**Process for Fake news analysis using Machine Learning**

**Tools and Language: -**

In this research python is used for the machine learning part. Python is one of the most popular

language scientific areas and ML. For running python first, we have to install Anaconda software. After installing this software, we will get Anaconda navigator. From this navigator we have to launch jupyter notebook. In jupyter notebook to run the codes first we have to install packages like numpy, pandas, nltk, sklearn by using the syntax

**!pip install ‘package name’**

For using nltk function like word tokenizer and stop words removal we have to download these also by using syntax:

**nltk.download(“stopwords”)**

**nltk.download(“punkt”)**

**Note: -To run the codes, we have to press Shift Enter**

## **Dataset Used: -**

The field of fake news detection is a relatively new area of research. Hence, few public datasets are available. We used in this work primarily a dataset from Kaggle which has two datasets one for real news and other one for fake news separately. Real news dataset has 15712 rows and 11 columns. Fake news data set has 12999 rows and 20 columns. For combining both the dataset we took some variables from real news dataset and same columns from fake news dataset and combined both to obtain a single dataset.

**Process in Fake news analysis: -**

* First, we will import both news data i.e. real and fake news data.
* We will add a new column **label**. Label values will be real for real news data and fake for

Fake news data.

* We will select same desired columns from both the dataset and combined them.
* Now, we are left with 4 columns in the final dataset: Title, content, publication and label. Title is the title of the news and content is the complete information on the news. Publication is the publication in which news was published and label is representing whether the news is true or false.
* Then we will change the values in label column by 1 and 0. We assigned 1 for fake news and 0 for true news because we have to detect fake news.
* Then, we will check the missing values in data. If there is any missing data, we will drop that values as the data is text.
* Next, we will change all the records in the **content** column to lower case.
* Then we will remove punctuation and tokenize the text data in content variable. The function used for this **RegexpTokenizer** is available in nltk library in python. This function remove punctuation and do world tokenizer of content variable at a time.
* After this we will remove stop words e.g. a, an, the, or etc.
* Then we spilt the dataset into training and testing data in 70:30.
* Content variable of final dataset work as X and response variable which have 1 and 0 values will be as Y.
* Then we will do feature selection by using two techniques i.e. TFIDF vectorizer and count vectorizer.
* Then I run 5 Machine learning algorithms on training data for both the feature selection techniques and evaluate the model performance on test data. The 5 ML algorithms are:

**Naïve Bayes**

**SVM**

**Random Forest**

**Decision trees**

**Logistic regression**

* Based on all evaluation metrices obtained from above ML algorithms like accuracy, recall score, precision and f1-score we got that SVM as the best performing algorithms for fake news analysis for our dataset on both type of feature selection techniques.