The aim of this project is to predict Coronary Heart Disease (CHD) in males within a high-risk area of the Western Cape, South Africa. The No Free Lunch theorem posits that no single algorithm can deliver optimal results in every scenario. Therefore, we employ a strategy that utilizes supervised machine learning algorithms. By methodically applying an array of chosen classifier algorithms to our dataset, we identify the algorithm that yields the best performance metrics.

1. EXPLORATORY DATA ANALYSIS

Table 1 presents the average characteristics of the dataset across various features. The analysis reveals several key insights regarding individuals with CHD. Firstly, those with CHD typically exhibit higher Systolic Blood Pressure (SBP). Additionally, tobacco consumption is notably higher among individuals diagnosed with CHD. Elevated concentrations of Low Density Lipoprotein (LDL) cholesterol, often referred to as "bad" cholesterol, are also observed in this group. Furthermore, adiposity is more marked in individuals with CHD and age, appears to be a critical element, with older individuals being more likely to have CHD.

	Overall Mean	Mean (Positive CHD)	Mean (Negative CHD)
sbp	138.33	143.74	135.46
tobacco	3.64	5.52	2.63
ldl	4.74	5.49	4.34
adiposity	25.41	28.12	23.97
typea	53.1	54.49	52.37
obesity	26.04	26.62	25.74
alcohol	17.04	19.15	15.93
age	42.82	50.29	38.85

Table 1: Comparison of mean features values for patients by CHD status

Additionally, we investigated the distribution of observations within the variable "family history" in relation to both negative and positive CHD events, as shown in Table 2. Our analysis revealed that among individuals without a family history of heart disease, there were 206 instances of negative CHD events and 64 instances of positive CHD events. Conversely, for those with a family history of heart disease, both negative and positive CHD events were observed 96 times each.

Family history	*Yes*	*No*
Absent	206	64
Present	96	96

Table 2: Observations counts for family history variable by CHD status

The box plots in Figure 1 depict the distributions of various factors in relation to CHD presence. It's evident that age, adiposity and sbp exhibit significant median differences between CHD-positive and -negative groups, suggesting these factors may have a stronger correlation with CHD. LDL, Type A and obesity also demonstrate disparities, but to a lesser extent. Alcohol and

tobacco consumption patterns show considerable variability and a right skew in both groups, with alcohol consumption outliers indicating a wide dispersion in higher consumption levels. Notably, the tobacco plot highlights significant variance in usage among CHD-positive individuals, underscored by the presence of outliers, pointing to diverse consumption patterns. We also evaluated the Pearson's Correlation Coefficient to assess the correlation between features. Although some features exhibited moderate to strong correlations, we chose to retain all features, believing they are critical for accurately predicting CHD.

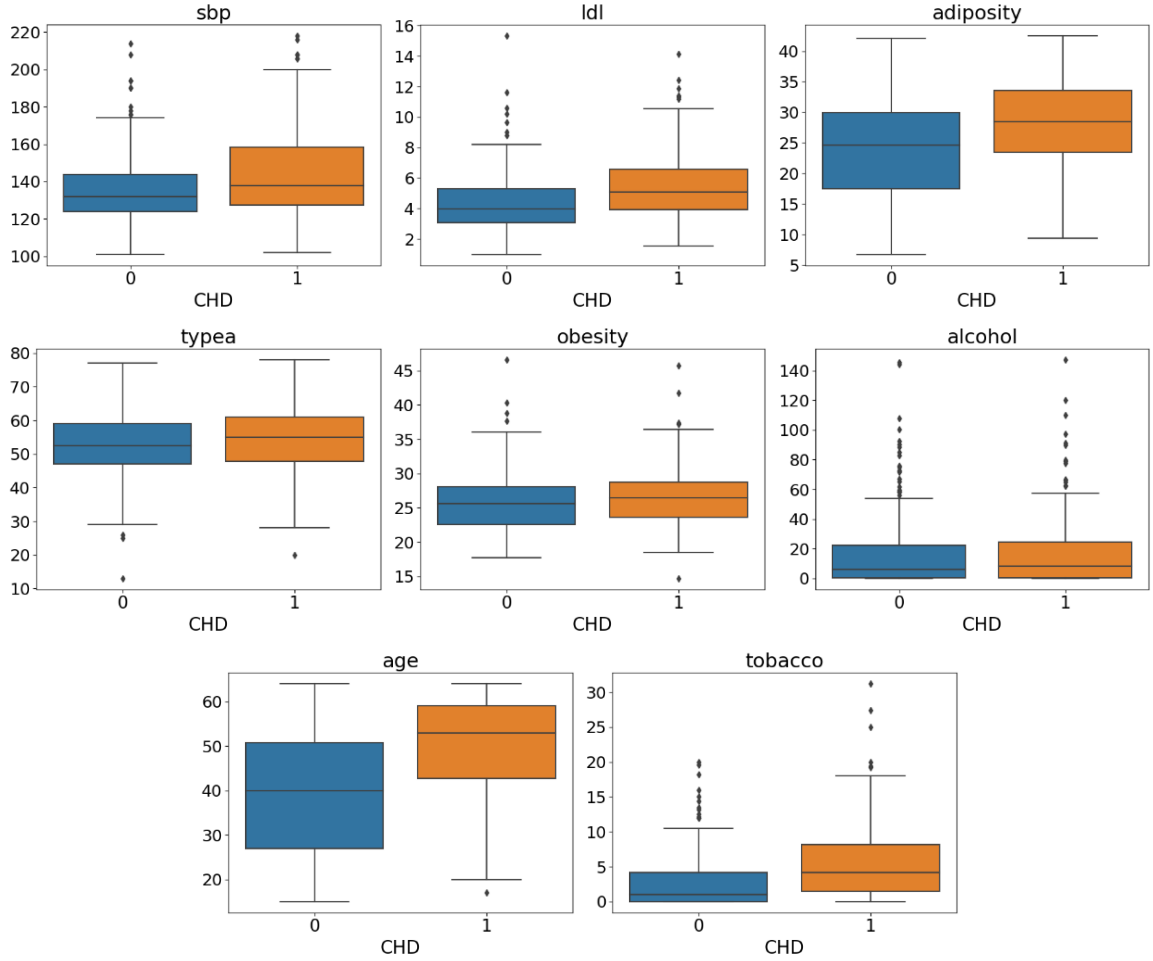

Figure 1: Distribution of health indicators by CHD status

2. LOGISTIC RIDGE REGRESSION (LRR)

In addressing our binary classification task, we evaluated six primary classifiers for their fit and effectiveness: LRR, Support Vector Machine (SVM), K-Nearest Neighbor (K-NN), Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA) and Gaussian Naïve Bayes (GaussianNB). Our choice was informed by the dataset's characteristics, which approximately align with the statistical needs of these models despite challenges, including right-skewed predictors and varying variance-covariance structures. The suitability of LDA and QDA is supported by the near-normal distribution of predictors within classes. SVM's adaptability, especially with non-linear relationships via the kernel trick, and its efficiency in

high-dimensional spaces, make it a strong contender. The applicability of GaussianNB is supported by the assumption of independent observations, as the features in the provided dataset—covering aspects of a person's physical habits and medical history—are assumed to be independent of one another, aligning with the fundamental assumption of the Naïve Bayes classifier. We employed ten-fold stratified cross-validation for robust training and evaluation, considering our dataset's moderate class imbalance—65.37% CHD negative and 34.63% CHD positive.

After training and testing the models, we evaluated their performance by calculating the average score across multiple metrics, including accuracy, F1 score, error rates, precision, recall and ROC AUC. Each metric provides insights into different aspects of the model's performance, from overall correctness (accuracy) to the balance between precision and recall (F1 score), the model's error tendencies (error rates), its capability to identify positive instances (precision and recall) and its performance across various threshold settings (ROC AUC).

For LRR, the optimal parameter C was identified using GridSearchCV, resulting in a value of 0.39. The regularisation approach, while slightly increasing bias, significantly reduces variance, improving model generalisability and mitigating overfitting risks. Table 3 reports the results of the LRR. Each coefficient represents the log odds change in the likelihood of the outcome variable (presence of CHD) for a one-unit increase in the predictor variable, assuming all other variables are held constant. For example, for ldl, a one-unit increase in this variable is associated with an increase in the log odds of CHD by approximately 0.43, which is statistically significant ($p < 0.05$). The constant represents the log odds of the outcome when all predictor variables are zero. The value -0.89 suggests that when all other variables are held at zero, the log odds of CHD are negative, indicating a lower likelihood of CHD occurrence under these conditions.

	Coef	Std Err	P>|z|
sbp	0.2	0.14	0.2
tobacco	0.28	0.14	>0.05
ldl	0.43	0.15	<0.05
adiposity'	0.2	0.27	0.5
typea	0.37	0.14	<0.05
obesity	-0.28	0.22	0.2
alcohol	0.02	0.14	0.9
age	0.61	0.21	<0.05
famhist_Present	0.37	0.14	<0.05
const	-0.89	0.147	0

Table 3: LRR results

3. ANALYSING OUTCOMES AND SELECTING THE OPTIMAL MODEL

Figure 2 depicts a set of bar plots for each performance metric to facilitate a straightforward comparison of each classifier's performance across different metrics. According to the graph, no single model consistently outperforms the others across all metrics, reflecting the principle of the No Free Lunch premise, which asserts that no one model is universally best for all problems. Among the 6 prediction techniques applied, LRR appears to be a strong contender for predicting CHD in this high-risk population. Its highest accuracy, low error rates, and leading precision make it particularly effective.

However, in a medical context, especially for diagnosing conditions like CHD, both high sensitivity (recall) and the ability to accurately distinguish between positive and negative cases (as measured by ROC-AUC) are crucial. High sensitivity is vital to prevent missing individuals who have the disease. ROC-AUC reflects a model's effectiveness in maintaining a balance between identifying true positives and avoiding false positives. Given the emphasis on these two metrics, NB excels in sensitivity, showing the highest F1 score and recall, ensuring that fewer cases of CHD go undiagnosed. The SVM classifier stands out in terms of ROC-AUC, indicating strong discriminatory power.

Given these considerations, we want to ensure that as many true cases of CHD as possible are identified (even at the risk of some false positives), therefore NB emerges as the most suitable choice due to its superior recall. This approach aligns with the principle of maximising patient safety by prioritising the detection of all potentially affected individuals.

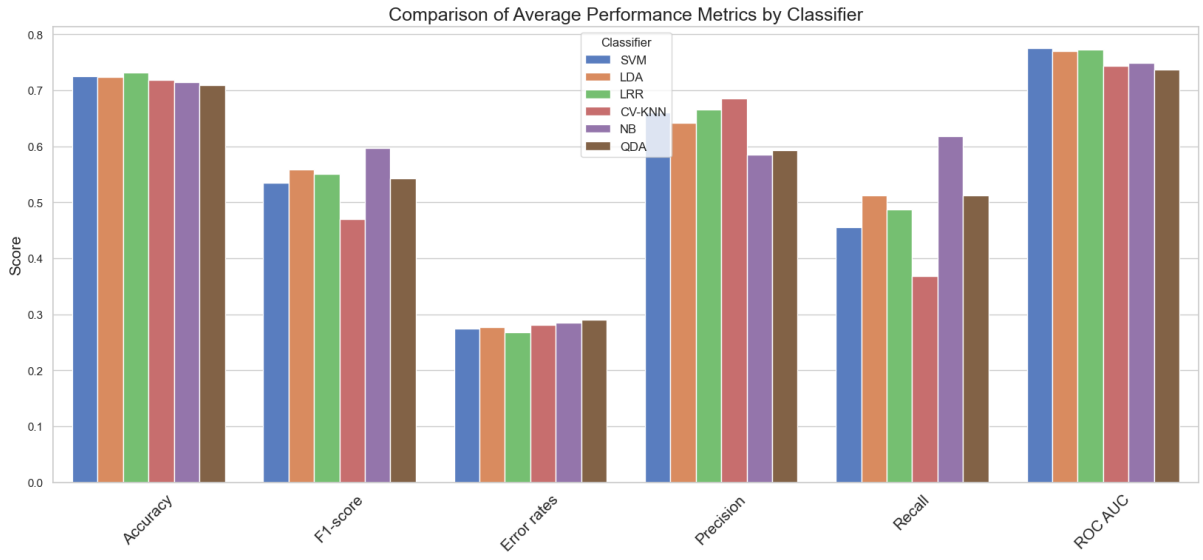

Figure 2: Comparison of average performance metrics by classifier