Heart Stroke Prediction: Analysis and Modeling

Tanvi Gandhi

HCS237023

Course: HSL 522

Supervisor: Prof. Ashwini Vaidya

Overview

- Motivation
- Dataset
- Background
- Exploratory Data Analysis
- Data Pre-processing
- Data Modeling and Results
- Future Work
- References

Motivation

- **Annually, 15 million people worldwide suffer a stroke.**
- **In this analysis, we explore the factors that influence stroke occurrences and employ machine learning techniques to predict stroke risks.**
- **Early prediction and identification of potential stroke cases are crucial for timely medical intervention and better health outcomes.**

Dataset

- **The dataset is available on Kaggle.**
- **It comprises 5110 observations and 11 attributes.**
- **Each row in the dataset contains relevant information about an individual person.**
- **The dataset includes input parameters such as gender, age, various diseases, and smoking status etc.**
- **The dataset can be used to predict whether a patient is likely to have a stroke.**

Dataset

- gender
- hypertension
- heart_disease
- ever_married
- work_type
- resident_type
- smoking_status

Categorical variables

- age
- avg_glucose_level
- bmi

Continuous variables

- stroke

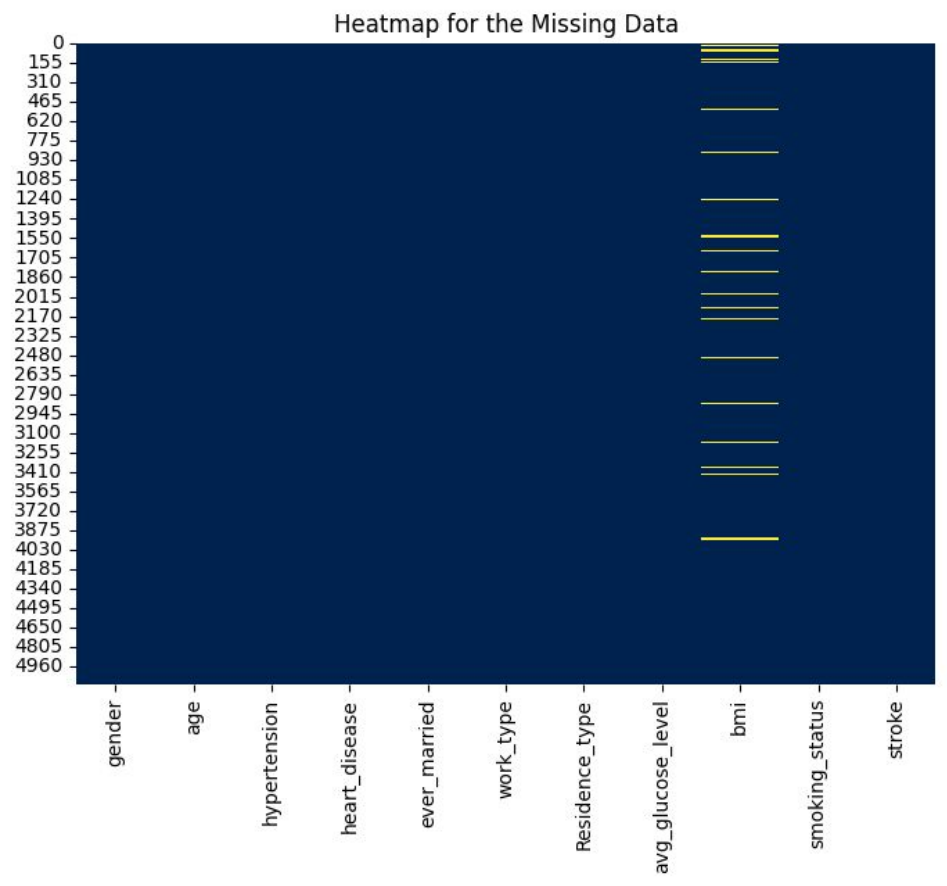
Target variable

	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

Background

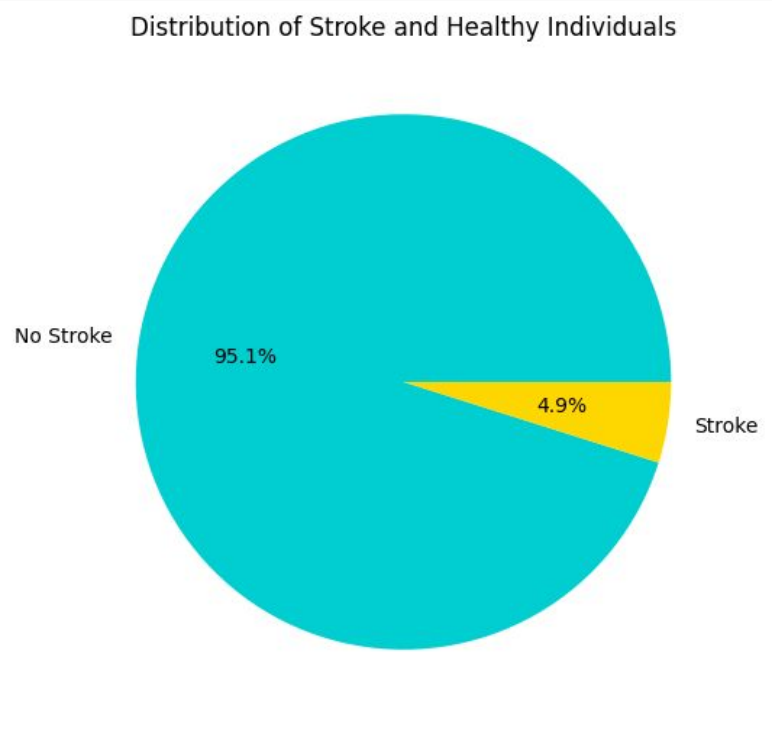
- Cynthia, Huang, Yizhi, Zhang, Yitian, Hu, Juien, Yang; Carnegie Mellon University: Visual Analysis and Prediction of Stroke [Electronic resource]. Access here.
- Research Questions:
  - What attributes are associated with stroke?
  - Can we reduce the linear dependence among the variables and build a good regression model for predicting whether one has stroke or not?
  - Will classification techniques be more effective in predicting if one has stroke?
- The training accuracy of the linear model using first five principal components was 73 % and for the classifier using decision tree model was 78.6%.

Exploratory Data Analysis




```
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
```

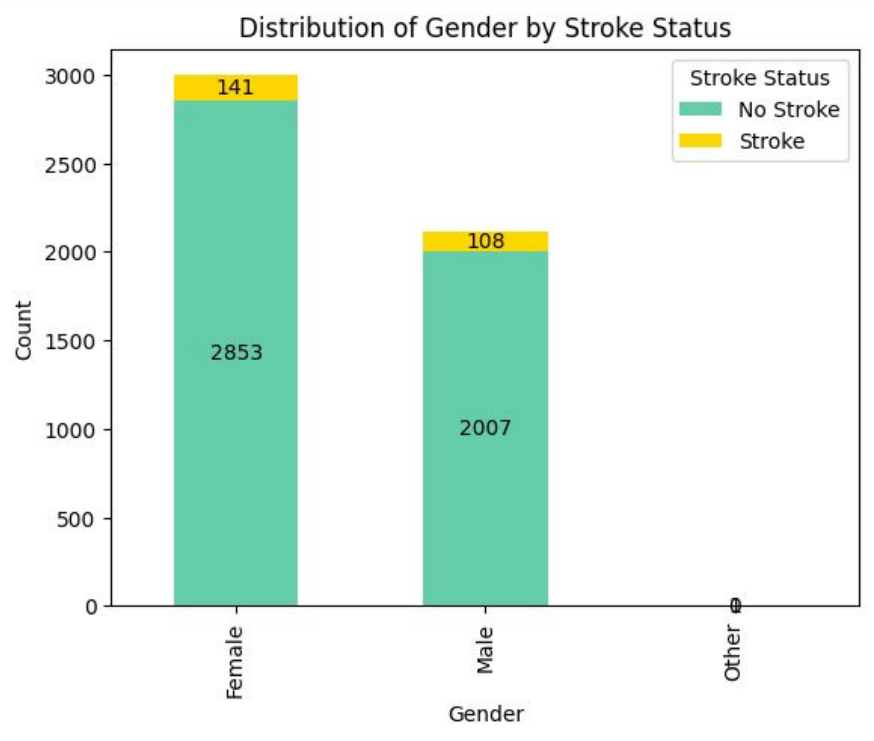
Exploratory Data Analysis - Stroke

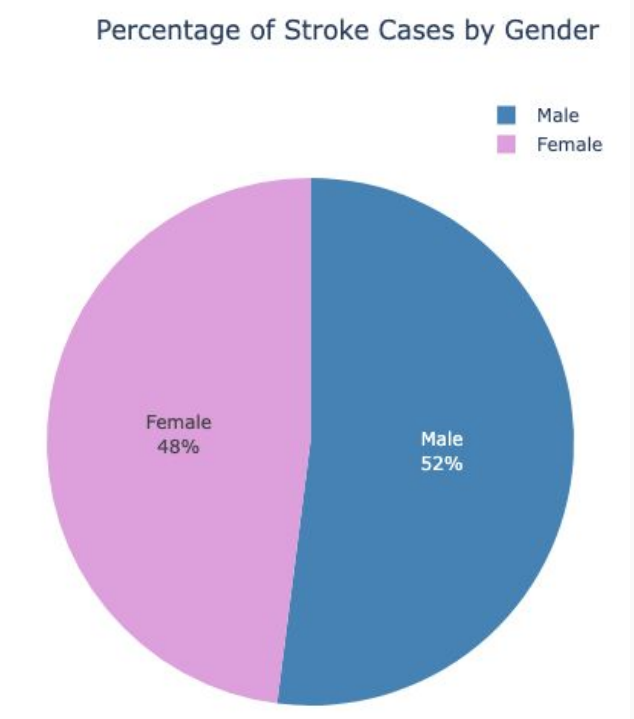


Exploratory Data Analysis

- **We performed hypothesis testing for each variable with stroke status, to determine whether they are significantly associated.**
- **Chi-Square test was used for categorical variables.**
- **T-test was used for continuous variables.**

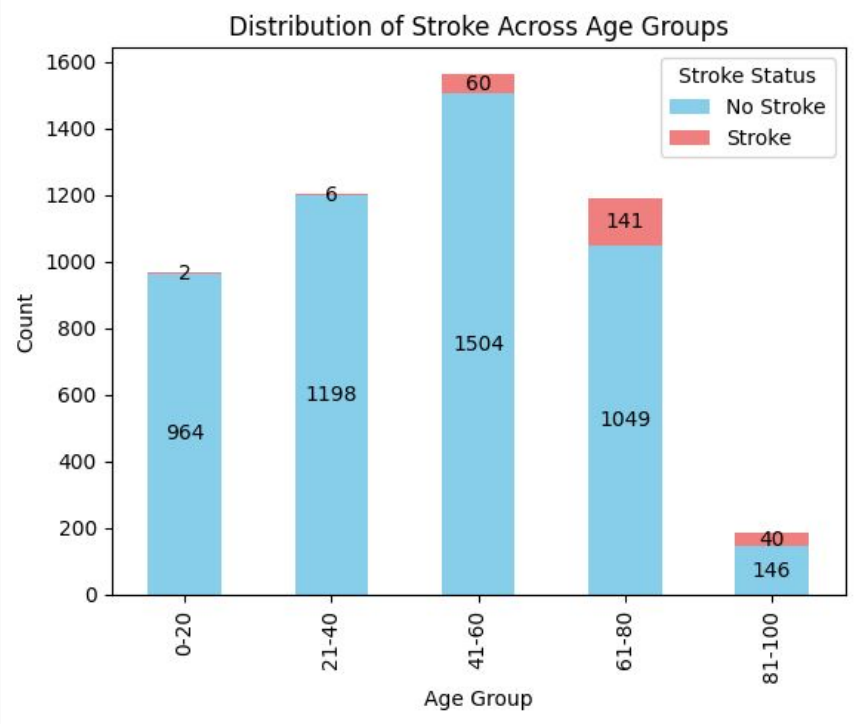
Exploratory Data Analysis - Gender





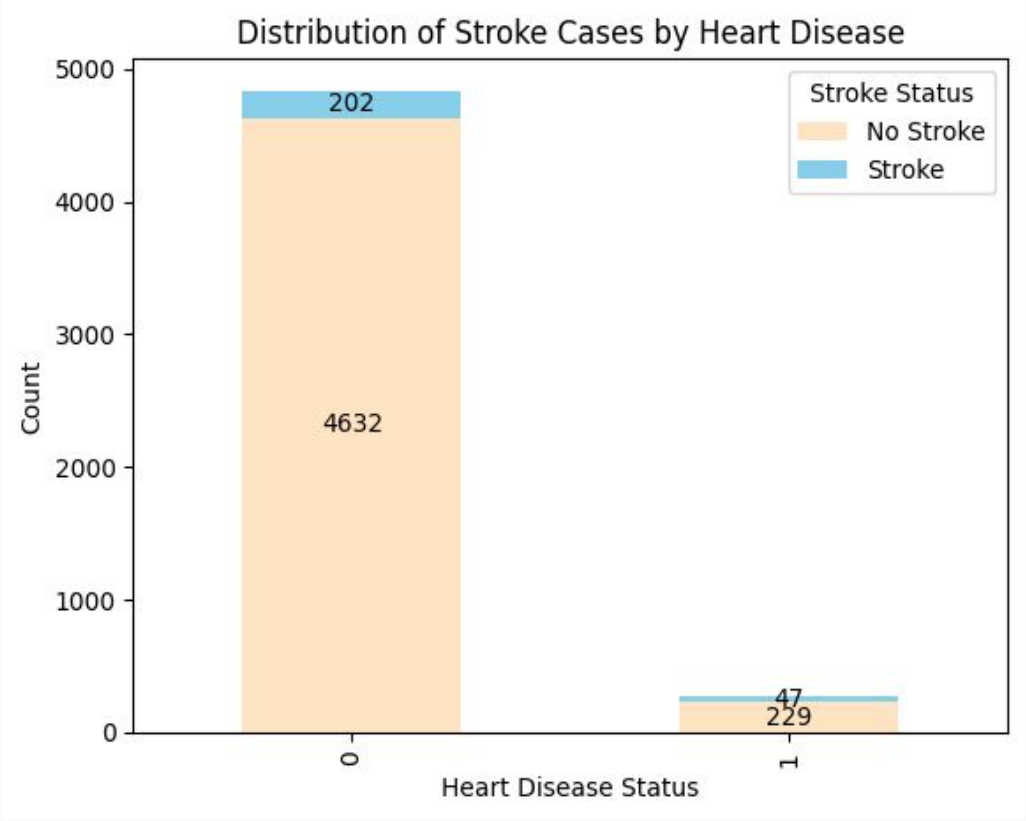
P-value: 0.78955
Failed to reject the Null hypothesis

Exploratory Data Analysis - Age



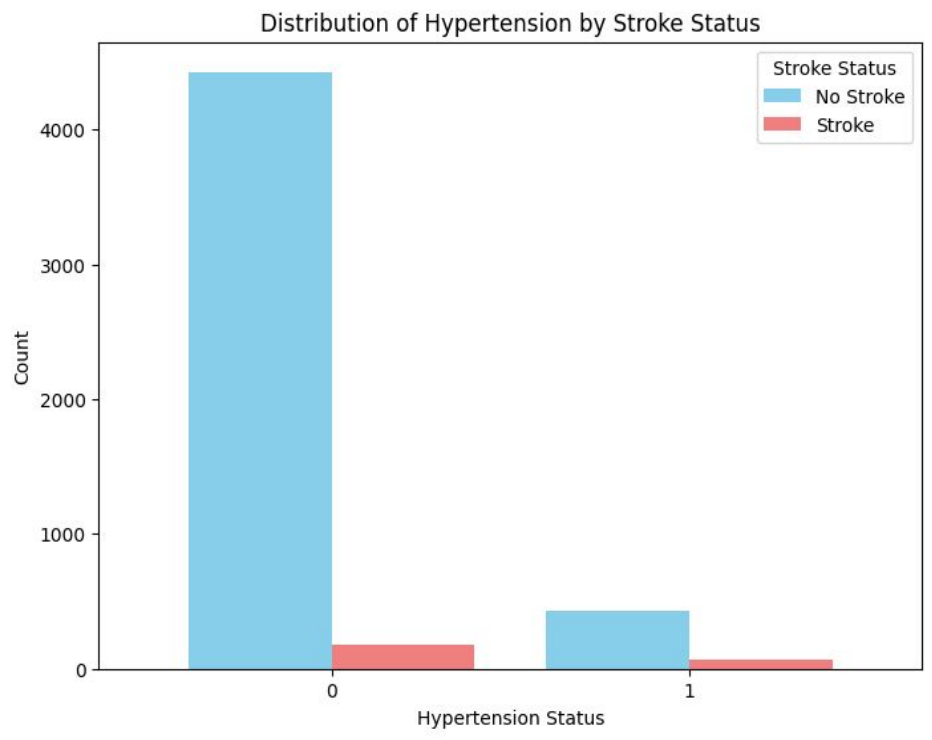
P-value: 2.1156848483472 72e-95
Reject the Null hypothesis

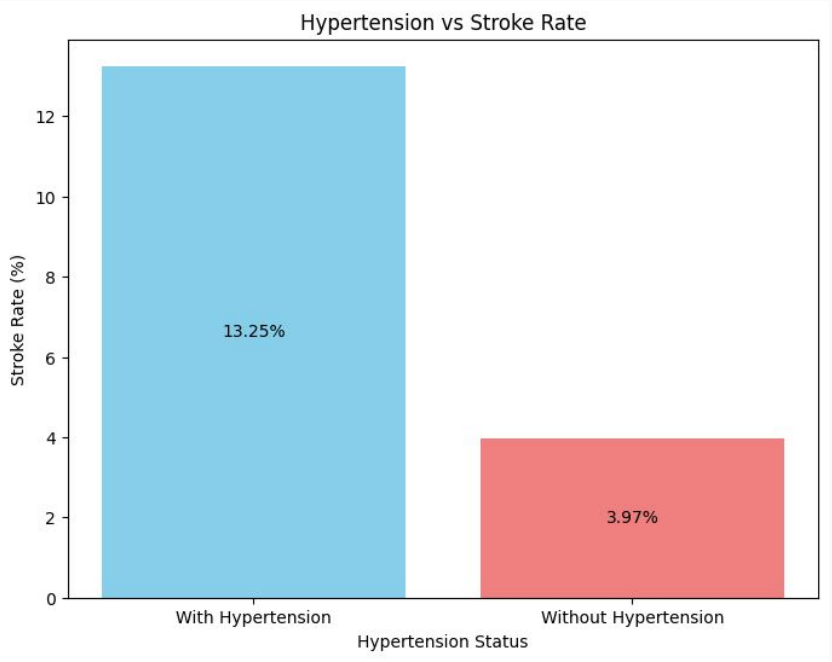
Exploratory Data Analysis - Heart disease



P-value: 2.0887845685229236e-21
Reject the Null hypothesis

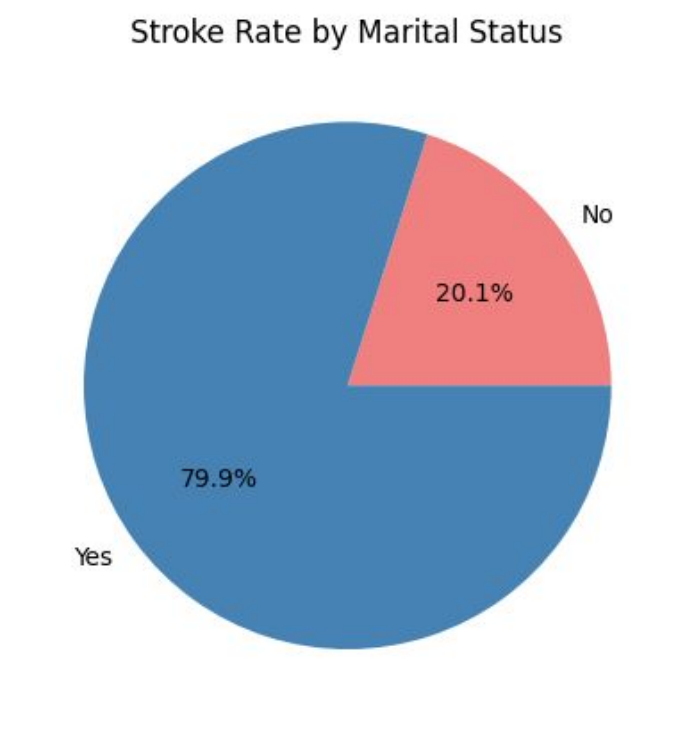
Exploratory Data Analysis - Hypertension





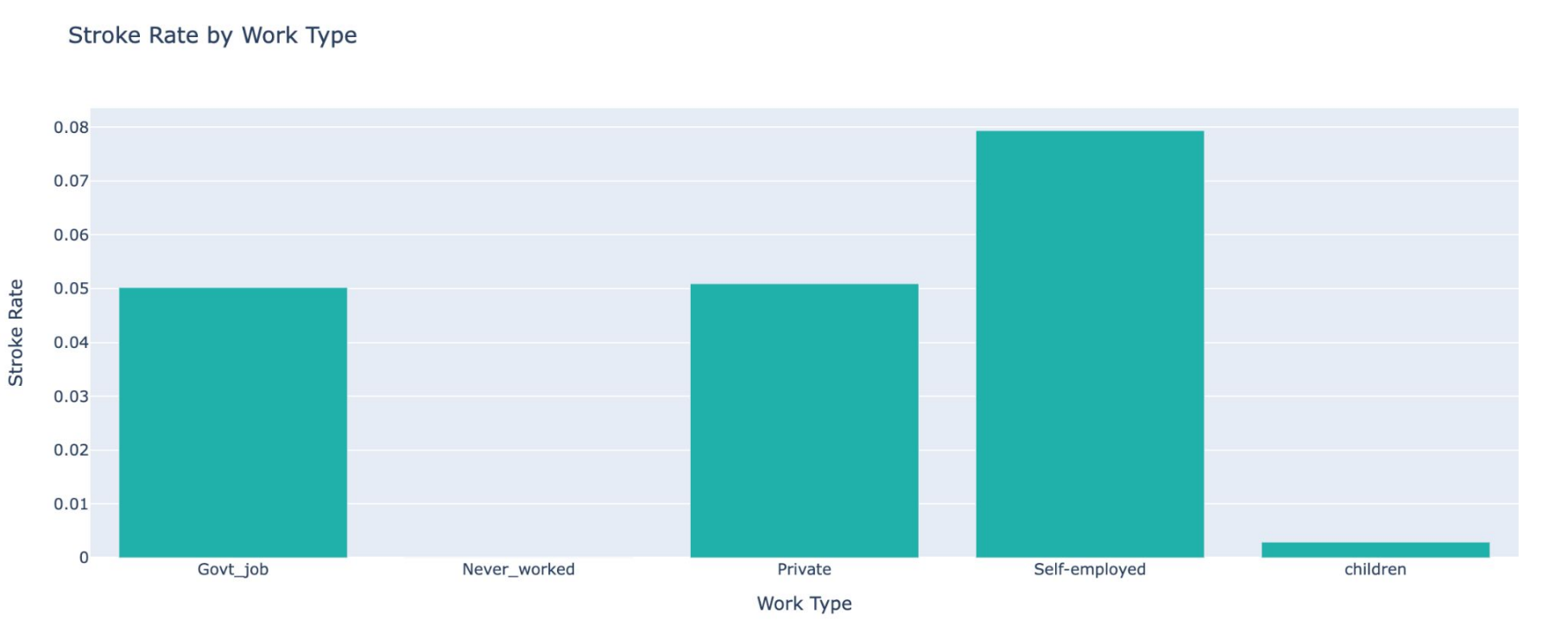
P-value: 1.6616219015118 23e-19
Reject the Null hypothesis.

Exploratory Data Analysis - Marital status



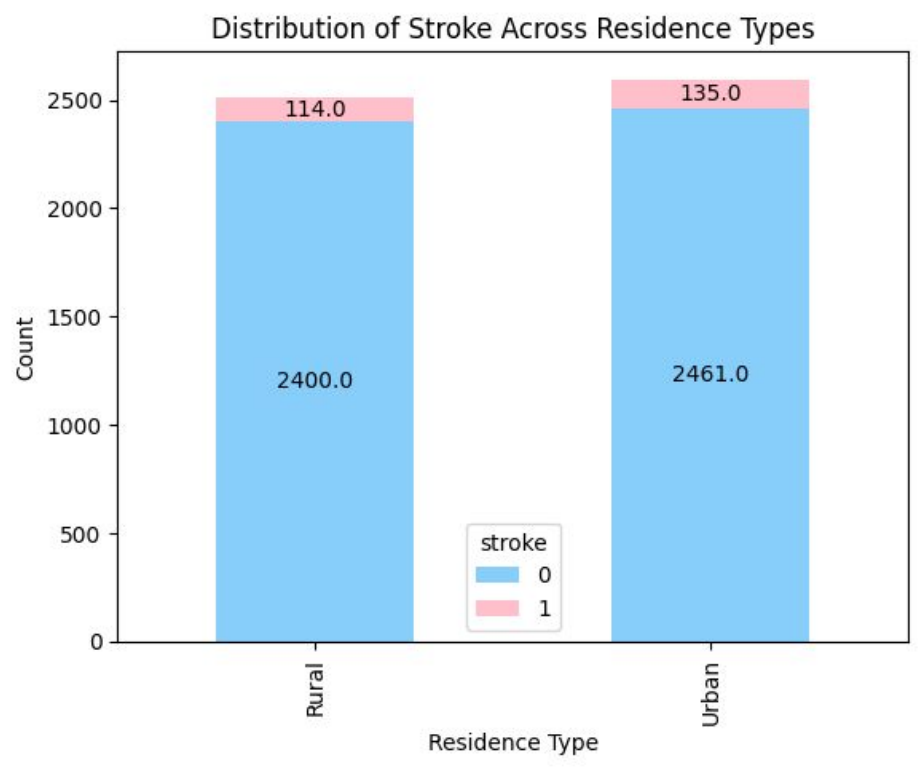
P-Value: 1.6389021142314745e-14
Reject the Null hypothesis

Exploratory Data Analysis - Work Type



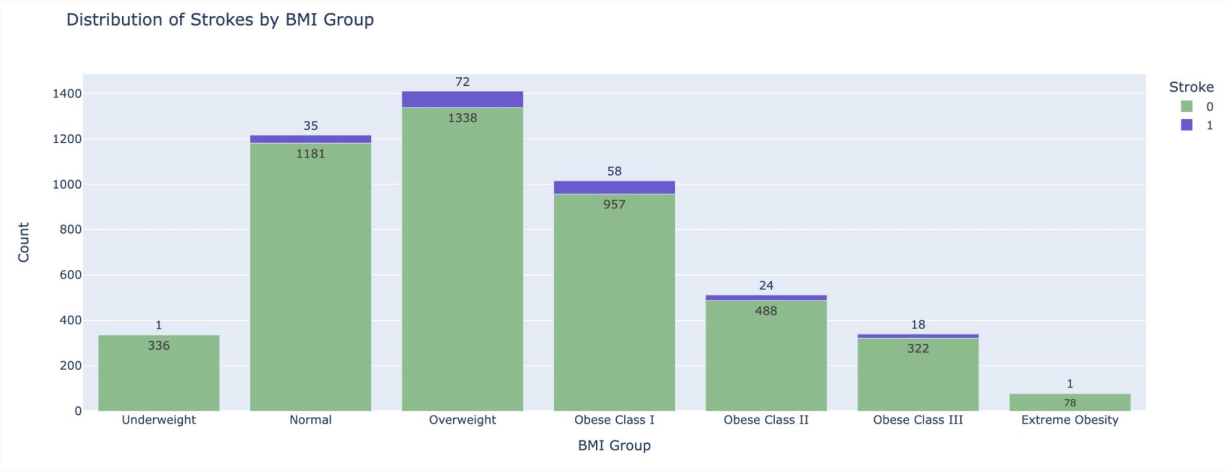
P-Value: 5.3977078018961l9e-10
Reject the Null hypothesis

Exploratory Data Analysis - Residence Type

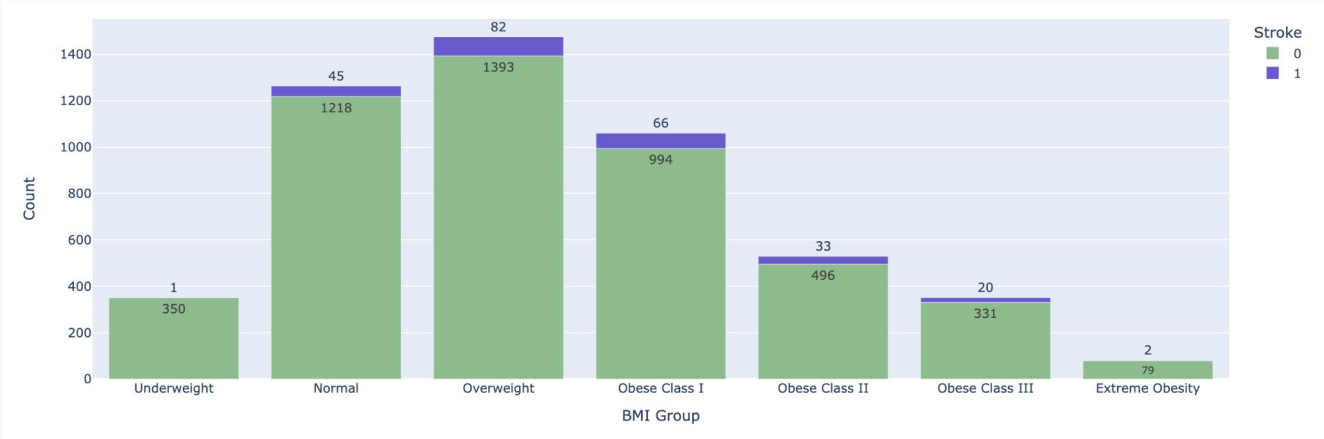


P-value: 0.29833169286876987
Fail to reject the Null hypothesis

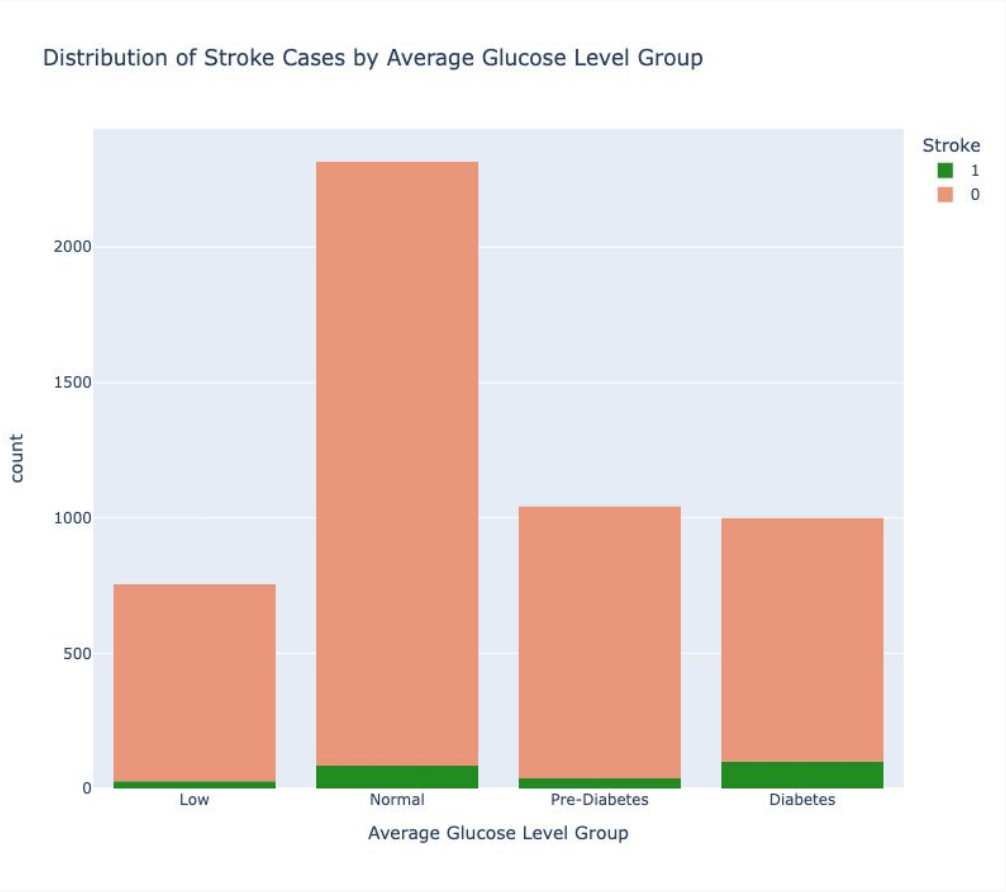
Exploratory Data Analysis - BMI



P-Value: 8.30260394684319e-05
Reject the Null hypothesis

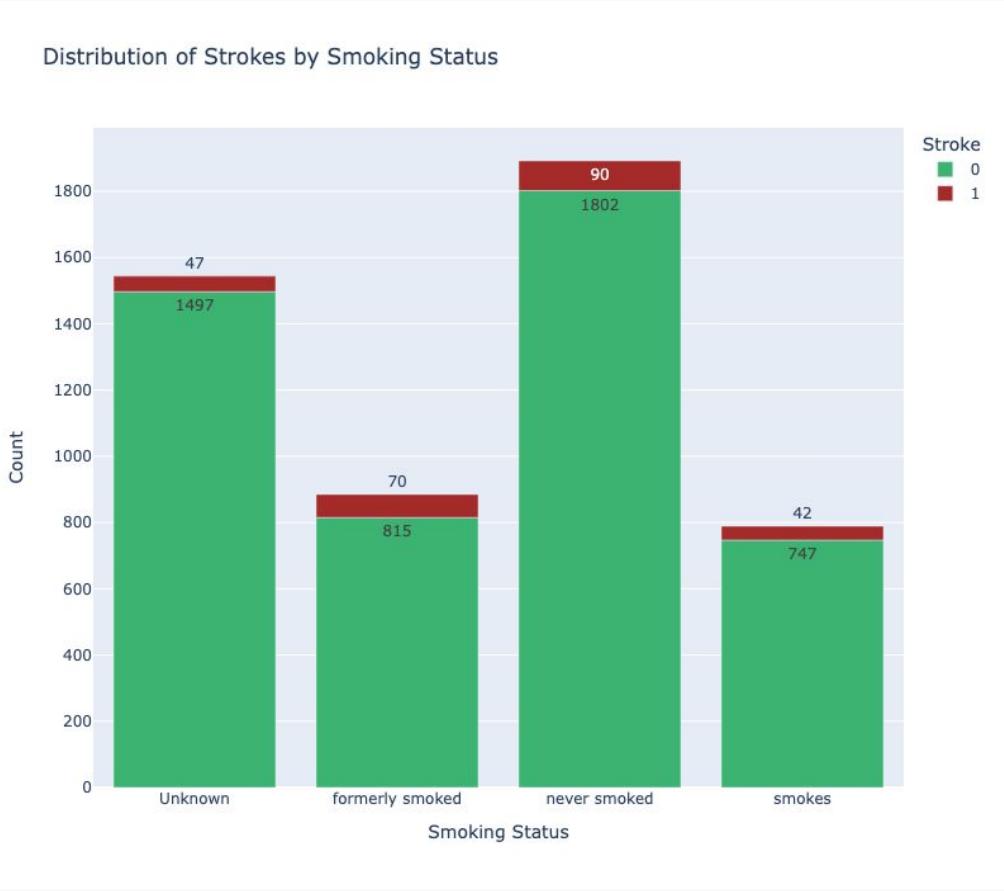


Exploratory Data Analysis - Average Glucose Level



P-Value: 2.4014366563697676e-11
Reject the null hypothesis.

Exploratory Data Analysis - Smoking Status



P-Value: 2.0853997025008455e-06
Reject the Null hypothesis

Data Pre-Processing

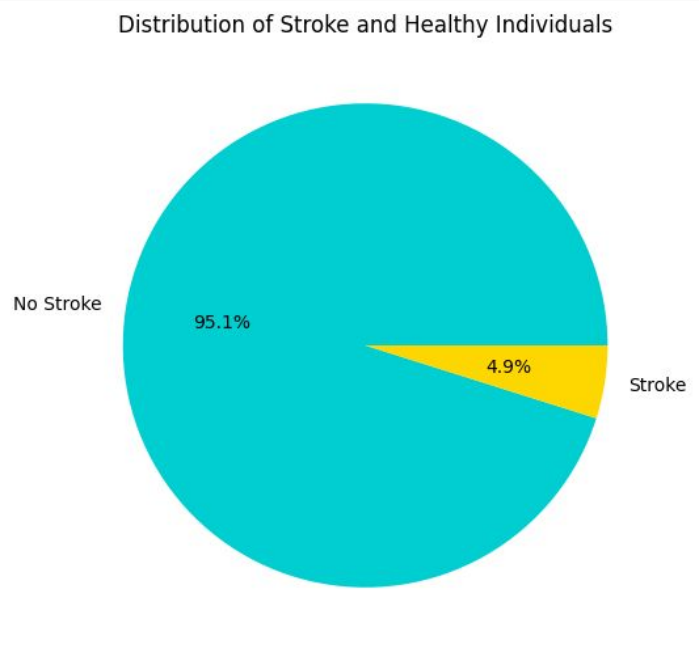
- The ‘gender’ and ‘residence_type’ variables were excluded from analysis.
- Since the ‘ever_married’, ‘work_type’ and ‘smoking status’ are categorical variables, we need to encode them numerically.
- I used label encoding for ‘ever_married’ and one-hot encoding for the other two.

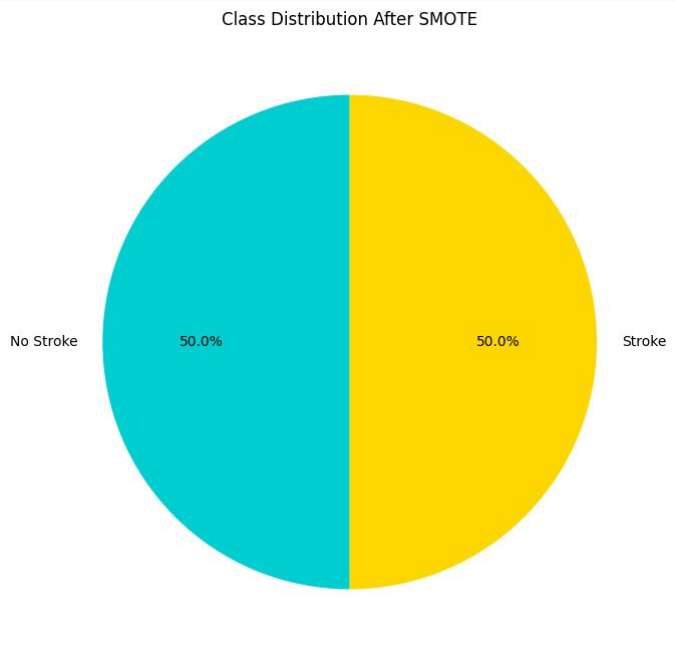
	age	hypertension	heart_disease	ever_married	avg_glucose_level	bmi	stroke	work_type_Never_worked	work_type_Private	work_type_Self-employed	work_type_children
0	67.0	0	1	1	228.69	36.6	1	0	1	0	0
1	61.0	0	0	1	202.21	28.8	1	0	0	1	0
2	80.0	0	1	1	105.92	32.5	1	0	1	0	0
3	49.0	0	0	1	171.23	34.4	1	0	1	0	0
4	79.0	1	0	1	174.12	24.0	1	0	0	1	0

smoking_status_formerly smoked	smoking_status_never smoked	smoking_status_smokes
1	0	0
0	1	0
0	1	0
0	0	1
0	1	0

Data Pre-Processing

- Since the data was highly imbalanced, we used SMOTE oversampling technique.
- First the data is split into training and testing, then SMOTE was used on the training data.





Data Modeling and Results

- To create an ML model that can predict the stroke status, we will use a logistic regression.
- The data was split in 8:2 ratio.
- We have already resampled the training data.

Accuracy: 0.7465753424657534

Classification Report:

	precision	recall	f1-score	support
0	0.98	0.74	0.85	960
1	0.17	0.81	0.28	62
accuracy			0.75	1022
macro avg	0.58	0.77	0.56	1022
weighted avg	0.93	0.75	0.81	1022

Future Work

- Spend time on feature engineering to discover new predictors or interactions that might enhance the model's predictive power.
- Experiment with advanced machine learning models beyond logistic regression.
- Additional datasets and other variables.
- Investigate advanced techniques for handling imbalanced data, beyond SMOTE, to further enhance the model's ability to generalize across classes.

References

- https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset
- https://www.stat.cmu.edu/capstoneresearch/spring2021/ 315files/team16.html
- Sailasya, G., & Kumari, G. L. A. (2021). Analyzing the performance of stroke prediction using ML classification algorithms. International Journal of Advanced Computer Science and Applications, 12(6).
- https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html