Assignment Title: Supervised Learning – Classification

Project Title: Fake News Detection

1. Reading Summary:

This week, I studied Supervised Learning – Classification and learned how machine learning algorithms can classify data into distinct categories.

- I learned the theory behind **Logistic Regression, Decision Trees, and Random Forests**.
- I understood how text features can be transformed into numeric vectors using **TF-IDF** for classification.
- I studied **evaluation metrics** like accuracy, precision, recall, F1-score, and confusion matrix to measure model performance.

2. Task Performed:

- Imported and explored the **Fake.csv** and **True.csv** datasets.
- Added a target column (`label`) to distinguish between **fake (1)** and **true (0)** news.
- Combined text columns (title + text) into a single feature for analysis.
- Split data into **training and testing sets** using train_test_split.
- Converted text data to numeric vectors using **TF-IDF Vectorizer**.
- Trained **Decision Tree** and **Random Forest** models.
- Applied **Logistic Regression** for classification comparison.
- Evaluated models using **accuracy, confusion matrix, and classification reports**.
- Visualized results with a **confusion matrix heatmap**.

3.Calculations and Analysis:

- Calculated mean, median, and mode for numeric columns (text_length, title_length, word_count).
- Found variance to measure the spread of data.
- Generated a correlation matrix and compared all numeric features with the target variable label.

4. Learning Outcome:

From this task, I learned :

- How to convert text data into numeric vectors using **TF-IDF**.· How to identify which features are most related to the target variable.
- How to train and evaluate **Decision Tree, Random Forest, and Logistic Regression** models.
- How to interpret **accuracy, precision, recall, F1-score, and confusion matrices**.
- Why Random Forest can perform better than Logistic Regression on text classification tasks.
- How to compare multiple models and select the best one for deployment.

5. Challenges Faced:

- Initially, the dataset had **only one class** when loading files incorrectly, causing model training errors.
- TF-IDF vectorization needed careful tuning to handle a large number of features.
- Balancing dataset and correctly labeling Fake vs True news was critical.
- Visualizing results clearly required adjusting heatmap parameters.

6. GitHub Repository Link:

GitHub repository link here:
https://github.com/AlmasMalik66/DataScience-AI-Assignment