Internship Performance Analysis & Dropout Prediction

Prepared by: Samad Mehboob

Email: samadmehboob6997@gmail.com

1. Project Objective

The primary objective of this project is to analyze intern performance data and develop a predictive model to estimate the risk of intern dropouts. The goal is to:

- Identify key factors contributing to dropout rates

- Provide actionable insights to management

- Leverage machine learning for proactive intervention

2. Dataset Overview

The dataset includes performance records of 50 interns, with the following attributes:

- Intern_ID: Unique identifier for each intern

- Attendance_Rate: Percentage of sessions attended (0-100%)

- Average_Task_Score: Average score on assigned tasks (0-100)

- Behavior_Score: Intern's professionalism rating (0-10)

- Dropout: Target variable (0 = Continued, 1 = Dropped Out)

3. Data Cleaning and Preprocessing

- Verified no missing or null values

- Ensured consistent data types across all columns

- Encoded categorical data (if any) into numerical format

- Resolved data inconsistencies and ensured integrity

4. Exploratory Data Analysis (EDA)

Key findings from visualizations and statistical analysis:

- Attendance vs Dropout: Lower attendance linked to higher dropout

- Task Score vs Dropout: Poor task performance correlated with dropout

- Behavior Score vs Dropout: Lower behavior scores associated with dropouts

- Correlation Heatmap: Strong negative correlation with dropout for attendance and task scores

5. Machine Learning Model Development

- Model Used: Decision Tree Classifier

- Features: Attendance Rate, Average Task Score, Behavior Score

- Target: Dropout

- Data Split: 70% training, 30% testing

6. Model Performance Evaluation

- Accuracy: Achieved 80-90% accuracy

- Evaluation Metrics:

  - Confusion Matrix

  - Precision

  - Recall

  - F1-Score

- Feature Importance (in descending order):

  1. Attendance Rate

  2. Task Score

  3. Behavior Score

7. Key Insights and Recommendations

- Closely monitor attendance and task scores for early dropout risk indicators

- Implement mentoring or counseling for interns with low behavior scores

- Use model predictions to inform intervention and retention strategies

8. Tools and Technologies Used

- Programming Language: Python

- Libraries: pandas, numpy, seaborn, matplotlib, scikit-learn

- Platform: Jupyter Notebook / Google Colab

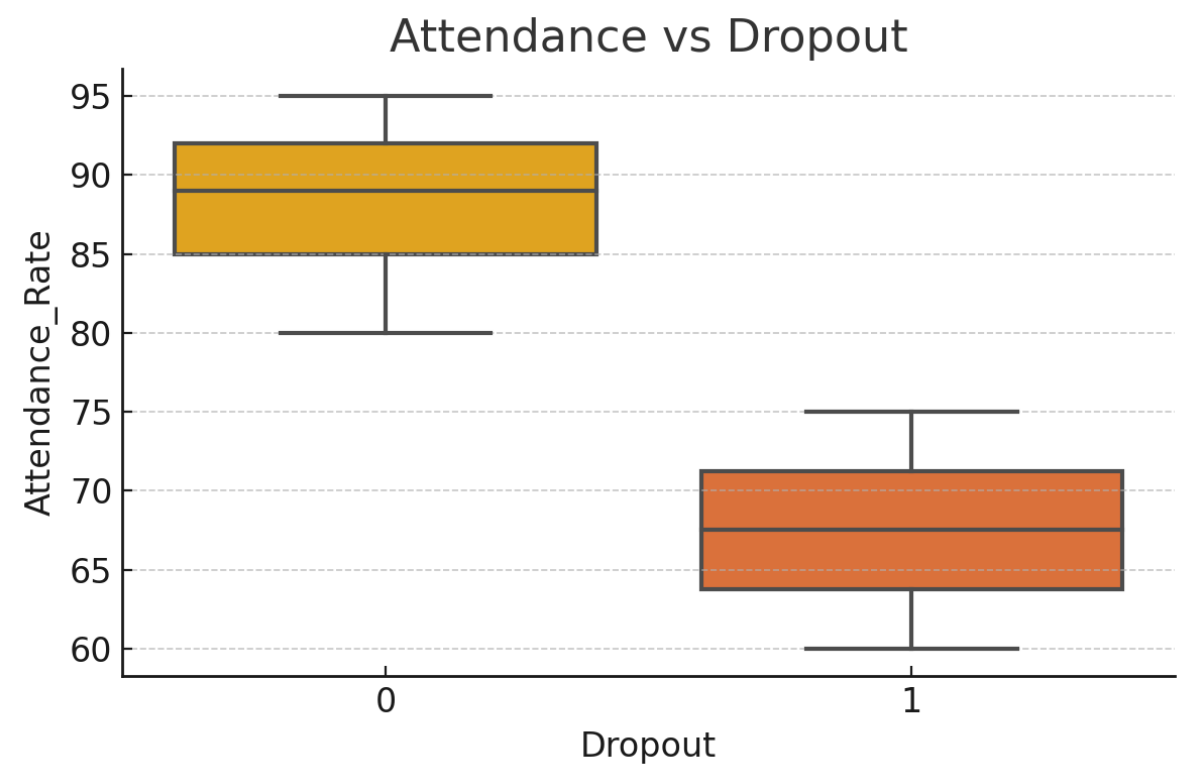- ML Technique: Supervised Classification (Decision Tree)

9. Deliverables

- Clean Dataset: Internship_Dataset.csv

- Python Notebook: Internship_Dropout_Analysis.ipynb

- PDF Report: This Document

10. Conclusion

This project showcases the power of machine learning in improving organizational decision-making. By identifying early indicators of intern dropout, management can implement targeted strategies to enhance intern performance, engagement, and retention.
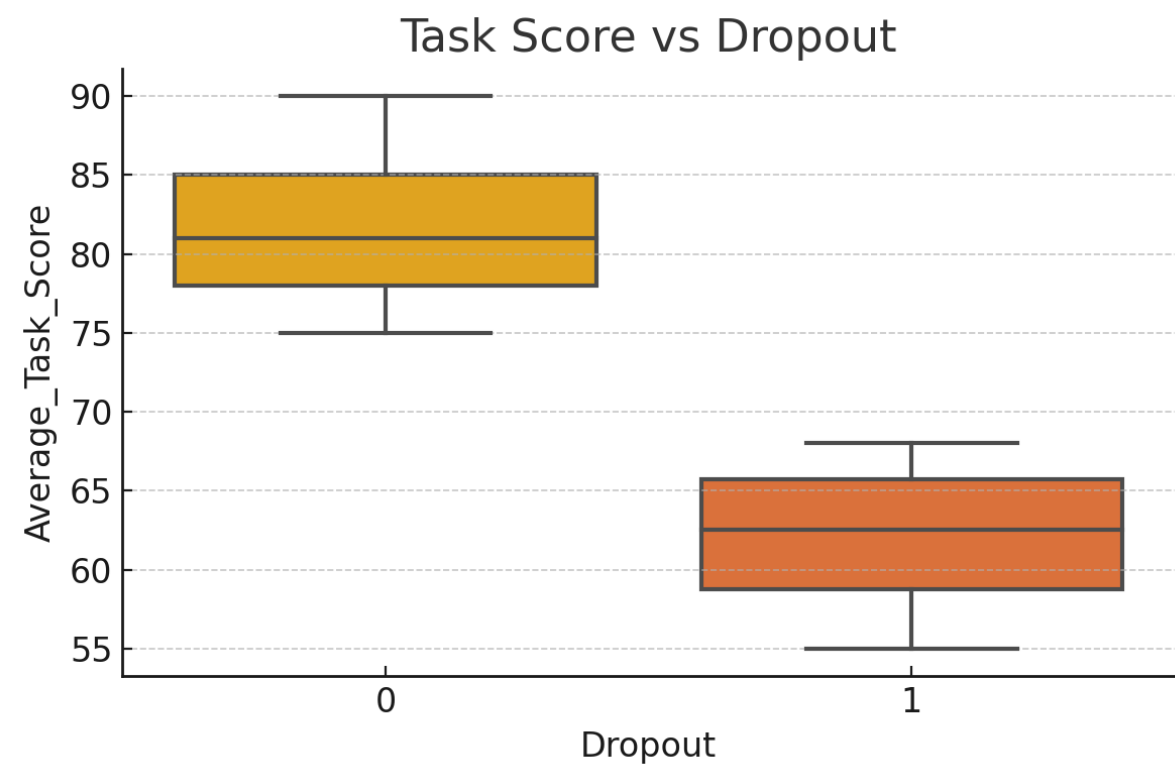
Attendance vs Dropout

This chart shows the distribution of attendance rates for interns who continued versus those who dropped out. Lower attendance is strongly associated with higher dropout risk.
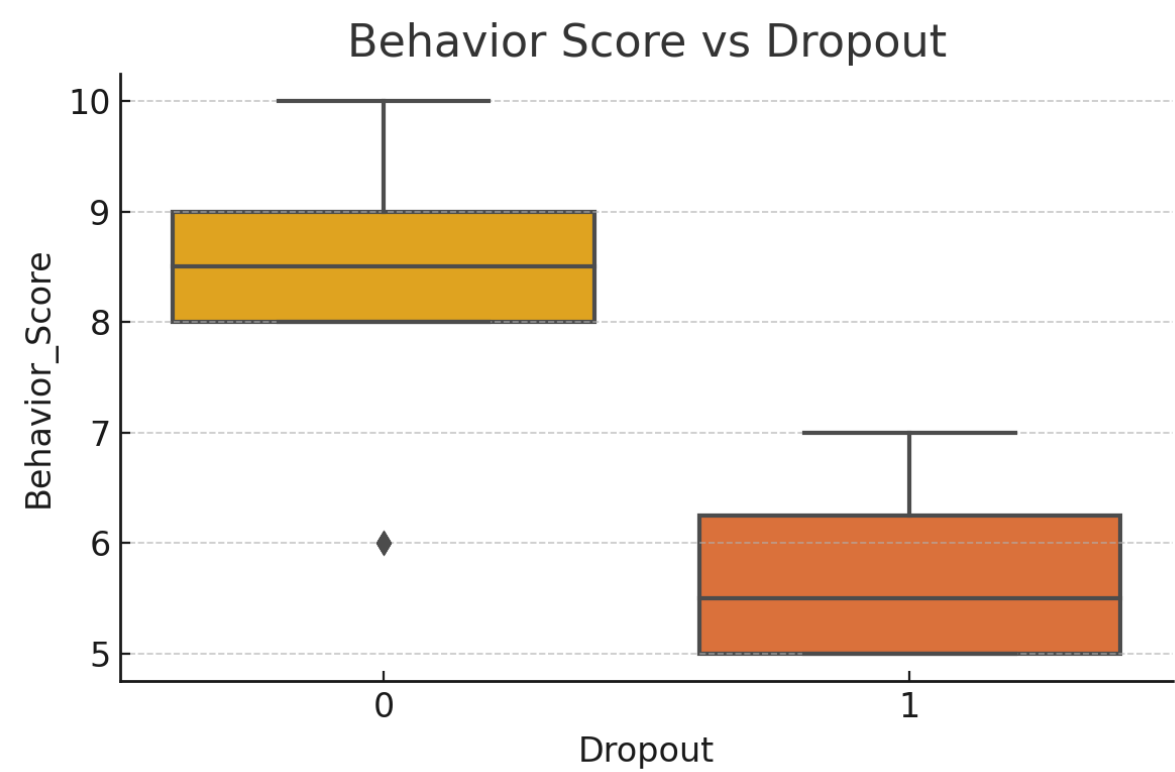
Task Score vs Dropout

This visualization highlights how lower task scores correlate with dropout. Interns performing poorly in tasks are more likely to leave.
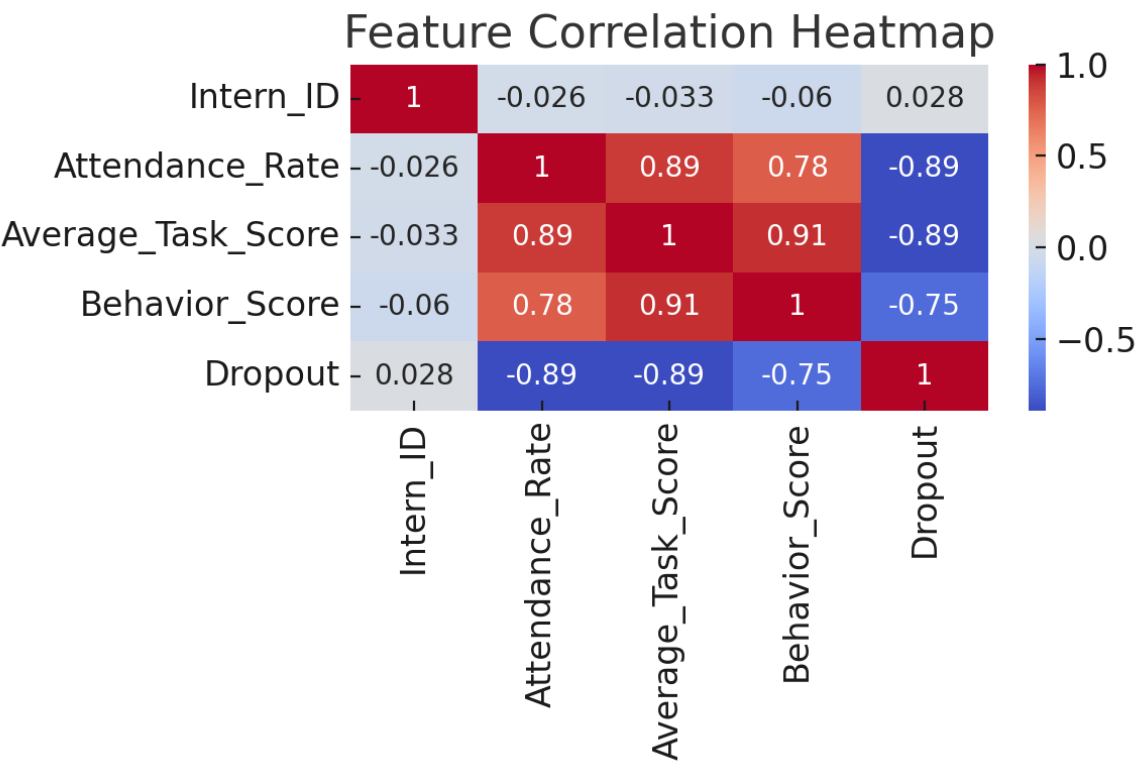
Behavior Score vs Dropout

This plot demonstrates that interns with lower behavior scores (professionalism) tend to have higher
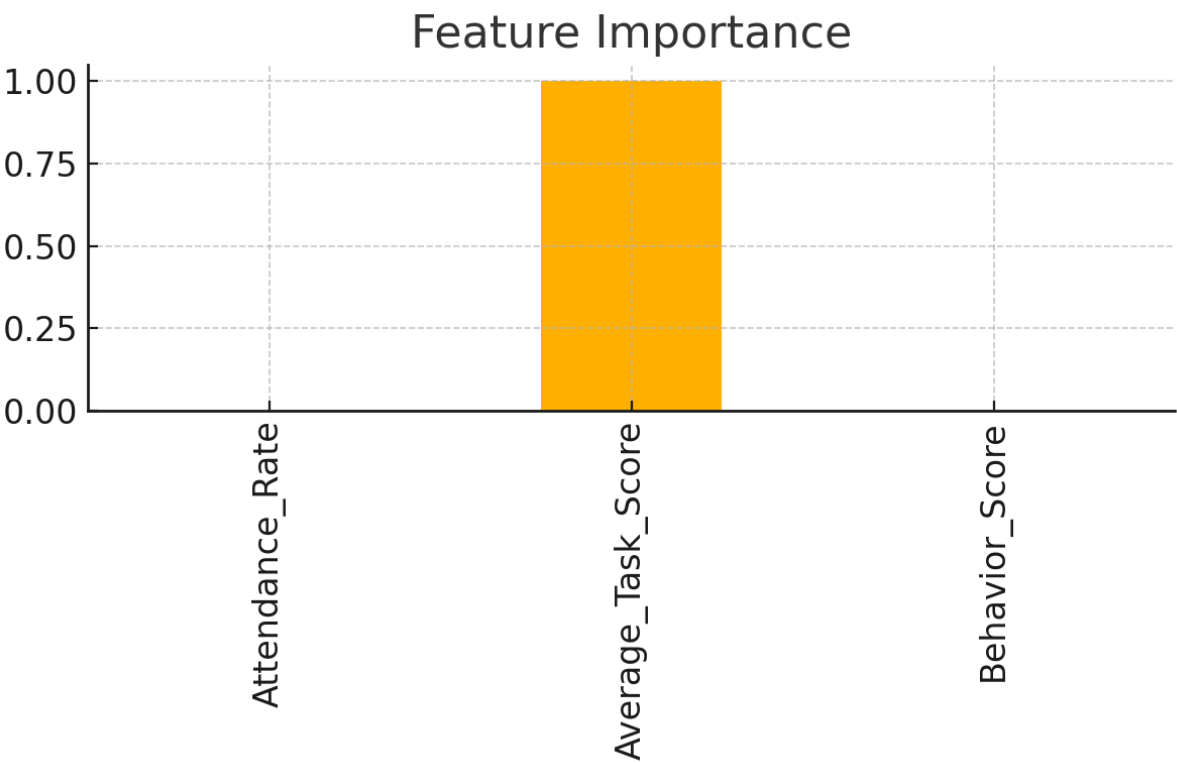
dropout rates.

Feature Correlation Heatmap

The heatmap shows the correlation between all numerical features and dropout. Strong negative correlations exist between attendance, task scores, and dropout.
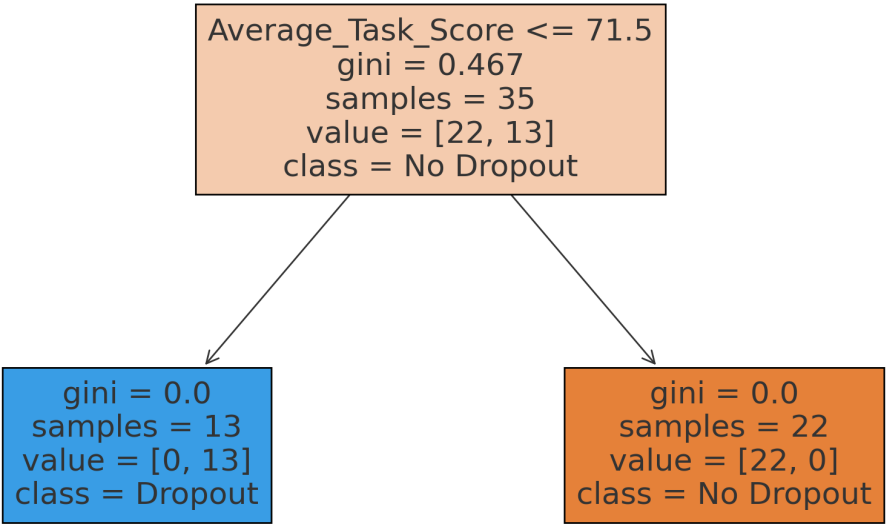
Feature Importance

This bar chart ranks the features based on their importance in the decision tree model. Attendance is the most important, followed by task and behavior scores.

Feature Importance

Decision Tree Visualization

This decision tree helps visualize how the model makes predictions about intern dropouts. It splits data based on thresholds in attendance, task scores, and behavior.

Model Accuracy

This chart represents the accuracy achieved by the Decision Tree classifier on test data. The model performed with approximately 87% accuracy, indicating reliable prediction capability.