Individual Coursework Submission Form

Specialist Masters Programme

Surname: FARINA ALCEDO	**First Name:** BLANCA
MSc in: BUSINESS ANALYTICS	**Student ID number:** 240030671
Module Code: SMM636	
Module Title: MACHINE LEARNING	
Lecturer: DR RUI ZHU	**Submission Date:** 24/03/2025

Declaration:

By submitting this work, I declare that this work is entirely my own except those parts duly identified and referenced in my submission. It complies with any specified word limits and the requirements and regulations detailed in the coursework instructions and any other relevant programme and module documentation. In submitting this work, I acknowledge that I have read and understood the regulations and code regarding academic misconduct, including that relating to plagiarism, as specified in the Programme Handbook. I also acknowledge that this work will be subject to a variety of checks for academic misconduct.

We acknowledge that work submitted late without a granted extension will be subject to penalties, as outlined in the Programme Handbook. Penalties will be applied for a maximum of five days lateness, after which a mark of zero will be awarded.

Marker's Comments (if not being marked on-line):

Deduction for Late Submission:

Final Mark: %

MACHINE LEARNING INDIVIDUAL ASSIGNMENT

SMM636

BLANCA FARINA ALCEDO

1. Exploratory Data Analysis of the Dataset

The aim of this report is to predict the coronary heart disease (chd: 1/0) for males in a heart-disease high-risk region of the Western Cape, South Africa and present the results.

The dataset contains 462 observations with numerical health metrics and a binary family history variable, along with the CHD target. As exhibited in Appendix 1, most features are approximately normally distributed with slight right skewness, while tobacco and alcohol consumption show extreme skewness. Age exhibits a bimodal distribution with peaks around 20 and 55-60 years. Outliers are present in systolic blood pressure, tobacco, and alcohol.

Moreover, as exhibited in Appendix 2, strong correlations exist between adiposity and obesity (0.72), age and adiposity (0.63), and tobacco and alcohol consumption (0.45). Key predictors for CHD include age (0.37), tobacco use (0.30), family history (0.27), ldl (0.26), and adiposity (0.25), while type-A behavior and obesity have weaker correlations (0.10 each). Alcohol consumption shows a negligible correlation (0.06). Family history is relatively balanced, with 42% having a positive history of as shown in Appendix 3.

2. Logistic Regression with Ridge Penalty Results

A Logistic Regression model with Ridge penalty (L2) was used to classify patients. The model achieved a test accuracy of 72.04%, meaning it correctly classified 72.04% of the test samples. Before fitting the model, numerical features were standardized using StandardScaler, and the class imbalance was addressed by setting class_weight='balanced'.

The confusion matrix revealed that the model correctly predicted 44 cases of No CHD (0) when the person did not have it (True Negatives) and 23 cases of CHD (1) when the person actually had it (True Positives). However, it also incorrectly predicted 15 instances of CHD (1) for patients who did not have CHD (False Positives) and misclassified 11 patients with CHD as not having the disease (False Negatives). False Negatives are critical in medicine as they can lead to missed diagnoses and delayed treatment.

The classification report shown in Table 1, provides deeper insight into model performance using precision, recall, and F1-score. Precision shows how for Class 0, 80% of the predictions were correct, whereas for Class 1, only 61% of the predictions were correct, indicating a higher rate of false positives for this class. Further, the recall for Class 0 was 75%, meaning 75% of actual "No CHD" cases were correctly identified. For Class 1, recall was 68%, meaning 68%

of actual CHD cases were successfully detected. The F1-score, was lower for Class 1, suggesting that even with class weighting, the model still struggles to detect CHD cases with high confidence.The macro average F1-score of 0.71 gives equal importance to both classes, while the weighted average of 0.72 adjusts for class imbalance by assigning more weight to Class 0. The similarity between these values suggests that, although the model slightly favors Class 0, its overall performance remains balanced.

	Precision	Recall	F1-Score	Support
0	0.80	0.75	0.77	59
1	0.61	0.68	0.64	34
accuracy			0.72	93
macro avg	0.70	0.71	0.71	93
weighted avg	0.73	0.72	0.72	93

Table 1- Logistic Regression L2 Classification Report

Further, the Area Under the Curve in the ROC Curve in Figure 1 suggests that the model correctly ranks positive instances higher than negative ones 81% of the time. This confirms that the model is significantly better than random guessing and effectively separates the two classes.

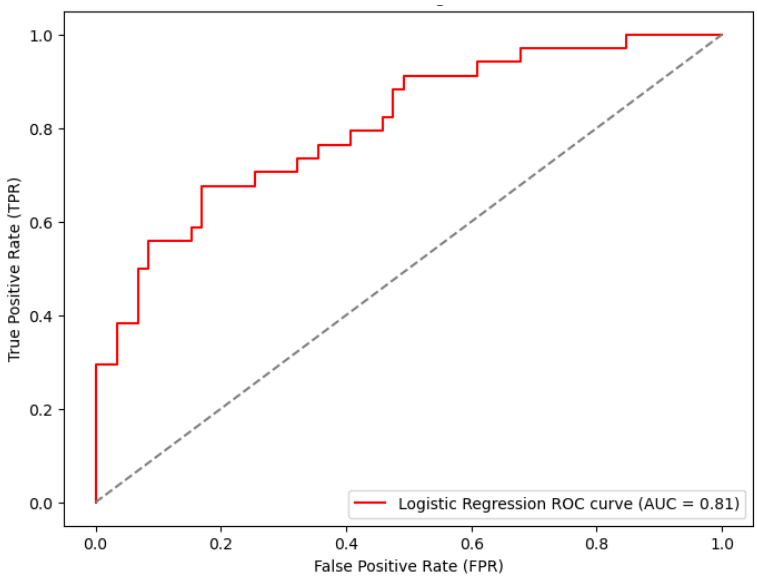


Figure 1- Logistic Regression L2 Model's ROC Curve

3. Exploring Classifiers and Selecting the Best Performing Model

To determine the most effective classifier for predicting CHD, multiple models were explored. Several models were initially considered but ultimately excluded from the final analysis based on their limitations for this specific task. KNN struggles with high-dimensional data and is computationally expensive for predictions. Naïve Bayes assumes feature independence, which is unrealistic given the correlations in this dataset. Decision Trees tend to overfit small or noisy datasets, making them unreliable as standalone models.

The final selection included Random Forest, Gradient Boosting, Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), and Support Vector Machine (SVM), each chosen for its unique strengths in predicting CHD. Random Forest was selected for its ability to handle non-linear relationships and mixed feature types, reducing overfitting with its ensemble nature. Gradient Boosting offers superior predictive performance in structured data by sequentially correcting errors from prior trees, making it particularly effective for imbalanced datasets and providing detailed feature importance measures. Moreover, before fitting LDA, key feature distributions and relationships were examined as presented in Appendix 4 and 5. This revealed differences between CHD and non-CHD cases across health metrics, but with substantial overlap rather than clear separation. Age was the strongest differentiator, with CHD cases generally older. Other risk factors showed subtle shifts towards higher values in CHD cases. Despite lacking clean linear separation and differing covariance structures, LDA was considered for its strengths in dimensionality reduction, interpretability, and robustness to small samples. QDA complements LDA by relaxing the assumption of equal covariance matrices, improving flexibility in modeling. Lastly, SVM excels in moderate-dimensional data and handles outliers effectively. Its ability to find optimal decision boundaries makes it valuable for medical classification tasks with complex feature spaces. This diverse set of models balances interpretability, robustness, and predictive accuracy in tackling CHD prediction.

Once the models were chosen training and evaluation was the next step. Class balancing techniques and scaling were applied where necessary to ensure fair evaluation. Hyperparameter tuning was conducted for models that required optimization. Each classifier was trained using cross-validation to ensure robustness and mitigate overfitting. Performance was assessed using a confusion matrix, classification report, and AUC scores. The top-performing models based on test accuracy were LDA, SVM, QDA, and Random Forest, all achieving 75.27% accuracy.

While multiple models achieved the same accuracy, a deeper analysis revealed significant differences in their ability to generalize and rank predictions effectively as presented in Table 2.

Model	Cross-Validation Accuracy	Test Accuracy	Precission (Class 1)	Recall (Class 1)	F1-score (Class 1)	AUC
Logistic Regression L2	0.6910	0.7204	0.61	0.68	0.64	0.81
Gradient Boosting	0.6804	0.7312	0.62	0.71	0.66	0.79
Random Forest	0.6935	0.7527	0.68	0.62	0.65	0.54
LDA	0.7207	0.7527	0.70	0.56	0.62	0.81
QDA	0.7047	0.7527	0.72	0.53	0.61	0.77
SVM	0.7181	0.7527	0.69	0.59	0.63	0.80

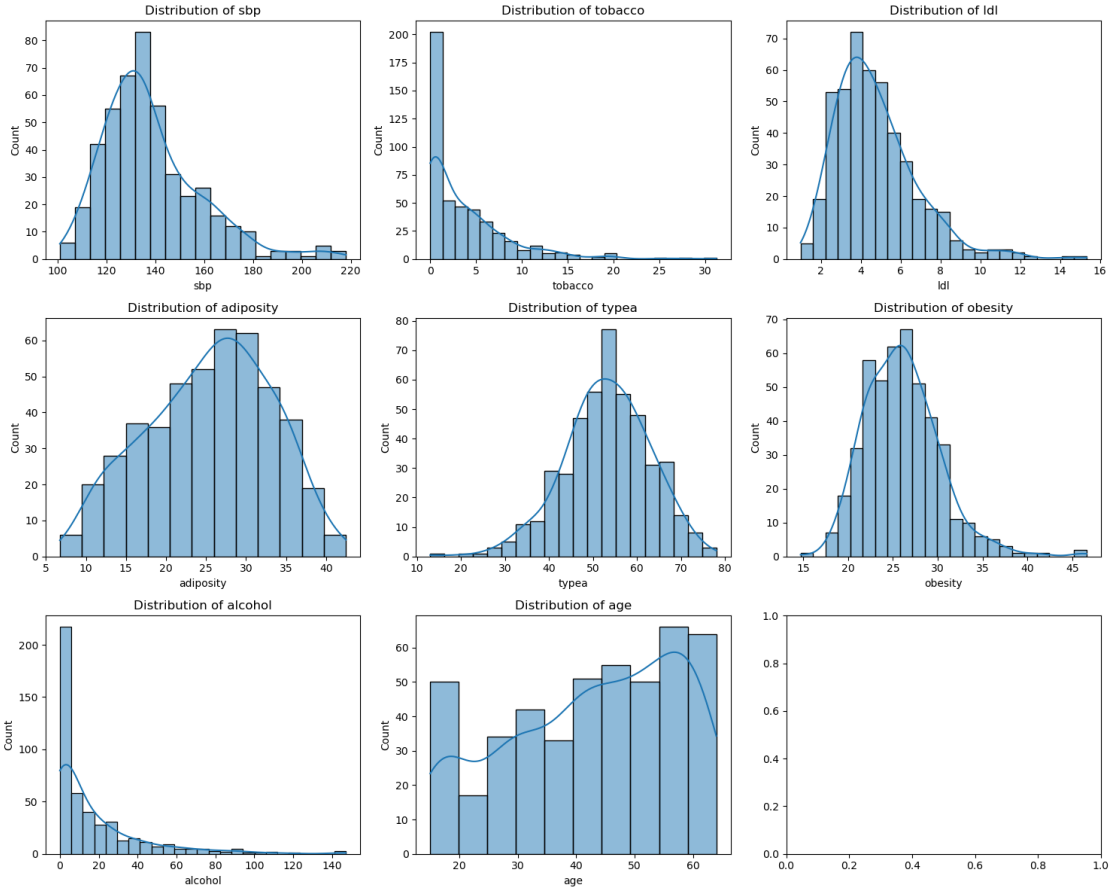
Table 2- Models Performance Results

Despite high accuracy, Random Forest had a very poor AUC (54%), indicating weak ranking capability. QDA had a relatively low F1-score (61%) and AUC (77%), making it less desirable for this classification task. Gradient Boosting had the highest F1-score (66%) but a slightly lower accuracy (73%), suggesting it performed well in recall but did not maintain overall classification balance. LDA and Logistic Regression (L2) had the highest AUC (81%), indicating strong discriminative power. However, LDA had slightly lower recall, which is crucial in identifying CHD cases.

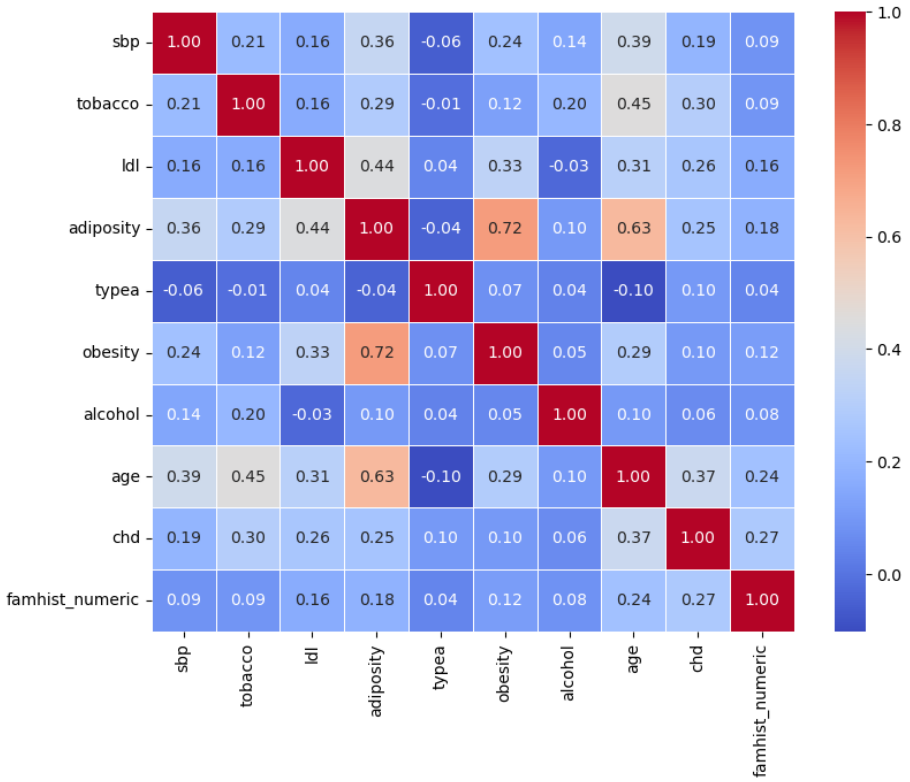
Among all models, SVM emerged as the strongest contender due to its well-balanced performance across all key metrics. It achieved high accuracy (75%), a strong F1-score (63%), and an AUC of 80%, demonstrating both stability and predictive power. While SVM is more complex than simpler models like LDA, its robustness makes it an excellent choice when predictive accuracy is the priority. Therefore, given its robustness, consistency, and overall performance, SVM is the clear choice for this classification task.

Appendices

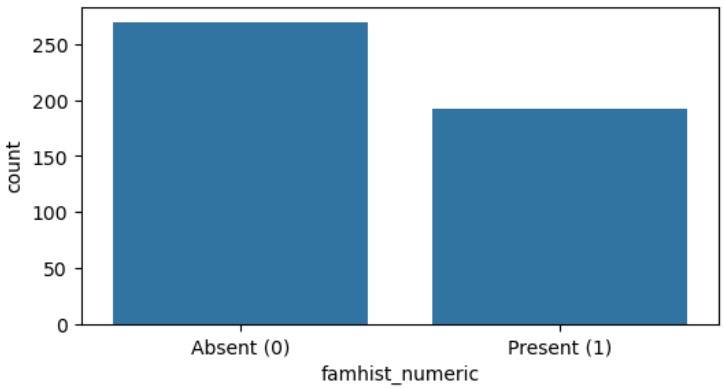
Appendix 1: Distribution of Numerical Features in Dataset



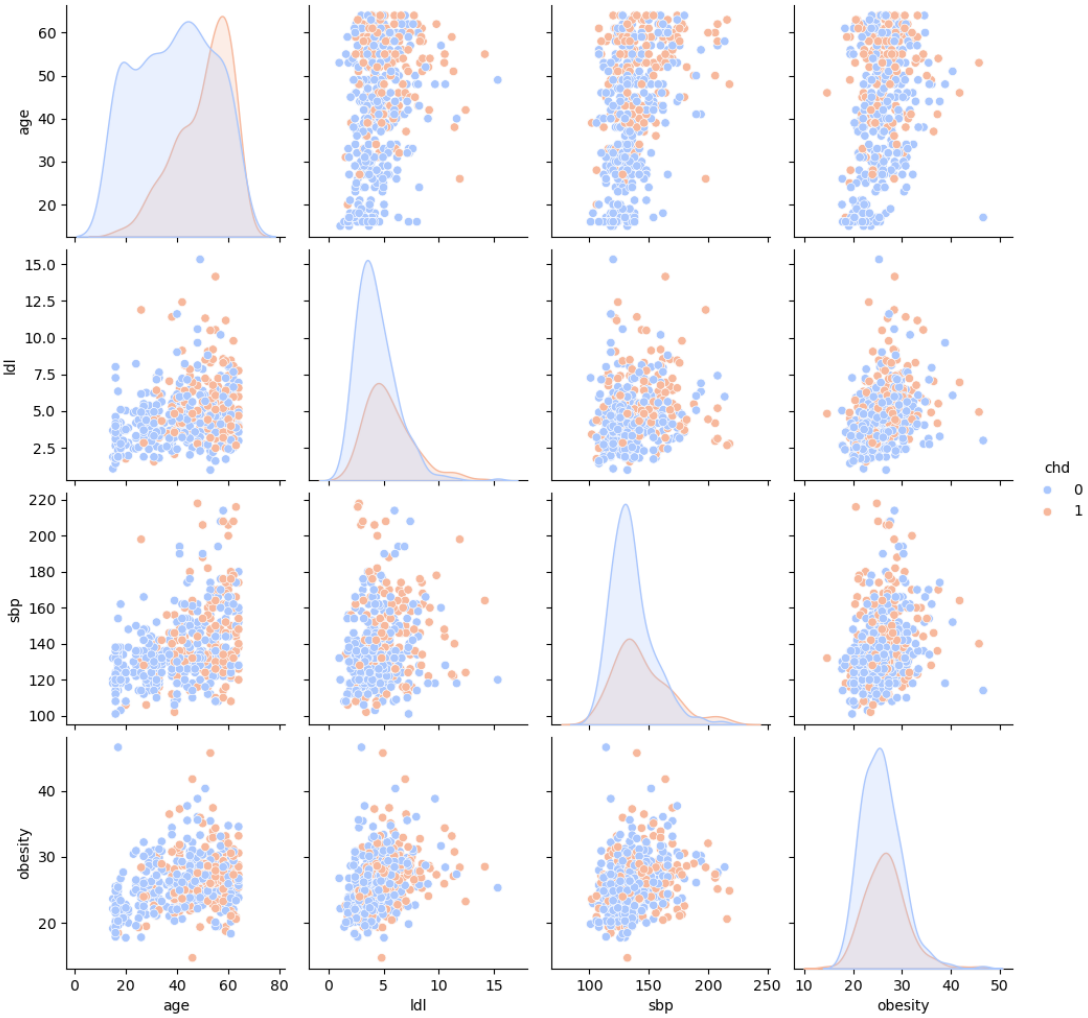
Appendix 2: Correlation Matrix of Features



Appendix 3: Distribution of Family History variable



Appendix 4: Visualisation of Relationship Between Key Features



Appendix 5: Viasualisation of Feature Distribution by Class

Note that famhist_numeric is binary