Analysis Report
Heart Disease Prediction and Blood Pressure Analysis

Krishna Sai Pendem

INTRODUCTION

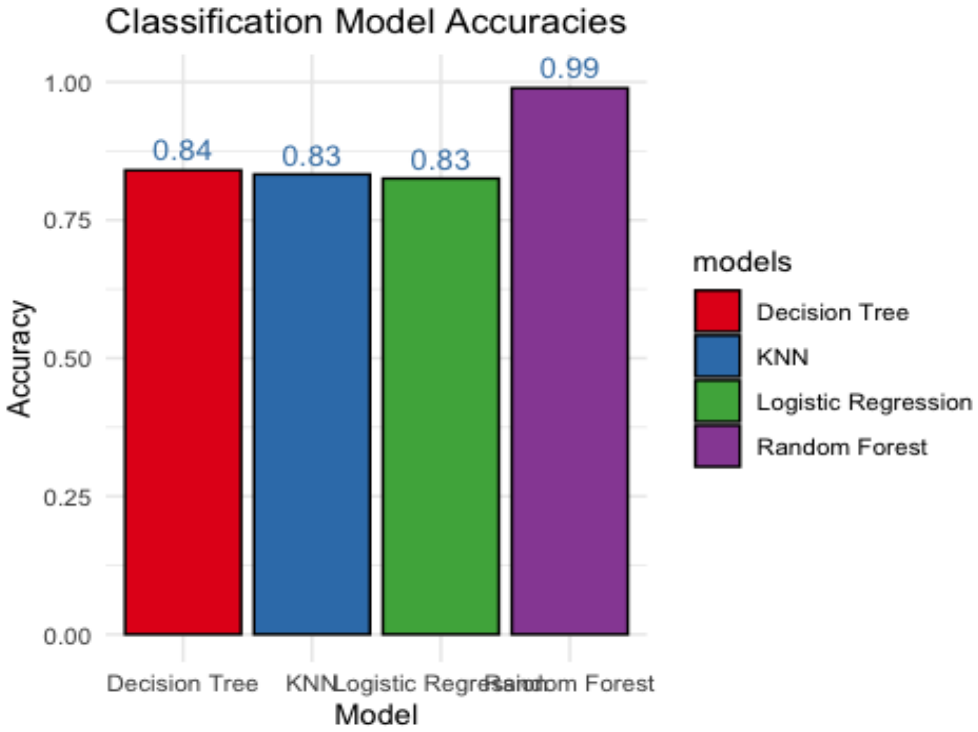
In this report, we present the results of our project aimed at predicting the presence of heart disease in individuals and analyzing the mean blood pressure difference between patients with and without heart disease. We addressed two primary research questions:

Research Question 1:

Which classification method resulted in the highest accuracy in predicting the presence of heart disease in individuals?

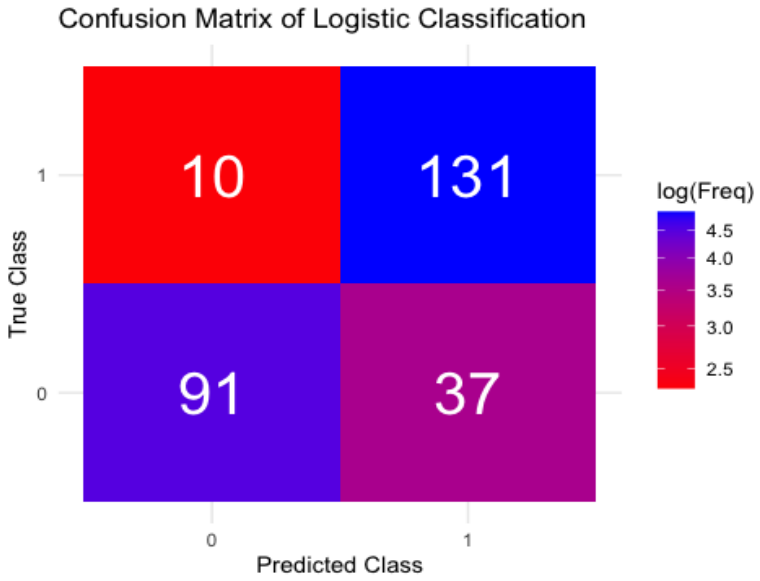
To answer this question, we explored the performance of four machine learning models, including Logistic Regression, K-Nearest Neighbors (KNN), Random Forest Classification, and Decision Trees Classification. The accuracy scores and confusion matrices for these models are as follows:

- Logistic Regression accuracy: 0.83
- KNN accuracy: 0.83
- Random Forest Classification accuracy: 0.99
- Decision Trees Classification accuracy: 0.84

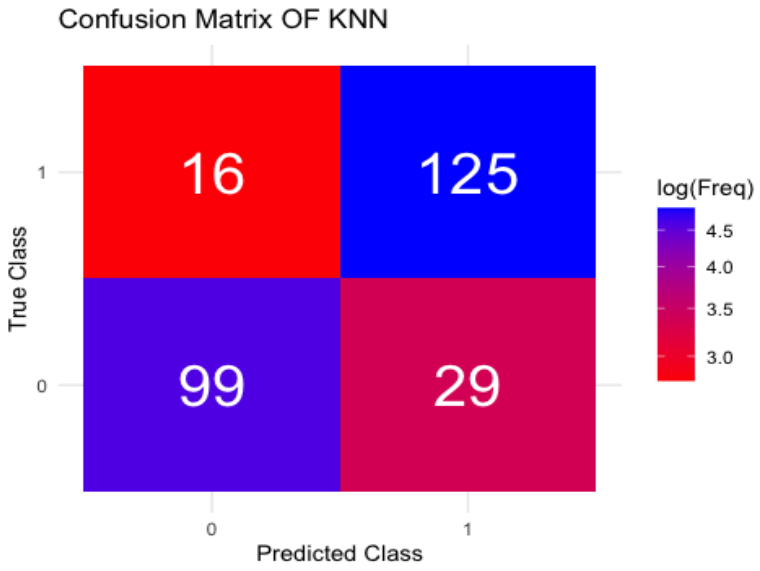


Confusion matrices:

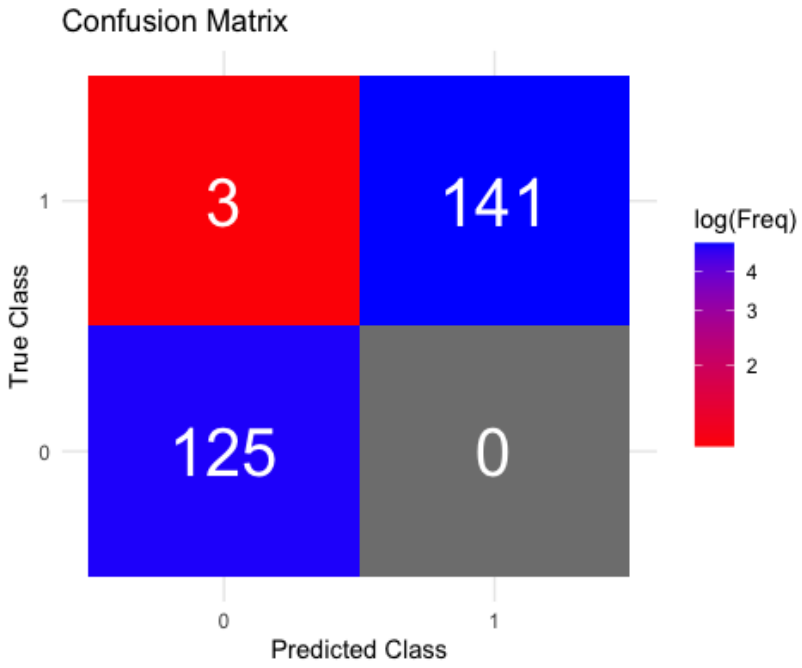
Logistic Regression Confusion Matrix



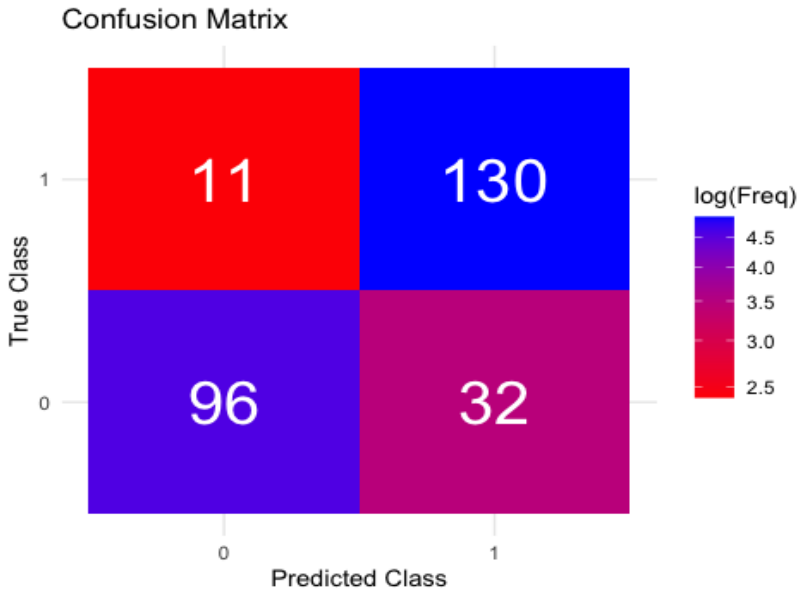
KNN Confusion Matrix



Random Forest Classification Confusion Matrix



Decision Trees Classification Confusion Matrix



Research Question 2:

Do patients with heart disease have significantly higher mean blood pressure compared to patients without heart disease?

To address this question, we performed a t-test to compare the mean blood pressure between patients with and without heart disease. The results of the t-test are as follows:

- t-Statistic: -4.465215
- p-value: 8.922492e-06

The t-test results indicate a statistically significant difference in mean blood pressure between the two groups, with patients without heart disease having slightly higher blood pressure on average. The calculated Cohen's d effect size was approximately -0.2800787, falling within the small effect size range.

DISCUSSION

Research Question 1

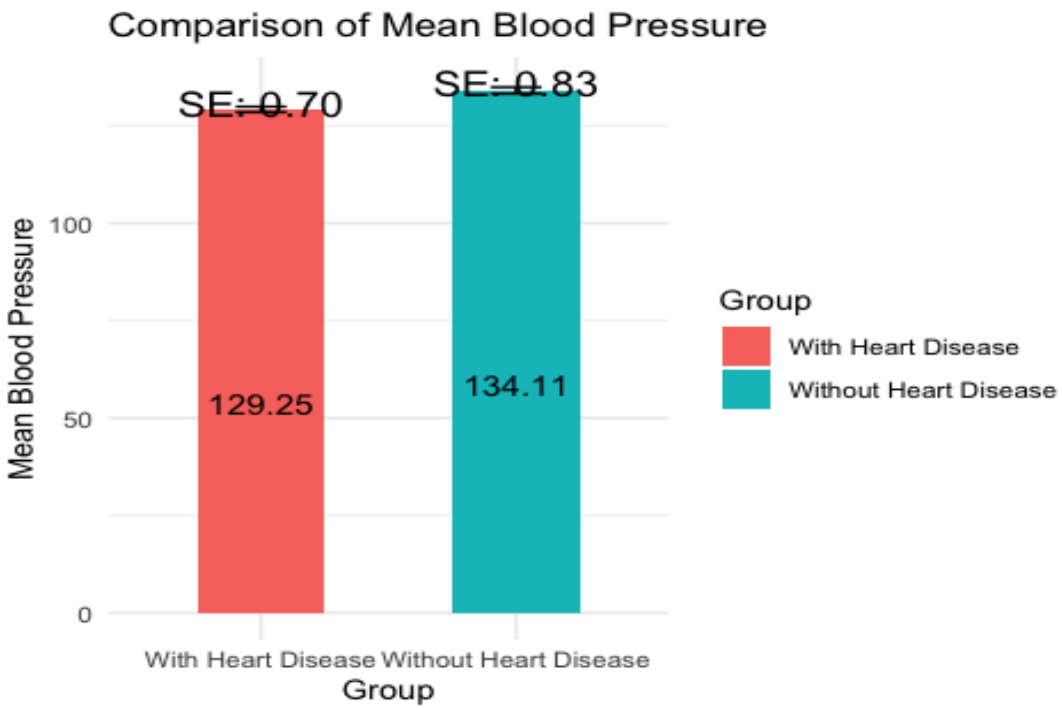
The analysis of the classification methods revealed that Random Forest Classification achieved the highest accuracy (0.99) in predicting the presence of heart disease, outperforming Logistic Regression, KNN, and Decision Trees Classification. These results suggest that Random Forest Classification is a robust method for this specific task.

Research Question 2

The t-test results provided strong evidence of a significant difference in mean blood pressure between patients with and without heart disease. The small effect size (Cohen's d = -0.2800787) suggests that the difference, while statistically significant, is relatively small.

VISUALIZATION

We presented the mean and standard error for each group using a bar plot, visually demonstrating the difference in mean blood pressure between patients with and without heart disease. The absence of overlap in the error bars emphasizes the significant difference.



CONCLUSION

In conclusion, our analysis indicates that Random Forest Classification is the most accurate method for predicting the presence of heart disease in individuals. Additionally, patients with heart disease have a statistically significant but relatively small difference in mean blood pressure compared to those without heart disease. These findings have important implications for healthcare and may aid in early detection and management of heart disease.

RECOMMENDATIONS

Further research can explore the use of other machine learning algorithms and additional features for heart disease prediction. Additionally, examining the clinical implications of the observed difference in blood pressure can provide valuable insights for medical practitioners.