“Cardiovascular Disease”

Abstract:

In the current area deaths due to cardiovascular disease have become a major issue. Approximately one person dies per minute due to cardiovascular disease. Data is generated and has to be stored daily because of fast growth in Information Technology. The data which is collected is converted into knowledge by data analysis by using various combinations of algorithms. Medical professionals working in the field of cardiovascular disease have their own limitations; they cannot predict the chance of getting cardiovascular disease up to high accuracy. These paper alms to improve Cardiovascular Disease predict accuracy using the Logistic Regression model of machine learning considering the health care dataset which classifies the patients whether they are having heart disease or not according to the information in the record

Keywords: - Cardiovascular Disease; Data Analysis; Machine Learning; Logical Regression Algorithms.

Introduction:

The load of cardiovascular disease is rapidly increasing all over the world from the past few years. Even if these diseases has found as the most important source of death, it has been announced as the most manageable and avoidable disease [1]. Mainly, blockage in arteries causes heart stroke. It occurs when heart does not pump the blood around the body efficiently.

Having high blood pressure is also one of the causes of getting a heart disease. A survey says that, in 2017 to 2020, the commonness of hypertension in the world was about 35%, which is also a cause of heart disease. Similarly, there are many more reasons for getting a heart disease such as obesity, not taking in proper nutrition, increased cholesterol and lack of physical activity. So, prevention is very necessary. For prevention, awareness of heart disease is important. Around 47% of people die outside the hospital and it shows that they don’t act on early warning signs.

Nowadays, lifespan of a human being is reduced because of heart diseases. So, World Health Organization (WHO) developed targets for prevention of non-communicable diseases (NCDS) in 2019, in which, 25% of relative reduction is form cardiovascular diseases and it is being ensured that at least 50% of patients with cardiovascular diseases have access to relevant drugs and medical counseling by 2025[2]. Around 17.9 million people died just because of cardiovascular disease in 2016, which is 31% of deaths around the world.

A major challenge in heart diseases is its detection [3]. It is difficult to predict that a person that a person has a heart disease or not. There are instruments available which can predict heart diseases but either they are expensive or are not efficient to calculate the chance of heart disease in human [4]. A survey of World Health Organization (WHO) says that medical professionals are able to predict just 67% heart disease, so there is vast scope of research in this field. In case of India, access to good doctors are hospitals in rural area is very low. A 2016 WHO report says that, just 58% of the doctors have medical degree in urban areas and 19% in rural areas.

In USA, someone has a heart attack every 40 second, that is, more than one person dies in USA due to heart attack. A part from this, Turkmenistan has the highest rate of deaths till 2012, with 712 deaths per 100,000 people. Whereas, Kazakhstan has the second highest rate of deaths due to heart diseases. India holds 56th position in this series. Study also shows that, at ages 30-69 years, 1.3 million cardiovascular deaths, 0.9 million (68.4%) were caused by coronary heart disease and 0.4 million (27%) by stoke.

Heart diseases are a major challenge in medical science; Machine Learning could be a good choice for predicting any heart disease in human. Heart diseases can be predicted using neural network, Decision Tree, KNN, etc. Later in this paper, we will see that how Logistic Regression is used to find the accuracy for that disease. It also shows that how ML will help in our future for heart disease.

Literature Review

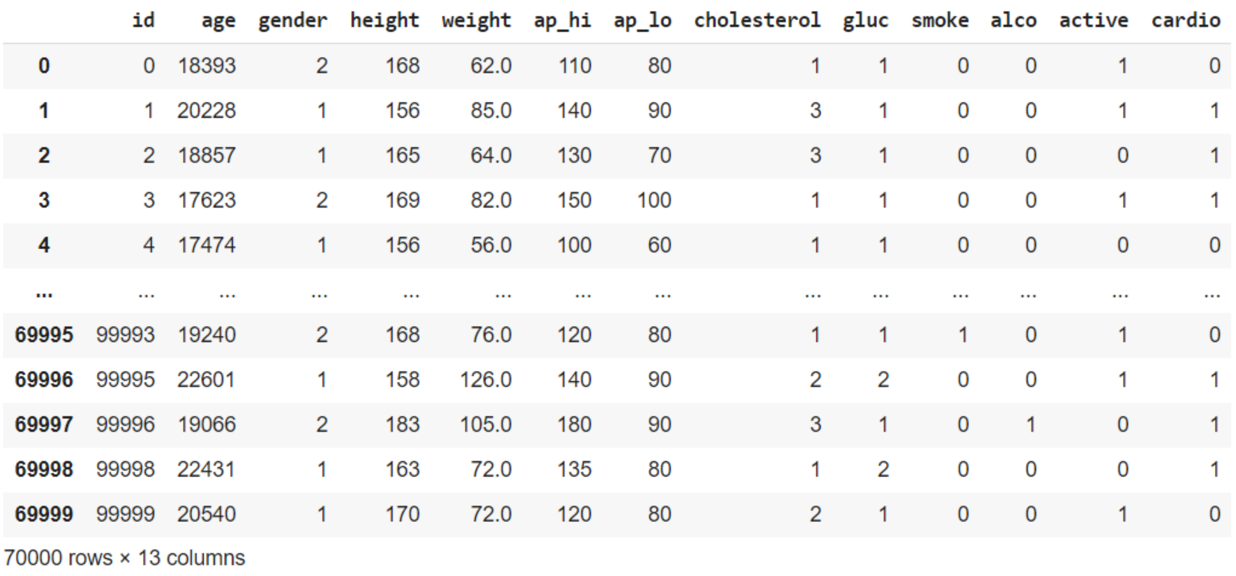
Description: It is a Machine Learning Project. Predicting Heart Disease using Random Forest. Cardiovascular Disease. Cardiovascular disease is a broad category for a range of diseases that are affecting heart and blood vessels. The early methods of forecasting cardiovascular diseases helped in making decisions about the changes to have occurred in high-risk patients which resulted in the reduction of their risks. The healthcare industry contains lots of medical data; therefore machine learning algorithms are required to make decisions effectively in the prediction of heart diseases. Recent research has delved into uniting these techniques to provide hybrid machine learning algorithms. In the proposed research, data preprocessing uses techniques like the removal of noisy data, removal of missing data, filling default values if applicable, and classification of attributes for prediction and decision making at different levels. The performance of the diagnosis model is obtained by using methods like classification, accuracy, sensitivity, and specificity analysis. This project proposes a prediction model to predict whether people have heart disease or not and to provide awareness or diagnosis on that. This is done by comparing the accuracies of applying rules to the individual results of Support Vector Machine, Gradient Boosting, Random Forest, Naive Bayes classifier, and logistic regression on the dataset taken in a region to present an accurate model of predicting cardiovascular disease.

Solution Approach: From the above statistics it is clear that the model is highly specific than sensitive. Men seem to be more susceptible to heart disease than men. Increase in age, number of cigarettes smoked per day and systolic Blood Pressure also show increasing odds of having heart disease. Total cholesterol shows no significant change in the odds of CHD. This could be due to the presence of good cholesterol (HDL) in the total cholesterol reading. Glucose too causes a very negligible change in odds (0.296). The model predicted with 0.87 accuracy. The model is more specific than sensitive. Overall model could be improved with more data.

How it will help them solve problem: From the data set of cardiovascular disease, I can find the solution through prediction and I can test the model and find the appropriate accuracy to solve the problem.

Conclusion: The amount of heart disease can excite the control line and reach a maximum point. Heart disease is complicated and each and every year lots of people are dying with this disease. It is difficult to manually determine the odds of getting heart disease based on risk factors. By using this system one of the major drawbacks of this work is that it’s main focus is aimed only at the application of classifying techniques and algorithms for heart disease prediction, by studying various data cleaning and mining techniques that prepare and build a dataset.

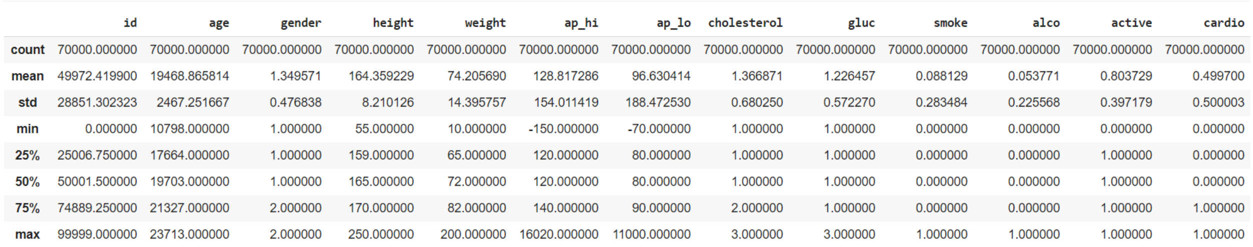
Methodology:



	id	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	cardio
0	0	18393	2	168	62.0	110	80	1	1	0	0	1	0
1	1	20228	1	156	85.0	140	90	3	1	0	0	1	1
2	2	18857	1	165	64.0	130	70	3	1	0	0	0	1
3	3	17623	2	169	82.0	150	100	1	1	0	0	1	1
4	4	17474	1	156	56.0	100	60	1	1	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
69995	99993	19240	2	168	76.0	120	80	1	1	1	0	1	0
69996	99995	22601	1	158	126.0	140	90	2	2	0	0	1	1
69997	99996	19066	2	183	105.0	180	90	3	1	0	1	0	1
69998	99998	22431	1	163	72.0	135	80	1	2	0	0	0	1
69999	99999	20540	1	170	72.0	120	80	2	1	0	0	1	0

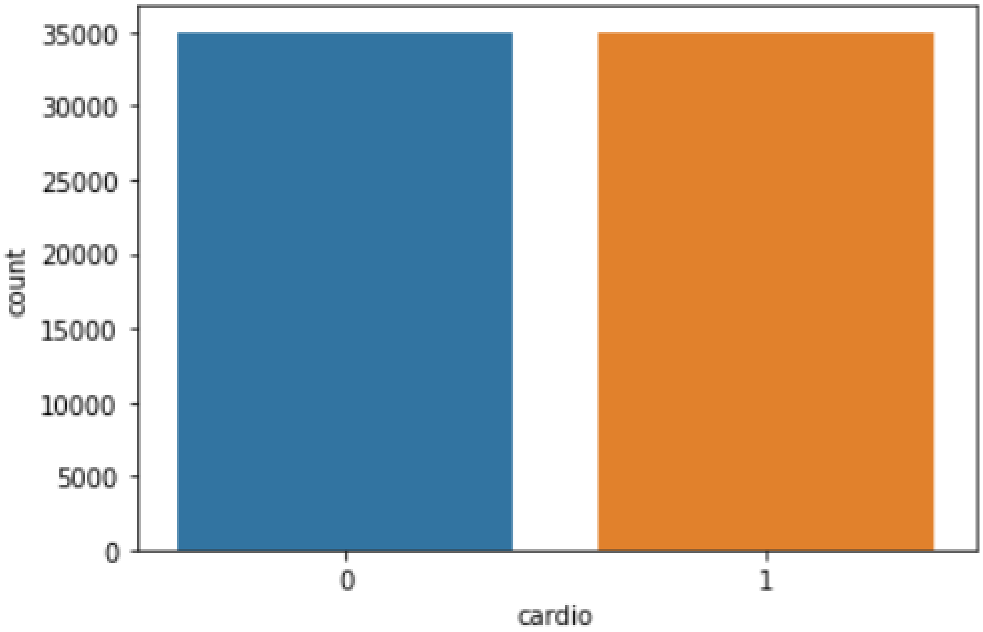
70000 rows × 13 columns

We can see that we get back seven rows or seven individual’s right that are classified as having a heart disease or not. so we have our id column here . We have our age column which looks a title strange right. You don’t know anybody who’s eighteen thousand three hundred and ninety three years old and that’s. Because the age is in days and not years. They we have the gender we have that person’s height in centimeters. We have that person’s weight and kilograms. We have this column called AP_ high and that’s the systolic blood pressure and we have this ap_lo which is the diastolic blood pressure and then we have that person’s cholesterol. We have a column called gluc or gl you see and that the person’s glucose. Then, we have a column called smoke which tells us if that individual smokes are not a common call Alco which tells us the person is alcohol intake. We active which tells us how physical how physically active that person. We have our target column called cardio. Which tells us if this individual has a cardiovascular disease or not. So, if they don’t then it’s zero and again if they do is one all right. So I think that’s pretty good kind of explaining the data here.

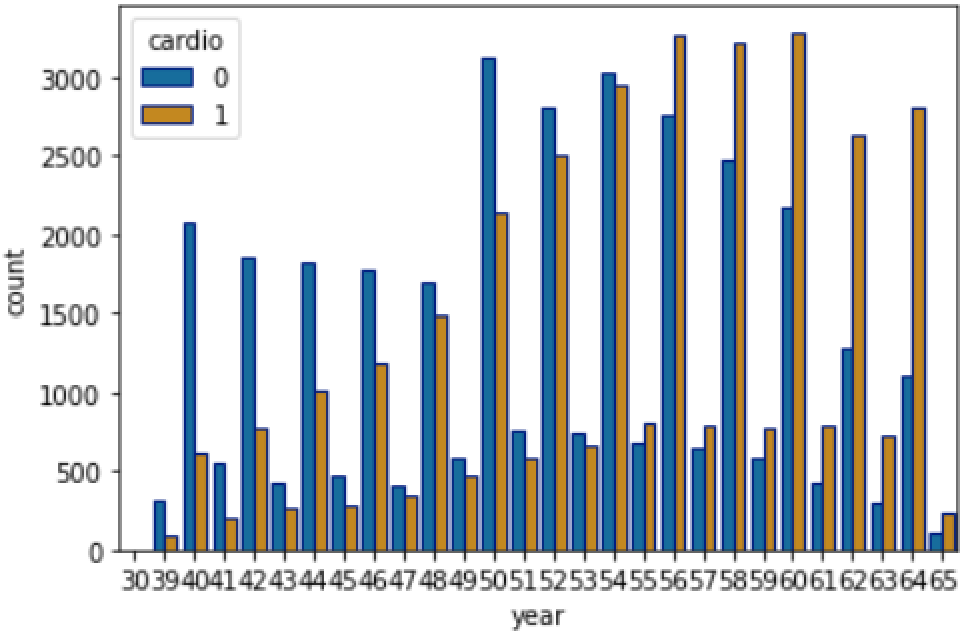


	id	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	cardio
count	70000.000000	70000.000000	70000.000000	70000.000000	70000.000000	70000.000000	70000.000000	70000.000000	70000.000000	70000.000000	70000.000000	70000.000000	70000.000000
mean	49972.419900	19468.865814	1.349571	164.359229	74.205690	128.817286	96.630414	1.366871	1.226457	0.088129	0.053771	0.803729	0.499700
std	28851.302323	2467.251667	0.476838	8.210126	14.395757	154.011419	188.472530	0.680250	0.572270	0.283484	0.225568	0.397179	0.500003
min	0.000000	10798.000000	1.000000	55.000000	10.000000	-150.000000	-70.000000	1.000000	1.000000	0.000000	0.000000	0.000000	0.000000
25%	25006.750000	17664.000000	1.000000	159.000000	65.000000	120.000000	80.000000	1.000000	1.000000	0.000000	0.000000	1.000000	0.000000
50%	50001.500000	19703.000000	1.000000	165.000000	72.000000	120.000000	80.000000	1.000000	1.000000	0.000000	0.000000	1.000000	0.000000
75%	74889.250000	21327.000000	2.000000	170.000000	82.000000	140.000000	90.000000	2.000000	1.000000	0.000000	0.000000	1.000000	1.000000
max	99999.000000	23713.000000	2.000000	250.000000	200.000000	16020.000000	11000.000000	3.000000	3.000000	1.000000	1.000000	1.000000	1.000000

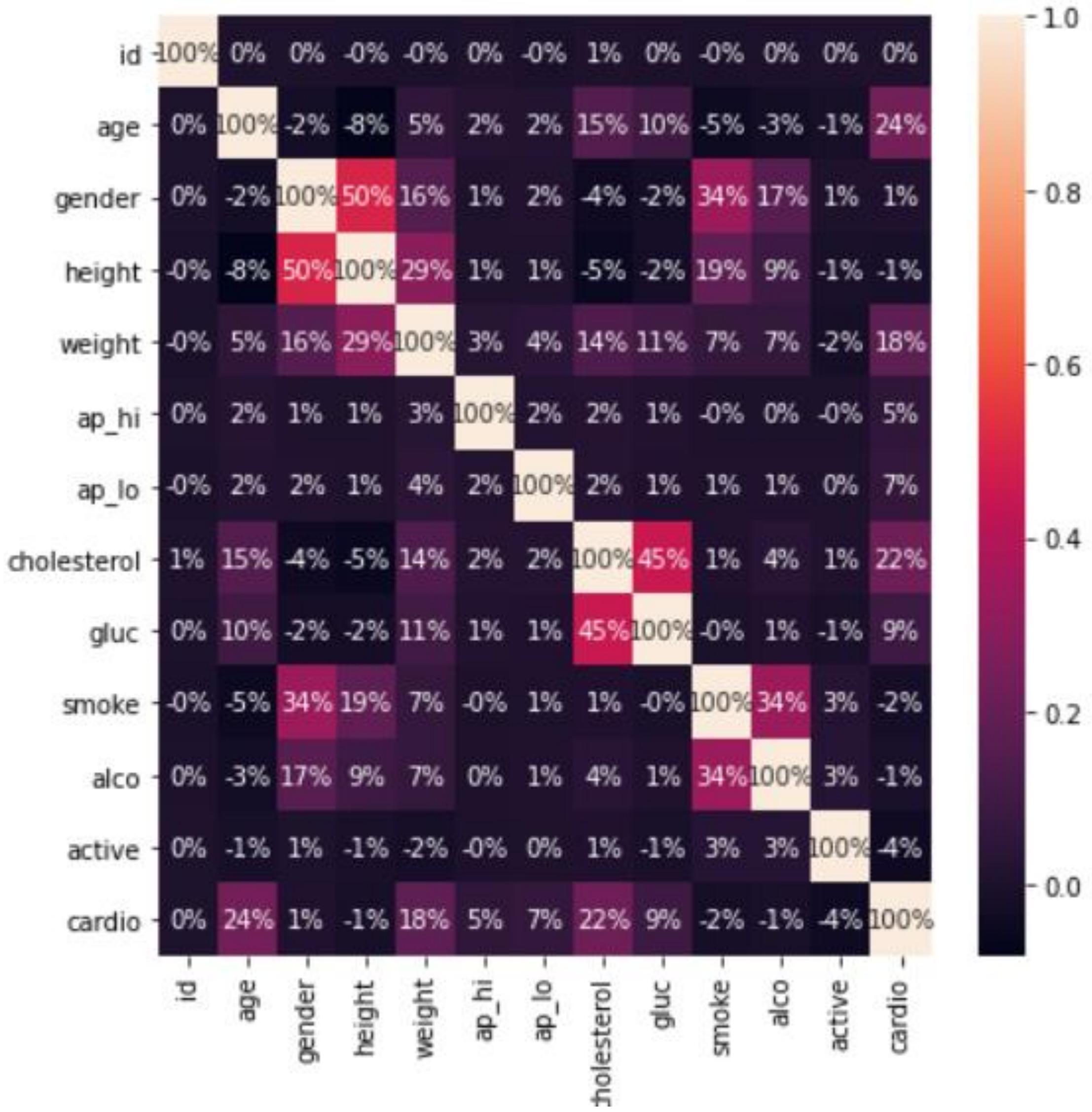
I want to get this standard deviation and I want see the mean. I want to see the maximum value and we can see all of that information for each of the columns here. So let’s take a look at the age column. We can see that the mean age is 19468.865814 days. The minimum value is 10798.000000 days. So I don’t know what that is in years but we have the days here. The maximum age is 23730 days old. We can see that there 35021 individuals in our data set that do not have a heart disease and there is about 34979 individuals that do have a heart disease or cardiovascular disease.



We can see that it is about even. I can’t really tell the difference here. So at least we have the numbers from our previous diagram to look at. But just from this chart here you can tell that’s it’s about even. We have about it you know half and half .



We can see the number of people with a heart disease and without a heart disease at that specific age. Now, what is interesting is here? We see that which is age 55 number of people with a cardiovascular disease exceeds that exceeds the number of people without a cardiovascular disease. So it’s kind of interesting to see that before age 55 that did not seem to be the case .If we look at the rest of these values it looks like those with a cardiovascular disease outnumber the people without a cardiovascular disease for that same age.



	id	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	cardio
id	100%	0%	0%	-0%	-0%	0%	-0%	1%	0%	-0%	0%	0%	0%
age	0%	100%	-2%	-8%	5%	2%	2%	15%	10%	-5%	-3%	-1%	24%
gender	0%	-2%	100%	50%	16%	1%	2%	-4%	-2%	34%	17%	1%	1%
height	-0%	-8%	50%	100%	29%	1%	1%	-5%	-2%	19%	9%	-1%	-1%
weight	-0%	5%	16%	29%	100%	3%	4%	14%	11%	7%	7%	-2%	18%
ap_hi	0%	2%	1%	1%	3%	100%	2%	2%	1%	-0%	0%	-0%	5%
ap_lo	-0%	2%	2%	1%	4%	2%	100%	2%	1%	1%	1%	0%	7%
cholesterol	1%	15%	-4%	-5%	14%	2%	2%	100%	45%	1%	4%	1%	22%
gluc	0%	10%	-2%	-2%	11%	1%	1%	45%	100%	-0%	1%	-1%	9%
smoke	-0%	-5%	34%	19%	7%	-0%	1%	1%	-0%	100%	34%	3%	-2%
alco	0%	-3%	17%	9%	7%	0%	1%	4%	1%	34%	100%	3%	-1%
active	0%	-1%	1%	-1%	-2%	-0%	0%	1%	-1%	3%	3%	100%	-4%
cardio	0%	24%	1%	-1%	18%	5%	7%	22%	9%	-2%	-1%	-4%	100%

We can see a little more visually the correlation between the columns alright. So we take a look at that cardio column here. We can see that it has about 24% positive correlation with age 18% positive correlation with weight 22% with cholesterol. So on so forth and also you will notice that. It has the same correlation number for both years and age has a 100% correlation.

Result Discussion:

The model is 0.9816% accurate on the training data set and the model got about 0.765% accuracy on our test data. So again that's not bad right it is not bad but when you're dealing with people and patient you still want your model to be much more accurate than that you know you really want it to be a 100% that’s not always. The case but we definitely want it to be better than 0.765% that’s it.