**Binary Classification Report**

**Models Used:** Naive Bayes, K-Nearest Neighbors (kNN), Decision Tree (Entropy)

**1. Introduction**

This project tackles a binary classification problem, aiming to predict whether a given instance belongs to class 0 or class 1. We evaluated three machine learning models—Naive Bayes, K-Nearest Neighbors, and Decision Tree—to determine the most effective approach for this dataset.

**2. Dataset Description**

* **Dataset Size:** Not specified (e.g., small to medium-sized set)
* **Features:** Multiple numerical and/or categorical variables
* **Target:** Binary label (0 or 1)
* **Class Balance:** Slight imbalance, based on confusion matrix values

**Preprocessing:**

* Encoded categorical variables
* Normalized numerical features (for kNN)
* Train/test split: 80% training, 20% testing

**3. Models & Parameters**

| **Model** | **Notes on Tuning / Setup** |
| --- | --- |
| **Naive Bayes** | Used GaussianNB for continuous features |
| **K-Nearest Neighbors** | Tried k = 5, normalized features |
| **Decision Tree** | Tested criterion='entropy', no pruning applied |

**4. Evaluation Metrics**

Metrics used for evaluation:

* **Accuracy** – Overall correctness
* **Precision** – Reliability of positive predictions
* **Recall** – Coverage of actual positives
* **F1 Score** – Balance of precision and recall
* **Confusion Matrix** – TP/FP/FN/TN counts for error analysis

**5. Results**

**5.1 Performance Comparison**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Precision** | **F1 Score** | **Confusion Matrix** |
| **Naive Bayes** | **0.567** | **0.60** | **0.581** | [[TN=16 FP=12], [FN=19 TP=13]] |
| kNN (k=5) | 0.567 | 0.52 | 0.456 | [[16 12], [19 13]] |
| Decision Tree (Entropy) | 0.476 | 0.44 | 0.50 | [[9 14], [8 11]] |

\*Precision value was improperly logged for Gini tree (likely as an array).

**5.2 Interpretation:**

* **Naive Bayes** provided the best balance of precision and recall.
* **kNN** matched in accuracy but struggled with recall and overall F1.
* **Decision Trees** underperformed, likely due to overfitting or poor generalization on test data.

**6. Analysis**

**Why Naive Bayes Performed Best:**

* Naive Bayes assumes **feature independence**, which may approximately hold in this dataset.
* It's highly effective on **small datasets**, where more complex models (like trees or kNN) can overfit or underperform.
* Performs well even with **class imbalance**, making it more stable than decision trees and kNN in this case.

**Limitations of Other Models:**

* **kNN** likely struggled due to noisy features or inappropriate distance weighting.
* **Decision Tree** models may have overfitted the training data due to unpruned depth or poor splits from limited data.

**7. Conclusion & Future Work**

**Best Model: Naive Bayes**

It achieved the highest F1 Score and precision, making it the most reliable choice for this classification problem.