Methodology

Dataset :

The dataset considered in the current research is collected from about 150 person . The dataset “ SMS spam collection dataset “ contains (150 rows and 5 columns) instances and five attributes “sample\_no , source , category , text\_content , spam or ham “.

The predicted label has two classes :

0 = non-spam and 1 = spam .

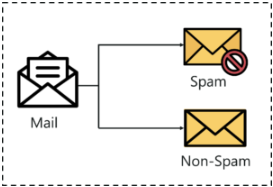
In the data , about 105 are non-spam samples and 45 are spam samples.

The dataset is given in table 1.

1 .Classification Algorithm :

In machine learning, classification is a supervised learning concept which basically categorizes a set of data into classes. The most common classification problems are – [speech recognition](https://www.edureka.co/blog/speech-recognition-python/), [face detection](https://www.edureka.co/blog/tensorflow-object-detection-tutorial/), handwriting recognition, document classification, etc. It can be either a binary classification problem or a multi-class problem too.

We choose only classification instead of linear regression , because our problem is to classify into ham or spam .



The following machine learning algorithm were considered for classification of ham and spam .

* 1. Logistic Regression :

It is a classifier problem in binary classification is computing the value of predictive y while

y ∈ [0, 1];

0 and 1 are for class negative and positive. The LR predicts the variable value of multiclassification such as

y ∈ [0, 1,2, 3].

* 1. Decision Tree.

It is a supervised machine learning algorithm . Its shape is like a tree in which each node is a decision node or leaf. The technique of DT is easily understandable and simple for making the decisions. A DT contains external and internal nodes interlinked with each other. Decision can be made based on the internal nodes and the child node to access the preceding node. There is no child of the leaf node and is linked with a label.

* 1. Support Vector Machine :

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning.

The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.

SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine.

SVM algorithm can be used for **Face detection, image classification, text categorization,** etc

Consider the below diagram in which there are two different categories that are classified using a decision boundary or hyperplane:



* 1. Random Forest Classifier :

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of **ensemble learning,** which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

**The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.**



Below are some points that explain why we use the Random Forest algorithm:

* It takes less training time as compared to other algorithms.
* It predicts output with high accuracy, even for the large dataset it runs efficiently.
* It can also maintain accuracy when a large proportion of data is missing.

1.5 Naïve Bayes Classifier Algorithm :

1 Naïve Bayes algorithm is a supervised learning algorithm, which is based on **Bayes theorem** and used for solving classification problems.

1. It is mainly used in *text classification* that includes a high-dimensional training dataset.
2. Naïve Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions.
3. **It is a probabilistic classifier, which means it predicts on the basis of the probability of an object**.
4. Some popular examples of Naïve Bayes Algorithm are **spam filtration, Sentimental analysis, and classifying articles**.

1.6 Division of Dataset.

The dataset were split into 30% and 70% for validation and training of the predictive model.

……………………………………………………………………………………………

table 1 .

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| sample\_no. | source | category | text\_content |  |
| 1 | education |  | dfsads |  |
| 2 | banking |  | fdfd |  |
| 3 | Insurance |  | gfs |  |
| 4 | shopping |  | sgs |  |

Confusion matrix 2 .

|  |  |  |
| --- | --- | --- |
|  | Predicted spam(1) | Predicted ham(0) |
| Actual spam(1) | TP | FN |
| Actual ham(0) | FP | TN |