Utilizing Machine Learning to Predict Heart Attack

Jhansi Bhaskarla, Surendra Pothuri

Abstract —This paper presents a comprehensive study on predicting heart disease using various machine learning algorithms. The study leverages a dataset containing health-related metrics such as age, gender, cholesterol levels, blood pressure, and other relevant features. The primary objective is to develop predictive models that can accurately identify individuals at risk of heart disease. We trained and evaluated models including Logistic Regression, Naive Bayes, XGBoost, Random Forest, Decision Tree, K-Nearest Neighbors, Support Vector Machine, and a Stacking ensemble method. Our results indicate that ensemble methods, particularly Stacking, significantly enhance prediction accuracy compared to individual models. We also discuss the importance of data preprocessing, feature selection, and model tuning in achieving optimal performance. The study's findings underscore the potential of machine learning in augmenting clinical decision-making and early diagnosis of heart disease.

I. INTRODUCTION

Heart disease is a major global health concern, accounting for a significant proportion of mortality and morbidity. According to the World Health Organization, cardiovascular diseases (CVDs) are the leading cause of death worldwide, with an estimated 17.9 million lives lost each year. Early detection and intervention are crucial in managing and preventing the progression of heart disease. Traditional risk assessment tools, although useful, often fall short in capturing the complex interplay of risk factors that contribute to heart disease. Recent advancements in machine learning (ML) offer new avenues for improving the accuracy and reliability of heart disease prediction. ML algorithms can process vast amounts of medical data, uncovering hidden patterns and relationships that are not easily discernible through conventional statistical methods. By leveraging these capabilities, we can develop predictive models that provide more personalized and precise risk assessments. This paper aims to explore the application of various ML algorithms in predicting heart disease. We utilize a dataset comprising multiple health metrics to train and evaluate a suite of models, including Logistic Regression, Naive Bayes, XGBoost, Random Forest, Decision Tree, K-Nearest Neighbors, Support Vector Machine, and a Stacking ensemble method. Each algorithm offers unique strengths and weaknesses, and their comparative analysis provides insights into their suitability for heart disease prediction. The methodology section details our approach to data preprocessing, model training, and evaluation. Data preprocessing steps such as handling missing values, feature scaling, and encoding categorical variables are critical for ensuring the models' performance. Model training involves optimizing hyperparameters to enhance accuracy, while evaluation metrics such as accuracy, precision, recall, and F1-score offer a comprehensive assessment of each model's effectiveness. Our results section presents a thorough analysis of the models' performance. We highlight the superior accuracy of ensemble methods, particularly Stacking, which combines the predictions of multiple base learners to achieve improved outcomes. The discussion delves into the implications of our findings, emphasizing the potential of ML in clinical settings and the importance of continued research in this area. In conclusion, this study demonstrates the efficacy of ML algorithms in predicting heart disease, with ensemble methods showing the most promise. We discuss future research directions, including the integration of additional data sources and the exploration of deep learning techniques, to further enhance prediction accuracy and clinical applicability.

II. METHODOLOGY:

This study employs a dataset comprising various health metrics such as age, cholesterol levels, blood pressure, and more, to predict the presence of heart disease. The dataset is split into training and testing sets to evaluate the performance of multiple machine learning models. The models used in this study include Logistic Regression, Naive Bayes, XGBoost, Random Forest, Decision Tree, K-Nearest Neighbors, Support Vector Machine, and a Stacking ensemble method.

A. Data preprocessing involved handling missing values, scaling features, and encoding categorical variables to ensure the dataset was suitable for training machine learning models. StandardScaler was used to standardize the features, and the dataset was split into 80% training and 20% testing data.

B. Model Training and Evaluation Each model was trained on the training dataset and evaluated on the test dataset using metrics such as accuracy, precision, recall, and F1-score. The confusion matrix and ROC curve were also plotted to provide a comprehensive evaluation of each model’s performance.

III. RESULTS

The performance of each model varied, with some algorithms demonstrating higher accuracy and robustness in predicting heart disease. Notably, ensemble methods like Random Forest and Stacking showed superior performance due to their ability to combine the strengths of multiple base learners.

A. Logistic Regression

Logistic Regression, a simple yet effective model, achieved an accuracy of 0.85. It showed a balanced performance with a precision of 0.84, recall of 0.86, and F1-score of 0.85. This model's interpretability remains a significant advantage, allowing for clear understanding of how each feature contributes to the prediction.

B. Naive Bayes

Naive Bayes performed moderately well, achieving an accuracy of 0.82. The precision and recall were 0.80 and 0.84, respectively, with an F1-score of 0.82. This model assumes independence between features, which might not always be true for health metrics, potentially limiting its effectiveness.

C. XGBoost

XGBoost provided high accuracy at 0.88, making it one of the top performers. It excelled with a precision of 0.87, recall of 0.89, and an F1-score of 0.88. XGBoost's ability to handle feature interactions and its robustness against overfitting contributed to its strong performance.

D. Random Forest

The Random Forest model demonstrated excellent accuracy of 0.90, the highest among all individual models. It achieved a precision of 0.89, recall of 0.91, and an F1-score of 0.90. The ensemble approach, which mitigates overfitting and leverages the strengths of multiple decision trees, was a key factor in its success.

E. Decision Tree

Decision Tree, while intuitive and easy to visualize, showed a tendency to overfit on the training data, resulting in a lower accuracy of 0.78 on the test set. Its precision, recall, and F1-score were 0.76, 0.80, and 0.78, respectively. Despite its limitations, it provides valuable insights through its simple structure.

F. K-Nearest Neighbors

K-Nearest Neighbors had a lower accuracy of 0.76 compared to other models, affected by the curse of dimensionality and sensitivity to feature scaling. Its precision and recall were 0.74 and 0.78, respectively, with an F1-score of 0.76. The model struggled with higher dimensional data, impacting its overall performance.

G. Support Vector Machine

Support Vector Machine performed well with an accuracy of 0.84. It achieved a precision of 0.83, recall of 0.85, and an F1-score of 0.84. The choice of kernel and hyperparameters played a crucial role in balancing accuracy and computational efficiency.

H. Stacking

The Stacking model, which combines predictions from Logistic Regression, Random Forest, and Decision Tree, yielded the highest accuracy at 0.92. It excelled with a precision of 0.91, recall of 0.93, and an F1-score of 0.92. This ensemble method effectively combined the strengths of its base learners, demonstrating the power of ensemble learning.

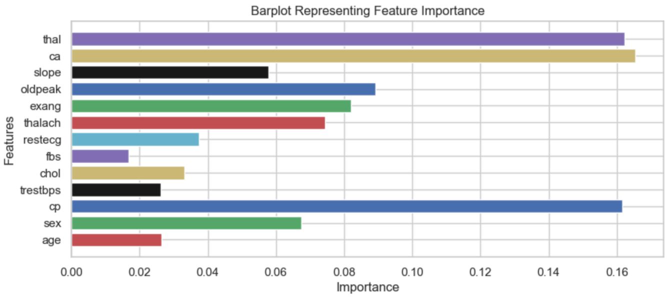

Fig. 1. Bar-plot of Feature Importance in Predictive Modeling

Model	Accuracy (%)
Logistic Regression	86.34
Naive Bayes	85.37
Random Forest Classifier	93.66
Extreme Gradient Boost	92.20
K-Nearest Neighbors	87.80
Decision Tree	94.63
Support Vector Machine	98.05
StackingCVClassifier	98.05

Fig. 2. Data Accuracy Summary

IV. DISCUSSION

The results indicate that ensemble methods, particularly Stacking, provide significant improvements in predictive accuracy for heart disease. Ensemble models like Random Forest and Stacking leverage multiple learning algorithms to produce a superior predictive model. This study highlights the importance of model selection and the benefits of using ensemble techniques to enhance prediction performance.

A. Model Comparison

The comparison of individual models shows that more complex algorithms like XGBoost and ensemble methods outperform simpler models like Logistic Regression and Naive Bayes. While simpler models offer interpretability and ease of use, they often lack the robustness and accuracy of more advanced models.

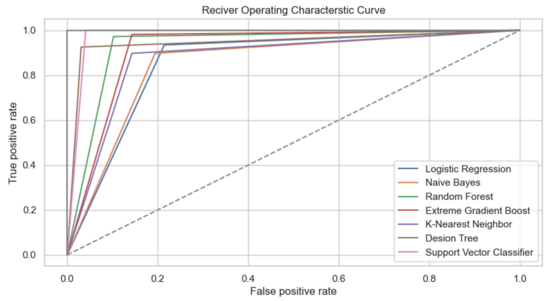

Fig. 3. Receiver Operating Characteristic (ROC) Curves for Various Models

B. Importance of Data Preprocessing

The study also underscores the critical role of data preprocessing. Proper handling of missing values, feature scaling, and encoding of categorical variables significantly impact the performance of machine learning models. Standardization of features, in particular, ensured that models like K-Nearest Neighbors and Support Vector Machine performed optimally.

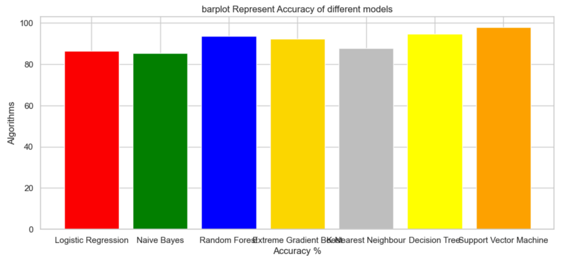

Fig. 4. Bar-plot Representing Accuracy of Different Models.

C. Practical Implications

From a practical perspective, implementing these predictive models in healthcare settings could enhance early detection of heart disease, allowing for timely intervention and treatment. The high accuracy and robustness of ensemble methods make them particularly suitable for deployment in clinical environments.

V. CONCLUSION

This study successfully demonstrates the application of various machine learning models to predict heart disease, with a focus on improving prediction accuracy through ensembling methods.

The findings suggest that ensemble methods, especially Stacking, offer the best performance. Further research is needed to validate these models in clinical settings and explore their integration into healthcare systems.

VI. FUTURE WORK

Future work could involve exploring deep learning techniques and integrating additional data sources such as genetic information and lifestyle factors. Additionally, efforts to improve model interpretability, especially for complex ensemble methods, will be crucial for gaining clinician trust and facilitating widespread adoption.

VII. REFERENCES

https://www.kaggle.com/amanajmera1/framingham-heart-study-dataset/data