Name: Minahil Irfan

Assignment 6: Supervised Learning Classification

Overview:
This assignment focused on building baseline classification models using Logistic Regression and Random Forest to classify tweet sentiments. The aim was to understand linear and non-linear classification approaches and compare model performance.

Key Concepts:

1. Logistic Regression:

- Supervised learning algorithm for classification.
- Predicts probabilities using the sigmoid function.
- Assigns input to classes based on a threshold (commonly 0.5).
- Works well with linear relationships between features and target.
- Strengths: Fast, interpretable.
- Weaknesses: Limited for non-linear data.

2. Random Forest:

- Ensemble method that builds multiple decision trees.
- Aggregates predictions to improve accuracy and stability.
- Reduces overfitting and handles non-linear relationships.
- Strengths: Handles complexity, reduces overfitting.
- Weaknesses: Slower, less interpretable.

3. Model Comparison:

- Accuracy and confusion matrices are used to determine which model better classifies the target variable.

Implementation Steps:

1. **Load Dataset:**
    - Loaded twitter_sentiment_cleaned.csv into Pandas.
    - Renamed columns for clarity.
2. **Data Preprocessing:**
    - Dropped missing values.
    - Sampled 5000 rows for faster computation.
    - Separated features (tweet) and target (sentiment).
3. **Feature Engineering:**
    - Converted text data into numeric features using CountVectorizer (max 2000 features).
4. **Train/Test Split:**
    - Split data into training (80%) and testing (20%) sets.
5. **Model Training:**

- Logistic Regression trained using Scikit-Learn.
- Random Forest trained with 100 estimators.

6. **Evaluation:**
   - Calculated accuracy for both models on the test set.
   - Compared performance to identify which model performed better.

Results:

Logistic Regression Accuracy: 0.xxx
Random Forest Accuracy: 0.xxx

- Higher accuracy indicates better classification performance.
- Comparison helps choose the most suitable baseline model for sentiment prediction.

Insights & Conclusion:

- Logistic Regression is faster and interpretable but may struggle with non-linear patterns.
- Random Forest is robust for complex relationships but less interpretable.
- Accuracy metrics guide the choice of baseline classification model.

Project Milestone:
Built and compared the first baseline classification models (Logistic Regression and Random Forest).

Github: DataScience-AI/assignment 6 at main · MinahilIrfan98/DataScience-AI