Jefferson Veliz

Professor Avinash Jairam

Data mining for business analytics

August 05th 2023

Project #1

My family's history of heart disease motivated me to take on this project. I wanted to understand the signs and factors linked to heart disease and learn how machine learning could help. I began by exploring the data, checking for missing values, and understanding the different types of information in each column. The dataset, obtained from Kaggle, contained 1025 rows and 14 columns, which formed the foundation of my analysis.

The visualizations performed in this project provide valuable insights into the relationships within the dataset and the predictive power of the machine learning models. The box plot depicting the distribution of age by heart disease status offers a clear overview of how age impacts heart disease occurrence. The correlation plot reveals that certain features like chest pain type (cp), maximum heart rate achieved (thalach), and slope are correlated with the target, providing hints of potential predictive power. The scatter plot showcasing Age vs Cholesterol color-coded by heart disease status allows us to visually grasp any trends or patterns in the data. The count plot detailing the number of individuals with heart disease gives us a sense of the data balance. Furthermore, the bar plot presenting feature importance scores offers a ranked view of clinical indicators that play pivotal roles in predicting heart disease. The subsequent bar plot presenting machine learning models based on multiple metrics highlights Decision Trees as the most successful model. Finally, the inclusion of a confusion matrix helps us understand how well the models performed in classifying instances. These visualizations collectively shed light on crucial aspects of the data and model performance.

The K-Nearest Neighbors (KNN) algorithm was utilized for binary classification to predict the presence or absence of heart disease. This report provides an in-depth evaluation of the model's performance, primarily focusing on the weighted average metrics and overall accuracy. The weighted average precision is 0.88, indicating an overall balanced accuracy of the model's predictions across both classes. The weighted average recall is 0.88, highlighting the model's ability to effectively capture instances from both the "heart disease" and "no heart disease" classes. The weighted average F1-score is 0.88, demonstrating a harmonious balance between precision and recall for both classes. The overall accuracy of the KNN model is 0.88. This suggests that the model correctly classified 88% of all instances in the dataset.

The Gaussian Naive Bayes (GNB) algorithm was employed for binary classification to predict the presence or absence of heart disease. This section presents a comprehensive assessment of the model's performance. The weighted average precision stands at 0.84, signifying a well-balanced accuracy of the model's predictions across both "heart disease" and "no heart disease" classes. Furthermore, the weighted average recall is measured at 0.84, suggesting that the model adeptly captures instances from both classes. The weighted average F1-score, a reflection of the harmonious equilibrium between precision and recall, reaches 0.84. In terms of overall accuracy, the GNB model achieves a accuracy rate of 0.84, implying that it accurately classified 84% of all instances in the dataset.

The Polynomial Kernel Support Vector Machine (SVM) algorithm was harnessed for binary classification, directed at predicting the presence or absence of heart disease. This section embarks on an extensive analysis of the model's performance. The weighted average precision reaches a significant 0.95, underscoring the model's mastery in achieving a balanced accuracy in its predictions across both "heart disease" and "no heart disease" classes. Simultaneously, the weighted average recall attains a value of 0.95, showcasing the model's adeptness in capturing instances from both categories. The weighted average F1-score, which encapsulates the harmonious trade-off between precision and recall, attains a substantial 0.95. The overall accuracy of the Polynomial Kernel SVM model is an impressive 0.95, highlighting its ability to accurately classify 95% of all instances in the dataset.

The Sigmoid Kernel Support Vector Machine (SVM) algorithm was employed in the realm of binary classification, with the aim of predicting the presence or absence of heart disease. This section provides a detailed analysis of the model's performance. The weighted average precision holds a value of 0.79, indicative of a balanced accuracy achieved by the model's predictions encompassing both "heart disease" and "no heart disease" classes. Concurrently, the weighted average recall stands at 0.79, illustrating the model's adeptness in capturing instances from both categories. The weighted average F1-score, which signifies the harmonious equilibrium between precision and recall, registers at 0.79. In the realm of overall accuracy, the Sigmoid Kernel SVM model reports an accuracy rate of 0.79, underscoring its capability to correctly classify 79% of all instances in the dataset.

The Radial Basis Function (RBF) Kernel Support Vector Machine (SVM) algorithm was harnessed for binary classification, specifically targeting the prediction of heart disease presence or absence. The weighted average precision achieves a remarkable score of 0.99, underlining the model's prowess in attaining balanced accuracy across both "heart disease" and "no heart disease" classes. Simultaneously, the weighted average recall demonstrates an equally impressive value of 0.99, showcasing the model's efficacy in capturing instances from both categories. The weighted average F1-score, symbolizing the balance between precision and recall, attains an outstanding 0.99. As for the overall accuracy, the RBF Kernel SVM model boasts an accuracy rate of 0.99, highlighting its exceptional ability to accurately classify 99% of all instances in the dataset.

The Decision Tree algorithm was employed to facilitate binary classification, with the primary goal of predicting the presence or absence of heart disease. This section offers a comprehensive evaluation of the model's performance. The weighted average precision achieves a perfect score of 1.00, signifying impeccable accuracy in the model's predictions spanning both "positive" (heart disease) and "negative" (no heart disease) classes. Simultaneously, the weighted average recall attains a perfect value of 1.00, demonstrating the model's exceptional ability to capture instances from both categories. The weighted average F1-score, serving as an indicator of the harmonious equilibrium between precision and recall, reaches a flawless 1.00. In terms of overall accuracy, the Decision Tree model attains a perfect accuracy rate of 1.00, certifying its accurate classification of 100% of all instances in the dataset.

The Random Forest algorithm was harnessed for binary classification, directed at predicting the presence or absence of heart disease. This section conducts a detailed analysis of the model's performance. The weighted average precision achieves an outstanding 0.99, signifying the model's mastery in achieving balanced accuracy across both "heart disease" and "no heart disease" classes. Simultaneously, the weighted average recall attains an impressive value of 0.99, underscoring the model's adeptness in capturing instances from both categories. The weighted average F1-score, attains a 0.99. The overall accuracy of the Random Forest model is an exceptional 0.99, spotlighting its ability to accurately classify 99% of all instances in the dataset.

Feature importance scores derived from machine learning models offer invaluable insights into the relative contributions of various clinical indicators. Here, we highlight why doctors should care about these importance scores, focusing on the significance of each feature in predicting heart disease: The highest importance score indicates that the type of chest pain a patient experiences plays a pivotal role in heart disease prediction. As a medical professional, recognizing distinct chest pain patterns can aid in diagnosing underlying cardiac issues and determining the urgency of interventions. With a significant importance score, the count of major vessels colored during fluoroscopy offers valuable insights into coronary artery health. This information enables doctors to make informed decisions about revascularization procedures and overall treatment strategies. Notable importance is attributed to serum cholesterol levels, which are closely linked to heart disease risk. For healthcare practitioners, this score emphasizes the need to manage cholesterol levels through targeted interventions and patient education. Overall, Feature importance scores highlight the variables that have the most influence on heart disease prediction. By focusing on these key factors, doctors can enhance diagnostic accuracy, ensuring that patients receive timely and appropriate interventions.

I chose to deploy the Decision Tree algorithm for its remarkable performance as demonstrated in the classification report. The precision, recall, and F1-score metrics indicated a perfect accuracy rate of 1.00, signifying impeccable classification of both positive and negative instances. By selecting the Decision Tree model, I aim to provide “healthcare professionals” with a reliable and easy-to-understand tool for predicting heart disease. This ensures that even those without extensive technical background can effectively utilize the model's insights and contribute to improved patient care decisions.

In summary, this project was motivated by a personal connection to heart disease within my family. The goal was to explore the features that contribute to heart disease and to learn about machine learning techniques. Through data exploration, I checked for missing values, data types, and correlations in the dataset. Several machine learning models were evaluated for predicting heart disease. The K-Nearest Neighbors, Gaussian Naive Bayes, Polynomial Kernel SVM, Sigmoid Kernel SVM, Decision Tree, and Random Forest, and logistic regression algorithms were analyzed based on their performance metrics. These models demonstrated varying levels of accuracy in predicting heart disease presence or absence. Additionally, feature importance scores shed light on the significance of different factors in predicting heart disease. Notably, chest pain type, major vessel health, and serum cholesterol levels emerged as crucial predictors. These insights have practical implications for healthcare professionals. By recognizing specific chest pain patterns, evaluating vessel health, and managing cholesterol levels, doctors can enhance diagnostic accuracy and improve patient care.

**Visualizations A chart of a distribution of age by heart disease status

Description automatically generated**

**Figure 1: Boxplot -**Distribution of age by heart disease Status. We can see the medium age for people with no heart disease is around 57 while for people with heart disease is around 52.

**A graph of percentages

Description automatically generated with medium confidence**

**Figure 2:Correlation plot-** individuals with cp, thalach, slope have a position correlation with the target. However, we can observe other features that negatively correlated with the target feature.

**A graph of red and blue dots

Description automatically generated**

**Figure 3: Scatter plot-** Age vs Cholesterol with target color of individuals with heart disease and no heart disease. Individuals aged from 50 to 70 seemed to be affect by heart disease than younger individuals

A graph of heart disease

Description automatically generated

**Figure 4: count plot-** Count of individual with heart disease and no heart disease in the dataset. There were more Individuals with heart disease in the dataset than with no heart disease.

A graph of different colored bars

Description automatically generated

**Figure 5 : countplot** individual’s age with no heart disease and heart disease**.** We can see the data more clearly. There are individuals with heart disease starting at the age of 29.

A blue squares with white text

Description automatically generated

**Figure 6:** confusion matrix- 117 instances where the actual outcome was positive (heart disease) and the model correctly predicted it as positive (TP).28 instances where the actual outcome was positive (heart disease), but the model incorrectly predicted it as negative (FN).13 instances where the actual outcome was negative (no heart disease), and the model incorrectly predicted it as positive (FP).150 instances where the actual outcome was negative (no heart disease), and the model correctly predicted it as negative (TN).

A graph with blue bars

Description automatically generated

**Figure 7:** Importance feature scores from high to low. We can observe the most important feature is CP and least important is fbs for the random forest model.

A graph of different colored bars

Description automatically generated

**Figure 8:** Machine learning models represented based on Accuracy, weighted recall, f1, and precision. Decision tree is the best performing model among all of the other machine learning models.

Link for Model

<http://jefferson1580.pythonanywhere.com/example_route>

References

<https://chat.openai.com> – To fix grammar and debug code. I learned to ask chatgpt the right questions. I used it as a tutor to refine my skills.