1. Introduction -

1.1 Overview -

Breast Cancer is the most leading malignancy affecting 2.1 million women each year which leads to greatest number of deaths among women. Early treatment not only helps to cure cancer but also helps in prevention of its recurrence. And hence this system mainly focuses on prediction of breast cancer where it uses different machine learning algorithms for creating models like decision tree, logistic regression, random forest which are applied on pre-processed data which suspects greater accuracy for prediction. Amongst all the models, Random Forest Classification leads to best accuracy with 98.6%. These techniques are coded in python and uses numpy, pandas, seaborn libraries.

1.2 Purpose -

According to World health organization, Breast cancer is the most frequent cancer among women and it is the second dangerous cancer after lung cancer. In 2018, from the research it is estimated that total 627,000 women lost their life due to breast cancer that is 15% of all cancer deaths among women. In case of any symptom, people visit to oncologist. Doctors can easily identify breast cancer by using Breast ultrasound, Diagnostic mammogram, Magnetic resonance imaging (MRI), Biopsy. Based on these test results, doctor may recommend further tests or therapy. Early detection is very crucial in breast cancer. If chances of cancer are predicted at early stage then survivability chances of patient may increase. An alternate way to identify breast cancer is using machine learning algorithms for prediction of abnormal tumor. Thus, the research is carried out for the proper diagnosis and categorization of patients into malignant and benign groups.

2. Literature Survey -

2.1 Existing Problem -

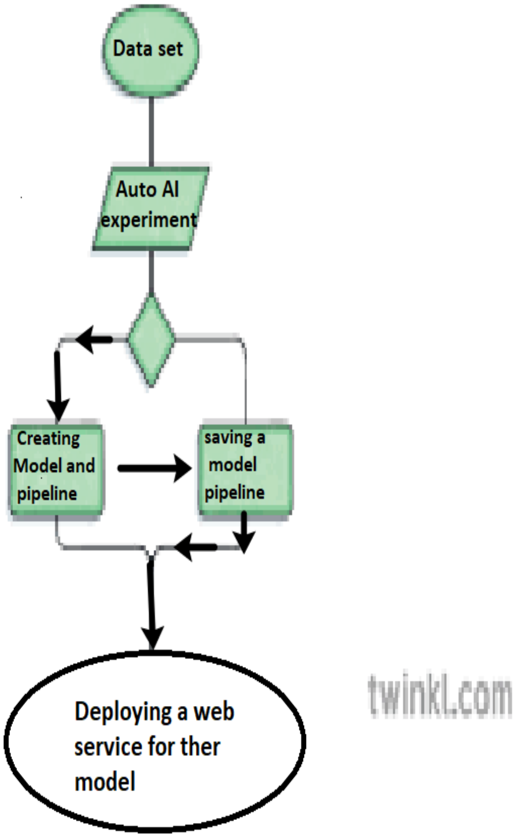
Over the past decades, a continuous evolution related to cancer research has been performed. Scientists applied different methods, such as screening in early stage, in order to find types of cancer before they cause symptoms. Moreover, they have developed new strategies for the early prediction of cancer treatment outcome. With the advent of new technologies in the field of medicine, large amounts of cancer data have been collected and are available to the medical research community. However, the accurate prediction of a disease outcome is one of the most interesting g and challenging tasks for physicians. As a result, ML methods have become a popular tool for medical researchers. These techniques can discover and identify patterns and relattionships between them, from complex datasets, while they are able to effectively predict future outcomes of a cancer type.

2.2 Proposed Solution -

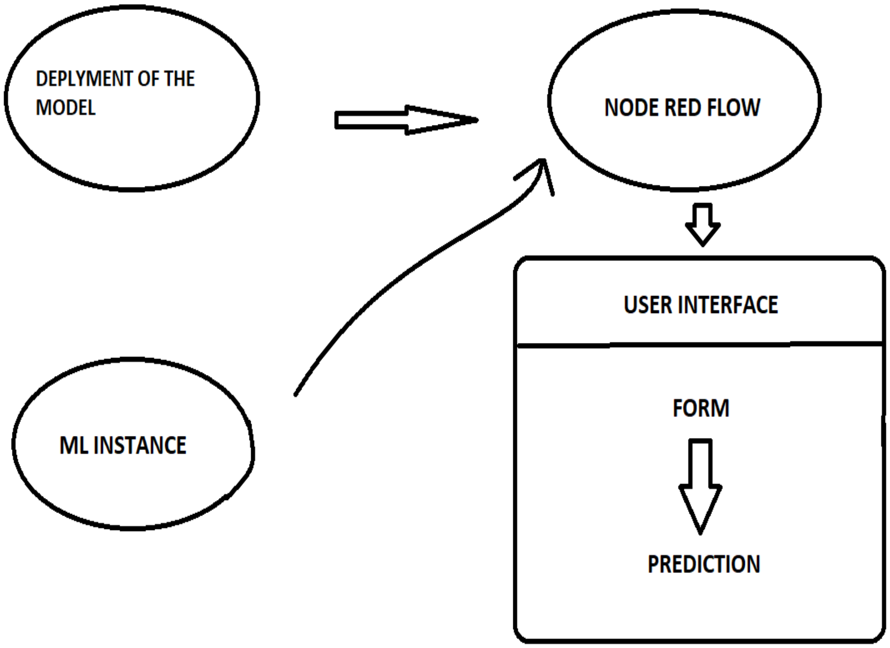
We obtained the breast cancer dataset of Wisconsin Breast Cancer diagnosis dataset and used IBM Auto AI as the platform for the purpose of development of model. The model is been trained using Auto AI service in IBMwatson cloud and that can be deployed in an application such as web or mobile application using Node-RED application. The AutoAI graphical tool in Watson Studio automatically analyzes your data and generates candidate model pipelines customized for your predictive modeling problem. These model pipelines are created over time as AutoAI analyzes your dataset and discovers data transformations, algorithms, and parameter settings that work best for your problem setting. Results are displayed on a leaderboard, showing the automatically generated model pipelines ranked according to your problem optimization objective.

3.THEORETICAL ANALYSIS

3.1 Block Diagram


Data set
Auto AI experiment
Creating Model and pipeline
saving a model pipeline
Deploying a web service for ther model
twinkl.com

3.2 Hardware/ Software Designing -



The AutoAI process follows this sequence to build candidate pipelines:

- Data pre-processing
- Automated model selection
- Automated feature engineering
- Hyperparameter optimization

Data pre-processing -

Most data sets contain different data formats and missing values, but standard machine learning algorithms work with numbers and no missing values. AutoAI applies various algorithms, or estimators, to analyze, clean, and prepare your raw data for machine learning. It automatically detects and categorizes features based on data type, such as categorical or numerical. Depending on the categorization, it uses hyper-parameter optimization to determine the best combination of strategies for missing value imputation, feature encoding, and feature scaling for your data. Automated model selection - The next step is automated model selection that matches your data. AutoAI uses a novel approach that enables testing and ranking ● Create a Node-RED starter application running in IBM Cloud, create machine learning instances of the Watson services, and connect the services to your Node-Red app. ● Launch and configure the Node-RED visual programming editor. ● Install additionalcandidate algorithms against small subsets of the data, gradually increasing the size of the subset for the most promising algorithms to arrive at the best match. This approach saves time without sacrificing performance. It enables ranking a large number of candidate algorithms and selecting the best match for the data.

Automated feature engineering -

Feature engineering attempts to transform the raw data into the combination of features that best represents the problem to achieve the most accurate prediction. AutoAI uses a novel approach that explores various feature construction choices in a structured, non-exhaustive manner, while progressively maximizing model accuracy using reinforcement learning. This results in an optimized sequence of transformations for the data

that best match the algorithms of the model selection step.
Hyperparameter optimization -
Finally, a hyper-parameter optimization step refines the best performing model pipelines. AutoAI uses a novel hyper-parameter optimization algorithm optimized for costly function evaluations such as model training and scoring that are typical in machine learning. This approach enables fast convergence to a good solution despite long evaluation times of each iteration.
Connecting to Node-Red - Node-RED nodes and create flows that use the Watson services to create the Breast cancer prediction model api

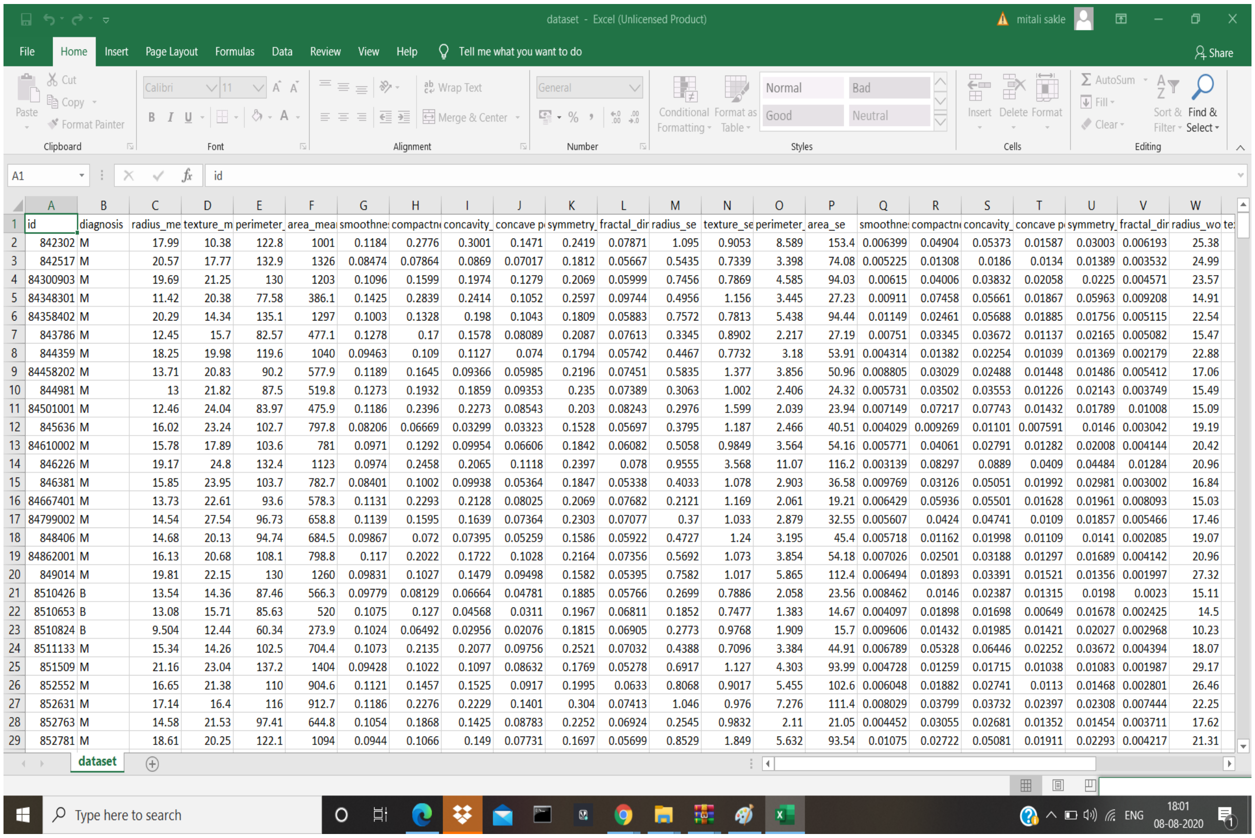
4. Experimental Investigation -

The Breast Cancer (Wisconsin) Diagnosis dataset contains the diagnosis and a set of 30 features describing the characteristics of the cell nuclei present in the digitized image of a of a fine needle aspirate (FNA) of a breast mass. Eight real-valued features are computed for each cell nucleus:

- *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 (perimeter^2 / area - 1.0);*
- *concavity (severity of concave portions of the contour);*

- *concave points (number of concave portions of the contour);*

The mean, standard error (SE) and “worst” or largest (mean of the three largest values) of these features were computed for each image, resulting in 24 features. We will analyze the features to understand the predictive value for diagnosis. We will then create models using IBM Auto AI and use the models to predict the diagnosis.

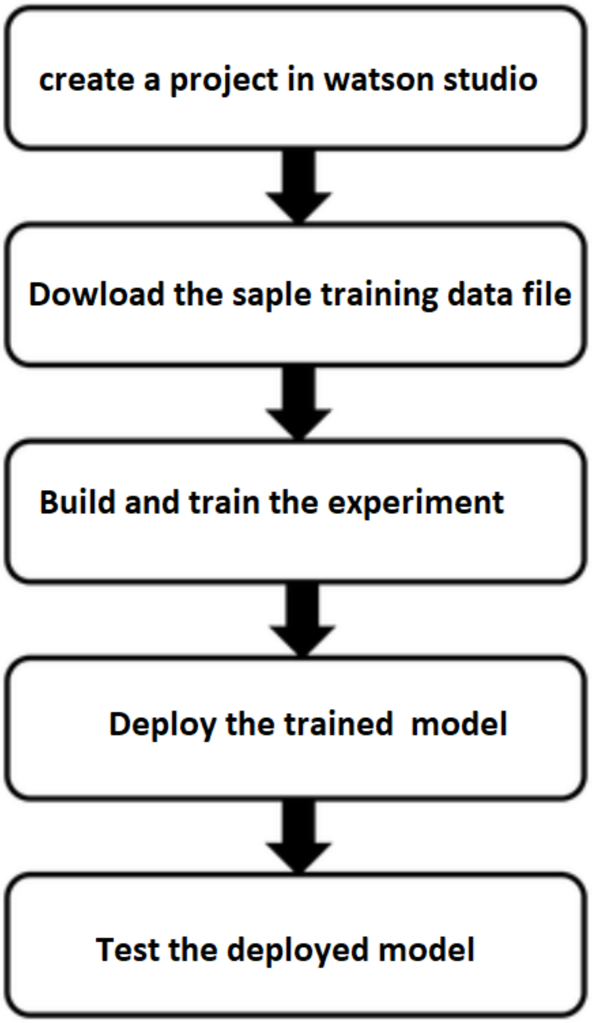


	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	
1	id	diagnosis	radius_me	texture_m	perimeter_	area_mea	smoothne:	compactn	concavity_	concave p	symmetry_	fractal_dir	radius_se	texture_se	perimeter_	area_se	smoothne:	compactn	concavity_	concave p	symmetry_	fractal_dir	radius_wo	te:
2	842302	M	17.99	10.38	122.8	1001	0.1184	0.2776	0.3001	0.1471	0.2419	0.07871	1.095	0.9053	8.589	153.4	0.006399	0.04904	0.05373	0.01587	0.03003	0.006193	25.38	
3	842517	M	20.57	17.77	132.9	1326	0.08474	0.07864	0.0869	0.07017	0.1812	0.05667	0.5435	0.7339	3.398	74.08	0.005225	0.01308	0.0186	0.0134	0.01389	0.003532	24.99	
4	84300903	M	19.69	21.25	130	1203	0.1096	0.1599	0.1974	0.1279	0.2069	0.05999	0.7456	0.7869	4.585	94.03	0.00615	0.04006	0.03832	0.02058	0.0225	0.004571	23.57	
5	84348301	M	11.42	20.38	77.58	386.1	0.1425	0.2839	0.2414	0.1052	0.2597	0.09744	0.4956	1.156	3.445	27.23	0.00911	0.07458	0.05661	0.01867	0.05963	0.009208	14.91	
6	84358402	M	20.29	14.34	135.1	1297	0.1003	0.1328	0.198	0.1043	0.1809	0.05883	0.7572	0.7813	5.438	94.44	0.01149	0.02461	0.05688	0.01885	0.01756	0.005115	22.54	
7	843786	M	12.45	15.7	82.57	477.1	0.1278	0.17	0.1578	0.08089	0.2087	0.07613	0.3345	0.8902	2.217	27.19	0.00751	0.03345	0.03672	0.01137	0.02165	0.005082	15.47	
8	844359	M	18.25	19.98	119.6	1040	0.09463	0.109	0.1127	0.074	0.1794	0.05742	0.4467	0.7732	3.18	53.91	0.004314	0.01382	0.02254	0.01039	0.01369	0.002179	22.88	
9	84458202	M	13.71	20.83	90.2	577.9	0.1189	0.1645	0.09366	0.05985	0.2196	0.07451	0.5835	1.377	3.856	50.96	0.008805	0.03029	0.02488	0.01448	0.01486	0.005412	17.06	
10	844981	M	13	21.82	87.5	519.8	0.1273	0.1932	0.1859	0.09353	0.235	0.07389	0.3063	1.002	2.406	24.32	0.005731	0.03502	0.03553	0.01226	0.02143	0.003749	15.49	
11	84501001	M	12.46	24.04	83.97	475.9	0.1186	0.2396	0.2273	0.08543	0.203	0.08243	0.2976	1.599	2.039	23.94	0.007149	0.07217	0.07743	0.01432	0.01789	0.01008	15.09	
12	845636	M	16.02	23.24	102.7	797.8	0.08206	0.06669	0.03299	0.03323	0.1528	0.05697	0.3795	1.187	2.466	40.51	0.004029	0.009269	0.01101	0.007591	0.0146	0.003042	19.19	
13	84610002	M	15.78	17.89	103.6	781	0.0971	0.1292	0.09954	0.06606	0.1842	0.06082	0.5058	0.9849	3.564	54.16	0.005771	0.04061	0.02791	0.01282	0.02008	0.004144	20.42	
14	846226	M	19.17	24.8	132.4	1123	0.0974	0.2458	0.2065	0.1118	0.2397	0.078	0.9555	3.568	11.07	116.2	0.003139	0.08297	0.0889	0.0409	0.04484	0.01284	20.96	
15	846381	M	15.85	23.95	103.7	782.7	0.08401	0.1002	0.09938	0.05364	0.1847	0.05338	0.4033	1.078	2.903	36.58	0.009769	0.03126	0.05051	0.01992	0.02981	0.003002	16.84	
16	84667401	M	13.73	22.61	93.6	578.3	0.1131	0.2293	0.2128	0.08025	0.2069	0.07682	0.2121	1.169	2.061	19.21	0.006429	0.05936	0.05501	0.01628	0.01961	0.008093	15.03	
17	84799002	M	14.54	27.54	96.73	658.8	0.1139	0.1595	0.1639	0.07364	0.2303	0.07077	0.37	1.033	2.879	32.55	0.005607	0.0424	0.04741	0.0109	0.01857	0.005466	17.46	
18	848406	M	14.68	20.13	94.74	684.5	0.09867	0.072	0.07395	0.05259	0.1586	0.05922	0.4727	1.24	3.195	45.4	0.005718	0.01162	0.01998	0.01109	0.0141	0.002085	19.07	
19	84862001	M	16.13	20.68	108.1	798.8	0.117	0.2022	0.1722	0.1028	0.2164	0.07356	0.5692	1.073	3.854	54.18	0.007026	0.02501	0.03188	0.01297	0.01689	0.004142	20.96	
20	849014	M	19.81	22.15	130	1260	0.09831	0.1027	0.1479	0.09498	0.1582	0.05395	0.7582	1.017	5.865	112.4	0.006494	0.01893	0.03391	0.01521	0.01356	0.001997	27.32	
21	8510426	B	13.54	14.36	87.46	566.3	0.09779	0.08129	0.06664	0.04781	0.1885	0.05766	0.2699	0.7886	2.058	23.56	0.008462	0.0146	0.02387	0.01315	0.0198	0.0023	15.11	
22	8510653	B	13.08	15.71	85.63	520	0.1075	0.127	0.04568	0.0311	0.1967	0.06811	0.1852	0.7477	1.383	14.67	0.004097	0.01898	0.01698	0.00649	0.01678	0.002425	14.5	
23	8510824	B	9.504	12.44	60.34	273.9	0.1024	0.06492	0.02956	0.02076	0.1815	0.06905	0.2773	0.9768	1.909	15.7	0.009606	0.01432	0.01985	0.01421	0.02027	0.002968	10.23	
24	8511133	M	15.34	14.26	102.5	704.4	0.1073	0.2135	0.2077	0.09756	0.2521	0.07032	0.4388	0.7096	3.384	44.91	0.006789	0.05328	0.06446	0.02252	0.03672	0.004394	18.07	
25	851509	M	21.16	23.04	137.2	1404	0.09428	0.1022	0.1097	0.08632	0.1769	0.05278	0.6917	1.127	4.303	93.99	0.004728	0.01259	0.01715	0.01038	0.01083	0.001987	29.17	
26	852552	M	16.65	21.38	110	904.6	0.1121	0.1457	0.1525	0.0917	0.1995	0.0633	0.8068	0.9017	5.455	102.6	0.006048	0.01882	0.02741	0.0113	0.01468	0.002801	26.46	
27	852631	M	17.14	16.4	116	912.7	0.1186	0.2276	0.2229	0.1401	0.304	0.07413	1.046	0.976	7.276	111.4	0.008029	0.03799	0.03732	0.02397	0.02308	0.007444	22.25	
28	852763	M	14.58	21.53	97.41	644.8	0.1054	0.1868	0.1425	0.08783	0.2252	0.06924	0.2545	0.9832	2.11	21.05	0.004452	0.03055	0.02681	0.01352	0.01454	0.003711	17.62	
29	852781	M	18.61	20.25	122.1	1094	0.0944	0.1066	0.149	0.07731	0.1697	0.05699	0.8529	1.849	5.632	93.54	0.01075	0.02722	0.05081	0.01911	0.02293	0.004217	21.31	

The basic steps for building and training a machine learning model using AutoAI:

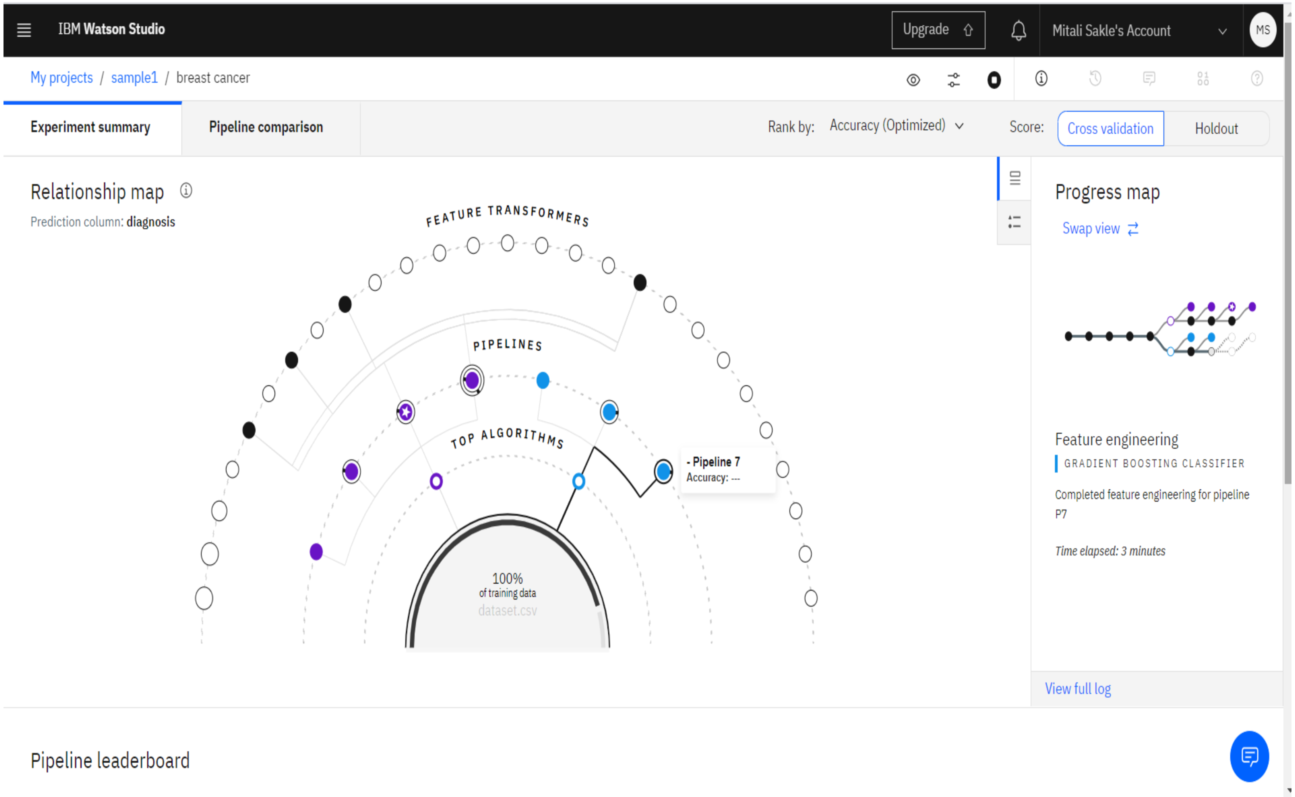
1. Build and train the model 2. Deploy the trained model 3. Test the deployed model

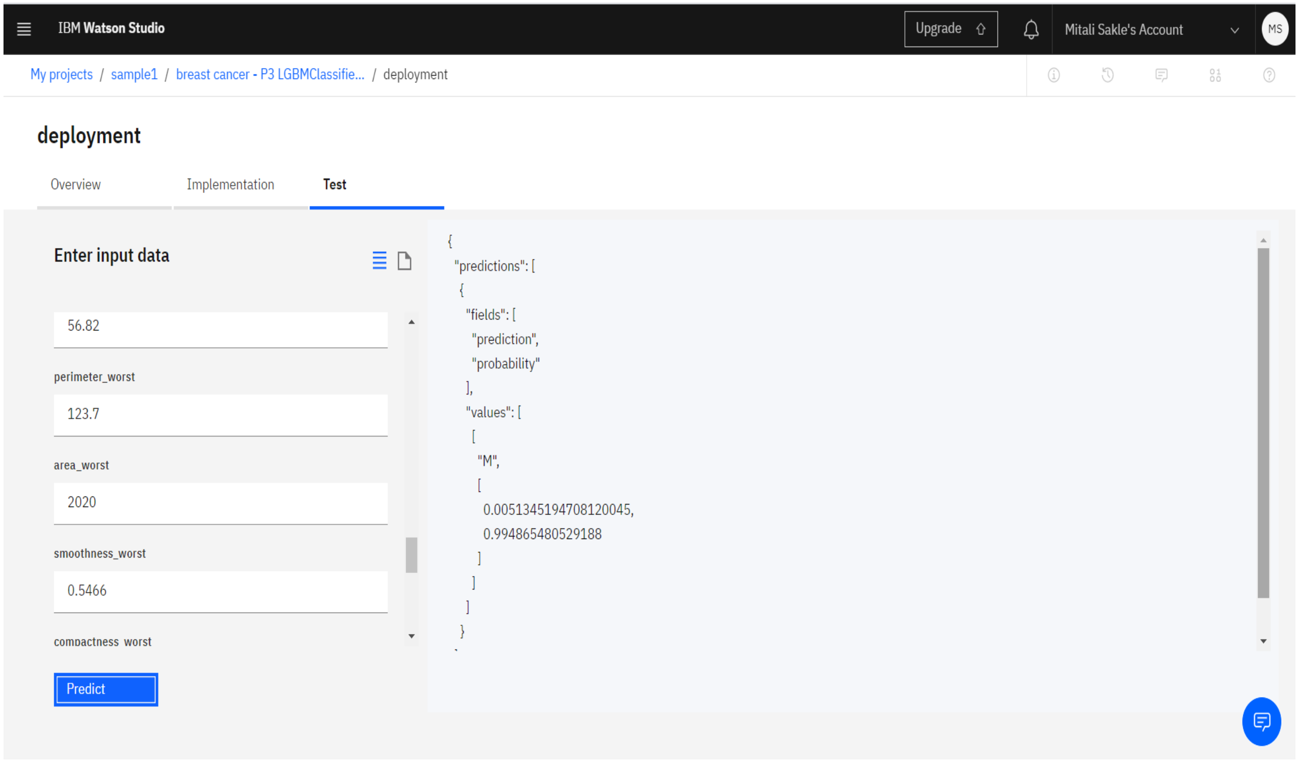
5. FlowChart -


create a project in watson studio
Dowload the saple training data file
Build and train the experiment
Deploy the trained model
Test the deployed model

6. Result -

We Successfully obtained the pipeline with an accuracy score of 0.965. As you can see from images below





56.82

perimeter_worst: 123.7

area_worst: 2020

smoothness_worst: 0.5466

compactness_worst

```
{
 "predictions": [
  {
   "fields": [
    "prediction",
    "probability"
   ],
   "values": [
    [
     "M",
     [
      0.0051345194708120045,
      0.994865480529188
     ]
    ]
   ]
  }
```

7. Advantages And Disadvantages

Advatages

1. Easily identifies trends and patterns
2. No human intervention needed (automation)
3. Continuous Improvement
4. Handling multi-dimensional and multi-variety data 5. Wide Applications

Disadvantages

1. Data Acquisition
2. Time and Resources
3. Interpretation of Results
4. High error-susceptibility

8. Applications -

Application of Breast Cancer Risk Prediction Models in Clinical PracticeBreast cancer risk assessment provides an estimation of disease risk that can be used to guide management for women at all levels of risk. In addition, the likelihood that breast cancer risk is due to specific genetic susceptibility (such as BRCA1 or BRCA2 mutations) can be determined. Recent developments have reinforced the clinical importance of breast cancer risk assessment. Tamoxifen chemoprevention as well as prevention studies such as the Study of Tamoxifen and Raloxifene are available to women at increased risk of developing breast cancer.

In addition, specific management strategies are now defined for BRCA1 and BRCA2 mutation carriers. Risk may be assessed as the likelihood of developing breast cancer (using risk assessment models) or as the likelihood of detecting a BRCA1 or BRCA2 mutation (using prior probability models). Each of the models has advantages and disadvantages, and all need to be interpreted in context. We review available risk assessment tools and discuss their application. As illustrated by clinical examples, optimal counseling may require the use of several models, as well as clinical judgment, to provide the most accurate and useful information to women and their families.

9. Conclusion

1.

Breast cancer if found at an early stage will help save lives of thousands of women or even men. This project will help the real world patients and doctors to gather as much information as they can. The research on nine papers has helped us gather the data for the project proposed by us. By using machine learning algorithms we will be able to classify and predict the cancer into benign or malignant.

2.

Machine learning algorithms can be used for medical oriented research, it advances the system, reduces human errors and lowers manual mistake

10. Future Scope

AI is set to change the medical industry in the coming decades — it wouldn’t make sense for pathology to not be disrupted too. Currently, ML models are still in the testing and experimentation phase for cancer prognoses. As datasets are getting larger and of higher quality, researchers are building increasingly accurate models. Here’s what a future cancer biopsy might look like: You perform clinical tests, either at a clinic or at home. Data is inputted into a pathological ML system. A few minutes later, you receive an email with a detailed report that has an accurate prediction about the development of your cancer. While you might not see AI doing the job of a pathologist today, you can expect ML to replace your local pathologist in the coming decades, and it’s pretty exciting! ML models still have a long way to go, most models still lack sufficient data and suffer from bias. Yet, something we are certain of is that ML is the next step of pathology, and it will disrupt the industry

11. Biblography -

https://www.kaggle.com/uciml/breast-cancer-wisconsin-data

https://www.researchgate.net/publication/337275322_Breast_Cancer_Prediction_using_Supervised_Machine_Learning_Algorithms
https://www.ijert.org/breast-cancer-classification-and-prediction-using-machine-learning#:~:text=Breast%20Cancer%20Prediction%20Using%20Genetic,detection%20of%20Breast%20Cancer%20Prediction.

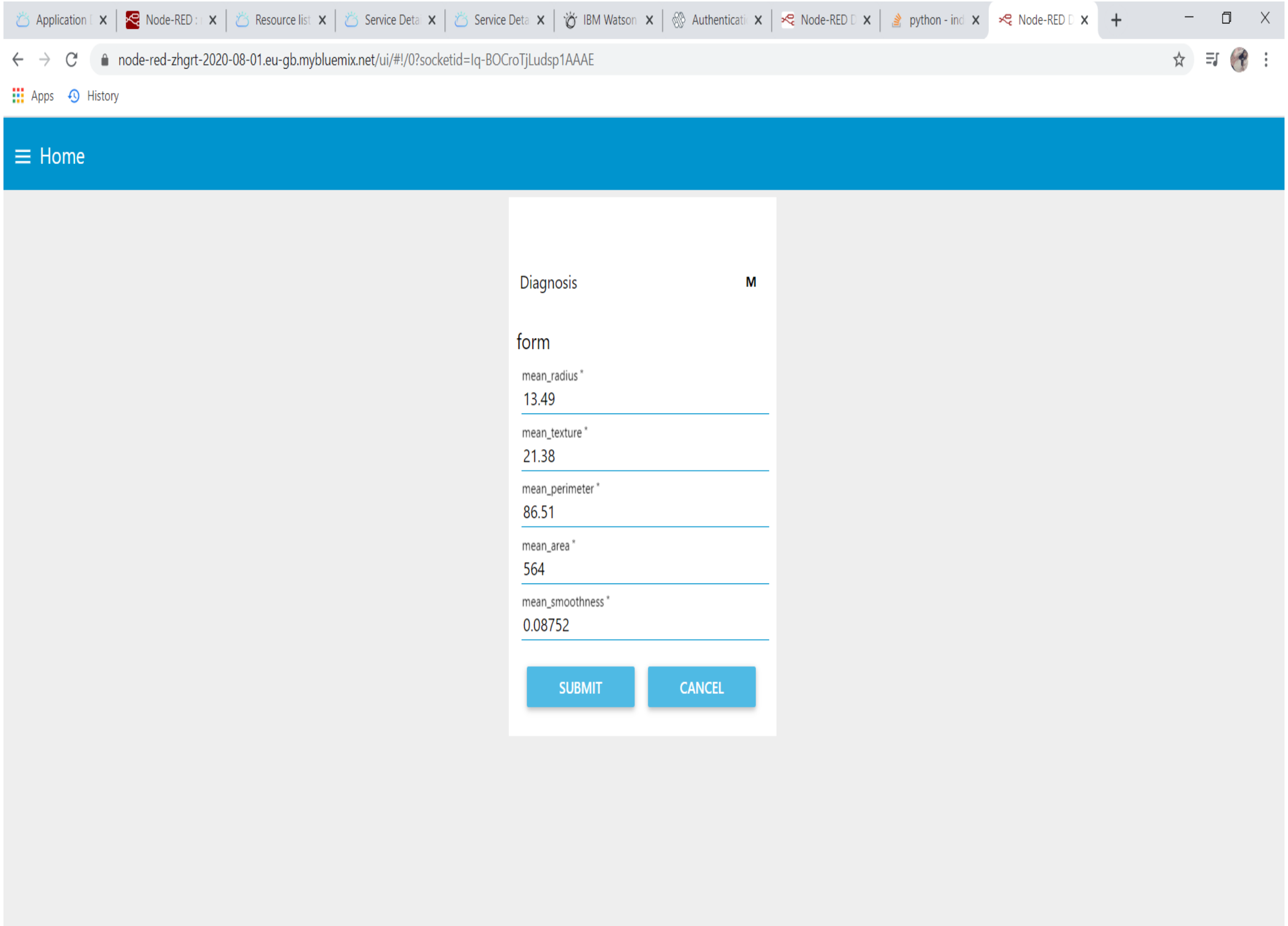
https://dataplatform.cloud.ibm.com/docs/content/wsj/analyze-data/autoai_example_binary_classifier.html

12.APPENDIX:-

A. SOURCE CODE

[{"id":"80fc8dba.aa05a","type":"tab","label":"Flow 2","disabled":false,"info":""},{"id":"3c2e1670.3a554a","type":"ui_form","z":"80fc8dba.aa05a","name":"","label":"","group":"acfa63c0.1b9b4","order":2,"width":0,"height":0,"options":[{"label":"mean_radius","value":"mr","type":"number","required":true,"rows":null},{"label":"mean_texture","value":"mt","type":"number","required":true,"rows":null},{"label":"mean_perimeter","value":"mp","type":"number","required":true,"rows":null},{"label":"mean_area","value":"ma","type":"number","required":true,"rows":null},{"label":"mean_smoothness","value":"ms","type":"number","required":true,"rows":null},{"label":"state1","value":"st1","type":"number","required":true,"rows":null},{"label":"state2","value":"st2","type":"text","required":true,"rows":null}],"formValue":{"mr":"","mt":"","mp":"","ma":"","ms":"","st1":"","st2":""},"payload":"","submit":"submit","cancel":"cancel","topic":"","x":150,"y":1060,"wires":[["217d2375.57153c"]]},{"id":"217d2375.57153c","type":"function","z":"80fc8dba.aa05a","name":"PreToken","func":"global.set(\"mr\",msg.payload.mr)\nglobal.set(\"mt\",msg.payload.mt)\nglobal.set(\"mp\",msg.payload.mp)\nglobal.set(\"ma\",msg.payload.ma)\nglobal.set(\"ms\",msg.payload.ms)\nglobal.set(\"st1\",msg.payload.st1)\nglobal.set(\"st2\",msg.payload.st2)\nvar apikey=\"Q0dqLIDyWxUgR4pFo3lxLPOmtNfmJ-ElCCVoMrBd80__\";\nmsg.headers={\"content-type\":\"application/x-www-form-urlencoded\"}\nmsg.payload={\"grant_type\":\"urn:ibm:params:oauth:grant-type:apikey\",\"apikey\":apikey}\nreturn msg;\n","outputs":1,"noerr":0,"x":220,"y":900,"wires":[["ed5d39e0.2ea778"]]},{"id":"ed5d39e0.2ea778","type":"http request","z":"80fc8dba.aa05a","name":"","method":"POST","ret":"obj","paytoqs":false,"url":"https://iam.cloud.ibm.com/identity/token","tls":"","persist":false,"proxy":"","authType":"","x":390,"y":1100,"wires":[["e80bd4bd.9b0628"]]},{"id":"e80bd4bd.9b0628","type":"function","z":"80fc8dba.aa05a","name":"Pre Prediction","func":"var mr=global.get(\"mr\")\nvar mt=global.get(\"mt\")\nvar mp=global.get(\"mp\")\nvar ma=global.get(\"ma\")\nvar ms=global.get(\"ms\")\nvar st1=global.get(\"st1\")\nvar st2=global.get(\"st2\")\nvar token=msg.payload.access_token\nvar instance_id=\"2d51206d-d636-4a6d-8348-5c806a2cd38c\"\nmsg.headers={'Content-Type': 'application/json',\"Authorization\":\"Bearer \"+token,\"ML-Instance-ID\":instance_id}\nmsg.payload={\"input_data\": [{\"fields\": [\"state1\",\"state2\",\"mean_radius\", \"mean_texture\", \"mean_perimeter\", \"mean_area\", \"mean_smoothness\"], \"values\": [[st1,st2,mr,mt,mp,ma,ms]]}]}\nreturn msg;","outputs":1,"noerr":0,"x":540,"y":880,"wires":[["55791bce.75da34"]]},{"id":"55791bce.75da34","type":"http request","z":"80fc8dba.aa05a","name":"","method":"POST","ret":"obj","paytoqs":false,"url":"https://eu-gb.ml.cloud.ibm.com/v4/deployments/15844736-fa9b-46a6-8c54-5815c9a3bd4e/predictions","tls":"","persist":false,"proxy":"","authType":"","x":730,"y":1020,"wires":[["b66853a5.2756a","cecf2c30.c5022"]]},{"id":"317fa518.e8e9da","type":"ui_text","z":"80fc8dba.aa05a","group":"acfa63c0.1b9b4","order":1,"width":0,"height":0,"name":"","label":"Dignosis","format":"{{msg.payload}}","layout":"row-spread","x":1040,"y":1140,"wires":[]},{"id":"b66853a5.2756a","type":"function","z":"80fc8dba.aa05a","name":"","func":"msg.payload=msg.payload.predictions[0].values[0][1][1]\nreturn msg;","outputs":1,"noerr":0,"x":770,"y":780,"wires":[["317fa518.e8e9da","7e1c0c6b.6ffd84"]]},{"id":"7e1c0c6b.6ffd84","type":"debug","z":"80fc8dba.aa05a","name":"","active":true,"tosidebar":true,"console":false,"tostatus":false,"complete":"payload","targetType":"msg","x":1010,"y":780,"wires":[]},{"id":"cecf2c30.c5022","type":"debug","z":"80fc8dba.aa05a","name":"","active":true,"tosidebar":true,"console":false,"tostatus":false,"complete":"payload","targetType":"msg","x":1046.892333984375,"y":954.1111450195312,"wires":[]},{"id":"acfa63c0.1b9b4","type":"ui_group","z":"","name":"Breast Cancer Risk Prediction using IBM Auto AI","tab":"a76f3181.b9c51","order":1,"disp":true,"width":"6","collapse":false},{"id":"a76f3181.b9c51","type":"ui_tab","z":"","name":"Home","icon":"dashboard","disabled":false,"hidden":false}]

B.UI OUTPUT SCREENSHOT



node-red-zhgrt-2020-08-01.eu-gb.mybluemix.net/ui/#!/0?socketid=Iq-BOCroTjLudsp1AAAE

Diagnosis **M**

form

mean_radius *
13.49

mean_texture *
21.38

mean_perimeter *
86.51

mean_area *
564

mean_smoothness *
0.08752