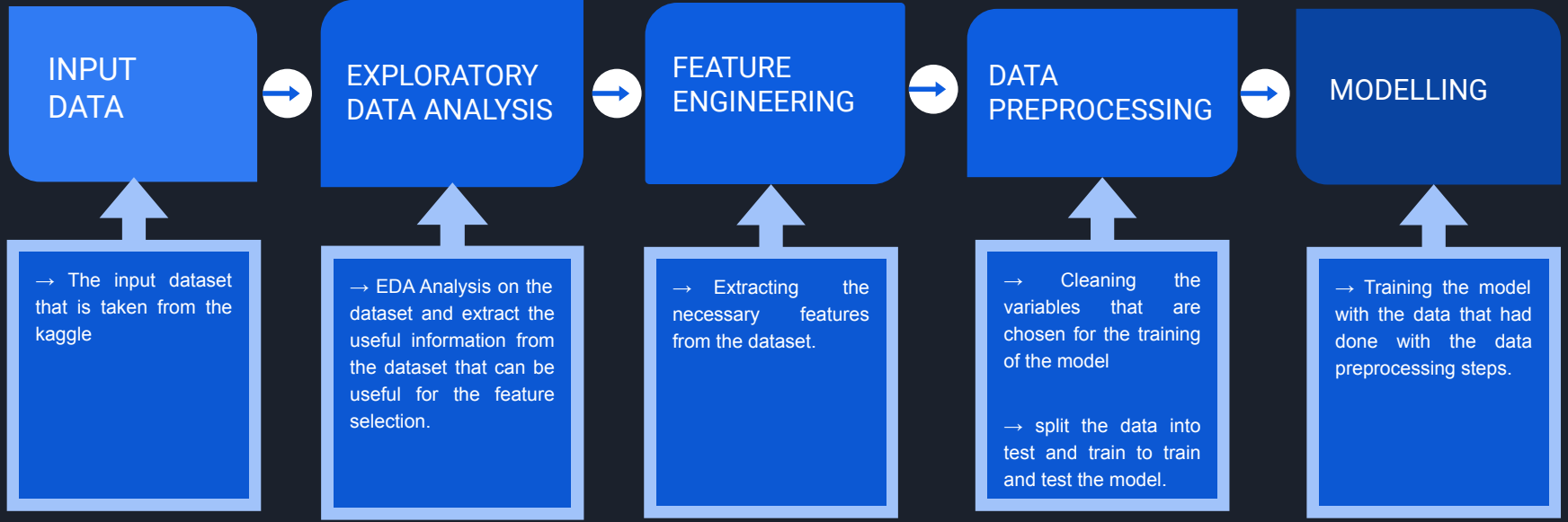
SPAM VS HAM CLASSIFICATION

SPAM VS HAM CLASSIFICATION



INPUT DATASET

- The Dataset is a SMS spam collection and consists of around 5500 messages this dataset is taken from kaggle.
  https://www.kaggle.com/uciml/sms-spam-collection-dataset

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

- Spam and Ham are the target variables used to train the Machine Learning Model.

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

→ The are no null values in Dataset

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

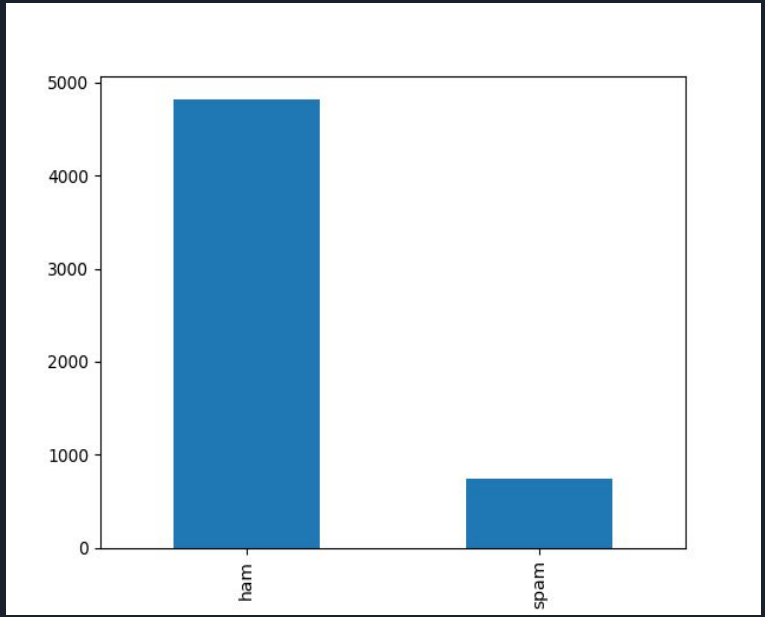
```
v1    0
v2    0
dtype: int64
```

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.



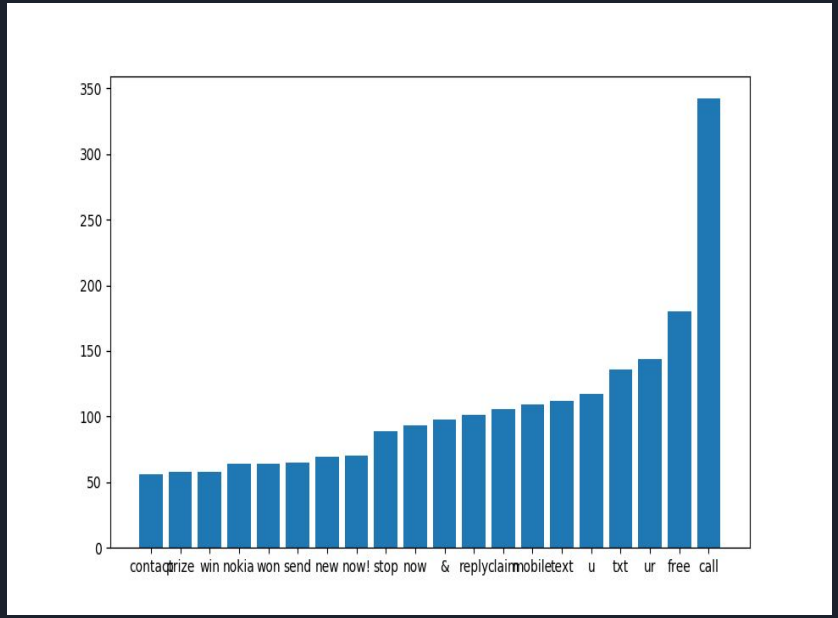
Word Cloud for SPAM and HAM

SPAM CLOUD

HAM CLOUD

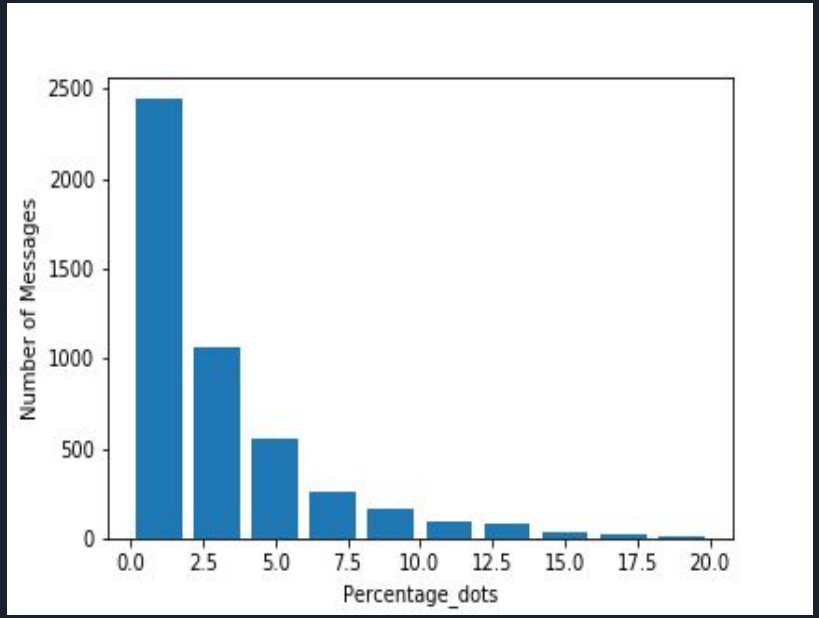
Frequent Words in SPAM

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



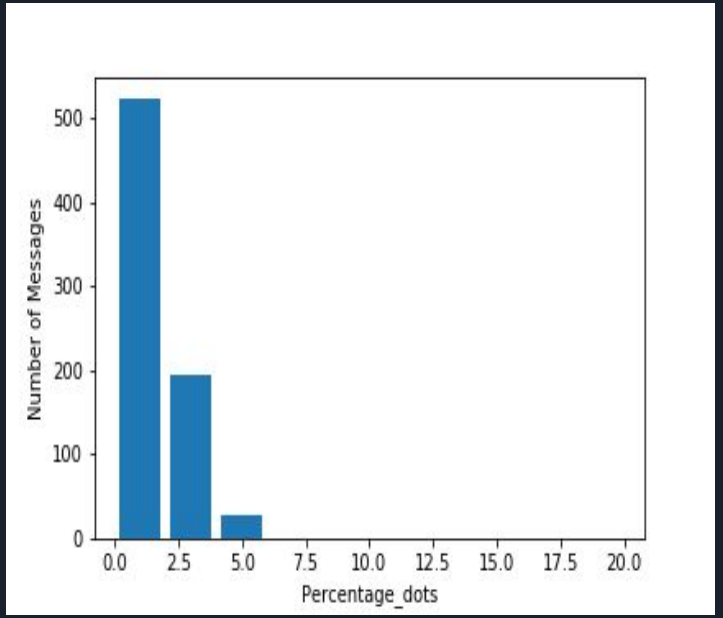
Percentage of Full stops ham

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



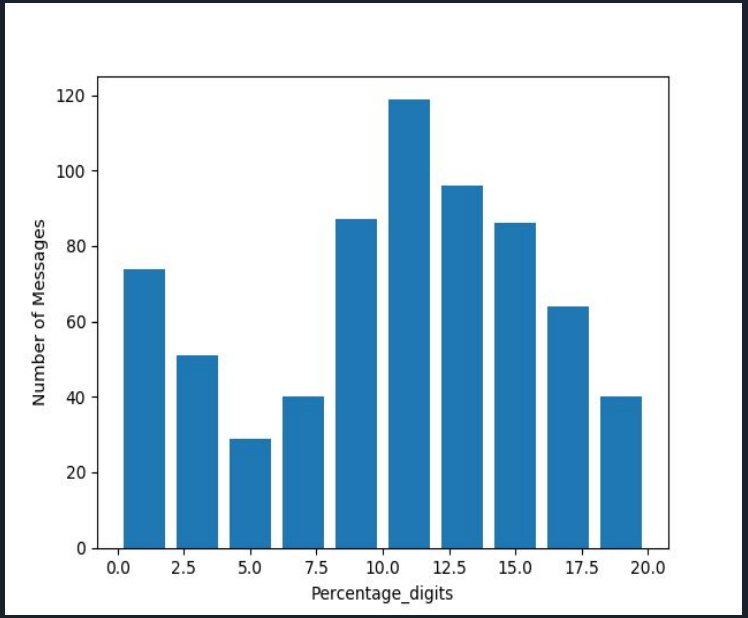
Percentage of Full stops Spam

→In Spam messages not more Full stops are used.



Percentage of DIGITS

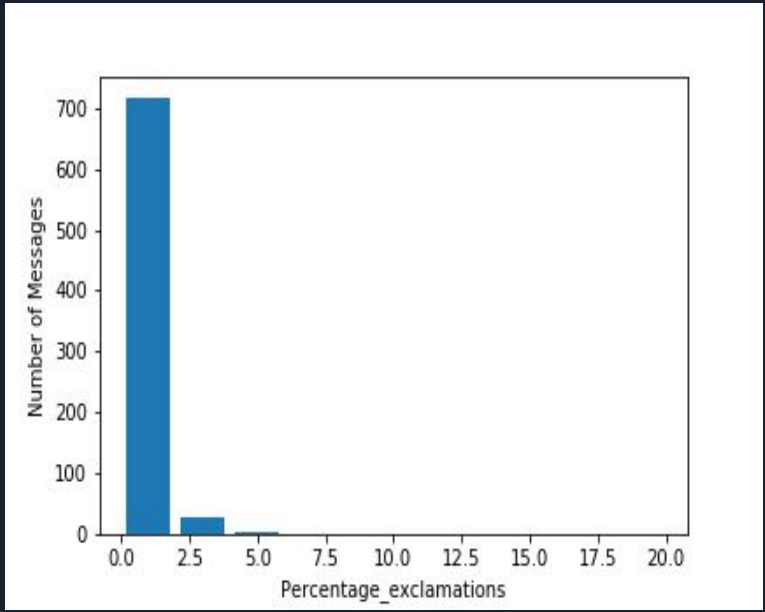
→ Usually in SPAM messages they ask for money and send Phone numbers in the message content, So there will be many digits in the spam messages as you can see in the bar graph.



Percentage of Exclamations

→ The exclamations are more used in the spam messages.

Ex: free!! , Claim!!!



Spelling mistake word cloud

→ Here you can see the spelling mistakes in the world cloud. Usually there are more spelling mistakes in the Spam.

THANK YOU