Breast Cancer Risk Prediction using IBM Auto AI

Introduction :-

Overview:

This project is about predicting the risk of Breast Cancer using IBM Watson Auto AI Experiment.This comes under the category of "Machine Learning".The main objective of project is to predict if diagnosis is required for Breast Cancer Patient or not based on the previously present data.

Purpose:

Worldwide, breast cancer is the most common type of cancer inwomen 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. . 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.

Literature Survey :-

Existing Problem:

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.

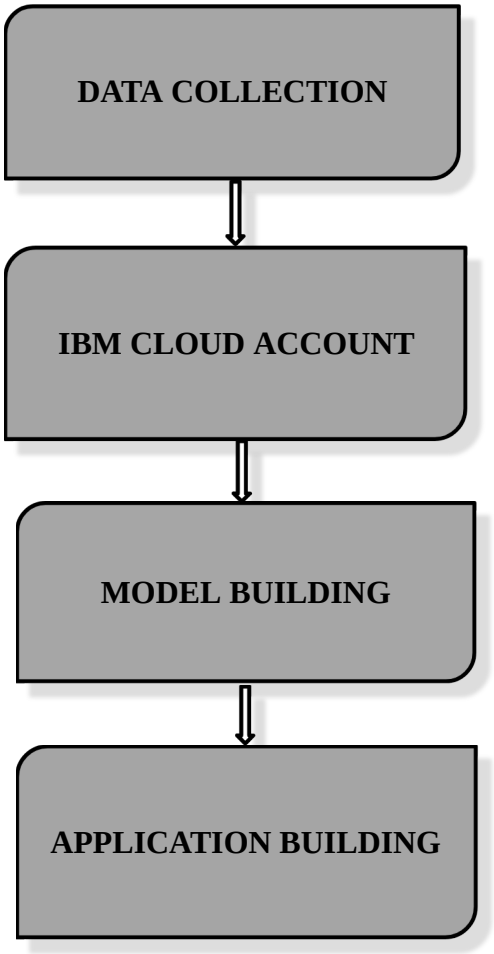
Proposed Solution:

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.

➢ Theoratical survey :-

➤Block Diagram:



➤Hardware/Software Designing:

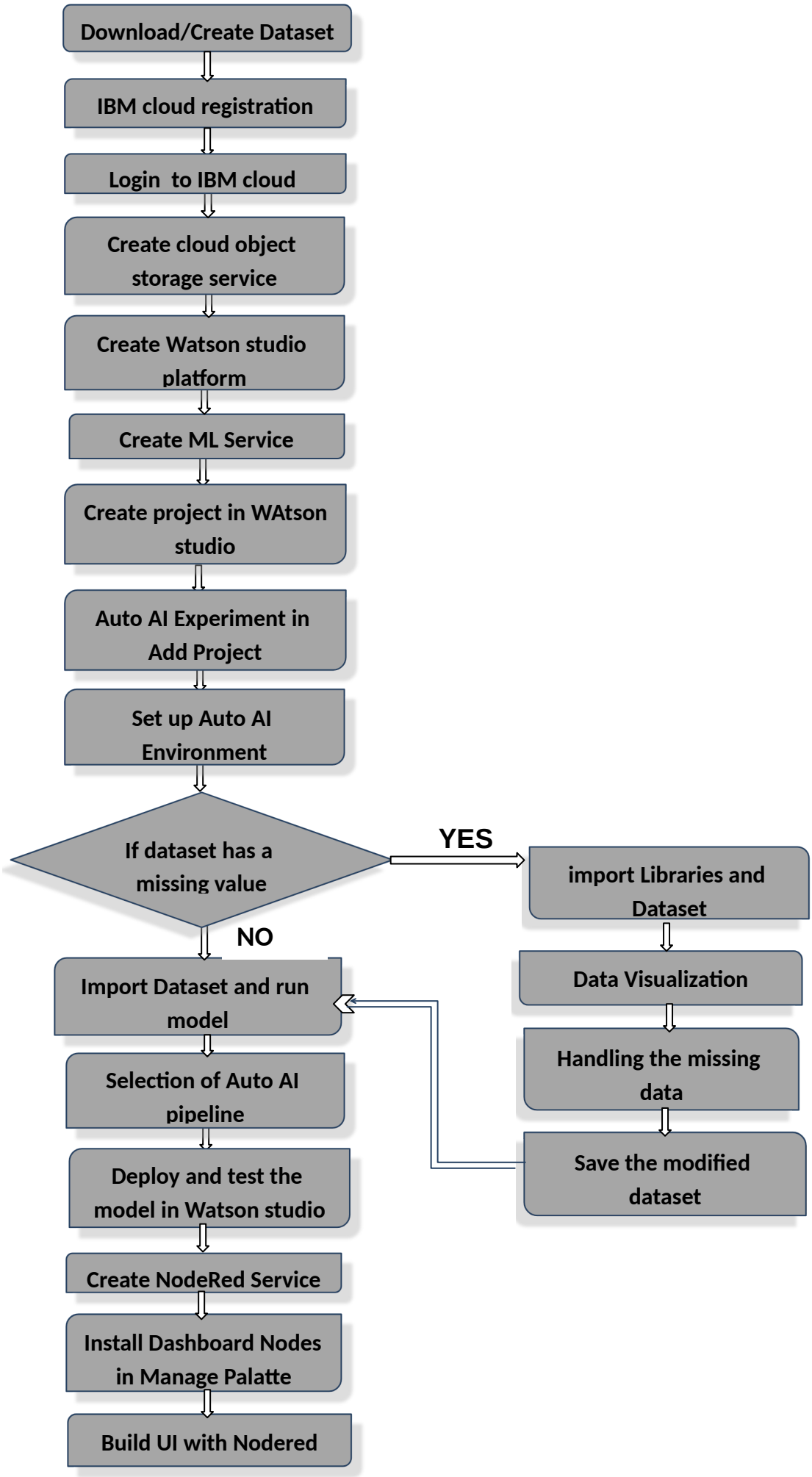
- Jupyter Notebook for Data Analysis
- IBM Cloud
- IBM Watson
- Auto AI Experiment
- NodeRed

➢ Experimental investigations :-

Upon analysing the dataset,we understand that -

- Since the dataset contains independent input parameters and an output is to be predicted,hence it comes under Supervised Learning.
- And since the output is categorical and not continuous,classification algorithm is to be used,but since we are using Auto AI here,any classification can not be applied externally.
- As the dataset contains all the important inputs to predict the output,there is no need to remove any input rows.

Flow chart :-


Download/Create Dataset
IBM cloud registration
Login to IBM cloud
Create cloud object storage service
Create Watson studio platform
Create ML Service
Create project in WAtson studio
Auto AI Experiment in Add Project
Set up Auto AI Environment
If dataset has a missing value
YES
NO
import Libraries and Dataset
Data Visualization
Handling the missing data
Save the modified dataset
Import Dataset and run model
Selection of Auto AI pipeline
Deploy and test the model in Watson studio
Create NodeRed Service
Install Dashboard Nodes in Manage Palatte
Build UI with Nodered

Result :-

The model is succesfully trained and deployed using Auto AI Experiment and Node-red Service.Now "Breast Cancer Risk Prediction" Machine Learning model can predict the risk of breast cancer.

Advantages and Disasvantages :-

➤ Advantages:

- This model helps in the prediction of diagnosis possibility for a patient with breast cancer by which we can easily stop the spread of cancer to different parts of the body.
- Improves the diagnostic accuracy.

➤ Disadvantages:

- Sometimes slow due to network glitch.
- Model cannot be edited yet in a more granular way.
- Limited in number of models it can choose from.
- Data must be good going in'
- If size of the dataset is large then the model takes a very large time for building.

Applications :-

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.This can be used in various hospitals to predict the breast cancer risk as early as possible due to which the deaths may be reduced.

Conclusion :-

Therefore the "Breast Cancer Risk Prediction" , Machine Learning model is created and the purpose of project is fulfilled.

Future Scope :-

This model is mainly used for predicting breast cancer diagnostic accuracy,by upgrading the dataset the results may be more precisely accurate.

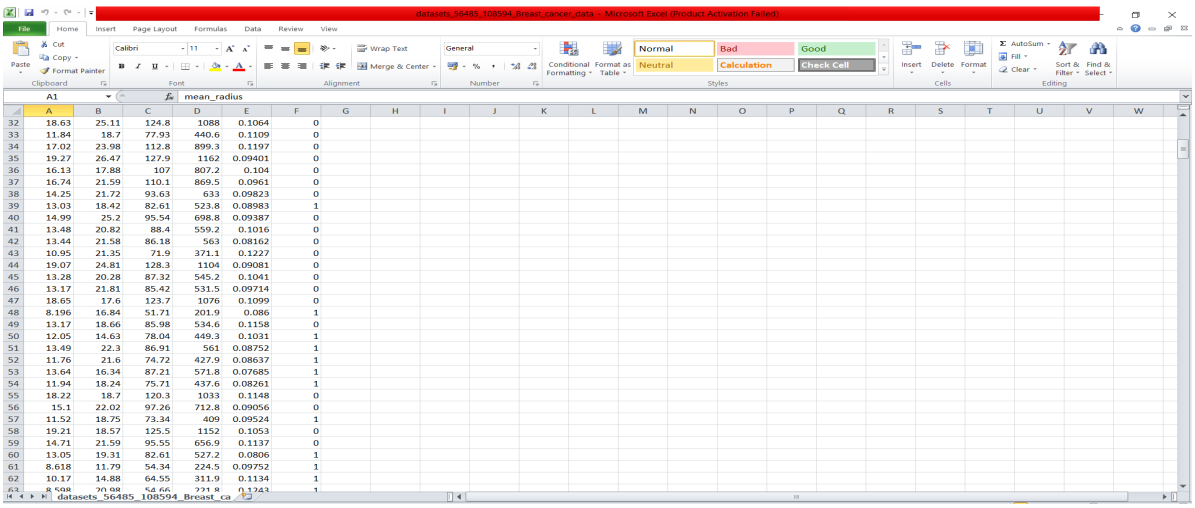
Bibilography :-

- github.com
- kaggle.com(for downloading dataset)
- smartinternz.com(for attaching links and pdfs in the platform)
- thesmartbridge.com

➤ Appendix :-

ScreenShots:

➤ Data Collection:



	mean_radius	mean_texture	mean_perimeter	mean_area	mean_smoothness	diagnosis
32	18.63	25.11	124.8	1088	0.1064	0
33	11.84	18.7	77.93	440.6	0.1109	0
34	17.02	23.98	112.8	899.3	0.1197	0
35	19.27	26.47	127.9	1162	0.09401	0
36	16.13	17.88	107	807.2	0.104	0
37	16.74	21.59	110.1	869.5	0.0961	0
38	14.25	21.72	93.63	633	0.09823	0
39	13.03	18.42	82.61	523.8	0.08983	1
40	14.99	25.2	95.54	698.8	0.09387	0
41	13.48	20.82	88.4	559.2	0.1016	0
42	13.44	21.58	86.18	563	0.08162	0
43	10.95	21.35	71.9	371.1	0.1227	0
44	19.07	24.81	128.3	1104	0.09081	0
45	13.28	20.28	87.32	545.2	0.1041	0
46	13.17	21.81	85.42	531.5	0.09714	0
47	18.65	17.6	123.7	1076	0.1099	0
48	8.196	16.84	51.71	201.9	0.086	1
49	13.17	18.66	85.98	534.6	0.1158	0
50	12.05	14.63	78.04	449.3	0.1031	1
51	13.49	22.3	86.91	561	0.08752	1
52	11.76	21.6	74.72	427.9	0.08637	1
53	13.64	16.34	87.21	571.8	0.07685	1
54	11.94	18.24	75.71	437.6	0.08261	1
55	18.22	18.7	120.3	1033	0.1148	0
56	15.1	22.02	97.26	712.8	0.09056	0
57	11.52	18.75	73.34	409	0.09524	1
58	19.21	18.57	125.5	1152	0.1053	0
59	14.71	21.59	95.55	656.9	0.1137	0
60	13.05	19.31	82.61	527.2	0.0806	1
61	8.618	11.79	54.34	224.5	0.09752	1
62	10.17	14.88	64.55	311.9	0.1134	1
63	8.598	20.98	54.66	221.8	0.1243	1

➤ Importing Libraries and Dataset:

```
In [1]: import numpy as np
        import pandas as pd
        import matplotlib.pyplot as plt
```

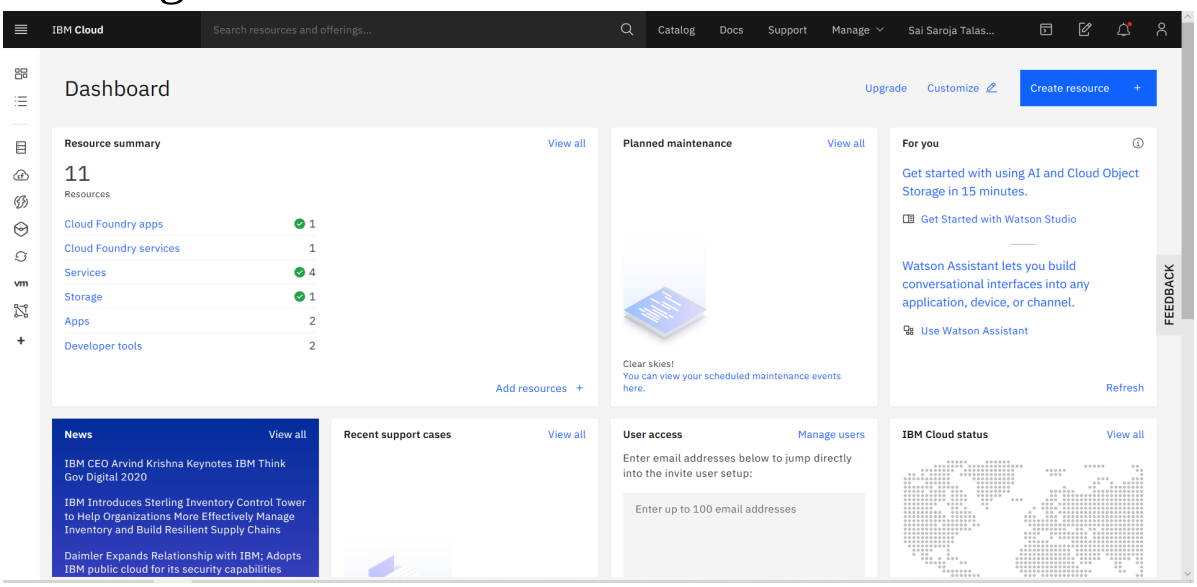
```
In [3]: dataset=pd.read_csv('datasets_56485_108594_Breast_cancer_data.csv')
        dataset
```

Out[3]:

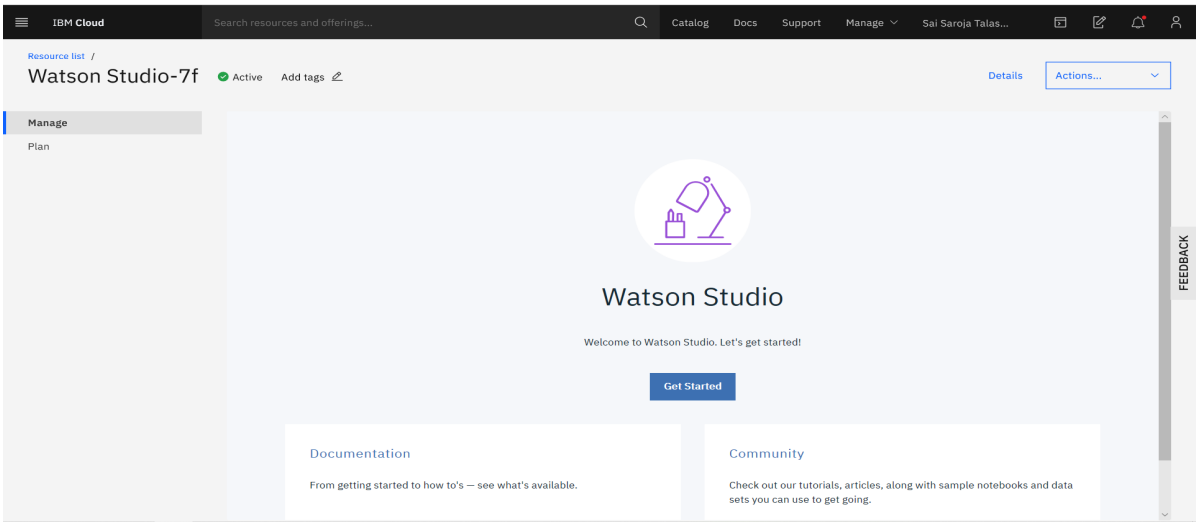
	mean_radius	mean_texture	mean_perimeter	mean_area	mean_smoothness	diagnosis
0	17.99	10.38	122.80	1001.0	0.11840	0
1	20.57	17.77	132.90	1326.0	0.08474	0
2	19.69	21.25	130.00	1203.0	0.10960	0
3	11.42	20.38	77.58	386.1	0.14250	0
4	20.29	14.34	135.10	1297.0	0.10030	0
...	...	...	...	...	...	...
564	21.56	22.39	142.00	1479.0	0.11100	0
565	20.13	28.25	131.20	1261.0	0.09780	0
566	16.60	28.08	108.30	858.1	0.08455	0
567	20.60	29.33	140.10	1265.0	0.11780	0
568	7.76	24.54	47.92	181.0	0.05263	1

569 rows × 6 columns

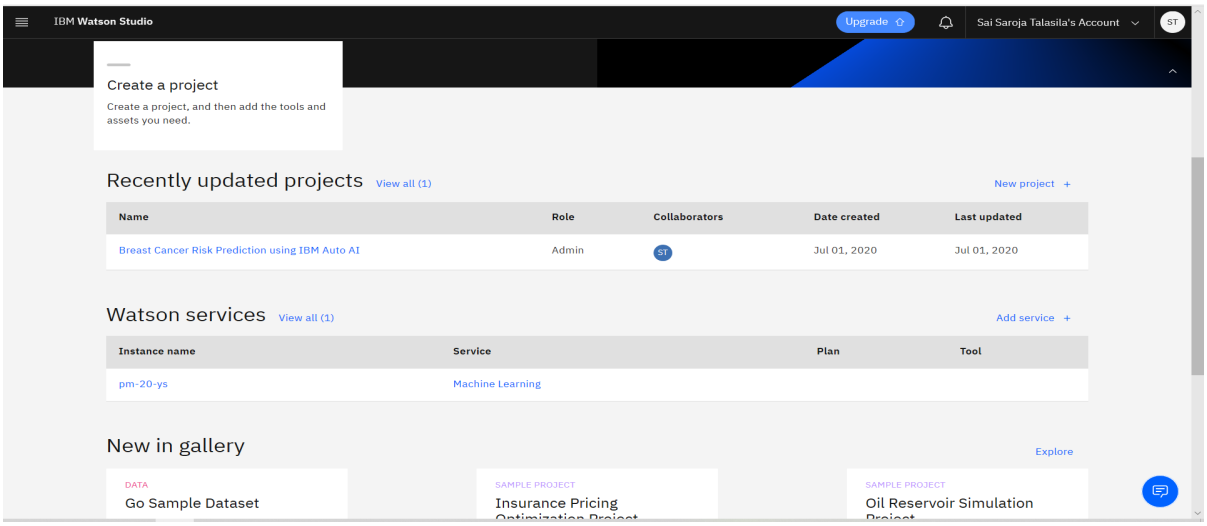
➤ Creating IBM Account:



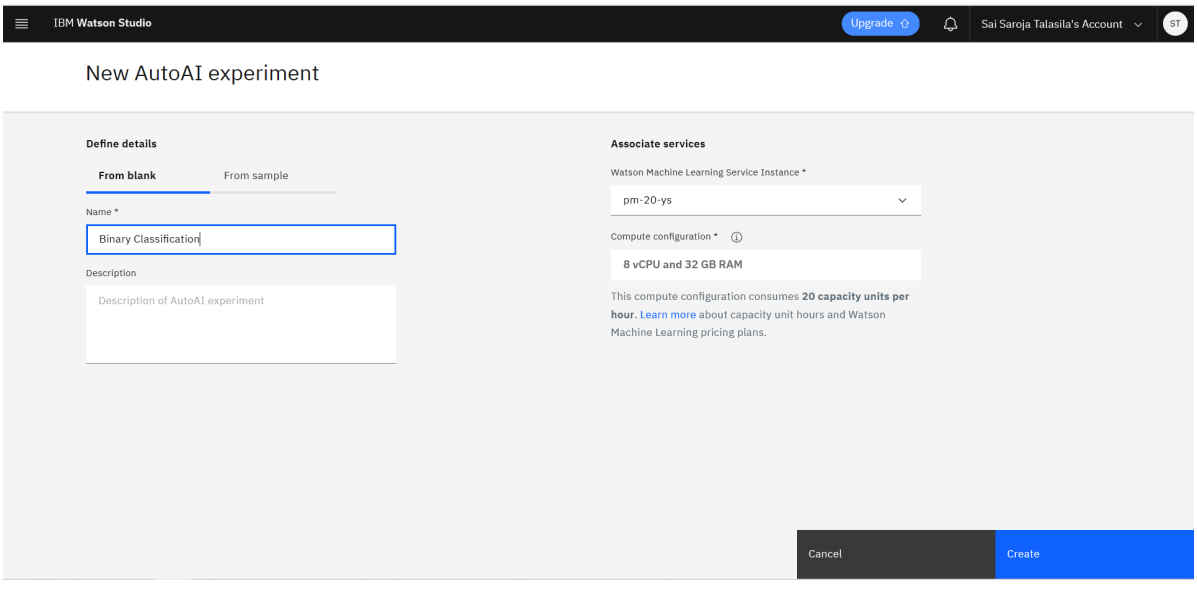
➤Creating Watson Studio Platform:



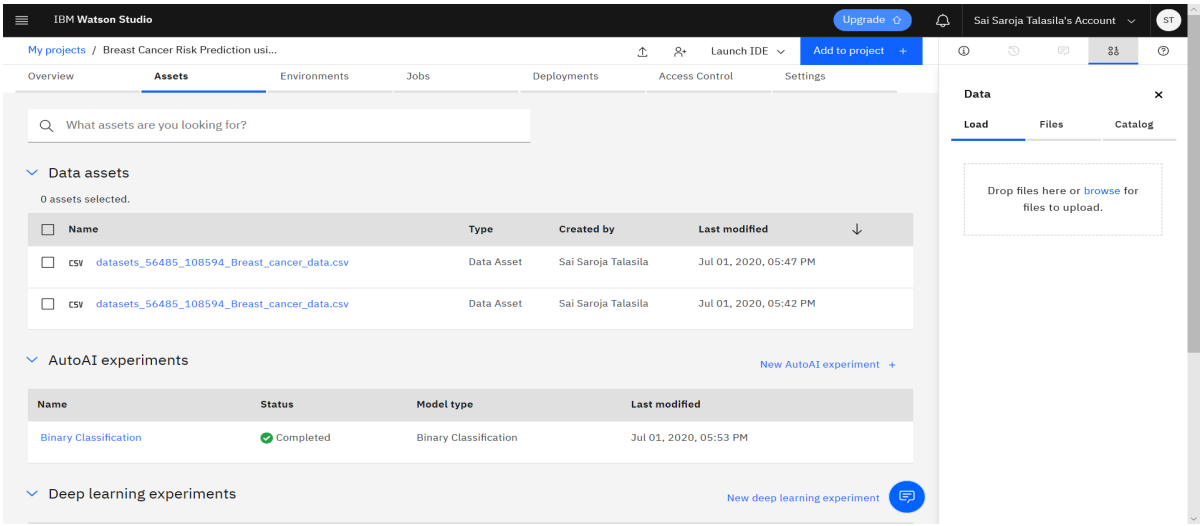
➤Creating Project and Watson Instance Service:



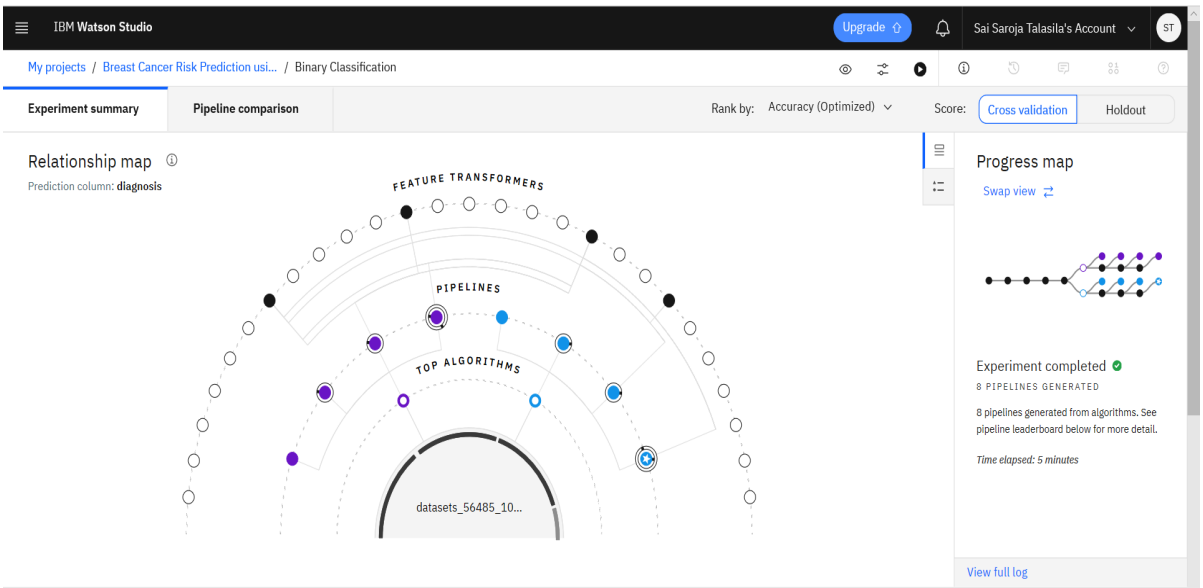
➤Auto AI Setup:



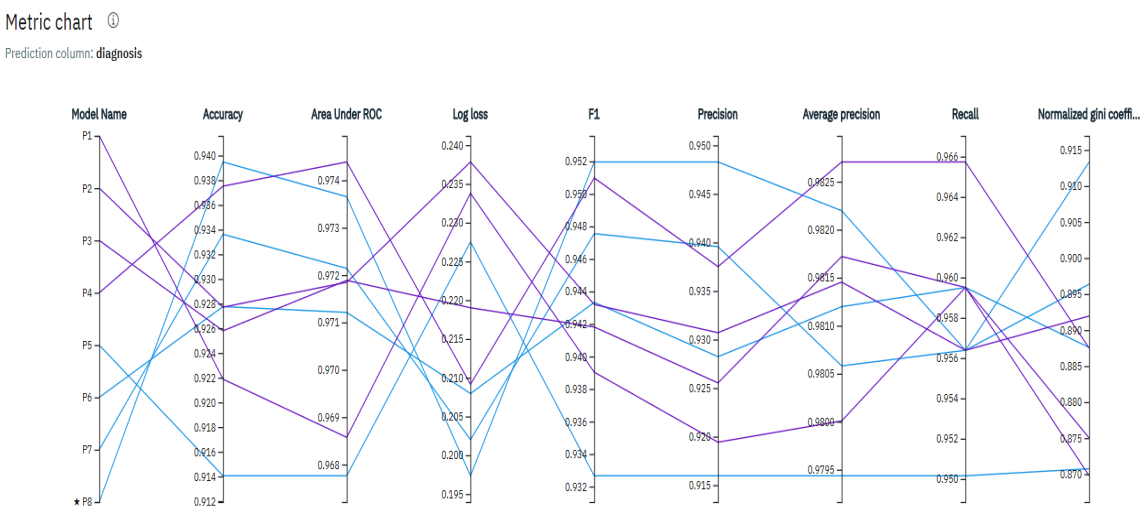
➤Importing Dataset:



➤Running Experiment:



➤Metric Analysis:

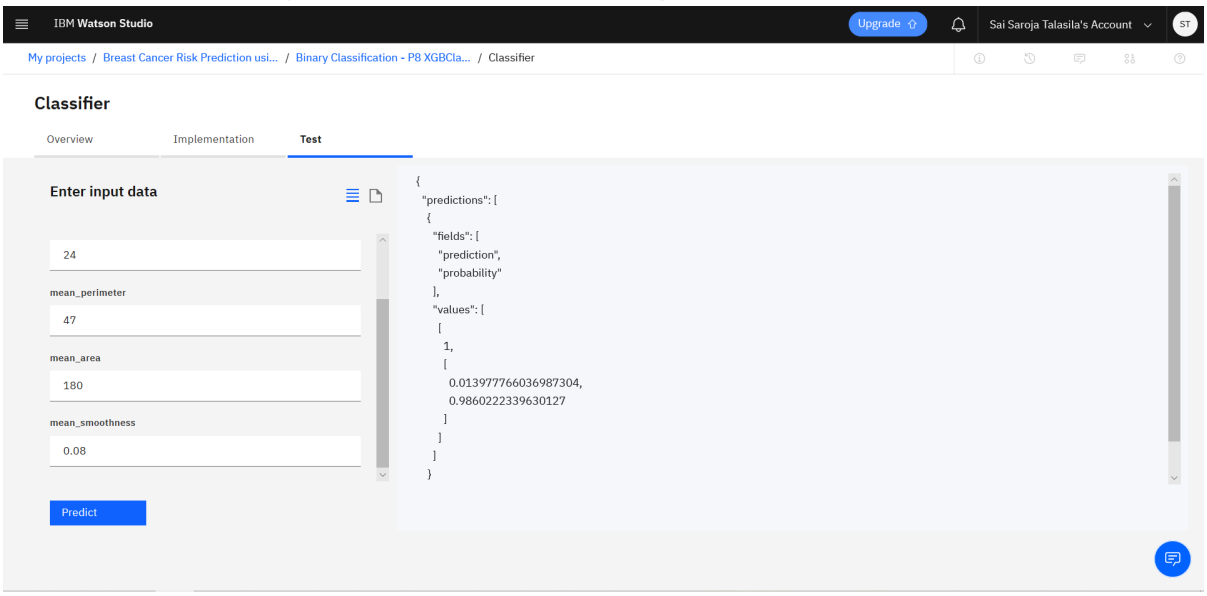


➤Selection of Auto AI Pipeline:

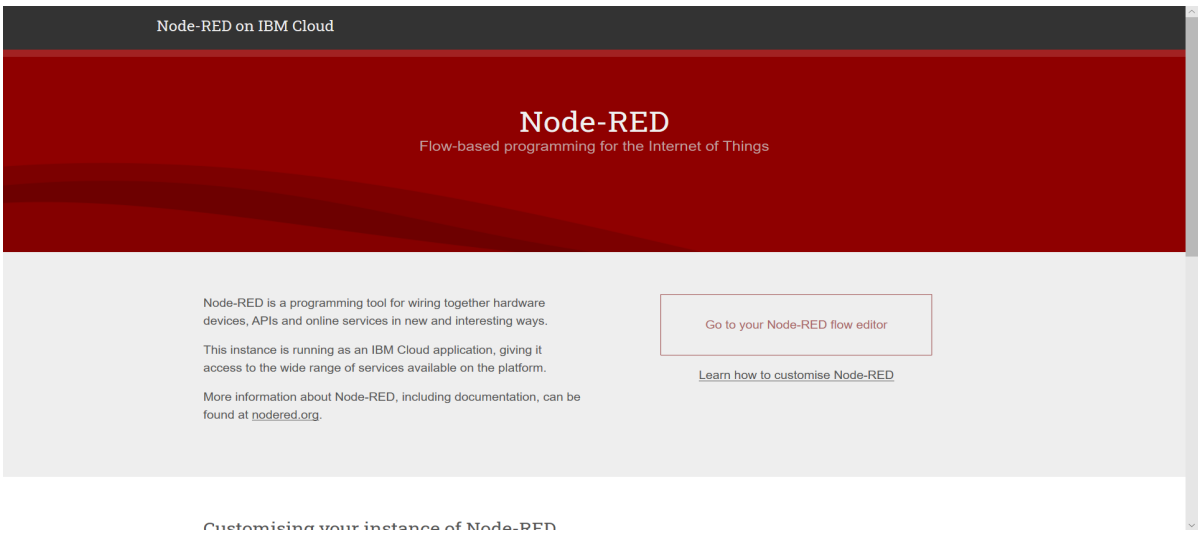
Pipeline leaderboard

	Rank	Name	Algorithm	Accuracy (Optimized)	Enhancements	Build time
>	★ 1	Pipeline 8	XGB Classifier	0.939	HPO-1 FE HPO-2	00:00:51
>	2	Pipeline 4	Gradient Boosting Classifier	0.938	HPO-1 FE HPO-2	00:00:09
>	3	Pipeline 7	XGB Classifier	0.934	HPO-1 FE	00:00:56
>	4	Pipeline 6	XGB Classifier	0.928	HPO-1	00:00:13
>	5	Pipeline 2	Gradient Boosting Classifier	0.928	HPO-1	00:00:04
>	6	Pipeline 3	Gradient Boosting Classifier	0.926	HPO-1 FE	00:00:32
>	7	Pipeline 1	Gradient Boosting Classifier	0.922	*None*	00:00:01
>	8	Pipeline 5	XGB Classifier	0.914	*None*	00:00:01

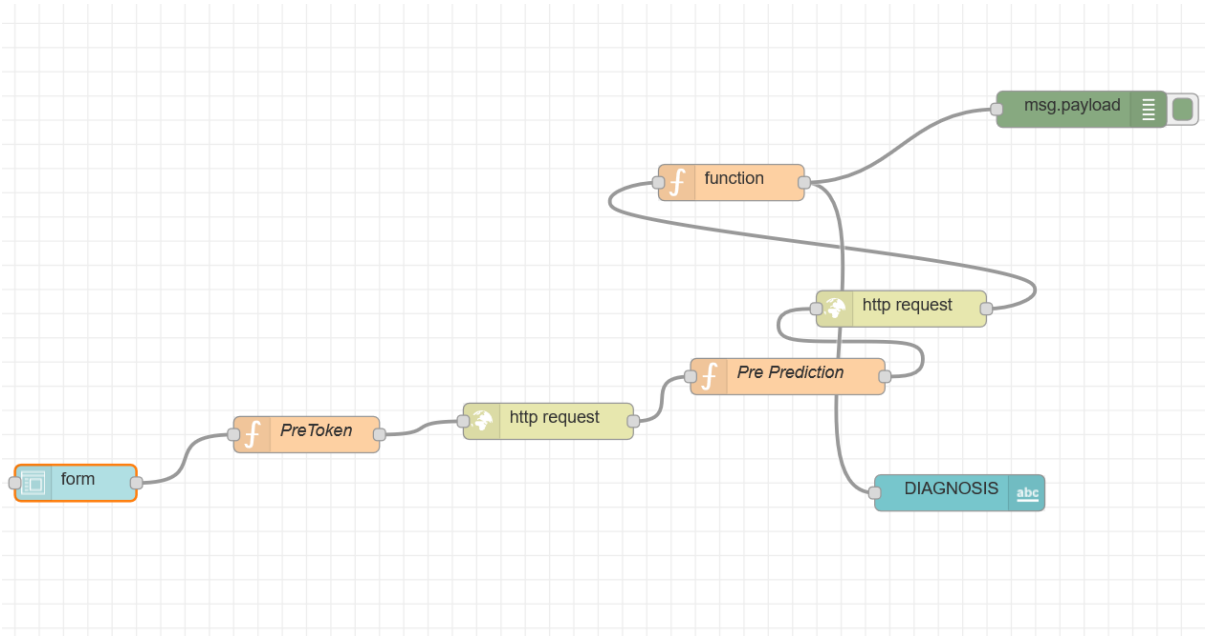
➤Model Deployment and Testing:



➤Node-Red Service creation:



➤Building UI(User Interface) with Node-Red:



➤Output in Node-Red App: