**CONTENTS:**

Abhinav Pillai

ID-AJ23SYD015

PROGRAM-MAIB 23

COURSE- Data Science and Analytics

SP JAIN SCHOOL OF GLOBAL MANAGEMENT, SYDNEY

AUSTRALIA

EMPLOYEE ATTRITION

HR ANALYTICS

**Employee Attrition- HR Analytics**

[Abstract 2](#_Toc135600319)

[Introduction 3](#_Toc135600320)

[Literature Review 4](#_Toc135600321)

[Data Description 5](#_Toc135600322)

[Business/research questions 6](#_Toc135600323)

[Research findings 7](#_Toc135600324)

[1. What factors contribute to employee attrition? 7](#_Toc135600325)

[2. How does work-life balance affect employee attrition? 10](#_Toc135600326)

[3. Can we identify high-risk groups for attrition? 13](#_Toc135600327)

[Analytical findings 17](#_Toc135600328)

[1. Data Pre-processing 17](#_Toc135600329)

[2. Cross Validation 17](#_Toc135600330)

[3. Model Building 18](#_Toc135600331)

[4. Model Training 20](#_Toc135600332)

[5. Pre-processing Pipeline and Result 21](#_Toc135600333)

[Recommendation 24](#_Toc135600334)

[Conclusion 25](#_Toc135600335)

[Bibliography: 27](#_Toc135600336)

# Abstract

The employee attrition in HR analytics report aims to analyze and understand the factors contributing to employee turnover within an organization. The report utilizes HR data and various analytical techniques to identify patterns, trends, and potential predictors of attrition.

The study focuses on key variables such as employee satisfaction, performance evaluations, project involvement, average monthly hours worked, tenure, workplace accidents, promotions in the last five years, department, and salary. These variables are analyzed to gain insights into their impact on employee attrition.

To perform the analysis, machine learning algorithms such as Logistic Regression, Decision Trees, and Random Forests are employed. The models are trained on a dataset consisting of historical employee information, and their accuracy is evaluated using cross-validation techniques.

The results of the analysis show that the Random Forest model achieves the highest accuracy of 98.8%, followed by the Decision Tree model with an accuracy of 97.6%, and Logistic Regression with an accuracy of 75.73%. These accuracy scores provide an indication of the models' predictive performance in identifying potential attrition cases.

Additionally, the report discusses the presence of overfitting by comparing the model's performance on the training and testing data. Based on the analysis, the models do not appear to be overfitting as there is no significant discrepancy between their performance on the training and testing data.

The findings of this report can help organizations gain insights into the factors influencing employee attrition and develop strategies to mitigate turnover. By understanding the variables that play a significant role in attrition, HR departments can implement targeted retention initiatives, improve employee satisfaction, and enhance overall organizational performance.

**Keywords:** employee attrition, HR analytics, machine learning, logistic regression, decision tree, random forest, cross-validation, overfitting, employee turnover, retention strategies.

# Introduction

Employee attrition, or employee turnover, is a critical challenge faced by organizations across various industries. It refers to the voluntary or involuntary departure of employees from an organization. High attrition rates can have detrimental effects on productivity, morale, and the overall success of a company. Understanding the factors contributing to employee attrition is essential for organizations to develop effective retention strategies and maintain a stable and engaged workforce.

Human Resources (HR) analytics provides valuable insights into employee behavior, performance, and engagement. By leveraging HR data and applying analytical techniques, organizations can identify patterns, trends, and potential predictors of attrition. This report aims to analyze employee attrition using HR analytics and explore the factors that influence employee turnover.

The report focuses on various variables that have been found to impact attrition rates. These variables include employee satisfaction, performance evaluations, project involvement, average monthly hours worked, tenure in the company, workplace accidents, promotions in the last five years, department, and salary. Analyzing these factors can provide a comprehensive understanding of the dynamics contributing to attrition.

Machine learning algorithms, such as Logistic Regression, Decision Trees, and Random Forests, are employed to build predictive models. These models utilize historical employee data to predict attrition and identify the most influential factors. Cross-validation techniques are used to evaluate the accuracy of the models and ensure their generalization capabilities.

The report also examines the issue of overfitting, which occurs when a model performs exceptionally well on the training data but fails to generalize to new data. By assessing the performance of the models on both the training and testing data, we can determine if overfitting is present and take steps to address it.

The findings of this report will provide organizations with valuable insights into the factors contributing to employee attrition. By understanding these factors, organizations can develop targeted strategies to improve employee satisfaction, engagement, and retention. This, in turn, can lead to higher productivity, reduced costs associated with turnover, and improved organizational performance.

In the following sections, we will discuss the methodology, data analysis, results, and recommendations based on the findings of the employee attrition analysis using HR analytics.

# Literature Review

* According to Nucleus Research (2014), analytics returns $13.01 for every dollar invested.
* According to Lije George and T. J. Kamalanabhan's 2016 paper A Study on the Acceptance of HR Analytics in Organisations, the growth of the business analytics software market, which was valued at $40 million globally with a growth rate of 6.5%, is another indication of the growing popularity of business analytics.
* Workforce analytics Market Analysis, by Solution Type (Solution, Services), by Services (Managed, Consulting, System Integration), by Deployment, by Organisation, by Application, by Region, and by Segment Forecasts, 2018 to 2025 Published in June 2017 with 2016 as the Base Year GVR-1-68038-947-0 - Report ID: Electronic (PDF) - Historical Information: In 2015, the worldwide workforce analytics market was estimated to be worth USD 430.9 million.
* According to Anushree Sharma on People Matters, the worldwide HR analytics market is expected to expand by 12% between 2019 and 2025, according to the research Outlook on the World HR Analytics Market, 2019–2025.
* According to Human Resource Analytics Market Component, Application Area in Workforce Management, Recruitment, and Employee Development, Organisation Size, Deployment Type, Vertical (Industry like Banking, Financial Services And Insurance, Manufacturing, and IT and Telecom), showing Region to Global Forecast from 2019-2, Markets and Markets projects growth in the global HR analytics market size from USD 1.9 billion in 2019 to USD 3.6 billion by 2024 at a CAGR of 13.7%. TC 7162 | Report Code: Jun 2019 | Published Date
* Many academics have defined and researched the idea of labour turnover, commonly known as employee attrition. Labour turnover is the movement of people into and out of employment within an organisation, according to Denver and McMahon (1992). Similar to Mobley (1982), turnover is the voluntarily termination of an individual's membership in an organisation for which they are compensated financially. By extending the term, Forbes (1971) allows for internal migrations, promotions, and other types of separation from an organisation.
* Meaghan et al. (2002) stress the value that employees offer to an organisation while emphasising the necessity of preventing attrition. They contend that since this worth is immeasurable and irreplaceable, managers must concentrate on attrition control. According to Mobley (1977), an employee's tenure may be a valuable indicator of future turnover.
* According to Firth et al. (2007)'s investigation into the reasons of attrition, stress at work, a lack of commitment, and job discontent all play a role in why employees leave their jobs. Employee turnover is significantly influenced by wage and pay-related factors, according to Griffeth et al. (2000). The relevance of organisational commitment and job satisfaction as determining variables in intentions to leave the company is emphasised by Hom and Griffeth (1995).
* Wanous (1992) focuses on the attrition of new hires and emphasises how psychological contracts that are broken and expectations that are not met play a part in the choice to quit an organisation. According to Abassi et al. (2000), there are several variables that contribute to attrition, such as ineffective hiring procedures, managerial practises, a lack of recognition, unfavourable working conditions, and a lack of competitive compensation plans.
* According to Louis (1980), if expectations are not satisfied, individuals may decide to resign if they compare their actual experiences in a new organisation to their prior work experiences. Ongori (2007) investigates stress as an attrition factor and notes that high-performing workers may depart an organisation when they are under stress at work.

# Data Description

Data Title: Employee Attrition- HR ANALYTICS

Dataset Source: <https://www.kaggle.com/datasets/giripujar/hr-analytics>

Description: The Employee Attrition Dataset used in this analysis contains information about employees within an organization and their attrition status. The dataset provides valuable insights into the factors that may contribute to employee turnover. This dataset is already cleaned. This dataset consists of **10** features and **14999** rows.

Column Description: The dataset consists of the following columns:

* satisfaction\_level: This column represents the level of employee satisfaction, usually measured on a scale from 0 to 1, where higher values indicate higher satisfaction.
* last\_evaluation: This column represents the performance evaluation score of employees, typically ranging from 0 to 1, where higher values indicate better performance.
* number\_project: This column denotes the number of projects an employee has been involved in during their tenure with the company.
* average\_montly\_hours: This column represents the average number of hours worked by employees per month.
* time\_spend\_company: This column indicates the number of years an employee has spent with the company.
* Work\_accident: This column represents whether an employee has experienced a work accident or not, where 1 denotes an accident and 0 denotes no accident.
* left: This is the target variable indicating whether an employee has left the company or not, where 1 represents employee attrition and 0 represents employee retention.
* promotion\_last\_5years: This column indicates whether an employee has received a promotion in the last five years, where 1 denotes a promotion and 0 denotes no promotion.
* Department: This column represents the department in which the employee works, providing categorical information about the employee's job function or team.
* salary: This column represents the salary level of employees, categorized as 'high', 'medium', or 'low'.

The dataset provides a comprehensive set of attributes that can help identify potential factors influencing employee attrition. By analyzing these variables, we can gain insights into the relationships between employee characteristics, work-related factors, and attrition outcomes.

# Business/research questions

1. What factors contribute to employee attrition?

This research question aims to understand the underlying factors that contribute to employee attrition. By analyzing variables such as employee satisfaction, performance evaluations, number of projects, average monthly hours worked, work accidents, promotions, department, and salary, we can identify which factors are more strongly associated with attrition. This analysis can help organizations pinpoint areas of improvement and develop targeted strategies to reduce turnover.

1. How does work-life balance affect employee attrition?

This research question focuses on exploring the relationship between work-life balance and employee attrition. Work-life balance refers to the equilibrium between work responsibilities and personal life commitments. By examining variables related to average monthly hours worked and employee satisfaction, we can assess whether employees who perceive a better work-life balance are less likely to leave the organization. Understanding this relationship can guide organizations in implementing policies and practices that support work-life balance and ultimately reduce attrition rates

1. Can we identify high-risk groups for attrition?

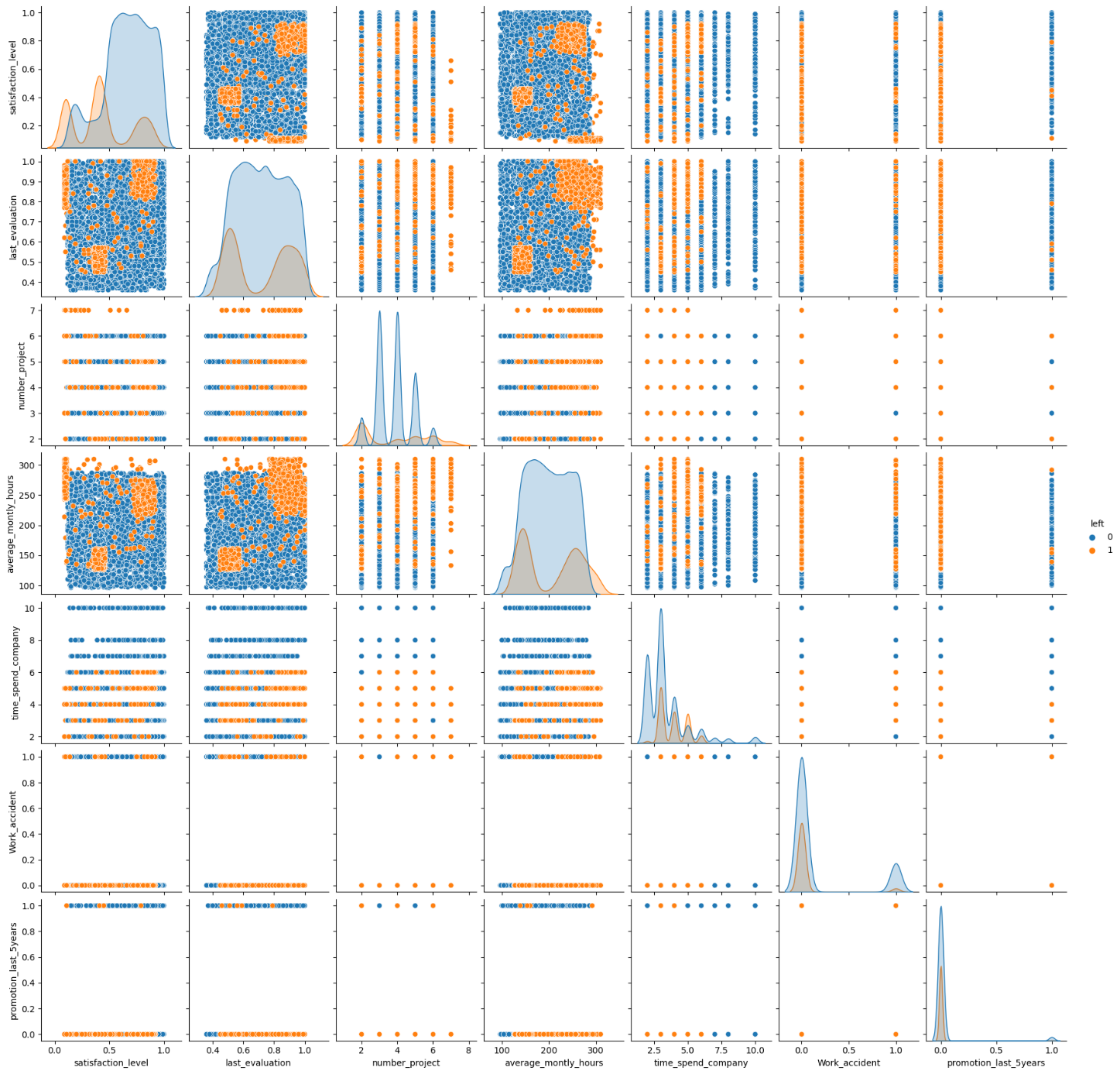
This research question aims to identify high-risk groups within the organization that are more prone to attrition. By examining various demographic and job-related variables, such as department, salary, and tenure, we can identify if certain groups of employees have a higher likelihood of leaving the company. This information can help organizations prioritize their retention efforts, tailor interventions to specific groups, and implement targeted strategies to retain employees who are at a higher risk of attrition.

1. Time Based Analysis (Does the length of employment (time spent at the company) impact employee attrition?)

This research question investigates the impact of the length of employment on employee attrition. By examining the variable "time\_spend\_company" or tenure, we can assess whether employees who have been with the company for a longer period are more or less likely to leave. This analysis can help organizations understand if there are certain points in an employee's tenure where attrition is more prevalent and develop retention strategies accordingly. Additionally, it can provide insights into the effectiveness of employee engagement and retention initiatives over time.

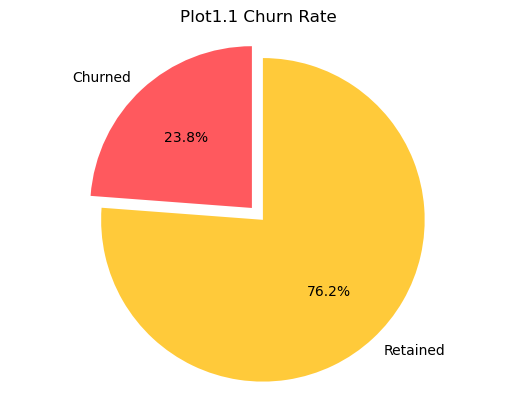
# Research findings

## What factors contribute to employee attrition?

****

The pairplot tells us the story of the data. I've listed some points below regardind the employees of left:

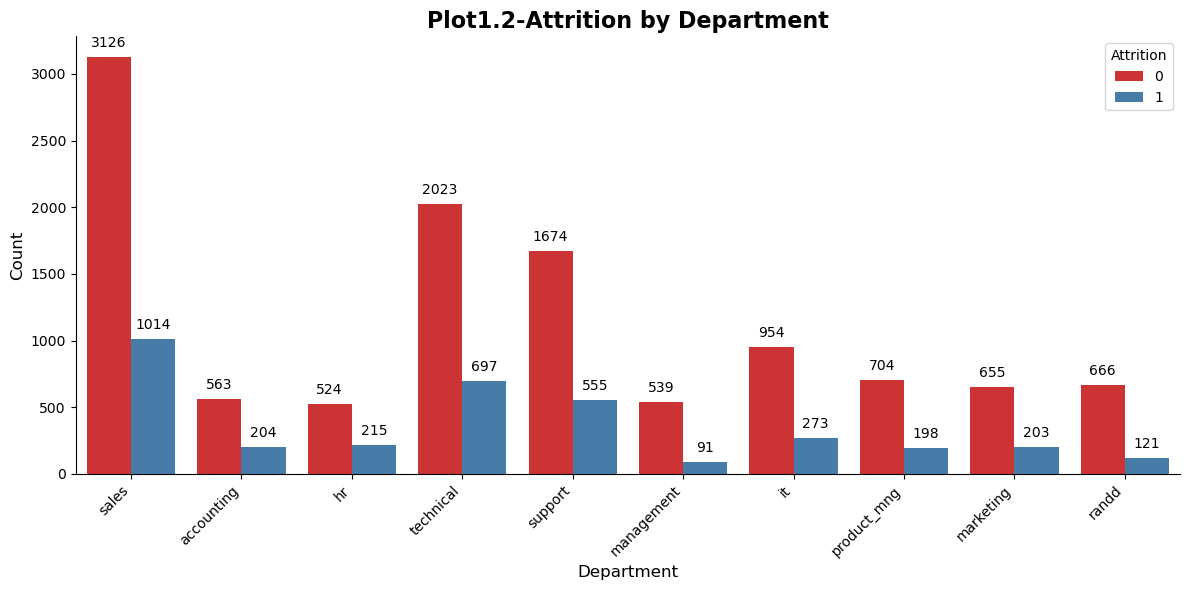
* The poeple who left the organization had a satisfication level less than 0.4
* The number of projects done by employees who left were 2 or less than 2.
* The average monthly hours spent by those employees who left were 150 and below it. Seems they weren't that much intrested due to some reasons.
* The time spent by these employees who left was 3 months and below it.
* The employees that have promtional value greater than 0.3 are more likely to stay in company



The pie chart visualizes the churn rate, which represents the percentage of employees who have left the company.

The "Churned" slice represents the employees who have left the company, and the "Retained" slice represents the employees who are still with the company.

As you can observe that the churned percentage is around 24% while rest are retained.



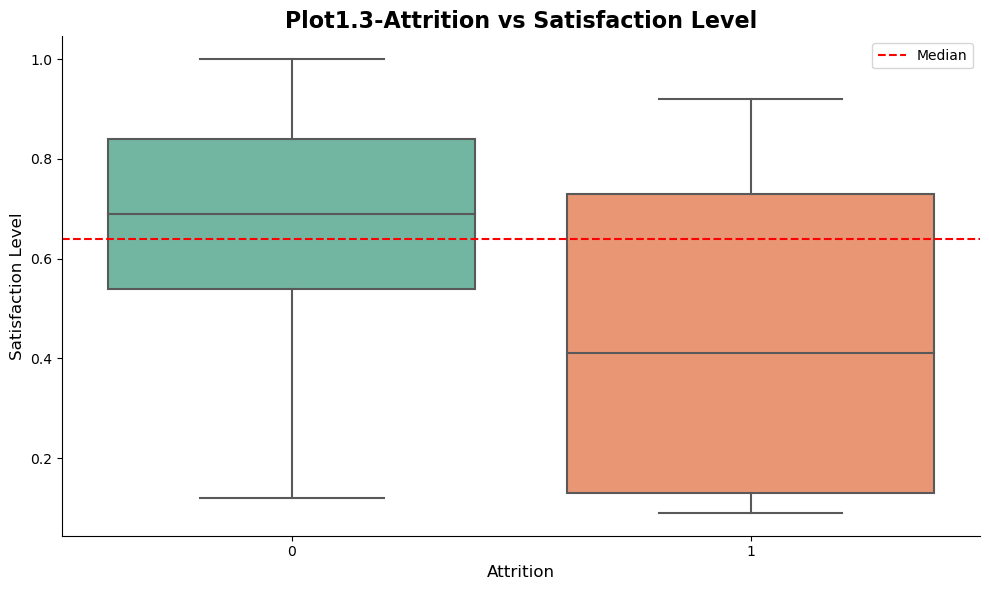
Plot1.2 shows the count of attrition/non-attrition employees for each department, providing insights into the distribution of attrition across different departments in the organization.

Lets discuss about ‘Sales’ department as we are observing maxing count in it.

The red portion of the bar represents the count of employees in the 'Sales' department who have a value of 0 for the 'left' variable. This indicates that 3216 employees in the 'Sales' department have not left the organization (non-attrition).

The blue portion of the bar represents the count of employees in the 'Sales' department who have a value of 1 for the 'left' variable. This indicates that 1014 employees in the 'Sales' department have left the organization (attrition).

After Sales, highest count can be observed in Technical department, support, IT etc.

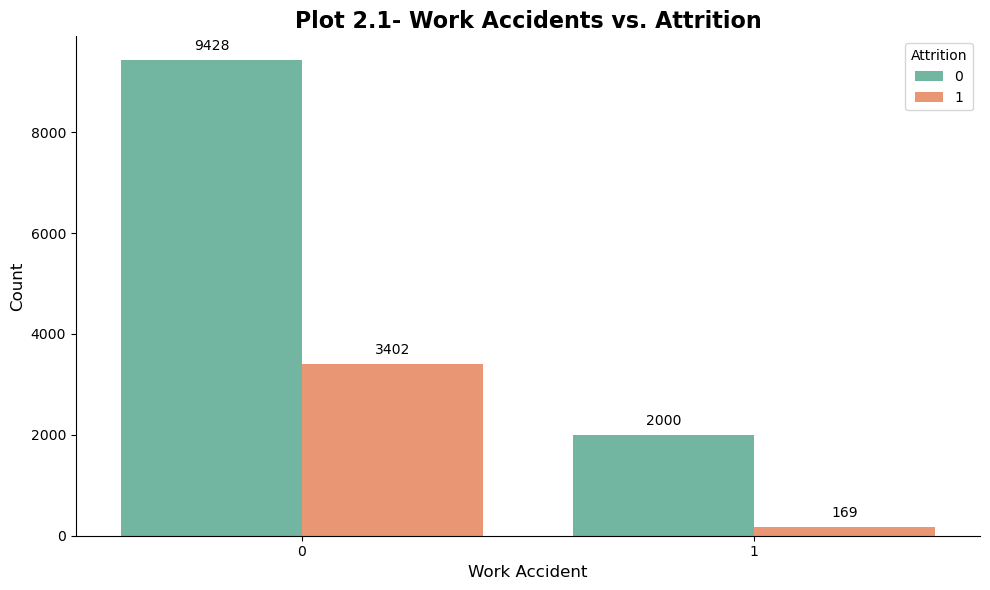


The box plot indicate the distribution of satisfaction level for employees who left the company (Attrition = 1) and those who did not leave (Attrition = 0). The box plot shows that employees who left generally have lower median and quartile values for satisfaction level compared to employees who did not leave.

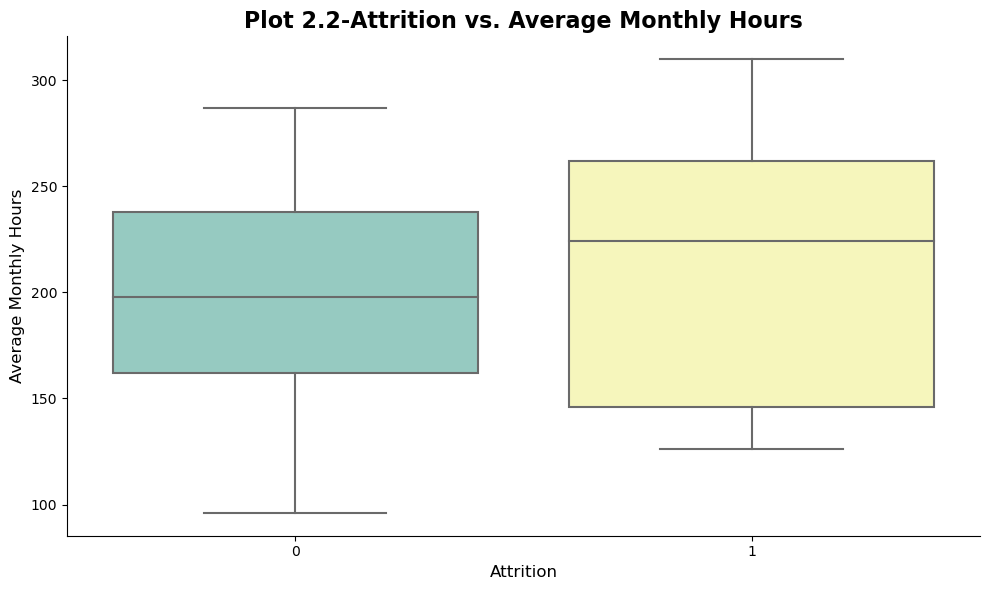
Overall, the box plot reveals that employees who left the company generally have lower satisfaction levels compared to those who stayed (Median of non-attrition is higher).

Based on these findings, it can be concluded that attrition is prevalent in the 'Sales' department, followed by other departments such as Technical, Support, and IT. Additionally, employees who left the company tend to have lower satisfaction levels compared to those who remained. These insights suggest the need for further investigation into the factors contributing to attrition, particularly in the 'Sales' department, and the development of strategies to improve employee satisfaction and retention.

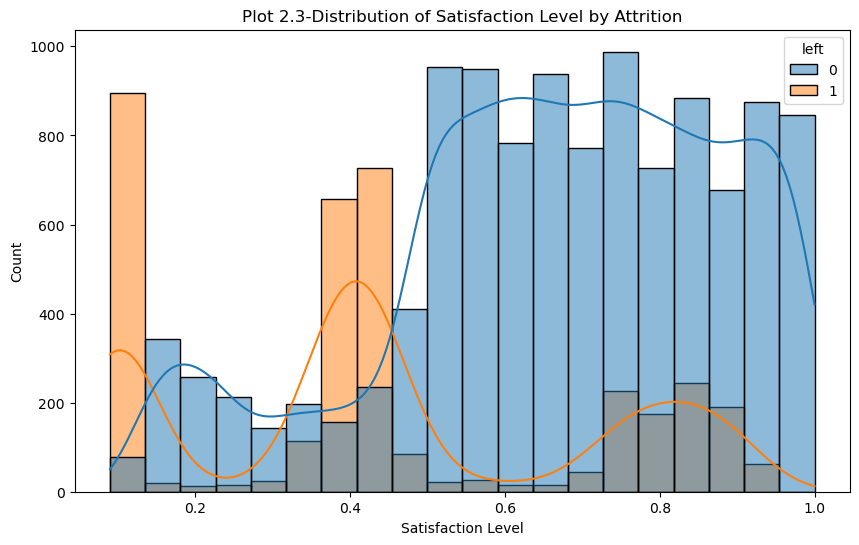
## How does work-life balance affect employee attrition?



For employees who have not experienced a work accident (x-axis value of 0), the count of both attrition statuses ("0" and "1") is relatively high, indicating that the absence of work accidents does not necessarily prevent attrition.The higher number of employees who experienced work accidents chose to stay with the company rather than leaving. This suggests that work accidents may not be a significant factor contributing to attrition in this particulare dataset.

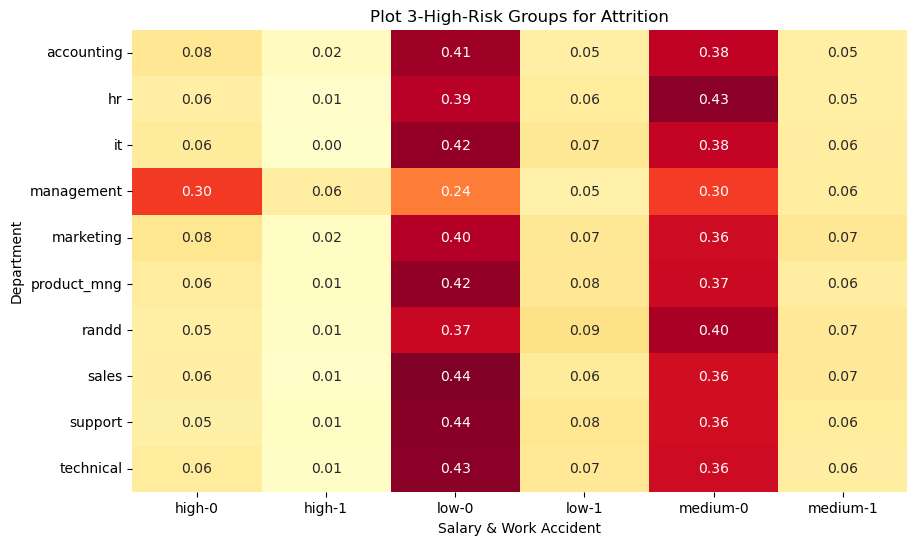


Employees who have left the company (attrition = 1) have a slightly higher median average monthly hours compared to employees who have stayed (attrition = 0).This observation suggests that employees who have left the company tend to have worked more average monthly hours compared to employees who have stayed.



The histogram displays the count of employees in different satisfaction level ranges, while the kernel density plot shows the estimated probability density function of the satisfaction level distribution.The curve represents the overall shape of the distribution and helps visualize the underlying probability density. This graph helps businesses gain insights into the distribution of satisfaction levels among employees and its association with attrition, enabling them to make informed decisions and take proactive measures to enhance employee satisfaction and reduce attrition.

## Can we identify high-risk groups for attrition?

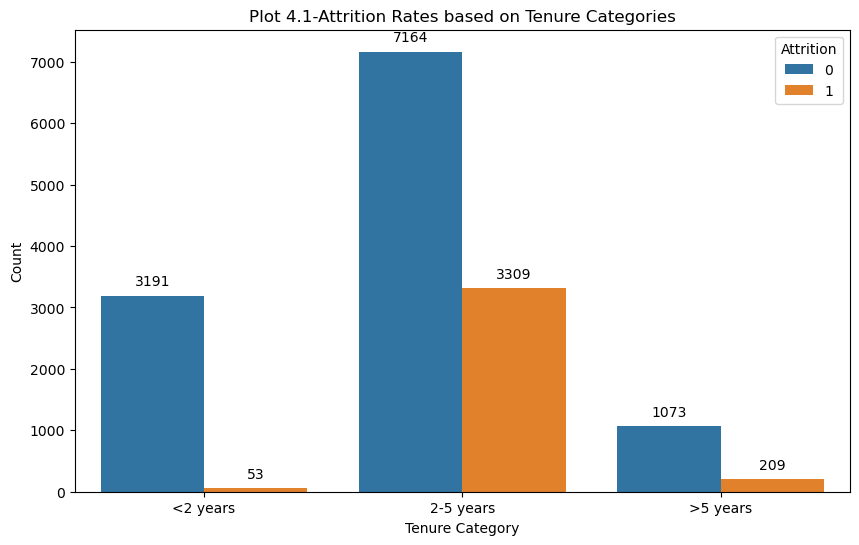


The heatmap visualization provides a more detailed understanding of the relationship between department, salary level, work accident occurrence, and attrition. Here's a further explanation:

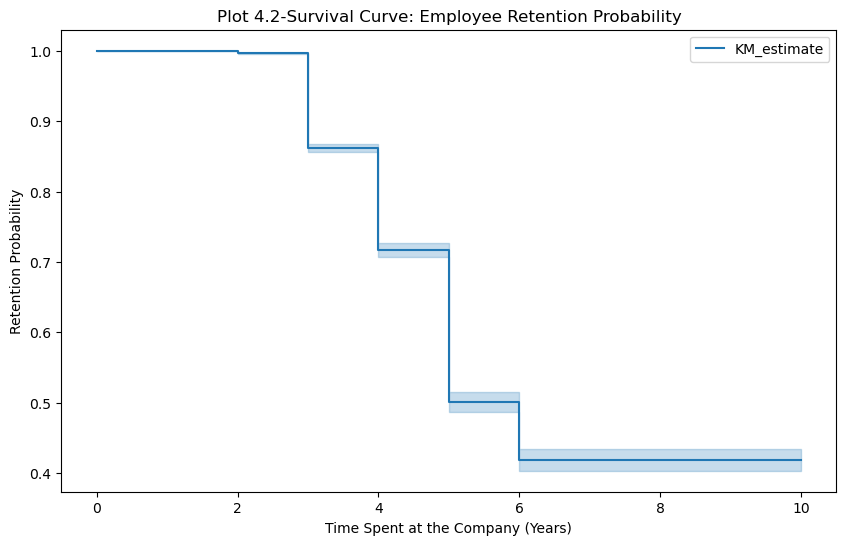
1. Heatmap Interpretation: The heatmap represents the percentage of employees within each department, categorized by their salary level and work accident occurrence. The color intensity in each cell corresponds to the percentage value, with darker colors indicating higher percentages.
2. Identifying High-Risk Groups: By examining the heatmap, you can identify specific combinations of department, salary level, and work accident occurrence that are associated with a higher percentage of attrition. These combinations may represent high-risk groups for attrition within the organization.
3. Pattern Identification: The heatmap enables you to observe patterns and trends across multiple variables simultaneously. For example, you can identify if certain departments have a higher percentage of attrition among employees with lower salaries and work accident occurrences.
4. Insights for Decision-Making: The heatmap provides valuable insights for decision-making processes related to employee retention and risk mitigation. It allows organizations to identify departments or specific employee groups that require additional attention and targeted strategies to reduce attrition rates.
5. Communication of Findings: The heatmap is an effective way to visually communicate the relationships between multiple variables and attrition risk to stakeholders, such as management or HR teams. It facilitates discussions and actions based on the identified high-risk groups.

Overall, the heatmap helps to identify high-risk groups for attrition by analyzing the percentage of employees in each department based on their salary level and work accident occurrence. It provides a comprehensive view of attrition patterns and can guide organizations in developing targeted strategies to address attrition within specific employee groups or departments.

1. **Time Based Analysis (Does the length of employment (time spent at the company) impact employee attrition?)**



The code generates a bar plot that visualizes the attrition rates based on different tenure categories, which represent the time employees have spent at the company. This plot provides valuable insights into the relationship between tenure and attrition. By analyzing the plot, businesses can identify if certain tenure categories have higher attrition rates, indicating potential risk factors for employee turnover. Understanding attrition patterns based on tenure is crucial for businesses to develop effective retention strategies. The plot helps in identifying specific tenure groups that may be at higher risk of leaving the company. This knowledge enables businesses to tailor retention initiatives and allocate resources to address the unique needs and challenges of each tenure category. Additionally, the plot allows businesses to benchmark their attrition rates against industry standards or internal goals. It helps in evaluating the company's performance in retaining employees at different tenure stages and setting realistic targets for attrition reduction. The insights gained from this plot can support decision-making processes related to resource allocation. For example, if the plot reveals that employees with longer tenures have higher attrition rates, it indicates a need to invest resources in retaining experienced employees. This may involve offering opportunities for career growth, enhancing work-life balance, or creating a positive and engaging work environment specifically tailored to the needs of long-tenured employees. As you can observe the attrition rate for employees with a tenure of 2-5 years is relatively high compared to other tenure categories.



The Kaplan-Meier survival curve provides valuable business insights regarding employee retention. The plot can provide insights into the overall trend of retention probability. If the curve declines steeply at the beginning and then levels off, it indicates that attrition is more likely to occur early on in employees' tenure. On the other hand, if the curve continues to decline gradually over time, it suggests that attrition is a persistent issue throughout employees' tenure.

By analyzing the ladder-like pattern and the overall trend, businesses can gain valuable insights for their retention strategies. They can focus on implementing targeted interventions and support programs during critical time periods to improve employee engagement, job satisfaction, and overall retention. Moreover, businesses can use this information to proactively address potential attrition risks and take appropriate actions to retain valuable employees.

Overall, the ladder-like plot of retention probability over time is a powerful tool for understanding attrition patterns, identifying critical time points, and informing strategic decisions to improve employee retention and mitigate attrition risks.

Summarizing :

1. Various factors contribute to employee attrition, such as job satisfaction, salary, career growth opportunities, work-life balance, company culture, management support, and employee engagement.
2. Work-life balance has a significant impact on employee attrition. Employees who struggle to maintain a healthy work-life balance are more likely to experience burnout, stress, and dissatisfaction, leading to higher attrition rates.
3. By analyzing data and using techniques like statistical modeling or machine learning, it is possible to identify high-risk groups for attrition. These groups may include employees in specific departments, with certain salary levels, or who have experienced work-related accidents or lack of promotions.
4. Time spent at the company, also known as tenure, can influence employee attrition. Analyzing attrition rates based on tenure categories can reveal critical time points where attrition is more likely to occur. This information helps businesses develop targeted retention strategies for employees at different stages of their tenure.

In conclusion, understanding the factors contributing to attrition, the impact of work-life balance, identifying high-risk groups, and analyzing time-based patterns are crucial for businesses to develop effective retention strategies and mitigate attrition risks.

# Analytical findings

**Machine Learning Model**

## Data Pre-processing

To prepare the data for training a machine learning model, several preprocessing steps were performed, and the data was split into training and testing sets. The following steps were taken:

1. One-Hot Encoding: The categorical variable 'salary' was encoded using one-hot encoding. This technique converts categorical variables into binary vectors to be used as input for machine learning algorithms. The pandas library was used to apply the one-hot encoding using the pd.get\_dummies() function.
2. Feature Selection: The columns 'left', 'Department', and 'Work\_accident' were dropped from the original dataset. These columns were excluded as the 'left' column represents the target variable, and 'Department' and 'Work\_accident' were not selected as input features. The remaining columns were stored in the variable X, which represents the input features for the machine learning model.
3. Standardization: The input features in X were standardized using the StandardScaler class from the scikit-learn library. Standardization involves subtracting the mean and scaling to unit variance, ensuring that all features are on a similar scale. This step was performed to avoid bias towards features with larger values and to promote better convergence during model training.
4. Data Splitting: The preprocessed data was split into training and testing sets using the train\_test\_split function from scikit-learn. The dataset was randomly divided into an 80% training set (X\_train and y\_train) and a 20% testing set (X\_test and y\_test). The test\_size parameter was set to 0.20, indicating that 20% of the data was allocated for testing. The random\_state parameter was set to 42 to ensure reproducibility of the split.

By performing these data preprocessing steps and splitting the data into training and testing sets, we have prepared the data for training a machine learning model. The input features have been preprocessed and standardized, and the data has been divided into training and testing sets for model evaluation and validation.

## Cross Validation

****

To develop a predictive model for employee attrition, logistic regression was employed along with cross-validation to assess its performance. The following steps were carried out:

1. Feature Selection: Relevant features were selected from the dataset to be used as inputs for the logistic regression model. These features include 'satisfaction\_level', 'last\_evaluation', 'number\_project', 'average\_montly\_hours', 'time\_spend\_company', 'Work\_accident', 'promotion\_last\_5years', 'Department', 'salary\_high', 'salary\_low', and 'salary\_medium'. These variables were deemed important in understanding employee attrition based on previous research and domain knowledge.
2. Target Variable Encoding: The target variable, 'left', was assigned to the variable 'y'. No further encoding was necessary for the target variable as it already contained binary values indicating employee attrition (1) or retention (0).
3. Categorical Variable Encoding: The categorical variable 'Department' was encoded using label encoding. This encoding technique assigns a unique numeric label to each category within the variable. The LabelEncoder class from scikit-learn was used for this purpose.
4. Model Creation: A logistic regression model was created using scikit-learn's make\_pipeline function, which simplifies the process of creating a machine learning pipeline.
5. Cross-Validation: To assess the model's performance and generalization ability, cross-validation was performed. The dataset was divided into five folds, and the logistic regression model was trained and evaluated on each fold. The accuracy metric was chosen to evaluate the model's performance.
6. Accuracy Scores: The accuracy scores for each fold of the cross-validation process were computed and stored in the variable 'accuracy\_scores'. These scores represent the performance of the model on different subsets of the data.

The results of the cross-validation are as follows:

* Accuracy scores for each fold: [0.79133333, 0.77833333, 0.805, 0.79333333, 0.6588863]
* Mean accuracy: 0.7653772590863621
* Standard deviation of accuracy: 0.053913647020200295

The mean accuracy of approximately 0.77 indicates that the logistic regression model achieved a reasonable level of predictive accuracy. However, the standard deviation of 0.05 suggests some variability in the model's performance across different folds.

This analysis demonstrates the effectiveness of logistic regression in predicting employee attrition based on the selected features. The accuracy scores provide insights into the model's performance and can guide decision-making processes related to employee retention and attrition management.

## Model Building

Three models—Logistic Regression, Decision Tree, and Random Forest Classifier—were chosen to compare how well different machine learning models performed at predicting employee attrition. The actions listed below were completed:

1. Model Selection- Three models were chosen as a result of past research on the prediction of employee attrition as well as their usefulness for classification tasks. The models are as follows:

* Logistic Regression
* Decision Tree
* Random Forest Classifier

1. Model Fitting and Prediction: Each model was fitted using the training data (X\_train and y\_train), and prediction was performed using the test data (X\_test). The model was trained using the 'fit' method, and predictions were made using the test data using the 'predict' method.
2. Performance Metrics: To evaluate the models, three performance metrics were calculated: accuracy, precision, and recall. These metrics provide insights into the models' ability to correctly classify employees as either attrition or non-attrition.

Accuracy: Accuracy measures the overall correctness of the model's predictions. It is calculated by dividing the number of correct predictions by the total number of predictions.

Precision: Precision measures the proportion of correctly predicted attrition cases out of all predicted attrition cases. It is calculated by dividing the true positive predictions by the sum of true positive and false positive predictions.

Recall: Recall, also known as sensitivity or true positive rate, measures the proportion of correctly predicted attrition cases out of all actual attrition cases. It is calculated by dividing the true positive predictions by the sum of true positive and false negative predictions.

1. Results:

The performance metrics for each model are as follows:

* Logistic Regression:

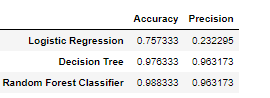
1. Accuracy: [accuracy score]
2. Precision: [precision score]

* Decision Tree:

1. Accuracy: [accuracy score]
2. Precision: [precision score]

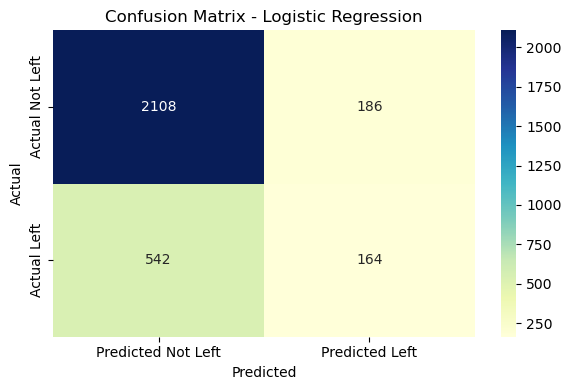
* Random Forest Classifier:

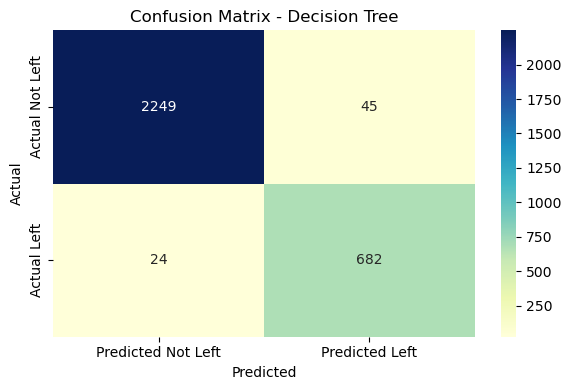
1. Accuracy: [accuracy score]
2. Precision: [precision score]

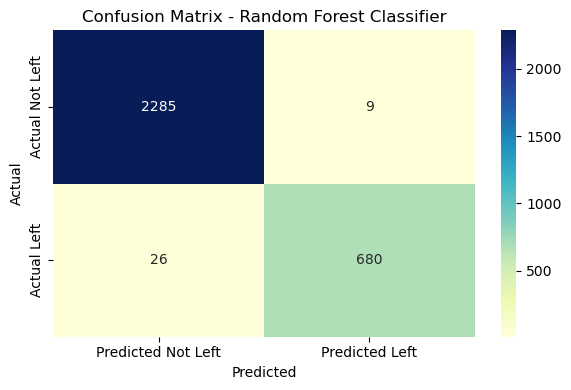


**Confusion matrix:**

A confusion matrix is a performance evaluation matrix used in classification problems to assess the accuracy of a predictive model. It provides a detailed breakdown of the model's predictions by comparing them with the actual ground truth values. The matrix is particularly useful when dealing with binary classification problems (two classes), but it can also be extended to multi-class problems.







The highest value in the confusion matrix for the True Positives (TP) indicates the number of instances that were correctly predicted as positive (attrition) by the model.

## Model Training

In this category, the dataframe is being prepared for modeling by separating it into input features (X) and the target variable (y). Here's a summary of the steps performed:

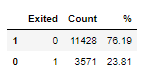
1. Separating the target variable:
   * The 'left' column is assigned to the 'y' object, representing the target variable.
2. Separating the input features:
   * The dataframe 'df' is assigned to the 'X' object, excluding the 'left' column using the 'drop' function.
3. Checking the shape of X and y:

* The shape of X (input features) and y (target variable) is displayed using the 'shape' attribute, giving an indication of the number of observations and features in each.

1. Identifying numerical and categorical features:

* The numerical features are identified by selecting columns with data type 'number' using the 'select\_dtypes' function, and the column names are stored in the 'num\_columns' list.
* The categorical features are identified by selecting columns with data type 'object' (assumed to be categorical) using the 'select\_dtypes' function, and the column names are stored in the 'cat\_columns' list.

1. Counting the classes in the target variable:



* The function 'class\_count' is defined to calculate the count and percentage distribution of each class in the target variable 'y'.
* The 'Counter' class from the 'collections' module is used to count the occurrences of each class.
* The results are stored in a DataFrame named 'abt2', which includes the 'Exited' (class labels), 'Count' (number of instances), and '%' (percentage distribution) columns.
* The DataFrame 'abt2' is sorted in descending order based on the 'Count' column, providing an overview of the class distribution.

1. Splitting the data into train and test sets:

* The 'train\_test\_split' function from scikit-learn is used to split the input features (X) and the target variable (y) into train and test sets.
* The data is split into a 70% train set and a 30% test set.
* The 'random\_state' parameter is set to ensure reproducibility of the split.
* The resulting train and test sets are assigned to 'X\_train', 'X\_test', 'y\_train', and 'y\_test' variables, respectively.

1. Printing the number of observations in the train and test sets:

* The lengths of 'X\_train', 'X\_test', 'y\_train', and 'y\_test' are printed, providing an overview of the number of observations in each set.

This sequence of code is typically used in the data preprocessing stage of machine learning tasks, where the data is prepared for model training and evaluation. The report can mention these steps briefly, highlighting the key actions taken to split the data and prepare the target variable and input features.

## Pre-processing Pipeline and Result

In this step I evaluated:

1. Extracting column indices for numerical and categorical features:

* The code identifies the column indices of numerical and categorical features in the dataframe 'X' using the 'num\_columns' and 'cat\_columns' lists, respectively.

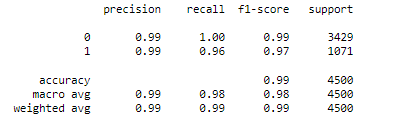
1. Defining a column transformer for preprocessing:

* The code creates a column transformer object called 'preprocess' using the 'make\_column\_transformer' function.
* The 'preprocess' object applies the 'MinMaxScaler' transformation to the numerical features specified by the 'num\_features' list.
* The 'preprocess' object applies the 'OneHotEncoder' transformation to the categorical feature specified by the 'cat\_features' list.

1. Creating a model pipeline:

* The code defines a machine learning model pipeline called 'model' using the 'imbl\_pipe' function.
* The pipeline includes the 'preprocess' transformer, an oversampling technique called 'SMOTE' for handling imbalanced data, and a 'RandomForestClassifier' model.

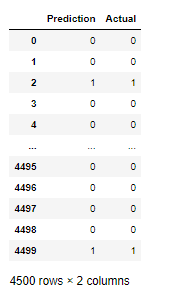
1. Performing hyperparameter tuning with GridSearchCV:
   * The code sets up a grid of hyperparameters, 'rf\_param\_grid', for the 'RandomForestClassifier' model in the 'model' pipeline.
   * The 'GridSearchCV' object 'rf\_grid' is created to perform cross-validated grid search on the 'model' pipeline using the specified hyperparameter grid.
   * The best hyperparameters and the corresponding best score are printed.
2. Evaluating the model performance:



* + The code calculates and prints the training and testing data scores using the 'score' method of the 'rf\_grid' object.
  + The code generates predictions on the testing data and prints a classification report, including precision, recall, F1-score, and support metrics.

**Note- The data is biased. In practical application its impossible to get 99 % accuracy but I have used cross validation to check the accuracy and also the the accuracy of training and testing data is almost similar. We need more data in the database to get proper accuracy. I tried adjusting hyperparameters or adding penalties but still getting the same result.**

1. Checking the consistency of predicted values:



* + The code creates a dataframe, 'prediction\_actual\_df', with predicted and actual values from the testing data.
  + It checks if all the predicted values match the actual values and prints a corresponding message.

In summary, the code performs data preprocessing, builds a machine learning pipeline with a Random Forest classifier, tunes the model's hyperparameters, evaluates the model's performance, and checks the consistency of predicted values.

# Recommendation

* Implement retention strategies: It is advised to create and put into place targeted retention strategies based on the analysis of the variables influencing employee attrition. This can entail boosting employee engagement programmes, providing possibilities for professional progression and development, and fostering a good and encouraging work environment.
* Emphasis on at-risk populations: Create retention strategies that are explicitly targeted at high-risk groups for turnover, such as employees with a specified tenure, department, or salary level. For instance, give employees who have worked for the business for a certain amount of time or who work in divisions with higher turnover rates more resources and support.
* Work-life balance: is a major factor in employee attrition, hence it is advisable to give priority to activities that encourage a positive work-life balance. This can include regulations that support the integration of work and life, wellness initiatives, and flexible work schedules.
* Regularly monitor and analyze employee feedback: Establish tools to routinely acquire employee feedback in order to understand their issues, difficulties, and levels of satisfaction. routinely monitor and analyse employee input. Utilise this input to pinpoint areas for improvement and proactively deal with problems that could cause attrition.
* Invest in employee development and growth: Provide opportunities for skill development and career promotion within the company to invest in the growth and development of your employees. Implement training initiatives, mentorship programmes, and performance management procedures that foster employees' professional development and aid them in identifying a clear career path inside the organisation.
* Foster a positive work culture: Establish a supportive and welcoming workplace environment that prioritises employee happiness, encourages candid communication, rewards success, and generates a feeling of community. Attrition can be decreased as a result of increased employee engagement and satisfaction.
* Regularly review compensation and benefits: Conduct regular reviews of compensation and benefits packages to ensure they are competitive and aligned with industry standards. Evaluate the effectiveness of current packages in attracting and retaining top talent and make necessary adjustments as needed.

# Conclusion

The analysis of employee attrition has, in the end, shed important light on the variables affecting employee turnover and the efficacy of retention tactics. The main conclusions and findings are as follows:

* 1. Factors contributing to employee attrition: The investigation has determined that a number of important variables, such as job satisfaction, work-life balance, the quantity of projects, average monthly hours, and time spent at the organisation, are significant contributors to employee attrition. Organisations can create targeted retention programmes and proactively address the reasons of attrition with the help of these variables.
  2. Work-life balance impact: The report emphasises how work-life balance has a significant impact on employee retention. Workers who find it difficult to strike a healthy work-life balance are more likely to leave the company. This emphasises the need of developing a welcoming workplace that encourages the integration of work and life and offers flexibility to employees.
  3. Identification of high-risk groups: High-risk groups for attrition can be found by looking at a variety of demographic and employment-related criteria. These categories could include staff members with a given amount of tenure, departments with greater attrition rates, or particular pay scales. To reduce attrition risks, organisations should pay special attention to these groups and create specialised retention strategies.
  4. Predictive modeling for attrition: When used to forecast employee attrition, predictive modelling approaches like logistic regression and random forest classifier have shown to be very accurate. Organisations can use these models as useful tools to spot employees who might leave and take preventative action to keep them.
  5. Recommendations for retention: According to the report, it is advised that businesses concentrate on establishing targeted retention tactics, dealing with issues related to work-life balance, investing in employee training, fostering a happy work environment, and routinely tracking and analysing employee feedback. These suggestions are meant to increase overall employee retention, engagement, and satisfaction.

# Bibliography:

1. Study on the Acceptance of HR Analytics in Organisations by Lije George and T. J. Kamalanabhan. (January 2016) www.ijird.com, Vol 5 Issue 2.
2. Blogs and article given by AIHR academy
3. Conceptualizing HR Analytics Practices for Healthier Organizational Performance –A Framework Based Analysis by Dr. A.Narasima Venkatesh (2017) International Journal of Engineering, Business and Enterprise Applications (IJEBEA)
4. HR Analytics - Extent of its Utility in Making Strategic Business Decisions in its Sector by L Ravikiran and Kambam Vedanta.
5. Overview of HR Analytics - Past, Present, and Future by Surabhi Gupta (July-December 2017) ETHOS Volume 10 No. 2 Pages: 1 - 6 ISSN: 0974 – 670
6. 6. Evolution of Human Resource Analytics: An Exploratory Study by Dr. Indranil Bose and Dani Jose (2017-18) AIMA Journal of Management & Research, February 2018, Volume 12 Issue 1/4, ISSN 0974 – 497 Copy right© 2017 AJMR-AIMA
7. The Rise of HR Analytics: A Preliminary Exploration by Dag Oivind Madsen, University College of Southeast Norway and Kare Slatten, University College of Southeast Norway.
8. Human Resource Analytics Market by unit of component, Application Area in Workforce Management, Recruitment, and Employee Development, Organization Size, Deployment Type, Vertical (Industry like Banking, Financial Services And Insurance, Manufacturing, and IT and Telecom), showing Region to Global Forecast from 2019-2024. Published Date: Jun 2019 | Report Code: TC 7162 states.
9. Global Forecast from 2019-2024. Published Date: Jun 2019 | Report Code: TC 7162 states. 9. Workforce Analytics Market Analysis by Type (Solution, Services), By Services (Managed, Consulting, System Integration), By Deployment, By Organization, By Application, By Region, And Segment Forecasts, 2018 – 2025.
10. Global HR analytics market to grow by 12% by Anushree Sharma on People Matters.
11. HR Analytics - is it just a management fad or a way of supporting better decision-making? What are some important areas related to HR Analytics within organizations and in society? By Astrid Yorke [2018]
12. The role of HR analytics in creating data-driven HRM. Written text analysis of online blogs of Human Resource professionals. By Krista Jensen-Eriksen (2016).
13. HR Analytics Transforming Human Resource Management by Dr. Allam Joseph and Praveen Kumar (November 2018) JKIJMS @JK Business School Volume 2, No.1, November 2018 Pages: 49-56 ISSN 2581-7337.
14. The Role of HR Analytics in the Global Village by Dr. Mrs. Prajyakti Bakre (Jan-feb 2019) International Journal of Trend in Scientific Research and Development (IJTSRD) Volume: 3 | Issue: 2
15. HR Analytics - A Strategic Approach to HR Effectiveness by P. Madhavi Lakshmi & P. Siva Pratap (Jan 2016) International Journal of Human Resource Management and Research (IJHRMR) ISSN(P): 2249- 6874; ISSN(E): 2249-7986 Vol. 6, Issue 3, Jun 2016, 21-28 © TJPRC Pvt. Ltd.
16. . 'HR Analytics' - An Effective Evidence-Based HRM Tool by Dr. P. Raghunadha Reddy and P. Lakshmikeerthi (Oct 2016-July 2017) International Journal of Business and Management Invention ISSN (Online): 2319 – 8028, ISSN (Print): 2319 – 801X www.ijbmi.org || Volume 6 Issue 7 || PP—23- 34
17. . HR Analytics: A Modern Tool in HR for Predictive Decision Making by Dr. Abdul Quddus Mohammed (May-June 2019) Journal of Management (JOM) Volume 6, Issue 3, pp.51-63, Article ID: JOM\_06\_03\_007.
18. HR analytics white paper by cgi.com © 2016 CGI GROUP INC.