Heart Disease Prediction and Analysis of key factors affecting the heart

Heart Disease is one of the major causes of death in the USA. Many factors contribute to the health of the heart. Understanding how different factors affect the heart can help in rectifying and make good choices to ensure a healthy heart and a long life. The dataset that is being analyzed is captured by the CDC. Although many factors influence the heart directly or indirectly, some of the most relevant factors are being captured in this dataset which include, Smoking, Alcohol consumption, General health, Physical and Mental health, Age, race, sleep patterns, other pre-existing diseases like diabetics, Kidney health, Asthma, Cancer etc.

Objective - The objecting of this project is to analyze the different factors that may contribute to the health of the heart and build a prediction model that can predict heart diseases in patients. The accurate prediction will help in identifying health issues in early stages and treating them in time.

Dataset - The dataset is retrieved from Kaggle. https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease

Attributes/Features –

1. HeartDisease : Respondents that have ever reported having coronary heart disease (CHD) or myocardial infarction (MI).
2. BMI : Body Mass Index (BMI).
3. Smoking : Have you smoked at least 100 cigarettes in your entire life? (The answer Yes or No).
4. AlcoholDrinking : Heavy drinkers (adult men having more than 14 drinks per week and adult women having more than 7 drinks per week
5. Stroke : (Ever told) (you had) a stroke?
6. PhysicalHealth : Now thinking about your physical health, which includes physical illness and injury, for how many days during the past 30 days was your physical health not good? (0-30 days).
7. MentalHealth : Thinking about your mental health, for how many days during the past 30 days was your mental health not good? (0-30 days).
8. DiffWalking : Do you have serious difficulty walking or climbing stairs?
9. Sex : Are you male or female?
10. AgeCategory: Fourteen-level age category.
11. Race : Imputed race/ethnicity value.
12. Diabetic : (Ever told) (you had) diabetes?
13. PhysicalActivity : Adults who reported doing physical activity or exercise during the past 30 days other than their regular job.
14. GenHealth : Would you say that in general your health is...
15. SleepTime : On average, how many hours of sleep do you get in a 24-hour period?
16. Asthma : (Ever told) (you had) asthma?
17. KidneyDisease : Not including kidney stones, bladder infection or incontinence, were you ever told you had kidney disease?
18. SkinCancer : (Ever told) (you had) skin cancer?

Categorical Features:
HeartDisease, Smoking, AlcoholDrinking, Stroke, DiffWalking, Sex, Race, Diabetic,

PhysicalActivity, GenHealth, Asthma, KidneyDisease, SkinCancer

Continuous Features:
BMI, PhysicalHealth, MentalHealth, AgeCategory, SleepTime

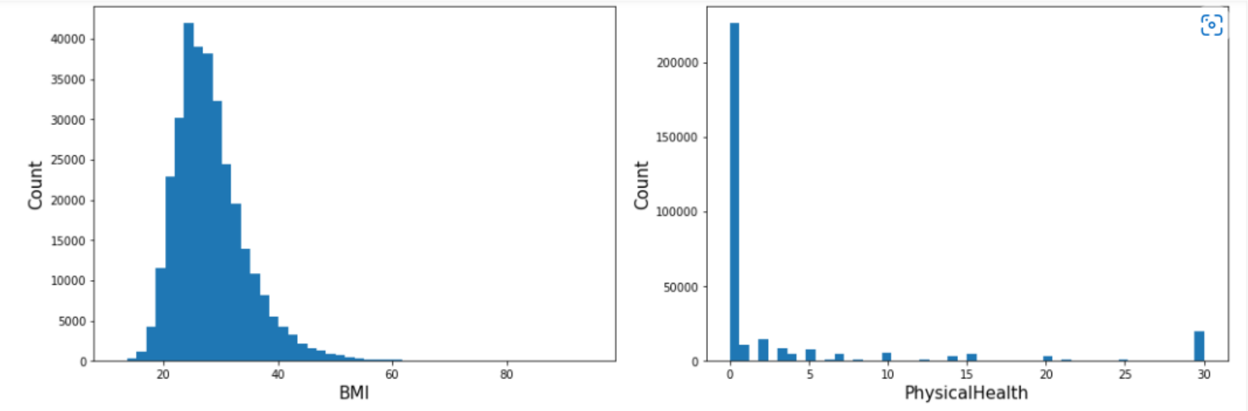
Data Analysis and EDA

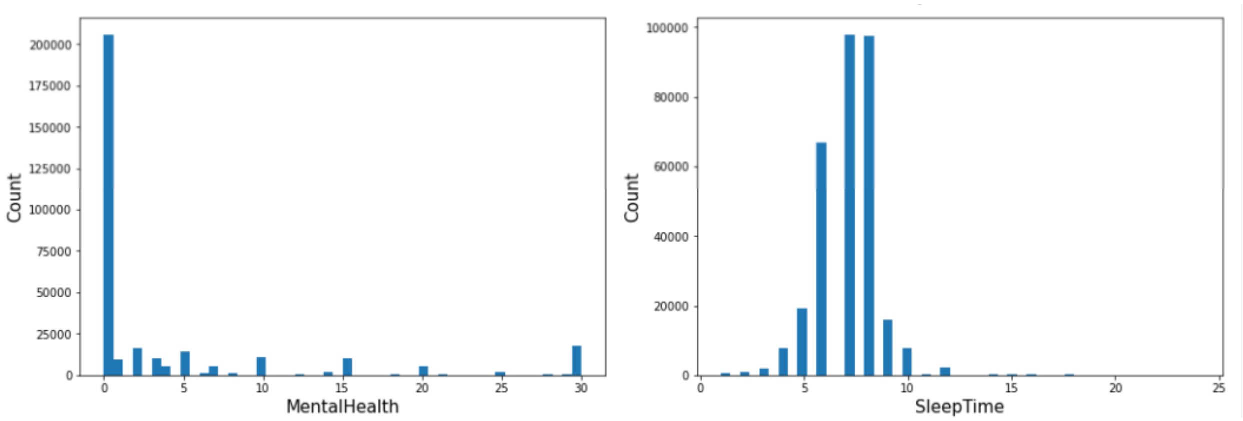
The dataset has 319795 rows with 18 columns.

HeartDisease column will be the target for this project. There are no null values in this dataset, which means we do not have to drop any rows.

Most of the columns have 2 unique values, Yes and No. We also can see that the HeartDisease column has 292422 values as No and only 27373 values as Yes - reported any HeartDisease.

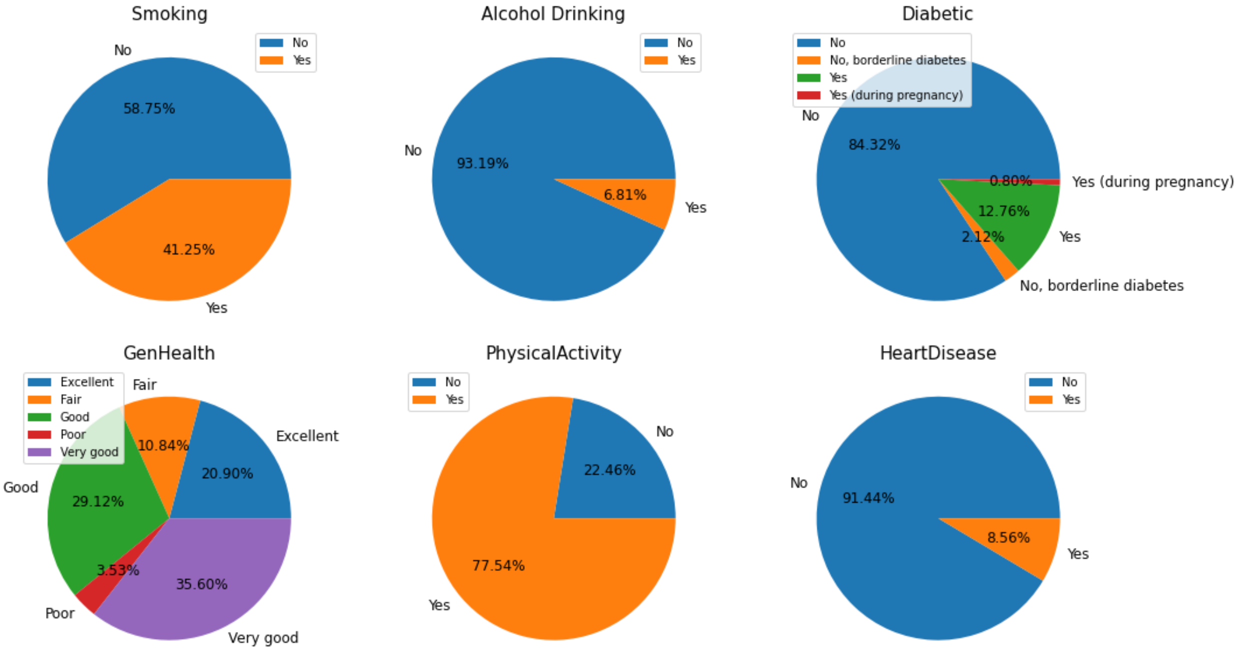
The other columns or features represent a very similar trend, making this dataset very unbalanced. We will need to either under-sample or oversampling during the data transformation phase. Analyzing the data distribution is one of the primary steps to perform. This will help in understanding the dataset better and any data cleaning or transformation can be easily performed. Plots like histograms, pie, bar graphs, scatterplots help in visualizing and understanding the data.





From the above histograms we can observe -

1. The BMI is concentrated between 25-35
2. Most of the respondents did not have any bad physical or mental health days
3. Most respondents were getting 6-8 hours of sleep per night.

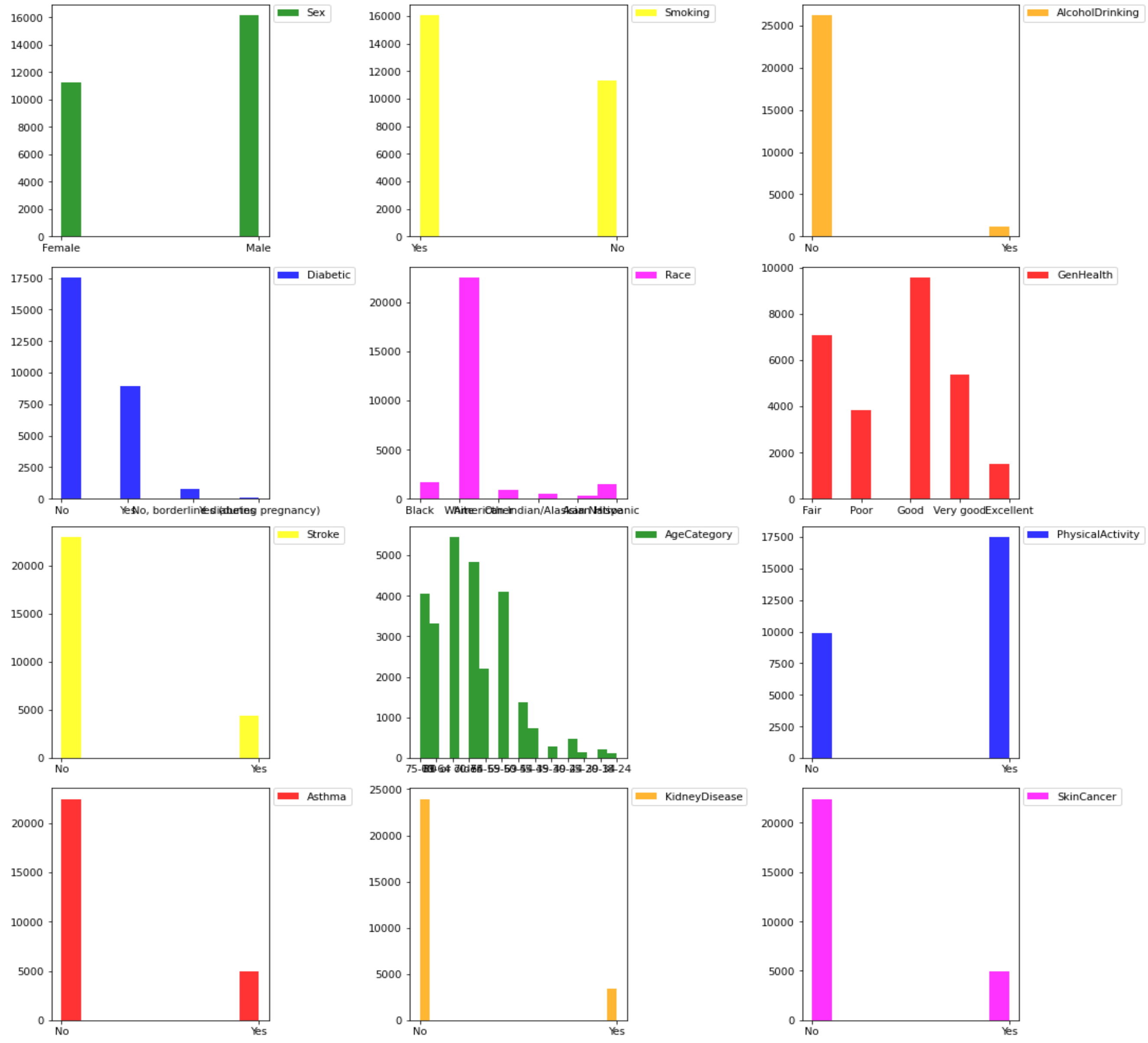


From the above pie graphs, the below observations can be inferred -

1. The Heart Disease distribution is very unbalanced, in that the dataset sample that we are considering does not have a balanced response. Only about 8.56% respondents have

reports heart disease. Therefore, to build a prediction model, the dataset may need to be under-sampled or oversampled.

Analyzing the dataset that has reported heart disease will provide a closer look at the factors that may influence the health of the heart.



From the above we observe,

1. Males are more susceptible to heart diseases.
2. Smoking seems to play a crucial role in the heart's health.
3. At a first glance, being diabetic may not translate to having poor heart health, but about 40% of respondents were diabetic. We may conclude that keep a check on the diabetes is important for health of the heart.

4. Interestingly, the White race seem to have reported the highest heart diseases. But one must wonder if the sampling or data collection was done in a heavy white population area.
5. General health also plays a role in determining the heart health. If you combine fair, poor, good general health, then it seems to affect the heart.
6. About 40% respondents that reported heart disease responded No to physical activity making it another crucial factor to consider.
7. Other diseases like Asthma, Kidney disease, Skin cancer do not seem to contribute significantly to determining the heart health.
8. Most respondents that reported heart disease are over 50 making age a factor in determining the heart's health.

As part of data preparation for the modeling and analysis, the original dataset has undergone the following steps –

1. Few columns like Smoking, AlcoholDrinking, Stroke etc. have only 2 distinct values. Converting them to 0 and 1 would be a sufficient, dummies are not needed for them. Smoking 2 AlcoholDrinking 2 Stroke 2 DiffWalking 2 PhysicalActivity 2 Asthma 2 KidneyDisease 2 SkinCancer 2
2. Columns like AgeCategory, Race, GenHealth, Diabetic - dummies can be created since they have more than 2 values but still are categorical columns. Race 6 Diabetic 4 GenHealth 5
3. The range of continuous features are different. Therefore, scaling them to be in-between 0 to 1 by dividing by the maximum value of the respective column.
4. The dataset is balanced with oversampling

Model Building and Analysis –

The cleaned and balanced (with oversampling) dataset is now split into training and test data sets with 30% being test data. The target is to predict whether the respondent have any complaint of heart disease based on the features.

Feature selection is crucial for predicting a target value. For this dataset, below are being selected as best features –

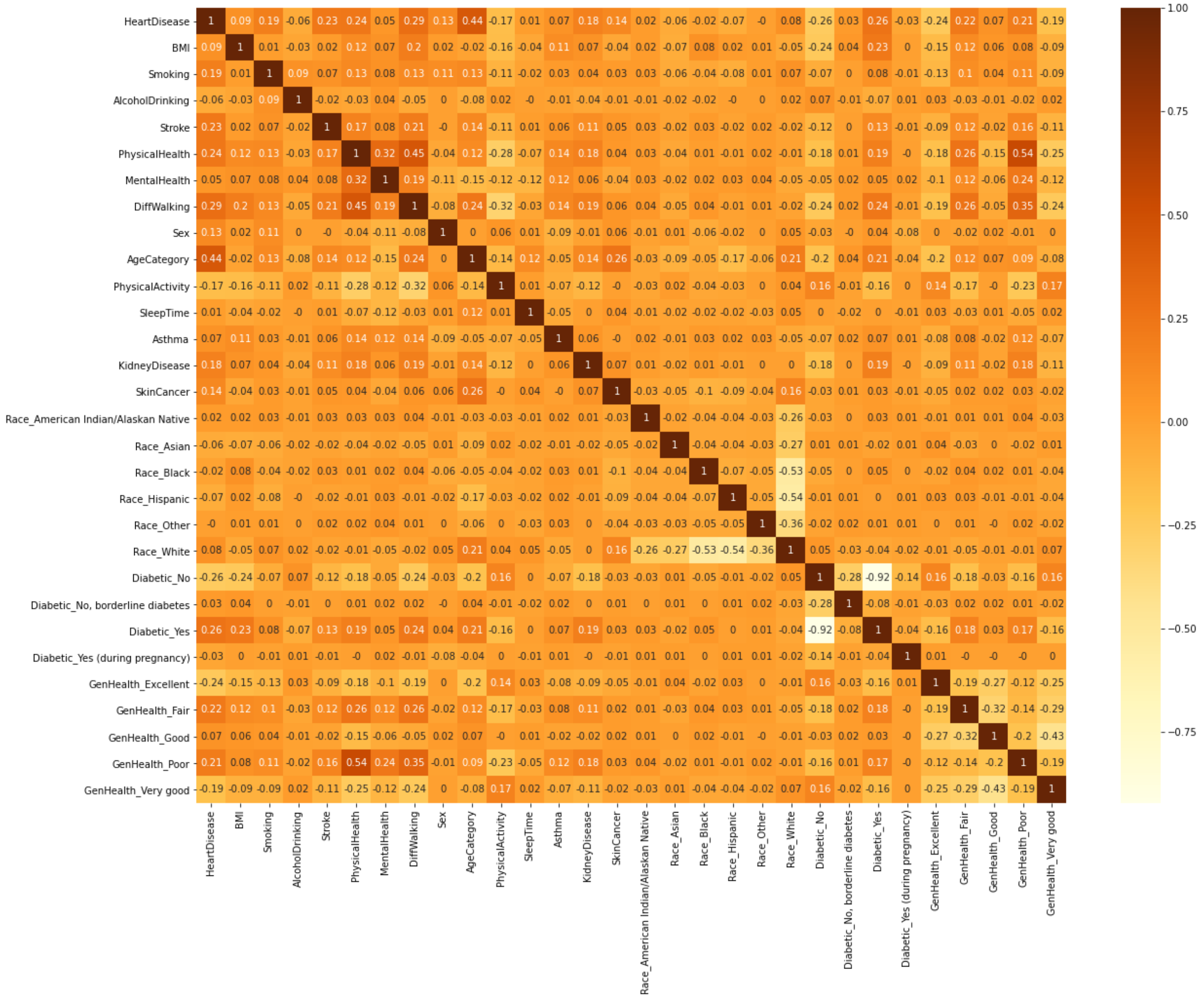
BMI,
Smoking,
AlcoholDrinking,
Stroke,
MentalHealth,
DiffWalking,
Sex,
AgeCategory,
SleepTime,
Asthma,
KidneyDisease,
Race_American Indian/Alaskan Native,
Race_Black,
Race_White,
Diabetic_No, borderline diabetes,
Diabetic_Yes (during pregnancy),
GenHealth_Excellent,

GenHealth_Fair,
GenHealth_Good,
GenHealth_Poor

Based on different models, the accuracy is as follows –

Sno	Model	Accuracy
1	Logistic Regression	76%
2	Random Forest Classifier	96%
3	KNeighborsClassifier	89%

Feature Correlation –



Conclusion -The features that most affect the heart are Age factor, Diabetes, have trouble in walking, physical health and General Health being less than good, if there was an incident of Stroke. Smoking, pre-existing Kidney and skin cancer issues may also contribute significantly to a weak heart. There also seems to be the factor of gender but that is quite low to be deterministic in our prediction.

Overall, the models predictions are varying largely to be able to choose the best. It is possible that the large difference is due to the oversampling process that was done on the highly unbalanced original dataset. All said and done, the important factors affecting the heart are the well-known offenders including age, general health conditions, and pre-existing conditions such as Diabetes. Whether someone has reported a heart disease or not, the above factors need to be taken care of for a healthy heart at any time.