HR Attrition Prediction Project Report

Problem Statement

The goal of this project is to predict whether an employee will leave (attrition) based on features like Monthly Income, OverTime, Work Experience, etc. This helps HR teams reduce employee turnover and plan retention strategies.

Step 1: Import Libraries

We used pandas for data handling, numpy for numerical operations, matplotlib and seaborn for visualization.

Step 2: Load Dataset

Used a CSV dataset containing HR records. Loaded using pandas and inspected with df.head(), df.info(), and df.describe().

Step 3: Data Cleaning

Handled null values and used one-hot encoding to convert categorical variables into numerical using pd.get_dummies().

Step 4: Exploratory Data Analysis

Visualized Attrition distribution and key relationships using sns.countplot and sns.histplot. Found that OverTime and low income are strongly linked with attrition.

Step 5: Train-Test Split

Split the data using train_test_split with 80% for training and 20% for testing.

Step 6: Handle Class Imbalance

Applied SMOTE (Synthetic Minority Over-sampling Technique) to balance the dataset since 'Leave' cases

were much fewer than 'Stay'.

Step 7: Model Training

Used RandomForestClassifier from sklearn. It's robust and handles tabular data well. Trained the model with balanced training data.

Step 8: Evaluation

Evaluated using classification report, confusion matrix, and ROC AUC score. AUC after SMOTE improved the balance between precision and recall.

Step 9: Model Saving

Saved the trained model using joblib.dump(), which allows easy loading in the Streamlit app.

Step 10: Streamlit App

Created an interactive UI using Streamlit to take user inputs and predict attrition in real-time using the trained model.

Final Summary

The project included data loading, cleaning, analysis, SMOTE, model training, and deployment. The final app helps HR predict attrition and act proactively.