AY 2022-23
PROJECT REPORT
ON

Breast Cancer Prediction

Submitted for

ECL305: PRACTICUM-III

By
DHRUV SRIVASTAVA (21215)
III SEMESTER
BTECH-ECE

SCHOOL OF ELECTRONICS

INDIAN INSTITUTE OF INFORMATION TECHNOLOGY UNA
HIMACHAL PRADESH

DECEMBER 2022

BONAFIDE CERTIFICATE

This is to certify that the project titled Breast Cancer Detection using machine and deep learning is a bonafide record of the work done by

Dhruv Srivastava (21215)

in partial fulfilment of the requirements for the award of the degree of Bachelor of Technology in ECE of the INDIAN INSTITUTE OF INFORMATION TECHNOLOGY UNA, HIMACHAL PRADESH, during the year 2022- 2023.

Under the guidance of

Ms. Pranshu Sharma

Faculty,

School of Electronics

Indian Institute of Information Technology Una

Saloh Una, H.P.

Project viva-voce held on:17/12/2022

Internal Examiner External Examiner

ORIGNALITY / NO PLAGARISM DECLARATION

We certify that this project report is our original report and no part of it is copied from any published reports, papers, books, articles, etc. We certify that all the contents in this report are based on our personal findings and research and we have cited all the relevant sources which have been required in the preparation of this project report, whether they be books, articles, reports, lecture notes, and any other kind of document. We also certify that this report has not previously been submitted partially or as whole for the award of degree in any other university in India and/or abroad.

We hereby declare that, we are fully aware of what constitutes plagiarism and understand that if it is found at a later stage to contain any instance of plagiarism, our degrees may be cancelled.

Dhruv Srivastava (21215)

ABSTRACT

In this period of time where every progressing day every technology is getting modernised and remote and same advancement are also felt in the field of medical sciences. But still in to this day the uncertainty of a disease as fatal as cancer, can be nerve-wracking as if left untreated may lead to untimely death.

This above mentioned issue can be helped with the help of clubbing data science with medical sciences and provide benefit to all free of cost by having a profound knowledge of cancer and machine learning.

***Keywords*:** Machine learning, Deep learning, cancer,healthcare.

ACKNOWLEDGEMENT

We would like to thank the following people for their support and guidance without whom the completion of this project in fruition would not be possible.

We would like to express our sincere gratitude and heartfelt thanks to Ms. Pranshu Sharma for their unflinching support and guidance, valuable suggestions and expert advice. Their words of wisdom and expertise in subject matter were of immense help throughout the duration of this project. And her continuous support in this project helped me for keeping the project motivated and emerging.

We also take the opportunity to thank our Director and all the faculty of School of Computing/Electronics, IIIT Una for helping us by providing necessary knowledge base and resources.

We would also like to thank our parents and friends for their constant support.

Dhruv Srivastava (21215)

TABLE OF CONTENTS

LIST OF ACRONYMS

ANN	Artificial Neural Network
DL	Deep learning
ML	Machine Learning
SVM	Support Vector Machine
KNN	K- nearest Neighbours

LIST OF FIGURES

Chapter 1

Introduction

1.1 Cancer Preface

Cancer is a disease in which some of the body's cells grow uncontrollably and spread to other parts of the body. Cancer can start almost anywhere in the human body, which is made up of trillions of cells. Normally, human cells grow and multiply (through a process called cell division) to form new cells as the body needs them. When cells grow old or become damaged, they die, and new cells take their place.

Figure 1: Cancer Patient

Chapter 2

Review of Literature

1. Project Objectives

a. To research on Cancer and predict Breast Cancer.

b. To understand the major features that acts as the root cause of breast cancer.

2. Academic Objectives

a. To explore the domain of ML and DL.

b. To learn new algorithms and understand the math behind it.

c. Will to put whatever learned to some practical use.

3. Rundown to Data Analysis and Machine Learning

The data analysis process is nothing more than gathering information using appropriate applications or tools that allow you to examine the data and find patterns. You can make decisions and draw final conclusions based on this information and data.

Machine Learning is a subbranch of artificial intelligence aims at making a computer learn something with the help of significant parameters by analyzing appropriate correlations and patterns in data and processing it.

Clubbing the two data analysis and ML and DL we can deploy efficient and helpful models and deduce paramount results for predicting and analyzing the information

Chapter 3

Prerequisites & Technologies Learnt

1. Mathematical Background

a. Statistics

b. Calculus

c. Linear Algebra

2. Python Programming

a. Preface

Python is a general-purpose high-level object oriented language used in variety of ways for automating tasks building backend of websites and provides data scientist a platform with huge libraries for Data analysis.

b. Libraries

i. Numpy: Numerical python the mathematical toolbox offering multitude of functions related to random number generator to multidimensional matrices with an efficient performance.

ii. Pandas: The python's data analysis library used for working with relational and labelled data's manipulation.

iii. Matplotlib: It helps in data visualization by numerous different graphs of any given data.

iv. Sklearn: Formally Sci-kit learn a package that acts as the backbone for machine learning or predictive analysis by providing profusion of algorithms, feature selection and engineering methods.

v. Tensorflow and Keras : A google created open source library used for establishing neural networks and building deep model projects.

3. Machine Learning Paradigms

a. Supervised Learning: Supervised learning is the machine learning task of learning a function that maps an input to an output example input output pairs.

b. Unsupervised Learning: Its is a training of machine using information that is neither classified nor labelled along the algorithm to act on that information without guidance

c. Reinforcement Learning: reinforcement learning is a feedback-based machine learning technique in which there is a continuous flow of information or actions which acts as a feed for the next output it is generally used in game playing and robotics.

d. Deep Learning: a type of machine learning paradigm based on artificial neural networks which is used to extract features of higher significance using the methods of dimensionality reduction.

4. General Procedure for Data Analysis

Figure 1: Data analysis Process

a. Data Collection

The data collected is required to be fed to the computer so it must be reliable and not noisy i.e., meaningless.

b. Feature Selection

   One of the most generic and significant process is selection of correct parameters from the dataset that will be used.

c. Feature Engineering

   In this process the null valued data is removed or modified with measures of central tendencies and scaled appropriately.

d. Training a model

   In this process after analysing the problem statement we decide weather it is a classification or regression problem and based on that we apply the algorithms and train the data and separate 2 portions of it for later use in testing the model accuracy.

e. Evaluating the statistically best Model

   After applying different algorithms we find the best and most optimal algorithm by accuracy measurement and chose it for predicting the required.

Chapter 4
Supervised Machine Learning

1. Regression Models

Statistical modelling regression analysis is a set of statistical process for predicting the relationship among different parameters.

It includes money techniques for modelling and analysing several variables when focus on relationship between a dependent variable and one or more independent variables. more specifically regression analysis helps one understand how the typical value of the dependent variable or criterion variable changes when any one of the independent variables is very read while other independent variables are held fixed.

A. Linear regression: is a linear approach for modelling the relationship between a scalar dependent variable y and an independent variable X.

B. Multiple linear regression: it is a Statistical technique used to predict the outcome of response variable through several explanatory variables and model the relationship between them.

C. Decision tree regression: a decision tree is a graphical representation of all the possible solutions or decision based on few conditions. Decision trees are non-parametric models which means the number of parameters is not predetermined prior to training such models will normally overwrite data in contrast of parametric models such as linear model has a 3 determined number of parameters there were reducing its degrees of freedom this in turn prevents overfitting.

D. Random forest regression: Ensemble learning uses the same algorithm multiple times or a group of different algorithms together to improve the prediction of model random forests are an ensemble decision trees to perform regression task continuously on different trees. Random forest combines

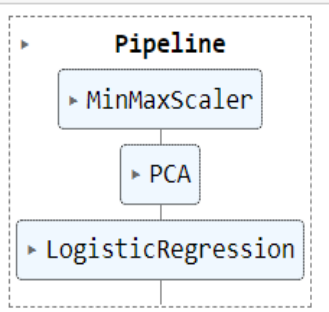
simplicity of decision trees with flexibility resulting in vast improvement in accuracy

2. Classification Models

Classification models are used to classify the data it's specifies the class to which data elements will long to or predicts the class Warren invariable it is best used on the output has finite and discrete values. It's of 2 types generally Linear and Non-linear Models classification.

A. Linear Models

a. Logistic regression: it is similar to linear regression, but it is used to classify the data into 2 parts and creates a squiggle shaped graph. N logistic regression we use the principle of maximum likelihood.




```
from sklearn.metrics import classification_report,accuracy_score
print(classification_report(yte,predpipe1))
print(accuracy_score(yte,predpipe2))
```

	precision	recall	f1-score	support
0	1.00	0.90	0.95	63
1	0.95	1.00	0.97	108
accuracy			0.96	171
macro avg	0.97	0.95	0.96	171
weighted avg	0.97	0.96	0.96	171

Figure 3: Logistic regression

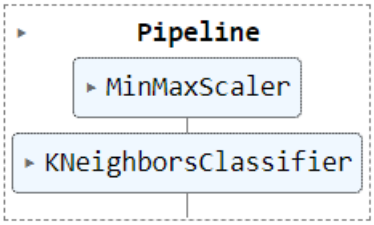
b. Support vector machines: support vector machines are very worst child are also capable of performing linear or non-linear classification regression outlier detection and then was detecting hyperplane with segregate data into different classes and the optimization objective is to find the maximum margin hyperplane that is the farthest from the

closest points in the 2 defined classes which are termed support vectors.

B. Non-Linear Models

a. Naïve Bayes: It works on the base model the conditional probability theorem. Naïve bayes is a generative model while gaussian naïve bayes is a model based on distribution. Naïve bayes interprets that each feature is independent.

b. K-Nearest Neighbours: It is used to assign the sample data point to the cluster based on the similarity measurement. A new data input point is classified in the category such that it has the highest number of neighbours from that category.



	precision	recall	f1-score	support
0	1.00	0.92	0.96	63
1	0.96	1.00	0.98	108
accuracy			0.97	171
macro avg	0.98	0.96	0.97	171
weighted avg	0.97	0.97	0.97	171

Figure 4: Classification Report (KNN)

c. Kernel support vector Machines: Kernel SVM are used for classification of nonlinear data. In the chart, nonlinear is projected to higher dimensional space through a mapping function which makes the function linearly separable using s hyperplane.

d. Decision tree classification: The advantage of decision trees is that they don't require any feature scaling. Start at the tree root and split the data on the feature using the decision algorithm, resulting in the largest information gain or entropy or Gini impurity criterion. It is a also a fundamental component of random forest.

e. Random Forest Classification: The random forest ensemble corrects the habit of decision trees' overfitting to the training set by generating

most efficient decision trees based on highest information gain or all possible decision trees and it is calculated upon their votes what is the most optimal output.

3. Deep Learning

It is a specialized form of machine learning that uses supervised, unsupervised or semi-supervised learning to learn data representations. It resembles to the structure and function of the human nervous system. The vast availability of big data enables machines to be trained. Multi layered learning networks can be trained.

Chapter 5

Conclusions and Future Work

1. Conclusions

a. Using these machine learning techniques I was able to deduce significant inferences of how machine learning predicts and improves accuracy of the model.

b. Different cases for applying various machine learning algorithms for prediction.

2. Future Work

a. To increase the performance, I can implement CNN and RNN the concepts of deep learning.

b. The model deployment can be better if the user interface is faster and more interactive and search engine optimised if hosted on web.

References

[1] Josh Stormer, Statistics – Statquest.

[2] Adichie, J. N. [1967], "Estimates of regression parameters based on rank tests," Ann. Math. Stat., 38, 894–904

[3] Fundamentals Concepts of Neural Networks & Deep Learning 'https://www.analytixlabs.co.in/blog/fundamentals-of-neural-networks/'.

[4] https://www.healthline.com/health/cancer

[5] https://www.g2.com/articles/data-analysis-process

Appendices

Appendix A

Code Attachments

The following is the subset of the code. Code of some module(s) have been wilfully supressed.

A.1 ANN Implementation

```
# Importing modules and functions from libraries
import tensorflow as tf
tf.random.set_seed(1)
from tensorflow import keras
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import Dense,LeakyReLU,ELU,PReLU
from tensorflow.keras.layers import Dropout
import tensorflow
# model building
model=keras.Sequential([
    keras.layers.Flatten(input_shape=(30,)),
    keras.layers.Dense(40,activation='relu'),
    keras.layers.Dense(2,activation='sigmoid'),
])
#Model Compilation
model.compile(optimizer='adam',loss='sparse_categorical_cross
entropy',metrics=['accuracy'])
# Training model
history=model.fit(xtr,ytr,validation_split=0.15,epochs=50)
loss,acc=model.evaluate(xte,yte)
print(acc)
```

Output

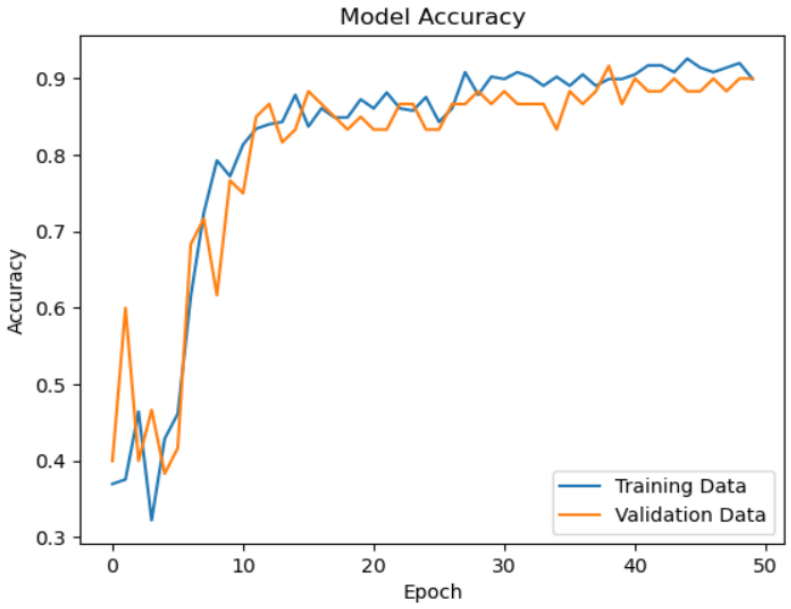


Figure 3: Learning rate

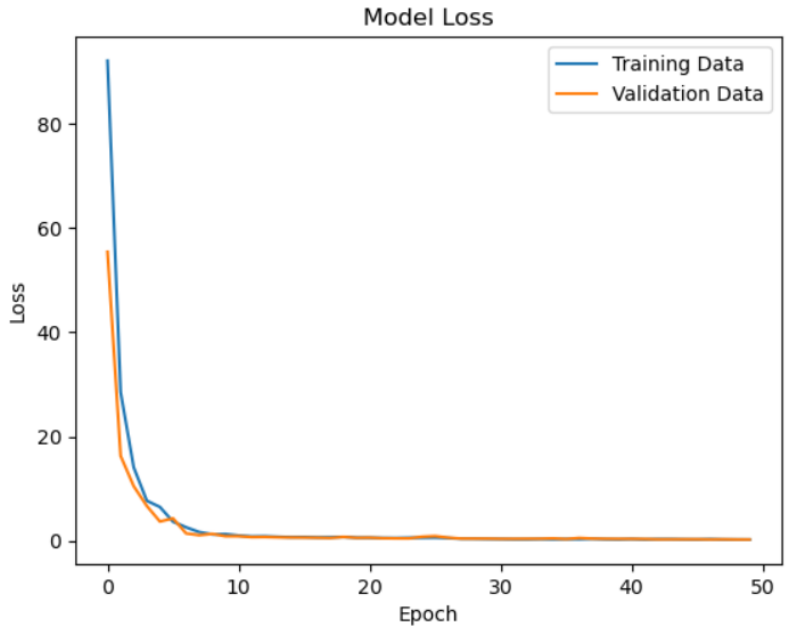


Figure 4: Loss rate

A.2 Logistic Regression

```
# sigmoid Function Declaration
def sigmoid(x):
    return 1/(1+np.exp(-x))
# Logistic class creating for the implementation of logistic
# methods
class LogisticRegression():

    def __init__(self, lr=0.001, itr=10000):
        self.lr = lr
        self.itr = itr
        self.weights = None
        self.bias = None
#         print("Learning Rate :",self.lr,"Number of
Iterations",self.itr,"Weights :",self.weights)

    def fit(self, X, y):
        n, feat = X.shape
        self.weights = np.zeros(feat)
        self.bias = 0
#         print("Weight Updates :",self.weights)

        for _ in range(self.itr):
            linear_pred = np.dot(X, self.weights) + self.bias
            predictions = sigmoid(linear_pred)

            dw = (1/n) * np.dot(X.T, (predictions - y))
            db = (1/n) * np.sum(predictions-y)

            self.weights = self.weights - self.lr*dw
            self.bias = self.bias - self.lr*db
#             print("Bias update :",self.bias)
```

```
#             print("Weight Updates :",self.weights,"Bias
Update :",self.bias)
        return self.weights,self.bias


    def predict(self, X):
        linear_pred = np.dot(X, self.weights) + self.bias
        y_pred = sigmoid(linear_pred)
        class_pred = [0 if y<=0.5 else 1 for y in y_pred]
        return class_pred
# model training
clflogr=LogisticRegression(lr=0.005)
k,m=clflogr.fit(xtr,ytr)
y_pred=clflogr.predict(xte)
# model Accuracy
```

```
def accuracy(y_pred, yte):
    return np.sum(y_pred==yte)/len(yte)
acc = accuracy(y_pred, yte)
print('Accuracy : ',acc*100)
from  sklearn.metrics import classification_report
print(classification_report(yte,y_pred))
```

```
Accuracy :  87.71929824561403
              precision    recall  f1-score   support

           0       0.78      0.94      0.85        63
           1       0.96      0.84      0.90       108

    accuracy                           0.88       171
   macro avg       0.87      0.89      0.87       171
weighted avg       0.89      0.88      0.88       171
```

Figure 7: Accuracy report (ANN)