Assessing Heart Disease Through Classification Models

Modifications to the Data:

* Encode all categorical variables that are binary to either 0 or 1
  + “sex”, “fasting\_blood\_sugar”, “exercise\_induced\_angina”, “vessels\_colored\_by\_flourosopy”
* One-hot encode the rest
  + “chest\_pain\_type”, “rest\_ecg”, “slope”, “thalassemia”

Interpretation of Results:

A chart with different colored squares

AI-generated content may be incorrect. The model (not a neural net) that had the best accuracy and the least amount of false positives/false negatives was the Support Vector Machines model. The model had an overall accuracy of 97% when predicting whether someone had heart disease. The worst model was the Naïve Bayes model with an overall accuracy of 47%. Though the model has 0 false negatives, the model had a total of 109 false positives. This is very bad as it had a very low number of actual true negatives. See the confusion matrix below:

A chart with different colored squares

AI-generated content may be incorrect.This would be the model to avoid when trying to classify heart disease. Now, when looking at the neural network, I iterated through several different parameter types. The function chose the least loss model based on the iteration and then this was used to predict. The neural net got a final accuracy of 85%. The confusion matrix is as follows:

Though when compared to SVM, there are still many false positives and false negatives. This is alarming as, within the healthcare industry, you would want the lowest amount of these predictions. In terms of other models that performed very well were the Random Forests model and KNN. When comparing the SVM to these, Random Forests produced 4 false positives and 4 false negatives. Both are distributed evenly, but it would be better to see a lower number of false positives, as this means the model predicted someone who does not have heart disease. In healthcare, you would want a small number of people to be diagnosed with heart disease, even if they do not have it (possibly as a safe measure). When looking at the KNN model, it predicted 3 people as false positives, and 6 as false negatives. The KNN model has the lowest number of false positives, so it may be possible to select KNN as one that predicts better in the scope of the healthcare industry.

A chart with different colored squares

AI-generated content may be incorrect. Only certain parameters were adjusted in some of the models, so further parameter adjustment can be done. The KNN confusion matrix: