BREAST CANCER RISK PREDICTION USING IBM AUTO AI

RSIP Career Basic ML 076

Yazhini P

Yazhiniyazhu2001@gmail.com

**INTRODUCTION**

Being the most frequently occurring cancer in women, breast cancer affects around 10% of women at some point in their life. It is the second leading contributor to women’s death after lung cancer. 25% of all cancers in women including 12% of all new cases are caused by breast cancer [9]. Big Data has seen a rise in value due to it being used in derivation of business intelligence, business analytics and data mining to obtain reports and result predictions. Topics like medial science rise rapidly when certain approaches like data mining is applied due to better possibility of prediction of diseases, reducing medicine costs, improving health of patient by revamping the quality of healthcare along with value by saving people’s lives through real time decisions. The paper provides you with a analysis of performance and comparison of accuracy in classification between the algorithms such as: Logistic Regression, SVM, Random Forest and Naïve Bayes, being the major influential algorithms of data mining used in the research community. Logistic Regression is used to perform regression analysis when the dependent variable is binary. It is a predictive analysis similar to all other regression analyses. Naïve Bayes is a powerful classification algorithm from machine learning, it is not a algorithm but a group of algorithm in which all of them have same fundamental principle. In that group of algorithms, they all classify the data independently such that no algorithm provides same classification result or analysis. Naïve Bayes can be seen using in places such as document classification.

Document classification is nothing but it is used to classify document based on fields Random forest algorithm basically creates multiple trees and select the best among them by voting methodology. Random Forest algorithm creates multiple tree, each tree having different decision independent to each other. In common words no two tree is similar in Random Forest such that every tree provides implies classification methodology. This can be useful in certain instances like these where we have to classify medical dataset. Random forest is also used in places such as ETM, detection and prediction of an object and in games where it replicates the actions produced by humans in games. SVM is classification and regression model. It can be used to both circumstances. SVM creates a hyperplane which is basically called as threshold limit to classify data. This limit is created by the dataset while training. What if there is an deviation in data, SVM creates soft threshold limit which is close near to the main threshold limit and near to the deviation. The more dataset given to the algorithm the proper the classification will be of hyperplane. The kernel used in this project is linear. The reason to use linear kernel is it is faster and it is preferred when the data can be linearly separable. Places where you can see SVM used is text classification, face detection etc. Face detection is one of the main reasons SVM is preferred widely. Algorithm’s efficiency evaluation is the primary objective of this project.

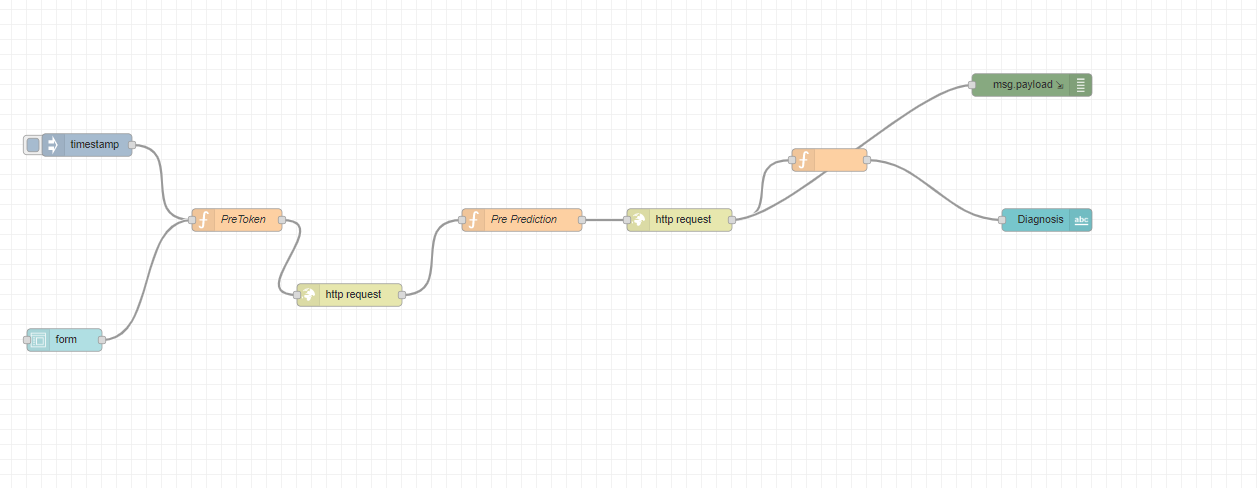
**Literature review**

In the literature, many works applied Artiﬁcial Intelligence (AI) technics for breast cancer diagnosis to improve classiﬁcation accuracy and time response. In this section, we give some works related to the solution of medical breast cancer diagnosis using machine and deep learning approaches. Arpit B. and Aruna T. proposed a genetically optimized neural network (GONN) for breast cancer classiﬁcation (malignant and benign). They optimized the neural network architecture by introducing new crossover and mutation operators. To evaluate their work, they used WBCD and compared the classiﬁcation accuracy, sensitivity, speciﬁcity, confusion matrix, ROC curves and AUC under ROC curves of GONN with classical model and classical Back propagation model. This method presents a good accuracy classiﬁcation. However, it can be improved by using a larger dataset than WBCD, feature extraction to make GONN more eﬃcient for real time diagnosis of Breast Cancer. Ashraf O. I. and Siti M. S. proposed a computer-based method to classify automatically breast cancer disease. The method applied multilayer perceptron (MLP) neural network based on enhanced non-dominated sorting genetic algorithm (NSGA-II) to optimize both the accuracy and network structure. Compared to other methods, this work improves classiﬁcation accuracy. However, MLP can get stuck in local minima. Na L. et al. proposed an intelligent classiﬁcation model for breast cancer diagnosis based on a hybrid feature selection approach: gain directed simulated annealing genetic algorithm wrapper (IGSAGAW), to remove redundant and irrelevant feature from the feature space and cost sensitive support vector machine (CSSVM) learning algorithm. This process can improve the classiﬁcation accuracy and reduce the computational cost. The proposed method is applied on Wisconsin Original Breast Cancer (WBC) and WBCD to verify its eﬀectiveness. The proposed work shows a good performance and decreases the calculation complexity. Nawel Z. et al. presented a conception and implementation of Computer Assisted Detection (CAD) for mammogram images classiﬁcation. The system is based on a GA-based features selection algorithm to reduce the dimensionality of the feature vector and semi supervised support vector machine (S3VM) for classiﬁcation. Experiments were validated on Digital Database for Screening Mammography (DDSM) dataset. The proposed approach improved accuracy. Abdulkader H. et al. developed an automated system for classiﬁcation of breast tissue. The system uses two machine learning techniques: Feed forward neural network using the back propagation learning algorithm (BPNN) and radial basis function network (RBFN). Breast cancer tissues were classiﬁed into 6 diﬀerent tissues, Carcinoma, Fibro-adenoma, Mastopathy, Glandular, Connective, and Adipose tissue. Data were acquired using an electrical impedance spectroscopy (EIS) method. The Radial basis function network outperformed the back propagation network for classifying six diﬀerent breast tissues in terms of accuracy, minimum error, maximum epochs and training time. The proposed system improved accuracy and decreases training time. However, learning with neural networks can be weak in generalizing and can get stuck in local optima. Haifeng W. et al. designed an SVM-based ensemble learning model for breast cancer diagnosis. The proposed ensemble model includes two types of SVM structures, i.e., a C-SVM and a -SVM, and six types of kernel functions. To import the expertise of diﬀerent base classiﬁers on diagnostic tasks, a Weighted Area Under the Receiver Operating Characteristic Curve Ensemble (WAUCE) mechanism is proposed for model hybridization. The model was evaluated based on two datasets: the Wisconsin Breast Cancer (WBC) dataset and the Wisconsin Diagnostic Breast Cancer (WDBC) dataset, and one large dataset, the Surveillance, Epidemiology, and End Results (SEER) dataset. The proposed model increases diagnosis accuracy compared to other works based on single SVM. However, it is a computationally expensive method and the training time is high. Kemal P. et al. proposed a hybrid approach based on mad normalization, KMC based feature weighting and AdaBoostM1 classiﬁer. The detection of the presence of breast cancer is done in three steps: In the ﬁrst step, the dataset was ﬁrst normalized by the MAD normalization method. In the second step, k-means clustering (KMC) based feature weighting has been used for weighting the normalized data. Finally, the AdaBoostM1 classiﬁer has been used to classify the weighted data set. The Breast Cancer Coimbra dataset (BCC) taken from UCI machine learning database was used. This method shows good results in terms of accuracy. However, it is a computationally expensive method. Teresa A. j. et al. proposed a classiﬁcation method of hematoxylin and eosin stained breast biopsy images using Convolutional Neural Networks (CNNs). They provide four classes of medical relevance: normal tissue, benign lesion, in situ carcinoma and invasive carcinoma. The proposed CNN architecture is designed to integrate information from multiple histological scales. The model is applied on image dataset composed of high resolution uncompressed, and annotated HE stain images from the Bioimaging 2015 breast histology classiﬁcation challenge. Fabio A. et al. used a deep learning approach to classify breast cancer histopathological images from BreaKHis, a public dataset. They proposed a method based on the extraction of image patches for training the Convolutional Neural Network (CNN) and the combination of these patches for ﬁnal classiﬁcation. This method allows avoiding adaptations of the model that can lead to a more complex and computationally costly architecture. However, the lack of data provides costly expensive experimentations. Hiba A. et al. presented a comparison between the performances of four classiﬁers: Support Vector Machine (SVM), Decision Tree (C4.5), Naive Bayes (NB) and k Nearest Neighbors (k-NN) on the Wisconsin Breast Cancer datasets

**THEORITICAL ANALYSIS**

**Diagrams:**

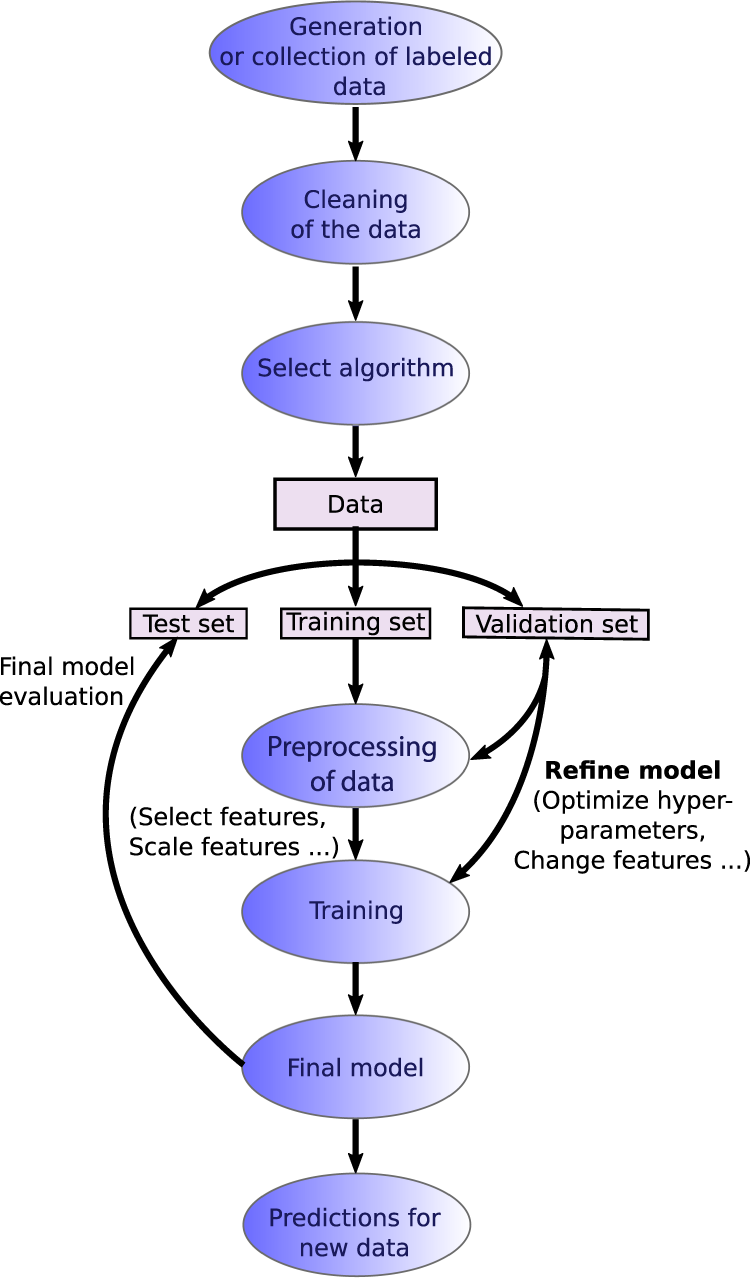
Hardware / Software Designing



**EXPERIMENTAL INVESTIGATIONS**

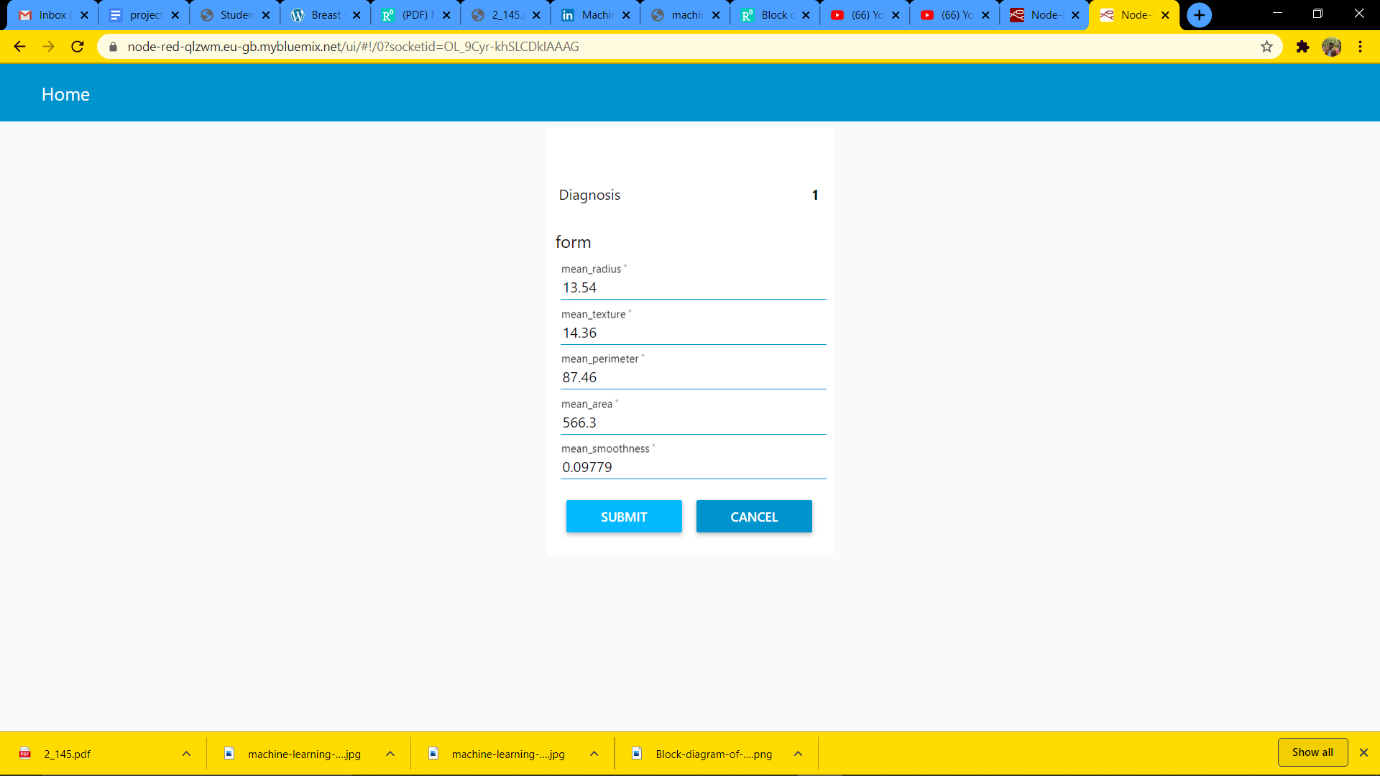
To compare the behaviour of LR, NB, SVM and Random Forest, the experiment conducted was focused on the evaluation of the algorithms. Questions raised from researchers were: Which algorithm is more effective? Which algorithm executes more efficiently? Which algorithm is more accurate in classifying? A. Experiment Environment The scikitlearn learn python libraries were used to conduct all experiments on classifiers explained in this paper. Sklearn is a collection of data mining, machine learning and deep learning algorithms used for classification, regression, data pre-processing and clustering. The sklearn libraries were used to implement machine learning algorithms for various real-world problems. Developers and practitioners can build and evaluate suitable models with this framework. The experiment conducted in which environment is conducted is ANACONDA. It contains various applications in that we have preferred Spyder, which is a development environment that supports python. It is a powerful IDE for python compared to others. It also has introspection features. Since our problem might require those features and also debugging is easier in this platform it is preferred. B. Breast cancer dataset The UCI machine learning repository consists of The Wisconsin Breast Cancer datasets [10] which is used in this study. There are 699 instances in which 458 are benign and 241 are malignant. In addition, there are two classes malignant which contributes to 65.5% of dataset and benign 34.5%. The breast cancer dataset is obtained in a csv format from their database. C. Data Visualization Data visualization is a key aspect of data science. It helps one to comprehend and also convey the data to another person in a meaningful manner. Matplotlib and Seaborn aresome of the several python data visualization libraries. It is essential in analysing large amounts of information and to make decisions. It employs the use of pictorial elements such as maps, plots, patterns, graph trends, etc. to provide the user with an easy method of comprehending the data.

**Flow chart**



**RESULT :**

The model has been created and tested using the machine learning technique and the output is predicted by deploying the model in a web application built by using Node Red App Service present in IBM Cloud.



**CONCLUSION**

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 study were NB, SVM, Random Forest and LR on Breast Cancer dataset. Our main aim for the research is to discover the algorithm which performs faster, accurate and efficiently. Random Forest surpasses all the other algorithms with an accuracy of 99.76%. In conclusion, the Random Forest algorithm achieves the lowest error rate along with highest precision which might be the best choice of algorithm for this problem and prediction of disease.

**ADVANTAGES & DISADVANTAGES**

**ADVANTAGES**

* It is easy to build
* It can easily predict the risk of the breast cancer diagnosis

**DISADVANTAGES**

* Choosing the inputs will be difficult because we can’t understand which inputs plays an major rule.
* The model can’t be prepared with 100 percent accuracy

**APPLICATIONS**

* This model can be deployed into an web application easily. Node Red Service can be used to built the web application’s easily.
* Any one can predict their chances of getting diagnosed by breast cancer and can take preventive measures