STUDY ON HOW DEMOGRAPHIC, BEHAVIOURIAL AND MEDICAL FACTORS AFFECT THE RISK OF CORONARY HEART DISEASE

ABSTRACT

The dataset is publicly available on the Kaggle website, and it is from an ongoing cardiovascular study on residents of the town of Framingham, Massachusetts. The classification goal is to predict whether the patient has 10-year risk of future coronary heart disease (CHD). The dataset provides the patients' information. It includes over 4,000 records and 15 attributes. This research intends to find the best suitable model in terms of prediction accuracy as well as pinpoint the most relevant/risk factors of coronary heart diseases.

The dataset was made free of missing values and then a 70% - 30% split for training and testing was used to fit a number of models. Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Logistic Regression, Decision tree classifier, Boosting, Random Forest, Bagging and Support Vector Machines with linear and Radial kernel was used for model building. Cross validation was used whenever deemed necessary for tuning the best parameters of the models.

As for model selection Training Accuracy, Testing Accuracy and Concordance Index (AUC-ROC) was used. The training and testing accuracy provide an overview of model fitting while the Concordance Index provides an idea of the predictive power of the model. All three parameters are best if higher.

LDA model came out to be the better performing model than all the other models. Boosting and SVM with linear kernel also performed well. Age and Systolic Blood Pressure turned out to be two important differentiators in developing the risk of coronary heart diseases in ten years.

Developing such models has a wide and extended application in the medical industry. It can help predict the risk of coronary heart diseases and help understand the root cause in terms of medical history, behavioural and demographic factors.

BACKGROUND AND INTRODUCTION

Congenital heart disease (CHD) is the most common congenital malformation diagnosed in new-borns,(1) with birth prevalence reported to be 10‰ of live births worldwide(2), (3) and 8.9‰ of live births in China.(4) Early diagnosis and advances in cardiac surgery and interventional cardiology have significantly increased survival of patients with CHD over the past several decades. As reported, the number of people with CHD who reach adulthood has also risen,(5) which is estimated to be >1 million in the United States(2),(5) and 1.2 million in Europe.(6)

World Health Organization has estimated 12 million deaths occur worldwide; every year due to heart diseases. Half the deaths in the United States and other developed countries are due to cardio vascular diseases. The early prognosis of cardiovascular diseases can aid in making decisions on lifestyle changes in high-risk patients and in turn reduce the complications. (7)

Individuals with congenital heart disease are at a higher risk of cardiovascular disease, and this suggests that the risk assessment for cardiovascular disease based on conventional cardiovascular risk factors as well as interventions for reducing cardiovascular disease risk should be considered in these patients. (8)

PURPOSE OF THE STUDY

The main Purpose and Objectives of the study are;

- To evaluate how different factors and variables affect the risk of CHD in Ten years.
- Understand the profile of the population who has developed the risk of CHD.
- Perform model building on various machine learning techniques and evaluate their performance.
- Choose a best fit model in terms of better accuracy and predictive power.
- Make inferences as to what factors are significant and which category is more susceptible than the other to develop the risk of CHD.
- Determine by how much one category is more susceptible than other.
- Discuss the applications and use cases of such models today and in near future.
- Evaluate the predictive power of the model and suggest any possible improvements in further studies.

DATASET AND VARIABLES

The dataset is publicly available on the Kaggle website, and it is from an ongoing cardiovascular study on residents of the town of Framingham, Massachusetts.

Variables;
Each attribute is a potential risk factor. There are both demographic, behavioural and medical risk factors.

Demographic:
• Sex: male or female (Nominal)
• Age: Age of the patient;(Continuous - Although the recorded ages have been truncated to whole numbers, the concept of age is continuous)

Behavioural
• Current Smoker: whether or not the patient is a current smoker (Nominal)
• Cigs Per Day: the number of cigarettes that the person smoked on average in one day.(can be considered continuous as one can have any number of cigarettes, even half a cigarette.)
Medical(history)
• BP Meds: whether or not the patient was on blood pressure medication (Nominal)
• Prevalent Stroke: whether or not the patient had previously had a stroke (Nominal)
• Prevalent Hyp: whether or not the patient was hypertensive (Nominal)
• Diabetes: whether or not the patient had diabetes (Nominal)

Medical(current)
• Tot Chol: total cholesterol level (Continuous)
• Sys BP: systolic blood pressure (Continuous)
• Día BP: diastolic blood pressure (Continuous)
• BMI: Body Mass Index (Continuous)
• Heart Rate: heart rate (Continuous - In medical research, variables such as heart rate though in fact discrete, yet are considered continuous because of large number of possible values.)
• Glucose: glucose level (Continuous)

Predict variable (desired target)
• 10-year risk of coronary heart disease CHD (binary: “1”, means “Yes”, “0” means “No”)

RESULTS, OUTCOMES AND INFERENCES

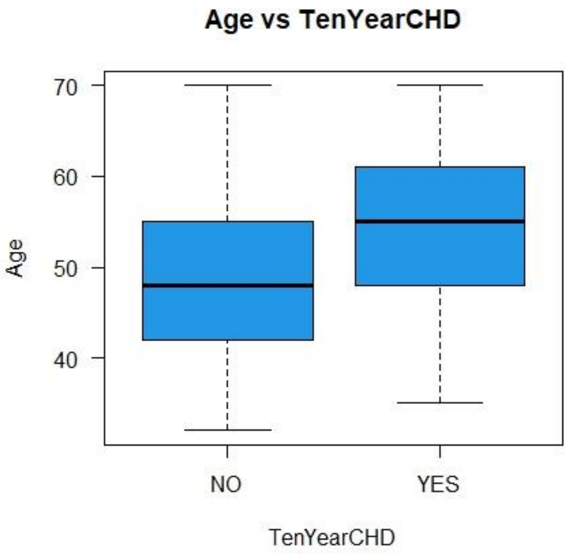
DESCRIPTIVE STATISTICS

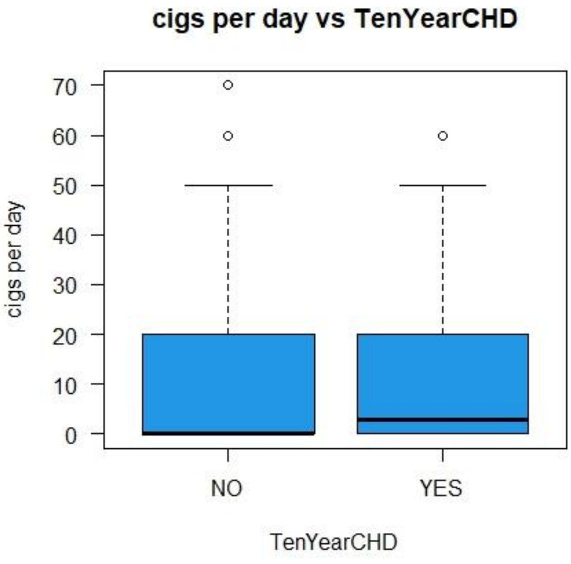
I. CONTINUOUS VARIABLES;

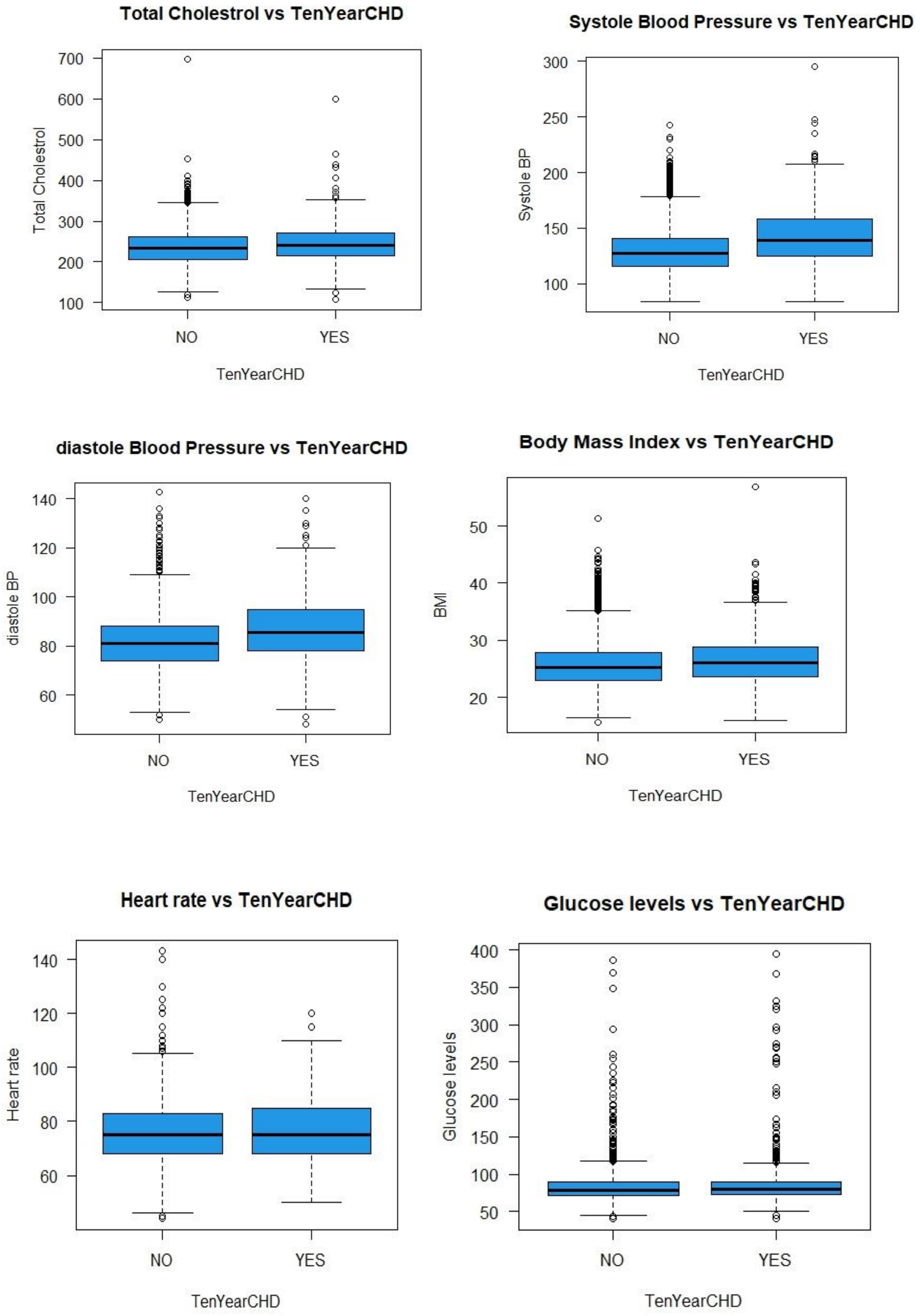
There are 8 continuous variables in the study and their summary is presented in Table-1 and visualised using Boxplots. They are plotted against the two categories of the response variable to give us an idea regarding their distribution for each category of the response variable.

Table-1; Summary Statistics of Continuous Variables

Variable	Min	1st Qua	Median	Mean	SD	3rd Qua	Max
Age	32	42	49	49.58	8.57	56	70
Cigs Per Day	0.0	0.0	0.0	9.01	11.88	20.00	70
Total Chol	107	206	234	236.9	44.35	262	696
Systole BP	83.5	117	128	132.4	22.04	144	295
Diastole BP	48	75	82	82.89	11.91	89.99	142.5
BMI	15.54	23.07	25.4	25.8	4.07	28.04	56.8
Heart rate	44	68	75	75.88	12.02	83	143
Glucose	40	72	80	82.7	22.95	90	394







figures (a – h)

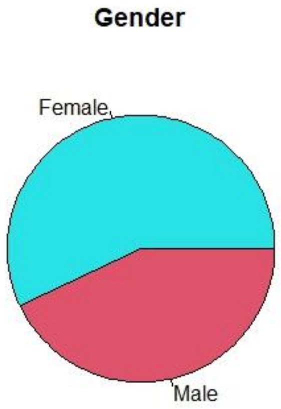
We can see from the table that mean age of all subjects was around 49 and half years with a Standard Deviation of 8 and half years. This means that most of the subjects were in their 40's and 50's where the chances of developing coronary heart diseases in the next Ten years is most likely. Of the subjects who developed risk of CHD in next ten years most of them were in their late 40's, 50's and early 60's. The median age of developing risk of CHD is higher than others.

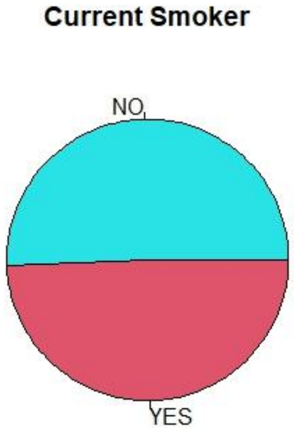
We can see that cigs per day and BMI have similar distribution with slightly higher median for subjects who developed risk of CHD.BMI has a very little Standard Deviation meaning most of the subjects were similar in terms of physique (height and weight). Median Blood Pressure levels of both systole and diastole were higher for subjects who developed risk of CHD. The mean Systole Blood pressure was 132.4 with a Standard Deviation of 22.4 and mean of Diastole blood Pressure was 82.89 with a Standard Deviation of 11.91.

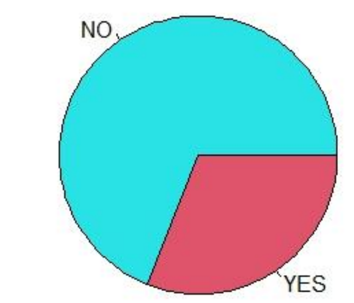
Other continuous variables like heart rate, glucose levels and total cholesterol levels have roughly similar distribution and median for both responses. The range of glucose levels in subjects is quite high compared to other variables.

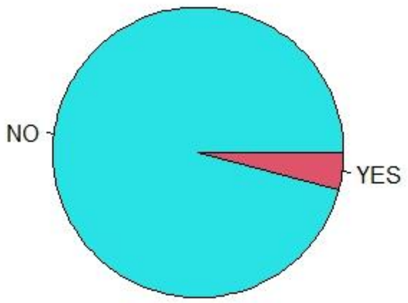
II. <u>CATEGORICAL VARIABLES;</u>

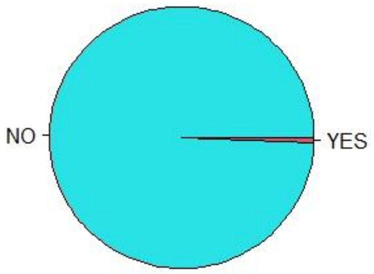
There are 7 continuous variables in the study. All the categorical variables in the study are binary in nature with success and failure (yes or no) as possible outcomes. They are summarised using Pie Charts and proportion of success. The success proportions are presented with 95% Wald Confidence Intervals to better understand the nature of subjects.

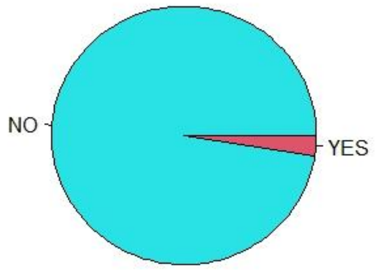





Prevalent Hypertension


BP Medication


Prevalent Stoke


Diabetes

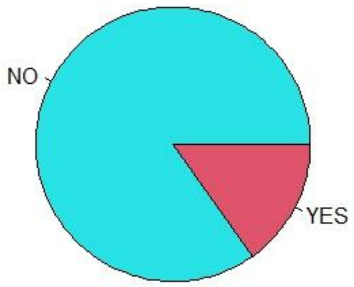

Ten Year CHD

Table-2; 95% Wald Confidence Intervals for Categorical Variables

Variable	Y=1 (success)	n (Total)	proportion	Lower limit	upperlimit
Male	1819	4238	0.429	0.414	0.444
Current Smoker	2094	4238	0.494	0.479	0.509
BP Medication	177	4238	0.042	0.036	0.048
Prevalent Stroke	25	4238	0.006	0.004	0.008
Prevalent Hypertension	1316	4238	0.310	0.296	0.324
Diabetes	109	4238	0.026	0.021	0.304
Ten Year CHD	644	4238	0.152	0.141	0.163

From the above table and Pie Charts we can say that among the subjects, females were slightly higher than males and almost half of the subjects were current smokers. While only (0.4 – 0.6) % of subjects had prevalent stroke (29.6-32.4) % had prevalent hypertension. Less than 5% of subjects were taking BP medication and less than 3% of subjects were Diabetic.

Of all the subjects, 15.2% of subjects were expected to develop the risk of CHD in ten years. With 95% Confidence we can say that (14.1 – 16.3) % of population will develop the risk of CDH in Ten years.

INFERENTIAL STATISTICS

I. LINEAR DISCRIMINANT ANALYSIS;

Training Accuracy: 0.854
Testing Accuracy: 0.856
AUC – ROC: 0.732

Training Accuracy and testing accuracy both are above 85%, which is reasonably good for a classification model. The AUC-ROC is also over 73% suggesting a good predictive power.

II. QUADRATIC DISCRIMINANT ANALYSIS;

Training Accuracy: 0.822
Testing Accuracy: 0.827
AUC – ROC: 0.704

Training and testing accuracy are over 82% and AUC-ROC is over 70%. The scores are slightly lower than Linear Discriminant analysis. Both LDA and QDA have testing accuracy slightly higher than their training accuracy.

III. LOGISTIC REGRESSION;

Training Accuracy: 0.851
Testing Accuracy: 0.851
AUC – ROC: 0.650

Training and testing accuracy is higher than QDA but less than LDA. Also, the AUC-ROC is lower than both the models suggesting lower predictive power. The logistic regression was done using all 15 independent variables.

IV. TREE BASED MODELS;

a) **Decision Tree Classifier;**
Training Accuracy: 1.0
Testing Accuracy: 0.763
AUC – ROC: 0.550

b) **Bagging;**
Training Accuracy: 0.971
Testing Accuracy: 0.845
AUC – ROC: 0.677

c) **Random Forest;**
Training Accuracy: 0.999
Testing Accuracy: 0.847
AUC – ROC: 0.694

d) **Boosting;**
Training Accuracy: 0.879
Testing Accuracy: 0.857
AUC – ROC: 0.705

e) **Boosting (CV for best params);**
Training Accuracy: 0.947
Testing Accuracy: 0.844
AUC – ROC: 0.695

We can see that most of the Tree based models have very high training accuracy and a significant dip in testing accuracy. Decision Tree Classifier, Bagging, Random Forest have all training accuracy of close to 1 while testing accuracy is about 85% or less.

This significant drop in testing accuracy suggests that the models are being overfit. The model is remembering too much from the dataset because of which training accuracy is close to 1, while testing accuracy is significantly lower. The model is not able to effectively generalise the data.

The same can be said of boosting with best parameters. When Cross Validation was done to find the best parameters for boosting the best parameters came out to be;
Learning rate: 0.01
Max depth: 3
No of estimators: 500
Subsample: 0.3

Whereas, boosting with default parameters turned out to be a better fit. The training and testing accuracy are close to each other with accuracy over 85% and also the AUC-ROC is over 70%. The default parameters for boosting are;
Learning rate: 0.1
Max depth: 3
No of estimators: 100
Subsample: 1

V. SUPPORT VECTOR MACHINES;

a. SVM with linear kernel;

Training Accuracy: 0.849
Testing Accuracy: 0.849
AUC – ROC: 0.717

b. SVM with radial kernel;

Training Accuracy: 0.850
Testing Accuracy: 0.849
AUC – ROC: 0.620

c. SVM with radial kernel (CV for best 'C');

Training Accuracy: 0.851
Testing Accuracy: 0.849
AUC – ROC: 0.606

We can see that all the support Vector machines performed well in terms of Training and testing accuracy. The AUC-ROC is higher of linear kernel while both radial kernel SVM's have lower AUC-ROC. The default value of 'C' used was 1 and the value of 'C' obtained from cross validation was 5.358.

MODEL SELECTION

I. TESTING AND TRAINING ACCURACY:

A higher training accuracy indicates that the model fits the training data well, while a higher testing accuracy suggests that the model generalizes well to new, unseen data. By comparing the training and testing accuracies of multiple models, practitioners can select the model that achieves the highest testing accuracy while also considering other factors such as model complexity, interpretability, and computational resources. This approach helps to mitigate overfitting and ensures that the selected model performs well on unseen data, making it suitable for real-world applications.

II. Area Under the ROC curve;

The Area Under the Receiver Operating Characteristic Curve (AUC-ROC) is a commonly used evaluation metric for binary classification models. It quantifies the model's ability to discriminate between positive and negative classes across different thresholds. The ROC curve plots the True Positive Rate (Sensitivity) against the False Positive Rate (1 - Specificity) for various threshold values.

The AUC-ROC represents the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance. A higher AUC-ROC value (closer to 1) indicates better discrimination performance of the model, while an AUC-ROC value of 0.5 suggests a random classifier. AUC-ROC is particularly useful for imbalanced datasets and provides a single scalar value to compare the performance of different classifiers.

Table- 3: Accuracy Scores and AUC-ROC

S.No	Models	Training Accuracy	Testing Accuracy	AUC-ROC
1	LDA	0.854	0.856	0.732
2	QDA	0.822	0.827	0.704
3	Logistic Regression	0.851	0.851	0.650
4	Decision Tree Classifier	1.0	0.757	0.550
5	Decision Tree Classifier (CV)	0.850	0.850	0.661
6	Bagging	0.980	0.841	0.677
7	Random Forest	1.0	0.849	0.694
8	Boosting	0.879	0.851	0.704
9	Boosting (CV)	0.943	0.842	0.687
10	SVM (Linear Kernel)	0.849	0.849	0.717
11	SVM (RBF Kernel)	0.850	0.849	0.620
12	SVM (RBF Kernel) (CV)	0.851	0.849	0.606

The model selection from the above fitted models were done based on the above criteria. LDA and QDA performed reasonably well in terms of accuracy and AUC-ROC. Logistic regression had lower AUC-ROC while the tree-based models were overfit. The training accuracy is close to 1 while testing accuracy is around 85%. The boosting model with default parameters performed very well. It had decent accuracy scores and Good predictive power. Support Vector machines performed well in terms of accuracy but AUC-ROC or radial kernel in both cases were lower.

Thus, based on above criteria and discussion Linear Discriminant Analysis model could be said as best fit model while boosting model with default parameters and Support Vector machine with linear Kernel performed reasonably well.

ASSESSING IMPORTANT FACTORS

I. RELATIVE RISKS / RISK RATIOS;

Relative risks were determined to evaluate how each of these Categorical variables affect the probability of risk of developing CHD in Ten years. 95% Confidence intervals for Relative risks were also determined for population. The Risk ratios and their respective 95% Wald CI can be summarised in table below.

Table-4; Relative risks and their 95% Wald CI

Variable	Risk Ratio	95% CI (lower)	95% CI (upper)
Male	1.079	1.050	1.108
Current smoker	1.016	0.991	1.042
BP medication	1.209	1.099	1.331
Prevalent stroke	1.517	1.071	2.148
Prevalent hypertension	1.183	1.144	1.223
Diabetes	1.348	1.168	1.558

The above table provides an insight as how success of each variable increases the chances of developing risk of CHD in Ten years. We can see that males have a (5-10.8) % higher chance of developing risk of CHD than females. Taking BP medication and having a Prevalent Stroke also increases the chance of developing risk of CHD by approximately (10-33)% and (7 – 15)% respectively.

Prevalent Hypertension and Diabetes have a significant effect among other variables. Having Prevalent Hypertension increases the chances of developing the risk of CDH by (14.4-22.3) % while being Diabetic increases the chances by (16.8-55.8) %.

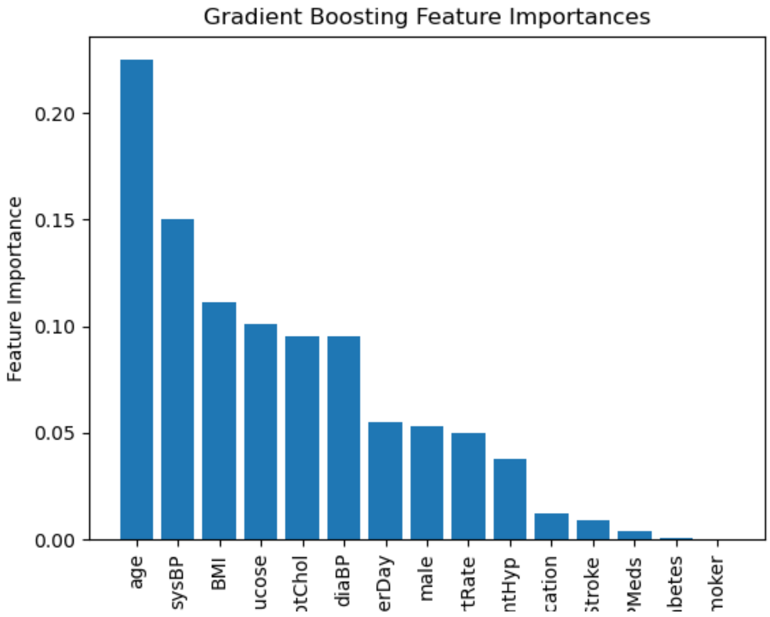
The 95% CI of Current smoker is not significant as it contains 1 in its interval. This means that the chances of developing the risk of CDH is independent of Smoking Status. Or, there is equal chance of developing the risk of CDH for both smokers and non-smokers.

II. FEATURE IMPORTANCES IN BOOSTING;

As we have seen during model selection that bagging model with default parameters performed reasonably well in terms of Accuracy scores and AUC-ROC, we can use model for inferences on important features in predicting the risk of coronary heart diseases.

Table-5: Feature Importances in Boosting

S.No	Variables	Feature Importances
1	Male	0.053
2	Age	0.224
3	Education	0.011
4	Current Smoker	0.001
5	Cigs Per Day	0.054
6	BP Meds	0.004
7	Prevalent Stroke	0.009
8	Prevalent Hyp	0.036
9	Diabetes	0.0007
10	Tot Chol	0.09
11	SYS BP	0.149
12	Dia BP	0.096
13	BMI	0.115
14	Heart Rate	0.047
15	Glucose	0.104



CONCLUSION AND RECOMMENDATIONS

The dataset contains a collection of Continuous variables and Categorical variables. It has variables corresponding to current medical status and medical history with Risk of CHD in Ten Years as the outcome variable. The proportion of females in dataset is slightly higher in than males and almost 50% were current smokers. Very few subjects were Diabetic, under BP medication and has history of Prevalent stroke. Age came out to be a significant variable with higher median for subjects who developed risk of CHD. Median levels of both Systole and Diastole Blood Pressures were higher for subjects who developed Risk of CHD in ten Years.

LDA and QDA performed reasonably well in terms of accuracy and AUC-ROC. Logistic regression had lower AUC-ROC while the tree-based models were overfit. The training accuracy is close to 1 while testing accuracy is around 85%. The boosting model with default parameters performed very well. It had decent accuracy scores and Good predictive power. Support Vector machines performed well in terms of accuracy but AUC-ROC or radial kernel in both cases were lower.
Linear Discriminant Analysis model could be said as best fit model while boosting model with default parameters and Support Vector machine with linear Kernel performed reasonably well.

Among all the features we can see that Age and Systolic blood Pressure were very significant in predicting the Risk of Coronary heart diseases. We can see that males have a (5-10.8) % higher chance of developing risk of CHD than females. Taking BP medication and having a Prevalent Stroke also increases the chance of developing risk of CHD by approximately (10-33)% and (7 – 15)% respectively.
Having Prevalent Hypertension increases the chances of developing the risk of CDH by (14.4-22.3) % while being Diabetic increases the chances by (16.8-55.8) %.

The major limitation of the study was equal weightage was given to all variables while from the knowledge of medical science we know that few variables are more important than others to certainly predict the Risk of Coronary Heart Diseases. The study can be continued further by looking more deeper into the association of the variables. If more Variables like daily physical activity, obesity, diet. Etc will be included the study could give more deeper insights.

The same process should be repeated over the same dataset with different entries. This can also be achieved by repeated measures methods. The models contain 15 features which may introduce noise. LDA, QDA and logistic regression could be further investigated with significant variables and Dimensionality reduction. Also, further complex algorithms like neural networks and deep learning algorithms could be deployed as well for better accuracy and predictive power.

USE CASES AND APPLICATIONS

1. **Early Diagnosis**: Predictive models can be used to identify individuals at high risk of developing heart diseases before symptoms appear. Early diagnosis allows for timely interventions and lifestyle modifications, potentially preventing the onset or progression of heart diseases.
2. **Personalized Risk Assessment**: By analysing various risk factors such as age, gender, family history, blood pressure, cholesterol levels, and lifestyle habits, predictive models can provide personalized risk assessments for individuals. This enables healthcare professionals to tailor preventive strategies and treatment plans based on individual risk profiles.
3. **Population Health Management**: Predictive models can analyse large-scale population health data to identify trends, patterns, and risk factors associated with heart diseases at the population level. This information can be used by public health agencies and policymakers to develop targeted interventions and preventive strategies for high-risk populations.
4. **Telemedicine and Remote Monitoring**: Remote monitoring devices and wearable sensors can collect real-time health data, such as heart rate, blood pressure, and physical activity levels. Predictive models can analyse this data to continuously assess an individual's risk of developing heart diseases and provide timely alerts or interventions, when necessary, even in remote or underserved areas.
5. **Clinical Decision Support**: Predictive models integrated into electronic health record systems can assist healthcare providers in making more informed clinical decisions. By providing risk predictions and recommendations based on patient data, these models can help prioritize resources, guide treatment decisions, and optimize patient outcomes.
6. **Drug Development and Clinical Trials**: Predictive models can be used in pharmaceutical research to identify potential drug targets, predict the efficacy and safety of new therapies, and stratify patient populations for clinical trials. By selecting the most promising candidates and optimizing trial design, predictive modelling can accelerate the development of new treatments for heart diseases.
7. **Healthcare Resource Allocation**: Predictive models can forecast future healthcare needs and resource requirements based on projected trends in heart disease prevalence and risk factors. This information can help healthcare organizations allocate resources more efficiently, plan preventive programs, and improve access to care for at-risk populations.

REFERENCES

1. Khoshnood B, Lelong N, Houyel L, Thieulin AC, Jouannic JM, Magnier S, Delezoide AL, Magny JF, Rambaud C, Bonnet D, Goffinet F; EPICARD Study Group . Prevalence, timing of diagnosis and mortality of newborns with congenital heart defects: a population-based study. *Heart*. 2012;98:1667–1673. [PubMed] [Google Scholar]
2. Hoffman JI, Kaplan S, Liberthson R. Prevalence of congenital heart disease. *Am Heart J*. 2004;147:425–439. [PubMed] [Google Scholar]
3. Tennant PW, Pearce MS, Bythell M, Rankin J. 20-year survival of children born with congenital anomalies: a population-based study. *Lancet*. 2010;375:649–656. [PubMed] [Google Scholar]
4. Zhao QM, Ma XJ, Ge XL, Liu F, Yan WL, Wu L, Ye M, Liang XC, Zhang J, Gao Y, Jia B, Huang GY; Neonatal Congenital Heart Disease screening group . Pulse oximetry with clinical assessment to screen for congenital heart disease in neonates in China: a prospective study. *Lancet*. 2014;384:747–754. [PubMed] [Google Scholar]
5. Marelli AJ, Mackie AS, Ionescu-Ittu R, Rahme E, Pilote L. Congenital heart disease in the general population: changing prevalence and age distribution. *Circulation*. 2007;115:163–172. [PubMed] [Google Scholar]
6. Moons P, Engelfriet P, Kaemmerer H, Oechslin E, Mulder B; Expert Committee of Euro Heart Survey on Adult Congenital Heart Disease . Delivery of care for adult patients with congenital heart disease in Europe: results from the Euro Heart Survey. *Eur Heart J*. 2006;27:1324–1330. [PubMed] [Google Scholar]
7. https://www.kaggle.com/datasets/dileep070/heart-disease-prediction-using-logistic-regression
8. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6585327/#jah34125-bib-0006