International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 08 | Aug 2020 www.irjet.net p-ISSN: 2395-0072

Breast Cancer Prediction using Supervised Machine Learning Algorithms

T.Gowri[1], Dr.S.Geetha[2]

[1]MCA, Department of Computer Science, University College of Engineering, Trichy, TamilNadu, India
[2]Asst.Professor Department of Computer Science, University College of Engineering, Trichy, TamilNadu, India

---

Abstract - *Breast Cancer is leading cause of death among women's. According to Cancer Report Breast cancer is seems constantly increasing all over worldwide in past years and It's a Most dreadful disease for women's. Even medical field has enormous amount of data, certain tools and techniques are needed to handle those data. Classification Techniques is one of main techniques often used. This system Predict arising possibilities of Breast Cancer using Classification Technique. This system provide the chances of occurring Breast cancer in terms of percentage. The real time dataset is used in this system in order to obtain exact prediction. The datasets are processed in Python Programming Language using three main Machine Learning Algorithms namely Naïve Bayes Algorithm, Decision Tree Algorithm and Support Vector Machine (SVM) Algorithm. The aim of the system to shows which algorithms are best to use in order perform prediction tasks in medical Filed. Algorithm results are calculated in terms of accuracy rate and efficiency and effectiveness of each algorithm.*

Key Words: **Machine Learning Algorithm, Classification Techniques, Python Programming, etc.**

1. INTRODUCTION

Machine learning is the study of algorithms and statistical models that the computer systems use to effectively perform a specific task without using any explicit instructions. Machine learning is one of the small part of intelligence, and refers to a specific sub-part of Artificial Intelligence is related to constructing algorithms that can make to accurate predictions about future results. Machine learning algorithms build a mathematical model of some data, known as "training set" in order to make predictions without being explicitly programmed to perform the task. Classification rules are typically useful for medical problems that have been applied mainly in the area of medical field.

Machine Learning is the field of study that gives computers the capability to learn without being explicitly programmed. ML is one of the most exciting technologies that one would have ever come across. As it is evident from the name, it gives the computer that which makes it more similar to humans: The ability to learn. Machine learning is actively being used today, perhaps in many more places than one would expect.

Data mining is extracting information and knowledge from huge amount of data. Data mining is an essential step in discovering knowledge from databases. There are numbers of databases, data marts, data warehouses all over the world. Data Mining is mainly used to extract the hidden information from a large amount of database. Data mining is also called as Knowledge Discovery Database (KDD).

The data mining has four main techniques namely Classification, Clustering, Regression, and Association rule. Data mining techniques have the ability to rapidly mine vast amount of data. Data mining is mainly needed in many fields to extract useful information from a large amount of data. The fields like the medical field, business field, and educational field have a vast amount of data, thus these fields data can be mined through those techniques more useful information. Data mining techniques can be implemented through a machine learning algorithm. Each technique can be extended using certain machine learning models.

2. LITERATURE SURVEY

Predicting the early detection of chronic Cancer disease also known as chronic renal disease for Cancer patients with the help of machine learning methods and finally suggests a decision tree to arrive at concrete results with desirable accuracy by measuring its performance to its specification and sensitiveness. In order to increase the accuracy of the prediction result, we have utilized algorithms such as neural network and clustering data which greatly helped in our mission and also gave scope for future work. [2]

This paper predict Breast Cancer and described symptoms and reason of causing Breast Cancer. This system is a result of comparative analysis. This system use Classification techniques for predicting process. The ML algorithms used are Support Vector Machine, Decision tree, Naïve Bayes k- nearest neighbors and conclude with the one algorithm in terms of accuracy level of results .[3]

This Paper predicts Breast Cancer for using Classification Techniques. The detailed information about Cancer diseases such as its Facts, Common Types, and Risk Factors has been explained in this paper. The Data Mining tool used is WEKA (Waikato Environment for Knowledge Analysis), a good Data Mining Tool for Bioinformatics Fields. The all three available Interface in WEKA is used here. Naive Bayes, Artificial Neural Networks and Decision Tree (J48) are Main Data Mining Techniques and through this techniques Breast Cancer is predicted in this System. [4]

3. BREAST CANCER

Breast cancer is the disease in which cells in the breast grow out of control .there are different kind of breast cancer. The kind of breast cancer depends on which cells in the breast turns into Cancer. Breast cancer can begin with different parts of the breast. A breast is made up of three main parts: lobules, Ducts, and Connective Tissue (which consists of fibrous and fatty tissue) surrounds and holds everything together. Most breast cancer begin in the ducts or lobules. Breast cancer can spread outside the breast through blood vessels and lymph vessels.

Breast Cancer could be an estrogen related cancer. Breast Cancer is the prime reason for demise of women. It is the second dangerous cancer after lung cancer. In the year 2019 according to the statistics provided by World Cancer Research Fund it is estimated that over 2 million new cases were recorded out of which 626,679 deaths were approximated. Of all the cancers, breast cancer constitutes of 11.6% in new cancer cases and come up with 24.2% of cancers among women.

Kinds of Breast Cancer

- Invasive Ductal Carcinoma
- Invasive Lobular carcinoma

Symptoms of Breast Cancer

- New Lump in the breast or Underarm(armpit)
- Irritation or dimpling of breast skin
- Redness or flaky skin in the nipple area.
- Change in Size or the shape of the breast
- Nipple Discharge.
- Thickening of the breast.
- Pain in any area of breast.

Risk Factor of Breast Cancer

- Not being physically Active
- Obese after Menopause
- Low Vitamin B levels
- Being over weight
- Having dense Breast
- Genetic mutations
- Light exposure at night
- Diethylstilbestrol Exposure

4. METHODOLOGY

4.1 Data Source

The dataset used here for predicting breast cancer is taken from UCI Machine Learning repository. UCI is collection of database that are used for implementing machine learning algorithms. The dataset used here is real dataset. The dataset consists of 300 instances of data with the appropriate 9 clinical parameters. The clinical parameters of dataset is to test which are taken to the breast cancer as like Age, Menopause, Tumor Size, Node-caps, Irradiation and etc.

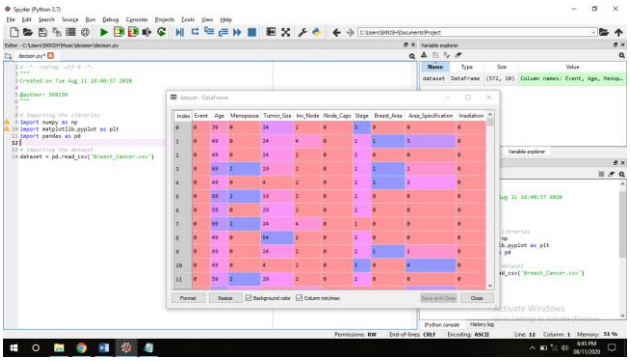

Fig -1: Visualizing Dataset in Python Environment

4.2 ALGORITHM DESCRIPTION

This section describes about three algorithms used in this system namely Naïve Bayes Classifier Algorithm, Decision tree Classification Algorithm and Support Vector Machine Algorithm (SVM).

4.2.1. Naïve Bayes Classifier Algorithm:

Naïve Bayes classifier is a supervised algorithm which classifies the dataset on the basis of Bayes theorem. The Bayes theorem is a rule or the mathematical concept that is used to get the probability is called Bayes theorem. Bayes theorem requires some independent assumption and it requires independent variables which is the fundamental assumption of Bayes theorem.

Naïve Bayes is a simple and powerful algorithm for predictive modeling. This model is the most effective and efficient classification algorithm which can handle massive, complicated, non-linear, dependent data. Naïve comprises two part namely naïve & Bayes where naïve classifier assumes that the presence of the particular feature in a class is unrelated to the presence of any other feature.

Bayes theorem on Mathematical Representation:

$$P(A\backslash B) = (P(B\backslash A) * P(A)) / (P(B))$$

Here,

P (A) => independent probability of A (prior probability)

P (B) => independent probability of B

P(B\A) => conditional probability of B given A (likelihood)

P(A\B) => conditional probability of A given B(posterior probability).

4.2.2. Decision tree Classification Algorithm:

The decision tree is a supervised machine learning algorithm. It handles both the categorical data and numerical

data. Based on certain conditions it gives a categorical solution such Yes/No, True or false, 1 or 0. For handling medical dataset the Decision tree Classification algorithm is widely used. The result of this model differing from the other models like the knn model, SVM model. The output consists of horizontal and vertical line splits based on the condition depends on the dependent variables. The accuracy level of this algorithm is quite higher than the other algorithms. The reason for the higher accuracy of this algorithm is this model analyses the dataset in the tree shape format. Thus each and every attribute of the dataset is been analyzed. Thus the accuracy rate of this model is higher. This model analyzes the data in the tree-shaped structure. Tree shaped diagram determines the course of actions. The decision tree model analyze the data on the basis of three nodes namely

• Root node - this main node, on basis of this node all other perform it function

• Interior node - the condition of dependent variables is handled by this node

• Leaf node - the final result is carried on a leaf node.

Formula for finding root node (Information Gain)

Information Gain = Class Entropy - Entropy Attributes

To find Class Entropy:

$$(Pi+Ni) = -\frac{P}{P+N}\log_2\frac{P}{P+N} - \frac{N}{P+N}\log_2\frac{N}{P+N}$$

Here => P, Possibilities of Yes.

=> N, Possibilities of No.

To find Entropy Attributes:

$$\text{Entropy attribute} = \sum\frac{Pi+Ni}{P+N}$$

4.2.2. Support Vector Machine Algorithm:

Support Vector Machine is usually represented as SVM. It is an elegant and Powerful Algorithm. The objective of the support vector machine algorithm is to find a hyperplane in an N-dimensional space (N — the number of features) that distinctly classifies the data points. To separate the two classes of data points, there are many possible hyperplanes that could be chosen. Our objective is to find a plane that has the maximum margin, i.e. the maximum distance between data points of both classes. Maximizing the margin distance provides some reinforcement so that future data points can be classified with more confidence.

Hyperplanes and Support Vector:

Hyperplanes are decision boundaries that help classify the data points. Data points falling on either side of the hyperplane can be attributed to different classes. Also, the dimension of the hyperplane depends upon the number of features. If the number of input features is 2, then the hyperplane is just a line. If the number of input features is 3, then the hyperplane becomes a two-dimensional plane. It becomes difficult to imagine when the number of features exceeds

Cost Function and Gradient Updates:

$$\omega = \omega + \alpha.\,(x_i + x_i - 2\lambda\,\omega)$$

$$\omega = \omega - \alpha.\,(2\,\lambda\,\omega)$$

In the SVM algorithm, we are looking to maximize the margin between the data points and the hyperplane. The loss function that helps maximize the margin is hinge loss.

5. RESULTS AND DISCUSSION

The aim of this project is to know whether the patient has Breast Cancer or not. The records in the datasets are divided into training set and test sets. After preprocessing the dataset, three data mining classification technique namely Naïve Bayes, Decision Tree and SVM Algorithm were applied. This section shows the results of those classification model done using Python Programming. The results are generated for both training datasets and test data sets.

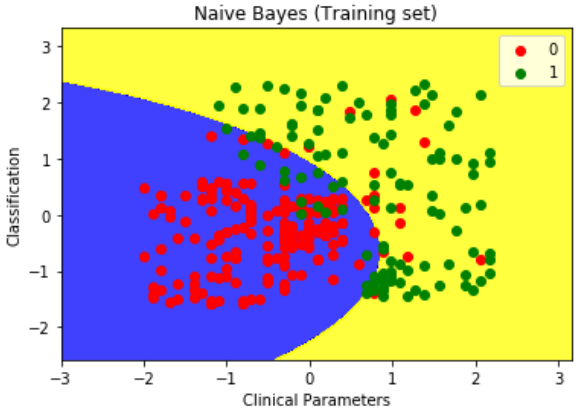

Fig-2: The Output shows Naïve Bayes Train set which classifies the 70% instances of dataset and displays Possibilities of having Breast cancer where Red denotes patients who not having heart disease (NO) Green denotes patients who had Breast cancer (YES)

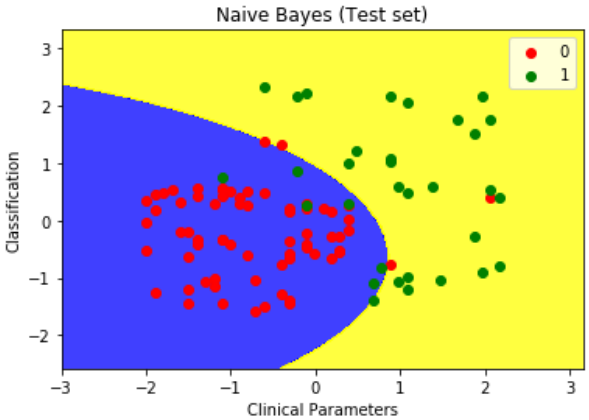

Fig-3: The Output shows Naïve Bayes Test set which classifies the 30% instances of dataset and displays

Possibilities of having Breast Cancer where Red denotes patients who not having Breast Cancer (NO), and Green denotes patients who had Breast Cancer(YES)

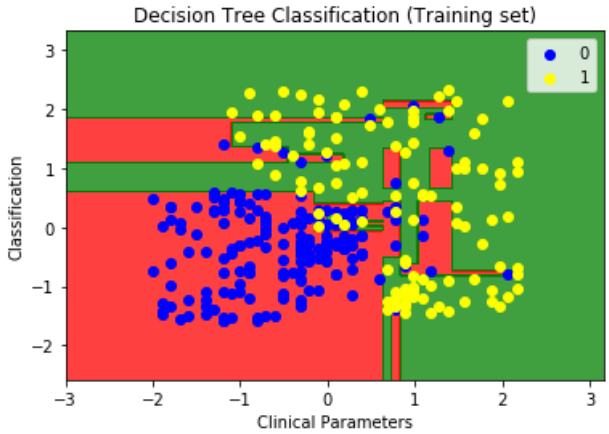

Fig-4: The Output shows Decision tree Train set which classifies the 70% instances of dataset and displays Possibilities of having Breast Cancer where Blue denotes patients who not having Breast Cancer (NO), and Yellow denotes patients who had Breast Cancer(YES)

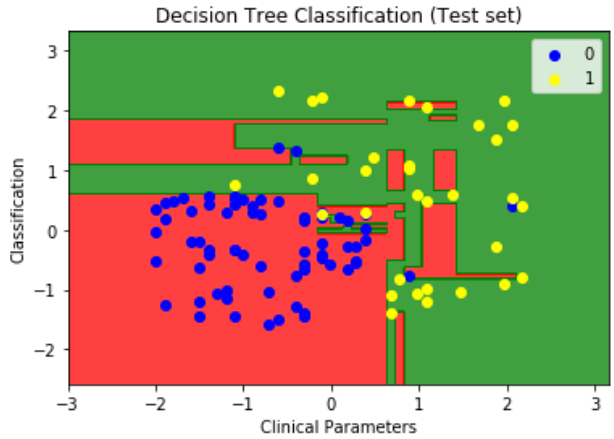

Fig-5: The Output shows Decision tree Test set which classifies the 30% instances of dataset and displays Possibilities of having Breast Cancer where Blue denotes patients who not having Breast Cancer (NO), and Yellow denotes patients who had Breast Cancer (YES)

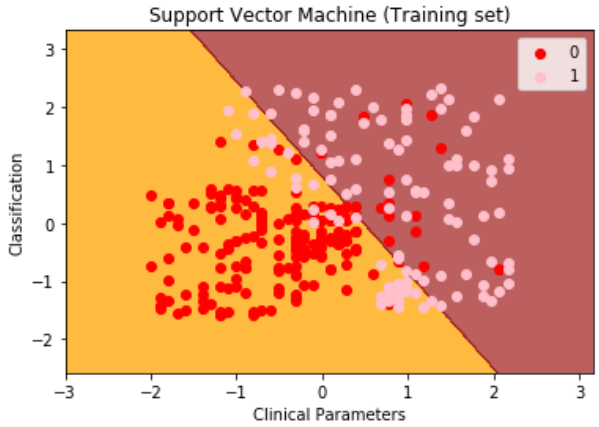

Fig-6 The Output shows SVM Train set which classifies the 70% instances of dataset and displays Possibilities of having Breast Cancer where Red denotes patients who not having Breast Cancer (NO), and Pink denotes patients who had Breast Cancer(YES)

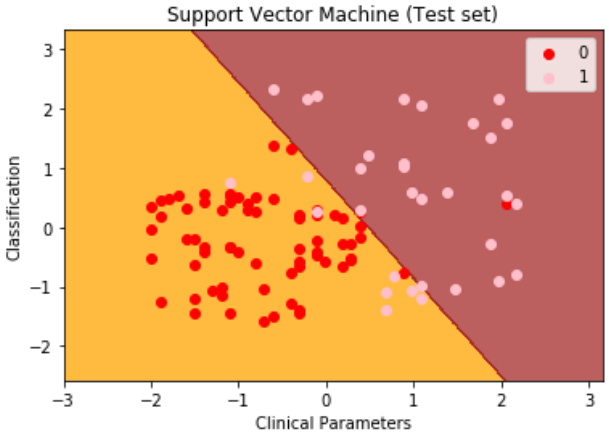

Fig-7: The Output shows SVM Test set which classifies the 30% instances of dataset and displays Possibilities of having Breast Cancer where Red denotes patients who not having Breast Cancer (NO), and Pink denotes patients who had Breast Cancer (YES).

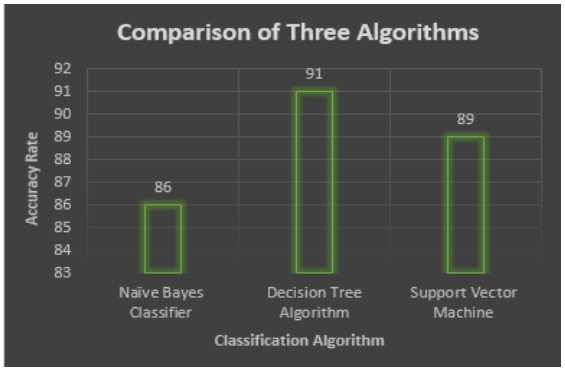

Fig-8: Comparison of three Algorithms used for Classification and Predictions

6. CONCLUSIONS

Medical dataset can not only be classified with the previously mentioned algorithms from machine learning, there are many algorithms and techniques which may perform better than these. Production of accurate classifier which perform efficiently for medicinal application is the main challenge we face in machine learning. Four main algorithms were implemented in this System were Naïve Bayes Algorithm, Decision Tree Algorithm and SVM Algorithm. Our main aim for the research is to discover the algorithm which performs faster, accurate and efficiently. Decision Tree Algorithm surpasses all the other algorithms with an accuracy of 91%.

Thus I Conclude, this project by saying Decision tree Classification algorithm is best and better for handling medical data set. In the future, the designed system with the used machine learning classification algorithm can be used to predict or diagnose other diseases. The work can be extended or improved for the automation of Breast cancer analysis including some other machine learning algorithms.

7. REFERENCES

[1] Shetty, Deeraj, Kishor Rit, Sohail Shaikh, and Nikita Patil. "Diabetes disease prediction using data mining." In 2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), pp. 1-5. IEEE, 2017

[2]Hiba Asria, Hajar Mousannif, Hassan Al Moatassime, Thomas Noel, "Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis", Procedia Computer Science 83, (2016), 1064-1069.

[3] Abien Fred M. Agarap, "On Breast Cancer Detection : An Application of Machine Learning Algorithms on the Wisconsin Diagnostic Dataset", arXiv:1711.07831v1 [cs.LG] 20 Nov 2017.

[4] K. R. Anantha Padmanaban and G. Parthiban Applying Machine Learning Techniques for Predicting the Risk of Chronic Cancer Disease August 2016 Vol 9(29),

[5] Suman Bala1, Krishan Kumar "A Literature Review on Kidney Disease Prediction using Data Mining Classification Technique " IJCSMC, Vol. 3, Issue. 7, July 2014, pg.960 – 967

[6] U.S. Cancer Statistics Working Group. United States Cancer Statistics: 1999–2008 Incidence and Mortality Web-based Report. Atlanta (GA): Department of Health and Human Services, Centers for Disease Control and Prevention, and National Cancer Institute; 2012.

[7] Siegel RL, Miller KD, Jemal A. Cancer Statistics , 2016. 2016;00(00):1-24. doi:10.3322/caac.21332

[8]"UCI Machine Learning Repository: Breast Cancer Wisconsin (Original) Data Set." [Online]. Available:https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original%29

[9] https://www.cancer.gov/

.