Cardiovascular Disease Prediction

Team Details

Ravi Shankar A V
Tintu B John
Abdul Rahman P J
Baskar D
Febin Thomas

Mentor

Vibha Santhanam

Problem Definition

- Heart disease is one of the leading causes of death for people in our country. Public health estimates indicate that India accounts for approximately 60% of the world's heart disease burden.
- About half of all Indians have at least 1 of 3 key risk factors for heart disease: high blood pressure, high cholesterol, and smoking. Other key indicator includes diabetic status, obesity (high BMI), not getting enough physical activity or drinking too much alcohol.
- We need a system to predict a patient's hearts condition using these factors.

Business Objective

- To create a solution for predicting a patient’s heart condition using the key indicators from the acquired data.

Approach

- Build a machine learning model to accurately detect "patterns" from the Key Indicators to predict a patient's heart condition.

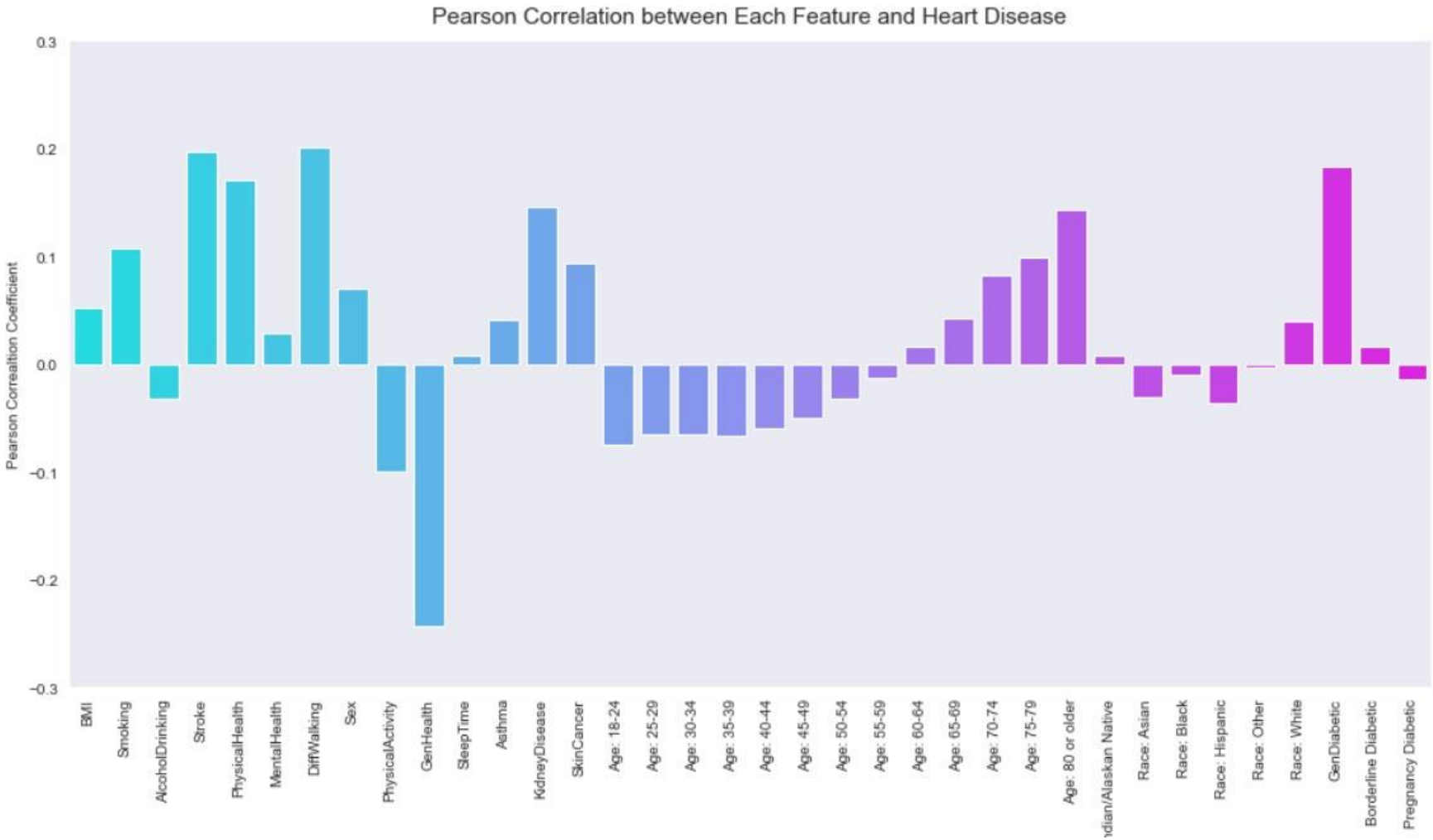
EDA

Dataset

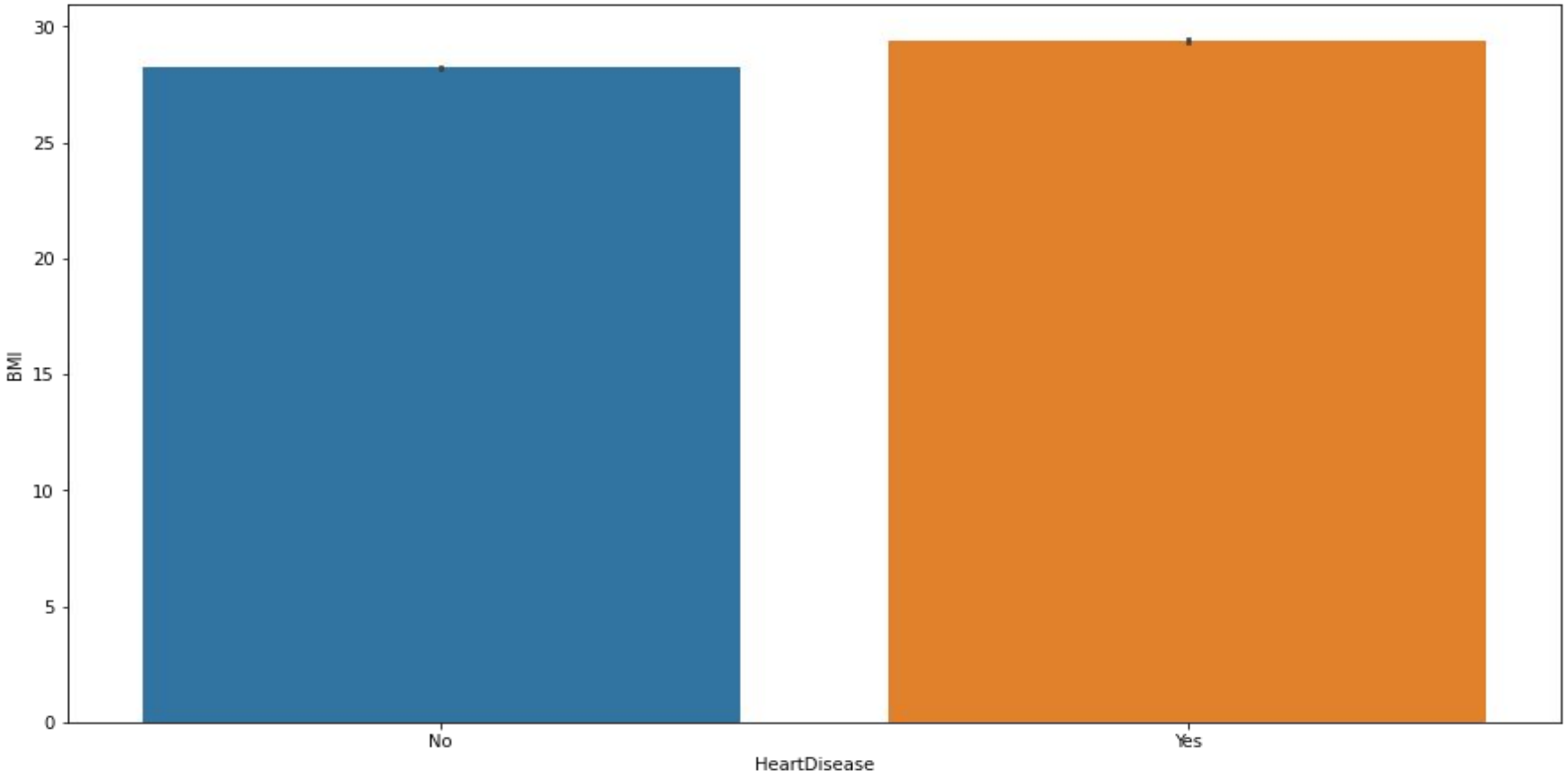
- The dataset provides us 319796 records with the various factors that lead to Heart Disease such as BMI, Smoking, Alcohol Drinking, Stroke, Physical Health, Mental Health, Diff Walking, Sex, Age Category, Race, Diabetic, Physical Activity, Gen Health, Sleep Time, Asthma, Kidney Disease, Skin Cancer.
- The dataset come from the CDC and is a major part of the Behavioral Risk Factor Surveillance System (BRFSS), which conducts annual telephone surveys to gather data on the health status https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease

	HeartDisease	BMI	Smoking	AlcoholDrinking	Stroke	PhysicalHealth	MentalHealth	DiffWalking	Sex	AgeCategory	Race	Diabetic	PhysicalActivity
0	No	16.60	Yes	No	No	3.0	30.0	No	Female	55-59	White	Yes	Yes
1	No	20.34	No	No	Yes	0.0	0.0	No	Female	80 or older	White	No	Yes
2	No	26.58	Yes	No	No	20.0	30.0	No	Male	65-69	White	Yes	Yes
3	No	24.21	No	No	No	0.0	0.0	No	Female	75-79	White	No	No
4	No	23.71	No	No	No	28.0	0.0	Yes	Female	40-44	White	No	Yes
5	Yes	28.87	Yes	No	No	6.0	0.0	Yes	Female	75-79	Black	No	No
6	No	21.63	No	No	No	15.0	0.0	No	Female	70-74	White	No	Yes
7	No	31.64	Yes	No	No	5.0	0.0	Yes	Female	80 or older	White	Yes	No
8	No	26.45	No	No	No	0.0	0.0	No	Female	80 or older	White	No, borderline diabetes	No
9	No	40.69	No	No	No	0.0	0.0	Yes	Male	65-69	White	No	Yes
10	Yes	34.30	Yes	No	No	30.0	0.0	Yes	Male	60-64	White	Yes	No
11	No	28.71	Yes	No	No	0.0	0.0	No	Female	55-59	White	No	Yes
12	No	28.37	Yes	No	No	0.0	0.0	Yes	Male	75-79	White	Yes	Yes
13	No	28.15	No	No	No	7.0	0.0	Yes	Female	80 or older	White	No	No
14	No	29.29	Yes	No	No	0.0	30.0	Yes	Female	60-64	White	No	No

CORRELATION GRAPH



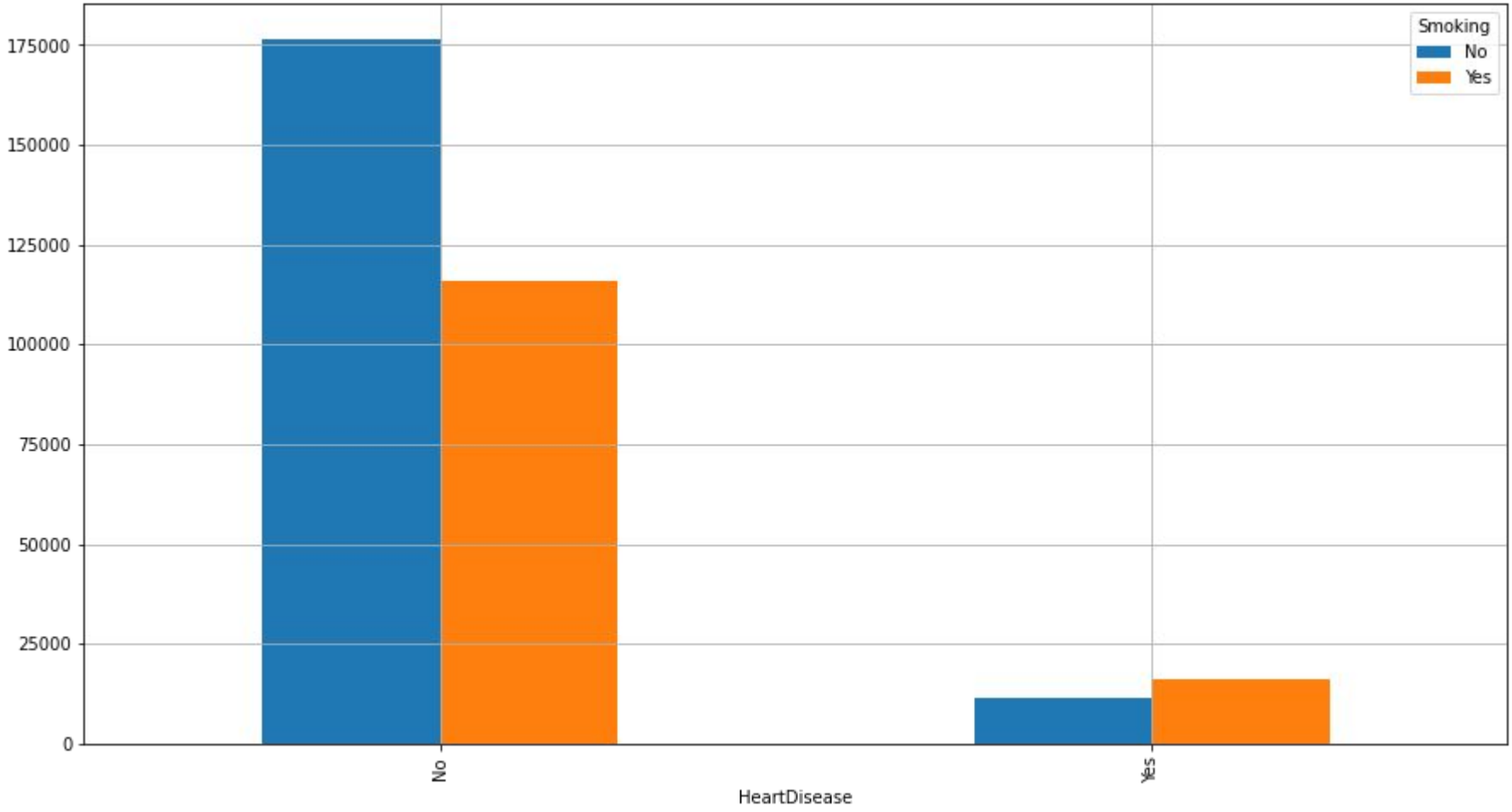
BMI



- BMI contributes to heart disease.

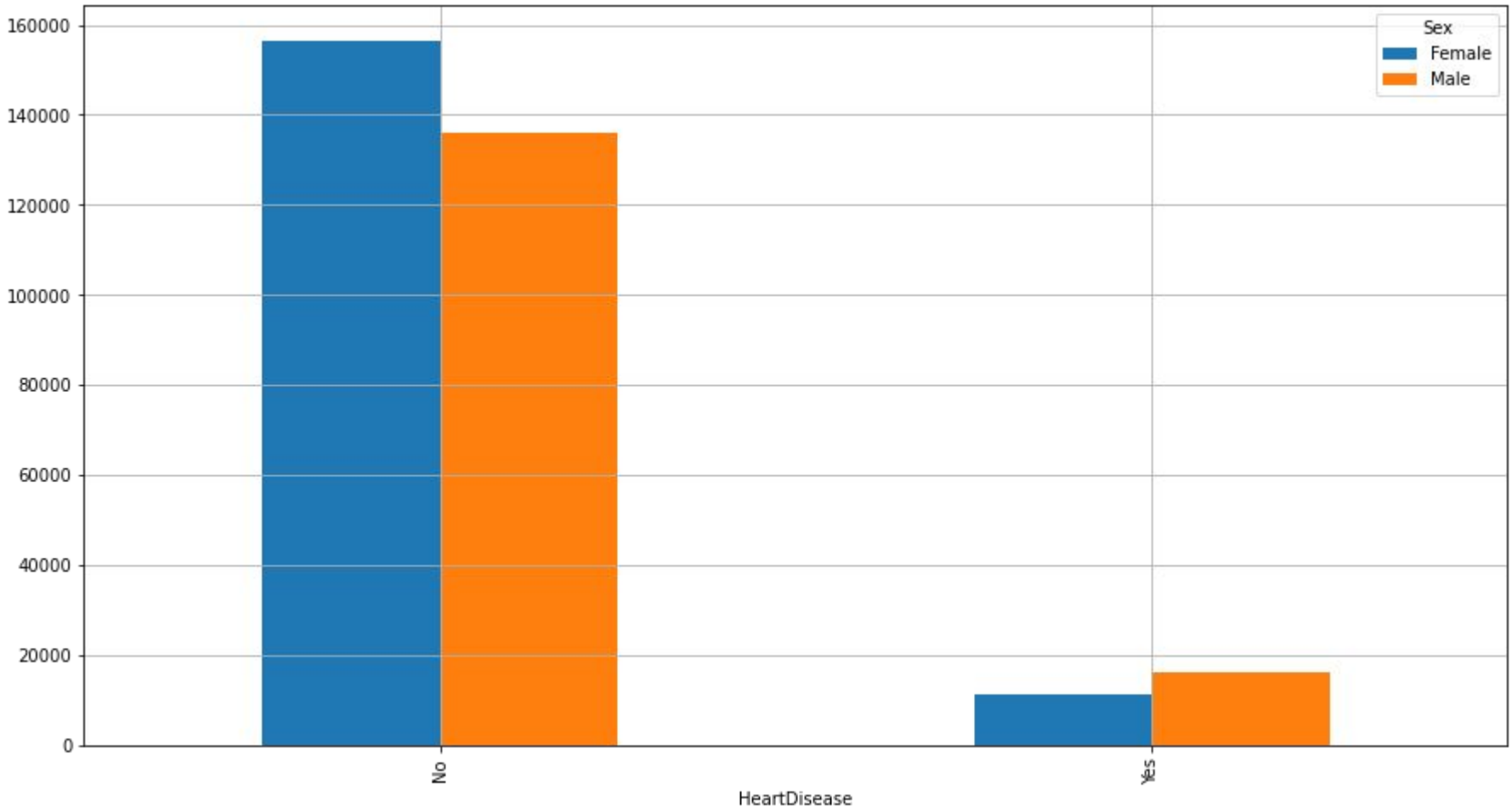
  People with BMI less than 20 or greater than 25 tend to have a higher chance of getting heart disease.

SMOKING



- Smoking habits contributes to heart disease.

SEX



- In terms of sexes, males tend to have a higher chance of heart disease than females.

Other Inferences from the data :

- Alcoholic drinking habits does not contribute to heart disease, which is quite counter-intuitive.
- Stroke contributes to heart disease.
- People older than 60 tend to have heart disease than those younger than 60.
- Kidney disease and diabetes contribute to heart disease.

- The amount of sleep time in a day does not seem to give a considerable contribution to heart disease. However, people without heart disease commonly have a better sleep time, that is about 6 to 8 hours in a day.
- Asthma only gives a small contribution.
- The amount of sleep time in a day does not seem to give a considerable contribution to heart disease. However, people without heart disease commonly have a better sleep time, that is about 6 to 8 hours in a day.

Algorithms, Solution and Conclusions

- GaussianNB :

  Naive Bayes methods are a set of supervised learning algorithms based on applying Bayes' theorem with the "naive" assumption of conditional independence between every pair of features given the value of the class variable.

  - Accuracy Score: 0.7619048562107172
  - Precision Score: 0.22755280407865987
  - Recall Score: 0.706827261761158
  - F1-Score: 0.339209790209790 2

- RandomForestClassifier :

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

  - Accuracy Score: 0.8935469412856085
  - Precision Score: 0.29183806018618746
  - Recall Score: 0.16260554885404102
  - F1-Score: 0.20884654117282517

Conclusion

The prediction exhibition of each technique and apply the proposed system for the area it required, The accuracy achieved for Naive Bayes is 76.19% and Random Forest showed 89.35%.

Thanks!