Breast Cancer Prediction System using Classification Algorithms

Sarthak Vinayaka
Department of Computer Science and Engineering
Jaypee University of Information Technology Waknaghat, India

Abstract— Breast cancer is the most common cancer in women worldwide, with nearly 1.7 million new cases diagnosed in 2012.It can be classified as Benign or Malignant. Classification and data mining methods are an effective way to classify data. Especially in medical field, where those methods are widely used in diagnosis and analysis to make decisions. In this paper we predict breast cancer with high accuracy by applying machine learning algorithm like K- Nearest neighbors on the Wisconsin Breast Cancer Database .In other word, we can anticipate the future for women diseases.

Keywords: K-Nearest Neighbors, Machine Learning, Accuracy

I. INTRODUCTION

Nowadays Brest cancer becomes very major disease in many women not only in India but also in other country. In 2017, around 252710 new diagnoses of breast cancer were expected in women, and around 40610 women almost died from the disease. Breast cancer can be divided into benign and malignant. Therefore, the selection of model for predicting the nature of breast tumor is significantly important. To predict breast cancer we will apply various machine learning classification algorithm on Wisconsin Breast Cancer Database and calculate the accuracy and in our proposed method we will modify K Nearest neighbor and the accuracy achieved is 98.28% and will also calculate Precision, Recall and F1 score.

II. DATA SOURCE

In this paper, we use the Breast cancer data obtained from the University of Wisconsin Hospitals [1], Madison from Dr. William H. Wolberg. There are total 699 instances with 11 attribute and class attribute has two values 2 and 4 which is 2 for benign, 4 for malignant.

A. Feature Description

Table 1 shows the 11 attributes and their description.

Attribute	Domain
Sample code number	-
Clump Thickness	1-10
Uniformity of Cell Size	1-10
Uniformity of Cell Shape	1-10
Marginal Adhesion	1-10
Single Epithelial Cell Size	1-10
Bare Nuclei	1-10
Bland Chromatin	1-10
Normal Nucleoli	1-10
Mitoses	1-10
Class	2 for benign, 4 for malignant

Table 1:

III. RELATED WORK

Gayathri Devi.S[2] have done research on the Breast cancer dataset and concluded that after reducing the attributes to 6, 5 and 4 using rank search, genetic search and greedy step wise search the performance of the classification algorithms are improved. It is noted that Logistic Classifiers performance is increased after Feature reduction.

A.Kathija,S. Shajun Nisha and Dr .M .Mohamed Sathik[3] have used the Neural Network Approach of MLP Algorithm to predict the breast cancer and concluded that the performance of MLP shows the high level accuracy with other classifiers. Therefore MLP is suggested for predict survivability of Breast Cancer disease based classification to get better results with accuracy and performance.

D.Lavanya and Dr.K.Usha Rani[4] have used Ensemble Decision tree classifier dataset and conclude that the experimental results of a hybrid approach with the combination of preprocessing, bagging with cart demonstrated the enhanced classification accuracy of the selected data sets.

Hiba Asria, Hajar Mousannifb, Hassan Al Moatassimec, Thomas Noeld[5] have implemented various machine learning algorithm and compare them and conclude that the SVM has proven its efficiency in Breast Cancer prediction and diagnosis and achieves the best performance in terms of precision and low error rate.

Vivek Kumar , Brojo Kishore Mishra , Manuel Mazzara, Dang N. H. Thanh , Abhishek Verma[6] have used the data mining approach in breast cancer data set and concluded that Only Naïve Bayes has underperformed compared to other with accuracy of 73.21%. Tree and Lazy classifier algorithms have performed exceptionally well; accuracy being close to 99%.

G. Ravi Kumar,Dr. G. A. Ramachandra[7] done research and apply various algorithm on dataset and concluded that SVM classifier is suggested for diagnosis of Breast Cancer disease based classification to get better results with accuracy, low error rate and performance. Yixuan Li, Zixuan Chen[8] done performance analysis on breast cancer dataset.

IV. OUR PROPOSED METHODOLOGY

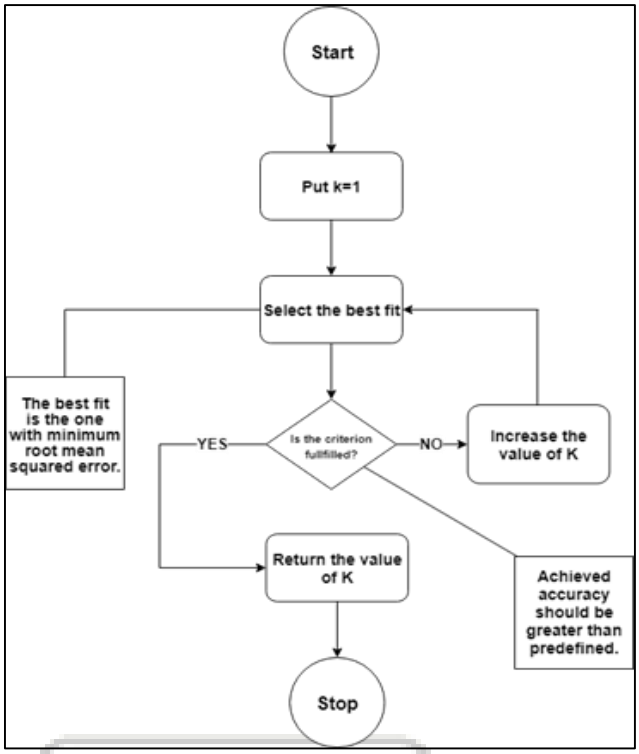

Fig. 1: Our proposed algorithm flow chart.

KNN (K — Nearest Neighbors) is one of many (supervised learning) algorithms used in data mining and machine learning, it's a classifier algorithm where the learning is based "how similar" is a data (a vector) from other. We have changed the value of K and start with K=1 and then calculate the accuracy then increase the value of K and then calculate the accuracy we have found that when K=3 the achieved accuracy is highest which is equal to 98.28%.

Value of K	Accuracy	(%) Time (Seconds)
1	97.71	0.022
2	97.14	0.024
3	98.28	0.024
4	97.71	0.25
5	97.71	0.23
6	97.71	0.25
7	97.71	0.26

Since by seeing the above table the accuracy is highest when k=3 and then it remains constant and in k nearest neighbor algorithm training time is approximately same or we can say it is independent from the value of k.

V. EXPERIMENTS AND RESULTS

We have applied our proposed method in our dataset and accuracy achieved is 98.28%.

F1 score = 0.98, Recall score= 0.98, Precision score = 0.99

We have used Python, jupyter notebook, pandas, Scikit-learn, matplotlib for applying the various algorithm in our dataset and plotting the result through graph.

VI. CONCLUSION AND FUTURE WORK

We have applied various algorithm in our dataset and measured the accuracy and highest accuracy achieved is 98.28% by keeping the value of k=3. The proposed algorithm can work very well on real world although the accuracy has been increased and can be increased more by collecting more data and implementing other algorithm like Neural network and other classification algorithm and collecting the value of incomplete dataset. This prediction through machine will help to reduce the errors made by doctors. However there will be still need of doctors.

REFERENCES

[1] Breast Cancer Wisconsin (Original) Data Set, https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(original)

[2] Gayathri Devi.S Breast Cancer Prediction System using Feature Se- lection and Data Mining Methods, International Journal of advanced research in computer science, ISSN No. 0976-5697

[3] A.Kathija, S. Shajun Nisha and Dr .M .Mohamed Sathik, Breast Cancer Data Classification Using Neural Network Approach of MLP Algorithm,International Journal of Trend in Research and Development, Volume 4(3), ISSN: 2394-9333

[4] D.Lavanya and Dr.K.Usha Rani ENSEMBLE DECISION TREE CLAS- SIFIER FOR BREAST CANCER DATA,International Journal of In- formation Technology Convergence and Services (IJITCS) Vol.2, No.1, February 2012

[5] Hiba Asria,Hajar Mousannifb, Hassan Al Moatassimec, Thomas Noeld Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis, The 6th International Symposium on Frontiers in Ambient and Mobile Systems (FAMS 2016)

[6] Vivek Kumar [0000-0003-3958-4704], Brojo Kishore Mishra [0000- 0002-7836-052X], Manuel Mazzara [0000-0002-3860-4948], Dang N. H. Thanh [0000-0003-2025-8319], Abhishek Verma Prediction of Ma- lignant Benign Breast Cancer: A Data Mining Approach in Healthcare Applications,

[7] G. Ravi Kumar Research Scholar,Dr. G. A. Ramachandra Associate Professor An Efficient Prediction of Breast Cancer Data using Data Mining Techniques, International Journal of Innovations in Engineering and Technology (IJIET) Vol. 2

[8] Yixuan Li, Zixuan Chen ,Performance Evaluation of Machine Learning Methods for Breast Cancer Prediction, Applied and Computational Math- ematics 2018; 7(4): 212-216,ISSN: 2328-5605 (Print); ISSN: 2328-5613 (Online)

