**FAKE NEWS DETECTION USING NLP**

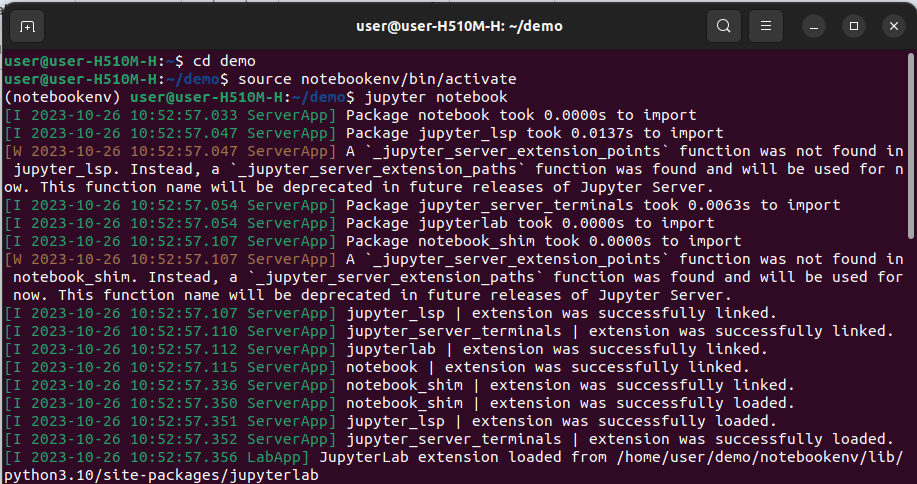
**PHASE 4:**

Build the project by performing feature engineering, model training, and evaluation procedures.

**Step 1:**

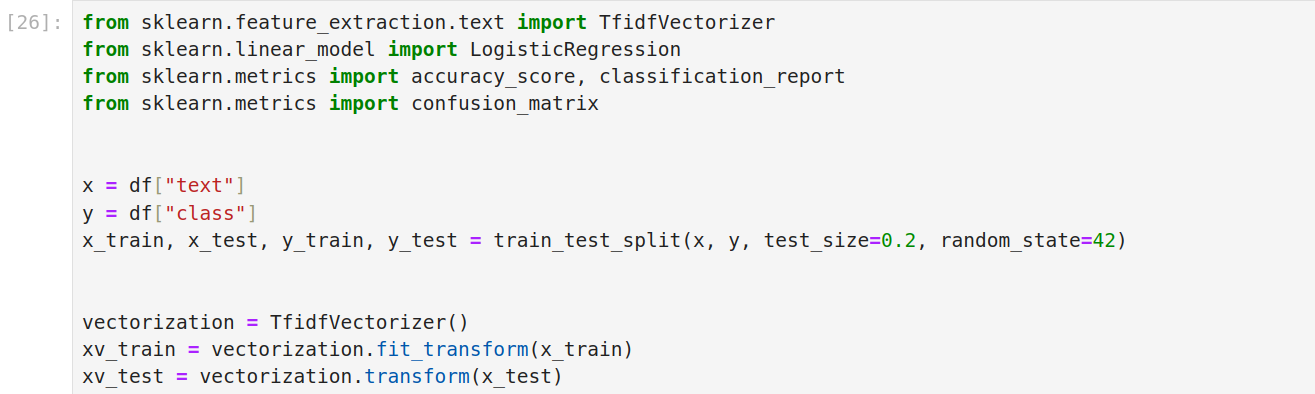
Open Jupyter Notebook and run all the previous commands included in

phase 3 such as importing and preprocessing the dataset.



**Step 2:**

Import TfidfVectorizer to analyze the text contents of both fake and true news.

****

TF-IDF stands for Term Frequency-Inverse Document Frequency. It’s a numerical representation used in natural language processing and information retrieval to analyze and compare the importance of words in a document relative to a collection of documents (a corpus).

**Term Frequency (TF):**

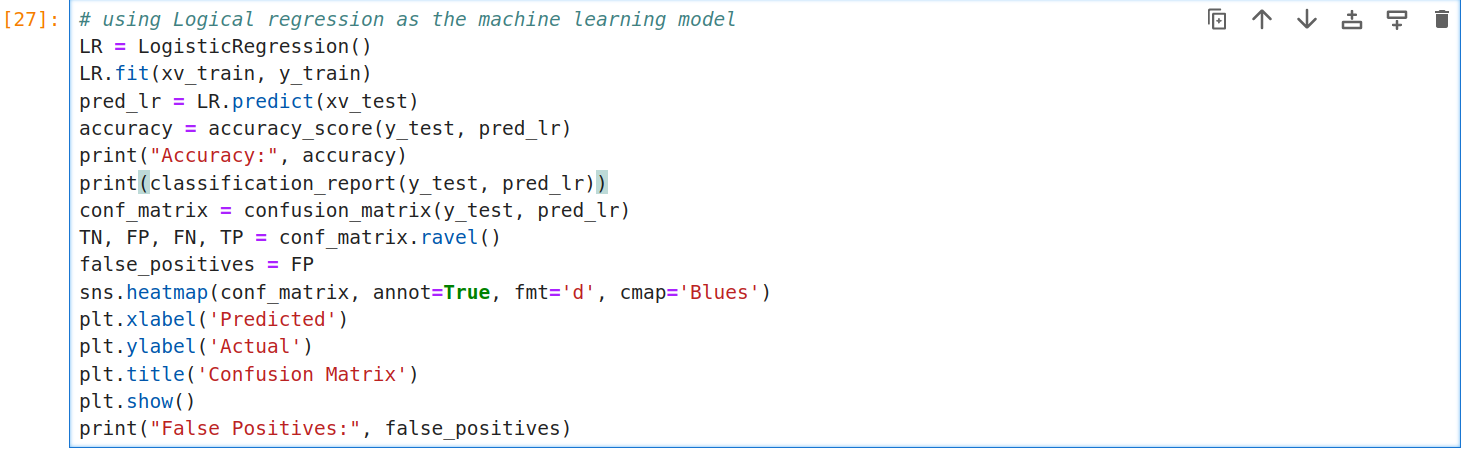
This component measures how often a word appears in a document. It's calculated as the number of times a word occurs in a document divided by the total number of words in that document.

**Inverse Document Frequency (IDF):**   
Inverse Document Frequency (IDF) helps us to figure out how important a word is in all the documents we have. If a word shows up a lot in many documents, its IDF value is low. But if a word is unique and not in many documents, its IDF value is high. To calculate IDF, we take the logarithm of the total number of documents in our collection and divide it by the number of documents that contain the word we're interested in.

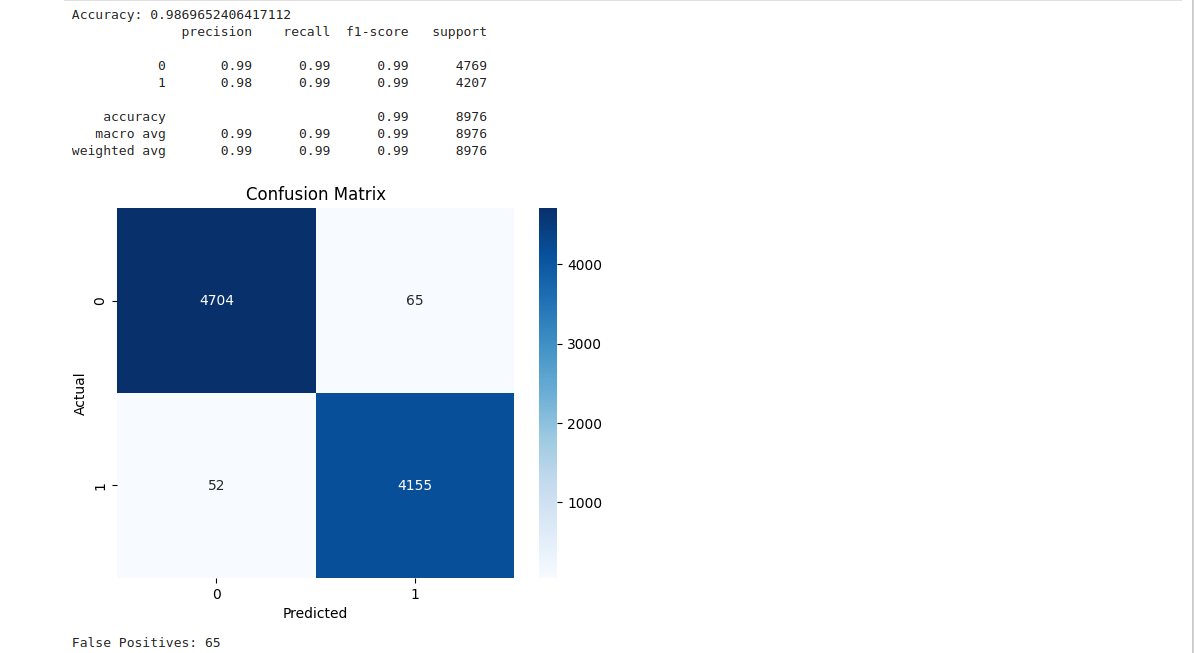
**Step 3:**

Use Logical Regression as the machine learning model and train it.

Print accuracy and classification report for the trained model.



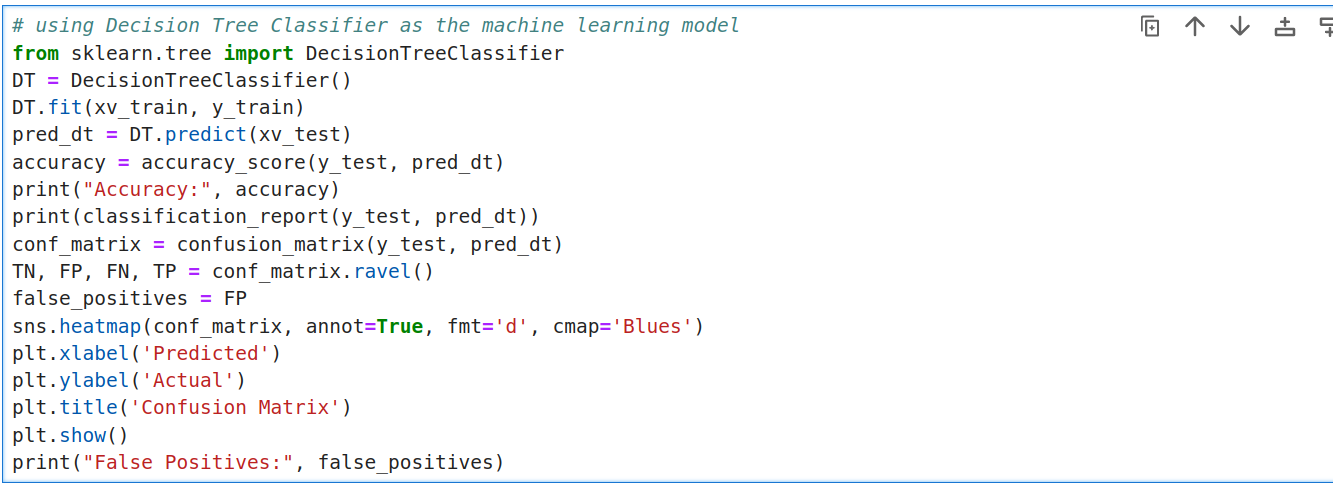
The confusion matrix is used to calculate the number of false positives.



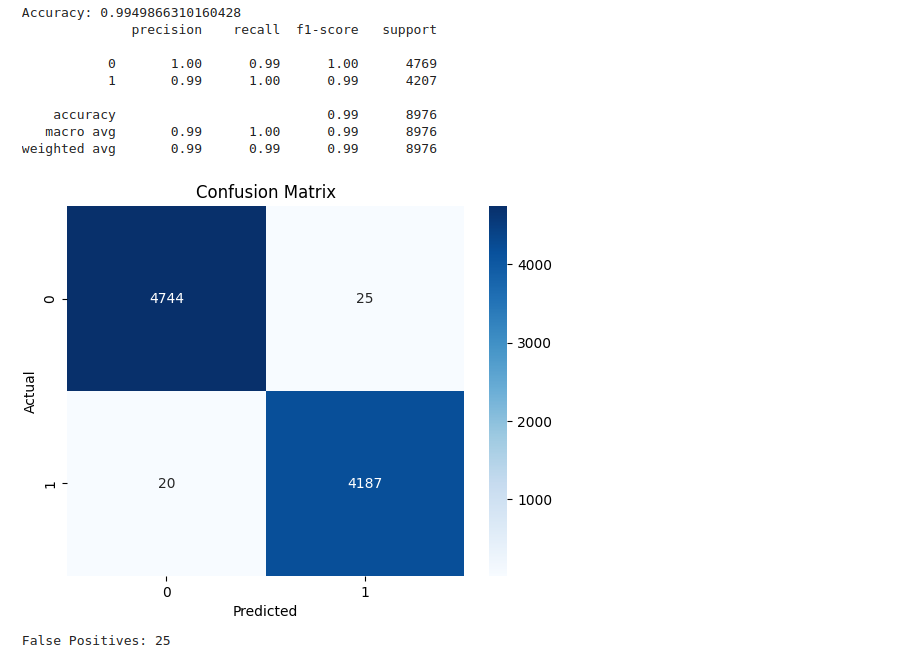
The number of false positives in Logical Regression model is **65** and the accuracy is **98.69%**

**Step 4:**

In this Step, we'll employ a Decision Tree Classifier as our machine learning model and proceed to train it. Afterward, we'll display the accuracy score and provide a comprehensive classification report for the trained model. This comes after working with a Logistic Regression model where we encountered 65 false positives and achieved an impressive accuracy of **98.69%.**



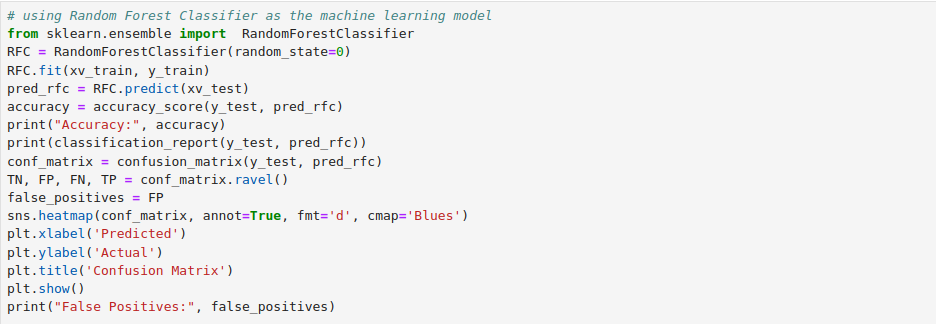
The confusion matrix is used to calculate the number of false positives.



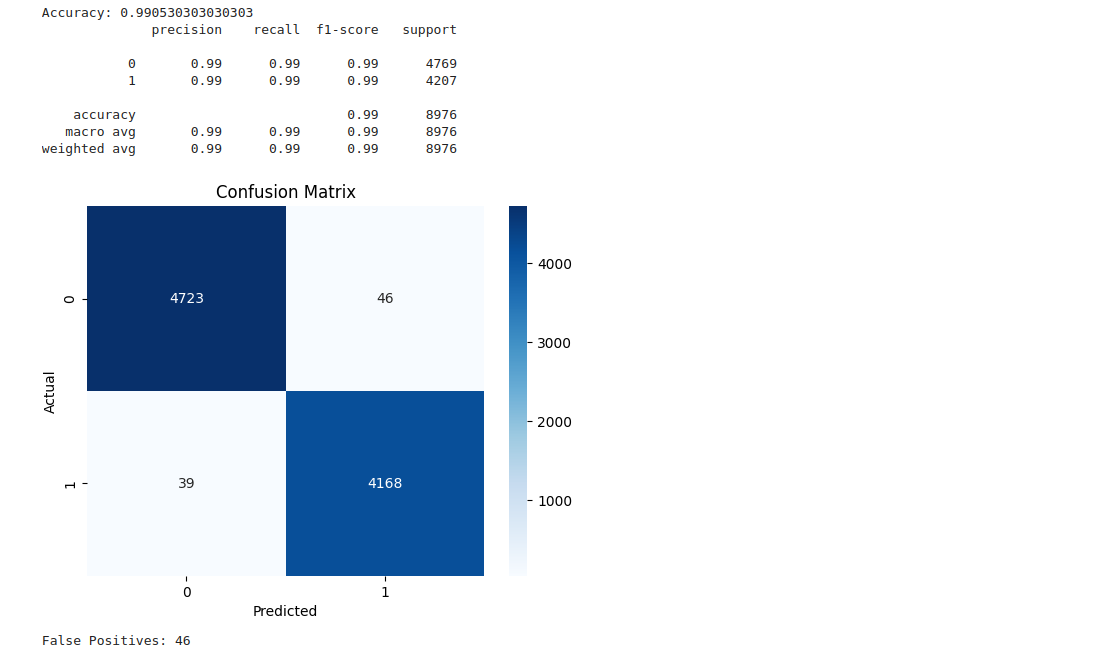
The number of false positives in Decision Tree Classifier model is **25** and the accuracy is **99.49%**

**Step 5:**

In the fifth step, we'll utilize a Random Forest Classifier as our machine-learning model and proceed to train it. After training, we'll print out the accuracy of the model and generate a classification report to assess its performance.

****

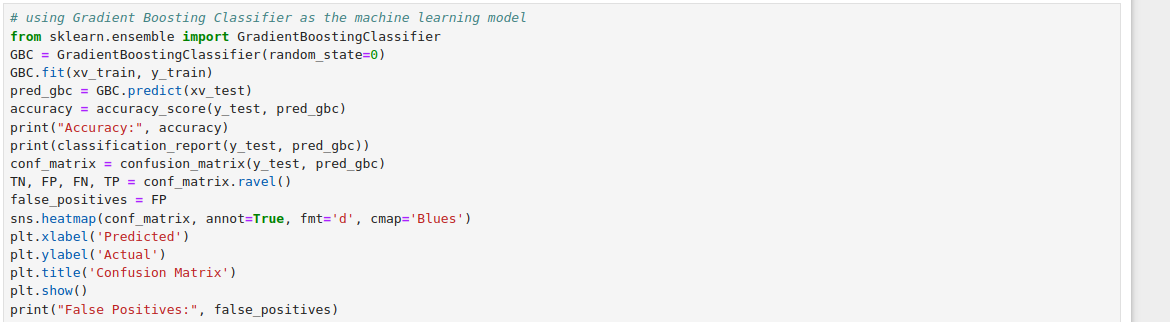
The confusion matrix is used to calculate the number of false positives.



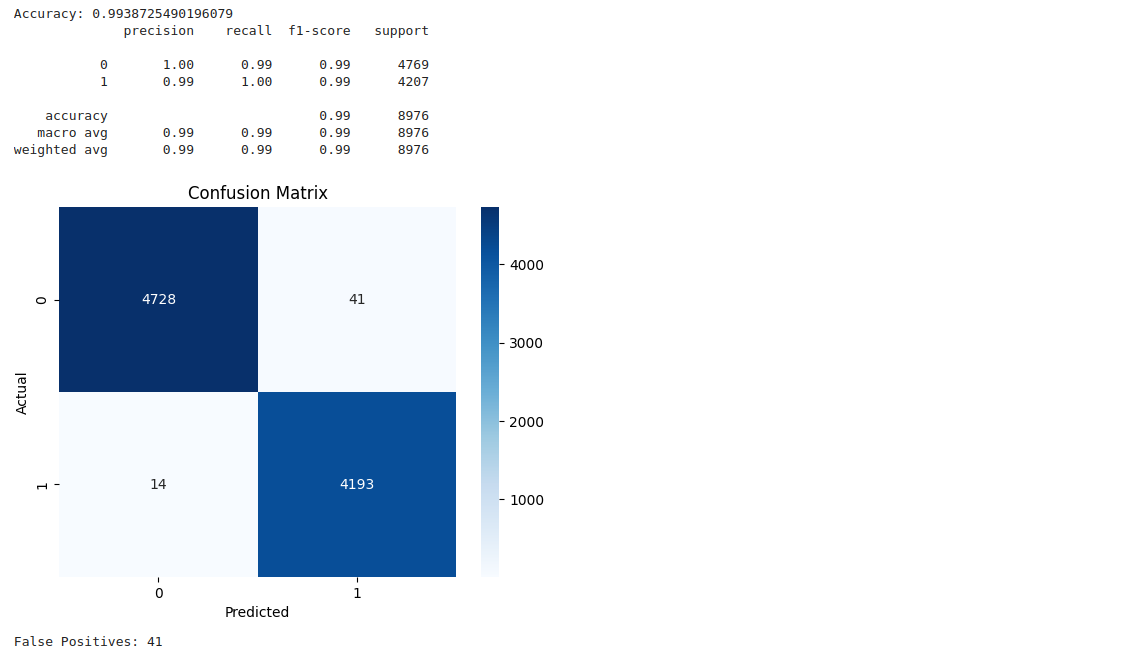
The number of false positives in Random Forest Classifier model is **46** and the accuracy is **99.05%**

**Step 6:**

Certainly! Let's use the Gradient Boosting Classifier as our machine learning model and train it. After training, we'll print the accuracy and classification report for the model.

****

The confusion matrix is used to calculate the number of false positives.

****

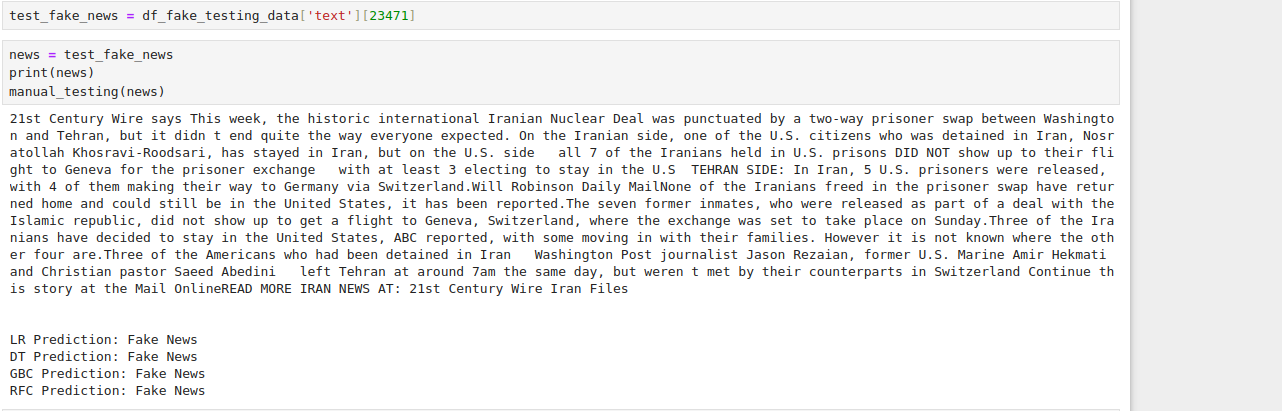
The number of false positives in the Random Forest Classifier model is **41** and the accuracy is **99.38%**

**Step 7:**

Create a function named manual\_testing( ) that eliminates stopwords and applies additional text preprocessing techniques to the testing data. Utilize all the machine learning models that have been trained thus far to classify whether the provided input text is categorized as Fake News or True News.

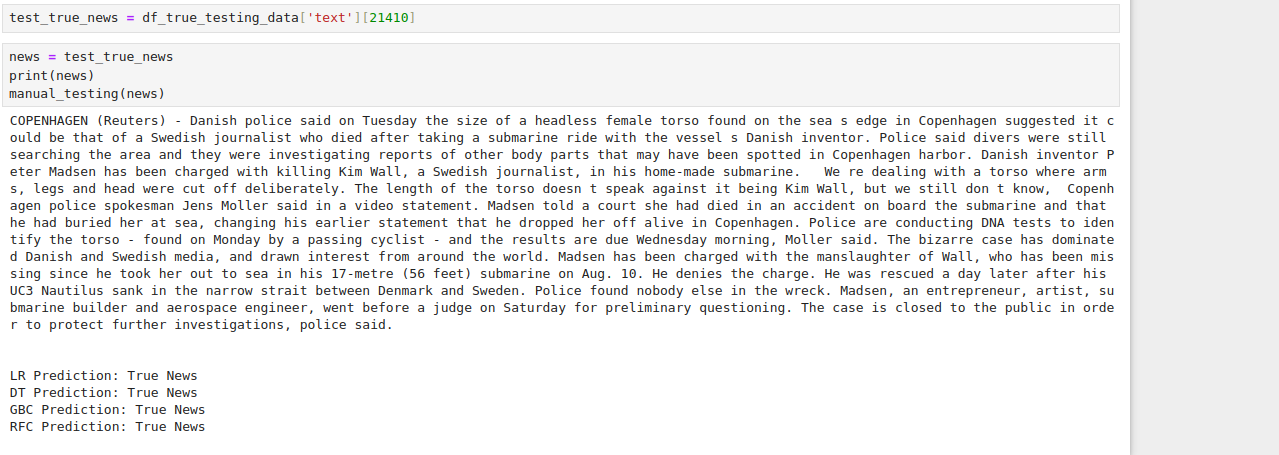
**Step 8:**

In order to evaluate the performance of the models we've developed, we will employ randomly generated fictitious news articles from the initial test dataset.

****

**Step 9:**

To assess the performance of the models developed, we will utilize a set of randomly selected, real news articles from the manual test data defined at the outset.

****

**INFERENCE:**

Among the various models developed for identifying fake news, the Decision Tree Classifier stands out with its remarkable performance. It achieves an impressive accuracy rate of 99.49% and boasts the lowest false positive count, which is only 25.