

Breast Cancer Risk Prediction Using IBM Auto AI

Internship Title : RSIP Career

Basic ML129 Duration: 1Month

ProjectID :SPS_PRO_300

ProjectTitle : Breast Cancer Risk Prediction using IBM Auto AI

UsingAuto AIService Team : GK

Team members : GOKUL K

Slot : June 22 - Slot(6)

Submitted by

GOKUL K

INDEX

Chapter No.	Title	Page No.
1	Introduction	3
	1.1 Overview	3
	1.2 Purpose	3
2	Literature Survey	4
	2.1 Existing problem	4
	2.2 Proposed Solution	4
3	Theoretical Analysis	5
	3.1 Block Diagram	5
	3.2 Hardware/Software Designing	5
4	Experimental Investigations	6 - 19
5	Flow Chart	20
6	Result	21
7	Advantages and Disadvantages	22
	7.1 Advantages	22
	7.2 Disadvantages	22
8	Applications	22
9	Conclusion	23
10	Future Scope	23
11	Bibliography	23

1.INTRODUCTION

1.1 Overview:

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 potential for improving the 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.

1.2 Purpose:

Early detection can give patients more treatment options. In order to detect signs of cancer, breast tissue from biopsies is stained to enhance the nuclei and cytoplasm for microscopic examination. Then, pathologists evaluate the extent of any abnormal structural variation to determine whether there are tumors.

Breast cancer may be noncancerous (benign) or cancerous (malignant). Most are noncancerous and not life threatening. Often, they do not require treatment. In contrast, breast cancer can mean loss of a breast or of life. Thus, for many women, breast cancer is their worst fear. However, potential problems can often be detected early when women regularly examine their breasts themselves, are examined regularly by their doctor, and have mammograms as recommended. Early detection of breast cancer can be essential to successful treatment.

Many women fear breast cancer, partly because it is common. However, some of the fear about breast cancer is based on misunderstanding. For example, the statement, "One of every eight women will get breast cancer," is misleading. That figure is an estimate based on women from birth to age 95. It means that theoretically, one of eight women who live to age 95 or older will develop breast cancer. However, a 40-year-old woman has only about a 1 in 70 chance of developing it during the next decade. But as she ages, her risk increases.

2.LITERATURE SURVEY

2.1 Existing Problem:

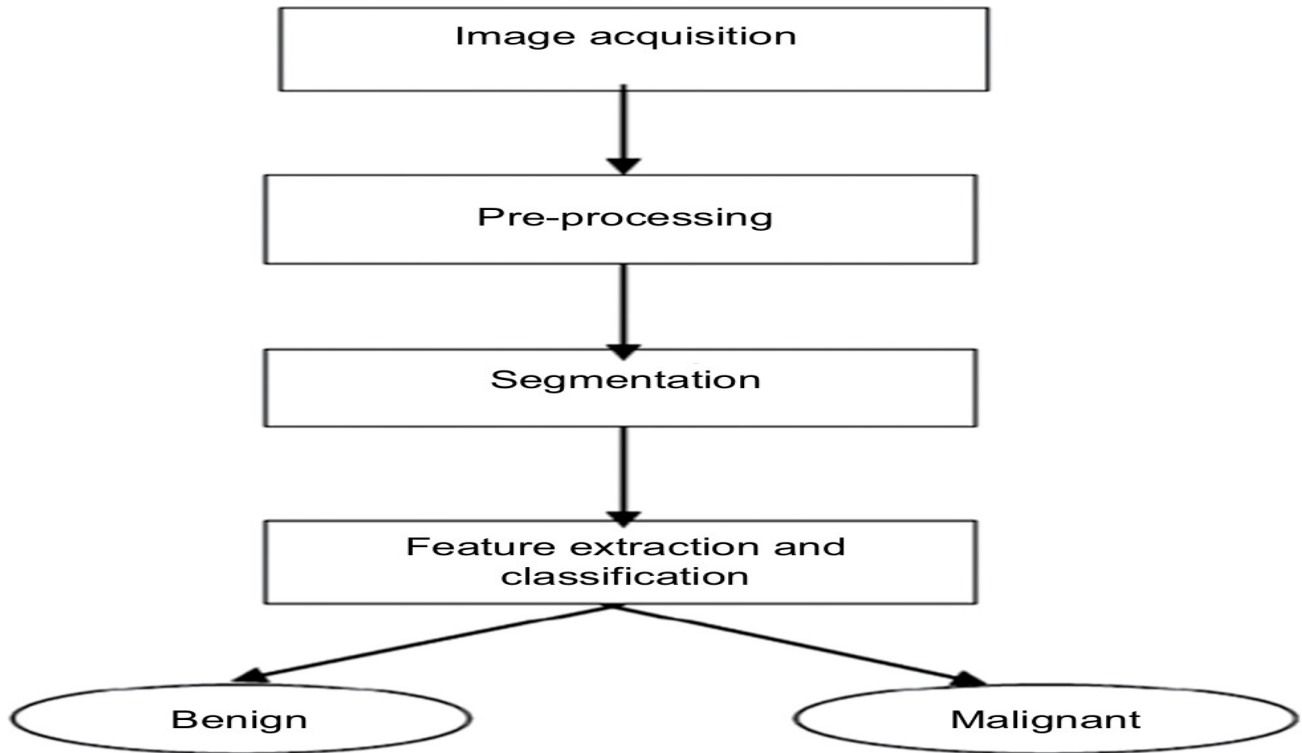
Among women, breast cancer is the most common cancer and the second most common cause of cancer deaths. Typically, the first symptom is a painless lump, usually noticed by the woman. Breast cancer screening recommendations vary and include periodic mammography, breast examination by a doctor, and breast self-examination. If a solid lump is detected, doctors use a hollow needle to remove a sample of tissue or make an incision and remove part or all of the lump and then examine the tissue under a microscope (biopsy). Breast cancer almost always requires surgery, sometimes with radiation therapy, chemotherapy, other drugs, or a combination. Outcome is hard to predict and depends partly on the characteristics and spread of the cancer.

2.2 Proposed Solution:

Here we are developing a machine learning model where in the model gets trained by considering the parameters such as: Radius ,Texture, Perimeter, Area, Smoothness, Concavity, Concaveness, Compactness here all these parameters are taken in mean, se and overall values are been taken. And the model is been trained using Auto AI service in IBM Watson cloud and that can be deployed in an application such as web or mobile applications.

3.THEORETICAL ANALYSIS

3.1BlockDiagram:



3.2 Hardware/ SoftwareDesigning:

This dataset is first tested by using various algorithms in our jupyter notebooks and then implemented in the IBM Cloud Platform. We upload our dataset in the cloud platform and choose the parameter to be predicted and we choose the number of algorithms and pipelines to be used. The cloud platform then predicts the best suited algorithm for our dataset along with the accuracy. We can also compare the performance of other algorithms used. The Auto AI function in the IBM cloud aids in deployment of our final machine learning models. This helps us to implement and test our model for our dataset. We have then, created a node red app for our deployed model. This UI will help us predict the avalanche in real time as we enter the details. This app aids in easy usage and better userinterface.

4.EXPERIMENTAL INVESTIGATIONS

Step_1 : Data Collection

We downloaded the dataset provided from Kaagle and did data pre-processing. We applied various algorithms in our jupyter notebook to the dataset to find the best one. We need to predict the Begnin or Malignant of the cancer .The given dataset is not a set of values which are categorical like 0 or 1. Hence we apply the regression algorithms to find the best fit algorithm for the given dataset.

Data pre-processing and analysing:

We import the dataset and find the correlation between the given values. and find if there are any null values.

First we will see the columns and several rows and their values.

```
df.head()
```

	mean_radius	mean_texture	mean_perimeter	mean_area	mean_smoothness	diagnosis
0	17.99	10.38	122.80	1001.0	0.11840	0
1	20.57	17.77	132.90	1326.0	0.08474	0
2	19.69	21.25	130.00	1203.0	0.10960	0
3	11.42	20.38	77.58	386.1	0.14250	0
4	20.29	14.34	135.10	1297.0	0.10030	0

Here we will check whether there are any missing values.

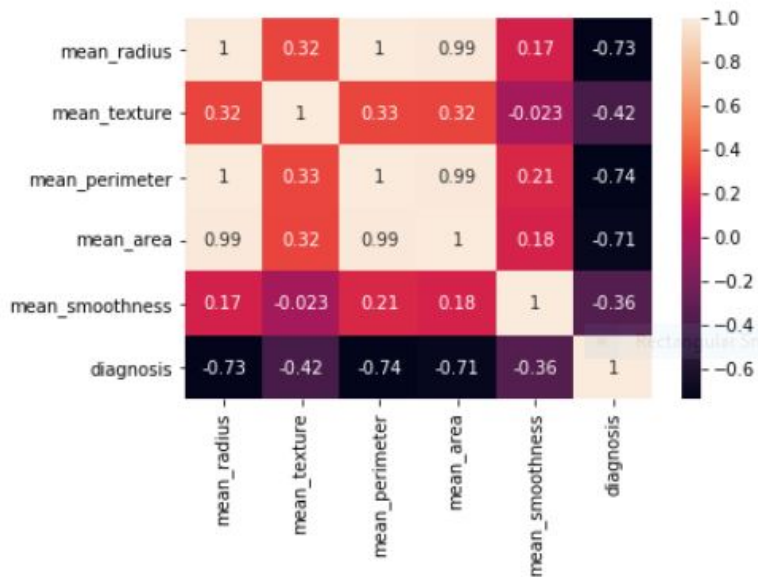
```
df.isna().any()
```

```
5]: mean_radius      False
    mean_texture     False
    mean_perimeter   False
    mean_area        False
    mean_smoothness  False
    diagnosis        False
    dtype: bool
```

As we can conclude there are no missing values from the above result, next we visualise the correlation between the features.

```
import seaborn as sns
sns.heatmap(df.corr(),annot=True)
```

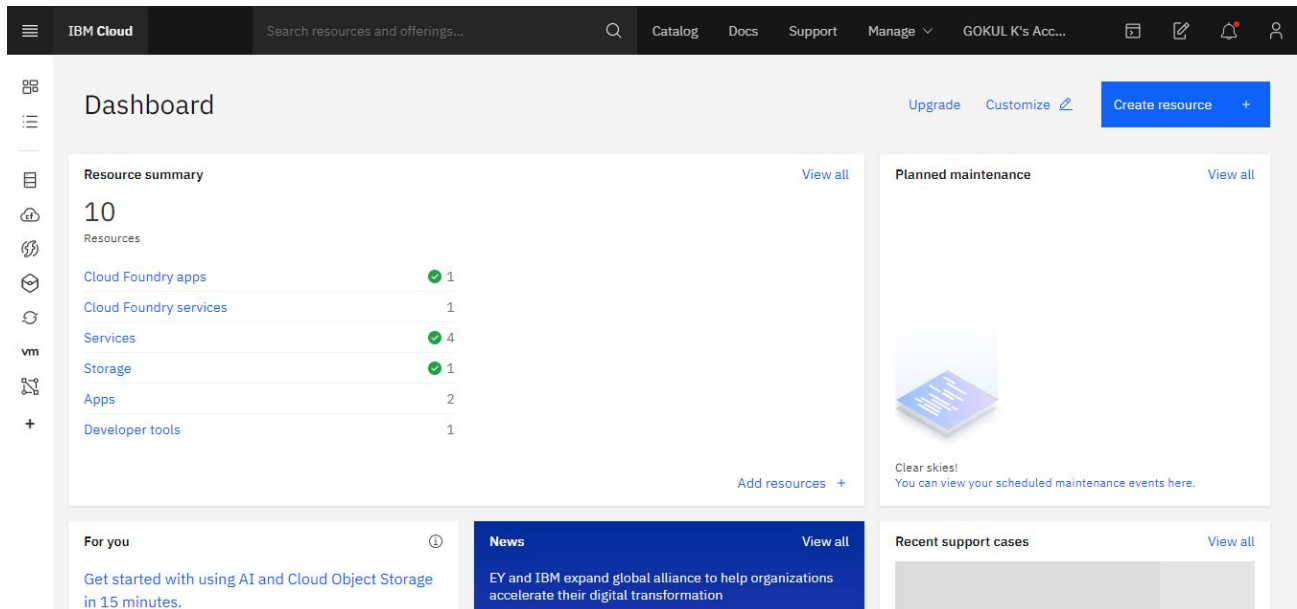
<matplotlib.axes._subplots.AxesSubplot at 0x20c33a586c8>



We are going to predict the diagnosis feature using the other features available in the dataset.

Step-3 : IBM Cloud Account: Creating an account

We have successfully created an IBM Cloud account

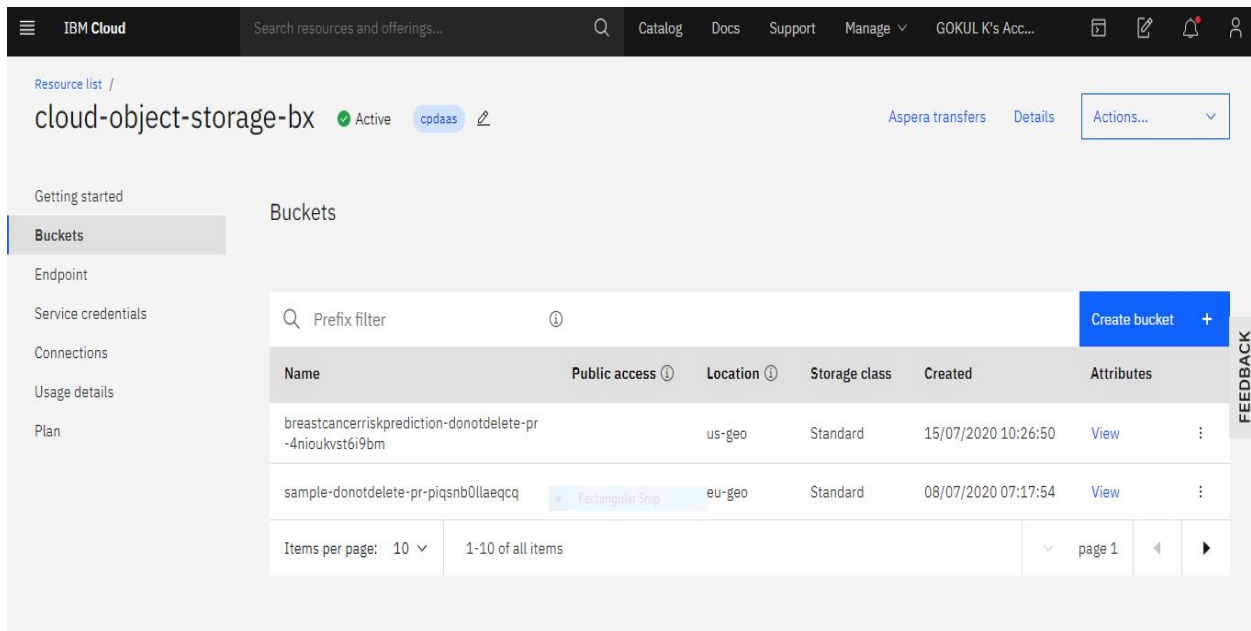


The screenshot shows the IBM Cloud Dashboard. The top navigation bar includes the IBM Cloud logo, a search bar, and links to Catalog, Docs, Support, Manage, and the user's account (GOKUL K's Acc...). The dashboard itself has a sidebar with icons for various services. The main content area is titled 'Dashboard' and includes a 'Resource summary' section showing 10 resources. Below this, there are sections for 'Planned maintenance', 'For you' (with a link to 'Get started with using AI and Cloud Object Storage in 15 minutes'), 'News' (with a link to 'EY and IBM expand global alliance to help organizations accelerate their digital transformation'), and 'Recent support cases'.

Resource	Count
Cloud Foundry apps	1
Cloud Foundry services	1
Services	4
Storage	1
Apps	2
Developer tools	1

Creating cloud object storage:

We have created a storage and have created a bucket to store our projects.



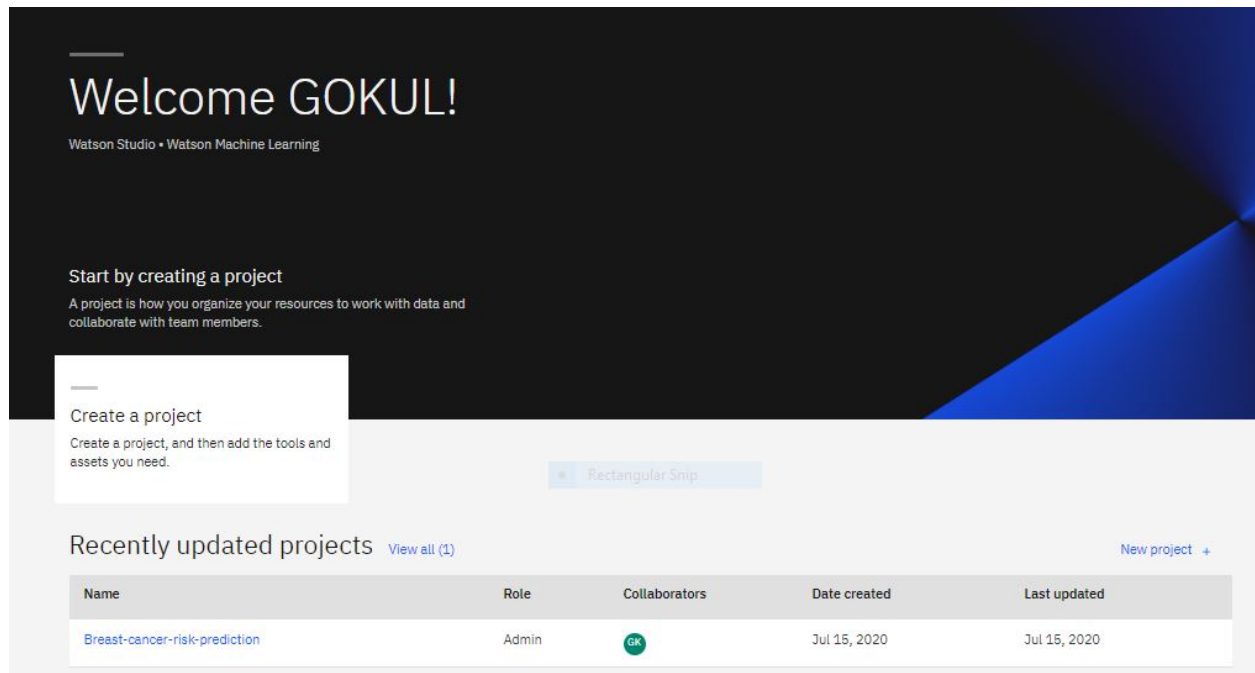
The screenshot shows the IBM Cloud Object Storage 'Buckets' page. The top navigation bar is the same as the dashboard. The page title is 'cloud-object-storage-bx' with a status of 'Active' and a link to 'cpdaas'. The left sidebar shows a list of options: 'Getting started', 'Buckets' (selected), 'Endpoint', 'Service credentials', 'Connections', 'Usage details', and 'Plan'. The main content area is titled 'Buckets' and includes a 'Prefix filter' search bar and a 'Create bucket' button. Below this is a table listing the buckets.

Name	Public access	Location	Storage class	Created	Attributes
breastcancerriskprediction-donotdelete-pr-4nioukvst6i9bm		us-geo	Standard	15/07/2020 10:26:50	View
sample-donotdelete-pr-piqsnb0llaecq		eu-geo	Standard	08/07/2020 07:17:54	View

Items per page: 10 | 1-10 of all items | page 1

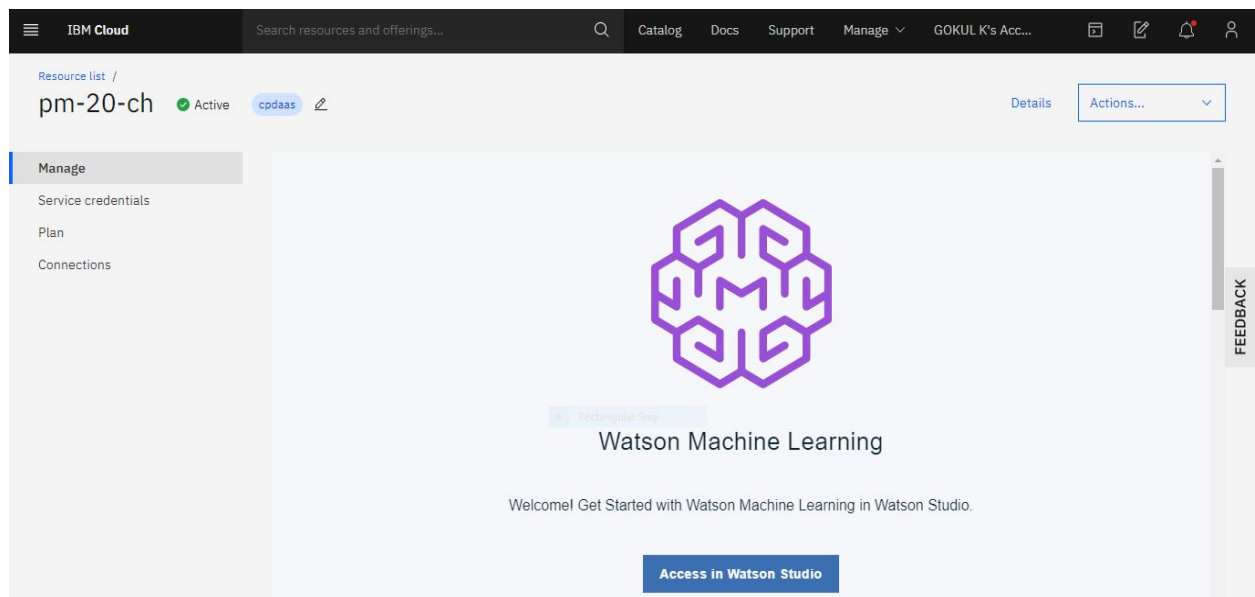
Creating Watson Studio Platform

We have created a Watson studio platform to predict our data and to implement, deploy and test our model in the real time.



Creating a ML Service:

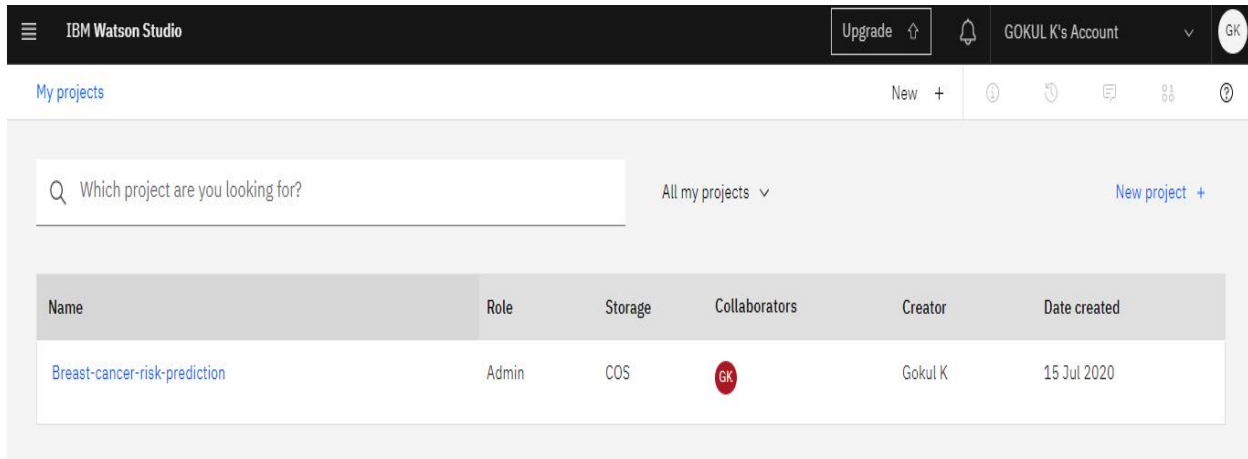
We have created a ML service to load and process our dataset



Step3:- Model Building

Creating a project in Watson Studio:

We have created a project in Watson studio namely Breast-cancer-risk-prediction.



IBM Watson Studio Desktop is a desktop client tool for solving your problems by analysing data with artificial intelligence. With Watson Studio Desktop, you can prepare data and build models on your desktop with visual drag and drop tools. You organize your resources for data analysis tasks in projects. Each project has its own directory on your computer. You can choose a standard project or to import a project that was previously exported from Watson Studio Desktop.

Auto AI Experiment in add Projects and set up AI environment:

We have created an auto AI experiment called Intern. AutoAI is available within [IBM Watson Studio](#) with one-click deployment through Watson Machine Learning. To help simplify an AI lifecycle management, AutoAI automates:

- Datapreparation
- Modeldevelopment
- Feature engineering
- Hyper-parameteroptimization

AutoAI experiments New AutoAI experiment +			
Name	Status	Model type	Last modified
risk-prediction	Completed	Binary Classification	Jul 15, 2020, 08:03 PM
risk-prediction	Completed	Binary Classification	Jul 15, 2020, 04:16 PM

Import Dataset:

We have then imported the dataset in the name of copy.csv in the assests section.

IBM Watson Studio

Upgrade

My projects / Breast-cancer-risk-prediction

Launch IDE

Add to project +

Overview

Assets

Environments

Jobs

Deployments

Access Control

Se

What assets are you looking for?

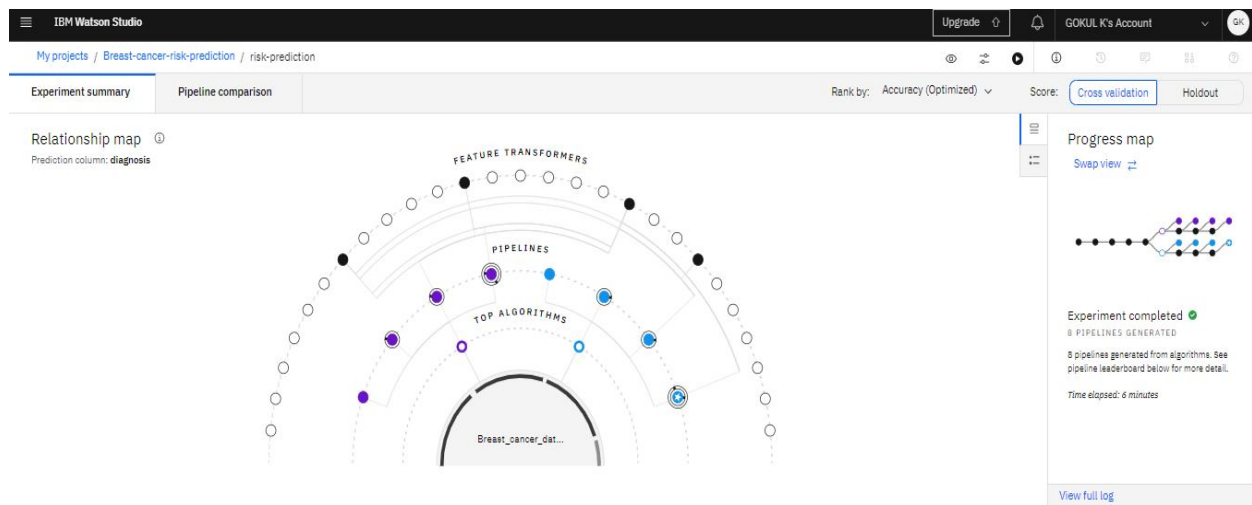
Data assets

0 assets selected.

<input type="checkbox"/>	Name	Type	Created by	Last modified
<input type="checkbox"/>	CSV Breast_cancer_data.csv	Data Asset	GOKUL K	Jul 15, 2020, 07:56 PM

Run the model and select the pipeline:

We then run the model, by choosing what to predict, the ratio of train and testing dataset and the number of pipelines to be used. We choose the default 90:10 ratio for training and testing and we chose 8 pipelines.

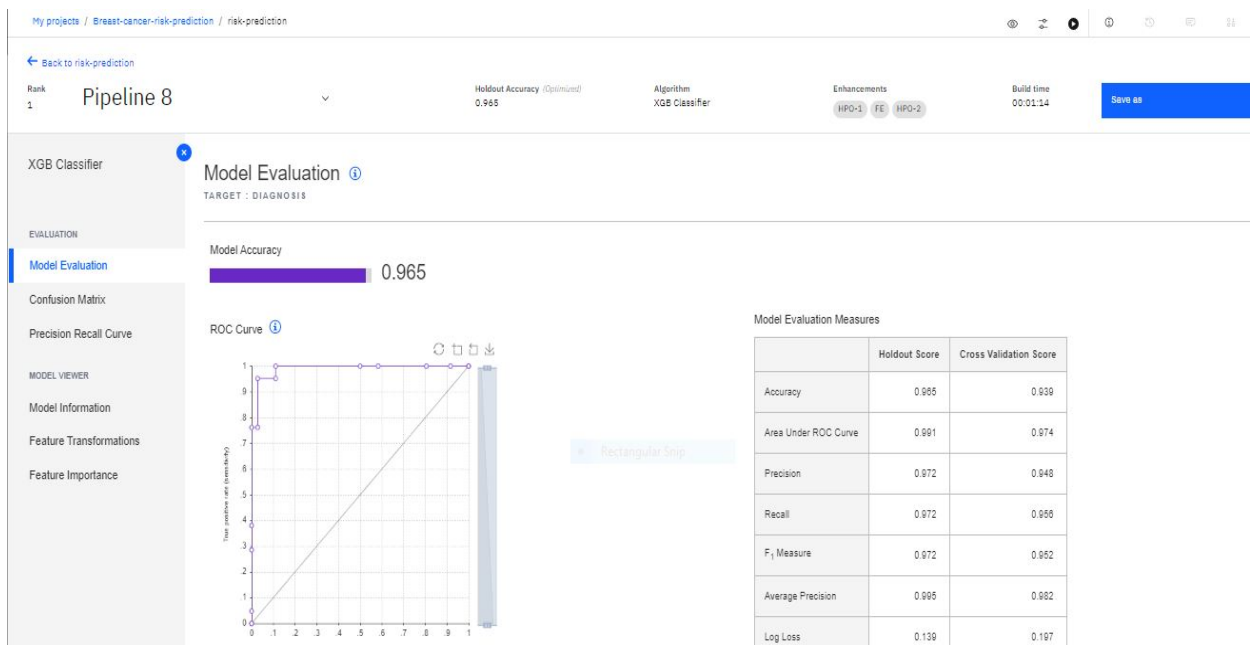
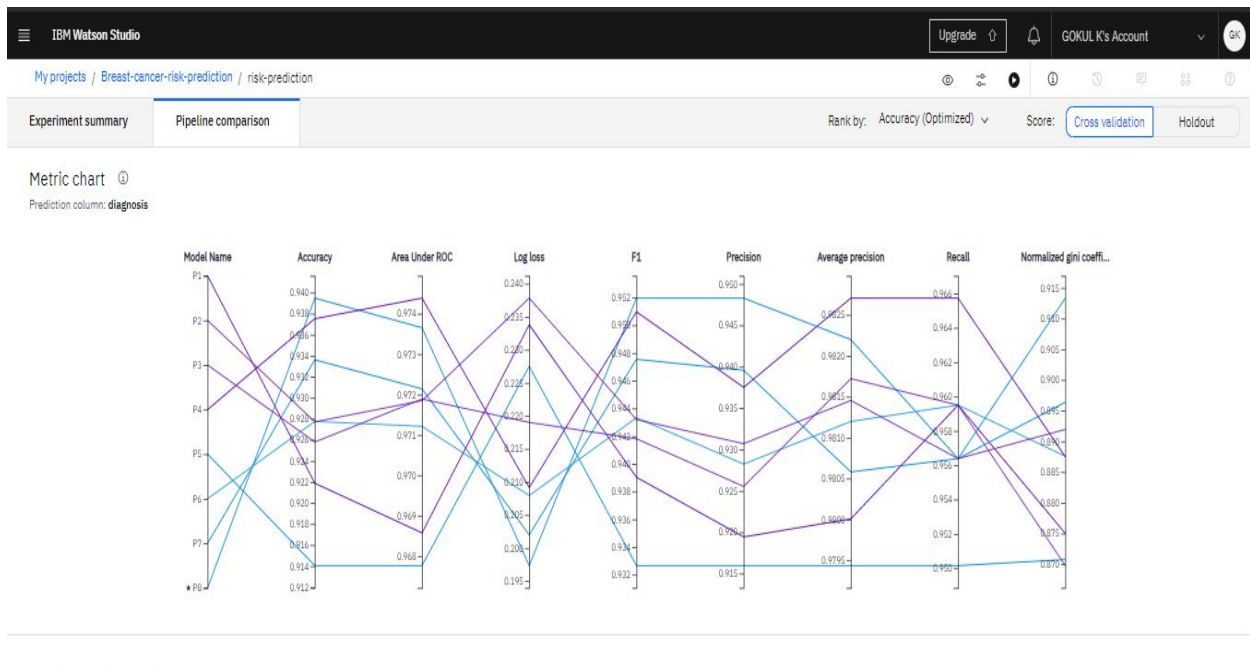


Pipeline leaderboard

Pipeline leaderboard

Rank	↑	Name	Algorithm	Accuracy (Optimized)	Enhancements	Build time	
>	★ 1	Pipeline 8	XGB Classifier	0.939	HPO-1 FE HPO-2	00:01:14	Save as
>	2	Pipeline 4	Gradient Boosting Classifier	0.938	HPO-1 FE HPO-2	00:00:13	
>	3	Pipeline 7	XGB Classifier	0.934	HPO-1 FE	00:01:16	
>	4	Pipeline 6	XGB Classifier	0.928	HPO-1	00:00:18	
>	5	Pipeline 2	Gradient Boosting Classifier	0.928	HPO-1	00:00:06	
>	6	Pipeline 3	Gradient Boosting Classifier	0.926	HPO-1 FE	00:00:46	
>	7	Pipeline 1	Gradient Boosting Classifier	0.922	None	00:00:01	
>	8	Pipeline 5	XGB Classifier	0.914	None	00:00:01	

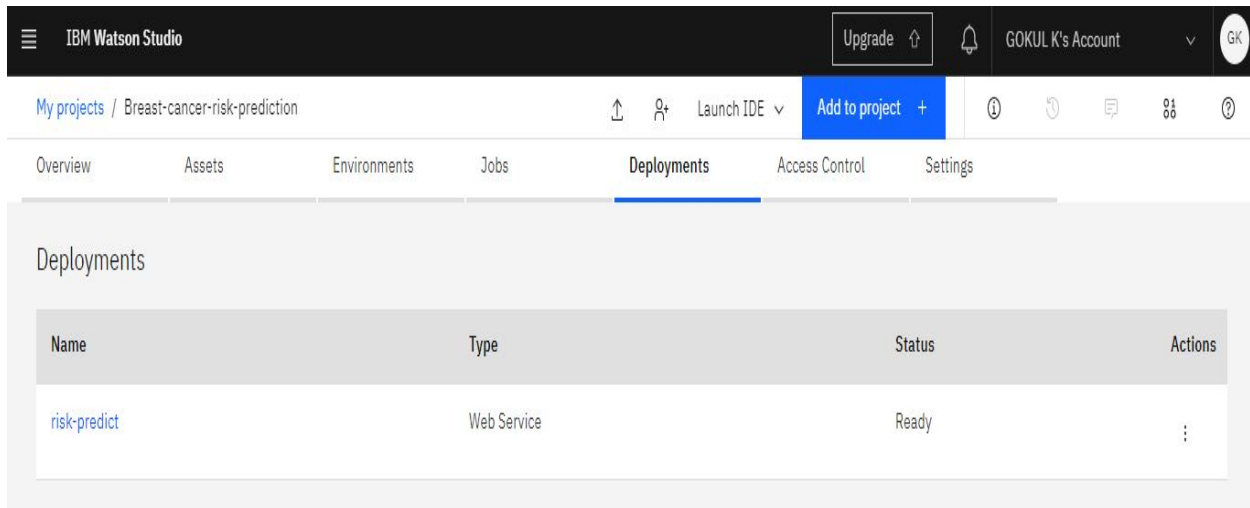
As we can see, the XGB Classifier ranks first as the best algorithm for the given dataset. It has the most less error, (i.e.) accuracy value in comparison to all the other algorithms.



As we can see, the XGB Classifier has a maximum accuracy of 96.5% in comparison to the other algorithms.

Deploy and test the model in Watson Studio:

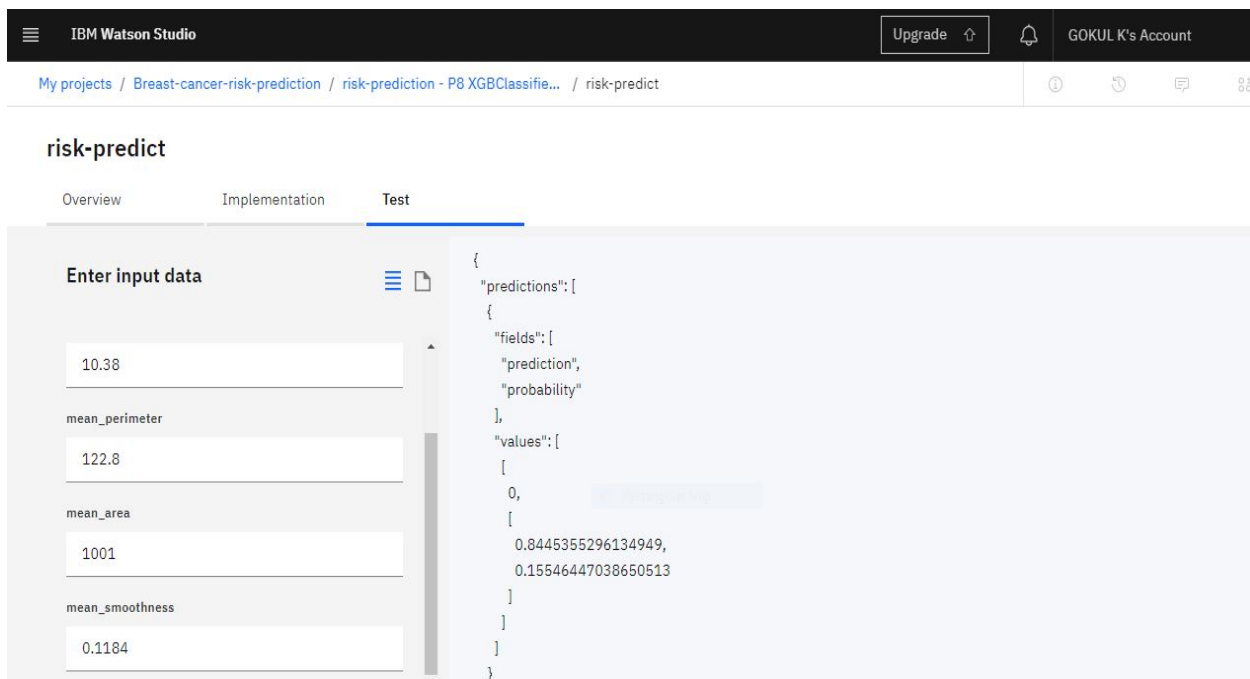
We then save this model and deploy it in the Watson studio in the name "Intern_RF".



The screenshot shows the IBM Watson Studio interface. The top navigation bar includes the IBM Watson Studio logo, an Upgrade button, a notification bell, and the user account "GOKUL K's Account". The breadcrumb trail indicates the current location: "My projects / Breast-cancer-risk-prediction". The main navigation tabs are Overview, Assets, Environments, Jobs, Deployments (selected), Access Control, and Settings. The Deployments section displays a table with the following data:

Name	Type	Status	Actions
risk-predict	Web Service	Ready	⋮

We can see that our model is successfully deployed and ready to implement and test. We need to click on the model and it will direct us to a page where we can find the model overview, implementation and test. We test our model before creating our app.



The screenshot shows the IBM Watson Studio interface for testing the "risk-predict" model. The top navigation bar is the same as the previous screenshot. The breadcrumb trail is "My projects / Breast-cancer-risk-prediction / risk-prediction - P8 XGBClassifie... / risk-predict". The main navigation tabs are Overview, Implementation, and Test (selected). The Test page has a section titled "Enter input data" with four input fields:

- 10.38
- mean_perimeter: 122.8
- mean_area: 1001
- mean_smoothness: 0.1184

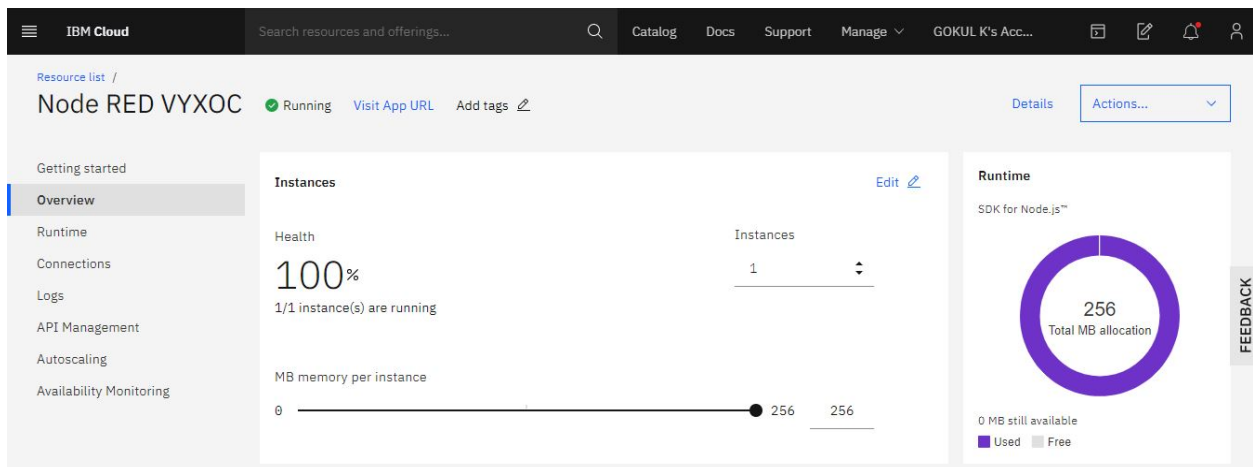
To the right of the input fields is a JSON output area showing the model's predictions:

```
{
  "predictions": [
    {
      "fields": [
        "prediction",
        "probability"
      ],
      "values": [
        0,
        0.8445355296134949,
        0.15546447038650513
      ]
    }
  ]
}
```

Step 4:- Application building:

Create a Node Red service:

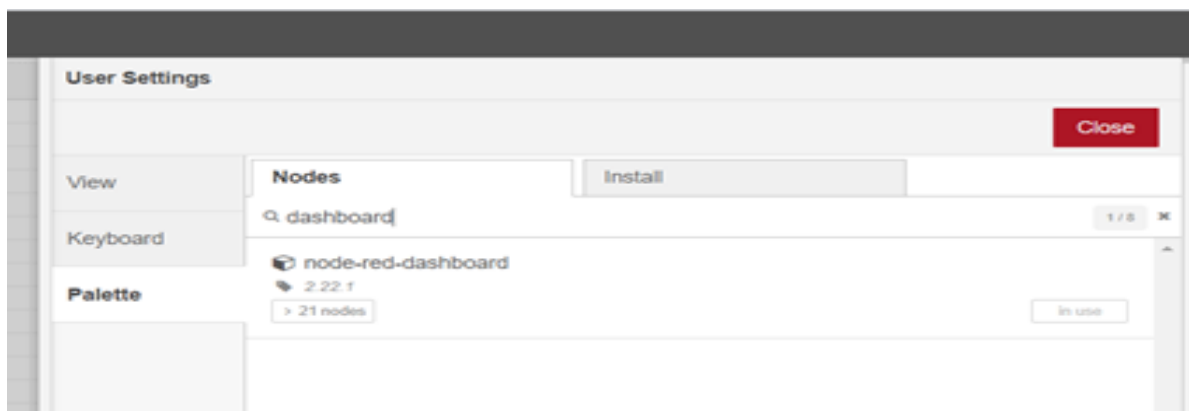
We have created a Node Red app and we can see that the app is running



Node-RED provides us a browser-based flow editor that makes it easy for us to wire together flows using the wide range of nodes in the palette. Flows can be then deployed to the runtime in a single-click. JavaScript functions can be created within the editor using a rich text editor. A built-in library allows you to save useful functions, templates or flows for re-use.

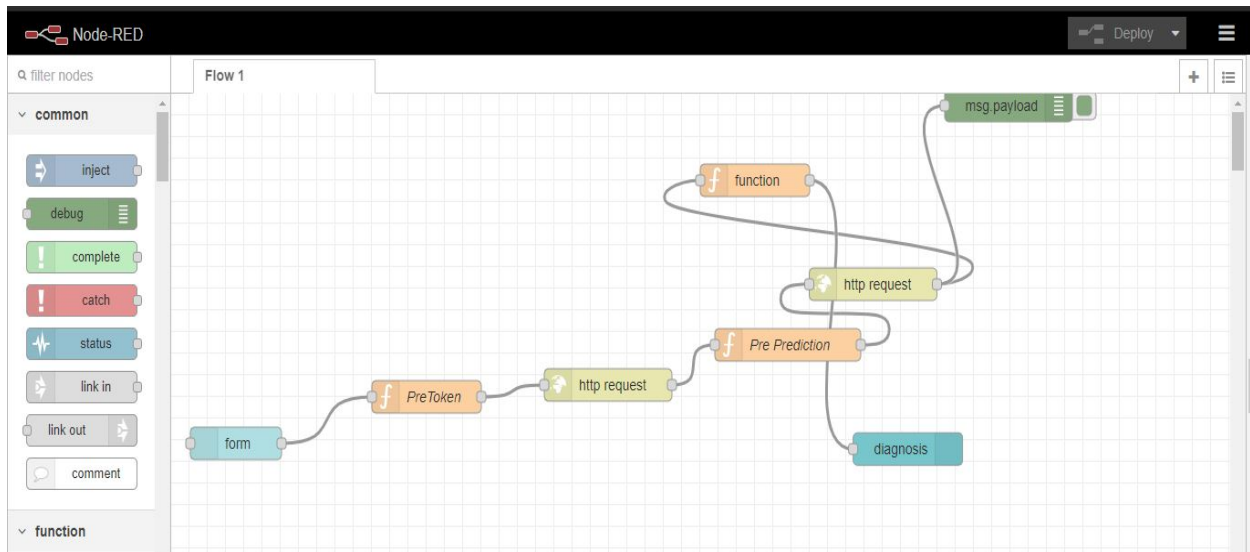
Install dashboard palette:

We have installed dashboard palette in the node red app and we use those nodes to build our app.



Building UI with Node Red:

We connect the following nodes in our node red flow.



Form node : In the form node, we give the titles and datatypes of our inputs

Edit form node

Delete Cancel Done

Properties

Group [Home] Default

Size auto

Label optional label

Label	Name	Type	Required	Rows	Remove
mean_radius	m_r	Number	<input checked="" type="checkbox"/>		
mean_texture	m_t	Number	<input checked="" type="checkbox"/>		
mean_perimeter	m_p	Number	<input checked="" type="checkbox"/>		
mean_area	m_a	Number	<input checked="" type="checkbox"/>		
mean_smoothness	m_s	Number	<input checked="" type="checkbox"/>		

+ element

Pre Token: The pre token is a function node. A JavaScript function to run against the messages being received by the node. The messages are passed in as a JavaScript object called msg. We link the api key of our deployment in this node. We get the input values from the user for the input parameters needed and then pass it on to our nextnode.

Http request: This node sends the http request and returns the response. The body of the response. The node can be configured to return the body as a string, attempt to parse it as a JSON string or leave it as a binary buffer.

Pre Prediction : The pre prediction node is also a function node. This node links our instance id to access the deployment of our model. The msg.payload in the code sends our fields as a dictionary format to our outputnode.

In our next http node we link the url of our app and in the next function node we link the path of our predicted output value from the debug message part. This helps us to view our output in our web page.

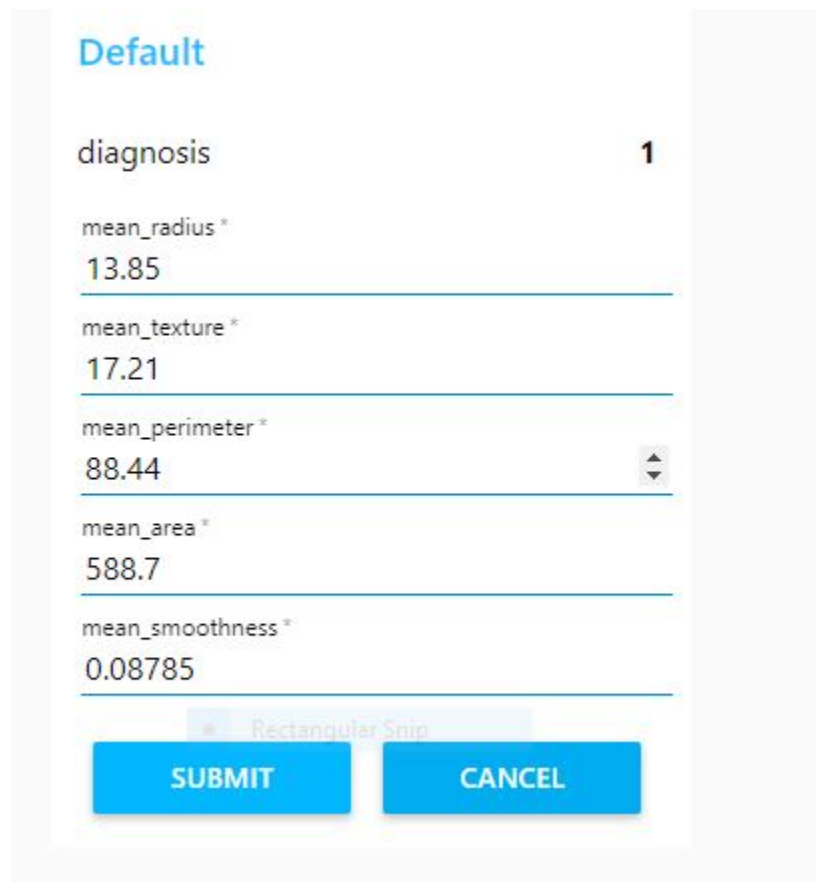


msg.payload : This node displays the value of our prediction (i.e) the depth of the snow in our case. This node displays selected message properties in the debug sidebar tab and optionally the runtime log. By default it displays

msg.payload, but can be configured to display any property, the full message or the result of a JSON data expression.

Deploy the app and run:

We then deploy our app and we load our ui web page.



The screenshot shows a web application interface for a snow depth prediction model. The interface is titled "Default" and displays a list of input features with their corresponding values. The features and values are:

- diagnosis: 1
- mean_radius*: 13.85
- mean_texture*: 17.21
- mean_perimeter*: 88.44
- mean_area*: 588.7
- mean_smoothness*: 0.08785

At the bottom of the form, there are two blue buttons labeled "SUBMIT" and "CANCEL". A small "Rectangular Snip" watermark is visible over the buttons.

We have given all the input values and our app has predicted the diagnosis to be 1. The probability of the occurrence is shown in the node red platform.

Default

diagnosis 0

mean_radius *
17.99

mean_texture *
10.38

mean_perimeter *
122.8

mean_area *
1001

mean_smoothness *
0.1184

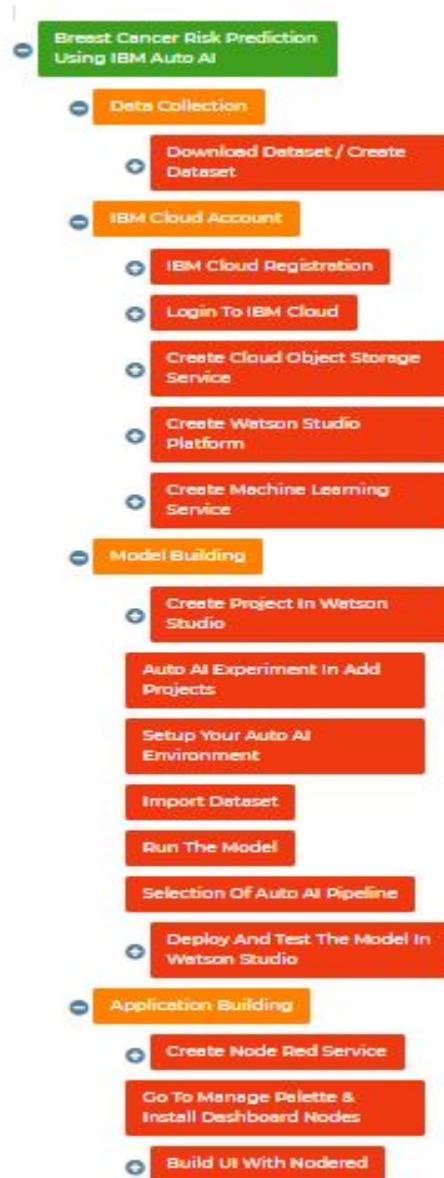
SUBMIT CANCEL

We changed our values and we can see that the diagnosis has changed to 0.

The url for our project is:

<https://node-red-vyxoc.eu-gb.mybluemix.net/ui/#!/0?socketid=Sms8gTBz9IzVrugvAAAV>

5. FLOW CHART



6. RESULT

Default

diagnosis **1**

mean_radius *
13.85

mean_texture *
17.21

mean_perimeter *
88.44

mean_area *
588.7

mean_smoothness *
0.08785

Rectangular Snip

SUBMIT CANCEL

As we can see, our app gives the output depth of the diagnosis if we enter the input values such as mean radius, mean texture, mean perimeter, mean area, mean smoothness. The value of diagnosis is 1 , which indicates the cancer is malignant (i.e) harmful. When the diagnosis value is 0. It is said to be non harmful.

7.ADVANTAGES AND DISADVANTAGES

7.1 Advantages:

- Risk prediction is an important building block of an individual.
- Effective risk prediction can improve attendance and confidence in screening programs.
- Our web app predicts the harmfulness of cancer with a good accuracy.
- Our web app is easy to access and provides us a good userinterface.
- The app gives results immediately without anydelay.

7.2 Disadvantages:

- The deep neural network overall was better than density-based models.
- Most existing breast cancer screening programs are based on mammography at similar time intervals—typically, annually or every two years—for all women. This "one size fits all" approach is not optimized for cancer detection on an individual level and may hamper the effectiveness of screening programs.

8.APPLICATIONS

- Breast cancer risk assessment provides an estimation of disease risk that can be used to guide management for women at all levels of risk.
- In today's world, breast cancer is one of the most widespread causes of death in women. According to an estimation, approximately 40,920 women would die in 2018 just because of breast cancer, which is a highly alarming number. Such alarming numbers could be reduced if the cancer is diagnosed at an early stage.
- With the advent of technology, making such predictions has become an easier task.

9.CONCLUSION

Worldwide, breast cancer is the most common type of cancer in women and the second highest in terms of mortality rates. Diagnosis of breast cancer is performed when an abnormal lump is found (from self-examination or x-ray) or a tiny speck of calcium is seen (on an x-ray). After a suspicious lump is found, the doctor will conduct a diagnosis to determine whether it is cancerous and, if so, whether it has spread to other parts of the body. Our Prediciton app helps the doctor to predict the patient conditon in an instance, at short period of time the app can predict the stage of cancer(malginant or benign). Using watson auto ai the machine learning model is built via testing the dataset with different algorithms and chosen the best model, in our case XGB Classifier is choosen the best model as it has the more accuracy compared to other models. Later the model is deployed in node red platform for creating user interface which is easy to access and if the doctor has internet connection he/she can give the input values of a patient and get the output themselves.

10.FUTURE SCOPE

As for now I have used the machine learning algorithms for predicting the cancer stage, in later period using deeplearning techniques the model can be improved and using computer vision maybe we can scan the image of cancer and process it for further research regarding the cancer. App also can be developed for patients for testing themselves with help of family doctor. The main aim is to reduce the graph of number of patients affected due to this cancer. In future there will be solutions found using the machine learning techniques regarding the origin of the cancer or how to prevent a patient get affected by cancer,etc.

11.BIBILOGRAPHY

- <https://www.kaggle.com/merishnasuwal/breast-cancer-prediction?>
- <https://www.msmanuals.com/home/women-s-health-issues/breast-disorders/breast-cancer>
- <https://medium.com/analytics-vidhya/detecting-breast-cancer-using-machine-learning-ab23e719f7fa>

