Heart Disease Prediction Model Report

Members: Justin Nichols, Patryk Czerniewicz

CSC 2730 Final Project

Instructor: Jian Zhang

5 December 2021

Table of Contents

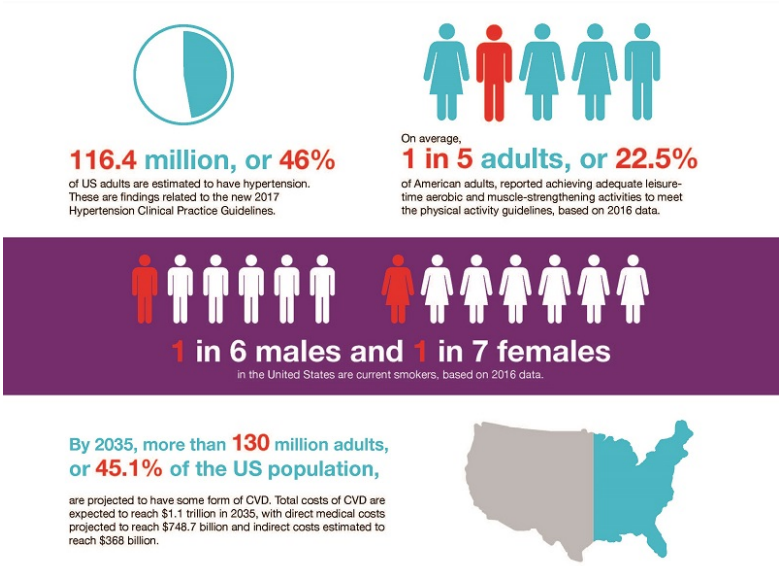
Preface

The ultimate goal of this project was to predict whether a patient has heart disease or not. The data was split into a training and testing set in order to compare the performance accuracy of each of the 5 predictive models used. The main objective was to find the accuracy score of each model and see which predictive analysis algorithm was the most accurate at predicting heart disease based on the data provided. The predictive models used include Linear Regression, Logical Regression, K-Nearest Neighbor, Decision Tree, and SVC. The data used within this predictive application contains patient health data from [Heart Disease UCI].

Motivation

The motivation behind this project is to provide health professionals the ability to more accurately determine whether a patient is showing signs of heart disease based on their personal health data. According to the CDC, the leading cause of death in the United States in 2019 was heart disease, with 659,041 people succumbing to it (CDC). Other than cancer, heart disease has claimed the leading spot of death in America for more than a decade.

To give a brief overview, heart disease refers to several serious health conditions mainly affecting the heart or the cardiovascular system. The three main symptoms of someone with heart disease include a heart attack, heart failure, or an arrhythmia (CDC). High blood pressure, high blood cholesterol, and smoking are key factors for heart disease, and about half (47%) of people in the U.S. have at least one of these risk factors (CDC). Heart disease is most common in men, people who smoke, overweight or obese people, individuals with a family history of it, and people over the age of 55 (Healthline).



American Heart Association

Problem Definition

For more than a decade, heart disease has been the primary cause of death in the United States. Hundreds of thousands of people each year die from heart disease. For many cases, the individual was unaware that he or she was showing signs of heart disease until after they had suffered a major health problem. To solve this problem, our group wanted to create a heart disease predictive model that will aid doctors, nurses, and other health professionals to determine whether a patient is likely to suffer from heart disease based on their current health status. With this prediction, a physician can further access the patient and determine whether further medical action is necessary such as medical scans and tests, medication, or lifestyle changes. By using 5 different predictive models, this project aims to find the most accurate prediction to use for a real-world setting.

Dataset

The dataset contains patient health data and was obtained from Kaggle using the link Heart Disease UCI. It is stored in a CSV formatted file. The dataset contains 303 unique entries. The attributes of the dataset include age, sex, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal, and target. Further explanation of each attribute is provided below. It is important to note that all participating individuals in this study consented under the Health Insurance Portability and Accountability Act of 1996 (HIPAA). Although the individuals consented to have their health data a part of this dataset, for privacy reasons, the names and social security numbers for each patient were removed.

Attributes

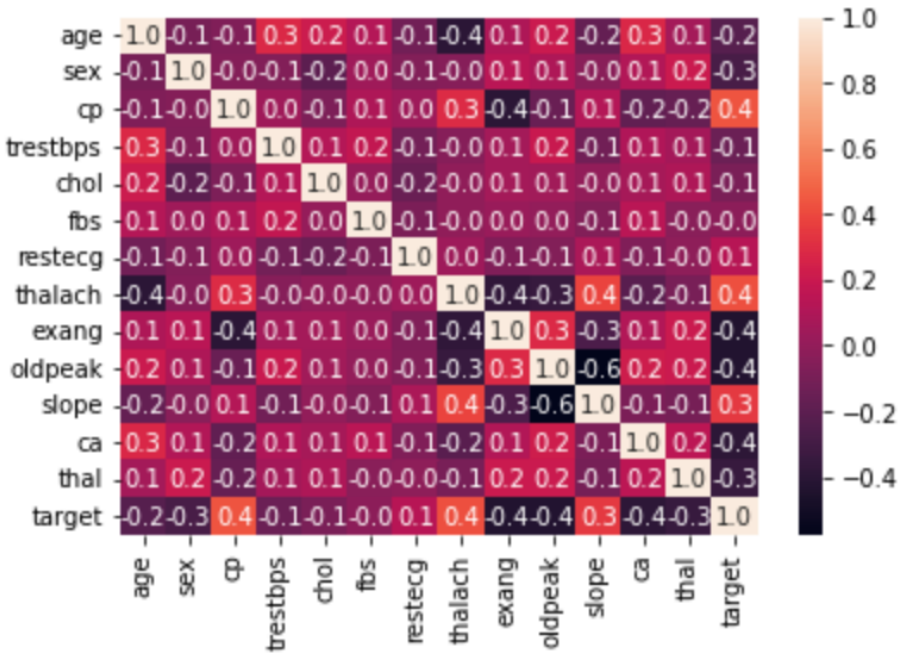
The dataset used contains 14 different attributes for 303 patients who were examined for heart diseases. Attribute 14 (target) is used for determining the prediction accuracy of a certain model. The attributes are:

1. Age - varying from 29 to 77, mean age is 54 with a standard deviation of 9 years
2. Sex - 1 for men, 0 for women
3. Chest pain type - 4 different values
4. Resting blood pressure - mean and median are around 130
5. Serum cholesterol in mg/dl - mean is 246, median is 240
6. Fasting blood sugar - 1 if value is larger than 120 mg/dl, 0 otherwise
7. Resting electrocardiographic results - 3 values
8. Maximum heart rate achieved
9. Presence of exercise induced angina - 1 if exists, 0 otherwise
10. Old peak - ST segment depression induced by exercise relative to rest
11. Slope of the peak exercise ST segment
12. Number of major vessels colored by fluoroscopy (ca) - 4 values (0-3)

13. Thalassemia - 3 = normal, 6 = fixed defect, 7 = reversible defect
14. Target (confirms whether a patient has heart disease or not)

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1

Attribute Correlations

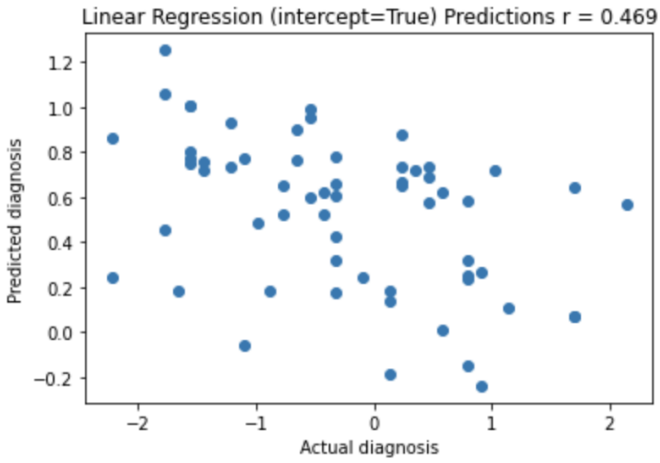


Data Analysis Algorithms

The 5 predictive models used include Linear Regression, Logical Regression, K-Nearest Neighbor, Decision Tree, and SVC. The data was split into two sets, training and testing, with the training set comprised of 80 percent of the data and the testing set containing 20 percent.

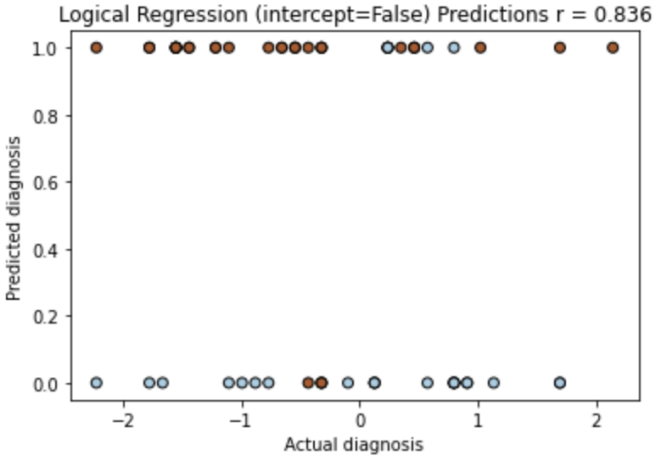
Linear Regression

For Linear Regression, the hyper-parameter fit_intercept was set to true. The Linear Regression algorithm yielded an accuracy score of 46.9%.



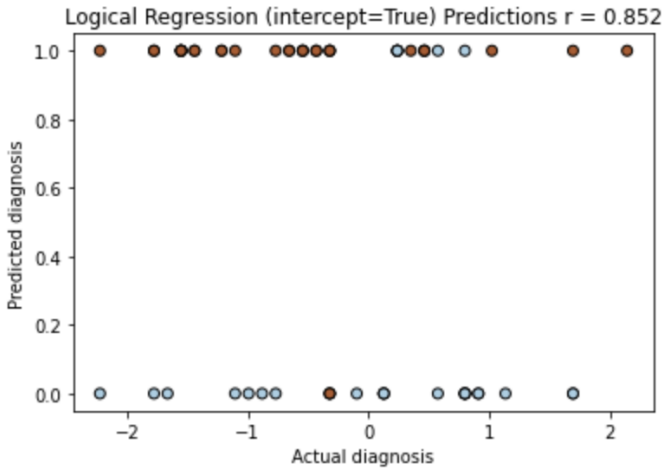
Logical Regression

For Logical Regression, the hyper-parameter fit_intercept was set to false and true. For the different fit_intercept values, the Logical Regression algorithm yielded an accuracy score of 83.61% and 85.25% respectively.



	precision	recall	f1-score	support
0	0.84	0.78	0.81	27
1	0.83	0.88	0.86	34
accuracy			0.84	61
macro avg	0.84	0.83	0.83	61
weighted avg	0.84	0.84	0.84	61

0.8360655737704918



	precision	recall	f1-score	support
0	0.88	0.78	0.82	27
1	0.84	0.91	0.87	34
accuracy			0.85	61
macro avg	0.86	0.84	0.85	61
weighted avg	0.85	0.85	0.85	61

0.8524590163934426

KNN

For K-Nearest Neighbor, the hyper-parameter n-neighbors was set to 5, 8, and 10. For the different n-neighbors values, the KNN algorithm yielded an accuracy score of 81.97%, 85.25%, and 88.52% respectively.

	precision	recall	f1-score	support
0	0.81	0.78	0.79	27
1	0.83	0.85	0.84	34
accuracy			0.82	61
macro avg	0.82	0.82	0.82	61
weighted avg	0.82	0.82	0.82	61

0.819672131147541

	precision	recall	f1-score	support
0	0.82	0.85	0.84	27
1	0.88	0.85	0.87	34
accuracy			0.85	61
macro avg	0.85	0.85	0.85	61
weighted avg	0.85	0.85	0.85	61

0.8524590163934426

```
                precision    recall  f1-score   support

           0       0.86      0.89      0.87        27
           1       0.91      0.88      0.90        34

    accuracy                           0.89        61
   macro avg       0.88      0.89      0.88        61
weighted avg       0.89      0.89      0.89        61

0.8852459016393442
```

Decision Tree

For Decision Tree, the hyper-parameter criterion was set to gini and entropy. For the different criterion values, the Decision Tree algorithm yielded an accuracy score of 75.41% and 77.05% respectively.

```
                precision    recall  f1-score   support

           0       0.69      0.81      0.75        27
           1       0.83      0.71      0.76        34

    accuracy                           0.75        61
   macro avg       0.76      0.76      0.75        61
weighted avg       0.77      0.75      0.75        61

0.7540983606557377
```

```
              precision    recall  f1-score   support

           0       0.71      0.81      0.76        27
           1       0.83      0.74      0.78        34

    accuracy                           0.77        61
   macro avg       0.77      0.78      0.77        61
weighted avg       0.78      0.77      0.77        61

0.7704918032786885
```

SVC

For SVC, the hyper-parameters kernel and C were altered. Kernel was set to linear for all predictions, however, C was set to 0.001, 0.01, 0.1, and 1. For the different C values, the SVC algorithm yielded an accuracy score of 65.57%, 83.61%, 80.33%, and 81.97% respectively.

	precision	recall	f1-score	support
0	1.00	0.22	0.36	27
1	0.62	1.00	0.76	34
accuracy			0.66	61
macro avg	0.81	0.61	0.56	61
weighted avg	0.79	0.66	0.59	61

0.6557377049180327

	precision	recall	f1-score	support
0	0.90	0.70	0.79	27
1	0.80	0.94	0.86	34
accuracy			0.84	61
macro avg	0.85	0.82	0.83	61
weighted avg	0.85	0.84	0.83	61

0.8360655737704918

	precision	recall	f1-score	support
0	0.83	0.70	0.76	27
1	0.79	0.88	0.83	34
accuracy			0.80	61
macro avg	0.81	0.79	0.80	61
weighted avg	0.81	0.80	0.80	61

0.8032786885245902

	precision	recall	f1-score	support
0	0.83	0.74	0.78	27
1	0.81	0.88	0.85	34
accuracy			0.82	61
macro avg	0.82	0.81	0.81	61
weighted avg	0.82	0.82	0.82	61

0.819672131147541

Conclusion

Conclusively, for the heart disease predictive model, K-Nearest Neighbor (n_neighbors=10), proved to be the most accurate and Linear Regression proved to be the least accurate. For each predictive model, different hyper-parameters were used and the accuracy score was found. Below is a list summarizing the performance of all the models dependent on the different hyper-parameters.

The overall goal of the project was to classify whether a patient has heart disease (1) or not (0). The prediction made by the Linear Regression model was the worst because it deals with continuous values and is not ideal for classification problems mandating discrete values. The prediction made by the K-Nearest Neighbor (n_neighbors=10) was the best because it is meant for classifying data into groups, i.e., whether heart disease (1) is present or not (0). The topmost accurate prediction models (K-Nearest Neighbor, Logical Regression, and SVC) are all classification models, which explains why they performed well.

To further improve the predictions for the heart disease predictive model, the different hyper-parameters and new values can be further explored and altered. Considering K-Nearest Neighbor proved to be the most accurate, it would be ideal to test whether changing other hyper-parameters would better the accuracy score from high 80% into greater than 90%.

```
Models sorted by accuracy score:
K-Nearest Neighbor: n_neighbors=10: 88.52%
K-Nearest Neighbor: n_neighbors=8: 85.25%
Logical Regression: fit_intercept=True: 85.25%
SVC: C=0.01: 83.61%
Logical Regression: fit_intercept=False: 83.61%
SVC: C=1: 81.97%
K-Nearest Neighbor: n_neighbors=5: 81.97%
SVC: C=0.1: 80.33%
Decision Tree: criterion='entropy': 77.05%
Decision Tree: criterion='gini': 75.41%
SVC: C=0.001: 65.57%
Linear Regression: fit_intercept=True: 46.90%
```

How to Run the Predictive Models

1. Go to Google Colab and upload the Jupyter notebook (i.e. heart_disease_prediction.ipynb).
2. If the dataset does not automatically load into the notebook, it will need to be manually added to Google Colab using the folder icon on the left-hand side.
3. Click “Runtime” from the navigation bar and hit “Run all” to execute all the cells within the notebook.
4. After all cells are executed, the 5 predictive models, their plots, and their accuracy scores will be generated.

Sources

https://www.kaggle.com/ronitf/heart-disease-uci
https://www.cdc.gov/nchs/fastats/leading-causes-of-death.htm
https://www.cdc.gov/heartdisease/about.htm
https://www.healthline.com/health/leading-causes-of-death
https://www.heart.org/en/news/2019/01/31/cardiovascular-diseases-affect-nearly-half-of-american-adults-statistics-show