📌 Advanced Data Analysis & Feature Selection - Explanation

📖 1. Overview

This document explains the advanced data analysis techniques and feature selection methods used to improve the predictive accuracy of the employee attrition model. The analysis involved statistical tests and machine learning-based feature selection to identify the most important factors influencing attrition.

📊 2. Statistical Tests for Feature Analysis

Statistical tests were performed to determine the relationship between various employee attributes and attrition. The following methods were used:

📍 A. T-Test (Independent Samples T-Test)

The T-test compares the means of numerical variables (e.g., salary) between two groups: employees who left and those who stayed. A low p-value (< 0.05) indicates that the variable significantly influences attrition. For example, a low p-value for 'Monthly Income' suggests that salary plays a crucial role in an employee’s decision to leave.

📍 B. ANOVA (Analysis of Variance)

ANOVA compares the means of multiple groups (e.g., job levels) to see if at least one differs significantly. This helps analyze whether factors like 'Job Level' or 'Performance Rating' impact attrition. A p-value < 0.05 confirms a significant difference among groups.

📍 C. Chi-Squared Test (For Categorical Features)

The Chi-Square test checks whether categorical features (e.g., Job Role, Remote Work) are related to attrition. A low p-value means the feature is significantly related to an employee’s decision to stay or leave. For example, if 'Remote Work' has a low p-value, it suggests that employees with flexible work options are less likely to leave.

🔍 3. Feature Selection Techniques

To enhance model accuracy, feature selection techniques were used to identify the most important predictors of attrition.

📍 A. Correlation Matrix

The correlation matrix measures relationships between numerical features and attrition. Highly correlated variables can be removed to avoid redundancy, improving model efficiency.

📍 B. Recursive Feature Elimination (RFE)

RFE iteratively removes less important features until only the best predictors remain. A machine learning model (Random Forest) was used to rank the most influential variables, ensuring the selection of features with the highest impact on attrition.

📍 C. SelectKBest (ANOVA F-test)

SelectKBest ranks features based on their statistical significance using ANOVA F-tests. The top-ranked variables are selected as the most relevant for predicting attrition. This method ensures that only the strongest predictors are included in the final model.

📌 4. Key Insights & Business Recommendations

🎯 Employees with lower salaries and fewer promotions are more likely to leave.

🎯 Remote Work reduces attrition, while longer commutes increase it.

🎯 Job Level is a strong predictor – higher positions have lower turnover.

🎯 The selected features should be used in the predictive model to enhance accuracy.