1.0 Introduction

Coronary heart disease (CHD) is a major cause of mortality worldwide, making early diagnosis crucial for prevention (Capotosto et al., 2018). This report compares various classifiers' ability to predict CHD based on nine risk factors such as age, tobacco use, and family history. Rather than relying on accuracy, which can be misleading for imbalanced datasets, AUC (Area Under the ROC Curve) is the primary evaluation metric.

The report follows a structured approach:

- Exploratory Data Analysis (EDA) to clean data, identify correlations, and assess numerical distributions.
- Fitting a baseline logistic regression model with ridge regularisation
- Comparison of alternative classifiers: kNN, Random Forest, XGBoost and SVM.
- Evaluates the overall fit and diagnosis accuracy of the model through ROC curve and confusion matrices

2.0 Exploratory Data Analysis

Boxplots *(Figure 2.0)* reveal outliers in Alcohol, LDL, and Tobacco, which may artificially inflate correlations with CHD. However, these values were retained to maintain real-world data integrity.

Histograms *(Figure 2.1)* show skewed distributions and extreme values, guiding normalisation choices like MinMax scaling for kNN and SVM.

The Correlation Matrix & Heatmap *(Figure 2.2)* detect multicollinearity, but no variables exceeded the 0.8 threshold, so none were removed.

The PCA Plot *(Figure 2.3)* shows CHD and non-CHD cases overlapping, suggesting linear models may struggle, while non-linear classifiers (Random Forest, XGBoost) may perform better.

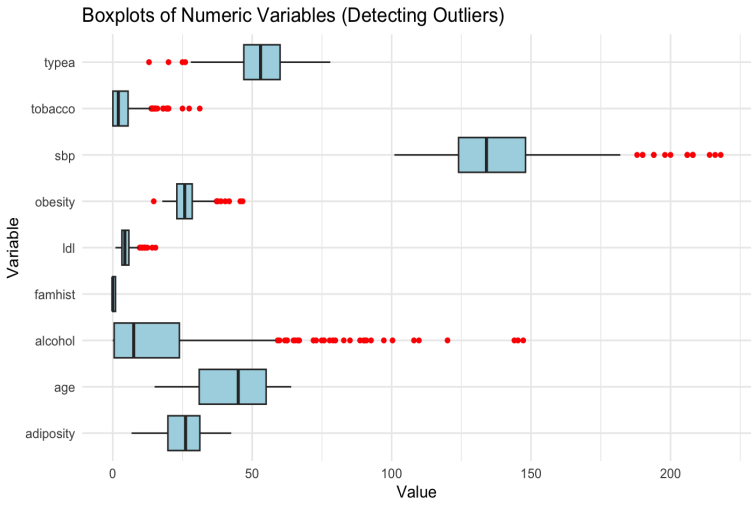

Figure 2.0

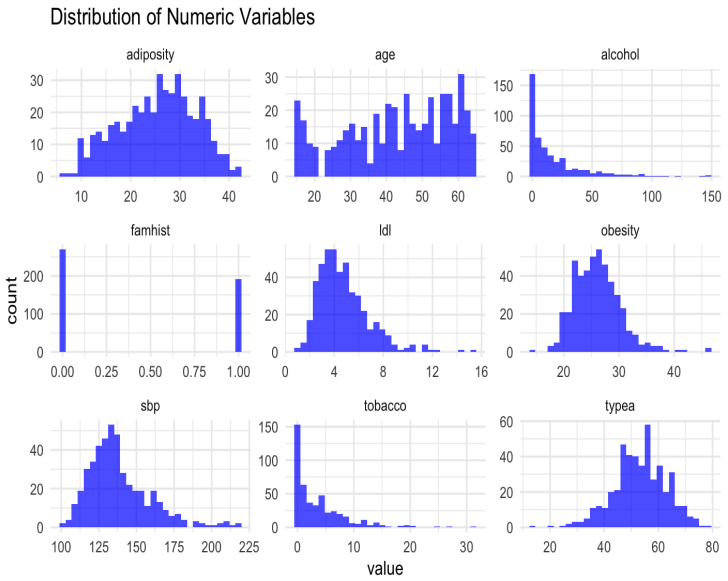

Figure 2.1

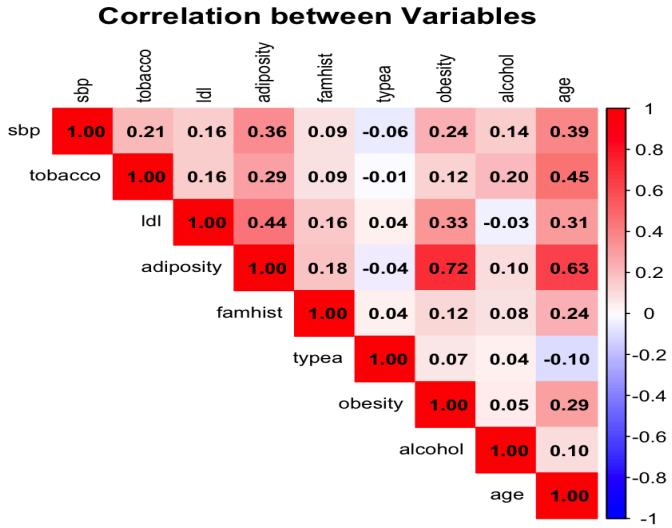

Figure 2.2

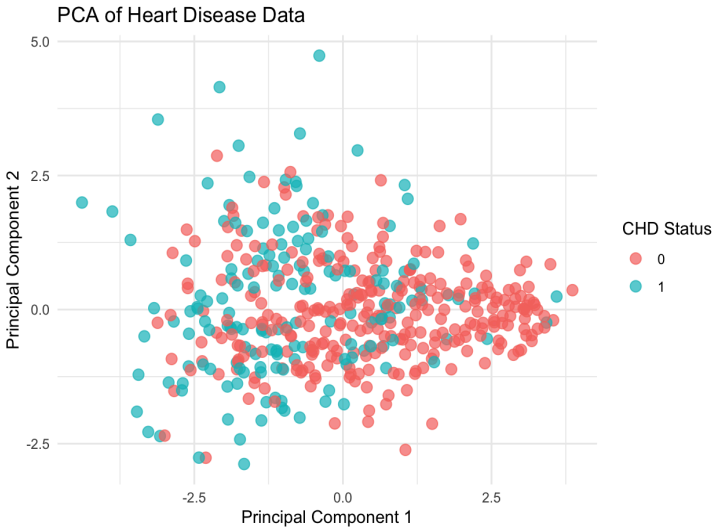

Figure 2.3

2.1 Data Preprocessing

Missing values were removed to ensure data integrity. The dataset was split into training and test sets, keeping the test set unbiased for evaluation. SMOTE was applied only to the training data to address class imbalance while preserving real-world distributions in the test set. To further maintain balance, controlled undersampling ensured both classes had equal representation, resulting in a more stable and unbiased model assessment.

3.0 Results

When diagnosing CHD, selecting the right model requires balancing predictive power, generalisation, and clinical reliability. While accuracy is often prioritised, AUC is a more holistic measure, assessing a model's ability to differentiate between CHD and non-CHD cases across all thresholds. Given the risks of misdiagnosis, reducing false negatives is crucial to prevent untreated patients.

Classifier	Test Accuracy (%)	Test AUC (%)	AUC Diff (Test - Train) (%)	True Pred.	False Neg.	False Pos.
Ridge Regression	71.7	79.5	1.6	62	6	24
kNN	69.6	74.7	-3.4	65	7	20
Random Forest	73.9	81.2	-12.8	67	8	17
XGBoost (Full)	69.6	75.8	-1.6	64	4	24
XGBoost (Selected)	71.7	76.4	-1.3	67	5	20
SVM Linear	71.7	81.4	2.5	66	4	22
SVM Radial	73.9	81.8	2.3	68	4	20

Table 1: Summary Statistic of Classification Models

Table 1 summarises test accuracy, AUC, and false negatives. RF had the highest accuracy (73.9%) but overfitted, making it unreliable. SVM Radial (AUC: 81.8%) performed best, maintaining low false negatives and strong generalisation, making it the most effective CHD predictor.

3.1 Ridge Regression

A ridge regression model was fitted as a baseline. L2 regularisation prevents overfitting by shrinking coefficients, reducing variance at the cost of bias. While stable, its linear nature limited its ability to capture complex patterns, as seen in its lower AUC (79.5%) compared to SVM and Random Forest. A higher false negative rate made it less favorable for CHD diagnosis.

3.2 Other Classifiers

All classifiers explored were included in the final report to ensure a complete evaluation. Ridge regression and kNN (AUC: 79.5%, 74.7%) struggled to capture complex relationships, aligning with PCA findings, which showed significant CHD/non-CHD overlap. From the PCA plot, we can expect that linear models won't be the best at classifying the data, as CHD cases do not separate cleanly along linear boundaries.

Random Forest (AUC: 81.2%) performed well but showed overfitting, with a significant AUC drop (-12.8%). This overfitting is likely due to decision trees' sensitivity to outliers in Alcohol, LDL, and Tobacco, leading to unnecessary splits & reduced generalisability

XGBoost (AUC: 75.8% full, 76.4% selected) underperformed compared to RF despite SMOTE balancing training data, likely due to class imbalance. While it captures interactions better than ridge regression, its moderate AUC and lower test accuracy suggest it is less effective for this dataset.

3.3 Best Classifier: SVM Radial

SVM (Linear: 81.4%, Radial: 81.8%) emerged as the strongest classifiers, with low false negatives (4 cases each) and stable AUC differences (Linear: +2.5%, Radial: +2.3%), indicating strong generalisation. Radial outperformed Linear, reinforcing PCA findings as CHD cases do not separate clearly along linear boundaries, a non-linear model like SVM Radial is better suited.

4.0 Feature Importance

Feature	Random Forest	XGBoost	SVM Models
Age	2.45	100.0	100.0
ldl	2.58	29.2	73.5
Tobacco	0.46	7.1	68.0
Adiposity	-0.11	2.2	64.6
Famhist1	0.21	11.1	56.7
sbp	0.50	0.0	48.8
Obesity	-1.46	0.0	27.9
Alcohol	-1.48	0.0	4.4
Type A	-0.72	0.0	0.0

Table 2: Feature Importance

Age and ldl consistently rank highest in SVM and XGBoost, aligning with medical findings that age-related vascular changes and high LDL contribute to arterial plaque buildup (Badimon & Vilahur, 2012), increasing CHD risk. Tobacco (68.0) and adiposity (64.6) are also key, reinforcing lifestyle-related risk factors. Tobacco damages arteries and increases LDL, while adiposity contributes to hypertension and metabolic disorders.

Random Forest exhibits overfitting by distributing feature importance too broadly, reducing emphasis on LDL (2.58) and family history (0.21). This suggests it captures

noise rather than meaningful patterns. SVM Radial and Linear assigning identical feature importance suggests a consistent classification approach, optimising CHD detection. XGBoost, though selective, overlooks SBP and obesity, likely due to interactions with stronger predictors. Type A personality is irrelevant across all models, aligning with research from Steptoe & Molloy (2007).

Ridge regression was excluded due to coefficient shrinkage limiting interpretability. Missing features like high-density lipoprotein (HDL) cholesterol and blood glucose could enhance risk assessment (Kannel, 1987).

5.0 Conclusion

SVM (Linear: 81.4%, Radial: 81.8%) demonstrated the strongest classification performance, with low false negatives (4 cases each) and minimal AUC variation, suggesting robust generalisation. Future improvements could enhance test accuracy and AUC by incorporating additional features such as sex and race, which account for 63-80% of the prognostic performance in cardiovascular risk models, providing a more comprehensive CHD risk assessment (Pencina et al., 2019). This model was trained solely on male patients from a specific region in South Africa, limiting its applicability to females and other populations. Incorporating diverse data, including females and individuals from different regions, would enhance the model's generalisation to a broader population. This analysis used 10-fold cross-validation to ensure a consistent comparison across all models. An improvement could be implementing dynamic cross-validation, optimising the number of folds for each model to maximise accuracy. Additionally, dynamic thresholds could optimise predictive ability by adjusting decision boundaries based on Youden's Index (see Appendix). Overall, improving data quality and model tuning will enhance CHD risk prediction, enabling earlier interventions and potentially reducing deaths associated with heart disease.

6.0 References

Badimon, L. and Vilahur, G., 2012. LDL-cholesterol versus HDL-cholesterol in the atherosclerotic plaque: inflammatory resolution versus thrombotic chaos. Annals of the New York Academy of Sciences, 1254(1), pp.18-32.

Capotosto, L., Massoni, F., De Sio, S., Ricci, S. and Vitarelli, A., 2018. Early diagnosis of cardiovascular diseases in workers: role of standard and advanced echocardiography. BioMed Research International, 2018(1), p.7354691.

Kannel, W.B., 1987. Hypertension and other risk factors in coronary heart disease. American Heart Journal, 114(4), pp.918-925.

Pencina, M.J., Navar, A.M., Wojdyla, D., Sanchez, R.J., Khan, I., Elassal, J., D'Agostino Sr, R.B., Peterson, E.D. and Sniderman, A.D., 2019. Quantifying importance of major risk factors for coronary heart disease. Circulation, 139(13), pp.1603-1611.

Steptoe, A. and Molloy, G.J., 2007. Personality and heart disease. Heart, 93(7), pp.783-784.

7.0 Appendix : Implementing Dynamic Threshold

Optimising the threshold using Youden's Index improves CHD diagnosis by balancing sensitivity and specificity. The SVM Linear (0.436) and SVM Radial (0.471) thresholds prioritise minimising false negatives, critical in medical settings where missed CHD cases can delay treatment and increase mortality risk. Adjusting the SVM Radial threshold from 0.5 to 0.471 reduced false negatives from 4.34% to 3.8%, enhancing recall while maintaining precision.

Lowering the threshold increases recall, detecting more CHD cases at the cost of slightly higher false positives. However, in clinical practice, additional testing is preferable to missed diagnoses, making this trade-off essential for early detection and reducing fatal cardiac events

Classifier	Optimal Threshold	AUC Score(%)	True Pred.	False Neg.	False Pos.
SVM Linear	0.436	81.4	66	2	24
SVM Radial	0.471	81.8	67	2	23

Table 3 : Dynamic Threshold on Classification