Predicting Coronary Heart Disease in High-Risk South African Males: A Comparative Analysis of Linear and Non-linear Classification Methods

Shaan Ali Remani

March 2025

Abstract

Men in the Western Cape region of South Africa suffer a *nearly-20% risk* of Coronary Heart Disease (CHD), placing them amongst the most high-risk individuals globally (Crowley et al. 2024). This report compares *linear* and *non-linear* classification methods to predict with the *highest accuracy* whether an individual suffers from CHD. Using a dataset of *462* observations with *9* clinical, demographic, and lifestyle features, and *11* models including *Logistic Regression with Ridge Penalty*, *Linear Discriminant Analysis (LDA)*, *Support Vector Machine with Linear Kernel (SVM)*, *AdaBoost*, *Gradient Boost*, *Decision Tree*, and *k-Nearest Neighbors (kNN)*, we find that linear models, particularly *LDA*, outperform non-linear models, achieving the highest F1-Score of **66%** with highest *recall* and *precision.* Performance was evaluated using *F1-Score*, *ROC-AUC*, *precision* and *recall*, and *sensitivity analyses*, removing alcohol from the predictors and excluding outliers, further confirmed the robustness of the linear decision boundary.

Contents

List of Abbreviations

CHD Coronary Heart Disease. 5, 9

kNN-CV k-Nearest Neighbours where k is selected through cross validation. 8

LDA Linear Discriminant Analysis. 5, 8, 9

QDA Quadratic Discriminant Analysis. 9

ROC-AUC Area Under ROC Curve. 5, 15

SVM Support Vector Machine. 8

List of Figures

List of Tables

1 Introduction

Our analysis serves two aims: firstly, to identify the best-performing predictive model (from a pool of 11 tested models); and, secondly, to determine if the classification boundary is linear or non-linear. We first explore the feature space through EDA, before introducing Logistic Regression with Ridge Penalty. We conclude, with high likelihood, that the decision boundary is linear, and that **Linear Discriminant Analysis (LDA)** is the best classifier for our data with an F1-score of **66%**.

2 Exploratory Data Analysis

The *Heart Disease* dataset contains *462* observations with *9* predictors spanning clinical, demographic, and lifestyle features, (Table 1), and a binary target variable, Coronary Heart Disease (CHD). Figure 1 suggests a mild class imbalance in CHD, and so simple *accuracy* or *classification error rate* cannot be used to measure model performance.

Feature	**Full Name & Units**	**Domain**	**Description**
SBP	Systolic Blood Pressure (mmHg)	Clinical Measurement	Blood pressure during heartbeats
LDL	LDL Cholesterol (mmol/L)	Clinical Measurement	Low-density lipoprotein cholesterol level
ADIPOSITY	Adiposity Index	Clinical Measurement	Body fat measure
OBESITY**	Body Mass Index (kg/m^2)	Clinical Measurement	Weight-to-height ratio
AGE	Age (Years)	Demographic	Age of the individual
FAMHIST	Family History of CHD	Demographic	Family history (1=Present, 0=Absent)
TOBACCO	Tobacco Consumption (kg)	Lifestyle	Tobacco consumed
ALCOHOL*	Current Alcohol Consumption	Lifestyle	Alcohol consumed
TYPEA	Type-A Behaviour Score	Lifestyle	Score assessing personality-driven lifestyle

Table 1: Feature descriptions with full names, units, domains, and explanations. A single asterisk (*) indicates that units were not available or easily decipherable. A double asterisk (**) denotes an inferred measure based on the data provided.

The highly right-skewed distributions of alcohol and tobacco (Figure 3), coupled with the outliers in alcohol and SBP (Figure 2) support the need for alternative performance metrics such as *Confusion Matrix Analysis* and *Area Under ROC Curve (ROC-AUC)*.[1][2]

[1] For a brief introduction to Confusion Matrices, see Appendix A.

[2] Despite the outliers in alcohol and SBP, these features are retained and scaled like all others since long-term heavy drinking is linked to heart disease (Drinkaware 2021), and British Heart Fondation (2025) suggests that alcohol consumption has a non-linear, indirect relationship to heart disease. See Appendix C for full Sensitivity Analysis.

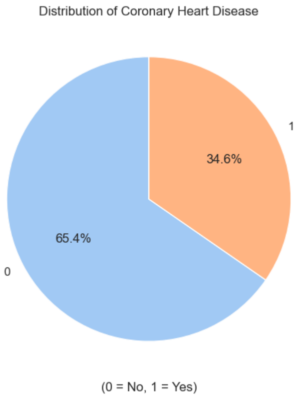

Figure 1: Distribution of CHD, the target variable.

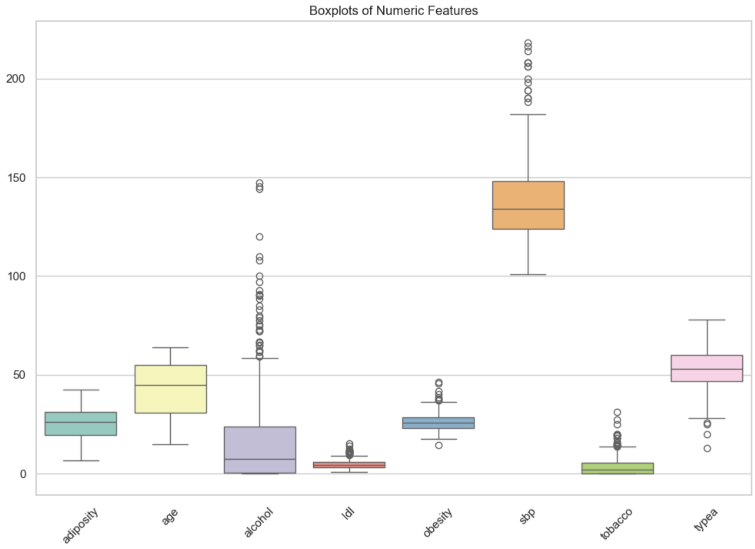

Figure 2: Boxplots of numeric features illustrating the spread of outliers.

Dimension reduction to the first two principal components (PC1 and PC2) shows no clear feature clustering (Figure 4), and so our analysis focuses on linear and non-linear classification models.[3]

3 Logistic Regression with Ridge Penalty

In multivariate linear regression, the model is estimated as $\hat{Y} = X\hat{\beta}$, where $\hat{Y}$ represents the predicted response, X is the matrix of predictors, and $\hat{\beta}$ is the vector of coefficients (one for each predictor and a constant). In logistic regression, $\hat{Y}$ is transformed using the logistic function:

$$P(X = x) = \frac{e^{\hat{Y}}}{1 + e^{\hat{Y}}}$$

[3] Pairplots in Appendix B reveal that no pair of features segregate CHD and a multivariate analysis is necessary.

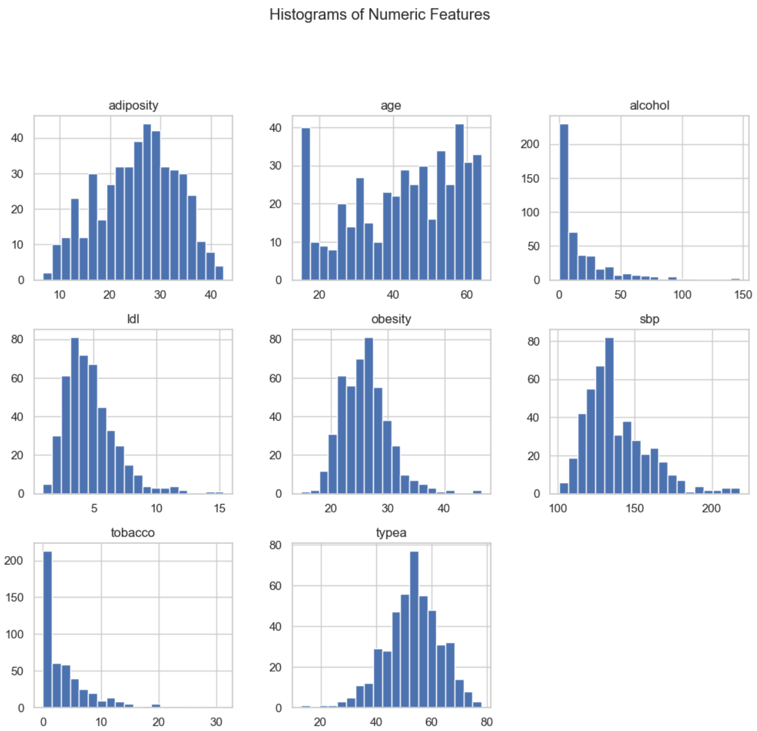

Figure 3: Empirical distribution of numeric features, represented with histograms, suggests a non-uniformally distributed feature space.

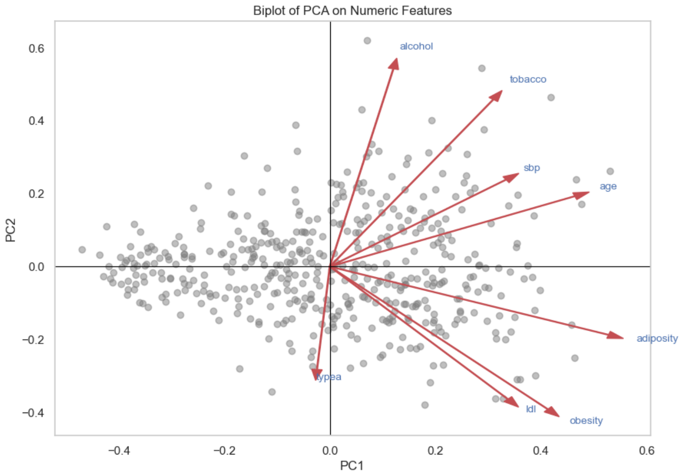

Figure 4: Biplot of Principal Component Analysis (PCA) on numeric features with arrows representing the contribution of each variable to PC1 and PC2. The lack of clear separation in the first two principal components suggests distances between points may not strongly correspond to class distinctions (and so kNN may struggle).

which naturally bounds the predicted probabilities between 0 and 1. For our binary classification task, if $P < 0.5$, we say a person does *not have CHD*, and if $P > 0.5$, we say a person *does have CHD*. To prevent overfitting the training data, and therefore reducing the generalisability of our model, a *penalty* is introduced which shrinks $\hat{\beta}$ to trade off an increase in bias for reduced variance. The *ridge* penalty, (*L2 regularisation*), adds a term proportional to the sum of squared coefficients, $\lambda \sum_{i=1}^{p} \beta_i^2$, to the *loss function*; where the magnitude of λ is determined through cross-validation.[4]

Logistic Regression (L2)'s ROC-AUC score of **0.82** and **62%** F1-Score (Table 2) indicate relatively strong performance in predicting CHD.[5]

4 Performance Evaluation

Model	F1-Score	ROC-AUC	Precision	Recall	Decision Boundary Type
LDA	66	0.81	63	69	Linear (Best Overall)
Logistic (L2)	62	0.82	61	63	Linear (Comparable)
SVM	62	0.80	61	63	Linear
AdaBoost	58	0.79	70	50	Nonlinear (Slightly Worse)
Gradient Boost	56	0.76	56	56	Nonlinear (Poorer)
Decision Tree	46	0.59	40	53	Nonlinear (Poor)
kNN (CV)	46	0.71	52	41	Nonlinear (Poor)

Table 2: Comparison of Key Classifier Results. F1-Score, Precision and Recall are reported in %.

To understand the effectiveness of including a ridge penalty, we first compared alternative logistic regressions; with no penalty, and with Lasso (L1). Figure 5 shows these ROC-AUC curves layer almost perfectly, suggesting there is minimal risk of overfitting; the dataset has relatively low noise and some predictive signal.

LDA, **our best model**, resulted in a similar ROC-AUC score, but with slightly better precision, **63%**, and notably better recall, **69%**, and a Support Vector Machine (SVM) with a linear kernel performed on par with logistic regression. Non-linear models, including Decision Tree, AdaBoost, Gradient Boost and k-Nearest Neighbours where k is selected through cross validation (kNN-CV) suffered poorer performance across all metrics, suggesting that a non-linear decision boundary fails to capture the feature space. That is, **the decision boundary is linear.** AdaBoost is a clear outlier, with high precision, **70%**, and lower recall **50%**; likely caused by a bias towards the majority

[4] Adding a penalty term leads to the optimisation problem: $\min_{\beta} [-\text{Loss}(\beta) + \text{Penalty}(\beta)]$. Ridge performs strongly in the presence of multicollinearity in X, stabilising $\hat{\beta}$ without forcing any of them to be exactly zero (Hastie et al. 2009). In contrast, *Lasso*, *L1 regularisation*, allows shrinking to zero (i.e. variable selection).

[5] Often, medical applications may favour high recall, to capture as many potential cases for initial screening, over precision, with highly precise secondary tests (e.g., ECG). However, given the high prevalence rate in the Western Cape region of South Africa (Crowley et al. 2024), a balanced approach, through F1-score, is prioritised.

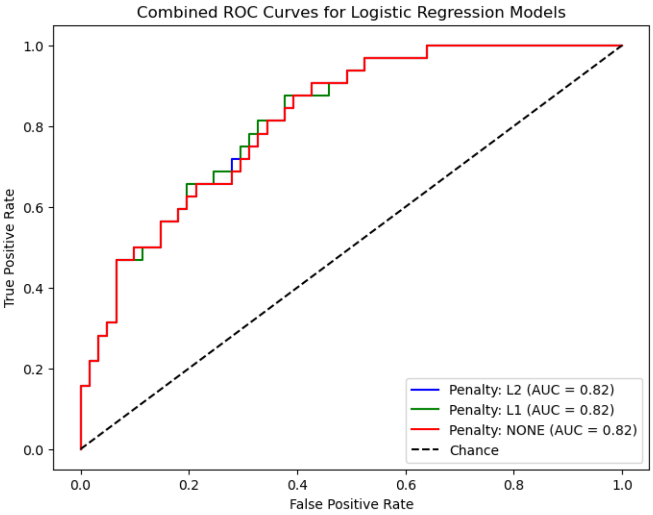

Figure 5: AUC of ROC Curves for Logistic Regression with: i) No Penalty, ii) Ridge Penalty, and iii) Lasso Penalty

class and aggressively filtering out uncertain cases.[6] Meanwhile, kNN lagged significantly behind, reinforcing that clustering-based approaches are not well-suited for this analysis.

A sensitivity analysis, *sans alcohol* and *sans all outliers*, is performed to determine whether these values contribute predictive value or merely add noise.[7]

Sans Alcohol model performance does increase marginally, suggesting alcohol may be a redundant feature. Surprisingly, the Naïve Bayes classifier achieves the highest F1-Score of **66%**; alcohol may be masking some underlying conditional dependency in the dataset. However, Naive Bayes does suffer lower precision than both other models in this pipeline and in the Full-Feature analysis. Linear models such as Logistic Regression with Ridge Penalty and SVM performed similarly to the full-feature pipeline, indicating that alcohol does not add significant noise to the dataset.

Sans Outliers performance is a very different story. All models perform worse overall. With an F1-score of **59%**, Quadratic Discriminant Analysis (QDA) performed the best, perhaps suggesting that the linear decision boundary observed in our main analysis is driven by outliers. Further analysis may consider optimal outlier thresholds to enhance diagnostic screening for men in the Western Cape.

5 Conclusion

CHD is among the 10 largest killers of South African men (WHO 2025). This report has shown that it is possible to achieve a good balance of precision and recall through **LDA**. Further analysis may explore whether classification performance remains robust in other regions across the world or across lower risk populations.

[6] Further analysis may apply *SMOTE* to address class imbalances.

[7] Outliers are removed using an Inter-Quartile Range method. See Appendix C).

References

British Heart Fondation (2025), 'High blood pressure (Hypertension)', https://www.bhf.org.uk/informationsupport/risk-factors/high-blood-pressure. Accessed: 16 March 2025.

Crowley, T., Francis, R., Ismail, T., Hoffman, J. et al. (2024), 'Cardiovascular risk among community members in three communities in the Cape Metropole of the Western Cape', *African Journal of Primary Health Care & Family Medicine* **16**(1), 4246.

Drinkaware (2021), 'Alcohol and the heart', https://www.drinkaware.co.uk/facts/health-effects-of-alcohol/alcohol-related-diseases-and-illnesses/alcohol-and-the-heart. Accessed: 16 March 2025.

Hastie, T., Tibshirani, R. & Friedman, J. (2009), *The Elements of Statistical Learning*, Springer Series in Statistics, Springer, New York, NY.

WHO (2025), 'South Africa [Country Overview]', https://data.who.int/countries/710. Accessed: 16 March 2025.

A Confusion Matrices and Performance Metrics

A.1 Confusion Matrix

Confusion matrices summarise classifier performance into a neat, easy-to-read table. For a binary classification problem, as in our case, it is usually represented as:

	Predicted	
Actual	Negative	Positive
Negative	TN	FP
Positive	FN	TP

Table 3: Confusion Matrix

where:

- **TP (True Positives)**: Correctly predicted positive cases.
- **FP (False Positives)**: Incorrectly predicted positive cases.
- **TN (True Negatives)**: Correctly predicted negative cases.
- **FN (False Negatives)**: Incorrectly predicted negative cases.

A.2 Precision

Precision quantifies the accuracy of positive predictions:

$$\text{Precision} = \frac{TP}{TP + FP}$$

A.3 Recall

Recall, also referred to as sensitivity, measures the proportion of actual positives that are correctly identified:

$$\text{Recall} = \frac{TP}{TP + FN}$$

A.4 F1-Score

The F1-score is the harmonic mean of precision and recall, offering a balance between the two:

$$F_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

A.5 ROC-AUC

The Receiver Operating Characteristic (ROC) curve plots recall (true positive rate) against the false positive rate. The area under this curve quantifies the ability of a model to correctly classify, in our case, *CHD-positive* and *CHD-negative* classes. $AUC = 1$ represents perfect discrimination, whereas $AUC = 0.5$ suggests no discriminative power, equivalent to 'chance' (i.e. a random guess).

B Pairplots

Pairplots of the numeric features in the heart disease dataset clearly visualise that the target cannot be separated in 2D space. This suggests LDA may perform reasonably well as performance is generally improved when class distributions overlap reasonably.

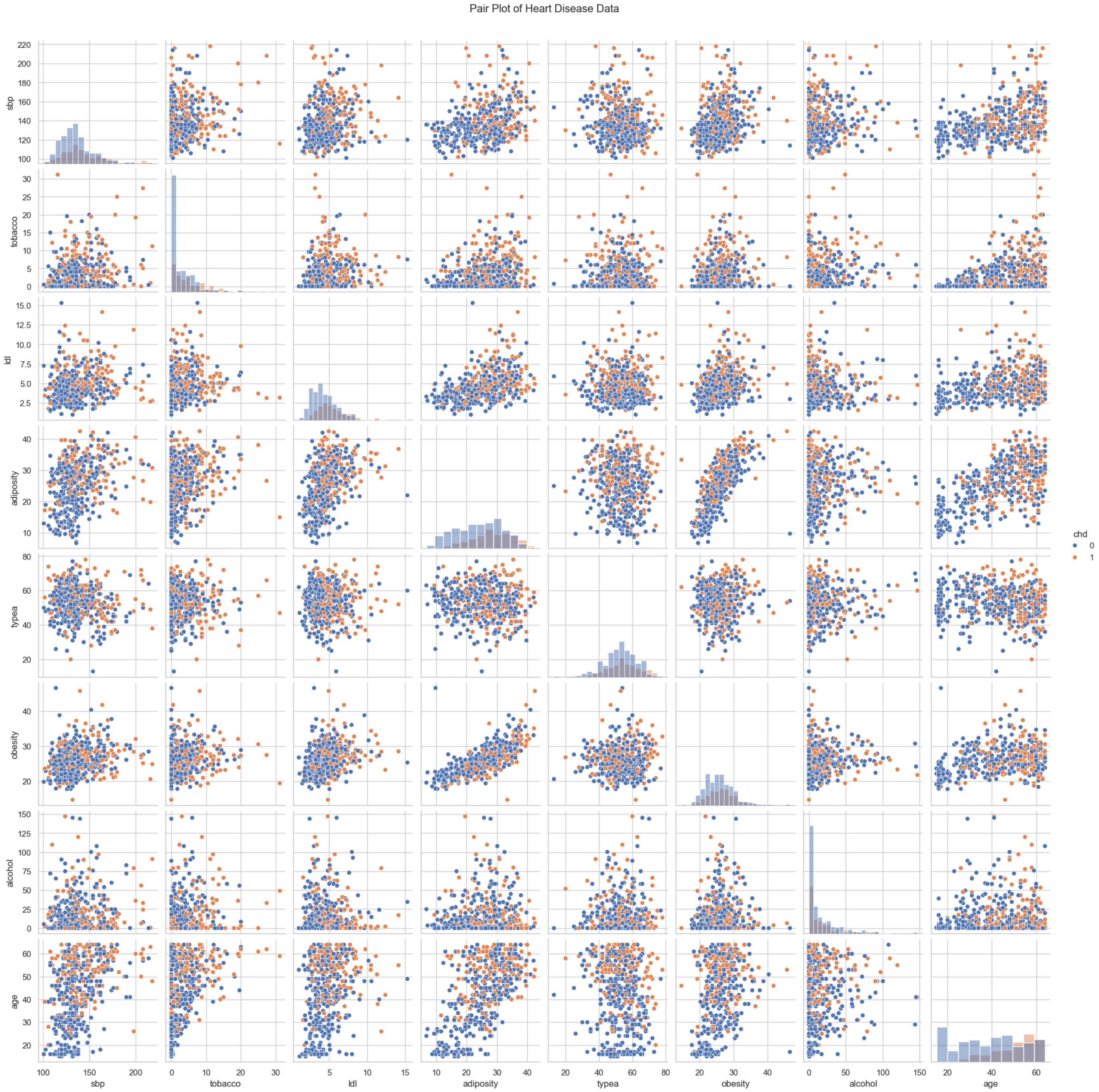

Figure 6: Ovelapping data across all pairplots suggest no two features can differentiate the target variable

C Sensitivity Analysis

To perform a sensitivity analysis, model performance is assessed under two further pipelines: i) *without alcohol* in the feature space, and ii) *removing all outliers* to determine whether these contribute predictive value or merely add noise.

Scenario	**Model**	**F1-Score (%)**	**ROC-AUC**	**Precision (%)**	**Recall (%)**
Sans Alcohol	Naïve Bayes	66	0.77	59	75
	Logistic (L2)	65	0.82	64	66
	SVM	65	0.81	64	66
Sans Outliers	QDA	59	0.74	56	63
	Decision Tree	55	0.73	57	54
	Naïve Bayes	54	0.77	54	54

Table 4: Sensitivity Analysis: Top 3 Models Under Each Scenario

C.1 IQR Method

Figure 2 identified the features with many outliers: *alcohol*, *ldl*, *obesity*, *tobacco*, and *typea*. We set a lower bound and upper bound:

$$\begin{aligned} \text{Lower Bound} &= Q_1 - 1.5 \times \text{IQR}, \\ \text{Upper Bound} &= Q_3 + 1.5 \times \text{IQR} \end{aligned}$$

where Q_1 and Q_3 are the first and third quartiles, and $\text{IQR} = Q_3 - Q_1$. Data beyond these bounds are removed.

D Tuned Hyperparameters

Hyperparamter tuning uses `GridSearchCV` with 10-fold cross-validation. The optimal values for each model is selected via the highest cross-validated ROC-AUC. The tables below show the parameter grids used for hyperparameter selection.

Summary Tables

Model	**Hyperparameter**	**Candidate Values**
Logistic Regression (L2)	C	$\{0.001, 0.01, 0.1, 1, 10, 100\}$
Logistic Regression (L1)	C	$\{0.001, 0.01, 0.1, 1, 10, 100\}$
Logistic Regression (No Penalty)	Penalty	`None`
Support Vector Machine (Linear Kernel)	C	$\{0.001, 0.01, 0.1, 1, 10, 100\}$

Table 5: Tuned Hyperparameters for Linear Models

Model	**Hyperparameter**	**Candidate Values**
Decision Tree	`max_depth`	$\{3, 5, 7, 10\}$
	`min_samples_split`	$\{2, 5, 10, 20\}$
	`ccp_alpha`	$\{0.0, 0.001, 0.01, 0.1\}$
Random Forest	`n_estimators`	$\{50, 100, 200\}$
	`max_depth`	$\{3, 5, 7, 10\}$
	`min_samples_split`	$\{2, 5, 10\}$
	`ccp_alpha`	$\{0.0, 0.001, 0.01, 0.1\}$
	`max_features`	{`'sqrt'`, `'log2'`}
AdaBoost	`n_estimators`	$\{50, 100, 200\}$
	`learning_rate`	$\{0.01, 0.1, 1, 10\}$
Gradient Boosting	`n_estimators`	$\{50, 100, 200\}$
	`learning_rate`	$\{0.01, 0.1, 0.2, 1\}$
	`max_depth`	$\{3, 5, 7, 10\}$

Table 6: Tuned Hyperparameters for Tree-Based and Ensemble Models

Hyperparameter	**Candidate Values**
k (number of neighbours)	$\{1, 2, \ldots, 20\}$

Table 7: Tuned Hyperparameters for k-Nearest Neighbours (kNN)

Explanations

Cost Complexity Pruning (α): For tree-based models, `ccp_alpha` controls cost complexity pruning. Increasing α reduces tree complexity (i.e., prunes larger trees), and reduces the risk of overfitting.

Regularisation in Logistic Regression: C is the inverse of the regularisation strength. Lower values of C increase regularisation to reduce the risk of overfitting.
SVM Regularisation: Similar to logistic regression, C in SVM controls the trade-off between smooth decision boundaries and correct classification of training data.

Decision Tree Parameters

`max_depth:` Limits the maximum depth of the tree, in turn, reducing how detailed the tree can become.
`min_samples_split:` Minimum number of samples required to split an internal node.

Random Forest Parameters

`n_estimators:` The number of trees in random forest. More trees improve performance but increase computational cost.
`max_features:` Number of features observed when looking for the best split.

Ensemble Model Parameters (AdaBoost and Gradient Boosting)

`learning_rate:` Controls contribution of each tree to final model. Lower values slow the learning process.
`n_estimators:` Sets the number of boosting iterations, i.e. trees, in the model.

k-Nearest Neighbours (kNN)

`k (number of neighbours):` Sets the number of closest data points used in the majority vote to classify a new observation.