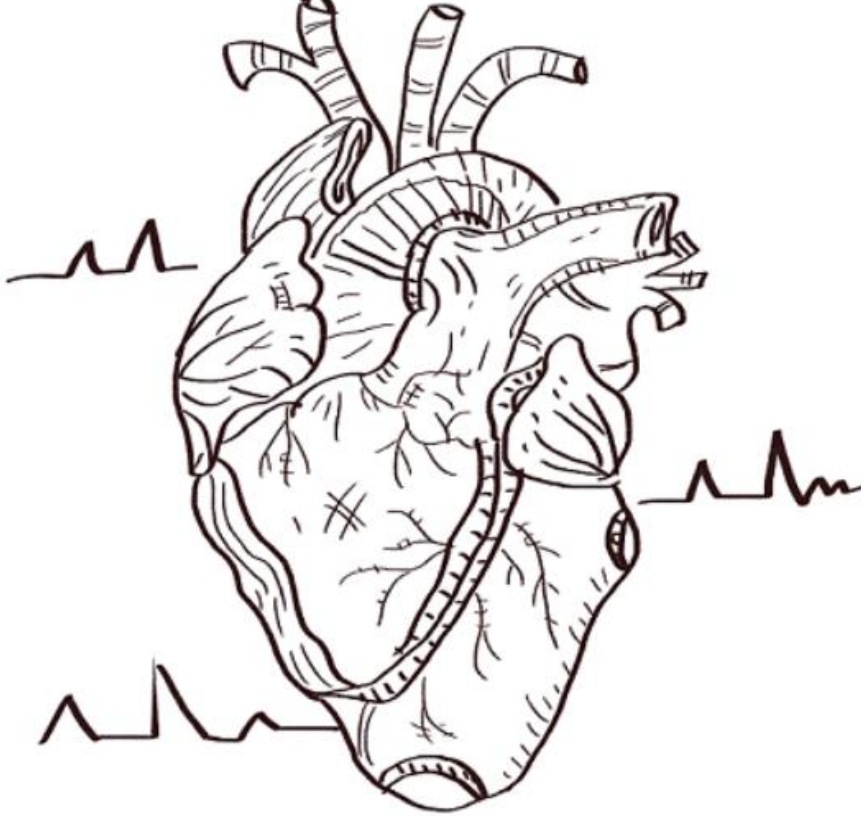
Heart Disease Prediction



GROUP-2

- PRIYANKA AKULA
- SOWJANYA PAMULAPATI
- SOMNA SATTOOR
- SHREYA CHIKATMARLA

Contents

Introduction

Heart disease is a significant health concern worldwide, and early detection plays a crucial role in improving patient outcomes.

In this project, we aim to utilize the power of visualizations and also develop a predictive model that can identify individuals at risk of developing heart disease.

By combining data analysis techniques with effective visual representations, we can enhance our understanding of the factors contributing to heart disease and provide valuable insights for healthcare professionals and patients alike.

Objectives

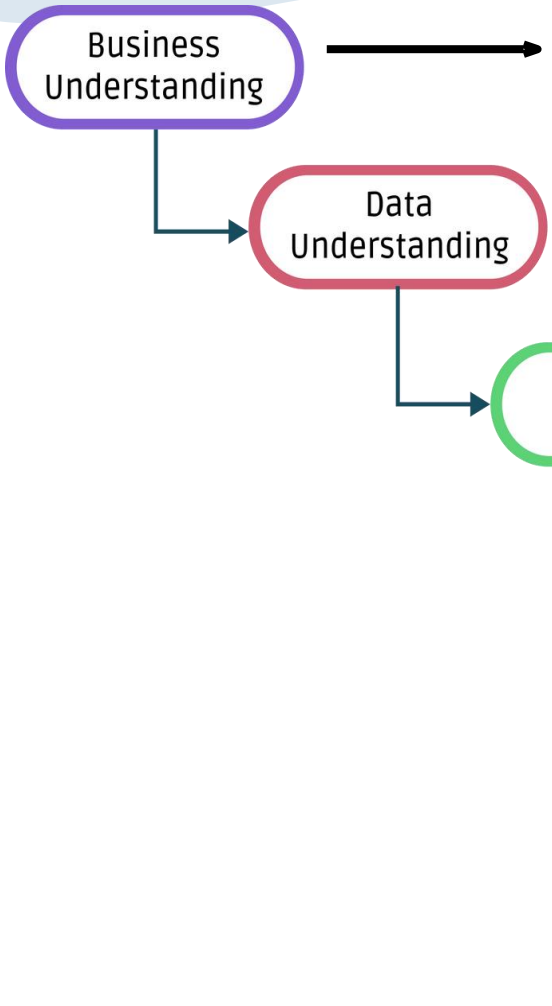

Investigating the relationship between specific attributes(such as sex, age, mental health, physical health)and the risk of heart disease.
Examining the impact of doing physical activity on heart disease risk.
Develop a machine learning model that can predict the likelihood of a heart attack based on different attributes from dataset.

Motivations

The motivation is using visualizations to better understand heart disease and create a predictive algorithm to detect at-risk individuals.

This can enhance heart disease prevention and management, improving patient outcomes and healthcare expenditures.

Hybrid Crisp-DM & Waterfall Model



Establish goals through team meetings and brainstorming for heart disease dataset analysis.
Evaluate research status and identify knowledge gaps through literature surveys.
Develop a data mining plan including dataset preparation, trend exploration, algorithm implementation, outcome assessment, and documentation.

Gather information for a historical survey.
Conduct exploratory data analysis.
Perform data quality checks.

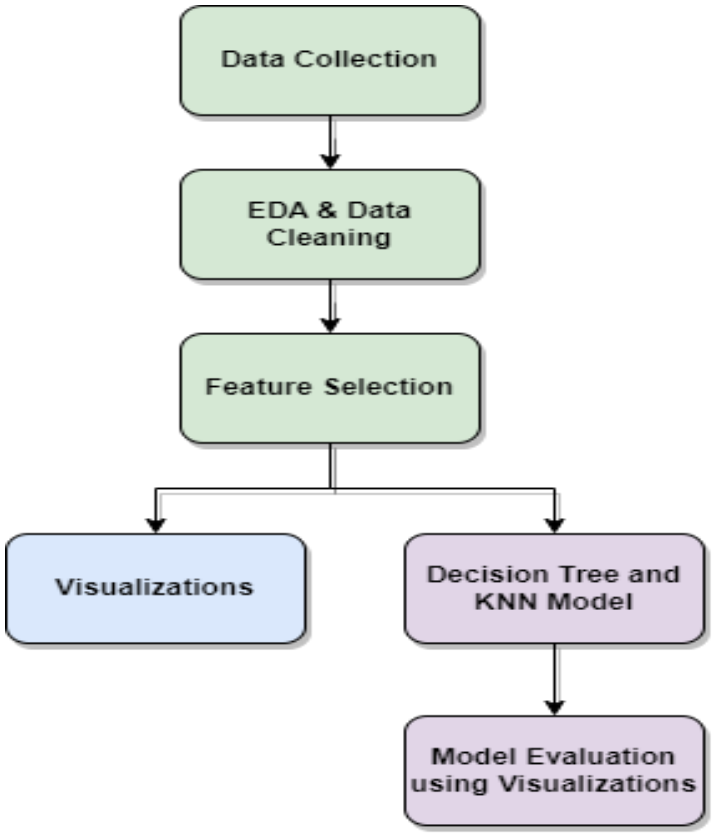
Data pre-processing
Feature selection for decision tree and KNN models.

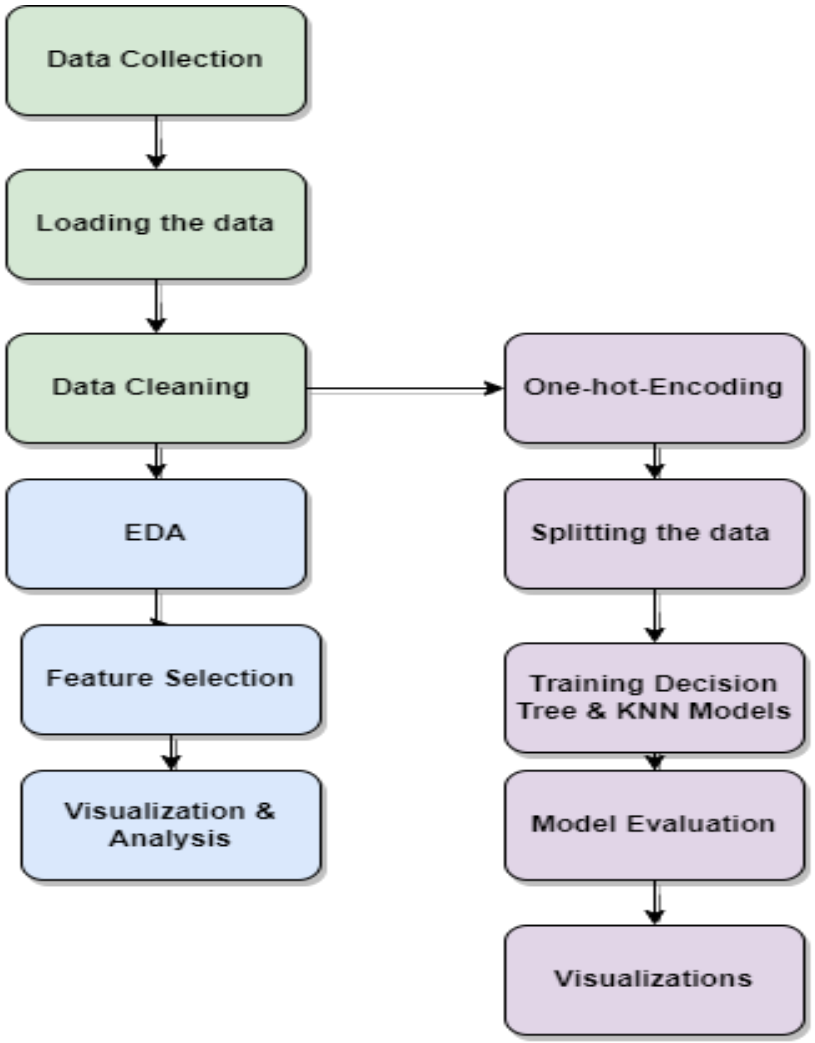
Split dataset set for training and testing
Create and visualize the model

Evaluate the model and visualize the results

Deploying all the files in git hub repository

Project Workflow





Dataset Description

The Heart Disease Prediction dataset was taken from Kaggle. It includes information from a study conducted on patients with suspected heart disease.

The dataset contains 18 columns of patient data, including age, BMI, blood pressure, cholesterol levels, heart disease, diabetic and other pertinent health indicators.

There are 319796 rows in the dataset, each of which represents a different patient. The target variable is heart disease which represents weather a patient having heart disease or not.

```
df1.describe().T
```

	count	mean	std	min	25%	50%	75%	max
Age	270.0	54.433333	9.109067	29.0	48.0	55.0	61.0	77.0
Sex	270.0	0.677778	0.468195	0.0	0.0	1.0	1.0	1.0
Chest pain type	270.0	3.174074	0.950090	1.0	3.0	3.0	4.0	4.0
BP	270.0	131.344444	17.861608	94.0	120.0	130.0	140.0	200.0
Cholesterol	270.0	249.659259	51.686237	126.0	213.0	245.0	280.0	564.0
FBS over 120	270.0	0.148148	0.355906	0.0	0.0	0.0	0.0	1.0
EKG results	270.0	1.022222	0.997891	0.0	0.0	2.0	2.0	2.0
Max HR	270.0	149.677778	23.165717	71.0	133.0	153.5	166.0	202.0
Exercise angina	270.0	0.329630	0.470952	0.0	0.0	0.0	1.0	1.0
ST depression	270.0	1.050000	1.145210	0.0	0.0	0.8	1.6	6.2
Slope of ST	270.0	1.585185	0.614390	1.0	1.0	2.0	2.0	3.0
Number of vessels fluro	270.0	0.670370	0.943896	0.0	0.0	0.0	1.0	3.0
Thallium	270.0	4.696296	1.940659	3.0	3.0	3.0	7.0	7.0

Data Cleaning

- Checked null values, duplicate entries, outliers and converted the categorical data into numeric and removed the unnecessary data.

```
df[['Diabetic','Sex']].T
```

	0	1	2	3	4	5	6	7	8	9	...	319785	319786	319787	319788	319789	319790	319791	319792	3
Diabetic	Yes	No	Yes	No	No	No	No	Yes	No, borderline diabetes	No	...	No	Yes	No	No	No	Yes	No	No	
Sex	Female	Female	Male	Female	Female	Female	Female	Female	Female	Male	...	Male	Female	Male	Female	Female	Male	Male	Female	F

2 rows × 319795 columns

```
df =  df[df.columns].replace({'Yes':1, 'No':0, 'Male':1,'Female':0,'No, borderline diabetes':'0','Yes (during pregnancy)':'1' })
df['Diabetic'] = df['Diabetic'].astype(int)
```

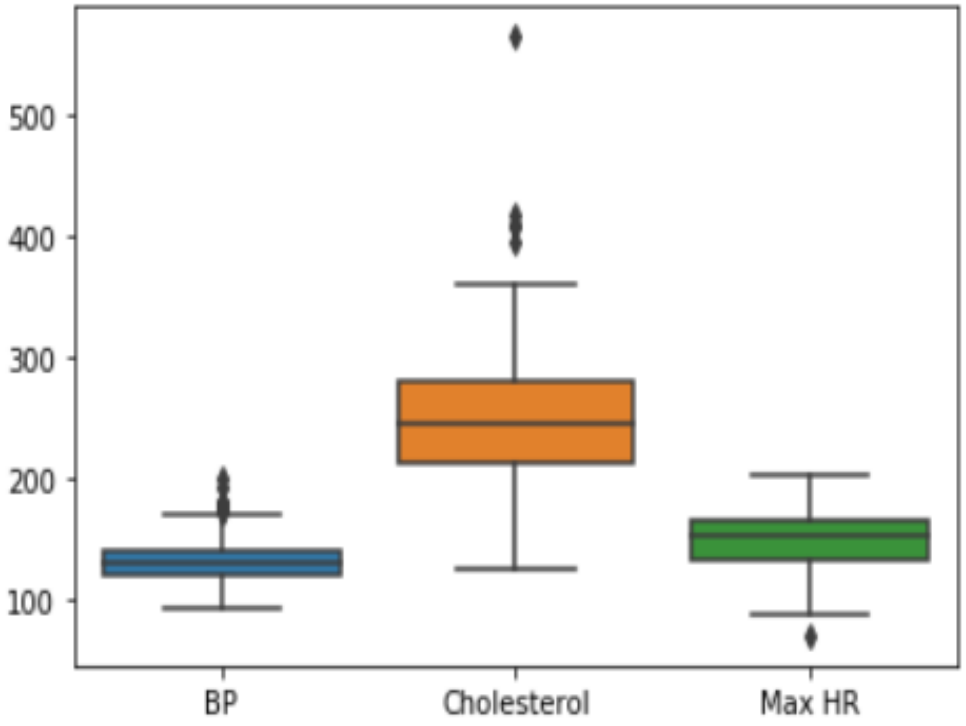
```
df[['Diabetic','Sex']].T
```

	0	1	2	3	4	5	6	7	8	9	...	319785	319786	319787	319788	319789	319790	319791	319792	319793	319794
Diabetic	1	0	1	0	0	0	0	1	0	0	...	0	1	0	0	0	1	0	0	0	0
Sex	0	0	1	0	0	0	0	0	0	1	...	1	0	1	0	0	1	1	0	0	0

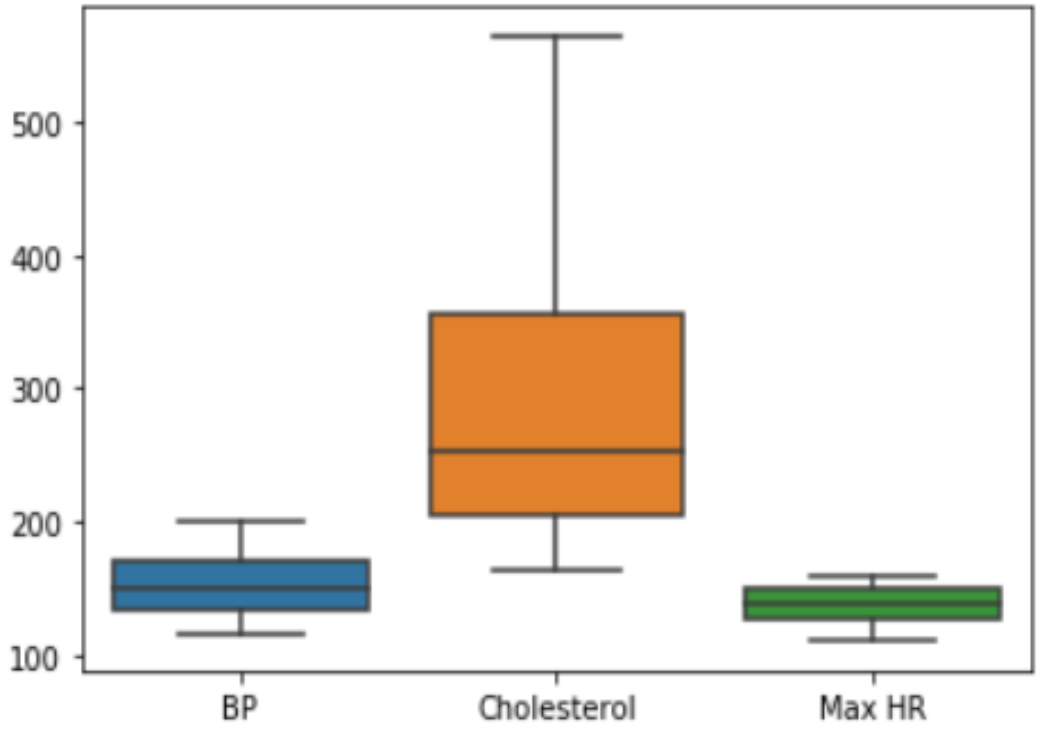
2 rows × 319795 columns

Data Cleaning

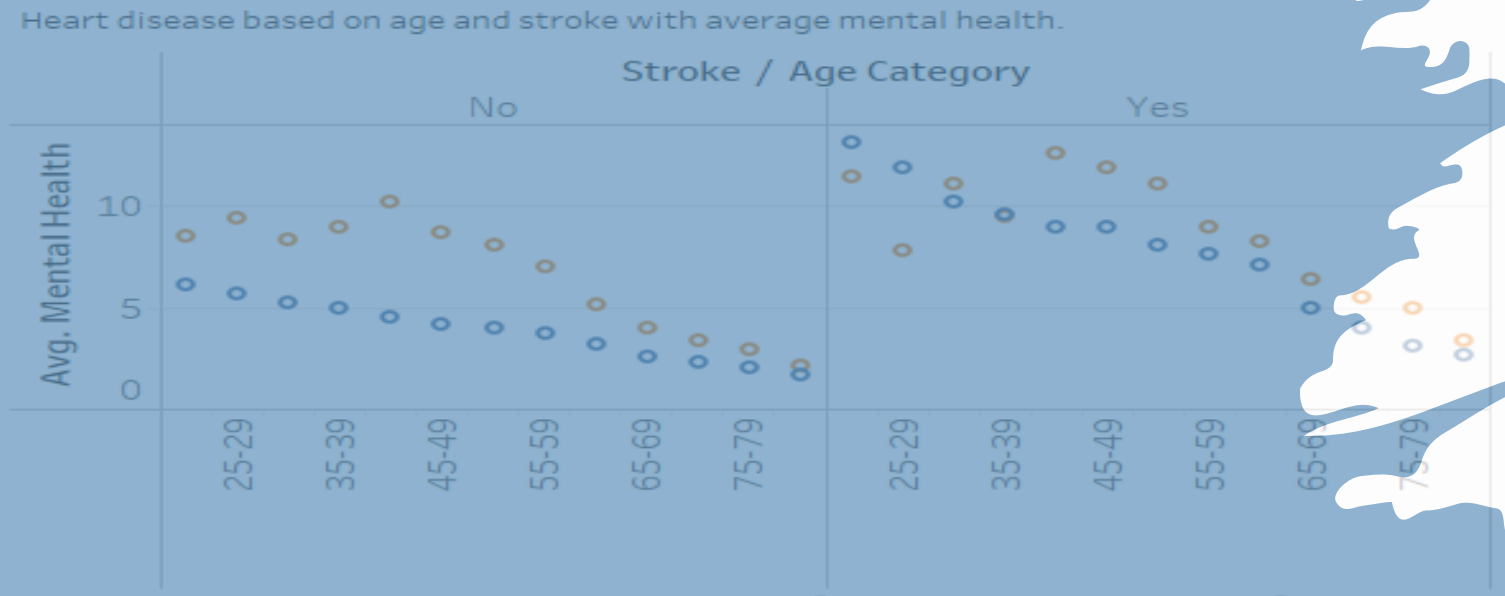
Box plot with outliers

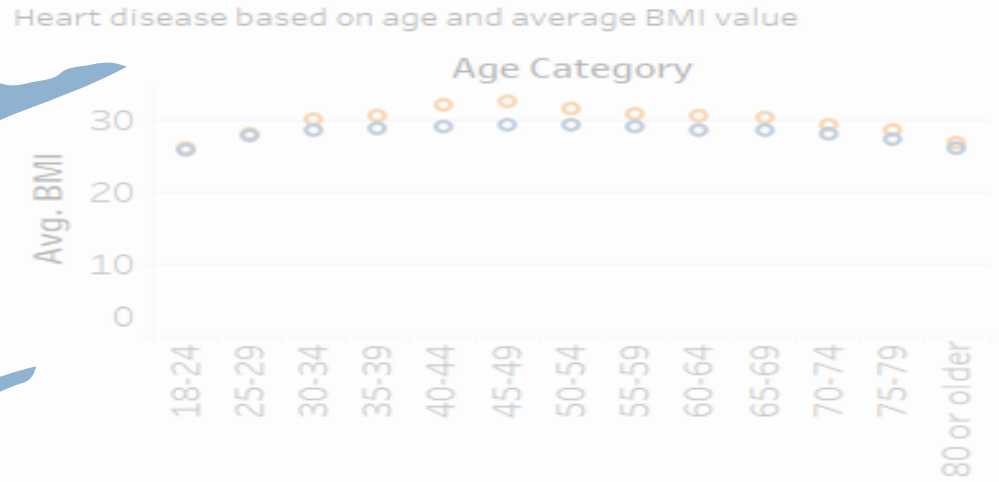


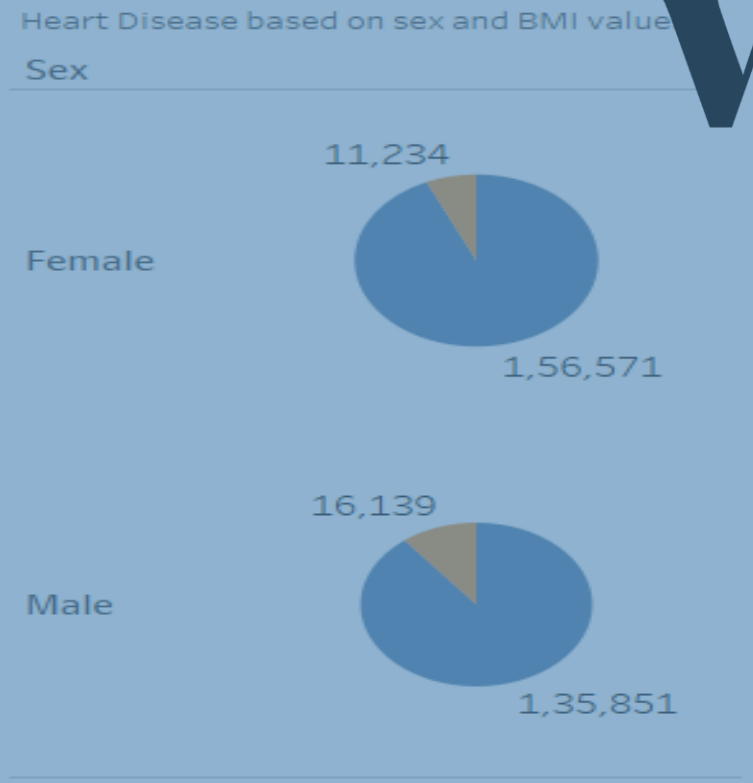
Box plot after removing outliers

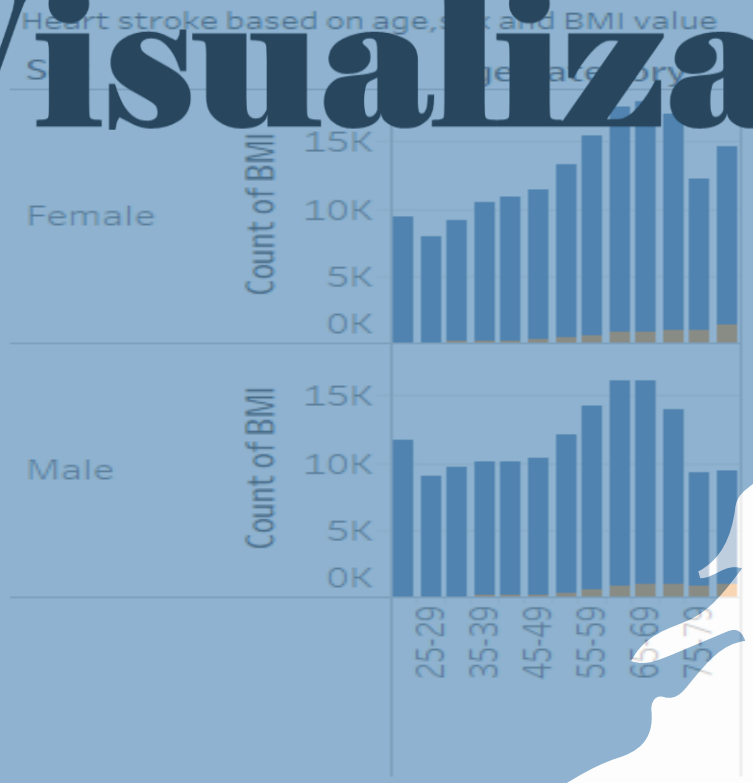


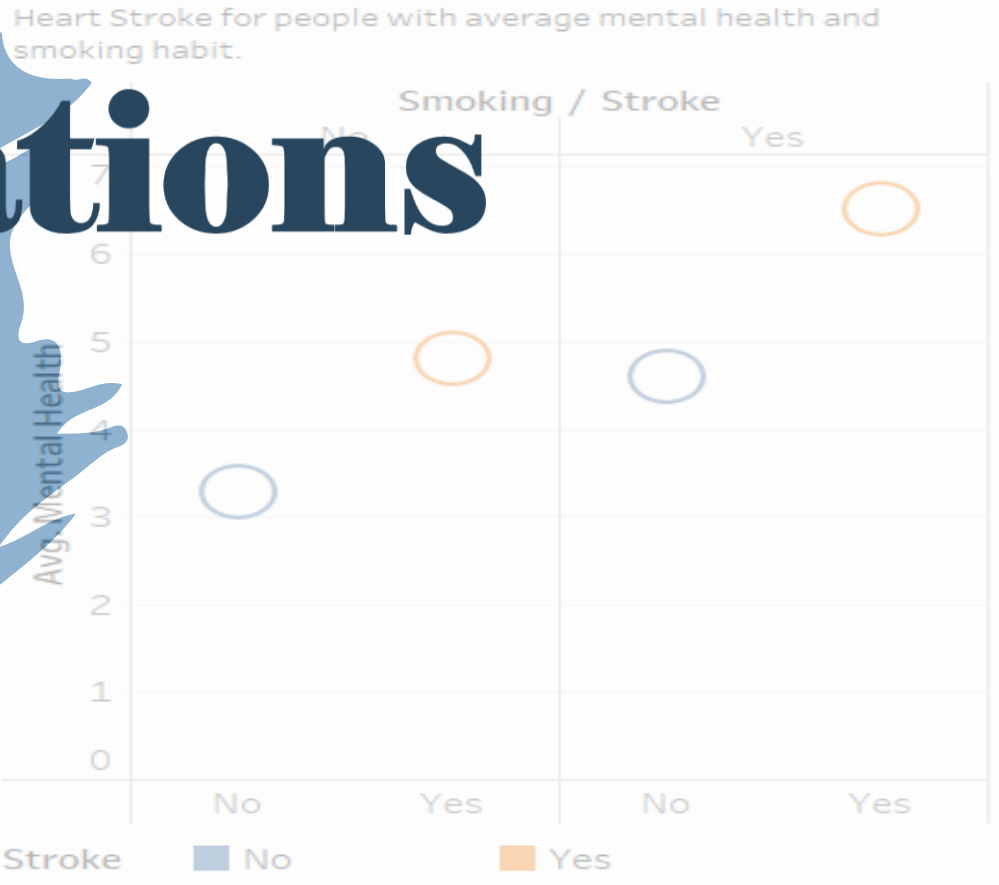
Visualizations

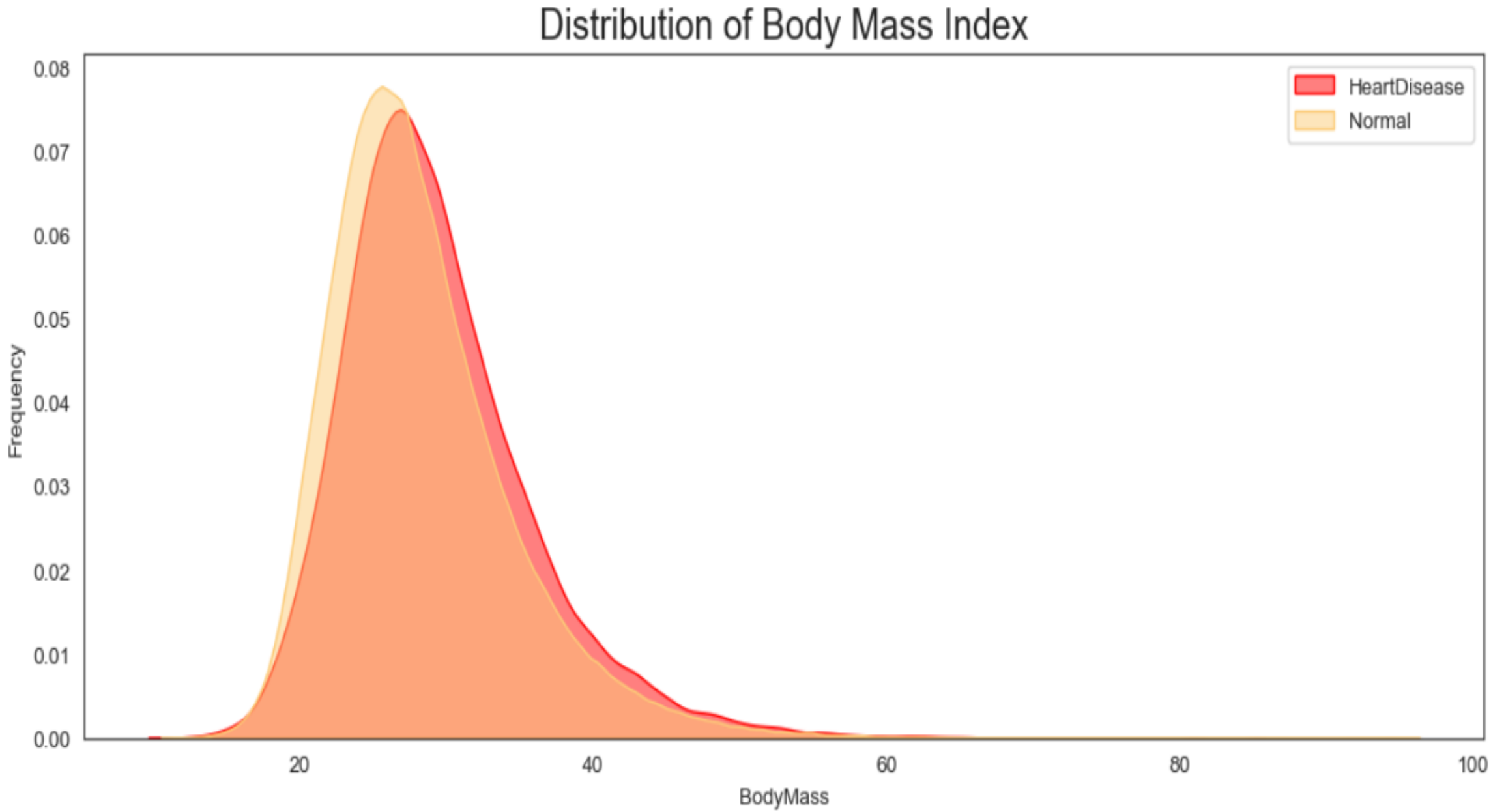


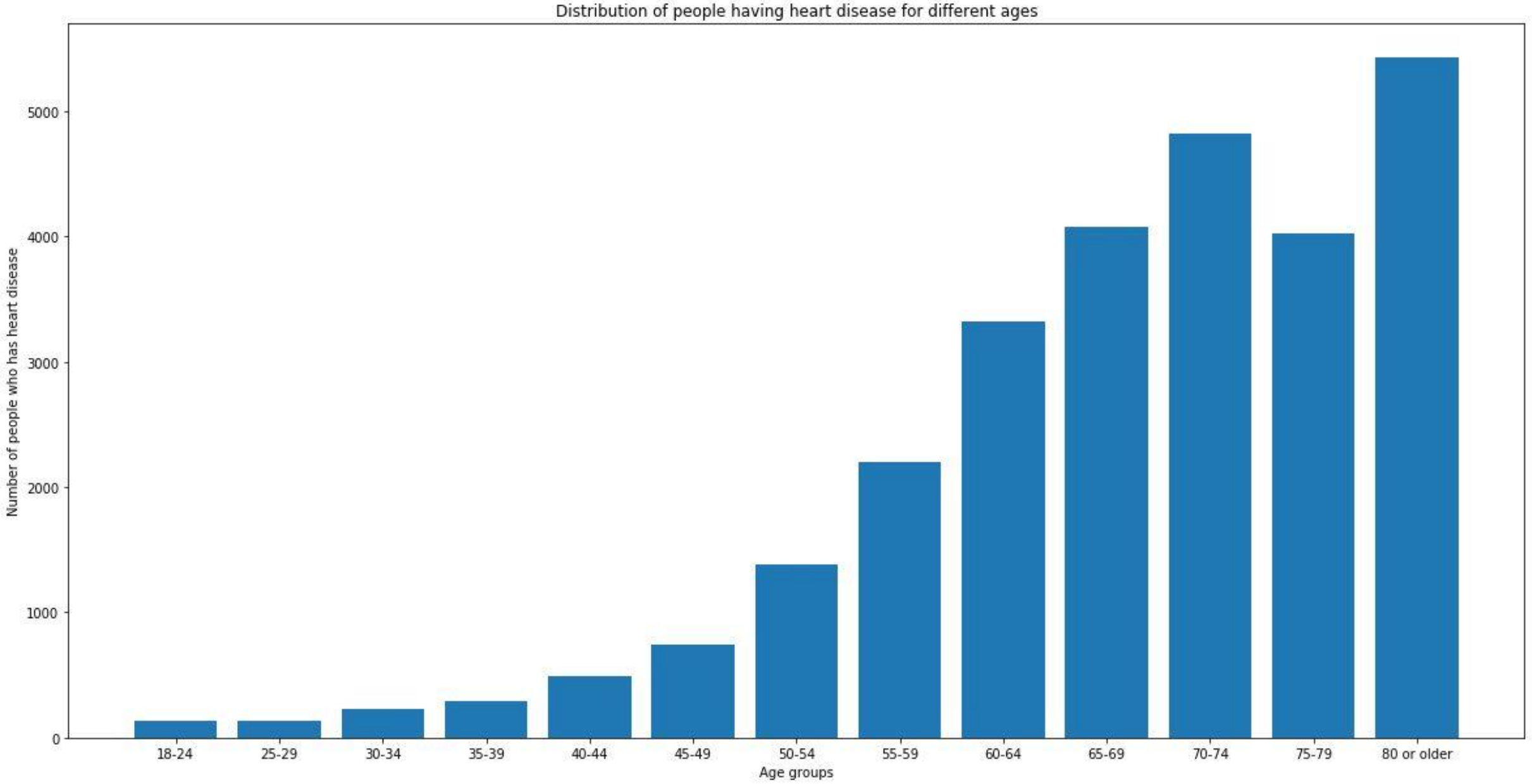




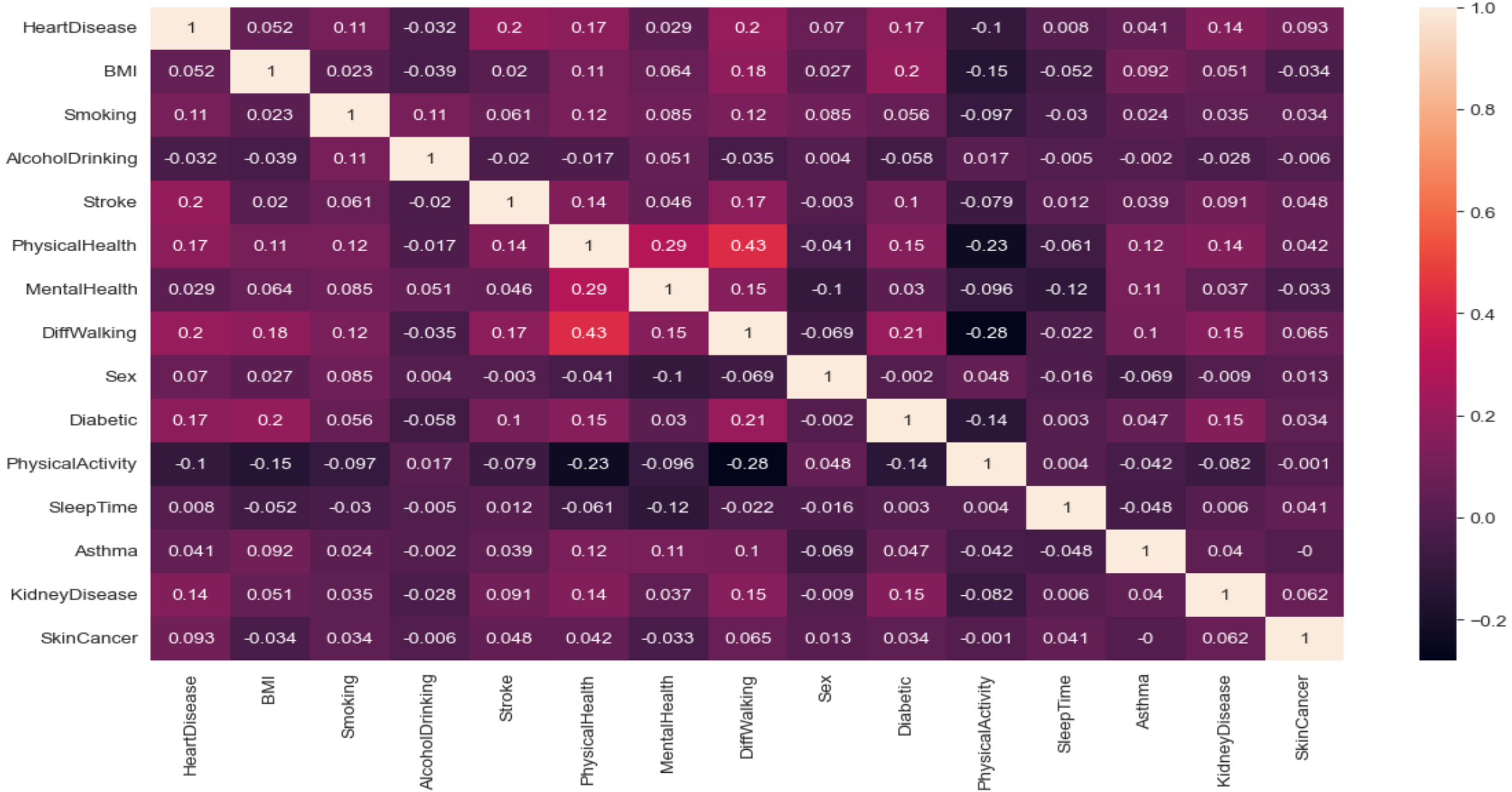




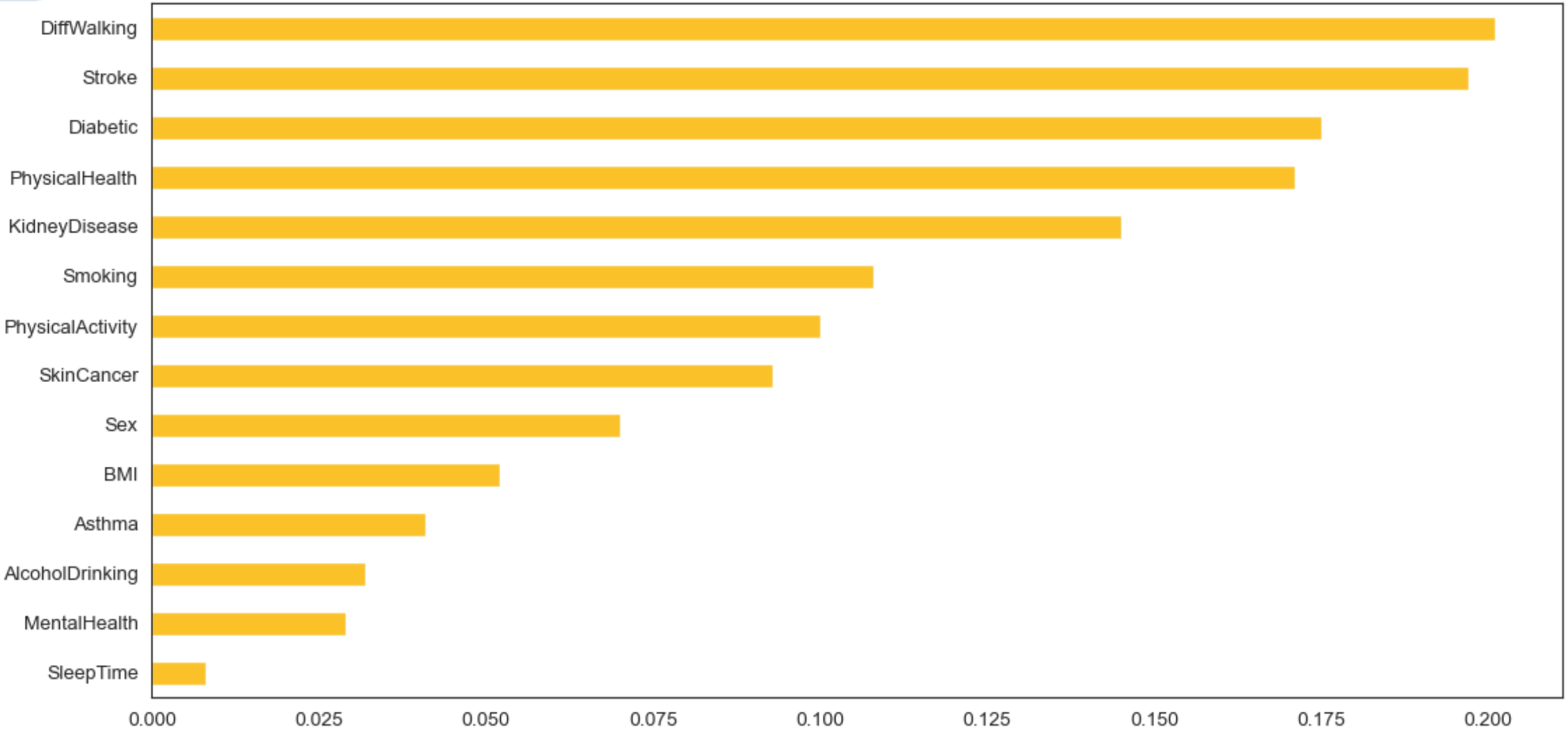

Distribution of Body Mass Index
HeartDisease
Normal
Frequency
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
20
40
60
80
100
BodyMass


Distribution of people having heart disease for different ages
Number of people who has heart disease
0
1000
2000
3000
4000
5000
18-24
25-29
30-34
35-39
40-44
45-49
50-54
55-59
60-64
65-69
70-74
75-79
80 or older
Age groups

Correlation between all set of features



Distribution of correlation between the features



Algorithms

The data is spilt into train and test sets with a ratio of 70:30.
Train set is used for both models while test set is used for calculating evaluation metrics.

Decision Tree

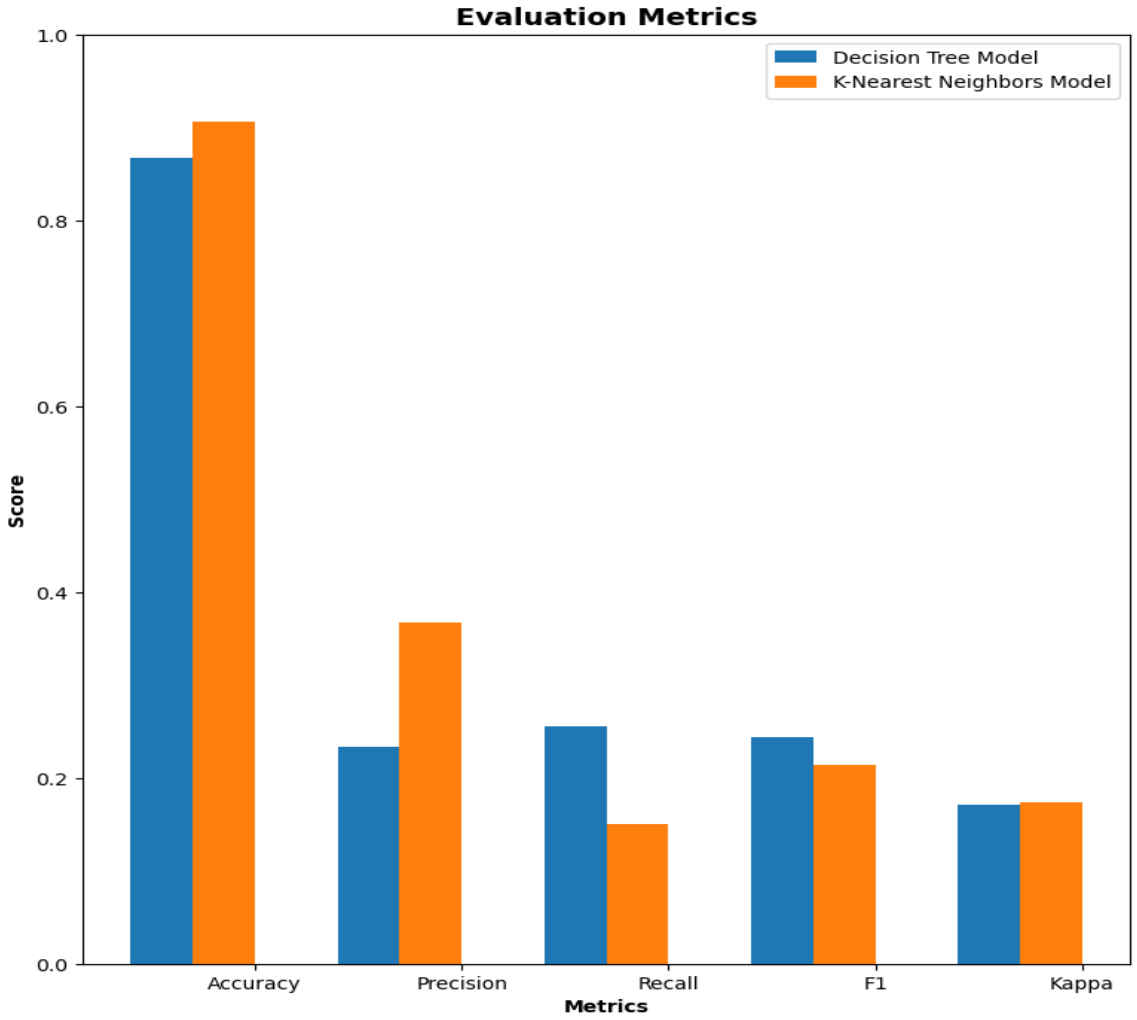
- Since the data set is having both categorical and numerical values where the categorical columns are containing less unique values like yes, no and female, male, we are training decision tree classification model to classify data into yes or no.
- F1 Score = 0.244
- Accuracy = 0.86

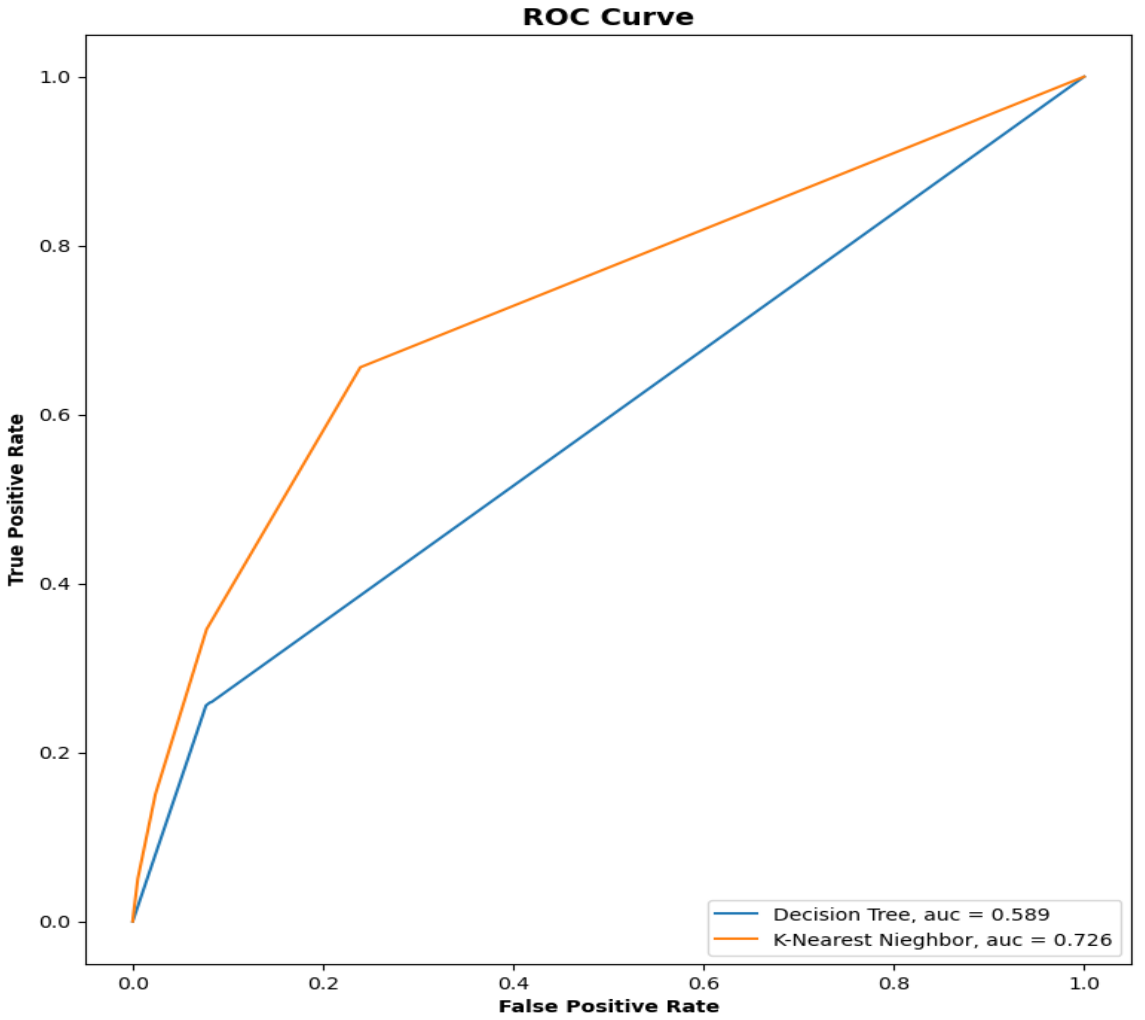
K- nearest Neighbors Model

- Another classification model that we taught would work best for this data is KNN since there is good correlation between different features and heart diseases.
- F1 Score = 0.213
- Accuracy = 0.90

Model Evaluation

Comparing the models





Conclusion

In conclusion, the heart disease study project yielded substantial insights regarding the incidence, causes, and mitigation strategies linked with heart disease.

It is possible to discover essential factors and develop efficient strategies for timely diagnosis and prevention using data analysis and statistical modeling.

It is critical to ensure the quality of data in order to generate reliable results.

In general, this also improves comprehension and results in the treatment of heart disease.

References

- https://github.com/Priyankaakula/DATA230-PROJECT
- https://www.kaggle.com/code/andls555/heart-disease-prediction/notebook
- https://link.springer.com/article/10.1007/s42979-020-00365-y


Thank you