# Heart Disease Detection Using Classification Methods

Navjot Singh
DEPT OF *APPLIED MODELLING & QUANTITATIVE METHOD*
*Trent University*
Peterborough, Canada
nasingh@trentu.ca

Umang Sood
DEPT OF *APPLIED MODELLING & QUANTITATIVE METHOD*
*Trent University*
Peterborough, Canada
umangsood@trentu.ca

Dhruvil Shah
DEPT OF *APPLIED MODELLING & QUANTITATIVE METHOD*
*Trent University*
Peterborough, Canada
dhruvilshah@trentu.ca

Manvesh Liddar
DEPT OF *APPLIED MODELLING & QUANTITATIVE METHOD*
*Trent University*
Peterborough, Canada
manveshliddar@trentu.ca

Samaun Sarwar Khan
DEPT OF *APPLIED MODELLING & QUANTITATIVE METHOD*
*Trent University*
Peterborough, Canada
samaunsarwarkhan@trentu.ca

***Abstract*—Heart diseases remain a prominent cause of death worldwide. This research addresses the urgent necessity for solutions that can identify individuals at risk of heart disease promptly and precisely. Leveraging data visualization techniques and machine learning models, a reliable and efficient tool will be developed to assist healthcare professionals in this task. Beginning with detailed Exploratory Data Analysis (EDA), visualizations will be employed to unravel complex patterns within the Cleveland Heart Disease dataset. Subsequently, multiple classification algorithms will be trained and evaluated using various metrics. The objective is to identify the most effective method for accurately detecting heart disease. The appropriate classification model will enable timely diagnosis and treatment of heart diseases.**

***Keywords — classification, logistic regression, XGBoost, SVM, heart disease, visualization, accuracy, precision, recall***

## I. Introduction

Cardiovascular diseases, particularly heart disease, stand out as a leading cause of death across the world. Heart disease often progresses silently, with symptoms appearing only in advanced stages. Early detection and accurate risk assessment are crucial in preventing and managing heart diseases. The main issue we are addressing is the urgent need to create solutions that can identify people at risk of heart disease early and accurately measure their risk factors. In our research, we aim to use data visualization techniques along with machine learning models to develop a reliable as well as efficient tool that can assist healthcare professionals in identifying individuals at risk of heart disease, ultimately leading to improved patient outcomes and reduced healthcare costs.

A detailed Exploratory Data Analysis (EDA) was conducted, focusing on visualizing the features available with the help of univariate and multivariate techniques such as violin plots and scatterplots to uncover complex patterns. The analysis provides crucial insights into the dataset. Subsequently, multiple classification algorithms – Logistic Regression, XGBoost, and Support Vector Machine (SVM) – were trained and evaluated using metrics like accuracy, precision, recall, and F1 score. The most effective method for accurately detecting heart disease in individuals was identified after comparing the classification algorithms. Given the life-or-death nature of the problem, our primary goal is to minimize misclassifications.

## II. Previous Work

In the realm of medical science, various applications employing diverse machine learning algorithms are actively utilized for data analysis and advancement. Recent research in healthcare has showcased instances of machine learning utilization, such as identifying COVID-19 through X-rays, detecting tumors via MRIs, predicting cardiac issues, dengue, strokes, and cancers.

In the study published in the Siberian Journal of Life Sciences and Agriculture, various machine learning algorithms were employed to predict cardiovascular diseases (CVDs). The algorithms included NBS, KNN, DT, Logistic Regression, SVM, RF, CNBC, LDA, RBF, and XGBoost. The research utilized the automatic search for hyperparameters MMO to optimize model performance. Results indicated that the Random Forest (RF) and XGBoost algorithms exhibited higher accuracy, achieving overall classification accuracies of 0.88 and 0.94, respectively [1].

To diagnose heart disease, Sreejit Ramakrishnan and Bhasutkar Mahesh used machine learning models such as Random Forest, Decision Tree Classifier, Multilayer Perceptron, and XGBoost in a study published in the International Journal of Engineering Technology and Management Sciences. The study concentrated on characteristics such as age, cholesterol levels, and chest discomfort. With Huang as the initialization method, the study used a k-modes clustering strategy to improve classification accuracy [2].

## III. Methodology

### A. Dataset

The Cleveland Heart Disease dataset, available in the UCI Machine Learning Repository (Heart Disease - UCI), has been used for the research. The dataset has 76 attributes, but all previously published studies focus on a subset of 13 features, which we will also use in our research. These attributes include age, sex, chest pain type (cp), resting blood pressure (trestbps), fasting blood sugar (fbs), resting electrocardiographic (restecg), serum cholesterol (chol), fasting blood sugar (fbs), maximum heart rate achieved (thalch), exercise-induced angina (exang), ST depression (oldpeak), slope of the peak exercise (slope), number of major vessels (ca) and types of defect (thal). In addition, we have a target variable, disease_present, which has 5 distinct values (0,1,2,3,4). The presence of disease is indicated by values ranging from 1 to 4, while 0 indicates its absence. The target variable was mapped to two values, 0 and 1, indicating the absence or presence of heart disease, respectively.

### B. Computational Resources and Tools Used

The project has been implemented using Python within a Jupyter Notebook environment, utilizing essential libraries such as matplotlib, seaborn, scikit-learn, numpy, and pandas. Matplotlib and Seaborn are popular Python libraries used for data visualization. On the other hand, scikit-learn is a powerful machine learning library in Python that provides a wide range of tools for various aspects of machine learning. Our analysis is conducted without the need for extra computational resources. The CPU cores that are readily accessible are sufficient for executing the project.

### C. Exploratory Data Analysis (EDA)

The project commenced with loading and understanding the dataset. The target variable was renamed to 'disease_present'. Some of the variables which were not required, including 'id', were dropped from the dataset. Following this, we dealt with missing values in each of the features.

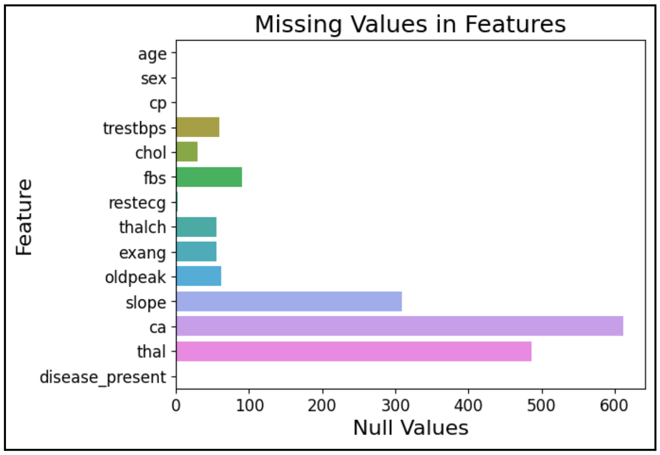

Figure 1. Count of Missing Values

The bar chart, shown in Figure 1, illustrates the count of missing values for each feature. Notably, certain features such as age, sex, cp, and restecg exhibited no missing values. Conversely, tretbps, chol, fbs, thalch, exang, and oldpeak had fewer than 100 null values each. Other features i.e. slope, ca, and thal had more than 300 missing values each, indicating a higher prevalence of missing data. Null or missing values for categorical variables such as fbs and exang were substituted using the backfill technique. Meanwhile, for numerical features, missing values were filled in by using the mean value of the respective feature.

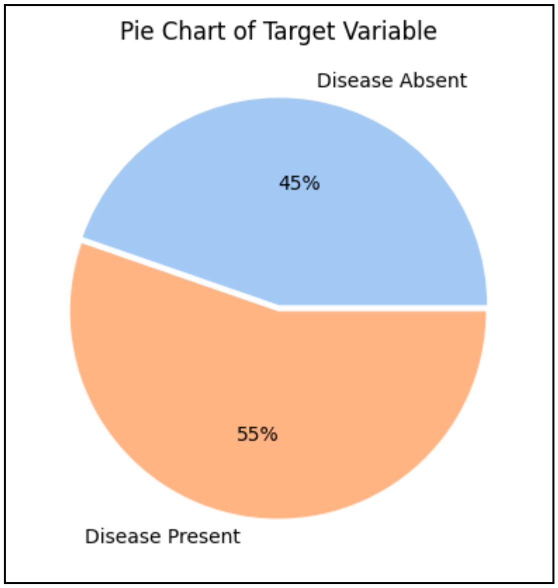

Figure 2. Count of Missing Values

Next, a pie chart (Figure 2) was employed to study the proportion of data points belonging to each category, disease_absent (0) and disease_present (1). The dataset consists of 45% of participants without the disease and 55% of participants with the disease.

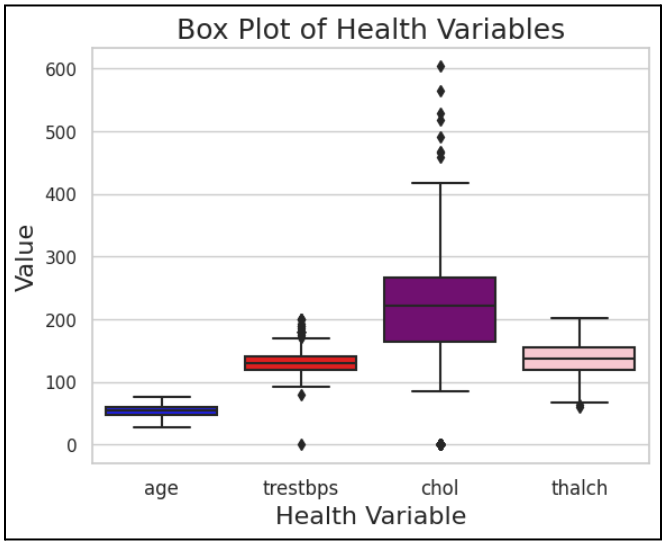

Figure 3. Box Plot for Health Variables

Box plots were employed to examine the distribution of crucial features such as age, trestbps, chol, and thal. Additionally, the boxplots were used to emphasize any outliers present in these features. The boxplot (Figure 3) displays distinctive characteristics for health variables. 'Chol' exhibits the highest skewness and interquartile range (IQR), indicating a significantly skewed distribution and a broad range of values. The existence of excessive cholesterol levels among individuals is demonstrated by the presence of outliers. On the other hand, the 'Age' variable has the shortest IQR, with no outliers and a median of 50. Age data distribution looks more compact, with less skewness and a tighter range. The variable 'thalch' exhibits the second-highest IQR and has outliers situated beneath the whisker. Its median, which is approximately 140, suggests a modest central tendency. Lastly, 'trestbps', which has a median of around 130, exhibits a sizable number of outliers, suggesting significant variances among individuals.

To examine the distribution of Maximum Heart Rate Achieved by sex, violin plots (Figure 4) were used. With its comprehensive display of the central tendency, dispersion, and density of values, this visualization offers insight into potential differences in the maximum heart rates of male and female.

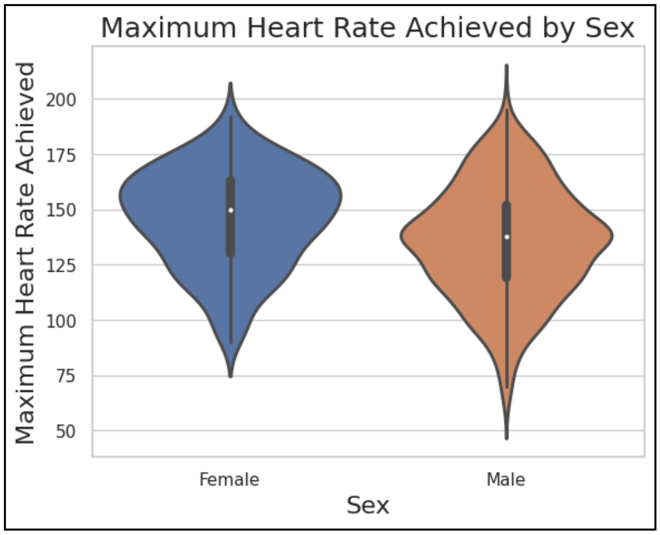

Figure 4. Violin Plot for Maximum Heart Rate Achieved by Sex

There are noticeable variances in the violin plot that compares the Maximum Heart Rate Achieved by sex. Females' median heart rate is 150, meaning that, on average, they tend to have a higher maximum heart rate than males, who have a median heart rate of 135. Although there is no difference in the Interquartile Range (IQR) between the sexes, the higher median for females points to a possible gender difference in cardiovascular responses. Maximum heart rates are centered for males and somewhat tilted to the left for females, according to the density plot concentration. This indicates that, whereas the distribution of males is more concentrated around the median, the distribution of females is broader and indicates a higher likelihood of reaching increased maximum heart rates. Males' longer whiskers suggest that there is more variation in maximal heart rates among the male population. In conclusion, the violin plot sheds light on the distribution and variability of maximum heart rate achieved among genders in addition to highlighting variations in central tendency.

Next, a pair plot (Figure 5) was used to examine the relationships between the Heart Disease dataset's health variables, including age, trestbps, and chol. The visualization, which uses color-coded markers to differentiate between the presence of the disease, reveals trends and possible correlations which increase our knowledge of the variables linked to heart disease.

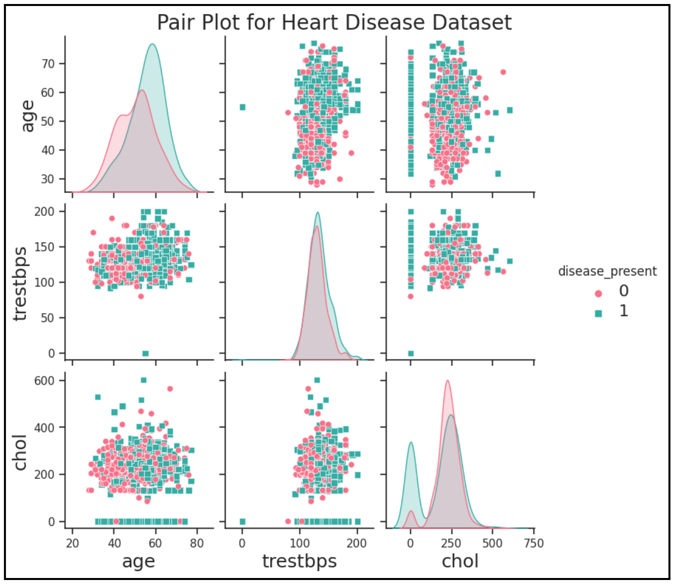

Figure 5. Pair plot for Health Variables

The pair plot provides detailed information about the associations between health variables. Interestingly, the age-disease relationship displays a unique pattern in which the disease absence density curve peaks approximately 60 years of age, but the presence curve exceeds it, suggesting a right-skewed distribution with higher rate of disease, especially in those 60 years of age and older. An analysis of the relationship between resting blood pressure (trestbps) and age reveals that the majority of disease presence occurs in the 50–80 age group, indicating a possible association between high blood pressure and disease in this population. Examining cholesterol (chol), the pair plot reveals that people with low cholesterol show signs of disease at any age, but people with moderate cholesterol show signs of disease in people 50 years of age and older.The plot also shows us that when cholesterol is at 0 and blood pressure is 100 or more, there is a higher chance of having heart disease. It's interesting that for people with cholesterol at 0, there is a lot more heart disease compared to those with cholesterol between 100 and 500, where both disease absence and no disease presence are seen while absence is more dominant. This makes it clear that understanding who might be at risk for heart disease is quite complicated and involves different factors like age and other health variables. The plot helps us see these connections and gives us information about potential reasons for heart disease.

### *D. Data Transformation for Modelling*

In the next step of preparing the data for modeling, categorical features in the dataset underwent label encoding. This transformation was executed using the scikit-learn library's LabelEncoder. The categorical features, namely 'cp' (chest pain type), 'restecg' (resting electrocardiographic results), 'slope' (slope of the peak exercise ST segment), 'thal' (thalassemia), 'fbs' (fasting blood sugar), and 'exang' (exercise-induced angina), were encoded into numerical representations. Additionally, the 'sex' feature was modified, with 'Male' and 'Female' being replaced by 1 and 0, respectively. This label encoding ensures that categorical variables are appropriately represented numerically, facilitating their incorporation into machine learning models for effective analysis and prediction.

The relationships between particular features in the dataset were visually represented using the correlation matrix heatmap to investigate the strength and direction of the relationship between the variables. The heatmap uses a color system, where neutral shades indicate little to no association and red indicates a stronger positive correlation and blue indicates a stronger negative correlation.

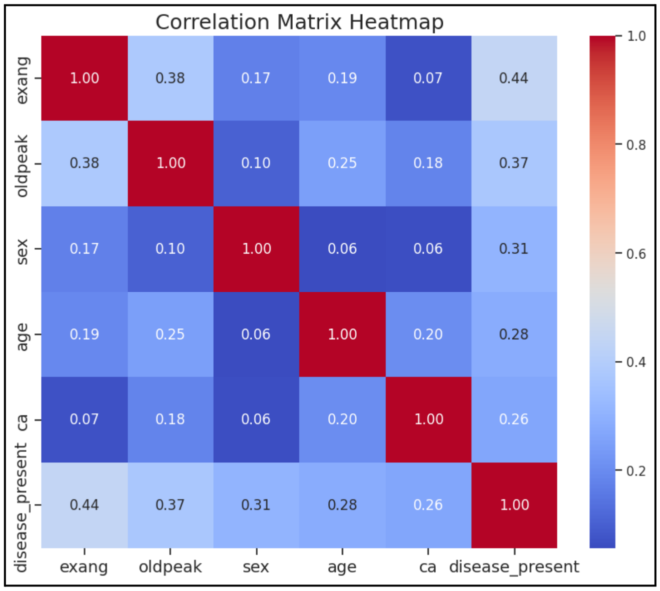

Figure 6. Correlation Matrix Heatmap

The correlation map demonstrates the relationship between multiple variables and the likelihood of heart disease. The correlation between 'disease_present' and 'exang' is 0.44, indicating a moderate positive relation between exercise-induced angina and cardiac disease. The degree of ST depression during activity, or 'oldpeak,' is moderately positively correlated (0.37) with the existence of disease. The variable 'sex' has a modestly positive correlation (0.31), indicating a moderate association between gender and heart disease. The presence of a disease is moderately positively correlated with age (0.38). There is a weak to moderate positive correlation (0.26) between the number of major vessels coloured by fluoroscopy ('ca') and the existence of heart disease. In short, the correlation map explains how these variables relate to the risk of heart disease.

### E. *Classification Modeling*

For predicting whether an individual is suffering from heart disease, a range of machine learning models are available. We have employed Logistic Regression, XGBoost, and Support Vector Machine (SVM) classification algorithms. Each of these algorithms provides unique advantages, and we have attempted to determine the most suitable one after training and evaluating their performance.

In the process of training and evaluating our model on the heart disease dataset, we divided the data into training and testing sets, with 80% (736) allocated to training and 20% (184) to testing. To ensure uniformity across features, standard scaling was applied. Subsequently, each model was trained using the training set (X_train), and predictions (y_pred) were generated using the test set (X_test).

*a) XGBoost:* Extreme Gradient Boosting, or XGBoost, is a distributed, scalable gradient-boosted decision tree (GBDT) machine learning framework. It is the top machine learning package for regression, classification, and ranking issues and offers parallel tree boosting. In our research, we have used XGBoost to train a binary classification model. Initially, an instance of the XGBClassifier, a classifier specifically designed for classification tasks, is created with default parameters. The default max_depth is set to 3, controlling the maximum depth of individual trees to prevent overfitting. The learning_rate, with a default of 0.1, determines the step size shrinkage in each boosting iteration. The number of boosting rounds is specified by n_estimators (default: 100), representing the quantity of trees added to the model. subsample (default: 1.0) and colsample_bytree (default: 1.0) introduce randomness during training by controlling the fraction of samples and features used for each tree, respectively. The default objective is 'binary:logistic,' aligning with the default task of binary classification. Additionally, parameters such as gamma (default: 0), reg_alpha (default: 0), and reg_lambda (default: 1) contribute to controlling tree complexity and introducing regularization. This classifier is then trained on the heart disease dataset.

*b) SVM:* The Support Vector Machine (SVM) is a supervised learning algorithm suitable for both classification and regression problems, capable of handling linear and nonlinear relationships. In our study, we have used the SVM kernel to do classification to train the model. It is critical to consider crucial features when diving into SVM's default parameters. With a default value of 1.0, the regularization parameter C is critical in balancing the trade-off between correctly categorizing training instances and maximizing the margin. The default kernel function is 'rbf', a versatile option suitable for a wide range of datasets, while alternatives such as 'linear,' 'poly,' and 'sigmoid' are also available. The model was then trained and the predictions were made.

*c) Logistic Regression:* Logistic Regression is a statistical method employed for binary classification problems, where the outcome variable is categorical with only two possible values (true or false). In our study, we applied a logistic regression model to address the classification task. The model utilized predictor variables such as age, cholesterol levels, exercise habits, and chest pain type to assess the likelihood of the binary outcome represented by the target variable (disease_present) — specifically, whether an individual has heart disease or not. As mentioned above, we enhanced data performance by splitting it into training and test sets and applying standard scaling. The logistic regression model was initialized with default parameters i.e., it used the default settings for the logistic regression algorithm without explicitly specifying custom values for its parameters. Afterwards, the model was trained on the scaled training data to make predictions on the test set.

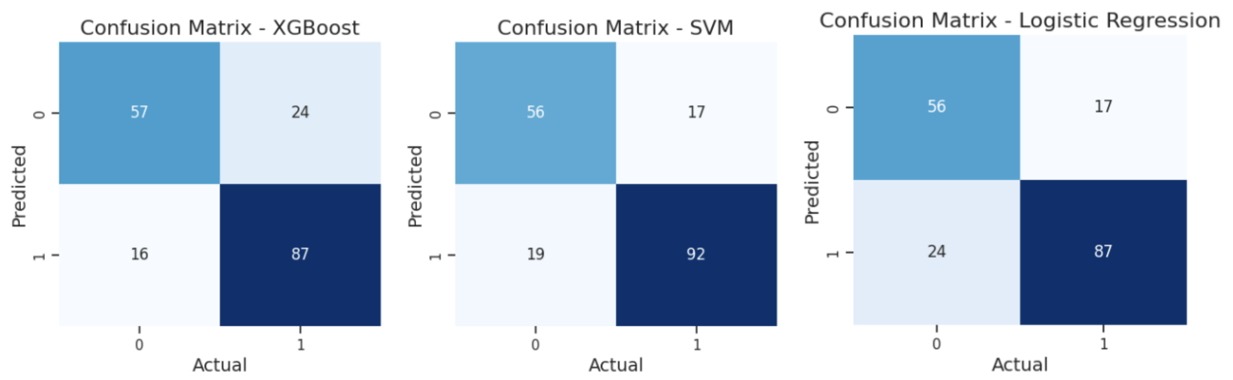

Fig 7. Confusion Matrix for all Classification Models

## IV. Results

To evaluate the classification models, we have constructed a confusion matrix for the test data (Figure 7). Using this matrix, we have calculated metrics such as accuracy, precision, recall, and the F1 score. Accuracy measures the model's overall performance. Recall assesses the model's ability to identify all patients with heart disease. Precision evaluates the model's accuracy in positive predictions. The F1 score balances precision and recall. Using these metrics, we have compared the different classification algorithms to determine which one accurately identifies individuals with heart disease.

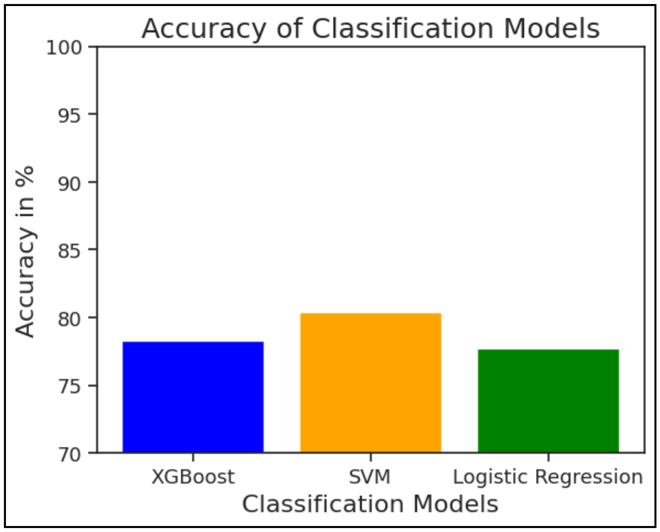

Fig 8. Accuracy of Classifiers

The accuracy of three different classification models used to identify heart disease is clearly represented in the bar chart (Figure 8). Support Vector Machine (SVM) demonstrated the highest accuracy at 80%, while XGBoost reached 78%. Another model under consideration, logistic regression, showed a 77% accuracy rate. Even though the accuracies are comparatively close, SVM is the most accurate in detecting heart disease.

In assessing the performance of models for disease presence (Class 1), we used three key metrics—Recall, Precision, and F1 Score—visualized through a line chart (Figure 9). For XGBoost, the Recall, representing the ability to capture actual disease cases, was 78%, with a Precision of 84% and an F1 Score of 81%. SVM demonstrated higher Recall at 83%, along with a Precision of 84% and an F1 Score of 84%. Logistic Regression mirrored XGBoost's Recall and Precision at 78% and 84%, respectively, with an F1 Score of 81%. SVM's higher Recall (83%) suggests that the model is effective at capturing a significant portion of individuals with the disease. Precision is the same for all the models resulting in fewer false positives. SVM's high F1 Score indicates it has a good balance between capturing disease cases and minimizing false positives.

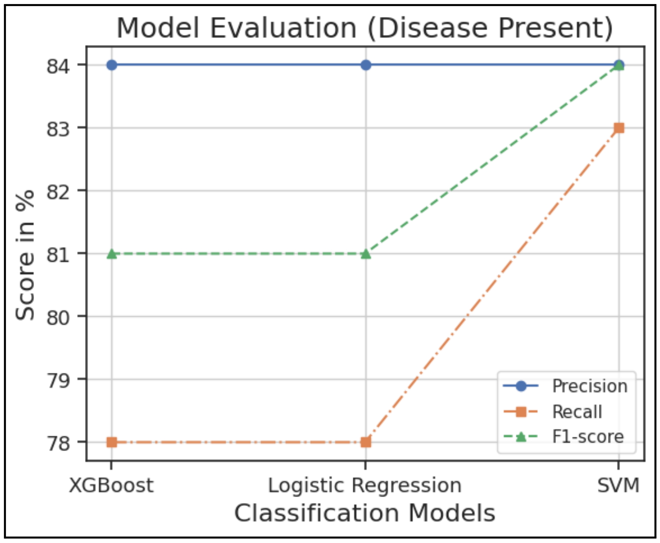

Fig 9. Model Performance in Predicting Disease Presence

In the assessment of models for identifying instances where disease is absent (Class 0), three crucial metrics—Recall, Precision, and F1 Score—were employed and visualized using a line chart (Figure 10). For XGBoost, the Recall, signifying the model's ability to correctly identify actual instances of disease absence, stood at 78%, accompanied by a Precision of 70% and an F1 Score of 74%. SVM exhibited a slightly lower Recall at 77%, coupled with a Precision of 75%, resulting in an F1 Score of 76%. Logistic Regression shared Recall with XGBoost at 77% and had a precision and F1 score of 70% and 73% respectively. The higher Recall for XGboost implies its effectiveness in capturing a significant portion of cases where disease is absent, while the high Precision of SVM indicates its ability to detect fewer false positives more efficiently. The F1 Score, as a balanced measure, underscores SVM's ability to strike a good equilibrium between capturing instances of disease absence and minimizing false positives in this specific evaluation context.

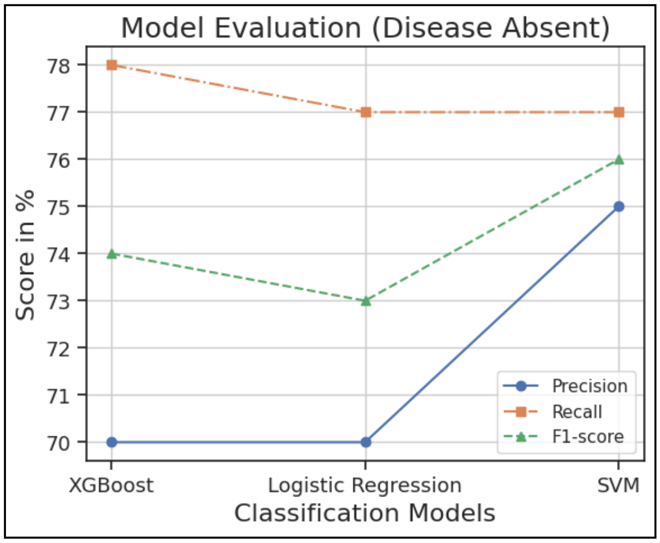

Fig 10. Model Performance in Predicting Disease Absence

## V. Conclusion

In summary, our evaluation of classification models for heart disease detection encompassed a range of metrics, including accuracy, precision, recall, and the F1 score. Support Vector Machine (SVM) emerged as the most accurate in identifying heart disease, showcasing superior overall performance. The metrics for disease presence demonstrated SVM's efficacy in capturing a significant portion of individuals with the disease, with an impressive F1 score reflecting a balance between precision and recall. However, it is crucial to acknowledge the competitive performance of XGBoost and Logistic Regression, each with its strengths. XGBoost displayed effectiveness in correctly identifying cases where disease is absent, while Logistic Regression demonstrated competitive performance, sharing similar recall and precision percentages with XGBoost. While SVM excelled in precision, detecting fewer false positives efficiently, Logistic Regression's balanced performance suggests its potential applicability in various scenarios. The choice among these models should consider specific priorities, such as the emphasis on minimizing false positives or capturing a broad spectrum of disease instances, highlighting the importance of context-specific considerations in selecting the most suitable model for heart disease prediction.

Fine-tuning the parameters employed during training can enhance the efficacy of these classification techniques. Another avenue for improvement lies in refining the preprocessing and treatment of data. This research can be extended to evaluate the performance of other classification techniques in predicting the presence or absence of heart disease in patients. A suitable classification model will facilitate prompt identification and treatment of heart conditions, alleviating strain on medical resources.

## References

1. Pavlova, A. I. (2023). Application of Machine Learning Algorithms for Heart Disease Prediction. Siberian Journal of Life Sciences and Agriculture, 15(3), 475-496.https://doi.org/10.12731/2658-6649-2023-15-3-475-496

2. Ramakrishnan, S., & Mahesh, B. (2023). Heart Disease Prediction Using Machine Learning. International Journal of Engineering Technology and Management Sciences, 7(6), DOI: 10.46647/ijetms.2023.v07i06.027.

3. Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J.-J., Sandhu, S., Guppy, K. H., Lee, S., & Froelicher, V. (1989). International application of a new probability algorithm for the diagnosis of coronary artery disease. The American Journal of Cardiology., 64(5),304–310.https://doi.org/10.1016/0002-9149(89)90524-9

4. Ayatollahi, H., Gholamhosseini, L., & Salehi, M. (2019). Predicting coronary artery disease: A comparison between two data mining algorithms. BMC Public Health, 19(1), 448–448. https://doi.org/ 10.1186/s12889-019-6721-5

5. Nashif, S., Raihan, M.R., Islam, M.R., & Imam, M.H. (2018). Heart Disease Detection by Using Machine Learning Algorithms and a Real-Time Cardiovascular Health Monitoring System. World Journal of Engineering and Technology, 06, 854-873. https://www.scirp.org/journal/paperinformation.aspx?paperid=88650

6. Liu, X., Wang, X., Su, Q., Zhang, M., Zhu, Y., Wang, Q., & Wang, Q. (2017). A Hybrid Classification System for Heart Disease Diagnosis Based on the RFRS Method. Computational and Mathematical Methods in Medicine,2017,8272091–11. https://doi.org/10.1155/2017/8272091