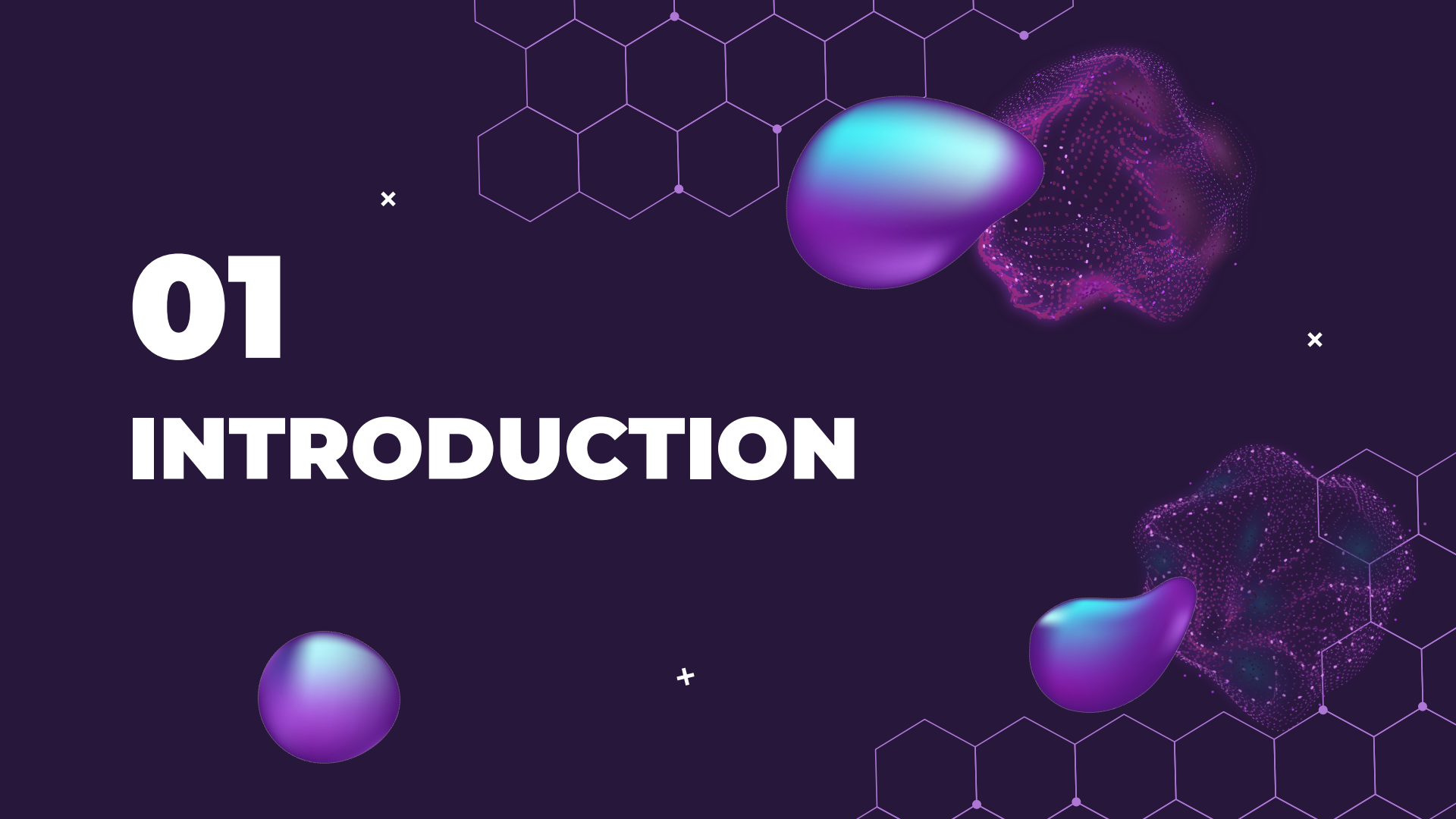
STROKE PREDICTION

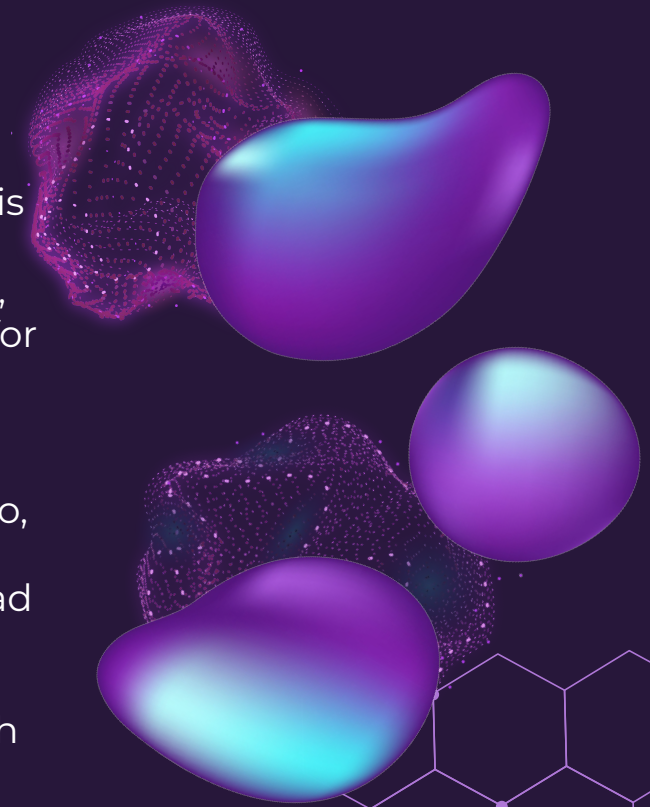
COMP 3608 - Intelligent Systems Project

Group: TECH TITANS

Members: Christin Brown • Shania Gajadhar • Dakota Sharma


01
INTRODUCTION

INTRODUCTION

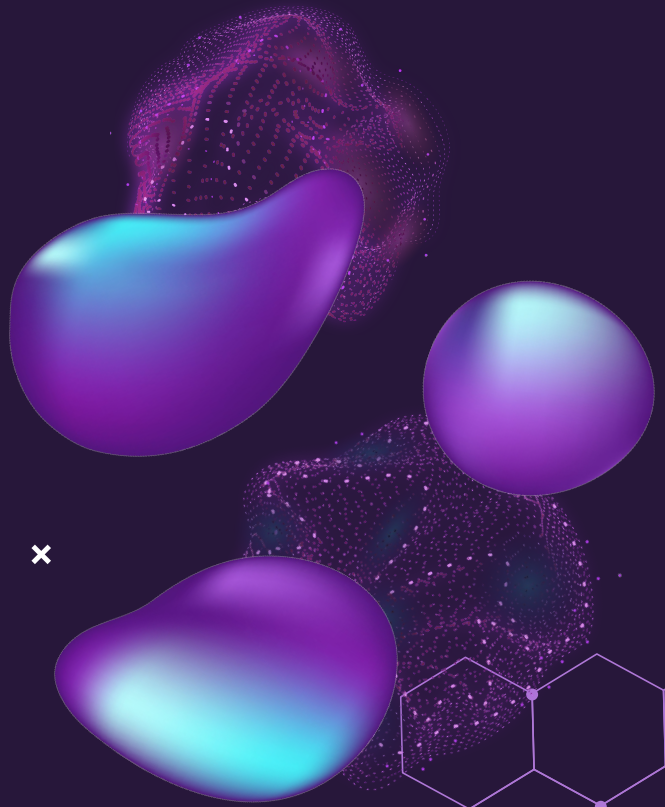


A stroke can cause death in minutes. It occurs because blood is blocked from flowing to the brain or the brain is suddenly bleeding. According to The World Health Organization (WHO), stroke is the 2nd leading cause of death globally, responsible for approximately 11% of total deaths. Strokes are also the leading cause of long-term disability.

in 2020, strokes accounted for 822 deaths in Trinidad in Tobago, or 9.56% of total deaths for the year. In 2019, strokes were recorded as the third highest leading cause of death in Trinidad and Tobago.

The prediction of strokes in advance will help reduce the death rate. In recent years, predicting stroke in the real-life medical area has not been easy. A dataset of stroke data was used for analysis.

INTRODUCTION



This project proposes predicting stroke using three (3) Machine Learning algorithms such as Logistics Regression, Random Forest Classifier and a Forward Feed Neural Network to predict whether a person is having a stroke or not.

After comparing the performance of each algorithm, we can find the best algorithm that can predict a stroke with the highest accuracy.

The main goal of this project is to determine a predictive model that is best suited for stroke prediction and prevention.

02
Dataset

Dataset

- Stroke Dataset
- Sourced from **Kaggle**
- Contains 5110 observations with 12 attributes.

Dataset Attributes

- **Id** - unique integer value
- **Gender** - male/female/other
- **Age** - age of the individual in years
- **Hypertension** - binary feature
- **Heart Disease** - binary feature
- **Marriage Status** - binary feature showing if ever married
- **Work Type** - private/self-employed/government/children/never
- **Residence Type** - urban or rural
- **Average Glucose Level** - avg. glucose level in blood
- **BMI** - body mass index
- **Smoking Status** - former/never/smoker/unknown
- **Stroke** - stroke event

Dataset Preprocessing

- Missing values in the BMI column were filled in by imputation
- The mean of all BMI values were calculated and given to all blank entries
- The data was then split into ‘features’ and ‘target’, the target being ‘stroke’.
- Categorical features were converted into numericals.
- Numerical features were normalized.
- The dataset was also balanced.
- Hyperparameter tuning was carried out.
- The ‘id’ attribute was dropped because it is a unique value for each entry and has no bearing on the likelihood of any other attribute.

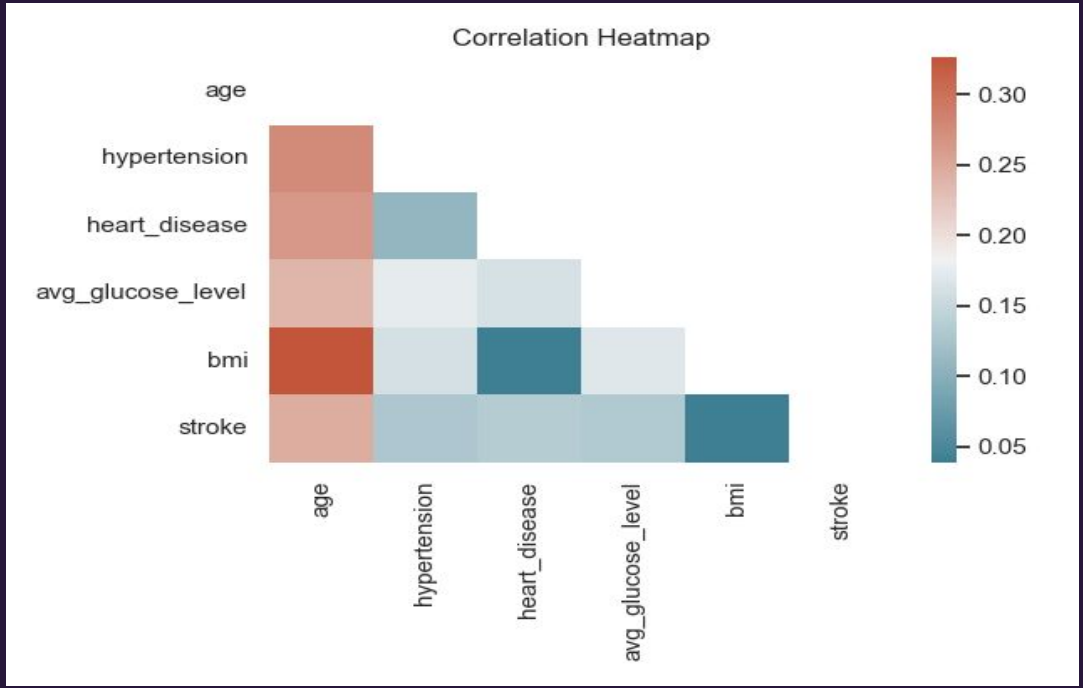
Dataset

id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
9046	Male	67	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
51676	Female	61	0	0	Yes	Self-emplo	Rural	202.21	N/A	never smoked	1
31112	Male	80	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1
60182	Female	49	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
1665	Female	79	1	0	Yes	Self-emplo	Rural	174.12	24	never smoked	1
56669	Male	81	0	0	Yes	Private	Urban	186.21	29	formerly smoked	1
53882	Male	74	1	1	Yes	Private	Rural	70.09	27.4	never smoked	1
10434	Female	69	0	0	No	Private	Urban	94.39	22.8	never smoked	1
27419	Female	59	0	0	Yes	Private	Rural	76.15	N/A	Unknown	1
60491	Female	78	0	0	Yes	Private	Urban	58.57	24.2	Unknown	1
12109	Female	81	1	0	Yes	Private	Rural	80.43	29.7	never smoked	1
12095	Female	61	0	1	Yes	Govt_job	Rural	120.46	36.8	smokes	1
12175	Female	54	0	0	Yes	Private	Urban	104.51	27.3	smokes	1
8213	Male	78	0	1	Yes	Private	Urban	219.84	N/A	Unknown	1
5317	Female	79	0	1	Yes	Private	Urban	214.09	28.2	never smoked	1
58202	Female	50	1	0	Yes	Self-emplo	Rural	167.41	30.9	never smoked	1
56112	Male	64	0	1	Yes	Private	Urban	191.61	37.5	smokes	1
34120	Male	75	1	0	Yes	Private	Urban	221.29	25.8	smokes	1
27458	Female	60	0	0	No	Private	Urban	89.22	37.8	never smoked	1
25226	Male	57	0	1	No	Govt_job	Urban	217.08	N/A	Unknown	1
70630	Female	71	0	0	Yes	Govt_job	Rural	193.94	22.4	smokes	1
13861	Female	52	1	0	Yes	Self-emplo	Urban	233.29	48.9	never smoked	1
68794	Female	79	0	0	Yes	Self-emplo	Urban	228.7	26.6	never smoked	1
64778	Male	82	0	1	Yes	Private	Rural	208.3	32.5	Unknown	1
4219	Male	71	0	0	Yes	Private	Urban	102.87	27.2	formerly smoked	1
70822	Male	80	0	0	Yes	Self-emplo	Rural	104.12	23.5	never smoked	1
38047	Female	65	0	0	Yes	Private	Rural	100.98	28.2	formerly smoked	1

03

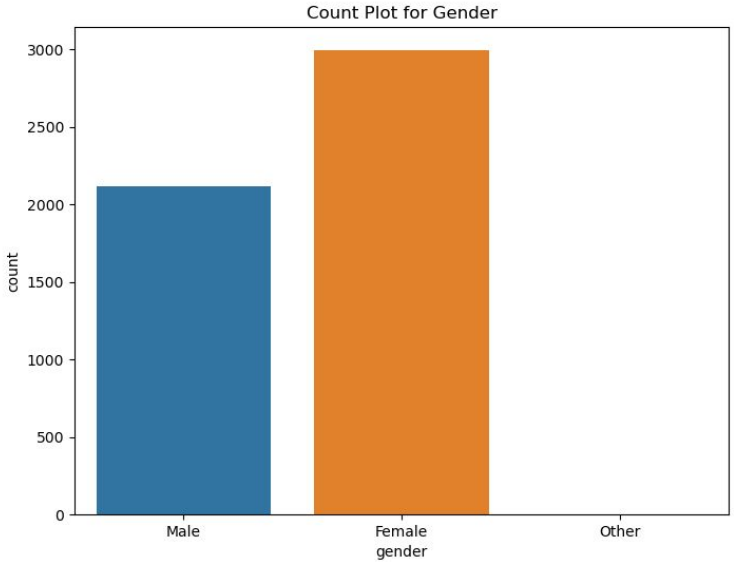
Exploratory Data Analysis (EDA)

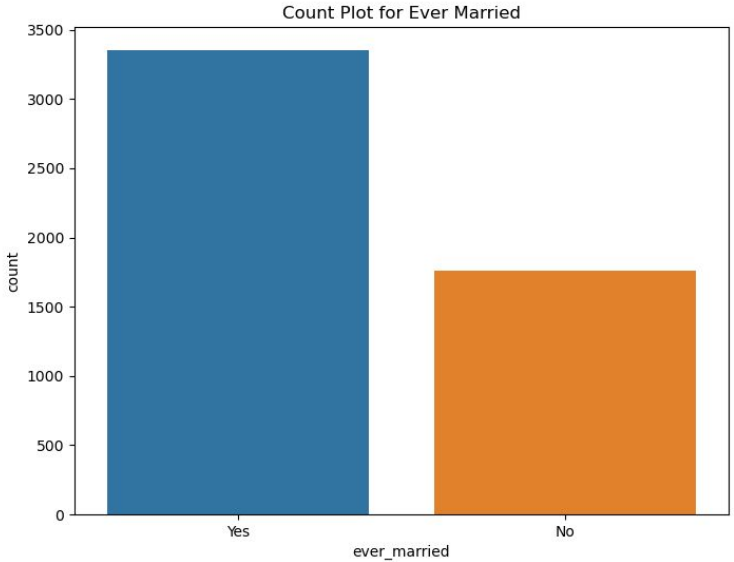
EDA



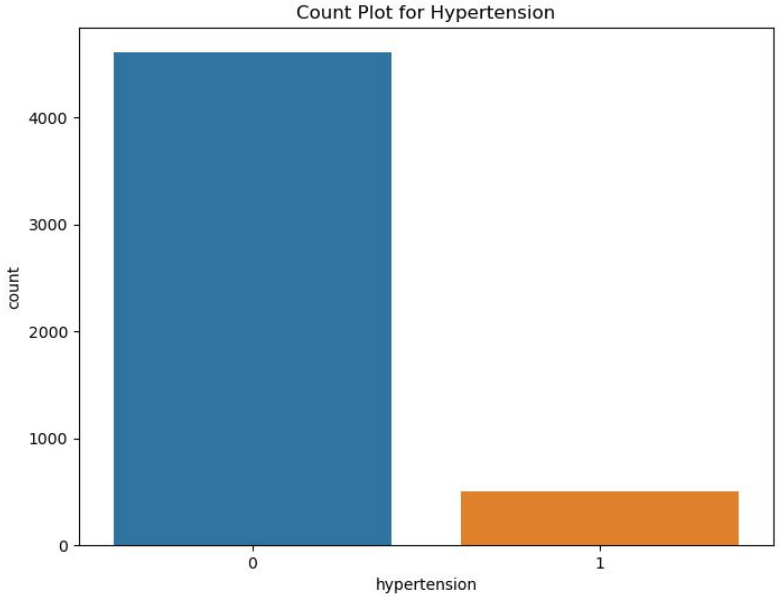
- This diagram demonstrates that age has the strongest positive correlation with stroke. This means that a person's age is more likely to influence weather or not they have a stroke.
- In contrast, BMI has the weakest correlation with stroke.

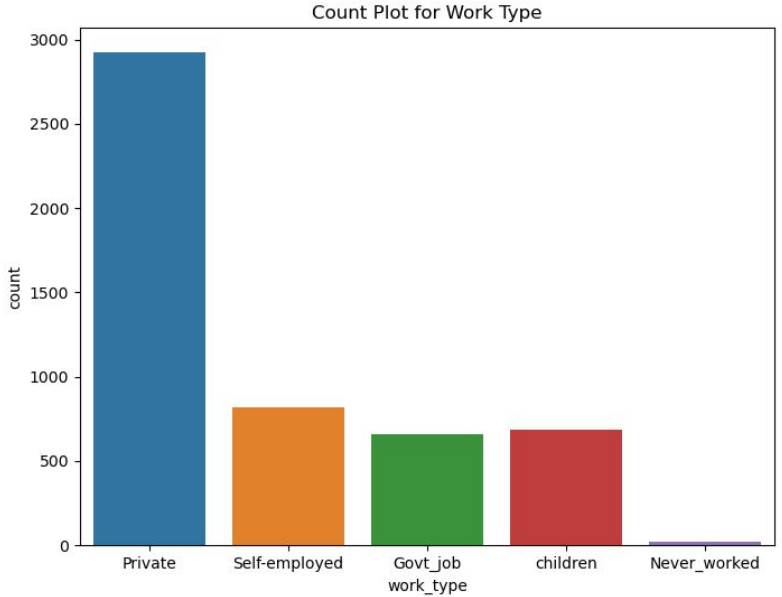
EDA


Count Plot for Gender
count
3000
2500
2000
1500
1000
500
0
Male
Female
Other
gender

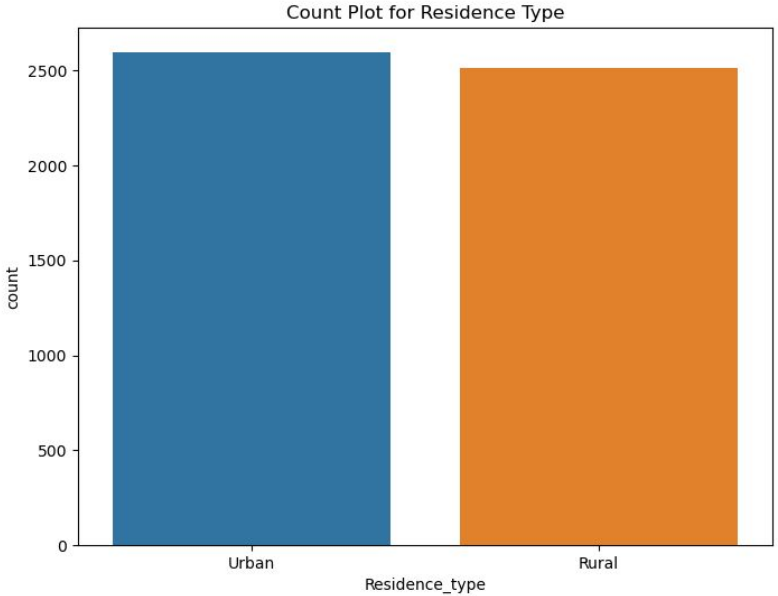

Count Plot for Ever Married
count
3500
3000
2500
2000
1500
1000
500
0
Yes
No
ever_married

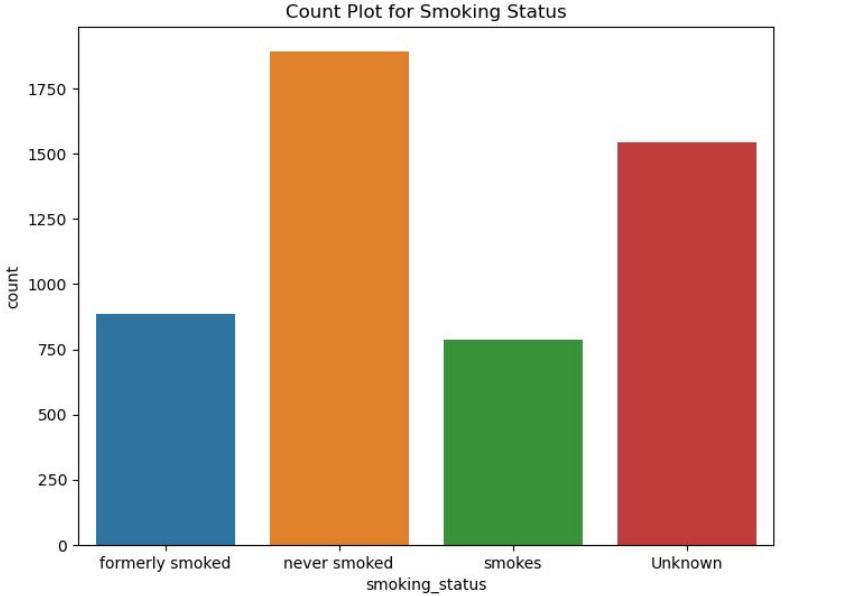
EDA


Count Plot for Hypertension
count
4000
3000
2000
1000
0
0
1
hypertension

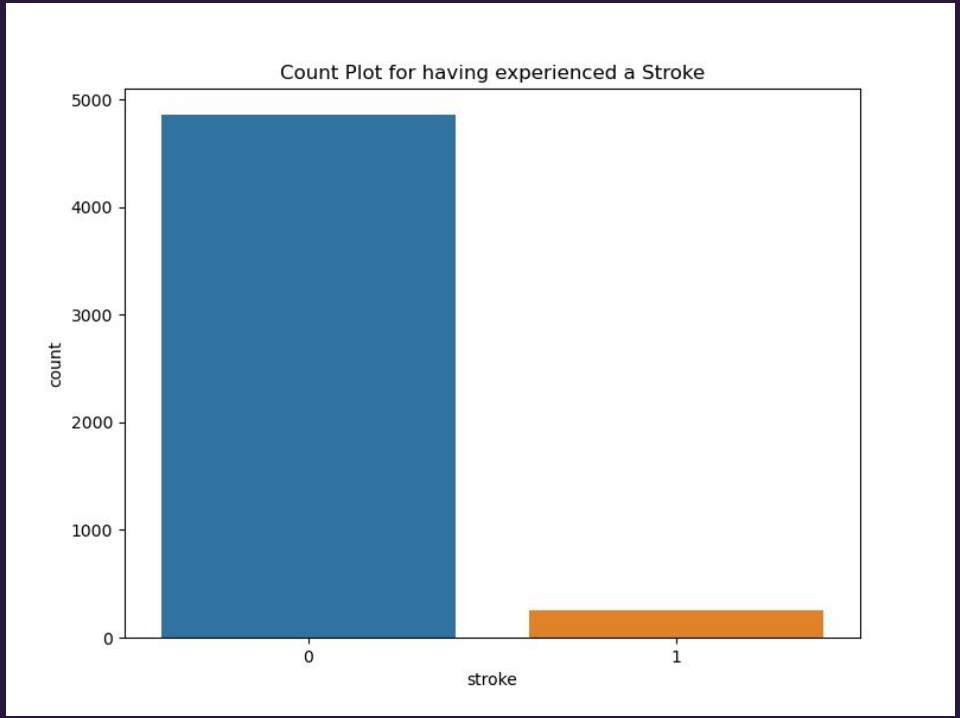

Count Plot for Work Type
count
3000
2500
2000
1500
1000
500
0
Private
Self-employed
Govt_job
children
Never_worked
work_type

EDA


Count Plot for Residence Type
count
2500
2000
1500
1000
500
0
Urban
Rural
Residence_type

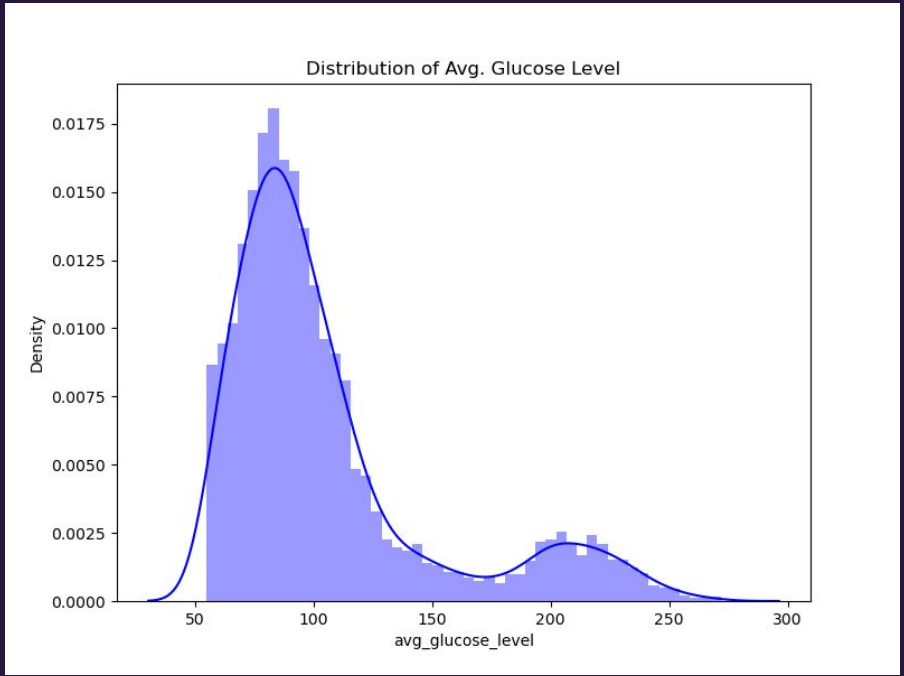

Count Plot for Smoking Status
count
1750
1500
1250
1000
750
500
250
0
formerly smoked
never smoked
smokes
Unknown
smoking_status

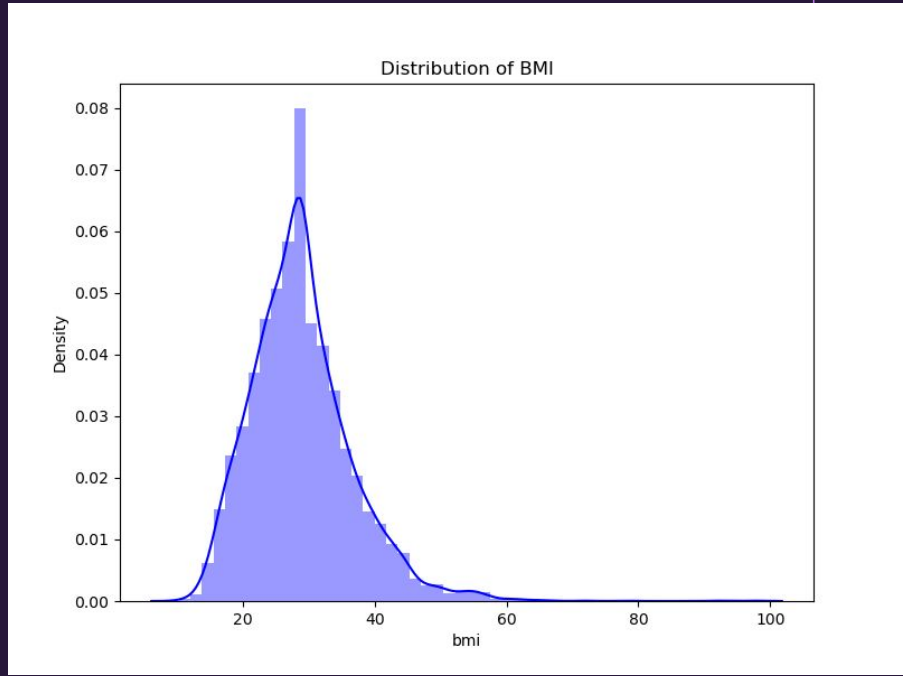
EDA



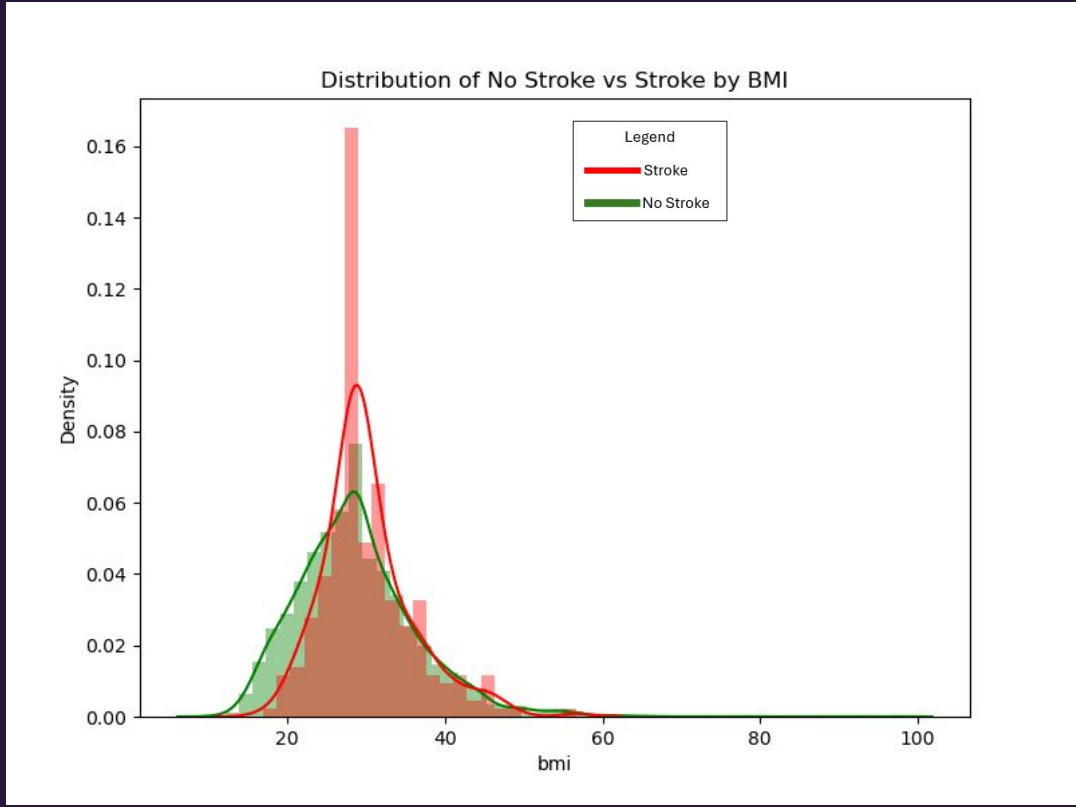
This indicates that the dataset is imbalanced as there are a lot less people have experienced a stroke.

EDA

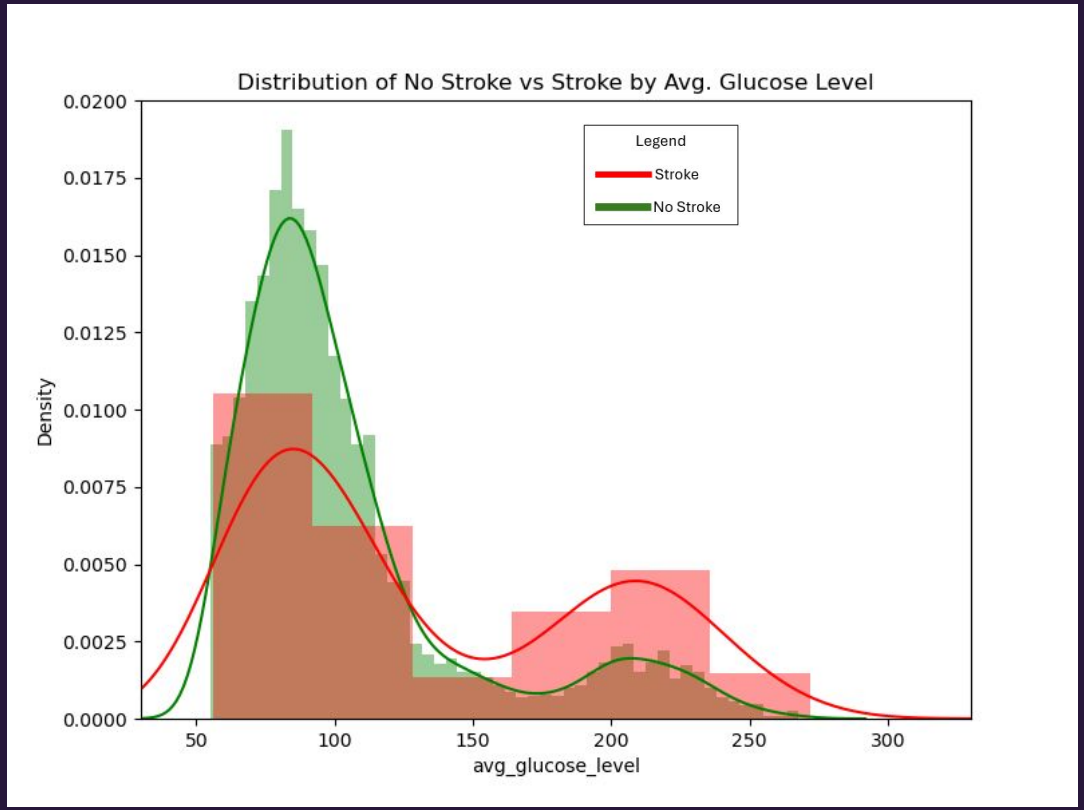

Distribution of Avg. Glucose Level
Density
0.0175
0.0150
0.0125
0.0100
0.0075
0.0050
0.0025
0.0000
50
100
150
200
250
300
avg_glucose_level


Distribution of BMI
Density
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
20
40
60
80
100
bmi

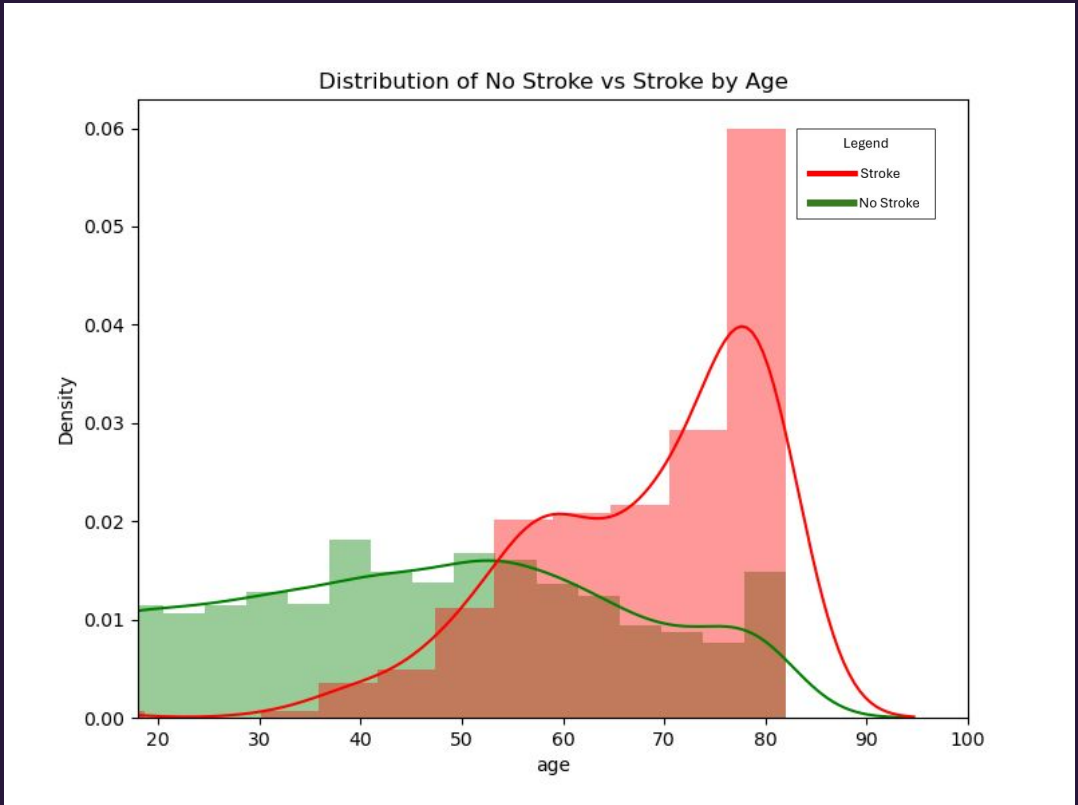
EDA


Distribution of No Stroke vs Stroke by BMI
Legend
Stroke
No Stroke
Density
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
20
40
60
80
100
bmi

EDA


Distribution of No Stroke vs Stroke by Avg. Glucose Level
Legend
Stroke
No Stroke
Density
0.0200
0.0175
0.0150
0.0125
0.0100
0.0075
0.0050
0.0025
0.0000
50
100
150
200
250
300
avg_glucose_level

EDA



- People over the age of 55 are likelier to have a stroke

04

Model Building

Model Selection

The three algorithms selected for the project were:

- **Logistic Regression**
- **Random Forest Classifier**
- **Feedforward Neural Network**

Logistic Regression

How it works:

- Linear classification algorithm
- Models probability using logistic function
- Calculates weighted sum of input features
- Produces probability score between 0 and 1

Logistic Regression

Why it was used:

- Simple and interpretable
- Well-suited for binary classification tasks
- Easy to understand relationship between features and stroke likelihood
- Performs well with small datasets
- Suitable for understanding the impact of individual features on stroke risk

Random Forest Classifier

How it works:

- Ensemble learning algorithm
- Constructs multiple decision trees
- Combines predictions through voting or averaging
- Reduces overfitting through randomness

Random Forest Classifier

Why it was used:

- Handles complex datasets with high dimensionality
- Captures nonlinear relationships and interactions
- Robust to outliers and noise
- Reduces risk of overfitting through ensemble learning

Feedforward Neural Network

How it works:

- Artificial neural network type
- Layers: input, hidden, output
- Adjusts weights using backpropagation
- Forward pass produces predictions

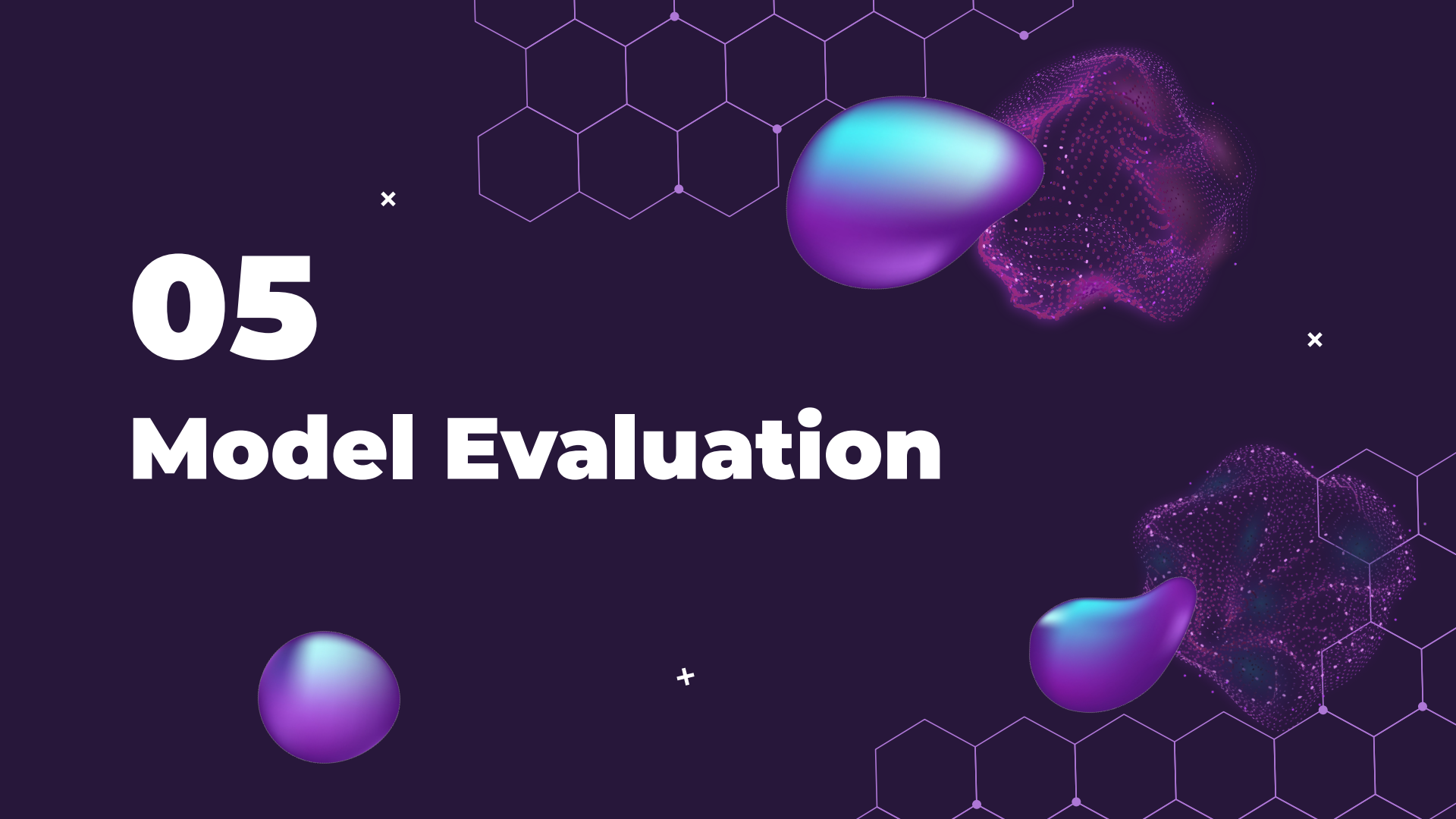
Feedforward Neural Network

Why it was used:

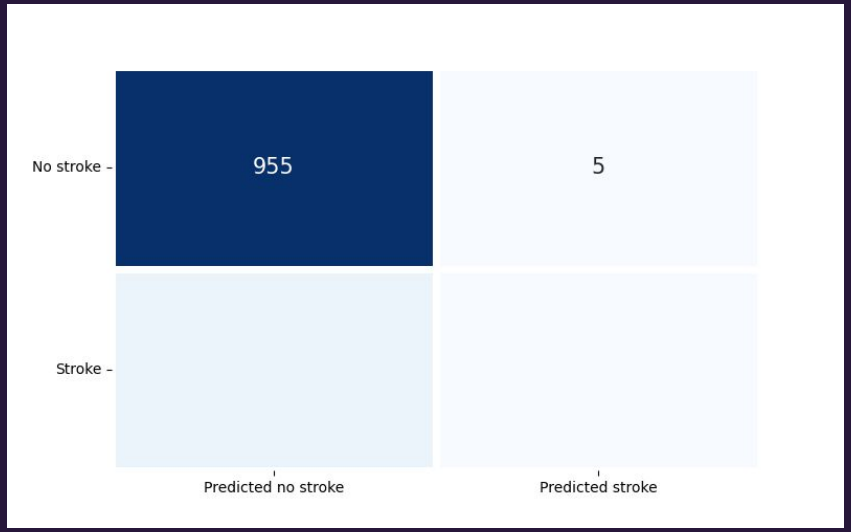
- Learns intricate patterns in data
- Automatically extracts features from raw data
- Adaptable to various data types and structures
- Continuously improves with more data

05

Model Evaluation



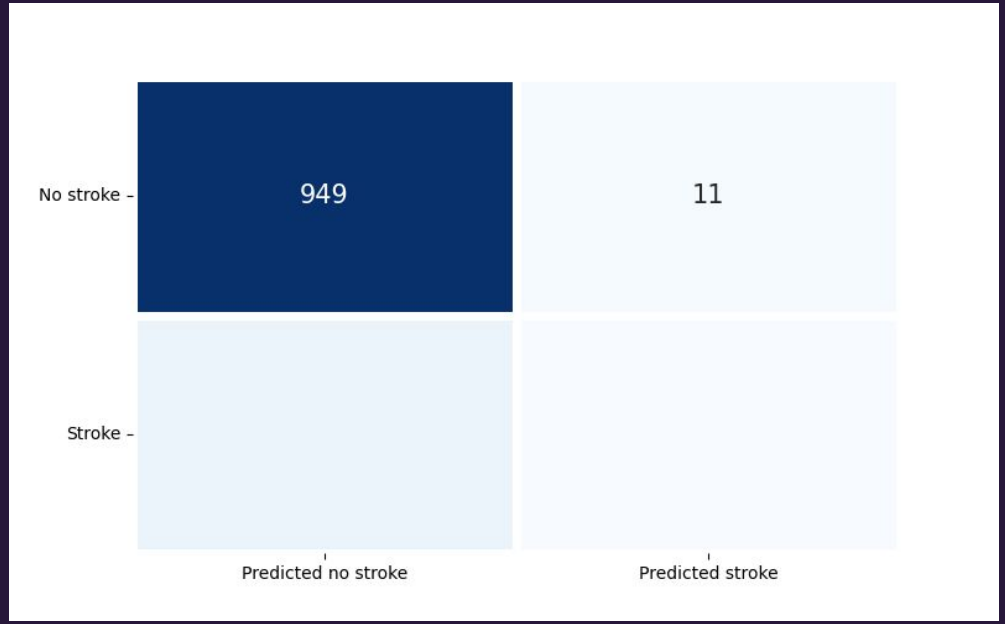
Model Evaluation



Logistic Regression
Confusion Matrix

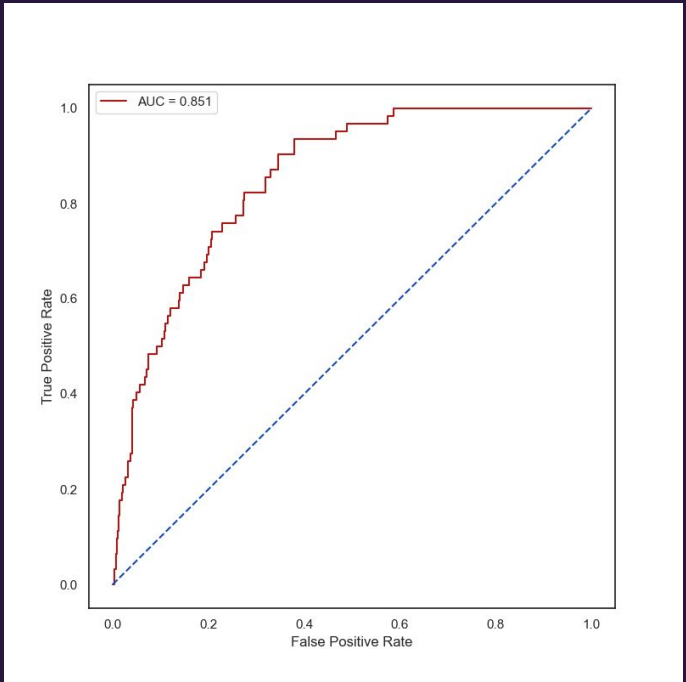
Random Forest Classifier
Confusion Matrix

Model Evaluation



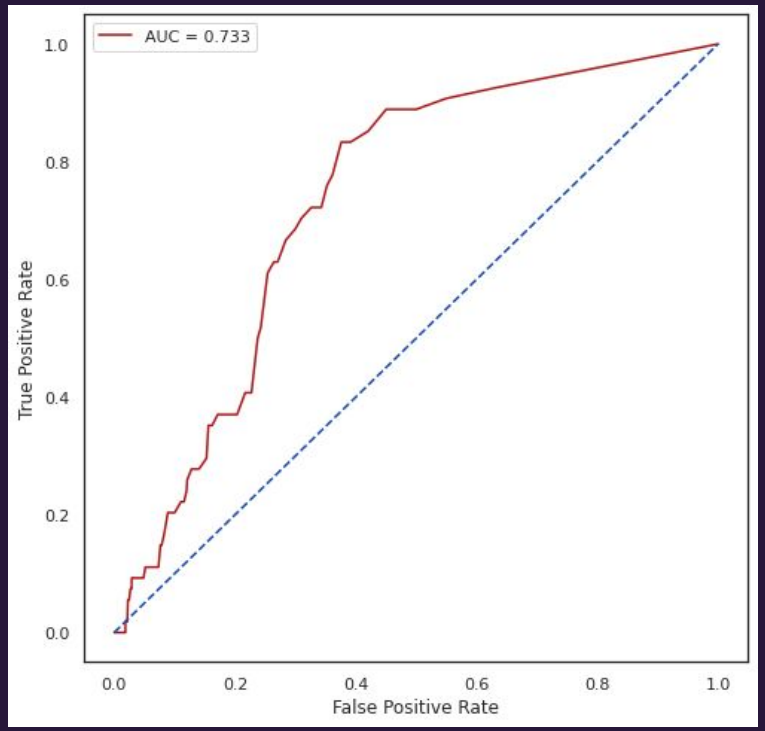
Feed Forward Neural Network
Confusion Matrix

Model Evaluation



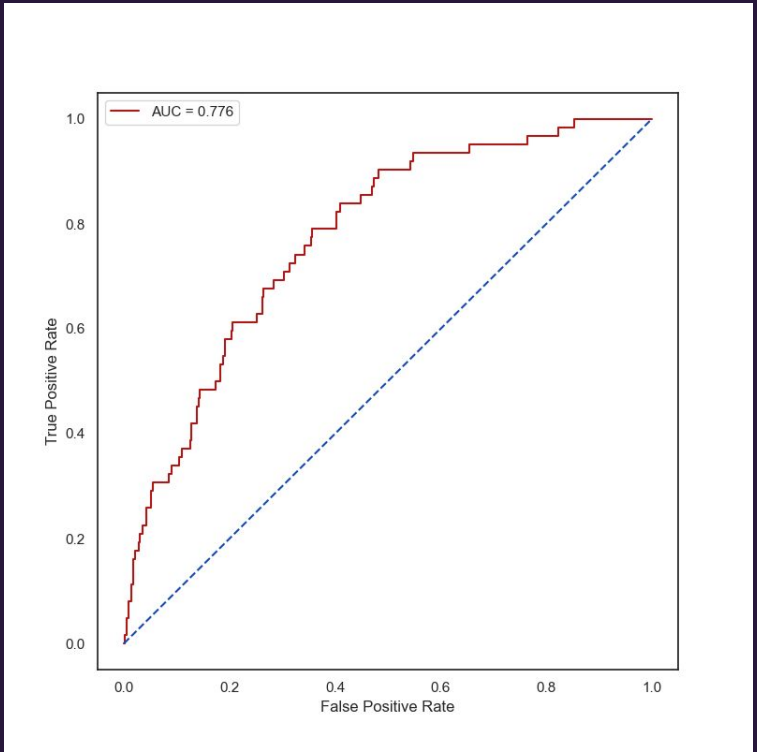
Receiver Operating Characteristics (ROC) Curve/Area Under Curve Graph for **Logistic Regression**

Model Evaluation



Receiver Operating Characteristics (ROC) Curve/Area Under Curve Graph for **Random Forest Classifier**

Model Evaluation



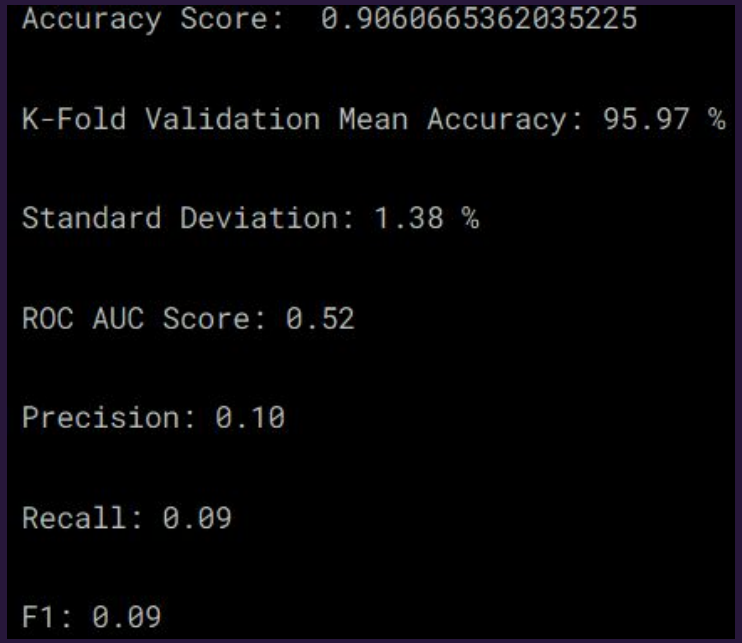
Receiver Operating Characteristics (ROC) Curve/Area Under Curve Graph for **Feedforward Neural Network**

Model Evaluation

```
Accuracy Score:  0.9363992172211351

K-Fold Validation Mean Accuracy: 95.11%

Standard Deviation: 0.06%

ROC AUC Score: 0.51

Precision: 0.29

Recall: 0.03

F1: 0.06
```

Logistic Regression
Evaluation & Performance Metrics

Model Evaluation



Random Forest Classifier
Evaluation & Performance Metrics

Model Evaluation

```
Accuracy Score: 0.9315068493150684

K-Fold Validation Mean Accuracy: 93.33%

Standard Deviation: 0.50%

ROC AUC Score: 0.52

Precision: 0.21

Recall: 0.05

F1: 0.08
```

Feed Forward Neural Network
Evaluation & Performance Metrics

Comparison

- **Accuracy:** Logistic Regression achieved the highest accuracy of 93.64%, followed by the Feedforward Neural Network (91.35%) and Random Forest Classifier (90.61%).
- **K-Fold Mean Accuracy:** Random Forest Classifier had the highest K-Fold Mean Accuracy of 95.97%, indicating better performance across different train-test splits.
- **Precision and Recall:** Logistic Regression had the highest precision (0.29), indicating a lower false positive rate, while Random Forest Classifier had the highest recall (0.09), indicating a lower false negative rate.
- **F1 Score:** Logistic Regression had the highest F1 Score (0.06), which balances precision and recall, followed by the Feedforward Neural Network (0.08) and Random Forest Classifier (0.09).

Decision

Based on these comparisons:

- **Logistic Regression** appears to be the best model for stroke prediction in this project.
- It achieves a good balance between accuracy, precision, recall, and F1 score.
- Logistic Regression achieved the highest accuracy (93.64%) among the three models
- Logistic Regression exhibited the highest precision (0.29), implying that it has the lowest rate of falsely predicting someone as at risk of stroke when they are not.
- Logistic Regression had the highest F1 score (0.06), which balances precision and recall. This indicates that it achieves a good balance between minimizing false positives and false negatives

06
Conclusion

Conclusion

Implications for Stroke Prediction and Prevention:

- Accurate prediction models like Logistic Regression can play a crucial role in early identification of individuals at a high risk of stroke.
- Timely interventions and preventive measures based on accurate predictions can potentially reduce the incidence and severity of strokes.
- This would lead to improved patient outcomes, prevent long term disability, and reduced healthcare costs.

Conclusion

Limitations:

- One limitation is the reliance on a single dataset, which may not capture the full complexity of stroke risk factors.
- Future research could explore additional datasets to validate the findings and improve model generalization.

Recommendations:

- Further research into feature engineering and selection techniques could enhance model performance and interpretability.
- Continuous monitoring and refinement of the predictive model with new data can ensure its relevance and effectiveness in medical practice.

Thank You!