**Case study – HR Churn Prediction**

**Introduction**

This case study focuses on predicting whether an employee will stay or leave the organization using a Logistic Regression model. Using a dataset containing various employee-related features such as salary, satisfaction level, average monthly hours, promotion last 5 years, and other relevant factors, we aim to build a predictive model that helps HR departments make data-driven decisions to improve employee retention strategies.

**Objective**

To perform Logistic Regression for predicting whether a employee will retain or not.

**Dataset link**

<https://drive.google.com/file/d/1DYKmXFqqFpXxpzx8t3yT3BRmMF1M38o-/view?usp=sharing>

**Understanding the data**

The dataset has 14,999 rows and 10 columns. Here's a breakdown of each column:

* satisfaction\_level - Employee’s level of satisfaction within the company.
* last\_evaluation - The most recent evaluation score of the employee's performance.
* number\_project - Number of projects the employee has worked on.
* average\_montly\_hours - Average number of hours worked per month.
* time\_spend\_company - Number of years the employee has been with the company.
* Work\_accident - Whether the employee has had a work accident (0 = No, 1 = Yes).
* left - Whether the employee left the company (0 = No, 1 = Yes).
* promotion\_last\_5years - Whether the employee was promoted in the last 5 years (0 = No, 1 = Yes).
* department - The department in which the employee works (e.g., sales, technical, HR).
* salary - Employee's salary level (categorical: low, medium, high).

**Procedure for coding**

* Import necessary libraries
* Load dataset
* Check the number of rows and columns
* Check for missing values
* Defining Independent Variables
* Creating dummy variable for Salary
* Creating independent (X) and dependent (y) dataframes
* Standardize independent variables
* Split data into training and test sets
* Train Logistic Regression model
* Make predictions
* Evaluate model performance (confusion matrix & accuracy)
* Display actual vs predicted program choices

**Code File Link**

<https://github.com/Ishita2003M/Logistic-Regression-Model-for-HR-Churn-Prediction/blob/main/Logistic_Reg_Emp_Retention.ipynb>

**Interpretation and conclusion**

1. The analysis aims to build a predictive model using Logistic Regression to determine whether an employee is likely to leave the company based on key factors such as salary, satisfaction level, average monthly hours and promotion last 5 years.
2. Model Performance:
   * The confusion matrix indicates:

[[3190 238]

[ 814 258]]

* + - True Negatives (3190): Employees correctly predicted to stay.
    - False Positives (238): Employees wrongly predicted to leave, but they actually stayed.
    - False Negatives (814): Employees wrongly predicted to stay, but they actually left.
    - True Positives (258): Employees correctly predicted to leave. The model achieved a classification accuracy of 82.67%, indicating a high level of prediction reliability in distinguishing between purchasers and non-purchasers.
  + Target Class Distribution:
    - Actual employees who left: 1072
    - Actual retained employees: 3428
    - Predicted to leave: 496
    - Predicted to stay: 4004

1. Insights:
   * The model tends to under-predict employee exits (only 496 predicted vs. 1072 actual).
   * This shows the model is a bit **conservative in predicting churn**, likely due to the imbalance in the dataset.