**Name: ROOP CHAND (65554)**

**Instructor: DR. FIDA HUSSAIN**

**Course: Data Science**

**Project: CCP**

**Marks: 10**

SMS Spam Classifier – One-Page Report

## Dataset:

• 5,572 SMS messages.  
• Features: raw text messages (vectorized into ~3,000 TF-IDF features).  
• Target: binary classification — ham (0) vs. spam (1).  
• Class balance: 4,825 ham (86.6%), 747 spam (13.4%) → imbalanced dataset.

## Preprocessing:

• Labels encoded (ham=0, spam=1).  
• Text converted into TF-IDF vectors (unigrams + bigrams, max 3,000 features).  
• Stratified 75/25 train-test split.

## Models:

• Baseline: Logistic Regression (max\_iter=500).  
• Tuned: Gradient Boosting (200 estimators, learning rate=0.05, depth=3).

## Metrics:

|  |  |  |
| --- | --- | --- |
| Model | Accuracy | Precision (macro) |
| Logistic Regression | 0.9727 | 0.9847 |
| Gradient Boosting | 0.9591 | 0.9707 |

## Loss Curves:

• Gradient Boosting train/test log-loss curves were plotted.  
• Both losses decreased and stabilized after ~50–100 iterations.  
• Test loss remained slightly higher than train loss, but no severe overfitting observed.

## Insights:

• Logistic Regression outperformed Gradient Boosting, achieving higher accuracy and precision — showing that linear models + TF-IDF work very well for SMS spam detection.  
• Gradient Boosting still performed competitively but added complexity with no accuracy gain.  
• Class imbalance (87% ham vs. 13% spam) may bias results toward the majority class, though macro-precision remains high (>0.97).  
• Next steps: experiment with recall-focused metrics (F1, confusion matrix), or apply class weighting/oversampling (SMOTE) to ensure spam detection is prioritized.