Data Collection and Preprocessing Phase

Date	June 2024
Team ID	739890
Project Title	The Language Of You tube: A Text Classification Approach To Video Descriptions
Maximum Marks	6 Marks

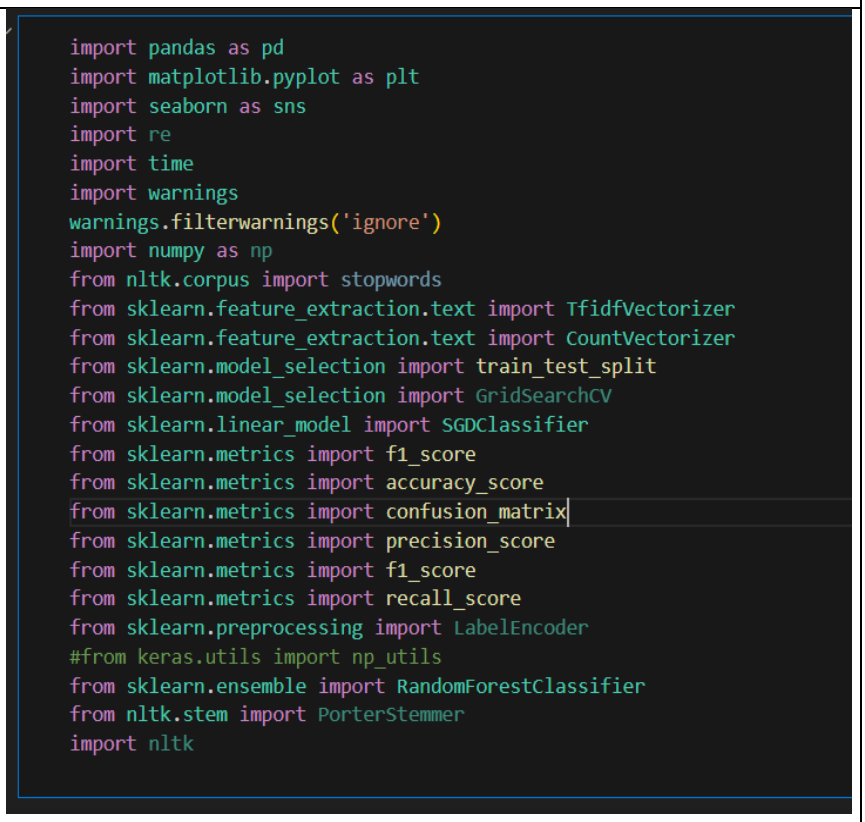
Preparation Template

The images will be pre processed by resizing, normalizing, augmenting, denoising, adjusting contrast, detecting edges, converting color space, cropping, batch normalizing, and whitening data. These steps will enhance data quality, promote model generalization, and improve convergence during neural network training, ensuring robust and efficient performance across various computer vision tasks.

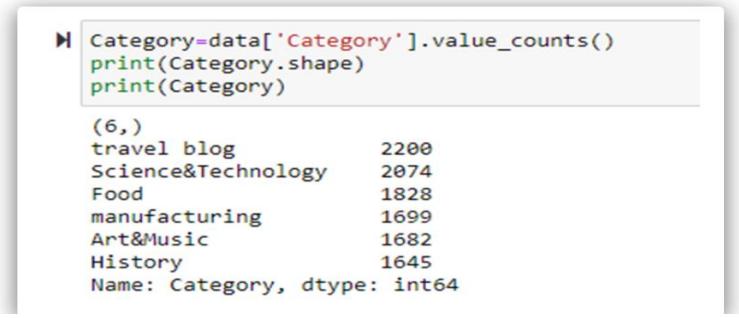
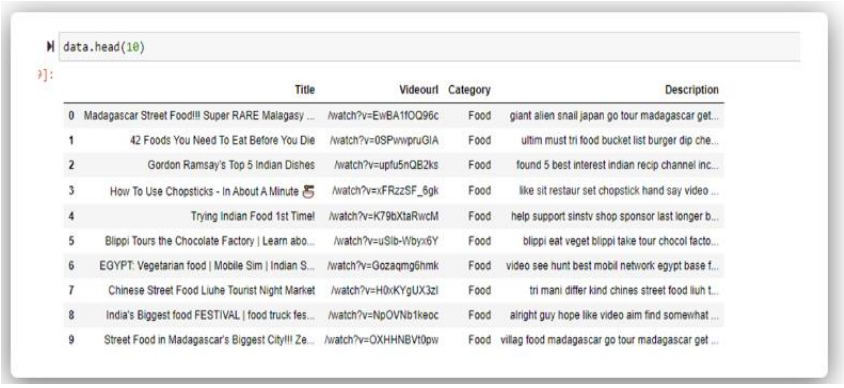
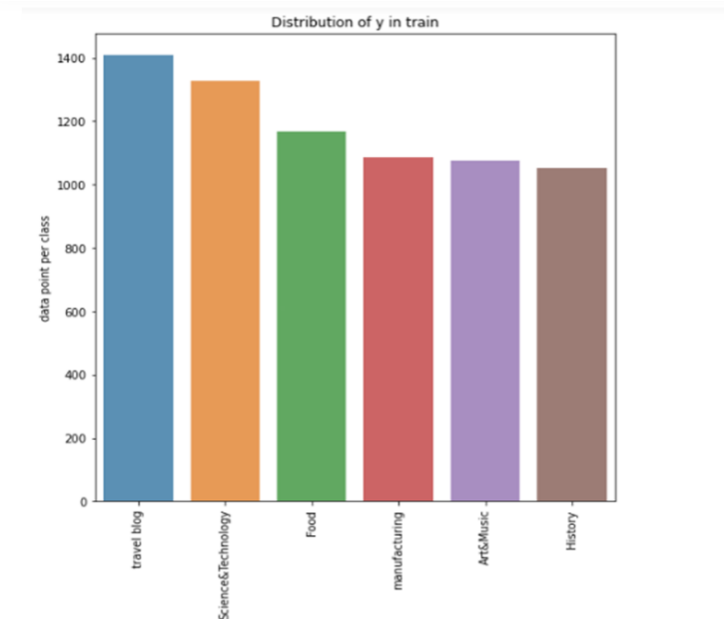
Section	**Description**
Data Overview	There are many popular open sources for collecting the data. Eg: kaggle.com, UCI repository, etc. In this project we have used .csv data.
Data Preparation	These are the general steps of pre-processing the data before using it for machine learning
Handling missing values	We use Handling missing values For checking the null values
Handling categorical data	As we can see our dataset has categorical data we must convert the categorical data to integer encoding or binary encoding

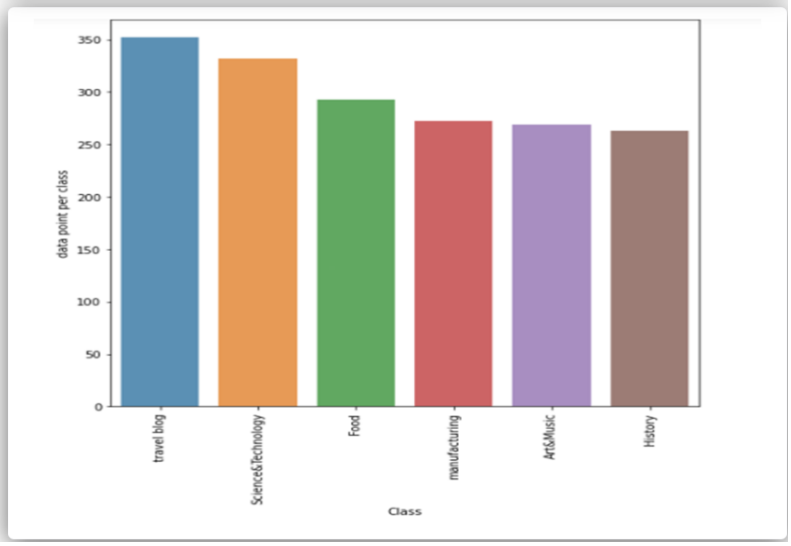
Handling Outliers in Data	With the help of boxplot, outliers are visualized. And here we are going to find upper bound and lower bound of numerical features with some mathematical formula.

Data Preparation

Collect the dataset	Please refer to the link given below to download the dataset. Youtube Videos Dataset (~3400 videos) (kaggle.com)
Importing the libraries	
Loading Data	We use the code Data =pd.read_csv('YoutubeDataSet.csv') For reading the dataset

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import re
import time
import warnings
warnings.filterwarnings('ignore')
import numpy as np
from nltk.corpus import stopwords
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.model_selection import train_test_split
from sklearn.model_selection import GridSearchCV
from sklearn.linear_model import SGDClassifier
from sklearn.metrics import f1_score
from sklearn.metrics import accuracy_score
from sklearn.metrics import confusion_matrix
from sklearn.metrics import precision_score
from sklearn.metrics import f1_score
from sklearn.metrics import recall_score
from sklearn.preprocessing import LabelEncoder
#from keras.utils import np_utils
from sklearn.ensemble import RandomForestClassifier
from nltk.stem import PorterStemmer
import nltk
```

Handling missing values	
Preprocessing of Description	
Visualizing the distribution of target variable in the test dataset	

data point per class
0
50
100
150
200
250
300
350
travel blog
Science&Technology
Food
manufacturing
Art&Music
History
Class