Rapport du Projet
Classification du Niveau de Risque d'Obésité

MOUACHA BASSOU

20 juin 2025

Introduction

Ce projet a pour objectif d'exploiter un ensemble de données démographiques et comportementales pour prédire le niveau d'obésité d'un individu à l'aide d'algorithmes d'apprentissage automatique. Le jeu de données comprend **20 758 échantillons**, avec des variables telles que l'âge, la taille, le poids, l'activité physique, les habitudes alimentaires, etc.

Les classes cibles sont :

- `Insufficient_Weight`
- `Normal_Weight`
- `Overweight_Level_I`
- `Obesity_Type_I`
- `Obesity_Type_II`
- `Obesity_Type_III`

Analyse Exploratoire des Données (EDA)

Résumé de l'EDA

- **Genre :** majorité de femmes ; elles consomment plus d'eau, font plus de sport, utilisent plus les transports en commun.
- **Âge :** moyenne de 24,4 ans ; valeurs aberrantes supprimées.
- **Poids :** distribution bimodale ; pics chez les femmes (55–60kg) et les hommes (80–85kg).
- **Activité physique :** modérée chez les hommes ; très peu de femmes inactives.

Prétraitement des Données

- Suppression de la colonne `id`
- Arrondi des âges
- Détection/suppression des *outliers* (IQR)
- Encodage via `LabelEncoder`
- Séparation 80% entraînement / 20% test

Modélisation et Résultats

2gray !10white

Modèle	Précision (%)	Commentaires
Random Forest	**90**	Excellente performance sur toutes les classes
SVC (SVM)	87	Bon compromis, nécessite normalisation
Decision Tree	84	Interprétable, mais moins stable
KNN	77	Sensible à l'échelle, résultats décevants
Naive Bayes	68	Hypothèses simplistes, moins adapté

TABLE 1 – Comparaison des performances des modèles

Commentaires détaillés

- **Random Forest** : robuste, précis, bien équilibré. Meilleur choix global.
- **SVC** : nécessite `StandardScaler`, bonnes performances sur classes minoritaires.
- **KNN** et **Naive Bayes** : résultats médiocres, inadéquats sans traitement avancé.

Conclusion

Le modèle **RandomForestClassifier** est le plus performant avec une précision globale de **90%**. Il surpasse les autres méthodes par sa stabilité, sa capacité à gérer des données complexes et sa robustesse sur les cas extrêmes.

Axes d'amélioration future :

- Optimisation par `GridSearchCV`
- Techniques d'équilibrage (SMOTE, class weights)
- Exploration de modèles profonds (Deep Learning)

Annexes

- **Dataset :** `ObesityDataSet_raw_and_data_sinthetic.csv`
- **Librairies :** `pandas, scikit-learn, seaborn, matplotlib`
- **Modèle sauvegardé :** `obesity_risk_model.pkl`