Heart Disease: Detection and Prevention

Winter 2023

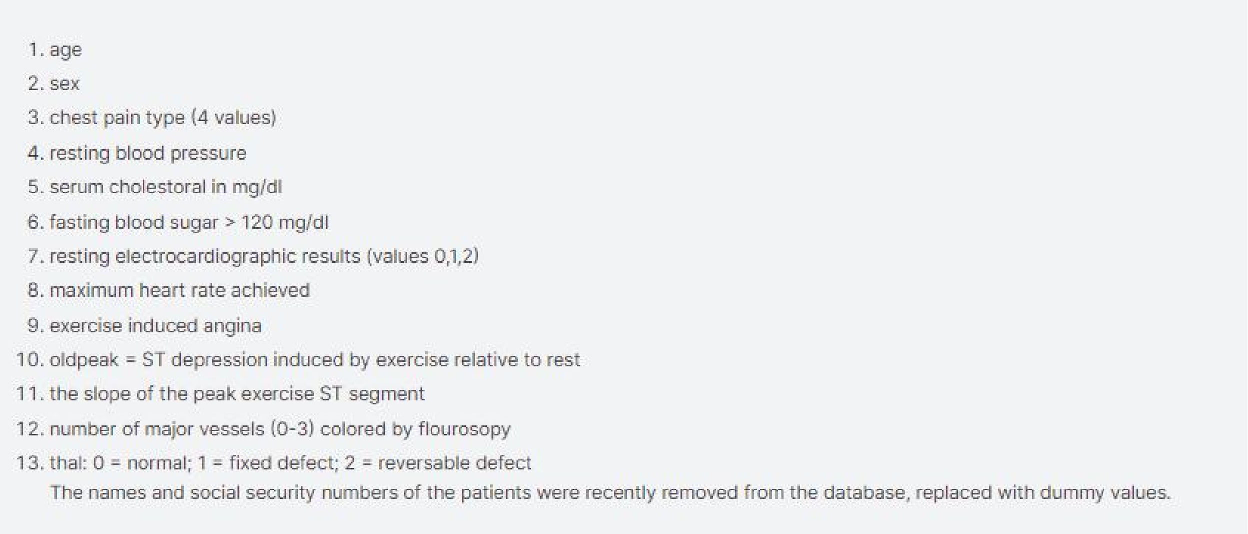
Will Wu, Drexel University

1. INTRODUCTION

The term “heart disease” refers to several types of heart conditions [1]. The most common type of heart disease in the United States is coronary artery disease (CAD), which affects the blood flow to the heart. Decreased blood flow can cause a heart attack. In 2020, about 697,000 people in the United States died from heart disease, that’s 1 in every 5 deaths. Heart disease is the leading cause of death for men, women, and people of most racial and ethnic groups in the United States. Some of the risk factors for heart disease include high blood pressure, high blood cholesterol, and smoking. In paper, we will work with an existing medical dataset and use machine learning models in hopes to find correlation between the different factors in detecting heart disease and thus help to prevent it.

2. DATASET

The dataset [3] we will be working with contains medical information that dates from 1988 and consists of four databases, Cleveland, Hungary, Switzerland and Long Beach V. It contains 12 possible predictive attributes and a “target” attribute which refers to the presence of heart disease in the patient. All columns are numerical, thus we will need to refer to the feature description to make sense of our findings. Figure 1 shows the description of each feature. Each row describes one patient’s information and there are a total of 1025 rows. Table 1 shows a sample collection of the data.



1. age
2. sex
3. chest pain type (4 values)
4. resting blood pressure
5. serum cholestoral in mg/dl
6. fasting blood sugar > 120 mg/dl
7. resting electrocardiographic results (values 0,1,2)
8. maximum heart rate achieved
9. exercise induced angina
10. oldpeak = ST depression induced by exercise relative to rest
11. the slope of the peak exercise ST segment
12. number of major vessels (0-3) colored by flourosopy
13. thal: 0 = normal; 1 = fixed defect; 2 = reversable defect

The names and social security numbers of the patients were recently removed from the database, replaced with dummy values.

Figure 1: Description of the features

age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
52	1	0	125	212	0	1	168	0	1.0	2	2	3	0
53	1	0	140	203	1	0	155	1	3.1	0	0	3	0
70	1	0	145	174	0	1	125	1	2.6	0	0	3	0
61	1	0	148	203	0	1	161	0	0.0	2	1	3	0
62	0	0	138	294	1	1	106	0	1.9	1	3	2	0

Table 1: Sample collection of the data

3. EXPLORATORY DATA ANALYSIS

The purpose of exploratory data analysis (EDA) is to investigate the dataset and to further understand the features and their relation to each other as well as to the target variable, without the use of machine learning algorithms. With EDA, we could look at the data before making any assumptions which helps with identifying errors, outliers, or anomalies. Figure 2 contains the distribution of all features, it provides an overview on the distribution of our data. Some notable things from this visualization will be 1) there are about twice the number of female patient data than male and 2) the target variable which is the identifier for heart disease, is approximately split 50/50.

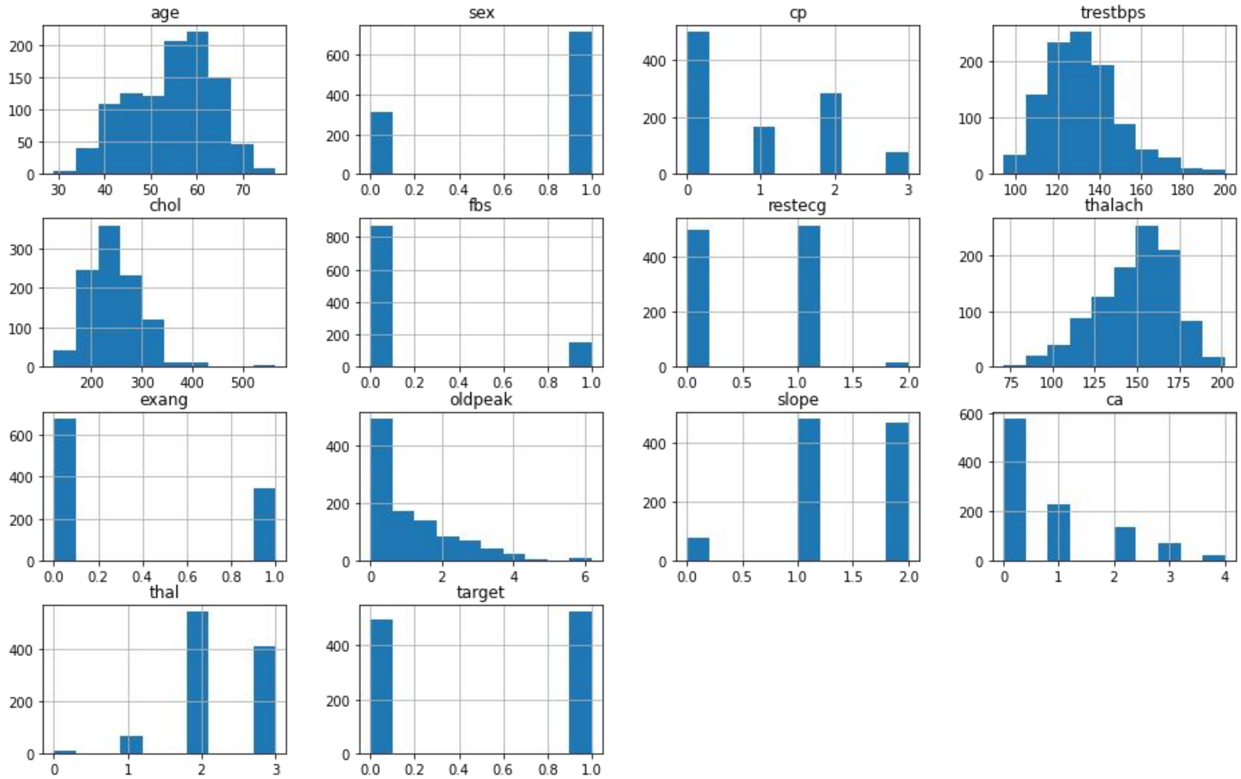


Figure 2: Distribution of all features

We can further investigate features and their relation to the target variable. Figure 3 shows the distribution of heart disease in relation to gender. We can see that although the number of female patients with heart disease is greater than that of male patients, the percentage of male patients with heart disease is largely greater than that of female patients. There are more than double the number of male patients with heart disease than male patients without heart disease.

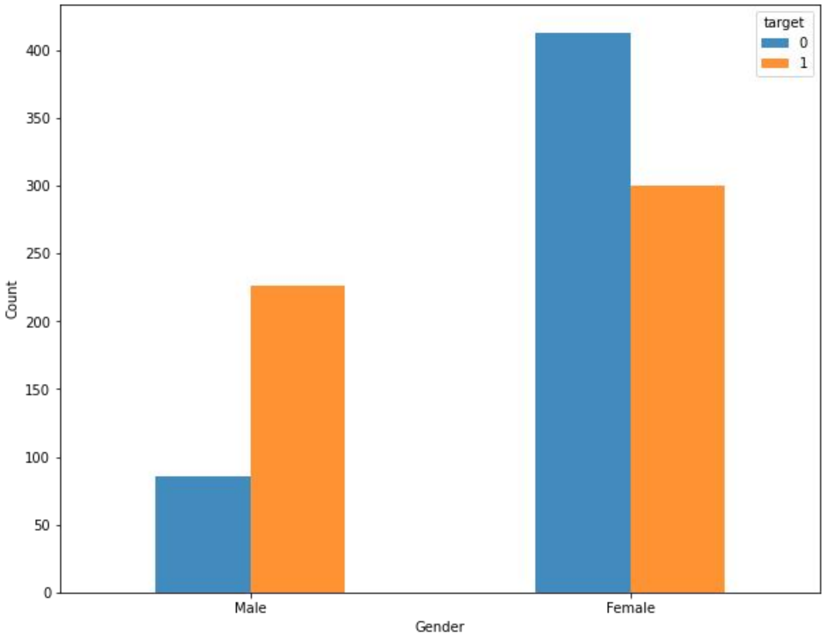


Figure 3: Distribution of heart disease vs. Gender

Another interesting relation was found by visualizing the distribution of heart disease in relation to age and the maximum heart rate (figure 4). A scatter plot was made with the x-axis as gender, the y-axis as the maximum heart rate(thalach), in relation to the target = 1 and target = 0. From the visualization, we can see that most patients with heart disease have a higher maximum heart rate (approximately above 140). We could also see that most patients with heart disease are between the age of 40 and 60, although the correlation isn't as significant as that of the heart rate.

Finally, we were able to examine the correlation between all features (including the target feature) by creating a visualization of the correlation heatmap (figure 5). The brighter the color, the higher the correlation (1 being the highest correlation, between the feature and itself). From the correlation heatmap, we're able to detect a couple features that have a strong correlation with the target feature: cp (chest pain type), thalach (maximum heart rate), exang(exercise induced angina), and oldpeak(ST depression induced by exercise).

After EDA, we are able to have an overall understanding of our dataset. We are now ready to utilize data algorithms in attempt to detect and predict heart disease existence. However, in order to apply machine learning algorithms, we'll need to preprocess which includes splitting our data and make sure everything is in line to set up for use of those algorithms.

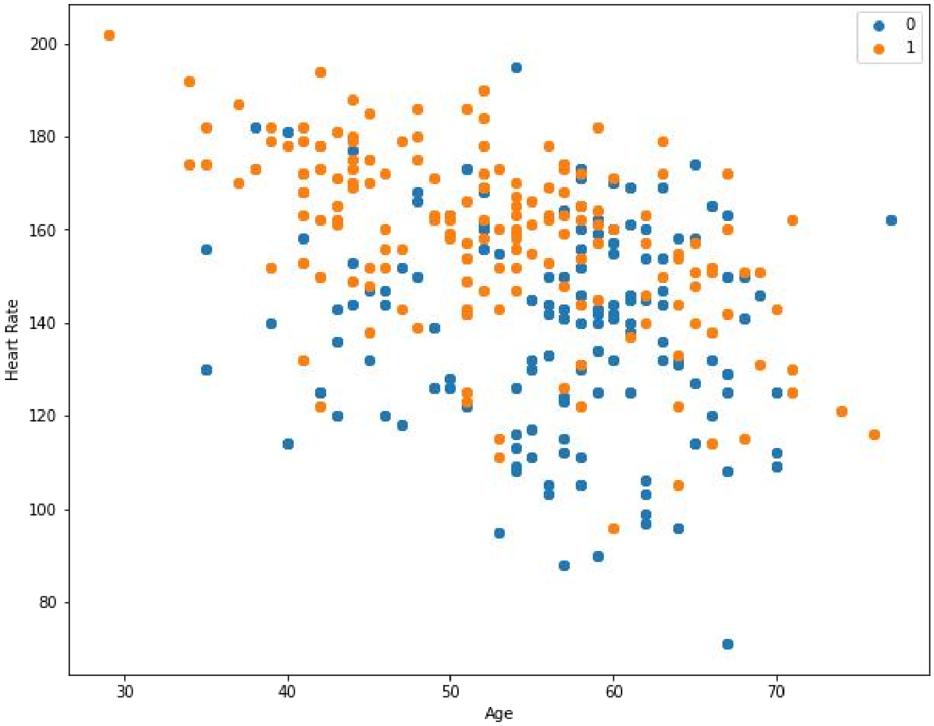


Figure 4: Distribution of heart disease vs. Age & Heart Rate

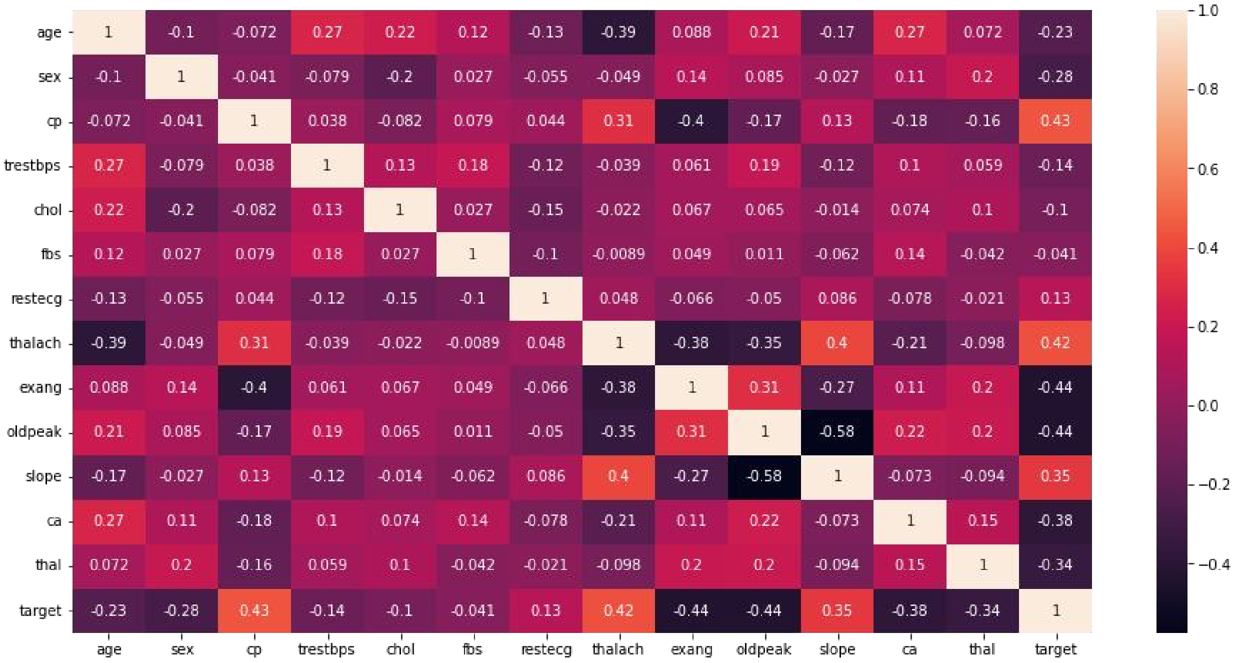


Figure 5: Correlation Heatmap between all features

4. METHODOLOGY

4.1 Preprocessing

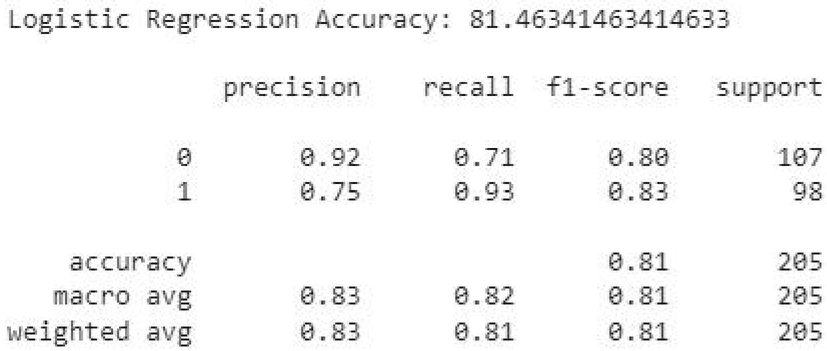
Preprocessing is an important and necessary step before utilizing machine learning algorithms to train and test the dataset. It involves (but not limited to) procedures such as dealing with null data, encoding categorical data, dealing with skewed data, feature selection, and scaling our dataset. In our dataset, we have no null data, no categorical features, the distribution of all features are not skewed, and we don't have a large number of features. Thus the only preprocess procedure we will implement will be scaling. Scaling our features is a critical step before creating machine learning models. Especially with our dataset, the unit and range of our values in each of the features are quite different. Thus we will perform standardization which transforms the data to have zero mean and a variance of 1, making our data unitless and uniform.

We will be splitting the data into 'train' and 'test' sets. The 'train' set will be used to train our machine learning model and the 'test' set will test our model for accuracy and performance. The split will be eighty-twenty, with 80% of our data in the train set and 20% of our data in the test set.

4.2 MACHINE LEARNING MODELS

4.2.1 Logistic Regression

The first machine learning model we will implement is logistic regression. It is a commonly used classification model for estimating the probability of an event occurring, which in our case, is the existence of heart disease in a patient. It is a great model to implement first, it is a relatively simple model that will give us a prediction and it helps set a standard for the next models we will implement. Figure 6 shows the model evaluation after testing the model on the test data. Comparing model predictions against the actual target value.

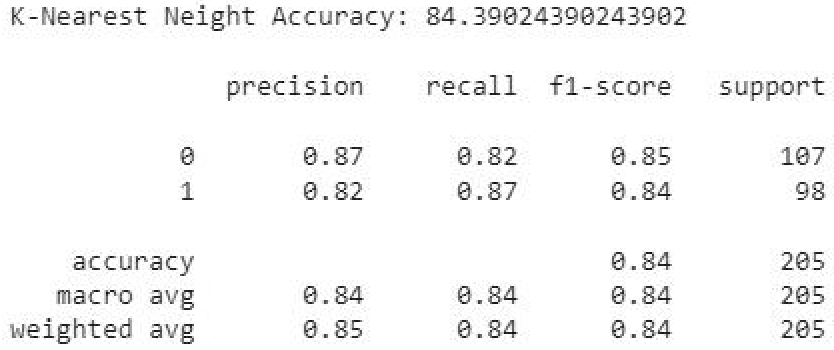


```
Logistic Regression Accuracy: 81.46341463414633

              precision    recall  f1-score   support

           0       0.92      0.71      0.80       107
           1       0.75      0.93      0.83        98

    accuracy                           0.81       205
   macro avg       0.83      0.82      0.81       205
weighted avg       0.83      0.81      0.81       205
```

Figure 6: Logistic Regression Model Evaluation

4.2.2 K-Nearest Neighbor

The second model we will implement is the K-Nearest Neighbor algorithm. The KNN algorithm can compete with the most accurate models because it makes highly accurate predictions. The algorithm works by first locating a starting point x, then selecting the specified number (K) of closest point to x, then resulting in the most frequent target label for the prediction. In our implementation, we chose K to be equal to 10. Figure 7 shows the model evaluation after testing the model on the test data.

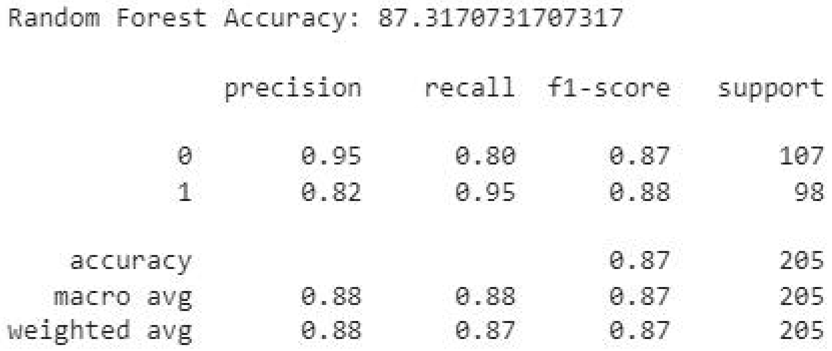


```
K-Nearest Neight Accuracy: 84.39024390243902

              precision    recall  f1-score   support

           0       0.87      0.82      0.85       107
           1       0.82      0.87      0.84        98

    accuracy                           0.84       205
   macro avg       0.84      0.84      0.84       205
weighted avg       0.85      0.84      0.84       205
```

Figure 7: K-Nearest Neighbor Model Evaluation

4.2.3 Random Forest

The last model we implemented was the Random Forest Classifier. It consists of a large number of individual decision trees that operate together. Decision trees model works by starting with the whole dataset, then dividing into branches (two at a time) to separate the classes and to predict the target class. The Random Forest model combines a large number of decision trees thus it will result in better performance. Figure 8 shows the model evaluation after testing the model on the test data.

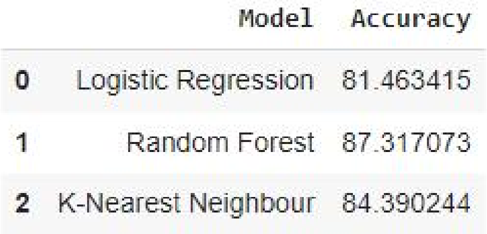


```
Random Forest Accuracy: 87.3170731707317

              precision    recall  f1-score   support

           0       0.95      0.80      0.87       107
           1       0.82      0.95      0.88        98

    accuracy                           0.87       205
   macro avg       0.88      0.88      0.87       205
weighted avg       0.88      0.87      0.87       205
```

Figure 6: Random Forest Model Evaluation

5. Results and Discussion

Figure 7 shows the visualization of the comparison between the results from the three different models we implemented. We received the highest accuracy of 87% from the Random Forest algorithm. The accuracy of 81% from the logistic regression model was surprising. Since the logistic regression method is supposed to be advantageous for datasets that are continuous and the target variable is binary.



	Model	Accuracy
0	Logistic Regression	81.463415
1	Random Forest	87.317073
2	K-Nearest Neighbour	84.390244

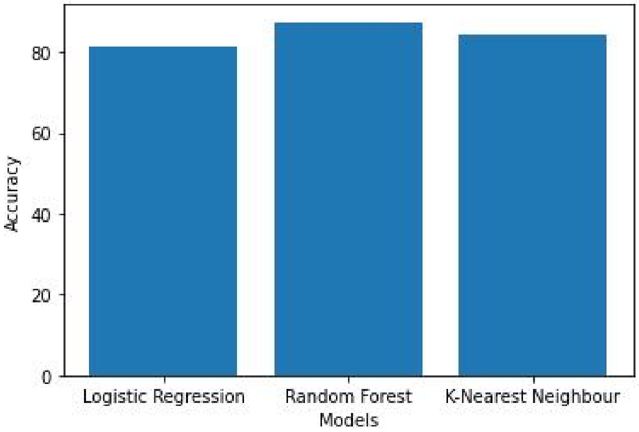


Figure 7: Model Evaluation Comparison

6. Conclusion

In our report, we selected a dataset that combines different sets of patient information from multiple databases with the target variable as the detection of heart disease in the patient. First we took an overall at the data and was able to grasp the overall theme and the meanings behind all the features. Then we performed exploratory data analysis to dive deeper in the dataset and found correlations between the possible predictive features and the target feature. Before we applied machine learning models to train and test on data, we performed the necessary preprocess step which was standardization scaling in order for our features to be unitless and uniform. Finally, we implemented three different machine learning models: Logistic Regression, Random Forest, and K-Nearest Neighbor. We trained our models on 80% of our dataset and tested them on the remaining 20%. We received the highest accuracy value at 87% with the Random Forest model. We could apply this model on any existing or future datasets with the patients information and optimistically predict the presence of heart disease in patients.

REFERENCES

1. Centers for Disease Control and Prevention (2022, July 12). *About Heart Disease*. Retrieved March 14, 2023, from https://www.cdc.gov/heartdisease/about.htm

2. BHAT, N. (2021). *HEART DISEASE ANALYSIS*. Retrieved March 14, 2023, from https://www.kaggle.com/code/nareshbhat/eda-classification-ensemble-92-accuracy#Model-Evaluation

3. Lapp, David. (2019). Heart Disease Dataset. Retrieved March 14, 2023, from https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset?datasetId=216167&sortBy=voteCount&select=heart.csv.