Final Project Report
Introduction to Data Analytics

Project Title:
Data Analysis on Stroke Predictions

Prepared by:
Mustafa Johar Udegadhwala (N01414702)

ITE 5201 – Winter 2022
Humber College

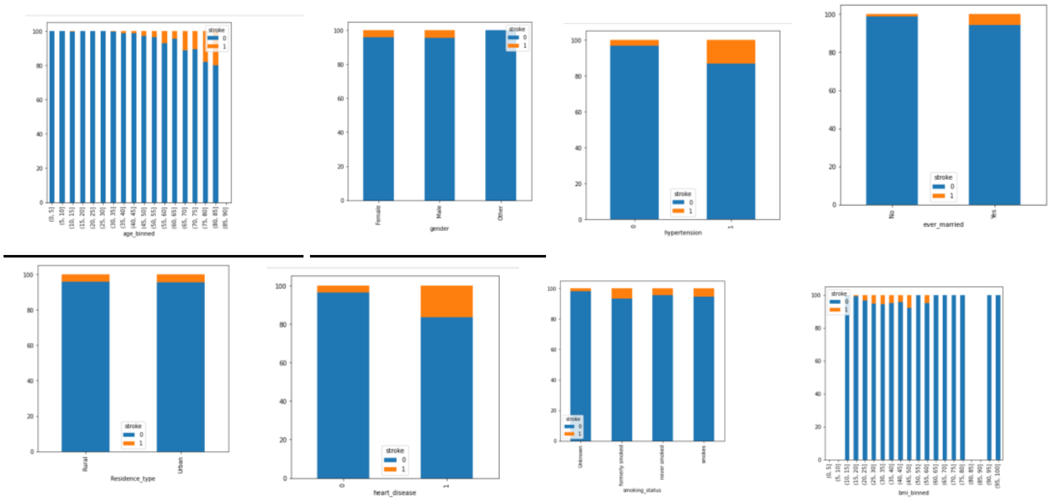
1. Problem Statement

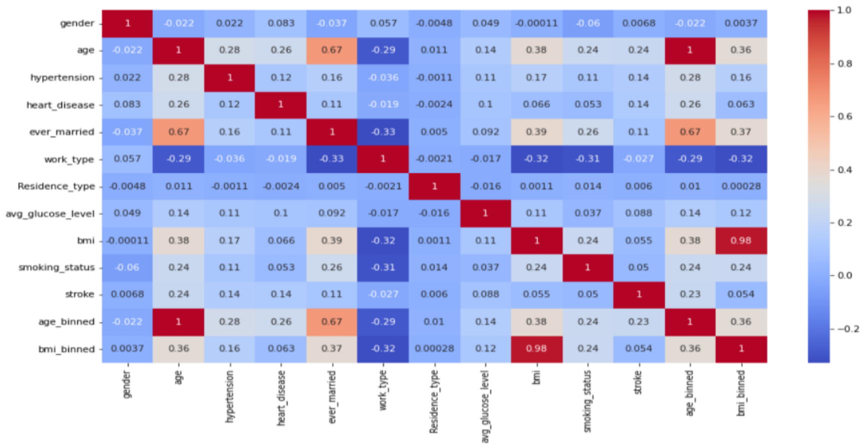
- **This dataset is used to predict whether a person can suffer from a stroke based on the input parameters like gender, age, marital status, work type, residence type , different health Issues, and smoking status.**

2. Dataset Description

The dataset consisted of 5110 records with 12 variables which includes people's demographic data (gender, age, marital status, type of work and residence type) and health records (hypertension, heart disease, average glucose, Body Mass Index (BMI), smoking status and experience of stroke).

3. Dataset Analysis and Observations





Observations:

a. Old people are more likely to suffer from a stroke compared to young people.
b. Gender Does not show much variation, Both Male and Female have equal chances of having stroke.
c. People with HyperTension has higher chances of stroke

d. From the analysis it can be said married people suffer from stroke more, but again that is a fact that married people are older in age compared to unmarried so we cannot infer much from it.
e. People from Rural areas as well as urban areas both suffer from strokes equally
f. People with BMI 45-50 suffer more from stroke
g. Even the Non Smokers can get a stroke so Smoking does not make any difference in chances of getting a stroke.
h. We could only see correlation between variables age and ever_married.

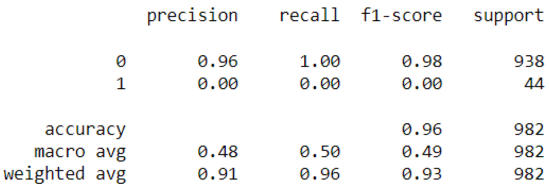
4. Proposed Analytical/Prediction Model

First of all I removed the Id column as it was not necessary for prediction and then I preprocessed the data and converted the categorical data to numbers. Then I did the Bivariate analysis. In the dataset Variable are ordinal; numeric, and categorical. Variables are related nonlinearly . Data is non-normally distributed so to find the co-relation i used Spearman's Rank Correlation. and after that i did the Prediction.

As a Prediction Model I have Used Logistic Regression Classification Model because in the selected data set most values are binary so Classification model is the best one to be used as Classification algorithms are used to predict/Classify the discrete values. And then there were only two classes so I felt it is best to use Logistic regression Classification Model.

5. Results and Discussions

Classification Report:

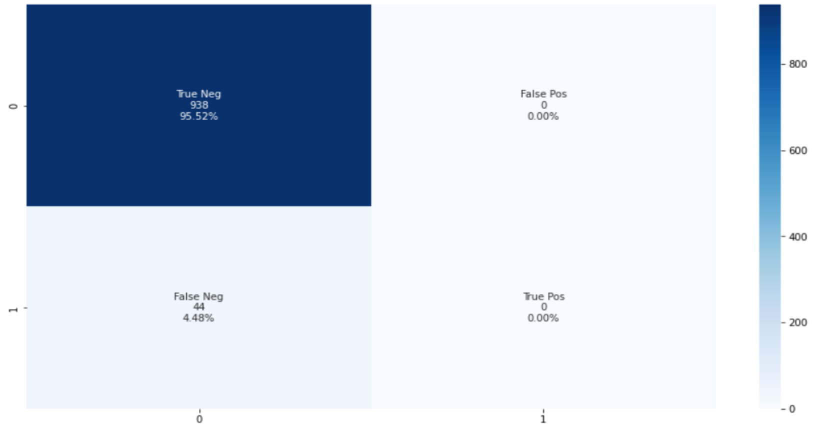


```
              precision    recall  f1-score   support

           0       0.96      1.00      0.98       938
           1       0.00      0.00      0.00        44

    accuracy                           0.96       982
   macro avg       0.48      0.50      0.49       982
weighted avg       0.91      0.96      0.93       982
```

Confusion Matrix:



The Model has achieved 96% accuracy and 0.98 f1-score.