THREE-WAY CHARACTERIZATION OF UNCERTAINTY ARISES DUE TO INFORMATION POLLUTION

Developed a supervised ML model that classifies the uncertainty arising from information pollution in mainstream Indian news channels into 3 categories: polluted, not polluted, or uncertain based on their ability to provide quality content for students.

TABLE OF CONTENTS

ABSTRACT

On a day-to-day basis, people consume information on a variety of subjects acquired from different sources, the worrying aspect is, consumers hardly look for the authenticity of the information, which otherwise can be polluted to a larger degree and in many ways. General and electronic media in particular has somewhere forgotten its role and responsibility by passing on fake news, sensationalizing news items, breaking news stories and uncredible information to society at large.

In order to improve the trustworthiness of information being disseminated and limit the damaging consequences of information pollution, quick identification is essential. Thus, our goal here is to show if the information in the data source is relevant to users and classify the information as polluted or not based upon the relevance such that the users can focus only on significant resources for consuming information and avoid the trivial resources that do not fulfill their requirements.

To demonstrate the concept, ML models namely decision tree classifier, random forest classifier, logistic regression, and support vector machine are used to determine if the debates across major Indian news channels shared on social media is pertinent to students and classify the data as polluted or not polluted, or not sure. Based on the debates conducted by mainstream news channels on certain events, it is determined if the particular news channel is intended to have quality content for students such that they can easily decide on which news channel can they rely on.

INTRODUCTION

People are well aware about pollutants from air, noise, soil, water, etc., which are very common. Information pollution has existed for a long time, but consumers of information have not yet realized the pollution associated with it and the worrying aspect here again is, by the time people will realize the need for and importance of pollution free information, the situation might become irreversible.

Internet has become a widespread, large scale and easy to use platform for real-time information dissemination with India being the world's second-largest internet user base. Social media offers its users a large-scale and easy-to-use platform, which cannot be provided using traditional media. Social Media users face two crucial problems when using this platform in regard of user generated content: one is that the indeterminable the quality of the information published in which that information may be false and the other is the difficulty of detecting copyright violations.

Media is gaining tremendous attraction and a huge user base from all sections and age groups of society. Despite the increasing use of media platforms for information and news gathering, its unmoderated nature often leads to the emergence and spread of trivial news, that is, items of information that are of less importance making it an opportunity for anyone to reach a broad audience very quickly; mainstream news channels companies are doing precisely that. The question that concerns us is to what extent the content that circulates among these platforms changes every second the mentality, the perceptions and the lives of billions of people are verified, authenticated and in compliance with standards.

Education, one of the most important areas of any country, and providing polluted knowledge among researchers and students might destroy the scientific temper of young brains. This actually has motivated us to study the possibility of helping limit student's information consumption to significant resources and avoid relying on trivial sources of information resulting in improvement of the decision capability and productivity.

The dataset taken consists of 2 attributes: topic and number of debates respectively. It is a list of last 262 debates across major Indian news channels Zee News (55 episodes of zee news show taal thok ke), India TV (56 episodes of INDIA TV show Kurukshetra), Aaj Tak (49 episodes of Aaj Tak show Dangal), News 18 India (57 episodes of News18India named Aar Paar), Republic Tv (45 episodes of Republic TV show The Debate) shared on social media till October 23, 2019.

According to the debates shared by mainstream media news channels, to show that the topics covered in the debates were supporting the ruling government, and it is ideology directly or indirectly which are irrelevant to our target audience - students which allows them to check for available relevant information from credible sources and avoid dilatory on superfluous sources making it more reliable and safer for decision making as well as for knowledge sharing. It is worth mentioning that all tests will be done using python language and Jupyter notebook.

LITERATURE REVIEW

A great number of researches and studies have been done on information pollution:

1. False Information Ecosystem

Fake information, which is present in the form of images, blogs, messages, stories, breaking news; referred to as information pollution has many formats that are not mutually exclusive, but also a certain heterogeneity which puts them into a specific category.

- Information pollution, outcome of an information revolution, where people receive contaminated information, which is of less importance, irrelevant, unreliable, and unauthentic,which lacks exactness or precision, which is always detrimental to society in general.
- Rumour, unverified information which is not necessarily false; can turn out to be true too.
- Fake news**, f**alse information spread under the guise of being an authentic news usually spread through news outlets or internet with an intention to gain politically or financially, increase readership,biased public opinion

Priyanka Meel, Dinesh Kumar Vishwakarma illustrated the Motivation behind information pollution.

- Political Intent: To smear the opponent's public image or to support a person or a political party.
- Financial Profit: False-positive information fosters large-scale investment and has an impact on stock prices. Fake product ratings and reviews written on purpose to boost sales.
- Passion for promoting an ideology: Large number of people are passionate about a specific organization, ideology, person, or philosophy and desire to propagate it through any means possible.
- Fun: Satirical websites create amusing information that is mistaken for true news for the sake of enjoyment and fun. As these are the least serious motivators, with little negative consequences because intentions are rarely erroneous.
- Increase customer base: Online news media is striving to gain viewership and expand their consumer- base in the era of Internet-based journalism. They are releasing stories of dubious quality and content in order to attract visitors to their websites and platforms.

- Rush to cover the latest news: Journalists frequently publish articles without fact-checking in a race to be the first to cover a story achieving millions of views. Today's web journalism, truth and truthfulness have become liabilities, with a goal of "publish first, fix later."
- Generate advertising revenues: During the 2016 US presidential elections, fake news creators profited from advertising engines such as AppNexus, Facebook Ads, and Google AdSense. Earning money by spreading false advertising news is a powerful motivator that a whole industry of practitioners has sprung up around it.
- Technological Reasons: Algorithms are designed to recommend items based on their popularity rather than their veracity. Some algorithmic flow models for skewed information circulation include echo chambers and search engine filter bubbles, using art to promote the spread of deception, as fake news is designed to attract greater user attention.
- Manipulate public opinion: In a consumer-driven economy, public opinion on a company, service, product, or person is crucial, as customers determine the fate of stocks, sales, election results, and a variety of other enterprises.

Ramesh Pandita demonstrated the Sources of information pollution.

- Internet became one of the time sources of information and extracting the most accurate, trustworthy, and authentic information from the web is nearly impossible because of the clutter of information.
- Explosion of information happens when the focus is on a quantitative aspect of producing information and the quality aspect is neglected, replicating the same information in different formats,duplicating information and other similar activities, which are the least desirable and lead to information pollution.
- Unorganized information, Internet has become one of the most important sources of unstructured information, malting it one of the most time-consuming databanks.
- World Wide Web (WWW):
  - Source of unstructured data.
  - Unsolicited information provider.
  - Unrestricted information generation and distribution.
  - No process in place to check for unnecessary or misleading information.

- Media in general, and electronic media in particular, have neglected their role and obligation by spreading fake news, sensationalizing news stories, and untrustworthy information to thegeneral public.
- Social media introduced us to the realm of true freedom of expression, but has also pushed us closer to the perils of absolute freedom of expression. Spreading blasphemous, insulting, sacrilegious, and communal messages, also content that is unfiltered and uncut on social media, is a source of concern.
- Spam, defined as unsolicited electronic messages from unknown senders, mostly are promotional in nature. Delivery of a lot of unwanted material that doesn't help readers, time- consuming activity, and an abundance of junk mail that consumers regard as a personal attack. By offering modem-day lottery hoaxes and prize money to recipients who are unconcerned about the annoyance caused to others, they are breaking communication rules.
- Mobile Phones, Service providers of business houses flashing instant messages on mobile phones to sell wares creates information pollution in its own right.
- Books and Journals, Journals issued with common titles and high prices make it difficult for the information seeker, and subscriptions, than assisting one in obtaining the relevant piece of information, cause one to become lost in a sea of unorganized information clutter.
- Plagiarism, some may refer to it as "stealing other people's work" or "intellectual theft" without acknowledging other people's work, but the fact is that reproducing other people's information and labeling it as one's own creates confusion about what to count and who is the authoritative source of the information.

Ramesh Pandita elucidated the Targets and Concerns of information pollution.

- In Society, Information conveyed in fields such as health, economics, politics, and religion, among others, is of widespread interest, and providing polluted information in these domains may be suicidal.
- In Business, operations are totally dependent on top management's decision-making capacity; there is always a need to give filtered information, which might transform productivity and growth into a loss.
- Education, one of the most important areas of any country, and providing polluted knowledge among researchers and students might destroy the scientific temper of young brains.

- Making a decision: Due to availability of contaminated information, information overload raises questions about the authenticity and trustworthiness of the information and impedes decision-making abilities of people, institutions, or organizations.
- Ecology: Study results and daily actions have a direct impact on ecosystem in which we live,and so information pollution has an indirect impact on our ecology and environment.
- Health: According to cognitive research, individuals have limit to amount of information they consume maintaining conscious concentration, and once the limit is reached (information intake saturation threshold), it can lead to choice paralysis.

Priyanka Meel, Dinesh Kumar Vishwakarma listed the public and commercial Social media analytics tools and Fact-checking platforms.

Public and commercial social media analytics tools that play a crucial role in providing suggestions and developing mass opinions are the main source of surveillance, analysis and management of floatinginformation on social networks in the public domain which statistically, behaviourally, and semantically analyse the data from different aspects to generate reports.

Talkwalker	Crowdbooster
Hootsuite	Snalytics
Vox Civitas	Whisper
Google analytics	

Depending on the users interests and ideas, they may either pass on the data on the assumption that it is true, reject it on the assumption that it is false or become neutral to the news according to their intelligence and their consciousness of the facts. Popular credibility assessment tools used to verify the authenticity of online content that reduce compromised social media accounts used to spread misinformation, tarnish opponents'reputations or cause monetary losses.

TwitterTrails	TweetCred
Hoaxy	Emergent
CredFinder	Snopes
FactCheck	FluxFlow
RumorLens	TruthOrFiction
COMPA	Fake News Tracker
ClaimBuster	PolitiFact
InVID	REVEAL

2. Methods of False Information Detection

- **Arkaitz Zubiaga, Ahmet Aker, Kalina Bontcheva, Maria Liakata, and Rob Procter developed Rumour classification system** for detection and resolution of rumours in social media despite the difficulties that rumours and misinformation offer for the system's development, breaking down the development process into smaller components can help (Detection, Tracking, Stance, Veracity) and making use of suitable techniques is the progress toward developing an effective rumour classification system that assists people in making decisions towards assessing the veracity of information gathered from social media.

  Twitter has become the go-to data source for the collection and analysis of rumours which are collected through large-scale datasets and emerging rumours via Twitter API to collect posts. Bayesian classifiers are applied in rumour tracking that collects and filters posts discussing the rumour once detected. The stance classification system is designed using Long/Short-Term Memory Networks (LSTMs) for sequential classification, where the stance of each tweet considers the features and labels of the previous tweets, and Support vector machines (SVM) are used in the final veracity classification component that determines the actual truth value of the rumour.

- **Mohamed Jehad Baeth, Mehmet Aktas implemented Provenance use in social media software to develop methodologies for detection of information pollution and violation of copyrights in the social web.**

Service Oriented Architecture based tool that consists of independent working layers to analyze data provenance (metadata that indicates the data's origin, validity, quality, and ownership) and monitoring of its spread among affected users to measure the quality of social data while developing algorithms and methodologies that utilizes this information to detect information pollution and detect violation of copyrights of the users' shared data.

A Web based user interface enables users to feed the tool with data object UR's for analysis and feedback shows extracted information of specific data and defines the scope of monitoring, in addition, retrieves user's identification and the values of provenance attributes, exploring various social networking sites for info. Attribute values acquired from different social networking sites by different information retrieval techniques suitable to platform, Sensors communicate directly with different media network API to detect events that occurs on data and provide unified form of data for the attribute engine to process.

A provenance repository to store collected data, provenance documents stored in repository may be transformed, visualized, or even shared, stores information of other collectible attribute valuesfrom the objects being monitored and visited sites. Metadata is applied to videos and photos, not restricted to social media, to give further validation for the collected attribute values, then, visualization of provenance to understand the data by generating the provenance graph or query to find details about a particular activity, entity or agent in areas including Databases and Semantic Web.

- Elmurngi and Gherbi (2017a, b) conducted a comparative research of supervised machine learning algorithms on movie reviews utilising sentiment analysis and text categorization. Elmurngi & Gherbi (2017b) worked on two different movie reviews datasets Vl.0 and V2.0, Nave Bayes (NB), Support Vector Machine (SVM), K-Nearest Neighbour (KNN-IBK), KStar (K), and Decision Tree (DT-148). Later, Elmurngi &Gherbi, 2018 conducted study on three separate movie review datasets, text classification combined with sentiment analysis shown to be an excellent method of false review identification.
- Castillo, Mendoza, and Poblete (2011) used propagation characteristics for the development of a decision tree inthe DT-J48 algorithm in addition to the User, Message, Content, Topic, and Sentiment features.

- Shah & Zaman, 2011proposed rumour centrality maximum likelihood estimator to reduce the source estimation error and analyse the asymptomatic behaviour of infected nodes in regular trees, general trees, andgeneral graphs for determining the single originating node of a rumour spread where all receivers are known ahead of time.
- Mondal, Pramanik, Bhattacharya , Boral, & Ghosh, 2018) developed a content-based probabilistic model for early detection of unverified tweets in the aftermath of a disaster.
- Fairbanks, Fitch, Knauf, & Briscoe, 2018 worked on bias detection and believability assessment, two independent approaches based on text content and structural are used: logistic regression and random forest using TF-IDF and doc2vec embeddings.

3. Preventive Measures

Some of the measures explained by **Ramesh Pandita**:

- Information Literacy (IL) leads to
  - Developing Professional Competence
  - Valuing Information
  - Skill Development
  - Bridge Knowledge Gap
  - Flow of lnformation from Information Haves to Information Have-nots
  - Information Economy
- Censorship: Censoring info especially sensitive areas like national security, cultural sensitivity & social values appear to be getting eroded by the free flow of unregulated information on the web that is sacrilegious, communal, insulting, or inflammatory in some way can cause societal upheavaland should not be considered a limitation on freedom of expression.
- Laws for Framing and Enforcing Information: Information in the form of a voice conversation, video call, photograph, cryptographic writing, or any other physical arrangement; to regulate information production and its communication behavior, put in place regulations, which if not adhered to, should be enforced to avoid information pollution or its contamination.
- Information Ethics: People may be sensitized about information ethics in the areas of Verbal Communication, Handling and use of Information Technology, Avoiding Plagiarism, Copyright Violation, Intellectual Property Rights, Censorship, Information Hacking, to various other information abuses and the ethical use of ICT.

Other Proposed Solutions:

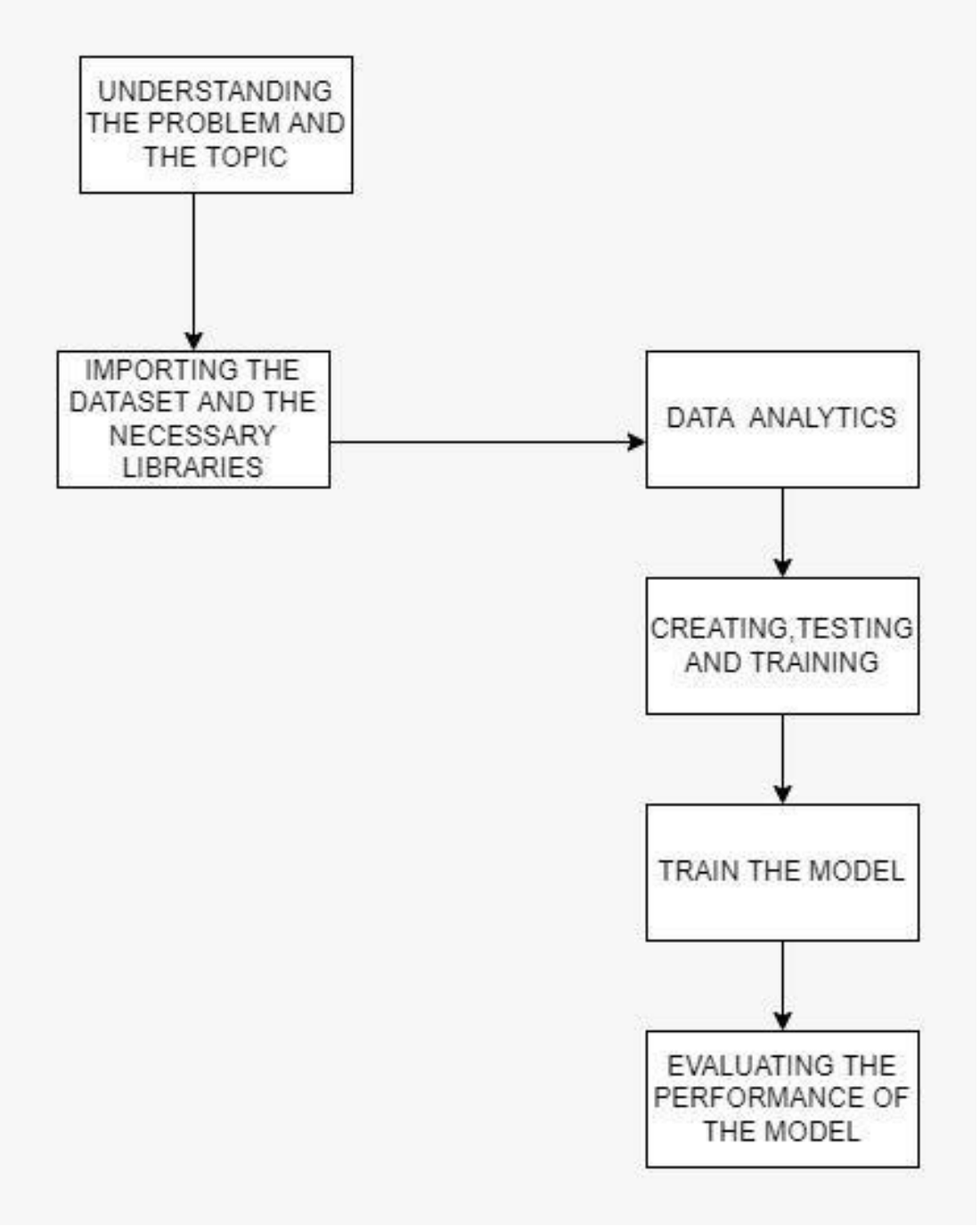
- Sustainable production and consumption of information is a key to proportionate growth and development, moreover, replacing technologies which frequently trouble, with lesser troubling technologies as a measure incurbing information pollution.
- Self-restraint in checking emails both at personal and professional level, in addition, conscious approach towards writing precisely emphasizing on the time and stress management among the personnel of an organization will help in tackling the information pollution toa great deal, and keeping customers informed through web portals.

PROBLEM IDENTIFICATION & OBJECTIVES

- The main objective is to show that the topics covered in the debates by mainstream media news channels were supporting the ruling government, and it is ideology directly or indirectly which are irrelevant to our target audience - students which allow them to check for available relevant information from credible sources and avoid dilatory on superfluous sources making it more reliable and safer for decision making as well as for knowledge sharing.
- Implementing ML models namely decision tree classifier, random forest classifier, logistic regression, and support vector machine to determine if the debates across major Indian news channels shared on social media are intended to have quality content for students and classify the data as polluted or not polluted, or not sure.
- The secondary objective is to compare the performance metrics and best fit model for the problem statement. Exploratory data analysis tasks will also be performed for data set verification.

SYSTEM METHODOLOGY

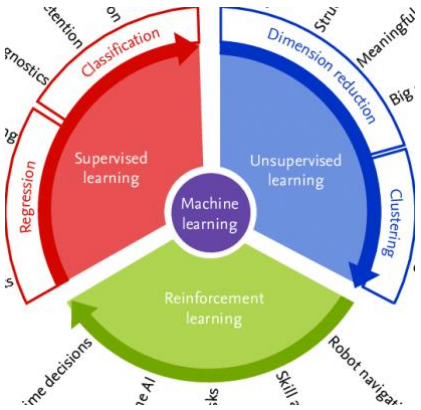
Flow Chart



OVERVIEW OF TECHNOLOGIES

Machine Learning

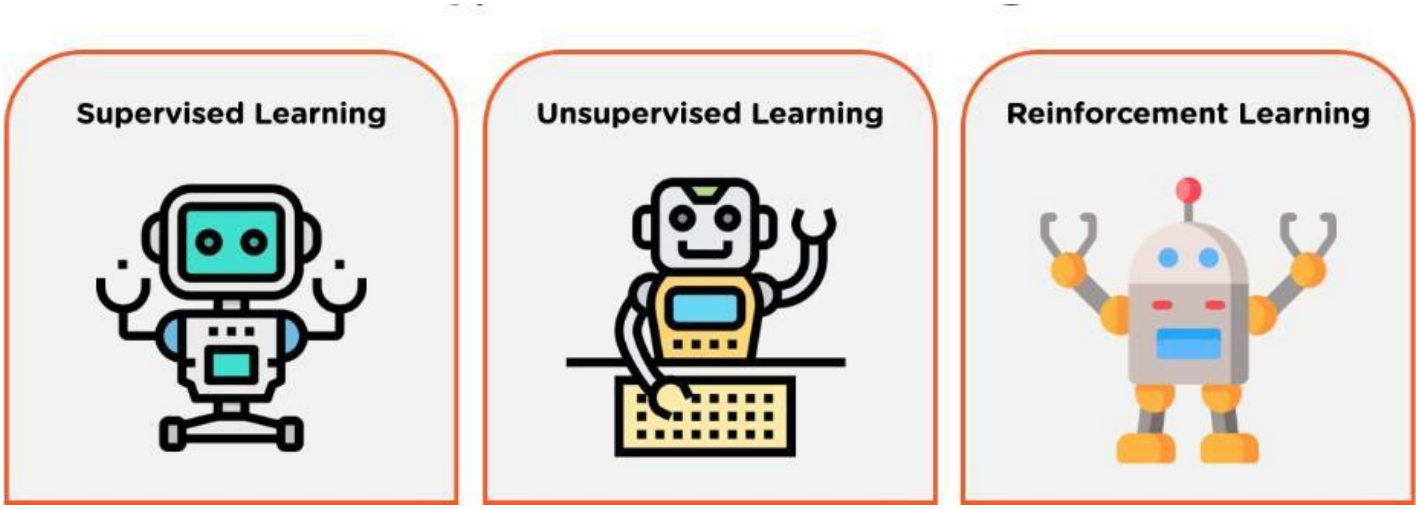
The use and development of computer systems capable of learning and adaptation without explicit instructions, using algorithms and statistical models to analyse and infer trends in the data.



Types of Machine Learning:

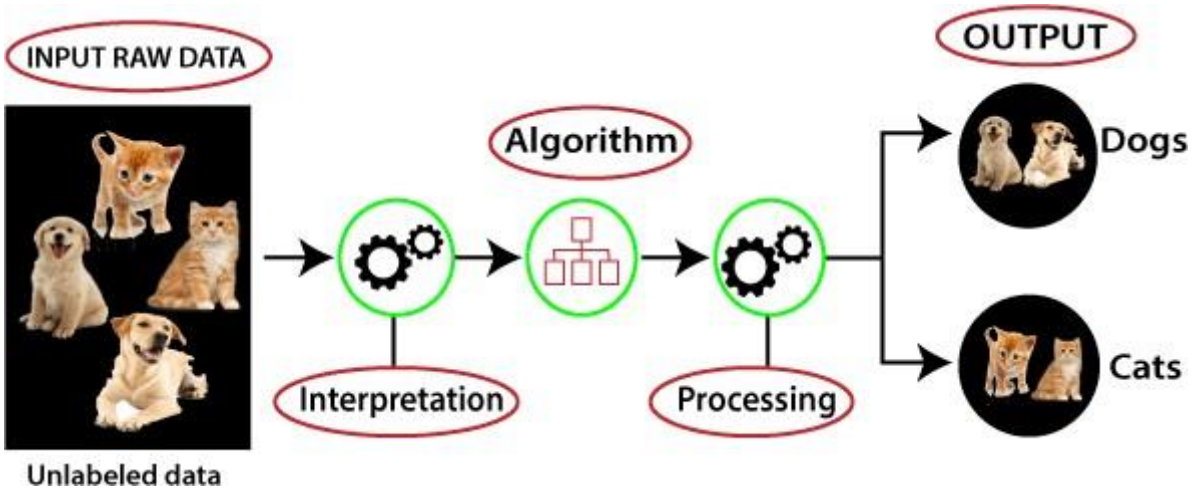
Based upon methods and mode of learning, machine learning is divided primarily into four types, which are:

- Supervised Machine Learning
- Unsupervised Machine Learning
- Reinforcement Learning



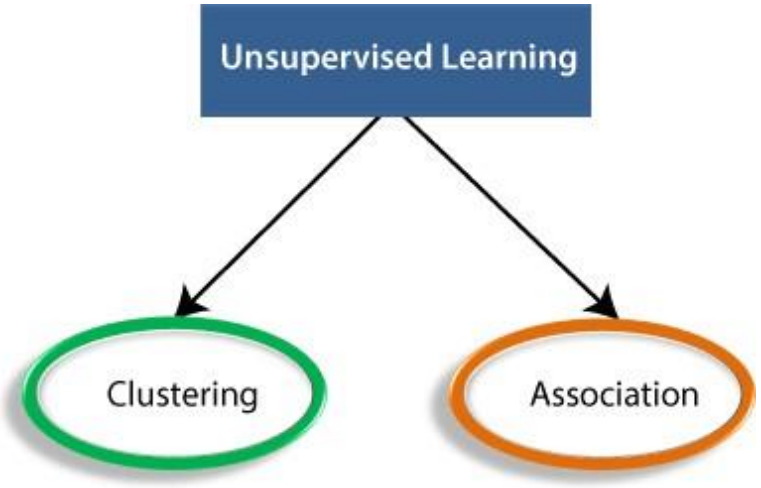
Unsupervised Machine Learning:

It is the type of machine learning in which machines are not supervised using training dataset i.e trained using unlabeled dataset, instead, models itself find the hidden patterns and insights from the given data.



Unsupervised machine learning can be classified into two types of problems, which are given below:

- Clustering
- Association



Clustering:

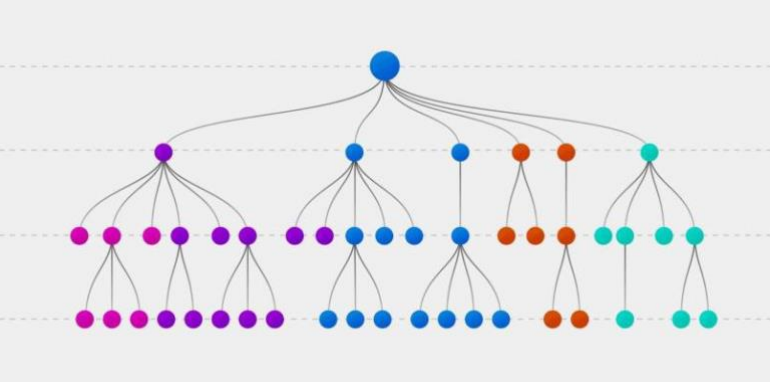
Group objects into clusters, so that objects with the most similarities stay in one group and have less or no similarities to objects in another group. It finds the commonalities between the data objects and categorizes based upon the presence and absence of those commonalities. Unsupervised clustering is a classification task.

Association:

Determines a set of items that together occur in the dataset. Association rule used for finding relationships between variables in the large database makes marketing strategy more effective. Market Basket Analysis is one of the best examples of Association rule.

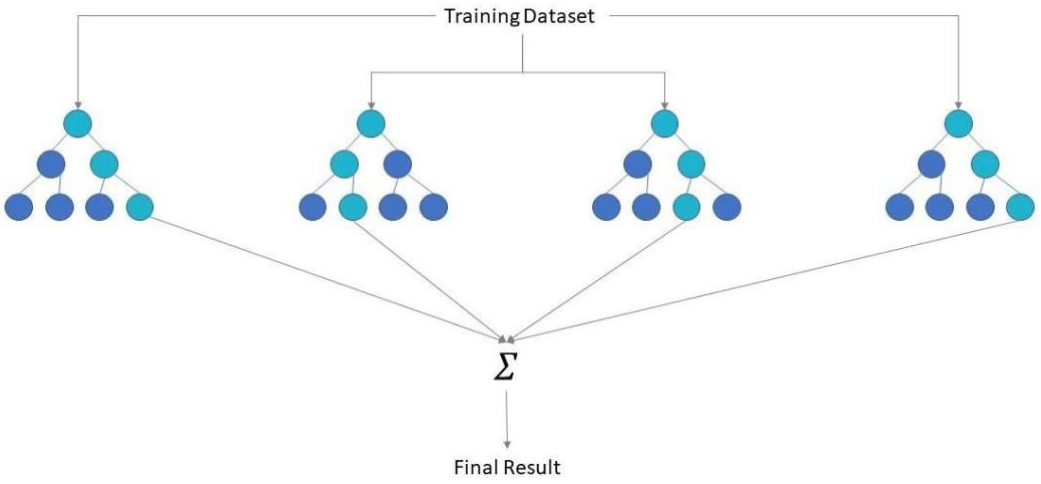
Decision Tree:

Decision Trees in which the data is continuously split according to a certain parameter, and can be explained by two entities, namely decision nodes and leaves. In unsupervised learning, any clustering algorithm that is sufficient for our data can be used to assume the resulting cluster are classes, and then further train a decision tree on the clusters.



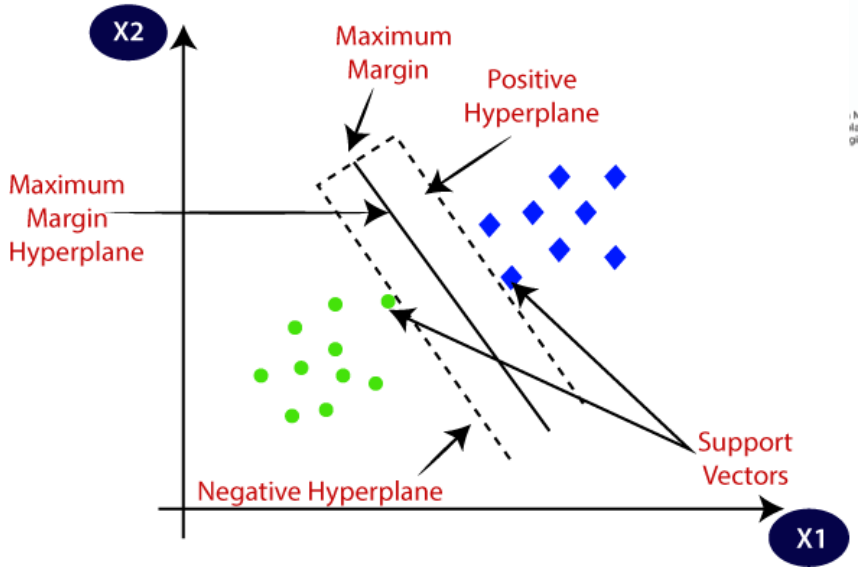
Random Forest:

Random Forest, a classifier that contains a number of decision trees on different subsets of the given dataset and uses the mean to improve the predictive accuracy of that dataset. Instead of relying on a decision tree, the random forest takes the prediction of each tree and, based on the majority votes of predictions, it predicts the final output.

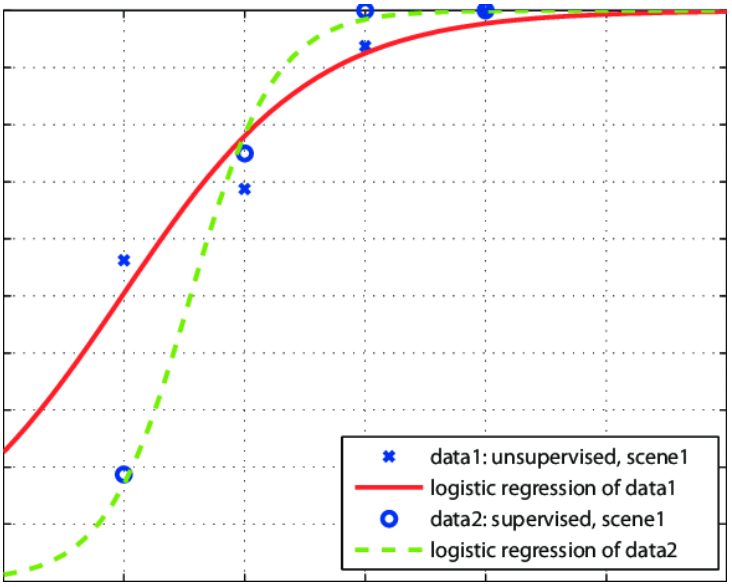


Support vector machine:

The goal of the SVM algorithm is to create the best line or decision boundary (hyperplane) that can segregate n-dimensional space into classes such that we can put the new data point in the correct category in the forthcoming.



Logistic Regression: It is an another technique taken by machine learning from the field of statistics, an go-to method for binary classification problems. Logistic Regression is applied across multiple areas and fields in the real world.


data1: unsupervised, scene1
logistic regression of data1
data2: supervised, scene1
logistic regression of data2

In the ancient days, individuals used to perform Machine Learning undertakings by manually coding all the algorithms and mathematical and statistical formulas. This made the process time consuming, dreary and inefficient. In any case, in the modern days, it has become very much easy and efficient compared to the olden days by various python libraries, frameworks, and modules. Today, Python is quite possibly the most famous programming language for this assignment and it has replaced many dialects in the industry, one of the explanations is its immense collection of libraries. Python libraries used in Machine Learning:

- NumPy
- Scipy
- Scikit-learn
- Theano
- TensorFlow
- Keras
- PyTorch
- Pandas
- Matplotlib

NumPy:

NumPy, a very popular python library for large multi-dimensional array and matrix processing, with the assistance of a large assortment of high-level mathematical functions. It is very valuable for fundamental logical computations in Machine Learning. It is particularly valuable for linear algebra, Fourier transform, and random number capabilities. High-end libraries like TensorFlow use NumPy internally for manipulation of Tensors.

Pandas:

Pandas, also a popular Python library for data analysis and is not directly related to Machine Learning. As we realize that the dataset should be prepared before training. In this case, Pandas was developed specifically for data extraction and preparation, in addition, provides high-level data structures anda wide variety of instruments for data analysis, also provides much inbuilt methods for grouping, combining and filtering data.

Matplotlib:

Matplotlib, a popular Python library for data visualization. Like Pandas, it is not directly related to Machine Learning. It particularly proves to be useful when a programmer wants to visualize the patterns in the data and is a 2D plotting library for creating 2D graphs or plots. A module named pyplot makes it easy for programmers for plotting as it provides features to control line styles, textual style properties, formatting axes, etc that provides different graphs and plots for data visualization.

Scikit-learn:

Scikit-learn, one of the most popular ML libraries for classical ML algorithms. It is based on top of two basic Python libraries, viz., NumPy and SciPy. Scikit-learn upholds most of the supervised and unsupervised learning algorithms. Scikit-learn can also be used for data-mining and data-analysis, which makes it a great instrument to start out with ML.

IMPLEMENTATION

Coding & Testing

The implementation of the project is divided into tasks namely:

Task 1: Understand the Problem Statement

Task 2: Import Libraries and data-sets Task

Task 3: Perform Exploratory Data Analysis

Task 4: Splitting into Training and Testing Datasets

Task 5: Train, Test and Evaluate four ML models (i.e., DecisionTree, Random Forest, Support Vector Machine, & Logistic Regression.)

Task 6: Classified Output of all ML models.

Task 1: Understand the problem statement

- To show that the topics covered in the debates by mainstream media news channels were supporting the ruling government, and it is ideology directly or indirectly which are irrelevant to our target audience - students which allow them to check for available relevant information from credible sources and avoid dilatory on superfluous sources making it more reliable, safer for decision making, and for knowledge sharing.
- Implementing ML models namely decision tree classifier, random forest classifier, logistic regression, and support vector machine to determine if the debates across major Indian news channels shared on social media are intended to have quality content for students and classify the data as polluted or not polluted, or not sure.

Task 2: Import Libraries and data sets

A perfect data taken from the paper publication “Use of Fake News and Social Media by Main Stream News Channels of India” written by Mohammed Hazim Alkawaz and Sayeed Ahsan Khan. The data set taken is a csv document.

CSV (Comma separated values) is a text document that has a particular configuration which permits data to be saved in a table organized arrangement as rows and columns. We will be taking this dataset into the jupyter note pad to play out the prediction errands on the data values present over the data set.

The dataset comprises of 2 columns namely topic, number of debates and 19 rows of debates across mainstream media news channels data.

```
In [2]: # read csv file
        df1 = pandas.read_csv("D:/SEM-8/PROJECT/news_change.csv",encoding='cp1252')
        df2= pandas.read_csv("D:/SEM-8/PROJECT/news_changeq.csv",encoding='cp1252')
        df1['topic']=df2['ï»¿topic']
```

Dataset: news_change.csv

Sample data of dataset:

	A	B
1	topic	number_of_debates
2	Attacking Opposition (including Article 370, Muslims, Political parties)	96
3	Attacking Pakistan	91
4	Praising Current Govt. & its Affiliates	45
5	Ram Mandir & Babri Masjid	19
6	Bihar Floods	4
7	PMC Bank Scam	4
8	Chandrayaan-2 (Moon Mission)	2
9	Rape Case	1
10	Economy	0
11	Unemployment	0
12	Education	0
13	Healthcare	0
14	Public Infrastructure	0
15	Farmer’s Distress	0
16	Poverty & Malnutrition	0
17	Women’s Safety	0
18	Environmental protection & pollution	0
19	Mob Lynching	0
20	Question governments decision	0
21		

Libraries imported in the jupyter notebook for this forecast are numpy, pandas, matplotlib, scikitlearn, seaborn. The reason these libraries are used:

PyDotPlus: PyDotPlus, an improved version of the old pydot project, provides a Python Interface to Graphviz's Dot language.

Pandas: Pandas, an open-source, BSD-licensed Python library providing high-performance, easy-to-use data structures and data analysis tools for the Python programming language.

Matplotlib: Matplotlib, one of the most popular Python packages used for data visualization and might be used in Python or IPython shells, Jupyter notebook, also web application servers.

ScikitLearn: Scikit-learn, free machine learning library for Python and features various algorithms like support vector machine, random forest, and k-neighbors, also supports Python numerical and scientific libraries such as NumPy and SciPy.

Statistics: Statistics, a built-in module that can be used to calculate mathematical statistics of numeric data with its functions.

```
In [1]: import pandas
        from sklearn import tree
        import pydotplus
        from pydotplus import graph_from_dot_data
        from sklearn.tree import DecisionTreeClassifier
        from sklearn.model_selection import train_test_split
        import matplotlib.pyplot as plt
        import matplotlib.image as pltimg
        import statistics
        from sklearn.metrics import classification_report
        import seaborn as sns
```

Task 3: Perform Exploratory Data Analysis

Exploratory Data Analysis is an approach to analyse data sets to summarize their main characteristics, often with the help of statistical graphs and other data visualisation methods. A statistical model may and may not be used, but the EDA is mainly to see what the data can tell us beyond the formal modelling, or hypothesis testing task.

The head() function makes it possible to obtain the first n rows. This function returns the first n rows of the object based on position. It is useful to rapidly test if your object contains the right type of data. Number of rows to select.

```
In [3]: df1.head()
```

Out[3]:

	topic	number_of_debates
0	1	96
1	2	91
2	3	45
3	4	19
4	5	4

Checking the null values in the dataset.

```
In [4]: # checking the null values
        df1.isnull().sum()
```

```
Out[4]: topic              0
        number_of_debates  0
        dtype: int64
```

Check the data frame information.

```
In [5]: # check the dataframe information
        df1.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 19 entries, 0 to 18
Data columns (total 2 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   topic              19 non-null     int64
 1   number_of_debates  19 non-null     int64
dtypes: int64(2)
memory usage: 432.0 bytes
```

Statistical summary of the dataframe.

```
In [6]:  # statistical summary of dataframe
         df1.describe()
```

Out[6]:

	topic	number_of_debates
count	19.000000	19.000000
mean	10.000000	13.789474
std	5.627314	30.099057
min	1.000000	0.000000
25%	5.500000	0.000000
50%	10.000000	0.000000
75%	14.500000	4.000000
max	19.000000	96.000000

Grouping by number_of_debates

```
In [7]:  # Grouping by number_of_debates
         df_debate=df1.groupby(by='number_of_debates').mean()
         df_debate
```

Out[7]:

	topic
number_of_debates	
0	14.0
1	8.0
2	7.0
4	5.5
19	4.0
45	3.0
91	2.0
96	1.0

Task 4: Splitting into Training and Testing Datasets

```
In [8]: df1.columns
```

```
Out[8]: Index(['topic', 'number_of_debates'], dtype='object')
```

```
In [9]: X = df1['topic']
        y = df1['number_of_debates']
        X=X.values.reshape(-1,1)
        print(X)
```

```
[[ 1]
 [ 2]
 [ 3]
 [ 4]
 [ 5]
 [ 6]
 [ 7]
 [ 8]
 [ 9]
 [10]
 [11]
 [12]
 [13]
 [14]
 [15]
 [16]
 [17]
 [18]
 [19]]
```

```
In [10]: print(y)
```

```
0     96
1     91
2     45
3     19
4      4
5      4
6      2
7      1
8      0
9      0
10     0
11     0
12     0
13     0
14     0
15     0
16     0
17     0
18     0
Name: number_of_debates, dtype: int64
```

In [11]:
```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.3, random_state = 44)
X_train
```

Out[11]:
```
array([[ 7],
       [10],
       [ 8],
       [ 9],
       [11],
       [15],
       [ 5],
       [ 1],
       [19],
       [12],
       [14],
       [18],
       [ 4]], dtype=int64)
```

Task 5: Train, Test and Evaluate five ML models (i.e, Decision Tree, Support Vector Machine, Random Forest, & Logistic Regression.

Decision Tree:

In [17]:
```
from sklearn.tree import DecisionTreeClassifier

model = DecisionTreeClassifier()
model.fit(X_train, y_train)
y_predict = model.predict(X_test)

y_predict = model.predict(X)
```

In [18]:
```
m=statistics.mean(y_predict)
l=[]
for i in y_predict:
    if i==0:
        l.append("NP")
    elif i>0 and i<=m:
        l.append("NS")
    elif i>m:
        l.append("P")
```

Support Vector Machine:

In [23]:
```
from sklearn.svm import SVC

svc_model = SVC(C= .1, kernel='linear', gamma= 1)
svc_model.fit(X_train, y_train)
```

Out[23]:
```
SVC(C=0.1, gamma=1, kernel='linear')
```

In [24]:
```
prediction = svc_model .predict(X)

m=statistics.mean(prediction)
l=[]
for i in prediction:
    if i==0:
        l.append("NP")
    elif i>0 and i<=m:
        l.append("NS")
    elif i>m:
        l.append("P")
```

Random Forest:

```
In [34]: from sklearn.ensemble import RandomForestClassifier
         classifier= RandomForestClassifier(n_estimators= 10, criterion="entropy")
         classifier.fit(X_train, y_train)

Out[34]: RandomForestClassifier(criterion='entropy', n_estimators=10)
```

```
In [36]: y_pred1 = classifier.predict(X)

         m=statistics.mean(y_pred1)
         l=[]
         for i in y_pred1:
             if i==0:
                 l.append("NP")
             elif i>0 and i<=m:
                 l.append("NS")
             elif i>m:
                 l.append("P")
```

Logistic Regression:

```
In [43]: #Fitting Logistic Regression to the training set
         from sklearn.linear_model import LogisticRegression
         classifier= LogisticRegression(random_state=0)
         classifier.fit(X_train, y_train)

Out[43]: LogisticRegression(random_state=0)
```

```
In [44]: y_predt = classifier.predict(X)

         m=statistics.mean(y_predt)
         l=[]
         for i in y_predt:
             if i==0:
                 l.append("NP")
             elif i>0 and i<=m:
                 l.append("NS")
             elif i>m:
                 l.append("P")
```

Task 6: Classified Output of all ML models.

Decision Tree:

```
In [19]: print(y_predict)
         print(l)
```

```
[96 96 19 19  4  4  2  1  0  0  0  0  0  0  0  0  0  0]
['P', 'P', 'P', 'P', 'NS', 'NS', 'NS', 'NS', 'NP', 'NP', 'NP', 'NP', 'NP', 'NP', 'NP', 'NP', 'NP', 'NP']
```

```
In [20]: df1['results']=l
         print(df1)
```

```
    topic  number_of_debates results
0       1                 96       P
1       2                 91       P
2       3                 45       P
3       4                 19       P
4       5                  4      NS
5       6                  4      NS
6       7                  2      NS
7       8                  1      NS
8       9                  0      NP
9      10                  0      NP
10     11                  0      NP
11     12                  0      NP
12     13                  0      NP
13     14                  0      NP
14     15                  0      NP
15     16                  0      NP
16     17                  0      NP
17     18                  0      NP
18     19                  0      NP
```

Support Vector Machine:

```
In [53]: print(prediction)
         print(l)
```

```
[96 96 19 19  4  4  0  0  0  0  0  0  0  0  0  0  0  0]
['P', 'P', 'P', 'P', 'NS', 'NS', 'NP', 'NP', 'NP', 'NP', 'NP', 'NP', 'NP', 'NP', 'NP', 'NP', 'NP', 'NP']
```

```
In [54]: df1['results']=l
         print(df1)
```

```
    topic  number_of_debates results
0       1                 96       P
1       2                 91       P
2       3                 45       P
3       4                 19       P
4       5                  4      NS
5       6                  4      NS
6       7                  2      NP
7       8                  1      NP
8       9                  0      NP
9      10                  0      NP
10     11                  0      NP
11     12                  0      NP
12     13                  0      NP
13     14                  0      NP
14     15                  0      NP
15     16                  0      NP
16     17                  0      NP
17     18                  0      NP
18     19                  0      NP
```

Random Forest:

In [59]:
```
print(y_pred1)
print(l)
```

```
[96 96 19 19  4  4  2  1  0  0  0  0  0  0  0  0  0  0]
['P', 'P', 'P', 'P', 'NS', 'NS', 'NS', 'NS', 'NP', 'NP', 'NP', 'NP', 'NP', 'NP', 'NP', 'NP', 'NP', 'NP']
```

In [60]:
```
df1['results']=l
print(df1)
```

```
    topic  number_of_debates results
0       1                 96       P
1       2                 91       P
2       3                 45       P
3       4                 19       P
4       5                  4      NS
5       6                  4      NS
6       7                  2      NS
7       8                  1      NS
8       9                  0      NP
9      10                  0      NP
10     11                  0      NP
11     12                  0      NP
12     13                  0      NP
13     14                  0      NP
14     15                  0      NP
15     16                  0      NP
16     17                  0      NP
17     18                  0      NP
18     19                  0      NP
```

Logistic Regression:

In [65]:
```
print(y_predt)
print(l)
```

```
[96 96 96 19  4  4  2  0  0  0  0  0  0  0  0  0  0  0]
['P', 'P', 'P', 'P', 'NS', 'NS', 'NS', 'NP', 'NP', 'NP', 'NP', 'NP', 'NP', 'NP', 'NP', 'NP', 'NP', 'NP']
```

In [66]:
```
df1['results']=l
print(df1)
```

```
    topic  number_of_debates results
0       1                 96       P
1       2                 91       P
2       3                 45       P
3       4                 19       P
4       5                  4      NS
5       6                  4      NS
6       7                  2      NS
7       8                  1      NP
8       9                  0      NP
9      10                  0      NP
10     11                  0      NP
11     12                  0      NP
12     13                  0      NP
13     14                  0      NP
14     15                  0      NP
15     16                  0      NP
16     17                  0      NP
17     18                  0      NP
18     19                  0      NP
```

RESULTS & DISCUSSIONS

- Classified into polluted, not polluted, or not sure based upon debates across major Indian news channels shared on social media. The models used are Decision Tree, Support Vector Machine, Random Forest, & Logistic Regression. Decision Tree and Random Forest has the highest classification rate due to more accuracy of 89%.
- It is worth mentioning that all tests will be done using Python.
- Topics listed below were not discussed at all:
    1. The economy which is in free fall from Q1, 2019
    2. Healthcare, education, public infrastructure which is in bad shape
    3. Farmer's distress and suicide
    4. Unemployment which is at 45 years high
    5. Poverty and malnutrition – India ranks 102 out of 117 Global Hunger index
    6. Pollution in India which is at an unbearable level
    7. Mob-lynching and attack against minorities in the name of cows and religion
    8. Women's safety in India
- And the ones which were discussed below account to 3.46% of debates:
    1. Rape case involving the current ruling government leader
    2. Bihar Flood
    3. A billion-dollar PMC bank scam
- 95% of the debates shared on social media by the mainstream news channels of India were either pro-government, or attacking the opposition, minorities, especially Muslims, Pakistan, Hindu vs. Muslims, Temple vs. Mosque and most, if not all, were either biased, slanted and misrepresented the data and the facts. The remaining less than 5% of the debates covered, flood in Bihar, PMC bank scam, rape case, and moon mission of India were biased to some extent.

Performance Metrics

Decision Tree:

In [21]:
```
print(classification_report(y, y_predict))
```

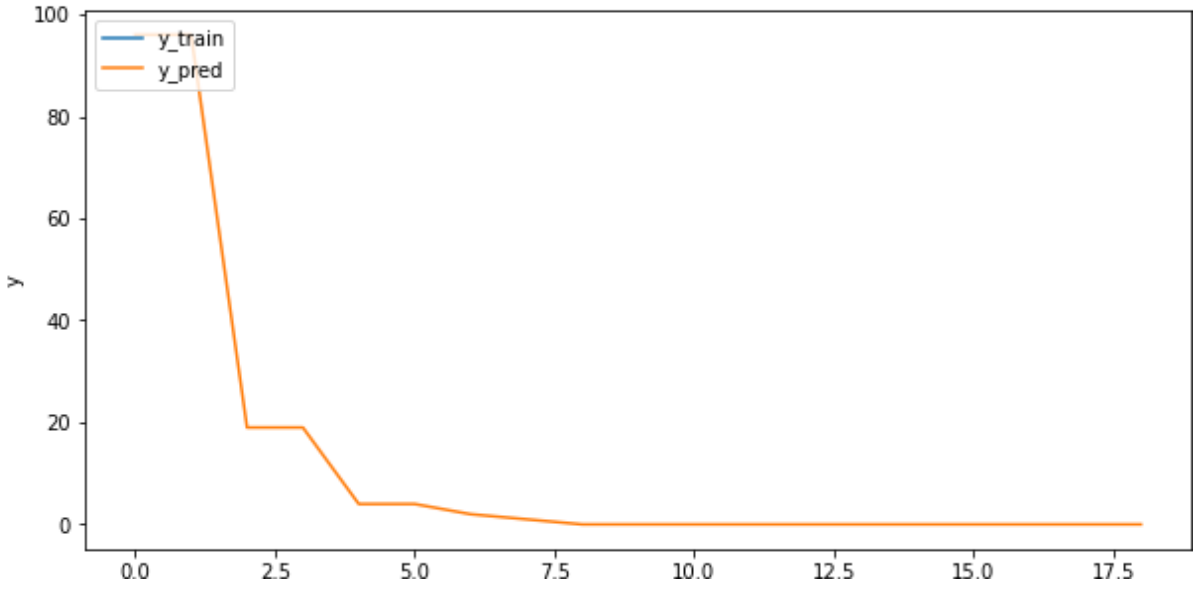
```
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        11
           1       1.00      1.00      1.00         1
           2       1.00      1.00      1.00         1
           4       1.00      1.00      1.00         2
          19       0.50      1.00      0.67         1
          45       0.00      0.00      0.00         1
          91       0.00      0.00      0.00         1
          96       0.50      1.00      0.67         1

    accuracy                           0.89        19
   macro avg       0.62      0.75      0.67        19
weighted avg       0.84      0.89      0.86        19
```

In [22]:
```
pred = pandas.DataFrame(y_predict, columns = ['y_pred'])
train = pandas.DataFrame(y,columns = ['y'])
final = pandas.concat([train,pred],ignore_index=True,sort=False)
final = final.set_index(df1.index)
plt.figure(figsize=(10,5))
sns.lineplot(x=final.index, y=final['y'])
sns.lineplot(x=final.index, y=final['y_pred'])
plt.legend(['y_train','y_pred','y_test'],
           loc='upper left')
plt.ylabel('y')
```

Out[22]: Text(0, 0.5, 'y')



Support Vector Machine:

In [55]:
```
print(classification_report(y, prediction))
```

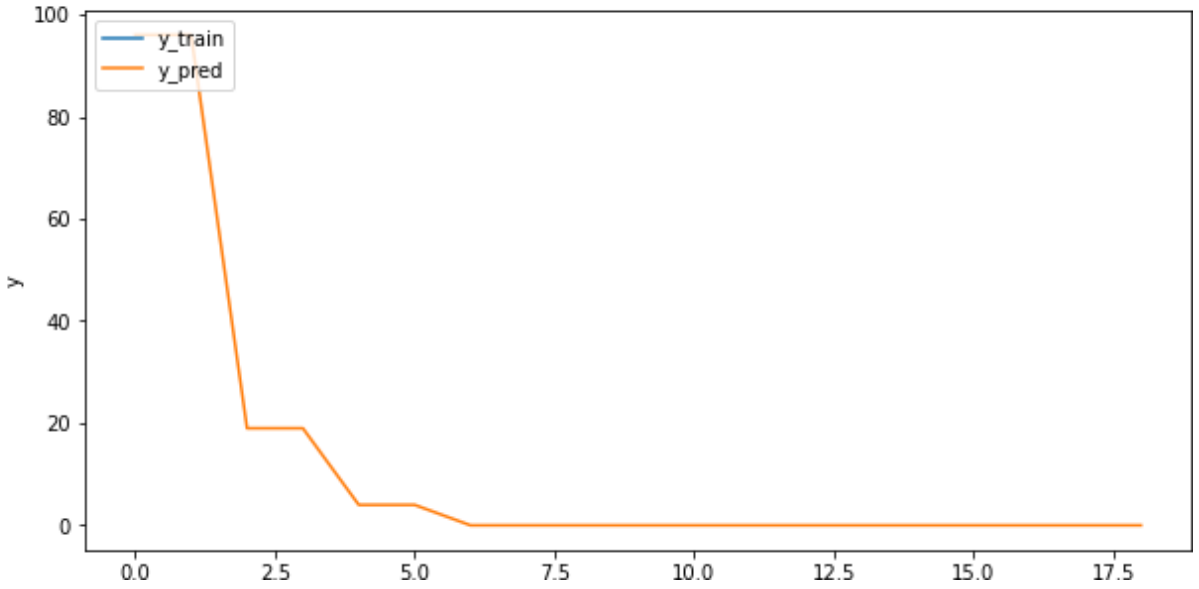
```
              precision    recall  f1-score   support

           0       0.85      1.00      0.92        11
           1       0.00      0.00      0.00         1
           2       0.00      0.00      0.00         1
           4       1.00      1.00      1.00         2
          19       0.50      1.00      0.67         1
          45       0.00      0.00      0.00         1
          91       0.00      0.00      0.00         1
          96       0.50      1.00      0.67         1

    accuracy                           0.79        19
   macro avg       0.36      0.50      0.41        19
weighted avg       0.65      0.79      0.71        19
```

In [56]:
```
pred = pandas.DataFrame(prediction, columns = ['y_pred'])
train = pandas.DataFrame(y,columns = ['y'])
final = pandas.concat([train,pred],ignore_index=True,sort=False)
final = final.set_index(df1.index)
plt.figure(figsize=(10,5))
sns.lineplot(x=final.index, y=final['y'])
sns.lineplot(x=final.index, y=final['y_pred'])
plt.legend(['y_train','y_pred','y_test'],
           loc='upper left')
plt.ylabel('y')
```

Out[56]: Text(0, 0.5, 'y')



Random Forest:

In [61]:
```
print(classification_report(y, y_pred1))
```

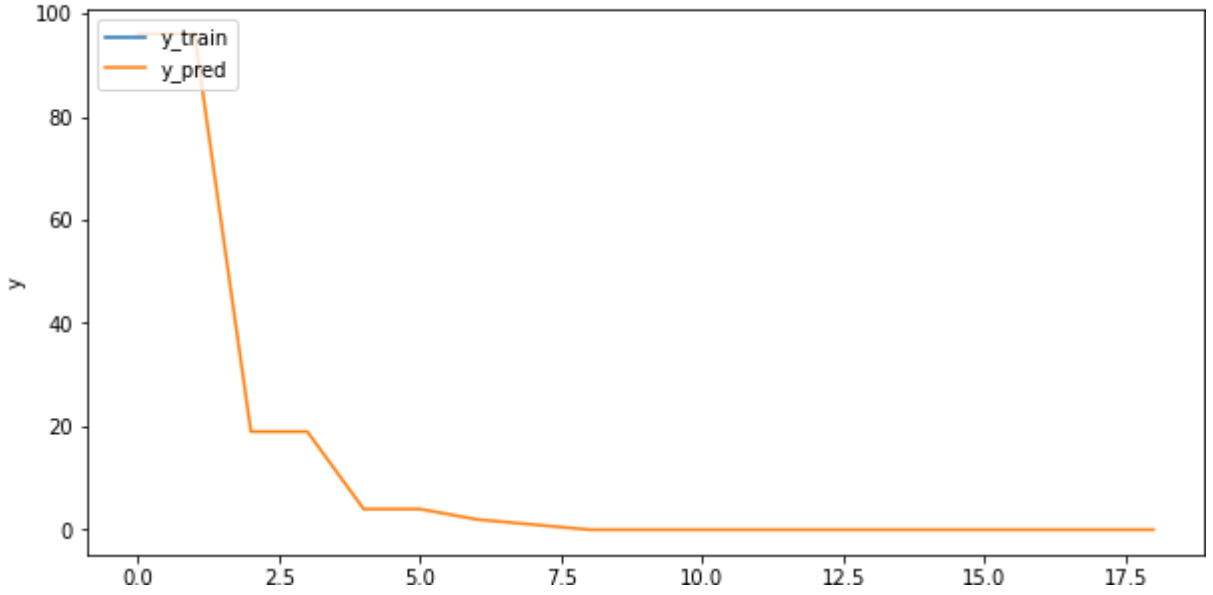
```
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        11
           1       1.00      1.00      1.00         1
           2       1.00      1.00      1.00         1
           4       1.00      1.00      1.00         2
          19       0.50      1.00      0.67         1
          45       0.00      0.00      0.00         1
          91       0.00      0.00      0.00         1
          96       0.50      1.00      0.67         1

    accuracy                           0.89        19
   macro avg       0.62      0.75      0.67        19
weighted avg       0.84      0.89      0.86        19
```

In [62]:
```
pred = pandas.DataFrame(y_pred1, columns = ['y_pred'])
train = pandas.DataFrame(y,columns = ['y'])
final = pandas.concat([train,pred],ignore_index=True,sort=False)
final = final.set_index(df1.index)
plt.figure(figsize=(10,5))
sns.lineplot(x=final.index, y=final['y'])
sns.lineplot(x=final.index, y=final['y_pred'])
plt.legend(['y_train','y_pred','y_test'],
           loc='upper left')
plt.ylabel('y')
```

Out[62]: Text(0, 0.5, 'y')



Logistic Regression:

In [73]:
```
print(classification_report(y, y_predt))
```

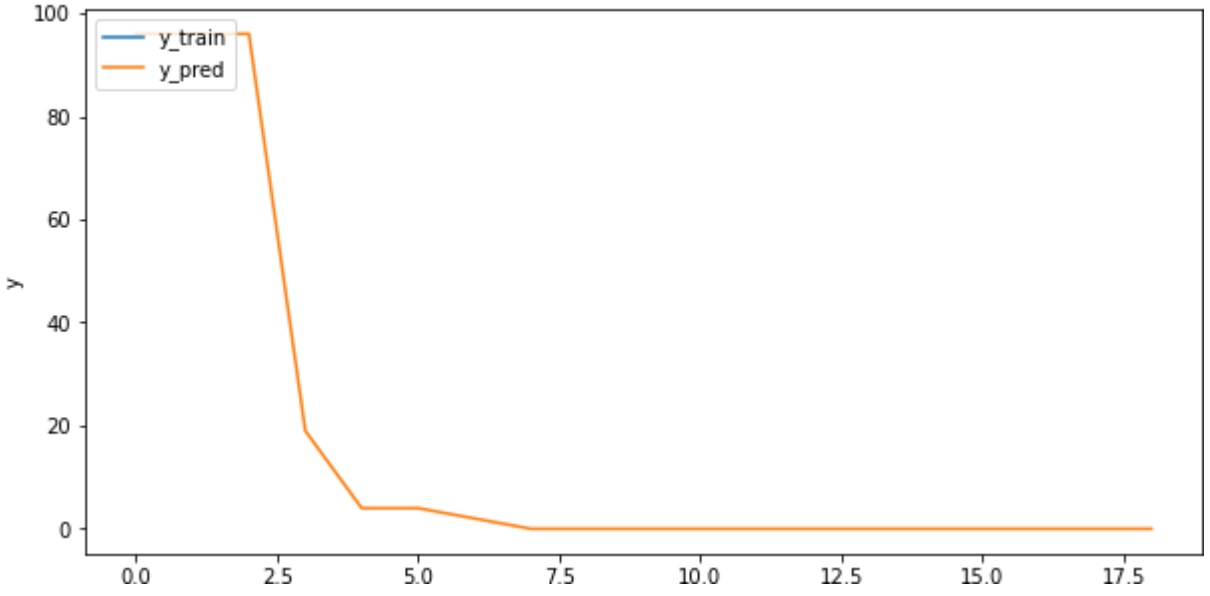
```
              precision    recall  f1-score   support

           0       0.92      1.00      0.96        11
           1       0.00      0.00      0.00         1
           2       1.00      1.00      1.00         1
           4       1.00      1.00      1.00         2
          19       1.00      1.00      1.00         1
          45       0.00      0.00      0.00         1
          91       0.00      0.00      0.00         1
          96       0.33      1.00      0.50         1

    accuracy                           0.84        19
   macro avg       0.53      0.62      0.56        19
weighted avg       0.76      0.84      0.79        19
```

In [74]:
```
pred = pandas.DataFrame(y_predt, columns = ['y_pred'])
train = pandas.DataFrame(y,columns = ['y'])
final = pandas.concat([train,pred],ignore_index=True,sort=False)
final = final.set_index(df1.index)
plt.figure(figsize=(10,5))
sns.lineplot(x=final.index, y=final['y'])
sns.lineplot(x=final.index, y=final['y_pred'])
plt.legend(['y_train','y_pred','y_test'],
           loc='upper left')
plt.ylabel('y')
```

Out[74]: Text(0, 0.5, 'y')



CONCLUSION & FUTURE SCOPE

In this project, ML models like Decision Tree, Random Forest, Support Vector Machine, & Logistic Regression are performed to show that topics covered in the debates by mainstream media news channels were supporting the ruling government, and it is ideology directly or indirectly which are intended to have no quality content to students.

As for future work, the domain of this study is highly pertinent to pragmatic human life, most of the work done to date in literature establishes reasonable theories by drawing cooperation between the problem domain and available methods. Few positive attempts to use the methods everyday have also been done, despite all these attempts, still there are several functionalities to be inbuilt for real time fact-checking by evolving all possible scenarios of information contamination.

Research on the development of false information detection and verification has become increasingly popular, enabling both ordinary users and professional practitioners to gather news and facts in a real- time fashion, however, r e s e a r c h i n t h i s d i r e c t i o n i s still in its infancy and more research is needed to best exploit this context for maximizing the performance of stance classifiers. Research in false information classification has largely depend on the content of social media posts, while further information extracted from user metadata and interactions may be of help to boost the performance of classifiers and help resolve the growing problem of hampering decision-making ability of individuals. Further research in this direction would then enable development of entirely automated rumor classification systems such that it enhances the performance of saving users time to refer credible and reliable resources.

REFERENCES

1. Ramesh Pandita (2014). "Information Pollution, a Mounting Threat: Internet a Major Causality". J.of infosci. theory and practice 2(4): 49-60, 2014
2. Mohamed Jehad Baeth, Mehmet Aktas (2015). "On the Detection of Information Pollution and Violation of Copyrights in the Social Web". DOI: 10.1109/SOCA.2015.27
3. Kai-Yuan Cai, Chao-Yang Zhang (1996). "Towards a research on information pollution". doi:10.1109/ICSMC.1996.561484
4. THE MIT PRESS READER (2019). "The Disturbing Power of Information Pollution".
5. Wikipedia (last edited in 2021). "Information Pollution".
6. Jakob Nielson (2004), Nielsen Norman Group. "Ten Steps for Cleaning Up Information Pollution".
7. Priyanka Meel, Dinesh Kumar Vishwakarma (2019). "Fake news, rumor, information pollution insocial media and web: A contemporary survey of state-of-the-arts, challenges and opportunities".
8. ARKAITZ ZUBIAGA, AHMET AKER, KALINA BONTCHEVA, MARIA LIAKATA and ROBPROCTER (2018). "Detection and Resolution of Rumours in Social Media: A Survey".
9. Mohammed Hazim Alkawaz, Sayeed Ahsan khan (2020). "Use of Fake News and Social Media byMain Stream News Channels of India".
10. Wikipedia (last edited in 2021). "Information Technology Act, 2000".