

Report on HR Employee Attrition Dataset

1. Brief Description of the Dataset and a Summary of Its Attributes

The dataset contains data regarding employee attrition from a company's HR department, with 1,470 entries and 35 columns. The attributes include a mix of demographic, job-related, and satisfaction-related information about the employees. Key attributes are:

- **Demographic:** Age, Gender, Marital Status, etc.
- **Job-Related:** JobRole, Department, JobLevel, etc.
- **Satisfaction and Performance:** JobSatisfaction, PerformanceRating, WorkLifeBalance, etc.
- **Target Variable:** Attrition (whether an employee left the company).

All columns are either integer or object types, with no missing values detected.

1. Initial Plan for Data Exploration

The exploration plan includes:

- **Data Cleaning:** Removing redundant or irrelevant columns and handling any duplicate entries.
- **Exploratory Data Analysis (EDA):** Generating visualizations to understand the distribution and relationships between variables, particularly those related to attrition.

1. Actions Taken for Data Cleaning and Feature Engineering

- **Redundant Columns Removed:** Columns such as EmployeeCount, EmployeeNumber, Over18, and StandardHours were identified as irrelevant to the analysis and dropped.
- **Duplicate Entries:** Any duplicate entries in the dataset were removed.
- **Missing Values:** No missing values were found in the dataset, so no further action was needed.
- **Feature Engineering:** The Attrition column was converted to binary (1 for 'Yes', 0 for 'No'). Min-max scaling was applied to numerical variables to prepare the dataset for potential machine learning applications. Transforming categorical/object variables using one-hot encoding.

1. Key Findings and Insights

- **Correlation Analysis:** Strong correlations were found between several job-related and satisfaction-related features and attrition. These insights can help identify key drivers of employee turnover.
- **Gender Distribution:** The company has a nearly even split between male and female employees.
- **Marital Status and Job Role:** The distribution of marital status and job roles was analyzed to understand their potential impact on attrition.
- **OverTime:** A significant proportion of employees work overtime, which might influence attrition.

1. Hypotheses Formulation

Based on the data, the following hypotheses were formulated:

1. **Hypothesis 1:** *Employees who work overtime are more likely to experience attrition.*
2. **Hypothesis 2:** *Job satisfaction levels are significantly associated with employee attrition.*
3. **Hypothesis 3:** *Employees in certain job roles have a higher probability of leaving the company compared to others.*

4. Formal Significance Test

A Chi-Square test was conducted to test the association between the categorical variables and attrition. For example, the relationship between OverTime and Attrition was tested, yielding a significant p-value, indicating a strong association. This suggests that employees who work overtime are indeed more likely to leave the company.

1. Suggestions for Next Steps

- **Further Analysis:** Consider deeper statistical analysis, such as logistic regression, to quantify the impact of different factors on attrition.
- **Model Building:** Develop predictive models to identify at-risk employees (using Attrition as response variable).
- **Employee Surveys:** Collect additional qualitative data to complement the current quantitative analysis, particularly regarding job satisfaction and work-life balance.

1. Data Quality and Request for Additional Data

The dataset is high quality, with no missing values in rows or columns or significant inconsistencies. However, to enhance the analysis, additional data on employee engagement, detailed reasons for attrition, and historical performance metrics would be valuable if available for collection. This could provide more granular insights into the causes of attrition and help refine the predictive capabilities of models.

Appendix 1

Detailed Description of Features

1. Age: The age of the employee.
2. Attrition: Indicates whether the employee left the company ('Yes') or not ('No').
3. BusinessTravel: Frequency of travel for business purposes, categorized as 'Travel_Rarely', 'Travel_Frequently', or 'Non-Travel'.
4. DailyRate: The daily rate of the employee's pay.
5. Department: The department in which the employee works (e.g., 'Sales', 'Research & Development', 'Human Resources').
6. DistanceFromHome: The distance between the employee's home and the workplace, measured in miles.
7. Education: The education level of the employee, represented as an integer (1: 'Below College', 2: 'College', 3: 'Bachelor', 4: 'Master', 5: 'Doctor').
8. EducationField: The field of education of the employee (e.g., 'Life Sciences', 'Medical', 'Marketing', 'Technical Degree').
9. EmployeeCount: A constant column with the same value (1) for all employees, which is not informative.
10. EmployeeNumber: A unique identifier assigned to each employee.
11. EnvironmentSatisfaction: Satisfaction level with the work environment, on a scale from 1 (Low) to 4 (Very High).
12. Gender: The gender of the employee ('Male' or 'Female').
13. HourlyRate: The hourly wage of the employee.
14. JobInvolvement: The level of involvement the employee has in their job, on a scale from 1 (Low) to 4 (Very High).
15. JobLevel: The level or rank of the employee within the organization.
16. JobRole: The specific role or job title of the employee (e.g., 'Sales Executive', 'Research Scientist').
17. JobSatisfaction: Satisfaction level with the job, on a scale from 1 (Low) to 4 (Very High).
18. MaritalStatus: The marital status of the employee ('Single', 'Married', 'Divorced').
19. MonthlyIncome: The monthly income of the employee.
20. MonthlyRate: The monthly rate of the employee's pay.
21. NumCompaniesWorked: The number of companies the employee has worked for prior to the current company.

22. Over18: A constant column indicating whether the employee is over 18 years old (all entries are 'Y').
23. OverTime: Indicates whether the employee regularly works overtime ('Yes' or 'No').
24. PercentSalaryHike: The percentage increase in salary in the last performance review.
25. PerformanceRating: The rating of the employee's performance, on a scale from 1 (Low) to 4 (Outstanding).
26. RelationshipSatisfaction: Satisfaction level with relationships at work, on a scale from 1 (Low) to 4 (Very High).
27. StandardHours: The standard number of working hours per week (all entries are 80).
28. StockOptionLevel: The level of stock options granted to the employee, on a scale from 0 (None) to 3 (High).
29. TotalWorkingYears: The total number of years the employee has been working in their career.
30. TrainingTimesLastYear: The number of times the employee participated in training sessions in the last year.
31. WorkLifeBalance: The level of balance the employee feels between their work and personal life, on a scale from 1 (Bad) to 4 (Best).
32. YearsAtCompany: The number of years the employee has been with the current company.
33. YearsInCurrentRole: The number of years the employee has been in their current role within the company.
34. YearsSinceLastPromotion: The number of years since the employee's last promotion.
35. YearsWithCurrManager: The number of years the employee has been working with their current manager.