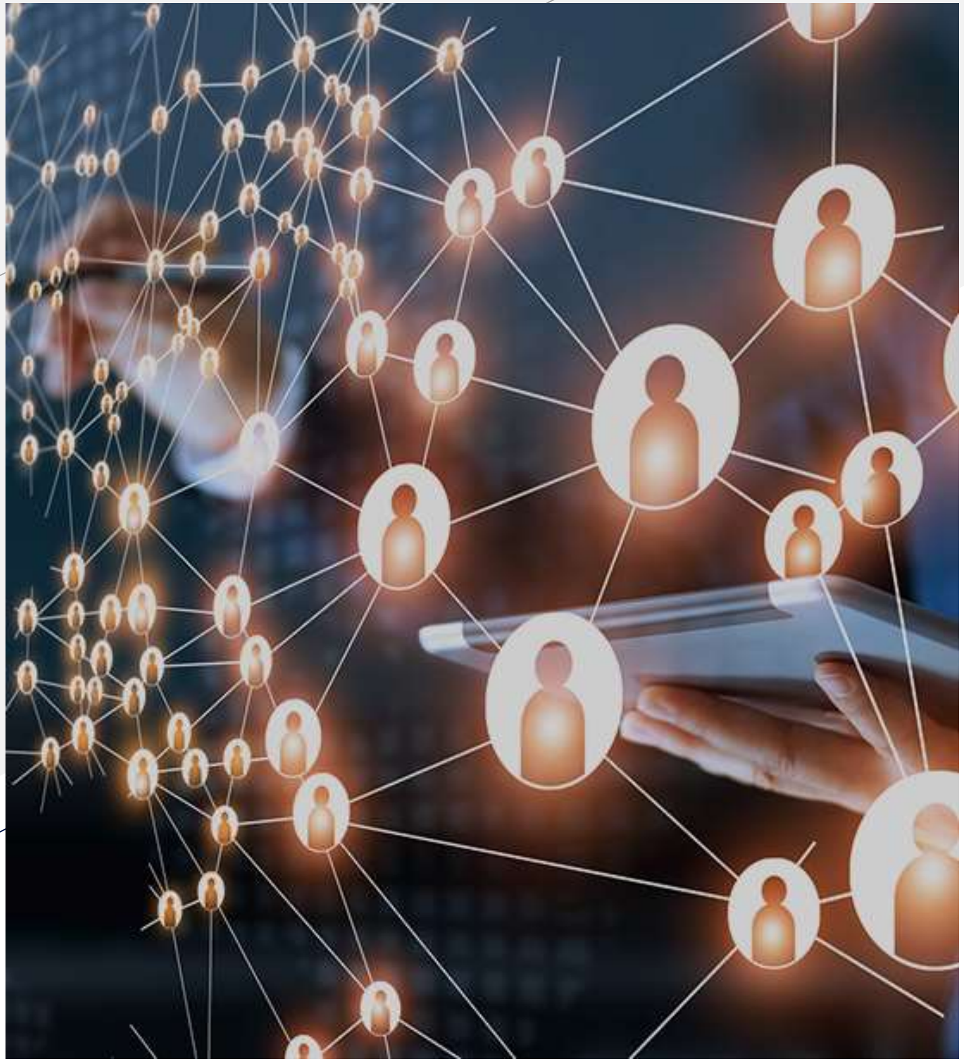


Heart Disease Analysis

Prepared by: Mrugandh Zodape

Group: 386

Introduction

Purpose

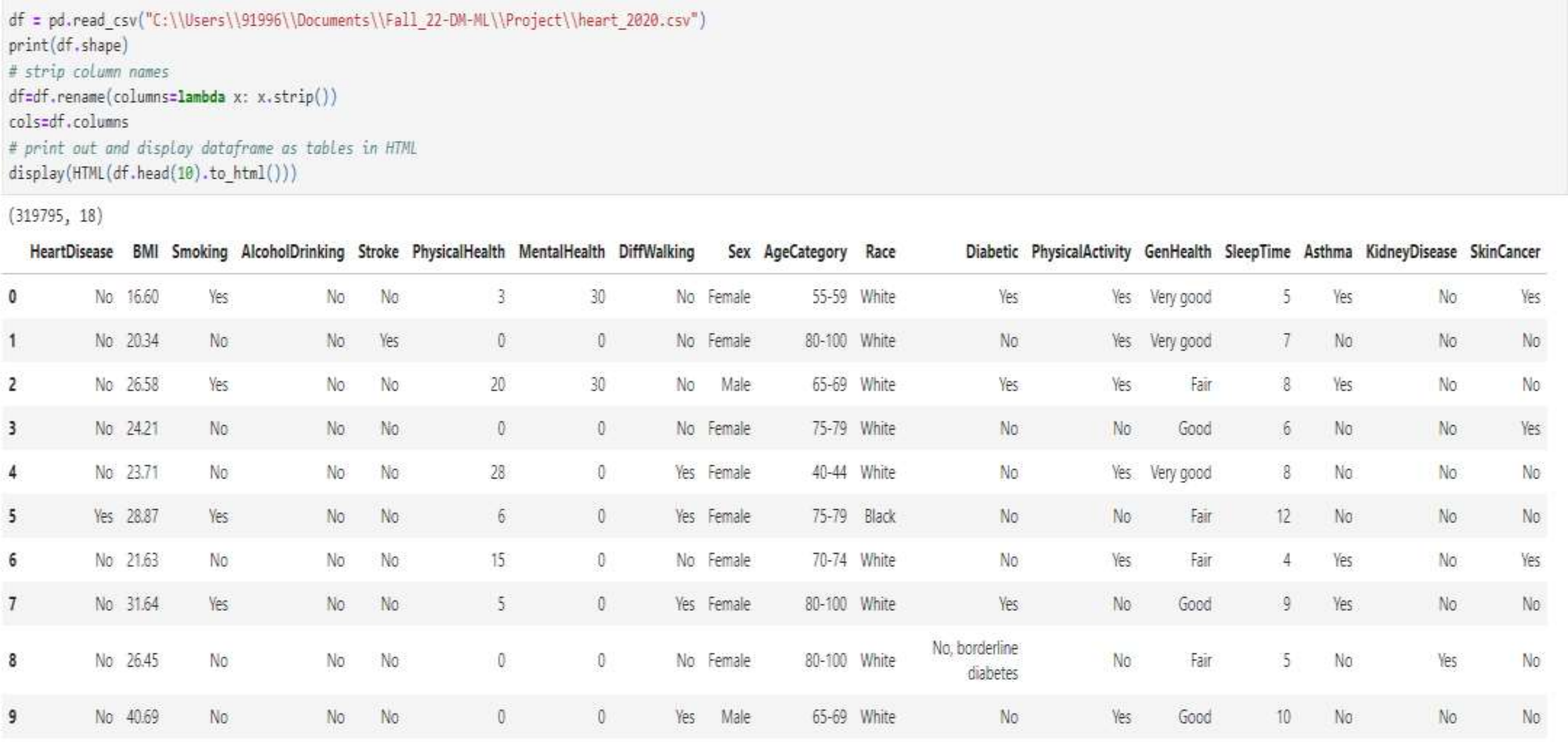
- Heart disease typically depends on various factors like physical health, mental health, sleep cycle, drug abuse, BMI, etc. In past few years it has seen that even a healthy person can suffer from various heart disease irrespective of their lifestyle.

Dataset is obtained from Kaggle:

Link: https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease

Dataset has 18 features and 319,796 rows.

Introduction



```
df = pd.read_csv("C:\\Users\\91996\\Documents\\Fall_22-DM-ML\\Project\\heart_2020.csv")
print(df.shape)
# strip column names
df=df.rename(columns=lambda x: x.strip())
cols=df.columns
# print out and display dataframe as tables in HTML
display(HTML(df.head(10).to_html()))
```

```
(319795, 18)
```

	HeartDisease	BMI	Smoking	AlcoholDrinking	Stroke	PhysicalHealth	MentalHealth	DiffWalking	Sex	AgeCategory	Race	Diabetic	PhysicalActivity	GenHealth	SleepTime	Asthma	KidneyDisease	SkinCancer
0	No	16.60	Yes	No	No	3	30	No	Female	55-59	White	Yes	Yes	Very good	5	Yes	No	Yes
1	No	20.34	No	No	Yes	0	0	No	Female	80-100	White	No	Yes	Very good	7	No	No	No
2	No	26.58	Yes	No	No	20	30	No	Male	65-69	White	Yes	Yes	Fair	8	Yes	No	No
3	No	24.21	No	No	No	0	0	No	Female	75-79	White	No	No	Good	6	No	No	Yes
4	No	23.71	No	No	No	28	0	Yes	Female	40-44	White	No	Yes	Very good	8	No	No	No
5	Yes	28.87	Yes	No	No	6	0	Yes	Female	75-79	Black	No	No	Fair	12	No	No	No
6	No	21.63	No	No	No	15	0	No	Female	70-74	White	No	Yes	Fair	4	Yes	No	Yes
7	No	31.64	Yes	No	No	5	0	Yes	Female	80-100	White	Yes	No	Good	9	Yes	No	No
8	No	26.45	No	No	No	0	0	No	Female	80-100	White	No, borderline diabetes	No	Fair	5	No	Yes	No
9	No	40.69	No	No	No	0	0	Yes	Male	65-69	White	No	Yes	Good	10	No	No	No

Research Problems

- The problem with heart disease is that, in past recent years it has been observed that even after maintaining the healthy lifestyle, people do suffer from heart disease.
- What are the reasons behind this problem?
- How prone are people to suffer from heart disease if they are into drug abuse?
- What percent of people suffer from heart disease if they maintain a healthy lifestyle?

Potential Solutions

- Exploratory Data Analysis (EDA)
- Various Classification model to compare accuracy like Logistic Regression
- K-NN (k-Nearest Neighbors)
- SVM
- Decision Trees
- Random Forest
- AdaBoosting
- XGBoost

Expected Outcomes

- At what level the drug abuse affects the heart health?
- People of which race suffer the most?
- What should be the ideal sleeping time?
- Is mental health related?
- Which age-category is highly prone?

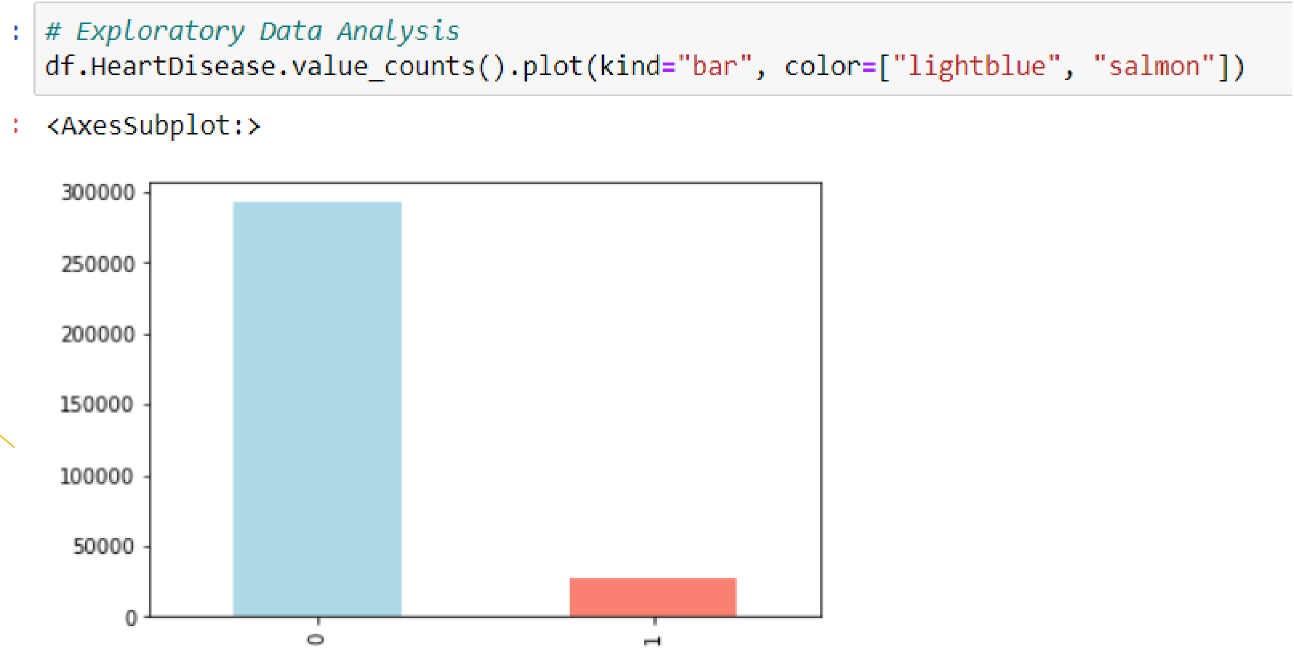
Potential Solutions

- Exploratory Data Analysis

- EDA helps us to observe and analyse the data to see what we are going to work with. The goal here is to learn more about the data.
- EDA also helps us find answers to some important questions such as: What kind of data do we have and how do we handle the different types? What is missing in the data and how do you deal with it? Etc.

Potential Solutions

- Exploratory Data Analysis



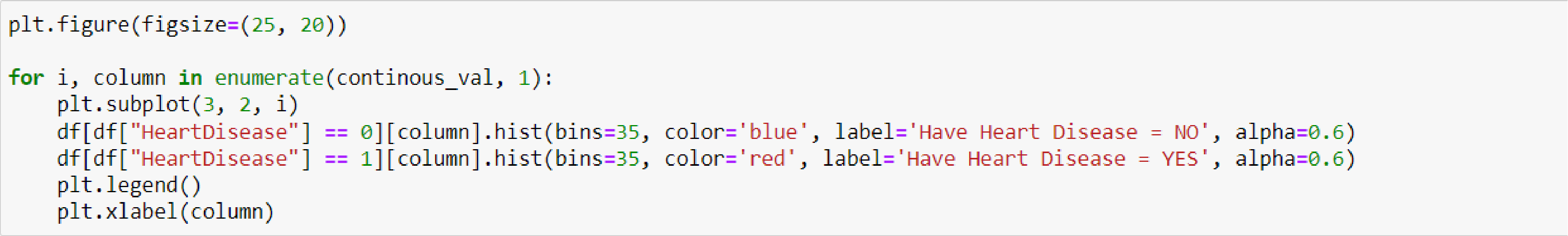
```
# Exploratory Data Analysis
df.HeartDisease.value_counts().plot(kind="bar", color=["lightblue", "salmon"])
```

```
<AxesSubplot:>
```

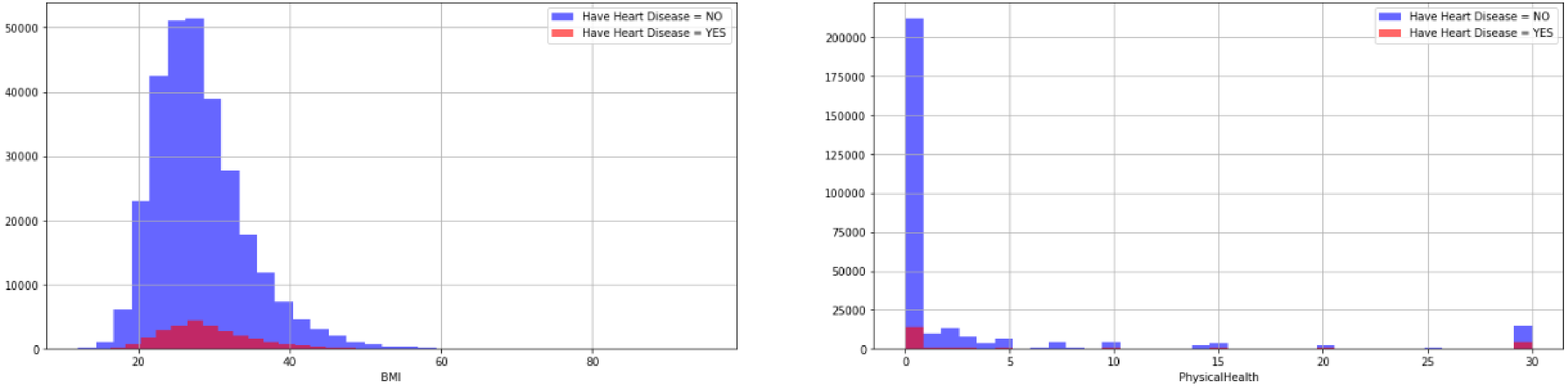
- We have close to 300,000 people with no heart disease and roughly around 25,000 people with heart disease.

Potential Solutions

- Exploratory Data Analysis

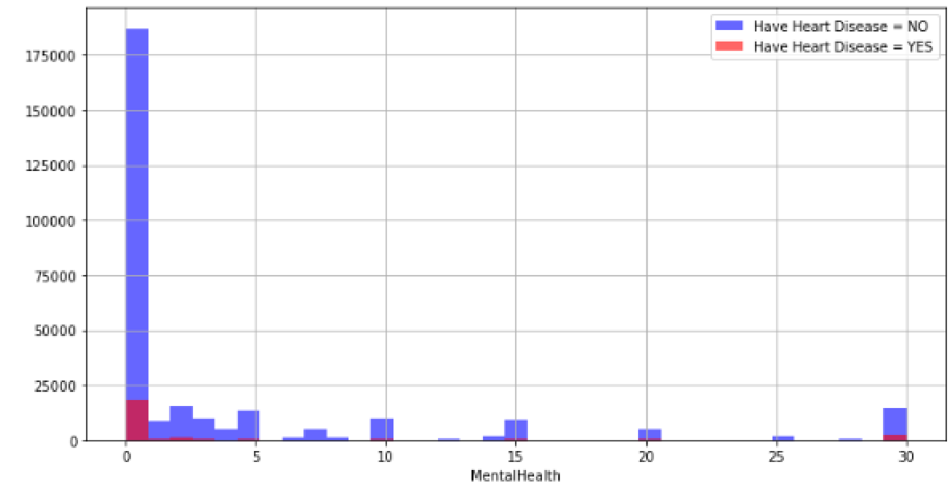


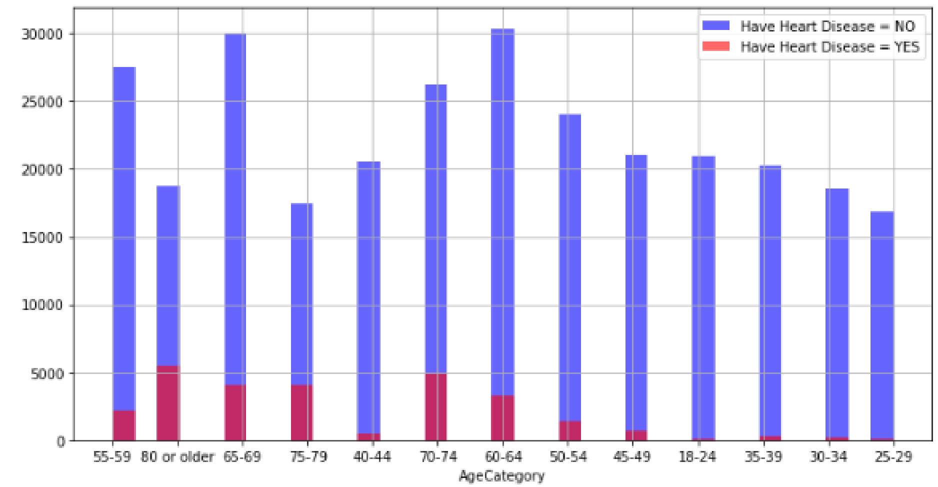
```
plt.figure(figsize=(25, 20))

for i, column in enumerate(continous_val, 1):
    plt.subplot(3, 2, i)
    df[df["HeartDisease"] == 0][column].hist(bins=35, color='blue', label='Have Heart Disease = NO', alpha=0.6)
    df[df["HeartDisease"] == 1][column].hist(bins=35, color='red', label='Have Heart Disease = YES', alpha=0.6)
    plt.legend()
    plt.xlabel(column)
```

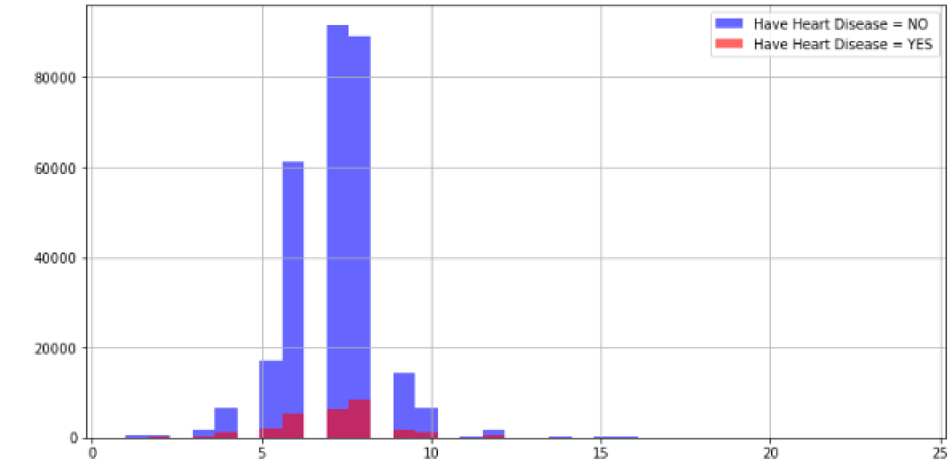


Potential Solutions

- Exploratory Data Analysis







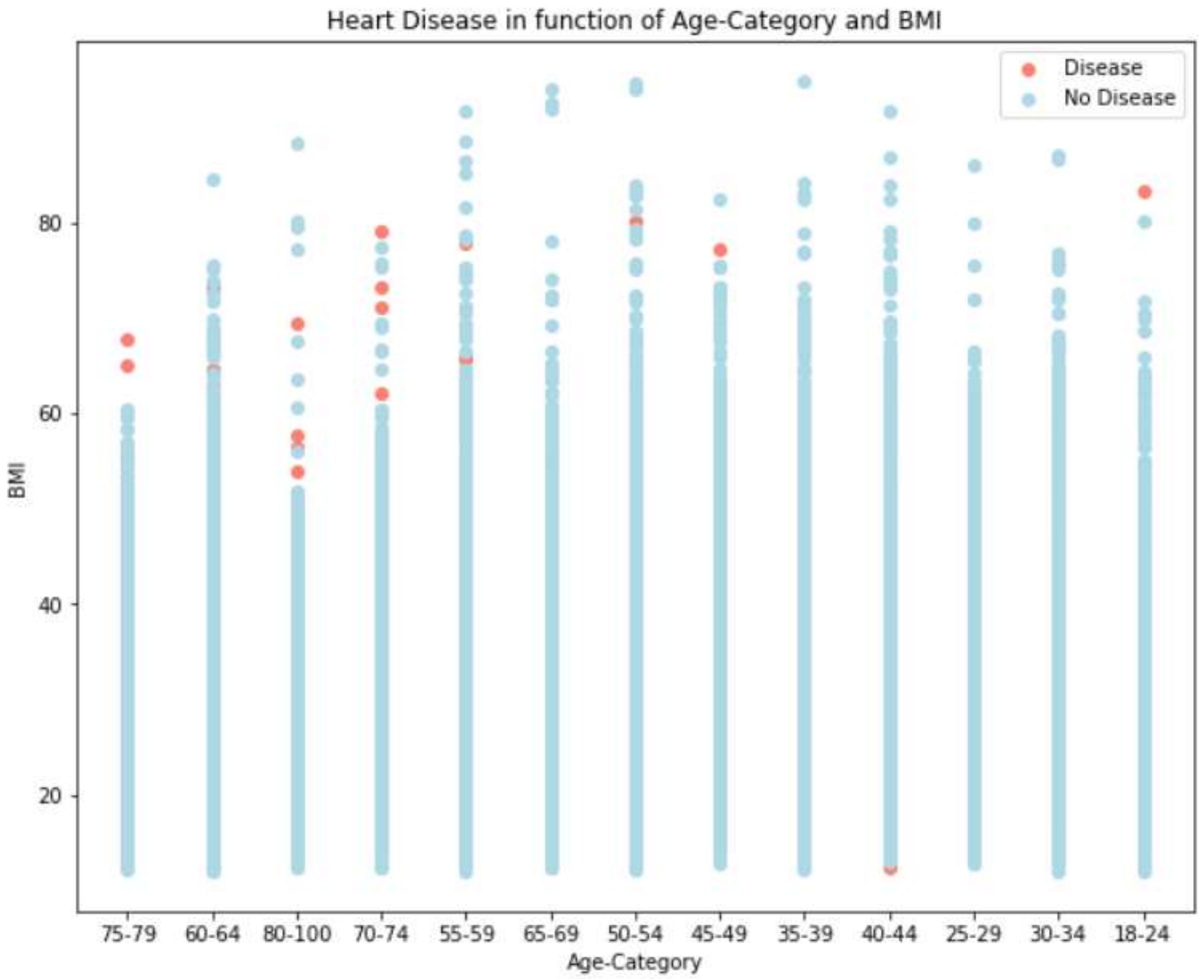
Potential Solutions

- Exploratory Data Analysis

- In Age-category, people above 60 years are more prone to disease.
- Physical health: With no to less exercise can cause heart problems.
- It is observed that BMI between 25.00 to 40.00 or more than 40.00 leads to heart issues.
- Sleep time: It is observed that, those individuals who has difficulty in sleeping early or individuals those who goes to bed after mid-night are on a higher risk of heart disease.

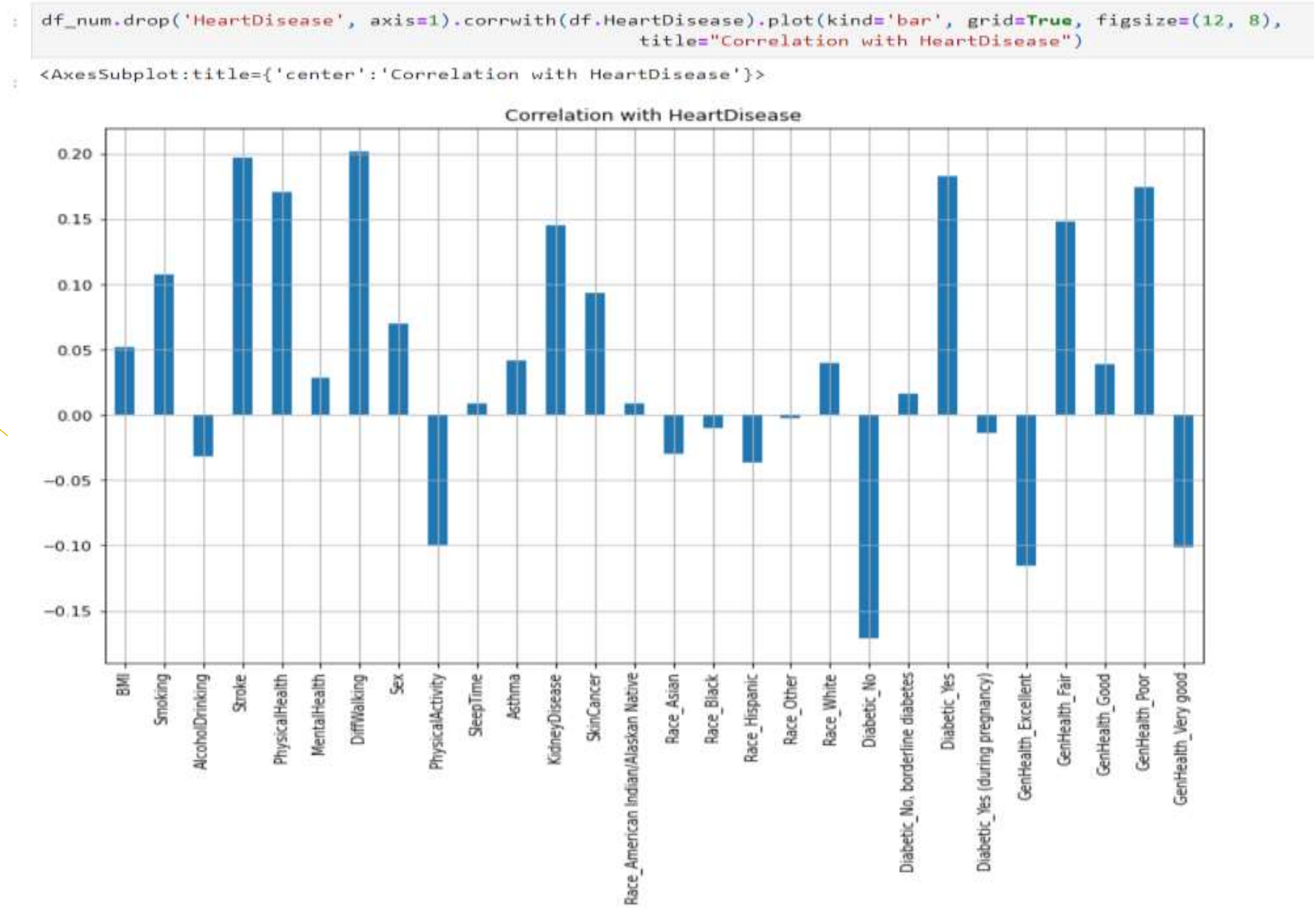
Potential Solutions

- Exploratory Data Analysis



Potential Solutions

- Exploratory Data Analysis



```
df_num.drop('HeartDisease', axis=1).corrwith(df.HeartDisease).plot(kind='bar', grid=True, figsize=(12, 8),
                                                                    title="Correlation with HeartDisease")
```

```
<AxesSubplot:title={'center':'Correlation with HeartDisease'}>
```

Potential Solutions

- Exploratory Data Analysis

Observations from the above correlation:

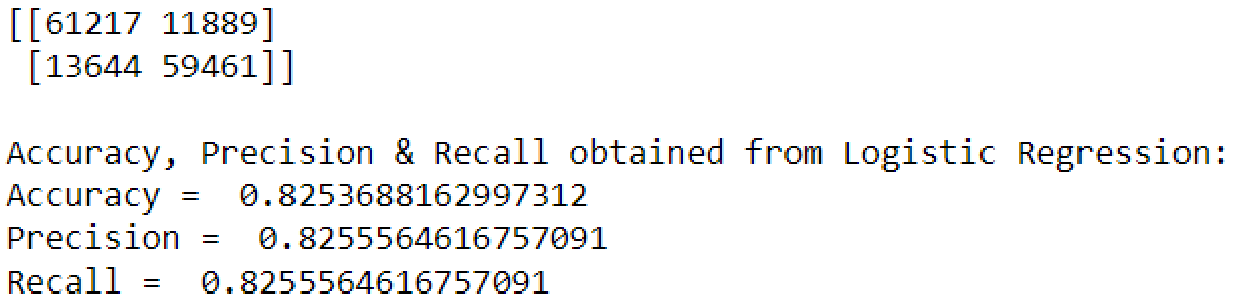
- Race_other is the least correlated with the HeartDisease variable.
- All other variables have a significant correlation with the HeartDisease variable.

Potential Solutions

Model 1: Logistic Regression

- Logistic regression is a simple and more efficient method for binary and linear classification problems.
- It is a classification model, and achieves very good performance with linearly separable classes.
- It is an extensively employed algorithm for classification in industry.

Output obtained from Logistic Regression model:



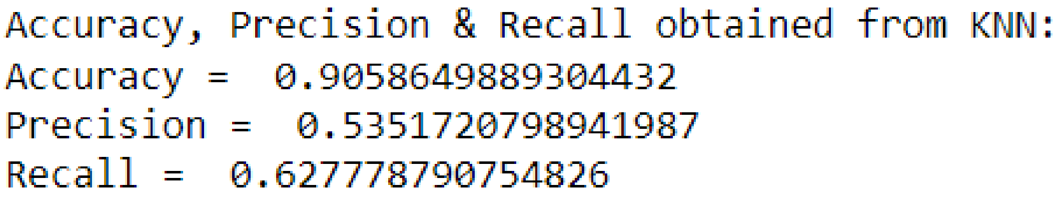
```
[[61217 11889]
 [13644 59461]]

Accuracy, Precision & Recall obtained from Logistic Regression:
Accuracy =  0.8253688162997312
Precision =  0.8255564616757091
Recall =  0.8255564616757091
```

Potential Solutions

Model 2: K-NN

- The k-nearest neighbours algorithm, also known as KNN or k-NN, is a non-parametric, supervised learning classifier.
- It uses proximity to make classifications or predictions about the grouping of an individual data point.

Output obtained from KNN model:



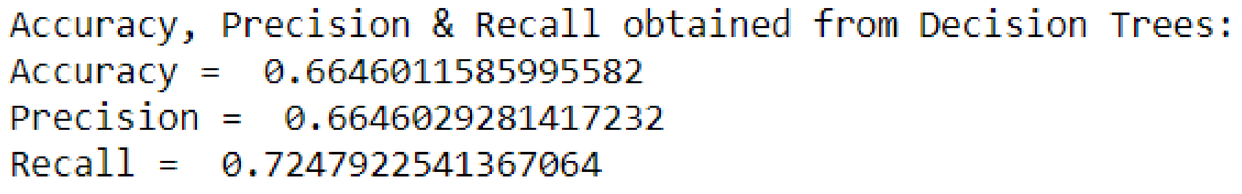
```
Accuracy, Precision & Recall obtained from KNN:
Accuracy =  0.9058649889304432
Precision =  0.5351720798941987
Recall =  0.627778790754826
```

Potential Solutions

Model 3: Decision Trees

- Decision Trees are supervised learning method used for classification and regression.
- Simple to understand and to interpret.
- Requires little data preparation.
- Able to handle both numerical and categorical data.
- Able to handle multi-output problems.

Output obtained from Decision Trees model:



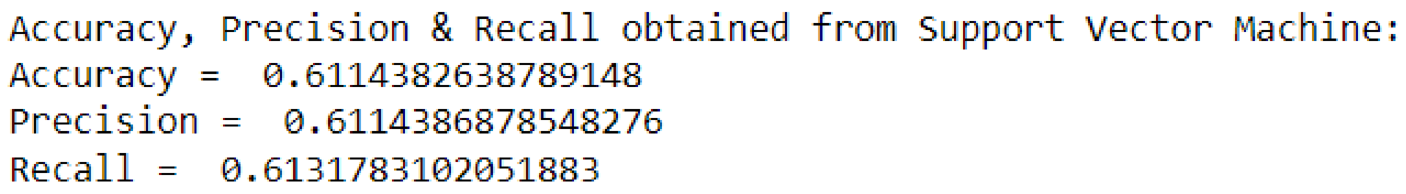
```
Accuracy, Precision & Recall obtained from Decision Trees:
Accuracy =  0.6646011585995582
Precision =  0.6646029281417232
Recall =  0.7247922541367064
```

Potential Solutions

Model 4: Support Vector Machine (SVM)

- Support vector machines are a set of supervised learning methods used for classification, regression and outliers detection.
- SVMs were originally designed for binary classifications.
- SVM can also be used for multi-class classifications.
- SVM require a numerical feature space to be run.
- Normalization is not required

Output obtained from Support Vector Machine model:



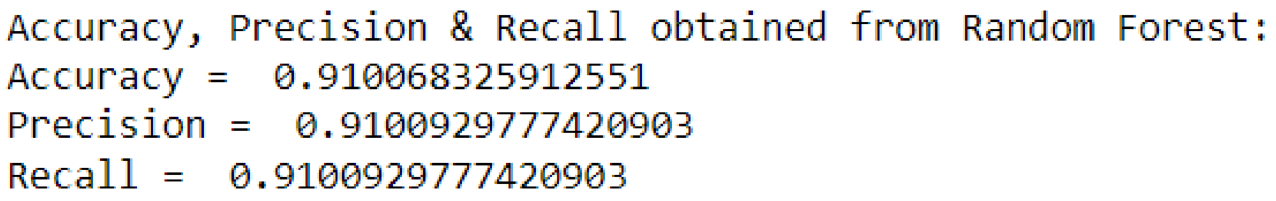
```
Accuracy, Precision & Recall obtained from Support Vector Machine:
Accuracy =  0.6114382638789148
Precision =  0.6114386878548276
Recall =  0.6131783102051883
```

Potential Solutions

Model 5: Random Forest

- Random forest is a estimator that fits a number of decision tree classifiers on various sub-samples of the dataset.
- Random forest uses averaging to improve the predictive accuracy and control over-fitting.

Output obtained from Random Forest model:



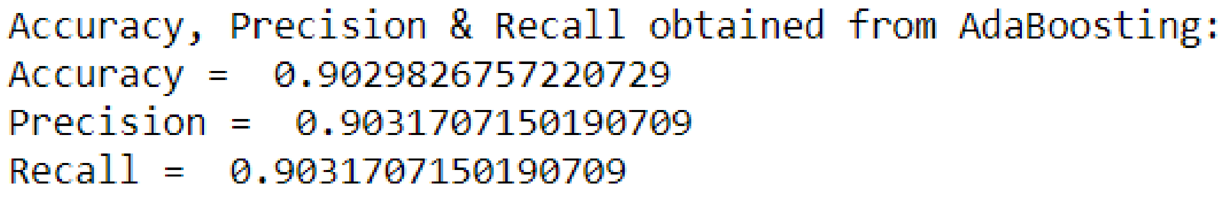
```
Accuracy, Precision & Recall obtained from Random Forest:
Accuracy =  0.910068325912551
Precision =  0.910092977742090З
Recall =  0.910092977742090З
```

Potential Solutions

Model 6: AdaBoosting

- AdaBoost classifier is a meta-estimator that begins by fitting a classifier on the original dataset and then fits additional copies of the classifier on the same dataset.
- The weights of incorrectly classified instances are adjusted such that subsequent classifiers focus more on difficult cases.

Output obtained from AdaBoosting model:



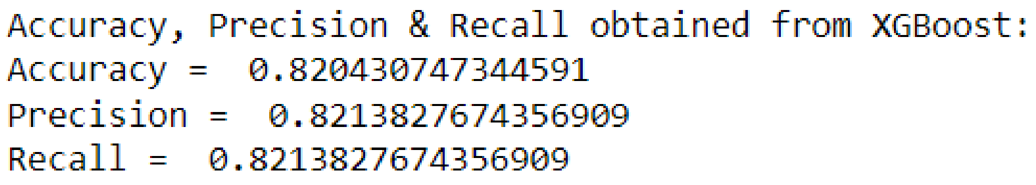
```
Accuracy, Precision & Recall obtained from AdaBoosting:
Accuracy =  0.9029826757220729
Precision =  0.9031707150190709
Recall =  0.9031707150190709
```

Potential Solutions

Model 7: XGBoost

- XGBoost is an optimized distributed gradient boosting library designed to be highly efficient and flexible.
- XGBoost provides a parallel tree boosting that solve many data science problems in a fast and accurate way.

Output obtained from XGBoost model:



```
Accuracy, Precision & Recall obtained from XGBoost:
Accuracy =  0.820430747344591
Precision =  0.8213827674356909
Recall =  0.8213827674356909
```

Result

By Comparing all models:

- Random Forest and AdaBooosting models yielded high accuracy.
- Whereas, XdBoost, Logistic Regression and SVM yielded less accuracy.

Conclusion

By Comparing all models:

- In this project, we have analysed heart disease dataset, which had 17 indicators of heart disease of 319,795 surveyed individuals in the United States.
- During our investigation we identified that age is a major factor in heart disease.
- Furthermore, heart disease is more prominent in those individuals who has no physical activity, has BMI between 25-40 and suffers from some sort of mental illness.

Future Scope:

- Unsupervised learning model can be implemented and compared with Supervised model.
- The models can be regressively tested by changing the parameters.

Questions?