B8IT156
Data and Network Mining
CA Group Project

Module Title: Data and Network Mining

Module Code: B8IT156

Lecturer Name: Terri Hoare

Student Names: Carla Cubes, Rodrigo Dominguez , Siobhan Hennessy

Table of Contents

1. Overview

The aim of this project was to analyse a public data set using the CRISP-DM methodology. Our group chose a data set with a classification problem and the project was broken down into tasks to be completed by each of the team members as described in the figure below.

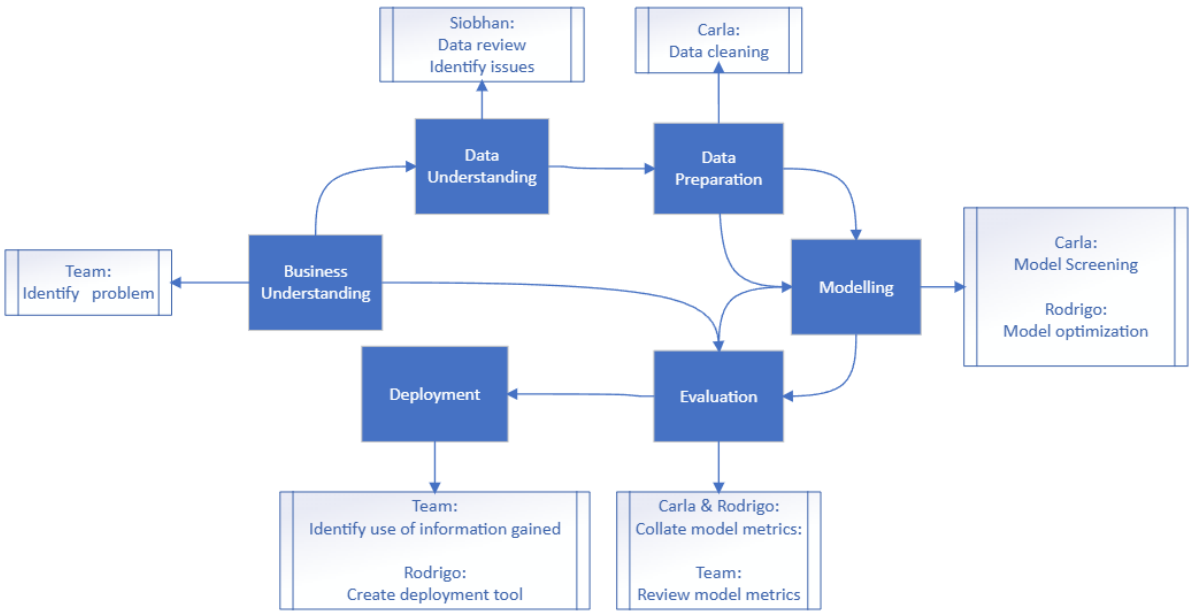

Figure 1: Task breakdown using CRSIP-DM approach

A data set was chosen from Kaggle[1] and the included data card indicated the target variable for prediction was a binary variable classifying whether each patient had heart disease or not. This meant that the objective of the modelling was to find the suitable to model that could best predict heart disease using the given features. These data are suitable for addressing the primary question of this study, enabling the classification of patients into two categories: "At-Risk Patients" and "Not-at-Risk Patients."

The data was explored and the quality was checked. The checked data was then prepared and classification models were screened using RapidMiner (v2024.1.0, Altair® AI Studio). The produced metrics were used to choose the best candidate models to perform optimization using Python. The final Python models were then reviewed to determine which was the most suitable model given that the data set was about disease identification and potential prevention rather than a simple business case.

As a final step, a tool was created in Python where the user could input patient information and be given a prediction on whether that patient had heart disease or not.

[1] Fedesoriano. (September 2021). Heart Failure Prediction Dataset. Retrieved [22nd Oct 2024] from https://www.kaggle.com/fedesoriano/heart-failure-prediction.

2. Business Understanding

According to the WHO, heart disease has been a leading cause of death for the past 20 years, accounting for 16% of deaths worldwide.[2] In order to reduce these figures, it is necessary to have predictive systems that can help prevent it and allow us to act more effectively and quickly. As a result, healthcare professionals will benefit from being able to identify patients at risk before they develop serious problems, thus reducing serious and costly cardiac events, improving patients' quality of life and reducing the percentage of mortality related to heart disease.

According to the Centers for Disease Control and Prevention (CDC), approximately 1.5 million heart attacks and strokes occur each year in the US. This contributes to an annual cost of more than $320 billion in medical care and lost productivity, reflecting a growing economic burden on the US healthcare system and economy.[3]

However, this problem is not unique to the US; the costs associated with cardiovascular disease are a significant burden on the health systems and economies of countries around the world. The care, treatment and lost productivity associated with these diseases affect governments, businesses and families worldwide, underscoring the need for preventive approaches and public health policies that comprehensively address this crisis.

Prediction of heart disease is an important topic to insurance companies, the medical community, public health bodies and the people who may be at risk. While the motivations of the interested groups may be different i.e. profit versus saving lives, all parties want a model that will best identify the disease and minimise false negatives.

A false positive, where a patient without heart disease is incorrectly predicted or diagnosed as having heart disease, is likely to be detected through continued patient testing. Any negative outcomes could be easily revered. On the other hand, a false negative will delay or even prevent treatment to the patient and this lack of accurate health information may lead to a financial cost to the insurance company in terms of a missed loading due to the disease or in an extreme case, a financial payout due to death.

The nature of this classification problem will have an influence on which model metrics to use when determining the best model.

[2] https://www.who.int/es/news/item/09-12-2020-who-reveals-leading-causes-of-death-and-disability-worldwide-2000-2019

[3] https://www.ahajournals.org/doi/abs/10.1161/circoutcomes.10.suppl_3.207

3. Data Understanding

The data understanding step was performed by importing the data straight into RapidMiner without making any modifications. The statistics and visualizations tabs were used for the initial data exploration.

The data set consisted of 918 entities and 11 variables which could be used to classify the variable of interest 'HeartDisease'. The predictor variables were both numerical and categorical.

The numerical variables were as follows:

- Age
- RestingBP
- Cholesterol
- MaxHR
- Oldpeak

The FastingBS was stated as an integer but as it only consists of 0s and 1s where 0 means non-diabetic and 1 means diabetic, this is a categorical variable. The list of categorical variables and their levels is given below.

- Sex (M, F)
- ChestPainType (ASY, NAP, ATA, TA)
- FastingBS (0, 1)
- RestingECG (Normal, LVH, ST)
- ExerciseAngina (Y, N)
- ST_Slope (Flat, Up, Down)

The 11 variables in this data set a mix of descriptive and baseline measurements of a patient (Age, Sex, RestingBP, Cholesterol, FastingBS, ChestPainType and RestingECG) and measurements from a test performed during exercise (MaxHR, OldPeak, ExerciseAngina and ST_Slope). Three of the variables are from an electrocardiogram (ECG) which is a test used to help diagnose and check conditions affecting the heart.[4] Annotated images to describe an ECG chart are given in the appendix to aid understanding of the variables related to an ECG.

There are 2 categories for the predicted variable HeartDisease, 1 to represent patients with heart disease and 0 to represent patients who do not have heart disease. The figure below shows that while they are not 50:50, neither category is over or under-represented so there was no requirement to use any techniques such as SMOTE to balance the data.

[4] https://www2.hse.ie/conditions/electrocardiogram-ecg/

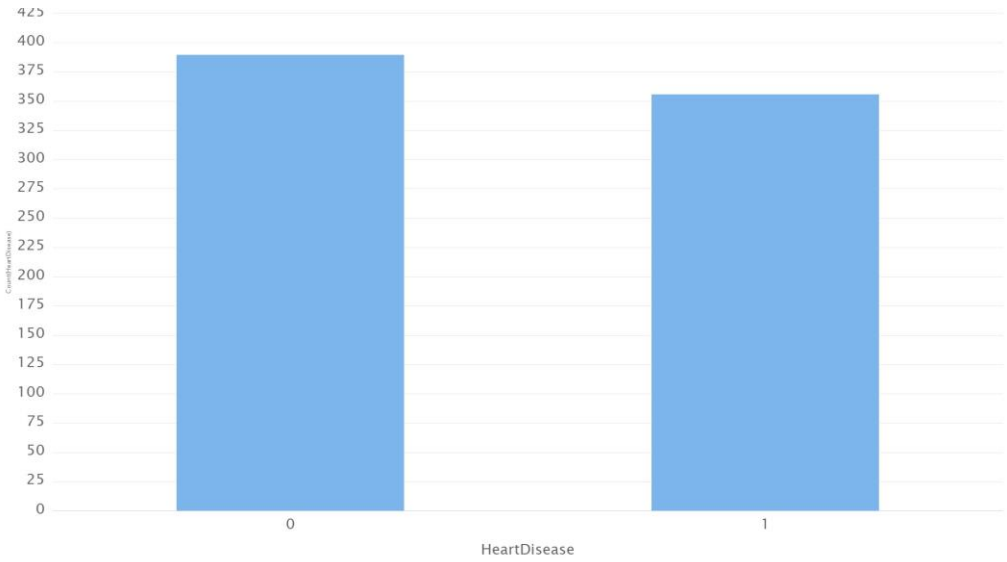

Figure 2: Examination of predictor variable HeartDisease

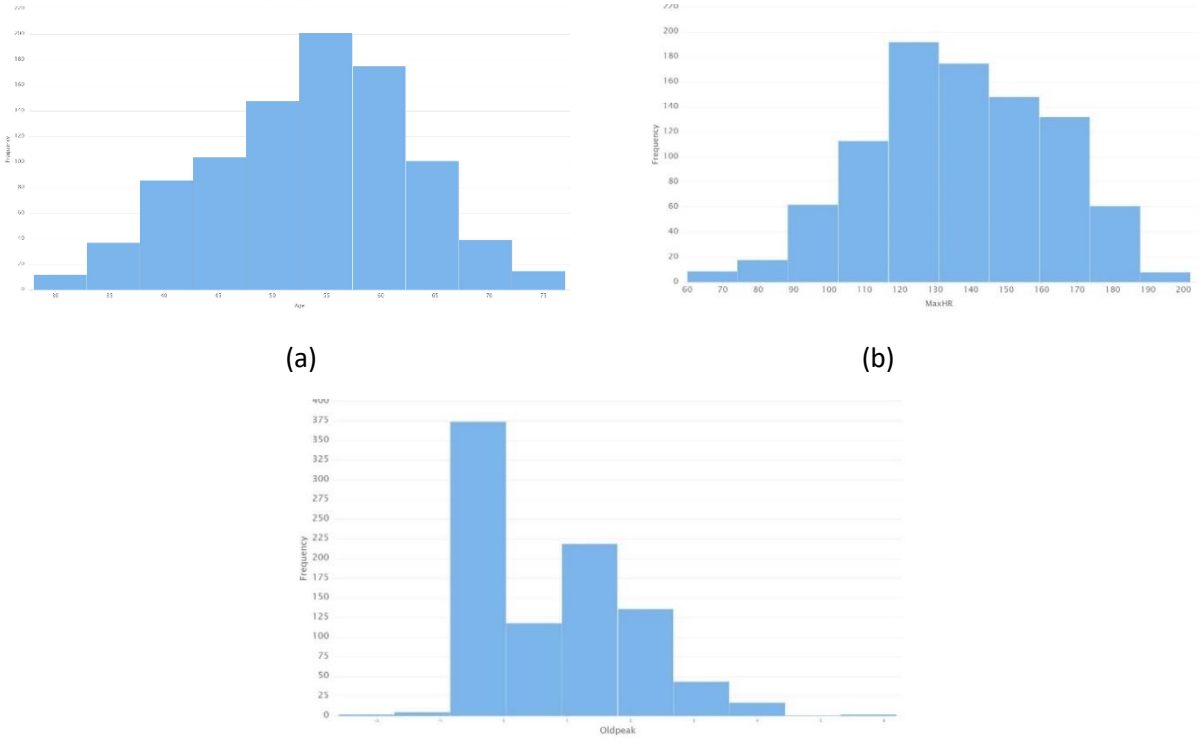

Figure 3: Histogram of numerical features (a) Age, (b) MaxHR and (c) OldPeak

The RapidMiner generated plots of the numerical features Age, MaxHR and OldPeak shown above indicate that these variables are of a relatively normal distribution. Values for age vary between 28

and 77 which seams reasonable for patient data for people at risk of heart disease. The values for MaxHR (Maximum heart achieved during an exercise test) ranged from 60 to 202 beats per minute (bpm). Again, these values appear to be within the expected region given that the average resting heart rate for adults is 60 to 100 bpm and is up to 200 bpm for adults during exercise.[5]

The OldPeak variable is a real number with values between -2.6 to 6.2. The numbers represent the magnitude of the change in ST depression on ECG chart between rest and exercise where ST depression is a segment of an ECG chart.

The other two numerical features were RestingBP (resting systolic blood pressure) and Cholesterol in units of mg/dL. Both of these numeric variables contained unexpected 0 values given the nature of these variable.[6] For this reason, the zero values within these variables were treated as missing values.

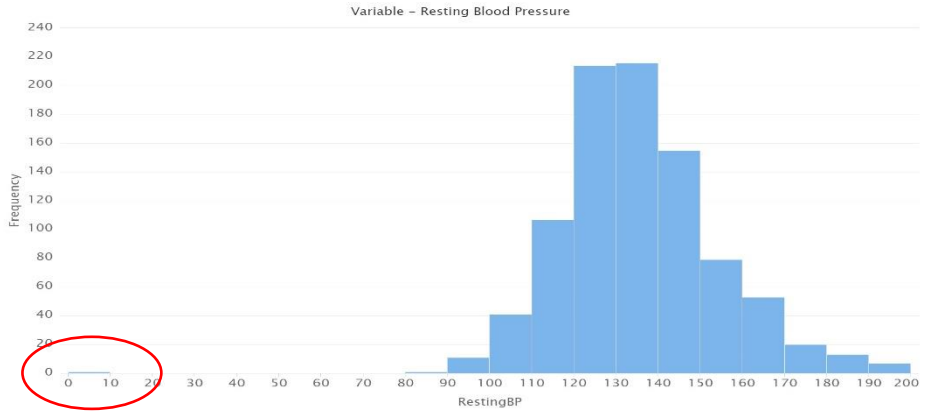

Figure 4: Histogram showing distribution of values for Resting Blood Pressure

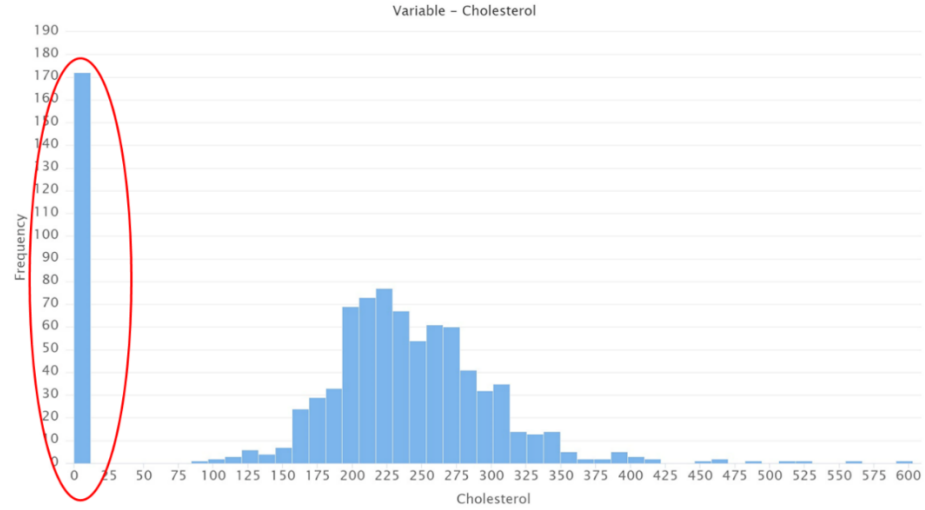

Figure 5: Histogram showing distribution of values for Cholesterol

[5] https://www.heart.org/en/healthy-living/fitness/fitness-basics/target-heart-rates

[6] https://irishheart.ie/wp-content/uploads/2024/09/Blood-Pressure-Ranges-768x432.png and https://www.verywellhealth.com/what-is-a-total-cholesterol-level-698073

The categorical variable Sex has 725 data points for the category M (Male) and 193 values for the category F (Female). This is a ratio of approximately 4:1, males to female. While this initially seems unbalanced, it represents the patient population for people with heart disease as acute coronary syndromes occur 3–4 times more often in men than in women below age 60, but after 75 years women represent the majority of patients.[7] The age range in this data set is 28 to 77 so this ratio of males to females is appropriate for this data set.

The variable ChestPainType has four categories (TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic). They describe chest pain in terms of angina which is chest pain caused by reduced blood flow to the heart muscles.[8] The category ASY or having no chest pain has 496 data points (54 % of data set) which makes sense for this data set which is relatively balanced between people with and without heart disease. The other categories describe different types of chest pain. There is a category for typical (46 data points) and atypical (173 data points) angina pain and another for non anginal pain (203 data points).

There is a variable (FastingBS) detailing whether a person is diabetic or not with the majority of patients (704 data points) being classed as non-diabetic. The RestingECG variable has three categories describing specific patterns on the ECG chart with normal being the shape for those without heart disease and the categories LVH and ST showing patterns associated with aspects of heart disease. The normal category represents 60 % of the data points while the other categories each represent about 20 % of the data.

Exercise angina is a binary variable to describe whether a person experiences angina while doing exercise. The categories are simply Y (Yes) and N (No) with a 40:60 yes:no ratio. The ST_Slope variable is the slope direction of ST segment on an ECG graph during exercise. The direction of the ST slope can be associated with certain conditions including heart disease.[9]

4. Data Preparation

The first step in building reliable predictive models is to properly preprocess the data.

In this case, the initial preprocessing will be done using RapidMiner and its AutoModel feature. While the AutoModel feature is technically part of the modelling step, it is gives direction for the main modelling optimization so it is described here. This tool will help us identify the most promising models for the dataset. Later, these insights will guide us to implement and fine-tune the selected models in Python deeper analysis and better performance.

This report focuses on the adjustments needed for the correct classification of data types in RapidMiner and explains why these changes are important for improving model accuracy and efficiency.

[7] Regitz-Zagrosek V. Sex and Gender Differences in Heart Failure. Int J Heart Fail. 2020 Apr 13;2(3):157-181. doi: 10.36628/ijhf.2020.0004. PMID: 36262368; PMCID: PMC9536682.

[8]https://www.nhs.uk/conditions/angina/#:~:text=Angina%20is%20chest%20pain%20caused,of%20these%20more%20serious%20problems.

[9] https://www.medicalnewstoday.com/articles/st-depression-on-ecg#causes

Issues with Data Classification

Sex:

- **Current Type**: Polynominal.
- **Correct Type**: Binominal.
- **Why**: This variable has only two categories (M/F), so it's binary.

ExerciseAngina:

- **Current Type**: Polynominal.
- **Correct Type**: Binominal.
- **Why**: This is a binary variable with "Yes" and "No" as possible values.

FastingBS (Fasting Blood Sugar):

- **Current Type**: Integer.
- **Correct Type**: Binominal.
- **Why**: Even though it's stored as a number, it represents binary categories (0: Normal, 1: Abnormal).

HeartDisease:

- **Current Type**: Integer.
- **Correct Type**: Binominal (Label).
- **Role**: This is the target variable that the model predicts:
  1: Presence of heart disease.
  0: Healthy.

When variables are classified incorrectly, it can lead to issues like:

- **Misinterpretation by the model**: Binary variables that are treated as polynominal or integers might not be processed correctly, which can make binary classification harder.
- **Unnecessary complexity**: If a binominal variable is marked as polynominal, the model may create extra categories (like one-hot encoding), which adds unnecessary complexity and slows down processing.

By correcting these classification errors in RapidMiner, we ensure that the models it generates perform better and provide us with a solid foundation for further implementation and improvement in Python.

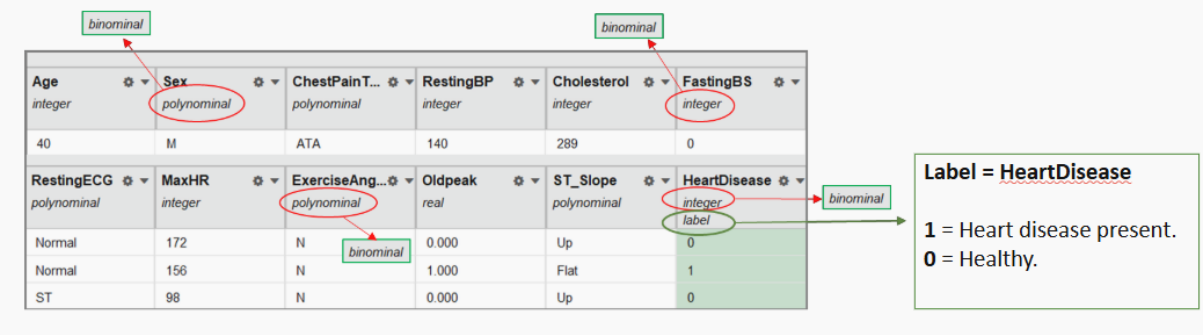

Figure 6: Data classification procedure in RapidMiner

Removing of Missing Values

During the preprocessing phase, we identified that some rows contained values of 0 for variables like cholesterol and resting blood pressure. These values are impossible in the given medical context, as neither cholesterol nor blood pressure can realistically be 0. As a result, these were classified as missing values and needed to be addressed.

Two potential strategies were considered:

- Replacing the missing values with the mean.
- Removing the rows containing missing values.

Given the nature of the data and its context in health analysis, replacing the missing values with the mean was deemed inappropriate, as it could distort the data distribution and compromise the integrity of the results. Removing these erroneous rows ensured the dataset remained as accurate and meaningful as possible, even if it meant reducing the total dataset size.

The steps taken to eliminate missing values were as follows:

Identification:
Rows where Cholesterol or RestingBP had a value of `0` were flagged as missing.

Action:
Using the **Transform** tool in RapidMiner, the rows containing these missing values were filtered out and deleted.

Outcome:
The dataset size was reduced from 918 rows to 746 rows, ensuring only valid and accurate data remained.

The filtered dataset was then used for further analysis in the **AutoModel** tool, maintaining data integrity for model training.

While removing missing data can reduce the overall dataset size, in this case, the reduction was manageable, and enough data remained to train robust models. Ensuring the dataset was clean and free of erroneous values is critical, particularly in health-related contexts, where even minor inaccuracies can lead to misleading conclusions.

By taking this approach, we ensured that the analysis stayed true to the data's real-world implications while maintaining sufficient data for effective modelling in subsequent steps.

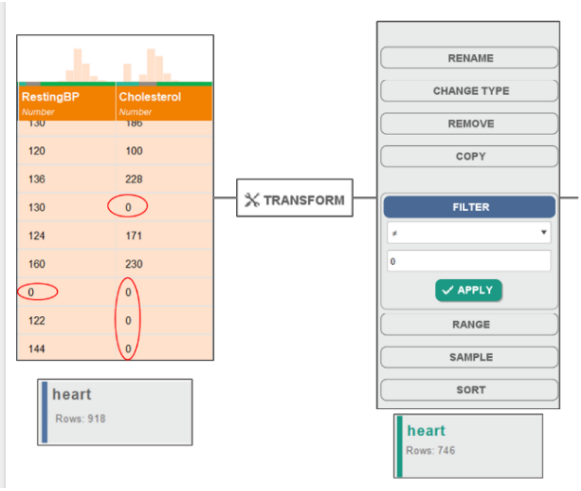

Figure 7: Missing data handling in RapidMiner

ROC Comparison

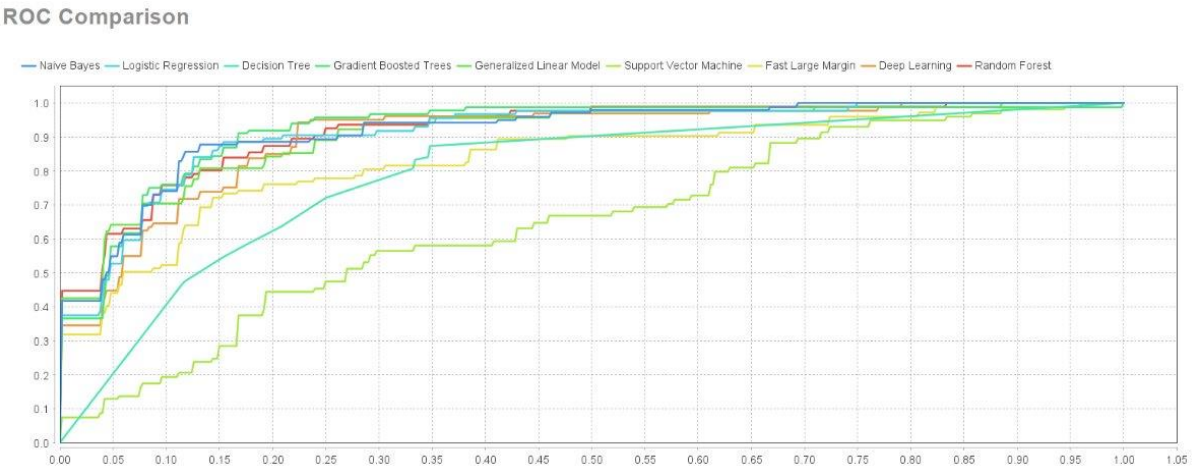

Figure 8: ROC comparison curve for models generated using AutoModel

The **ROC curve** (Receiver Operating Characteristic) is used to check how well different classification models perform. It shows the balance between sensitivity (true positives) and specificity (false positives) for each model.

The chart compares the performance of several models, including:

- **Naïve Bayes**
- **Logistic Regression**
- **Decision Tree**
- **Gradient Boosted Trees**
- **Random Forest**

Looking at the curves, the models that perform the best are:

1. **Naïve Bayes**
2. **Logistic Regression**
3. **Gradient Boosted Trees**

These models have the highest curves, meaning they are the best at predicting correctly for our dataset.

Selecting the best predictive Model

Table 1: Summary of model metrics from RapidMiner

	NB	SD+-	GLM	SD+-	Logit	SD+-	FastLargeM	SD+-	DeepLearnin	SD+-	Dec. Tree	SD+-	Random For	SD+-	GBT	SD+-	SVM	SD+-
accuracy	0.85370985 60354375	0.02644624 003851269	0.81129568 10631229	0.03297805 909522573	0.85470653 37763012	0.03721943 581361532	0.78781838 31672204	0.04053519 2183478924	0.82635658 91472868	0.01143246 6049427261	0.79822812 84606866	0.01974268 778352723	0.83023255 81395349	0.00925374 9136681578	0.83488372 09302327	0.00212295 5649223518	0.61926910 29900331	0.05213936 1721208505
classificatio n_error	0.14629014 396456258	0.02644624 0038519496	0.18870431 893687706	0.03297805 909522586	0.14529346 62236988	0.03721943 581361759	0.21218161 683277964	0.04053519 218348021	0.17364341 085271315	0.01143246 6049420813	0.20177187 153931345	0.01974268 7783530086	0.16976744 186046508	0.00925374 9136677125	0.16511627 90697674	0.00212295 5649245989	0.38073089 70099668	0.05213936 172120717
AUC	0.91689462 93062786	0.03846095 259910698	0.90854884 6693521	0.03337337 416763366	0.91284170 43830577	0.03344620 5762342786	0.83514789 10228647	0.06297239 830000022	0.90078684 45212974	0.04920528 074868442	0.79373606 76800205	0.03451558 0635432905	0.91629885 95074244	0.04191339 409147924	0.92271947 41653042	0.04008622 279718255	0.65029512 19624152	0.07575777 63043945
precision	0.84491231 15210071	0.08434642 939861402	0.79887370 60041406	0.06243706 002733598	0.86320346 32034632	0.05857628 3702108996	0.81095238 09523809	0.06031850 458578187	0.81966708 21438623	0.07603534 176729813	0.76791316 10077901	0.06713445 277150792	0.80418992 77193394	0.08294523 237971513	0.82175070 02801121	0.07444731 19395161	0.60429487 17948718	0.12839961 21018385
recall	0.87998833 62251784	0.06993239 25469146	0.81480865 62560246	0.11410805 93440344	0.80605359 55272797	0.09665957 693842903	0.74228985 92635435	0.12427579 99718738	0.82949595 14170042	0.07759665 945530268	0.85024089 06882589	0.06777271 719786077	0.85419076 53749758	0.07139271 214329458	0.83414064 00616928	0.04948697 32948786	0.45145112 78195489	0.09147690 506622545
f_measure	0.85683717 35791092	0.02090176 3425373105	0.80031055 90062113	0.04477067 099478563	0.83072915 38603716	0.05929946 492401098	0.77073942 62052647	0.07895130 127382202	0.82012644 99191734	0.03310310 0224210896	0.80297636 0522873	0.02037932 707647539	0.82385494 80272367	0.02904003 0294526935	0.82488254 06272214	0.02487767 376714384	0.51350174 21602788	0.09300068 148855628
sensitivity	0.87998833 62251784	0.06993239 25469146	0.81480865 62560246	0.11410805 93440344	0.80605359 55272797	0.09665957 693842903	0.74228985 92635435	0.12427579 99718738	0.82949595 14170042	0.07759665 945530268	0.85024089 06882589	0.06777271 719786077	0.85419076 53749758	0.07139271 214329458	0.83414064 00616928	0.04948697 32948786	0.45145112 78195489	0.09147690 506622545
specificity	0.82937241 78634665	0.11175461 572336831	0.81074356 16387022	0.07988711 333340598	0.89705702 79393808	0.03453234 777008223	0.82654756 43992268	0.05710408 9133272064	0.82509891 35899621	0.08120059 627189057	0.74770476 75257394	0.08823290 740669748	0.81517827 86693273	0.08459892 631203615	0.84173107 37003831	0.06441360 178336003	0.75938135 93813594	0.04745278 293952862

To decide which metrics will be the ones with most weight for a predictive model, it is crucial to consider the context of the problem, that in this case is Health.

Predicting heart disease is a critical task where the consequences of errors are significant:

- **False negatives** (not detecting a real case) can prevent necessary treatment, putting lives at risk.
- **False positives** (diagnosing a condition that doesn't exist) can cause unnecessary stress or treatment.

Given this, certain metrics take priority like Recall/Sensitivity, F-Measure, AUC, specificity or precision.

Results of the Top Models

From the models suggested by **RapidMiner**, the three best performers were evaluated based on their metrics:

Naive Bayes:

- **Recall**: 0.8799 (good ability to detect positive cases).
- **Precision**: 0.8449 (fairly accurate predictions).
- **Specificity**: 0.8293 (lower than others).
- **Error Analysis**:
  - **True Negatives**: 94
  - **True Positives**: 87
  - **False Positives**: 13
  - **False Negatives**: 18

Observation: Although it balances recall and precision well, the 18 false negatives could pose a serious issue in this health-related problem.

Logistic Regression:

- **Specificity**: 0.8970 (best at identifying negatives).
- **Precision**: 0.8632 (accurate predictions).
- **Recall**: 0.8060 (lower ability to detect all positives).

- **Error Analysis**:
  True Negatives: 79
  True Positives: 103
  False Negatives: 12
  False Positives: 19

Observation: Fewer false negatives compared to Naive Bayes, reducing false alarms. Performs well in terms of specificity but is slightly weaker in recall.

Gradient Boosted Trees:

- **AUC**: 0.9227 (best overall performance across thresholds).
- **Recall**: 0.8341 (highest among the three).
- **Precision**: 0.8217 (slightly lower than others).
- **Error Analysis**:
  True Positives: 93
  True Negatives: 84
  False Positives: 17
  False Negatives: 18

Observation: The model provides the best balance of metrics, with the highest AUC and competitive recall. However, it still shares the same number of false negatives as Naive Bayes.

Based on the priority metric (recall) to minimize false negatives, **Gradient Boosted Trees** emerged as the most suitable model due to its high recall and AUC. However, the performance of all three models was reasonably close.

To ensure the most accurate and reliable results, we decided to perform a more in-depth evaluation using **Google Colab**, where additional testing and refinement were conducted to confirm which model best aligns with the objectives of the analysis.

Building on the insights gained from RapidMiner, we identified the best-performing models and implemented them in Python. Both standard and optimized versions of these models were developed and evaluated to achieve the best results.

The models we focused on will be:

- **Logistic Regression**
- **Naïve Bayes**
- **Gradient Boosted Trees**

5. Modelling

To analyse the risk of heart disease and predict potential cases, we employed three distinct machine learning models: **Logistic Regression**, **Naive Bayes**, and **Gradient Boosted Trees**. These models were chosen for their complementing characteristics, allowing us to compare their strengths and weaknesses in terms of predictive accuracy, interpretability, and computational efficiency.

1. Logistic Regression

Logistic Regression is a widely used, interpretable model that estimates the probability of an outcome using a logistic function. It assumes a linear relationship between the input features and the log-odds of the target variable. This simplicity makes it an excellent baseline model, particularly for medical datasets where interpretability and clarity of feature contributions are essential. It also performs well when relationships in the data are linear or close to linear.

2. Naive Bayes

Naive Bayes is a probabilistic model based on Bayes' Theorem, assuming conditional independence between features. Despite its simplicity, it is robust and effective for binary classification tasks like this one. It provides probabilistic predictions, making it suitable for understanding the likelihood of heart disease while requiring minimal computational resources. This model is particularly effective when the dataset features are nearly independent.

3. Gradient Boosted Trees

Gradient Boosted Trees (GBT) is an advanced ensemble learning method that builds multiple decision trees sequentially, with each tree improving on the mistakes of the previous one. It is highly flexible and capable of capturing complex, non-linear relationships in the data. Although computationally intensive compared to Logistic Regression and Naive Bayes, GBT often delivers superior predictive performance, making it a powerful choice for medical risk prediction.

4. Optimizing the Models and Grid Search

To optimise our models, we used **Grid Search**, a systematic approach to find the best hyperparameter combinations for each algorithm. By defining a range of hyperparameters, such as regularisation strength (C) and solver type for Logistic Regression, or learning rate and number of estimators for Gradient Boosted Trees, Grid Search evaluates all possible combinations to identify the configuration that yields the best performance. We used **cross-validation** to validate each combination, ensuring robust results across different subsets of the data.

This process was essential to enhance the models' predictive power, as hyperparameters significantly influence their performance. For example, tuning Logistic Regression improved recall, aligning with our goal of minimising false negatives in health risk prediction. Overall, Grid Search allowed us to systematically optimise the models, balancing accuracy, precision, and recall to create reliable and effective predictors.

6. Evaluation

A summary of all metrics optimization candidate models is shown in Table 1 below with the details described in the following sections.

Table 2: Model metrics for Standard and Optimized models using Python

Metric	Logistic Regression - Standard	Logistic Regression - Optimized	Naive Bayes - Standard	Naive Bayes - Optimized	Gradient Boosted Trees - Standard	Gradient Boosted Trees - Optimized
Accuracy	0.89	**0.90**	0.89	0.89	**0.90**	0.89
Precision	0.93	0.91	0.91	0.91	**0.94**	0.93
Recall	0.86	**0.90**	0.87	0.87	0.86	0.86
F1 Score	0.89	**0.90**	0.89	0.89	**0.90**	0.89
AUC Score	**0.96**	**0.96**	0.96	0.96	**0.96**	0.94

1. Logistic Regression – Standard

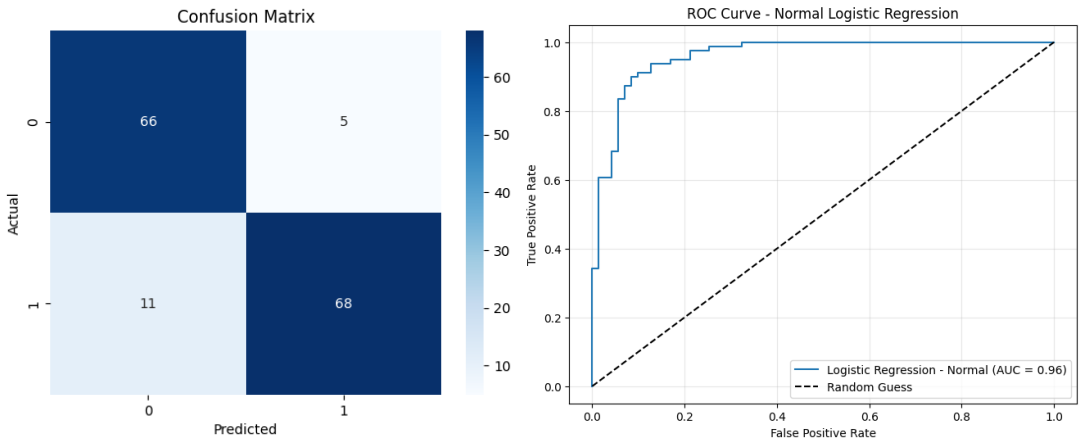

Figure 9: Confusion matrix (a) and ROC curve (b) for Standard Logistic Regression model

a. The **standard Logistic Regression model** performed well as a baseline for heart disease prediction, achieving an **accuracy of 89%**, a **precision of 0.93**, a **recall of 0.86**, and an **F1 score of 0.89**, indicating a strong balance between correctly identifying patients at risk and avoiding false positives. The **AUC score of 0.96** highlights the model's excellent capability in distinguishing between at-risk and not-at-risk patients.

b. The confusion matrix reveals that the model correctly classified **66 patients without heart disease (true negatives)** and **68 patients with heart disease (true positives)**. However, it misclassified **5 patients as at risk (false positives)** and **11 patients as not at risk (false negatives)**, showing a slight tendency to miss positive cases.

c. The **ROC curve** for the standard Logistic Regression model confirms its effectiveness, with a steep curve and an **AUC score of 0.96**, reflecting a high true positive rate and a low false positive rate. This indicates the model is reliable in predicting the likelihood of heart disease across different thresholds.

2. Logistic Regression – Optimized

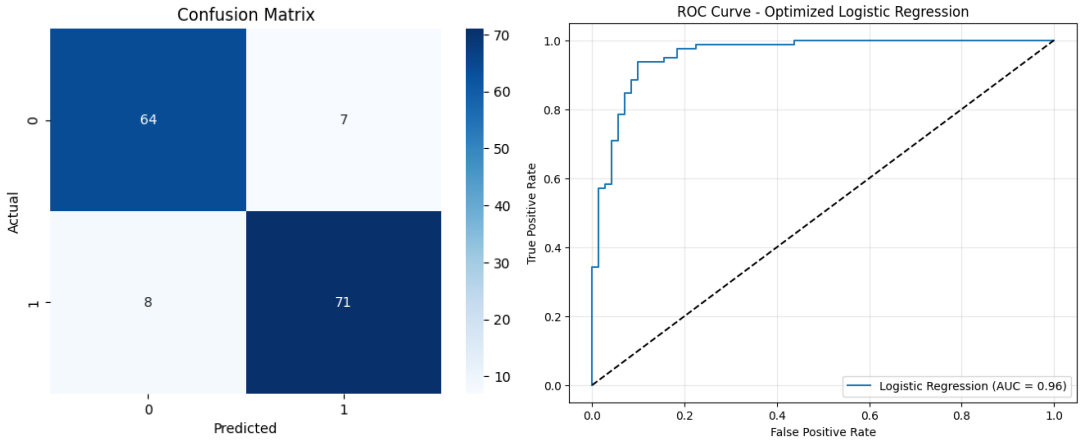

Figure 10: Confusion matrix (a) and ROC curve (b) for Optimized Logistic Regression model

a. The **optimized Logistic Regression model** demonstrated improved performance compared to the standard version. It achieved an **accuracy of 90%**, a **precision of 0.91**, a **recall of 0.90**, and an **F1 score of 0.90**, highlighting its ability to balance true positive and true negative predictions effectively. The **AUC score of 0.96** further reinforces the model's strong ability to distinguish between patients at risk and those not at risk of heart disease.

b. The confusion matrix shows that the optimized model correctly classified **64 patients without heart disease (true negatives)** and **71 patients with heart disease (true positives)**. However, it incorrectly classified **7 patients as at risk (false positives)** and missed **8 patients with heart disease (false negatives)**. This reflects an improvement in recall compared to the standard version.

c. The **ROC curve** for the optimized Logistic Regression model remains robust, with an **AUC score of 0.96**, demonstrating a high true positive rate across different classification thresholds. This confirms that the optimization process enhanced the model's overall predictive capabilities while maintaining its discriminative power.

3. Naïve Bayes – Standard

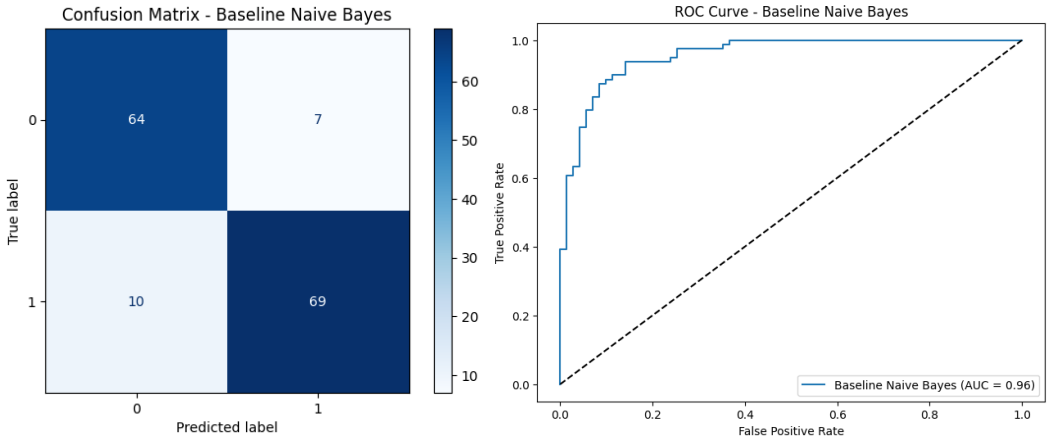

Figure 11: Confusion matrix (a) and ROC curve (b) for Naive Bayes model

a. The **baseline Naive Bayes model** provided solid predictive performance for heart disease classification. It achieved an **accuracy of 89%**, a **precision of 0.91**, a **recall of 0.87**, and an **F1 score of 0.89**, demonstrating a balanced ability to correctly identify at-risk patients while minimising false positives. The **AUC score of 0.96** indicates the model is highly capable of distinguishing between patients with and without heart disease.

b. The confusion matrix shows the model correctly classified **64 patients without heart disease (true negatives)** and **69 patients with heart disease (true positives)**. However, it misclassified **7 patients as at risk (false positives)** and missed **10 patients with heart disease (false negatives)**. These metrics reflect the model's slight tendency to under-detect true positives compared to Logistic Regression.

c. The **ROC curve** confirms the strength of the baseline Naive Bayes model, with an **AUC score of 0.96**, highlighting its ability to maintain a high true positive rate while keeping the false positive rate low. This makes the model a reliable choice for probability-based predictions in heart disease classification.

4. Naïve Bayes – Optimized

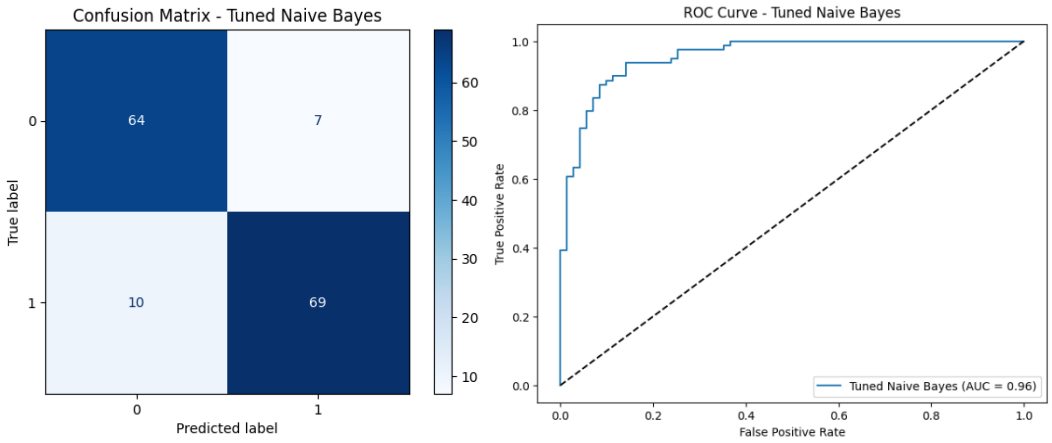

Figure 12: Confusion matrix (a) and ROC curve (b) for Optimized Naive Bayes model

a. The **Optimized Naive Bayes model** demonstrated identical performance to the baseline model. It achieved an **accuracy of 89%**, a **precision of 0.91**, a **recall of 0.87**, and an **F1 score of 0.89**, maintaining a strong balance between identifying true positives and avoiding false positives. The **AUC score of 0.96** highlights its continued strength in distinguishing between patients with and without heart disease.

b. The confusion matrix for the tuned Naive Bayes model remains unchanged, with **64 true negatives** and **69 true positives**, along with **7 false positives** and **10 false negatives**. This consistency suggests that hyperparameter tuning did not significantly affect the model's performance, as it was already well-calibrated in its baseline configuration.

c. The **ROC curve** confirms that the optimized Naive Bayes model retains its robust performance, with an **AUC score of 0.96**, further validating its reliability in classifying heart disease risk. The model remains a dependable choice for probabilistic predictions with minimal computational cost.

5. Gradient Boosted Trees – Standard

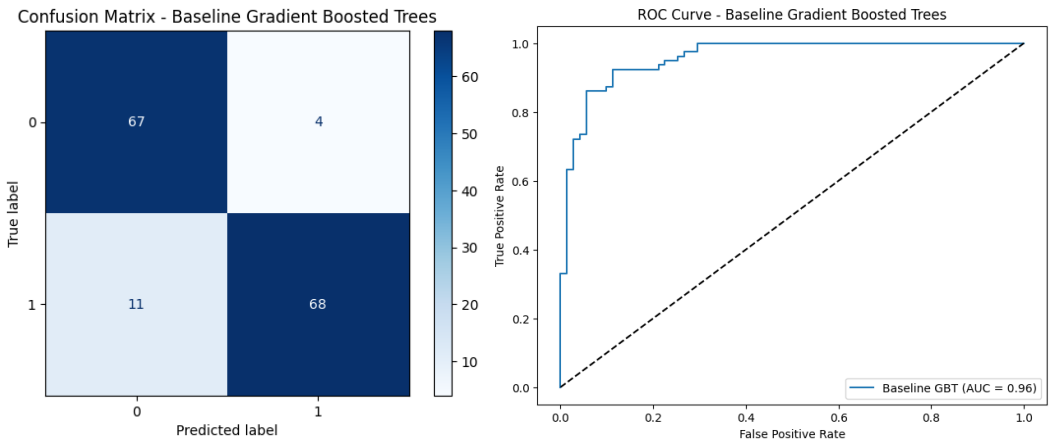

Figure 13: Confusion matrix (a) and ROC curve (b) for Standard Gradient Boosted Trees model

a. The **baseline Gradient Boosted Trees (GBT) model** delivered strong predictive performance with an **accuracy of 90%**, a **precision of 0.94**, a **recall of 0.86**, and an **F1 score of 0.90**. The high precision indicates the model is particularly effective at minimising false positives. The **AUC score of 0.96** demonstrates the model's excellent ability to distinguish between at-risk and not-at-risk patients.
b. The confusion matrix shows that the model correctly classified **67 patients without heart disease (true negatives)** and **68 patients with heart disease (true positives)**. However, it misclassified **4 patients as at risk (false positives)** and missed **11 patients with heart disease (false negatives)**. These results highlight the model's strength in precision but show a slight trade-off in recall.
c. The **ROC curve** for the baseline Gradient Boosted Trees model confirms its strong performance, with an **AUC score of 0.96**, reflecting a high true positive rate and a low false positive rate across different thresholds. This makes the baseline GBT model a robust choice for heart disease prediction.

6. Gradient Boosted Trees – Optimized

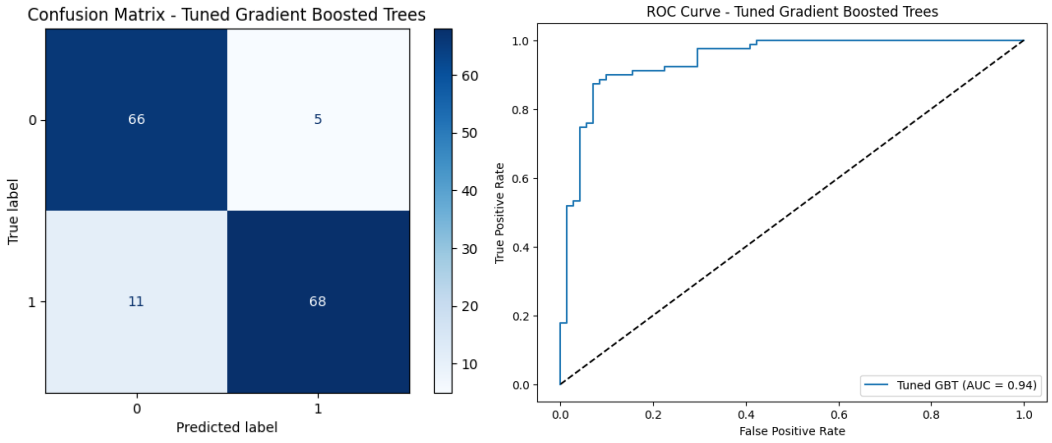

Figure 14: Confusion matrix (a) and ROC curve (b) for Optimized Gradient Boosted Trees model

a. The **optimized Gradient Boosted Trees (GBT) model** achieved slightly different performance metrics compared to its baseline version. It recorded an **accuracy of 89%**, a **precision of 0.93**, a **recall of 0.86**, and an **F1 score of 0.89**, maintaining a balance between precision and recall while slightly lowering precision compared to the baseline. The **AUC score of 0.94** highlights the model's ability to distinguish between at-risk and not-at-risk patients, although it saw a slight decrease from the baseline.
b. The confusion matrix shows that the optimized GBT model correctly classified **66 patients without heart disease (true negatives)** and **68 patients with heart disease (true positives)**. It misclassified **5 patients as at risk (false positives)** and missed **11 patients with heart disease (false negatives)**, indicating no significant change in the distribution of errors post-optimization.

c. The **ROC curve** for the optimized Gradient Boosted Trees model, with an **AUC score of 0.94**, confirms that the optimization retained the model's strong discriminative power. However, the slight decrease in AUC compared to the baseline model suggests that the optimization may not have improved performance in this specific scenario.

7. Deployment

Applying Logistic Regression and Gradient Boosted Trees to Classify New Data

To demonstrate the predictive capabilities of our models, we tested them on new patient data with the same set of features used in training. Both the **Optimized Logistic Regression** and **Gradient Boosted Trees (Standard)** models predicted that the individual was **NOT at risk of heart disease**, with prediction probabilities of **0.23** and **0.24**, respectively. These results show that both models provide similar predictions and probabilities, reflecting their comparable performance in classifying unseen data.

While the prediction probabilities are close, we chose the **Optimized Logistic Regression model** as the preferred model for this task. This decision was driven by our goal to prioritise **minimising false negatives**, which is crucial in health-related predictions to ensure that actual at-risk individuals are not overlooked. The Logistic Regression model demonstrated better recall during evaluation, meaning it is more sensitive to identifying true cases of heart disease. This makes it a safer and more reliable choice when the consequences of missing a true case can be severe.

The similarity in probabilities underscores the robustness of both models, but the context of prioritising health risks guided our final selection. The Optimized Logistic Regression model aligns better with our objective of reducing missed cases of heart disease.

8. Conclusion

A model that could predict heart disease risk would optimize hospital resources in a number of ways, including reducing the number of unnecessary consultations, prioritising care for high-risk patients and improving resource allocation in intensive care units or specialised cardiology teams. This would not only facilitate more efficient and preventative care by improving the personalisation of medical care, but also reduce waiting times for patients, improve customer satisfaction, improve health outcomes and reduce the workload of emergency departments. In addition, by preventing hospitalisation and intensive treatment through early diagnosis and management, it would significantly reduce the operating costs of the healthcare system, allowing resources to be redirected to other critical areas and promoting a more sustainable and equitable healthcare system.

A predictive model would also raise social awareness of cardiovascular health risks and motivate people to adopt healthier lifestyles. With a better understanding of risk factors, such as physical inactivity or an unbalanced diet, people would be more inclined to make positive changes that would contribute to a lower incidence of heart disease and a better overall quality of life. In addition, this awareness would facilitate the development of public health policies focused on prevention and collective well-being.

In conclusion, developing a model capable of predicting the risk of heart disease not only offers the potential to save lives but also, as noted in the previous points, presents an opportunity to reduce costs and optimize resources within healthcare systems. Additionally, it promotes the adoption of healthier lifestyles, supported by the development of preventive public health policies. Therefore, this model could become an essential tool in addressing the healthcare crisis and mitigating the economic impact posed by cardiovascular diseases.

9. Appendix

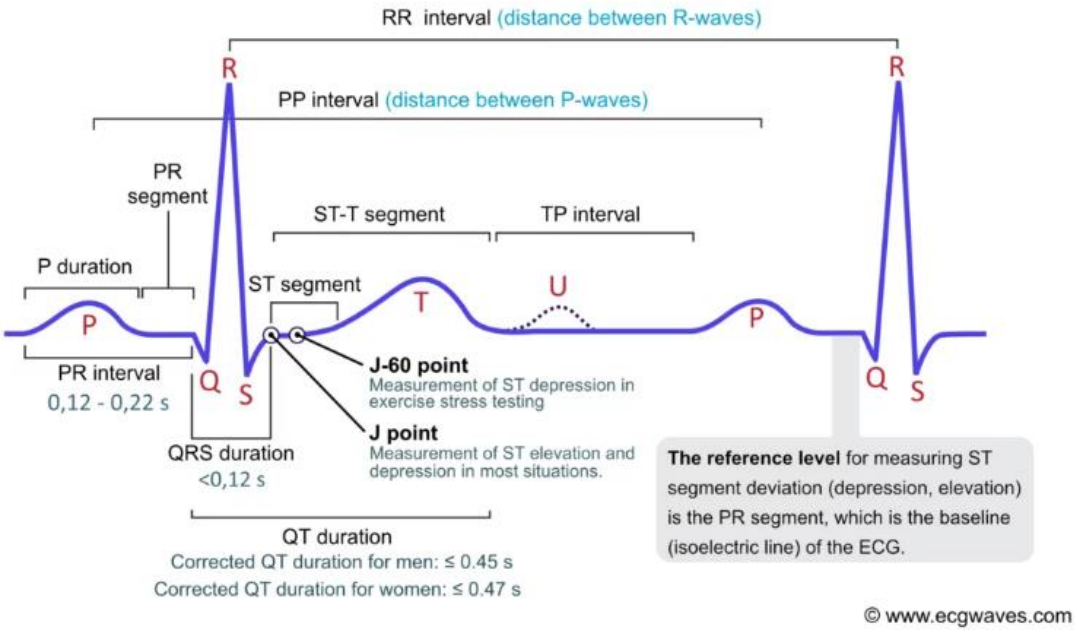

Figure A.1: Explanation of ECG chart sections from www.ecgwaves.com

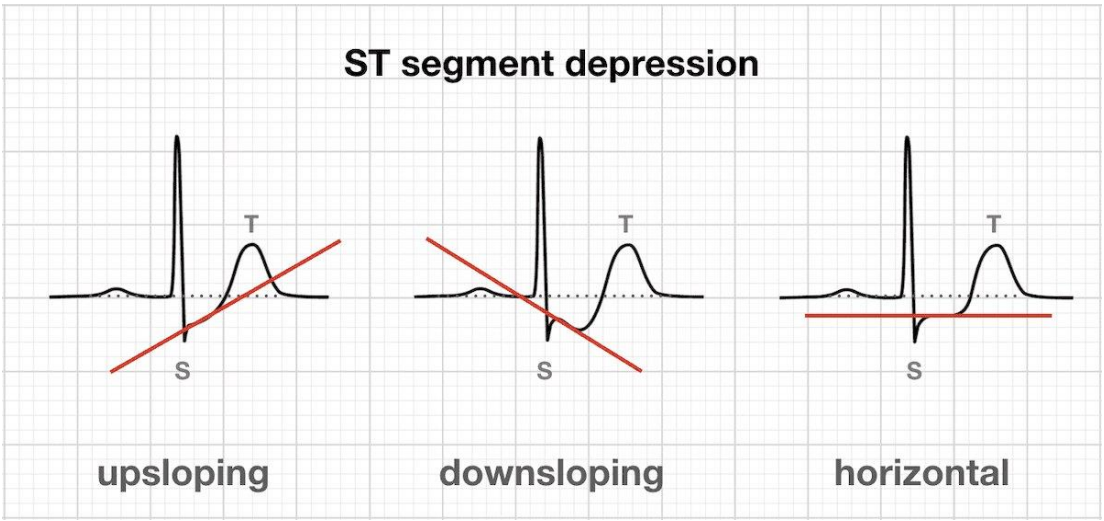

Figure A.2 Explanation of ST segment depression (https://litfl.com/wp-content/uploads/2018/10/ST-segment-depression-upsloping-downsloping-horizontal.png)