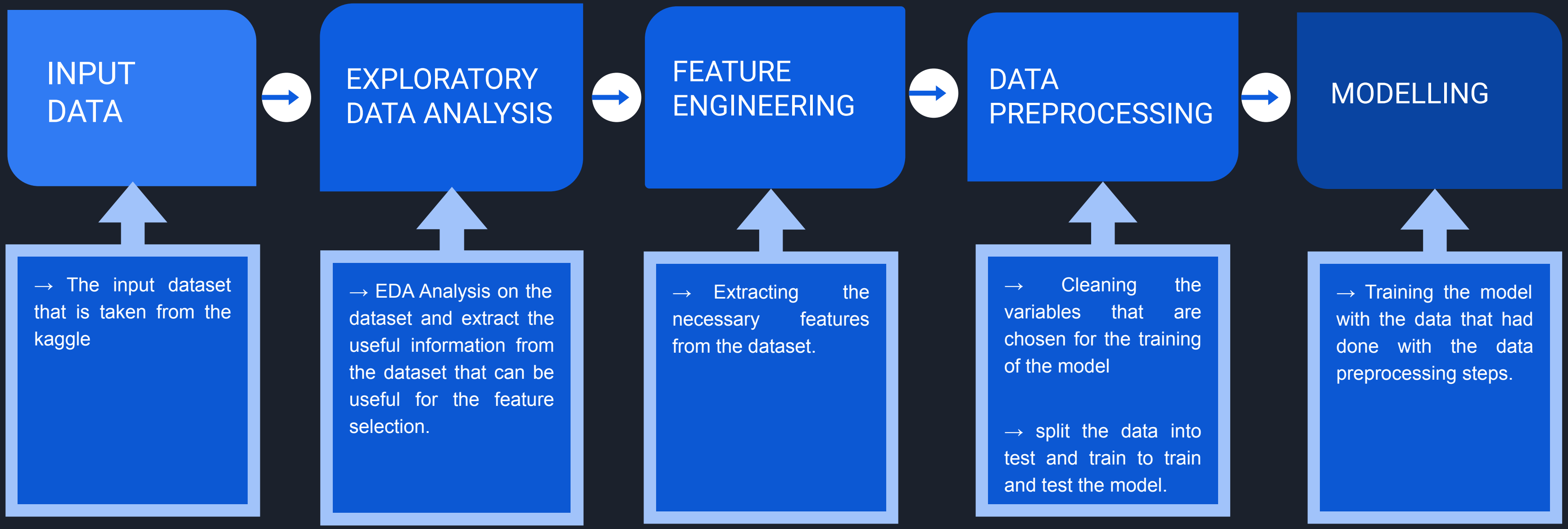
SPAM VS HAM CLASSIFICATION

SPAM VS HAM CLASSIFICATION



INPUT DATA

→ The input dataset that is taken from the kaggle

EXPLORATORY DATA ANALYSIS

→ EDA Analysis on the dataset and extract the useful information from the dataset that can be useful for the feature selection.

FEATURE ENGINEERING

→ Extracting the necessary features from the dataset.

DATA PREPROCESSING

→ Cleaning the variables that are chosen for the training of the model

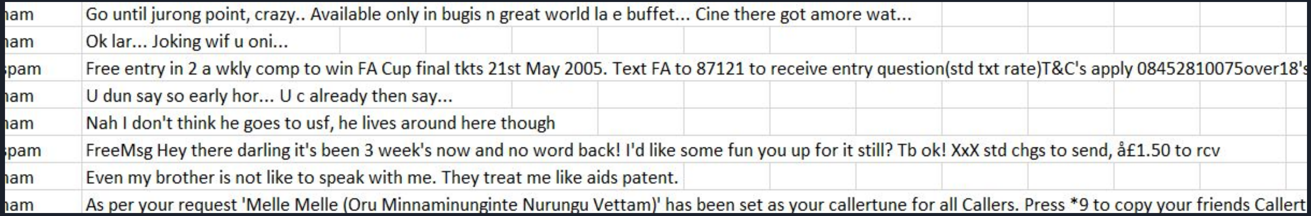
→ split the data into test and train to train and test the model.

MODELLING

→ Training the model with the data that had done with the data preprocessing steps.

INPUT DATASET

- The Dataset is a SMS spam collection and consists of around 5500 messages this dataset is taken from UCI Repository. https://archive.ics.uci.edu/ml/datasets/SMS+Spam+Collection



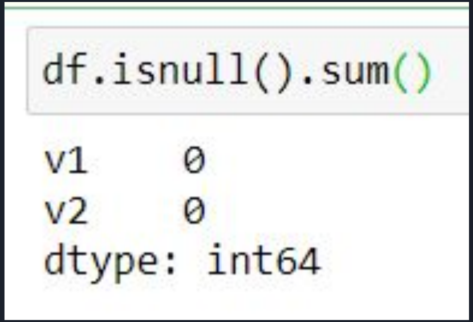
ham	Go until jurong point, crazy.. Available only in bugis n great world la e buffet... Cine there got amore wat...
ham	Ok lar... Joking wif u oni...
spam	Free entry in 2 a wkly comp to win FA Cup final tkts 21st May 2005. Text FA to 87121 to receive entry question(std txt rate)T&C's apply 08452810075over18's
ham	U dun say so early hor... U c already then say...
ham	Nah I don't think he goes to usf, he lives around here though
spam	FreeMsg Hey there darling it's been 3 week's now and no word back! I'd like some fun you up for it still? Tb ok! XxX std chgs to send, å£1.50 to rcv
ham	Even my brother is not like to speak with me. They treat me like aids patent.
ham	As per your request 'Melle Melle (Oru Minnaminunginte Nurungu Vettam)' has been set as your callertune for all Callers. Press *9 to copy your friends Callert

- Each row consist of one message and labelled it as spam or ham(normal message).

EXPLORATORY DATA ANALYSIS

- In this EDA part let’s know more about the data we are going to use to train the model.
- First, let’s Check of any null values and remove them.
- Next, check if it is a balanced data set or not.
- And continue with the analysis of the Spam and Ham Messages.

Presence of Null Values

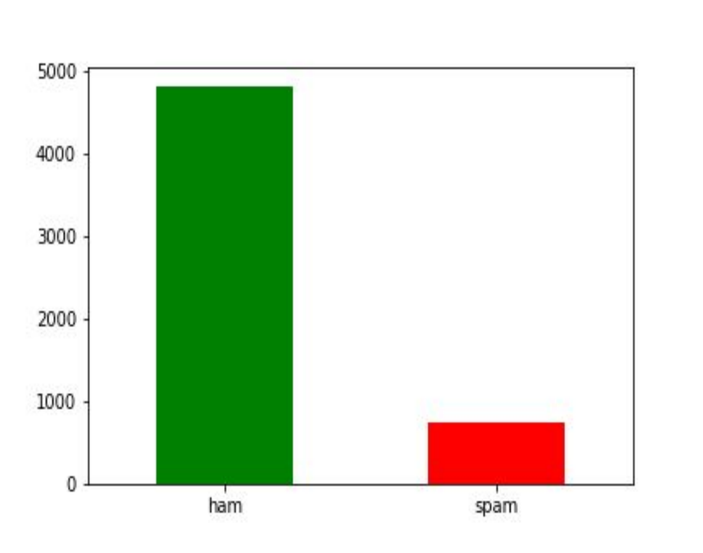


```
df.isnull().sum()
```

```
v1    0
v2    0
dtype: int64
```

→ The are no null values in Dataset ,So data augmentation is not required.

SPAM vs HAM Imbalance



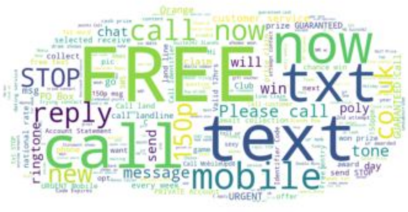
→ Here we can clearly see that the Ham is more than Spam in the dataset.

→ Here the SPAM is 15.50 % of the HAM.

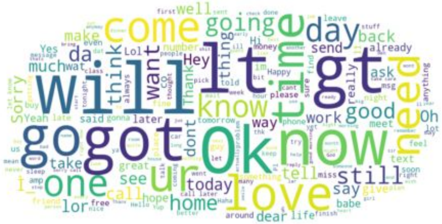
→ So,we can say that it is an imbalance dataset.

Word Cloud for SPAM and HAM

SPAM WORD CLOUD



HAM WORD CLOUD

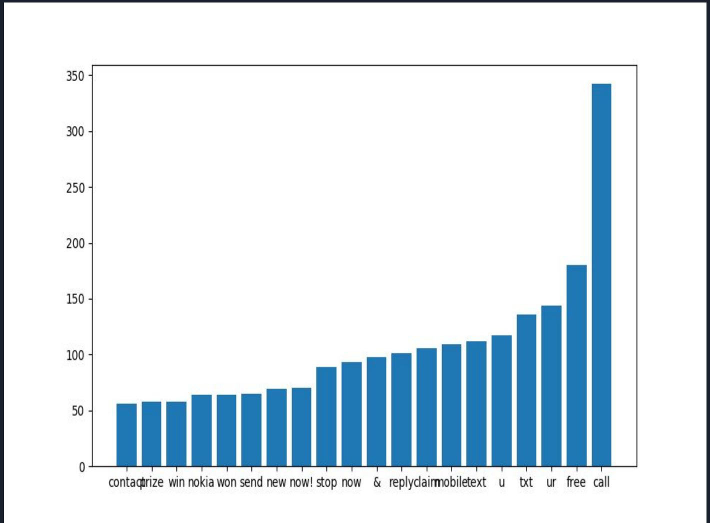


→ Here You can see the most prominent words used in SPAM and HAM messages.

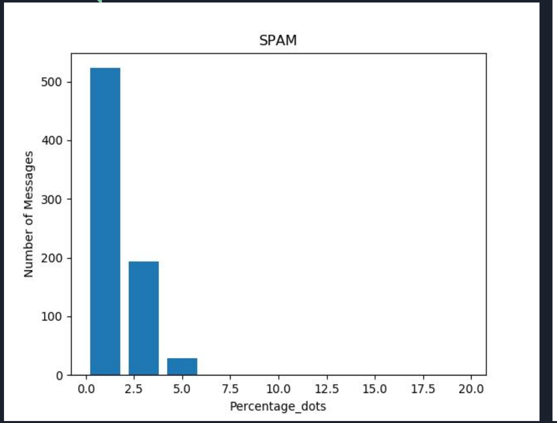
Frequent Words in SPAM

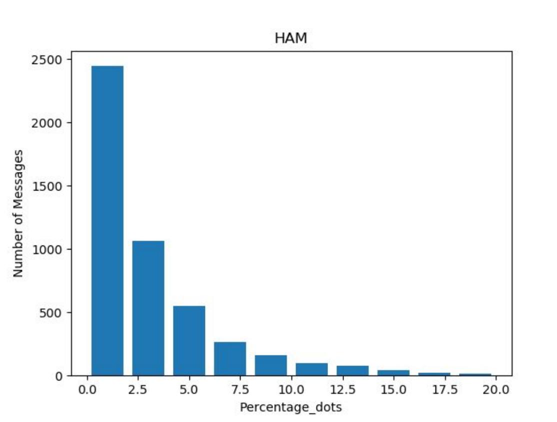
→Here we can see the frequently used words in the SPAM messages.

→Spam_Message: “SIX chances to win CASH! From 100 to 20,000 pounds txt> CSH11 and send to 87575. Cost 150p/day, 6days, 16+ TsandCs apply Reply HL 4 info”



Percentage of Full stops SPAM vs HAM

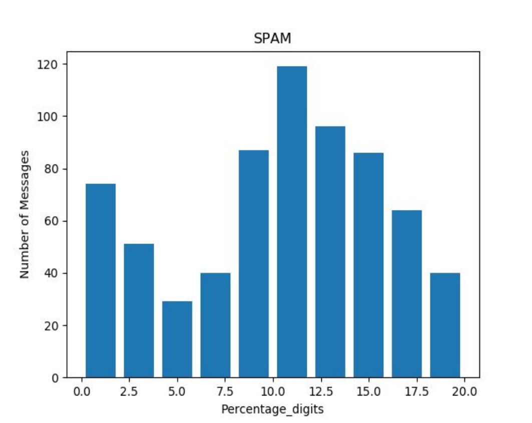


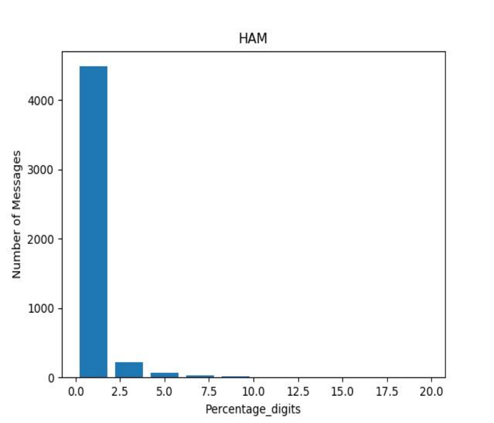


→ Normally in the Messages we use so many dots in the messages.

→ Ham_Message:
“Ok lar... Joking wif u oni…”

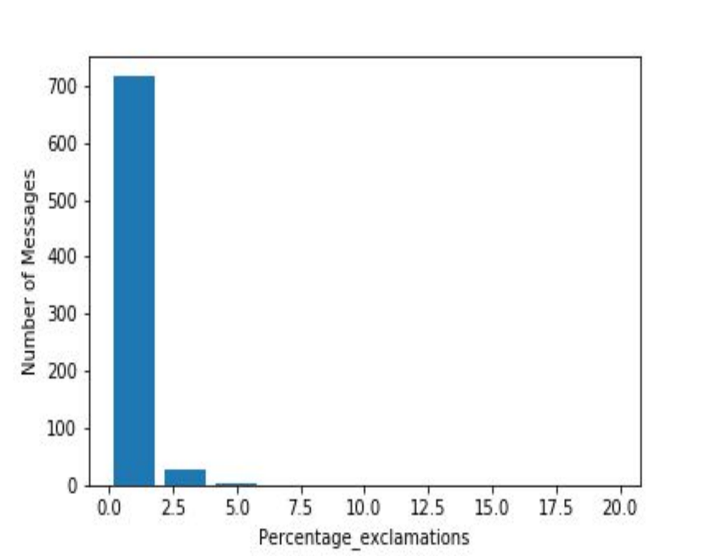
Percentage of DIGITS SPAM vs HAM





→ Usually in SPAM messages they ask for money and send Phone numbers in the message content, So there will be many digits in the spam messages as you can see in the bar graph.

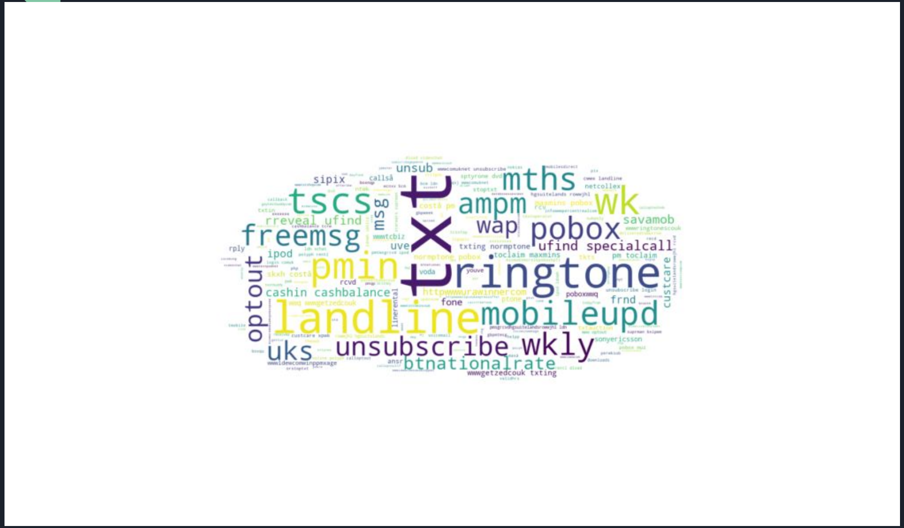
Percentage of Exclamations



→ The exclamations are more used in the spam messages.

Ex: Free!! , Claim!!!

More Slang Words are used in spam



→ Here you can see the like in HAM more Slang words are used in SPAM messages.

THANK YOU