# 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.

***Index Terms***- Bioinformatics, Evolutionary Algorithm, Genetic Algorithm, Machine Learning

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

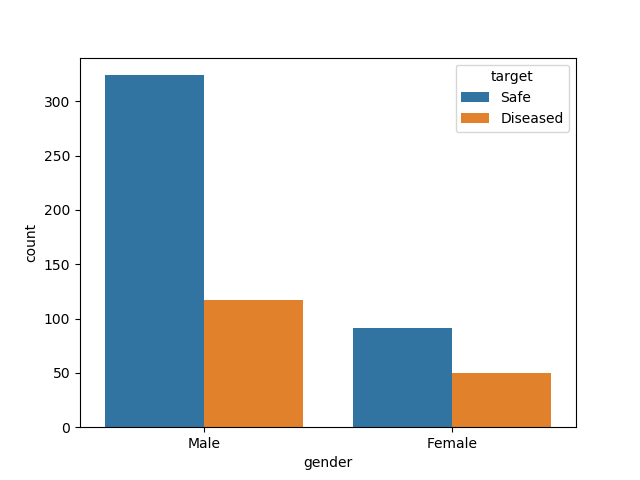
The Dataset that we would be using is called *Indian Liver Patient Dataset*. It is a popular dataset referenced in literature various times through different research papers on the classification algorithms. There are total 583 records out of which 416 have the disease and 167 do not.

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. GET PEER REVIEWED

Here comes the most crucial step for your research publication. Ensure the drafted journal is critically reviewed by your peers or any subject matter experts. Always try to get maximum review comments even if you are well confident about your paper.

**For peer review send you research paper in IJSRP format to** [**editor@ijsrp.org**](mailto:editor@ijsrp.org)**.**

1. IMPROVEMENT AS PER REVIEWER COMMENTS

Analyze and understand all the provided review comments thoroughly. Now make the required amendments in your paper. If you are not confident about any review comment, then don't forget to get clarity about that comment. And in some cases there could be chances where your paper receives number of critical remarks. In that cases don't get disheartened and try to improvise the maximum.

**After submission IJSRP will send you reviewer comment within 10-15 days of submission and you can send us the updated paper within a week for publishing.**

This completes the entire process required for widespread of research work on open front. Generally all International Journals are governed by an Intellectual body and they select the most suitable paper for publishing after a thorough analysis of submitted paper. Selected paper get published (online and printed) in their periodicals and get indexed by number of sources.

**After the successful review and payment, IJSRP will publish your paper for the current edition. You can find the payment details at:** [**http://ijsrp.org/online-publication-charge.html**](http://ijsrp.org/online-publication-charge.html)**.**

1. CONCLUSION

A conclusion section is not required. Although a conclusion may review the main points of the paper, do not replicate the abstract as the conclusion. A conclusion might elaborate on the importance of the work or suggest applications and extensions.

Appendix

Appendixes, if needed, appear before the acknowledgment.

Acknowledgment

The preferred spelling of the word “acknowledgment” in American English is without an “e” after the “g.” Use the singular heading even if you have many acknowledgments.

References

1. G. O. Young, “Synthetic structure of industrial plastics (Book style with paper title and editor),” in *Plastics*, 2nd ed. vol. 3, J. Peters, Ed. New York: McGraw-Hill, 1964, pp. 15–64.
2. W.-K. Chen, *Linear Networks and Systems* (Book style)*.* Belmont, CA: Wadsworth, 1993, pp. 123–135.
3. H. Poor, *An Introduction to Signal Detection and Estimation*. New York: Springer-Verlag, 1985, ch. 4.
4. B. Smith, “An approach to graphs of linear forms (Unpublished work style),” unpublished.
5. E. H. Miller, “A note on reflector arrays (Periodical style—Accepted for publication),” *IEEE Trans. Antennas Propagat.*, to be published.
6. J. Wang, “Fundamentals of erbium-doped fiber amplifiers arrays (Periodical style—Submitted for publication),” *IEEE J. Quantum Electron.*, submitted for publication.

Authors

**First Author** – Author name, qualifications, associated institute (if any) and email address.

**Second Author** – Author name, qualifications, associated institute (if any) and email address.

**Third Author** – Author name, qualifications, associated institute (if any) and email address.

**Correspondence Author** – Author name, email address, alternate email address (if any), contact number.