MACHINE LEARNING PROJECT

RANDOM FOREST

**Group 2**

Dan Pham | Kostiantyn Reva | Manish Dua | Precious Osarenren | Suhong Liang

Business Intelligence – BDAT1010

Professor Pedram Habibi

Tuesday, April 16th, 2019

Contents

[**Abstracts** 3](#_Toc5713842)

[Project Objectives & Scope 3](#_Toc5713843)

[About Random Forest Model 3](#_Toc5713844)

[**Dataset & Scripts** 4](#_Toc5713845)

[Chosen dataset 4](#_Toc5713846)

[Scripts 4](#_Toc5713847)

[**Model Implementation** 5](#_Toc5713848)

[Implementation on AWS 5](#_Toc5713849)

[Implementation on Azure 11](#_Toc5713850)

[Implementation on Google Cloud Platform 14](#_Toc5713851)

[**Model Result & Analysis** 18](#_Toc5713852)

[**Key Takeaways** 19](#_Toc5713853)

# **Abstracts**

## Project Objectives & Scope

This project aims to develop a learning method to classify objects based on their features.

Random Forest is chosen as the machine learning model to apply on the dataset and is implemented on three platforms: Azure, AWS and Google Cloud Platform.

## About Random Forest Model

Ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time. It corrects for decision trees habit of overfitting to their training set.

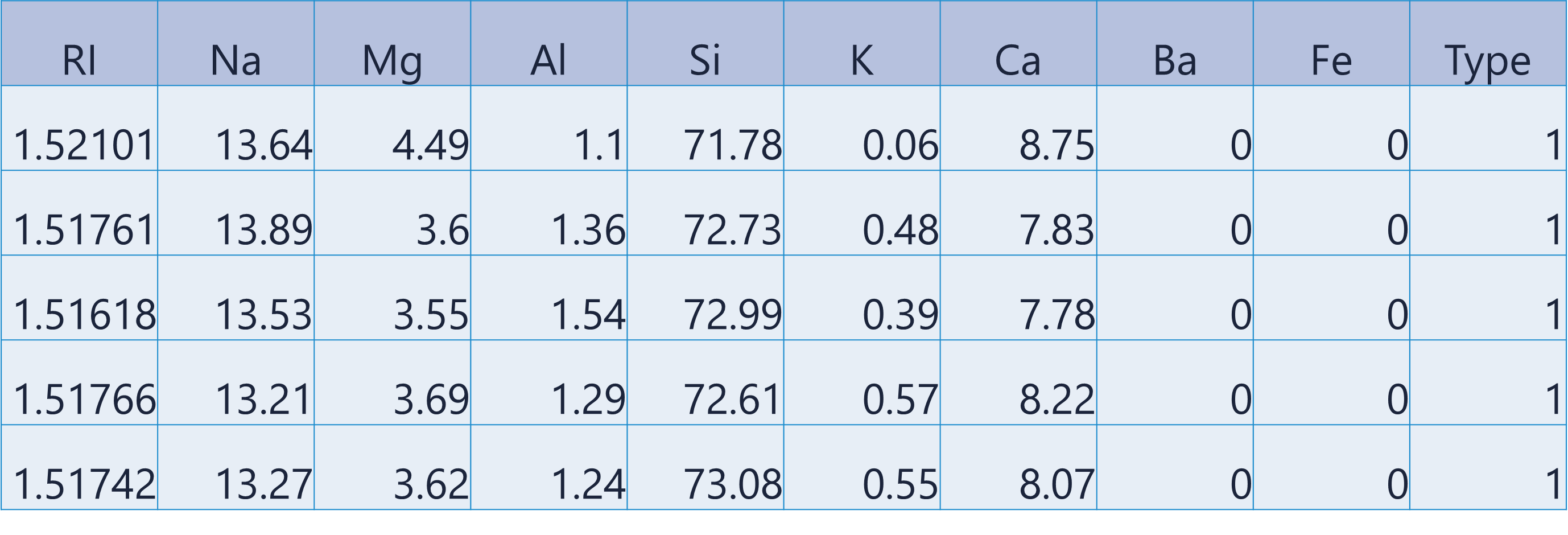
# **Dataset & Scripts**

## Chosen dataset

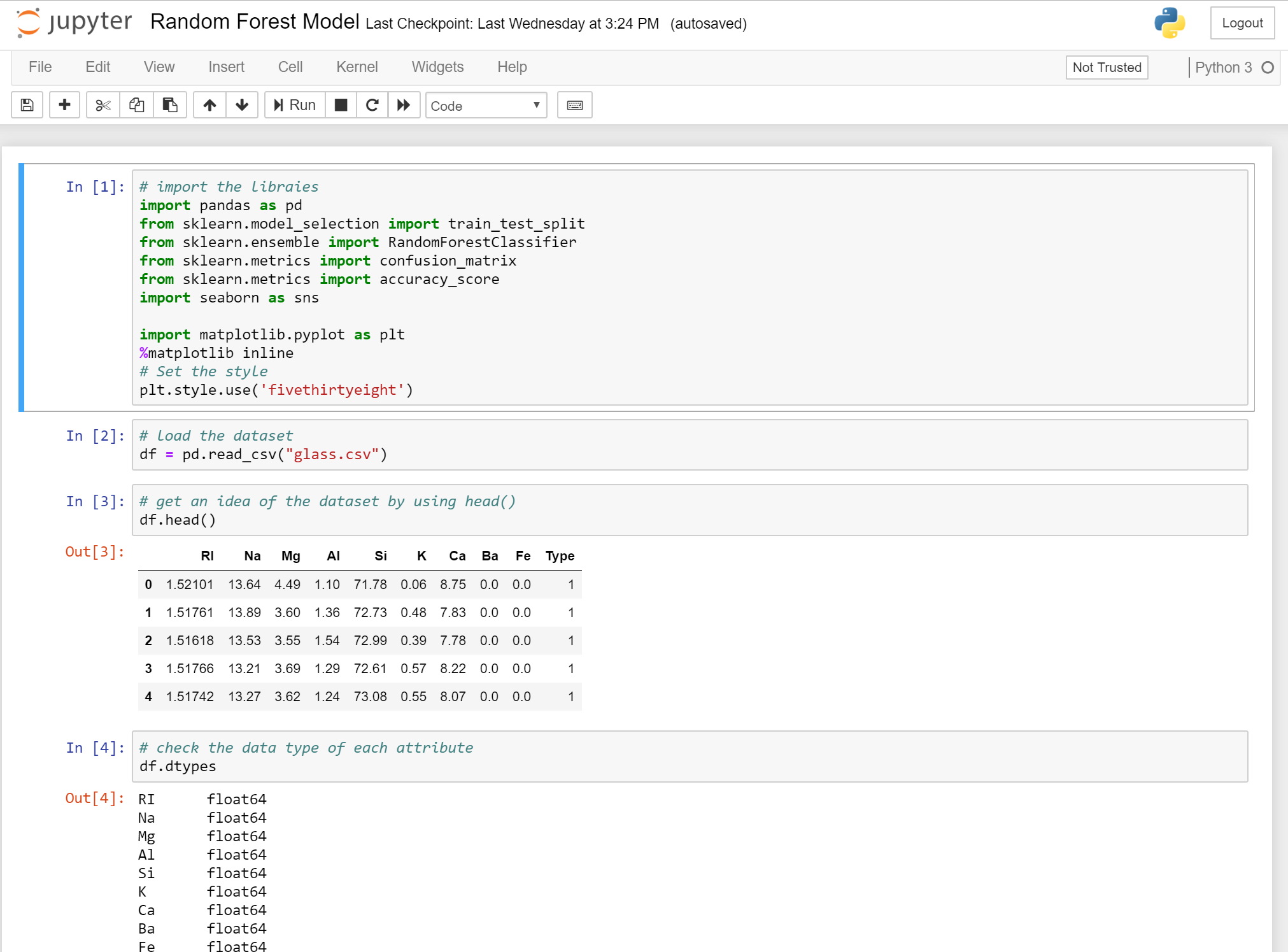
GLASS CLASSIFICATION

* Containing 10 attributes including id
* The response is glass type (discrete 7 values)
* 214 records x 10 attributes

Data source: [https://www.kaggle.com/uciml/glass#glass.csv](https://www.kaggle.com/uciml/glass)



Scripts



*\* For full script of the project, please find attached the .ipynb file*

# **Model Implementation**

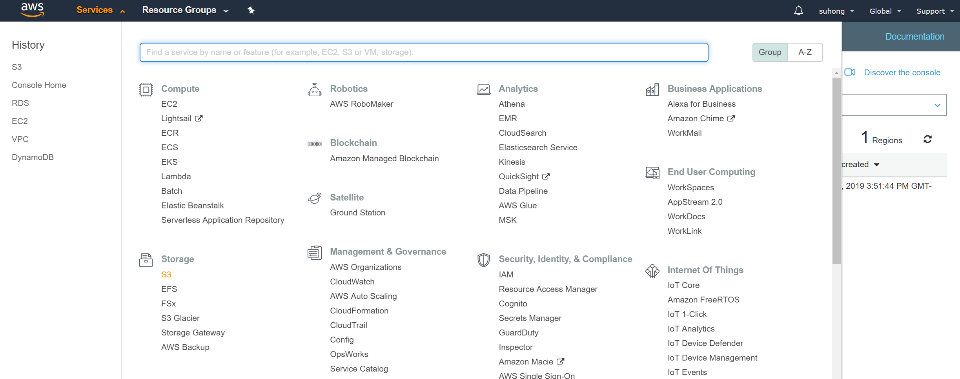
## Implementation on AWS

Amazon SageMaker is chosen to implement the model using Notebook Instance.

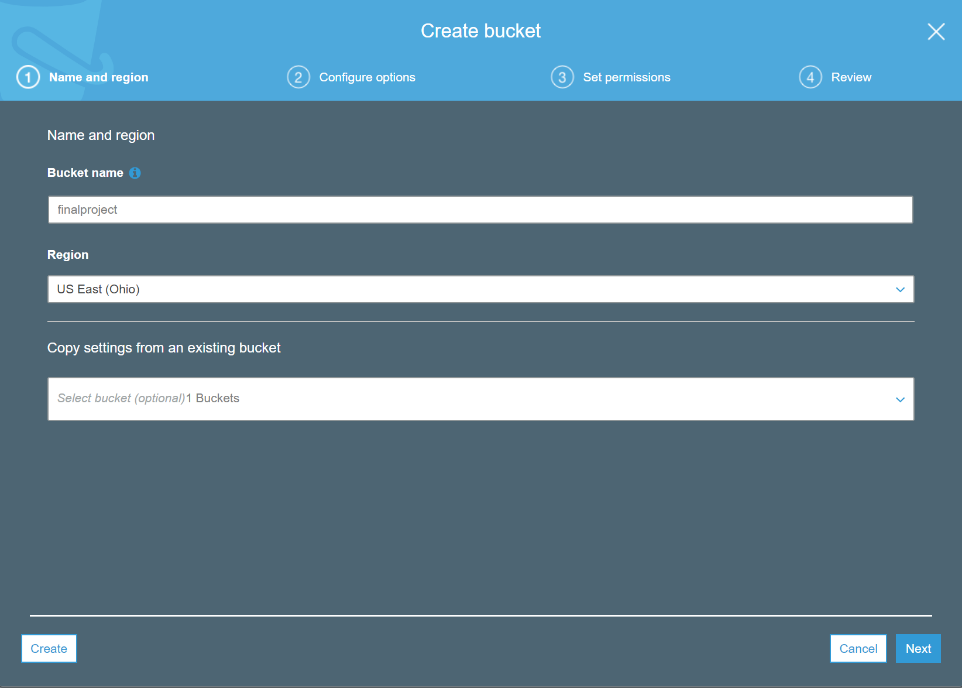
Amazon S3 is used for data storage.

Implementation steps on AWS:

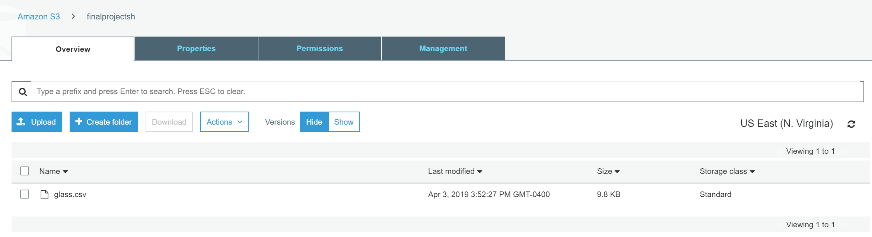
* Open AWS Management Console and select S3 from the list of services



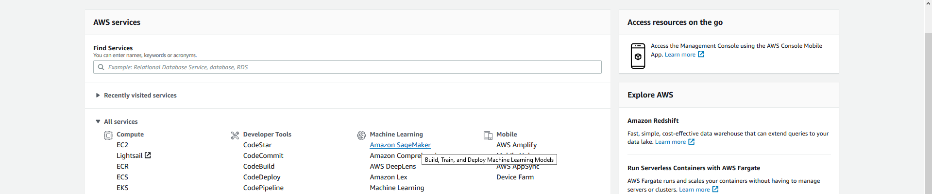
* Create a bucket



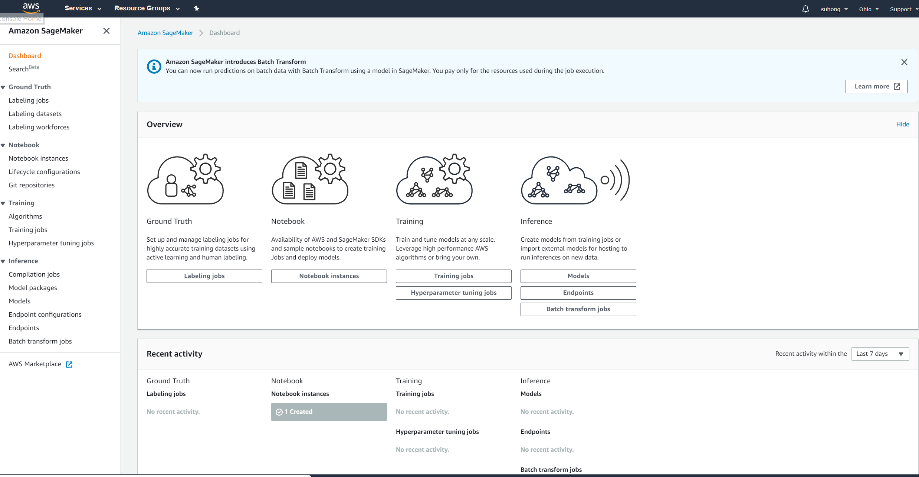
* Upload the datasets to the bucket



* Open AWS Management Console and select Amazon SageMaker



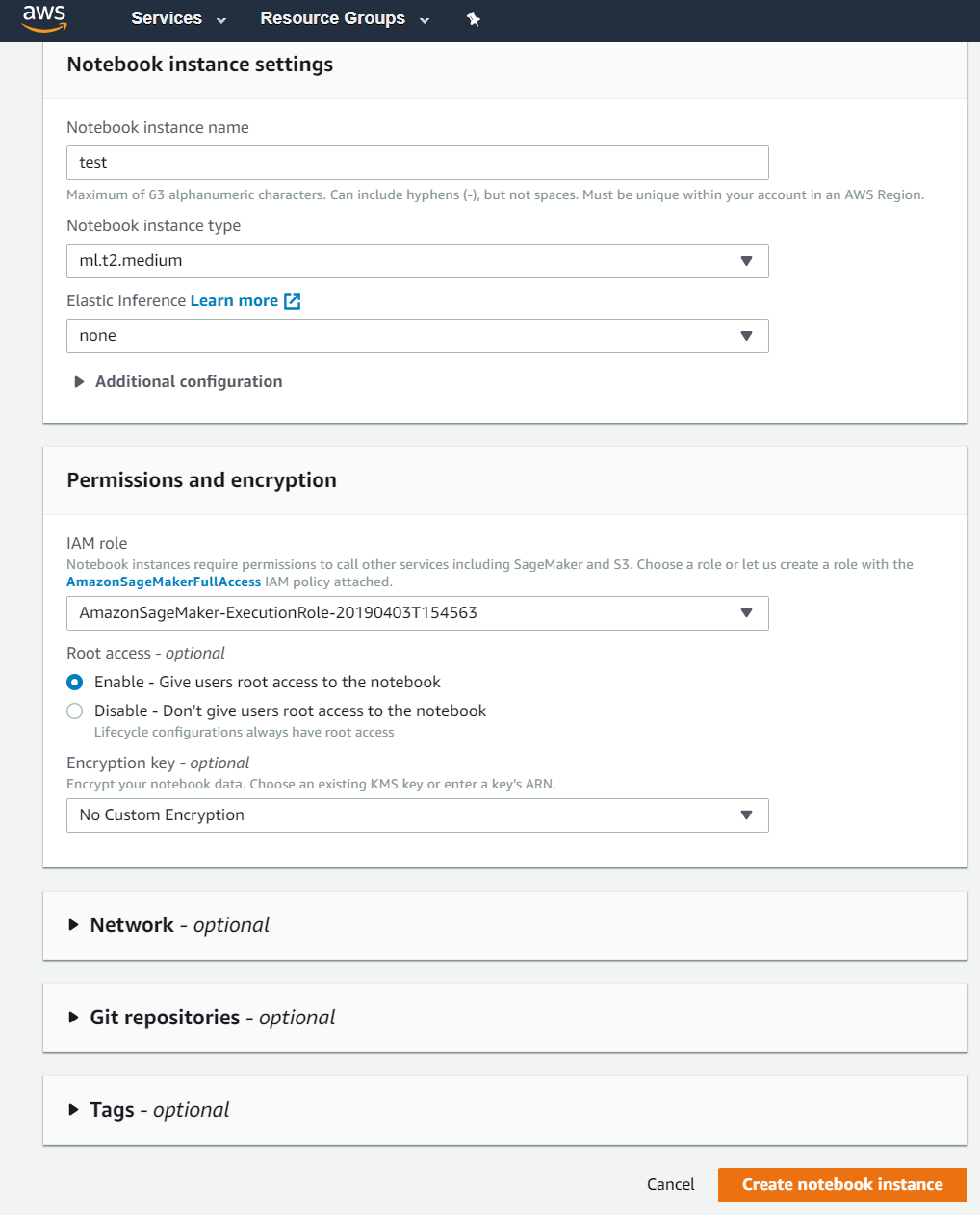
* Select Notebook Instance from the left panel

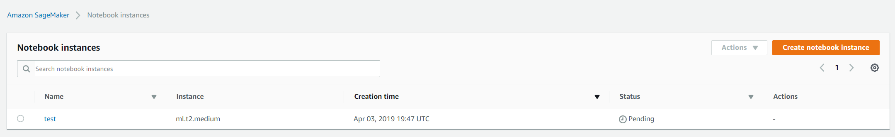


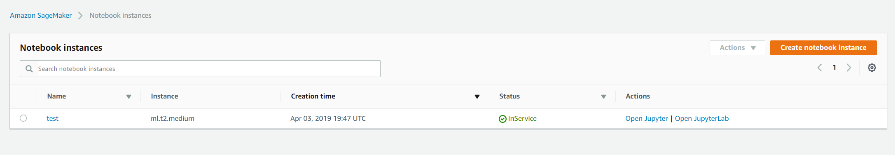
* Create a notebook instance



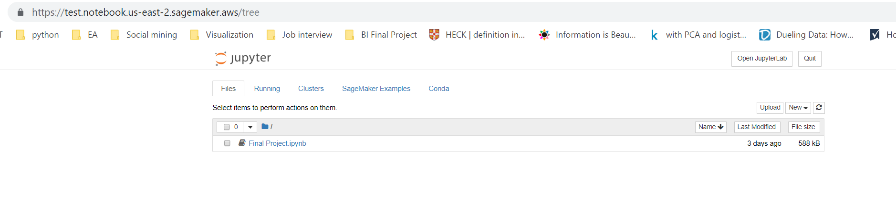
* Create and launch the notebook instance



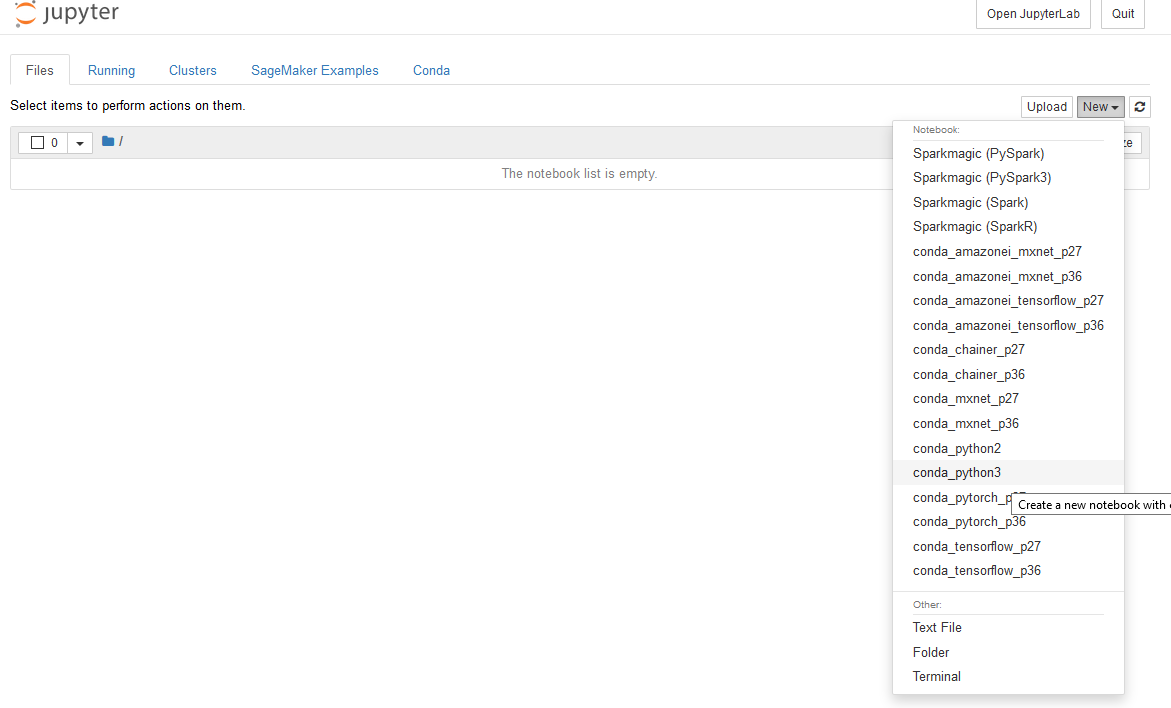




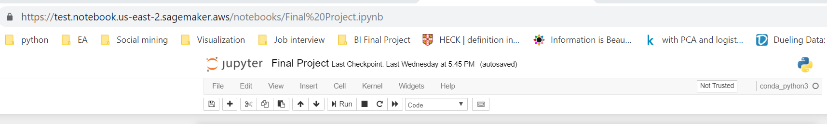
* Open Notebook



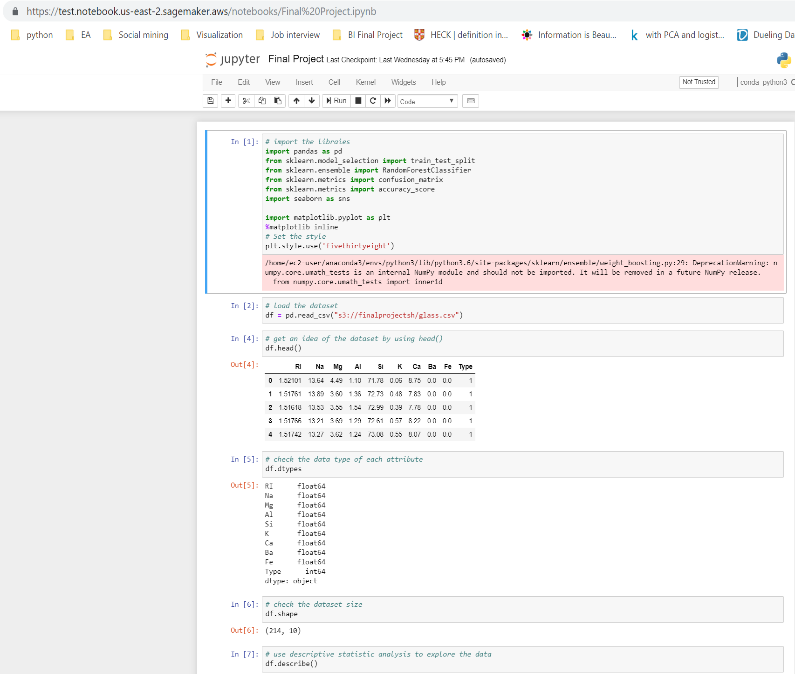
* Create a new Ipython Notebook using conda\_python3

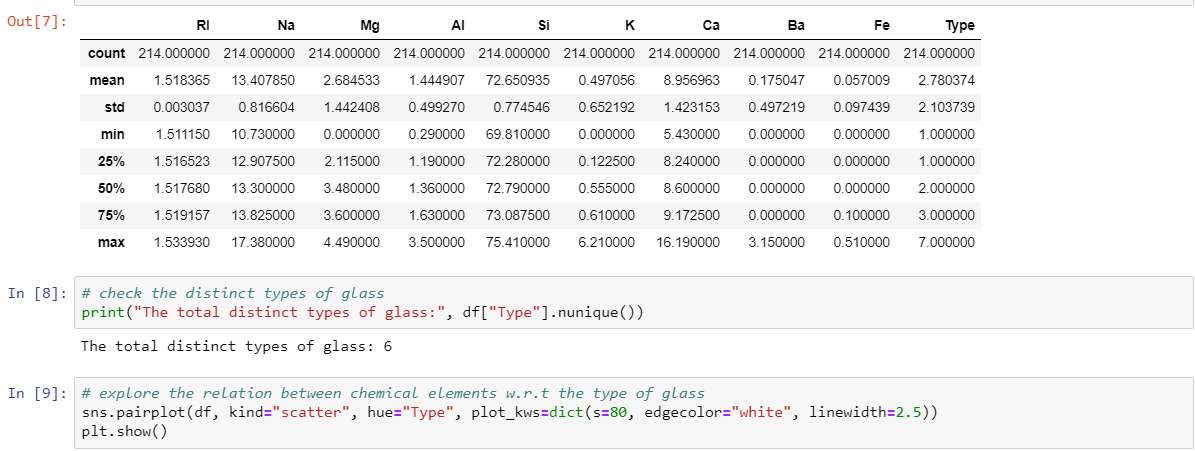


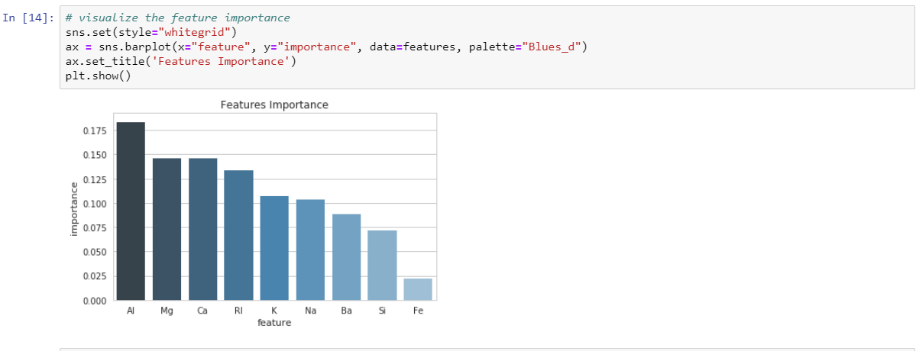
* Call it Final Project



* Run the random forest classification model using python package







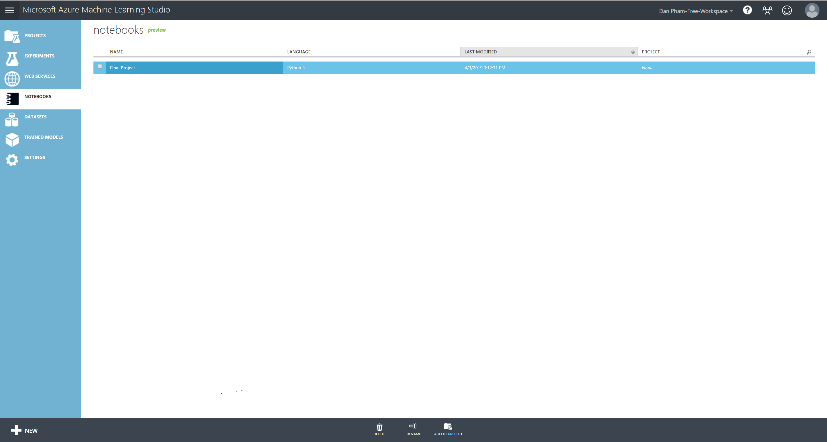
## Implementation on Azure

Azure Machine Learning Studio is chosen to implement the model using Notebook.

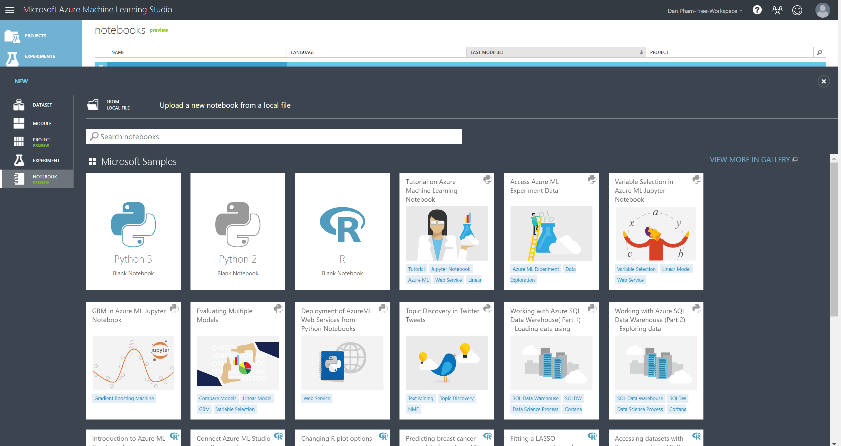
The dataset is uploaded directly into the Studio.

Implementation Process:

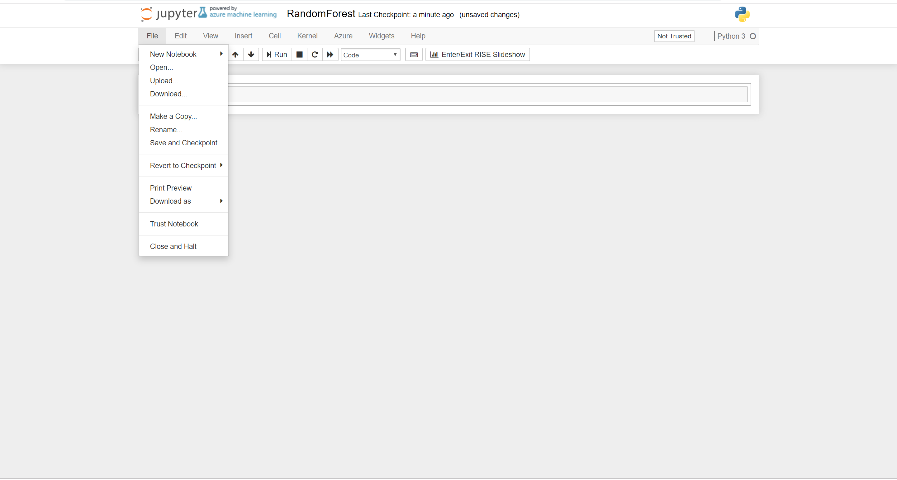
* Navigate to Notebook in Azure Machine Learning Studio



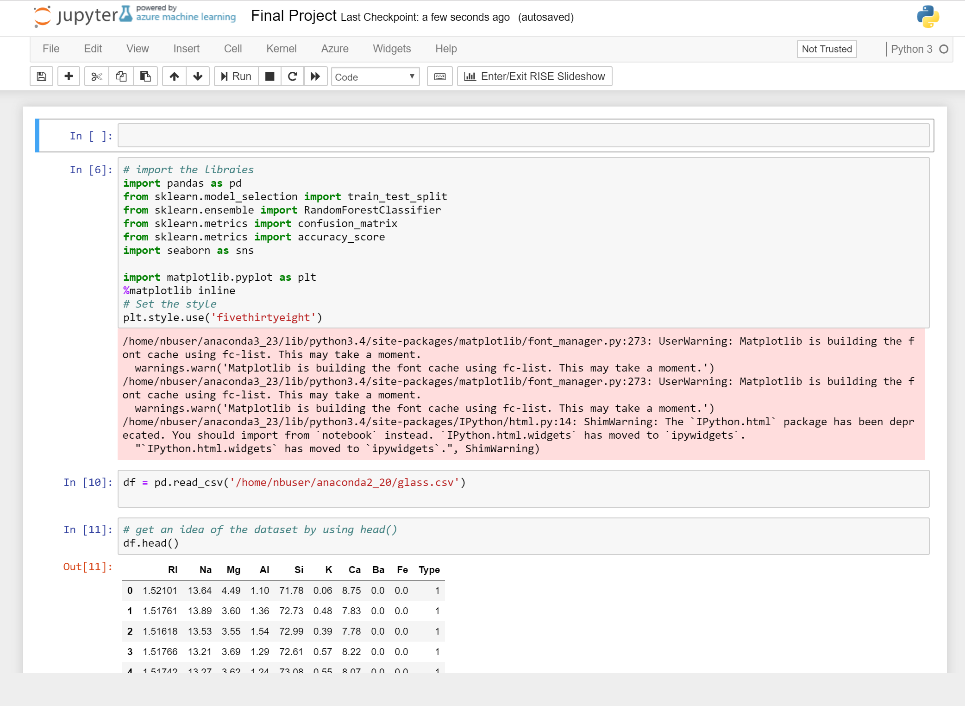
* Create a new Notebook



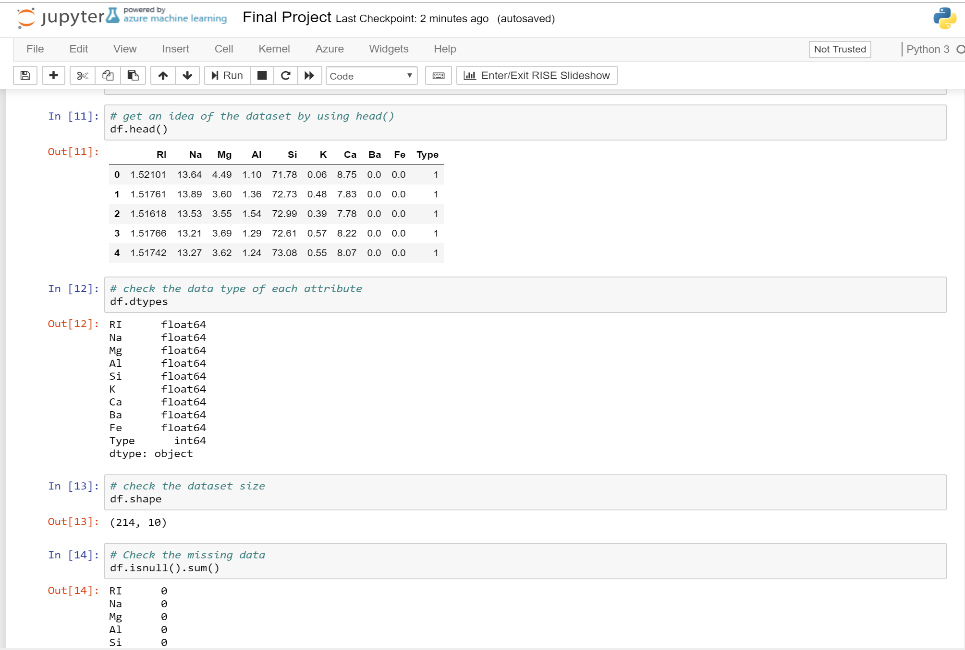
* Upload the dataset to Notebook



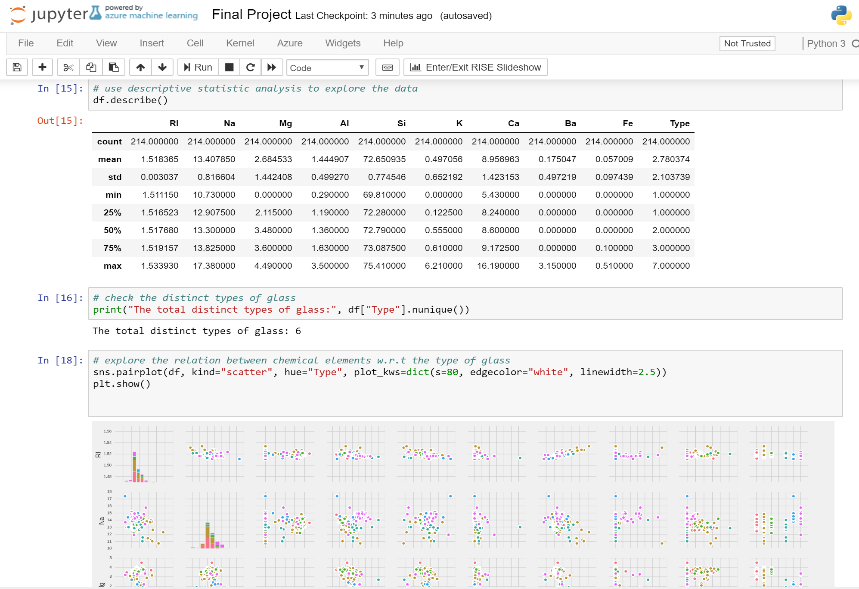
* Insert the script to Notebook, change the path of the dataset



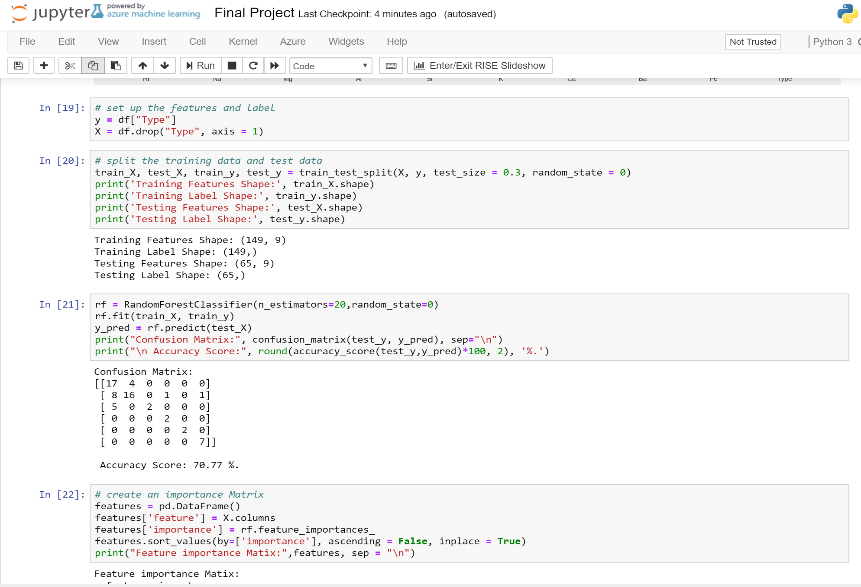
* Inspect the dataset



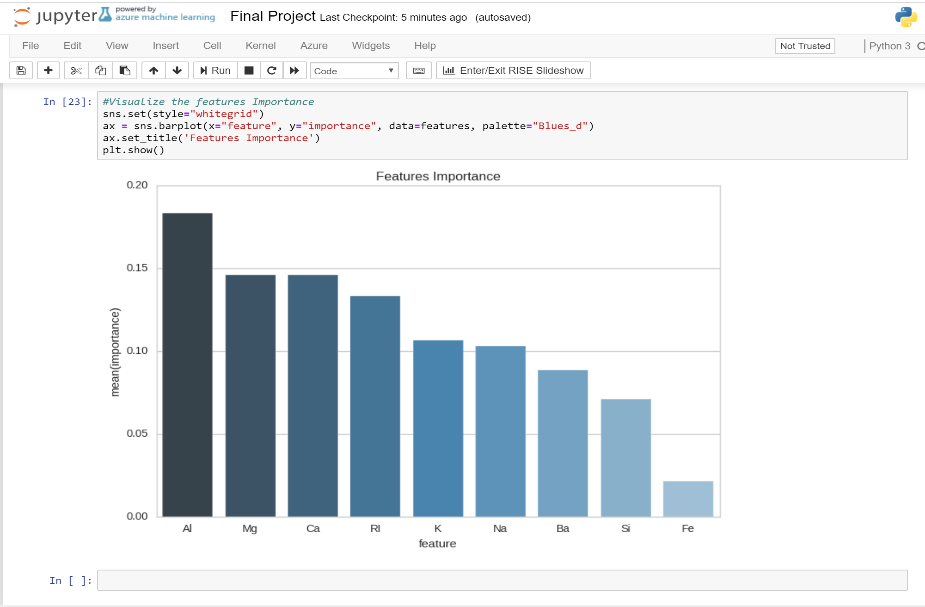
* Explore the data and relation of attributes



* Execute the model



* Visualize the features importance



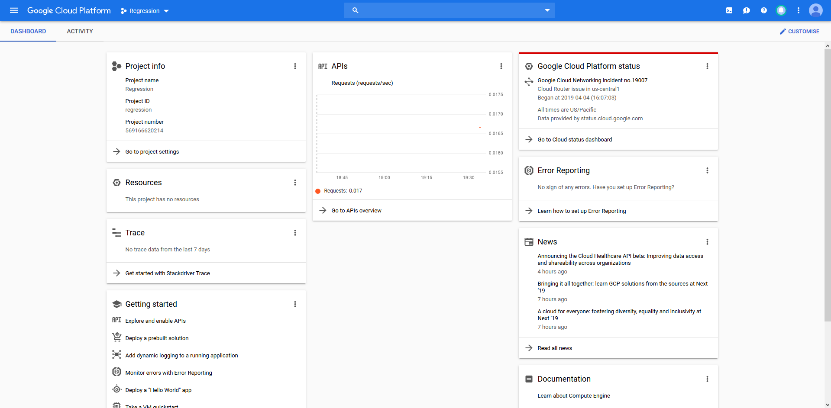
## Implementation on Google Cloud Platform

Datalab is chosen to implement the model using Notebook.

Bucket is used for data storage.

Implementation steps on Google Cloud Platform:

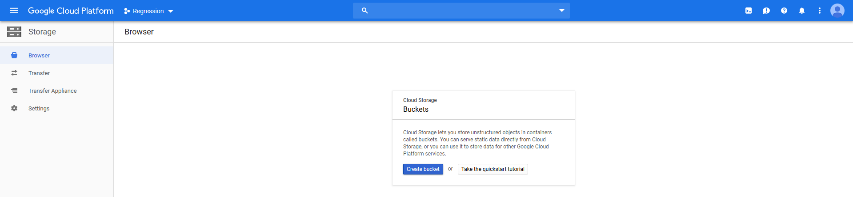
* Open Google Cloud Platform Console

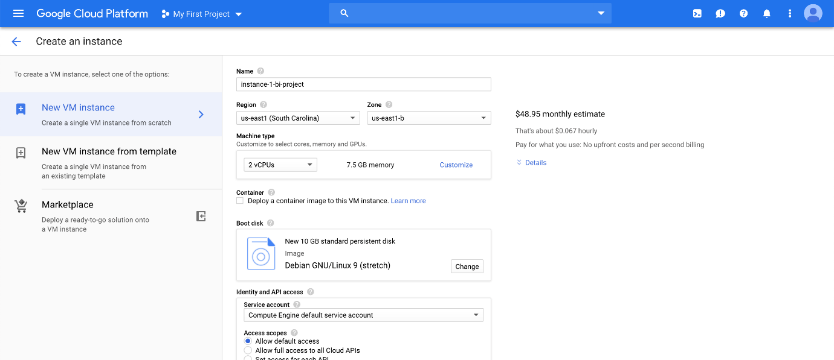


* Click on the navigation menu and select **Storage**

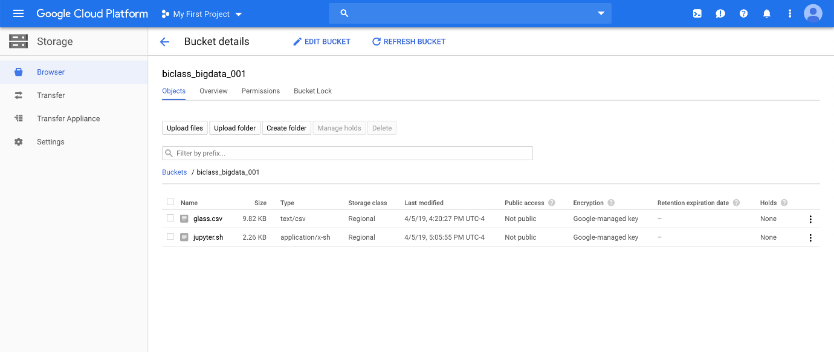


* Create a Bucket



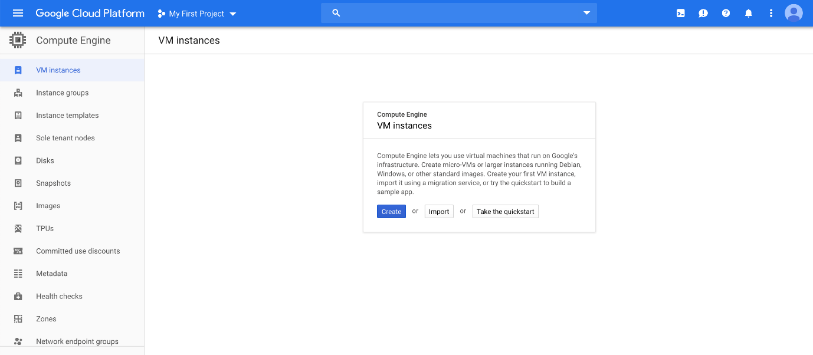


* Upload your datasets

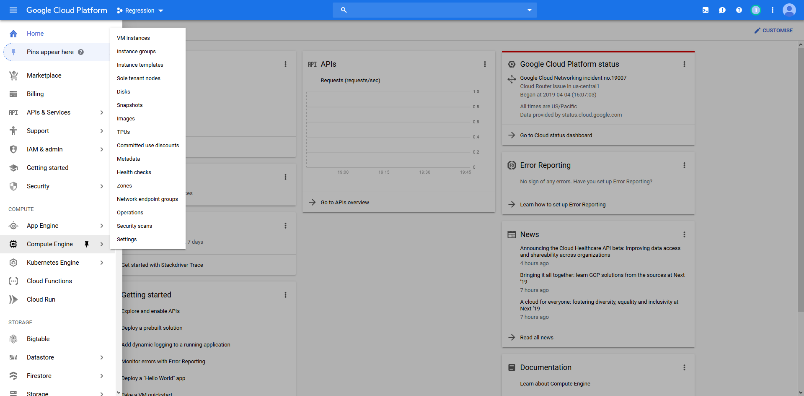


Creating VM Instance using Compute Engine and performing Linear and Logistic Regression on Google Cloud Datalab

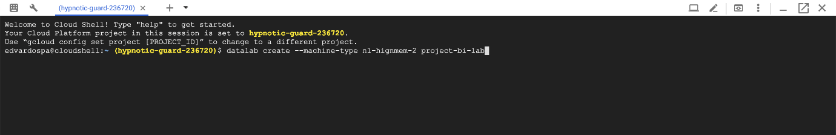
* Open Google Cloud Platform console



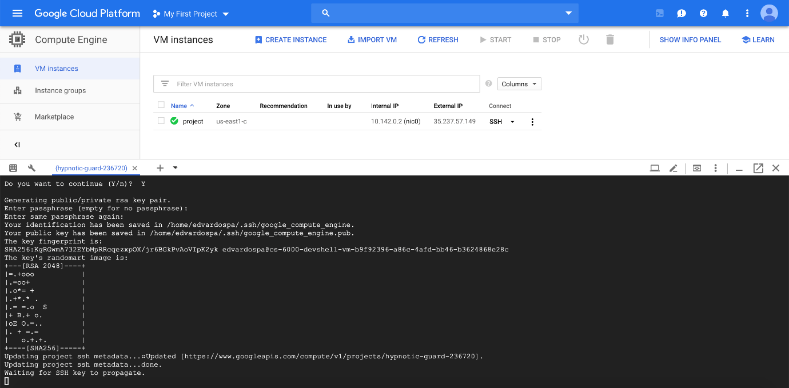
* Click on the navigation menu and select **Compute Engine**



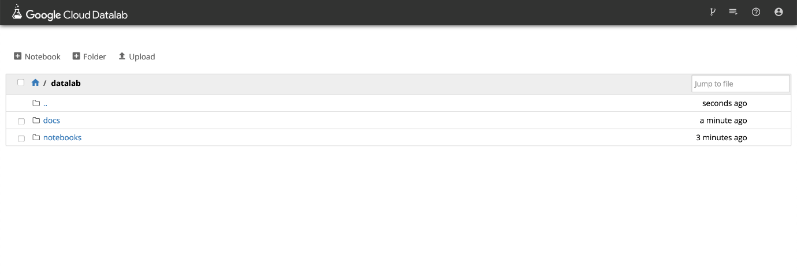
* Click on the Cloud Shell on the top and it should open a cloud shell terminal



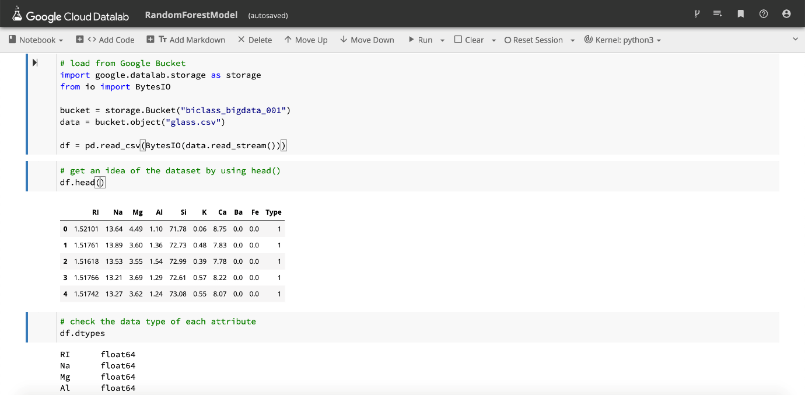
* Create a new VM Instance with high memory machine type with 2 vCPUs and 13 GB of memory named ‘project’

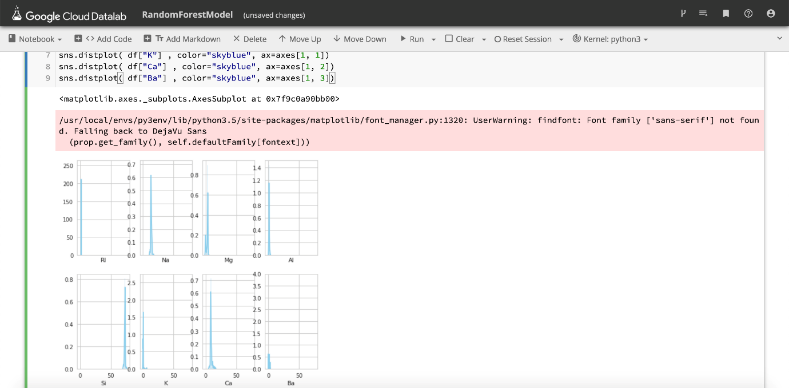


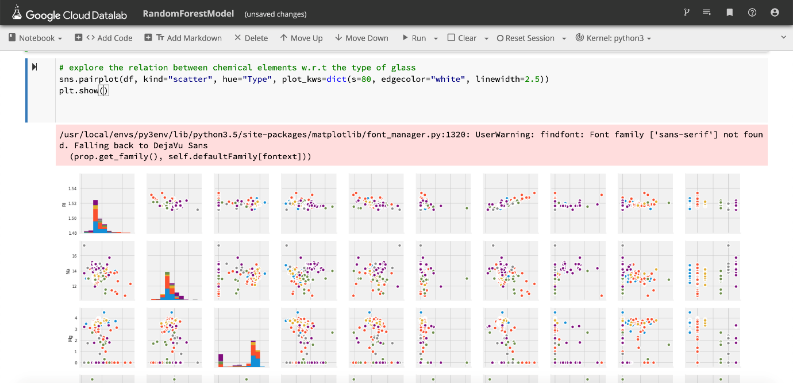
* DataLab is now open on localhost

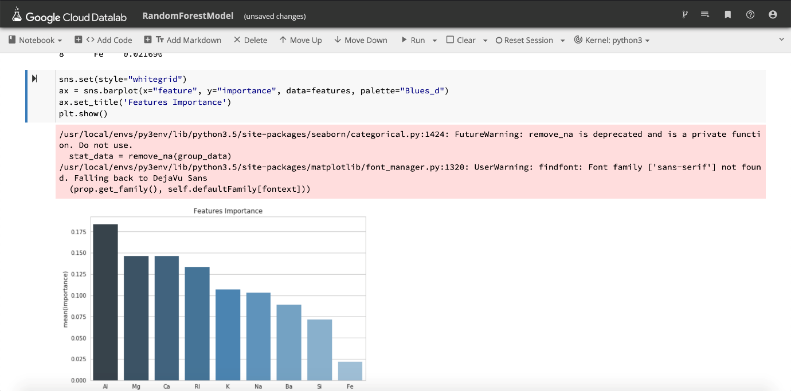


* Create a new notebook and name it









# **Model Result & Analysis**

A screenshot of a cell phone

Description generated with very high confidence

The accuracy of random forest model on the glass dataset is 71%, which is acceptable. This result is the same across all three platforms. The implementation on Azure is preferable because of its simple set up, uploading data and implementation.

A screenshot of a cell phone

Description generated with very high confidence

The features importance shows that features Al, Mg, Ca and RI are the top 4 most important features for the model.

# **Key Takeaways**

What we have learned about cloud platforms:

* How to set up storages from which we can load data for future analysis:
  + A s3 bucket for AWS
  + A bucket for Google Cloud Platform
  + Upload a dataset to Machine Learning Studio
* How to run a Project in Google Cloud and run a Datalab instance there in order to run a python notebook
* How to use AWS SageMaker to run a notebook instance so that we can run the python notebook
* How to use Azure Machine Learning Studio create notebook in order to load and run a python notebook