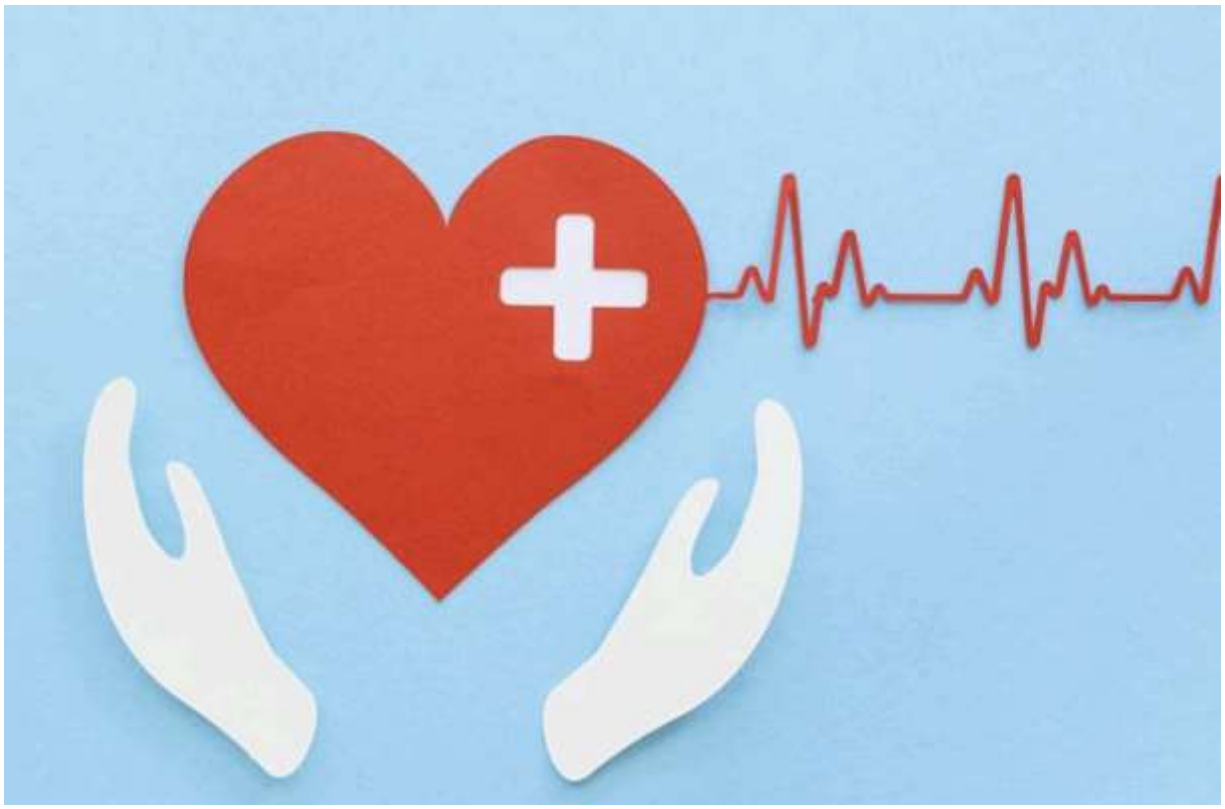
Heart Attack Risk Prediction: Machine Learning Project Report



12.03.2024

Melisa Lara Denizoglu

Part 1 : Explanatory analysis (EDA)

This dataset focuses on predicting coronary heart disease (CHD) in high-risk males from the Western Cape, South Africa, representing a classic supervised learning classification problem. The Exploratory Data Analysis (EDA) conducted on a dataset of 462 individuals aims to unearth risk factors for coronary heart disease (CHD). It reveals an imbalanced distribution with 34.63% diagnosed with CHD and 65.37% without, potentially impacting the accuracy of future models. This phase handles diverse data types, transforming 'famhist', a categorical variable indicating a family history of CHD, into a numerical one for logistic regression compatibility. The EDA ensures no missing values or extreme outliers, maintaining dataset integrity for reliable predictive modelling.

Through detailed analyses using histograms, box plots, and scatter plots, we've gained a deeper understanding of the distributions, central tendencies, and variability of variables related to coronary heart disease (CHD). These visual explorations, along with summary statistics, have provided a detailed overview of the dataset. Individuals with a family history of CHD have a 50% incidence rate of the disease, which is more than double the rate of those without a family history (23.7%). The target variable analysis versus other variables also offers profound insights into CHD risk factors. Individuals diagnosed with CHD, on average, exhibit higher systolic blood pressure (143.74 mmHg vs. 135.46 mmHg), greater tobacco use (5.52 units vs. 2.63 units), increased LDL cholesterol levels (5.49 vs. 4.34), and higher adiposity (28.12 vs. 23.97) compared to those without CHD.

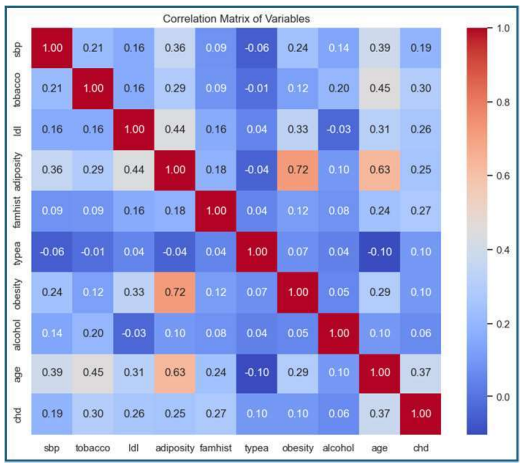


Table 1: Correlation Matrix

The correlation heatmap insights from the dataset's analysis reveal significant relationships, such as the strong positive correlation between adiposity and obesity (0.72), indicating that an increase in adiposity is typically associated with higher levels of obesity. Additionally, there is a notable positive correlation between adiposity and age (0.63), suggesting that adiposity tends to increase with age, while age and tobacco use display a moderately strong positive correlation (0.45), implying that tobacco usage may be more prevalent among older individuals.

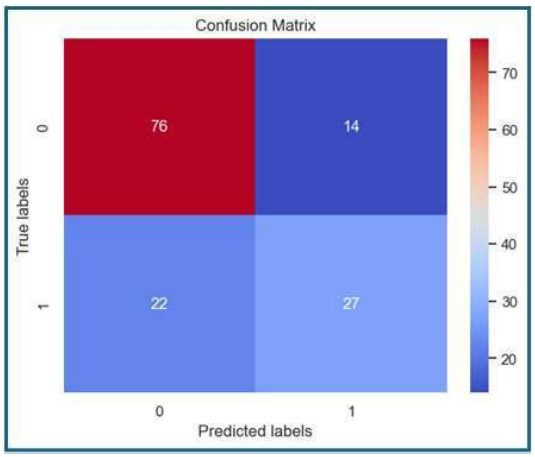
Part 2: Logistic Regression with Ridge Penalty

In this part, we utilise logistic regression with a ridge penalty to differentiate patients into CHD-positive or CHD-negative categories. By incorporating ridge regularization, the model combats overfitting, making it instrumental in pinpointing CHD risk factors within binary classification scenarios. For our logistic regression model, the equation based on the fitted model can be cited as follows:

$$\textit{LogOdds(CHD)} = -0.8607747 + 0.19569403 \times \textbf{sbp} + 0.27777013 \times \textbf{tobacco} + 0.41754378 \times \textbf{ldl} + 0.19339188 \times \textbf{adiposity} + 0.36367854 \times \textbf{famhist} + 0.35949123 \times \textbf{typea} - 0.2634047 \times \textbf{obesity} + 0.02193324 \times \textbf{alcohol} + 0.59344898 \times \textbf{age}$$

Table 2: Logistic Regression Model

Each coefficient in this model represents the change in log odds of CHD for a one-unit increase in the predictor variable, holding all other variables constant. In the importance feature analysis, it was determined that age had the highest coefficient with 0.59344898. This shows that as age increases, the likelihood of coronary heart disease (CHD) increases significantly.



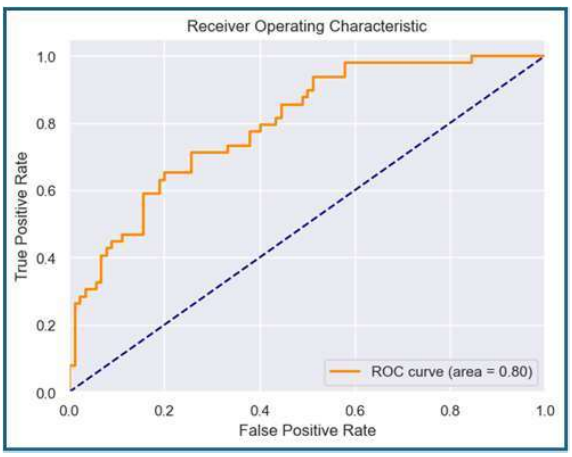


Table 3: Confusion Matrix and Table 4: ROC Curve Table

The model's performance in predicting Coronary Heart Disease (CHD) was evaluated using accuracy, precision, recall, F1 Score, ROC-AUC score, and Log Loss. With a dataset imbalance (35% CHD-positive cases), accuracy alone is not sufficient. Precision (66%) indicates reliability in identifying CHD-positive cases, while recall (55%) shows its ability to capture actual cases. The F1 Score (0.60) balances precision and recall moderately well. A ROC-AUC score of 0.80 indicates effective differentiation between positive and negative cases. The Log Loss value (0.5083) provides a consolidated measure of predictive precision.

Additionally, 10-fold cross-validation confirmed the model's generalizability, while hyperparameter tuning identified C = 0.1 as optimal for the Ridge penalty, enhancing model balance. The optimized model showed strong performance in identifying non-CHD cases (precision: 0.76, recall: 0.84, F1-score: 0.80) but was less effective for CHD-positive cases (F1-score: 0.57), supporting its use for reliable CHD prediction. These outcomes confirm our choice of the initial model with 74% accuracy level.

Part 3: Other Classifiers and Finding the Best Model

In our analysis, we thoroughly investigated seven distinct models to identify the most effective approach for predicting Coronary Heart Disease (CHD), each offering unique advantages for binary classification. Gradient Boosting Machines master complex pattern detection, and Support Vector Machines adeptly classify data into distinct groups, needing precise kernel choices for non-linear datasets. Meanwhile, K-Nearest Neighbors relies on instance proximity with a must

for correct data scaling, whereas the CART algorithm excels in CHD prediction through its interpretable binary trees, capable of processing various feature types.

During the exploratory data analysis (EDA) phase, we performed essential data preprocessing to ensure the integrity and relevance of our dataset. In the modelling phase, we carefully fitted each model, ensuring precision and consistency in our approach. Furthermore, in the model evaluation phase, we employed advanced techniques such as cross-validation and GridSearch for hyperparameter tuning, aiming to maximize the performance of each model. We took careful steps to fairly compare different models, which was especially important because we were working with imbalanced data. The performance levels of accuracy for these models are summarized in the following table:

Ranking Based on Accuracy Level	Classification Type	Accuracy Level	Precision (Class 1)	Recall (Class 1)	F-1 Score (Class 1)	ROC-AUC Score
1	Decision Tree (CART)	0.80.58 %	0.89%	0.51%	0.65%	0.88%
2	Logistic Regression with Ridge	0.74 %	0.66%	0.55%	0.60%	0.80%
3	KNN	0.73.16 %	0.64%	0.39%	0.48%	0.66%
4	Gradient Boosting	0.72.66 %	0.73%	0.63%	0.55%	0.76 %
5	Support Vector Machine	0.71.94 %	0.62%	0.51%	0.56 %	0.79 %
6	Random Forest	0.70.5 %	0.73%	0.87%	0.79%	0.74%
7	Naïve Bayesian	0.70 %	0.57 %	0.61%	0.59%	0.74%

Table 5: Classification Reports

According to the metrics, The Decision Tree (CART) model stands out as the top performer with an 80.58% accuracy, 89.9% precision for class 1, and an 88% ROC-AUC score, showcasing its superior classification capabilities. Logistic Regression offers respectable performance with a 74% accuracy and balanced metrics, including a 66% precision and a 55.5% recall for class 1. In contrast, KNN, despite its fairly high 73.16% accuracy, falls short in precision (64%), recall (39%) for class 1, and an F1 score of 48%, highlighting its struggle with effectively balancing precision and recall. This comparison once again highlights the importance of evaluating various

metrics beyond accuracy. While the SVM and Gradient Boosting models show solid and balanced performances respectively, conversely the Naïve Bayesian classifier, despite its lowest accuracy, shows a competitive ROC-AUC score, indicating a decent class discrimination ability. Notably, the Random Forest model, while lower in overall accuracy, demonstrates a remarkable recall for class 1 (87.9%).

The Best Classification: Decision Tree with CART model

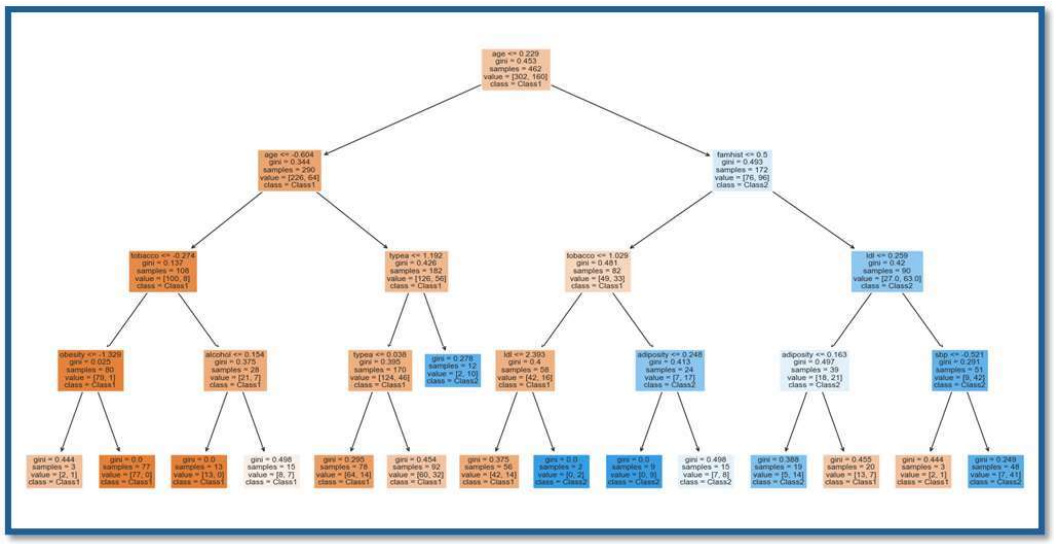


Table 6: CART's Visualisation