**CT College Dublin**

**Assessment Cover Page**

|  |  |
| --- | --- |
| **Module Title:** | Machine Learning |
| **Assessment Title:** | CA 2 Integrated |
| **Lecturer Name:** | Muhammad Iqbal  David McQuaid  Marina Soledad Iantorno |
| **Student Full Name:** | Federico Ariton |
| **Student Number:** | sba22090 |
| **Assessment Due Date:** | 22nd Nov 2023 |
| **Date of Submission:** | 5rd Jan 2024 |

**Declaration**

|  |
| --- |
| By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution. |

# Introduction

In this project, I will analyzing the dataset that containing employee information to improve employee satisfaction and productivity. I will use data preparation, statistical techniques, and machine learning models to identify patterns and trends. The impact of data preprocessing on analysis outcomes will be assessed, and results will be communicated clearly

Explaining the difference between Lda and pca,

Formulate and test hypotheses within a business context using appropriate statistical techniques

like t-tests or ANOVA to identify significant relationships between variables

discuss the pros and cons (supervised/ Unsupervised)

discuss comparison of ML modelling outcomes using a Table or graph visualization.

# 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:

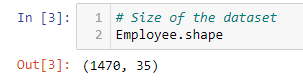
Can we predict which employees are at risk of low satisfaction or productivity based on the available data?

What factors are predictive of future employee turnover?

How we are going to deal with the missing values?

# Characterization of the data set

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



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.

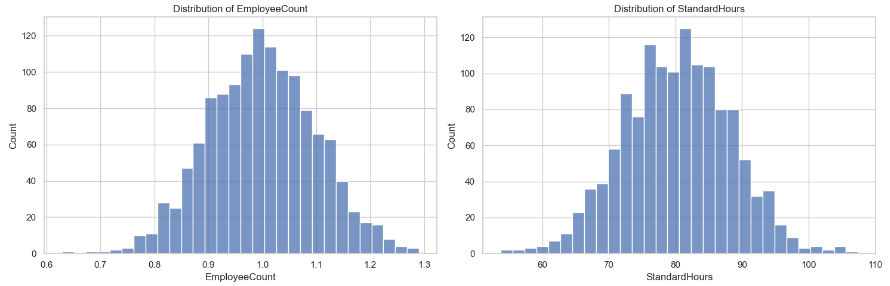
EmployeeNumber: A unique identifier for each employee, not useful for pattern analysis.

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 Over, but the column EmployeeCount and StandardHours ewe have to analayze because

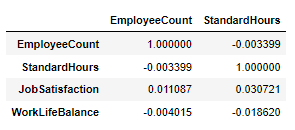
The histograms provide insight into the distributions of 'EmployeeCount' and 'StandardHours':



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. ​​



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:

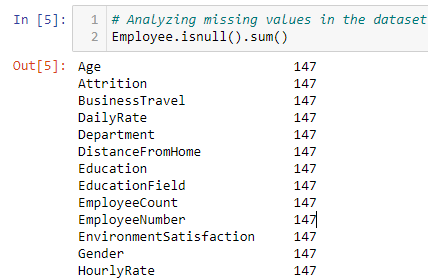
Low correlation with 'JobSatisfaction' (0.030721).

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.

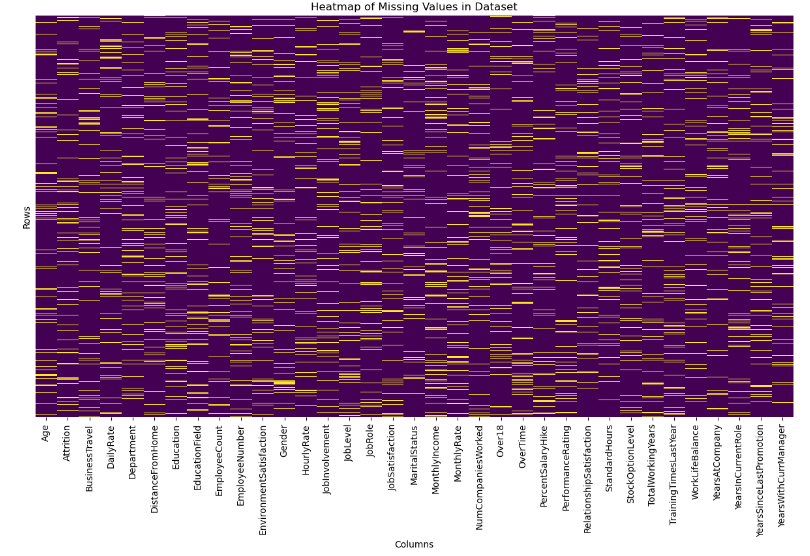
Based on this analysis, it seems reasonable to consider dropping both 'EmployeeCount' and 'StandardHours' from our dataset, along with 'Over18' and 'EmployeeNumber', as they are unlikely to contribute meaningful insights for your objectives

## Missing values



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.

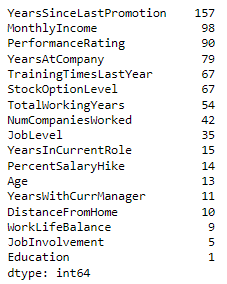


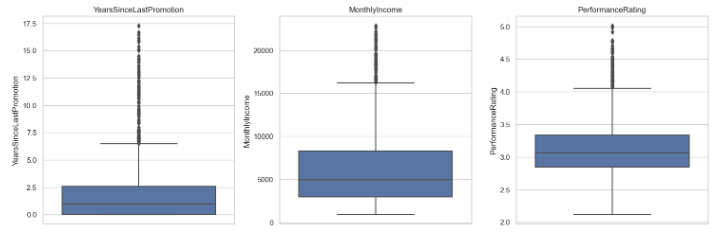
What we can observe is that the missing values in the data set are very distibuited, so remove the columns with missing values is not a good choice because we can lose drastically rows on the data set

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



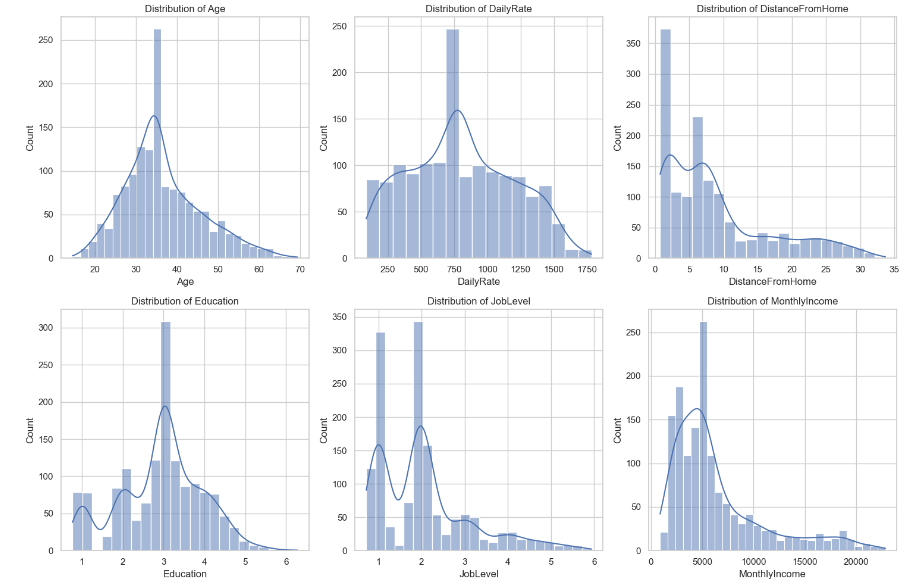


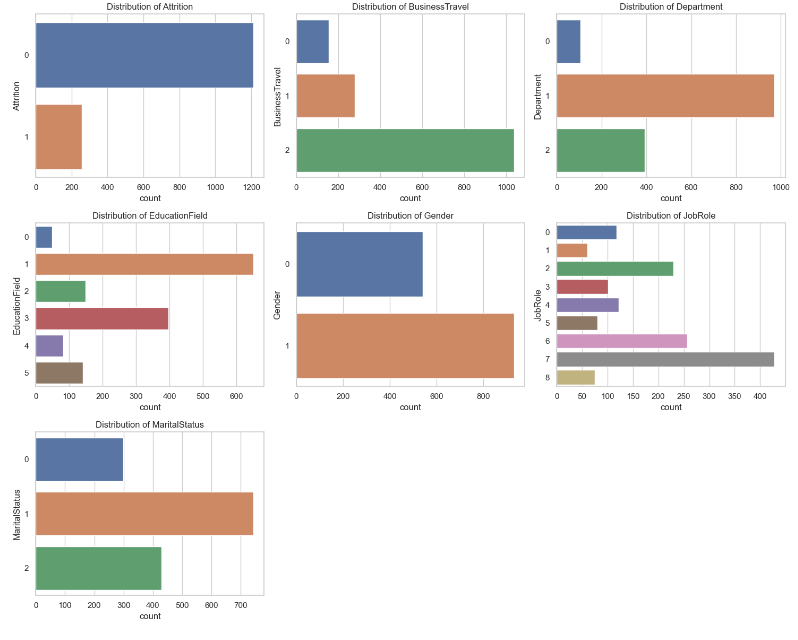
So based on the numbers of outliers and the distribution the are not representing anomalies in the data set, the decision for keep the outliers is allows us to understand the full spectrum of employee experiences, including those who are extremely satisfied or dissatisfied.

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





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.

## 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

# 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.

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.

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

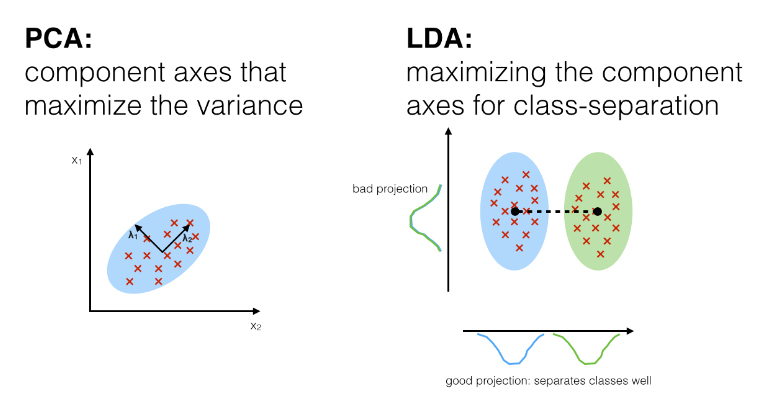
PCA is focused on capturing the directions of maximum variance in the data, irrespective of class labels, LDA concentrates on finding the feature space that best separates the classes.

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.

(Raschka, 2014)



# 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:

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:

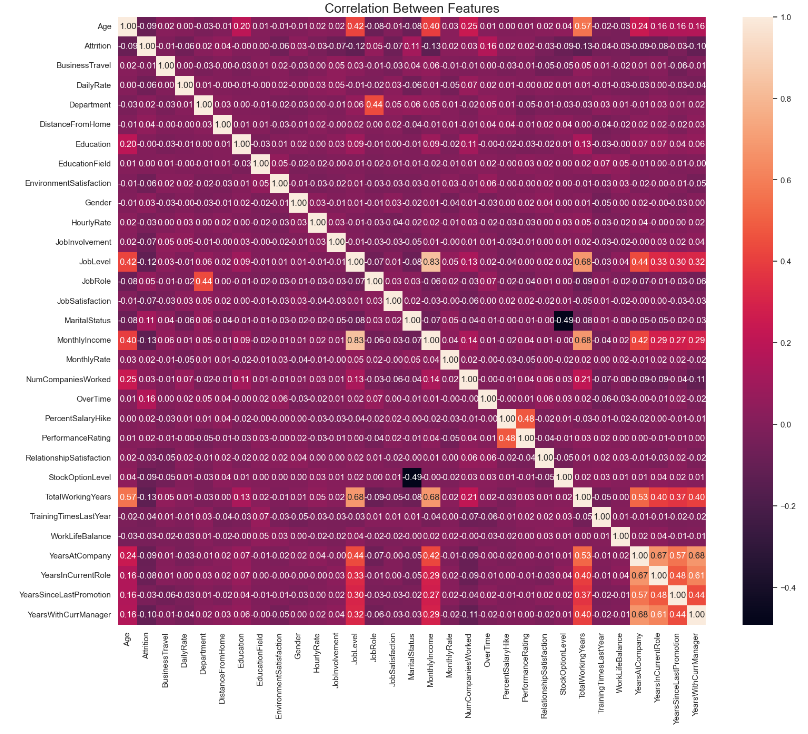
'No': 70.22%

'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.





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.

Summary of Descriptive Statistical Analyses (Continued)

Frequency Distributions for Categorical Variables:

Gender: Out of the respondents, 540 are female, and 783 are male.

Department: The majority are in 'Research & Development' and 'Sales'.

EducationField: 'Life Sciences' and 'Medical' fields are predominant.

JobRole: Diverse roles are represented, with 'Sales Executive', 'Research Scientist', and 'Laboratory Technician' being more common.

MaritalStatus: Distribution across 'Married', 'Single', and 'Divorced' categories.

Attrition: There are 258 cases of attrition.

OverTime: 394 employees have reported working overtime.

These frequencies provide insight into the demographic and professional composition of the workforce.

# 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.

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

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.

# 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.

# 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 R&D 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.

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.

# Reference

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

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/.