**Performance Analysis of Machine Learning Models for Breast Cancer**

Rishabh Giri

Department of Information Technology

ABES Engineering College, Ghaziabad, UP, India

2000rishabhgiri@gmail.com

**Abstract:**

Breast cancer is an all too common disease in women, making it the major source of mortality between women. The intention of this study is to design a prediction system that can predict the incidence of the breast cancer at early stage. Wisconsin breast cancer dataset (WBCD) have been used to conduct the proposed experiment. It is based on classification problem where we need to classify the disease as malignant (harmful for woman) or benign (not harmful) In this work we use six classification providing best accuracy for classification problem for example SVM have accuracy of 87.7%, KNN have accuracy of 85% etc. Ada Boost classifier provides us the best result of 94% accuracy of the whole sample. These findings will help in early stage breast cancer prediction.

**Keywords**- machine learning, Breast Cancer, performance analysis, Prediction, Classification.

1. **Introduction**

Cancer begins in the cells which are the basic building blocks that make up tissue. Tissue is found in the breast and other parts of the body. Sometimes, the process of cell growth goes wrong and new cells form when the body doesn’t need them and old or damaged cells do not die as they should. When this occurs, a build up of cells often forms a mass of tissue called a lump, growth, or tumor.

Breast cancer occurs when malignant tumors develop in the breast. These cells can spread by breaking away from the original tumor and entering blood vessels or lymph vessels, which branch into tissues throughout the body. When cancer cells travel to other parts of the body and begin damaging other tissues and organs, the process is called metastasis. When this Cancer develops in the breast tissue it is termed as Breast Cancer, there are several causes that include drinking alcohol, family history of breast cancer, Menstrual and Reproductive History, Certain Genome Changes, Having dense breast tissue can increase your risk for breast cancer and make lumps harder to detect, Lack of Physical Activity, Poor Diet, Being Overweight.

According to the American Cancer Society, when breast cancer is detected early, and is in the localized stage, the 5-year relative survival rate is 99%. Early detection includes doing monthly breast self-exams, and scheduling regular clinical breast exams and mammograms. With the help of latest and advanced screening methods majority of such cancers are diagnosed when the disease is still at a localized stage. As Machine Learning Techniques are widely used in healthcare sector. Patient’s data and physician’s prescription are the most considerable features in diagnosis.

1. **Literature Review**

* Madhu Kumari, Vijendra Singh has implemented Decision tree for breast cancer and got very good performance.
* Wei-chao Lin has implemented SVM and SVM Ensembles in breast cancer.

1. **Proposed methodology**

Feature Extraction

Feature selection

Data Preprocessing

Data Collection

* + **Data Collection**

For this project the data has been downloaded from [Kaggle](https://www.kaggle.com/code/buddhiniw/breast-cancer-prediction/notebook) . It consist of 29 classes namely radius, texture, perimeter, area, smoothness, compactness, concativity etc .There are total of 569 number of samples and divided in 2 classes as shown in figure 1.

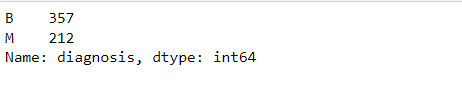


Figure 1

* + **Data Preprocessing**

The data is preprocessed it does not require to make any change in dataset. Id is not considered as any valid parameter and diagnosis is our final output hence we are not considering both these as input parameter.

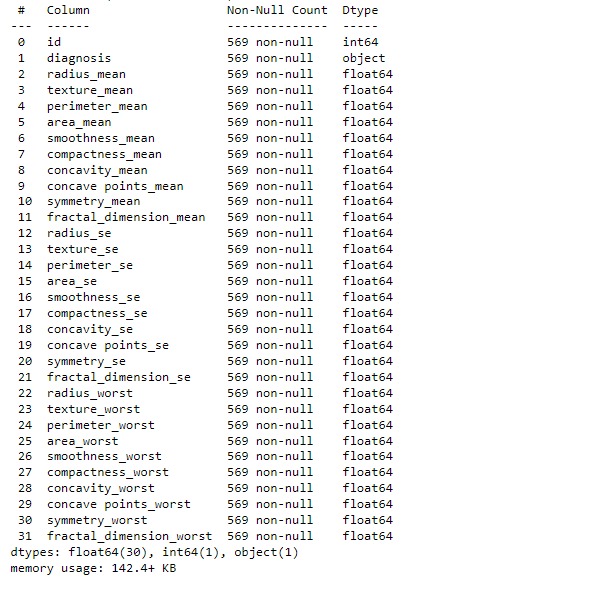


Figure 2

* + **Feature selection**

For feature selection we take parameters with mean values and draw a heat-map between these parameters of the dataset.

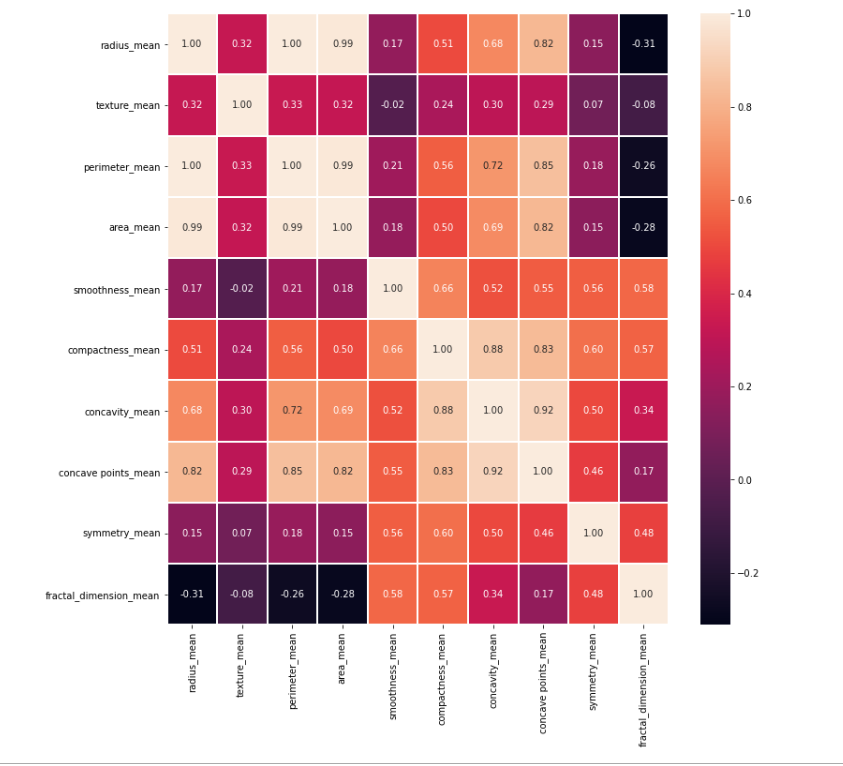


Figure 3

* + **Feature Extraction**

PCA (Principle Component Analysis) is used to extract features by reducing the numbers of input variables, we use Random forest with PCA and got 91.8% accuracy score, Feature importance is also used to find importance of every parameter as shown in figure 4.

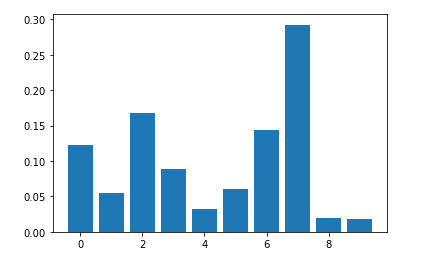


Figure 4

1. **Result & Discussion**

All experiment has been carried out in widows 10 with Intel(R) Core(TM) i5-8265U CPU @ 1.60GHz 1.80 GHz, RAM 8 GB. In this analysis we worked on six Machine Learning Models providing accuracy 87.71% by SVM, 90% by Naïve Bayes, 85.9% by KNN, 90% by Decision Tree, 92.3% by Random forest, 93.5% by Extra tree classifier and highest 94% by Ada Boost.

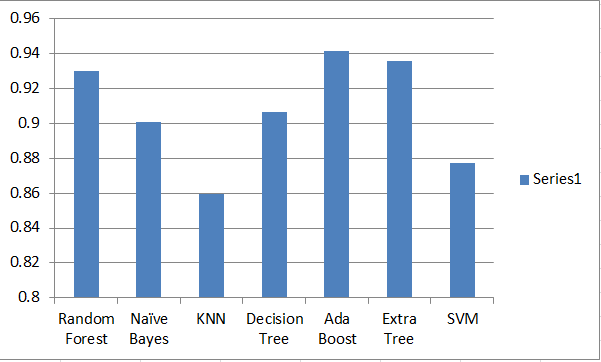


Figure 5

Table 1. Shows the result comparision of various ML techniques

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Accuracy | Recall | Precision |
| SVM | 0.8771 | 0.90 | 0.84 |
| Naïve Bayes | 0.9008 | 0.90 | 0.89 |
| KNN | 0.8596 | 0.85 | 0.84 |
| Decision Tree | 0.9064 | 0.90 | 0.89 |
| Random Forest | 0.9239 | 0.92 | 0.91 |
| Ada boost | 0.9415 | 0.94 | 0.93 |
| Extra tree | 0.9356 | 0.94 | 0.93 |

1. **Conclusion & Future work**

In this study we illustrate six machine learning models in which Ada boost classifier provides the best accuracy score of 94% followed by Extra tree classifier and Random forest classifier.

Future work or scope for this project is that it can further implemented in the form of smart device which can be deployed in hospitals or other health care centers for early stage breast cancer detection and these models need to be more trained by large data sets to depict the problem more precisely.

**References:**

[1]. "WHO | Breast Cancer: Preventation and contro", www.

who.in/cancer/detection/breastcancer/en/index3.html

[Accessed: 01-Nov-2019].

[2]. M. R. Al-Hadidi, A. Alarabeyyat, and M. Alhanahnah,

“Breast Cancer Detection Using K-Nearest Neighbor

Machine Learning Algorithm,” in 2016 9th International

Conference on Development.

[3]. L. Abdel-Ilah and H. Šahinbegović, “Using machine

learning tool in classification of breast cancer,” Springer,

Singapore, 2017, pp. 3–8.

[4]. M. Kumari, “Breast Cancer Prediction system,” Procedia

Comput. Sci., vol. 132, pp. 371–376, Jan. 2018.