Stroke Prediction

Prepared by:
Jefn Alshammari
Abdulaziz Almass

Overview

1. Objective
2. Data Cleaning
3. Exploratory Data Analysis
4. ML Classification Models
5. Models Evaluation
6. Conclusion

Objective

According to the World Health Organization (WHO),**15 million people worldwide suffer** a stroke annually.

5 million die and another 5 million are left permanently disabled. By predicting stroke, it can help healthcare organizations and practitioners to save millions of lives and prevent disability.

Subsequently, this project aimed to predict stroke.

Dataset

- 5110 observations

- 12 Features

From kaggle

Data Cleaning

Handle Missing

Replaced all values containing Nan with mean

1

2

Feature Dropping

Dropped all rows with irrational values

One Hot Encoding

Encoded categorical features as one-hot numeric array

3

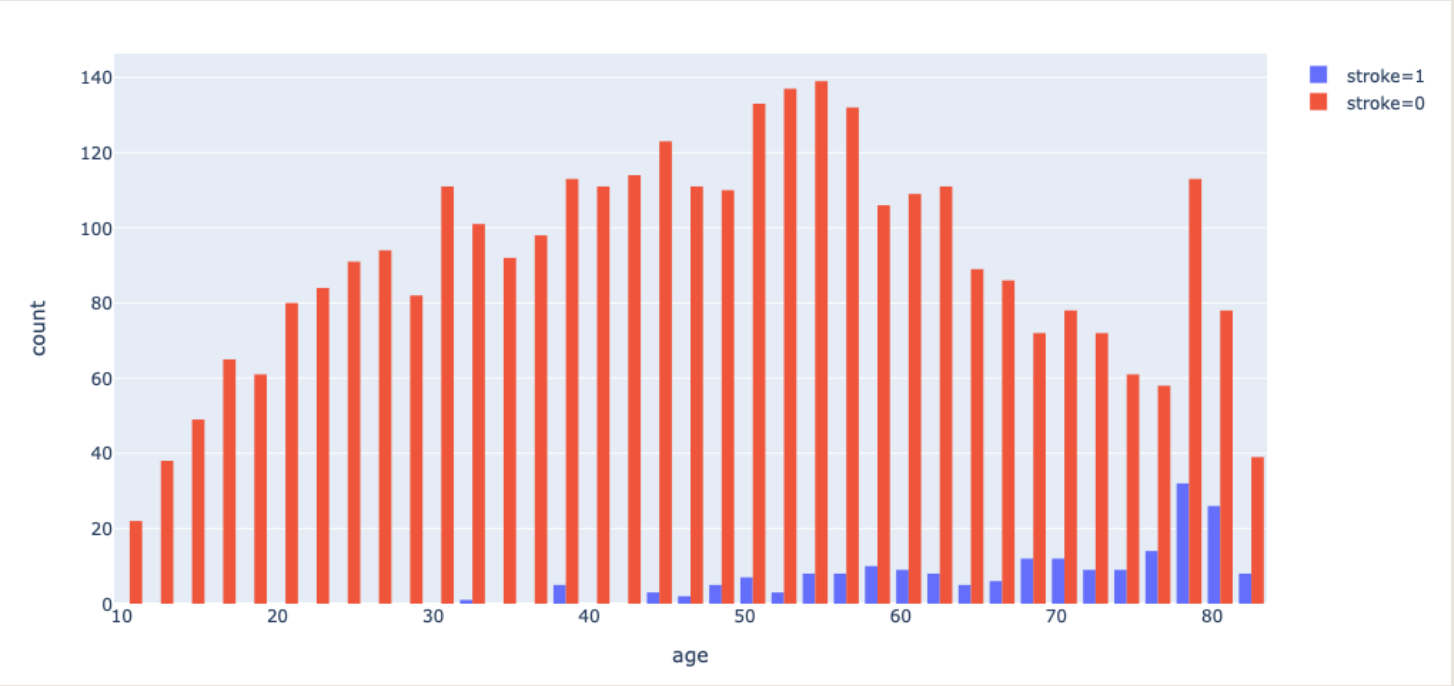
4

Concatenation

Inserted all new columns to dataset

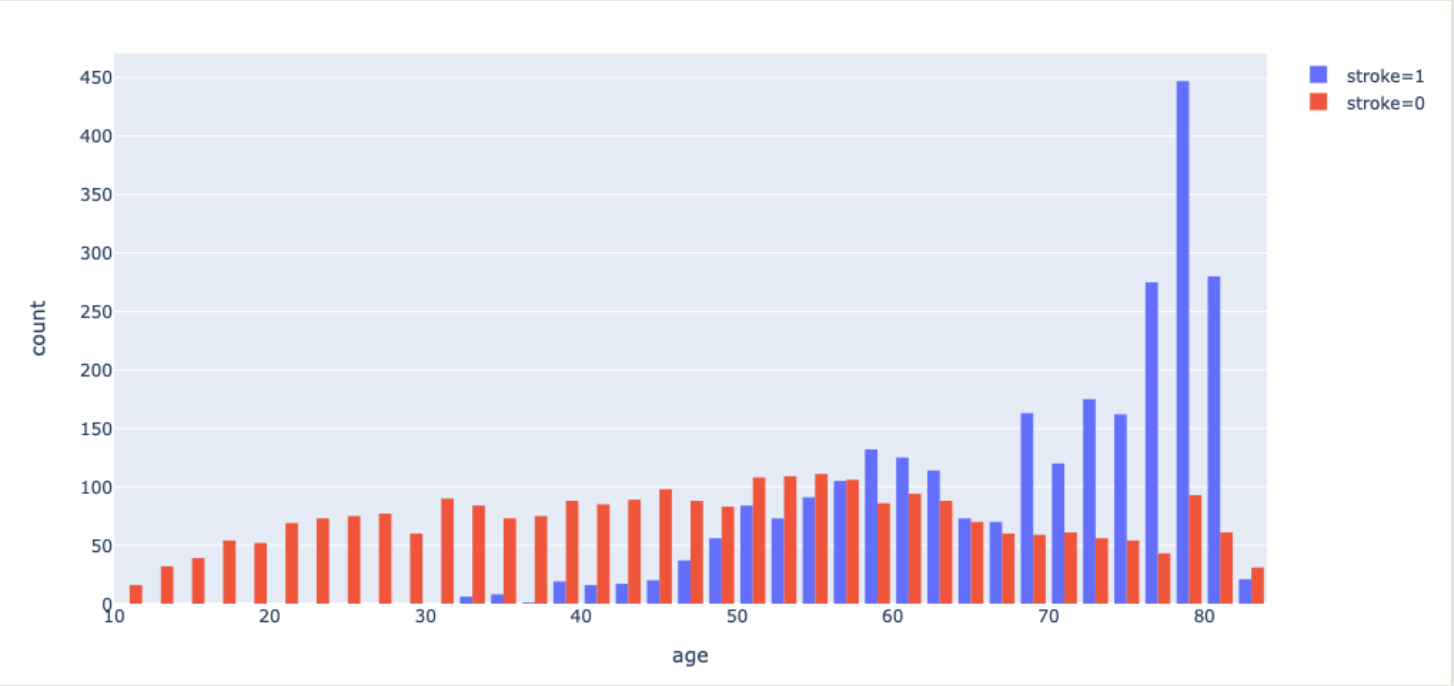
EDA

By using charts, imbalanced data has been analyzed for any important feature. e.g “age”



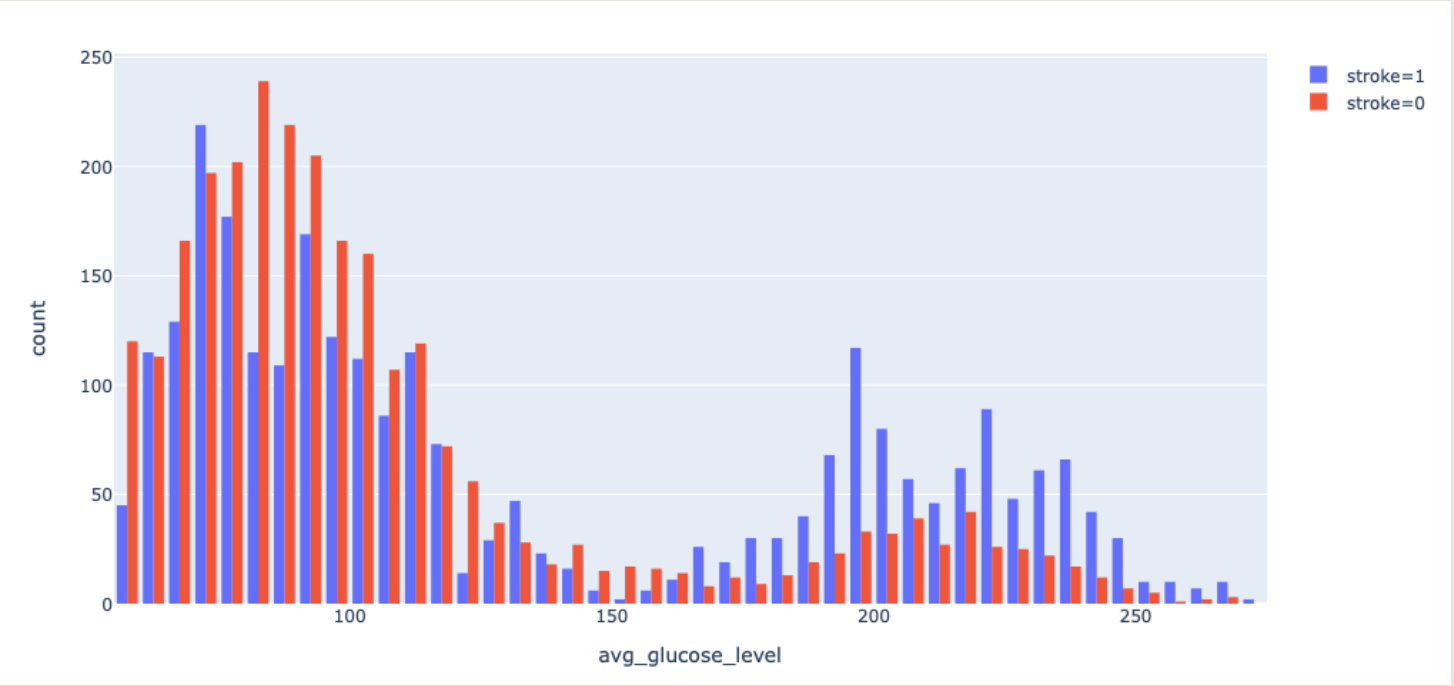
EDA with Balanced Dataset

The age is significantly noticeable as an important feature

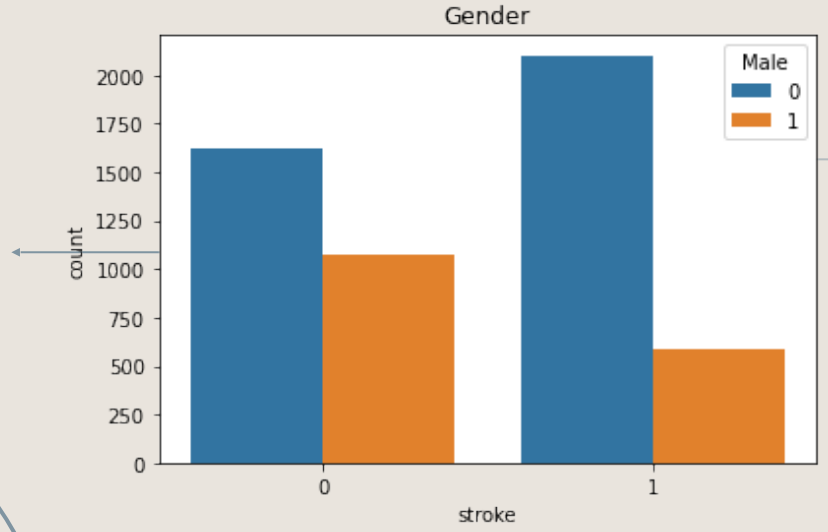


EDA with Balanced Dataset

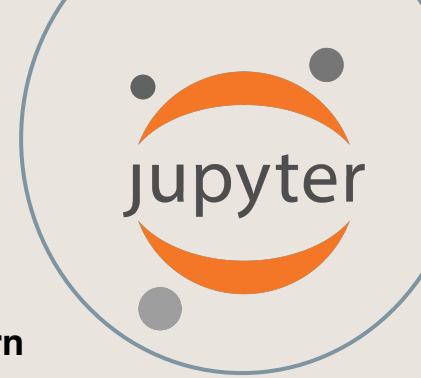
Stroke has been noticed in avg_glucose_level is getting higher from 150 onwards



EDA with Balanced Dataset



Tools



1. Numpy
2. Pandas
3. Seaborn
4. Matplotlib
5. Plotly
6. Sklearn
7. Imblearn (SMOTE)
8. XGBoost

ML Classification Models

Classification Algorithms

- Logistic Regression
- K-Nearest Neighbor (KNN)
- Decision Tree
- Support Vector Machine (SVM)
- Random Forest
- XGBoost

Models Evaluation

		Macro Avg	Accuracy	Precision	Recall	F1 Score	Stroke
Logistic Regression	Imbalanced	0.50	0.92	0.93 0.00	1.00 0.00	0.96 0.00	0 1
Logistic Regression	Not Scaled	0.91	0.90	0.89 0.92	0.93 0.89	0.91 0.90	0 1
	Scaled	0.92	0.91	0.88 0.97	0.97 0.87	0.92 0.92	0 1
	Tuned & Scaled	0.92	0.91	0.89 0.94	0.94 0.89	0.92 0.91	0 1
KNN	Scaled	0.94	0.94	0.96 0.93	0.93 0.96	0.94 0.94	0 1
Decision Tree	Scaled	0.92	0.92	0.94 0.91	0.90 0.95	0.92 0.93	0 1
SVM	Not Scaled	0.92	0.92	0.89 0.96	0.97 0.88	0.92 0.92	0 1
	Scaled	0.92	0.92	0.88 0.97	0.97 0.87	0.92 0.92	0 1
Random Forest	Scaled	0.97	0.96	0.96 0.97	0.97 0.96	0.97 0.97	0 1
XGBoost	Scaled	0.93	0.93	0.91 0.95	0.95 0.91	0.93 0.93	0 1

Conclusion

From model predictions, it was noticed that the most important feature in stroke prediction dataset is age. As the best model to be considered for deployment is Random Forest due to the highest scores amongst other classification models.

Questions & Discussion