PREDICTION OF NEONATAL HYPERBILIRUBINEMIA

Pulkit Pandey
2K19/EP/076 Engineering Physics
Delhi Technological University
New Delhi,110042, India
ppulkit9.pp@gmail.com

Karan
2K19/EC/087 Electronics and Communication
Delhi Technological University
New Delhi,110042, India
kk85968@gmail.com

ABSTRACT **- Neonatal jaundice also known as hyperbilirubinemia is a phenomenon that occurs in 60% of all term new-borns. Severe cases of this may lead to chronic encephalopathy, kernicterus, neurologic deficits, and, in some cases, death. Intensive blue light phototherapy involves irradiation of a patient with high intensity light and is widely used for treatment of severe hyperbilirubinemia. The aim of this study to predict whether a person suffers from hyperbilirubinemia and require phototherapy treatment or not**.

Index Terms – Hyperbilirubinemia, Phototherapy, Logistic Regression.

I. INTRODUCTION

Hyperbilirubinemia is a medical condition in which the content of bilirubin exceeds the normal range in a neonate body. Bilirubin (BR) is a yellow compound that occurs in the normal catabolic pathway that breaks down heme. This catabolism is a necessary process in the body's clearance of waste products that arise from the destruction of aged or abnormal red blood cells (RBCs). It's hard for babies to get rid of bilirubin at first. It can build up in their blood, tissues, and fluids. Bilirubin has a colour. It makes a baby's skin, eyes, and other tissues turn yellow (jaundice). The aim of this study was to establish a predicting model by using Indian liver patient data set for prediction of subsequent hyperbilirubinemia in humans.

II. LITERATURE REVIEW

A. Background Theory

Unconjugated 4Z,15Z-bilirubin is a toxic and insoluble product of heme-catabolism that does not readily exit the body. Bilirubin leaves the body after binding to glucuronic acid in the liver to form a water-soluble compound (i.e. conjugated bilirubin), which can be excreted into the bile. Blood vessels transport unconjugated bilirubin to the liver as bilirubin bound to albumin. Bilirubin dissociates from albumin in the liver, is rapidly internalized by hepatocytes, and binds to ligandin molecules inside the cells. Bilirubin and glucuronic acid conjugate in a reaction catalyzed by diphosphoglucuronyltransferase (UDPGT).

Indirect hyperbilirubinemia is defined as an abnormally elevated concentration of unconjugated bilirubin in the blood. Neonatal hyperbilirubinemia occurs in the majority of newborns. Two concurrent factors contribute to this prevalence. Firstly, fetal erythrocytes are rapidly broken down shortly postpartum, resulting in a rapid production of unconjugated bilirubin. Secondly, ligandin concentration and UDPGT activity are both below mature levels in the neonatal liver. Neonates with congenital liver disease, such as Gilbert Syndrome, are at an increased risk for severe hyperbilirubinemia. Severe hyperbilirubinemia can cause chronic brain damage. Unconjugated bilirubin in the blood can transcend the blood brain barrier and accumulate in the central nervous system (CNS). Slightly more than half of all neonates become visibly jaundiced in the first week of life. Almost all hyperbilirubinemia in the immediate neonatal period is unconjugated, which is termed indirect bilirubin, based on older laboratory measurement methods; conjugated bilirubin is termed direct bilirubin.

Accumulation of bilirubin in the CNS can lead to permanent brain lesions, resulting in long-term neurologic disease and possibly death.

B. Fundamentals of Phototherapy

Phototherapy is the use of visible light for the treatment of hyperbilirubinemia in the new born. This relatively common therapy lowers the serum bilirubin level by transforming bilirubin into water-soluble isomers that can be eliminated without conjugation in the liver. The dose of phototherapy is a key factor in how quickly it works; dose in turn is determined by the wavelength of the light, the intensity of the light (irradiance), the distance between the light and the baby, and the body surface area exposed to the light [2]. Commercially available phototherapy systems include those that deliver light via fluorescent bulbs, halogen quartz lamps, light-emitting diodes, and fiber-optic mattresses. Proper nursing care enhances the effectiveness of phototherapy and minimizes complications. Caregiver responsibilities include ensuring effective irradiance delivery, maximizing skin exposure, providing eye protection and eye care, careful attention to thermoregulation, maintaining adequate hydration, promoting elimination, and supporting parent-infant interaction

Phototherapy converts bilirubin that is present in the superficial capillaries and interstitial spaces of the skin and subcutaneous tissues to water-soluble isomers that are excretable without further metabolism by the liver. [1] Neonatal jaundice expert Maisels suggests that phototherapy is much like a percutaneous drug. When phototherapy illuminates the skin, an infusion of discrete photons of energy are absorbed by bilirubin much like a drug molecule binds to a receptor. Bilirubin molecules in light-exposed skin undergo relatively quick photochemical

reactions—configurational isomerization, structural isomerization, and photooxidation—to form nontoxic, excretable isomers. These bilirubin isomers have different shapes than the native isomer, are more polar, and can be excreted from the liver into the bile without undergoing conjugation or requiring special transport for their excretion. Urinary and gastrointestinal elimination remain important to the process of reducing the bilirubin load.

C. *Machine Learning Method*

Classification is one of the most important aspects of supervised machine learning. In machine learning, classification refers to a predictive modelling problem where a class label is predicted for a given example of input data. Logistic Regression, a method for classifying the data in Machine Learning. Logistic regression is generally used where we have to classify the data into two or more classes. One is binary and the other is multi-class logistic regression. As the name suggests, the binary class has 2 classes that are Yes/No, True/False, 0/1, etc. In multi-class classification, there are more than 2 classes for classifying data. [4] Logistic Regression was used in the biological sciences in early twentieth century. It was then used in many social science applications. Logistic Regression is used when the dependent variable(target) is categorical. Consider a scenario where classification is to be done whether an email is spam or not. If linear regression is used for this problem, there is a need for setting up a threshold based on which classification can be done. Say if the actual class is malignant, predicted continuous value 0.4 and the threshold value is 0.5, the data point will be classified as not malignant which can lead to serious consequence in real time. From this example, it can be inferred that linear regression is not suitable for classification problem. Linear regression is unbounded, and this brings logistic regression into picture. Their value strictly ranges from 0 to 1.

III. Analysis/Implementation

A. *Fixing of Data:*

Missing values in dataset were removed. For easier readability and for better understanding some column names were changed.

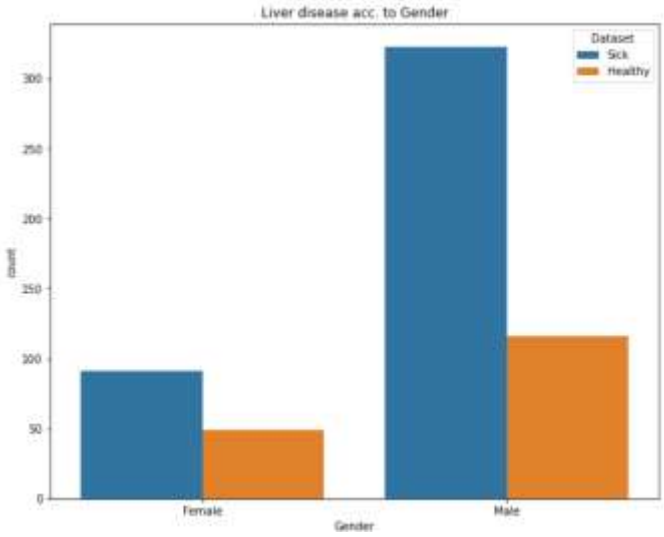
B. *Division of Data*

Division of data into categorical and quantitative variables so as to perform mathematical operations on quantitative variable leaving categorical ones for grouping and fraction operations.

C. *Statistical analysis on study group*

i. Categorical data:

Data was grouped according to gender and further divided under sick and healthy category to make visualization of data through a different perspective.



Dataset	Healthy	Sick
Gender		
Female	35.00000	65.00000
Male	26.42369	73.57631

Fig.3- Operations on categorical data

ii. Quantitative data:

Coefficient of variation: The coefficient of variation is often used to determine if a variable will be important to the model. In our case, the variable: Protein and Albumin do not have a very high V, compared to the other variables.

```
Age                         0.033591
Total_Bilirubin             4.890768
Direct_Bilirubin            3.199163
Alkaline_Phosphotase        3.753582
Alamine_Aminotransferase    6.527575
Aspartate_Aminotransferase 10.512251
Total_Protiens             -0.292433
Albumin                     0.048516
Albumin_and_Globulin_Ratio  0.992299
dtype: float64
```

(a)

Kurtosis: Kurtosis is a measure of outliers. The higher its value, the more likely there are outliers in the database. The lower the value, the more the results are clustered around the mean. In our case, five variables exceeded the safe threshold of K=|3|K=|3|, of which Aspartate and Alamine very strongly. This meant that there were many outliers in the database.

Age	46.192293
Total_Bilirubin	187.681424
Direct_Bilirubin	188.341676
Alkaline_Phosphotase	83.520836
Alamine_Aminotransferase	225.605118
Aspartate_Aminotransferase	262.284033
Total_Protiens	16.719460
Albumin	25.290574
Albumin_and_Globulin_Ratio	33.716416
dtype: float64	

(b)

Skewness: The skewness for most variables is positive, indicating that the distribution has an extended right arm. This proved our distribution was not normal due to heavy outliers.

Age	-0.564927
Total_Bilirubin	36.921240
Direct_Bilirubin	11.252207
Alkaline_Phosphotase	17.634683
Alamine_Aminotransferase	50.234929
Aspartate_Aminotransferase	149.939487
Total_Protiens	0.239473
Albumin	-0.390890
Albumin_and_Globulin_Ratio	3.281900
dtype: float64	

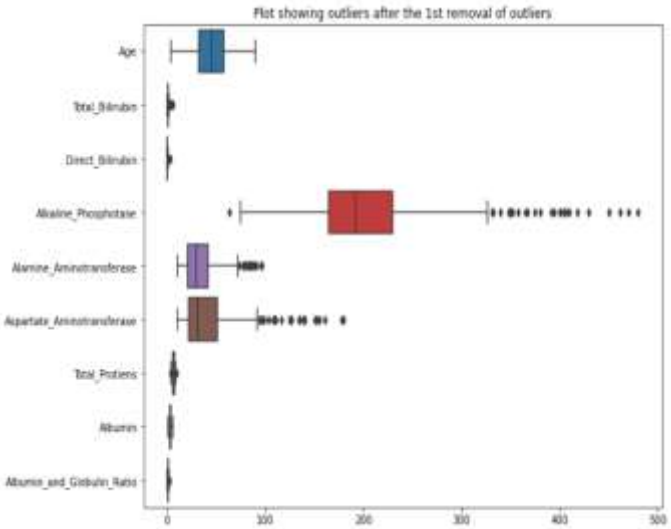
©

Fig.4- (a): Variation; (b): Kurtosis; (c): Skewness

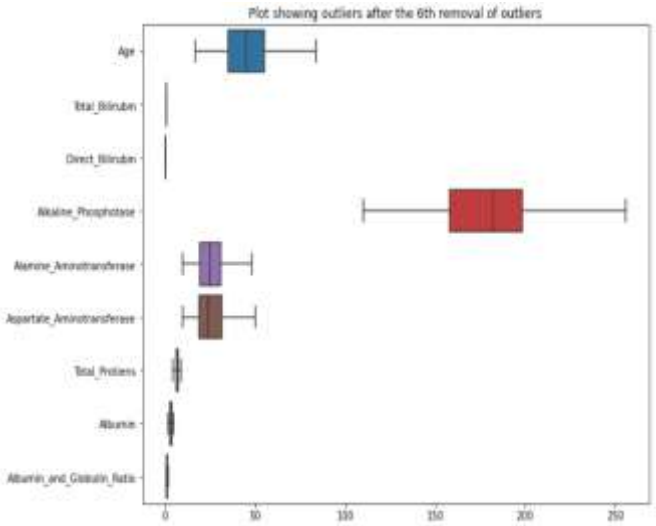
D. Cleaning of Data

i. Removal of Outliers:

Database had a significant number of outliers so to deal with it the interquartile range $IQR = Q_3 - Q_1$ was applied. The outlier observations that are below the lower bound defined as $lb = Q_1 - 1.5 * IQR$ and above the upper bound defined as $ub = Q_3 + 1.5 * IQR$.



(a)

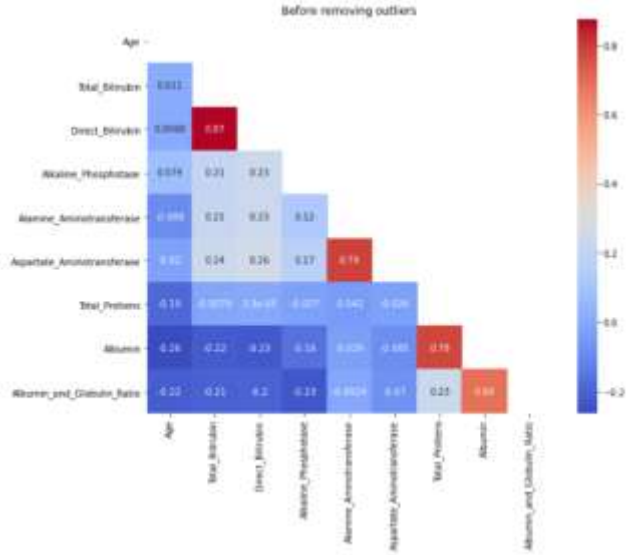


(b)

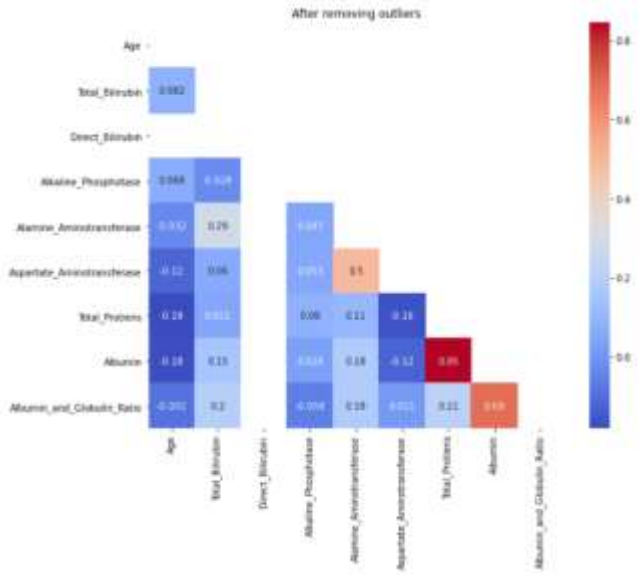
Fig.5- (a): Database after running the outlier removal function once; (b): Database after running the outlier function 6 times for complete removal of outliers

ii. Removal of Highly correlated variable:

To further increase model's accuracy and we decided to avoid the problem the multicollinearity. To address this, we used Pearson Correlation Coefficient both before and after removing outliers. From the correlation results, it was observed removing outliers reduced the correlations in the database which is a positive effect considering the Logistic Regression model. Also, the only highly correlating variable was Albumin therefore we removed it from the database before building the logistic regression model.



(a)



(b)

Fig.6- Heatmap for correlation (a): Before removing outliers; (b): after removing outliers

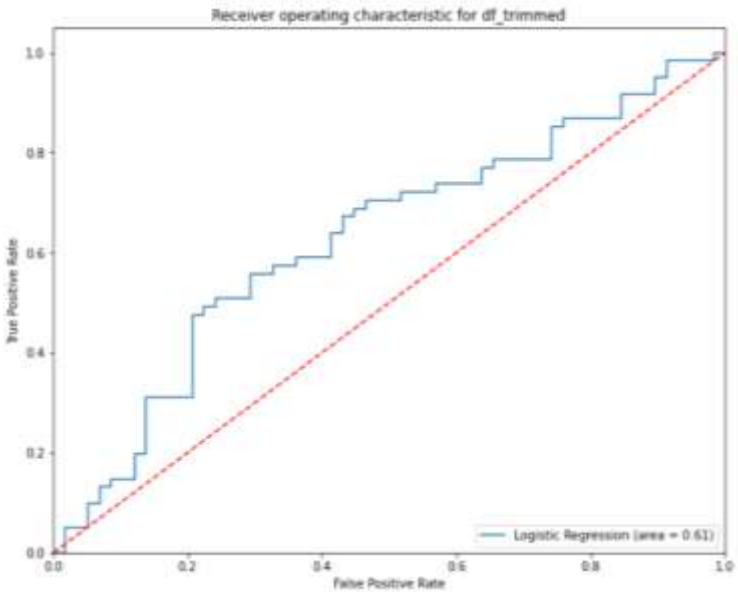
E. Machine learning model:

In this paper, two models were made, one with the cleaned & trimmed dataset and the other one with the original dataset (one with outliers). Since, clean and trimmed dataset was not big enough we directly used it logistic regression model without splitting it while splitting was done on original dataset into training and testing and then used it in model. But to compensate for less data the number of iterations of model was increased.

Performance was measured as AUC, i.e., the area under the receiver operator characteristic (ROC) curve, which evaluates a model based on the true positive rate versus false positive rate of its predictions at different threshold values. It is a default metric for the evaluation of binary classifiers and it has been used in related studies, thus allowing us to draw a comparison between the two models.

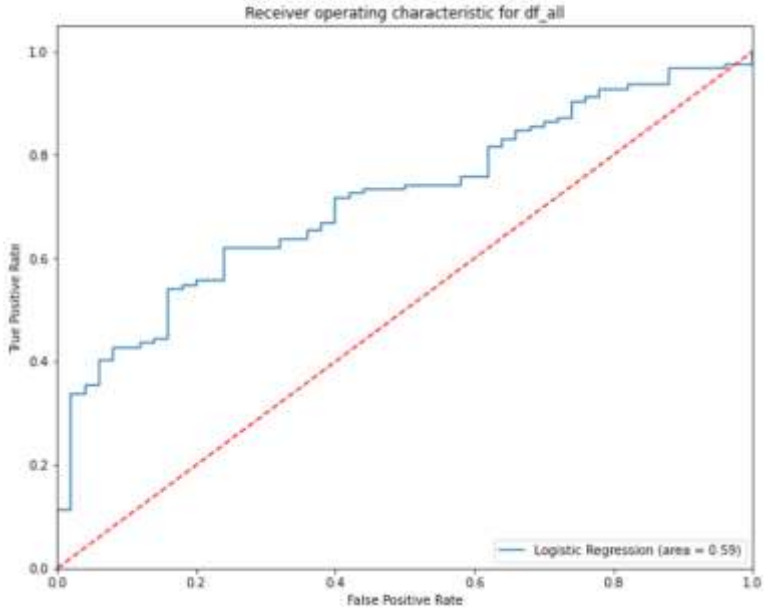
IV. RESULT

After applying the logistic regression on both the models it was found that AUROC of our trimmed data was 0.61 which slightly better that than the AUROC of our original data which is 0.51. The result was poor because of the lack of sufficient data to train our model. Almost 80% of data was lost after removing the outliers which became our model's prominent limitation.



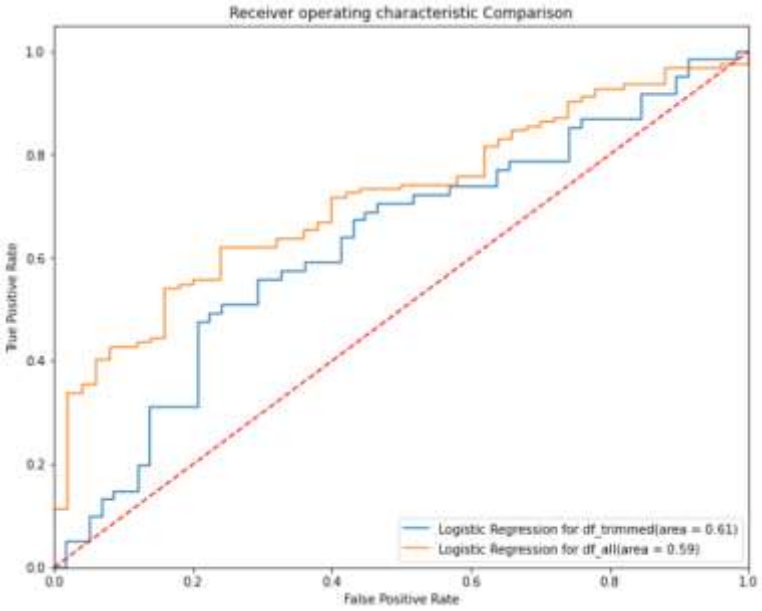

```
Accuracy: 0.6134453781512605
Precision: 0.6056338028169014
Recall: 0.7049180327868853
Balanced accuracy score: 0.6110797060486151
```

(a)




```
Accuracy: 0.6954022988505747
Precision: 0.7591240875912408
Recall: 0.8387096774193549
Balanced accuracy score: 0.5893548387096774
```

(b)



	df_trimmed	df
Accuracy	0.613445	0.695402
Precision	0.605634	0.759124
Recall	0.704918	0.838710
Balanced accuracy	0.611080	0.589355

(c)

Fig.7- (a): Receiver operating characteristic (ROC) curve for the prediction model in trimmed dataset. The area under the ROC (AUROC) curve was 0.61; (b): Receiver operating characteristic (ROC) curve for the prediction model in original dataset. The area under the ROC (AUROC) curve was 0.59; (c): Comparison of ROC curve of both models

V. Acknowledgment

Presentation, Inspiration and Motivation have always played a key role in success of any project.

A deep sense of gratitude to Dr Prashant Giridhar Shambharkar, Assistant Professor, Department of Computer Engineering, DTU to encourage us to highest peak and to provide with the opportunity to prepare the project. We are immensely obliged to him for his elevating inspiration, encouraging guidance and kind supervision in the completion of the project.

We are also thankful for his invaluably constructive criticism and advises throughout the working of the project. The accomplishment of the project was due to our sole efforts and contribution. We are highly indebted towards our University, that is, Delhi Technological University, New Delhi

VI. References

[1] https://www.nature.com/articles/jp201271 H. Simpson, *Dumb Robots*, 3rd ed., Springfield: UOS Press, 2004, pp.6-9.

[2] https://www.nursingcenter.com/ce_articleprint?an=00149525-201110001-00003.

[3] https://machinelearningmastery.com/types-of-classification-in-machine-learning/

[4] https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc

[5] Model to Predict Hyperbilirubinemia in Healthy Term and Near-Term Newborns with Exclusive Breast Feeding - ScienceDirect

[6] https://pubmed.ncbi.nlm.nih.gov/32156481/

[7] https://pubmed.ncbi.nlm.nih.gov/30928997/