Cardiovascular Disease Prediction

Vidyasagar Kummarikunta
DSC 680 – Applied Data science

Professor Catie Williams
Bellevue University

Contents

Introduction

According to American Heart Association (AHA), Cardiovascular disease (CVD) is the number one cause of deaths in the US. CVD accounted for 859,125 deaths in the US in 2017. Cardiovascular diseases claim more lives each year than all forms of cancer and Chronic Lower Respiratory Disease combined. In 2017, Coronary Heart Disease was the leading cause (42.6%) of deaths attributable to cardiovascular disease in the US, followed by stroke (17.0%), High Blood Pressure (10.5%), Heart Failure (9.4%), diseases of the arteries (2.9%), and other cardiovascular diseases (17.6%) (AHA, 2020).

CVD is the leading global cause of death and accounted for approximately 17.8 million deaths globally in 2017. This number is expected to grow to more than 22.2 million by 2030, according to a 2014 study. Total direct medical costs of CVD are projected to increase to $749 billion in 2035, according to a 2016 study.

The American Heart Association gauges the cardiovascular health of the nation by tracking seven key health factors and behaviors that increase risks for heart disease and stroke. These key health factors are called "Life's Simple 7" and they are measured to track progress of improving the cardiovascular health. Life's Simple 7 are: not-smoking, physical activity, healthy diet, body weight, and control of cholesterol, blood pressure, and blood sugar.

The silver lining is that CVD is highly preventable and simple lifestyle modifications (such as reducing alcohol and tobacco use; eating healthily and exercising) coupled with early treatment greatly improves its prognosis. It is, however, difficult to identify high risk patients because of the multi-factorial nature of several contributory risk factors such as diabetes, high blood pressure, high cholesterol, etc. This is where machine learning and data mining come to the rescue (Mordecai, n.d.).

Predictive analytics is being slowly and steadily embraced by the healthcare industry. There are many areas in healthcare that are using machine learning methods to improve patient care and reduce costs. One such application of machine learning in healthcare is predictive diagnosis. The goal of this project is to predict whether an individual has cardiovascular disease or not based on various parameters and health conditions.

Data sources

The dataset is obtained from Kaggle website. (Link: https://www.kaggle.com/sulianova/cardiovascular-disease-dataset). The dataset has 12 columns and 70,000 rows. The column attributes include age, gender, height, weight, systolic blood pressure, diastolic blood pressure, cholesterol, glucose, smoking, alcohol intake, physical activity and cardiovascular disease. The target variable is the last column which is a binary categorical variable of whether a person will have the heart disease or not. The advantage with this dataset is that it has reasonable amount of data (70,000 rows) to explore and build a really well working classification model.

Problem statement

The goal of this project is to predict whether an individual has cardiovascular disease or not based on various parameters and health conditions. For the predictive analysis, we will also compare the performance of classification algorithms to understand which one works better.

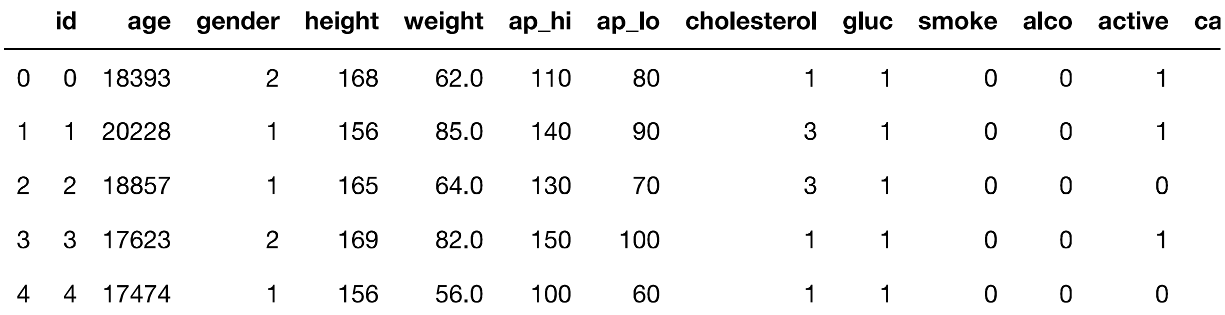
Approach

For this project, I will be using Python program, Jupyter notebook and sikit learn classification algorithms for the analysis. First step is to understand the dataset. Next, I will be performing exploratory analysis. The dataset has a few issues that need to be fixed. For example, the age is represented in days instead of years. I will also include the BMI column. Next step will be feature engineering and selecting dependent and independent variables to build a model. I will be using Logistic regression, Support Vector Machine (SVM) and Naïve Bayes Classifier algorithms to fit the data and evaluate performance.

Methods and Results

Loading dataset –

Shown below is the preview of the dataset before making any changes.



	id	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	ca
0	0	18393	2	168	62.0	110	80	1	1	0	0	1	
1	1	20228	1	156	85.0	140	90	3	1	0	0	1	
2	2	18857	1	165	64.0	130	70	3	1	0	0	0	
3	3	17623	2	169	82.0	150	100	1	1	0	0	1	
4	4	17474	1	156	56.0	100	60	1	1	0	0	0	

Shown below is the preview of dataset after making the following changes - ‘age’ column to age in years, adding ‘BMI’ column, dropping ‘id’ column, and rearranging the column locations.

Also, the datatypes of columns cholesterol, gluc, smoke, alco, active and cardio are changed to categorical.

	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	age_years	BMI	cardio
0	2	168	62.0	110	80	1	1	0	0	1	50	22.0	0
1	1	156	85.0	140	90	3	1	0	0	1	55	35.0	1
2	1	165	64.0	130	70	3	1	0	0	0	52	24.0	1
3	2	169	82.0	150	100	1	1	0	0	1	48	29.0	1
4	1	156	56.0	100	60	1	1	0	0	0	48	23.0	0

Fig.1 below, shows the distribution of all numerical columns in the dataset.

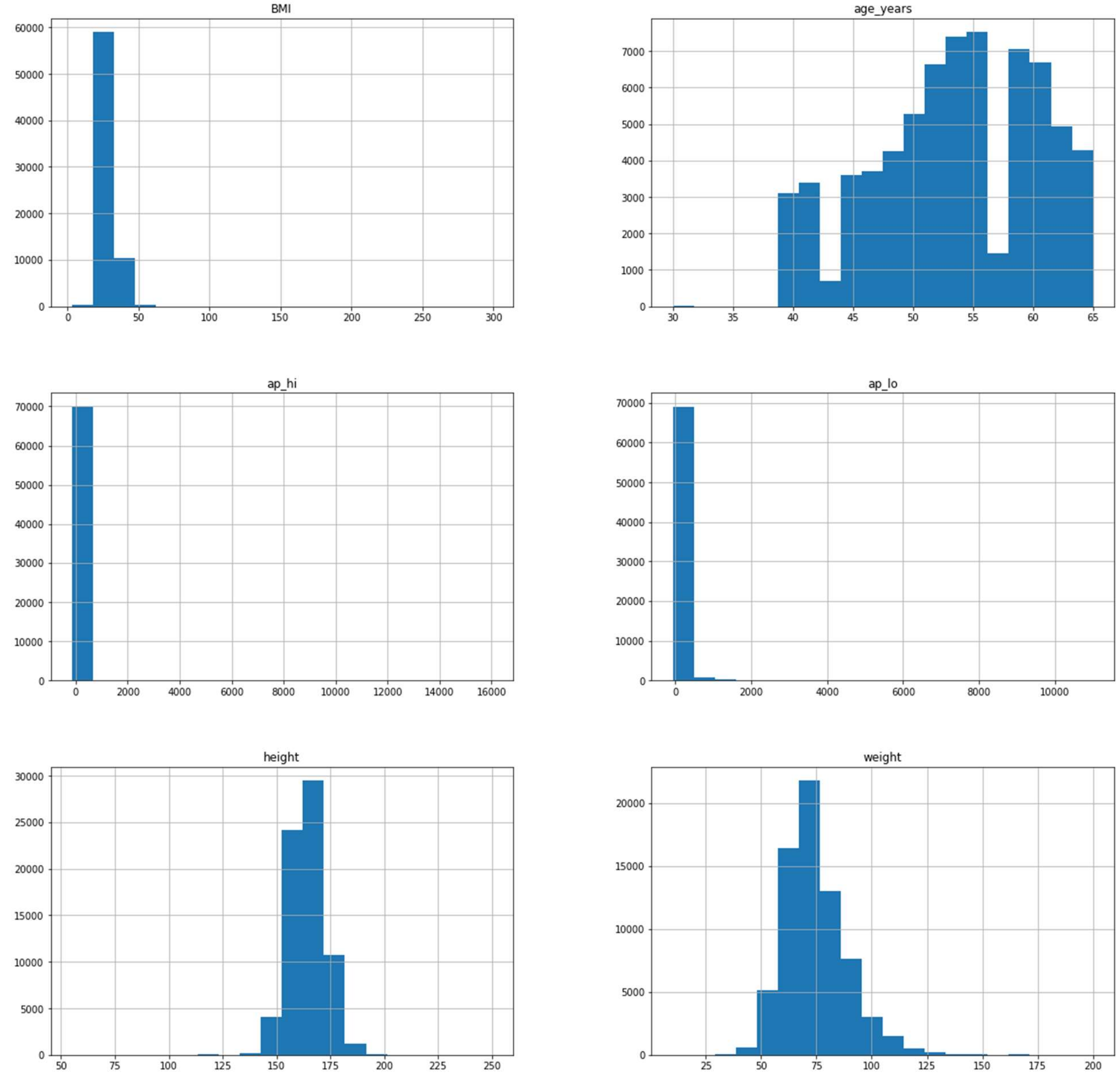


Fig.1: Histogram of all numerical data

Plotting age in years vs cardio to see at what age does the number of people with CVD exceed the number of people without CVD? From the plot (Fig.2) it appears that the switch happens at age 55.

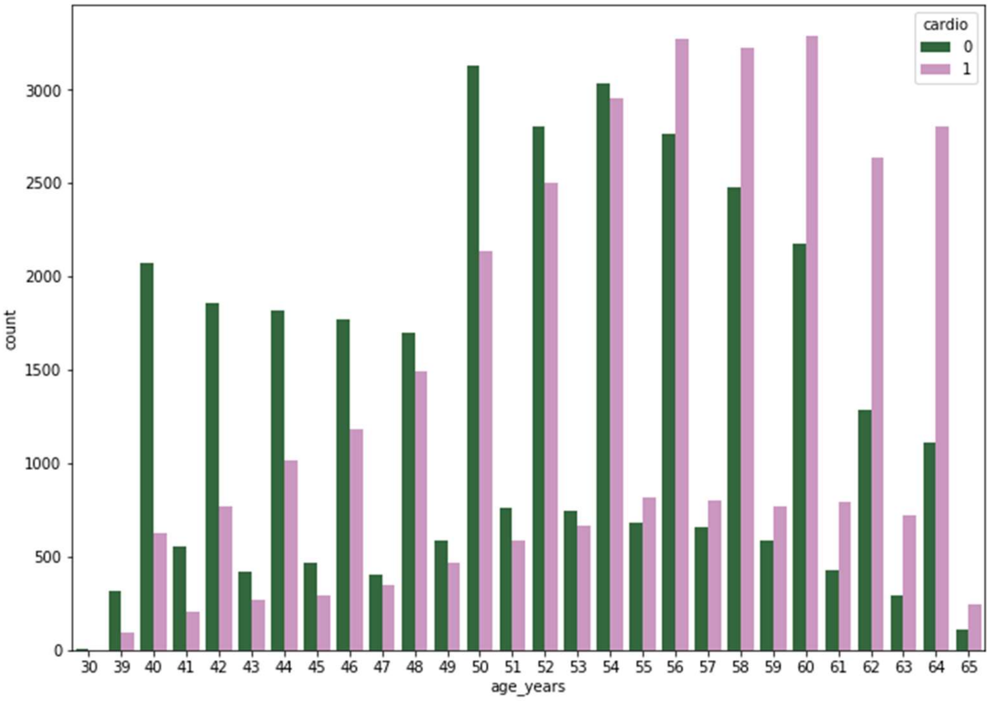


Fig.2: Age vs Cardio

Fig.3 shows the distribution of categorical variables. Cholesterol has categories 1, 2, 3; gluc has categories 1, 2, 3; smoke, alco and active all have categories 0 and 1.

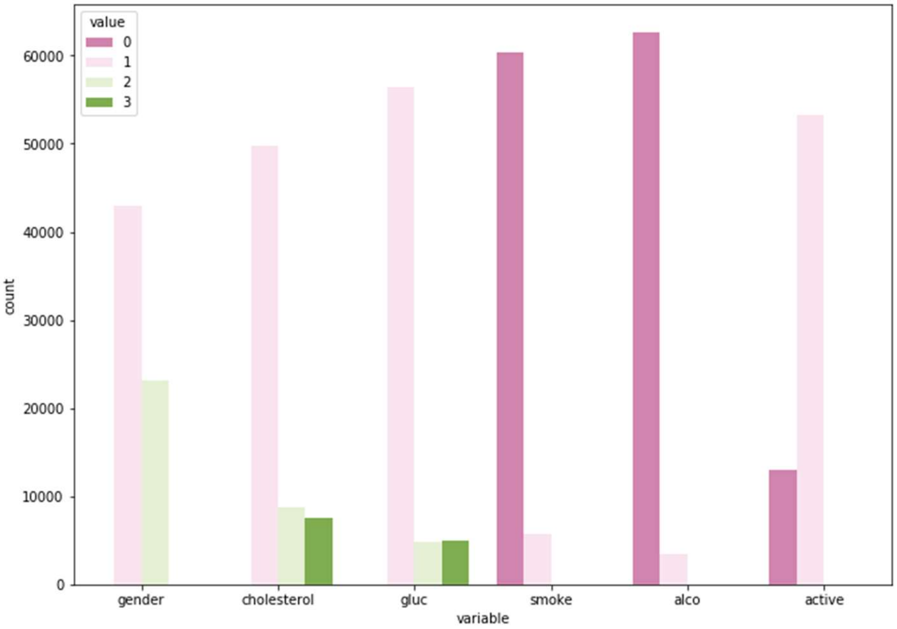


Fig.3: Bar chart of categorical variables

From Fig.4, It can be clearly seen that patients with CVD have higher cholesterol (increase in categorical values 2 and 3) and blood glucose level. And, generally speaking less active.

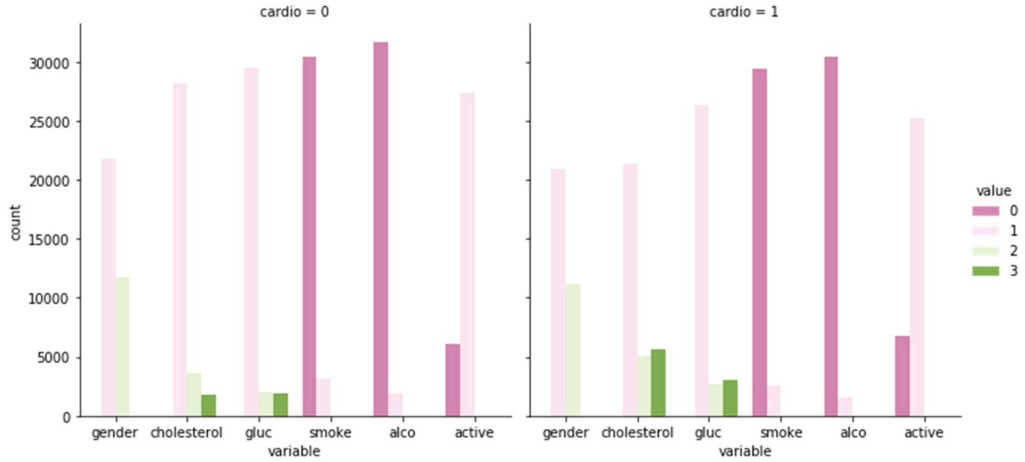


Fig.4: Plotting categorical variables by separating them with cardio positive and negative

According to Fig.5, the diastolic pressure (ap_lo) is low compared to systolic pressure (ap_hi) which is the expected trend.

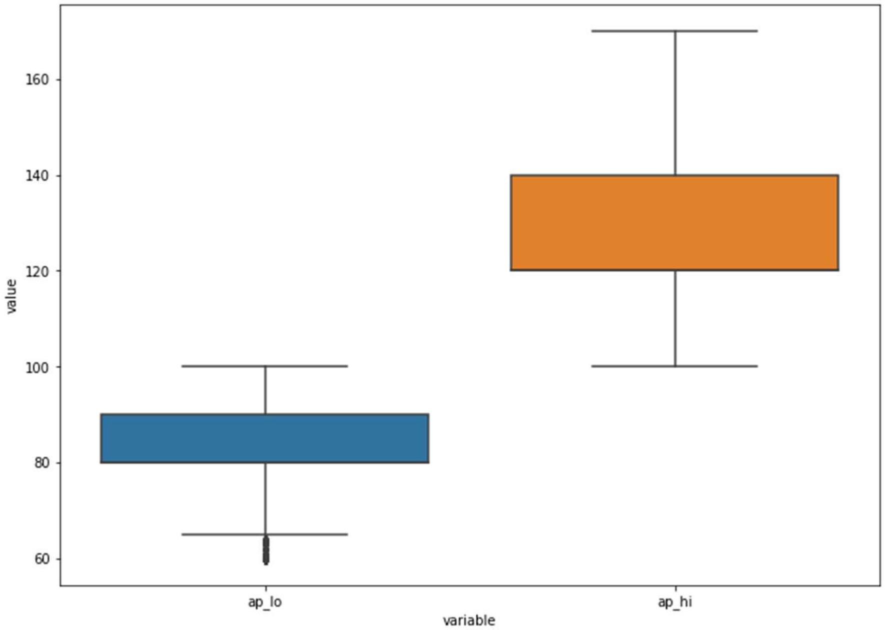


Fig.5: Plotting distribution of diastolic and systolic pressure

In Fig.6, As we can see weight and BMI are highly correlated.

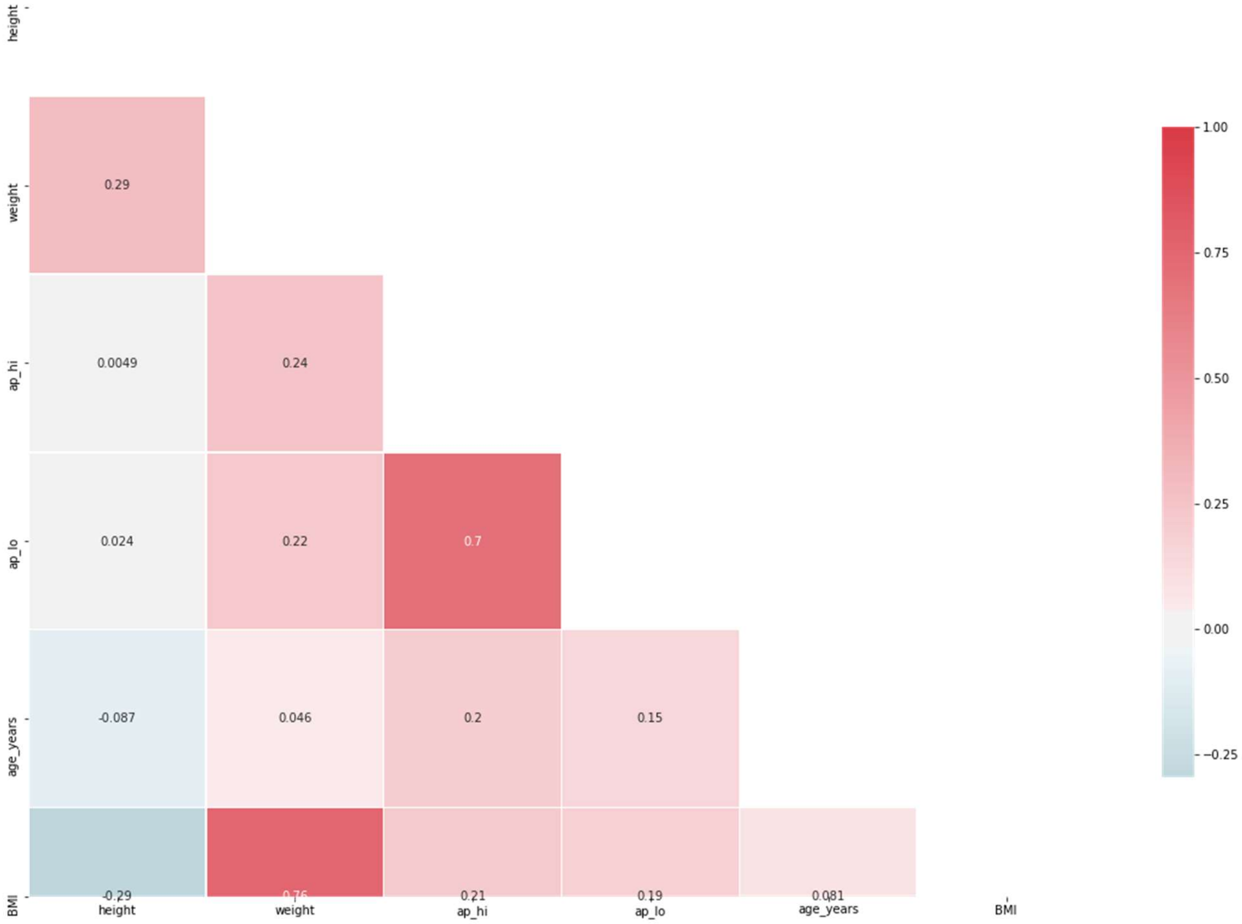


Fig.6: Correlation matrix

After the exploratory analysis, the data is split into predictor variables and target variables. The target variable is ‘cardio’ to determine whether a patient has CVD or not. For training the model, data is split into 80:20 ratio of train and test sets. Using the sikit learn library, the train and test sets were applied to Logistic regression, Support Vector Machine (SVM) and Naïve Bayes Classifier algorithms to fit the data and evaluate performance.

Shown below are the results:

The Logistic regression model demonstrates an accuracy of 72%
The Support Vector Machine model demonstrates an accuracy of 71%
The Naïve Bayes Classifier model demonstrates an accuracy of 72%

It is evident from the results that Logistic regression model and Naïve Bayes classifier gives highest accuracy. However, all three models seem to perform at the same level. To further this project, extensive feature selection needs to be performed to achieve better accuracy.

Q & A

With the results/findings in hand, here are some questions and answers that I can derive from the observations of the project -

Was the objective of this project achieved?

Yes, the objective of this project is to apply machine learning algorithms to help predict cardio vascular diseases. The results of this project show that there is scope to build an accurate predictive model to predict CVD.

Are the results of this project good enough to start implementing the algorithm for real world CVD predictions?

While all three models have shown similar accuracy results, it only suggests that these models can be useful if and when the accuracy is improved. There is still room for improvement before a model can be deployed in the real world.

Are there any issues with the data? Do you think the data used for this project is accurate for this type of analysis?

Upon loading the data and inspecting it, I found some anomalies. There is not much supporting information given for the dataset. For example, while preparing data I had to assume that the categorical values of 'smoke', 'alcohol', 'active' and 'cardio' are 0 and 1 where zero stands for absence and 1 stands for presence of the event. Also, 'gender' has categorical values as 1 and 2 and I had to figure out which category does 1 and 2 assigns to male or female? I also had to include a new column and compute BMI values for analysis. In addition, in some cases diastolic pressure is higher than systolic, which is also incorrect.

Is there any relation to CVD and cholesterol?

Yes, as we can see in Fig.4, for patients who have CVD (cardio=1), the levels of cholesterol is higher compared to cardio=0.

Is age a factor for cardio vascular disease?

Yes, it is. Fig.2, suggests that after age 55 there are more cardio cases than before.

Which model performs better?

All three models show similar performance.

Concluding Remarks

According to the American Heart Association, cardiovascular disease is accounted for over eight hundred thousand deaths in the US in 2017. Cardiovascular diseases claim more lives each year than all forms of cancer and chronic lower respiratory diseases combined. The use of machine learning in healthcare was made possible when patient data is being stored in the form of electronic records. Of the many applications of machine learning in healthcare industry, disease prediction is gaining popularity.

In this project, it is demonstrated that we could predict whether a patient has cardiovascular disease or not by applying various machine learning algorithms to the patient data. The three models Logistic regression, Support Vector Machine (SVM) and Naïve Bayes Classifier have shown accuracy of above 70% which can be promising for further studies.

References

AHA. (2020). *American Heart Association.* Retrieved from heart.org: https://www.heart.org/-/media/files/about-us/statistics/2020-heart-disease-and-stroke-ucm_505473.pdf?la=en

Mordecai, A. (n.d.). *towards data science*. Retrieved from https://towardsdatascience.com/heart-disease-risk-assessment-using-machine-learning-83335d077dad

Appendix A - Web resources

- First, this is the link to the dataset for this project - https://www.kaggle.com/sulianova/cardiovascular-disease-dataset
- The CDC website has compiled data about the causes, statistics and trends in cardiovascular diseases. This is a great resource to explore that provides almost everything that is needed to understand this disease. https://www.cdc.gov/heartdisease/statistics_maps.htm
- The American Heart Association provides a latest report on statistics, and it is reported that the leading causes of cardiovascular disease are smoking, physical inactivity, poor nutrition, obesity, high blood pressure, diabetes and high cholesterol. https://www.heart.org/-/media/files/about-us/statistics/2020-heart-disease-and-stroke-ucm_505473.pdf?la=en
- For the exploratory analysis, I will be applying the techniques discussed in this text book. This text book is a great source for me to perform basic exploratory analysis to determine the nature of data. Think stats – Exploratory data analysis by Allen Downey
- Story telling with data, a data visualization guide for business professionals by Coke Nussbaumer Knaflic – For ideas to present the project that appeals to audience
- I will be referring to this article to help me with the choice of machine learning algorithms that are suitable for this project. https://www.edureka.co/blog/classification-in-machine-learning/
- I will be referring to this book to help with building a model for the predictive analysis. – Applied Predictive Analytics - Principles and Techniques for the Professional Data Analyst by Dean Abbott.
- Finally, I will be referring to this article that demonstrates a model to heart attacks using machine learning. https://towardsdatascience.com/heart-disease-risk-assessment-using-machine-learning-83335d077dad

Appendix B – Data description

Features:

1. Age | Objective Feature | age | int (days)
2. Height | Objective Feature | height | int (cm) |
3. Weight | Objective Feature | weight | float (kg) |
4. Gender | Objective Feature | gender | categorical code |
5. Systolic blood pressure | Examination Feature | ap_hi | int |
6. Diastolic blood pressure | Examination Feature | ap_lo | int |
7. Cholesterol | Examination Feature | cholesterol | 1: normal, 2: above normal, 3: well above normal |
8. Glucose | Examination Feature | gluc | 1: normal, 2: above normal, 3: well above normal |
9. Smoking | Subjective Feature | smoke | binary |
10. Alcohol intake | Subjective Feature | alco | binary |
11. Physical activity | Subjective Feature | active | binary |
12. Presence or absence of cardiovascular disease | Target Variable | cardio | binary |