# CAPSTONE PROJECT

HEALTHCARE

# DESCRIPTION

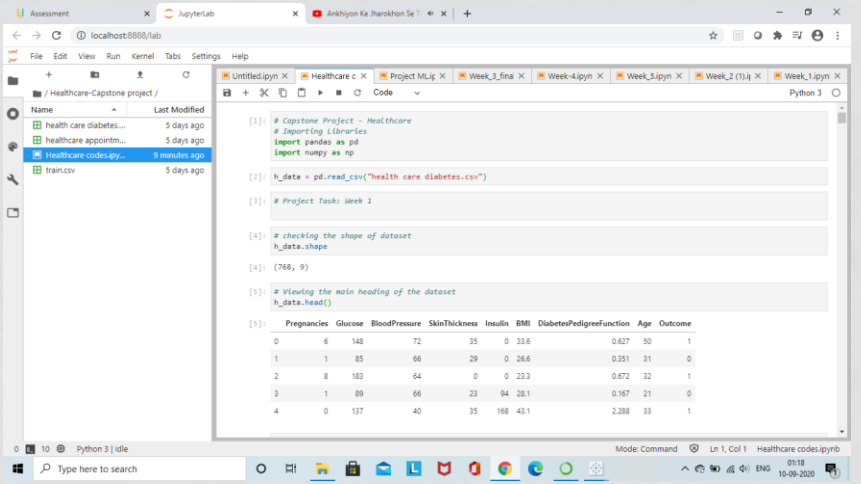
- NIDDK (National Institute of Diabetes and Digestive and Kidney Diseases) research creates knowledge about and treatments for the most chronic, costly, and consequential diseases.
- The dataset used in this project is originally from NIDDK. The objective is to predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset.
- Build a model to accurately predict whether the patients in the dataset have diabetes or not.

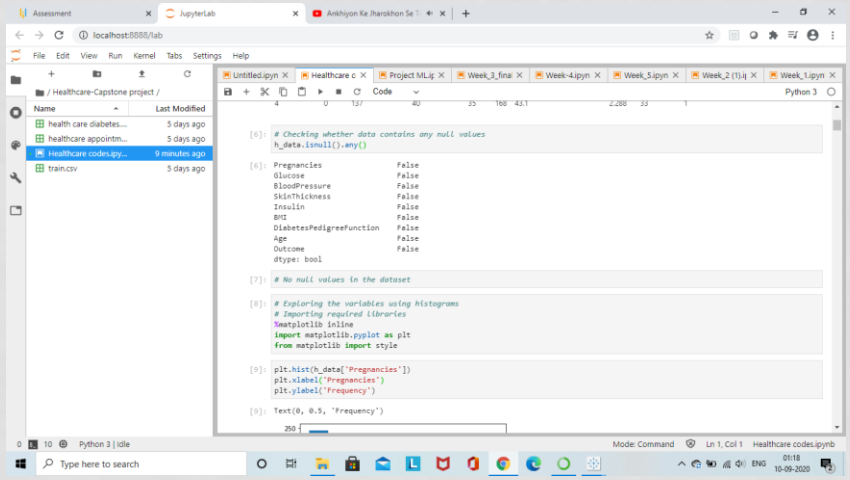
# PROJECT TASK

WEEK 1

# DATA EXPLORATION(TASK 1)

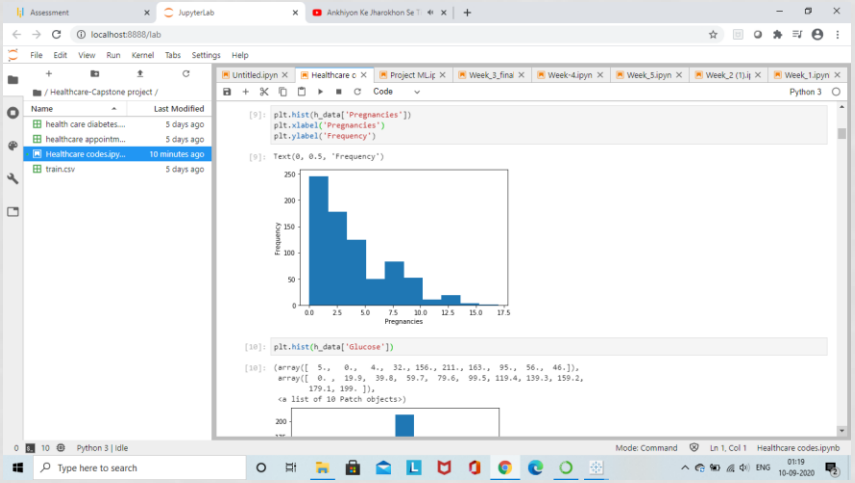
- Perform descriptive analysis. Understand the variables and their corresponding values. On the columns below, a value of zero does not make sense and thus indicates missing value:
- Glucose
- Blood Pressure
- Skin Thickness
- Insulin
- BMI

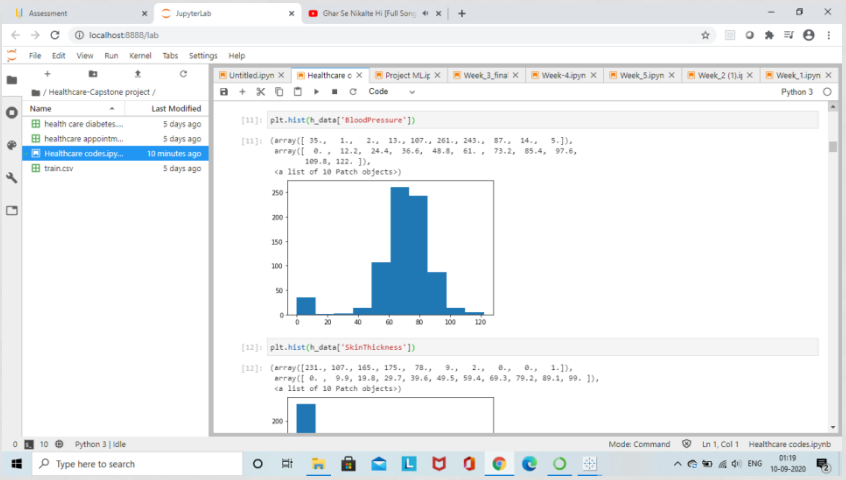


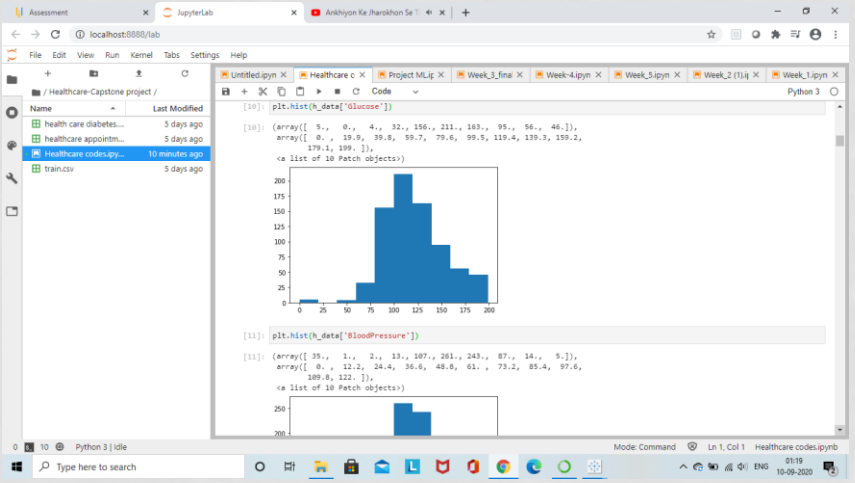



# TASK 2

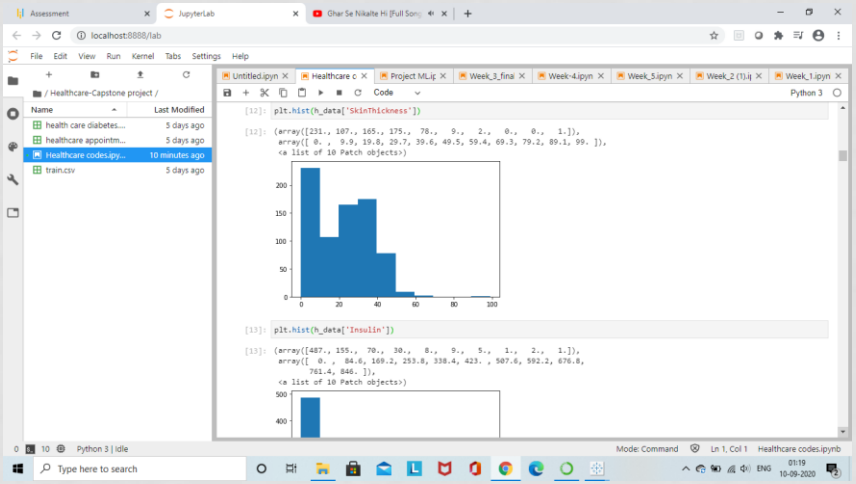
- Visually explore these variables using histograms. Treat the missing values accordingly.

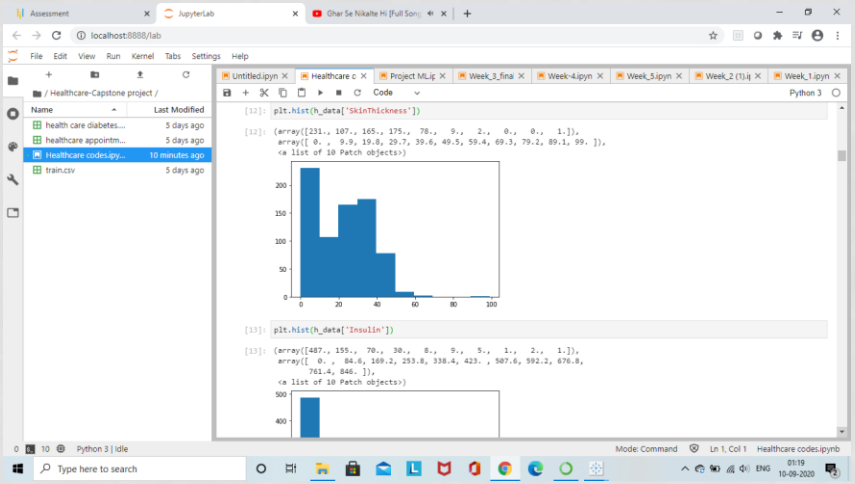


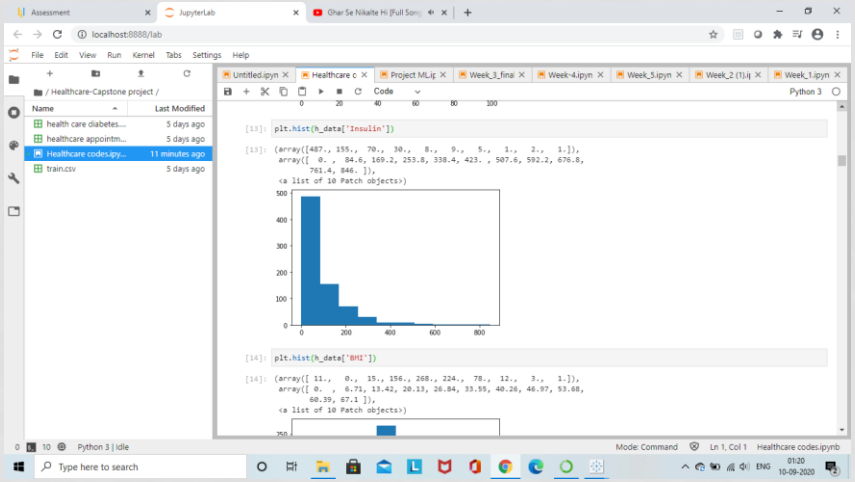



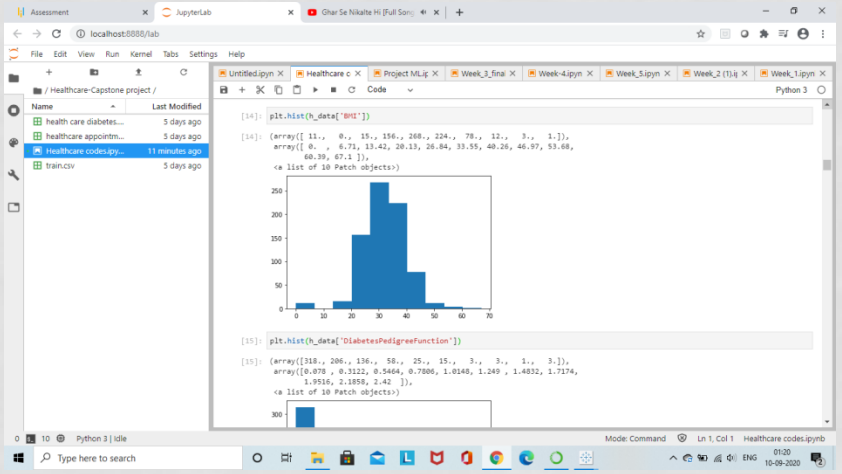



# TASK 2

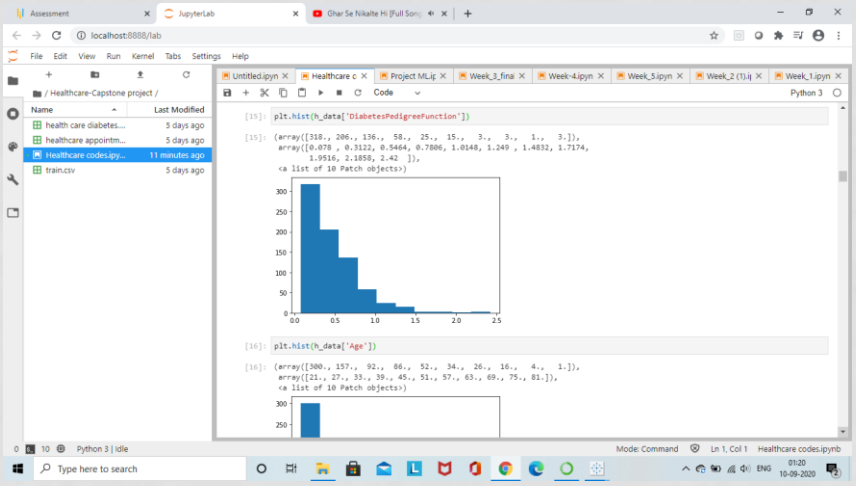

plt.hist(h_data['SkinThickness'])
plt.hist(h_data['Insulin'])

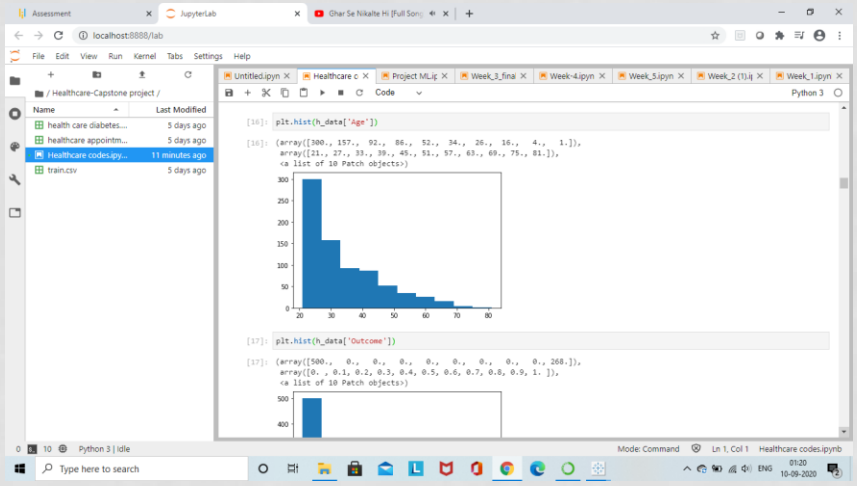

plt.hist(h_data['SkinThickness'])
plt.hist(h_data['Insulin'])

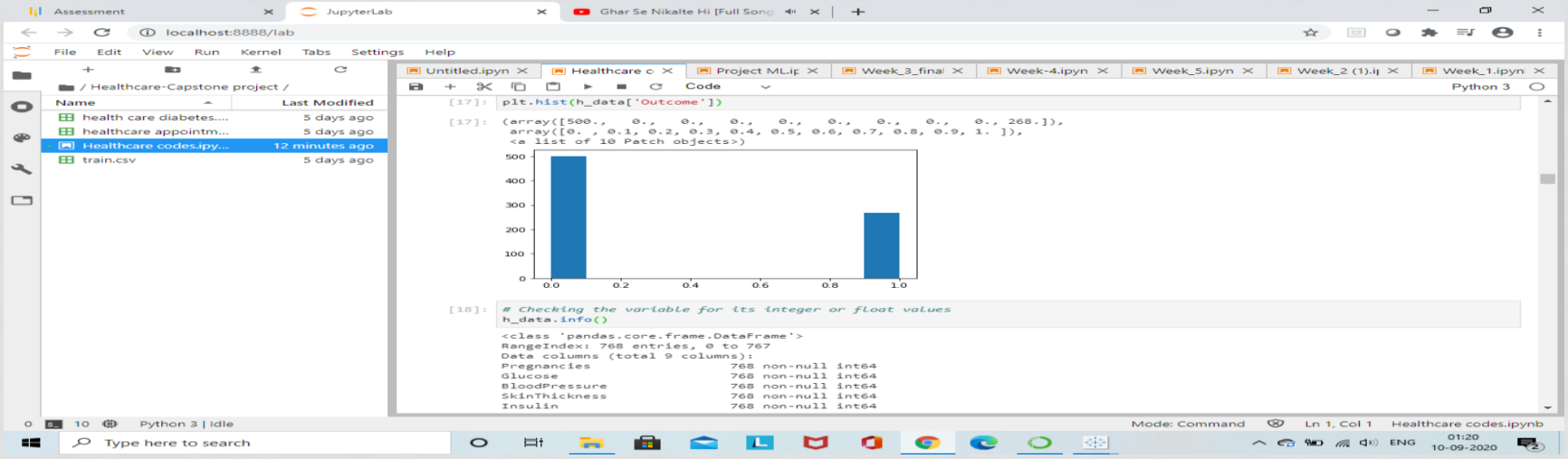

plt.hist(h_data['Insulin'])
plt.hist(h_data['BMI'])


plt.hist(h_data['BMI'])
plt.hist(h_data['DiabetesPedigreeFunction'])

# TASK 2

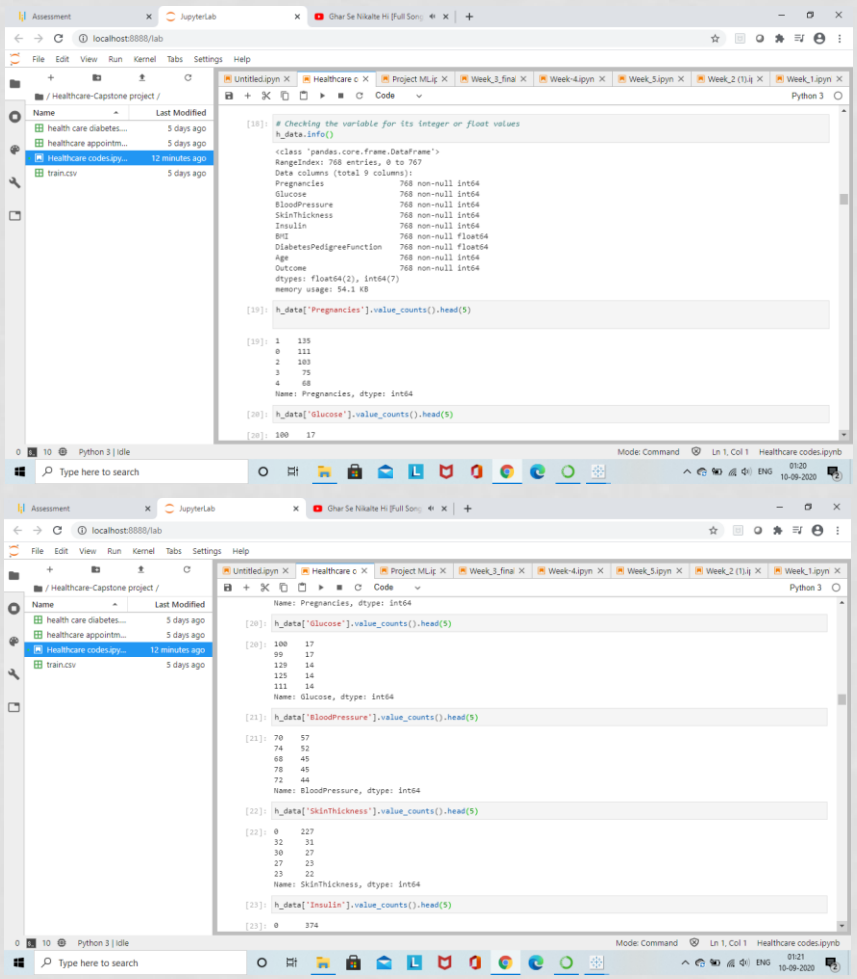





Assessment
JupyterLab
Ghar Se Nikalte Hi [Full Song]
localhost:8888/lab
File Edit View Run Kernel Tabs Settings Help
/ Healthcare-Capstone project /
Name
Last Modified
health care diabetes.... 5 days ago
healthcare appointm... 5 days ago
Healthcare codes.ipy... 12 minutes ago
train.csv 5 days ago
Untitled.ipyn
Healthcare c
Project ML.ip
Week_3_final
Week-4.ipyn
Week_5.ipyn
Week_2 (1).ip
Week_1.ipyn
Code
Python 3
[17]: plt.hist(h_data['Outcome'])
[17]: (array([500., 0., 0., 0., 0., 0., 0., 0., 0., 268.]),
array([0. , 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1. ]),
<a list of 10 Patch objects>)
[18]: # Checking the variable for its integer or float values
h_data.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 768 entries, 0 to 767
Data columns (total 9 columns):
Pregnancies 768 non-null int64
Glucose 768 non-null int64
BloodPressure 768 non-null int64
SkinThickness 768 non-null int64
Insulin 768 non-null int64
Python 3 | Idle
Mode: Command
Ln 1, Col 1
Healthcare codes.ipynb
Type here to search
ENG
01:20
10-09-2020

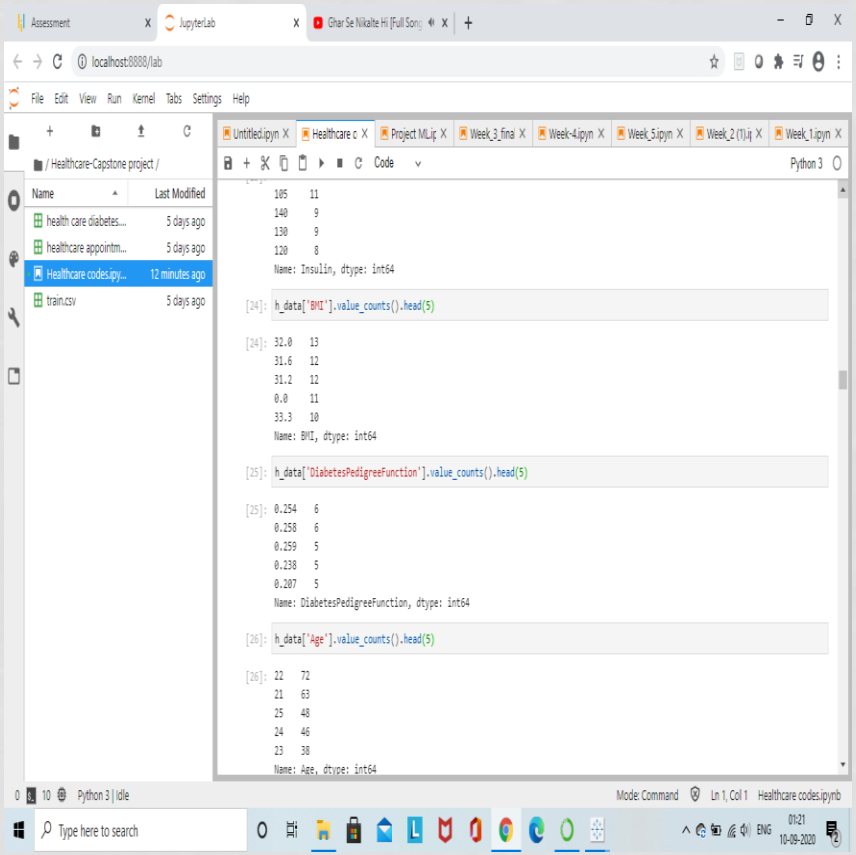
# TASK 3

- There are integer and float data type variables in this dataset. Create a count (frequency) plot describing the data types and the count of variables.



# TASK 3

- There is no null values in the dataset.
- There are two float variables and rest are integer variables including Target variable.

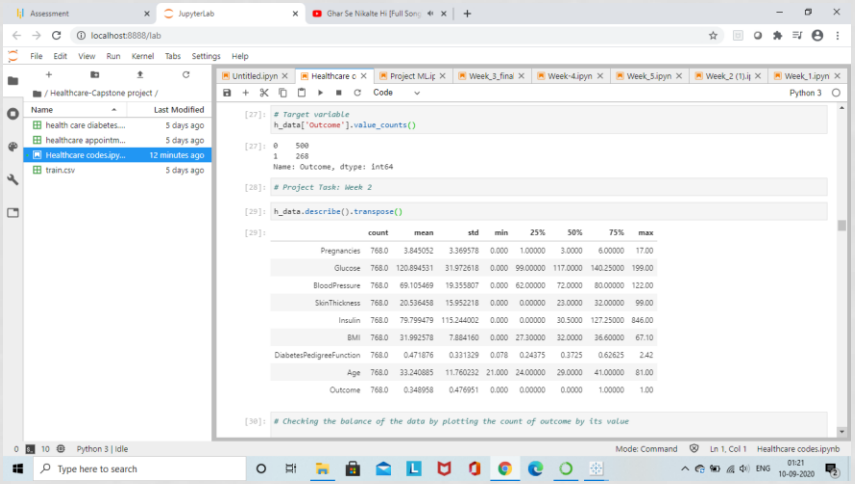


# PROJECT TASK

WEEK 2

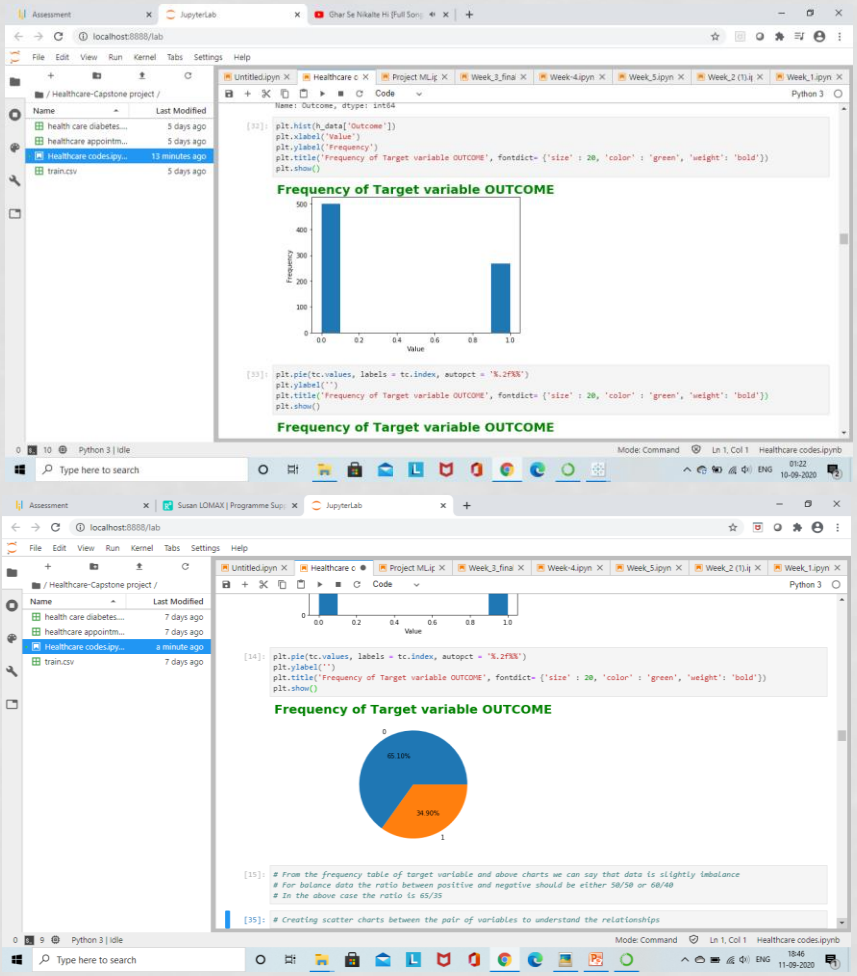
# TASK 1

- Check the balance of the data by plotting the count of outcomes by their value. Describe your findings and plan future course of action.



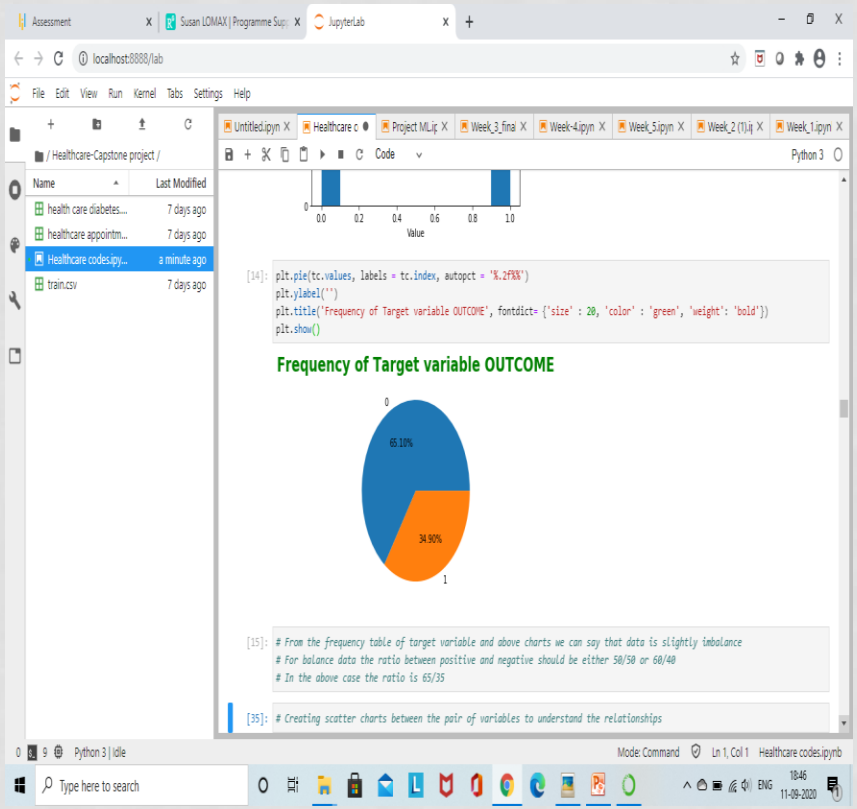
# TASK 1

- Out of 768 people, without diabetes people are 500 and with diabetes people are 268.
- From the pie chart it is clear that without diabetes population is 65% and with diabetic population is 35% approx.



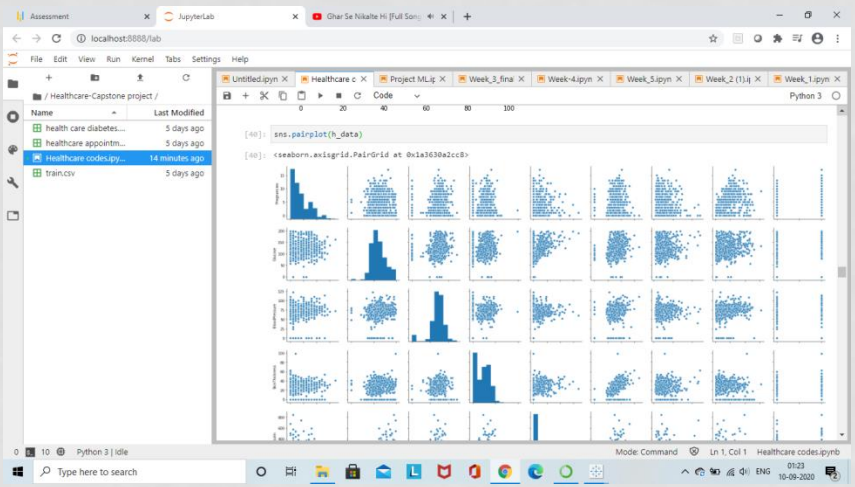
# TASK 1

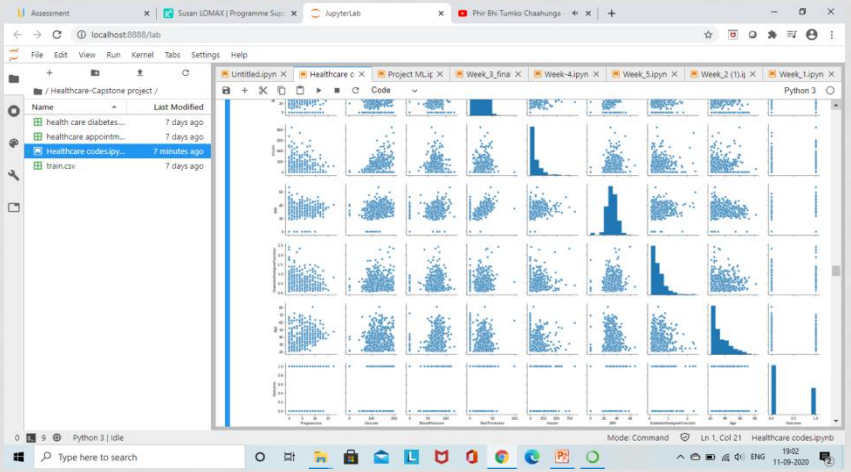
- As we can see in this pie chart, data is slightly imbalanced, as perfectly balanced data has 50/50 or 60/40 distribution. Here, the ratio in data is 65/35.



# TASK 2

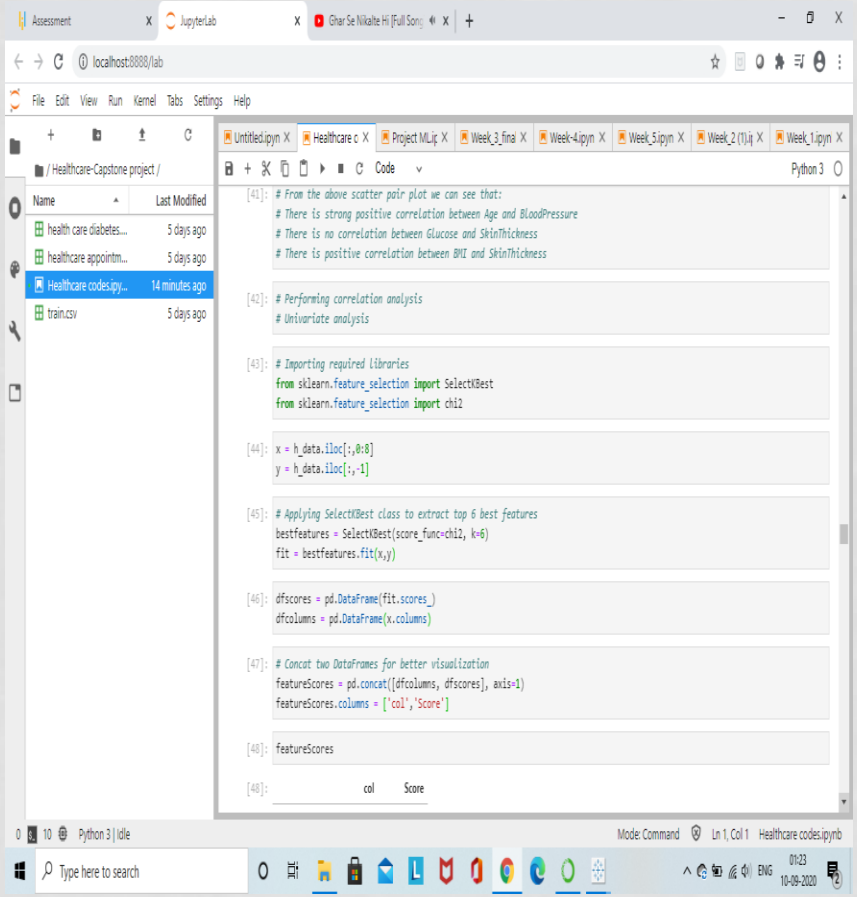
- Create scatter charts between the pair of variables to understand the relationships. Describe your findings.



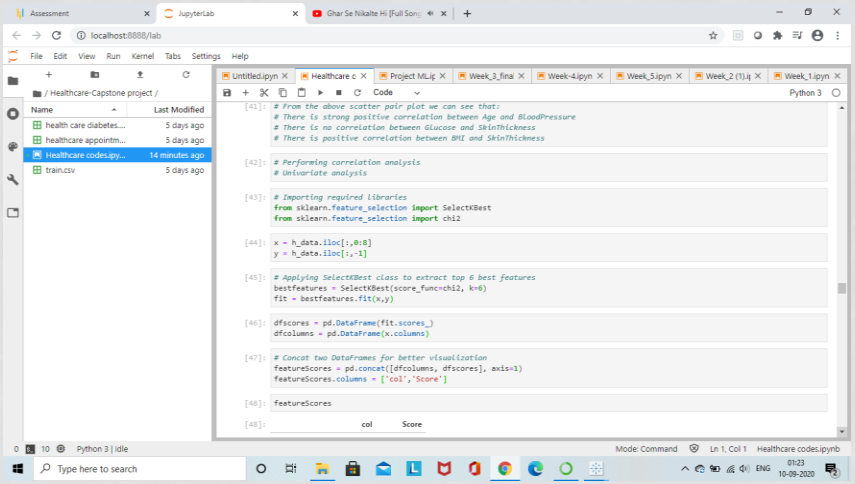
# TASK 2

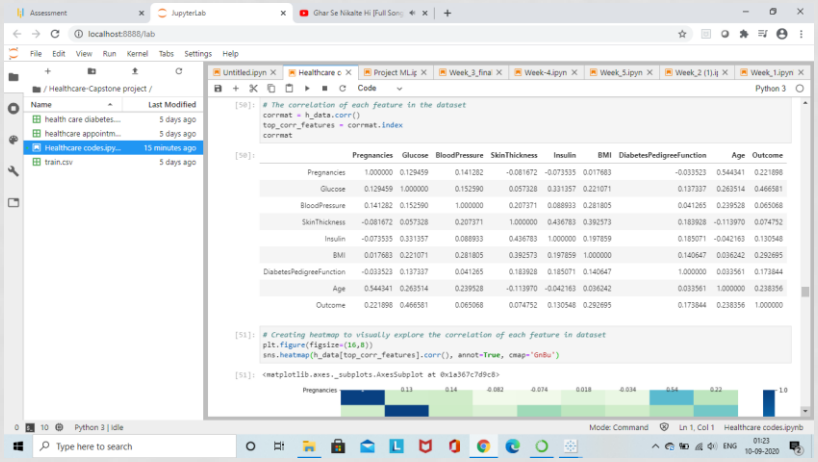
- There is a strong positive correlation between Age and Blood Pressure.
- There is no correlation between Glucose and Skin Thickness.

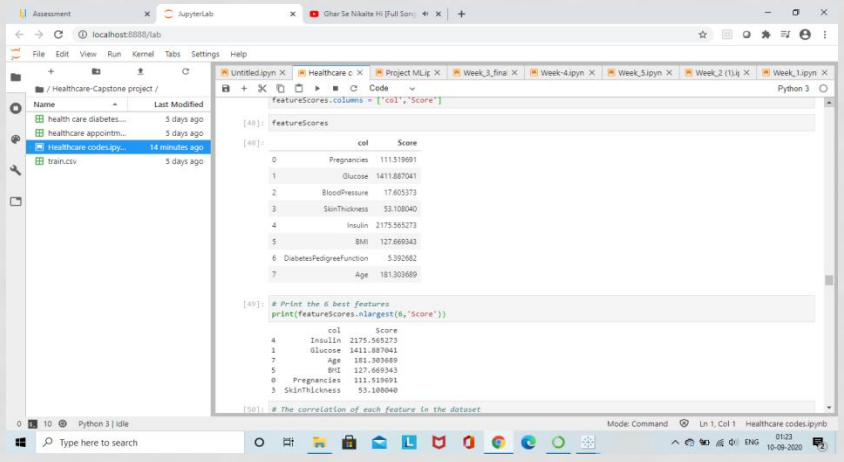


# TASK 3

- Perform correlation analysis. Visually explore it using a heat map.

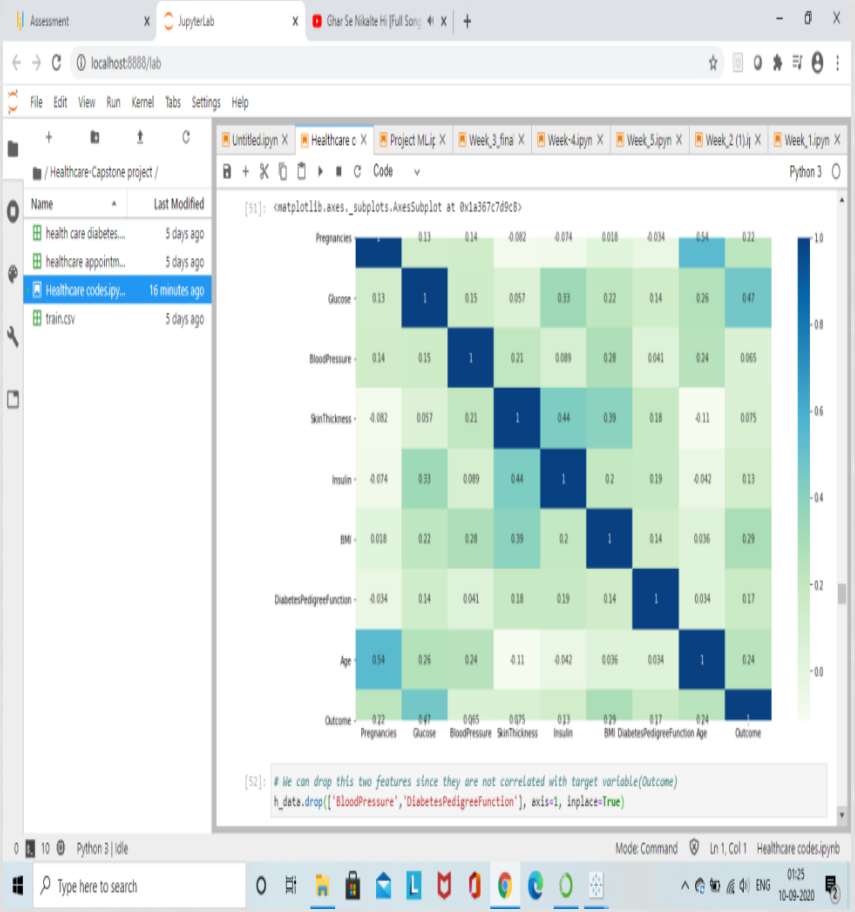


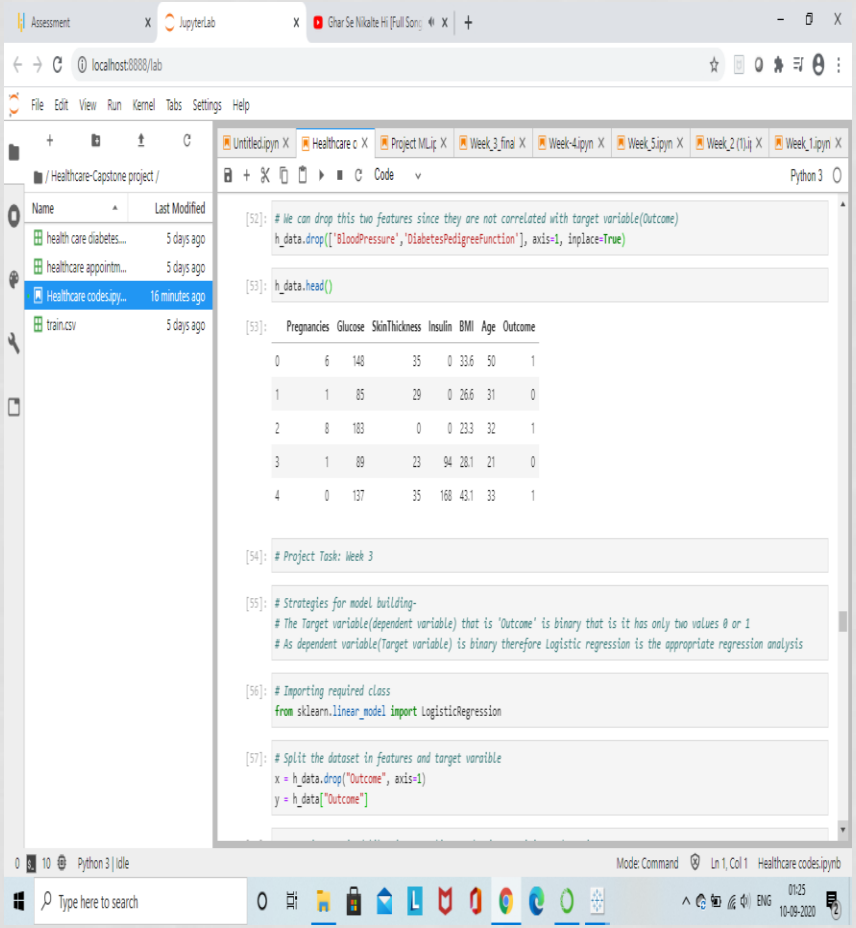
# TASK 3

- Strong correlation between independent features which are Insulin and Glucose, BMI and Skin Thickness, Age and Pregnancies.
- Skin Thickness and Insulin has almost no correlation with Pregnancies.



# TASK 3

- Outcome has the strong positive correlation with Glucose.
- Outcome has almost no correlation with Diabetes Pedigree Function.

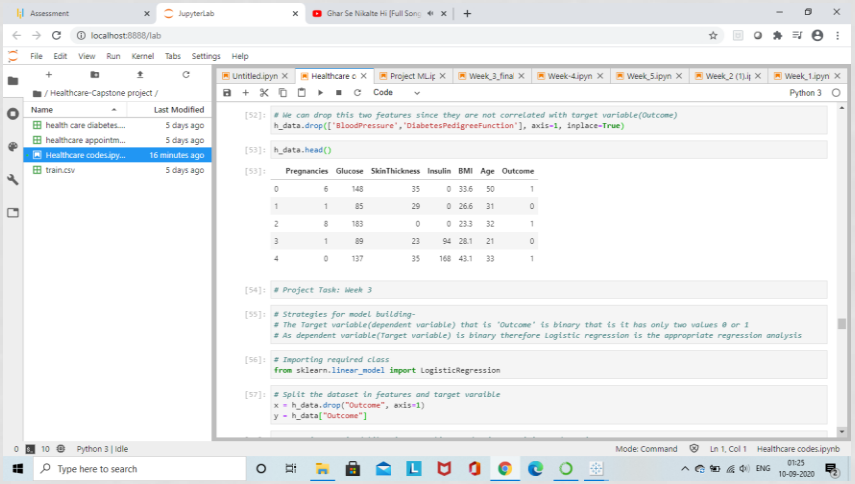


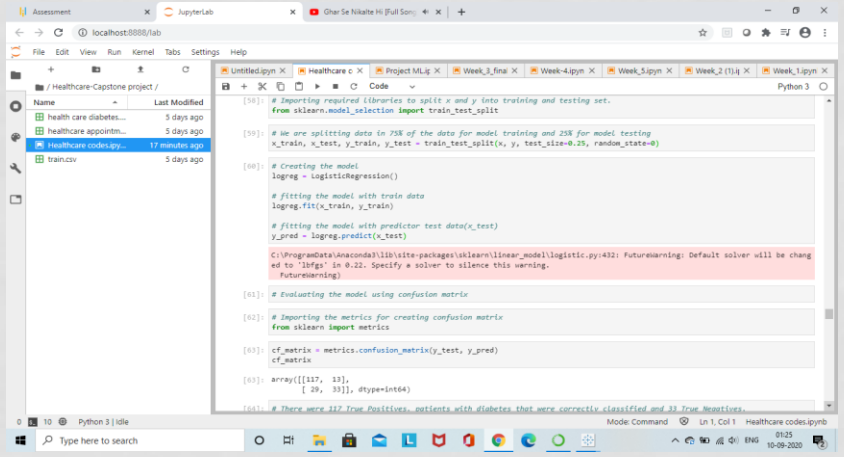
# PROJECT TASK

WEEK 3

# DATA MODELING (TASK 1)

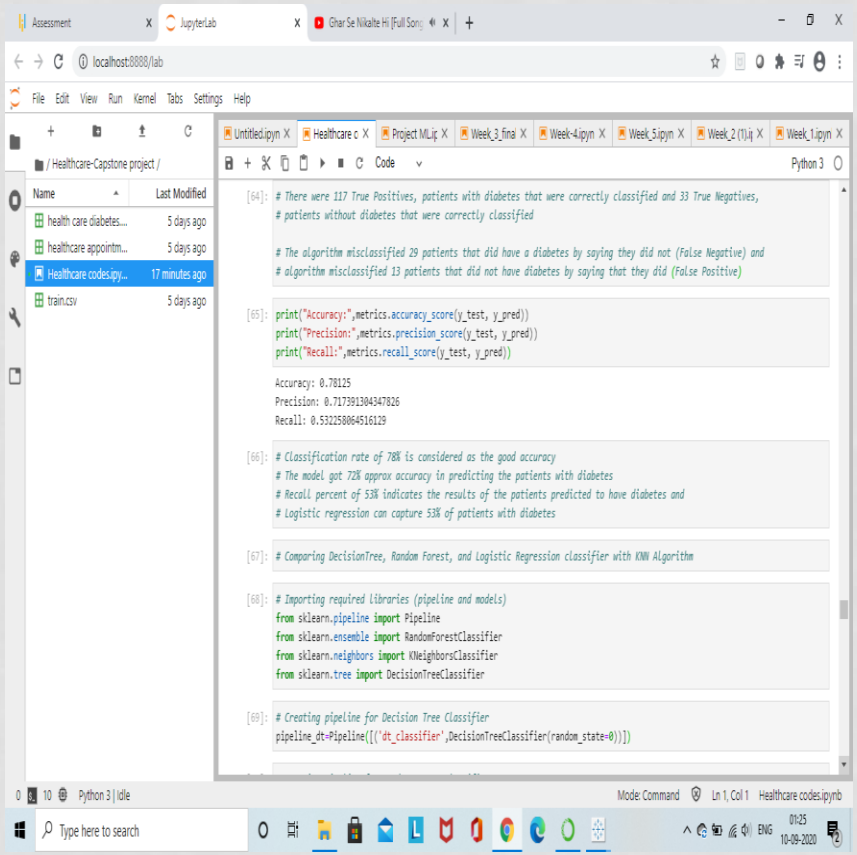
- Devise strategies for model building. It is important to decide the right validation framework. Express your thought process.



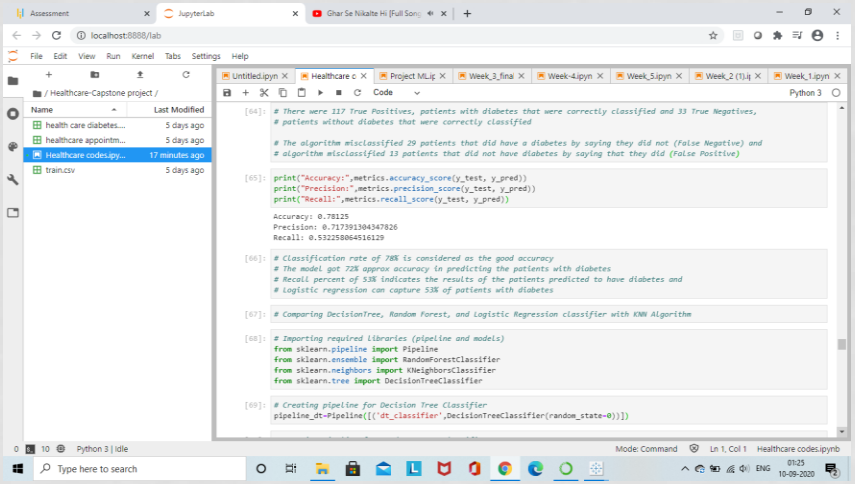
# TASK 1

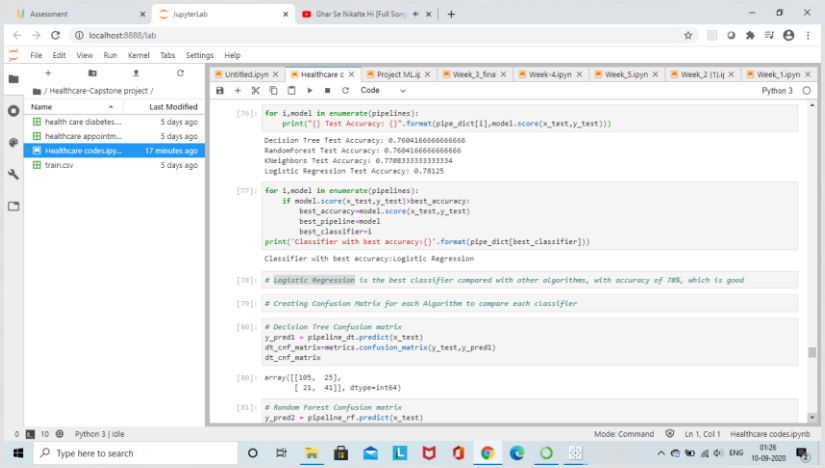
- Strategies for model building-
- The Target variable(dependent variable) that is 'Outcome' is binary that is it has only two values 0 or 1
- As dependent variable(Target variable) is binary therefore Logistic regression is the appropriate regression analysis

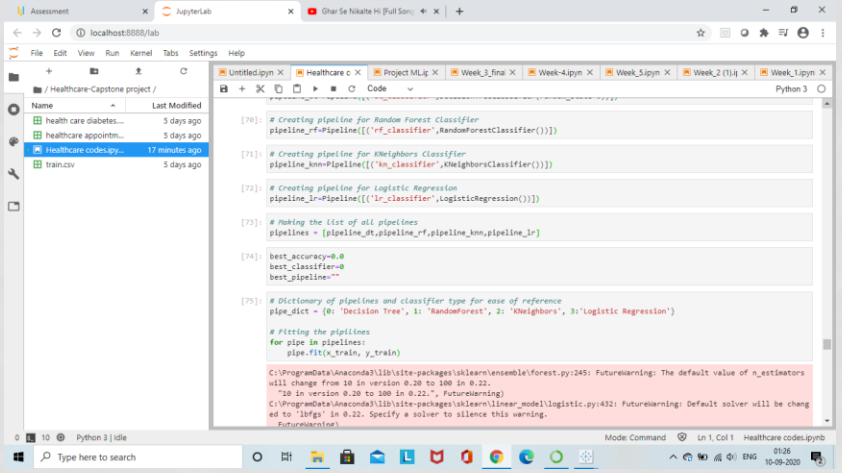


# TASK 2

- Apply an appropriate classification algorithm to build a model. Compare various models with the results from KNN algorithm.

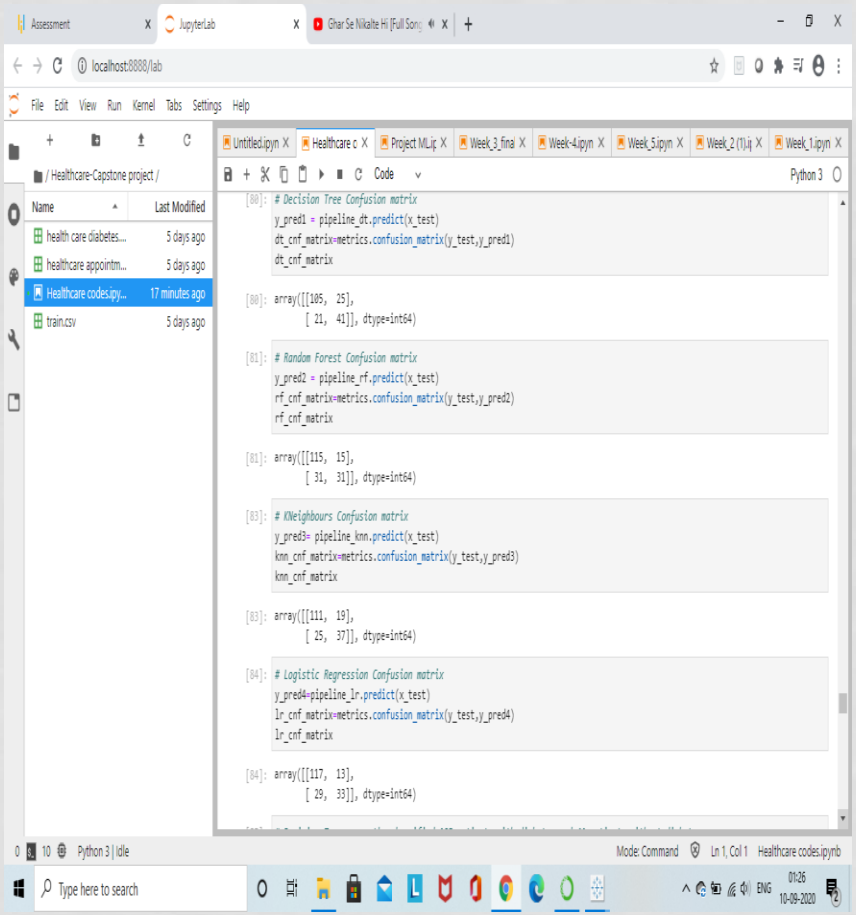


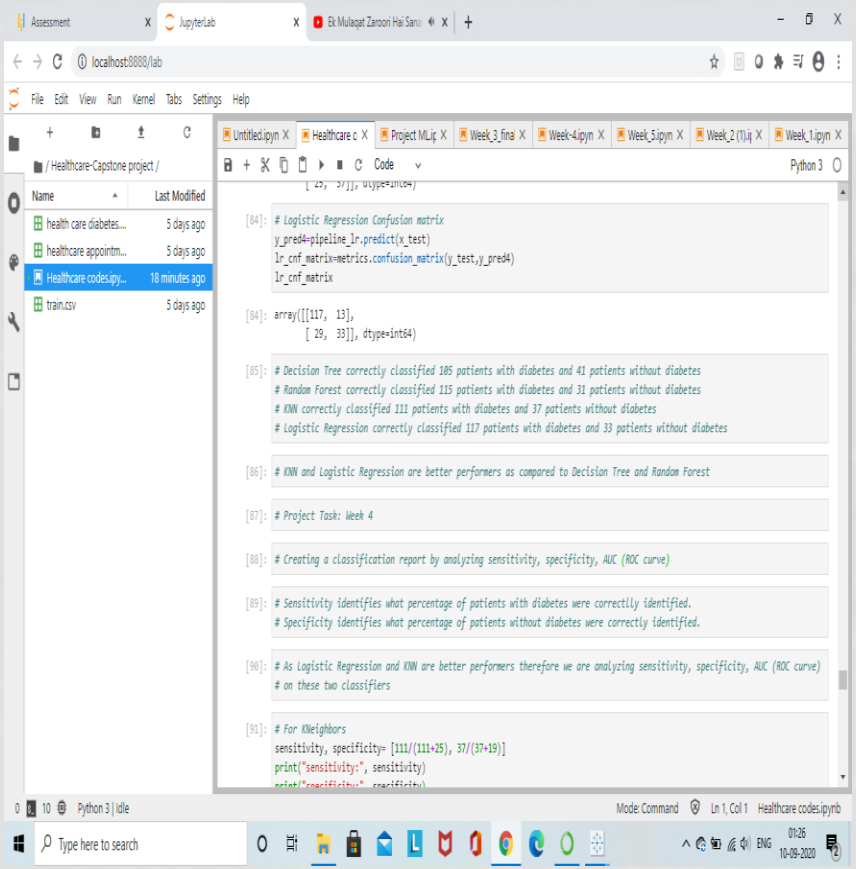
# TASK 2

- Decision Tree correctly classified 105 patients with diabetes and 41 patients without diabetes
- Random Forest correctly classified 115 patients with diabetes and 31 patients without diabetes
- KNN correctly classified 111 patients with diabetes and 37 patients without diabetes
- Logistic Regression correctly classified 117 patients with diabetes and 33 patients without diabetes



# TASK 2

- KNN and Logistic Regression are better performers as compared to Decision Tree and Random Forest

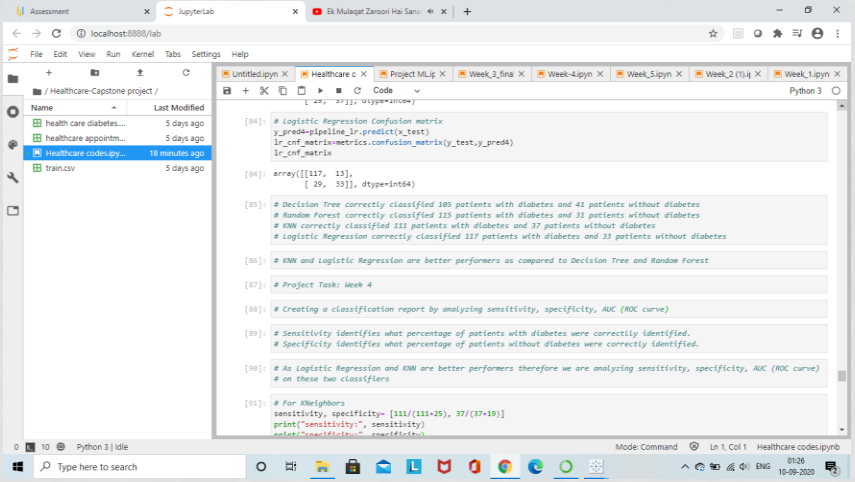


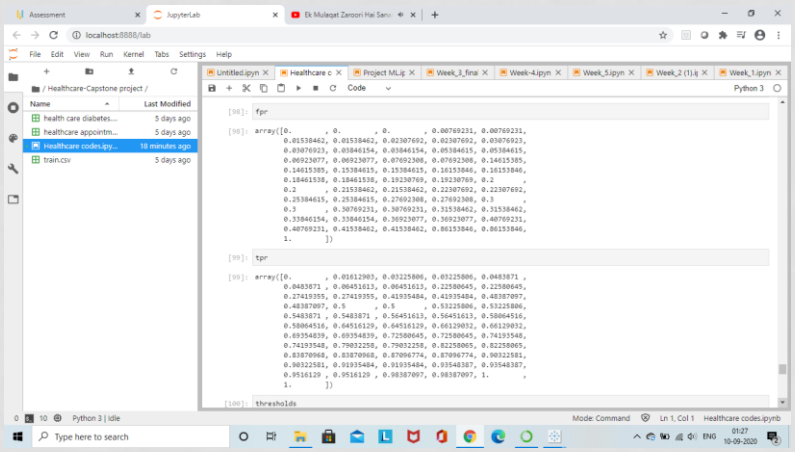
# PROJECT TASK

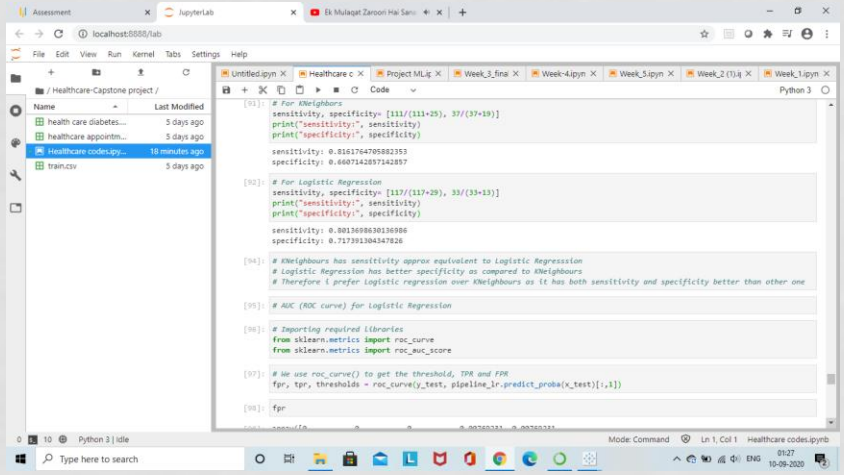
WEEK 4

# DATA MODELING (TASK 1)

- Create a classification report by analyzing sensitivity, specificity,AUC (ROC curve). Please be descriptive to explain values of the parameter you have used.

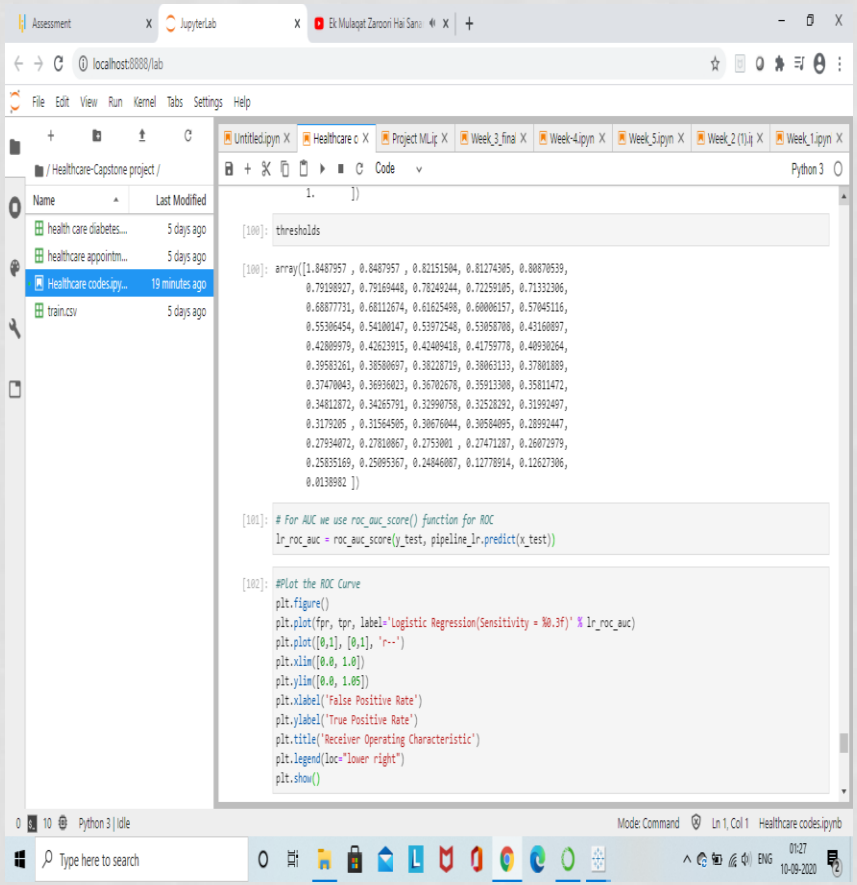


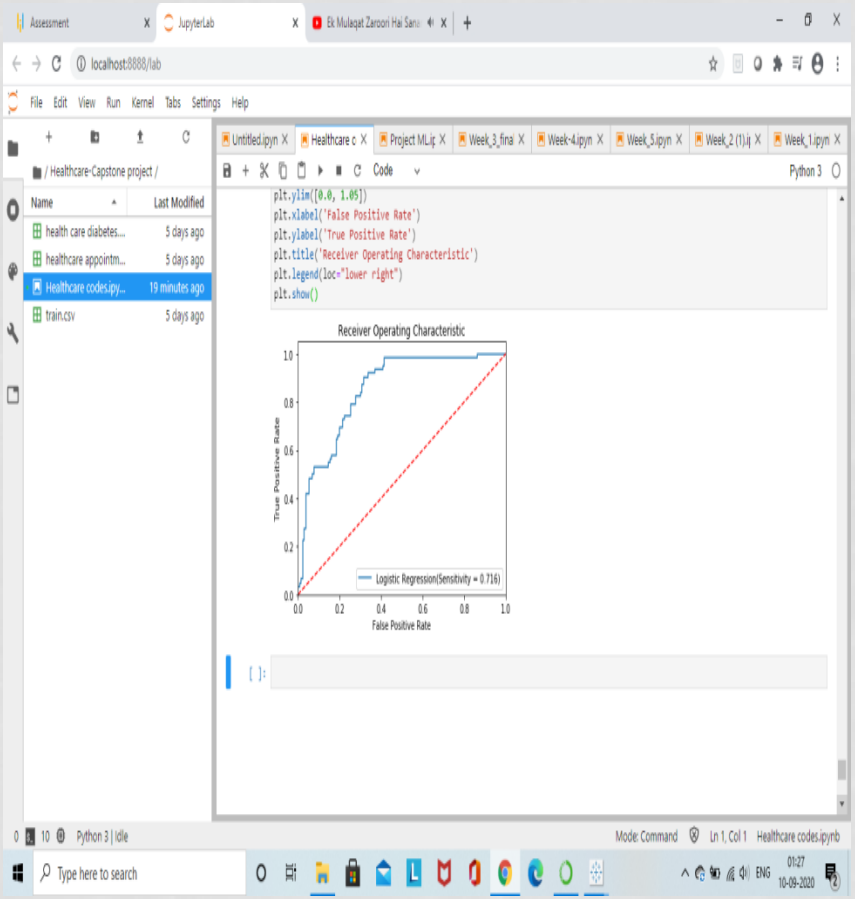
# TASK 1

- As Logistic Regression and KNN are better performers therefore we are analyzing sensitivity, specificity, AUC (ROC curve)on these two classifiers.



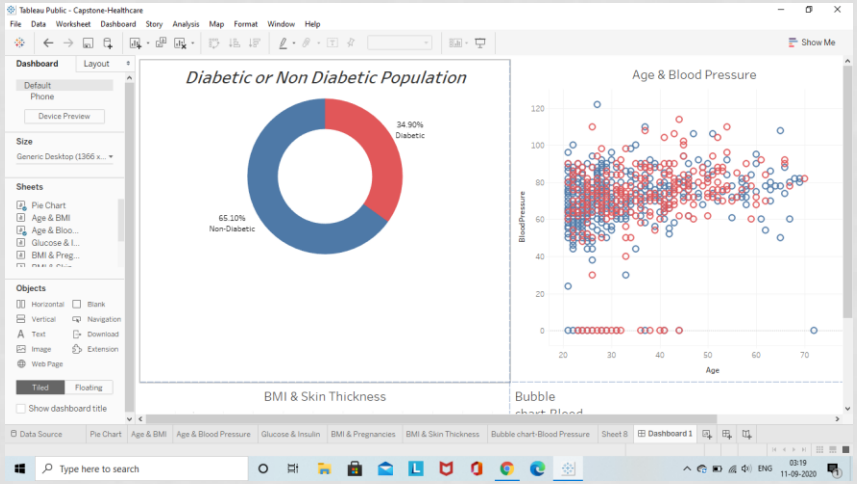
# TASK 1

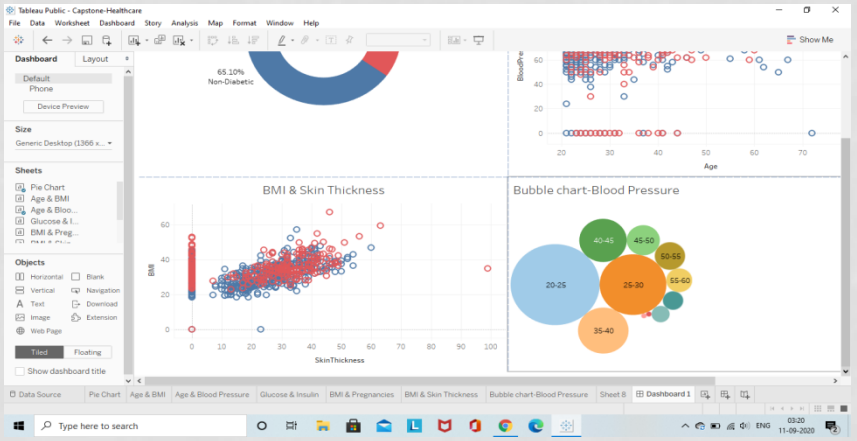
- KNeighbours has sensitivity approx equivalent to Logistic Regresssion
- Logistic Regression has better specificity as compared to KNeighbours
- Therefore i prefer Logistic regression over KNeighbours as it has both sensitivity and specificity better than other one



# DATA REPORTING (TASK 2)

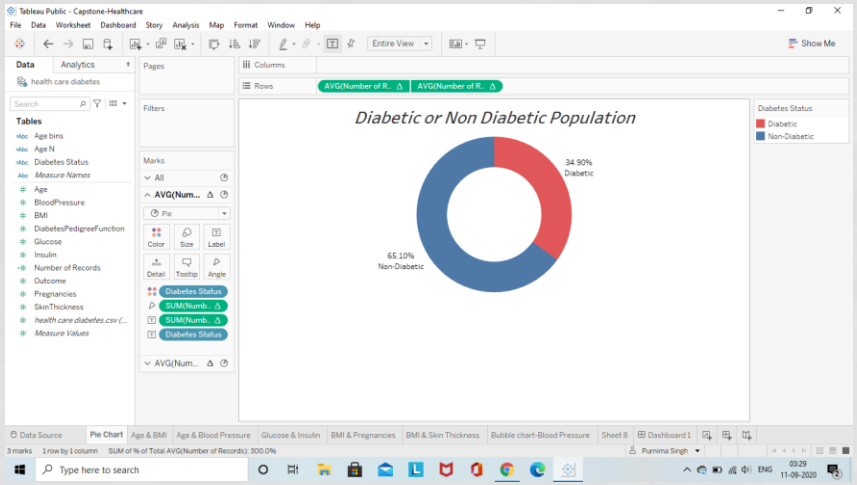
- Create a dashboard in tableau by choosing appropriate chart types and metrics useful for the business. The dashboard must entail the following:
- Pie chart to describe the diabetic or non-diabetic population.
- Scatter charts between relevant variables to analyze the relationships

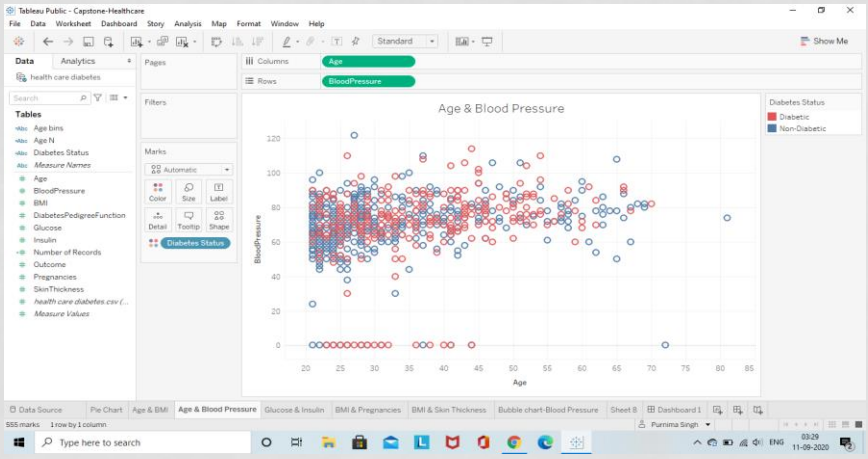


# TASK 2

- Histogram or frequency charts to analyze the distribution of the data.
- Heatmap of correlation analysis among the relevant variables.
- Create bins of these age values: 20-25, 25-30, 30-35, etc. Analyze different variables for these age brackets using a bubble chart.



# TASK 2

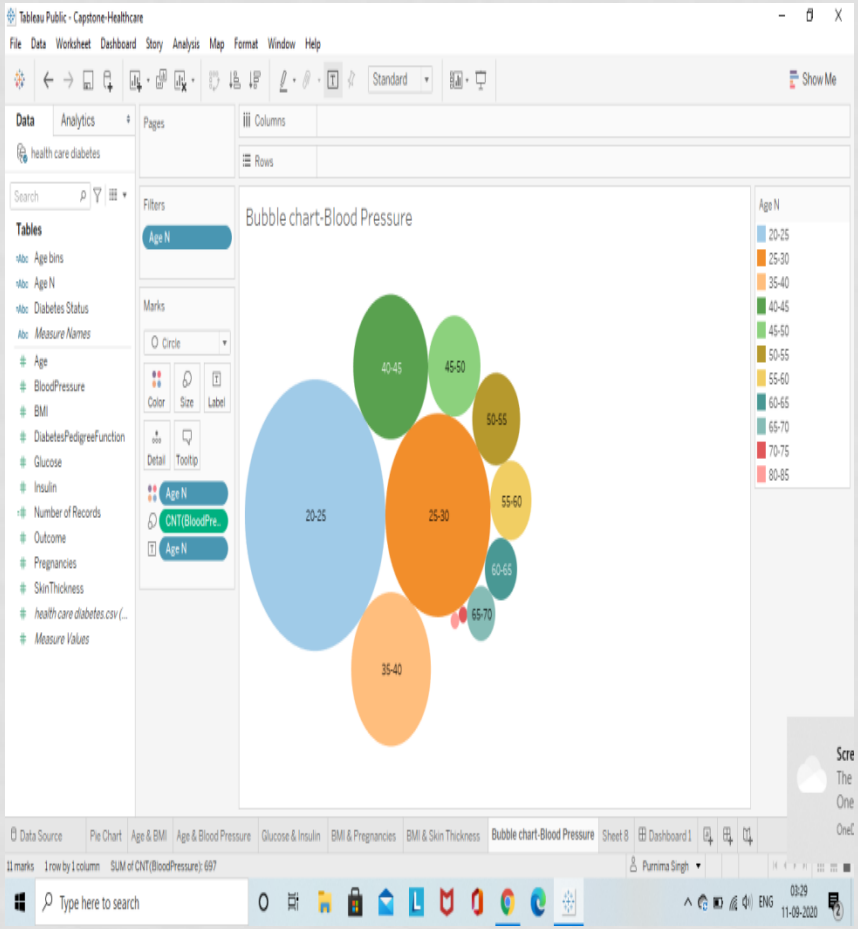


Tableau Public - Capstone-Healthcare
Bubble chart-Blood Pressure
20-25
25-30
35-40
40-45
45-50
50-55
55-60
60-65
65-70
Age N
BMI & Pregnancies
Bubble chart-Blood Pressure

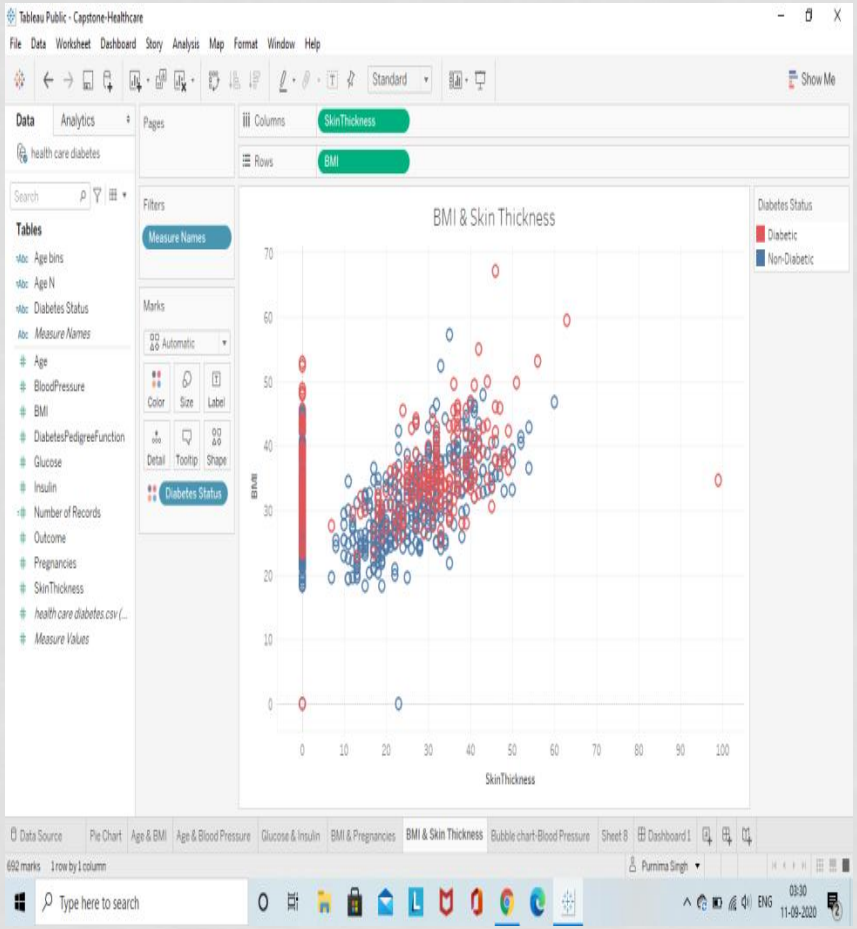


Tableau Public - Capstone-Healthcare
BMI & Skin Thickness
SkinThickness
BMI
Diabetes Status
Diabetic
Non-Diabetic
BMI & Skin Thickness