DSTI School of Engineering

Applied MSc in Data Analytics
Applied MSc in Data Science & Artificial Intelligence
Applied MSc in Data Engineering & Artificial Intelligence

Project: Data-Driven Retail Insights for Cost Savings & Revenue Growth

Instructor: Assan Sanogo

1. Dataset Overview

We will use the **Instacart Online Grocery Basket Analysis Dataset** (Kaggle link), which contains:

- **3+ million orders** from **200,000+ customers**.
- **Customer behavior data**: repeat purchases, shopping frequency, and product preferences.
- **Product details**: categories, aisles, and departments.

This dataset has already been used in **academic and applied research** to study customer behavior, shopping patterns, and retail optimization — making it a solid foundation for educational exploration and business-focused insights.

2. Project Objective

The goal is to **help shop owners** answer a critical question: 👉 *“How much money can I save or earn if I listen to the insights from this DSTI team ;) ?”*

Our team will provide actionable recommendations that enable shop owners to:

- **Boost sales** through smart product bundling and upselling.
- **Predict customer purchases**

- **Segment customers** to tailor marketing strategies.
- **Design bundles** that maximize revenue.

3. Data science exploration

- The usage of Github is highly recommended (because of teamwork)
- The usage of mini project management is recommended too

Data scientists will apply **association rule mining algorithms** to uncover product relationships:

- **Apriori** → classic algorithm for frequent itemset mining.
- **Eclat** → efficient depth-first search approach.
- **FP-Growth** → scalable algorithm for large datasets.
- **UP-Tree (Utility Pattern Tree)** → goes beyond frequency by incorporating **utility (value/profit)** of items, identifying not just what products are bought together but which combinations **generate the most revenue or savings**.

Challenge: Apriori, Eclat, and FP-Growth focus only on frequency, ignoring product value and sequence. UP-Tree introduces the **notion of value**, making it more relevant for monetization strategies. Students will be encouraged to compare these approaches and highlight the added business impact of utility-aware mining.

4. Data Enrichment

Data engineers could/may enrich the dataset by crawling **real-world product prices** from Open Food Facts Prices (for example).

- Inputing decisions should/may be made when prices are not available
- This would allow us to calculate the **actual monetary value** of associations.
- The result could be a **clean, organized dataset** combining transactional data with pricing and product attributes.

5. Business Application

Business analysts will design a **dashboard/tool** that shows :

- **Customer segmentation** (e.g., budget vs. premium shoppers).
    - Type of buyer (frequent/irregular)
    - Basket size
    - Variability in basket size
- **Product associations** (frequent bundles and co-purchases).
    - Type of bundles…
- **Revenue simulations** (how much money could be lost/earned with current practices).
- **Promotion efficiency** (ROI of targeted discounts vs basic untargeted discount).

This dashboard will directly answer the shop owner's key question:

NB: Data people (us) are not listened to at the beginning of a data project, so you must work on the project trying to answer : *"If I adopt these insights, how much money do I save, and why?"*

6. Educational Value

Students will collaborate across roles:

- **Data Scientists** → uncover associations & predictive models (including utility-aware mining with UP-Tree).
- **Data Engineers** → explore external pricing data integration.
- **Business Analysts** → translate insights into **financial impact**.

By the end, the team will deliver a **prototype tool (simple for the user)** that demonstrates how **data science can help shop owners save money and grow their business**.

7. Project Evaluation

The project will be evaluated using the following rubric. It contains the required items for a complete submission as well as bonus elements. The grading system is over 5 and the final grade will be transformed to a grade over 100.

- Jupyter Notebook (or Python script) containing entire machine learning pipeline **[1 point]**
- Complete web application code **[1 point]**
- Report (in PDF format) **[1 point]**
- Web application short demo video **[1 point]**
- GitHub repository **[1 point]**
- **BONUS:** Best group model performance in class **[1/2 point]**