**Detecting email spam or ham using Machine Learning**

In addition to the undeniable benefits, the development of the Internet has led to many undesirable security effects. Spam emails are one of the most challenging issues faced by the Internet users. Spam refers to all emails of unsolicited content that arrive in a user’s email box. Spam can often lead to network congestion and blocking or even damage to the system for receiving and sending electronic messages. Thus, appropriate classification of spam email from legitimate email has become quite important. This paper presents a new approach for feature selection and Iterative Dichotomiser 3 (ID3) algorithm designed to generate the decision tree for email classification. The experimental results indicate that the proposed model achieves very high accuracy.

**EXISTING SYSTEM:**

Unfortunately, the continuous rise of email users has led to a massive increase of spam emails. Whether it is commercial in nature or not, spam emails can cause serious problems in electronic communication. Spam emails produce huge amount of unsolicited data and thus affect the network bandwidth and storage capacity. Due to the large number of spam emails to users of email services it is difficult to distinguish useful from unsolicited emails. Thus, managing and filtering emails is an important challenge. The filtering purpose is to detect and isolate spam emails.

**DISADVANTAGES OF EXISTING SYSTEM:**

* Spam emails are usually sent in bulk and do not target individual recipient.
* he aim is to preserve the most important features and to reduce computations demand.

**Algorithm**: Linear Discriminant Analysis, Classification.

**PROPOSED SYSTEM:**

In the model development, the dataset is consistently split into train and test set of 80% and 20%. Train set has 400 profiles and test set has 100 profiles. The dataset used for modelling looks like this. Preprocessing is a crucial step in method. The aim is to clean the data and prepare it for use in a prediction algorithm. Few improvements are required for the data obtained from Occidental College in order to make it suitable for the proposed machine learning algorithms. Determining how to deal with missing data is a common problem in data cleaning. Since the function in question could be a good predictor of the algorithm's outcome, it's critical to find missing entries, locate them, and apply a treatment based on the variable form that enables us to use the data in the model. The data was pre-processed and split into two classes at random: a training set and a testing set. We selected 80 percent of the 7976 entries in our dataset as our training collection.

**ADVANTAGES OF PROPOSED SYSTEM:**

* The proposed approach is evaluated using accuracy, precision and recall.
* In the near future, it is planned to incorporate other classifiers and to compare their performances with the proposed approach.

**Algorithm**: Machine learning, ID3 algorithm, classification algorithm.

**SYSTEM REQUIREMENTS:**

**HARDWARE REQUIREMENTS:**

System : Intel i5 6 core.

Hard Disk : 500 GB SSD.

Monitor : 15’’ LED

Input Devices : Keyboard, Mouse

Ram : 32 GB.

**SOFTWARE REQUIREMENTS:**

Operating system : Windows 10.

Coding Language : Python

Tool : PyCharm, Visual Studio Code

Database : SQLite