**ML Assignment 6: Supervised Model Comparison   
and k-Means Clustering**

**Supervised Learning Model Comparison**

**Life Expectancy (WHO) Analysis**

The WHO Life Expectancy dataset provides global statistics on life expectancy at birth and various ages, categorized by gender and health indicators. This analysis will involve comparing both regression and classification models. For regression, we will predict exact life expectancy values based on various features, while for classification, we will categorize countries as "developed" or "developing" based on the status column. We will evaluate model performance using metrics such as R² for regression and F1 score for classification. This dual approach will enhance our understanding of the factors influencing life expectancy and inform targeted public health strategies.

**Dataset Link:**

<https://raw.githubusercontent.com/ArchanaInsights/Datasets/refs/heads/main/Life%20Expectancy%20Data.csv>

**Model Comparison Steps:-**

1. **Initial Exploration and Pre-Processing**
   1. Load the dataset and perform an initial exploration to understand its structure and content.
   2. Determine the number of unique values in the *'Country'* and *'Year'* columns.
   3. Analyze the distribution of values in the *'Status'* column. If the dataset is imbalanced, consider using appropriate metrics for evaluation.
   4. Check for any missing values. Impute missing values for all numerical columns using the median or another suitable strategy.
   5. Optionally, you may perform a comprehensive EDA to visualize relationships, distributions, and patterns in the data.
   6. Encode categorical columns appropriately.
2. **Classification Models Comparison**
   1. Define *'Status'* as the target variable and use all other columns as features.
   2. Split the data into training and testing sets.
   3. Scale all feature columns, excluding *'Country'* and *'Year'*.
   4. Compare classification models including ***Logistic Regression***, ***K-Neighbors Classifier***, ***Decision Tree Classifier***, ***Random Forest Classifier***, ***Support Vector Classifier***, ***Gaussian Naive Bayes***,and ***XGBoost***. Report both ***accuracy*** and ***F1-scores*** for each model. Optionally, you may perform hyperparameter tuning to optimize model performance.
   5. Identify and report the best-performing classification model for this dataset.
3. **Regression Models Comparison**
   1. Define *'Life expectancy'* as the target variable and use all other columns as features.
   2. Split the data into training and testing sets.
   3. Scale all feature columns, excluding *'Country'*, *'Year'*, and *'Status'*.
   4. Compare regression models including ***Linear Regression***, ***Decision Tree Regressor***, ***Random Forest Regressor***, ***Support Vector Regressor***,and ***K-Neighbors Regressor.*** Report ***R² score*** and ***Mean Absolute Error*** for each model. Optionally, you may perform hyperparameter tuning to enhance performance.
   5. Identify and report the best-performing regression model for this dataset.

**Unsupervised Learning – k-Means**

**Customer Segmentation**

The mall-customers dataset provides demographic and spending information about customers at a shopping mall, including features such as age, income, and spending score. This analysis will utilize k-Means clustering to segment customers into distinct groups based on their behaviors and characteristics. By identifying these segments, we can uncover patterns and insights that can inform targeted marketing strategies and improve customer experience. The results will help in understanding customer preferences and optimizing services to enhance engagement and satisfaction.

**Dataset Link:**

<https://raw.githubusercontent.com/ArchanaInsights/Datasets/refs/heads/main/Mall_Customers.csv>

**k-Means Clustering Steps:**

* **Initial Visualization:** Create a scatter plot using *'Annual Income (k$)'* on the x-axis and *'Spending Score (1-100)'* on the y-axis to visualize the data distribution.
* **Feature Selection:** Select *'Annual Income (k$)'* and *'Spending Score (1-100)'* as the features for k-Means clustering.
* **Determine Optimal k-Value:** Calculate the ***Within-Cluster Sum of Squares (WCSS)*** for a range of ***k-values***.
* **Plot the Elbow Chart:** Plot the elbow method graph to visually determine the optimal k-value.
* **Build k-Means Model:**
  + Construct the k-Means model using the ***optimal k-value*** determined in the previous step.
  + Identify the ***clusters*** and ***centroids*** from the model.
* **Visualization of Clusters:** Plot the scatter plot again, this time including the centroids. Color the data points according to their respective clusters to visualize the clustering results effectively.