Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 05, May - 2017

Predicting Breast Cancer using Novel Approach in Data Analytics

Ms. L. Sankari
PG – Scholar, Department of CSE
Manakula Vinayagar Insitute of Technology
Puducherry, India

Mr. R. Rajbharath,
“Research – Scholar, Rayalaseema University”
Assistant Professor, Department of CSE
Manakula Vinayagar Insitute of Technology
Puducherry, India

Dr. G. Tholkappia Arasu
Principal, AVS College of Technology
Salem, India.

***Abstract*— Cancer has been characterized as a heterogeneous disease consisting of many different subtypes. The early diagnosis and prognosis of a cancer type have become a necessity in cancer research, as it can facilitate the subsequent clinical management of patients. The importance of classifying cancer patients into high or low risk groups has led many research teams, from the biomedical and the bioinformatics field, to study the application of machine learning (ML) methods. In this work it has been proposed to use a hybrid of Random Forest and Logistic Regression algorithms for building a breast cancer survivability prediction model. The Random Forest Technique is used to perform a preliminary screening of variables and to receive important ranks. Then, the new data set is extracted from initial WDBC dataset according to top-k important predictors and is input into the Logistic Regression procedure, which is responsible for building interpretable models for predicting breast cancer survivability.**

Index Terms*— RF Random Forest, LR Logistic Regression, WDBC Wisconsin Breast Cancer Data, ROC Receiver Operating Characteristic, AUC Area Under Curve

1. INTRODUCTION:

Cancer has been characterized as a heterogeneous disease consisting of many different subtypes. The early diagnosis and prognosis of a cancer type have become a necessity in cancer research, as it can facilitate the subsequent clinical management of patients. The importance of classifying cancer patients into high or low risk groups has led many research teams, from the biomedical and the bioinformatics field, to study the application of machine learning (ML) methods. Therefore, these techniques have been utilized as an aim to model the progression and treatment of cancerous conditions. In addition, the ability of ML tools to detect key features from complex datasets reveals their importance. A variety of these techniques, including Support Vector Machines (SVMs) and Logistic Regression and Random Forest have been widely applied in cancer research for the development of predictive models, resulting in effective and accurate decision making.In this project we are going to predict the breast cancer for a particular region. We are storing the inputs that are taken from organization which consists of survey of breast cancer result in the particular region.

1.1.1 Breast Cancer

- Breast cancer can occur in women and rarely in men.
- Breast cancer usually starts off in the inner lining of milk ducts or the lobules that supply them with milk.
- A malignant tumor can spread to other parts of the body.
- **Breast cancer is the most common invasive cancer in females worldwide**. It accounts for 16% of all female cancers and 22.9% of invasive cancers in women. 18.2% of all cancer deaths worldwide, including both males and females, are from breast cancer.
- In this proposed system we are trying to predict whether the sample observation is malignant or not.
- The first sign of breast cancer often is a breast lump or an abnormal mammogram. Breast cancer stages range from early, curable breast cancer to metastatic breast cancer, with a variety of breast cancer treatments.

Male breast cancer is not uncommon and must be taken seriously.

1.1.2 Symptoms

Breast cancers in their early stages are usually painless. Often the first symptom is the discovery of a hard lump. Fifty percent of such masses are found in the upper outer quarter of the breast. The lump may make the affected breast appear elevated or asymmetric. The nipple may be retracted or scaly. Sometimes the skin of the breast is dimpled like the skin of an orange. In some cases there is a bloody or clear discharge from the nipple.Many cancers, however, produce no symptoms and cannot be felt on examination. With an increase in the use of mammogram screening programs during the last several decades, more breast cancers are being discovered before there are any symptoms.

- A lump in a breast
- A pain in the armpits or breast that does not seem to be related to the woman's menstrual period
- Pitting or redness of the skin of the breast; like the skin of an orange
- A rash around (or on) one of the nipples
- A swelling (lump) in one of the armpits
- An area of thickened tissue in a breast
- One of the nipples has a discharge; sometimes it may contain blood
- The nipple changes in appearance; it may become sunken or inverted
- The size or the shape of the breast changes
- The nipple-skin or breast-skin may have started to peel, scale or flake.

1.1.3 Causes of Breast Cancer

Over the course of a lifetime, 1 in 8 women will be diagnosed with breast cancer.

Risk factors you cannot change include:

- Age and gender -- Your risk of developing breast cancer increases as you get older. Most advanced breast cancer cases are found in women over age 50. Men can also get breast cancer. But they are 100 times less likely than women to get breast cancer.

  Family history of breast cancer -- You may also have a higher risk of breast cancer if you have a close relative who has had breast, uterine, ovarian, or colon cancer. About 20 - 30% of women with breast cancer have a family history of the disease.
- Genes -- Some people have genetic mutations that make them more likely to develop breast cancer. The most common gene defects are found in the BRCA1 and BRCA2 genes. These genes normally produce proteins that protect you from cancer. If a parent passes you a defective gene, you have an increased risk of breast cancer. Women with one of these defects have up to an 80% chance of getting breast cancer sometime during their life.
- Menstrual cycle -- Women who got their periods early (before age 12) or went through menopause late (after age 55) have an increased risk of breast cancer.

  Other risk factors include:
- Alcohol use -- Drinking more than 1 - 2 glasses of alcohol a day may increase your risk of breast cancer.
- Childbirth -- Women who have never had children or who had them only after age 30 have an increased risk of breast cancer. Being pregnant more than once or becoming pregnant at an early age reduces your risk of breast cancer.
- DES -- Women who took diethylstilbestrol (DES) to prevent miscarriage may have an increased risk of breast cancer after age 40. This drug was given to the women in the 1940s - 1960s.
- Hormone replacement therapy (HRT) -- You have a higher risk of breast cancer if you have received hormone replacement therapy with estrogen for several years or more.
- Obesity -- Obesity has been linked to breast cancer, although this link is not completely understood. The theory is that obese women produce more estrogen. This can fuel the development of breast cancer.
- Radiation -- If you received radiation therapy as a child or young adult to treat cancer of the chest area, you have a very high risk of developing breast cancer. The younger you started such radiation and the higher the dose, the higher your risk. This is especially true if the radiation was given during breast development.

1.2 DATABASE USED

MySQL is an open source relational database management system (RDBMS) based on Structured Query Language (SQL).MySQL runs on virtually all platforms, including Linux, UNIX, and Windows. Although it can be used in a wide range of applications, MySQL is most often associated with web-based applications and online publishing and is an important component of an open source enterprise stack called LAMP.MySQL is a key component of many big data platforms, with Hadoop vendors estimating that 80% of deployments are integrated with MySQL.

1.3 CONNECTIVITY TOOL USED

The Jupyter Notebook:

The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and explanatory text. Uses include: data cleaning and transformation, numerical simulation, statistical modeling, machine learning and much more.

2.1 EXISTING SYSTEM

In the existing system all the data will be visualized manually which means that there must be the storage of data in files and the data can be viewed only manually. It is difficult to visualized the data properly. And the existing system takes more time in predicting whether the patient is affected by breast cancer or not. The exact stage of the breast cancer cannot be predicted in earlier stage. There is high risk in prediction.

2.1.1 Drawback of Existing System

- The data are visualized by manual work.
- The time taken for predicting tumor is high.
- The correct stage of Breast Cancer is not predicted accurately.
- The risk of prediction is high and the data are not visualized properly

2.2 PROPOSED SYSTEM

- In the proposed system the data are visualized properly by exact attributes.
- The prediction stages are predicted accurately and proper treatment can be taken.
- The time taken for visualization is reduced and grouped by the four factors
- The cancer has been predicted by the logistic regression.

2.2.1 List of Symptoms of Breast Cancer

- Lump in the breast
- Skin dimpling
- Change in skin color or texture
- Change in how nipple looks, like pulling in of the nipple
- Clear or bloody fluid that leaks out of the nipple

2.2.2 Attributes used

There are some of the tissue characteristics which we are considered as attributes for predicting the tumor with the help of the values of each of them.

- Clump Thickness
- Uniformity of Cell Size

- Uniformity of Cell Shape
- Marginal Adhesion
- Single Epithelial Cell Size
- Bare Nuclei
- Bland Chromatin
- Normal Nucleoli
- Mitoses.

2.3 PROPOSED SYSTEM ARCHITECTURE

This describes the architecture of the system which takes the inputs and classifies it into training and testing data. The prediction is done for the both training and testing data.

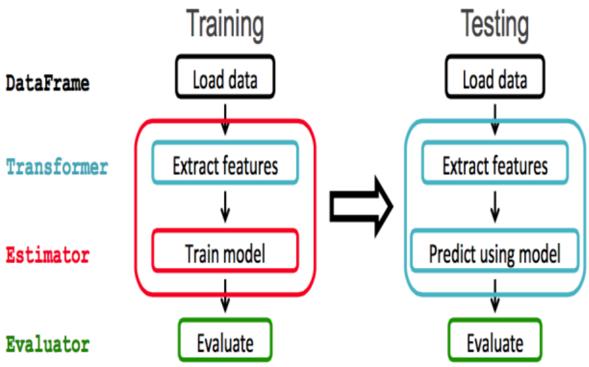

Fig 2.3 Architecture of breast cancer prediction

- DataFrame: The ML API uses DataFrames from Spark SQL as an ML dataset.
- Transformer: A Transformer is an algorithm which transforms one DataFrame into another DataFrame. For example, turning a DataFrame with features into a DataFrame with predictions.
- Estimator: An Estimator is an algorithm which can be fit on a DataFrame to produce a Transformer. For example, training/tuning on a DataFrame and producing a model.
- Evaluator: Metric to measure how well a fitted Model does on held-out test data.

Training:

The data are first loaded. Then it is send to the transformer and estimator circuit. In the transformer it extract the features and send it to the estimator. The estimator consist of the train model which performs on the train data using its model. Then the result is given out to the evaluates which evaluates and produce the final output.

Testing:

It is similar to the training phase. Here the data frame loads the data. And the data is passed to the transformer and the estimator circuit. In the transformer circuit it extract the features of the data that it taken. In the estimator it predicts the result for the input obtained. Then the result till now is given to the evaluator. The evaluator evaluates and produce the output.The training data is analyzed first then the test data is executed and thus the prediction is made.

3. ALGORITHM USED

There are three algorithms used for predicting the breast cancer.

- Logistic Regression
- Random Forest

3.1 Logistic Regression

Logistic regression is a popular method to predict a binary response. It is a special case of Generalized Linear models that predicts the probability of the outcome. Logistic regression measures the relationship between the Y "Label" and the X "Features" by estimating probabilities using a logistic function. The model predicts a probability which is used to predict the label class.

3.2 Random Forest

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of over fitting to their training set.

4. *The Cancer Observation Schema*

In this module we load the data from the csv file into an RDD of Strings. Then we use the map transformation on the rdd, which will apply the ParseRDD function to transform each String element in the RDD into an Array of Double. Then we use another map transformation, which will apply the ParseObs function to transform each Array of Double in the RDD into an Array of Cancer Observation objects. The toDF() method transforms the RDD of Array Cancer Observation into a Dataframe with the Cancer Observation class schema.In this module the cancer is observed by using different type of schemas which will give the result that the cancer is observed in the correct stage of patient.

6. Implementing Logistic Regression for Prediction

- Logistic regression is a popular method to predict a binary response. It is a special case of Generalized Linear models that predicts the probability of the outcome. For more background and more details about the implementation, refer to the documentation of the logistic regression in spark.mllib.
- The current implementation of logistic regression in spark.ml only supports binary classes. Support for multiclass regression will be added in the future.
- When fitting LogisticRegressionModel without intercept on dataset with constant nonzero column, Spark MLlib outputs zero coefficients for constant nonzero columns. This behavior is the same as R glmnet but different from LIBSVM.
- A common metric used for logistic regression is area under the ROC curve (AUC). We can use the BinaryClassificationEvaluator to obtain the AUC.

- A Precision-Recall curve plots (precision, recall) points for different threshold values, while a receiver operating characteristic, or ROC, curve plots (recall, false positive rate) points. The closer the area Under ROC is to 1, the better the model is making predictions.

7. Output - Logistic regression

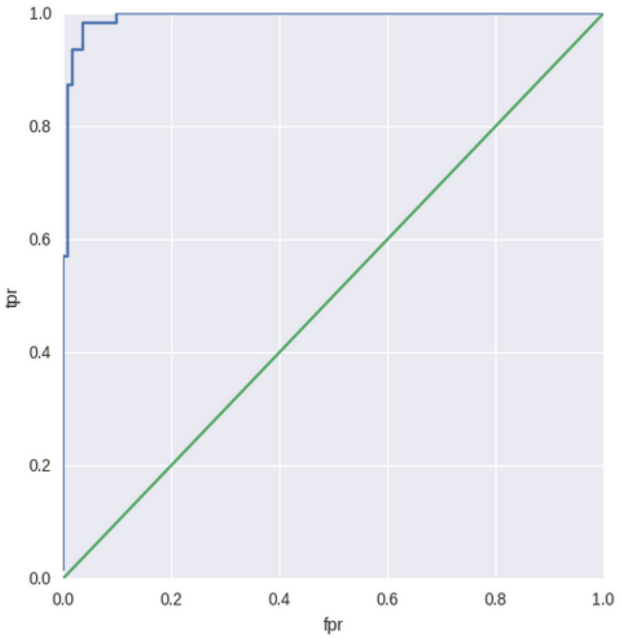

Fig. 7.1. ROC- Logistic Regression

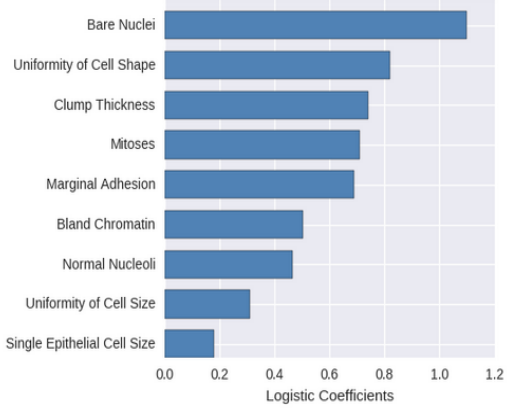

Fig. 7.2. Attribute classifier- Logistic Regression

8. Output - Random Forest

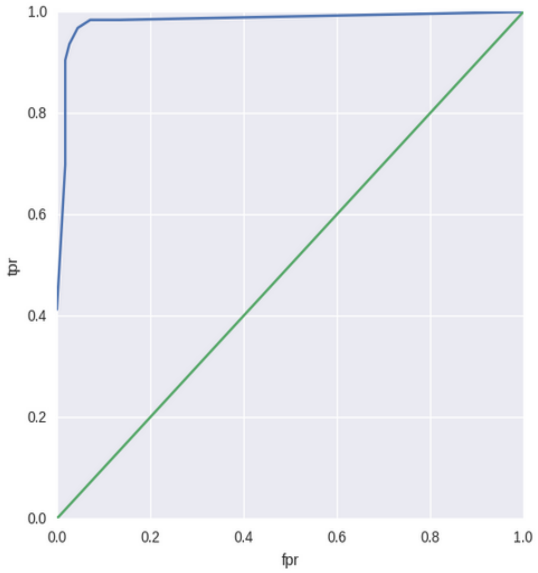

Fig. 8.1. ROC- Random Forest

9. RESULTS

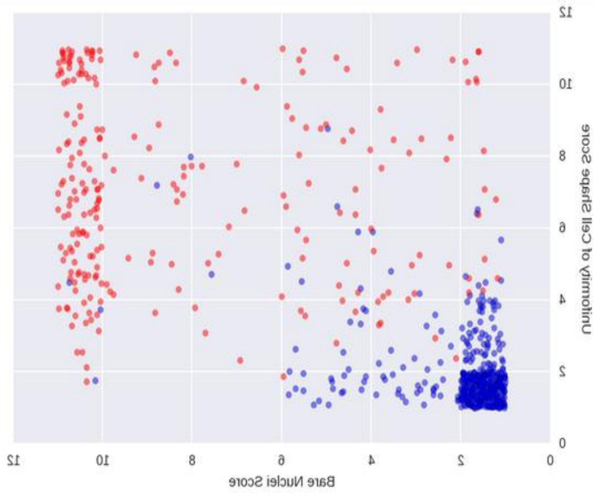

Fig.9.1: Bare Nuclei and Uniformity of Cell Shape scores

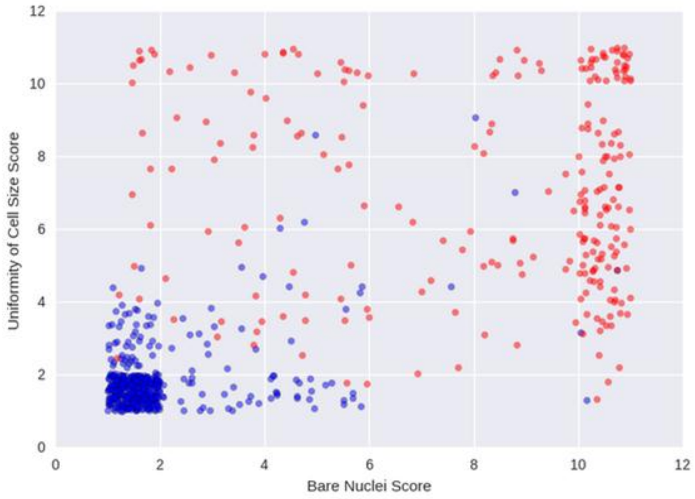

Fig.9.2: Bare Nuclei and Uniformity of Cell Size scores:

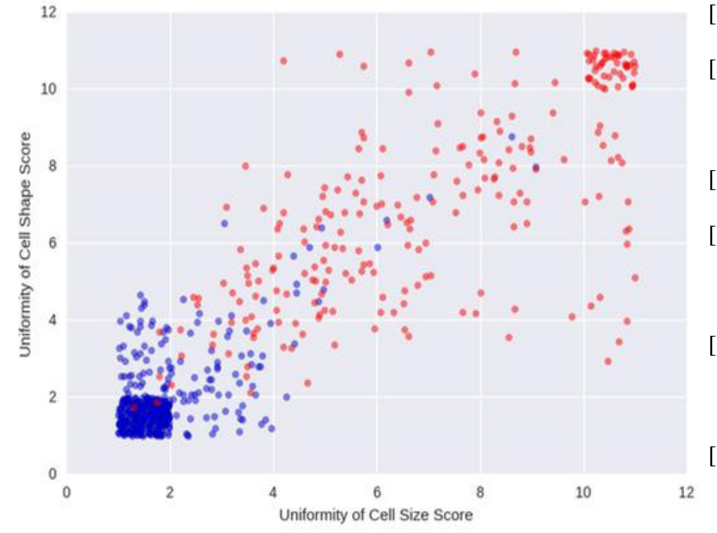

Fig.9.3 Uniformity of cell size and Uniformity of Cell shape scores

10. CONCLUSION:

Implementation of this ”BREAST CANCER PREDICTION” with biological test results of the individual provides us to depict the appropriate stage of disease for each and every individual which in turn helps to predict the average of people affected in the same stage in a particular area in a graphical representation. Based on the above prediction awareness can be created along the environmental changes which causes such problems. Here based on the figure it can be concluded that in that particular area the persons with benign and malignant stages. The graph is plotted for different attributes which will give the result for the lab observation. And certain measures can be taken to avoid the occurrences of the breast cancer.

REFERENCES

[1] A. SoltaniSarvestani, A. A. Safavi, N.M. Parandeh, M.Salehi., “Predicting Breast Cancer Survivability Using Data Mining Techniques,” IEEE Transl., vol. 2, PP. 227-231, 2010

[2] D. Delen, G. Walker and A. Kadam, “Predicting breast cancer survivability: a comparison of three data mining methods,” J.Artificial Intelligence in Medicine, vol. 34, pp. 113-127, 2005.

[3] L. Breiman, J. Friedman, R Olshen, C. Stone, “Classification and Regression Trees”, Wadsworth, 1984.

[4] Wang Yi and Wang Fuyong, “Breast Cancer Diagnosis via Support Vector Manchines,” Proceedings of the 25th Chinese Control Conference, Haebin, Heilongjiang, pp. 1853-1856. August 2006

[5] L. Breiman, “Random Forests,” J. Machine Learning, vol. 45, pp. 5 -32, 2001.

[6] S. Wei, B. T. Greer, F. Westermann*et al*., “Prediction of clinical outcome using gene expression profiling and artificial neutral networks for patients with neuroblastoma,” *Cancer research*, vol. 64, no. 19, pp. 6883-6891, 2004.

[7] G. C. Hon, R. D. Hawkins, O. L. Caballero *et al*., “Global DNA hypomethylation coupled to repressive chromatin domain formation and gene silencing in breast cancer,” *Genome research*, vol. 22, no. 2, pp. 246-258, 2012.

[8] B. P. Berman, D. J. Weisenberger, J. F. Aman*et al*., “Regions of focal DNA hypermethylation and long-range hypomethylation in colorectal cancer coincide with nuclear lamina-associated domains,” *Nature genetics*, vol. 44, no. 1, pp. 40-46, 2012.

[9] L. Tolosi, and T. Lengauer, “Classification with correlated features: unreliability of feature ranking and solutions,” *Bioinformatics*, vol. 27, no. 14, pp. 1986-1994, 2011.

[10] H. Binder, and M. Schumacher, “Incorporating pathway information into boosting estimation of high-dimensional risk prediction models,” BMC *bioinformatics*, vol. 10, no. 1, pp. 18, 2009.

[11] X. Wan, C. Yang, Q. Yang *et al*., “BOOST: A fast approach to detecting gene-gene interactions in genome-wide case-control studies,” *The American Journal of Human Genetics*, vol. 87, no. 3, pp. 325-340, 2010.

[12] C. Winter, G. Kristiansen, S. Kersting*et al*., “Google goes cancer: improving outcome prediction for cancer patients by network-based ranking of market genes,” PLoS*computational biology*, vol. 8, no. 5, pp. e1002511, 2012.

[13] C. Porzelius, M. Johannes, H. Binder *et al*., “Leveraging external knowledge on molecular interactions in classification methods for risk prediction of patients,” *Biometrical Journal*, vol. 53, no. 2, pp. 190-201, 2011.

[14] S. Zhang, Q. Li, J. Liu *et al*., “A novel computational framework for simultaneous integration of multiple types of genomic data to identify micro RNA-gene regulatory modules,” *Bioinformatics*, vol. 27, no. 13, pp. i401-, 2011.

[15] M. Sang, Y. Lian, X. Zhou *et al*., “MAGE-A family: attractive targets for cancer immunotheraphy,” Vaccine, vol. 29, no. 47, pp. 8496-8500,