Predicting Heart Diseases with Machine Learning

Syed Shahzain, David Morales, Nathan Ritchie,
Yuke Hou, Jasleen Bakshi

Abstract

Cardiovascular diseases are the leading cause of death worldwide, accounting for 17.9 million annual fatalities (World Health Organization, 2021). This study leverages a Heart Failure Prediction dataset from Kaggle to develop machine learning models for early detection of heart disease. Out of the five models chosen (Logistic Regression, Random Forest, XGBoost, Gradient Boosting, Support Vector Machine (SVM)), Random Forest emerged as the best-performing model, with an AUC-ROC of 0.947. When feature importance was analyzed, ST_Slope_Up identified as the most influential predictor. These findings underscore the potential to improve early detection and intervention of cardiovascular diseases through machine learning. Recommendations for future research include focusing on incorporating additional features and validating models across diverse datasets.

Introduction

Machine learning (ML) is emerging as a powerful tool in healthcare research to analyze complex dataset and uncover patterns of multifaceted health issues (Habehh & Gohel, 2021). Cardiovascular diseases, the leading cause of mortality around the world, contributes to 17.9 million annual deaths (World Health Organization, 2021), highlighting potential use of ML for predictive models to identify at-risk individuals and improve early detection. This study uses the Heart Failure Prediction dataset from Kaggle, which contains 918 records and 11 attributes crucial for predicting heart disease (see Appendix A) as well as the output value of the presence of heart disease (1: heart disease, 0: normal). The dataset combines five diverse sources (Fedesoriano, 2020), offering a comprehensive mix of demographic, clinical, and symptom data. Its diversity enhances model performance and reduces potential biases compared to smaller, single-source datasets, making it highly suitable for machine learning approaches focusing on heart disease prediction.

Methods

Data Preprocessing

In this project, preprocessing involved handling missing values, feature scaling, and encoding categorical variables to prepare the dataset for predictive modeling. In any real-world dataset, missing values are a common occurrence, often due to incomplete data collection (Joel et al., 2024). However, machine learning algorithms cannot work directly with null (NaN) values, as these can cause errors or reduce the predictive quality of the model. To address this, we performed an initial examination of the dataset for missing values. Our dataset, sourced from Kaggle and containing 918 records, did not exhibit any missing values in its attributes so we proceeded with feature scaling. If features have widely varying scales, those with larger magnitudes may unduly influence the model, introducing bias (Analytics Vidhya, 2020). To address this, we applied feature scaling to continuous variables in our dataset, particularly those representing clinical measurements such as cholesterol levels and blood pressure. This step ensures that these clinical features contribute comparably within the model and minimizing potential bias. Our scaling method of choice was robust scaling, since it centers values around the median and scales them within the interquartile range when working with significant outliers, therefore mitigating the effect of extreme values. Next, we took care of handling categorical variables. The dataset includes categorical variables, such as *Sex, ChestPainType, RestingECG, ExerciseAngina*, and *ST_Slope.* These variables represent qualitative characteristics that machine learning models cannot interpret in text form. Therefore, converting these variables to a numeric format is essential. We utilized one-hot encoding to transform categorical features into binary columns, ensuring each category is represented as a separate, binary-valued feature. This approach is well-suited to non-tree-based algorithms, as it prevents any imposed ordinal relationship between categories. For tree-based algorithms (e.g., Random Forests or Decision Trees), label encoding was our other option. Label encoding assigns integer values to categories but introduces an implicit ordinal relationship, which can lead to unintended bias when used with algorithms that treat feature values as ranks (Brownlee, 2020).

Visualization

We begin interpreting our dataset by using distributional visualizations, such as histograms and bar charts, to illustrate the frequency and spread of key variables. Histograms show the distributions of continuous characteristics—age, cholesterol levels, resting blood pressure, and maximum heart rate—categorized by heart disease status (1 or 0). These visualizations help reveal patterns and differences between individuals with or without heart disease.

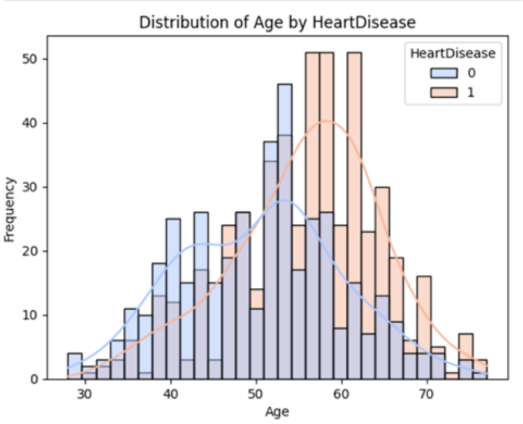


Figure 1: This histogram displays the distribution of age by heart disease.

Scatter plots allowed us to examine relationships between key variables, such as MaxHR versus Oldpeak (ST depression), categorized by heart disease status. This visualization facilitated the illustration of the relationship between higher heart rates and stress-related abnormalities. Similarly, a scatter plot of age versus cholesterol, also distributed by heart disease status, provided additional insights into potential age-related trends in cholesterol levels.

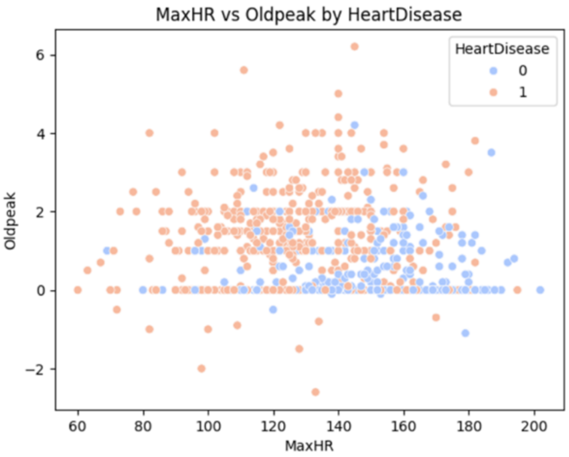


Figure 2: This scatter plot displays the relationship between max heartrate and ST depression.

Correlation heatmaps were used to highlight the strength of correlations between variables. Strong correlations helped identify key predictors of heart disease, while highlighting multicollinearity issues that could affect model reliability. This visual aid also contributed to a deeper understanding to select relevant variables for analysis.

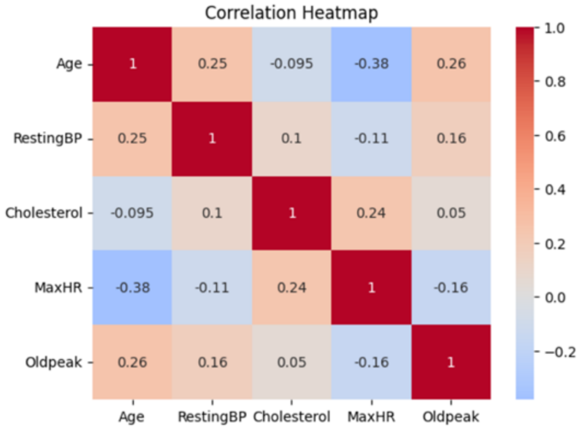


Figure 3: This correlation heatmap displays the correlation between variables.

Model Selection

We employed both baseline and advanced models to analyze heart disease prediction, focusing on their predictive power and interpretability. The baseline model chosen was Logistic Regression, which provides a simple and interpretable approach for binary classification tasks. For advanced models, we used Random Forest, XGBoost, SVM and Gradient Boosting, as they excel at capturing non-linear relationships and interactions among variables, especially for tabular medical datasets (Liu et al., 2024).

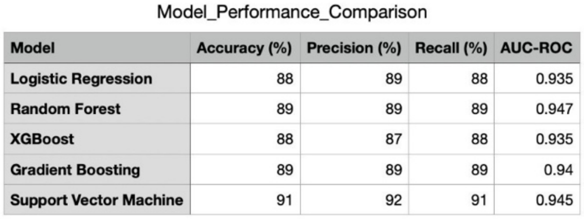
Predictor Variables

The predictor variables were selected for their relevance to heart disease risk and availability in the dataset. Key demographic variables included age and sex to take an equity approach to the prediction of heart disease. Chest pain type was selected due to its relevance in symptomatic factors (Leistner et al., 2014). RestingBP, cholesterol, and fasting blood sugar (FastingBS) variables were also selected to establish cardiovascular risk factors (Roth et al., 2020). MaxHR and Oldpeak were used to reflect exercise-related heart stress and abnormal function. Lastly, ExerciseAngina and ST_Slope provided additional insights into exertion-related symptoms and recovery (Roth et al., 2020).

<u>Results</u>

Model Performance Comparison

Figure 4 compares the results of the five models (Logistic Regression, Random Forest, XGBoost, Gradient Boosting, and SVM) used based on Accuracy, Precision, Recall, and AUC-ROC. A ROC curve comparison was also used to illustrate the strong discriminative power of all models.



Model_Performance_Comparison

Model	Accuracy (%)	Precision (%)	Recall (%)	AUC-ROC
Logistic Regression	88	89	88	0.935
Random Forest	89	89	89	0.947
XGBoost	88	87	88	0.935
Gradient Boosting	89	89	89	0.94
Support Vector Machine	91	92	91	0.945

Figure 4: This table compares the performance of the five models we used. See Appendix B for a more detailed comparison of the performance metrics.

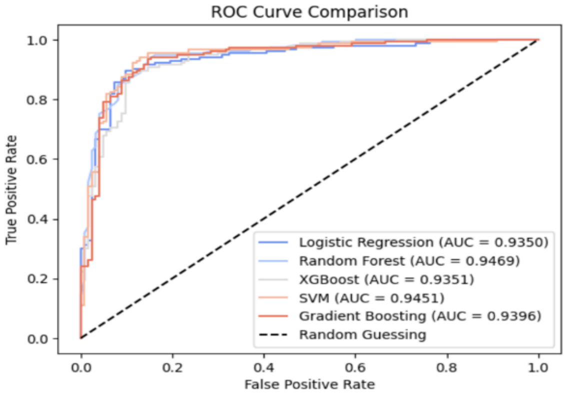


Figure 5: This is a ROC curve comparison of the five models we used

Random Forest demonstrated its strength and reliability by having the greatest AUC-ROC score and the most balanced performance across all measures. SVM also proved to be an effective model, resulting with the highest accuracy, precision and recall, making it ideal for high-risk patient identification. All other models also demonstrated strong discriminative ability, with their performance metrics only slightly trailing behind (Figure 4).

Feature Importance

Feature importance was completed for all models except SVM, as it does not inherently provide feature importance for non-linear kernels such as radial basis functions (RBFs) because it relies on support vectors and kernel functions rather than direct feature weights (Sanz et al., 2018). While methods such as SHapley Additive exPlanations (SHAP) or permutation importance can estimate feature importance, they require additional computation. Given this complexity, feature importance is omitted for SVM. In both the Random Forest and XGBoost models, ST_Slope_Up emerged as the most influential predictor, indicating that the upward slope of the ST segment during stress tests is a critical factor in assessing heart disease risk. For Random Forest, additional important features included ST_Slope_Flat, Oldpeak, and MaxHR, reflecting the relevance of ST segment characteristics, exercise-induced heart stress (MaxHR), and ST depression (Oldpeak) in predicting heart disease. Meanwhile, XGBoost showed a sharper focus on ST_Slope_Up, with a significantly higher importance score for this feature compared to others, followed by ExerciseAngina_Y and ChestPainType_ATA. This difference suggests that XGBoost places more weight on ST segment slope and exercise-induced angina in its decision-making process. Although both models agree on the significance of ST segment features, Random Forest provides a more balanced importance distribution across various features, whereas XGBoost heavily prioritizes a few key predictors. This insight highlights the differing ways these models prioritize risk factors and reinforces the importance of ST segment characteristics in heart disease diagnosis.

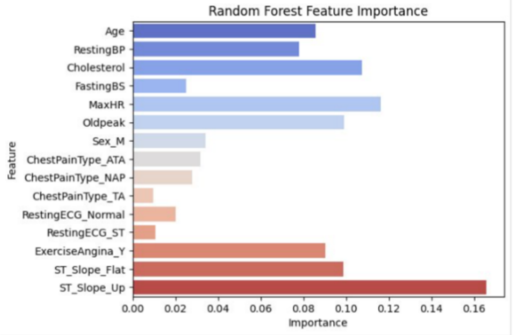


Figure 6: This bar graph shows the feature importance of Random Forest

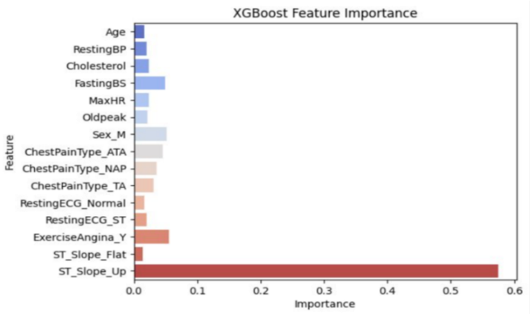


Figure 7: This bar graph shows the feature importance of XGBoost

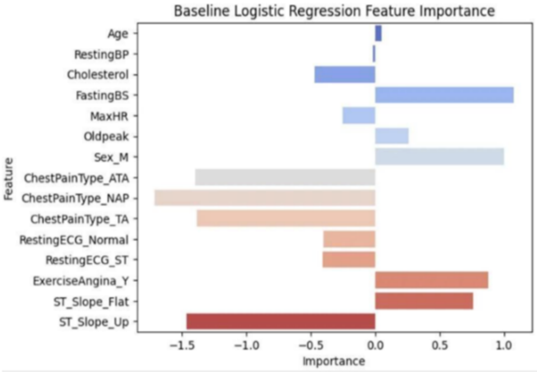


Figure 8: This bar graph shows the feature importance of Logistic Regression

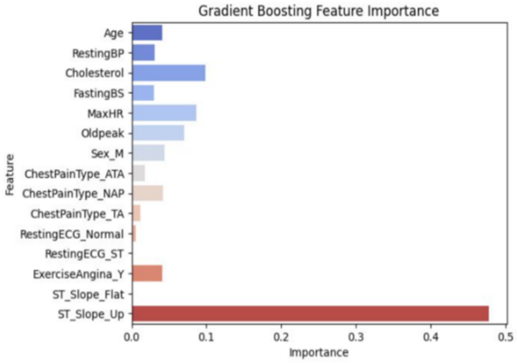


Figure 9: This bar graph shows the feature importance of Gradient Boosting

Optimizing Random Forest

Considering all performance metrics, Random Forest emerged as the best-performing model, achieving 89% accuracy, recall and precision and the highest AUC-ROC score out of all the models (0.947), essential for reliable prediction. Random Forest's balanced feature importance distribution indicates its ability to utilize a broader range of predictors, providing a more comprehensive approach to heart disease assessment. In contrast, XGBoost's focused prioritization of a few key features suggests a more targeted but less generalizable approach. Overall, Random Forest stands out as the most powerful model, combining high accuracy and interpretability, making it well-suited for clinical applications where both precision and a comprehensive assessment of risk factors are essential. Therefore, we prioritize further optimization of the Random Forest model to enhance its performance. The purpose is to optimize Random Forest hyperparameters to enhance predictive performance. After optimization, the AUC-ROC improved to 0.948. Feature importance analysis identified ST_Slope_Up as the most critical predictor, with MaxHR and ST_Slope_Flat also being key factors are essential. The ROC curve illustrates superior discriminative power compared to baseline models. Our optimization process involved tuning hyperparameters such as the number of trees and maximum depth, significantly improving the model's ability to differentiate between heart disease and no heart disease cases. The enhanced feature importance plot reinforces the clinical relevance of the top predictors, aligning with established cardiovascular risk factors.

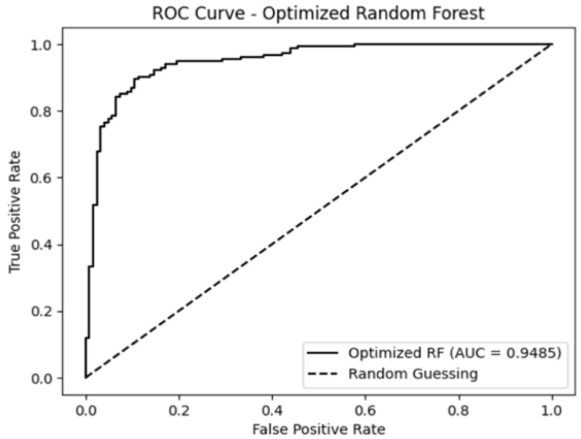


Figure 10: This is the ROC curve of the optimized Random Forest model

Discussion

While our models showed promising results in predicting heart disease risk, several enhancements could further improve their accuracy and reliability. One primary area for improvement involves refining the feature set by incorporating additional attributes relevant to heart disease. Features such as lifestyle factors (e.g., diet, physical activity levels), socioeconomic status, and genetic predispositions could add valuable context that is not captured in the existing dataset. To further enhance model generalizability, testing the model on diverse datasets from multiple sources (e.g., clinical trials, hospital records, or larger population studies) could validate its effectiveness across different demographics and regions. Datasets from these varied sources would also allow the model to adjust for geographic or demographic variables, making it more widely applicable.

Conclusion

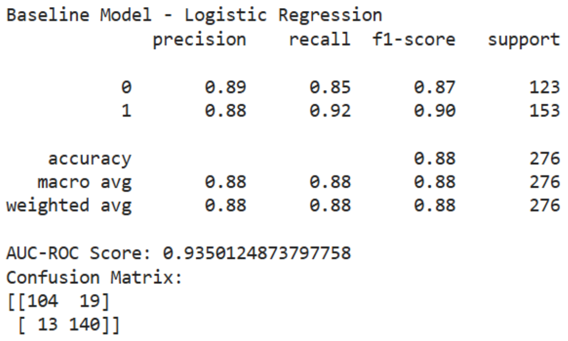
The objective of this study was to develop a machine learning-based predictive model for heart disease. This model was developed using demographic, clinical, and symptomatic data to assess patient risk and enhance early detection. The dataset was subjected to rigorous preprocessing, including the handling of missing values, scaling of features, and encoding of categorical variables, to ensure that it was optimized for analysis. Several machine learning models were implemented and evaluated, including Logistic Regression, Random Forest, XGBoost, SVM, and Gradient Boosting. The Random Forest model exhibited the best-balanced performance, with an optimized AUC-ROC of 0.947, underscoring its reliability and interpretability. Furthermore, SVM exhibited excellent performance, attaining an accuracy and recall of 91% and a precision of 92%, thereby demonstrating an exceptional capacity to identify individuals at considerable risk. The feature importance analysis consistently identified ST_Slope_Up and MaxHR as critical predictors, thereby underscoring the models' relevance in highlighting key risk factors. These findings illustrate the potential of machine learning to assist healthcare professionals in identifying high-risk patients at an early stage, thereby enabling prompt interventions that could enhance outcomes and reduce the burden of late-stage treatment. Future research should concentrate on incorporating additional predictive features and validating the models across diverse populations to enhance their applicability in real-world clinical settings. This work represents a significant step towards leveraging machine learning for preventive healthcare and reducing the global burden of cardiovascular disease.

Appendices

Appendix A: Dataset Composition

Category Type	Variable	Description
Demographic Variables	Age	The age of the patient in years, ranging from young adults to elderly individuals. This variable allows for analysis of how heart disease risk correlates with age.
Demographic Variable	Sex	Gender of the patient, where 1 denotes male and 0 denotes female. This enables the examination of gender-specific risk factors and their impact on heart disease prevalence.
Clinical Measurement	RestingBP	Measured in millimeters of mercury (mm Hg), this indicates the blood pressure of the patient at rest. Elevated resting blood pressure is a well-known risk factor for cardiovascular diseases.
Clinical Measurement	Cholesterol	Serum cholesterol level in milligrams per deciliter (mg/dl). High cholesterol levels can lead to plaque buildup in arteries, increasing the risk of heart failure.
Clinical Measurement	FastingBS	Indicates if the patient's fasting blood sugar is greater than 120 mg/dl (1 = True, 0 = False). Elevated fasting blood sugar levels are associated with diabetes, which is a significant risk factor for heart disease.
Clinical Measurement	MaxHR	The highest heart rate reached during exercise. This metric is useful for assessing cardiovascular fitness and identifying potential anomalies in heart function.
Clinical Measurement	Oldpeak	Represents ST depression induced by exercise relative to rest, measured in depression units. It reflects potential abnormalities in heart activity during stress tests.
Symptomatic and Lifestyle Factor	ChestPainType	Categorized into: TA (Typical Angina), ATA (Atypical Angina), NAP (Non-Anginal Pain), ASY (Asymptomatic). This variable helps in understanding the type and severity of chest pain symptoms experienced by patients.
Symptomatic and Lifestyle Factor	ExerciseAngina	Indicates whether the patient experiences angina during exercise (1 = Yes, 0 = No). This provides insights into exercise-induced symptoms and cardiac stress response.
Electrocardiogram Result	RestingECG	Results of the resting electrocardiogram test, categorized as: Normal, ST (ST-T wave abnormalities), LVH (Left Ventricular Hypertrophy). This attribute aids in detecting electrical abnormalities and structural changes in the heart.
Electrocardiogram Result	ST_Slope	The slope of the peak exercise ST segment, categorized as: Up, Flat, Down. The ST slope provides information on the heart's ability to handle increased workload during exercise.

Appendix B: Performance Metrics of Models

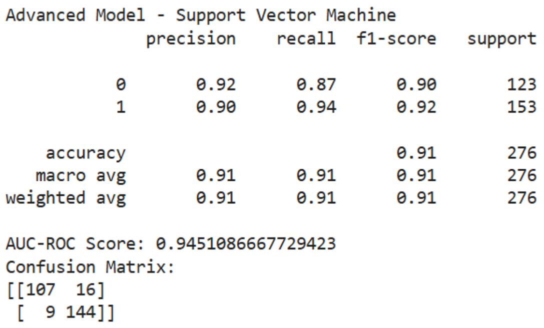


```
Baseline Model - Logistic Regression
              precision    recall  f1-score   support

           0       0.89      0.85      0.87       123
           1       0.88      0.92      0.90       153

    accuracy                           0.88       276
   macro avg       0.88      0.88      0.88       276
weighted avg       0.88      0.88      0.88       276

AUC-ROC Score: 0.9350124873797758
Confusion Matrix:
[[104  19]
 [ 13 140]]
```

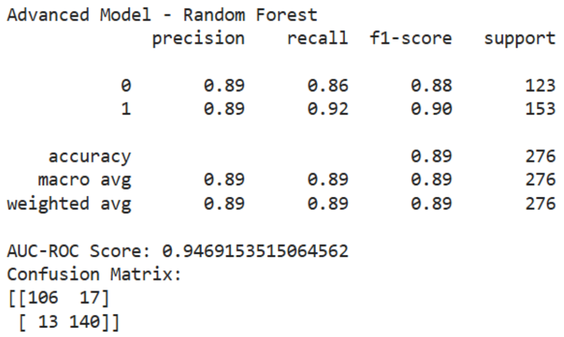


```
Advanced Model - Support Vector Machine
              precision    recall  f1-score   support

           0       0.92      0.87      0.90       123
           1       0.90      0.94      0.92       153

    accuracy                           0.91       276
   macro avg       0.91      0.91      0.91       276
weighted avg       0.91      0.91      0.91       276

AUC-ROC Score: 0.9451086667729423
Confusion Matrix:
[[107  16]
 [  9 144]]
```

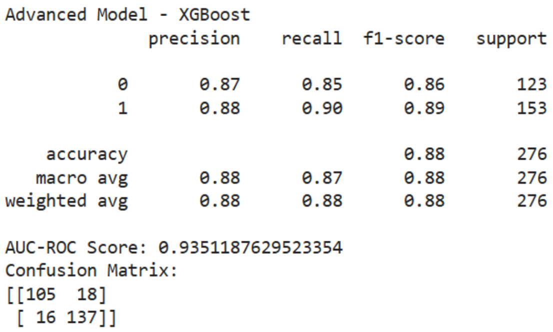


```
Advanced Model - Random Forest
              precision    recall  f1-score   support

           0       0.89      0.86      0.88       123
           1       0.89      0.92      0.90       153

    accuracy                           0.89       276
   macro avg       0.89      0.89      0.89       276
weighted avg       0.89      0.89      0.89       276

AUC-ROC Score: 0.9469153515064562
Confusion Matrix:
[[106  17]
 [ 13 140]]
```

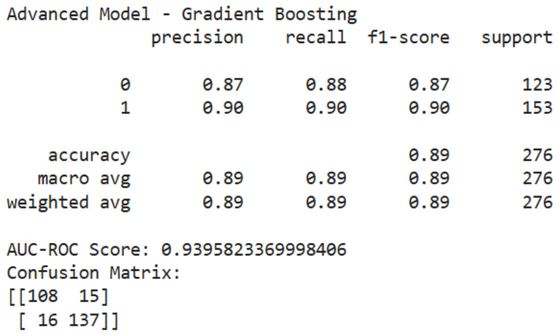


```
Advanced Model - XGBoost
              precision    recall  f1-score   support

           0       0.87      0.85      0.86       123
           1       0.88      0.90      0.89       153

    accuracy                           0.88       276
   macro avg       0.88      0.87      0.88       276
weighted avg       0.88      0.88      0.88       276

AUC-ROC Score: 0.9351187629523354
Confusion Matrix:
[[105  18]
 [ 16 137]]
```



```
Advanced Model - Gradient Boosting
              precision    recall  f1-score   support

           0       0.87      0.88      0.87       123
           1       0.90      0.90      0.90       153

    accuracy                           0.89       276
   macro avg       0.89      0.89      0.89       276
weighted avg       0.89      0.89      0.89       276

AUC-ROC Score: 0.9395823369998406
Confusion Matrix:
[[108  15]
 [ 16 137]]
```

References

Analytics Vidhya. (2020, April 27). Feature scaling in machine learning: Normalization and standardization. Retrieved from https://www.analyticsvidhya.com/blog/2020/04/feature-scaling-machine-learning-normalization-standardization/

Brownlee, J. (2020). How to prepare data for machine learning. *Machine Learning Mastery*. Retrieved from https://machinelearningmastery.com/how-to-prepare-data-for-machine-learning/

Brownlee, J. (2020, September 16). A gentle introduction to one-hot encoding for categorical data. *Machine Learning Mastery*. Retrieved from https://machinelearningmastery.com/one-hot-encoding-for-categorical-data/

Fedesoriano. (2020). Heart failure prediction dataset. *Kaggle*. Retrieved from https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction/data

Habehh, H., & Gohel, S. (2021). Machine learning in healthcare. *Current Genomics, 22*(4), 291–300. https://doi.org/10.2174/1389202922666210705124359

Joel, L. O., Doorsamy, W., & Sena Paul, B. (2024). On the performance of imputation techniques for missing values on healthcare datasets. *arXiv*. Retrieved from https://arxiv.org/abs/2403.14687

Leistner, D. M., Klotsche, J., Palm, S., Pieper, L., Stalla, G. K., Lehnert, H., Silber, S., März, W., Wittchen, H.-U., & Zeiher, A. M. (for the DETECT Study Group). (2014). Prognostic value of reported chest pain for cardiovascular risk stratification in primary care. *European Journal of Preventive Cardiology, 21*(6), 727–738. https://doi.org/10.1177/2047487312452503

Liu, T., Krentz, A., Lu, L., & Curcin, V. (2024). Machine learning-based prediction models for cardiovascular disease risk using electronic health records data: Systematic review and meta-analysis. *European Heart Journal - Digital Health*. Advance online publication. https://doi.org/10.1093/ehjdh/ztae080

Roth, G. A., Mensah, G. A., Johnson, C. O., Addolorato, G., Ammirati, E., Baddour, L. M., Barengo, N. C., Beaton, A. Z., Benjamin, E. J., Benziger, C. P., Bonny, A., Brauer, M., Brodmann, M., Cahill, T. J., Carapetis, J., Catapano, A. L., Chugh, S. S., Cooper, L. T., Coresh, J., … Fuster, V. (2020). Global burden of cardiovascular diseases and risk factors, 1990–2019: Update from the GBD 2019 study. *Journal of the American College of Cardiology, 76*(25), 2982–3021. https://doi.org/10.1016/j.jacc.2020.11.010

Sanz, H., Valim, C., Vegas, E., Oller, J. M., & Reverter, F. (2018). SVM-RFE: Selection and visualization of the most relevant features through non-linear kernels. *BMC Bioinformatics, 19*(432). https://doi.org/10.1186/s12859-018-2451-4

World Health Organization. (2021). Cardiovascular diseases (CVDs). Retrieved from https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)