Predictive Analysis of Cardiovascular Disease Using Python

Ty Johnson

tjohnson@bellarmine.edu

8 February 2024

Introduction

The dataset used in this predictive analysis project is about cardiovascular disease. It can be found on kaggle at Risk Factors for Cardiovascular Heart Disease (kaggle.com). After taking one too many trips to the cardiologist in previous years, I stumbled upon this dataset and found interest in it due to my personal connection to cardiovascular disease (nothing serious thankfully!). This project presented an opportunity to understand cardiovascular disease even more. This report will investigate the details of the dataset, providing a closer look into what the dataset contains such as the definitions and statistics within it, visualizing these findings, and providing a summary of them in the concluding section.

Dataset Description

Overall, the dataset contains information on around 70,000 patients. The data includes information on the patient such as their age, gender, height, and weight. Based on this information, an additional BMI field was calculated which will be used to understand the categories an individual may fall into from the range of underweight to class III obesity. There are measures that were collected from the patients in the dataset, blood pressure levels (systolic and diastolic) are the measures that were collected as ratios. Cholesterol and glucose levels were tracked as an ordinal types (refer to the table for more information). There were no missing values in any of the columns which means that the dataset was processed prior to being uploaded onto Kaggle. The table below displays the column names, description of data, data types, value range, and % missing for each column within the dataset. The table has been adjusted to represent additional processing that has been done in this project. Age was originally represented in days, but has since been changed to be displayed as years. In the next section, we will look into the summary statistics of the dataset.

Figure 1 – Dataset Description

Column Names	Description	Date Type	Value Range	% Missing
age	Age of patient	Ratio	29 - 64	0
gender	Gender of patient, represented by 0 or 1	Nominal	N/A	0

height	height of the patient measured in centimeters	Ratio	120-207	0
weight	weight of the patient measured in kilograms	Ratio	28-135	0
ap_hi	Systolic blood pressure reading taken from patient	Ratio	70-200	0
ap_lo	Diastolic blood pressure reading taken from patient	Ratio	40-140	0
cholesterol	Based on cholesterol level taken from patient, range of 1-3. 1 being normal levels, 2 being borderline high, 3 being high levels	Ordinal	N/A	0
gluc	Total glucose level taken from patient. Range of 1-3. 1 being normal levels, 2 being borderline high levels, 3 being high levels	Ordinal	N/A	0
smoke	Whether or not a person smokes. 0 represents the patient not smoking, 1 represents the patient smoking	Nominal	N/A	0
alco	Whether or not a person consumes alcohol. 0 represents no alcohol consumption for a patient, 1 represents alcohol consumption for a patient	Nominal	N/A	0
active	Whether or not a person identifies as being physically active. 0 represents the patient being not physically active, 1 represents the patient as physically active	Nominal	N/A	0
bmi	Body mass index, computated based on patient height and weight. Measure that can categorize individuals into underweight, normal, obese, etc.	Ratio	14.5-39.48	0

cardio	Whether or not a person suffers from a cardiovascular disease. 0 represents cardiovascular disease being absent in the patient, 1 represents the presence of cardiovascular disease in the patient	Nominal	N/A	0

Dataset Summary Statistics

In the table below, the summary statistics generated in the Python exploration is shown. The statistics are for the ratio data types that are included in the dataset. Beginning with age, it can be seen as patients in the dataset fall between the ages of 29 and 64. The median age is 53 with a standard deviation of 6.77. The minimum height is at 120cm (3' 11") and the maximum at 207cm (6' 9"). The median height of a patient is 165cm (5' 5"). The minimum weight is 28kg (61lb) and the maximum is 135kg (297lb). As shown in the table below, the count of rows in the dataset has been reduced from 70,000 to 66,676. Rows that had irregular blood pressure readings for systolic or diastolic were removed from the dataset as there were questions of validity. Furthermore, there were some readings that fell into the negative ranges that needed to be removed from the dataset. To make this happen, a lower bound (70 for systolic and 40 for diastolic) and upper bound (200 for systolic and 140 for diastolic) value was set in order to determine how to handle these values. The lower and upper bound values are based on Figure 3 below. Anything outside of this range was deleted.

The column for cardiovascular disease is relatively proportional between the amount of each outcome. 34090 patients had the absence of cardiovascular disease and 32586 had one present (51% vs 49%). The number of patients who were active outweighed the ones who were inactive, with 80% being active. Only 5.3% of patients identified as alcohol consumers, while the amount of smokers was slightly higher at 8.9%. Male was the predominant gender included in this dataset with around 65% of patients being male. These proportions are later visualized in the graphical exploration section.

Previously shown in the section, there were no missing values included in the dataset. There were many outliers present in each of the columns (specifically ones with the ratio data types). In order to deal with these outliers, a function was created that would identify them, print them out, and if needed, remove them from the dataset. There were 4 outliers for age, 507 for height, 1741 for weight, and 2020 for bmi. Since a lower and upper

bound range was preset for systolic and diastolic blood pressure levels, that section was not examined. As for handling these outlier values, given that bmi was calculated using both the height and weight columns, outliers in the dataset were deleted ONLY according to that column. Given that one of the major factors we are concerned with is the characterization of the body, extremely irregular cases would likely do more harm to the models than good.

Before moving onto the next section, the correlation between variables will be mentioned. In order to look into the correlation, a correlation heatmap was generated (See Figure 4 Below). From this heatmap, as we are concerned with the outcome of having a cardiovascular disease or not, the most noticeable takeaway would be that both systolic and blood pressure levels are the most correlated to cardiovascular disease in comparison to others. The rest of the variables have a weaker correlation (Though no relationship of variables in this dataset can be defined as strong).

Figure 2 – Summary Statistics

Summary	age	height	weight	ap_hi	ap_lo	bmi
count	66676	66676	66676	66676	66676	66676
mean	52.799028	164.57818	73.032593	126.23628	81.168801	26.972406
std	6.776008	7.801685	12.708398	16.386771	9.401935	4.409611
min	29	120	28	70	40	14.52
25%	48	159	64	120	80	23.81
50%	53	165	71	120	80	26.17
75%	58	170	80	140	90	29.74
max	64	207	135	200	140	39.48

Figure 3 – Blood Pressure Levels

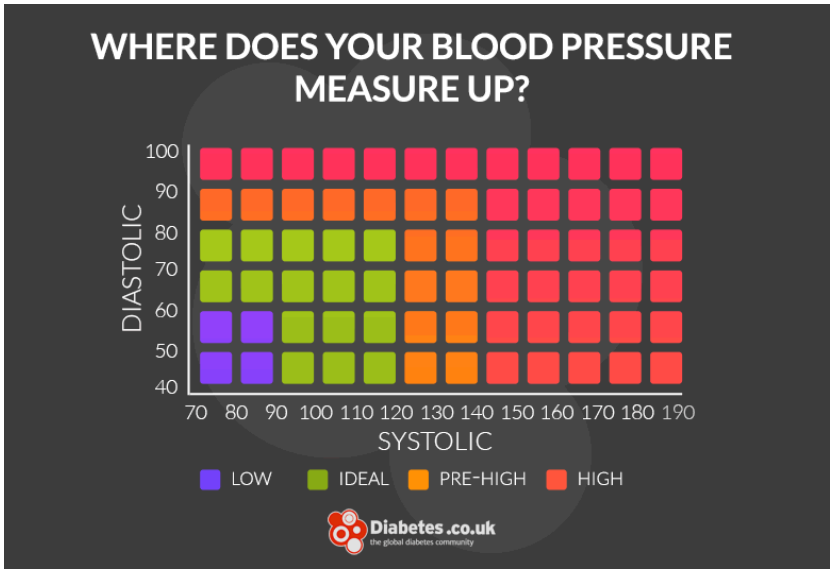
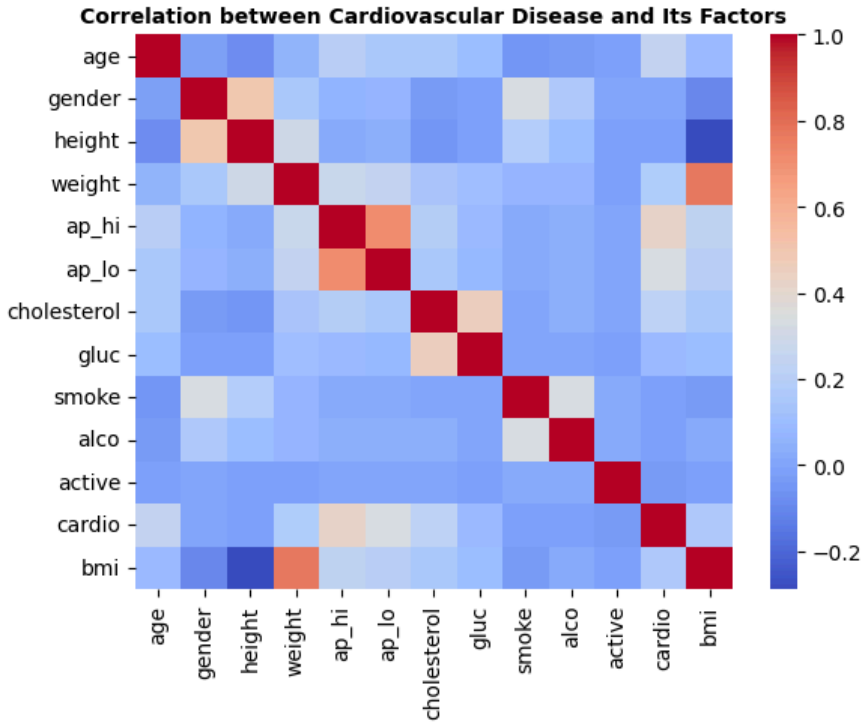


Figure 4 – Correlation Heatmap



Dataset Graphical Exploration

Figure 5 – Patient Information Dashboard

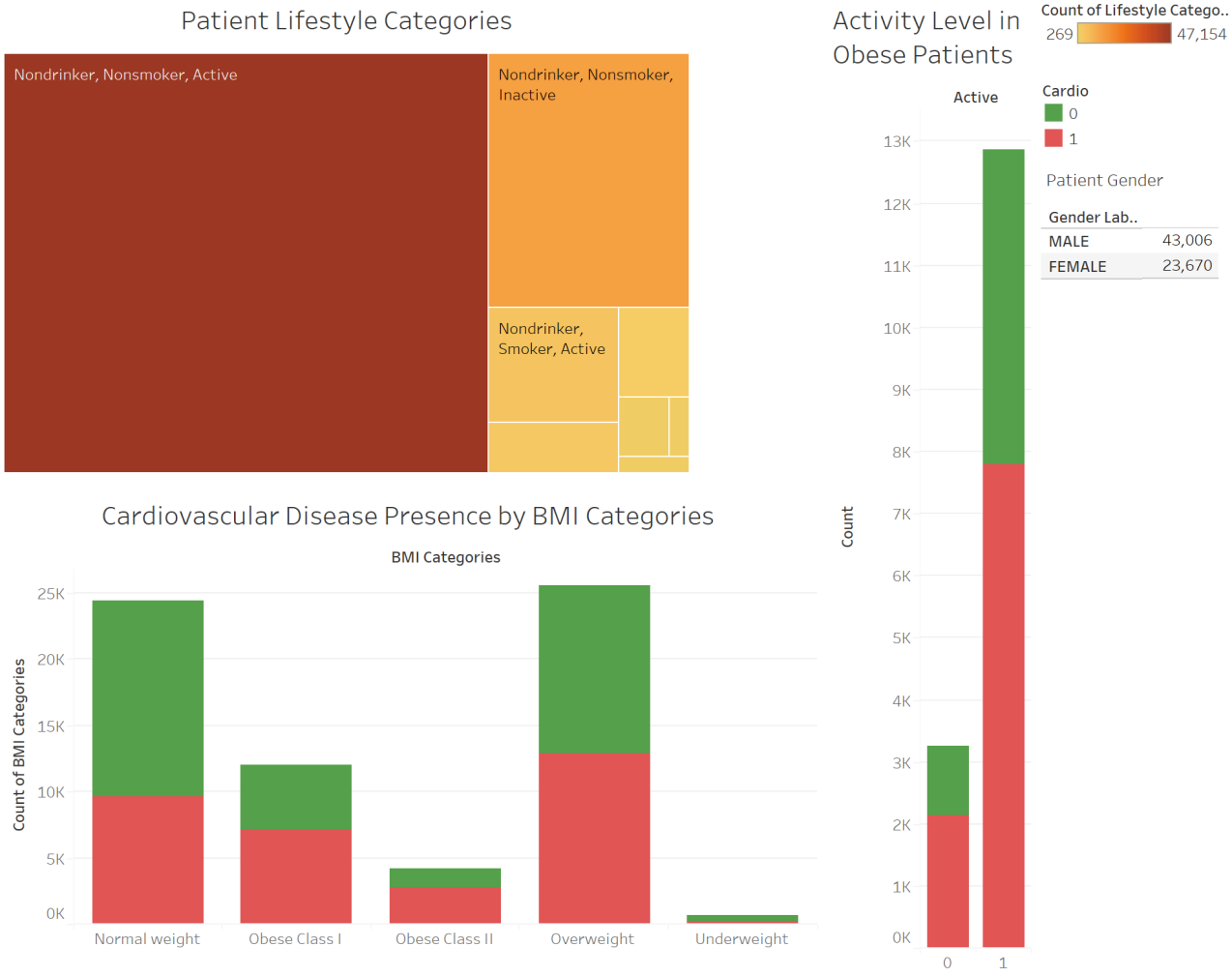
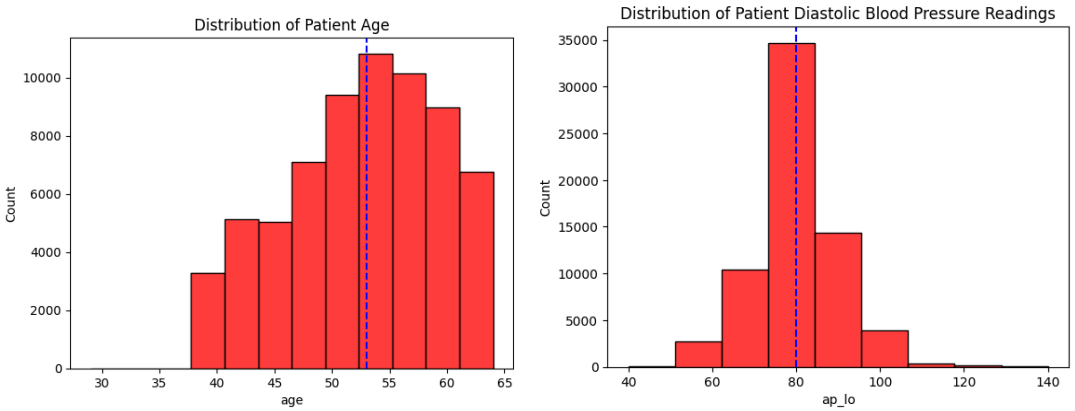


Figure 6 – Distributions of Variables



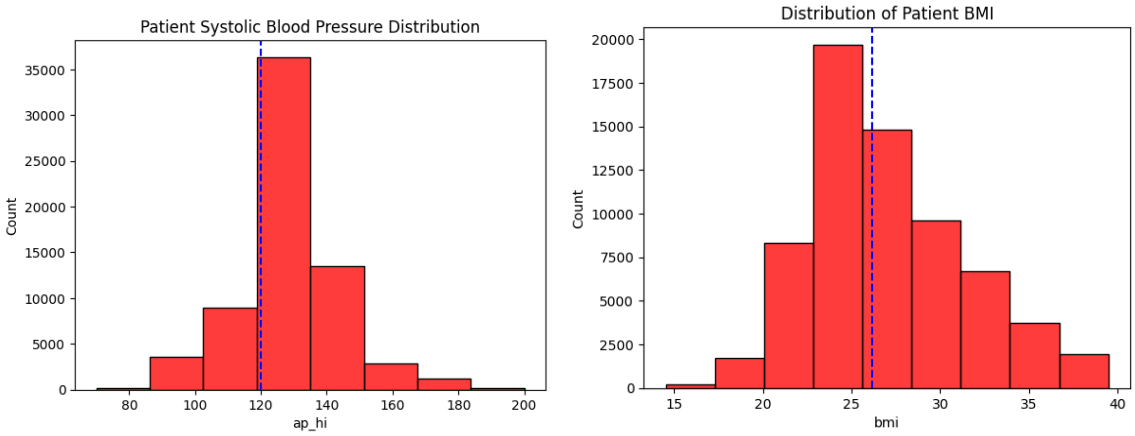
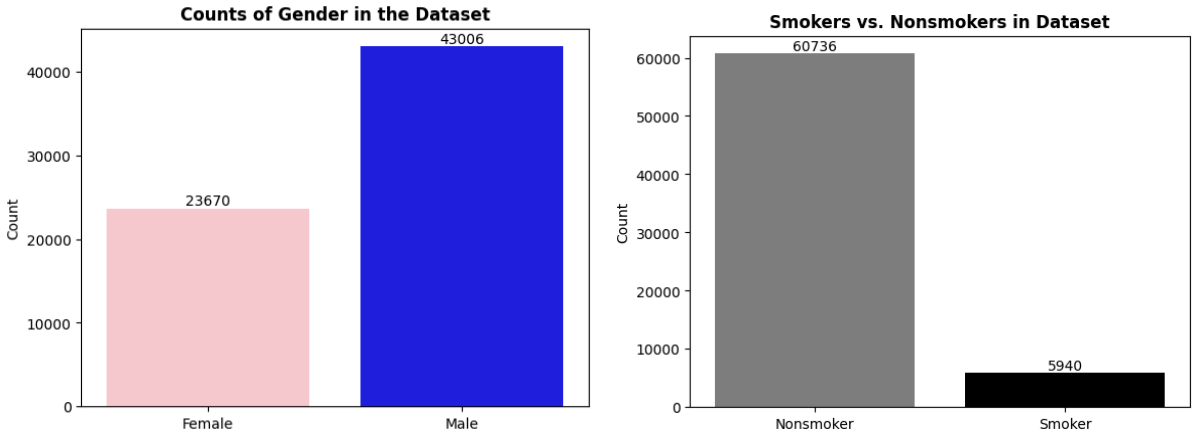
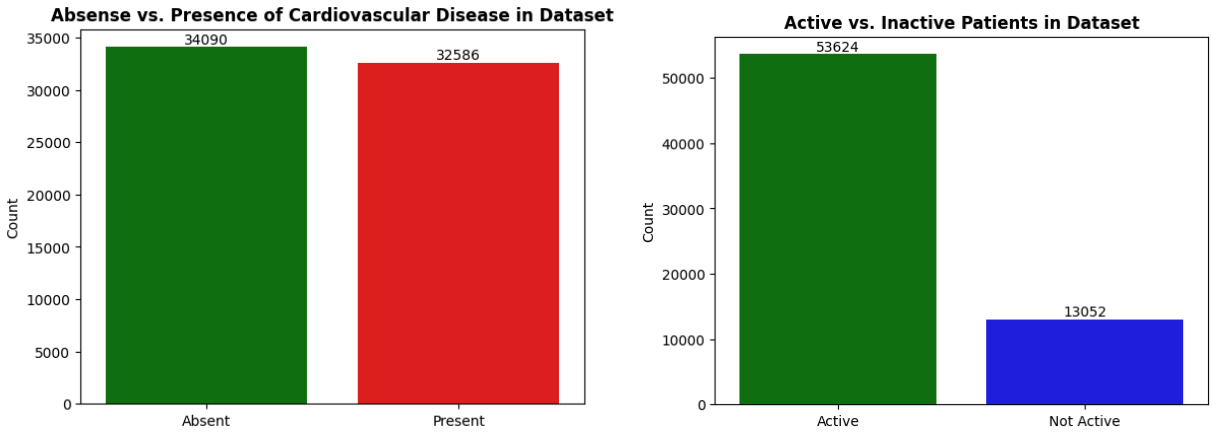
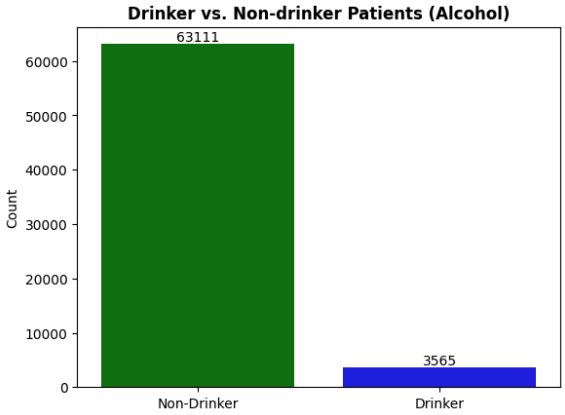


Figure 7 – Counts of Variables







Summary of Findings

In terms of model building, this dataset should serve good for finding the desired outcome. The amount of patients with cardiovascular disease present vs. absent is nearly even, reducing the concern of the model being bias towards one outcome over another. There may be some issues when it comes to identifying and signifying how smoking, activity level, and alcohol consumption will play into developing cardiovascular disease, given the small percentage of patients that fall into the categories. The creation of multiple models could be a way to address this issue, with having some models dedicated towards "leveling out" the data, trying to find an equilibrium when it comes to entries that have those factors. With over 60,000 entries, this outcome should be possible to achieve and it would be interesting to see if it would have any effect on the accuracy of the model. Since there are outliers that remain in a few columns, it could be possible that the removal of them will become necessary when it comes to building the models. As it came, this dataset was relatively clean and set up for model building right away, with the exception of the erroneous values that came up in a few columns such as in the blood pressure levels.

A disappointing finding in this dataset is the lack of depth for different features. For instance, activity level, smoking, and alcohol consumption are all columns that are represented with 0 and 1s. The source of the dataset does not specify how this data was collected and there is always the question of the truthfulness of the level. Furthermore, these categories cannot measure the extensiveness of each issue, as an individual could have an opinion when it comes to what being active or any of the other categories mean. In order to answer some of the bigger picture questions related to these factors and developing cardiovascular disease, the calculation of an "intensity score" for each of the categories could prove useful in determining "how much" of each (although this is impossible for this project, it gives a good idea of the shortcomings of data collection). Also, the tree map reveals that most patients in

the dataset fall into the perfect combination of their interaction with the risk factors of cardiovascular disease, which prevents an issue for really understanding the factors through this data.