

**ADVANCED DATA SCIENCE AND ARCHITECTURE (INFO 7390)**

**BREAST CANCER DETECTION AND ANALYSIS**

**Final Project Report**

**Team 5**

**DHRUV KANAKIA**

**AKILAN RAJENDIRAN**

**Under the guidance of**

**Sri Krishnamurthy**

SUMMARY

This report summarizes the analysis performed on two Breast Cancer Datasets. The links for two are:

1. <http://www.cbioportal.org/study?id=brca_metabric#clinical>
2. <http://www.bcsc-research.org/data/bcsc_data_definitions.html>

Using the two datasets we created two use cases:

1. Predict the risk of having Breast Cancer based on the input parameters given by the user.
2. Classify the Tumor based on different user vitals and cancer cells report.

The problem statement is divided into 4 sections:

1. **Data Wrangling**

* Web scraping the data from the above links and pre-processing
* Exploratory Data Analysis on Tableau, Python.

1. **Dockerizing the process:**

* Two docker images for web scraping and uploading the data on Azure Blob.
* Airflow pipeline

1. **Building and Evaluating models:**

* Prediction using Random Forest, SVM, Neural Network, Logistic Regression
* Classification using Random forest, SVM, Neural Network

1. **Creating User interface for Oncologists and Females (User)**

* **User:** will be able to check the risk/probability of her having breast cancer based on the prediction model and the input parameters.
* **Oncologist(Doctor):** will be able to classify what type of tumor is it based on classification model.

**PART 1: DATA INGESTION AND WRANGLING**

* 1. Data Wrangling and Pre-Processing

**11. Breast Cancer Surveillance Consortium Dataset(**The dataset has around 2,392,998 mammograms from women who participated in the study). The dataset consists of a .txt file and it includes information about women during the time of the study like:

Age, Menopause status, density, agegrp, race, bmi, cancer status etc.

The picture attached below shows the web url from which we are Web Scraping the Data.



CHALLENGE:

The challenge was to programmatically get through the login page and extract the dataset. The link to the landing page of the above screenshot : <http://www.bcsc-research.org/rfdataset/app2/protected/Logon.aspx?ReturnUrl=%2frfdataset%2fapp2%2fprotected%2frisk_dataset.zip>

SOLUTION:

* We tried going pass the login page using request libraries but couldn’t get pass it. Finally, we managed to extract the dataset using the mechanical soup library and here’s the code for it:



* The file downloaded by the above code was in .zip format and used the zipfile library to extract the zip file. The .txt file present in the zipfile was read using the pandas module, column names were allotted and then the file was written on local as ‘Risk.csv’.
* Below is the screenshot which shows how we gave column names to the file replaced the missing values with space characters(because we are going to clean the file using MICE on Azure).



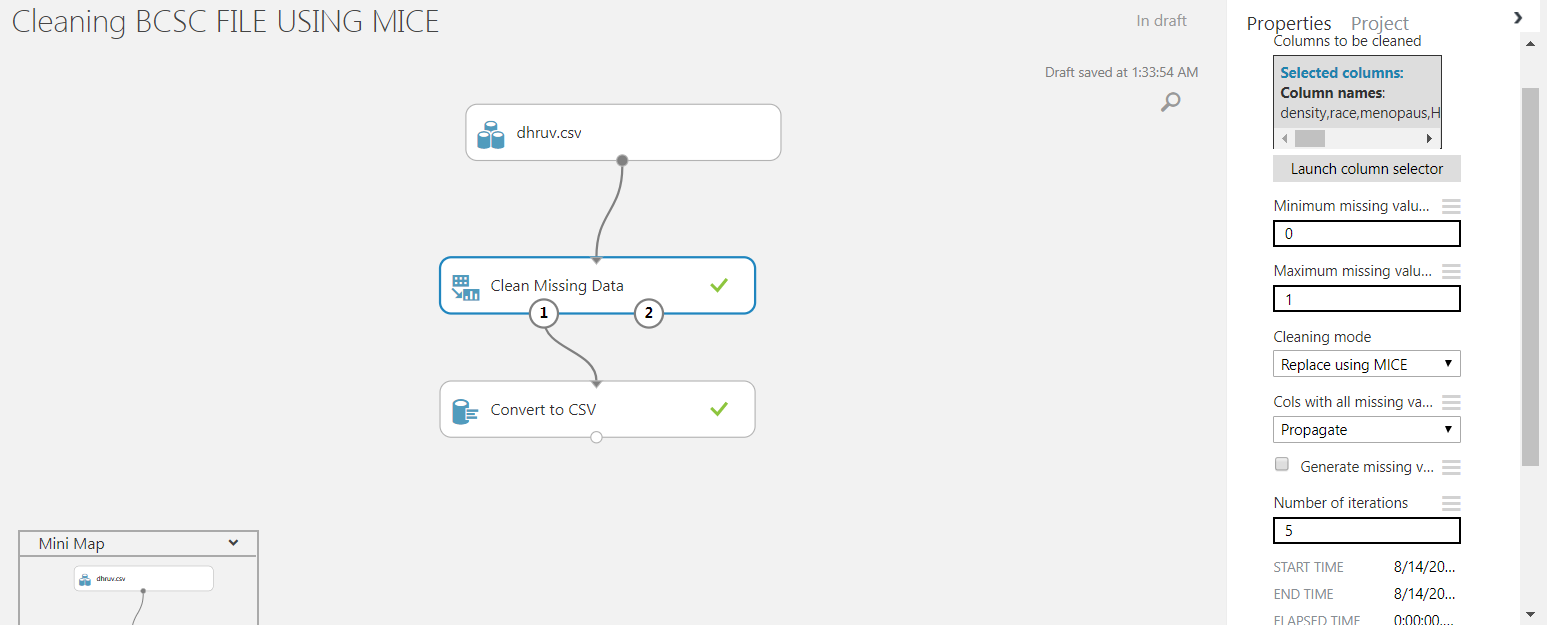
**PRE-PROCESSING:**

**PROBLEM:**

Since there are lot of missing values in this file we can neither ignore them nor just fill in mean/mode/random value as it is a patient dataset and nothing can be generalized.

**SOLUTION:**

After looking at resources online we came across something call DATA IMPUTATION techniques and how missing data can be handled by looking at patterns/ predicting them. Reading about different techniques we found MICE would suit the best for this problem and we implemented it on Azure.

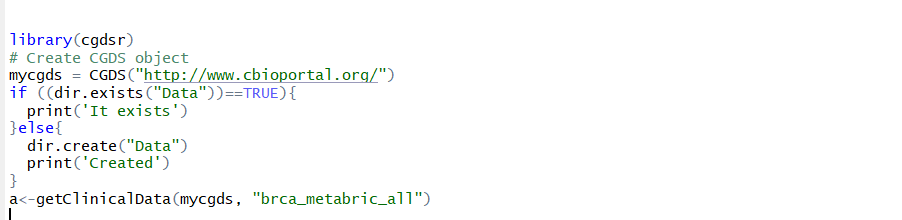
**Screenshot shows how the missing value is handled using Azure block Handling missing values(MICE)**

**2. Breast Cancer (METABRIC, Nature 2012 & Nat Commun 2016:**

This dataset consists of 2509 patients who have participated in the study and it includes details like type of cancer, treatments used, gene and tumor information.

We used their library(cgdsr) to web scrape the data using R

The screenshot attached below shows how we are web scraping the data using its WEB API.



**Pre-Processing:**

There are few categorical columns which needs to be converted into numbers for modeling.

Besides, we have also created manual clusters for age group <=62 and age group >62.

Following is the screenshot for pre processing of this file and storing it in local.



Part 2: Exploratory Data Analysis

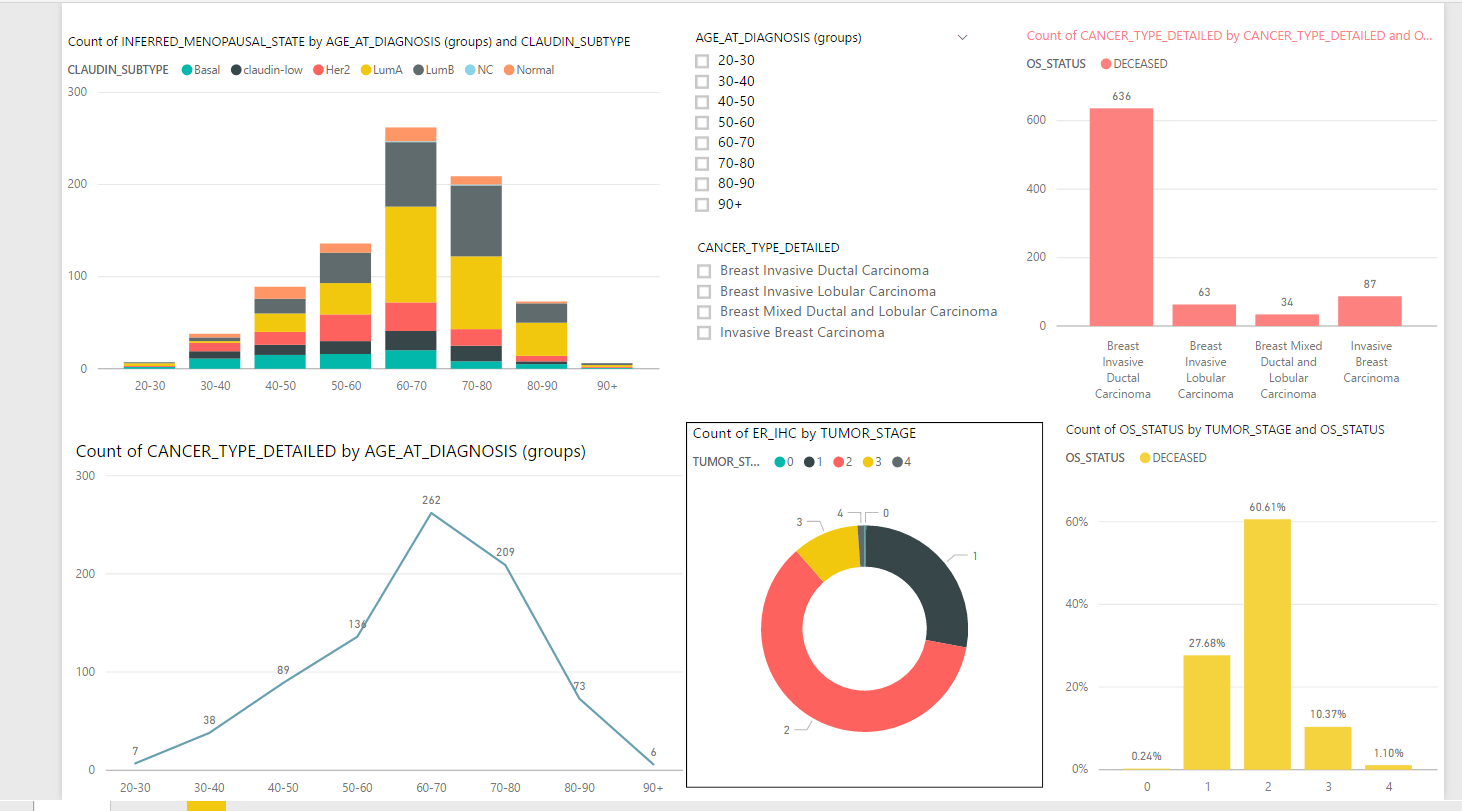
2.1 Analysis – Power BI

EDA on Metabric Dataset:

The first dashboard has 5 visualizations that explore the different aspects in the Metabric Dataset.

* The first plot shows different Claudin subtypes over the age group. The claudin low subtype tumors one of the important factors for evaluation of unique biology of structures and heterogeneity of Breast Cancer.
* The second plot shows the count of Different cancer subtypes in this dataset. Our multi class classification is built on this subtype attribute and it was necessary to explore that column.
* Age also played an important role among females getting breast cancer. Thus, the third plot analyzes the age group present in the dataset in the form of bins.
* Just like how we have Claudian subtype, the other important attribute in the dataset is ER\_IHC. This is a test that is performed to analyze whether or not the cancer cells have HER2. Based on this information further steps are taken.
* The last plot shows vital information about tumor stages. According to the plot(dataset) tumor stages at level 2 have the highest deceased rate.

The dashboard screenshot is attached below:

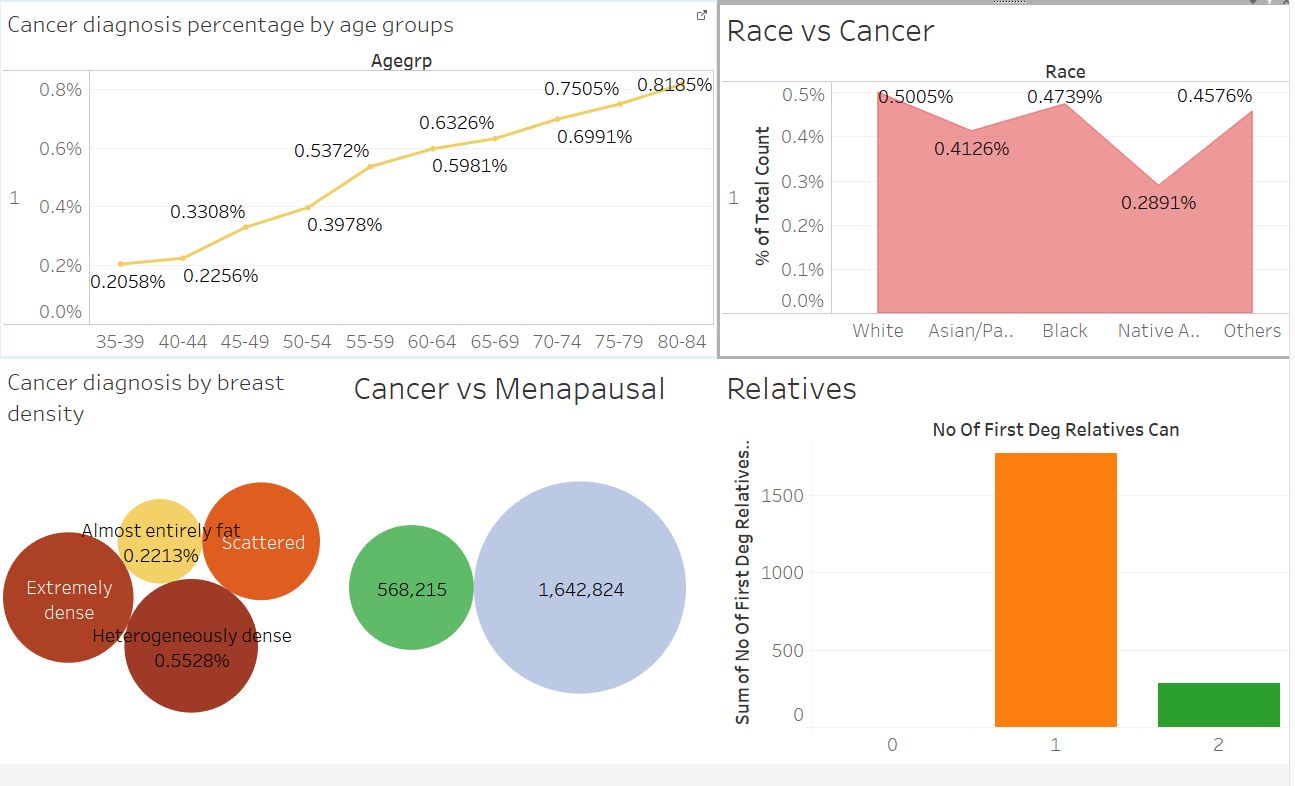


1. The second dashboard

EDA on BCSC Dataset(Power BI):

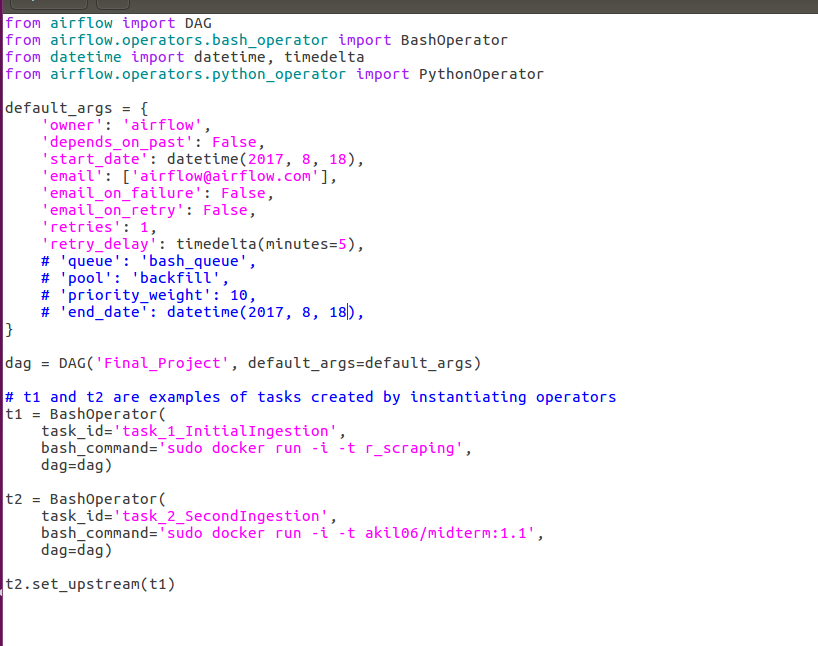
This dashboard also has five visualizations for exploring the Risk Dataset.

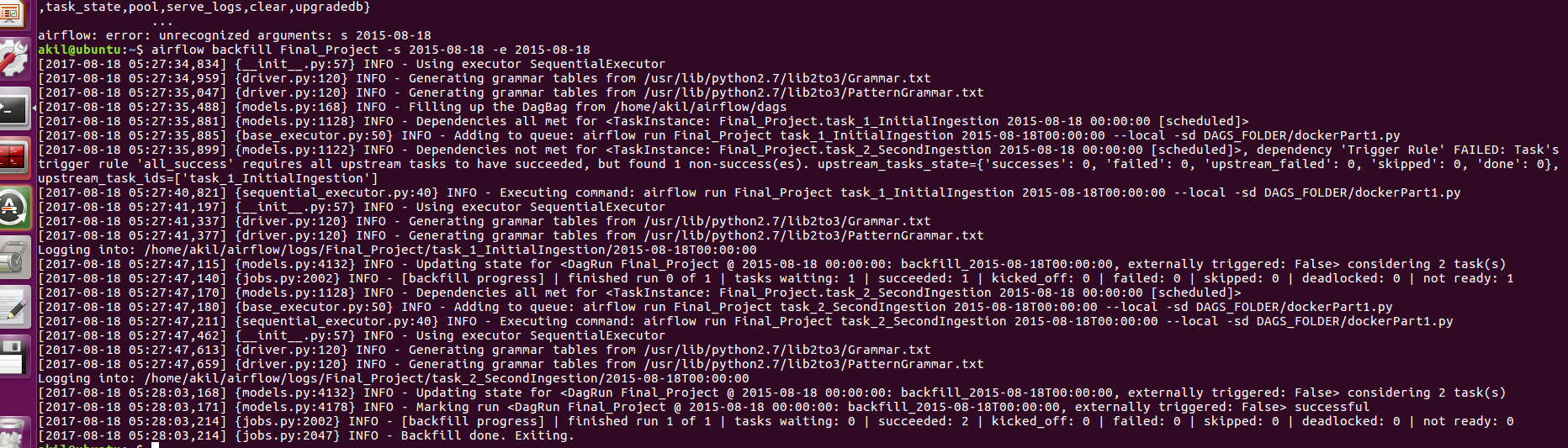
* The first visualization summarizes the percentage of cancer diagnosis over the years. We can clearly see that with increase in age there is high probability of getting cancer.
* The second visualization summarizes the information of race and cancer diagnosis. From, the graph we can clearly see that Whites and Black have the highest percentage of getting bgreast cancer as per dataset.
* The third visualization is the cancer diagnosis visualization which measures the density of breast cancer. From the dataset we can conclude that majority of samples are from Extremely and Heterogeneously data
* The fourth visualization analyzes the menopause status in the dataset where in it indicates number of patients who in are post and pre menopause and are suffering breast cancer.
* The last visualization describes the risk of breast cancer on the basis of number of first degree Relatives

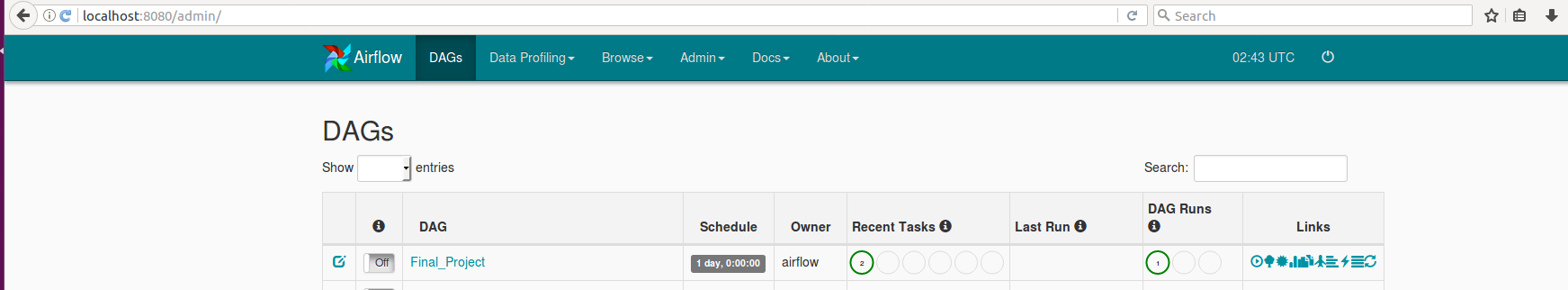


PART II: DOCKERIZING THE PROCESS

PIPELINING THROUGH Airflow





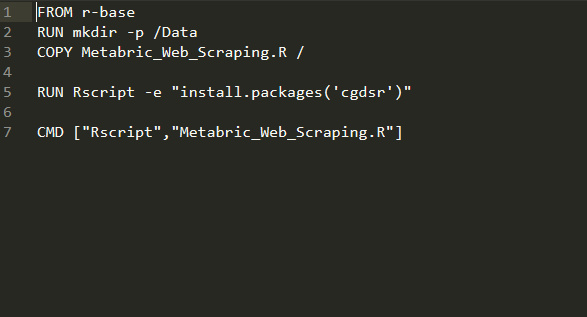


DOCKER STEPS AND PROCESS

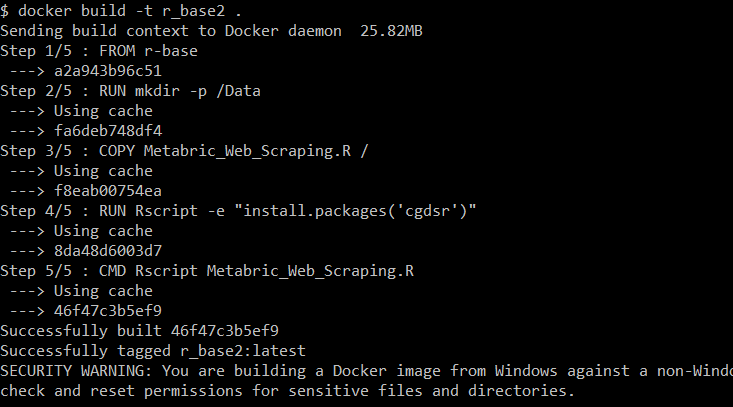
IMAGE 1:

The first image is a R image that web scrapes Metabric Data, preprocesses it and stores it in data container of the image.

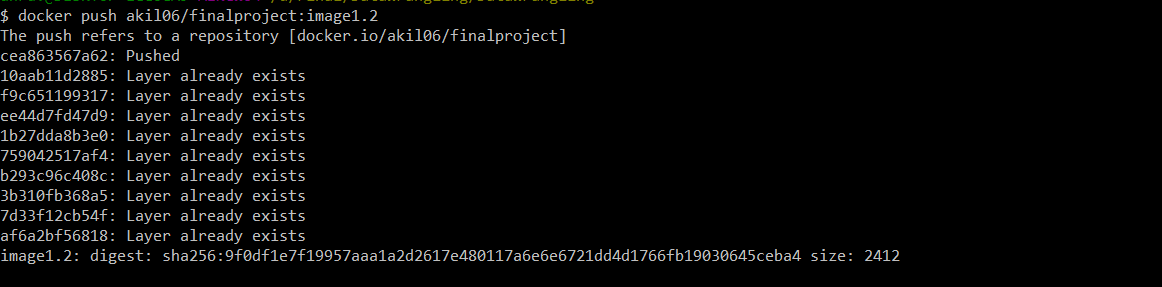
Here’s the DOCKERFILE:



Built the docker image using with name r\_base



Pushing the image to the Docker hub. The image name is [**akil06**](https://hub.docker.com/u/akil06/)**/**[**finalproject**](https://hub.docker.com/r/akil06/finalproject/)**:image1.2**



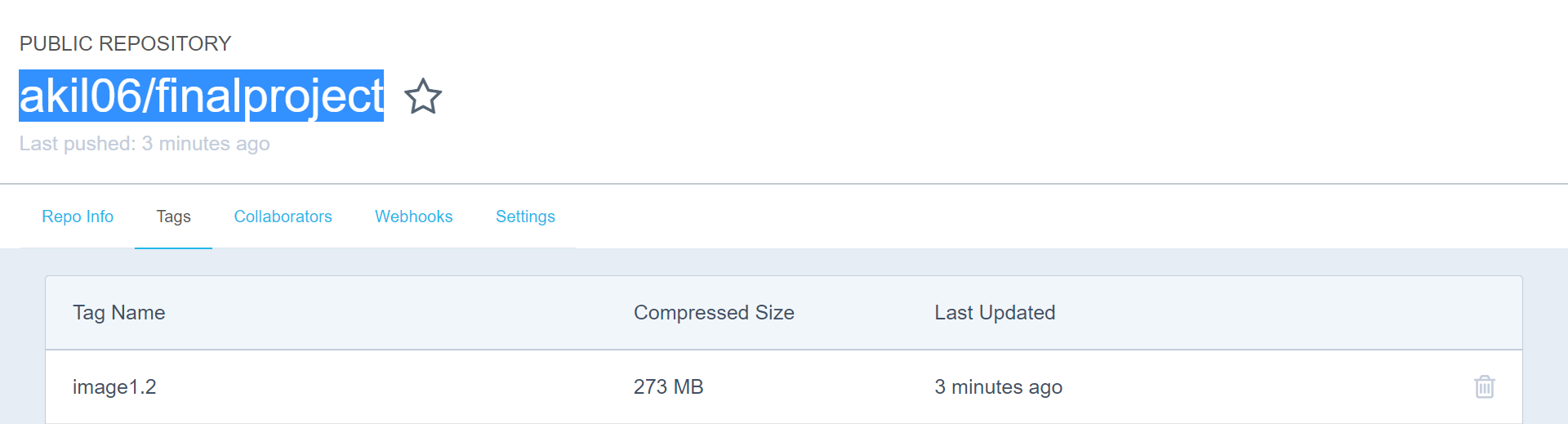
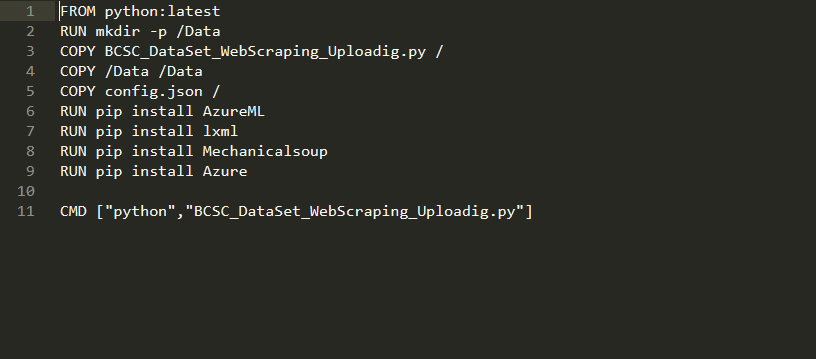


Image can be called using command docker pull [**akil06**](https://hub.docker.com/u/akil06/)**/**[**finalproject**](https://hub.docker.com/r/akil06/finalproject/)**:image1.2**

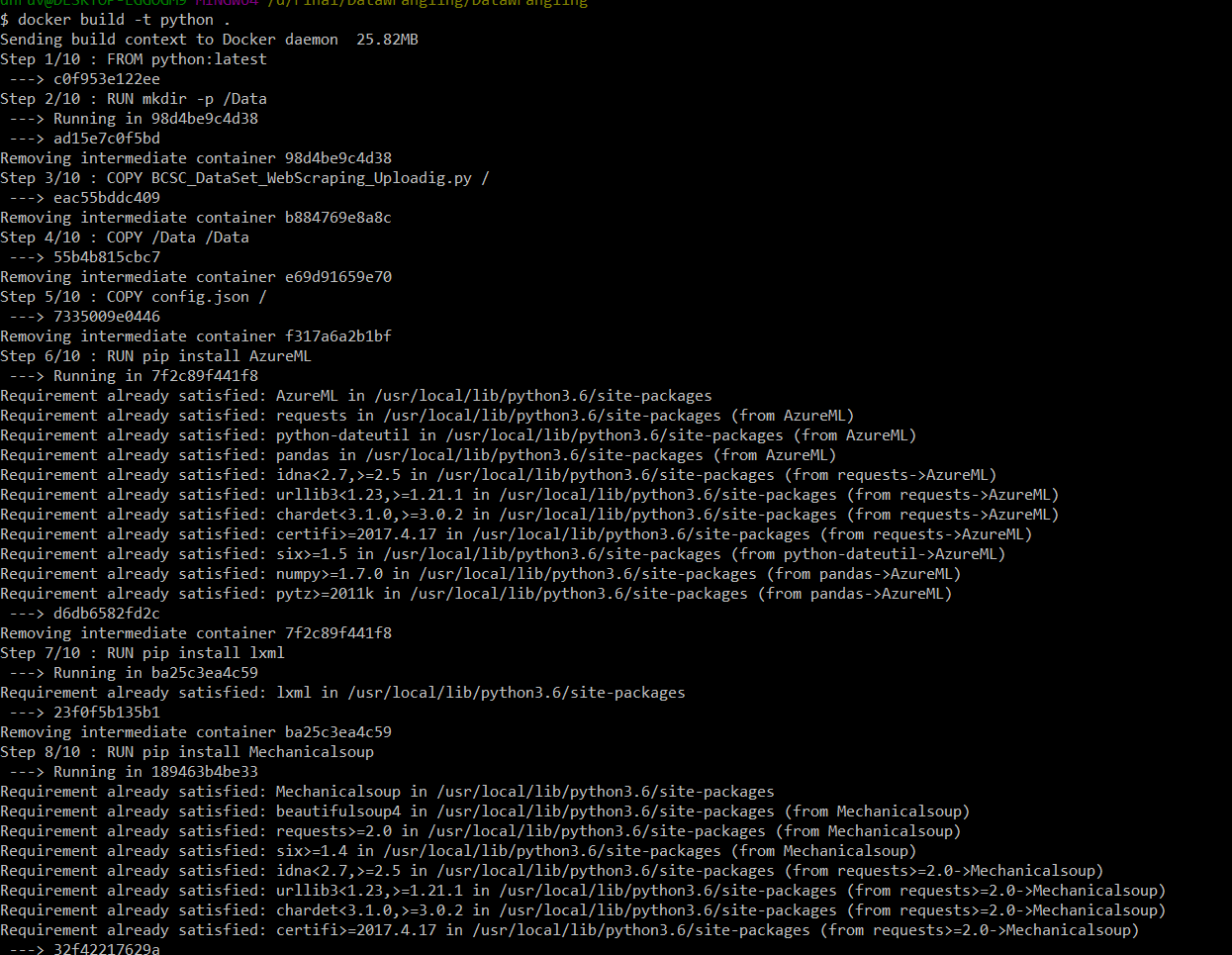
IMAGE 2:

The second image is on python base image. It web scrapes the BCSC data and uploads the files onto blob.

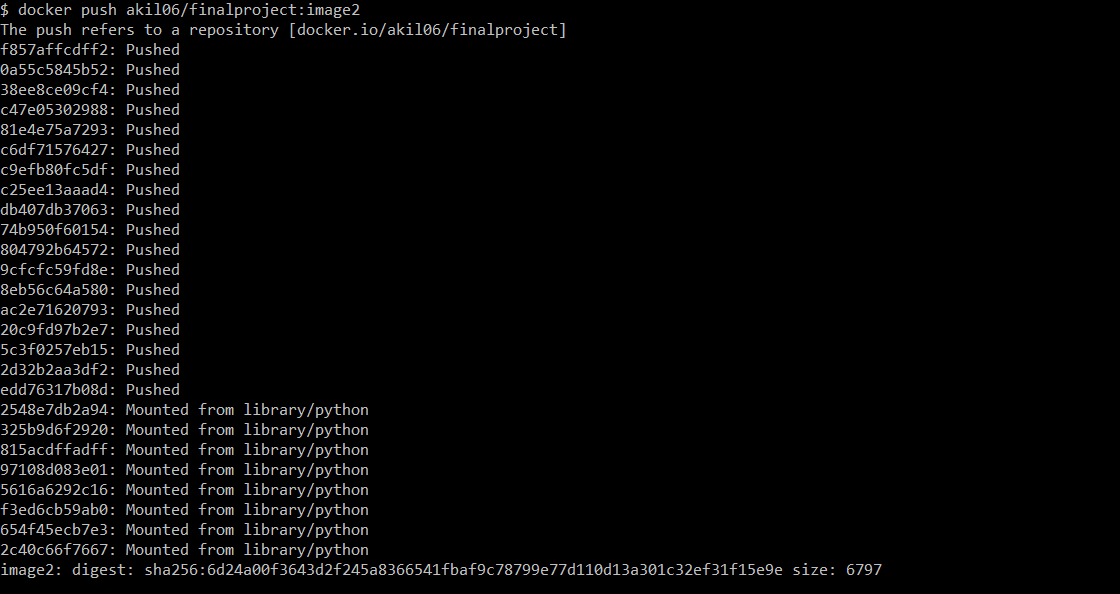
Here’s the Docker file:



Built the image using python base image. Here’s the command and screenshot of it.



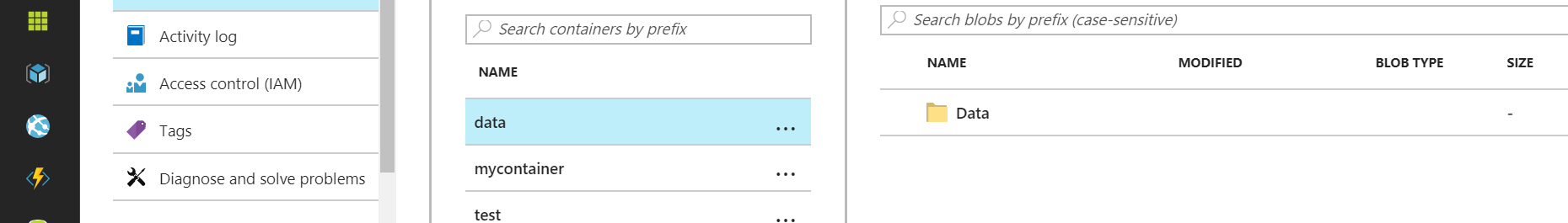
Pushed the image on docker hub with tag image2

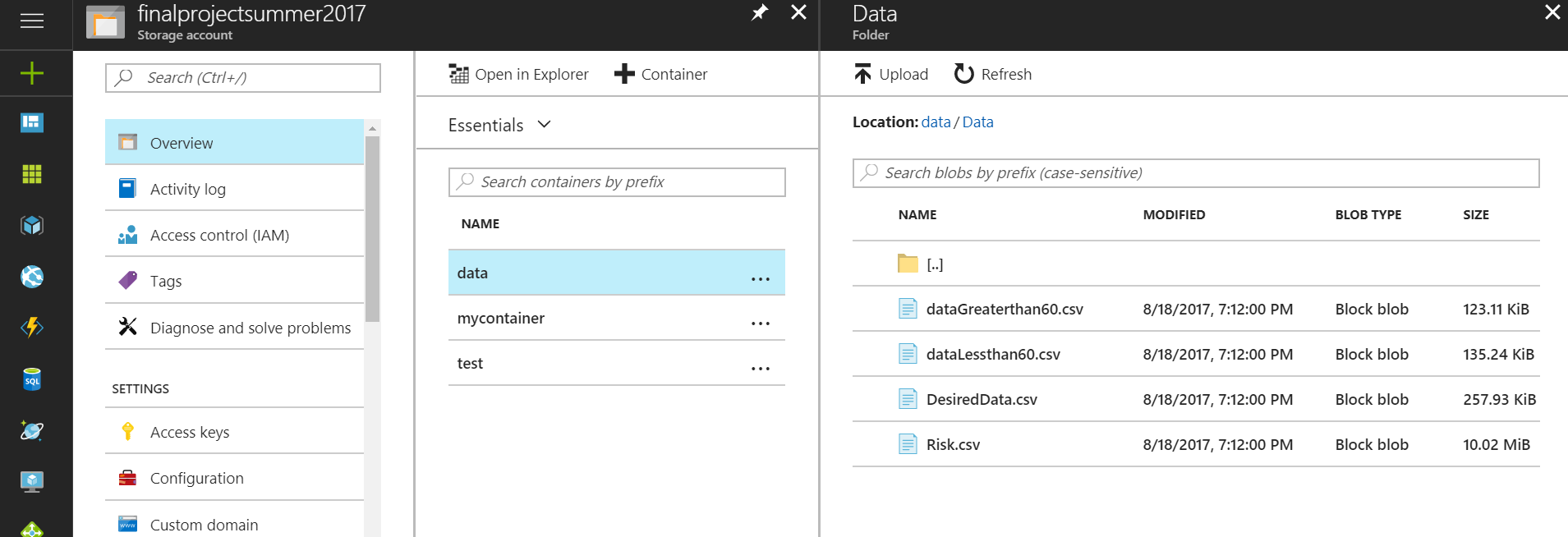


To pull the image you can simply run the command docker pull [**akil06**](https://hub.docker.com/u/akil06/)**/**[**finalproject**](https://hub.docker.com/r/akil06/finalproject/)**:image2**

Once, both the images are up and running we can see the files being transferred to the **AZURE BLOB.**

The screenshot shows presence of file on azure blob from where we can directly use the data into Azure Machine learning Studio.

Image showing a folder being created on the Azure blob. 

The next picture shows the csv files that got transferred form local to a foreign student. m

**PART 3. Building and Evaluating Model**

**Breast Cancer Risk Assessment**

``

**What is Breast Cancer Risk Assessment?**

This model is built upon the Breast Cancer Surveillance Consortium (BCSC Dataset).

**WORKFLOW:**

Web Scraping &

Pre-processing

Balanced Data using Over and Under Sampling

Deployed Neural Network(Over Sampling) with Acc. 88.61%

Two Class SVM

Neural Network

Naïve Bayes

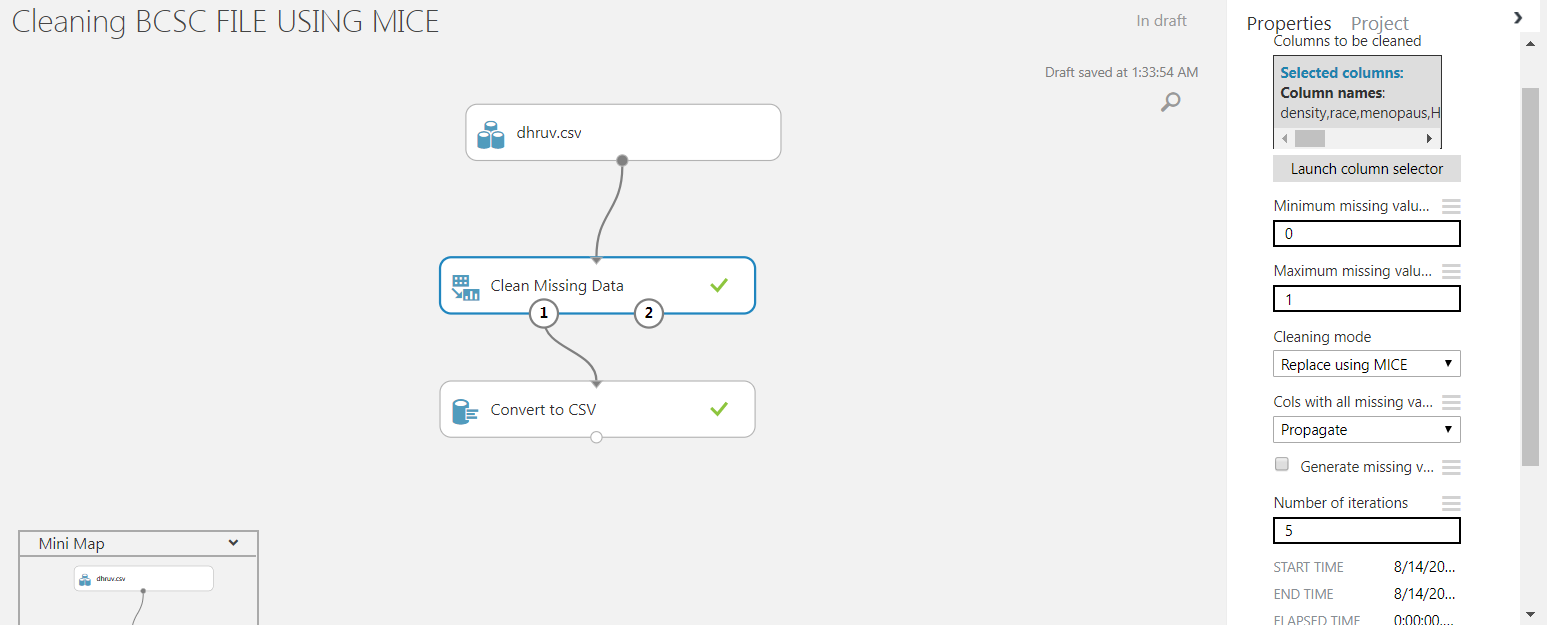
Random Forest

Logistic

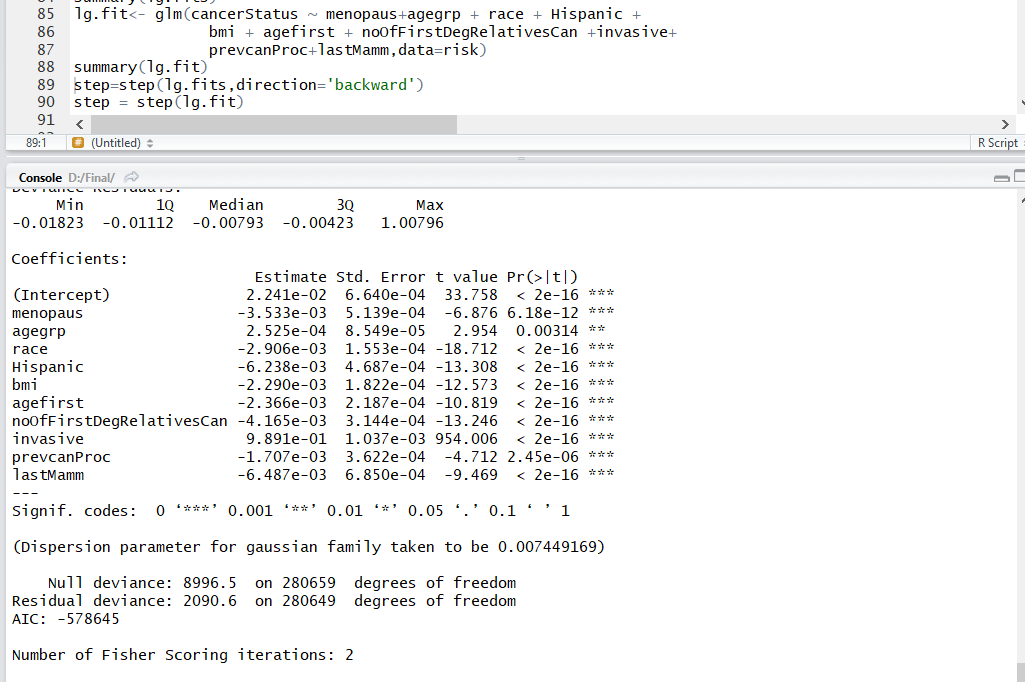
Regression

Handled Missing data using MICE

* Handing Missing values using MICE: As mentioned above in the Pre processing part due to lot of missing values we implemented MICE to fill up the missing values.



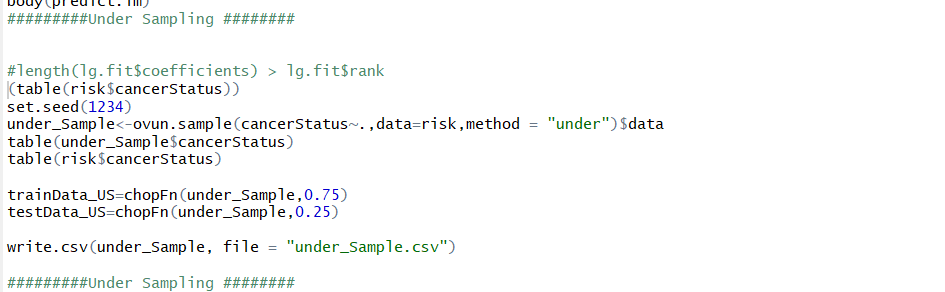
* Feature Selection: Using Stepwise regression we selected the features with the highest significance(\*\*\*). Below is the screenshot which shows the attributes we have selected based on the significance.



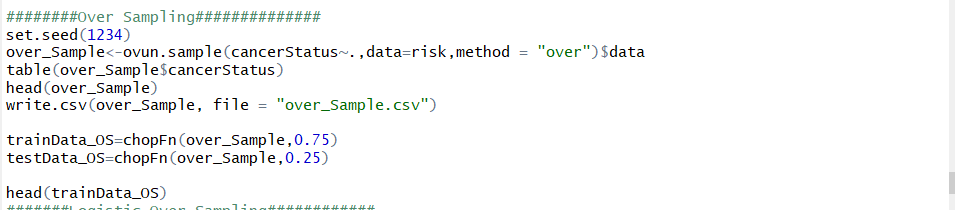
* Balancing the data: After analyzing at the attribute that we are classifying we see that the attribute is highly imbalanced and we need to balance it for better results.

We used two methods for balancing 1. Over Sampling and 2. Under Sampling.

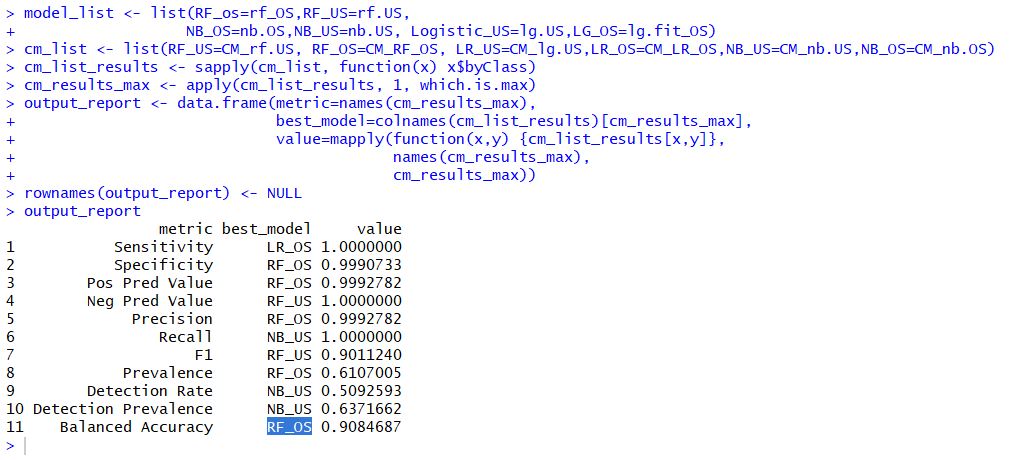
Below is the screenshot for under sampling.



Here is the screenshot for Over Sampling .



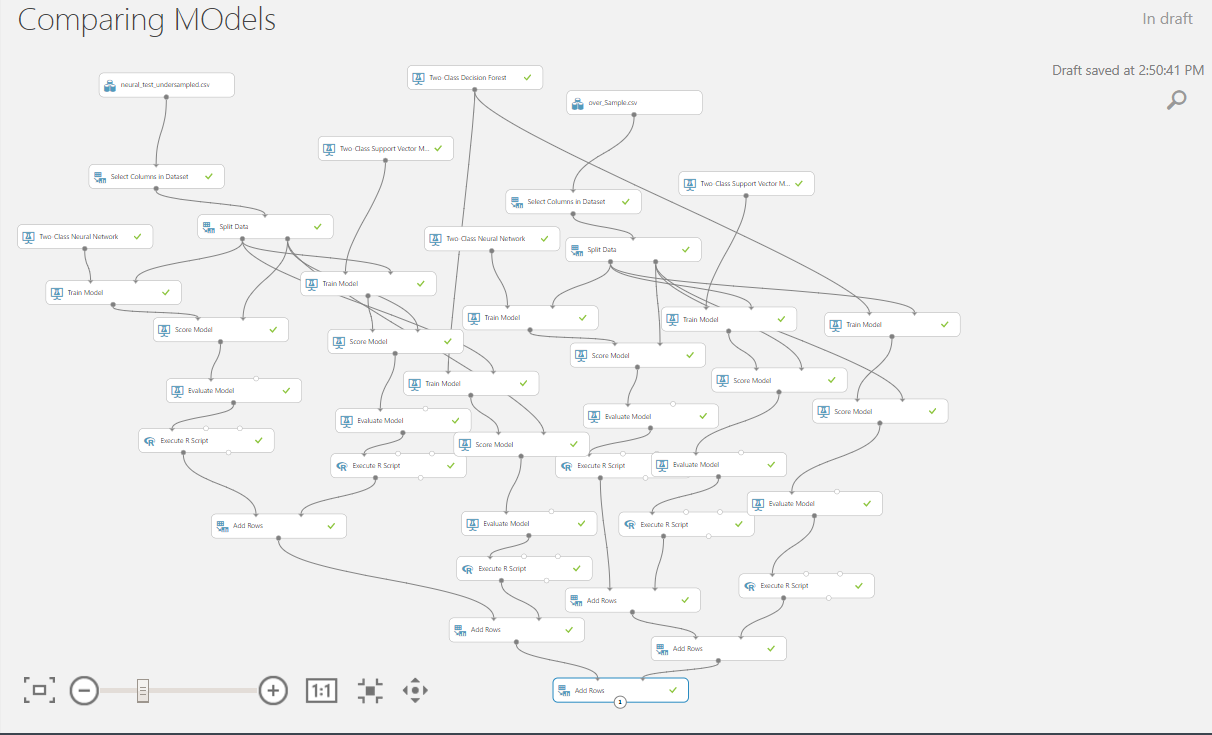
* We implemented RandomForest, Logistic Regression, Naïve Bayes in R and here is the comparison amongst them



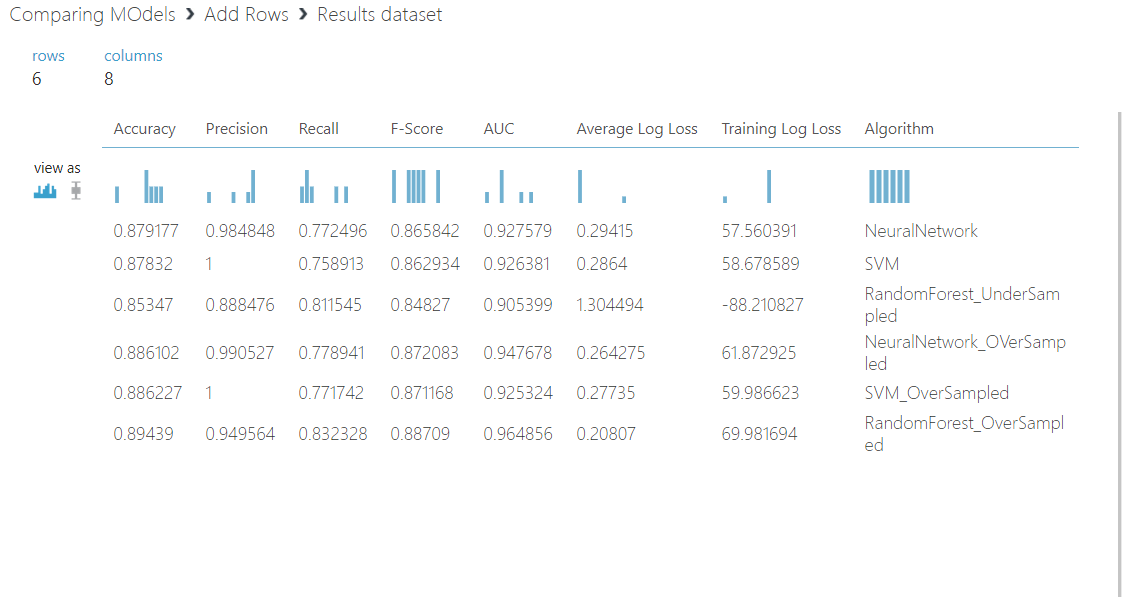
The above method returns the method which is the best for each testing parameter. Looking at it we realize that Random Forest Over Sampled Model is the best amongst the all.

Now we will compare the Random Forest Over Sampled with Neural network and SVM in Azure ML.

* Comparing RandomForest, Neural networks and SVM in Azure ML

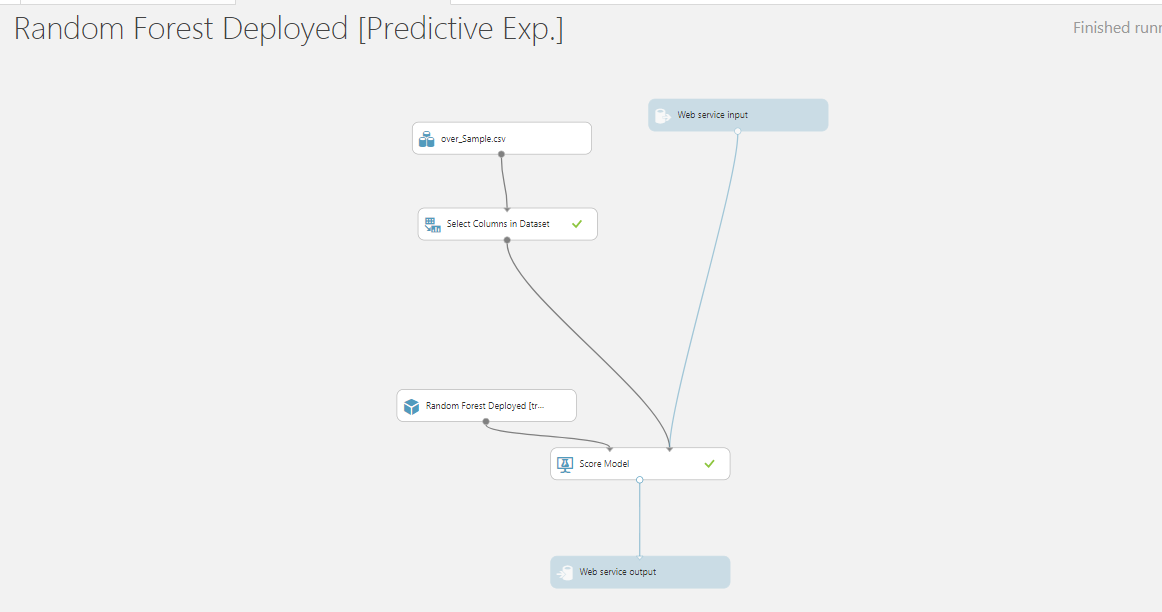


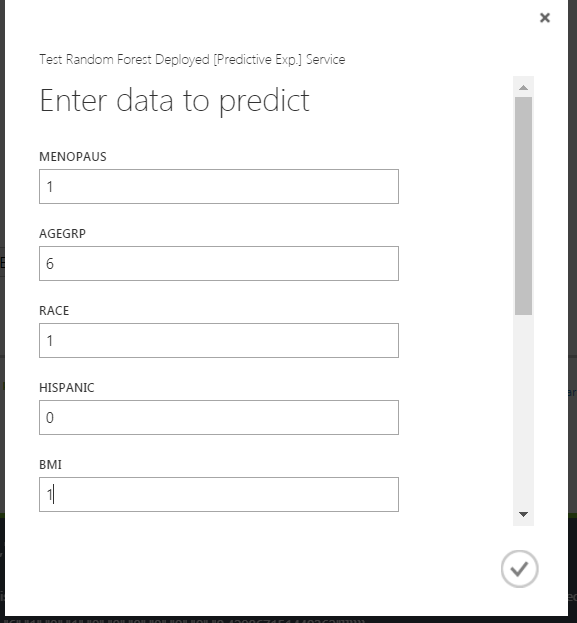
Here’s the comparison matrix to compare the test parameters of different models:



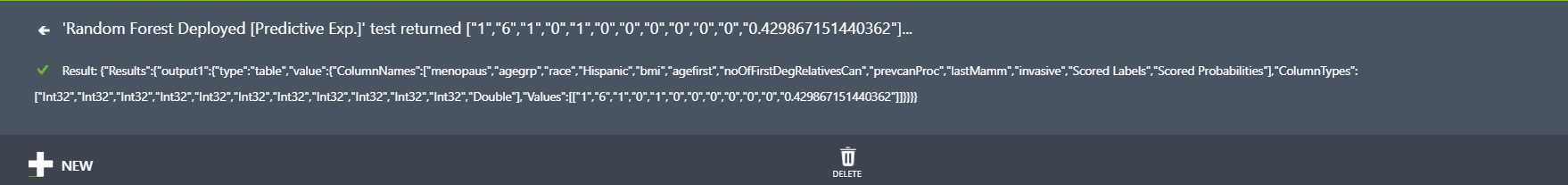
Looking at the comparison matrix we can see that RandomForest OverSampled is giving the best results. So, we deploy Random Forest oversampled and use the API of it in our application.

Here’s the experiment that shows how we deployed it:



Testing the Model on AzureML: 

Output:



**Classification of Breast cancer subtype:**

There are various types of Breast Cancer and it is difficult to classify what type of Cancer it is unless we perform rigorous tests on it. We have come up with a model that can classify the lump into what type of cancer it is based on several input parameters.

**WORKFLOW:**

Clustered the data into two clusters on the basis of age and then deployed Decision forest and One vs Multiclass SVM as best models for respective clusters.

SVM

One vs All multi class

Neural Network

Decision Forest

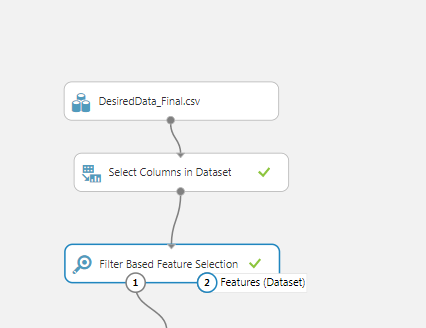
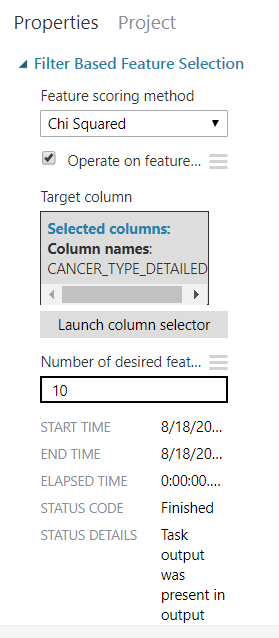
Web Scraping &

Pre-processing

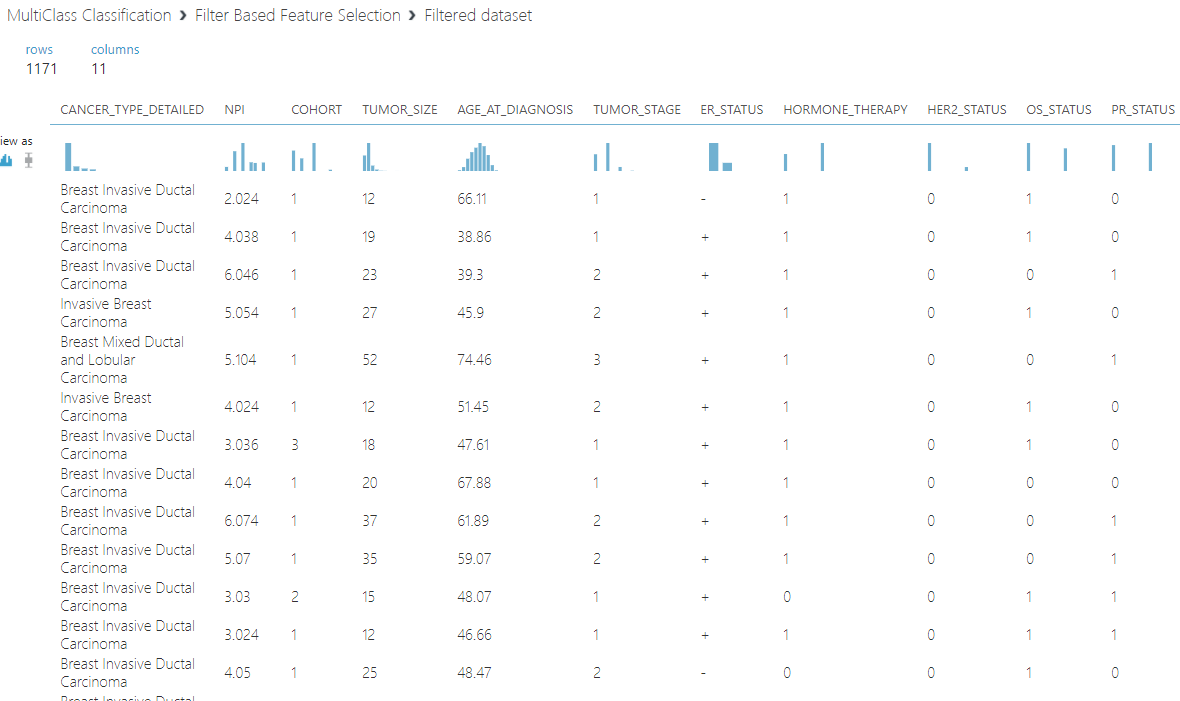
**3.1.3 FEATURE SELECTION:**

Before proceeding with our models, we have done feature selection using Chi Squared method.

The best features that add to the predictive power of the model will be retained and irrelevant features removed from the model.



We selected 10 features using the Filter Based Feature Selection block. The statistical method used for scoring method was Chi Squared. The features selected are:

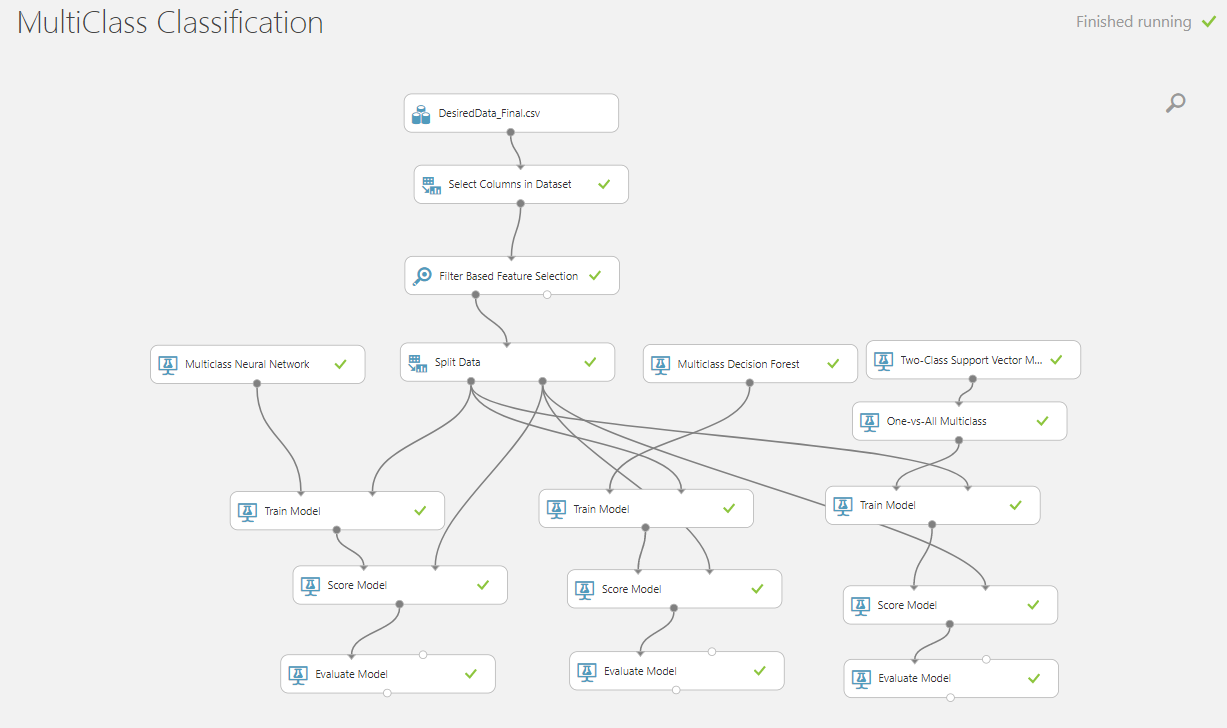


**Machine Learning Algorithms and Output:**

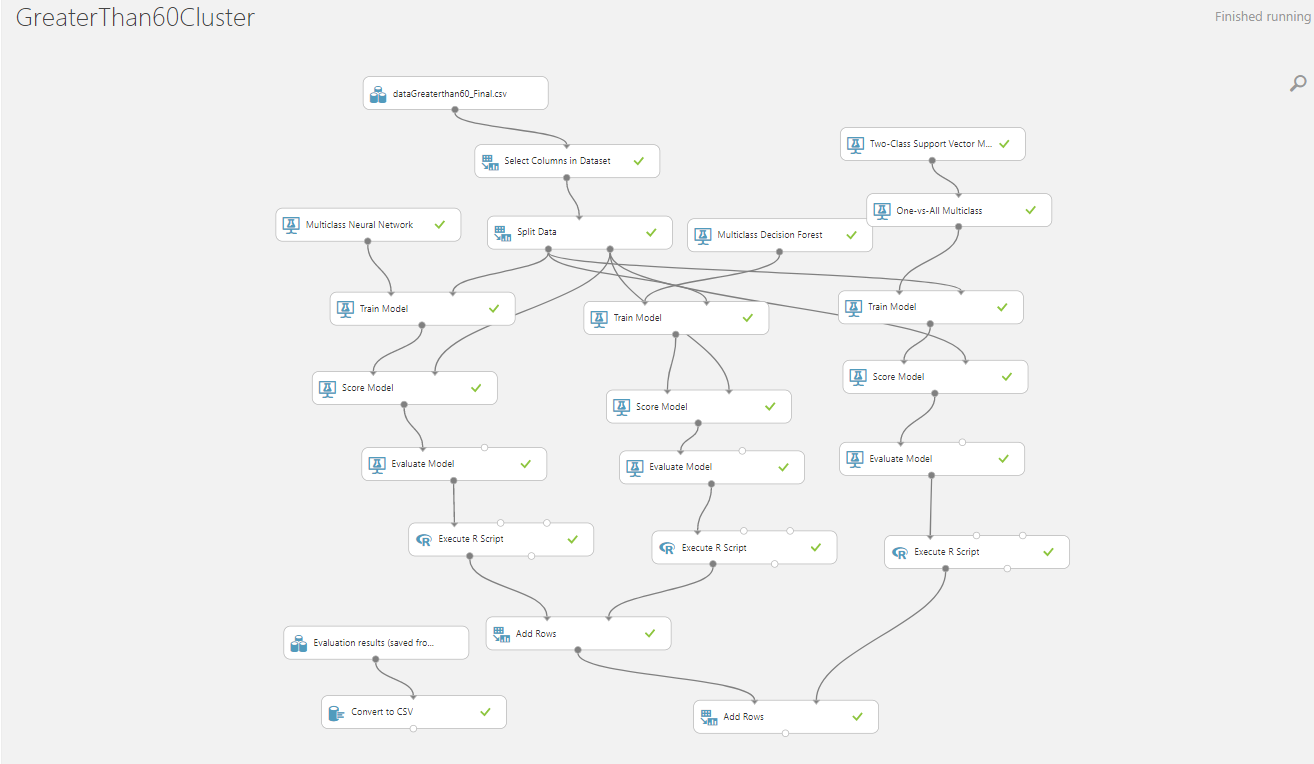
To classify the types of the Breast Cancer with the data of results of patients vitals we found that the age at diagnosis of the patient is normally distributed . So we clustered the dataset into two clusters (age group <= 62 and age group >62) and then build the two models using this clustered data. The algorithms used are:

1. Multi Class Neural Networks
2. Multi Class Decision Forest
3. 1 vs all Multi class SVM

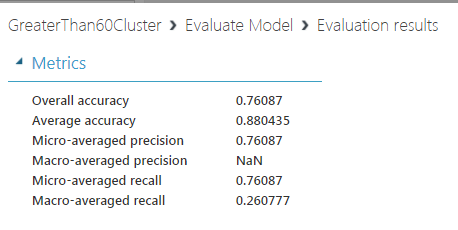
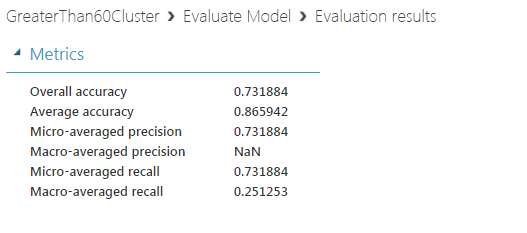
Following is the screenshot attached that shows the three algorithms implemented on the entire dataset:



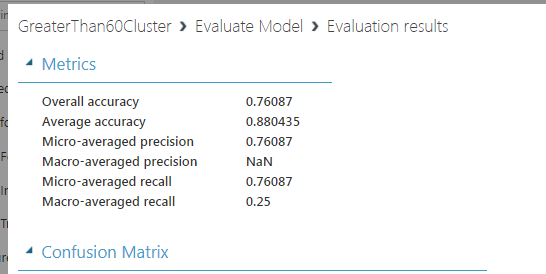
Following is the screenshot for the cluster with age greater than 62 cluster :



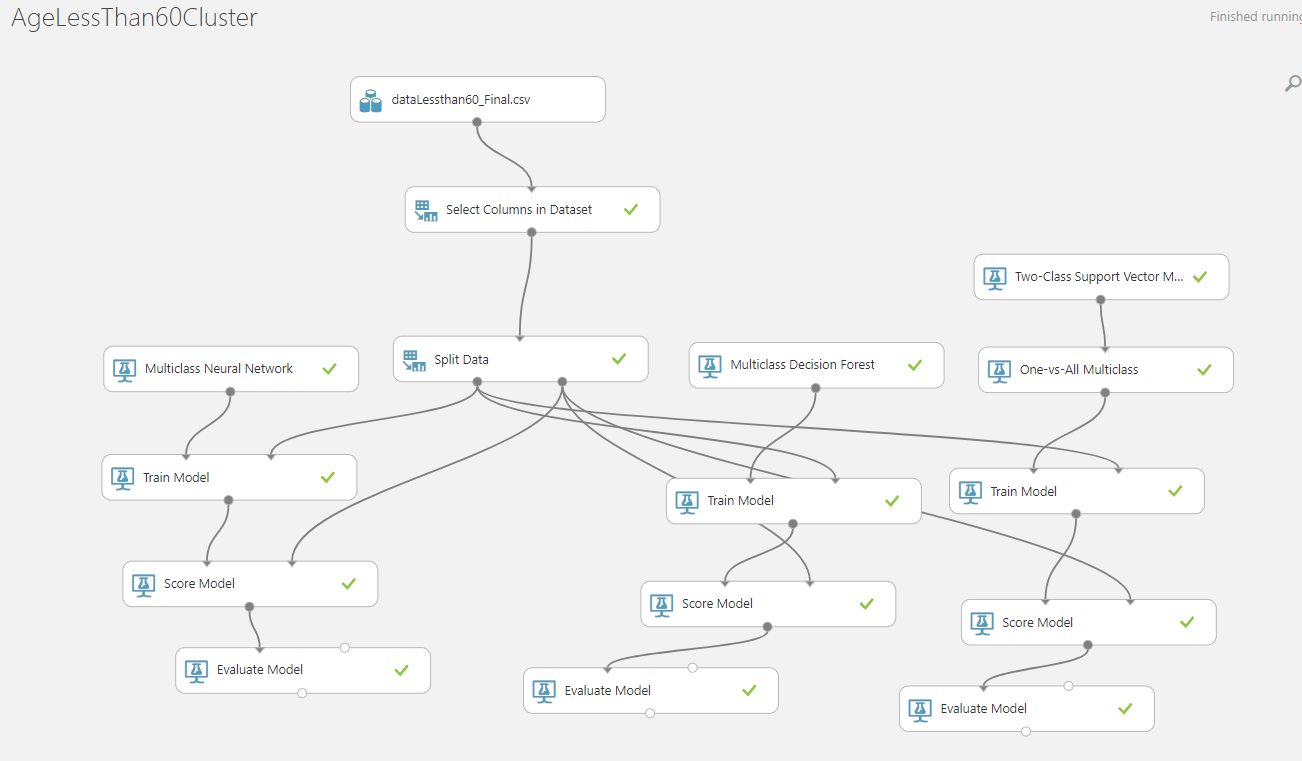
**Multiclass Neural networks** **Decision Forest**



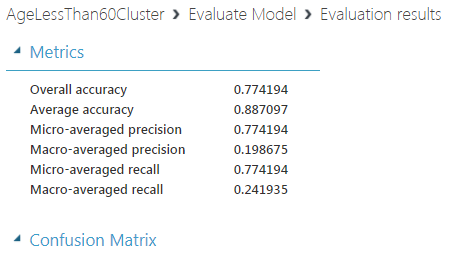
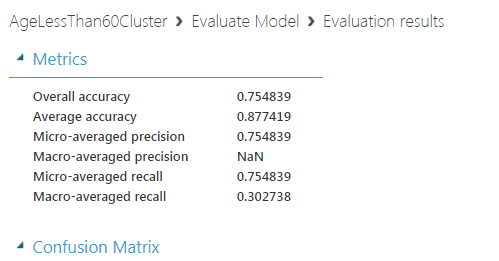
**1 vs all Multiclass SVM**



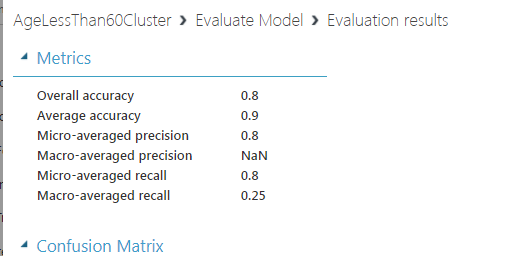
From these three models Decision Forest and SVM models gave similar results but we finally deployed with Decision Forest though either one of the above mentioned algorithms can be used.



**Neural network** **Decision forest**



**SVM**



The accuracy of the SVM gave better accuracy than the other models so we deployed SVM model for this category of cluster.

**Model Performance**

The model gave good performance in classifying **Invasive Ductal Carcinoma** and **Invasive Lobular Carcinoma** which is widely occurring type of cancer among women even over the years.

But it was not in giving similar results for other two types of cancer. We tried to balance the data using SMOTE and ROSE but the accuracy and precision was getting low when using sampled data.

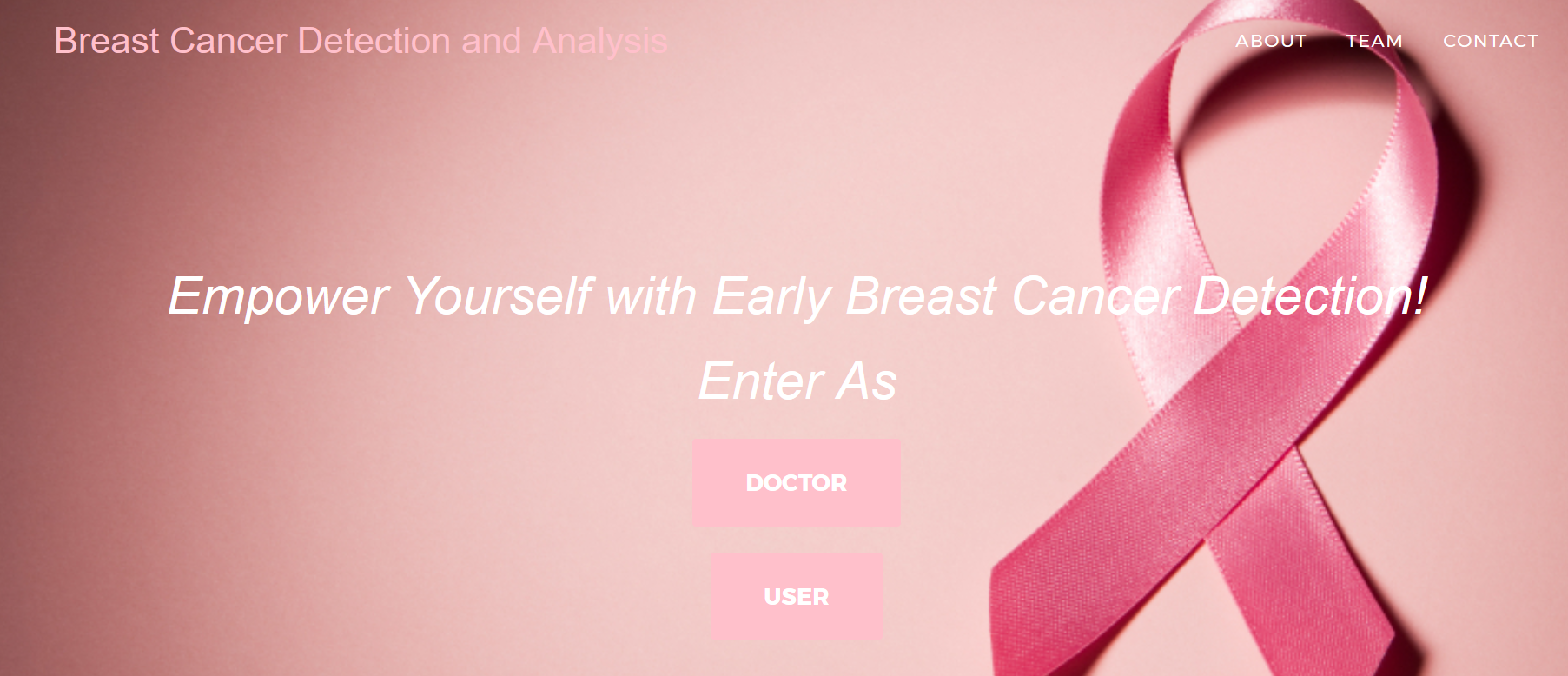
The app for dealing the two above mentioned use cases is developed in

* Python Flask
* HTML
* CSS
* Java Scripts

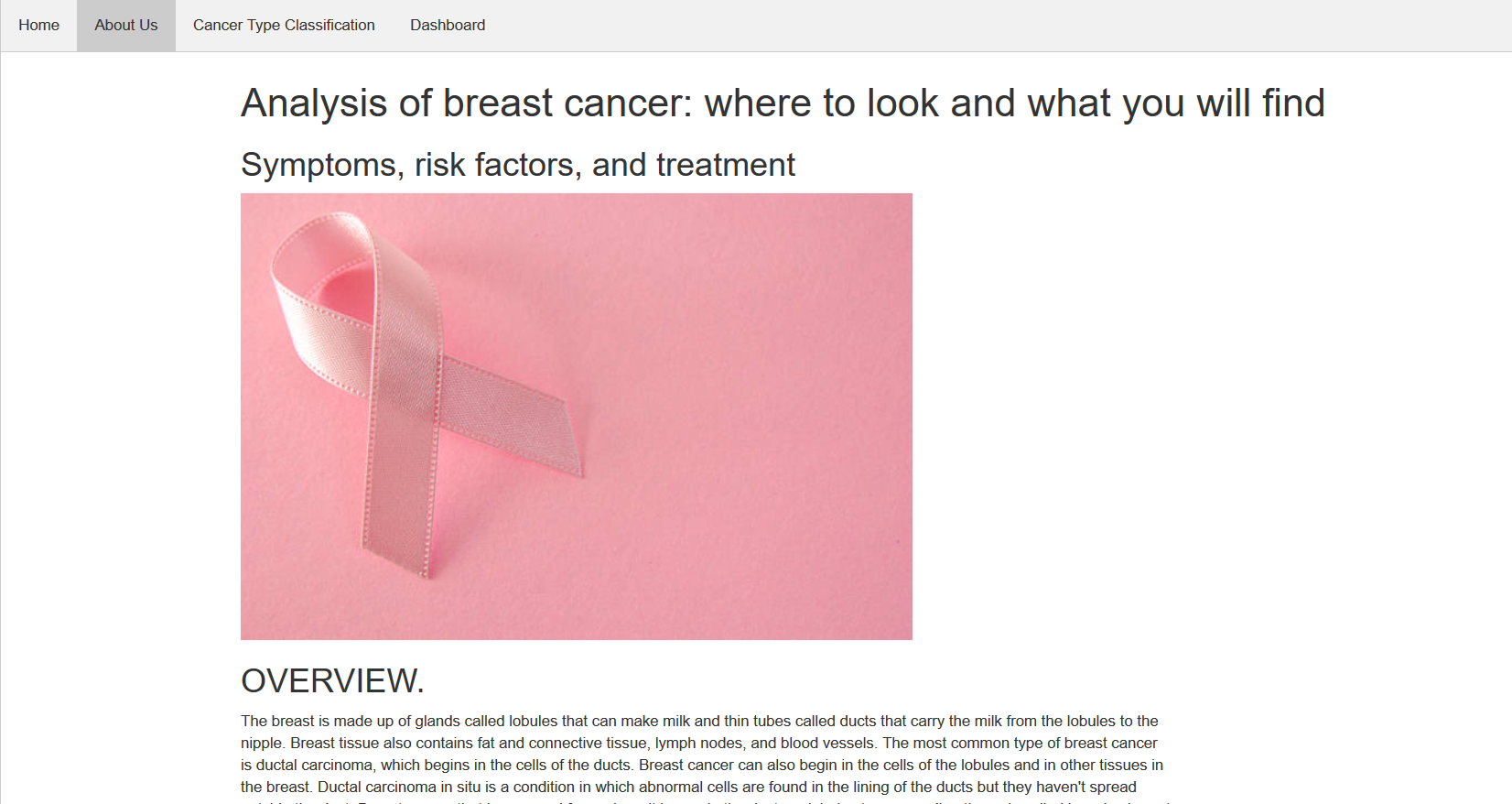
And deployed as a Web App in IBM Bluemix

<http://breastcancerassesment.mybluemix.net/>

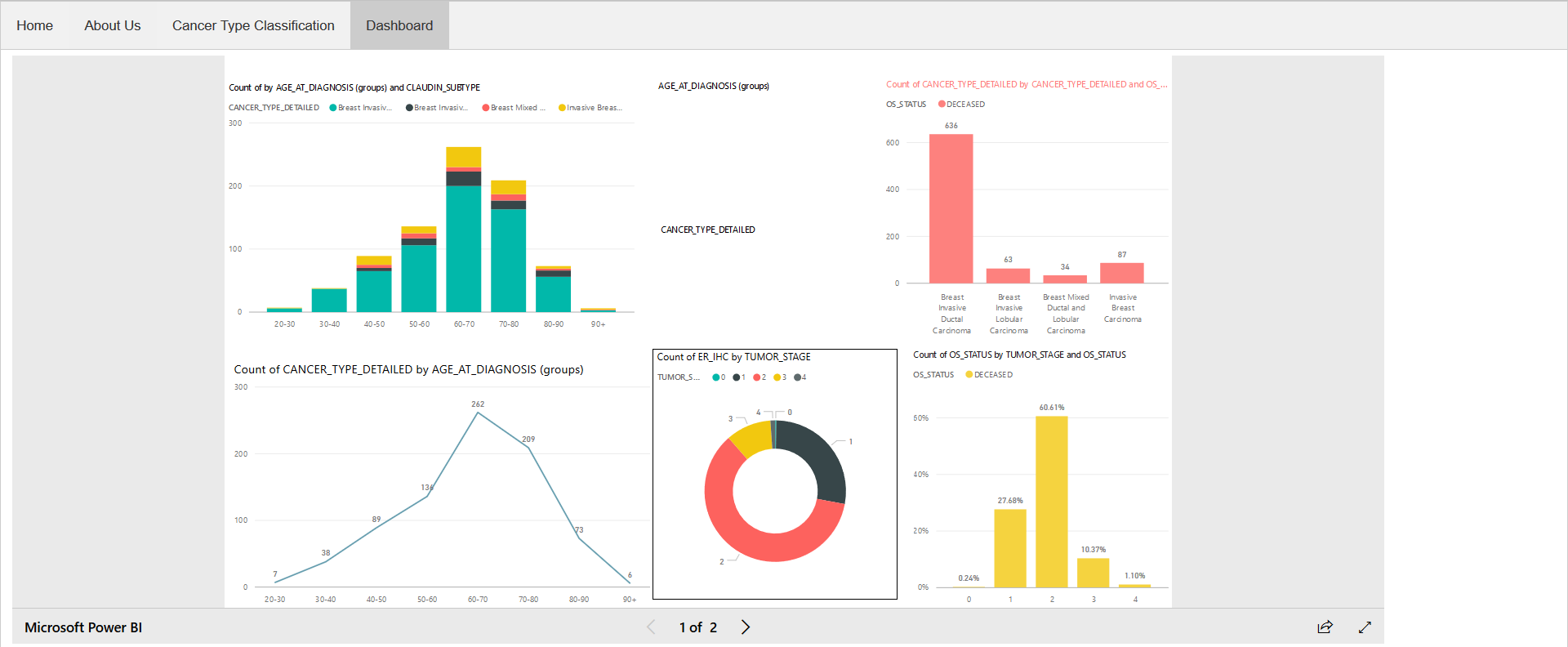
**The sample screens for the App looks like :**



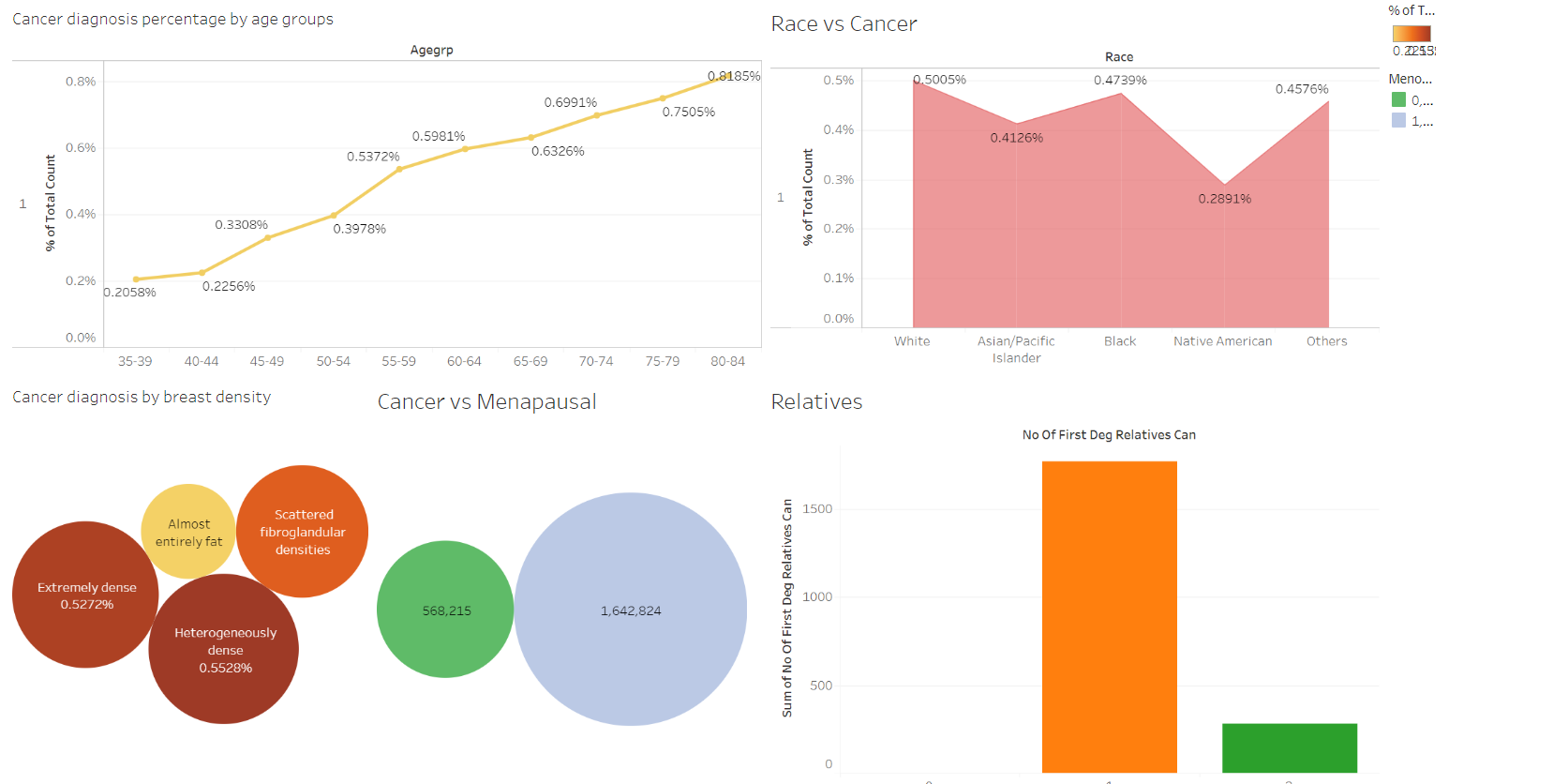
**One of the tabs of our application that describes about our app**



**POWER BI DASHBOARD INTEGRATED WITH THE APP**

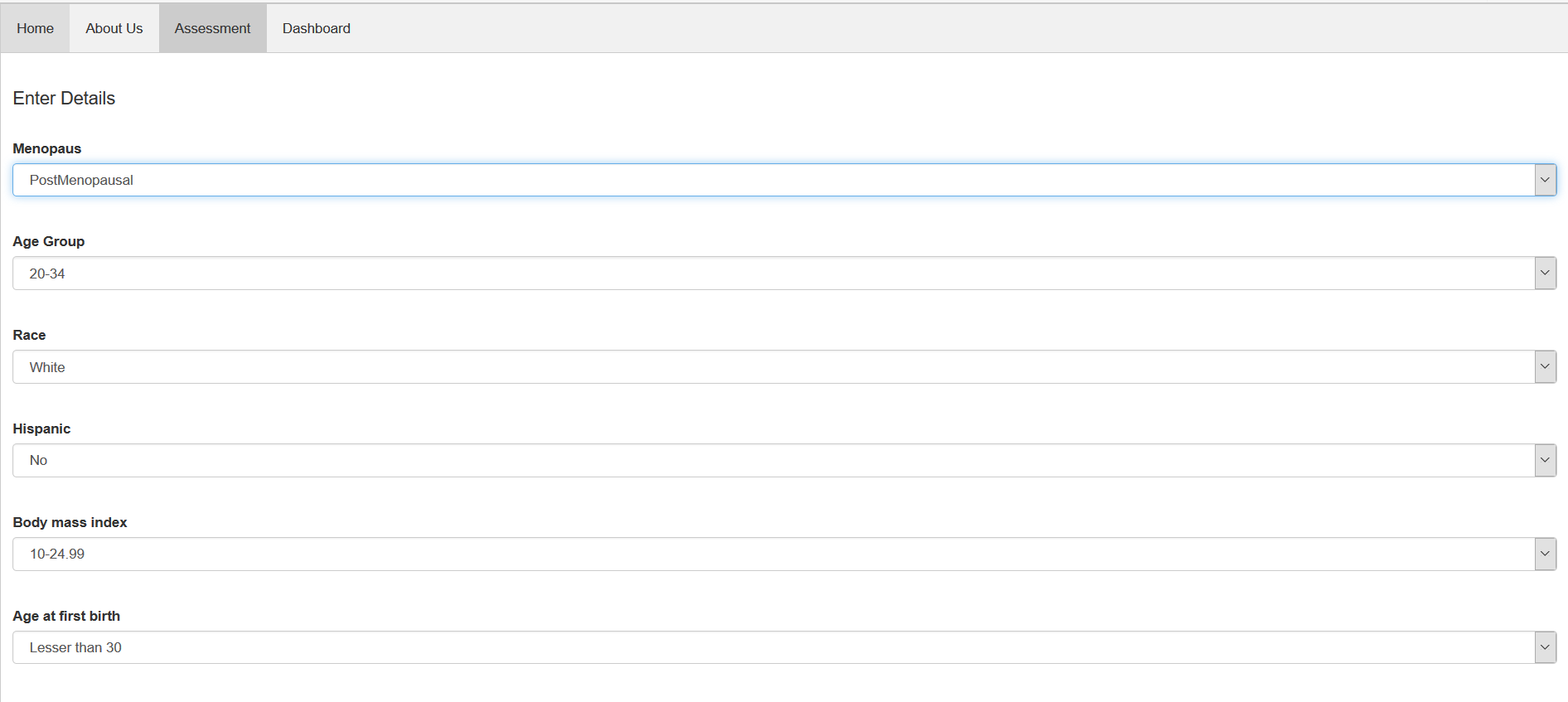
44

**TABLEAU DASBOARD INTEGRATED WITH THE APP**

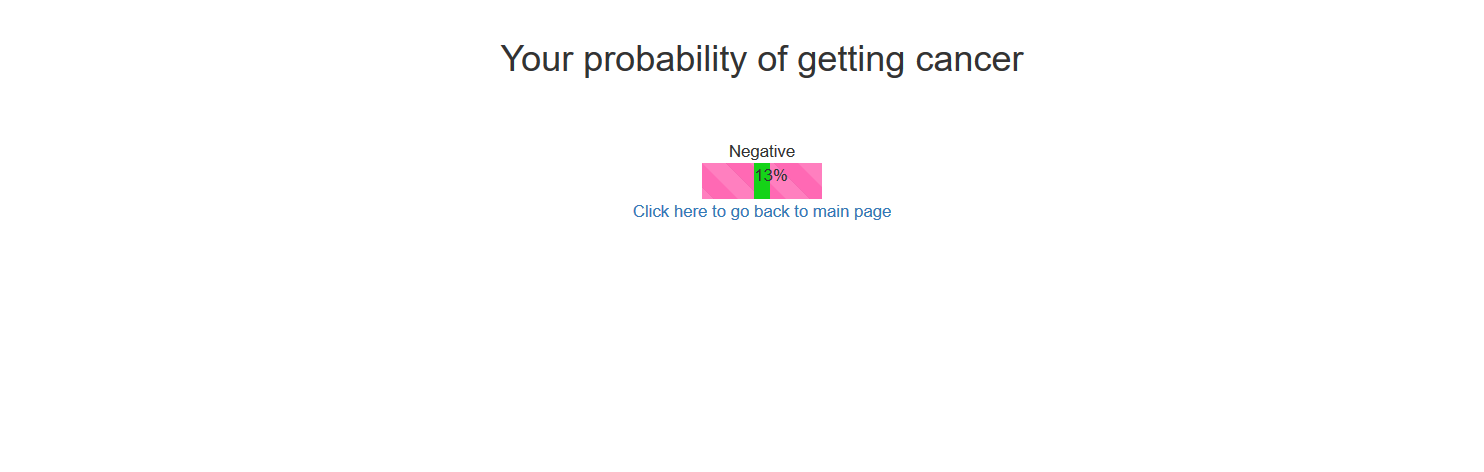
****

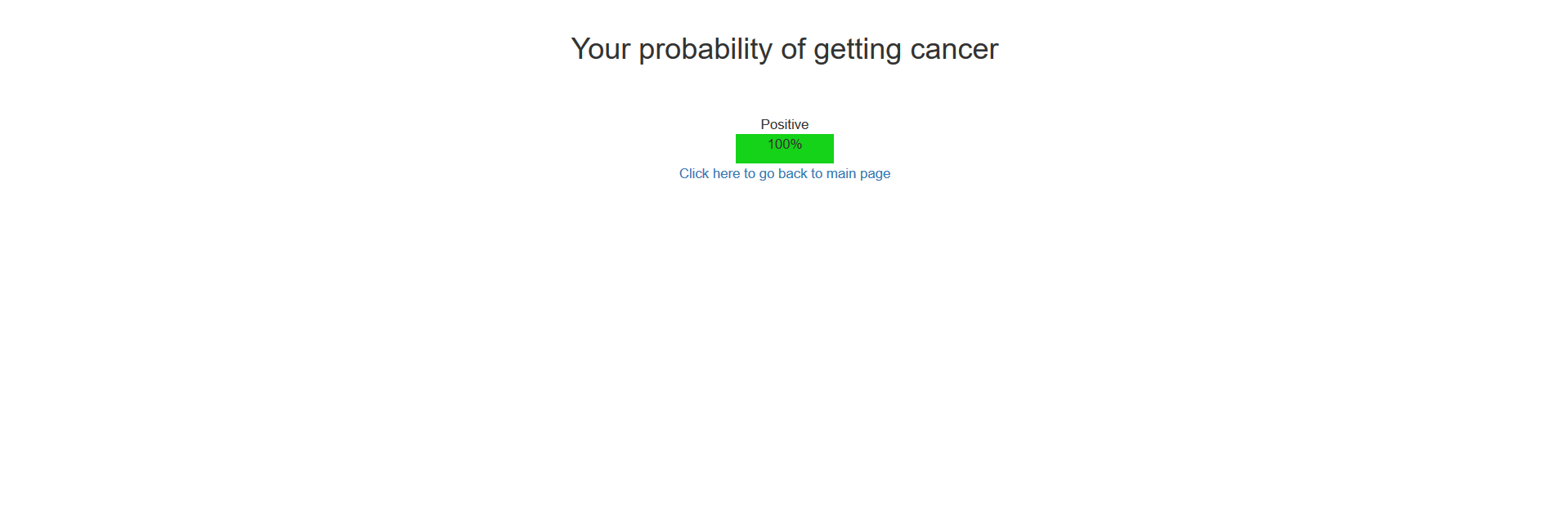
**USE CASE1:**

**Estimating risk of getting Breast Cancer**



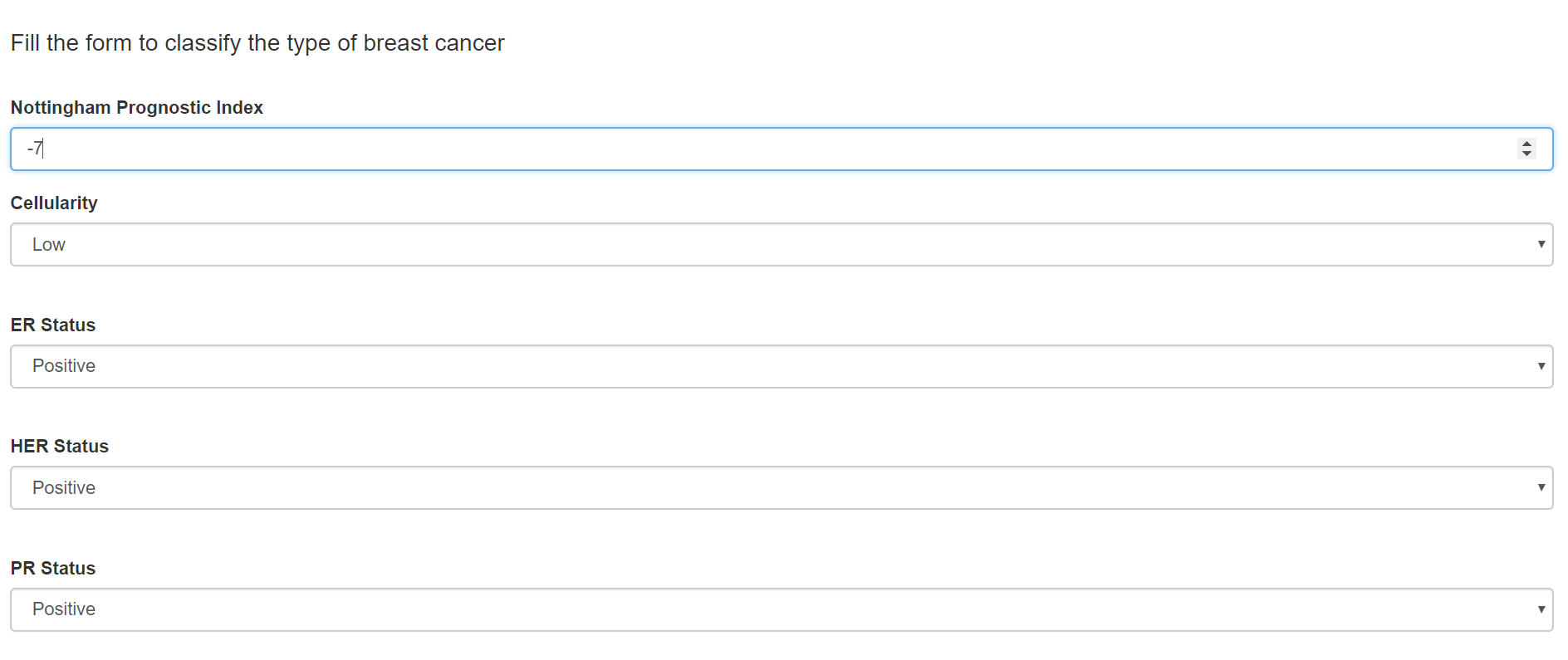
**OUTPUT:**





**USE CASE 2:**

**Classifying the tumor into types of Cancer:**

****

