Chronic Heart Disease Prediction

Final Project - Data Mining (CSCI-B 565 Fall 2022)

Project by:

Krishna Teja Jillelamudi (kjillela)
Royce Dcunha (rdcunha)
Siona Crissel DSouza (sidsouza)

Problem Statement

According to the **Centers for Disease Control and Prevention**, *“One person dies every 34 seconds in the United States from cardiovascular disease”*

One major challenge is the detecting of the disease. There are tools that can forecast heart disease, but they are either expensive or ineffective in calculating the likelihood of heart disease in a human. Since there is a lot of data available nowadays, we can use data mining methods to search for hidden patterns. In medical data, the hidden patterns might be exploited for health diagnosis.

Motivation

One way to bring down such a huge number of losses would be to use Data Mining techniques for the Prediction of Chronic Heart Disease, which in turn can be of help to the government and hospitals.

Proposed Solution

We intend on using the *Behavioral Risk Factor Surveillance System (BRFSS)* survey data of the year 2021 to design a classification model that will help us in determining whether a person has heart disease or not, which will aid in early diagnosis and intervention while also lowering medical costs.

Implementation

Data Collection

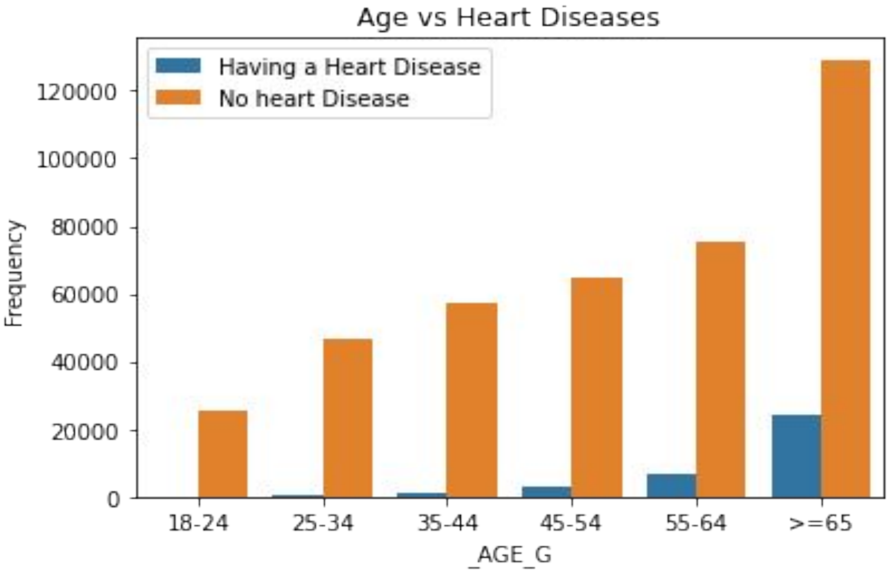
- We used the Behavioral Risk Factor Surveillance System (**BRFSS**) dataset
- Downloaded from CDC official site(https://www.cdc.gov/brfss/index.html)
- This dataset is the most comprehensive collection of state information on U.S. citizens' risk behaviors, chronic health issues, and utilization of preventive care in the country.
- The District of Columbia, three U.S. territories, and all 50 states are currently included in the data collection areas of BRFSS.
- The BRFSS is the biggest continually running health survey system in the world, conducting over 400,000 adult interviews each year.

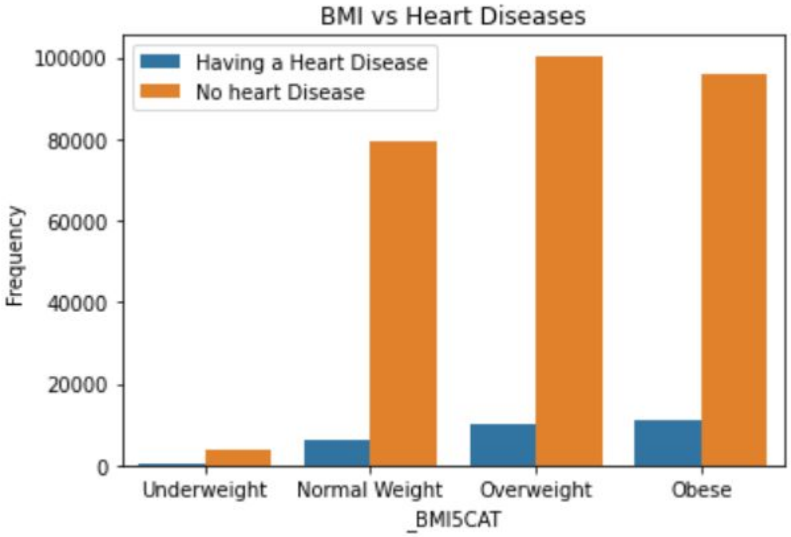
Pre-Processing

- Null Values/Missing Values: This part included identifying incomplete, inaccurate, duplicated or null values in the data. These values were then deleted as removing a few missing records removing wouldn’t impact the distribution of your dataset.
- Encoding: We used one-hot encoding for converting the data to prepare it for an algorithm and get a better prediction
- Noisy data: removed values which were refused/don’t know/not asked to the person.

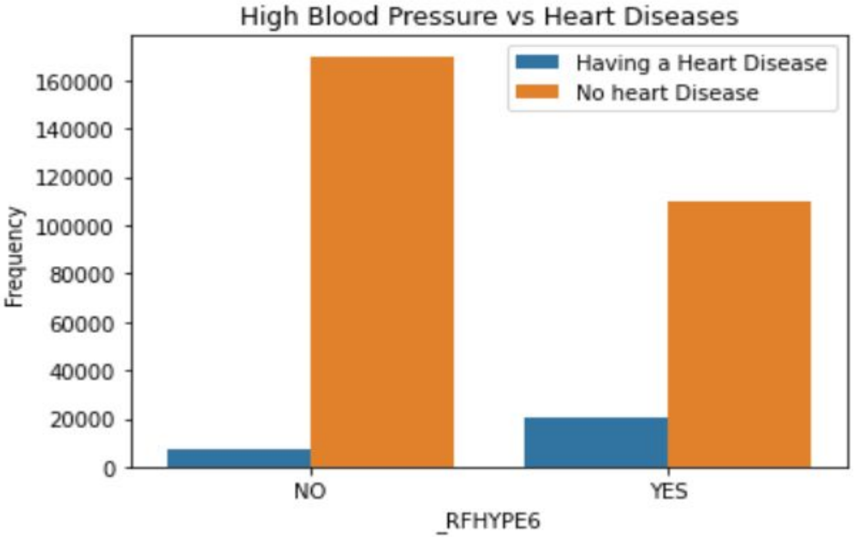
Data Visualization

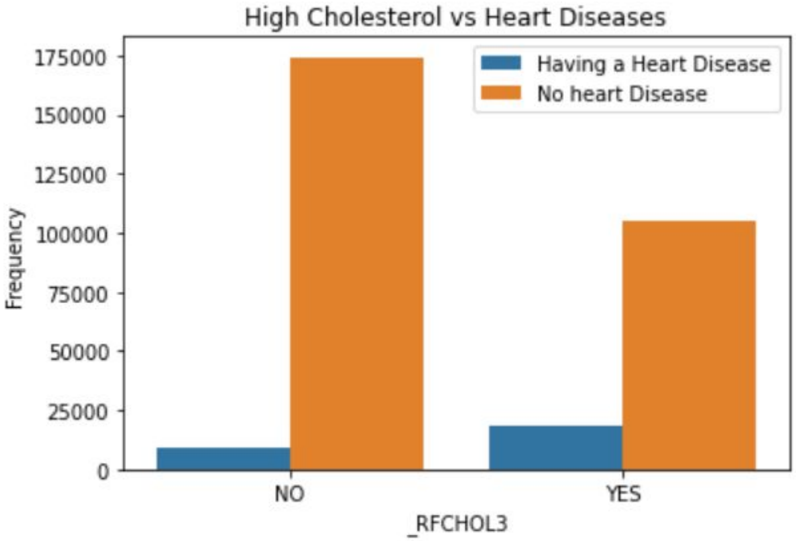
- The graphs below provide information about the dataset and how it compares to heart diseases.
- We have taken the following factors into account
  - Gender
  - Age
  - Mental Health
  - Smoking Habits
  - BMI





The highest percentage of people suffering from heart diseases are the people belonging to ‘>=65’ age group and ‘Overweight’ BMI which accounts for 26.5%





12.6% of people having heart disease are found to have normal cholesterol level and no high blood pressure.

Classification

- Sampling
- Models used:
    - Logistic Regression
    - Naive Bayes
    - Decision Tree
    - Bagging Classifier
    - AdaBoost Classifier
    - KNN
    - Neural Network

Performance Evaluation Metrics

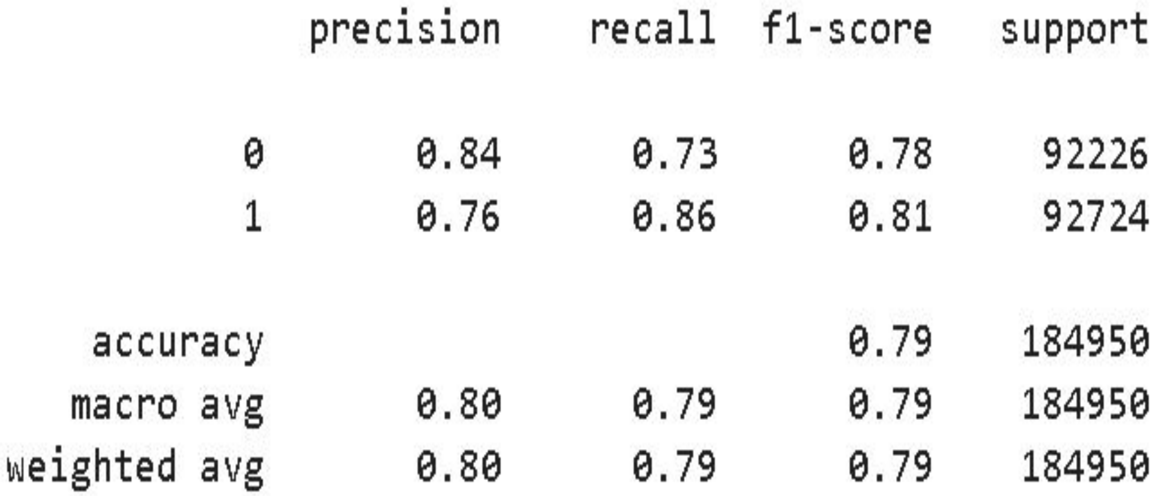
The following metrics/plots were used to analyze the classification models used:

- Accuracy
- F1-score
- Precision
- Recall
- Support
- Confusion Matrix
- ROC curve

Results

Sl.No.	Model Used	Accuracy
1	Logistic Regression	73%
2	Naive Bayes	72%
3	Decision Tree	79%
4	AdaBoost Classifier	76%
5	KNN	74%
6	Neural Network	74%

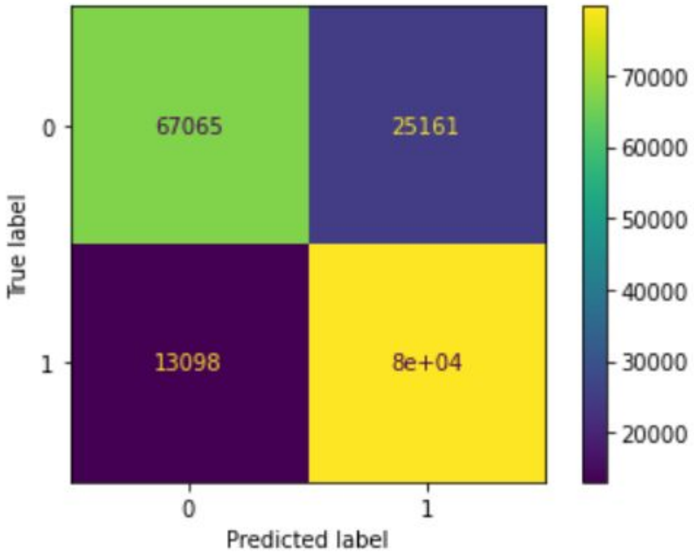
Evaluation metrics used for decision tree



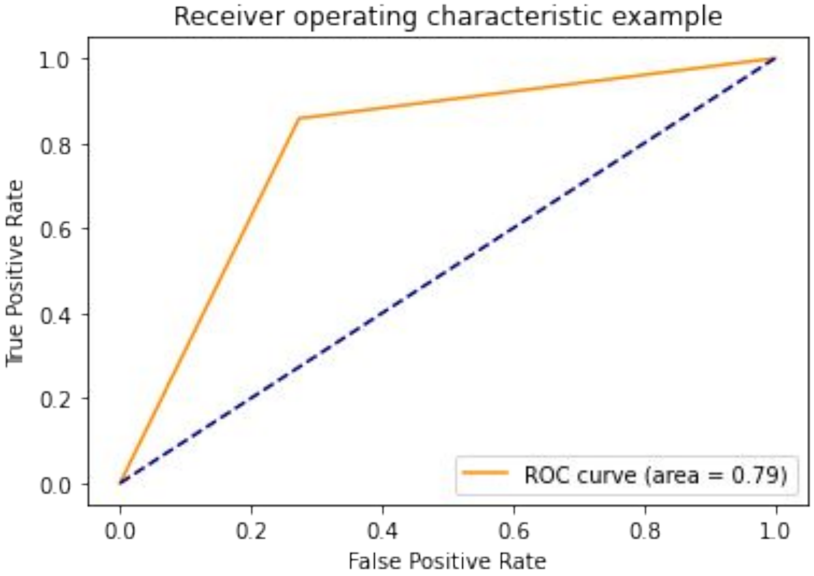
```
              precision    recall  f1-score   support

           0       0.84      0.73      0.78     92226
           1       0.76      0.86      0.81     92724

    accuracy                           0.79    184950
   macro avg       0.80      0.79      0.79    184950
weighted avg       0.80      0.79      0.79    184950
```

Classification Report

Evaluation metrics used for decision tree (contd.)



Confusion Matrix



ROC Curve

Conclusion and Future Scope

The future scope includes fine tuning the parameters of the final classification model to improve its accuracy and performance.

In addition, we can plan on implementing the classification model for other chronic diseases along with leveraging the data from previous years.

References

1) https://www.cdc.gov/brfss/annual_data/2021/pdf/2021-calculated-variables-version4-508.pdf
2) https://imbalanced-learn.org/stable/over_sampling.html#from-random-over-sampling-to-smote-and-adasyn
3) https://www.w3schools.com/python/python_ml_confusion_matrix.asp
4) https://python.plainenglish.io/how-to-use-pandas-profiling-on-google-colab-e34f34ff1c9f
5) https://www.w3schools.com/python/python_ml_confusion_matrix.asp
6) https://scikit-learn.org/stable/auto_examples/model_selection/plot_roc.html