# **DATA INTERPRETATION**

## **DOMAIN ANALYSIS:**

**EmpNumber:** Represents a unique identifier for each employee. It serves as a primary key for identifying individual employees in the dataset.

**Age:** Represents the age of each employee. It is essential for analyzing demographics, workforce distribution, and understanding age-related factors affecting performance or attrition.

**Gender:** Indicates the gender of each employee. Gender diversity analysis can provide insights into gender balance within the organization and its impact on various aspects such as job roles, promotions, and salaries.

**EducationBackground:** Describes the educational background of employees.

**MaritalStatus:** Indicates the marital status of employees.

**EmpDepartment:** Specifies the department in which each employee works. Analysis of department-wise performance, attrition rates, and other metrics can provide insights into the organization's structure and functioning.

**EmpJobRole:** Describes the specific job role or position held by each employee within their department.

**BusinessTravelFrequency:** Indicates how frequently employees travel for business purposes. Analysis of travel patterns can help in managing travel-related expenses, work-life balance considerations, and understanding job demands.

**DistanceFromHome:** Represents the distance between an employee's home and the workplace. It can be relevant for assessing commuting challenges, relocation opportunities, and their impact on employee satisfaction and retention.

**EmpEducationLevel:** Specifies the education level attained by each employee, ranging from 'Below College' to 'Doctor'.

**EmpEnvironmentSatisfaction:** Indicates the level of satisfaction employees have with their work environment, ranging from '1' to '4'.

**EmpHourlyRate:** Represents the hourly rate of pay for each employee.

**EmpJobInvolvement:** Describes the level of involvement or engagement employees have in their jobs, ranging from '1' to '4'.

**EmpJobLevel:** Specifies the hierarchical level or rank of each employee's job within the organization.

**EmpJobSatisfaction:** Represents the level of satisfaction employees have with their jobs, ranging from '1' to '4'.

**NumCompaniesWorked:** Represents the number of companies employees have worked for previously.

**OverTime:** Indicates whether employees work overtime or not.

**EmpLastSalaryHikePercent:** Specifies the percentage of the last salary hike received by each employee.

**EmpRelationshipSatisfaction:** Indicates the level of satisfaction employees have with their work relationships, ranging from '1' to '4'.

**TotalWorkExperienceInYears:** Represents the total work experience (in years) of each employee.

**TrainingTimesLastYear:** Specifies the number of training sessions attended by employees in the last year. Analysis of training effectiveness and professional development opportunities can be performed using this column.

**EmpWorkLifeBalance:** Indicates the perceived level of work-life balance by employees, ranging from '1' to '4'.

**ExperienceYearsAtThisCompany:** Represents the number of years employees have been working at the current company.

**ExperienceYearsInCurrentRole:** Specifies the number of years employees have been in their current job role.

**YearsSinceLastPromotion:** Indicates the number of years since the last promotion received by employees.

**YearsWithCurrManager:** Represents the number of years employees have been working with their current manager.

**Attrition:** Indicates whether employees have left the company (attrition) or are still employed.

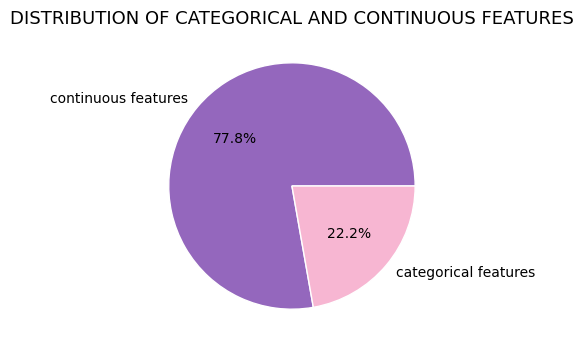
**PerformanceRating:** Specifies the performance rating assigned to each employee, ranging from '2' to '4'.

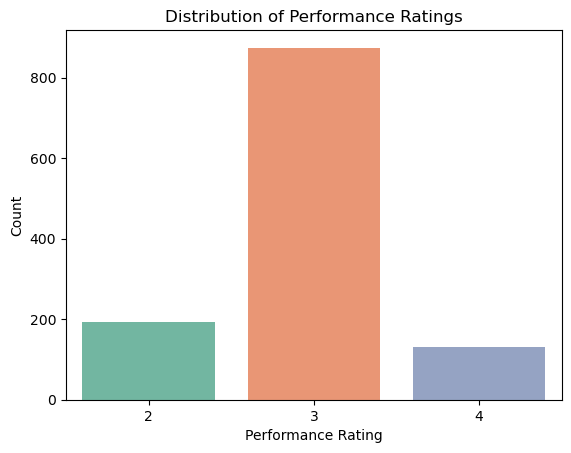
# **SUMMARY OF IMPORTANT ASPECTS OF THE MODEL**

## **DATA ANALYSIS**

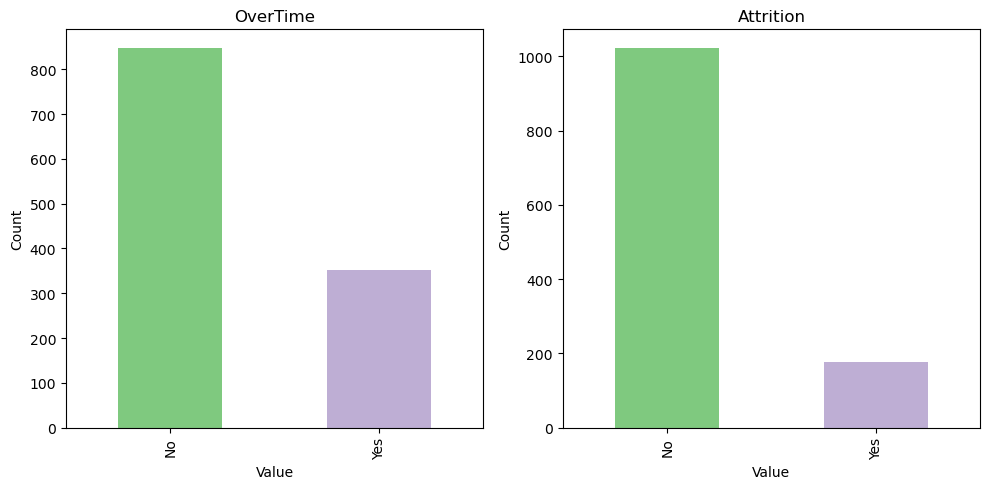
### **VISUALIZATION**

To effectively explain analysis findings, we used visualisations such as **histograms, count plots, pie charts, box plots, pair plots, and heatmaps**. Data visualisations can help make intricate relationships and patterns easier to see and comprehend.

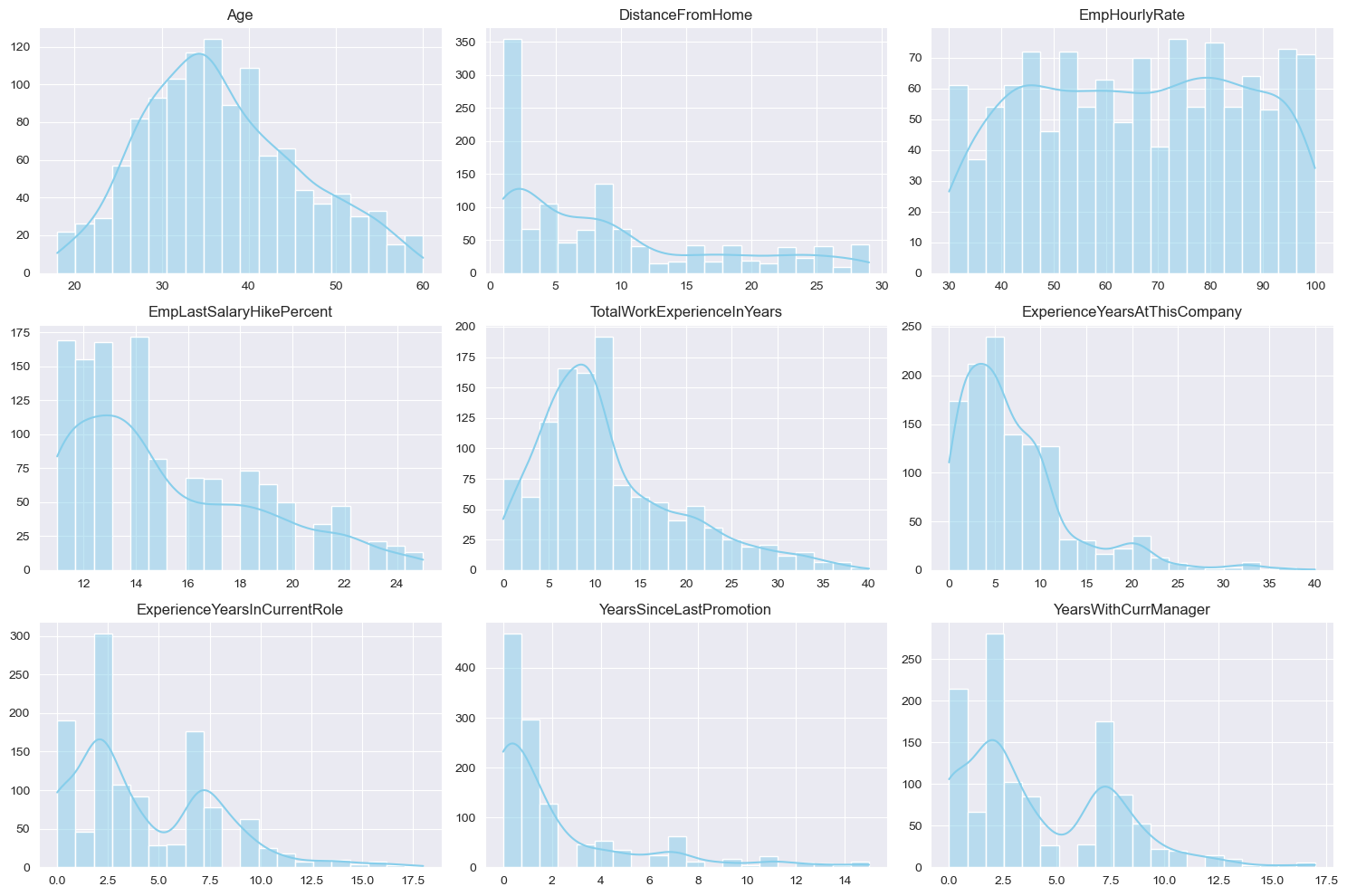




**BINARY FEATURES**



**CONTINOUS COLUMNS**



**INSIGHTS FROM CONTINOUS COLUMNS**

**Age** The distribution is fairly normal, peaking between 30-40 years old, indicating that the majority of employees fall within this age range. This could suggest that the company has a diverse age demographic but with a concentration in the mid-career stage.

**Distance From Home** There's a significant peak at a very short distance, likely indicating that a considerable portion of employees live very close to the workplace, within 0-5 units of distance. This might imply that the company's location is convenient for many of its employees.

**EmpHourlyRate** The distribution appears to be somewhat positively skewed, with a range from 30 to 100. This suggests that most employees have hourly rates between these values, with a wider spread towards the higher rates.

**EmpLastSalaryHikePercent** The distribution is concentrated between 12-14 percent, indicating that most employees received a salary hike within this range. It might suggest a consistent policy or structure regarding salary increments within the company.

**TotalWorkExperienceInYears** The distribution shows a peak between 5-12 years, indicating that a significant portion of employees have this range of total work experience. This suggests that the company might attract mid-career professionals.

**ExperienceYearsAtThisCompany** There's a peak at the lower end of the range, indicating that many employees have relatively low tenure at the company, likely due to a combination of factors such as turnover or a younger workforce.

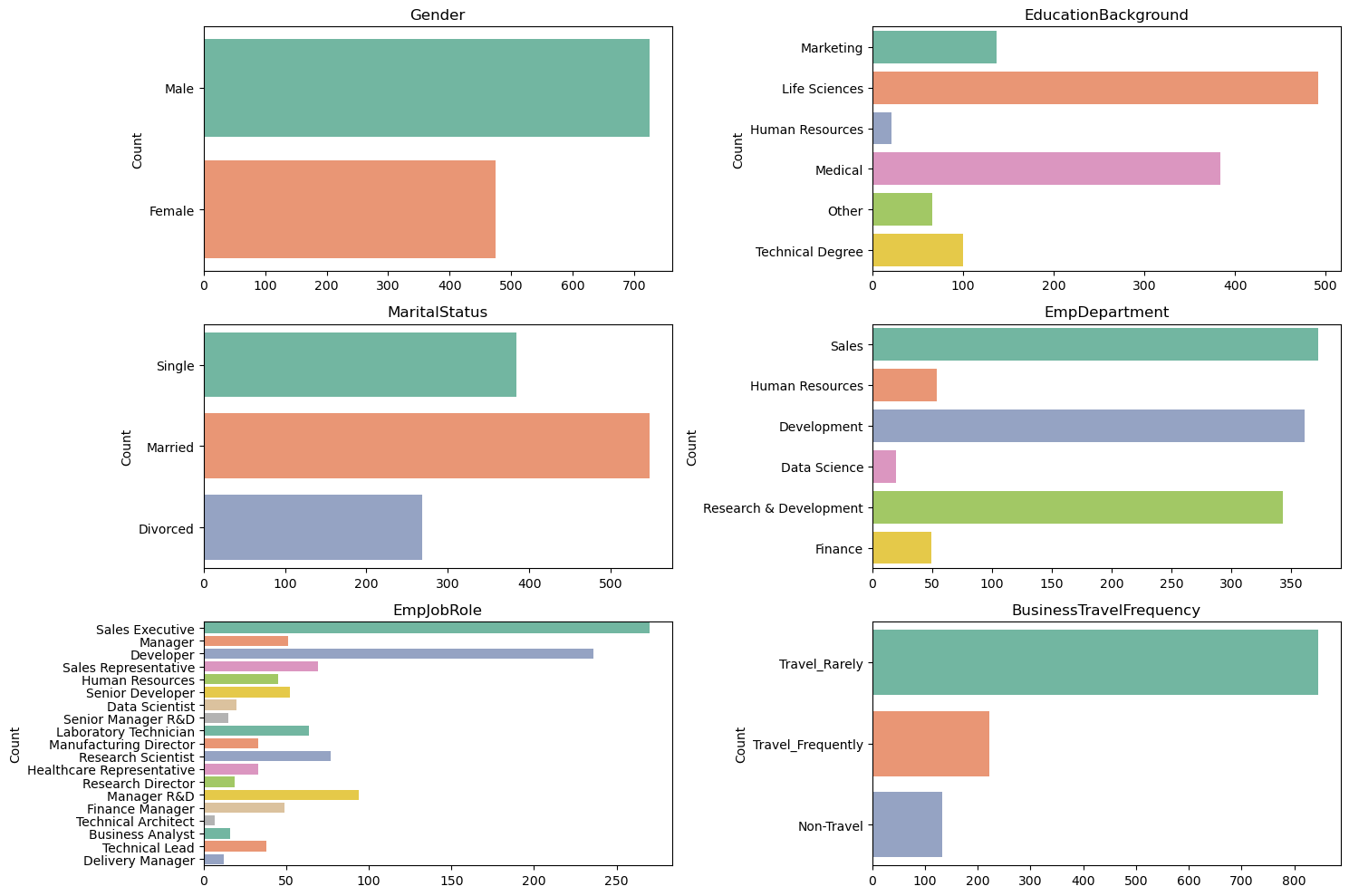
**ExperienceYearsInCurrentRole** The distribution peaks at around 2.5 years, indicating that many employees have relatively short tenures in their current roles. This could suggest either frequent role changes or promotion opportunities within the company.

**YearsSinceLastPromotion** There's a significant peak at the lower end of the range, particularly at 0-2 years, indicating that many employees have been promoted recently or have not had a promotion for a while. This could reflect the company's promotion policies or growth opportunities.

**YearsWithCurrManager** Similar to the distribution for experience in the current role, there's a peak at around 2.5 years, indicating that many employees have relatively short tenures with their current managers. This might suggest either frequent manager changes or a relatively flat organizational structure.

These insights provide a glimpse into the demographic and employment characteristics of the company's workforce, which can be further analyzed to understand trends and make strategic decisions regarding employee retention, career development, and organizational structure.

**CATEGORICAL COLUMNS**



**INSIGHTS FROM CATEGORICAL COLUMNS**

**Gender** The workforce seems to be predominantly male, indicating a potential gender imbalance within the organization. This could prompt further investigation into diversity and inclusion initiatives.

**Education Background** The most common education backgrounds among employees are Life Sciences, Medical, and Technical Degrees. This suggests that the company may attract individuals with specialized knowledge and skills in these fields, which could be relevant to the nature of the company's operations or industry focus.

**Marital Status** The majority of employees are married, followed by single and divorced individuals. This distribution might indicate the need for different types of benefits or support programs tailored to employees with varying marital statuses.

**EmpDepartment** Sales appears to be the largest department within the company, followed by Development and Research & Development. This could suggest a focus on sales-driven operations, with significant resources allocated to sales activities.

**EmpJobRole** The most common job roles are Sales Executive, Developer, and Manager R&D, indicating a mix of sales, technical, and managerial positions within the organization. This distribution might reflect the organizational structure and hierarchy of the company.

**BusinessTravelFrequency** The majority of employees travel rarely for business purposes. This could indicate that the company's operations are predominantly localized or that remote work is common, with less frequent need for business travel. It could also suggest the company's emphasis on work-life balance and minimizing employee travel requirements.

**DISCRETE COLUMNS**

**EmpEducationLevel** Employees with an education level of 3 have the highest performance rating of 3. This suggests that employees with a moderate level of education tend to perform better.

**EmpEnvironmentSatisfaction** Employees with environment satisfaction ratings of 3 and 4 have the highest performance rating of 3. This indicates a positive correlation between environment satisfaction and better performance.

**EmpJobInvolvement** Employees with a job involvement rating of 3 have the highest performance rating of 3. This implies that employees who are moderately involved in their jobs tend to perform better.

**EmpJobLevel** Employees with job levels of 1 followed by 2 have the highest performance rating of 3. This suggests that employees at lower job levels perform better, potentially indicating effectiveness in entry-level positions.

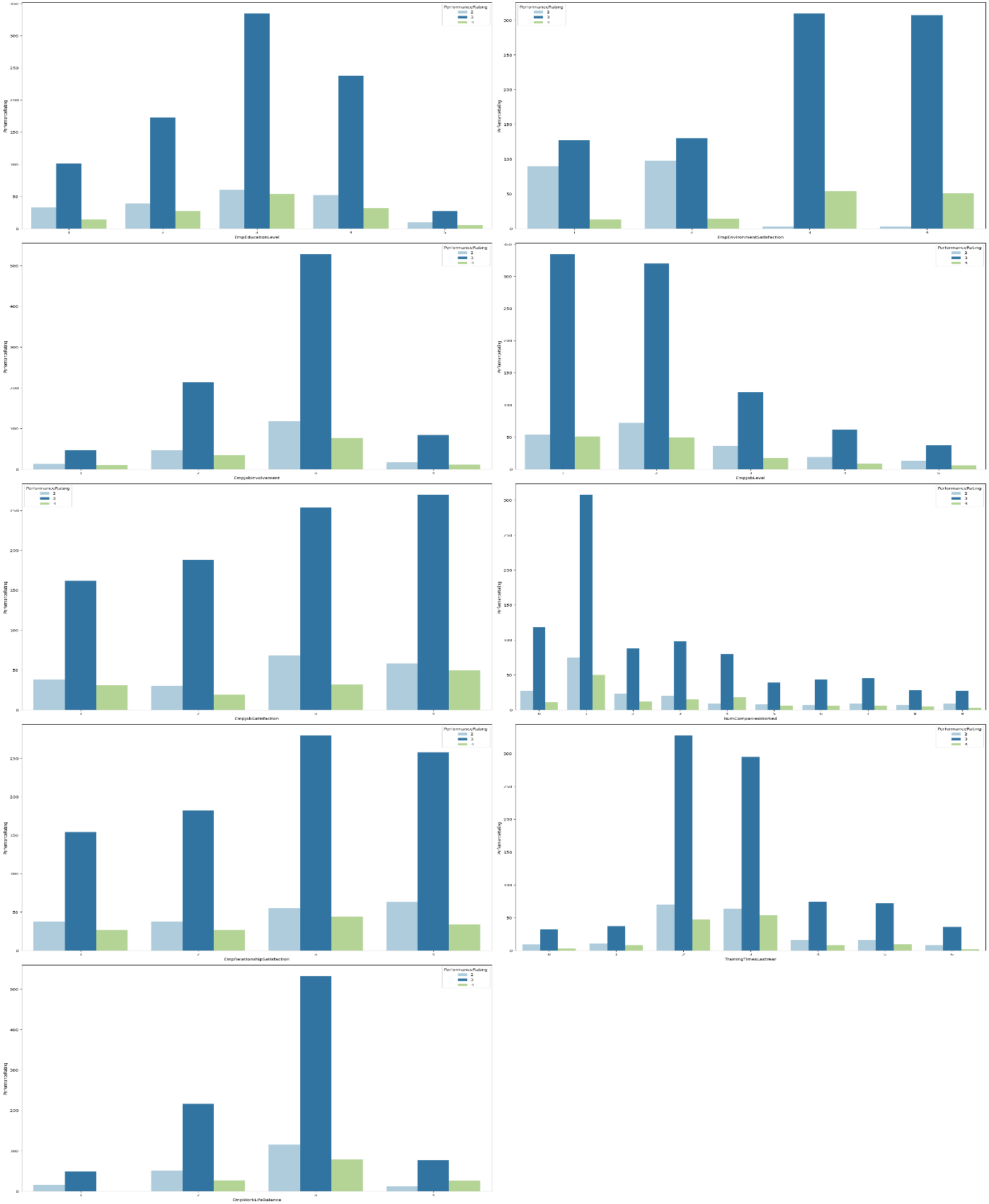
**EmpJobSatisfaction** Employees with job satisfaction ratings of 4 followed by 3, 2, and 1 have the highest performance rating of 3. This indicates a positive correlation between job satisfaction and better performance.

**NumCompaniesWorked** Employees who have worked for only 1 company have the highest performance rating of 3. This suggests that stability in employment history might correlate with better performance.

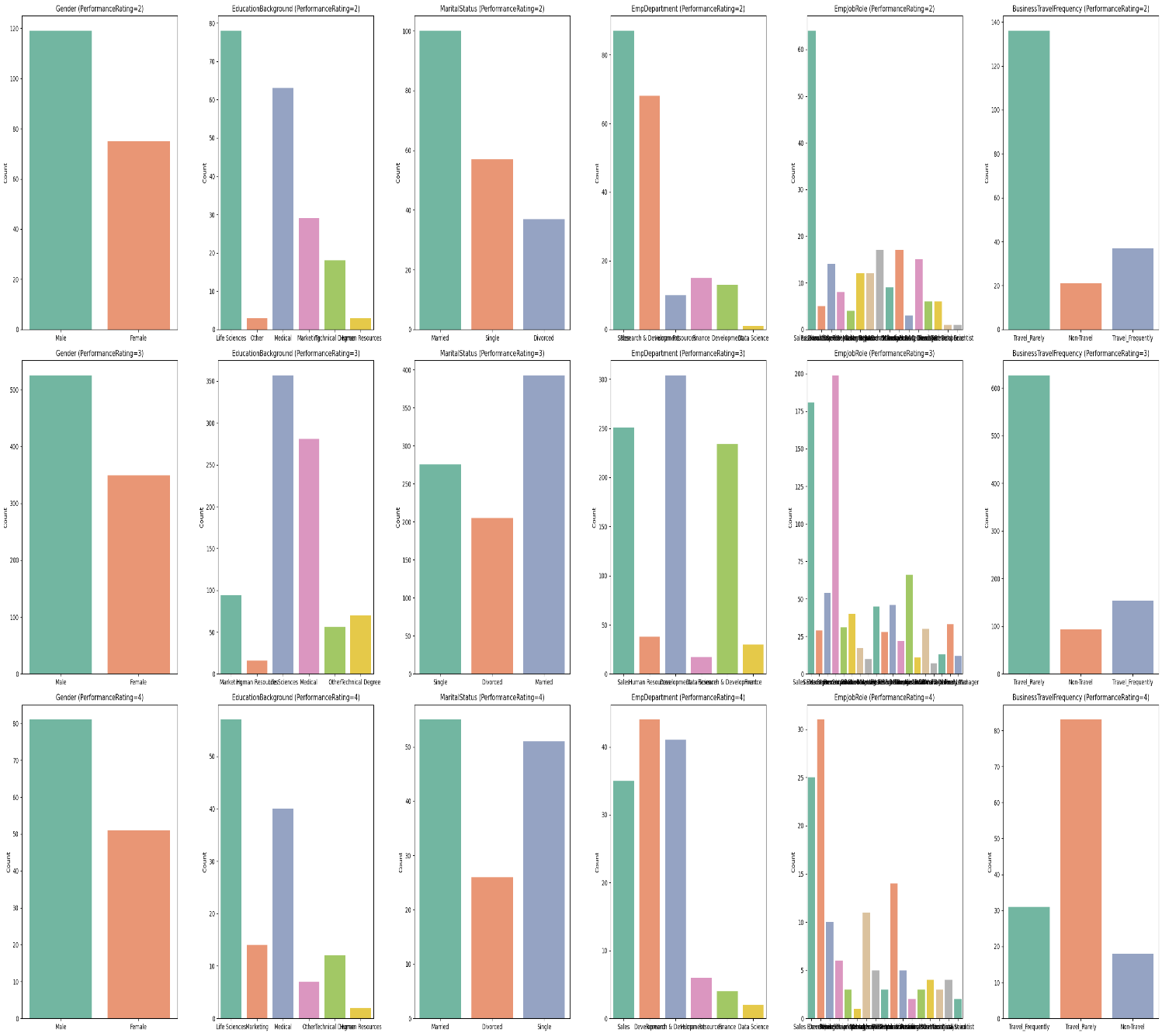
**EmpRelationshipSatisfaction** Employees with relationship satisfaction ratings of 3 have the highest performance rating of 3. This indicates a positive correlation between relationship satisfaction and better performance.

**TrainingTimesLastYear** Employees who underwent training 2 times last year, followed by 3 times, have the highest performance rating of 3. This implies that moderate to frequent training correlates with better performance.

**EmpWorkLifeBalance** Employees with a work-life balance rating of 3 have the highest performance rating of 3. This suggests that employees who perceive a balance between work and personal life tend to perform better.



**CATEGORICAL (VS) PERFORMANCE RATING**



**INSIGHTS**

**Gender** The distribution of performance ratings across genders can reveal if there is a gender bias in performance evaluations.

**EducationBackground** Understanding how employees from different educational backgrounds are rated can help in assessing if educational diversity impacts performance perceptions.

**MaritalStatus** This plot can show if marital status influences stability and performance ratings, potentially indicating if personal life stability translates into professional performance.

**EmpDepartment** Different departments might have varying benchmarks for performance ratings. This plot helps in identifying if certain departments rate their employees more stringently or leniently.

**EmpJobRole** Similar to departments, this can show if specific job roles are associated with higher or lower performance ratings, which might reflect on the expectations and pressures associated with those roles.

**BusinessTravelFrequency** Employees who travel more or less frequently might experience different stress levels and work challenges, which could affect their performance ratings.

**EmpEducationLevel, EmpEnvironmentSatisfaction, EmpJobInvolvement, EmpJobLevel, EmpJobSatisfaction** These plots can highlight how intrinsic job factors and personal employee satisfaction levels correlate with performance ratings.

**NumCompaniesWorked** Insights from this plot can indicate if having experience in multiple companies affects performance positively or negatively.

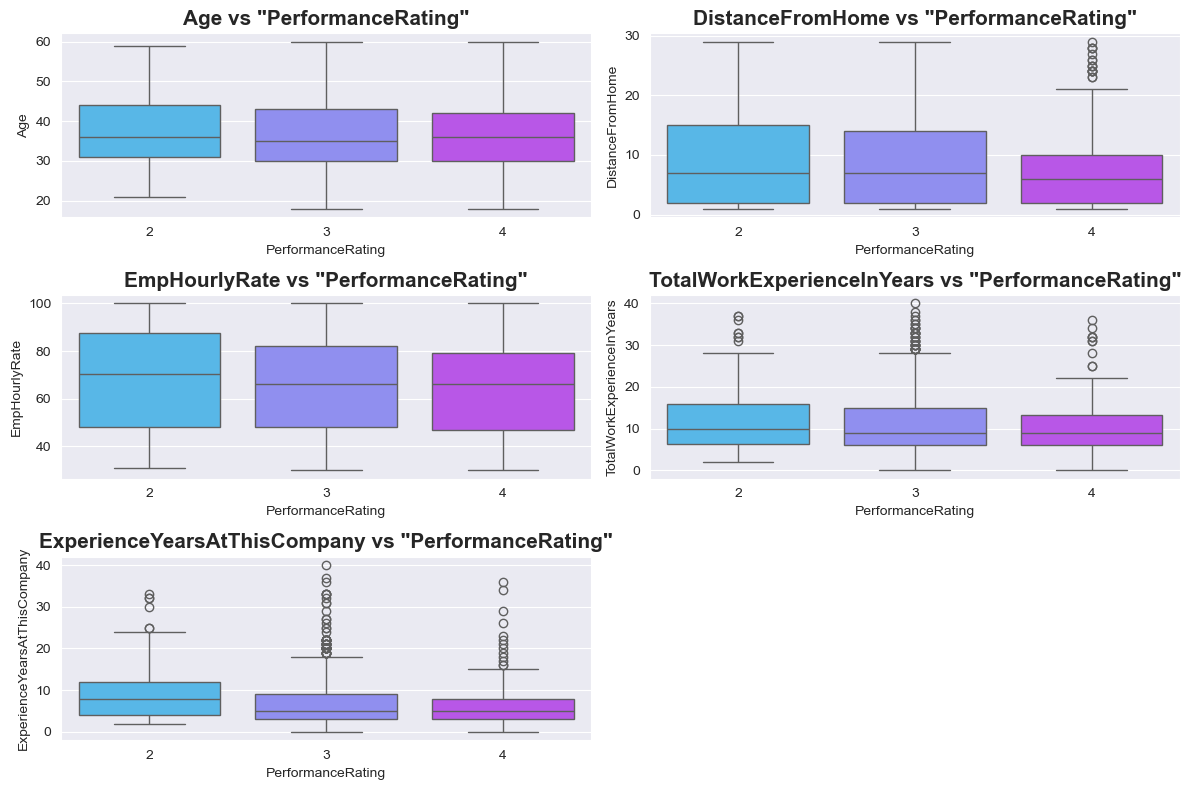
**OverTime** Overworking can either be seen as a sign of dedication or a route to burnout. This plot can help understand how overtime is affecting employee performance ratings.

**EmpLastSalaryHikePercent** This could show if salary hikes are aligned with performance ratings, potentially indicating if financial rewards are being used effectively as a motivational tool.

**EmpRelationshipSatisfaction, TrainingTimesLastYear, EmpWorkLifeBalance** These factors contribute to an employee's overall work satisfaction and could directly impact their performance ratings.

**ExperienceYearsInCurrentRole, YearsSinceLastPromotion, YearsWithCurrManager** These plots can provide insights into career progression and its impact on performance ratings.

**CONTINOUS COLUMNS (VS) PERFORMANCE RATING**

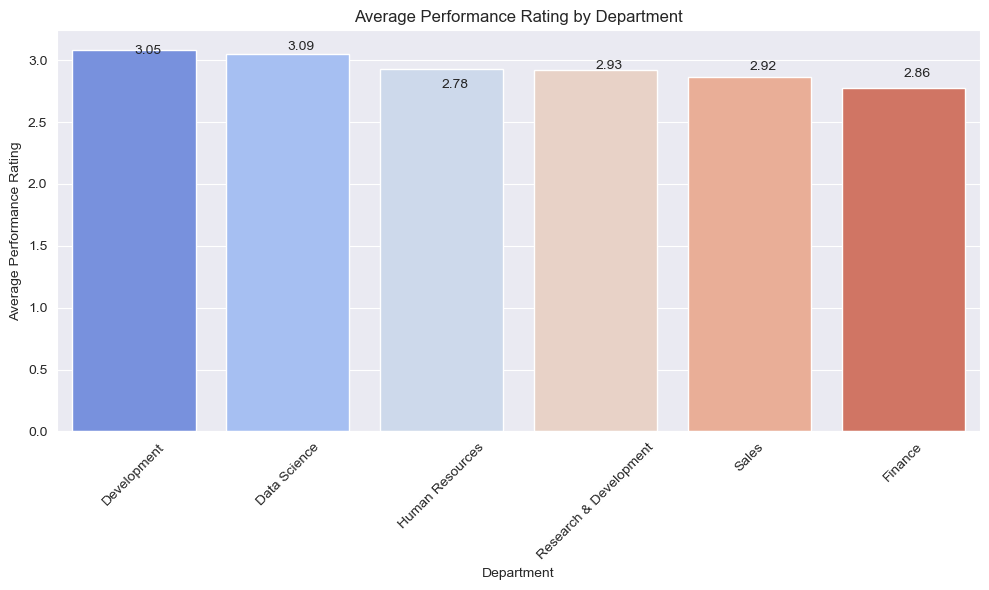


**Insights from bivariate analysis:**

* **EmpEducationLevel vs PerformanceRating**: Employees with higher education levels tend to have higher performance ratings, suggesting a positive correlation between education and performance.
* **EmpEnvironmentSatisfaction vs PerformanceRating:** Higher environment satisfaction correlates with higher performance ratings, indicating that a better work environment may enhance employee performance.
* **EmpJobInvolvement vs PerformanceRating**:Greater job involvement is associated with higher performance ratings, supporting the idea that more engaged employees perform better.
* **EmpJobLevel vs PerformanceRating:** Higher job levels generally see higher performance ratings, possibly reflecting more responsibilities and expectations being met.
* **NumCompaniesWorked vs PerformanceRating**: Employees who have worked with many companies do not necessarily have higher performance ratings, suggesting that frequent company changes might not impact performance significantly.
* **EmpLastSalaryHikePercent vs PerformanceRating:** Significant salary hikes are often given to employees with higher performance ratings, indicating a reward system based on performance.
* **TrainingTimesLastYear vs PerformanceRating**: More training does not clearly correlate with higher performance ratings, suggesting that the quality or relevance of training might be more important than quantity.
* **ExperienceYearsInCurrentRole vs PerformanceRating**: Longer tenure in the current role tends to correlate with higher performance ratings, possibly due to greater experience and familiarity with the role.
* **YearsWithCurrManager vs PerformanceRating**: Longer durations with the same manager are associated with higher performance ratings, which might reflect stable and effective managerial relationships.
* **Attrition vs PerformanceRating:** Employees who are not leaving the company generally have higher performance ratings, which could indicate satisfaction and commitment influencing performance.
* **Gender vs PerformanceRating:** The distribution of performance ratings across genders shows no significant bias, suggesting equitable performance evaluation processes.
* **EducationBackground vs PerformanceRating:** Employees from different educational backgrounds have varying performance ratings, with some backgrounds like 'Marketing' showing higher ratings, possibly reflecting the relevance of education to job roles.
* **MaritalStatus vs PerformanceRating**: Marital status shows varied performance ratings, with married employees slightly more likely to have higher ratings, potentially indicating stability.
* **EmpDepartment vs PerformanceRating**: Certain departments like 'Development' and 'Sales' have higher performance ratings, which might reflect department-specific performance criteria or work dynamics.
* **EmpJobRole vs PerformanceRating**: Specific job roles within departments show different performance ratings, highlighting the impact of job role expectations and responsibilities on performance.
* **BusinessTravelFrequency vs PerformanceRating**: Frequent travelers have varied performance ratings, suggesting that travel demands do not uniformly affect employee performance.

#### **DEPARTMENT WISE PERFORMANCE ANALYSIS:**

The study evaluates worker performance in many departments in order to spot any differences or patterns. When in case of Average performance, the below image

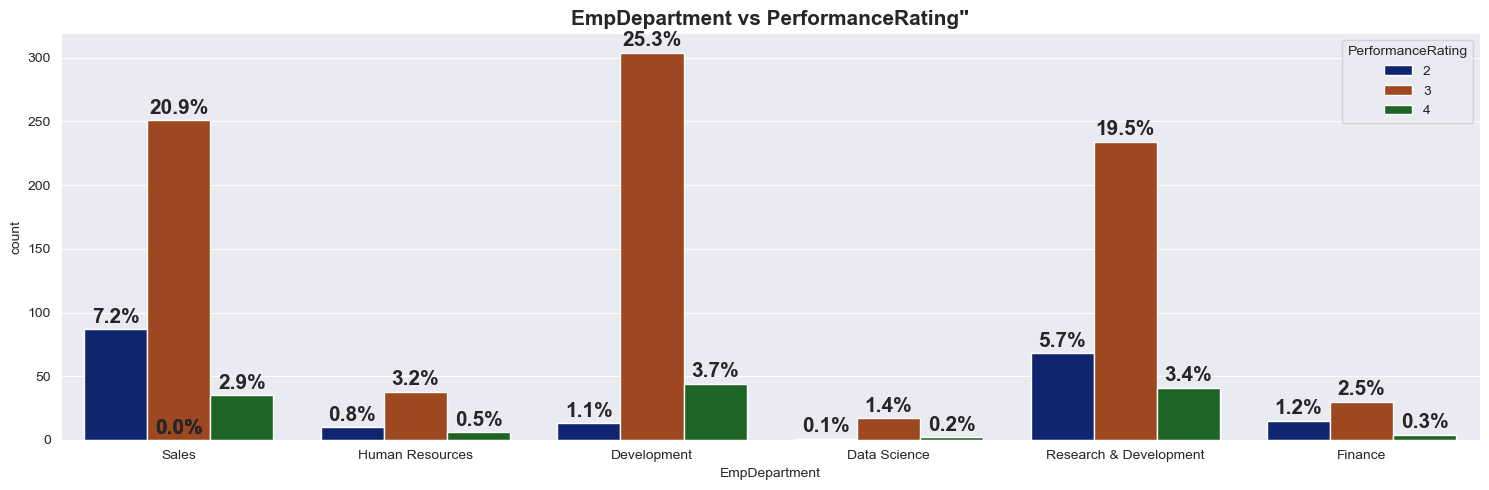


Implies **Data Science** group has high average performance.

**Insights:**

* High Performing Departments: Data Science has the highest average performance rating at 3.09, closely followed by Development at 3.05. This suggests that these departments are likely meeting or exceeding performance expectations, possibly due to effective management, skilled employees, or alignment with business goals.
* Mid-Range Performance: Human Resources, Research & Development, and Sales have performance ratings ranging from 2.78 to 2.93. These departments are performing adequately but might benefit from targeted improvements or interventions to boost their ratings closer to the leading departments.
* Lower Performance: Finance shows the lowest average performance rating at 2.86. This department might be facing challenges that could be impacting its performance, such as resource constraints, alignment with business objectives, or employee engagement issues.

When In case of Numbers comparison, the below image



Implies Development Team or Department Has Highest Performance Rating.

**Insights:**

* In terms of number, the development department outnumbers the other departments.
* Employees in the development department score highly on performance evaluations, with a resounding 25% of ratings of 2.
* Therefore, high-performing personnel are found in departments like development, sales, and research & development.

### **FEATURE SELECTION/ FEATURE ENGINEERING:**

Feature selection techniques determine the relative importance of different features in predicting employee performance.

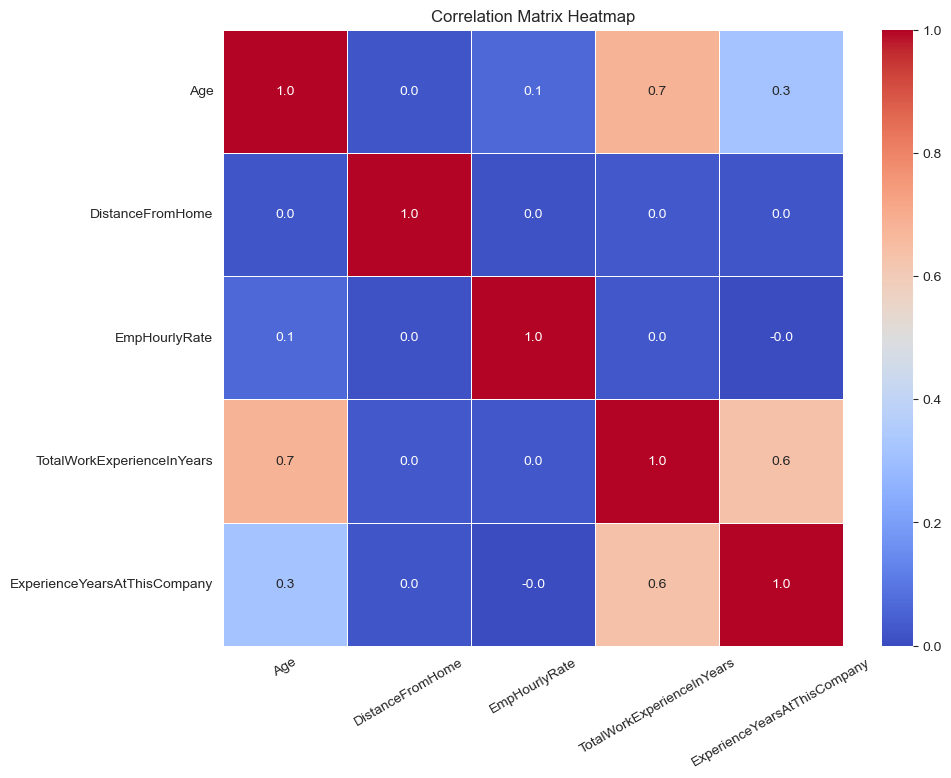
These techniques:

1. Ensure that visual analysis results are rigorously validated through appropriate methodologies and techniques.
2. Continuously monitor and evaluate analysis outcomes to ensure consistency and validity over time.

#### **THE TECHNIQUES USED IN THIS PROJECT ARE:**

#### **Correlation analysis:**

Pearson’s correlation coefficient is calculated to find out the correlation coefficients between two continuous variables.



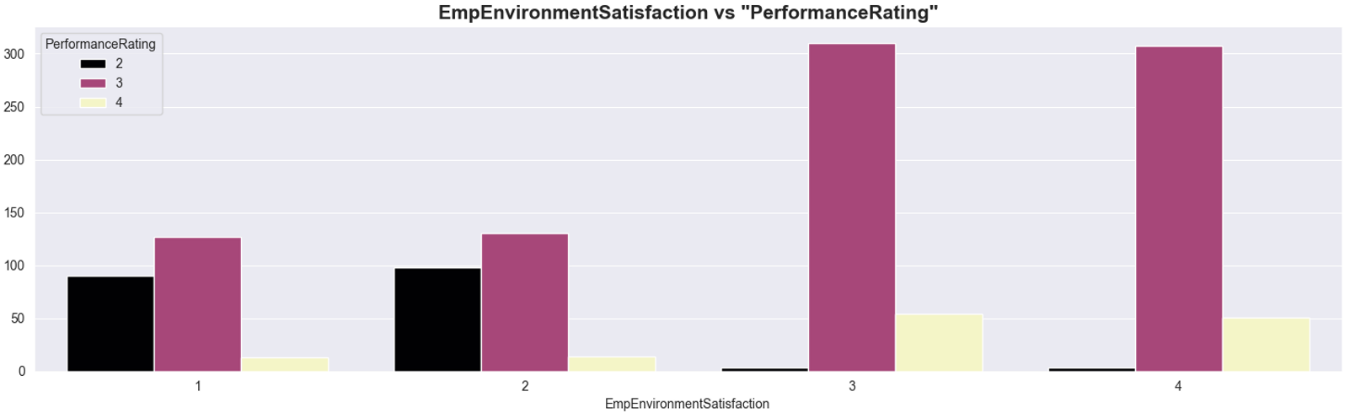
#### **TOP 3 FACTORS AFFECTING THE EMPLOYEE’S PERFORMANCE:**

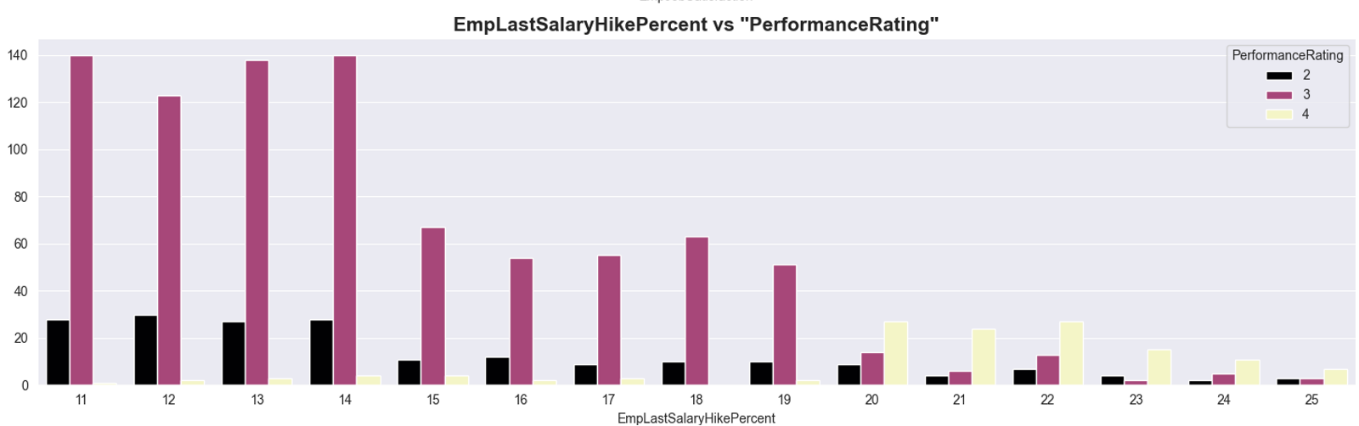
From the data analysis and the feature engineering, there are some of the factors which proved crucial in predicting the employee’s performance.

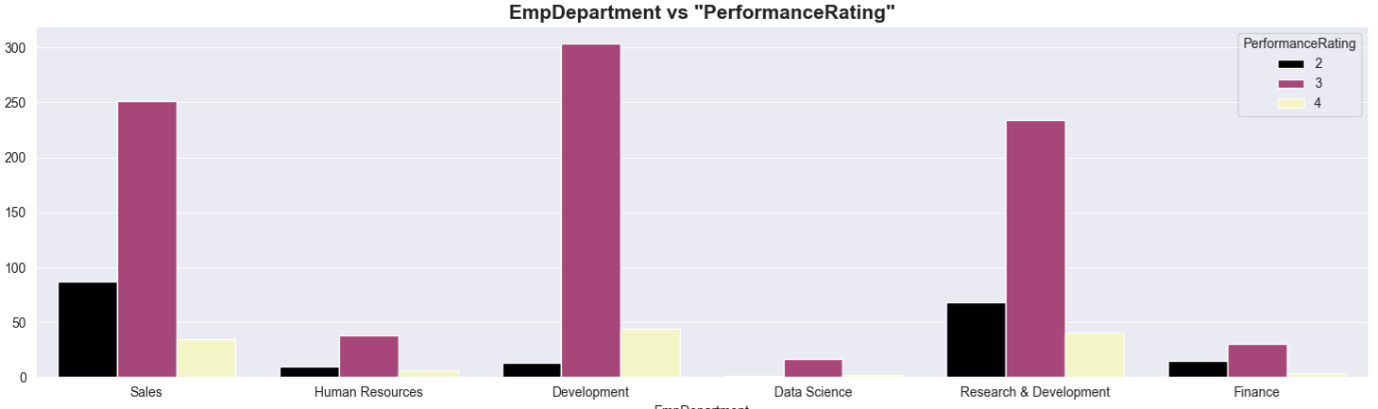
**Among them the top three factors are as factors:**

1. EmpEnvironmentSatisfaction
2. EmpLastSalaryHikePercent
3. EmpDepartment\_Development

**The visual analysis of these three factors are as follows:**







**Insights:**

### There exists a positive correlation between the aim and work environment satisfaction. Employee performance ratings rise in tandem with an increase in workplace happiness.

### There is a negative correlation between the feature "EmpLastSalaryHikePercent" and the target. For a defined range of wage raise percent of 11–14%, the performance rating is consistently extremely high. Thereafter, the performance falls off. Other elements that may contribute to this include the number of departments and their productivity, the scarcity of high-level posts, etc.

### In comparison to other departments, the employee's department of development has a high-performance rating.

## **PREPROCESSING TCHNIQUES USED**

### **ENCODING TECHNIQUES:**

binary\_features= ['OverTime', 'Attrition']

categorical\_features= ['Gender', 'EducationBackground', 'MaritalStatus', 'EmpDepartment', 'EmpJobRole', 'BusinessTravelFrequency']

These features are the categorical types to be encoded in the dataset.

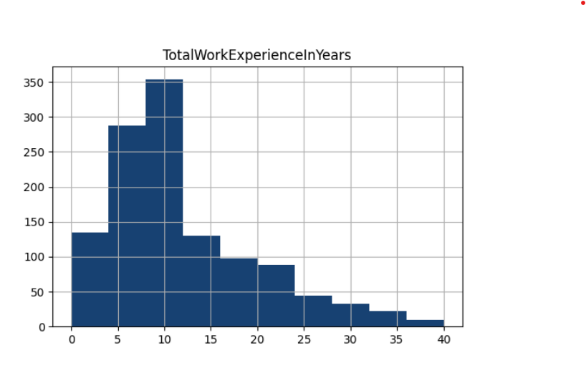
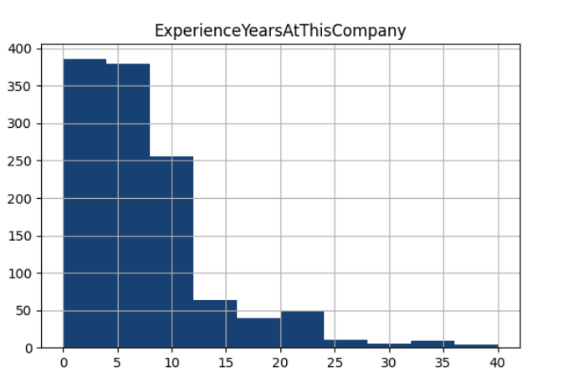
### **OUTLIERS HANDLING:**

**Features with outliers are as follows:**

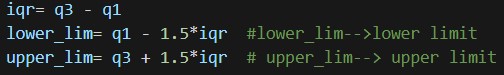
* TotalWorkExperienceInYears
* ExperienceYearsAtThisCompany

**Technique used to handle the outliers:**

**Interquartile method (IQR)** is used to find the outliers since these features are having skewed distribution.



**The formulae used in the IQR method are as follows:**



***# iqr🡪 inter-quartile range***

***# q1🡪 25th percentile***

***# q3🡪 75th percentile***

After finding the outliers, it would be replaced by the **median value** of the corresponding feature or column.

**The techniques utilized are as follows:**

* **For binary features:**
* Manual encoding is done, typically mapping is used ***(“No”🡪 0, “Yes”🡪 1)***
* **For nominal features:**
* One-hot encoding is done which assigns binary value to a category with the rest of the categories as zero in a categorical feature.

### **SCALING:**

* There are two types of scaling in common practice.
* They are minmax scaler and the standard scalerMethods.
* However, in this project, **MinMaxScaler** is utilized from **scikit library**.

**MinMaxScaler:**

Scales the features to a specified range, typically between 0 and 1.

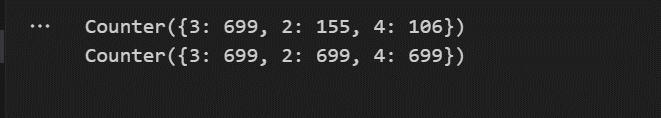
**The continuous features that are scaled using MinMaxScaler are as follows:**

1. TotalWorkExperienceInYears
2. ExperienceYearsAtThisCompany

### **BALANCING THE DATASET:**

* Oversampling technique is used to balance the classes of the target variable.
* **SMOTE function** is used from **imblearn library**, which is one of the oversampling techniques.
* This algorithm helps to overcome the overfitting problem posed by random oversampling.

**Classes of the target of the training sample before and after oversampling:**



## **ALGORITHMS AND TRAINING METHODS USED**

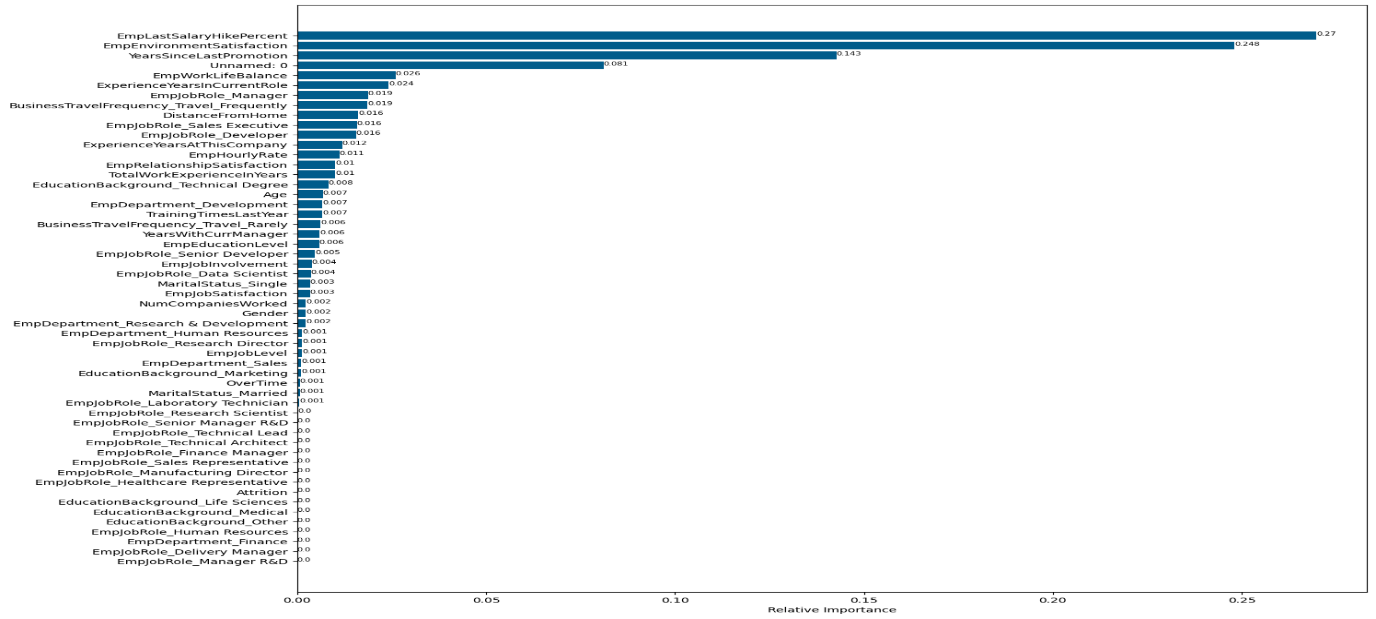
### **Algorithms used in this project:**

* Decision Tree classifier
* Random Forest classifier
* Support Vector Machine
* Extreme Gradient Boosting

### **Model Training Methods:**

#### **DECISION TREE CLASSIFIER:**

* This algorithm works on the principle of if-else condition and make decisions by partitioning the input features into regions and assigning a class label to each region.
* It is preferred in this dataset due to its flexibility and easy model interpretability.
* Since the given dataset is medium-sized, decision tree can handle it well.



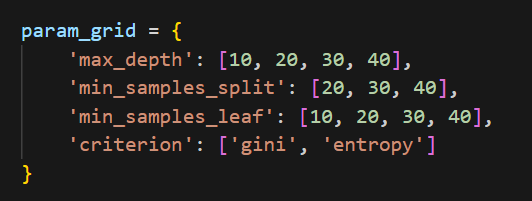
From the **feature importances** arrived at by this model, it is clear that the top three relevant features are:

* EmpLastSalaryHikePercent
* EmpEnvironmentSatisfaction
* YearsSinceLastPromotion

#### **TUNED DECISION TREE CLASSIFIER:**

This model refers to a decision tree algorithm that has been optimized for better performance.

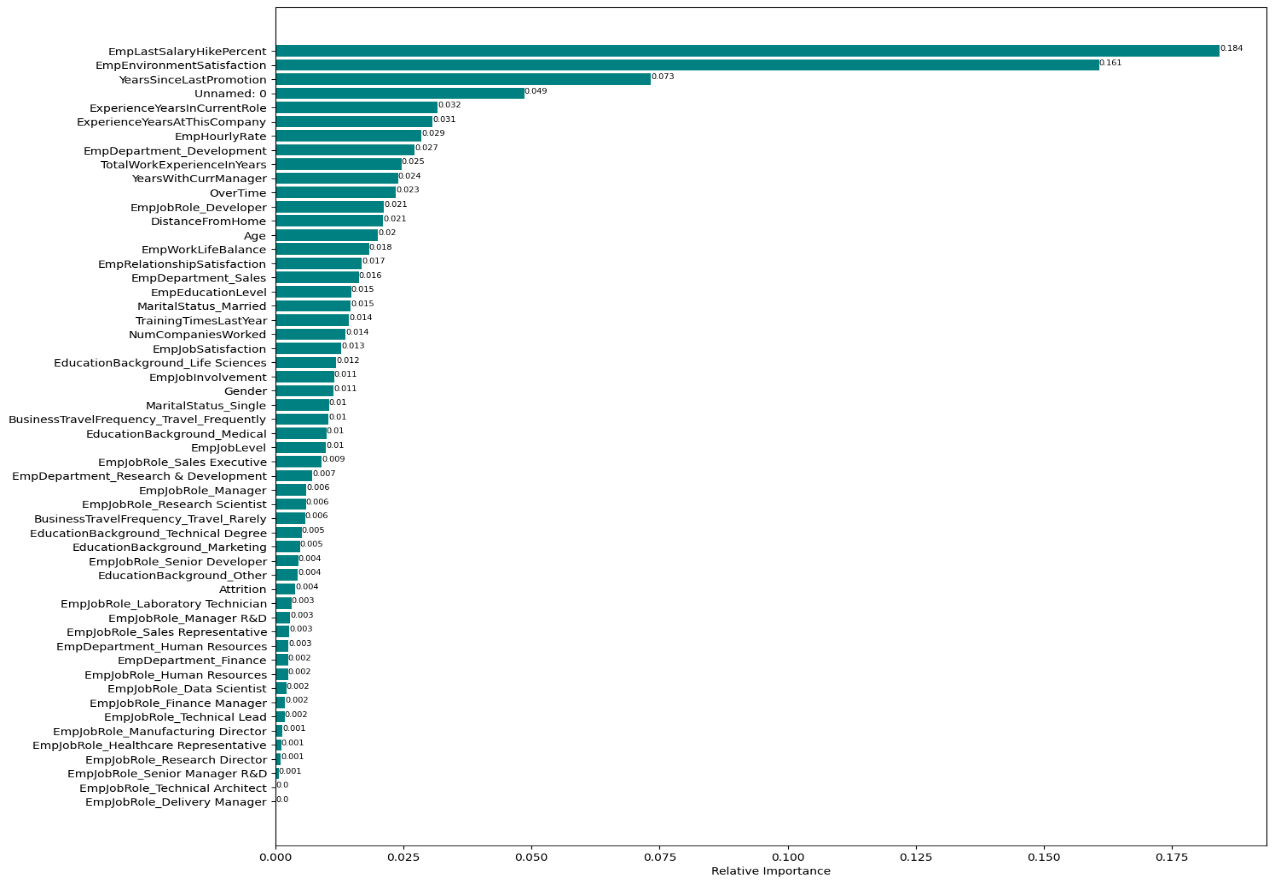
The hyperparameters tuned in this model are as follows:



Cross validation technique used is **Grid Search CV** which exhaustively searches through a specified subset of hyperparameter combinations.

#### **RANDOM FOREST CLASSIFIER:**

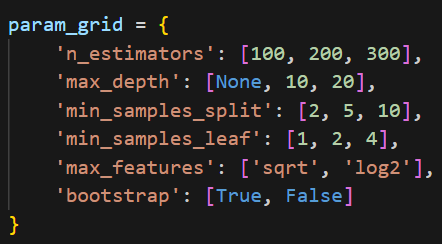
* Random forest is an ensembling learning method that functions by constructing multiple decision trees during training and outputting the majority of the classes (classification) or the average prediction (regression) of the individual trees.
* This model is used here due to it’s ability to handle High-dimensional data, imbalanced datasets and capability of capturing complex relationships in the dataset.
* Moreover, it overcomes the overfitting issue of the decision tree model.



The feature importances in this model also gives the same top three features as the relevant features similar to that of the decision tree classifier.

#### **TUNED RANDOM FOREST CLASSIFIER:**

The hyperparameters tuned in this model are as follows:



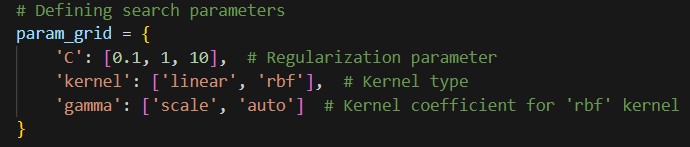
Cross validation technique used is **Grid Search CV** which exhaustively searches through a specified subset of hyperparameter combinations.

1. **SUPPORT VECTOR MACHINE:**

* A **Support Vector classifier** (SVC) works by finding the hyperplane that best separates the classes in the feature space.
* This model is trained in this project due to suitability to handle high-dimensional data.

1. **TUNED SUPPORT VECTOR CLASSFIER:**

The hyperparameters tuned in this model are as follows:



Cross validation technique used is **Grid Search CV** which exhaustively searches through a specified subset of hyperparameter combinations.

1. **TUNED XG BOOSTING:**

* XGBoost (eXtreme Gradient Boosting) is an efficient implementation of gradient boosting which builds a series of weak learners, with each new learner correcting errors made by the previous ones.
* This contains in-built regularization techniques and pruning to prevent overfitting and control the model complexity.
* This model is highly scalable due to parallelization and distributed computing.

