Employee Attrition Prediction Model Performance Report

1. Objective:

To develop and evaluate a classification model capable of predicting employee attrition based on historical HR data, enabling proactive retention efforts.

2. Models Selected:

- **Model:** Logistic Regression *and Decision Tree*
- **Rationale:** Logistic Regression was chosen for its interpretability and effectiveness as a baseline classification model and *Decision Trees generate a set of explicit if-then rules that are relatively easy for humans (like HR professionals or managers) to understand. This allows for clear insights into the potential factors or combinations of factors the model identifies as leading to attrition.*

3. Data & Methodology:

- The dataset contained 1000 employee records with various demographic, job-related, and satisfaction features.
- The data was preprocessed, including handling categorical features (One-Hot Encoding) and scaling numerical features (StandardScaler).
- The data was split into a 75% training set and a 25% testing set, stratified by the 'Attrition' target variable to maintain class proportions.

4. Performance Evaluation:

4.1 Logistic Regression

- **Parameters:** random_state=42, max_iter=1000, class_weight='balanced'
- **Overall Accuracy: 50.80%**
    - *Interpretation:* The model correctly predicted whether an employee would leave or stay for approximately 50.8% of the employees in the test set, which is barely better than random guessing.
- **Confusion Matrix:**

Predicted: NO Predicted: YES

Actual: NO [[110]] [[93]] <-- TN, FP

Actual: YES [[29]] [[18]] <-- FN, TP

    - **Interpretation:**
        - TN: 110 employees correctly predicted to stay.
        - TP: 18 employees correctly predicted to leave.
        - FP: 93 employees incorrectly predicted to leave (High False Alarms).
        - FN: 29 employees incorrectly predicted to stay (Missed Attrition).
- **Class-Specific Metrics (for Attrition = YES - Class 1):**

- **Precision: 16.22%** *(Approx. 18 / (18 + 93) - using 0.16 from report)*
    - *Interpretation:* Very low; when the model predicts attrition, it's correct only ~16% of the time.
- **Recall (Sensitivity): 38.30%** *(Approx. 18 / (18 + 29) - using 0.38 from report)*
    - *Interpretation:* Identified only ~38% of employees who actually left.
- **F1-Score: 22.78%** *(Approx. using 0.23 from report)*
    - *Interpretation:* Very low score, indicating poor performance for the attrition class.

4.2 Decision Tree

- **Parameters:** random_state=42, max_depth=5, class_weight='balanced'
- **Overall Accuracy: 52.40%**
    - *Interpretation:* Slightly better than Logistic Regression but still very low, correctly predicting ~52.4% of outcomes.
- **Confusion Matrix:**
- Predicted: NO Predicted: YES
- Actual: NO [[108]] [[95]] <-- TN, FP
- Actual: YES [[24]] [[23]] <-- FN, TP
    - **Interpretation:**
        - TN: 108 employees correctly predicted to stay.
        - TP: 23 employees correctly predicted to leave.
        - FP: 95 employees incorrectly predicted to leave (High False Alarms).
        - FN: 24 employees incorrectly predicted to stay (Missed Attrition).
- **Class-Specific Metrics (for Attrition = YES - Class 1):**
    - **Precision: 19.49%** *(Approx. 23 / (23 + 95) - using 0.19 from report)*
        - *Interpretation:* Very low; when the model predicts attrition, it's correct only ~19.5% of the time.
    - **Recall (Sensitivity): 48.94%** *(Approx. 23 / (23 + 24) - using 0.49 from report)*
        - *Interpretation:* Identified ~49% of employees who actually left – better than Logistic Regression but still misses about half.
    - **F1-Score: 27.88%** *(Approx. using 0.28 from report)*
        - *Interpretation:* Low score, better than Logistic Regression but still indicates poor overall performance for the attrition class.

4.3 Model Comparison

Metric	Logistic Regression	Decision Tree	Winner	Notes
Overall Accuracy	50.80%	52.40%	Decision Tree	Both are very low, close to random chance.
Precision (Attrition)	16.2%	19.5%	Decision Tree	Both extremely low (many false alarms).
Recall (Attrition)	**38.3%**	**48.9%**	**Decision Tree**	Decision Tree identifies more actual leavers (less missed).
F1-Score (Attrition)	22.8%	27.9%	Decision Tree	Both very low, indicating poor balance.

- The **Decision Tree** shows marginally better performance across most metrics, notably achieving higher Recall for the attrition class (identifying almost 49% of leavers vs. 38% for Logistic Regression).
- However, both models suffer from extremely low Precision, meaning their predictions of attrition are highly unreliable (mostly false positives).
- Neither model demonstrates strong predictive capability with the current setup.

5. Summary:

Both the Logistic Regression and Decision Tree models, despite using balanced class weights, exhibit very low predictive performance on the test data, with accuracies barely above 50%. The Decision Tree performs slightly better, particularly in identifying a larger proportion (Recall ≈ 49%) of employees who actually leave compared to Logistic Regression (Recall ≈ 38%). However, this comes at the cost of extremely poor Precision (≈ 19.5% for Decision Tree, ≈ 16.2% for Logistic Regression), indicating that a large majority of positive attrition predictions are incorrect (false alarms).

The low F1-scores for the attrition class (0.23 for Logistic Regression, 0.28 for Decision Tree) highlight that neither model is currently effective at reliably predicting employee turnover. This suggests potential issues with:

a) **Feature Predictiveness:** The available features may not contain strong enough signals to accurately predict attrition.

b) **Model Complexity/Tuning:** The simple models used might not capture complex relationships, or more extensive hyperparameter tuning is needed.

c) **Data Quality/Quantity:** Issues not apparent in initial checks or simply insufficient data might be limiting factors.

Recommendation: Significant improvement is needed. Focus should be on **feature engineering** (creating new, more predictive variables), exploring more **complex models** (e.g., RandomForest, Gradient Boosting, XGBoost), performing rigorous **hyperparameter tuning**, and potentially acquiring **additional relevant data** (e.g., engagement survey details, manager feedback scores, more granular compensation data) before deploying any model for practical use in predicting attrition. The current models should **not** be used for decision-making due to their low reliability.