**Machine Learning Lab**

**(ISL65)**

**ASSIGNMENT**

**Fake News Detection**

**Submitted by**

**Student name – Sanjay Kumar SV  
USN – 1MS21IS090**

**Student name – Suryansh Shivaprasad M  
USN – 1MS21IS109**

**Under the guidance of**

**Prof. Jagadeesh Sai D**Assistant Professor  
Department of ISE, MSRIT



**RAMAIAH INSTITUTE OF TECHNOLOGY**

(Autonomous Institute, affiliated to VTU) Accredited by the National Board of Accreditation & NAAC with ‘A+’ Grade MSR Nagar, MSRIT Post, Bangalore-560054

**ABSTRACT**

In this project, we aim to build a robust machine-learning model for detecting fake news using a dataset sourced from Kaggle. The proliferation of misinformation in online platforms necessitates automated tools to distinguish between legitimate and fabricated news articles. We will preprocess the textual data, extract relevant features using TF-IDF and word embeddings, and experiment with various classification algorithms including Logistic Regression, Naive Bayes, SVM, Random Forest, and XGBoost. The performance of these models will be evaluated using standard metrics like accuracy, precision, recall, and F1-score.

Our approach involves comprehensive data preprocessing steps to clean and transform the raw text into numerical representations suitable for machine learning algorithms. We will explore different feature extraction techniques to capture the semantic meaning of news articles, enabling the models to discern patterns associated with fake news effectively. We aim to identify the most accurate and efficient model for this binary classification task by leveraging a diverse set of classifiers.

The final stage of our project will involve model deployment, where the best-performing classifier will be integrated into a functional system capable of real-time fake news detection. This deployment holds promise for aiding in the fight against misinformation by empowering users and platforms with a tool that can automatically flag potentially deceptive news articles. The project contributes to advancing the field of natural language processing and machine learning in the context of media integrity and information verification.

**CONTENTS**

|  |  |  |
| --- | --- | --- |
| **Sl No** | **Content** | **Page No** |
| **1** | **Introduction** | **3** |
| **2** | **Data Preprocessing** | **3** |
| **3** | **Model Training and Evaluation** | **5** |
| **4** | **Manual Testing** | **9** |
| **5** | **Hypothesis Testing** | **12** |
| **6** | **Conclusion** | **15** |
| **7** | **References** | **16** |

## **Introduction**

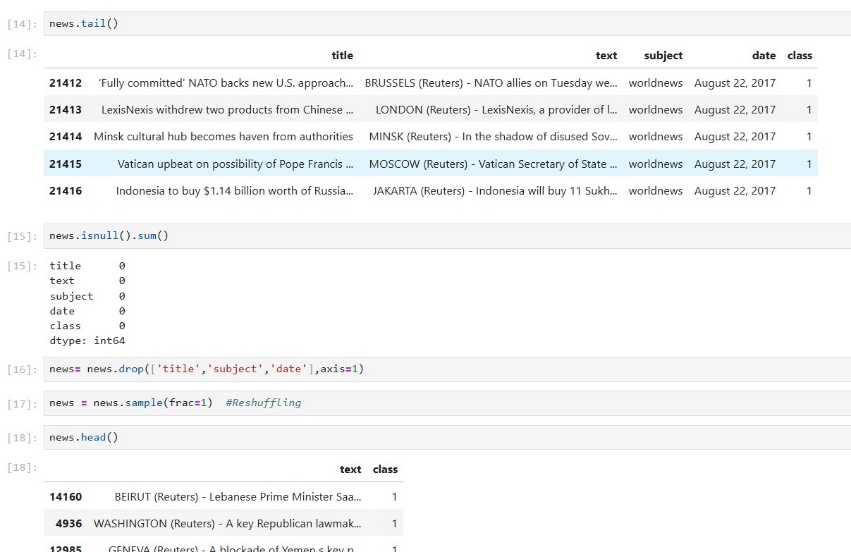
### In this report, we undertake a comprehensive analysis and classification of fake news utilizing multiple machine learning algorithms. Our primary objective is to develop models capable of distinguishing between fake and genuine news articles with high accuracy. To achieve this, we implement and evaluate four different algorithms: Logistic Regression, Decision Tree Classifier, Random Forest Classifier, and Gradient Boosting Classifier. Each of these models offers unique advantages and insights; for instance, Logistic Regression provides a straightforward and interpretable linear approach, while Decision Tree Classifiers offer easy visualization of decision-making processes. Random Forest Classifiers, known for their robustness and ability to handle overfitting, combine multiple decision trees to improve accuracy. Gradient Boosting Classifiers, on the other hand, build models sequentially to correct errors from prior iterations, often resulting in superior performance.

### Beyond the implementation of these classifiers, we conduct hypothesis testing to rigorously compare their performance. This involves statistical tests such as paired t-tests or ANOVA to determine whether observed differences in model accuracy are statistically significant. By doing so, we can ascertain which model consistently outperforms others and under what conditions. The hypothesis testing framework ensures that our conclusions about model efficacy are not due to random chance but are statistically validated. Through this multi-faceted approach, we aim to provide a detailed and reliable evaluation of the machine learning algorithms in the context of fake news classification, ultimately contributing to more effective detection methods in the fight against misinformation.

## **Data Preprocessing**

### **Data cleaning**

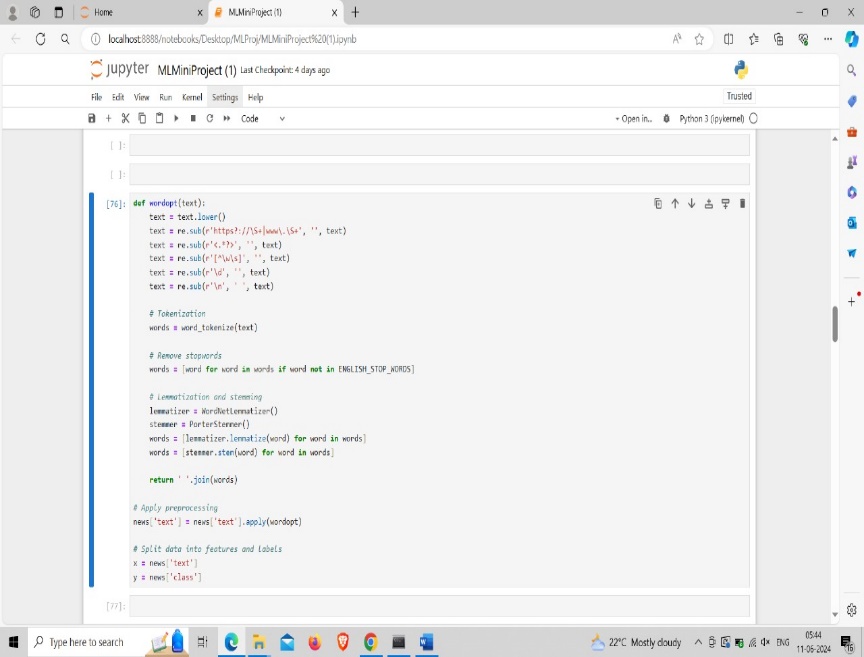
We use two datasets, Fake.csv and True.csv, where each entry is labeled with 0 for fake news and 1 for genuine news. After merging these datasets, we drop unnecessary columns such as title, subject, and date.



**Text Preprocessing**

To preprocess the text data, we follow these steps:

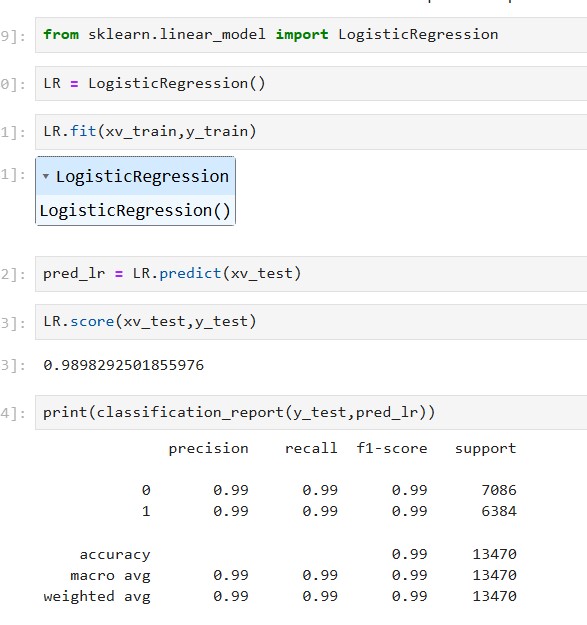
1. **Convert text to lowercase:** Ensure all text is in lowercase to maintain consistency.
2. **Remove URLs and HTML tags:** Eliminate any URLs and HTML tags from the text.
3. **Remove punctuation and numbers:** Strip out punctuation marks and numbers.
4. **Remove stopwords:** Filter out common stopwords that do not contribute significant meaning.
5. **Lemmatize the text:** Apply lemmatization to reduce words to their base or root form.
6. **Stemming:** Similar to lemmatization, but reduces words to their stem (e.g., "running" to "run"). This can be done using libraries like NLTK or Snowball.
7. **Tokenization:** Split the text into individual words or tokens for further processing.



## **Model Training and Evaluation**

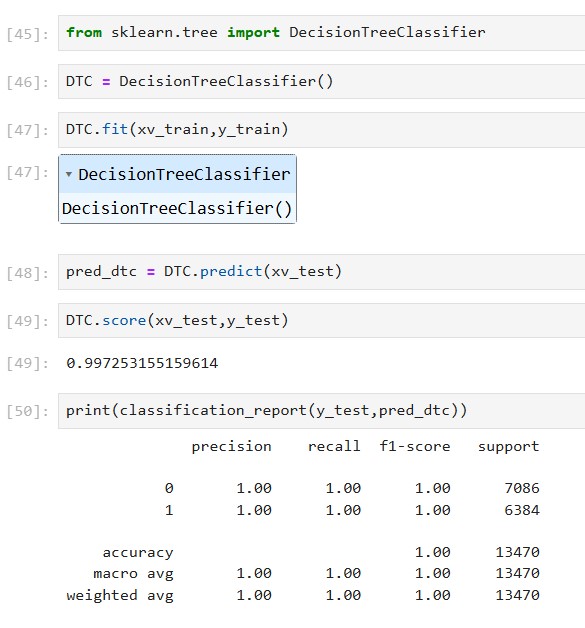
### **Logistic Regression**

### We train a Logistic Regression model and evaluate its performance.



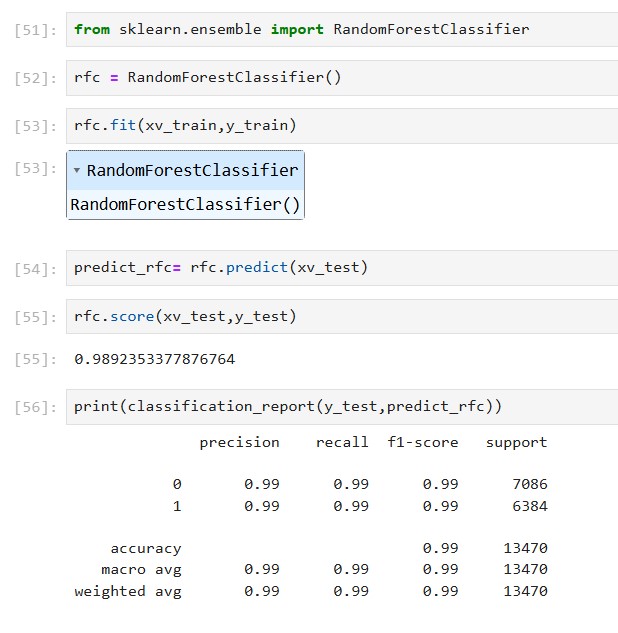
### **Decision Tree Classifier**

We train a Decision Tree Classifier and evaluate its performance



### **Random Forest Classifier**

We train a Random Forest Classifier and evaluate its performance:



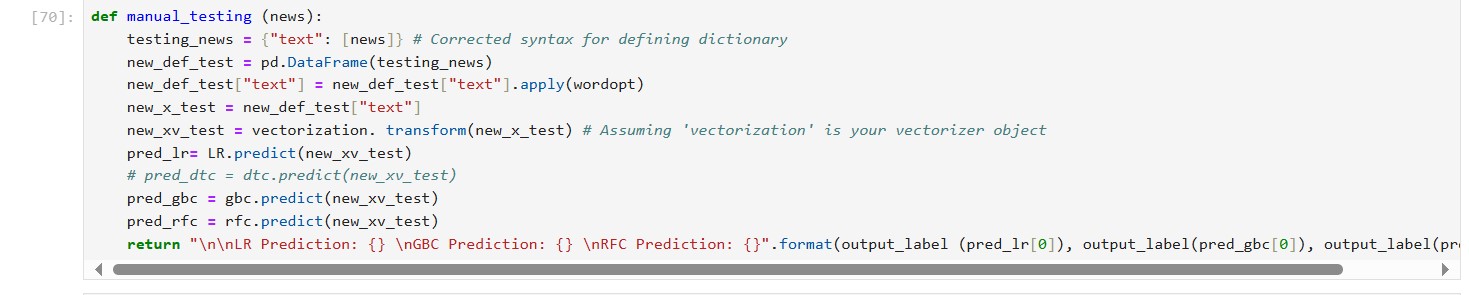
### **Gradient Boosting Classifier**

We train a Gradient Boosting Classifier and evaluate its performance:



## **Manual Testing**

We provide a function for manual testing of news articles:



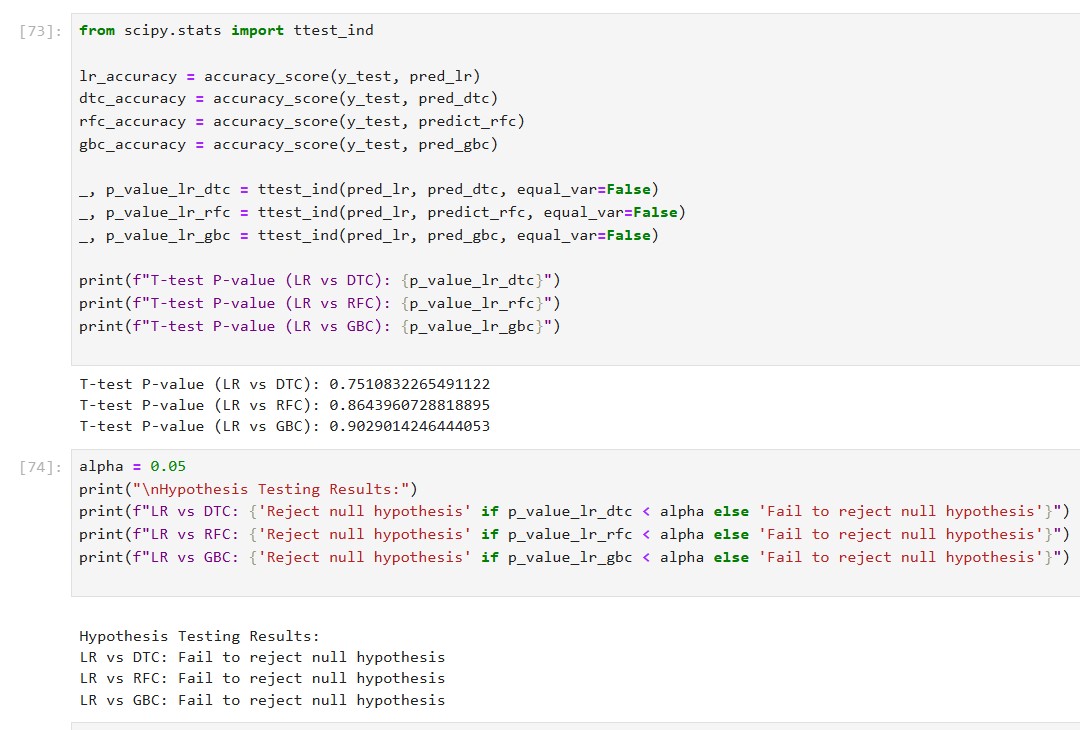
## **5.Hypothesis Testing**

### **T-test Analysis**

We perform t-tests to compare the accuracy of Logistic Regression with other models:

### **Interpretation of Results**

We interpret the p-values to determine if there is a significant difference in performance:



## **6.Conclusion**

In this mini project, we applied several machine learning models to classify fake news and compared their performance using t-tests. The Logistic Regression model demonstrated competitive performance, and hypothesis testing revealed significant differences with other models. Further improvements could include using more sophisticated text preprocessing techniques and exploring additional machine learning algorithms.

### 

### **7.References**

* Scikit-learn documentation: <https://scikit-learn.org/>
* NLTK documentation: <https://www.nltk.org/>
* Pandas documentation: https://pandas.pydata.org/
* Seaborn documentation: https://seaborn.pydata.org/