International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 03 | Mar 2021 www.irjet.net p-ISSN: 2395-0072

Analysis of Machine Learning Algorithms for Breast Cancer Prediction

Dev Prakash[1], **Sindhu K**[2]

[1] *Student, Dept. of Information Science, BMS College of Engineering, Bangalore, Karnataka, India*
[2] *Assistant Professor, Dept. of Information Science, BMS College of Engineering, Bangalore, Karnataka, India*

Abstract - *Breast cancer represents one of the diseases that makes many deaths every year. Breast cancer accounts for the second major cause of death in women. Several machine learning algorithms have been used to develop a prediction model. Among them Logistic Regression, Decision Tree, KNN, SVM are the most used techniques. However, there have been very few studies about the performance of SVM based on kernel functions used in the breast cancer prediction. Apart from these, classifier ensemble which is a powerful technique can also be employed in this scenario. Therefore, the aim of this paper is to fully assess the prediction performance of these algorithms in breast cancer prediction considering accuracy score, precision score, recall score and f1 score in the evaluation metrics. The experimental results show an ensemble of Logistic regression + Decision tree + KNN outperforms all the other classifiers. Individual models of RBF kernel and logistic regression also perform much better than other models.*

Key Words: **Breast Cancer Prediction, SVM, Cross Validation, Feature Selection, Classifier Ensemble.**

1.INTRODUCTION

Breast Cancer is a quite common disease in women all over the world. It is one of the main causes of women's death all over the world. The cancer is developed in the breast tissue. Several risk factors for breast cancer are lack of physical exercise, obesity, hormone replacement therapy during menopause, early age at first menstruation, ionizing radiation, having children late or not at all [1]. We can measure the seriousness of this disease by this report given by Siegal et al. [2] which states that Breast cancer contributes around 12% of the cancer cases and 25% of all cancers in women and that's why breast cancer prediction becomes an important research problem both in the medical as well as in healthcare communities.

Many machine learning algorithms and statistical techniques have been employed to develop a large variety of breast cancer prediction models. Some of the most used techniques are Logistic Regression, Decision Trees, K nearest Neighbors KNN), Support vector machine (SVM), etc. In the proposed work the performance of these algorithms has been analyzed, compared and the best suited algorithm for breast cancer prediction is identified. The focus of this study is the SVM algorithm since accuracy of an SVM model depends largely on kernel functions used. These Kernel functions include Linear, RBF (Radial Basis Function), Poly and Sigmoid functions.

Section 2 of the paper discusses the algorithms used in the proposed work. Previous research work is discussed in Section 3. The experimental methodologies are presented in Sections 4. The results of the work are highlighted in Section 5 while Section 6 gives the concluding remarks.

2. BACKGROUND

All the algorithms used in this study have been discussed below:

2.1 Logistic Regression

Logistic Regression is one of the most fundamental classification algorithms used in ML(Machine Learning). Logistic Regression can be used for both binary classification as well as multi-class classification. In fact, logistic regression predicts the probability of different samples and then these samples are mapped to a discrete class based on that probability. It uses Logistic/Sigmoid Function at its core, that's why it has been named Logistic regression.

2.2 Decision tree

In decision tree all possible decision paths are mapped out in the form of a tree. The training set is split into distinct nodes to build the decision tree. To classify data, it uses recursive partitioning.

2.3 K-Nearest Neighbors(KNN)

Is a classification algorithm that uses a bunch of labelled points to learn how to label other points. This algorithm classifies cases based on their similarity to other cases. Data points that are near each other are called neighbors. K in K nearest neighbors stand for the number of nearest neighbors to examine. K value plays a major role in determining how accurate our model is. So, what value of K to select becomes a big issue here. For extremely low values of K, it is considered that noise has been captured

in the data or one of the points that was an anomaly in the data has been chosen. On the other side of the spectrum, a higher value of K causes the model to become overly generalized.

2.4 Support Vector Machine (SVM)

SVM uses classification algorithms for two-group classification problems by finding a separator. Each data item is plotted as a point in n dimensional space with the value of each feature treated as the value of a particular coordinate. Then a hyperplane is found which differentiates the two classes. Anything that falls on one side of a hyperplane will be considered in one class while anything that falls on the other side will be considered in the other class.

There can be many possible hyperplanes. But the one whose distance to the nearest points of each tag is the largest are considered the best hyperplane. For finding the best hyperplane, support vectors are used. Most of the time finding the hyperplane becomes quite a tough job since not every time all data points can be separated by linearly separable lines. Therefore, the concept of kernel functions comes into the picture. Kernelling basically means mapping data into a higher-dimensional space and the mathematical functions that are used for this purpose are called kernel functions. There are many kernel functions out there. Linear, RBF (Radial basis functions), Poly and Sigmoid are some of the most used kernel functions [3].

2.5 Classifier ensemble

Classifier ensemble is an advanced technique used in machine learning. It is used mainly to solve complex problems. In the classifier ensemble many different and independent models (also called base models) are created and then the results of each model are combined to produce a better model. Usually, classifier ensembles have higher accuracy as compared to single model (base model). The three most popular methods for combining the model predictions are:

1. Bagging Method
2. Boosting Method
3. Voting and averaging Method

Voting and averaging are two of the easiest ensemble methods. Voting is mainly used for classification. Averaging is used mainly for regression. Predictions from multiple machine learning algorithms are combined in voting methods. A vote is considered as a prediction form each model. The final prediction is considered based on the predictions of most of the models [4].

3. RELATED WORK

Many machine learning models have been created using the above algorithms. There have been several research works done which deal with the performance of different algorithms being used in breast cancer prediction.

One such work has been carried out by Asri et. al. where they have compared the performance of four classifiers SVM, C4.5, Naive Bayes (NB) and KNN on the Wisconsin Breast cancer dataset. The algorithm is evaluated by considering sensitivity, specificity, precision and accuracy. SVM proved to be the most effective, outperforming the rest of them by reaching the highest accuracy of 97.13% while C4.5, Naive Bayes and KNN had accuracies that varied between 95.12% and 95.28%. In addition, SVM achieved the best performance both in terms of precision as well as in terms of low error rate [5].

Another work has been done by Vikas et al. where they have used Decision tree, RBF kernel and Logistic Regression algorithms on the dataset obtained from University Medical Centre, Institute of Oncology, Ljubljana. They conducted this experiment using libraries obtained from the Weka machine learning environment. They concluded that simple Logistic Regression performed better than all with an accuracy of 74.47% [6].

Ahmad et al. used three very popular machine learning techniques, Decision Tree (C4.5), SVM and Artificial Neural Networks (ANN). The performance of the algorithms was compared through sensitivity, specificity and accuracy on the dataset obtained from Iranian center for Breast cancer (ICBC). The predictions given by the SVM model were 95.7% accurate which is more as compared to the predictions obtained by other algorithms [7].

Delen et al. used ANN, Decision Trees along with Logistic regression to develop the model on a dataset that contained more than 200,000 instances. They made use of 10-fold cross validation in order to evaluate the unbiased estimate of three models obtained and carried out their performance comparison. Their results concluded Decision Tree to be the best predictor which gave an accuracy of 93.6% followed by ANN and Logistic classifier which predicted the cases with the accuracies 91.2% and 89.2% respectively [8].

Huang et al. studied the accuracies of SVM classifiers based on the different kernel functions used. They used 10-fold cross validation for better estimates of a model and feature selection as a pre-processing step. Linear kernel

based SVM ensembles based on the bagging method and RBF kernel based SVM ensembles with the boosting method were considered better for the small dataset. RBF kernel based SVM ensembles based on the boosting method performs better for larger dataset [9].

4. EXPERIMENT METHODOLOGY

In the proposed work, multiple experiments were conducted to evaluate the performance of the various classifiers and then examined their performance.

4.1 Datasets Used

The dataset used in this experiment is The Wisconsin Breast Cancer(original) dataset from the UCI Machine Learning Repository. It contains 569 instances and has 33 features. Out of which 357 samples are benign and 212 samples are malignant [10].

4.2 Experimental Procedure

The data cannot be directly used to train the models. Therefore, it must be passed through the pre-processing step. In the pre-processing step, the values of each feature are scaled in such a way that it has a mean of 0 and step deviation of 1. This is required because features having more variance can dominate other features having low variance. The experiment has been divided into four analyses. Each analysis has been further divided into subparts. All of them are discussed below:

The first part of the first analysis was for comparing the performances of Logistic Regression, Decision Tree and KNN. The first step was to clean the data to get rid of any missing values. In the next step the data is pre-processed. In the following step the dataset is split into training and testing data then individual models based on these techniques were created. First, we trained each of these models on training data then we checked their accuracy scores on testing data.

The second part of the first analysis was for comparing the performances of SVM based on the kernel functions used. So, here also individual models based on Linear, RBF, Poly and Sigmoid kernel functions were created. After splitting the dataset into training and testing data, each of these models were trained on the training data and then calculated their accuracy scores on the testing data.

In the second analysis we made use of 10-fold cross validation. Cross validation helps in estimating the skill of machine learning models and how it gives more estimates of out-of-sample accuracy. In the first part, after preprocessing the data, models based on Logistic Regression, Decision tree and KNN were created. Then cross validation is used for training and testing. In the end, the accuracies of each model obtained in this analysis were compared with the accuracies obtained in the first part of the first analysis. Same process was repeated in the second part of the second analysis for SVM kernels.

In the third analysis we used univariate elimination technique for feature selection and selected only 25 features out of 30 to be included in the training and testing part. Feature selection helps in getting rid of unnecessary features which do not play much role in predictions. In the first part of the third analysis individual models of Logistic Regression, DT and KNN were created. Using 10-fold cross validation we trained each model and calculated their individual accuracies. Then we compared their accuracies not only among them but also with the accuracies obtained in the first part of the second analysis. This comparison gave a broad perspective of whether feature selection has helped improve the performance of a model or not. In its second part same process was repeated for SVM kernels

In the Fourth analysis, we used classifier ensembles. First the data is pre-processed, then using feature selection unnecessary features were removed. Post that individual models based on Logistic regression, DT, KNN, Linear, RBF, Poly and Sigmoid were created. After that using the voting method five classifier ensembles were created. Those ensembles are as follows:

- Logistic Regression + Decision Tree + KNN
- Linear + RBF + Poly + Sigmoid
- All (combining all classifiers)
- KNN + Decision Tree + RBF
- KNN + Tree + Linear

Each of these ensembles using 10-fold cross validation were trained and tested and their accuracies were compared.

5. EXPERIMENT RESULTS

In the first part of the first analysis highest accuracy is shown by KNN for k = 12. Its accuracy stands at 99.47%, precision score, recall score and f1 score are also 99.47. Logistic regression gives an accuracy of 96.28%, precision score is 96.5, recall score is 96.28, f1 score is 96.31. For Decision tree, accuracy is 94.15% at depth = 5. Its precision score is 94.28, recall score is 94.15, f1 score is 94.19. This comparison has been shown in Chart 1.

In the second part, we compared different kernels of SVM. RBF kernel gives the highest accuracy of 99.47%. Its precision score is 99.48, recall score is 99.47 and f1 score is 99.47. This is followed by the Sigmoid kernel which has done the prediction with an accuracy of 97.87%. Then linear and Poly kernels have predicted 96.28% and 95.21% of the test data, respectively. chart 2 gives a pictorial representation of this comparison.

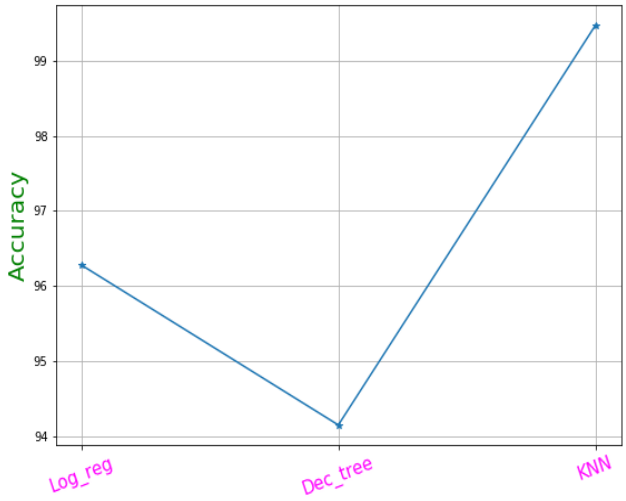

Chart -1: Logistic Regression, Decision Tree and KNN

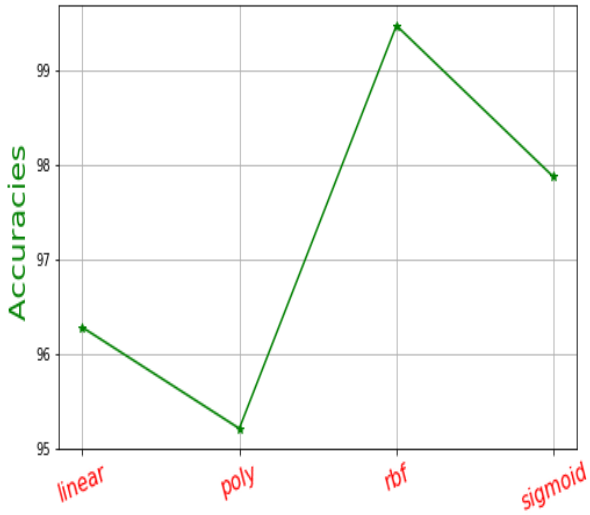

Chart -2: SVM Kernels

In the second analysis, we used 10-fold cross validation. Logistic Regression model not only gives more accuracy as compared to itself in first analysis but also outperforms all other classifiers in second analysis. It has predicted 98.24% of the results correctly. Its precision score recall score and f1 scores are also 98.24. For KNN, the results have been opposite. Its accuracy decreased to 96.84% against 99.47% in the first analysis. The accuracy of the decision tree model has been recorded to be 93.85%. Chart 3 gives the pictorial representation of this comparison.

The second part of the second analysis was for kernels of SVM. Sigmoid kernel tops the list with an accuracy of 97.72% while linear, poly and RBF get the accuracy of 97.36%, 96.31% and 97.19% respectively. Pictorial comparison is shown in Chart 4.

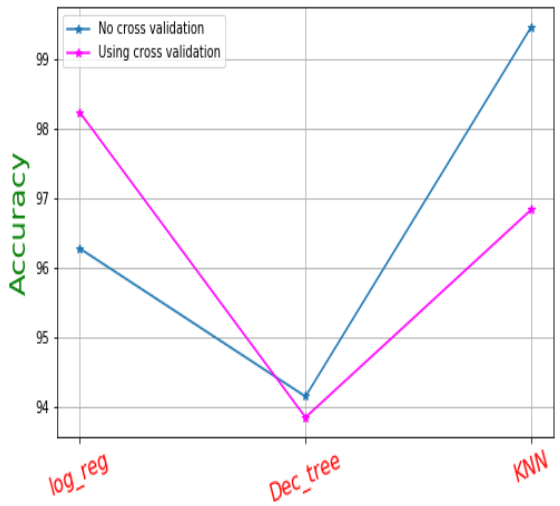

Chart -3: Comparison of accuracies of second analysis with first analysis (for Logistic Regression, Decision tree and KNN)

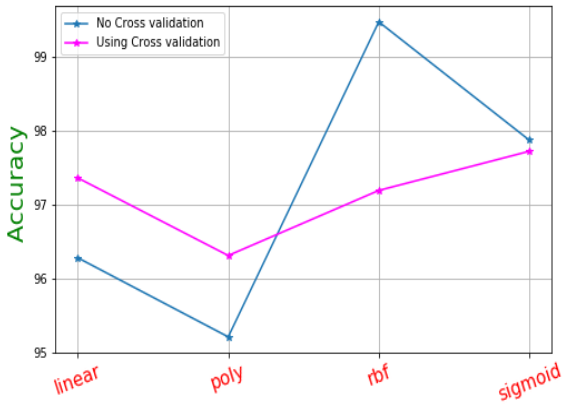

Chart-4: Comparison of accuracies of second analysis with first analysis for different kernels of SVM

In the third analysis we found that many classifiers encountered an increase in the performance as compared to their respective accuracy in the second analysis. In the first part of the third analysis, Logistic model outperforms the rest of them with an accuracy of 98.07% while KNN and Decision tree have done prediction with 97.19% and 94.02% accuracies. The precision score recall score and f1 score of logistic regression are also 98.07.

In its second part, RBF kernel dominates with an accuracy of 98.07% followed by linear (accuracy of 97.36%), sigmoid (accuracy of 97.19%) and poly (accuracy of 95.78 %).
Chart 5 and 6 gives pictorial representation of these comparisons.

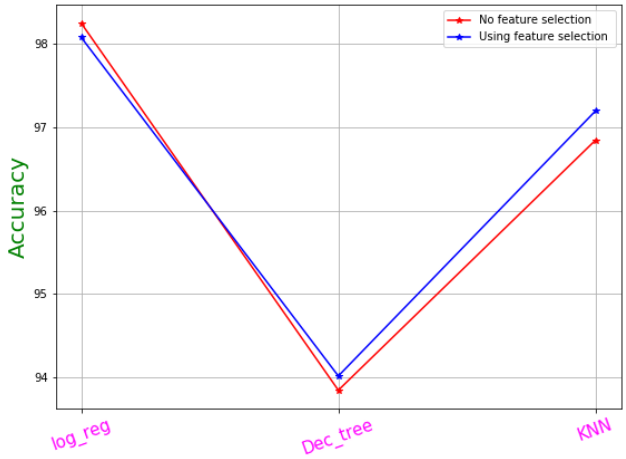

Chart -5: Accuracy comparison using feature selection for log_reg, DT and KNN

The fourth analysis was for classifier ensembles. An ensemble consisting of Logistic Regression, Decision tree and KNN records the highest accuracy of 98.24%. Its precision score is 98.27, recall and f1 scores are both 98.24. Another ensemble composed of RBF, KNN and Decision Tree performs quite well with an accuracy of 98.07%. Its precision, recall and f1 scores are 98.09, 98.07 and 98.06 respectively. Chart 7 gives the pictorial comparison of the above comparison.

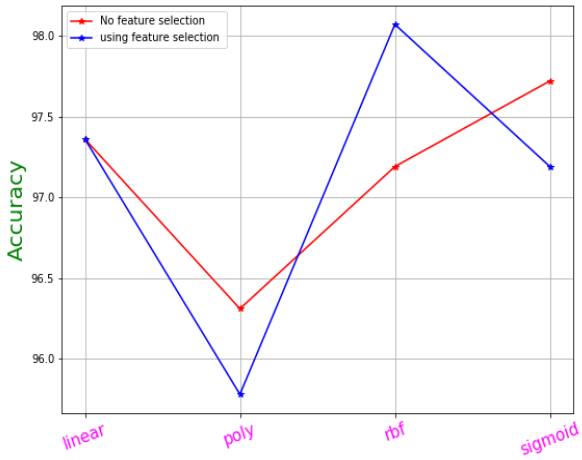

Chart -6: Accuracy comparison using feature selection For SVM kernels .

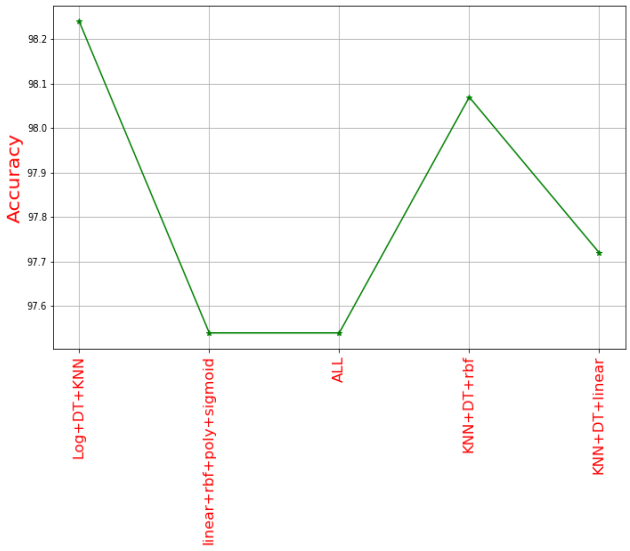

Chart -7: Comparing accuracies of different classifier ensembles

6. CONCLUSIONS

In the proposed work an analysis of different algorithms for breast cancer prediction was carried out. The experiment results indicate that the feature selection has improved the accuracy of the models. Highest accuracy was achieved by the classifier ensemble consisting of Logistic regression + Decision Tree + KNN. It outperforms all the other classifiers by predicting 98.24% of the test cases correctly. Logistic Regression and RBF kernel performs also quite better with an accuracy of 98.24% and 98.07% respectively. In conclusion, these three classifiers have proven their efficiency in Breast Cancer prediction and achieves best performance in terms of accuracy.

REFERENCES

[1] US Cancer Statistics Working Group. "United States cancer statistics: 1999–2012 incidence and mortality web-based report." Atlanta (GA): department of health and human services, centers for disease control and prevention, and national cancer institute (2015).

[2] Siegal, Rebecca, K. D. Miller, and Ahmddin Jemal. "Cancer statistics, 2012." Ca cancer J clin 64.1 (2014): 9-29.

[3] Hofmann, Thomas, Bernhard Schölkopf, and Alexander J. Smola. "Kernel methods in machine learning." The annals of statistics (2008): 1171-1220.

[4] Leon, Florin, Sabina-Adriana Floria, and Costin Bădică. "Evaluating the effect of voting methods on ensemble-based classification." 2017 IEEE International Conference on INnovations in

Intelligent SysTems and Applications (INISTA). IEEE, 2017.

[5] Asri, Hiba, et al. "Using machine learning algorithms for breast cancer risk prediction and diagnosis." Procedia Computer Science 83 (2016): 1064-1069.

[6] Chaurasia, Vikas, and Saurabh Pal. "Data mining techniques: to predict and resolve breast cancer survivability." International Journal of Computer Science and Mobile Computing IJCSMC 3.1 (2014): 10-22.

[7] Eshlaghy, A. T., Poorebrahimi, A., Ebrahimi, M., Razavi, A. R., & Ahmad, L. G. (2013). Using three machine learning techniques for predicting breast cancer recurrence. J Health Med Inform, 4(2), 124.

[8] Delen, Dursun, Glenn Walker, and Amit Kadam. "Predicting breast cancer survivability: a comparison of three data mining methods." Artificial intelligence in medicine 34.2 (2005): 113-127.

[9] Huang, Min-Wei, Chih-Wen Chen, Wei-Chao Lin, Shih-Wen Ke, and Chih-Fong Tsai. "SVM and SVM ensembles in breast cancer prediction." *PloS one* 12, no. 1 (2017).

[10] Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.