Anekant Education Society's

Tuljaram Chaturchand College of Arts, Science, and Commerce, Baramati 413102

Department of Statistics

A PROJECT REPORT

ON

"CARDIOVASCULAR DISEASE ANALYSIS"

A PROJECT REPORT SUBMITTED

By

Ms. Dhame Puja Shivaji — Roll No. 17031

Ms. Gawade Monali Balu — Roll No.17032

Ms. Komkar Vaishnavi Rahul — Roll No.17033

TO
DEPARTMENT OF STATISTICS
UNDER THE GUIDANCE OF
Mr. Swami. C. P.
2022-23

Department of Statistics

CERTIFICATE

This is to certify that, Miss. Dhame Puja Shivaji, Miss. Gawade Monali Balu and Miss. Komkar Vaishnavi Rahul are regular students of the Department of Statistics. A project on "**CARDIOVASCULAR DISEASE ANALYSIS**" Is submitted in partial fulfillment of the program in M.Sc. – II to the Department of Statistics, Tuljaram Chaturchand College of Arts, Science and Commerce, Baramati.

This project has been conducted under my supervision and guidance.

Place: Baramati

Mr. Swami C. P. EXAMINER Prof. Dr. Jagtap A. S.

Project Guide (Head of Department of Statistics)

ACKNOWLEDGEMENT

In the accomplishment of this project successfully, many people have best owned upon me their guidance and assistance. All the work that we have done in this project is only due to such supervision and assistance, this time we are utilizing this to thank all the people who have been concerned with this project. First of all, we wish to express my sincere gratitude and due respect to our Principal Prof Dr. C.V. Murumkar Sir for their continuous encouragement, positive support, and inspiration. Also, I thank Prof. Dr. A. S. Jagtap Sir, Head of the Department of Statistics for providing us with an opportunity to do this project and giving us all the support and valuable guidance for the successful completion of this project work duly.

The timely and successful completion of this project could hardly be possible without the guidance and assistance of our guide Mr. C. P . Swami sir, whose valuable guidance has been the ones that helped us patch this project work and make it a full-proof success.

We are thankful to all the Teaching Staff of the Department of Statistics for giving their priceless guidance and encouragement to embark on this project work. Also, we would like to express wholehearted thanks to my friends for their help and support in the completion of the project work.

INDEX

ABSTRACT

The use of machine learning in medical science is increasing particularly. We aim to build optimized ensemble models with ML models for disease prediction. In this project, we analyzed Cardiovascular Disease (CVD) data set to find which attributes affect the disease. Then we used different machine learning models to predict whether the patient has cardiovascular disease or not.

For this study, data is taken from the Kaggle and the dataset contains information about a patient who has cardiovascular disease. Data contains 70000 cases and 12 attributes, for this data we use machine learning algorithms (Logistic Regression, Decision Tree, Random Forest, KNN.) The accuracy of Random Forest is 72 %, precision is 72.98 % and recall is 70.17 %. So, we conclude that the Random Forest is the best-fit model.

Key Words: Cardiovascular disease, Machine learning algorithms, Logistic Regression, Decision Tree Classifier, Random Forest Classifier, K-Nearest Neighbor Classifier.

Software:

- MS-Excel
- Python
- R-Software

INTRODUCTION

Cardiovascular diseases (CVDs) are the leading cause of death globally. An estimated 17.9 million people died from CVDs in 2019, representing 32% of all global deaths. Of these deaths, 85% were due to heart attack and stroke. Over three-quarters of CVD deaths take place in low- and middle-income countries. Out of the 17 million premature deaths (under the age of 70) due to non-communicable diseases in 2019, 38% were caused by CVDs.

Cardiovascular diseases can be prevented by addressing behavioral risk factors such as tobacco use, unhealthy diet and obesity, physical inactivity, and harmful use of alcohol. It is important to detect cardiovascular disease as early as possible so that management with counseling and medicines can begin.

The most important behavioral risk factors of heart disease are physical inactivity, glucose, blood pressure, Smoke, and harmful use of alcohol. The effects of behavioral risk factors may show up in individuals as raised blood pressure, raised blood glucose, and overweight. Reduction of salt in the diet, eating more fruit and vegetables, regular physical activity, and avoiding harmful use of alcohol has been shown to reduce the risk of cardiovascular disease.

In this, we split data into 80-20 patterns i.e. 80% of the data is in the training dataset and 20% of the data is in the test dataset then we use the training dataset to fit the models and the test dataset is used for the prediction purpose of the models. We fit the different types of models. The accuracy of Random Forest is 72 %, precision is 72.98 % and recall is 70.17 %. So, we conclude that the Random Forest is the best-fit model.

OBJECTIVES

- Prediction of cardiovascular disease using machine learning algorithms
- Which factors affect the cardiovascular disease
- Whether men or women affect the cardiovascular disease
- To find high-performance predictive models that classified cardio-vascular disease
- To analyze which machine learning model is best fitted.

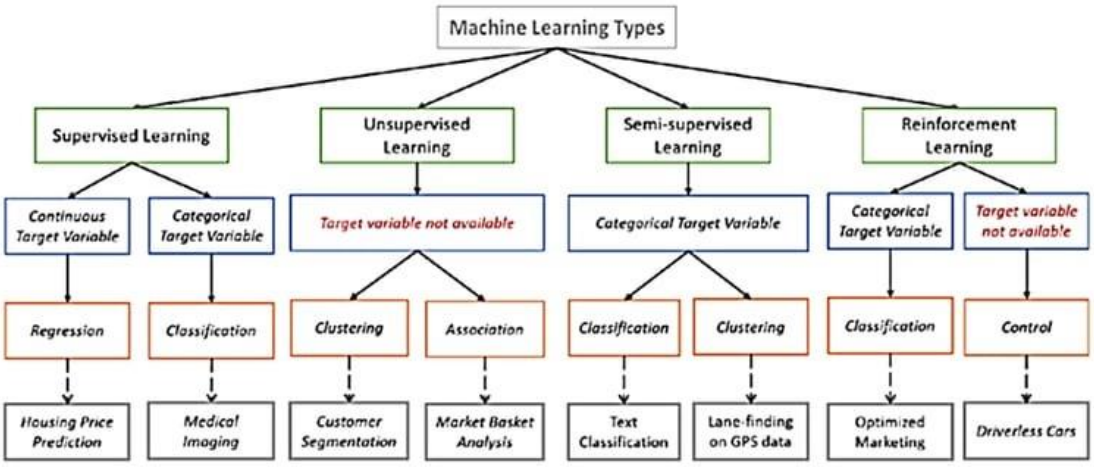
METHODOLOGY

We collected data from Kaggle. The data contains 12 attributes which are age, gender, Glucose, Cholesterol, Ap_hi (systolic blood pressure), ap_lo (Diastolic blood pressure), smoke, alcohol, active, cardio, height, and weight. In our dataset, there are 70,000 cases. the pre-processing of data had done by Python software and R Software. After data pre-processing, we fit various machine learning models to predict the cardiovascular disease

Technical Terms:

MACHINE LEARNING:

What is machine learning? Machine learning is a branch of artificial intelligence (AI) and computer science that focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy.



SUPERVISED LEARNING

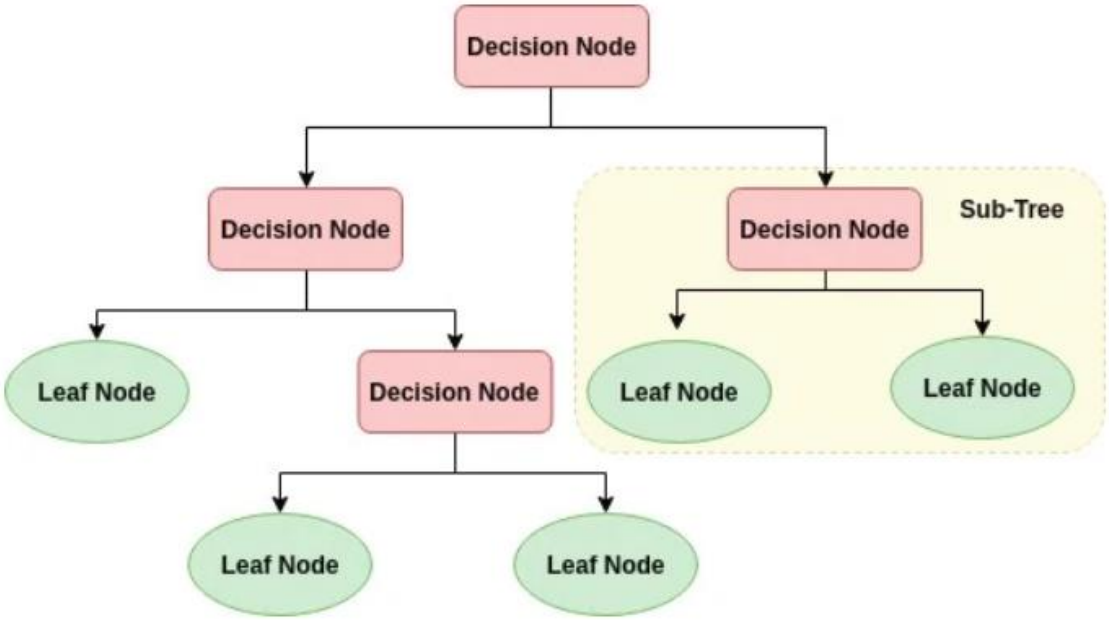
Supervised learning, also known as supervised machine learning, is a subcategory of machine learning and artificial intelligence. It is defined by its use of labeled datasets to train algorithms that classify data or predict outcomes accurately. As input data is fed into the model, it adjusts its weights until the model has been fitted appropriately, which occurs as part of the cross-validation process. Supervised learning helps organizations solve a variety of real-world problems at scale, such as classifying spam in a separate folder from your inbox. Classification uses an algorithm to accurately assign test data into specific categories. It recognizes specific entities within the dataset and attempts to draw some conclusions on how those entities should be labeled or defined. Common classification algorithms are linear classifiers, support vector machines (SVM), decision trees, k-nearest neighbors, and random forests.

CLASSIFICATION ALGORITHMS

1. Logistic Regression:

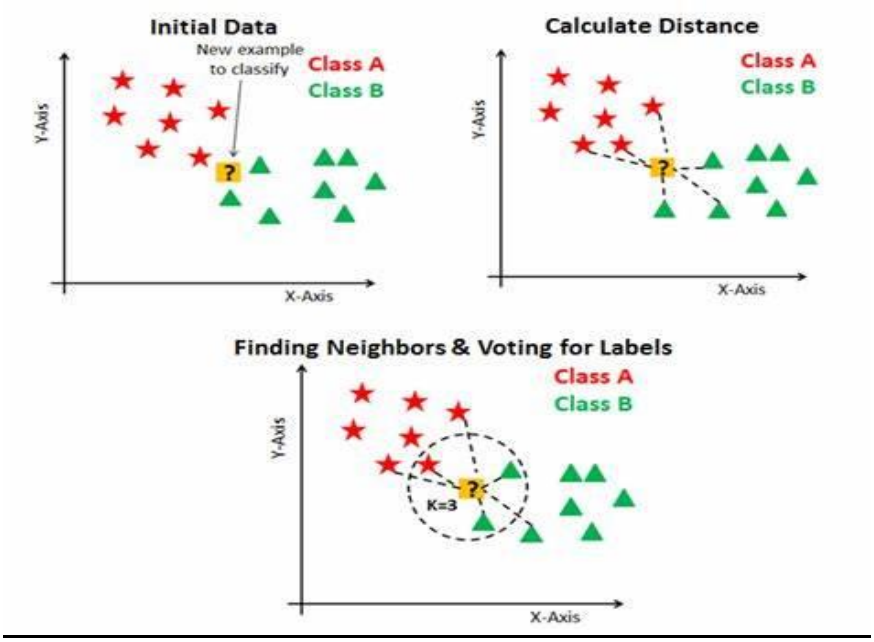
Logistic regression is a classification algorithm. It is used to predict a binary outcome based on a set of independent variables. the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more independent variables.

2. Decision Tree:



a decision tress is a structure that includes a root node, branches, and leaf nodes. A decision tree is a supervised machine-learning algorithm that is used in both classification and regression tasks. It is a powerful tool for modeling and predicting outcomes in a wide range of domains, including business, finance, healthcare, and more. A decision tree is a flowchart like a tree structure. Decision trees can handle high-dimensional data. Decision trees can handle missing values in the data, making them a suitable choice for datasets with missing or incomplete data.

3. **KNN Algorithm**

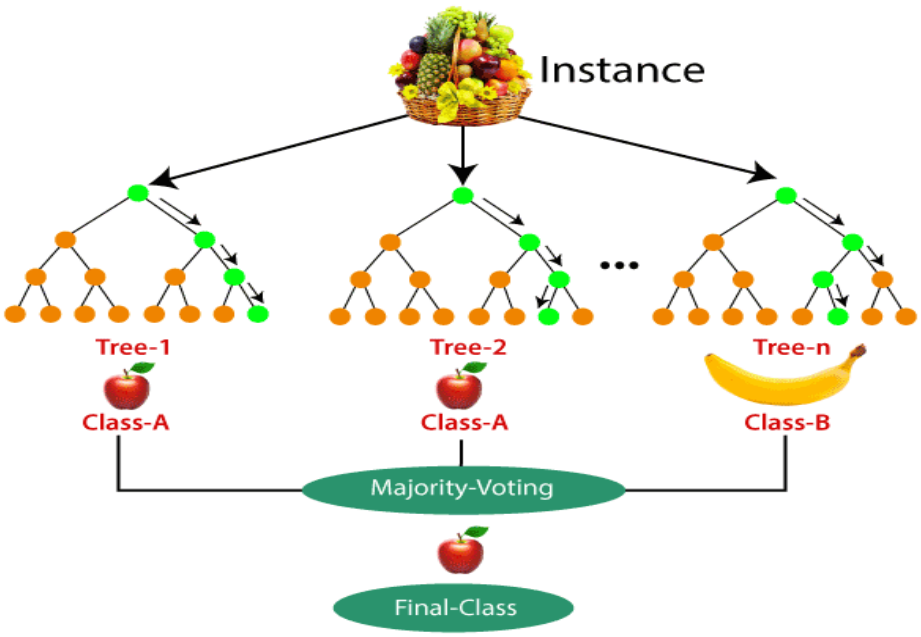


K-nearest neighbors (KNN) algorithm is a type of supervised ML algorithm that can be used for both classifications as well as regression predictive problems. However, it is mainly used for the classification of predictive problems in the industry.

- Lazy learning algorithm − KNN is a lazy learning algorithm because it does not have a specialized training phase and uses all the data for training while classification.
- Non-parametric learning algorithm − KNN is also a non-parametric learning algorithm because it doesn't assume anything about the underlying data.
- KNN tries to predict the correct class for the test data by calculating the distance between the test data and all the training points. Then select the K number of points which is closest to the test data

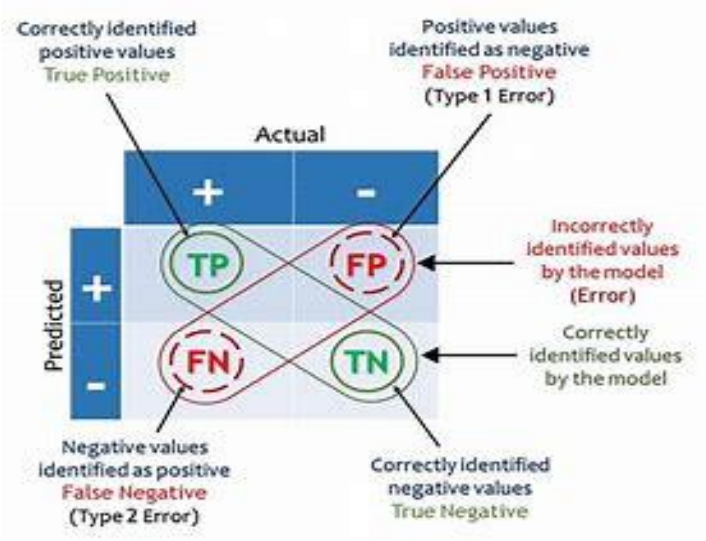
4. Random Forest:

- A Random Forest Algorithm is a supervised machine learning algorithm that is extremely popular and is used for Classification and Regression problems in Machine Learning.
- We know that a forest comprises numerous trees, and the more trees more it will be robust. Similarly, the greater the number of trees in a Random Forest Algorithm, the higher it's accuracy and problem-solving ability.
- Random Forest is a classifier that contains several decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset.
- It is based on the concept of ensemble learning which is a process of combining multiple classifiers to solve a complex problem and improve the performance of the model.
- It builds decision trees on different samples and takes their majority vote for classification and average in case of regression**.**



Need for Confusion Matrix in Machine Learning:

It evaluates the performance of the classification models when they make predictions on test data and tells how well our classification model does not only tell the error made by the classifiers but also the type of errors such as it is either type-I or type-II error. with the help of the confusion matrix, we can calculate the different parameters for the model, such as accuracy, precision, recall, etc.



Accuracy:

It is one of the important parameters to determine the accuracy of classification problems. It defines how often the model predicts the correct output. It can be calculated as the ratio of the number of correct predictions made by the classifier to the number of predictions made by the classifiers.

Accuracy=(TP+TN)/(P+N)

Misclassification rate/Error Rate:

It is also termed an Error rate, and it defines how often the model gives the wrong predictions. The value of the error rate can be calculated as the number of incorrect predictions to all the predictions made by the classifier.

Error rate=1-Accuracy=(FP+FN)/(P+N)

Precision:

It can be defined as the number of correct outputs provided by the model or out of all positive classes that have been predicted correctly by the model, how many of them were true.

Precision=TP/TP+FP

Recall:

It is defined as the out of total positive classes, how our model predicted correctly. The recall must be as high as possible. The recall is a measure of completeness and it is similar to the sensitivity

Recall=TP/TP+FN=TP/P

F-measure:

If two models have low precision and high recall or vice versa, it is difficult to compare these models. So, for this purpose, we can use F-score. This score helps us to evaluate the recall and precision at the same time. The F-score is maximum if the recall is equal to the precision.

DATA DESCRIPTION

Feature	Variable Type	Variable	Value Type
Age	Objective Feature	age	int (days)
Height	Objective Feature	height	int (cm)
Weight	Objective Feature	weight	float (kg)
Gender	Objective Feature	gender	categorical code
Systolic blood pressure	Examination Feature	ap_hi	int
Diastolic blood pressure	Examination Feature	ap_lo	int
Cholesterol	Examination Feature	cholesterol	1: Normal 2: above normal 3: well above normal
Glucose	Examination Feature	glucose	1: Normal 2: above normal 3: well above normal
Smoking	Subjective Feature	smoke	Binary
Alcohol intake	Subjective Feature	alcohol	Binary 0:No 1:Yes
Physical activity	Subjective Feature	active	Binary 0: No 1: Yes
Presence or absence of cardiovascular disease	Target Variable	cardio	Binary 0: No 1: Yes

DESCRIPTIVE STATISTICS

Descriptive statistics involves summarizing and organizing the data so it can be easily understood. Descriptive statistics is the term given to the analysis of data that helps describe, show, or summarize data in a meaningful way

Data describe:

	Count	Mean	SD	Min	Q_1	Median	Q_3	Max
age	70000	19468.9	2467.25	10798	17664	19703	21327	23713
gender	70000	1.35	0.4768	1	1	1	2	2
height	70000	164.359	8.2101	55	159	165	170	250
weight	70000	74.206	14.3958	10	65	72	82	200
ap_hi	70000	128.817	154.011	150	120	120	140	360.20
ap_lo	70000	96.63	188.473	70	80	80	90	150.00
cholesterol	70000	1.367	0.6803	1	1	1	2	3
glucose	70000	1.226	0.5723	1	1	1	1	3
smoke	70000	0.088	0.2835	0	0	0	0	1
alcohol	70000	0.054	0.2256	0	0	0	0	1
active	70000	0.804	0.3972	0	1	1	1	1
cardio	70000	0.5	0.5	0	0	0	1	1

Categorical values are not meant to have calculations performed on them so, we can ignore those. What we want to focus on is the unique count and frequency of the categorical features

Conclusion:

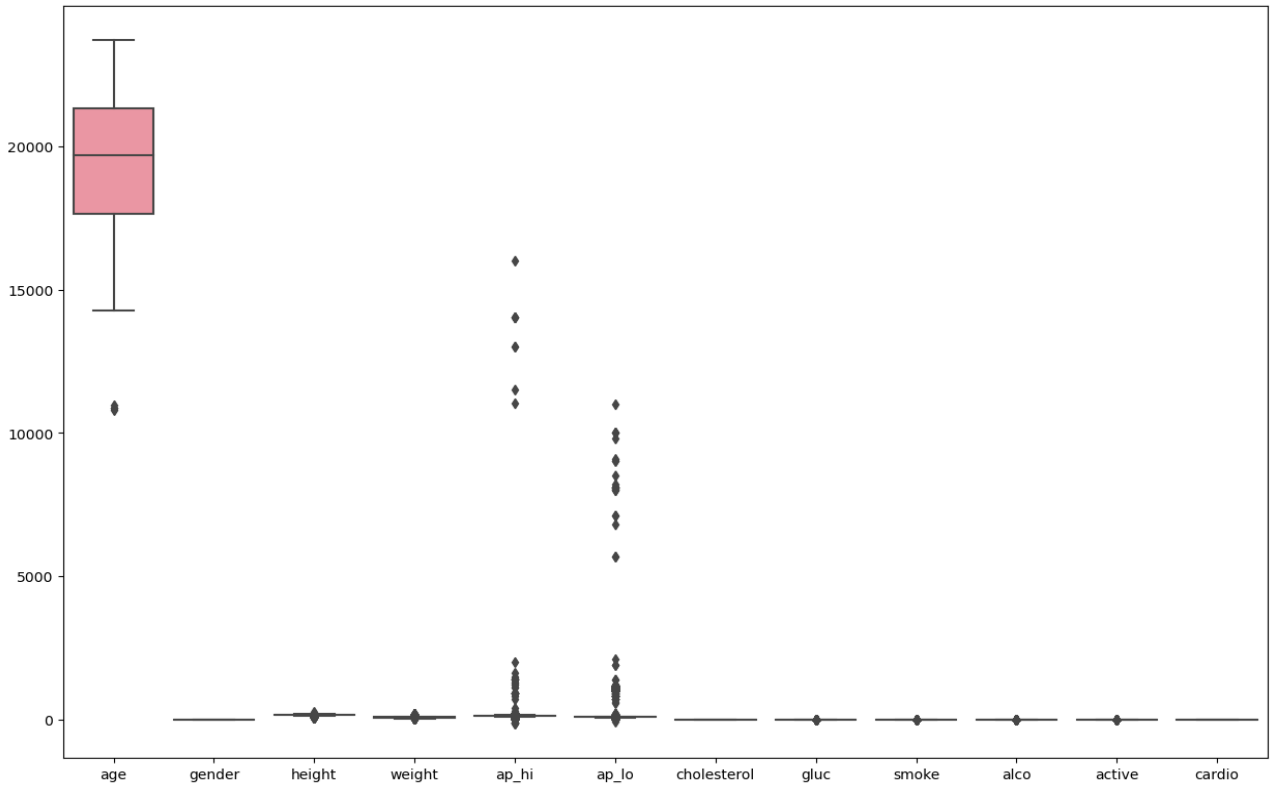
- From the table, we observe that the maximum age of the patient is 23713 (in days) and the minimum age is 10798 (in days)
- We observe that the maximum ap_ hi (Systolic blood pressure) is 360.20 and the minimum ap_hi is 150
- From the table we can say that the maximum ap_ lo is (Diastolic blood pressure) is 150.00 and the minimum ap_lo is 70

Data Preprocessing:

Data cleaning is the process of preparing data for analysis by removing or modifying data that is incorrect, incomplete, irrelevant, duplicated, or improperly formatted. Data cleaning is one of the important parts of machine learning. It plays a significant part in building a model. It surely isn't the fanciest part of machine learning. The goal of data cleaning is to ensure that the data is accurate, consistent, and free of errors, as incorrect or inconsistent data can negatively impact the performance of the ML model.

Outlier:

Outliers are nothing but data points that differ significantly from other observations. They are the points that lie outside the overall distribution of the dataset. Outliers, if not treated, can cause serious problems in statistical analyses.

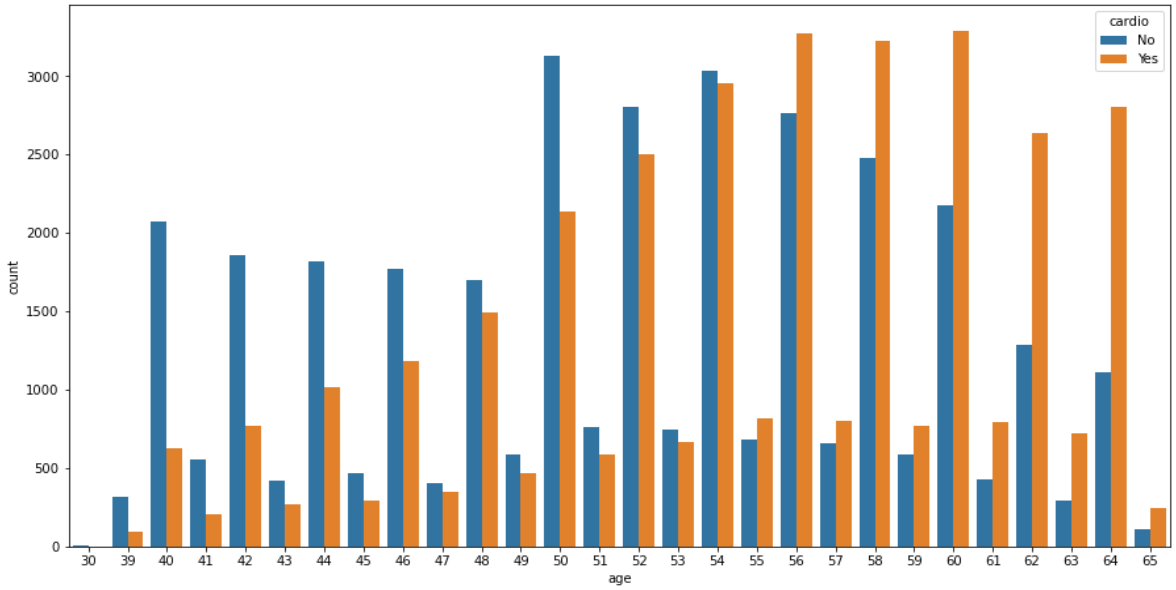


- From the graph, we can conclude that the attributes age, height, weight, ap-hi, and ap-lo have the outliers.
- We treated the outliers present in the data by using appropriate ways.

EXPLORATORY DATA ANALYSIS

Exploratory Data Analysis (EDA) is a process of describing the data using statistical and visualization techniques to bring important aspects of that data into focus for further analysis.in EDA, we find the patterns and relationships between the variables.

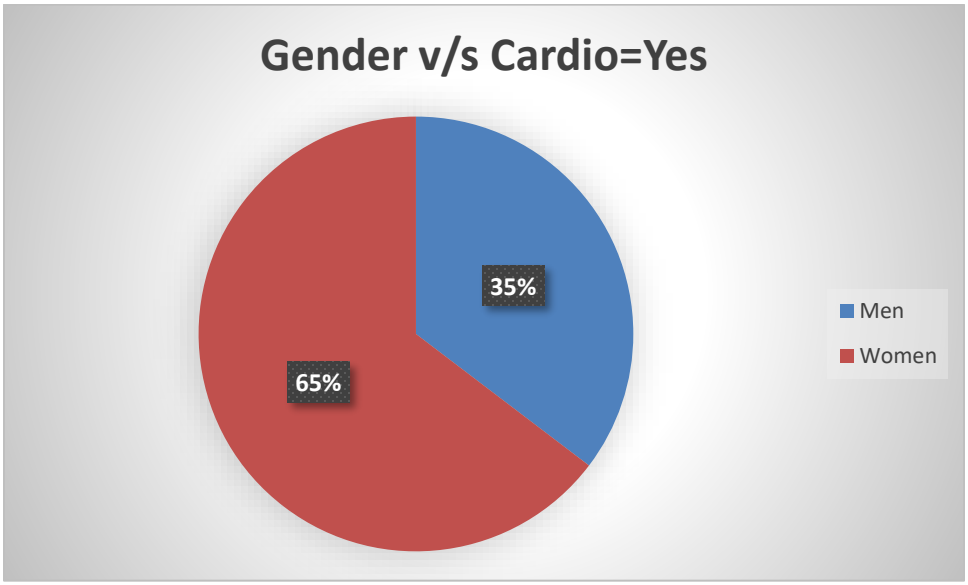
Distribution of Age



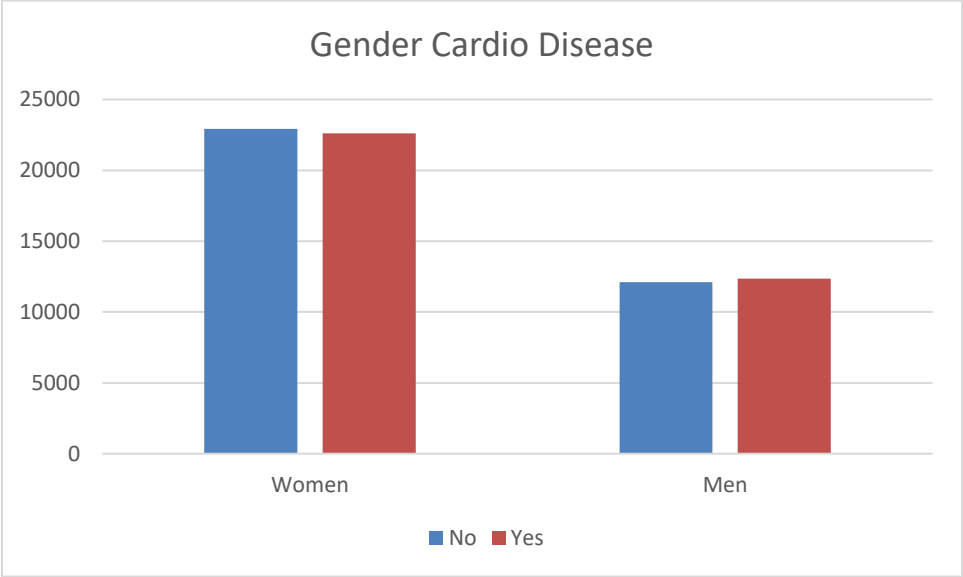
Here, the blue shaded part shows people don't have the cardio disease, orange shade shows people having Cardio disease at a particular age

- we can conclude that the chance of getting cardiovascular disease becomes larger as the person gets older i.e. above 55
- From the age of 40 to 54, the chance of getting cardiovascular disease is smaller

Gender-wise distribution of cardiovascular disease

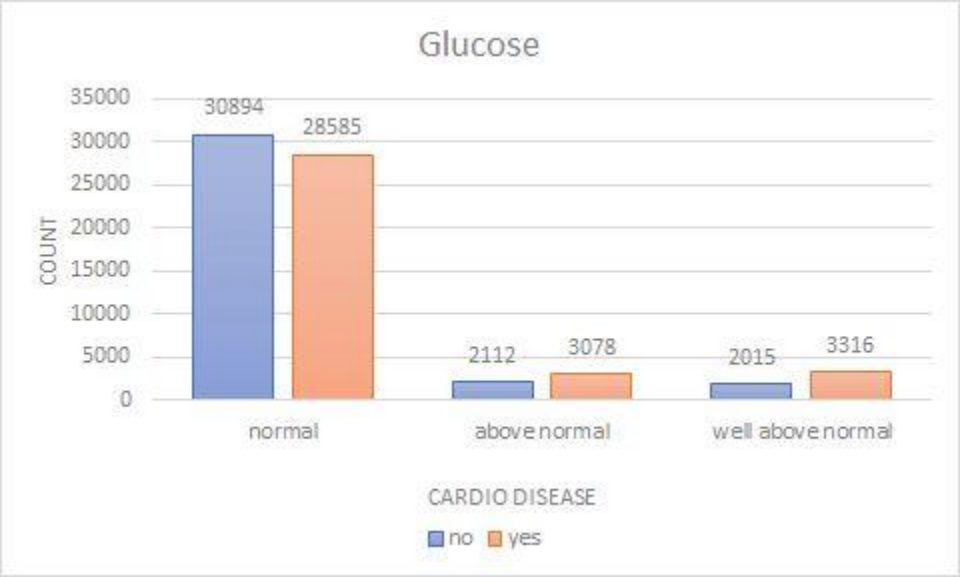


- From the above graph we can say that the out of total strength of 65% women and 35 % men having cardiovascular disease.



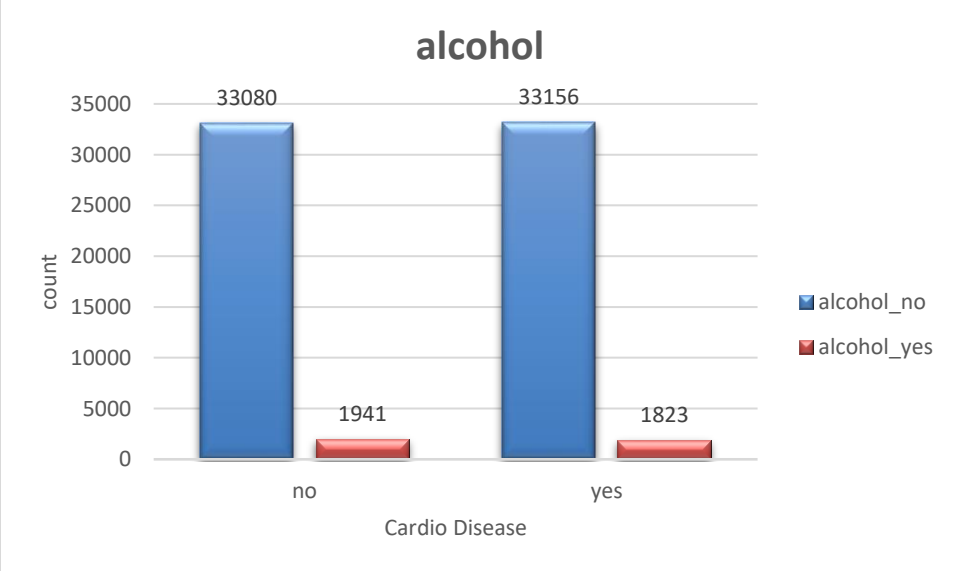
- We can say that men and women both have a 50% of chance getting the cardio disease.
- While women make up slightly more of the respondents there are identical chances of having a cardio disease or not having the cardio disease.

Glucose v/s cardiovascular disease



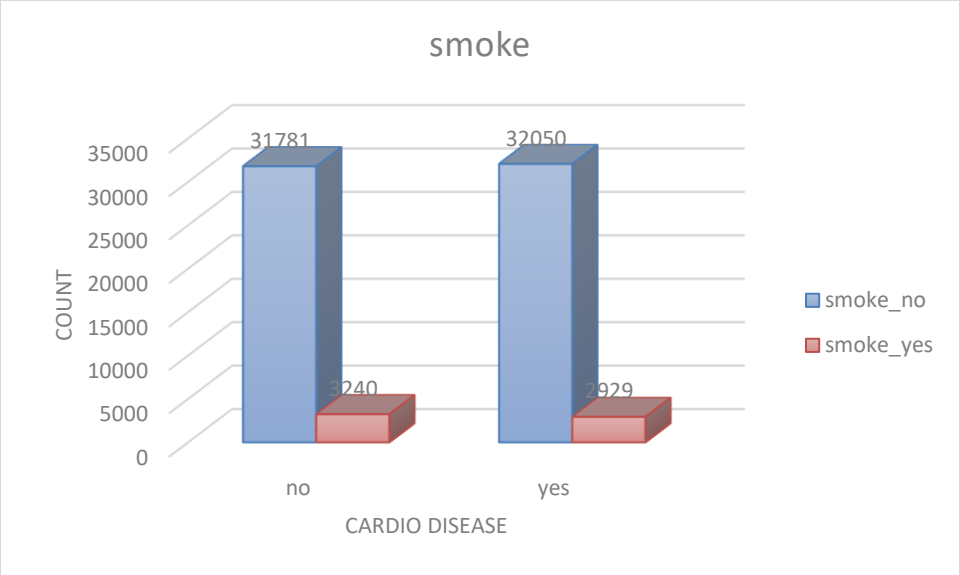
Interpretation: people who have normal glucose, they are getting more chances of cardiovascular disease

Alcohol-wise cardiovascular disease



From the above graph, we can see that there is no effect of alcohol on cardiovascular disease

The plot of smoke v/s cardiovascular disease



From the above graph, we can see that there is no effect of smoke on cardiovascular disease.

STATISTICAL ANALYSIS

Hypothesis testing: To check the association between attributes:

1. Test for the association between cardiovascular disease and gender

Ho: There is no association between gender and cardiovascular disease

V/S

H1: There is an association between gender and cardiovascular

			Cardio	
		No	Yes	Total
	Women	22914	22616	45530
Gender	Men	12107	12363	24470
	Total	35021	34979	70000

X^2_{Cal}=**4.6034** X^2_{Tab}=3.84 p-value=0.032

Decision criteria:

X^2_{Cal}>X^2_{Tab}, therefore we Reject Ho which means we Accept H1

Here, X^2_{Cal}=4.6034 > 3.84

Interpretation: There is an association between gender and cardiovascular disease.

For All Attributes Chi-Square Test Table

	Chi-square calculated	tabulated	p-value	Decision	Conclusion
Gender	**4.6034**	**3.84**	**0.032**	**Reject H_0**	**association**
Alcohol	**3.7612**	**3.84**	**0.0546**	**Accept H_0**	**No association**
Smoke	**16.7869**	**3.84**	**0.00004**	**Reject H_0**	**association**
Active	**88.9807**	**3.84**	**2.2e-16**	**Reject H_0**	**association**
cholesterol	**3423.4**	**3.84**	**2.2e-16**	**Reject H_0**	**association**
Glucose	**586.91**	**3.84**	**2.2e-16**	**Reject H_0**	**association**

Interpretation :

From the above table, we observe that the cardiovascular disease is associated with the attributes gender, smoke, cholesterol, physical activity, glucose.

FITTING OF CLASSIFICATION MODELS

Logistic Regression:

- **Logistic model:**

Here we split our data into 80-20 pattern i.e. 80% of the data is in train data and 20% of the data are in the test dataset

Coefficients:

(Intercept)	age	gender	height	weight	ap_hi
-8.4963	0.05411	0.01558	-0.00575	0.01534	0.03953
ap_lo	cholesterol	glucose	smoke	alcohol	active
0.0003	0.5233	-0.1184	-0.1315	-0.1690	-0.2098

THE GIVEN LOGISTIC MODEL IS :

$$\Pi(x)= \frac{\text{EXP}(8.4963+0.05411X_1-0.01558X_2+0.01534X_3+0.0395X_4+0.0003X_5 +0.5233X_6-0.1184X_7-0.1315X_8-0.1691X_9-0.2098X_{10})}{1+\text{EXP}(8.4963+0.05411X_1-0.01558X_2+0.01534X_3+0.0395X_4+0.0003X_5 +0.5233X_6-0.1184X_7-0.1315X_8-0.1691X_9-0.2098X_{10})}$$

i.e.

$$h(x)= \mathbf{EXP}(\mathbf{8.4963+0.05411X_1-0.01558X_2+0.01534X_3+0.0395X_4+0.0003X_5 +0.5233X_6-0.1184X_7-0.1315X_8-0.1691X_9-0.2098X_{10}})$$

Summary(model) :

Deviance Residuals:

Min	1Q	Median	3Q	Max
-8.4904	-0.9638	-0.0984	0.9897	4.6658

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-8.50E+00	2.14E-01	-39.67	< 2e-16	***
Age	5.41E-02	1.30E-03	41.765	< 2e-16	***
Gender	1.56E-02	2.11E-02	0.74	0.459	
Height	-5.75E-03	1.23E-03	-4.674	2.95E-06	***
Weight	1.54E-02	6.59E-04	23.277	< 2e-16	***
ap_hi	3.95E-02	6.05E-04	65.319	< 2e-16	***
ap_lo	3.00E-04	6.74E-05	4.452	8.49E-06	***
cholesterol	5.23E-01	1.50E-02	34.923	< 2e-16	***
Glucose	-1.19E-01	1.70E-02	-6.968	3.21E-12	***
Smoke	-1.32E-01	3.32E-02	-3.966	7.31E-05	***
Alcohol	-1.69E-01	4.02E-02	-4.203	2.63E-05	***
Active	-2.10E-01	2.11E-02	-9.973	< 2e-16	***

Significant codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 97041 on 69999 degrees of freedom

Residual deviance: 80931 on 69988 degrees of freedom

AIC: 80955

Number of Fisher Scoring iterations: 25

Age, Height, Weight, ap_hi, ap_lo, cholesterol, glucose, smoke alcohol, and active are significant variables.

Now we want to test the significance of regression coefficient β_j where, j=1, 2, …11

i.e. we want to test:

$H_0: \beta_j = 0$ V/S $H_0: \beta_j \neq 0$

Test of significance of regression:

∴ G = Null Deviance - Residual Deviance

From the above output

Null Deviance = Deviance for model excluding regressor X = 97041

Residual Deviance = Deviance for the model including regressor X = 80931

∴ G = 97041-80931 = 16110

$\chi^2_{11,0.05}$ =19.675

Here G exceeds the value of $\chi^2_{2,0.05}$

∴ We make reject H_0 at a 5% level of significance i.e., at least one regressor is significant.

The odds ratio estimate is given by

Ψ_1 hat = $e^{0.0541}$ = 0.0541 <1 Ψ_5 hat = $e^{0.0395}$ = 1.04029> 1 Ψ_9 hat = $e^{-0.1315}$ = 0.87677<1

Ψ_2 hat = $e^{0.0155}$ = 1.0153> 1 Ψ_6 hat = $e^{0.0003}$ = 1.0003> 1 Ψ_{10} hat = $e^{-0.1691}$ = 0.844424<1

Ψ_3 hat = $e^{-0.0057}$ = 0.994< 1 Ψ_7 hat = $e^{0.5233}$ = 1.68758> 1 Ψ_{11} hat = $e^{-0.2098}$ = 0.8107<1

Ψ_4 hat = $e^{0.0153}$ = 1.0154> 1 Ψ_8 hat = $e^{-0.118}$ = 0.88816<1

Interpretation: Here Ψ_i hat>1 i.e. for every unit increase in Xi the chances of getting cardio disease increases, when other effects of Xi are kept constant, i =1, 2, …..11

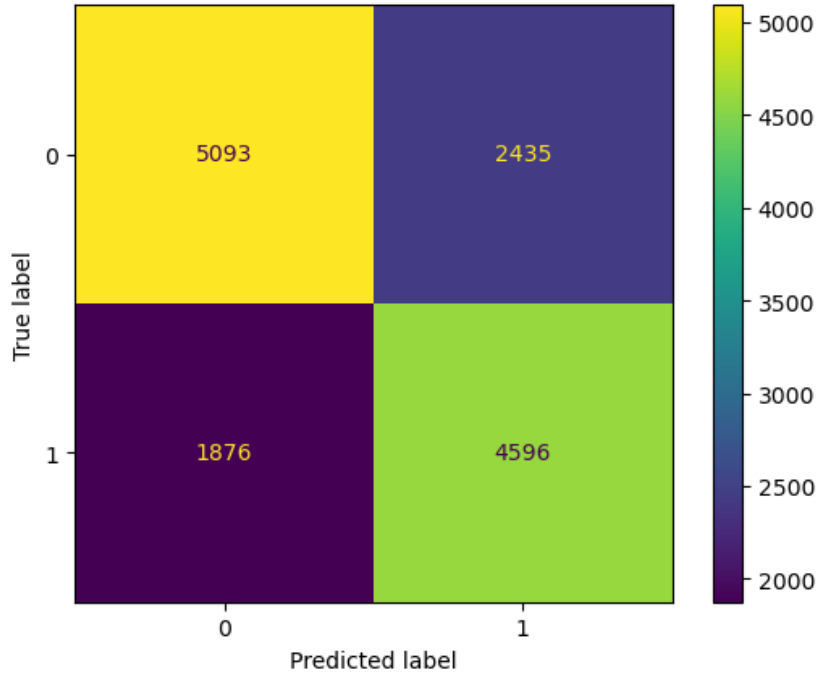
And Ψ_i hat< i.e. For every unit decrease in Xi the chances of getting cardio disease decrease, when other effects of Xi are kept constant.

Where i = 1, 2, ….., 11

Classification Report:

	precision	Recall	f1-score	support
0	0.73	0.68	0.70	7528
1	0.65	0.71	0.68	6472
accuracy			0.69	14000
macro_avg	0.69	0.69	0.69	14000
weighted_avg	0.70	0.69	0.69	14000

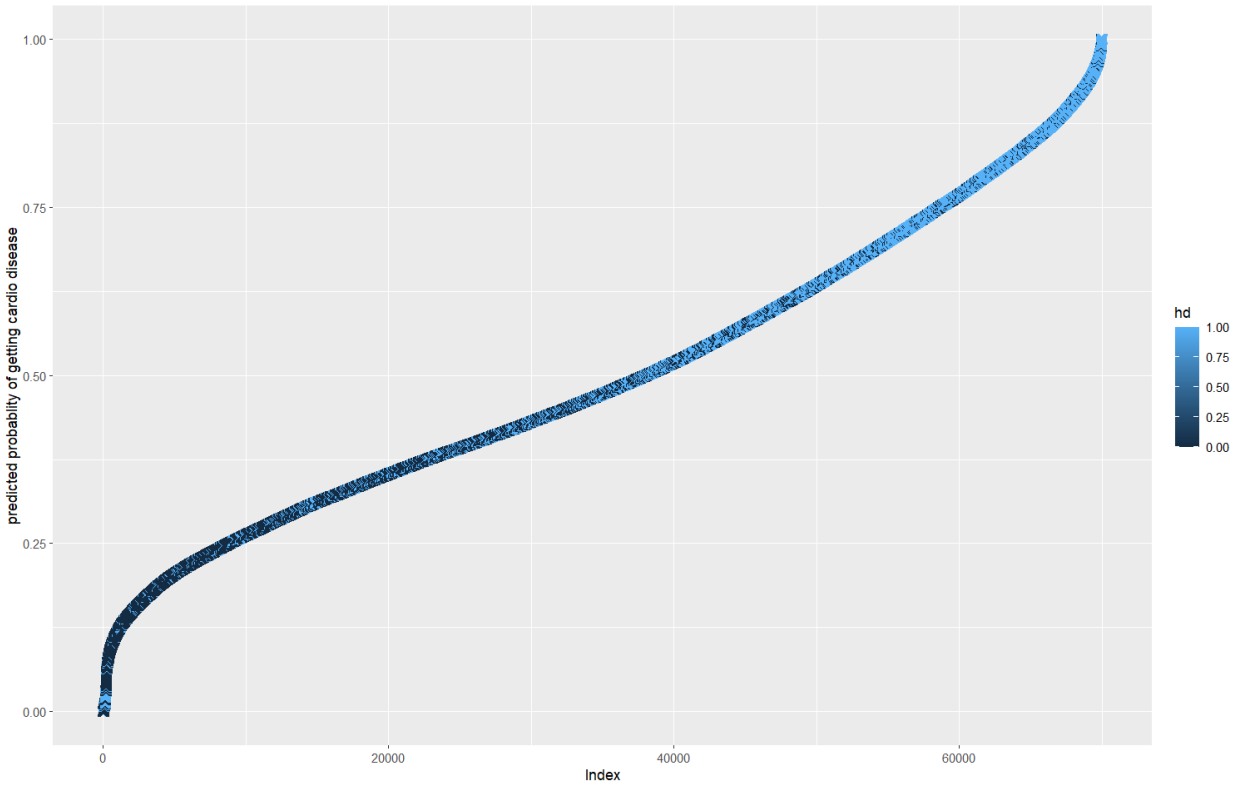
Logistic Regression Confusion Matrix



Conclusion: Logistic regression gives 69% accuracy.

Accuracy	69%
Precision	73%
Sensitivity	67%
Specificity	71%
F1 Score	69%

The logistic regression probability curve



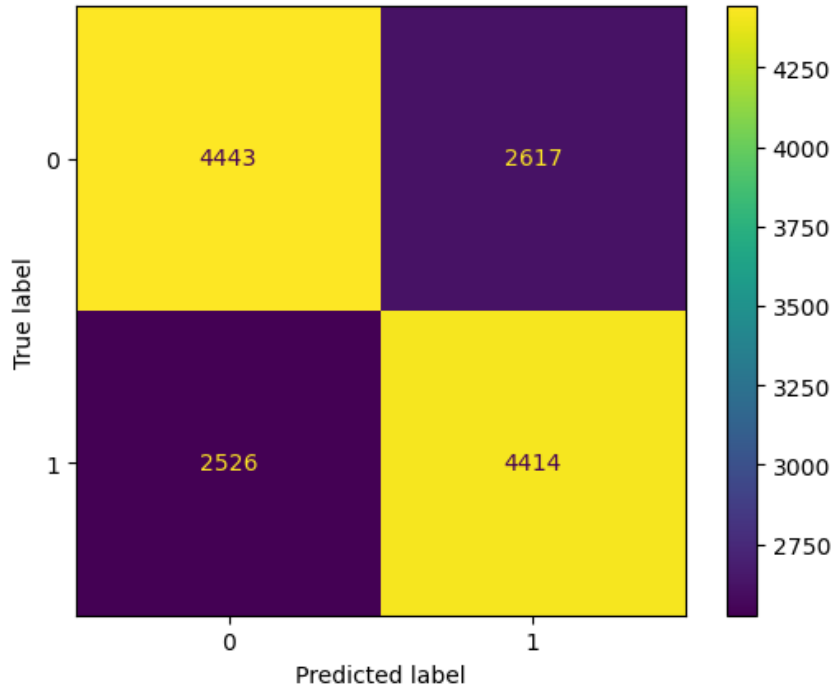
Interpretation: the given graph contain the probability of cardiovascular disease along with an actual cardiovascular disease status. From the above graph, we can see that the logistic regression has done a pretty good job.

2. Decision Tree Algorithm:

Classification Report of Decision Tree:

	precision	Recall	f1-score	support
0	0.64	0.63	0.63	7060
1	0.63	0.64	0.63	6940
accuracy			0.63	14000
macro_avg	0.63	0.63	0.63	14000
weighted_avg	0.63	0.63	0.63	14000

Confusion Matrix:

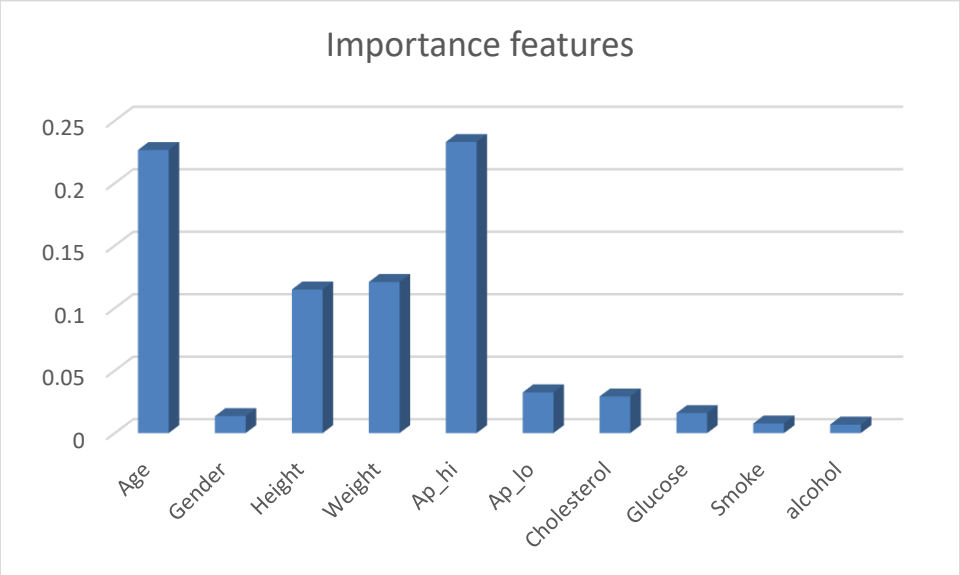


Conclusion: The decision Tree gives 63% accuracy.

Accuracy	63%
Precision	63%
Sensitivity	62%
Specificity	63%
F1score	62%

Important Features:

Attributes	Importance	percentage
Age	0.2265	22%
Gender	0.0136	1.30%
Height	0.1149	11%
Weight	0.1209	12%
Ap_hi	0.2329	23%
Ap_lo	0.0326	3%
Cholesterol	0.0293	3%
Glucose	0.016	2%
Smoke	0.0077	0%
alcohol	0.0068	0%



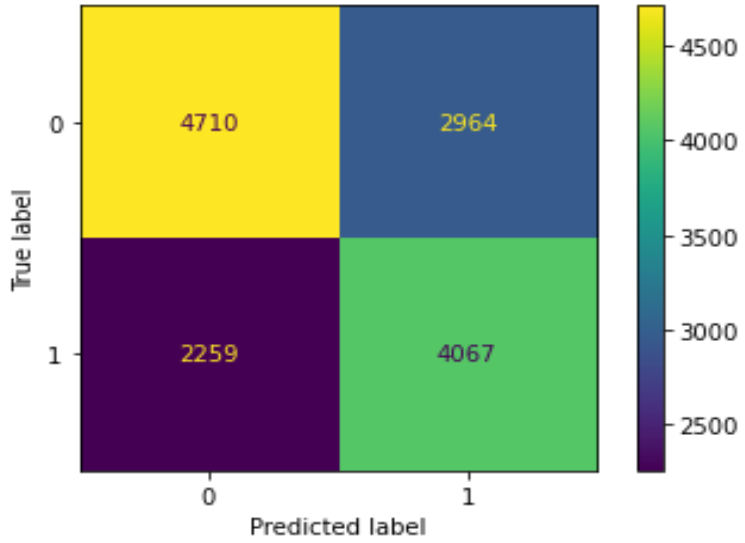
Conclusion: from the table, we get Age (22.65%), Height (11.49%), Weight (12.09%), Systolic blood pressure (23.29%) are significant features associated with cardiovascular disease.

3. K-Nearest Neighbor:

- **Classification report:**

	Precision	Recall	F1-Score	Support
0	0.68	0.61	0.64	7674
1	0.58	0.64	0.61	6326
accuracy			0.63	14000
macro avg	0.63	0.63	0.63	14000
weighted avg	0.63	0.63	0.63	14000

- **Confusion Matrix:**

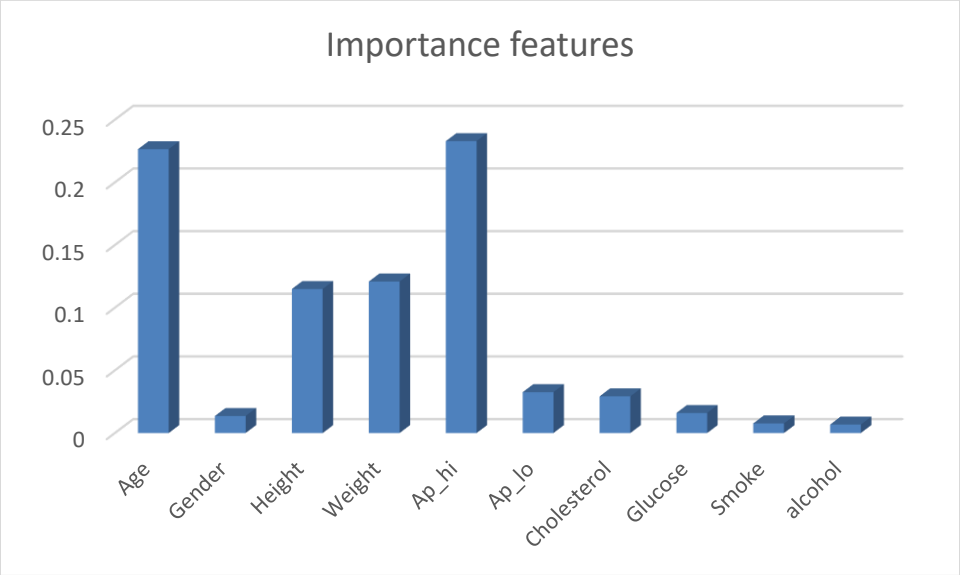


Conclusion: KNN gives 63% Accuracy

Accuracy	63%
Precision	67%
Sensitivity	61.31%
Specificity	64.29%
F1score	63.85%

Important Features:

Attributes	Importance	percentage
Age	0.2265	22%
Gender	0.0136	1.30%
Height	0.1149	11%
Weight	0.1209	12%
Ap_hi	0.2329	23%
Ap_lo	0.0326	3%
Cholesterol	0.0293	3%
Glucose	0.016	2%
Smoke	0.0077	0%
alcohol	0.0068	0%



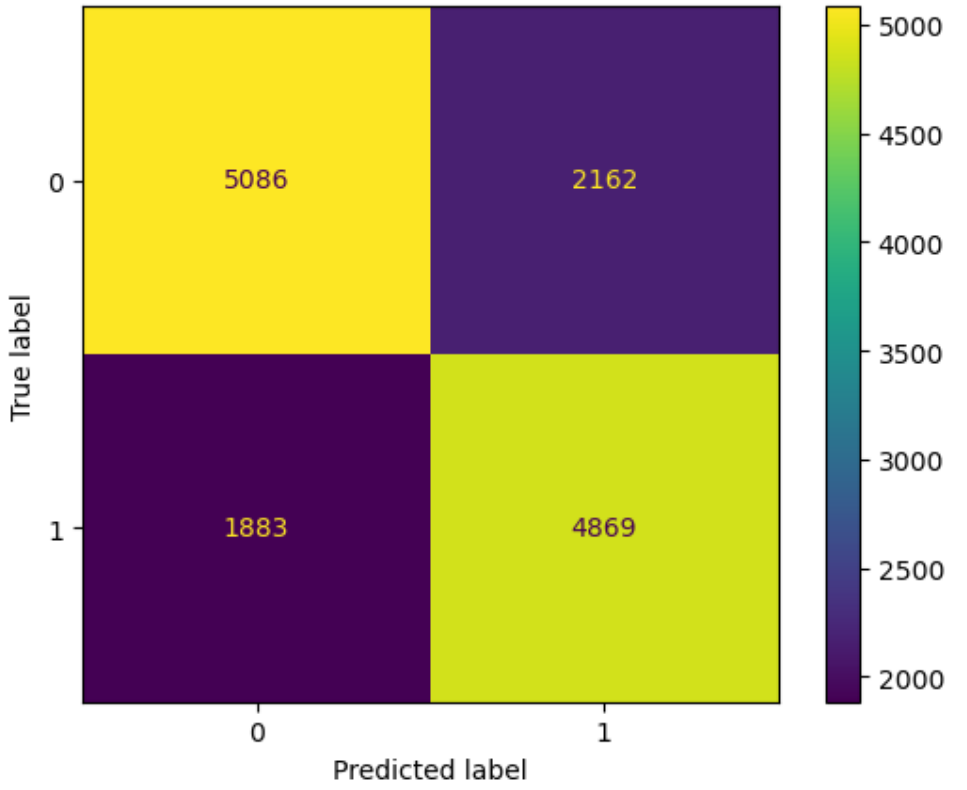
Conclusion: from the table, we get Age (22.65%), Height (11.49%), Weight (12.09%), Systolic b lood pressure (23.29%) are significant features associated with cardiovascular disease.

4. Random Forest:

- **Classification report:**

	precision	Recall	f1-score	support
0	0.73	0.70	0.72	7248
1	0.69	0.72	0.71	6752
accuracy			0.72	14000
macro_avg	0.71	0.71	0.71	14000
weighted_avg	0.71	0.71	0.71	14000

- **Confusion Matrix:**

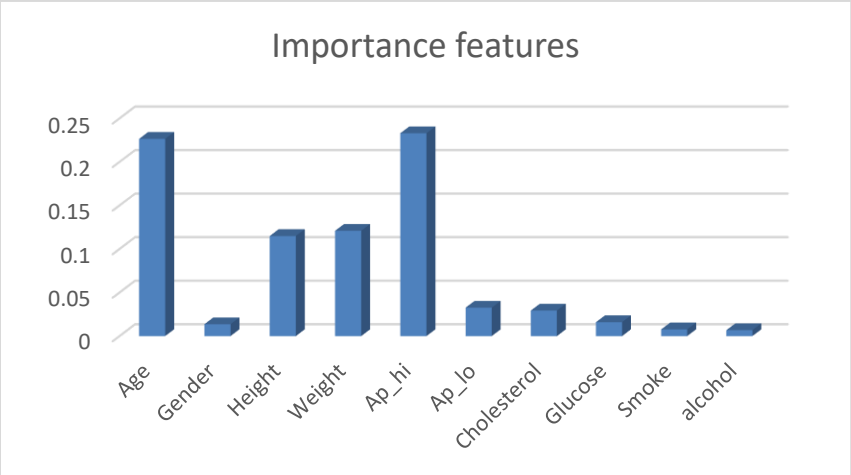


Conclusion: The Random Forest gives an Accuracy of 71.214%

Accuracy	72%
Precision	72.98%
Sensitivity	70.17%
Specificity	72.18%
F1score	71.54%

Important Features:

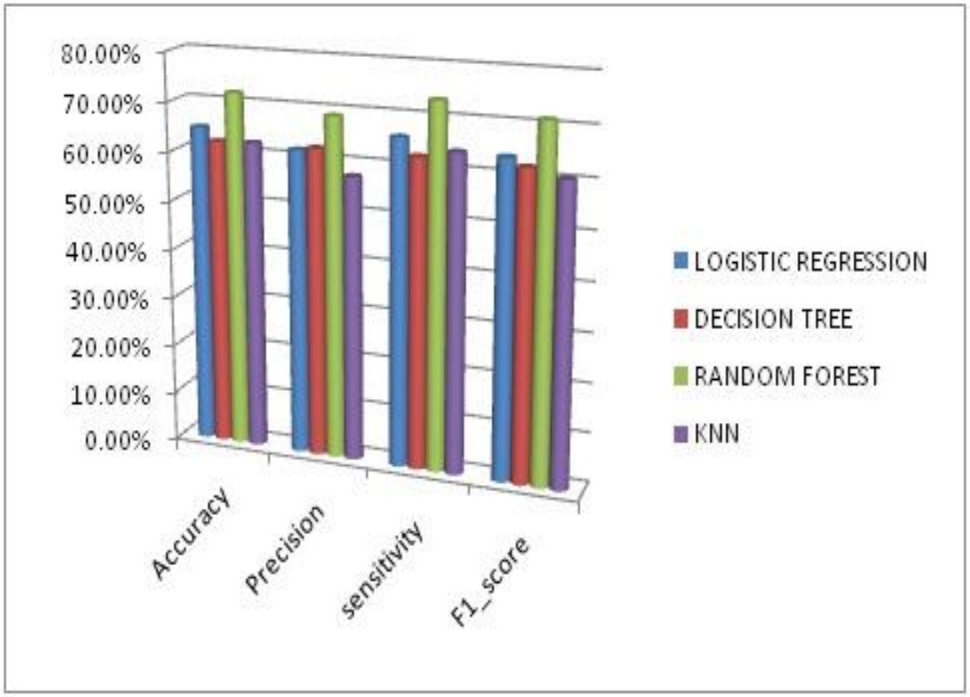
Attributes	Importance	percentage
Age	0.2265	22%
Gender	0.0136	1.30%
Height	0.1149	11%
Weight	0.1209	12%
Ap_hi	0.2329	23%
Ap_lo	0.0326	3%
Cholesterol	0.0293	3%
Glucose	0.016	2%
Smoke	0.0077	0%
alcohol	0.0068	0%



Conclusion: from the table, we get Age (22.65%), Height (11.49%), Weight (12.09%), Systolic blood pressure (23.29%) are significant features associated with cardiovascular disease.

Overall Comparison:

	Model Name	Accuracy	Precision	sensitivity	F1_score
1	LOGISTIC REGRESSION	69%	73%	67%	69%
2	DECISION TREE	63%	63%	62%	62%
3	RANDOM FOREST	72%	72.98%	70.17%	71.54%
4	KNN	63%	67%	61.31%	63.85%



Interpretation: From the above table we can observe that the accuracy of random forest is 72%, precision is 72.98%, sensitivity is 70.17% and F1 score is 71.54% which are higher than other models. Therefore, we can say that the Random Forest is the best model other three models for the prediction of cardiovascular disease.

CONCLUSION

- From the statistical models, we observe that the accuracy of random forest is 72%, precision is 72.98%, sensitivity is 70.17% and F1 score is 71.54%. The accuracy of KNN is 63 %, precision is 67 %, sensitivity is 61.31 % and F1 score is 63.85 %. The accuracy of decision tree is 63 %, precision is 63 %, sensitivity is 62 % and F1 score is 62 %. The accuracy of logistic regression is 69 %, precision is 73 %, sensitivity is 67 % and F1 score is 69 %.
- The chances of getting cardiovascular disease for men and women are equal.
- From the chi-square test, we observe that the cardiovascular disease is associated with the attributes gender, smoke, cholesterol, physical activity, glucose.
- From the statistical models, we observe that the accuracy of random forest is 72%, precision is 72.98%, sensitivity is 70.17% and F1 score is 71.54% which are higher than other models. Therefore, we can say that the Random Forest is the best model than the other three models for the prediction of cardiovascular disease.

SCOPE

- This project will be useful in identifying the possible patients who may suffer from cardiovascular disease. This may help take preventive measures and hence try to avoid the possibility of cardiovascular disease.

LIMITATIONS

- If we get more responses for men, it will be more helpful for us
- We fit the model by using available attributes but if we get more attributes then the performance of the model will be better
- In our data, we have the attribute alcohol which is having categories only Yes and No but if we get the proper specification for attribute alcohol then it will be more effective

REFERENCES

1) Cardiovascular diseases (CVDs) prediction using machine learning algorithms http://www.who.int/newsroom/factsheets/detail/cardiovascular-diseases-(cvds accessed on 30/9/2018. [Google Scholar]
2) Patel B, Sengupta P. Machine learning for predicting cardiac events: what does the future hold? Expert Rev Cardiovasc Ther. 2020;18(2):77–84. [PMC free article] [PubMed] [Google Scholar]

3) Risk prediction of cardiovascular disease using machine learning classifiers: 2022; 17(1): 1100–1113. Published online 2022 Jun 17. Doi: 10.1515/med-2022-0508
(https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9206502/)
4) Breiman, L., Friedman, J.H. Olshen, R.A. and Stone, C.J. (1984). Classification and Regression Trees. (Wadsworth and Brooks/Cole).
5) Kaggle link: https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset