Table of Contents

1. Introduction

In this machine learning project, we will explore real-world datasets through the lenses of unsupervised learning, regression, and classification to uncover patterns, predict continuous outcomes, and categorize distinct groups. Employing a variety of machine learning techniques, we will compare their predictive performances to identify the most effective methods for our data-driven inquiries, ensuring a nuanced understanding of the complex dynamics at play within our chosen datasets.

2. Country Development (Unsupervised Learning)

HELP International, a non-governmental organization dedicated to combating poverty and supplying essential resources to underprivileged nations, plans to distribute its $10 million funding strategically. Identifying the countries most in need of aid includes clustering them based on socio-economic and health factors, thus determining their overall development status. This approach will yield crucial insights into prioritizing assistance and attention from HELP International.

2.1 Research Questions

The research questions (RQ) that have been identified for the issue are as follows:

- RQ1: How many clusters are ideal for evaluating a country's level of development?
- RQ2: How should we interpret the clusters that have been identified?
- RQ3: Which countries are more likely to require financial assistance?

2.2 Methodology

Principal component analysis (PCA), K-means clustering and hierarchical clustering with Euclidean Dendrogram are used to analyze the dataset to understand the pattern and relationship between variables with the aim of clustering countries with different level of financial aid requirement. Silhouette, elbow methods and gap statistics are also used in determining the optimal number of clusters.

2.3 Dataset and Variables

The dataset consists of 167 rows and 10 columns. There are no NAs and missing values. Country variable was removed in numeric data and we will retain the outliers as they could indicate a country's dire situation and need for financial aid.

2.4 Analysis

Correlation Matrix is first performed to examine the relationship between variables. From Figure 2.1, it is observed that the most strongly correlated pairs are income/GDP, followed by child mortality/total fertility and exports/import which would likely serve as prominent indicators for categorizing countries into clusters later on. PCA is a dimensionality-reduction technique that reduces the dataset's variables while preserving significant information from the original data. Too few variables would result in a loss of information while too many variables may increase complexity. From Figure 2.2, 3 components will be chosen as they amount to 76.14% of the total variance and their respective variance are more than 1 (Figure 2.3). PC1 represents "quality of life" since child_mort, total_fer, life_expec, gdpp, and income

load heavily on PC1. PC2 represents “trading condition” since imports and exports load heavily on PC2. PC3 represents “inflation rate” as inflation and health load heavily on PC3.

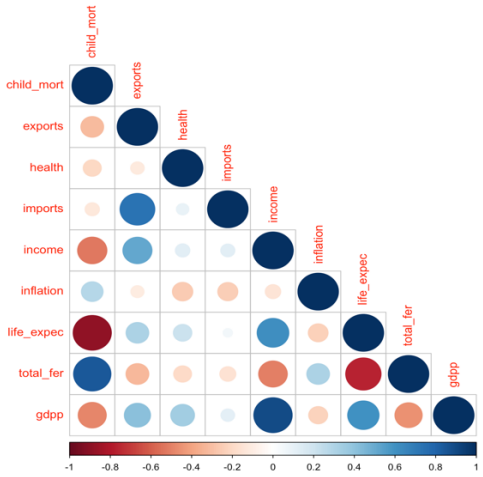

Figure 2.1: Correlation Matrix

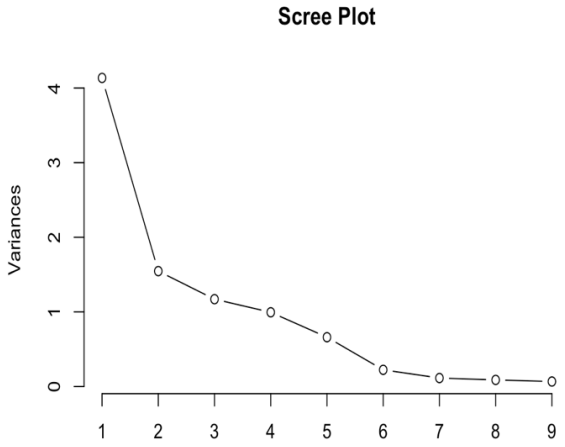

Figure 2.2: PCA Scree Plot

```
Importance of components:
                          PC1    PC2    PC3    PC4    PC5     PC6    PC7     PC8     PC9
Standard deviation     2.0336 1.2435 1.0818 0.9974 0.8128 0.47284 0.3368 0.29718 0.25860
Proportion of Variance 0.4595 0.1718 0.1300 0.1105 0.0734 0.02484 0.0126 0.00981 0.00743
Cumulative Proportion  0.4595 0.6313 0.7614 0.8719 0.9453 0.97015 0.9828 0.99257 1.00000

                PC1         PC2         PC3         PC4         PC5         PC6         PC7
child_mort -0.4195194 -0.192883937  0.02954353  0.370653262 -0.16896968 -0.200628153 -0.07948854
exports     0.2838970 -0.613163494 -0.14476069  0.003091019  0.05761584  0.059332832 -0.70730269
health      0.1508378  0.243086779  0.59663237  0.461897497  0.51800037 -0.007276456 -0.24983051
imports     0.1614824 -0.671820644  0.29992674 -0.071907461  0.25537642  0.030031537  0.59218953
income      0.3984411 -0.022535530 -0.30154750  0.392159039 -0.24714960 -0.160346990  0.09556237
inflation  -0.1931729  0.008404473 -0.64251951  0.150441762  0.71486910 -0.066285372  0.10463252
life_expec  0.4258394  0.222706743 -0.11391854 -0.203797235  0.10821980  0.601126516  0.01848639
total_fer  -0.4037290 -0.155233106 -0.01954925  0.378303645 -0.13526221  0.750688748  0.02882643
gdpp        0.3926448  0.046022396 -0.12297749  0.531994575 -0.18016662 -0.016778761  0.24299776
                PC8         PC9
child_mort  0.68274306 -0.32754180
exports     0.01419742  0.12308207
health     -0.07249683 -0.11308797
imports     0.02894642 -0.09903717
income     -0.35262369 -0.61298247
inflation   0.01153775  0.02523614
life_expec  0.50466425 -0.29403981
total_fer  -0.29335267  0.02633585
gdpp        0.24969636  0.62564572
```

Figure 2.3: PCA

Hierarchical clustering employs both Euclidean and Manhattan distances, utilizing four dissimilarity measures: single, complete, average, and Ward's method. Assessing the aggregation coefficient, which gauges the clarity of the clustering structure (where proximity to 1 indicates stronger structure), reveals that Manhattan Hierarchical Clustering, when coupled with Ward's Linkage, exhibits the highest performance, achieving 97.55%.

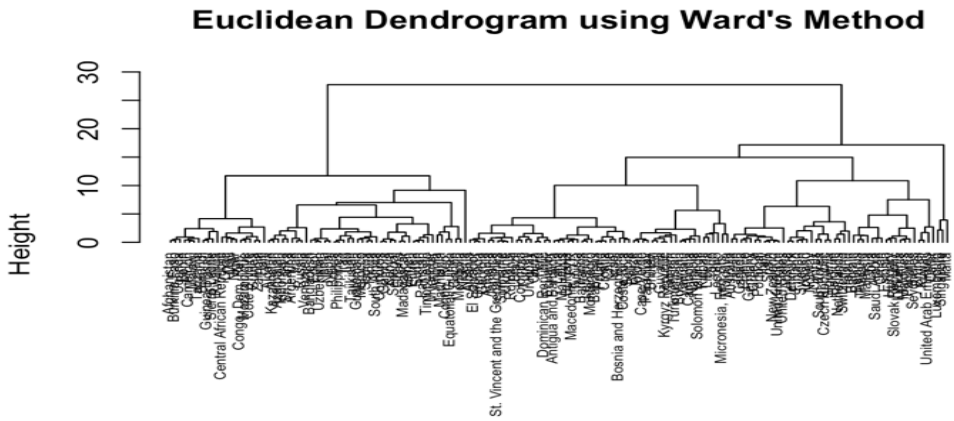

Figure 2.4: Hierarchical Clustering: Euclidean Dendrogram

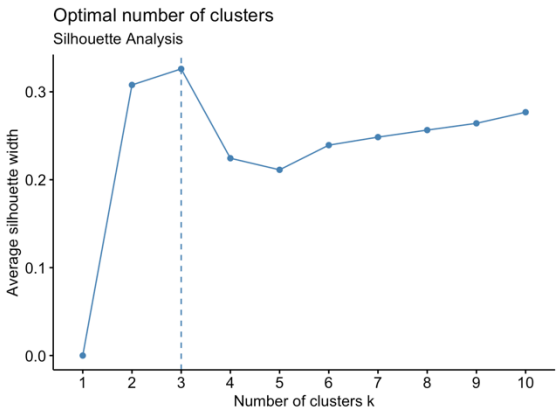

Figure 2.5: Silhouette Analysis

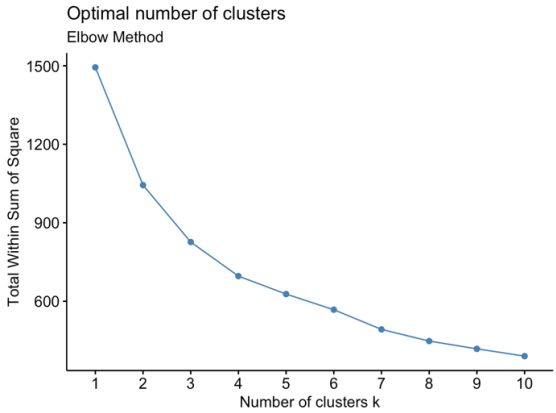

Figure 2.6: Elbow Method

Through Silhouette Analysis and Elbow Method (Figure 2.5 & 2.6), we are able to determine 3 as the optimal number of clusters for clustering.

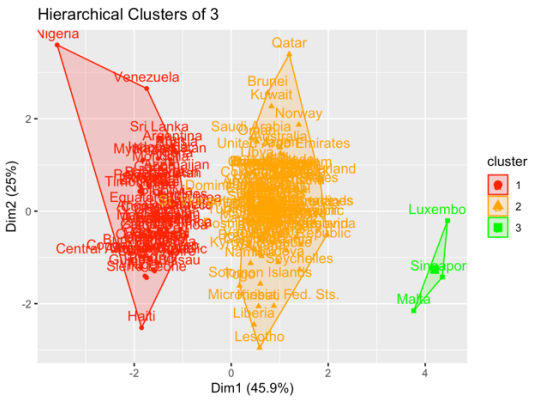

Figure 2.7: Hierarchical Clusters

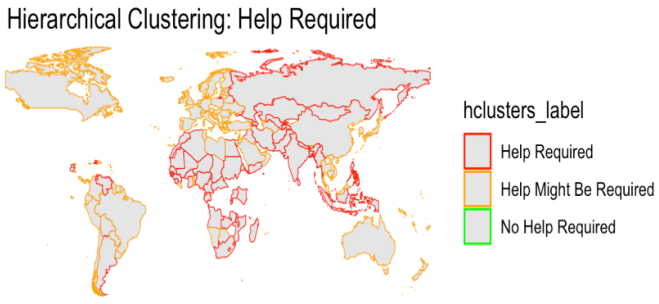

Figure 2.8: Hierarchical – World Map

Hierarchical clustering organizes similar data points into a hierarchical structure. Illustrated in Figure 2.7, the red cluster denotes countries requiring assistance, the orange cluster signifies those possibly requiring assistance, while the green cluster comprises of countries not in need. Figure 2.8 presents a visual depiction of these clustered countries on a world map, enhancing comprehension. Utilizing the PC1 score, a composite measure incorporating factors like child mortality rate, low income, GDP, life expectancy, and fertility rate, aids in identifying the cluster most in need of financial assistance.

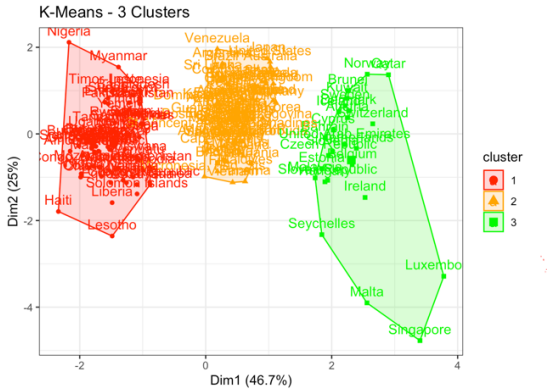

Figure 2.9: K-Means Clusters

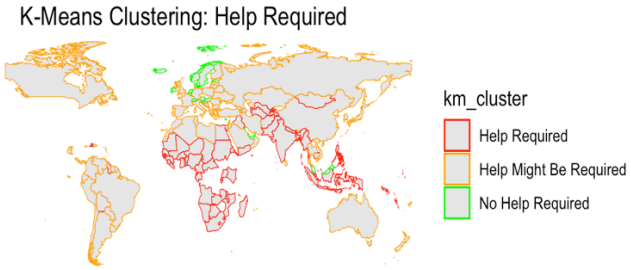

Figure 2.10: K-Means - World Map

In K-means clustering, our objective is to partition n observations into k clusters based on their proximity to cluster centroids. Analyzing the sum of squares within groups plot, we opt to partition the data into 3 clusters. The representation of each cluster in Figure 2.9 mirrors that of hierarchical clustering.

2.5 Results

RQ1: How many clusters are ideal for evaluating a country's level of development?
Based on the silhouette and elbow method, it is shown that 3 clusters was ideal with it being "Help Required", "Help Might Be Required" and "No Help Required".

RQ2: How should we interpret the clusters that have been identified?
For each clusters, the level of income, gpp, child mortality are assessed to determine if financial help is required. For cluster 3, the countries have the lowest income, gpp and highest child mortality rate therefore, they can be considered as the priority country that require assistance from HELP.

RQ3: Which countries are more likely to require financial assistance?
It is shown in Figure 2.8 and 2.10 that most African countries are in need of assistance where they are facing struggles stabilizing their economy (IMF, 2023)

3. California Houses (Regression)

Predicting California housing prices is essential for informed decision-making in real estate and urban planning. With various factors influencing the market, including economic conditions and population trends, developing an accurate predictive model is crucial for stakeholders to allocate resources effectively.

3.1 Research Questions

The research questions (RQ) that have been identified for the issue are as follows:
• RQ1: Which machine learning model performs the best in predicting housing price?
• RQ2: What are the prominent factors that influence housing price?

3.2 Methodology

After we perform exploratory data analysis (EDA) by visualizing the distribution of each numerical variable through histograms, we then split our dataset into training and testing set for model training and evaluation. Our analysis includes fitting a baseline linear regression model and training regularization techniques such as Ridge, Lasso, and Random Forest

regression using cross-validation to address potential multicollinearity and overfitting. The performance of each model is assessed on the test set, primarily using the R-Squared (R^2) and Root Mean Square Error (RMSE) as a metric to evaluate predictive accuracy, enabling us to identify the most effective model for predicting housing prices and understand the impact of various features on the housing market.

3.3 Dataset and Variables

The housing dataset consists of 20640 rows and 10 columns. Missing values are handled by imputing it with median while categorical features are converted to dummy variables. Columns are also renamed to lowercase for consistency and numeric features are scaled.

3.4 Analysis

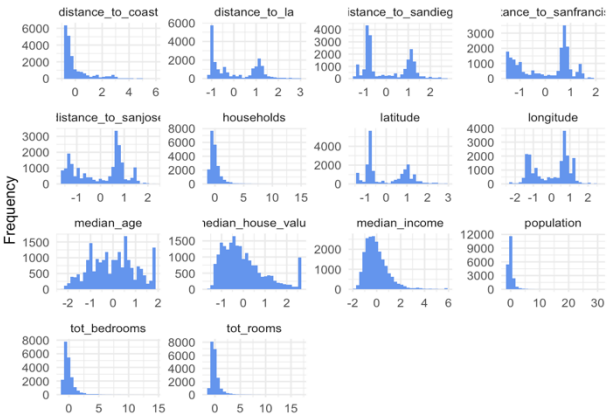

Figure 3.1: Distribution Plot

From Figure 3.1, it can be seen that most of the distributions are right-skewed, indicating that while the majority of observations fall within lower range values, there are significant outliers with very high values, particularly in variables like population, total rooms, and median house value. Geographic features like distance to major cities and coastlines show that many observations are concentrated close to these points, suggesting that geographical clustering could impact housing dynamics.

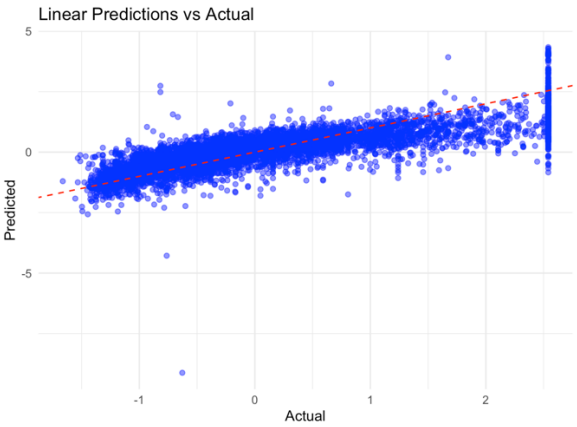

Figure 3.2: Linear Regression

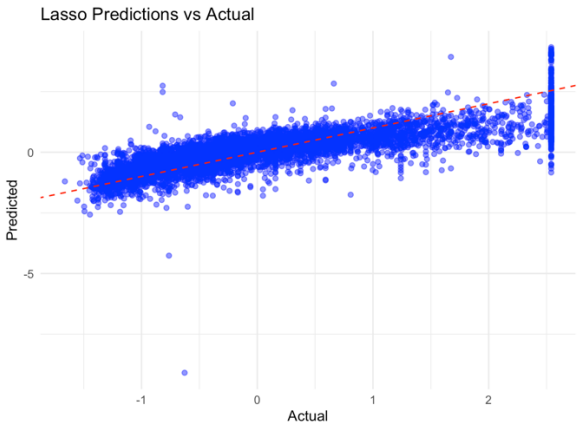

Figure 3.3: Lasso Regression

Both Lasso and Linear Regression model illustrates a general alignment with a few deviations. The dense clustering along the line y = x (represented by the red dashed line) suggests that for many observations, the Lasso model predicts values close to their actual values, indicating good model performance. However, there are noticeable outliers, particularly for higher actual values where the model seems to underpredict. The vertical spread of points, especially in the middle of the x-axis, indicates variance in the model's accuracy, with the spread increasing as actual values rise.

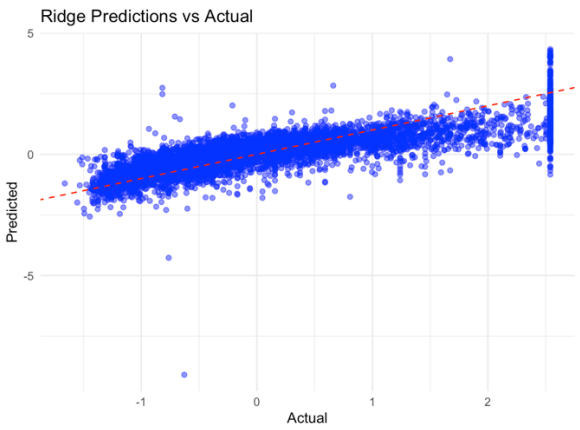

Figure 3.4: Ridge Regression

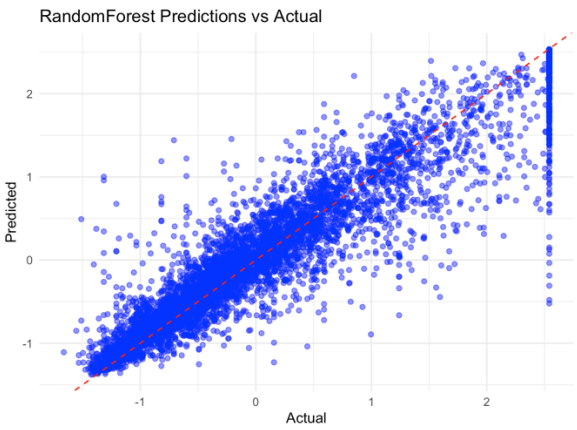

Figure 3.5: Random Forest

The Random Forest model shows a strong correlation between predicted and actual values, as seen by the close clustering of points around the line of perfect prediction. The plot reveals an effective prediction across the range but with a notable dispersion at higher values. The Random Forest model, leveraging its ensemble method, appears robust and versatile across different data points, although the increase in variance at the higher end of the spectrum suggests that extreme values are less accurately predicted.

The Ridge regression model, which includes regularization to reduce overfitting, predicts values that closely follow the actual values, indicated by the clustering around the diagonal line. The spread of predictions around this line is fairly uniform as seen in Figure 3.4, suggesting that the model handles variance in the data effectively. However, similar to other models, predictions for higher values show some discrepancies, hinting at potential limitations of the model in dealing with high-end values or possible anomalies in the data affecting model performance.

Method	RMSE	R^2
Linear	0.6016060	0.6395898
Ridge	0.6043852	0.6359997
Lasso	0.6011731	0.6400304
Random Forest	0.3822636	0.8548564

Figure 3.6: Accuracy Comparison

It can be seen that Random Forest model outperforms the rest where this indicates that the housing price variables are mostly non-linear and involve interactions that tree-based methods can exploit better than linear models. The modest differences between the linear models suggest that simple linear relationships are not sufficient to model the data accurately.

3.5 Results

RQ1: Which machine learning model performs the best in predicting housing price?
From Figure 3.6, The Random Forest model is substantially more effective with the lowest RMSE of 0.3823 and the highest $R^2$of 0.8549. This suggests that the Random Forest, with its ensemble approach, is much better at capturing complex patterns in the data, leading to higher accuracy, and explaining about 85% of the variance.

RQ2: What are the prominent factors that influence California housing price?

From Figure 3.1, it can be seen that geographic features like distance to major cities and coastlines playing a big part in affecting housing price.

4. Heart Disease Prediction (Classification)

The prediction of heart disease plays a pivotal role in the landscape of healthcare and preventive medicine, acting as a cornerstone in mitigating the global burden of cardiovascular diseases (Smith, 2020). Considering the critical importance of detecting heart disease accurately, we aim to focus on maximizing the model's sensitivity to the positive class—namely, its ability to detect all true cases of heart disease. Thus, the recall, or true positive rate, will be one of the key performance indicator for our model's effectiveness.

4.1 Research Questions

The research questions (RQ) that have been identified for the issue are as follows:

- RQ1: Which machine learning model performs the best in predicting heart disease?
- RQ2: Which health metric(s) are the most significant predictors of heart disease in the population studied?

4.2 Methodology

Exploratory Data Analysis is performed to understand the variables' relationship and distribution. After one-hot encoding is performed, classification techniques, such as K-Nearest Neighbour and Random Forest are used to compare the predictive performance and to find the best model to predict the presence of heart disease in patients. The dataset will be split into train set and test set with a ratio of 70:30 respectively.

4.3 Dataset and Variables

The heart dataset consists of 303 rows and 14 columns. There are no missing values or duplicated data. Based on EDA, no columns seem irrelevant and given the small size of our dataset, we will be applying transformations like Box-Cox to reduce the impact of outliers instead of opting for direct removal approach.

4.4 Analysis

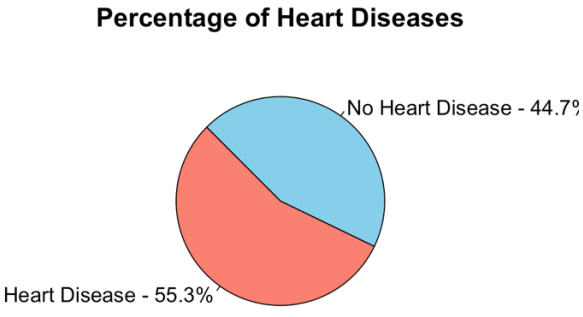

Figure 4.1: Pie Chart: Heart Diseases %

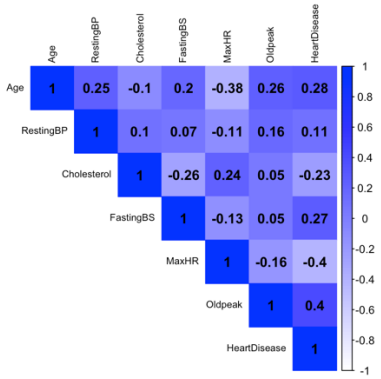

Figure 4.2: Correlation Matrix

As shown in Figure 4.1, the dataset is almost balanced with 45% not having heart disease and 55% having heart disease. In the correlation matrix, there is a moderate positive

correlation between old peak and heart disease, signifying that age is one of the important factors relating to heart disease.

$$Precision = \frac{True\ Positive}{True\ Positive+False\ Postitive}\ Recall = \frac{True\ Positive}{True\ Positive+False\ Negative}\ F1\ Score = 2 * \frac{Precision*Recall}{Precision+Recall}$$

Given the close to balanced nature of the dataset, both the F1 score and ROCAUC will serve as evaluation metrics. The F1 score, encompassing precision and recall, will primarily assess the classifier's performance. Additionally, ROCAUC will complement the evaluation, providing insight into the model's accuracy in predicting true positives and negatives. As for the confusion matrix, True Negative denotes instances where no heart disease is present and correctly predicted, while True Positive signifies the accurate prediction of heart disease. False Positive indicates cases where no heart disease exists but mistakenly predicted as such and False Negative represents heart disease cases mistakenly predicted as absent.

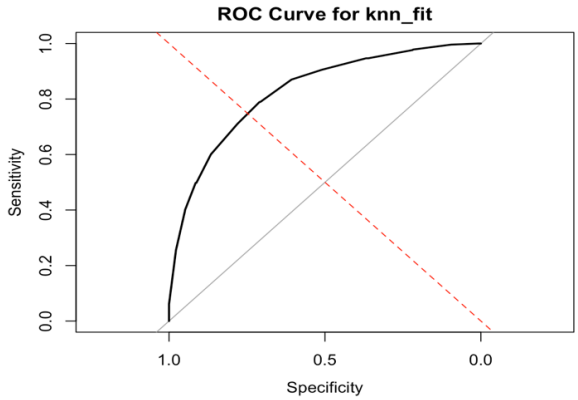

Figure 4.3: K-Nearest Neighbours (KNN)

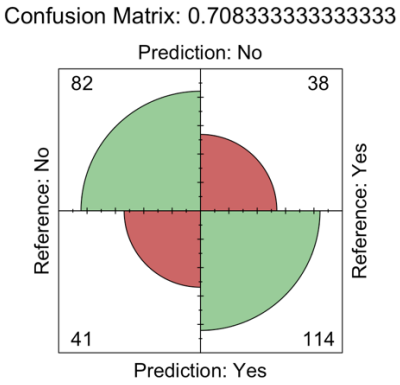

Figure 4.4: Confusion Matrix

K-Nearest Neighbours involves computing the Euclidean distance for a specified number, K, of neighbouring points. From Figure 4.9, The F1 score stands at 0.67, while the ROCAUC score reaches 0.75. Examining the confusion matrix depicted in Figure 4.4, there were 38 instances of False Positives and a count of 41 False Negatives.

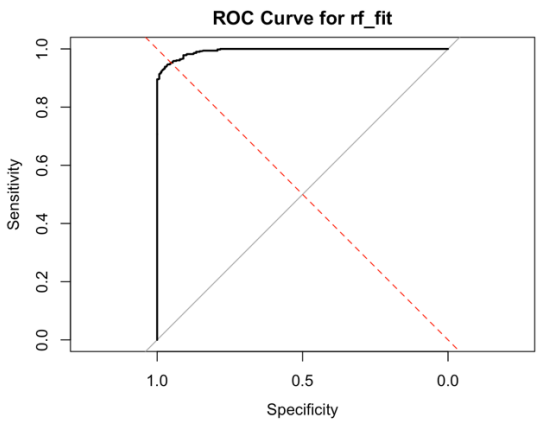

Figure 4.5: Random Forest

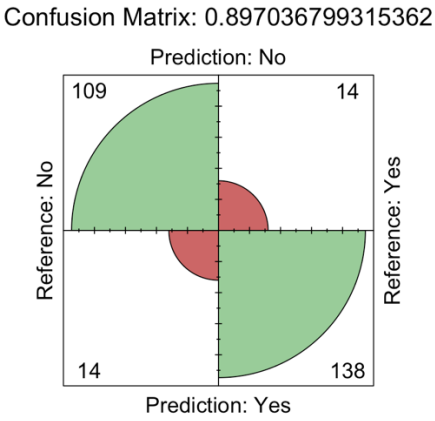

Figure 4.6: Confusion Matrix

Random Forest achieves a F1 score of 0.88, alongside a ROCAUC score of 0.94 where the confusion matrix in Figure 4.8 shows a False Negatives of 14 and False Positives of 14. Notably, both the F1 score and ROCAUC score surpass those of KNN.

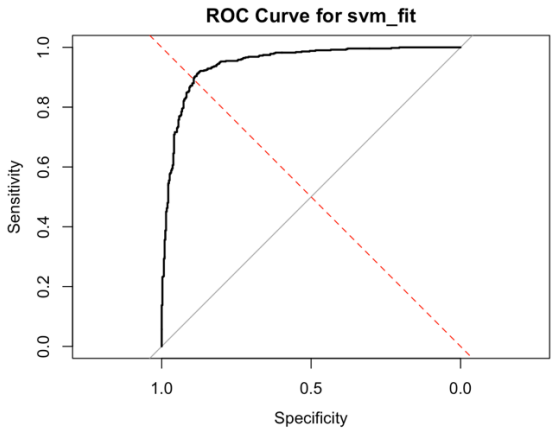

Figure 4.7: Support Vector Classifier

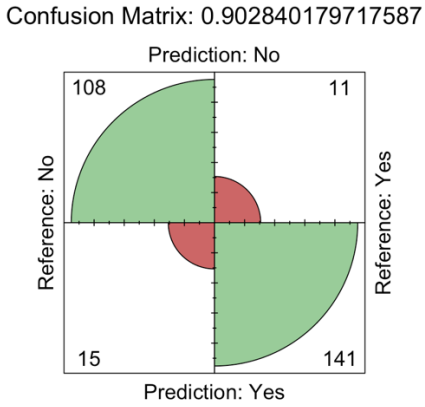

Figure 4.8: Confusion Matrix

Support Vector Machine employs data points to construct a hyperplane that effectively segregates and categorizes the data into two distinct classes. Achieving an F1 score of 0.89 and a ROCAUC score of 0.94, this model demonstrates notable performance (Figure 4.9). Figure 4.6's confusion matrix indicates 15 instances of False Negatives and 11 False Positives.

Model	F1	ROCAUC	Accuracy	TimeTaken
K-Nearest Neighbors	0.6721311	0.7520860	0.7090909	0.02658606
Random Forest	0.8861789	0.9476359	0.8981818	0.03666401
Support Vector Machine	0.8925620	0.9508451	0.9054545	0.07860780

Figure 4.9: Summary

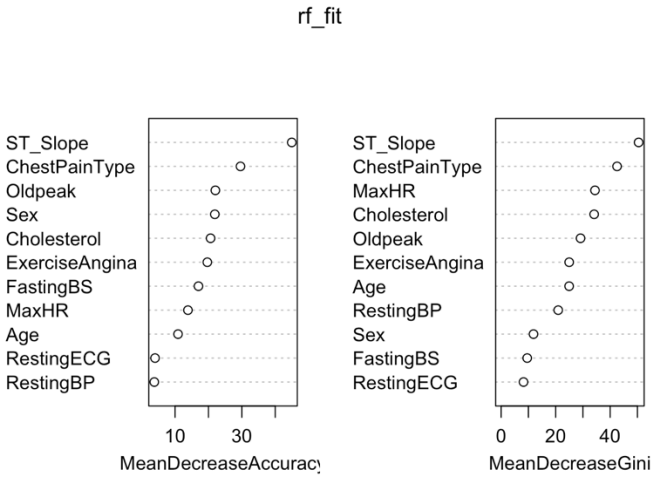

Figure 4.10: Feature Importance plot

4.5 Results

RQ1: Which machine learning model performs the best in predicting heart disease?

KNN model should be eliminated due to its lowest F1 score, while the Random Forest Classifier emerges as the preferred overall model with a notable F1 score, ROCAUC, Accuracy and Time Taken indicators, rendering it the optimal choice for heart disease prediction. Although the support vector machine indexes performed slightly better, its longer runtime compared to the Random Forest model deems the latter more favourable for practical implementation. Therefore, Random Forest is selected as the preferred predictive model.

RQ2: Which health metric is the most significant predictors of heart disease in the population studied?

Figure 4.10 shows the significance of the different variables as predictors. ST_Slope has the highest importance, indicating its strong influence on both the accuracy of the model and in reducing node impurity, followed by ChestPainType and MaxHR.

References
Smith, E., 2020. *Ergothioneine is associated with reduced mortality and decreased risk of cardiovascular disease.* [Online]
Available at: https://pubmed.ncbi.nlm.nih.gov/31672783/
[Accessed 18 March 2024].
IMF, 2023. *IMF Ghana / Ethiopia / Zambia.* [Online]
Available at: https://mediacenter.imf.org/news/imf-ghana---ethiopia---zambia/s/5921f6db-e27f-4d8d-a66c-066254a53d05
[Accessed 30 March 2024].

Unsupervised Learning Dataset (Country):
https://www.kaggle.com/code/maryamnoroozi68/unsupervised-learning-on-country-data/input

Regression Dataset (California Houses):
https://www.kaggle.com/code/evillionkeng/california-housing-prices-data-extra-features/input

Classification Dataset (Heart Disease):
https://www.kaggle.com/code/farzadnekouei/heart-disease-prediction/input#Step-3.1-|-Dataset-Basic-Information