Implementation of Python for Data Analysis:

Examining Heart Disease Data

Joshua Cook

Dept. of Economics and Decision Sciences, Western Illinois University

DS 490: Statistics with Python

Dr. Chauhan

May 03, 2020

Abstract

Python is a powerful programming language that can be used for data analysis. The purpose of this project is to utilize a variety of Python skills covered throughout the course to examine a real-world dataset. This will be done utilizing the various Python libraries to assist with the data analysis. The dataset chosen for this project details information about the occurrence of heart disease in individuals for a specific population. The libraries used throughout this project include NumPy, Pandas, Scikit-Learn (Sklearn), SciPy, Matplotlib, StatsModels, and Seaborn. Some of the Python and related libraries skills include for loops, data manipulation, creating informative visualizations, and implementing basic machine learning principles. Through the use of machine learning principles, the aim of the project is to answer three questions about the dataset. The questions will be answered utilizing different machine learning approaches, such as classification, regression, and clustering.

The Data:

The data for this project originated from a 1988 study conducted by physicians on selected populations in Hungary, Switzerland, and the United States of America. The original dataset contained 72 features but has been reduced to 14 features, including the target feature – the presence of heart disease. This dataset was donated to the UCI Machine Learning Repository in 1988 and has been redistributed on several other data resource platforms, such as Kaggle (Lapp, David). The data for this project was downloaded from Kaggle because the data file was already in a .csv format and had descriptions for the 14 different features. The 14 features are Age, Sex, Chest Pain Type, Resting Blood Pressure, Total Cholesterol, Fasting Blood Sugar, Resting ECG, Max Heart Rate, Exercised Induced Angina, Old Peak, Slope, Number of Colored Blood Vessels, Thalassemia, and Presence of Heart Disease (Target Variable). The following list describes 9 of the 14 features:

1. *Age.* A numerical feature that details the age of the participants when the study was conducted. It has a minimum value of 29 years and a maximum value of 77 years.
2. *Sex.* The sex feature, labeled as "Male" in the notebook, is a bivariate categorical feature that asks, "is male?" A 0 indicates the participant is a female, whereas 1 represents a male.
3. *Chest Pain Type*. The chest pain type is a categorical feature with four different possibilities. 0 = Typical Angina, 1 = Atypical Angina, 2 = Non-Anginal Pain, 3 = Asymptomatic. Angina is a condition categorized by a "squeezing" type of chest pain which is caused by poor blood flow to the heart (Mayo Clinic).
4. *Resting Blood Pressure.* This numerical feature ranges from 94 to 200 mm Hg and is determined by the amount of arterial pressure exerted by blood being pumped from the heart for a patient while at rest. High levels of resting blood pressure can make an individual more susceptible to developing heart conditions (New Health Advisor).
5. *Total Cholesterol.* Cholesterol is a type of fat that can restrict blood flow within the body. A healthy total cholesterol level is that of 200 mg/dL or less. The dataset ranges from 126 to 394 (after the remover of outliers).
6. *Fasting Blood Sugar.* This categorical feature denoted whether or not the individual had An abnormal fasting blood sugar (levels greater than 120 mg/dL) by having the values of 0 = False/No and 1 = True/Yes
7. *Max Heart Rate.* A numerical feature listing the maximum heart rate in beats per minute (BPM) for those in the dataset. Max heart rate has an inverse relationship with age and is correlated with fitness levels.

8. *Thalassemia.* The categorical feature denotes whether or not the patient has thalassemia, a blood disorder cauterized by the abnormal formation of hemoglobin (Mayo Clinic). 1 = normal; 2 = fixed defect; 3 = reversible defect.
9. *Presence of Heart Disease*. A bivariate categorical feature which denotes whether or not the patient has heart disease. A 0 = no heart disease and a 1 = the patient has heart disease.

Exploratory Data Analysis (EDA) revealed that the downloaded dataset contained 1,025 rows of data. However, the data was reduced to 298 rows after removing duplicate values and outliers. The outliers were identified via a boxplot of the Total Cholesterol feature and removed based on having a z-score not within the acceptable range of [-3,3]. Additionally, the feature names were updated during the EDA phase to allow for easier interpretation.

Questions To Be Explained:

Classification Based: Left untreated, heart disease can be detrimental to one's health. Can a prediction model be designed to predict with 70% accuracy whether or not someone is likely to have heart disease based on their age, their sex, their chest pain type, resting blood pressure, and total cholesterol?

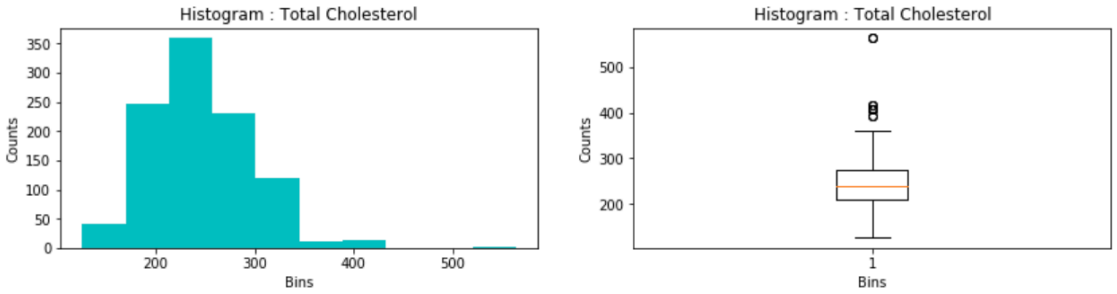
Regression Based: It is natural to see increased levels of cholesterol as one ages. Can total cholesterol (dependent variable) be predicted based on the following independent variables: age, sex, chest pain type, resting blood pressure, max heart rate, and heart disease status?

Clustering Based: Grouping individuals based on characteristics that are consistent with observations towards increased likelihoods of heart disease may be useful in identifying those at risk for developing the disease. Can clusters (groups) be easily identified based on the following features: Age, Resting Blood Pressure, and Total Cholesterol?

Visualizations:

A variety of visualization techniques were employed to learn more about the dataset. The first visualization technique used was creating a histogram and boxplot. Analysis of the .describe() method revealed that Total Cholesterol had a high standard deviation (45.76). The visualizations below confirmed this suspicion (Figure 1.1).

Figure 1.1



After reviewing the distributions, the data was modified to remove outliers from the dataset. The histogram and boxplot below show the data after outliers were dropped (Figure 1.2).

Figure 1.2

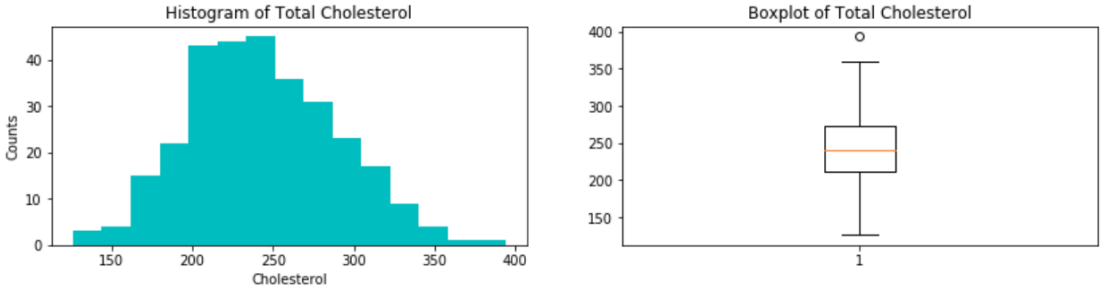
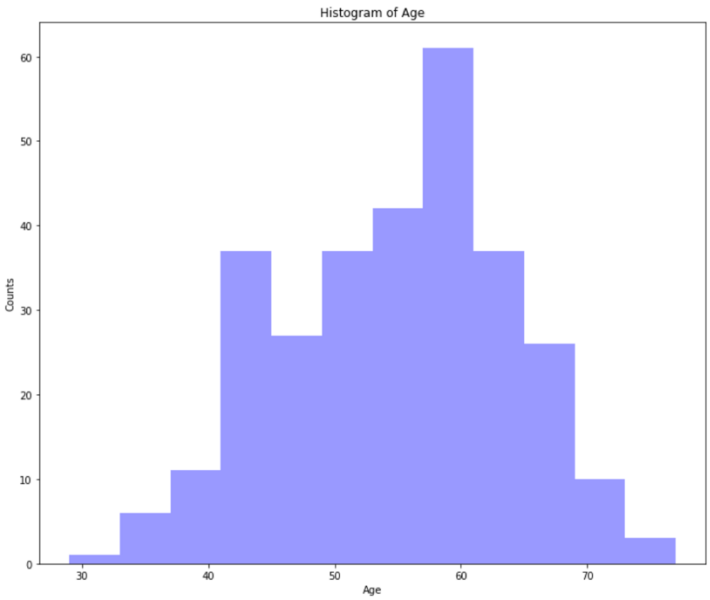


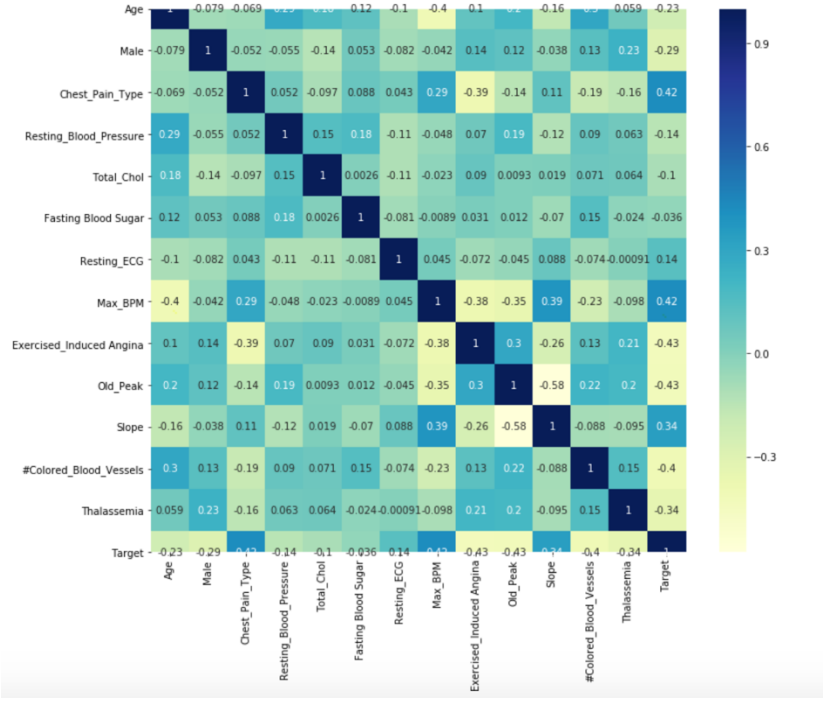
Figure 2

Figure 2 shows the age distribution of the dataset after modification were made. The data is slightly positively skewed, but not enough to cause concern.



Another visual aid that was created was a heatmap of the features. This visualization was created utilizing the Seaborn library. Dark blue boxes indicate a strong correlation between features. Too strong of a correlation would prompt the data scientist to do further investigation and may result in the dropping of one of the features. This dataset showed no signs of a strong correlation between features (i.e. the highest correlation between features was approximately 42%). As a result, no modifications were necessary.

Figure 3



Feature Engineering:

Often the dataset to be examined contains items that need to be altered so that the dataset can provide additional information. The concept of altering the raw data for the purpose of enhancing the information is known as *feature engineering.* Feature engineering can take many forms, including filling in missing values (imputation), feature transformation, or feature selection.

For this dataset, I was interested in creating a new categorical feature which was a concatenation of the individual's sex and heart disease status. For example, a male with heart disease would have the value of "11" and a female with heart disease would be assigned the

category of "01." I created this new feature by adding an additional column to my processed Pandas DataFrame (processed_df) and then assigned the new feature to the concatenated values. By doing this, it was easy to count and distinguish between those with and without heart disease based on their gender. A for loop, conditional statements, and initiators were used to count the frequencies (although the .value_counts() method could also have been used). The new feature allowed for the easy creation of a count/frequency plot visualization with Seaborn.

The second feature engineering task that was done was creating a new variable to provide general insight into the cardiovascular health of the participants. None of the original features provided a good measure of heart health; however, one could be created by subtracting their age from 220. This number results in a general/"rule-of-thumb" for an individual's potential maximum heart rate. A comparison was then done between the participant's max heart rate and their potential max heart to get an idea of their cardiovascular health. Lastly, a scatter plot visualization was created using Seaborn to examine any correlations between those with heart disease and those without heart disease.

Results:

Classification Results. Three attempts were made to identify models that would accurately predict whether or not someone had heart disease. The model assumed the following explanatory features: Age = 55, Male =1, Chest Pain Type = 0, Resting Blood Pressure = 125, and Total Cholesterol = 220.

The first model was based on a Naïve Bayes classifier. The data was split into test 70% training data and 30% test data. The model produced results with an approximately 75% accuracy in the training data and 72% accuracy in the test data. A confusion matrix revealed that the model produced 65 true results and 25 false results. Further, a classification report revealed

recall percentages of 76% and 75% for precision and accuracy, respectively, with regards to predicting the true positive cases of heart disease.

Next, a logistical regression model was conducted. This model was able to predict with 75% accuracy whether or not someone had heart disease based on the aforementioned features. Further, StatsModels was utilized to gather further information on the data. From the StatsModels, it was inferred that one's gender can have a significant impact on the likelihood of having heart disease.

The final classification model used to predict the target variable was the Trees Classification Model. This model was able to predict with 100% accuracy whether or not heart disease would be present in a person (Note that this could be misleading due to the limitations inherent in this specific algorithm).

The above models proved that classification prediction algorithms can be designed to predict the target variable at a 70% or greater level of accuracy.

Regression Results. A linear regression model was designed to predict Total Cholesterol based on Age, Gender, Resting Blood Pressure, and Max Heart Rate. Although a model can be made, the results of this model were inconclusive due to a low R-Squared/Adj. R-Squared value. This is because the dataset contained too much variability amongst the values for the selected features. As total cholesterol is heavily dependent on one's genetics and lifestyle, it makes sense that an accurate model would be difficult to design given the limitations of the dataset.

Clustering Results. The clustering results proved inconclusive. Three numerical features were used as values for the scatter plot (Age, Total Cholesterol, and Resting Blood Pressure); however, the resulting 3D scatter plot visualization did little to provide informative details on the clusters..

References

Detrano, R., Janosi, A., Pfisterer, M. & Steinbrunn, W. Heart Disease Data Set. (1988). [Public Dataset]. UCI Machine Learning Repository. Retrieved from https://archive.ics.uci.edu/ml/datasets/Heart+Disease

Lapp, David. (2019). Heart Disease Dataset [Public Dataset]. Kaggle. Retrieved from https://www.kaggle.com/johnsmith88/heart-disease-dataset

Mayo Clinic. (n.d.). [Website]. https://www.mayoclinic.org/

New Health Advisor. (n.d.). *Resting Blood Pressure.* Retrieved from https://www.newhealthadvisor.org/Resting-Blood-Pressure.html