Heart attack Prediction

Pari Katyal

Introduction

In this project we will predict heart disease and find key variables that can lead to early prevention, detection and personalized healthcare. The data used for this project was collected from Kaggle and includes 14 variables and 303 observations. The variables included in this data set are: *age*, *sex*, *cp* (Chest pain type ranging from 0 to 3) ,*trtbps* (Resting blood pressure), *chol* (cholesterol in mg/dl), *fbs* (Fasting blood sugar) > 120 mg/dl (1 = true; 0 = false), *restecg* (Resting electrocardiographic results (values 0,1,2)), *thalachh* (Maximum heart rate achieved), *exng* (Exercise-induced angina (1 = yes; 0 = no)), *oldpeak* (ST depression induced by exercise relative to rest), *slp* (slope of the peak exercise ST segment), *caa* (Number of major vessels (0-3) colored by fluoroscopy), *thall (*Thalium stress test result (3 = normal; 6 = fixed defect; 7 = reversible defect)), and lastly *output* which is the predicted attribute - a diagnosis of heart disease with 1 = heart attack and 0 = no heart attack.

Something notable about this dataset is that the variable output is relatively balanced making it good for modeling and leaves us without worry of skewed results. There were no missing values, and all data types were appropriate for their variables as well. Lastly, the dataset has notable medical test results that could lead us to create many insightful models.

Question #1: Can we predict the likelihood of a heart attack based on variables like age, sex, chest pain type, cholesterol levels, and maximum heart rate achieved? How are the variables most correlated to heart attack related to each other?

My aim with this question was to learn more about my dataset and the relationships between both the variables and the outcome itself, see if predicting heart attacks would be possible and if so, how accurate it would be, and lastly, to set up a basic understanding of my data for the next questions.

Methods

My first step after data preprocessing was to create a correlation plot that shows how linearly correlated each variable is to each other and the outcome on a -1 to 1 scale. If a variable is positively correlated with the outcome, then as the variable increases, so does the likelihood of the outcome, and vice versa. Second, the predictive models I created included random forest and logistic regression, and logistic regression outperformed by ~ 1%. I wanted to create a feature importance graph so I continued further with RF. A random forest model is built up of many decision trees, each of which is trained on samples of data. Each tree in RF makes decisions by splitting data based on features values and has the goal of separating data into groups that are all the same class. Figure 2 shows an ROC curve which is the predictive capability of our model and helps us understand the ability of our model in terms of classification. It is built on the numbers shown in Figure 3. As I mentioned above a feature importance graph is important for my analysis so I can know what features may be most important in predicting the outcome and helps me understand non-linear relationships my variables may have with the outcome.

Results

As we can see in Figure 1, Thaalach (maximum heart rate achieved), chest pain and slp (slope of the peak exercise ST segment) are most positively correlated to the outcome. But, they are also highly correlated with each other which means there is multicollinearity and this can be tricky when looking at the correlation between outcome and variables because we don't know what the actual driving variable is. As we can see the variables with the highest amount of negative correlation are exng and oldpeak.

Figure 2 shows that the random forest model had an AUC of 92% which means that our model is good at classifying between true and false positives as a perfect score would equal 100%. This is vital in a medical context because accuracy is key in order to reach the goal of early prevention or detection. In figure 3 we have the confusion matrix which shows our true positive, negative, false positive and negative. As we can see we have a nice balanced confusion matrix again indicating that the model does a good job in classification for both sides. These numbers are also used to calculate accuracy, precision, sensitivity and f-1 score. Our accuracy was ~ 83.5% meaning that the Random Forest model correctly predicted whether a heart attack occurred in approximately 83.5 out of 100 cases. Precision at 84.37 lets us know about: the proportion of positive identifications that were actually correct. As mentioned this is extremely crucial in medical predictions because it means that there are few instances where a heart attack is predicted, but the patient does not actually have one. Sensitivity was also 84.37 and shows us the proportion of actual positives that were identified

correctly. A recall of 84.38% means that the model correctly identified approximately 84% of all patients who actually had heart attacks. Lastly, the F1 score which is the mean of precision was at 84.37, a pretty high number and tells us that the model has a good balance between precision and recall and class distribution is balanced.

Lastly, we have figure 4 which is a graph of features and their importance in predicting outcome. What was interesting to me was that oldpeak was very high in feature importance, but had a negative correlation with outcome. This is why we can not solely rely on linear relationships because although the two don't have one, oldpeak is still very important to predicting our output. Our other two most important variables were thaalach, which had high positive correlation with outcome, and caa which had negative correlation with output.

Overall, all our models indicate very clearly that we are able to predict if one will have a heart attack with 84% accuracy. Our metrics also let us know our model is very balances and has a strong ability to identify patients who are likely to have a heart attack, while minimizing the number of false alerts and missed cases. And in terms of our question about relationships between variables, through our graphs we were able to find linear and nonlinear relationships which give is better understanding of our data and can help greatly in early detection and prevention

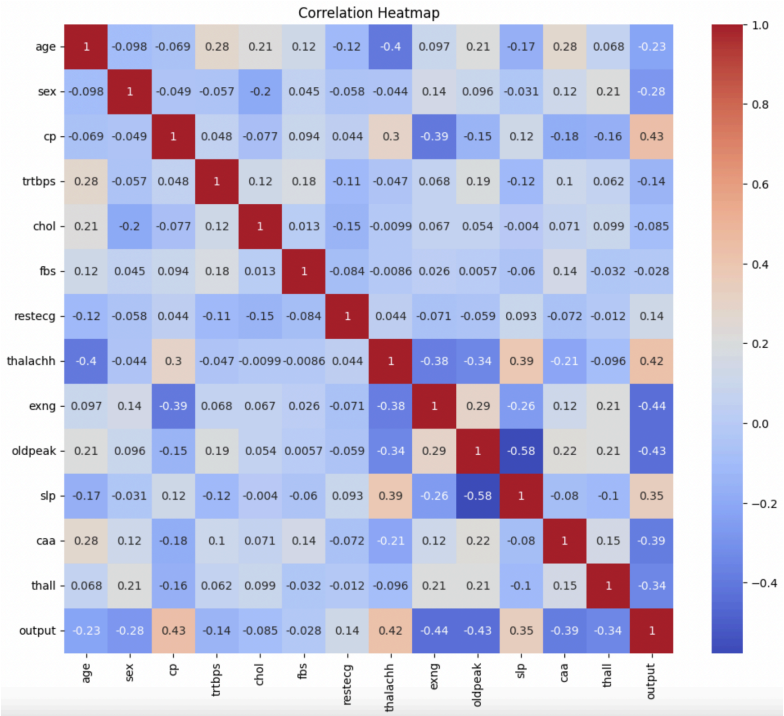

Figure 1: Correlation Plot

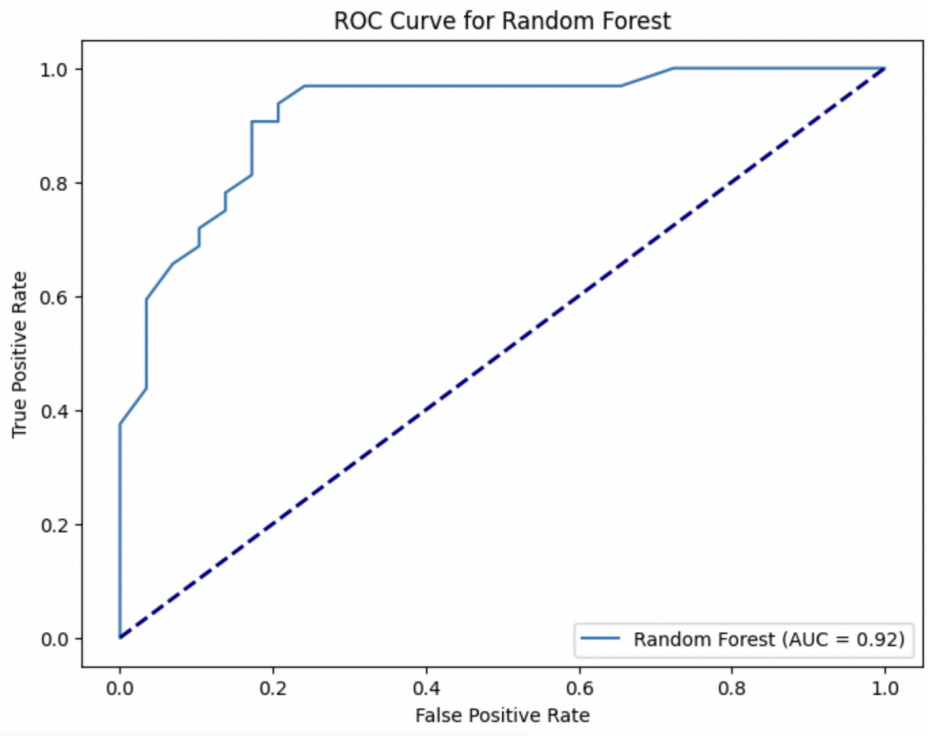

Figure 2: Roc curve

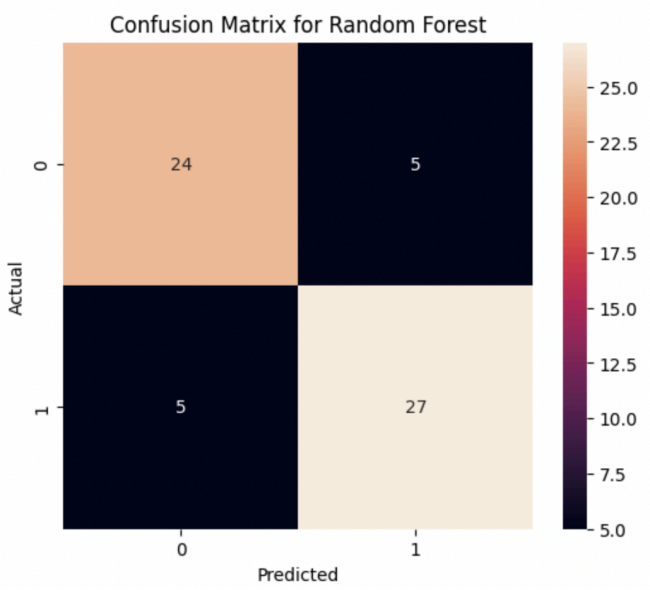

Figure 3: Confusion Matrix

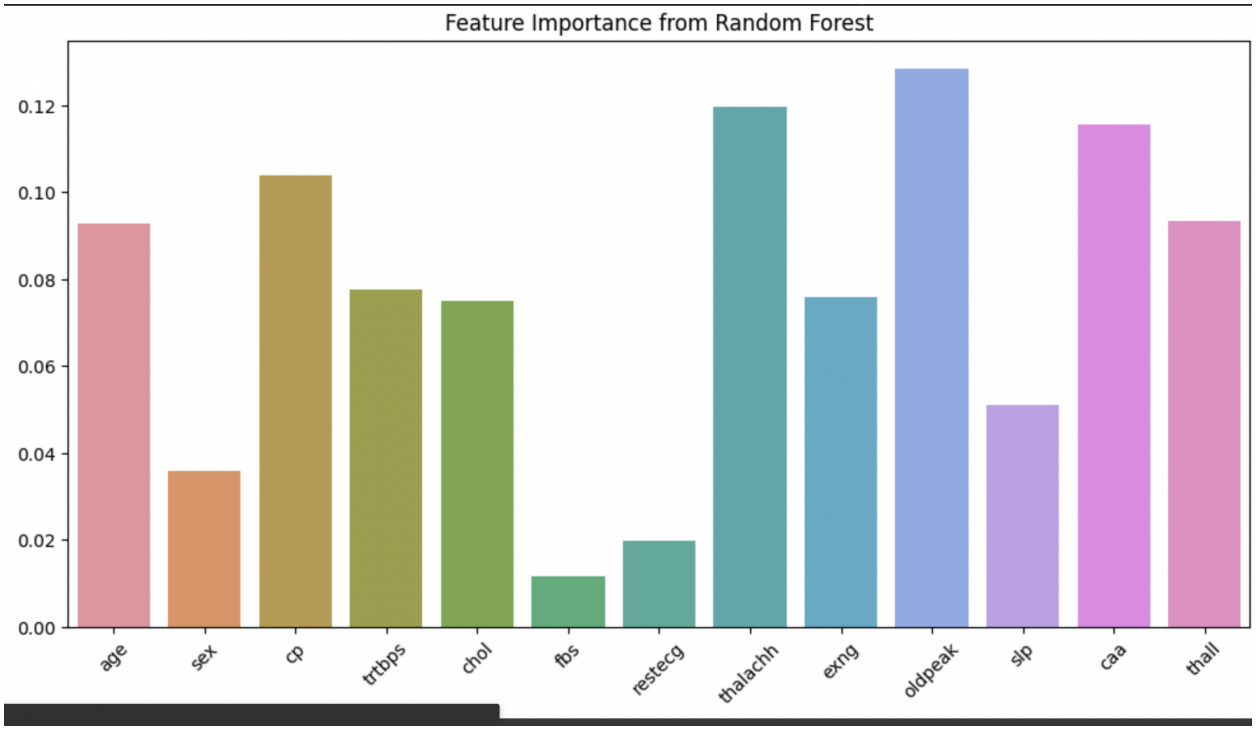

Figure 4: Feature importance

Question #2: Question 2: Can we identify distinct clusters based on demographics or health attributes? What characterizes each cluster in terms of heart health risk factors?

My reason for this question was I believe clustering can be invaluable for a number of reasons in the medical field. Some of which include: putting patients in subgroups which can lead to personalized care, risk assessment for these subgroups, effective medicine for these subgroups. By identifying hidden patterns in groups of people we are able to see what the reasons behind diseases may be and therefore be able to give patients with those attributes personalized effective care, early.

Methods

To start off I normalized my data and this is a preprocessing step to ensure that each feature contributes equally to the distance computations in clustering and without it we can get biased clustering results. My first step was using the Elbow Method to determine the optimal number of clusters. This works by fitting the KMeans algorithm with a range of cluster counts and plotting the within-cluster sum of squares against the number of clusters. When inertia decreases at a slower rate and creates a bent elbow shape, we know it is a good number to

use as our optimal number of clusters as shown in Figure 2. Figure 1 shows the silhouette score and this is a measure of how similar an object is to its own cluster compared to other clusters. A high silhouette score indicates that objects are well matched to their own clusters and poorly matched to neighboring clusters. A high score indicates the optimal number of clusters. Lastly, to find the optimal number of clusters I used a Dendrogram which is when one uses Hierarchical clustering to build a hierarchy of clusters and the dendrogram visualizes this hierarchy. The dendrogram shows how each cluster is made by drawing links and the height of it shows the distance between that cluster and the smaller ones. It can also help in determining the number of clusters by looking for the longest vertical lines that are not crossed by extended horizontal lines.

My next step was using KMeans, a clustering algorithm that partitions the data into k clusters by minimizing the variance within each cluster.. The code assigns each data point to the nearest cluster center. Once this happens we are able to calculate the mean of the features for each cluster, which lets us know how each cluster differs by variable. Lastly I put a Scatter Plot of oldpeak and cp because they were the most intriguing based on our last question. The colors represent different clusters, providing a visual representation of how the clusters differ according to these two features.

Results

The dendrogram is a hierarchical clustering visualization that helps determine the number of clusters that can naturally be formed based on the data. In the dendrogram, there is a noticeable gap between the second and third level from the top, and this suggests a natural division of the data into two or three clusters. Next I looked at the Elbow Method which is a plot of inertia against the number of clusters used to determine the optimal number of clusters for k-means clustering. The elbow is the point where the graph bends, and indicates the optimal number of clusters which in this case is 3 or 4. Lastly, I looked at the Silhouette Scores which measure how similar an object is to its own cluster compared to other clusters. This model suggested 2 was the best. Since our methods are contradicting a bit I decided the optimal number would be an average of the three which in turn gave me 3 clusters shown in Figure 4.

Now if we look at Figure 4 we see the clusters and their characteristics. To start off, Cluster 0 - has the highest `thalachh` value on average, which is the maximum heart rate. This cluster has an average age of approximately 55.6 years and a slightly higher proportion of males. This group also has the lowest average chest pain type value and average resting blood pressure

(trtbps) is around 131.6 mm Hg with an average cholesterol level that is the highest among the clusters at approximately 238.8 mg/dL. Other important characteristics to note include that exercise-induced angina (exng) is present in about 34% of patients and The average ST depression is low at 1.03. With all of these factors, with the rest being average, the likelihood of a heart attack is 0.53.

Cluster 1 has the youngest average age at approximately 50.2 years and has the highest average chest pain type value at around 1.28, indicating more severe pain types. The maximum heart rate achieved is the highest at 165.6 bpm, which may indicate better cardiovascular fitness. Oldpeak is at .85 which could let us know patients may have exercise-related cardiac issues. Also, the slope of the peak exercise ST segment is steeper at 1.55 and this group had the highest output at 0.70.

Lastly for Cluster 2 we have an average age of approximately 56.8 years and highest average cholesterol levels at around 305.5 mg/dL, a number significantly higher than the other clusters. Oldpeak is highest at 1.25 which means that there is a much higher likelihood of heart disease because we know this variable is important. The output is the lowest at 0.40.

These clusters represent different patient profiles at risk for heart disease, with the most high risk group being cluster 1. Knowing the output with those variables is high, medical professionals are able to go straight to the source of all the things that could increase likelihood and help with prevention, again signaling the importance of finding patterns in data especially in the medical field.

Figure 5 shows us the distribution of the dataset across two dimensions: oldpeak and cp. Since these two had high feature importance I wanted to confirm that they would drive distinct groups. As we can see on the scatter plot the clustering algorithm has definitely found patterns in the dataset that separate individuals into distinct groups based on these two features which lets professionals know that these variables are ones to look out for when looking at heart disease.

In conclusion these clusters can help in understanding patient subgroups and this analysis can uncover patterns that might not be apparent when considering the population as a whole. By understanding these clusters, healthcare providers can develop targeted strategies for prevention and treatment, and improve patient outcomes.

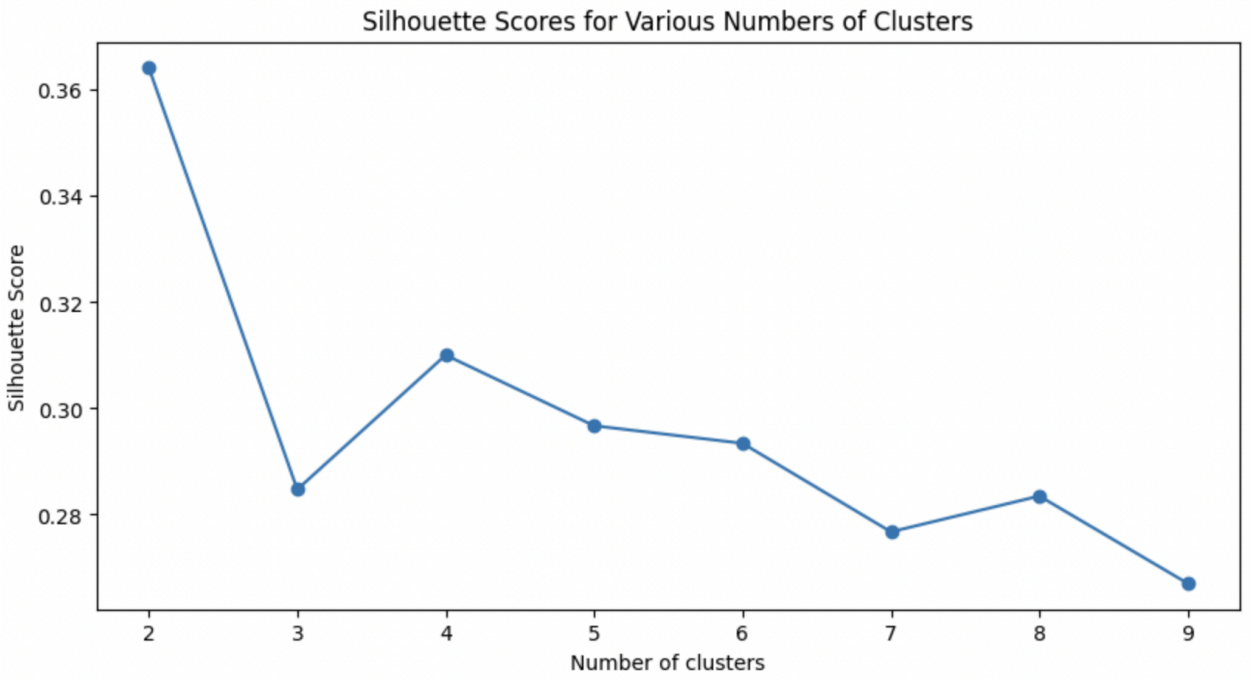

Figure 1: Silhouette Score

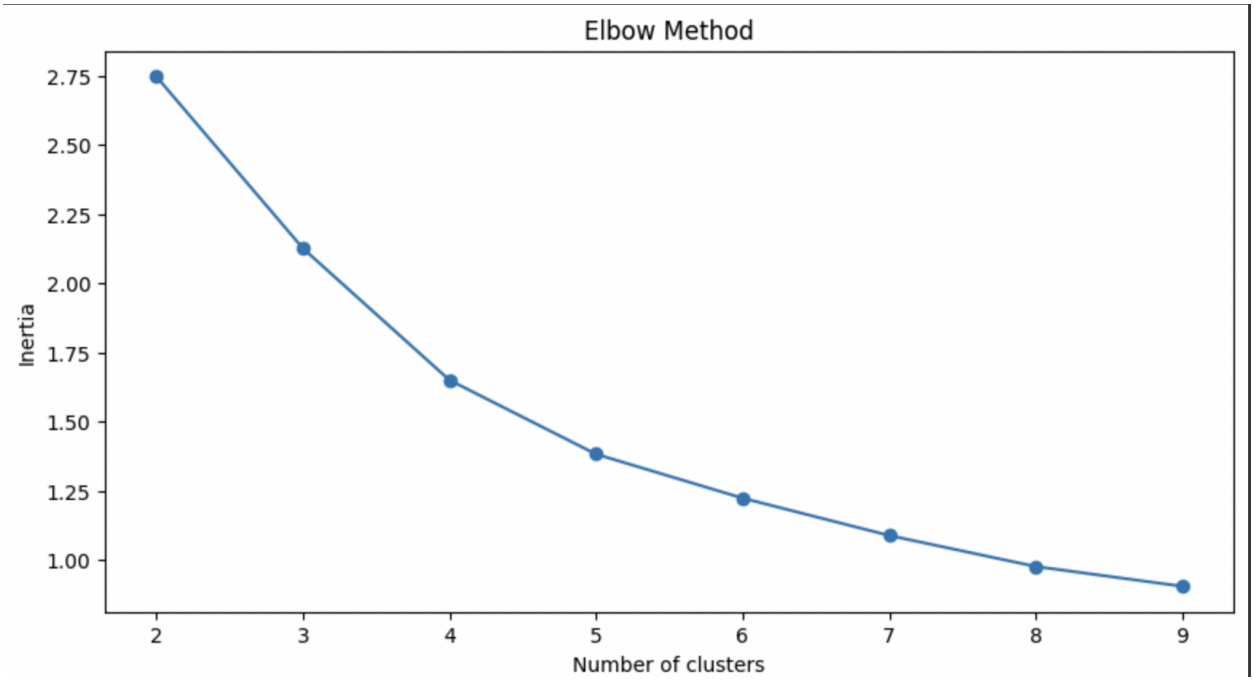

Figure 2: Elbow Method

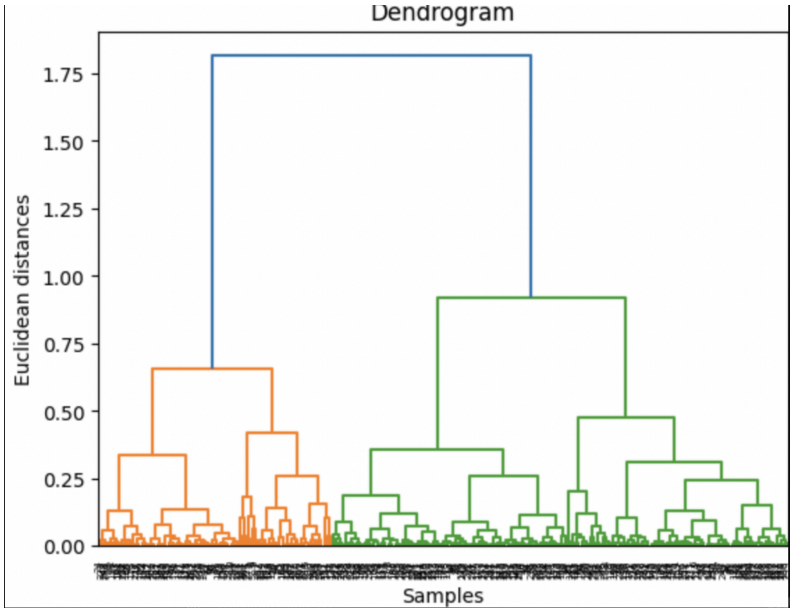

Figure 3: Dendrogram

Cluster	age	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng	oldpeak	slp	caa	thall	output
0	55.574627	0.694030	0.955224	131.626866	238.835821	0.171642	0.492537	146.723881	0.343284	1.033582	1.320896	0.723881	2.291045	0.529851
1	50.172414	0.747126	1.275862	132.574713	201.839080	0.137931	0.632184	165.551724	0.172414	0.849425	1.551724	0.563218	2.275862	0.701149
2	56.841463	0.597561	0.658537	130.609756	305.536585	0.121951	0.475610	137.548780	0.463415	1.251220	1.365854	0.914634	2.390244	0.402439

Figure 4: Clusters

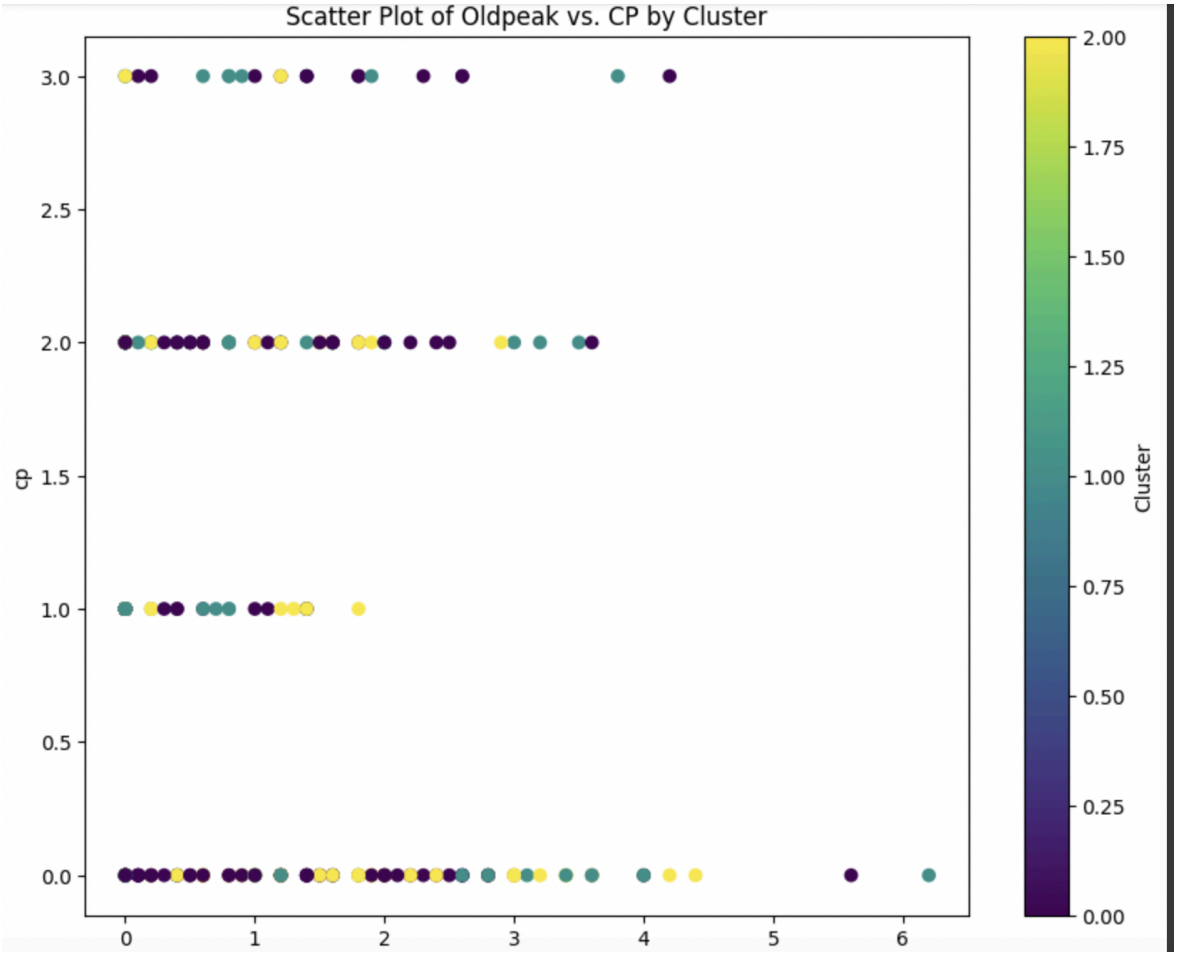

Figure 5: Scatter Plot

Question #3: What is the effect of dimensionality reduction on our ability to predict heart attacks? Which reduced features are most informative for this prediction?

The reason I asked this question is to : 1. Find out what variables were most important when predicting heart disease ; 2. Simplifying and creating an easier model for professionals. Many medical professionals don't have the time to ask about all the variables in our model so by reducing the number we need we can provide efficient health care that is realistic.

Methods

In this question I performed dimensionality reduction using Principal Component Analysis followed by predictive modeling with Gradient Boosting and Random Forest classifiers to evaluate the ability to predict heart attacks from a dataset. To start off for data preparation I

created X and Y which was variables vs output. For PCA, a process that decomposes the matrix into vectors that then represent the directions of maximum variance in the data. Basically this model lets us know how much variance is in each PCA which is done through explaining the variance ratio in the next step. Once I had this model I created a scree plot shown in Figure 1. This is a bar chart visualizing the variance for each principal component. It lets us know how many components to retain when we need 95%. Once I had this I reduced the model to just significant components, those that explained 95% of the total variance. I used both a Gradient Boosting Classifier and a Random Forest Classifier to predict outcomes based on significant variables.

The scatter plot was made to visualize the two PC of the data to show how well separated data points are after reduction as we can see in Figure 2. Lastly we have the feature importance graph once again to show us the importance of variables for prediction in order. In Figure 3.

Results

To start off the scree plot is a representation of the variance explained by each principal component extracted from PCA. This plot lets us know how many principal components to retain for further analysis. When the variance starts to level off, we know the optimal number of components to keep. In my scree plot, the first few components explain a significant amount of variance, with a sharp drop after the first component and a gradual leveling off after the second. This lets us know that most of the variance in my data was captured by the first two. The first principal component explains approximately 74.76% of the variance. The second principal component accounts for about 15.04% with the third component explaining around 8.46%.

The GBC and RFC model was composed of the three most important components which were thaalach, oldpeak and caa. Random forest achieved an accuracy of 69% while GBC achieved 65% just using 3 components. Precision is .69 and sensitivity is .66 which shows a good balance between the two and is relatively high for just 3 components. The scatter plot as mentioned visualizes the data created by the first two principal components from PCA. The distribution of points can give an idea of how well-separated the data is when reduced to two dimensions. As we can see in Figure 2 since there is a lot of overlap, we know that PCA captures varinave, but it is not enough to separate the two outcomes and we need more variables for better classification. Looking at this scatter plot explains why even if we used

three variables that made up most of our variance, our accuracy and the rest of our metrics were still low. It teaches an important lesson in the differentiation between varinave and classification and how those two can not be seen as one.

Overall through our results we were able to answer our question and realize that with dimensionality reduction we can still predict heart rates but at much lower accuracy rates. But in medical cases when fill history isn't available, research like this lets medical professionals know out of the 13 variables, what three they need to have about their patients for a decent chance of predicting heart disease. Our most informative features were thaalach, oldpeak and caa, something that has been reaffirmed through all of our different questions.
In summary we were able to come up with a simplified model that can still provide some insight and prediction. We were able to find our key variables and increase the focus around these variables.

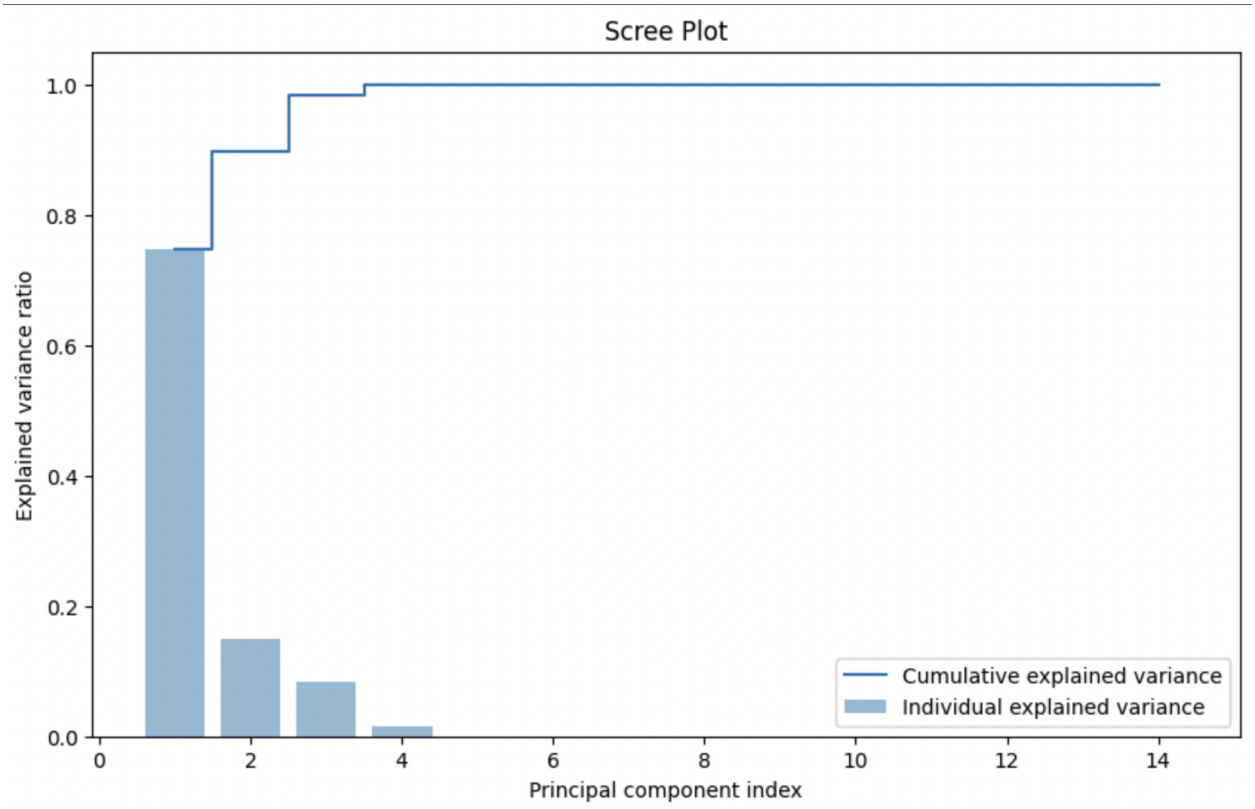

Figure 1: Scree Plot

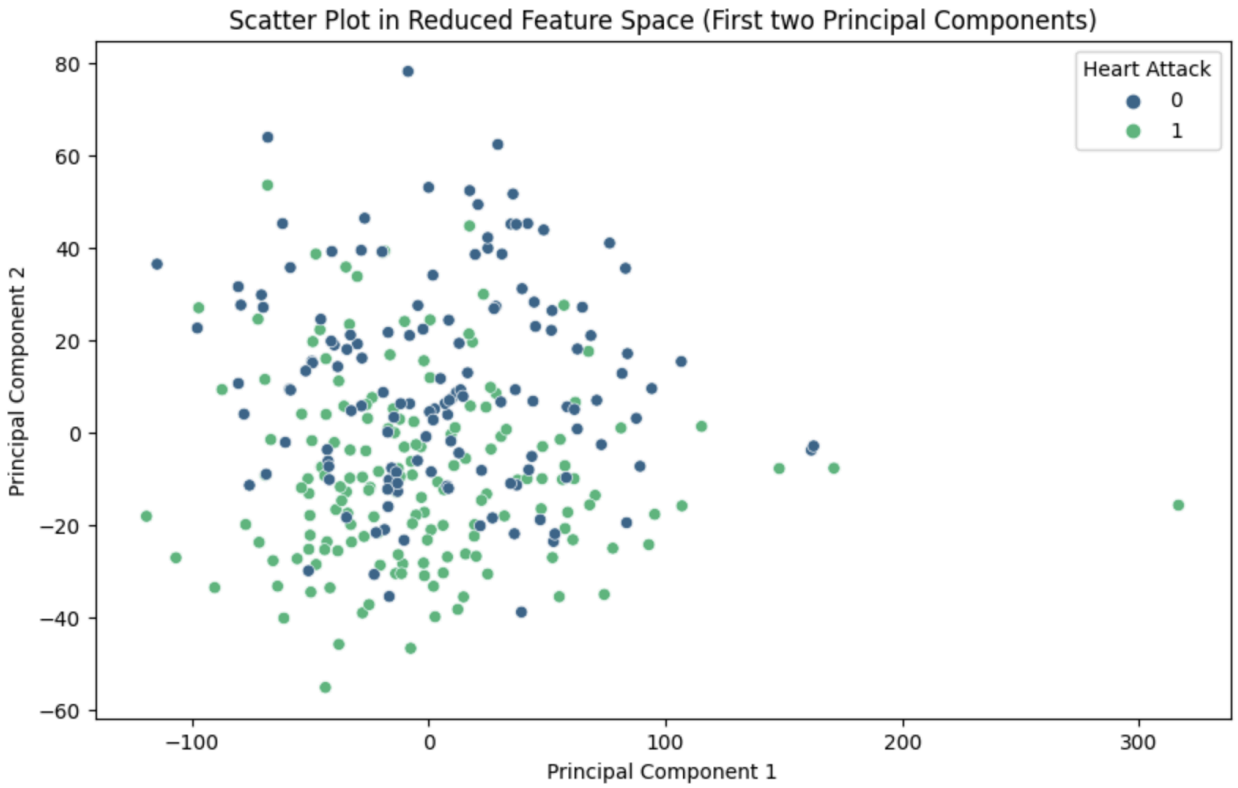

Figure 2: Scatter plot

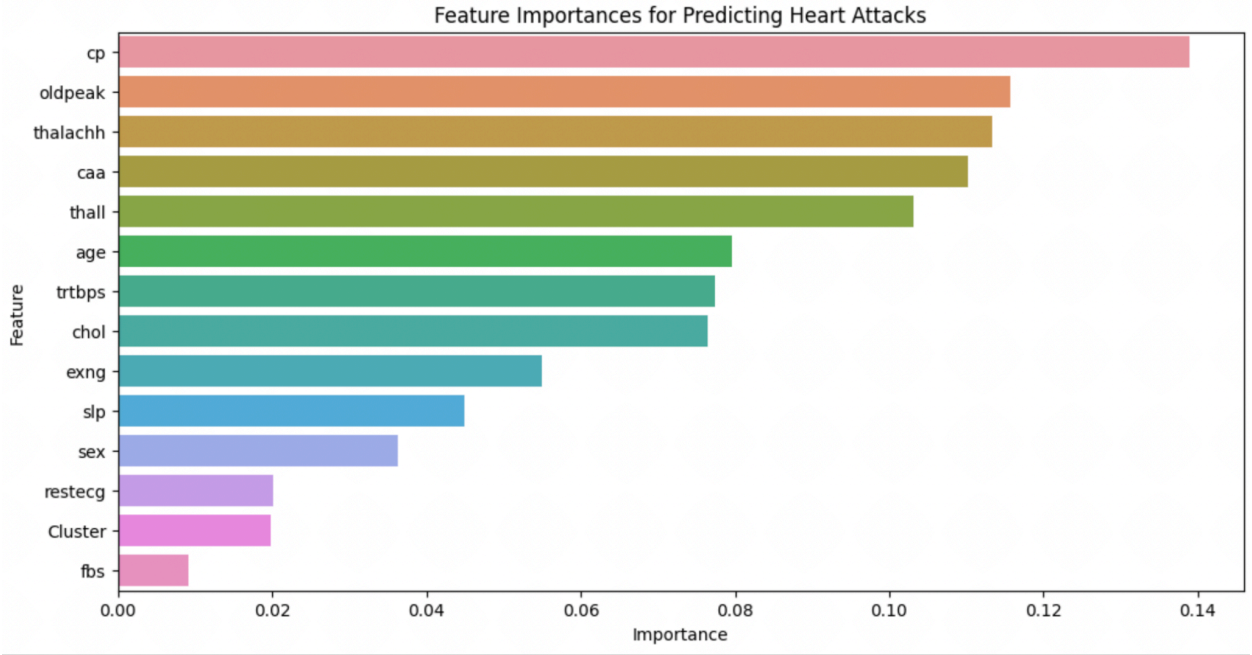

Figure 2: Feature Importance

Conclusion & Takeaways

Through this project I was able to answer three vital questions regarding heart attack prediction through machine learning. The first question focused on determining the predictability of heart attacks by using our variables and understanding variable relationships. Notably, the random forest model exhibited an AUC of 92%, affirming its ability to classify true positive and negative heart attack cases. The second question focused on clustering techniques to find distinct patterns within the data based on demographics and health attributes. We ended up with three distinct clusters, each with unique characteristics. These allow for personalized healthcare and risk assessment and with the application of our clusters there can be targeted intervention strategies. The third question focused on dimensionality reduction. We used Principal Component Analysis to find our three most important variables. We were able to highlight critical variables for heart attack prediction, providing medical professionals with ways of early detection and prevention, things that are pivotal in improving patient outcomes. This project shows how machine learning is vital in all industries and needs to be taken into consideration.