Abstract/Executive Summary

This study addresses employee attrition to tackle labour shortages and enhance talent retention. The report utilises a synthetic dataset of 500 employees, applying CRISP-DM as the analysis framework. Key features include Job Satisfaction, Monthly Income, Years at Company, and Work-Life Balance. Analytical methods encompass clustering (K-Means, DBSCAN), classification (SVM, Naïve-Bayes), and predictive modelling (Linear and Logistic Regression). Data transformations ensured consistency, and proximity analysis guided feature refinement.

Findings reveal that K-Means offers practical segmentation, highlighting tenure as a critical retention factor. SVM identified Work-Life Balance as the strongest predictor of attrition, while logistic regression confirmed poor Work-Life Balance and shorter tenure as the main attrition drivers. These insights align with literature, demonstrating the importance of tenure and work-life balance in retention strategies. The study underscores the value of predictive analytics in mitigating attrition and fostering workforce stability.

Table of contents

List of Tables

List of Figures

Introduction

Employee attrition is the gradual reduction of workforce within an organisation caused by the employees who leave voluntarily by resigning, or by force arising from layoffs, retirements, or dismissals (Frye *et al.*, 2018). High attrition rates can disrupt workforce stability and negatively impact business performance (Jorgensen, 2005). In industries like IT and healthcare, where attrition rates can reach 23% and 12% respectively, retaining skilled employees has become a top priority (Dharmadhikari, 2013). Figure 1 (Steinfeld, 2024) clearly illustrates the growing attrition rate in companies over the years. Understanding the key drivers of attrition is essential for developing effective retention ideas.

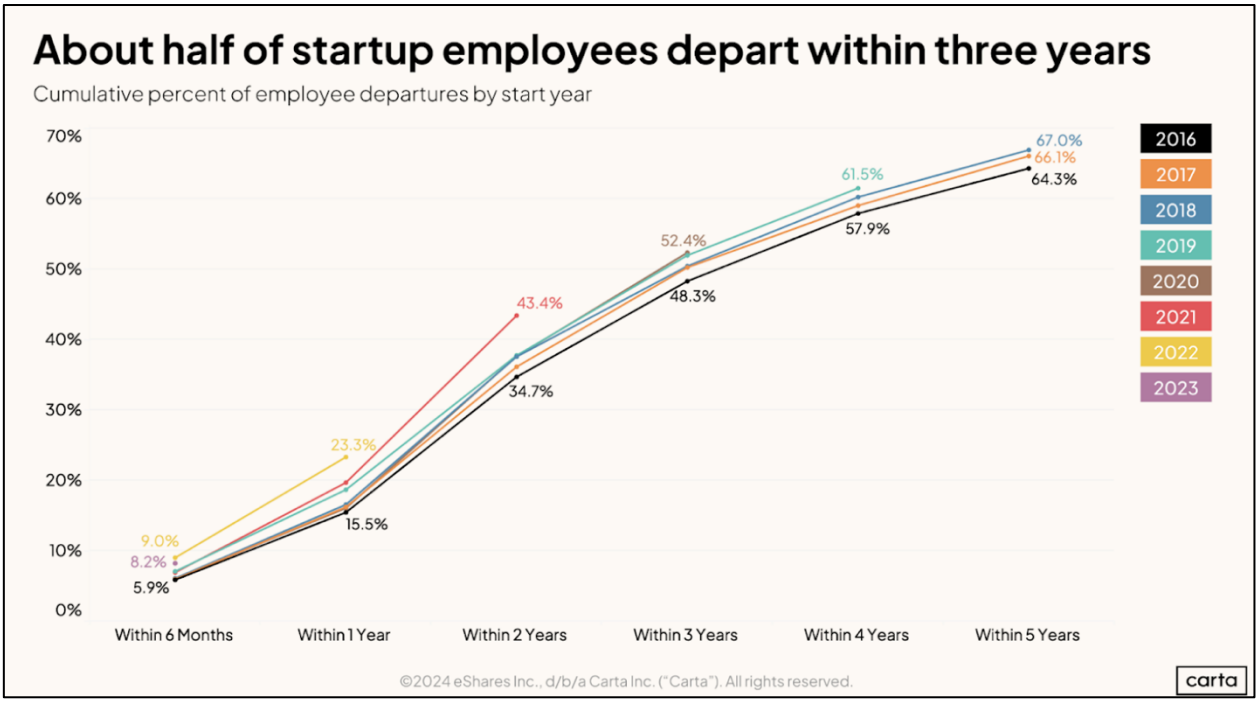


Figure 1: Employee Attrition growth over the years (Source: (Steinfield, 2024)

Business Problem

In today's competitive and dynamic business environment, retaining high-performing employees paves the path for long-term success. Increased employee turnover poses significant challenges to organisations such as increased recruitment and training costs, loss of institutional knowledge, and reduced morale. Replacing an employee can cost approximately 50% of their annual salary, further straining the budget (Bacha, 2016). To combat these issues, traditional retention strategies have proven to be reactive as they fail to address the underlying causes of attrition. Failure to address employee attrition properly can result in operational inefficiencies, reduced productivity, and long-term damage to an organisation's competitive advantage (Goswami and Jha, 2012). In fact, data analytics has proven to be a strong tool in identifying attrition drivers and improving workforce stability; hence, it has the potential to transform Human Resources (HR) decision-making (Hancock and Schaninger, 2022).

Research Questions

The **primary objective** of this study is to fill the gap by leveraging predictive analytics to study employee attrition analysis and generate actionable insights to improve talent retention. The following research questions will guide this analysis:

- What clusters exist among employees based on factors like job satisfaction, income, tenure and work-life balance, and how do these clusters relate to attrition?
- Which factors are the most accurate predictors of employee attrition?
- What factors are most strongly correlated with employee tenure, and which best predict the likelihood of attrition?

Clustering, classification, and regression techniques will be utilized to address these questions, and the answers will shed light on employee behaviour and attrition patterns.

Data Gathering and Data Review

The dataset used for this study is a synthetic employee attrition dataset sourced from Kaggle [Link]. The dataset has been reviewed for its relevance and compatibility with the research objectives. Data quality and reliability have been ensured by performing initial checks for missing values and inconsistencies. The dataset's structured format and completeness with an overall score of 100% on Kaggle make it ideal for clustering, classification, and regression models used in predictive analytics, ensuring accurate insights on employee behaviour and attrition patterns.

Variables	Description	Measurement Unit	Data level	Data Type	References
Employee ID (Identifier)	A unique number assigned to each employee	N/A	Nominal	Qualitative	
Age	The age of the employee in years	Years	Ratio	Quantitative	(Raza *et al.*, 2022)
Job Role	The specific role or department the employee works in (e.g., Technology, Finance)	N/A	Nominal	Qualitative	(Latha, 2013)
Monthly Income	The employee's monthly salary.	Currency (GBP)	Ratio	Quantitative	(Raza *et al.*, 2022), (Latha, 2013), (Farkiya, 2014)
Job Satisfaction	A rating of the employee's job satisfaction level (e.g., High, Medium)	Rating Scale (1 to 4)	Ordinal	Qualitative	(Latha, 2013), (Deery, 2008)
Years at Company	The number of years the employee has worked at the company	Years	Ratio	Quantitative	(Raza *et al.*, 2022)
Number of Promotions	The total number of promotions the employee has received	Count (Integer)	Ratio	Quantitative	(Latha, 2013)
Work-Life Balance	A rating of the employee's work-life balance (e.g., Fair, Good).	Rating Scale (1 to 4)	Ordinal	Qualitative	(Farkiya, 2014), (Deery, 2008)
Attrition (Class Label)	Indicates whether the employee has left the company (Left) or stayed (Stayed)	Yes/No	Nominal	Qualitative	(Raza *et al.*, 2022), (Farkiya, 2014), (Deery, 2008)

Table 1: Descriptions of the features selected from the dataset (Source: Author)

In Table 1, the **Data Time** and **Data Structure** for all the variables is cross-sectional and structured respectively.

Data Analysis Framework

CRISP-DM was chosen as the framework (Figure 2) for its structured approach to address employee attrition and labour shortages. It ensures clear alignment with business goals by guiding the understanding, preparation, and modelling of data for tasks such as identifying high-risk employees, clustering by engagement factors, and analysing predictors of turnover. Evaluation ensures accuracy, while deployment translates insights into actionable retention strategies. This framework ensures data-driven decisions to mitigate attrition, retain talent, and address organisational challenges effectively.

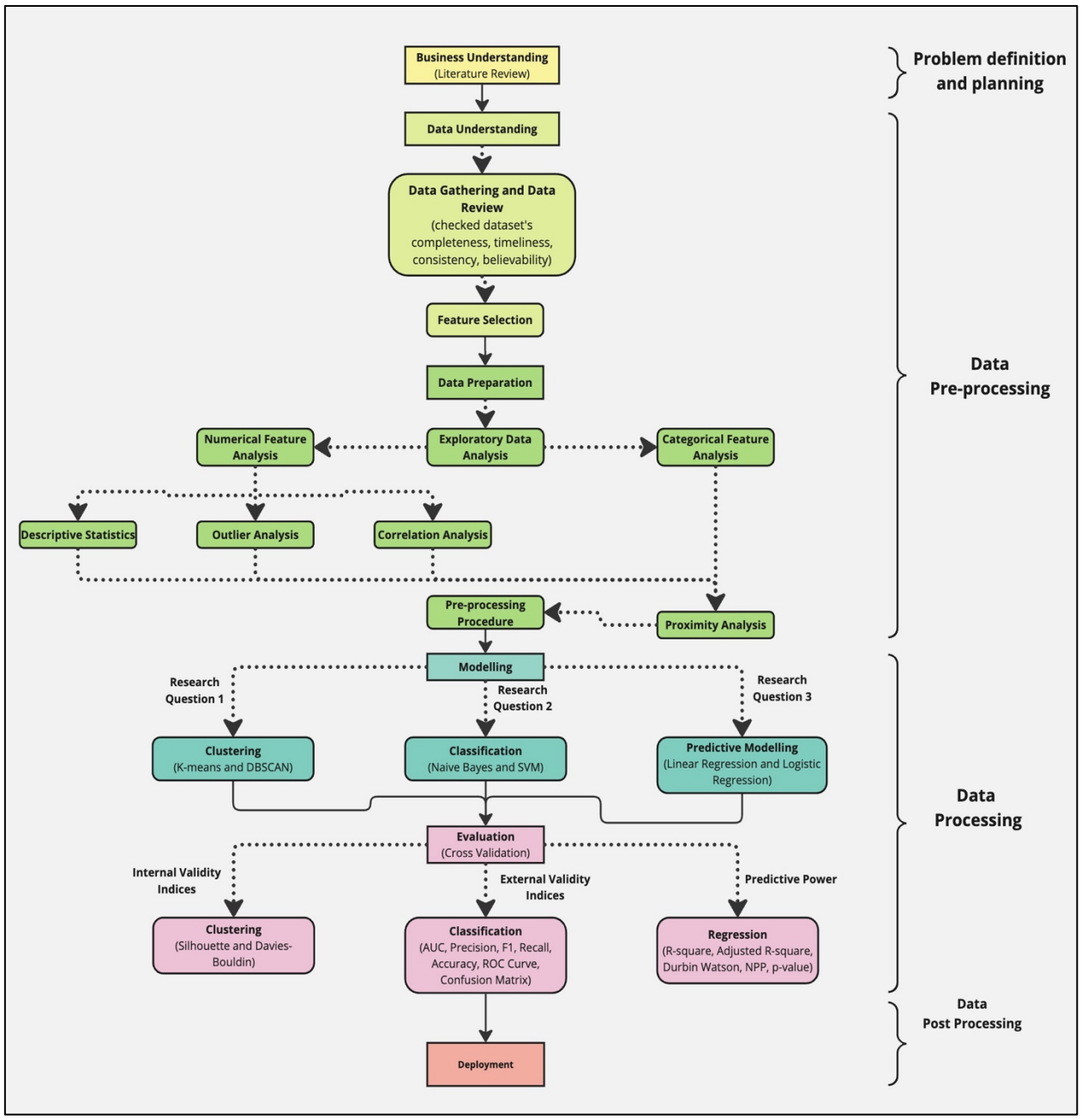


Figure 2: Data Analysis Framework (Source: Author)

Data Preprocessing

Data preprocessing is a fundamental step in data analysis that involves transforming raw data into a clean, structured, and usable format (Alexandropoulos, Kotsiantis and Vrahatis, 2019). The dataset, shown in Table 1, requires thorough exploration to establish a solid foundation for subsequent data pre-processing steps, enabling the extraction of meaningful insights.

1. Exploratory Data Analysis (EDA):

EDA is the process of summarising and visualising data to uncover patterns, detect anomalies, and identify relationships between variables (Komorowski *et al.*, 2016). Exploring the underlying structure of the dataset is essential for informed decision-making during the modelling phase, as it provides initial insights that guide data cleaning, feature selection, and transformation processes to ensure reliable outcomes.

A. Numerical Feature Analysis:

Numerical Feature Analysis involves assessing the primary numerical variables within a dataset using statistical measures and visualizations. This process is important for understanding value distributions, identifying potential outliers, and recognizing trends that may necessitate transformations to enhance the accuracy and reliability of the analysis (Wu and Zhang, 2004).

	Age	Years at Company	Monthly Income	Number of Promotions
Count	500	500	500	500
Mean	37.61	14.84	7301	0.84
Minimum	18	1	2793	0
Q1	27	6	5592	0
Median	37	12	7409	1
Q3	47.25	21	8869	2
Maximum	59	51	13431	4
IQR	20.25	15	3277	2
Range	41	50	10638	4
Variance	145.63	121.21	4671520	0.94
Standard Deviation	12.07	11.01	2161	0.97
Skewness	0.14	1	0.12	0.86
Kurtosis	-1.11	0.60	-0.59	-0.23

Table 2: Analysis of Numerical Variables (Source: Author)

Table 2 and Figure 3 provide insights into the numerical features. The average employee **Age** is 37.61 years (SD: 12.07), ranging from 18 to 59, with a near-normal distribution (skewness: 0.14, kurtosis: -1.11), indicating lighter tails and fewer extremes. The mean of **Years at Company** is 14.83 years, with a right-skewed distribution (skewness: 1.00, kurtosis: 0.59); over 40% of employees have less than 10 years, while a small proportion exceeds 40 years, showing variability in retention.

The average **Monthly Income** is £7,301.45 (range: £2,793–£13,431, SD: £2,161.36), with a near-normal distribution (skewness: 0.12, kurtosis: -0.59), clustering around the mean and median (£7,408.50). The **Number of Promotions** averages 0.84, with a right-skewed distribution (skewness: 0.86, kurtosis: -0.23); most employees received zero or one promotion, with few achieving three or more, reflecting limited variability in career progression.

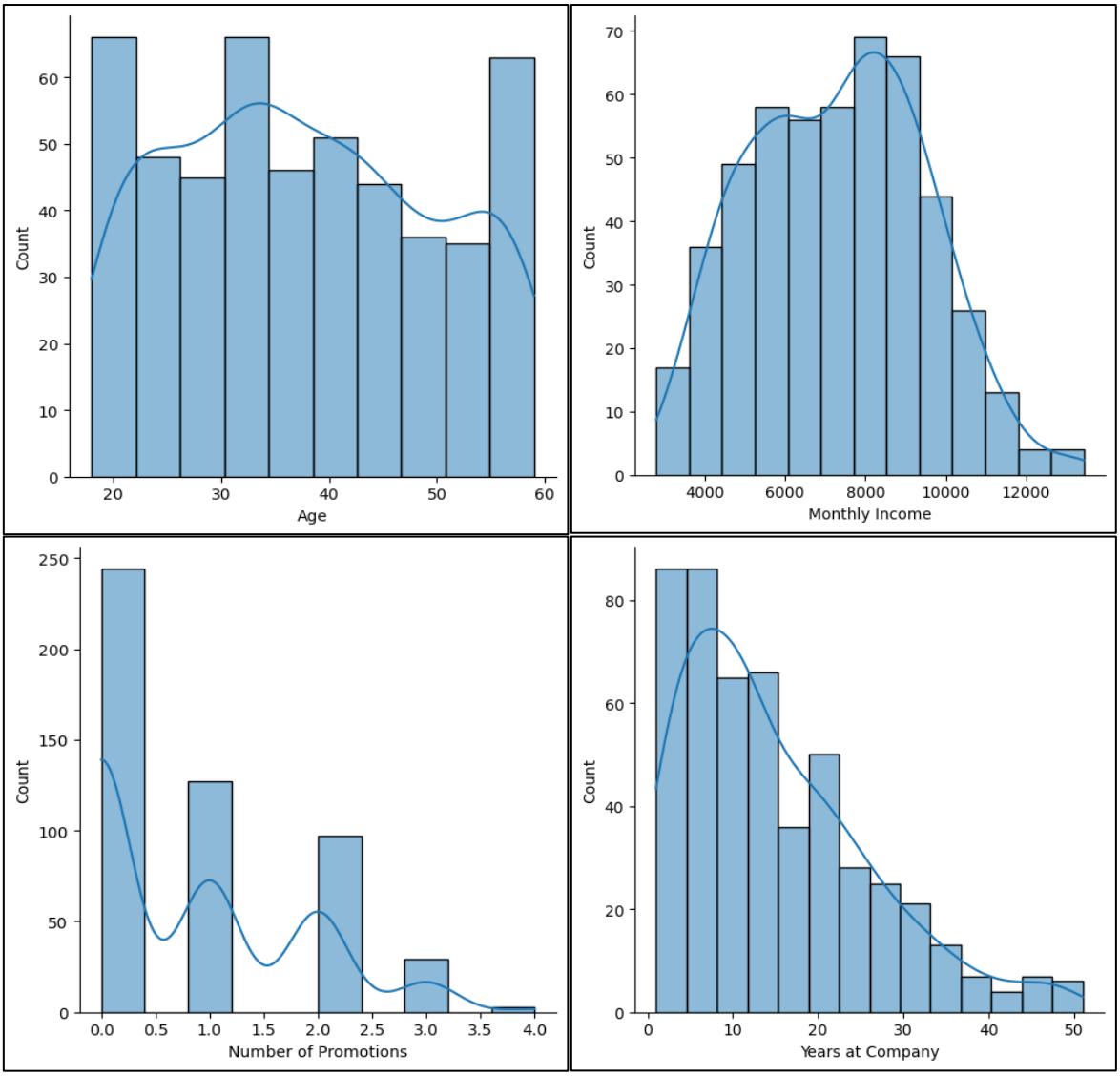


Figure 3: Histograms of Numerical Variables (Source: Author)

Outlier Analysis

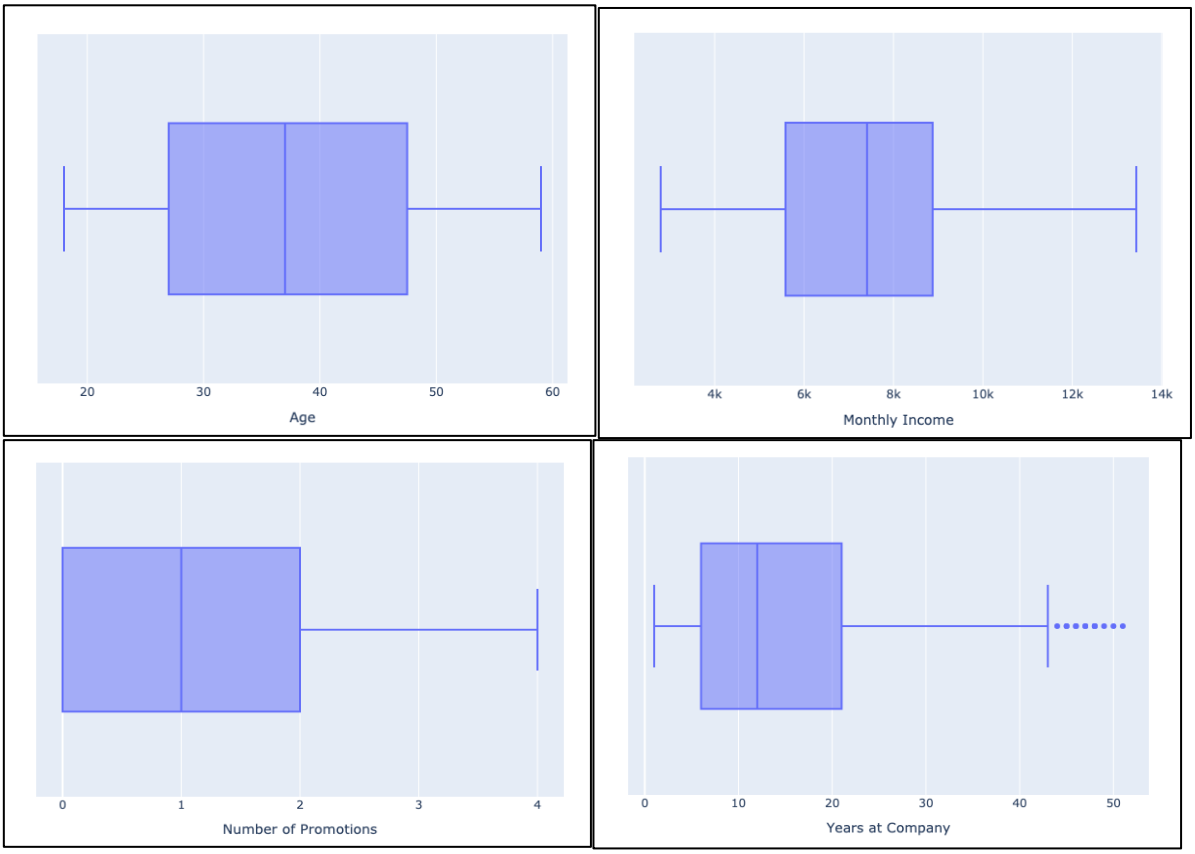


Figure 4: Outlier Analysis of Numerical Variables (Source: Author)

Figure 4 highlights potential outliers in **Years at Company**, with 12 employees having worked for more than 40 years. The **Number of Promotions** and **Monthly Income** boxplots show no significant outliers, while **Age** displays a balanced spread with no extreme values. The overall proportion of outliers is low, indicating that outliers are not a major issue for data reliability.

Correlation Analysis

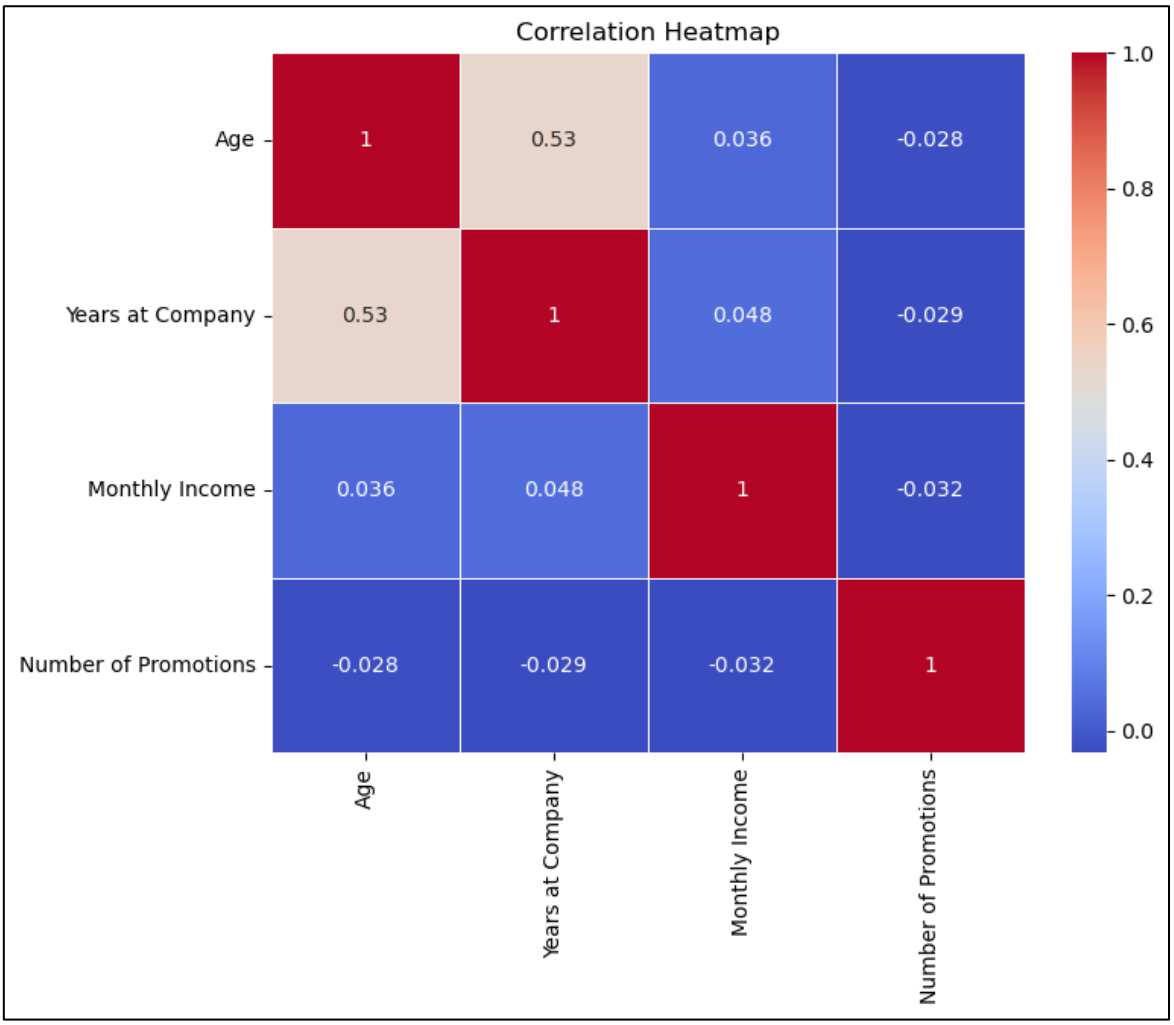


Figure 5: Correlation Heatmap of Numerical Variables (Source: Author)

Figure 5 shows a moderate positive correlation (0.53) between "Age" and "Years at Company," indicating that older employees tend to have longer tenures. Other variables, such as Monthly Income and Number of Promotions, show weak correlations, suggesting minimal linear relationships. This implies that attrition is influenced by multiple factors rather than any single variable.

B. Categorical Feature Analysis:

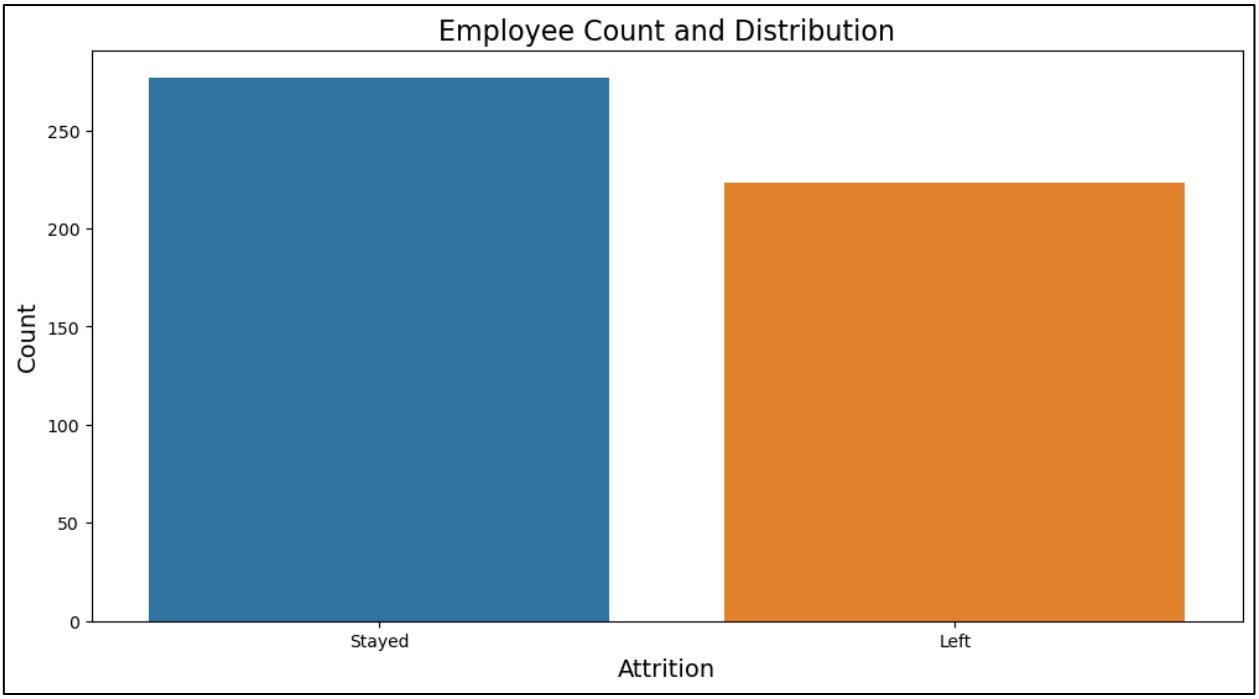


Figure 6: Customer Count and Distribution by Attrition (Source: Author)

Figure 6 shows that out of 500 employees, **265 stayed** the company and **235 left**, resulting in an attrition rate of **47%**. The high attrition rate underscores the importance of identifying factors that drive employee turnover.

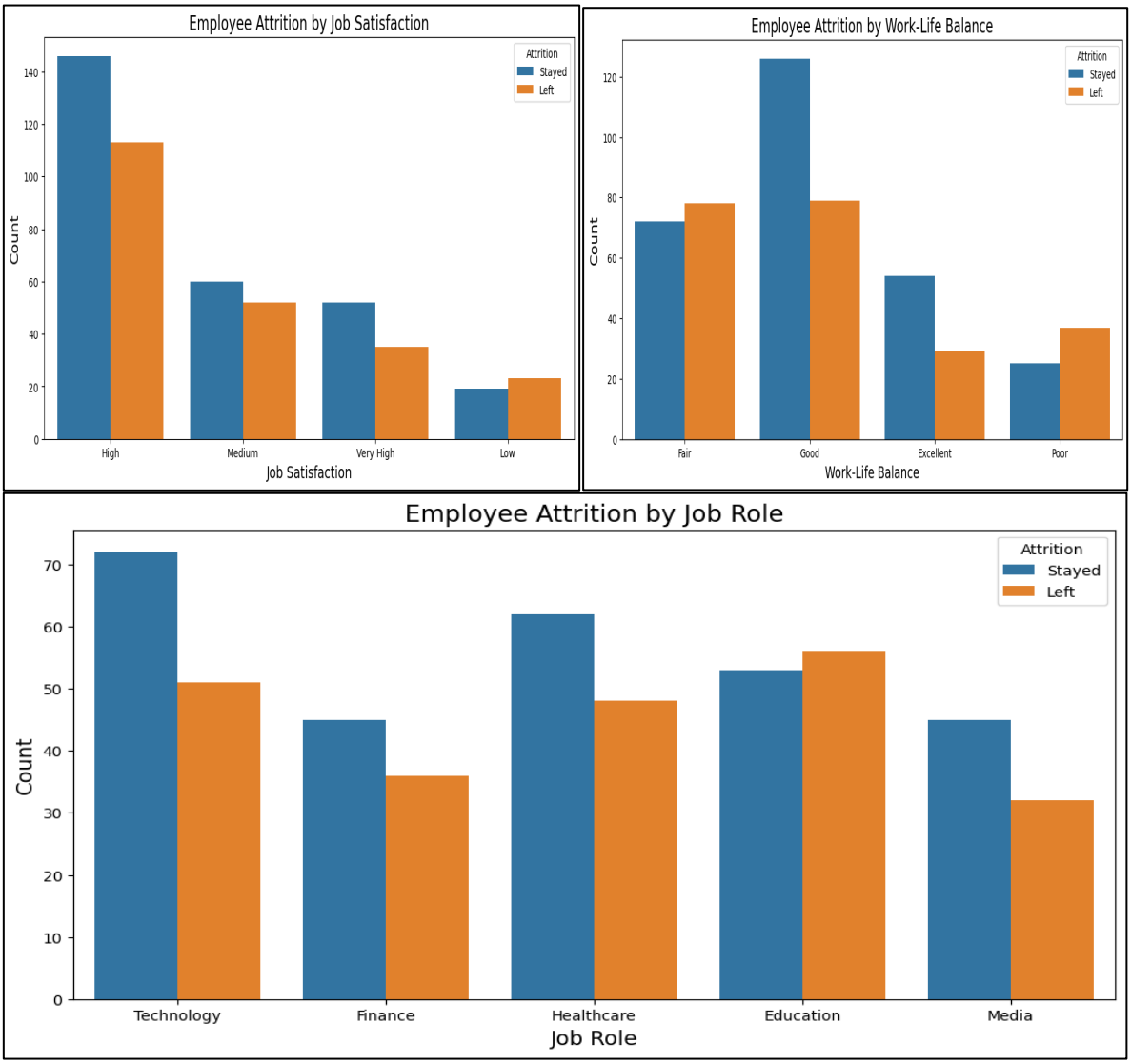


Figure 7: Categorical Variables Count by Attrition (Source: Author)

Figure 7 reveals that attrition for **job role** is highest in the technology sector, where nearly 40% of employees left, followed by the education sector. These roles may face unique challenges contributing to higher turnover. Employees with poor **work-life balance** have the highest attrition rate, with over 55% leaving, compared to lower rates for those with better balance, suggesting improvements in work-life balance could reduce turnover. **Job satisfaction** also impacts attrition, with 60% of dissatisfied employees leaving, compared to 30% of those highly satisfied, indicating that enhancing job satisfaction may help retain staff.

2. Proximity Analysis:

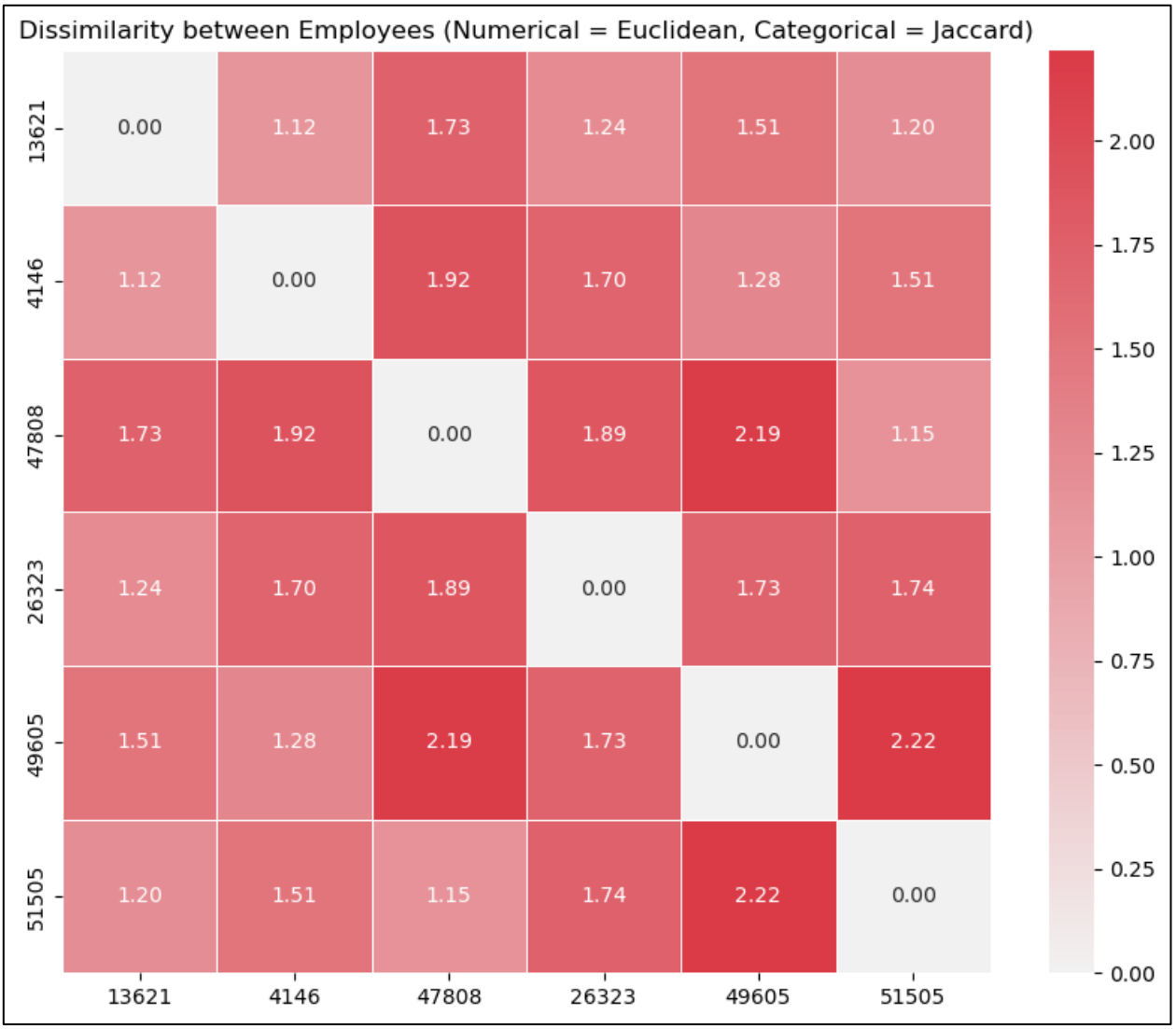


Figure 8: Dissimilarity Matrix – Employee Comparison (Source: Author)

Proximity analysis measures the dissimilarity between employees based on both numerical and categorical variables (Shi, Lee and Whinston, 2016). The dissimilarity matrix uses **Euclidean distance** for numerical features to capture geometric differences and **Jaccard similarity** for categorical features to compare shared attributes. Manhattan distance was avoided as it overemphasised small differences, and cosine similarity was unsuitable as it measures vector angles. The matrix values range from 0 to 2.22, with 0 indicating identical profiles. The highest dissimilarity (2.22) is between employees 49605 and 51505, reflecting significant differences. High dissimilarity scores indicate diverse profiles, requiring personalised retention strategies. Identifying clusters through proximity analysis can help HR tailor strategies to reduce attrition effectively.

3. Pre-Processing Procedure

Data cleaning was not necessary as no missing or null values were found in Excel. The dataset was reduced from 14,900 to 500 rows [Link] for efficient analysis in Orange, retaining key patterns.

For data transformation, numerical columns (**Age, Years at Company, Monthly Income, and Number of Promotions**) require normalisation and standardisation and will be done in Data Processing. Categorical features like **Job Role** and **Work-Life Balance** will be label encoded in Data Processing for machine learning. PCA was not applied to preserve interpretability. SMOTE was also not applied, but class imbalance will be assessed during classification to ensure fair performance.

	Softwares Used In Data Processing					
	Clustering		Classification		Regression	
Softwares	K-Means	DBSCAN	Naïve-Bayes	Support Vector Machine (SVM)	Linear Regression	Logistic Regression
Minitab	✓				✓	✓
Python	✓	✓	✓	✓	✓	
Orange	✓	✓	✓	✓		✓

Table 3: Software used for further analysis in Data Processing (Source: Author)

Data Processing

Clustering

Cluster analysis, a widely used method for identifying hidden patterns in data (Pradana and Ha, 2021), helps organisations categorise employees into distinct groups based on key factors like job satisfaction, income, tenure, and work-life balance. In this research, **K-Means** and **DBSCAN** clustering algorithms were chosen.

K-Means Clustering:

K-Means was chosen because the selected features—**Job Satisfaction, Monthly Income, Years at Company, and Work-Life Balance**—lack significant outliers, making it suitable for a centroid-based method that minimises variance (Na, Xumin and Yong, 2010). It efficiently handles moderate datasets, quickly identifying distinct employee segments.

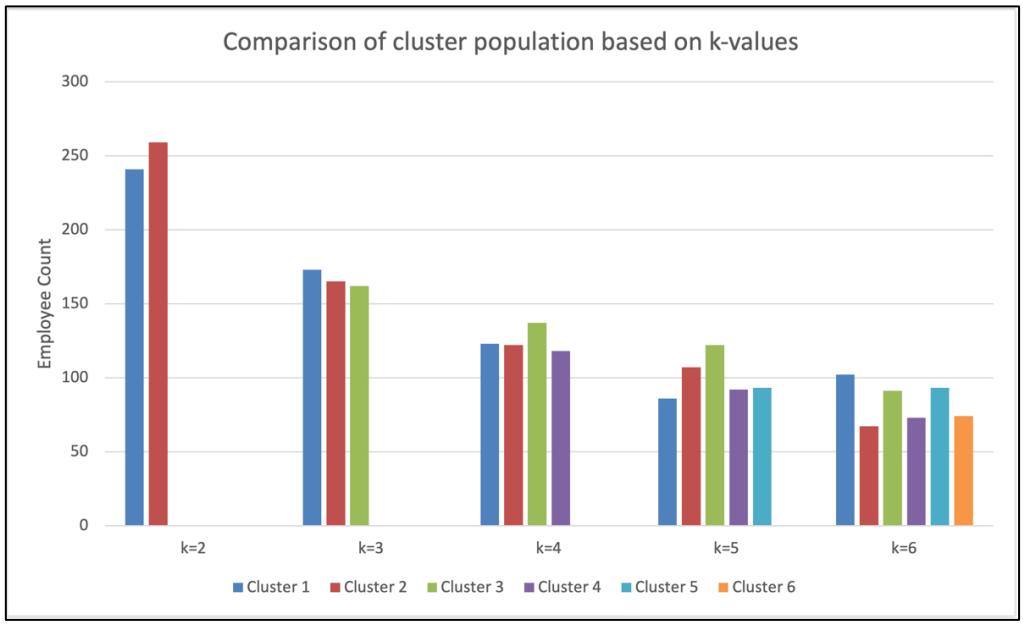


Figure 9: Comparison of cluster population based on k-values (Source: Author)

	k=2	k=3	k=4	k=5	k=6
Cluster 1	241	173	123	86	102
Cluster 2	259	165	122	107	67
Cluster 3		162	137	122	91
Cluster 4			118	92	73
Cluster 5				93	93
Cluster 6					74
Total	500	500	500	500	500

Table 4: Data points clustered based on k-values (Source: Author)

Euclidean distance was chosen as it scales all features equally and aligns with K-Means' goal of minimising squared distances, enhancing interpretability. Figure 4 confirmed the dataset has few outliers, making Euclidean an appropriate choice despite its sensitivity to outliers.

According to Table 4, **k=2**, **k=3**, and **k=4** show balanced cluster populations, indicating that these values create equally distributed segments. Higher values of **k** lead to smaller, uneven clusters, suggesting diminishing returns in segmentation.

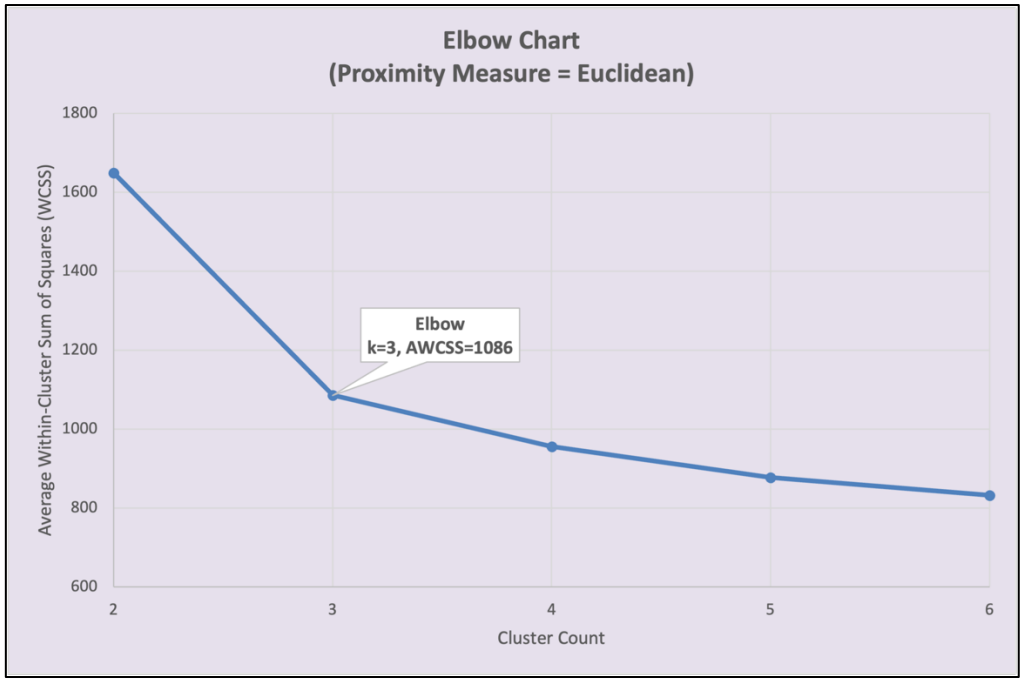


Figure 10: Elbow Chart showing optimal k-value (Source: Author)

Validity Indices	Interpretations	k=2	**k=3**	k=4	k=5	k=6
Silhouette Index	Compares cohesion and distinction (higher is better)	0.193	**0.207**	0.183	0.177	0.172
Davies-Bouldin Index	Similarity of a cluster with the closest cluster (lower is better)	2.057	**1.214**	1.359	1.492	1.516

Table 5: Silhouette and Davies-Bouldin indexes for the k-values (Source: Author)

Figure 10 and Table 5 indicate that **k=3** is the optimal cluster count. At **k=3**, the average within-cluster sum of squares (AWCSS) shows a significant reduction, and the silhouette score (0.207) and Davies-Bouldin index (1.214) suggest balanced clusters with good separation and cohesion, making it a better choice for meaningful segmentation.

DBSCAN Clustering Analysis:

Following K-Means, DBSCAN was applied to explore clusters based on density rather than centroids. This method handles noise and varying cluster shapes effectively (Deng, 2020).

Epsilon	Minimum Points	Clusters Formed	Noise Points	Silhouette Index	Davies-Bouldin Index
0.1	3	34	162	0.307	0.635
0.1	4	24	209	0.308	0.660
0.1	5	13	263	0.400	0.679
0.2	3	17	41	0.281	1.207
0.2	**4**	**14**	**61**	**0.300**	**1.245**
0.2	5	15	68	0.279	1.182
0.3	3	17	14	0.267	1.383
0.3	4	16	20	0.274	1.397
0.3	5	14	30	0.280	1.457
0.4	3	1	1	NaN	NaN
0.4	4	1	1	NaN	NaN
0.4	5	1	1		
0.5	3	1	0	•	•
0.5	4	1	0		
0.5	5	1	0	•	•
0.6	3	1	0		
0.6	4	1	0	•	•
0.6	5	1	0		
0.7	3	1	0	•	•
0.7	4	1	0		
0.7	5	1	0	•	•
0.8	3	1	0		
0.8	4	1	0	•	•
0.8	5	1	0		
0.9	3	1	0	•	•
0.9	4	1	0		
0.9	5	1	0	NaN	NaN

Table 6: Silhouette and Davies-Bouldin indexes for clusters formed (Source: Author)

Euclidean distance was again chosen due to its consistency across multi-dimensional spaces. Table 6 shows various epsilon (ε) and minimum points combinations were tested. Higher ε values resulted in null silhouette and Davies-Bouldin scores, indicating no clusters were formed. The optimal configuration was found at **ε=0.2** and **Minimum Points=4**, forming **14 clusters** with 61 noise points, meaning that 439 data points were used for clustering. The Silhouette Score of 0.3 indicates moderately cohesive clusters, while the Davies-Bouldin Index of 1.245 shows reasonable separation between clusters, supporting this choice of parameters.

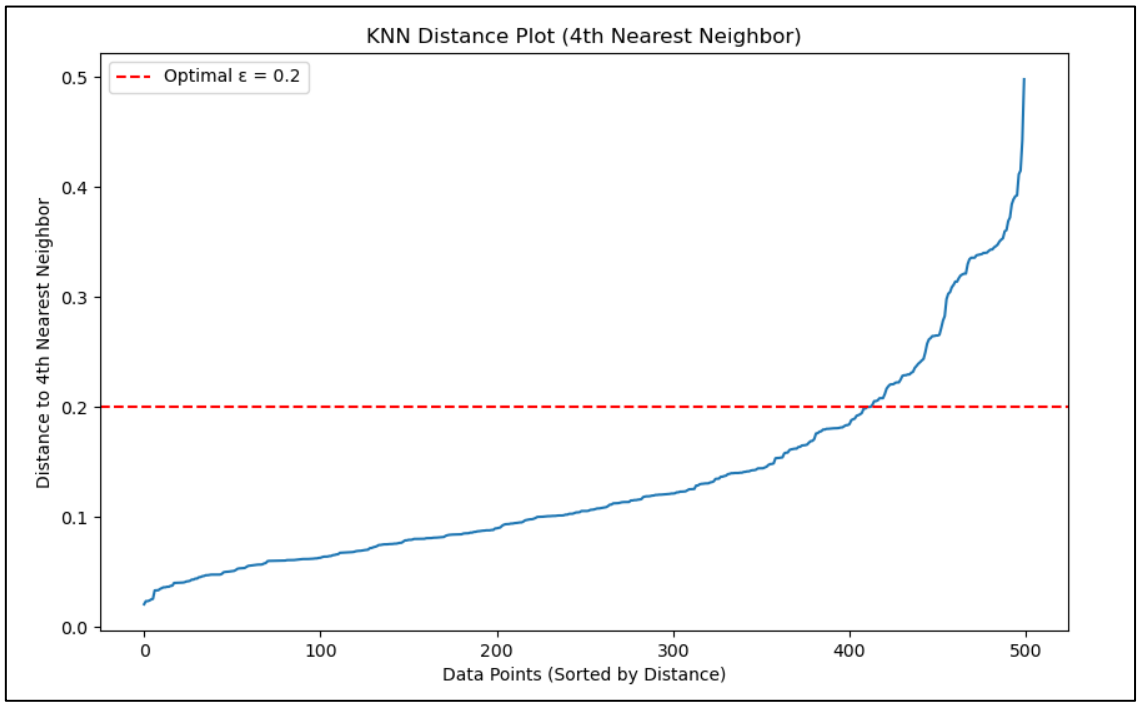


Figure 11: kNN Distance Plot showing Optimal Epsilon on Elbow Chart (Source: Author)

Figure 11 confirmed ε=0.2 as optimal, showing an elbow point indicating the best density threshold (Figure 1). This setting minimised noise while capturing meaningful clusters.

Summary:

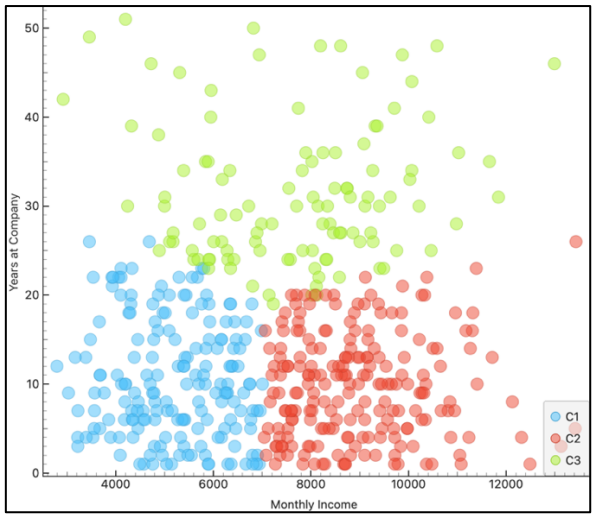


Figure 12: K-Means Scatter Plot (Source Author)

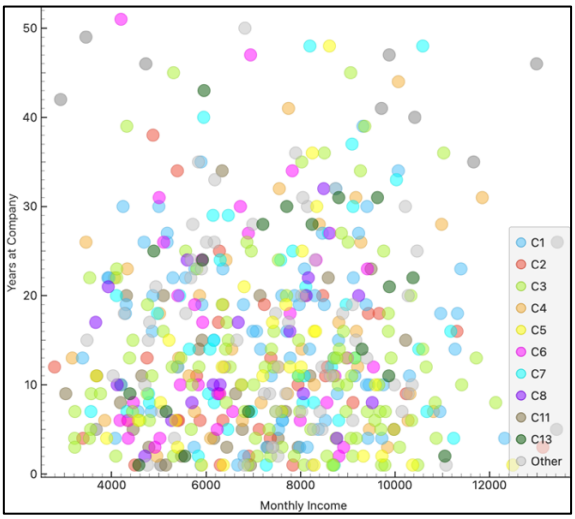


Figure 13: DBSCAN Scatter Plot (Source: Author)

Distances Between Cluster Centroids

	Cluster1	Cluster2	Cluster3
Cluster1	0.0000	2.2873	2.5214
Cluster2	2.2873	0.0000	2.0249
Cluster3	2.5214	2.0249	0.0000

Final Partition

	Number of observations	Within cluster sum of squares	Average distance from centroid	Maximum distance from centroid
Cluster1	144	394.308	1.557	3.077
Cluster2	237	491.011	1.363	3.168
Cluster3	119	304.416	1.526	2.635

Table 7: Twin tables showing cluster distances for K-Means (Source: Author)

Although DBSCAN achieved a higher Silhouette Score, its 14 clusters are impractical for this research. K-Means produced a more interpretable scatter plot (Figure 12) with distinct, well-separated clusters suited for analysing attrition patterns, compared to DBSCAN (Figure 13). Table 7 confirm Cluster 3 as the most distinct, with the greatest distance from Cluster 1 and a good average distance from its centroid. Thus, K-Means is better for addressing employee attrition analysis and the research problem.

Classification

The classification analysis aimed to identify the most accurate predictors of employee attrition using **Naïve-Bayes** and **Support Vector Machines (SVM)** models. Classification is essential for addressing the research question as it enables organisations to predict employee turnover based on key features, allowing for proactive retention strategies (Krishnaiah, Narsimha and Chandra, 2014).

Model Selection

Categorical features were l**abel encoded**, numerical features **standardised**, and the dataset split **70:30** for training and testing, ensuring balanced model evaluation. Class imbalance was **addressed** for fair performance.

Model	AUC	Accuracy	F1	Precision	Recall
Tree	0.510	0.520	0.519	0.519	0.520
kNN	0.459	0.474	0.473	0.472	0.474
Naive Bayes	**0.575**	**0.570**	**0.558**	**0.562**	**0.570**
SVM	**0.535**	**0.556**	**0.504**	**0.540**	**0.556**

Table 8: Model Summary for 70:30 split (Source: Author)

According to Table 8, Naïve-Bayes achieved an AUC of 0.575, accuracy of 57.0%, and F1 score of 0.558, effectively distinguishing attrition cases. SVM showed balanced performance with an AUC of 0.535, accuracy of 55.6%, and F1 score of 0.504.

Model	AUC	Accuracy	F1	Precision	Recall
Tree	0.517	0.524	0.522	0.521	0.524
kNN	0.492	0.494	0.492	0.491	0.494
Naive Bayes	**0.552**	**0.545**	**0.534**	**0.536**	**0.545**
SVM	**0.511**	**0.554**	**0.526**	**0.542**	**0.554**

Table 9: Model Summary for 80:20 split (Source: Author)

Analysis and Validity

Compared with the 80:20 split (Table 9), Naïve-Bayes' AUC dropped to 0.552, and its F1 score fell to 0.534, while SVM's AUC decreased to 0.511, with accuracy improving to 55.4%, indicating better performance with more training data. These results show the impact of train-test split ratios on model performance.

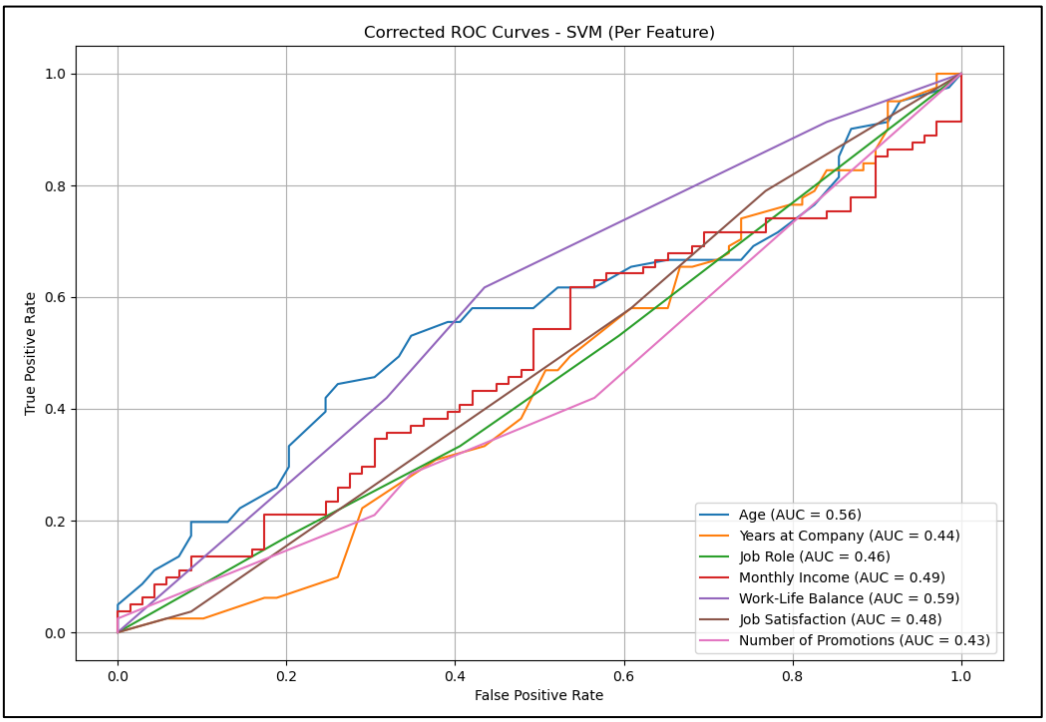


Figure 14: ROC Curve for SVM with AUC values (Source: Author)

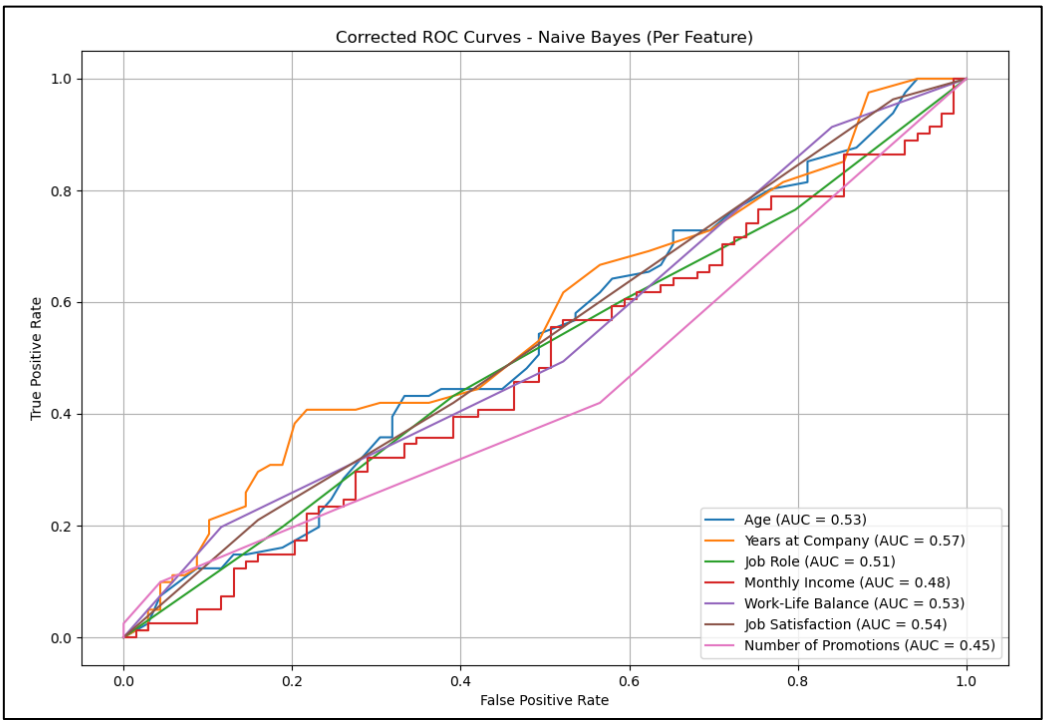


Figure 15: ROC Curve for Naïve-Bayes with AUC values (Source: Author)

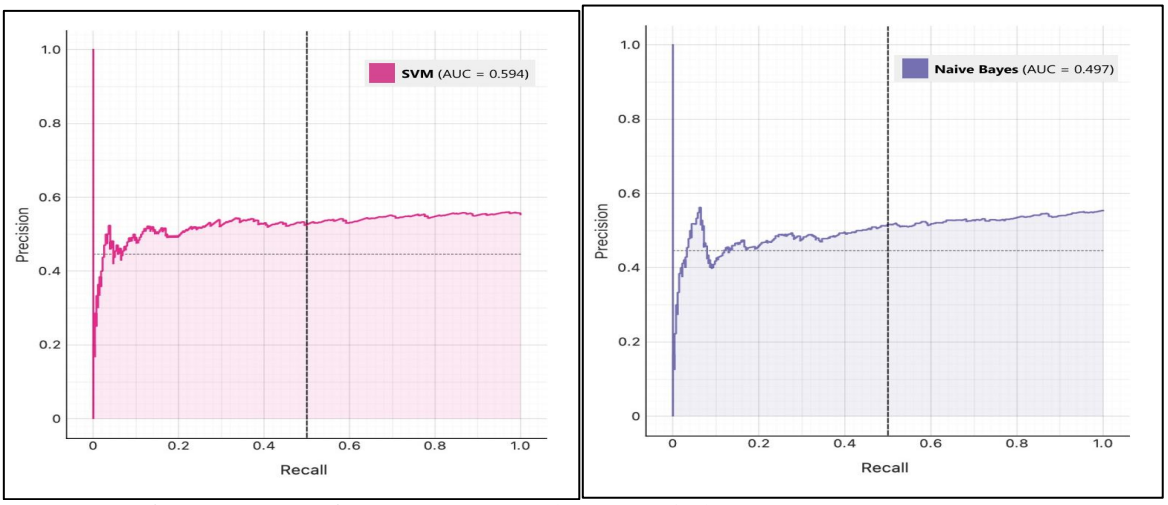


Figure 16: Performance Curves of SVM vs Naïve-Bayes (Source: Author)

Figure 14 for SVM indicate that **Work-Life Balance** (AUC = 0.59) and **Age** (AUC = 0.56) are the best predictors for classifying employee attrition. Figure 16 shows an overall AUC of 0.594 for SVM, confirming it as the more reliable model compared to Naive Bayes (AUC = 0.497). Based on these findings, only Work-Life Balance and Age will be considered for **further analysis** using confusion matrices for **SVM**.

Figure 17: Confusion Matrices of Work-Life Balance vs Age using SVM (Source: Author)

In Figure 17, for **Age**, the model identified 33 true positives and 229 true negatives but had 190 false positives and 48 false negatives, showing Age is a poor predictor with many false positives.

For **Work-Life Balance**, the model achieved 108 true positives and 120 true negatives but had 115 false positives and 157 false negatives. Reducing false negatives is crucial, as they represent employees predicted to stay but who leave, aligning with the goal of identifying at-risk employees.

Based on the analysis, Work-Life Balance is the best predictor with SVM. Prioritising the reduction of false negatives can help organisations identify and retain employees at risk of leaving.

Predictive Modelling

Linear Regression

Linear regression can help identify key factors that influence employee tenure by analysing their impact on the duration of employment. This analysis aims to formulate a predictive equation using continuous features (Uyanık and Güler, 2013), such as Age, Monthly Income, and Number of Promotions to better understand tenure patterns.

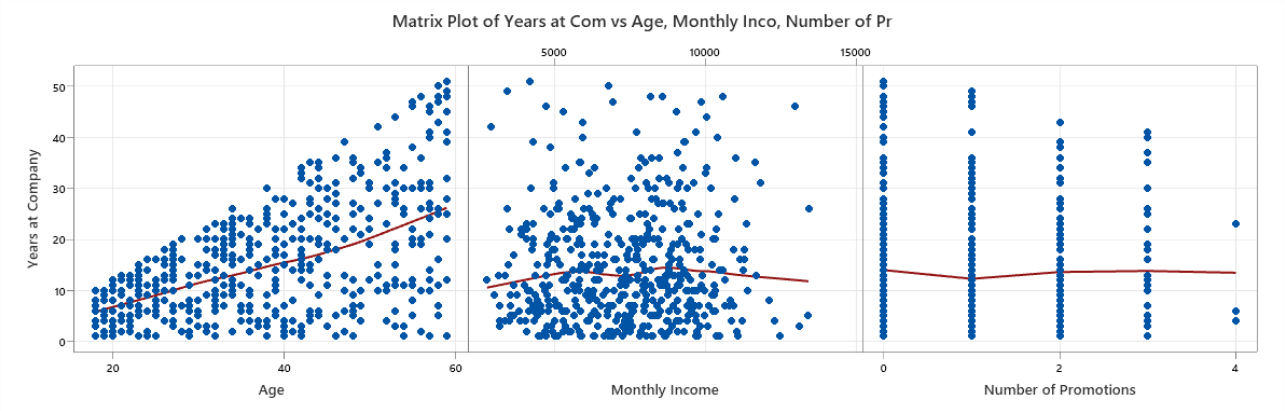


Figure 18: Matrix Plot of Years at Company by other numerical features (Source: Author)

Figure 18 shows a linear relationship between Age and Years at Company, while Monthly Income and Number of Promotions display weaker correlations. The trends suggest non-linear regression is unsuitable, supporting the use of linear regression to formulate a predictive equation for tenure patterns.

Regression with all the features

Model Summary:	
R-square	0.281
Adjusted R-square	0.277
Durbin-Watson	1.942

Coefficients Table:

Variable	Coefficient	Standard Error	t-value	P>|t|	VIF
Intercept	-4.208	1.988	-2.117	0.03	
Age	0.482	0.035	13.857	0	1.002
Monthly Income	0	0	0.744	0.46	1.002
Number of Promotions	-0.15	0.432	-0.346	0.73	1.002

Regression after backward elimination

Model Summary:	
R-square	0.28
Adjusted R-square	0.279
Durbin-Watson	1.939

Coefficients Table:

Variable	Coefficient	Standard Error	t-value	P>|t|	VIF
Intercept	-3.326	1.37	-2.428	0.015521	
Age	0.483	0.035	13.925	1.83E-37	1

Table 10: Model Summary and Coefficients Table before and after backward elimination (Source: Author)

According to Table 10, the initial regression model used Age, Monthly Income, and Number of Promotions to predict Years at Company, achieving an R-square of 0.281 and an Adjusted R-square of 0.277, explaining 28.1% of tenure variation. Monthly Income and Number of Promotions had p-values above 0.05 (0.457 and 0.729), indicating they were not significant predictors. The null hypothesis suggests no impact on tenure for these variables, cannot be rejected.

Initial Linear Regression Equation:
Years at Company = -4.208 + 0.482*Age + 0.000*Monthly Income + -0.150*Number of Promotions

Backward elimination refined the model, retaining only Age as a significant predictor with a p-value of 1.83e-37 and a coefficient of 0.483, indicating a strong positive relationship with Years at Company. The adjusted R-square remained 0.279, and the Durbin-Watson statistic was 1.939, confirming minimal autocorrelation. VIF scores below 5 indicate no multicollinearity. Figure 19 shows residuals aligned with theoretical quantiles, satisfying the normality assumption. The final model confirms Age as the most significant predictor of employee tenure, with no multicollinearity issues present.

Linear Regression Equation after backward elimination:
Years at Company = -3.326* + 0.483*Age

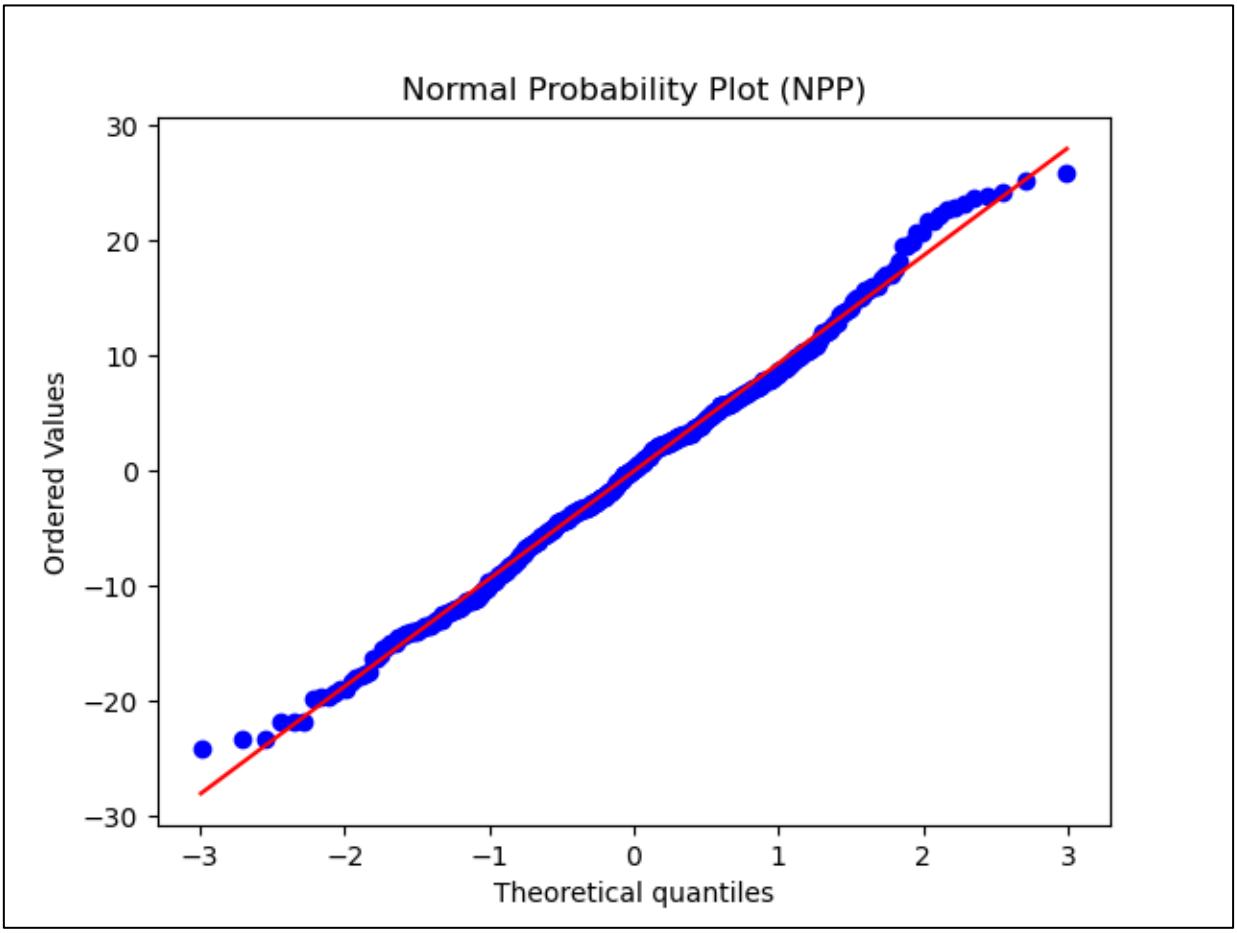


Figure 19: Normal Probability Plot after backward elimination (Source: Author)

Logistic Regression

Logistic regression effectively models binary outcomes (Choi *et al.*, 2005), such as predicting whether an employee will leave or stay. The Attrition class "Left" was set as the event of interest to analyse attrition. The dataset was split 70:30 for training and testing.

Model Summary

Deviance R-Sq	Deviance R-Sq(adj)	AIC	AICc	BIC	Test Area Under ROC Curve	Test Deviance R-Sq	Test Area Under ROC Curve
5.61%	2.73%	487.27	488.71	545.14	0.6560	0.00%	0.5773

Coefficients

Term	Coef	SE Coef	Z-Value	P-Value	VIF
Constant	0.062	0.654	0.10	0.924	
Age	0.0127	0.0111	1.14	0.255	1.48
Years at Company	-0.0315	0.0125	-2.52	0.012	1.50
Monthly Income	-0.000010	0.000091	-0.11	0.909	3.09
Number of Promotions	-0.123	0.117	-1.05	0.292	1.04
Job Role					
Finance	-0.559	0.525	-1.07	0.287	2.77
Healthcare	-0.456	0.465	-0.98	0.327	2.99
Media	-0.589	0.399	-1.48	0.139	1.65
Technology	-0.312	0.530	-0.59	0.556	4.38
Work-Life Balance					
Fair	0.652	0.354	1.84	0.065	2.12
Good	0.113	0.332	0.34	0.734	2.20
Poor	1.069	0.459	2.33	0.020	1.50
Job Satisfaction					
Low	0.694	0.436	1.59	0.112	1.07
Medium	0.032	0.290	0.11	0.913	1.14
Very High	-0.252	0.314	-0.80	0.422	1.10

Regression Equation

P(Left) = exp(Y')/(1 + exp(Y'))

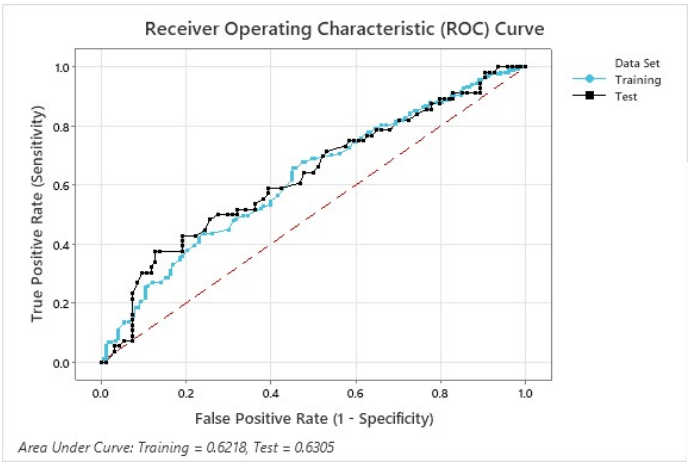
Y' = 0.062 + 0.0127 Age - 0.0315 Years at Company - 0.000010 Monthly Income
- 0.123 Number of Promotions + 0.0 Job Role_Education - 0.559 Job Role_Finance
- 0.456 Job Role_Healthcare - 0.589 Job Role_Media - 0.312 Job Role_Technology
+ 0.0 Work-Life Balance_Excellent + 0.652 Work-Life Balance_Fair
+ 0.113 Work-Life Balance_Good + 1.069 Work-Life Balance_Poor
+ 0.0 Job Satisfaction_High + 0.694 Job Satisfaction_Low + 0.032 Job Satisfaction_Medium
- 0.252 Job Satisfaction_Very High

Table 11: Model Summary, Coefficients Table and Regression Equation before backward elimination (Source: Author)

Table 11 shows the initial model included predictors such as Age, Years at Company, Monthly Income, Number of Promotions, Job Role, Work-Life Balance, and Job Satisfaction. It yielded a deviance R-squared of 5.61% and an AUC of 0.6560. However, predictors like Monthly Income and Job Role categories had p-values over 0.05, indicating a lack of statistical significance.

Model Summary

Deviance R-Sq	Deviance R-Sq(adj)	AIC	AICc	BIC	Area Under ROC Curve	Test Deviance R-Sq	Test Area Under ROC Curve
3.47%	2.65%	477.64	477.81	496.93	0.6218	0.00%	0.6305



Coefficients

Term	Coef	SE Coef	Z-Value	P-Value	VIF
Constant	-0.064	0.328	-0.20	0.844	
Years at Company	-0.0244	0.0102	-2.40	0.017	1.01
Work-Life Balance					
Fair	0.649	0.345	1.88	0.060	2.07
Good	0.092	0.325	0.28	0.777	2.18
Poor	1.070	0.447	2.39	0.017	1.48

Table 12: Model Summary, ROC Curve and Coefficients Table after backward elimination (Source: Author)

Table 12 shows backward elimination refined the model to include Years at Company and Work-Life Balance_Poor ($p<0.05$) as the only significant predictors. Employees with poorer work-life balance and fewer years at the company were more likely to leave. The final model's deviance R-squared was 3.47%, with an AUC of 0.6218 for training data and 0.6305 for test data, showing moderate predictive performance.

Logistic Regression Equation after backward elimination:
P(Left) = exp(Y')/(1 + exp(Y'))
Y' = 1.005 + 1.070 Work-Life Balance_Poor - 0.0244 Years at Company

Model	AUC	Accurary	F1	Precision	Recall
Logistic Regression	0.622	0.635	0.598	0.629	0.635

Table 13: Model Summary of Logistic Regression (Source: Author)

According to Table 13, the model achieved 63.5% accuracy, an F1 score of 0.598, precision of 62.9%, and recall of 63.5%, indicating balanced performance in identifying employees likely to leave.

Conclusion

This study sought to address attrition through a multifaceted approach, combining clustering, classification, and predictive modelling to identify actionable insights. The findings, compared with existing literature, offer critical perspectives on attrition dynamics and practical strategies for talent retention.

Recap of Findings

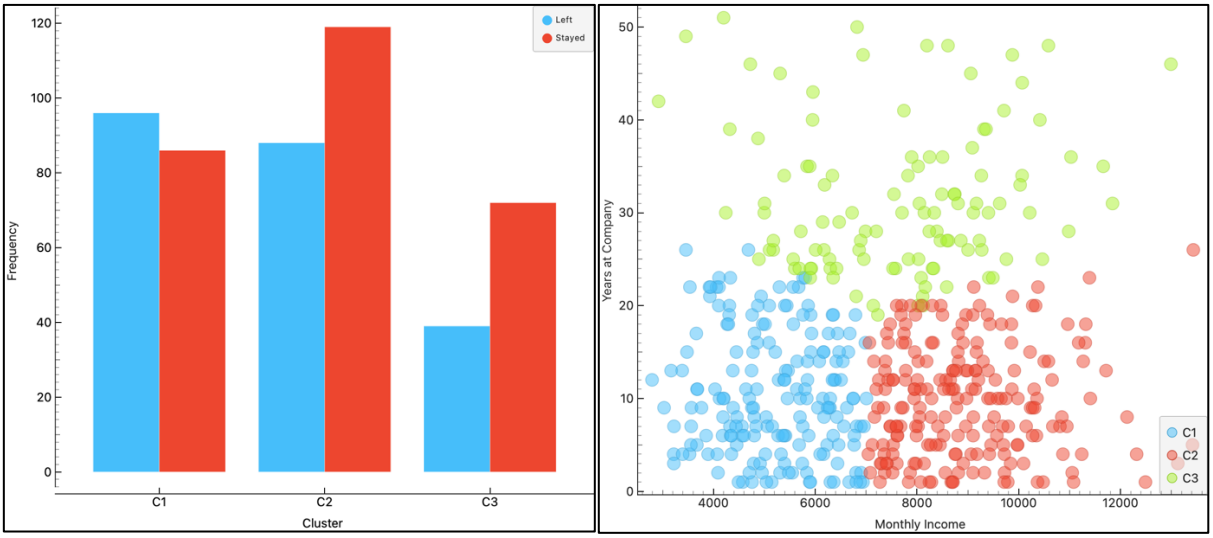


Figure 20: K-Means clustering scatter plot correlating to attrition (Source: Author)

The clustering analysis employed K-Means and DBSCAN algorithms to segment employees based on Job Satisfaction, Monthly Income, Years at Company, and Work-Life Balance. In Figure 20, K-Means, optimised at k=3, revealed that employees with longer tenure were more likely to stay at the company. DBSCAN, although achieving a higher silhouette score, produced an impractical number of clusters (14), making K-Means more suitable for actionable insights. The clustering findings highlighted tenure as a critical factor influencing attrition patterns, providing a foundation for further analysis.

Classification analysis using SVM identified Work-Life Balance as the most significant predictor of attrition, with Age also providing a good AUC score compared to other features. These findings built upon clustering insights by further emphasising tenure's role in understanding attrition risk and highlighting Work-Life Balance as a key determinant.

Predictive modelling further solidified these conclusions. Linear regression highlighted Age as the sole significant predictor of employee tenure, reinforcing the clustering finding that tenure can indicate which employees are likely to stay. Logistic regression identified poor Work-Life Balance and shorter tenure as the strongest predictors of attrition. This connected directly to both clustering's emphasis on tenure and classification's identification of Work-Life Balance as a critical factor. Referring to Table 13, logistic regression outperformed SVM and Naïve Bayes, achieving higher AUC, accuracy, F1, precision, and recall scores, making it the most effective model for predicting attrition.

Comparison with Current Literature

This study's findings align with and extend key themes in current literature. (Deery, 2008) emphasised Work-Life Balance as a cornerstone for retention, linking it to job satisfaction and organisational commitment through comprehensive industry reviews. (Latha, 2013) used descriptive analysis to show the role of Monthly Income in retention, which aligns with this study's clustering insights on income disparities and attrition risk. (Raza *et al.*, 2022) identified Years at Company and Age as critical attrition predictors through machine learning models, corroborating this study's linear regression and clustering results. However, while (Farkiya, 2014) validated Work-Life Balance and income disparities using secondary data, her findings lacked predictive modelling depth, which this study addresses comprehensively. These studies employed methods ranging from hypothesis testing to machine learning, validating this study's robust analytical approach and highlighting the generalisability of its conclusions.

Practical and Managerial Implications

Implementing these findings poses challenges. Enhancing Work-Life Balance through flexible arrangements may face cultural resistance and uneven benefits across clusters. Adjusting compensation for lower-income employees risks budgetary strain and must prioritise high-risk clusters. Predictive models, though effective, introduce ethical concerns like perceived surveillance or decision biases, potentially undermining trust. Successful implementation requires aligning strategies with organisational culture and ensuring transparency. Tailored interventions based on cluster characteristics can mitigate risks and foster equity, maximising organisational impact.

Limitations and Future Recommendations

This study has notable limitations. The synthetic dataset used constrains the applicability of findings to real-world scenarios, as it may oversimplify the nuanced nature of employee behaviours. Additionally, the exclusion of critical exogenous factors, such as economic conditions and personal circumstances, limits the comprehensiveness of the analysis. Potential biases in data selection, including omitted variables like leadership style or employee engagement, may skew results. Methodological assumptions in clustering and regression further constrain accuracy, especially if real-world data deviates from expected distributions.

Future research should address these limitations by utilising real-world datasets that reflect organisational complexities. Incorporating qualitative insights, such as employee interviews, could enrich the analysis of attrition drivers. Advanced machine learning models, such as ensemble techniques, and longitudinal studies can refine predictive accuracy while examining changes in attrition over time. Such efforts will ensure more robust, actionable insights for addressing employee turnover. In conclusion, combining clustering and predictive analytics with targeted actions helps mitigate attrition, retain talent, and secure a competitive edge.

Bibliography

Alexandropoulos, S.-A.N., Kotsiantis, S.B. and Vrahatis, M.N. (2019) 'Data preprocessing in predictive data mining', *The Knowledge Engineering Review*, 34, p. e1. Available at: https://doi.org/10.1017/S026988891800036X.

Bacha, S. (2016) 'Antecedents and Consequences of Employee Attrition: A Review of Literature'. Rochester, NY: Social Science Research Network. Available at: https://doi.org/10.2139/ssrn.2868451.

Choi, L. *et al.* (2005) 'Estimating treatment efficacy over time: a logistic regression model for binary longitudinal outcomes', *Statistics in Medicine*, 24(18), pp. 2789–2805. Available at: https://doi.org/10.1002/sim.2147.

Deery, M. (2008) 'Talent management, work-life balance and retention strategies', *International Journal of Contemporary Hospitality Management*. Edited by N. D'Annunzio-Green, G. Maxwell, and S. Watson, 20(7), pp. 792–806. Available at: https://doi.org/10.1108/09596110810897619.

Deng, D. (2020) *DBSCAN Clustering Algorithm Based on Density | IEEE Conference Publication | IEEE Xplore*. Available at: https://ieeexplore.ieee.org/abstract/document/9356727?casa_token=mARrhyAR_EQAAAAA:EvquD4ZRTVuu1-K3Ww99qMOSpu03DwVLHuusmuty385ri7Vh8MjMoNI-ZvY4425UflwTm6SO4Dg (Accessed: 16 January 2025).

Dharmadhikari, S. (2013) 'Cost Optimization Through "Internal Talent Retention Strategies": An Analytical Study', *CLEAR International Journal of Research in Commerce & Management*, 4(12), pp. 37–41.

Farkiya, R. (2014) 'A Study on Overview of Employee Attrition Rate in India'. Available at: https://d1wqtxts1xzle7.cloudfront.net/57144859/A_Study_on_Overview_of_Employee_Attrition_Rate_in_India_150759937-libre.pdf?1533623385=&response-content-disposition=inline%3B+filename%3DA_Study_on_Overview_of_Employee_Attritio.pdf&Expires=1732761367&Signature=Z3aFuD3zfah4Svm9yxkhUrBdTpCagnNs7LZDsJ8458SXftG0i826KE7tjnpUNPSuCHYKuS~0DkzZYNrtmM6HQso3cQXawiLTPwz~hZtIDtoFR9imsvmt5doGcVDskTTS4P35rd7SikCoqzU0mzGefW34Fg-Z8zObJ9siHFkB2V3bKCkyYuxSkccjMeCuXqfxRguIuwgY5Bx8oBSaJYNj7Pyy61T1-NAhfWcw23paG7y-5p0EMqWp635vJ-9LBUKdlYz0n6ROWhIYExigGjx75sVfphFb~idPRyd8lcUqvpS7hgKfCqNKKI7-gIqGxTjnPcIFp1lWuGU8fvB0Vozdcg__&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA.

Frye, A. *et al.* (2018) 'Employee Attrition: What Makes an Employee Quit?', *SMU Data Science Review*, 1(1). Available at: https://scholar.smu.edu/datasciencereview/vol1/iss1/9.

Goswami, B.K. and Jha, S. (2012) 'Attrition Issues and Retention Challenges of Employees'. Available at: https://d1wqtxts1xzle7.cloudfront.net/32561604/researchpaper-Attrition-Issues-and-Retention-Challenges-of-Employees-libre.pdf?1391123610=&response-content-disposition=inline%3B+filename%3DAttrition_Issues_and_Retention_Challenge.pdf&Expires=1733181297&Signature=XhT4tN8YCQ8yox9VFzuUe229RFtr07f07QAkvkFr8-ccEJ9MGzrdya8eCct7c20ZTDy6YBDdT1h5QzizPASJ-

ds4Zdq7m7VkqgfNZsLBJMYEz991XoyIm8-Cc0kzFwXBfUn2Fk3MkC0su9GUW6g-fxRM~Fkrz751MUlZQNTDmTgVv0X-Vag3hSZvgGjWY1JgaaM-o-ed2u947bt19VMJd-64BF7VIATIcRk64OVqelm2FAwtCY32EZ8uBF9AWp9T5YGT20HBRXiRIRd~zXndIz8K-I3axuERV0IyWYHRDwrREnZ8m0jg4G3nPDempB3EKwkR4YMqVjwvSeYGYh~nEg__&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA.

Hancock, B. and Schaninger, B. (2022) *Talent at a turning point: How people analytics can help*, *McKinsey & Company*. Available at: https://www.mckinsey.com/capabilities/people-and-organizational-performance/our-insights/talent-at-a-turning-point-how-people-analytics-can-help#/ (Accessed: 26 November 2024).

Jorgensen, B. (2005) 'Attract, retain and innovate: a workforce policy architecture adapted to modern conditions', *Foresight*, 7(5), pp. 21–31. Available at: https://doi.org/10.1108/14636680510622163.

Komorowski, M. *et al.* (2016) 'Exploratory Data Analysis', in MIT Critical Data (ed.) *Secondary Analysis of Electronic Health Records*. Cham: Springer International Publishing, pp. 185–203. Available at: https://doi.org/10.1007/978-3-319-43742-2_15.

Krishnaiah, V., Narsimha, D.G. and Chandra, D.N.S. (2014) 'Survey of Classification Techniques in Data Mining', 2. Available at: https://d1wqtxts1xzle7.cloudfront.net/80766836/IJCSE-libre.pdf?1644822348=&response-content-disposition=inline%3B+filename%3DSurvey_of_Classification_Techniques_in_D.pdf&Expires=1737014368&Signature=K8qXVtSuHm-zxbCg0Bwi1IvuuxM23JE3OM78lzBPTL5dwgeVCWF90XXy-cShG0NBEKj5dVH1D0E5cVeEwYbZiYEAJoH-AHGG~0UK8vXGchXnsUidy3fEKBph2fkw2Z2QtxwjEC1~dxzJYtAv64bFfNia04sLLOHr1NC5abNmhpSKWAVbAEo4~8YJf5eVfbLJHgc3A0YmF5PyHjrWj3k~d8XbO-3~CHjWToXCdutJxcXeuId-BBltZ6dZTlcbzLVFhA8LmwYaOFuQQaazVei~MRzYrGf1sk7qSSkwjF3UAtcCmJfAnijn0KP2ulZBABktttUmI~z69em-K13Fx4IHDQ__&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA.

Latha, D.K.L. (2013) 'A STUDY ON EMPLOYEE ATTRITION AND RETENTION IN MANUFACTURING INDUSTRIES'. Available at: https://d1wqtxts1xzle7.cloudfront.net/47852658/09-libre.pdf?1470552222=&response-content-disposition=inline%3B+filename%3DA_STUDY_ON_EMPLOYEE_ATTRITION_AND_RETENT.pdf&Expires=1732760159&Signature=ZZRbln4U1mYdOi-TtlzMzIMQtY42qQxHxpWY6yrLYc-dxL~kTmOKcj~t4UpjxLHMUYnvI8c56~PD3vhWIpKDqW5gQC72iOXL3M~vzes8K7wCumbtstOXWIgK2Zb9zXU0cZ9FKrkg5mrjEraBunY5TrRd7yJ0JkUcjf~J1a2~zi9wsZcVQBv2hFxsXD9FfDiGcA2LmrBvMjiDfemcOTsMaIRN9FqG8wzg4E7WhAhzfjssLkqgHH6P2Q7WvO6LiXezxRfEBJfyvXyJjOBXjg7CvZNJJR1NnZFM-5trLXhX9Q~PUhZmxPh4xHIPEGZa9pH0L0aAz4PS4ttRlEcUVWAppg__&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA.

Na, S., Xumin, L. and Yong, G. (2010) 'Research on k-means Clustering Algorithm: An Improved k-means Clustering Algorithm', in *2010 Third International Symposium on*

Intelligent Information Technology and Security Informatics. *2010 Third International Symposium on Intelligent Information Technology and Security Informatics*, pp. 63–67. Available at: https://doi.org/10.1109/IITSI.2010.74.

Raza, A. *et al.* (2022) 'Predicting Employee Attrition Using Machine Learning Approaches', *Applied Sciences*, 12(13), p. 6424. Available at: https://doi.org/10.3390/app12136424.

Shi, Z. (Michael), Lee, G.M. and Whinston, A.B. (2016) 'Toward a Better Measure of Business Proximity: Topic Modeling for Industry Intelligence', *MIS Quarterly*, 40(4), pp. 1035–1056.

Steinfeld, J. (2024) *Employee Attrition: Types, Causes & Trends for Startups*, *Carta*. Available at: https://carta.com/uk/en/learn/startups/compensation/employee-attrition/ (Accessed: 16 January 2025).

Uyanık, G.K. and Güler, N. (2013) 'A Study on Multiple Linear Regression Analysis', *Procedia - Social and Behavioral Sciences*, 106, pp. 234–240. Available at: https://doi.org/10.1016/j.sbspro.2013.12.027.

Wu, Y. and Zhang, A. (2004) 'Feature selection for classifying high-dimensional numerical data', in *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, p. II–II. Available at: https://doi.org/10.1109/CVPR.2004.1315171.