Date: 3/08/2022

Name: Jack Sydenham StudentID: s3841816

Foundations of AI

Module 2B Assessment

Report on

Investigation Into Safety of CV99 Pill for Patients with Issues Related to the Heart

TABLE OF CONTENTS

1. DATA SUMMARY AND INVESTIGATION STATEMENT

The given data is said to be comprised of 4 different datasets, though there are 303 entries given and we are told that 1 of the 4 datasets (Cleveland database) contains 303 entries. So, we can assume that the data given is comprised only of the Cleveland database records, rather than 4 different datasets. We are told that there are a recorded 75 attributes relating to heart health within patients, along with a final predicted attribute pertaining to the diagnosis of heart disease in any given patient based off the recorded 75 attributes, though the data given only contains 14 attributes out of this 76 (including the predicted value of diagnosis of heart disease). This data was collected in July of 1988, though may still prove useful in today's field of health. The goal of this investigation is to classify patients as suitable to consume the newly developed CV99 pill, which is only to be taken by patients with a healthy heart. So then, based on the given data, we aim to establish an understanding of what constitutes a patient who is healthy enough to be administered the pill.

It should be understood that in all past public experiments utilizing this data, only the given 14 of the 76 attributes were taken into consideration, including the predicted value, diagnosis of heart disease. The reason for this is not specified, though it is safe to assume that the previous investigators viewed the said 14 attributes as the most significant out of the given attributes. For this reason, we can assume also that only these 14 values are needed for our investigation. The values to be used are as follows: (['column name'] Description)

1. age — Age of patient
2. sex — Gender of patient
3. cp — Chest Pain
4. trestbps — Resting Blood Pressure (mm)
5. chol — Serum Cholesterol (mg/dl)
6. fbs — Fasting Blood Pressure (> 120 mg/dl ➔ True)
7. restecg — Resting Electrocardiographic Results
8. thalach — Max Heart Rate
9. exang — Exercise Induced Angina
10. oldpeak — ST Depression Induced by Exercise Relative to Rest
11. slope — The Slope of the Peak exercise ST Segment
12. ca — Number of Major Vessels (0-3) Colored by Fluoroscopy
13. thal — Thalassemia
14. num — Diagnosis of Heart Disease (Predicted Value)

Throughout this report, columns will be referred to by their respective column name.

2. DATA WRANGLING

Here, we aim to identify any missing values or outliers in the given data and deal with them using various transformations. We will look at the data description given to determine where missing values or outliers exist.

First things first it's important to understand that our dataset only has 303 entries, making it a very small set. This means that we can safely fill missing continuous values using imputation by the mean/median/mode.

MISSING VALUES

Using Weka Explorer, we can view the dataset and find missing values of each column easily. Looking at the columns one by one, we can see that there are 2 missing values in the 'ca' column and 5 missing values in the 'thal' column. These columns are both categorical, so we may replace the missing value with the most likely input rather than applying mean/mode/median. Since we have very little missing values (2% in 'ca' and 1% in 'thal'), substituting the most frequently occurring value of each column will not be a problem and will to keep predictions accurate later.

So then, using Weka, we can see that the 'ca' column has the most frequent entry of '0.0', so we will replace the missing values of this column with '0.0'.

We can also see that the 'thal' column has the most frequent entry of 'normal', so we will replace the missing values of this column with 'normal'

OUTLIERS

Now we look for outliers. Going through the column one at a time on Weka, there are some clear outliers. The first of which is in the 'chol' column. There is a single entry with a value of 564, while the second highest value sits at 417. As this column holds continuous values, we replace this value of 564 with the mean of the column, 246. Now there is a gap between the highest four values in the column and the rest, though these values are relatively close together so we will not classify them as outliers.

The 'restecg' column contains 3 attribute values of 'left_vent_hyper', 'normal', and 'st_t_wave_abnormality'. Here, the attributes of 'left_vent_hyper' and 'normal' have 147 inputs and 152 inputs respectively, while the 'st_t_wave_abnormality' attribute only has 4 inputs. This is clearly an outlier and we will replace all values of 'st_t_wave_abnormality' with 'normal', as it is the most common value in this column.

There are two outliers in the 'oldpeak' column which we will replace with the column's rounded mean as the data is continuous.

Similar to the 'restecg' column, both the 'slope' and 'thal' columns have 3 attribute values where one of the values has significantly less inputs than the other two. However, these less inputted attributes still have around 20 inputs so they should not be regarded as outliers as they may have a noticeable impact on our prediction later on.

NORMALIZATION

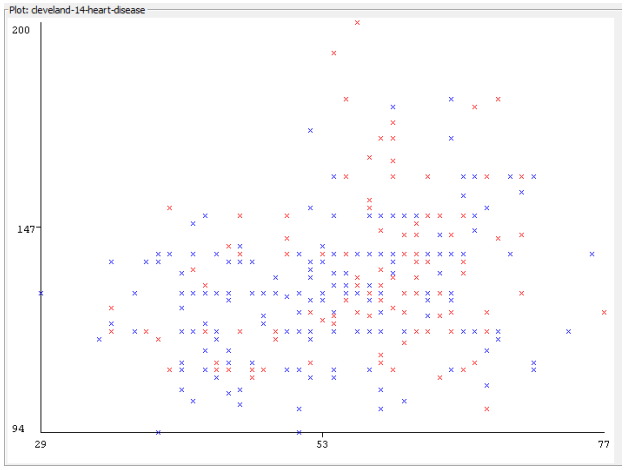
We will also apply normalization to our data, as there are many different numerical columns operating on different scales. We will normalize the given numeric attributes using Weka. Although it may change some values now – for example the mean age decreased by a very little amount after applying the normalization – we will see

more accurate results later, as the machine learning algorithm will be able to predict values more effectively if all the numerical data is measured in the same range.

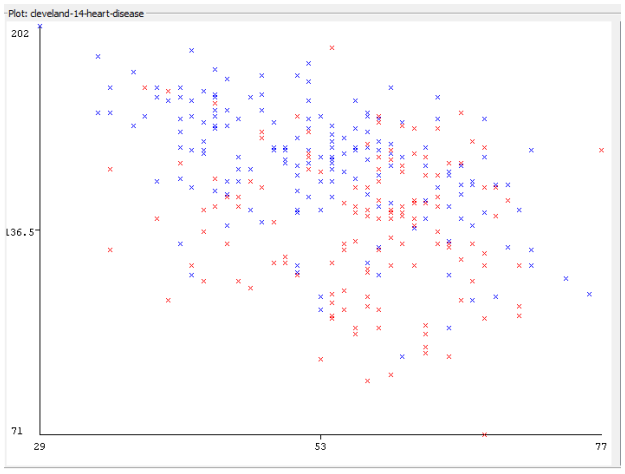
3. CHARTS REPRESENTING PATIENT HEART HEALTH

In the following charts/plots, the color of visualized inputs represents whether or not a patient was predicted to have heart disease (based off the Diagnosis of Heart Disease (Predicted Value) attribute), with red inputs representing patients that were and blue inputs representing patients that were not. The plots compare attributes to the age attribute. This way, visualization is clear, and the colored inputs allow us to easily analyze the heart disease attribute as we go through each attribute.

We will first compare some of the numeric attributes to the age.

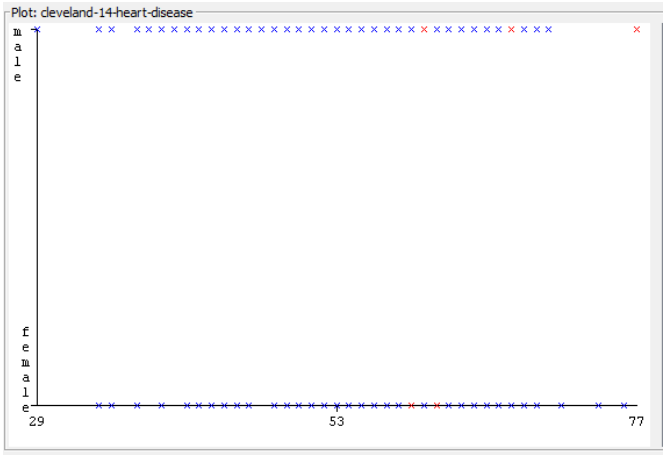


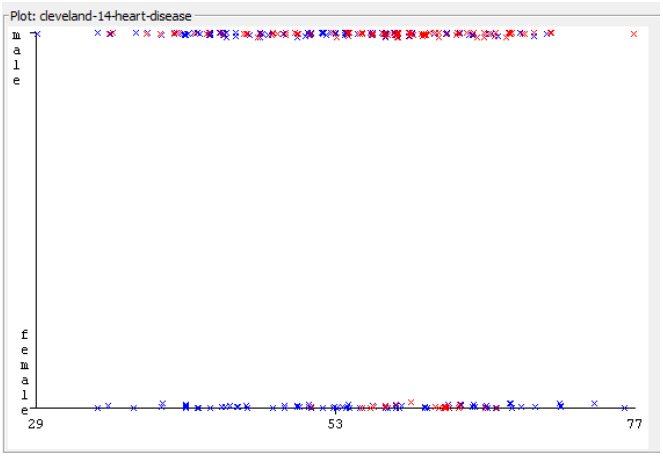
Looking at the above chart, where x = Age and y = Resting Blood Pressure of the patient, we can see clearly that between the ages of 29 - ~50, patient's resting blood pressure follows no real pattern, though as the age passes 50, we see many patients with elevated resting blood pressure. This is important to keep in mind, as an elevated resting blood pressure may be a sign or rather a result of an unhealthy heart.



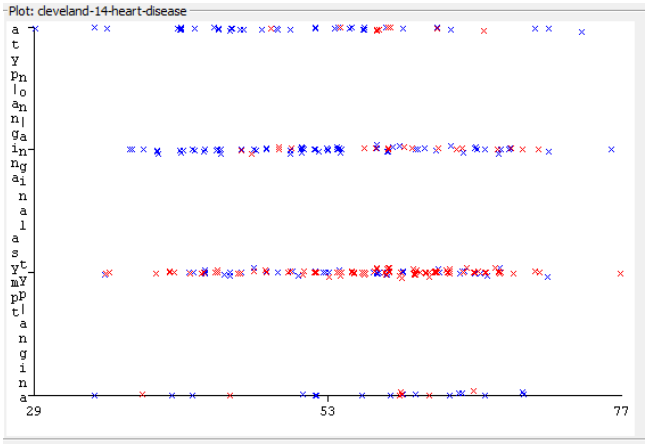
Looking now at comparing Age (x value) and Max Heart Rate (y value), we can see as age increases, the max heart rate of patients tends to decrease, with older patients having potentially very low max heart rates. It should be noted here that we see throughout all ages, the majority of patients who are predicted to have heart disease have lower heart rates than those predicted to not have heart disease. This tells us that for one, a high max heart rate may be a strong indicator of a 'healthy heart', but also that age is a very significant attribute, as older patients are more likely to have lower heart rates and in turn an unhealthier heart.

Now we compare some of the categorical attributes to the age.





The above plot shows age compared to gender (left). If we use the jitter feature in Weka to slightly separate the input points, we get a better idea of where the data lies on each category (right). While this is most likely a coincidence, we can see that interestingly, the majority of patients who were predicted to have heart disease were among male patients.



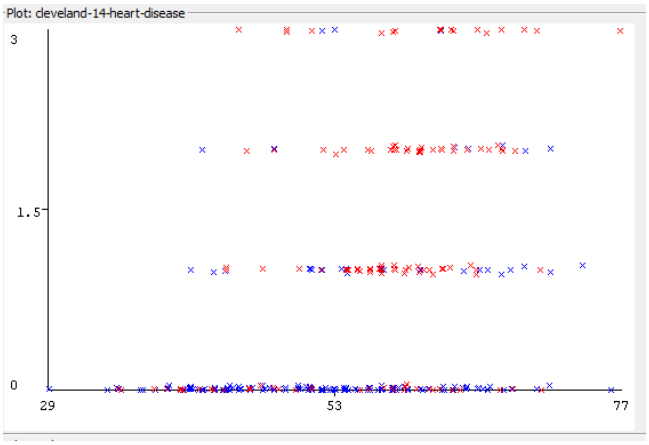
Atypical

Non-anginal

Asymptomatic

Typical

In this graph, we compare age with chest pain. We first apply the jitter to see the input groupings more easily. There are four categories of chest pain stated in the given data description document: Typical Angina, Atypical Angina, Non-anginal pain, and Asymptomatic. Quite surprisingly, we see that asymptomatic patients have far more predicted heart disease values than that of the other categories. Perhaps chest pain does not correlate strongly with heart disease likelihood.



This plot compares Number of Major Vessels (0 - 3 ascending) Colored by Fluoroscopy. We can see that patients who have 0 vessels colored tend to be predicted not to have heart disease, while patients who do have any number of vessels colored tend to be predicted to have it. We can also see that as the age increases, patients tend to have more colored vessels, with each category's median age being higher as the number of colored vessels increases.

4. EXPLANATION OF PRODUCED MODELS

The data given aims to make a prediction for the categorical attribute 'Diagnosis of Heart Disease'. Since this is indeed categorical, we will apply logistic regression modelling to predict for whether a patient has heart disease or not and in turn whether or not they can be classified as healthy enough to receive the CV99 pill. We will then evaluate the performance of the models produced using this method.

So then, we apply logistic regression on the data with the target attribute set to 'num' (the diagnosis attribute column) with a split of 80% train and 20% test.

```
=== Summary ===

Correctly Classified Instances          51               83.6066 %
Incorrectly Classified Instances        10               16.3934 %
Kappa statistic                          0.6724
Mean absolute error                      0.0886
Root mean squared error                  0.2193
Relative absolute error                 43.6641 %
Root relative squared error             68.7722 %
Total Number of Instances               61
```

The summary we receive tells us that out of the test data, we had an ~83.6% accuracy, with 51 of the total 61 test values being identified correctly.

```
=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 0.867     0.194     0.813      0.867    0.839      0.674    0.910     0.914     <50
                 0.806     0.133     0.862      0.806    0.833      0.674    0.910     0.908     >50_1
                 ?         0.000     ?          ?        ?          ?        ?         ?         >50_2
                 ?         0.000     ?          ?        ?          ?        ?         ?         >50_3
                 ?         0.000     ?          ?        ?          ?        ?         ?         >50_4
Weighted Avg.    0.836     0.163     0.838      0.836    0.836      0.674    0.910     0.911
```

The model also output a class-by-class accuracy record. We see that as in the data, there are five output variables, though we only refer to '<50' and '>50_1' corresponding with 'Does not have heart disease' and 'Does have heart disease', respectively.

The TP Rate here tells us the percentage of positives that were correctly predicted. We can see that predictions on patients with heart disease were slightly more accurate than on patients without, though this may be a result of the relatively small amount of test data.

We can also see here that we have a high precision rate, meaning that the model was able to (to an extent) correctly identify which attributes were most relevant and predict based off the most significant data.

```
=== Confusion Matrix ===

  a  b  c  d  e   <-- classified as
 26  4  0  0  0 |  a = <50
  6 25  0  0  0 |  b = >50_1
  0  0  0  0  0 |  c = >50_2
  0  0  0  0  0 |  d = >50_3
  0  0  0  0  0 |  e = >50_4
```

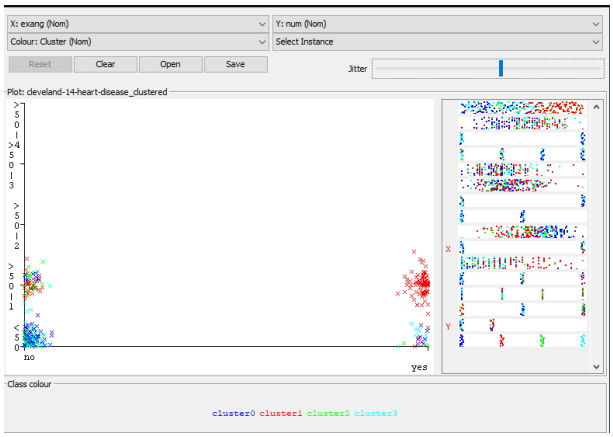
Finally, the model produced a confusion matrix based on the model. This is a great piece of information, as it more clearly identifies which output values the model had the most success with. We of course refer only to columns/rows 'a' and 'b' since the rest of the data was not measured.

We can see from value (a,a) that 26 positives were correctly predicted while (a,b) tells us that 4 negative values were predicted to be positive. Similarly, we see at (b,b) that 25 negative values were correctly identified, while at (b,a), 6 values were predicted to be negative when they were actually positive.

Now that we have evaluated the prediction model, we will see if there are any clustering tendencies within the given data, we will use the SimpleKMeans algorithm. By running the SimpleKMeans algorithm multiple times with different k values, we see that the 'within cluster sum of squared erroes' value which is output with each clustering model begins to diminish once we use any value higher than 4. Therefore we will use k=4 for our model.

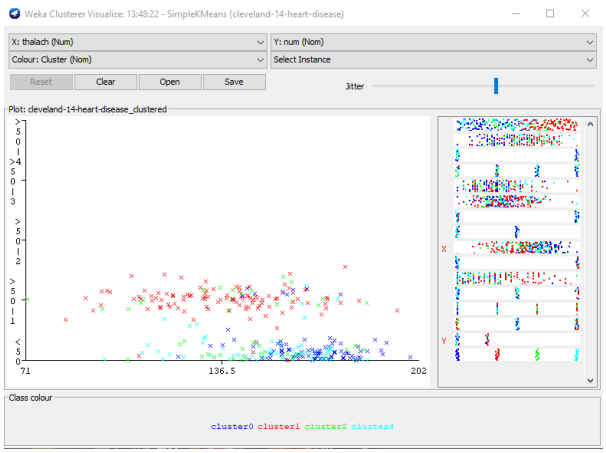
```
Final cluster centroids:
                                        Cluster#
Attribute                 Full Data             0                1                2                3
                            (303.0)          (86.0)          (108.0)           (43.0)           (66.0)
=============================================================================================================
age                         54.3663         47.1279          57.0093          56.7674          57.9091
sex                            male            male             male             male           female
cp                           asympt      non_anginal          asympt           asympt       non_anginal
trestbps                   131.6238         125.9535         135.463          129.3721          134.197
chol                       245.2739         229.5814         252.9167          231.186          262.3939
fbs                               f                f                f                f                f
restecg                      normal           normal   left_vent_hyper           normal   left_vent_hyper
thalach                    149.6469         166.3488         137.1296          143.6279          152.2879
exang                            no               no              yes               no               no
oldpeak                      1.0073           0.4547           1.6472           1.1953           0.5576
slope                            up               up             flat             flat               up
ca                           0.6634           0.1628           1.1019           0.7674           0.5303
thal                         normal           normal reversable_defect reversable_defect           normal
num                              <50              <50              >50_1              <50              <50
```

We can see from this that out of the 4 clusters created, only one of the clusters (2nd cluster) has an average of patients being diagnosed with heart disease. This cluster also contains the most patients, so there are definitely clustering tendencies within this 2nd cluster.



We can see from this graph that patients with Exercise Induced Angina ('exang' = yes) are extremely likely to fall into cluster 1 with a high chance of being diagnosed with heart disease.

We also see from the output that patients in this 2nd cluster have a lower Max Heart Rate ('thalach') than that of patients in other clusters



Looking then at the graph of 'thalach' against the predicted value of diagnosis, we see that clusters 1, 2, and 3, with higher thalassemia values are more likely to not be diagnosed with heart disease, while the points which represent a diagnosis of heart disease are comprised almost entirely of red, 2nd cluster patients, with lower max heart rates than that of other patients.

5. FINDINGS AS A RESPONSE TO THE INVESTIGATED PROBLEM

Based off the information displayed throughout this report and the findings of this investigation, we can assume that the two most significant factors towards determining whether or not a patient has a 'healthy heart' and should be administered the CV99 pill are the patient's state of exercise induced angina and their maximum heart rate. While there are countless factors which can determine a patient's health, especially related to their heart, and we cannot truly know a patient is healthy enough to consume the CV99 pill, these two significant attributes should guide doctors primarily in making informed decisions about the pill's administration and use by covid-19 patients.