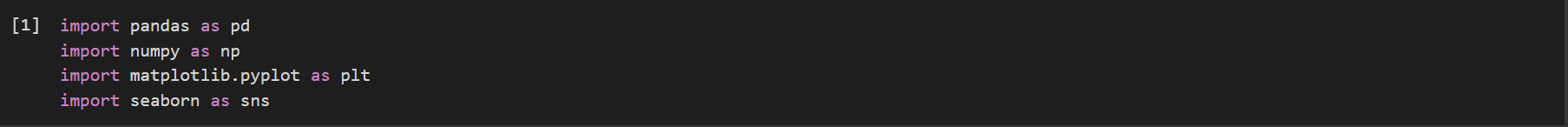
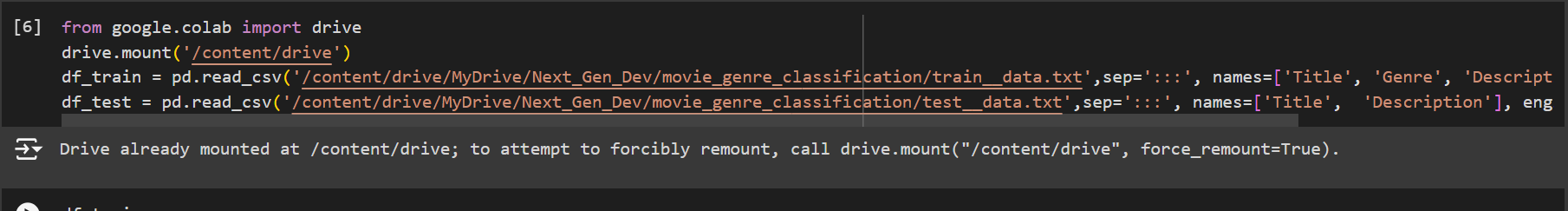
**Movie Genre Classification**

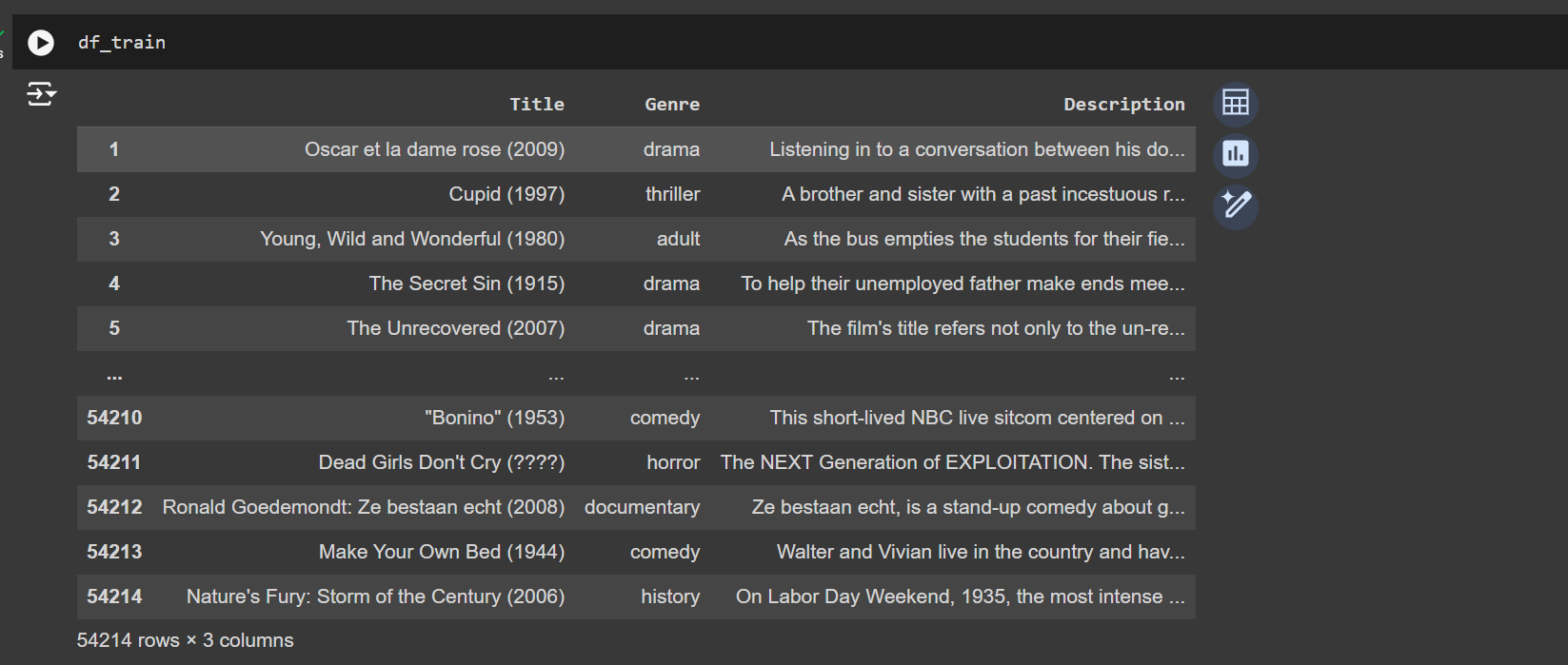
**Importing required modules and libraries: Here we imported numpy, pandas, matplotlib, and seaborn.**

****

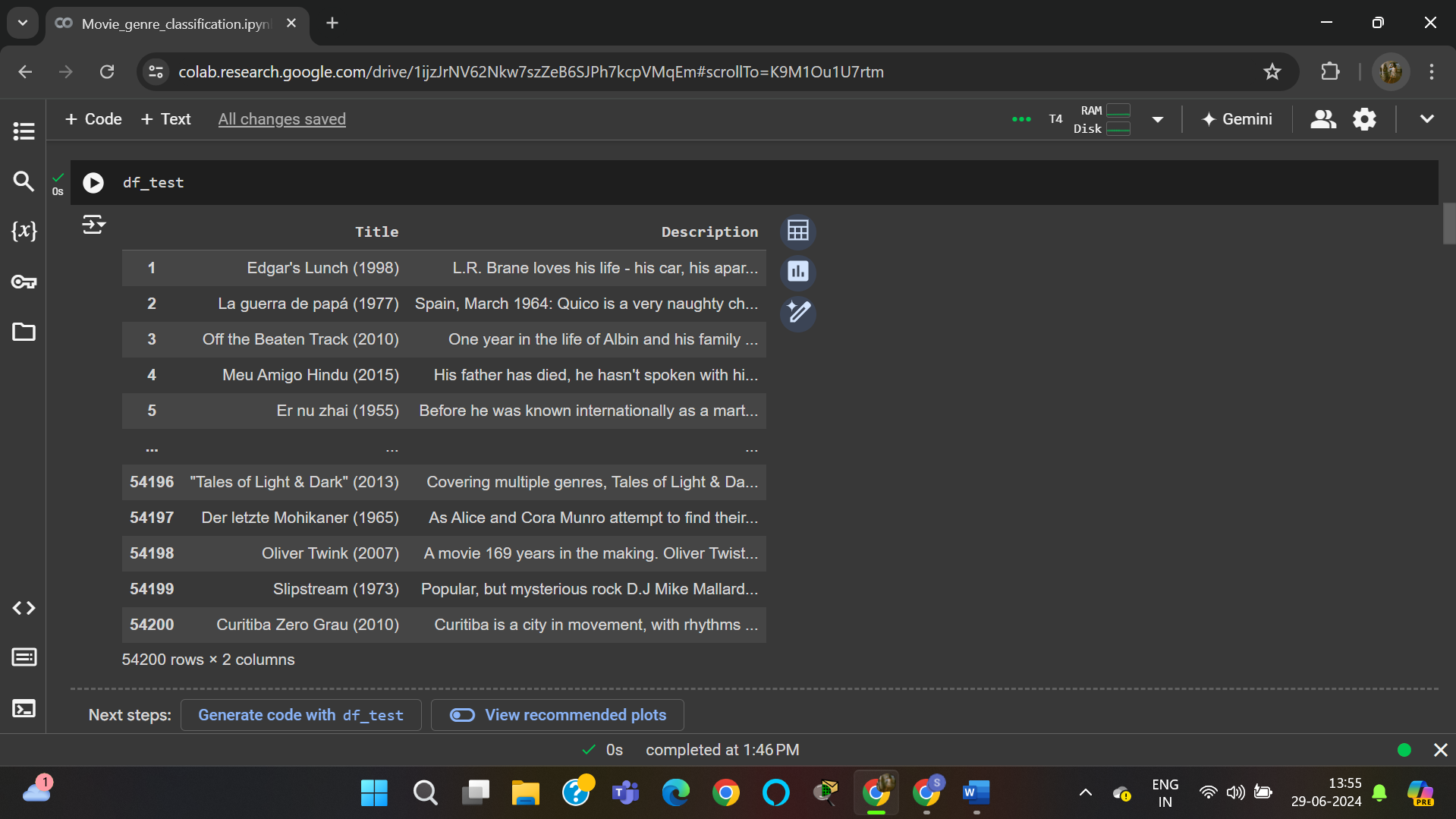
**Dataset Loading**: The data set is divided into 2 parts training data set and testing data set. Here we read both csv format data files and separate the title, genre and the description of the movie.

****

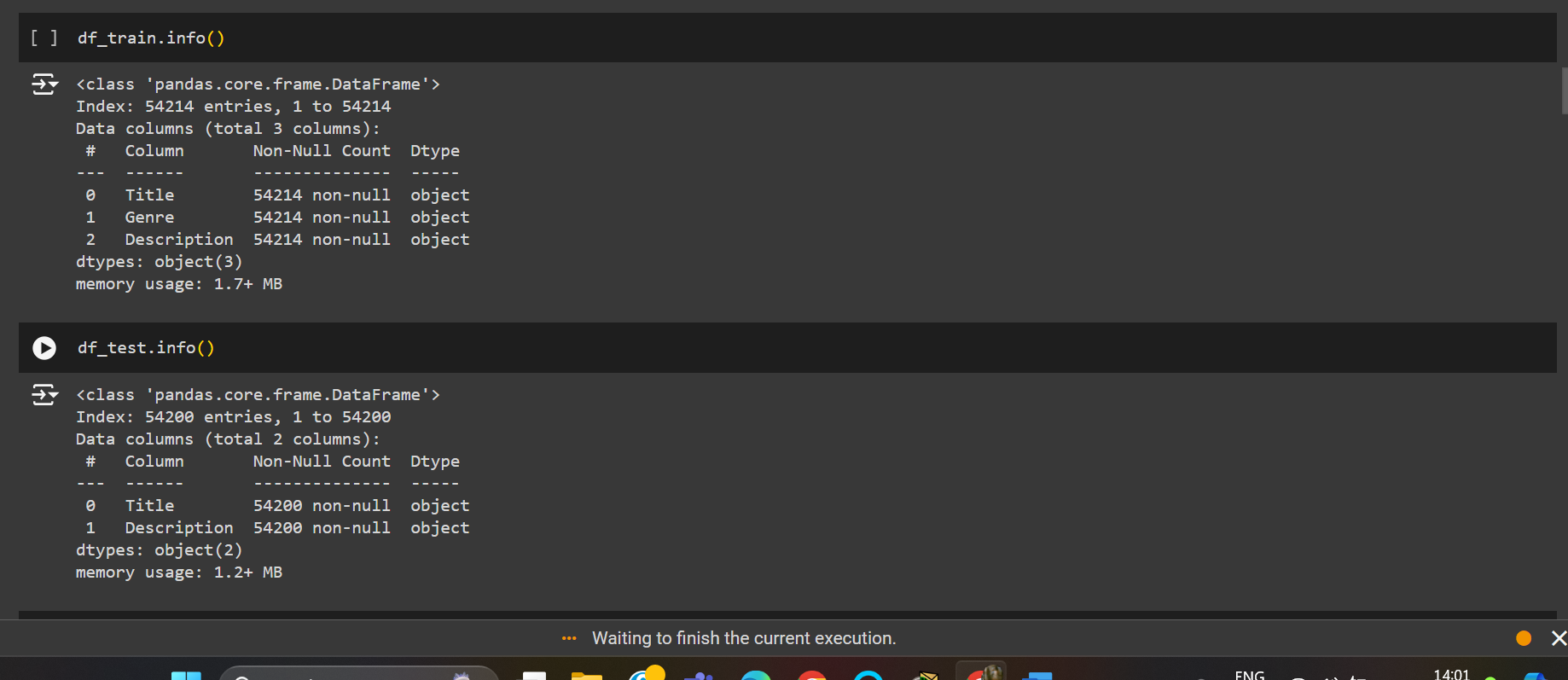
**Training dataset stored in data frame:** This command shows us all the data stored in the training file according to the format we read it with titles, genre and description separated.



Testing data set stored in data frame: This command shows us all the data stored in the testing file according to the format we read it with titles, genre and description separated.



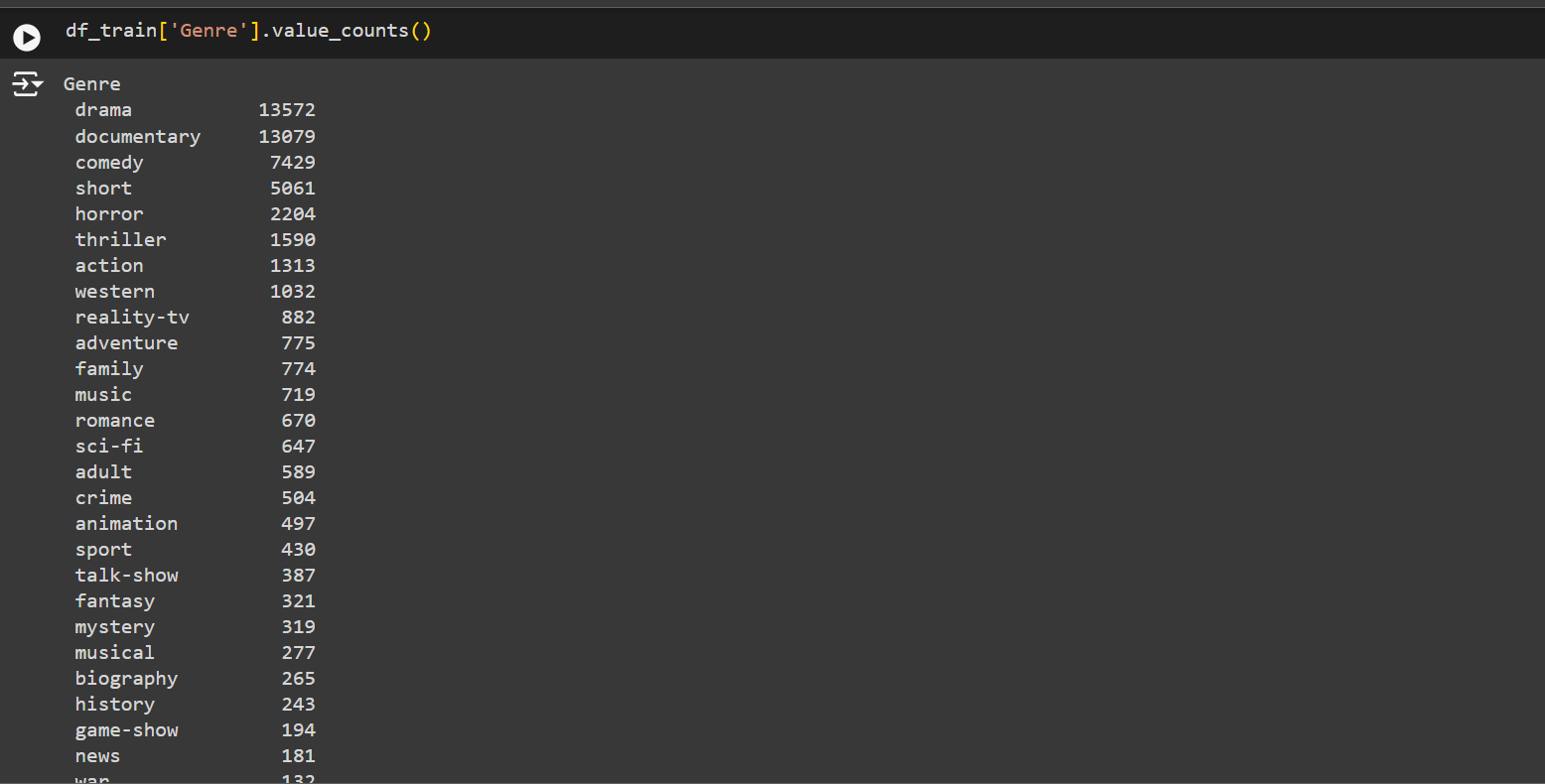
Displaying info about training and testing dataset:



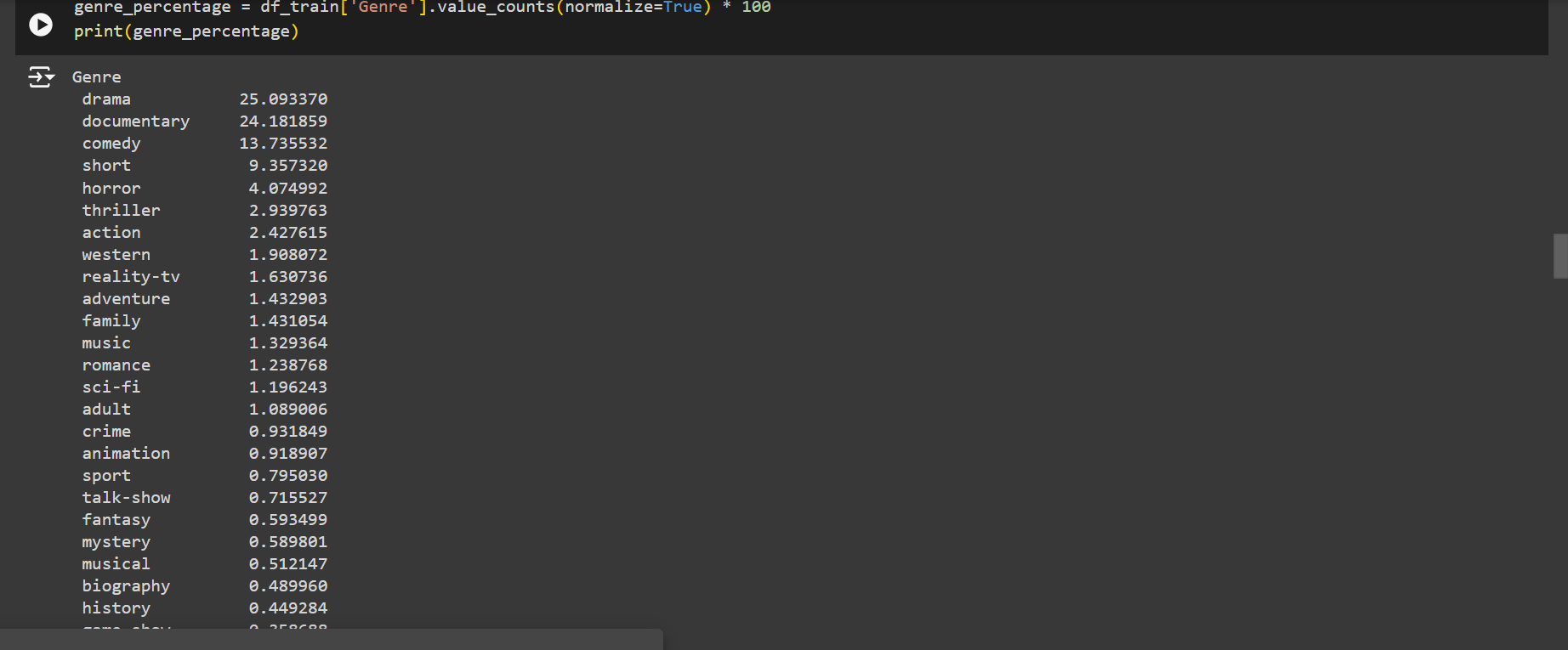
Removing duplicate values and displaying Training info without duplicate values:



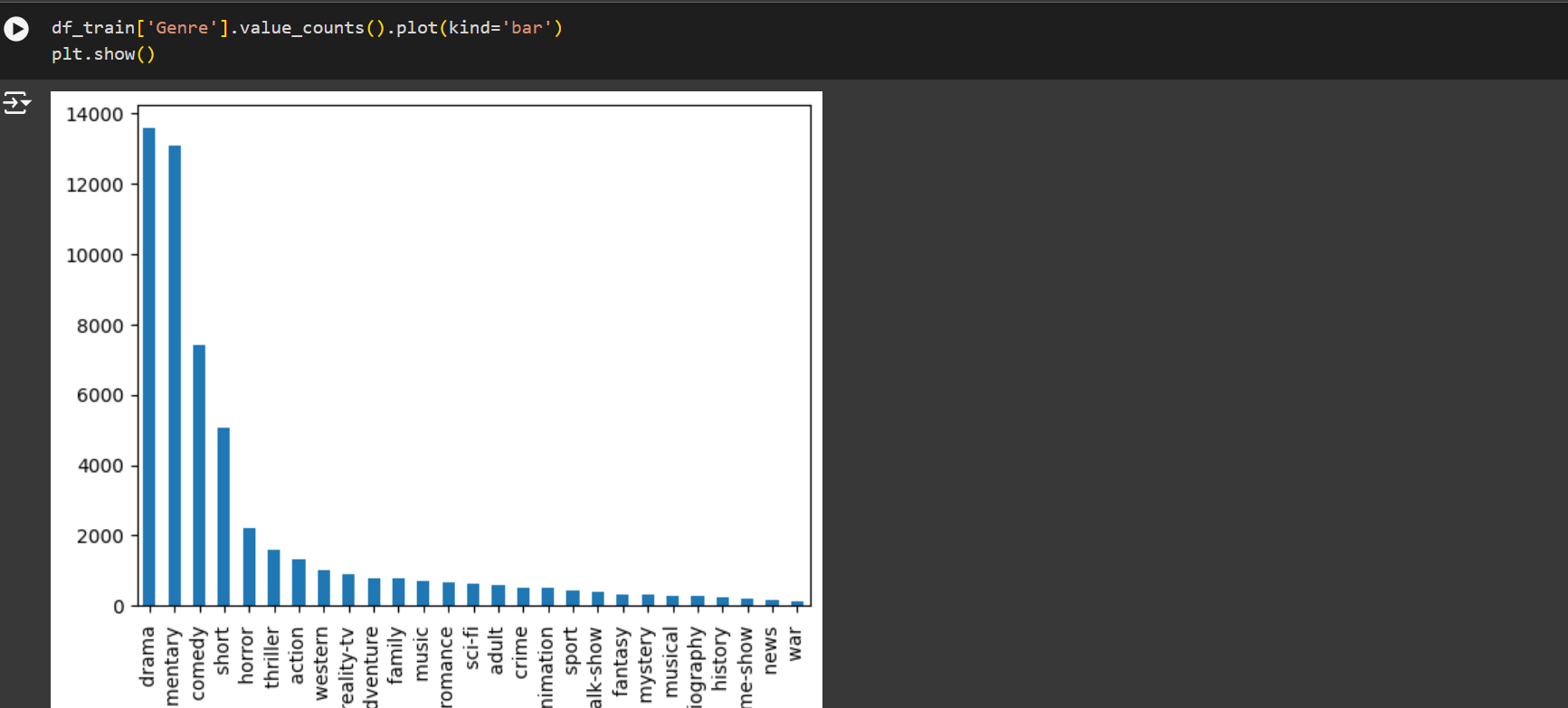
Total types and number of each genre in training set:



Showing percentages of different genres in the training data set (like drama genre is of 25% in whole data):



Displaying genre and the number of times the genre is found in training data set in a bar graph:

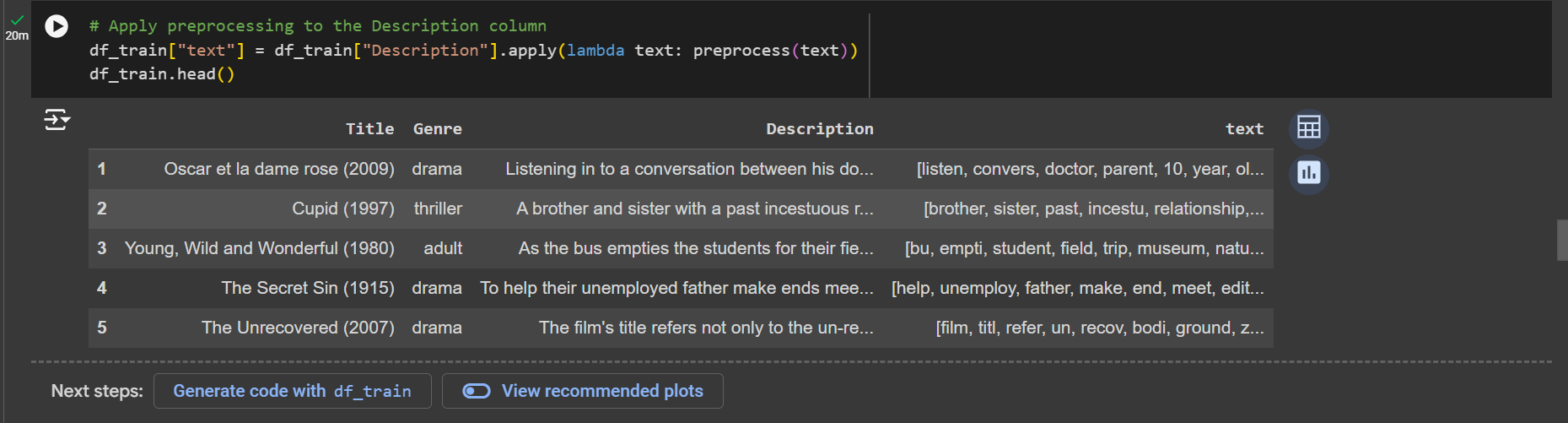


Downloading the natural language toolkit, and its modules required for reading and training of data:



Preprocessing the data:

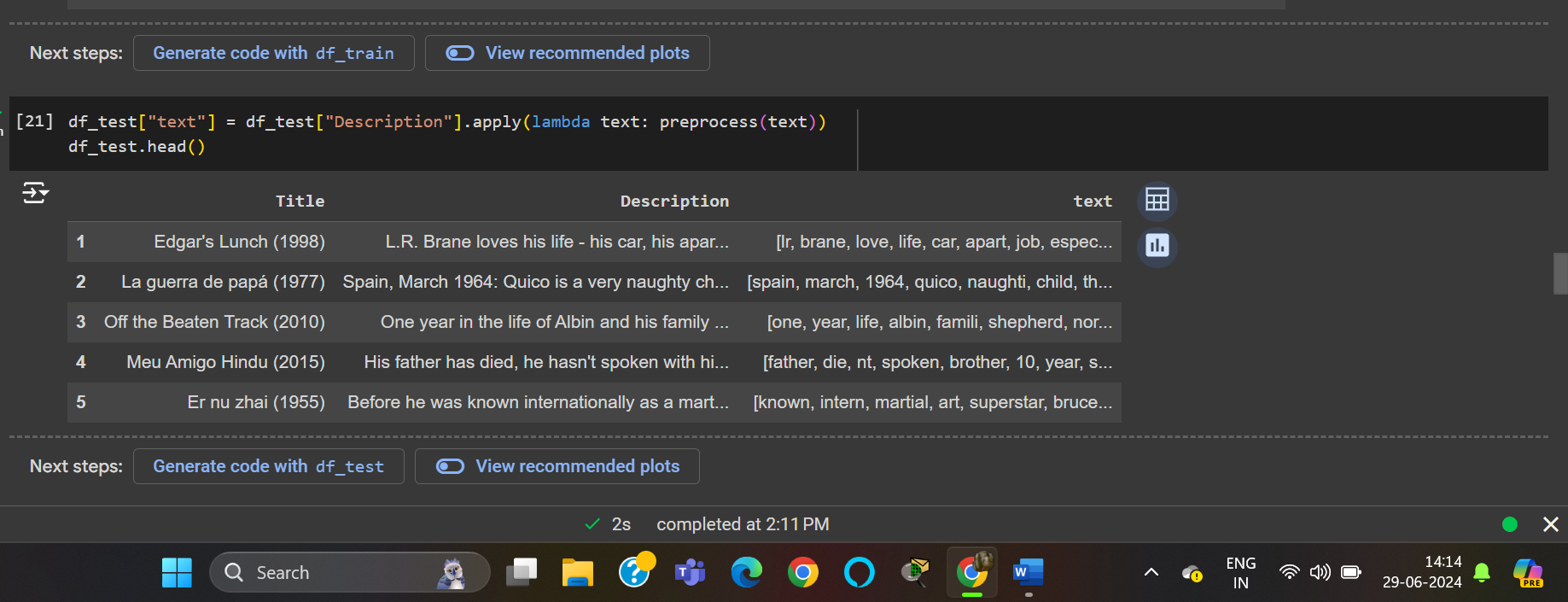
The train data file is in the preprocessed and separated description according to space, with comma into a data frame.

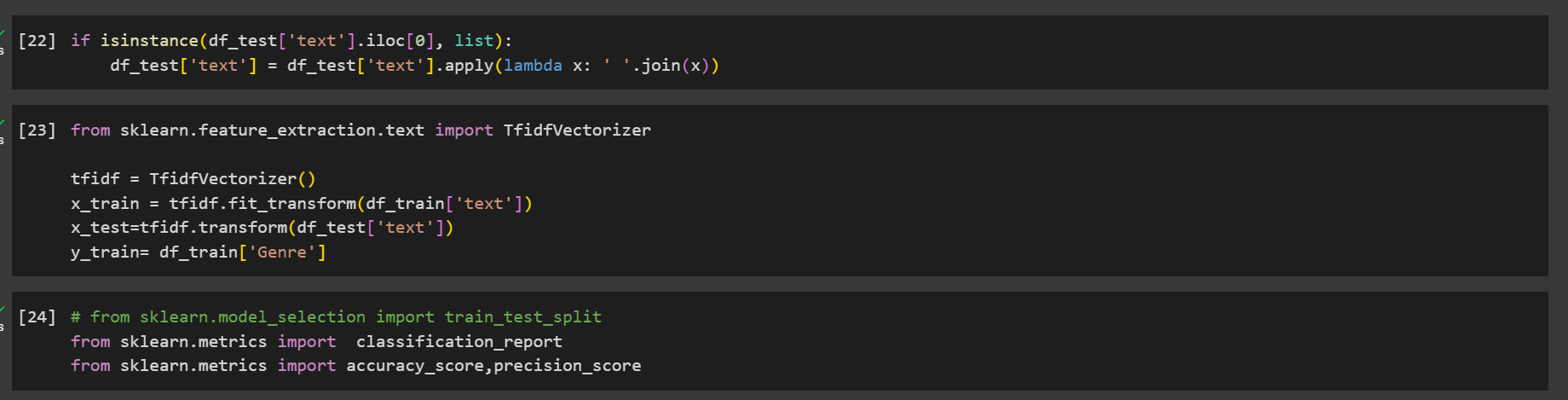


Now we are converting the description to lower case, removing punctuations and further preprocessing.

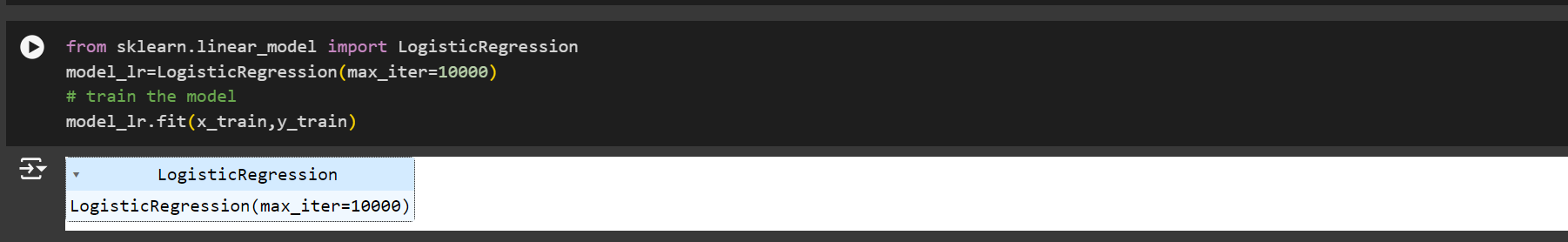




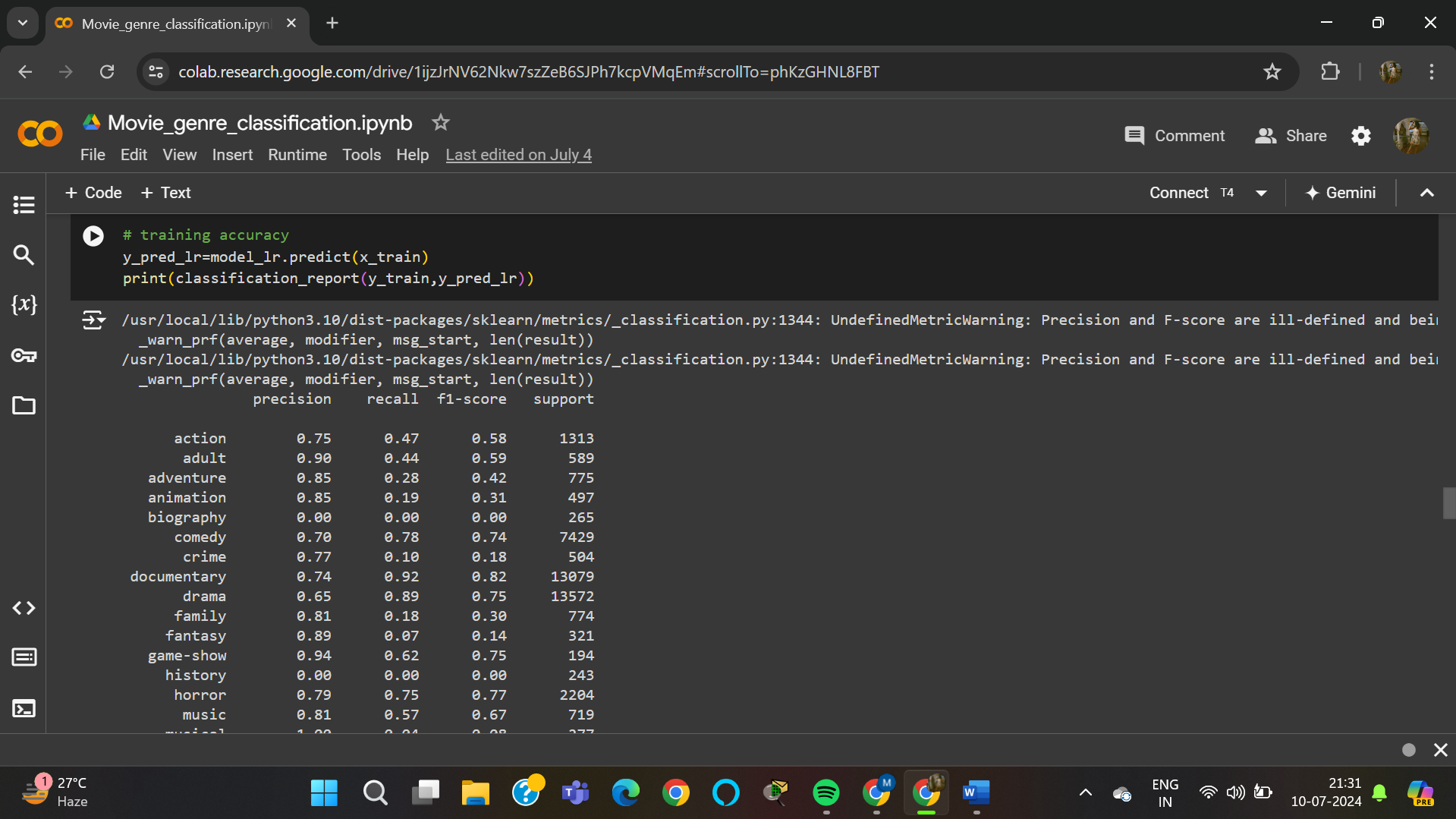


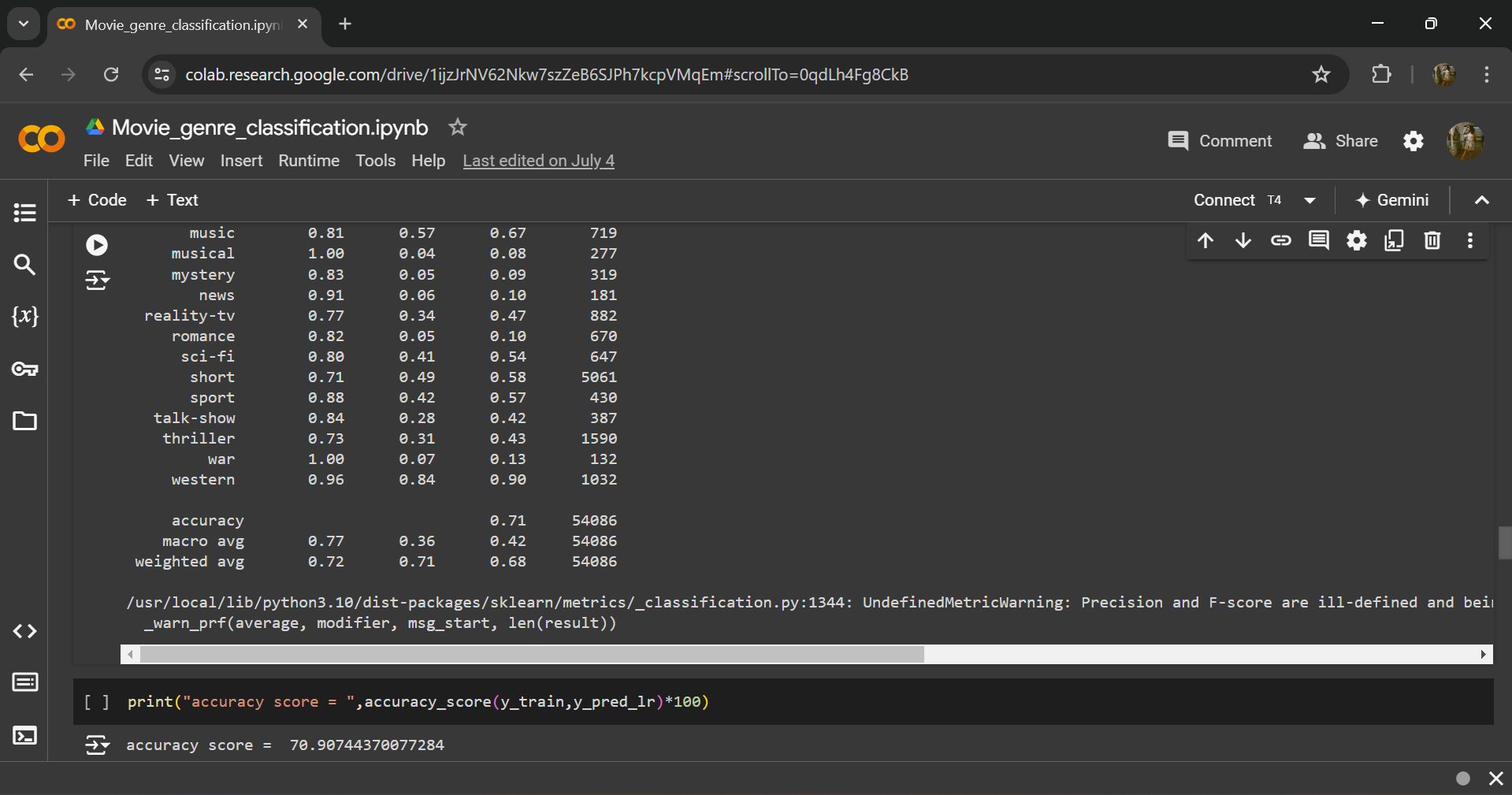


Logistic Regression model is applied from sklearn model so to fit the training data:



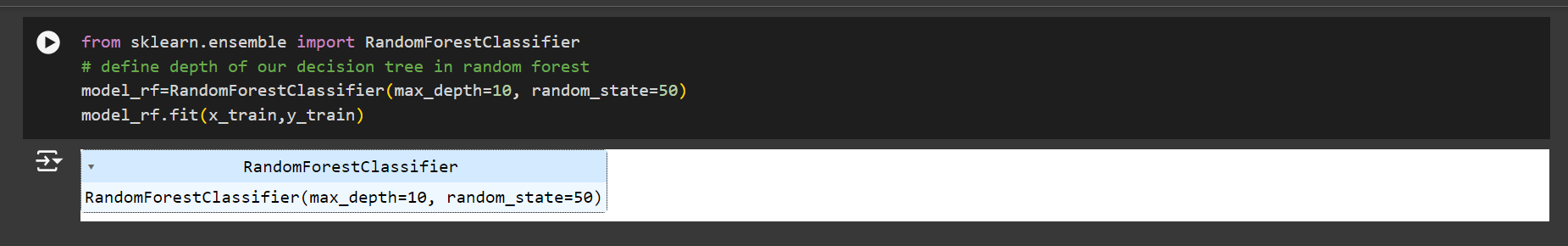
Now we use the model to predict the train and find out accuracy on the training model.



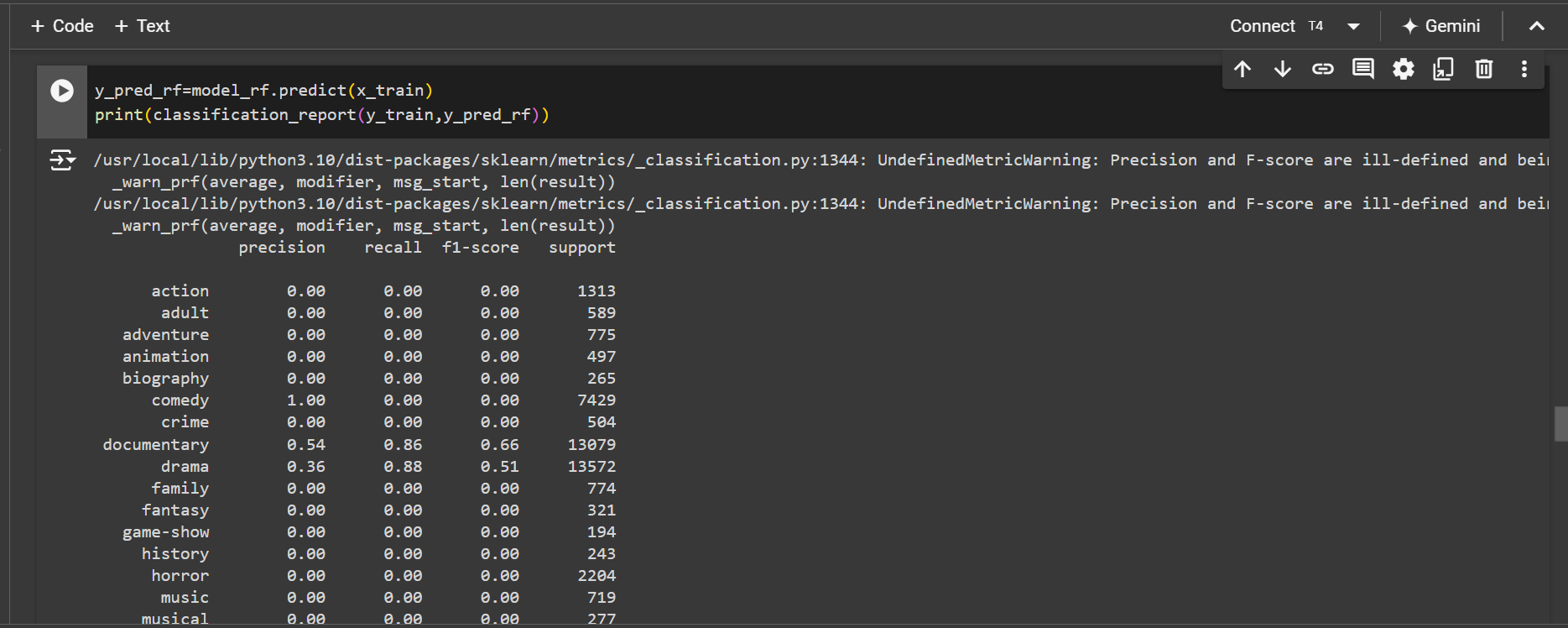


Here , we get the accuracy of 70.90 on prediction on training data.

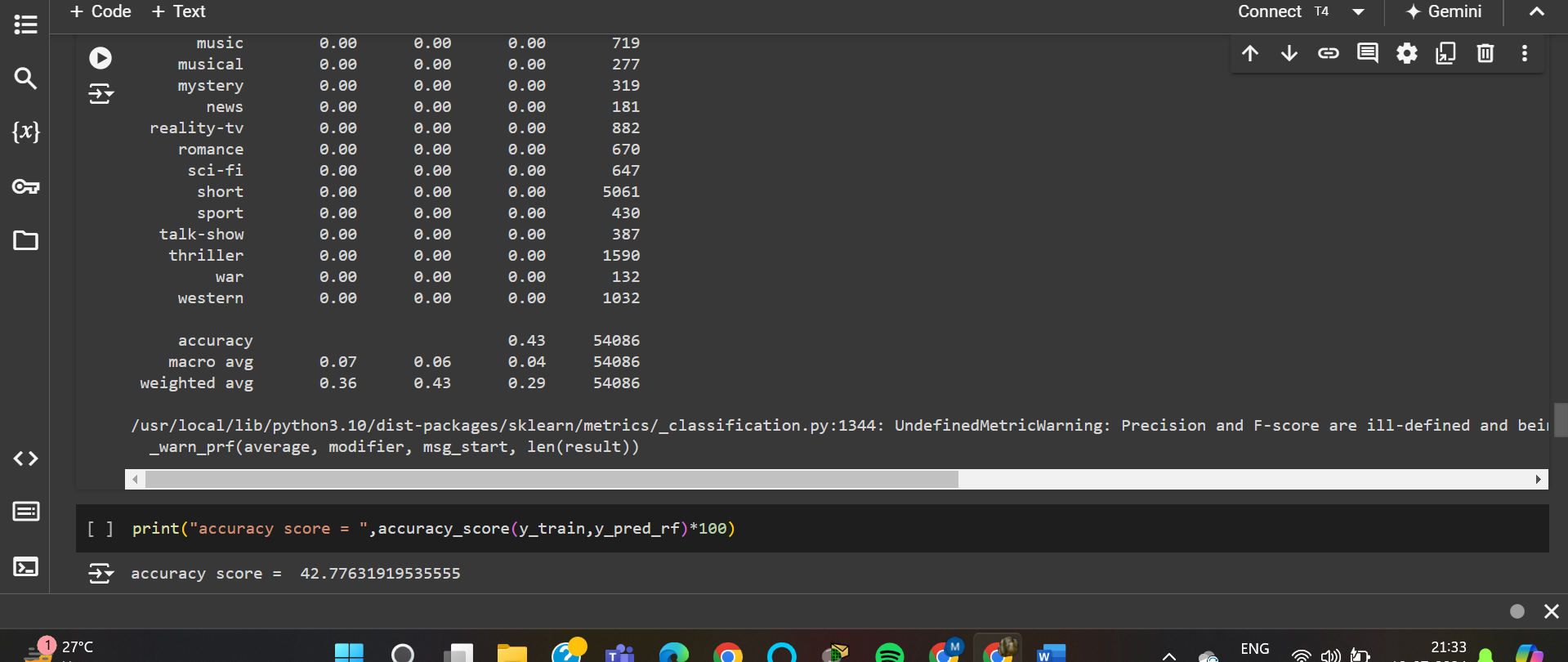
Now we are using Random Forest Classifier to decide where or which category does a particular description belong to:



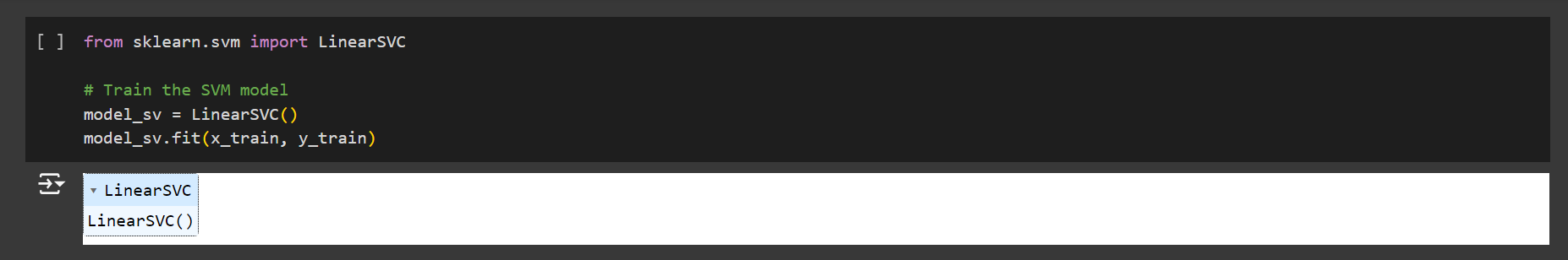
Now we measure the precision, recall, f1-score, support:

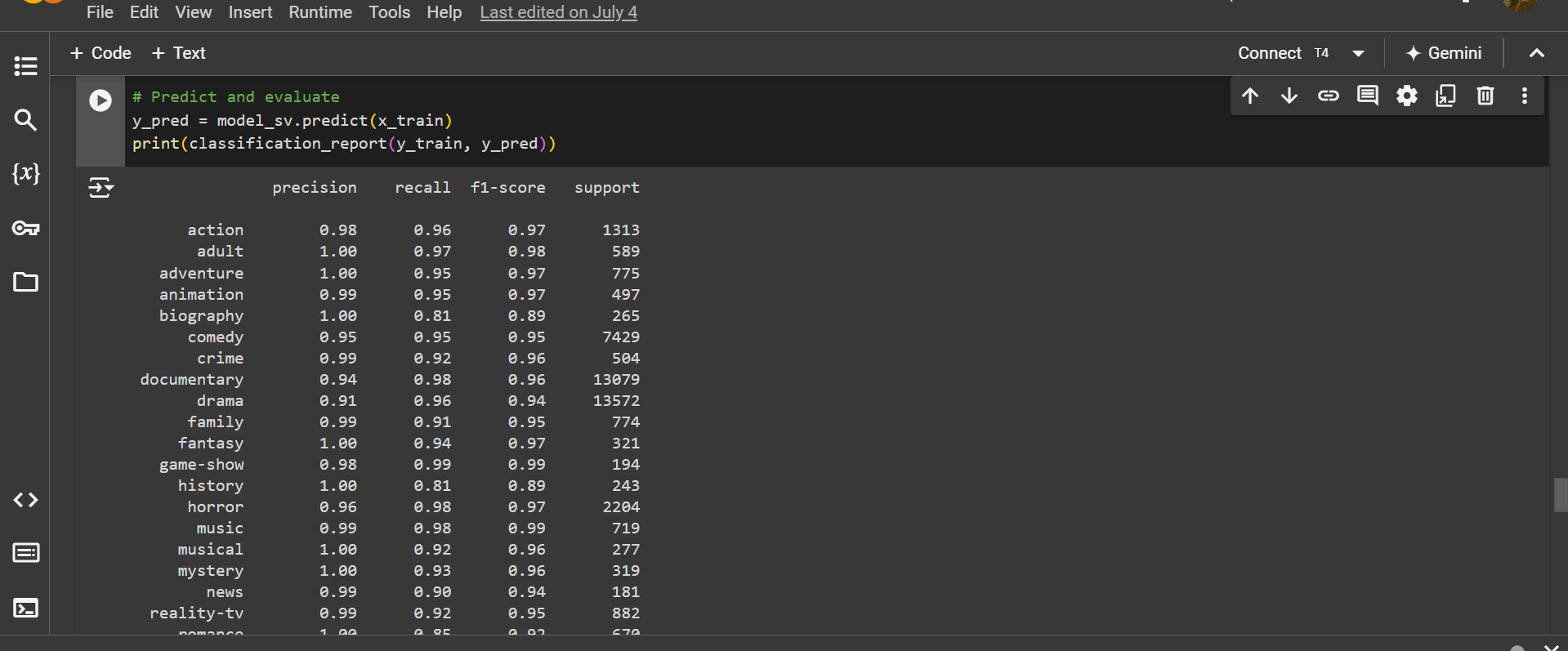


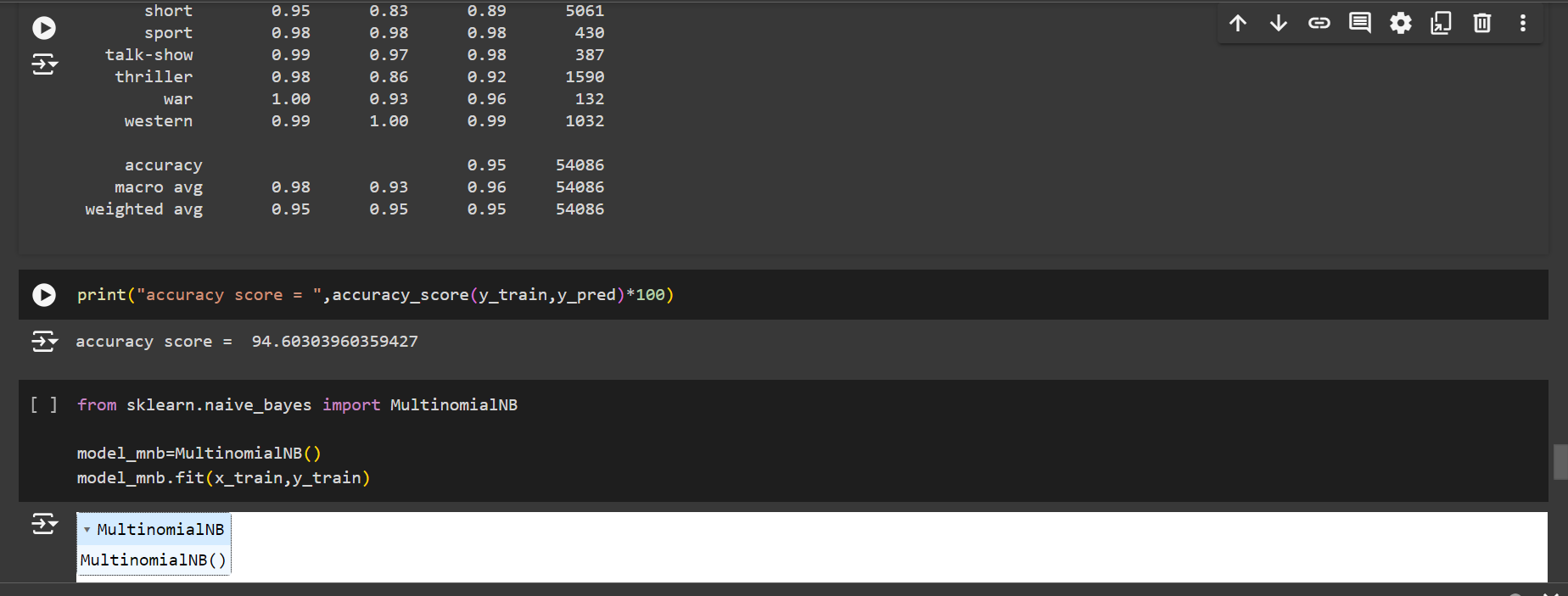
Accuracy score on the prediction:



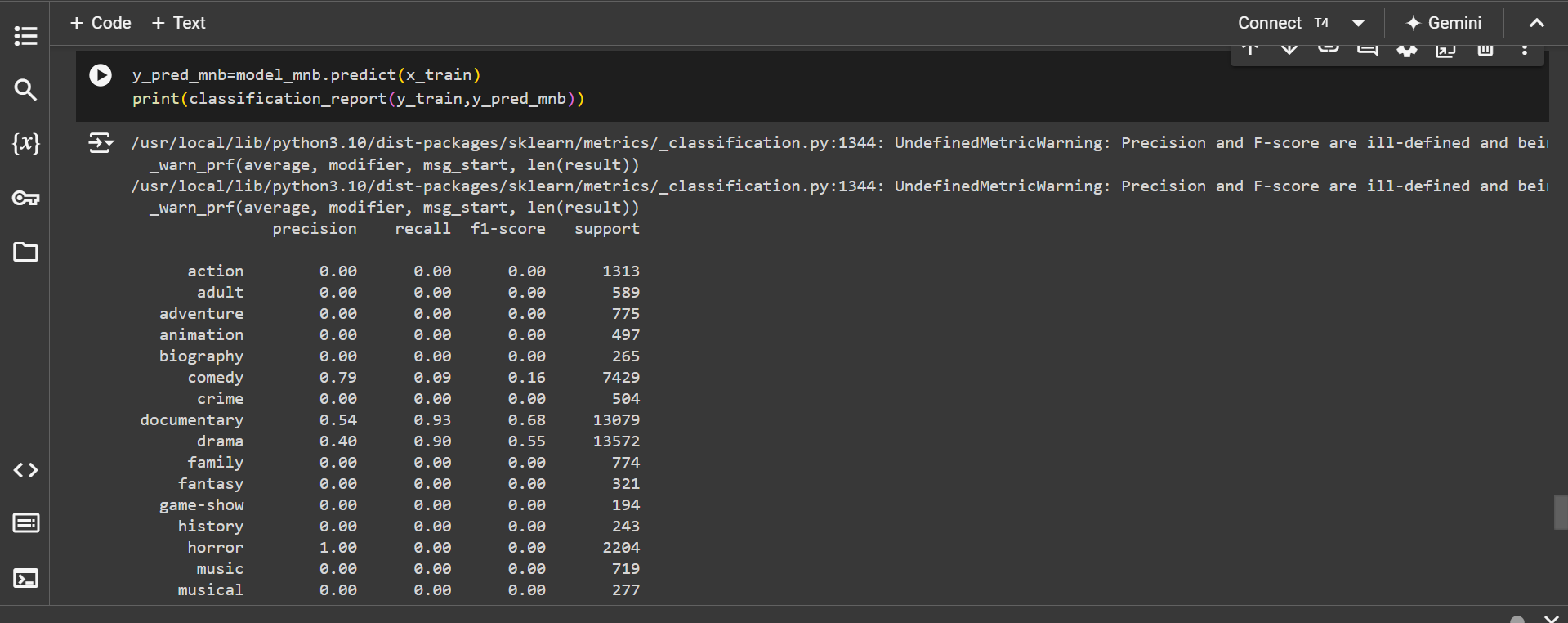
Applying Linear Support vector classification algorithm which reduces the distance between the classified data and collects more similar type of data with less distance into same class/ group. Hence it improves accuracy.

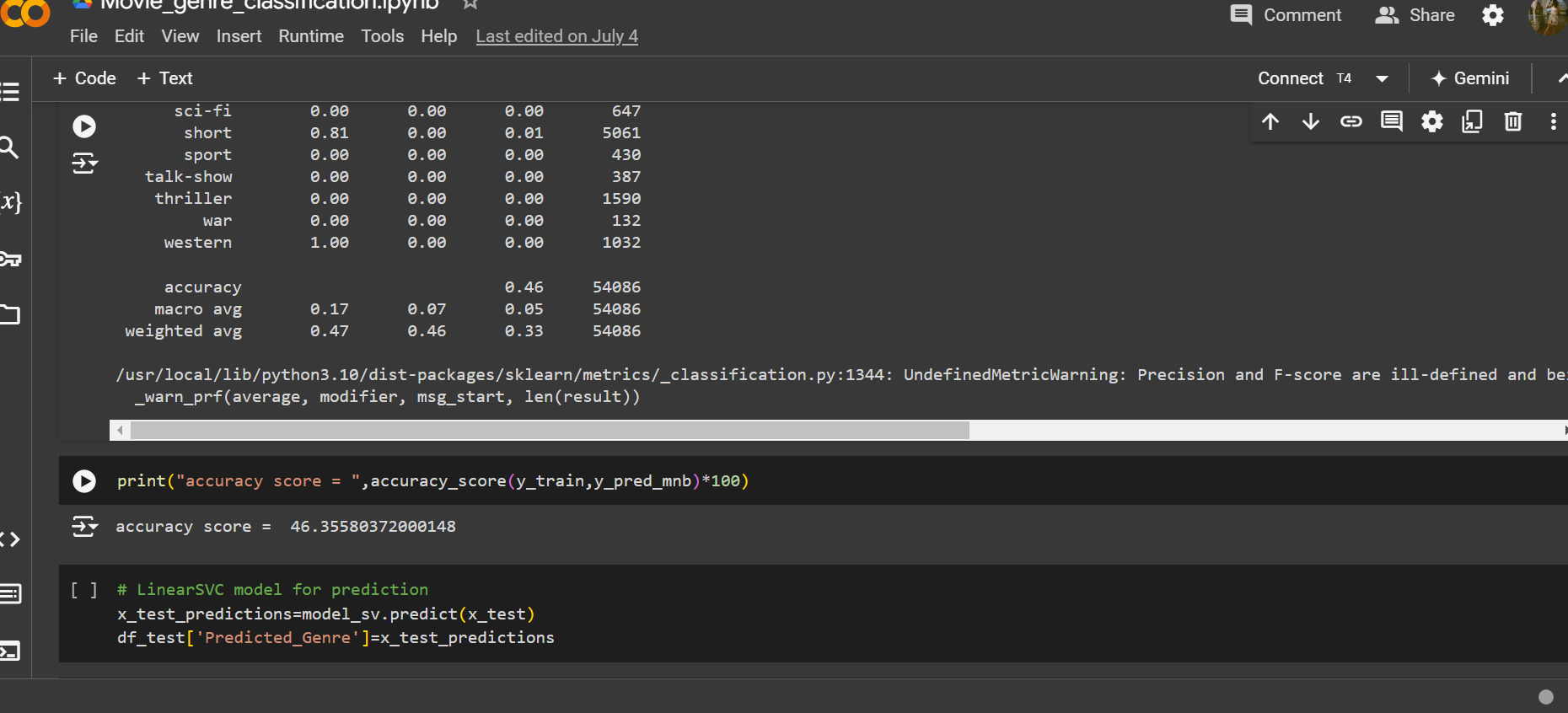




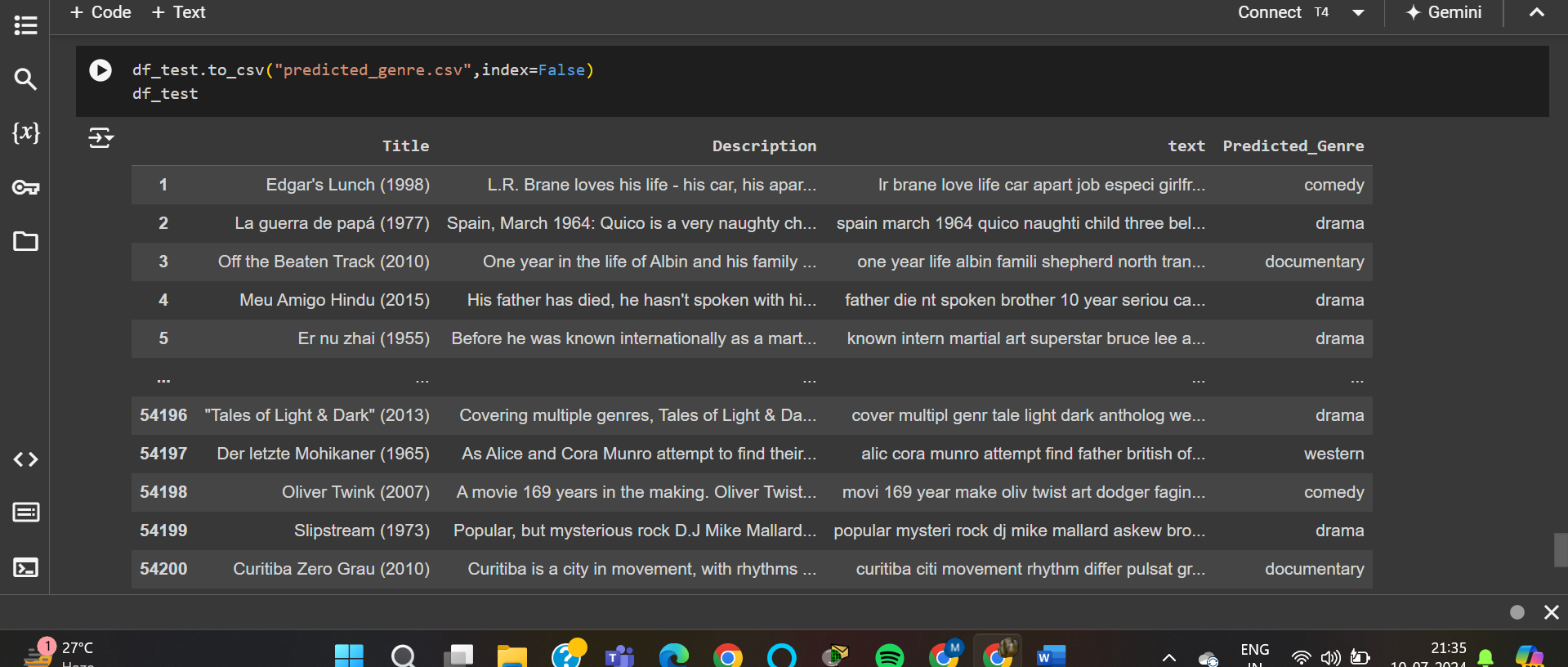


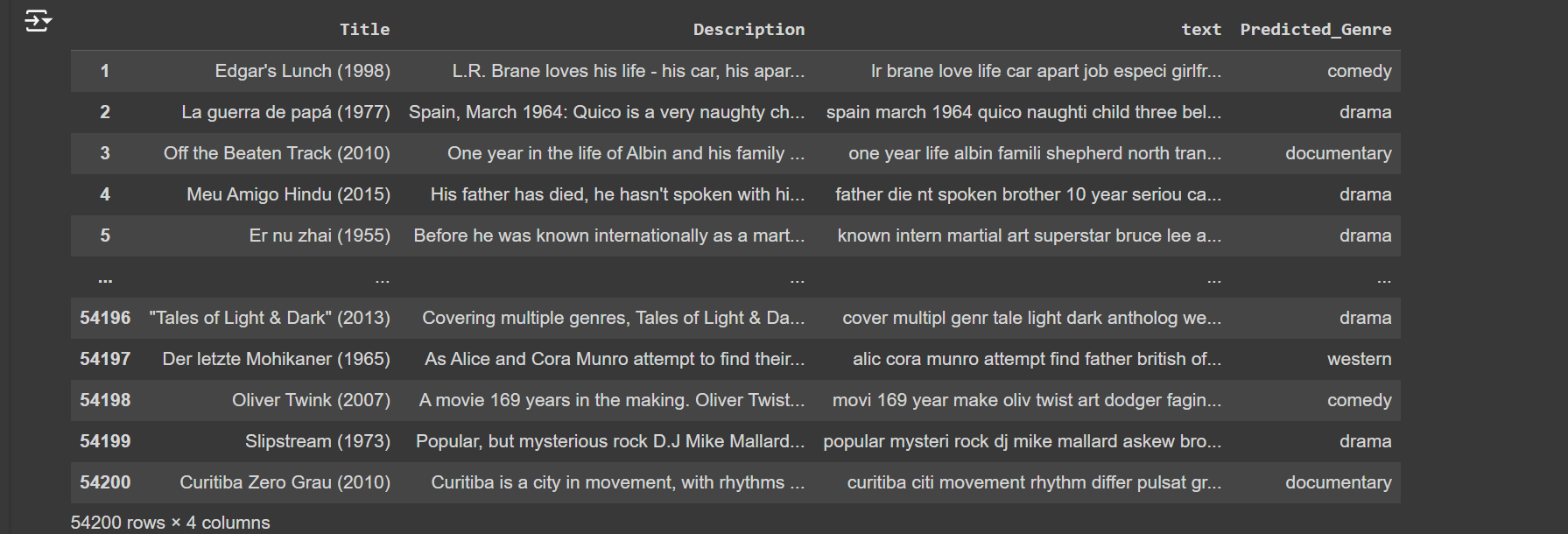
Here after achieving 94.60 accuracy on prediction of trained data now we import Naïve Baise model in which we us MultinomialNB to again test and predict data and test the accuracy.





Since accuracy was highest in Linear SVC for which it is considered so much, we test out model on same as you can see.





These are the predicted output using the linear SVC model.