**Email filter project:**

**Goal:** Everyday we receive a huge chunk of mails, It would be easier to sort out the useful mails from the spam ones, as this could be time saving. In this project my main focus is to filter out the spam mails, from useful ones.

**Approach and techniques used :**Since ML is a becoming really popular these days, this would be a useful implementation of ML.

I used the following steps to work on my project:

As a part of data collection, I downloaded the dataset from kaggle.

Then comes, data preprocessing. As a part of data pre-processing, I made sure to replace all the NULL values by a specific value.

Then I performed Label Encoding. Assigned label 0 to one of the categories from spam /not spam and 1 to the other. Label encoding produces numerical values from text values, thus giving continuous values, which are easier to work upon.

Then I performed feature extraction in which all the un important features were removed, which have no contribution in determining whether the email is spam or not.

The next important step was splitting the data set into training and test data.

Training data is one on which our model will be trained and test data will be one which is new to our model and used for prediction purpose.

The model I used was Logistic Regression since it works best for binary classification problems.

**Project Outcomes:**

I received an accuracy of 97% on training data and 96% on test data.

**Challenges:**

Data Preprocessing was really a very time consuming process since it required removal of unecessary email headers.

Handling missig data: By using stastical measures like mean and max value.

**Future scope:**

Multilingual support can be added and NLP can be used in text preprocessing.

Privacy and security of users is most important and this can be achieved to an extent by removing usernames and email ids.