# Analyzing Health Care Dataset using Machine Learning Techniques

B. Tamilvanan
Research and Development Centre,
Bharathiar University,
Coimbatore-641046,
TN, India.

V. Murali Bhaskaran, PhD
Principal, Dhirajlal Gandhi
College of Technology, Salem-636290,
TN, India

## ABSTRACT
This paper mainly deals with different classification algorithms techniques namely Navie Bayes, Sequential Minimal Optimization, Multilayer Perception, and Random Forest. It analyses the breast cancer from UCI machine learning repository. The result of the classification model is precision, recall, F-Measure, time, accuracy. From theses measure, it is observed that naive Bayes algorithms are able to achieve high accuracy and consumed very less time when compare other algorithms.

## General Terms
Data Mining, Classification, Breast Cancer

## Keywords
Navie Bayes, Sequential Minimal Optimization, Multilayer Perception, Random Forest.

## 1. INTRODUCTION
Breast cancer is a harmful cell development in the bosom. On the off chance that left untreated, the malignancy spreads to different territories of the body. Excluding skin cancer, breast cancer is the most common type of cancer in women in the world. The incidence of breast cancer rises after age 40. The highest incidence (approximately 80% of invasive cases) occurs in women over age 50. Ninety percent of breast cancer is adenocarcinomas, which arise from glandular tissue. Within this broad category, there is a great degree of variation. For instance, there are about 30 different subtypes of adenocarcinoma.

The soonest type of the infection, ductal carcinoma in situ, includes around 15-20% of all bosom malignancies and grows exclusively in the drain channels. The most common type of breast cancer, invasive ductal carcinoma, develops from ductal carcinoma in situ, spreads through the duct walls, and invades the breast tissue. The Breast Cancer that begins in the lobes or lobules is called lobular (small cell) carcinoma and is more likely to be found in both breasts. Invasive lobular carcinoma originates in the milk glands and accounts for 10-15% of invasive breast cancers. Both ductal and lobular carcinomas can be either in situ or independent; or invading, which means entering the mass of the channel or flap and spreading to adjoining tissue. Nowadays health care industry generates a large amount of data about patients, disease diagnosis, etc. Some extraordinary sorts of ways to deal with building exact groupings have been proposed Naive Bayes, Multi-Layer Perception, Sequential Minimal Optimization and Random Forest.

This paper is organized accordingly: the relates works and depiction of the specialized parts of the utilized information mining techniques in section 1. The elaborates with classification algorithms like navie bayes, multi-layer perception, Sequential minimal optimization and random forest in section 2.The introduction of the dataset for Breast Cancer in section 3. The Experiment Results and Discussion in section 4. And finally, conclude the paper and future works.

## 2. CLASSIFICATION
### 2.1 Naive Bayes
A Naive Bayesian classifier in light of Bayes hypothesis is a probabilistic factual classifier. Here, the expression "credulous" demonstrates contingent autonomy among components or traits. The "credulous" presumption enormously diminishes calculation intricacy to a straightforward augmentation of probabilities. The real preferred standpoint of the Naive Bayesian classifier is its quickness of utilization. This rate happens in light of the fact that it is the least complex calculation among grouping calculations. On account of this effortlessness, it can promptly handle information set with many properties. Moreover, the guileless Bayesian classifier needs the just little arrangement of preparing information to create exact parameter estimations since it requires just the computation of the frequencies of characteristics and trait result combines in the preparation information set. A noteworthy downside of this calculation is its major presumption that all characteristics are free each other. As a rule, this supposition is impossible. For instance, in the restorative field, numerous patient indications and wellbeing conditions are unequivocally related each other (e.g., pulse and body mass record (BMI)), which may bring about some deviation in the subsequent characterization. By and large, notwithstanding, the utilization of the gullible Bayesian classifier delivers great execution as far as arrangement exactness, in spite of infringement of the characteristic autonomy supposition and is, all things considered, generally utilized as a part of medicinal information mining[6, 7].

### 2.2 Multi Layer Perception
The multilayer perceptron (MLP) consists of multiple layers of simple, two-state, sigmoid processing elements (nodes) or neurons that interact using weighted connections. After a lowermost input layer, there are usually any number of intermediate, or hidden, layers followed by an output layer at the top. There exist no interconnections within a layer while all neurons in a layer are fully connected to neurons in adjacent layers. Weights measure the degree of correlation between the activity levels of neurons that they connect.[17]

### 2.3 Random Forest
Random Forest developed by Leo Breiman [15] is a group of un-pruned classification or regression trees made from the random selection of samples of the training data. Random features are selected in the induction process. Prediction is made by aggregating (majority vote for classification or

averaging for regression) the predictions of the ensemble. Each tree is grown as described in [14]:

Step 1: By Sampling N randomly, If the number of cases in the training set is N but with replacement, from the original data. This sample will be used as the training set for growing the tree.

Step 2: For M number of input variables, the variable m is selected such that m<<M is specified at each node, m variables are selected at random out of the M and the best split on this m is used for splitting the node. During the forest growing, the value of m is held constant.

Step 3: Each tree is grown to the largest possible extent. No pruning is used.

Random Forest generally exhibits a significant performance improvement as compared to single tree classifier such as C4.5. The generalization error rate that it yields compares favorably to Adaboost, however, it is more robust to noise[16].

## 3. DATA PREPROCESSING

The dataset utilized as a part of this model ought to be more exact and precise so as to enhance the prescient time and exactness of information mining. The dataset which is gathered may have lost or insignificant traits.

### 3.1 Attribute Identification

**Table 1. Attribute details of the Brest Cancer**

| Attributes Name | Description |
|---|---|
| Age | Age (years) |
| Inv-Nodes | 0-2, 3-5, 6-8, 9-11, 12-14, 15-17, 18-20, 21-23, 24-26, 27-29, 30-32, 33-35, 36-39 |
| Node-Caps | yes, no. |
| Menopause | lt40, ge40, premeno |
| Tumor-Size | 0-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59 |
| Deg-Malig | 1, 2, 3. |
| Breast | left, right |
| Breast-Quad | left-up, left-low, right-up, right-low, central. |
| Irradiat | yes, no. |
| Class | no-recurrence-events, recurrence-events |

## 4. EXPERIMENTAL RESULTS AND DISCUSSIONS

Some measure of evaluating performance has to be introduced. One common measure in the literature is accuracy defined as correct classified instances divided by the total number of instances.

**A. Precision, Recall, F-Measure, Accuracy**

A single prediction has the four different possible outcomes shown in Table 2 for the true positives (TP) and true negatives (TN) are correct classifications. A false positive (FP) occurs when the outcome is incorrectly predicted as yes (or positive) when it is actually no (negative). A false negative (FN) occurs when the outcome is incorrectly predicted as no when it is actually yes. In this paper, we use following equation to measure the Precision Eq. (1), Recall Eq. (2), F-Measure Eq. (3), Accuracy Eq. (4).

$$\text{Precision} = TP / (TP + FP) \quad (1)$$

$$\text{Recall} = TP / (TP+FN) \quad (2)$$

$$\text{F-Measure} = 2 * (\text{PRECISION} * \text{RECALL}) / (\text{PRECISION} + \text{RECALL}) \quad (3)$$

$$\text{Accuracy} = (TP+TN) / (TP+TN+FP+FN) \quad (4)$$

**Table 2. Different Outcome of Two Class Prediction**

| Actual Class | | Predicated Class | |
|---|---|---|---|
| | | a | b |
| | a | Ture Positive | False Negative |
| | b | False Positive | True Negative |

The Precision, Recall, F-Measure and Accuracy for Naïve Bayes, SMO, Multilayer Perception and Random Forest are shown in Table 3 and Table 4.

**Table 3. Comparison of Precision, Recall, F-Measure**

| Classifier | Precision | Recall | F-Measure |
|---|---|---|---|
| NavieBayes | 83% | 71% | 70% |
| SMO | 85% | 69% | 67% |
| MLP | 74% | 64% | 64% |
| Random Forest | 86% | 69% | 67% |

**Table 4. Comparison of Accuracy**

| Classifier | Time Taken (Seconds) | Accuracy |
|---|---|---|
| NavieBayes | 0.00 | 71% |
| SMO | 0.09 | 69% |
| MLP | 2.18 | 64% |
| Random Forest | 0.1 | 69% |

**B. Confusion Matrix**

A confusion matrix is calculated for Naive Bayes classifiers to interpret the results. The confusion matrix is shown in table 5.

**Table 5. Confusion matrix for Naive Bayes**

| a | b | Classified as |
|---|---|---|
| 168 | 33 | a = no-recurrence-events |
| 48 | 37 | b = recurrence-events |

**C. Graph Results**

The graph results in Figure 1. Shows performance analysis related to the accuracy of various algorithms.

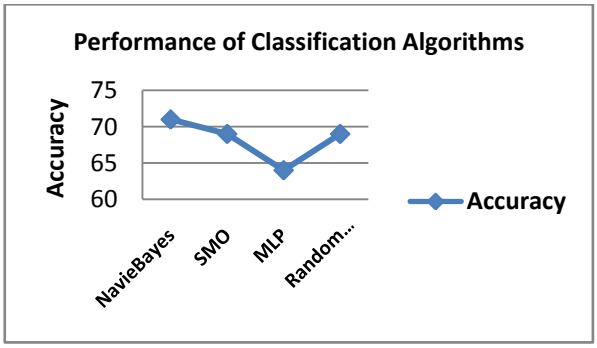


**Figure 1. Performance related to Accuracy**

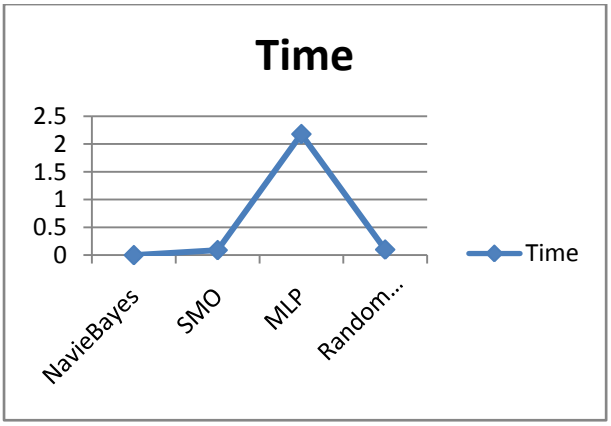


**Figure 2. Performance related to Time**

## 5. CONCLUSION

In this work conducted and analyzed breast cancer dataset to determine their classification accuracy and calculate the time taken to build the model and found the same statistics. The breast cancer data used in this research were selected for completeness and for each classification of breast cancer patients. From the experiments of results prove the Navie Bayes high accuracy and consumed very less time using WEKA tool. The limitation is that it consider the only small amount of dataset to detect the virus. The complex terminology is required to predict the results more accuracy. The recommendations arise from this research implies the data mining algorithms may be applied in the field of medical research in future as the will provide another research tool for comparison of the huge amount of the dataset.

In future, it is possible to extend the research work by using different clustering, association rule mining for a huge amount of data available in the healthcare industry.

## 6. REFERENCES

[1]. Arun K. Pujari, Data Mining Techniques, *University Press (India) Ltd*, 2001.

[2]. Jiawei Han, Micheline Kamber, Data Mining Concepts and Techniques, *Elsevier*.

[3]. Klosgen W, Zytkow JM, Handbook of Data mining and Knowledge Discovery, *Oxford University Press*, 2002.

[4]. M.S.Chen, J.hans, P.SYu, Data mining: A overview from a data base perspective, *IEEE transaction on Knowledge and data engineering* 8(6), pp. 866-883, 1996.

[5]. Quinlan, J. R., C4.5: programs for machine learning. Morgan Kaufmann, Amsterdam, 1993.

[6]. Wu X, Kumar V, Quinlan JR, Ghosh J, Yang Q,MotodaH, McLachlan GJ, Ng A, Liu B, Yu PS, Zhou Z-H, Steinbach M, Hand DJ, Steinberg D, Top 10 algorithms in data mining. *Knowl Inf Syst* 14, pp.1–37, 2008.

[7]. Diana Dumitru, Prediction of recurrent events in breast cancer using the Naive Bayesian classification, *Annals of University of Craiova, Math. Comp. Sci. Ser.* Volume 36(2), 2009.

[8]. Nurnberger A, Pedrycz W, Kruse R, Neural network approaches. In: Klosgen W, Zytkow JM (eds) Handbook of data mining and knowledge discovery. *Oxford University Press*, 2002.

[9]. Hammerstrom D, Neural networks at work. *IEEE Spectr:*pp.26–32 (June), 1993.

[10]. Delen, D., Walker, G., and Kadam, A., Predicting breast cancer survivability: a comparison of three data mining methods. *Artif. Intell. Med.* 34, pp.113–127,2005.

[11]. Kaur, H., and Wasan, S. K., Empirical study applications of data mining techniques in healthcare. *J. Comput. Sci.* 2(2), pp. 194–200, 2006.

[12]. Ubeyli, E. D., Comparison of different classification algorithms in clinical decision making. *Expert Syst* 24(1), pp. 17–31, 2007.

[13]. Schwarzer, G., Vach, W., and Schumacher, M., On the misuses of artificial neural networks for prognostic and diagnostic classification in oncology. *Stat. Med.* 19, pp. 541–561, 2000.

[14]. http://www.stat.berkeley.edu/~breiman/Random Forests/cc_home.htm#prox Symposium, volume 1, July, 2005.

[15]. Breiman, L., Random Forests, Machine Learning 45(1), 5-32,2001.

[16]. Jehad Ali, Rehanullah Khan, Nasir Ahmad, Imran Maqsood, Random Forests and Decision Trees, International Journal of Computer Science Issues, Vol. 9, Issue 5, No 3, September 2012.

[17]. Sankar K. Pal, Multilayer Perceptron, Fuzzy Sets, and Classification, IEEE transactions on neural networks, vol. 3, no. 5, September 1992.

[18]. https://training.seer.cancer.gov/breast/intro/types.html.