Stroke Prediction EDA & ML

Petropoulos Panagiotis

Student, MSc in Artificial Intelligence,

University of Piraeus & NCSR Demokritos,

Athens, Greece, June 2022

1. Abstract

Stroke is quite harmful for an organism, especially for the older. As a result, there is a need to be able to predict the probability of risk in a direct and easy way for a stroke to occur. Using big data through sensors, we can now record data for each patient, to perform analysis. The specific research based on medical data, focuses on creating machine learning models to predict the likelihood of a stroke occurring. There is also a Mockup from the creation of the corresponding application and the analysis results from EDA procedure.

2. Introduction

Stroke, also known as brain attack, occurs when blood, flowing to the brain, is blocked. There are many possible reasons that may lead to this. In this project an Exploratory Data Analysis (EDA) has been performed and a predictive model for the risk of Stroke has been created. In addition, many times there is a need to be able to explain the predicted results. For this reason, the creation of a tool that explains the results, could not be missing.

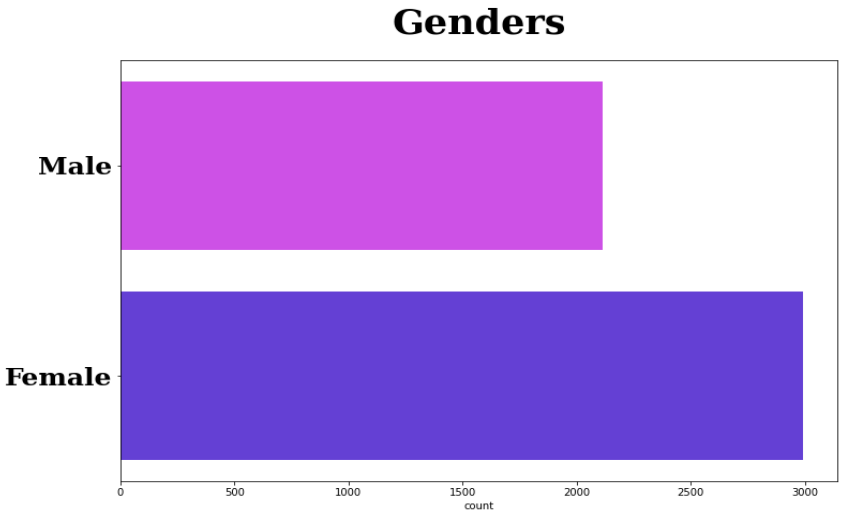
3. Related Work

Several Data Scientists have dealt with this project such as the publication of José Alberto Tavares Rodríguez where he achieved a 92% F-score with Random Forest Classifier [1]. Also, good results can be seen in the publication entitled "Stroke Prediction Using Machine Learning" [2]. The present project experiments with different methods and proposes the best of them. Finally, a multi-platform tool was created and focuses on the explainability of the Classifier's decision through visualizations.

4. Data

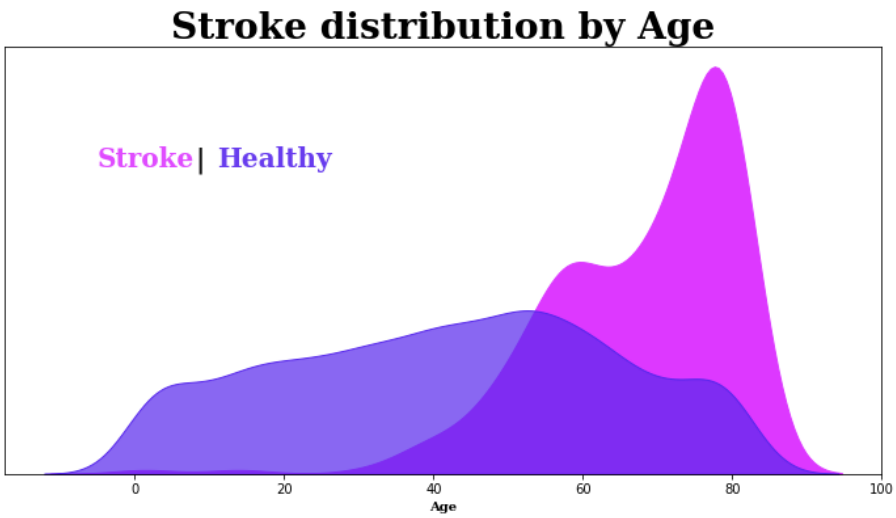
The Data, for this project, were collected from Kaggle in csv format and they are annotated. This csv file consists of 5110 rows (samples) with the following column as features:

1. Gender:
    a. Male: almost 2100 samples
    b. Female: almost 3000 samples



Plot-1: Distribution of Gender.

2. Age: a distribution for all ages.



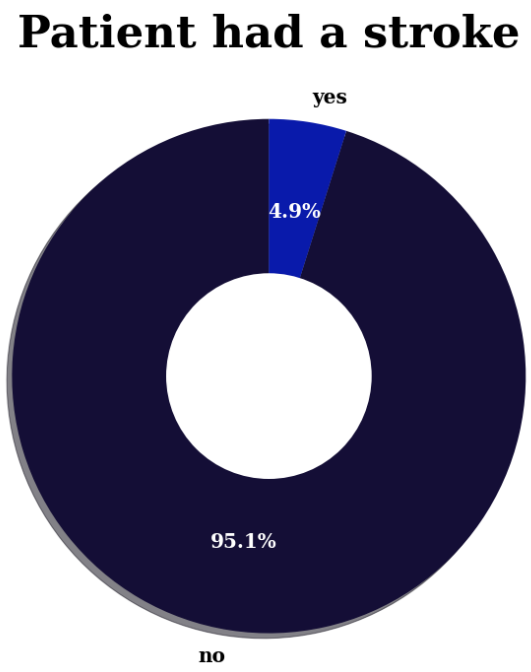
Plot-2 Distribution of Stroke per age.

3. Hypertension:
    a. Yes
    b. No

4. Heart Disease:
    a. Yes
    b. No
5. Ever Married:
    a. Yes
    b. No
6. Work Type:
    a. Private
    b. Self-employed
    c. Govt-job

    d. Children
    e. Never Worked
7. Residence Type:
    a. Urban
    b. Rural
8. Average Glucose Level
9. BMI: Here we had our first challenge as you can see further down in this paper, due to the fact, that many values of this feature were null.
10. Smoking Status:
    a. Formerly Smoked
    b. Never Smoked
    c. Smokes
    d. Unknown

The Dataset contains measurements from both gender types Male and Female. Also, the data had a big imbalanced problem in target label (Stroke), as we can see below:



Plot-3 Distribution of Target Label.

To deal with this problem, I performed the SMOTE technique for oversampling.

Finally, many values from BMI were missing and filled with 2 different approaches:

1. Multivariate Linear Regression based on features.
2. Mean Value of BMI.

5. Exploratory Data Analysis (EDA)

To make the research more meaningful, an analysis of the data is presented below.

General Information about our data is shown below:

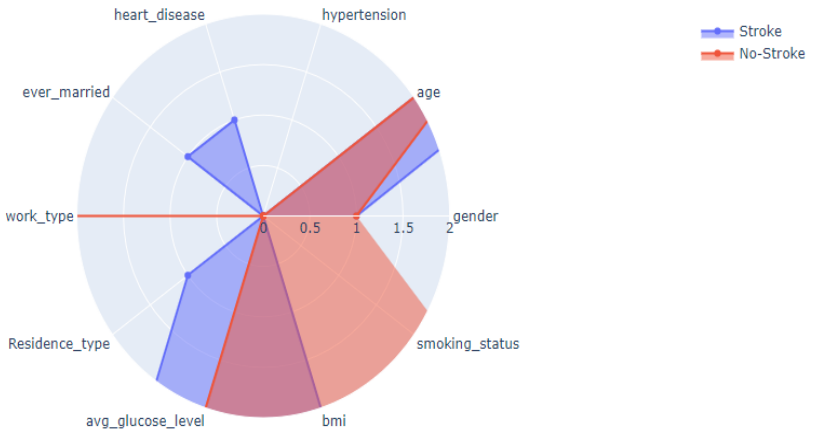
	count	mean	std	min	25%	50%	75%	max
id	5110.0	36517.829354	21161.721625	67.00	17741.250	36932.000	54682.00	72940.00
age	5110.0	43.226614	22.612647	0.08	25.000	45.000	61.00	82.00
hypertension	5110.0	0.097456	0.296607	0.00	0.000	0.000	0.00	1.00
heart_disease	5110.0	0.054012	0.226063	0.00	0.000	0.000	0.00	1.00
avg_glucose_level	5110.0	106.147677	45.283560	55.12	77.245	91.885	114.09	271.74
bmi	4909.0	28.893237	7.854067	10.30	23.500	28.100	33.10	97.60
stroke	5110.0	0.048728	0.215320	0.00	0.000	0.000	0.00	1.00

Table-1: General information about the values of Data.

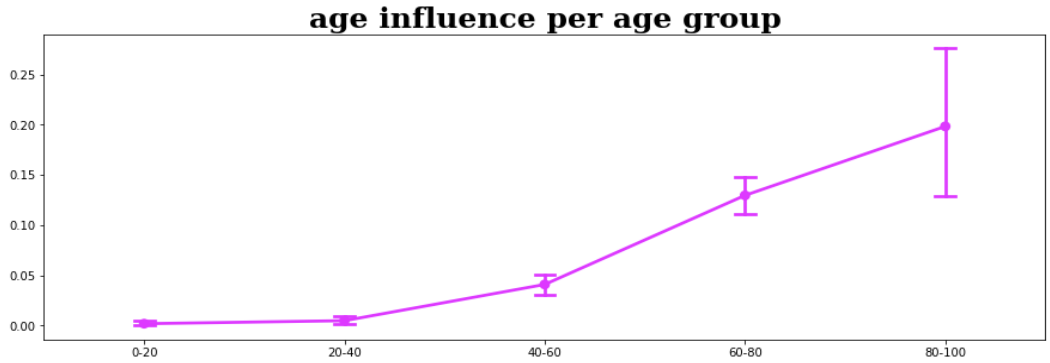
In below radar plot we can observe that features:

- Age
- Average glucose level
- BMI

Can affect the risk of Stroke more than the other features.

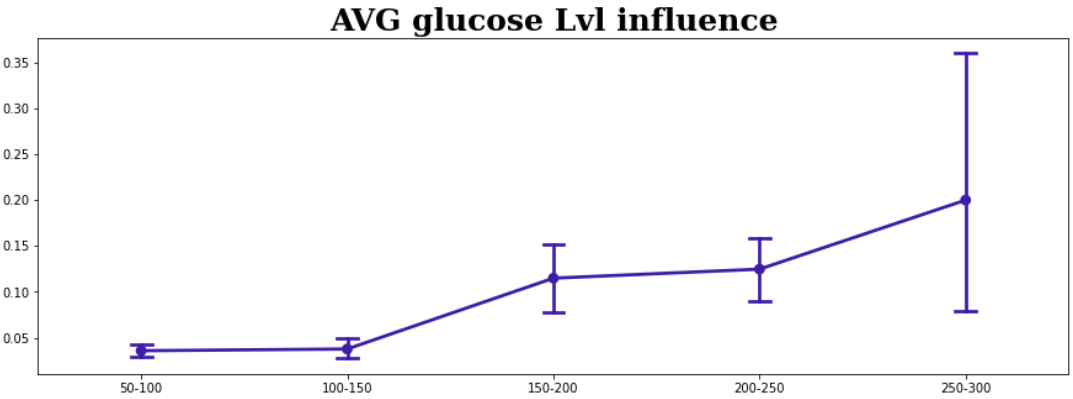


Plot-4: Influence of features for having Stroke (blue) and no Stroke (Red).



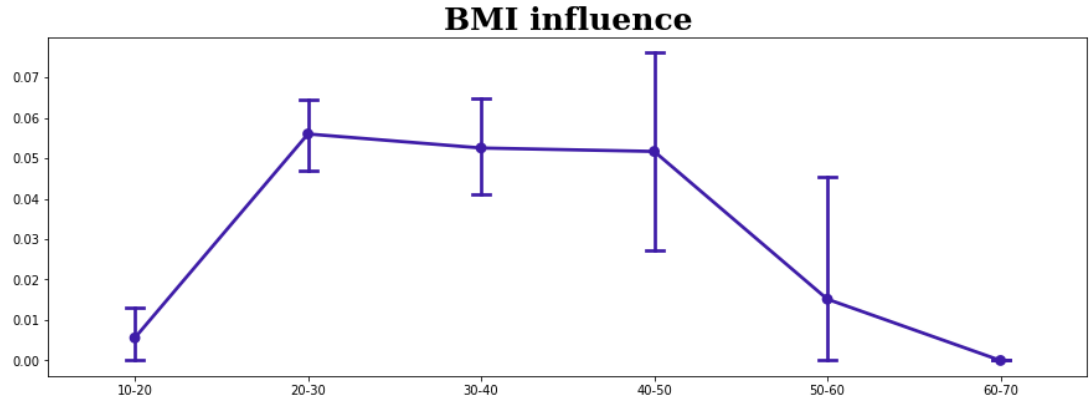
Plot-5: Influence of Age for having Stroke.

In the above plot, people who had Stroke aged in 80-100, appear in the 20%-25% of the Dataset. That means that the age affects the possibility of Stroke.



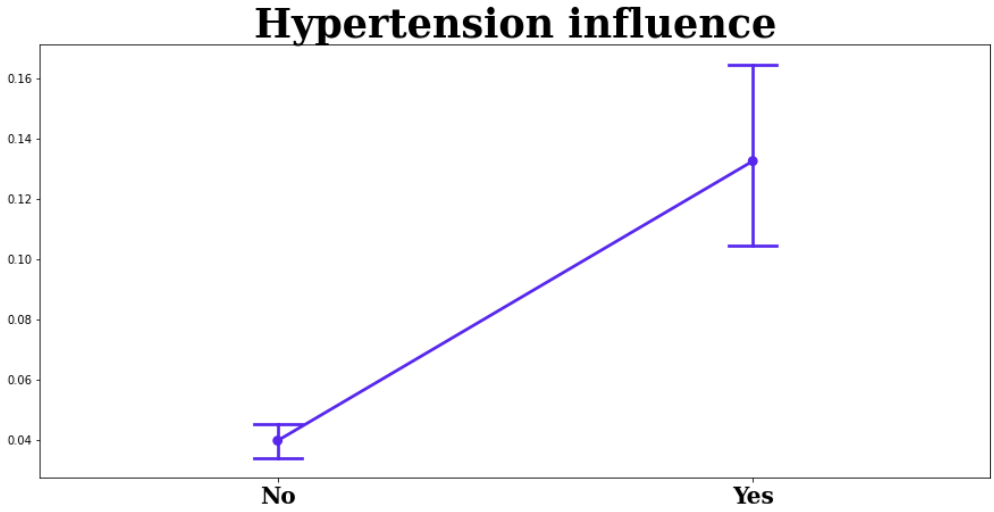
Plot-6: Influence of average glucose level for having Stroke.

Most people having a stroke usually have a higher Average glucose level than those who never had a stroke.



Plot-7: Influence of BMI for having Stroke per group of BMI values.

Average Glucose level and BMI values can affect the possibility of Stroke, as we can see from previous plots. Especially people with BMI values between 30 and 50 and a high glucose level, have high probability of having a Stroke. It would be helpful to look at the other features too.



Plot-8: Influence of having hypertension for Stroke.

The number of people who has hypertension is much lower than the number of people who do not. The proportion of people who had a stroke in the hypertension category is much higher than the proportion of people who had a stroke in the No hypertension category. But the number of people who had hypertension is significantly lower than the number of people who didn't, so it cannot be confidently concluded that people with hypertension is more likely to suffer from a stroke than people with no hypertension

As an extra note, we can observe that someone who has younger age, low glucose level, lower than 30 BMI value and lives in Rural region, has low risk of having a Stroke.

6. Machine Learning (ML)

6.1 Method

First of All, due to the fact that the dataset has values in different scales, I performed normalization with the MinMaxScaler library from the sklearn package. After that, two approaches were followed to deal with the missing values of BMI.

1. Multivariate Linear Regression based on features.
2. Mean Value of BMI.

For each of the above methods, I performed the below experiments:

1. An approach with feature selection with X^2 statistical test.
2. An approach without feature selection.

The above experiments were executed for the below Classifiers:

1. Naïve Bayes.
2. Logistic Regression.
3. SVM.
4. Random Forest.
5. Gradient Boosting classifier.

For the procedure of K-fold cross validation with K=5 folds, I split the dataset in a stratified way with 80% for training and 20% for validation procedure.

As an additional experiment, I trained a Gradient Boosting Classifier without feature selection and normalization procedure, filling the missing values of BMI with its mean value.

Finally, an online, multi-platform and easy-to-use tool has been created, in order to predict the Risk of Stroke and explain the results. The explainability procedure is performed with SHAP[1] values. Based on them, for each feature a metric is calculated. This metric shows to the end-user how much the corresponding feature affects the prediction, calculating the log-odds using Shapley values from game theory.

7. Application

Dash package[2] and framework is used to create this application and the user interface. A user interface mockup is shown below. In the below image we can see some predicted results in a dashboard. The dashboard is an easy-to-use tool. In this way, we are able to have a useful and simple user interface. The group of Risk level (Low, Medium, High) are designed and created based on the optimal threshold of ROC curves from the corresponding Classifier. Due to the fact that we are interesting about the Stroke, classifier predicts only the probability of having Stroke, based on the patient profile that user will give as input.

[1] https://shap.readthedocs.io/en/latest/index.html
[2] https://dash.plotly.com/

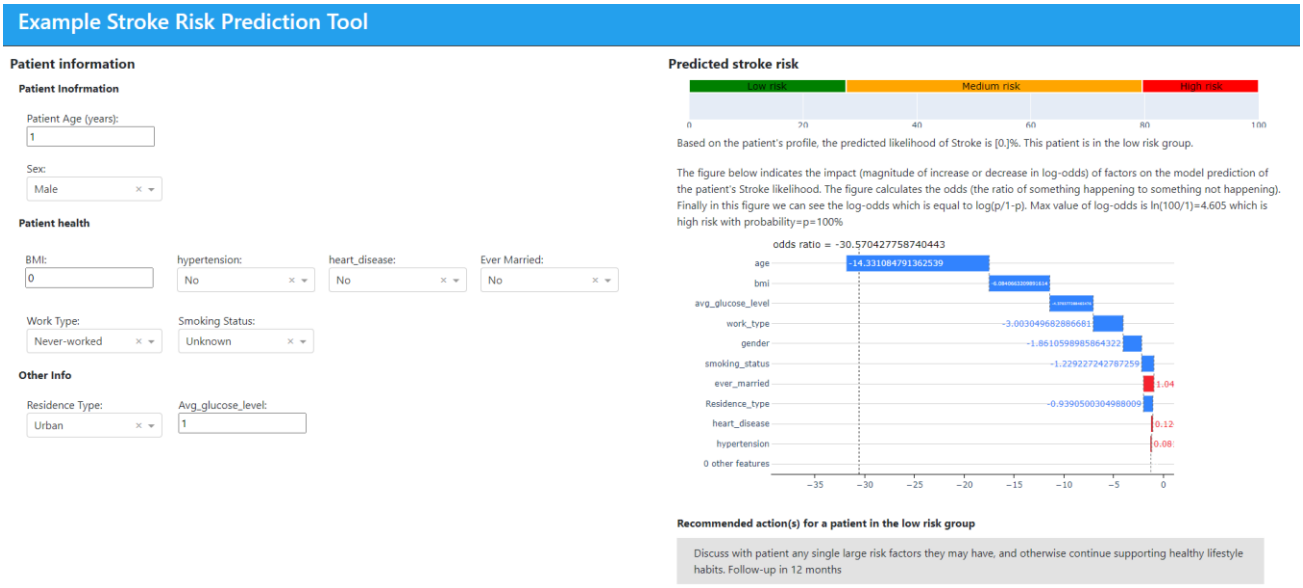


Image-1: Default Home Page before prediction.

The above image shows us the default User interface before a user puts the values in the corresponding fields and the classifier makes a prediction. Red bars show us that the corresponding feature affects the likelihood of a Stroke to happen. In the other side the blue bars show us the opposite.

Image-2: Low Risk Prediction.

Image-2 shows us a Low-risk prediction with likelihood 4.4%, based on the user's profile.

- Age: 59
- Gender: Male
- BMI: 36.6
- Hypertension: No
- Heart Disease: No
- Ever Married: No
- Work Type: Private
- Smoking Status: Formerly smoked
- Residence Type: Urban
- Average Glucose Level: 228.69

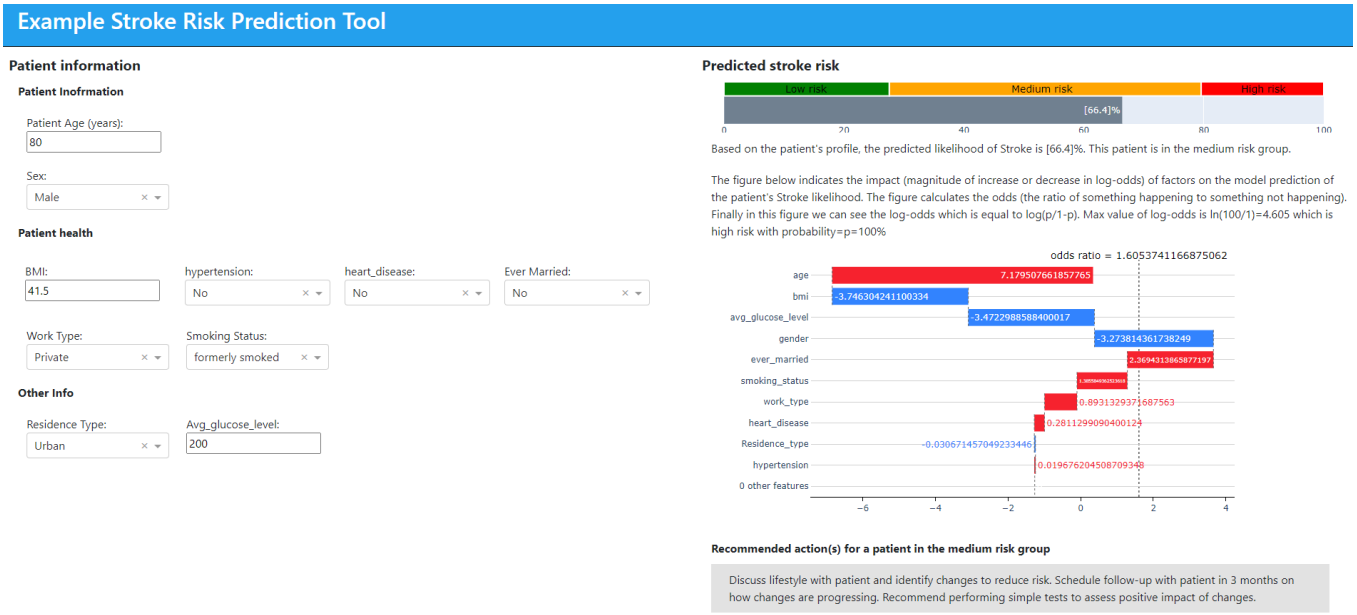


Image-3: Medium Risk Prediction.

Image-3 shows us a medium-risk prediction with likelihood 66.4%, based on the user's profile.

- Age: 80
- Gender: Male
- BMI: 41.5
- Hypertension: No
- Heart Disease: No
- Ever Married: No
- Work Type: Private
- Smoking Status: Formerly smoked
- Residence Type: Urban
- Average Glucose Level: 200

As we can see from the user's profile the feature Age affects the prediction.

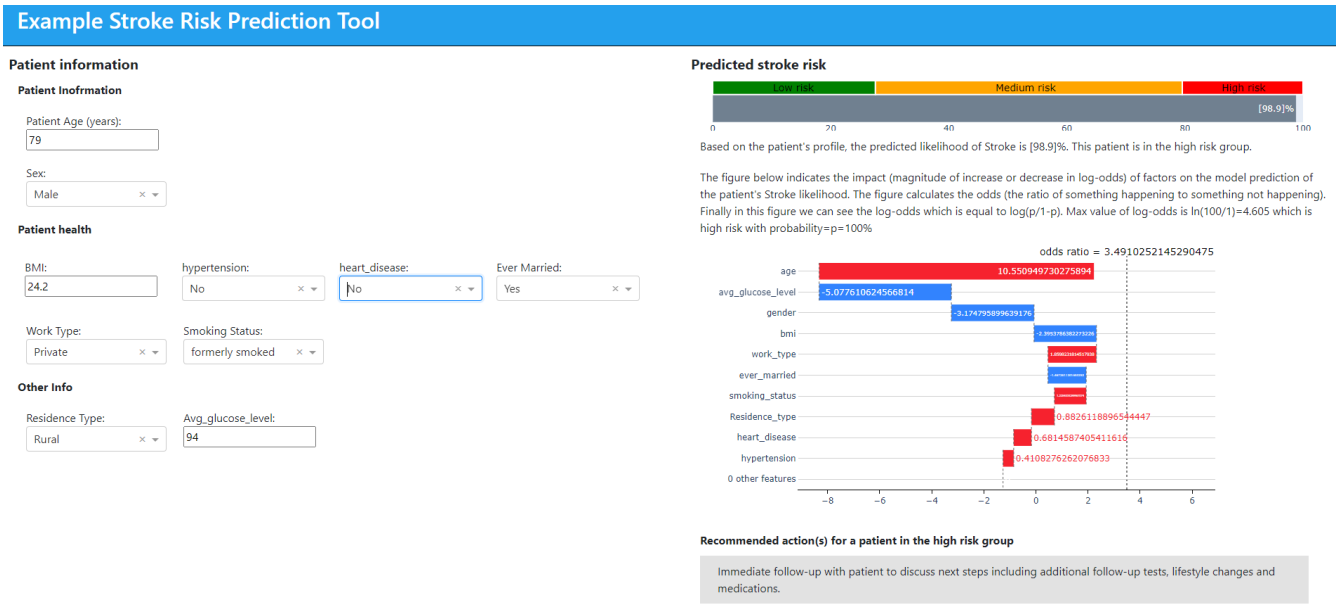


Image-4: High Risk Prediction.

Image-4 shows us a high-risk prediction with a likelihood of 98.9%, based on the user's profile.

- Age: 79
- Gender: Male
- BMI: 24.2
- Hypertension: No
- Heart Disease: No
- Ever Married: Yes
- Work Type: Private
- Smoking Status: Formerly smoked
- Residence Type: Rural
- Average Glucose Level: 94

8. Results

In the below table, we can see **Accuracies**, **F1-scores**, **precision, Recall** and **prediction time** for each experiment, using validation dataset. For F1-scores and Accuracies standard deviations (std) from cross validation are reported too.

Method	Model	Accuracy (std)	F1-Score (std)	Precision	Recall	Duration for Prediction (seconds)
With Chi2 and mean value of BMI	GradientBoosting	0.94 (+/- 0.019)	**0.94** (+/- 0.012)	0.94	**0.94**	**0.02**
	RandomForest	0.88 (+/- 0.002)	0.88 (+/- 0.016)	0.9	0.86	0.018
	SVM	0.81 (+/- 0.017)	0.82 (+/- 0.09)	0.77	0.88	0.4
	Naïve Bayes	0.74 (+/- 0.004)	0.75 (+/- 0.006)	0.72	0.8	0.0002
	Logistic Regression	0.78 (+/- 0.016)	0.79 (+/- 0.018)	0.77	0.8	0.0011
Without Chi2 and mean value of BMI	GradientBoosting	0.97 (+/- 0.004)	**0.97** (+/- 0.013)	0.96	**0.98**	**0.011**
	RandomForest	0.94 (+/- 0.008)	0.94 (+/- 0.011)	0.92	0.96	0.15
	SVM	0.81 (+/- 0.018)	0.82 (+/- 0.023)	0.77	0.88	0.4
	Naïve Bayes	0.72 (+/- 0.004)	0.74 (+/- 0.008)	0.7	0.78	0.0012
	Logistic Regression	0.8 (+/- 0.008)	0.8 (+/- 0.011)	0.79	0.81	0.001
With Chi2 and Linear Regression	GradientBoosting	0.94 (+/- 0.015)	**0.94** (+/- 0.016)	0.94	**0.94**	**0.017**
	RandomForest	0.88 (+/- 0.008)	0.87 (+/- 0.012)	0.9	0.84	0.034
	SVM	0.81 (+/- 0.017)	0.82 (+/- 0.018)	0.77	0.88	0.39
	Naïve Bayes	0.75 (+/- 0.002)	0.76 (+/- 0.004)	0.72	0.81	0.000001
	Logistic Regression	0.79 (+/- 0.013)	0.79 (+/- 0.010)	0.77	0.81	0.001
Without Chi2 and Linear Regression	GradientBoosting	0.97 (+/- 0.017)	**0.97** (+/- 0.033)	0.97	**0.97**	**0.04**
	RandomForest	0.94 (+/- 0.011)	0.94 (+/- 0.013)	0.92	0.96	0.095
	SVM	0.87 (+/- 0.025)	0.87 (+/- 0.024)	0.84	0.91	0.32
	Naïve Bayes	0.73 (+/- 0.002)	0.74 (+/- 0.004)	0.7	0.78	0.000001
	Logistic Regression	0.8 (+/- 0.024)	0.81 (+/- 0.021)	0.79	0.83	0.005

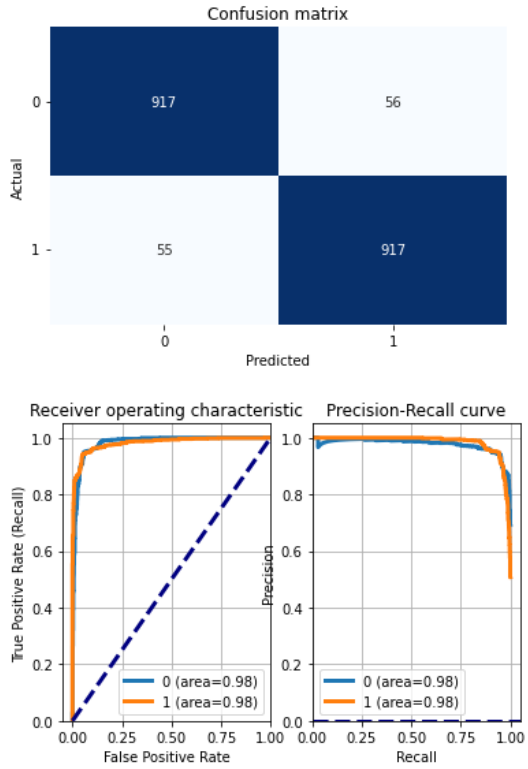
Table-2: Metrics and prediction time for all Experiments and approaches.

As we can see from Table-2, the best model is a Gradient Boosting Classifier, without X^2 test for feature selection and filling the missing values of BMI with its mean value. This model, with this approach achieves **97% F1-score** with prediction time to be too small (~0.02 seconds). Now let's look, if the ROC Curves, Precision-Recall Curves and Confusion matrices, will confirm the above results.

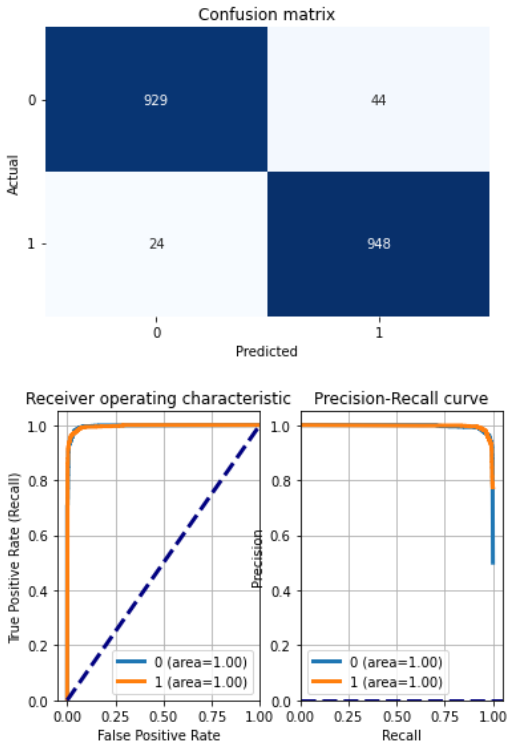
Only the best models will be reported here. Further analysis of results is reported in Jupyter notebook[3].

3 https://nbviewer.org/github/icsd13152/StrokePrediction_EDA_ML/blob/main/StrokePrediction/src/Stroke%20Prediction%20EDA%20and%20ML%20prediction.ipynb

```
For  GB
Duration for prediction:  0.022367238998413086
(1945,)
Accuracy: 0.94
macro F1: 0.94
Recall: 0.94
Precission: 0.94
Detail:
Optimal Threshold value is: 0.5348978680545282
```




```
For  GB
Duration for prediction:  0.011368036270141602
(1945,)
Accuracy: 0.97
macro F1: 0.97
Recall: 0.98
Precission: 0.96
Detail:
Optimal Threshold value is: 0.9998329934895377
```

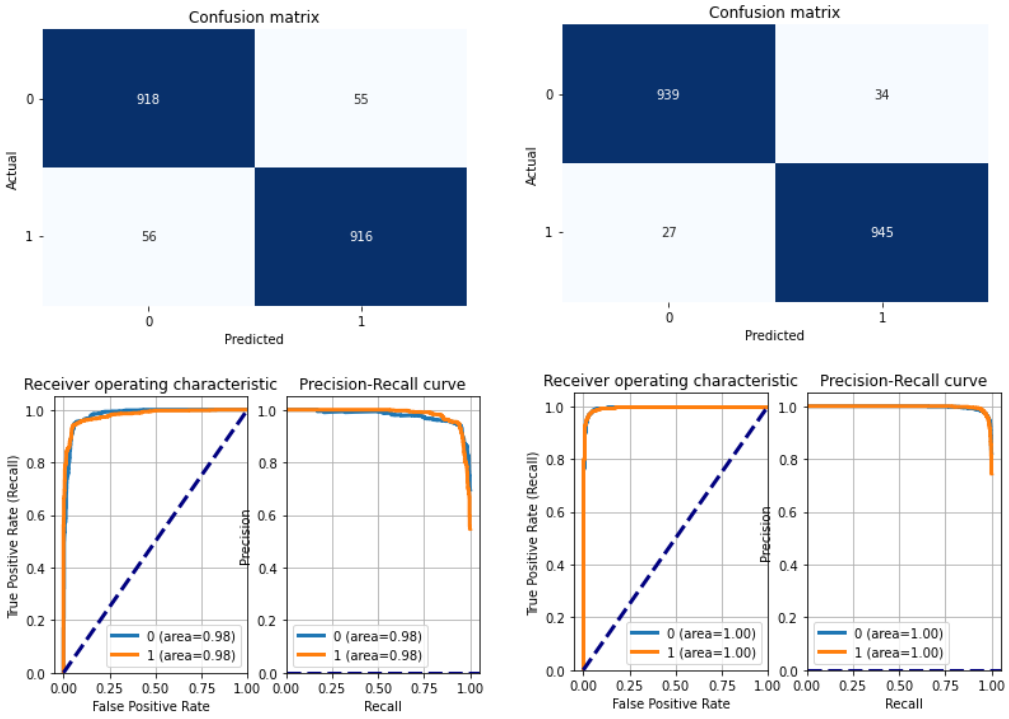


Plot-9: Gradient Boosting with X^2 test and mean value of BMI (left). Gradient Boosting with mean value of BMI and no X^2 test (right).

The AUC scores are up to 98%. That means we have good models that their predictions is too accurate.

```
For  GB
Duration for prediction:  0.017075300216674805
(1945,)
Accuracy: 0.94
macro F1: 0.94
Recall: 0.94
Precission: 0.94
Detail:
Optimal Threshold value is: 0.5051248241488054
```

```
For  GB
Duration for prediction:  0.043999671936035156
(1945,)
Accuracy: 0.97
macro F1: 0.97
Recall: 0.97
Precission: 0.97
Detail:
Optimal Threshold value is: 0.9307101788816952
```

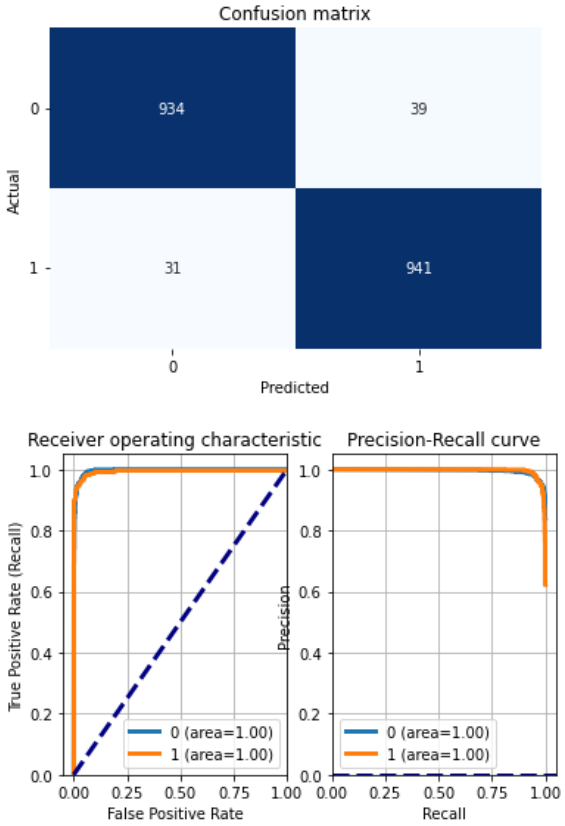


Plot-10: Gradient Boosting with X^2 test and linear Regression (left). Gradient Boosting with linear Regression and no X^2 test (right).

Considering the best approach from the previous experiments, I performed one more experiment with the below results.

Method	Model	Accuracy (std)	F1-Score (std)	Precision	Recall	Duration for Prediction (seconds)
Without Chi2 and mean value of BMI and No scaling	GradientBoosting	0.96 (+/- 0.028)	**0.96** (+/- 0.020)	0.96	0.97	**0.011**
	RandomForest	0.94 (+/- 0.007)	0.95 (+/- 0.016)	0.93	0.97	0.133
	SVM	0.96 (+/- 0.007)	0.96 (+/- 0.015)	0.95	0.96	1.087
	Naïve Bayes	0.73 (+/- 0.025)	0.75 (+/- 0.025)	0.71	0.79	0.001
	Logistic Regression	0.8 (+/- 0.019)	0.8 (+/- 0.018)	0.79	0.81	0.0009

Table-3: Metrics and prediction time for the final experiment.



Plot-11: Gradient Boosting with mean value of BMI and no X^2 test and no data scaling.

Again, the best classifier is a Gradient Boosting model. Considering not only the performance in F1-score but the performance in prediction time.

9. Conclusion

In this project different experiments and approaches have been performed, in order to solve the problem of Stroke prediction and explainability of the results. The best results came up from the below approach:

- Without feature selection using X^2 statistical test.
- Deal with missing values of BMI, by replacing them with the mean value of this feature.
- Gradient Boosting Classifier.

10. Future Work

Apart from this approach and method, reported in this paper, it would be challenging as future work to perform some experiments with simple Neural Networks even though with Deep learning techniques. In addition, someone can keep those approaches and perform oversampling with another way or collect more data from a survey or sensors. Finally, it would be meaningful if another method for feature selection would be performed.

References

[1]. José Alberto Tavares Rodríguez " Stroke prediction through Data Science and Machine Learning Algorithms" June 2021.

[2]. Harshitha K V, Harshitha P, Gunjan Gupta, Vaishak P, Prajna K B "Stroke Prediction Using Machine Learning Algorithms" International Journal of Innovative Research in Engineering & Management (IJIREM), July 2021.

[3]. Sailasya, Gangavarapu, and Gorli L. Aruna Kumari. "Analyzing the performance of stroke prediction using ML classification algorithms." *Int. J. Adv. Comput. Sci. Appl* 12.6 (2021): 539-545.

[4]. Tasfia Ismail Shoily, Tajul Islam, Sumaiya Jannat and Sharmin Akter Tanna "Detection of stroke using machine learning algorithms", 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT), IEEE, July 2019.

[5].https://nbviewer.org/github/icsd13152/StrokePrediction_EDA_ML/blob/main/StrokePrediction/src/Stroke%20Prediction%20EDA%20and%20ML%20prediction.ipynb

[6]. https://shap.readthedocs.io/en/latest/index.html

[7]. Lundberg, Scott M., et al. "From local explanations to global understanding with explainable AI for trees." *Nature machine intelligence* 2.1 (2020): 56-67.

[8]. Lundberg, Scott, et al. "Explainable machine-learning predictions for the prevention of hypoxaemia during surgery, Nature Biomedical Engineering, v. 2, no. 12." (2018): 749-760.