# Employee Performance Evaluation

Analysing key workplace factors and predicting employee performance using Machine Learning

JIFAN THOTTUNGAL

## Objective

This analysis aims to develop a predictive model that evaluates how different factors such as compensation, work environment, and employee growth opportunities correlate with an employee's performance score by leveraging machine learning techniques.

## Methodology

- Data Acquisition
- Exploratory Data Analysis (EDA)
- Data Preprocess
- Model Training & Evaluation
- Hyperparameter Tuning
- Model Deployment

# Understanding The Problem

**Problem Definition:**

Employee performance is a critical factor in organizational success. Various elements, such as salary, team size, and education level, influence an employee's productivity and overall performance. However, identifying the most impactful factors and quantifying their influence remains a challenge.

**Importance of the Analysis:**

This analysis helps optimize workforce productivity by identifying key performance drivers. It provides insights into how salary, training, and workplace factors impact employee satisfaction and retention. By leveraging data-driven insights, HR can make fair decisions on promotions, rewards, and career growth. Additionally, predictive workforce planning enables organizations to invest in high-potential employees and drive long-term success.

# Dataset Overview

**Summary:**

This dataset contains 100,000 rows of data capturing key aspects of employee performance, productivity, and demographics in a corporate environment. It includes details related to the employee's job, work habits, education, performance, and satisfaction. The dataset is designed for various purposes such as HR analytics, employee churn prediction, productivity analysis, and performance evaluation.

**Source:**

The dataset used in this analysis has been sourced from Kaggle.

Employee Performance and Productivity Dataset

**Columns:**

- Employee_ID: Unique identifier for each employee.
- Department: The department in which the employee works (e.g., Sales, HR, IT).
- Gender: Gender of the employee (Male, Female, Other).
- Age: Employee's age (between 22 and 60).
- Job_Title: The role held by the employee (e.g., Manager, Analyst, Developer).
- Hire_Date: The date the employee was hired.
- Years_At_Company: The number of years the employee has been working for the company.
- Education_Level: Highest educational qualification (High School, Bachelor, Master, PhD).
- Performance_Score: Employee's performance rating (1 to 5 scale).
- Monthly_Salary: The employee's monthly salary in USD, correlated with job title and performance score.
- Work_Hours_Per_Week: Number of hours worked per week.

- Projects_Handled: Total number of projects handled by the employee.
- Overtime_Hours: Total overtime hours worked in the last year.
- Sick_Days: Number of sick days taken by the employee.
- Remote_Work_Frequency: Percentage of time worked remotely (0%, 25%, 50%, 75%, 100%).
- Team_Size: Number of people in the employee's team.
- Training_Hours: Number of hours spent in training.
- Promotions: Number of promotions received during their tenure.
- Employee_Satisfaction_Score: Employee satisfaction rating (1.0 to 5.0 scale).
- Resigned: Boolean value indicating if the employee has resigned.

**Target Feature:**

- Performance_Score: Employee's performance rating (1 to 5 scale).

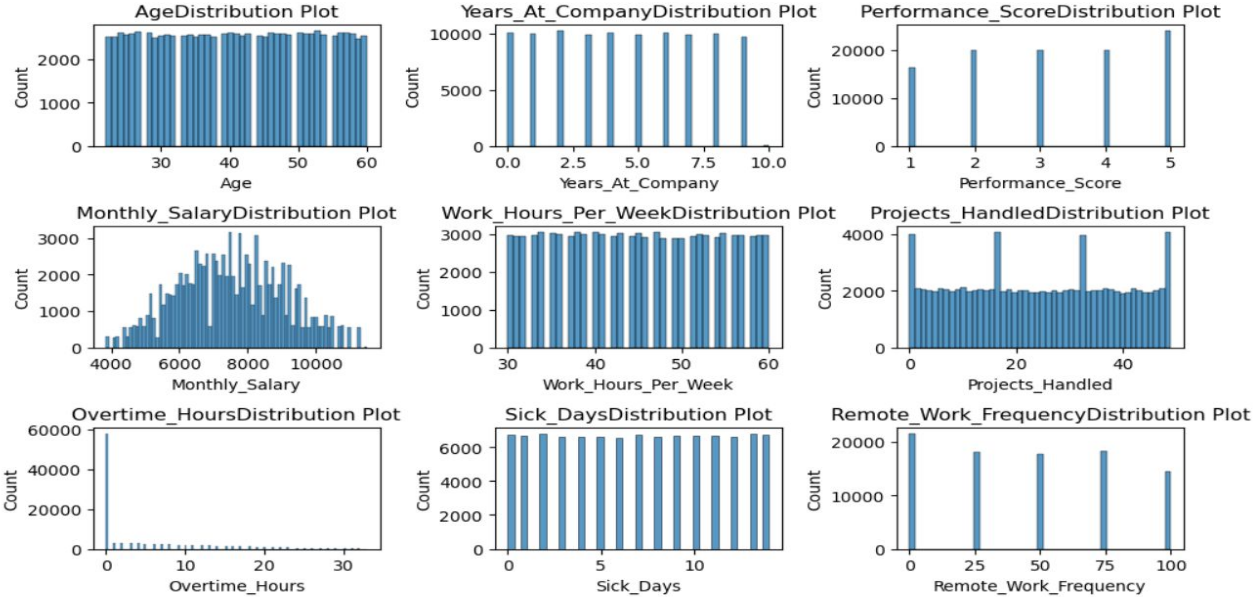
**Dataset Dimension:**

- Rows - 100000 | Columns - 20

# Exploratory Data Analysis (EDA)

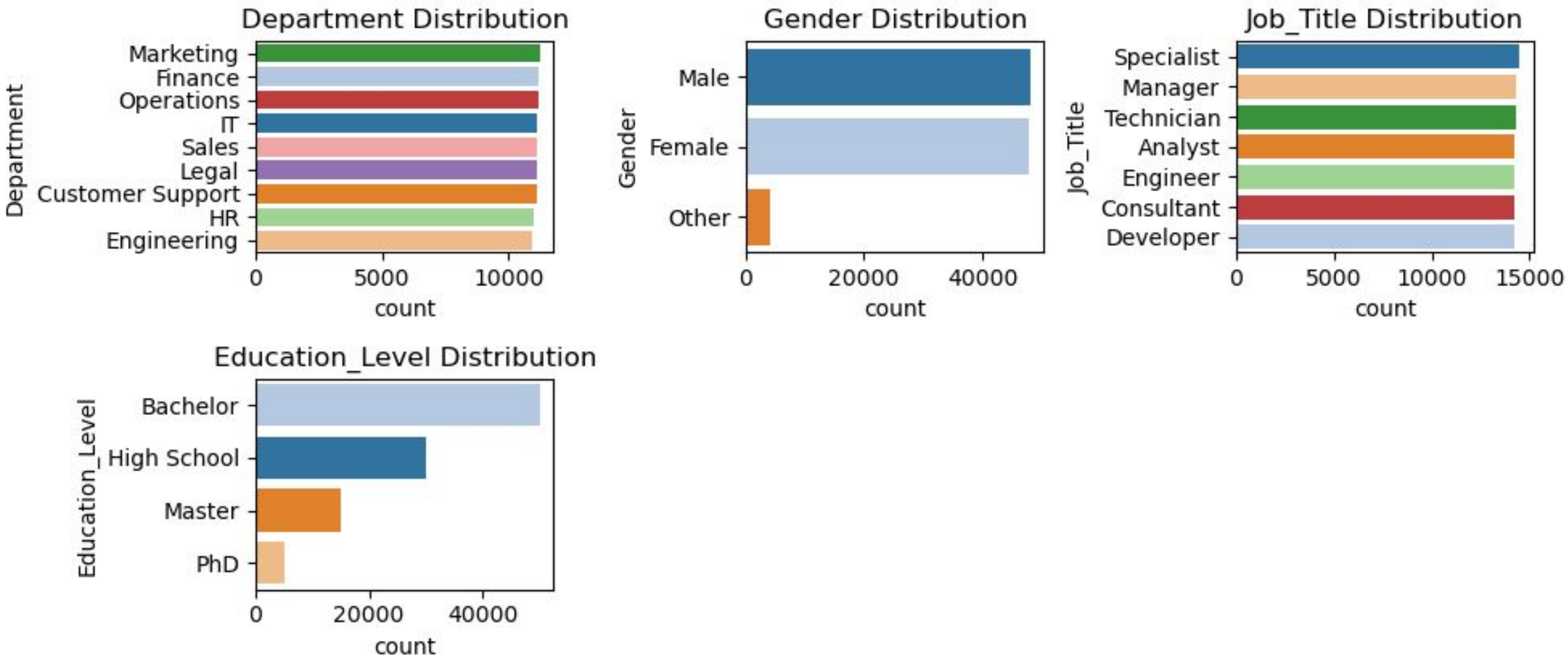
## Distribution Plot:

This visualization displays the distribution of numerical columns, with each subplot providing insights into the distribution of numerical features.



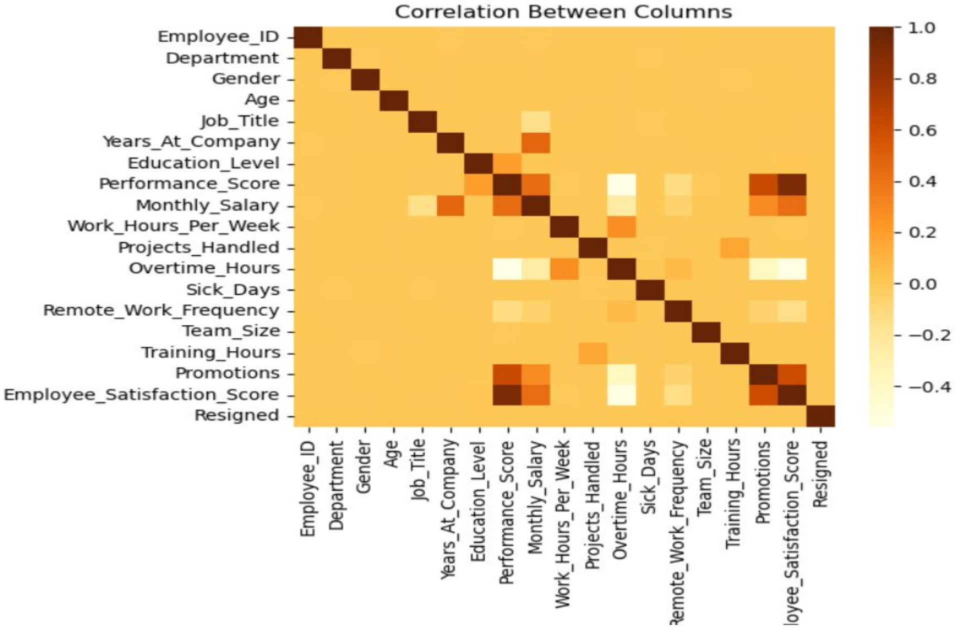
## Distribution Plot:

This visualization presents the distribution of employees across departments, gender, job titles, and education levels. Each subplot provides insights into the workforce composition.



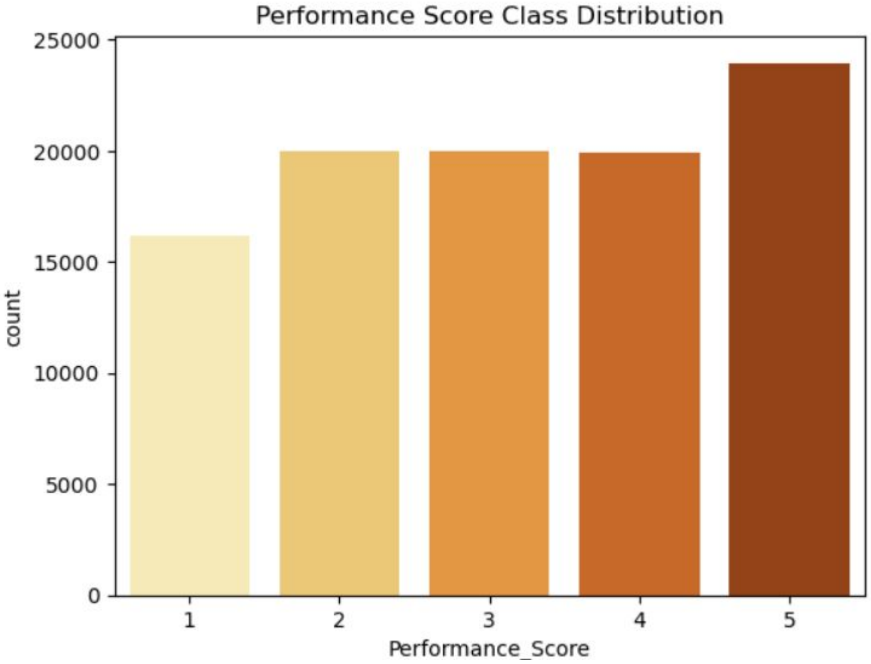
## Heatmap:

This is a correlation heatmap representing the relationships between different numerical variables in the dataset. The heatmap uses a color gradient where darker shades of blue indicate stronger positive correlations, while lighter shades (or white areas) indicate weak or no correlation.



## Bar Chart:

This bar chart illustrates the value counts for each class in the target column, "Performance Score."



# Data Preprocessing

**Handling Missing Values:**

Null values were identified in *Work_Hours_Per_Week, Remote_Work_Frequency,* and *Training_Hours*. To address this, missing values in *Work_Hours_Per_Week* and *Training_Hours* were imputed using the mean, while *Remote_Work_Frequency* null values were filled using the mode.

**Handling Outliers:**

The dataset was analyzed for outliers by using box plot, And treated using IQR method.

**Label Encoding:**

It is a technique that assigns numerical values to categorical data, making it suitable for machine learning models. In this process, features such as *Education_Level* were transformed using label encoding.

### Feature Selection:

Feature selection involves choosing the most important features (columns) from a dataset to enhance model performance and minimize overfitting. I applied the correlation method for this purpose.

### Balancing Data:

The Imbalance Ratio (IR) was calculated and found to be less than 1.5, indicating that balancing the data is not necessary.

### Standard Scaling:

It is a technique where data is transformed to have a mean of 0 and a standard deviation of 1. All the features to be added in the x_train is Standard Scaled.

### Splitting Data:

Train-Test Split is a crucial step in machine learning for assessing model performance. It helps ensure that the model generalizes well to unseen data. The train_test_split function is used to achieve this.

# Model Training & Evaluation

**Model Selection:**

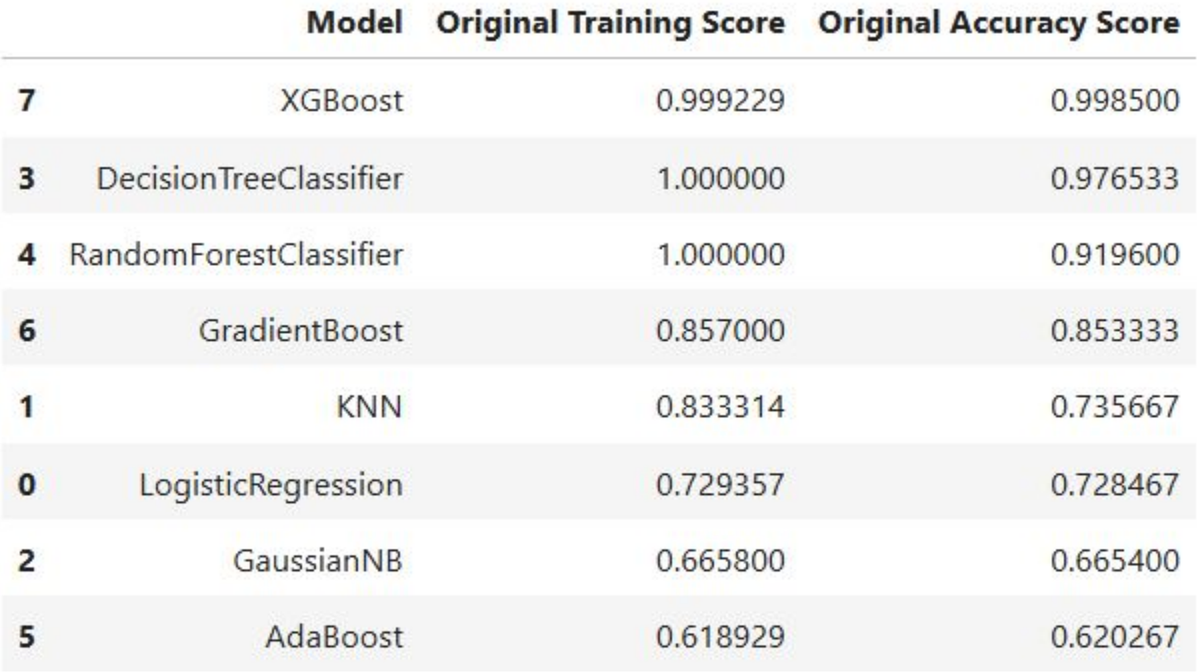
Classification models were used for the analysis since the target variable is categorical.

**Model Training:**

The analysis was conducted using various classification models, including Logistic Regression, KNN, Gaussian Naïve Bayes, Decision Trees, Random Forest, AdaBoost, XGBoost, and Gradient Boosting.

**Model Evaluation:**

Evaluated the models using key metrics such as accuracy, precision, recall and F1-score to assess predictive performance. Additionally, a confusion matrix was used to analyze errors.

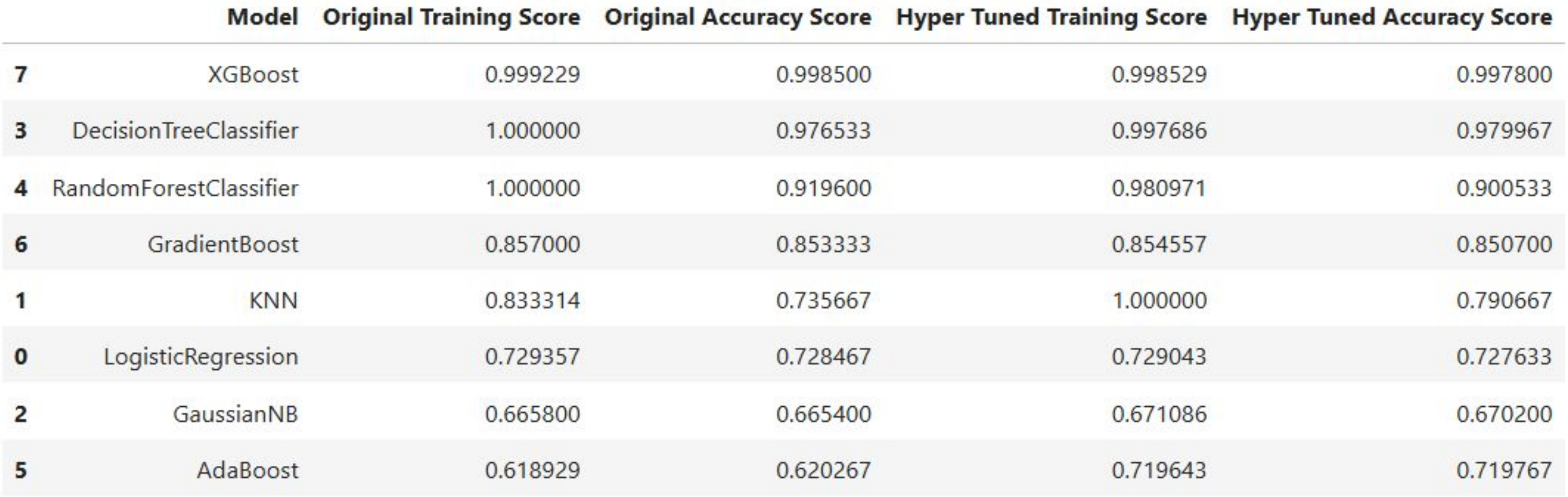


| | Model | Original Training Score | Original Accuracy Score |
|---|---|---|---|
| **7** | XGBoost | 0.999229 | 0.998500 |
| **3** | DecisionTreeClassifier | 1.000000 | 0.976533 |
| **4** | RandomForestClassifier | 1.000000 | 0.919600 |
| **6** | GradientBoost | 0.857000 | 0.853333 |
| **1** | KNN | 0.833314 | 0.735667 |
| **0** | LogisticRegression | 0.729357 | 0.728467 |
| **2** | GaussianNB | 0.665800 | 0.665400 |
| **5** | AdaBoost | 0.618929 | 0.620267 |

Among all the models tested, XGBClassifier achieved the best performance scores.

# Hyperparameter Tuning

Hyperparameter tuning was conducted on Logistic Regression, KNN, Decision Tree, Random Forest, AdaBoost, XGBClassifier, and Gradient Boosting to enhance model performance. GridSearchCV was utilized to identify the optimal parameters.



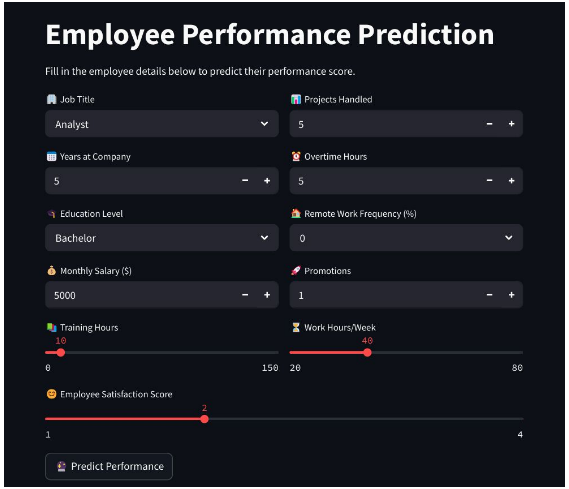
| | Model | Original Training Score | Original Accuracy Score | Hyper Tuned Training Score | Hyper Tuned Accuracy Score |
|---|---|---|---|---|---|
| **7** | XGBoost | 0.999229 | 0.998500 | 0.998529 | 0.997800 |
| **3** | DecisionTreeClassifier | 1.000000 | 0.976533 | 0.997686 | 0.979967 |
| **4** | RandomForestClassifier | 1.000000 | 0.919600 | 0.980971 | 0.900533 |
| **6** | GradientBoost | 0.857000 | 0.853333 | 0.854557 | 0.850700 |
| **1** | KNN | 0.833314 | 0.735667 | 1.000000 | 0.790667 |
| **0** | LogisticRegression | 0.729357 | 0.728467 | 0.729043 | 0.727633 |
| **2** | GaussianNB | 0.665800 | 0.665400 | 0.671086 | 0.670200 |
| **5** | AdaBoost | 0.618929 | 0.620267 | 0.719643 | 0.719767 |

# Model Deployement

The performance prediction app was developed using Streamlit, with the model saved using Joblib. It was then deployed on the Streamlit Community Server via GitHub.

Streamlit App

## User Interface:



Employee Performance Prediction

Fill in the employee details below to predict their performance score.

Job Title: Analyst

Projects Handled: 5

Years at Company: 5

Overtime Hours: 5

Education Level: Bachelor

Remote Work Frequency (%): 0

Monthly Salary ($): 5000

Promotions: 1

Training Hours: 10 (0–150)

Work Hours/Week: 40 (20–80)

Employee Satisfaction Score: 2 (1–4)

# Insights On Project

## Feature Importance:

- Factors like monthly salary, overtime hours, promotions, training hours, and employee satisfaction seem crucial in predicting performance.
- Remote work frequency may have less impact but could still contribute.

## Overfitting Analysis:

To evaluate the possibility of overfitting, the model's training and testing performance were compared. The training accuracy (0.9985) and testing accuracy (0.9978) are nearly identical, indicating strong generalization. Additionally, the model achieved high precision, recall, and F1-scores across all classes, without favoring any specific category. Cross-validation using GridSearchCV also confirmed consistent performance (CV accuracy: 0.9973). These observations suggest that the model is not overfitting and is performing robustly on unseen data.

**Hyperparameter Tuning:**

GridSearchCV was used for optimization, but execution time was high.

# Recommendations On Project

**Optimize Model Training:**

- Instead of exhaustive GridSearchCV, try RandomizedSearchCV or Bayesian Optimization for faster hyperparameter tuning.

**Consider Alternative Models:**

Try LightGBM, or CatBoost, which often outperform traditional models in structured data classification.

# Impact of The Project

**Business Impact:**

Improved HR Decision-Making:

- HR teams can make data-driven decisions on promotions, training, and employee retention strategies.
- Identifies high-potential employees for leadership development programs.

**Social & Workplace Culture Impact:**

Fair & Transparent Evaluations:

- Reduces subjectivity in performance evaluations, promoting fairness in promotions and salary hikes.

Employee Satisfaction & Engagement:

- Organizations can improve work-life balance strategies based on performance trends.
- Helps in creating a more engaged, motivated, and satisfied workforce.

# Conclusion

The Employee Performance Prediction project successfully demonstrates the power of machine learning in evaluating and forecasting employee performance. By leveraging classification models, feature engineering, and hyperparameter tuning, the model provides valuable insights into the key drivers of employee productivity.

With a well-balanced and preprocessed dataset, the best-performing model, XGBoost (XGBClassifier), achieved the highest accuracy of 99%, proving its reliability in real-world applications. The project effectively identifies crucial performance factors such as years at company, promotions, training hours, and employee satisfaction, enabling HR teams to make data-driven decisions on talent management, promotions, and retention strategies.

This solution empowers organizations to make proactive, strategic decisions, leading to a more engaged workforce, improved productivity, and a positive workplace culture. Moving forward, the model can be enhanced with additional behavioral and sentiment analysis features for even deeper insights into employee performance trends.

# Thank You