

**INTERNSHIP REPORT APPROVAL FORM**

1 July, 2019

With immense pleasure, this is to approve that the students of Sona College of Technology .i.,e

**Deepika Gurung(17ECEBE103)**

**Nagarjun S(17ECEBE001)**

**Varundeepak B.A(17ITBE009)**

successfully completed their Project and Project Report on **“Breast cancer detection using machine learning”** under our guidance.

We are highly impressed with the work that they have done and commend them on their quick grasping skills. They have shown good intent to learn and have put the knowledge gained into application in the from of this project. We appreciate the hard work and commitment shown by them.

We, hereby approve that this document is completely checked and accepted by SmartBridge Technical Team. Its been an absolute pleasure to educate and mentor these students. We hope that this document will also serve as a Letter of Recommendation, to whomsover applied.

We wish them success in all future endeavors and a great career ahead.

**GD Abhishek**

AI Developer

Breast cancer detection

using machine learning algorithm

**1.Introduction**

**Artificial Intelligence**

Artificial intelligence is when we train any machine to behave like a human being . We define AI as the study of agents that receive percepts from the environment and perform actions.

**Python**

Python is an object oriented , interpreted language which was developed by Guido van Rossum in 1991.

It is an open source and platform independent software , which allows you to run on a wide variety of systems.

Python makes the development and debugging fast. Since there is no compilation process, edit-test-debug cycle is incredibly fast. Python supports GUI applications; libraries for major operating systems.

**Machine learning**

ML is an application of artificial intelligence that is based on statistics to provide systems ability to automatically learn and improve from experience without being explicitly programmed.

**Watson Machine Learning**

It provides a full range of tools and services so you can build, train and deploy Machine Learning models. Choose from tools that fully automate the training process for rapid prototyping to tools that give you complete control to create a model that matches your needs.

**Node Red**

Node red is a programming tool for wiring together hardware devices ,API’S and online services in a new and interesting ways.

**SQL**

SQL stands for Structured Query Language. It is designed for managing data in a relational database management system (RDBMS). It is pronounced as S-Q-L or sometime See-Qwell. SQL is a database language , it is used for database creation, deletion, fetching rows, and modifying rows, etc.

**1.2. Objective of the Research**

* 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.
* The main objective of this research is 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 parameters selection. Selection of suitable model is important ,especially in medical fields where those methods are widely used in diagnosis and analysis to make the decisions.
* The goal is to classify whether the breast cancer is benign or malignant. To achieve this we have used machine learning algorithms to train our model to classify the discrete class of new inputs. Comparing the various algorithm ,depending on their accuracy ,correlation, area under curve(auc).
* And to make it user friendly we have used Node red to create a user-interface , using which the users can interact with the design or get the classification for their given inputs.
* We have also developed an android app using which a user can edit the input parameters ,delete it and add any new values. Using the app , the user need not to wait for longer period of time for the results to come which may the doctor a leap of time before the cancer spreads to other tissues.

**1.3. Problem statement**

Breast cancer has become one the most common factor nowadays. Depite the fact , not all general hospitals have facilities to diagnose breast cancer to be malignant or benign through mammograms. The time taken for diagnosis may result in increase of the possibility of cancer spreading(if the cancer is of malignant type). Thus, we use machine learning models to diagnose and classify the breast cancer and reduce the death rate. To classify the benign and malignant cancer we have used machine learning algorithms and have provided a simpler access to the result using either an user-interface or an app.

**1.4. Related works**

There are many existing works for detection of breast cancer using machine learning techniques. The performance comparision of support vector machine (svm), k\_nearest neighbor (knn), decision trees ,naïvebayes machine learning techniques were shown.Winconsin’s breast cancer diagnosis

breast cancer (WDBC) dataset was use for this work.

|  |  |  |
| --- | --- | --- |
| Name | Algorithm | Accuracy |
| Alireza Osarech | Support Vector Machine | 96.5% |
| Vikas C , BB Tiwari , Saurabh Pal | Naïve Bayes | 97.3% |
| Sajjad Salaria | Support Vector Machine | 96.6% |

**2. Data Colletection:**

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.

Dataset is collected from the Kaggle, Winconsin Breast Detection Dataset.

**2.1. Input parameters**:

* ID number
* Diagnosis (M = malignant, B = benign) 3–32)
* radius (mean of distances from center to points on the perimeter)
* texture (standard deviation of gray-scale values)
* perimeter
* area
* smoothness (local variation in radius lengths)
* compactness (perimeter² / area — 1.0)
* concavity (severity of concave portions of the contour)
* concave points (number of concave portions of the contour)
* symmetry
* fractal dimension (“coastline approximation” — 1)
* The above parameters are then used to take out mean, standard deviation and worst(which is the mean of the three largest values)

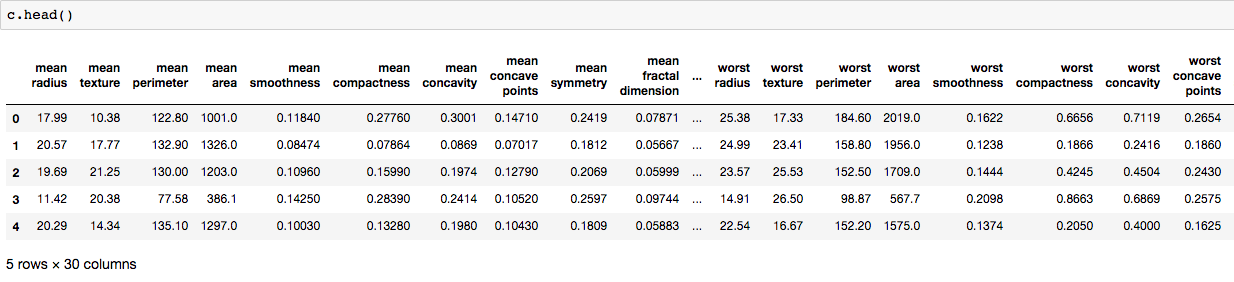
**2.2. Output parameters** : Malignant(M) or Benign(B)[categorical output]

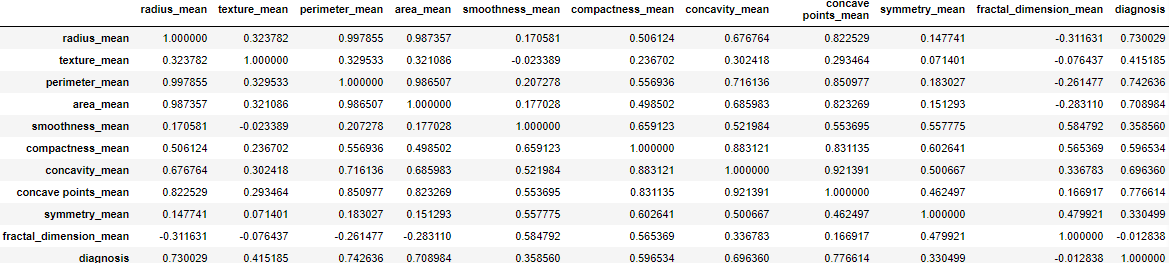
**3. Methodology**

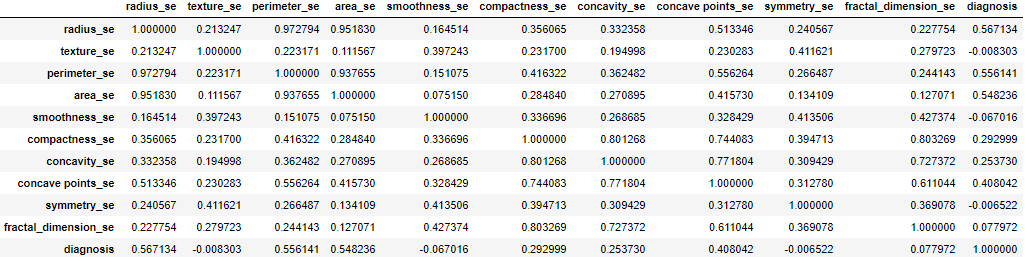
* When we analyze the dataset, we notice that the output is categorical and all the data has been labelled .Thus, we have to approach using supervised learning under which we will use classification methods.
* We then correlated all the data that we have to the output.
* Take out the columns that have lesser correlation to the output.
* Checked the accuracy rate of is algorithm using accuracy \_score ,confusion\_matrix and Region of Interest.
* We have choosed logistic regression to classify the type of breast cancer.
* Using IBM cloud (Watson Machine Learning) to deploy the trained model and also Node Red to create an user-interface for user access the model.
* Also , we have created an android app using android studio.

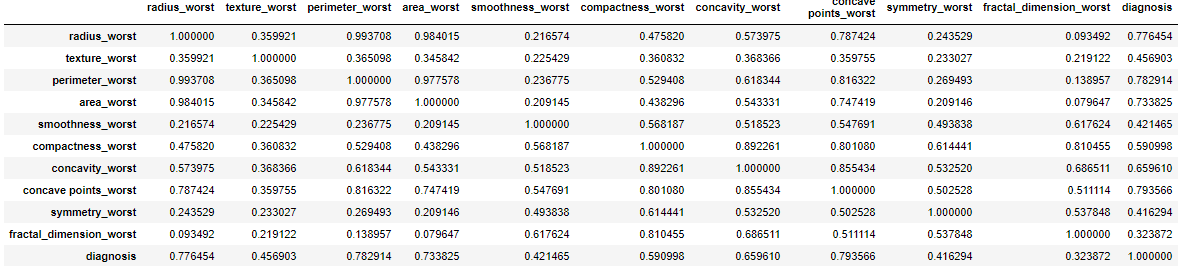
**4. Exploratory Data Analysis**:

**4.1.Dataset figures and tables**:

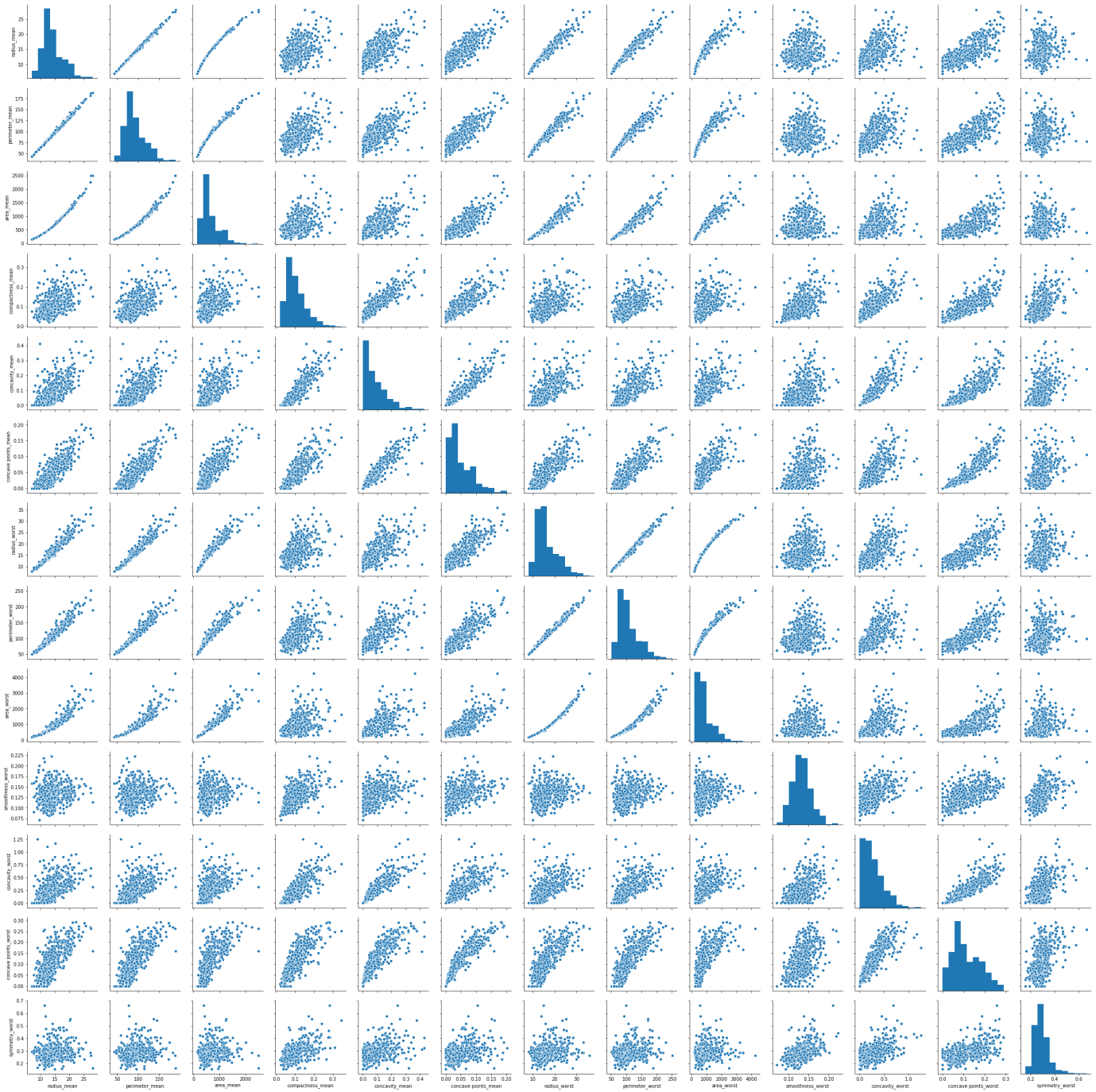


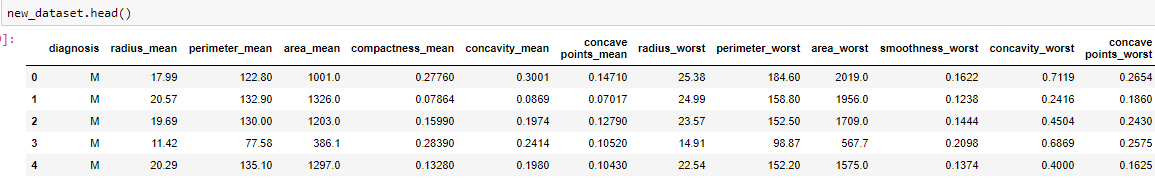
**4.2. Correlation Tables**:



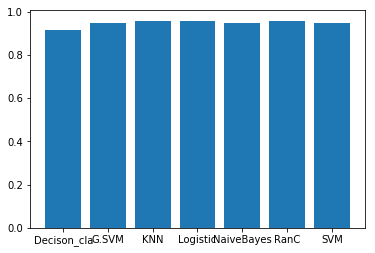


**4.3** **New Dataset figure and table:**

****



**4.4 Algorithm comparision:**

****

|  |  |  |
| --- | --- | --- |
| Algorithm | Accuracy | AUC |
| Support Vector Machine (SVM) | 95.61 | 94.6 |
| Naïve Bayes | 94.3 | 94.5 |
| Logistic regression | 96.49 | 95.8 |
| Gaussian SVM | 95.61 | 94.6 |
| KNN | 96.491 | 95.7 |
| RandomForestClassifier | 96.49 | 95.6 |
| Decision Tree Classifier | 93.4 | 91.2 |

5. **Data Modeling:**

We have our final dataset. And to get the accurate result we have now realized that, how important the training data phase is. We train the model in a way that it can predict(almost) correct results. In this dataset, we have split the data into train and test and rescaled it using pipelining. We will train the model on training data and predict the results on the test data. For this project we will use all the algorithms and find best algorithm i.e.) Logistic Regression. Before jumping into the code, let's get a little background about the Logistics Regression. Logistic regression can be used also to solve problems of classification. In general, logistic regression classifier can use a linear combination of more than one feature value or explanatory variable as argument of the sigmoid function. The corresponding output of the sigmoid function is a number between 0 and 1. The middle value is considered as threshold to establish what belong to the class 1 and to the class 0. In particular, an input producing an outcome greater than 0.5 is considered belong to the class 1.

**Implementation:**

import pandas as pd

import numpy as np

data=pd.read\_csv(“path~”)

x=dataset.iloc[:,2:].values

y=dataset.iloc[:,1].values

from sklearn.preprocessing import LabelEncoder

lb=LabelEncoder()

y=lb.fit\_transform(y)

dataset=pd.concat([data,y\_data],axis=1)

new\_dataset=dataset.drop(["id","texture\_mean","smoothness\_mean","symmetry\_mean","fractal\_dimension\_mean","radius\_se","texture\_se","perimeter\_se","area\_se","smoothness\_se","compactness\_se","concavity\_se","concavepoints\_se","symmetry\_se","fractal\_dimension\_se","texture\_worst","compactness\_worst", "fractal\_dimension\_worst"],axis=1)

x\_new=new\_dataset.iloc[:,1:]

y\_new=new\_dataset.iloc[:,0:1]

from sklearn.preprocessing import LabelEncoder

lb=LabelEncoder()

y\_new=lb.fit\_transform(y\_new)

y\_new=pd.DataFrame(y\_new)

y\_new

from sklearn.model\_selection import train\_test\_split

X\_train,X\_test,Y\_train,Y\_test=train\_test\_split(x\_new,y\_new,test\_size=0.2,random\_state=0)

from sklearn.preprocessing import MinMaxScaler

sc =MinMaxScaler(feature\_range=(0, 1))

from sklearn.linear\_model import LogisticRegression

model\_logr=LogisticRegression()

from sklearn.pipeline import Pipeline

pipeline=Pipeline([('scaler',sc),('logis',model\_logr)])

model1=pipeline.fit(X\_train,Y\_train)

y\_predlogr=pipeline.predict(X\_test)

from sklearn.metrics import accuracy\_score

from sklearn.metrics import confusion\_matrix

from sklearn.metrics import roc\_auc\_score

accuracy\_score(Y\_test,y\_predlogr)

from sklearn.metrics import confusion\_matrix

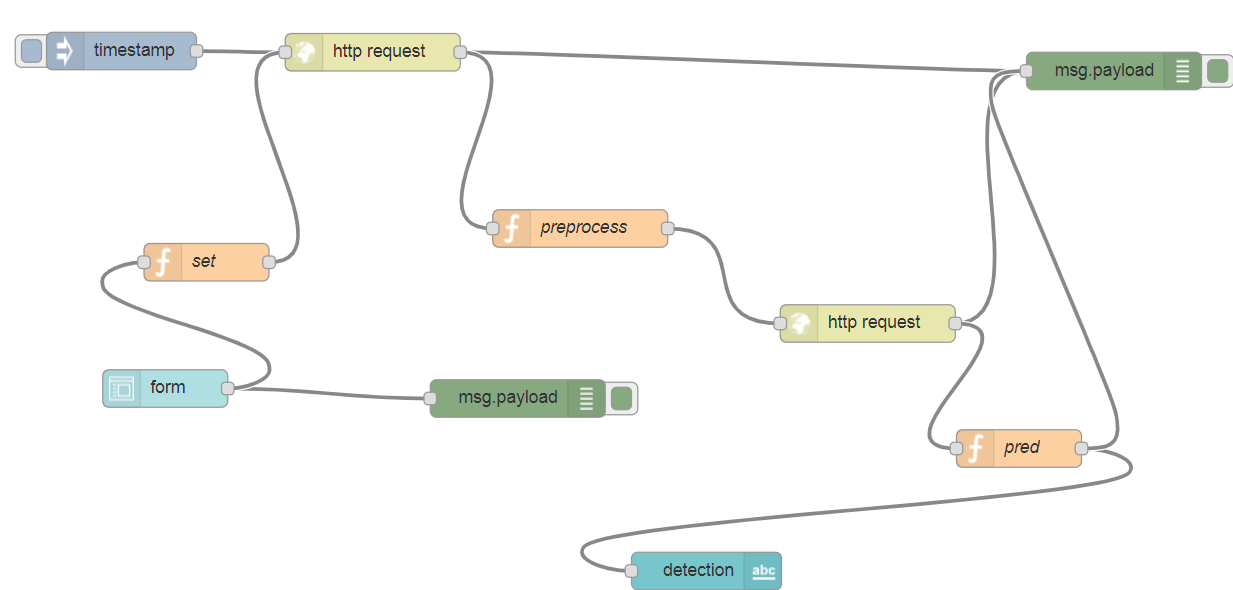
pd.DataFrame(confusion\_matrix(Y\_test,y\_predlogr),columns=["Prediction - 0","Prediction - 1"],index = ["Actual - 0","Actual - 1"])

import sklearn.metrics as metrics

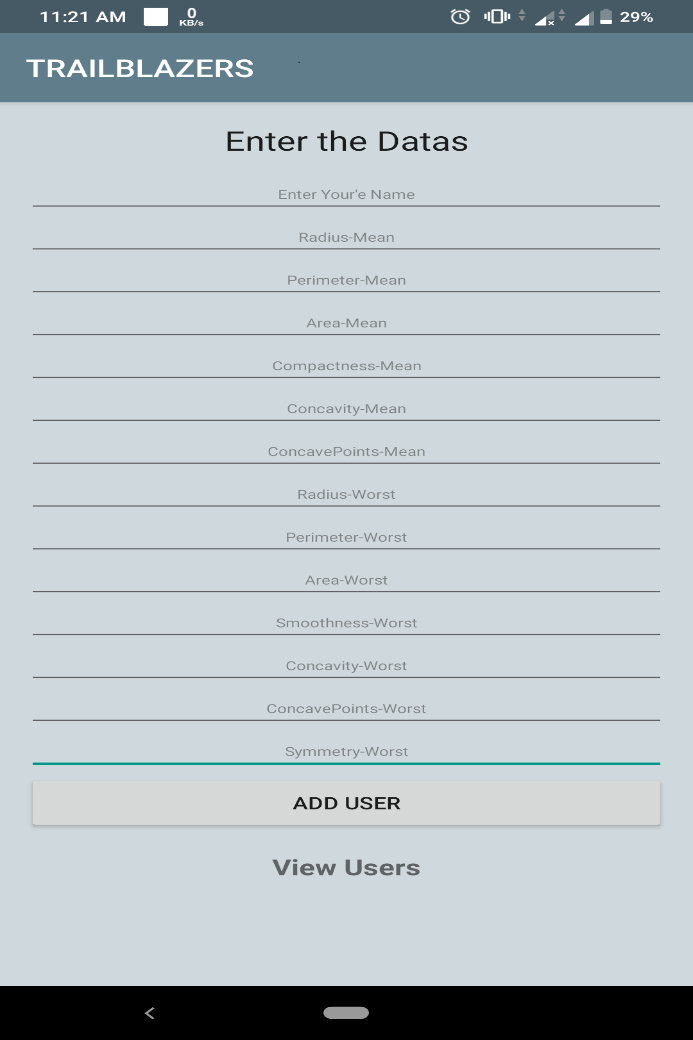
fpr,tpr,threshold=metrics.roc\_curve(Y\_test,y\_predlogr)

auc\_log=metrics.auc(fpr,tpr)

**6.Node Red**



7.**Android**



**References**

* Using\_Machine\_Learning\_Algorithms\_for\_Breast\_Cancer\_Risk\_Prediction\_and\_Diagnosis

<https://www.researchgate.net/publication/303028543_>

* <https://www.kaggle.com/junkal/breast-cancer-prediction-using-machine-learning>
* <https://towardsdatascience.com/building-a-simple-machine-learning-model-on-breast-cancer-data-eca4b3b99fa3>
* <https://developer.ibm.com/open/projects/node-red/>
* <https://developer.ibm.com/clouddataservices/docs/ibm-watson-machine-learning/docs/>

**Conclusion**

Our work mainly focused in the advancement of predictive models to achieve good accuracy in predicting valid disease outcomes using supervised machine learning methods. The analysis of the results signify that the integration of multidimensional data along with different classification, feature selection and dimensionality reduction techniques can provide auspicious tools for inference in this domain.