Python Programming

Machine Learning Assignment

Predict whether a news article is **Fake** or **Real** using text classification techniques. This assignment demonstrates the power of **ensemble learning** using a **Voting Classifier** with models like Logistic Regression, Decision Tree.

Dataset Information:

Dataset Name: Fake News Dataset
Columns include:

- `title` – Title of the news article
- `text` – Main content of the article
- `label` – 0 = Fake, 1 = Real

Part 1: Data Preprocessing

1. Load the dataset using Pandas
2. Drop null values and select useful columns (`title` or `text`)
3. Convert the target variable (`label`) to binary (0 or 1)

Part 2: Feature Extraction

1. Use **TF-IDF Vectorization** to convert text into numerical features

Part 3: Model Training

1. Train individual models:
   - Logistic Regression
   - Decision Tree Classifier
2. Combine them using:
   - **Hard Voting** (majority rule)
   - **Soft Voting** (average predicted probabilities)

Part 4: Evaluation

1. Compare accuracies of all models
2. Display confusion matrices
3. Soft vs hard voting

Note : Dataset is divided into 2 parts as fake.csv and true.csv

1. Load both CSV files

Each CSV represents a class:

- `fake.csv` contains **fake news articles**
- `true.csv` contains **real news articles**

2. Add a 'label' column to both

We need to **combine** the two datasets, so we must label them first:

- 0 = Fake
- 1 = Real

3. Combine the datasets

Now concatenate them into one DataFrame:

4. Use only the relevant columns

You may use either `title`, `text`, or both combined.