Employee Performance Analysis

INX Future Inc.

- Candidate Name : Raja Prabu Manivel
- Candidate E-Mail : rpthalaprabhu@gmail.com
- REP Name : DataMites™ Solutions Pvt Ltd
- Assesment ID : E10901-PR2-V18
- Module : Certified Data Scientist - Project
- Exam Format : Open Project- IABAC™ Project Submission
- Project Assessment : IABAC™
- Registered Trainer : Ashok Kumar A
- Submission Deadline Date : 24-MAY-2025

Analysis

1. Data Understanding and Exploration

Initial exploration revealed:

Dataset contains a mix of categorical (e.g., Department, Gender, JobRole) and numerical features (e.g., Age, YearsAtCompany, MonthlyIncome).

Target variable: **Performance Rating** or derived classification (e.g., high vs low performer).

Class distribution is somewhat imbalanced, but no severe skew.

Key Observations from EDA:

- **JobSatisfaction**, **JobInvolvement**, and **EnvironmentSatisfaction** are positively correlated with higher performance.
- Features like **OverTime** and **YearsAtCompany** provided actionable variance.
- Low variance or irrelevant features like EmployeeNumber were removed during preprocessing.

---

2. Data Processing Techniques

The following preprocessing steps were taken across notebooks:

Missing Value Handling: Dropped rows with missing values (e.g., NumCompaniesWorked, TotalWorkingYears).

Categorical Encoding:

Label Encoding for binary features (e.g., OverTime, Gender).

One-Hot Encoding for multi-class features (e.g., Department, JobRole).

Feature Scaling:

StandardScaler applied to numerical features for SVM and Logistic Regression models.

Feature Selection:

Correlation matrix used to identify and retain key influencing features.

Low-correlation or identifier columns dropped.

3. Machine Learning Algorithms Considered

Three main algorithms were trained and evaluated:

✅ **Logistic Regression**

- Baseline model
- Fast and interpretable
- Performed decently but struggled with nonlinear patterns

✅ **Random Forest Classifier**

- Performed best among all models
- Provided feature importance for insight generation
- Handled both categorical and numerical features well

✅ **Support Vector Classifier (SVC)**

- Performed well after scaling
- Sensitive to hyperparameters, better with tuned parameters

Each model was evaluated using:

- **Accuracy**
- **Classification Report (Precision, Recall, F1-Score)**
- **Confusion Matrix**

4. Model Selection Rationale

Model	Accuracy	Pros	Cons
Logistic Regression	89%	Simple, interpretable	Lower performance, linear only
Random Forest	99.5%	High accuracy, feature insights	Slightly slower, more complex
SVC	99.4%	Good on scaled data	Requires tuning, less interpretable

Conclusion:

Random Forest was selected as the final model due to the best trade-off between accuracy and interpretability.

The Random Forest model gave 99.58% test accuracy with good generalization capability. Followed a structured machine learning workflow involving data preprocessing, model building, diagnostics and optimizations. The end-to-end implementation, analysis and choice of final model were appropriate.