Nolan Barry, Ryan Green, Eliza Villaroman
Frank Armour
ITEC-321-001
8 December 2022

Predicting Heart Disease

The Dataset

The dataset we chose is called Heart Failure Prediction Dataset with which we found on Kaggle. It is a csv file that contains data on heart disease and some characteristics of people such as vitals like resting blood pressure, and general information such as age and gender. There are 918 observations and 12 variables. These variables are: age, sex, chest pain type, resting blood pressure, cholesterol, fasting blood sugar, resting ECG, max heart rate, exercise angina, oldpeak, ST slope, and heart disease. This data set has a mix of categorical and numeric variables. Additionally, there are some binary variables such as heart disease, where a 1 equates to having heart disease and 0 equates to not having it. Most of the variables are straight forward as their name sounds. Some that could be further explained are resting ECG, exercise angina, oldpeak, ST slope, and chest pain type. Resting ECG is the resting electrocardiogram results. There are three options for this variable: normal, ST which is having ST-T wave abnormality, or LVH which is showing probable or definite left ventricular hypertrophy by Estes' criteria. Exercise angina is a categorical variable of yes or no if the observation has chest angina while exercising. Oldpeak is a numeric measurement of heart activity depression. ST slope if a categorical variable describing if the slope of the ST segment is upward, downward, or flat. Additionally, the variable chest pain type is a categorical variable describing the observation type of chest pain, such as typical angina, atypical angina, and more. This dataset did not require any additional cleaning on our group's end.

We picked this data set primarily for three main reasons. There first was that it had a good number of observations to work with. Ideally, the dataset would have more, but the one that was chosen was still satisfactory. Secondly, our group found that the heart disease data set had clear variables that could be predicted and be used to predict a variable by machine learning processes. Lastly, we decided to pick this dataset because it has some real world value. Heart disease is one of the most common lethal diseases. This makes it so important for specialists to be able to catch it early.

The Business Case and Hypothesis:

Business Question

Predict whether or not a patient will have heart disease with the risk factors or variables provided. Analyze the results from the perspective of different stakeholders.

What are we Predicting?

How accurately can we predict heart failure with the variables provided, do any variables provide better predictive power than others. Which predictive model provided the best overall accuracy and which model best fits what we are trying to answer within the business question. Are they the same Model? The reason these models might not be the same depends on who we would be providing analysis for. If we were consulting for a heart specialist we would most likely use the model with the highest class recall for having heart failure. This is because the opportunity cost of predicting heart failure when the patient is not actually at risk is much lower than if we were to predict the inverse. Inversely, if we were to consult for a healthcare provider we would most likely want the model with a higher class precision and overall accuracy as we would be looking for a model that correctly predicts whether or not heart disease happens in order to predict what potential procedure/costs that might have to provide to the selected customer.

Hypothesis

We believe that the more complex the data model the better it will perform with our data set. We also believe that all variables will have predictive capabilities as they are all defined as risk factors.

Method Approach

Main Model and Process Documents from Data Subprocess

Figure 1

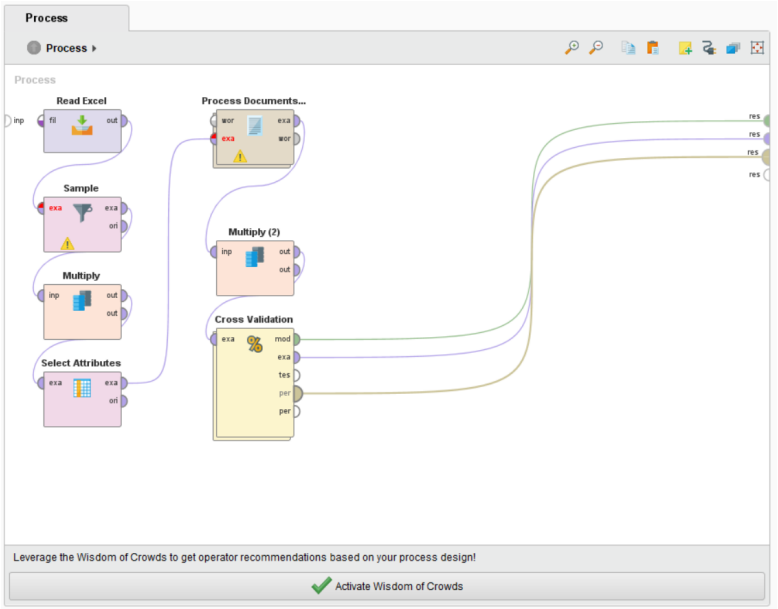
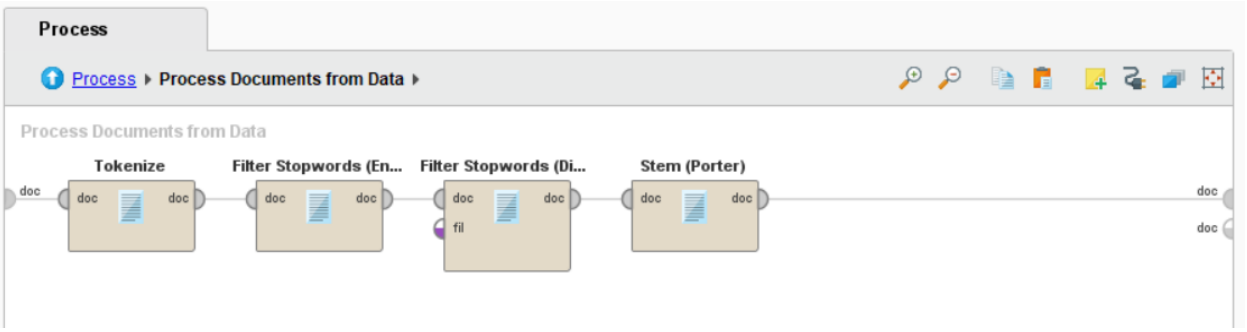


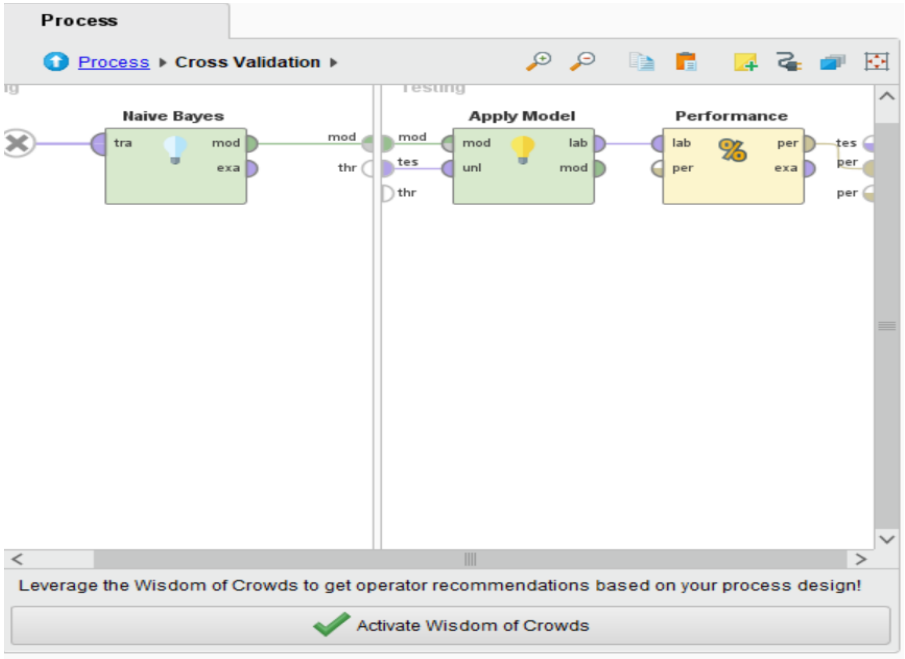
Figure 2



In our general model, we had to import and read in the dataset into RapidMiner. We set the binary variable of "heart disease" as the label since we aim to predict the likelihood of someone having heart disease. Once the data was imported and the label had been set, we set the sample size to 100. We used the 'select attributes' operator to aid us in choosing the variables to make our predictions. Then, we used the 'process documents from data' operator to tokenize nonletters, stem, and filter any stopwords in our data, so that any text data is cleaned and will output more meaningful results. Finally, we cross validated in order to check the accuracy of the performance of our models as seen in the next screenshot.

Cross Validation Subprocess

Figure 3



The Analysis Method

Finally, we chose to record a summary of our results in a classification matrix in order to compare accuracies, class recalls, and class precisions for the various types of models, variables and parameters we experimented with.

Results & Experimentation

In order to answer the question we set out to analyze, we first began by determining the best prediction model to use. This is because to determine the risk factors that are the best predictors of heart disease, the best way to do this is by determining this in the lens of the best model. By creating this, it should yield the best predictors as well. We also wanted to do this in order to build the best model to predict heart disease in general. To do this, we experimented with multiple different prediction models, tested them, and compared their results using cross validation. These different models included: decision trees, random forests, boosted trees, K nearest neighbors with different k parameters, naive bayes, deep learning, and logistic regression. The only changes to parameters we did in this step was for k nearest neighbors. For this, we used a k of 5, 7, and 8. We had to exclude operators, such as neural networks and support vector machines, because they did not have sufficient capabilities for the given data set due to polynomial data sets not being supported with these models. We also included logistic regression due to the nature of the variable we are predicting. Because it is binomial, it made sense to include this to see if it had any merit.

For this initial experimentation, we tested each model and compiled all the results into one large classification matrix. The results include total accuracy, class recall, and class precision which is included below. Looking at the results of this matrix, there are a couple models worth discussing. These would be logistic regression, random forests, deep learning, and naive bayes. Logistic regression had the worst total accuracy across these four models at 81%. 'Random Forests' is the next highest at 82%. This means that these two models predicted the test set correctly 81% and 82% correctly, respectively. Interestingly, although logistic regression had a lower total accuracy than random forests, it had higher class recall for predicting not having heart disease and class precision for having heart disease. This means that logistic regression was able to predict not having heart disease when the data point actually does not have heart disease at a higher rate and when it predicted having heart disease it was right more often. The last two are the strongest

Figure 4

Predicitive Model	Total Accuracy	Class Recall true 0	Class Recall true 1	Class Precision Pred 0	Class Precision Pred 1	Model Parameters
Decision Trees	77.00%	68.29%	83.05%	73.68%	79.03%	
Random Forests	82.00%	75.61%	86.44%	79.49%	83.61%	
Boosted Trees	77.00%	68.29%	83.05%	73.68%	79.03%	
K-NN	75.00%	68.29%	79.66%	70.00%	78.33%	k=5
K-NN	76.00%	70.73%	79.66%	70.73%	79.66%	k=7
K-NN	75.00%	65.85%	81.36%	71.05%	77.42%	k=8
Naïve Bayes	89.00%	85.37%	91.53%	87.50%	90.00%	
Deep Learning	86.00%	80.49%	89.83%	84.62%	86.89%	
Logistic Regression	81.00%	78.05%	83.05%	76.19%	84.48%	

models we tested. Deep learning had a total accuracy of 86% while Naive Bayes had an accuracy of 89%. Naive bayes had higher accuracies for all class recall and class precision as well. This means that this model was the best performing out of all the models that were tested. This led us to the discovery that Naive Bayes was the best model to use.

From here we had to test which variables made this model stronger or weaker. To do this we used the RapidMiner tool, 'select attributes' and tested the model using all the variables except one for each one. For example, the first model we tested was naive bayes using all variables except age. We compiled all the classification matrix results into one large matrix again and used it to see which variables had the largest decrease in total accuracy, class recall, and class precision when they weren't included. This is because if the prediction accuracy drops significantly when a variable is not included, it is clearly important for the accuracy. The results are shown below. Looking at the results, there were three variables that had some significant effect on the model when excluded. These were chest pain type, cholesterol, and ST slope. The largest change from these was chest pain type which decreased the prediction total accuracy from 89% to 83%. Additionally, it had the largest decreases to class precision and recall accuracies, with the largest one being to class precision for predicting not having heart disease which decreased by 9% when excluded from the model. The other notable changes were excluding cholesterol and ST slope both decreased accuracy from 89% to 87%. To confirm these findings were true, we then repeated the same process but only included each variable as the only predictor. For example we used select attributes to only include age as a predictor variable. This matrix is included to the right. For these results, a higher accuracy means that it is a stronger predictor. This is because it is now the only variable in the model, so if it is more accurate there is most likely a correlation between the variable and heart disease. From these results, we can see that this generally confirmed our findings except for cholesterol. When this was the only variable included, the accuracy was only 57%. Another important thing to note was although Exercise Angina was not noted as an important variable but when it is the only one included the accuracy was 73%, tied as the fourth highest accuracy. This led us to the answer that important variables are chest pain type and ST

Predicitive Model	Total Accuracy	Class Recall true 0	Class Recall true 1	Class Precision Pred 0	Class Precision Pred 1
Age	88.00%	85.37%	89.83%	85.37%	89.83%
chest pain type	83.00%	80.49%	84.75%	78.57%	86.21%
cholesterol	87.00%	85.37%	88.14%	83.33%	89.66%
ex angina	89.00%	85.37%	91.53%	87.50%	90.00%
fasting BS	89.00%	85.37%	91.53%	87.50%	90.00%
heart disease	88.00%	82.93%	91.53%	87.18%	88.52%
Sex	90.00%	85.37%	93.22%	89.74%	90.16%
max HR	88.00%	85.37%	89.83%	85.37%	89.83%
old peak	88.00%	85.37%	89.83%	85.37%	89.83%
restingBP	89.00%	85.37%	91.53%	87.50%	90.00%
resting ECG	88.00%	85.37%	89.83%	85.37%	89.83%
ST slope	87.00%	82.93%	89.83%	85.00%	88.33%

Figure 5

	Total Accuracy
age	64%
chest pain type	83%
cholesterol	57%
ex Angina	73%
fasting BS	58%
max HR	67%
Oldpeak	73%
resting BP	56%
Resting ECG	59%
Sex	68%
ST Slope	75%

Figure 6

slope. Cholesterol and Exercise Angina were not included in this answer as although they had merit either by themselves or when excluded, their accuracies decrease significantly when experimented with. This led to the conclusion that they are strong predictors in certain situations such as in combination with other variables or by itself, but not in any given situation.

Takeaways

From the results we just covered, we came to the conclusion that chest pain type is the most important risk factor for predicting if someone has heart disease. ST slope is another important factor to focus on when predicting heart disease. However, it is important to note that this is only true in a vacuum. If someone had to selectively choose which variables to collect given this goal, these should be considered. When they are the only variable, they have the best accuracy. Additionally, if someone wanted to increase their model accuracy, these two variables are the best from our data set at doing so. This is because when excluded, the accuracy of the model drops fairly significantly. If all the variables were available, these variables should be included with the others. Referring to figure 5, we can see that when all the variables are included except for sex, the prediction accuracy is the highest at 90%. It also increased accuracy for class recall and precision meaning it got better at predicting heart disease when it was heart disease and was correct more often. This means that if all the variables are available, including all of them except sex would create the highest prediction accuracy. If given more time, our next steps would be to analyze why including sex along with all other variables decreases accuracy and experiment with other variables not included in the data set and analyze if exercise angina and cholesterol are important variables because there was a degree of ambiguity here when we confirm our findings.