**Prompt — Projet complet STA211 à coller dans Cursor ou ChatGPT :**

Tu es un assistant expert en data science spécialisé dans les méthodes statistiques enseignées dans le cours **STA211 du CNAM**.

Tu dois m’aider à construire **un Jupyter Notebook en Python** pour prédire si une image est une publicité ou non, à partir du dataset *Internet Advertisements*.

Voici les attentes pédagogiques à respecter dans le notebook :

**🔹 Exploration et nettoyage :**

* Charger les données ad.data (avec noms de colonnes si disponibles)
* Gérer les **valeurs manquantes** : identifier le mécanisme (MCAR/MAR/MNAR), appliquer des imputations (moyenne, kNN, multivariée)
* Vérifier les **distributions** : histogrammes, boxplots, qq-plots pour tester la normalité
* Appliquer des **transformations** : log, Box-Cox, etc. pour linéarité ou normalité
* Discrétiser les variables si utile (quantiles, CAH, MDLPC)

**🔹 Analyse exploratoire :**

* Réaliser des analyses **univariées**, **bivariées** (avec la cible) et **multivariées**
* Utiliser **ACP/AFM** ou **classification de variables** pour explorer la structure
* Si applicable, tester les **cartes de Kohonen**

**🔹 Modélisation supervisée :**

* Mettre en œuvre les modèles du cours :
  + **Régression logistique**
  + **CART (arbre de décision)**
  + **Forêt aléatoire**
  + **KNN ou SVM** (si souhaité)
* Appliquer la **validation croisée** ou split (80/20) stratifié
* Évaluer les modèles : **précision, rappel, F1-score, AUC, confusion matrix**
* Gérer les classes déséquilibrées : **pondération**, **SMOTE**

**🔹 Interprétation & conclusion :**

* Visualiser les **importances de variables**
* Expliquer le **compromis biais-variance**
* Comparer les modèles selon **performance et complexité**
* Présenter les résultats avec des **cellules Markdown claires**

Génère un **notebook Python structuré, commenté**, en utilisant : pandas, scikit-learn, seaborn, matplotlib, prince, imblearn.  
Si besoin, propose une **analyse factorielle multiple**, ou des regroupements de modalités pour faciliter la lecture des données.