Individual Coursework Submission Form

Specialist Masters Programme

Surname: Karaman	**First Name: Noor**
MSc in: Business Analytics	**Student ID number: 190014984**
Module Code: SMM636	
Module Title: Machine Learning	
Lecturer: Dr Rui Zhu	**Submission Date: 24.03.2024**

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:** **%**

Coronary Heart Disease Prediction in High-Risk Males: Classification Using Ridge-Penalised Logistic Regression and Comparative Analysis

Noor Karaman (190014984)

1. Introduction

This report aims to present an analysis of prediction of coronary heart disease (CHD) in South African males using nine clinical and behavioral features. The primary objective is to to develop an accurate CHD prediction model for early intervention and improved health outcomes and includes an exploratory data analysis (EDA), a logistic regression model with ridge penalty, and evaluating other machine learning classifiers.

2. Exploratory Data Analysis (EDA)

2.1 Dataset Overview

The dataset contains 462 observations with 9 predictors (sbp, tobacco, ldl, adiposity, famhist, typea, obesity, alcohol, age) and a binary target variable (chd). The summary statistics (appendix 1) shows approximately 34.6% of the subjects have CHD, indicating a somewhat imbalanced dataset.

2.2 Key Findings from EDA

2.21 Feature Distributions Analysis

The boxplots (appendix 2) reveal several important patterns in how features are distributed between patients with and without CHD. Age showed the clearest distinction between groups, with CHD patients being significantly older (median ~52-55 years) than non-CHD patients (median ~40 years). CHD prevalence increases substantially in subjects over 50. Patients with CHD also showed higher values for sbp, LDL, and adiposity. Family history was an important risk factor, with higher CHD rates among those with family history of heart disease. Obesity showed slight differences between groups, while alcohol consumption and type-A behavior patterns showed minimal differences. Tobacco consumption, the distribution shows that patients with CHD have higher tobacco consumption, with some extreme values above 25kg.

2.22 Correlation analysis

Age (0.37), tobacco (0.30), LDL (0.26), and family history (0.27) showed the strongest correlations with CHD. Several predictors showed correlations with each other (appendix 3): adiposity and obesity (0.72), age and adiposity (0.63), age and tobacco (0.45), and LDL and

adiposity (0.44). Type-A behavior showed weak correlations with other variables, including CHD (0.10).

2.23 Insights for Modelling

Age and family history appear to be crucial risk factors for CHD, aligning with clinical knowledge, (Pencina et al., 2009; Lloyd-Jones et al., 2004). The strong correlation between adiposity and obesity (0.72) suggests potential multicollinearity issues that can impact regression based approaches. Variable importance, based on correlation strengths, age, tobacco, family history, and LDL appear to be promising predictors for CHD classification. No missing values were found, but standardisation was needed for skewed variables (tobacco, alcohol, LDL).

3. Logistic Regression with Ridge Penalty

3.1 Model Tuning and Optimisation

To start, a baseline model was created before optimising (appendix 4). This model controls the trade-off between fitting the training data and keeping model coefficients small to prevent overfitting. A range of C values was explored using 5-fold cross-validation, the grid search revealed that C=0.034 provides the best balance for our dataset.

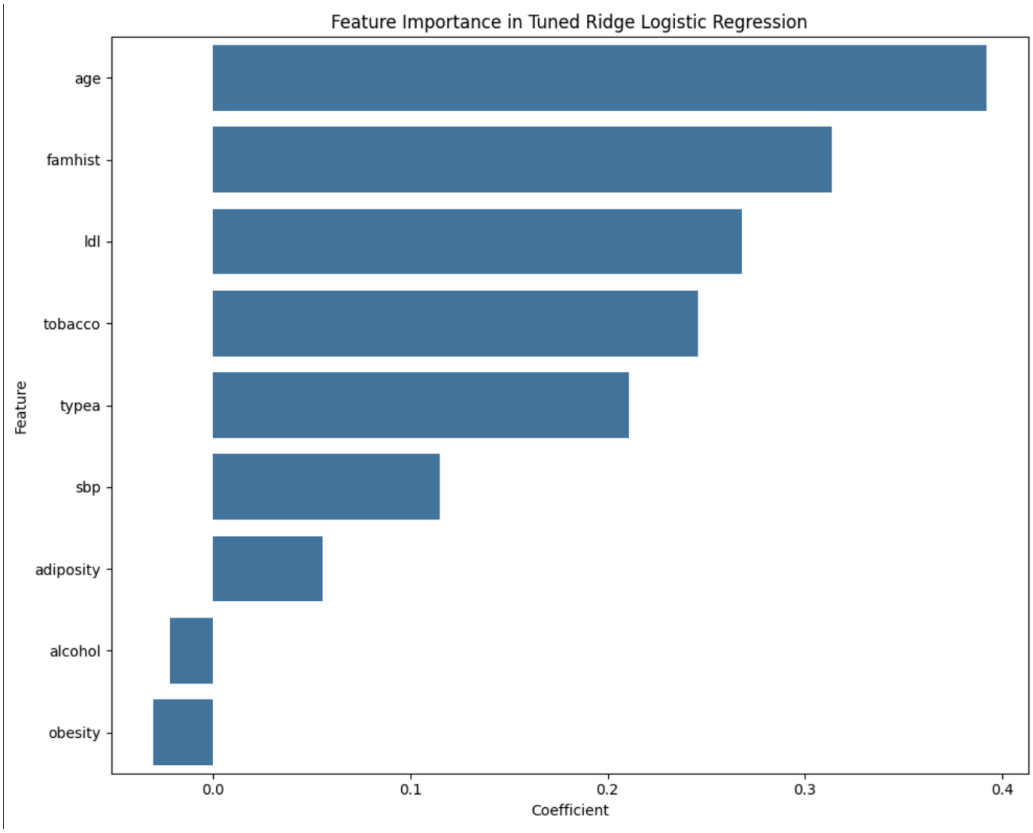
3.2 Model Performance

The model achieved a test accuracy of 74.19%, showing a balanced performance profile, with somewhat stronger precision for identifying non-CHD cases (0.82) and reasonable recall for detecting CHD cases (0.69). The AUC value of 0.824 indicates good discriminative ability, substantially better than random classification (0.5).

3.3 Feature Importance Analysis.

The ridge regression coefficients show the relative importance of different features for predicting CHD. Age (0.392) is the strongest predictor, consistent with established medical knowledge that cardiovascular disease risk increases with age, (Pencina et al., 2009). Family History (0.314), the presence of family history of heart disease is confirmed as a significant risk factor, indicating the genetic component of CHD, (Lloyd-Jones et al., 2004). Elevated levels of low-density lipoprotein (LDL) cholesterol show a substantial positive association with CHD risk (0.268),

aligning with its well-known role in atherosclerosis development, (Ference et al., 2017). Tobacco use emerges as the fourth most influential predictor (0.246), reinforcing its status as a major modifiable risk factor for heart disease, (Banks et al., 2019). Type-A Behavior (0.211), the positive coefficient for Type-A behavior patterns suggests psychological factors may contribute to heart disease risk.



4. Alternative Classification models

To validate model selection, other classification algorithms were explored to improve prediction accuracy and performance. Each classifier was first evaluated with default parameters and then optimised using grid search with 5-fold cross-validation. The models included random forest, gradient boosting, support vector machine (SVM) and K nearest neighbours. SVM performed the best, to see other model performance see appendix 5.

4.1 Support Vector Machine (SVM)

SVM attempts to find the optimal hyperplane that separates classes. The model achieved an accuracy of 76.34% and outperformed tree-based ensemble methods and achieved performance comparable to the ridge logistic regression. It correctly identified 50 out of 61 patients without CHD and 21 out of 32 patients with CHD. This balanced performance across both classes, combined with the highest overall accuracy among all models tested (76.34%), suggests that SVM was particularly effective for this dataset.

5. Model Comparison and Selection

Model	Accuracy	AUC	Precision (no CHD)	Recall (no CHD)	Precision (CHD)	Recall (CHD)
Basic Ridge	74.19%	0.818	0.80	0.80	0.62	0.62
Tuned Ridge	74.19%	0.824	0.82	0.77	0.61	0.69
SVM	76.43%	0.796	0.82	0.82	0.66	0.66
Random Forest	69.89%	0.739	0.77	0.77	0.56	0.56
Gradient Boosting	69.89%	0.761	0.77	0.77	0.56	0.56
KNN	67.74%	0.748	0.73	0.80	0.54	0.44

Table 1 - Summary of Model Metrics

5.1 Performance comparison

SVM achieved the highest accuracy (76.34%) and exhibited strong, balanced performance across both classes. Its AUC of 0.796 was second only to the ridge logistic regression models. Both ridge logistic regression models performed well, with the tuned model showing improved recall for CHD cases (0.69 vs. 0.62) and achieved the highest AUC value (0.824), indicating superior ranking capability. Ensemble methods (Random Forest and Gradient Boosting) underperformed relative to both SVM and logistic regression, suggesting the decision boundaries in this dataset are smoother than what tree-based models efficiently capture. KNN showed an interesting pattern with the highest recall for non-CHD cases (0.80) but the lowest recall for CHD cases (0.44), indicating a substantial bias toward the majority class.

5.2 Best Model

Based on the evaluation, the SVM emerges as the best performing classifier, with several compelling advantages. Highest overall accuracy (76.34%) among all tested models, most balanced performance across both classes, with identical precision and recall values (0.82 for non-CHD and 0.66 for CHD) and strong discriminative ability indicated by an AUC of 0.796.

The tuned ridge logistic regression is a close second choice, with specific advantages in certain scenarios, highest AUC (0.824), indicating superior probability ranking, better recall for CHD cases (0.69). Additionally it has greater interpretability due to the direct relationship between coefficients and features, offering insights into risk factors.

6. Conclusion and Recommendations

This study demonstrates that both SVM and ridge logistic regression provide effective approaches for predicting CHD risk based on several factors. The optimisation of model hyperparameters yielded meaningful improvements in model performance, particularly in terms of class-specific metrics and AUC values.

Key risk factors consistently identified across models include age, family history, LDL cholesterol, and tobacco consumption, which aligns with established medical knowledge (Banks et al., 2019; Wilson et al., 1998). The relatively modest performance of tree-based ensemble methods suggests that the relationship between these risk factors and CHD may be more linear or smoothly non-linear than complex and hierarchical. Future work could explore combining multiple models through ensemble methods and investigating more sophisticated approaches for handling the class imbalance in the dataset.

Bibliography

Banks, E. *et al.* (2019) ‘Tobacco smoking and risk of 36 cardiovascular disease subtypes: Fatal and non-fatal outcomes in a large prospective Australian study’, *BMC Medicine*, 17(1). doi:10.1186/s12916-019-1351-4.

Ference, B.A. *et al.* (2017) ‘Low-density lipoproteins cause atherosclerotic cardiovascular disease. 1. evidence from genetic, epidemiologic, and clinical studies. A consensus statement from the European Atherosclerosis Society Consensus Panel’, *European Heart Journal*, 38(32), pp. 2459–2472. doi:10.1093/eurheartj/ehx144.

Lloyd-Jones, D.M. *et al.* (2004) ‘Parental cardiovascular disease as a risk factor for cardiovascular disease in middle-aged adults’, *JAMA*, 291(18), p. 2204. doi:10.1001/jama.291.18.2204.

Pencina, M.J. *et al.* (2009) ‘Predicting the 30-year risk of cardiovascular disease’, *Circulation*, 119(24), pp. 3078–3084. doi:10.1161/circulationaha.108.816694.

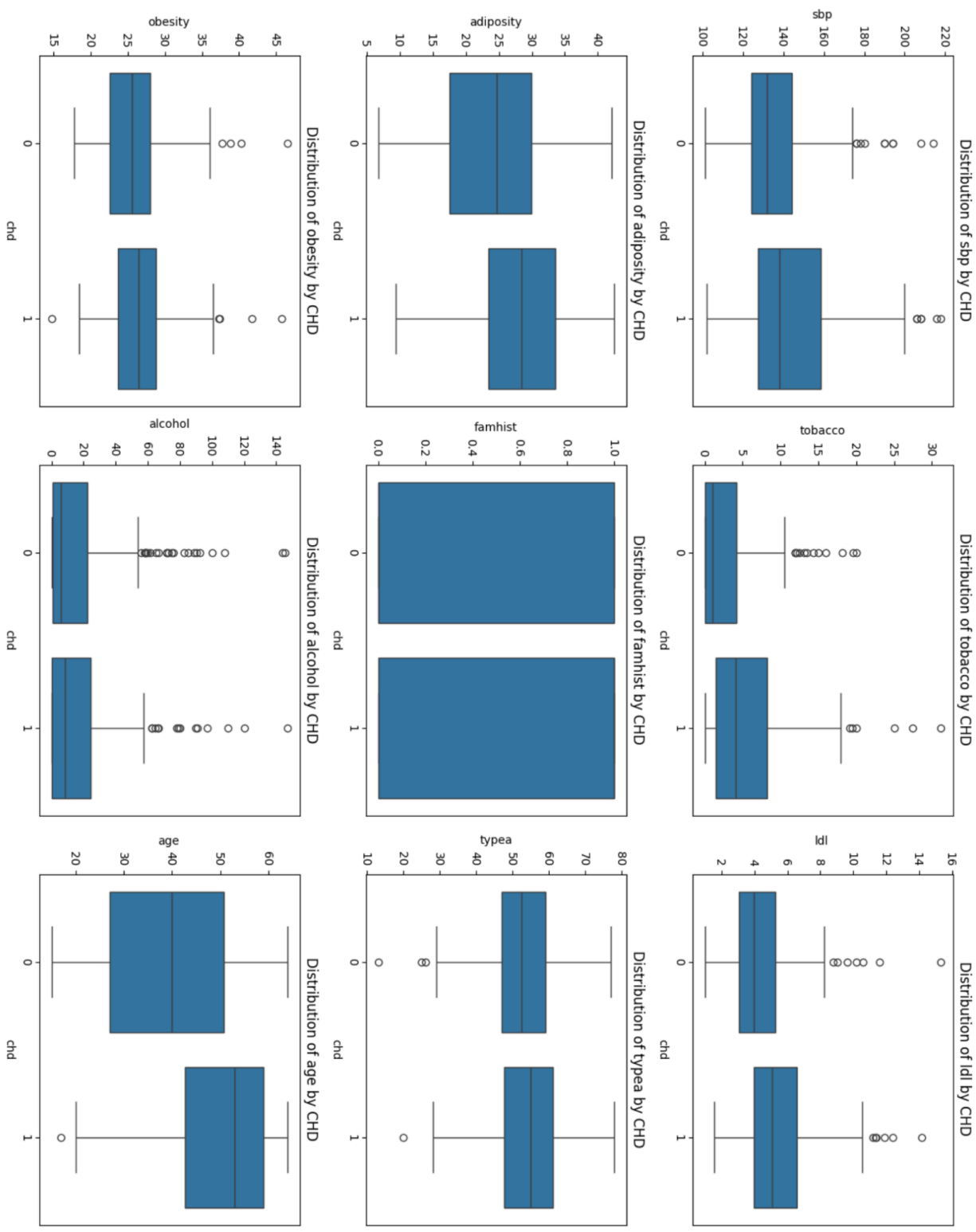
Wilson, P.W. *et al.* (1998a) ‘Prediction of coronary heart disease using risk factor categories’, *Circulation*, 97(18), pp. 1837–1847. doi:10.1161/01.cir.97.18.1837.

Appendix

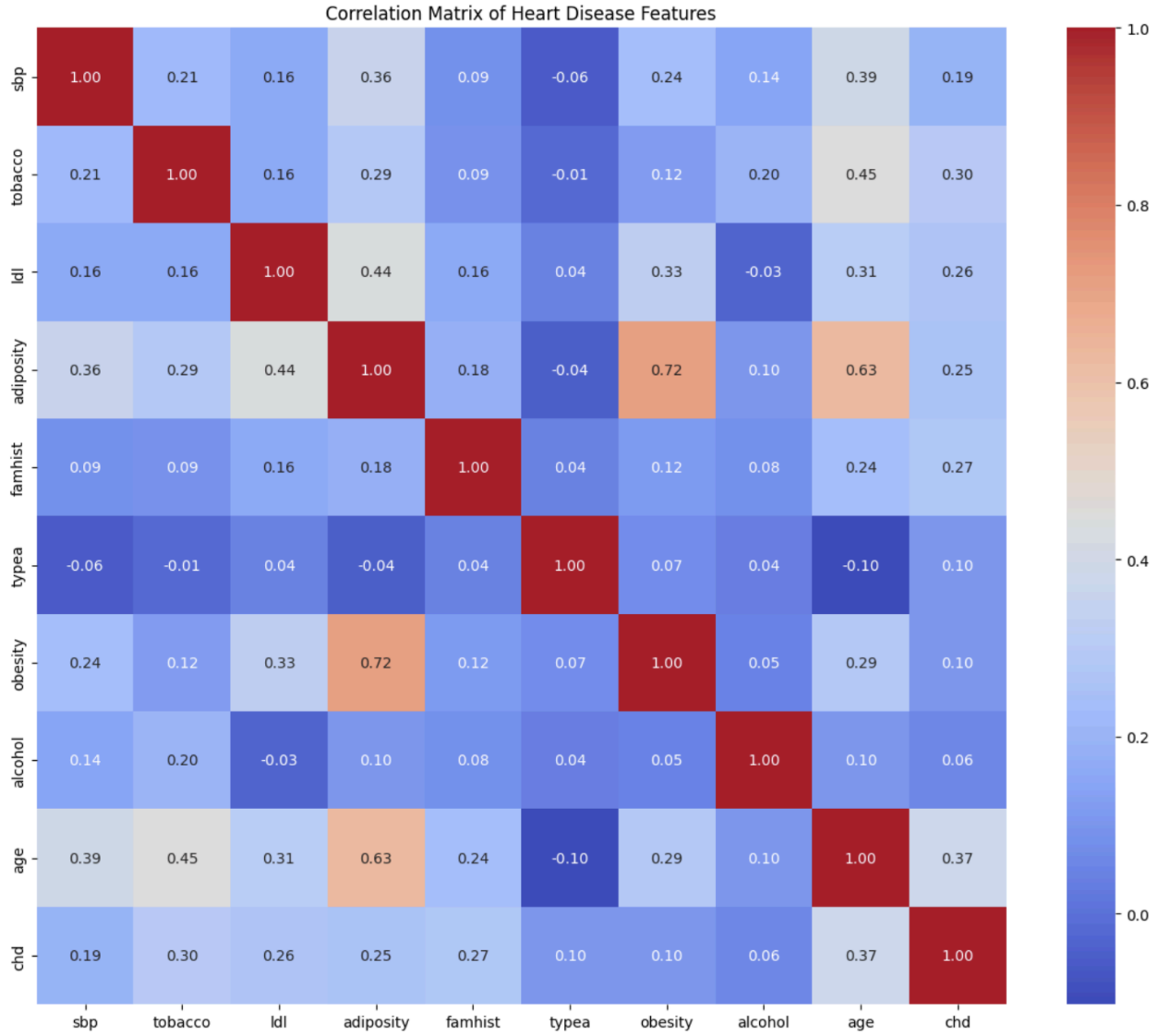
Appendix 1 - Summary Statistics

	sbp	tobacco	ldl	adiposity	typea	obesity	alcohol	age	chd
count	462.00	462.00	462.00	462.00	462.00	462.00	462.00	462.00	462.00
mean	138.33	3.64	4.74	25.41	53.10	26.04	17.04	42.82	0.35
std	20.50	4.59	2.07	7.78	9.82	4.21	24.48	14.61	0.48
min	101.00	0.00	0.98	6.74	13.00	14.70	0.00	15.00	0.00
25%	124.00	0.05	3.28	19.78	47.00	22.99	0.51	31.00	0.00
50%	134.00	2.00	4.34	26.12	53.00	25.81	7.51	45.00	0.00
75%	148.00	5.50	5.79	31.23	60.00	28.50	23.89	55.00	1.00
max	218.00	31.20	15.33	42.49	78.00	46.58	147.19	64.00	1.00

Appendix 2 - Boxplots of Features by CHD class



Appendix 3 - Correlation Matrix of Features



Appendix 4 - Baseline Logistic Regression

4.1 Baseline Logistic Regression with Ridge Penalty

Logistic regression with ridge penalty (L2 regularization) was implemented as our initial classification approach. Ridge regularisation helps prevent overfitting by penalizing large coefficients, which is particularly useful for our dataset where some features show moderate correlations. We started with a basic model, implemented using a default regularisation strength parameter (C=1.0) without any hyperparameter tuning to serve as our baseline model before accuracy optimisation.

4.11 Baseline Model Performance

The basic ridge logistic regression model achieved an accuracy of 72.19% on the test set, when examining the performance by class (table 1) we observe a precision and recall of 0.80 for the negative class (no CHD) and 0.62 for the positive class (CHD). This imbalance in performance between classes is expected given the moderate class imbalance in our dataset (approximately 34.6%) positive cases.

The confusion matrix (table 1) provides further insights into the model's predictions. Out of 61 patients without CHD, the model correctly identified 49 (80.3%) and misclassified 12. For the 32 patients with CHD, 20 (62.5%) were correctly classified, while 12 (37.5%) were incorrectly predicted as not having CHD. These results suggest that the model performs better as identifying non-CHD cases than detecting CHD cases.

The ROC curve (figure A1) shows the trade-off between sensitivity (true positive rate) and specificity (1-false positive rate) at various threshold settings. The area under the curve (AUC) for our basic model is 0.818, indicating good discriminative ability, an AUC of above 0.8 suggests the model has strong predictive power for distinguishing between patients with and without CHD.

4.12 Baseline Feature Importance Analysis

An advantage of logistic regression is its interpretability, the coefficient values provide insights into the importance and direction of each feature's relationship with the target variable, figure 2A illustrates the feature coefficients of the basic ridge model.

The most influential predictors for CHD, in order of importance, were:

1. **Age** (0.710), age stands out as the strongest predictor of CHD, with a substantial positive coefficient indicating higher risk with increasing age. This aligns with clinical knowledge that CHD risk increases significantly with age.

2. **Family History** (0.435), the presence of family history of heart disease is the second most important risk factor, confirming the genetic component of CHD susceptibility.
3. **LDL Cholesterol** (0.394), higher levels of low-density lipoprotein (LDL) cholesterol are associated with increased CHD risk, consistent with established medical understanding.
4. **Type-A Behavior** (0.360), interestingly, Type-A behavioral patterns show a positive association with CHD in our model, suggesting psychological factors may play a role in heart disease risk.
5. **Tobacco Consumption** (0.316), the model confirms tobacco use as a significant risk factor for CHD.

Several features showed negative coefficients, including, adiposity (-0.114), obesity (-0.012), alcohol consumption (-0.047). These negative associations appear counterintuitive, as these factors are typically considered risk factors for heart disease. This could potentially be due to multicollinearity among features (particularly between adiposity and obesity, which showed high correlation in our exploratory analysis) or the impact of the ridge penalty constraining coefficient values. It highlights the importance of further model refinement and careful interpretation of coefficients in the presence of correlated features.

The basic model demonstrates reasonable predictive performance with an accuracy of 74.19% and AUC of 0.818. However, there is room for improvement, particularly in the detection of positive CHD cases. The next step is to fine-tune the model by optimizing the regularization parameter to potentially enhance its predictive capability.

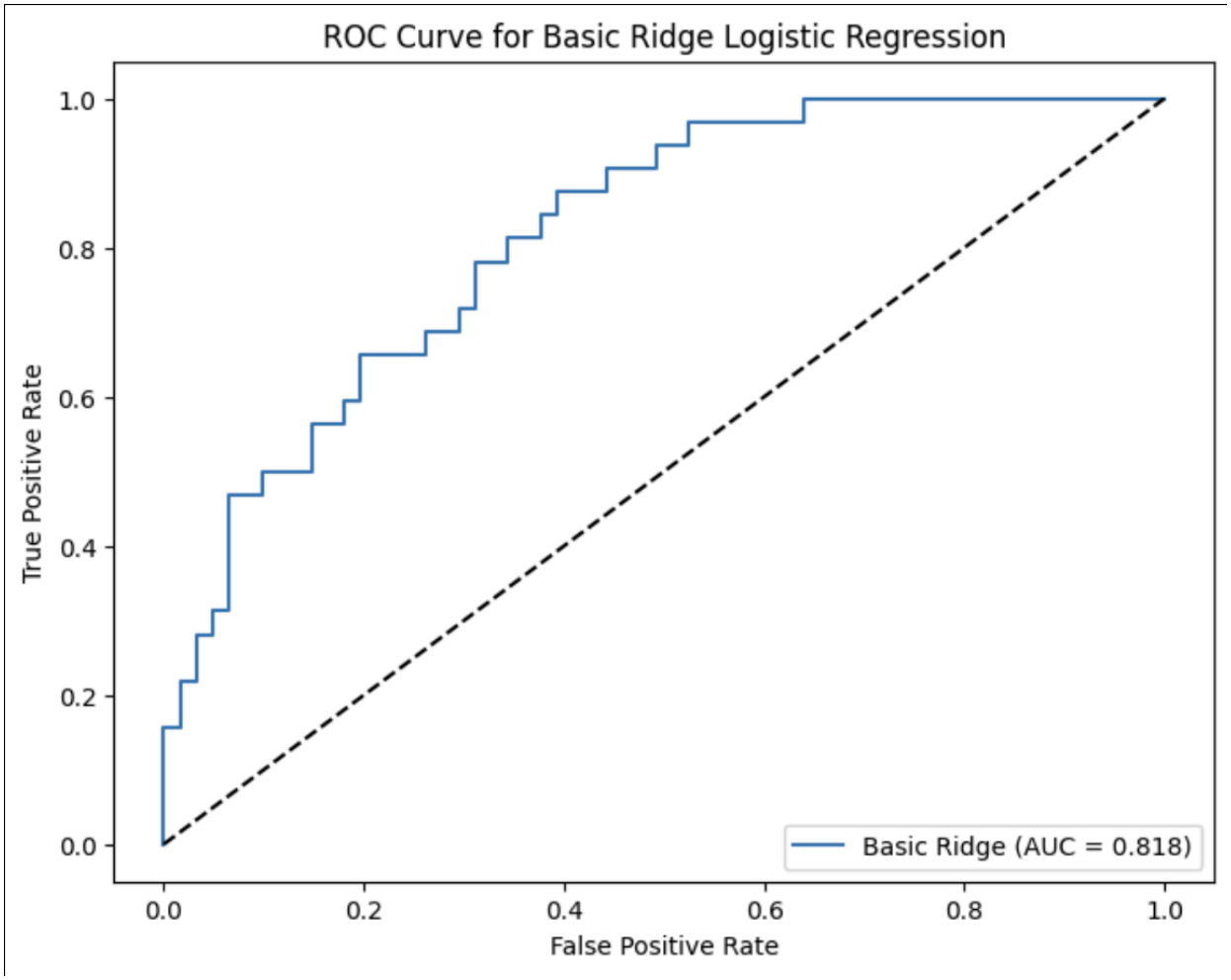


Figure 1A - ROC Curve Basic Ridge Logistic Regression

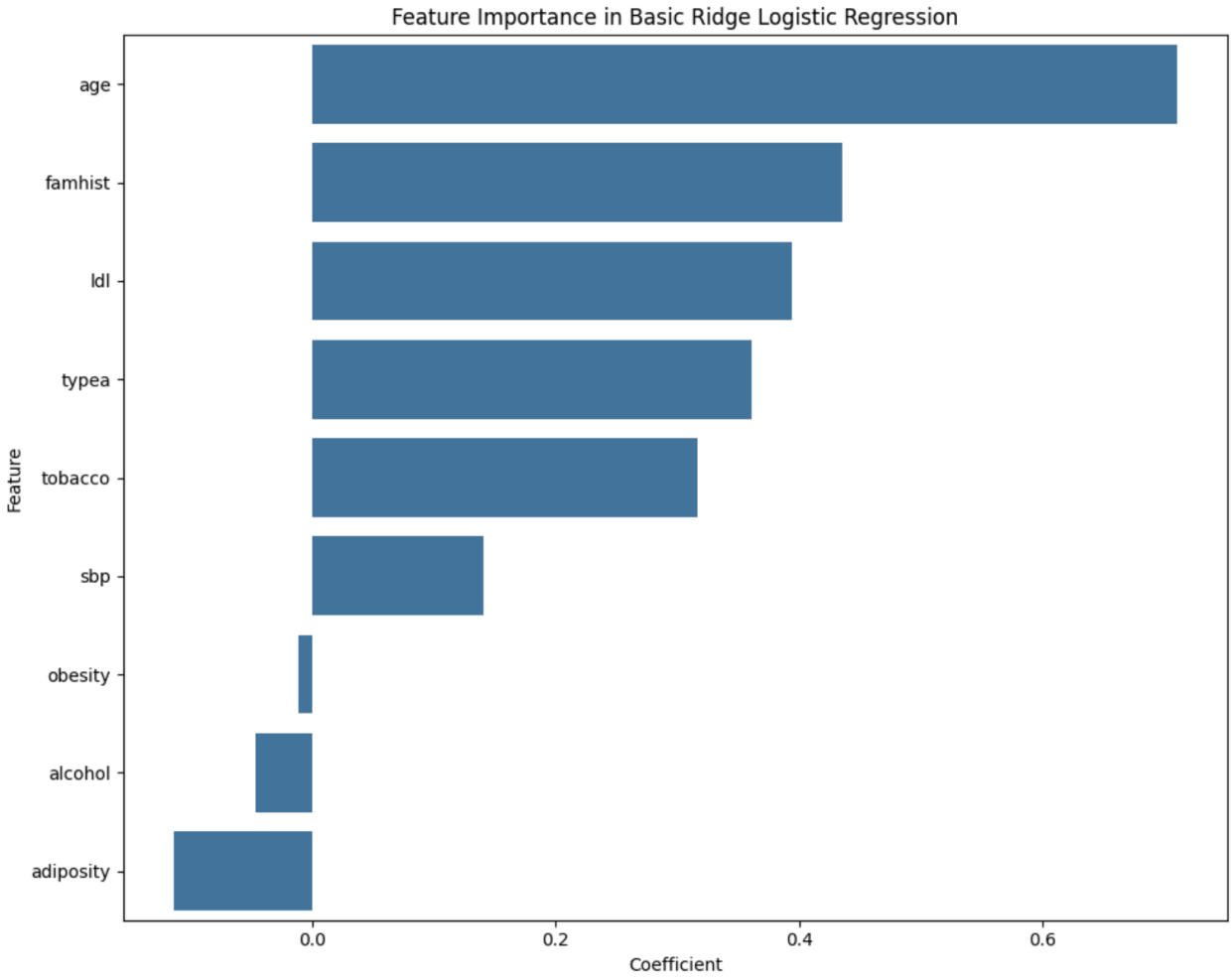


Figure 2A - Feature Importance Basic Ridge Logistic Regression

Appendix 5 - Other Classifiers

5.1 Random Forest

Random Forest is an ensemble method that builds multiple decision trees and aggregates their predictions to reduce overfitting. The model was tuned by optimising parameters such as the number of trees, maximum depth, and minimum samples per leaf. The best Random Forest model achieved a test accuracy was 69.89% and AUC value of 0.739. For patients without CHD recieved precision = 0.77, recall = 0.77 and F1 score = 0.77. For class 1 patients (with CHD), precision = 0.56, recall = 0.56, F1 score = 0.56. The Random Forest model correctly identified 47 out of 61 patients without CHD and 18 out of 32 patients with CHD. The performance was relatively balanced between classes but notably weaker than the ridge logistic regression for identifying CHD cases.

5.2 Gradient Boosting

Gradient Boosting builds trees sequentially, with each tree correcting errors made by previous trees. The model achieved an accuracy of 69.89% and AUC of 0.761. For patients without CHD received precision = 0.77, recall = 0.77 and F1 score = 0.77. For class 1 patients (with CHD), precision = 0.56, recall = 0.56, F1 score = 0.56. Interestingly, the Gradient Boosting model produced identical classification results to the Random Forest model on the test set despite having different hyperparameters. This suggests that both ensemble methods found similar patterns in the data but struggled to improve beyond a certain performance level. The Gradient Boosting model did achieve a slightly higher AUC (0.761 vs. 0.739), indicating somewhat better ranking of probabilities.

5.3 K-Nearest Neighbours (KNN)

KNN classifies based on the majority class of the K nearest neighbours. The optimized KNN model achieved an accuracy of 67.74% and AUC of 0.748. For patients without CHD recieved precision = 0.73, recall = 0.80 and F1 score = 0.77. For class 1 patients (with CHD), precision = 0.54, recall = 0.44, F1 score = 0.48. KNN exhibited the lowest overall accuracy (67.74%) among all the models tested. While its ability to identify non-CHD cases was reasonable (recall of 0.80), it struggled significantly with identifying CHD cases (recall of only 0.44). This imbalance

suggests that in the feature space created by our dataset, the distribution of CHD cases was not well captured by proximity-based classification.