DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
ASANSOL ENGINEERING COLLEGE
VIVEKANANDA SARANI, ASANSOL - 713305

Minor Project Report

a NEW PARADIGM of MACHINE LEARNING in BIOINFORMATICS

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. 161080110028***),***
Farooq Ansari (Roll No. 10800116102, Reg No. 161080110034***),***
Rohan Kumar Singh (Roll No. 10800116062, Reg. No. 161080110074***) and***
Rohit Kumar Majee (Roll No. 10800116063, Reg. No. 161080110075***),***

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.

CONTENTS

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.

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

Definitions and Theories:

What is machine Learning?

Machine Learning enables IT systems to recognize patterns on the basis of existing algorithms and data sets and to develop adequate solution concepts. Therefore, in Machine Learning, artificial knowledge is generated on the basis of experience.

Arthur Samuel's Definition:

Machine Learning is a science of getting computers to learn without being explicitly programmed.

Tom Mitchell's Definition:

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

Advantages of Machine Learning:

Self-learning machines can perform complex tasks. These include, for example, the recognition of error patterns. This is a major advantage, especially in areas such as the manufacturing

industry. The industry relies on continuous and error-free production. While even experts often cannot be sure where and by which correlation a production error in a plant fleet arises, Machine Learning offers the possibility to identify the error early - this saves downtimes and money.

Self-learning programs are now also used in the medical field. In the future, after "consuming" huge amounts of data (medical publications, studies, etc.), apps will be able to warn a in case his doctor wants to prescribe a drug that he cannot tolerate. This "knowledge" also means that the app can propose alternative options which for example also take into account the genetic requirements of the respective patient.

Types of Machine Learning:

Supervised Learning

Supervised learning is when the model is getting trained on a labelled dataset. **Labelled** dataset is one which have both input and output parameters. In this type of learning both training and validation datasets are labelled
Example:
Linear Regression
Logistic Regression

Types of Supervised Learning:

Classification:

It is a Supervised Learning task where output is having defined labels (discrete value).

Example:

Given a patient with a tumor, we have to predict whether the tumor is malignant or benign.

Regression:

Regression predictive modeling is the task of approximating a mapping function (f) from input variables (X) to a continuous output variable (y).

Example:

Given data about the size of the house on the real estate market, try to predict their price. Price as a function of size is a continuous output, so this is a regression problem.

Logistic regression:

The logistic regression paradigm is defined as where $\boldsymbol{x}$ represents an instance to be classified, and $\beta_0, \beta_1, \ldots, \beta_n$ are the parameters of the model. These parameters should be estimated from the data in order to obtain a concrete model. The parameter estimation is performed by means of the maximum likelihood estimation method. The system of $n + 1$ equations and $n + 1$ parameters to be solved does not have an analytic solution. Thus, the maximum

likelihood estimations are obtained in an iterative manner. The Newton–Raphson procedure is a standard in this case.

The modelling process is based on the Wald test and on the likelihood ratio test. The search in the space of models is usually done with forward, backward or stepwise approaches.

Unsupervised Learning

Unsupervised learning is the training of machine using information that is neither classified nor labeled and allowing the algorithm to act on that information without guidance. Here the task of machine is to group unsorted information according to similarities, patterns and differences without any prior training of data.

Example:
Let's, take the case of a baby and her family dog. She knows and identifies this dog. Few weeks later a family friend brings along a dog and tries to play with the baby. Baby has not seen this dog earlier. But it recognizes many features (2 ears, eyes, walking on 4 legs) are like her pet dog. She identifies the new animal as a dog. This is unsupervised learning, where you are not taught but you learn from the data (in this case data about a dog.) Had this been supervised learning, the family friend would have told the baby that it's a dog.

Types of Unsupervised Learning

Unsupervised learning problems further grouped into clustering and association problems.

Clustering

Clustering is an important concept when it comes to unsupervised learning. It mainly deals with finding a structure or pattern in a collection of uncategorized data. Clustering algorithms will process your data and find natural clusters (groups) if they exist in the data. You can also modify how many clusters your algorithms should identify. It allows you to adjust the granularity of these groups.

K-means clustering

K means it is an iterative clustering algorithm which helps you to find the highest value for every iteration. Initially, the desired number of clusters are selected. In this clustering method, you need to cluster the data points into k groups. A larger k means smaller groups with more granularity in the same way. A lower k means larger groups with less granularity.

The output of the algorithm is a group of "labels." It assigns data point to one of the k groups. In k-means clustering, each group is defined by creating a centroid for each group. The centroids are like the heart of the cluster, which captures the points closest to them and adds them to the cluster.

Data Pre-processing

Pre-processing refers to the transformations applied to our data before feeding it to the algorithm.
Data Pre-processing is a technique that is used to convert the raw data into a clean data set. In other words, whenever the data is gathered from different sources it is collected in raw format which is not feasible for the analysis.

Need of Data Pre-processing

• For achieving better results from the applied model in Machine Learning projects the format of the data has to be in a proper manner. Some specified Machine Learning model needs information in a specified format, for example, Random Forest algorithm does not support null values, therefore to execute random forest algorithm null values have to be managed from the original raw data set.
• Another aspect is that data set should be formatted in such a way that more than one Machine Learning and Deep Learning algorithms are executed in one data set, and best out of them is chosen.

Best practices of Data Cleaning:

- Setting up of quality plan
- Fill out missing values

- Removing rows with missing values.
- Fixing errors in the structure.
- Reducing data for a proper data handling.

Data Cleaning is a critical process for the success of any machine learning function. For most machine learning projects, about 80 percent of the effort is spent on data cleaning

Data Visualization

Data visualization is a technique that uses an array of static and interactive visuals within a specific context to help people understand and make sense of large amounts of **data**.
The **data** is often displayed in a story format that visualizes patterns, trends and correlations that may otherwise go unnoticed.

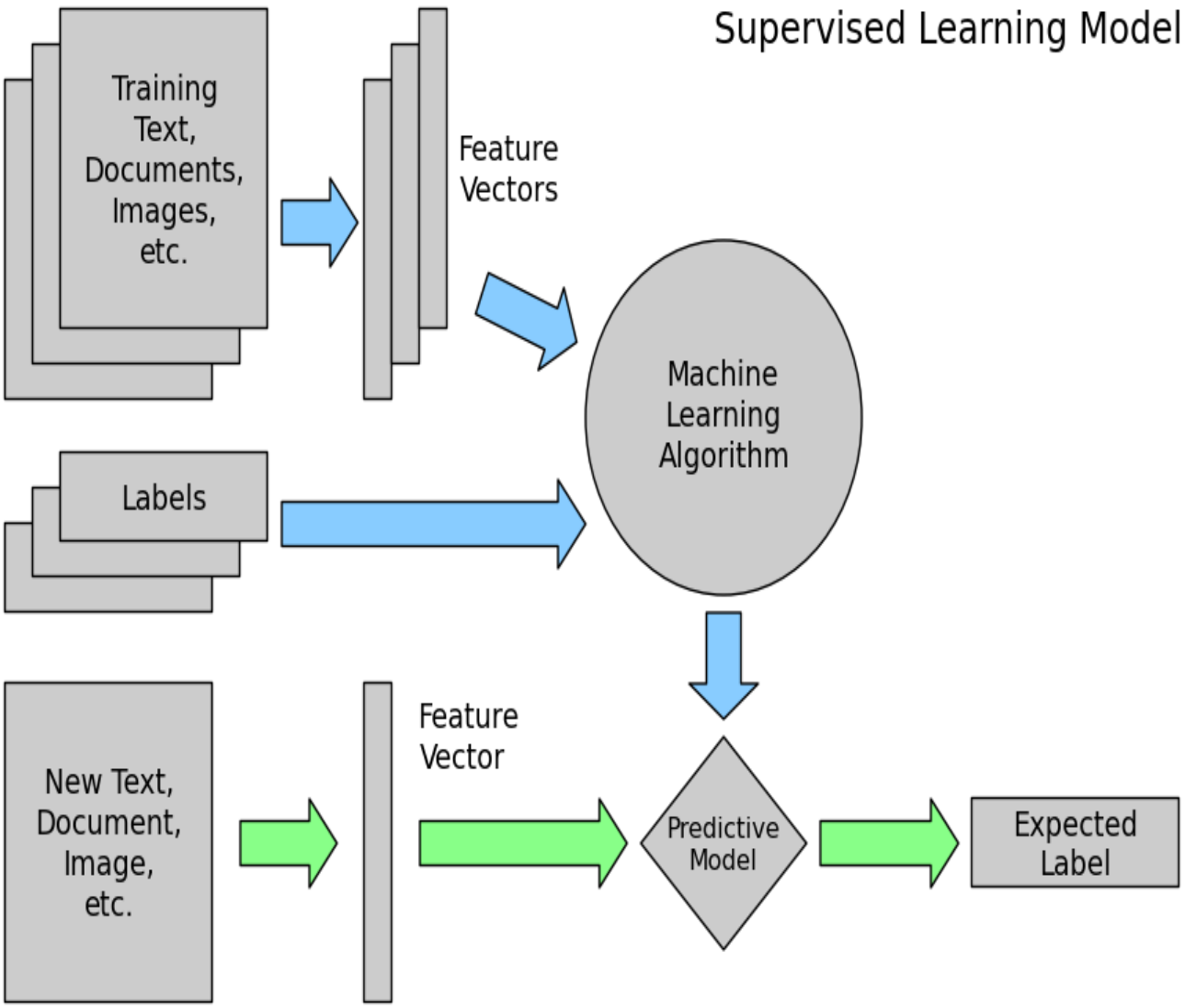


*Fig: **Data workflow for Machine Learning***

Logistic Regression

It is a classification algorithm.

A solution for classification is logistic regression. Instead of fitting a straight line or hyperplane, the logistic regression model uses the logistic function to squeeze the output of a linear equation between 0 and 1.

In such problems we have to predict whether the result is 0 or 1.

Logistic regression: $0 <= h\ \theta(x) <= 1$

$Y \in \{0,1\}$

0: for negative class (Diagnosis: Healthy)

1: for positive class(Diagnosis: Unhealthy)

Hypothesis representation for Logistic regression

$h\ \theta(x) = \theta^T x$ such that $0 <= h\ \theta(x) <= 1$

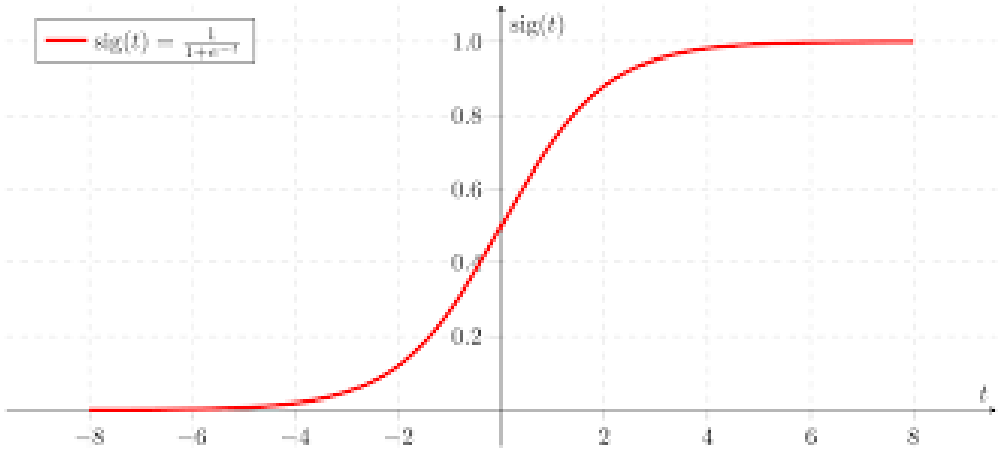
this is accomplished by putting $\theta^T x$ into logistic function.

New form uses sigmoidal function or logistic function

$h\ (\theta) = g(\theta^T x)$

$g(z) = 1/1+e^z$ where $z = \theta^T x$

the function g(z) looks like this:

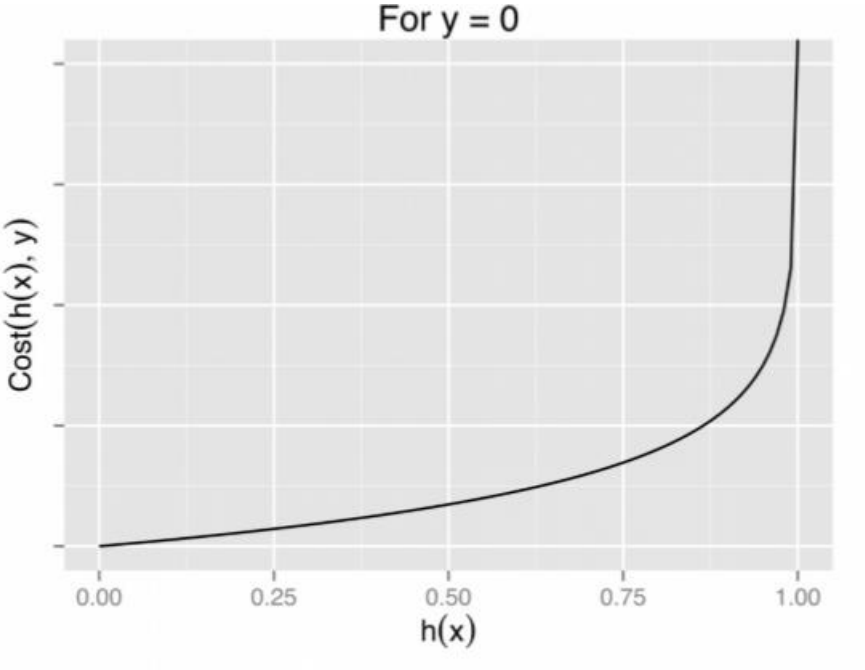


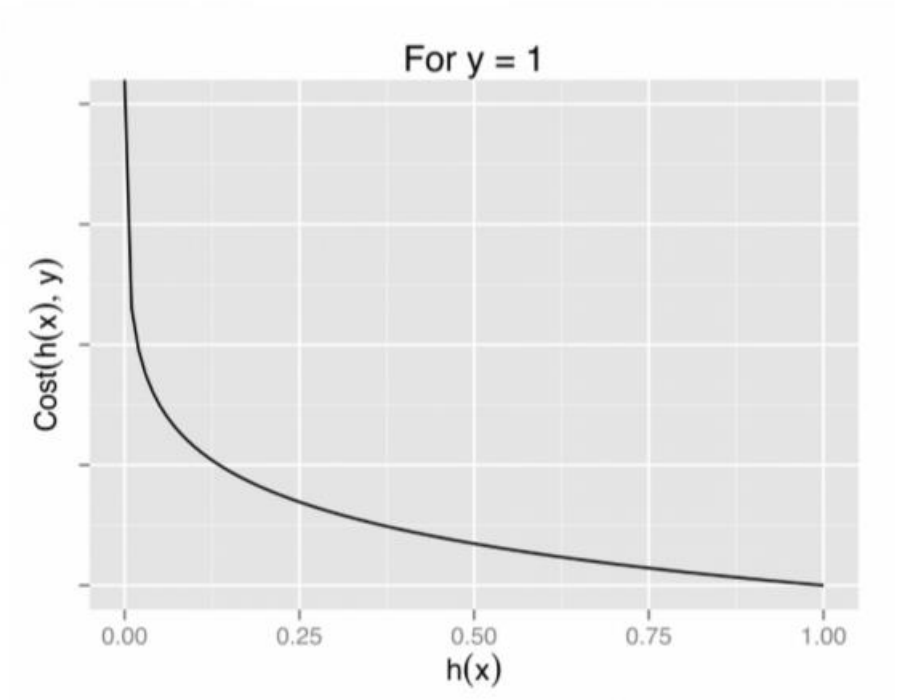
Threshold classifier output h θ(x) at 0.5:
If h θ(x) >= 0.5, predict y=1
If h θ(x) <= 0.5, predict y=0

Now, if hθ (x) will give the probability that our output is 1 or 0.
Example: 0.7 means the probability of having output 1 is 70%

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

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


For y = 0
Cost(h(x), y)
0.00
0.25
0.50
0.75
1.00
h(x)


For y = 1
Cost(h(x), y)
0.00
0.25
0.50
0.75
1.00
h(x)

Want $\min_\theta J(\theta)$:

Repeat $\{$

$$\theta_j := \theta_j - \alpha \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) x_j^{(i)}$$

(simultaneously update all θ_j)

$\}$

System Requirements

We ran our model in a 64-bit Intel i5 processor with 8 GB or RAM running Windows OS.

The program was written in Python 3.7, with the help of various modules such as scikit-learn, pandas, numpy and seaborn.

The Editor used was Sublime Text 3.

IMPLEMENTATION

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

```
import pandas as pd
import numpy as np
from sklearn.impute import SimpleImputer
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split
import seaborn as sns
import matplotlib.pyplot as plt
import sklearn.metrics as metrics

# info about the features used

header_row = ['age','sex','pain','BP','chol','fbs','ecg', \
'maxhr','eiang','eist','slope','vessels','thal','diagnosis']

# Loading the dataset

dataset = pd.read_csv("processed.hungarian.data")

print(dataset.shape)

# Replacing the missing values with the Mean  value of its column

dataset = dataset.replace("?",np.NaN)
imp = SimpleImputer(missing_values=np.NaN, strategy='mean')
imp.fit(dataset)
dataset = pd.DataFrame(imp.transform(dataset))
dataset.columns = header_row

df = dataset.iloc[:,:-1]
corr = df.corr()


f, ax = plt.subplots()
sns.heatmap(corr, vmax=.8,annot_kws={'size': 20}, annot=False);
plt.show()

# Extracting the features and the results

X, y = dataset.iloc[:,:-1], dataset.iloc[:, -1]

# Splitting the dataset into training and test sets to remove bias
```

```
f, ax = plt.subplots()
sns.heatmap(corr, vmax=.8,annot_kws={'size': 20}, annot=False);
plt.show()

# Extracting the features and the results

X, y = dataset.iloc[:,:-1], dataset.iloc[:, -1]

# Splitting the dataset into training and test sets to remove bias

X_train, X_test, y_train, y_test = train_test_split(X, y, \

test_size=1/3, random_state=2)

# Using Logistic Regression as the algorithm for classification

model = LogisticRegression()
model.fit(X_train, y_train)
score = model.score(X_test, y_test)

y_pred = model.predict(X_test)
confusion_matrix = metrics.confusion_matrix(y_test,y_pred)

print(len(y_pred))

# The output suggests the % of times the model can correctly
# predict the outcome  i.e. the characteristics would lead
# to a heart disease or not

print(round(score*100,2), "% Accuracy") # Output: 83.67 % Accuracy

tn = confusion_matrix[0][0]
tp = confusion_matrix[1][1]
fp = confusion_matrix[0][1]
fn = confusion_matrix[1][0]

result = pd.DataFrame({
    "True Negative": [tn,tp],
    "True Positive" :[fp,fn]},
    index=["False Positive", "False Negative"]
    )
```

```
y_pred = model.predict(X_test)
confusion_matrix = metrics.confusion_matrix(y_test,y_pred)

print(len(y_pred))

# The output suggests the % of times the model can correctly
# predict the outcome  i.e. the characteristics would lead
# to a heart disease or not

print(round(score*100,2), "% Accuracy") # Output: 83.67 % Accuracy

tn = confusion_matrix[0][0]
tp = confusion_matrix[1][1]
fp = confusion_matrix[0][1]
fn = confusion_matrix[1][0]

result = pd.DataFrame({
    "True Negative": [tn,tp],
    "True Positive" :[fp,fn]},
    index=["False Positive", "False Negative"]
    )

f, ax = plt.subplots()
sns.heatmap(result, vmax=62,annot_kws={'size': 10}, annot=False);

plt.show()

print ("True Negative =", tn, ", False Positive =", fp)
print ("False Negative =", fn, ", True Positive =", tp)
```

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.

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
University Roll No. 10800116108
University Reg. No. 161080110028
Signature

Farooq Ansari
University Roll No. 10800116102
University Reg. No. 161080110034
Signature

Rohan Kumar Singh
University Roll No. 10800116062
University Reg. No. 161080110074
Signature

Rohit Kumar Majee
University Roll No. 10800116061
University Reg. No. 161080110075
Signature

BIBLIOGRAPHY

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