University of Cape Town

EMPLOYEE ATTRITION MULTIVARIATE ANALYSIS

Sanana Mwanawina (MWNSAN002)

Department of Statistical Sciences

April 7, 2025

Introduction

Companies dedicate significant time and resources to their employees, and when employees leave for other organisations, replacing them incurs additional costs and effort. Accurately predicting employee turnover can help businesses mitigate these losses and optimise retention strategies.

Research Question

What are the key factors influencing employee turnover, and can we build a predictive model to identify employees at risk of leaving?

Hypotheses to Test

1. **Employees with low job satisfaction are more likely to leave.**
   A study by Rakhra (2018) provided strong evidence to support the idea that organisations aiming for high employee retention must focus on ensuring their staff are happy and satisfied.
2. **Salary hikes have a significant impact on retention.**
   There is strong evidence that suggests that organisations that succeed in the marketplace must maintain attractive salary packages to increase employee loyalty and higher retention rates (Iqbal et al., 2017).
3. **Salary is a stronger predictor of attrition than work-life balance for young employees or people early in their career.**
   Research shows that pay and benefits have greater influence on young professionals' retention compared to work-life balance (Chua, 2023).

Data Description

The IBM HR Analytics Employee Attrition and Performance dataset is a synthetic dataset created by the International Business Machines (IBM) Corporation to model employee attrition factors in a corporate setting. The dataset contains 30 variables (21 numeric and 9 categorical) related to employee demographics, job roles, and performance metrics.

Description of Observations

The dataset consists of 1,470 employee records, each represented by multiple attributes that capture various aspects of their employment experience. The variables include:

- **Demographic Information:** Age, gender, marital status, and whether the employee is over 18.
- **Job-Related Features:** Job role, department, job level, and business travel frequency.
- **Workplace Environment:** Job satisfaction, work-life balance, overtime status, and relationship satisfaction.
- **Performance Metrics:** Performance rating, job involvement, and training attendance.
- **Employment History:** Years at company, years in the current role, and distance from home.
- **Education and Qualifications:** Highest education level and field of study.
- **Attrition**: A categorical variable indicating whether an employee left the company (Yes/No). The dataset is imbalanced, with 83.9% labeled as "No" and 16.1% as "Yes". This imbalance is typical in real-world attrition datasets, as most employees in an organisation remain employed over a given period.

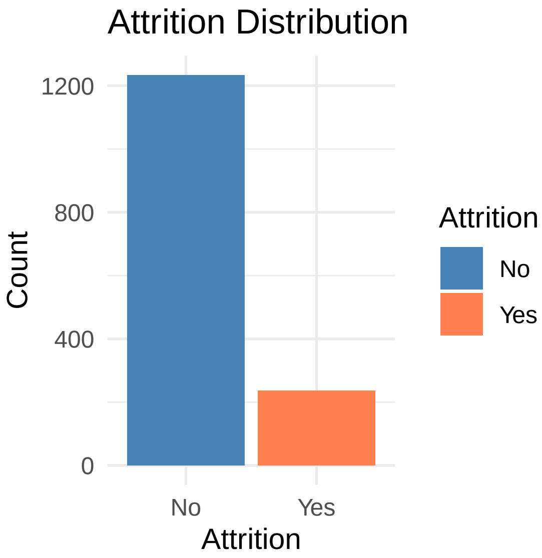


Figure 1: Distribution of target variable (Attrition) classes

More details about key features of the dataset can be found in the appendix.

Data Preprocessing

Before applying multivariate analysis techniques, the dataset was preprocessed as follows:

1. Removal of non-informative columns. Columns such as EmployeeCount, EmployeeNumber, StandardHours, and Over18 were removed as they did not provide useful predictive information.
2. Categorical variable encoding to convert their respective levels into numeric representations.
3. Continuous numeric variables were standardised to ensure that all features contributed equally.

Methodology

To analyse the research question, we employed multiple multivariate statistical methods. We aim to:

1. Use multidimensional scaling (MDS) to reduce the dimensions of the dataset.
2. Use clustering analysis to see if meaningful subgroups emerge, and visualise these in the lower dimensional space created using MDS.
3. Use Attrition as the target variable and train a support vectors machines (SVM) classifier and a random forest classifier to predict attrition.

Multidimensional Scaling (MDS)

To visualise the structure of the dataset, we applied MDS which reduced the dataset to a two-dimensional space. Given the presence of both numerical and categorical variables, we used **Gower's distance** to ensure a meaningful representation. We assessed the quality of the MDS projection using a **stress test**, which measures how well the lower-dimensional representation preserves the original distances. A low stress value indicates that the lower-dimensional mapping retains much of the original data's structure, making it suitable for visualisation.

Feature Importance

The precursor to clustering was an exploration of feature importance. Since the dataset contains many variables, identifying the most important ones can greatly aid **cluster profiling** by reducing complexity and highlighting key dimensions to focus on. To achieve this, a **feature importance** plot was generated by training a random forest classifier on the dataset. The classifier was optimised using a grid search over the following hyperparameter values:

- **mtry**: the number of randomly sampled features
- **ntree**: the number of trees

By analysing feature importance scores, we aimed to refine our cluster profiling by focusing on the most relevant attributes. An important caveat is that the random forest is a supervised algorithm, and may therefore capture a different structure from that revealed by the clustering algorithm. For this reason, the results from the feature importance plot are treated as a suggestive guide rather than a definitive one. The random forest classifier will be compared to the SVM classifier, which will be discussed later.

For **using the feature importance to determine the top features**, we followed the approach suggested by Prasetiyowati et al. (2021) which is to apply a relative threshold to the feature importance scores obtained from the Random Forest model. Features with importance values below this threshold can be excluded which could potentially improve model performance and efficiency. This threshold was determined using the mean of the scores, which is consistent with the authors' approach to identifying less informative attributes.

Clustering

We explored whether distinct subgroups exist within the dataset by applying different clustering techniques.

1. K-Means Clustering

To determine the optimal number of clusters, **K**, we used two methods. The **average silhouette score** approach measures how well each data point fits within its assigned cluster compared to other clusters. A higher average score indicates better-defined clusters. The **gap statistic** approach compares the within-cluster dispersion to that expected under a reference null distribution, allowing us to identify the K at which the observed clustering is substantially better than random. It is the smallest K where the gap value is within one standard error of the gap at K + 1. In the figure that follows, the optimal K for both approaches was 3 and it is indicated with a vertical dotted line:

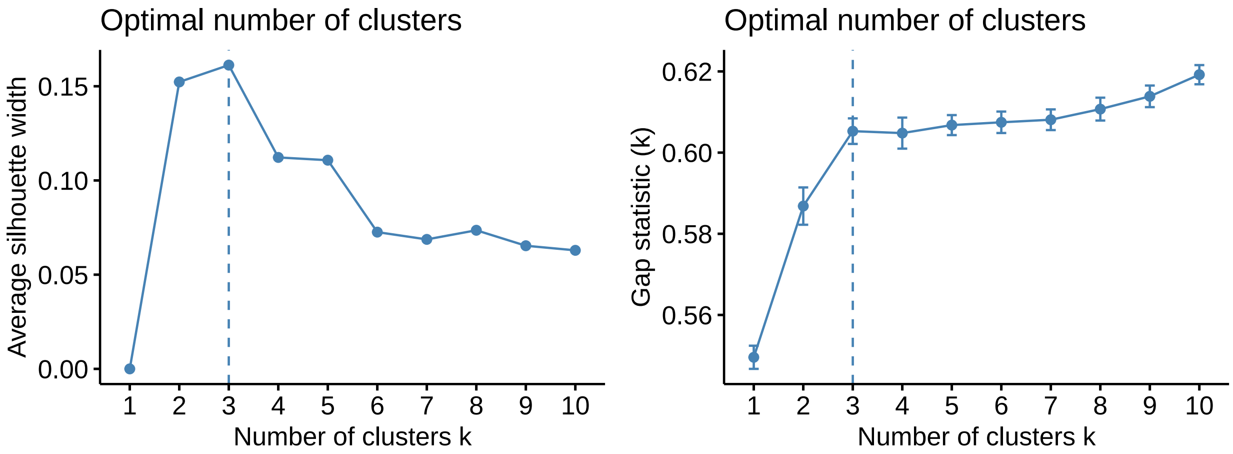


Figure 2: Results of average silhouette scores and gap statistics techniques for choosing the optimal number of clusters

2. Hierarchical Clustering

We used Ward's method to minimise variance within clusters, and a dendrogram was analysed to confirm the optimal number of clusters. A large vertical gap indicates that two relatively dissimilar clusters are being combined, and by cutting the dendrogram at a height just before such a large jump, we can identify a natural separation in the data. Based on this, cutting the tree at a height of 70 would be appropriate which results in 3 clusters, as shown in Figure 3 on the next page.

3. Cluster Profiling

We proceeded to obtain cluster centroids for the resulting profiles. These were calculated separately: the mean for continuous columns and the mode for categorical columns. To better understand these subgroups within the context of our attrition exploration, we also calculated the percentage of observations in each cluster that originally had a value of "Yes" (represented as 2) for the Attrition attribute.

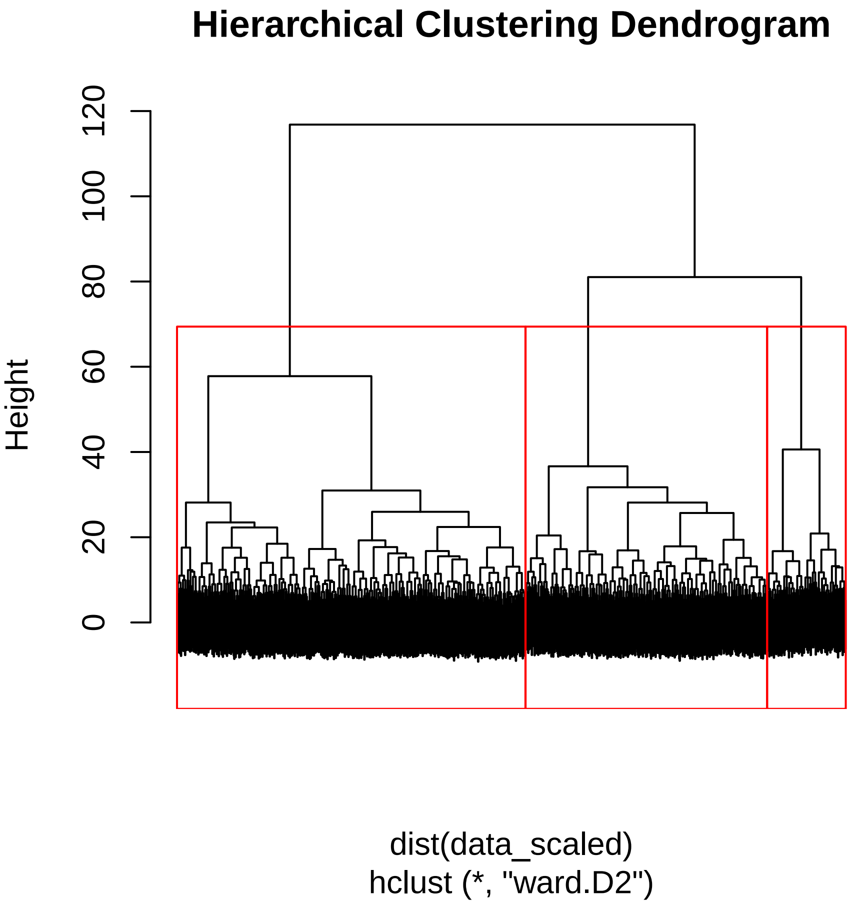


Figure 3: Dendrogram cut at height 70 to yield 3 clusters

Support Vector Machines

We aimed to predict the risk of Attrition and, to achieve this, we trained an SVM classifier. SVM was chosen over other classification models due to its ability to handle high-dimensional data and class imbalance effectively. Unlike logistic regression, which assumes linear relationships, SVM can capture complex patterns using kernel methods. While tree-based models like random forests provide feature interpretability, which was utilised in this analysis, SVM demonstrated strong predictive performance with proper hyperparameter tuning despite dataset imbalance, making it the optimal choice for this project (Abdullah & Abdulazeez, 2021).

As mentioned, the target variable is highly imbalanced. So, we considered different approaches to handle this issue. One approach is assigning weights to the target variable classes and fitting the model on the imbalanced dataset, while another involves using sampling techniques to balance the classes. Mohammed et al. (2020) found that oversampling performs better than undersampling for different classifiers and obtains higher scores in different evaluation metrics. So for this project, we chose to oversample the minority class (Attrition = "Yes"), resulting in the balanced distribution shown below:

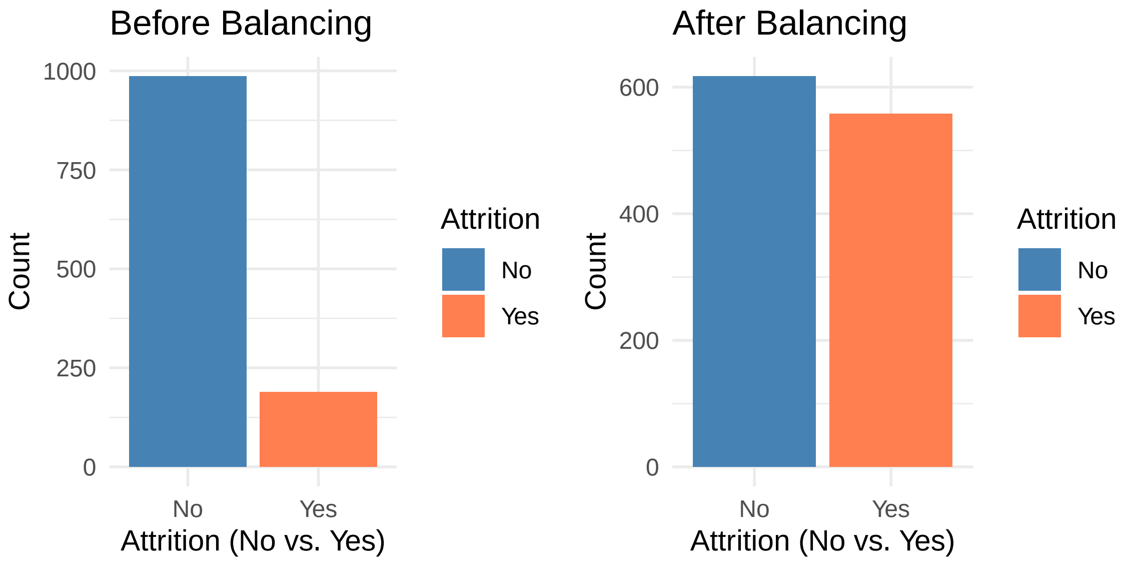


Figure 4: Comparison between imbalanced and over-sampled balanced distributions of the target variable.

We experimented with different approaches to improve model performance and handle class imbalance:

1. Training the model on the unbalanced dataset without adjusting class weights.
2. Training the model on a balanced dataset created by oversampling the minority class (Attrition = "Yes").
3. Training the model on the unbalanced dataset while assigning different class weights.
4. Training the model on the unbalanced dataset tuning hyperparameters (cost, gamma, and class weight) to optimise model performance.
5. Training the model on the unbalanced dataset using only the 12 most important features, as identified by the feature importance plot, and tuning cost, gamma, and class weight parameters.

Results

Multidimensional Scaling

The MDS projection provided a two-dimensional representation of the dataset.

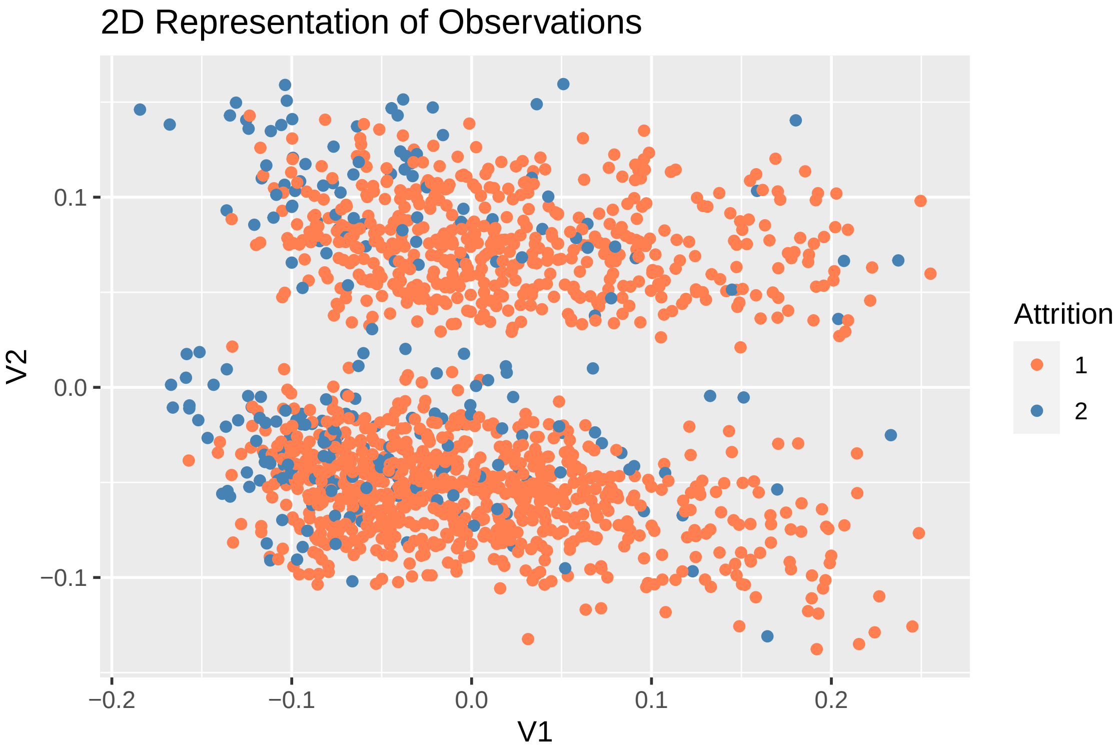


Figure 5: Two-dimensional MDS projection of the dataset, with observations colored by the target variable (attrition).

The computed stress value for the above projection was **0.0974** which indicates a reasonably good ft, and therefore suggests the original distance relationships were well preserved.

Feature Importance Results

For the feature importance plot, the optimised random forest classifier had the following parameter values: `mtry = 6` and `ntree = 300`. The feature importance plot is shown in figure 6 on the next page.

According the results, `MonthlyIncome`, `Age` and `DailyRate` are the 3 most important features. We expect to see great separation in the centroids of the resulting clusters based on these features. The top features were chosen using the approach suggested by Prasetiyowati et al. (2021) discussed in the methodology. The features above the mean threshold of 10.6, and thus deemed important, are the top 11: `MonthlyIncome`, `Age`, `TotalWorkingYears`, `PercentSalaryHike`, `RelationshipSatisfaction`, `TrainingTimesLastYear`, `JobInvolvement`, `MaritalStatus`, `YearsSinceLastPromotion`, `JobLevel`, `BusinessTravel`.

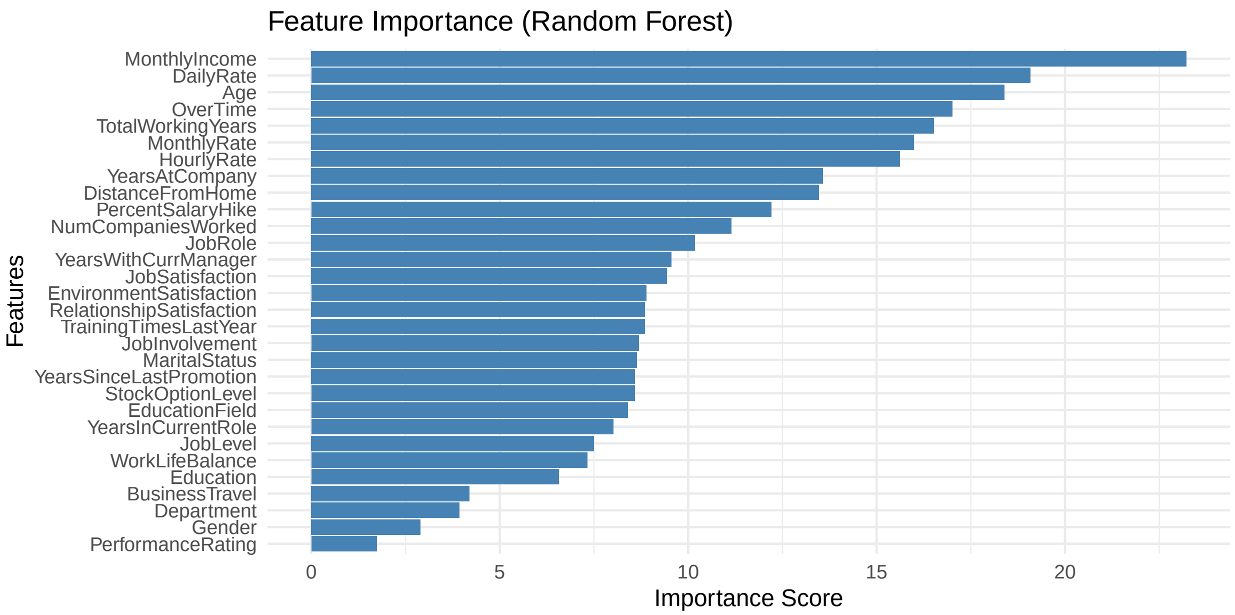


Figure 6: Random forest feature importance plot showing the relative contribution of each variable to the model's predictions.

Clustering

K-Means vs. Hierarchical Clustering Results

To compare the results of the two clustering algorithms, the average silhouette scores, which is a measure of how similar each point is to its own cluster compared to other clusters, were used. The **K-means algorithm resulted in a higher score**, and so it was the chosen algorithm for further exploration.

Exploring Subgroups

We first visualise the resulting clusters in the lower dimensional space we created by implementing classical MDS. As shown in figure 7 on the next page, there is great overlap between observations in clusters 1 and 2. The observations in cluster 3 are considerably separated from the rest of the observations.

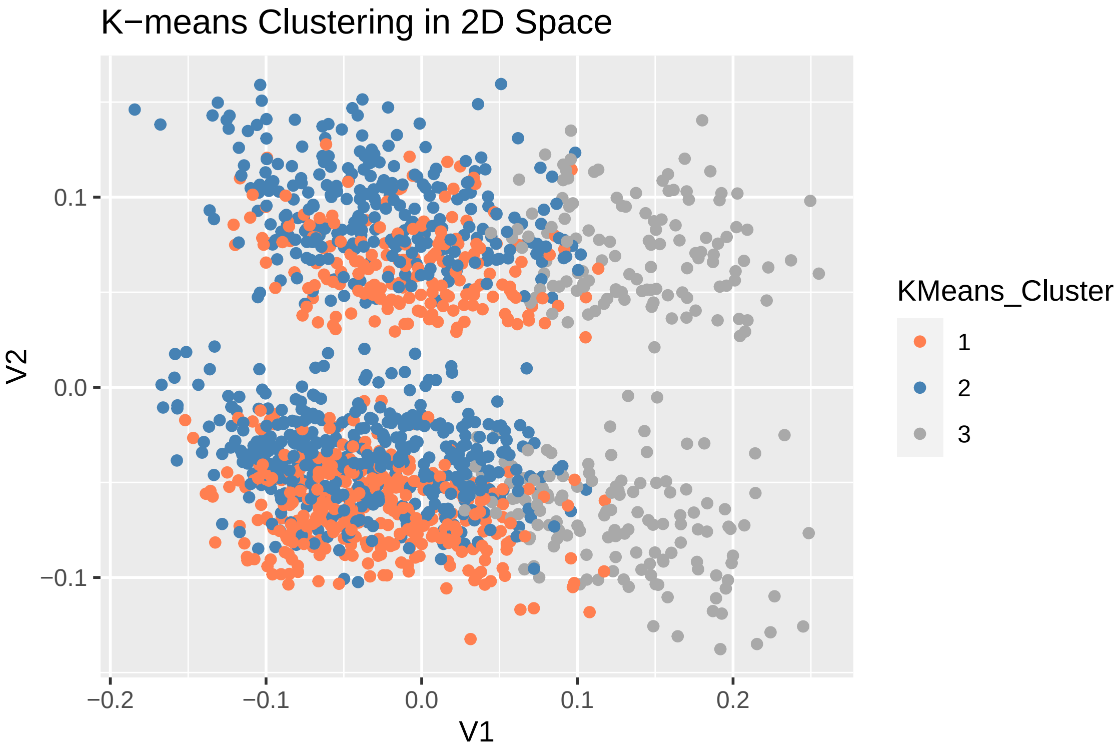


Figure 7: Two-dimensional MDS projection of the dataset with colors indicating cluster assignments from k-means clustering.

A closer look at the cluster centroids provided insight into the overlap and separation of the clusters observed above. We analysed the centroids across all features, identifying those with significant separation and different interpretations. Many of the important features, as indicated by the feature importance plot, were monetary. To simplify the analysis and reduce crowding the results with repetitive information, we focused only on the most important monetary feature, **MonthlyIncome**. The resulting profiles across other features with considerable separation are summarised in the table below:

Table 1: Cluster Profiles for Selected Features

Cluster	MonthlyIncome	Age	TotalWorkingYears	DistanceFromHome	YearsAtCompany
Cluster 1	4767.76	35.55	8.93	9.40	5.26
Cluster 2	4802.71	34.60	8.44	9.18	5.47
Cluster 3	14411.63	45.74	23.40	8.82	14.52

We can see that for cluster 1 and 2, their centroid values are quite close while cluster 3's value are significantly different, hence the observed separation. The attrition percentages for each cluster are summarised below:

Table 2: Percentage of 'Yes' Attrition in Each Cluster

Cluster	Attrition %
1	16.31
2	19.51
3	6.87

According to the above percentages, observations in cluster 3 are far less likely to leave the workplace compared to those in clusters 1 and 2.

Cluster Profiles and Subgroups Discussion

Cluster 1: This group consists of young professionals in their mid-to-late thirties, with nearly 9 years of experience. They have the shortest average tenure at their current companies, approximately 5 years, and earn the lowest monthly income, around 4,700 monetary units. They are the second most likely to leave their jobs.

Cluster 2: Comprising individuals in their early-to-mid thirties, this cluster has an average work experience of 8 years and a slightly longer tenure at their current company, around 5.5 years. They leave furthest away from their workplaces, which means they have the longest commute times. Their monthly income, just under 5,000 monetary units, is comparable to that of Cluster 1. However, they have the highest likelihood of leaving their workplace.

Cluster 3: This is the oldest group, with members in their mid-to-late forties. They have extensive experience, exceeding 20 years, and have been with their companies for nearly 15 years on average. Their homes are closest to work, so they have the shorted commute distances. Their earnings are significantly higher, at over 14,400 monetary units per month. This cluster is the least likely to leave, with an attrition rate of just 6%.

Support Vector Machines

As discussed in the methodology, 5 different SVM models were fitted to explore if balancing the dataset, hyperparameter tuning, and reducing the feature set would improve model performance. We also evaluated the performance of the random forest classifier that was built for feature importance interpretation. The confusion matrices in figure 8 on the next page summarise the classifications of each respective approach:

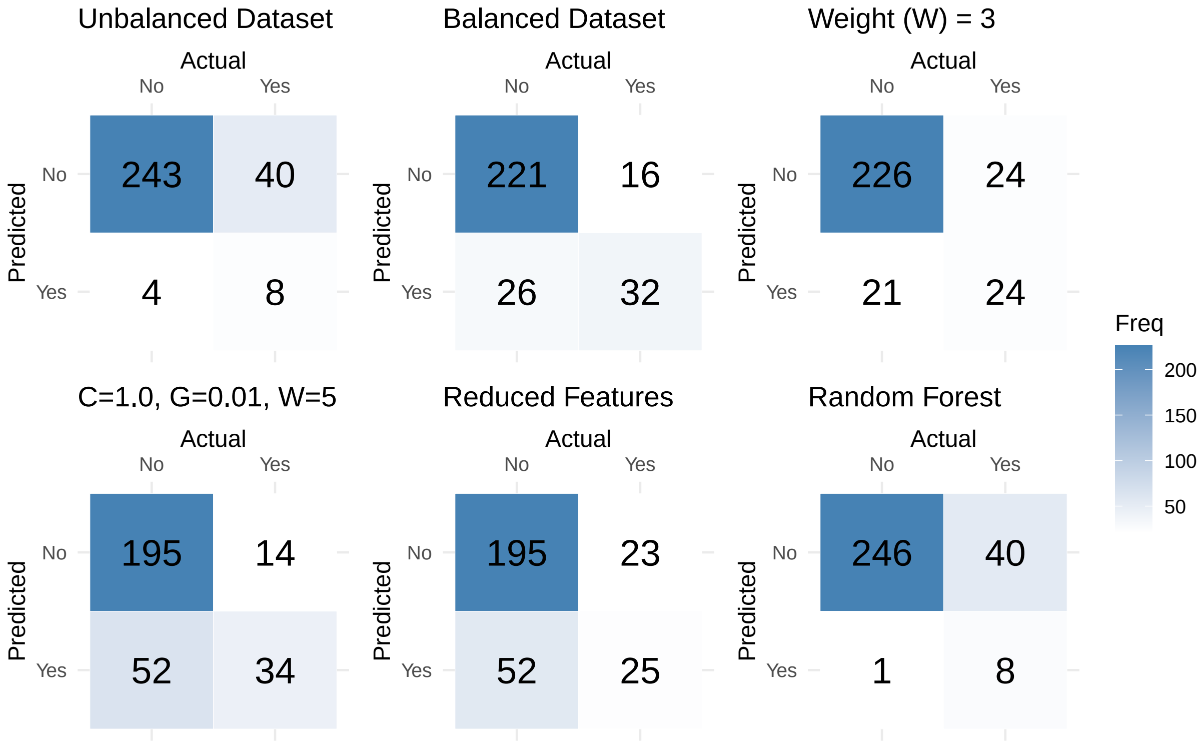


Figure 8: Confusion matrices for the different SVM configurations and the random forest classifier.

The different models were evaluated based on precision, recall, and accuracy. Table 3 shows the generated performance metrics:

Table 3: Classification Metrics for Different Models

Model	Precision	Recall	Accuracy	F1_Score
Unbalanced	0.846	0.229	0.868	0.361
Balanced	0.517	0.625	0.844	0.566
Weight (W) = 3	0.585	0.500	0.861	0.539
C=1.0, G=0.01, W=5	0.481	0.771	0.827	0.592
Reduced Features	0.440	0.771	0.803	0.561
Random Forest	0.889	0.167	0.861	0.281

The performance of the SVM model varies significantly depending on data balancing techniques and parameter tuning. The unbalanced model achieves the highest accuracy (86.8%) and high precision (84.6%), but its recall is extremely low (22.9%), indicating poor detection of the minority class. Balancing the dataset improves recall to 62.5%, but precision drops to 51.7%, leading to a more balanced F1-score of 0.566. Assigning weights (W=3) slightly improves recall (50.0%) while maintaining a moderate precision (58.5%), but the F1-score remains lower than the balanced model. Further parameter tuning (C=1.0, G=0.01, W=5) leads to an even higher recall (77.1%) but at the expense of precision (48.1%), resulting in a slightly improved F1-score (0.592). Reducing features leads to the lowest precision (44.0%) and accuracy (80.3%), but recall remains high (77.1%), suggesting that feature reduction may have removed important predictive variables.

The random forest classifier has the highest precision (88.9%) and a high accuracy (86.1%), but its recall is the lowest observed (16.7%) which means it detects the minority class poorly. The F1-score for this model is the lowest observed (0.281). The **best performing model** is the one with the tuned parameters (C=1.0, G=0.01, W=5) because it has the best balance of precision and recall (highest F1-score, 0.592), and it performed comparatively well across all other metrics.

Discussion

From the clustering analysis, we can conclude that there are significant subgroups within our dataset. The cluster assignments reveal the following key insights:

1. **Higher income and longer tenure** correlate with **lower attrition**.
2. **Younger employees with lower earnings** are more likely to leave.
3. **Career stability** seems to **increase with age and earnings**.
4. Employees who **live closer to work** and have **shorter commute times** tend to stay longer at their workplaces.

Our findings support the hypothesis that salary is a stronger predictor of attrition than work-life balance, not just for younger professionals but across all age groups. Some features initially considered important, such as job satisfaction and percentage salary hikes, turned out to be less influential. This suggests that for our dataset, higher salaries play a dominant role in retention, outweighing other factors. Surprisingly, commute distance proved to be more influential than expected. Employees who lived further from work were more likely to leave, suggesting that long commutes contribute to dissatisfaction and retention challenges. This highlights the potential for organisations to reduce attrition by offering flexible work arrangements.

Conclusions to Hypotheses

Hypothesis 1: Employees with low job satisfaction are more likely to leave.

Our analysis did not strongly support the hypothesis that job satisfaction is a primary driver of attrition. Instead, salary and tenure emerged as the dominant factors, suggesting that employees may tolerate lower job satisfaction if they are well-compensated and have job stability.

Hypothesis 2: Salary hikes have a significant impact on retention.

While compensation is a critical factor in attrition, our findings indicate that absolute salary levels are more influential than periodic salary hikes. This suggests that employees are more likely to stay if they start with a competitive salary rather than relying on future raises to maintain satisfaction.

Hypothesis 3: Salary is a stronger predictor of attrition than work-life balance for young employees.

Our findings confirm that salary is a stronger predictor of attrition than work-life balance, and not just for younger professionals. Employees with lower salaries were significantly more likely to leave, whereas work-life balance factors did not contribute as strongly to the attrition model.

The **key takeaways** from these hypotheses conlusions are:

- Monthly income, age, and career tenure are the most critical factors affecting attrition.
- More experienced professionals, who earn higher salaries, are less likely to leave.
- Younger professionals tend to be more volatile, possibly seeking better-paying opportunities.

However, while these features strongly influence attrition, the process of trying to build a model to predict attrition showed that **alone, the most influencial features are not enough for accurate prediction**. Reducing the feature set led to a decline in model performance. Instead of focusing solely on feature selection, hyperparameter tuning by adjusting cost, weight, and gamma parameters proved to be a more effective strategy for improving predictive performance.

Ultimately, while certain features stand out as key predictors of attrition, the best results are achieved through an approach that **considers all features alongside careful hyperparameter tuning** to refine predictive accuracy.

References

Abdullah, D.M. and Abdulazeez, A.M., 2021. Machine learning applications based on SVM classification a review. Qubahan Academic Journal, 1(2), pp.81-90.

Chua, W.Y., 2023. The factors affecting employee retention among young graduates (Doctoral dissertation, UTAR).

Iqbal, S., Guohao, L. and Akhtar, S., 2017. Effects of job organizational culture, benefits, salary on job satisfaction ultimately affecting employee retention. Review of Public Administration and Management, 5(3), pp.1-7.

Mohammed, R., Rawashdeh, J. and Abdullah, M., 2020, April. Machine learning with oversampling and undersampling techniques: overview study and experimental results. In 2020 11th international conference on information and communication systems (ICICS) (pp. 243-248). IEEE.

Prasetiyowati, M.I., Maulidevi, N.U. and Surendro, K., 2021. Determining threshold value on information gain feature selection to increase speed and prediction accuracy of random forest. Journal of Big Data, 8(1), p.84.

Rakhra, H.K., 2018. Study on factors influencing employee retention in companies. International journal of public sector performance management, 4(1), pp.57-79.

Appendix

Exploratory Data Analysis of Key Features

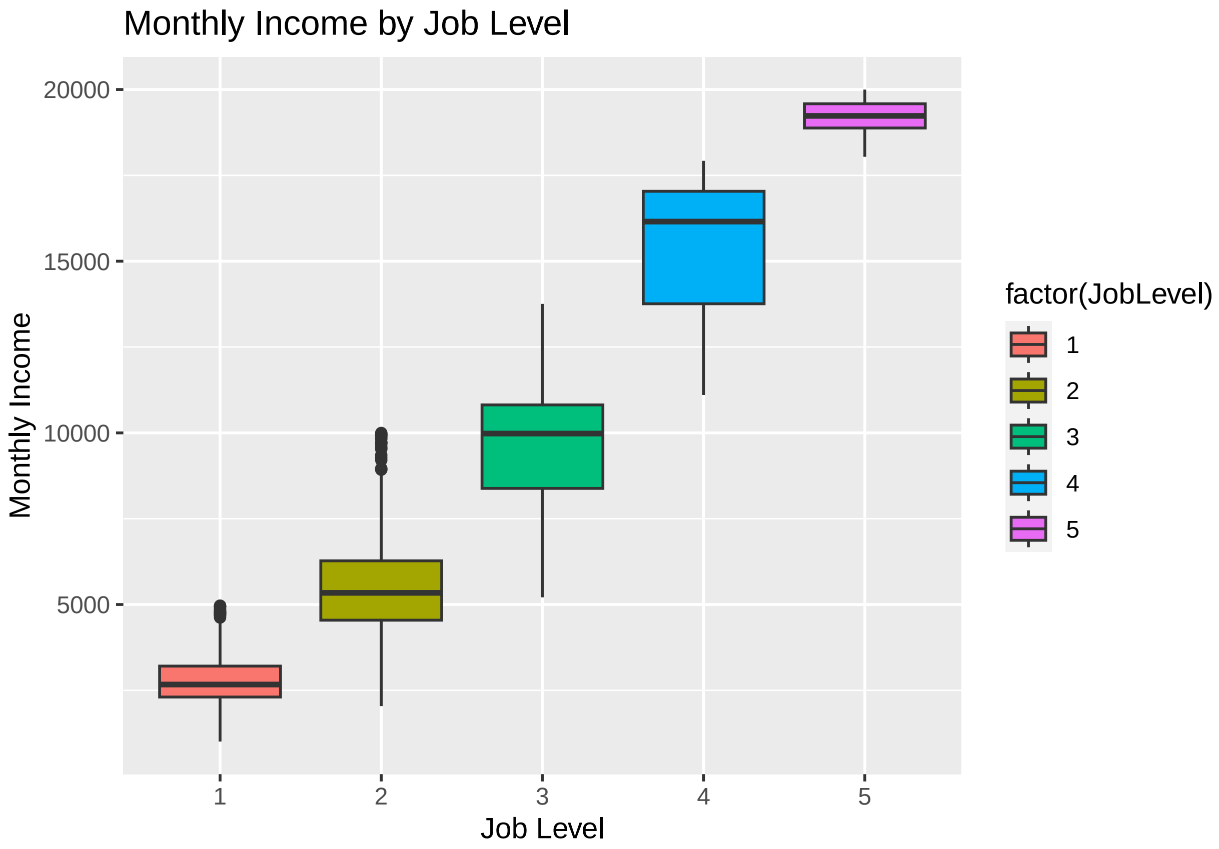


Figure 9: Box plot comparing monthly income across job levels

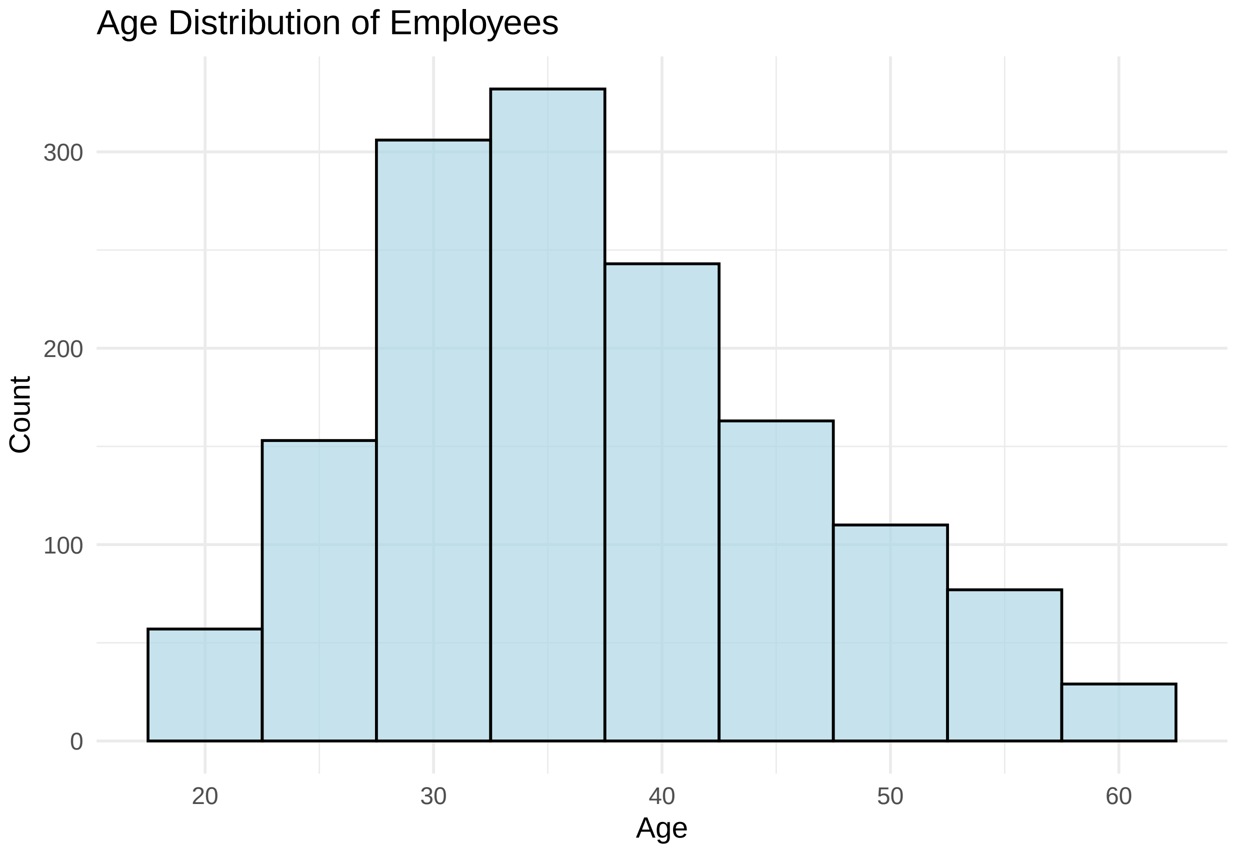
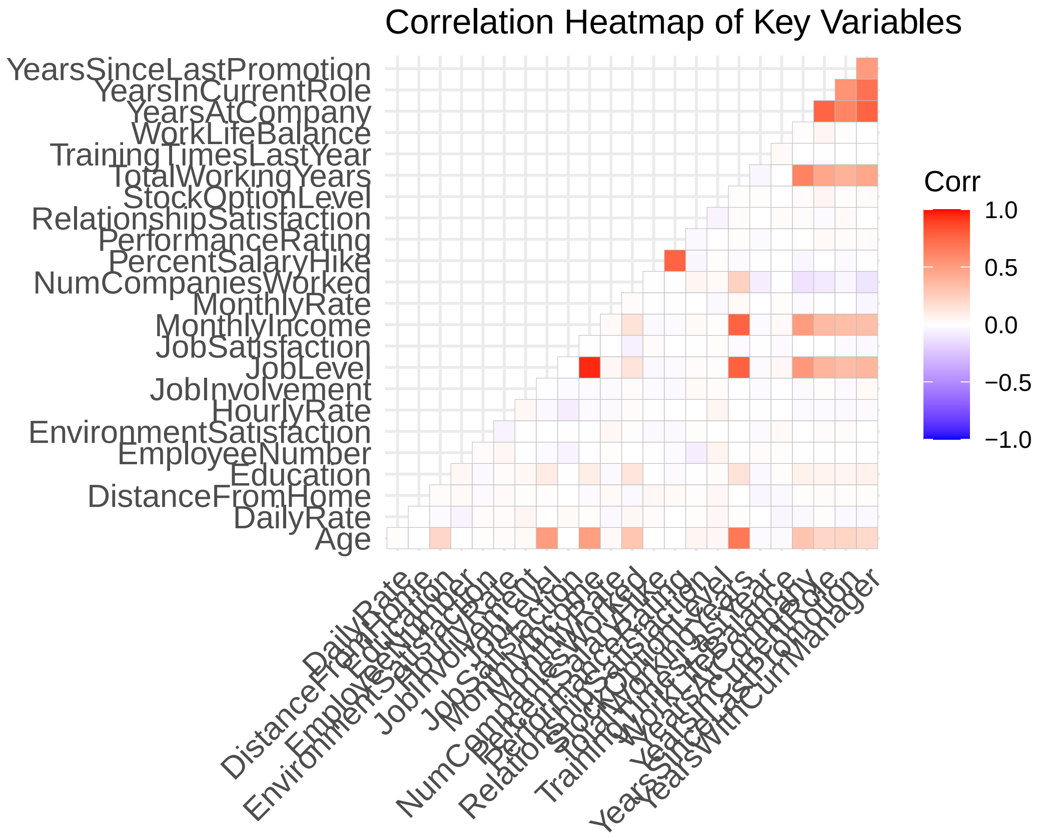


Figure 10: Histogram of employee age distribution


Correlation Heatmap of Key Variables
YearsSinceLastPromotion
YearsInCurrentRole
YearsAtCompany
WorkLifeBalance
TrainingTimesLastYear
TotalWorkingYears
StockOptionLevel
RelationshipSatisfaction
PerformanceRating
PercentSalaryHike
NumCompaniesWorked
MonthlyRate
MonthlyIncome
JobSatisfaction
JobLevel
JobInvolvement
HourlyRate
EnvironmentSatisfaction
EmployeeNumber
Education
DistanceFromHome
DailyRate
Age
DailyRate
DistanceFromHome
Education
EmployeeNumber
EnvironmentSatisfaction
HourlyRate
JobInvolvement
JobLevel
JobSatisfaction
MonthlyIncome
MonthlyRate
NumCompaniesWorked
PercentSalaryHike
PerformanceRating
RelationshipSatisfaction
StockOptionLevel
TotalWorkingYears
TrainingTimesLastYear
WorkLifeBalance
YearsAtCompany
YearsInCurrentRole
YearsSinceLastPromotion
YearsWithCurrManager
Corr
1.0
0.5
0.0
−0.5
−1.0