IMDB SENTIMENT ASSIGNGMENT

NAME: EMMANUEL ANTWI OSEI
STUDENT NUMBER: C0932615

IMDB Sentiment Analysis Report

1. Objective

The goal was to perform sentiment classification on movie reviews from the IMDB dataset using Natural Language Processing (NLP) and machine learning techniques.

2. Data Preparation

Dataset: IMDB Dataset.csv (from Kaggle) containing movie reviews labeled as positive or negative.

Processing Steps:

1. Imported dataset into Pandas for exploration.
2. Preprocessed text using NLTK (tokenization, stopword removal, stemming/lemmatization).
3. Converted text into numerical features using TF-IDF Vectorization.

3. Model Development

Algorithms Used:

- Logistic Regression (baseline model).
- Support Vector Machine (SVM) via GridSearchCV for hyperparameter tuning.
- Pipeline: Combined preprocessing (TF-IDF) and classifier into a single scikit-learn Pipeline for cleaner training and evaluation.

4. Model Evaluation

Logistic Regression Results:

- Accuracy: 89.22%
- Precision: 89%
- Recall: 89%
- F1-score: 89%

SVM Results:

- Achieved competitive accuracy after hyperparameter tuning.
- GridSearchCV identified optimal parameters for better generalization.

5. Key Insights

- TF-IDF vectorization worked effectively in capturing important terms for sentiment classification.
- Logistic Regression provided strong baseline performance with minimal tuning.
- SVM showed potential for slightly improved results after parameter optimization.

6. Conclusion

The project successfully built and evaluated models that achieved high accuracy in predicting sentiment from movie reviews. Logistic Regression was simple yet effective, while SVM offered scope for marginal improvements with more tuning and computational resources.