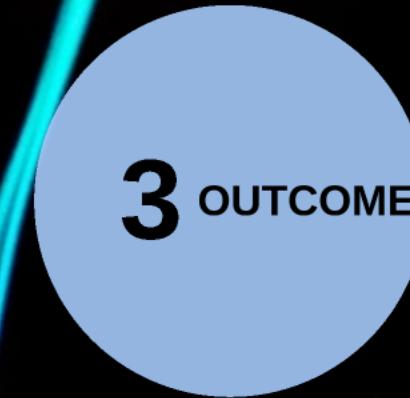
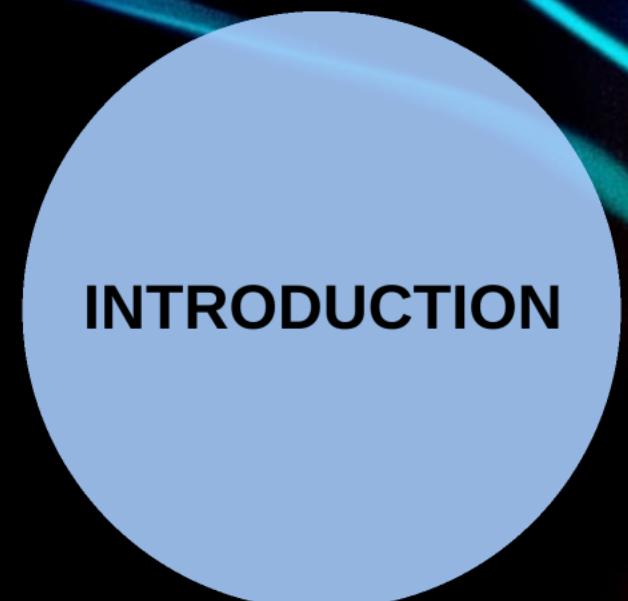


PROJECT 4

Stroke Prediction

Bec B & Bec N



29 August 2023

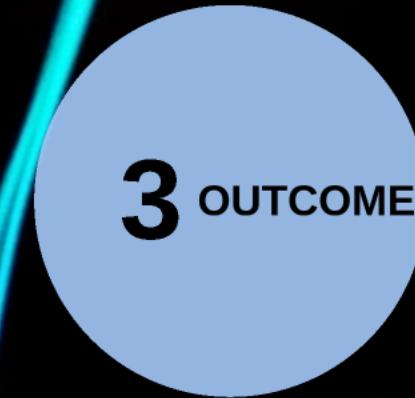
AGENDA

- Project Overview
- Data techniques
- Project implementation
- Outcome
- Visualisation

PROJECT 4

Stroke Prediction

Bec B & Bec N



29 August 2023

1 PROJECT OVERVIEW

DEFINITIONS

DATA
SOURCE

PROJECT
OBJECTIVE

What is a stroke?

- Insufficient blood to areas of the brain by:
 - Blood vessel being blocked; or
 - Blood vessel ruptures and bleeds.
- Brain cells die due to lack of oxygen.
- Globally 1 in 4 adults over the age of 25 will have a stroke in their lifetime*



DATA SOURCE

The stroke prediction dataset was obtained from Kaggle.

The data contains information on:

- Gender
- Age
- Hypertension
- Heart disease
- Marriage status
- Work type
- Residence type
- Glucose levels
- BMI
- Smoking status
- Stroke status

The Kaggle dataset was obtained from patient records

A	B	C	D	E	F	G	H	I	J	K	L
1	id	gender	age	hypertens	heart_dise	ever_marri	work_type	Residence	avg_gluc	bmi	smoking_stroke
2	9046	Male	67	0	1	Yes	Private	Urban	228.69	36.6	formerly smokes
3	51676	Female	61	0	0	Yes	Self-emplo	Rural	202.21	N/A	never smokes
4	31112	Male	80	0	1	Yes	Private	Rural	105.92	32.5	never smokes
5	60182	Female	49	0	0	Yes	Private	Urban	171.23	34.4	smokes
6	1665	Female	79	1	0	Yes	Self-emplo	Rural	174.12	24	never smokes
7	56669	Male	81	0	0	Yes	Private	Urban	186.21	29	formerly smokes
8	53882	Male	74	1	1	Yes	Private	Rural	70.09	27.4	never smokes
9	10434	Female	69	0	0	No	Private	Urban	94.39	22.8	never smokes
10	27419	Female	59	0	0	Yes	Private	Rural	76.15	N/A	Unknown
11	60491	Female	78	0	0	Yes	Private	Urban	58.57	24.2	Unknown
12	12109	Female	81	1	0	Yes	Private	Rural	80.43	29.7	never smokes
13	12095	Female	61	0	1	Yes	Govt_job	Rural	120.46	36.8	smokes
14	12175	Female	54	0	0	Yes	Private	Urban	104.51	27.3	smokes
15	8213	Male	78	0	1	Yes	Private	Urban	219.84	N/A	Unknown
16	5317	Female	79	0	1	Yes	Private	Urban	214.09	28.2	never smokes
17	58202	Female	50	1	0	Yes	Self-emplo	Rural	167.41	30.9	never smokes
18	56112	Male	64	0	1	Yes	Private	Urban	191.61	37.5	smokes
19	34120	Male	75	1	0	Yes	Private	Urban	221.29	25.8	smokes
20	27458	Female	60	0	0	No	Private	Urban	89.22	37.8	never smokes
21	25226	Male	57	0	1	No	Govt_job	Urban	217.08	N/A	Unknown
22	70630	Female	71	0	0	Yes	Govt_job	Rural	193.94	22.4	smokes
23	13861	Female	52	1	0	Yes	Self-emplo	Urban	233.29	48.9	never smokes
24	68794	Female	79	0	0	Yes	Self-emplo	Urban	228.7	26.6	never smokes

Data entries before sanitisation:

0 (non stroke): 4861
1 (stroke): 249

Data entries after sanitisation:

0 (non stroke): 4235
1 (stroke): 165



Project Objective

Understanding risk factors are important in the prevention of strokes, their identification early on is crucial for preventing strokes.

The objective of this project is to use a machine learning algorithm to develop an early detection model for strokes, using a stroke prediction dataset

PROJECT 4

Stroke Prediction

Bec B & Bec N



29 August 2023

2 IMPLEMENTATION

APPROACH

DATA
EXPLORATION

STATISTICAL
ANALYSIS

DATA
TECHNIQUES

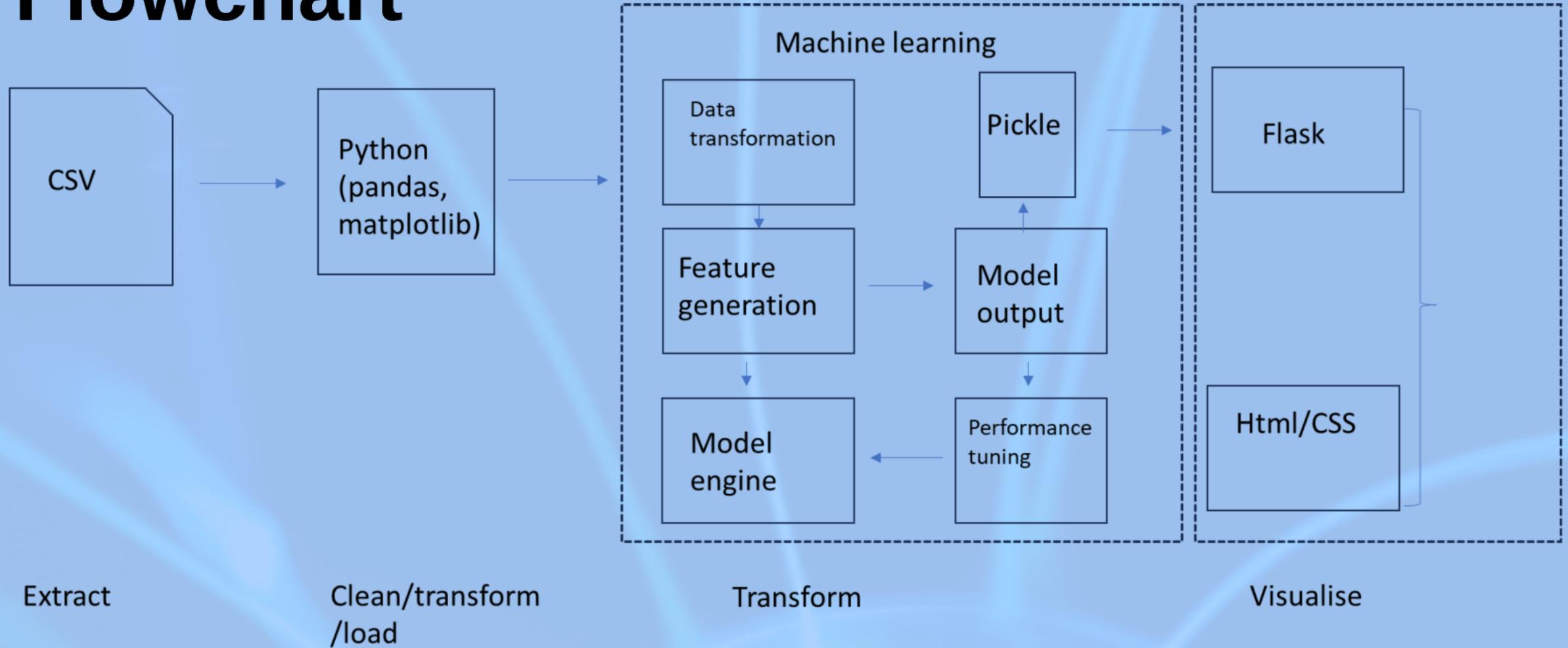
MODEL
SELECTION

APPROACH



- Identify data sources and import dependencies
- Perform Exploratory Data Analysis (**EDA**), determine feature set and transform the stroke data
- Compile, train and evaluate the model
- Compare models for optimization of accuracy metric
- Store the transformed dataset into pickle
Create Flask App, import data in via pickle and connect routes to model
- Create interactive web app using pickle, html and css

Flowchart



DATA EXPLORATION AND CLEANING

- Investigate NaN cells
- Remove unnecessary columns ("ID")
- Dropping duplicates

DATA CLEANING

```
n [44]: stroke_df.isnull().sum()

Out[44]: id          0
gender        0
age           0
hypertension   0
heart_disease 0
ever_married   0
work_type      0
Residence_type 0
avg_glucose_level 0
bmi           201
smoking_status 0
stroke         0
dtype: int64
```

```
n [45]: #remove unnecessary columns
stroke_df.drop(['id'],axis=1,inplace = True)

n [46]: #remove duplicates values
stroke_df.drop_duplicates(inplace=True)
```

```
n [47]: #null data handled
stroke_df.isnull().sum()

Out[47]: gender        0
age           0
hypertension   0
heart_disease 0
ever_married   0
work_type      0
Residence_type 0
avg_glucose_level 0
bmi           201
smoking_status 0
stroke         0
dtype: int64
```

```
# Replace NaN values in the "bmi" column with the average BMI of the corresponding age
def replace_bmi(row):
    if pd.isna(row['bmi']):
        return avg_bmi_by_age[row['age']]
    else:
        return row['bmi']

stroke_df['bmi'] = stroke_df.apply(replace_bmi, axis=1)

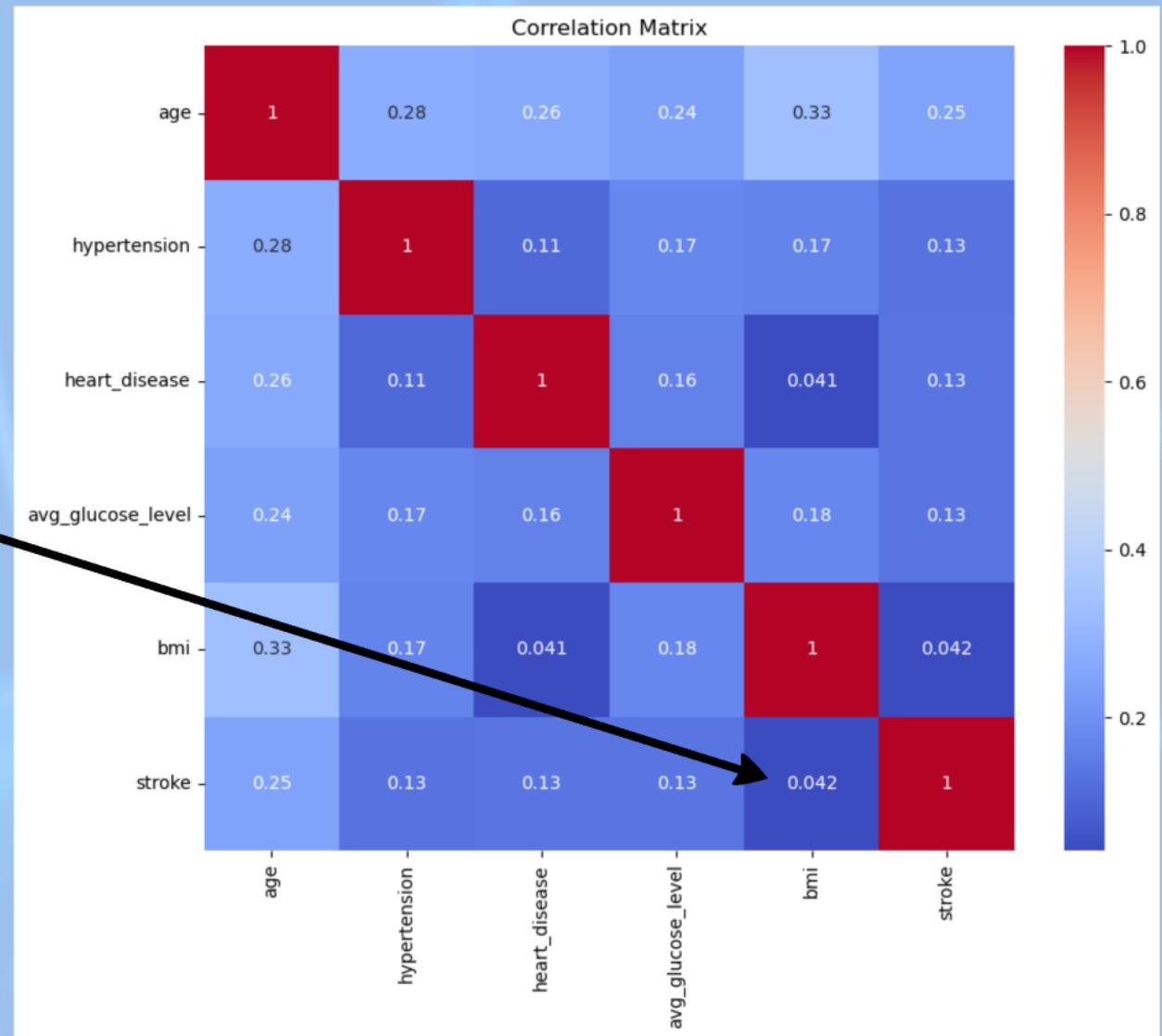
# Check the info of the dataframe (if the NaN values in the 'bmi' column are replaced)
stroke_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 4401 entries, 2 to 5109
Data columns (total 11 columns):
 #   Column            Non-Null Count Dtype  
 ---  -- 
 0   gender            4401 non-null  object  
 1   age               4401 non-null  float64 
 2   hypertension       4401 non-null  int64  
 3   heart_disease     4401 non-null  int64  
 4   ever_married      4401 non-null  object  
 5   work_type          4401 non-null  object  
 6   Residence_type    4401 non-null  object  
 7   avg_glucose_level 4401 non-null  float64 
 8   bmi                4401 non-null  float64 
 9   smoking_status     4401 non-null  object  
 10  stroke             4401 non-null  int64  
dtypes: float64(3), int64(3), object(5)
memory usage: 541.6+ KB
```

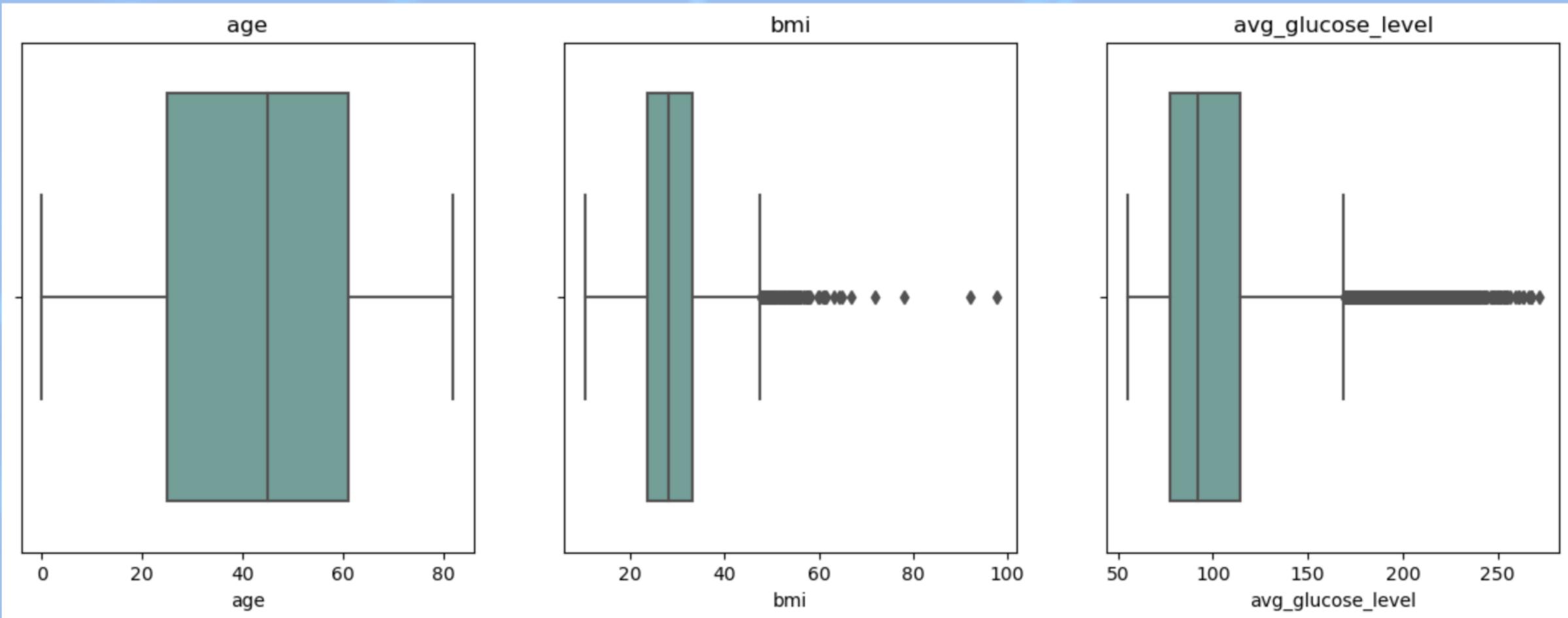
Correlation Matrix

Most of the features are not highly correlated with any other features

BMI has a weakly positive correlation with stroke

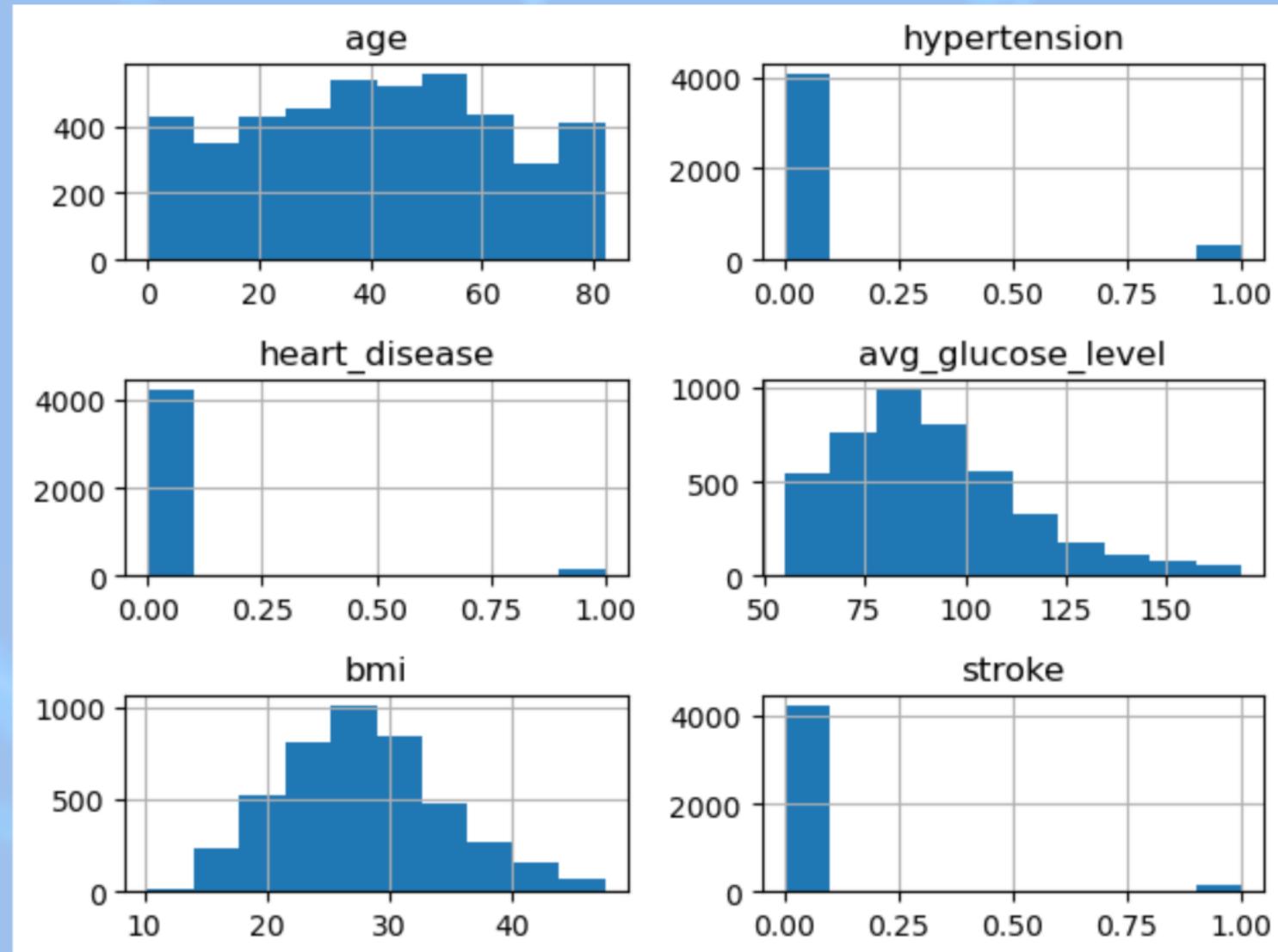


Removing Outliers

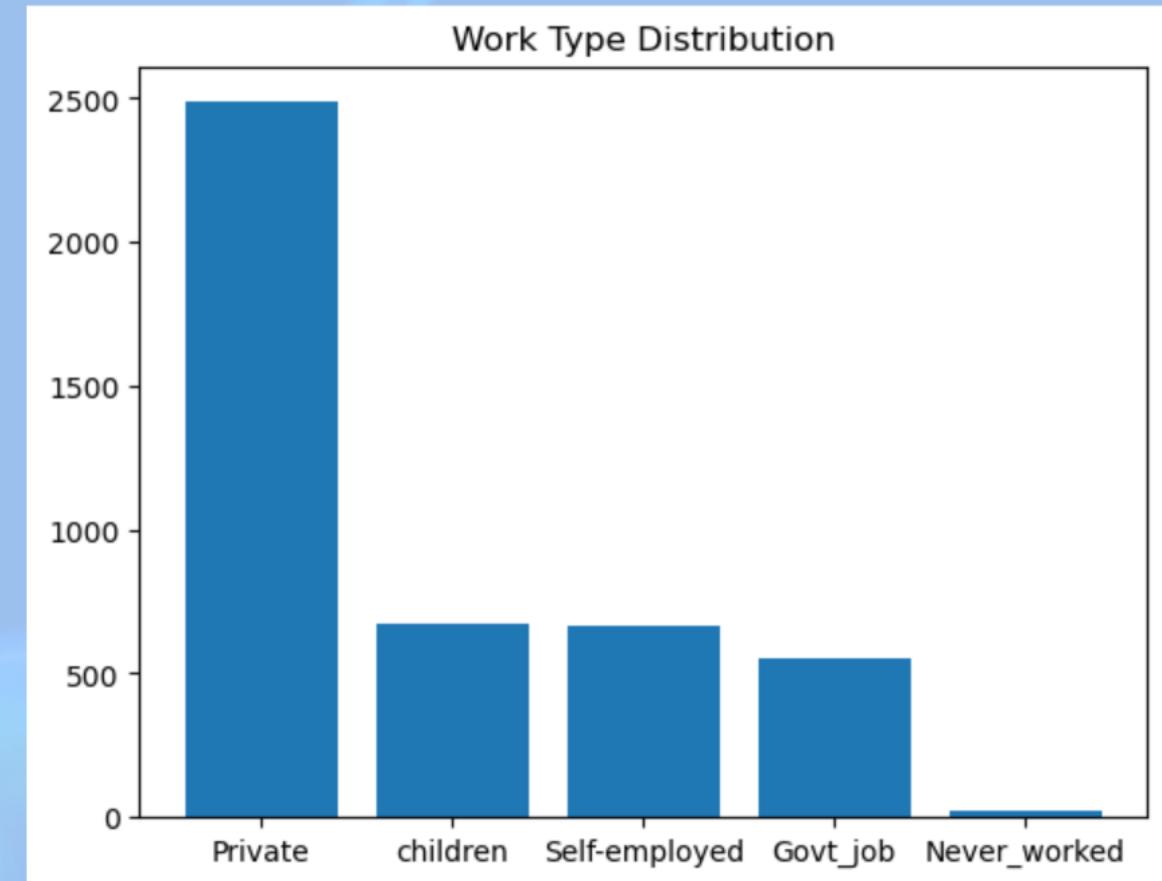
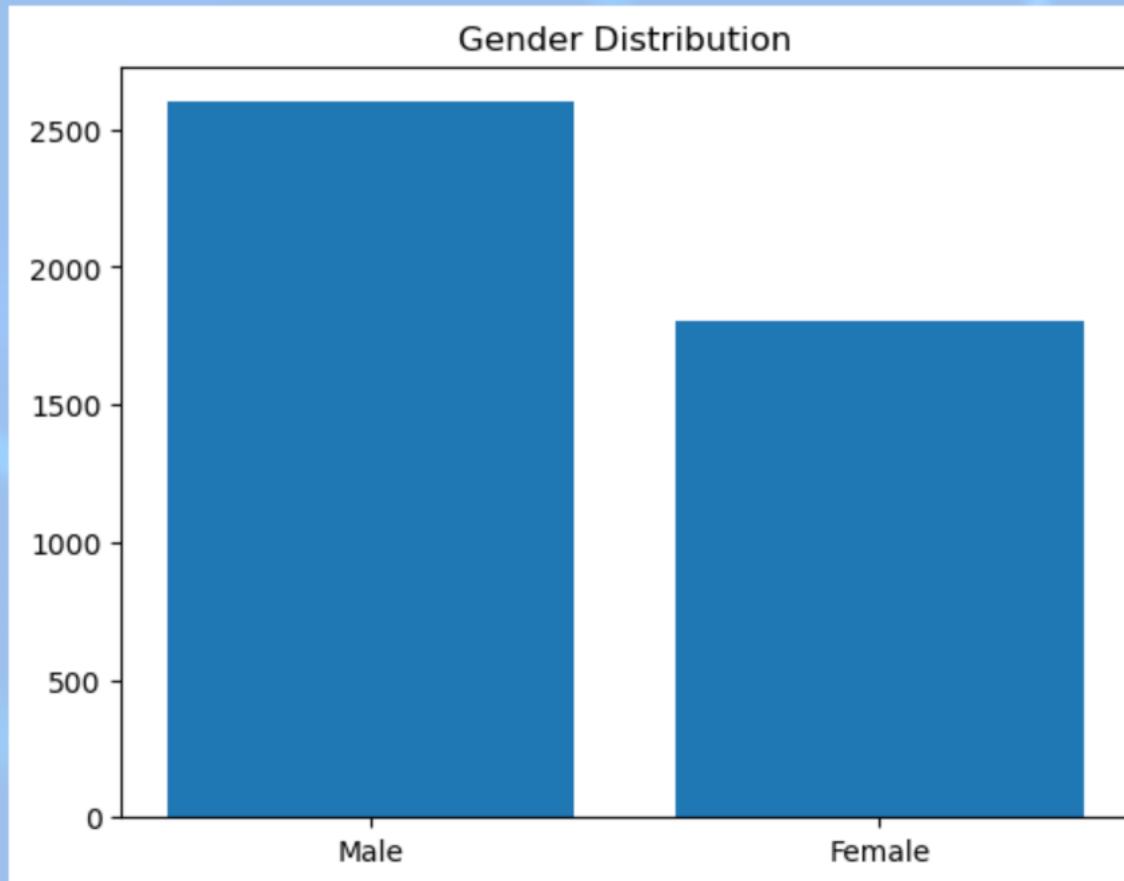


Outliers were removed from the dataset. BMI and average glucose level fields contained numerous outliers therefore were removed from the dataset, particularly as BMI >60 is very rare

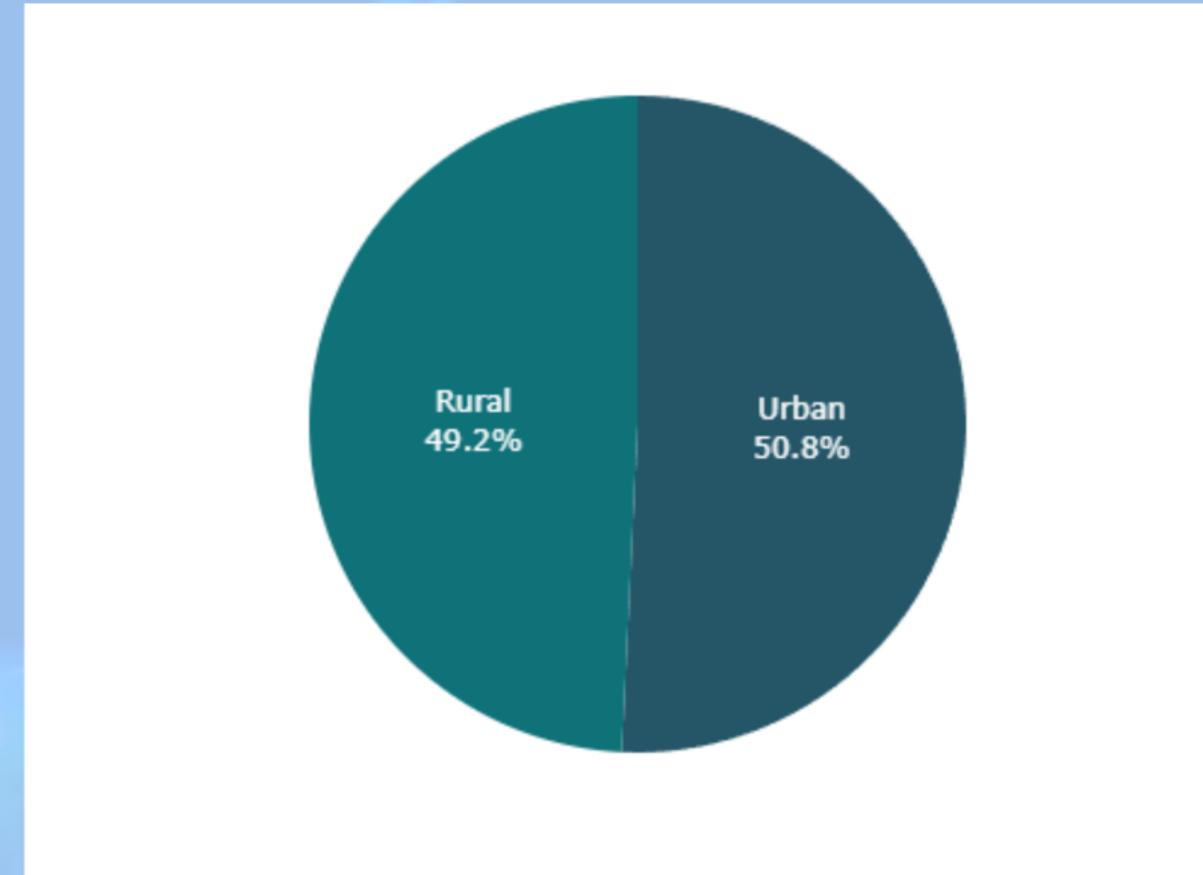
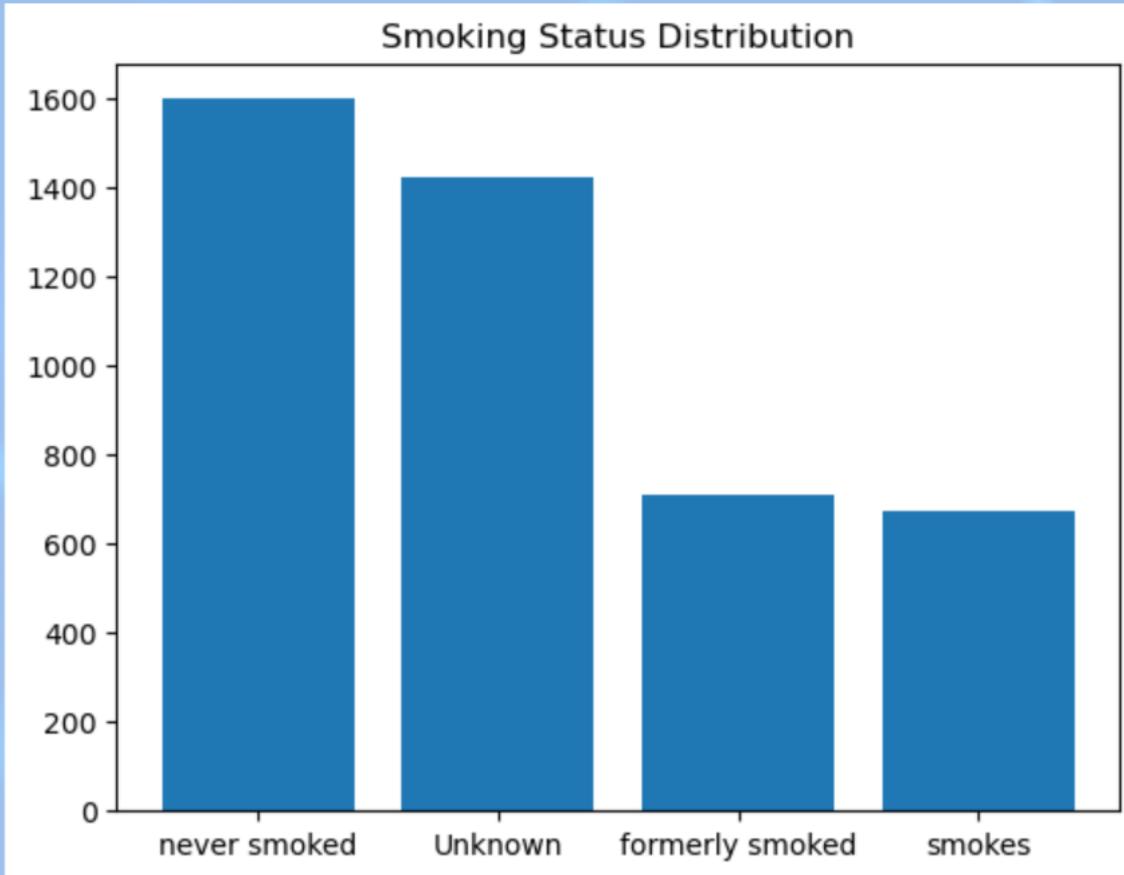
Data Distribution



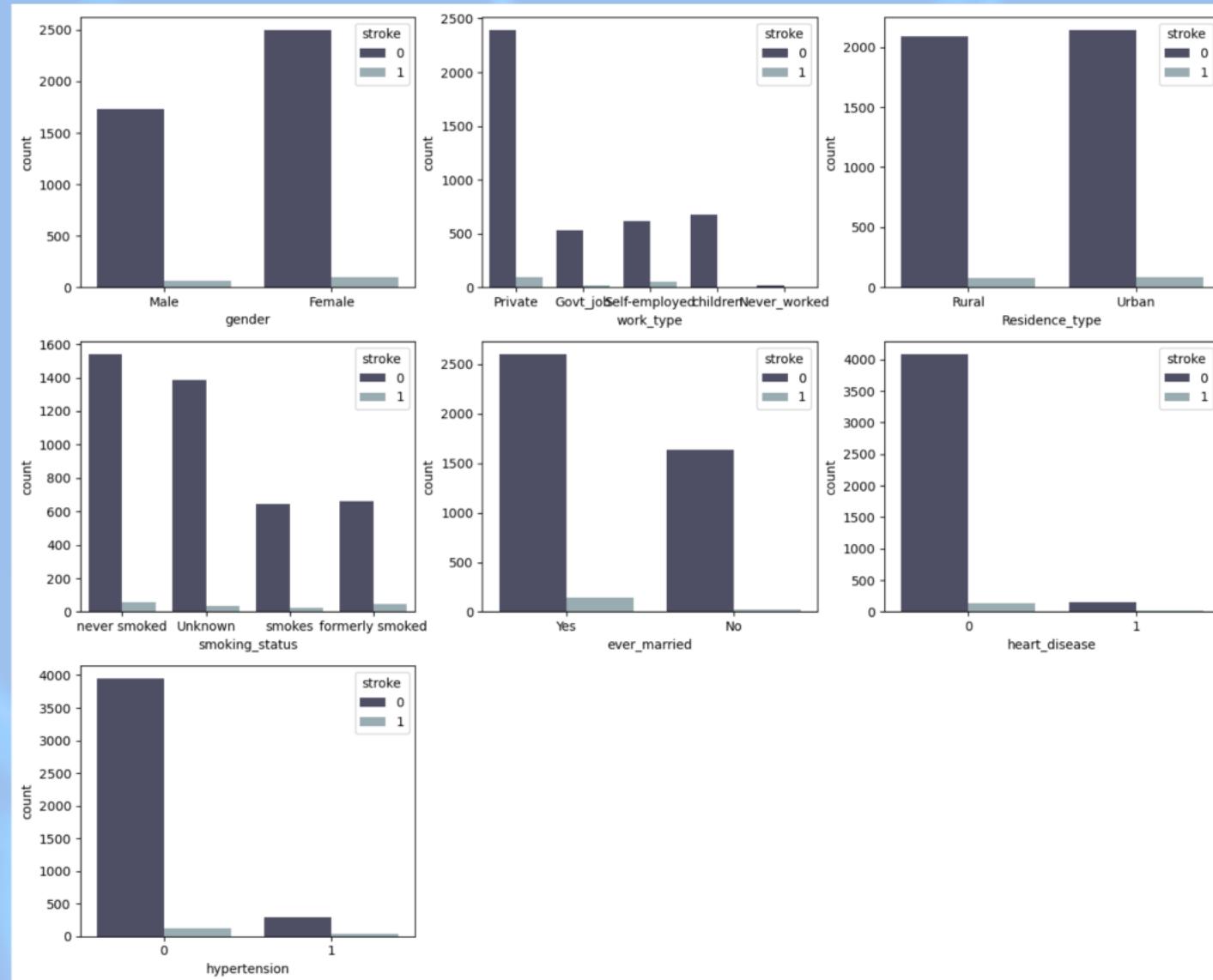
Data Distribution



Data Distribution



Data Distribution



DATA TECHNIQUES

The following techniques were used:

- Python Packages including scikit-learn, matplotlib, Pandas, seaborn
- pickle
- Flask
- HTML
- CSS



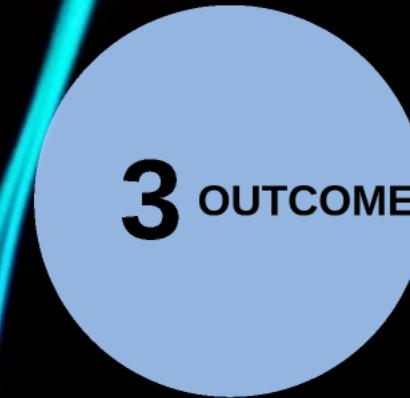
MODEL SELECTION

- Random Forest model with Synthetic Minority Oversampling Technique (**SMOTE**)
- Random Forest and binning
- Random Forest
- Decision Tree
- Logistic Regression

PROJECT 4

Stroke Prediction

Bec B & Bec N



29 August 2023

3 OUTCOME

MODEL
ACCURACY

DECISION
TREE

WEBPAGE

CONCLUSION

MODEL ACCURACY

Model 1: SMOTE and Random Forest Classifier

	TRAINING		TESTING	
	Predicted 0	Predicted 1	Predicted 0	Predicted 1
Actual 0	3,385	-	848	2
Actual 1	-	3,391	27	817

Accuracy 98.2%

Classification Report

	Precision	Recall	F1-Score	Support
0	0.97	1.00	0.98	850
1	1.00	0.97	0.98	844
Accuracy				0.98 1,694
Macro avg	0.98	0.98	0.98	1,694
Weighted avg	0.98	0.98	0.98	1,694

Model 2: Random Forecast Classifier with binning

	TRAINING		TESTING	
	Predicted 0	Predicted 1	Predicted 0	Predicted 1
Actual 0	3,385	-	844	6
Actual 1	-	3,391	-	844

Accuracy 99.6%

Classification Report

	Precision	Recall	F1-Score	Support
0	1.00	0.99	1.00	850
1	0.99	1.00	1.00	844
Accuracy				1.00 1,694
Macro avg	1.00	1.00	1.00	1,694
Weighted avg	1.00	1.00	1.00	1,694

Model 3: Random Forecast Classifier

	TRAINING		TESTING	
	Predicted 0	Predicted 1	Predicted 0	Predicted 1
Actual 0	3,385	-	843	7
Actual 1	-	3,391	-	844

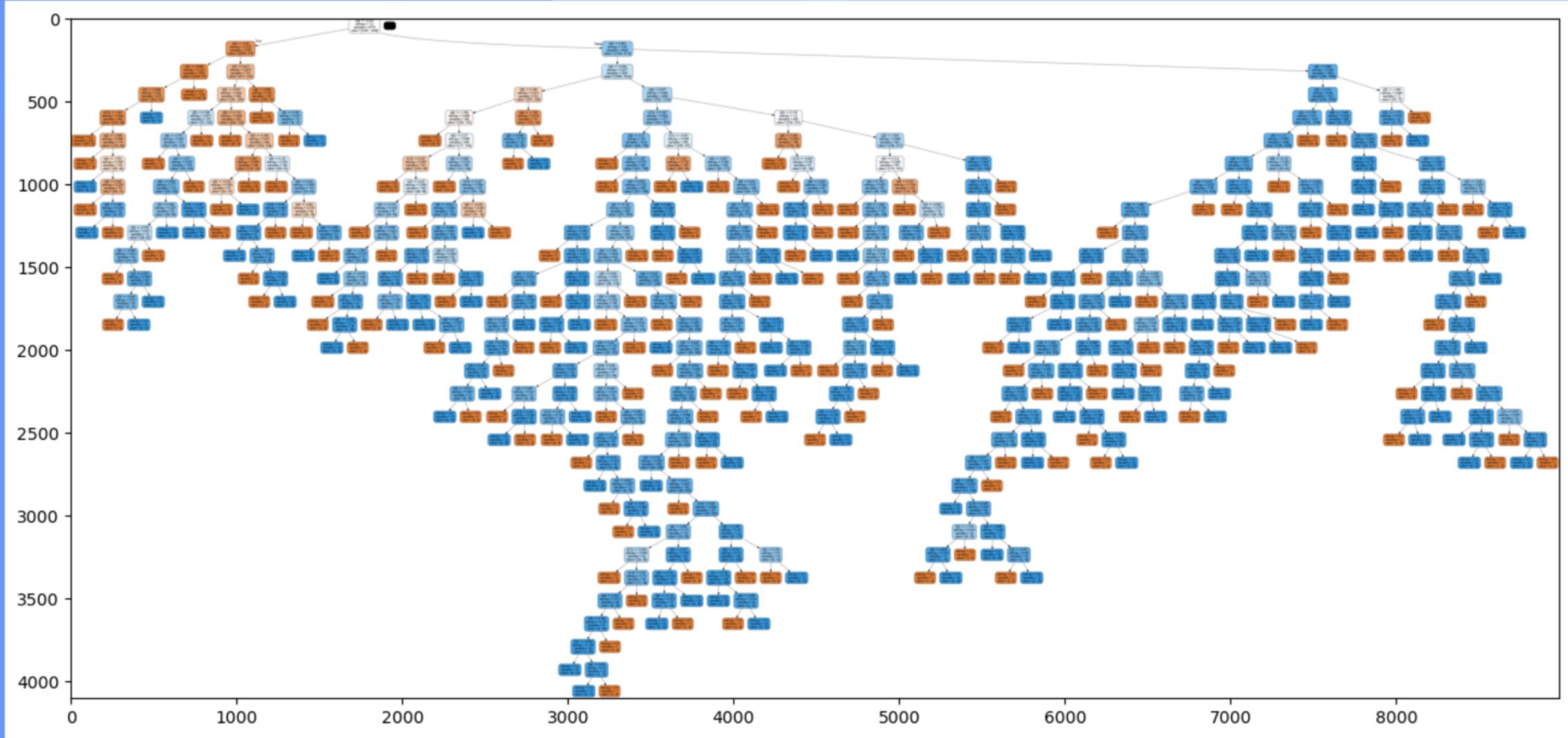
Accuracy 99.6%

Classification Report

	Precision	Recall	F1-Score	Support
0	1.00	0.99	1.00	850
1	0.99	1.00	1.00	844
Accuracy				1.00 1,694
Macro avg	1.00	1.00	1.00	1,694
Weighted avg	1.00	1.00	1.00	1,694

DECISION TREE

F1 Score: 66%



WEBPAGE

STROKE PREDICTION

Gender:

Age:

Hypertension:

Heart Diseases:

Marital Status:

Work-Type:

Residency Type:

Glucose Levels:

BMI:

Smoking Status:

THANKS FOR LISTENING

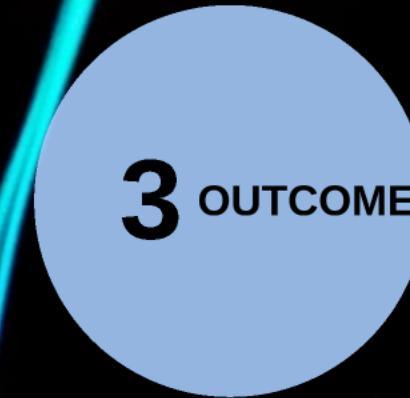


Happy graduation :) Thanks James and Shriya!

PROJECT 4

Stroke Prediction

Bec B & Bec N



29 August 2023