# A New Paradigm of Machine Learning in the Field of Bioinformatics

##### **Farooq Ansari, Rohan Kumar Singh, Rohit Kumar Majee and Arkadeep Bagal**

Department of Computer Science and Engineering

ASANSOL ENGINEERING COLLEGE

***Abstract* -** Chronic liver diseases impact humans around the globe. Through the research done in the medical field, the professional are able to diagnose it properly due to the indicators that the disease leaves in the body.

Through this project, we are trying to predict the diagnosis of liver disease based on those indicating parameters as accurately as possible.

Using the model machine learning techniques implemented in Classification algorithms, we're trying to come up with an optimal model that provides maximum performance with minimal feature engineering.

1. Introduction

T

his project mainly deals with the ways of how we can improve upon the existing models for this particular dataset implemented using Logistic Regression algorithm, and achieving a better Test Set accuracy than them.

As the field of study has observed, the proper way to achieving a well performing model is to feed it the most cleaned and processed dataset. Feature Engineering in that context is a field on itself. It has been observed in the past that for most, if not all, of the popular machine learning dataset, some noise in form of either missing or false values, or irrelevant features are present.

This particular dataset was mostly used in literature in its raw form i.e. without the exclusion of any of its features in the input dataset. What we have tried to do is select the best possible set of features from this dataset that outputs the highest accuracy.

Our resolve is that even a small dataset with proper feature selection and engineering and with a simple machine learning algorithm such as Logistic Regression can result in high model performance.

1. DATASET

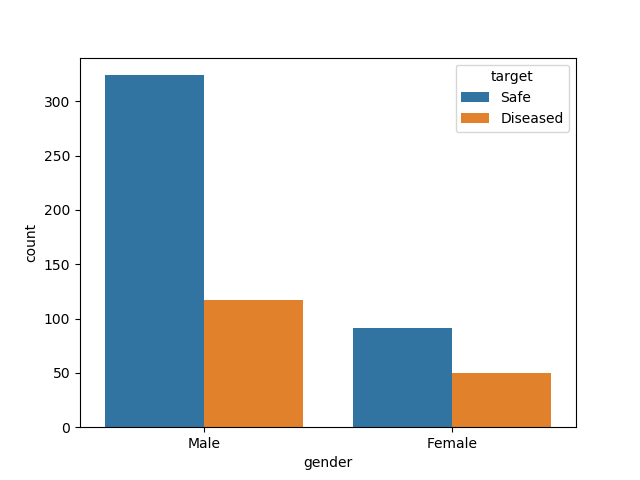
The Dataset that we would be using is called *Indian Liver Patient Dataset*. It is a popular dataset referenced in literature various times through different research papers on the classification algorithms. There are total 583 records out of which 416 have the disease and 167 do not.

The features present in the dataset are as follows:

1. **AGE** - Age of the patient  
2. **GENDER** - Gender of the patient  
3. **TB** - Total Bilirubin  
4. **DB** - Direct Bilirubin  
5. **ALKPHOS** - Alkaline Phosphotase  
6. **SGPT** - Alamine Aminotransferase  
7. **SGOT** - Aspartate Aminotransferase  
8. **TP** - Total Protiens  
9. **ALB** - Albumin  
10. **A/G** - Ratio Albumin and Globulin Ratio  
11. Selector field used to split the data into two sets (labeled by the experts)

1. DATA VISUALISATION

We used popular data visualization tools like seaborn and matplotlib in Python to plot some relevant information that we could gather from the dataset.



1. FEATURE ENGINEERING

We used the following methods.

## Imputation

In the dataset, some fields are missing for some records. We are using an imputer class called Simple Imputer that inserts in those places, the mean value of that particular field.

## B. Feature Normalization

Since all the fields are measured and documented in different units, we are using a mean normalization technique to put all the values in the same scale for better performance of our model.

## C. Feature Selection

Many of the popular dataset, though painstakingly acquired through different means and put together to further the research happening in the field of Data Analysis and prediction, have fields and values that might not prove relevant or useful for the optimum prediction model.   
  
Here, we tried to find the optimum combination among the different features of the dataset that results in the maximum accuracy of the model.  
  
For that we are iterating through all the possible combination of all the possible sizes of the features, and choosing the combination that results in the optimum accuracy.

Following is the pseudo code for our approach.

data = open('dataset.csv)  
  
X = data.features  
y = data.target  
  
max\_columns = data.feature\_length  
max\_accuracy = -1  
no\_of\_columns = 1  
  
while no\_of\_columns < max\_columns  
  
 all\_combinations = every combination of that size

*# if no\_of\_columns = 3, then*

*# all\_combinations = [1,2,3], [2,3,4], [5,6,7] etc*  
 for combination in all\_combinations  
  
 X = X[combination]

***# selects just those columns from X***

model = LogisticRegression(X,y)  
 accuracy = model.accuracy()  
  
 if accuracy > max\_accuracy  
 max\_accuracy = accuracy  
 best\_combination = combination  
  
 no\_of\_columns++

return max\_accuracy, best combination

1. MODEL OPTIMIZATION

We are using the Logistic Regression algorithm as our prediction model.

We also tried using different combination of values for the parameters that goes in the library

Function that we are using for Logistic Regression. The parameters that we are using are:

C = Inverse of regularization strength; must be a positive float. Like in Logistic Regression, smaller values specify stronger regularization.

tol = Tolerance for stopping criteria.

Here’s a pseudo code of our implementation:

for value of C in [0.1, 1, 10, 100]

for value of tol in [0.001, 0.1, 0.01]

model = LogisticRegression(C, tol)

If model.accuracy > max\_acc:

max\_acc = model.accuracy

best\_C = C

best\_tol = tol

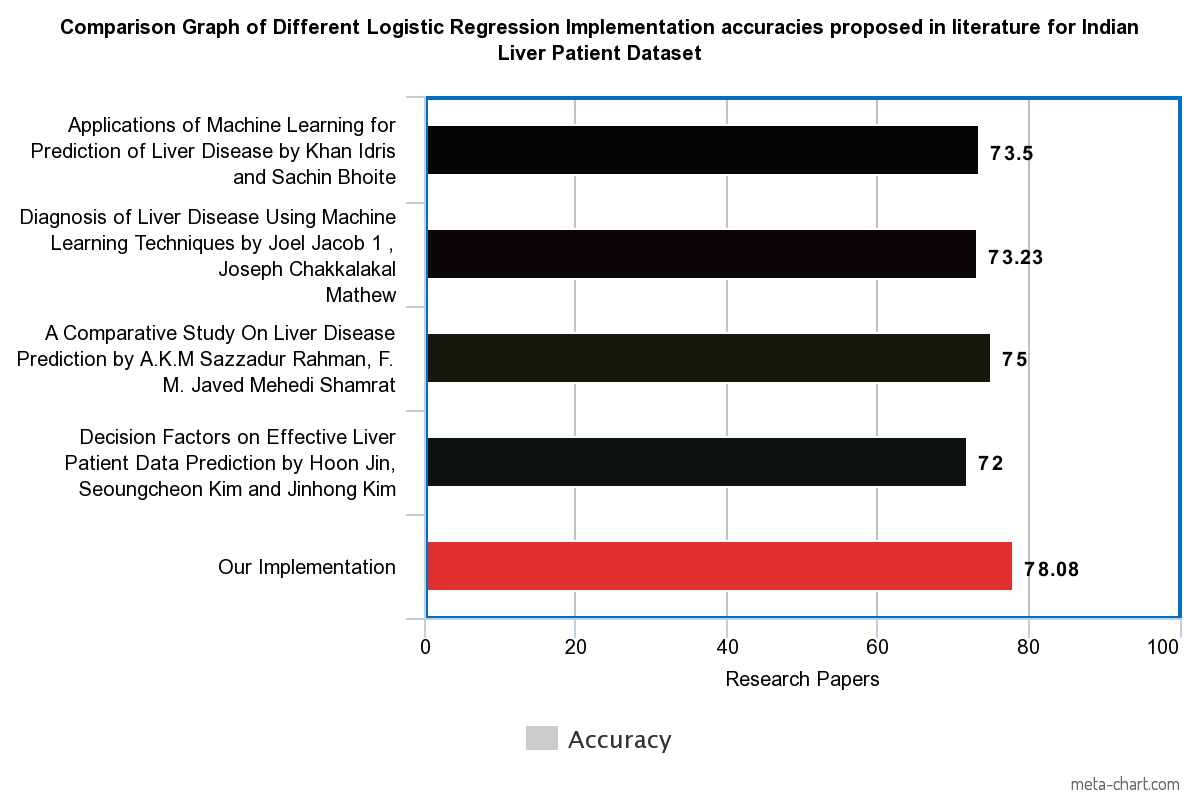
return best\_C, best\_tol

1. RESULTS

Applying the mentioned methods, the best accuracy that we managed to get was **78.08%,** which was significantly better than any other logistic regression approach that we could find in the literature at the moment. Our findings made us aware that choosing the columns [0, 1, 2, 3, 5, 6, 8] or AGE, GENDER, TB, DB, SGPT, SGOT and TP give us the best model to predict liver disease through this dataset.

It could be inferred that the other features only provide noise and irrelevant data for a Logistic Regression model resulting in lower accuracy.

Here’s a comparative study of other research done in the field.



1. FUTURE WORK

* We plan to incorporate an Evolutionary Machine Learning approach using Genetic Algorithm, that would take as input different iterations of Logistic Regression and would choose the best possible model for our dataset.
* We will be exploring the dataset with some data visualization tools.
* We will be adding a Graphical User Interface that any user can interact with to input some required features, and get a prediction in terms of percentage of the chances of those features resulting in a Liver Disease.
* We will discard the library function that we’re using at the moment and will implement our own Logistic Regression function.

1. CONCLUSION

We applied an iterative approach for feature selection to find the optimum combination of features for a predictive model, and we further tried to optimize that model through hyper-parameter tuning. We found a test set accuracy of 79.45%, which was better than most logistic regression implementation that we could find in the literature.

Acknowledgment

We would like to thank out mentor Prof. Sabyasachi Mukherjee for his continuous guidance and support though this project.

References

1. Applications of Machine Learning for Prediction of Liver Disease by Khan Idris and Sachin Bhoite
2. Diagnosis of Liver Disease Using Machine Learning Techniques by Joel Jacob 1 , Joseph Chakkalakal Mathew
3. A Comparative Study On Liver Disease Prediction by A.K.M Sazzadur Rahman, F. M. Javed Mehedi Shamrat
4. Decision Factors on Effective Liver Patient Data Prediction by Hoon Jin, Seoungcheon Kim and Jinhong Kim