TASK 4

SUBMITTED BY:

Anshi Debata

Project Summary :

### **Data Preprocessing:**

****Missing Values:**** Numerical columns were filled with median values, and categorical columns were filled with the mode.

****Categorical Encoding:**** Label Encoding was applied to all categorical features.

****Feature & Target Separation:**** The dataset was split into features (X) and target (y), followed by an 80-20 train-test split.

### **Models Trained by:**

1. **Decision Tree Classifier**
2. **Random Forest Classifier**
3. **Gradient Boosting Classifier**

Each model was trained using default hyperparameters and evaluated on the test set.

### **Evaluation Metrics:**

**Each model was assessed using:**

* **Accuracy**
* **Precision (weighted)**
* **Recall (weighted)**
* **Classification Report**
* **Feature Importance Analysis**

**Model Performance Summary:**

| Model | Accuracy | Precision | Recall |
| --- | --- | --- | --- |
| Decision Tree | 0.9742388758782201 | 0.9742828417054793 | 0.9742388758782201 |
| Random Forest | 0.9765807962529274 | 0.9765619791194287 | 0.9765807962529274 |
| Gradient Boosting | 0.9754098360655737 | 0.9755768172517093 | 0.9754098360655737 |

* **Random Forest** and **Gradient Boosting** typically outperformed the Decision Tree in terms of generalization and stability.
* Feature importance plots revealed the most influential variables in model predictions.

### **Conclusion**

Among the three models, **Random Forest** (especially after tuning) provided the best balance between performance metrics and interpretability. It was chosen as the final model for deployment or further usage.