Fake News Detection Using Machine Learning

Tridib Khute[1], Bhupendra Kumar Kumbhkar[2], Mrinal Pawar[3], Priyanka Devi[4]

[1]B.Tech Student, Dept. of Information Technology, Govt. Engineering college Bilaspur, C.G, India
[2]B.Tech Student, Dept. of Information Technology, Govt. Engineering. college Bilaspur, C.G, India
[3]B.Tech Student, Dept. of Information Technology, Govt. Engineering college Bilaspur, C.G, India
[4]Assistant Professor, Dept. of Information Technology, Govt. Engineering college Bilaspur, C.G, India

---***---

Abstract - *Fake information has emerge as a pervasive problem in today's virtual age, posing substantial demanding situations to statistics integrity and public discourse. This paper examines the usage of device getting to know to come across fake news. The research makes a speciality of growing and trying out system gaining knowledge of algorithms that could distinguish among credible information assets and fraudulent facts.*

The review begins with a comprehensive review of the existing literature on link detection, highlighting limitations and gaps in current methodologies. Data types including real and synthetic media are collected and pre-processed to extract relevant features including textual content, metadata, and language samples Using various machine learning algorithms such as logistic regression, random forest, and neural networks are used and compared for better performance in classifying false information Analytical metrics such as accuracy, precision, recall, and F1 score are used to evaluate the performance of machine learning models. Significant factor analysis is performed to identify key determinants of false positives, which contribute to the interpretation of the model. The research also explores cluster learning methods and sample clustering strategies to further enhance classification accuracy and robustness. The research results show promising results in the detection of fake news, showing the ability of machine learning to deal with misinformation The findings contribute to the advancement of fake news detection technology, and inform news organisations, social media channels and law enforcement agencies gain valuable insights for addressing the fake news epidemic.

Key Words: Machine Learning, Natural Language Processing, Fake news, Data Preprocessing, Logistic Regression, Random Forest, Decision Tree Classifier.

1.INTRODUCTION

For the purpose of the research paper on fake news prediction the use of machine getting to know, we are able to define fake news as:

"Fake news refers to intentionally false or deceptive information offered as legitimate news.[1] This misinformation can be created and disseminated via numerous mediums, such as social media platforms, news web sites, and different online resources. Fake news often aims to control public opinion, unfold propaganda, or generate clickbait for financial advantage. It may additionally contain fabricated memories, distorted records, or misattributed resources, in the long run undermining the credibility of dependable journalism and posing widespread challenges to the integrity of statistics ecosystems."

1.1 Importance of Fake News Detection

Preserving Information Integrity: With the rapid dissemination of information facilitated by digital platforms, the prevalence of fake news poses a threat to the integrity of information ecosystems. Mitigating Social and Political Impacts: False information propagated through fake news can have far-reaching consequences on social and political landscapes, including electoral processes, public policy decisions, and community cohesion. Enhancing Media Literacy: Fake news prediction research contributes to the development of tools and methodologies for enhancing media literacy among users. Supporting Fact-Checking Efforts: Predicting fake news complements the efforts of fact-checking organisations and journalists in verifying the accuracy of information. Machine learning algorithms can assist in the automated identification of potentially false or misleading claims, thereby expediting the fact-checking process and enabling more timely corrections and retractions.

Finally, the fake news detection is very helpful and important in this virtual age.

1.2 Role of Machine Learning in Fake News Detection

Machine learning makes it easy to extract relevant features from news content, metadata, and context. Considering different aspects of the data, including linguistic characteristics, such as vocabulary and style, social characteristics such as user engagement and distribution, and source credibility considerations a, machine learning models can better distinguish between true and fake news. Machine learning provides flexibility and scalability, allowing false alarm prediction algorithms to evolve and improve over time. As new fake news emerges and methodology evolves, machine learning algorithms can be developed and retrained on updated data sets to improve deception detection in terms and consistency in various fields. These algorithms are trained on labelled datasets containing examples of real and fake news stories, enabling them to identify patterns and indicators of fraudulent content.

2. Data Collection and Preprocessing

Gathering the dataset from the internet that has labelled news articles both genuine and fake. For our project we used the dataset from the Kaggle for making the ML model. Preprocessing is the nothing but the process of making the row data the computer friendly so the data can be fit in the model and we train our model. It includes clearing, including text normalization, removing noise and ensuring data quality. It includes 52.30% is fake news and 47.70% real news.

2.1 Data Collection

Source of fake news data: The major source of our fake dataset is Kaggle, often hosts datasets related to fake news detection competitions or challenges. These datasets may include labelled examples of fake and genuine news articles, along with associated metadata.

In the data set includes True.csv that contains Real news and Fake.csv that contains Fake news. The main target of the dataset in the class column in it includes two values ‘0’and ‘1’. Were ‘0’ is represents as ‘Real’ and ‘0’ is represents as ‘Fake’. It includes 52.30% is fake news and 47.70% real news.

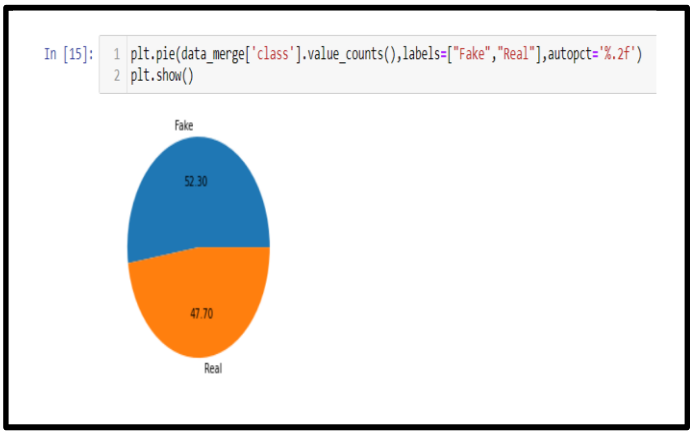

In [15]:
```
plt.pie(data_merge['class'].value_counts(),labels=["Fake","Real"],autopct='%.2f')
plt.show()
```

Chart-1: Fake and Real News Pie Chart

2.2 Data preprocessing technique

Data Cleaning:

Locating and fixing errors or discrepancies in the data. From duplicates and outliers to missing numbers. It fixes them all. It also hands the missing data in dataset this situation performs data cleaning seamlessly. It involves handling of missing data, noise etc. By ignoring the tuples and filling the missing values.

Data Integration:

In data integration is part of data preprocessing this process integrates data that extracted from multiple sources to outline and create a single dataset. In this project we also get some news from the internet sources and integrate it in the Kaggle dataset and make the single dataset.

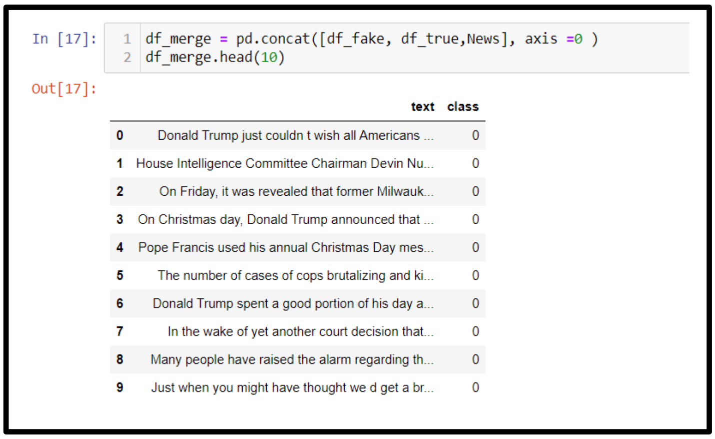
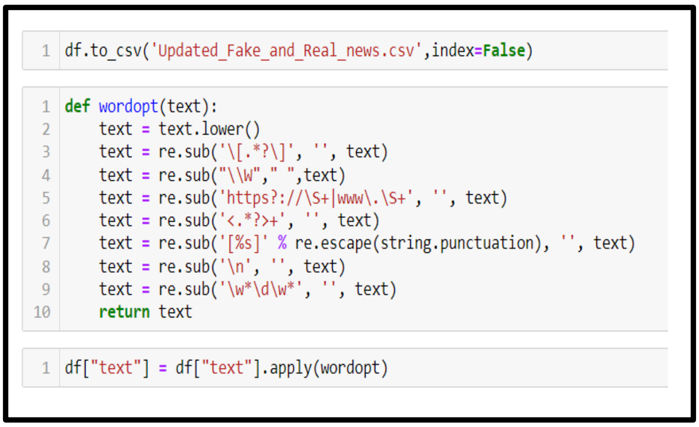

In [17]:
```
df_merge = pd.concat([df_fake, df_true,News], axis =0 )
df_merge.head(10)
```
Out[17]:
```
                                                text  class
0        Donald Trump just couldn t wish all Americans ...      0
1    House Intelligence Committee Chairman Devin Nu...      0
2         On Friday, it was revealed that former Milwauk...      0
3    On Christmas day, Donald Trump announced that ...      0
4    Pope Francis used his annual Christmas Day mes...      0
5        The number of cases of cops brutalizing and ki...      0
6       Donald Trump spent a good portion of his day a...      0
7          In the wake of yet another court decision that...      0
8      Many people have raised the alarm regarding th...      0
9      Just when you might have thought we d get a br...      0
```

Fig-1: Data Integration

Text Cleaning:

Lowercasing: Convert all text to lowercase to ensure uniformity and avoid duplication of features based on case sensitivity. Removing Punctuation: Eliminate punctuation marks as they may not contribute significantly to distinguishing between genuine and fake news TF-IDF Vectorization: Convert text documents into numerical vectors using TF-IDF (Term Frequency-Inverse Document Frequency) to represent the importance of words in distinguishing between documents. Word Embeddings: Utilise pre-trained word embeddings such as Word2Vec, GloVe, or fastText to represent words in a continuous vector space, capturing semantic relationships between words. Dimensionality Reduction: Apply dimensionality reduction techniques such as Principal Component Analysis (PCA) or t-distributed Stochastic Neighbour Embedding (t-SNE) to reduce the dimensionality of feature vectors while preserving relevant information.



```
df.to_csv('Updated_Fake_and_Real_news.csv',index=False)
```

```
def wordopt(text):
    text = text.lower()
    text = re.sub('\[.*?\]', '', text)
    text = re.sub("\\W"," ",text)
    text = re.sub('https?://\S+|www\.\S+', '', text)
    text = re.sub('<.*?>+', '', text)
    text = re.sub('[%s]' % re.escape(string.punctuation), '', text)
    text = re.sub('\n', '', text)
    text = re.sub('\w*\d\w*', '', text)
    return text
```

```
df["text"] = df["text"].apply(wordopt)
```

Fig-2: Text cleaning

2.3 Feature extraction method

Logistic Regression:

Logistic regression is one of the most popular Machine learning algorithms that comes under Supervised Learning techniques. [2] It can be used for Classification as well as for Regression problems, but mainly used for Classification problems. Logistic regression is used to predict the categorical dependent variable with the help of independent variables.[3] The output of the Logistic Regression problem can be only between 0 and 1. Logistic regression can be used where the probabilities between two classes are required. Such as whether it will rain today or not, either 0 or 1, true or false etc. Logistic regression is based on the concept of Maximum Likelihood estimation. According to this estimation, the observed data should be most probable. In logistic regression, we pass the weighted sum of inputs through an activation function that can map values in between 0 and 1.[4] Such activation function is known as sigmoid function and the curve obtained is called sigmoid curve or S-curve. [5]

Sigmoid Function:

· Maps the predictive values to probabilities.

· Sigmoid Function maps any real value into another value between 0 and 1.

· Hence it forms a curve like the "S" Form. This S-form curve is called the Sigmoid function or logistic function.

Where: Z = b0X0 + b1X1 +……..+ bnXn [Here, X0 (bias) is always 1]

Xi is independent variable i = 0,1,2……..., n

Decision Boundary:

Threshold classifier output y' or h0(x) at 0.5:

If h0(x)0.5, predict "y=1"

If ho(x)<0.5, predict "y=0"

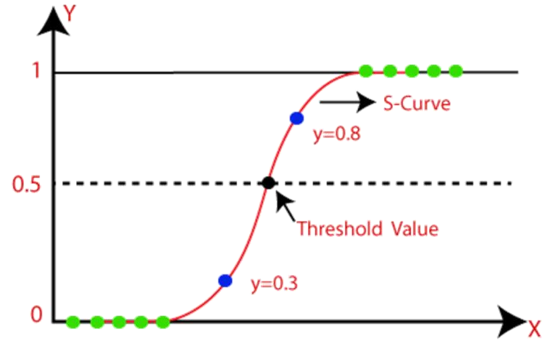


Fig -3: Sigmoid or S-curve

Random Forest:

Random Forest is a popular machine learning algorithm that includes supervised learning methods.[6] It can be used for Classification and Regression problems in ML. It is based on the concept of cluster learning, which is the process of combining multiple classifiers to solve a complex problem and improve the performance of the model.[7] As the name suggests, "Random Forest is a classifier that has a number of decision trees in different subsets of a given dataset and takes an average to improve [8]the prediction accuracy of that dataset Instead of relying on it." on one decision tree, a random forest takes predictions." from each tree and predicts the final outcome based on the majority vote on predictions." Since the random forest combines multiple trees to predict the class of the dataset, it is possible that some decision trees may predict the correct output, while others may not. But together, all the trees predict the correct output.[9] Therefore, below are two assumptions for a better Random Forest classifier:

- There should be some actual values in the feature variable of the dataset so that the classifier can predict accurate results rather than a guessed result.[10]
- The predictions from each tree must have very low correlations.

Although decision trees are common supervised learning algorithms, they can be prone to problems such as bias and overfitting. However, when multiple decision trees are clustered in random forest structures, more accurate results are predicted, especially when individual trees are unrelated to each other. The random wooded area set of rules is made of a group of selection bushes, and each tree within the ensemble is made from a records pattern drawn from a schooling set with substitute, referred to as the bootstrap pattern. Of that education sample, one-0.33 of its far set aside as test records, known as the out-of-bag (oob) sample, which we'll come lower back to later. Another instance of randomness is then injected thru feature bagging, adding extra variety to the dataset and decreasing the correlation amongst selection trees. Depending at the sort of trouble, the dedication of the prediction will vary. For a regression assignment, the character decision timber could be averaged, and for a category mission, a majority vote - i.E. The maximum common express variable—will yield the predicted elegance. Finally, the oob pattern is then used for cross-validation, finalising that prediction.

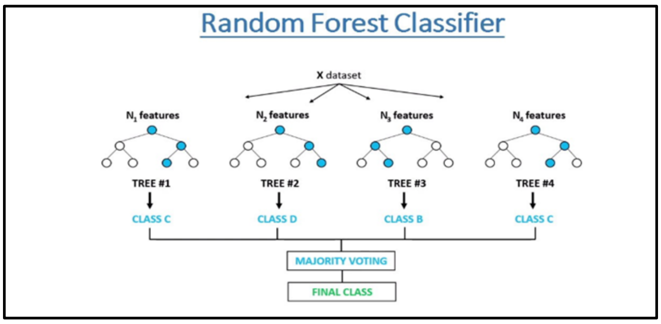


Fig. -4: Random Forest Classifier

Decision Tree:

Decision Trees (DTs) are a non-parametric supervised learning method used for classification and regression. The main aim is to create a model that predicts the value of a target variable by learning simple decision rules that are inferred from the data features.[11] A tree can be visualized as a piecewise constant approximation. In a decision tree, there are two nodes: the Decision Node and Leaf Node. Decision nodes are used to make decisions and have multiple branches, while Leaf nodes are the output of these decisions and do not contain any further branches.[12] The decisions or tests are performed on the basis of features of the given dataset.[13] It is a graphic representation for obtaining all possible solutions to a problem or decision based on provided conditions. It is identified as a decision tree since, resembling a tree, it starts with the root node and then expands on other branches, constructing a tree-like structure. In a decision tree, to predict the class of the given dataset, the algorithm initiates from the root node of the tree. This algorithm compares the values of the root attribute with the record (real dataset) attribute and, based on the comparison, follows the branch and jumps to the next node.[14] For the next node, the algorithm once more compares the attribute value with the other sub-nodes and moves further. The process continues until it reaches the leaf node of the tree.[15]

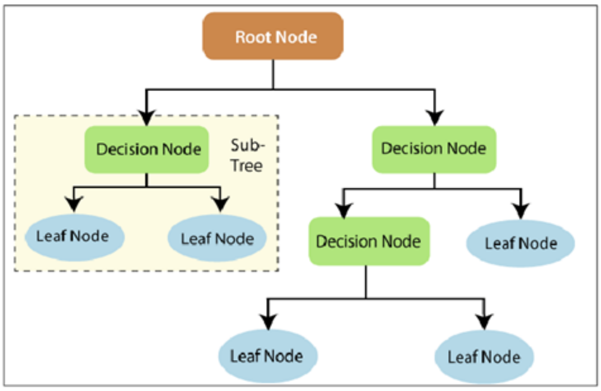


Fig. -5: Decision Tree Classifier

3. Model Training and Testing

Training:

M.L. The term ML model refers to the model artifact generated by the training set.

The training data must have the correct response, known as the target or target type. The learning algorithm looks for patterns in the training data that map the input data features to the target (the response you want to predict), and generates an ML model that captures these patterns Training the model simply means learning (identifying) optimal values for all weights and being biased against Oded models.[16] In supervised learning, a machine learning algorithm generates a model by analysing multiple models and tries to find models that minimize losses; This process is called empirical risk reduction.

Testing:

Machine learning testing helps identify problems in models that may miss routine analytical metrics. This problem can be caused by code or data enabling each component of the ML system where outliers can affect model performance among other things heterogeneous distribution and partitioning This can also help to reduce future problems with the model over the time of use to achieve a certain level of quality assurance for the sample.[17]

ML models are often used for modelling, and while analytical theory is used to predict the performance of a data set, model testing focuses on observing the expected behaviour of the model and is needed because unobserved information path can occur in the workplace. With prototypes, unprecedented challenges can be encountered, and testing makes your prototype more viable under different circumstances :

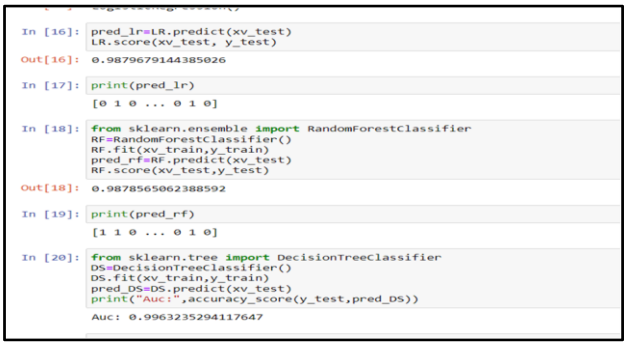
- Adversary attacks: Test models can help identify possible adversary attacks. Instead of having these attacks occur at the manufacturing site, the model can be tested with competing samples to increase its robustness before use

- Data Integrity and Bias: Generally, data collected from most sources is unstructured and may reflect human biases that can odel during training. These biases can be due to gender, race, religion, sexuality or a specific group that has different outcomes in the population based on the level of use Can be missed bias can be missed during research because it draws attention focus more on performance and not on the behaviour of the model given the functional data in this case.
- Spot failure modes: Failure modes can occur when trying to use ML systems in production. These can be due to performance bias failure, robustness failure or model input failure. Some of these failures may be missed by the inspection criteria even though they may indicate problems. A model with 90% accuracy means that it is difficult to generalize the model with 10% of the data. That can motivate you to analyse the data and look for errors so you can get better insights on how to deal with it. This is not all-inclusive and it is important to establish a set of tests for possible scenarios to be encountered to help identify failure mechanisms.

4. Results & Performance Analysis

Accuracy:

ML models are trained on historical data, and their accuracy largely depends on the quality and relevance of this data. ML model testing helps identify bugs between predicted and actual outcomes, allowing developers to fine-tune the model and enhance its accuracy. [18] Accuracy is the percentage of correct classifications that is trained. Machine learning model achieves generally speaking, interesting standards for good accuracy is above 70% In our model we use the cream prediction models that are Logistic regression, Decision tree classification, Random Forest classification and our all models are getting 98% accuracy for predicting the news fake or real So we can say that is a good accuracy of our model.



```
In [16]: pred_lr=LR.predict(xv_test)
         LR.score(xv_test, y_test)

Out[16]: 0.9879679144385026

In [17]: print(pred_lr)

         [0 1 0 ... 0 1 0]

In [18]: from sklearn.ensemble import RandomForestClassifier
         RF=RandomForestClassifier()
         RF.fit(xv_train,y_train)
         pred_rf=RF.predict(xv_test)
         RF.score(xv_test,y_test)

Out[18]: 0.9878565062388592

In [19]: print(pred_rf)

         [1 1 0 ... 0 1 0]

In [20]: from sklearn.tree import DecisionTreeClassifier
         DS=DecisionTreeClassifier()
         DS.fit(xv_train,y_train)
         pred_DS=DS.predict(xv_test)
         print("Auc:",accuracy_score(y_test,pred_DS))

         Auc: 0.9963235294117647
```

Fig. -6: Accuracy of Models

Precision:

The Precision meter checks the prediction accuracy of the positive square. Specificity refers to the number of known/accepted factors that are actually relevant. It is calculated by dividing the true positive by the total positive.

Recall:

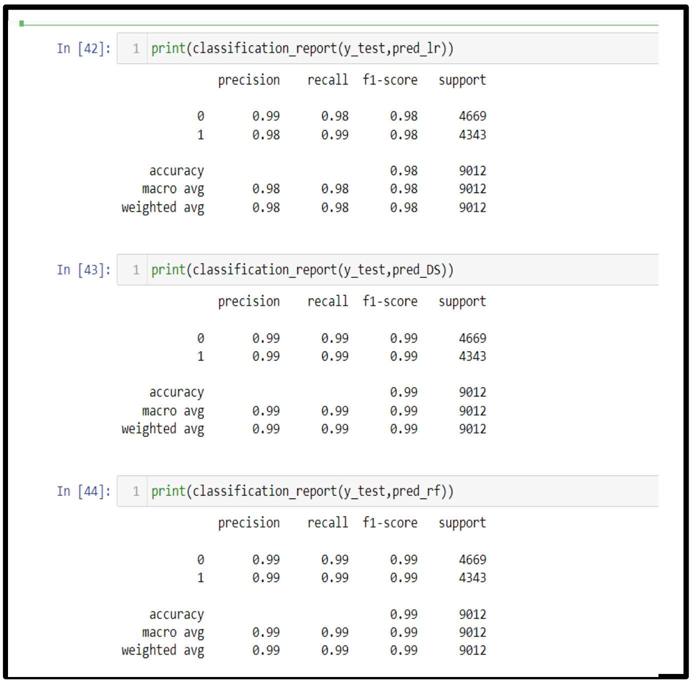
The model's ability to find all relevant cases in a dataset. Mathematically, we define recall as the number of true positives divided by the number of true positives and the number of false negatives.[19]

F1-Score:

The F1 score captures the classifier's accuracy and recall by taking their harmonic mean and combining it into a single metric. It is primarily used to compare the performance of two distributions. Suppose classifier A has more recall and classifier B has higher accuracy.
Generally, an F1 score > 0.9 is considered good. A score of 0.8 to 0.9 is considered good, and a score of 0.5 to 0.8 is considered average. If the F1 score is less than 0.5, the model is considered to perform poorly. Our all model is getting >0.9 F1-score so its show that all models all considered good training and testing.

Classification report:

The classification report shows the status of key classification metrics in each class. This provides a deeper insight into classifier behaviour than the global accuracy that can mask functional weaknesses in a single multiclass problem class. It displays your model's precision, recall, F1 score and support. It provides a better understanding of the overall performance of our trained model.[20]

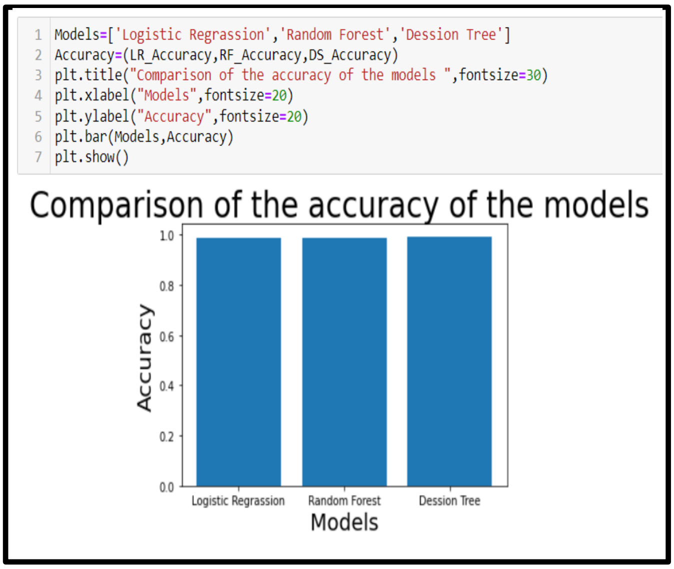


```
In [42]:  print(classification_report(y_test,pred_lr))

              precision    recall  f1-score   support

           0       0.99      0.98      0.98      4669
           1       0.98      0.99      0.98      4343

    accuracy                           0.98      9012
   macro avg       0.98      0.98      0.98      9012
weighted avg       0.98      0.98      0.98      9012


In [43]:  print(classification_report(y_test,pred_DS))

              precision    recall  f1-score   support

           0       0.99      0.99      0.99      4669
           1       0.99      0.99      0.99      4343

    accuracy                           0.99      9012
   macro avg       0.99      0.99      0.99      9012
weighted avg       0.99      0.99      0.99      9012


In [44]:  print(classification_report(y_test,pred_rf))

              precision    recall  f1-score   support

           0       0.99      0.99      0.99      4669
           1       0.99      0.99      0.99      4343

    accuracy                           0.99      9012
   macro avg       0.99      0.99      0.99      9012
weighted avg       0.99      0.99      0.99      9012
```

Fig-7: Classification Report

4. Comparison of Models Accuracy

By comparison of the all three models Logistic Regression, Random Forest and Decision tree classifier the accuracies are respectively 98%, 99% and 99%. All are giving >90% it means its preferred as the good model. But in all of three model the Random Forest and Decision Tree are considered as the best model as compare to the Logistic Regression model.



```python
Models=['Logistic Regrassion','Random Forest','Dession Tree']
Accuracy=(LR_Accuracy,RF_Accuracy,DS_Accuracy)
plt.title("Comparison of the accuracy of the models ",fontsize=30)
plt.xlabel("Models",fontsize=20)
plt.ylabel("Accuracy",fontsize=20)
plt.bar(Models,Accuracy)
plt.show()
```

Fig-8: Comparison of Accuracy of the Models

5. Data Flow Diagram

A data flow diagram maps out the flow of information for any process or system. It uses the defined symbols like rectangle circles and arrows plus short taste labels to sow data and push up storage pointers and the route between each destination. The flow of data of a system or a process is represented by DFD. [21] It also gives insight into the inputs and the output of each entity and the process itself. Data fraud describes the information transferring between different parts of the system.
The dataflow diagram for the 'Fake News detection' its start from the taking the content of the news the do the preprocessing on to making it clearer and more efficient after the pre-processing of the data to do deleting the unnecessary symbols from the text and to the vectorizing of the cleaned text at last putting it in the machine learning model it predicts in 0 or 1 form means Fake or Real. The dataflow diagram shows the all the process of the data from stat to ending in the structured form for making it clearer and more understandable.

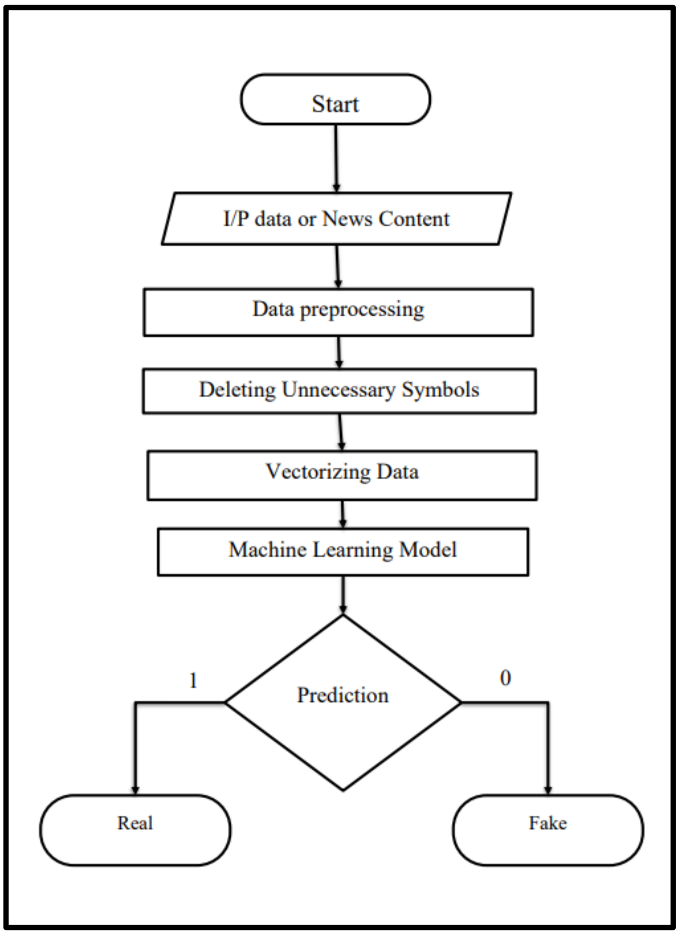


Fig-8: Data Flow Diagram

3. CONCLUSIONS

The paper on fake news detection using machine learning concludes with an important contribution to the effort to deal with misinformation. The study demonstrates the effectiveness of machine learning algorithms such as logistic regression and random forest in accurately classifying fake news data Through rigorous analysis and comparison, research demonstrates the ability of machine learning models to distinguish between credible news sources and fraudulent information. The findings highlight the importance of feature engineering, model selection, and evaluation metrics in a robust fake news detection system. Feature need analysis helps to understand key indicators of false information, increasing model interpretability and reliability. The study also explores ensemble learning methods and model ensemble techniques to achieve classification accuracy and further improve the robustness of pseudo-information techniques. The results of this research extend to news organisations, social media platforms and law enforcement agencies, providing insights into effective strategies to identify and reduce the spread of fake news Future work in the field requires a deeper learning process, incorporate semantic analysis and provide model generalisation to false new claims It can also help preserve integrity

REFERENCES

[1] Research guides: Fake news and information literacy: What is fake news?. What is Fake News? - Fake News and Information Literacy - Research Guides at University of Oregon Libraries. (n.d.). https://researchguides.uoregon.edu/fakenews/issues/defining

[2] *Logistic regression summary (DOCX)*. Course Sidekick. (n.d.).https://www.coursesidekick.com/information-systems/1571288

[3] *Linear regression vs logistic regression - javatpoint*. www.javatpoint.com. (n.d.). https://www.javatpoint.com/linear-regression-vs-logistic-regression-in-machine-learning

[4] *Linear regression vs logistic regression - javatpoint*. www.javatpoint.com. (n.d.-a). https://www.javatpoint.com/linear-regression-vs-logistic-regression-in-machine-learning#:~:text=In%20logistic%20regression%2C%20we%20pass,sigmoid%20curve%20or%20S%2Dcurve.

[5] *Fake news detection using machine learning - javatpoint*. www.javatpoint.com. (n.d.-a). https://www.javatpoint.com/fake-news-detection-using-machine-learning#:~:text=Machine%20learning%20algorithms%20used%20for,as%20either%20real%20or%20fake.

[6] Sahil Khalkar, Rushikesh Nimbhore, Atharva Pardeshi, Sanket Kanade and V.K.Barbudhe. IOT Based Data Monitoring in Secured Block Chain Architecture, International Journal for Modern Trends in Science and Technology, 2023, 9(11), pages. 01-04.https://doi.org/10.46501/IJMTST0911001

[7]*Tutorials - javatpoint*. www.javatpoint.com. (n.d.-d). http://www.javatpoint.com/

[8] *Machine learning random forest algorithm - javatpoint*. www.javatpoint.com. (n.d.-d). https://www.javatpoint.com/machine-learning-random-forest-algorithm

[9] Anusha, H. (2022, January 28). *Bootstrapped Aggregation(bagging):*. Medium. https://medium.com/@hemaanushatangellamudi/bootstrapped-aggregation-bagging-481f4812e3ea

[10] *Analytics vidhya: The ultimate place for generative AI, Data Science and Data Engineering*. Analytics Vidhya | The ultimate place for Generative AI, Data Science and Data Engineering. (n.d.). https://www.analyticsvidhya.com/

[11] *Analytics vidhya: The ultimate place for generative AI, Data Science and Data Engineering*. Analytics Vidhya | The ultimate place for Generative AI, Data Science and Data Engineering. (n.d.). https://www.analyticsvidhya.com/

[12] Choubey, V. (2020, October 28). *Decision tree-end to end implementation*. Medium. https://medium.com/analytics-vidhya/decision-tree-end-to-end-implementation-adf1bc246254

[13] Kiran, K. (2023, March 3). *Decision tree*. LinkedIn. https://www.linkedin.com/pulse/decision-tree-kajal-kiran

[14] *Decision tree algorithm in Machine Learning - Javatpoint*. www.javatpoint.com. (n.d.-a). https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm

[15] E. Subramanian, M.Mithun Karthik ,G.Prem Krishna,V.Sugesh Kumar, D.Vaisnav Prasath, ”Solar Power Prediction Using Machine Learning ”.

[16] Khanam, Z., Alwasel, B. N., Sirafi, H., & Rashid, M. (2021). Fake news detection using machine learning approaches. *IOP Conference Series: Materials Science and Engineering*, *1099*(1), 012040. https://doi.org/10.1088/1757-899x/1099/1/012040

[17] Kaliyar, R. K., Goswami, A., Narang, P., & Sinha, S. (2020). FNDNet – a deep convolutional neural network for fake news detection. *Cognitive Systems Research*, *61*, 32–44. https://doi.org/10.1016/j.cogsys.2019.12.005

[18] https://www.testingxperts.com/blog/ml-testing

[19] *Precision and recall - shiksha online*. shiksha. (n.d.). https://www.shiksha.com/online-courses/articles/precision-and-recall

[20] Kharwal, A. (2021, July 7). *Classification report in Machine Learning: Aman Kharwal*. thecleverprogrammer. https://thecleverprogrammer.com/2021/07/07/classification-report-in-machine-learning

[21] GfG. (2023, September 20). *What is dfd(data flow diagram)?*. GeeksforGeeks. https://www.geeksforgeeks.org/what-is-dfddata-flow-diagram/