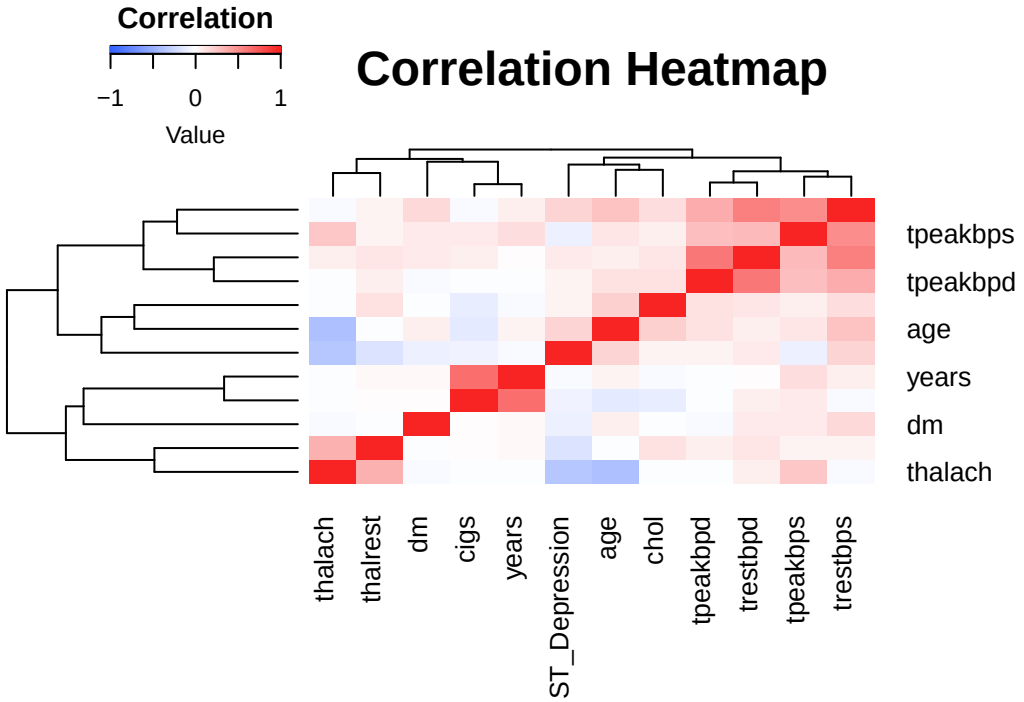
Final Paper: Heart Disease

Aditya Krishna, Thomas Huo, Varun Divi, Pranay Rana

May 02, 2024

INTRODUCTION

As people navigate their everyday lives, their heart is hard at work. However, the very organ that supports the core of everyone's existence is also the leading cause of death across the United States. Aside from biological factors that impacts heart health, smoking is a prominent external factor that leads to heart disease. Our research allows a better understanding of the multifaceted nature of heart disease and offers the potential for earlier interventions before it becomes too late. Ultimately, our research can shed some light on how to broaden patient care and systemic improvement toward precision medicine and evidence-based healthcare practices. Our study leverages the Heart Disease dataset sourced from the Cleveland Clinic in Cleveland, Ohio, available through the University of California Irvine's dataset repository. The data was derived from clinical, noninvasive research trials on patients undergoing angiography at the clinic. Angiography is a medical imaging technique that focuses on visualizing the arteries, veins, and heart chambers by injecting a contrasting agent that the X-rays used will read (NHS). Within the data, there are two different settings that the patients were asked to do. One subset of the data contained observations from regular angiography and tracking, whereas the second subset contained observations where the patients were put through trials of different exercises before conducting the angiography. Therefore, the data was cleaned to only contain values pertaining to the non-exercise-induced statistics.



To explore the potential relationship between all the continuous variables, a correlation heatmap was used. This heatmap encompasses the entire set of continuous variables within the dataset, providing a visual

representation that accentuates the correlations between these variables. Ranging from strongly negative to strongly positive, the correlations are vividly illustrated. Specifically, the variable `trestbps`, which denotes the resting blood pressure of the patient, exhibits varying degrees of correlation with each variable across the map. Notably, the strength of these correlations intensifies as we traverse further to the right and upper portions of the map, while it diminishes in a negative direction towards the left and lower regions.

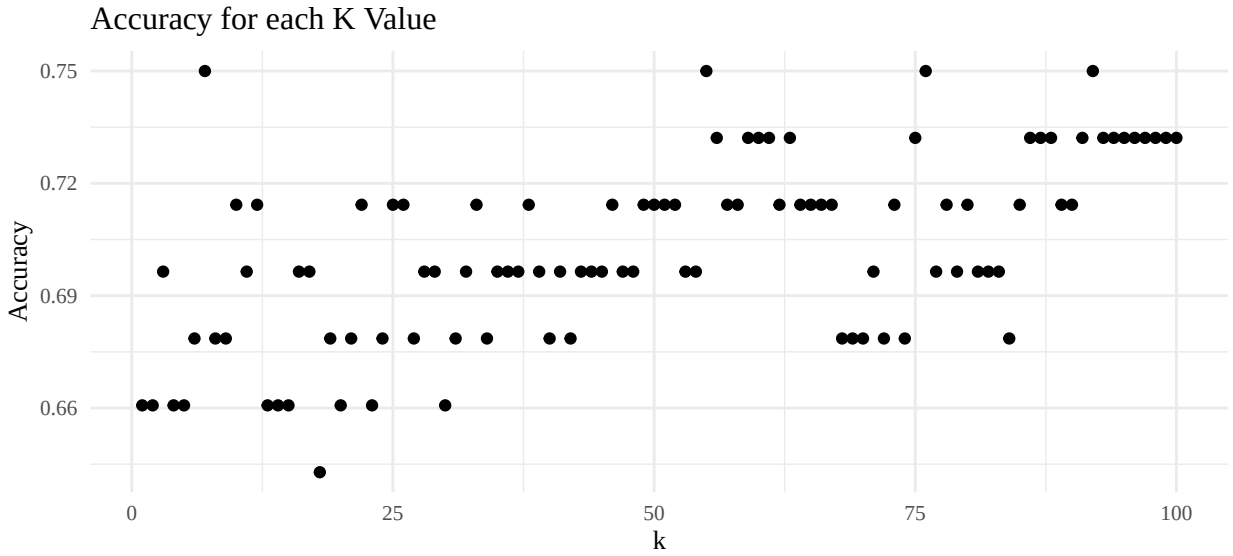
RESULTS

Question 1: Identifying the Optimal Variables for Predicting Heart Disease in Patients

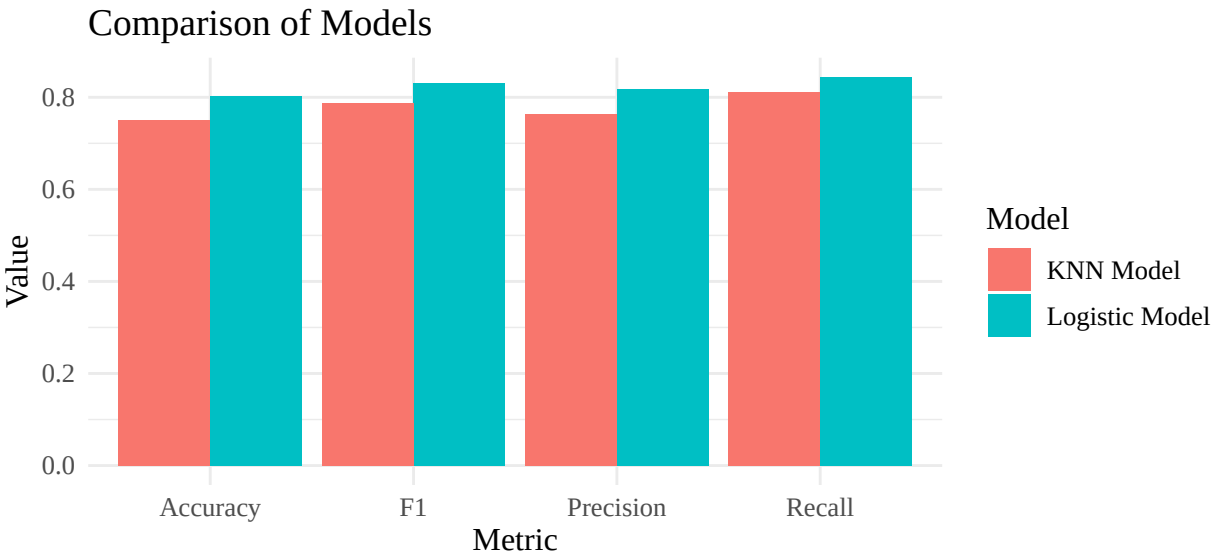
The first approach we took was to use a fairly straightforward Logistic Regression model on the data to first find the most significant variables, and then using those variables to predict and find results. The code we used is shown below:

First, in order to train the model, we split the data between a training set and a testing set, which was later used to find the results. We fitted a logistic regression model using the training data (glm() function), specifying the formula "presence ~" to predict the presence of heart disease based on all other variables. To find the most significant variables in predicting heart disease, we created a stepwise model using the step() function. From there, ran the predict() function on the stepwise model to create predictions, and evaluated the performance of the model using a confusion matrix to calculate metrics such as accuracy, precision, recall, and F1-score.

When coming up with goals for this data, we decided to focus on a more general task and a more specific task. Our more specific task involved performing multi-class classification to determine exactly which stage of heart disease was present. Our more general goal was to create a model that could simply detect whether a patient had heart disease, turning the problem in a binary classification. KNN models are very effective approaches to creating a binary classification because of their ability to be hypertuned easily. The essence of the KNN model is as follows: if the majority of the k nearest neighbors belongs to class 1, the new data point will be predicted as class 1, and the same for class 0.



We first scaled the data because it is important to have a proportional distance metric. We then created a loop for hyperparameter tuning. The code loops over possible values of k from 0 to 100 to find the optimal number of neighbors that would yield the highest accuracy. Once the k value was found, we used the knn() function on the training data set in preparation for predicting our test set. For each iteration, a simple accuracy metric was used as a measure of selectability. Accuracy is the proportion of correct predictions. Once the optimal k is found, we use one final knn() function and evaluate predictions. We use accuracy, recall, precision, and F1 score as our performance metrics.



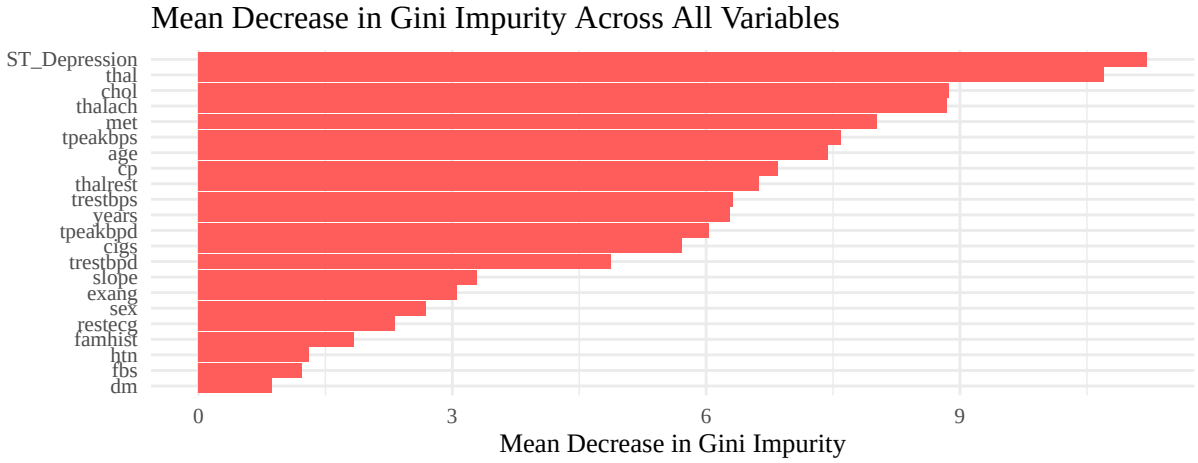
Comparing our Logistics model to our KNN model we see that on all aspects of metrics our Logistics models outperforms the KNN model. As seen in the graph, our KNN F1 score is 0.76 and our Logistic F1 score is 0.85, meaning that we still have 2 decently performing models since it is above 0.7 which is the general rule of thumb for F1 scores. However, all of these metrics showcase why the logistics model is superior to the KNN model and is better at predicting the presence of Heart Disease within patients.

Question 2: Exploring Key Factors Influencing the Accuracy of Predicting Heart Disease Stages in Patients

As mentioned earlier, we had a more specific goal of performing a multi-class classification to determine exactly which stage of heart disease the patient is in.

We split the dataset into 70% training and 30% testing. We used the randomForest Packages and specified stage as our response variable while all the other columns on the training data were used as predictors. We convert the output into an importance matrix using gini impurity for measuring the purity of a node. Our decision tree uses ensemble methods meaning that it combines the results of multiple decision trees to improve accuracy and avoid overfitting.

The accuracy for our random forest model is 0.61. Kappa is a metric designed to analyze the inter-rater reliability for categorical variables. Our random forest model kappa was 0.32 which indicates that our model was "fair" in terms of inter-rater reliability and could be improved.



Based on the figure above we see that the top four variables for predicting whether or not a patient has heart disease are `ST_Depression`, `thal`, `chol`, and `thalach`. Due to these variables all having lower gini impurities

within the random forest model compared to the ordinal classification model. While we see that the variables `dm`, `fbs`, `htn`, `famhist`, and `restecg` gini impurities in the random forest model are relatively similar to the ordinal classification model and therefore are not as important for predicting heart disease within a patient. We can therefore answer our question using the Mean Decrease in Gini Impurities because as shown by our model the best predictors for predicting Heart Disease are `ST_Depression`, `thal chol`, and `thalach`. While the variables `dm`, `fbs`, `htn`, `famhist`, and `restecg` are some of the worst predictors for predicting Heart Disease.

CONCLUSION

In our study, we sought to identify key factors that contribute to the onset and progression of heart disease. Using metrics such as logistic regression and random forest models, we analyzed a myriad of variables derived from real-world patient data ranging from demographic information to medical indicators. Our findings revealed significant predictors of heart disease such as sex, chest pain type, cholesterol levels, diabetes, and specific ECG readings like ST depression and maximum heart rate. These results, while aligning with already understood medical understandings, also offered new insights into heart disease, specifically the relationships between various factors and heart disease.

The real-world implications of our findings are significant. For healthcare professionals, these insights can provide a more refined lens to view and evaluate a patient's heart disease risk which can, in turn, enhance diagnostic accuracy and improve patient care. These findings can also lead to more informed public health strategies that can be used in preventative measures. Understanding the interconnectedness of these factors that contribute to heart disease can allow for more effective measures. For example, emphasizing the increased risk in men and stressing the importance of reporting their heart pain could lead to earlier diagnoses and better patient outcomes. This knowledge can also guide resource allocation, which can ensure that funds and efforts are directed toward the most impactful treatments. Beyond healthcare and public policy, the study's implications can broadened to be impact society as a whole. By reducing the incidence and improving the management of heart disease there can be an overall impact on humanity. Communities with better heart health can experience a reduction in healthcare costs, increased productivity, and overall improved quality of life.

References

Angiography. (2017, October 19). nhs.uk. https://www.nhs.uk/conditions/angiography/#:~:text=Angiography%20is%20a%20type%20of,doctor%20to%20see%20any%20problems

Diabetes. (n.d.). Cleveland Clinic. https://my.clevelandclinic.org/health/diseases/7104-diabetes

Heart disease facts. (2023, May 16). Centers for Disease Control and Prevention. https://www.cdc.gov/heartdisease/facts.htm

Loaiza, S. (2020, March 23). Gini impurity measure. Medium. https://towardsdatascience.com/gini-impurity-measure-dbd3878ead33

Janosi,Andras, Steinbrunn,William, Pfisterer,Matthias, and Detrano,Robert. (1988). Heart Disease. UCI Machine Learning Repository. https://doi.org/10.24432/C52P4X.