*International Journal for Modern Trends in Science and Technology*, 6(12): 282-285, 2020

ISSN: 2455-3778 online
DOI: https://doi.org/10.46501/IJMTST061252
Available online at: http://www.ijmtst.com/vol6issue12.html

# Predicting Breast Cancer Classification Using Various Machine Learning Classification Algorithm

Nishant Bansal[1] | Nidhi Sengar[2] | Amita Goel

[1]B. Tech Scholar, Information Technology Department, Maharaja Agrasen Institute of Technology, New Delhi, India
[2]Assistant Professor, Department of IT, Maharaja Agrasen Institute Of Technology, Delhi
[3]Professor, Department of IT, Maharaja Agrasen Institute Of Technology, Delhi

**To Cite this Article**
Nishant Bansal, Nidhi Sengar and Amita Goel, "Predicting Breast Cancer Classification Using Various Machine Learning Classification Algorithm", *International Journal for Modern Trends in Science and Technology*, 6(12): 282-285, 2020.

**Article Info**
Received on 10-November-2020, Revised on 02-December-2020, Accepted on 06-December-2020, Published on 11-December-2020.

## ABSTRACT

*Cancer diagnosis is one among the foremost studied problems within the medical domain. Several researchers have focused so as to enhance performance and achieve to get satisfactory results. Breast cancer[1] represents the second primary explanation for cancer deaths in women today and has become the foremost common cancer among women both within the developed and therefore the developing world in the last years. Breast cancer diagnosis is used to categorize the patients among benign (lacks ability to invade neighbouring tissue) from malignant (ability to invade neighbouring tissue) categories. In this study, the diagnosis of breast cancer from mammograms is complemented by using various classification techniques. In artificial intelligence, machine learning is a discipline which allows to the machine to evolve through a process. Machine learning[2] is widely utilized in bio-informatics and particularly in carcinoma diagnosis. This paper explores the various data processing approaches using Classification which may be applied on carcinoma data to create deep predictions. Besides this, this study predicts the simplest Model yielding high performance by evaluating dataset on various classifiers.[4-8] The results that are obtained through the research are assessed on various parameters like Accuracy, RMSE Error, Sensitivity, Specificity etc. Our work is going to be performed on the WBCD database (Wisconsin carcinoma Database) [12]obtained by the university of Wisconsin Hospital.*

***KEYWORDS:****Breast Cancer Data Classification, Decision Trees (DT), Logistic Regression, Prediction, K-Nearest Neighbors, mammograms.*

## I. INTRODUCTION

Breast Cancer is the most identified cancer and is the prime source of cancer demise amid women. Early detection of cancer is important for a rapid response and better chances of cure. First procedure in breast examination is to undergo mammographic screening which is to assist with the diagnosis of the disease in women. Mammography is of the most renowned screening method as it is widely available and is also relatively fast. Diagnostic mammography is employed to gauge a patient with abnormal clinical examination results. The results detected on the mammogram are mass, architectural distortion, skin thickening, and calcification (Balleyguier, 2007; Eltoukhy, 2010; Moezzi, 1996). Radiologists can predict the condition of the patient from the results of the mammogram. Unfortunately, early detection of cancer is usually difficult because the symptoms of the disease at the start are absent. Thus, cancer remains one among the topics of health research, where many researchers have invested the aim of making evidence which will improve treatment, preventions and diagnostics.

Earlier detection of cancer can save many lives during a best effective manner.

**Structure of Paper**

The paper is organized as follows: In Section 1, the introduction of the paper is provided along with the structure, important terms, objectives and overall description. In Section 2 we have the complete information about the k-Nearest Neighbour Algorithm. Section 3 and 4 share information about Decision Trees algorithm and Logistic Regression algorithm respectively. Section 5 concludes the paper with results and conclusions and Section 6 ending the paper with references.

**Important Terms Used**

Decision Trees[10]- It may be a decision support tool that uses a tree-like model of selections and their possible consequences, including utility, resource costs and chance event outcomes. It is a method to display an algorithm that only contains conditional control statements.

Logistic Regression- Logistic regression[9] can be defined as a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although more complex extensions exist. In multivariate analysis, logistic regression (or logit regression) is estimating the parameters of a logistic model (a sort of binary regression).

Mammograms- Mammography is the process of using low-energy X-rays to study human breast for diagnosis and screening. The goal of mammography is that the early detection of carcinoma, typically through detection of characteristic masses or microcalcifications. A mammogram is an X-ray of the breast. Doctors use a mammogram to see for early signs of carcinoma. Regular mammograms are the simplest tests doctors need to find carcinoma early, sometimes up to 3 years before it is actually felt.

K-Nearest Neighbours- K nearest neighbors can be defined as a simple algorithm that stores all available cases and classifies new cases on the basis of a similarity measure (e.g., distance functions). KNN has been utilized in statistical estimation and pattern recognition already in the beginning of 1970's as a non-parametric technique

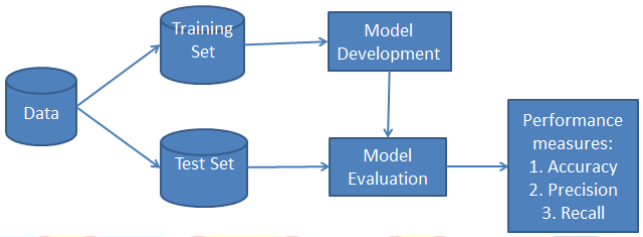
**Objectives**

The present work is intended to meet the following objectives:

1. Summarize Breast cancer – Types, Risk Factors, Symptoms, Diagnosis and Treatment.
2. Test on a number of possible classification techniques that can be applied on diagnosis of breast cancer.

**Overall Description**

In this Paper we have selected the open source software Jupyter Notebook, proficiently works with limited data . Few tasks offered by data mining like pre-processing of data, Classification, Clustering, Association and Visualization can be performed. Data set is fed in the form of Comma Separated Values-CSV format. In general Jupyter Notebook is used in analysis for preliminary collection of data . A decompositional method is applied to discover information from the Breast Cancer dataset. The extracted knowledge is used for prediction purposes .




## II. K-NEAREST NEIGHBORS METHOD

The k-nearest neighbors algorithm is one of the most used algorithms in machine learning . It is a learning method

bases on instances that does not required a learning phase.

$$E(S) = \sum_{i=1}^{c} - p_i \log_2 p_i$$

| Play Golf | |
|---|---|
| Yes | No |
| 9 | 5 |

**Entropy(PlayGolf)** = Entropy (5,9)
= Entropy (0.36, 0.64)
= - (0.36 $\log_2$ 0.36) - (0.64 $\log_2$ 0.64)
= 0.94

The training sample, associated with a distance function and therefore the

choice function of the class based on the classes of nearest neighbors be classified. The neighbors are weighted by the distance that separate it to the new elements to classify.

**The K-Nearest Neighbors Algorithm**

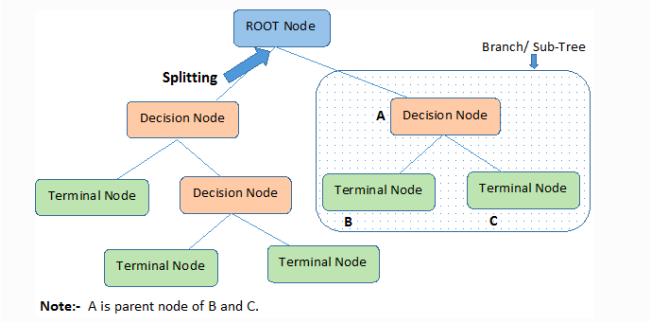
Choose a value for the parameter k:

Input : Give a sample of N examples and their classes.

The class of a sample x is c(x):
Give a new sample y:
Determine the k-nearest neighbors of y by calculating the distances.
Combine classes of these y examples in one class c
Output : The class of y is c(Y ) = c

**THE CHOICE OF THE PARAMETER K (THE NUMBER OF NEAREST NEIGHBORS)**

The choice of the parameter k (k 2 N) is determined by the user, this choice depends on the data. The effect of noise on the classification is reduced when the values chosen for k is bigger, but this makes the boundaries between classes less distinct. A good choice of the value of k can be selected by different heuristic techniques such as cross-validation. In this study we choose the value of k that minimizes the classification error. In the case of a binary classification, it is more inviting to choose an odd value for k, it avoids the equal




votes. In case of equality, we can increase the value of k from 1 to decide [11].

## III.Decision Trees

Decision tree is a classifier that is expressed as a recursive partition of the instance space. It creates a predictive model, which maps observations of a few node to conclusions about the nodes' target value. In a tree structure leaves represent the class labels and branches represent conjunctions of feature leading to the class labels. The word "data" is plural, not singular.

**Procedure Of Decision Tree Type ID3 Algo:**

1. It begins with the original set S as the root node.

2. On each iteration of the algorithm, it iterates through the very unused attribute of the set S and calculates Entropy(H) and Information gain (IG) of this attribute.

3. It then selects the attribute which has the smallest Entropy or Largest Information gain.

4. The set S is then split by the selected attribute to produce a subset of the data.

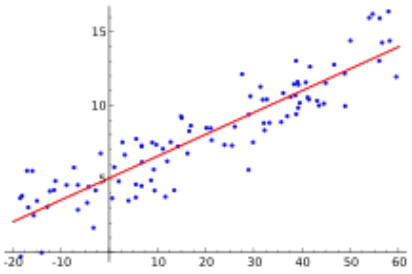
5. The algorithm continues to recur on each subset, considering only attributes never selected before.

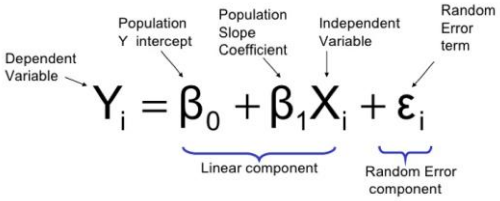
$$Information\ Gain = Entropy(before) - \sum_{j=1}^{K} Entropy(j, after)$$

## IV. Logistic Regression

The logistic model is employed to model the probability of a particular class or event existing like pass/fail, win/lose, alive/dead or healthy/sick. This can be extended to model several classes of events like determining whether a picturecontains a cat, dog, lion, etc. Each object being detected within the image would be assigned a probability between 0 and 1, with a sum of 1 .

Mathematically, a binary logistic model features a variable with two possible values, like pass/fail which is represented by an indicator variable, where the 2 values are labelled "0" and "1". In the logistic model, the log-odds (the logarithm of the odds) for the value labelled "1" is a linear combination of one or more independent variables ("predictors"); the independent variables can each be a binary variable (two classes, coded by an indicator variable) or endless variable (any real value). The corresponding probability of the worth labelled "1" can vary between 0 (certainly the worth "0") and 1 (certainly the worth "1"), hence the labelling; the function that converts log-odds to probability is that the logistic function, hence the name. The unit of measurement for the log-odds scale is named a logit, from logistic unit, hence the choice names. Analogous models with a special sigmoid function rather than the logistic function also can be used, like the probit model; the defining characteristic of the logistic model is that increasing one among the independent variables multiplicatively scales the chances of the given outcome at a continuing rate, with each experimental variable having its own parameter; for a binary dependent variable this generalizes the odds ratio.






**How Puppeteer Outweighs Selenium**

The logistic curve relates the independent variable, X, to the rolling mean of the DV, P ( $\overline{Y}$ ). The formula to do so may be written as

$$P = \frac{1}{1 + e^{-(a+bX)}}$$

where P is the probability of a 1 (the proportion of 1s, the mean of Y), e is the base of the Napierian logarithm (about 2.718) and a and b are the parameters of the model. The value of ayields P when X is zero, and b adjusts how quickly the probability changes with changing X one unit (we can have standardized and unstandardized b weights in logistic regression, even as in ordinary linear regression). Because the relation between X and P is nonlinear, b doesn't have an easy interpretation during this model because it does in ordinary rectilinear regression.

## V. Results and Conclusion

| *Methods* | *Accuracy* | *Precision (0)* | *Precision (1)* |
|---|---|---|---|
| ***K-NN*** | ***0.95*** | ***0.94*** | ***0.97*** |
| ***Decision Tree*** | ***0.95*** | ***0.94*** | ***0.96*** |
| ***Logistic Regression*** | ***0.96*** | ***0.95*** | ***0.97*** |

Data Mining might not be able to predict the type of tumor the patient is affected by within the Medical field as there's increase within the number of females suffering from Breast cancer. To predict the class of cancer to which a patient may classified, we need to extract the hidden knowledge pertaining to various attributes that could be used to boost the efficiency in general by utilizing the best resources available. In our paper, Comparing the efficiency of different classifiers; Logistic regression, Decision Tree, K-Nearest neighbours, Results conclude that Simple Logistic regression method obtains the Best Model to predict breast cancer by means of different data mining techniques.

In future work rule extraction part can be improved, also try to improve accuracy in Decision Tree. To obtain absolute best rules with the assistance of RSES tool, Rough Sets algorithms like LEM2 are often utilized in DT, NBTREE. Future research work also suggests that this work are often extended to three-class Classification. from left to right then moving right down to subsequent line

## VI.REFERENCES


[1] National Cancer Institute: http://www.cancer.gov/cancertopics/types/breast.

[2] Han J., Kamber M., Data Mining Concepts and Techniques. Morgan Kaufman Publishers, 2001.

[3] McCarthy et al. Applications of Machine Learning and High -Dimensional Visualization in Cancer Detection, Diagnosis, and Management. (2004).

[4] Kesavaraj G, Sukumaran S. A Study on Classification Techniques in Data Mining. 1 4th ICCCNT (2012).

[5] Soundarya M, Balakrishnan R. Survey on Classification Techniques in Data mining. International Journal of Advanced Research in Computer and Communication Engineering Vol.3:7550–7552 (2014).

[6] Li J, Wong L. Rule-Based Data Mining Methods for Classification Problems in Biomedical Domains; 15th European Conference on Machine Learning (ECML) (2004).

[7] Kumar D, Beniwal S. Genetic Algorithm and Programming Based Classification: A Survey. Journal of Theoretical and Applied Information Technology. 54:48–58 (2013).

[8] Mansuri AM, Verma M, Laxkar P. A Survey of Classifier Designing Using Genetic Programming and Genetic Operators. International Journal of Engineering Research and Reviews (IJERR) Vol. 2:16–22 (2014).

[9] le Cessie, S., van Houwelingen, J.C. (1992). Ridge Estimators in Logistic Regression. Applied Statistics. 41(1):191-201.D. Aha, D. Kibler Instance-based learning algorithms. Machine Learning. 6:37-66 (1991).

[10] Kohavi R. The Power of Decision Tables. In: 8th European Conference on Machine Learning, 174-189, (1995)

[11] D. Bremner, E. Demaine, J. Erickson, J. Iacono, S. Langerman, P. M., and Godfried. Output-sensitive algorithms for computing nearest-neighbour decision boundaries. Discrete and Computational Geometry, 33(4), 2005.

[12] O. L. Mangasarian and W. H. Wolberg. Cancer diagnosis