A logo of a company

Description automatically generated A close-up of a logo

Description automatically generated

**Joint Tech Internship Community Program**

**Assignment: Predicting Employee Attrition in a Corporate Organization**

**Problem Statement:**

A large corporate organization is experiencing a high rate of employee attrition. Your task is to develop a machine learning model that predicts whether an employee will leave the organization or stay, based on various factors related to their job and personal background.

**Objective:**

Build a predictive model to classify whether an employee will leave the organization (attrition) or stay. Evaluate the model using appropriate metrics and provide insights into the factors contributing to employee attrition.

**Dataset:**

You are provided with a dataset containing the following columns:

1. **EmployeeID**: Unique identifier for each employee.
2. **Age**: Age of the employee.
3. **Gender**: Gender of the employee (Male/Female).
4. **MaritalStatus**: Marital status of the employee (Single, Married, Divorced).
5. **Education**: Level of education (e.g., High School, Bachelor, Master, PhD).
6. **Department**: Department where the employee works (e.g., HR, IT, Finance, Sales).
7. **JobRole**: Specific job role of the employee (e.g., Manager, Analyst, Developer).
8. **YearsAtCompany**: Number of years the employee has been with the company.
9. **YearsInCurrentRole**: Number of years the employee has been in their current role.
10. **MonthlyIncome**: Monthly income of the employee.
11. **JobSatisfaction**: Job satisfaction level on a scale of 1-5.
12. **WorkLifeBalance**: Work-life balance on a scale of 1-5.
13. **TrainingTimesLastYear**: Number of training sessions the employee attended last year.
14. **OverTime**: Whether the employee works overtime frequently (Yes/No).
15. **DistanceFromHome**: Distance from the employee's home to the workplace in miles.
16. **Attrition**: Whether the employee has left the company (Yes/No).

**Tasks:**

1. **Data Exploration and Preprocessing:**
   * Load the dataset and perform initial exploration to understand the data.
   * Identify and handle any missing values appropriately.
   * Detect and remove outliers using appropriate statistical methods (e.g., Z-score, IQR).
   * Convert categorical variables into numerical ones using techniques such as One-Hot Encoding.
2. **Feature Engineering:**
   * Perform feature scaling (e.g., Standardization or Normalization) on continuous variables.
   * Create new features if relevant, such as average tenure per department.
   * Use dimensionality reduction techniques (e.g., PCA) to reduce the feature space if necessary.
3. **Model Building:**
   * Split the dataset into training and testing sets (e.g., 80-20 split).
   * Train different classification models (e.g., Logistic Regression, Decision Trees, Random Forest, Gradient Boosting).
   * Perform hyperparameter tuning using techniques like Grid Search or Random Search to optimize model performance.
4. **Model Evaluation:**
   * Evaluate your models using appropriate metrics such as Accuracy, Precision, Recall, F1-Score, and AUC-ROC.
   * Compare the performance of different models and select the best one.
   * Analyze feature importance to understand the most significant factors contributing to employee attrition.
5. **Insights and Recommendations:**
   * Provide insights based on your model analysis regarding the factors that influence employee attrition.
   * Suggest actionable strategies for the company to reduce attrition based on the findings.
6. **Documentation:**
   * Document your process, including data exploration, preprocessing steps, model selection, and evaluation.
   * Include visualizations where necessary to support your findings.

**Deliverables:**

* A Jupyter notebook (or Python script) with the entire workflow.
* A report summarizing your findings, including the model's performance and recommendations for the business.

This assignment problem focuses on employee attrition in a corporate organization, covering various aspects of a typical machine learning workflow, including data preprocessing, outlier removal, feature engineering, dimensionality reduction, model training, and evaluation. It also emphasizes providing business insights based on the model's output.

Top of Form

Bottom of Form