

| Section | Description |
|---------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Data Overview | Collect and describe the data available for predicting promotions. This might include employee demographics, performance scores, tenure, education, department, previous promotions, and other relevant features. |
| Univariate Analysis | <p>Explore each variable individually to understand its distribution and key statistics.</p> <ul style="list-style-type: none"> - Calculate and plot mean, median, mode, standard deviation, histograms, and box plots for variables like age, performance score, tenure, etc.. |
| Bivariate Analysis | <ul style="list-style-type: none"> - Investigate relationships between pairs of variables. - Use correlation coefficients and scatter plots to explore the relationship between variables like tenure and promotion status, performance score and promotion status, etc. - Example: A scatter plot of performance scores vs. the number of promotions received. |

| | |
|-----------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Multivariate Analysis | <p>Explore patterns and relationships involving multiple variables.</p> <ul style="list-style-type: none"> - Use techniques like multiple regression analysis, logistic regression, or machine learning models (e.g., decision trees, random forests) to understand how combinations of variables |
|-----------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|

Data Collection and Preprocessing Phase

| | |
|---------------|-------------------------------------------------------------------|
| Date | 10 JUNE 2024 |
| Team ID | 740008 |
| Project Title | Human resource management: predicting employee promotion using ML |
| Maximum Marks | 6 Marks |

Data Exploration and Preprocessing Template

Identifies data sources, assesses quality issues like missing values and duplicates, and implements resolution plans to ensure accurate and reliable analysis.

| | |
|--------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| | predict promotion. |
| Outliers and Anomalies | <p>Identify and address outliers that could skew the analysis.</p> <ul style="list-style-type: none"> - Use techniques such as Z-scores, IQR, or robust statistical methods to detect and handle outliers. |
| Data Preprocessing Code Screenshots | |
| Loading Data | <pre> > df=pd.read_csv("/content/emp_promotion (1).csv") print('shape of train data {}'.format(df.shape)) df [0] ... shape of train data (54808, 14) </pre> |
| Handling Missing Data | <pre> > df.isnull().sum() [17] ... department 0 education 2409 gender 0 no_of_trainings 0 age 0 previous_year_rating 4124 length_of_service 0 KPIs_met >80% 0 awards_won? 0 avg_training_score 0 is_promoted 0 dtype: int64 </pre> |
| Data Transformation | <pre> > handling outliers q1 = np.quantile(df['length_of_service'],0.25) q3 = np.quantile(df['length_of_service'],0.75) IQR = q3-q1 upperBound = (1.5*IQR)+q3 lowerBound = (1.5*IQR)-q1 print('q1 :',q1) print('q3 :',q3) print('IQR :',IQR) print('upper Bound :',upperBound) print('Lower Bound :',lowerBound) print('skewed data :',len(df[df['length_of_service']>upperBound])) [22] ... q1 : 3.0 q3 : 7.0 IQR : 4.0 upper Bound : 13.0 Lower Bound : 3.0 skewed data : 3489 </pre> |

Save Processed Data

```
▶  
[41] pickle.dump(rf,open('model.pkl','wb'))
```