Advance Project - 9: Classification using Random Forest on Company\_Data.csv

# 1. Objective

Classify sales performance (HighSales) based on various store and demographic attributes using a Random Forest classification model to help the cloth manufacturing company identify key factors that lead to high unit sales of car seats.

# 2. Problem Statement

A cloth manufacturing company wants to understand which attributes contribute the most to high sales. We classify the Sales variable into High or Low sales and train a model to predict the segment.

# 3. Dataset Description

|  |  |
| --- | --- |
| Feature | Description |
| Sales | Unit sales (in thousands) |
| CompPrice | Price charged by competitor |
| Income | Community income level |
| Advertising | Local advertising budget |
| Population | Population in the region |
| Price | Price company charges |
| ShelveLoc | Shelf location quality (Bad/Medium/Good) |
| Age | Average age of the local population |
| Education | Education level |
| Urban | Whether store is in urban area (Yes/No) |
| US | Whether store is in the US (Yes/No) |
| HighSales | Derived target variable: 'High' if Sales ≥ 8, else 'Low' |

# 4. Methodology (Solution Architecture)

1. Data Understanding & Cleaning  
2. Exploratory Data Analysis (EDA)  
3. Data Preprocessing  
4. Model Building (Random Forest Classifier)  
5. Model Evaluation (Accuracy, Confusion Matrix)  
6. Feature Importance Analysis

# 5. Tools & Technologies Used

Python (Pandas, NumPy, Scikit-learn, Seaborn, Matplotlib), Jupyter Notebook, GitHub, MS Word / PowerPoint

# 6. Business Impact

Helps identify top contributing factors to high sales and optimize decisions related to pricing, advertisement, and product placement.

# 7. Time Taken

Total Time: ~6 Hours  
- Data Understanding & EDA: 2.5 Hours  
- Preprocessing & Modeling: 2 Hours  
- Documentation & Presentation: 1.5 Hours

# 8. Challenges Faced

Handling categorical variables, defining a sales threshold, and ensuring class balance.

# 9. Project Complexity

Level: Intermediate  
Type: Classification  
Model: Random Forest  
Key Challenge: Interpreting multi-feature influence

# 10. Results & Insights

Top features influencing High Sales:  
- Shelf Location (Good shelves boost sales)  
- Advertising Budget  
- Price Sensitivity  
Urban and US stores tend to perform better.

# 11. Future Improvements

Hyperparameter tuning, exploring boosting models, building dashboards, and addressing class imbalance.