DataScienceCapstoneIDV

Richa Gautam

February 1st 2020

Introduction

Liver disease in India is a growing concern. Rising rates of chronic illnesses in general and poor development of public health programs have taxed caregivers. In this scenario, being able to predict the occurence of a disease, in this case liver disease, using machine learning algorithms and comprehensive health information is becoming more and more essential.

The dataset contains a total of 583 cases, with 416 liver patients and 167 non-liver-patients, collected from North East of Andhra Pradesh, India. The "Dataset" column indicates "Liver patient" or "non-liver-patient." The data set contains 441 male patient records and 142 female patient records. Any patient whose age exceeded 89 is listed as being of age "90".

Methods

Pre-processing

Dataset was cleaned to exclude cases with missing values, leading to a final dataset of 414 liver patient records and 165 non-liver-patient records. This cleaned dataset contained 140 female patients and 439 male patients.

Columns in Dataset:

1. Age of the patient
2. Gender of the patient
3. Total Bilirubin
4. Direct Bilirubin
5. Alkaline Phosphotase
6. Alamine Aminotransferase
7. Aspartate Aminotransferase
8. Total Protiens
9. Albumin
10. Albumin and Globulin Ratio
11. Dataset

Demographic Characteristics:

```
## # A tibble: 4 x 4
## # Groups:   Gender [2]
##   Dataset           Gender SampleSize Percentage
##   <fct>             <chr>       <int>      <dbl>
## 1 Liver patient     Female         91       65  
## 2 Liver patient     Male          323       73.6
## 3 Non-liver patient Female         49       35  
## 4 Non-liver patient Male          116       26.4
```

Table 1. Percentage of liver patients and non-liver patients by gender.

Because the dataset has more men than women, we computed the incidence rate of liver disease within each gender. We noticed that rates of liver disease are higher in men compared to women by almost 9 points (73.6% vs 65%).

Machine Learning Algorithm

My plan is to analyze the dataset using an ensemble of algorithms applicable to categorical prediction. I then planned to use PCA and re-analyze the dimensionally-reduced dataset using the same ensemble.

Analysis without PCA

To begin, we used an ensemble method that combined 6 algorithms applicable to categorical prediction - svmLinear, svmRadial, svmRadialCost, svmRadialSigma, and lda.

```
##            Model  Accuracy
## 1      svmLinear 0.7137931
## 2            gbm 0.6724138
## 3      svmRadial 0.7137931
## 4  svmRadialCost 0.7137931
## 5 svmRadialSigma 0.7137931
## 6            lda 0.7206897
```

Table 2. Accuracy of each algorithm used in the ensemble.

The accuracy of the individual methods were not very high, with lda leading to the highest accuracy (0.721). The ensemble's average accuracy was 0.708046. After using the majority vote method to select predictions, the ensemble's accuracy was 0.7137931.

Analysis with PCA

While the dataset does not have a lot of variables to need dimension reduction, I applied dimension reduction to see if it improves accuracy. The ensemble of algorithms remained the same with one exception - lda was removed as PCA had been applied and there was no need for the algorithm to reduce dimensions.

```
## Importance of components:
##                             PC1      PC2      PC3      PC4     PC5     PC6
## Standard deviation     351.2416 202.1807 97.29941 16.78591 7.03067 1.65897
## Proportion of Variance   0.7088   0.2349  0.05439  0.00162 0.00028 0.00002
## Cumulative Proportion    0.7088   0.9437  0.99807  0.99969 0.99997 0.99999
##                            PC7    PC8    PC9   PC10     PC11
## Standard deviation     1.19310 0.5986 0.4233 0.1343 1.17e-15
## Proportion of Variance 0.00001 0.0000 0.0000 0.0000 0.00e+00
## Cumulative Proportion  1.00000 1.0000 1.0000 1.0000 1.00e+00
```

Table 3. Standard Deviation, Proportion of Variance and Cumulative Proportion explained by each component.

Plotting the Cumulative Proportion of variance we see that the first seven components explain all the variance in the dataset.

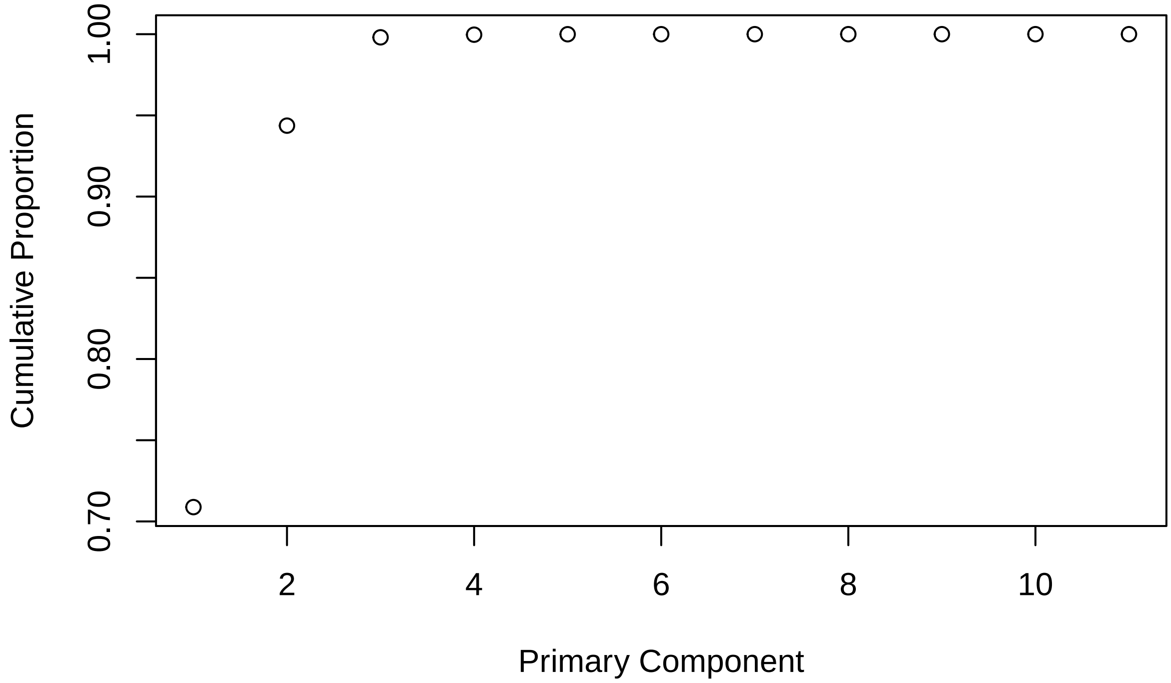

Figure 1. Cumulative proportion of variance explained by each primary component.

Therefore, I created a pca_train and pca_test set with the first seven primary components and the correponding Dataset labels.

Plotting the first three primary components for each dataset, we see a noticeable difference in distribution.

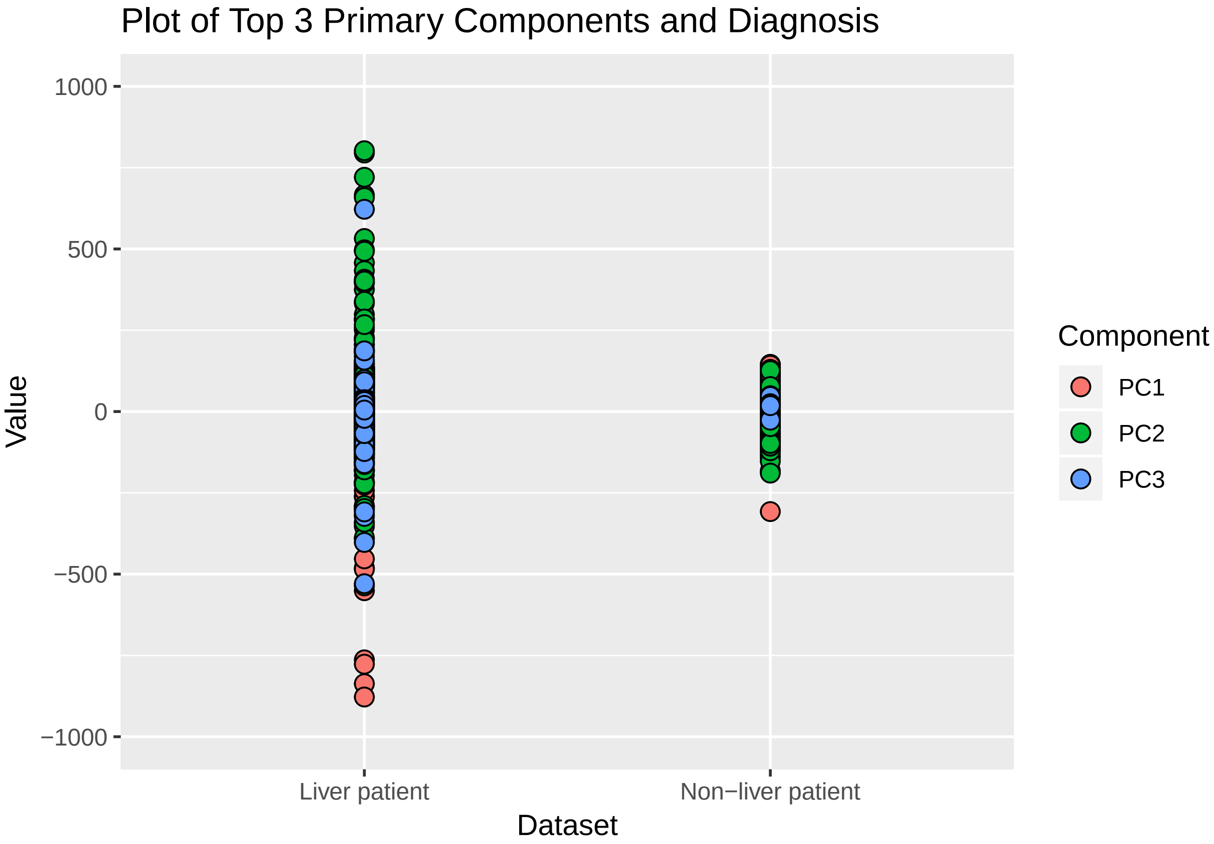

Figure 2. First three primary components for each dataset.

```
##             Model  Accuracy
## 1       svmLinear 0.7137931
## 2             gbm 0.7068966
## 3       svmRadial 0.7137931
## 4   svmRadialCost 0.7137931
## 5 svmRadialSigma 0.7137931
```

Table 4. Post-PCA accuracy of each algorithm used in the ensemble.

Unsurprisingly, due to the lower number of dimensions the accuracy did not change much. The mean accuracy of the ensemble was 0.7124138 and the accuracy after majority votes was 0.7137931.

Conclusion

Predicting liver disease from comprehensive health information is less accurate than we would hope for, but is nevertheless possible. The high variance within the population with liver disease, as seen in Table 3 and visualized in Figure 2, is likely the reason behind the lower accuracy of machine learning algorithms. Nevertheless, the ability to predict liver disease by 0.714 accuracy can prove to be a useful tool that can be used to ease the burden on healthcare providers.

Acknowledgements:

This dataset was downloaded from the UCI ML Repository: Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.