

---

# Machine Learning

Course-End Project Problem Statement



# Employee Turnover Analytics

## **Project Statement:**

Portobello Tech is an app innovator who has devised an intelligent way of predicting employee turnover within the company. It periodically evaluates employees' work details, including the number of projects they worked on, average monthly working hours, time spent in the company, promotions in the last five years, and salary level.

Data from prior evaluations shows the employees' satisfaction in the workplace. The data could be used to identify patterns in work style and their interest in continuing to work for the company.

The HR Department owns the data and uses it to predict employee turnover. Employee turnover refers to the total number of workers who leave a company over time.

As the ML Developer assigned to the HR Department, you have been asked to create ML programs to:

1. Perform data quality checks by checking for missing values, if any.
2. Understand what factors contributed most to employee turnover at EDA.
3. Perform clustering of employees who left based on their satisfaction and evaluation.
4. Handle the left Class Imbalance using the SMOTE technique.
5. Perform k-fold cross-validation model training and evaluate performance.
6. Identify the best model and justify the evaluation metrics used.
7. Suggest various retention strategies for targeted employees.

## **Data will be modified from:**


<https://www.kaggle.com/liujiaqi/hr-comma-sepcsv>


| Column Name        | Description                                  |
|--------------------|--|
| satisfaction_level | Satisfaction level at the job of an employee |

|                       |   |
|-----------------------|---|
| last_evaluation       | Rating between 0 and 1, received by an employee at his last evaluation                      |
| number_project        | The number of projects an employee is involved in   |
| average_monthly_hours | Average number of hours in a month spent by an employee at the office                       |
| time_spend_company    | Number of years spent in the company  |
| Work_accident         | 0 - no accident during employee stay, 1 - accident during employee stay                     |
| left                  | 0 indicates an employee stays with the company and 1 indicates an employee left the company |
| promotion_last_5years | Number of promotions in his stay  |
| Department            | Department to which an employee belongs to  |
| salary                | Salary in USD   |

### **Perform the following steps:**

1. Perform data quality checks by checking for missing values, if any.
2. Understand what factors contributed most to employee turnover at EDA.
  - 2.1. Draw a heatmap of the correlation matrix between all numerical features or columns in the data.
  - 2.2. Draw the distribution plot of:
    - Employee Satisfaction (use column satisfaction\_level)
    - Employee Evaluation (use column last\_evaluation)
    - Employee Average Monthly Hours (use column average\_monthly\_hours)
  - 2.3. Draw the bar plot of the employee project count of both employees who left and stayed in the organization (use column number\_project and hue column left), and give your inferences from the plot.
3. Perform clustering of employees who left based on their satisfaction and evaluation.
  - 3.1. Choose columns satisfaction\_level, last\_evaluation, and left.

- 
- 3.2. Do K-means clustering of employees who left the company into 3 clusters?
    - 3.3. Based on the satisfaction and evaluation factors, give your thoughts on the employee clusters.
  4. Handle the left Class Imbalance using the SMOTE technique.
    - 4.1. Pre-process the data by converting categorical columns to numerical columns by:
      - Separating categorical variables and numeric variables
      - Applying `get_dummies()` to the categorical variables
      - Combining categorical variables and numeric variables
    - 4.2. Do the stratified split of the dataset to train and test in the ratio 80:20 with `random_state=123`.
    - 4.3. Upsample the train dataset using the SMOTE technique from the `imblearn` module.
  5. Perform 5-fold cross-validation model training and evaluate performance.
    - 5.1. Train a logistic regression model, apply a 5-fold CV, and plot the classification report.
    - 5.2. Train a Random Forest Classifier model, apply the 5-fold CV, and plot the classification report.
    - 5.3. Train a Gradient Boosting Classifier model, apply the 5-fold CV, and plot the classification report.
  6. Identify the best model and justify the evaluation metrics used.
    - 6.1. Find the ROC/AUC for each model and plot the ROC curve.
    - 6.2. Find the confusion matrix for each of the models.
    - 6.3. Explain which metric needs to be used from the confusion matrix: Recall or Precision?
  7. Suggest various retention strategies for targeted employees.
    - 7.1. Using the best model, predict the probability of employee turnover in the test data.
    - 7.2. Based on the probability score range below, categorize the employees into four zones and suggest your thoughts on the retention strategies for each zone.
      - Safe Zone (Green) (Score < 20%)
      - Low-Risk Zone (Yellow) (20% < Score < 60%)

- 
- Medium-Risk Zone (Orange) ( $60\% < \text{Score} < 90\%$ )
  - High-Risk Zone (Red) ( $\text{Score} > 90\%$ ).