Project Report on " *Breast Cancer Risk Prediction using IBM Python notebook and Machine learning Service"*

Project Title: Breast Cancer Risk Prediction System

Project Description:

Breast cancer is one of the main causes of cancer death worldwide. Early diagnostics significantly increases the chances of correct treatment and survival, but this process is tedious and often leads to a disagreement between pathologists. Computer-aided diagnosis systems showed the potential for improving diagnostic accuracy. But early detection and prevention can significantly reduce the chances of death. It is important to detect breast cancer as early as possible.

This project builds a model in Watson Studio and deploys the model in IBM Watson Machine Learning.

IBM Watson Services:

1. IBM Services Used:
2. IBM Nodered
3. IBM Watson Studio
4. IBM Machine Learning
5. IBM Cloud Object Storage

Project Architecture:

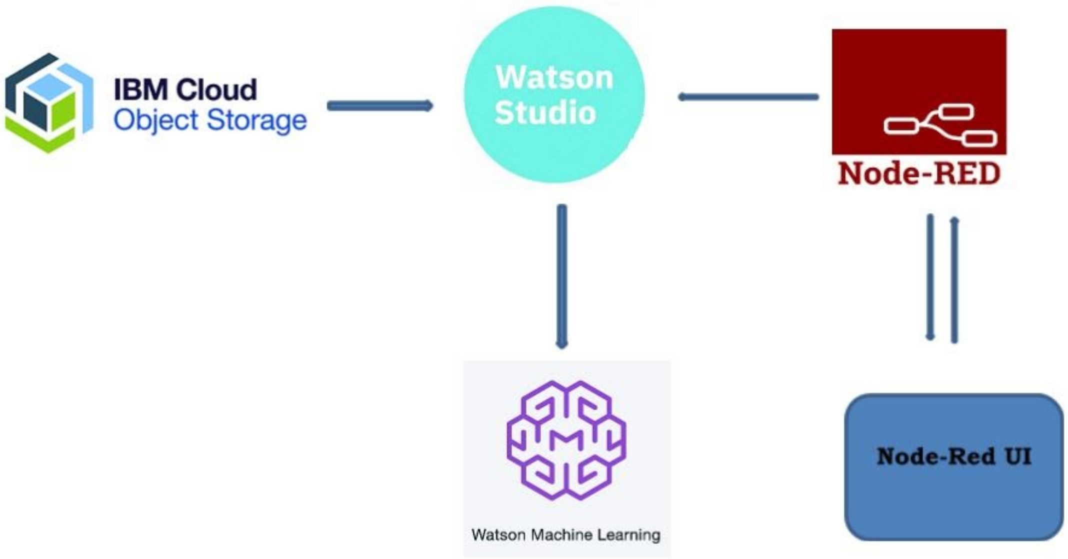


Figure 1 Project Architecture

Project Objectives:

1. To design a python notebook code for breast cancer prediction
2. To perform comparative analysis of various machine learning algorithms
3. Design a User Interface for prediction of Breast Cancer from various health

parameters

Dataset Description:

The Breast Cancer Prediction model is designed using Wisconsin cancer dataset. This dataset consists of 569 patient information. There are total 30 health attributes for each patient. Table 1 describes all the 30 attributes of the dataset.

Table 1 Wisconsin Cancer Dataset

Attribute	Numeric values (Min,Max)	Description
radius_mean	6.98, 28.1	Mean of distances centre point to points on perimeter
Texture_mean	9.71,39.3	Standard deviation of gray-scale pixel values
Perimeter_mean	43.8,189	Mean size of tumor
Smoothness_mean	0.05,0.16	Mean of local variations in radius
Compactness_mean	0.02, 0.35	Mean of compactness
Concavity_mean	0,0.43	Mean value of severity of concave proportions of contour
Concave points_mean	0,0.2	Mean of total number of concave proportions in contour
Symmetry_mean	0.11,0.3	
Factal_dimension_mean	0.05,0.1	Mean value of coastline approximation -1
Radius_se	0.11,2.87	Standard error for the mean of the distances from centre to points on the perimeter
Texture_se	0.36,4.88	Standard error of standard deviation of gray-scale pixel values

Perimeter_se	0.76,22	Standard error of size of tumor
Area_se	6.8,5.42	
Smoothness_se	0,0.03	Standard error for local variation in radius length
Compactness_se	0,0.14	Standard error for compactness
Concavity_se	0,0.4	Standard error for concavity proportions
Concave_points_se	0,0.05	Standard error of no of concave points in contour
Symmetry_se	0.01,0.08	Standard error of symmetry
Fractal_dimension_se	0,0.03	Standard error for coastline approximation-1
Radius_worst	7.93,36	Worst case mean of distances from centre to points on perimeter
Texture_worst	12,49.5	Worst case mean of standard deviation of gray scale pixel values
Perimeter_worst	50.4,251	Worst case mean of perimeter
Area_worst	85,4250	Worst case area mean
Smoothness_worst	0.07,0.22	Worst case value of local variation in radius lengths
Compactness_worst	0.03,1.06	Worst or largest mean of compactness
Concavity_worst	0.1, 25	Worst case mean of severity of concave proportions of countour
Concave points_worst	0,0.29	Worst or largest mean for number of concave proportions of countour
Symmetry_worst	0.16,0.66	Largest mean value of symmetry

Fractal_dimenstion worst	0.06,0.21	Worst or largest mean of coastline approximation-1

Design of Classification Model

The Breast Cancer Risk Prediction User Interface is designed with a purpose to help the doctor or physians in making a right prediction about breast cancer risk from the patient's health reports. Figure 2 shows the design methodology of the entire system. The design is divided in three sections.

1. Import dataset and train machine learning dataset using Python Notebook and Scikit Learn library
2. Deploy the trained model in IBM Cloud space
3. Interface model API with nodeRED flow and generate UI

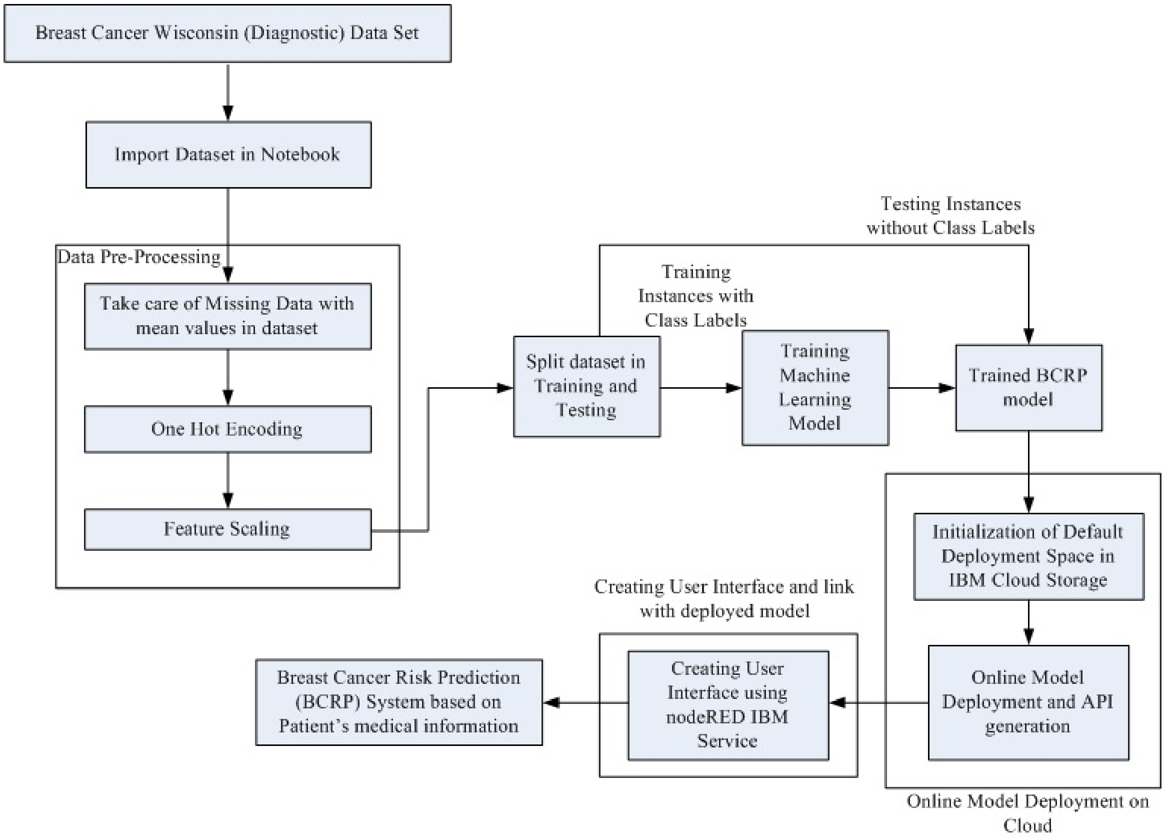

Breast Cancer Wisconsin (Diagnostic) Data Set
Import Dataset in Notebook
Data Pre-Processing
Take care of Missing Data with mean values in dataset
One Hot Encoding
Feature Scaling
Split dataset in Training and Testing
Training Instances with Class Labels
Testing Instances without Class Labels
Training Machine Learning Model
Trained BCRP model
Initialization of Default Deployment Space in IBM Cloud Storage
Online Model Deployment and API generation
Online Model Deployment on Cloud
Creating User Interface and link with deployed model
Creating User Interface using nodeRED IBM Service
Breast Cancer Risk Prediction (BCRP) System based on Patient's medical information

Figure 2 Project Design Methodology

Comparative Analysis:

In order to analyse the performance of various machine learning models various quantitative measures are used. These measures demonstrates the effectiveness of machine learning model for the prediction of diabetes mellitus based on patient's health parameters. The quality measuers used for the comparative analysis are as follows:

1. Accuracy : Total number of correct classifications out of total test instances.
2. Precision: Correctness over the positive detections.
3. Recall: Total positive detections out of total positive instances.
4. ROC Curve: The plot between True Positive Rate (TPR) and False Positive Rate (FPR) while changing the threshold for positive classification.

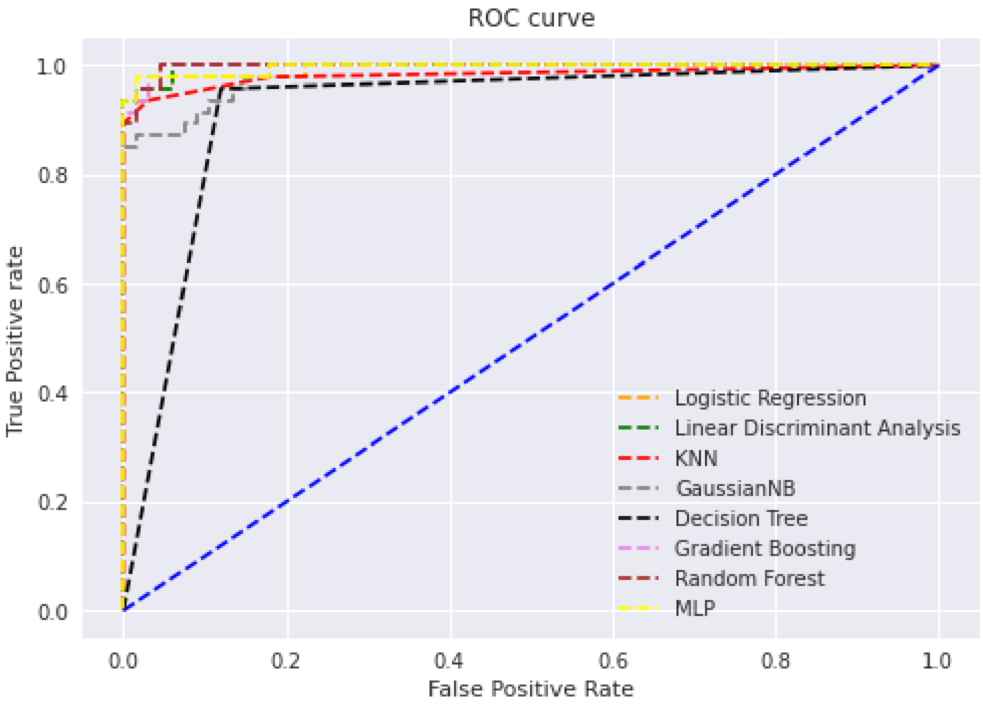


Figure 3 ROC for various Machine Learning Models

5. Area Under Curve (AUC) : It is obtained from the graph ploatted between Precision and Recall.
6. F_measure: It represents the weighted harmonic mean of the Precision and Recall.
7. Confusion Matrix: Table 2 shows the confusion matrix.
   TP: True Positive i.e. Positive sample are classified as Positive
   TN: True Negative i.e. Negative samples are classified as Negative
   FP: False Positive i.e. Negative samples are classified as Positive
   FN: False Negative i.e. Positive samples are classified as Negative

Table 2 Confusion Matrix

Confusion Matrix			
Total Number of Samples		Predicted Class	
		Benign	Malignant
True Class	Benign	True Negative (TN)	False Positive (FP)
	Malignant	False Negative (FN)	True Positive (TP)

Table 3 Comparative Performance of Various Machine Learning Algorithms

Performance parameter	Logistic Regression	Linear Discriminant Analysis	KNN	Gaussian Naïve Bayes	Decision Tree	Gradient Boosting	Random Forest	Multilayer Perceptron(MLP)
Accuracy	96.5	96.5	95.6	90.4	91.2	97.4	95.6	97.4
Precision	95.7	100	100	87.5	84.9	95.8	93.8	95.8
Recall	95.7	91.5	89.4	89.4	95.7	97.9	95.7	97.9
F_Measure	95.7	95.6	94.4	88.4	90.0	96.8	94.7	96.8
Area Under Curve (AUC)	99.333	99.714	4	98.412	91.902	99.745	99.714	99.555

Design of User Interface

Once the prediction model is designed using IBM Auto AI service, it is deployed using IBM Cloud storage and API is obtained. In order to make the UI another IBM service named nodeRED App is used. nodeRED contains various types of nodes which helps to design the flow and user interface. The flow of proposed model designed in nodeRED is shown in Figure 4.

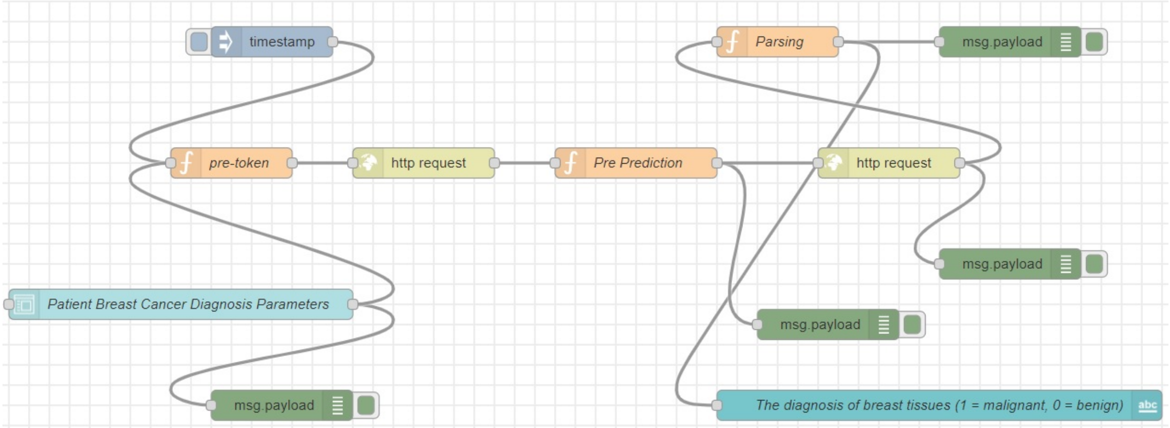
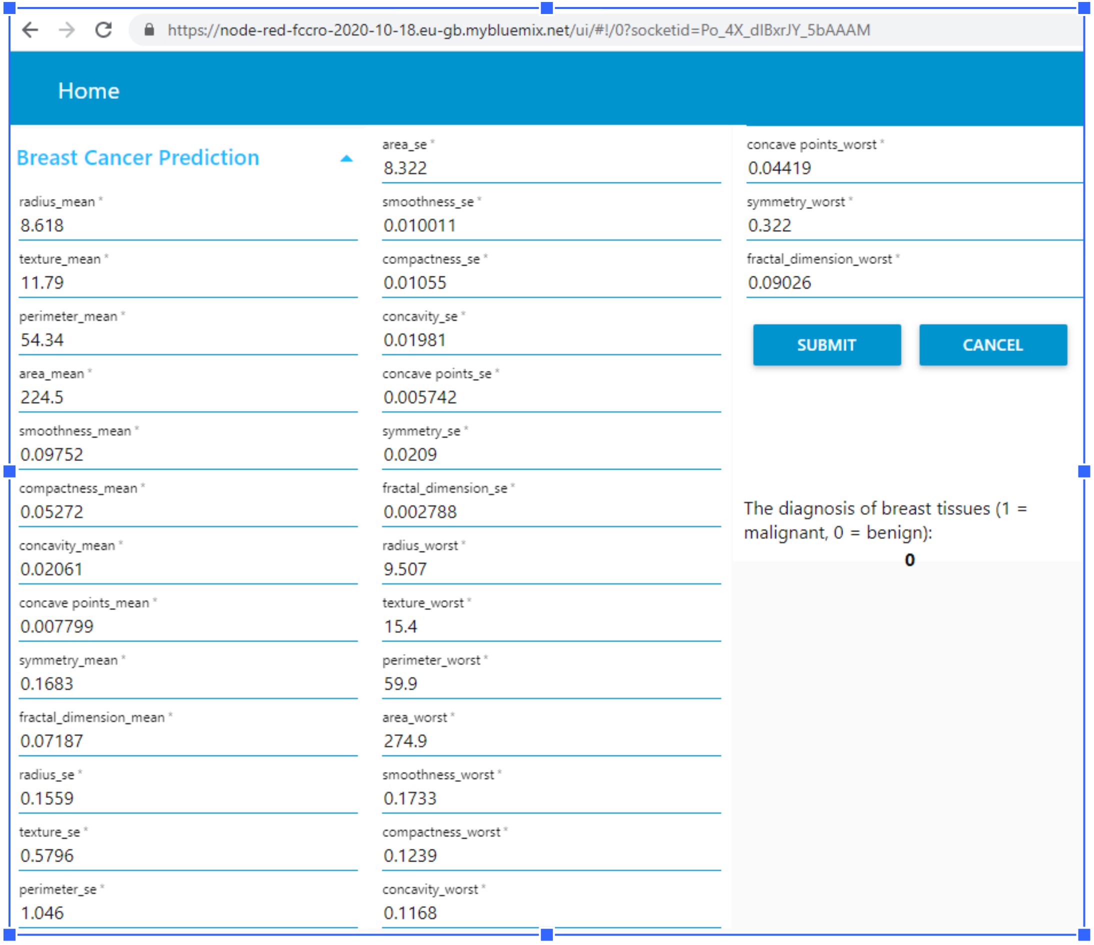


Figure 4 nodeRED flow for Breast Cancer Risk Prediction Model UI

User Interface Output:

Figure 5 shows the output of User Interface for Breast Cancer Risk Prediction system.


https://node-red-fccro-2020-10-18.eu-gb.mybluemix.net/ui/#!/0?socketid=Po_4X_dlBxrJY_5bAAAM
Home
SUBMIT
CANCEL

Breast Cancer Prediction

radius_mean *
8.618

texture_mean *
11.79

perimeter_mean *
54.34

area_mean *
224.5

smoothness_mean *
0.09752

compactness_mean *
0.05272

concavity_mean *
0.02061

concave points_mean *
0.007799

symmetry_mean *
0.1683

fractal_dimension_mean *
0.07187

radius_se *
0.1559

texture_se *
0.5796

perimeter_se *
1.046

area_se *
8.322

smoothness_se *
0.010011

compactness_se *
0.01055

concavity_se *
0.01981

concave points_se *
0.005742

symmetry_se *
0.0209

fractal_dimension_se *
0.002788

radius_worst *
9.507

texture_worst *
15.4

perimeter_worst *
59.9

area_worst *
274.9

smoothness_worst *
0.1733

compactness_worst *
0.1239

concavity_worst *
0.1168

concave points_worst *
0.04419

symmetry_worst *
0.322

fractal_dimension_worst *
0.09026

The diagnosis of breast tissues (1 = malignant, 0 = benign):
0

Figure 5 User Interface for Breast Cancer Risk Prediction

Conclusion:

After analyzing the quantitative results gradient boosting algorithm based machine learning classification model is deployed using IBM Cloud services. Gradient boosting algorithm provides the accuracy of 97.4%, recall of 97.9% and F_measure of 96.8% along with AUC of 99.74%. The model is capable of correctly predicting the malignant and benign classes from the patient's health features.

Acknowledgement:
I express my since thanks to smartinternz for providing insights of using IBM cloud services during the FDP. The experts have supported immensly while designig of project on IBM Cloud platform.

Author Details:
Name: Dr. Vipul H. Mistry
Designation: Assistant Professor and Head of Electronics and Communication Engineering Department
Institute Name: S. N. Patel Institute of Technology and Research Centre, Umrakh affiliated with Gujarat Technological University, Chandkheda, Gujarat
IBM Cloud Userid: vipul.vpl@gmail.com