**PROJECT REPORT**

**TOPIC: - E-MAIL CLASSIFICATION**

**1. Introduction**

Email classification, also known as email categorization or email tagging, is the process of automatically organizing and categorizing emails into predefined categories or labels. This process is often used in email filtering to separate important emails from spam or to route emails to the appropriate department or person within an organization. Email classification can be done using various machine learning algorithms, such as Naive Bayes, Random Forest, Decision Tree or Logistic Regression, which are trained on a labelled dataset containing emails and their corresponding categories.

**2. Research Objective**

Make a model which correctly classify the ham and spam emails separately with high accuracy, high precision value with smooth upward trajectory of ROC curve. In this classification we have to focus on Type-I Error. It is that, when the mail is spam and model detect it ham. We have to focus on Precision Score. It should be high as possible as.

**3. Data Collection**

We took dataset from Kaggle. Dataset contain 4 columns and 5171 rows. Text column contains description of mails whereas label column contains mail's details  
i.e., mail is spam or ham.

**4. Data & Text Preprocessing**

(i) Find Null and duplicates values in the dataset.

(ii) Drop unnecessary columns.

(iii) Remove stop words.

(iv) Remove Special characters & numbers.

(v) Converting text to lowercase and done stemming.

(vi) Do Count vectorizer – Count Vectorizer is a method used in natural language processing to convert a collection of text documents into a matrix of token counts. In simpler terms, it takes a bunch of text and turns it into numbers, counting how often certain words appear in each document.

(vii) Made a word cloud for ham and spam mails.

**5. Used Machine Learning Algorithms: -**

**(i) Logistic Regression: -**

Logistic regression in text classification of emails is a statistical method used to model the relationship between the textual features of an email (such as word frequencies) and the probability that the email belongs to a particular category (spam or ham). It calculates the likelihood of an email being spam based on the features extracted from its text, allowing for binary classification into spam or ham categories.

**(ii) Random Forest Classifier: -**

Random Forest Classifier in text classification of emails is a machine learning algorithm that uses an ensemble of decision trees to classify emails into spam or ham categories. It works by constructing multiple decision trees during training and outputs the class that is the mode of the classes (classification) of the individual trees.

**(iii) Decision Tree Classifier: -**

Decision Tree Classifier in text classification of emails is a machine learning algorithm that uses a tree-like model of decisions to classify emails into spam or ham categories. It breaks down the email text into a series of questions based on the features (words or phrases) in the text, leading to a decision (spam or ham) at the end of the tree.

**(iv) Naive Bayes: -**

Naive Bayes in text classification of emails is a probabilistic machine learning algorithm that uses the Bayes theorem to predict the probability that a given email belongs to a particular category (spam or ham) based on the words or features present in the email. It assumes that the presence of each word in the email is independent of the presence of other words, which is why it's called "naive."

**6. Validation Technique/Evaluation Technique**

**(i) Accuracy Score: -**

Accuracy score is a measure of how often a model's predictions are correct, calculated as the number of correct predictions divided by the total number of predictions.

**(ii) Precision Value: -**

Precision is a measure of how many correctly predicted positive instances (true positives) there are among all predicted positive instances, indicating the model's ability to avoid false positives.

**(iii) Recall Value: -**

Recall, also known as sensitivity or true positive rate, is a measure of how many correctly predicted positive instances (true positives) there are among all actual positive instances, indicating the model's ability to find all positive instances.

**(iv) F-1 Score: -**

The F-1 score is a single metric that combines both precision and recall into a single value. It is the harmonic mean of precision and recall, giving a balanced measure of a model's performance on both positive and negative classes.

**(v) Classification Report: -**

A classification report is a summary of the performance of a classification model, showing metrics such as precision, recall, F1- score, and support for each class in the dataset. It helps in evaluating how well the model is performing for each class.

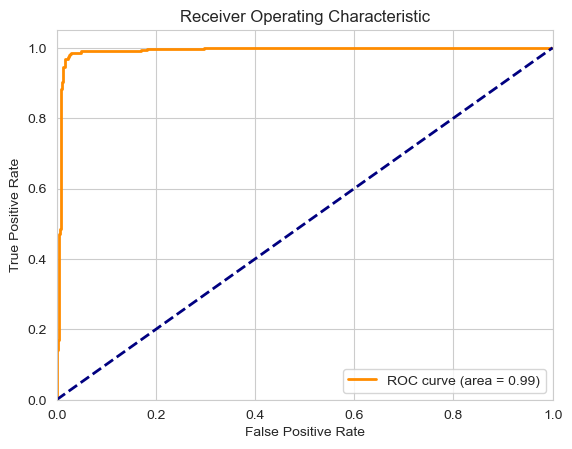
**(vi) ROC Curve: -**

The ROC (Receiver Operating Characteristic) curve is a graphical representation of the performance of a binary classification model. It shows the trade-off between the true positive rate (sensitivity) and the false positive rate (1 - specificity) as the classification threshold is varied.

**7. Finalize Model**

Logistic Regression is our final model because it gives high accuracy (97.87%), high precision value (99%) with smooth upward trajectory of ROC Curve.

**8. Visualization of ROC Curve: -**



### 9. Purpose of the Project: -

### This model helps detect and filter out spam, fraud, and legitimate (ham) emails, protecting us from cyber attackers and scammers who try to deceive or defraud us through email communication.