A Study on Different Machine Learning Models To Predict Breast Cancer

Shashank Pathak(15BCE1287)*,
*School of Computing Sciences and Engineering, VIT Chennai, Tamilnadu, India 600127
Email: shashank.pathak2015@vit.ac.in*

***Abstract*—Breast Cancer is one of the most widespread disease among woman worldwide. Early and Accurate diagnosis of breast cancer is an extremely important step in rehabilitation and treatment. However, it is not an easy task due to several factors that play a big role in detection of a cancerous tumour. Machine learning techniques are prominently used in disease detection phases as it gives quite accurate results. Various techniques may provide different desired accuracies and it is therefore imperative to use the most suitable method which provides the best desired results. This research paper seeks to provide comparative analysis of Artificial Neural Network,Support Vector Machine,Decision Tree(ID3),K Means,K nearest Neigbours,Logistic Regression and Perceptron on the Wisconsin breast cancer dataset**

1. Introduction

According to different survey and WHO 25% of the females in the US are diagnosed with breast cancer at some stage in their life. Predicting a cancerous tumor remains a challenging task for many doctors.Thus the early detection of cancerous tumour is very important as it eases the removal of tumour and chances of survival increases.There are many factors which play an important role in the detection of cancer but the accurate detection is not dependent on any one factor or there is no matematical formula or ratio that predicts the Breast cancer.Therefore a learning approach is required to tackle this problem. Machine Learning (ML), is a subfield of Artificial Intelligence (AI) that allows machines to learn without explicit programming by exposing them to sets of data allowing them to learn a specific task through experience.Nowadays Machine Learning is used every where from social networks to self driving car.So we will use different machine learning models like Artificial Neural Network,Support Vector Machine,Decision Tree(ID3),K Means,K nearest Neigbours,Logistic Regression and Perceptron inorder to predict that a tumour is cancerous or non-cancerous that corresponds to Malignant or Benign tumour respectively.

2. Methodology

We took the Breast cancer Wisconsin dataset and realized that this dataset needed cleaning.So we cleaned the dataset by removing NaN and '?' enteries .We then analysed the dataset and calculated various stastical measures for analysing the data distribution like mean,mode,median,total valid instances etc.After this we played with the data to get some insight on data and any kind relation between class and attributes .After this we applied different Machine Learning Models like Artificial Neural Network,Support Vector Machine,Decision Tree(ID3),K Means,K nearest Neigbours,Logistic Regression and Perceptron on the dataset.We got the insample and test sample accuracies for all the models which helped us to comparatively analyze the models for the Breast Cancer predication problem.

3. Database - Breast Cancer Wisconsin Dataset(Original)

The Wisconsin Breast Cancer datasets from the UCI Machine Learning Repository is used,to distinguish malignant (cancerous) from benign (non-cancerous) samples.This dataset consist of 699 instances and 11 attributes that would help the classiy the data into the two classes.458 of the cases are benign and 241 are malignant.The attributes are described in the fig.1

#	Attribute	Domain
1.	Sample code number	id number
2.	Clump Thickness	1 - 10
3.	Uniformity of Cell Size	1 - 10
4.	Uniformity of Cell Shape	1 - 10
5.	Marginal Adhesion	1 - 10
6.	Single Epithelial Cell Size	1 - 10
7.	Bare Nuclei	1 - 10
8.	Bland Chromatin	1 - 10
9.	Normal Nucleoli	1 - 10
10.	Mitoses	1 - 10
11.	Class:	(2 for benign, 4 for malignant)

Figure 1. Breast Cancer Wisconsin Dataset Attributes

The above attributes except for the class are of integer type, and their values range between 1 and 10 both inclusive. The class attribute has a value of 2 for benign and 4 for malignant.

4. Classification Algorithms

The various algorithms used in this study are explained briefly in this section of the document

i)Decision Tree

Decision tree is used in supervised learning in machine learning for classification.Decision tree is like a normal tree that is used in data structures whose nodes represent as features or attributes,branches are the various conditions on which classification is done,leaves represent the classes of the dataset.So decision tree starts from start node which are non terminal goes through the branches which are the various classification conditions and reaches the leaves which are the classes of the data.Each node contains a condition which decide which branch to follow next.

The nodes are created using the entropy or information gain values.The condition on the nodes decide the path if the condition is satisfied then that path is followed otherwise different path is taken by the algorithm. Algorithm starts from the root node and follows a top down approach to traverse through the tree to reach the nodes.

We have used ID3 Algorithm to create the decision tree for classification.In ID3 algorithm it searches for that attribute which classify the most number of data instances that is it chooses that attribute which is able to classify most efficiently and then it makes tree subsequently in order in which these attributes classify the data using entropy and information gain of each attribute.The algorithm continues the process until finding subset instances belonging to the same class, and so the leaf node is created, and the algorithm stops when it has handled all the attributes.

ii)K Nearest Neighbor

$$d(x,x') = \sqrt{(x_1 - x_1')^2 + (x_2 - x_2')^2 + \ldots + (x_n - x_n')^2}$$

Figure 2. Euclidean Distance used in KNN

(KNN) K Nearest Neighbor classifier is one of the basic classifiers in machine learning algorithms.It classifies the data based on the majority vote of the K nearest points of a data point.Basically what it does is that it maps all the data instances in the n dimension space and calculates the distance between a given point and other points k and find out the K nearest neighbours and then it takes votes of the classes of these neighbours and if there are more number of points which are belonging to a particular class t then the class of the point is t. It is a supervised learning mechanism so we need to decide the value of K which is used by the algorithm for classification.

iii)Support Vector Machines

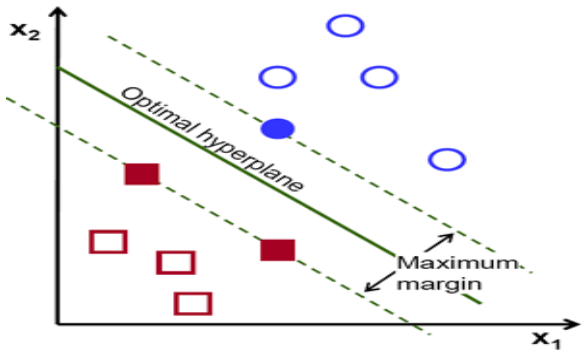

Figure 3. Support Vector Machine

SVM(Support Vector Machines) is a very important and efficient suprevised learning algorithm in machine learning.Support Vectors are the critical sample points which are used to generate a linear function which divides the points as broadly as possible.These support vectors are used for deciding the width of the boundary of the linear function.The more will be the boundary of this linear function the more better will be the classification.So we aim for maximizing the boundary and this is also known as maximum margin classifier.SVM aims to find the most suitable hyperplane that divided the dataset into different classes.

$$\min P(\boldsymbol{w},\, b) = \underbrace{\frac{1}{2}\,\|\boldsymbol{w}\|^2}_{\text{maximize margin}} + \underbrace{C\sum_i H_1[\,y_i\, f(\boldsymbol{x}_i)\,]}_{\text{minimize training error}}$$

Figure 4. SVM Maximize the margin

iv)Multi Layer Perceptron

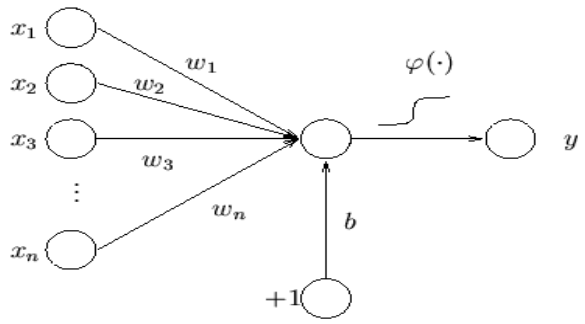

Figure 5. Multi Layered Perceptron

MLP(Multi Layer Perceptron) is a supervised classification algorithm used in machine learning.Multi Layer perceptron is a feedforward neural network with one or more layers between input and output layer.The algorithm uses a feedforward approach in which data starts from the input layer and flows till it reaches the output layer. This type of network is trained with the backpropagation

learning algorithm.Once a feedforward cycle is complete we calculate the error between the predicted value and the value of the class at the output layer and then this error is propagated backward till it reaches the last layer and the weights of the neural nets are modified according to the error.

$$y = \varphi(\sum_{i=1}^{n} w_i x_i + b) = \varphi(\mathbf{w}^T \mathbf{x} + b)$$

Figure 6. Forward propogation in MLP

So the algorithm perform in cycles where one complete cycle consist of feedforward the network calculating the error and backpropagation of error and the updation of weights of neurons.These are the basic three steps which complete a given cycle.Linearly inseperable problems can be solved using Multi Layer Perceptron.MLP uses stochastic gradient descent for classification.

v)K Means

K means algorithm is distance based classification.There are two main step in K means classification first step is cluster assignment step and the second step is updation of centroid.We need to define K for the algorithm.In K means algorithm distance of the given data point is taken with all the other data points and the centroid of the group is decided using the data points that belong to the same class and then we update the centroid with each step.K means is an iterative algorithm.

vi)Logistic Regression

Logistic Regression is a supervised machine learning algorithm which gives a probabilistic approach to machine learning.Logistic regression values range between zero and one.Logistic Regression gives the probability of a point belonging to a particular class.We estimate teh value of the coefficients using the stochastic gradient descent. We can estimate the values of the coefcients using stochastic gradient descent. This is a simple procedure that can be used by many algorithms in machine learning. It works by using the model to calculate a prediction for each instance in the training set and calculating the error for each prediction. We can apply stochastic gradient descent to the problem of finding the coefcients for the logistic regression model as follows by calculating for each instance a prediction using the current value of coefficents and calculating new coefficeints based on the error in the prediction.The process is repeated until the model is accurate enough or for a fixed number of iterations.

$$E_{\text{in}}(\mathbf{w}) = \frac{1}{N} \sum_{n=1}^{N} \underbrace{\ln\left(1 + e^{-y_n \mathbf{w}^{\intercal} \mathbf{x}_n}\right)}_{\mathrm{e}\left(h(\mathbf{x}_n), y_n\right)}$$

Figure 7. Logistic Regression

vii)Perceptron

Perceptron is the most basic algorithm under supervised classification. It involves the mapping of a function on its input x (a real- valued vector) to an output f(x) (a single binary value): where w is the vector of a real-valued weights and b is the bias. The bias shifts the decision boundary away from the origin and does not depend on any input value.

This linear function is used to classify the in-sample as well as out-sample data and the score of this algorithm can be calculated using the number of misclassification done by the best hypothesis.

5. Experiments

We took the Breast cancer Wisconsin dataset and realized that this dataset needed cleaning.So we cleaned the dataset by removing NaN and '?' enteries.After this we used standard scaler to scale the values although this step is not necessary as the range of attributes are already between 1 to 10. After this we changed the label value of benign and Malignant from 2 and 4 to -1 and 1 repectively as this is a case of binary classification and it would be better to have a positive and a negative case.Then we used train_test_split from cross_validation of scikit learn to split the dataset into X_train,y_train X_test and y_test. From plotting and infographics we used seaborn and matplotlib for all the models.

After this we made a perceptron from scratch for our dataset and trained it on the X_train and y_train.We got a seprating hyperplane as in fig. .After this we analysed the performance of perceptron for differnt value of epochs.

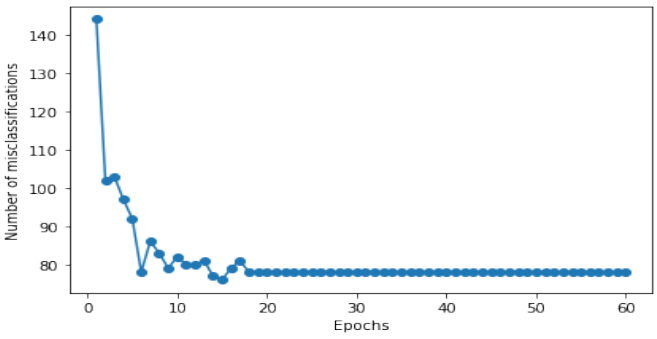

Figure 8. Misclassification vs errors

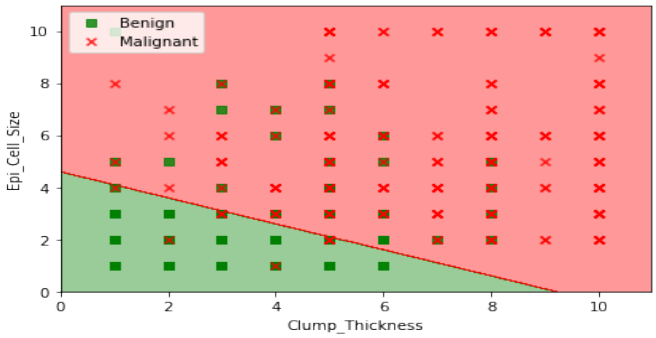

Figure 9. Classification by Perceptron

After this we trained Decision tree on our training dataset for all attributes.After doing some hyperparameter tuning we concluded the optimum value of max_depth equal to 4.After this we used the testing dataset to test the ID3 based Decision tree and got a accuracy of 95.9%.

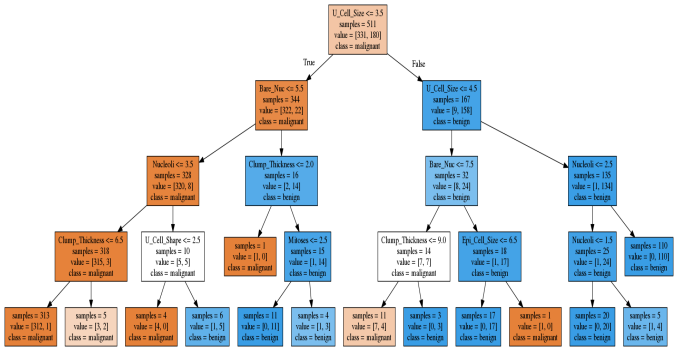

Figure 10. Classification by Decision Tree

Then we applied Logistic Regression model from sklearn models and trained it on X_train and y_train. After this we used the score method to test the trained Logistic Regressor and got around 32% accuracy which is the lowest of all the model we used.

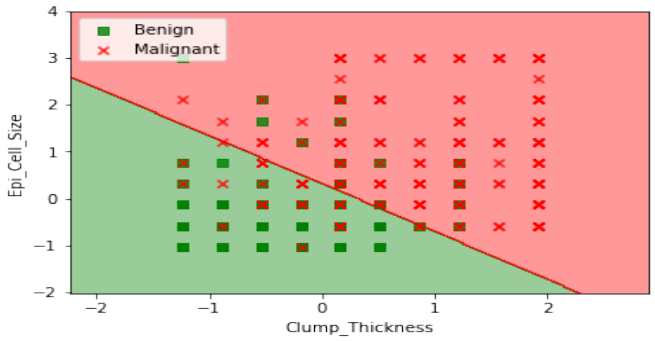

Figure 11. Classification by Logistic Regression

After this we tried an unsupervised learning model i.e. K Means to find the clusters in our dataset and got 2 clusters and a distotion of aroun 347.49.The hyper parameter values are mentioned in the table.

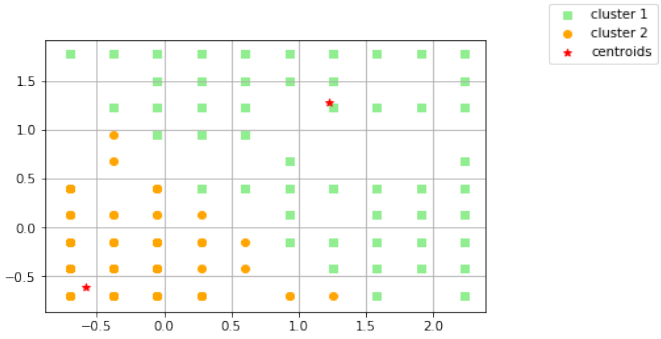

Figure 12. Clustering by K Means

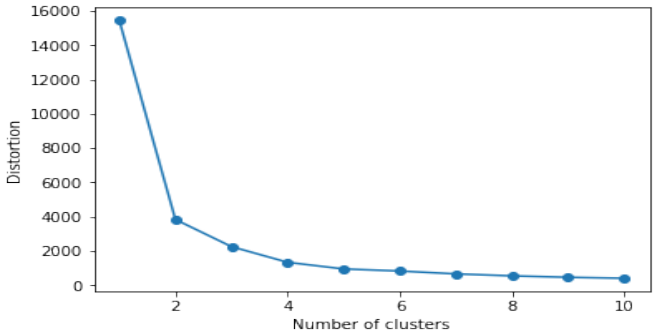

Figure 13. no.of clusters vs Distortion

After we used multi layered perceptron or more commonly known as Artificial neural network.We trained our model for 1000 epochs and 25 hidden layers to get a 2 neuron output layer.Then we tested the model on X_test and y_test and got 97.56% accuracy.

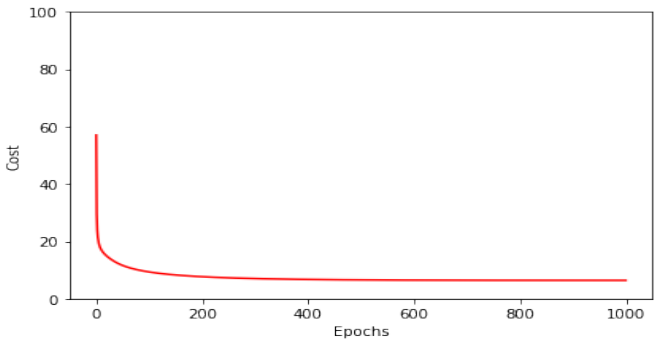

Figure 14. Cost vs Epochs in MLP

Then we tried support vector machine for the Breast cancer prediction problem. This is the most widely used model for this problem and we used the SVM from sklearn model library.We tried different kernels like radial bias,polynomial and linear.We used Gridsearch and pipeline to do hyper

parameter tuning and check the best kernel.We got best results for rbf kernel and for C i.e. the cost of error value equal to 10.We got a accuracy of 94.63%

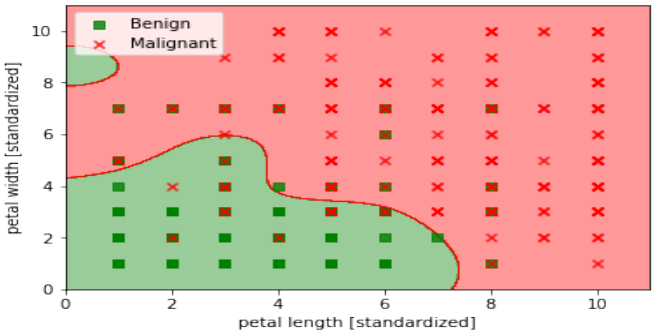

Figure 15. Classification by SVM

The last model that we tried was K Nearest Neighbour.After training on our dataset we did the hyperparameter tuning using cross_val_score of scikitlearn to get the best parameters.We got no. of neighbours equal to 2.We used Minkowski distance as the distance parameter .After testing the model on X_test and y_test we got a accuracy of 96.09%

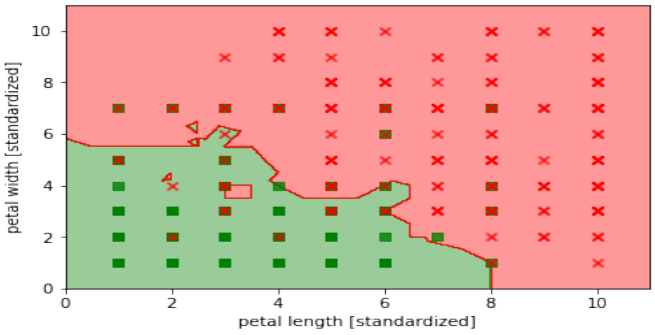

Figure 16. Classification by KNN

6. Conclusion

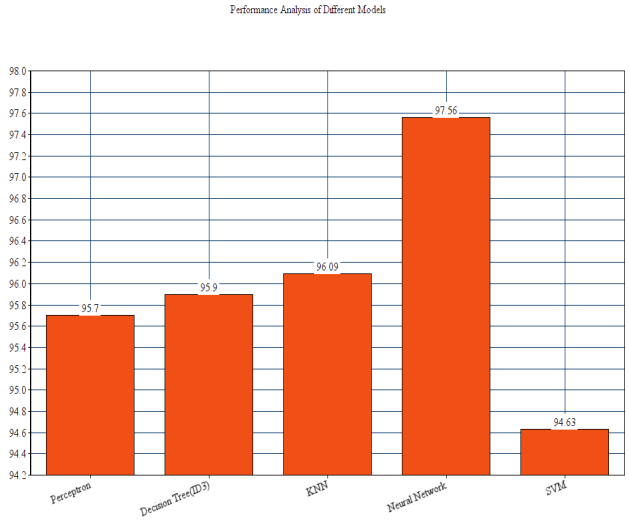

Figure 17. Accuracies for different models

We trained some of the Machine Learning Models like Artificial Neural Network,Support Vector Machine,Decision Tree(ID3),K Means,K nearest Neigbours,Logistic Regression and Perceptron and got differnt accuracies.We got accuracies that are mentioned in the fig17.We can clearly see that Neural Network performed the best with accuracy 97.63%.Most of the algorithm performed very well in the learning process but Logistic regression performed below par with wort accuracy of 32.1%.So in this we have seen that the breast cancer dataset can be learned by most of these models very effectively.Hence the Breast Cancer Dataset is easily classifiable using htese models.

7. References

[1] WHO Breast Cancer: Prevention and Control (2015) Retrieved 20 Jan 2015, from WHO World Health Organization. http://www.who.int/cancer/detection/breastcancer/en/index1.html

[2] Y. Elobaid, T.-C. Aw, J. N. W. Lim, S. Hamid, and M. Grivna, Breast cancer presentation delays among Arab and national women in the UAE, a qualitative study, SSM - Popul. Heal., Mar. 2016.

[3] E. D. Michie, D. J. Spiegelhalter, and C. C. Taylor, Machine Learning , Neural and Statistical Classification, Proceeding, 1994.

[4] I. Kononenko, Machine learning for medical diagnosis: history , state of the art and perspective, vol. 23, 2001.

[5] G. Williams, Descriptive and Predictive Analytics, Data Min. with Ratt. R Art Excav. Data Knowl. Discov. Use R, pp. 193-203, 2011.

[6] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, Machine learning

applications in cancer prognosis and prediction, Comput. Struct. Biotechnol. J., vol. 13, pp. 8-17, 2015.
[7] T. J. Cleophas and A. H. Zwinderman, Machine Learning in Medicine, pp. 1-271, 2013.
[8] Y. Yasui and X. Wang, Statistical Learning from a Regression Perspec- tive by BERK, R. A., vol. 65, no. 4. 2009.
[9]M. Lichman, UCI Machine Learning Repositry, 2013. [Online]. Available: https://archive.ics.uci.edu/.
[10] T. Fushiki, Estimation of prediction error by using K-fold cross- validation, Stat. Comput., vol. 21, no. 2, pp. 137-146, 2011.