Business Problem:

The objective was to predict employee performance ratings based on various features.

Data Overview:

- The dataset contains 1200 records and 27 features (after dropping the EmpNumber column).
- Features include demographic data, job-related information, satisfaction scores, and historical data such as the number of companies worked for and years of experience.
- The target variable is Performance Rating, a categorical variable with values 2, 3, and 4.

Data Preprocessing:

1. **Handling Duplicates:**
   - Checked for duplicate records; none were found.
2. **Exploratory Data Analysis (EDA):**
   - Conducted univariate and bivariate analysis to understand the distribution of various features.
   - Analyzed categorical variables through count plots.
3. **Handling Missing Values:**
   - No missing values were found in the dataset.
4. **Handling Imbalanced Data:**
   - Used oversampling techniques to balance the target variable classes.
5. **Feature Selection:**
   - Utilized correlation matrices to identify key features.
   - Selected features based on correlation coefficients greater than 0.1 with Performance Rating.
6. **Data Transformation:**
   - Converted categorical variables into numerical format using label encoding.
7. **Data Scaling:**
   - Applied standard scaling to standardize the features.
8. **PCA (Principal Component Analysis):**
   - Conducted PCA for feature dimensionality reduction.

Model Development:

(I) Logistic Regression

- **Precision:** Ranged from 0.71 to 0.87 for different classes.
- **Recall:** Ranged from 0.70 to 0.83 for different classes.
- **F1-Score:** Ranged from 0.70 to 0.83 for different classes.
- **Accuracy:** 78%.

(II) Support Vector Machine (SVM)

- **Precision**: Ranged from 0.88 to 0.95 for different classes.
- **Recall**: Ranged from 0.81 to 0.99 for different classes.
- **F1-Score:** Ranged from 0.87 to 0.96 for different classes.
- **Accuracy:** 92%.

(III) Decision Tree

- **Precision:** Ranged from 0.88 to 0.99 for different classes.
- **Recall:** Ranged from 0.74 to 1.00 for different classes.
- **F1-Score:** Ranged from 0.85 to 0.95 for different classes.
- **Accuracy:** 91%.

(IV) Random Forest

- **Precision:** Ranged from 0.93 to 1.00 for different classes.
- **Recall:** Ranged from 0.93 to 1.00 for different classes.
- **F1-Score:** Ranged from 0.96 to 1.00 for different classes.

Accuracy: 98%.

(V) XGBoost

- **Precision:** Ranged from 0.95 to 0.99 for different classes.
- **Recall:** Ranged from 0.90 to 1.00 for different classes.
- **F1-Score:** Ranged from 0.94 to 0.98 for different classes.
- **Accuracy**: 97%.

(VI) K-Nearest Neighbor (KNN)

- **Precision:** Ranged from 0.76 to 0.86 for different classes.
- **Recall:** Ranged from 0.56 to 0.95 for different classes.
- **F1-Score:** Ranged from 0.68 to 0.86 for different classes.
- **Accuracy:** 81%..

(VII)Naive Bayes (Bernoulli)

- **Precision:** Ranged from 0.61 to 0.70 for different classes.
- **Recall:** Ranged from 0.61 to 0.75 for different classes.
- **F1-Score:** Ranged from 0.63 to 0.72 for different classes.
- **Accuracy:** 67%.

Cross-Validation Results

Logistic Regression: Mean Accuracy: 67%

KNN: Mean Accuracy: 81%

Decision Tree: Mean Accuracy: 73%

Naive Bayes (Bernoulli): Mean Accuracy: 59%

SVM: Mean Accuracy: 87%

Random Forest: Mean Accuracy: 80%

XGBoost: Mean Accuracy: 85%

Conclusion:

- The Random Forest model achieved the highest accuracy at 98%, indicating outstanding performance in classifying instances into their respective classes. It displayed excellent precision, recall, and F1-scores across all classes.
- XGBoost also performed exceptionally well with an accuracy of 97% and demonstrated a strong balance between precision and recall for all classes.
- Support Vector Machine (SVM) showed robust performance with an accuracy of 92% and maintained high precision, recall, and F1-scores for all classes.
- Decision Tree and K-Nearest Neighbors (KNN) models performed reasonably well with accuracies of 91% and 81%, respectively, and balanced precision and recall metrics.
- Logistic Regression provided a good baseline with 78% accuracy, while Naive Bayes (Bernoulli) showed moderate performance with 67% accuracy.

Based on these results, the Random Forest, XGBoost, and SVM models are recommended for this classification task due to their high accuracy and balanced performance across all classes.