MACHINE LEARNING FOR BIOINFORMATICS

PREDICTING LIVER DISEASE USING LOGISTIC REGRESSION AND GENETIC ALGORITHM

Presented by

Arkadeep Bagal
Farooq Ansari
Rohan Kumar Singh
Rohit Kumar Majee

OUTLINE

- Introduction
- What is Machine Learning?
- Dataset
- Implementation
- GUI
- Results
- Comparative Study
- Conclusion

INTRODUCTION

In this short presentation, we would like to talk about and explore the various ways popular machine learning algorithms can help us extract useful and actionable data from seemingly ordinary datasets with its implementation.

We would also explore some popular machine learning algorithms and how they work.

WHAT IS MACHINE LEARNING?

DEFINITION

- Some popular definitions are as follows

- **Arthur Samuel(1959):** Field of study that gives computers the ability to learn without being explicitly programmed.

- **Tom Mitchell(1998):** A computer program is said to learn from experience**(E)** with respect to some task **(T)** and some performance measure **(**P) , if its performance on **T** , as measured by **P** , improves with experience **E.**

TYPES OF MACHINE LEARNING

- Supervised Learning
  - Linear Regression
  - Logistic Regression
- Unsupervised Learning
  - Clustering

DATASET

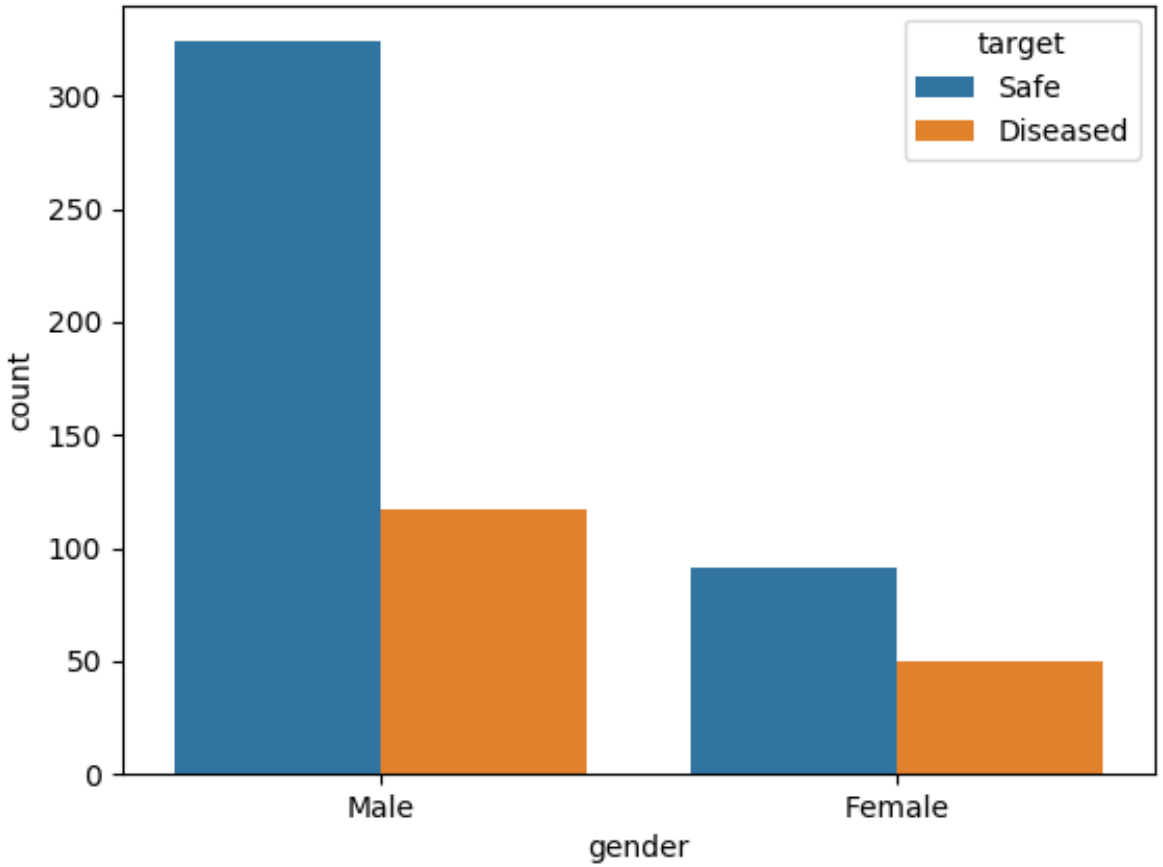
- The Dataset that we would be using is called **Indian Liver Patient Dataset**. It is a popular dataset referenced in literature various times through different research papers on the classification algorithms. There are total **583** records out of which **416** have the disease and **167** do not.

DATASET

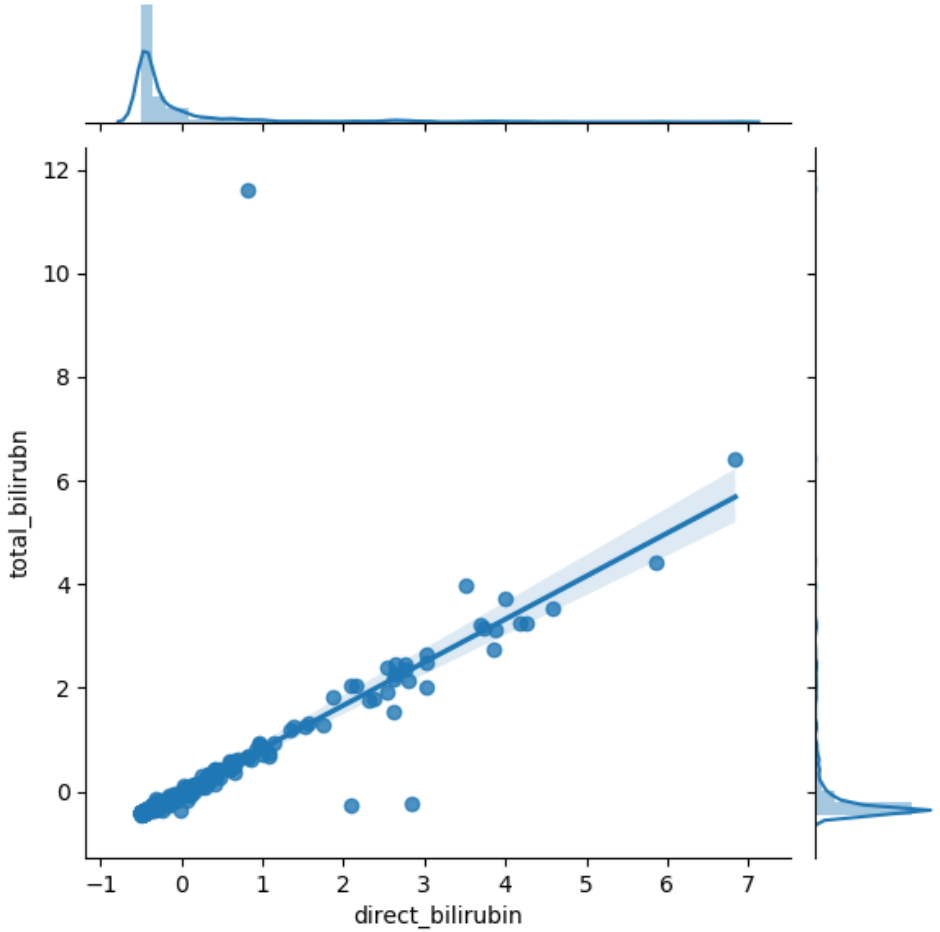
- The features present in the dataset are as follows:

- 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 Proteins
- 9. **ALB** - Albumin
- 10. **A/G** - Ratio Albumin and Globulin Ratio
- 11. **Target** - Selector field used to split the data into two sets (labeled by the experts)

GENDER DISTRIBUTION

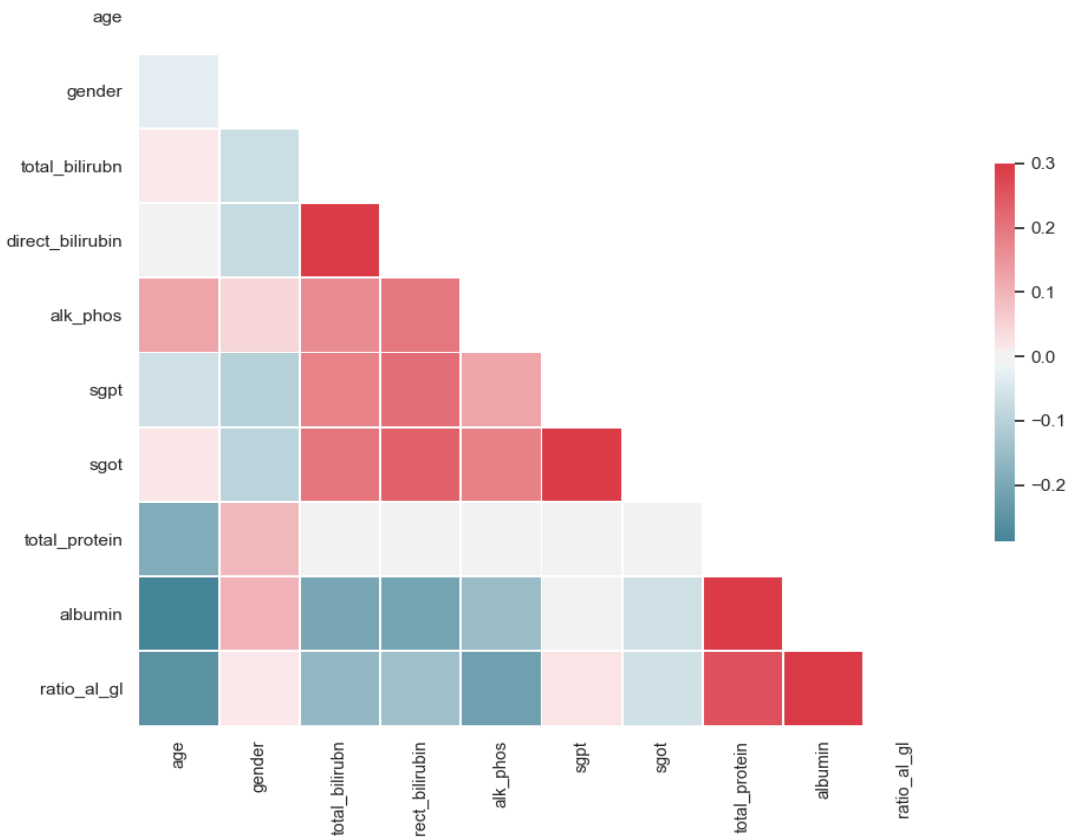


TOTAL BILIRUBIN V DIRECT BILIRUBIN



CORRELATION MATRIX

- We made a correlation matrix that depicts how some of the features are related to each other, if at all. High correlation among many features imply redundancy in the dataset.



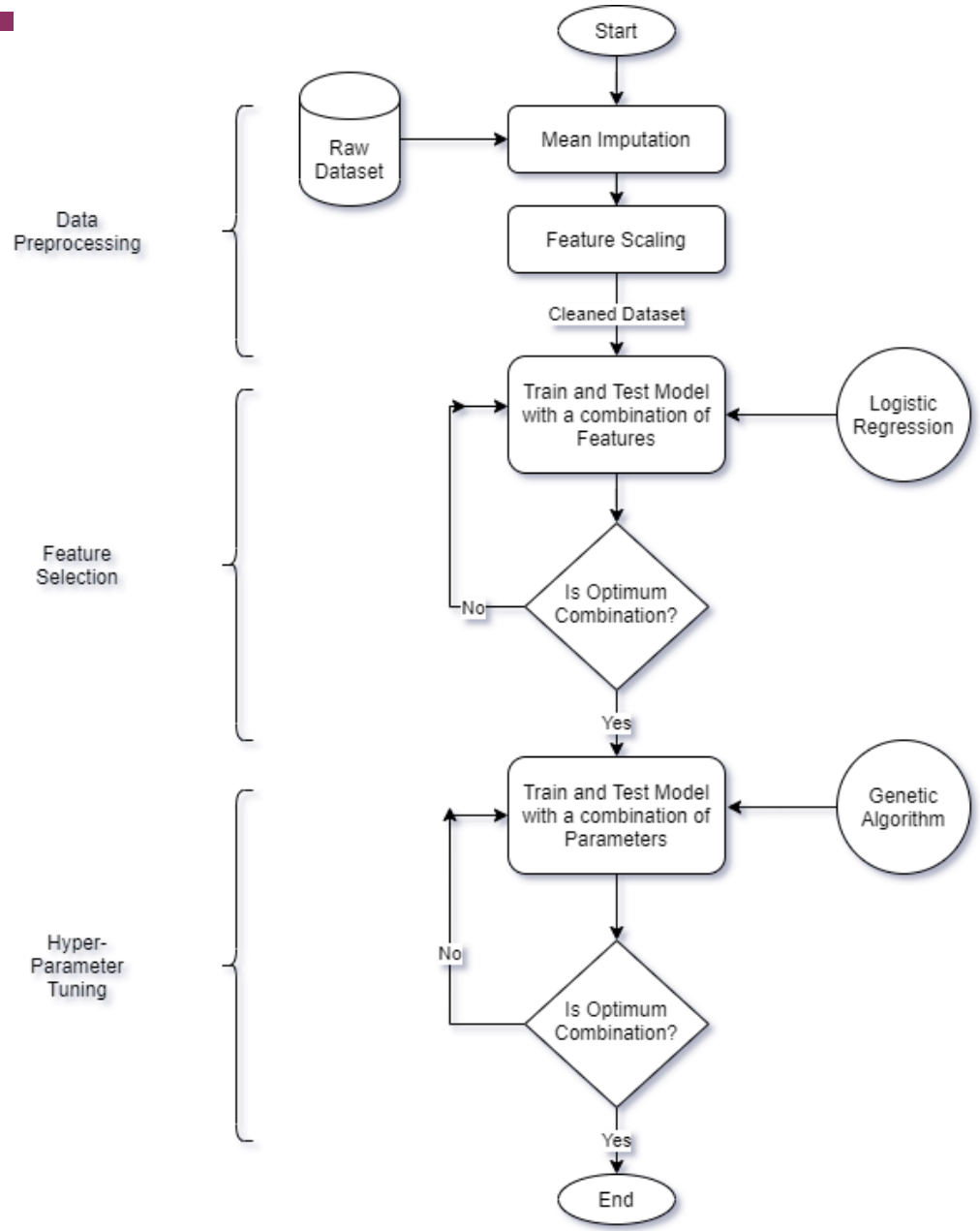
IMPLEMENTATION

PIPELINE

- Components of our pipeline:

1. **Data Pre-processing**: In this step, the input data is cleaned and processed in a way that enables our machine learning model to run efficiently.
2. **Feature Selection**: In this step, 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.
3. **Model Optimization:** In this step, we carried out the process of hyper-parameter tuning that searches for the optimum combination of the logistic regression parameters. We are using Genetic Algorithm to get the best model(which consists of the logistic regression parameters) in reasonable amount of time.

DIFFERENT STAGES OF THE PIPELINE



DATA PRE-PROCESSING

In this stage, we clean our dataset and transform it in a way that helps the model.

It involves:

- Encoding String values to Integer
- Mean Imputation
- Feature Scaling using a Standard Scaler

FEATURE SELECTION

- 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.

```
while no_of_columns < max_columns

    all_combinations = every combination of
                       the current no_of_columns

    # 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
```

LOGISTIC REGRESSION

- Logistic regression is a supervised machine learning algorithm used for solving classification problems.

For example,

1. To predict whether an email is spam (1) or (0)
2. Whether the tumor is malignant (1) or not (0)

the logistic regression model uses the sigmoid function to squeeze the output of a linear hypothesis (θTx) between 0 and 1.

$$\text{sig}(t) = \frac{1}{1+e^{-t}}$$

Here "sig(t)" represents the sigmoid function and "t" is the linear hypothesis,

The output of sigmoid funcition is the estimated probabilty that the "y = 1" (i.e the prediction is positive)

for a given "x" (input)

LOGISTIC REGRESSION - COST FUNCTION

- The Cost Function represent the optimization objective. Our aim is to minimize this cost function so that we can develop an accurate model with minimum error.

$$Cost(h_\theta(x), y) = \begin{cases} -log(h_\theta(x)) & \text{if } y = 1 \\ -log(1 - h_\theta(x)) & \text{if } y = 0 \end{cases}$$

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m} \text{Cost}(h_\theta(x^{(i)}), y^{(i)})$$
$$= -\frac{1}{m}[\sum_{i=1}^{m} y^{(i)} \log(h_\theta(x^{(i)})) + (1 - y^{(i)}) \log(1 - h_\theta(x^{(i)}))]$$

LOGISTIC REGRESSION - MODEL

After selecting features our logistic regression model consist of the following parameters:

- **C** ***default value=1.0 :*** Inverse of regularization strength; must be a positive float. Like in support vector machines, smaller values specify stronger regularization.
- **tol*float, default=1e-4 :***Tolerance for stopping criteria.
- **penalty{*'l1', 'l2', 'elasticnet', 'none'*}, *default='l2' :***Used to specify the norm used in the penalization.
- **solver{*'newton-cg', 'lbfgs', 'liblinear', 'sag', 'saga'*}, *default='lbfgs':*** Algorithm to use in the optimization problem.

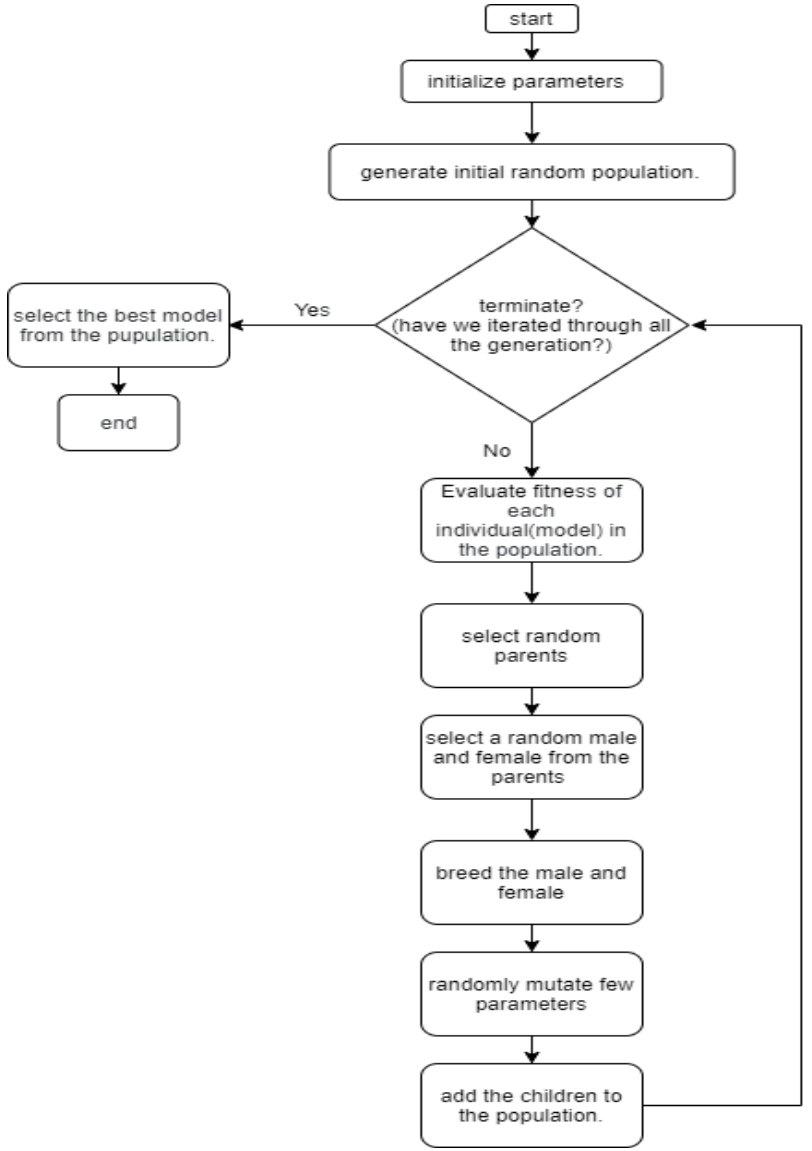
GENETIC ALGORITHM

- A Genetic Algorithms is a type of optimization algorithms. It is a stochastic method for function optimization based on the mechanics of natural genetics and biological evolution.
- A **genetic algorithm** is a search heuristic that is inspired by Charles Darwin's theory of natural evolution. This algorithm reflects the process of natural selection where the fittest individuals are selected for reproduction in order to produce offspring of the next generation.

HYPER-PARAMETER TUNING USING GENETIC ALGORITHM

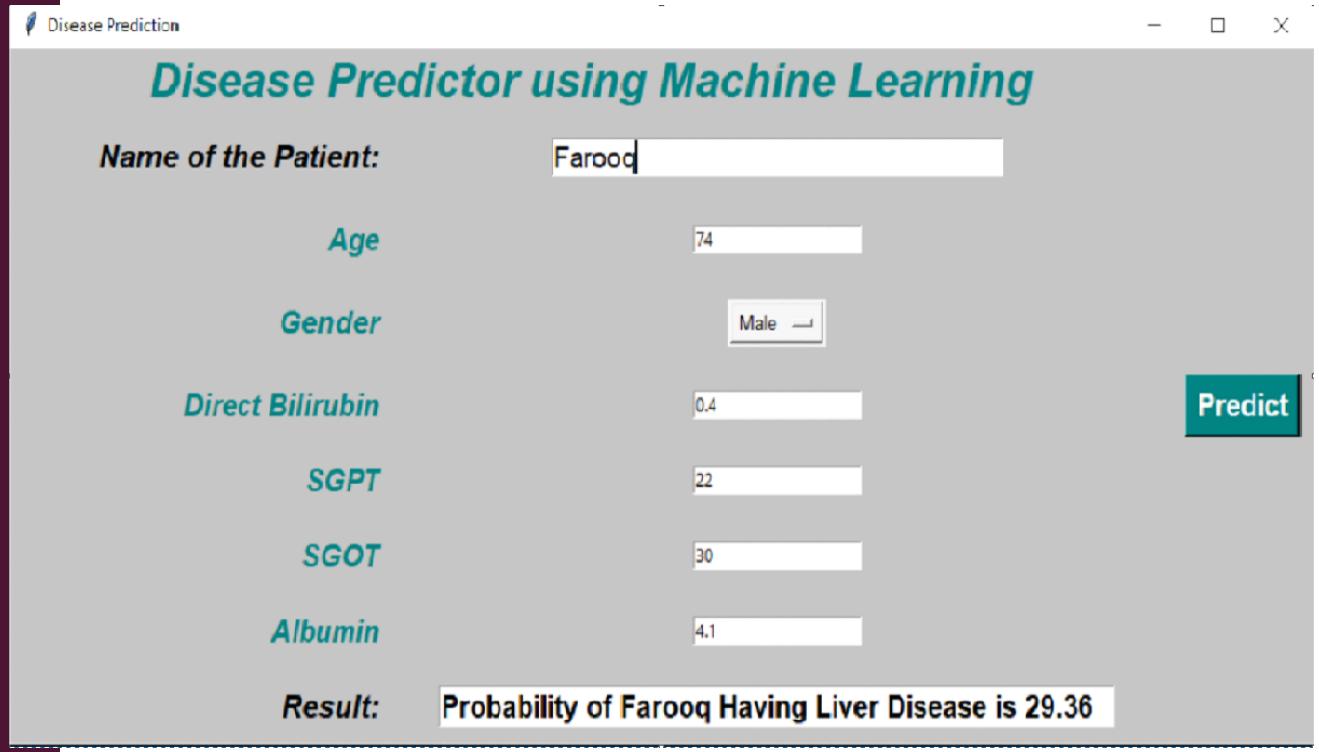
- Components of Genetic Algorithm.

1. Random initial population.
2. Fitness Function
3. Selection of random parents
4. Breeding
5. Mutation



GRAPHICAL USER-INTERFACE

- In order to make our prediction model interactive and responsive to a query based interface, we used ‘tkinter’ to make a Graphical User Interface to the model



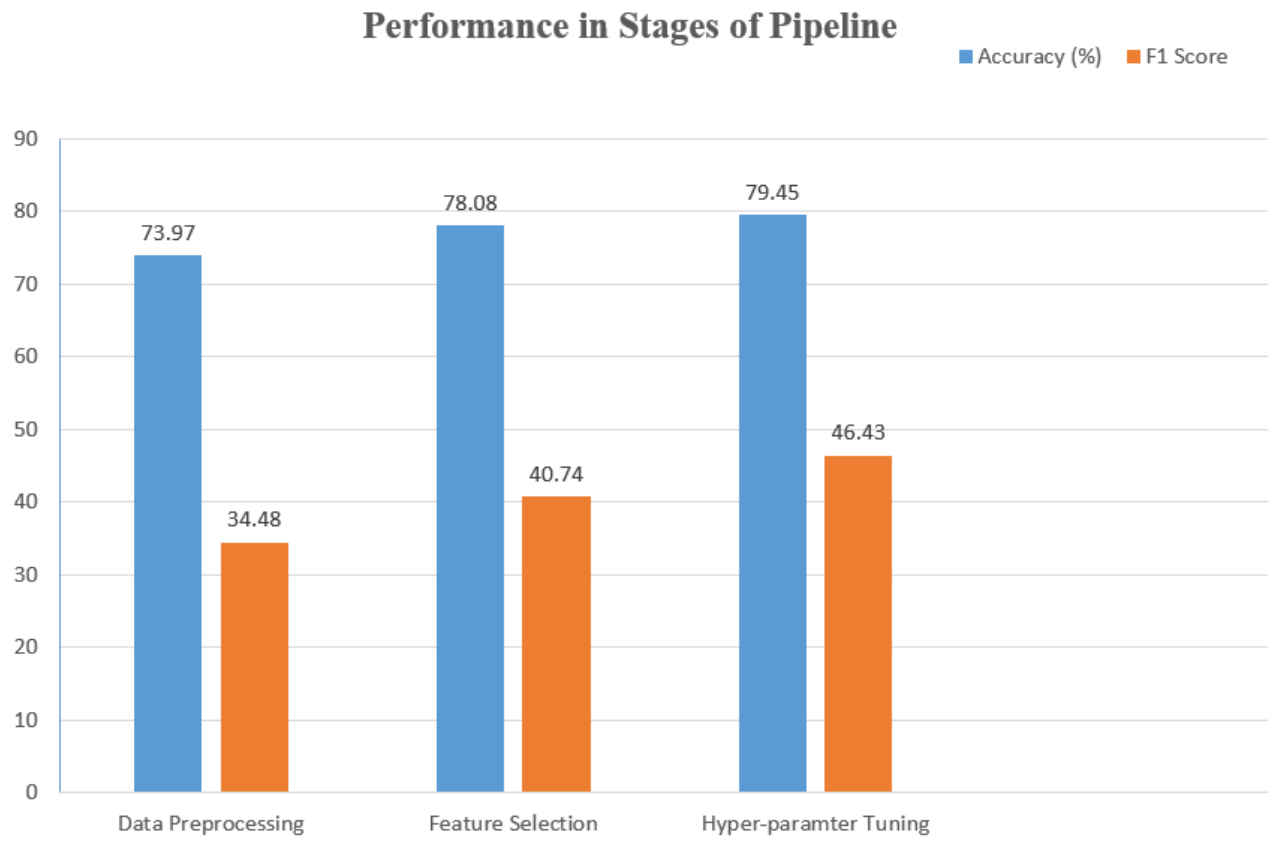
RESULTS

ACCURACY

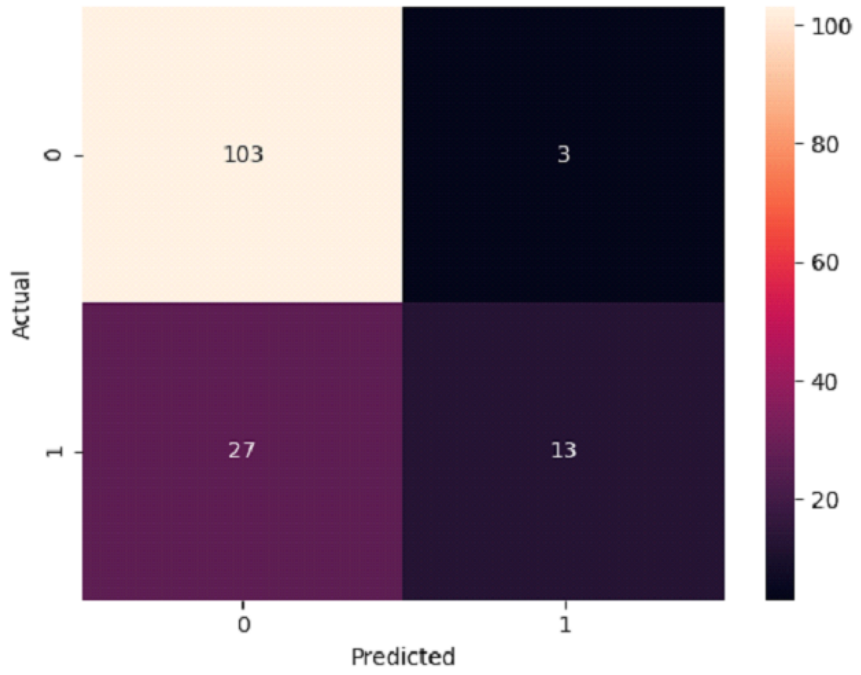
- The model was trained on the aforementioned 436 data points, and then tested on the 'unknown' 146 data points.

- When we ran our optimized Logistic Regression model, it returned with a 79.45% accuracy. It implies that out of the 146 cases that we tested, the model could correctly predict whether the person with those characteristics would get a heart disease or not 116 times

RESULTS


Performance in Stages of Pipeline
Accuracy (%)
F1 Score
90
80
70
60
50
40
30
20
10
0
73.97
34.48
78.08
40.74
79.45
46.43
Data Preprocessing
Feature Selection
Hyper-paramter Tuning

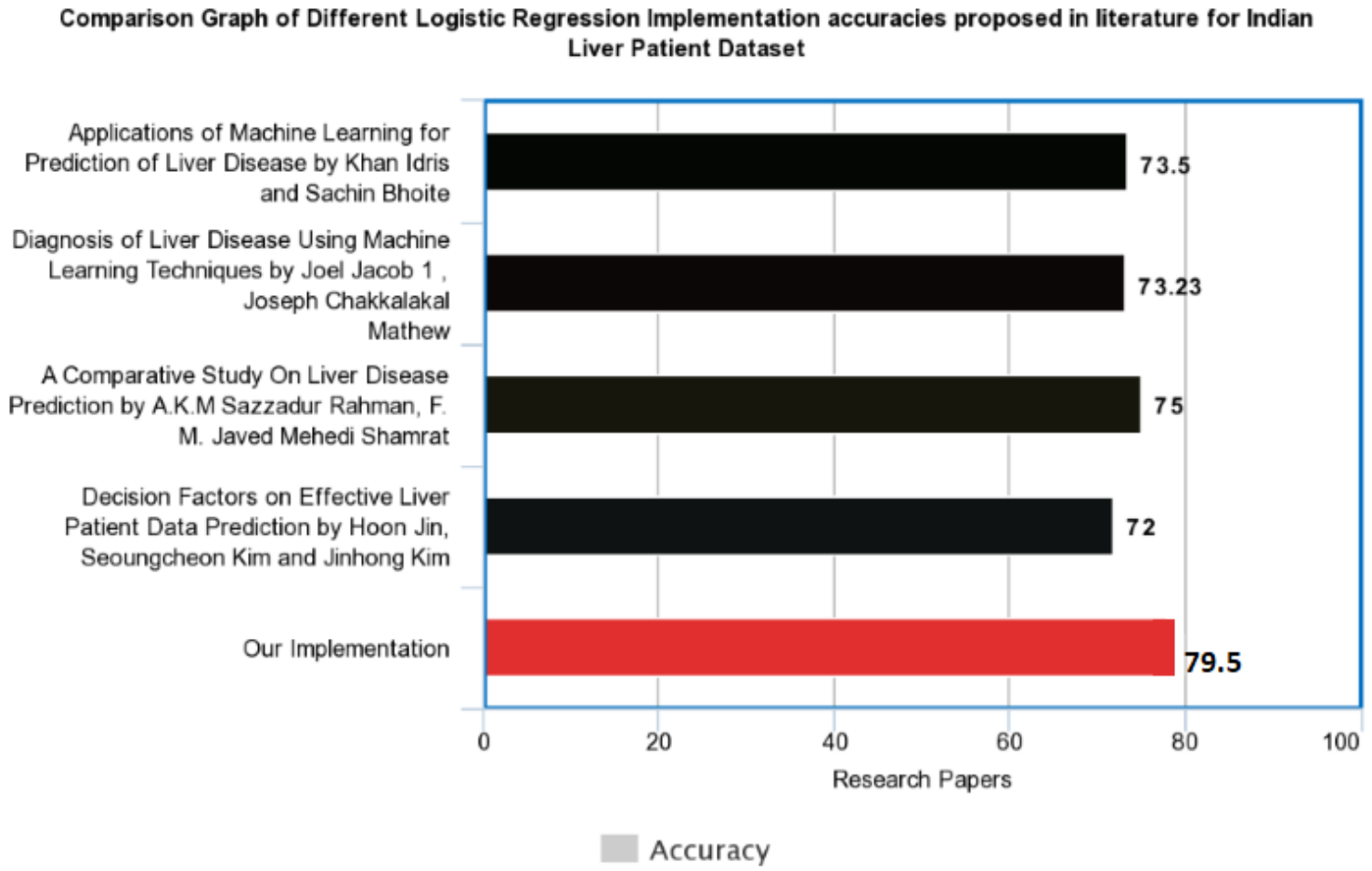
CONFUSION MATRIX



Confusion Matrix is a metric used for prediction models, some relevant terms are as follows:

- True Positive: Actually positive, and predicted to be positive.
- True Negative: Actually negative, and predicted to be negative
- False Positive: Actually negative, and predicted to be positive.
- False Negative: Actually positive, and predicted to be negative.

COMPARATIVE STUDY



CONCLUSION

- As we can observe, there is plenty of scope in the field of Bio-Medical Informatics for us, as budding Computer Science students to explore and implement ground-breaking algorithms & machine learning pipelines to extract meaningful inferences that can directly help people.
  It is our conviction that we would contribute to this ever-growing field and improve our initial understandings.

ACKNOWLEDGMENT

We would like to thank our mentor Mr. Sabyasachi Mukherjee Sir, for his continuous guidance during the preparation for this topic.
We would like to thank the Head of Computer Science & Engineering department, Dr. Monish Chatterjee for providing us with the opportunity to explore such technological fields.

REFERENCES

1. Dataset was taken from the popular UCI Machine Learning Repository
http://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/

2. The Python in-built machine learning module "Scikit-Learn"
https://scikit-learn.org/stable/

3. Understanding of the Algorithms and various Data Science costructs from the popular MOOC by Andrew Ng
https://www.coursera.org/learn/machine-learning-with-python

4. Genetic Algorithm Implementation
https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-15-S16-S11

5. More information on Genetic Algorithm
https://blog.coast.ai/lets-evolve-a-neural-network-with-a-genetic-algorithm-code-included-8809bece164