Chronic Heart Disease Prediction

Course Name: Data Mining
Course Code: CSCI B565
Professor: Yuzhen Ye

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

Abstract:

According to the recent survey, it is found that heart disease is the leading cause of death in the United States. Almost half of all Americans are at risk for heart disease because of one of the following risk factors: High Blood pressure, High cholesterol levels, Smoking, Obesity, Insufficient physical activity and excessive alcohol consumption. One major challenge is the detecting of the disease. Heart disease forecasting tools are available, but they are either expensive or inefficient at estimating the likelihood of heart disease.

In this project, we will use algorithms to forecast potential heart diseases in people. Logistic Regression, Naive Bayes, Decision Tree, AdaBoost Classifier, KNN and Neural Networks are the algorithms used. We will examine heart disease prediction models with a larger set of input attributes.

The system uses attributes, including medical terms like Sex, Age, Blood Pressure, and Cholesterol, to predict the likelihood that a patient will develop heart disease. Through the use of classification models, the model is able to determine whether a person could have heart disease or not, enabling early detection and treatment while also reducing medical expenses.

Keywords:

Logistic Regression - Naive Bayes - Decision Tree - AdaBoost Classifier – KNN - Neural Networks – Prediction – Accuracy - Data Mining Techniques

Introduction:

The amount of data available today allows us to use data mining techniques to look for hidden patterns. Everyone needs to be aware of the seriousness and risks of heart disease as these problems become more severe. This project forecasts the risk of developing heart disease in order to reduce such risks.

The model is based on the BRFSS survey data. 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.

The user's criteria can determine how the rules are prioritized. The performance of the system is evaluated on the basis of classification accuracy and the results show that the system has great potential in predicting the heart disease risk level more accurately.

We will evaluate the accuracy of each algorithm used to predict heart disease.

Methods:

Preprocessing: Data preprocessing is a phase in the data mining and data analysis process that converts raw data into a format that computers and machine learning algorithms can understand and evaluate. Garbage in, garbage out is a saying that is frequently used while training machine learning models utilizing data sets. This implies that if you train your model using faulty or "dirty" data, the model will be poor and inadequately trained and won't truly be useful for your research.

These are the following pre-processing methods we have applied to our dataset:

1) 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 record wouldn't impact the distribution of your dataset. It doesn't provide any value if the entire row contains NaN values. Therefore, such rows and columns must be promptly removed.
2) Encoding: We used one-hot encoding for converting the data to prepare it for an algorithm and get a better prediction. When the characteristics are nominal, we employ this method of categorical data encoding (do not have any order). We produce a new variable in a single hot encoding for each level of a category feature. A binary variable with the values 0 or 1 is assigned to each category. In this case, 0 denotes the lack of that category while 1 denotes its existence.
3) Noisy data: Noisy data is useless information that computers are unable to understand. Poor data collection, data entry issues, and other factors may be to blame. For this dataset, we removed values which were refused/don't know/not asked to the person.

Data Visualization: The graphic display of information and data is known as data visualization. Data visualization tools offer an easy approach to observe and analyze trends, outliers, and patterns in data by utilizing visual components like charts, graphs, and maps.

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

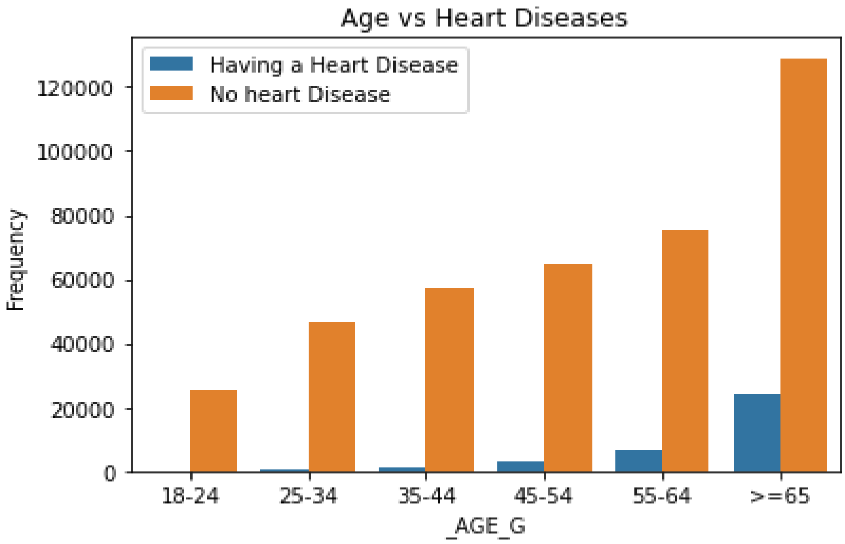


Figure 1: Age vs Heart Diseases Graph

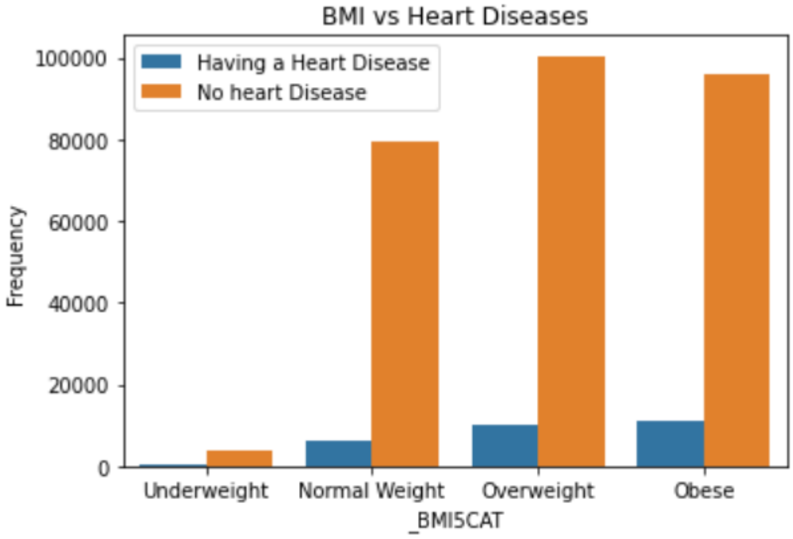


Figure 2: BMI vs Heart Diseases Graph

By analysing the data for these variables, it is evident that 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%

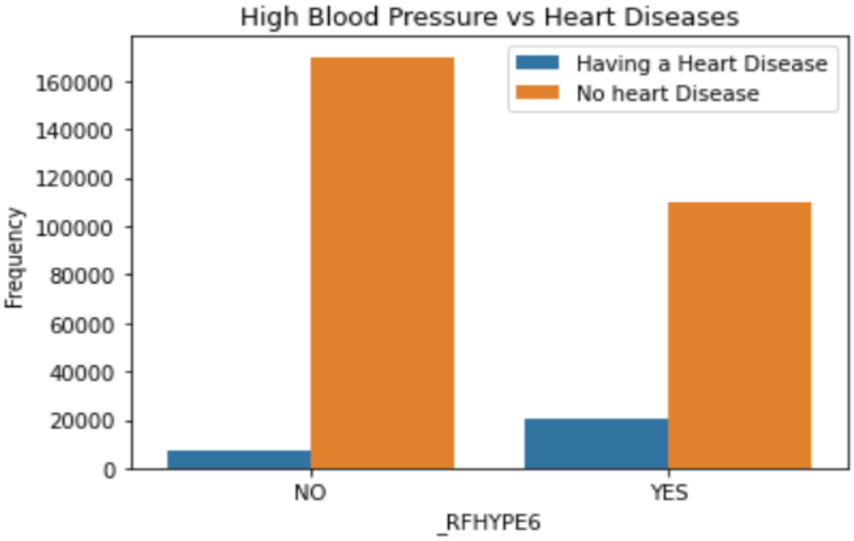


Figure 3: High Blood Pressure vs Heart Diseases Graph

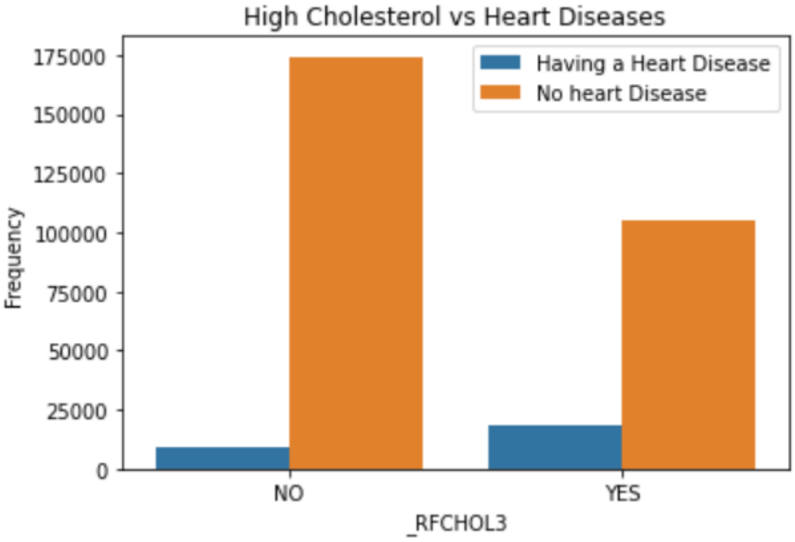


Figure 4: High Cholestrol vs Heart Diseases Graph

By analysing the data for these variables, we observe that 12.6% of people having heart disease are found to have normal cholesterol level and no high blood pressure.

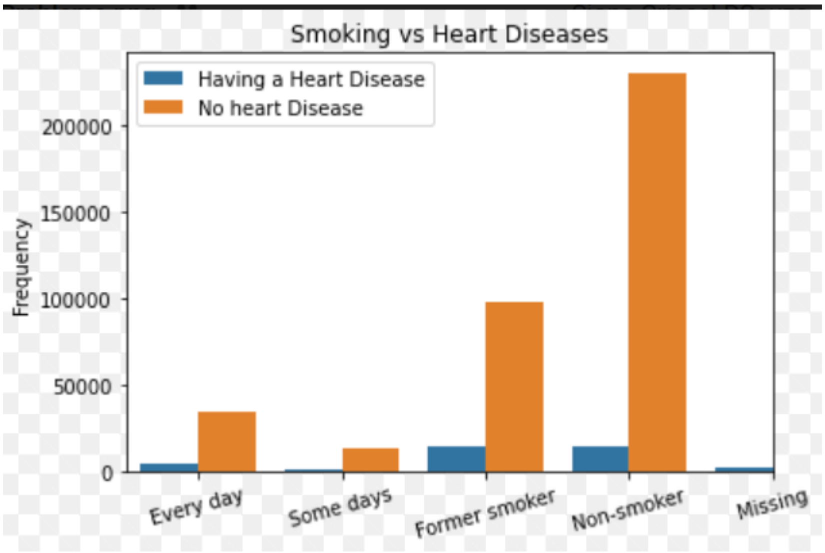


Figure 5: Smoking vs Heart Diseases Graph

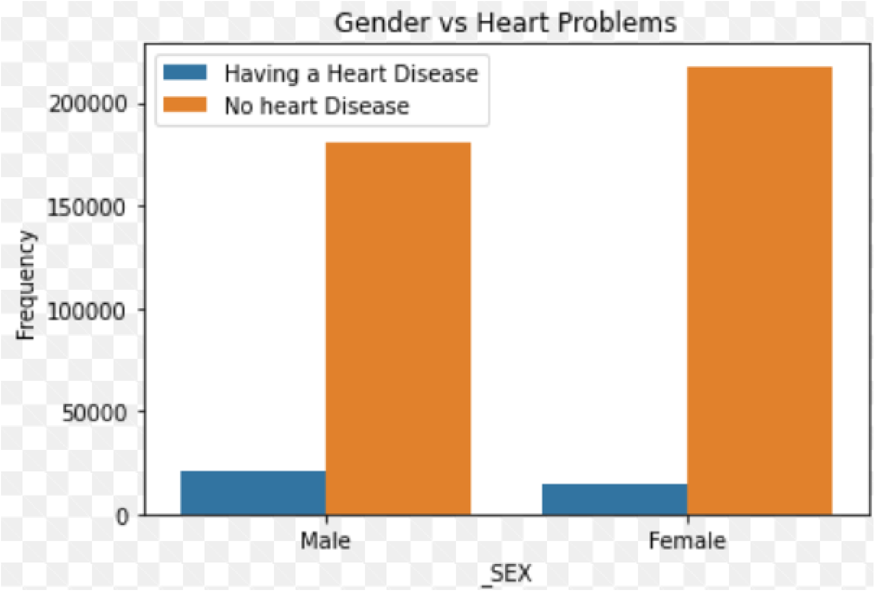


Figure 6: Gender vs Heart Diseases Graph

Classification/Various models used:

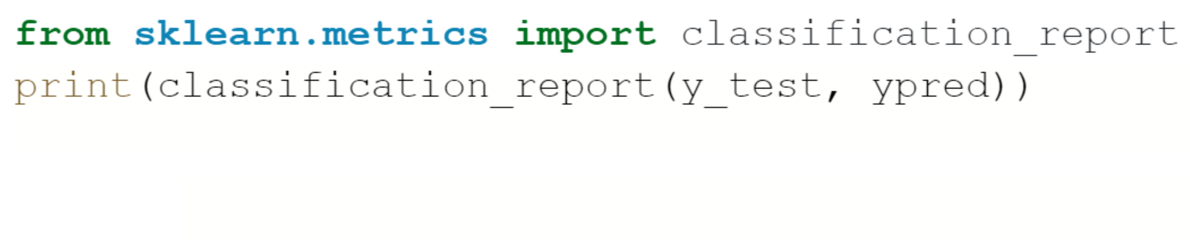
Sampling: To find patterns and trends in the bigger data set being reviewed, data sampling is a statistical analysis approach that is used to choose, modify, and analyze a representative selection of data points.For this project, we performed oversampling using ADASYN. Producing a suitable number of synthetic alternatives for each observation that belongs to the minority class is the main goal of ADASYN. An observation from the minority class is particularly "hard to learn" if there are several instances from the majority class that share characteristics with that observation (i.e. if drawn in the features space, an hard observation looks surrounded by elements from the majority class, as shown in the image below).

Models used:

1. Logistic Regression : Predictive analytics and categorization frequently make use of this kind of statistical model, commonly referred to as a logit model. Based on a given dataset of independent variables, logistic regression calculates the likelihood that an event will occur, such as voting or not voting. The following formulae are used to express this logistic function: Logit(pi) = 1/(1+ exp(-pi))
2. Naive Bayes: It is a classification method built on the Bayes Theorem and predicated on the idea of predictor independence. A Naive Bayes classifier, to put it simply, believes that the existence of one feature in a class has nothing to do with the presence of any other feature.
3. Decision Tree: A decision tree uses a tree structure to develop classification or regression models. It progressively develops an associated decision tree while segmenting a dataset into smaller and smaller sections. The outcome is a tree containing leaf nodes and decision nodes.
4. Bagging Classifier: A kind of ensemble machine learning method called bagging combines the results from several learners to enhance performance. The way these algorithms work is to divide the training set into smaller subsets, put each subset through a different machine learning model, and then combine the results to produce a final prediction for each instance in the original data.
5. AdaBoost Classifier: To improve classifier accuracy, this algorithm combines a number of weak classifiers. An iterative ensemble algorithm is AdaBoost. AdaBoost classifier combines a number of ineffective classifiers to create a powerful classifier that has a high degree of accuracy. The fundamental idea underlying Adaboost is to train the data sample and adjust the classifier weights in each iteration in a way that provides accurate predictions of uncommon observations. Any machine learning method that accepts weights from the training set can be used as the basis classifier.
6. KNN: The k-nearest neighbors algorithm, sometimes referred to as KNN or k-NN, is a supervised learning classifier that employs proximity to produce classifications or predictions about the grouping of a single data point. Although it may be applied to classification or regression issues, it is commonly employed as a classification method since it relies on the idea that comparable points can be discovered close to one another.
7. Neural Network: An input vector is transformed into an output by layers of units (neurons) that make up a neural network. Each unit receives an input, processes it through a (often nonlinear) function, and then transfers the result to the following layer. In general, networks are characterised as being feed-forward, meaning that there is no feedback to the previous layer and each unit feeds its output to all the units on the layer above it. Signals travelling from one unit to another are given weightings, and it is these weightings that are tweaked throughout the training phase to adapt a neural network to the specific issue at hand.

Results:

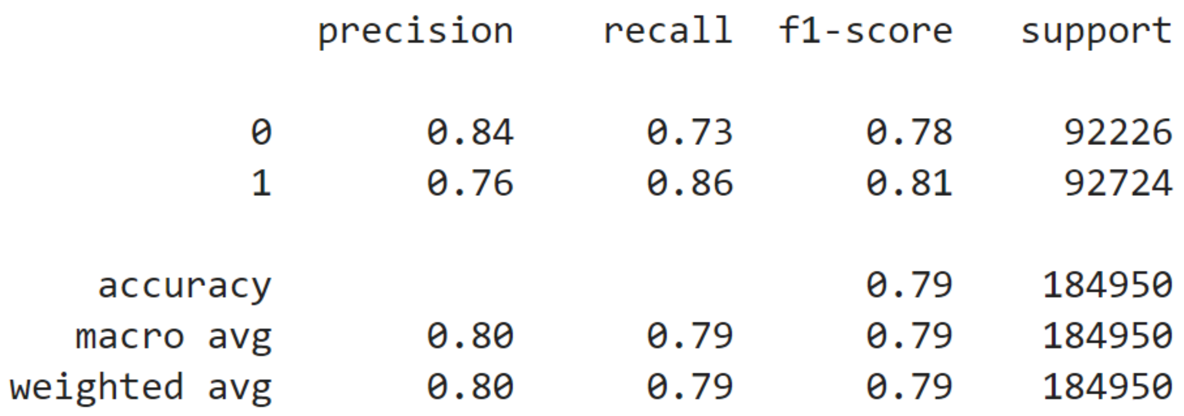
For choosing the best model, a performance evaluation metric has to be used. For this specific implementation, accuracy was used.



```
from sklearn.metrics import classification_report
print(classification_report(y_test, ypred))
```

Figure 7: Code to demonstrate classification_report

Figure 7 is the snippet from the code that shows classification report imported from scikit learn. It outputs the data about various evaluation metrics along with accuracy, the one we are particularly interested in. y_test and ypred represent y values of testset and predicted y values by the model respectively.



```
              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
```

Figure 8: Output for the decision tree's classification report

Figure 8 shows accuracy among other metrics to be as 0.79 i.e. 79% accuracy for Decision Tree classifier.

In order to understand more about the performance of the specific classifier, confusion matrix and ROC curve were also extracted.

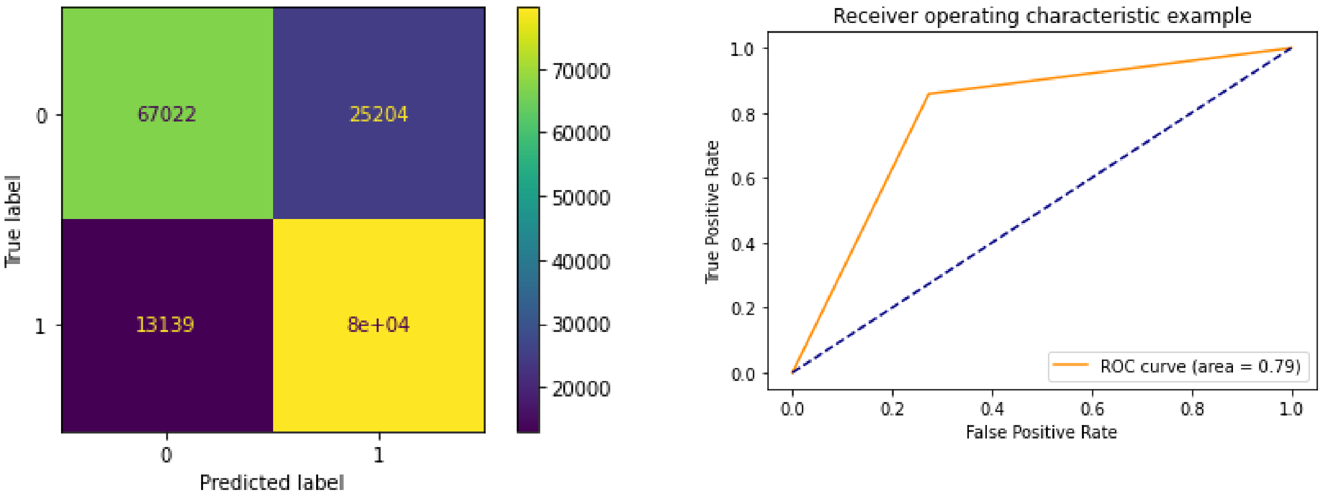
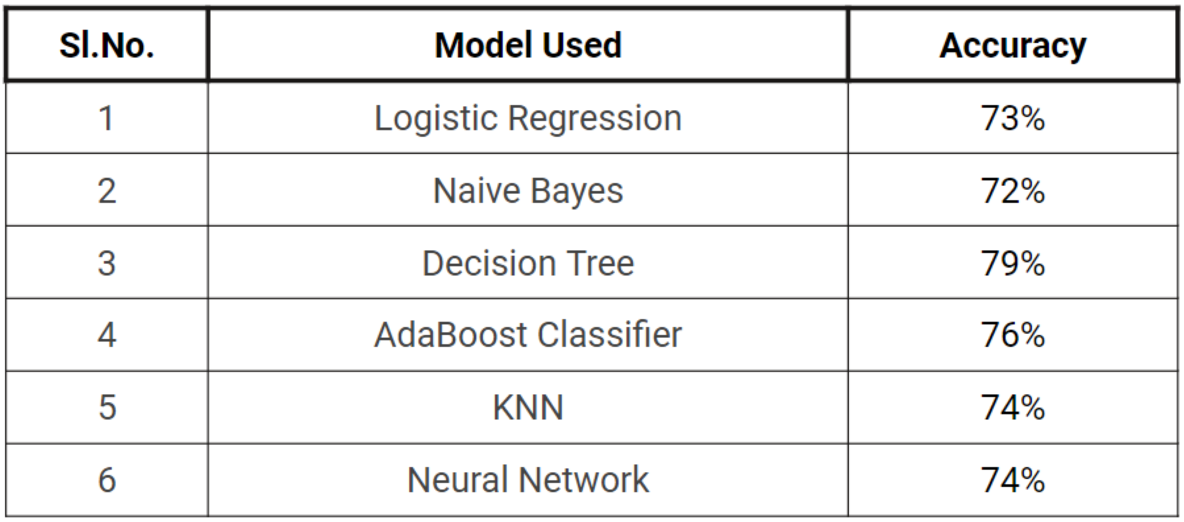


Figure 9a (left) and Figure 9b (right) shows Confusion Matrix and ROC curve for Decision Tree Classifier respectively



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%

Figure 10: Comparison of various classification models

Among all the models used for analysis as shown in Figure 10, Decision Tree classifier had the highest accuracy of 79% and hence was the obvious choice.

Discussion and Future work:

Although we proceeded with the Decision Tree Classifier, the other models were not far off in terms of accuracy. The major problem encountered in terms of prediction is the highly class unbalance of the data. Using oversampling, that problem is effectively addressed improving the implementation of the models.

With the help of the finalised model, given the values related to a person i.e. values that correspond to the features considered, it is possible to predict whether the person could potentially be affected by a chronic heart disease or not. This can be further used in understanding more about the risk of the heart disease in a larger sample setting helping decision makes efficiency.

Further more experimentation in terms of hyper parameter tuning can increase the model accuracy and performance. In addition, using the BRFSS data from previous years by figuring out a decent strategy on how to approach the different number of features collected for each year can uncover useful insights and inturn improve the current approach. The true potential of this prediction can be truely discovered if combined with other forms of heart disease prediction approaches.

References:

1. https://www.cdc.gov/brfss/annual_data/2021/pdf/2021-calculated-variables-version4-508.pd
2. https://python.plainenglish.io/how-to-use-pandas-profiling-on-google-colab-e34f34ff1c9f
3. https://imbalanced-learn.org/stable/over_sampling.html#from-random-over-sampling-to-smote-and-adasyn
4. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.classification_report.html
5. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.ConfusionMatrixDisplay.html
6. https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html
7. https://www.w3schools.com/python/python_ml_confusion_matrix.asp
8. https://scikit-learn.org/stable/auto_examples/model_selection/plot_roc.html
9. https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html
10. https://www.tensorflow.org/guide/keras/sequential_model
11. https://www.ibm.com/topics/logistic-regression
12. https://scikit-learn.org/stable/visualizations.html
13. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5863635/