**Performance Evaluation Report**

**1. Introduction**

This report evaluates the performance of multiple machine learning models, including **Random Forest (RF), Decision Tree (DT), XGBoost, Artificial Neural Network (ANN), and BERT**, on a given classification problem. The key metrics analyzed are **accuracy, precision, recall, F1-score, and AUC-ROC score**.

**2. Model Performance Overview**

**Random Forest (RF)**

* **Accuracy**: 28.05%
* **Precision**: 27% (macro avg)
* **Recall**: 28% (macro avg)
* **F1-Score**: 27% (macro avg)
* **Observations**:
  + Performed poorly across all classes.
  + Model misclassified many instances, leading to an overall weak performance.

**Decision Tree (DT)**

* **Accuracy**: 32.83%
* **Precision**: 33% (macro avg)
* **Recall**: 33% (macro avg)
* **F1-Score**: 33% (macro avg)
* **Observations**:
  + Improved performance compared to RF, but still underwhelming.
  + Overall, the model lacked generalization.

**XGBoost**

* **Accuracy**: 34.49%
* **Precision**: 34% (macro avg)
* **Recall**: 34% (macro avg)
* **F1-Score**: 33% (macro avg)
* **Observations**:
  + Performed slightly better than DT and RF.

**Artificial Neural Network (ANN)**

* **Training Accuracy**: 94.2%
* **Validation Accuracy**: 23.31%
* **Loss (Train / Validation)**: 0.1383 / 4.0864
* **Observations**:
  + The model shows extreme **overfitting**, with a huge performance gap between training and validation accuracy.
  + Indicates that the ANN learned training patterns well but failed to generalize.

**BERT Model**

* **Accuracy**: 33.96% (Training), 32.05% (Validation)
* **Loss**: 1.1003 (Training), 1.0986 (Validation)
* **Observations**:
  + Performance remains close to **random chance (33%)** for a three-class classification problem.

**Optimized Random Forest (RF)**

* **Accuracy**: 36.34%
* **Precision**: 36% (macro avg)
* **Recall**: 35% (macro avg)
* **F1-Score**: 28% (macro avg)
* **Observations**:
  + Outperformed all other models in terms of accuracy.
  + Class 0 showed significant improvement in recall (80%), but Class 1 had very low recall (3%), indicating poor performance for certain categories.

**Confusion Matrix Analysis**

[[287 195 369]

[361 106 306]

[353 160 287]]

* A significant number of misclassifications across all three classes.
* High confusion between **Class 0 and Class 2**, leading to lower precision and recall.
* Class 1 has the lowest correctly classified instances.

**AUC-ROC Score**

* **0.431** (poor discrimination capability).