**Prompt — Projet complet STA211 :**

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 data\_train.csv
* Gérer les **valeurs manquantes** : identifier le mécanisme (MCAR/MAR/MNAR), appliquer des imputations en utilisant des méthodes avancées (moyenne, médiane, 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.

Préparer le fichier de prédiction prêt à être soumis en respectant les consignes du challenge :

Soumissions

Les soumissions se font sous forme de fichier texte portant l’extension .csv, que vous pouvez exporter avec la commande suivante :

**write**(y\_pred, file = "my\_pred.csv")

Le fichier doit contenir 820 lignes contenant uniquement un des mots ad. ou noad..

Tous les fichiers .csv placés dans votre répertoire Dropbox partagé seront automatiquement importés grâce à la fonction read\_pred.

read\_pred <- **function**(file, n = 820) {

y\_pred <- **scan**(file, what = "character")

**if** (**length**(y\_pred) **!=** n)

**stop**("Le nombre de prédictions est incorrect")

**if** (**any**(**is.na**(y\_pred)))

**stop**("Le vecteur de prédiction contient des valeurs manquantes (NA°")

**if**(**any**(**!**(**unique**(y\_pred)**%in%c**("ad.","noad.")))){

**stop**(**paste0**("Le nom des modalités prédites doit être ad. ou noad. au lieu de ",**paste**(**unique**(y\_pred),collapse = " ")))}

**if**(**length**(**unique**(y\_pred))**==**1){

y\_pred <- **as.factor**(y\_pred)

}**else**{

y\_pred <- **factor**(y\_pred, levels = **c**("ad.", "noad."))}

**return**(y\_pred)

}

Utilisez cette fonction pour vérifier que votre fichier sera correctement importé.

Les erreurs de lecture lors de l’import sont affichées à la section [Erreurs de lecture](file:///C:\Users\aella\OneDrive\Documents\Projects\STA211_Challenge_2025\description_du_challenge.html#erreurs-de-lecture).

Une fois un fichier importé, son score est calculé et stocké. Vous pouvez effacer ou remplacer des soumissions, l’historique est conservé.