# Projet : Analyse prédictive du VIH à partir de KmerData.csv

## Compréhension des affaires

**Q : Quel problème essayez-vous de résoudre ou à quelle question essayez-vous de répondre ?**
R : Prédire le statut VIH (positif ou négatif) chez les individus à partir de séquences génomiques codées en k-mers, afin de détecter les risques de manière rapide et précise.

**Q : Pourquoi ce sujet ?**
R : Le VIH reste un problème majeur de santé publique. L'analyse des k-mers permet d'extraire des informations génétiques pertinentes pour la classification. Ce projet est motivant car il combine bioinformatique, apprentissage automatique et impact réel sur la santé.

**Q : À quel secteur/domaine cela s'applique-t-il ?**
R : Santé publique, bioinformatique, génomique, data science appliquée à la médecine.

**Q : Quel est votre public cible ?**
R : Chercheurs en VIH/SIDA, institutions de santé, laboratoires cliniques, ONG et programmes de prévention du VIH.

**Q : Quel impact aurait votre réponse/solution sur le monde réel ?**
R : Détection précoce du VIH, meilleure allocation des ressources médicales et aide à la recherche sur les mutations génétiques liées au VIH.

**Q : Quels projets/recherches/articles préexistants dans ce domaine avez-vous explorés ?**
R : Études utilisant les k-mers pour la classification virale et l'apprentissage supervisé pour prédire le statut VIH. Ce projet applique ces méthodes avec une pipeline complète de machine learning pour améliorer la prédiction.

## Compréhension des données

**Q : Quelles données allez-vous collecter ?**
R : Le fichier `KmerData.csv`, contenant des séquences génétiques transformées en k-mers et la variable cible : statut VIH.

**Q : D'où viennent vos données brutes ?**
R : Fournies sous forme de CSV (`KmerData.csv`).

**Q : Existe-t-il un plan pour obtenir les données ?**
R : Les données sont déjà disponibles localement. Une vérification d'intégrité et un prétraitement seront réalisés.

**Q : Les fonctionnalités qui seront utilisées sont-elles décrites clairement ?**
R : Oui, chaque k-mer correspond à une colonne. La variable cible est le statut VIH (positif/négatif).

**Q : Quelqu'un d'autre a-t-il travaillé sur ce problème ?**
R : Oui, des recherches ont utilisé des k-mers pour la classification virale. Ce projet s'appuie sur ces travaux mais introduit un pipeline supervisé complet avec comparaison de modèles.

# Préparation des données

**Q : Sous quelle forme les données sont-elles stockées ?**
R : CSV, avec lignes = individus et colonnes = k-mers + variable cible.

**Q : Quels sont les types de données des variables ?**
R : Numériques pour les k-mers, catégorielle binaire pour la variable cible.

**Q : Quelles étapes de prétraitement prévoyez-vous ?**
R : Nettoyage des valeurs manquantes, standardisation ou normalisation, encodage de la variable cible, réduction de dimension éventuelle (PCA ou sélection de caractéristiques).

**Q : Quels sont les défis liés au nettoyage et au prétraitement ?**
R : Haute dimensionnalité (beaucoup de k-mers), risque de surapprentissage, gestion des valeurs manquantes.

**Q : Quel est le nombre minimum de lignes ?**
R : Estimation : plusieurs milliers de lignes.

**Q : Comment comptez-vous visualiser les aspects importants de ces données ?**
R : Histogrammes des k-mers, carte de chaleur des corrélations, PCA ou t-SNE pour visualiser la séparation des classes.

# Modélisation

**Q : Quelles techniques de modélisation sont les plus appropriées ?**
R : Régression logistique, Random Forest, XGBoost ou LightGBM.

**Q : Quelle est votre variable cible ?**
R : Statut VIH (binaire : positif / négatif).

**Q : Quel modèle prévoyez-vous d'utiliser comme base de référence ?**
R : Régression logistique simple avec quelques k-mers sélectionnés.

**Q : S'agit-il d'un problème de régression ou de classification ?**
R : Classification supervisée binaire.

# Évaluation

**Q : Quelles mesures utiliserez-vous pour déterminer le succès ?**
R : Accuracy, Precision, Recall, F1-score, ROC-AUC.

**Q : En quoi consiste le produit minimum viable (PMV) ?**
R : Modèle de classification simple utilisant un sous-ensemble réduit de k-mers pour prédire le statut VIH avec validation croisée.

**Q : Quels sont vos objectifs de progression ?**
R : Ajouter tous les k-mers pertinents, optimiser les hyperparamètres, comparer plusieurs modèles, déployer un notebook interactif pour visualiser les prédictions.

---

# Déploiement

**Q : La méthode de communication des résultats finaux est-elle décrite ?**
R : Oui, notebook interactif sur GitHub avec visualisations et rapport PDF.

**Q : Existe-t-il un plan de déploiement ?**
R : Optionnel : interface web avec Streamlit ou Dash pour tester le modèle avec de nouvelles données.

**Q : Quelle est la fonctionnalité ?**
R : Prédiction du statut VIH pour de nouvelles séquences k-mer et visualisation des k-mers les plus importants pour la prédiction.

# Outils / Méthodologies

**Q : Quelles bibliothèques Python prévoyez-vous d'utiliser ?**
R : `pandas`, `numpy`, `scikit-learn`, `xgboost`, `lightgbm`, `matplotlib`, `seaborn`, `plotly`, `imblearn` (si classes déséquilibrées).

**Q : Quels algorithmes de modélisation prévoyez-vous d'utiliser ?**
R : Régression logistique, Random Forest, XGBoost, LightGBM.

**Q : Où allez-vous effectuer votre analyse ?**
R : Local avec Anaconda ou Google Colab si nécessaire.

**Q : Vos données seront-elles stockées sur votre machine ou dans le cloud ?**
R : Localement pour le fichier `KmerData.csv`. GitHub sera utilisé pour le code et les instructions de téléchargement si d'autres données sont nécessaires.