# Projet de Machine Learning : Analyse Prédictive et Visualisation des Données sur la Satisfaction Client

Master I en Data Science & IA - Départepment IA & ingénierie des données

Ibrahima SY

Institut Superieur Informatique

## Objectifs du Projet

1. **Comprendre et utiliser les techniques d'analyse prédictive** avec des algorithmes supervisés (KNN, régression linéaire, Naïve Bayes).
2. **Effectuer une réduction de dimensionnalité** avec l'ACP pour mieux visualiser et interpréter les données.
3. **Explorer et nettoyer des données** pour créer un pipeline complet de traitement des données.
4. **Améliorer la précision des modèles** en optimisant les paramètres et en comparant les performances des modèles.
5. **Utiliser Kaggle pour le partage des données** et éventuellement pour l'obtention d'un dataset riche, adapté aux objectifs.

## Dataset

Utiliser un **dataset de satisfaction client** ou un dataset sur **l'évaluation des services (ex : satisfaction des clients d'une compagnie aérienne)** disponible sur Kaggle, tel que Customer Satisfaction for Airlines.

---

## Structure du Projet

### 1. Exploration des Données (EDA)

- Importer et visualiser les premières lignes du dataset pour comprendre les variables.
- Analyser les types de données, détecter les valeurs manquantes, les valeurs aberrantes, et identifier les corrélations entre les variables.
- Visualiser la distribution des données pour les différentes classes (par exemple, satisfaction/non-satisfaction).
- **Livrable :** Rapport d'analyse préliminaire incluant des graphiques et des statistiques descriptives.

### 2. Prétraitement des Données

- **Nettoyage :** Imputer les valeurs manquantes et traiter les valeurs aberrantes.
- **Encodage des variables catégorielles :** Appliquer l'encodage par variables factices (dummy variables) pour les données non numériques.
- **Standardisation :** Standardiser les données numériques pour améliorer les performances des algorithmes.
- **Livrable :** Code et documentation du pipeline de prétraitement des données.

**3. Réduction de Dimensionnalité avec l'ACP**

- Appliquer l'ACP pour réduire les dimensions du dataset à 2 ou 3 dimensions.
- Visualiser les données projetées pour explorer les groupes et les clusters éventuels.
- Interpréter les résultats de l'ACP pour comprendre quelles variables influencent le plus la satisfaction client.
- **Livrable :** Visualisation des données projetées en 2D ou 3D avec interprétation des résultats.

**4. Modélisation et Prédiction**

- Diviser le dataset en ensemble d'entraînement et de test.
- **Régression Linéaire :** Utiliser la régression pour prédire une note de satisfaction (si applicable).
- **KNN (K-Nearest Neighbors) :** Appliquer KNN pour classer la satisfaction (satisfait/non-satisfait).
- **Naïve Bayes :** Appliquer Naïve Bayes pour comparer les performances et évaluer la probabilité d'appartenance aux classes.
- **Livrable :** Rapport de comparaison des performances de chaque modèle (précision, rappel, F1-score).

**5. Évaluation et Optimisation des Modèles**

- Utiliser la validation croisée pour évaluer les modèles et éviter le surapprentissage.
- Effectuer une recherche par grille (Grid Search) pour optimiser les hyperparamètres (ex. nombre de voisins pour KNN).
- Comparer les performances avant et après optimisation.
- **Livrable :** Code d'optimisation et tableau récapitulatif des performances optimisées des modèles.

**6. Déploiement et Visualisation Interactive**

- Utiliser Streamlit pour créer une interface où l'utilisateur peut tester les modèles avec de nouvelles données.
- Visualiser la probabilité de satisfaction prédite et explorer les différents modèles.
- **Livrable :** Interface interactive en Streamlit ou en notebook Jupyter, incluant les prédictions en temps réel.

**7. Documentation et Présentation Finale**

- Expliquer les choix d'algorithmes et les étapes de prétraitement.
- Discuter des résultats obtenus, des limites et des potentielles améliorations.
- Publier les résultats et le code sur un notebook Kaggle pour partager le projet avec la communauté.
- **Livrable :** Présentation PowerPoint ou PDF avec tous les résultats, interprétations, et code final.

---

## Outils et Technologies

- **Python (pandas, numpy, scikit-learn, matplotlib, seaborn)** pour la manipulation des données, la modélisation, et la visualisation.
- **Kaggle** pour télécharger et partager les datasets.
- **Streamlit** pour l'interface interactive.
- **Jupyter Notebook** pour l'organisation des étapes et des résultats.