Employee Resignation Prediction Using Machine Learning

By: Nitin Raj

1. Objective

The goal of this project is to predict whether an employee is likely to resign from the company.
Using a dataset of **100,000 employees**, the model analyzes factors such as:

- Age
- Department
- Job Title
- Salary
- Working hours
- Overtime
- Training hours
- Performance score
- Satisfaction score

This project helps HR teams understand employee behavior and identify early resignation risks.

2. Dataset Summary

- **Rows:** 100,000
- **Columns:** 20
- **Target variable:** `Resigned` (0 = No, 1 = Yes)
- **Missing values:** 0

- **Duplicates:** 0

The dataset is clean and balanced across multiple departments, genders, and job roles.

3. Tools and Technologies Used

Category	Tools
Programming	Python
Data Handling	Pandas, NumPy
Visualization	Matplotlib, Seaborn
ML Models	KNN, Logistic Regression, Neural Network (MLP)
Evaluation	Accuracy Score
Scaling	MinMaxScaler
Encoding	LabelEncoder
Model Saving	Pickle

4. Workflow Summary

```
Load Data → Explore → Clean → Encode → Scale → Train/Test Split
→ Build Models (KNN, Logistic Regression, MLP)
→ Evaluate Accuracy → Save Final Model
```

5. Step-by-Step Implementation

Step 1: Load the Data

Dataset loaded directly from GitHub using pandas.

Step 2: Exploratory Data Analysis

- Checked null values → found **zero missing**.
- Checked duplicates → **0 duplicates**.
- Observed distribution of:

    - Education level
    - Gender
    - Salary
    - Age via boxplot

Step 3: Preprocessing

- **LabelEncoder** used to convert categorical text columns into numeric values:
    - Department
    - Job Title
    - Gender
    - Education Level
    - Resigned
- **Dropped Hire_Date** because it is not useful for training.
- **MinMaxScaler** used to scale all features between 0 and 1.

Step 4: Splitting the Data

- 80% for training
- 20% for testing
- Used **stratify=y** to keep target distribution balanced.

6. Machine Learning Models Used

Model 1: K-Nearest Neighbors (KNN)

- Simple distance-based algorithm.
- **Accuracy: 89.27%**

Model 2: Logistic Regression

- Fast and interpretable model.
- **Accuracy: 89.99%**

Model 3: Neural Network (MLPClassifier)

- 2 hidden layers (5 neurons, 2 neurons).
- Optimizer: Adam
- Max iterations: 2000
- **Accuracy: 89.99%**

7. Results and Insights

✔ Best Model: Logistic Regression + MLP (Tie)

Both models achieved the highest accuracy of **~90%**.

✔ Trends Observed

- High **satisfaction score** reduces resignation probability.
- Employees with:
  - High overtime
  - Low training hours
  - Low performance
    are more likely to resign.

✔ Why Model Works Well

- Data is clean
- Good number of numeric features
- Large dataset (100k rows) helps model learn patterns clearly

8. Saving the Model

You saved the final **Neural Network (MLP)** model using Pickle:

```
pickle.dump(clf, open('model.pkl', 'wb'))
```

This model can now be used in:

- Web apps
- HR dashboards
- Flask API or Streamlit interface

9. What I Learned

This project helped me strengthen my skills in:

- Handling large datasets
- Feature encoding and scaling
- Training multiple ML models
- Comparing model performance
- Saving a trained model for deployment

10. Future Improvements

If time allowed, I would improve the project by:

- Adding feature importance
- Using advanced models like Random Forest or XGBoost
- Creating a Streamlit prediction app
- Using SHAP values to interpret predictions
- Balancing classes if data becomes skewed

11. GitHub / Demo Links

Machine-learning-PROJECT-HUB/Resigned_prediction.ipynb at main · Student-NitinRaj/Machine-learning-PROJECT-HUB