Stroke Prediction

Random Forest Classifier

Team 6:
Yeyan Wang, Meera G K, Shweta J,
Reed Zimpfer, Dang Tran

Project Overview

- **Background:** Stroke is a severe condition caused by interrupted blood supply to the brain, leading to brain damage, disability, or death. Various factors, such as age, gender, hypertension, heart disease, obesity, and smoking, influence stroke risk
- **Objective:** Develop a machine learning model for early stroke prediction
- **Dataset:** Utilized a healthcare dataset containing various features related to stroke risk

  https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset?select=healthcare-dataset-stroke-data.csv
- **Target Audience:** Clinical Providers, Medical Experts can exploit the established model and use it as an additional resource to access stroke occurrence risk

Data Overview

```
# Import and read the healthcare-dataset-stroke-data.csv.
import pandas as pd
stroke_df = pd.read_csv("data/healthcare-dataset-stroke-data.csv")
stroke_df.head()
```

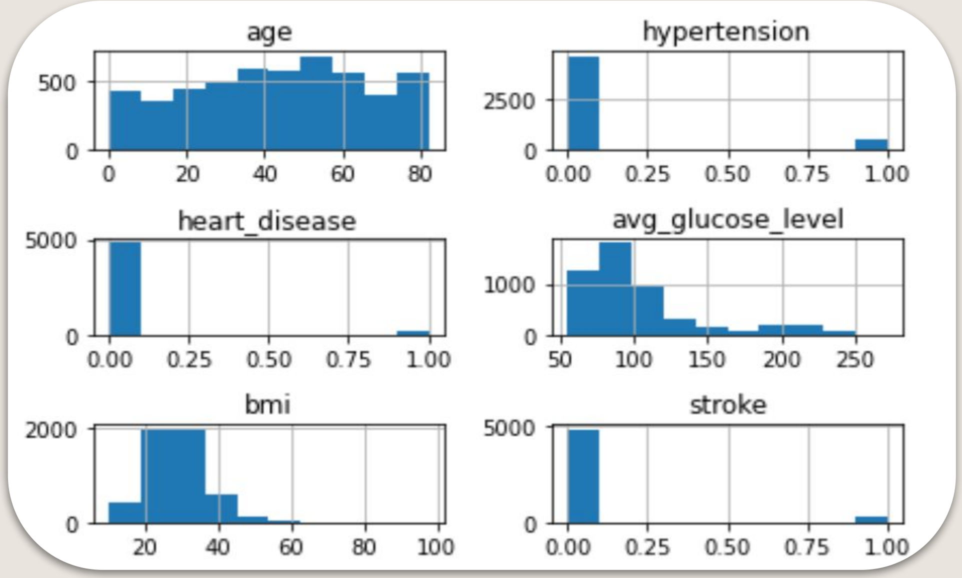
	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
0	9046	Male	67.0	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
1	51676	Female	61.0	0	0	Yes	Self-employed	Rural	202.21	NaN	never smoked	1
2	31112	Male	80.0	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1
3	60182	Female	49.0	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
4	1665	Female	79.0	1	0	Yes	Self-employed	Rural	174.12	24.0	never smoked	1

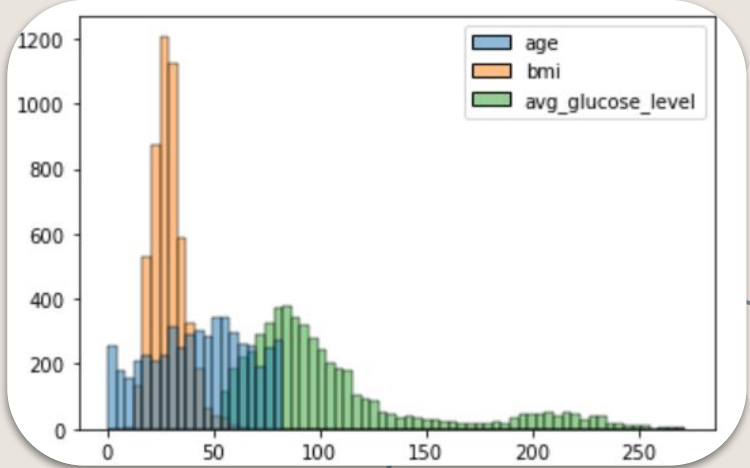
Model Selection

- Looking at the data where the stroke column has either 0 or 1 values, it indicates that stroke prediction is a classification problem.

- The **Random Forest Classifier** was selected for this problem due to its reputation for achieving high accuracy in classification tasks. It is a popular choice in the healthcare and medical industry, where precise and reliable predictions are crucial.

Examine Data Distribution



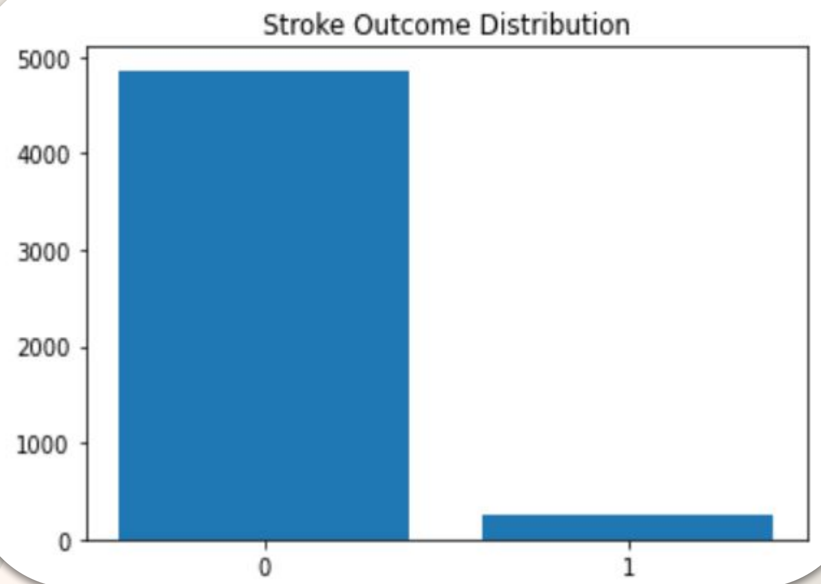


Examine Data Distribution on Target Variable

```
# Look at the stroke outcome value counts
stroke_counts = stroke_df['stroke'].value_counts()
stroke_counts
```

```
0    4860
1     249
Name: stroke, dtype: int64
```

Findings: The `0` s and `1` s in stroke column is highly imbalanced



Examine & Impute Missing Values

```python
# 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)
```

```python
# Examine the total NaN values for each column
stroke_df.isnull().sum()
```

```
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
```

Examine & Handle Singleton Record

```python
# Look at gender value counts
gender_counts = stroke_df['gender'].value_counts()
gender_counts
```

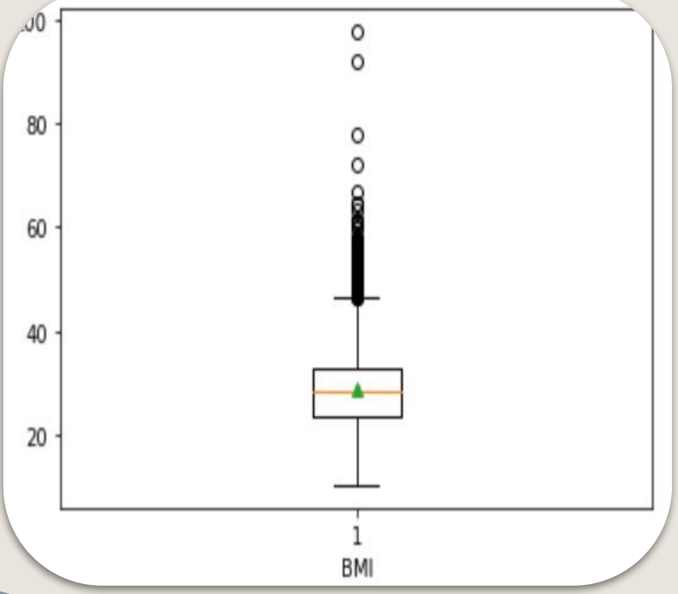
```
Female    2994
Male      2115
Other        1
```

```python
# Drop the record with gender = 'Other' (since there is only 1 record)
stroke_df = stroke_df.drop(stroke_df[stroke_df['gender'] == 'Other'].index)
```

```python
# Check if 'Other' is dropped on gender column
stroke_df['gender'].unique()
```

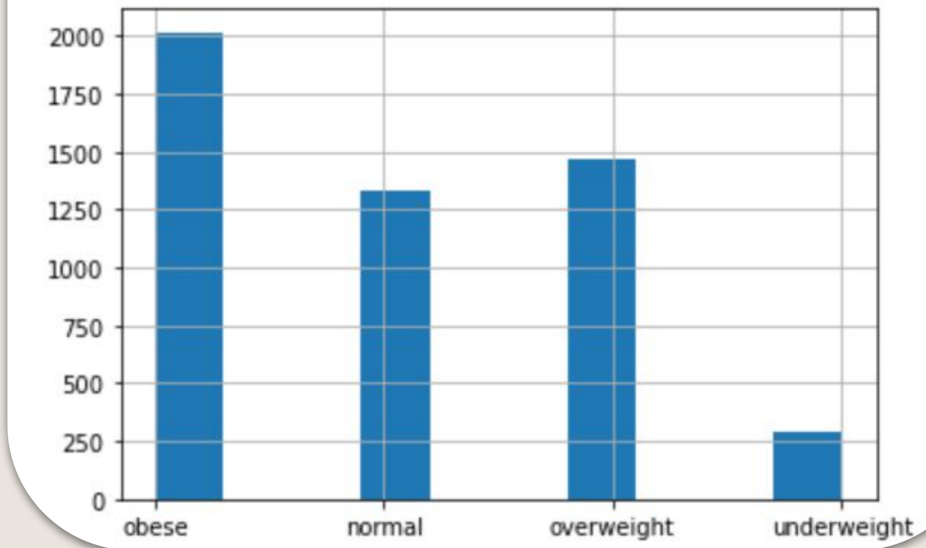
```
array(['Male', 'Female'], dtype=object)
```

Binning




```
# Check the distribution of the bmi bins
stroke_df['binned_bmi'].hist()
```

<AxesSubplot:>



Data Preprocessing

1. Handle Imbalance Data
2. Feature Scaling
3. Encode Categorical Variables
4. Train-Test Split

Evaluation

	Accuracy	FPR	FNR
Model 1 RandomOverSampler	***99.1%***	***1.82%***	***0%***
Model 2 SMOTE	***97.3%***	***0.82%***	***4.68%***
Model 3 ROS + binning	***98.9%***	***0.22%***	***0%***

Limitations

Data/Model limitation

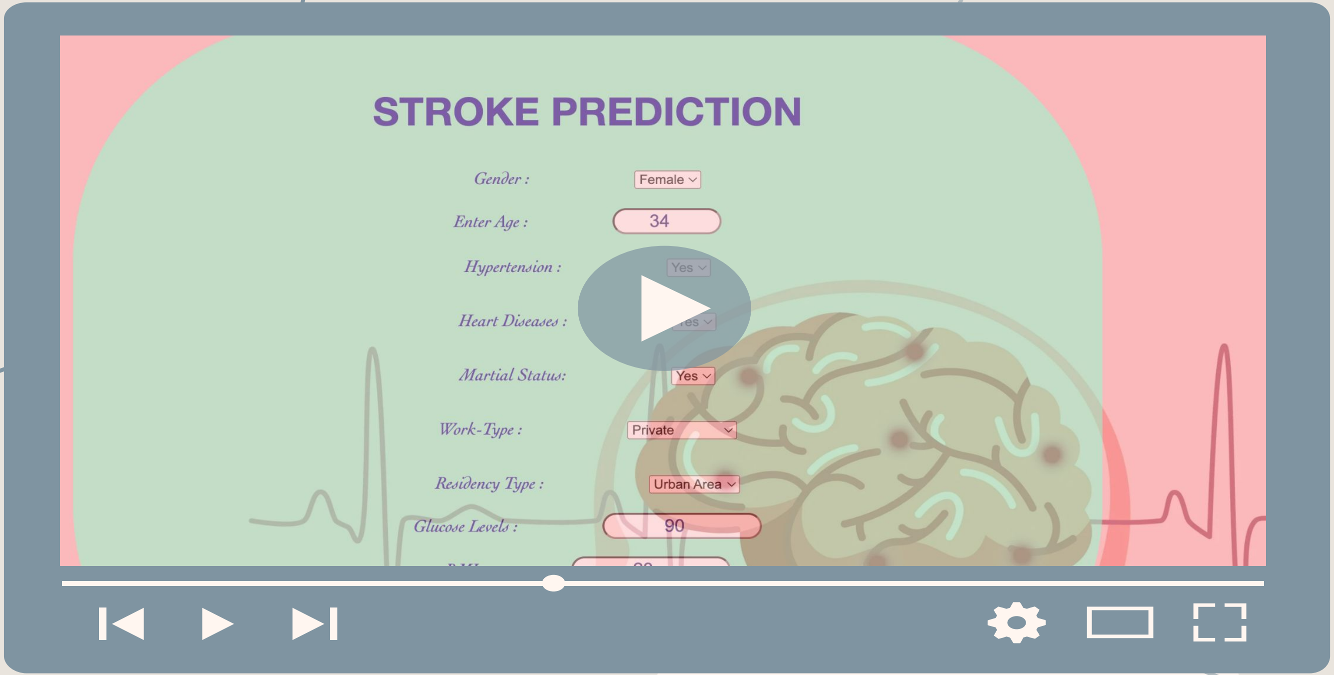
- A limitation of this study is that it was based on a publicly available dataset. These data are of specific size and features as opposed to data from a hospital or institute. Although the latter could give more rich information data models with various features capturing a detailed health profile of the participants, acquiring access to such data is usually time-consuming and difficult for privacy reasons
- Some parts of the dataset were incomplete and was presented with NaN values. To combat this, we've replaced the missing values with the mean or the average of the data that was available. This is an assumption that we've inserted into the model and is not a representation of the actual data itself. Therefore, this can skew the result and causes inaccurate reading from the model.

Conclusion

- **Model Performance:** Our final model achieved an impressive accuracy rate of around 99.1%, with minimal false positives and false negatives.
- **Refinement Strategies:** To further improve the model, we can consider introducing more unseen data during the training process and integrating additional data sources. Additionally, exploring alternative algorithms and approaches may also enhance its performance.

Developing a machine learning model is an ongoing and iterative process that involves continuous experimentation and fine-tuning. This study serves as a crucial initial step in that journey.

DEMO


STROKE PREDICTION
Gender :
Female
Enter Age :
34
Hypertension :
Yes
Heart Diseases :
Martial Status:
Yes
Work-Type :
Private
Residency Type :
Urban Area
Glucose Levels :
90

THANK YOU KEVIN!
HAPPY GRADUATION EVERYONE!

👏👏👏