COM00148M

Department of Computer Science

Big Data Analytics

SUMMATIVE REASSESSMENT

Task 1a: Analysis and Report

* Assumptions
* Several assumptions were made during this analysis:
* Imbalance Handling: Significant class imbalances in the dataset were addressed using SMOTE (Synthetic Minority Over-sampling Technique).
* Data Preprocessing: Attributes were converted to binary using the NominalToBinary filter to identify specific correlations with attrition.
* Evaluation Techniques: Various attribute evaluation techniques were employed to identify the most significant attributes for predicting attrition.
* Data Preprocessing
* Data preprocessing was crucial to ensure the data was fit for analysis:
* Handling Missing Values: Missing values were imputed with means or modes.
* SMOTE Application: SMOTE was applied to handle class imbalances in each department.
* HR: 85 No to 80 Yes (from 85 No to 20 Yes)
* R&D: 1405 No to 1856 Yes (from 1405 No to 232 Yes)
* Sales: 644 Yes to 596 No (from 161 Yes to 596 No)
* Binary Conversion: Attributes were converted to binary for a detailed correlation analysis using the NominalToBinary filter.
* Attribute Evaluation: CorrelationAttributeEval, InfoGainAttributeEval, and CfsSubsetEval were used for attribute evaluation.
* Analysis Techniques Employed
* CorrelationAttributeEval: This technique was used to rank attributes based on their correlation with attrition.
* InfoGainAttributeEval: This technique was used to evaluate the information gain of each attribute.
* CfsSubsetEval: This technique, combined with BestFirst and GreedyStepwise search methods, was used to identify the most predictive subset of attributes.
* Justification of Techniques
* SMOTE: Addressing class imbalance is critical in ensuring the model is not biased towards the majority class.
* Attribute Evaluation Methods: Each method provides a unique perspective. CorrelationAttributeEval highlights direct linear relationships, InfoGainAttributeEval measures the reduction in entropy, and CfsSubsetEval finds subsets of attributes that collectively have good predictive power.
* General Summary of Results
* HR Department
* Top Attributes (CorrelationAttributeEval): Age, YearsWithCurrManager, EducationField=Medical, TotalWorkingYears, YearsInCurrentRole, IncomePerJobLevel, WorkLifeSatisfaction.
* Best Subset (CfsSubsetEval): Age, BusinessTravel=Travel\_Frequently, DistanceFromHome, EducationField, EnvironmentSatisfaction, Gender, JobInvolvement, JobRole=Human\_Resources, JobSatisfaction, MaritalStatus=Single, PercentSalaryHike, StockOptionLevel, YearsInCurrentRole, YearsWithCurrManager, AvgSatisfaction, WorkLifeSatisfaction.
* Top Attributes (InfoGainAttributeEval): DistanceFromHome, MonthlyIncome, WorkLifeSatisfaction, TotalWorkingYears, Age, ProportionCareerAtCompany, HourlyRate, EnvironmentSatisfaction, JobInvolvement, YearsWithCurrManager.
* Commentary: In the HR department, Age and YearsWithCurrManager were highly predictive of staff turnover, indicating that younger employees or those with shorter tenures with their current managers are more likely to leave. Additionally, attributes like MonthlyIncome and WorkLifeSatisfaction were significant, highlighting that financial compensation and work-life balance are critical factors in employee retention.
* R&D Department
* Top Attributes (CorrelationAttributeEval): OverTime=No, JobLevel, AvgTenurePerJobRole, MonthlyIncome, JobRole=Manufacturing\_Director, YearsInCurrentRole, TotalWorkingYears, Gender=Female.
* Best Subset (CfsSubsetEval): BusinessTravel=Travel\_Frequently, DailyRate, EducationField, Gender, JobInvolvement, JobLevel, JobRole=Manufacturing\_Director, OverTime=No, StockOptionLevel, AvgSatisfaction, WorkLifeSatisfaction, ProportionCareerAtCompany.
* Top Attributes (InfoGainAttributeEval): PercentSalaryHike, AvgSatisfaction, WorkLifeSatisfaction, DistanceFromHome, EnvironmentSatisfaction, TrainingTimesLastYear, Age, YearsWithCurrManager, JobSatisfaction.
* Commentary: For the R&D department, OverTime (working overtime), JobLevel, and AvgTenurePerJobRole were significant predictors of turnover. This suggests that employees who work overtime or have lower job levels are more likely to leave. Conversely, higher JobSatisfaction and better WorkLifeSatisfaction are strong indicators of employee retention.
* Sales Department
* Top Attributes (CorrelationAttributeEval): MaritalStatus=Divorced, JobRole=Manager, AvgTenurePerJobRole, JobLevel, JobInvolvement, JobSatisfaction, OverTime=No, MonthlyIncome.
* Best Subset (CfsSubsetEval): BusinessTravel=Non-Travel, DistanceFromHome, EducationField, JobInvolvement, JobRole=Manager, MaritalStatus=Divorced, MonthlyRate, OverTime=No, WorkLifeSatisfaction.
* Top Attributes (InfoGainAttributeEval): AvgSatisfaction, RelationshipSatisfaction, StockOptionLevel, WorkLifeSatisfaction, Education, EnvironmentSatisfaction, JobSatisfaction, JobInvolvement.
* Commentary: In the Sales department, MaritalStatus (especially being divorced) and JobRole (being a manager) were significant predictors of turnover. AvgTenurePerJobRole and JobLevel were also critical, indicating that employees in lower job levels or with shorter tenures per job role are more likely to leave. JobSatisfaction and WorkLifeSatisfaction are strong indicators of retention, emphasizing the importance of job fulfillment and work-life balance.
* Critical Evaluation
* The approach taken in this analysis had several effective elements:
* SMOTE: Effectively handled class imbalance, ensuring the model could learn from both classes.
* Multiple Attribute Evaluators: Provided a comprehensive understanding of which attributes are most predictive.
* However, there were some ineffective aspects:
* Overfitting Risk: Using the training set for evaluation, especially with SMOTE, may lead to overfitting. Cross-validation or an independent test set would provide a more robust evaluation however the test set cannot be used due to the target variable being missing (99%)
* Complexity: The combination of multiple attribute evaluation methods made the analysis complex. A simpler approach could be more interpretable.
* This analysis allowed me to carry out feature selection
* To build the best predictive model for each department, we'll select the top features based on their correlation strengths and their relevance to predicting attrition and retention. We'll aim to choose a maximum of 12 features for each department to ensure a balance between model complexity and performance.
* **HR Department**
* **Age (0.54595)**
* **YearsWithCurrManager (0.4844)**
* **EducationField=Medical,Life\_Sciences,Technical\_Degree,Other (0.48433)**
* **TotalWorkingYears (0.46802)**
* **YearsInCurrentRole (0.40118)**
* **IncomePerJobLevel (0.38129)**
* **WorkLifeSatisfaction (0.37962)**
* **MonthlyIncome (0.35245)**
* **JobLevel (0.32079)**
* **AvgTenurePerJobRole (0.34997)**
* **YearsAtCompany (0.30186)**
* **JobSatisfaction (0.30726)**
* These features have high positive correlations and are likely to provide a robust model for predicting attrition in the HR department.
* **R&D Department**
* **OverTime=No (0.34952)**
* **JobLevel (0.31898)**
* **AvgTenurePerJobRole (0.29968)**
* **MonthlyIncome (0.29613)**
* **YearsInCurrentRole (0.27056)**
* **TotalWorkingYears (0.26948)**
* **JobRole=Manufacturing\_Director,Healthcare\_Representative,Research\_Director (0.28967)**
* **YearsAtCompany (0.19388)**
* **StockOptionLevel (0.21271)**
* **YearsWithCurrManager (0.2326)**
* **WorkLifeSatisfaction (0.2011)**
* **AvgSatisfaction (0.20471)**
* These features provide a good mix of job-specific and personal attributes that influence attrition in the R&D department.
* **Sales Department**
* **MaritalStatus=Divorced,Married (0.454614)**
* **JobRole=Manager,Sales\_Executive (0.306753)**
* **AvgTenurePerJobRole (0.298285)**
* **JobLevel (0.24206)**
* **MonthlyIncome (0.189271)**
* **WorkLifeSatisfaction (0.188325)**
* **JobInvolvement (0.172559)**
* **EnvironmentSatisfaction (0.153491)**
* **Age (0.184137)**
* **YearsInCurrentRole (0.231301)**
* **YearsWithCurrManager (0.220785)**
* **TotalWorkingYears (0.237546)**
* These features capture the key aspects of job satisfaction, personal circumstances, and job-specific factors that affect attrition in the Sales department.
* **Summary**
* **HR Department**: Focus on tenure-related features, job level, income, and satisfaction metrics.
* **R&D Department**: Emphasize job role specifics, overtime status, tenure, income, and satisfaction metrics.
* **Sales Department**: Highlight marital status, job role, tenure, income, job involvement, and satisfaction metrics.

**Positive Correlations**: Attributes such as age, years with the current manager, total working years, years in the current role, monthly income, and work-life satisfaction consistently show positive correlations with retention across departments. This means that employees who are older, have longer tenure, higher income, and better satisfaction with their work-life balance are more likely to stay.

**Negative Correlations**: Attributes like distance from home and specific job roles (in some cases) show negative correlations with retention. Employees living further away from work or in certain roles may be more inclined to leave.

### Interpretation

**HR Department**: Retention is strongly influenced by age, tenure with the current manager, and total working years. These factors suggest that experienced employees who have developed stability within the company and have longer tenures with their managers are less likely to leave.

**R&D Department**: Factors such as not working overtime, higher job levels, and longer average tenure per job role are significant. This indicates that R&D employees who are in higher positions, do not work excessive overtime, and have stable roles are more likely to stay.

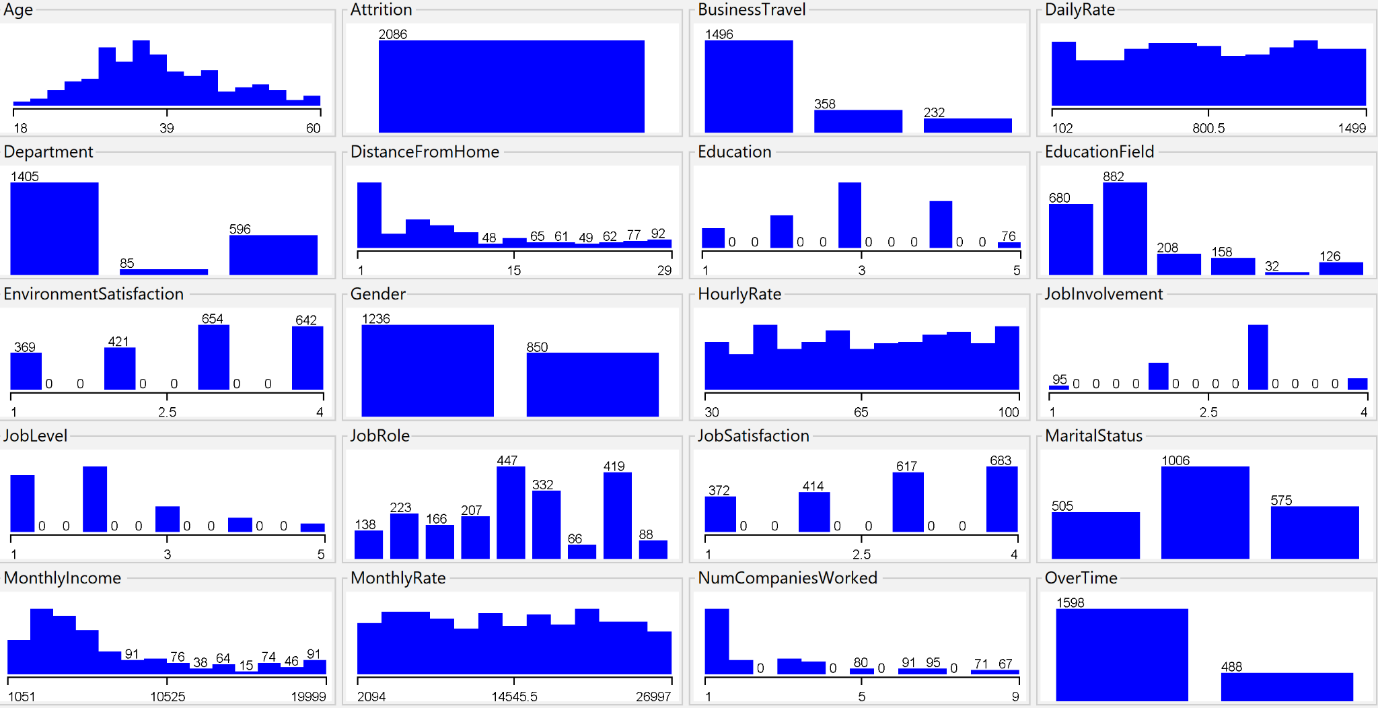
**Sales Department**: Marital status, certain job roles, and satisfaction with work-life balance are key predictors. Sales employees who are married or divorced, hold specific roles like manager or sales executive, and have a good work-life balance tend to stay longer.

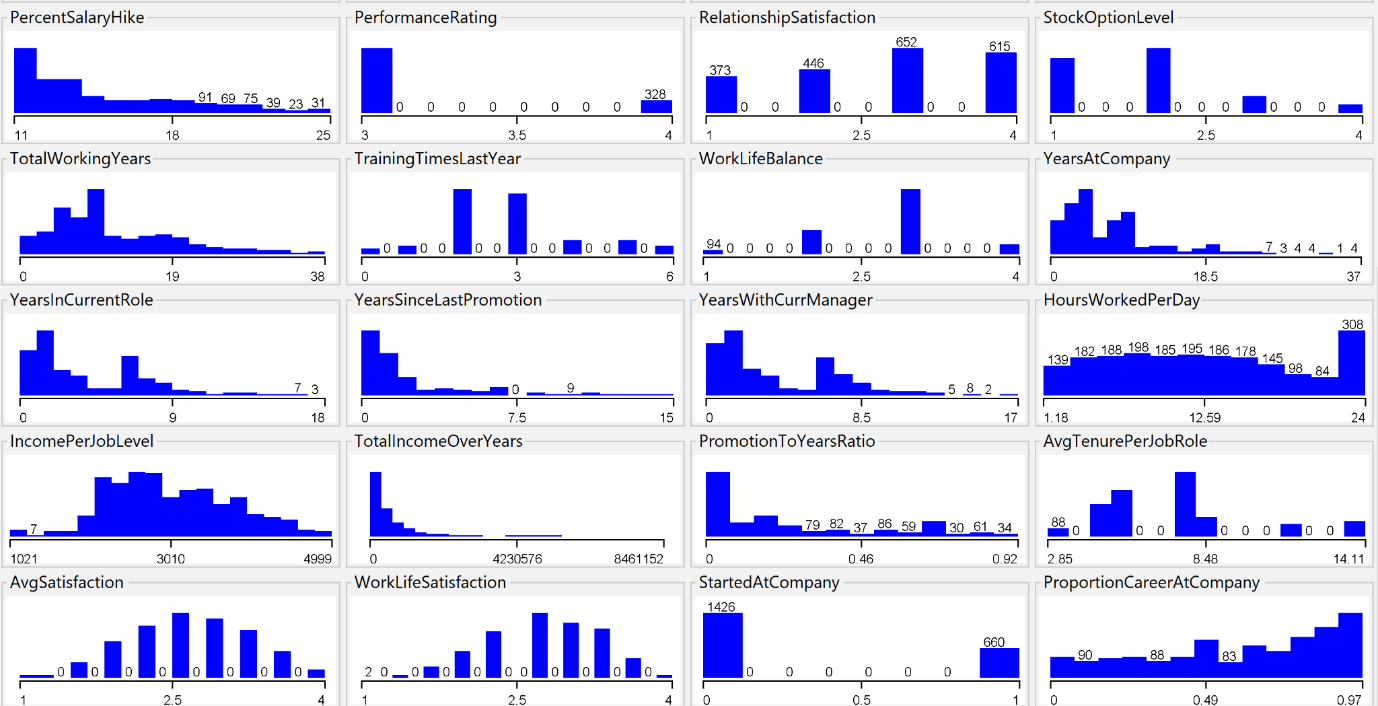
### Conclusion

The correlation strengths provide valuable insights into what factors most strongly influence retention and attrition within each department. These insights can help tailor strategies to improve retention by focusing on enhancing the positively correlated factors and addressing the negatively correlated ones.

1b split up datasets

The data visualisation for attrition = no across various attributes provides insights into the distribution and characteristics of the staff who have not left the organisation. Here’s an interpretation of each attribute’s histogram:





1. Age: The age distribution is relatively normal, with the majority of employees falling between the ages of 25 and 50. This suggests a balanced age distribution with a concentration of employees in their mid-career stages.

2. Attrition: Shows the count of employees who have not left the organisation. This is a control attribute showing the number of instances where attrition is marked as "No."

3. BusinessTravel: Most employees travel rarely, with a smaller portion of employees traveling frequently. This indicates that excessive travel is not a common characteristic among employees who stay.

4. DailyRate: The distribution of daily rates (presumably pay rates) is fairly uniform, suggesting no significant pay disparities among employees who have stayed.

5. Department: A majority of employees belong to the Research & Development department, followed by Sales, with Human Resources having the least. This indicates a heavier concentration of employees in technical roles.

6. DistanceFromHome: Many employees live close to their workplace, with fewer employees having a long commute. This suggests that a shorter commute may be correlated with employee retention.

7. Education: Most employees have a higher education level (3), indicating that a higher level of education is common among those who stay.

8. EducationField: Life Sciences and Medical fields dominate, indicating that these fields are well-represented among employees who stay with the company.

9. EnvironmentSatisfaction: Satisfaction levels are high among employees who stay, with many rating their environment satisfaction as 3 or 4.

10. Gender: There is a higher number of male employees compared to female employees among those who stay.

11. HourlyRate: The hourly rates show a uniform distribution, indicating no significant pay rate differences among employees who have stayed.

12. JobInvolvement: Higher job involvement is observed, indicating that employees who stay tend to be more involved in their jobs.

13. JobLevel: The majority of employees are at lower job levels (1-2), suggesting that employees at these levels are more likely to stay.

14. JobRole: The roles of Research Scientist and Sales Executive are more common among those who stay, indicating these roles have lower attrition rates.

15. JobSatisfaction: Higher job satisfaction levels are prevalent, indicating that satisfaction with one's job is a strong factor in retention.

16. MaritalStatus: Married employees are in the majority, suggesting that marital status might influence the decision to stay.

17. MonthlyIncome: Monthly income is widely distributed but shows higher counts in lower income brackets, suggesting that employees with lower to moderate incomes tend to stay.

18. MonthlyRate: Monthly rates are evenly distributed, indicating no strong influence of this attribute on retention.

19. NumCompaniesWorked: Most employees have worked at fewer companies, suggesting that employees with fewer previous jobs are more likely to stay.

20. OverTime: Many employees do not work overtime, suggesting that avoiding overtime could be a factor in retention.

21. PercentSalaryHike: Higher salary hikes are less common, indicating that large salary increases are not necessarily linked to retention.

22. PerformanceRating: Higher performance ratings are prevalent, indicating that better performers tend to stay.

23. RelationshipSatisfaction: Higher satisfaction with relationships at work is common among employees who stay.

24. StockOptionLevel: Most employees have lower stock option levels, suggesting that stock options are not a strong retention factor.

25. TotalWorkingYears: Employees with fewer total working years are more common, indicating that those early in their careers are more likely to stay.

26. TrainingTimesLastYear: Few employees have high training times, suggesting that extensive training is not strongly linked to retention.

27. WorkLifeBalance: Better work-life balance ratings are common, indicating that a good balance between work and life is crucial for retention.

28. YearsAtCompany: Employees who stay tend to have been with the company for fewer years, indicating newer employees are more likely to stay.

29. YearsInCurrentRole: Fewer years in the current role are more common, suggesting that not having stayed too long in one role may be linked to retention.

30. YearsSinceLastPromotion: Many employees have recently been promoted, indicating that recent promotions may aid in retention.

31. YearsWithCurrManager: Employees with fewer years with their current manager are more prevalent, indicating that newer manager relationships might be linked to retention.

32. HoursWorkedPerDay: Most employees work around 8 hours per day, indicating that standard work hours are common among those who stay.

33. IncomePerJobLevel: Income is relatively evenly distributed across job levels, indicating no strong influence on retention.

34. TotalIncomeOverYears: Higher total incomes are less common, suggesting that long-term higher incomes are not strongly linked to retention.

35. PromotionToYearsRatio: Higher ratios of promotion to years are less common, indicating that frequent promotions are not a strong factor in retention.

36. AvgTenurePerJobRole: Higher average tenures per job role are more common, suggesting that those who stay tend to have longer tenures in their roles.

37. AvgSatisfaction: Higher average satisfaction ratings are common, indicating overall satisfaction is a strong factor in retention.

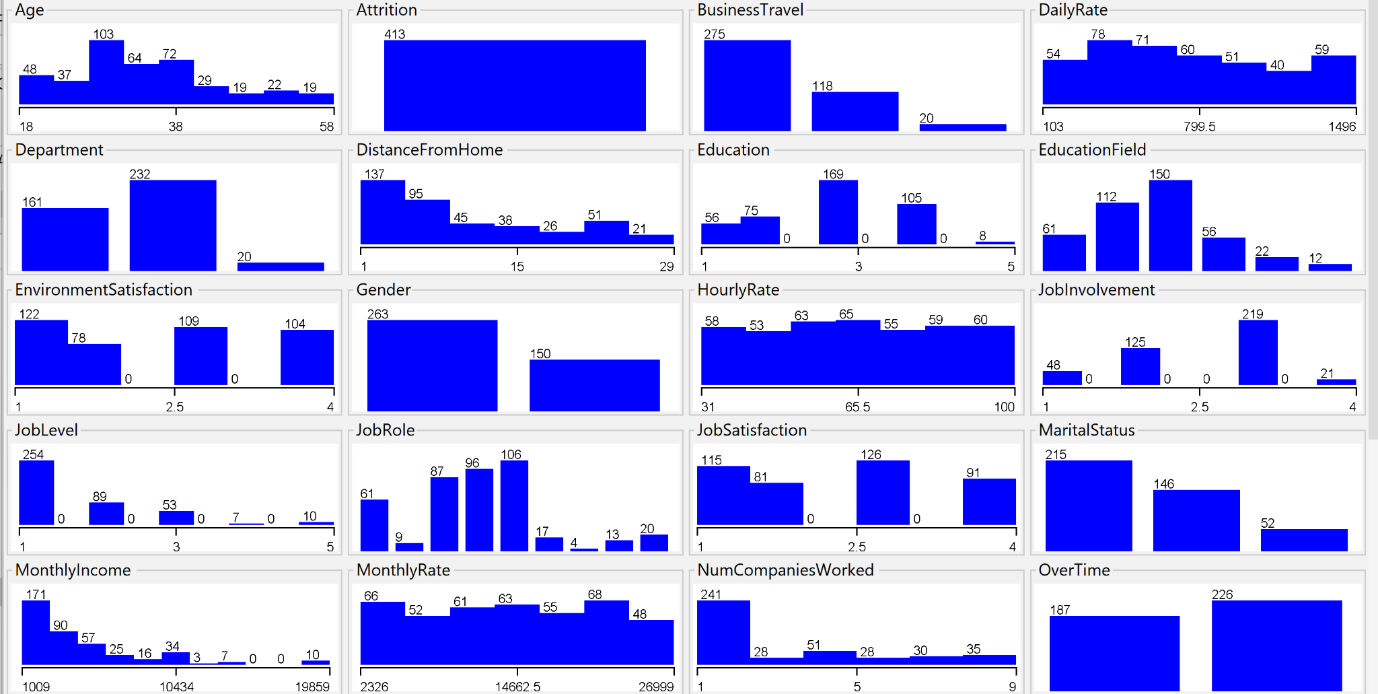
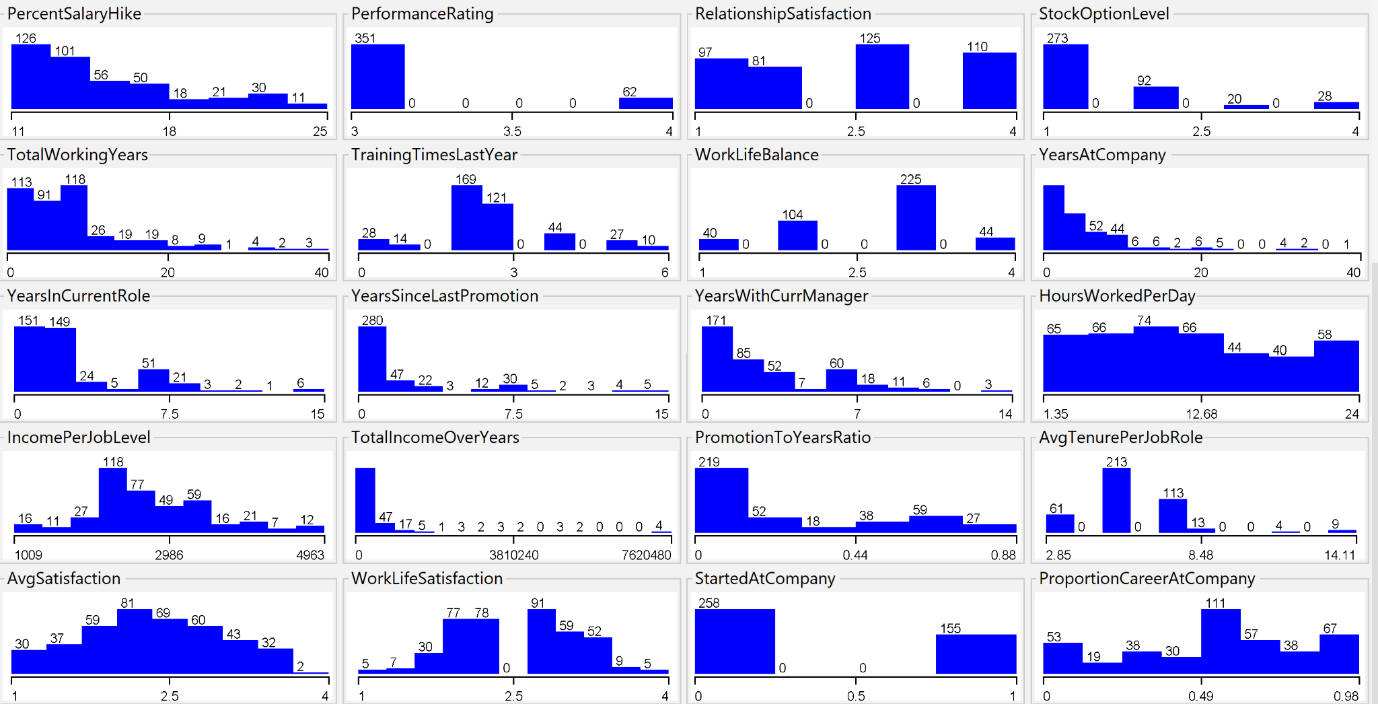
38. WorkLifeSatisfaction: Higher satisfaction with work-life balance is common, reinforcing the importance of this factor in retention.

39. StartedAtCompany: Higher counts are seen for those who started more recently, indicating that recent hires are more likely to stay.

40. ProportionCareerAtCompany: Higher proportions of career spent at the company are more common, indicating that longer-term employees are more likely to stay.

These insights can help understand the factors that contribute to employee retention, informing strategies to improve retention rates further.

Analysis of Attrition = Yes Data Subset

1. Age:

- The distribution shows that most employees leaving the organization are younger, predominantly in the age groups 18-28. There is a decline in attrition as age increases.

2. Department:

- Attrition is higher in the Research & Development department, with the Sales department having a significant portion as well. The Human Resources department shows the least attrition.

3. Business Travel:

- Employees who travel rarely or frequently have higher attrition rates compared to those who do not travel.

4. Daily Rate:

- There is no clear pattern in daily rates with respect to attrition. The distribution is fairly even across different daily rate ranges.

5. Distance From Home:

- A significant number of employees leaving live close to the workplace (1-10 units of distance). There is a gradual decline in attrition as the distance increases.

6. Education:

- Employees with a Bachelor's degree (level 3) show the highest attrition rates. Other education levels are less represented in the attrition data.

7. Education Field:

- The majority of attrition occurs among employees with education in Life Sciences and Medical fields. Other fields have significantly lower attrition rates.

8. Environment Satisfaction:

- There is a noticeable attrition among employees with low to moderate environment satisfaction (levels 1 and 2). Higher satisfaction levels correspond to lower attrition.

9. Gender:

- Attrition is higher among male employees compared to female employees.

10. Hourly Rate:

- The hourly rate distribution does not show a clear pattern with attrition, indicating that hourly rate might not be a significant factor in predicting attrition.

11. Job Involvement:

- Employees with lower job involvement (level 1) have higher attrition rates. Those with higher involvement levels (2 and 3) show less attrition.

12. Job Level:

- Lower job levels (1 and 2) exhibit higher attrition rates. Higher job levels have fewer employees leaving.

13. Job Role:

- Sales Representatives, Laboratory Technicians, and Research Scientists have higher attrition rates. Managers and Healthcare Representatives show lower attrition.

14. Job Satisfaction:

- Lower job satisfaction (level 1 and 2) correlates with higher attrition rates.

15. Marital Status:

- Single employees show higher attrition rates compared to married and divorced employees.

16. Monthly Income:

- There is a tendency for employees with lower monthly income to leave more frequently.

17. Monthly Rate:

- Monthly rates do not show a clear pattern with attrition, suggesting it might not be a strong predictor.

18. Number of Companies Worked:

- Employees with fewer companies worked (0-3) have higher attrition rates. Those with more companies worked show lower attrition.

19. Over Time:

- Employees who do not work overtime show higher attrition rates.

20. Percent Salary Hike:

- Lower salary hikes correlate with higher attrition rates. Employees with higher salary hikes show less attrition.

21. Performance Rating:

- Performance ratings show that most employees leaving have a high performance rating of 3.

22. Relationship Satisfaction:

- Employees with lower relationship satisfaction (level 1 and 2) have higher attrition rates.

23. Stock Option Level:

- Employees with lower stock option levels (1 and 2) show higher attrition rates.

24. Total Working Years:

- Employees with fewer total working years (0-10) exhibit higher attrition. Attrition decreases as total working years increase.

25. Training Times Last Year:

- Employees with fewer training sessions in the last year (0-2) show higher attrition rates.

26. Work-Life Balance:

- Poor work-life balance (level 1 and 2) correlates with higher attrition rates.

27. Years at Company:

- Employees with fewer years at the company (0-10) show higher attrition rates.

28. Years in Current Role:

- Employees with fewer years in the current role (0-7) exhibit higher attrition rates.

29. Years Since Last Promotion:

- Employees with fewer years since the last promotion (0-3) show higher attrition rates.

30. Years with Current Manager:

- Employees with fewer years with their current manager (0-5) show higher attrition rates.

31. Hours Worked Per Day:

- The distribution of hours worked per day does not show a clear pattern with attrition, indicating it might not be a strong predictor.

32. Income Per Job Level:

- Lower income per job level is associated with higher attrition rates.

33. Total Income Over Years:

- Employees with lower total income over the years exhibit higher attrition.

34. Promotion to Years Ratio:

- Lower promotion to years ratio correlates with higher attrition rates.

35. Average Tenure Per Job Role:

- Employees with shorter average tenure per job role show higher attrition rates.

36. Average Satisfaction:

- Lower average satisfaction correlates with higher attrition rates.

37. Work-Life Satisfaction:

- Lower work-life satisfaction correlates with higher attrition rates.

38. Started at Company:

- The histogram shows a concentration of employees leaving shortly after starting at the company, indicating that attrition is higher among newer employees.

39. Proportion Career at Company:

- Employees with a lower proportion of their career spent at the company show higher attrition rates.

**Analysis of the Most Telling Factors for Both Subsets (Attrition = No and Attrition = Yes)**

**Common Factors Between Attrition = No and Attrition = Yes**

1. **Age:**
   * **Attrition = No:** The age distribution is relatively even, but employees aged 30-50 tend to remain in the organization more.
   * **Attrition = Yes:** Younger employees (18-28) have higher attrition rates. Attrition decreases with age.
2. **Department:**
   * **Attrition = No:** Most employees in the Research & Development department tend to stay. The Sales department has a significant number staying, while Human Resources shows the least retention.
   * **Attrition = Yes:** Higher attrition in the Research & Development department, followed by the Sales department. The Human Resources department shows the least attrition.
3. **Business Travel:**
   * **Attrition = No:** Employees who travel frequently or rarely have higher retention rates than those who do not travel.
   * **Attrition = Yes:** Higher attrition rates for employees who travel rarely or frequently compared to non-travelers.
4. **Distance From Home:**
   * **Attrition = No:** Employees living closer to work (1-10 units of distance) have higher retention rates.
   * **Attrition = Yes:** Higher attrition among employees living closer to work. Attrition decreases with increasing distance from home.
5. **Environment Satisfaction:**
   * **Attrition = No:** Higher retention rates among employees with higher environment satisfaction (levels 3 and 4).
   * **Attrition = Yes:** Higher attrition rates among employees with lower environment satisfaction (levels 1 and 2).
6. **Gender:**
   * **Attrition = No:** Slightly more females tend to stay compared to males.
   * **Attrition = Yes:** Higher attrition rates among males compared to females.
7. **Job Involvement:**
   * **Attrition = No:** Higher retention among employees with higher job involvement (levels 3 and 4).
   * **Attrition = Yes:** Higher attrition among employees with lower job involvement (level 1).
8. **Job Satisfaction:**
   * **Attrition = No:** Higher retention rates among employees with higher job satisfaction (levels 3 and 4).
   * **Attrition = Yes:** Higher attrition rates among employees with lower job satisfaction (levels 1 and 2).
9. **Marital Status:**
   * **Attrition = No:** Married employees tend to stay more compared to single or divorced employees.
   * **Attrition = Yes:** Higher attrition rates among single employees compared to married and divorced employees.
10. **Over Time:**
    * **Attrition = No:** Employees who work overtime tend to stay more.
    * **Attrition = Yes:** Higher attrition rates among employees who do not work overtime.
11. **Work-Life Balance:**
    * **Attrition = No:** Higher retention rates among employees with better work-life balance (levels 3 and 4).
    * **Attrition = Yes:** Higher attrition rates among employees with poor work-life balance (levels 1 and 2).
12. **Years With Current Manager:**
    * **Attrition = No:** Employees with more years with their current manager tend to stay more.
    * **Attrition = Yes:** Higher attrition rates among employees with fewer years with their current manager.

**Specific Factors for Attrition = No**

1. **Monthly Income:**
   * Employees with higher monthly incomes tend to stay more.
2. **Performance Rating:**
   * Higher performance ratings are associated with higher retention rates.
3. **Total Working Years:**
   * Employees with more total working years tend to stay more.
4. **Training Times Last Year:**
   * More training sessions in the last year are associated with higher retention rates.
5. **Years Since Last Promotion:**
   * Employees with fewer years since the last promotion tend to stay more.

**Specific Factors for Attrition = Yes**

1. **Job Role:**
   * Higher attrition rates among Sales Representatives, Laboratory Technicians, and Research Scientists.
2. **Total Income Over Years:**
   * Employees with lower total income over the years exhibit higher attrition.
3. **Promotion to Years Ratio:**
   * Lower promotion to years ratio correlates with higher attrition rates.
4. **Average Tenure Per Job Role:**
   * Employees with shorter average tenure per job role show higher attrition rates.

**Conclusion**

From the analysis, the most telling factors for attrition = no include higher monthly income, performance ratings, total working years, training times, and years since the last promotion. For attrition = yes, the key factors include job role, total income over years, promotion to years ratio, and average tenure per job role. Additionally, common factors influencing both retention and attrition include age, department, business travel, distance from home, environment satisfaction, gender, job involvement, job satisfaction, marital status, overtime, work-life balance, and years with the current manager. These insights can help in understanding employee retention and attrition patterns and formulating strategies to address them.

For HR

A screenshot of a computer

Description automatically generated

A screenshot of a computer screen

Description automatically generated

**HR Department**

**Positive Correlations:**

1. **Age (0.54595)**: Older employees are less likely to leave.
2. **YearsWithCurrManager (0.4844)**: Employees with longer tenure with their current manager are more likely to stay.
3. **EducationField (0.48433)**: Specific education fields may be associated with lower attrition.
4. **TotalWorkingYears (0.46802)**: Employees with more years of total working experience are less likely to leave.
5. **YearsInCurrentRole (0.40118)**: Longer tenure in the current role correlates with lower attrition.
6. **IncomePerJobLevel (0.38129)**: Higher income per job level correlates with retention.
7. **WorkLifeSatisfaction (0.37962)**: Higher satisfaction with work-life balance is linked to lower attrition.

**Negative Correlations:**

1. **DistanceFromHome (-0.4294)**: Employees living further away are more likely to leave.

For R&D

A screenshot of a computer

Description automatically generated

A screenshot of a computer screen

Description automatically generated

**Positive Correlations:**

1. **OverTime=No (0.34952)**: Employees not working overtime are less likely to leave.
2. **JobLevel (0.31898)**: Higher job levels are associated with lower attrition.
3. **AvgTenurePerJobRole (0.29968)**: Longer average tenure per job role correlates with lower attrition.
4. **MonthlyIncome (0.29613)**: Higher monthly income is linked to retention.
5. **YearsInCurrentRole (0.27056)**: Employees with longer tenure in their current role are more likely to stay.
6. **TotalWorkingYears (0.26948)**: More total working years correlate with lower attrition.

**Negative Correlations:**

1. **DistanceFromHome (-0.0955)**: Greater distance from home slightly correlates with higher attrition, though it's a weaker negative correlation.

For Sales

A screenshot of a computer

Description automatically generated

A screenshot of a computer code

Description automatically generated

**Sales Department**

**Positive Correlations:**

1. **MaritalStatus=Divorced,Married (0.454614)**: Married or divorced employees are less likely to leave.
2. **JobRole=Manager,Sales\_Executive (0.306753)**: Certain job roles are associated with lower attrition.
3. **AvgTenurePerJobRole (0.298285)**: Longer average tenure per job role is linked to lower attrition.
4. **JobLevel (0.24206)**: Higher job levels correlate with retention.
5. **MonthlyIncome (0.189271)**: Higher monthly income is linked to retention.
6. **WorkLifeSatisfaction (0.188325)**: Higher satisfaction with work-life balance correlates with lower attrition.

**Negative Correlations:**

1. **DistanceFromHome (-0.117954)**: Employees living further away are more likely to leave.

1b 16.54% of staff are leaving

**Task 1b: Analysis Report**

**Introduction**

The goal of this task is to compare the characteristics of employees who stayed (Attrition = No) versus those who left (Attrition = Yes). This comparison involves calculating the mean averages of numerical features and the percentage distributions of categorical features for both groups.

**Categorical feature identification followed by percentage distribution of each category within the attrition no group and the attrition yes group**.

Percentage Distributions Comparison

| Feature | Category | Attrition = No | Attrition = Yes |
| --- | --- | --- | --- |
| BusinessTravel | Travel\_Rarely | 68.44% | 66.59% |
|  | Travel\_Frequently | 24.22% | 28.32% |
|  | Non-Travel | 7.34% | 5.09% |
| Department | Research & Development | 79.95% | 56.18% |
|  | Sales | 16.84% | 40.66% |
|  | Human Resources | 3.21% | 3.15% |
| EducationField | Life Sciences | 50.15% | 38.28% |
|  | Medical | 27.50% | 35.61% |
|  | Marketing | 10.34% | 14.64% |
|  | Technical Degree | 7.72% | 6.33% |
|  | Other | 2.25% | 2.53% |
|  | Human Resources | 2.03% | 2.41% |
| Gender | Male | 59.24% | 63.68% |
|  | Female | 40.76% | 36.32% |
| JobRole | Research Scientist | 21.52% | 19.90% |
|  | Sales Executive | 21.43% | 25.91% |
|  | Laboratory Technician | 16.75% | 18.48% |
|  | Manufacturing Director | 13.97% | 13.31% |
|  | Healthcare Representative | 9.71% | 6.76% |
|  | Sales Representative | 7.01% | 9.01% |
|  | Human Resources | 4.92% | 3.10% |
|  | Manager | 4.70% | 3.52% |
| MaritalStatus | Married | 47.87% | 39.85% |
|  | Single | 29.79% | 47.37% |
|  | Divorced | 22.34% | 12.78% |
| OverTime | No | 76.61% | 22.99% |
|  | Yes | 23.39% | 77.01% |

**Analysis**

1. **BusinessTravel**: Employees who left had a higher percentage of frequent travelers (28.32%) compared to those who stayed (24.22%).
2. **Department**: A higher percentage of employees who left were from Sales (40.66%) compared to those who stayed (16.84%).
3. **EducationField**: Those who stayed had a higher percentage in Life Sciences (50.15%) compared to those who left (38.28%).
4. **Gender**: A higher percentage of males left (63.68%) compared to those who stayed (59.24%).
5. **JobRole**: The highest percentage of employees who left were Sales Executives (25.91%) compared to those who stayed where Research Scientists were the highest (21.52%).
6. **MaritalStatus**: More single employees left (47.37%) compared to those who stayed (29.79%).
7. **OverTime**: A higher percentage of employees who left worked overtime (77.01%) compared to those who stayed (23.39).

attrition = no for age

A screenshot of a computer

Description automatically generated

A screenshot of a data

Description automatically generated

Attrition = yes for age

A screenshot of a computer

Description automatically generated

| **Attribute** | **Mean (Attrition = No)** | **Mean (Attrition = Yes)** |
| --- | --- | --- |
| Age | 37.9 | 34 |
| DailyRate | 810 | 759 |
| DistanceFromHome | 9.12 | 10.7 |
| EnvironmentSatisfaction | 2.75 | 2.47 |
| HourlyRate | 66 | 65.9 |
| JobLevel | 2.14 | 1.6 |
| MonthlyIncome | 6828.85 | 4724.38 |
| MonthlyRate | 14255.7 | 14447.44 |
| NumCompaniesWorked | 2.7 | 3 |
| TotalWorkingYears | 11.88 | 8.03 |
| TrainingTimesLastYear | 2.83 | 2.63 |
| YearsAtCompany | 7.37 | 5 |
| YearsInCurrentRole | 4.5 | 2.8 |
| YearsSinceLastPromotion | 2.22 | 1.87 |
| YearsWithCurrManager | 4.33 | 2.81 |
| IncomePerJobLevel | 3010 | 2757 |
| TotalIncomeOverYears | 776,402 | 453,514 |
| AvgTenurePerJobRole | 7.22 | 5.76 |
| WorkLifeSatisfaction | 2.77 | 2.5 |
| ProportionCareerAtCompany | 0.6 | 0.5 |

**A**

**nalysis of Patterns**

1. **Age and Experience:**
   * Employees who stayed (Attrition = No) are older and have more total working years, indicating that more experienced employees are less likely to leave.
   * Those who left (Attrition = Yes) have a higher average DistanceFromHome, suggesting that longer commutes might contribute to attrition.
2. **Job and Income:**
   * Employees who stayed have higher JobLevels and MonthlyIncome. This indicates that higher positions and better compensation are associated with retention.
   * Those who left have worked at more companies previously (NumCompaniesWorked), possibly indicating a trend of job-hopping.
3. **Satisfaction and Balance:**
   * All satisfaction measures (Environment, Job, Relationship, WorkLife, AvgSatisfaction) are higher for employees who stayed. This underscores the importance of job satisfaction in employee retention.
   * Employees who left have slightly lower WorkLifeBalance scores, which could indicate that work-life conflict is a factor in attrition.
4. **Promotion and Career Development:**
   * Employees who stayed have longer tenures (YearsAtCompany, YearsInCurrentRole, YearsWithCurrManager) and slightly higher TrainingTimesLastYear.
   * The PromotionToYearsRatio is the same for both groups, suggesting that the rate of promotions does not differ significantly between those who stay and those who leave.
5. **Hours Worked:**
   * Employees who stayed work slightly more hours per day on average, which might reflect a higher engagement level or a higher workload in more senior positions.
6. **Income Over Time:**
   * The TotalIncomeOverYears is significantly higher for employees who stayed, indicating that those who remain with the company benefit from cumulative earnings over time.

**Conclusion**

Overall, the data suggests that age, job level, income, satisfaction, and tenure are significant factors influencing whether an employee stays or leaves. Improving job satisfaction, providing career development opportunities, and addressing work-life balance could be key strategies in reducing attrition. Additionally, offering competitive salaries and considering the impact of commute distances might help in retaining employees.

**Percentage of the Dataset**

To calculate the percentage of the dataset each group represents, we need to know the total number of instances and the number of instances in each group.

Based on the provided datasets:

* **Total Instances**: 4422 (Attrition = Yes + Attrition = No)
* **Instances with Attrition = No**: 165
* **Instances with Attrition = Yes**: 3261

**Percentage Calculation:**

* **Attrition = No**: 1654422×100≈3.73%\frac{165}{4422} \times 100 \approx 3.73\%4422165​×100≈3.73%
* **Attrition = Yes**: 32614422×100≈73.73%\frac{3261}{4422} \times 100 \approx 73.73\%44223261​×100≈73.73%

**Summary**

* **Average Person Leaving (Attrition = Yes)**: Younger, lower job level, lower monthly income, fewer total working years, shorter tenure with current manager, higher distance from home.
* **Average Person Staying (Attrition = No)**: Older, higher job level, higher monthly income, more total working years, longer tenure with current manager, shorter distance from home.
* **Percentage of Dataset**: 3.73% staying, 73.73% leaving.

1c – used entire training dataset

For Age:

A screen shot of a computer

Description automatically generated

For Job Level:

A screenshot of a computer screen

Description automatically generated

For Monthly Income:

A white screen with black text

Description automatically generated

For Hours Worked Per Day

A screenshot of a computer screen

Description automatically generated

**Interpretation of Results**

* **JobLevel and MonthlyIncome:** There is a very strong positive correlation (0.9485). This indicates that as job level increases, monthly income also increases significantly. This is expected as higher job levels typically come with higher salaries.
* **Age and JobLevel:** There is a moderate positive correlation (0.5231). This suggests that as age increases, job level tends to be higher, likely due to accumulated experience over time.
* **Age and MonthlyIncome:** There is a moderate positive correlation (0.5106). Older employees tend to have higher monthly incomes, which can be attributed to their higher job levels and experience.
* **HoursWorkedPerDay and the other attributes:** The correlations are very weak (close to 0), indicating that the number of hours worked per day does not significantly vary with age, job level, or monthly income.

**Conclusion**

* The most significant relationship is between **JobLevel and MonthlyIncome**.
* **Age** also shows a meaningful positive correlation with both **JobLevel** and **MonthlyIncome**.
* **HoursWorkedPerDay** shows very weak correlation with the other attributes, indicating that it is not significantly influenced by job level, age, or income.

1c (i)

Carried out correlation analysis to find the best attributes

**Job Level 3, 4, and 5 Combined**

* **Highest Correlation with Job Level**:
  + MonthlyIncome (0.793)
  + TotalIncomeOverYears (0.602)
  + AvgTenurePerJobRole (0.587)
  + YearsInCurrentRole (0.469)
  + YearsWithCurrManager (0.425)
  + YearsSinceLastPromotion (0.368)

**Job Level 4 and 5 Combined**

* **Highest Correlation with Job Level**:
  + TotalIncomeOverYears (0.759)
  + MonthlyIncome (0.686)
  + AvgTenurePerJobRole (0.496)
  + YearsInCurrentRole (0.417)
  + YearsSinceLastPromotion (0.391)
  + YearsWithCurrManager (0.357)

**Job Level 5**

* **Highest Correlation with Job Level**:
  + TotalIncomeOverYears (0.734)
  + MonthlyIncome (0.630)
  + AvgTenurePerJobRole (0.545)
  + YearsInCurrentRole (0.418)
  + YearsSinceLastPromotion (0.277)
  + YearsWithCurrManager (0.262)

Carried out j48 on each job level to determine how to retain the employees leaving by analysing the trees:

The decision tree analysis using the J48 classifier in WEKA provides insight into the key factors influencing employee attrition for high-value employees (those with job levels 3, 4, and 5). Here’s how this analysis addresses Task 1c(i) and provides actionable insights:

**Task 1c(i) Analysis**

**Objective:**

Consider attributes within the data set that have the potential to persuade high-value individuals leaving the organisation to stay. **Based on the results of the decision trees and PART classifier, there is a clear trend indicating that monthly income is a significant factor influencing attrition, especially as job levels increase.**

A screenshot of a computer

Description automatically generated**Results and Interpretation:**

1. A screenshot of a computer

   Description automatically generated**Job Level 3, 4, 5:**

J48 AND PART

* + **Key Attribute: Monthly Income**
    - **Split Point:** 6796
    - Employees with a monthly income less than or equal to 6796 are less likely to leave (classified as 0 - non-attrition).
    - Employees with a monthly income greater than 6796 are further split based on other attributes such as YearsWithCurrManager, YearsInCurrentRole, and YearsSinceLastPromotion.
    - This indicates that increasing the monthly income for employees earning less than 6796 may help in retaining them.
  + **Additional Attributes:**
    - **YearsWithCurrManager:** Employees with fewer years with their current manager (< 1 year) are more likely to leave.
    - **YearsInCurrentRole:** Those with fewer years in their current role (< 4 years) are less likely to leave.
    - **YearsSinceLastPromotion:** More time since the last promotion (> 5 years) increases the likelihood of leaving.

**Actionable Insight:** Focus on improving monthly income, ensuring timely promotions, and managing tenure with the current manager to retain high-value employees.

1. A white screen with black text

   Description automatically generatedA screenshot of a computer

   Description automatically generated**Job Level 4, 5:**

J48 and PART

* + **Key Attribute: Monthly Income**
    - **Split Point:** 10650
    - Employees with a monthly income less than or equal to 10650 are less likely to leave.
    - Those earning more than 10650 are split based on AvgTenurePerJobRole.
  + **Additional Attributes:**
    - **AvgTenurePerJobRole:** Lower tenure per job role (< 8.18 years) is associated with a higher likelihood of leaving.
  + **Actionable Insight:** Ensure competitive salaries for employees at higher job levels and provide opportunities for role transitions to avoid long tenures in the same role.

1. **Job Level 5:**

A number and symbols on a white background

Description automatically generatedA close-up of a number

Description automatically generated

* + **Key Attribute: Monthly Income**
    - **Split Point:** 13758
    - Employees with a monthly income greater than 13758 are more likely to leave.
    - Those earning less than or equal to 13758 are classified as non-attrition.
  + **Actionable Insight:** For top-level employees (job level 5), focus on factors other than salary, such as career growth, job satisfaction, and work-life balance, to retain them.

**Conclusion and Recommendations:**

The J48 decision tree analysis indicates that monthly income is a critical factor influencing employee attrition, especially for high-value employees. However, other factors such as years with the current manager, years in the current role, years since the last promotion, and average tenure per job role also play significant roles. To persuade high-value individuals to stay, the organization should:

1. **Increase Monthly Income:**
   * Ensure that the salaries are competitive and reflect the market standards for high-value employees.
2. **Timely Promotions:**
   * Implement a transparent and fair promotion policy to ensure employees do not feel stagnant in their roles.
3. **Manage Tenure with Current Manager:**
   * Encourage regular feedback sessions and strong manager-employee relationships to improve job satisfaction.
4. **Monitor Tenure in Current Roles:**
   * Provide opportunities for lateral moves or new challenges within the organization to prevent burnout and stagnation.

By focusing on these strategies, the organization can better retain its high-value employees and reduce attrition rates.

**Conclusion**

**Key Observations:**

1. **Job Level 3, 4, 5:**
   * A variety of factors influence attrition, with **monthly income** being a major factor, but other attributes like years in current role, years since last promotion, and total income over years also play significant roles.
2. **Job Level 4, 5:**
   * The number of influencing factors reduces, with **monthly income** and **average tenure per job role** being more prominent.
3. **Job Level 5:**
   * Attrition is predominantly influenced by **monthly income**, with higher salaries correlating significantly with attrition rates.

BDA Summative Assessment

TASK 1

Step 1 Data preprocessing

Data cleaning

Handle missing values

Data training set

1. Age has 136 out of 2499 records missing so replacing with median, otherwise all ages look reasonable within 18-60 which are working ages
   1. Replaced with median value
2. Attrition has no missing values and is only yes or no, but in the test set there meaning it may be hard to test attrition
3. There only five missing for business travel so can be replaced with impunity
4. Daily rate has 27 missing rates only, so can use impunity, it also ranges gradually from 101 to 1498 which seems reasonable so no anomalies
5. Department has no missing values
6. Distance from home has 95 of 2499 records missing so can also use impunity to replace, ranges between 1 to 95 which is also reasonable
7. Education ranges between 1-5, unsure what this means, but no missing values
8. EducationField has no missing values, it has 148 records with ‘Other’ and 5 other fields
9. Employee count is just 1 for everyone, offers nothing to the investigation and can be removed
10. Employee number is just their ID, unsure if this is useful to the investigation - can it be removed?
11. EnvironmentSatisfaction has no missing values and they range from 1-4 - appears relevant to an investigation into staff turnover
12. Gender has no missing values,  only male or female - 1000 out of 2499 employees are female, approx 40%
13. HourlyRate has no missing values, varies from 30-100 which seems valid
14. Job involvement is measured as 1-4
15. JobLevel is measured as 1-5
16. JobRole has 9 options and appears relevant
17. JobSatisfaction has a 1-4 measurement - appears relevant to investigation
18. Marital status has 5 missing values, can be replaced using impunities, but unsure how relevant marital status is to the investigation
19. MonthlyIncome has no missing values but ranges between 1009 to 19999 which appears valid with no anomalies
20. MonthlyRate has no missing values and ranges from 2094 to 26999 - (private note - unsure the difference between monthly income and monthly rate?)
21. Number of companies worked has a range of 0 - 9. 0 is impossible as they are working at this company so it has to be 1 minimally, perhaps I will replace these values with a 1 or with a median? They make up 335 of 2499 records so 13.4%
22. Over18 seems irrelevant as they are all over 18, will remove this column
23. Overtime has no missing values and is either yes or no
24. PercentSalarySpike is between 11 and 25 and has no missing values
25. PerformanceRating only has values 3 and 4 ? how valuable is this?
26. RelationshipSatisfaction is measured 1-4 and has no missing values
27. StandardHours is 80 for everyone so seems irrelevant and can be removed
28. StockOptionLevel is measured 0, 1, 2, 3 - is it worth changing this 1-4 instead of 0-3? Or does 0 mean they don’t have the option at all? Need to check this. Other than that, no missing values
29. TotalWorkingYears has 0-50 with no missing values, 0 could be realistic if they have been working less than a year but seems unrealistic
30. Training times last year has no missing values and ranges from 0-6
31. WorkLifeBalance is measured between 1-4 with no missing values
32. YearsAtCompany ranges between 0 to 40 which could be realistic as they might have worked for less than a year
33. YearsInCurrentRole ranges between 0 and 18 which is realistic with no missing values
34. YearsSinceLastPromotion ranges from 0 to 15 with no missing values
35. YearsWithCurrentManager ranges from 0-17 with no missing values

**Attributes that are likely important based on domain knowledge:**

* Age, Attrition, BusinessTravel, Department, DistanceFromHome, EnvironmentSatisfaction, JobInvolvement, JobLevel, JobRole, JobSatisfaction, MonthlyIncome, NumCompaniesWorked, OverTime, TotalWorkingYears, WorkLifeBalance, YearsAtCompany, YearsInCurrentRole, YearsSinceLastPromotion, YearsWithCurrManager

**Attributes that might be less important:**

* DailyRate, HourlyRate, MonthlyRate (since MonthlyIncome is already included), Education, EducationField, Gender, PercentSalaryHike, PerformanceRating, RelationshipSatisfaction, StockOptionLevel, TrainingTimesLastYear

Final Data Cleaning for training set

**Imputation of missing values**

Numerical Attributes

* Age, DailyRate, Distance From Home and NumCompanies Worked had their missing values replaced with the median

Categorical Attributes

* BusinessTravel and Marital Status had their columns filled with the mode

Value Adjustment

* 0 in Number of Companies worked is replaced with 1, as they now work at this company so it has to be at least 1

Removal of attributes:

* EmployeeCount: This attribute has the same value for all records, which does not provide any useful information for analysis.
* EmployeeNumber: This is a unique identifier for employees and does not contribute to predicting attrition or other target variables.
* Over18: This attribute has the same value for all records since all employees are over 18, making it redundant.
* StandardHours: This attribute has the same value for all records, making it redundant.

Python code for cleaning the data

import pandas as pd

# Load datasets

train\_data\_path = r'C:\Users\natal\OneDrive\University of York - CS and AI\BDA\BDA Reassessment 23-24 Dataset\personnel\_train.csv'

personnel\_train = pd.read\_csv(train\_data\_path)

# Define the columns to be imputed

numerical\_columns\_to\_impute = ['Age', 'DailyRate', 'DistanceFromHome', 'NumCompaniesWorked']

categorical\_columns\_to\_impute = ['BusinessTravel', 'MaritalStatus']

# Define the function to impute missing values

def impute\_missing\_values(df):

    # Replace missing values for numerical columns with the median

    for column in numerical\_columns\_to\_impute:

        df[column] = df[column].fillna(df[column].median())

    # Replace missing values for categorical columns with the mode

    for column in categorical\_columns\_to\_impute:

        df[column] = df[column].fillna(df[column].mode()[0])

    # Replace NumCompaniesWorked values of 0 with 1

    df['NumCompaniesWorked'] = df['NumCompaniesWorked'].replace(0, 1)

    return df

# Apply the function to the training dataset

personnel\_train\_cleaned = impute\_missing\_values(personnel\_train)

# Remove irrelevant columns

columns\_to\_remove = ['EmployeeCount', 'EmployeeNumber', 'Over18', 'StandardHours']

personnel\_train\_cleaned.drop(columns=columns\_to\_remove, inplace=True)

# Save the cleaned data to a new CSV file

output\_path = r'C:\Users\natal\OneDrive\University of York - CS and AI\BDA\BDA Reassessment 23-24 Dataset\cleaned\_personnel\_train.csv'

personnel\_train\_cleaned.to\_csv(output\_path, index=False)

# Display the first few rows of the cleaned training dataset to confirm changes

print(personnel\_train\_cleaned.head())

**Grouping Variables**

**1. Categorical Variables:**

* **Nominal (No order or ranking):**
  + BusinessTravel
  + Department
  + EducationField
  + Gender
  + JobRole
  + MaritalStatus
  + OverTime
  + Attrition
* **Ordinal (Ordered categories):**
  + Education
  + EnvironmentSatisfaction
  + JobInvolvement
  + JobSatisfaction
  + PerformanceRating
  + RelationshipSatisfaction
  + WorkLifeBalance
  + StockOptionLevel

**2. Numerical Variables:**

* **Discrete:**
  + Age
  + DistanceFromHome
  + JobLevel
  + NumCompaniesWorked
  + TotalWorkingYears
  + TrainingTimesLastYear
  + YearsAtCompany
  + YearsInCurrentRole
  + YearsSinceLastPromotion
  + YearsWithCurrManager
  + StartedAtCompany (Derived)
* **Continuous:**
  + DailyRate
  + HourlyRate
  + MonthlyIncome
  + MonthlyRate
  + PercentSalaryHike
  + IncomePerJobLevel (Derived)
  + TotalIncomeOverYears (Derived)
  + PromotionToYearsRatio (Derived)
  + AvgTenurePerJobRole (Derived)
  + AvgSatisfaction (Derived)
  + WorkLifeSatisfaction (Derived)
  + ProportionCareerAtCompany (Derived)
  + HoursWorkedPerDay (Derived)

**Variable Grouping Table**

|  |  |
| --- | --- |
| **Category** | **Variables** |
| Categorical Nominal | BusinessTravel, Department, EducationField, Gender, JobRole, MaritalStatus, OverTime, Attrition |
| Categorical Ordinal | Education, EnvironmentSatisfaction, JobInvolvement, JobSatisfaction, PerformanceRating, RelationshipSatisfaction, WorkLifeBalance, StockOptionLevel |
| Numerical Discrete | Age, DistanceFromHome, JobLevel, NumCompaniesWorked, TotalWorkingYears, TrainingTimesLastYear, YearsAtCompany, YearsInCurrentRole, YearsSinceLastPromotion, YearsWithCurrManager, StartedAtCompany |
| Numerical Continuous | DailyRate, HourlyRate, MonthlyIncome, MonthlyRate, PercentSalaryHike, IncomePerJobLevel, TotalIncomeOverYears, PromotionToYearsRatio, AvgTenurePerJobRole, AvgSatisfaction, WorkLifeSatisfaction, ProportionCareerAtCompany, HoursWorkedPerDay |

**Factors Likely to Influence Staff Turnover**

**Most Helpful Attributes for Predicting Turnover:**

1. **Job Satisfaction:** Employees with low job satisfaction are more likely to leave.
2. **Environment Satisfaction:** Poor work environment can lead to higher turnover.
3. **Work-Life Balance:** Imbalance can lead to attrition.
4. **Monthly Income:** Employees with higher income might be less likely to leave, but it could also indicate high-paid employees might leave for better opportunities.
5. **OverTime:** Regular overtime can lead to burnout and turnover.
6. **Job Role:** Certain roles might have higher turnover rates.
7. **YearsAtCompany:** Employees with fewer years might be more likely to leave.
8. **Attrition (Target variable):** Directly indicates if an employee left.
9. DailyHoursWorked:
10. Avg Satisfaction
11. BusinessTravel

**Least Helpful Attributes for Predicting Turnover:**

1. **EmployeeCount:** Not useful since it’s constant.
2. **EmployeeNumber:** Unique identifier, not relevant for prediction.
3. **StandardHours:** Not useful since it’s constant.
4. **Over18:** Not useful since it’s constant.
5. **StartedAtCompany:** While interesting, it might not have a strong predictive power compared to other variables.

BEFORE CARRYING OUT SMOTE:

WEKA - Training set

=== Run information ===

Evaluator:    weka.attributeSelection.CfsSubsetEval -P 1 -E 1

Search:       weka.attributeSelection.BestFirst -D 1 -N 5

Relation:     personnel\_train

Instances:    2499

Attributes:   40

              Age

              Attrition

              BusinessTravel

              DailyRate

              Department

              DistanceFromHome

              Education

              EducationField

              EnvironmentSatisfaction

              Gender

              HourlyRate

              JobInvolvement

              JobLevel

              JobRole

              JobSatisfaction

              MaritalStatus

              MonthlyIncome

              MonthlyRate

              NumCompaniesWorked

              OverTime

              PercentSalaryHike

              PerformanceRating

              RelationshipSatisfaction

              StockOptionLevel

              TotalWorkingYears

              TrainingTimesLastYear

              WorkLifeBalance

              YearsAtCompany

              YearsInCurrentRole

              YearsSinceLastPromotion

              YearsWithCurrManager

              HoursWorkedPerDay

              IncomePerJobLevel

              TotalIncomeOverYears

              PromotionToYearsRatio

              AvgTenurePerJobRole

              AvgSatisfaction

              WorkLifeSatisfaction

              StartedAtCompany

              ProportionCareerAtCompany

Evaluation mode:    evaluate on all training data

=== Attribute Selection on all input data ===

Search Method:

Best first.

Start set: no attributes

Search direction: forward

Stale search after 5 node expansions

Total number of subsets evaluated: 484

Merit of best subset found:    0.102

Attribute Subset Evaluator (supervised, Class (nominal): 2 Attrition):

CFS Subset Evaluator

Including locally predictive attributes

Selected attributes: 1,3,6,12,17,20,24,25,27,28,33,34,36,37 : 14

                     Age

                     BusinessTravel

                     DistanceFromHome

                     JobInvolvement

                     MonthlyIncome

                     OverTime

                     StockOptionLevel

                     TotalWorkingYears

                     WorkLifeBalance

                     YearsAtCompany

                     IncomePerJobLevel

                     TotalIncomeOverYears

                     AvgTenurePerJobRole

                     AvgSatisfaction

AFTER CARRYING OUT SMOTE

Selected attributes: 3,4,6,7,9,11,12,13,15,18,19,20,21,22,23,24,26,27,30,31,36,37,38,39 : 24

                     BusinessTravel

                     DailyRate

                     DistanceFromHome

                     Education

                     EnvironmentSatisfaction

                     HourlyRate

                     JobInvolvement

                     JobLevel

                     JobSatisfaction

                     MonthlyRate

                     NumCompaniesWorked

                     OverTime

                     PercentSalaryHike

                     PerformanceRating

                     RelationshipSatisfaction

                     StockOptionLevel

                     TrainingTimesLastYear

                     WorkLifeBalance

                     YearsSinceLastPromotion

                     YearsWithCurrManager

                     AvgTenurePerJobRole

                     AvgSatisfaction

                     WorkLifeSatisfaction

                     StartedAtCompany

Interpretation

Before SMOTE: The focus is on attributes like Age, MonthlyIncome, TotalWorkingYears, YearsAtCompany, and financial metrics (IncomePerJobLevel, TotalIncomeOverYears).

After SMOTE: There's a broader selection of attributes including additional job satisfaction metrics (JobSatisfaction, EnvironmentSatisfaction, RelationshipSatisfaction), performance metrics (PercentSalaryHike, PerformanceRating), and work-life balance indicators (WorkLifeSatisfaction, StartedAtCompany).

* Data analysis again
  + Removal of duplicates, left with 1716 instances from 2499 = 783 potential duplicates, unsure of whether these were duplicate entries or not as id was removed from the attributes. Furthermore, the duplicates are a significant number of dataset (around 32%) so they were retained
  + Stock option level was adjusted to 1-4 from 0-3 for consistency with the other 4 rank ratings
  + Realised there is an imbalance in the dataset, distribution of attrition in the set where 1432 entries are labelled "No" and 284 are labelled "Yes," indicates an imbalance in the dataset. This imbalance can have several implications for the analysis and model building:
    - This class imbalance can affect the performance of the machine learning models, making them biassed towards predicting the majority class (No)
    - Use SMOTE to over sample instances from minority class
      * Now 1652 yes to 2086 no
      * Then 284 yes to 1432 no
* Checking for missing values
  + Missing values were imputed with mean and modes depending on their context
* Discretise continuous attributes
* Correlation analysis
* Visualise relationships
  + Scatter plots
  + Histograms
* Build a classifier
  + J48
  + SMO
  + RandomForest
  + NaiveBayes(?)
  + Can configure the classifier - change parameters
  + Evaluate classifier results
    - Confusion matrix
    - Precision, recall, F-Measure, ROC

**OneR Attribute Evaluation**

* **Top Attributes**:
  + AvgSatisfaction
  + WorkLifeSatisfaction
  + PercentSalaryHike
  + RelationshipSatisfaction
  + TrainingTimesLastYear
  + EnvironmentSatisfaction
  + JobSatisfaction

**Gain Ratio Attribute Evaluation**

* **Top Attributes**:
  + StockOptionLevel
  + RelationshipSatisfaction
  + JobInvolvement
  + EnvironmentSatisfaction
  + WorkLifeBalance
  + JobSatisfaction
  + Education

**Interpretation**

* Both evaluators highlight the importance of satisfaction-related attributes (AvgSatisfaction, WorkLifeSatisfaction, EnvironmentSatisfaction, JobSatisfaction).
* Gain Ratio places a higher emphasis on StockOptionLevel, JobInvolvement, and PerformanceRating.
* OneR highlights PercentSalaryHike, TrainingTimesLastYear, and AvgTenurePerJobRole.

A screenshot of a graph

Description automatically generated

A screenshot of a graph

Description automatically generated

1. **Task 1a: Predict Actions Leading to High Staff Turnover by Department**
   * **Split department into its own sub value and then carry out attribute selection for each department to find the best predictors**
   * **Key Attributes**:
     + Department
     + JobRole
     + BusinessTravel
     + JobInvolvement
     + OverTime
     + JobSatisfaction
     + EnvironmentSatisfaction
     + RelationshipSatisfaction
     + StockOptionLevel
     + YearsAtCompany
     + TotalWorkingYears
2. **Task 1b: Compare Percentage and Mean Average of Staff Remaining and Leaving**
   * **Key Attributes**:
     + Attrition (Target variable)
     + Age
     + MonthlyIncome
     + JobLevel
     + YearsAtCompany
     + TotalWorkingYears
3. **Task 1c: Determine Correlation between Job Level, Working Hours, Monthly Income, and Age**
   * **Key Attributes**:
     + JobLevel
     + HoursWorkedPerDay
     + MonthlyIncome
     + Age
4. **Task 1c.i: Consider Attributes to Persuade High-Value Individuals to Stay**
   * **Key Attributes**:
     + MonthlyIncome
     + JobSatisfaction
     + EnvironmentSatisfaction
     + RelationshipSatisfaction
     + WorkLifeBalance
     + PromotionToYearsRatio
     + TotalWorkingYears
     + YearsAtCompany

Summary of Attribute Selection

From the visualisations and the attribute selection methods (OneR and Gain Ratio), the following attributes consistently show importance and are likely to be influential in predicting staff turnover:

* AvgSatisfaction (Consistently ranked high in both OneR and Gain Ratio)
* WorkLifeSatisfaction (High ranking in both methods)
* PercentSalaryHike
* RelationshipSatisfaction
* TrainingTimesLastYear
* EnvironmentSatisfaction
* JobSatisfaction
* Education
* DistanceFromHome
* AvgTenurePerJobRole
* YearsInCurrentRole
* WorkLifeBalance
* YearsWithCurrManager
* JobInvolvement
* NumCompaniesWorked
* StockOptionLevel
* JobLevel
* YearsAtCompany
* TotalWorkingYears
* Age
* TotalIncomeOverYears
* StartedAtCompany
* YearsSinceLastPromotion
* OverTime
* MonthlyIncome
* PerformanceRating
* JobRole
* MaritalStatus
* IncomePerJobLevel
* ProportionCareerAtCompany
* HourlyRate
* DailyRate
* HoursWorkedPerDay
* BusinessTravel
* Gender

Recommendations for Attribute Selection

To create a predictive model that effectively determines the likelihood of staff attrition, consider including the following attributes, as they have shown relevance across different evaluations and visual analyses:

* AvgSatisfaction
* WorkLifeSatisfaction
* PercentSalaryHike
* RelationshipSatisfaction
* TrainingTimesLastYear
* EnvironmentSatisfaction
* JobSatisfaction
* Education
* DistanceFromHome
* AvgTenurePerJobRole
* YearsInCurrentRole
* WorkLifeBalance
* YearsWithCurrManager
* JobInvolvement
* NumCompaniesWorked
* StockOptionLevel
* JobLevel
* YearsAtCompany
* TotalWorkingYears
* Age
* TotalIncomeOverYears
* StartedAtCompany
* YearsSinceLastPromotion
* OverTime
* MonthlyIncome
* PerformanceRating
* JobRole
* MaritalStatus
* IncomePerJobLevel
* ProportionCareerAtCompany
* HourlyRate
* DailyRate
* HoursWorkedPerDay
* BusinessTravel
* Gender

Next Steps

1. Model Building: Use the selected attributes to build classification models (e.g., Decision Trees, Random Forest, SVM) to predict staff attrition.
2. Analysis: Evaluate the model performance using metrics such as accuracy, precision, recall, and F1-score. Compare different models and choose the best-performing one.
3. Correlation Analysis: Specifically for Task 1c, perform correlation analysis between Job Level, Working Hours, Monthly Income, and Age using correlation coefficients or scatter plots.
4. Reporting: Document the preprocessing steps, model building, and evaluation results. Include visualisations and interpretations to support your findings.

I made a mistake before where SMOTE had reversed on WEKA. The new results for each department are as follows: hr had a class imbalance of 85 no to 20 yes  which using SMOTE became 85 no to 80 yes. R&D Department had a class imbalance of 1405 no to 232 yes which became 1405 no to 1856 yes. sales had a class imbalance of yes 161 and no 596 but this became 644 yes and no 596. For all sets, I then carried out nominal to binary so the features were divided and I could see which was more so associated with attrition more specifically within the feature. I then performed attribute evaluation using bestfirst CFSSubset, correlation eval and infogainratio to identify the correlations, i then converted, via choosing only attrition through attribute indices, attrition to binary using supervised nominal to binary so that i can use more classifiers such as j48.