Intern Dropout Prediction Report

Project Workflow

- 1. Dataset Creation: Generated a synthetic dataset of 50 intern records, including performance and behavior metrics.
- 2. Data Loading & Exploration: Loaded the dataset, reviewed the structure, types, and class distribution.
- 3. Preprocessing: Verified no missing values and encoded target column (Dropout) as binary values (Yes → 1, No → 0).
- 4. Data Visualization: Used seaborn/matplotlib to visualize attendance, task scores, and dropout distributions.
- 5. Model Building: Trained a Decision Tree Classifier to predict dropout based on intern features.
- 6. Model Evaluation: Calculated accuracy and classification report. Analyzed feature importance from the trained model.
- 7. Reporting: Summarized the findings and potential data-cleaning techniques.

Objective

The aim of this project was to analyze internship performance data (attendance, task scores, behavior, communication) and build a machine learning model to predict the likelihood of an intern dropping out.

Dataset Summary

- Total Records: 50
- Features:
  - Attendance_Rate (0–100%)
  - Avg_Task_Score (0–100)
  - Behavior_Score (1–10)
  - Communication_Score (1–10)
- Target Variable: Dropout (Yes or No, encoded as 1 or 0)

✅No null values were present in this dataset.

Model Training

A Decision Tree Classifier was trained on the dataset. It was selected for its simplicity and ability to reveal which features drive decisions.

- Train-Test Split: 80/20
- Target: Predict if an intern will drop out (1) or not (0)

❓ Feature Importance (Results)

The chart below shows which features the model found most useful:

Top 3 Features:
✅Attendance Rate — Most important predictor
✅Behavior Score — Second highest influence
✅Avg Task Score — Also a key factor

Least impactful:
❌Communication Score

❓ Handling Missing Values (If Present)

Although this dataset had no missing values, here are standard methods to handle them:

1. Mean / Median Imputation:
   For numeric columns – replace missing values with the average or median.

2. Mode Imputation:
   For categorical values – replace with the most frequent value.

3. Model-Based Imputation:
   Use techniques like KNN Imputer or Iterative Imputer to fill in values based on other features.

✅Conclusion

The machine learning model revealed that interns with:
- Low attendance
- Poor behavior score
- Weak task performance
are more likely to drop out.

These insights can help HR teams monitor at-risk interns early and apply interventions.