BAYES
BUSINESS SCHOOL
CITY UNIVERSITY OF LONDON

Individual Coursework Submission Form

Specialist Masters Programme

Surname: Martin	**First Name: Nathaniel**
MSc in: Business Analytics	**Student ID number: 240046745**
Module Code: SMM636	
Module Title: Machine Learning	
Lecturer: 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: □ %

Coronary Heart Disease (CHD) is a leading cause of death globally, requiring accurate prediction models. This study aims to develop a classification model that achieves the highest accuracy in predicting CHD presence, based on nine health-related predictor variables. Secondary considerations include recall and F1-score, given their importance in medical diagnostics.

The dataset contains 462 observations and nine predictor variables including age, tobacco use, systolic blood pressure (sbp), and low-density lipoprotein (ldl). With an imbalanced class distribution (65.37% non-CHD vs. 34.63% CHD), class imbalance was addressed using model-based class weighting to minimize false negatives. Ridge Logistic Regression, SVM, Random Forest, and k-Nearest Neighbors were evaluated, with Ridge Regression ultimately selected due to its superior accuracy.

Data Preprocessing & Exploratory Data Analysis

Exploratory Data Analysis identified significant skew in 'tobacco' and 'alcohol', prompting log transformations. Moderate collinearity between 'adiposity', 'obesity', and 'ldl' guided targeted feature selection. Continuous variables were standardized, and 'famhist' was encoded to binary. Low-impact features ('typea', 'alcohol_log') were excluded after initial testing. Class imbalance was addressed through model-based class weighting.

In contrast, adiposity and ldl show more symmetrical distributions, while age appears bimodal or unevenly distributed, possibly indicating multiple age cohorts within the data. These patterns support the need for scaling and careful preprocessing before training distance- or margin-based classifiers.

A correlation heatmap highlighted moderate associations between variables such as adiposity and obesity, which informed subsequent feature selection. The high positive correlation between adiposity and ldl suggests potential multicollinearity, which could reduce model interpretability and stability. This reinforces the decision to use Ridge Logistic Regression, which helps mitigate such issues by penalizing large coefficients.

The categorical variable 'famhist' was encoded to binary values (Present=1, Absent=0). Log transformation was applied to 'tobacco' and 'alcohol', followed by standardization of continuous variables including 'sbp', 'ldl', 'adiposity', and 'age' using StandardScaler. Feature selection was guided by correlation analysis and initial model tests. Variable typea was excluded due to weak correlation with CHD occurrence, and alcohol_log was removed after preliminary analyses demonstrated minimal improvement in model accuracy. Class imbalance was addressed using class weighting (rather than oversampling), ensuring models could learn effectively without overfitting to the majority class.

Ridge Logistic Regression

Ridge Logistic Regression was selected as the final model primarily due to its superior accuracy (0.720) and highest AUC (0.800). Although its recall (0.625) and F1-score (0.606) were moderate, the emphasis on overall predictive accuracy made it most suitable for reliably predicting CHD. Ridge regularization applies an L2 penalty to the loss function, effectively reducing the impact of less influential predictors. This approach helps mitigate overfitting and improves generalization, particularly important when predictors exhibit multicollinearity or when working with limited data. The final model parameters were optimized using GridSearchCV, and performance was evaluated comprehensively using accuracy, precision, recall, and F1-score metrics.

This performance represents a solid balance across metrics. While recall is crucial in medical contexts to minimize false negatives, the achieved accuracy and highest AUC justify selecting Ridge Logistic Regression as the optimal classifier. Additional threshold tuning could be considered if future improvements in recall become necessary.

Alternative Classifiers for Comparison

We evaluated Ridge Logistic Regression, SVM, Random Forest, and kNN classifiers due to their recognized effectiveness, interpretability, and relevance based on course materials. Other classifiers, such as Decision Trees and Naive Bayes, were briefly considered but excluded due to inherent limitations in handling multicollinearity and continuous predictors.

The Support Vector Machine (SVM) with an RBF kernel was tuned using GridSearchCV across various 'C' and 'gamma' values, with class weights set to 'balanced' to address dataset imbalance. It achieved an accuracy of 0.688, with a notably high recall of 0.875 and an F1-score of 0.659, indicating its strength in identifying positive CHD cases despite slightly lower overall accuracy compared to Ridge Logistic Regression.

Random Forest, using 100 estimators, produced an accuracy of 0.667, recall of 0.469, and an F1-score of 0.492, reflecting moderate predictive capabilities but weaker performance overall.

The K-Nearest Neighbors (KNN) classifier achieved the lowest performance metrics among the evaluated models, with an accuracy of 0.591, recall of 0.438, and an F1-score of 0.424. These results highlight the variability in performance outcomes influenced by model complexity and parameter tuning. All models were rigorously assessed using stratified 5-fold cross-validation to ensure reliable and robust evaluation of predictive accuracy, precision, recall, and F1-score metrics.

Model Selection

Ridge Logistic Regression achieved the highest accuracy (0.720) and highest AUC (0.800), reflecting superior predictive capability. While SVM had notably higher recall (0.875) and strong F1-score (0.659), its lower accuracy (0.688) was less aligned with the primary objective. Random Forest and kNN underperformed in accuracy and recall, further supporting Ridge Logistic Regression as the final model.

Although SVM showed promise, its slightly lower stability and interpretability made Ridge Regression more suitable for real-world deployment. Additionally, logistic regression offers higher transparency and interpretability, which is valuable in healthcare applications. Coefficients in logistic models can provide insights into influential CHD feature prediction, helping support clinical decision-making.

Limitations & Future Work

This analysis is limited by the relatively small dataset size (462 observations). Future work could validate findings on larger or diverse datasets, explore hyperparameter optimization further for SVM or Random Forest, or investigate advanced ensemble or deep learning models to potentially improve accuracy. Moreover, integrating explainability tools like SHAP or LIME could help interpret individual predictions, increasing trust and usability of the model in real-world healthcare settings. Additional improvements may be realized through tuning Random Forest, trying gradient boosting methods, or incorporating ensemble models. Exploring model calibration and explainability tools like SHAP or LIME could further enhance trust and transparency in clinical decision-making.

References

Wakankar, R., Khangembam, B., Patel, C. & Kumar, R. (2024). Machine learning can predict M1 disease in treatment-naïve prostate cancer patients undergoing 68Ga-PSMA-11 PET/CT. IAEA. Available at: INIS Database [Accessed 18 March 2025].

Figures

Figure 1: Histograms of predictor variables

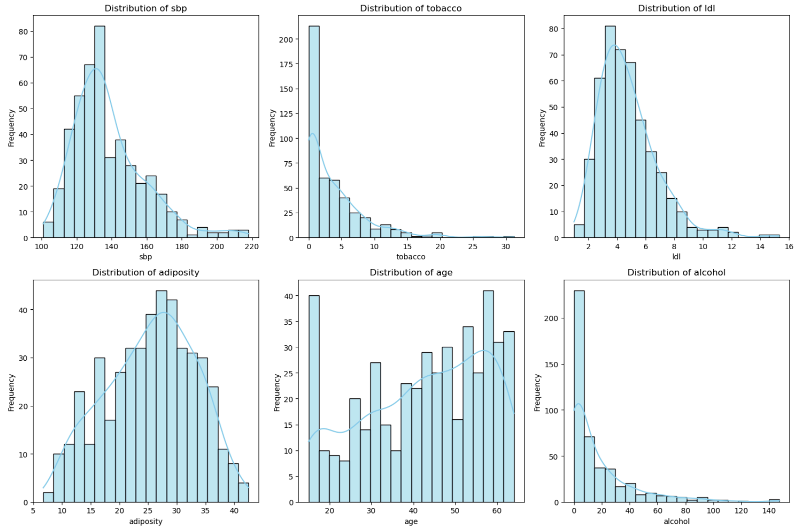
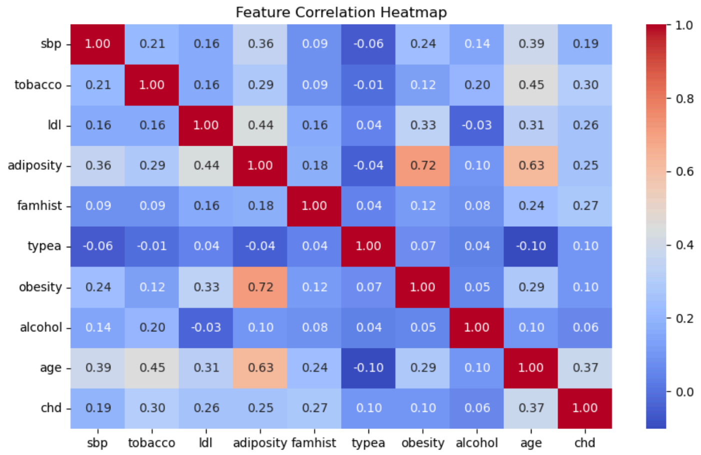


Figure 2: Correlation heatmap



	sbp	tobacco	ldl	adiposity	famhist	typea	obesity	alcohol	age	chd
sbp	1.00	0.21	0.16	0.36	0.09	-0.06	0.24	0.14	0.39	0.19
tobacco	0.21	1.00	0.16	0.29	0.09	-0.01	0.12	0.20	0.45	0.30
ldl	0.16	0.16	1.00	0.44	0.16	0.04	0.33	-0.03	0.31	0.26
adiposity	0.36	0.29	0.44	1.00	0.18	-0.04	0.72	0.10	0.63	0.25
famhist	0.09	0.09	0.16	0.18	1.00	0.04	0.12	0.08	0.24	0.27
typea	-0.06	-0.01	0.04	-0.04	0.04	1.00	0.07	0.04	-0.10	0.10
obesity	0.24	0.12	0.33	0.72	0.12	0.07	1.00	0.05	0.29	0.10
alcohol	0.14	0.20	-0.03	0.10	0.08	0.04	0.05	1.00	0.10	0.06
age	0.39	0.45	0.31	0.63	0.24	-0.10	0.29	0.10	1.00	0.37
chd	0.19	0.30	0.26	0.25	0.27	0.10	0.10	0.06	0.37	1.00

Figure 3: ROC curve comparison

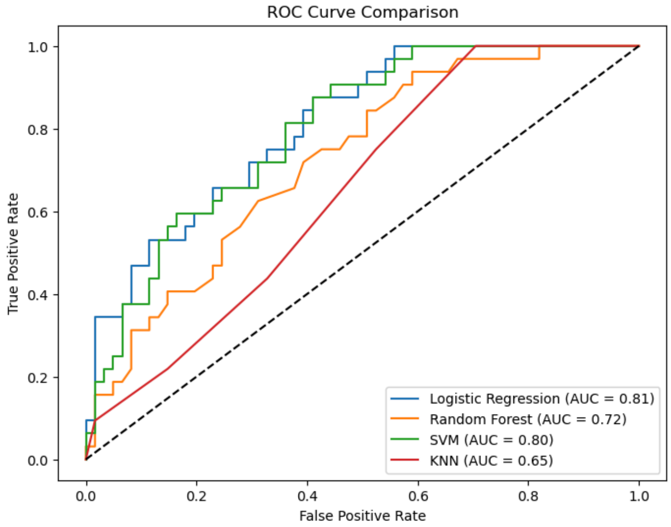
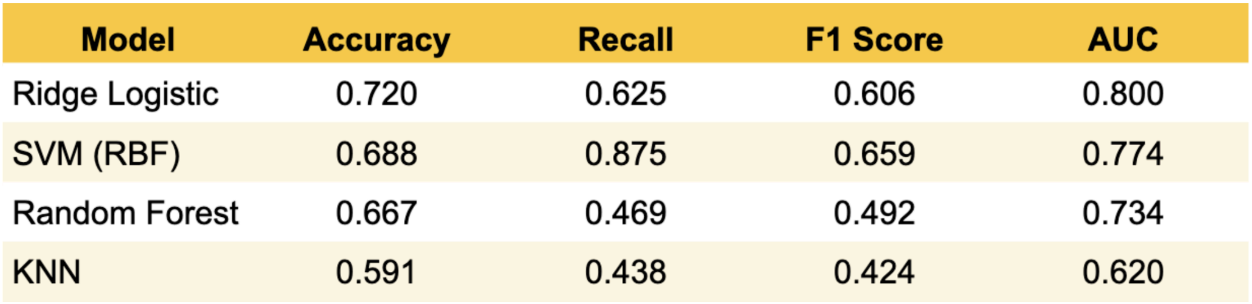


Figure 4: Summary Table



Model	Accuracy	Recall	F1 Score	AUC
Ridge Logistic	0.720	0.625	0.606	0.800
SVM (RBF)	0.688	0.875	0.659	0.774
Random Forest	0.667	0.469	0.492	0.734
KNN	0.591	0.438	0.424	0.620

Coronary Heart Disease (CHD) is a leading cause of death globally, requiring accurate prediction models. This study aims to develop a classification model that achieves the highest accuracy in predicting CHD presence, based on nine health-related predictor variables. Secondary considerations include recall and F1-score, given their importance in medical diagnostics.

The dataset contains 462 observations and nine predictor variables including age, tobacco use, systolic blood pressure (sbp), and low-density lipoprotein (ldl). With an imbalanced class distribution (65.37% non-CHD vs. 34.63% CHD), class imbalance was addressed using model-based class weighting to minimize false negatives. Ridge Logistic Regression, SVM, Random Forest, and k-Nearest Neighbors were evaluated, with Ridge Regression ultimately selected due to its superior accuracy.

Data Preprocessing & Exploratory Data Analysis

Exploratory Data Analysis identified significant skew in 'tobacco' and 'alcohol', prompting log transformations. Moderate collinearity between 'adiposity', 'obesity', and 'ldl' guided targeted feature selection. Continuous variables were standardized, and 'famhist' was encoded to binary. Low-impact features ('typea', 'alcohol_log') were excluded after initial testing. Class imbalance was addressed through model-based class weighting.

In contrast, adiposity and ldl show more symmetrical distributions, while age appears bimodal or unevenly distributed, possibly indicating multiple age cohorts within the data. These patterns support the need for scaling and careful preprocessing before training distance- or margin-based classifiers.

A correlation heatmap highlighted moderate associations between variables such as adiposity and obesity, which informed subsequent feature selection. The high positive correlation between adiposity and ldl suggests potential multicollinearity, which could reduce model interpretability and stability. This reinforces the decision to use Ridge Logistic Regression, which helps mitigate such issues by penalizing large coefficients.

The categorical variable 'famhist' was encoded to binary values (Present=1, Absent=0). Log transformation was applied to 'tobacco' and 'alcohol', followed by standardization of continuous variables including 'sbp', 'ldl', 'adiposity', and 'age' using StandardScaler. Feature selection was guided by correlation analysis and initial model tests. Variable typea was excluded due to weak correlation with CHD occurrence, and alcohol_log was removed after preliminary analyses demonstrated minimal improvement in model accuracy. Class imbalance was addressed using class weighting (rather than oversampling), ensuring models could learn effectively without overfitting to the majority class.

Ridge Logistic Regression

Ridge Logistic Regression was selected as the final model primarily due to its superior accuracy (0.720) and highest AUC (0.800). Although its recall (0.625) and F1-score (0.606) were moderate, the emphasis on overall predictive accuracy made it most suitable for reliably predicting CHD. Ridge regularization applies an L2 penalty to the loss function, effectively reducing the impact of less influential predictors. This approach helps mitigate overfitting and improves generalization, particularly important when predictors exhibit multicollinearity or when working with limited data. The final model parameters were optimized using GridSearchCV, and performance was evaluated comprehensively using accuracy, precision, recall, and F1-score metrics.

This performance represents a solid balance across metrics. While recall is crucial in medical contexts to minimize false negatives, the achieved accuracy and highest AUC justify selecting Ridge Logistic Regression as the optimal classifier. Additional threshold tuning could be considered if future improvements in recall become necessary.

Alternative Classifiers for Comparison

We evaluated Ridge Logistic Regression, SVM, Random Forest, and kNN classifiers due to their recognized effectiveness, interpretability, and relevance based on course materials. Other classifiers, such as Decision Trees and Naive Bayes, were briefly considered but excluded due to inherent limitations in handling multicollinearity and continuous predictors.

The Support Vector Machine (SVM) with an RBF kernel was tuned using GridSearchCV across various 'C' and 'gamma' values, with class weights set to 'balanced' to address dataset imbalance. It achieved an accuracy of 0.688, with a notably high recall of 0.875 and an F1-score of 0.659, indicating its strength in identifying positive CHD cases despite slightly lower overall accuracy compared to Ridge Logistic Regression.

Random Forest, using 100 estimators, produced an accuracy of 0.667, recall of 0.469, and an F1-score of 0.492, reflecting moderate predictive capabilities but weaker performance overall.

The K-Nearest Neighbors (KNN) classifier achieved the lowest performance metrics among the evaluated models, with an accuracy of 0.591, recall of 0.438, and an F1-score of 0.424. These results highlight the variability in performance outcomes influenced by model complexity and parameter tuning. All models were rigorously assessed using stratified 5-fold cross-validation to ensure reliable and robust evaluation of predictive accuracy, precision, recall, and F1-score metrics.

Model Selection

Ridge Logistic Regression achieved the highest accuracy (0.720) and highest AUC (0.800), reflecting superior predictive capability. While SVM had notably higher recall (0.875) and strong F1-score (0.659), its lower accuracy (0.688) was less aligned with the primary objective. Random Forest and kNN underperformed in accuracy and recall, further supporting Ridge Logistic Regression as the final model.

Although SVM showed promise, its slightly lower stability and interpretability made Ridge Regression more suitable for real-world deployment. Additionally, logistic regression offers higher transparency and interpretability, which is valuable in healthcare applications. Coefficients in logistic models can provide insights into influential CHD feature prediction, helping support clinical decision-making.

Limitations & Future Work

This analysis is limited by the relatively small dataset size (462 observations). Future work could validate findings on larger or diverse datasets, explore hyperparameter optimization further for SVM or Random Forest, or investigate advanced ensemble or deep learning models to potentially improve accuracy. Moreover, integrating explainability tools like SHAP or LIME could help interpret individual predictions, increasing trust and usability of the model in real-world healthcare settings. Additional improvements may be realized through tuning Random Forest, trying gradient boosting methods, or incorporating ensemble models. Exploring model calibration and explainability tools like SHAP or LIME could further enhance trust and transparency in clinical decision-making.

References

Wakankar, R., Khangembam, B., Patel, C. & Kumar, R. (2024). Machine learning can predict M1 disease in treatment-naïve prostate cancer patients undergoing 68Ga-PSMA-11 PET/CT. IAEA. Available at: INIS Database [Accessed 18 March 2025].

Figures

Figure 1: Histograms of predictor variables

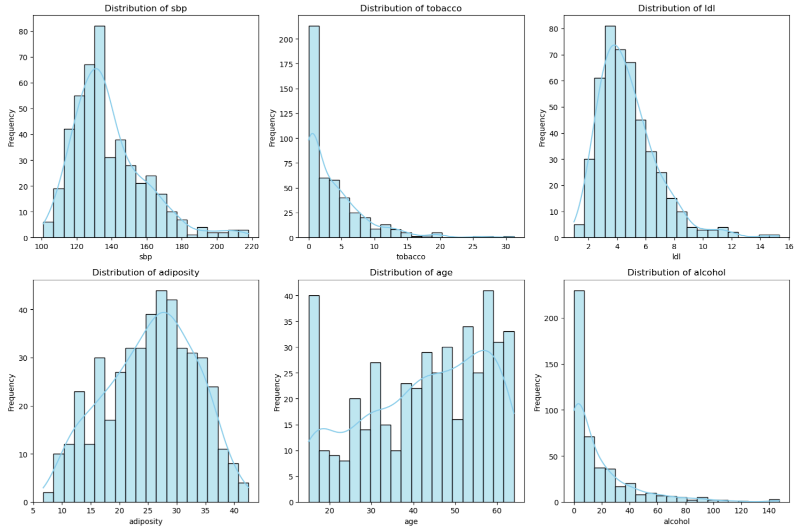


Figure 2: Correlation heatmap

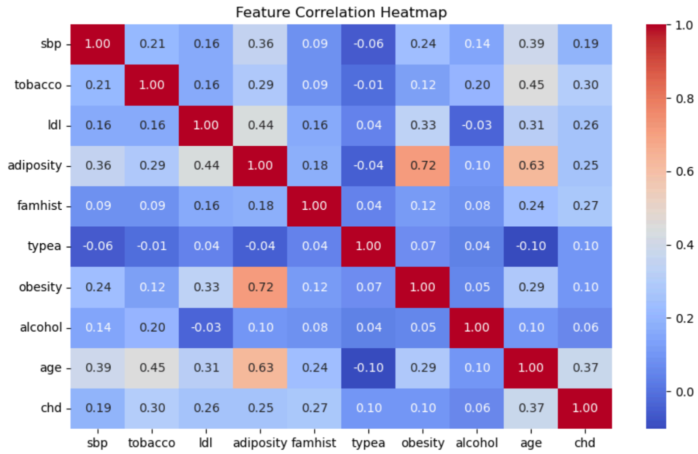


Figure 3: ROC curve comparison

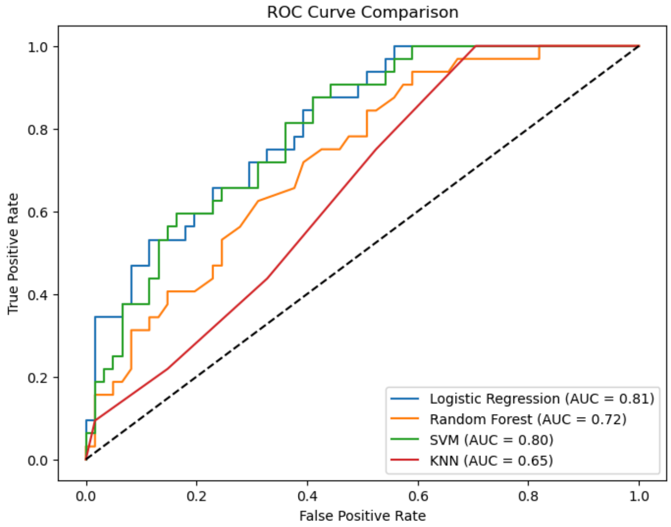
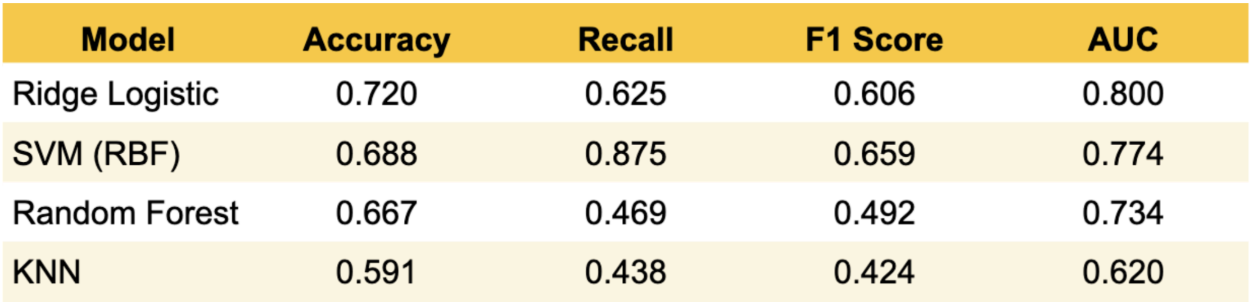


Figure 4: Summary Table



Model	Accuracy	Recall	F1 Score	AUC
Ridge Logistic	0.720	0.625	0.606	0.800
SVM (RBF)	0.688	0.875	0.659	0.774
Random Forest	0.667	0.469	0.492	0.734
KNN	0.591	0.438	0.424	0.620