**CS 677 Data science with Python**

**Project Report**

**Stroke Data Classification**

**By**

**Akshay Ankush Kapare**

[**Akshay08@bu.edu**](mailto:Akshay08@bu.edu)

**BU ID :U50485521**

**Introduction**

Project is Stroke data classification. In this project I have taken health care stroke data set and using different kind of classifiers I am trying to predict that Patient will have stroke or not .To predict the data I am using different type of classifiers such as Logistic Regression , Random Forest , Support vector machine(SVM), Naive-Bayes , Decision Tree etc. with this I am printing classification report like precision ,recall ,F1 score . With all of this I am also presenting graph of some attributes with respect to stroke.

**Dataset:**

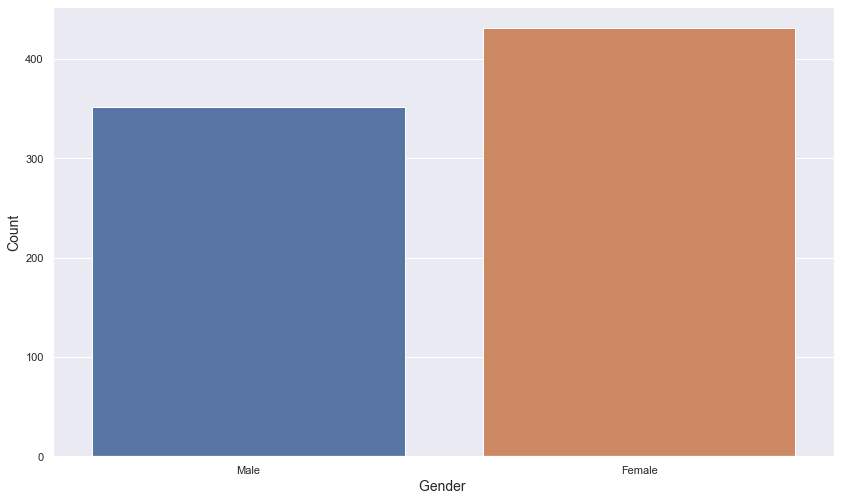
The dataset is taken form Kaggle:

<https://www.kaggle.com/asaumya/healthcare-dataset-stroke-data#train_2v.csv>

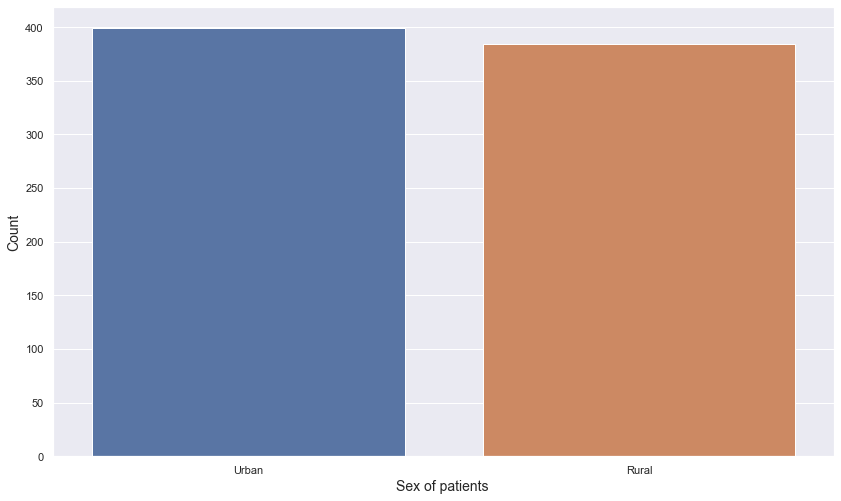
The dataset contains approximately 44,000 rows, and 12 columns. Each row contains data for a single patient. The columns include attributes for patients such as their age, gender, BMI, smoking status, etc, with which we can predict the ﬁnal attribute: “stroke”, which predicts if the patient might suﬀer from a stroke or no.

I have created plots to visualize the all the attributes with stroke data(1 : Patient has stroke; 0: Patient does not have stroke).I have use Seaborn and Matplot library to show the result.

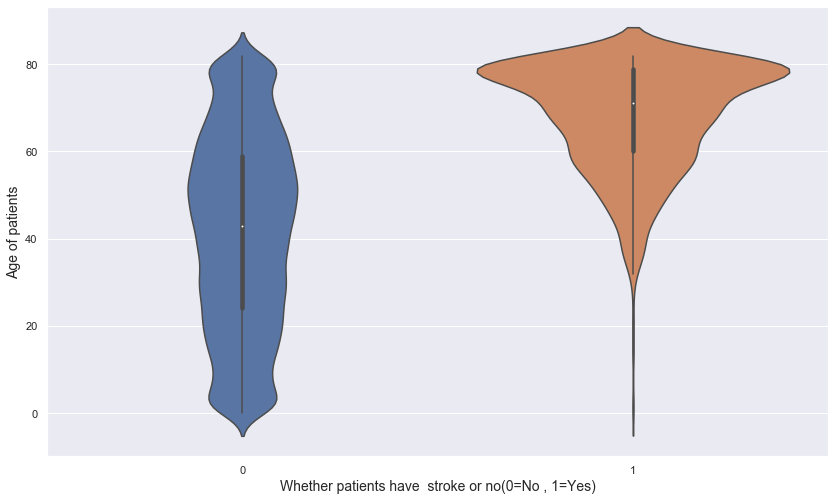
1.Count of Gender w.r.t patient has stroke:



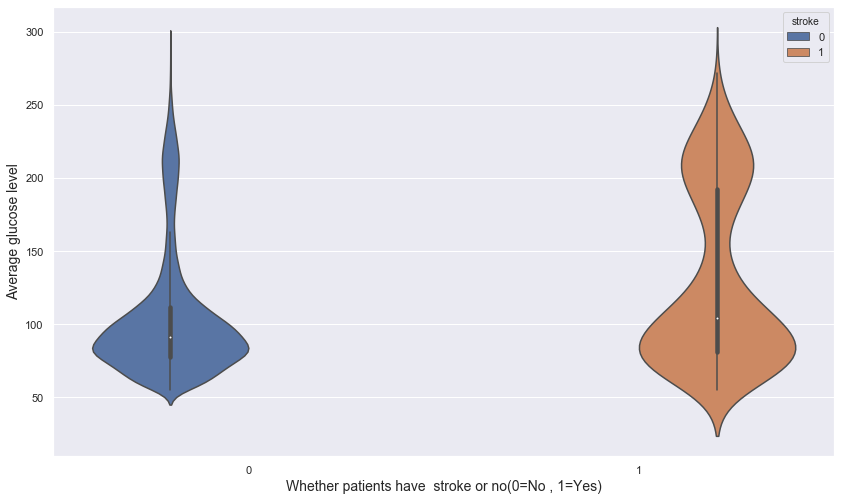
2. Count of Residence type w.r.t patient has stroke:



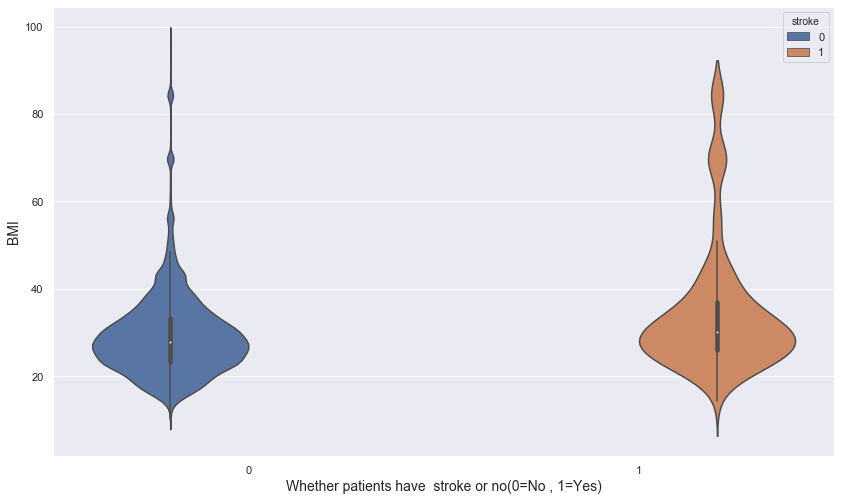
3. Age w.r.t Stroke (0: Patient does not have stroke 1: patient had stroke):



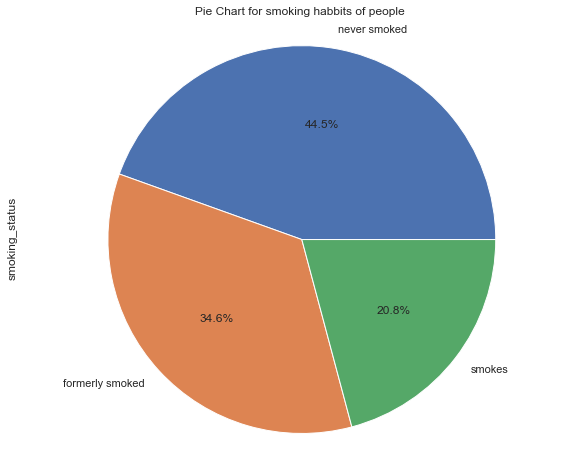
4. Average glucose level w.r.t Stroke (0: Patient does not have stroke 1: patient had stroke):



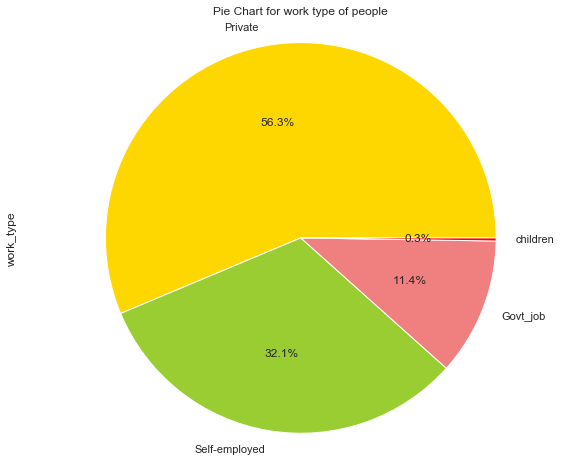
5. BMI w.r.t Stroke(0:Patient does not have stroke 1:patient had stroke):



6. Number for people having stroke with smoking status:



7.Number for people having stroke with their work type:



**Classification:**

I have build some classification models on the dataset considering stroke as a target variable and Gender, Age, Hypertension, Heart disease, Average Glucose level, BMI etc. as labels.

Classifiers are as follows:

1. **Naive-Bayes**
2. **Logistic Regression**
3. **AdaBoost**
4. **Random Forest**
5. **Decision Tree**
6. **Support Vector Machine(SVM)**

**Accuracy I have obtained from each classifier are as follows:**

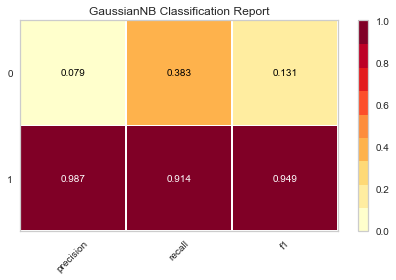
|  |  |
| --- | --- |
| **Classifiers** | **Accuracy** |
| Naive-Bayes | 90.3514% |
| Logistic Regression | 98.1048% |
| AdaBoost | 98.0588% |
| Random Forest | 98.0991% |
| Decision Tree | 96.3191% |
| Support Vector Machine(SVM) | 98.1048% |

**Evaluation of classifiers:**

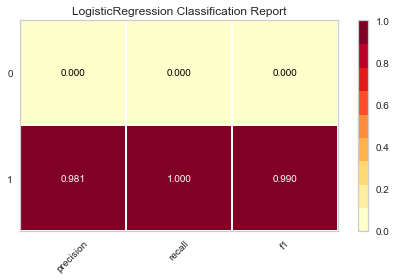
Using Yellowbrick python library I have evaluated every classifier and generated graph that shows precision ,f1 score and recall.

Following are the plots for every classifier:

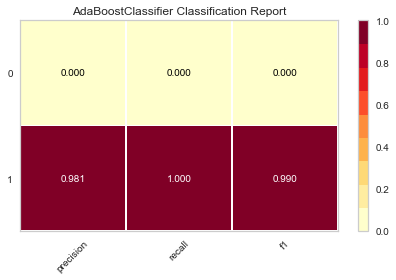
1. Naive-Bayes

1

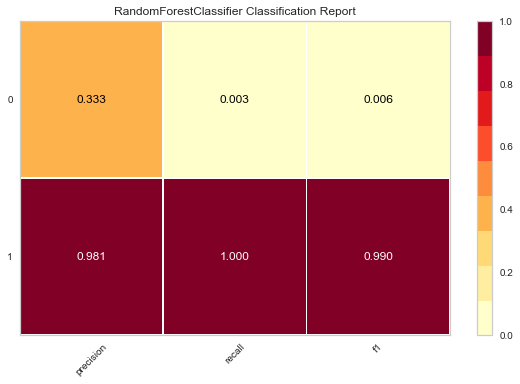
2. Logistic Regression



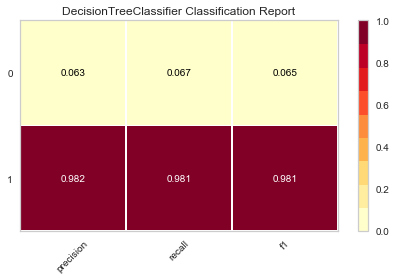
3. Adaboost



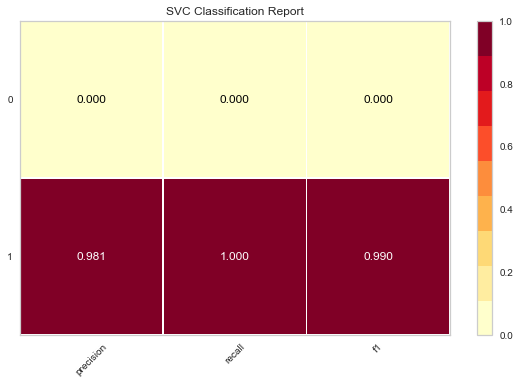
4 Random Forest



5. Decision Tree



6.Support Vector Machine



**Conclusion:**

Logistic regression and Support Vector Machine classifier works best with accuracy of 98.1048%

**References:**

[1] https://towardsdatascience.com/implementing-binary-logistic-regression-in-r-7d802a9d98fe

[2] Hastie TJ, Tibshirani RJ, Friedman JH. The Elements of Statistical Learning: Data Mining Inference and Prediction. Second Edition. Springer; 2009. ISBN 978-0-387-84857-0

[3] <https://scikit-learn.org/stable/modules/neural_networks_supervised.html>

[4] <https://gaming.stackexchange.com/questions/167318/what-do-fifa-14-position-acronyms-mean>