652.0916
Date:               2022-04-14 04:01 BIC:              1743.6342
No. Observations:   5109             Log-Likelihood:   -812.05
Df Model:           13               LL-Null:          -995.14
Df Residuals:       5095             LLR p-value:      3.0437e-70
Converged:          1.0000           Scale:            1.0000
No. Iterations:     10.0000
----------------------------------------------------------------
                  Coef.  Std.Err.    z     P>|z|   [0.025  0.975]
----------------------------------------------------------------
gender           -0.0981   0.1393  -0.7041 0.4813 -0.3712  0.1750
age               0.0688   0.0055  12.4760 0.0000  0.0580  0.0796
hypertension      0.4818   0.1636   2.9447 0.0032  0.1611  0.8024
heart_disease     0.2625   0.1942   1.3521 0.1763 -0.1180  0.6431
ever_married     -0.1107   0.2248  -0.4924 0.6224 -0.5514
0.3299
Residence_type   -0.1365   0.1365  -1.0000 0.3173 -0.4041
0.1311
avg_glucose_level  0.0034   0.0012   2.8375 0.0045  0.0011
0.0058
bmi              -0.0327   0.0103  -3.1832 0.0015 -0.0529
-0.0126
Govt_job         -5.8177   0.4885 -11.9101 0.0000 -6.7751
-4.8603
Private          -5.6676   0.4616 -12.2782 0.0000 -6.5724
-4.7629
Self-employed    -6.0306   0.4958 -12.1633 0.0000 -7.0024 -5.0589
children         -5.2674   0.7497  -7.0265 0.0000 -6.7367 -3.7981
formerly smoked  -0.3637   0.1965  -1.8506 0.0642 -0.7489  0.0215
never smoked     -0.6453   0.1749  -3.6904 0.0002 -0.9880 -0.3026
================================================================
```

real	0	1
prediction		
0	1443	88
1	1	1

Figure 2: Logistic Regression (Balanced data) - Model Summary

```
                           Results: Logit
================================================================
Model:              Logit            Pseudo R-squared: 0.650
Dependent Variable: stroke           AIC:              3345.6768
Date:               2022-04-14 04:01 BIC:              3441.2880
No. Observations:   6832             Log-Likelihood:   -1658.8
Df Model:           13               LL-Null:          -4735.6
Df Residuals:       6818             LLR p-value:      0.0000
Converged:          1.0000           Scale:            1.0000
No. Iterations:     9.0000
----------------------------------------------------------------
                  Coef.  Std.Err.    z     P>|z|   [0.025  0.975]
----------------------------------------------------------------
gender           -0.6345   0.0946  -6.7104 0.0000 -0.8199 -0.4492
age               0.1134   0.0036  31.2733 0.0000  0.1063  0.1205
hypertension     -0.6049    0.1401   -4.3188 0.0000 -0.8794
-0.3304
heart_disease    -0.9741    0.1814   -5.3692 0.0000 -1.3297
-0.6185
ever_married     -1.0048    0.1233   -8.1485 0.0000 -1.2465
-0.7631
Residence_type   -0.9598    0.0943  -10.1769 0.0000 -1.1447
-0.7750
avg_glucose_level  0.0036    0.0009    4.1949 0.0000  0.0019
0.0053
```

real	0	1
prediction		
0	1267	54
1	177	35

```
bmi                0.0209    0.0064    3.2656 0.0011   0.0084   0.0335
Govt_job          -6.5106    0.2951  -22.0649 0.0000  -7.0889  -5.9323
Private           -5.1162    0.2574  -19.8749 0.0000  -5.6207  -4.6117
Self-employed     -6.5086    0.2915  -22.3286 0.0000  -7.0800  -5.9373
children          -2.4866    0.3399   -7.3153 0.0000  -3.1528  -1.8204
formerly smoked   -2.3653    0.1339  -17.6603 0.0000  -2.6278  -2.1028
never smoked      -2.2922    0.1077  -21.2835 0.0000  -2.5033  -2.0811
==============================================================================
```

Figure 3: Logistic Regression (Balanced data) - Logit (p) chart

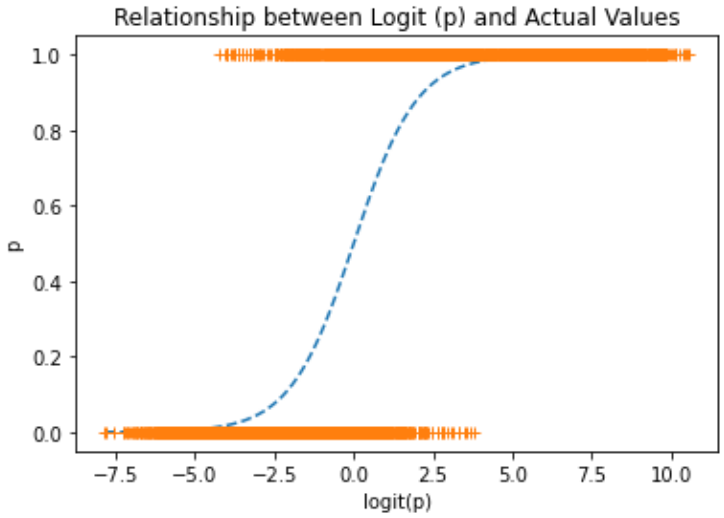
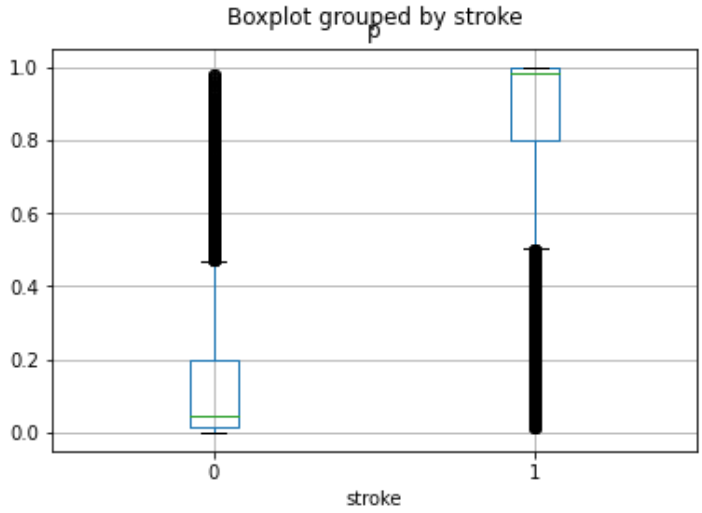


Figure 4: Logistic Regression (Balanced data) - distribution of the estimated odds for both values of the actual response

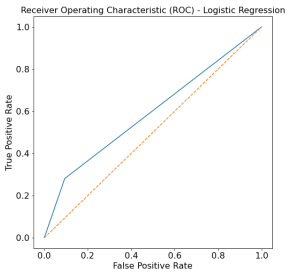


Other Classification Models

The following sections show the results from the alternate models.

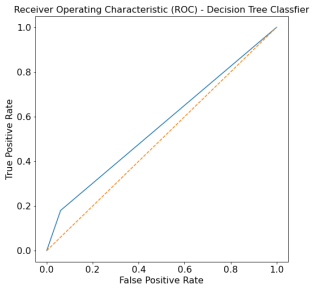
Balanced Data Logistic Regression Model using SKLearn

real	0	1
prediction		
0	1305	63
1	139	26



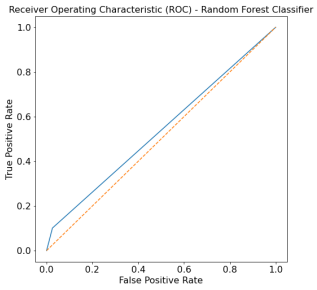
Decision Tree Classifier

stroke_pred	0	1
stroke		
0	1355	89
1	73	16



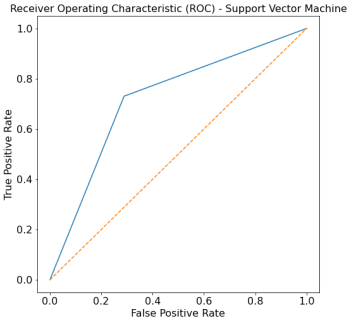
Random Forest Classifier

stroke_pred	0	1
stroke		
0	1404	40
1	79	10



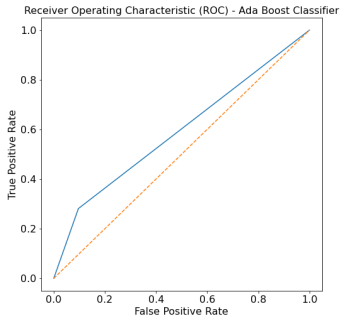
Support Vector Machine Model

prediction	0	1
real		
0	1027	417
1	24	65



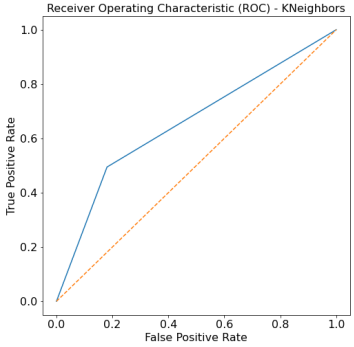
AdaBoost Regression Model

stroke_pred	0	1
stroke		
0	1290	154
1	62	27



K Neighbors Classifier

stroke_pred	0	1
stroke		
0	1190	254
1	46	43



References

[1] Ontario Stroke Network. (n.d.). *Fact sheet: Stroke statistics*. Retrieved April 12, 2022, from http://www.ontariostrokenetwork.ca/pdf/Final_Fact_Sheet_Stroke_Stats_3.pdf

[2] (Confidential Source) - Use only for educational purposes, Stroke Prediction Dataset, Unknown, from https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset?select=healthcare-dataset-stroke-data.csv