Employee Attrition

➢ *Objectives: -*

The goal of this project is to analyze employee attrition data to understand the factors contributing to employee turnover and to develop a predictive model that can identify employees who are likely to leave the company. This will help the organization implement effective strategies to reduce attrition and improve employee retention.

➢ *Data Overview: -*

The dataset includes various features related to employee demographics, job roles, satisfaction levels, and other work-related attributes. Key features include:

- Attrition: Whether the employee left the company (Yes/No)
- Age, MonthlyIncome, JobLevel, TotalWorkingYears, YearsAtCompany, etc.
- Categorical features like BusinessTravel, Department, Gender, JobRole, etc.

➢ *Steps follows: -*

Step-1: Load and Inspect Data

Step-2: Data Cleaning and Preprocessing

Step-3: Exploratory Data Analysis (EDA)

Step-4: Data Visualization

Step-5: Model Building and Evaluation

➢ *Technical tools:-*

- Python Programming
- Seaborn,
- Matplotlib,
- Pandas,
- Numy,
- Machine Learning algorithms: Random Forest Classifier

Step-1: Load and Inspect Data: -

Load the data given in the (.csv) file.

Step-2: Data Cleaning and Preprocessing: -

Before performing EDA, let's address any missing values and ensure our data is clean and ready for analysis.

I. *Handle Missing Values:*

"NumCompaniesWorked",'"TotalWorkingYears", "EnvironmentSatisfaction"," JobSatisfaction", and "WorkLifeBalance" have some missing values.

II. *Encode Categorical Variables:*

Convert categorical variables like "Attrition", "BusinessTravel", "Department", "EducationField", "Gender", "JobRole", "MaritalStatus", and "Over18" into numerical values.

III. *Drop Irrelevant Columns:*

Drop columns like "EmployeeCount", "Over18", and "StandardHours" as they have no variance or add little value to the analysis.

Step-3: Exploratory Data Analysis: -

EDA helps us understand the data distribution, relationships between variables, and key patterns. Here are the steps and findings from the EDA:

1. Data Cleaning and Preprocessing:

- ✓ Handling missing values (if any)
- ✓ Encoding categorical variables using Label Encoding
- ✓ Normalizing numerical features

2. Univariate Analysis:

- ✓ **Attrition Distribution**: The dataset is imbalanced with more employees staying than leaving.
- ✓ **Age Distribution**: Age is normally distributed.
- ✓ **Monthly Income**: Income distribution is right skewed with most employees earning in the lower range.

3. Bivariate Analysis:

- ✓ **Attrition vs. Monthly Income**: Employees with lower monthly income are more likely to leave.
- ✓ **Attrition vs. Age**: Younger employees show a higher tendency to leave compared to older employees.
- ✓ **Attrition vs. Job Role**: Some job roles have higher attrition rates (e.g., Sales Executive, Research Scientist).

4. Correlation Analysis:

- ✓ **Features with Strong Correlation**: Identifying features with strong correlations to "Attrition" can guide you in feature selection for machine learning models.
- ✓ **Potential Predictors**: Features like "JobLevel", "MonthlyIncome", or" TotalWorkingYears" might be significant predictors of Attrition.
- ✓ **Redundant Features**: Features that are strongly correlated with each other (like "Age" and "TotalWorkingYears") might be redundant and could be considered for removal to avoid multicollinearity.

Step-4: Data Visualization: -

1. Attrition Distribution:

- The dataset shows that the number of employees who did not leave is higher compared to those who did.
- **Implication**: This indicates an imbalanced dataset, which should be considered when building predictive models.

2. Age Distribution:

- Age is normally distributed.
- **Implication**: Age diversity in the workforce, which is useful for understanding age-related attrition patterns.

3. Monthly Income Distribution:

- Monthly income is right-skewed, with most employees earning in the lower range.
- **Implication**: Income disparity may affect attrition, as seen in later analyses.

4. Attrition vs. Monthly Income:

- Employees with lower monthly incomes are more likely to leave the company.
- **Implication**: Compensation could be a significant factor in employee retention.

5. Attrition vs. Age:

- Younger employees tend to have higher attrition rates.
- **Implication**: Retention strategies might need to be tailored for younger employees.

6. Attrition vs. Job Role:

- Certain job roles, such as Sales Executive and Research Scientist, have higher attrition rates.
- **Implication**: Job-specific interventions might be necessary to address high turnover roles.

7. Correlation Heatmap:

- Shows the relationships between numerical features. High correlation between TotalWorkingYears and YearsAtCompany.
- **Implication**: Redundant features can be considered for removal or further analysis.

8. Feature Importance:

- Key features contributing to attrition include MonthlyIncome, OverTime, Age, and JobLevel.
- **Implication**: Focus on these features for developing retention strategies and improving predictive models.

Feature Engineering:

- Creating new features if necessary (e.g., combining related features or creating ratios)
- Encoding categorical variables using Label Encoding or One-Hot Encoding

Step-5: Model Building:

Several machine learning models can be used to predict employee attrition. In this project, we use a Random Forest Classifier.

Steps:

1. **Splitting Data**: Split the data into training and testing sets.
2. **Training the Model**: Train the Random Forest Classifier on the training data.
3. **Evaluating the Model**: Use the testing data to evaluate model performance using metrics like accuracy, precision, recall, F1-score, and the confusion matrix.

Confusion Matrix and Classification Report:

- ***Confusion Matrix: -***
  - True Positives (TP): Employees correctly predicted to leave.
  - True Negatives (TN): Employees correctly predicted to stay.
  - False Positives (FP): Employees incorrectly predicted to leave.
  - False Negatives (FN): Employees incorrectly predicted to stay.
- **Classification Report: -**
  - **Precision**: Ratio of correctly predicted positive observations to the total predicted positives.
  - **Recall**: Ratio of correctly predicted positive observations to all observations in actual class.
  - **F1-Score**: Weighted average of Precision and Recall.
  - **Support**: The number of actual occurrences of the class in the dataset.

CONCLUSION

1. **Key Drivers of Attrition**:
   - High attrition is observed in lower income brackets, specific job roles (e.g., Sales Executive), and younger employees.
   - Overtime is a significant factor, indicating work-life balance issues.
2. **Model Performance**:
   - The Random Forest model performs well with an accuracy of around 83%.
   - Precision and recall for predicting attrition (employees leaving) are moderately high, indicating a balanced model.

Strategies to Reduce Attrition:

1. **Improve Compensation**:
   - ✓ Evaluate and adjust compensation structures to ensure competitive salaries, particularly for roles with high attrition.
2. **Work-Life Balance**:
   - ✓ Address issues related to overtime. Implement policies that promote a healthy work-life balance, such as flexible working hours and workload management.
3. **Career Development**:
   - ✓ Provide clear career progression paths and professional development opportunities, especially for younger employees and those in high-risk roles.

4. **Job Role-Specific Interventions**:
    - ✓ For roles with high attrition rates, such as Sales Executive and Research Scientist, conduct deeper investigations to understand role-specific challenges and address them.
5. **Employee Engagement**:
    - ✓ Increase engagement through regular feedback, recognition programs, and team-building activities to enhance job satisfaction.
6. **Onboarding and Training**:
    - ✓ Strengthen onboarding programs to better integrate new hires and provide ongoing training to enhance their skills and satisfaction.

By focusing on the insights gained from the EDA and model, the organization can implement targeted strategies to reduce attrition, improve employee satisfaction, and retain valuable talent. Continuous monitoring and updating of the model with new data will further enhance the predictive power and effectiveness of retention strategies.