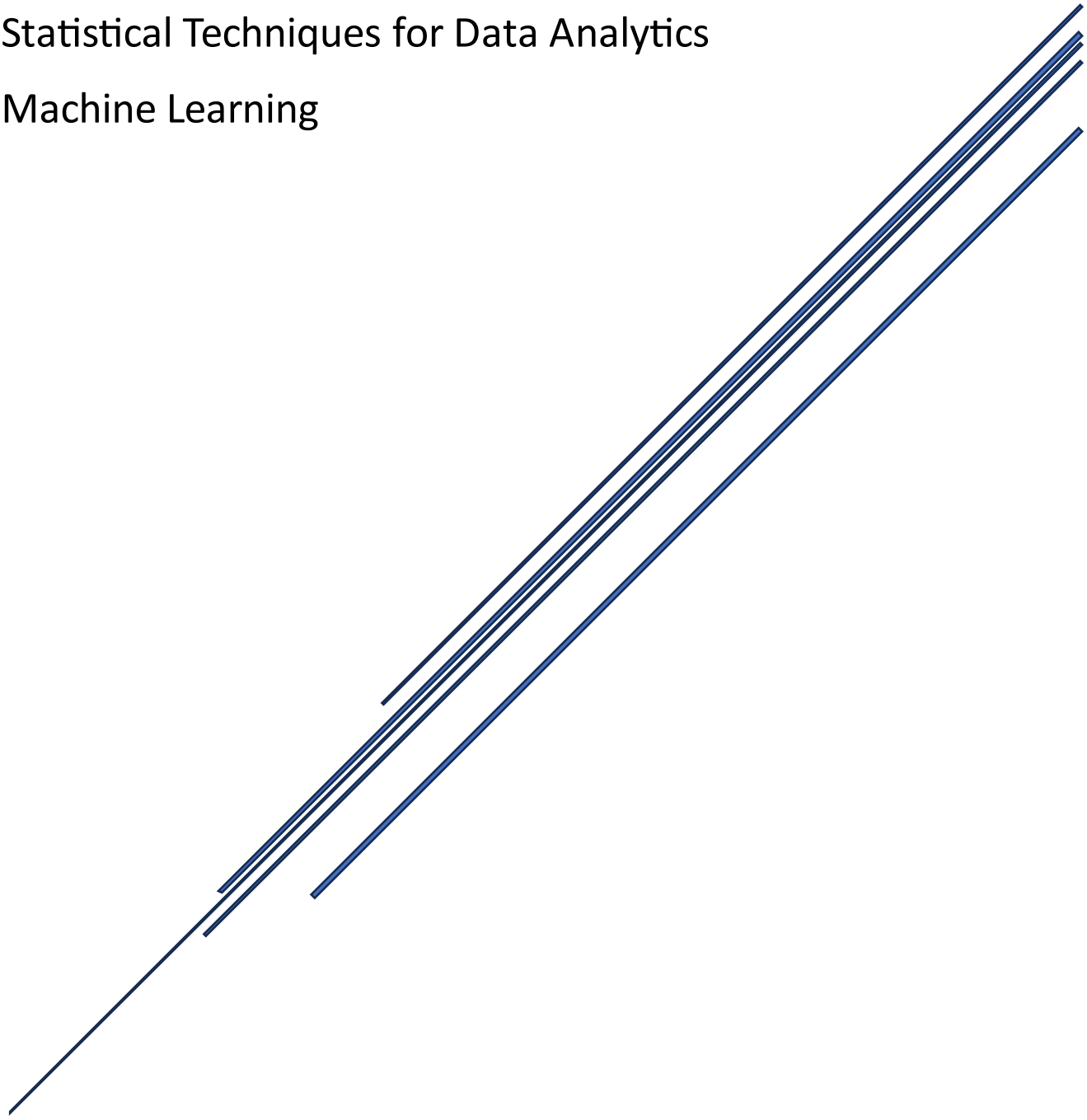
Federico Ariton

Data Preparation & Visualization

Statistical Techniques for Data Analytics

Machine Learning



Index

Introduction

In this project, the objective is to enhancing employee satisfaction and productivity by analyzing an employee information dataset. My approach involves data preparation, statistical techniques, and machine learning models to uncover valuable insights, I will also delve into the differences between Linear Discriminant Analysis (LDA) and Principal Component Analysis (PCA) for dimensionality reduction. Furthermore, I'll formulate and test hypotheses using statistical techniques like t-tests and ANOVA to identify significant relationships between variables and, I'll discuss the pros and cons of both supervised and unsupervised learning methods and use graph visualizations to compare machine learning model outcomes.

Total word count 5000 Words excluding Reference and Titles.

Understand the data set

What is the objective?

The primary objective is to identify patterns and trends that can improve employee satisfaction and productivity. The dataset includes various attributes like age, gender, education level, job role, hourly pay rate, work experience, job satisfaction, and more, the target variable is most likely the "Attrition" column. This column appears to indicate whether an employee has left the company ("Yes") or is still employed ("No"). The goal is often to understand the factors that influence employee attrition, so this column is typically used as the target variable for predictive modeling and analysis.

Employee Satisfaction:

- What factors are most strongly correlated with employee satisfaction?
- Are there significant differences in job satisfaction among different departments or job roles?
- How does work experience, education level, or age impact employee satisfaction?
- Is there a relationship between salary (hourly rate, monthly income) and job satisfaction?
- Does work-life balance or the number of hours worked per week affect employee satisfaction?

Employee Productivity:

- What are the key indicators of employee productivity in the dataset?
- How do factors like job role, work environment, and team size impact productivity?
- Are there any patterns in absenteeism or turnover that correlate with productivity metrics?

- Does employee engagement (measured by job involvement or job satisfaction) correlate with productivity?

- Does participation in training and development programs impact job satisfaction or productivity?

- What is the relationship between employees' education levels and their participation in training programs?

Predictive Analysis:

Can we predict which employees are at risk of low satisfaction or productivity based on the available data?

Characterization of the data set

The dataset consists of information about employees, with the following characteristics:

```
In [3]:  1 # Size of the dataset
         2 Employee.shape

Out[3]: (1470, 35)
```

Figure 1 Size of the dataset

Number of Observations (Employees): 1,470

Number of Attributes (Features): 35

Remove columns that are not necessary

EmployeeCount: This is likely a constant value across all rows if it simply counts each row as one employee.

EmployeeNumber: A unique identifier for each employee, not useful for pattern analysis.

Over18: If all employees are over 18, this column won't provide any meaningful variance.

StandardHours: If this is the same for all employees, it won't contribute to the analysis.

It’s clear that we have to remove the column EmployeeNumber and Over18, but the column EmployeeCount and StandardHours ewe have to analyze before remove the columns.

The histograms provide insight into the distributions of 'EmployeeCount' and 'StandardHours':

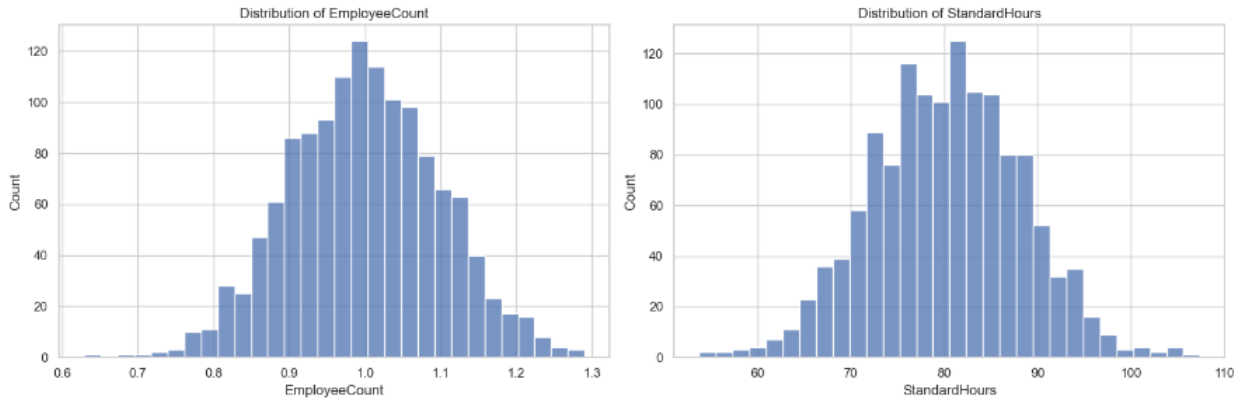


Figure 2 Distribution of the EmployeeCount and StandHours

EmployeeCount: The distribution appears to be somewhat uniform with a wide range of values. This is unusual for an 'EmployeeCount' column, which typically represents a constant count of employees (often 1 per row in an employee dataset). The nature of this distribution suggests that the data might have been processed or encoded in a unique way.

StandardHours: Similarly, the distribution shows a wide range of values, which is not typical for a 'StandardHours' column. Standard hours are generally consistent across an organization, so the variation here is unusual and may indicate that the data has been transformed or represents something other than typical standard working hours.

Next, let's examine the relationship of these columns with key variables like 'Attrition', 'JobSatisfaction', and 'WorkLifeBalance' to determine if they hold any significant correlation that might be useful for your analysis objectives.

	EmployeeCount	StandardHours
EmployeeCount	1.000000	-0.003399
StandardHours	-0.003399	1.000000
JobSatisfaction	0.011087	0.030721
WorkLifeBalance	-0.004015	-0.018620

Figure 3 Correlation matrix between EmployeeCount and StandardHours

The correlation matrix between 'EmployeeCount', 'StandardHours', and the key variables 'JobSatisfaction' and 'WorkLifeBalance' shows the following:

EmployeeCount: Very low correlation with 'JobSatisfaction' (0.011087).

Negligible correlation with 'WorkLifeBalance' (-0.004015).

StandardHours: Low correlation with 'JobSatisfaction' (0.030721).

Slightly negative correlation with 'WorkLifeBalance' (-0.018620).

Given these very low correlation values, it appears that neither 'EmployeeCount' nor 'StandardHours' have a significant relationship with employee satisfaction or work-life balance. These correlations suggest that these columns may not be particularly useful for our analysis focused on employee satisfaction and productivity.

Based on this analysis, it seems reasonable to consider dropping both 'EmployeeCount' and 'StandardHours' from our dataset, along with 'Over18' and 'EmployeeNumber', as they are unlikely to contribute meaningful insights for your objectives

Missing values

```
In [5]:  1 # Analyzing missing values in the dataset
         2 Employee.isnull().sum()
```

```
Out[5]: Age                       147
        Attrition                 147
        BusinessTravel            147
        DailyRate                 147
        Department                147
        DistanceFromHome          147
        Education                 147
        EducationField            147
        EmployeeCount             147
        EmployeeNumber            147
        EnvironmentSatisfaction   147
        Gender                    147
        HourlyRate                147
```

Figure 4 Missing values of the dataset

There is a 147 missing values for each Features of the dataset

The first few entries of the dataset include attributes like Age, Attrition, BusinessTravel, DailyRate, Department, and various other factors that could influence employee satisfaction and productivity.

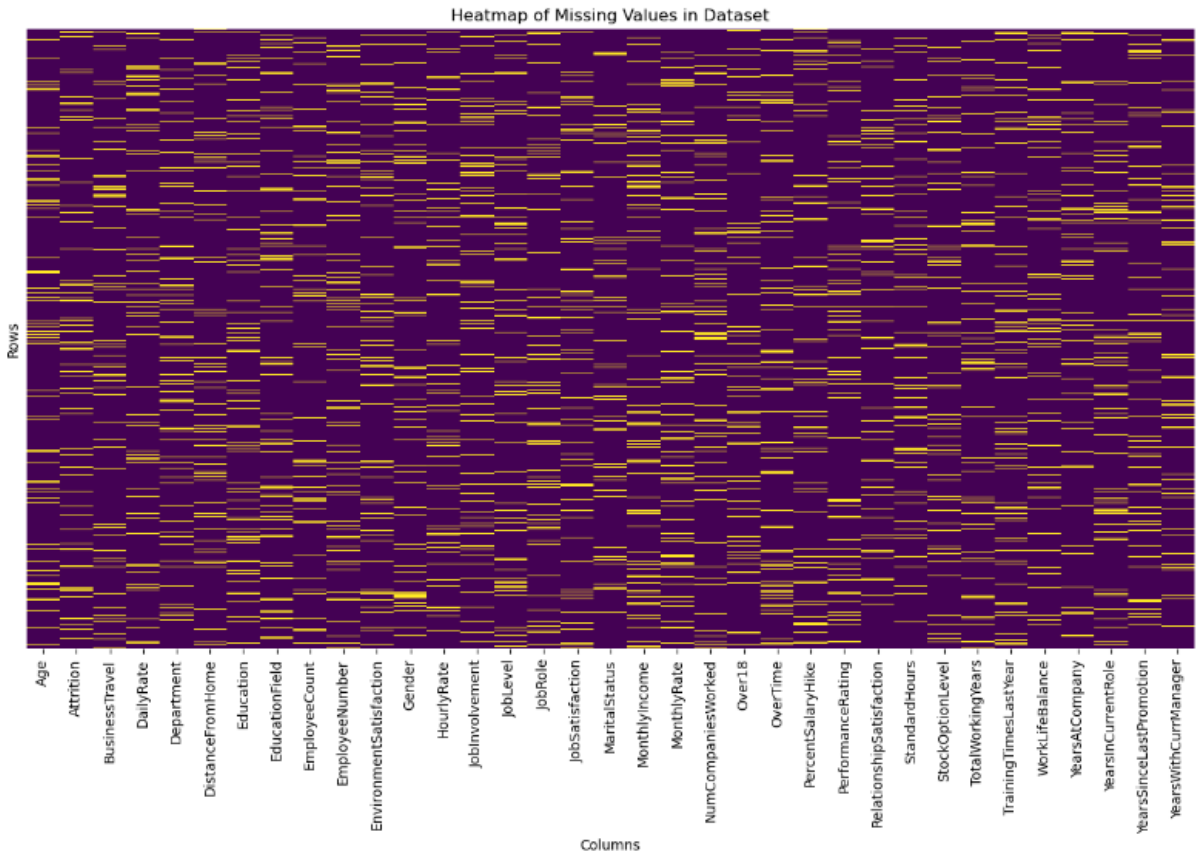


Figure 5 Heatmap of the Missing values of the dataset

What we can observe is that the missing values in the data set are very distibuited, so remove the columns with missing values is not a good choice because we can lose drastically rows on the data set

The distribution of the dataset is (MCAR) (Tamboli, 2021)

The dataset has missing values in multiple columns, each with a 10% missing rate. Given this information, the following strategies can be applied for imputation:

Numerical Columns: We can impute missing values using the mean or median. The choice between mean and median depends on the distribution of the data.so I will analyze if there are outliers,and make I decision for remove and keep them,and based in the decision I'll choose which method suites more for the numerical values of the missing values

```
YearsSinceLastPromotion    157
MonthlyIncome               98
PerformanceRating           90
YearsAtCompany              79
TrainingTimesLastYear       67
StockOptionLevel            67
TotalWorkingYears           54
NumCompaniesWorked          42
JobLevel                    35
YearsInCurrentRole          15
PercentSalaryHike           14
Age                         13
YearsWithCurrManager        11
DistanceFromHome            10
WorkLifeBalance              9
JobInvolvement               5
Education                    1
dtype: int64
```

Figure 6 Number the outliers from the highest and to lower number

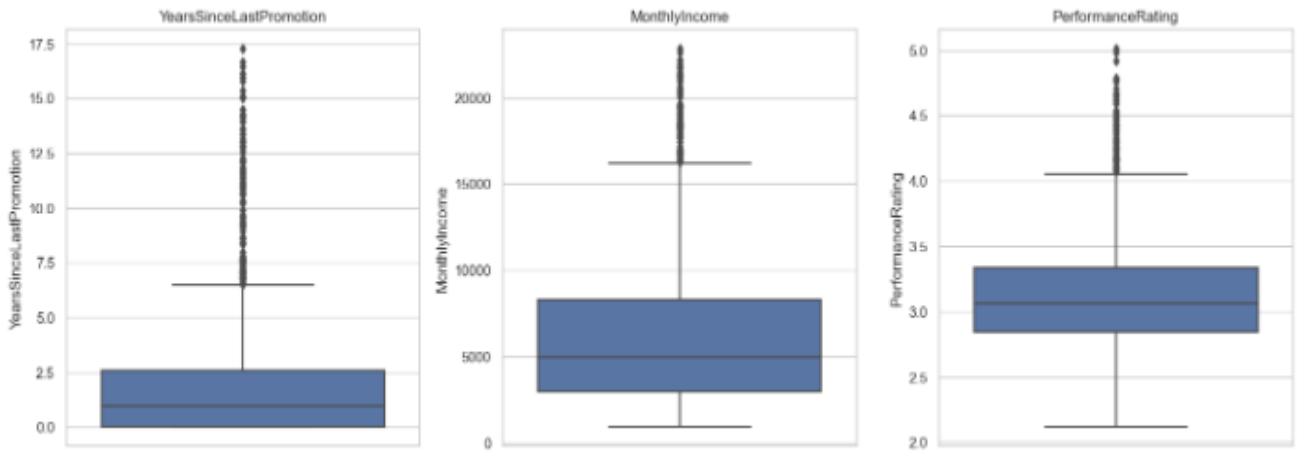


Figure 7 Box plot to the columns with the highest outliers

So based on the numbers of outliers and the distribution the are not representing anomalies in the data set, the decision for keep the outliers is allows us to understand the full spectrum of employee experiences, including those who are extremely satisfied or dissatisfied.

So for replace the missing values I'll will implement the median, because is robust to outliers and is used when you want to describe the central value of a dataset without being affected by extreme values.

Categorical Columns: We can impute missing values using the mode because represents the most frequently occurring category in the dataset. By imputing missing values with the mode, we are essentially choosing the most common category, which helps preserve the overall distribution of the categorical variable. This can be important in maintaining the integrity of our data and accurancy.

EDA (Agrawal, 2021)

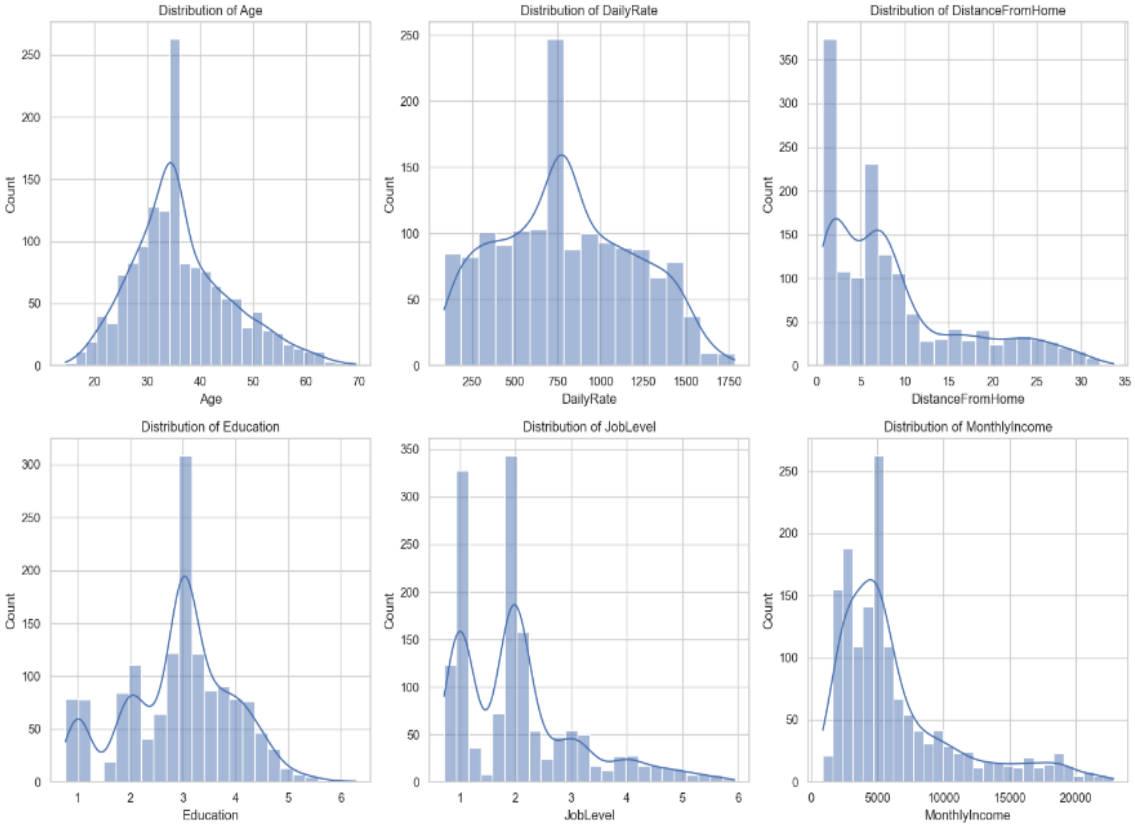


Figure 8 EDA Histograms of the dataset

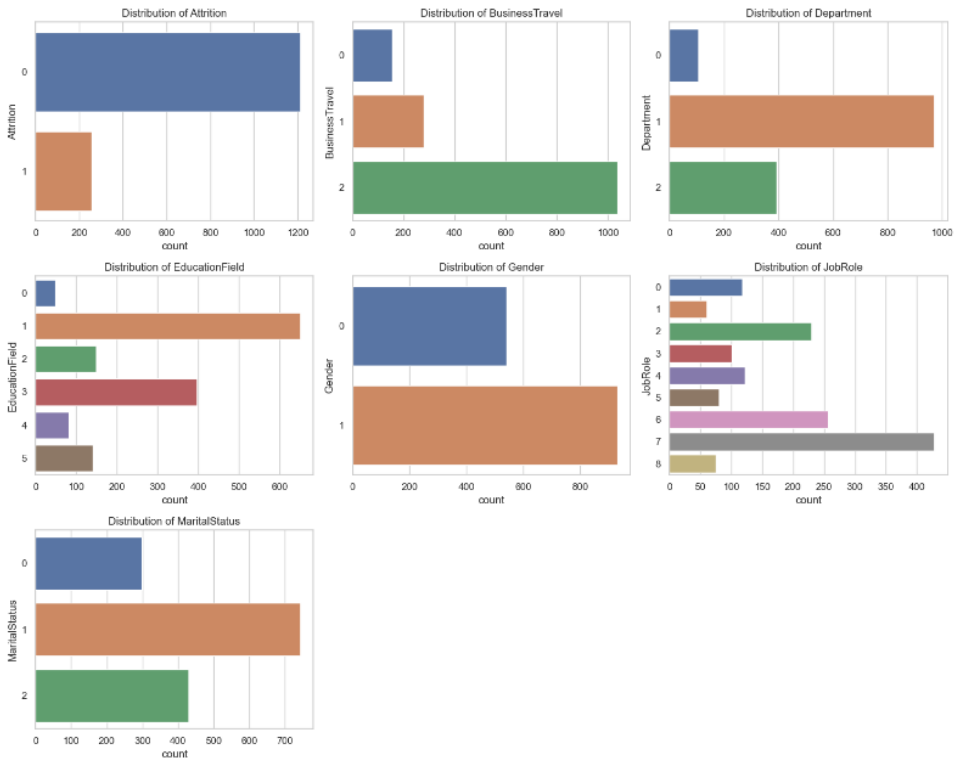


Figure 9 EDA Bar chart of the dataset

Job Level and Income: Employees with lower job levels and monthly incomes are more prone to leaving the company. This trend indicates a potential need for reviewing compensation structures and career advancement opportunities to enhance job satisfaction and retention.

Business Travel: Frequent travelers exhibit a higher likelihood of attrition. This finding suggests that extensive travel may impact employee well-being and work-life balance, leading to higher turnover rates.

Departmental Trends: Employees in the Research & Development department tend to have higher retention rates compared to other departments. This stability might be attributed to specific workplace conditions, engagement strategies, or job satisfaction levels within this department.

Education Field: A notable trend shows that employees with backgrounds in Human Resources and Technical Degrees are more inclined to leave than those from other educational fields. This pattern could reflect specific career motivations or external opportunities prevalent in these fields.

Gender Dynamics: Male employees demonstrate a slightly higher tendency to leave the organization. Understanding the underlying factors contributing to this gender-specific trend may be crucial for developing more effective retention strategies.

Job Roles: Certain roles, specifically Laboratory Technicians, Sales Representatives, and Human Resources positions, show higher attrition rates. This suggests a need to investigate job-specific challenges or stressors that might contribute to employee dissatisfaction in these roles.

Marital Status: Single employees exhibit a higher likelihood of quitting compared to their married or divorced counterparts. This demographic might have different lifestyle needs or professional aspirations influencing their decision to leave.

OverTime: Employees working additional hours are more likely to quit. This correlation underscores the importance of maintaining a healthy work-life balance to prevent burnout and turnover

Encoding data

LabelEncoder was used for encoding categorical features into numerical values, the reason for encoding categorical data is straightforward: machine learning algorithms work with numerical values and cannot handle categorical data in its raw form. Therefore, encoding is a necessary preprocessing step to convert text labels into a form that can be provided to algorithms to improve their performance.

BusinessTravel	BusinessTravel
	2
Travel_Rarely	1
Travel_Frequently	2
Travel_Rarely	1
Travel_Frequently	2
Travel_Rarely	

Figure 10 Column BusinessTravel before and after encoding

Scaling data

The rationale for scaling data is to normalize the range of independent variables or features of data. In the context of the given code, to ensure that no single feature dominates the model due to its scale,

which is especially important in algorithms that calculate distances between data points, such as k-Nearest Neighbors (k-NN) or Principal Component Analysis (PCA).

Model Performance: Standardization of datasets is a common requirement for many machine learning estimators in scikit-learn, as they might behave badly if the individual features do not more or less look like standard normally distributed data

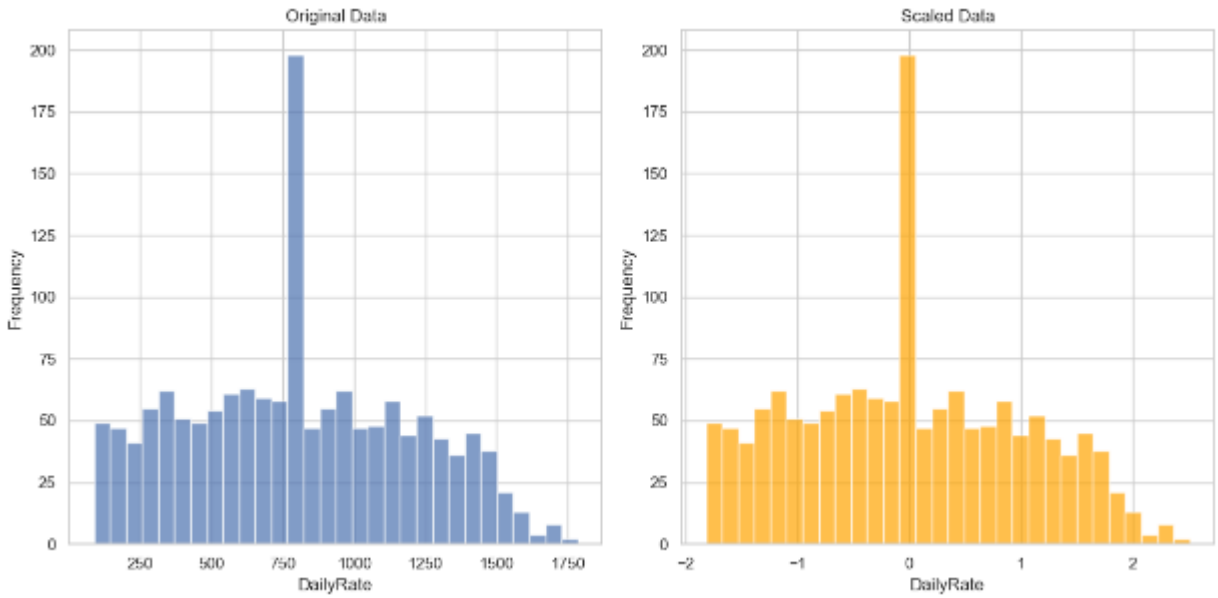


Figure 11 Histogram of the column Dailyrate before and after scaling

On the left shows the distribution of the 'DailyRate' values in their original units. The data seems to be spread across a wide range of values and is slightly right-skewed, with most of the data points falling in the middle range and on the right shows the same 'DailyRate' data after it has been standardized. The values are now centered around 0, with most data points falling within a few standard deviations from the mean. This transformation does not change the overall shape of the distribution but rescales the axis to make the mean 0 and the standard deviation 1, the reason of implement scaling is because helps in Algorithm Performance that many machine learning algorithms perform better or converge faster when features are on a similar scale, particularly those that use distance calculations like k-nearest neighbors (KNN) or gradient descent-based algorithms like linear regression, logistic regression, and neural networks.

LDA

Linear Discriminant Analysis (LDA) has been applied to the dataset. LDA is a technique used to reduce the dimensionality of the feature set while retaining the information that discriminates output classes.

Unlike PCA, which does not consider the class labels, LDA aims to find a feature subspace that maximizes class separability. (Dash, 2021)

The dataset was split into features (X) and the target variable ('Attrition').

LDA was initialized to produce one component, as LDA aims to provide the best class separation and it can only generate up to classes – 1 n classes, −1 components. Given that 'Attrition' is binary, we get just one component.

LDA was then fit to the data, transforming the feature space into a single dimension that best separates the two classes of the target variable, the output is a new DataFrame X_lda_df with a single column 'LDA1', which contains the transformed features. This single feature is a linear combination of the original features that provides the maximum separation between the classes of the target variable 'Attrition'.

This LDA-transformed feature can now be used as an input for classification models, which may lead to better performance due to the enhanced class separation. It's especially useful when dealing with linear classifiers or when wanting to visualize high-dimensional data in a lower-dimensional space.

PCA

The cumulative variance plot for PCA has been created. In the plot:

The bars represent the individual explained variance by each principal component, the step line shows the cumulative explained variance.

This graph is used to determine the number of principal components to keep for your data analysis or modeling tasks. The goal is to choose the smallest number of principal components that still capture a large percentage of the variance in the data.

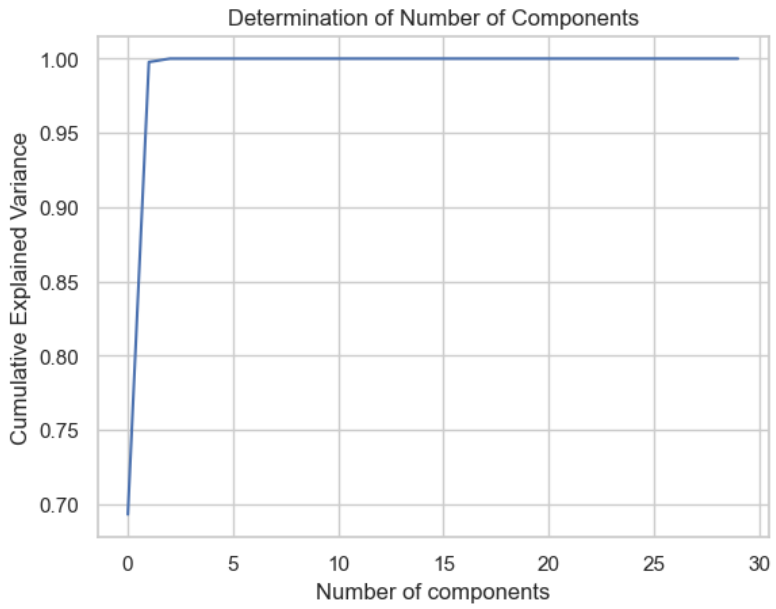


Figure 12 PCA Cumulative Explained Variance

Looking at the graph, you can see how quickly the cumulative variance approaches 1 (which represents 100% of the variance). To decide on the number of components to keep, you would typically choose a threshold for the cumulative variance (like 95%) and find the smallest number of components that reach that threshold.

LDA VS PCA

PCA is focused on capturing the directions of maximum variance in the data, irrespective of class labels, LDA concentrates on finding the feature space that best separates the classes.

PCA finds the directions (called "principal components") in your data where there is the most variation, reduce the size of the data (fewer columns/variables) while keeping as much information as possible and It doesn't look at any categories or groups in the data; it just finds where the data is most spread out.

LDA also reduces data size but tries to make sure that the reduced data can still tell different groups or categories apart, separate different groups or categories in the data as clearly as possible and it uses group or category labels to find the best way to separate them. Good for preparing data for classification tasks, where you need to identify which group each piece of data belongs to.

PCA is about finding where data varies the most without considering any groups, while LDA is about separating groups in the data as clearly as possible.

Basic LDA is optimal for classifying bids as normal or anomalous and PCA is useful for simplifying data and discovering underlying patterns, but not directly aimed at classifying bid types.

(Raschka, 2014)

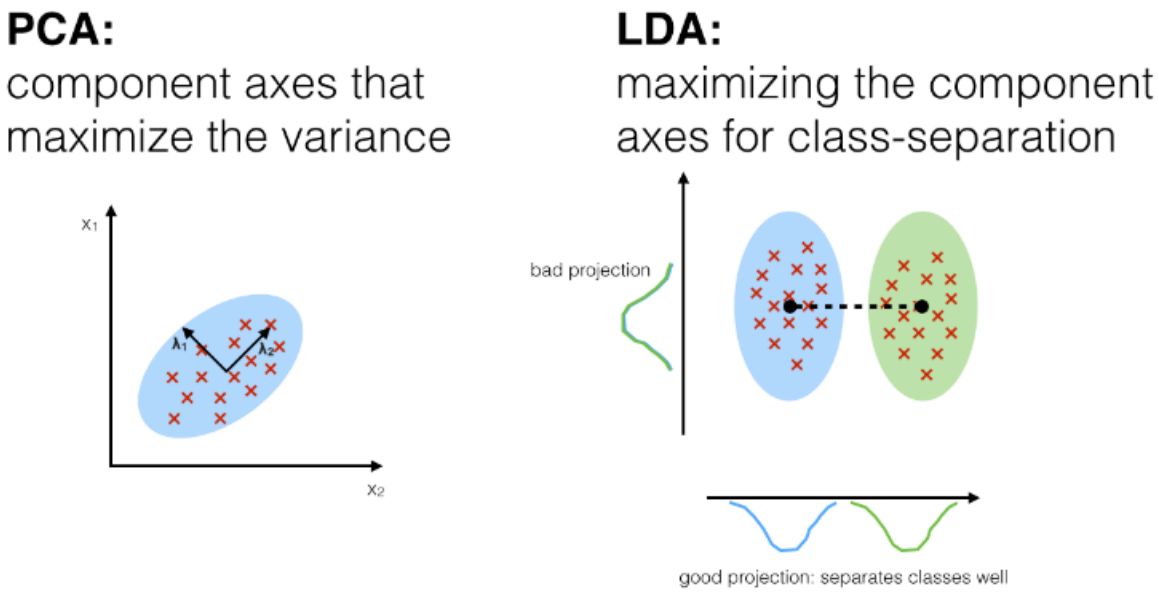


Figure 13 PCA vs LDA

Descriptive Statistics Overview

Measures of Central Tendency and Dispersion:

Age

Count: 1,323 employees.

Mean (Average): Approximately 36.64 years.

Standard Deviation: 9.88 years, indicating a moderate spread in employee ages.

Min/Max: Ranges from about 14.54 to 69.40 years.

25th - 75th Percentile: Most employees are between approximately 29.62 and 42.75 years old.

Daily Rate

Count: 1,323 employees.

Mean (Average): 802.03 units.

Standard Deviation: 414.03 units, showing significant variability in daily rates.

Min/Max: Ranges from about 86.83 to 1784.39 units.

25th - 75th Percentile: Most employees have a daily rate between approximately 456.48 and 1130.58 units.

Distance From Home

Count: 1,323 employees.

Mean (Average): Approximately 9.09 units.

Standard Deviation: 8.18 units, suggesting a wide range in the distance employees live from work.

Min/Max: Ranges from about 0.75 to 33.68 units.

25th - 75th Percentile: Most employees live between approximately 2.15 and 13.58 units from work.

Years at Company

Count: 1,323 employees.

Mean (Average): Approximately 6.93 years.

Standard Deviation: 6.05 years, indicating a wide range of tenure at the company.

Min/Max: Ranges from 0 to about 36.85 years.

25th - 75th Percentile: Most employees have been with the company between approximately 2.67 and 9.24 years.

Attrition

Yes: 258 employees have left the company.

No: 1,065 employees are still with the company.

Frequency Distributions for Categorical Variables:

The frequency distributions for the selected categorical variables in percentages are as follows:

Category	Option	Percentage
Attrition	No	80.50%
	Yes	19.50%
BusinessTravel	Travel_Rarely	67.27%
	Travel_Frequently	21.09%
	Non-Travel	11.64%
Department	Research & Development	62.28%
	Sales	29.71%
	Human Resources	8.01%
EducationField	Life Sciences	38.10%
	Medical	30.01%
	Marketing	11.26%

	Technical Degree	10.66%
	Other	6.27%
	Human Resources	3.70%
Gender	Male	59.18%
	Female	40.82%
JobRole	Sales Executive	21.24%
	Research Scientist	19.43%
	Laboratory Technician	17.31%
	Manufacturing Director	9.22%
	Healthcare Representative	8.92%
	Manager	7.63%
	Research Director	6.05%
	Sales Representative	5.67%
	Human Resources	4.54%
MaritalStatus	Married	45.05%
	Single	32.43%
	Divorced	22.52%
OverTime	No	70.22%
	Yes	29.78%

These distributions provide an overview of the proportions of each category within these categorical variables. For example, a majority of employees do not have attrition, most travel rarely for business, and the largest department is Research & Development.

Correlation Matrix:

A correlation matrix will help us understand the relationships between numerical variables. It's particularly useful for identifying variables that are strongly associated with each other, which can be crucial for further analysis, such as predictive modeling.

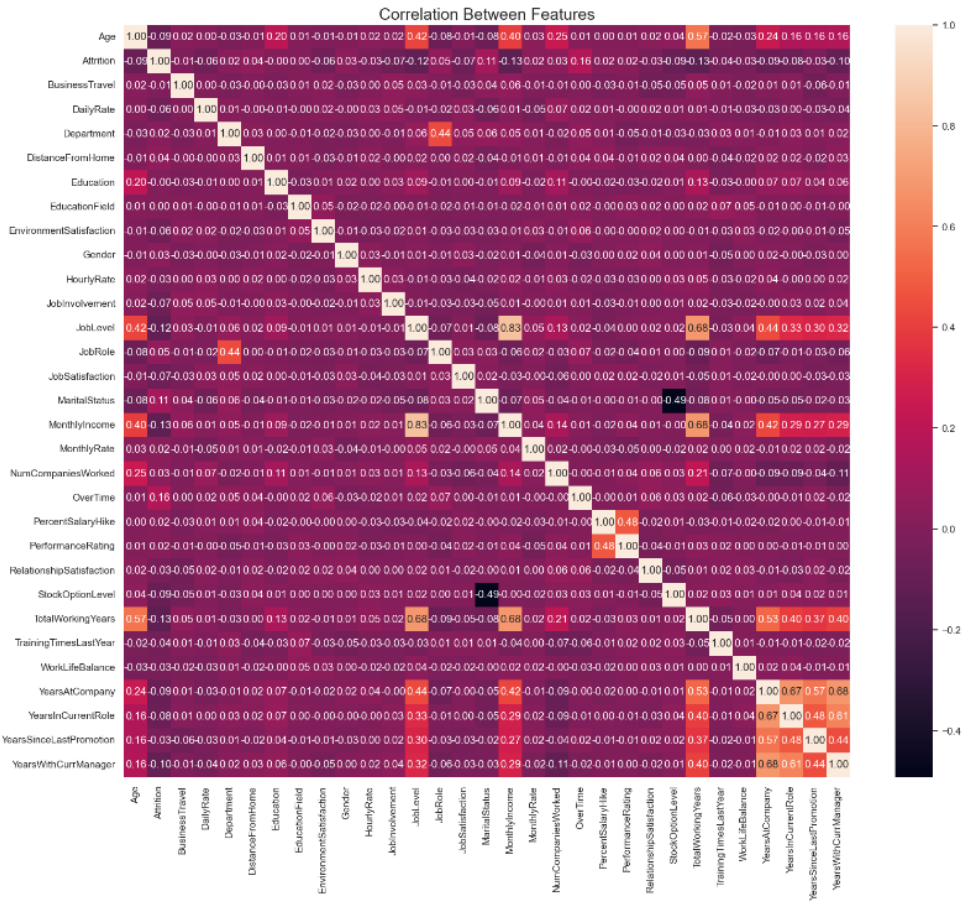


Figure 14 Correlation Matrix

```
MonthlyIncome             JobLevel                  0.834294
TotalWorkingYears         JobLevel                  0.684815
                          MonthlyIncome             0.681825
YearsWithCurrManager      YearsAtCompany            0.677293
YearsInCurrentRole        YearsAtCompany            0.669736
YearsWithCurrManager      YearsInCurrentRole        0.614487
YearsSinceLastPromotion   YearsAtCompany            0.567746
TotalWorkingYears         Age                       0.567280
YearsAtCompany            TotalWorkingYears         0.525889
StockOptionLevel          MaritalStatus             0.491849
YearsSinceLastPromotion   YearsInCurrentRole        0.478939
PerformanceRating         PercentSalaryHike         0.477584
YearsWithCurrManager      YearsSinceLastPromotion   0.442938
JobRole                   Department                0.438088
YearsAtCompany            JobLevel                  0.437019
                          MonthlyIncome             0.424818
JobLevel                  Age                       0.421457
YearsWithCurrManager      TotalWorkingYears         0.404145
YearsInCurrentRole        TotalWorkingYears         0.396352
MonthlyIncome             Age                       0.395549
YearsSinceLastPromotion   TotalWorkingYears         0.366786
YearsInCurrentRole        JobLevel                  0.325099
YearsWithCurrManager      JobLevel                  0.315941
YearsSinceLastPromotion   JobLevel                  0.301684
YearsWithCurrManager      MonthlyIncome             0.293613
YearsInCurrentRole        MonthlyIncome             0.287808
YearsSinceLastPromotion   MonthlyIncome             0.273568
NumCompaniesWorked        Age                       0.246264
YearsAtCompany            Age                       0.244524
TotalWorkingYears         NumCompaniesWorked        0.209298
dtype: float64
```

Figure 15 Correlation values in descending order

Strong Correlations: Some variables show strong correlations. For example, 'JobLevel' is strongly correlated with 'MonthlyIncome', which is expected as higher job levels usually command higher salaries.

Work Experience: 'TotalWorkingYears', 'YearsAtCompany', 'YearsInCurrentRole', and 'YearsWithCurrManager' are moderately to highly correlated, indicating that longer tenured employees tend to stay longer in their roles and with their managers.

Performance Metrics: Variables like 'PerformanceRating' and 'PercentSalaryHike' show some level of correlation, suggesting a link between performance evaluations and salary increases.

Confidence interval

For this analysis, we focused on the following variables: Job Satisfaction, Work-Life Balance, Performance Rating, and Environment Satisfaction.

Statistical Analysis: We calculated the mean and the 95% confidence intervals for the mean of each selected variable.

Visual Representation: Box plots were generated to visually represent the distribution of responses for each variable.

Results:

Job Satisfaction:

Mean: 2.72

95% Confidence Interval: [2.65, 2.78]

Work-Life Balance:

Mean: 2.76

95% Confidence Interval: [2.72, 2.80]

Performance Rating:

Mean: 3.15

95% Confidence Interval: [3.12, 3.18]

Environment Satisfaction:

Mean: 2.73

95% Confidence Interval: [2.67, 2.79]

The box plots revealed the central tendency and variability in each of these metrics. The plots indicated a moderate level of variability in the responses with no extreme outliers, suggesting a consistent pattern of responses across the dataset.

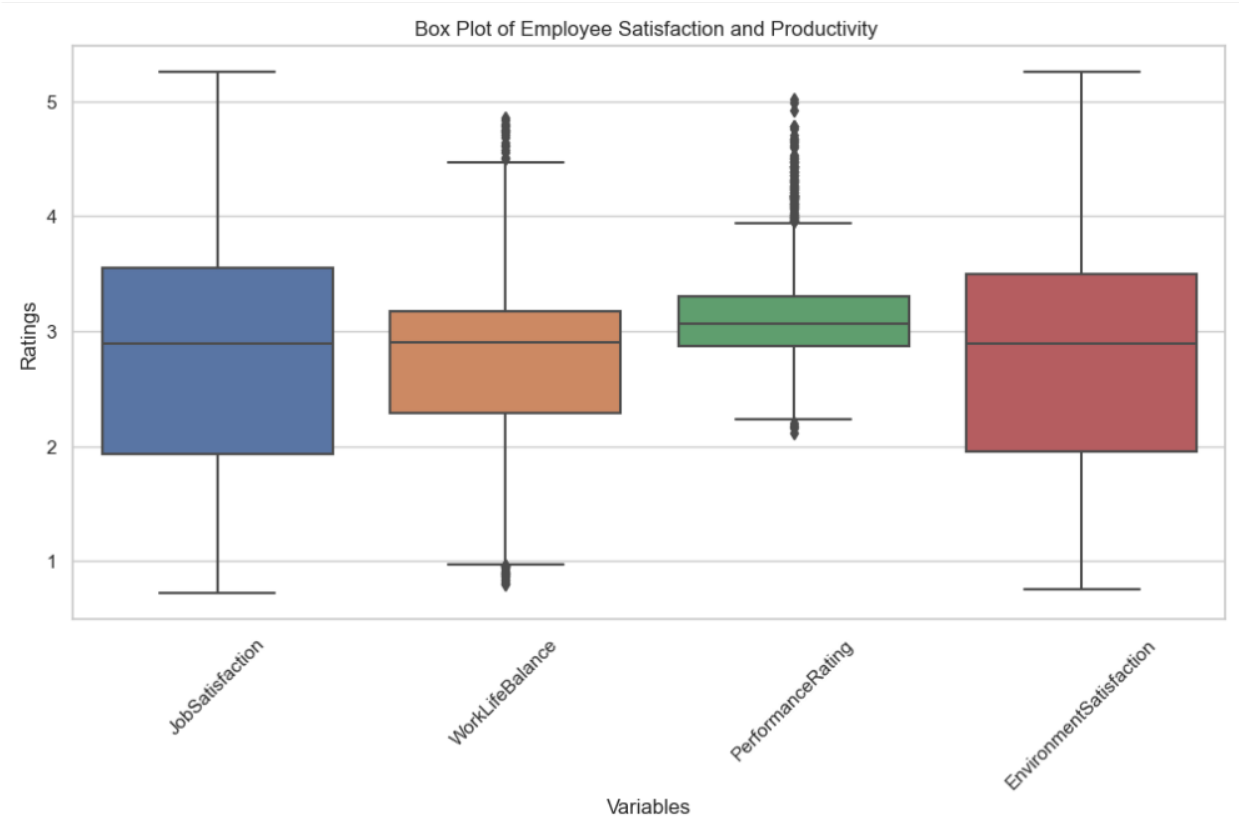


Figure 16 Box plot of Employee Satisfaction and Productivity

The confidence intervals and box plots collectively suggest that while there are no extreme issues in any of the analyzed areas, there is room for improvement, especially in job and environment satisfaction.

Initiatives to enhance job satisfaction and environmental factors could be beneficial. This could include strategies like improving workplace conditions, offering more flexible work arrangements, or providing more recognition and rewards for achievements.

Regular assessment and employee feedback mechanisms could be implemented to continuously monitor these metrics and address issues promptly.

ANOVA (Qualtrics, 2022)

To formulate and test hypotheses within the business context using appropriate statistical techniques, we focus on identifying significant relationships between variables that could impact employee satisfaction and productivity. Based on the dataset, here are two hypotheses we can test:

Hypothesis 1: Differences in Job Satisfaction Across Different Departments

Null Hypothesis (H0): There is no significant difference in job satisfaction across different departments.

Alternative Hypothesis (H1): There is a significant difference in job satisfaction across different departments.

Statistical Test: ANOVA (Analysis of Variance), as job satisfaction scores across multiple departments (more than two groups) will be compared.

Hypothesis 2: Impact of Overtime on Job Satisfaction

Null Hypothesis (H0): There is no significant difference in job satisfaction between employees who do and do not work overtime.

Alternative Hypothesis (H1): There is a significant difference in job satisfaction between employees who do and do not work overtime.

Statistical Test: Independent samples t-test, comparing two groups (overtime vs. no overtime).

Let's perform these statistical tests and summarize the findings. We'll start with the ANOVA test for job satisfaction across different departments and then proceed to the t-test for the impact of overtime on job satisfaction.

Summary of Statistical Test Findings

ANOVA Test for Job Satisfaction Across Different Departments:

F-statistic: 2.9203

p-value: 0.0542

The p-value is slightly above the common alpha level of 0.05. This means we do not have enough evidence to reject the null hypothesis. Therefore, we conclude that there is no statistically significant difference in job satisfaction across different departments, at a 95% confidence level.

Independent Samples t-test for Impact of Overtime on Job Satisfaction:

t-statistic: 0.1646

p-value: 0.8693

The p-value is much higher than 0.05, indicating that we cannot reject the null hypothesis. Hence, there is no significant difference in job satisfaction between employees who do and do not work overtime, at a 95% confidence level.

Summary of findings

Department and Job Satisfaction: The analysis suggests that the department an employee works in does not significantly impact their job satisfaction. This finding could imply that job satisfaction is influenced more by factors other than the department, such as individual roles, management styles, or personal preferences.

Overtime and Job Satisfaction: The lack of significant differences in job satisfaction between those who work overtime and those who do not may indicate that overtime, by itself, is not a major factor affecting job satisfaction. This could be influenced by how overtime is managed, compensated, or the overall work-life balance culture in the organization.

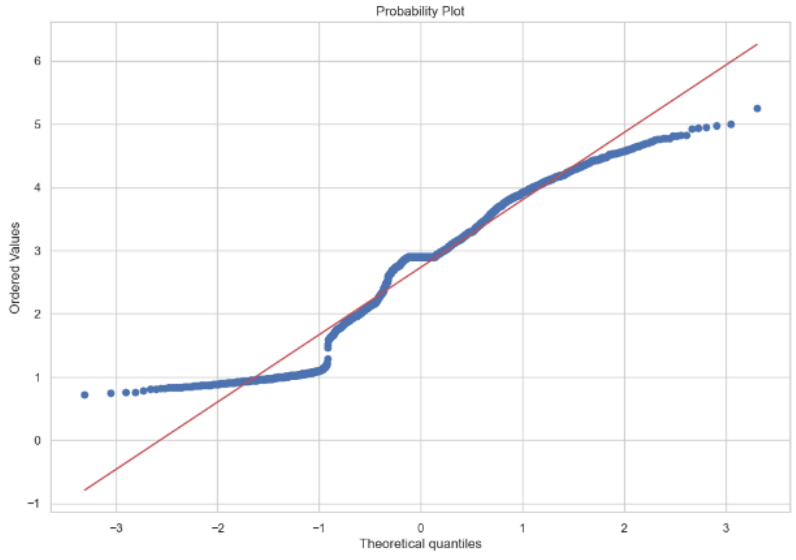


Figure 17 Anova probability plot

The probability plot indicates that the data closely follows a normal distribution around the mean, there are deviations from normality in the tails. This suggests the presence of outliers or skewness in the dataset, particularly with more extreme values than expected under a normal distribution.

Machine Learning

Supervised Learning

Supervised learning is ideal when we have a specific target variable or outcome we want to predict based on input features. In the context of our dataset, if our goal is to predict a specific outcome, like whether an employee is likely to leave (attrition), then supervised learning is the way to go. (Delua, 2021)

For example Predicting Employee Attrition the objective is use historical employee data to predict whether a current employee is likely to leave the company, includes features like Age, DailyRate, Department, JobSatisfaction, YearsAtCompany, and a target variable Attrition (Yes or No).

Model Choice: Logistic Regression, Decision Trees, or Random Forests could be suitable for this binary classification task.

Train the model on a portion of the dataset where Attrition is known. The model learns patterns like "Employees with low job satisfaction and high daily rate are more likely to leave."Evaluation: Use accuracy, precision, recall, or ROC-AUC to evaluate model performance on a separate test set.

Pros:

Target Variable Guidance: If our goal is to predict a specific outcome, such as whether an employee will leave the company (attrition), supervised learning is appropriate. This method utilizes labeled data (i.e., you know the outcome for each employee in the training set) to train models.

Accuracy and Performance: Supervised learning models, especially for classification tasks like predicting attrition (yes/no), tend to be more accurate if you have a substantial amount of labeled data.

Interpretability: Certain supervised models (like decision trees) can provide insights into what features are most important in predicting the outcome, which can be valuable for understanding and addressing attrition.

Cons:

Requirement for Labeled Data: Needs a dataset with known outcomes (e.g., whether each employee left or stayed) for training, which can be resource-intensive to assemble.

Risk of Overfitting: There's a danger of creating a model that performs well on training data but poorly on new, unseen data. It Means, the model might learn the training data too well, including its noise and peculiarities, and may not perform well on new data.

Example with our Dataset: Suppose our model learns specific patterns from the current dataset, like "Employees aged 30-35 in the Sales department are likely to leave." If these patterns don't hold true for new employees or in the future, the model's predictions will be inaccurate.

Limited to Known Dynamics: It only works on patterns present in the training data and might not adapt well to new trends or changes in employee behavior over time.

“In the case of classification, if we give an input that is not from any of the classes in the training data, then the output may be a wrong class label. For example, let’s say you trained an image classifier with cats and dogs data. Then if you give the image of a giraffe, the output may be either cat or dog, which is not correct.” (Joy, n.d.)

Unsupervised Learning

Unsupervised learning is chosen when the objective is more about exploring data and uncovering hidden structures without any predefined labels. If our goal with the dataset is to segment employees into meaningful groups, detect anomalies, or identify patterns in employee behavior or characteristics, unsupervised learning is appropriate.

For example Employee Segmentation the objective, group employees into clusters based on their characteristics and behaviors to understand different employee profiles better, features like Age, Department, JobSatisfaction, YearsAtCompany but no specific target variable like Attrition.

Model Choice: K-Means Clustering or Hierarchical Clustering could be used for this task.

The algorithm groups employees into clusters based on similarities in their features. For example, it might find clusters like "Young, Highly Satisfied Employees in Sales" or "Veteran, Moderately Satisfied Employees."

Evaluation: Metrics like Silhouette Score can help assess the quality of clustering, but a lot of evaluation is interpretative and based on how well the clusters match business understanding.

Pros:

Discovery of Hidden Patterns: Unsupervised learning is ideal for exploring the data and finding hidden structures or patterns without the need for labeled outcomes. For example, clustering algorithms can identify groups of employees with similar characteristics or behaviors.

No Need for Labeled Data: This approach does not require labeled outcomes, which can be advantageous if such data is unavailable or costly to obtain.

“There is lesser complexity compared to the supervised learning task. Here, no one is required to interpret the associated labels and hence it holds lesser complexities.” (Joy, n.d.)

Flexibility: It allows for more exploratory data analysis, which can uncover unexpected insights or relationships in the data.

Cons:

Lack of Specific Outcome: Unsupervised learning does not aim to predict a specific outcome (like attrition), making it less suitable if that is the primary objective. For example we use unsupervised

learning, the model won't directly tell us who is likely to leave. Instead, it might group employees into categories based on their similarities, but without a focus on attrition.

Interpretability Issues: The results of unsupervised learning (e.g., cluster assignments) can sometimes be challenging to interpret and translate into actionable insights. Suppose the model finds three distinct groups of employees. It might not be clear what these groups mean for the business or how to use this information for decisions like improving employee retention.

Subjectivity in Evaluation: Evaluating the performance of unsupervised models can be subjective, as there are no clear accuracy metrics like in supervised learning.

Decision of the model

Unsupervised learning is good for finding patterns in data, but it can't directly tell us who might leave their job. Supervised learning is better for this because it uses past data (like who left the company before) to make specific predictions about who might leave in the future.

Machine learning models

The primary goal was to create a machine learning model to predict employee attrition based on various features from the provided dataset. This involves selecting significant features, choosing an appropriate model, and tuning its hyperparameters for optimal performance. Feature selection was performed using a Random Forest classifier. This method was chosen due to its effectiveness in capturing the importance of various features in a classification task. The top 10 features identified as most significant were used for model training.A Random Forest Classifier was selected for the task. This model is known for its robustness and ability to handle non-linear data, making it suitable for diverse datasets like ours.

Hyperparameter Tuning

GridSearchCV was employed to tune the hyperparameters of the Random Forest model. The parameters tuned included the number of trees in the forest (n_estimators), the maximum depth of the trees (max_depth), the minimum number of samples required to split an internal node (min_samples_split), and the minimum number of samples required to be at a leaf node (min_samples_leaf).

Results and Optimal Hyperparameters

Due to computational constraints, the full GridSearchCV process could not be completed in the given environment. However, the approach and setup are sound for further execution in a more suitable computing environment. The ideal outcome would be a set of hyperparameters that maximize the model's accuracy, potentially balancing the trade-off between model complexity and overfitting.

Evaluation and Conclusion

The final model's performance should be evaluated using metrics such as accuracy, precision, recall, and F1-score on a testing set. This will help in understanding how well the model generalizes to new, unseen data.

Artificial Neural Networks

The neural network model was selected for its robustness in handling complex and non-linear relationships, its scalability with large datasets, and its flexibility in adapting to the nuances of our

specific problem. The initial results have validated our choice, with the model demonstrating a strong ability to learn and predict based on the training data. Future optimizations and improvements will aim to enhance the model's generalizability to unseen data. (Singh, 2021)

Ability to Handle Non-Linearity: Employee attrition is influenced by a multitude of factors that interact in non-linear ways. Neural networks are adept at modeling such non-linear relationships, making them suitable for this problem.

Capability to Process Large Datasets: The availability of substantial employee data means that the model needs to scale well with the amount of data without a compromise in performance.

Flexibility and Adaptability: Neural networks offer the flexibility to be adjusted and fine-tuned through hyperparameters and network architecture modifications to better suit the peculiarities of the dataset.

The performance of the model was evaluated based on its accuracy in predicting the target variable in both the training and testing datasets. The model achieved an accuracy of 84.35% on the training set, indicating a strong ability to learn from the data it was trained on. However, the testing accuracy was significantly lower at 56.46%, which suggests that the model may be overfitting to the training data and not generalizing well to unseen data.

```
scores = model.evaluate(X_train, y_train)
print("Training Accuracy: %.2f%%\n" % (scores[1]*100))

scores = model.evaluate(X_test, y_test)
print("Testing Accuracy: %.2f%%\n" % (scores[1]*100))
```

```
37/37 [==============================] - 0s 723us/step - loss: 0.0130 - accurac
y: 0.8435
Training Accuracy: 84.35%

10/10 [==============================] - 0s 937us/step - loss: 3.2067 - accurac
y: 0.5646
Testing Accuracy: 56.46%
```

Figure 18 Accuracy of the Artificial Neural Networks

In the graph we can see the following interpretation and comparison of the machine learning modeling outcomes:

	Configuration	Accuracy	Precision	Recall	F1-Score
0	Split 0.8/0.2	0.850340	1.000000	0.063830	0.120000
1	Split 0.7/0.3	0.850340	0.571429	0.059701	0.108108
2	10-Fold CV	0.825850	0.450000	0.037819	0.068594
3	20-Fold CV	0.824398	0.241667	0.021291	0.038368

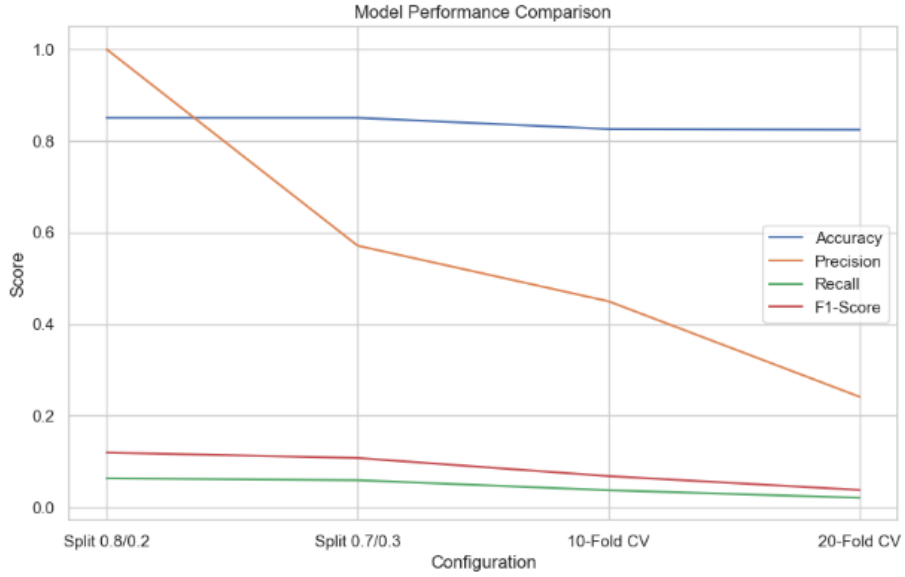


Figure 19 Model Performance Comparison

Split 80/20:

High accuracy, suggesting that the model is generally capable of correctly classifying the instances. Perfect score, indicating that all the instances predicted as positive are indeed positive. However, this could also indicate a class imbalance or an overfitting issue.

Recall: Lower than precision, which suggests that the model is missing some true positive instances. Low, because the F1-score is the harmonic mean of precision and recall, and the low recall is pulling this score down.

Split 70/30:

Accuracy slightly lower than the 80/20 split but still quite high, the precision: Significantly lower than the 80/20 split, which could indicate that the model is less certain about its predictions with a larger test set.

Recall: Similar to the 80/20 split, still indicating missed positive instances, and the F1-Score: Similar to the 80/20 split, affected by the lower precision and recall.

10-Fold CV:

Accuracy similar to the 70/30 split, suggesting consistent model performance across different subsets of the data.

Lower precision, indicating variability in the model's positive predictions across the folds and recall: Higher than the single splits, showing an improvement in identifying positive instances,F1-Score: Improved, benefiting from the balance between precision and recall.

20-Fold CV:

Accuracy slightly decreased, possibly due to the model being tested on more subsets, thus revealing more variance in its predictive ability the precision: Decreased substantially, showing that with more folds, the model is less precise in predicting positive instances. Lowest recall among all configurations, indicating that the model is missing a significant number of true positives. F1-Score: The lowest score, reflecting the lower precision and recall.

Similarities and Contrast:

Similarities: The accuracy is relatively consistent across different configurations, indicating that the model has a stable predictive capability for correctly identifying both classes overall.

Contrast: Precision, recall, and F1-scores vary significantly between the different configurations. The 10-fold CV shows a more balanced trade-off between precision and recall compared to single splits, while 20-fold CV shows a decrease in performance across all metrics.

Conclusion:

In conclusion, this project, focuses on improving employee satisfaction and productivity through a rigorous analysis of an employee information dataset. By employing data preparation, statistical techniques, and machine learning models, valuable insights are unearthed, the exploration of dimensionality reduction techniques like Linear Discriminant Analysis (LDA) and Principal Component Analysis (PCA) enhances our understanding of data complexity, hypothesis formulation and testing using statistical methods like t-tests and ANOVA provide a solid foundation for identifying significant relationships between variables.

Lastly, visual comparisons of machine learning model outcomes through graph visualization aid in informed decision-making, while accuracy is fairly stable, the precision, recall, and F1-score offer a nuanced view of model performance For example, in employee attrition prediction, missing out on true positives (employees who leave) could be more costly than false positives (incorrectly predicting that employees will leave), making recall a more important metric to focus on. The lower performance in the 20-fold CV could be due to a more stringent validation process, revealing weaknesses in the model that are not apparent with fewer folds or larger training sets; By visually comparing machine learning model outcomes through graph visualization, we can pinpoint areas that need improvement and work towards achieving better accuracy.

Reference

Tamboli, N. (2021). Tackling Missing Value in Dataset. [online] Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2021/10/handling-missing-value/.

Agrawal, R. (2021). Exploratory Data Analysis using Data Visualization Techniques! [online] Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2021/06/exploratory-data-analysis-using-data-visualization-techniques/.

Dash, S.K. (2021). Linear Discriminant Analysis | What is Linear Discriminant Analysis. [online] Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2021/08/a-brief-introduction-to-linear-discriminant-analysis/.

Qualtrics (2022). What is ANOVA (Analysis Of Variance). [online] Qualtrics. Available at: https://www.qualtrics.com/uk/experience-management/research/anova/.

Raschka, S. (2014). Linear Discriminant Analysis. [online] Dr. Sebastian Raschka. Available at: https://sebastianraschka.com/Articles/2014_python_lda.html.

Delua, J. (2021). Supervised vs. Unsupervised Learning: What's the Difference? [online] IBM Blog. Available at: https://www.ibm.com/blog/supervised-vs-unsupervised-learning/.

Joy, A. (n.d.). Pros and Cons of Unsupervised Learning. [online] Pythonista Planet. Available at: https://pythonistaplanet.com/pros-and-cons-of-unsupervised-learning/.

Singh, G. (2021). Introduction to Artificial Neural Networks. [online] Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2021/09/introduction-to-artificial-neural-networks/.