Adam Lang

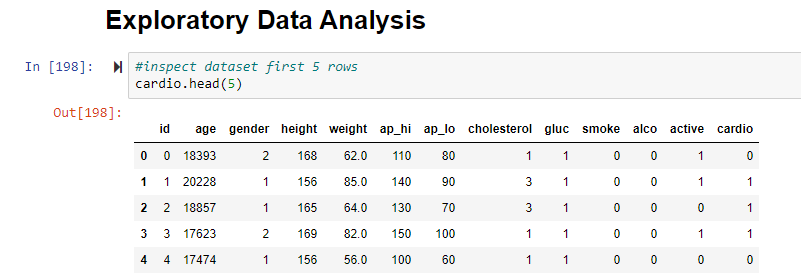
9/13/20

IBM Data Science-Final Project

**Data Understanding**

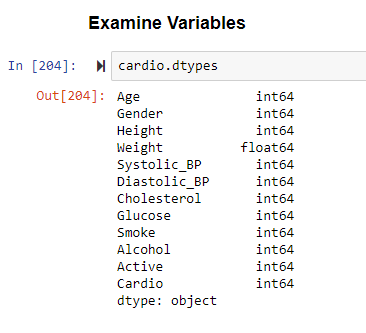
This data entitled the “Cardiovascular Disease dataset” was obtained from Kaggle and is ideal for performing data mining and machine learning activities to predict Cardiovascular Disease (CVD). The dataset contains 70,000 records, 11 features, and 1 target variable (0 having no CVD, 1 having CVD). There are 3 types of input features for each variable: Objective (factual information), Examination (results of medical exam), and Subjective (information given by the patient). All dataset values were entered after a patient’s medical exam (Ulianova, 2018).

I began exploratory data analysis and started to find some interesting things about the dataset features. I first saw that there are 70,000 rows and 13 columns. We can see the data frame below:



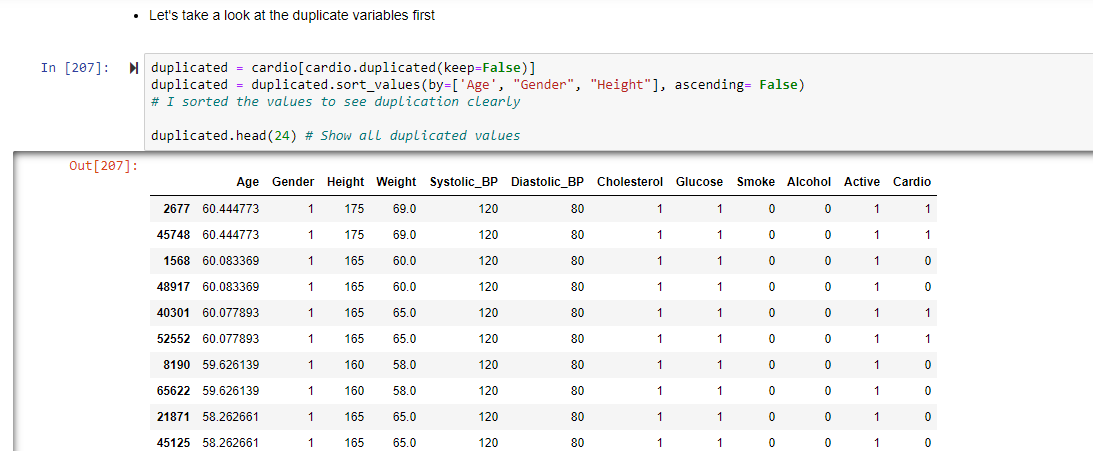
Right away we can tell that there are some interesting features that need to be dealt with from the dataset. The age is in number of days which will need to be changed to years to fully understand and work with. I am not too concerned about height which is in cm and weight which is in kg as this can be useful to calculate the body mass index (BMI) which is an important risk factor that I will add as a separate column to the dataset for predicting CVD. The columns “api\_hi” and “ap\_lo” correspond to “systolic blood pressure” and “diastolic blood pressure.” I am going to change these headings so we can better understand these values. The final variables are cholesterol, glucose, smoking, alcohol intake, activity, and the target variable cardio which 0 stands for no disease and 1 for having CVD.

The data types are seen below:

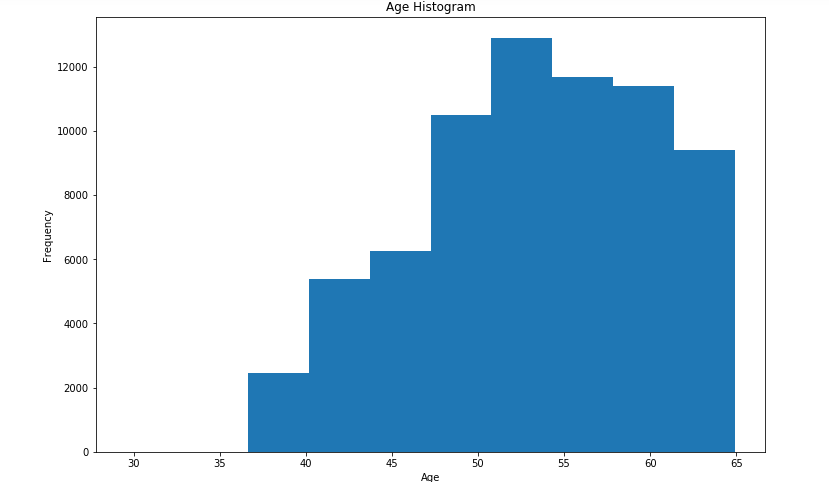


What we can see is that all variables are numeric. However, not all variables are continuous as there are categorical variables present which have been encoded for analysis. The numeric variables are: Age, Height, Weight, Systolic\_BP, and Diastolic\_BP. The categorical variables are: Gender, Cholesterol, Glucose, Smoke, Alcohol, Active, and Cardio. Looking closer at these variables Cholesterol and Glucose are scaled from 1 to 3 with 1 being “Normal”, 2 being “Above Normal” and 3 being “Well Above Normal”. Smoke, Alcohol, Active, and Cardio are all 0 or 1 for not present vs. present. Since the variables are already encoded, we will not have to do much for preparing them for machine learning.

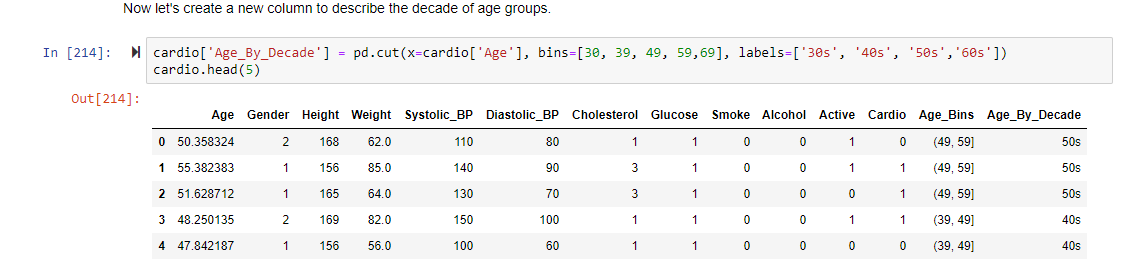
In terms of data integrity, there are no missing values but there were 24 duplicate records as seen in this data frame from my Jupyter notebook:



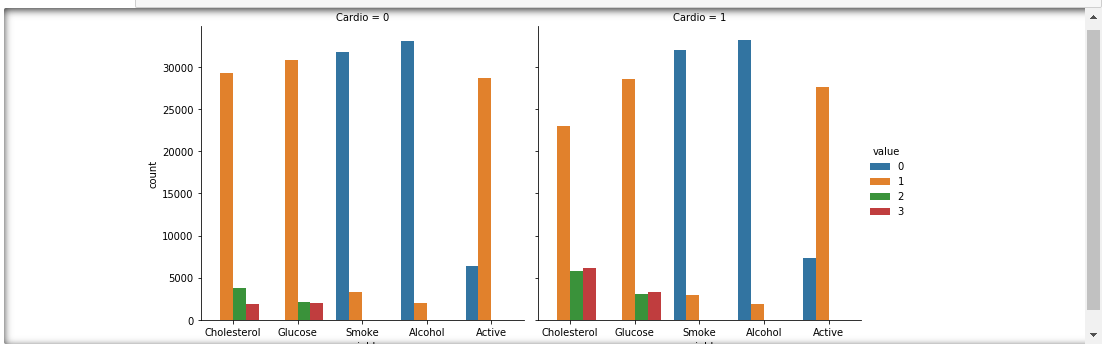
I ended up dropping all 24 duplicate records from the dataset. Next, I evaluated the variables. First, I looked at the Age variable.



We can see that the age distribution is predominantly around age 50-55 but all below age 65. I decided to bin the ages to make the data more manageable for analysis.

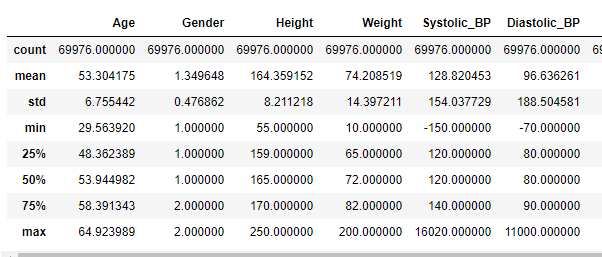


This made it easier to see how many age groups we are working with. The next thing I did was evaluate all categorical variables. It was interesting to me that there were more “Normal” (Group 1) findings for Glucose, Cholesterol; more smokers and non-drinkers; and more active individuals overall in the dataset. This makes it interesting to automatically assume we have patients with heart disease as those alone are high risk factors. I did however compare the dataset based on those with and without CVD and saw that those with CVD do have higher glucose and cholesterol levels as seen below:



As we can see the dataset does reveal there are more people with cardiovascular disease that have higher glucose and cholesterol levels as compared to those without CVD. Every other variable is virtually the same and normal.

Another major issue with this dataset is there are significant outliers. We can see this in the data frame below:

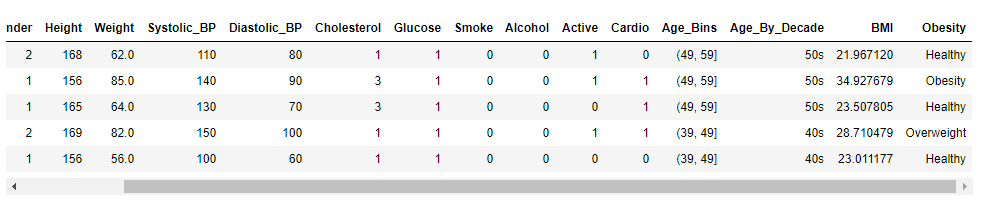


The values outlined in red are maximums for the respective columns. Most concerning is that the highest Systolic blood pressure is 1,6020 and the highest Diastolic blood pressure is 11,000 both of which are not possible as the highest recorded blood pressure on planet earth was in an exercise study at 370/360. A doctor will diagnose you with hypertension if your blood pressure is above 120/80 and is considered a “hypertensive crisis” if it is above 200/120mm/Hg (Narloch et al. 1995 and Unger et al. 2020).

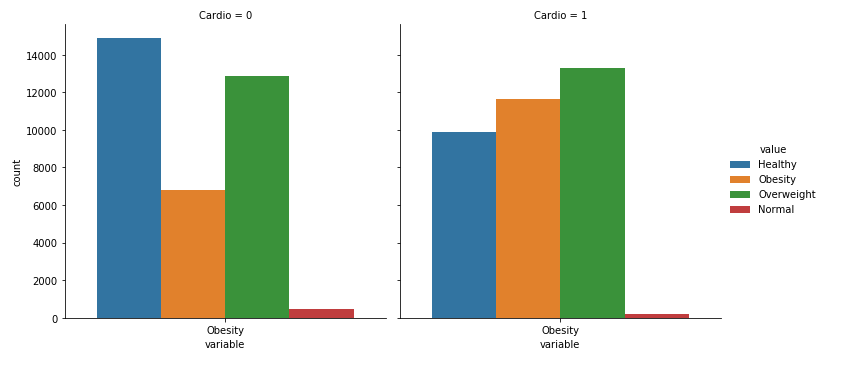
To deal with these extreme outliers I had to perform data normalization to better quantify the extremeness of the values. I did this by performing scaling of the data. I was then able to quantify appropriate quartiles for the upper levels of systolic and diastolic blood pressure as 250 and 200, respectively.

As for the weight outlier this is 200kg and corresponds to about 440lbs which may be an outlier but is still important consideration and reasonable to consider when diagnosing obesity which is a major risk factor for CVD so I left it as is. I did however remove the height of 250cm which corresponds to over 8 feet tall which is a definite outlier.

The next thing I did to the data was create columns for obesity classification and BMI. This can be seen below:



You can also see that I created columns for Age bins and Age classification by decade. This made it easier to visualize the data and classify each group. I was then able to quantify this with a bar graph.



What is interesting is that we can see there are more obese and overweight individuals in those with CVD (1) than without (0). This could prove to be beneficial in prediction of CVD as obesity is one of the highest risk factors.

**Next Steps**

My next steps are going to begin by performing descriptive and inferential statistical analysis of the numeric data. I am going to look for correlations that may stand out and make a correlation matrix. I also intended to perform an ANOVA. ANOVA (which stands for ANalysis Of VAriance) is a technique for testing whether the continuous outputs depend on the inputs. Equivalently, it tests whether different input categories have significantly different values for the output variable (MIT, n.d.). The point would be to look at the different levels of categorical variables and their respective influence on the other variables. Before doing this, I will have to transform the categorical variables I created (i.e. Obesity, Age bins) using one hot encoding or pd.get\_dummies in python.

The final piece of this whole data project is modeling. Prior to modeling I will have to make sure that all data has been feature scaled. Feature scaling is essential for machine learning algorithms that calculate **distances between data** (Roy, 2020). I know that I performed this earlier to remove the outliers in the data, but it may be necessary again as I did leave an outlier in the weight category. It is also important for my machine learning models. Prior to machine learning I will have to split the data into training and testing sets appropriately.

For machine learning I intend to perform Decision Tree Analysis, Naïve Bayes, and Support Vector Machines. Decision Trees are good at handling categorical data such as this data set has. In fact, I may perform this algorithm first so we can see how it responds to the categorical inputs and then perform feature scaling and the other algorithms. Naïve Bayes Classifier uses the Bayes’ theorem to predict membership probabilities for each class such as the probability that given record or data point belongs to a particular class. The class with the highest probability is considered as the most likely class. Naïve Bayes uses Bayes Theorem which considers that all events are independent of each other, basically that the presence of a particular feature in a class is unrelated to the presence of any other feature (Ray, 2017). Lastly, the SVM algorithm will find a hyperplane that distinctly classifies the patients as having CVD or not (Gandhi, 2018).

Finally, I plan to perform k-folds cross validation to validate the accuracy of the models I perform on unseen data. I also intend to validate the models accuracy with the Jaccard index, F1 score, confusion matrix, and the AUC-ROC curves. The final step will be to translate the model results into actionable insights on what was learned regarding the independent variables ability to predict cardiovascular disease and which variables are most important in the prediction process.

**References**

Gandhi, R. (2018) Support Vector Machine – Introduction to Machine Learning Algorithms. Retrieved from: https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

# MIT (n.d.) Categorical Data. Retrieved from: http://www.mit.edu/~6.s085/notes/lecture6.pdf

# Narloch et al. (1995) Influence of breathing technique on arterial blood pressure during heavy weightlifting. Arch Phys Med Rehabil. Vol. 76(5): 457-62.

# Ray, S. (2017) 6 Easy Steps to Learn Naive Bayes Algorithm with codes in Python and R. Retrieved from: https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/

Roy, B. (2020) All about Feature Scaling. Retrieved from: https://towardsdatascience.com/all-about-feature-scaling-bcc0ad75cb35

Ulianova, S. (2018) Cardiovascular Disease dataset. Retrieved from: <https://www.kaggle.com/sulianova/cardiovascular-disease-dataset>

# Unger et al. (2020) 2020 International Society of Hypertension Global Hypertension Practice Guidelines. Hypertension. Vol. 75(6):1334-1357.