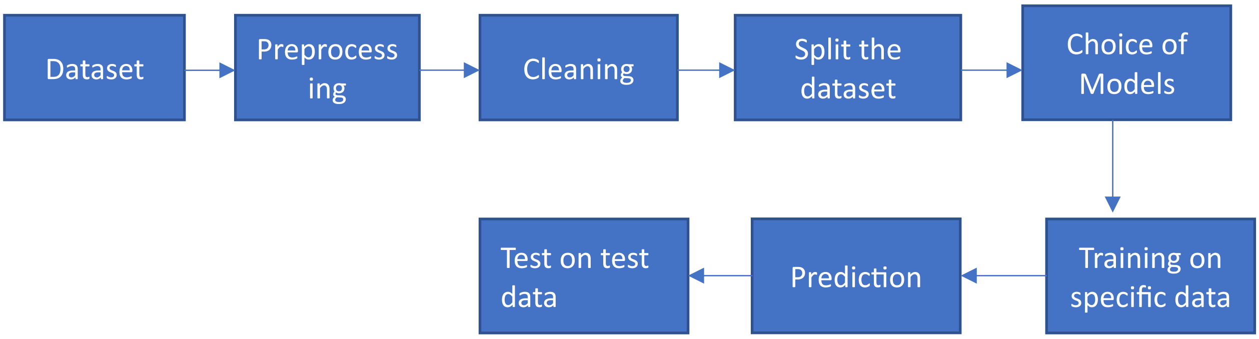
Amazon ML Challenge 2023

About the Dataset:

The dataset consists of around 2.9 million data entries with includes various unique, null, duplicate entries. Even the data in the each entry consists of various numerical numbers, HTML Tags, Hyperlinks. Further, it is sensed that elimination of data entries and choice of feature inorder to train the model is given priority. Inaddition to this, the data even consists of NULL values. Even, it is atmost important to check for the unique values for and I have observed the following PRODUCT_TYPE_ID, DESCRIPTION, BULLET_POINTS, TITLE are 12907, 745276, 965331, 2210763 respectively.

Architecture Adapted:

Inorder to understand the data and construct the models I have followed the below steps for the model construction.



Preprocessing of the data:

Inorder to select the right feature for training, ere each and every attribute accounted inorder to predict the PROCDUCT_LENGTH. So, in the preprocessing the major function is defined that includes the following:

1. Created a new column named data in the dataframe that is mix of all other column without any NULL values and it takes cares to include only unique values. By doing so the shape of the dataframe got reduced to around 1.03 million.
2. Now, the data attribute is read to get cleaned.

Cleaning of the data:

Inorder to clean the data columns nltk is incorporated. Stopwords, Punctations, Numbers, special Symbols, HTML tags are all removed. Then this model is ready for use.

Split the dataset:

The main reason to split the dataset is it consisted of around 1.03 million data entries so, here the data is split apart into 5 entries for the ease of training.

Choice of Model:

Inoder to account for precision and accuracy, the models proposed are DistillBert and TF-IDF based LinearSVM() model.

The models are trained and the predicted on the test data.