SUMMER INTERNSHIP

PROJECT REPORT

"DETECTING FAKE NEWS"

SUBMITTED BY:

SARVAGYA BANSAL

Data Science June Batch

What is Data Science?

Data science is the domain of study that deals with vast volumes of data using modern tools and techniques to find unseen patterns, derive meaningful information, and make business decisions. Data science uses complex machine learning algorithms to build predictive models. The data used for analysis can come from many different sources and presented in various formats.

Now that you know what data science is, let's see why data science is essential to today's IT landscape.

What Does a Data Scientist Do ?

Before tackling the data collection and analysis, the data scientist determines the problem by asking the right questions and gaining understanding. The data scientist then determines the correct set of variables and data sets. When the data has been completely rendered, the data scientist interprets the data to find opportunities and solutions.

The data scientists finish the task by preparing the results and insights to share with the appropriate stakeholders and communicating the results.

Applications of Data science

- **Image Recognition**:- Identifying patterns in images and detecting objects in an image is one of the most popular data science applications.

- **Gaming** :- Video and computer games are now being created with the help of data science and that has taken the gaming experience to the next level

- **Healthcare**: - Healthcare companies are using data science to build sophisticated medical instruments to detect and cure diseases.

- **Recommendation Systems**: - Netflix and Amazon give movie and product recommendations based on what you like to watch, purchase, or browse on their platforms.

- **Logistics**: - Data Science is used by logistics companies to optimize routes to ensure faster delivery of products and increase operational efficiency.

- **Fraud Detection: -** Banking and financial institutions use data science and related algorithms to detect fraudulent transactions.

Main Components of Data science

1. Data Exploration

It is the most important step, as this step consumes the most amount of time. Around 70 per cent of the time is spent on data exploration. The main ingredient for data science is data, so when we get data, it is seldom that data is in a correct structured form. There is a lot of noise present in the data. The noise here means a lot of unwanted data that is not required.

2. Modeling

So, by now, our data is prepared and ready to go. This is the second step, where we actually use Machine Learning algorithms. Here we actually fit the data into the model. The selection of a model depends on the type of data we have and the business requirement.

For example, the model selection for recommending an article to a customer will be different than the model required for predicting the number of articles that will be sold on a particular day.

3. Testing the Model

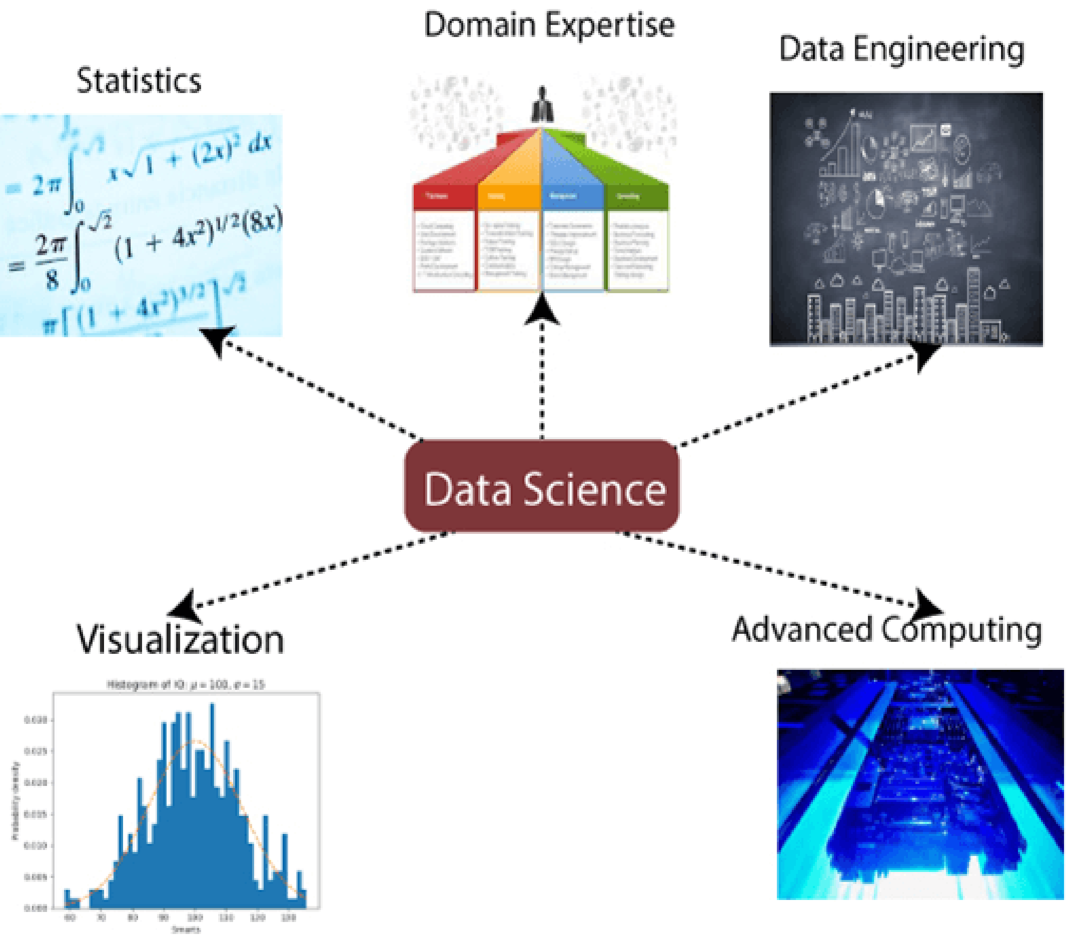
It is the next step and very important concerning the performance of the model. The model is tested with test data to check the model's accuracy and other characteristics and make the required changes in the model to get the desired result.

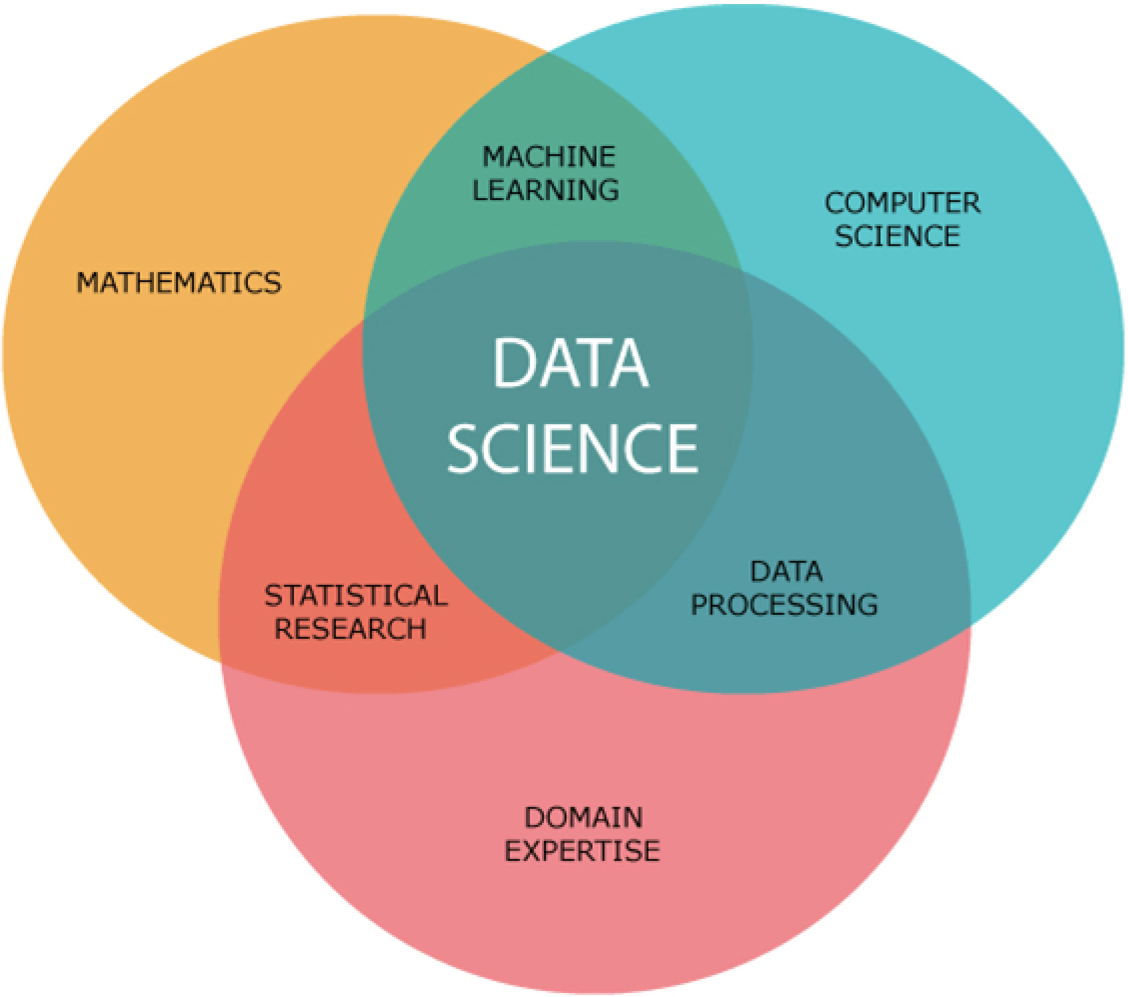
In case we do not get the desired accuracy, we can again go to step 2 (modelling), select a different model, and then repeat the same step 3 and choose the model which gives the best result as per the business requirement.

4. Deploying Models

Once we get the desired result by proper testing as per the business requirements, we finalize the model, which gives us the best result as per testing results and deploys the model in the production environment.

Data Science Component

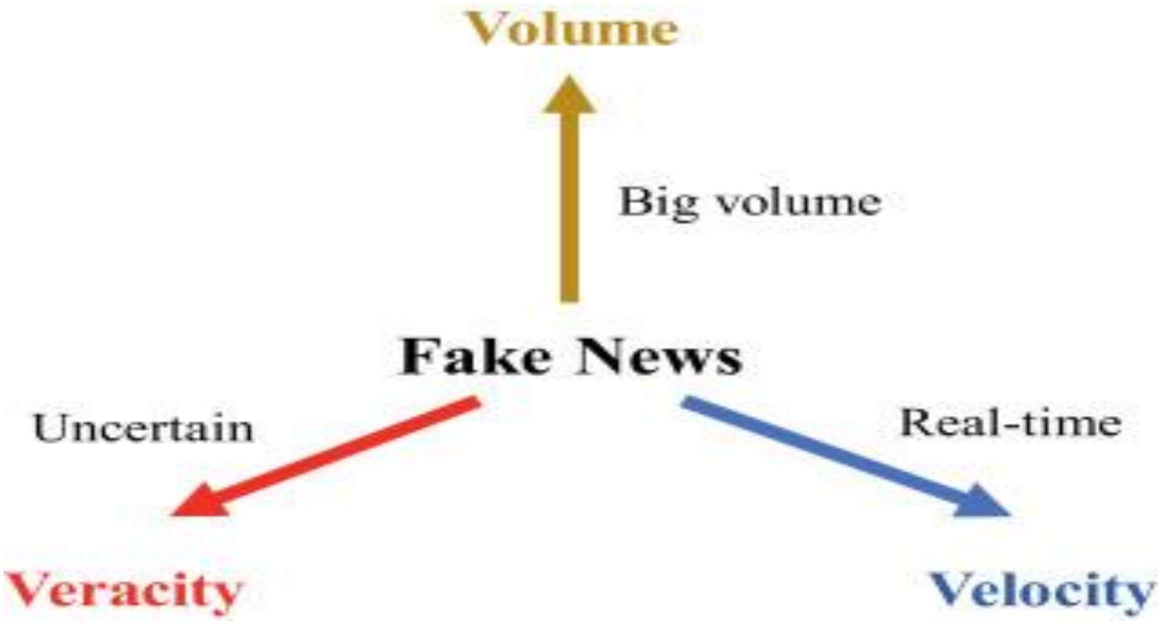



MACHINE LEARNING
MATHEMATICS
COMPUTER SCIENCE
DATA SCIENCE
STATISTICAL RESEARCH
DATA PROCESSING
DOMAIN EXPERTISE

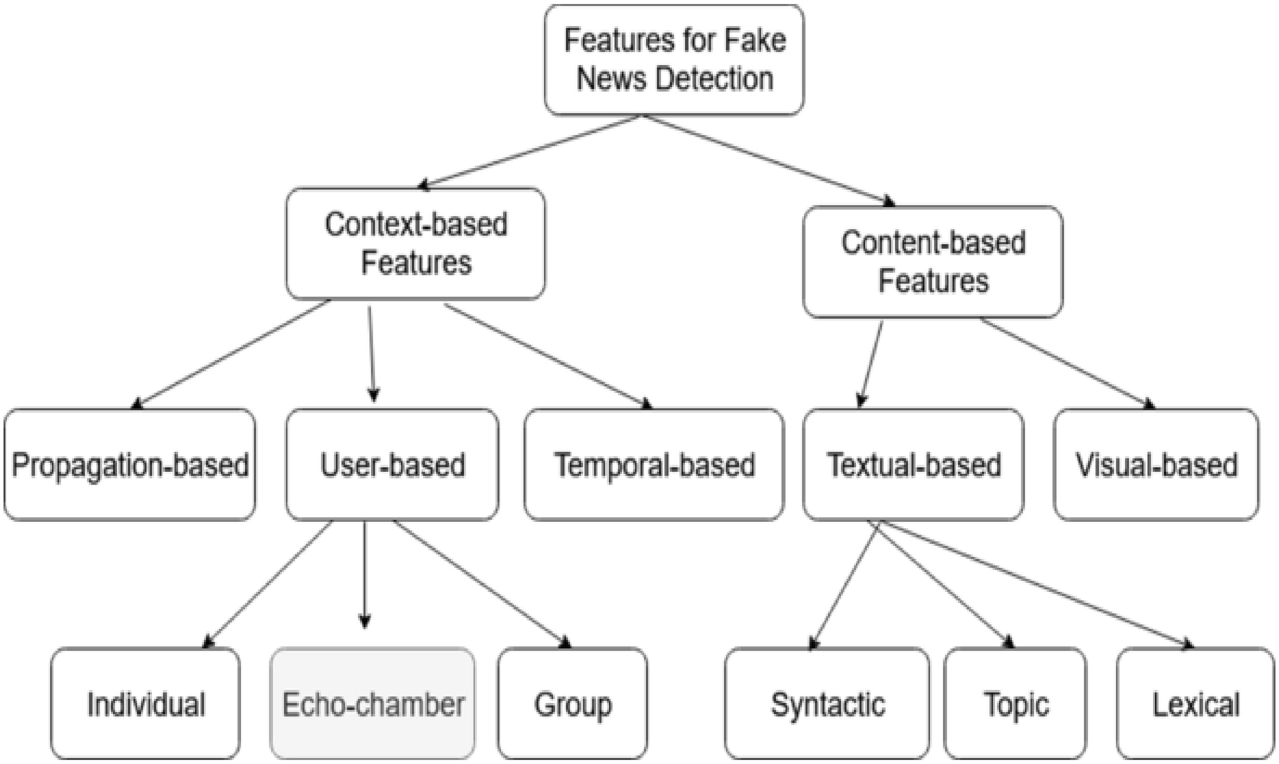
Fake News Detection

It is evident that the maximum accuracy achieved on DS1 (ISOT Fake News Dataset) is 99%, achieved by random forest algorithm and Perez-LSVM. Linear SVM, multilayer perceptron, bagging classifiers, and boosting classifiers achieved an accuracy of 98%. The average accuracy attained by ensemble learners is 97.67% on DS1, whereas the corresponding average for individual learners is 95.25%. The absolute difference between individual learners and ensemble learners is 2.42% which is not significant. Benchmark algorithms Wang-CNN and Wang-Bi-LSTM performed poorer than all other algorithms.

On DS2, bagging classifier (decision trees) and boosting classifier (XGBoost) are the best performing algorithms, achieving an accuracy of 94%. Interestingly, linear SVM, random forest, and Perez-LSVM performed poorly on DS2. Individual learners reported an accuracy of 47.75%, whereas ensemble learners' accuracy is 81.5%.A similar trend is observed for DS3, where individual learners' accuracy is 80% whereas ensemble learners' accuracy is 93.5%. However, unlike DS2, the best performing algorithm on DS3 is Perez-LSVM which achieved an accuracy of 96%. On DS4 (DS1, DS2, and DS3 combined), the best performing algorithm is random forest (91% accuracy).



On average, individual learners achieved an accuracy of 85%, whereas ensemble learners achieved an accuracy of 88.16%. The worst performing algorithm is Wang-Bi-LSTM which achieved an accuracy of 62%.



Source Code

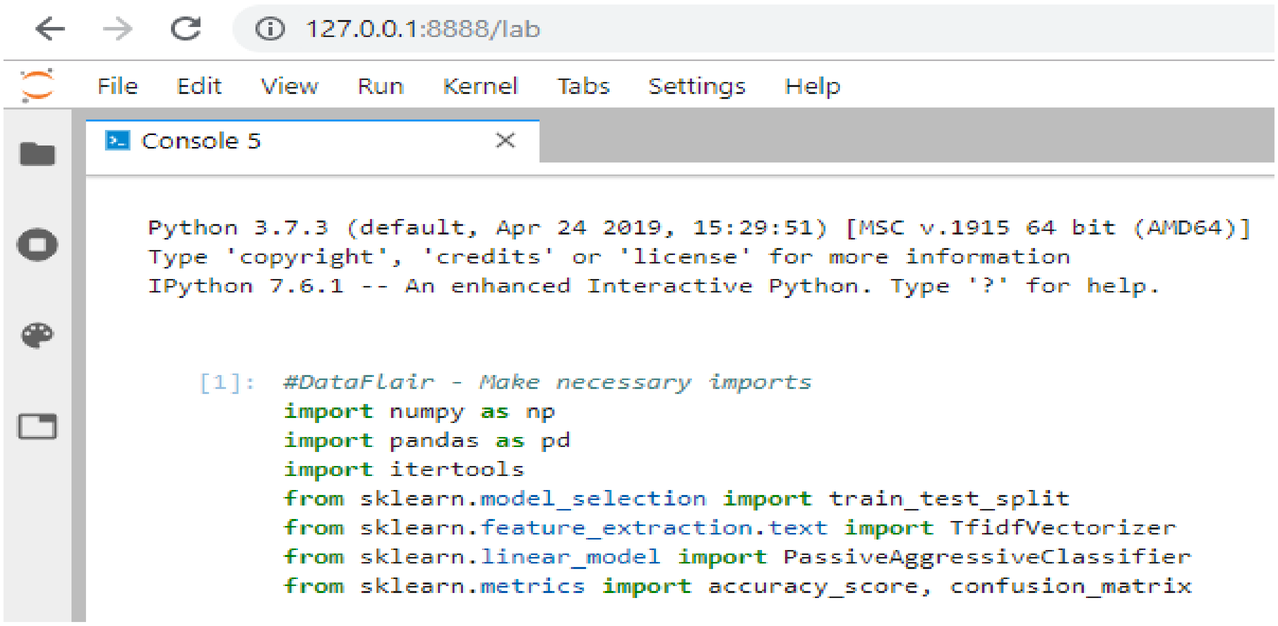
### Steps for Detecting Fake News

Prerequisites

```
pip install numpy pandas sklearn
C:\Users\DataFlair>jupyter lab
```

1) Make necessary imports:

```
import numpy as np
import pandas as pd
import itertools
from sklearn.model_selection import train_test_split
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.linear_model import PassiveAggressiveClassifier
from sklearn.metrics import accuracy_score, confusion_matrix
```

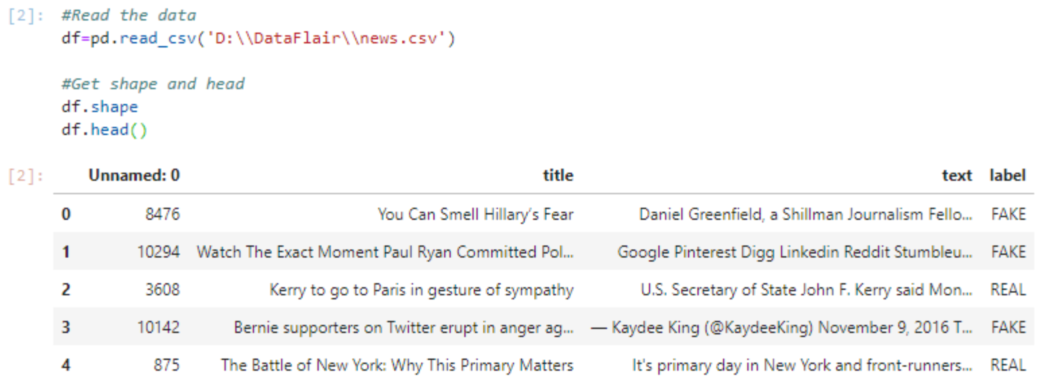


```
Python 3.7.3 (default, Apr 24 2019, 15:29:51) [MSC v.1915 64 bit (AMD64)]
Type 'copyright', 'credits' or 'license' for more information
IPython 7.6.1 -- An enhanced Interactive Python. Type '?' for help.

  [1]: #DataFlair - Make necessary imports
       import numpy as np
       import pandas as pd
       import itertools
       from sklearn.model_selection import train_test_split
       from sklearn.feature_extraction.text import TfidfVectorizer
       from sklearn.linear_model import PassiveAggressiveClassifier
       from sklearn.metrics import accuracy_score, confusion_matrix
```

2) Now, let’s read the data into a DataFrame,.

```
#Read the data
df=pd.read_csv('D:\\DataFlair\\news.csv')
```

```
#Get shape and head
df.shape
df.head()
```



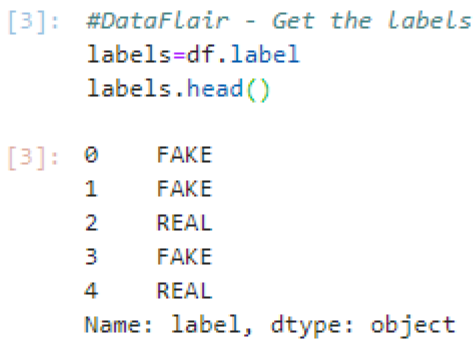
```
[2]: #Read the data
     df=pd.read_csv('D:\\DataFlair\\news.csv')

     #Get shape and head
     df.shape
     df.head()
```

[2]:

	Unnamed: 0	title	text	label
0	8476	You Can Smell Hillary's Fear	Daniel Greenfield, a Shillman Journalism Fello...	FAKE
1	10294	Watch The Exact Moment Paul Ryan Committed Pol...	Google Pinterest Digg Linkedin Reddit Stumbleu...	FAKE
2	3608	Kerry to go to Paris in gesture of sympathy	U.S. Secretary of State John F. Kerry said Mon...	REAL
3	10142	Bernie supporters on Twitter erupt in anger ag...	— Kaydee King (@KaydeeKing) November 9, 2016 T...	FAKE
4	875	The Battle of New York: Why This Primary Matters	It's primary day in New York and front-runners...	REAL

3) And get the labels from the DataFrame.

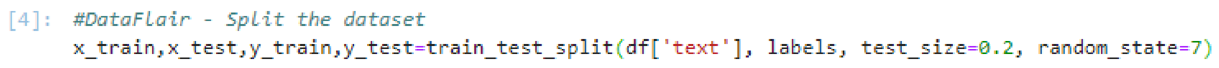
```
#DataFlair - Get the labels
labels=df.label
labels.head()
```



```
[3]: #DataFlair - Get the labels
     labels=df.label
     labels.head()

[3]: 0    FAKE
     1    FAKE
     2    REAL
     3    FAKE
     4    REAL
     Name: label, dtype: object
```

4) Split the dataset into training and testing sets.

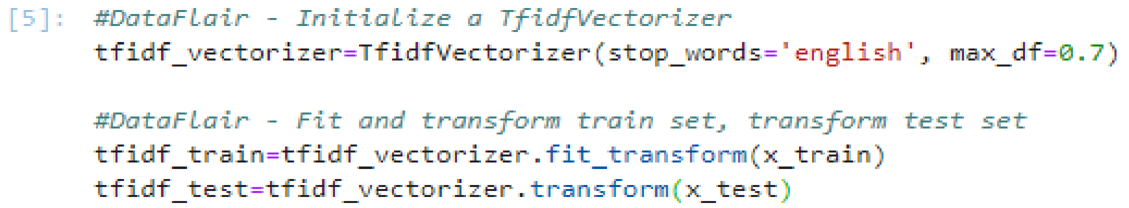
```
#DataFlair - Split the dataset
x_train,x_test,y_train,y_test=train_test_split(df['text'], labels, test_size=0.2, random_state=7)
```



```
[4]: #DataFlair - Split the dataset
     x_train,x_test,y_train,y_test=train_test_split(df['text'], labels, test_size=0.2, random_state=7)
```

5) Let’s initialize a Tfidf-Vectorizer with stop words from the English language and a maximum document frequency of 0.7 (terms with a higher document frequency will be discarded). Stop words are the most common words in a language that are to be filtered out before processing the natural language data. And a Tfidf-Vectorizer turns a collection of raw documents into a matrix of TF-IDF features.

```
#DataFlair - Initialize a TfidfVectorizer
tfidf_vectorizer=TfidfVectorizer(stop_words='english', max_df=0.7)
#DataFlair - Fit and transform train set, transform test set
tfidf_train=tfidf_vectorizer.fit_transform(x_train)
tfidf_test=tfidf_vectorizer.transform(x_test)
```

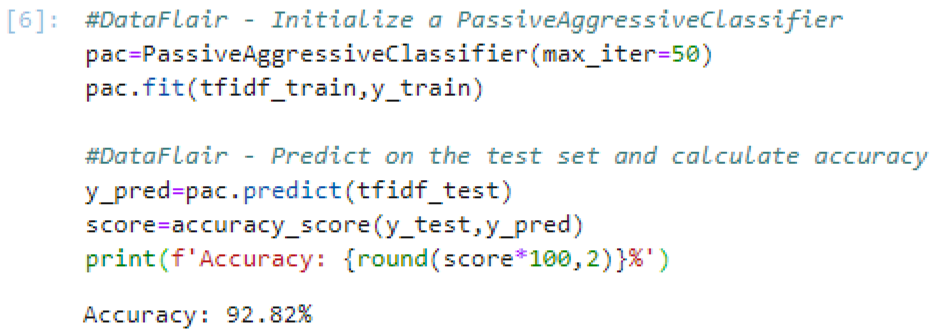


```
[5]: #DataFlair - Initialize a TfidfVectorizer
     tfidf_vectorizer=TfidfVectorizer(stop_words='english', max_df=0.7)

     #DataFlair - Fit and transform train set, transform test set
     tfidf_train=tfidf_vectorizer.fit_transform(x_train)
     tfidf_test=tfidf_vectorizer.transform(x_test)
```

6) Next, we’ll initialize a Passive-Aggressive-Classifier. This is. We’ll fit this on tfidf_train and y_train. Then, we’ll predict on the test set from the Tfidf-Vectorizer and calculate the accuracy with accuracy_score() from sklearn.metrics.

```
#DataFlair - Initialize a PassiveAggressiveClassifier
pac=PassiveAggressiveClassifier(max_iter=50)
pac.fit(tfidf_train,y_train)
#DataFlair - Predict on the test set and calculate accuracy
```

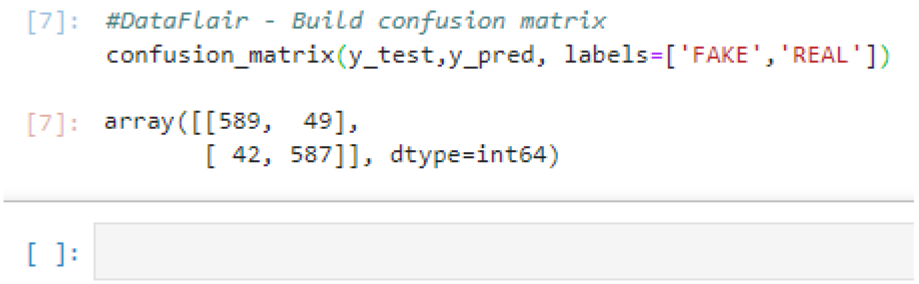
```
y_pred=pac.predict(tfidf_test)
score=accuracy_score(y_test,y_pred)
print(f'Accuracy: {round(score*100,2)}%')
```



```
[6]: #DataFlair - Initialize a PassiveAggressiveClassifier
     pac=PassiveAggressiveClassifier(max_iter=50)
     pac.fit(tfidf_train,y_train)

     #DataFlair - Predict on the test set and calculate accuracy
     y_pred=pac.predict(tfidf_test)
     score=accuracy_score(y_test,y_pred)
     print(f'Accuracy: {round(score*100,2)}%')

     Accuracy: 92.82%
```

7) We got an accuracy of 90.18% with this model. Finally, let's print out a confusion matrix to gain insight into the number of false and true negatives and positives.

```
#DataFlair - Build confusion matrix
confusion_matrix(y_test,y_pred, labels=['FAKE','REAL'])
```



```
[7]: #DataFlair - Build confusion matrix
     confusion_matrix(y_test,y_pred, labels=['FAKE','REAL'])

[7]: array([[589,  49],
            [ 42, 587]], dtype=int64)

[ ]:
```

Some Research on the News

Title	label
Kerry to go to Paris in of sympathy gesture	REAL
How women lead differently gesture	REAL
Trump takes on Cruz, but lightly	REAL
Tehran, USA	FAKE

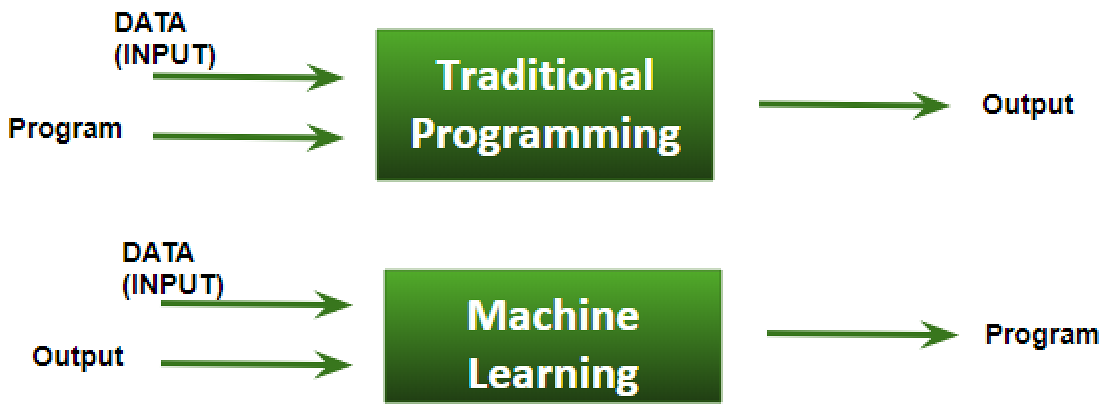
Python libraries for Machine Learning

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

Machine Learning

Machine Learning is the field of study that gives computers the capability to learn without being explicitly programmed. ML is one of the most exciting technologies that one would have ever come across.

As it is evident from the name, it gives the computer that makes it more similar to humans: The ability to learn. Machine learning is actively being used today, perhaps in many more places than one would expect.



How ML works ?

Gathering past data in any form suitable for processing. The better the quality of data, the more suitable it will be for modeling.

Data Processing – Sometimes, the data collected is in the raw form and it needs to be pre-processed.

Example: Some tuples may have missing values for certain attributes, an, in this case, it has to be filled with suitable values in order to perform machine learning or any form of data mining. Missing values for numerical attributes such as the price of the house may be replaced with the mean value of the attribute whereas missing values for categorical attributes may be replaced with the attribute with the highest mode.

This invariably depends on the types of filters we use.

- If data is in the form of text or images then converting it to numerical form will be required, be it a list or array or matrix. Simply, Data is to be made relevant and consistent. It is to be converted into a format understandable by the machine
- Divide the input data into training cross-validation and test sets. The ratio between the respective sets must be 6:2:2. Testing our conceptualized model with data which was not fed to the model at the time of training and evaluating its performance using metrics such as F1 score, precision and recall.

Summary

Today, We detected fake news with Python. We took a political dataset, implemented a Tfidf-Vectorizer, initialized a Passive-Aggressive-Classifier, and fit our model. We ended up obtaining an accuracy of 90.18% as magnitude.

SUBMITTED BY:

SARVAGYA BANSAL

Data Science June Batch