Summary Report: Machine Learning Classification Project

Problem Description:
The objective of this project is to build a classification model that predicts target labels using given features. The dataset was preprocessed, encoded, balanced, and used to train multiple machine learning models to evaluate performance.

Methodology:
1. Loaded the dataset using pandas.
2. Checked for correlation between features and target variable; removed low-correlation features.
3. Encoded categorical features using appropriate encoding techniques.
4. Dropped missing values to ensure data consistency.
5. Checked for imbalanced data and applied SMOTE to balance the dataset.
6. Split the data into training (80%) and testing (20%) using train-test split.
7. Trained four classification models — Logistic Regression, Random Forest, XGBoost, and Gradient Boosting.
8. Evaluated models using accuracy, precision, recall, F1-score, ROC curve, and AUC score.

Results:
Logistic Regression:
Accuracy: 0.86
Precision (Class 0/1): 0.93 / 0.55
Recall (Class 0/1): 0.89 / 0.69
F1-score (Class 0/1): 0.91 / 0.61

Random Forest:
Accuracy: 0.89
Precision (Class 0/1): 0.94 / 0.64
Recall (Class 0/1): 0.92 / 0.72
F1-score (Class 0/1): 0.93 / 0.68

Gradient Boosting:
Accuracy: 0.88
Precision (Class 0/1): 0.95 / 0.62
Recall (Class 0/1): 0.91 / 0.74
F1-score (Class 0/1): 0.93 / 0.68

Insights:
- All models achieved good accuracy (above 85%), indicating strong predictive capability.
- Random Forest performed slightly better overall with the highest balanced precision and recall.
- Gradient Boosting showed consistent results, handling class imbalance effectively.
- The recall score of 0.75 indicates the model is effectively identifying positive cases.

Conclusion:
Based on model comparison, Random Forest provides the most reliable and balanced performance across all metrics. Future improvements could include hyperparameter tuning and testing additional ensemble models.