Diagnostics for Detecting Cardiovascular Abnormalities

BY: ARYAN DAHIYA

Abstract: *This study determines heart disease risk variables using a patient information dataset. I used machine learning techniques and exploratory data analysis to investigate the relationships between various medical traits and the diagnosis of heart disease. The research has revealed significant correlations between the risk of heart disease and factors like blood pressure, sex, age, and type of chest discomfort. Support vector machines, random forests, and logistic regression were the three machine learning models that were utilized to forecast cardiac illness. With an accuracy of 98.54%, the random forest classifier has the highest performance. I used a thorough methodology that included accuracy scores as the main metric and extensive classification reports with precision, recall, and F1-scores for every class to assess the models' performance. Furthermore, I used confusion matrices to display the models' performance and provide insight into the different kinds of errors that each model makes. All models perform well, according to the examination, but the random forest classifier stands out for its overall accuracy and its balanced performance across classes. While the support vector machine achieves 86.83% accuracy, the logistic regression model only manages 80.49%.*

Key Words: Exploratory Data Analysis, Correlations, Support vector machines, Random forests, Logistic regression, Recall, F1-scores

INTRODUCTION:

Heart disease is still one of the major causes of death globally, so prevention and early detection are essential. A third of all deaths worldwide are thought to be caused by cardiovascular disorders, with coronary heart disease being the primary cause of these deaths, according to the World Health Organization. Improved techniques for early diagnosis and risk assessment are desperately needed, as this startling data shows. By utilizing patient data, this study seeks to pinpoint important risk variables and create heart disease prediction models. In order to identify trends and connections suggestive of cardiac illness, I examine a large dataset comprising several physiological and test result characteristics.

The dataset used in this study includes a wide range of factors, including clinical measurements like blood pressure and cholesterol levels, findings from several cardiac tests, and demographic data like age and sex. Through an analysis of these various characteristics, the goal is to pinpoint the known and maybe unknown elements that raise the risk of heart disease. By using data visualization approaches, I hope to make complex relationships understandable and visually appealing. These visualizations are useful tools for communicating findings to patients and medical professionals alike, in addition to supporting our study. By using methods like distribution plots, correlation heatmaps, and comparison analysis, we try to draw attention to the most important patterns in the data.

Additionally, this study uses a variety of machine learning methods to create heart disease prediction models. The goal is to determine the best strategy for early detection by contrasting the results of support vector machine, random forest, and logistic regression models. Healthcare providers may find it easier to prioritize patients for additional testing or intervention and make better judgments with the help of these models. I aim to add to the expanding body of knowledge on data-driven strategies for managing and preventing heart disease by fusing exploratory data analysis with predictive modelling. The research could contribute to the improvement of risk assessment techniques, which could result in more specialized and successful responses. The ultimate aim of this study is to better understand heart disease risk factors and increase our capacity to anticipate and prevent this ubiquitous health issue by utilizing the power of Exploratory data analysis and machine learning.

APPROACH AND OUTCOMES:

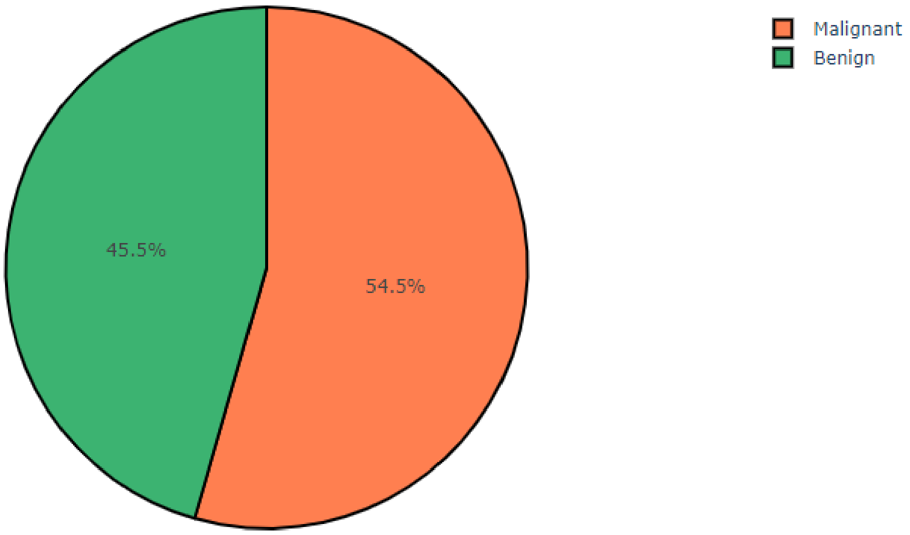
The approach involves several key steps:

Data Preprocessing:

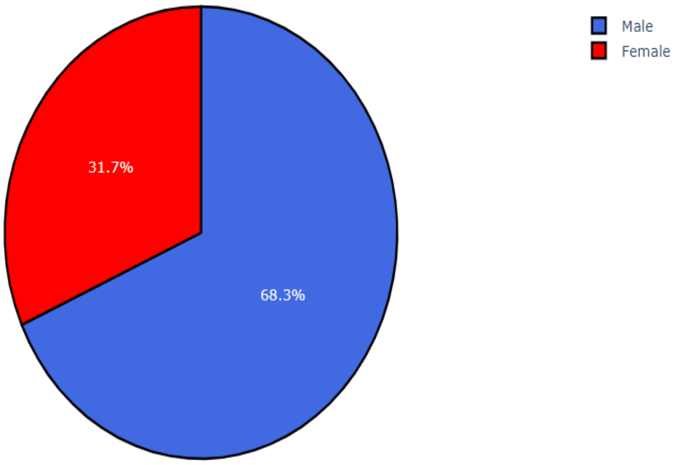
Loaded the heart disease dataset, handle missing values, and normalize the features using Min-Max scaling.

Exploratory Data Analysis (EDA):

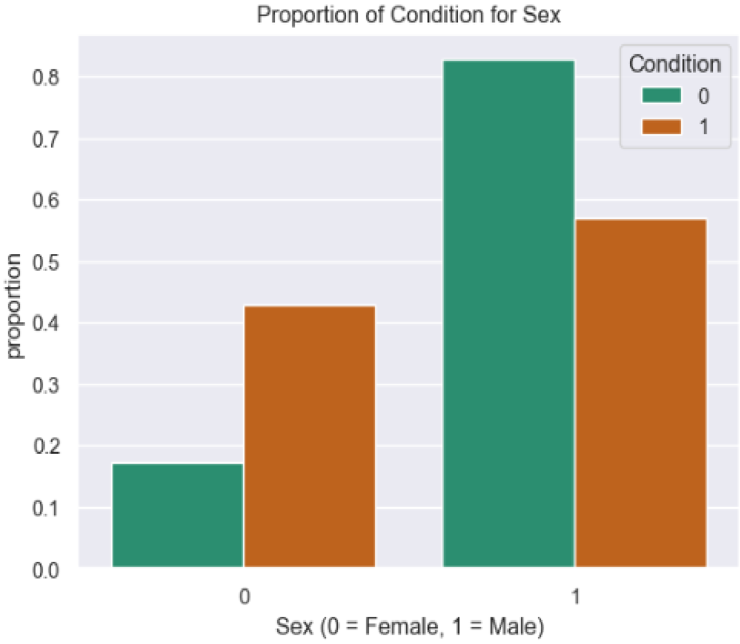
Conducted extensive visualizations to understand the distributions and relationships between variables. This includes pie charts for categorical variables, distribution plots for continuous variables, and comparisons across different subgroups (e.g., by gender and condition).



Benign tumours, which are non-cancerous and do not spread, are greatly outnumbered by malignant tumours, which are cancerous and have the potential to spread to other parts of the body. The information shows that benign tumours accounted for less than half of the instances examined, with malignant tumours accounting for 54.5% of the cases.

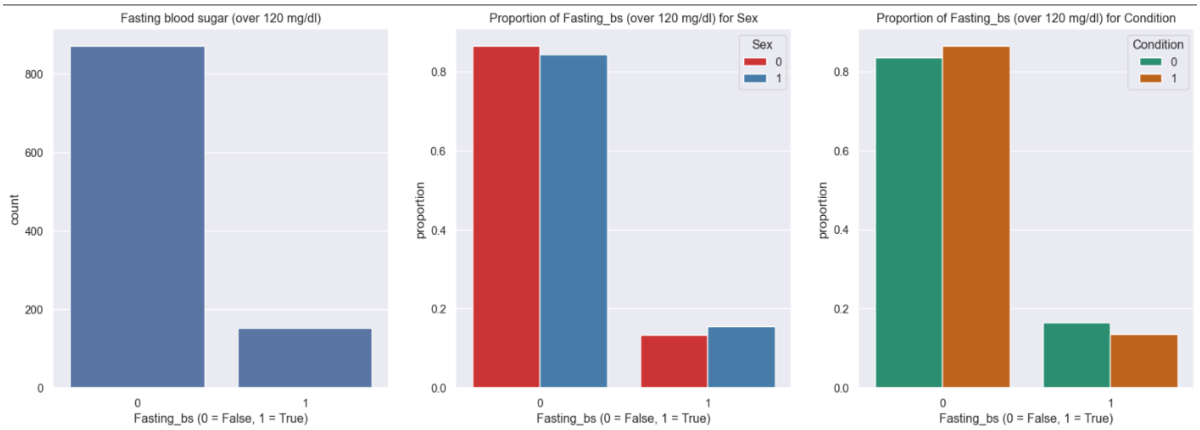


According to the data, there is a notable gender disparity, with men making up almost twice as many people as women. More specifically, 68.3% of the population is male and only 31.7% is female. The given pie chart, which graphically depicts the gender distribution, makes this discrepancy clear.

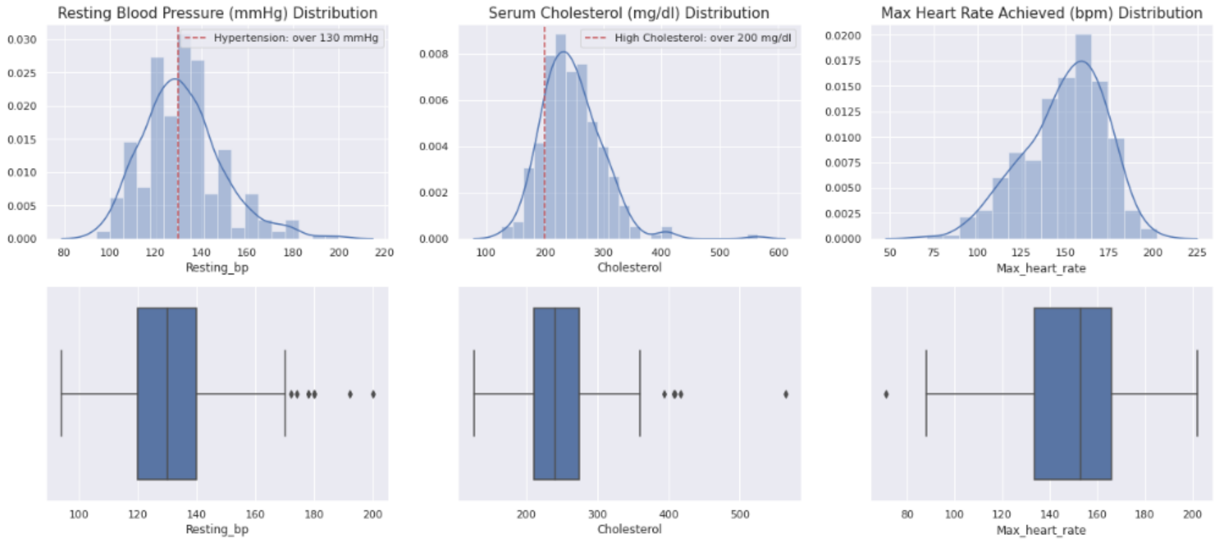


This graph shows the percentage of heart disease conditions by sex. Benign conditions (no heart disease) are represented as 0, whereas malignant conditions (heart disease) are represented as 1. Approximately 18% of females (0 on the x-axis) have benign disorders, whereas 43% have malignant conditions. Approximately 83% of males (1 on the x-axis) have benign illnesses, while 57% have malignant conditions.

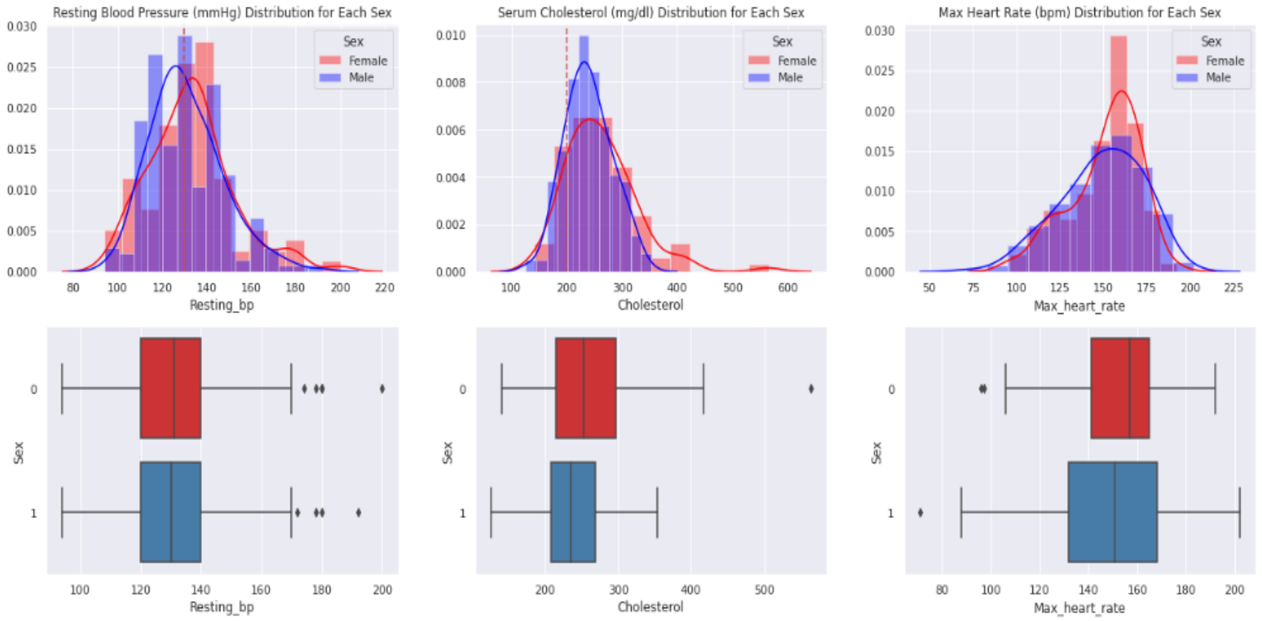
Crucially, these figures show the share of cancer cases for every gender in their respective categories. Sex is a significant risk factor for heart disease, as this visualization plainly shows that men in our dataset have a higher proportion of malignant heart disorders than women do. This finding aligns with the study's approach of analysing various demographic and physiological factors to develop accurate predictive models for heart disease.



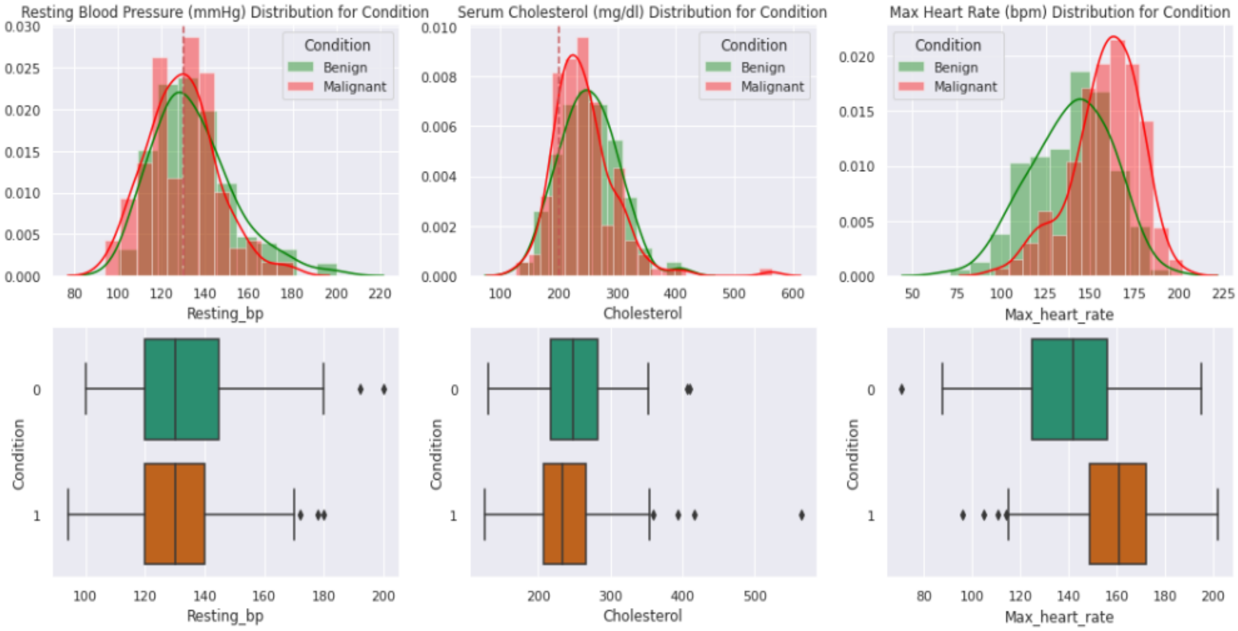
The feature Fasting_bs indicates whether subjects have elevated blood sugar levels before meals. Normal fasting blood sugar levels are below 100 mg/dl, while levels above 120 mg/dl indicate prediabetes or diabetes, potentially leading to pancreatic dysfunction and atherosclerosis, which can cause heart disease, kidney disease, and stroke. Observations reveal that the number of subjects with a fasting blood sugar level below 120 mg/dl is five times higher than those with elevated levels, suggesting most subjects maintain a healthy blood sugar level. While both females and males generally exhibit healthy fasting blood sugar levels, males are more prone to elevated levels above 120 mg/dl. Interestingly, individuals with normal fasting blood sugar levels are more frequently associated with malignant tumors compared to those with higher fasting blood sugar levels, highlighting a surprising relationship between blood sugar levels and the occurrence of malignant tumors.



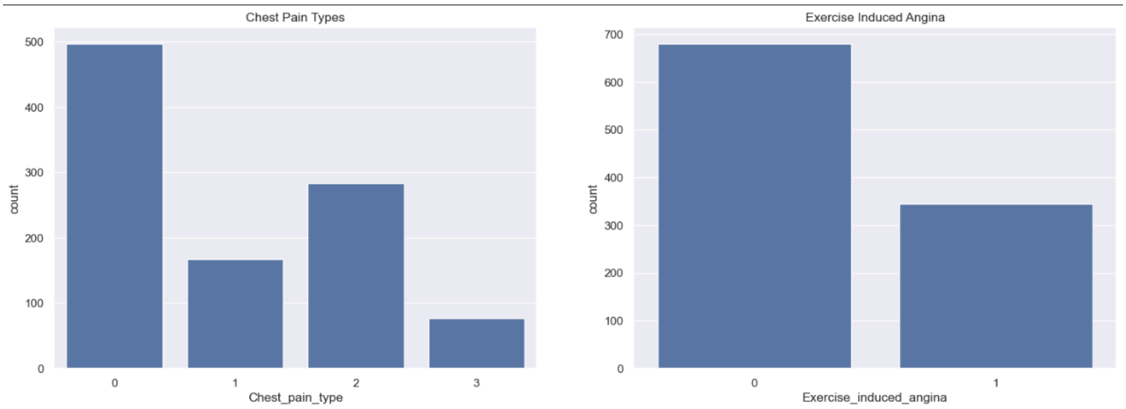
Blood pressure is divided into five levels by the American Heart Association: Normal, Elevated, Hypertension Stage 1, Hypertension Stage 2, and Hypertensive Crisis. A reading of less than 120 mmHg indicates normal blood pressure, while a reading of more than 130 mmHg indicates a stage of hypertension. Triglycerides, HDL ('harmless' cholesterol), and LDL ('dangerous' cholesterol) make up cholesterol, a kind of fat that travels through the bloodstream. Since LDL accumulation can obstruct blood flow, measuring serum cholesterol—which should be within the normal range of 125–200 mg/dL for people 20 years of age and above—is essential for determining the risk of heart disease. Adults typically have a resting heart rate between 60 and 100 bpm, and their maximum heart rate is 220 minus their age (100 bpm for a 70-year-old, and 200 bpm for a 20-year-old). The distributions of serum cholesterol (Cholesterol) and resting blood pressure (Resting_bp) are both somewhat right-skewed, with obvious outliers, particularly an extreme outlier in the cholesterol values. The majority of resting blood pressure values cluster around 130 mmHg, while the maximum heart rate distribution (Max_heart_rate) is more left-skewed than right. A high range of cholesterol is indicated by the mean serum cholesterol level. Even though the typical adult heart rate ranges from 150 to 200 bpm, the distribution of Max_heart_rate shows that certain people have exceptionally low maximum heart rates.



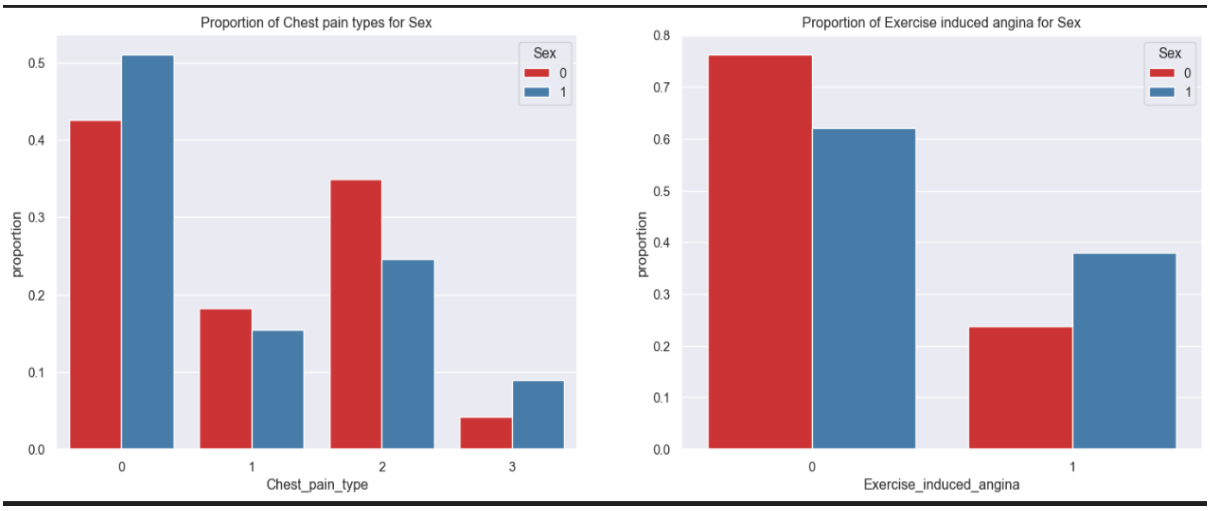
Compared to men, the majority of females had somewhat higher distributions in the over 130 mmHg blood pressure range, according to observations. Male and female cholesterol levels have right-skewed distributions, with females having a higher likelihood of having excessive cholesterol levels. The distribution of maximum heart rates for both genders is left-skewed, with most falling into the typical range of 150 to 200 bpm. Nonetheless, there are a few outliers in both the male and female populations that have abnormally low maximal heart rates.



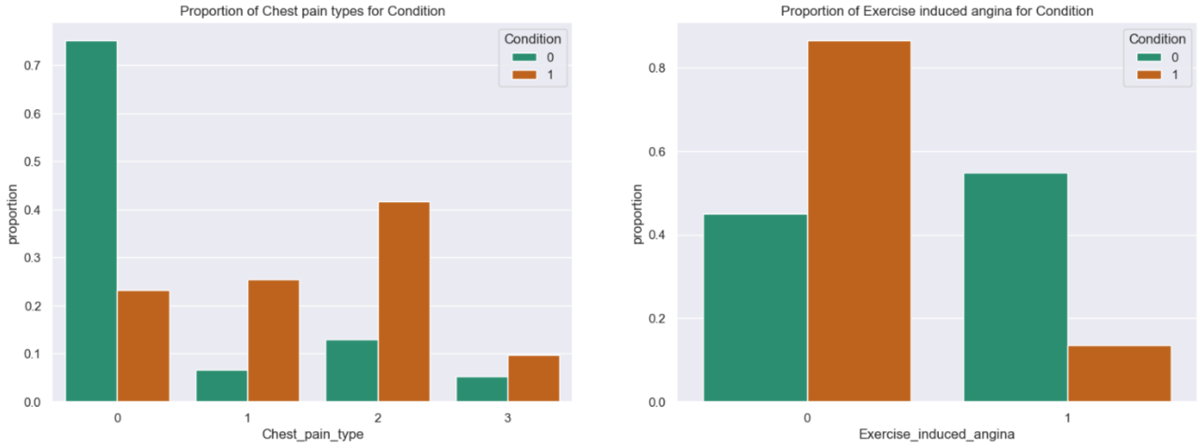
Evidence suggests that, when it comes to resting blood pressure, benign and malignant disorders primarily cluster around 130 mmHg. Since benign illnesses tend to have greater blood pressure more often, there is no obvious tendency for malignant conditions to have higher resting blood pressure. Malignant disorders are more likely to have greater serum cholesterol levels, even though most patients with benign and malignant conditions have cholesterol levels over the healthy range.



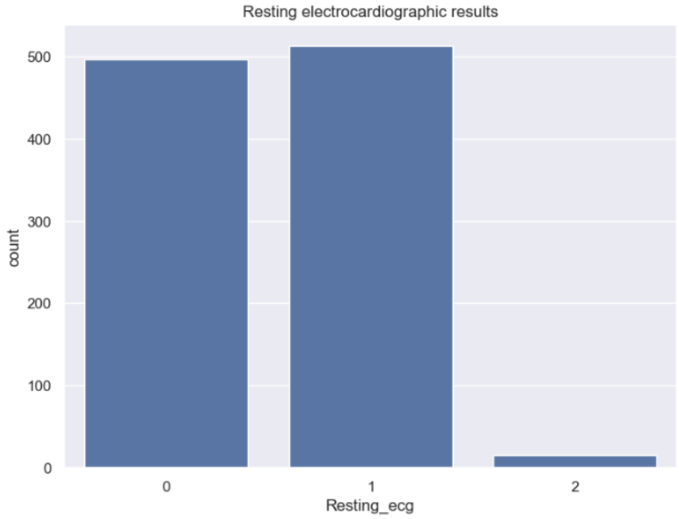
In this dataset, the symptoms of heart disease are influenced by the types of chest pain and exercise-induced angina. A common cardiovascular symptom, chest discomfort comes in four varieties: 1 is atypical angina (sudden, worsening pain owing to plaque), 2 is non-angina discomfort (associated to esophageal difficulties), 3 is asymptomatic (silent myocardial infarction, modest and often overlooked), and 0 is typical angina (induced by physical activity or stress). One kind of stable angina that happens during physical activity when the heart needs extra oxygen is exercise-induced angina. According to observations, asymptomatic angina is the least prevalent but most hazardous type of chest pain, whereas typical angina is the most common. Half of the cases have exercise-induced angina, and these cases fall under the category of typical angina.



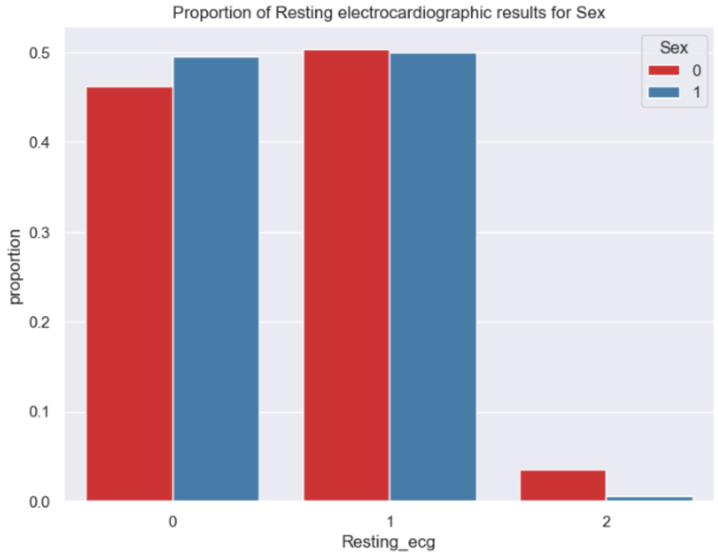
Approximately 40% of females have typical angina (type 0) and approximately 35% have non-angina pain (type 2), making them the majority. Compared to men, women are more prone to experience non-angina discomfort (2) and atypical angina (1). Typical angina is present in 50% of males (0). For both sexes, asymptomatic instances (3) are the least frequent; nonetheless, the proportion in males is twice that of females. In addition, men are more prone than women to get angina brought on by activity.



It was observed that non-angina pain (2), which was not related to heart attacks or heart disease, was surprisingly significantly associated with malignant tumours, with a proportion almost three times higher than in benign circumstances. Other than typical angina (0), there was a substantial correlation between malignant illnesses and several types of chest pain. Furthermore, there was no consistent association between the occurrence of exercise-induced angina (1) and malignant diseases.

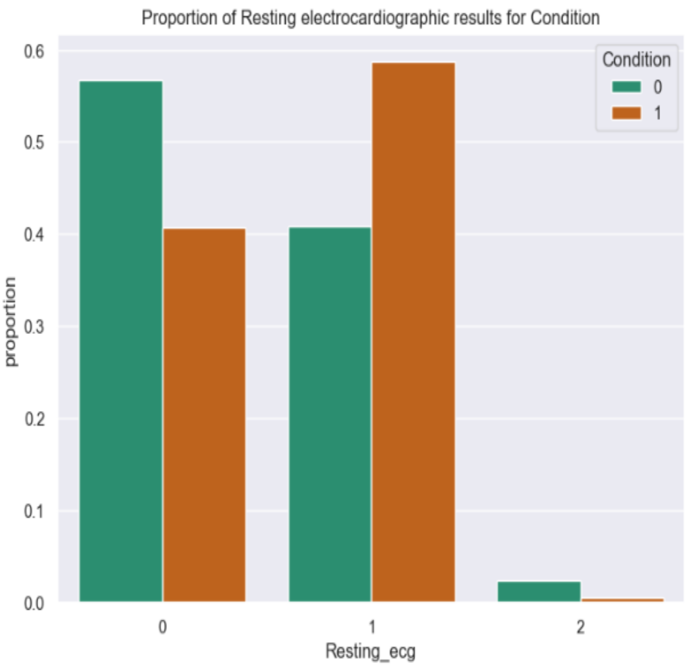


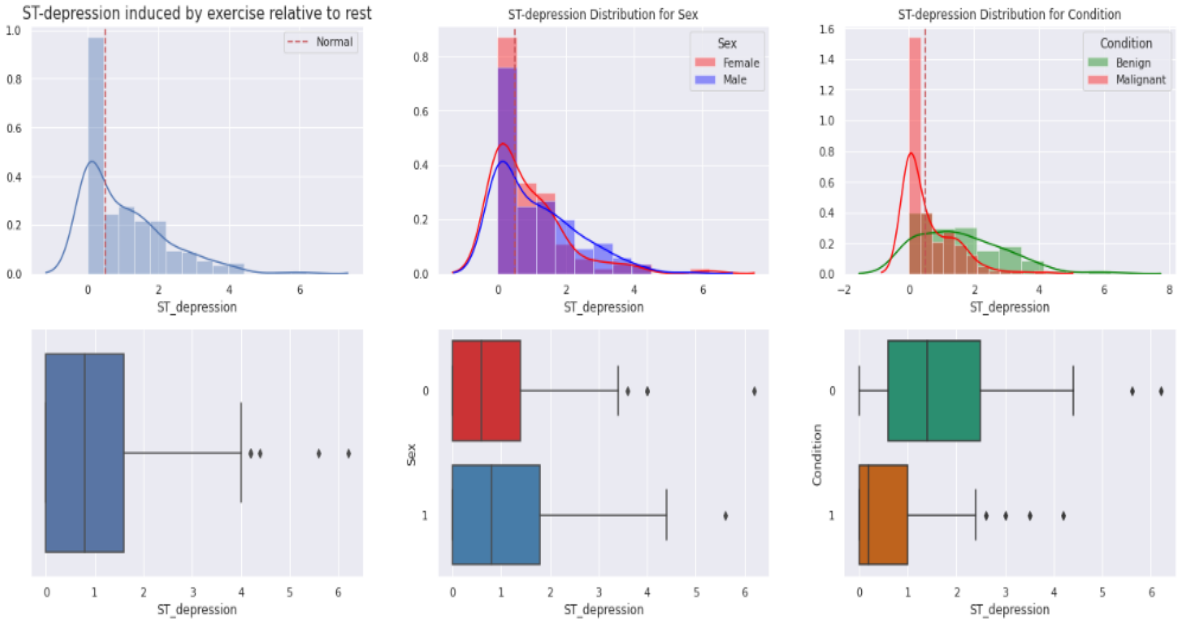
A number of techniques, including thallium stress tests, fluoroscopy, and electrocardiography (ECG), are used to monitor heart function. An irregular heart rhythm, heart attack, enlarged heart, and cardiac illness can all be identified by an ECG, which measures heart rate and rhythm. The results of the resting electrocardiogram (ECG) are Normal (0), Abnormal ST-T wave (1), and Probable or definite left ventricular hypertrophy (2). The latter refers to the thickening and expansion of the left ventricle of the heart, which puts undue strain on the heart muscle. Despite the rarity of left ventricular hypertrophy (2), observations show that half of the individuals had aberrant ST-T waves (1).



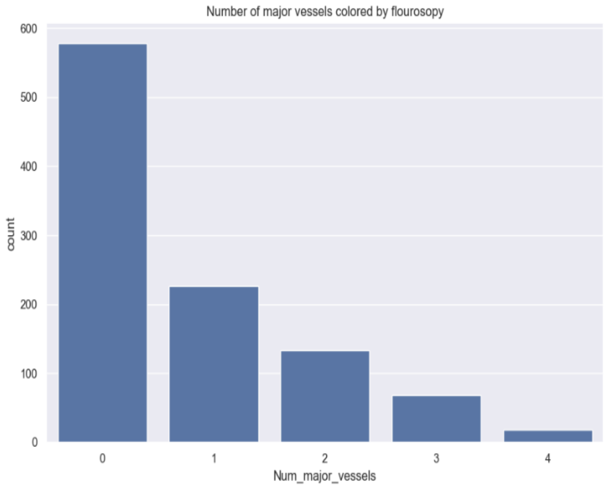
It can be shown that a somewhat higher percentage of females than males have abnormal ST-T waves (1) and probable or definite left ventricular hypertrophy (2). Specifically, more than 50% of females have abnormal ST-T waves.

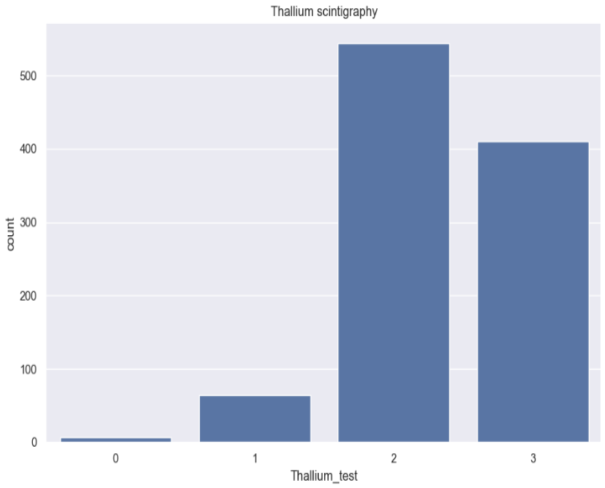
Although more than half of the subjects had a normal ST-T wave (0), observations show that this still has a role in malignant diseases, meaning that those with normal ST-T waves can still have heart disease. Furthermore, a malignant disease is explicitly linked to an abnormal ST-T wave (1), impacting approximately 58% of the participants.



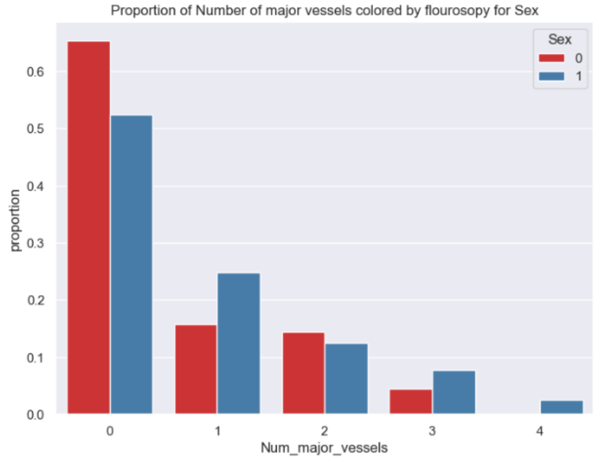


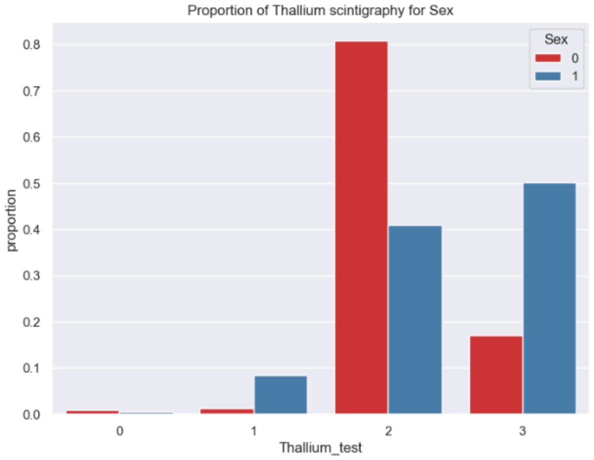
With most values falling below 1, ST_depression, which measures ST-segment depression brought on by exercise compared to rest, exhibits a distribution that is skewed to the right. The general pattern is reflected in the distributions of males and females. Malignant instances are sharply skewed to the right and contain multiple outliers, whereas benign cases are broadly distributed. The chance of a malignant condition is not clearly affected by the ST_depression values, even though typical ST-segment depression is flat or up to 0.5 mm.



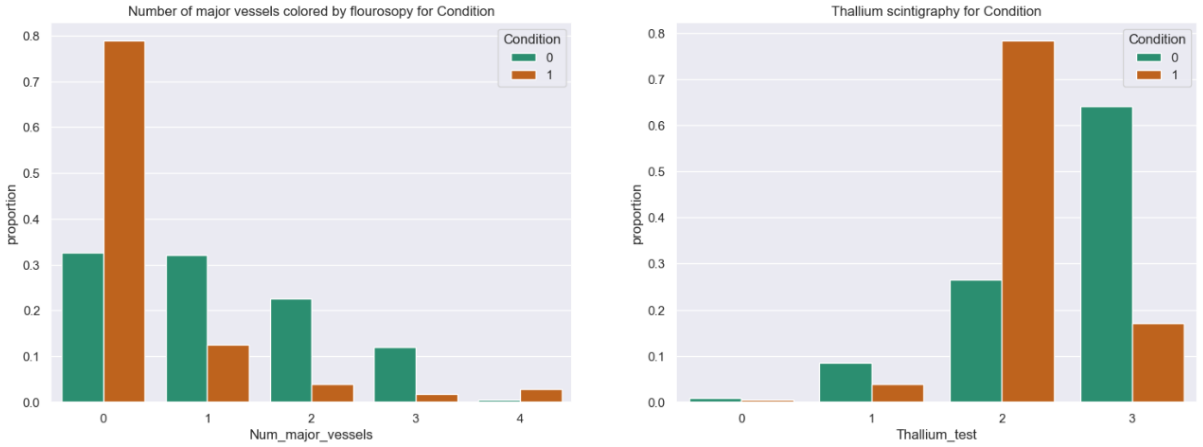


The number of major vessels (0–3) that are colored by fluoroscopy, an imaging technique used to examine blood flow through the coronary arteries and other body systems, is indicated by the variable num_major_vessels. Although nearly half of the participants in this sample had at least one clot, the majority do not have any colored clots in their blood arteries. According to the data description, the usual values for Thallium scintigraphy, which is utilized in the Thallium_test, are three (3 = Normal, 6 = Fixed defect, and 7 = Reversible defect). However, the actual coded results include four values (0, 1, 2, and 3), presenting an inconsistency.

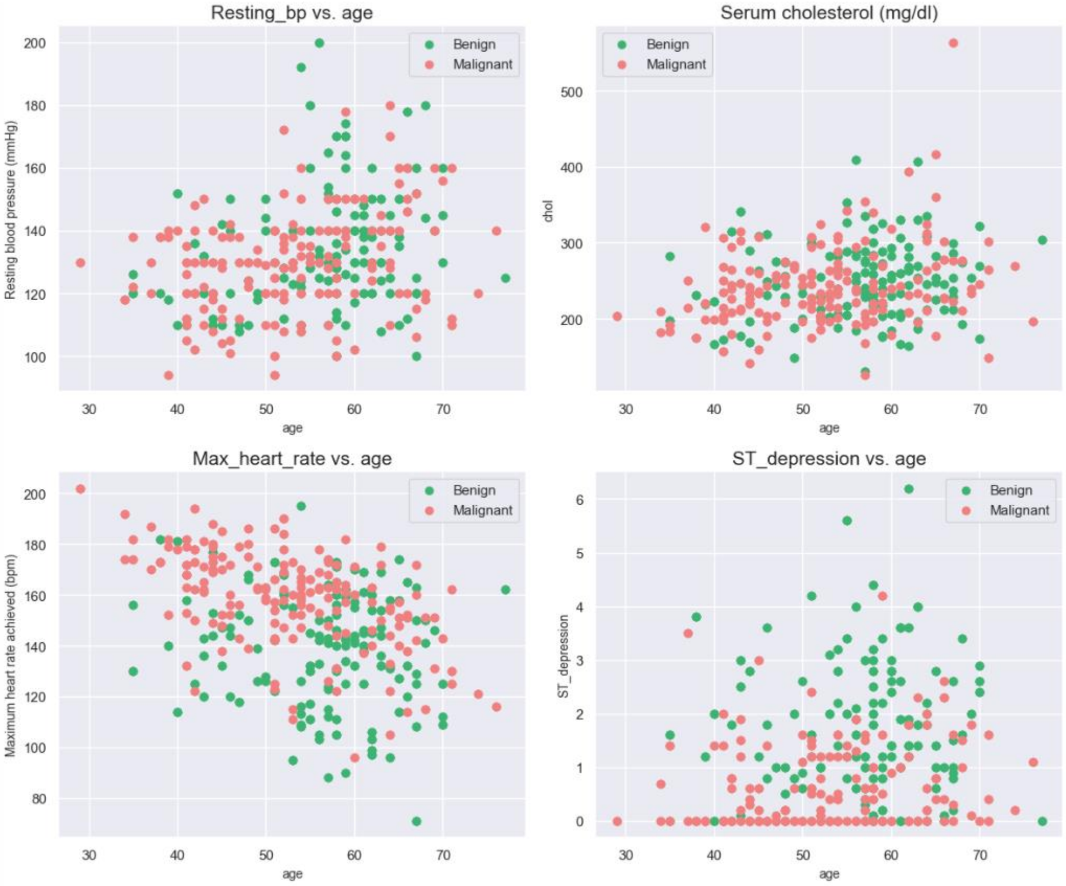




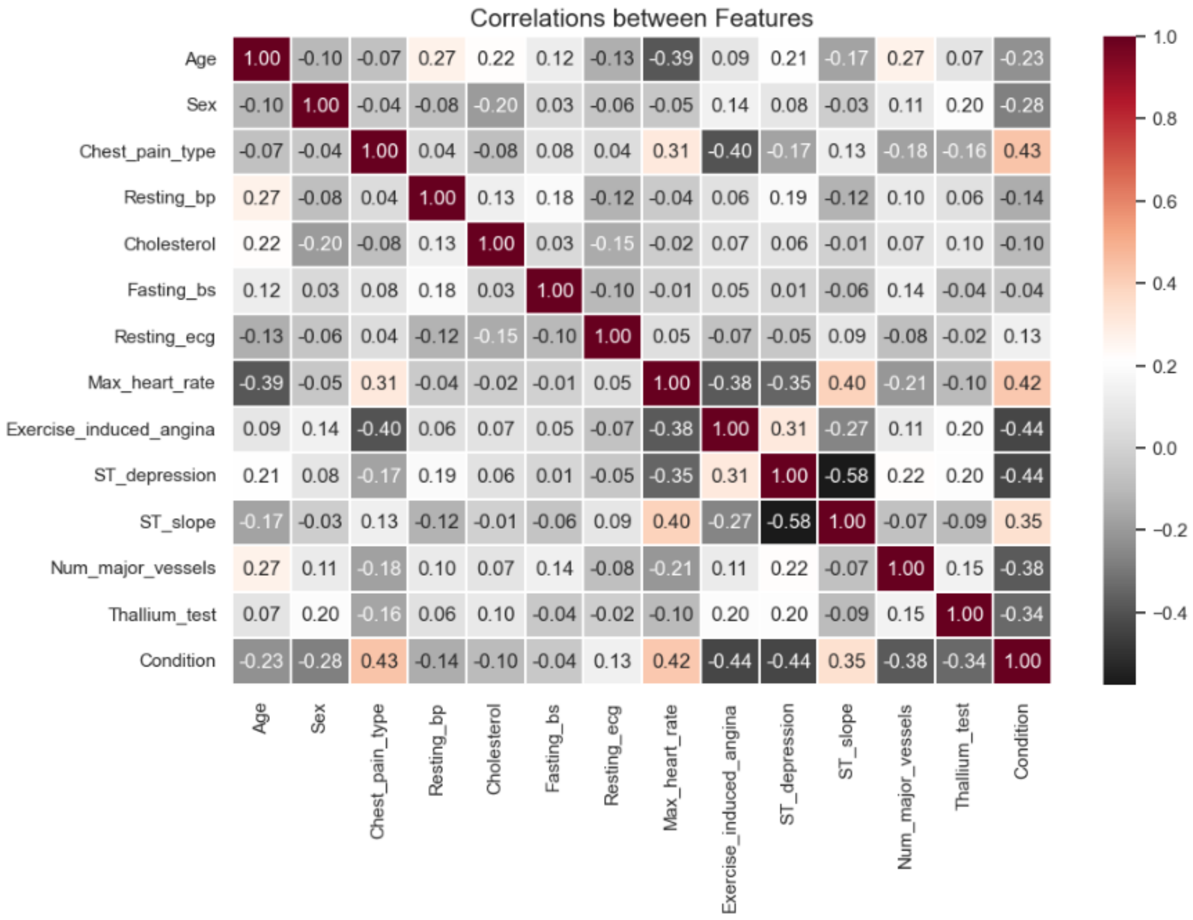
The given charts show possible differences in cardiovascular health across genders. When it comes to aberrant Thallium scintigraphy results and numerous arteries colored by fluoroscopy, which are frequently indicators of coronary artery disease, men are more likely than women to have these conditions. These results point to a higher risk of heart-related problems in the male group under study.



The given bar charts show a possible correlation between the quantity of large vessels stained by fluoroscopy, the findings of Thallium scintigraphy, and the existence of cancerous diseases. Across all categories of vessel count and Thallium test results, there is a notably larger proportion of patients with malignant diseases (Condition=1), even if the majority of people in both groups had no major vessels colored. This implies that there may be a connection between the emergence of cancerous illnesses and these cardiovascular markers.



Findings show that although blood pressure varies more widely in people over 50, there is no discernible correlation between resting blood pressure and age-related conditions. A higher occurrence of malignant illnesses is correlated with middle-aged people's increased cholesterol levels. While elderly people typically have benign tumors and lower maximal heart rates, younger populations are more likely to have both of these disorders. People with shallow ST-segment depression are more prone to have cancerous diseases regardless of their age.



Correlations between Features

	Age	Sex	Chest_pain_type	Resting_bp	Cholesterol	Fasting_bs	Resting_ecg	Max_heart_rate	Exercise_induced_angina	ST_depression	ST_slope	Num_major_vessels	Thallium_test	Condition
Age	1.00	-0.10	-0.07	0.27	0.22	0.12	-0.13	-0.39	0.09	0.21	-0.17	0.27	0.07	-0.23
Sex	-0.10	1.00	-0.04	-0.08	-0.20	0.03	-0.06	-0.05	0.14	0.08	-0.03	0.11	0.20	-0.28
Chest_pain_type	-0.07	-0.04	1.00	0.04	-0.08	0.08	0.04	0.31	-0.40	-0.17	0.13	-0.18	-0.16	0.43
Resting_bp	0.27	-0.08	0.04	1.00	0.13	0.18	-0.12	-0.04	0.06	0.19	-0.12	0.10	0.06	-0.14
Cholesterol	0.22	-0.20	-0.08	0.13	1.00	0.03	-0.15	-0.02	0.07	0.06	-0.01	0.07	0.10	-0.10
Fasting_bs	0.12	0.03	0.08	0.18	0.03	1.00	-0.10	-0.01	0.05	0.01	-0.06	0.14	-0.04	-0.04
Resting_ecg	-0.13	-0.06	0.04	-0.12	-0.15	-0.10	1.00	0.05	-0.07	-0.05	0.09	-0.08	-0.02	0.13
Max_heart_rate	-0.39	-0.05	0.31	-0.04	-0.02	-0.01	0.05	1.00	-0.38	-0.35	0.40	-0.21	-0.10	0.42
Exercise_induced_angina	0.09	0.14	-0.40	0.06	0.07	0.05	-0.07	-0.38	1.00	0.31	-0.27	0.11	0.20	-0.44
ST_depression	0.21	0.08	-0.17	0.19	0.06	0.01	-0.05	-0.35	0.31	1.00	-0.58	0.22	0.20	-0.44
ST_slope	-0.17	-0.03	0.13	-0.12	-0.01	-0.06	0.09	0.40	-0.27	-0.58	1.00	-0.07	-0.09	0.35
Num_major_vessels	0.27	0.11	-0.18	0.10	0.07	0.14	-0.08	-0.21	0.11	0.22	-0.07	1.00	0.15	-0.38
Thallium_test	0.07	0.20	-0.16	0.06	0.10	-0.04	-0.02	-0.10	0.20	0.20	-0.09	0.15	1.00	-0.34
Condition	-0.23	-0.28	0.43	-0.14	-0.10	-0.04	0.13	0.42	-0.44	-0.44	0.35	-0.38	-0.34	1.00

Observations reveal that the top correlated variables for Condition are:

Chest_pain_type: 0.43
Max_heart_rate: 0.42
ST_slope: 0.35

MODEL TRAINING:

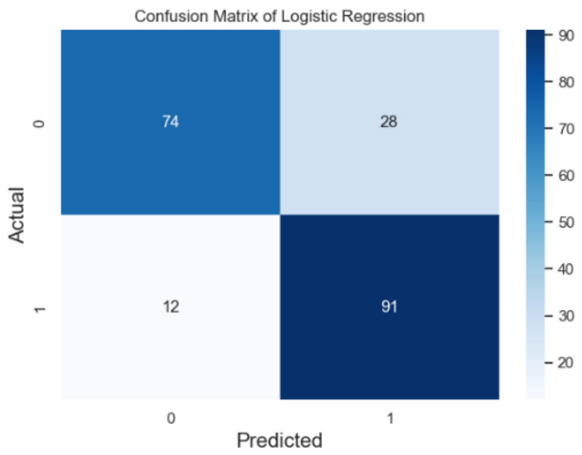
To divide the dataset into the training and testing data subsets, the code uses a train-test split. This is accomplished by using the scikit-learn train_test_split function, which assigns 80% of the data to the training set and the remaining 20% to the testing set. The unseen testing set acts as a rigorous evaluation ground to examine the model's predictive accuracy on fresh, unobserved data, offering an unbiased estimate of its generalization capabilities. The training set is crucial for model building and parameter adjustment.

Three primary classification algorithms are implemented and evaluated in the code:

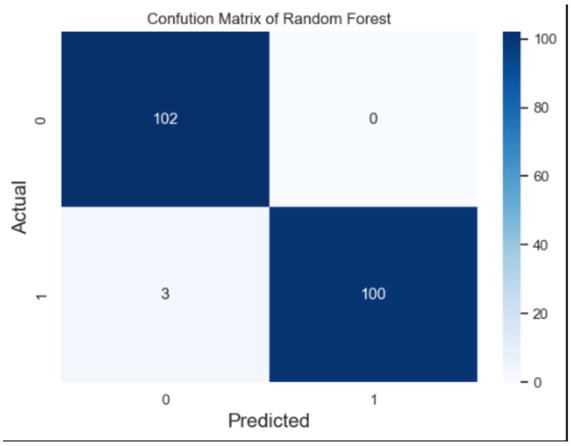
1. **Logistic Regression:** This model is suitable for binary classification problems like bankruptcy prediction. It estimates the probability of a company going bankrupt based on the input features.
2. **Support Vector Machine (SVM):** SVM is known for its effectiveness in classification tasks, particularly when dealing with complex decision boundaries. It aims to find the optimal hyperplane that separates the two classes (bankrupt and non-bankrupt).
3. **Random Forest:** Random Forest is an ensemble learning method that combines multiple decision trees to improve predictive accuracy. It's robust to overfitting and can handle various data types. Random Forest is known for its versatility and performance in various classification tasks.

With an astounding accuracy of 98.54%, Random Forest was determined to be the best accurate classifier for this dataset based on the evaluation. Compared to Logistic Regression (80.49%) and SVM (86.83%), Random Forest's ensemble learning approach appears to have been more successful in capturing the underlying patterns in the data, as seen by its improved prediction performance.

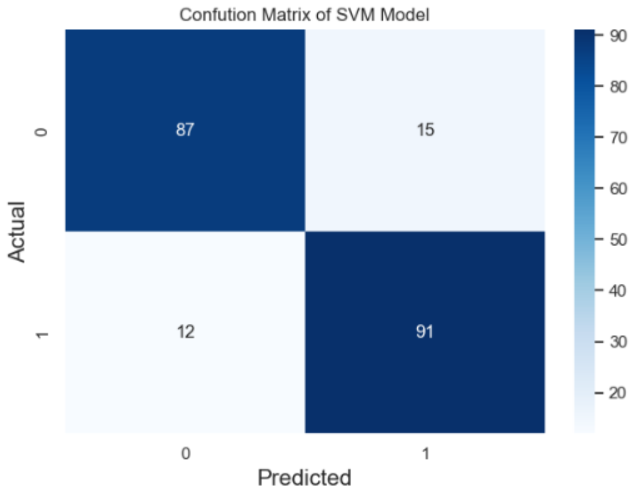
Despite being a more straightforward model, logistic regression might not have been able to adequately capture intricate relationships in the data. Despite its reputation for handling non-linearity, SVM may not have identified the best hyperplane for this particular dataset, giving Random Forest a higher accuracy.



A logistic regression model's performance is summarized in this confusion matrix. It reveals that the model misclassified 40 cases (28 false positives and 12 false negatives) out of 205 total cases, successfully categorizing 165 instances (74 true negatives and 91 true positives). Higher counts are represented by darker blue hues, which emphasize the model's superior ability to accurately detect both positive and negative cases.

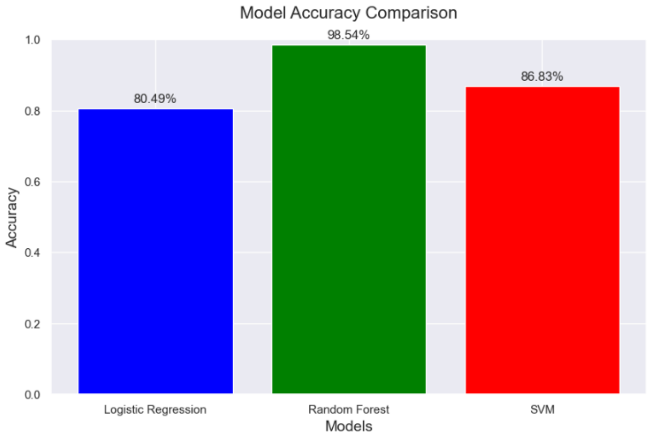


The Random Forest model's performance is shown in this confusion matrix. With only 3 misclassifications (3 false negatives and 0 false positives) out of 205 total cases, the model properly identified 202 occurrences (102 true negatives and 100 true positives). The top-left and bottom-right quadrants' dark blue hues draw attention to how well the model predicts both good and negative scenarios. This Random Forest model outperforms the other classifiers by a large margin, exhibiting outstanding predictive performance with zero false positives and very few false negatives.



The Support Vector Machine (SVM) model's performance is demonstrated by this confusion matrix. Out of 205 total examples, 177 occurrences (87 true negatives and 91 true positives) were correctly identified by the model, while 27 instances (15 false positives and 12 false negatives) were incorrectly classified. The top-left and bottom-right quadrants' dark blue hues imply excellent performance in accurately detecting both positive and negative cases. Though the SVM model performs well, the existence of misclassifications—specifically, the 15 false positives and 12 false negatives—indicates that, while it is equivalent to the logistic regression model, it is not as accurate as the Random Forest model that was previously observed.

CONCLUSION:



The study highlighted the efficacy of machine learning models in the prediction of heart disease. Specifically, the Random Forest classifier outperformed Logistic Regression (80.49%) and SVM (86.83%), obtaining the greatest accuracy of 98.54%. Age, the type of chest discomfort, maximal heart rate, ST depression, and the number of main vessels colored by fluoroscopy are among the important risk variables that have been found. Risk variables were shown to differ according to gender, which raises the possibility that customized preventative measures could be beneficial. Even though Random Forest performed better in terms of prediction, the study highlights how crucial it is to strike a balance between interpretability and accuracy in clinical contexts. The thorough examination of feature distributions and correlations opens the door to new understandings of heart disease risk factors and the development of more effective diagnostic and preventative cardiology techniques.