Classification Analysis of heart disease presence in patients
Final Project of Spring 2019 Data Mining
Jean Paul Uwimana

Introduction

The human heart is an organ that pumps blood throughout the body via the circulatory system, supplying oxygen and nutrients to the tissues and removing carbon dioxide and other wastes. As complex and paramount that sounds everyone would want to maintain a healthy lifestyle and to make sure that anything that is not going right with our heart is diagnosed as early as possible so it can be treated before it's too late. To that end, it would be nice for healthcare providers to check our clinical measurements and be able to tell us whether they've found presence of heart diseases in our system. In this project I will study a dataset that is composed of multiple clinical measurements that were collected by healthcare providers and will try to deduce whether these measurements are likely to predict the presence of heart disease in the patient.

I will use two binary classification algorithms and a decision tree algorithm to predict the presence or absence of heart disease using a small dataset from the UCI Machine Learning Repository provided via Kaggle.com. Specifically, I will use Support Vector Machine, Naïve Bayes and J48 to perform the binary classification on this data set.

Dataset Description

The dataset provided by the UCI machine Learning Repository through Kaggle is composed of 303 observations of which 165 are patients whose results were positive for heart diseases classified as 1 and 138 for those whose results were negative, classified as 0. The dataset is pretty representative and contains the following 14 attributes:

Table 1: Dataset description

Description	Abbr	Data Type
age	age	integer
sex	sex	values 0 or 1
chest pain	cp	values 0 to 4
resting blood pressure	trestbps	integer
serum cholesterol in mg/dl	chol	integer
fasting blood sugar > 120 mg/dl	fbs	integer in mg/dl
resting electrocardiographic results	restecg	values 0 to 2
maximum heart rate achieved	thalach	integer
exercise induced angina	exang	values 0 or 1
ST depression induced by exercise relative to rest	oldpeak	decimal
slope of the peak exercise ST segment	slope	values 0 to 2
number of major vessels colored by flouroscopy	ca	values 0 to 3
thallium stress test	thal	values 0 to 3
target	target	values 0 or 1

Data Pre-processing

1. No NA values
2. Since all the attributes are of numeric types, I will convert the following from integer to categorical variables:
    a. Cp
    b. Exang
    c. Slope
    d. Ca
    e. Thal
    f. Target
3. I will also rename the *age* attribute from *A…age* to *age*
4. Train and test datasets: I will randomly shuffle the data to ensure there's no bias in the order of the records, and then will assign 70% of the data to train set and 30% to test set. This will result in 212 and 91 observations for train and test datasets, respectively.
5. Finally, I will remove attributes of less importance

Information Gain

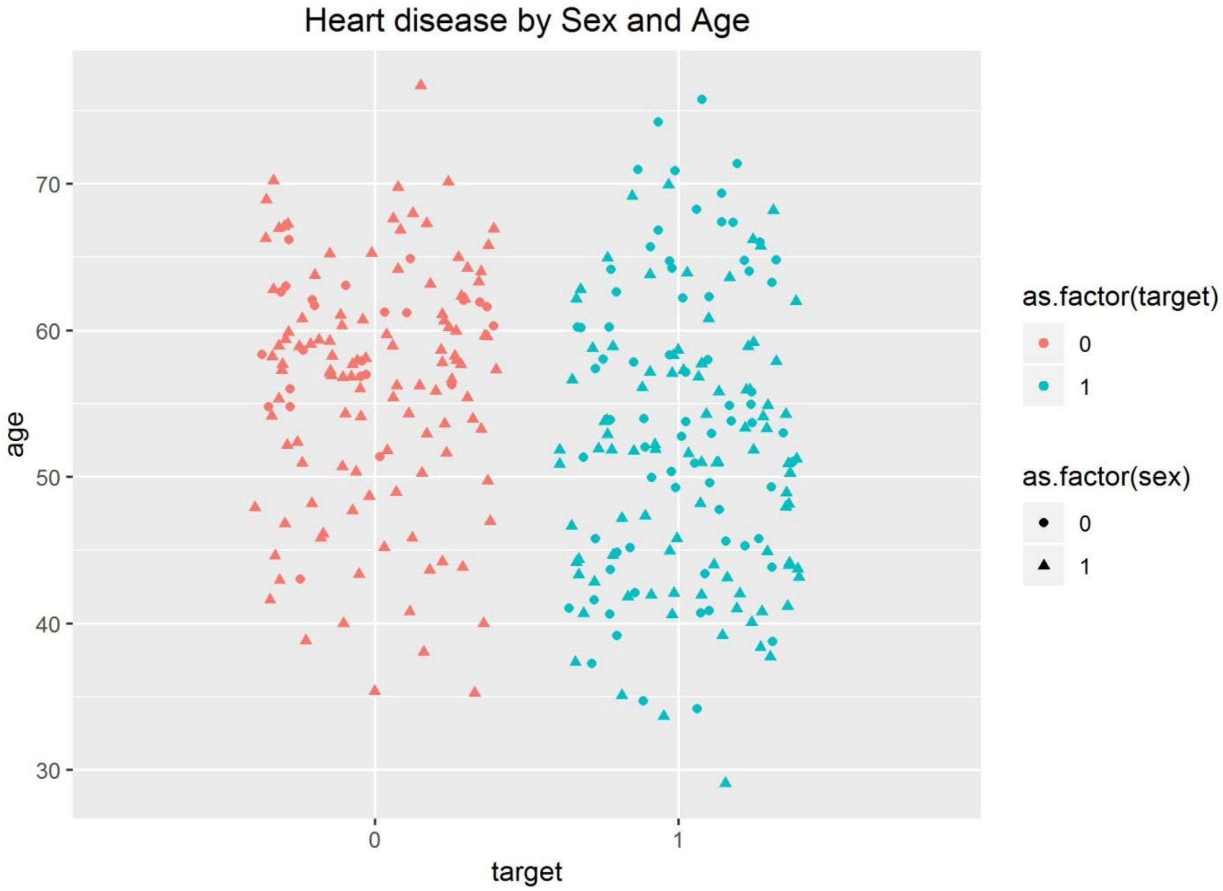
The following table illustrates how much each variable contributes to predicting our target variable. It's interesting how *age,* sex, and *cholesterol total* contribute very little to our heart disease classification model. Below is a breakdown of the information gain ratios extracted from the dataset using *gain.ratio* algorithm from *FSelector* package. These ratios were consistent with those observed from *RWeka's algorithm* for information gain analysis

Table 2: Variable Information Gain Ratios

```
> kable(FSelector::gain.ratio(target ~., data = heart))

| attr     | attr_importance|
|:--------|---------------:|
|age      |       0.0000000|
|sex      |       0.0000000|
|cp       |       0.1716812|
|trestbps |       0.0000000|
|chol     |       0.0000000|
|fbs      |       0.0000000|
|restecg  |       0.0000000|
|thalach  |       0.1047133|
|exang    |       0.1474028|
|oldpeak  |       0.1669423|
|slope    |       0.0954908|
|ca       |       0.1279407|
|thal     |       0.1861782|
```

Heart disease outcome by Sex and Age

In addition to the gain ratio analysis above, we can also visualize our data to see if we can visually observe any trends of gender or age in predicting the heart disease outcome.



One thing first, our data was not evenly distributed in terms of gender. There were 207 male observations and 96 female observations. Hence, we see more male than female data points on the graph. Second, the age, although our data age ranges between 29 and 77, it would have been very helpful in our analysis to have a dataset that is more all-age representative to see whether a particular age range might be more affected by heart diseases than others.

As we can see from the above graph, in addition to the gain ratio analysis we performed previously, there’s really no trend to be observed about *age* and *sex* in relation to heart disease.

Given the above insight from our dataset, I will only be using the following attributes for our binary classification model: *cp, thalach, exang, oldpeak, slope, ca, thal*, to predict the *target* variable*.*

Predictive Analytics

1. Support Vector Machine (SVM):

The first algorithm I will use is *SVM* which seeks to find the widest margin between 2 object classes. I will use the *caret* wrapper package which makes it easier to build models, evaluate and compare them. Additionally, since our data is not linear, I will use the *svmRadial* method within our *train* function. I will also use the following parameters: Cross Validation with k-fold = 4 and TuneLength = 10. Below are the performance results of the SVM model

Model Performance:

```
> svm_heart_model
Support Vector Machines with Radial Basis Function Kernel

212 samples
  7 predictor
  2 classes: '0', '1'

Pre-processing: centered (15), scaled (15)
Resampling: Cross-Validated (4 fold)
Summary of sample sizes: 159, 160, 159, 158
Resampling results across tuning parameters:

  C       Accuracy   Kappa
    0.25  0.8156447  0.6218109
    0.50  0.8155573  0.6209623
    1.00  0.7967767  0.5835195
    2.00  0.7870774  0.5663083
    4.00  0.7824477  0.5563349
    8.00  0.7922412  0.5718792
   16.00  0.7732758  0.5336290
   32.00  0.7639292  0.5141187
   64.00  0.7828946  0.5551756
  128.00  0.7782649  0.5447495

Tuning parameter 'sigma' was held constant at a value of 0.04494817
Accuracy was used to select the optimal model using the largest value.
The final values used for the model were sigma = 0.04494817 and C = 0.25.
```

As seen from the above model performance evaluation, the highest accuracy score achieved by our model was 81.56% at *Cost = 0.25* and *sigma = 0.04*

2. Naïve Bayes:

The second algorithm I am going to experiment with for this project is the Naïve Bayes which assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature (Ray, 2017). Again I will utilize Cross Validation with 4 folds, laplace = 1, useKernel = T.

Model Performance:

```
> nb_heart_model
Naive Bayes

212 samples
  7 predictor
  2 classes: '0', '1'

No pre-processing
Resampling: Cross-Validated (4 fold)
Summary of sample sizes: 159, 160, 159, 158
Resampling results:

  Accuracy   Kappa
  0.7926813  0.5575516

Tuning parameter 'fL' was held constant at a value of 1
Tuning parameter 'usekernel' was held constant at a value of TRUE
Tuning parameter 'adjust' was held constant at a value of 1
```

Naïve Bayes model achieved an overall accuracy of 79.26% with cross validation and k-folds at 4. I am utilizing 4 folds for all the models so that during model comparison I am comparing the algorithm with the same parameter values

3. J48 - Decision Tree:

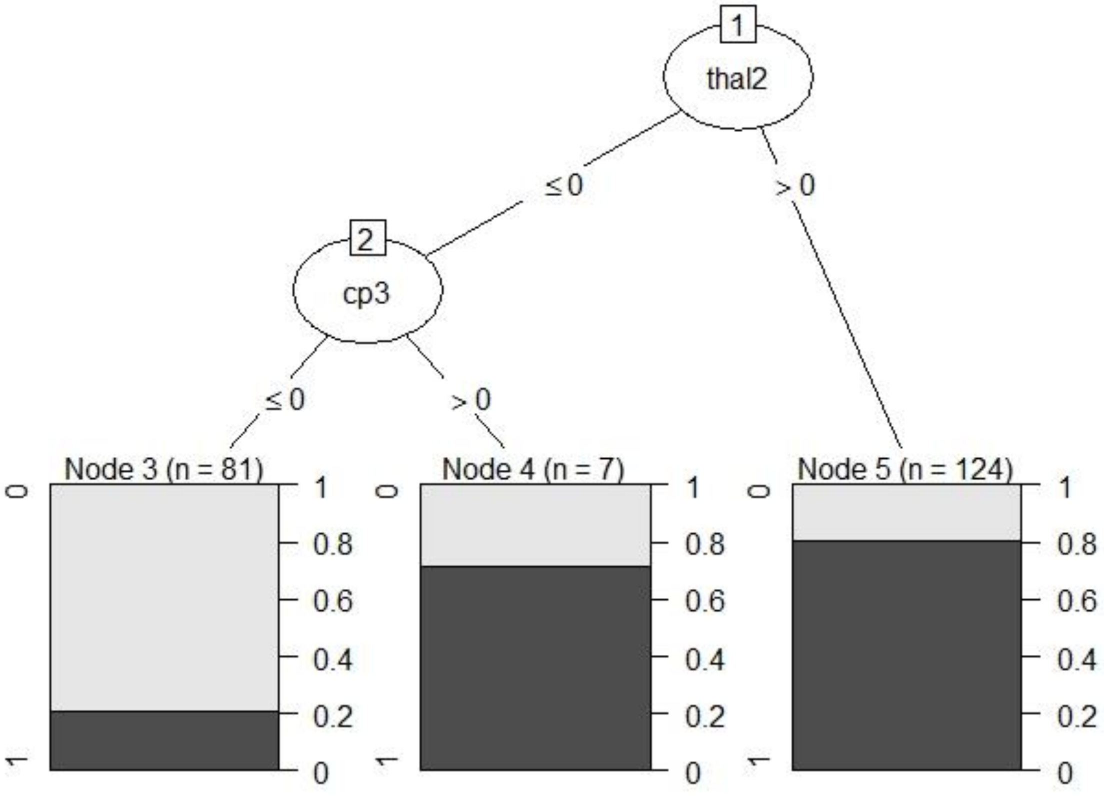
Just like we saw it earlier in the information gain ratio table, all of our algorithms show that the *thallium stress test (thal)* is the best predictor of heart disease. Therefore, it is the root node for our *J48* model and the *chest pain (cp)* is nested inside it. The J48's decision tree model achieved 76.36% accuracy

Model Performance:

```
> J48_heart_model()
C4.5-like Trees

212 samples
  7 predictor
  2 classes: '0', '1'

No pre-processing
Resampling: Cross-Validated (4 fold)
Summary of sample sizes: 159, 160, 159, 158
Resampling results:

  Accuracy   Kappa
  0.7636638  0.5173126

Tuning parameter 'C' was held constant at a value of 0.005
Tuning parameter 'M' was held constant at a value of 5
```

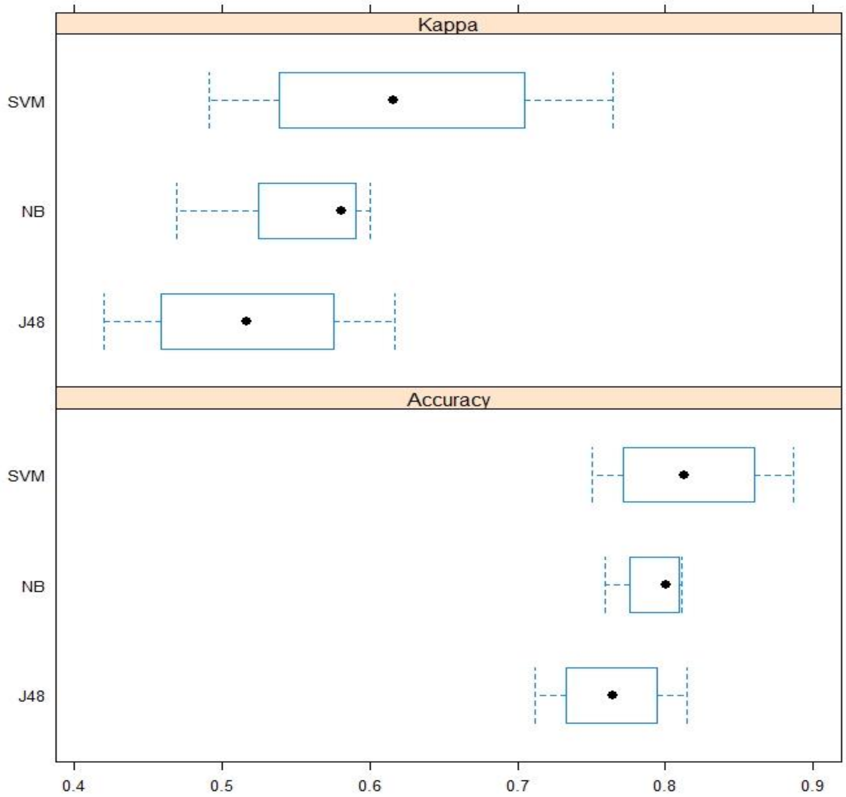
J48 Decision Tree:

```
> J48_heart_model$finalModel

J48 pruned tree
------------------

thal2 <= 0
|   cp3 <= 0: 0 (81.0/17.0)
|   cp3 > 0: 1 (7.0/2.0)
thal2 > 0: 1 (124.0/24.0)

Number of Leaves  :       3

Size of the tree :        5
```



The SVM and Naïve Bayes model accuracy scores were pretty similar with both classifiers achieving the overall accuracy score of 81% and 79%, respectively. On the other hand J48 model performance was slightly lower with accuracy of 76%. Below is the graph of how these models look graphically using resamples algorithm from caret package:

Model comparison using *resample* algorithm:



Now that we just evaluated our models, next we will see what our **prediction** performance look like using the test dataset that we preserved for this purpose.

Below is the table for all the prediction performance metrics:

Table 3: Prediction Performance for SVM and Naïve Bayes

Algorithm	Parameter Tuning	Accuracy	Kappa	Precision	Recall	F-measure
SVM	CV, 4 folds, tuneLength = 10	80.22%	60.13%	82.05%	74.42%	78.05%
Naïve Bayes	CV, 4 folds, useKernel = T laplace = 1	78.02%	55.15%	89.66%	60.47%	72.22%
J48	CV, 4 folds, tuneLength = 10, M = 5, C = 0.005	70.33%	40.26%	70.00%	65.12%	67.47%

The above was produced based on the SVM prediction results. As we can see, **from** our models are not perfect. You can observe a few data points from the *positive* class that were classified as *negative* and vice versa.

Conclusion

As we have observed in the binary classification results obtained utilizing Support Vector Machine, Naïve Bayes and J48, all these algorithms showed that given the results of the above clinical measurements, we can predict whether there's presence or absence of heart diseases in a patient with a good degree of accuracy. All the three prediction performance showed scores of 0.80, 0.78 and 0.70 for SVM, Naïve Bayes and J48, respectively. In terms of accuracy, there was no overfitting for SVM and Naïve Bayes, there was overfitting, however, for J48 as the performance on the train dataset dropped from 76% to 70% on the test dataset.

Although our dataset had 13 predictor variables in total: *age (age), sex (sex), chest pain (cp), resting blood pressure (trestbps), serum cholesterol (chol), fasting blood sugar (fbs), resting electrocardiographic results (restecg), maximum heart rate (thalach), exercise induced angina (exang), ST depression induced by exercise relative to rest (oldpeak), slope of the peak exercise ST segment (slope), number of major vessels (ca) and thallium stress test (thal),* not all of them were meaningful in prediction. And as a result only seven were used: **cp, thalach, exang, oldpeak, slope, ca,** and **thal**.

Both gain ratio analysis from *FSelector* algorithm and the decision tree model by *J48* algorithm showed that the strongest predictors for this dataset were *thallium stress test (thal), chest pain (cp)* and the *number of major vessels (ca).* No one or two predictor variables alone could produce an accuracy score close to what's reported in the above table. It seems to be a combination of all these seven variables working together to predict the heart disease presence or absence in patients.

With that being said, given the amount of observations used in this study (303), I am unable to conclude that we can predict the presence or absence of heart disease in patients with a high degree of accuracy. Given that healthcare is a very complex domain, and people's lives depend on it, we would want a model with a much better accuracy, precision and recall as we would not want a model that may lead us into treating patients for diseases they don't have or miss out on the treatment window should our model tell us there's no presence of heart disease while there actually is. I would say there's insufficient amount of observations to definitively conclude that our model will accurately predict the presence of heart disease in patients.

References

Works Cited

Kaggle. (n.d.). *Heart Disease UCI*. Retrieved June 10, 2019, from https://www.Kaggle.com: https://www.kaggle.com/ronitf/heart-disease-uci

Lewis, T. (n.d.). *Human Heart: Anatomy, Function & Facts*. Retrieved June 9, 2019, from Live Science: https://www.livescience.com/34655-human-heart.html

Ray, S. (2017). *6 Easy Steps to Learn Naive Bayes Algorithm.* https://www.analyticsvidhya.com.