BREAST CANCER ANALYSIS USING MACHINE LEARNING ALGORITHMS

[1]Sanjana Govindaswamy, [2]K. Ramani Reddy, [3]Lingala Sai Saketh Reddy

[1]Student, [2]Student, [3]Student
[1]Computer Science and Engineering,
[1]CMR Technical Campus, Hyderabad, India

Abstract

Breast cancer represents one of the diseases that make a high number of deaths every year. It is the most common type of all cancers and the main cause of women's deaths worldwide. Classification and data mining methods are an effective way to classify data. Especially in medical field, where those methods are widely used in diagnosis and analysis to make decisions. In this paper, a performance comparison between different machine learning algorithms: Logistic regression, K-Nearest Neighbour, Support Vector Machine (SVM), Kernel Support Vector Machine, Decision Tree, Random forest, Naive Bayes is conducted. The main objective is to assess the correctness in classifying data with respect to efficiency and effectiveness of each algorithm in terms of accuracy, precision, sensitivity and specificity. Experimental results show that Random Forest gives the highest accuracy (98.60%) with lowest error rate. All experiments are executed within a simulation environment and conducted in Anaconda3.

I. INTRODUCTION

Breast cancer is a heterogeneous tumour that has various subtypes with different biological behaviours and clinicopathological and molecular characteristics. In the last 20 years, there has been an increase in the understanding of multistep carcinogenesis and the leading role of genetic change in the diagnosis, treatment and prevention of breast cancer. This leads to an increase in prevention, detection and treatment strategies in breast cancer patients.

The cause of breast cancer is multifactorial. Several risk factors for breast cancer have been known nowadays. The risk factors are classified into non modifiable risk factors: age, sex, genetic factors (5-7%), family history of breast cancer, history of previous breast cancer and proliferative breast disease; modifiable risk factors: menstrual and reproductive factors, radiation exposure, hormone replacement therapy, alcohol and high fat diet; and environmental factors: organochlorine exposure, electromagnetic field and smoking.

Increasing comprehensive knowledge and awareness of breast cancer risk could facilitate its early detection. It can be more effectively treated in earlier stage than when clinical signs and symptoms present, justifying early detection efforts. Based on those studies, it is necessary to do the calculation of risk factors through an algorithm that can assist to determine whether a person has risk factors for breast cancer so it can help the early detection of breast cancer. An algorithm system using real scoring can support people to perform routine checks for early detection of breast cancer and help healthcare workers to find people at risk of developing breast cancer.

Through measurement of breast cancer risk, it can be seen whether a person has a safe risk to breast cancer, adequate for breast cancer prevention or harmful to the occurrence of breast cancer. If someone in the high risk category then the action should be hastened to do screening to ascertain whether someone is likely to have breast cancer or not, whereas if someone in the prevention behaviour is adequate then it is advisable to keep the behaviour to avoid breast cancer, and on the other hand, when entering in the safe category then someone will be recommended to maintain the behaviour and avoid risk factors for breast cancer to avoid breast cancer.

The calculation of breast cancer risk factors can be determined by using the algorithm or early detection model of breast cancer risk through determinant factors is used to detect breast cancer risk itself and is a preventive action, using machine learning by classifying the risk of breast cancer of the variable predictors, making it is easier to classify. Classification using machine learning is a type of artificial intelligence (AI) that provides computers with the ability to learn without being explicitly programmed. Machine learning has been used in cancer detection and diagnosis for a score. Machine learning methods have been used to identify, classify, detect, or distinguish tumours and other malignancies. In other words, machine learning has been used primarily as an aid to cancer diagnosis and detection.

Machine learning algorithms are effective because their process of searching for a model function can explain and differentiate the class and concept data, which the model is determined based on the data training analysis that is class object data whose label class is already known. The types of learning algorithm are Naive Bayes, Decision Tree, Logistic Regression, Super Vector Machine, Kernel Support Vector Machine and K-Nearest Neighbour

II. DATA COLLECTION TECHNIQUE

The data collection done through online search by entering keywords as follows: ((breast cancer risk OR breast cancer risk calculation OR breast cancer prediction) AND (machine learning OR algorithms OR Naive Bayes OR Decision Tree OR Logistic Regression OR Linear Discriminant Analysis OR Super Vector Machine OR K-Nearest Neighbour)).

The search was limited to English language articles. The article type was limited to journal articles. The research subject was limited to research with human subject. The time of publication was limited from January 2000 to May 2018. The abstract of articles with potentially relevant titles were reviewed, while the irrelevant articles were excluded. Furthermore, articles that have potentially relevant abstracts will be reviewed in full-text, while the irrelevant articles were excluded. The inclusion criteria of this study sample were research on machine learning algorithms for breast cancer risk calculations using Naive Bayes, Decision Tree, Random Forest, Logistic Regression, Super Vector Machine, Kernel Support Vector Machine and K-Nearest Neighbour with prognostic model study. The dataset was created by the University of Wisconsin which has 569 instances (rows—samples) and 32 attributes (features—columns).

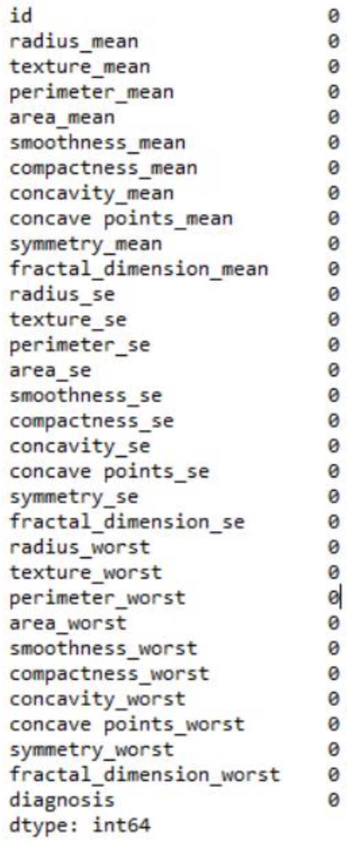
III. Research Methodology

This analysis aims to observe which features are most helpful in predicting malignant or benign cancer and to see general trends that may aid us in model selection and hyper parameter selection. The goal is to classify whether the breast cancer is benign or malignant. To achieve this, I have used machine learning classification methods to fit a function that can predict the discrete class of new input.

The dataset used in this story is publicly available and was created by Dr. William H. Wolberg, physician at the University Of Wisconsin Hospital at Madison, Wisconsin, USA. To create the dataset Dr. Wolberg used fluid samples, taken from patients with solid breast masses and an easy-to-use graphical computer program called Xcyt, which is capable of perform the analysis of cytological features based on a digital scan. The program uses a curve-fitting algorithm, to compute ten features from each one of the cells in the sample, than it calculates the mean value, extreme value and standard error of each feature for the image, returning a 30 real-valuated vector

3.1 Importing the dataset

We will be using Jupyter notebook to work on this dataset. We will first go with importing the necessary libraries and import our dataset Visualization of data is an imperative aspect of data science. It helps to understand data and also to explain the data to another person. Python has several interesting visualization libraries such as Matplotlib, Seaborn etc. In this we will use pandas' visualization which is built on top of matplotlib, to import the dataset. The attributes used in the dataset are as follows,



```
id                         0
radius_mean                0
texture_mean               0
perimeter_mean             0
area_mean                  0
smoothness_mean            0
compactness_mean           0
concavity_mean             0
concave points_mean        0
symmetry_mean              0
fractal_dimension_mean     0
radius_se                  0
texture_se                 0
perimeter_se               0
area_se                    0
smoothness_se              0
compactness_se             0
concavity_se               0
concave points_se          0
symmetry_se                0
fractal_dimension_se       0
radius_worst               0
texture_worst              0
perimeter_worst            0
area_worst                 0
smoothness_worst           0
compactness_worst          0
concavity_worst            0
concave points_worst       0
symmetry_worst             0
fractal_dimension_worst    0
diagnosis                  0
dtype: int64
```

Figure 1: Attributes used in the dataset

3.2 Categorical Data

Categorical data are variables that contain label values rather than numeric values. The number of possible values is often limited to a fixed set. For example, users are typically described by country, gender, age group etc. We will use Label Encoder to label the categorical data. Label Encoder is the part of Scikit Learn library in Python and used to convert categorical data, or text data, into numbers, which our predictive models can better understand

3.3 Splitting the dataset

The data we use is usually split into training data and test data. The training set contains a known output and the model learns on this data in order to be generalized to other data later on. We have the test dataset (or subset) in order to test our model's prediction on this subset. We will do this using Scikit-Learn library in Python using the train_test_split method

3.4 Feature Scaling

Most of the times, your dataset will contain features highly varying in magnitudes, units and range. But since, most of the machine learning algorithms use Euclidian distance between two data points in their computations. We need to bring all features to the same level of magnitudes. This can be achieved by scaling. This means that you're transforming your data so that it fits within a specific scale, like 0–100 or 0–1.

3.5 Model Selection

This is the most exciting phase in Applying Machine Learning to any Dataset. It is also known as Algorithm selection for Predicting the best results. Usually Data Scientists use different kinds of Machine Learning algorithms to the large data sets. But, at high level all those different algorithms can be classified in two groups: supervised learning and unsupervised learning. Without wasting much time, I would just give a brief overview about these two types of learnings.

3.5.1 Supervised learning

Supervised learning is a type of system in which both input and desired output data are provided. Input and output data are labelled for classification to provide a learning basis for future data processing. Supervised learning problems can be further grouped into regression and Classification problems. A regression problem is when the output variable is a real or continuous value, such as "salary" or "weight". A classification problem is when the output variable is a category like filtering emails "spam" or "not spam"

3.5.2 Unsupervised Learning

Unsupervised learning is the algorithm using information that is neither classified nor labelled and allowing the algorithm to act on that information without guidance. In our dataset we have the outcome variable or Dependent variable i.e Y having only two set of values, either M (Malign) or B(Benign). So we will use Classification algorithm of supervised learning. We have different types of classification algorithms in Machine Learning. They are Logistic Regression, Nearest Neighbour, Support Vector Machines, Kernel SVM, Naïve Bayes, Decision Tree Algorithm, and Random Forest Classification.

IV. Results

To check the accuracy, we need to import confusion matrix method of metrics class. The confusion matrix is a way of tabulating the number of mis-classifications, i.e., the number of predicted classes which ended up in a wrong classification bin based on the true classes.

We will use Classification Accuracy method to find the accuracy of our models. Classification Accuracy is what we usually mean, when we use the term accuracy. It is the ratio of number of correct predictions to the total number of input samples.

$$Accuracy = \frac{Number\ of\ Correct\ predictions}{Total\ number\ of\ predictions\ made}$$

To check the correct prediction, we have to check confusion matrix object and add the predicted results diagonally which will be number of correct prediction and then divide by total number of predictions. After applying the different classification models, we have got below accuracies with different models:

- Logistic Regression—95.8%
- Nearest Neighbour—95.1%
- Support Vector Machines—97.2%
- Kernel SVM—96.5%
- Naive Bayes—91.6%
- Decision Tree Algorithm—95.8%
- Random Forest Classification—98.6%

So finally, we have built our classification model and we can see that Random Forest Classification algorithm gives the best results for our dataset.

V. Related work

Classification is one of the most important and essential tasks in machine learning and data mining. About a lot of research has been conducted to apply data mining and machine learning on different medical datasets to classify Breast Cancer. Many of them show good classification accuracy. Vikas Chaurasia and Saurabh Pal11 compare the performance criterion of supervised learning classifiers; such as Naïve Bayes, SVM-RBF kernel, RBF neural networks, Decision trees (J48) and simple CART; to find the best classifier in breast cancer datasets. The experimental result shows that SVM-RBF kernel is more accurate than other classifiers; it scores accuracy of 96.84% in Wisconsin Breast Cancer (original) datasets. Djebbari et al.12consider the effect of ensemble of machine learning techniques to predict the survival time in breast cancer. Their technique shows better accuracy on their breast cancer data set comparing to previous results. S. Aruna and L. V Nandakishore13, compare the performance of C4.5, Naïve Bayes, Support Vector Machine (SVM) and K- Nearest Neighbour (K-NN) to find the best classifier in WBC. SVM proves to be the most accurate classifier with accuracy of 96.99%. Angeline Christobel. Y and Dr. Sivaprakasam14, achieve accuracy of 69.23% using decision tree classifier (CART) in breast cancer datasets. The accuracy of data mining algorithms SVM, IBK, BF Tree is compared by A. Pradesh15. The performance of SMO shows a higher value compared with other classifiers. T.Joachims16 reaches accuracy of 95.06% with neuron fuzzy techniques when using Wisconsin Breast Cancer (original) datasets. In this study, a hybrid method is proposed to enhance the classification accuracy of Wisconsin Breast Cancer (original) datasets (95.96) with 10fold cross validation. Liu Ya-Qin's, W. Cheng, and Z. Lu17 experimented on breast cancer data using C5 algorithm with bagging; by generating additional data for training from the original set using combinations with repetitions to produce multisets of the same size as you're the original data; to predict breast cancer survivability. Delen et al. Lu18 take 202,932 breast cancer patients records, which then pre-classified into two groups of "survived" (93,273) and "not survived" (109,659). The results of predicting the survivability were in the range of 93% accuracy. With respect to all related work mentioned above, our work compares the behaviour of data mining algorithms SVM, NB, k-NN and C4.5 using Wisconsin Breast Cancer (original) datasets in both diagnosis and analysis to make decisions. The goal is to achieve the best accuracy with the lowest error rate in analysing data. To do so, we compare efficiency and effectiveness of those approaches in terms of many criteria, including: accuracy, precision, sensitivity and specificity, correctly and incorrectly classified instances and time to build model, among others. Our experimental results show that SVM achieves the highest accuracy (97.13%) with the lowest error rate (0.02%) unlike C4.5, Naïve Bayes and k-NN that have an accuracy that varies between 95.12 % and 95.28 % and an error rate that varies between 0.03 and 0.06

VI. Conclusion

To analyse medical data, various data mining and machine learning methods are available. An important challenge in data mining and machine learning areas is to build accurate and computationally efficient classifiers for Medical applications. In this study, we employed main algorithms on the Wisconsin Breast Cancer (original) datasets. We tried to compare efficiency and effectiveness of those algorithms in terms of accuracy, precision, sensitivity and specificity to find the best classification accuracy. of Random Forest classification algorithm reaches and accuracy of 98.6% and outperforms, therefore, all other algorithms. In conclusion, Random Forest classification algorithm has proven its efficiency in Breast Cancer prediction and diagnosis and achieves the best performance in terms of precision and low error rate.