Predictive Analysis
Liver Disease

Barbara Payne
Bellevue University
DSC680-T302
Summer 2021

Abstract

The liver is an important organ in the human body. It regulates the blood and helps with the production of some proteins for blood. The liver needs to be taken care of, as a person only has one. Too much consumption of alcohol, the consumption of contaminated food, a poor diet, or the use of drugs can contribute to liver disease. As with any diseases, it is best to catch it early to begin treatment or have preventive measures from the disease even happening.

Machine learning and predicative analysis are very beneficial to the medical field. By knowing the signs of what to look for, machine learning can be able to predict if a patient has a disease early than what a doctor can. It can look at certain trends of past patients to help future patients. Liver disease is just one of the diseases that can be used by machine learning in the health industry.

Introduction

Liver Disease

The liver is used to filter blood that comes from the digestive track in the body. If this is compromised, the liver could pass blood that is contaminated. The liver needs to be protected and taken care of as it is a vital part of the human body. It can also detoxify chemicals and metabolizes drugs that pass through. Proteins are also created that can prevent blood clotting (Hoffman, 2021). There are many important functions that the liver does for the body. One of which is that it produces bile, which helps the small intestine during digestion by removing waste and breaking down fats. Cholesterol is also produced by the liver, which is used to spread special fats in the body. It also helps regulate blood flow (*Liver: Anatomy and Functions*, n.d.).

As mentioned above, the liver is vital to the body. It must be protected and taken care of. There are many factors that can affect the liver. Certain over the counter drug combinations can affect the liver. Poor diet is another factor. If the liver is disturbed, then a disease can form. It is the largest solid organ in the body. Typically, more than 75% of the liver needs to be affected before a decrease in liver function occurs. There are signs to look for in liver disease. Some of these signs are the skin being yellow (jaundice), abdominal pain, itchy skin, dark yellow urine, and loss of appetite (Wedro, B., 2021).

Predictive Analysis

Predictive analysis is used by looking at past trends of data to be able to predict what would happen in the future. Statistical algorithms and machine learning techniques are used to identify likelihood of future outcomes on historical data. There are many benefits to using predictive analytics. Some of these benefits are detecting fraud, optimizing campaigns, improving operation, and reducing risks (*Predictive analytics: What it is and why it matters*, n.d.).

The quantitative prediction has a certain workflow that it must follow for it to be accurate. It begins with a business goal that must be defined. The large amount of data is first sorted through to see what is valuable. This is done in the exploratory data analysis. The data is preprocessed. Predictive models are then developed. The model is created, optimized and validated. The analytics is integrated with the systems (*What is PREDICTIVE Analytics?*, n.d.).

Predictive analytics is great for the health industry. It can answer the question of what could happen next. By watching trends of what has happened to other patients, the likelihood of the trend to continue within other patients is high. It can be used for prognosis, diagnosis,

treatments, improving care quality, reducing costs, and reducing adverse effects (Birkmeyer, 2020). This is all very important for the patient. Improving the quality of life for the patient is the overall goal in the healthcare industry. Predictive analytics can help with this.

Dataset

https://www.kaggle.com/uciml/indian-liver-patient-records?select=indian_liver_patient.csv

The dataset I chose can be found on Kaggle at the above link. It contains information on the patient, such as their age and their sex. There is also some medical information, such as the levels of bilirubin, protein levels, and albumin levels. Some of the patients have a liver disease and some do not. Majority of the patients do have some type of liver disease and there are more males than females in the dataset.

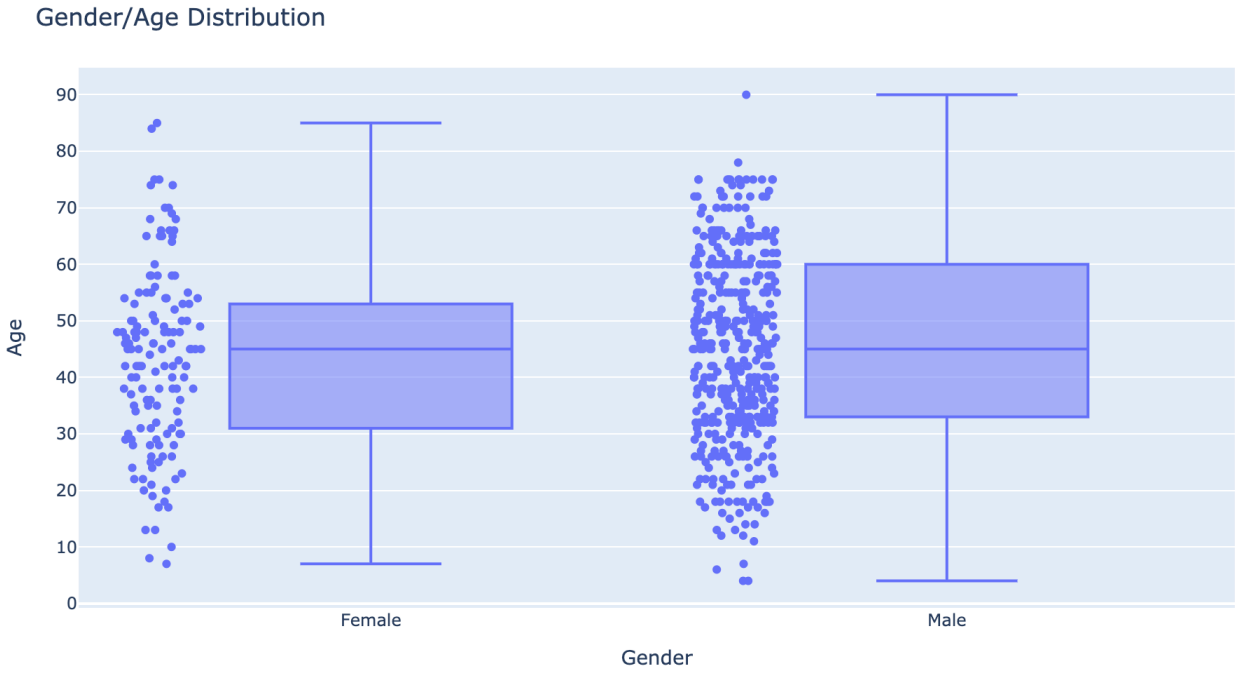
Methods

With any data science project, the first step would be to conduct Exploratory Data Analysis (EDA). Within this phase the outliers and any anomalies will be called out. Any bad data would be removed within this phase. Bad data could be missing values and nulls within the dataset.

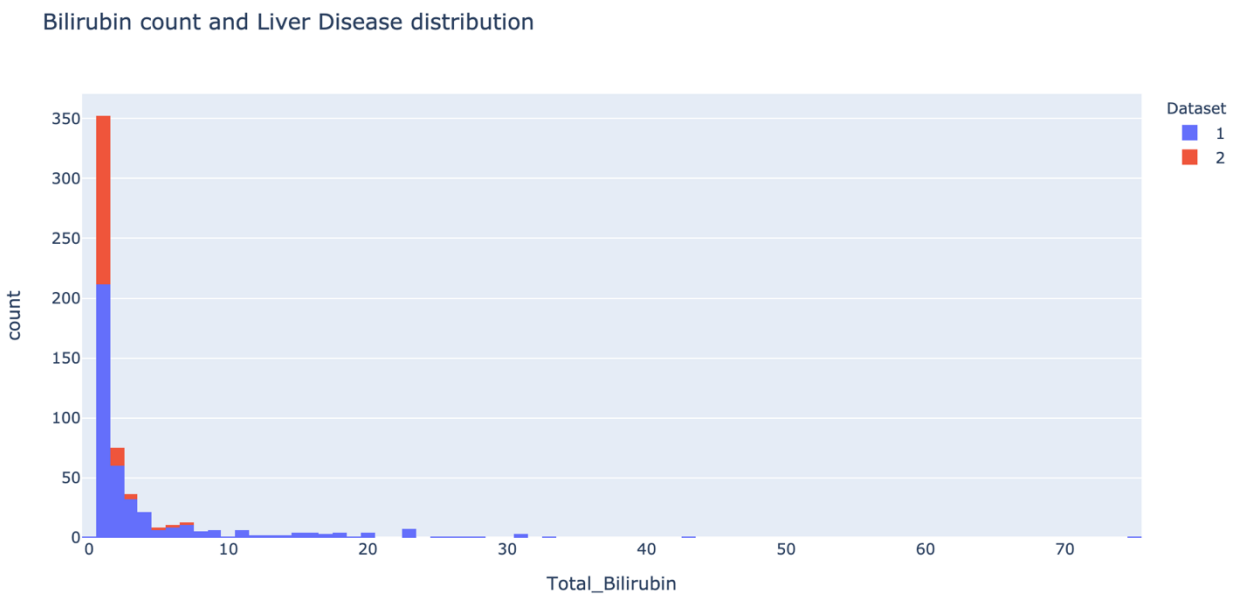
After EDA, I can then do the predictive analysis. I can get an accuracy score of how well my model can predict liver disease by training, testing and validating the data. I can then use different types of data models to get the accuracy score of each. The models I will be using are Logistic Regression, K-Nearest Neighbors (KNN), Decision Tree, Random Forest, and Support Vector Machine (SVC). I can then see which of these models will give the best accuracy score from my dataset.

Results

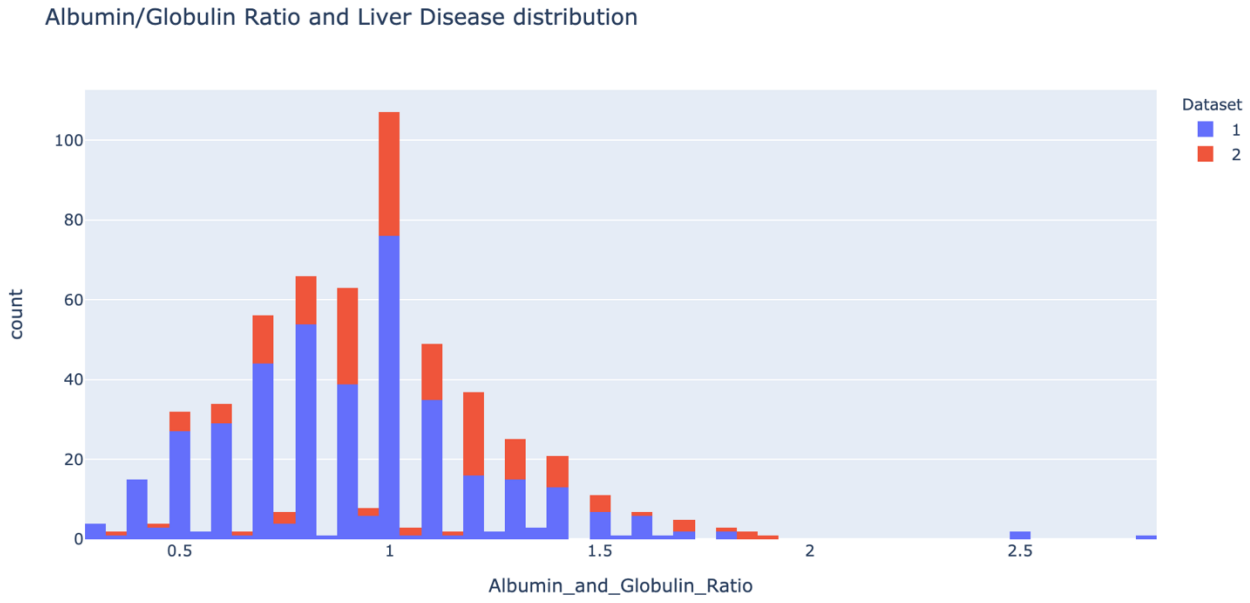
At this point, I have conducted mostly the EDA. I have found that in my dataset there are a lot more male patients than there are females. This is inconclusive that male patients are more likely to have liver disease than females because there is a skew in the data. I have also removed any null values from the dataset. I have also started comparing some of the variables with each other through distribution. Even though there is more males in the dataset, I wanted to see the gender/age distribution. The average age among both groups is 45.



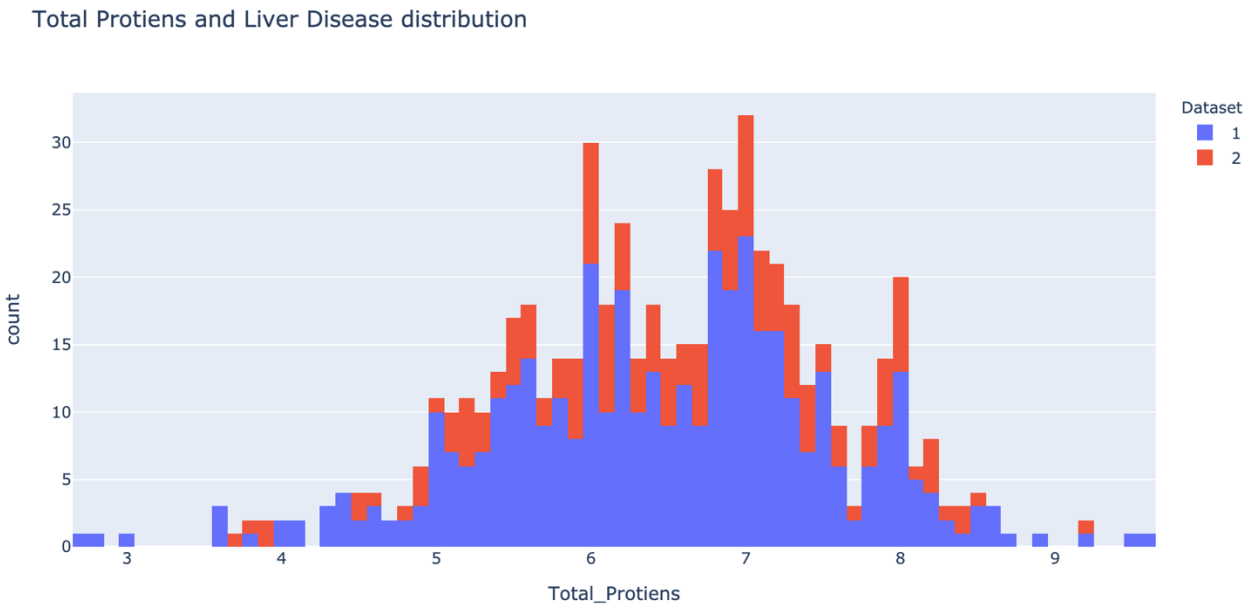
Bilirubin is a compound that helps break down food for your body to pass it. Higher levels of bilirubin are an indicator that the patient may have some type of liver disease. The bilirubin should be at a lower level as too much bilirubin can have your red blood cells break down at an unusual rate (Felson, 2021). Below is the distribution of bilirubin count with patients that have liver disease (1) and those that do not (2).



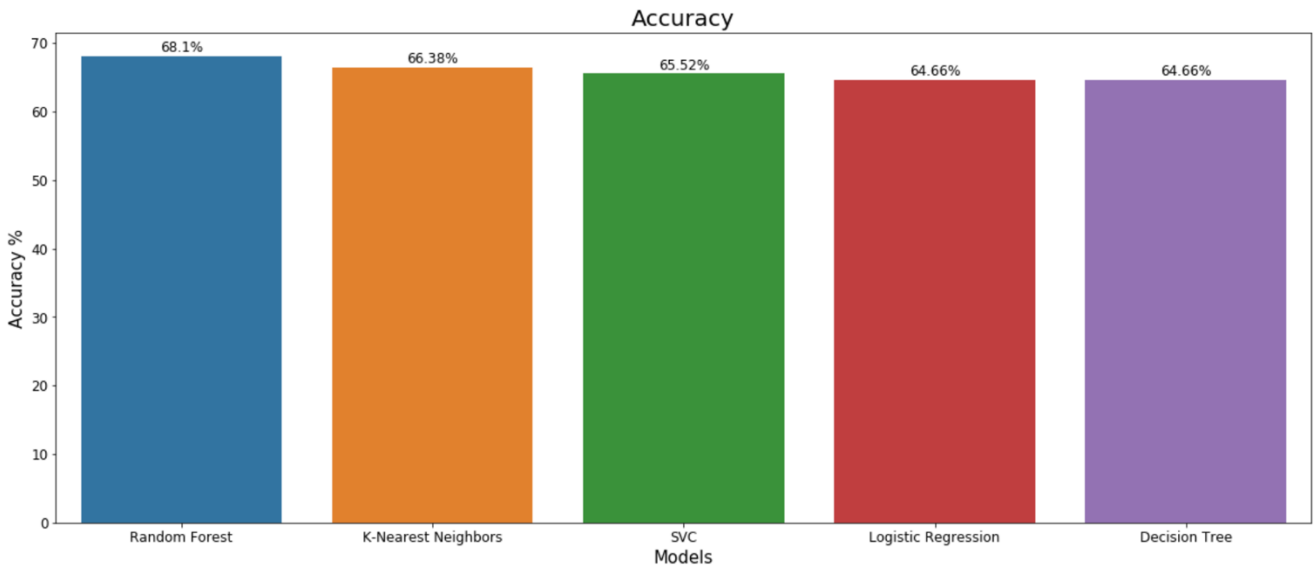
Bilirubin is not the only reason why a patient may have liver disease. Albumin is a protein made by the liver. It keeps fluids from leaking out of the blood vessels, transports hormones and vitamins, and nourishes tissue. Globulins is the other type of protein that is found in the blood. This helps fight off infections and transports nutrients. Albumin makes up 60% of the protein and Globulin makes up the other 40% (*Total protein, albumin-globulin (a/g) ratio*, 2021). The average Albumin/Globulin is just over 1.



Having too low of protein levels can be an indicator for liver problems. When the protein levels are too high, could mean that you are dehydrated or cancer. The normal range of protein levels is 6.0 to 8.3 g/dl.



As mentioned earlier, I decided to use the following models to compare the accuracy scores of the predictive model. I used logistic regression, KNN, Decision Tree, Random Forest, and SVC. The average accuracy model was at 65%. The highest accuracy score was with the Random Forest at 68.1%. The lowest accuracy score was between the Decision Tree and Logistic Regression, both with a score of 64.66%.



Conclusion

The liver is a vital organ. It is important to keep it healthy as it helps break down waste in the body. Without a liver, you would not survive. It regulates the chemicals in the blood and produces bile. Ensuring that the liver is healthy is extremely important. If liver disease is detected, you would want to get treatments to help maintain healthy levels or have a lifestyle change.

Machine learning and predictive analysis are beneficial to the healthcare industry. By taking trends from previous patients, it can be inferred of what could happen to future patients. It can help doctors look for signs and can give good recommendations on what has been done for previous patients. Technology is always advancing, and predictive modeling is proof that technology is beneficial to people as it can help save lives.

The dataset could use some more distributed data as there were some patients that did not have liver disease have high levels of bilirubin or a high albumin and globulin ratio. There was also some skew in the data between men and women and there was a lot more patients

that have liver disease than those that do not. If these were more evenly distributed, the accuracy scores of the models could have increased. Though my accuracy scores across all models are not ideal, it was a good learning opportunity to see the different ways of using predictive modeling.

References

Birkmeyer, C. (2020, November 16). *An intro to predictive analytics in Healthcare [2020] // ArborMetrix*. RSS. https://www.arbormetrix.com/blog/intro-predictive-analytics-healthcare.

Felson, S. (2021, February 6). *Bilirubin test: High vs. low Levels, direct vs. indirect*. WebMD. https://www.webmd.com/a-to-z-guides/bilirubin-test.

Hoffman, M. (2021, June 23). *Liver (anatomy): Picture, FUNCTION, CONDITIONS, Tests, Treatments*. WebMD. https://www.webmd.com/digestive-disorders/picture-of-the-liver.

Liver: Anatomy and functions. Johns Hopkins Medicine. (n.d.). https://www.hopkinsmedicine.org/health/conditions-and-diseases/liver-anatomy-and-functions.

Predictive analytics: What it is and why it matters. SAS. (n.d.). https://www.sas.com/en_us/insights/analytics/predictive-analytics.html.

Total protein, albumin-globulin (a/g) ratio. Lab Tests Online. (2021, August 6). https://labtestsonline.org/tests/total-protein-albumin-globulin-ag-ratio.

Wedro, B. (2021, March 18). *Liver disease symptoms, treatment, stages, signs, types, diet*. MedicineNet. https://www.medicinenet.com/liver_disease/article.htm.

What is PREDICTIVE Analytics? - 3 things you need to know. What Is Predictive Analytics? - 3 Things You Need to Know - MATLAB & Simulink. (n.d.). https://www.mathworks.com/discovery/predictive-analytics.html.

10 Questions

1. Why did you want to do a predictive analysis?
2. Why did you choose the dataset you chose?
3. What were you hoping to gain from this project?
4. What challenges did you face while doing the project?
5. What went right in the project?
6. Is there anything you would have done differently?
7. Why did you choose these methods?
8. Was there any data that was excluded from the dataset?
9. How much time did you spend on the analysis?
10. What you know now, what would you tell your past self when starting the project?