COMP 562 Final Project

Jacob Dang	Shreya Gundam	Nathan Jacobs	Madhumitha Gopinath
730558339	730483758	730400716	730475197

GitHub Repo: https://github.com/madhu260/Comp562-Final-Project.git

1 Introduction

Cardiovascular diseases are the leading cause of death globally, accounting for approximately 17.9 million deaths each year, which is 32% of all global deaths according to the World Health Organization (WHO) [4]. These conditions are influenced by a combination of risk factors, including age, lifestyle, genetic predisposition, and clinical parameters. Given that a majority of these deaths are due to heart attacks and strokes which show early signs through intermediate risk factors, it underscores the importance of early detection and targeted intervention. The primary objective of this study is to leverage machine learning techniques to predict heart disease based on clinical parameters. By analyzing features from a comprehensive dataset of heart disease indicators, this project aims to determine the model that most accurately predicts heart disease and identify which parameters impact model prediction the most.

The intermediate risk factors in our dataset include Resting Blood Pressure (RestingBP), Fasting Blood Sugar (FastingBS), and Maximum Heart Rate (MaxHR), among other relevant attributes combined from various datasets. The accuracy and importance of each factor is evidenced from their recording in a primary care facility. We will preprocess the included features to understand their roles in predicting heart disease and explore feature importance and interpretability using machine learning algorithms, ultimately identifying the best one based on accuracy, sensitivity, and specificity. The results of this project will offer valuable perspectives for improving early diagnosis and intervention strategies in cardiovascular care.

2 Related Works

The prevalence of heart disease as a leading cause of mortality worldwide makes it a common area of interest for improving prediction of its diagnosis. There are many approaches when addressing this problem, starting off by analyzing clinical and demographic data then leveraging various machine learning techniques. Previously, researchers such as El-Sofany et al. (2024) and Srinivasan et al. (2023) have evaluated various algorithms such as LR, RF, and KNN to enhance classification accuracy for heart disease prediction. El-Sofany et al. compared ten machine learning classifiers, including Naive Bayes, SVM, voting, XGBoost, AdaBoost, bagging, KNN, DT, RF, and LR, achieving an accuracy of 97% using ensemble techniques like AdaBoost, DT, and RF on a private dataset from an Egyptian hospital. [1] Meanwhile, Srinivasan et al. employed an HRFLM method which used an artificial neural network (ANN) with 13 clinical features and achieved 88.7% effectiveness rate. [3]

These studies used various ways of analyzing the effect of the features on the model and predictions, such as analysis of variance (ANOVA), chi-square, and mutual information methods in order to aid in feature selection for more accurate models. Our work builds on these contributions by focusing not only on model accuracy but also on understanding feature importance. By incorporating feature importance, we are able to provide a comprehensive list of which clinical and demographic factors contribute most significantly to heart disease predictions.

3 Approach

3.1 Data and Feature Encoding

We sourced our dataset, named "Heart Failure Prediction Dataset," from kaggle and it contains 12 variables and 918 observations. [2] The features are Age, Sex, ChestPainType, RestingBP, Cholesterol, FastingBS, RestingECG, MaxHR, ExerciseAngina, Oldpeak, ST_Slope, whereas the response variable is HeartDisease. Sex, ChestPainType, FastingBS, RestingECG, ExerciseAngina, ST_Slope, and HeartDisease are categorical variables, with Sex, FastingBS, RestingECG and ExerciseAngina being binary. Our first step when cleaning up the data was to convert these categorical variables into numerical values so that the machine learning models can interpret them without assuming any ordinal relationship between categories. Features that had more than two unique categories, such as ChestPainType, RestingECG, and ST_Slope, were encoded using one-hot encoding, while an ordinal encoder was used for binary categorical features such as Sex and ExerciseAngina.

3.2 Description of models

Because the response variable is categorical, we needed to use classifier models to predict the variable. We decided to focus on logistic regression (LR) and random forests (RF), since both have their own merits. LR is simplistic and easy to interpret, making it a reliable choice to understand the relationships between the features and the target variable. We also used a RF model, as they are ideal for capturing nonlinear relationships and interactions between features, particularly in tabular data, making them well-suited for our dataset. We found that the numerical features in the dataset have different scales, so we passed them through a standard scalar, making means zero and variances one, before creating an LR with the data. This helped the solvers converge on a solution. This scaling was omitted for RF, since it can inherently handle this issue well.

3.3 Hyperparameter tuning

To create both types of models, we used scikit-learn's LogisticRegression and RandomForestClassifer classes. We performed hyperparameter tuning for each model, using 5-fold cross-validation to evaluate different hyperparameter values and selecting the best parameter based on the mean. For RF, we averaged cross validation results over 15 different generation seeds to reduce variance and obtain usable results, as random forests are non-deterministically generated. A 95% confidence interval was also created around each mean to find the difference between the parameter values.

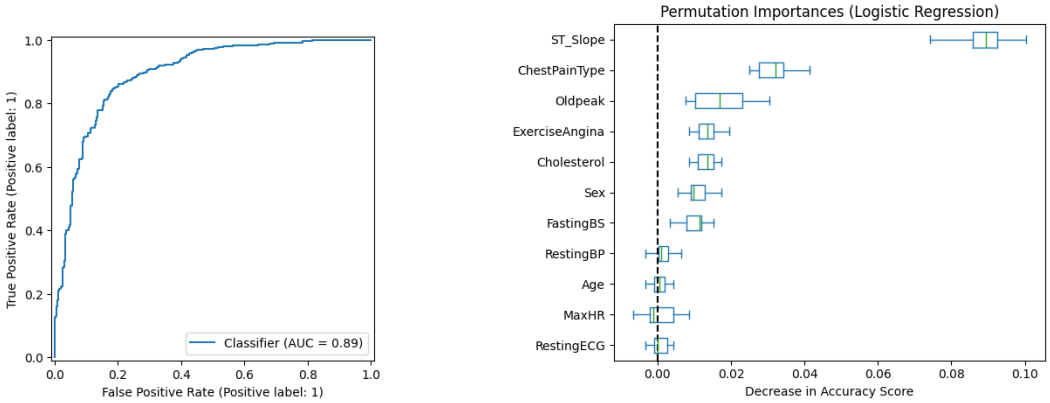
For the LR model, we tuned the solver, penalty term, and regularization strength (C). All scikit-learn solvers converged to the same solution, so the default value (lbfgs) was used. While no statistically significant differences were found between other hyperparameter values, adding an L2 penalty and setting $C = 0.1$ yielded the best cross-validation accuracy. For the RF model, we tuned n_estimators, max_features, min_samples_split, and min_samples_leaf.[1] Increasing n_estimators improved accuracy significantly until 100, beyond which the rate of improvement diminished. Since compute time is long for models with multiple estimators, we chose 100 to balance performance and time. For max_features, values of 1-2 outperformed higher values (>13), so we selected 2. min_samples_split and min_samples_leaf were set to 2 and 10, respectively, as no significant differences were observed for them.

3.4 Evaluation and obtaining feature importances

Final model performance was evaluated by using the average accuracies over a 5-fold cross-validation, confusion matrices, and receiver operating characteristic (ROC) curves. To evaluate how important a feature was to each model, we decided to use scikit-learn's implementation of permutation importances. Permutation importances involve randomly shuffling features one at a time and then comparing the accuracy of a baseline performance to the accuracy after the feature has been shuffled. Each model was pipelined so that the features would be shuffled before encoding was applied, which preserves the original features groups. This is important because permutation importance does not

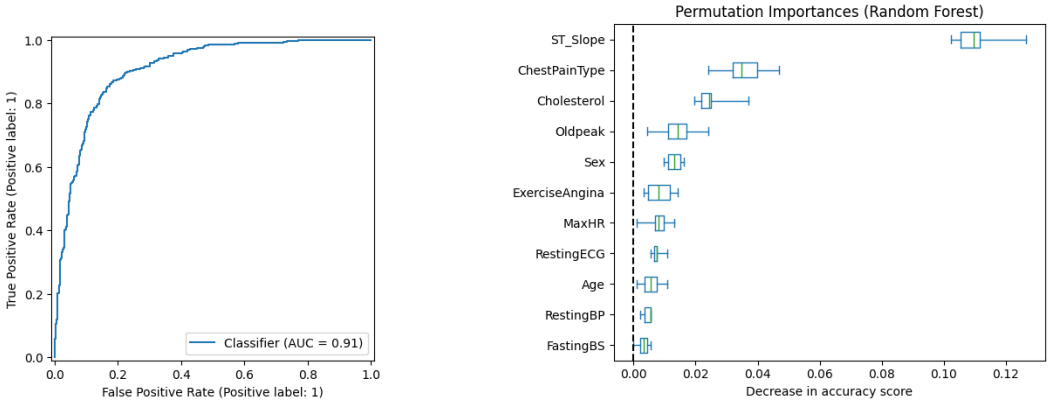
[1] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html contains descriptions of each parameter

give reliable scores if the features are correlated, and it is likely this would be the case for one hot encoded features. [2]



(a) ROC Curve (b) Permutation Importance

Figure 1: Logistic Regression: ROC Curve and Permutation Importance



(a) ROC Curve (b) Permutation Importance

Figure 2: Random Forests: ROC Curve and Permutation Importance

4 Results

4.1 Performance

The LR model performs well in predicting heart disease, achieving a 5-fold cross validation mean accuracy of 83.1%. The confusion matrix illustrates the model's ability to predict both classes effectively, though it suggests improvement in handling ambiguous cases. It reveals that the model correctly identifies 332 cases of "No Heart Disease" and 431 cases of "Heart Disease," while misclassifying 77 false positives and 78 false negatives. This gives the model a false positive rate of 19%, and a false negative rate of 15.2%. The confusion matrix shows the model's performance on a threshold of p=0.5 for deciding whether a set of features should be categorized as having heart disease or not. An ROC curve can be applied to see how the model performs in terms of true positive rate and false positive rate under different thresholds, which can also help in understanding how good the model is at distinguishing between classes. In particular, the area under the ROC curve (AUC) for the logistic regression model is 0.89, which indicates the model performs well (Figure 1a).

[2]https://scikit-learn.org/1.5/modules/permutation_importance.html

The final RF model is slightly more accurate overall than the logistic regression, with overall 5-fold cross validation mean accuracy of 84%. The confusion matrix shows that the model identified 326 negative cases and 446 positive cases of heart disease correctly, while misclassifying 84 negative cases and 62 positive cases. We can deduce then that the model has a 12.2% false negative rate and a 20.5% false positive rate. This would indicate that the model is slightly better at identifying cases of people without heart disease, but is slightly worse at identifying people with heart disease. The AUC for the RF model is 0.91, which is slightly better performing than the logistic model (Figure 2a).

Overall, the results demonstrate that the models are effective and reliable for heart disease prediction, though additional methods, such as fine-tuning or employing ensemble methods, could further improve performance.

4.2 Permutation Importance

The permutation importance plots (Figures 1b and 2b) visualize the significance of the individual features in the model's predictions. ST Slope is the most influential feature, as permutations of the feature degrade the predictive ability of both the LR and RF models more than Chest Pain Type, Cholesterol, and Oldpeak, which also appear to be influential in the models. Features like Sex and Exercise Angina, have moderate influence, while Max HR, Fasting BS, Resting ECG, Resting BP, and Age have lower significance relative to the other features. The analysis underscores the importance of specific features, namely ST changes and chest pain characteristics, in accurately predicting heart disease.

5 Conclusion

When comparing both models, RF has slightly better overall accuracy and a better AUC value, suggesting it might be a better choice for our dataset. However, if we consider minimizing false positives to be more important—which would avoid over-diagnosis—LR model would be preferable due to its higher precision for the "Heart Disease" class and fewer false positives. This would also help in reducing costs resulting from unnecessary medical tests.

This study provided many insights on the potential prevention of cardiovascular disease, and by focusing on analyzing features, we were able to find the warning signs and conditions that the machine learning models found the most predictive of heart disease. We also determined the most effective machine learning models, providing a foundation of tools that have the potential to be used for real-world analysis in the medical field. The models' ease of interpretability helped show a clear understanding of the relationships between features and their impact on predictions, making the insights readily usable for medical professionals

These insights can also be used on a broader scale. By determining the most significant features in the problem of heart disease, certain risk factors and early warning signs can be prioritized. This means that people can be more accurately assessed on their risk for heart disease based on data, which could save lives on a much larger scale. Many potential solutions can be drawn, for example, applications that can better determine the risk of a large number of people based on data.

References

[1] El-Sofany, H., Bouallegue, B. & El-Latif, Y.M.A. (2024). A proposed technique for predicting heart disease using machine learning algorithms and an explainable AI method. *Sci Rep* 14, 23277. https://doi.org/10.1038/s41598-024-74656-2

[2] fedesoriano. (2021) Heart Failure Prediction Dataset. Retrieved November 19, 2024 from https://www.kaggle.com/fedesoriano/heart-failure-prediction.

[3] Srinivasan, S., Gunasekaran, S., Mathivanan, S.K. et al. (2023) An active learning machine technique based prediction of cardiovascular heart disease from UCI-repository database. *Sci Rep* 13, 13588. https://doi.org/10.1038/s41598-023-40717-1

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