Heart Attack Analysis: Predictive Modeling for Heart Attack Classification

Brad Washburne & Dakota Doyle

November 17th, 2024

Abstract

The underlying goal throughout this project was to fit and compare classification models to predict the presence of heart disease based on patient health data. We used a public dataset from Kaggle, with 303 samples and 14 features, to explore 6 different classification models to identify the most robust model for heart attack prediction. We initially performed exploratory data analysis and preprocessed the data to handle missing values and standardize the features using the StandardScaler from sklearn. From here, we fit 6 different classification models from the sklearn package: Decision Tree, Logistic Regression, Neural Network, Gaussian Naïve Bayes, Bernoulli Naïve Bayes, and k-Nearest Neighbors – utilizing GridSearchCV to find the best hyperparameters for each model. After fitting each model, we created a data frame and visualizations to compare the metrics for each model. In our findings, the Gaussian Naïve Bayes, Bernoulli Naïve Bayes, and k-Nearest Neighbors classifiers achieved the highest test accuracies of 86.9%, followed by Logistic Regression and Neural Network at 85.2%, and finally, Decision Tree with 83.6% accuracy. Beyond accuracy, we also performed cross validation for each model, and visualized the ROC curves to further validate the performance of the models. Lastly, we created a heat map to visualize the feature importance across the different models, identifying key predictors like chest pain type and age. Overall, our results established the relatively high accuracy of machine learning classifiers to effectively predict heart disease – findings that could be utilized in medical domains and research.

Introduction

Our project focused on evaluating and comparing various machine learning classifiers to predict heart disease by using patient health data. Heart disease is the leading cause of death in the United States, accounting for 1 in 5 deaths in adults. With heart disease being a widespread

issue, plaguing hundreds of thousands of Americans each year, we wanted to apply our knowledge with machine learning to predict the presence of heart disease from easily attainable health data. Although heart disease is not curable, lifestyle adjustments can significantly slow and even halt its progression, making early detection crucial in saving lives. In the domain of data mining, heart disease prediction lends itself perfectly to classification models. Key health indicators like age, gender, blood pressure, and cholesterol levels can all be used to predict the presence of heart disease in individuals. Essentially, our classification models exemplify the power of extracting insights from real-world datasets, which can in turn enhance decision making in healthcare. Predicting heart disease through patient data not only aids in early detection but also optimizes healthcare resources to target those most at risk.

Literature Review

Early detection is a crucial component for effective treatment regarding heart disease, and as a result, the use of machine learning models for heart disease prediction has been widely explored in recent years. In fact, much like the goal of our project in comparing the performance of multiple different classification models, several studies have demonstrated the effectiveness of different machine learning approaches for heart disease classification. For example, one study published in 2022 by a team of researchers found that machine learning has shown promise in improving cardiovascular risk prediction by integrating data from EHRs and wearable device metrics, which allowed more personalized and accurate metrics than traditional methods. Another study published in 2023 by Ahmad Ayid Ahmad and Huseyin Polat utilized the Jellyfish optimization algorithm – a metaheuristic optimization technique that is modeled after the natural behavior of jellyfish in the ocean – significantly improving the accuracy of machine learning models for predicting heart disease. In fact, the SVM model they trained was able to achieve

98.47% accuracy, significantly beating the performance of any of our models, and showcasing the potential of integrating intelligent algorithms to enhance machine learning models. Lastly, we found a study that used the exact same dataset we are using in our project, which found the k-Nearest neighbor algorithm to be the best approach with an accuracy of 90.16%. What most interests us about this study was the choice of metrics: in their evaluation, they did not use GridSearchCV or cross-validation scores. Thus, we are intrigued to see how well their k-NN model would generalize against our most robust models. In any sense, machine learning applications for predicting heart disease is an extremely well-studied domain with plenty of different research in attempting to find the best classifiers.

Methodology

Throughout our project, we utilized a variety of built-in libraries to effectively preprocess our data, fit machine learning classifiers, and visualize the findings. The main libraries we used were Pandas, NumPy, Matplotlib, Seaborn, and scikit-learn.

To import the dataset and preprocess the data, we used Pandas to create a data frame. From here, we preformed exploratory data analysis, which mostly utilized Matplotlib and Seaborn to visualize the distributions and correlations between features. Next, we used train test split and StandardScaler from the sklearn package to split our data into training and testing groups and standardize the data. In addition, we checked for missing data and outliers, however, the data was extremely clean as there was no missing data.

After preprocessing, we fit 6 different classification models from the sklearn package: Decision Tree, Logistic Regression, Neural Network, Gaussian Naïve Bayes, Bernoulli Naïve Bayes, and k-Nearest Neighbors. Furthermore, we imported functions from the sklearn metrics and model

selection packages to find the accuracy, classification report, and cross validation score for each model, as well as importing GridSearchCV to find the best hyperparameters for each model.

After fitting and finding the metrics associated with each model, we again used Pandas and Matplotlib to organize the metrics of each model and visualize the performance of each model for comparison. Lastly, we imported AUC score, ROC curve, and permutation importance from sklearn to visualize the ROC curve for each model and create a heatmap to compare feature importance between the different classifiers. Overall, the combination of techniques and built in python libraries allowed us to benchmark a range of approaches, and compare different classifiers to determine the best model for predicting heart disease.

Dataset

In our analysis, we utilized a dataset from Kaggle titled *Heart Attack Analysis & Prediction* published by Rashik Rahman - a dataset master on Kaggle. This dataset earned a perfect usability score of 10.0 on Kaggle, indicating its credibility. The data was stored as a CSV, and contains 303 rows or different patients, with 14 different features – one being the target feature identifying whether a patient had heart disease or not. The predictive features all consisted of patient data, including age, gender, chest pain type, resting blood pressure, cholesterol level, fasting blood sugar rate, resting ECG, max heart rate, exercise induces angina, previous peak, slope, number of major vessels, and thal rate.

In terms of preprocessing, we first checked for missing values, however, we found none which simplified the cleaning process. Additionally, all categorial features were already encoded, reducing the need for additional preprocessing. After an initial feature analysis, we renamed the features columns to improve interpretability as the original column names were unclearly

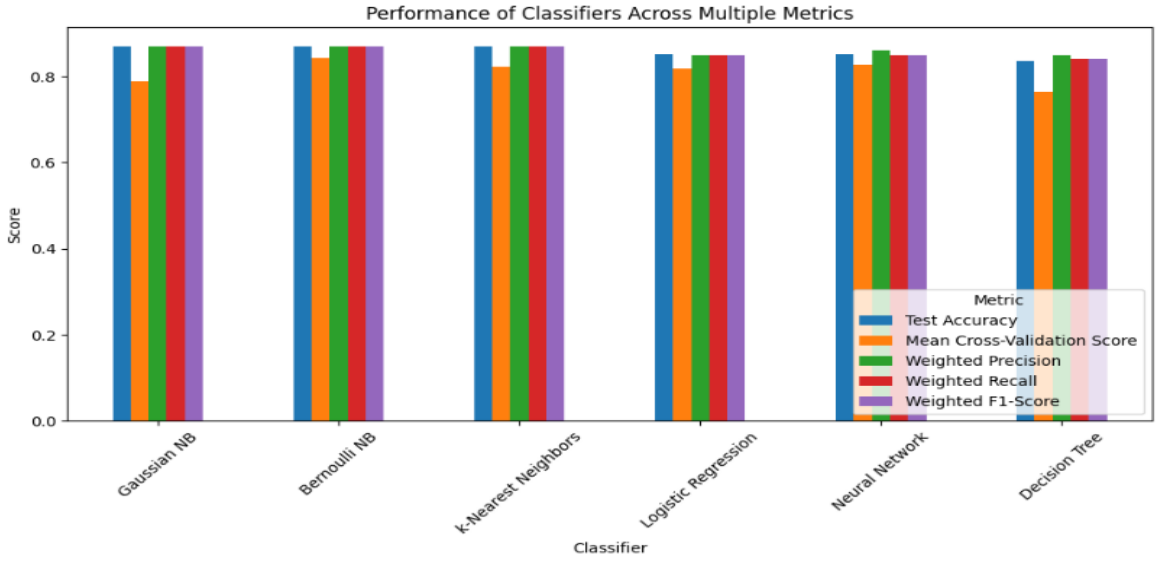
abbreviated. We then used StandardScaler to standardize the features and in turn optimize model performance. From here, we split the data into training and testing sets with an 80/20 split. Overall, our dataset was both reliable and clean, which allowed us to better fit the classification models and compare results without having to focus entirely on cleaning the data.

Results

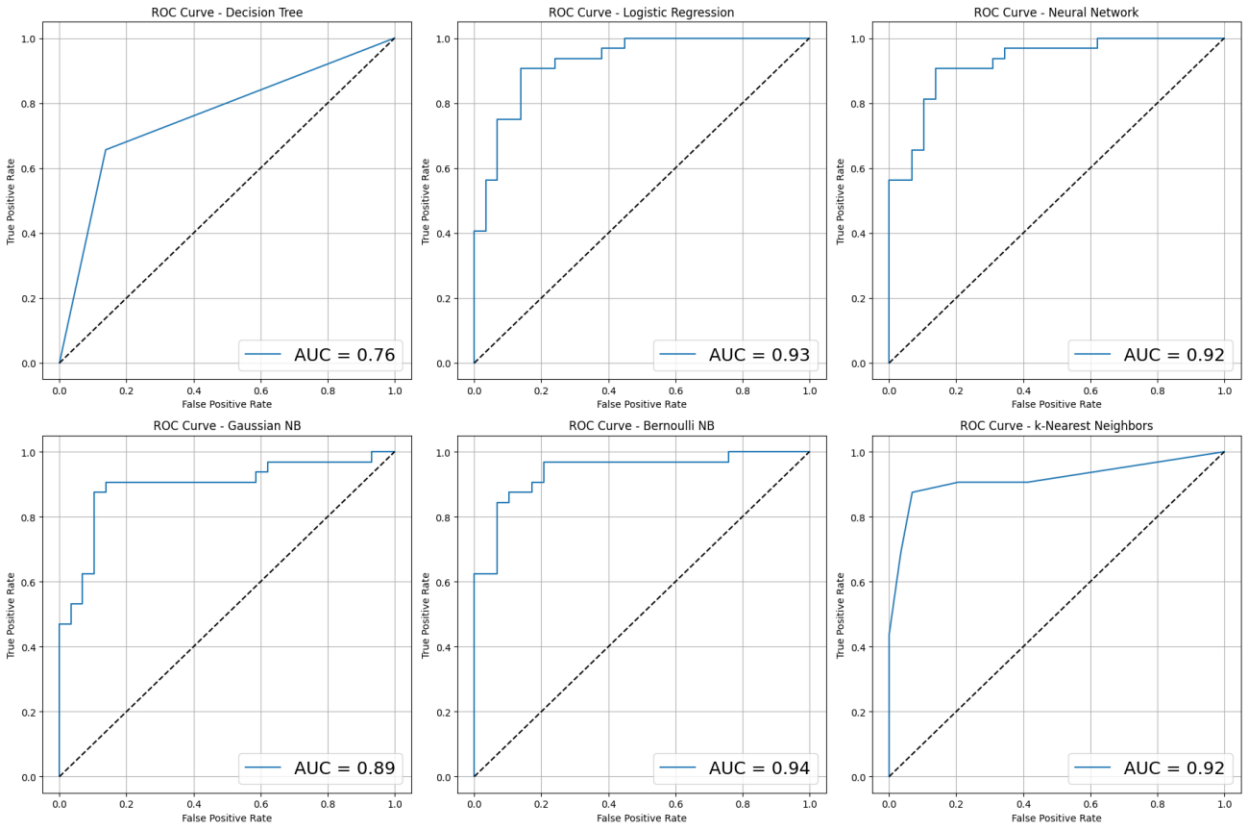
To evaluate our models, we used serval metrics, including accuracy, precision, recall, F1-score, and mean cross validation score. We applied these metrics to each of these models, summarized in the table below:

	Model	Test Accuracy	Mean Cross-Validation Score	Weighted Precision	Weighted Recall	Weighted F1-Score
0	Gaussian NB	0.8689	0.7890	0.87	0.87	0.87
1	Bernoulli NB	0.8689	0.8429	0.87	0.87	0.87
2	k-Nearest Neighbors	0.8689	0.8224	0.87	0.87	0.87
3	Logistic Regression	0.8525	0.8180	0.85	0.85	0.85
4	Neural Network	0.8525	0.8263	0.86	0.85	0.85
5	Decision Tree	0.8361	0.7647	0.85	0.84	0.84

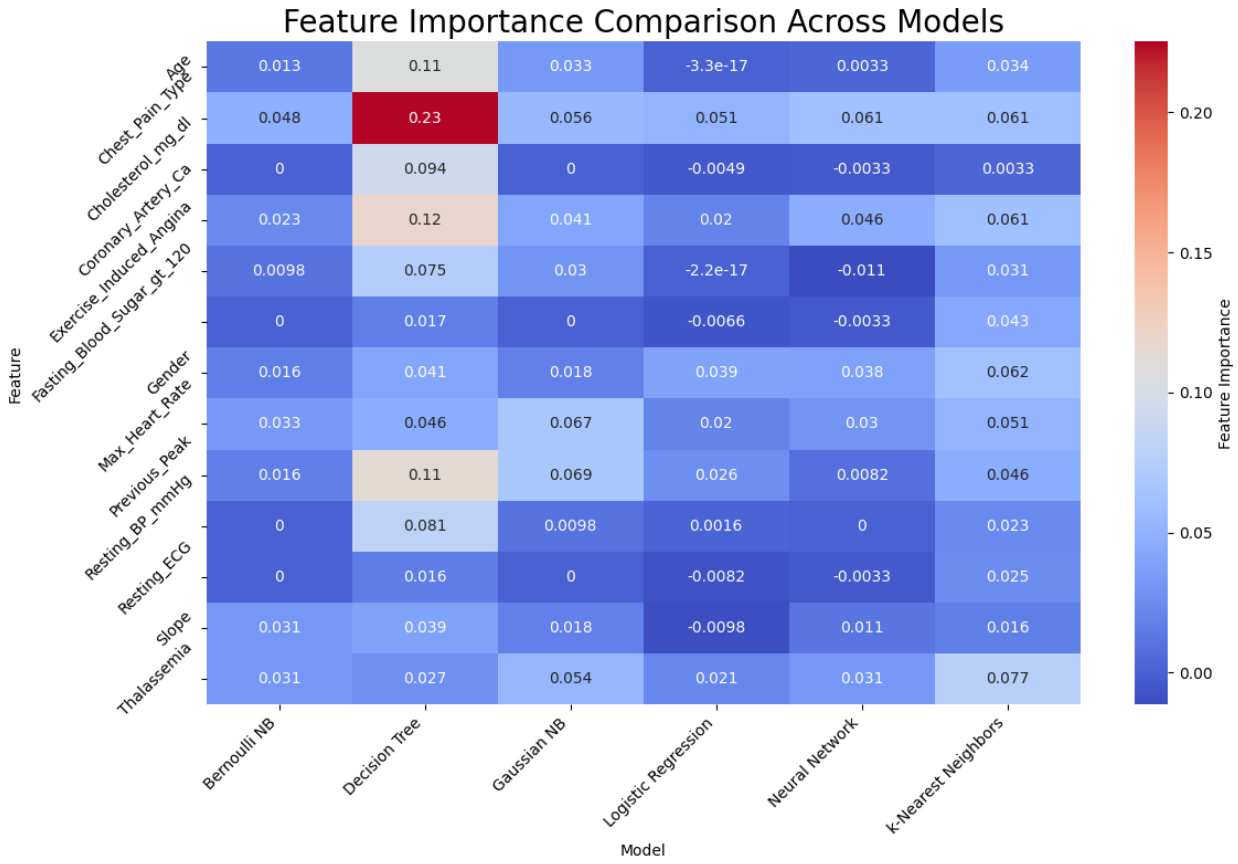
To visualize the metrics, we also generated a bar chart for each model as seen below:



Beyond these metrics, we also plotted the ROC curve for each model and calculated the AUC:



Lastly, we analyzed the feature importance by creating a heatmap to compare the different features across models:



Discussion

Overall, all the models performed well, with the Gaussian Naïve Bayes, Bernoulli Naïve Bayes, and k-Nearest Neighbors classifiers achieving the highest test accuracies of 86.9%, followed by Logistic Regression and Neural Network at 85.2%, and finally, Decision Tree with 83.6% accuracy. Amongst these models, however, the Bernoulli Naïve Bayes classifier had the highest cross validation score, indicating that it likely generalizes the best of the six classifiers. For the other metrics of precision, recall, and F1-score, all of the models had very similar scores – all within the 0.84-0.87 range, further indicating reliability in predicting heart disease.

Regarding the ROC curves and AUC scores, the Bernoulli Naïve Bayes model performed the best with an AUC score of 0.94, followed by Logistic Regression with a score of 0.93, Neural Network and k- Nearest Neighbors with 0.92, Gaussian Naïve Bayes with 0.89, and the lowest score being the Decision Tree with 0.76. Combining both the metrics and ROC curves, we can conclude that the Bernoulli model is the most robust for classifying heart disease, while the Decision Tree is the weakest performer. These results align with the strengths and weakness of each model, as the Bernoulli model is particularly strong for binary classifications, like predicting whether someone has heart disease, while the Decision Tree model is prone to overfitting and lacks smooth predictions.

As seen in the feature importance heatmap above, chest pain type stands out as the most significant predictor. Other significant predictors include max heart rate and age, which displayed high importance across all models. This makes sense, as age, chest pain, and heart rate are three main indicators used by medical professionals to diagnose heart disease. Even so, the variability in feature importance between each of the models reflects how different classifiers interpret and utilize features.

In terms of challenges, one of our primary goals was to reduce the risk of overfitting to ensure our models would generalize effectively. To address this, we utilized GridSearchCV to tune the hyperparameters of each model, ultimately finding the best parameters for each model, and used cross-validation scores to quantify how well the models would generalize – not just how well the model performed on the test set. Another challenge was feature scaling, specifically to reduce the computational time needed to fit and tune six different models. Admittedly this process was made easier because of the inherently clean nature of the data, yet had to be addressed. As a solution, we used StandardScaler to standardize the features and ensure these models performed

optimally. Lastly, interpreting the results of the feature importance across different models is inherently difficult due to the nature of different models. For example, models like Decision Trees have highly interpretable results, compared to Neural Networks due to how the different algorithms handle features. Regardless, we found that the heatmap was an effective way to try to quantify how important certain features were in predicting heart disease amongst the patient data. Even so, by addressing all our challenges, we were able to build strong classifiers, and ultimately derive meaningful insights.

Conclusion

In this project, we were successfully able to fit and compare six different machine learning models to predict heart disease using patient data. The results demonstrated that the Bernoulli Naïve Bayes, Gaussian Naïve Bayes, and k-Nearest Neighbors models were the most accurate, followed by the Logistic Regression and Neural Network models, and lastly the Decision Tree model. Furthermore, with other metrics including the ROC curve, the Bernoulli model proved to be the most robust, excelling in both accuracy, cross-validation scores, and having the highest AUC score. On the other hand, the Decision Tree classifier was clearly the weakest performer, having the lowest score for each of the above metrics. Lastly, our feature importance analysis highlighted Chest Pain Type, Max Heart Rate, and Age to be key predictors in classifying for heart disease.

In the future, we could explore fitting different models – one specifically that seems relevant being random forest. Because the Decision Tree model performed poorly in comparison to the other models, the random forest may improve results due to the more powerful nature of the model. Furthermore, expanding our dataset could also improve the model's generalization. While our models performed well, the dataset only included 303 samples. Thus, if we would want to

generalize our findings to a large domain, like all American adults, we would want to have a more representative sample. Along with this, we would likely want to incorporate the knowledge of a domain expert in heart disease to determine if there are other important features to include. Lastly, deploying our models in the real world with live clinical data could provide valuable insight into how well our models generalize, and could eventually provide support to healthcare providers in early diagnosis and treatment planning.

Resources

Ahmad, Ahmad Ayid, and Huseyin Polat. “Prediction of Heart Disease Based on Machine Learning Using Jellyfish Optimization Algorithm.” *Diagnostics (Basel, Switzerland)*, U.S. National Library of Medicine, 17 July 2023, pmc.ncbi.nlm.nih.gov/articles/PMC10378171/.

Heart Attack Prediction Using Machine Learning Approach | IEEE Conference Publication | IEEE Xplore, ieeexplore.ieee.org/document/9969395/. Accessed 10 Nov. 2024.

“Heart Disease Facts.” *Centers for Disease Control and Prevention*, Centers for Disease Control and Prevention, www.cdc.gov/heart-disease/data-research/facts-stats/index.html#:~:text=Heart%20disease%20in%20the%20United%20States&text=One%20person%20dies%20every%2033,1%20in%20every%205%20deaths. Accessed 10 Nov. 2024.

Javaid, Aamir, et al. “Medicine 2032: The Future of Cardiovascular Disease Prevention with Machine Learning and Digital Health Technology.” *American Journal of Preventive Cardiology*, U.S. National Library of Medicine, 29 Aug. 2022, pmc.ncbi.nlm.nih.gov/articles/PMC9460561/#:~:text=In%20one%20study%2C%20an%20ensemble,sensitivities%20of%200.20%20and%200.38%2C.

Rahman, Rashik. *Heart Attack Analysis & Prediction Dataset*. Kaggle, 2020, https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset.