IMDB Sentiment Analysis – Machine Learning Assignment

1. Overview

The objective of this assignment was to build a machine learning model that can classify movie reviews from the IMDB dataset as either positive or negative. The dataset consists of 50,000 balanced movie reviews. This project covered text preprocessing, feature extraction, model building, evaluation, and pipeline development.

2. Data Loading & Preprocessing

The dataset was provided in Excel format. After loading it into a Pandas DataFrame, the structure was explored to check for class balance. Preprocessing steps included converting all text to lowercase, removing punctuation, eliminating stopwords, and applying lemmatization to reduce words to their base forms. The clean reviews were then split into training (80%) and testing (20%) sets.

3. Feature Extraction

TF-IDF Vectorization was applied to transform the text data into numerical form. The vectorizer was set to use a maximum of 5000 features and an n-gram range of (1,2) to capture both unigrams and bigrams.

4. Model Building & Evaluation

Three models were trained and evaluated: Logistic Regression, Naive Bayes, and Linear SVM. Performance was measured using accuracy, precision, recall, and F1-score. The Logistic Regression model achieved the highest overall performance.

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.89	0.89	0.89	0.89
Naive Bayes	0.85	0.85	0.85	0.85
SVM	0.88	0.88	0.88	0.88

5. Pipeline & Optimization

An end-to-end scikit-learn pipeline was created combining the TF-IDF vectorizer and the classifier. This allowed for cleaner code, easier experimentation, and consistent preprocessing for training and testing.

6. Inference & Results

The final Logistic Regression pipeline was tested on 5 unseen reviews. The predictions aligned well with expected sentiments, demonstrating the model's generalization capability.
Final model accuracy on the test set: 0.89.

7. Learning Points

- Text preprocessing significantly improves sentiment classification performance.
- TF-IDF is effective in representing movie review text.
- Logistic Regression works well for balanced datasets like IMDB.
- Pipelines make the workflow more maintainable and reproducible.