A picture containing logo

Description automatically generated

**Classifying Personalities (4P)**

Jessica Abueledam 20200137

Sondos Ali 20200438

**Abstract**

The aim of this project is to do classification on types of personalities. We did a stratified sample to generate 6000 records as the dataset was too large.

And We divided the personalities into 4 types (Analyst ,Diplomats, Sentinels and Explorers) from 16 Personalities types.

**Introduction**

The dataset consists of 16 types of personalities ; however, we replaced each 4 personalities with 1 personality to reduce the data and to assemble them into different single personalities.

For example(we replaced 'INTJ', 'INTP','ENTJ','ENTP' with 'Analyst' ).

So the analyst is a Mastermind, Thinker , Commander and Visionary.

The meanings of personality types are:

ESTJ - The Supervisor

ENTJ - The Commander

ESFJ - The Provider

ENFJ - The Giver

ISTJ - The Inspector

ISFJ - The Nurturer

INTJ - The Mastermind

INFJ - The Counselor

ESTP - The Doer

ESFP - The Performer

ENTP - The Visionary

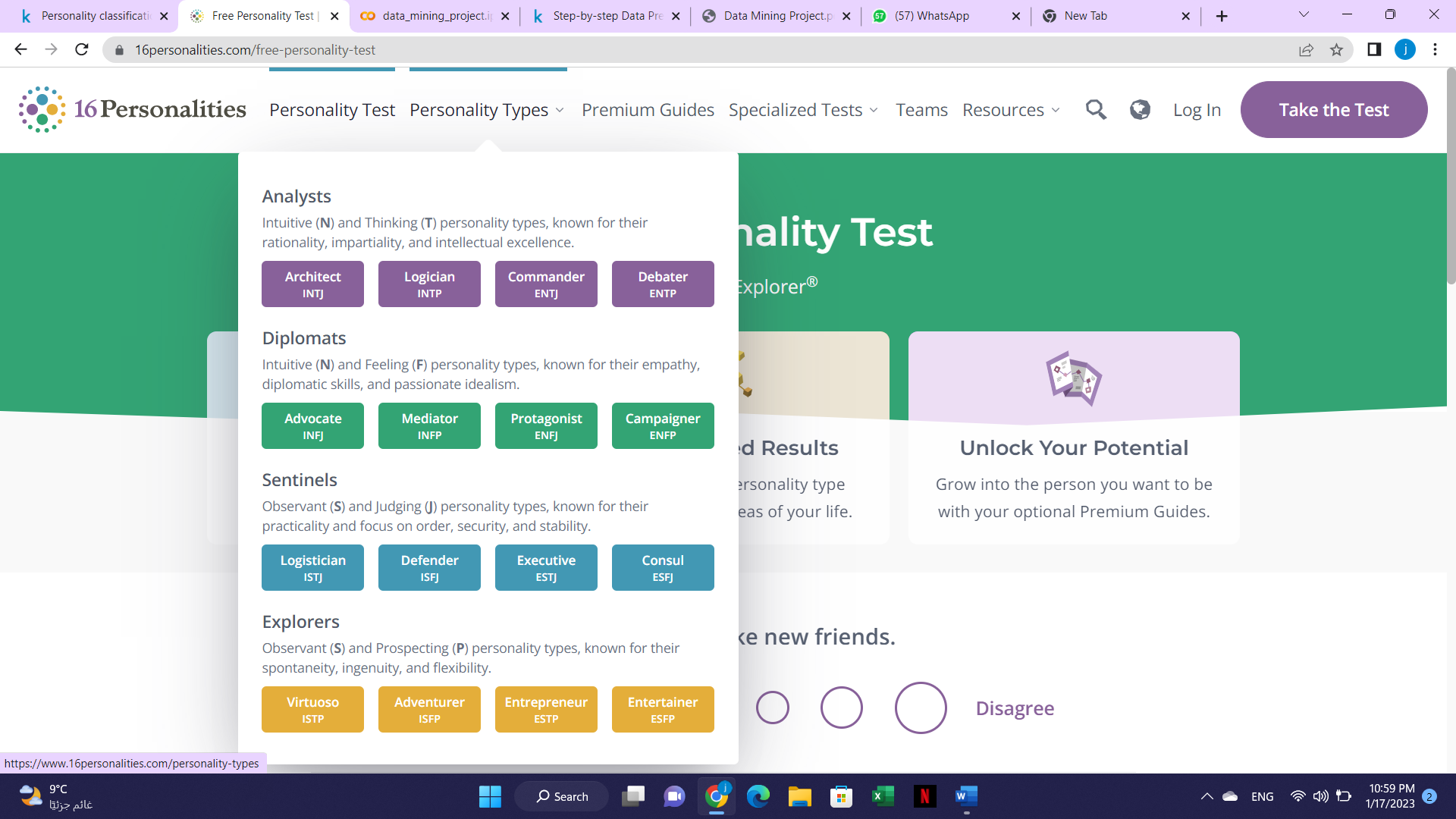
ENFP - The Champion

ISTP - The Craftsman

ISFP - The Composer

INTP - The Thinker

INFP - The Idealis



The dataset contains answers in the Scale from -3 to 3 and their meanings are as follow:

Fully Agree: 3  
Partially Agree: 2  
Slightly Agree: 1  
neutral : 0  
Slightly disagree: -1  
Partially disagree: -2  
Fully disagree: -3

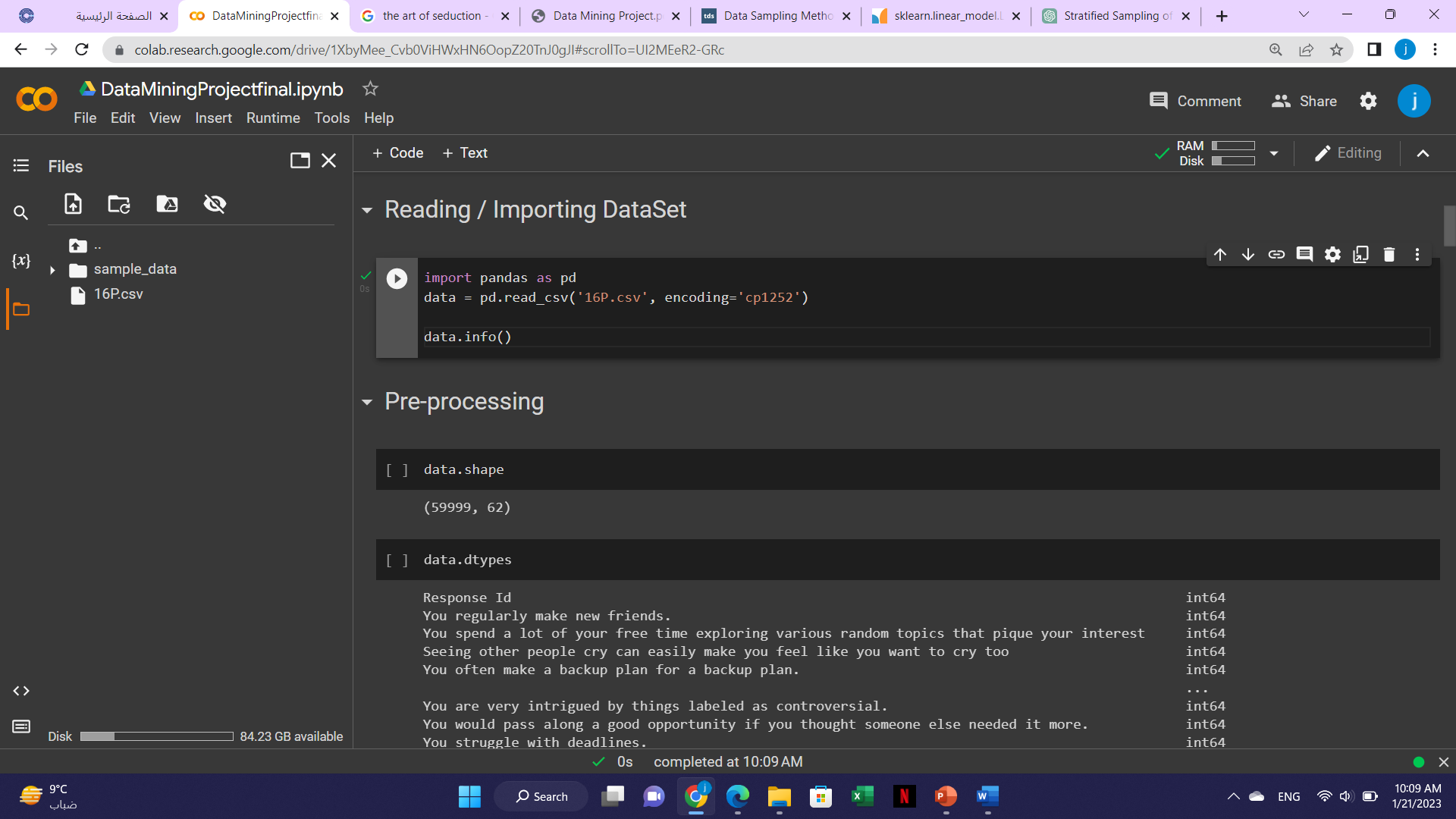
**Proposed Approach**

**Data**

1. We imported the necessary libraries that we want to use

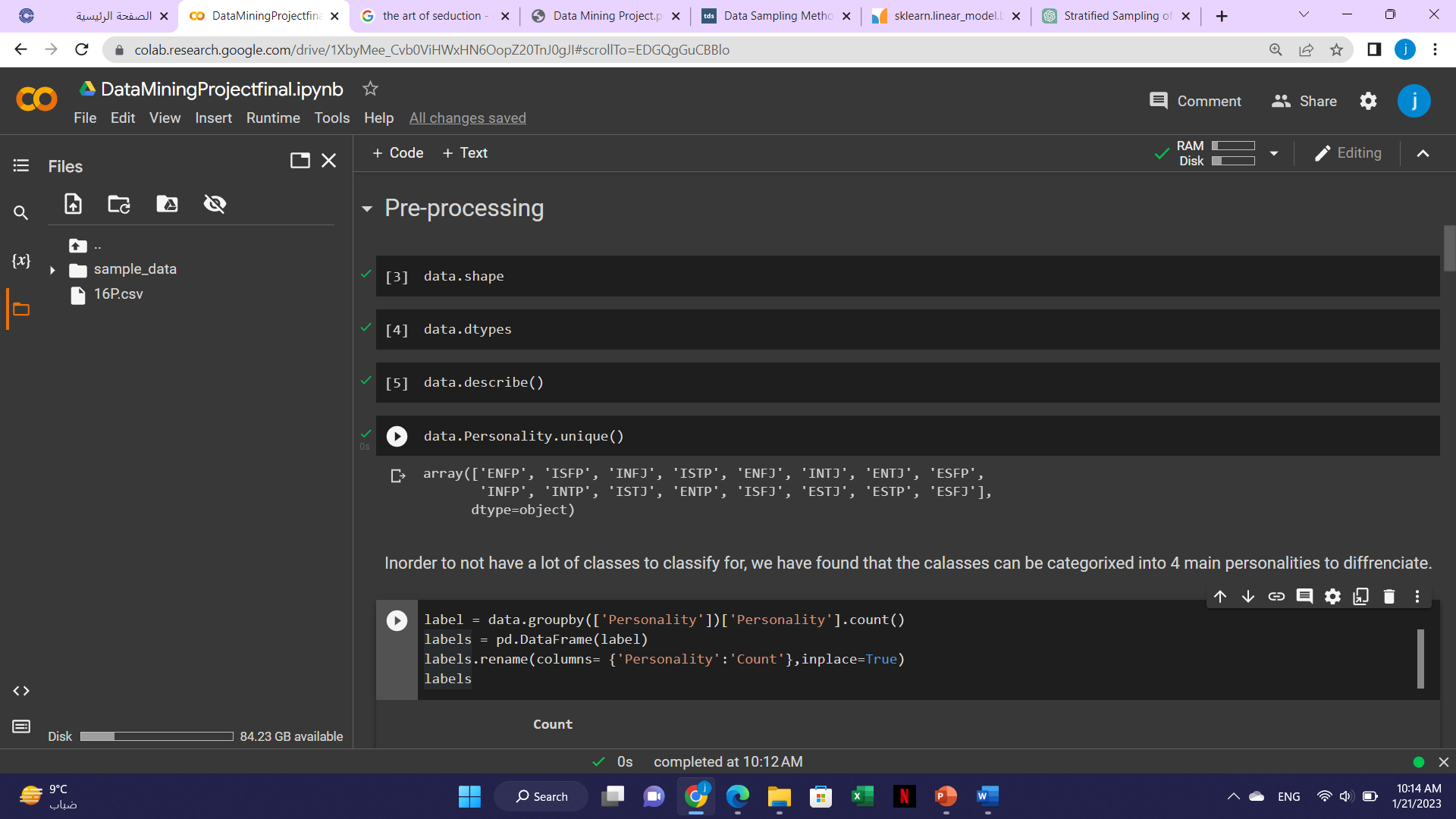


1. We loaded the dataset from excel into a panda dataframe



**Preprocessing**

We started with viewing the data and data cleaning:



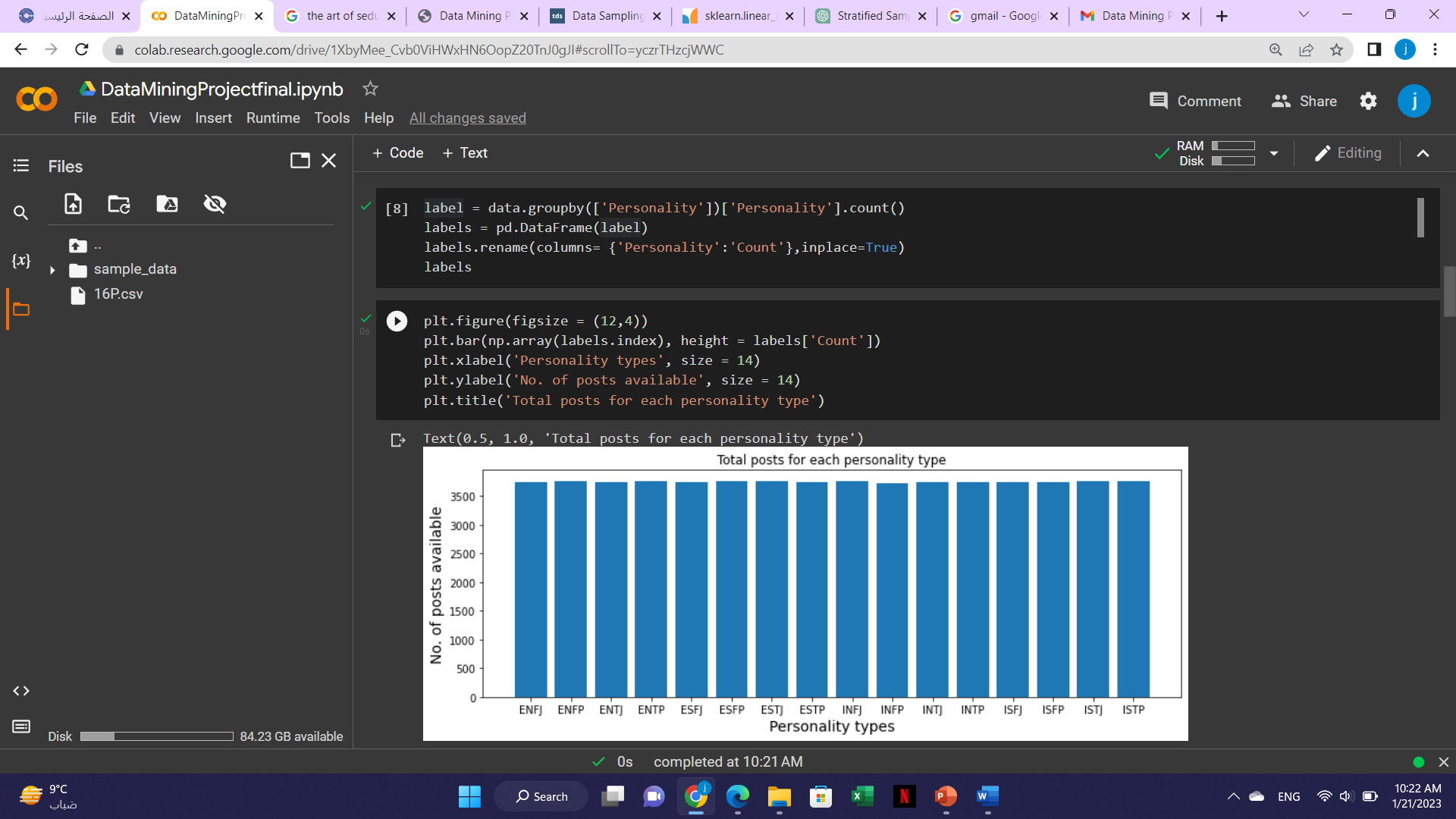
In order not to have a lot of classes to classify ,we found that we can categorize the classes into 4 main personalities .

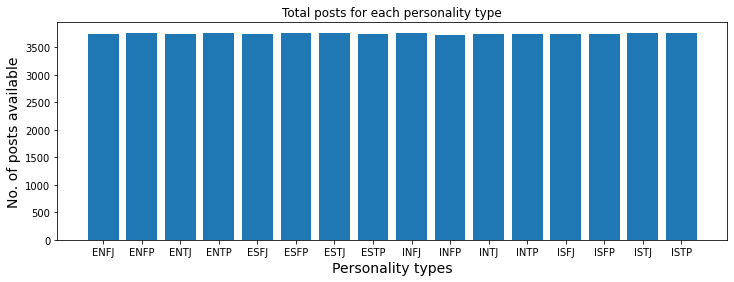
So, an analyst is 'INTJ', 'INTP','ENTJ' and 'ENTP’.

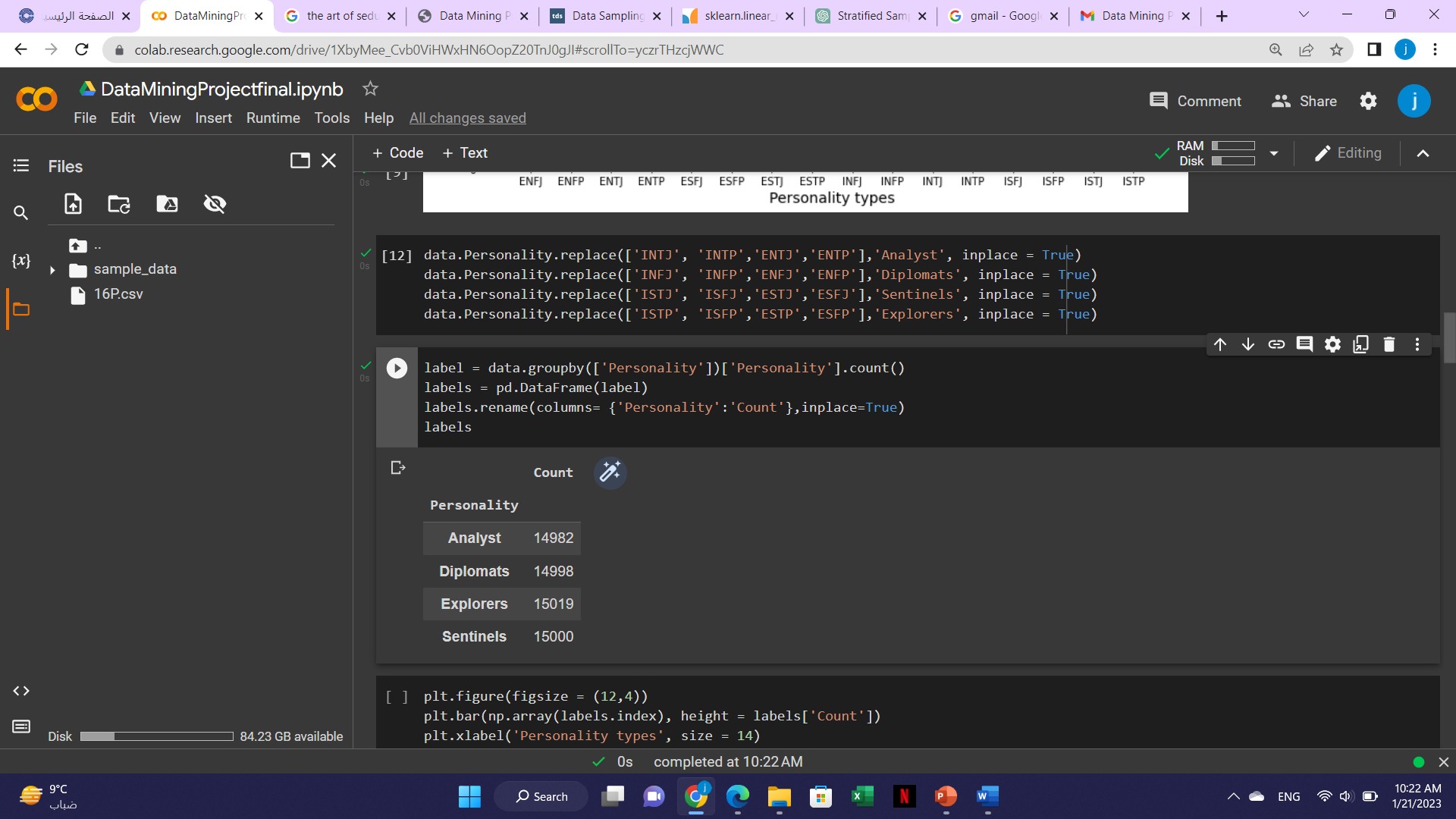
A diplomat is 'INFJ', 'INFP','ENFJ' and 'ENFP'.

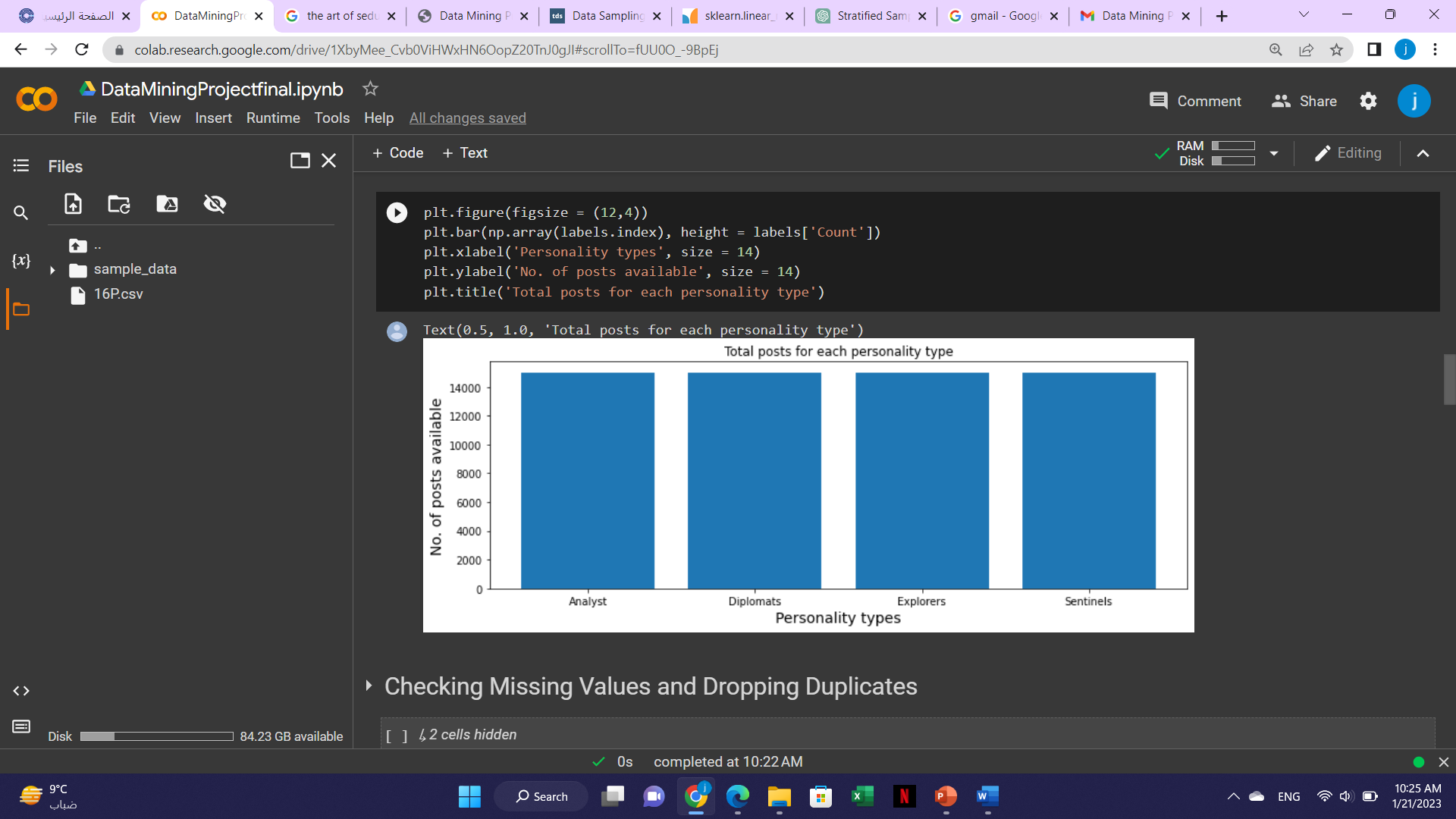
A Sentinel is 'ISTJ', 'ISFJ','ESTJ' and 'ESFJ'.

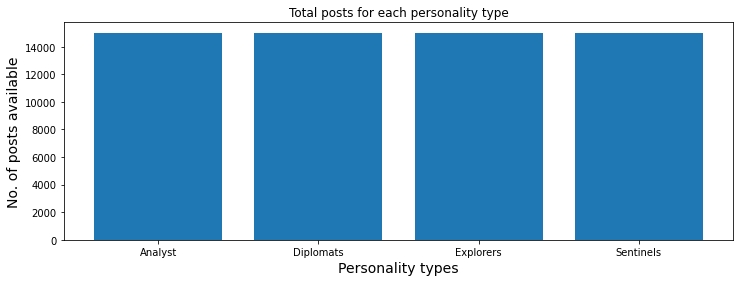
An Explorer is 'ISTP', 'ISFP','ESTP' and 'ESFP'.



****





****

After that we checked for any missing values and dropped duplicates :

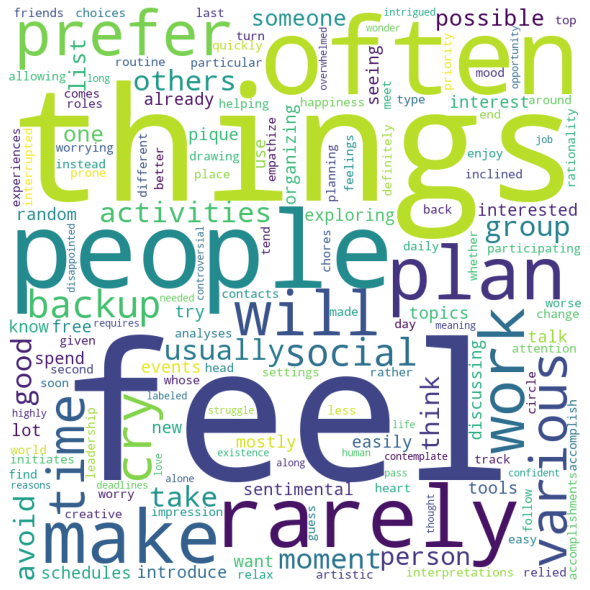


We then used the word cloud to quickly identify the most common words that describe each personality type :

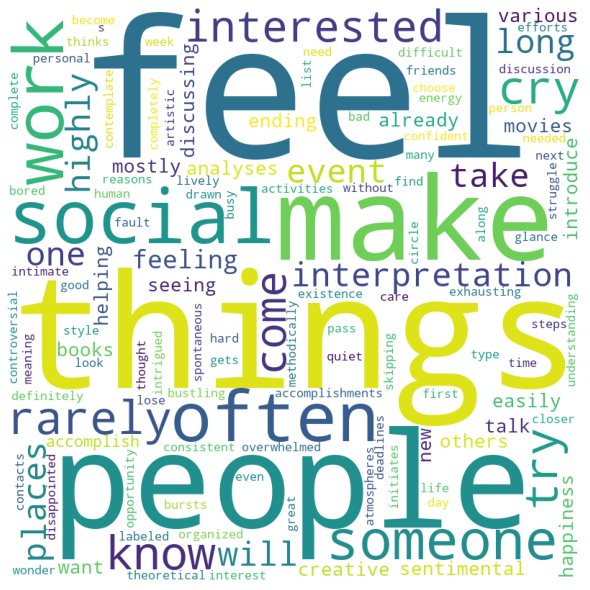
As for the analyst the most common words are :



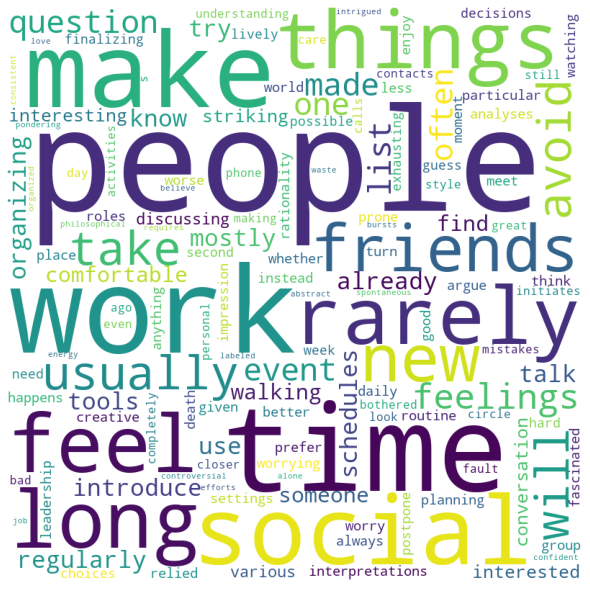
As for the Diplomats the most common words are :



As for the Sentinels the most common words are :



As for the Explorers the most common words are :



**Modelling**

As for the modelling we applied:

1)Decision tree

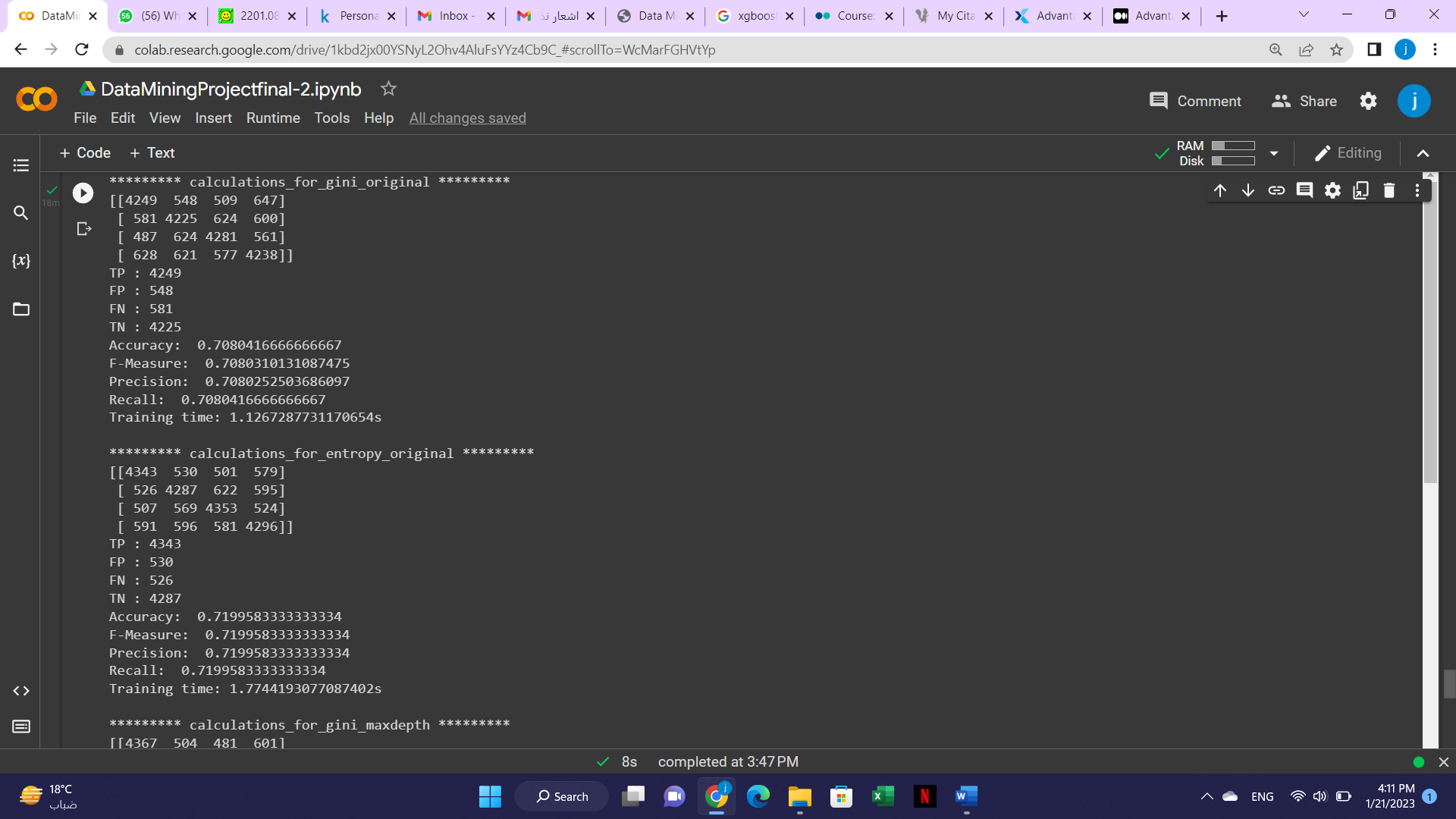
2) XGBoost

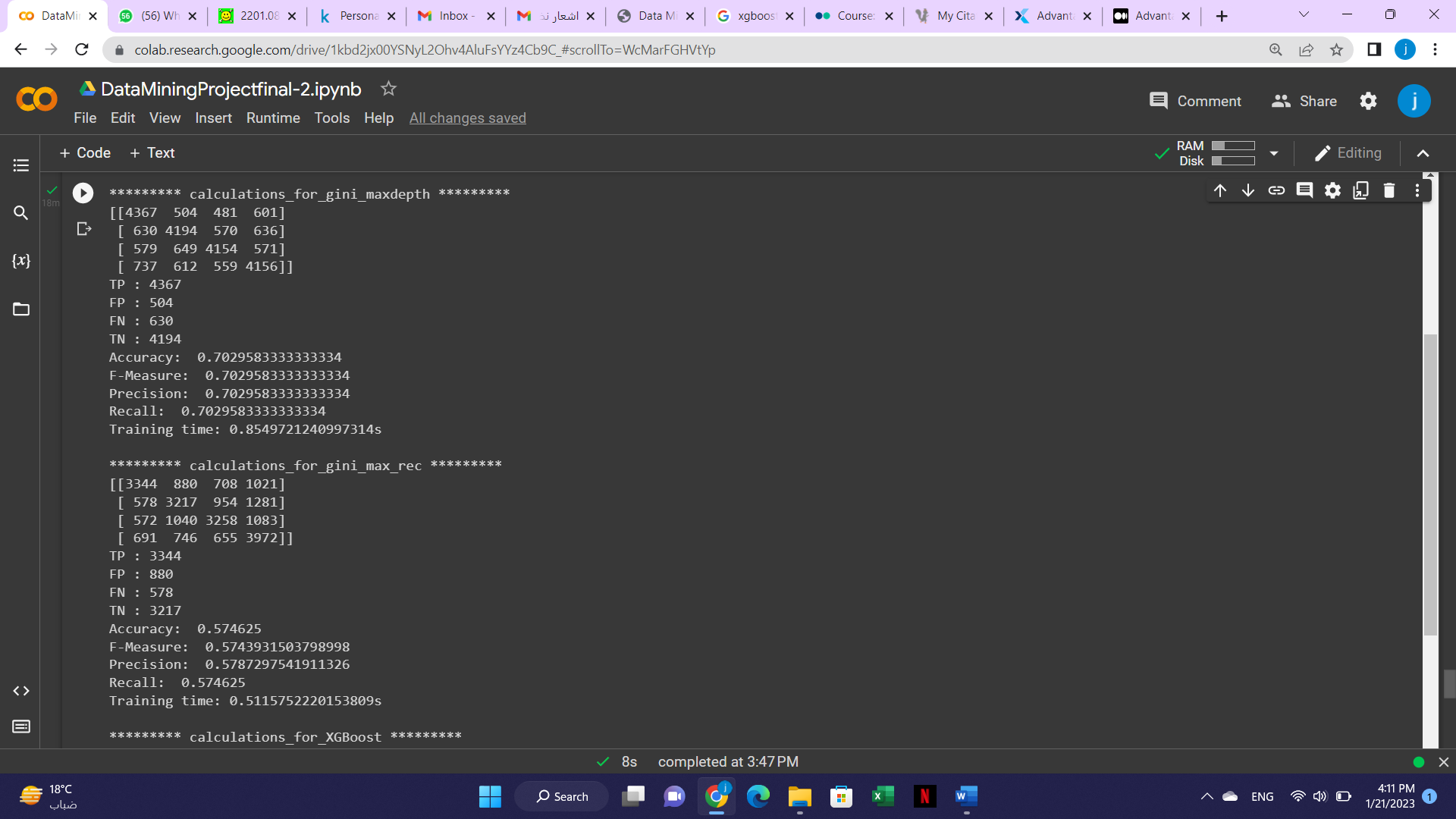
3)SVM

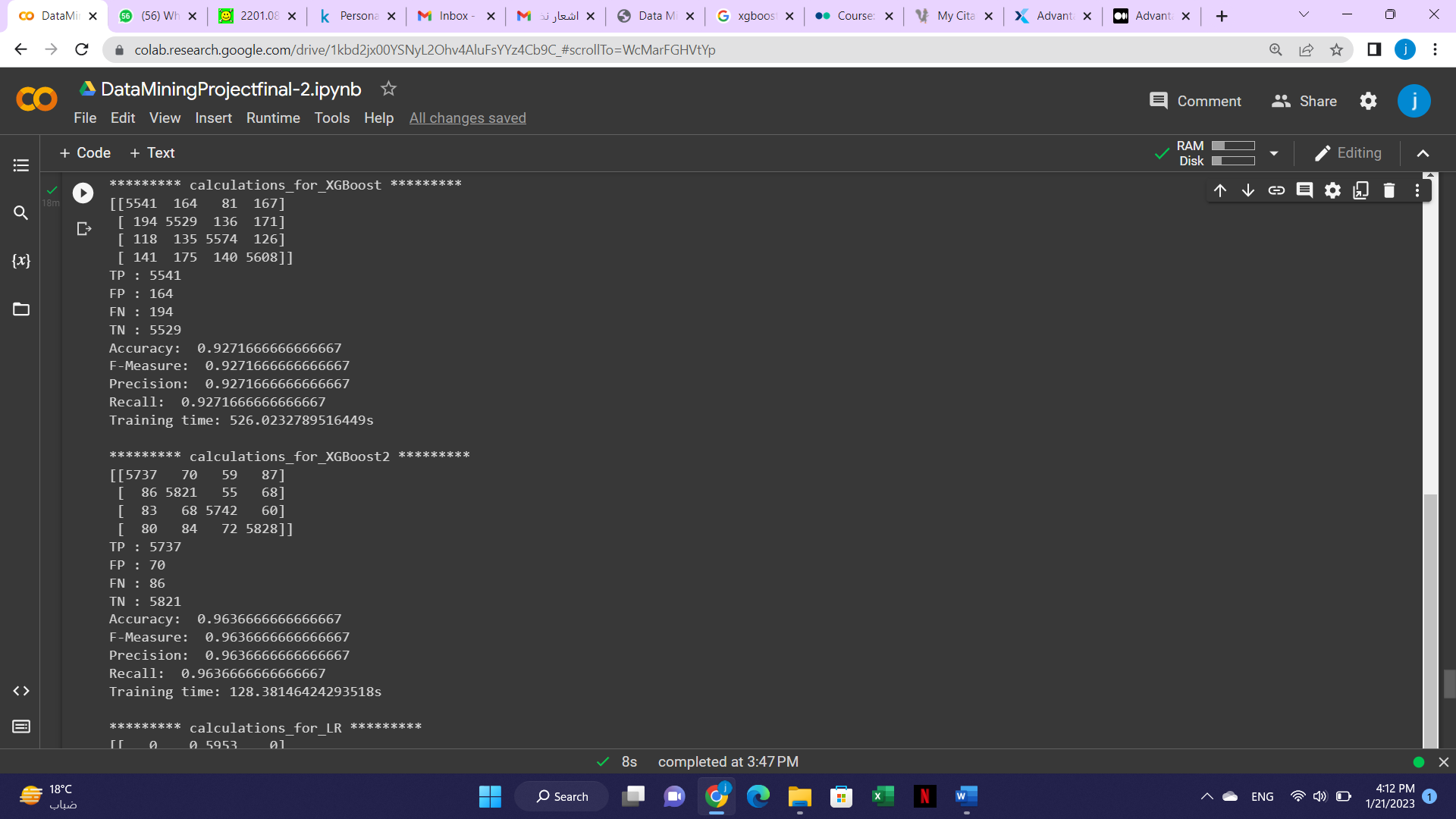
4)Naïve Bayes

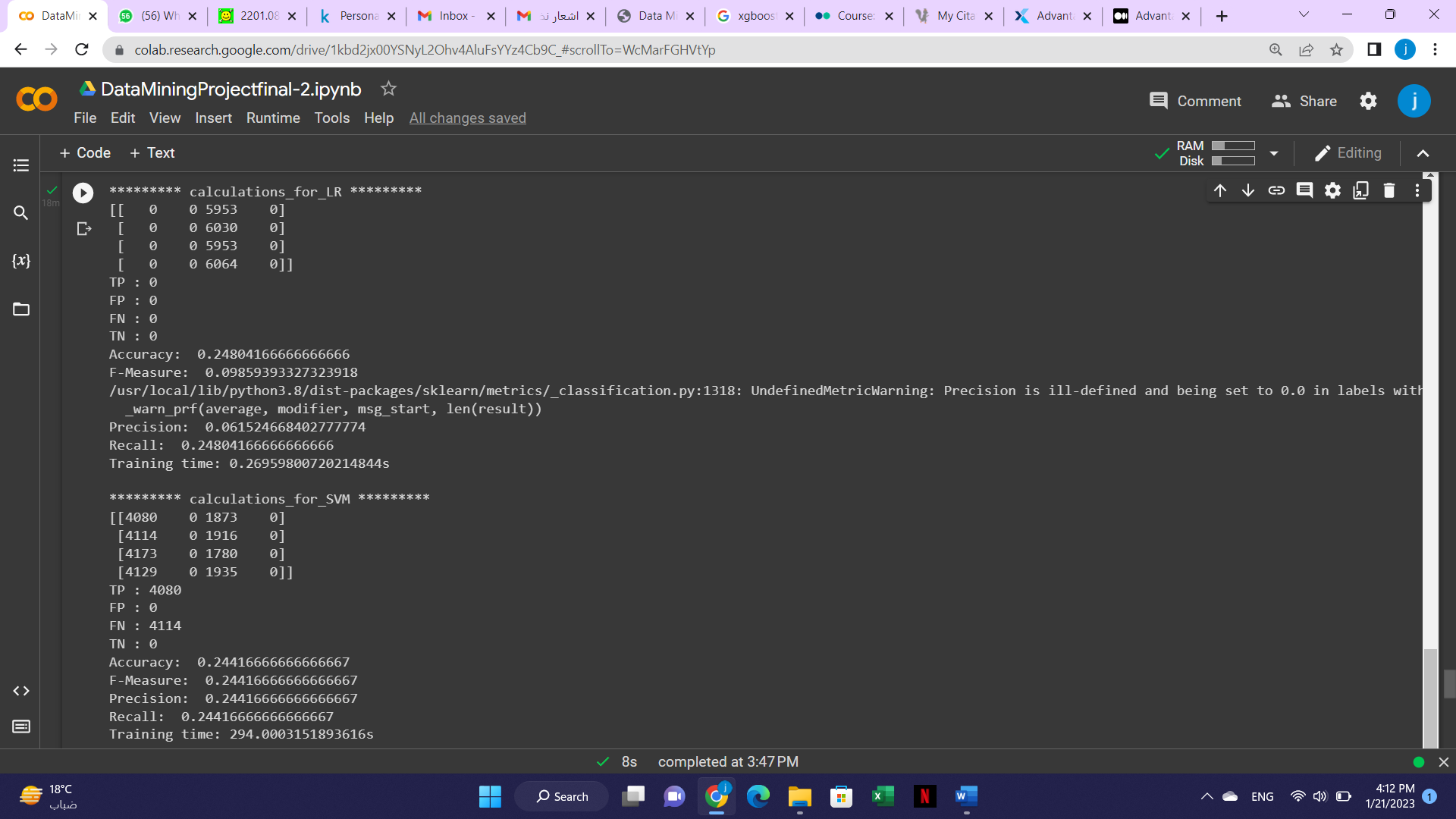
5)KNN

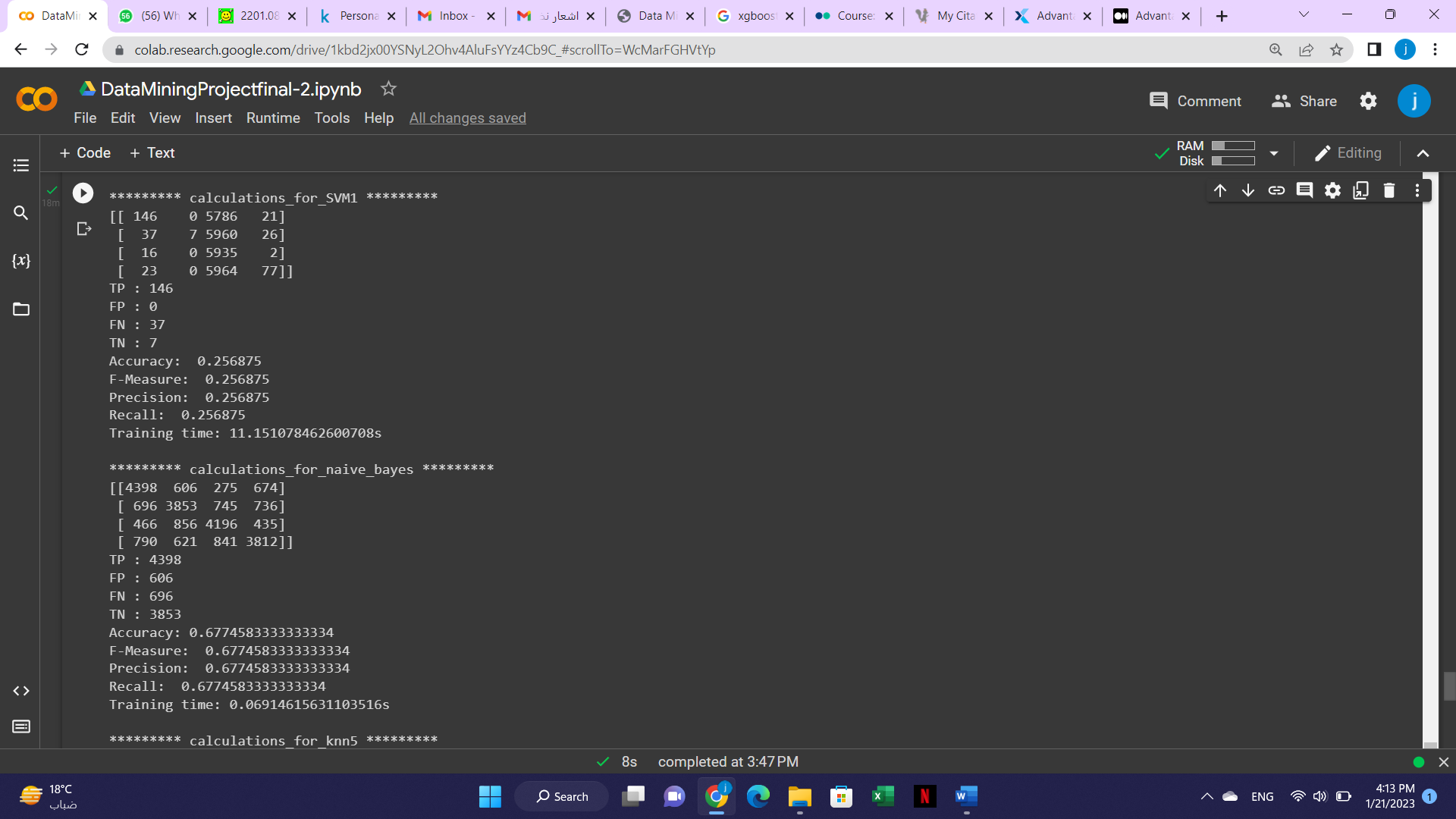
**Model performance evaluation measures table below:**

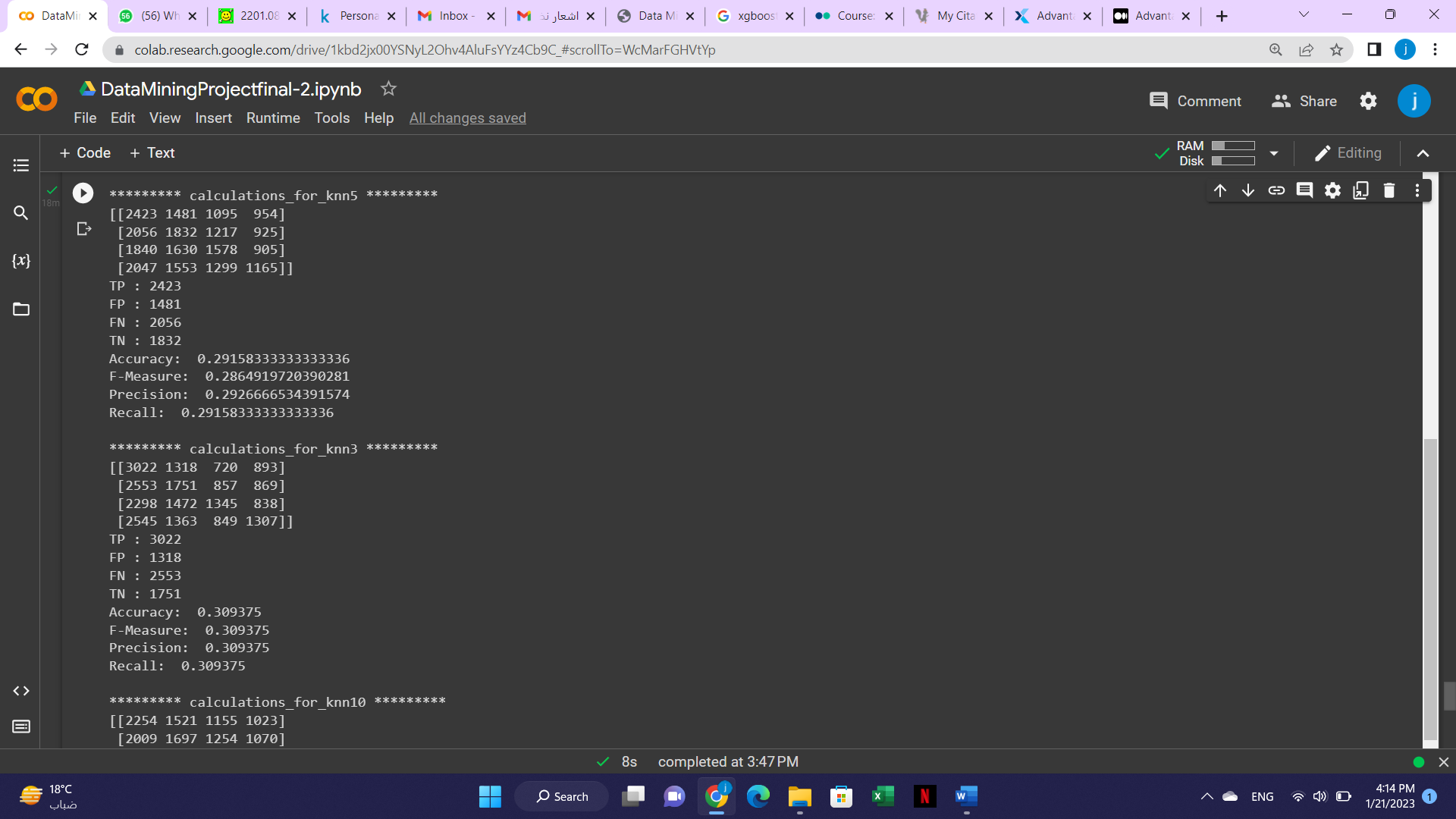


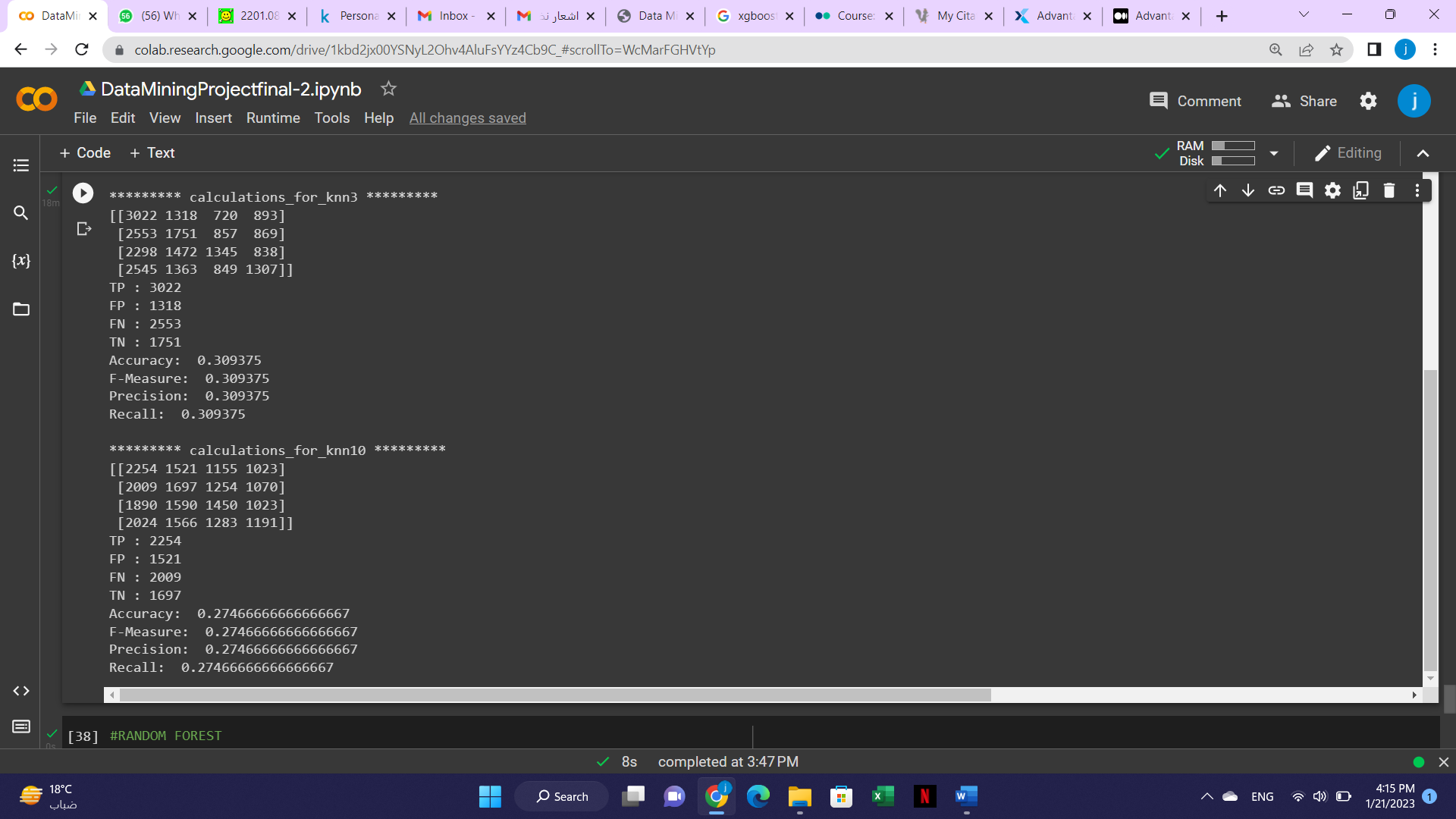


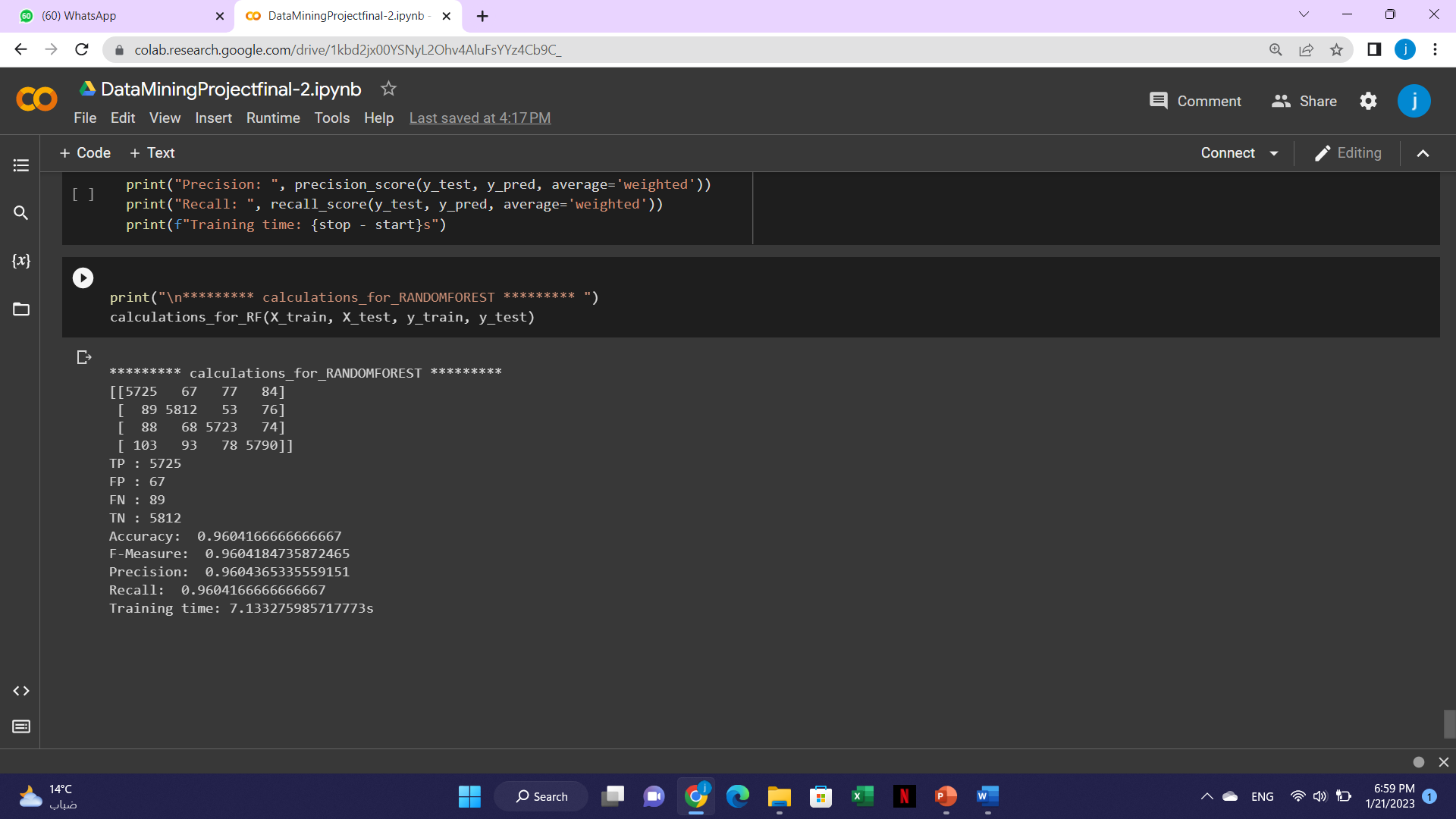










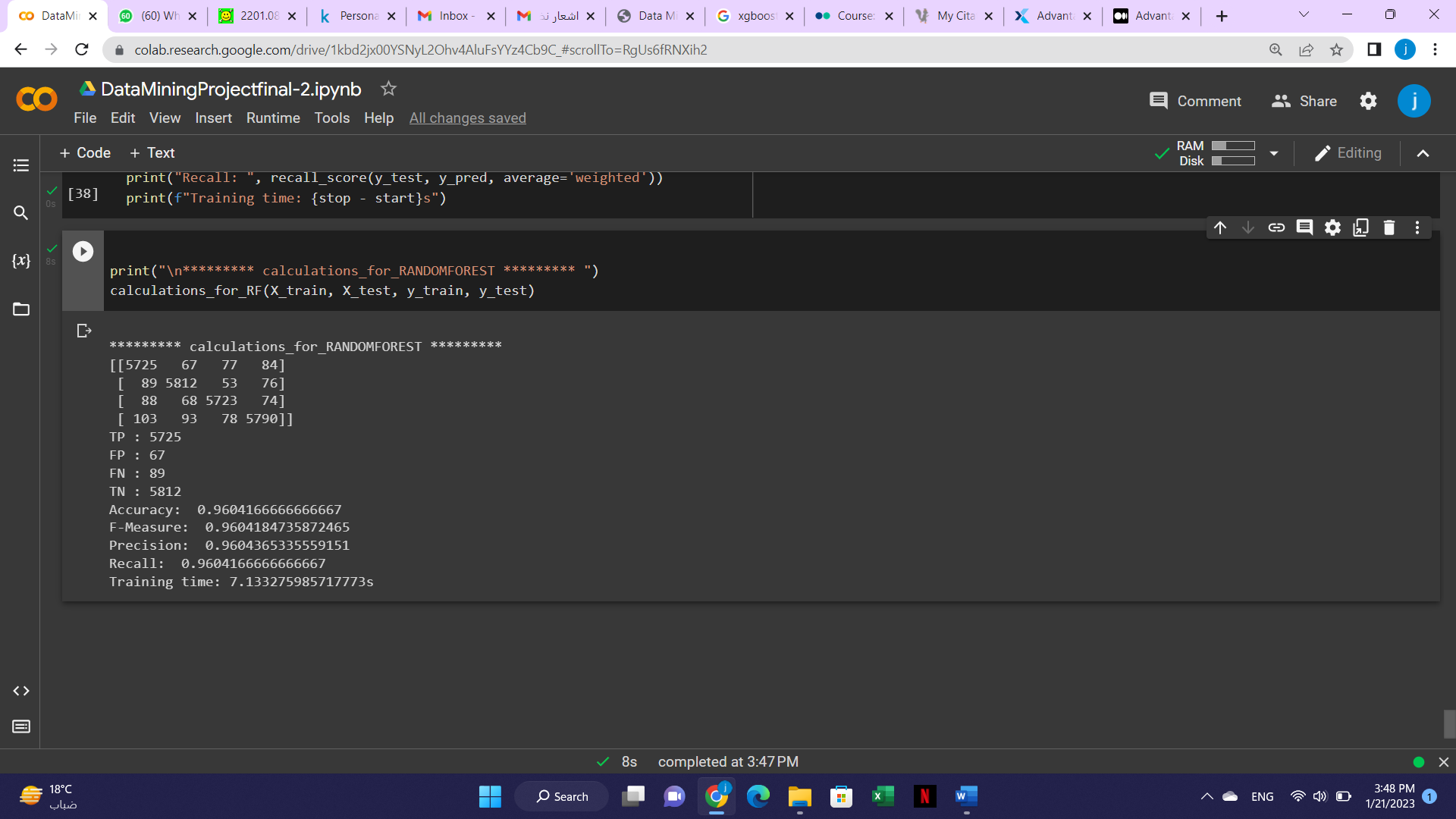


**The best mode was the XGBoost with the highest accuracy of 0.963**

**Conclusion :**

After we have evaluated the classification models that we applied to the dataset, XGBoost model gave the best accuracy, which is 96.7%, the second-best model was the random forest with accuracy of 96.2%. Unfortunately, the problem that we faced was the time consumption it took the XGBoost model to fit it to the testing data. In contrast, random forest model took a lot less time to fit the model.

In conclusion, we have found that the best classification model for the personality dataset was the random forest model.



|  |  |  |
| --- | --- | --- |
| **Method** | **Advantages** | **Disadvantages** |
| Decision tree | •Interpretability •Less Data Preparation•Non-Parametric | •Overfitting•Feature Reduction & Data Resampling |
| XGBoost | •It is Highly Flexible.  •It uses the power of parallel processing.  •It supports regularization | •XGBoost does not perform so well on sparse and unstructured data. |
| SVM | • In high-dimensional spaces, “Support Vector Machines” are more effective.  • effective when the number of dimensions exceeds the number of samples .  • use small amount of memory | • The “Support Vector Machines” algorithm usually perform poorly for large datasets.  • don’t work with noisy data  • may perform poorly when the number of features in each data point outstrips the sample size of the training data . |
| Naïve Bayes | • used to solve multi-class prediction problems quickly and efficiently. | • presupposes that all characteristics are independent, which is seldom the case. This restricts the algorithm’s usability in real-world use applications  • This approach encounters the “zero-frequency problem” in which it assigns zero probability to a categorical variable whose category is in the test data set . Also, it was not included in the training dataset .  • Because its predictions might be inaccurate in some instances, its probability outputs should not be taken too seriously. |
| KNN | •Easy Implementation  •No Training Period | •Does not work well with large dataset  •Sensitive to noisy and missing data |

**References**

* Mehta, A. (2022, June 15). *Personality Classification Data: 16 personalities*. Kaggle. Retrieved January 21, 2023, from https://www.kaggle.com/datasets/anshulmehtakaggl/60k-responses-of-16-personalities-test-mbt?select=16p-Mapping.txt
* Personality Type Based on Myers-Briggs Type Indicator with Text Posting Style by using Traditional and Deep Learning PDF
* Kapil, A. R. (2022, October 1). *Advantages and disadvantages of Decision Tree in machine learning*. Blogs & Updates on Data Science, Business Analytics, AI Machine Learning. Retrieved January 21, 2023, from https://www.analytixlabs.co.in/blog/decision-tree-algorithm/