Data Science exercise

🧪 Exercise Title: “Smart Risk & Fit Prediction Engines” (2 - 4h)

You are tasked with designing the foundations of a dual-purpose AI system for a product agency:

1. **Fraud Detection Model** for payment logs
2. **Profile-based Ranking Algorithm** for recommending health insurance plans

Your objective is to choose the best approach for each problem, justify your choices, and outline the pipeline to go from raw data to working models (both offline and online). You are encouraged to use AI tools to accelerate your work.

✅ Deliverables

The candidate should submit a **notebook or PDF** containing:

1. **Problem Formulation**
    - Rephrase and frame each problem.
    - Identify metrics of success and potential challenges.
2. **Data Preparation Plan**
    - Simulate what raw data may look like.
    - Outline how they would load, clean, transform, and enrich the data.
    - Discuss how to handle missing values, outliers, feature engineering.
3. **Modeling Strategy**
    - For **Fraud Detection**: suitable supervised or unsupervised models, class imbalance handling, explainability.
    - For **Ranking**: choice of ranking models (e.g., Learning to Rank, logistic regression + sorting logic, embeddings), how user and insurance profile data might be structured.

4. **Evaluation & Testing**
    - Define offline evaluation strategies.
    - Describe how online testing would be designed.
5. **GenAI & LLM Integration**
    - Suggest ways to use GenAI/LLMs to enhance the models or data pipeline.
6. **Bonus (if time permits)**
    - Prototype snippets: show basic data schema, simulate inputs/outputs, or sketch a model training flow using pseudocode or notebooks.