



SCHOOL OF COMPUTER SCIENCE

ASSIGNMENT 4 (Weightage 30%) AUGUST 2023 SEMESTER



MODULE NAME	: DATA MINING
MODULE CODE	: ITS61504
DUE DATE	: 20.09.2023 – 20 November 2023
PLATFORM	: <u>MyTIMES</u>

This paper consists of NINE (9) pages, inclusive of this page.

Group No 5

Project Title: Employee Attrition Factors

STUDENT DECLARATION

- 1. I confirm that I am aware of the University's Regulation Governing Cheating in a University Test and Assignment and of the guidance issued by the School of Computing and IT concerning plagiarism and proper academic practice, and that the assessed work now submitted is in accordance with this regulation and guidance.*
- 2. I understand that, unless already agreed with the School of Computing and IT, assessed work may not be submitted that has previously been submitted, either in whole or in part, at this or any other institution.*
- 3. I recognise that should evidence emerge that my work fails to comply with either of the above declarations, then I may be liable to proceedings under Regulation*







No	Student Name	Student ID	Work Breakdown	Signature	Score
1	Felicia Dossou	0364975	Coding part Chapter 2,3,4,5		
2	Linda Lischetti	0364980	Project background Problem Statement & Objectives Conclusion		
3	Andrea Monoli	0364981	Results, interpretation Chapter 3,4,5		
4	Ng Wua Jzim	0352406	Results, interpretation Chapter 3,4,5		
5	Zaireen Adeena	0348269	Introduction Chapter 1		
6	Dylan Ho Wei Shearn	0350074	Results, interpretation Chapter 3,4,5		

Table of contents

Table of contents	3
Abstract	4
Chapter 1 – Introduction	5
I. Project background	5
II. Problem Statement & Objectives	5
III. Literature Review	6
Chapter 2 – Overview of the datasets	7
I. Dataset issues	7
II. EDA and descriptive statistics	7
A. Data exploration	7
B. Distribution visualisations	8
Chapter 3 – Pre-processing Techniques	10
I. Issue in the dataset	10
II. Steps in preprocessing	10
A. Data reduction	10
B. Data transformation	10
C. Data cleaning	11
Missing values	11
Outliers	11
III. Correlation analysis	12
Chapter 4 – Data Mining Techniques	13
I. Prediction models and results	13
A. Logistic regression	13
B. Decision tree regression	15
C. Association Rule Mining	16
Chapter 5 – Performance evaluation	18
Logistic regression	18
Decision tree regression	18
Association rule mining	19
Conclusion of the results	20
Appendixes	22
References	22
Similarity Report	22

Abstract

Employee attrition has been an important concern for businesses since it decreases the number of employees in a company. It is mainly caused by resignations and retirements and as a result, companies have to go through several detrimental effects in terms of morale and profitability.

The main aim of this study is to analyse employee attrition and its causes by using Data Mining techniques. We implemented Data Mining techniques like Association Rule Mining, Decision Tree Regression and Logistic Regression. Multiple factors of employee attrition such as age, monthly income, job satisfaction and more are taken into consideration and analysed to provide us with the results.

Chapter 1 – Introduction

I. Project background

Employee attrition, which refers to the gradual reduction in employee number within a company whether it be from resignations or retirements, has always been an important concern for businesses. Recently, there has been a massive uptick in employee turnover rates across the business landscape, with a reported average employee drop off rate in Malaysia reaching as high as 15% as of 2022[1]. This means that, on average, companies lose 1 out of 7 employees every year. High attrition rates can cause several detrimental effects to a company, such as greatly hindering a company's productivity and morale, as well as profitability.

In response to this dilemma, the use of Data Mining, a powerful analytical technique, can be harnessed to gain valuable insights into the factors contributing to employee attrition.

II. Problem Statement & Objectives

Which factors lead to employee attrition?

This project aims to employ data mining methodologies towards a dataset to identify the various factors that may contribute to employee attrition within an organisation, ultimately helping businesses reduce the rate of employee turnover and enhance workforce stability. In essence, the problem statement of our project is to make use of data mining techniques to determine the factors that contribute to employee attrition. In accordance with our goal, we have chosen to make use of the Employee Attrition dataset[2] to apply a multitude of data mining techniques, with the intent to transfer successful techniques to real world examples.

In the current competitive job market, understanding the underlying causes of employee attrition is paramount for organisations, if they are to retain their talent pool. Data mining provides an opportunity to analyse large datasets encompassing employee information, such as demographics, roles, performance, and workplace satisfaction. By making use of various data mining techniques such as logistic regression, decision tree mode and association mining rule it is possible to discover new correlations and dependencies within this data, providing greater insights on the main causes of attrition.

III. Literature Review

Employee Attrition Statistics [3]

According to the Bureau of Labor Statistics, it is stated that in the year 2021, the attrition rate is 57.3%. After further analysis and observation of the data, it can be seen that the employee attrition rate tends to be higher when the age is lower. This is mainly because younger employees are more likely to switch their jobs to gain more experience compared to sticking with one organisation or company. Furthermore, the age factor also relates to the monthly income that is provided by the company to their employees. Younger employees are more likely to earn a lower income compared to a senior employee. To further explain, the employee attrition rate is seen to be lower when the employees are older as they gain a more stable monthly income.

Study on the Most Determining Factor of Employee Attrition [4]

One of the most common forms of employee attrition is voluntary attrition where an employee leaves the company on their own will and decision. This is more likely done to seize better opportunities that will benefit the employee's career. It can be shown that age is one of the factor or attributes that influences employee attrition. According to the article, new employees who are younger usually lose their work-life balance which eventually will lead to them leaving their company. Compared to a senior employee, they have much more experience in a working environment and have a stable work-life balance.

Data Mining Technique: Logistic Regression [5]

To approach our issue and meet our aim, there are a variety of Data Mining techniques that can be implemented. To further explain this, in order for us to analyse the relationship between input features like age or work-life balance and a binary outcome, data mining techniques like Logistic Regression can be used since it helps to predict the approximate probability of an employee staying or leaving based on the independent variable. By calculating values like estimate, z values and standard error, we would be able to meet our aim and identify a significant factor that leads to employee attrition.

Data Mining Technique: Decision Tree Regression [6]

Decision Tree Regression will also be suitable for this study since it helps to easily identify relationships between variables and it allows us to understand and visualise the results better. By implementing Decision Tree Regression, the data can be split into smaller subsets that only contain one class and enable us to recognize the likelihood that an employee will leave the company. We would also implement Association Rule Mining or ARM to better analyse the cause-and-effect linkages between the variables. By using the ARM algorithm, it helps us to identify patterns and recognise which elements commonly occur together which will suggest possible correlations.

Chapter 2 – Overview of the datasets

I. Dataset issues

Our dataset is a fictional dataset on HR employee attrition and performance, obtained from the website Kaggle.com, which is a website made for the purpose of sharing and distributing machine learning data, techniques and technologies, among other things. The dataset is a fictional dataset designed by an HR analyst to simulate employee attrition and performance.

Age	Attrition	BusinessTravel	DailyRate	Department	DistanceFromHome	Education	EducationField	EmployeeCount	EmployeeNumber	EnvironmentSatisfaction	Gender
41	Yes	Travel_Rarely	1102	Sales	1	2	Life Sciences	1	1	2	Female
49	No	Travel_Frequently	279	Research & Development	8	1	Life Sciences	1	2	3	Male
37	Yes	Travel_Rarely	1373	Research & Development	2	2	Other	1	4	4	Male
33	No	Travel_Frequently	1392	Research & Development	3	4	Life Sciences	1	5	4	Female
27	No	Travel_Rarely	591	Research & Development	2	1	Medical	1	7	1	Male
32	No	Travel_Frequently	1005	Research & Development	2	2	Life Sciences	1	8	4	Male
59	No	Travel_Rarely	1324	Research & Development	3	3	Medical	1	10	3	Female
30	No	Travel_Rarely	1358	Research & Development	24	1	Life Sciences	1	11	4	Male
38	No	Travel_Frequently	216	Research & Development	23	3	Life Sciences	1	12	4	Male
36	No	Travel_Rarely	1299	Research & Development	27	3	Medical	1	13	3	Male
35	No	Travel_Rarely	809	Research & Development	16	3	Medical	1	14	1	Male
29	No	Travel_Rarely	153	Research & Development	15	2	Life Sciences	1	15	4	Female
31	No	Travel_Rarely	670	Research & Development	26	1	Life Sciences	1	16	1	Male
34	No	Travel_Rarely	1346	Research & Development	19	2	Medical	1	18	2	Male
28	Yes	Travel_Rarely	103	Research & Development	24	3	Life Sciences	1	19	3	Male

Sample of the Dataset

Figure 1.0

II. EDA and descriptive statistics

A. Data exploration

By using some basic functions such as `dim(df)` or `colnames(df)`, we find the dimensions of our dataframe: 1470 observations and 35 attributes. Moreover, the use of the `sum(is.na(df))` function gives us the confirmation of no missing value (NA)

In order to have an overall statistical view of our dataset, we make use of the `summary(df)` function to display information about the dataset, such as the minimum, maximum, mean, median and quartiles of our numerical variables columns.

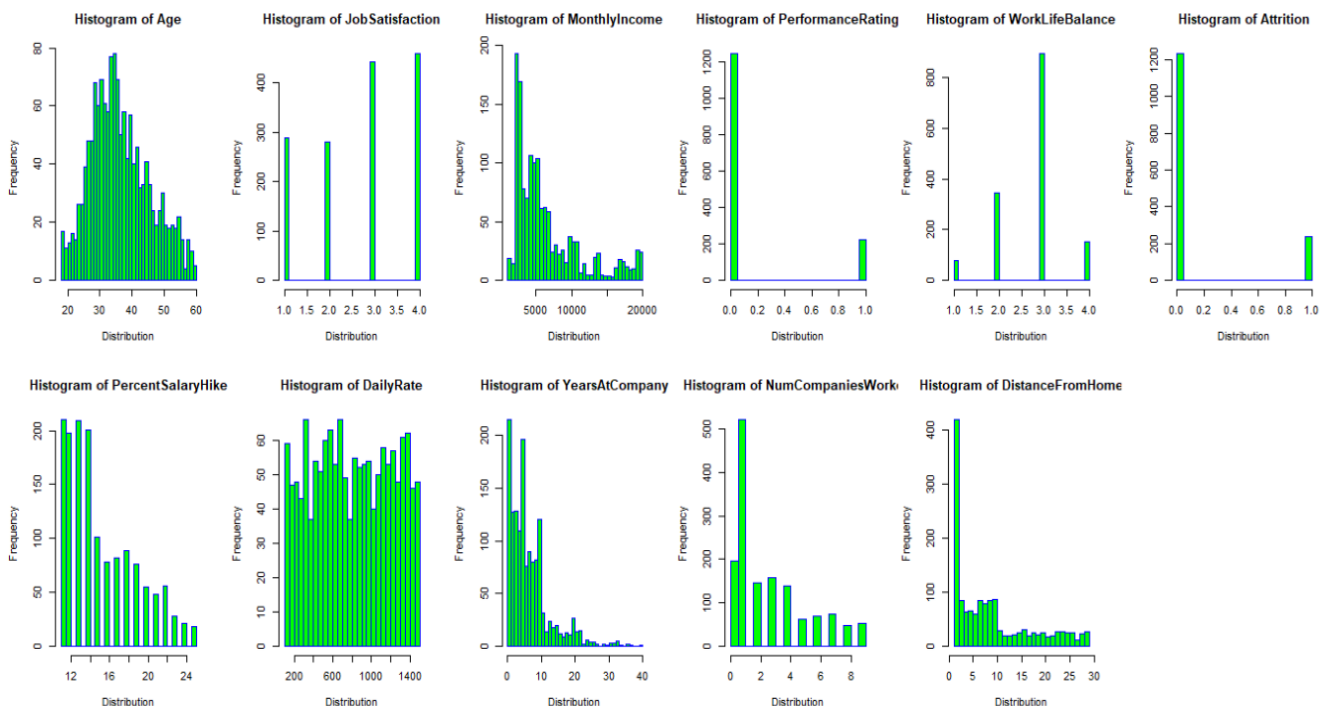
Age	JobSatisfaction	MonthlyIncome	PerformanceRating	WorkLifeBalance	Attrition
Min. :18.00	Min. :1.000	Min. : 1009	Min. :0.0000	Min. :1.000	Min. :0.0000
1st Qu.:30.00	1st Qu.:2.000	1st Qu.: 2911	1st Qu.:0.0000	1st Qu.:2.000	1st Qu.:0.0000
Median :36.00	Median :3.000	Median : 4919	Median :0.0000	Median :3.000	Median :0.0000
Mean :36.92	Mean :2.729	Mean : 6503	Mean :0.1537	Mean :2.761	Mean :0.1612
3rd Qu.:43.00	3rd Qu.:4.000	3rd Qu.: 8379	3rd Qu.:0.0000	3rd Qu.:3.000	3rd Qu.:0.0000
Max. :60.00	Max. :4.000	Max. :19999	Max. :1.0000	Max. :4.000	Max. :1.0000
PercentSalaryHike	DailyRate	YearsAtCompany	NumCompaniesWorked	DistanceFromHome	
Min. :11.00	Min. : 102.0	Min. : 0.000	Min. :0.000	Min. : 1.000	
1st Qu.:12.00	1st Qu.: 465.0	1st Qu.: 3.000	1st Qu.:1.000	1st Qu.: 2.000	
Median :14.00	Median : 802.0	Median : 5.000	Median :2.000	Median : 7.000	
Mean :15.21	Mean : 802.5	Mean : 7.008	Mean :2.693	Mean : 9.193	
3rd Qu.:18.00	3rd Qu.:1157.0	3rd Qu.: 9.000	3rd Qu.:4.000	3rd Qu.:14.000	
Max. :25.00	Max. :1499.0	Max. :40.000	Max. :9.000	Max. :29.000	

Summary of the dataset

Figure 1.1

B. Distribution visualisations

In order to have a better understanding on how the data is distributed, we created histograms with the function `hist()` for each of our attributes, that we placed into 6x2 subsets.



The histograms of our dataset

Figure 1.2

When analysing each of the histograms in Figure 1.2, there are noticeable trends to talk about.

Continuous Variables:

- The age spans from 20 to 80 years. The peak frequency is around 35 to 36 years old.
- Monthly Income: Most employees earn more than 2000. The frequency declines when the income exceeds 7000, the average income is around 4900.

Discrete Variables:

- Job Satisfaction: The histogram shows at a high frequency around 3.0 and 4.0, indicating a high job satisfaction rate. Satisfaction levels at 1.0 and 2.0 have a lower frequency.
- Percent Salary Hike: Most employees receive a salary hike between 10% to 14%. The frequency decreases significantly beyond the 14% salary hike.
- Performance Rating: The histogram suggests that a rating of 3.0 is the most common, surpassing the frequency of the rating 4.0.
- Work-Life Balance: The histogram suggests that a rating of 3.0 is the most common, followed by 2.0 and 4.0, the least frequent rating is 1.0.

Chapter 3 – Pre-processing Techniques

This second step is performed to clean, transform and organise the data for modelling.

I. Issue in the dataset

During the processing of the dataset, a number of issues arose, 3 of which are most noteworthy. Firstly, there were too many attributes within the datasets. Because of this, a number of issues may arise, such as overfitting, data sparsity and an increased risk of noise, among other challenges. Aside from this, another issue that occurred during pre-processing was the abundance of data points that were qualitative rather than quantitative. While this in itself is not a problem, to ease the task of pre-processing, the use of integers for data is preferred. To aid in this, the columns which consist of qualitative data are reduced to integers, so that they can be manipulated and visualised easier. Lastly, the third issue that was found during preprocessing is the number of binary character data points within the dataset. To fix this, the use of binary transformation and re-coding was implemented, which converts these data points into integers, for ease of manipulation.

II. Steps in preprocessing

A. Data reduction

According to our study, we don't need that many columns. This is why we will filter some of them to reduce the dimensionality. We will be running a code which will be retaining some specific columns of interest which in this case will help us to process the data with ease. Additionally, this can help improve the performance of the data mining algorithms as the data mining algorithms can perform well when they are trained on smaller datasets.

B. Data transformation

To make manipulations easier and to allow us to visualise character columns, we first convert char columns into integers. It's a simplified version of hot encoding.

Moreover, to detect outliers, we adjust some variables.

The attrition value is set to 0 if the employee has not left, else it would be set 1 if the employee has left the company. In any other case, the value will be set to 2.

The same way for the Employee performance attribute, when the value is 3, the variable is set to 0. If the employee rating is 4, the variable is set to 1. Otherwise, the variable is set to 2.

C. Data cleaning

Data cleaning is the process of identifying and correcting errors to improve the quality of the data, making it suitable for analysis and ensuring that the results obtained from the data are reliable and meaningful. This is done by the handling of missing values, outliers and data types.

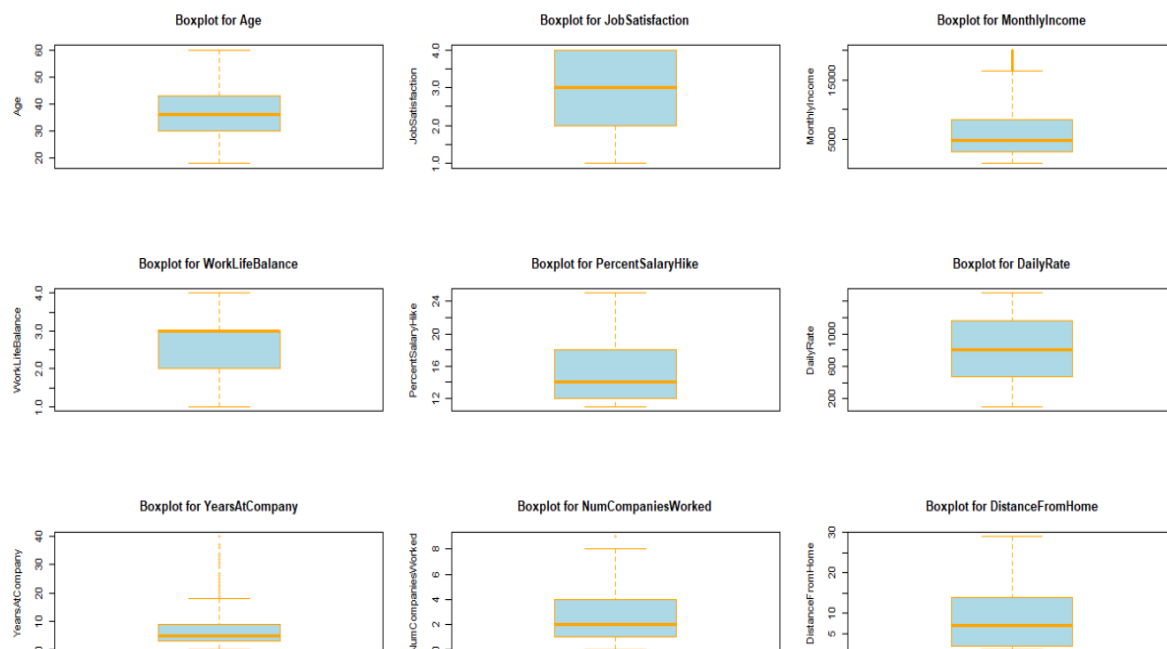
Missing values

As is evident within the EDA, there are no missing values (NA) in our dataset!

Outliers

As we set the outliers value of Attrition and Performance rating attributes earlier, rows that have the value of 2 will be removed from the dataset.

Afterwards, a box plot grid with a 2x3 layout will be created to visualise the distribution of each variable, this excludes attrition and performance rating which are binary variables.



Boxplot graph

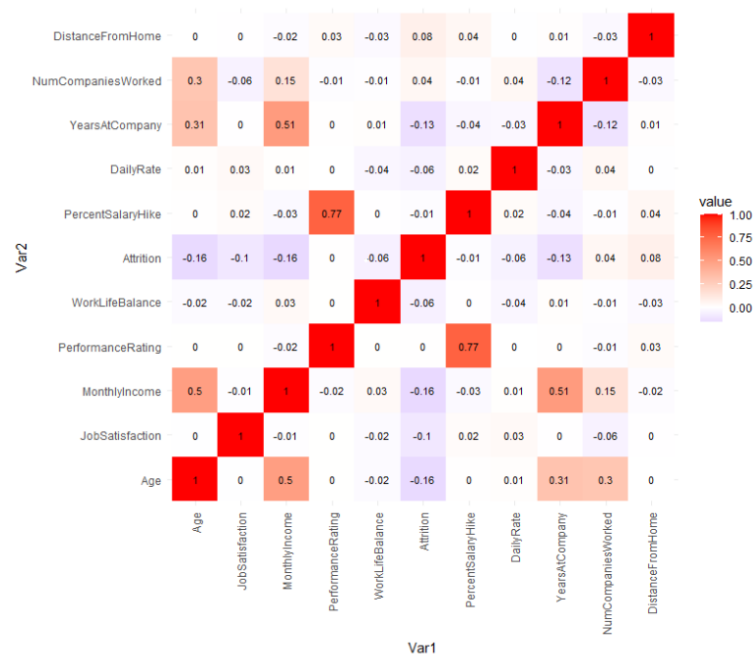
Figure 1.3

In Figure 1.3, the box plot illustrates the distribution of the data we have selected and analysed for outliers. Looking at the boxplots, some of the graphs show that the distribution is rather uniform and consistent. However, some attributes have outliers but they each can be explained: MonthlyIncome has outliers due to the possibility that some employers have relatively high or low income compared to the majority, YearsAtCompany have outliers due to some employers having spent an unusually long or short time in the company.

We decide not to remove these outliers as they can be relevant to our study of Attrition.

III. Correlation analysis

We will be using a correlation matrix to display the variables patterns. A heatmap will be implemented to further enhance the visualisation of the data, with deeper colour representing a stronger present, while a lighter colour represents a weaker presence.



Correlation Matrix with a heatmap

Figure 1.4

Considering that most of the coefficients are pretty low, we decide to remove attributes that have a negligible correlation to Attrition : we take out the columns ‘PerformanceRating’ and ‘PercentSalaryHike’.

Chapter 4 – Data Mining Techniques

I. Prediction models and results

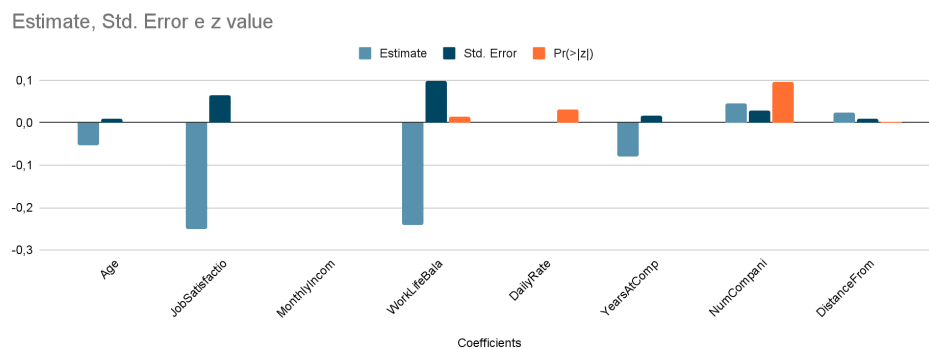
A. Logistic regression

Logistic regression is a statistical method that is designed for binary classification, meaning that it is the optimal choice of prediction models when the dataset in question has an outcome variable that is categorical and is limited to only 2 classes. As such, the use of logistic regression is ideal for exploring correlations between employee Attrition and various variables, aligning with our aim of identifying significant predictors.

To start and have a better understanding of the relationship between each numerical attribute and Attrition, we print out the model summary. Here a summary table of the outcome:

Coefficients	Estimate	Std. Error	z value	Pr(> z)
Age	-0.05225	0.00870	-6.006	1.91e-09
JobSatisfaction	-0.2510	0.0637	-3.940	8.16e-05
MonthlyIncome	-1.271e-04	2.162e-05	-5.879	4.12e-09
WorkLifeBalance	-0.23956	0.09796	-2.445	0.014467
DailyRate	-0.0003834	0.0001769	-2.168	0.0302
YearsAtCompany	-0.08076	0.01594	-5.068	4.03e-07
NumCompaniesWorked	0.04565	0.02742	1.665	0.096
DistanceFromHome	0.024710	0.008312	2.973	0.00295

Additionally and to better visualise the variability of data, we use a bar chart.



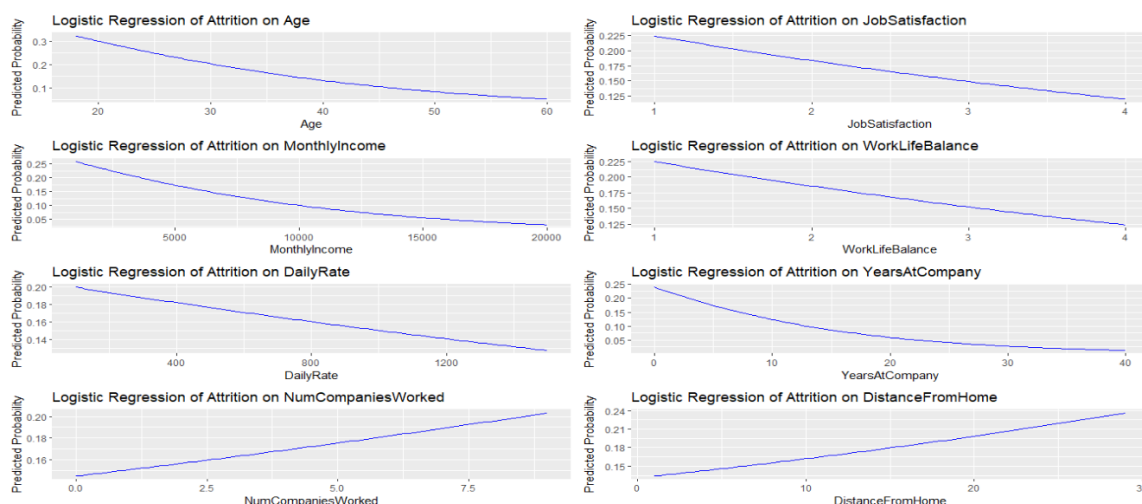
*For scalability purposes the Z_value attribute has not been mentioned in the graph

This analysis reveals two major trends: firstly, factors such as Age, JobSatisfaction, MonthlyIncome, Work-lifeBalance, DailyRate, and YearsAtCompany are all significantly associated with a decrease in the dependent variable, indicating their strong influence. Secondly, NumCompaniesWorked and DistanceFromHome show a tendency to increase the dependent variable, albeit with varying degrees of significance, suggesting these factors also play a notable role in the outcomes.

Predictions

Now, we generate predictions building the same logistic regression model but instead for new data points created within the range of each attribute. For this purpose, we create **x_values** spanning from the minimum to the maximum value.

We create a `plot()` for each of my attributes to have a better understanding of evolution of our predictions :

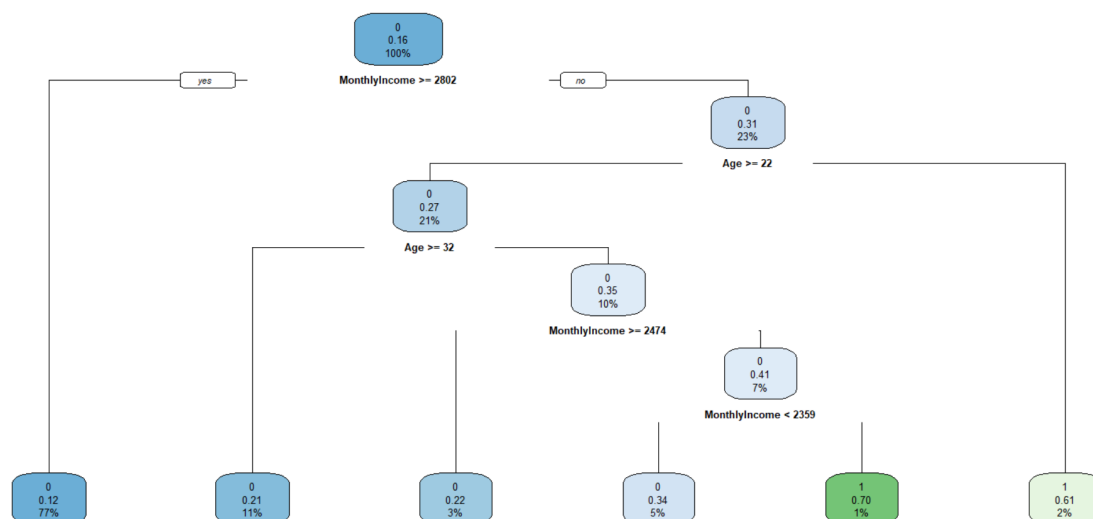


Based on the observed results, some of the attributes might be significant predictors of employee attrition such as NumCompaniesWorked and DistanceFromHome.

B. Decision tree regression

The decision tree model for predicting job attrition is a well structured model that will divide feature values into subsets to create a tree-like structure that predicts the target variable's value based on the paths through these subsets. It is notable due to its interpretability, ability to capture non-linear relationships.

In this section we want to calculate the probability of an employee staying in the company using the following features : Age and Salary. We first start printing out a summary tree of our actual dataset using `rpart()` function:



Predictions

Using `predict()` and after converting the prediction results into probabilities, we obtain the following output :

	Age <int>	MonthlyIncome <int>	Attrition <factor>	Probability_Stay <dbl>
1	41	5993	1	0.7310586
3	37	2090	1	0.7310586
15	28	2028	1	0.7310586
22	36	3407	1	0.7310586
25	34	2960	1	0.7310586
27	32	3919	1	0.7310586

6 rows

The consistent Probability_Stay values of approximately 0.7311 across all instances suggest that the model is assigning a similar probability of employees staying in their current roles, with some minor variations in Age and MonthlyIncome.

While the age of the employees have a vast range from 28 to 41, and monthly income of the chosen employees are of greatly varying figures, the probability of the employees suffering attrition stays consistent, with a 73.1% across all the board. This implies that age and monthly income does not have a strong effect on the probability of employees staying within the company. As such, all of the predictions made have the employees being lost to attrition, regardless of the 2 variables.

C. Association Rule Mining

In the context of job attrition, ARM can be employed to generate association rules that provide insights into the cause-and-effect relationships between the variables. In our case, it will be useful to analyse the reasons behind employee attrition, in fact the algorithm works by examining the items that frequently appear together in the transaction data, indicating potential correlations.

For our analysis we setted support = 0.05 and confidence = 0.8 this choice is driven by the need to strike a balance between finding meaningful patterns and reducing the risk of false positives using `apriori()` to generate the following rule:

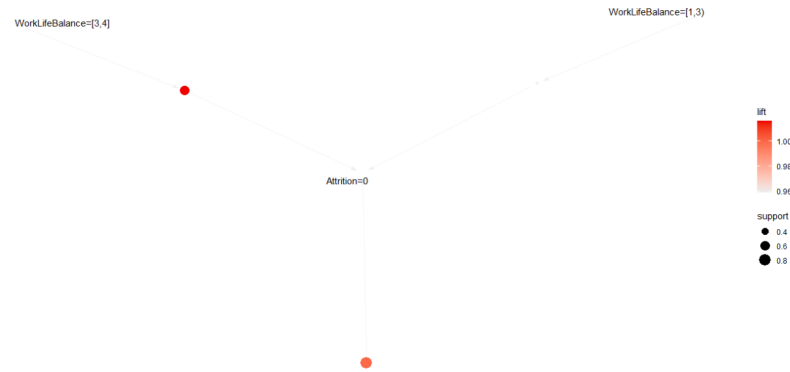
	confidence <dbl>	minval <dbl>	smax <dbl>	arem <chr>	aval <lg>	originalSupport <lg>	maxtime <dbl>	support <dbl>	minlen <int>	maxlen <int>	target <chr>	ext <lg>
	0.8	0.1	1	none	FALSE	TRUE	5	0.05	1	10	rules	TRUE

1 row

	lhs <chr>	rhs <chr>	support <dbl>	confidence <dbl>	coverage <dbl>	lift <dbl>	count <int>
[1]	0	=> (Attrition=0)	0.8387755	0.8387755	1	1	1233

1 row

As usual, for a better understanding, we use a graph to displays the relationships between the rules :



Predictions

Based on the results, the apriori algorithm found 3 rules with high confidence. The support values for these rules are also high, ranging from 0.2319728 to 0.6068027. For the first rule, the lift value is 1.0, indicating independence between job satisfaction and attrition. This means that job satisfaction, as considered in this rule, does not have a direct impact on job attrition. For the other two rules, the lift values are 0.9588326 and 1.0166874, suggesting a negative and positive association, respectively, with job attrition when the rules are satisfied (lift value <1 = negative association , lift value >1 = positive association , lift value =1 =no direct impact “independence”).

We can observe that the first rule has the highest support and confidence, suggesting that job satisfaction has a strong relationship with Attrition. The second rule has the lowest support but also the highest confidence, giving more precise indications that job satisfaction also has a significant relationship with attrition even when other factors, such as work-life balance, are not considered. The third rule indicates that both job satisfaction and work-life balance have a strong relationship with attrition.

Overall, the results of this analysis suggest that job satisfaction and work-life balance play a significant role in attrition, with job satisfaction being the most influential factor.

Chapter 5 – Performance evaluation

Logistic regression

After splitting the data into **training set** and **testing set**, we make prediction on both and print the `table()` of both predicted and actual classes :

```
Distribution of Predicted Classes:      Distribution of Actual Classes:
  0   1                                0   1
293   1                             246  48
```

We observe that the distribution of predicted classes closely matches the distribution of actual classes, which is a first good indicator of model accuracy.

Moreover once we check the predictions, we use `confusionMatrix()` to evaluate the model performance.

	Reference	
Prediction	0	1
0	246	47
1	0	1

According to our data

- TP = 246 cases correctly predicted as Attrition
- TN = 1 case correctly predicted as non Attrition
- FP = 0 case wrongly predicted as Attrition
- FN = 47 cases wrongly predicted as non Attrition

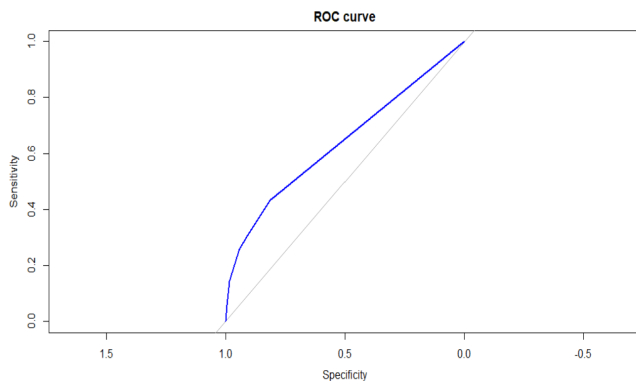
The formula of accuracy is $A = (TP+TN)/(TP+TN+FP+FN)$

$$A = (246+1)/(246+1+0+47) = 0.8401361$$

The accuracy shows 84% on average which is a fair score of prediction.

Decision tree regression

The model demonstrates an accuracy of approximately 84.9% and an AUC-ROC score of 0.6372. The accuracy indicates the model's overall effectiveness in predicting whether employees stay or leave, while the AUC-ROC score, a measure for binary classification tasks, reflects its capability to distinguish between employees who stay (class 0) and those who leave (class 1).



The model performs moderately well in differentiating between these two classes, meanwhile for AUC-ROC, a value of 0.5 would represent random chance, so our model is performing better than random. However, an ideal model would have an AUC-ROC close to 1.0, indicating perfect discrimination.

Association rule mining

There are a number of evaluation metrics that are used in order to properly determine the reliability and validity of any rules obtained in ARM. Support, confidence, lift, conviction and leverage, can be used to evaluate different aspects of the rules.

Support is used to determine the frequency of items in the rules within the dataset. This helps measure the significance of rules, based on the number of times they appear within the dataset, which allows the filtering out of rules that are not significant or important. The output of this metric for our dataset was found to be 0.8387755, 0.6068027 and 0.2319728. With these being able to be converted into percentages, with them being 83.9%, 60.7% and 23.2% respectively.

Confidence, on the other hand, is used to determine the rate of occurrence of the consequential part of the rule in transactions that contain the antecedents. This comparison allows the measurement of the conditional probability of rules. The output of this metric for our dataset was found to be 0.8387755, 0.8527725, and 0.8042453. With these being able to be converted into percentages, with them being 83.9%, 85.3% and 80.4% respectively.

Lastly, the lift metric is used to assess the strength of a rule by testing the results of the rules unbiasedly. This is done by assuming the two halves of the rule are instead independent of each other. From our project, the results of this metric were 1.0000000, 1.0166874, and 0.9588326. For the Lift metric, 1 implies that the antecedent of the rule does not have any effect on the consequent, while more than 1 being a positive influence from the antecedent to the consequent, and inversely, an output of less than 1 implies a negative influence from the antecedent on the consequent.

Conclusion of the results

In our HR attrition study using logistic regression, decision tree regression, and association rule mining, we gained key insights into employee turnover.

Logistic regression showed that older, more experienced employees are less likely to leave, highlighting age as a stabilising factor. Lower job satisfaction and monthly income were linked to increased attrition, while a good work-life balance was crucial in reducing it. Daily rates and longer tenures at the company also influenced attrition, showing the impact of both recent and long-term factors.

The decision tree model, focusing on age and monthly income, was effective in predicting employee retention, achieving 84.9% accuracy and an AUC-ROC score of 0.6372. This model could differentiate between employees likely to stay or leave. For instance, employees aged 37 earning 2090 showed a higher likelihood of leaving, similar to 25-year-olds earning 2960, both with only a 1% probability of staying. The model consistently estimated the likelihood of employees staying, with the average Probability_Stay around 0.7311, suggesting a need for further analysis for potential biases.

Association Rule Mining helped understand the factors behind employee turnover. The Apriori algorithm found strong links between job satisfaction, work-life balance, and attrition, with high confidence and support values. The most robust rule highlighted the strong connection between job satisfaction and attrition. These insights are valuable for organisations looking to improve their work environment and retain employees.

Implications and Suggestions:

- *Strategic Retention Initiatives*

Armed with insights into influential factors, organisations can strategically implement initiatives to fortify job satisfaction and work-life balance, yielding a significant reduction in attrition rates.

- *Tailored Interventions*

Acknowledging the diverse impact of various variables, interventions should be tailored to address specific concerns. Tailoring solutions for age-related considerations and focusing on work-life balance interventions for others can be particularly effective.

- *Continuous Monitoring*



Regular reassessment of employee satisfaction, income structures, and work-life balance is imperative. This iterative process ensures the adaptability of strategies to the evolving dynamics of the workforce.

Appendixes

References

1. Tan, J. (2022, December 12). *South-East Asia expected to see salary hike in 2023*. HRM Asia. <https://hrmasia.com/south-east-asia-expected-to-see-salary-hike-in-2023/>
2. *Employee attrition*. (2018, February 7). Kaggle. <https://www.kaggle.com/datasets/patelprashant/employee-attrition>
3. Bradshaw, R. (2023, October 21). *19 Employee Retention Statistics That Will Surprise you*. Apollo Technical LLC. <https://www.apollotechnical.com/employee-retention-statistics/#:~:text=In%20the%202021%20Bureau%20of,looking%20at%20only%20high%2Dperformers>
4. Khan, S. Y. (2019). Study on the most determining factor of employee attrition I.E. age factor. *IJERT*. <https://doi.org/10.17577/IJERTCONV7IS12015>
5. *What is Logistic regression?* | IBM. (n.d.). <https://www.ibm.com/topics/logistic-regression>
6. Xu, M., Watanachaturaporn, P., Varshney, P. K., & Arora, M. K. (2005). Decision tree regression for soft classification of remote sensing data. *Remote Sensing of Environment*, 97(3), 322–336. <https://doi.org/10.1016/j.rse.2005.05.008>

Similarity Report

	Submission Title	Turnitin Paper ID	Submitted	Similarity	Grade	
 View Digital Receipt	Group5_EmployeeAttritionFactors	2239302230	29/11/23, 14:02	7% 	–/100	Submit Paper 