**Title: Analysis of Employee Attrition & Performance using Machine learning**

**Problem Statement:** Employee attrition is a significant concern for organizations, leading to productivity loss, increased recruitment costs, and a negative impact on overall performance. It can be costly to replace employees, and it can also disrupt business operations. In the case of IBM, employee attrition has been a significant problem in recent years. In 2021, IBM's attrition rate was 14%, which is higher than the industry average. Using synthetic data created by IBM Watson, machine learning experiments will be conducted to predict employee attrition and will provide a clear perspective to top management in making key decisions

**Dataset:** The dataset used for this project is the IBM HR Analytics Employee Attrition & Performance dataset, which is available on Kaggle. The dataset consists of a substantial 1,470 rows and 35 columns, making it suitable for comprehensive big data analytics including their age, gender, job title, salary, performance ratings, and whether they have left the company.

Link: <https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset>

**Aims and Objectives:** The aim of this project is to identify the factors that are most likely to lead to employee attrition using big data analytics. The objectives of the project are to:

1. To Perform exploratory data analysis (EDA) to understand the data distribution and data visualization using Tableau to identify the key features
2. To Develop predictive models that can be used to predict employee attrition.
3. To Evaluate the performance of the models and identify the best model for predicting employee attrition.

**Work-Plan:**

1. Week 1: Data acquisition and understanding the dataset. Perform initial data cleaning and pre-processing steps, such as handling missing values and data type conversions.
2. Week 2: Conduct Exploratory Data Analysis (EDA) using PySpark to gain insights into the dataset. Visualize data distributions, and correlations, identify potential patterns, and Integrate Tableau for interactive data visualization to create insightful dashboards for presenting the findings.
3. Week 3: Develop predictive models for employee attrition like Random Forest and Logistic Regression. Evaluate the models' performance and fine-tune them to achieve higher accuracy

**Tools & Frameworks:** Pyspark, Python (3.7), Jupyter IDE, Pandas, Tableau and sci-kit-learn