COM00148M

Department of Computer Science

Big Data Analytics

SUMMATIVE REASSESSMENT

Word Count: 2994

Task 1: 1668 words

Task 2: 707 words

Task 3: 619 words

Contents

[Task 1: Staff Turnover Investigation 3](#_Toc171858933)

[Task 1a: Key Attributes Predicting Staff Turnover and Retention 5](#_Toc171858934)

[Task 1b: Comparison of Percentage and Mean Average of Staff Remaining and Leaving 6](#_Toc171858935)

[Task 1c: Correlation Between Job Level, Working Hours, Monthly Income, and Age 7](#_Toc171858936)

[Task 1c(i): Persuading High-Value Individuals to Stay 7](#_Toc171858937)

[Task 2: Storing Data and Possible Solutions 8](#_Toc171858938)

[Part 1: Designing a Relational Database 8](#_Toc171858939)

[Part 2: Scaling the Database 8](#_Toc171858940)

[Task 3: Considering a Human Resource Management System 10](#_Toc171858941)

[Appendices 12](#_Toc171858942)

[References 17](#_Toc171858943)

## 

## Task 1: Staff Turnover Investigation

Assumptions

Several assumptions were made during this analysis. Firstly, significant class imbalances in the dataset were addressed using SMOTE (Synthetic Minority Over-sampling Technique) to ensure balanced representation of both attrition classes [1]. Secondly, it was assumed that converting attributes to binary form using the NominalToBinary filter would facilitate detailed correlation analysis with attrition. Lastly, various attribute evaluation techniques were assumed to effectively identify the most significant attributes for predicting attrition.

Data Preprocessing

The data preprocessing phase involved several critical steps to prepare the dataset for analysing staff turnover. Initially, Excel was used to review the dataset, identifying columns with missing values and redundant features. Redundant features such as Over18 and StandardHours, which had the same value for all employees, were removed as they provided no useful information. EmployeeNumber, serving as a unique identifier, was deemed irrelevant for the attrition investigation and was also removed. Continuous attributes with missing values, including Age, DistanceFromHome, and DailyRate, were identified, alongside nominal attributes like BusinessTravel and MaritalStatus.

Python was employed for data cleaning, where missing values for continuous attributes were replaced with the median, ensuring the central tendency was maintained without being skewed by extreme values [2]. Nominal attributes had missing values replaced with the mode, the most frequently occurring value in the dataset. For the test set, missing values were replaced with '?', which later led to the test set being unused due to the lack of the target variable Attrition, resulting in time loss. Duplicate removal was considered but not pursued after the removal of EmployeeNumber as it was unclear if the duplicates being deleted were duplicates.

Histograms were used to visualise attribute distributions and check for outliers, revealing no significant outliers. However, entries with impossible HoursWorkedPerDay (calculated from HourlyRate and DailyRate resulting in more than 24 hours) were removed. Class imbalance was identified in Weka, with significant disparities between Attrition = Yes and Attrition = No. SMOTE was applied to balance the classes, adjusting the proportions within HR, R&D, and Sales departments to ensure a more balanced dataset for analysis.

Data transformations using NominalToBinary and NumericToNominal were applied to better evaluate categorical attributes like BusinessTravel, allowing for a more detailed analysis of specific categories’ correlation with attrition. The dataset was split into subsets for each department (HR, R&D, Sales) and for Attrition = Yes and Attrition = No groups, facilitating focused and comparative analysis in Task 1b. Mean averages for continuous variables were calculated using Weka’s summary feature when Attrition was set as the class, and percentage distributions for categorical features were calculated for both Attrition = Yes and Attrition = No groups. This involved identifying categorical features, calculating percentage distributions within each group, and comparing these distributions.

Justification for Using J48 and SMO Classifiers

J48 Decision Trees were chosen for their ability to uncover interpretable patterns within the data for correlations and independencies [3]. They provide a visual representation of the decision rules, making it easier to understand and communicate the factors influencing employee attrition and retention. J48 is particularly effective in handling categorical data and capturing non-linear relationships. SMO (Support Vector Machine) was selected for its robustness in handling high-dimensional data and its ability to classify instances accurately [4]. SMO is effective in finding the optimal hyperplane that separates the classes, making it a powerful tool for binary classification tasks like predicting employee attrition. It helps in capturing the complex relationships between the attributes and the target variable. A significant limitation of this analysis was the inability to use the test set due to the target variable having a high percentage of missing values. This limitation means the machine learning models could not be validated on an independent test set, potentially affecting the reliability of the predictive models. To address this, cross-validation and percentage split methods were employed to evaluate the classifiers, but the lack of an independent test set remains a caveat. To further evaluate which predictive model performs best, a third classifier, such as NaiveBayes or LogisticRegression, could be introduced. These additional classifiers would provide further insights and comparisons, ensuring a more robust evaluation of the models’ predictive power.

### Task 1a: Key Attributes Predicting Staff Turnover and Retention

The evaluation of the data across the HR, Sales, and R&D departments identified key attributes that predict employee attrition and retention. The results were obtained using various evaluation techniques InfoGainAttributeEval and CfsSubsetEval however CorrelationAttributeEval was focused upon (Appendix A, figures 4-9). Additionally, classifiers J48 and SMO were employed to further analyse the data (Appendix A, figures 1-3).

In the HR department, attributes like Age and YearsWithCurrManager were highly predictive of staff turnover (Appendix), 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. SMO classifier performed best for the HR department with a high accuracy rate.

For the R&D department, attributes like OverTime, JobLevel, and AvgTenurePerJobRole were significant predictors of turnover (Appendix). 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. J48 classifier performed best for the R&D department with high accuracy and precision.

In the Sales department, MaritalStatus and JobRole 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, emphasising the importance of job fulfillment and work-life balance. J48 classifier performed best for the Sales department with excellent classification accuracy.

Splitting the data by departments (HR, Sales, and Research &Development) was crucial for identifying department-specific factors influencing attrition and retention. Different departments have unique characteristics and job roles, making it essential to analyse them separately to derive meaningful insights. CorrelationAttributeEval was selected for its ability to identify linear relationships between attributes and the target variable. This method provided a clear understanding of which factors are most strongly associated with attrition and retention. InfoGainAttributeEval was used to assess the importance of each attribute based on the information gain metric, which helps in understanding the reduction in entropy or uncertainty [5]. CfsSubsetEval was employed to identify the best subset of attributes that together have the highest predictive power, considering the inter-correlation among attributes.

The combined use of CorrelationAttributeEval, InfoGainAttributeEval, and CfsSubsetEval, along with classifier techniques like J48 and SMO, provided a thorough analysis of the factors affecting employee attrition and retention across different departments.

### Task 1b: Comparison of Percentage and Mean Average of Staff Remaining and Leaving

To compare the percentage and mean average of staff remaining in the organisation versus those leaving it, a table was created with percentage distributions for categorical features and mean averages for numeric features (Appendix A, figure 10-13).

For categorical features, the comparison revealed significant differences. For instance, employees who left had a higher percentage of frequent travelers (28.32%) compared to those who stayed (24.22%). A higher percentage of employees who left were from Sales (40.66%) compared to those who stayed (16.84%). Additionally, more single employees left (47.37%) compared to those who stayed (29.79%).

For numeric features, the analysis showed that employees who stayed had higher mean values for attributes such as Age (37.88 vs. 34), MonthlyIncome (6828.85 vs. 4724.38), and TotalWorkingYears (11.88 vs. 8.03). Conversely, those who left had higher mean values for DistanceFromHome (10.7 vs. 9.12) and NumCompaniesWorked (3 vs. 2.7).

These comparisons highlighted the importance of attributes such as age, job level, income, and satisfaction in predicting staff retention and turnover.

### Task 1c: Correlation Between Job Level, Working Hours, Monthly Income, and Age

The task aimed to determine the presence of a correlation between job level, working hours, monthly income, and age. The correlation analysis revealed several significant relationships (Appendix A, figures 16-19). There was a very strong positive correlation between JobLevel and MonthlyIncome (0.9485), indicating that as job level increases, monthly income also increases significantly. Age showed a moderate positive correlation with both JobLevel (0.5231) and MonthlyIncome (0.5106). However, the correlation between HoursWorkedPerDay and other attributes was weak, suggesting that working hours do not significantly vary with job level, age, or income.

These findings provided insights into the relationships between key attributes, which can inform targeted interventions to improve employee retention.

### Task 1c(i): Persuading High-Value Individuals to Stay

This task aimed to identify attributes that could persuade high-value individuals to stay with the organisation. Correlation analysis was carried out to find the best attributes for retaining high-value individuals.

The J48 classifier in Weka provided insights into key factors influencing employee attrition and were supported by the rules provided by PART (Appendix, figures 14 and 15). For job levels 3, 4, and 5, monthly income was a significant factor, with split points indicating that employees with lower incomes were less likely to leave. Additional attributes such as YearsWithCurrManager, YearsInCurrentRole, and YearsSinceLastPromotion also played significant roles. For job levels 4 and 5, lower monthly income and shorter average tenure per job role were associated with a higher likelihood of leaving. For job level 5, the analysis indicated that factors other than salary, such as career growth, job satisfaction, and work-life balance, were more significant in retaining employees. These insights suggest that improving monthly income, ensuring timely promotions, and managing tenure with the current manager can help retain high-value employees.

This report identifies key attributes influencing staff turnover and retention across three departments. The preprocessing steps and analysis techniques provided a comprehensive understanding of the factors contributing to employee attrition. Significant attributes were identified, and their correlations with attrition were analysed. The results indicate that improving job satisfaction, providing career development opportunities, and addressing work-life balance are key strategies in reducing attrition. Additionally, offering competitive salaries and considering the impact of commute distances can help retain employees. Future work should focus on validating these findings with real-world data and exploring additional preprocessing techniques to further enhance model performance.

## Task 2: Storing Data and Possible Solutions

### Part 1: Designing a Relational Database

In developing a relational database to manage the 'Personnel' dataset, several key considerations were made to ensure optimal data management, scalability, and compliance with normalisation principles. The goal was to create a structure that minimises redundancy and maintains data integrity.

A significant feature of the database design and ER diagram (Appendix B1) is the separation of distinct entities, such as Employee, Job, Financials, Satisfaction, Department, WorkLocation, and HomeLocation. This strategy was chosen to reduce redundancy and enhance data integrity. For example, separating the Job entity allows us to accurately capture job-specific attributes, which can be linked to employees without duplicating data. This separation aligns with the principles used in SQL database management, where each table represents a single entity or concept, ensuring a maintainable database structure (Appendix B2).

The normalisation process played a crucial role in refining the database structure and reducing data redundancy. By decomposing larger tables into smaller, manageable ones, each adhering to 3NF, we ensured that the data was organised logically and consistently. For instance, the Employee table includes core employee details and links to other tables through foreign keys, facilitating a clean and efficient data structure.

Selecting appropriate primary keys was critical for maintaining data integrity. Primary keys, such as EmployeeID for the Employee table and composite keys for location entities, uniquely identify each record, ensuring efficient indexing and querying. For example, using EmployeeID as the primary key in the Employee table helps to connect each employee's records across various related tables seamlessly.

### Part 2: Scaling the Database

In response to the HR manager's long-term goals for improving human resources across the nation, adopting a cloud-based relational database management system (RDBMS) emerges as a robust solution for efficiently handling large-scale datasets [6]. This proposal advocates for leveraging Amazon Web Services (AWS) RDS, coupled with PostgreSQL, a powerful open-source RDBMS renowned for its reliability, scalability, and extensive feature set [7].

The AWS RDS solution with PostgreSQL offers scalability by providing scalable compute and storage resources, enabling the HR manager to adjust capacity based on workload demands. As the dataset grows, PostgreSQL's scalability ensures seamless handling of these variables across various locations. PostgreSQL's ACID compliance ensures data integrity and reliability, critical for managing sensitive HR data across diverse offices [8] [9]. With features like transactions and constraints, PostgreSQL enforces data consistency and integrity at the database level.

Real-time event processing can be seamlessly integrated with PostgreSQL using AWS services such as Amazon Kinesis Data Streams and AWS Lambda [10]. For instance, automated analysis can detect a significant increase in staff turnover rates, triggering immediate actions. AWS Lambda functions can process the analysis results in real-time, determining if predefined thresholds have been surpassed [11]. If so, the system can automatically generate and send notifications to relevant stakeholders, enabling the business to respond promptly to changing dynamics and make informed decisions.

When comparing the proposed cloud-based RDBMS solution with alternative approaches, notable differences emerge. Unlike flat-file systems, which often lack scalability and robust data management features, the AWS RDS with PostgreSQL solution offers dynamic scalability and comprehensive data management capabilities tailored to the complexities of the dataset. Traditional relational databases provide robust data management capabilities but may not match the scalability and flexibility offered by cloud-based RDBMS platforms like AWS RDS.

On the other hand, compared to big data solutions like Hadoop and Kafka, the cloud-based RDBMS approach offers a more structured and familiar environment for data management and analysis. While Hadoop excels in distributed data processing and storage, its learning curve and infrastructure requirements may pose challenges for organisations not well-versed in big data technologies [12]. Similarly, Kafka's real-time event processing capabilities are powerful but may require additional complexity and overhead for integration with traditional RDBMS systems [13].

In conclusion, the deployment of AWS RDS with PostgreSQL presents a compelling solution for the HR manager’s long-term expansion efforts. By leveraging cloud-based infrastructure and a proven open-source RDBMS, the solution offers scalability, reliability, and real-time capabilities tailored to the challenges of handling large datasets. With the ability to manage data efficiently, ensure reliability, and react to real-time events, the HR manager can navigate the complexities of human resource management effectively and drive business growth.

## Task 3: Considering a Human Resource Management System

The HR manager is contemplating the development of a smart human resource management system (HRMS) aimed at enhancing staff learning and professional development in a personalised and engaging manner. This system is envisioned to facilitate the easy scheduling of mandatory training while integrating long-term objectives such as leadership development into employees' calendars. Additionally, the HR manager intends to optimise the recruitment process to attract top talent by capturing personal details of potential candidates via an online form. This HRMS would support applicant tracking and streamline the succession planning process. Before initiating this venture, it is essential to identify and address significant privacy issues to ensure the protection of personal data and compliance with legal requirements.

One of the most critical privacy issues is ensuring the lawful collection, storage, and processing of personal data in compliance with data protection laws, particularly the General Data Protection Regulation (GDPR). GDPR sets forth stringent guidelines on handling personal data, emphasising principles such as data minimisation, purpose limitation, and robust data security measures. Non-compliance with these regulations can result in hefty fines and reputational damage as in 2022 when Google were fined $57M [14]. To address this issue, the HR manager must implement several strategies, including data anonymisation to minimise the risk of exposure. Conducting regular compliance audits ensures ongoing adherence to GDPR and other relevant data protection laws. Establishing clear data policies that outline how personal data is collected, used, stored, and disposed of is also crucial.

Storing sensitive personal details in a digital environment heightens the risk of data breaches and unauthorised access. Data breaches can lead to identity theft, financial loss, and significant legal repercussions as in the Equifax data breach of 2017 which led to a $425M settlement [15]. To mitigate these risks, the HR manager should implement data encryption to ensure that all personal data is encrypted both in transit and at rest. Stringent access control mechanisms should be established to limit access to sensitive data to authorised personnel only. Regular security audits and staff training are essential to maintain a high level of data security awareness and readiness to address potential threats [16].

Securing informed and explicit consent from employees and job applicants before collecting their personal data is imperative to uphold privacy standards, a failure to obtain consent whilst allegedly collecting and storing biometric data, resulted in the previously cited lawsuit against Google in 2022 [17]. To address this concern, the HR manager should inform individuals about how their data will be used, who will have access to it, and their rights regarding their data. Implementing user-friendly mechanisms for individuals to give and withdraw consent ensures compliance with privacy standards. Developing comprehensive privacy policies that are easily accessible and understandable by all users is also essential.

Implementing data anonymisation techniques can significantly reduce the risk of exposure if the data is compromised [18]. Regular compliance audits will help ensure that the organisation adheres to GDPR guidelines. Establishing clear data policies will provide a framework for how personal data is managed, ensuring consistency and legal compliance. Encrypting personal data ensures that even if data is intercepted, it cannot be read without the decryption key [19]. Access controls limit data access to authorised personnel only, reducing the risk of internal breaches [20]. Regular security audits and staff training will help maintain a high level of data security awareness and readiness to address potential threats. Transparent information practices will help build trust with employees and applicants by clearly communicating how their data will be used. User-friendly consent mechanisms will ensure that individuals feel in control of their personal information, and detailed privacy policies will provide the necessary information to make informed decisions.

## Appendices

**Appendix A**

**A screenshot of a computer

Description automatically generated**

Figure 1 - SMO was the most predictive on the full training set

**A screenshot of a computer

Description automatically generated**

Figure 2 j48 was the most predictive on the full training set

**A white screen with text

Description automatically generated**

Figure 3 - j48 was the most predictive on the full training set

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

Figure 4 (left above) HR department, figure 5 (right above Researcb and Development), figure 6 (left below) Sales using CfsSubsetEval

A screenshot of a computer

Description automatically generated

Figure 7 (right) HR department, figure 8 (left below) Research & Development, figure 9 (right below) Sales using InfoGainRatio

A screenshot of a computer

Description automatically generated

**1b**

Figure 10: Mean Averages on the discrete features, comparing Attrition = No dataset and Attrition = Yes dataset

Figure 11: percentage distribution for categorical features comparing Attrition = No dataset and Attrition = Yes dataset

A screenshot of a document

Description automatically generatedA screenshot of a white sheet

Description automatically generatedA screenshot of a phone

Description automatically generated

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

Figure 12 and 13 Finding the mean average of age for attrition = no (above) and attrition = yes (right). The older employees were more likely to stay with the mean for attrition = no being approximately 38 and for leaving was approximately 34.

A screenshot of a computer

Description automatically generated**1c**A screenshot of a computer

Description automatically generated

Figure 15 PART rule list for job level 3, 4 and 5. Highly supportive of the j48 tree

Figure 14 J48 tree when job Level 3, 4 and 5 were selected as the class to see which attributes could offer staff retention

Figure 16, 17, 18, 19 from top left to right below highlighting the correlation analysis for Age, MOB Level, Monthly Income and Hours Worked Per day

A screenshot of a computer screen

Description automatically generatedA screen shot of a computer

Description automatically generated

A white screen with black text

Description automatically generatedA screenshot of a computer screen

Description automatically generated

**Appendix B**A screenshot of a computer

Description automatically generated**.1**

Figure 20 UML Diagram for Employee Management System

This UML diagram illustrates the relational database design for an Employee Management System. It comprises several entities with their respective attributes and relationships with Employee as the central entity. Each entity is normalised to ensure data integrity and minimise redundancy, with foreign keys establishing relationships between entities to support comprehensive data management and analysis.

A black screen with white text

Description automatically generatedA screenshot of a computer program

Description automatically generatedA screen shot of a computer

Description automatically generated**Appendix B.2: SQL Statements**

Figure 23 SQL query for extracting the description of a job

Figure 21 SQL queries for inserting a new line of data

Figure 22 SQL queries for extracting the average monthly income for each job role.

# References

|  |  |
| --- | --- |
| [1] | A. A. Dina Elreedy, “A Comprehensive Analysis of Synthetic Minority Oversampling Technique (SMOTE) for handling class imbalance,” *Information Sciences,* vol. 505, pp. 32-64, 2019. |
| [2] | S. B. A. C. Z. Mundargi, “PrePy - A Customize Library for Data Preprocessing in Python,” *International Conference for Advancement in Technology (ICONAT),* pp. 1-5, 2023. |
| [3] | A. K. J. A. S. Dheeraj Pal, “Comparing Various Classifier Techniques for Efficient Mining of Data,” Singapore, 2016. |
| [4] | S. S. W. S. Zulfiqar Ali, “Performance analysis of support vector machine based classifiers,” *International Journal of ADVANCED AND APPLIED SCIENCES,* 2018. |
| [5] | Z. Hang, “Algorithm for Attributes Reduction Based on Information Gain,” *Journal of Chongqing Institute of Technology,* 2007. |
| [6] | S. C.-L. M. G.-R. D. Trillo-Montero, “Design and Development of a Relational Database Management System (RDBMS) with Open Source Tools for the Processing of Data Monitored in a Set of Photovoltaic (PV) Plants,” *Applied Sciences,* 2023. |
| [7] | J. Nadon, “Database Services in AWS,” in *Website Hosting and Migration with Amazon Web Services*, 2007, pp. 127-151. |
| [8] | R. M. Lerner, “Open-source databases, Part II: PostgreSQL,” *Linux Journal,* vol. 16, 2007. |
| [9] | F. L. Peng Lu, “The Design of Postgresql’s Security Protection Scheme,” *IEEE 7th Information Technology and Mechatronics Engineering Conference (ITOEC),* pp. 2023-2030, 2023. |
| [10] | A. G. A. Z. M. Malawski, “Serverless execution of scientific workflows: Experiments with HyperFlow, AWS Lambda and Google Cloud Functions,” *Future Generation Computer Systems,* vol. 110, pp. 502-514, 2020. |
| [11] | P. A. Sudipto Chakraborty, “Let Us Create A Lambda Function for Our IoT Device In The AWS Cloud Using C#,” *International Journal of Management, Technology, and Social Sciences,* vol. 8, no. 2, 2023. |
| [12] | C. Liu, “Application of Hadoop-Based Cloud Computing in Teaching Platform Research,” *Journal of Interconnection Networks,* vol. 22, 2022. |
| [13] | S. R. V. P. H. Sanjana N, “Real-time Event Streaming for Financial Enterprise System with Kafka,” in *3rd Asian Conference on Innovation in Technology (ASIANCON)*, Ravet, 2023. |
| [14] | C. Brook, “Google Fined $57M by Data Protection Watchdog Over GDPR Violations , DATA INSIDER, Digital Guardian's Blog,” 28 December 2022. [Online]. Available: https://www.digitalguardian.com/blog/google-fined-57m-data-protection-watchdog-over-gdpr-violations. [Accessed 10 July 2024]. |
| [15] | “CFPB, FTC and States Announce Settlement with Equifax Over 2017 Data Breach,” The Consumer Financial Protection Bureau , 22 July 2019. [Online]. Available: https://www.consumerfinance.gov/about-us/newsroom/cfpb-ftc-states-announce-settlement-with-equifax-over-2017-data-breach/. [Accessed 11 July 2024]. |
| [16] | R. Böhme, “Security Audits Revisited,” in *Keromytis, A.D. (eds) Financial Cryptography and Data Security*, Berlin, Springer, 2012, pp. 129-147. |
| [17] | P. L. A. /. B. N. a. D. V. L. Offices, “Law Firm in Metro Manila, Philippines, Corporate, Family, IP law, and Litigation Lawyers > Philippine Legal Advice Consent in the Collection of Data: What Type of Consent Does the Data Privacy Law Require?,” [Online]. Available: https://ndvlaw.com/consent-in-the-collection-of-data-what-type-of-consent-does-the-law-require/. |
| [18] | F. P. C. E. F. Kohlmayer, “A flexible approach to distributed data anonymization,” *Journal of biomedical informatics,* vol. 50, pp. 62-76, 2014. |
| [19] | V. K. Sudeva, “Symmetric Key Encryption With Many Secret Keys,” in *Proceedings of the Second International Conference on Research in*, Pathamuttom, Kottayam, Kerala, 2017. |
| [20] | P. S. R. Sandhu, “Access control: principles and practice,” *IEEE Communications Magazine,* vol. 32, no. 9, pp. 40-48, 1994. |