 Department of Computer Science and Engineering

*Minor Project Report*

A NEW PARADIGM of MACHINE LEARNING in BIOINFORMATICS

Asansol Engineering College

Vivekananda Sarani, Asansol - 713305

Submitted to Asansol Engineering College in Partial Fulfilment for the degree of

**Bachelor of Technology**

(Computer Science and Engineering)

of

Maulana Abul Kalam Azad University of Technology

KOLKATA – 700064

*By*

**Arkadeep Bagal** – 10800116108

**Farooq Ansari** – 10800116102

**Rohan Kumar Singh** – 10800116062

**Rohit Kumar Majee** – 10800116061

*Under the guidance of*

**Mr. Sabyasachi Mukherjee**

*Assistant Professor*

*Department of Computer Science and Engineering*

**Department of Computer Science and Engineering**

**Asansol Engineering College**

**Asansol**

*Certificate*

***I hereby recommend that the thesis entitled “A New Paradigm of Machine Learning in Bioinformatics” submitted by***

***Arkadeep Bagal (Roll No. – 10800116108, Reg. No. - 16100),***

***Farooq Ansari (Roll No. 10800116102),***

***Rohan Kumar Singh (Roll No. 10800116062) and***

***Rohit Kumar Majee (Roll No. 10800116063),***

***has been carried out under my guidance and supervision and may be accepted in partial fulfillment for the award of the degree of Bachelor of Technology in Computer Science and Engineering of Maulana Abul Kalam Azad University of Technology, Kolkata-700064.***

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Mr. Sabyasachi Mukherjee

Assistant Professor

Department of Computer Science and Engineering

Asansol Engineering College

Asansol-713305, West Bengal.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Dr. MONISH CHATTERJEE

(Head and Associate Professor)

Department of Computer Science and Engineering

Asansol Engineering College

Asansol-713305, West Bengal.

SYNOPSIS

**Scope:**

To help medical professionals diagnose diseases in an improved using statistical analysis done by the help of modern Computer Science tools.

**Domain:**

Machine Learning, Bioinformatics.

**Objective:**

Bioinformatics is a field of study that uses computation to extract knowledge from biological data. It includes the collection, storage, retrieval, manipulation and modelling of data for analysis, visualization or prediction through the development of algorithms and software.

This project aims to leverage the rapidly growing repository of information related to molecular biology by the application of Machine Learning, supervised machine learning specifically,

**Technical Details:**

We have developed a hypothesis that predict the possibility of having a heart disease based one the features provided in the dataset, we have developed an logistic regression model with high accuracy and aim to optimize our model in future to obtain credible predictions.

We started the process of developing our model with Data Preprocessing and Data Visualization,

Data preprocessing is an integral step in Machine Learning as the quality of data and the useful information that can be derived from it directly affects the ability of our model to learn

Data visualization is a technique that uses an array of static and interactive visuals within a specific context to help us understand our dataset better and make sense of large amounts of data. It helps in visualizing patterns, trends and correlations that may otherwise go unnoticed.

We selected a Supervised Machine Learning algorithm i.e. Logistic Regression, trained our model on the cleaned and formatted dataset to obtain the hypothesis.

ACKNOWLEDGEMENT

We express my sincere gratitude to Mr. Sabyasachi Mukherjee, our guide for his affectionate and valuable guidance without whose help the present work could not have been successful. We are also indebted to him as a teacher who introduced us to the topics related to the project.

We thank Dr. Monish Chatterjee, the Head of the Department of Computer Science and Engineering for his constant encouragement and permission to work in the Departmental laboratory and use the various resources of the Department of CSE whenever required without which our work would not have been possible.

We are also grateful to the other teachers of the Department of CSE who have taken the pain of teaching us various core subjects of Computer Science for the last few years. We are also thankful to all other staffs of the Department for clearing the various technical doubts during the Laboratory Sessions which boosted our confidence.

**Arkadeep Bagal Farooq Ansari**

University Roll No. 10800116108 University Roll No. 10800116102

University Reg No. 161080110028 University Reg No. 161080110034

Signature Signature

**Rohan Kumar Singh Rohit Kumar Majee**

University Roll No. 10800116062 University Roll No. 10800116061

University Reg. No. 1610801100074 University Reg. No. 161080110075

Signature Signature

CONTENTS

INTRODUCTION

The exponential growth of the amount of biological data available raises two problems: on one hand, efficient information storage and management and, on the other hand, the extraction of useful information from these data.

The second problem is one of the main challenges in computational biology, which requires the development of tools and methods capable of transforming all these heterogeneous data into biological knowledge about the underlying mechanism. These tools and methods should allow us to go beyond a mere description of the data and provide knowledge in the form of testable models. By this simplifying abstraction that constitutes a model, we will be able to obtain predictions of the system.

There are several biological domains where machine learning techniques are applied for knowledge extraction from data. The figure below shows a scheme of the main biological problems where computational methods are being applied.

Complex experimental data raise two different problems. First, data need to be pre-processed, i.e. modified to be suitably used by machine learning algorithms. Second, the analysis of the data, which depends on what we are looking for. In the case of microarray data, the most typical applications are expression pattern identification, classification and genetic network induction.

In a modelling problem, the ‘learning’ term refers to running a computer program to induce a model by using training data or past experience. Machine learning uses statistical theory when building computational models since the objective is to make inferences from a sample. The two main steps in this process are to induce the model by processing the huge amount of data and to represent the model and making inferences efficiently. It must be noticed that the efficiency of the learning and inference algorithms, as well as their space and time complexity and their transparency and interpretability, can be as important as their predictive accuracy.

The process of transforming data into knowledge is both iterative and interactive. The iterative phase consists of several steps. In the first step, we need to integrate and merge the different sources of information into only one format. By using data warehouse techniques, the detection and resolution of outliers and inconsistencies are solved. In the second step, it is necessary to select, clean and transform the data. To carry out this step, we need to eliminate or correct the uncorrected data, as well as decide the strategy to impute missing data. This step also selects the relevant and non-redundant variables; this selection could also be done with respect to the instances.

In the third step, called data mining, we take the objectives of the study into account in order to choose the most appropriate analysis for the data. In this step, the type of paradigm for supervised or unsupervised classification should be selected and the model will be induced from the data. Once the model is obtained, it should be evaluated and interpreted—both from statistical and biological points of view—and, if necessary, we should return to the previous steps for a new iteration.

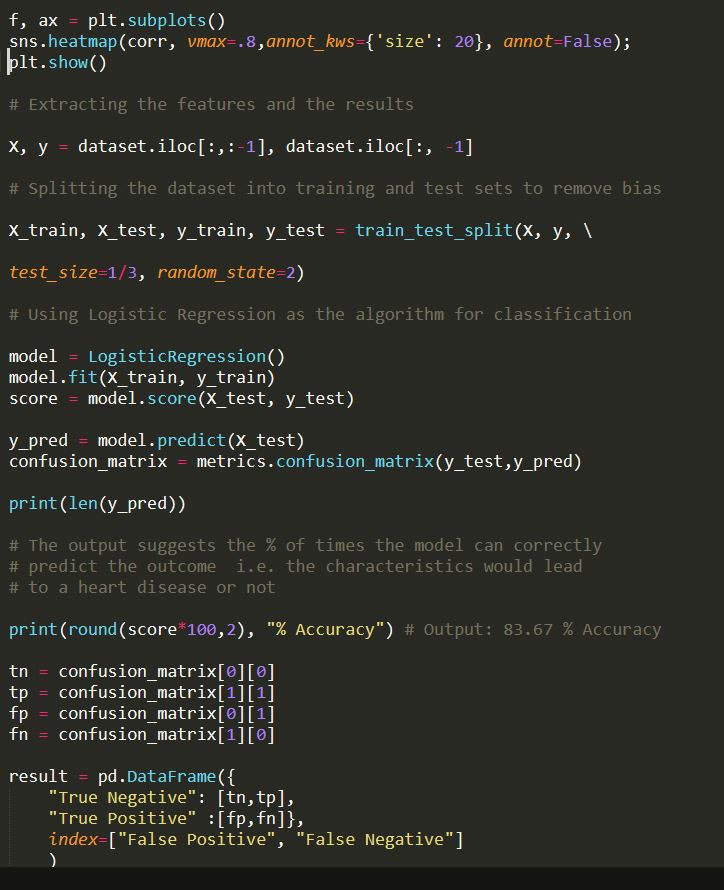
This includes the solution of conflicts with the current knowledge in the domain. The model satisfactorily checked—and the new knowledge discovered—are then used to solve the problem.

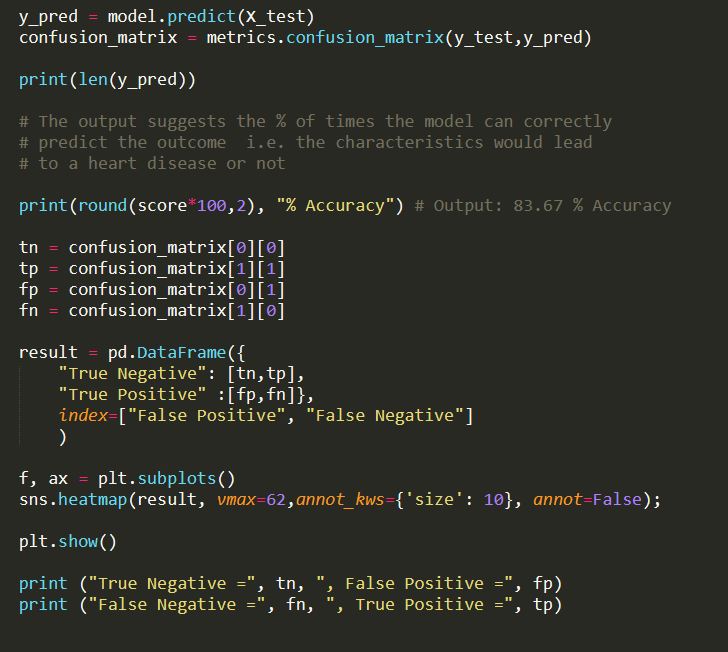
PROJECT DETAILS

IMPLEMENTATION

To demonstrate the intention of this project, we implemented a predictive model on medical dataset using Logistic Regression.







*Figure: Source code of the implementation using Python*

RESULT

In our implementation, we obtained some positive results.

We split our original dataset of **293** data points (with 14 attributes each) into two sets of data, a training set of size **195** data points and a test set of **98**. We followed a split ratio of 1/3.

The model was trained on the aforementioned 195 data points, and then tested on the ‘unknown’ 98 data points.

When we ran a Logistic Regression model, it returned with a **83.67%** accuracy. It implies that out of the 98 cases that we tested, the model could **correctly** predict whether the person with those characteristics would get a heart disease or not 82 times.

We can see a breakdown of those results in the following Heat-map of the **Confusion Matrix.**

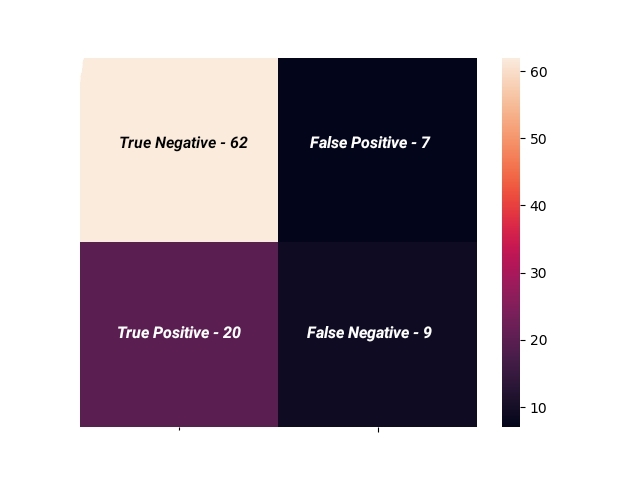
*Confusion Matrix is a metric used for prediction models, some relevant terms are as follows:*

**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.



*Figure: Confusion Matrix for the predicted results*

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.

With the help of Modern Computer Science tools, we can use the abundance of data available in this “Age of Information” to

Find patterns in raw numbers and help Medical Professionals take potentially life-saving decisions.

Not only in disease diagnosis; we can take the help of Machine Learning and Data Analysis in the field of Genetics too. As the DNA information that defines the blueprint of the Human Body, is a seemingly never-ending well of raw data and information that is in dire need of prediction models and analytical tools.

It is our conviction that we would contribute to this ever-growing field and improve our initial understandings by doing further research in the next phase of the project.

FUTURE SCOPE

Some possibilities for the future include:

1. ***Visualization:*** We intend to make the results, our analysis, the dataset and the Algorithm more transparent using various modules that are available in Python Languege.
2. ***Accuracy:*** Even though the accuracy 83.67% in the field of Data Science is commendable, we would like to use different machine learning algorithms with proper parameters to see how they perform, and if there’s room for further improvement.
3. ***Data Cleaning:*** We intend to apply methods such as Feature Scaling and Regularization and remove outliers to utilize our dataset in a better way.
4. ***Domain Specific Research:*** We would also like to understand the “bio” of the Bioinformatics better, and learn how different features of the dataset correlated with each other so we can implement a model in way that would be most suited for the subject matter.