In advance of your submission, please:

1. Complete the following details and include as the first page of your submission;
2. Read the Turnitin section and check that you understand how Turnitin is used to assess your work;
3. Read the declaration to check that your submission conforms with the listed requirements before you submit your work.

Name(s) and Student Number(s):	N0964115
Module Title:	Principles of Machine Learning in Finance ACCA40671
Title of Coursework:	Comparison of different Machine Learning Models for Predicting heart disease.
Word Count (see declaration below):	2885 words

Required Format

Font: Verdana

Points: 10

Paragraph line spacing: 1.5

Page Numbers: Included

Turnitin Similarity Check

Where appropriate to the coursework assignment, your document will be submitted to Turnitin to generate a similarity report for review by your tutor. This report will compare your work against millions of previously submitted student papers and online resources (e.g. web sites, journal articles) in the Turnitin database and will highlight any text that matches your submission.

Declaration

By submitting your work you are certifying that:

1. The submission is the result of your own work and does not contravene the University Regulation on Academic Integrity. You must ensure that you have referred to valid sources of information to support your work, and that these are properly referenced in the required format (i.e. using Harvard Referencing style). This includes, but is not limited to, use of Generative Artificial Intelligence (GenAI), such as the ChatGPT app, to complete all or sections of your work.

Please select **ONE** of the options below which is the most appropriate. You must select at least **ONE** option:

☐ No content generated by AI tools have been used in the development of my work.

☐ I acknowledge the use of AI generated materials for background research and self-study in the drafting of this assessment. *If this option has been selected, please retain your outputs as these could be requested by the tutor grading your work.*

☐ I acknowledge the use of AI in my final assessment, and I have made it clear that this has not been included as my own work by referencing the source correctly. *If this option has been selected, please retain your outputs as these could be requested by the tutor grading your work.*

2. The word count included on this cover sheet is accurate and follows the guidelines outlined in the assignment brief (failure to include an accurate word count will be treated as a minor academic offence as defined in the Academic Integrity Policy).
3. Your ability to complete your assessment has NOT been adversely impacted by circumstances beyond your control.
4. Once you have submitted your work, any such circumstances would need to be disclosed through the Academic Appeals Policy and process and not through the Notification of Extenuating Circumstances Policy and process.

Comparison of different Machine Learning Models for Predicting heart disease.

Introduction

At the fundamental core of modern medicine, accurate and easily accessible diagnoses is vitally important in ensuring timely treatment, improving patient outcomes, and reducing the burden of diseases on both individuals and healthcare systems. Richens, Lee, Johri (2020) Noted that in the United States alone approximately 5% of outpatients receive a wrong diagnosis every year, a common trend being seen with diagnoses of serious medical conditions. With that said 20% of such are misdiagnosed at the level of primary care resulting in one in three misdiagnoses causing serious patient harm. This void in effective diagnoses of various diseases stems from the complexity of different disease mechanisms and underlying symptoms of the patient population Ahsan, Luna, Siddique (2022). In the medical industry, machine learning primarily focuses on developing adaptive algorithms and techniques to determine a disease diagnosis based on an individuals' symptoms. Whilst typically this is done through patient history and physical examination, it is increasingly difficult as symptoms are ambiguous and can only be diagnosed by trained medical professionals. In areas such as developing countries, individuals may face difficulty when seeking medical aid due to a lack of health professionals, therefore shedding light on the need for such Machine learning algorithms for medical diagnosis.

According to the World Health Organization (WHO), cardiovascular disease (also referred to as heart disease) is the leading cause of death globally. It is estimated that 17.9 million deaths are attributed to cardiovascular disease annually. Within this figure 4 out of 5 deaths are due to heart attacks and strokes with one-third of these deaths occurring prematurely in people under 70. Effective ways of identifying individuals who are at highest risk of cardiovascular disease is imperative so that it can be ensured that they receive appropriate treatment in a timely manner to prevent these premature deaths.

The aim of this study is to classify the target variable using different Machine Learning Models and show which of the algorithms is best at predicting the possibility of an individual having cardiovascular disease. The comparative analysis is done between, Logistic Regression, Decision Tree Learning and Random Forest Learning models. The dataset is taken from Kaggle and feature selection technique Recursive Feature Elimination with Cross-Validation is applied to the dataset. The objective of this paper can be achieved by evaluating the models against their respective, confusion matrixes, F1, Recall, Accuracy and Precision scores. Sections in the paper are as follows: Section 2 will review the existing literature in the field of Logistical regression, Decision Tree Learning, Random Forest and other machine learning models within the field of medical diagnoses. Section 3 presents and in-depth exploratory analysis of the dataset. Section 4 will describe the research methodology of the analysis. Section 5 details the outcomes of our analysis, while section 6 offers a summary of our conclusions.

Literature Review

The use of machine learning to predict cardiovascular disease is not new. Researchers have applied various models to try and capture the relationships between demographic, physiological and clinical factors, with the aim to the ultimately enhance early detection and improve patient healthcare. Whilst extensive research has been undergone to find the best predictive model for cardiovascular disease there has not been a unanimous consensus on which predictive model best achieves a balance between predictive accuracy, interpretability, and clinical applicability.

Rajdhan et al. (2020) recognised the need to develop a predictive system that was both effective and accurate. To achieve this, they set out to identify the most efficient machine learning algorithm through a comparative study of several models using the UCI repository dataset – a methodology like that employed by Shah et al. (2020) and Asif et al. (2021) Their research ultimately revealed that the Random Forest model outperformed the other, achieving an accuracy score of 90.16%

Sharma et al. (2020) agreed with Rajdhan et al. (2020) in their comparative analysis of various machine learning models. Using the Cleveland heart disease dataset, Sharma et al.'s findings corroborated that the Random Forest algorithm achieved the highest accuracy at 99%, while Decision Trees recorded a lower accuracy of 85%.

Shah, Patel, Bharti (2020) implement a variety of supervised learning algorithms such as Naïve Bayes, Decision trees and K-nearest neighbour on a well-established Cleveland database of UCI repository of heart disease patients. Of the 76 attributes available only 14 were considered for testing. Their results concluded that K-nearest neighbour achieved the highest accuracy score in successful predictions of heart disease.

Asif et al. (2021) contributed further to the early detection of cardiovascular disease with their research into twelve machine learning algorithms. In parallel to Shah el al., the UCI dataset was also used to assess the predictive accuracy of the twelve various machine learning algorithms. Asif et al., research broadened the scope of early cardiovascular disease detection with a more comprehensive comparison on the UCI dataset in comparison to Shah et al., earlier research. Their findings concluded an accuracy of 92% was found in both hard and soft voting ensemble classifiers (EVCH and EVCS) Asif et al., (2021). However, they did remark that the Adaboost algorithm outperformed the ensemble classifiers in both precision and specificity.

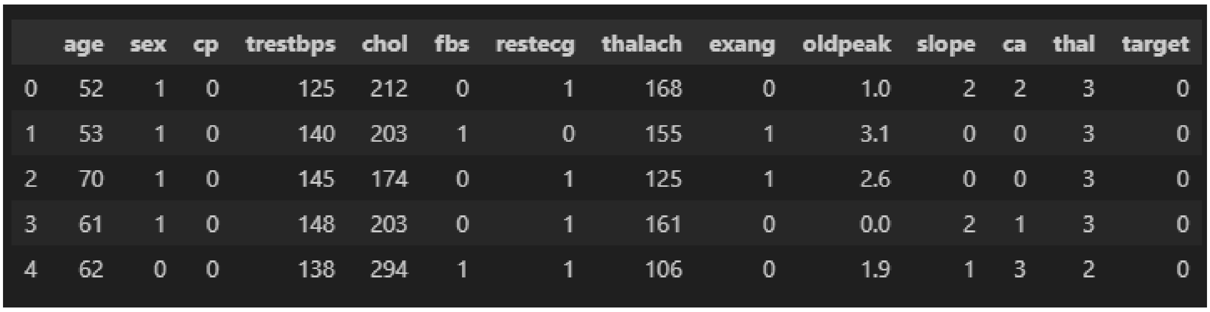
Arumugam et al. (2021) highlighted the limitations of current data mining approaches in healthcare for accurately predicting heart disease in diabetic individuals. To address this gap, they fine-tuned a decision tree model, which consistently outperformed both the Naïve Bayes and Support Vector Machine models in their comparisons.

In addition, Mahendran et al. (2023) evaluated several models with the goal of developing a heart disease prediction model that is both more accurate and reliable. Having employed the Kaggle Framingham dataset, the most accurate model was found to be logistical regression with an accuracy rate of 92%.

Methodology

The most significant role in any machine learning model belongs to the dataset. In this paper, we have used a Heart Disease Dataset from Kaggle for the training and testing of our models. The Heart Disease Dataset consists of 14 Features and 1025 data-points. Table 1 shows a sample of the dataset.

Table 1. First 5 rows of Heart Disease Dataset.



	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	52	1	0	125	212	0	1	168	0	1.0	2	2	3	0
1	53	1	0	140	203	1	0	155	1	3.1	0	0	3	0
2	70	1	0	145	174	0	1	125	1	2.6	0	0	3	0
3	61	1	0	148	203	0	1	161	0	0.0	2	1	3	0
4	62	0	0	138	294	1	1	106	0	1.9	1	3	2	0

Visual Studio Code was used as a coding platform in our research. We first examined the dataset for missing values and found none, ensuring that subsequent analyses were not skewed by incomplete data. Figure 1 illustrates the missing values heatmap.

Figure 1. Missing values Heatmap

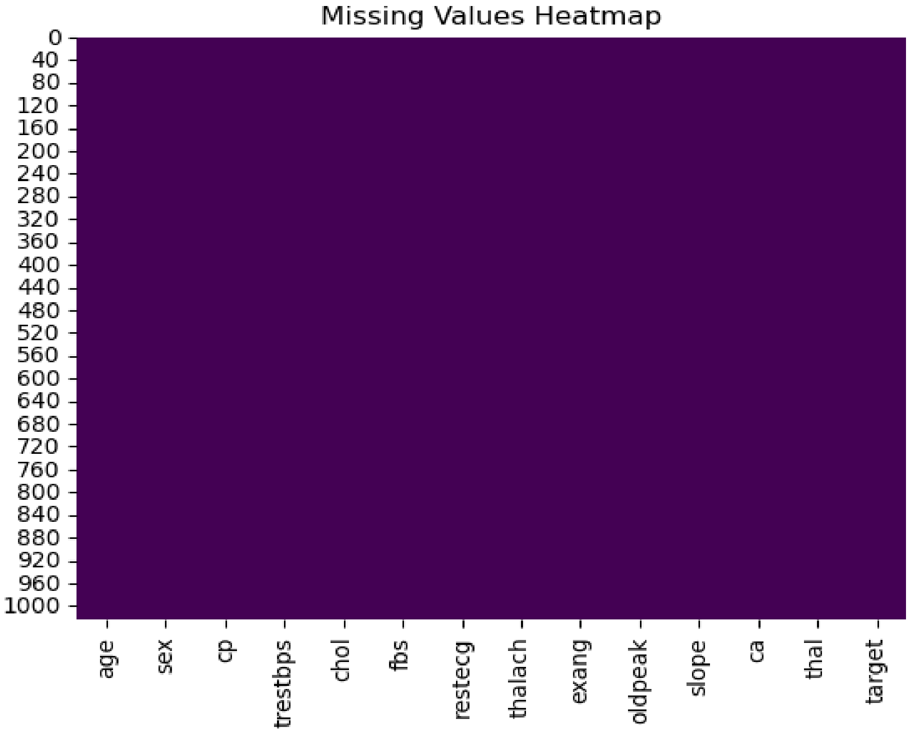
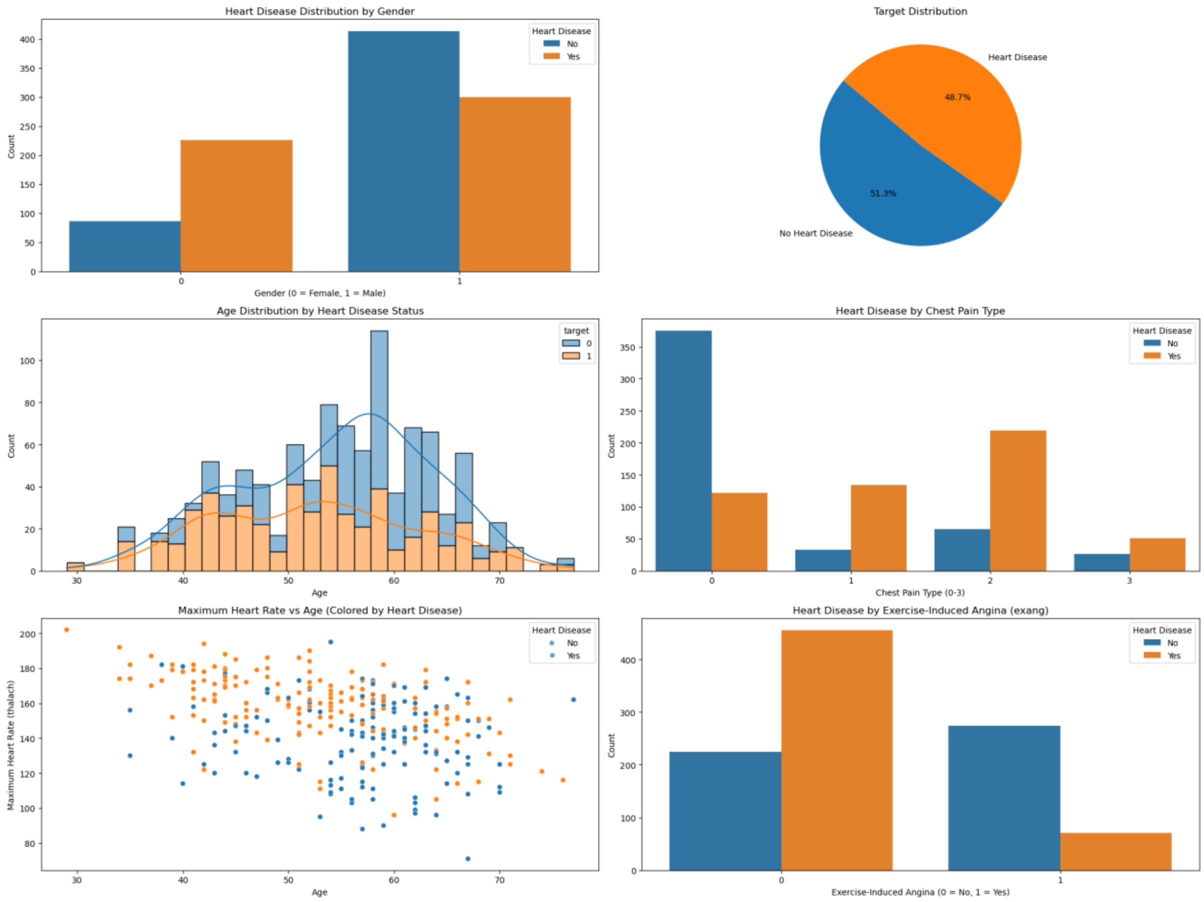


Figure 2. Explanatory Data Analysis



The visualisation presented here in Figure 2, offers a comprehensive exploration into the distribution and interrelationships amongst key variables of the heart disease dataset. A description of each subplot and their importance is given below:

Subplot 1; depicts ‘Heart Disease Distribution by Gender’. This highlights any disparities between heart disease prevalence and the two genders. Mapping the counts by gender reveals that the disease disproportionately affect males compares to females.

Subplot 2; utilises a pie chart to detail the overall target distribution. This visualisation is crucial in understanding the class balance within the dataset which have significant implications for both model training and evaluation.

Subplot 3; examines the ‘Age Distribution by Heart Disease Status’ with 1 (orange) meaning positive for the presence of heart disease and 0 (blue) indicating a negative presence of heart disease. This visual aid allows for an immediate visual comparison of how age correlates with heart disease.

Subplot 4; displays of count plot categorising heart disease by Chest pain type. The differing chest pain types 0-3 are examined to determine the frequency of heart disease associated to the respective types of chest pain. The visualisation shows that heart disease is most prevalent in chest pain types 1,2 and 3 with individuals suffering from type 2 chest pains having the highest count of heart disease. This visualisation reinforces the clinical importance of chest pain as a symptom and serves as an initial check for potential for potential predictors in a predictive model.

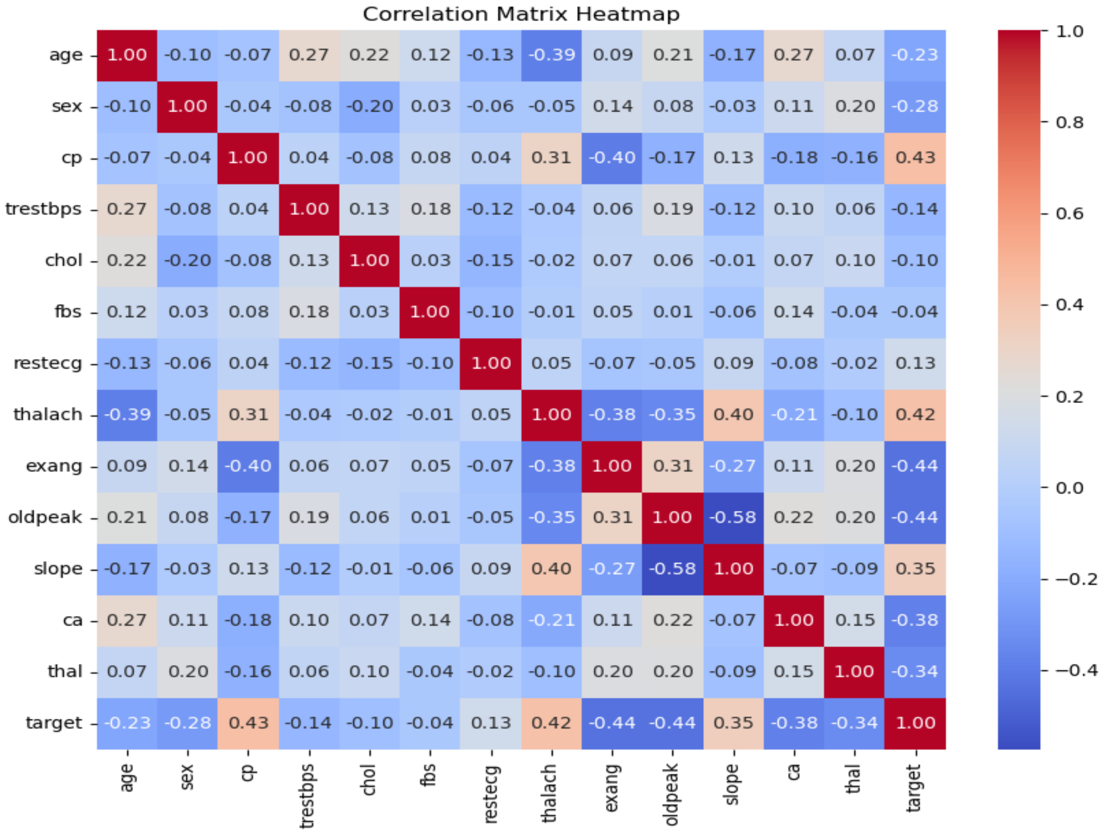
Subplot 5; a scatter plot that examines the relationship between age and maximum heart rate (thalach). The points are colour coded according to heart disease status. The dual-variable plot offers insight into the inverse relationship often observed between age and maximum heart rate but also helps to distinguish any distinct patterns linking cardiac performance to heart disease.

Subplot 6; investigates heart disease by exercise-induced angina (exang). Exercise-induced angina indicates a type of chest pain that occurs when your heart is requiring more oxygen than usual but is unable to do so because of heart disease. Since exercised-induced angina is a known risk factor, visualising its distribution among patients with and without heart disease offers further validation of clinical expectations and may guide our feature selection for predictive modelling.

Collectively, these subplots offer a comprehensive snapshot of the data underpinning the rationale for subsequent model development. The exploratory data analysis represented by these graphs highlights key patterns that inform both feature selection and the overall modelling strategy.

To guide our feature selection process and ensure the efficiency of our machine learning models the correlation heatmap was plotted to visually assess the pairwise relationships among the dataset features. Figure 3 illustrates the correlation matrix heatmap.

Figure 3.) Correlation Matrix Heatmap

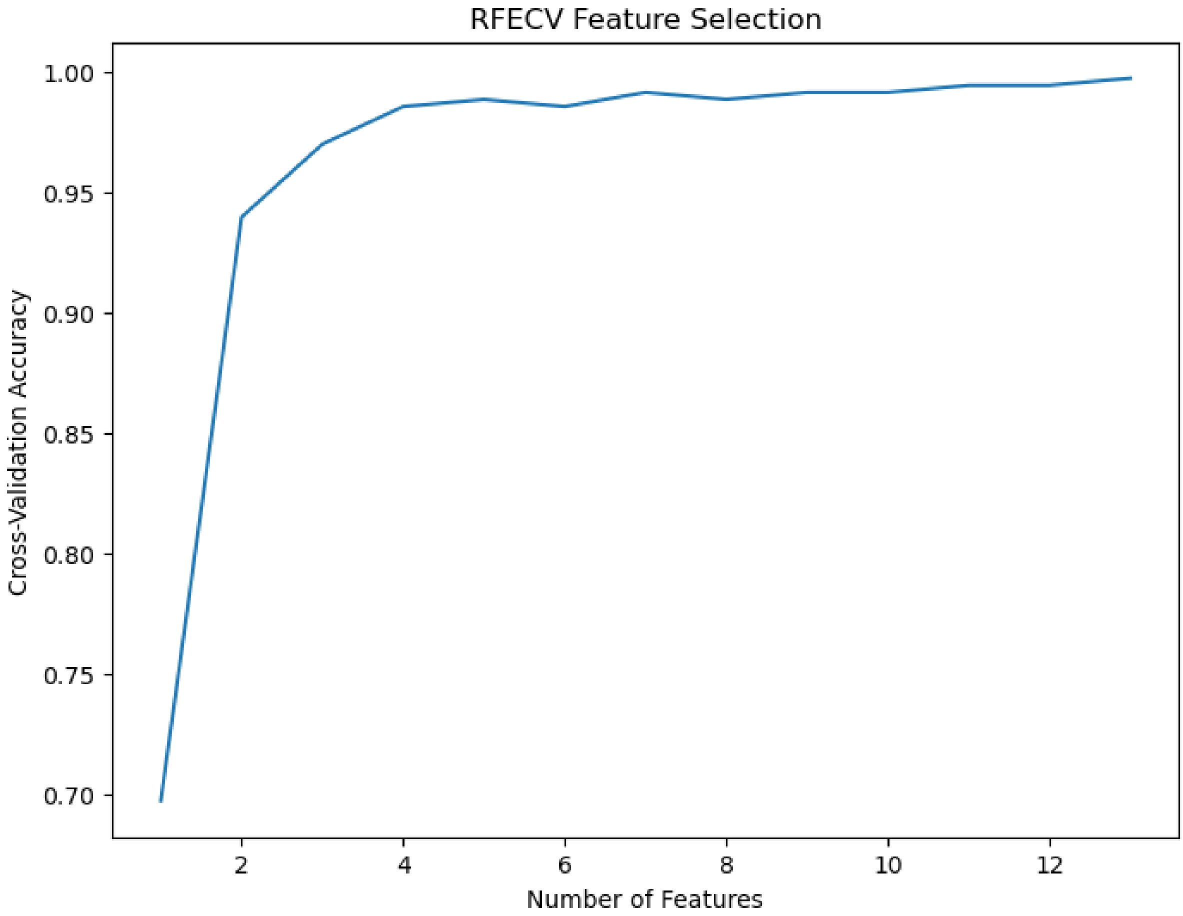


The visualisation has helped identify potential multicollinearity between and the features and understand the strength and direction of the relationships between variables. According to the correlation heat map features, cp, thalach, slope and restecg are highly correlated with the target variable.

After visualisation and analysis, the data was processed via a recursive feature elimination with 5-fold cross-validation (RFECV) to select the most optimal features. The RFECV found 11 optimal features to be used in our analysis. The features to be removed from our analysis are fbs and restecg suggesting that these features are either redundant or contributed noise.

Figure 4 illustrates the RFECV plot. This shows the models cross-validation performance as features are eliminated. The x-axis represents the number of features retained in the model at each iteration of the recursive elimination process, whilst the y-axis shows the corresponding cross-validation score. The removal of these features is expected to reduce complexity, mitigate overfitting and enhance predictive accuracy.

Figure 4. RFECV Feature Selection



To standardise the dataset an ensure that all features contributed equally to the learning process of the model, A MinMax scaler was applied. This Normalised the 11 features to a 0,1 range to ensure that no single feature disproportionately influenced the model by its magnitude improving the predictive accuracy of the models.

We then split the data into a training set and testing set utilising 70% of the total data as our training set and 30% of the total data as our testing set. Three machine learning algorithms were investigated. Performance of the prediction models are assessed using measures such as accuracy, precision, recall, F1 score and confusion matrixes. The evaluation measures capture how accurate the machine learning models predications appear. They are represented mathematically by the following equations.

Equation 1.

$$Precision = \frac{True\ Positive\ (TP)}{True\ Positive\ (TP) + False\ Positive\ (FP)}$$

Equation 2.

$$Recall = \frac{True\ Positive\ (TP)}{True\ Positive\ (TP) + False\ Negative\ (FN)}$$

Equation 3.

$$Accuracy = \frac{True\ Positive\ (TP) + True\ Negative\ (TN)}{TP + TN + FP + FN}$$

Equation 4.

$$F1\ Score = 2\ \cdot \frac{(Precision \cdot Recall)}{Precision + Recall}$$

- (TP) True positive: the patient has the disease and the test is positive.
- (FP) False Positive: the patient does not have the disease, but the test is positive.
- (TN) True Negative: the patient does not have the disease, but the test is negative.
- (FN) False Negative: the patient has the disease and the test is negative.

Machine Learning Algorithms

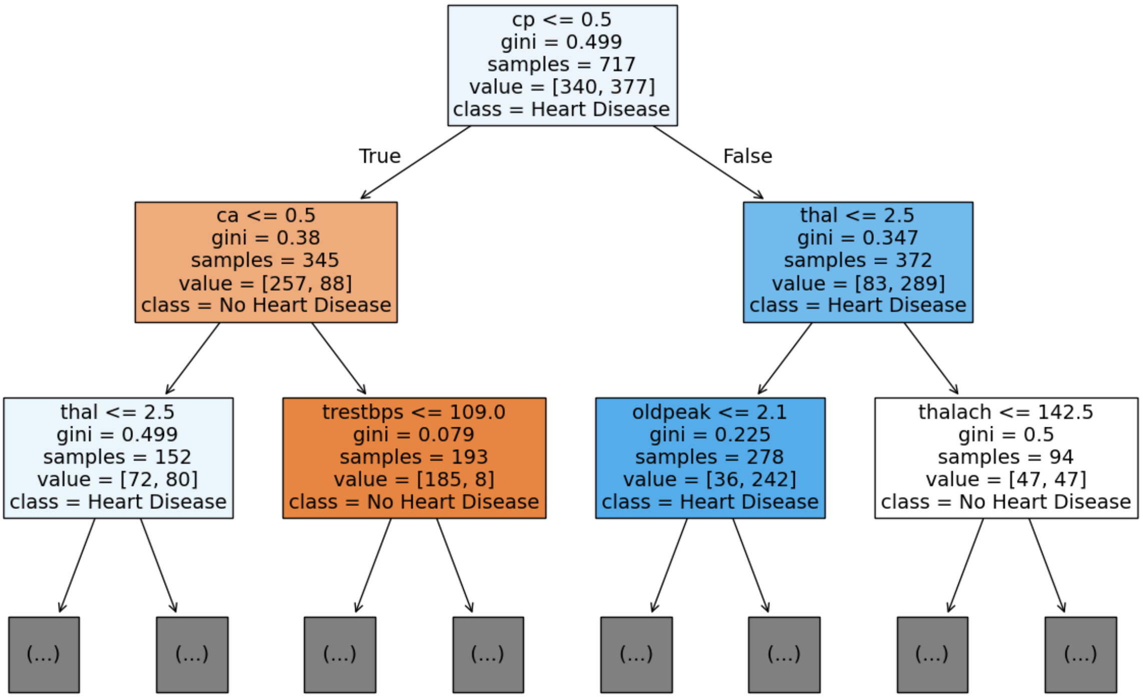
Logistic Regression

Logistic Regression is a widely used classification algorithm, they are particularly well-suited for binary outcomes. Rather than fitting a straight line like linear regression, this method employs the sigmoid function to constrain outputs between 0 and 1, effectively converting them into probabilities. It models the relationship between one or more independent variables and a categorical dependent variable by applying the logit transformation—essentially forecasting the log-odds of the target outcome. This process enables the algorithm to estimate the probability that an instance belongs to a particular classification. One of the key merits of Logistic Regression is its interpretability, as it provides clear probability estimates that facilitate informed decision-making. However, the model can encounter challenges, such as complete class separation, where the predictors perfectly discriminate between classes, potentially affecting the stability of the estimates. Overall, Logistic Regression remains a fundamental and effective tool within machine learning for addressing classification problems in various domains.

Decision Tree

Decision Trees provide a straightforward, interpretable approach to classification by representing the decision process as a flowchart as depicted in figure 5. In this model, each inner node corresponds to a dataset attribute, while the branches represent the outcomes, culminating in leaf nodes that deliver the final classification. Their inherent simplicity and minimal data preparation requirements make them highly attractive for handling both categorical and numerical data. The decision-making process starts at the root node, where the value of a key attribute is compared against each observation. Based on metrics such as maximum information gain or minimum entropy, subsequent branches are followed until a leaf node is reached. For instance, in our analysis 1025 datapoints were navigated through the tree to predict heart disease, with the model's performance evaluated in terms of F1, Recall, Precision and Accuracy scores. To acquire efficient prediction, hyperparameter optimisation through GridSearchCV was conducted which identified the optimal set of parameters to be a maxdepth of 11. Although decision trees often yield high accuracy and are favoured for their interpretability, they may sometimes suffer from overclassification, as only one attribute is assessed at each decision point.

Figure 5. Baseline Decision Tree with maxdepth = 2



Random Forest

Random Forest is a robust supervised learning technique applicable to both classification and regression tasks. By building an ensemble of decision trees, the algorithm aggregates their predictions—typically through majority voting—to generate a final output. This process not only mitigates the overfitting that can challenge single decision trees but also ensures reliable performance even when dealing with large datasets or missing values. Essentially, Random Forest leverages methods such as random input selection and blending, which help tune its performance and improve accuracy. Overall, the Random Forest algorithm's ability to combine multiple predictive models makes it a compelling choice for developing reliable, high-accuracy predictive systems in the healthcare domain.

Analytical Findings

The results obtained by applying Logistic Regression, Decision Tree and Random Forest are shown in this section. In the experiment the pre-processed dataset is used to carry out the experiments and the machine learning models are applied. The previously mentioned evaluation metrics are obtained using the confusion matrix. A confusion matrix is a graphical representation of how accurate a classifier is at predicting the labels for a categorical variable. In our experiment that is how accurate our models are at predicting correctly whether an individual has or does not have a heart disease. The confusion matrix obtained by the proposed models applied to our processed dataset is shown in Figure 6. The accuracy score obtained from the Logistic Regression, Decision Tree and Random Forest classification techniques is shown in Table 2.

Examining the confusion matrixes, we can see that the Random Forest classifier model reveals exceptional predictive performance and is the highest performer out of the models. Out of 308 samples, the model correctly identified all 159 negative cases. Additionally, it correctly classified 146 out of 149 positive cases. Overall, the classifier achieved an accuracy of about 98.7%, with 305 correct predictions out of 308 samples. The absence of any false positives further highlights the model's reliability, as it achieved a precision of 100% for the positive class. This high performance indicated by the models respective F1 score, demonstrates that the Random Forest algorithm minimised misclassifications effectively.

The comparative evaluation of these models as shown in figure 2 reveals distinct performance differences in predicting heart disease. The Random Forest model stands out with an F1 score of 0.971, recall of 0.978, precision of 0.963, and an overall accuracy of 96.9%, indicating that it excels in both detecting positive cases and minimizing false alarms. In contrast, the Tuned Decision Tree model, while still robust, registers slightly lower scores across the board with an F1 of 0.939, recall of 0.925, precision of 0.955, and an accuracy of 93.8%, suggesting it may miss a few more positive cases compared to the ensemble approach.

Logistic Regression falls behind these tree-based methods with an F1 score of 0.819, recall of 0.879, precision of 0.766, and an accuracy of 81.2%, implying that its linear boundaries might not capture the data complexity as effectively. Meanwhile, the Base Decision Tree model achieves an impressive 100% precision and 97.1% accuracy, paired with an F1 score of 0.969 and recall of 0.94. Although its perfect precision is praiseworthy, the slightly lower recall compared to Random Forest CV indicates some sensitivity is sacrificed.

These comparisons highlight the superior performance of tree-based approaches especially Random Forest over Logistic Regression in this application.

Figure 6. Correlation Matrix

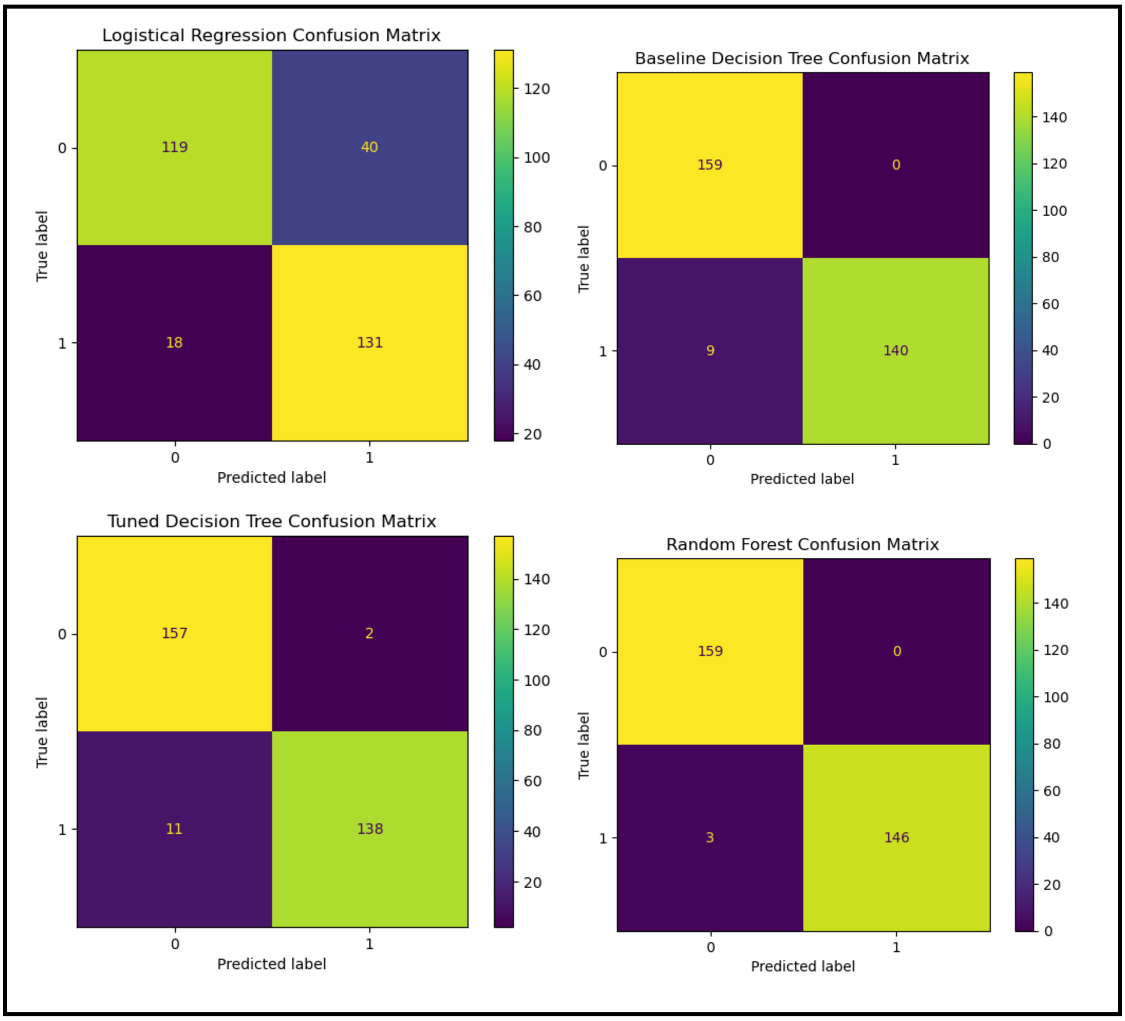
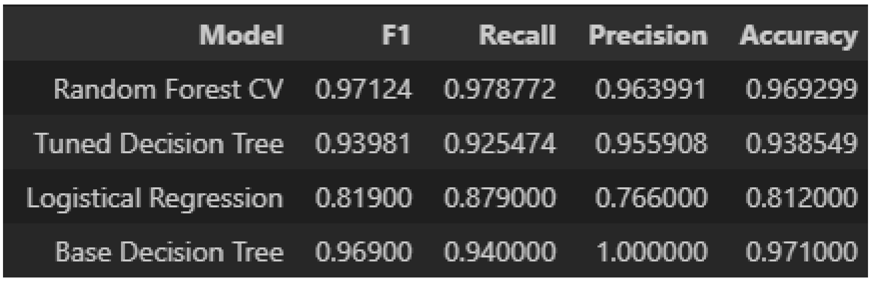


Table 2. Analysis of Machine Learning Algorithms



Model	F1	Recall	Precision	Accuracy
Random Forest CV	0.97124	0.978772	0.963991	0.969299
Tuned Decision Tree	0.93981	0.925474	0.955908	0.938549
Logistical Regression	0.81900	0.879000	0.766000	0.812000
Base Decision Tree	0.96900	0.940000	1.000000	0.971000

Conclusion

In conclusion, this paper provides a comparative analysis of the effectiveness of various machine learning models on detecting heart disease within individuals. The analysis reveals that Random Forest modelling proves to be the most effective when compared to Decision Tree and Logistical Regression models. These findings focuses in on the potential of ensemble-based approaches for reliable clinical classification and highlights the importance of careful model selection to guide feature research and practical applications in healthcare.

References:

Ahsan, Md., Luna, S., Siddique, Z., 2022, Machine Learning Based Disease Diagnosis: A Comprehensive Review

Arumugam, K., Naved, M., Shinde, P., Leiva-Chauca, O., Huaman-Osorio, A., Gonzales-Yanac, T., 2021. Multiple disease prediction using Machine Learning algorithms, Materials Today: Proceedings, https://doi.org/10.1016/j.matpr.2021.07.361

Asif, Md., Nishat, M., Faisal, F., Dip, R., Udoy, M., Shikder, Md., Ahsan, R., 2021, Performance Evaluation and Comparative Analysis of Different Machine Learning Algorithms in Predicting Cardiovascular Disease, Engineering Letters.

Mahendran, K., Dharshini, J., Dharshini, S., Anitha, A., 2023, Comparative Analysis Of Cardiovascular Disease Using Machine Learning Techniques.

Patel, Jaymin, Dr TejalUpadhyay, and Samir Patel, "Heart disease prediction using machine learning and data mining technique," Heart Disease, vol. 7, no.1, pp 129-137, 2015

Rajdhan, A., Agarwal, A., Sai, Milan., Ravi, D., Ghuli, P., 2020, Heart Disease Prediction using Machine Learning, International Journal of Engineering Research & Technology, Vol. 9.

Richens, J., Lee, C., Johri, S., 2020, Improving the accuracy of medical diagnosis with causal machine learning, Nature Communications.

Shah, D., Patel, S., Bharti, S.K., 2020. Heart Disease Prediction using Machine Learning Techniques. SN Computer Science.

Sharma, P., Agarwal, S., 2024. Cardiovascular Disease Analysis Using Different Machine Learning Techniques.

Sharma, V., Yadav, S., Gupta, M., 2020, Heart Disease Prediction Using Machine Learning Techniques.

Singh, A., 2023, Comparative Performance Analysis of Cardiovascular Disease Prediction by Machine Learning Techniques.

World Health Organization (WHO). (n.d.) Cardiovascular diseases. Available at: https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1 (Accessed: 13 March 2025)

Appendices.

Dataset; Heart Disease DataSet.csv

Processed data; df_heart.csv

Machine Learning Code; Data_Modelling.ipynb