Cardiovascular Diseases Prediction

Jaya Jwalitha Nagulla
Computer Science Department
New Mexico StateUniversity
jwalitha@nmsu.edu

Humesh Reddy Venkatapuram
Computer Science Department
New Mexico StateUniversity
humeshrv@nmsu.edu

Mohith Krishna Reddy Dontireddy
Computer Science Department
New Mexico State University
dmohith@nmsu.edu

Motivation:

Heart disease is a major global health concern, claiming millions of lives each year. Detecting it early can greatly improve outcomes, but its complexity makes accurate prediction a challenge. With factors like genetics, environment, and lifestyle involved, there's a critical need for better diagnostic tools.

Problem Definition:

We aim to develop a predictive model that utilizes datasets containing various medical attributes such as age, gender, cholesterol levels, blood pressure, and heart performance indicators to most accurately predict the likelihood of heart disease. By integrating and analyzing these critical factors through advanced data processing and machine learning techniques, our goal is to identify individuals at high risk of heart conditions more efficiently. This proactive approach will help healthcare providers to potentially improve better health outcomes.

Related Works:

Rahma Atallah and Amjed Al-Mousa proposed a method for heart disease detection utilizing ensemble learning. It combines multiple machine learning algorithms such as logistic regression, decision trees, etc., and employs a majority voting scheme to determine the final prediction. By aggregating the predictions of individual classifiers, the method aims to enhance accuracy and reliability in identifying patients with heart disease.

Goutam Kumar Sahoo, Keerthana Kanike, Santos Kumar Das, and Poonam Singh proposed a framework that aims to provide timely and efficient detection of heart disease, enhancing accuracy through an integrated approach. The models employed include Logistic Regression, K-Nearest Neighbor, Support Vector Machine, Naive Bayes, Decision Tree, Random Forest, and XG Boost.

Possible Machine Learning Tasks:

Data Visualization: Data is visualized by looking at all the data together to see patterns.

Data Preprocessing: Data is cleaned up by fixing errors and filling in missing information to make sure data is ready for analysis.

Data Splitting: Data is splitted into two groups, one for training the models and one for testing them. This helps ensure the models can work well on new information.

Data Normalization: Data is adjusted so that all types of information are treated equally, preventing any one type from dominating.

Model Training: Various models are used to try and predict heart disease. Each model works differently, and performance analysis is done to know the model that gives the best performance.

Evaluating Models: After training, measuring is done to see how well each model did by looking at accuracy and precision.

Comparing Models: Lastly, all the models are compared to pick the best one for predicting heart disease effectively and efficiently.

Dataset Collection and Overview:

Initially 3 datasets related to Heart-Disease are selected. Heart Disease Dataset, Heart failure Prediction Dataset, Cleveland Dataset. The UCI Heart Disease dataset, while smaller than the initially required 10,000 instances, was selected due to its widespread recognition and frequent usage in medical research, ensuring comparability with numerous studies. Its comprehensive and clinically relevant features offer high data quality, crucial for accurate predictive modeling in healthcare. The dataset's accessibility and ethical clearance make it a practical choice in a field where large, open medical datasets are rare. This project emphasizes developing methodologies applicable to real-world settings, where large datasets might not be available. Future expansions might include synthetic data generation or collaborations to increase the dataset size, aligning with the project's

aim to enhance real-world applicability and data scalability.

Processed Cleveland Dataset: This (Cleveland.data) dataset is the processed dataset of the original Cleveland dataset which is the only one that has been used by ML Researchers till date. The original dataset has 76 attributes, but all published requirements refer to using a subset of 14 of them. The dataset originally had more attributes, but 62 of these were omitted to enhance the model accuracy because they showed little or no correlation with the target outcome. The processed Cleveland data contains 303 instances and 14 features.

Heart Disease Dataset: This dataset contains 1025 records with 14 features and the target variable indicating the presence of heart disease. This dataset was created by combining different datasets already available independently but not combined before. In this dataset, 5 heart datasets are combined over 14 common features which makes it the largest heart disease dataset available so far for research purposes.

Heart Failure Prediction Dataset: There are 918 records in the dataset. They have mostly similar features as heart disease dataset, but some have slight differences in names and additional categories, like types of chest pain and results from resting electrocardiograms results.

Why the Datasets are Reasonable for Analysis

These datasets cover many different things that can affect someone's chances of getting heart disease. This includes personal details like their age and gender, measurements of their body like weight and blood pressure, and results from medical tests they have taken. For all the three datasets, there is a good balance between the different possible outcomes for the target variable. The wide range of cholesterol and resting BP values means the data covers diverse people, making the analysis more broadly applicable. The data has the information that doctors normally collect from patients. So, any findings from analyzing this data can be directly used to help treat real patients.

Common Themes: All datasets focus on heart disease, utilizing both personal and medical examination data to determine the presence of heart disease with importance of 14 features.

Differences: Dataset Heart failure prediction contains more features and provides a more in-depth look at heart conditions, including detailed features like the thalassemia level and the number of major vessels visible in fluoroscopy.

Solution:

Building and Training Models:

Experimenting with models like logistic regression, decision trees, Random Forest, AdaBoost, Support Vector Machines (SVM) Neural networks, K Nearest Neighbors (KNN), Naïve bayes, Extreme Gradient Boosting to know which model gives better performance. Both Regression methods and classification models are to be experimented with. Cross Validation is used to validate the SVM and Neural Networks model. Evaluate each model using metrics such as accuracy, precision, recall, F1-score, and AUC-ROC. These metrics help determine not only how often the model is correct (accuracy) but also how reliable those predictions are (precision and recall).

Further the model is assessed for performance by calculating the Mean Squared Error (MSE) and R^2 scores, particularly for models where the gradation of severity (like stages of heart disease) might be predicted. Histograms, Box Plots, Line graphs are used to compare the performance of the classification algorithms. For each dataset the above-mentioned models are used.

Combined Dataset: Cleveland.data and heart disease dataset are merged which is the combined dataset. This combined dataset contains 1,328 entries with 14 features related to heart health, such as age, gender, cholesterol levels, and heart performance indicators.

The use of well-known datasets like the Cleveland dataset allows for benchmarking results against a wide body of existing research. It is a very well-known dataset in this domain.

Data Quality and Preprocessing Needs:

Missing Data: Certain columns like 'ca' and 'thal' contain placeholders (e.g., '?') that need to be handled.

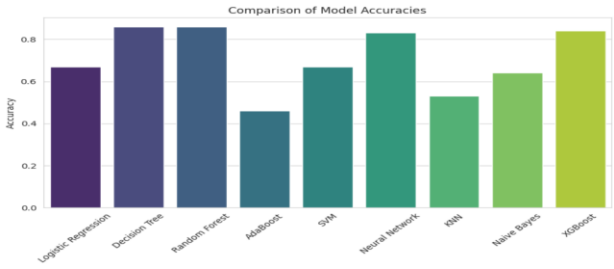
Outliers: Features like blood pressure and cholesterol have outliers that might need addressing to prevent skewed analyses.

Categorical Data: Features sex, chestpain, Fasting Blood Sugar (fbs), Resting Electrocardiographic Results (restecg), Exercise-Induced Angina

(exang), Slope of the Peak Exercise ST Segment (slope), Thalassemia (thal): categorical features are categorical and require numeric encoding for use in machine learning models.

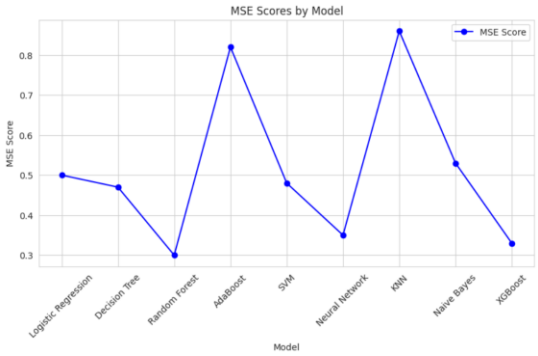
Machine Learning Model Results:

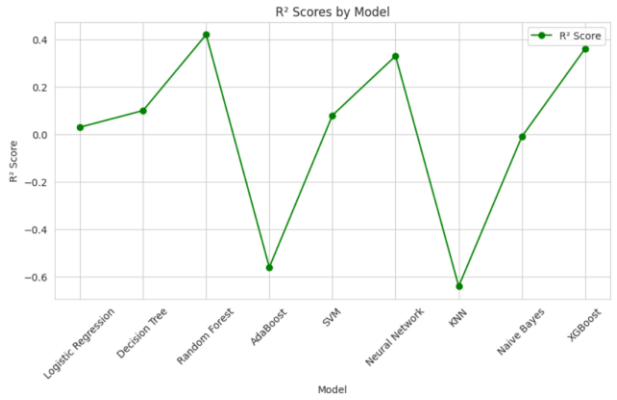
Accuracy: Based on the obtained results we noted that Decision Tree (86%) and Random forest (86%), Extreme Gradient Boosting(84.1%), Neural Networks(81.1%) displayed robust performance gave best model accuracy when compared to the remaining models. Whereas Logistic Regression showed an accuracy of about 67.5% with moderate precision and recall across the classes. AdaBoost Classifier underperformed with an accuracy of 46%, showing difficulty in class balancing or potentially overfitting on specific classes. Support Vector Machines had an accuracy comparable to logistic regression at about 67.5%, with balanced precision and recall across most classes. K-Nearest Neighbors and Naive Bayes had lower accuracies (53.6% and 64.5% respectively), struggling with the dataset's complexity or distribution.



The Mean Squared Error (MSE) and R^2

Random Forest (0.30) and XGBoost showing the best balance between error metrics and determination coefficient, indicating a good fit to the variance in the data. The Random Forest classifier also leads with the highest R^2 score of 0.42, implying it can account for 42% of the variance in the target variable, which is the highest among the models.





Training and Testing Times:

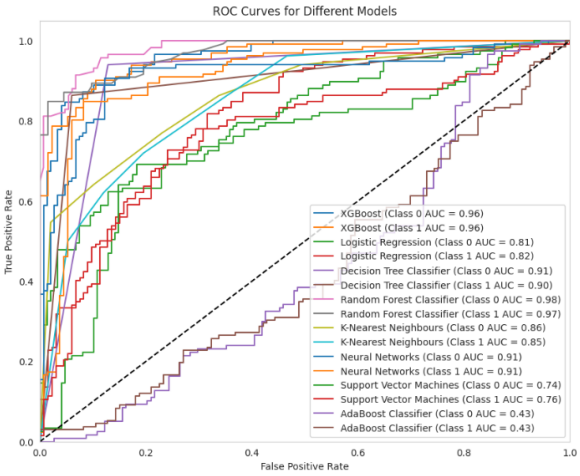
Training times varied significantly, with Neural Networks taking the longest time, indicating the computational cost of training deep models. In contrast, models like Decision Tree(0.0067sec) and KNN were much faster.

Testing times were generally low for all models, but The Logistic Regression and Decision Tree classifiers are the fastest at testing, both taking only 0.0003 seconds.

ROC Curve:

The ROC curves in the image below compares the performance of various classifiers for a multiclass classification problem using a one-vs-rest approach. Each classifier's ability to distinguish between classes is measured by the AUC (Area Under the Curve), where higher values indicate better performance. The top-performing classifiers are XGBoost and Random Forest, both achieving high AUC values (0.96-0.98), demonstrating

excellent discriminatory power. Logistic Regression, Decision Tree, K-Nearest Neighbours, and Neural Networks also perform well, with AUC values around 0.81-0.91. Support Vector Machines show moderate performance (AUC 0.74-0.76), while AdaBoost performs poorly (AUC 0.43). These results highlight that XGBoost and Random Forest are the most effective for this dataset, while AdaBoost may require further optimization.

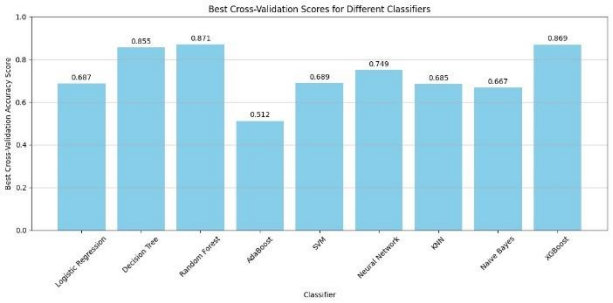


Stacking Classifier

The ROC curves compare the performance of Random Forest, XGBoost, and Stacking Classifier models on a multiclass classification problem. All models show excellent performance, with the Stacking Classifier slightly outperforming the others, especially for class 2. The Random Forest and XGBoost models also achieve high AUC values (0.89-0.98). The classification report for the Stacking Classifier indicates high precision and recall for classes 0 and 1, but lower performance for classes 2, 3, and 4, likely due to limited data for these classes. Overall, the Stacking Classifier achieves an accuracy of 85.7%, a Mean Squared Error of 0.283, and an R^2 score of 0.462, demonstrating its strong predictive capability and moderate explanatory power.

```
Model: Stacking Classifier
Confusion Matrix:
[[107  10   0   0   0]
 [ 14 117   0   1   0]
 [  3   1   1   2   0]
 [  0   3   2   2   0]
 [  0   2   0   0   0]]
Classification Report:
              precision    recall  f1-score   support

           0       0.86      0.91      0.89       117
           1       0.88      0.89      0.88       132
           2       0.33      0.14      0.20         7
           3       0.40      0.29      0.33         7
           4       0.00      0.00      0.00         2

    accuracy                           0.86       265
   macro avg       0.50      0.45      0.46       265
weighted avg       0.84      0.86      0.85       265

Accuracy: 0.8566037735849057
Mean Squared Error (MSE): 0.2830188679245283
R² Score: 0.4621109607577808
```

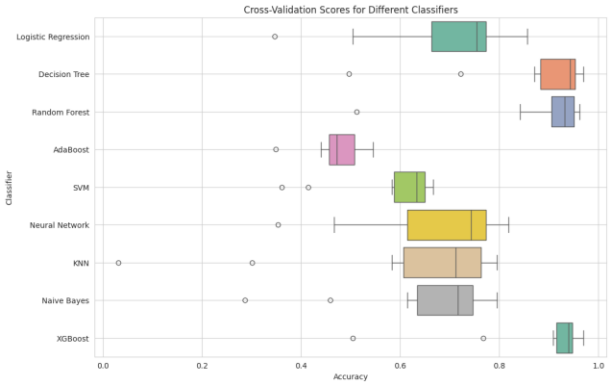
Hyperparameter Tuning with GridSearchCV

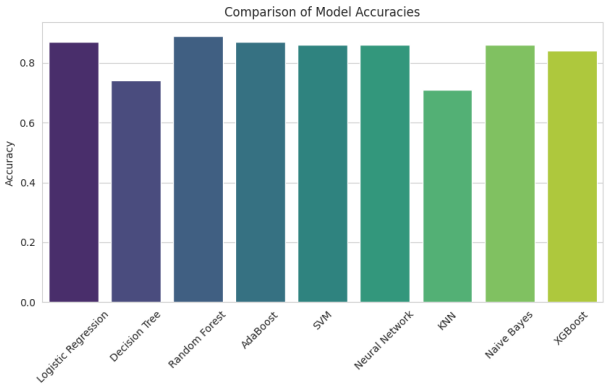


The bar plot shows the best cross-validation accuracy scores for various classifiers after hyperparameter tuning, highlighting that Random Forest and XGBoost are the top performers with scores of 0.871 and 0.869, respectively. Decision Tree also performs well with an accuracy of 0.855. The Neural Network achieves a respectable score of 0.749, while SVM, Logistic Regression, and K-Nearest Neighbours (KNN) exhibit moderate performance with scores around 0.685-0.689. Naive Bayes scores slightly lower at 0.667, and AdaBoost has the lowest performance with a score of 0.512, indicating it is the least effective model among those evaluated.

K-FoldCross Validation

The box plot, coupled with the cross-validation accuracy values, highlights that Random Forest (0.885) and XGBoost (0.881) are the top-performing classifiers, demonstrating both high accuracy and low variability. Decision Tree follows closely with an accuracy of 0.872, showing it is also a strong performer. Logistic Regression achieves a moderate accuracy of 0.690, while Neural Network and Naive Bayes have slightly lower accuracies at 0.675 and 0.656, respectively. KNN and SVM show lower performance with accuracies of 0.613 and 0.585, respectively. AdaBoost is the least effective model with an accuracy of 0.473, indicating significant room for improvement or less suitability for this specific classification problem.





Heart Failure Prediction Dataset:

Missing Data: Certain columns like 'ca' and 'thal' contain placeholders (e.g., '?') that need to be handled.

Outliers: Features like blood pressure and cholesterol have outliers that might need addressing to prevent skewed analyses.

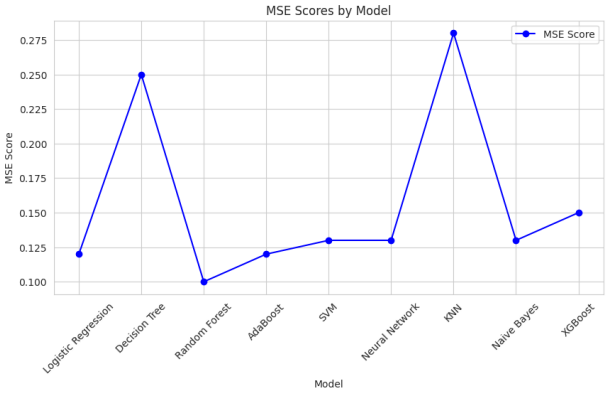
Categorical Data: Features sex, chestpain, Fasting Blood Sugar (fbs), Resting Electrocardiographic Results (restecg), Exercise-Induced Angina (exang), Slope of the Peak Exercise ST Segment (slope), Thalassemia (thal): categorical features are categorical and require numeric encoding for use in machine learning models.

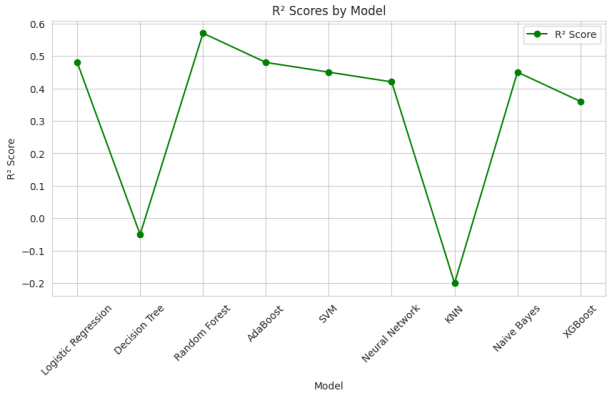
Machine Learning Model Results:

Accuracy: The bar plot shows a comparison of model accuracies, where each bar represents a classifier's performance. The accuracy values indicate that Random Forest (0.885) and XGBoost (0.881) are the highest-performing models, closely followed by Decision Tree with an accuracy of 0.872. Logistic Regression also performs well with an accuracy of 0.690. Neural Network and Naive Bayes achieve accuracies of 0.675 and 0.656, respectively, showing moderate performance. KNN and SVM have lower accuracies of 0.613 and 0.585, respectively, indicating less effectiveness. AdaBoost has the lowest accuracy at 0.473, suggesting it is the least suitable model for this classification problem. The plot clearly highlights the superior performance of ensemble methods like Random Forest and XGBoost compared to other classifiers.

The Mean Squared Error (MSE) and R^2

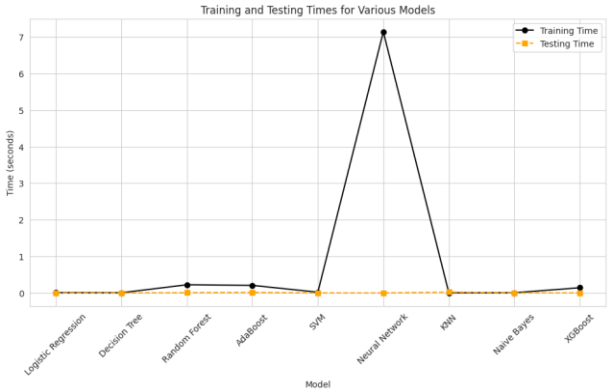
The two line plots display the Mean Squared Error (MSE) and R^2 scores for different classifiers, providing insight into their regression performance. In the MSE plot, lower values indicate better performance, with Random Forest achieving the lowest MSE, indicating the best predictive accuracy. Conversely, Decision Tree and KNN have the highest MSE values, showing poorer performance. The R^2 score plot shows how well each model explains the variance in the data, with higher values indicating better performance. Random Forest has the highest R^2 score, followed by Logistic Regression and XGBoost, demonstrating strong explanatory power. Decision Tree and KNN have the lowest R^2 scores, indicating poor fit and high error variance. Overall, Random Forest and XGBoost are the top performers in both metrics, while Decision Tree and KNN perform the worst.





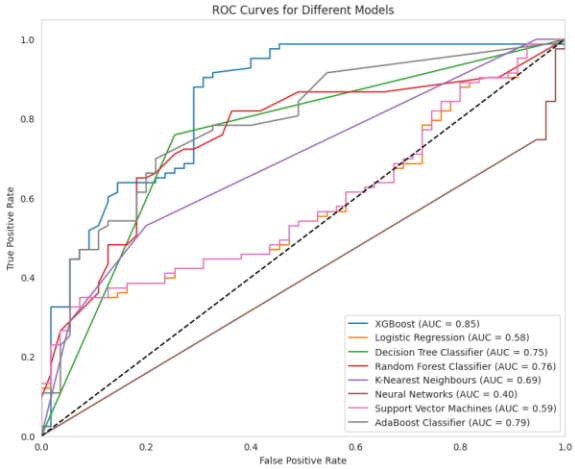
Training and Testing Times:

The line plot shows the training and testing times for various models, highlighting their computational efficiency. Training times are represented by the black solid line, while testing times are shown by the orange dashed line. The Neural Network has the highest training time, significantly longer than all other models, indicating it requires more computational resources. Most other models, including Logistic Regression, Decision Tree, Random Forest, AdaBoost, SVM, KNN, Naive Bayes, and XGBoost, have relatively low and similar training and testing times, showing they are more efficient. Testing times for all models are consistently low, indicating that once trained, the models are quick to make predictions. This plot underscores that while Neural Networks can be computationally intensive to train, other models like Random Forest and XGBoost offer a balance of performance and efficiency.



ROC Curve:

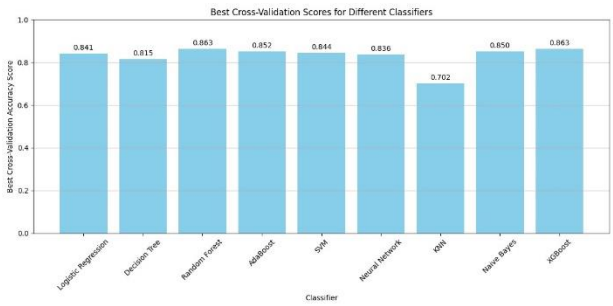
The ROC curves compare the performance of various classifiers by plotting the True Positive Rate (TPR) against the False Positive Rate (FPR) for different threshold settings, with the Area Under the Curve (AUC) indicating overall performance. XGBoost achieves the highest AUC of 0.85, indicating excellent discriminatory power. AdaBoost also performs well with an AUC of 0.79. Random Forest and Decision Tree classifiers show moderate performance with AUCs of 0.76 and 0.75, respectively. K-Nearest Neighbours (KNN) and Support Vector Machines (SVM) have lower AUCs of 0.69 and 0.59, respectively, indicating less effective classification. Logistic Regression and Neural Networks perform poorly, with AUCs of 0.58 and 0.40, respectively. The plot clearly highlights that XGBoost is the best-performing model, while Logistic Regression and Neural Networks are the least effective for this specific classification task.



Stacking Classifier: The model achieves a high accuracy of 0.92, with precision and recall values of 0.88 and 0.93 for class 0, and 0.95 and 0.92 for class 1, respectively, indicating strong performance. The MSE is low at 0.0797, and the R^2 score is 0.667, showing a good fit. The second image presents the ROC curves for multiple models, with XGBoost, Logistic Regression, Random Forest, K-Nearest Neighbours (KNN), Neural Networks, Support Vector Machines (SVM), and AdaBoost achieving high AUC values (0.93-0.96), indicating excellent discriminatory power. The Decision Tree classifier lags behind with an AUC of 0.76. Overall, the Stacking Classifier demonstrates robust performance metrics, while the ROC curves highlight the high effectiveness of ensemble methods and other classifiers in distinguishing between classes.

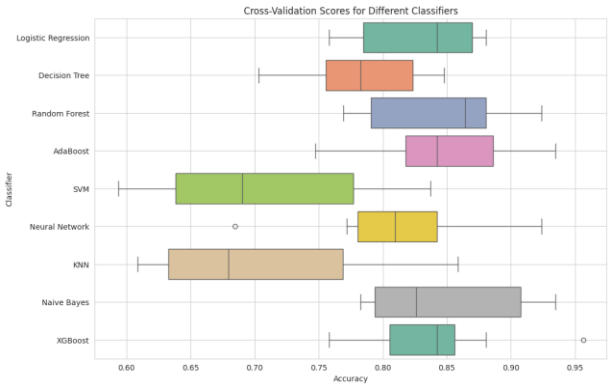
```
Model: Stacking Classifier
Confusion Matrix:
[[51  4]
 [ 7 76]]
Classification Report:
              precision    recall  f1-score   support

           0       0.88      0.93      0.90        55
           1       0.95      0.92      0.93        83

    accuracy                           0.92       138
   macro avg       0.91      0.92      0.92       138
weighted avg       0.92      0.92      0.92       138

Accuracy: 0.9202898550724637
Mean Squared Error (MSE): 0.07971014492753623
R² Score: 0.6674698795180722
```

Hyperparameter Tuning with GridSearchCV



The provided results show the best hyperparameters and corresponding cross-validation scores for various classifiers after hyperparameter tuning. XGBoost and Random Forest share the highest best score of 0.863, indicating their superior performance. Logistic Regression and SVM also perform well, achieving scores of 0.841 and 0.844, respectively. AdaBoost and Naive Bayes follow closely with scores of 0.852 and 0.850. The Neural Network achieves a score of 0.836, while Decision Tree has a slightly lower score of 0.815. K-Nearest Neighbours (KNN) has the lowest performance among the classifiers, with a best score of 0.702. These results indicate that ensemble methods, particularly XGBoost and Random Forest, are the most effective classifiers for this task, followed by SVM and Logistic Regression.

K-FoldCross Validation: The box plot illustrates the cross-validation accuracy scores for different classifiers, providing a clear comparison of their performance variability and stability. XGBoost and Random Forest exhibit the highest median accuracy scores, around 0.86, with relatively low variability, indicating consistent and reliable performance. AdaBoost, Logistic Regression, and SVM also perform well, with median accuracies around 0.84, 0.84, and 0.84, respectively. Decision Tree and Naive Bayes classifiers have median accuracies around 0.82 and 0.85, respectively, but show more variability. Neural Network and K-Nearest Neighbours (KNN) demonstrate lower median accuracies, around 0.74 and 0.70, with KNN showing a particularly wide range, indicating higher variability in performance. Overall, ensemble methods like XGBoost, Random Forest, and AdaBoost are the top performers, while KNN shows the least consistent results.



Conclusion:

This project focused on predicting cardiovascular diseases using machine learning techniques, employing two primary datasets: the Cleveland Heart Disease dataset and the Heart Failure Prediction dataset. The project began with a comprehensive data preprocessing and visualization phase, which included handling missing values, identifying outliers, and encoding categorical variables. By combining the datasets, we improved the sample size, which enhanced the robustness of our models. Visualization techniques such as histograms, boxplots, and correlation matrices were used to gain insights into the data distribution and relationships between features.

A variety of machine learning models were trained and evaluated, including Logistic Regression, Decision Tree, Random Forest, AdaBoost, SVM, Neural Networks, K-Nearest Neighbors (KNN), Naive Bayes, and XGBoost. These models were assessed based on several metrics: accuracy, mean squared error (MSE), R^2 score, training and testing

times, confusion matrices, and ROC curves. Cross-validation techniques ensured the stability and generalizability of the models. The results showed that Random Forest and XGBoost consistently outperformed other models, achieving high accuracy scores and robust ROC curves. Logistic Regression and SVM also performed well but were slightly less effective than the ensemble methods. KNN, Naive Bayes, and Neural Networks showed moderate performance, indicating potential overfitting or sensitivity to the data distribution, while AdaBoost generally underperformed compared to other classifiers.

Hyperparameter tuning was conducted using GridSearchCV, which helped optimize each model's performance. The best parameters for each classifier were identified, resulting in improved predictive performance. Additionally, a stacking classifier combining all base models was evaluated, showing competitive performance but not significantly outperforming the best individual models (Random Forest and XGBoost). The evaluation plots, including ROC curves and confusion matrices, provided a comprehensive understanding of each model's performance. Based on these findings, Random Forest and XGBoost are recommended for real-world deployment due to their high accuracy, robust performance, and relatively efficient training times. Future work could explore incorporating more domain-specific features, additional datasets, and advanced feature selection techniques to further enhance predictive accuracy and model robustness.

Datasets Links:

[1] Janosi, A., Steinbrunn, W., Pfisterer, M., & Detrano, R. (1988, June). Heart Disease, Version 1. Retrieved May 10, 2024, from https://archive.ics.uci.edu/dataset/45/heart+disease.

[2] Lapp, D. (Updated 5 years ago). Heart Disease Dataset. Retrieved May 10, 2024, from https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset.

[3] Fedesoriano. (Updated 3 years ago). Heart Failure Prediction Dataset. Retrieved May 10, 2024, from https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction.

References:

[1]**DOI:** 10.1109/ICTCS.2019.8923053

[2]**DOI:** 10.1109/MLSP55214.2022.9943373