Lab 3 Report - Task 1

Classical Machine Learning & Data Conversion

1. Objective

The objective of this lab was to apply classical machine learning models to the **Fake and Real News Dataset** to evaluate their effectiveness on text data. The tasks included:

- Converting text to numerical features using TF-IDF
- Applying Linear Regression and Random Forest Regressor
- Evaluating performance using R^2, RMSE, and MAE
- Applying Logistic Regression for classification
- Converting structured JSON data to a custom TOON format

---

2. Dataset

The dataset consists of news articles labeled as:

- **0 – Fake**
- **1 – Real**
  A subset of 5000 samples was used for modeling. The text field was vectorized using TF-IDF with 5000 features and English stopword removal.

```
Dataset shape: (5000, 3)
```

	title	text	label
0	German Social Democrats face pressure over coa...	BERLIN (Reuters) - Germany s Social Democrats ...	1
1	U.S. diplomatic delays, Trump agenda snarl Ita...	ROME (Reuters) - Italy's preparations for host...	1
2	Trump taps Retired General Kelly to lead Homel...	WASHINGTON (Reuters) - Republican U.S. Preside...	1
3	Texas Republicans Cut Environmental Regulatio...	A disgusting black sludge is coming out of res...	0
4	Trump attacks FBI on leakers of Russia reports...	WASHINGTON (Reuters) - U.S. President Donald T...	1

```
Train size: 4000
Test size: 1000
```

---

3. Regression Models

Linear Regression

Linear Regression was trained to predict the binary label.
Performance showed moderate R^2 values, indicating partial explanatory power.

```
===== LINEAR REGRESSION RESULTS =====
R2: 0.3815
RMSE: 0.3932
MAE: 0.3069
```

Random Forest Regressor

Random Forest performed better than Linear Regression due to its ability to model non-linear relationships in textual features.
However, regression is not ideal for binary classification problems.

```
===== RANDOM FOREST RESULTS =====
R2: 0.9682
RMSE: 0.0891
MAE: 0.015
```

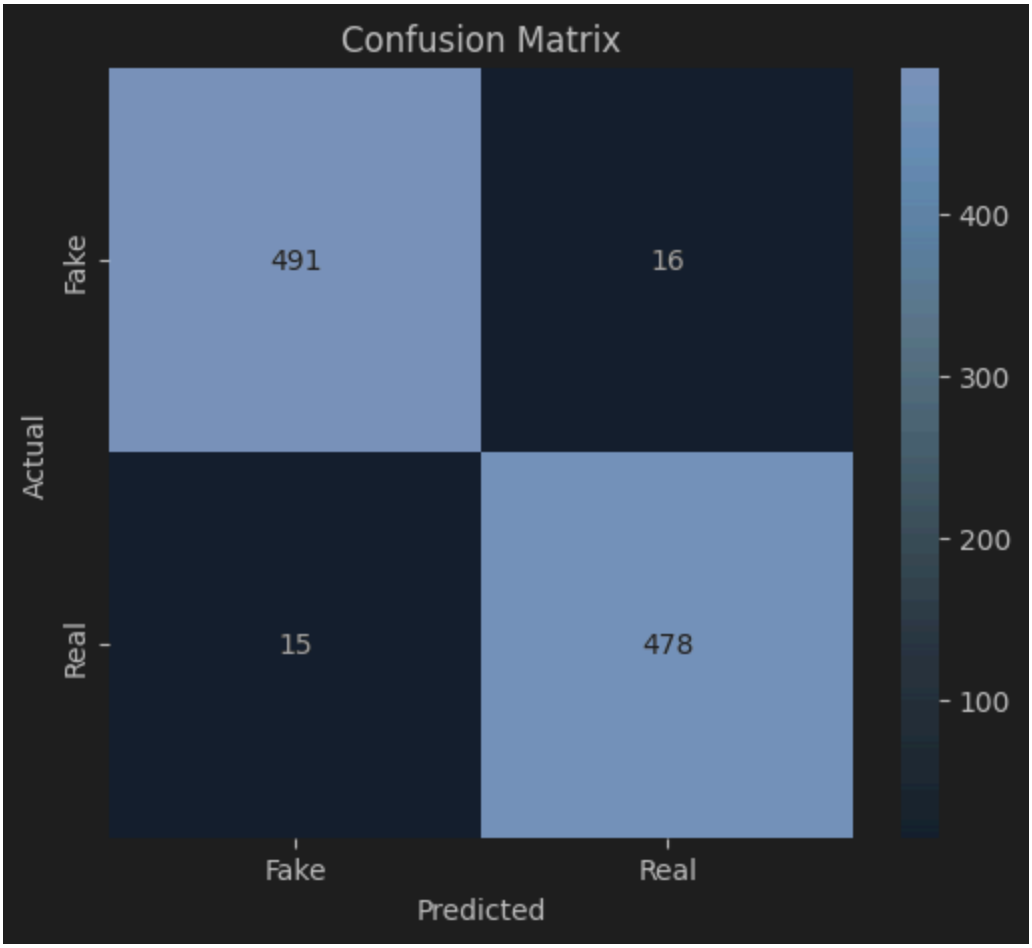
---

4. Classification Model

Logistic Regression was applied as a proper classification model.
Results showed high accuracy and strong precision-recall performance.
The confusion matrix indicated clear separation between fake and real news articles.
Classification proved significantly more suitable than regression for this task.

```
===== LOGISTIC REGRESSION (CLASSIFICATION) =====
Accuracy: 0.969

Classification Report:

              precision    recall  f1-score   support

           0       0.97      0.97      0.97       507
           1       0.97      0.97      0.97       493

    accuracy                           0.97      1000
```

```
   macro avg       0.97      0.97      0.97      1000
weighted avg       0.97      0.97      0.97      1000
```



---

5. TOON Format Conversion

The dataset was converted into a structured TOON format:
news[300]{label,text}:
0,Sample fake news text...
1,Sample real news text...

```
1,BERLIN (Reuters) - Germany s Social Democrats (SPD) faced pressure on
Wednesday to consider offering...
1,ROME (Reuters) - Italy’s preparations for hosting this year’s Group of Seven
major powers meetings h...
1,WASHINGTON (Reuters) - Republican U.S. President-elect Donald Trump on
Monday formally announced Ret...
0,A disgusting black sludge is coming out of residents  faucets in Crystal
City, Texas, and the people...
1,WASHINGTON (Reuters) - U.S. President Donald Trump
```

This demonstrated structured data transformation and formatting.

---

6. Conclusion

This lab demonstrated:

- TF-IDF effectively converts text into numerical form.
- Regression models can approximate binary labels but are not optimal.
- Logistic Regression is more appropriate and achieves superior performance.
- Proper model selection is critical in NLP tasks.
  The lab successfully integrated feature engineering, model evaluation, and data transformation within a classical NLP pipeline.

7. Git Link:-

- (Lab 3 - Task 1 - Git Repo)[https://github.com/MYTH-il/NLP-Lab/blob/master/lab3_task1.ipynb]