INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact factor: 4.295
(Volume 5, Issue 3)
Available online at: www.ijariit.com

Breast cancer prediction algorithms analysis

Pooja Mudgil
engineer.pooja90@gmail.com
Bhagwan Parshuram Institute of Technology, New Delhi, Delhi

Mohit Garg
mohitgarg701@gmail.com
Bhagwan Parshuram Institute of Technology, New Delhi, Delhi

Vaibhav Chhabra
vaibhavchhabra97@gmail.com
Bhagwan Parshuram Institute of Technology, New Delhi, Delhi

Parikshit Sehgal
parikshit4497@gmail.com
Bhagwan Parshuram Institute of Technology, New Delhi, Delhi

Jyoti
jyoti.mj868@gmail.com
Bhagwan Parshuram Institute of Technology, New Delhi, Delhi

ABSTRACT

Machine learning which an application of Artificial intelligence (AI) is makes the system capable to automatically learn through the environment without being explicitly programmed. It is widely used in various domains like classification and prediction processes. This paper basically compares classifier algorithms like-Naïve Bayes, K Nearest Neighbour, Decision tree, Logistic Regression, Random Forest, Support Vector Machine (SVM). These algorithms predict chances of breast cancer and are programmed in python language. The implementation procedure shows that the performance of any classification algorithm is based on the type of attributes of datasets and their characteristics. The main aim of this paper is to do the comparison of these algorithms on the basis of the accuracy. The goal is to classify whether breast cancer is "Benign" or "Malignant".

Keywords— *Machine learning, Classification, Naïve Bayes, K Nearest Neighbour, Decision tree, Logistic regression, Random forest*

1. INTRODUCTION

Breast cancer (BC) is considered as the most common cancers, resulting majority of new cancer cases and cancer-related deaths according to global statistics, making it a significant public health problem in today's society [1]. The early diagnosis of BC can improve the prognosis and chance of survival significantly, as it can promote timely clinical treatment to patients. Benign Tumours can be classified in such a way that can prevent patients. So, the diagnosis of BC and the classification of patients into malignant or benign is really a matter of concern. Because of its unique advantages in critical features detection from complex BC datasets, machine learning (ML) is widely recognized as the methodology of choice in BC pattern classification and forecast modelling. Especially in the medical field, where those methods are widely used in diagnosis and analysis to make decisions. There are various risk factors for BC such as Age, family history, genetic factors, menstrual or childbearing history, etc. The most important screening test that is Mammogram is an X-ray of the breast, it can detect the risk of cancer 2 years before any doctor or felt by the patient, it should be done by women of 40-45 years once in a year. All the algorithms which are taken have their own strengths and weaknesses based on the type of data input and tools which are used for the implementation of the algorithms. There are various tools which are available for implementing machine learning. Scikit-learn is a very powerful python library which features various classification, regression and clustering algorithms including support vector machines, random forests, gradient boosting, k-means.

2. CLASSIFICATION ALGORITHMS

2.1 Naive Bayes

This model comes under the classification technique. It is based on the Bayes Theorem which is in probability with an assumption that independence between predictors will be there. A Naive Bayes classifier assumes that the presence of a specific feature in a class is not related to the presence of any other feature [2]. It is a probability or statistical-based approach which comes under the concept of supervised learning. In this basically guessing part is done for example a diagnosis done by a doctor. Bayes theorem tells the probability of an event based on conditions which are priory known that might be related to that event. Mathematically (E/F) = (P(F/E)*P(E))/P(F). P(E/F) is the probability of E occurring if F has already occurred. The Bayesian theorem is proved better than any other probabilistic approach this is the reason it is used in machine learning. This model is easy to build and it is beneficial where a large set of data is present. A dependency graph is first made in a Naive Bayes model after that implementation is done. For example,

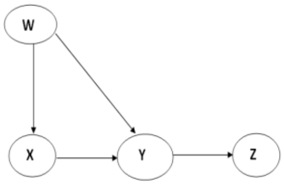

Fig. 1: Example to show a dependency graph

Here, W is dependent on the occurrence of X.
W is dependent on the occurrence of Y.
X is dependent on the occurrence of Y.
Y is dependent on the occurrence of Z.
All W, X, Y, Z are independent events.

2.2 Decision Tree

It comes under the category of supervised learning. Regression and classification problems can be solved by a Decision tree. It represents a problem in the form of a tree in which internal nodes of the tree serve as attributes and leaf node serve as class label[3]. Splitting is done in this model to divide a node into two or more sub-nodes. Below are some assumptions for making a decision tree:

- At first, the whole training set is considered as root.
- Secondly, Values are preferred to be categorical.
- Records are distributed recursively on the basis of attribute values.
- Statistical methods are used for ordering attributes as internal node or root.

There are two types of decision trees:
(a) Categorical Variable Decision Tree: Decision Tree which has a categorical target variable. In simple words, where Yes or NO values are there.
(b) Continues Variable Decision Tree: Decision Tree which has a continuous target variable. From the below example, it will be clearer.

Example: Let’s say we have a problem to predict whether a customer will pay his renewal premium with an insurance company (yes/ no). Here we know that the income of customer is a significant variable but the insurance company does not have income details for all customers. Now, as we know this is an important variable, then we can build a decision tree to predict customer income based on occupation, product and various other variables. In this case, we are predicting values for continuous variable [4].

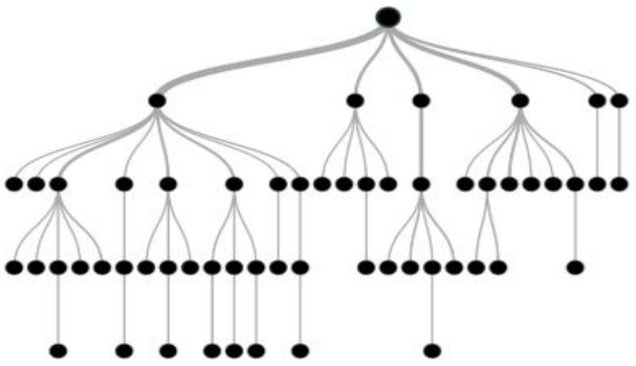

Fig. 2: Pictorial view of decision tree [4]

2.3 Random Forest

A decision tree is basically a set of decisions which is used to classify the datasets. It is also known as a classification and regression tree (CART) [3]. Random forest is a collection of decision trees. Random forest iteratively asks a series of questions and based on that answer it will ask another set of questions to classify the data.

Prediction using train random forest algorithm
(a) Take the test data sets and then randomly creates the decision trees.
(b) Find the decision of each decision tree according to the majority vote and then take a decision.
(c) Choose the high voted class as a final decision.

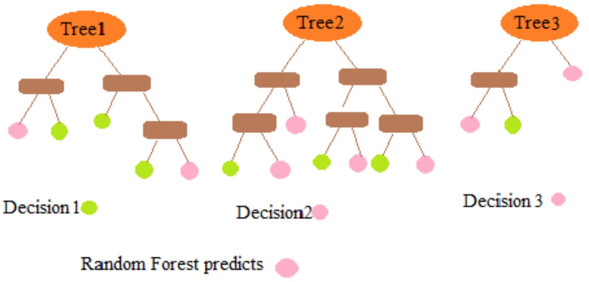

Fig. 3: Pictorial view of Random Forest [6]

Why the random forest?

If we build a decision tree that the decision tree can be over fit. So, multiple decision trees build a random forest, the way random forest works is that each of the decision tree that is generated once we give data points as inputs which process through all the decision trees and they all make a prediction and then make a majority decision or calculation. It gives higher accuracy which is not possible in just using a single decision tree. The more decision trees the better accuracy we get. Random forest is a more efficient and better option for the larger datasets [6].

2.4 K-nearest Neighbour

K nearest neighbour is used for the classification as well as regression. It is a supervised learning algorithm. It is an instance-based learning in which we take the training examples but don't process them instead we store the training example when we need to classify an instance at that time we do classification, that is why it is also known as a lazy algorithm. It uses the Stored instance in order to find the possible output [8][9]. Phases of KNN:
(a) Training phase:
– Save the training sets.
(b) Prediction
– Get the test instance $x_{i.}$
– Find K training sets $\{(x_1,y_1),(x_{2,}y_2),...(x_{k,}\ y_k)\}$Which are closest to x.
– Predict y_1 as the output of $y_{i.}$
(c) Classification
Predict the majority class $(y_1,y_{2,...}y_k)$.
(d) Regression: Predict the average of $(y_1,y_{2,...}y_k)$. Euclidean distance is the Sum of squared difference. we will go for averaging in the following condition:
– Noise in the attribute.
– Noise in class labels.
– Classes may be overlapping.

Averaging is needed in the case of the regression in KNN. KNN regression uses the same distance functions as the KNN Classification. Example:

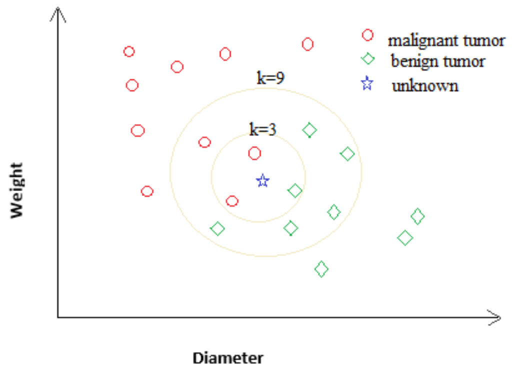

Fig. 4: Example graph Weight V/s Diameter [8]

In the given example we have two classes of cancer tumour i.e. malignant and benign, if we have an unknown data point then by taking the value of k, we can choose the class of that unknown data point.

In this particular example if the value of k=3 then the unknown data point's class will be a malignant tumour. If the value of k=9 then it will be a benign tumour.

- If the value of k is larger then we get better and smoother classifier but less sensitive for discrete classes.
- If the value of k is smaller than it captures the fine structure of problem space better.
- It may be necessary for small training sets.

2.5 Logistic Regression

Logistic Regression is a subcategory of regression analysis. Regression Analysis consists of a group of machine learning algorithms that enables us to make continuous predictions regarding outcome/ dependent variable (y) on the basis of one or multiple independent/predictor variable (x).

Regression Analysis is a form of predictive modelling technique which investigates the relationship between the dependent variable by defining a mathematical equation showing the defining the dependent variable as a function of the independent variable. Subsequently the equation can be used to predict the outcome/dependent variable (y) on the basis of the new independent/predictor variable (x). Three major uses of regression analysis are:

(a) Determines the strength of the predictor
(b) Forecasting effects
(c) Trend Forecasting

Logistic Regression is derived from the concepts of Linear Regression and concepts of odds. Hence a brief overview of the same is required. Linear Regression is a statistical model that shows the relationship between two variables with a linear equation.

$$y=mx + c$$

Logistic Regression is basically a supervised classification algorithm which produces results in binary format which is used to predict the output of the categorical dependent variable for a given set of features or input. So the outcome should be discreet categorical. Based on categories, classification of Logistic Regression are:

(a) Binomial: Only two values are possible for the target variable "0" or "1" representing various values "True" vs. "False", "dead" vs. "alive".
(b) Multinomial: Three or more possible values for the target variable which are not ordered.
(c) Ordinal: The output target variables with ordered categories.

Equations for various logistic regression models:

Binomial:

$$P = e^{mx+c}/1+e^{mx+c}$$

Multinomial:

For k class scenario

$$P(r) = exp(RHS\_R)/(1+exp(RHS\_A)+exp(RHS\_B)...exp(k-1))$$

Ordinal:

For n possible outcomes

$$P(Y=N) = (1/1+e^{-(a_n +b_1x_1 b_2x_2+b_3x_3)})$$
$$P(Y=N-1) = (1/1+e^{-(a_n +b_1x_1 b_2x_2+b_3x_3)}) - P(Y=N)$$
$$P(Y=N-2) = (1/1+e^{-(a_n +b_1x_1 b_2x_2+b_3x_3)}) - P(Y=N) - P(Y=N-1)$$

2.6 Support Vector Machine

Support Vector Machine (SVM) is a supervised machine learning algorithm which is used for classification, regression (time series prediction etc.), outlier detection, clustering. But it is usually used for classification. SVM is great for small data set while other algorithms (Random forest, Deep Neutral Networks, etc.) require more data set but almost always come up with a very robust model. A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyper plane. In two dimensional space, this hyper plane is a line dividing a plane into two parts where each part defines a different category of class. Dimensions of the hyper plane depend upon the input features. If input features are 3 then the hyper plane is 2-dimensional. It becomes very hard to imagine when the number of features exceeds 3.

To separate two classes of data points, there are many hyper planes. In the SVM algorithm, our aim is to find the hyper plane which has a maximum margin. Data points that are closer to the hyper plane are known as support vector and influence the position and orientation of the hyper plane.

3. METHODOLOGY

To compare the algorithms namely Naïve Bayes Nearest Neighbour, Decision tree, Logistic Regression, Random Forest, Support Vector Machine (SVM) implementation procedure, tools used and data set information is discussed here.

3.1 Tools used

Scikit-learn is probably the most useful library for machine learning in Python. It is on NumPy, SciPy and matplotlib, this library contains a lot of efficient tools for machine learning and statistical modelling including classification, regression, clustering, and dimensionality reduction [5].

3.2 Data Set

UCI machine learning repository is used for this work, this dataset is publicly available and was created by Dr William H. Wolberg, a physician at the University of Wisconsin Hospital at Madison, Wisconsin, USA. Fluid samples, taken from patients with solid breast masses and an easy-to-use computer program called Xcyt were used to create the dataset.

3.3 Attribute Information

1.ID number
2. Diagnosis (M = malignant, B = benign) .
3. Ten real-valued features are computed for each cell nucleus:

(a) Radius (mean of distances from the centre to points on the perimeter)
(b) Texture (standard deviation of grey-scale values)
(c) Perimeter
(d) Area
(e) Smoothness (local variation in radius lengths)
(f) Compactness (perimeter² / area — 1.0)
(g) Concavity (severity of concave portions of the contour)
(h) Concave points (number of concave portions of the contour)
(i) Symmetry
(j) Fractal dimension ("coastline approximation" — 1)

4. RESULTS AND DISCUSSION

After implementing the algorithms in python, all the algorithms were given same input three times and then mean is calculated to get how accurate a particular algorithm is given below is the table of the calculated values:

Table 1: Accuracy values for the algorithms and their mean value

Algorithm name	Value1 (in %)	Value2 (in %)	Value3 (in %)	Mean (in %)
Logistic Regression	96.42	95	98.57	96.33
Decision Tree	94.28	96.42	91.42	94.04
Naive Bayes	97.1428	90.714	93.87	93.90
Random Forest	92.85	96.42	92.142	93.80
SVM	90.71	95.714	95.7	93.86
KNN	97.85	97.142	99.088	98.033

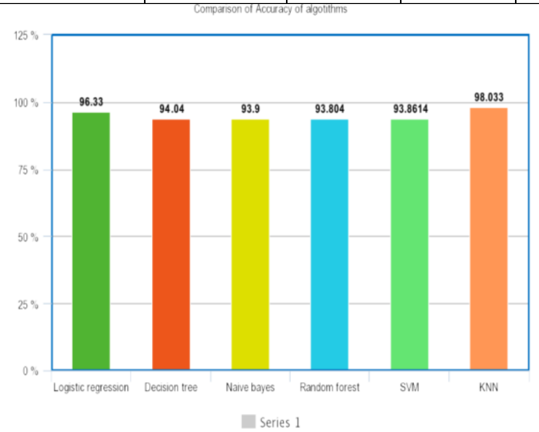

Fig. 5: Bar Graph for a mean accuracy percentage of algorithms

5. CONCLUSION

By doing all this work it is concluded that KNN (k-nearest neighbour) gives the best accuracy that is 98.033% in predicting breast cancer whereas Random forest gives the worst results with 93.80%.

6. REFERENCES

[1] Vishabh Goel,” Building a simple machine learning model on Breast Cancer Data”.

[2] Comparative Analysis of Classification Methods in R Environment with two Different Data Sets B Nithya, Dr V Ilango

[3] Abhishek Sharma, Geeks for geeks,” Decision Tree introduction with example”.

[4] Analytics Vidhya content team,” A Complete Tutorial on Tree Based Modeling from Scratch (in R & Python)”

[5] Kunal Jain, Analytics Vidhya content team,” Scikit-learn in Python – the most important Machine Learning tool I learnt last year!”.

[6] Gowgi, Coding Game,” Machine learning with Java - Part 6 (Random Forest)”.

[7] Dr William H. Wolberg, W. Nick Street, Olvi L. Mangasarian,” Breast Cancer Wisconsin (Diagnostic) Data Set”.

[8] Odajima, K., & Pawlovsky, A. P. (2014). A detailed description of the use of the kNN method for breast cancer diagnosis. 2014 7th International Conference on Biomedical Engineering and Informatics**.**

[9] Mandeep Rana, “Breast cancer diagnosis and recurrence prediction using machine learning techniques”, International Journal of research in Engineering and Technology, Vol.4, No.4, pp.372-376, April 2015.