SMS Spam Classifier

Data Set: 5 Columns -> v1, v2, unnamed:2, unnamed:3, unnamed:4

V1 – spam/ham , 1/0

V2 – SMS

Unnamed 2 – N/A (50 non null values)

Unnamed 3 – N/A (12 non null values)

Unnamed 4 – N/A (6 non null values)

Shape of data (5572, 5)

Structure of this project

1. Data cleaning
2. EDA
3. Text Preprocessing
4. Model building
5. Evaluation of model
6. Improvement
7. Website
8. Deployment
9. Data cleaning:

* Will check other 3 columns that we need it or not
* These 3 columns have missing values so we will remove it

I will drop last 3 columns

Always check 2 things in data cleaning

1. Missing values
2. Duplicate values
3. EDA:

* first thought in my mind is that how much data is spam and how much data is spam.
* And I found there are 4516 messages are ham and remaining’s are spam
* 87 % is ham and 13 % is spam
* Data is imbalanced

Now ill check how many alphabets, words & sentences are using in SMS.

Now ill add 3 columns

1. Number of characters in the SMS
2. Number of words in the SMS
3. Number of sentences in the SMS

For this we will use NLTK library

3. Data Preprocessing / Textual preprocessing

* Lower case: converts data into lower case
* Tokenization: break it into words
* Removing special characters: removes @, $, # etc.
* Removing stop words and punctuation: stop words like is, of, the, are etc. no role in determining meaning of the sentence.
* Stemming/lemmatization: dance – dancing – danced = dance (converting into single meaning)

Now we will make word cloud of most used words in spam and ham

Also while building machine learning model interpretability is most important for better understanding.

1. Model Building

* Will apply naïve bayes model because its works best on textual data.
* Also will try to see other model on this data too.
* We need to convert text data into numeric – vectorization
* Importing all naïve bayes model because we don’t know the distribution of the data
* Precision score is high for multinomial naïve bayes so we will go with that