STAT 208 Statistical Data Mining Methods - Project Proposal

Team: Breaking Bias

Team Leader: Sunny Khade

Team Members: Ashwin Satra, Dushyant Vaishnaw, Atharva Pawar, Devayani Pawar

Topic Title:
Predicting Employee Attrition Using HR Analytics

Topic Motivation:
Our team is interested in exploring employee attrition because it directly affects company performance, morale, and recruitment costs. In today's competitive market, retaining talent is more crucial than ever. By predicting attrition, HR teams can take proactive steps to improve employee satisfaction and reduce turnover.

Research

Employee attrition is when workers voluntarily or involuntarily leave a company. Numerous factors contribute to attrition: age, job satisfaction, department, overtime, and income levels. Scientifically, attrition is often analyzed using classification models, and it involves both categorical (e.g., gender, department) and numerical (e.g., age, income) variables. Machine learning techniques such as decision trees and logistic regression help uncover hidden patterns and relationships in this type of workforce data.

Objective Proposal

Objective 1: Predict which employees are most likely to leave the company using classification models.

Objective 2: Identify the top features that influence attrition using model interpretability tools.

Objective 3: Analyze patterns in attrition across different departments and age groups to suggest policy changes.

Data

Link: Synthetic Employee Attrition Dataset – Kaggle

- **Source:** Kaggle
- **Rows:** ~1,500
- **Columns:** Includes age, job role, department, overtime status, income, years at company, and attrition status
- **Type:** Mix of categorical and numerical data suitable for supervised learning

---

Data Gaps

This dataset is synthetic, so while it's realistic, it may not include real-world complexities like abrupt policy changes or economic shifts. It also lacks real-time updates and qualitative feedback from exit interviews. Additionally, cultural or psychological aspects of attrition (like engagement or leadership quality) are not captured, potentially limiting model accuracy.