Two Stage Job Title Identification System for Online Job Advertisements

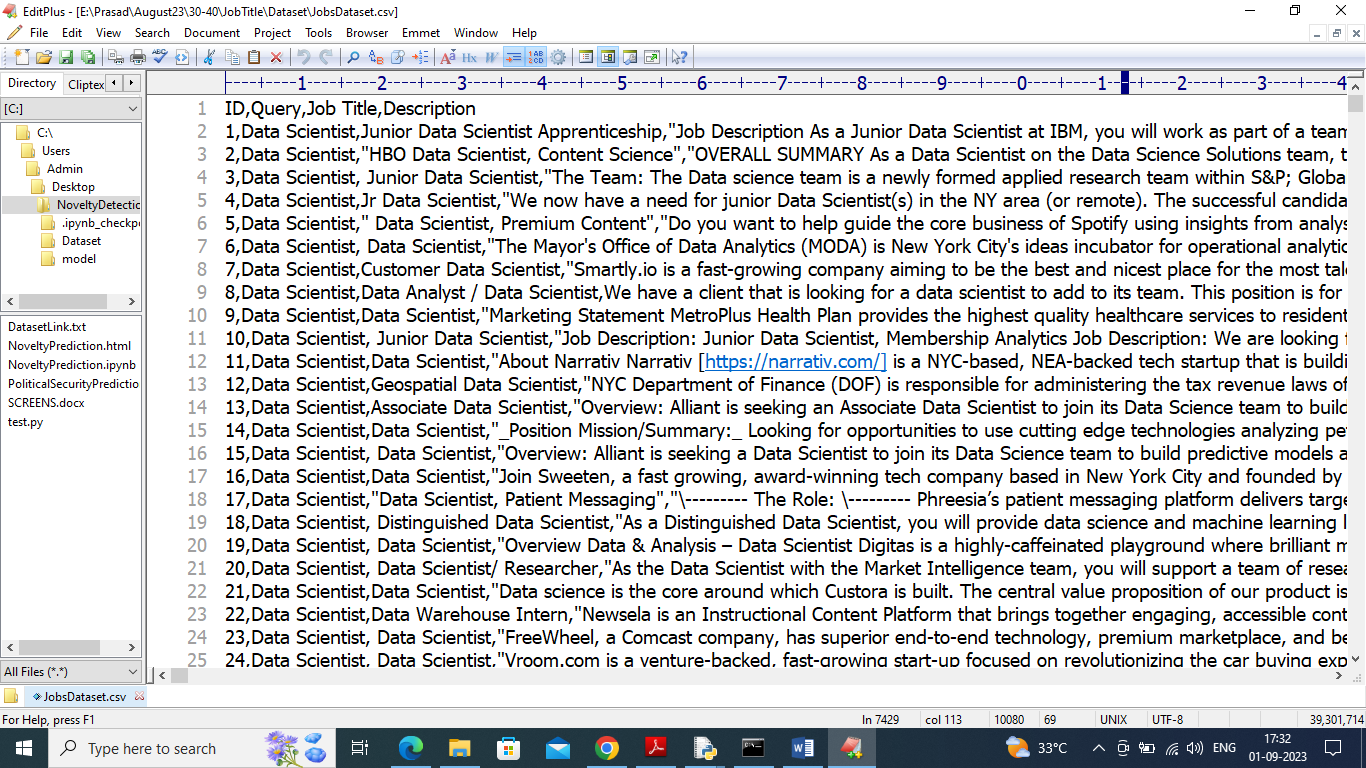
Data science algorithms often used to extract useful knowledge from unstructured text data such as Identifying Job Title by analysing Job Text Description. All existing algorithms are heavily dependent on large Label data for perfect classification and gathering huge label require lots of experience and time. All existing algorithms were using Occupational Information Network (O\*NET) data from US job market and this existing algorithm were not applying any additional technique to improve accuracy. So author of this paper employing two stage Job Title Identification and this stages include

1. In first Stage author using Bidirectional Encoder Representations from Transformers (BERT) to first classify the job ads according to their corresponding sector (e.g., Information Technology, Agriculture). BERT is used to convert unstructured text data into numeric vector by considering semantic similarity.
2. In stage two author employing Euclidean Distance algorithm to measure similarity between train and test data to find closest matching Job Title. This similarity measure will work with small amount of labels and doesn’t require huge labels of data.

Propose BERT model has compared with many other vector models like TFIDF, WORD2VEC etc. Compare to TFIDF and WORD2VEC propose BERT with Euclidean distance is giving high accuracy.

All vector models data will be input to CHI2 features selection algorithm to find only those features with highest weight values. TFIDF and WORD2VEC vector features trained with SVM, Naïve Bayes and Logistic Regression and propose BERT is trained with Euclidean Distance similarity measure. Each algorithm performance is evaluated in terms of accuracy, precision, recall and FSCORE.

To train all algorithms author has generated his own dataset but not publish on internet so we have used Job Title Description dataset from KAGGLE and below are the dataset details



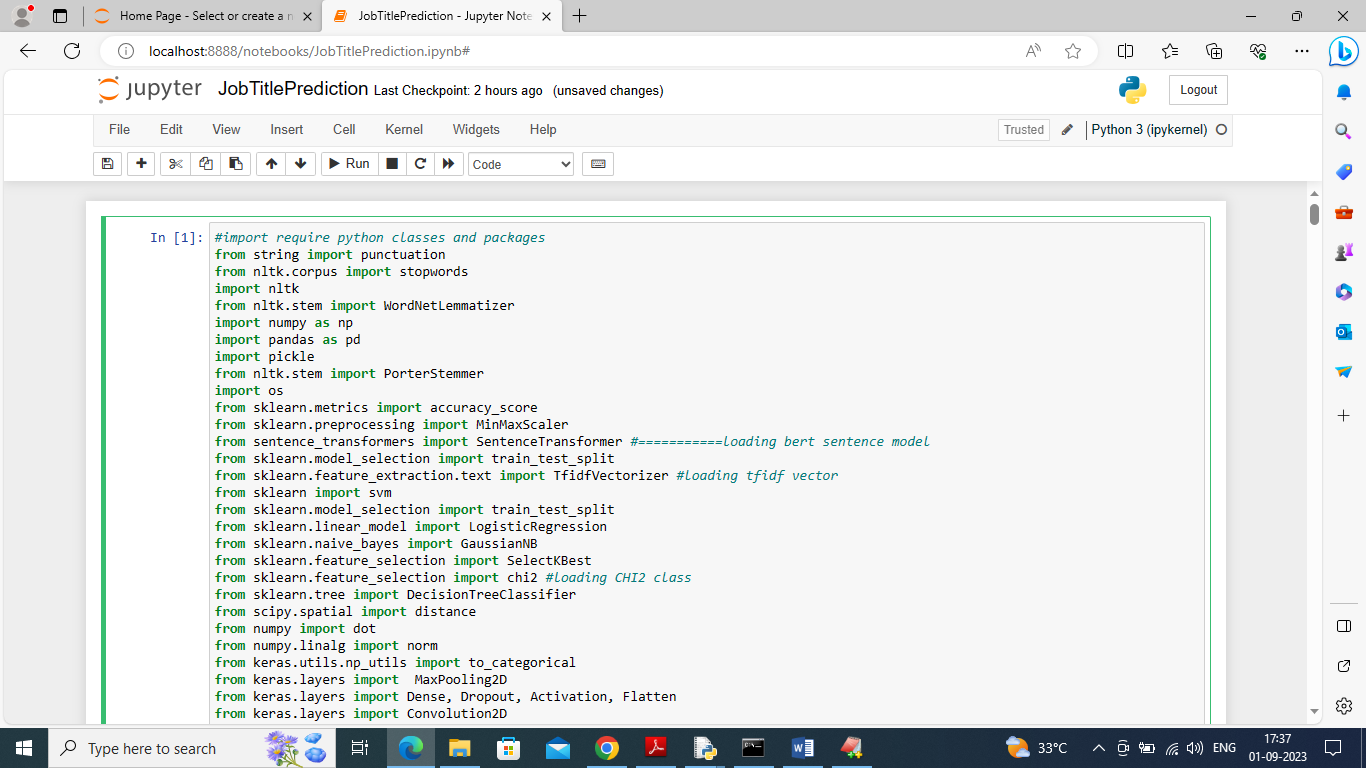
In above dataset screen first row contains dataset column names and remaining rows contains dataset values and in dataset we can see Job Title, Name and Description and by using above dataset we will train and test all algorithm performance. In propose paper we have applied various TEXT processing techniques such as removing stop words, special symbol, applying stemming, Lemmatization etc.

Extension Concept

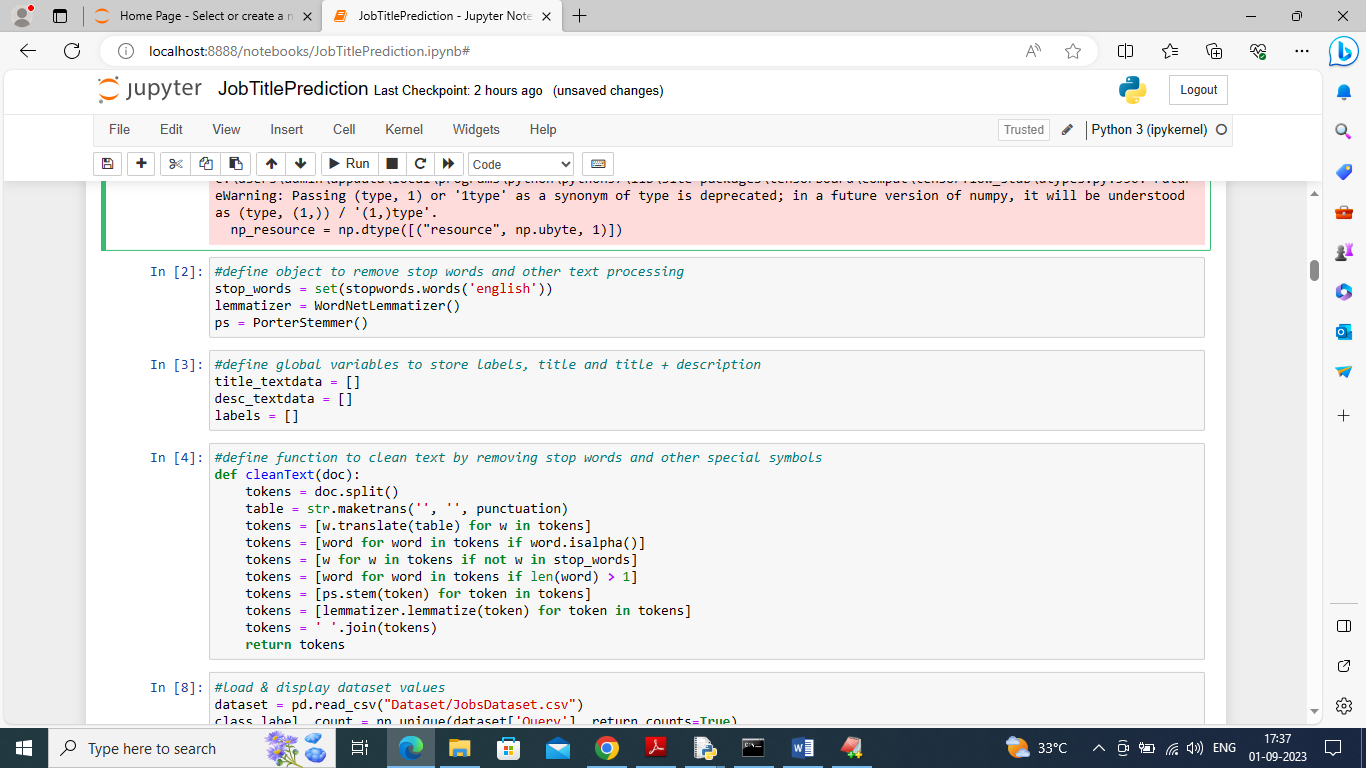
In propose paper author has used all traditional machine learning algorithms such as SVM, Naïve Bayes and Logistic Regression and not experimented with any other advanced algorithms like CNN2D, BI-LSTM etc. So as extension we have experimented with CNN2D algorithm as this algorithm filtered features at multiple neurons iterations to train model with best features and this best features help CNN in getting high accuracy.

SCREEN SHOTS

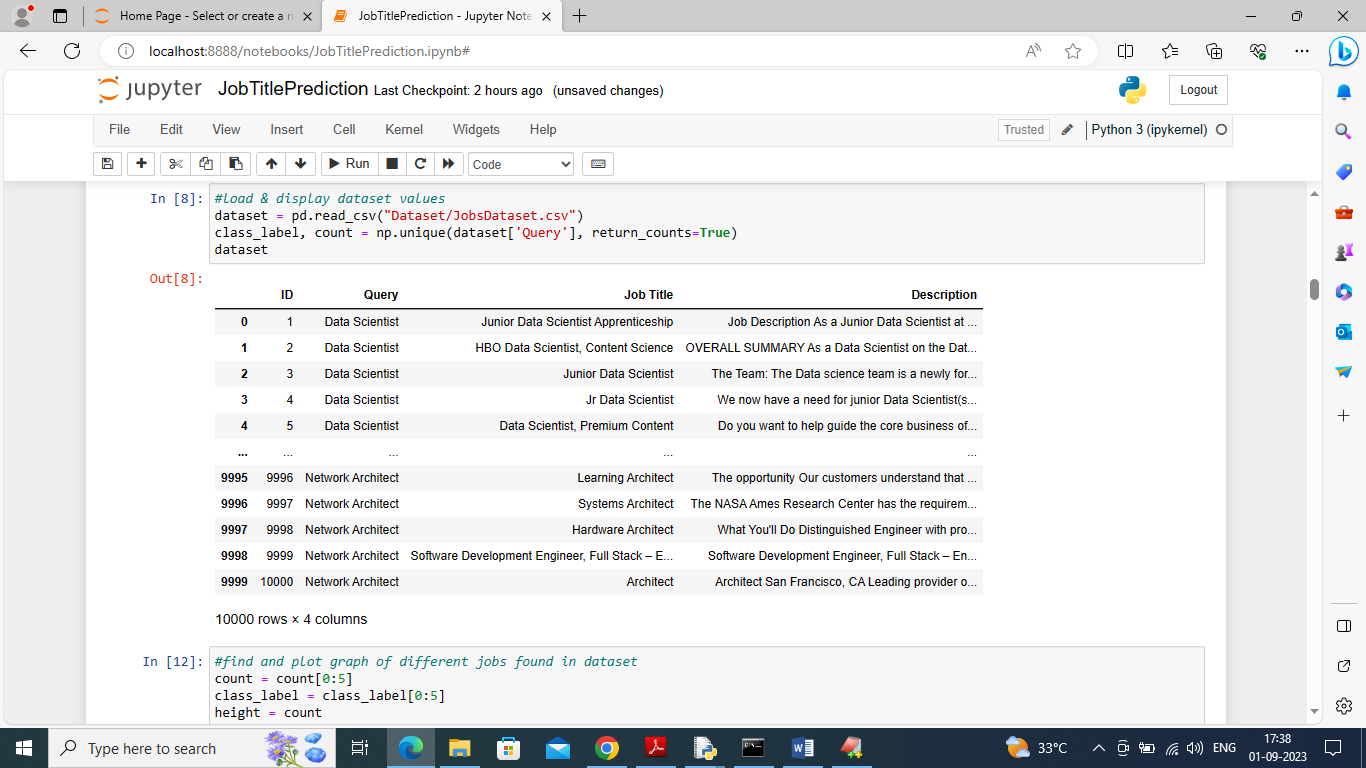
We have coded this project using JUPYTER notebook and below are the code and output screens with blue colour comments



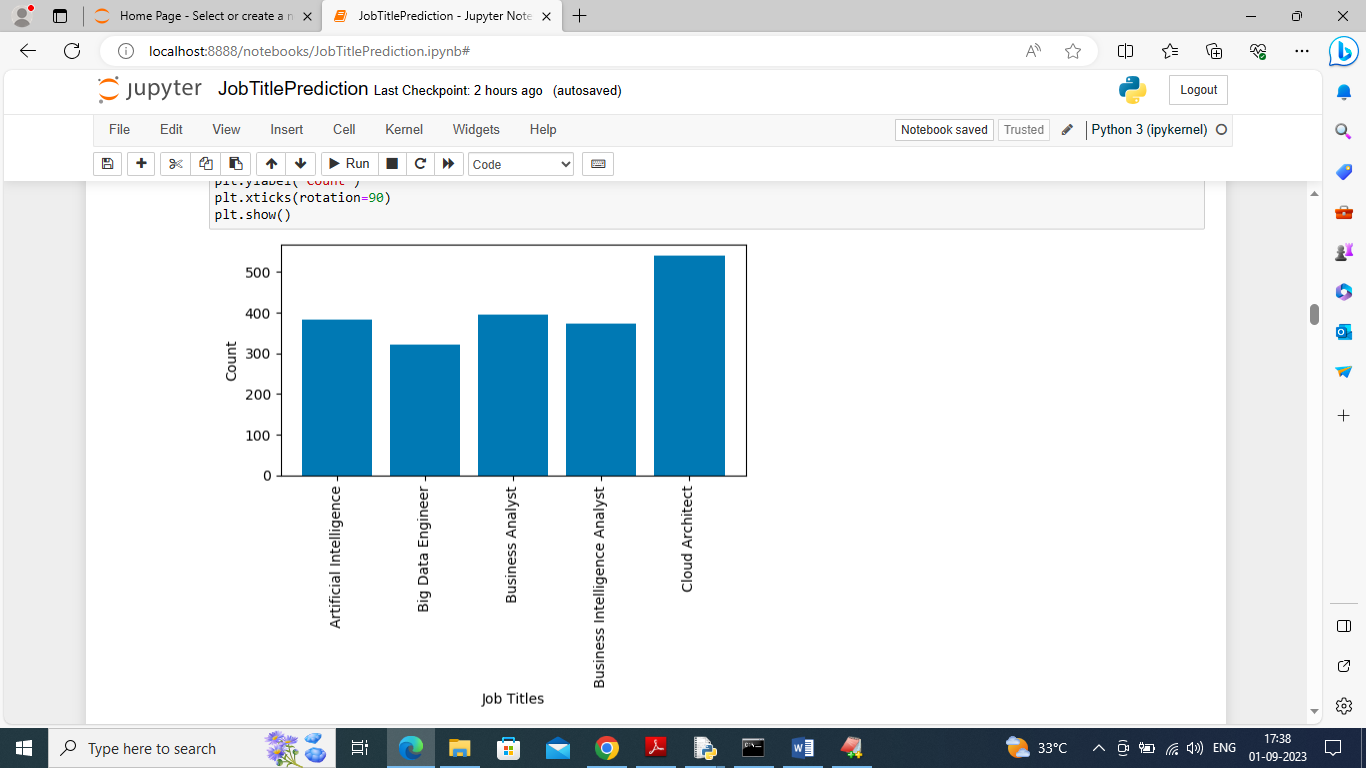
In above screen importing require python classes and packages



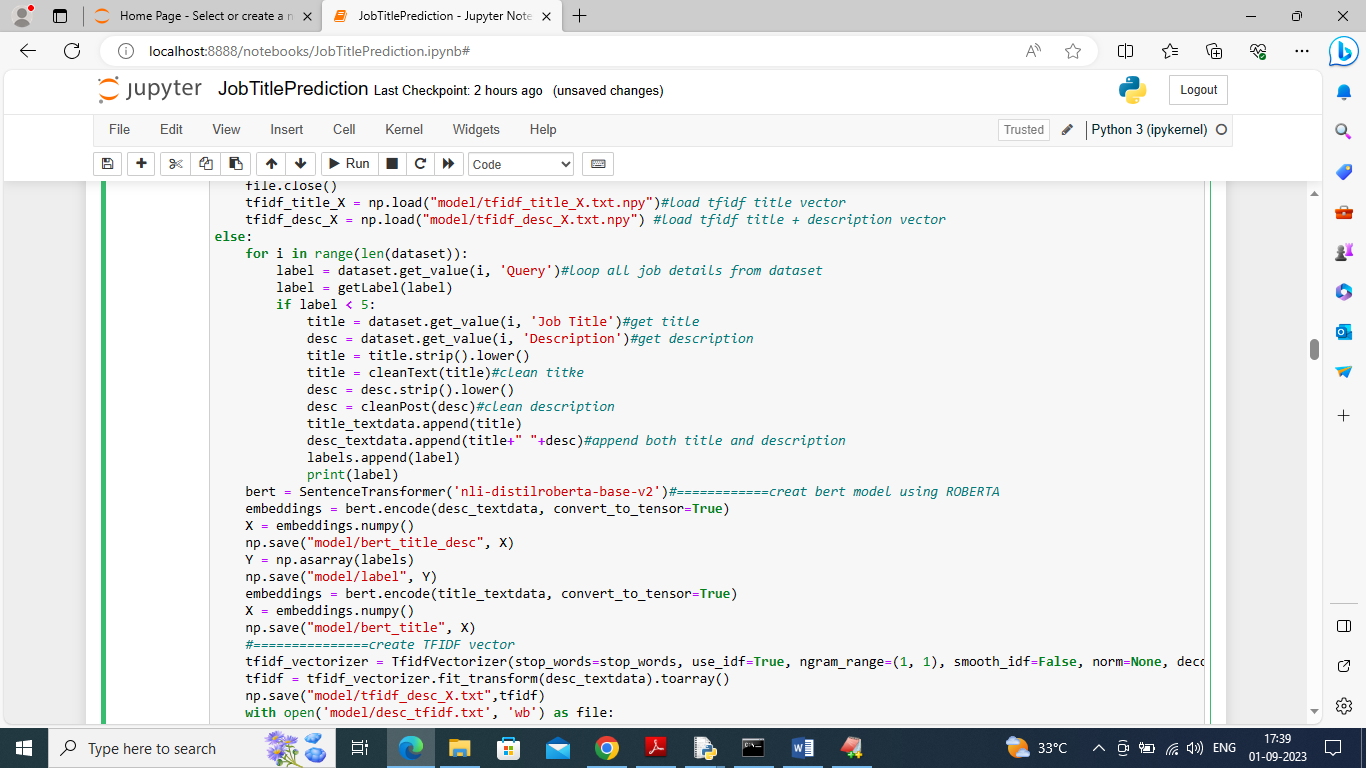
In above screen defining code to remove stop words, special symbols etc.



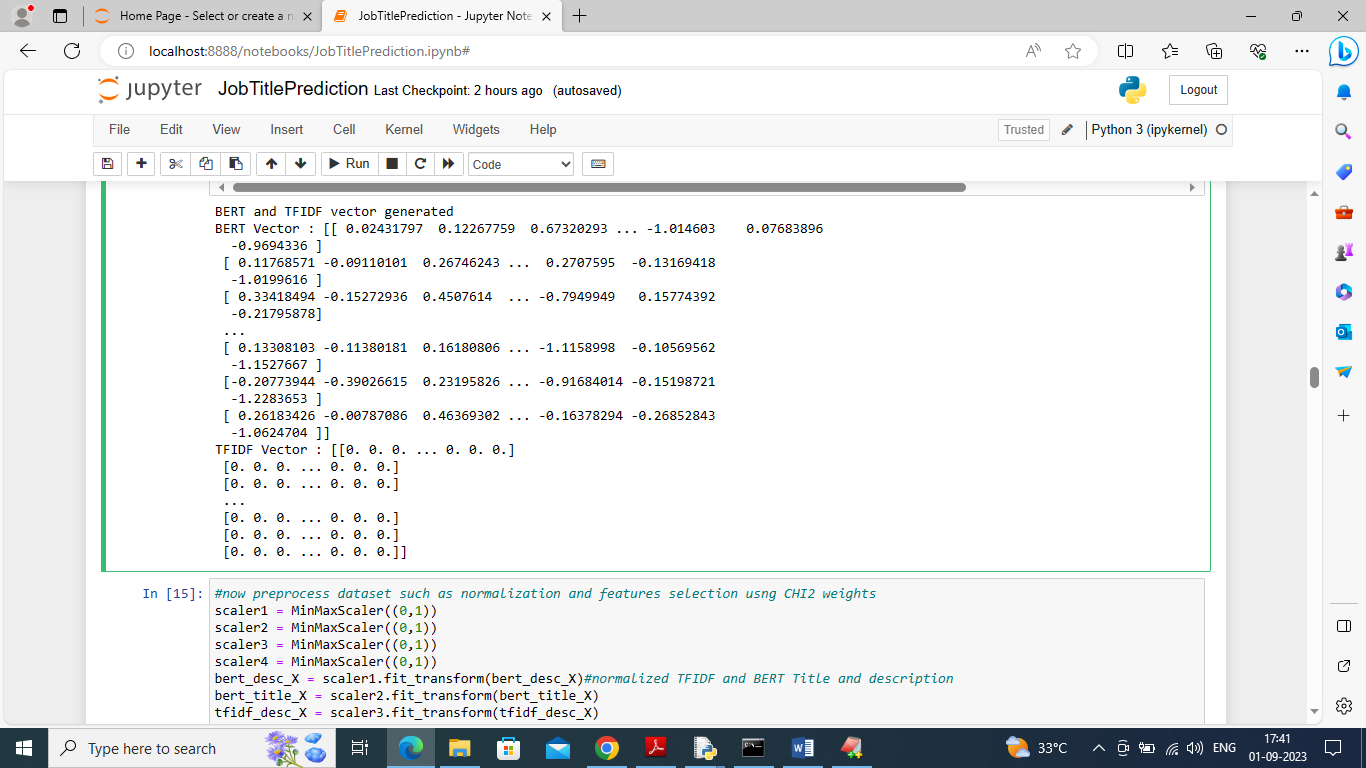
In above screen reading and displaying dataset values



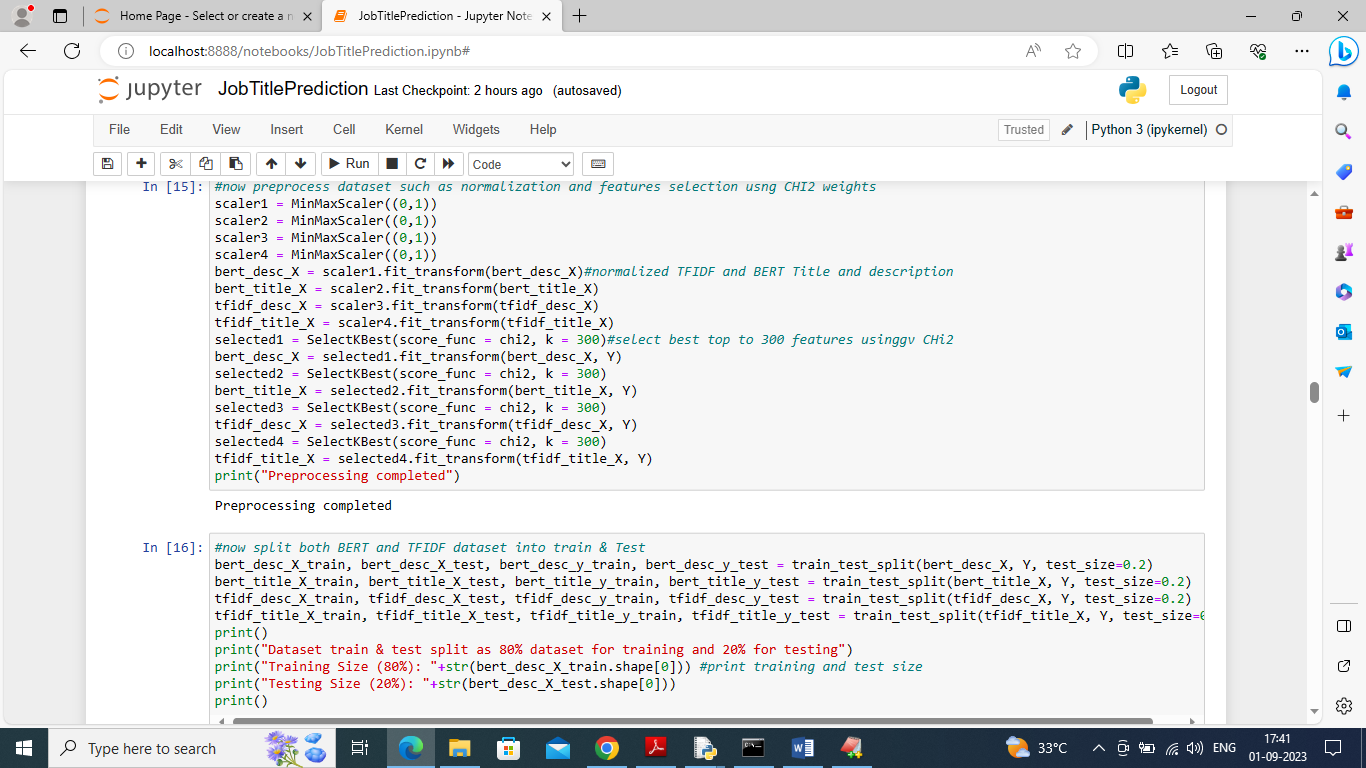
In above screen finding and plotting graph of various JOBS found in dataset where x-axis represents JOB TITLE and y-axis represents counts



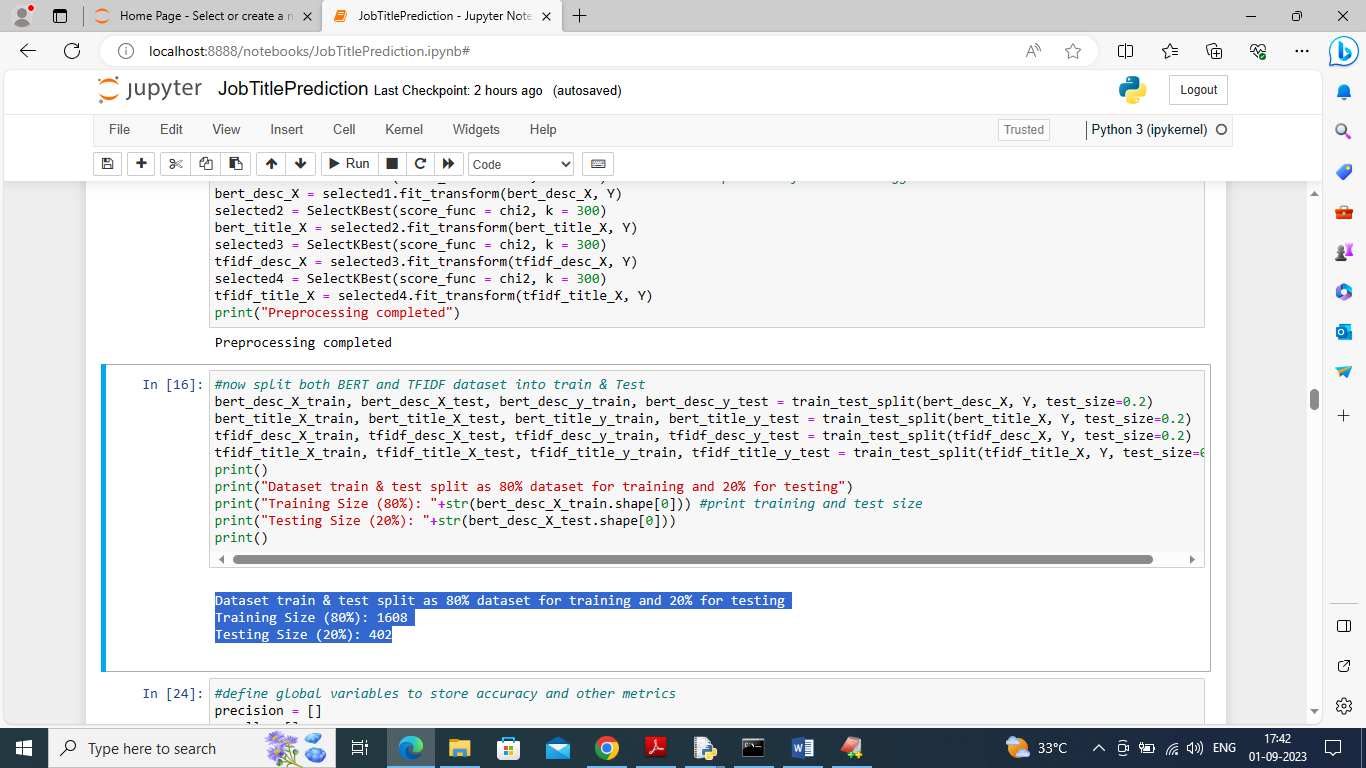
In above screen we are reading each JOB description and then cleaning and adding to array variable and then creating BERT and TFIDF object to convert all JOB description into numeric vector. In above screen =========== before dashed lines you can see we are creating BERT and TFIDF vectors and after executing above block will get below vector



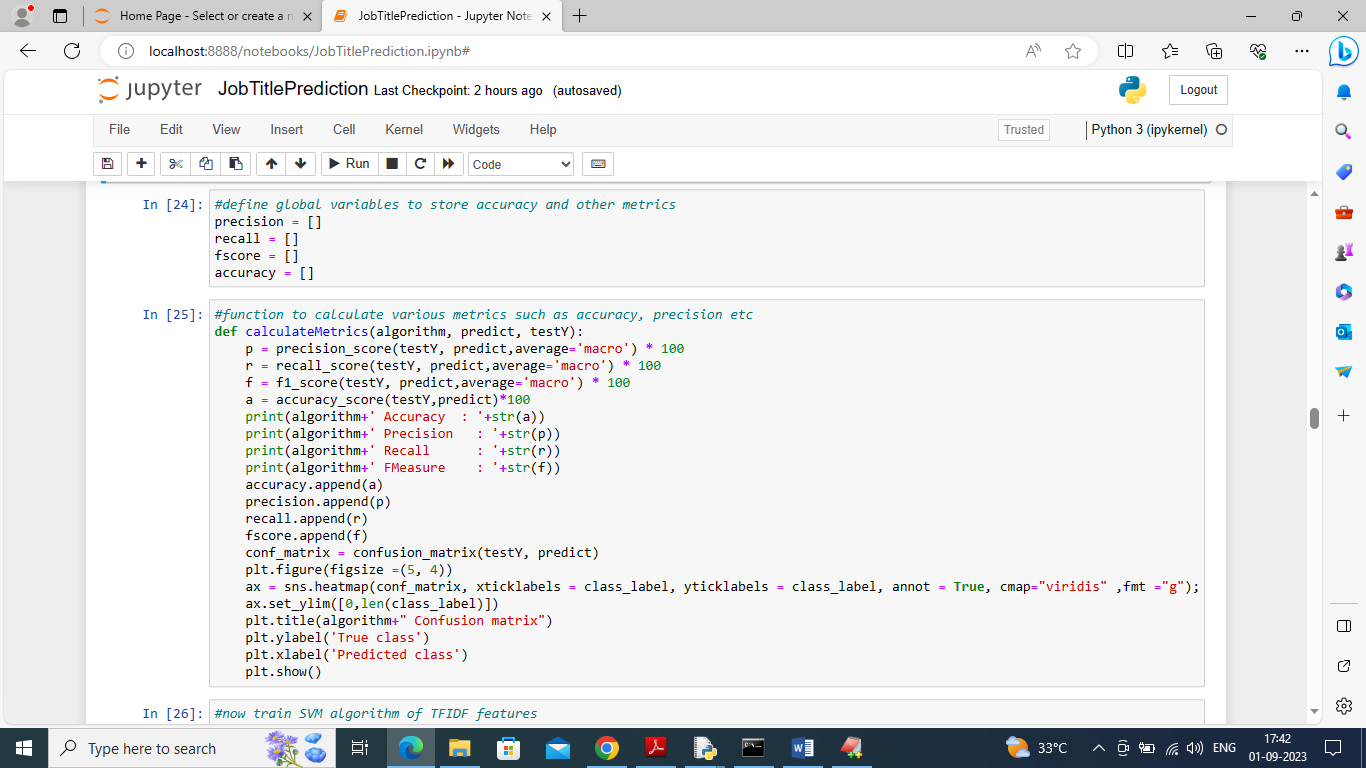
In above screen BERT and TFIDF vector created



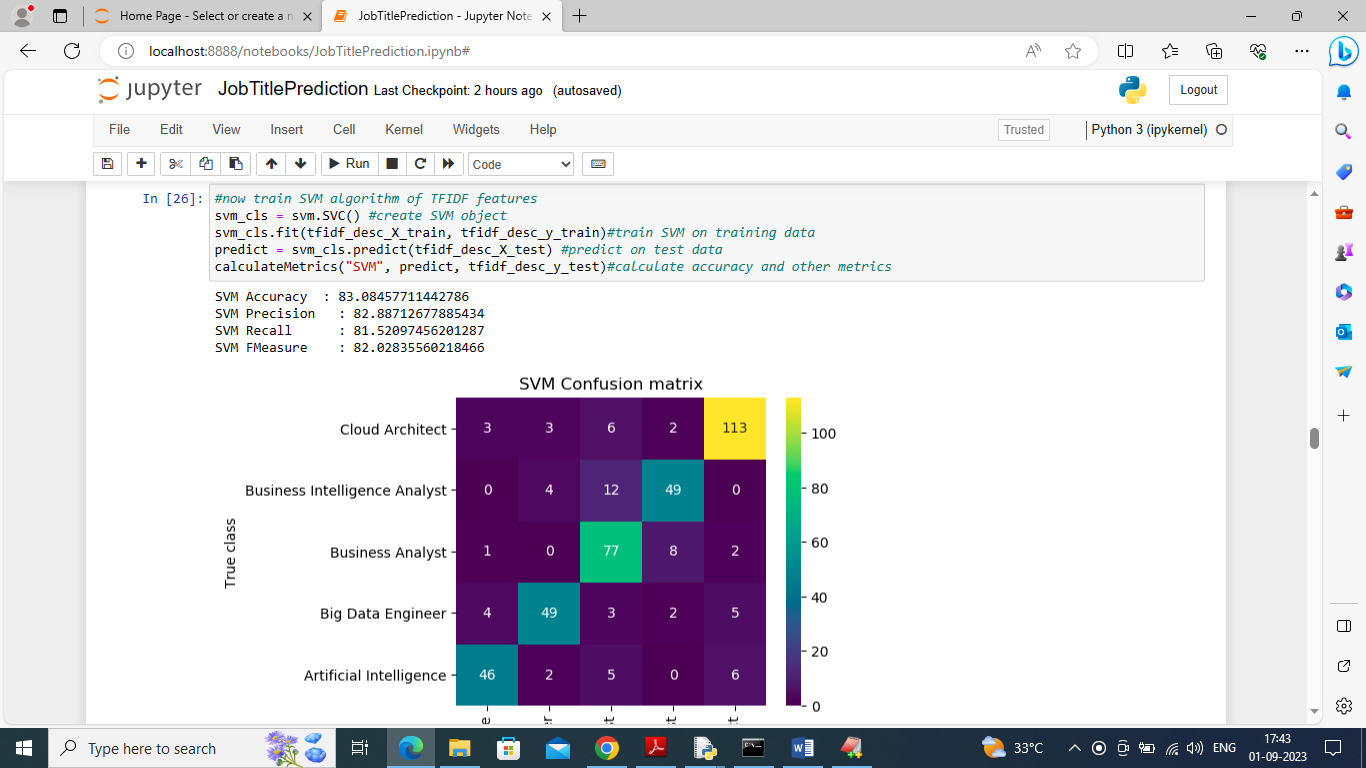
In above screen normalizing and applying CHI2 algorithm on both BERT and TFIDF vector



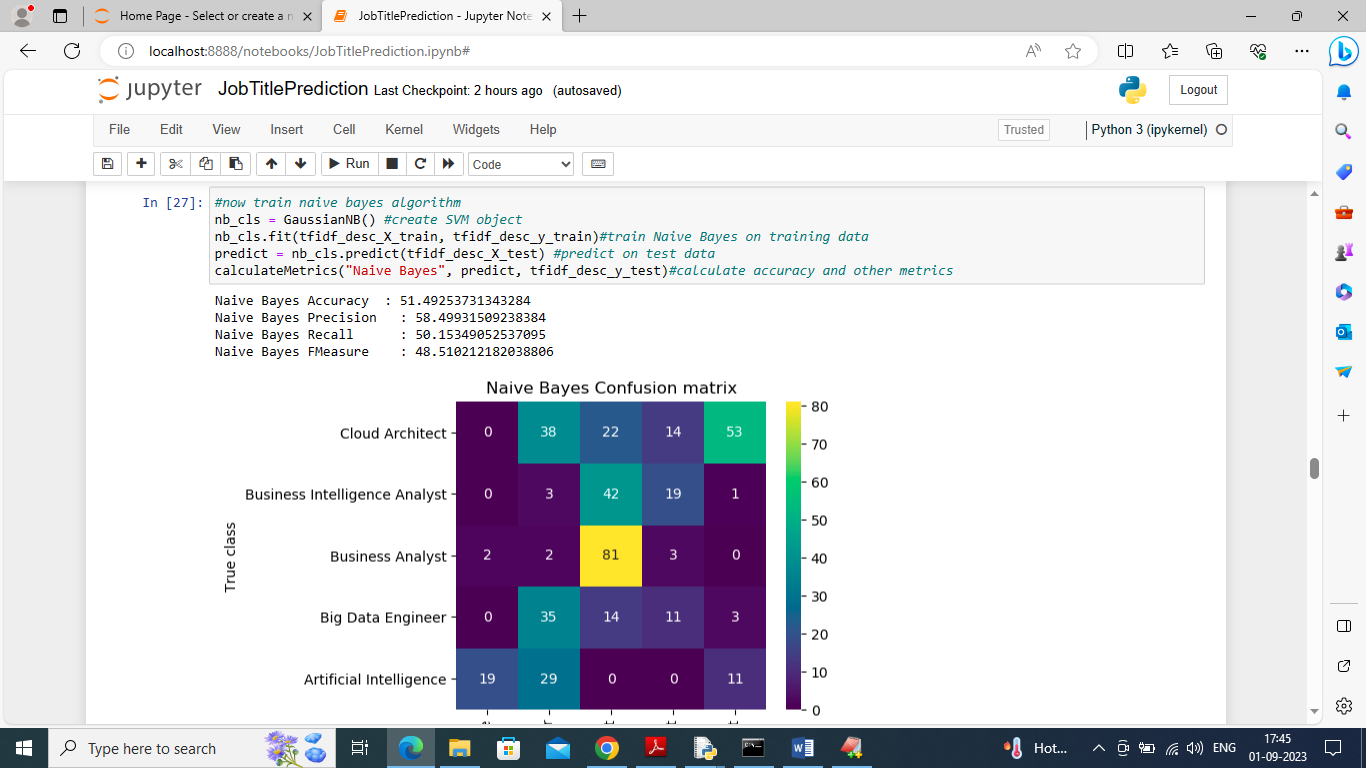
In above screen splitting data into train and test



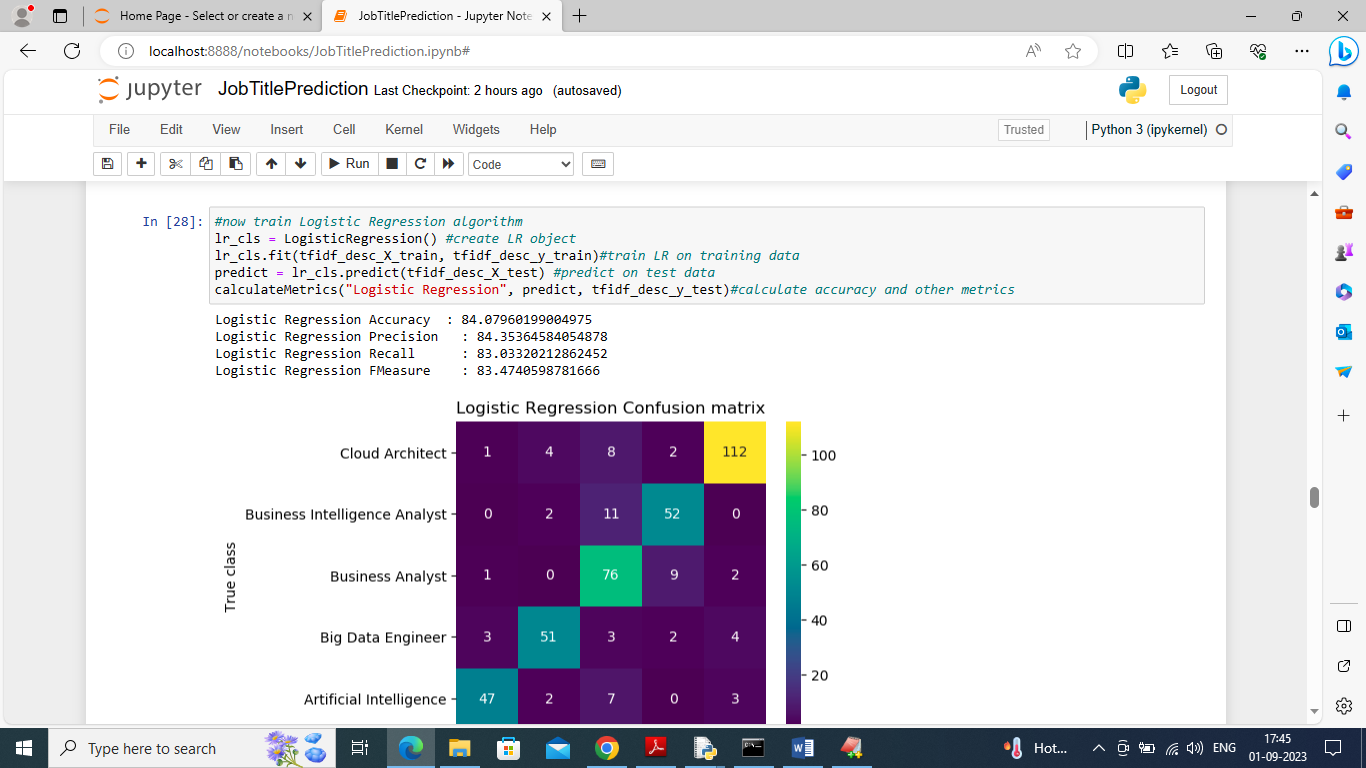
In above screen defining function to calculate accuracy and other metrics



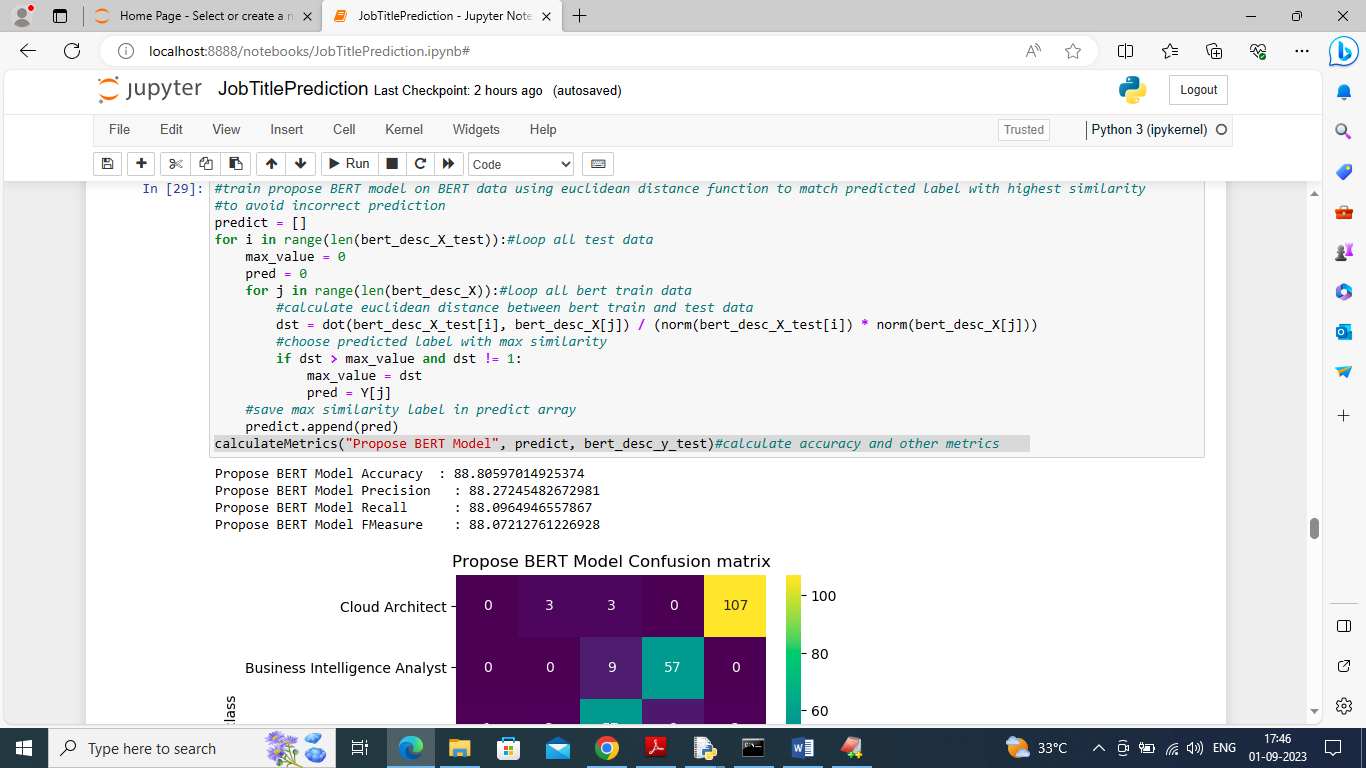
In above screen training SVM on TFIDF features and it got 83% accuracy and can see other metrics also and in confusion matrix graph x-axis represents True Job Title and Y-axis represents Predicted Job Title and all different colour boxes in diagnol represents correct prediction count and remaining blue boxes contains incorrect prediction count



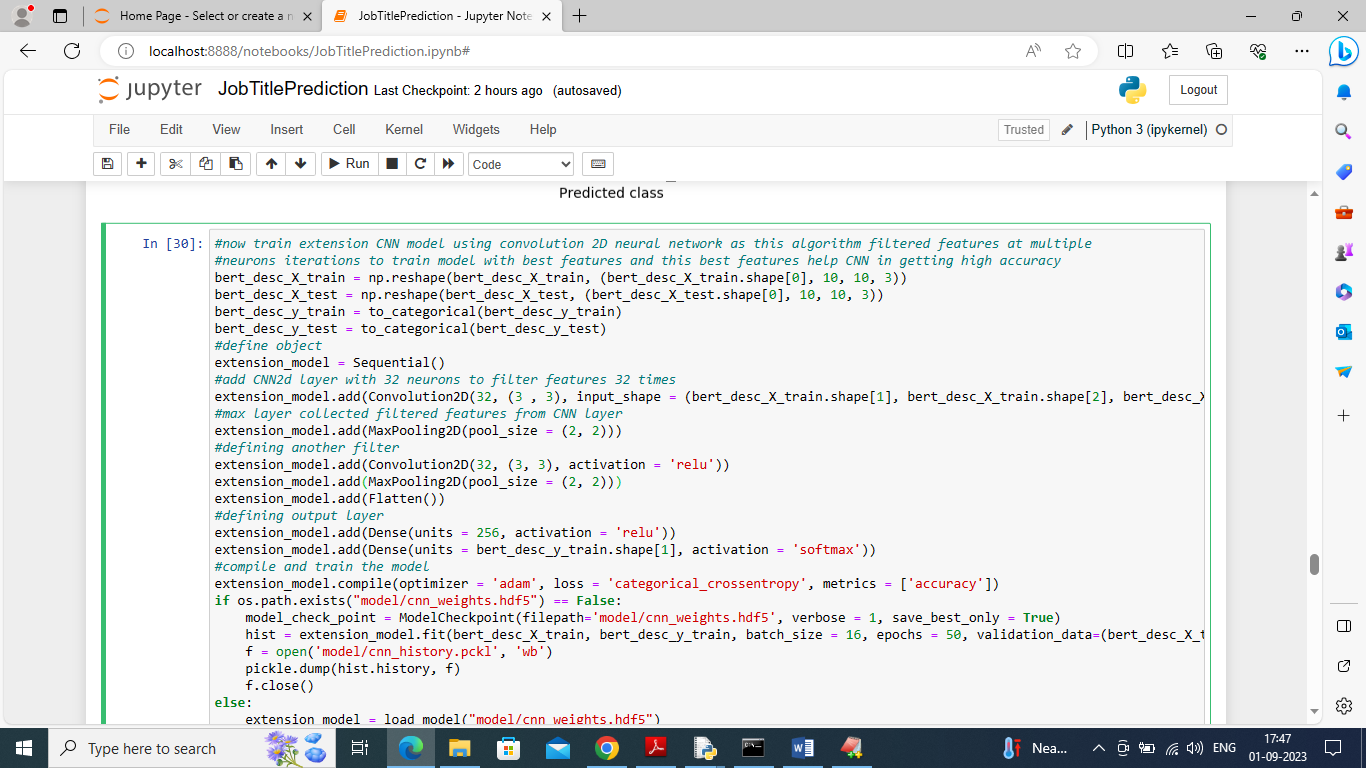
In above screen Naïve Bayes got 51% accuracy



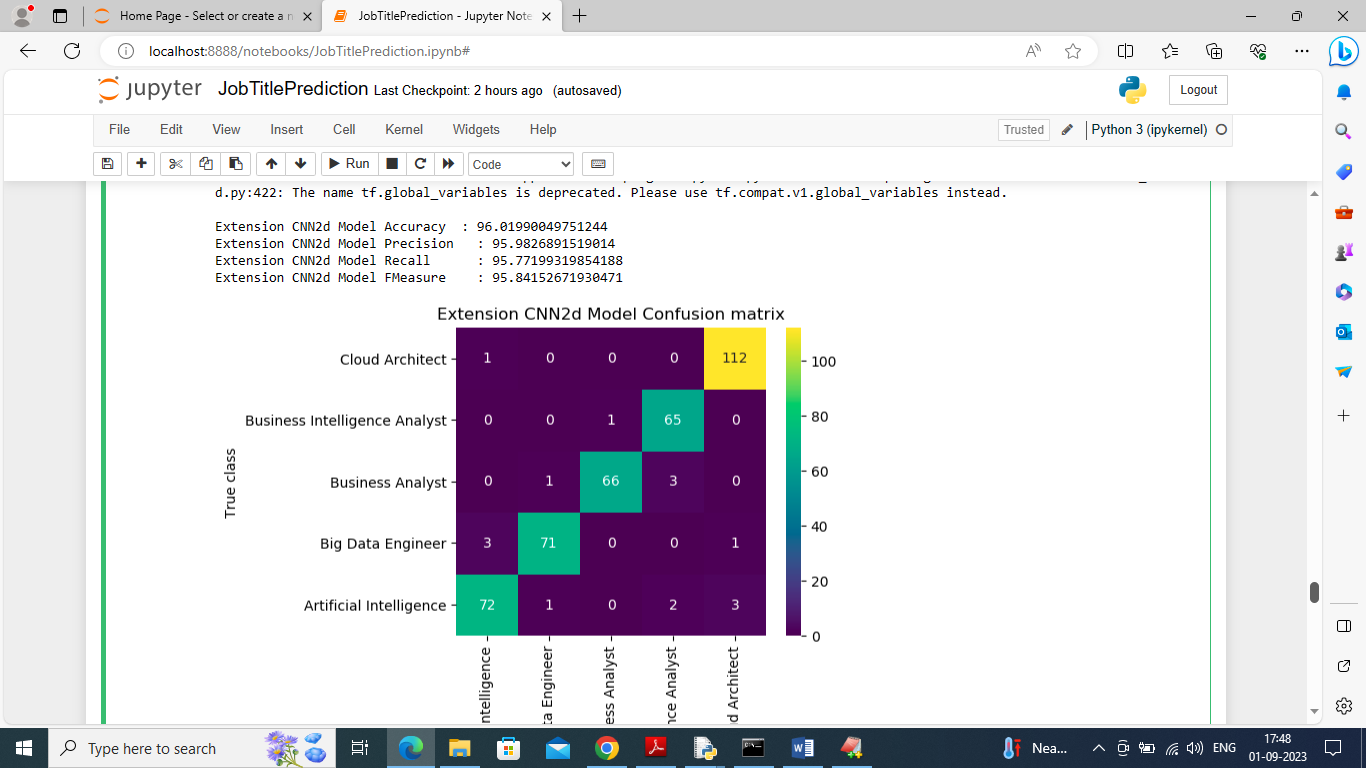
In above screen Logistic Regression got 84% accuracy



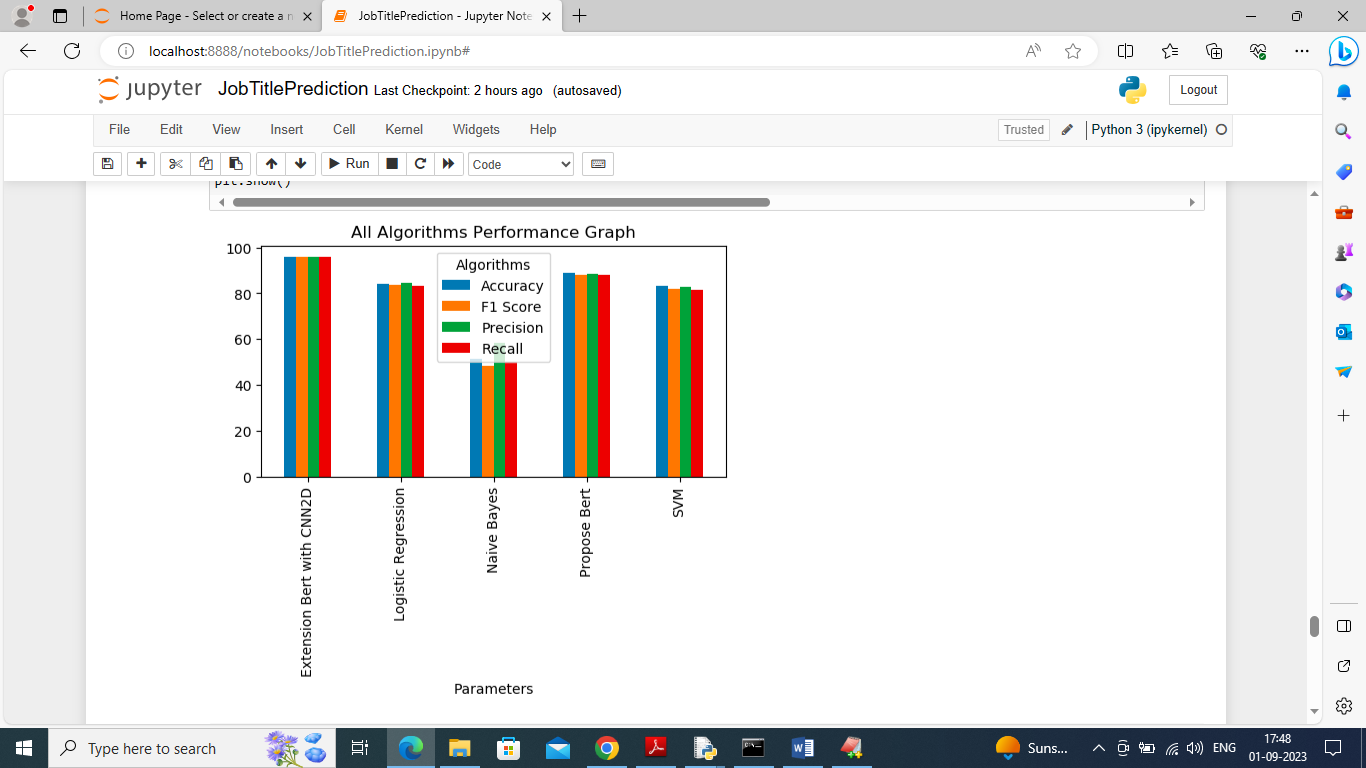
In above screen propose BERT model with max similarity measure got 88% accuracy which is higher than existing algorithms and can see other metrics also



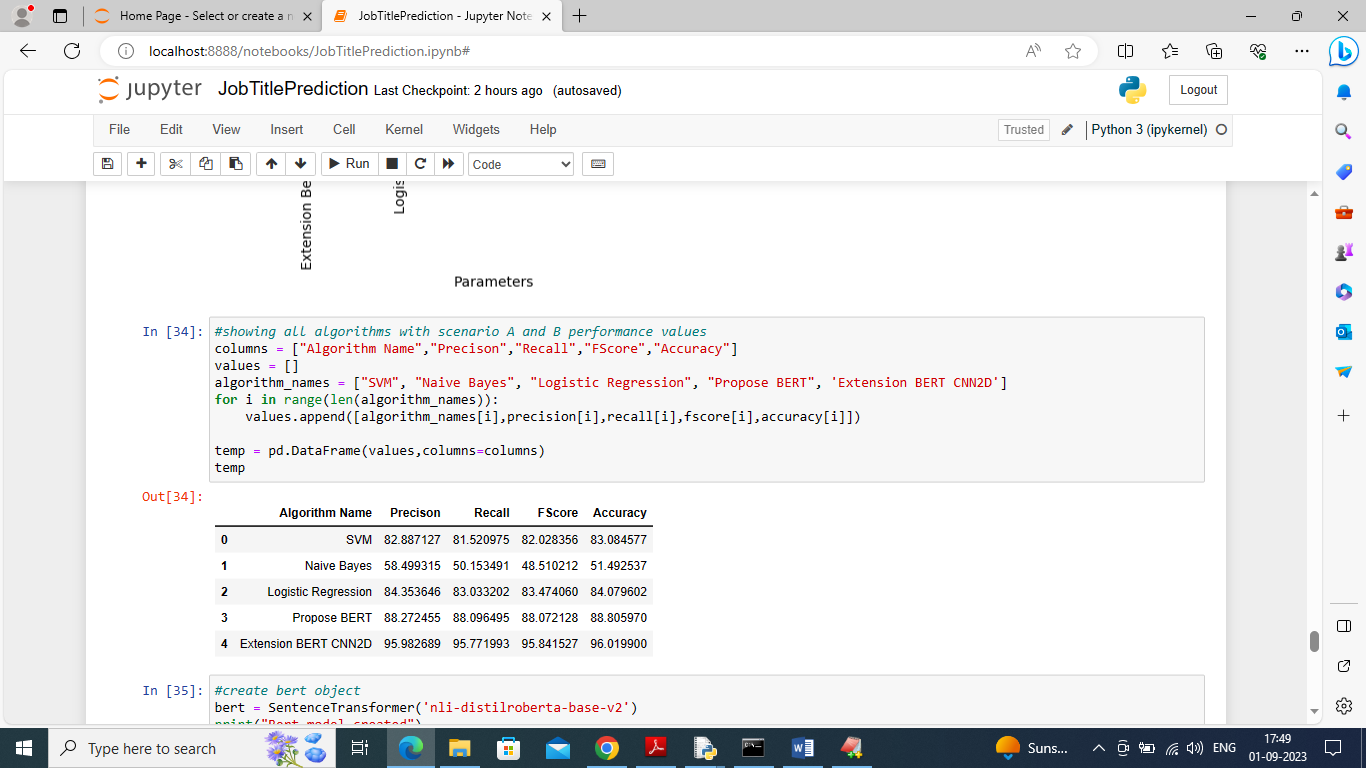
In above screen train extension CNN2D algorithm and after executing above block will get below output



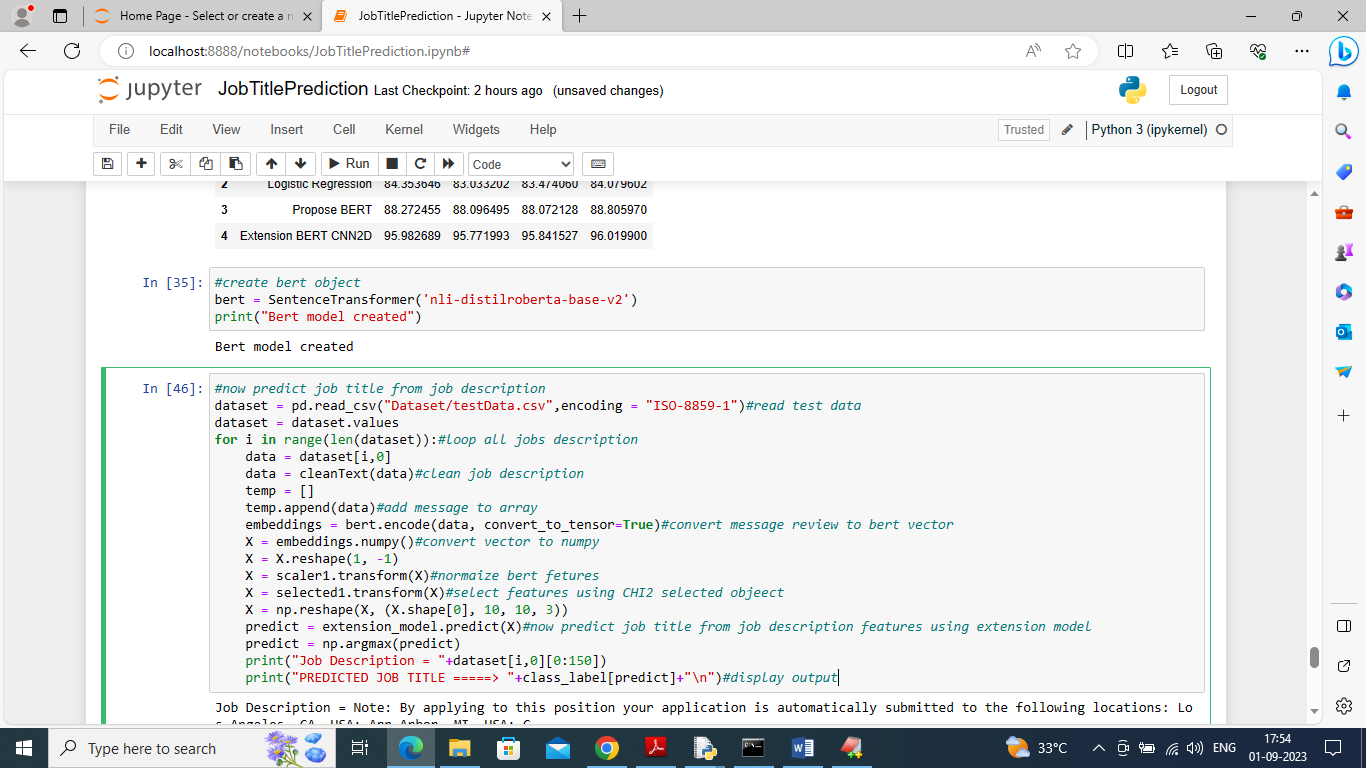
In above screen extension CNN2D model got 96% accuracy and can see other metrics also



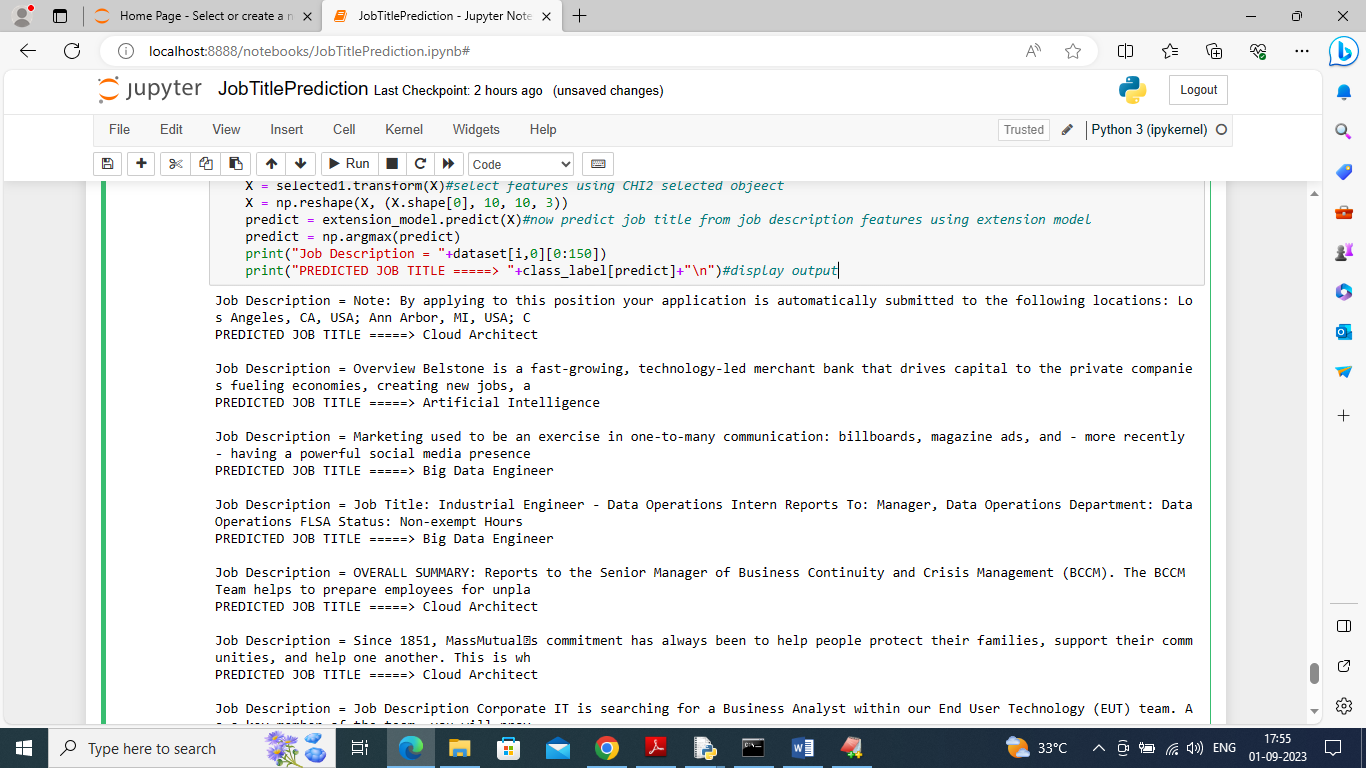
In above screen displaying all algorithm performance where x-axis represents algorithm names and y-axis represents accuracy and other metrics in differnet colour bars



In above screen displaying all algorithms performance in tabular format



In above screen reading JOB description from TEST data and then predicting JOB TITLE and below is the output



In above screen in first line we can see JOB Description and then after ==🡺 arrow symbol can see predicted JOB title as Big data Engineer or Cloud Architect