**Spam detection**

**Software requirements**

**• Programming Language : Python, Colab**

**• Packages : Numpy,Matplotlib, SKLearn, Pandas, Flask Framework**

**• Tool : Python 3.7**

**Dataset description:**

Dataset The data set is collected from UCI Machine Learning Repository. It is a public set of comments collected for spam research. It has five datasets composed by 1,957 real messages extracted from five videos that were among the 10 most viewed on the collection period.

**COMMENT\_ID , AUTHOR, DATE, CONTENT, CLASS**

LZQPQhLyRh80UYxNuaDWhIGQYNQ96IuCg-AYWqNPjpU, Julius NM, 2013-11-07T06:20:48,

Huh, anyway check out this you[tube] channel: kobyoshi02 , 1

z122wfnzgt30fhubn04cdn3xfx2mxzngsl40k, Bob Kanowski, 2013-11-28T12:33:27, i turned it on mute as soon is i came on i just wanted to check the views...ï»¿, 0

z13ttt1jcraqexk2o234ghbgzxymz1zzi04, Cony, 2013-11-28T16:01:47, You should check my channel for Funny VIDEOS!!ï»¿, 1

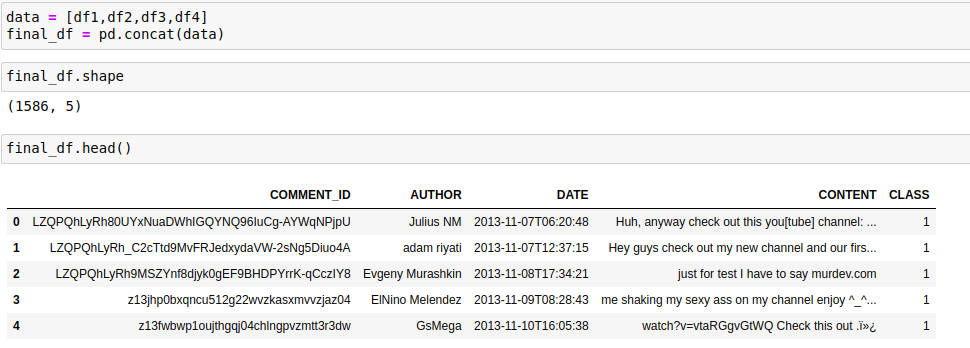
z122wfnzgt30fhubn04cdn3xfx2mxzngsl40k, Bob Kanowski, 2013-11-28T12:33:27, i turned it on mute as soon is i came on i just wanted to check the views...ï»¿, 0

**Above all bold names are the dataset column names and all values are the dataset values and last columns contains either value 0 or 1 where 0 means that comment contains not spam and 1 means spam.**

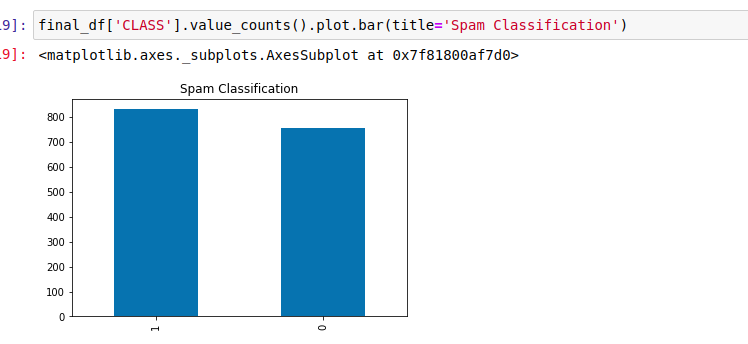
**Import required package and extracting the dataset**

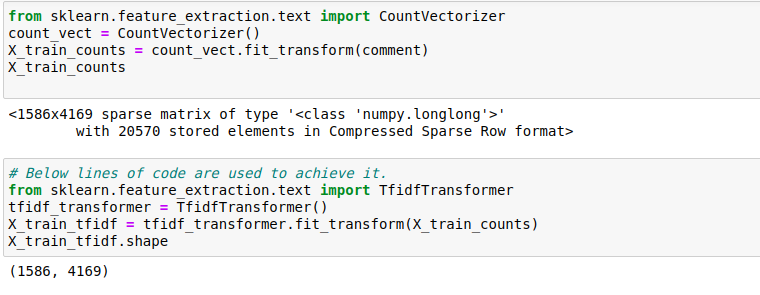
****

**Data Processing**

****

**Visualization of outcome**

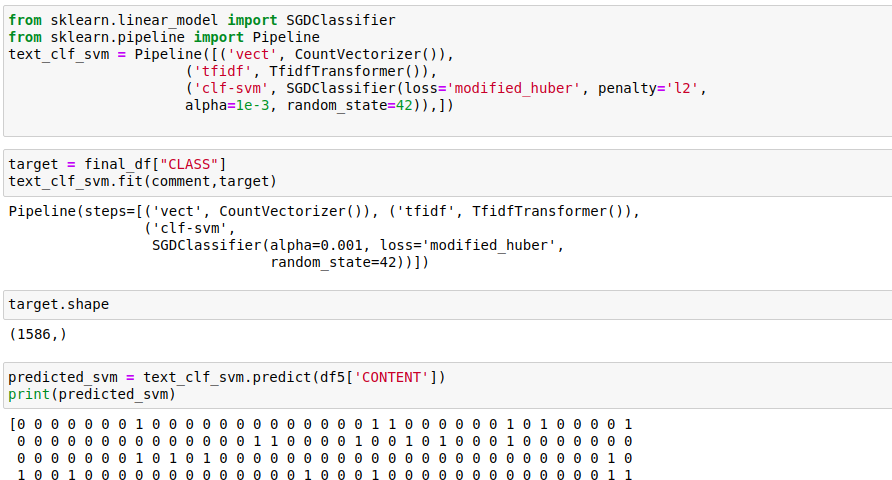
****

****

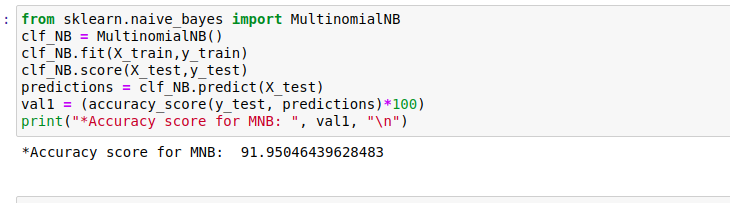
**Feature extraction with count vectorizer & TFid**

**ML deploying**

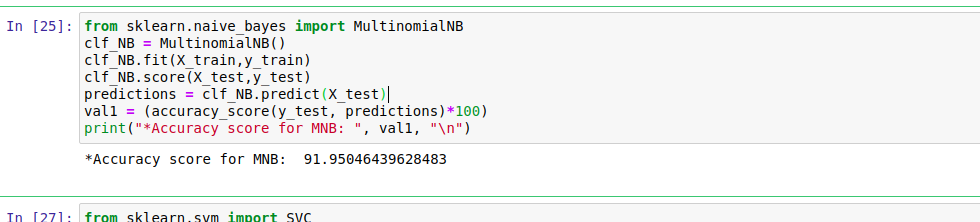
**SGD**

****

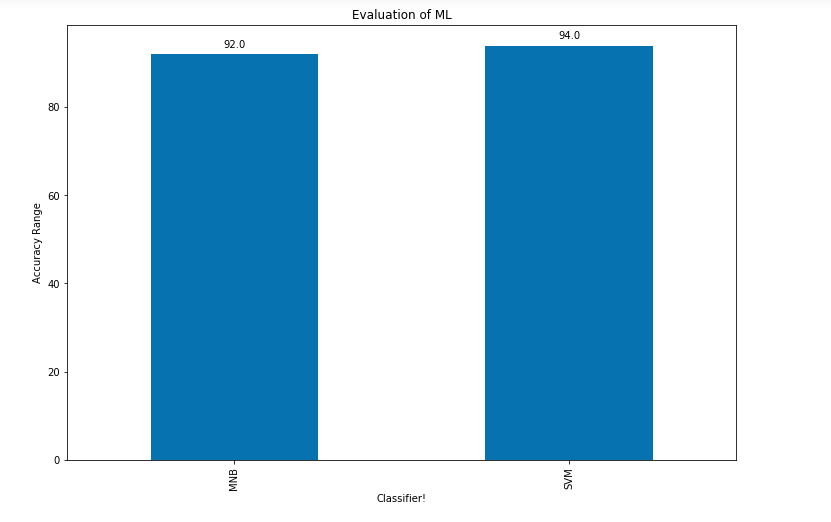
**MNB**

****

**SVM**

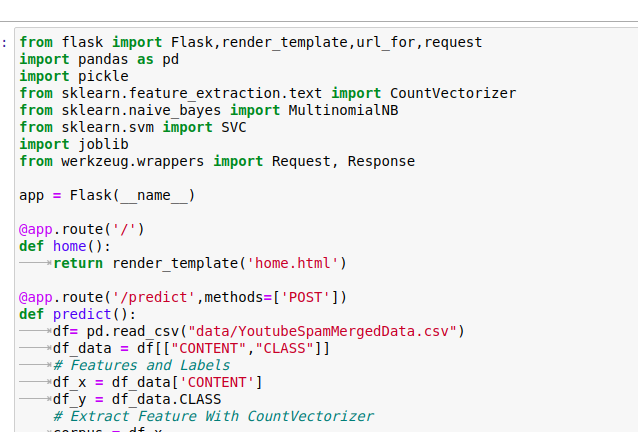
****

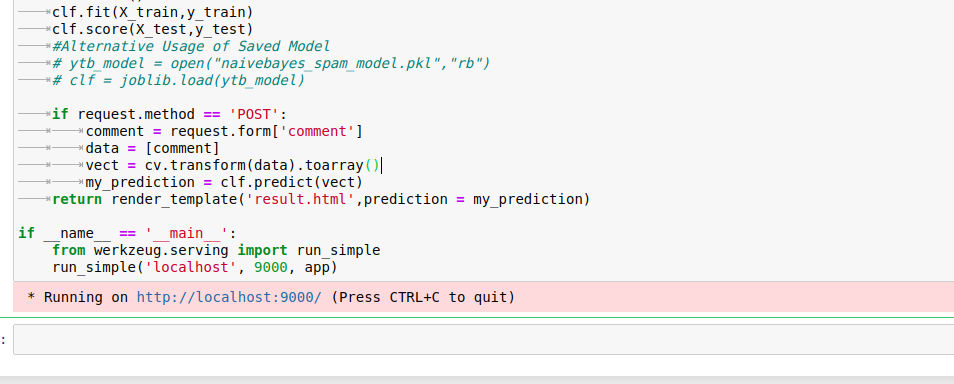
**Model Comparison**

****

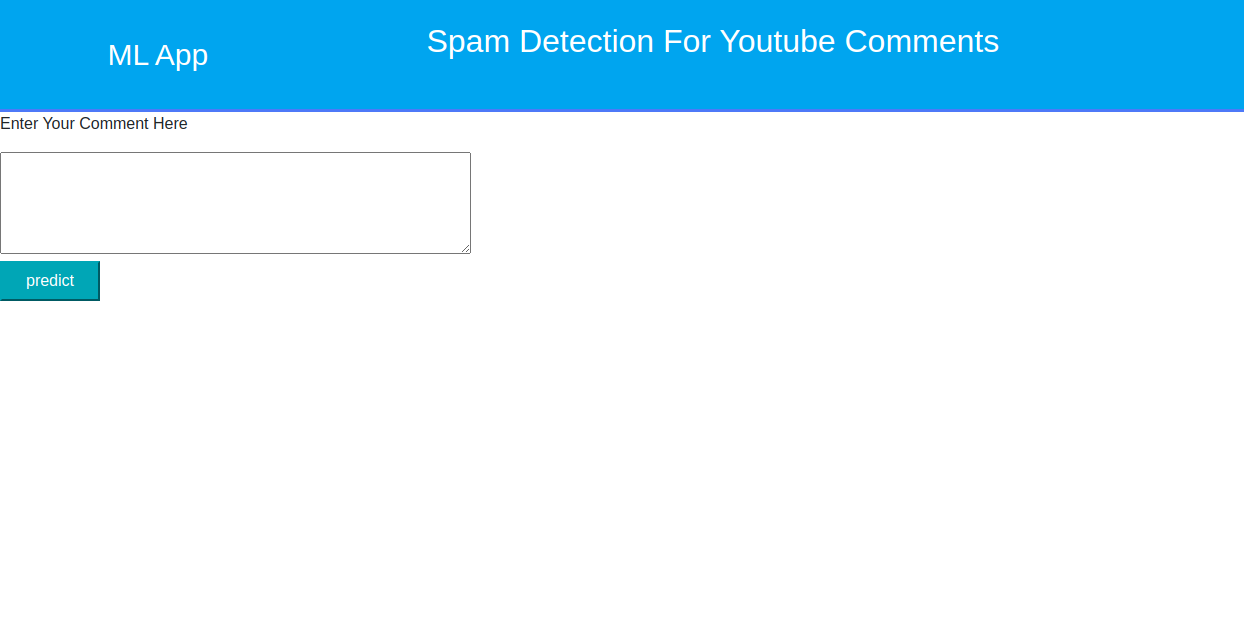
**From the above the SVM are giving better accuracy for prediction**

**SO we are implementing the flask framework for user base youtube spam detection with svm classifier**

****

****

**Loaclhost**

****