**Case study – Predicting Car Purchase Decision**

**Introduction**

This case study focuses on predicting car purchase decisions using Linear Discriminant Analysis (LDA). The analysis aims to classify customers based on their likelihood to purchase a car. By analyzing customer data, including age and annual salary, the LDA model seeks to provide insights into the factors that influence purchasing behavior.

**Objective**

To perform Linear Discriminant Analysis (LDA) for predicting whether a customer will purchase a car or not.

**Dataset link**

<https://drive.google.com/file/d/1yYjjJiRN7RKQKyrGLwTl8LVly4Dsfz5Y/view?usp=sharing>

**Understanding the data**

The dataset has 1,000 rows and 5 columns. Here's a breakdown of each column:

* User ID -A unique identifier for each user.
* Gender - Categorical variable indicating the user's gender (e.g., "Male", "Female").
* Age - Age of the user in years.
* AnnualSalary - The user's annual salary.
* Purchased - A binary variable indicating whether the user purchased the car: 1 = Purchased, 0 = Not purchased

**Procedure for coding**

* Import necessary libraries
* Load dataset
* Check the number of rows and columns
* Check for missing values
* Define independent (X) and dependent (y) variables
* Standardize independent variables
* Split data into training and test sets
* Train LDA model
* Make predictions
* Evaluate model performance (confusion matrix & accuracy)
* Display actual vs predicted program choices

**Code File Link**

<https://github.com/Ishita2003M/Predicting-Car-Purchase-Decision/blob/main/car_LDA.ipynb>

**Interpretation and conclusion**

1. The analysis aims to predict whether a customer is likely to purchase a car using Linear Discriminant Analysis (LDA), a supervised classification technique that models class separation based on linear combinations of features.
2. The dataset consists of 1,000 customer records with the following attributes:
   * User ID (unique identifier – not used in modeling)
   * Gender (categorical – not used in LDA)
   * Age (used as predictor)
   * AnnualSalary (used as predictor)
   * Purchased (target variable: 1 = Purchased, 0 = Not purchased)
3. The LDA model uses Age and Annual Salary as the independent variables (X) to predict the binary target variable Purchased (Y). This setup captures the relationship between a user’s demographic and financial profile and their purchasing behavior.
4. Model Performance:
   * The confusion matrix indicates:

[[158 14]

[ 38 90]]

* + - True Negatives (TN): 158 — Correctly predicted not purchased
    - False Positives (FP): 14 — Predicted purchased, but not actually purchased
    - False Negatives (FN): 38 — Predicted not purchased, but actually purchased
    - True Positives (TP): 90 — Correctly predicted purchased
  + The model achieved a classification accuracy of 82.67%, indicating a high level of prediction reliability in distinguishing between purchasers and non-purchasers.

1. Target Class Distribution:
   * Actual Purchased (Y=1): 128 users
   * Actual Not Purchased (Y=0): 172 users
   * Predicted Purchased: 104 users
   * Predicted Not Purchased: 196 users
2. Insights:
   * The model performs well overall, with strong predictive power and a relatively balanced classification performance.
   * While the number of false negatives (38) suggests some underestimation of buyers, the relatively low false positive rate (14) is favorable in contexts where over-targeting uninterested customers can be costly.
   * The use of only two predictors (Age and Salary) achieved a notable level of accuracy, suggesting these variables have strong discriminative power in purchase behavior.