

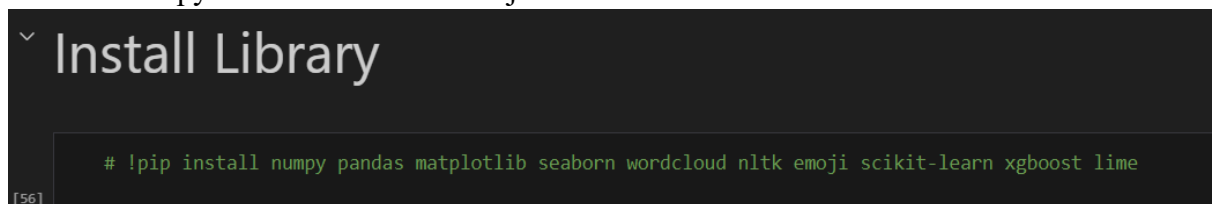
Demo test for the Machine Learning Engineer position

I worked in **python version 3.10.13** so please install this version. Other version can be worked but if you use this version, you won't face any error.

Installation:

Create a virtual environment:

- If you have Conda in your system, just write below command in VS code terminal
 - **conda create -n venv python=3.10.13**
- For activate the venv environment
 - **conda activate venv**
- Another approach for creating and activate the virtual environment windows is:
 - **python -m venv venv**
 - **cd venv**
 - **cd scripts**
 - **activate**
 - **cd ..**
 - **cd ..**
- Go to the folder path where you code, dataset and requirements files are, and install the requirements if you want install from terminal otherwise you can install requirements from the ipynb code notebook just uncomment the first line of code.



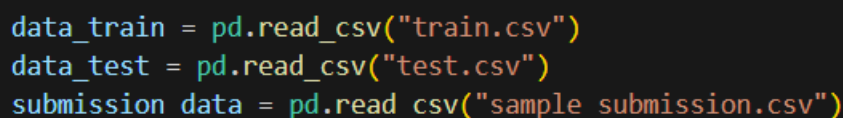
```
~ Install Library

# !pip install numpy pandas matplotlib seaborn wordcloud nltk emoji scikit-learn xgboost lime
```

- **pip install -r requirements.txt**

Read the Dataset:

Please make sure the dataset path, it can be absolute path or relative path, I worked in windows and took the “**relative path**”.



```
data_train = pd.read_csv("train.csv")
data_test = pd.read_csv("test.csv")
submission_data = pd.read_csv("sample_submission.csv")
```

Questions Answers:

Question 1: Data handling

Answer 1.1: I have used pandas for read the dataset.

Answer 1.2: Preprocess email column text, remove unnecessary text from the original text. As it's a classification so i have used stemming for obtaining the base word. Then use TF-IDF vectorizer for converting the text into numeric value and it's a matrix. TF-IDF based on the frequency of a word in the corpus but it also provides a numerical representation of how important a word is for statistical analysis.

Answer 1.3: In EDA, I have illustrated the class number and word cloud. I have demonstrated number of spam and not_spam class are in the dataset using bar plot and pie chart. Bar plot represents number of each class and pie chart represent the percentage of each class. Word cloud represent the most frequency word is in the text which word size is large that word frequency is more than others word frequency.

Question 2: Spam detection

For Spam detection, I have used Multinomial Naïve Bayes algorithm.

Explainable AI(XAI):

When performing inference or when users utilize our service, we often encounter a situation where we don't understand why an email is classified as spam or not spam, as machine learning and deep learning models are typically considered "**black boxes**." To gain insights into the reasons behind these classifications, we can employ Explainable AI (XAI) techniques. XAI helps us understand the decision-making process of the model, providing explanations for why an email is categorized as spam or not spam.