CHD Prediction using Machine Learning

Introduction:

In this analysis, our aim was to predict Coronary Heart Disease (CHD) for males in a high-risk region of the Western Cape, South Africa. The chosen dataset includes nine features, such as systolic blood pressure (sbp), cumulative tobacco use (tobacco), low-density lipoprotein cholesterol (ldl), adiposity, family history of heart disease (famhist), type-A behavior (typea), obesity, current alcohol consumption (alcohol), and age.

Exploratory Data Analysis (EDA):

Before beginning our modelling, we conducted an exploratory data analysis to gain insights into the dataset. This involved checking the structure of the dataset, obtaining the summary statistics and examining the distribution of each feature and its relationship with the target variable, CHD.

Factors	Minimum	Maximum	1st Quartile	Median	Mean	3rd Quartile
sbp	101.00	218.00	124.00	134.00	138.30	148.00
tobacco	0.00	31.20	0.0525	2.00	3.6356	5.50
ldl	0.98	15.33	3.283	4.34	4.74	5.79
adiposity	6.74	42.49	19.77	26.11	25.41	31.23
typea	13.00	78.00	47.00	53.00	53.10	60.00
obesity	14.70	46.58	22.98	25.80	26.04	28.50
alcohol	0.00	64.00	0.51	7.51	17.04	23.89
age	15.00	1.00	31.00	45.00	42.82	55.00
chd	0.00	1.00	0.00	0.00	0.3463	1.00

Table 1: Summary Statistics of Dataset

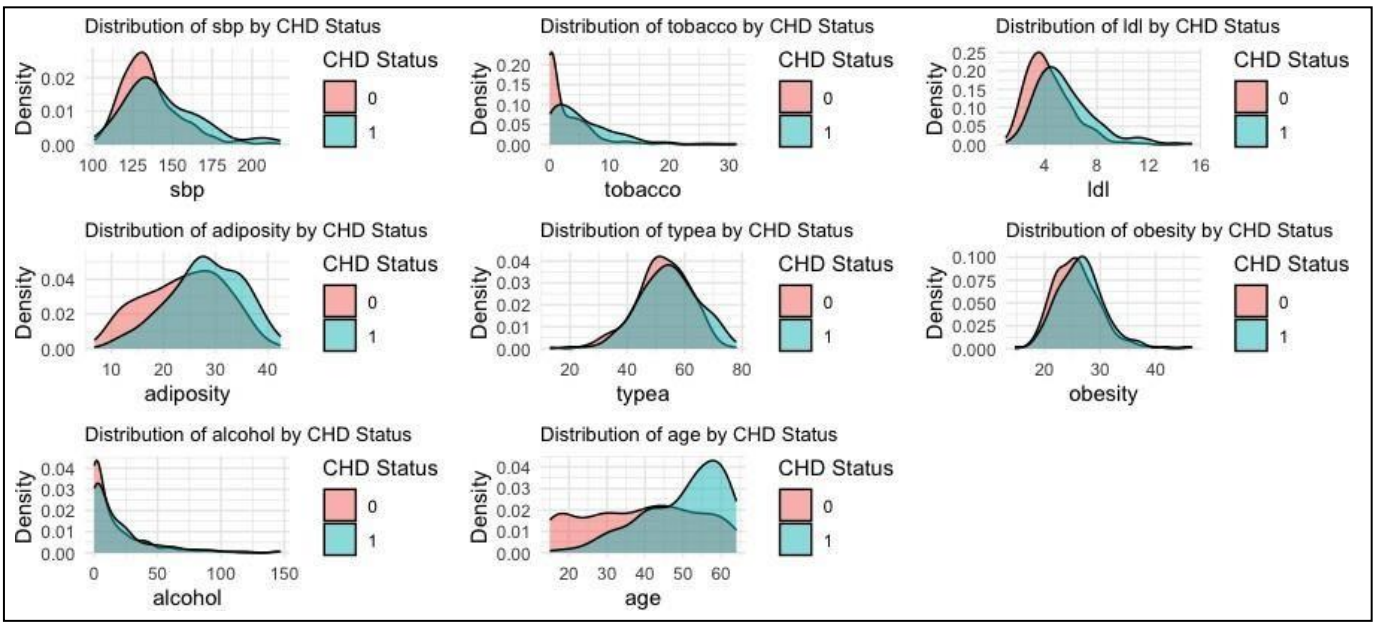

Figure 1: Distribution-Comparison Multi-Plot

The density plot for each factor is shown in the subplots in Figure 1 above; the density (probability distribution) is shown on the y-axis, while the values of the factor are represented on the x-axis.

1. Systolic Blood Pressure (sbp): Compared to people without CHD, the distribution of systolic blood pressure readings seems to be skewed higher in those with CHD, with average sbp being around 138.3.

2. Tobacco (tobacco) and Alcohol Consumption (alcohol): The distributions of these two variables indicate that those without CHD tend to have a much broader distribution with greater levels of consumption, whereas those with the disease typically have similar and higher densities around lower values.

3. Low-Density Lipoprotein (ldl) Cholesterol: The distributions of LDL cholesterol levels in the two groups are very similar, with a small density for higher values in those who have CHD, with values skewed greater than the average of 4.74.

4. Obesity and Adiposity: Compared to individuals without the condition, those with CHD have skewed distributions that lead to higher values of obesity and adiposity. This suggests that those with higher levels may be more likely to get CHD.

5. Type A Behavior (typea): The distribution of type A behavior is typically higher in individuals with CHD.

6. Age: The age distributions suggest that older individuals may be more prone to CHD.

Ridge Penalty Regression:

For modelling, we employed logistic regression with a ridge penalty which helps prevent overfitting by penalizing large coefficients. Post fitting the ridge regression model, we evaluated its performance with accuracy as the primary metric. The ridge penalty regression model achieved an **accuracy** of approximately **0.7467 (74.67%)**, indicating correctly predict CHD status for about 74.67% of the dataset samples.

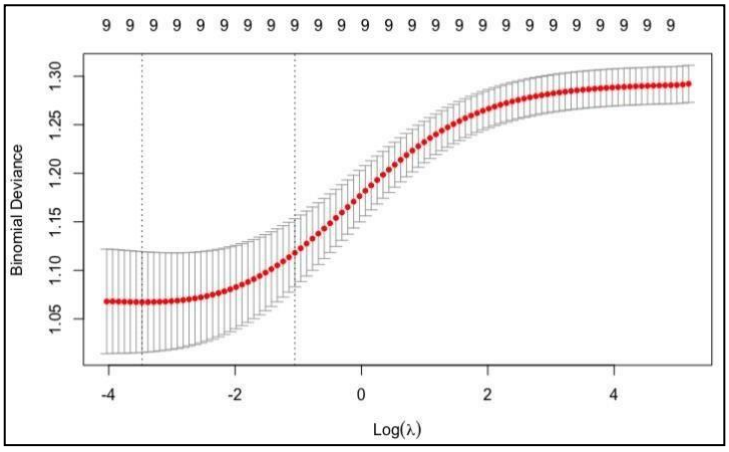

Figure 2: Binomial Deviance Plot for Ridge Regression

The plot in Figure 2 shows the Binomial Deviance across different log (λ) values. The leftmost vertical line at log (λ) = -4 indicates the minimum deviance with all predictors included. The "1se" line at log (λ) ≈ -0.65 represents the largest lambda within one standard error of the minimum, by removing less important predictors. The numbers at the top (9…) indicate the log (λ) values where predictor coefficients become zero.

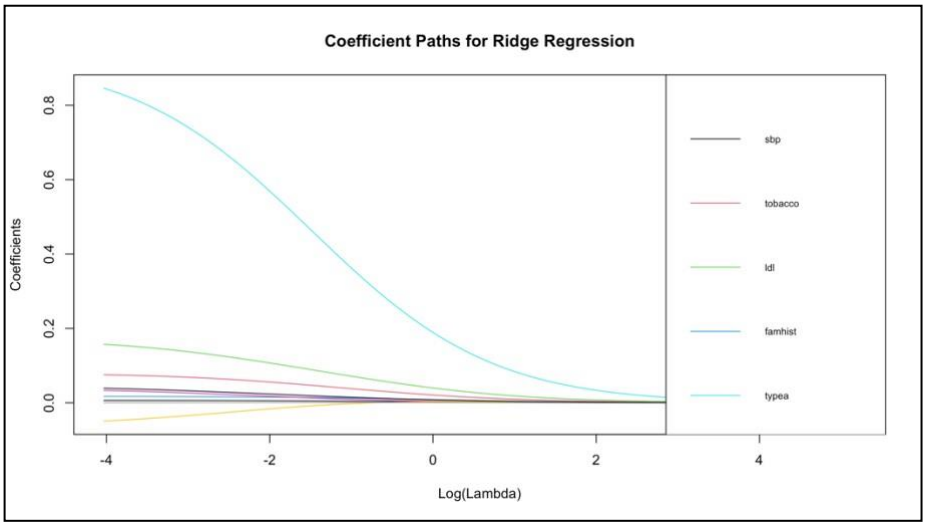

Figure 3: Coefficient Paths

The plot in Figure 3 shows the coefficient paths of the logistic regression model with ridge penalty as a function of log(lambda). Each line represents the coefficient for one predictor, and it demonstrates how the coefficients shrink as our regularization parameter lambda increases. The predictors with coefficients that are initially higher are more significant to the model. In this case these are: typea (neon blue line), ldl (green line) and tobacco (pink line). As lambda increases, these coefficients decrease but do not reach zero, indicating that ridge regression does not force coefficients to exactly zero, unlike a lasso regression (L1 penalty).

Interpretation:

- The leftmost part of the plot (low log(lambda)) includes more variance with larger coefficients, implying overfitting.
- The rightmost part of the plot (high log(lambda)) shows increased bias with coefficients close to zero, implying underfitting.
- The optimal lambda (around log(lambda) ≈ -4, as per the earlier deviance plot) balances this trade-off, providing a good balance between bias and variance.

Support Vector Machines (SVM):

The SVM model was trained and evaluated for predicting CHD using various physiological features. After splitting the dataset into training and testing sets, the SVM model was fitted with a linear kernel. SVM is particularly well-suited for this task due to its ability to handle high- dimensional data and capture complex relationships between predictors and the target variable. However, with an accuracy of approximately **69.57%**, the SVM model's performance was slightly lower than that of the ridge penalty regression model previously evaluated. This suggests that while SVM offers advantages in handling complex data structures, further optimization may be necessary to overcome the predictive performance of ridge regression in this specific context.

Decision Tree:

The decision tree model provides a concise and interpretable framework for predicting the likelihood of heart disease based on various physiological features. Starting with age as the root node, the tree branches out into distinct groups based on factors such as family history, tobacco use, LDL cholesterol levels, adiposity, and type A behaviour, ultimately assigning probabilities of heart disease to leaf nodes.

With a reported accuracy of 78.78%, the decision tree model could outperform ridge penalty regression for this dataset, particularly due to its ability to capture nonlinear relationships and prioritize interpretability. Moreover, the hierarchical structure of the decision tree offers insights into feature importance, aiding in feature selection.

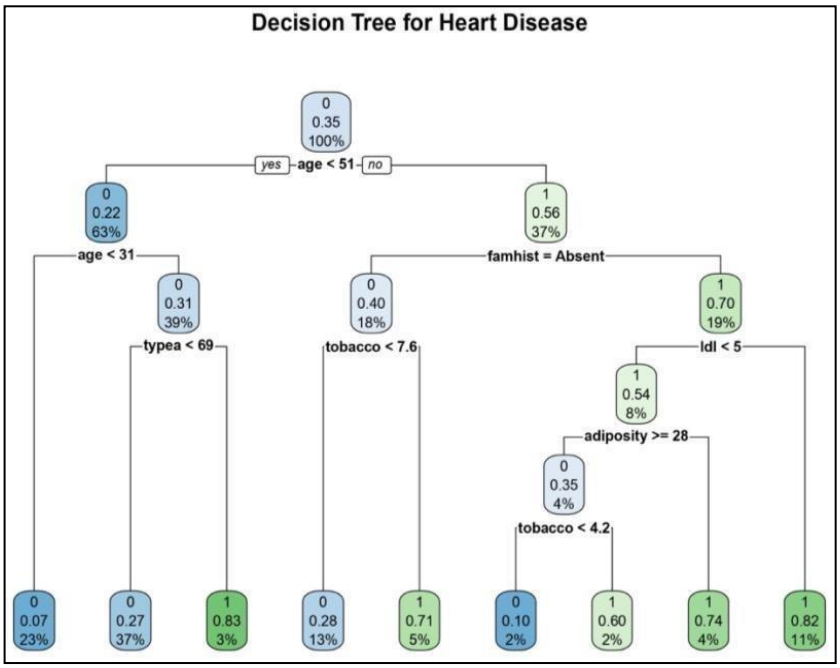

Figure 4: Decision Tree for Heart Disease Dataset

However, we need to note that decision trees are susceptible to outliers and noise, which sometimes can result in overfitting, high variance, and instability. Thus, as a suggestion for further analysis, random forests can assist us– acting as a robust solution that mimics a team of doctors, with each tree in the forest providing its expert opinion. By aggregating the predictions of multiple decision trees, random forests reduce the impact of individual errors and improve the overall predictive accuracy for CHD risk assessment.

Random Forest:

The Random Forest works by combining predictions from several individual decision trees, each trained on different parts of the data. This method is particularly in our case useful because it can handle complex relationships between various factors that might influence CHD risk, like age, cholesterol levels (ldl), and lifestyle habits (alcohol & tobacco). By analyzing these factors together, we can get insights into which aspects have the greatest impact on CHD risk.

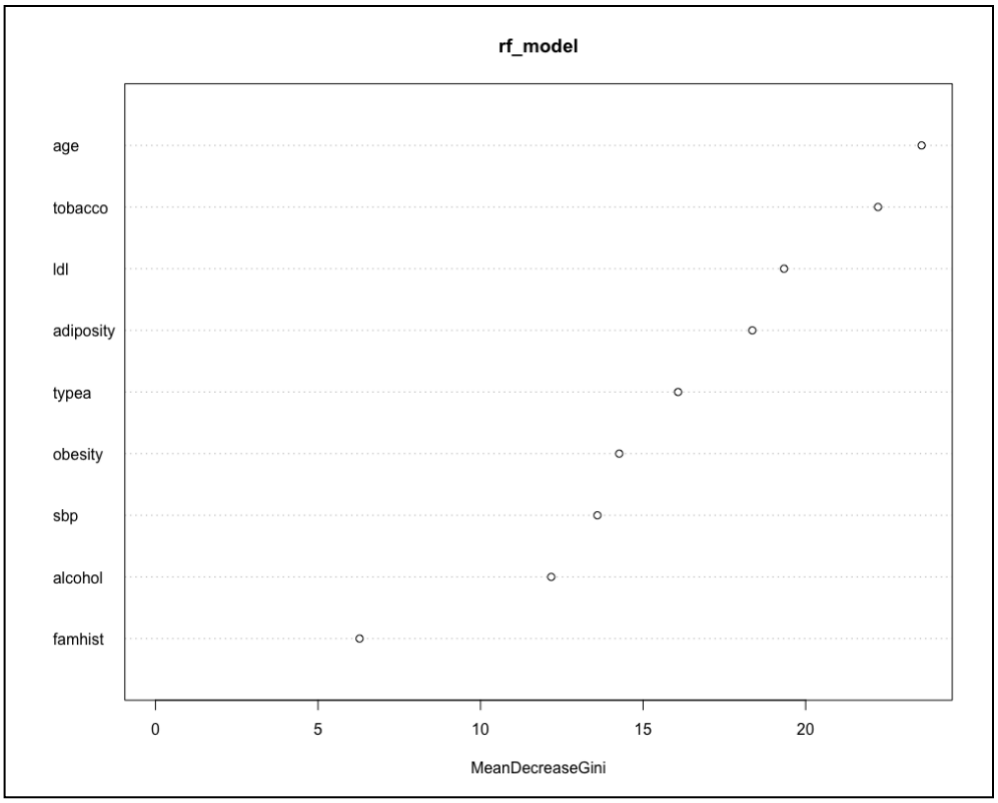

Figure 5: Variable Importance Plot of the Random Forest Model

The plot in Figure 4 Represents the importance scores of the variables. The 'Mean Decrease Gini' score on the X-axis measures the impact of excluding each variable on the homogeneity of the nodes and leaves in the trees. A higher value suggests that the variable is crucial for making pure splits in the decision trees.

Variables at the top of the plot with higher importance scores contribute more significantly to the model's predictions. These are the features that the Random Forest model relies on the most to make accurate predictions. This implies that **age**, **tobacco** and **ldl** are critical predictors of CHD risk. The model's accuracy shall decrease significantly if these factors are excluded.

Prediction

We check the Decision Tree model by inputting new data of a person having the following values for each respective feature: age = 45, sbp = 130, tobacco = 12, ldl = 5, typea = 50, obesity = 25, alcohol = 14, adiposity = 23, famhist = Present. By running the Decision Tree model, we get the Prediction result of '**1**'. This indicates that as per our Model, this person with these features will end up **having** Coronary Heart Disease (CHD).

Conclusion

In conclusion, we can see that the factors of Age, Tobacco consumption, LDL and adiposity as important predictive factors for CHD risk.
We can consider comparing the ridge regression performed with a lasso regression (L1 penalty) plot to further see which predictors are entirely removed by lasso and retained by ridge.

We arrived at an accuracy of **74.67%** for **ridge penalty regression**, **69.57%** for **Support Vector Machines**, **69.56%** for **Random Forests** model and the highest of **78.78%** for the **Decision Tree** analysis.