Breast Cancer Risk Prediction System

1.Introduction

Breast cancer (BC) is one of the most common cancers among women worldwide, representing the majority of new cancer cases and cancer-related deaths according to global statistics, making it a significant public health problem in today's society.

The early diagnosis of BC can improve the prognosis and chance of survival significantly, as it can promote timely clinical treatment to patients. Further accurate classification of benign tumors can prevent patients undergoing unnecessary treatments. Thus, the correct diagnosis of BC and classification of patients into malignant or benign groups is the subject of much research. Because of its unique advantages in critical features detection from complex BC datasets, machine learning (ML) is widely recognized as the methodology of choice in BC pattern classification and forecast modelling.

1.1 Overview

In the conventional way of diagnosing breast cancer some tests and procedures are carried out. These tests include Breast exam Mammogram Breast ultrasound Biopsy. As an alternative we can also use Machine Learning techniques for the classification of benign and malignant tumors. The prior diagnosis of Breast Cancer can enhance the prediction and survival rate notably, so that patients can be informed to take clinical treatment at the right time. Classification of benign tumors can help the patients avoid undertaking needless treatments. Thus the research is to be carried for the proper diagnosis of Breast Cancer and categorization of patients into malignant and benign groups. Machine Learning, with its advancements in detection of critical features from the complex datasets is largely acknowledged as the method in the prediction of breast cancer. Application of data mining techniques in the medical field can help in prediction of outcomes, minimizing the cost of medicines, aid people's health, upgrade the healthcare value and to rescue lives of people. This process of classifying benign and malignant tumors can be best done by the application of Classification techniques of machine learning. Lot of research is being conducted in this area by the application of various machine learning and data mining techniques for many different datasets on Breast Cancer. Most of them show that classification techniques give a good accuracy in prediction of the type of tumor.

1.2 PURPOSE

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.

We are building a model in Watson Studio and deploying the model in IBM Watson Machine Learning. To interact with the model we will be using Node-Red and scoring Endpoint.

2.LITERATURE SURVEY

2.1 Existing problem

Breast cancer is the most common malignancy among women, accounting for nearly 1 in 3 cancers diagnosed among women in the United States, and it is the second leading cause of cancer death among women. Breast Cancer occurs because of abnormal growth of cells in the breast tissue, commonly referred to as a Tumor. A tumor does not mean cancer - tumors can be benign (not cancerous), pre-malignant (pre-cancerous), or malignant (cancerous). Tests such as MRI, mammogram, ultrasound, and biopsy are commonly used to diagnose breast cancer performed.

2.1 PROPOSED SOLUTION

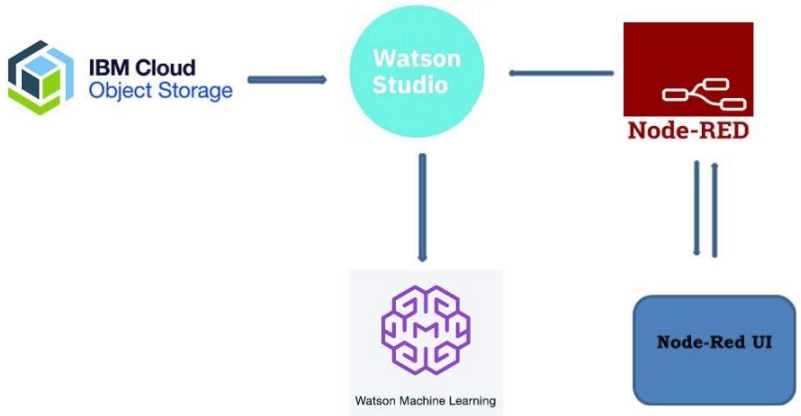
Building a model in Watson Studio and deploying the model in IBM Watson Machine Learning. To interact with the model, I will be using Node-Red and scoring Endpoint.

Solution Requirements:

Develop a model that is capable of detecting the Breast Cancer in early stages. The Machine learning model is trained and deployed on IBM Watson Studio and an endpoint is created. The web application is built using IBM Node-Red.

3.THEORITICAL ANALYSIS

3.1 BLOCK DIAGRAM



3.2 Hardware / Software designing

IBM Node red, IBM Watson Studio, IBM Machine Learning, IBM Cloud Object Storage

3.3 SOLUTION REQUIREMENTS

Develop a model that is capable of detecting the Breast Cancer in early stages. The Machine learning model is trained and deployed on IBM Watson Studio and an endpoint is created. The web application is built using IBM Node-Red.

4.EXPERIMENTAL INVESTIGATION

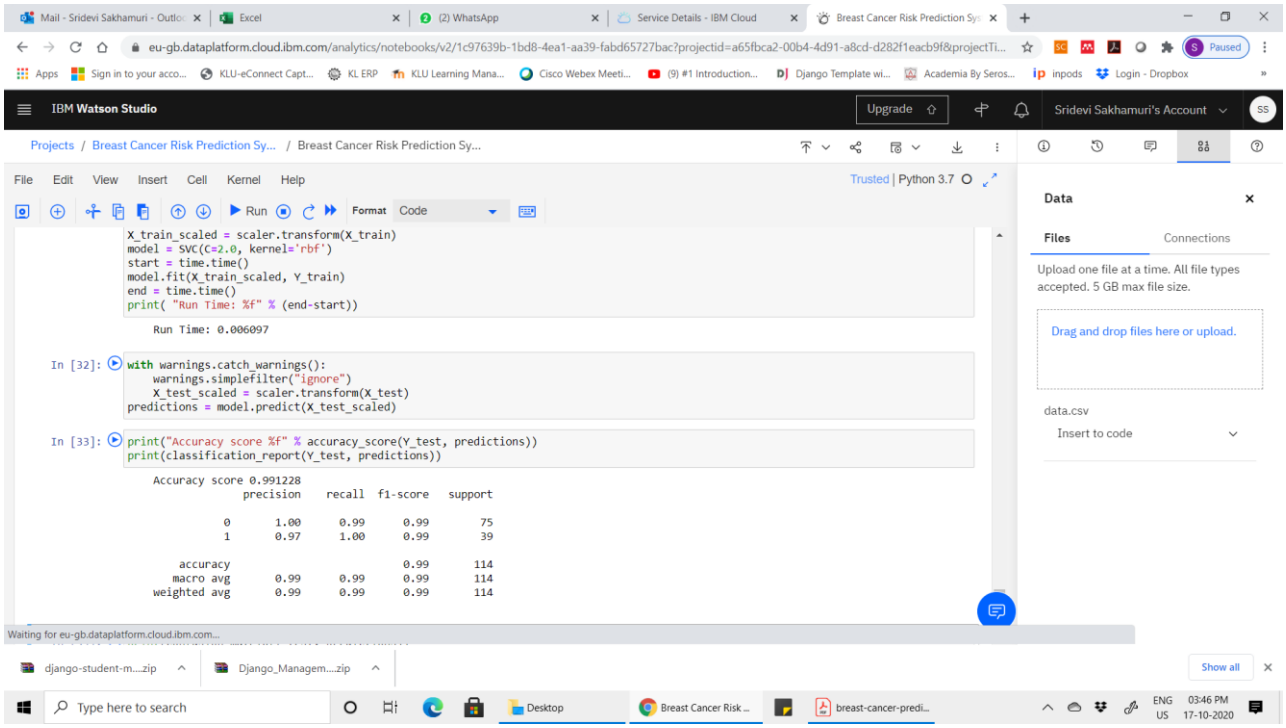
The dataset in the given problem consist of 32 columns and 569 rows. Totally there are 569 records with 30 featured values.

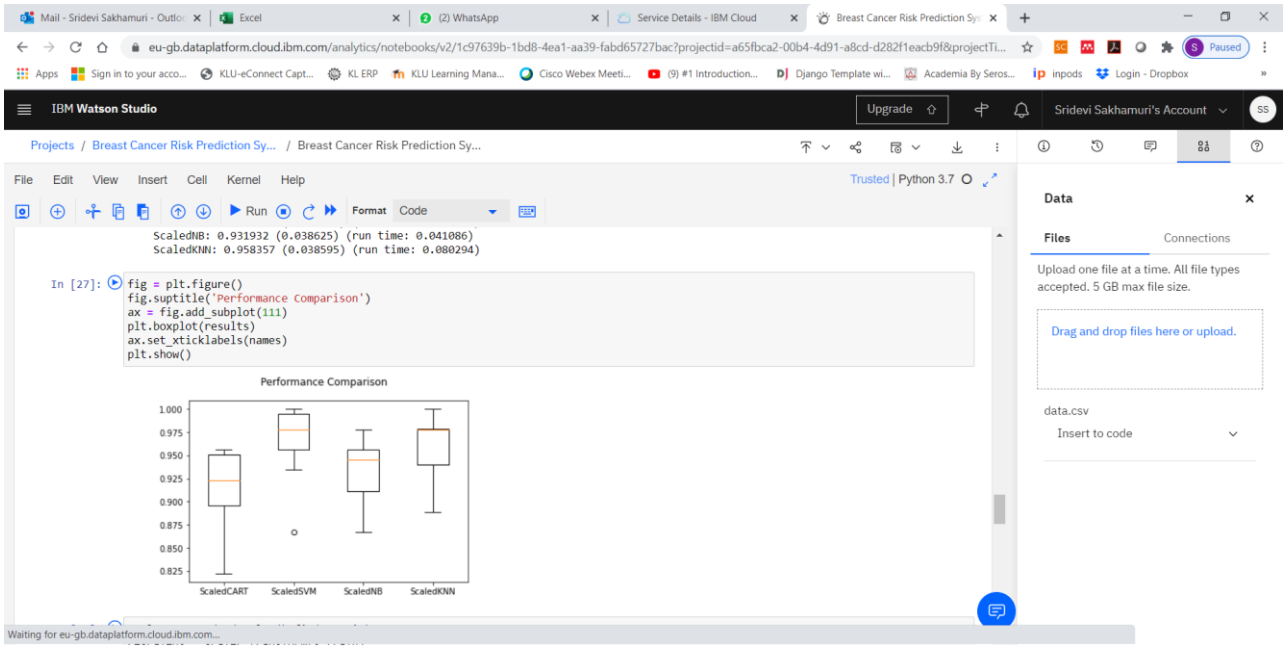
The prediction was to make whether according to the data from the previous data, is the new patient likely to have breast cancer or not.

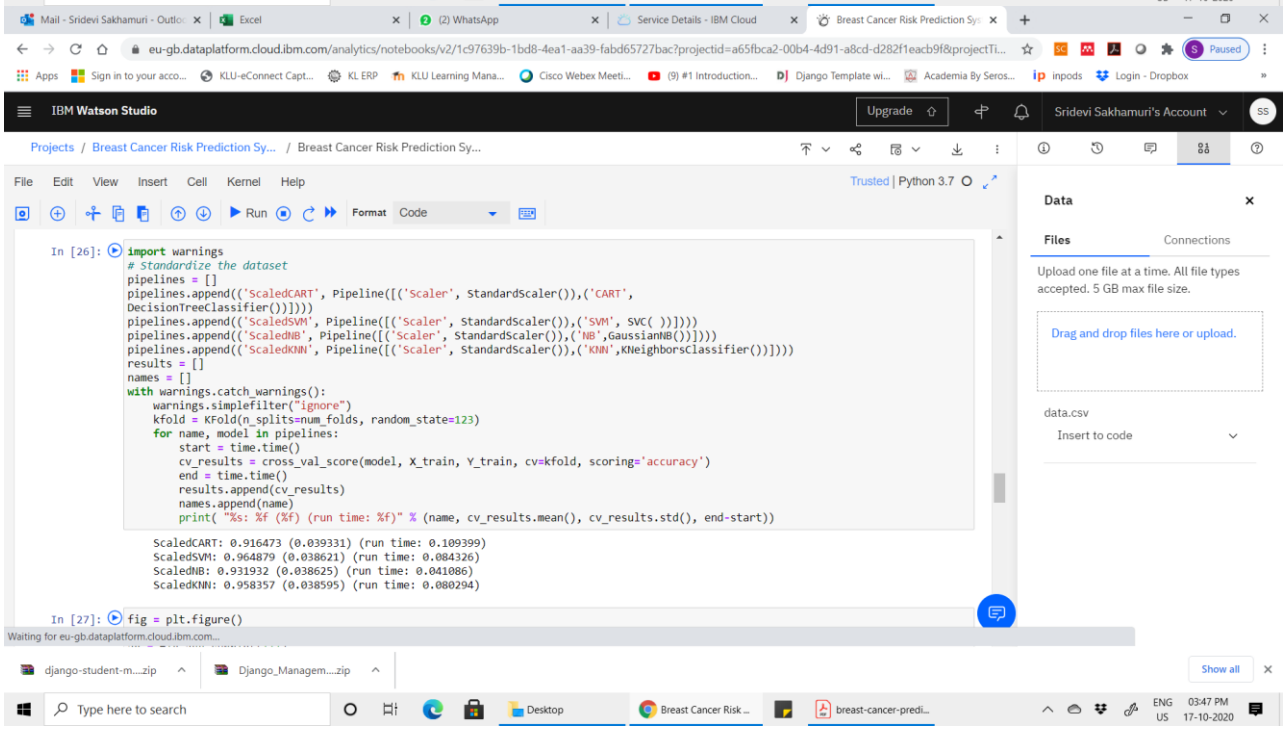
Diagnosis column has the output as M and B, where M represents patient with the cancer, whereas B represents no cancer. During analysis of the dataset, and studying the heat map of the correlation between different features, I came to a conclusion that no part of the dataset should not be ignored/deleted.

There is no categorical data in the dataset. Batch of the data is bisected into 80:20 ratio of training test: test data.

5. RESULTS:




IBM Watson Studio
Upgrade
Sridevi Sakhamuri's Account
Projects / Breast Cancer Risk Prediction Sy... / Breast Cancer Risk Prediction Sy...
File Edit View Insert Cell Kernel Help
Trusted | Python 3.7
Run Format Code
ScaledNB: 0.931932 (0.038625) (run time: 0.041086)
ScaledKNN: 0.958357 (0.038595) (run time: 0.080294)
In [27]:
fig = plt.figure()
fig.suptitle('Performance Comparison')
ax = fig.add_subplot(111)
plt.boxplot(results)
ax.set_xticklabels(names)
plt.show()
Performance Comparison
1.000
0.975
0.950
0.925
0.900
0.875
0.850
0.825
ScaledCART
ScaledSVM
ScaledNB
ScaledKNN
Data
Files
Connections
Upload one file at a time. All file types accepted. 5 GB max file size.
Drag and drop files here or upload.
data.csv
Insert to code
Waiting for eu-gb.dataplatform.cloud.ibm.com...


IBM Watson Studio
Upgrade
Sridevi Sakhamuri's Account
Projects / Breast Cancer Risk Prediction Sy... / Breast Cancer Risk Prediction Sy...
File Edit View Insert Cell Kernel Help
Trusted | Python 3.7
Run Format Code
In [26]:
import warnings
# Standardize the dataset
pipelines = []
pipelines.append(('ScaledCART', Pipeline([('Scaler', StandardScaler()),('CART',
DecisionTreeClassifier())])))
pipelines.append(('ScaledSVM', Pipeline([('Scaler', StandardScaler()),('SVM', SVC( ))])))
pipelines.append(('ScaledNB', Pipeline([('Scaler', StandardScaler()),('NB',GaussianNB())])))
pipelines.append(('ScaledKNN', Pipeline([('Scaler', StandardScaler()),('KNN',KNeighborsClassifier())])))
results = []
names = []
with warnings.catch_warnings():
warnings.simplefilter("ignore")
kfold = KFold(n_splits=num_folds, random_state=123)
for name, model in pipelines:
start = time.time()
cv_results = cross_val_score(model, X_train, Y_train, cv=kfold, scoring='accuracy')
end = time.time()
results.append(cv_results)
names.append(name)
print("%s: %f (%f) (run time: %f)" % (name, cv_results.mean(), cv_results.std(), end-start))
ScaledCART: 0.916473 (0.039331) (run time: 0.109399)
ScaledSVM: 0.964879 (0.038621) (run time: 0.084326)
ScaledNB: 0.931932 (0.038625) (run time: 0.041086)
ScaledKNN: 0.958357 (0.038595) (run time: 0.080294)
In [27]:
fig = plt.figure()
Data
Files
Connections
Upload one file at a time. All file types accepted. 5 GB max file size.
Drag and drop files here or upload.
data.csv
Insert to code
Waiting for eu-gb.dataplatform.cloud.ibm.com...

Mail - Sridevi Sakhamuri - Outlo | Excel | (2) WhatsApp | Service Details - IBM Cloud | Breast Cancer Risk Prediction Sy...

eu-gb.dataplatform.cloud.ibm.com/analytics/notebooks/v2/1c97639b-1bd8-4ea1-aa39-fabd65727bac?projectid=a65fbca2-00b4-4d91-a8cd-d282f1eacb9f&projectTi...

Apps | Sign in to your acco... | KLU-eConnect Capt... | KL ERP | KLU Learning Mana... | Cisco Webex Meeti... | (9) #1 Introduction... | Django Template wi... | Academia By Seros... | inpods | Login - Dropbox

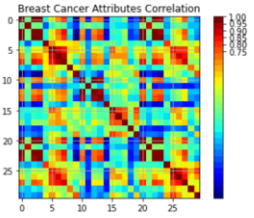
IBM Watson Studio | Upgrade | Sridevi Sakhamuri's Account

Projects / Breast Cancer Risk Prediction Sy... / Breast Cancer Risk Prediction Sy...

File Edit View Insert Cell Kernel Help | Trusted | Python 3.7

Run | Format Code

```
fig = plt.figure()
ax1 = fig.add_subplot(111)
cmap = cm.get_cmap('jet', 30)
cax = ax1.imshow(data.corr(), interpolation="none", cmap=cmap)
ax1.grid(True)
plt.title('Breast Cancer Attributes Correlation')
# Add colorbar, make sure to specify tick locations to match desired ticklabels
fig.colorbar(cax, ticks=[.75,.8,.85,.90,.95,1])
plt.show()
```



In [19]:
```
Y = data['diagnosis'].values
X = data.drop('diagnosis', axis=1).values
```

Waiting for eu-gb.dataplatform.cloud.ibm.com...

Data

Files | Connections

Upload one file at a time. All file types accepted. 5 GB max file size.

Drag and drop files here or upload.

data.csv

Insert to code

django-student-m....zip | Django_Managem....zip | Show all

Type here to search | Desktop | Breast Cancer Risk ... | breast-cancer-predi... | ENG US 03:47 PM 17-10-2020

Mail - Sridevi Sakhamuri - Outlo | Excel | (2) WhatsApp | Service Details - IBM Cloud | Breast Cancer Risk Prediction Sy...

eu-gb.dataplatform.cloud.ibm.com/analytics/notebooks/v2/1c97639b-1bd8-4ea1-aa39-fabd65727bac?projectid=a65fbca2-00b4-4d91-a8cd-d282f1eacb9f&projectTi...

Apps | Sign in to your acco... | KLU-eConnect Capt... | KL ERP | KLU Learning Mana... | Cisco Webex Meeti... | (9) #1 Introduction... | Django Template wi... | Academia By Seros... | inpods | Login - Dropbox

IBM Watson Studio | Upgrade | Sridevi Sakhamuri's Account

Projects / Breast Cancer Risk Prediction Sy... / Breast Cancer Risk Prediction Sy...

File Edit View Insert Cell Kernel Help | Trusted | Python 3.7

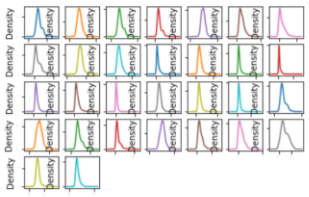
Run | Format Code

```
During handling of the above exception, another exception occurred:
```

In [14]:
```
print(data.groupby('diagnosis').size())
```

```
diagnosis
0    357
1    212
dtype: int64
```

In [16]:
```
data.plot(kind='density', subplots=True, layout=(5,7), sharex=False, legend=False, fontsize=1)
plt.show()
```



Waiting for eu-gb.dataplatform.cloud.ibm.com...

Data

Files | Connections

Upload one file at a time. All file types accepted. 5 GB max file size.

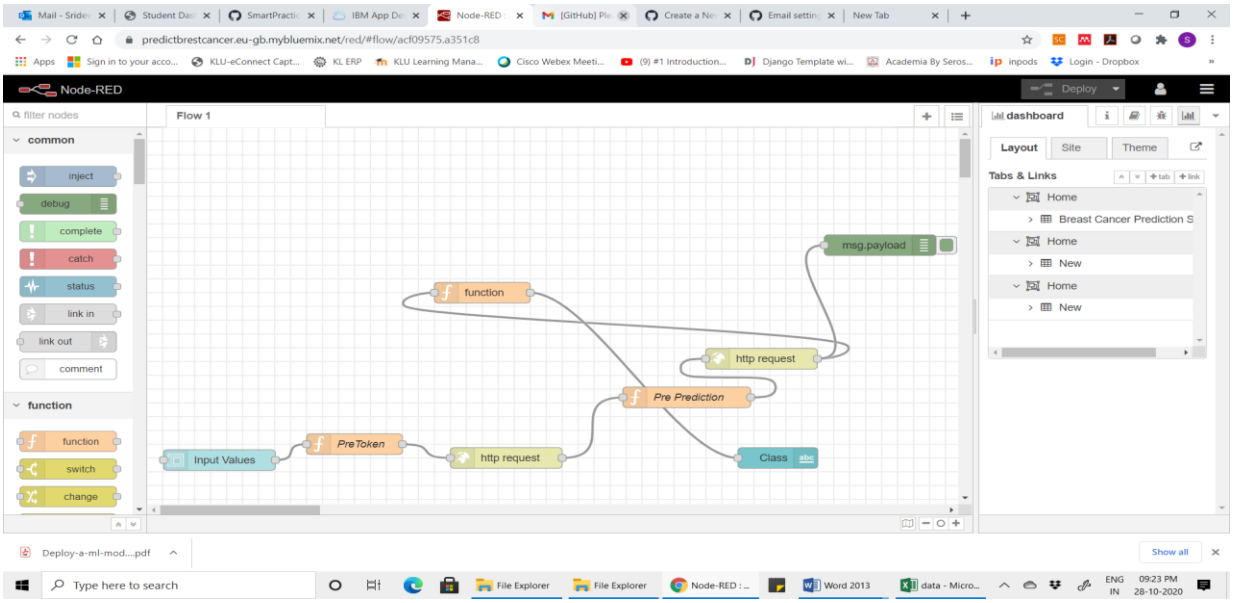
Drag and drop files here or upload.

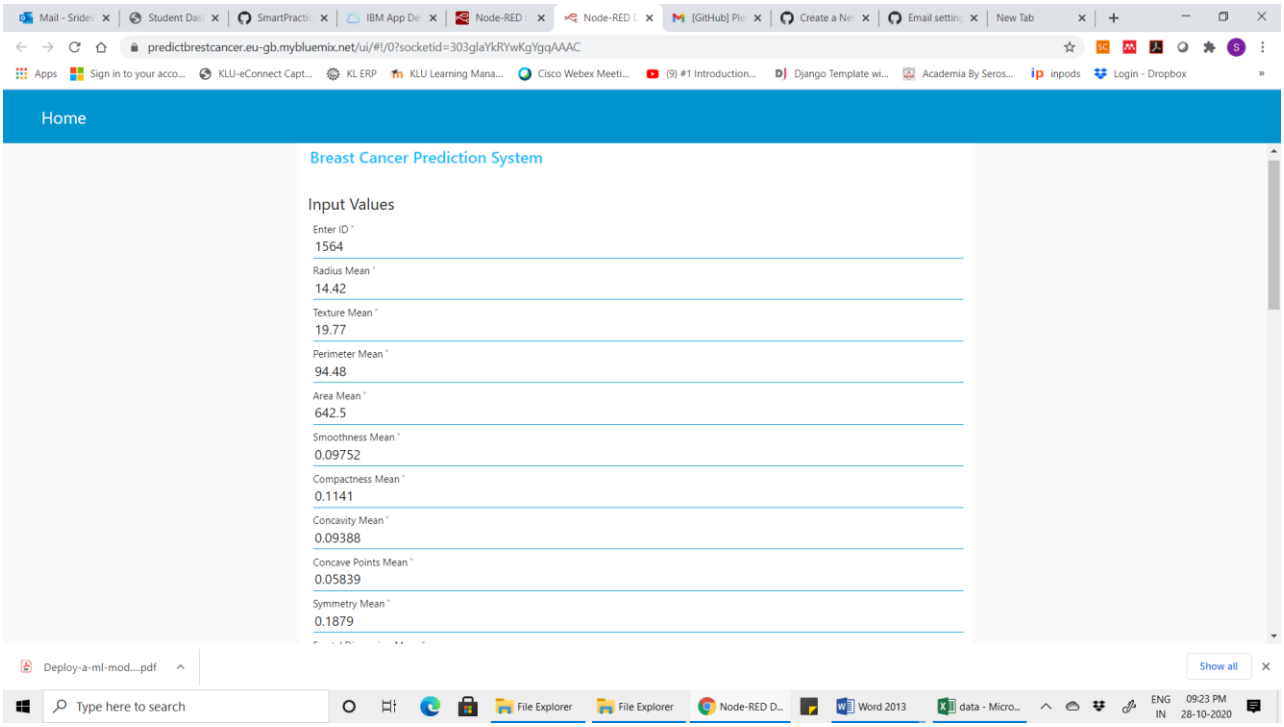
data.csv

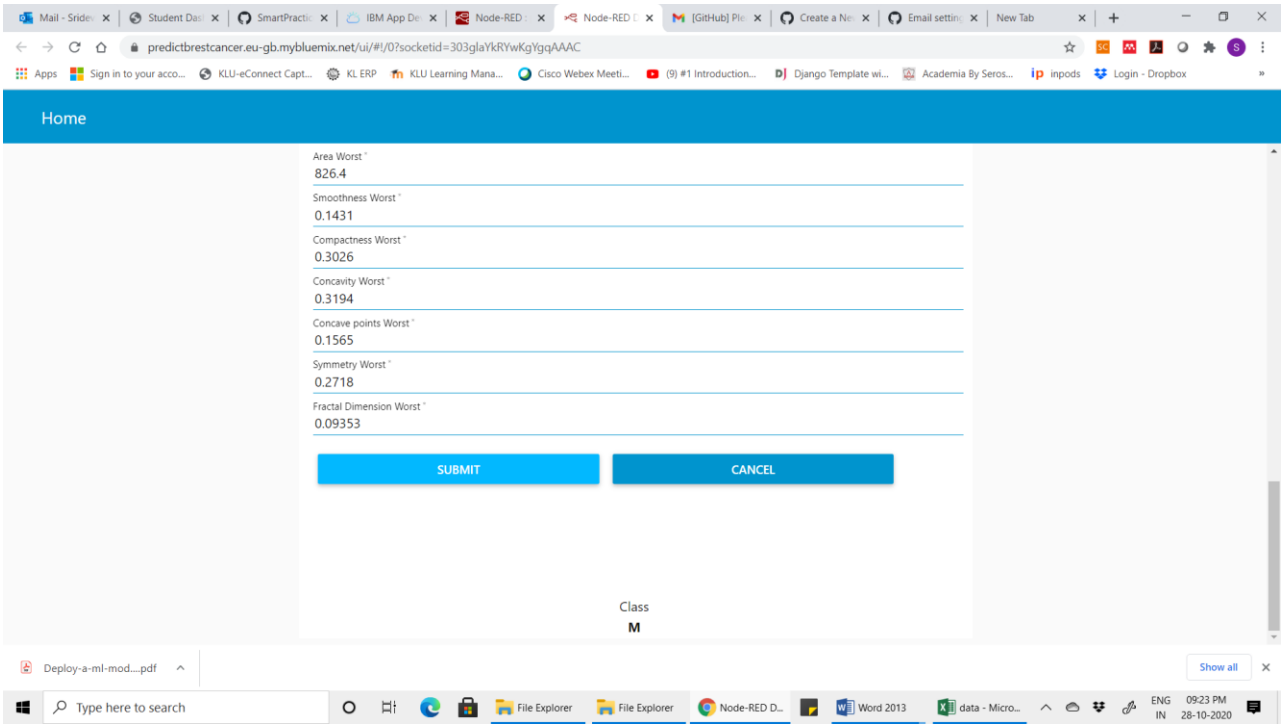
Insert to code

django-student-m....zip | Django_Managem....zip | Show all

Type here to search | Desktop | Breast Cancer Risk ... | breast-cancer-predi... | ENG US 03:48 PM 17-10-2020


predictbrestcancer.eu-gb.mybluemix.net/red/#flow/acf09575.a351c8
Node-RED
Deploy
Flow 1
common
inject
debug
complete
catch
status
link in
link out
comment
function
function
switch
change
msg.payload
function
http request
Pre Prediction
Input Values
PreToken
http request
Class
dashboard
Layout
Site
Theme
Tabs & Links
Home
Breast Cancer Prediction S
Home
New
Home
New


predictbrestcancer.eu-gb.mybluemix.net/ui/#!/0?socketid=303glaYkRYwKgYgqAAAC
Home
Breast Cancer Prediction System
Input Values
Enter ID *
1564
Radius Mean *
14.42
Texture Mean *
19.77
Perimeter Mean *
94.48
Area Mean *
642.5
Smoothness Mean *
0.09752
Compactness Mean *
0.1141
Concavity Mean *
0.09388
Concave Points Mean *
0.05839
Symmetry Mean *
0.1879



6. ADVANTAGES & DISADVANTAGES

ADVANTAGES:

1. Both the model and application is lightweight.
2. Prediction speed is high.
3. Server side is authenticated.
4. The prediction is helpful in educational purposes.

DISADVANTAGES:

1. Node-red is not suitable for commercial purposes.
2. Predictions on missing feature can be inaccurate.

7. APPLICATIONS

1. Through this model we can predict, whether a patient is likely to have cancer or not, without even doing medical tests.
2. Medical test can be modified and optimized.
3. We can analyze which starts of population (in women) are more likely to have it in the future.
4. The application can be run on android by SL4A.
5. It would be very useful and handy tool in healthcare.
6. It can run on PC server very fast.
7. It bypasses the first level of manual inspection.

8. CONCLUSION

Because of this ongoing covid-19 pandemic, now people have an urge to have pre knowledge of all medical ailments and advancements. For better working of the model we would be needed actual and large dataset. Since by the level of dataset in the repository, the results are "good" .In the source code I have 8 types of models, but on the IBM cloud, Node-red I have deployed it by using only LGBM classification algorithm. The model was trained on a dataset of 569 patients, the total number of features were 30. Feature scaling was very important in this problem set ,as the classification algorithm used demands uniform distancing between all features.

The model after initial testing was deployed IBM Watson and NODE-RED.

This intermediate level of machine learning is very necessary to understand to leave scope for potential developments.

9. FUTURE SCOPE

As per WHO breast cancer is a deadly cancer, which develops inside the human body without even showing symptoms. Since in the initial days of the disease, no symptoms are witnessed by the patient, the diseases develops into later stages of the deadly cancer .If by regular examination ,we can deploy the dataset into the model to predict ,to very good accuracy we can find about the cancer ,without any medical tests. That is why, this prediction algorithm has great future ahead, if it keeps on learning from bigger dataset. `