**[Video link](https://drive.google.com/file/d/1NBRGciwZ2qyf-BAyKp5t4f2KtbuhX12u/view?usp=sharing)**

# Project Part A: IMDb Movie Review Sentiment Analysis

REPORT

## Step-1 : environment

Genism is older library, it will create issues in google colab.

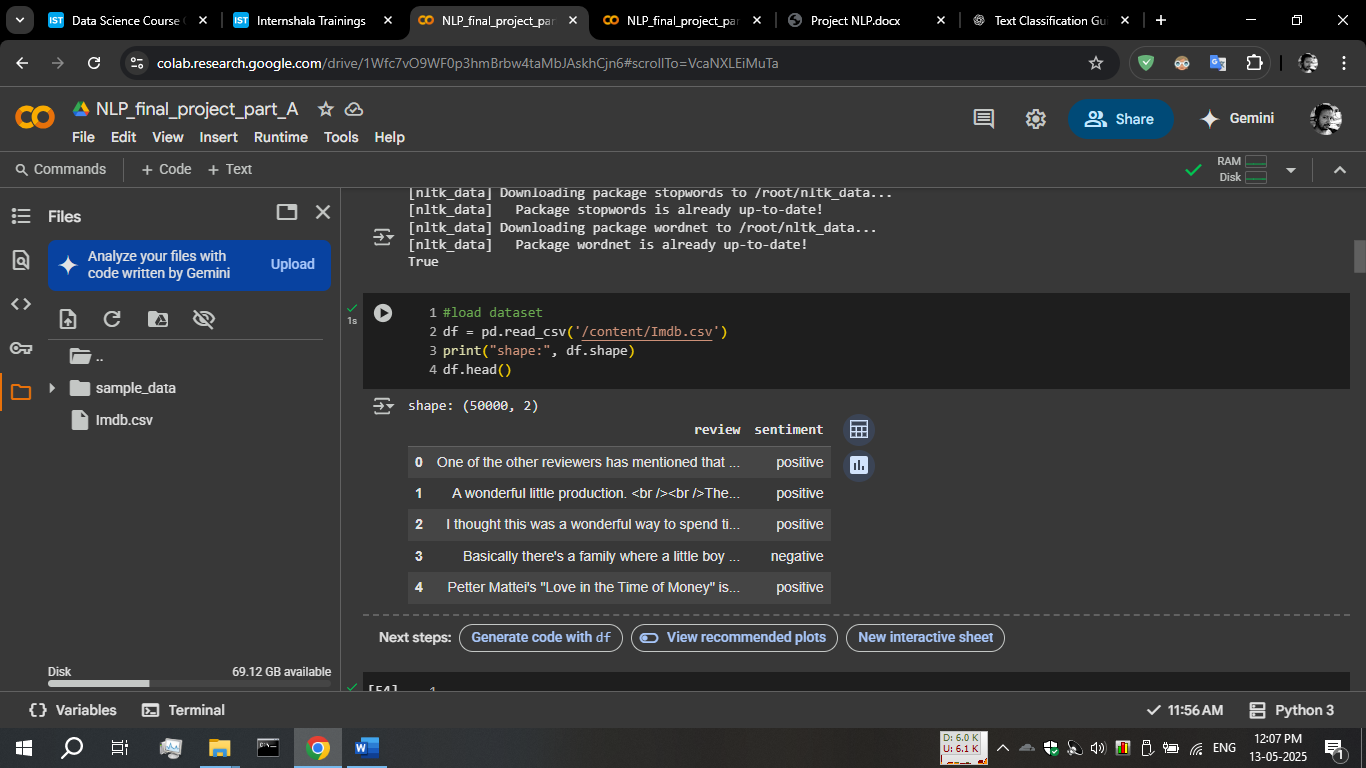
I downgraded version of some of the packages. so the gensim lib is compatible with other libraries.

first executing these commented code lines and restating the session.

!pip uninstall -y gensim numpy scipy

!pip install numpy==1.23.5 scipy==1.10.1 gensim==4.3

## Step-2: loading the dataset

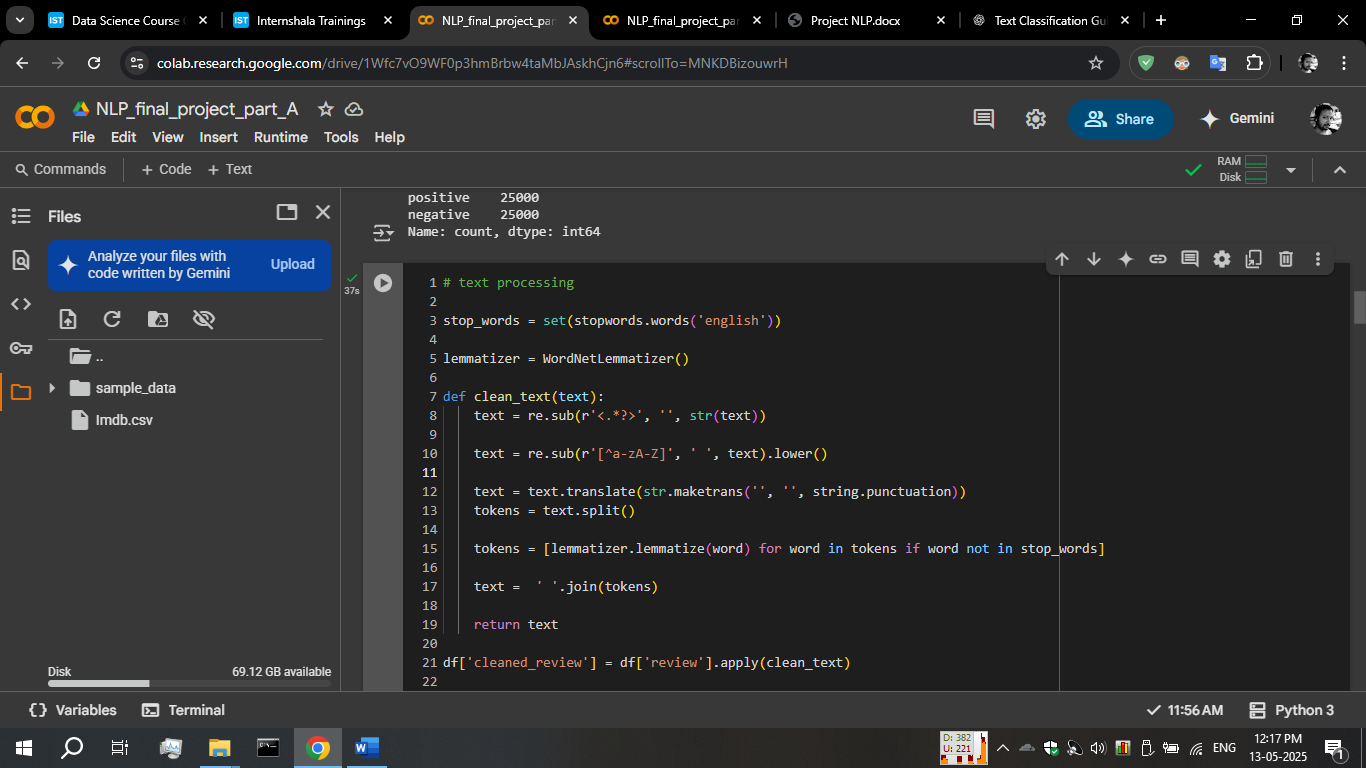


I used upload option in colab,

It appears orange on left side of the image.

I found that this dataset contains 50000 records of data containing reviews and their labelled sentiments.

## Step-3: text pre-processing



First I loaded the lemmmatizer using WordNetLemmatizer()function

then I created a function clean\_text(),

it is used for removing entire html tags.

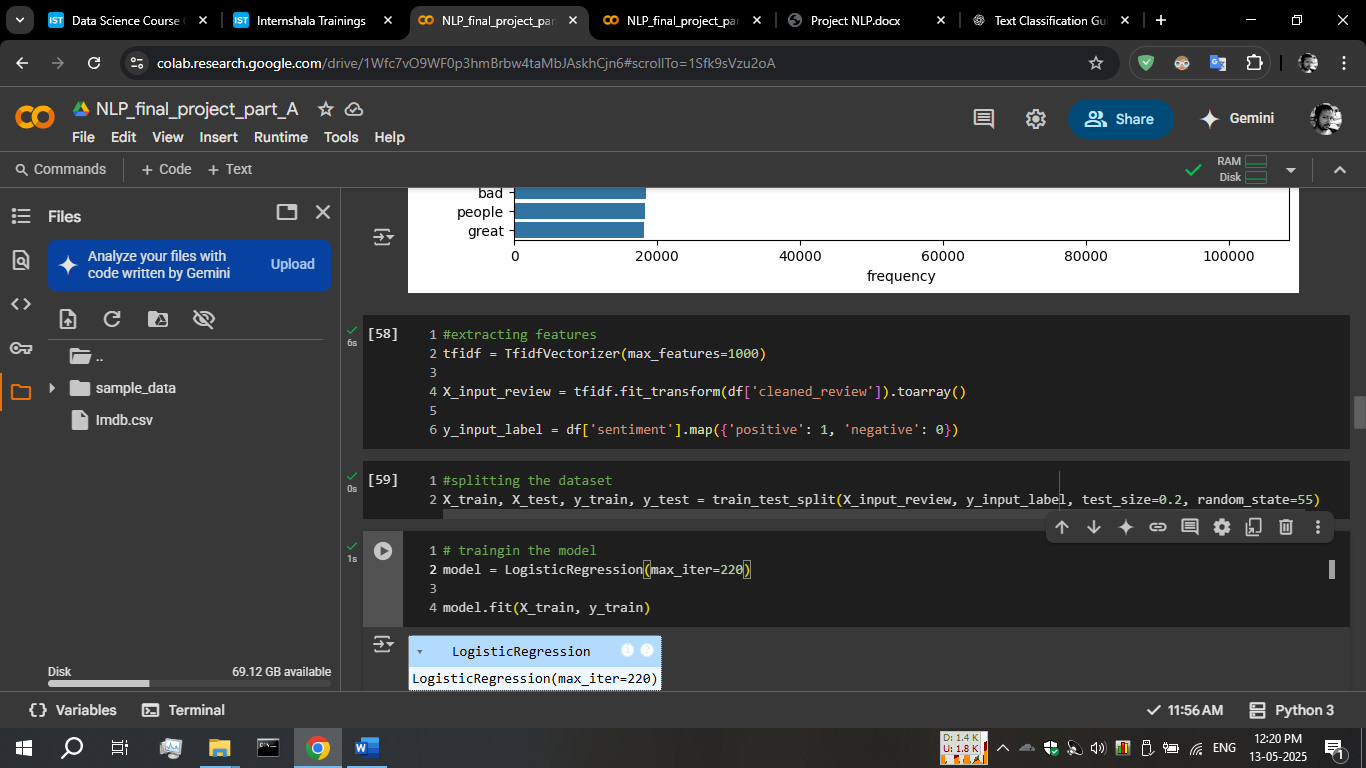
Then I used maketrans and translate function for removing any punctuation from the given text.

Then I applied lemmatizer for reducing the dimantionality

finally creating new column in dataframe and applying clean\_text() function into review column

df['cleaned\_review'] = df['review'].apply(clean\_text)

## step-4: extracting features and vectorization



I used TfidfVectroizer for vectorization and then I applied fit\_trnsform() function to vectorize the review column

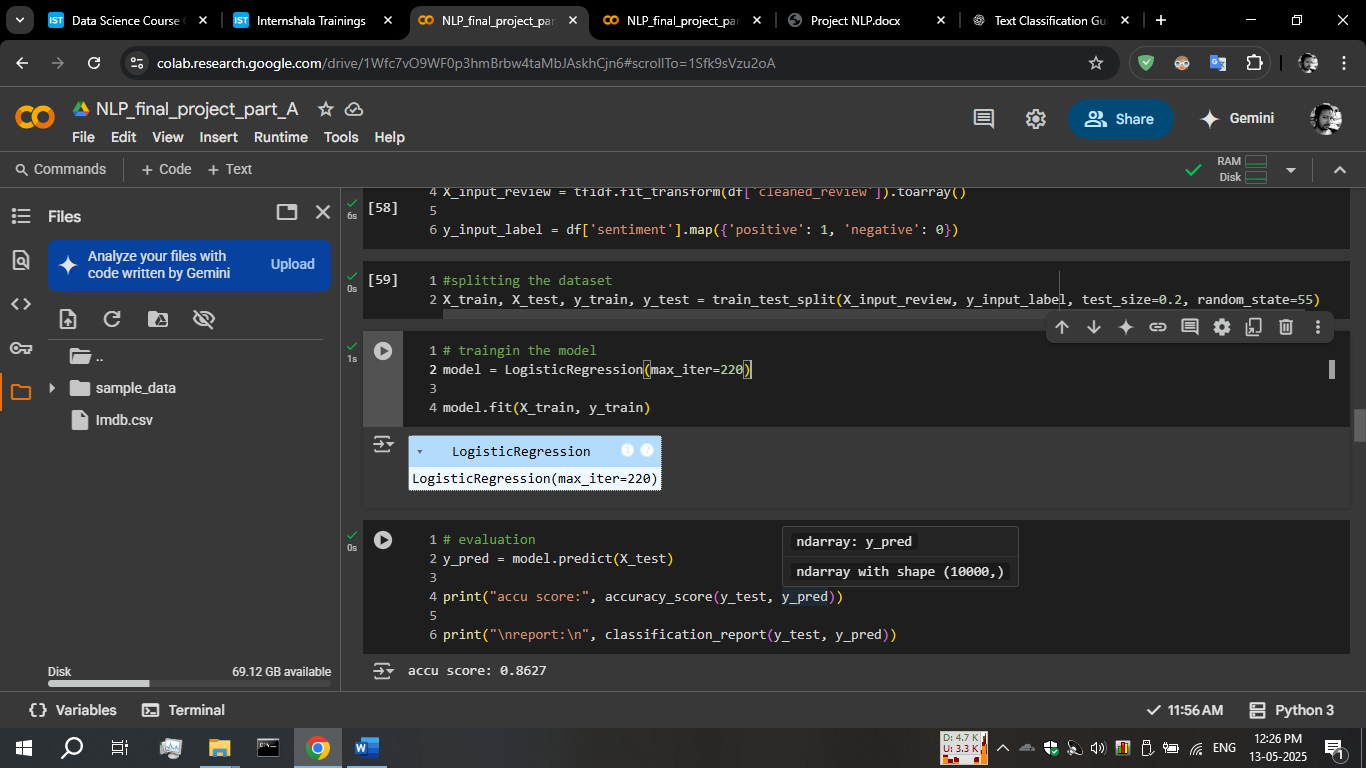
After vectorization, toarray() function is used to make model understand the input data.

In y\_input\_label,

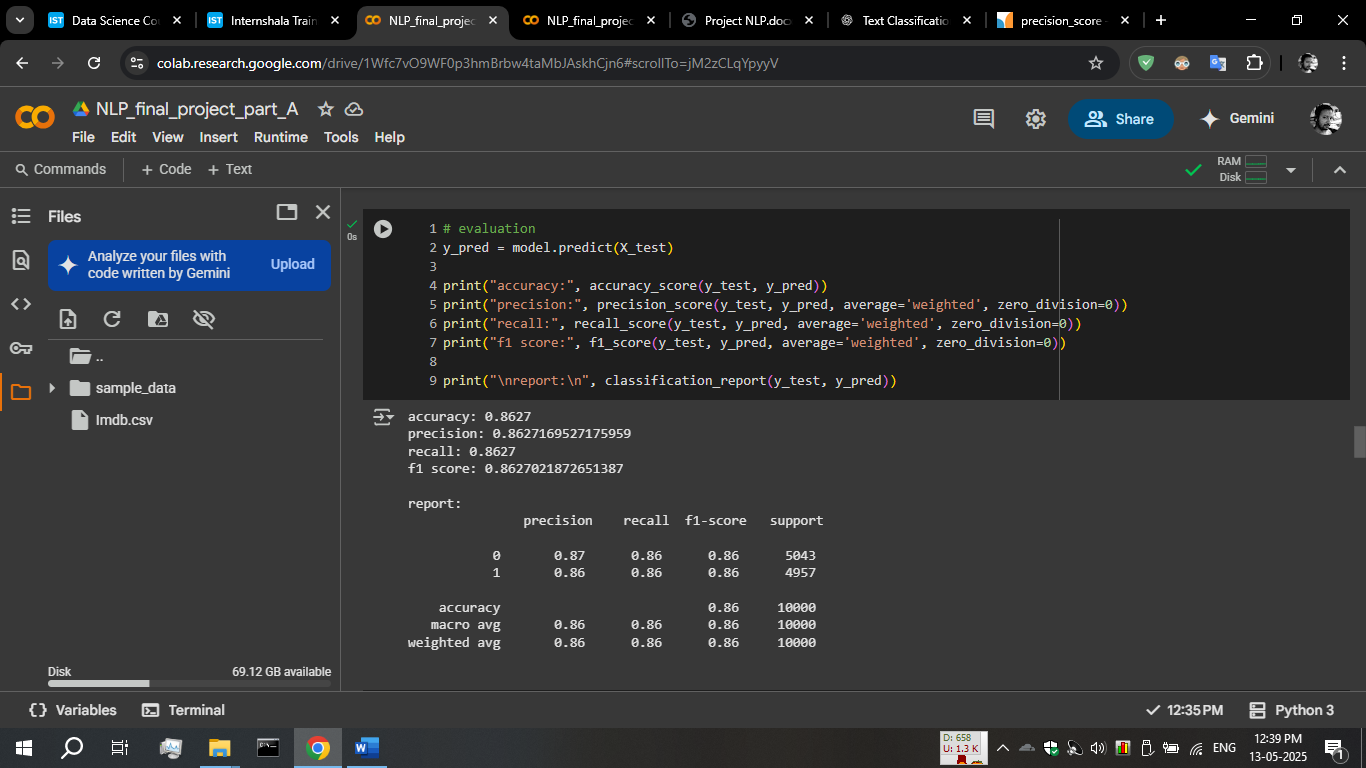
Sentiment is mapped to its numerical values for the same, so model can understand the input data

## Step-5: model training

I’m using LogisticRegression model with iteration of 220.



## Step-6: evaluating the model

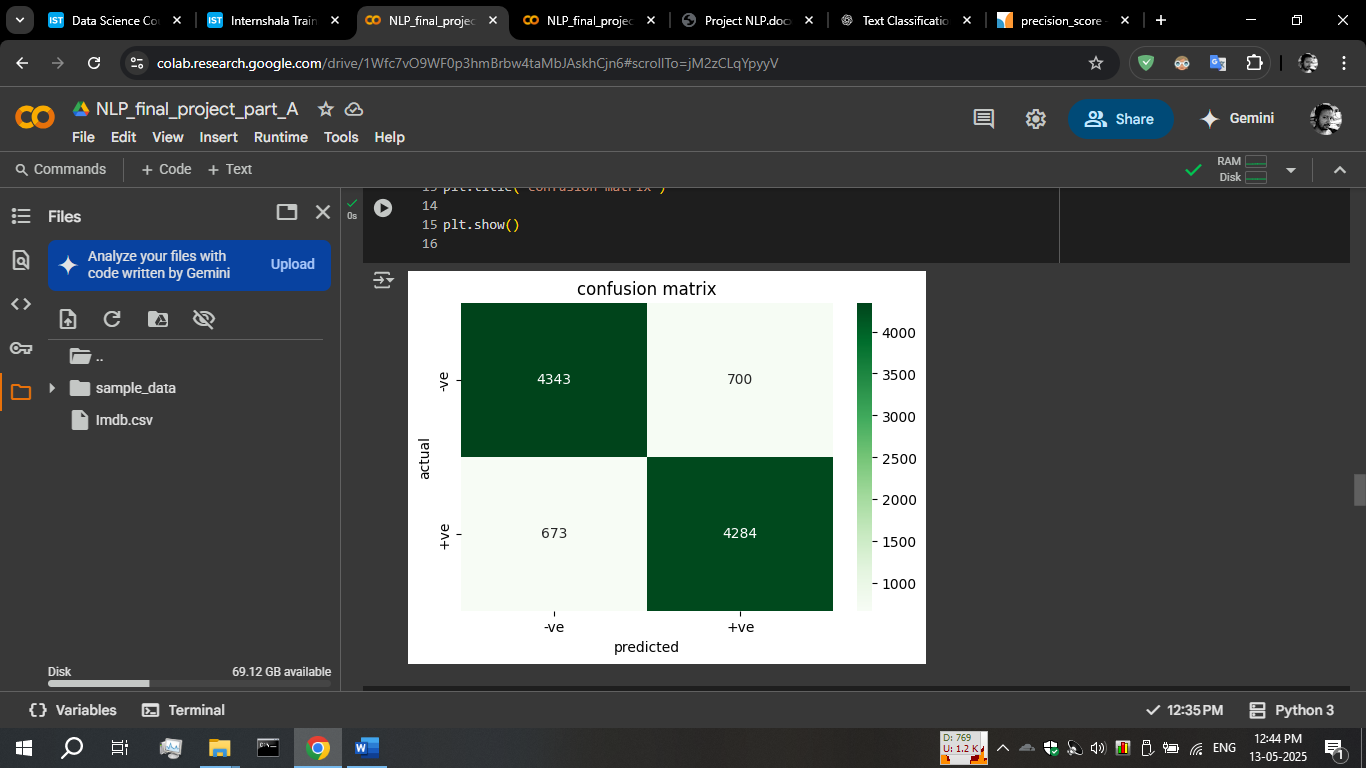


As appears model accuracy is 0.86, model predicts the 86% of the samples correctly.

Precision, recall and f1-score both appears around 86% as well.

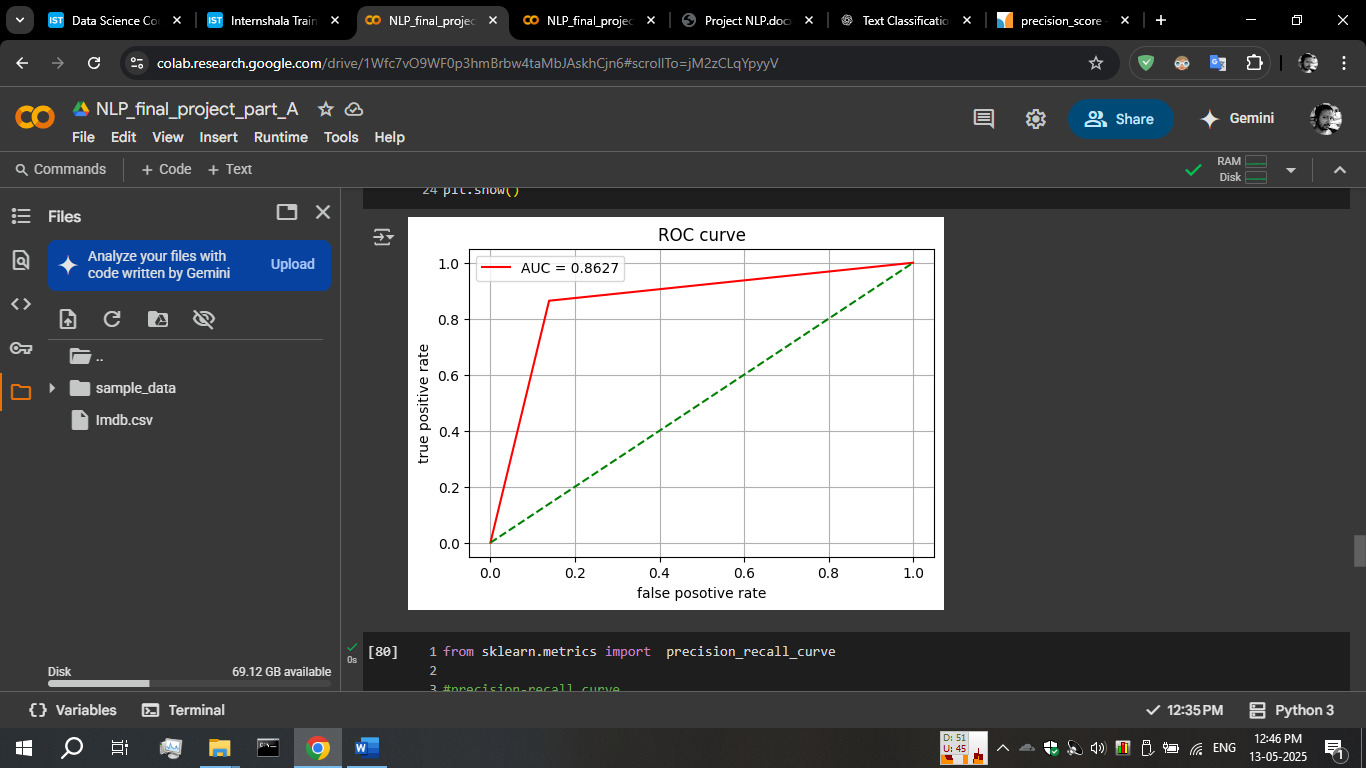
So, model performance and prediction capability are fine

Confusion matrix:



Most of the time model predicting accurate negative and accurate positives, with low false-negatives and false-positives.

ROC curve: our model performing well in distinguishing between two classes (positive or negative in sentiment).



Thank you