REPORT ON

Cardiovascular Disease Prediction using ML Techniques

EEN-300: INDUSTRY ORIENTED PROBLEM

Submitted by

Aditya Karad(18115010) Aditya Karad

Aman Tiwari(18115015) Aman

Ritik Mathur(18117083) Ritik Mathur

Yash K Gandhi(18115124)

Advisor: Prof. Ambalika Sharma

Department of Electrical Engineering
Indian Institute of Technology
Roorkee ROORKEE-247667, UK
(INDIA)

Spring 2021-22

Table of Contents

Abstract

Cardiovascular disease (CVD) is a very important matter to be considered as it is the biggest cause of death across the globe. CVD is not a disease. Rather it is a group of various diseases of the heart and blood vessels (the cardiovascular system). Usually, these are diseases which affect the heart and blood vessels of the heart and brain. There are a lot of factors which increase the risk of getting it. These factors are known as the “risk factors”.

Many lives can be saved from premature deaths by identifying people in the early stage of CVD or even people with a high risk of getting it. In this study, we aim to do so using ML techniques and algorithms. We will use three different ML algorithms namely Logistic Regression, K-Nearest Neighbours and Decision Tree Classifier to predict if a person has a CVD or not, based on their medical reports for various risk factors such as age, gender, blood pressure, etc. We will be comparing the results from these methods and conclude which is the best method to predict the presence/absence of CVD in a person.

Introduction

Cardiovascular Disease or CVD is a generic term for diseases or disorders related to heart or blood vessels. CVD includes heart attack, stroke, heart failure, hypertensive heart disease, rheumatic heart disease, cardiomyopathy, abnormal heart rhythms, congenital heart disease, valvular heart disease, carditis, aortic aneurysms, peripheral artery disease, thromboembolic disease and venous thrombosis. Generally, affected people do not have any symptoms in the early stages. A heart attack or stroke is one of the first symptoms which suggest that there is an underlying disease of the blood vessels. People can also have symptoms including (but not limited to) shortness of breath, light-headedness or faintness, nausea or vomiting, chest pain and fatigue.

The dataset used to train and validate our ML models contains 70,000 records of patients’ data and can be found at this Kaggle source [4]. Each record has twelve features namely age, gender, height, weight, systolic blood pressure, diastolic blood pressure, cholesterol value, glucose level, whether the person smokes or not, whether the person is alcoholic, whether the person is involved in physical activities, and lastly whether the person has a CVD. We use the first 11 features to predict the last feature. All the values in the dataset were collected during medical examination.

Table 1. Structure of the dataset

Feature	Description
id	Serial number of the current row. Not relevant
age	Age of the person in days
gender	Gender of the person. 1 for women, 2 for men
height	Height of the person in cm
weight	Weight of the person in kg
ap_hi	Systolic blood pressure in mmHg, i.e., the pressure exerted when blood is ejected into arteries
ap_lo	Diastolic blood pressure in mmHg, i.e., the pressure blood exerts within arteries between heartbeats
cholesterol	Cholesterol value in the blood. 1 if it is normal, 2 if it is above normal & 3 if it is well above normal
gluc	Glucose level in the blood. 1 if it is normal, 2 if it is above normal & 3 if it is well above normal
smoke	It's a binary value. 1 if the person smokes, 0 if not
alco	It's a binary value. 1 if the person is alcoholic, 0 if not
active	It's a binary value. 1 if the person is involved in physical activities, 0 if not
cardio	It's our target binary value. 1 if the person has CVD, 0 if not

CVDs can be prevented by improving upon the risk factors through a healthy diet, exercise, avoiding alcohol consumption and tobacco smoke. Maintaining blood pressure and managing risk factors like blood lipids and diabetes are proved to be fruitful. The risk of rheumatic heart disease can be decreased by treating people suffering from strep throat (streptococcal pharyngitis) with antibiotics. According to some estimates, it is said that about 90% of CVDs may be preventable. Therefore, it is of utmost importance to identify those at high risk of getting CVD as early as possible to start with time counselling and proper medication. Advancements in data science and machine learning have enabled us to do so in a quite reliable manner.

Following are some visualizations for the dataset distribution.
To understand the dataset distribution, we have considered the following visualizations which depict how pathological parameters are distributed with respect to other variables.

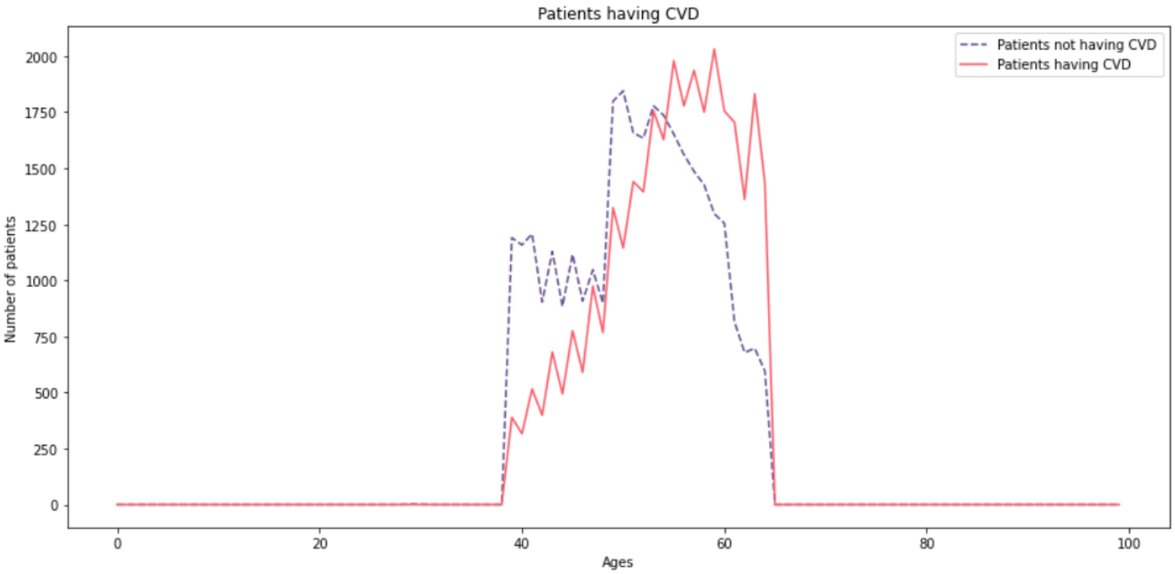


Figure 1. Number of patients having CVD across different age groups

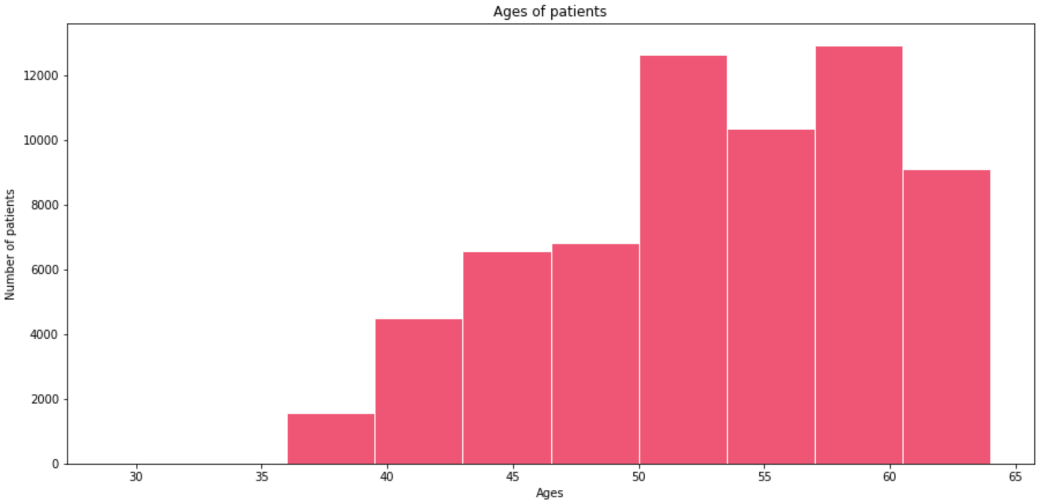


Figure 2. Age distribution of people in the dataset

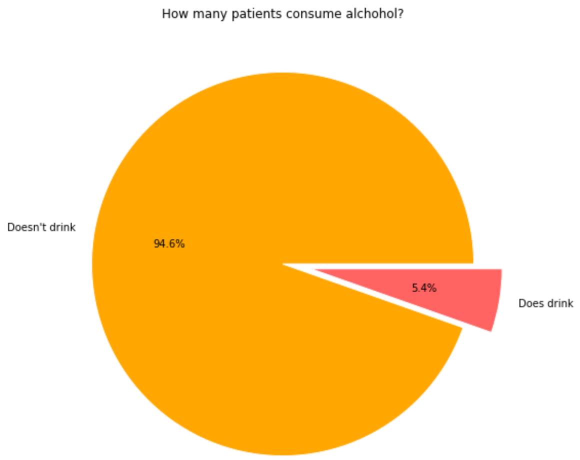


Figure 3. Percentage of people consuming alcohol

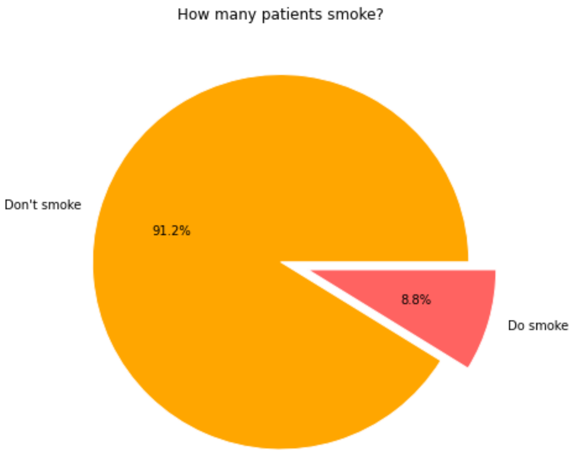


Figure 4. Percentage of people who smoke

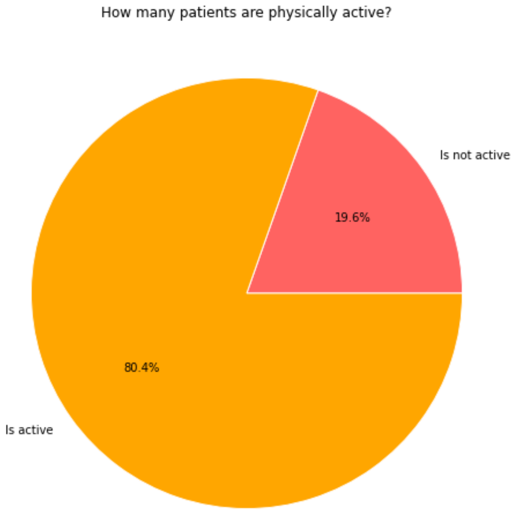


Figure 5. Percentage of people who are physically active

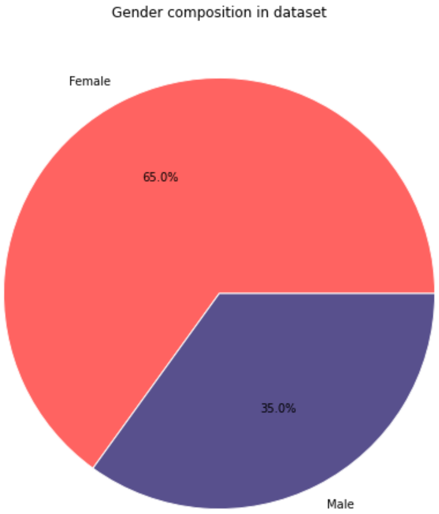


Figure 6. Gender composition in the dataset

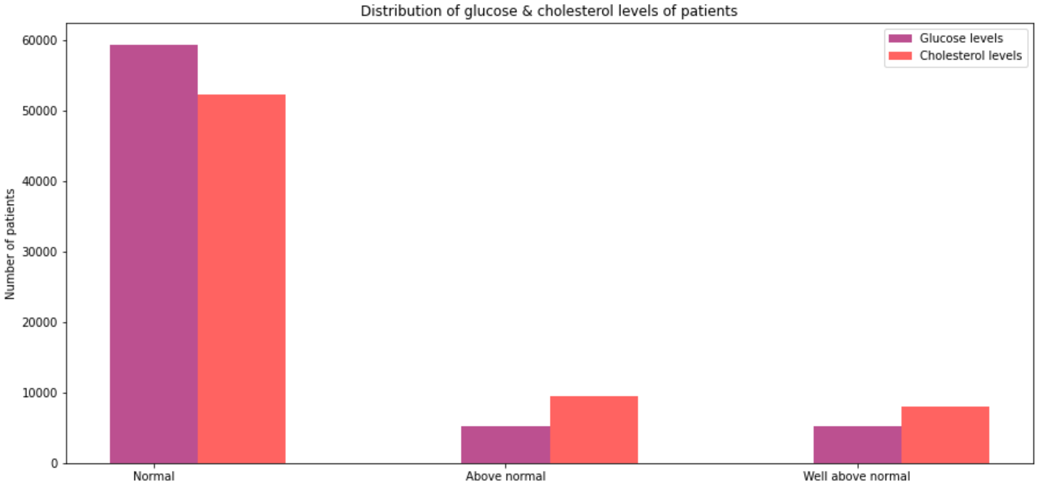


Figure 7. Distribution of glucose and cholesterol levels of patients

Motivation

Cardiovascular diseases (CVDs) are the leading cause of death globally. According to the World Health Organization, Deaths from CVDs are 17.9 million in 2019, which is 32% of all deaths in the world. Heart attack and stroke cause 85% of deaths out of all these deaths due to cardiovascular diseases. In 2019, 17 million premature deaths (under the age of 70) due to non-communicable diseases, out of which 38% deaths were due to cardiovascular diseases. Addressing behavioural risk factors such as tobacco use, unhealthy diet and obesity, physical inactivity, and harmful alcohol use can prevent most cardiovascular diseases.

At least three-quarters of cardiovascular deaths in the world occur in low and middle income countries. Middle-income countries suffering from cardiovascular diseases and other non-communicable diseases don’t have many opportunities to obtain effective and fair health services that meet their needs. As a result, for several people within these countries, detection is late in the course of the disease, and people die at a younger age from cardiovascular diseases.

In low and middle income countries, the poorest people are hit the hardest. There is evidence that cardiovascular disease and other non-communicable diseases cause poverty through catastrophic health and personal expenses at the household level.

It is essential to detect cardiovascular diseases as early as possible to start with time counselling and proper medication.

Data science analysis methods learn from historical data and make accurate predictions. They process patient data, analyse clinical records, discover interactions, symptom associations, common adjectives, habits, diseases, and make predictions. According to medical science, the effect of certain biological factors such as genome structure or clinical variability can predict certain diseases. The common reason is to predict the course of the disease and prevent it to reduce risks and side effects. The main benefit is to improve the quality of life of patients and the quality of medical conditions. Nowadays, machine learning algorithms help in the field of medical science, meet growing medical demands, improve operations, and reduce testing costs. Machine Learning can positively improve patient care delivery strategies. For example, it can help doctors to identify diseases on time, diagnose and treat disease.

Objectives

Train machine learning models by using different classification algorithms on the given dataset, predicting for sample data, and finding the accuracy of different models.

In this project, we use three classification machine learning algorithms [1] and predict values for the CVDs dataset. We represent the performance of different models with the help of confusion matrices and find out the accuracy of models. We can see the best model for the CVDs dataset among those three models is one that has a good accuracy.

We use correlation heatmap to find dependencies or relations between cardiovascular disease results and other factors like smoking, person's age, blood pressure, etc. Heatmap is a graphical representation of data in which data values are represented as colors. Correlation heatmap shows correlation between different numerical value columns in the dataset. Correlation is the measure of how two or more variables are related to each other. Correlation does not tell us about cause and effect; it refers to the degree to which two variables are related.

Discussion

As mentioned in the previous sections, early prediction of cardiovascular diseases is of utmost importance to detect the cause and initiate untimely treatment of the patient. In our study, we have identified machine learning as an efficient and robust method to predict the presence or absence of cardiovascular disease [3] in a patient on the basis of numerical pathological data available from the patients' clinical tests and history. In this section, we have discussed the distribution and visualization of the dataset used for cardiovascular disease prediction and the machine learning models applied for the same. The dataset consists of 70,000 rows where each row includes the following pathological information about the patient: patient id, age, height, weight, gender, systolic blood pressure, diastolic blood pressure, whether or not the person smokes on a regular basis, whether or not the person drinks alcohol on a regular basis, whether or not the person is physically active, cholesterol levels and glucose levels. Further, each row includes information about whether or not the person has suffered from cardiovascular disease. Together this will form the training data for supervised learning of the machine learning model. Following graphs depict the distribution of training parameters such as weight, glucose intake, cholesterol levels, and blood pressure with respect to presence or absence of cardiovascular disease.

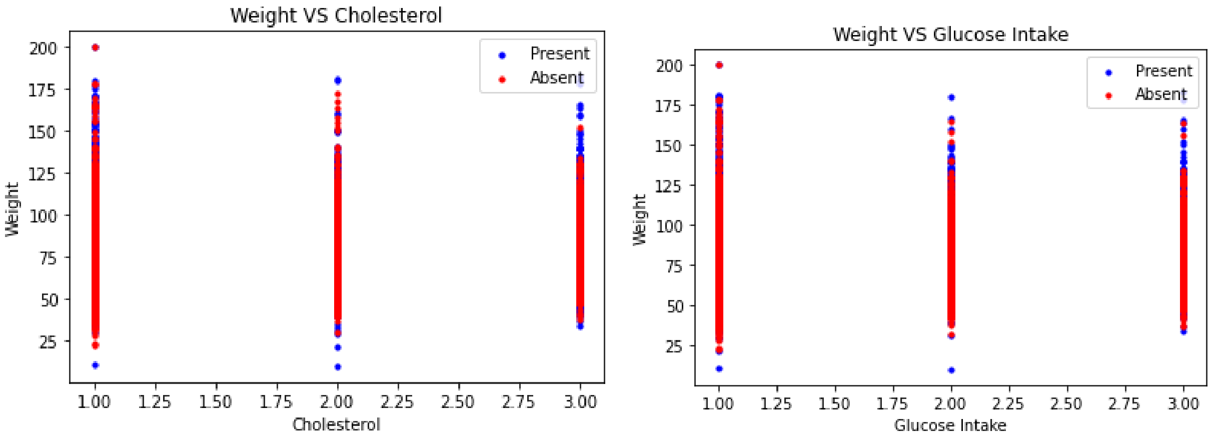


Figure 8, 9. Plot showing relation between weight, cholesterol, glucose & presence of CVD

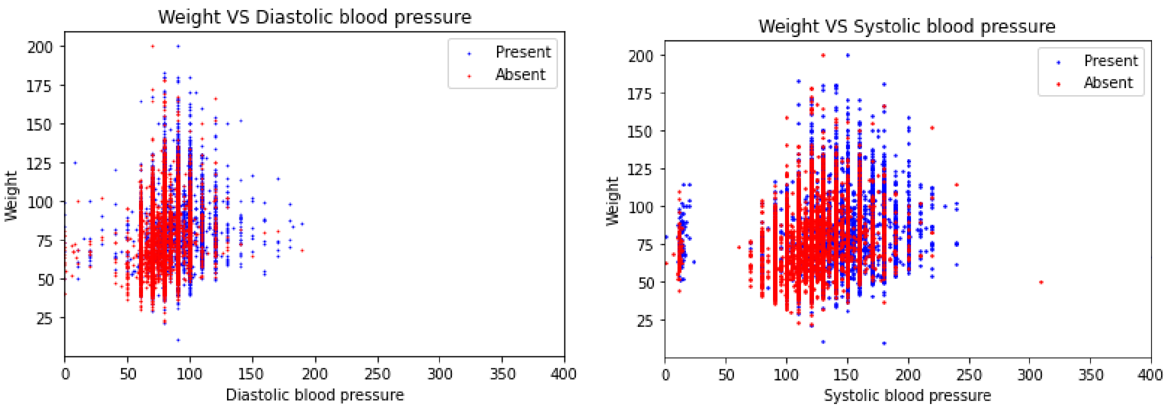


Figure 10, 11. Plot showing relation between weight, cholesterol, glucose & presence of CVD

The dataset is provided and imported in CSV format for model training, followed by data preprocessing where missing or null values are removed from the dataset rows. In addition to this, we have dropped the “patient id” column since we have to consider only features required for binary classification model training. To maintain consistency in our dataset, we have performed outlier removal where we removed rows having value of systolic blood pressure lesser than diastolic blood pressure. Further we removed outliers found by the box plot distribution of systolic and diastolic blood pressure which is shown in the following figure.

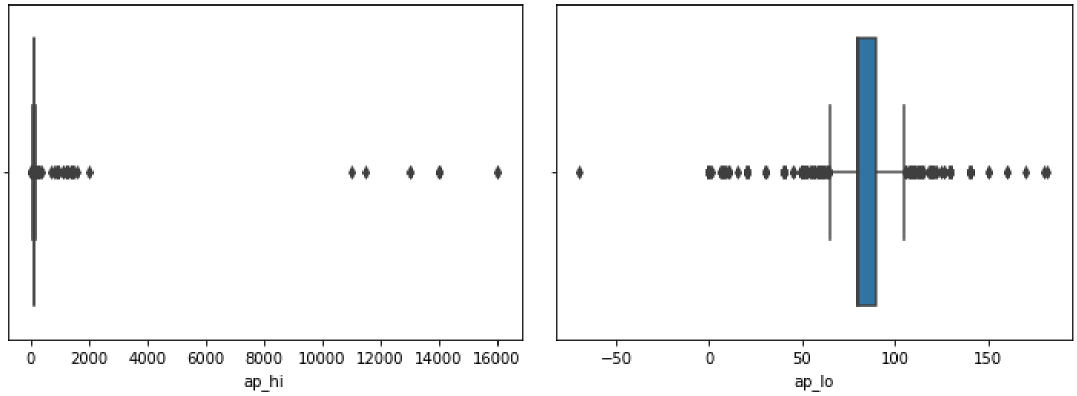


Figure 12, 13. Box plot showing outliers in ap_hi and ap_lo

The dataset is then split into dependent and independent variables, here independent variables include pathological information about the patient such as age, height, weight, gender, systolic blood pressure, diastolic blood pressure, whether or not the person smokes on a regular basis, whether or not the person drinks alcohol on a regular basis, whether or not the person is physically active, cholesterol levels and glucose levels and dependent variable is a binary variable indicating whether or not the patient has cardiovascular disease. To split the dataset into train-test in a ratio of 80:20 we have used the train_test_split library. The numeric values are then normalized using StandardScaler library to remove redundancy and complexity in the dataset and to convert all values to a similar scale. Following graph depicts the distribution of patient height and weight with respect to presence or absence of cardiovascular disease. Here data points marked in orange indicate the presence of cardiovascular disease while the data points marked in blue indicate the absence of cardiovascular diseases.

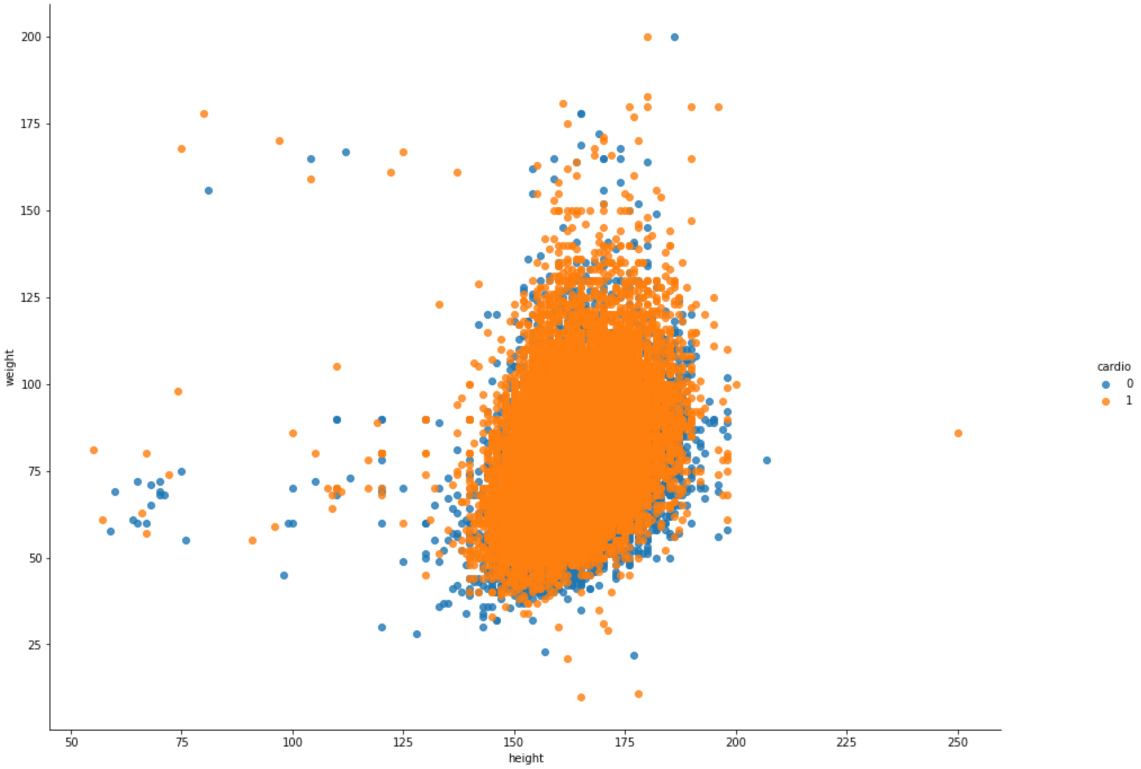


Figure 14. Plot showing relation between height, weight & presence of CVD

We have applied 3 supervised learning-based classification models to analyze the dataset: Logistic Regression [2], K-Nearest Neighbours and Decision Tree Classifier. Based on the model training, we have used accuracy scores and confusion matrix to establish a comparative opinion about the model performance.

K value is an important parameter in the K-Nearest Neighbours algorithm which greatly impacts the classification accuracy. To determine the most favourable value of K, we have analysed the False Negatives VS K value as well as the Accuracy VS K value curves which are shown below. On closer analysis of the following curves, we can conclude that K=13 is the most suitable candidate for our KNN classification problem as it simultaneously minimizes the false negatives and has good accuracy.

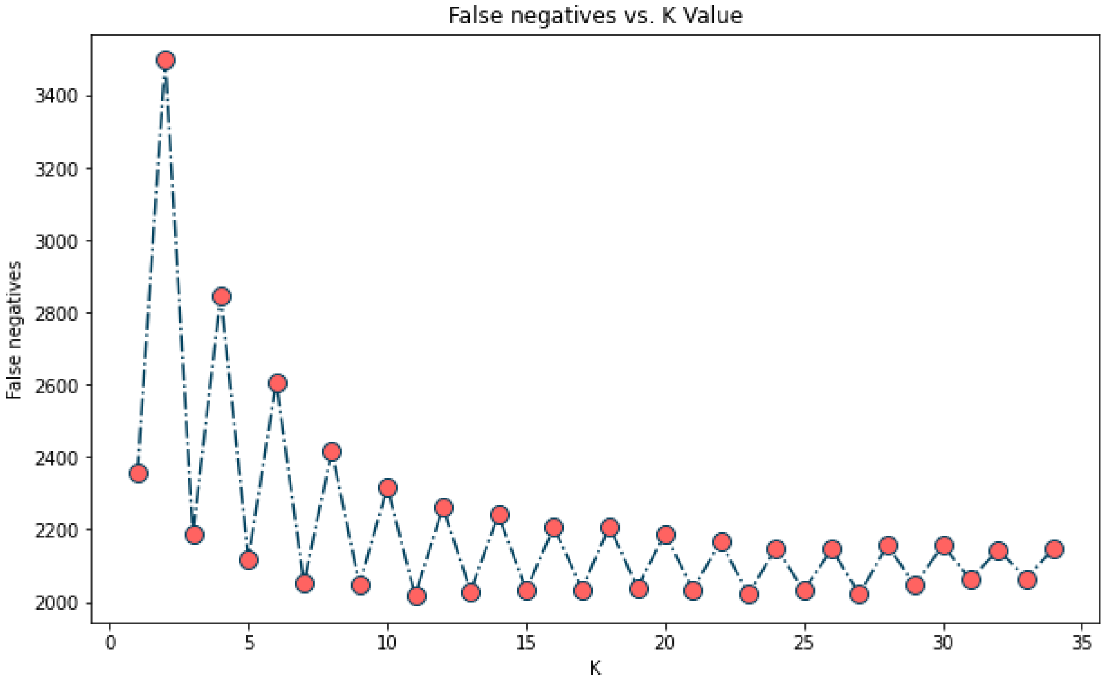


Figure 15. False negatives VS K value

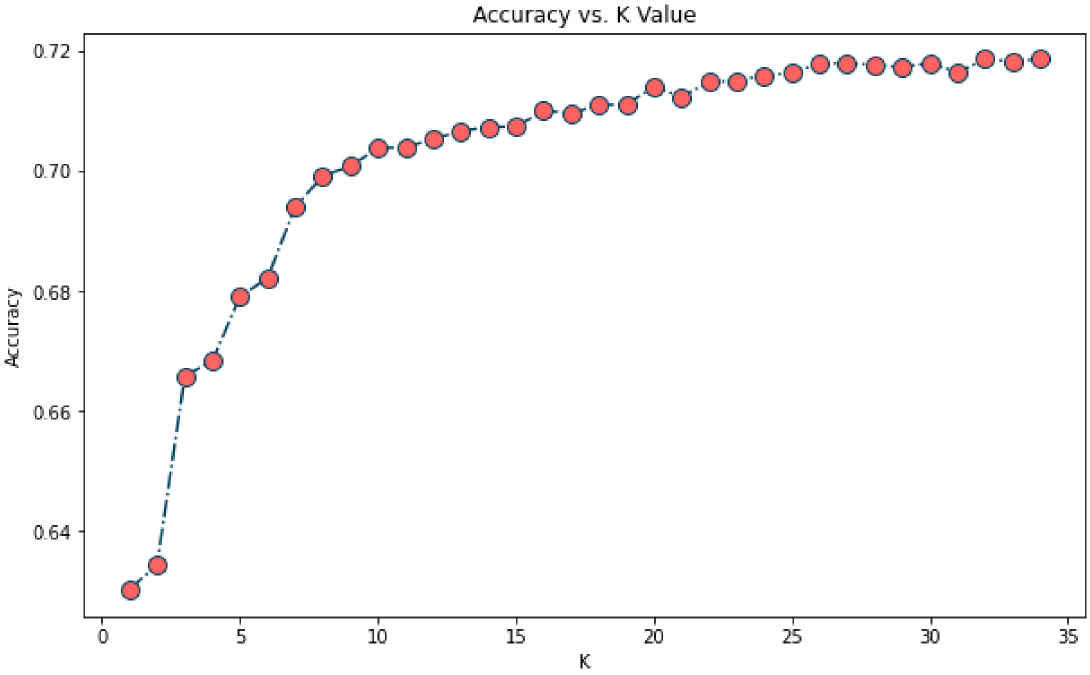


Figure 16. Accuracy VS K value

Decision tree is an algorithm which is very prone to overfitting. Overfitting can be handled by tuning some hyperparameters or by pruning methods. We tuned various hyperparameters such as criterion(gini/entropy), min_smaples_split, max_depth and ccp_alpha and compared the results to choose the best tuning.

We plotted the variation of accuracy and the number of false negatives with the variation in the max_depth hyperparameter as shown in the figure. We can see that at max_depth of 9, the accuracy is high(71.58%) and also the number of false negatives are very less(1895) thus making it an optimal depth.

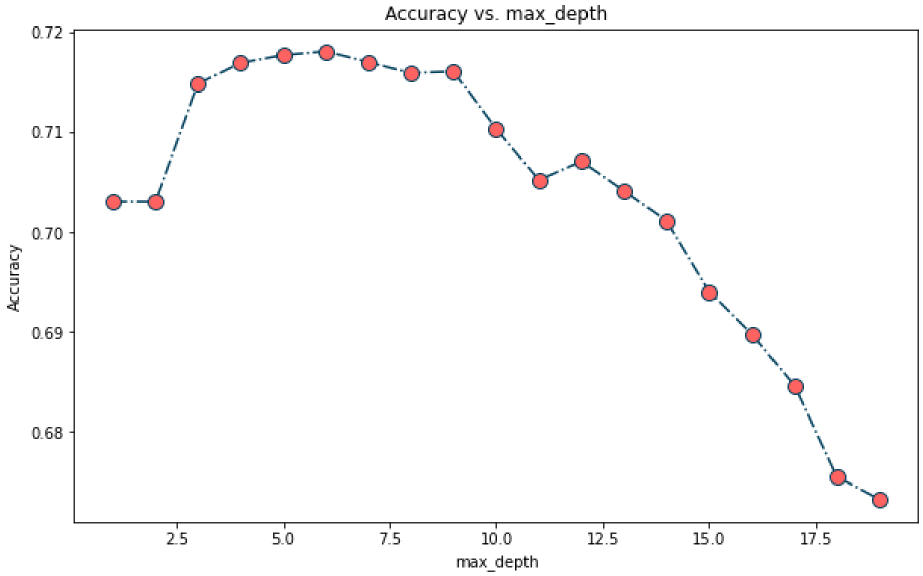


Figure 17. Variation of accuracy with max_depth

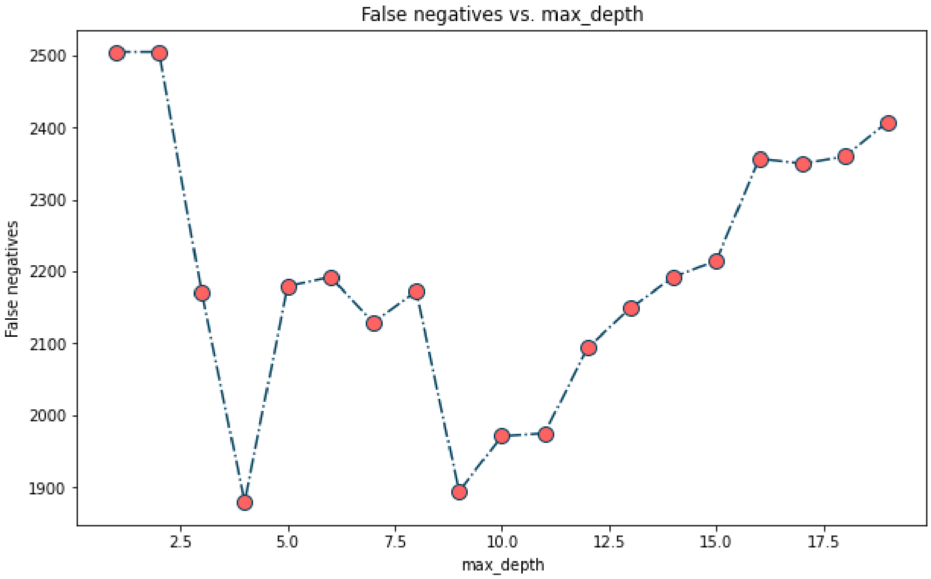


Figure 18. Variation of false negatives with max_depth

Pruning the tree without tuning other hyperparameters gave us similar results. The number of false negatives were slightly higher in this case and thus we had to remove it from the model. The plot for variation in training and testing accuracies with different levels of pruning (alpha) is shown below.

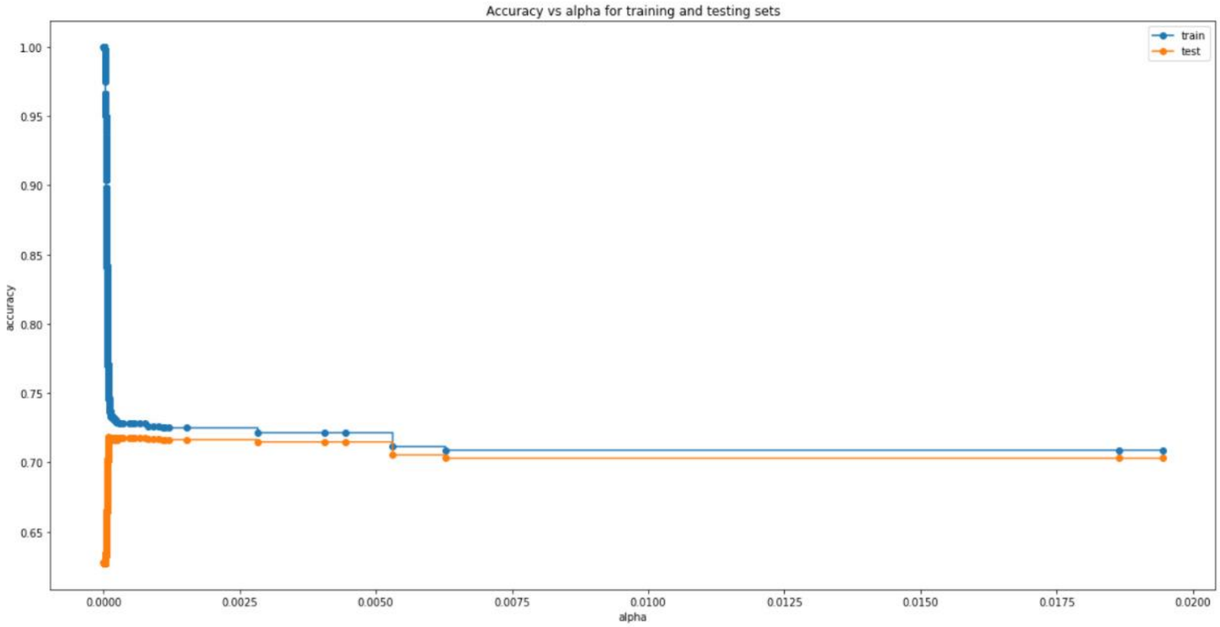
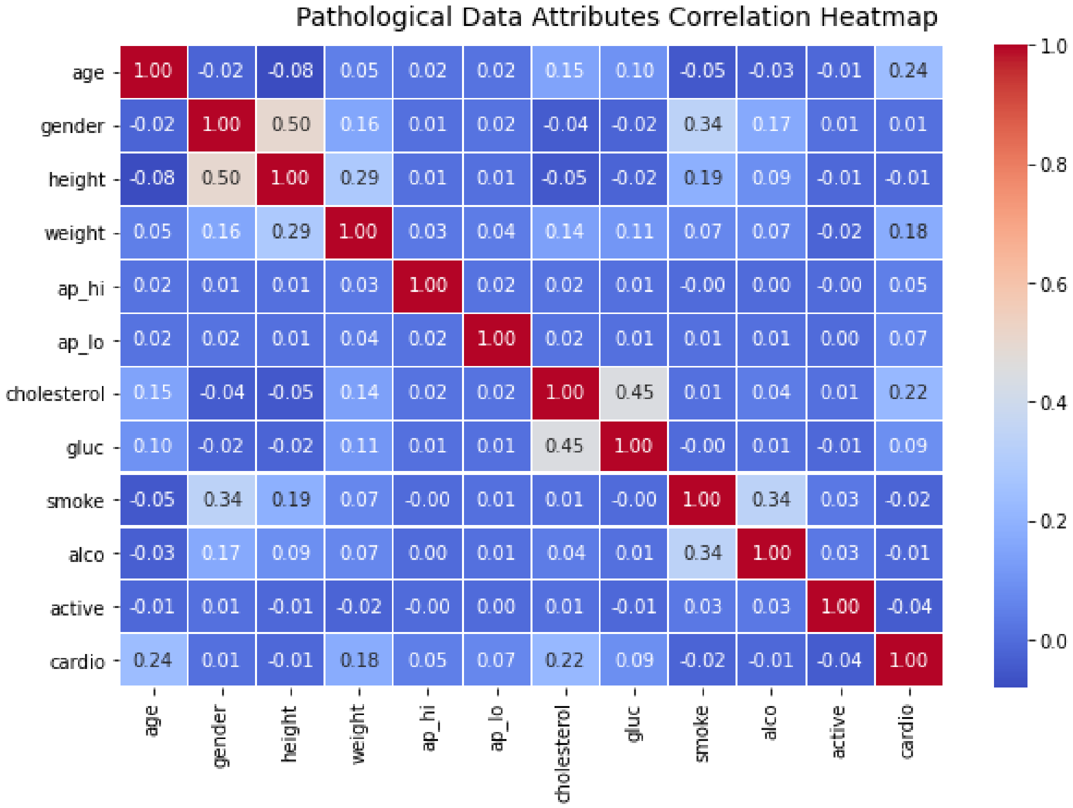


Figure 19. Variation of training and testing accuracies with alpha

To get an idea about the correlation between different pathological parameters in the dataset, we plotted the following correlation heat map.



	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	cardio
age	1.00	-0.02	-0.08	0.05	0.02	0.02	0.15	0.10	-0.05	-0.03	-0.01	0.24
gender	-0.02	1.00	0.50	0.16	0.01	0.02	-0.04	-0.02	0.34	0.17	0.01	0.01
height	-0.08	0.50	1.00	0.29	0.01	0.01	-0.05	-0.02	0.19	0.09	-0.01	-0.01
weight	0.05	0.16	0.29	1.00	0.03	0.04	0.14	0.11	0.07	0.07	-0.02	0.18
ap_hi	0.02	0.01	0.01	0.03	1.00	0.02	0.02	0.01	-0.00	0.00	-0.00	0.05
ap_lo	0.02	0.02	0.01	0.04	0.02	1.00	0.02	0.01	0.01	0.01	0.00	0.07
cholesterol	0.15	-0.04	-0.05	0.14	0.02	0.02	1.00	0.45	0.01	0.04	0.01	0.22
gluc	0.10	-0.02	-0.02	0.11	0.01	0.01	0.45	1.00	-0.00	0.01	-0.01	0.09
smoke	-0.05	0.34	0.19	0.07	-0.00	0.01	0.01	-0.00	1.00	0.34	0.03	-0.02
alco	-0.03	0.17	0.09	0.07	0.00	0.01	0.04	0.01	0.34	1.00	0.03	-0.01
active	-0.01	0.01	-0.01	-0.02	-0.00	0.00	0.01	-0.01	0.03	0.03	1.00	-0.04
cardio	0.24	0.01	-0.01	0.18	0.05	0.07	0.22	0.09	-0.02	-0.01	-0.04	1.00

Figure 20. Correlation heat map between all pairs of parameters

We tried feature-selection by dropping the least correlated features with presence of cardiovascular disease. These dropped fields were “height”, “active”, “smoke” and “alco”. Upon running the same models for it, we observed an insignificant increase in accuracy & a significant increase in false negatives.

Results

We trained Logistic regression, K-nearest neighbors & decision tree classifier which received similar results in terms of accuracy. Decision tree classifier fit our model the best with an accuracy of ~ 71.6%.

Table 2. Accuracy & confusion matrix values of all classification models

Model	Accuracy (%)	True Positives	True Negatives	False Positives	False Negatives
Logistic Regression	71.38	4242	4967	1489	2202
K-Nearest neighbors	70.66	4414	4701	1755	2030
Decision Tree Classifier	71.58	4685	4685	1771	1895

The table also shows the confusion matrix for all the models, which consists of 4 values: true positives, true negatives, false positives & false negatives.

Table 3. Table explaining confusion matrix values

	Patient has disease	Patient doesn't have disease
Model predicts patient has disease	True Positive	False Positive
Model predicts patient doesn't have disease	False Negative	True Negative

Since we are building a model which can predict a disease which impacts a human's life, we must optimise to reduce the false negatives. The reason is because even if a patient is wrongly tested positive from the model, he/she can perform an actual test and test negative in it. But if the patient is wrongly tested negative from our model, then he/she might not take a test, causing harm to his/her life.

Hence we must optimise to minimise false negatives. For this purpose, Decision tree classifier seems to be the best method for our purpose, due to highest accuracy & lowest false negatives.

Conclusion

Cardiovascular diseases are a leading cause of death and early prediction is of utmost importance to detect the cause and initiate treatment of a patient. We used supervised learning-based classification algorithms to classify if a patient has cardiovascular disease from his/her pathological data.

We tried to experiment with 3 models to fit the data. Since the accuracy of all models was roughly the same, we decided to choose a model which minimised false negatives. For this purpose, decision tree classifier is the best model, due to it’s high accuracy as well as low false negatives. Making the dataset more diverse by including more clinical information pertaining to cardiovascular systems as well as enhancing the model complexity can be the future scope for the project.

References

1. A Method of Cardiovascular Disease Prediction using Machine Learning | IJERT | A. Devi
2. Logistic Regression Analysis To Determine Cardiovascular Diseases Risk Factors A Hospital-Based Case-Control Study, 2019. | Ebtehag Mustafa
3. K. Hariharan, W. S. Vigneshwar, N. Sivaramakrishnan, V. Subramaniyaswamy. A Comparative Study on Heart Disease Analysis using Classification Techniques. International Journal of Pure and Applied Mathematics, Academic Publishing Ltd, 2018, 119 (12e), pp.13357-13366. ffhal-01826700f
4. Cardiovascular disease dataset. Contains 11 features for 70000 records of patient data - https://www.kaggle.com/sulianova/cardiovascular-disease-dataset