**Analysis:**

**About INX Future Inc:** INX Future Inc (INX), a global leader in data analytics and automation solutions with over 15 years of industry presence, has consistently ranked among the top 20 employers for five years. This success is attributed to their commitment to employee-centric HR policies.

**Challenge:** In recent years, INX has encountered a decline in employee performance indicators, raising concerns among senior management.

**CEO's Dilemma:** Mr. Brain, the CEO, recognizes the issue but is cautious about penalizing underperforming employees, fearing potential negative impacts on overall employee morale and the company's reputation as a premier employer.

**Proposed Solution:** Mr. Brain, a data scientist himself, initiated a data science project to analyze employee data and uncover the root causes of performance issues. The project aims to provide actionable insights that guide decision-making while safeguarding employee morale and the company's esteemed reputation.

**Key Project Highlights:**

* Dataset with 1200 employee records and 28 features, including 16 qualitative and 11 quantitative features.
* Data meticulously cleaned, revealing no null values, duplicates, or NaN entries.
* Exploratory Data Analysis revealed critical insights.

**Categorical Features**

These values classify the samples into sets of similar samples. Within categorical features are the values nominal, ordinal, ratio, or interval based. The categorical features as follows,

* Gender
* EducationBackground
* MaritalStatus
* EmpDepartment
* EmpJobRole
* BusinessTravelFrequency
* EmpEducationLevel
* EmpEnvironmentSatisfaction
* EmpJobInvolvement
* EmpJobLevel
* EmpJobSatisfaction
* OverTime
* EmpRelationshipSatisfaction
* EmpWorkLifeBalance
* Attrition
* PerformanceRating

**Distribution of Categorical Features:**

* Gender ratio: 60% male, 40% female.
* Six unique educational backgrounds.
* Nineteen unique employee job roles.
* Predominance of education level 3.
* High job satisfaction levels.
* 85% of employees without attrition.
* Only 11% achieved level 4 performance rating.
* 30% of employees doing overtime.

Numerical Features

These values change from sample to sample. Within numerical features are the values discrete, continuous, or timeseries based. The Numerical Features as follows,

- Age

- DistanceFromHome

- EmpHourlyRate

- NumCompaniesWorked

- EmpLastSalaryHikePercent

- TotalWorkExperienceInYears

- TrainingTimesLastYear

- ExperienceYearsAtThisCompany

- ExperienceYearsInCurrentRole

- YearsSinceLastPromotion

- YearsWithCurrManager

**Distribution of Numerical Features:**

* Age distribution peaks between 30 and 40.
* Majority of employees live within 0-5 units from work.
* Employees have worked in up to 8 companies, but most worked in 2.
* Hourly rate ranges from 65 to 95.
* Most employees have up to 5 years of company experience.
* Salary hike percentages are typically 11% to 15%.Top of Form

**Data Visualization:**

Data visualization helps machine learning analysts to better understand and analyze complex data sets by presenting them in an easily understandable format. Visualizations consisting of histograms, box plots, scatter plots, line plots, heatmaps, and bar charts assist in identifying styles, trends, and relationships within the facts.

**Univariate Analysis:**

* Distribution of Categorical Features the Gender variance is divided by 60% of Male employees and 40% of Female employees in the company.
* The number of the educational backgrounds present in the employees is six unique backgrounds.
* nineteen unique employee job roles are present in this company.
* The most of the employees are having the education level of 3.
* The Job satisfaction level in this company is high level for the majority of employees.
* The 85% of employees are not having attrition in their work .
* Only 11% of employees in the company were achieved level 4 - performance rating .
* The overall percentage of employees doing overtime is 30%

Distribution of Numerical Features This helps us determine, among other early insights, how representative is the training dataset of the actual problem domain.

* The age distribution is starting from 18 to 60 where the most of the employees are lying between 30 to 40 age count.
* The distance from home to office is distributing from 0 unit to 30 unit which can be kilometre or mile. The most of the employees are coming from the range of 0 to 5 units.
* Employees are worked in the multiple companies up to 8 companies where most of the employees worked up to 2 companies before getting to work here.
* The hourly rate range is 65 to 95 for majority employees work in this company. •
* In General, Most of Employees work up to 5 years in this company. • Most of the employees get 11% to 15% of salary hike in this company.

**Bivariate Analysis with Scatter Plots:**

* Effective tool for visualizing relationships between numerical variables.
* Revealed insights such as age vs. total experience correlation and distance from home vs. working hours.
* Showed how factors like job satisfaction, tenure, education level, and salary hike rate relate to employee performance.

**Department-Wise Performance Analysis:**

* Development department led with a mean performance rating of approximately 3.09.
* Data Science department followed with a rating of approximately 3.05.
* Notable variations in performance across different departments.

**Multivariant Analysis:**

Multivariate analysis typically involves exploring relationships between multiple variables in a dataset. To do this, various types of graphs and plots can be used to visualize these relationships. Here we used pair plot.

**Label Encoding:**

#### Label Encoding is a technique that is used to convert categorical columns into numerical ones so that they can be fitted by machine learning models which only take numerical data. It is an important pre-processing step in a machine-learning project.

**Correlation Heatmap:**

* Emp Environment Satisfaction (0.396): Employees with higher environmental satisfaction tend to have higher performance ratings, suggesting a positive work environment's impact on performance.
* Emp Last Salary Hike Percent (0.334): Employees receiving higher salary hikes tend to have higher performance ratings, indicating a positive association between compensation increases and performance.
* Emp Work Life Balance (0.124): There is a moderate positive correlation between work-life balance and performance rating. Better work-life balance is associated with higher performance.
* OverTime (0.050): Working overtime has a slight positive correlation with performance rating, implying that longer working hours may lead to higher performance in some cases.
* Emp Department (-0.163): The department an employee works in is negatively correlated with performance rating, suggesting that certain departments may have lower performance ratings.
* Years Since Last Promotion (-0.168): The longer the time since an employee's last promotion, the lower their performance rating tends to be.
* Experience Years In Current Role (-0.148): The number of years an employee spends in their current role negatively correlates with performance rating. Longer tenures in the same role may lead to lower performance ratings.
* Years With Curr Manager (-0.122): The duration of an employee's working relationship with their current manager is negatively correlated with performance rating. Longer relationships with managers may be associated with lower performance ratings.
* Emp Job Role (-0.096): Specific job roles are negatively correlated with performance rating, with certain roles having lower performance ratings.
* Emp Job Level (-0.077): Employee job level is negatively correlated with performance rating, indicating that lower job levels may be associated with lower performance ratings.
* Total Work Experience In Years (-0.068): Total work experience negatively correlates with performance rating, suggesting that more experienced employees may have slightly lower performance ratings.
* Distance From Home (-0.046): The distance an employee lives from their workplace has a slight negative correlation with performance rating. Longer commutes may lead to lower performance ratings.
* Age (-0.040): Age is negatively correlated with performance rating, implying that older employees may have slightly lower performance ratings.
* Attrition (-0.040): Employees who haven't left the organization tend to have higher performance ratings, as indicated by the negative correlation with attrition status.
* Business Travel Frequency (-0.031): Higher business travel frequency negatively correlates with performance rating. Employees who travel more frequently may have lower performance ratings.
* Emp Relationship Satisfaction (-0.020): Employee relationship satisfaction has a slight negative correlation with performance rating, suggesting that those reporting lower relationship satisfaction may have slightly lower performance ratings.
* Marital Status (0.024): Marital status is weakly positively correlated with performance rating, indicating that married employees may have slightly higher performance ratings.
* Num Companies Worked (0.021): The number of companies an employee has worked for is weakly positively correlated with performance rating, suggesting that those with more work experience may have slightly higher performance ratings.

**Outliers:**

In the given dataset numerical columns such as Number of Companies Worked, TrainingTimesLastYear, Total Work Experience in Years, Experience Years at This Company, Experience Years in Current Role, Years Since Last Promotion, Years with Current Manager, Performance Rating have some outliers.

some of the columns are right skewed.

**Handling Outliers:**

Winsorizing or winsorization is the transformation of statistics by limiting extreme values in the statistical data to reduce the effect of possibly spurious outliers. It is averaging method that involves replacing the smallest and largest values of a data set with the observations closest to them. It mitigates the effects of outliers by replacing them with less extreme values.

**Checking Feature Importance:**

**Top 3 correlated features with Performance Rating are**

Emp Environment Satisfaction

Emp Last Salary Hike Percent

Emp Work Life Balance

**Standard Scaler:**

For numerical features, we perform scaling to ensure that all numerical features have a similar scale. This is important because many machine learning algorithms are sensitive to the scale of features. StandardScaler or MinMaxScaler are commonly used for numerical features.

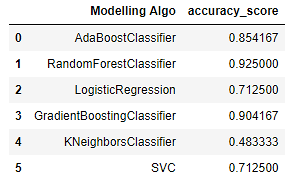
**SMOTE to oversample due to the skewness in target**

Given the evident imbalance in the values within the target variable, we will proceed to address this skewed distribution by applying the SMOTE method using the imbalanced-learn (imblearn) Python package.

SMOTE (Synthetic Minority Over-sampling Technique) is typically applied after splitting the data into training and testing sets. SMOTE is a technique used to address class imbalance, particularly in the training dataset, by generating synthetic examples of the minority class.

**Machine Learning Model Building:**

* Utilized various models like AdaBoost Classifier, Random Forest Classifier, Logistic Regression, Gradient Boosting Classifier, KNeighbors Classifier, Support Vector Classifier (SVC).
* Achieved high accuracy with the KNN as per below



**Random Forest Classifier has highest accuracy\_score of 92.50%**

Random Forest Classifier has achieved highest training accuracy of 1.00 and performed exceptionally well on the test data with an accuracy of 0.925 and Gradient Boosting classifier achieves accuracy of 0.9041. These two are the models which are showing strong performance on both training and test data, suggesting it is a good choice for this dataset.

**Conclusion:** The Employee Performance Analysis project at INX Future Inc is equipped to identify and address performance issues while maintaining the company's reputation as an exceptional employer. By leveraging data analytics, INX can enhance employee performance, attract top talent, and increase client satisfaction, further solidifying its position as an industry leader.