cahier des charges :

2-LBC-BI 4

*

Encadré par: Dr. Eya JEBALI

- Réalisé par :Souha Salhi

1. Introduction

Contexte:

- Le projet s'inscrit dans le cadre de la mise en place d'un système décisionnel qui intègre des analyses de données et des prévisions basées sur des modèles de machine learning.
- L'objectif est de fournir des outils d'aide à la décision pour améliorer la compréhension des facteurs influençant la santé des individus.

Objectif :

- Créer un entrepôt de données qui stocke des informations sur la santé et le mode de vie.
- Élaborer un tableau de bord interactif pour visualiser les données.
- Appliquer des techniques de machine learning pour extraire des connaissances pertinentes et utiles à la prise de décision.

Problématique :

Comment peut-on prédire efficacement l'état de santé global d'un individu à partir de ses habitudes de vie telles que l'alimentation, le sommeil, l'activité physique, la consommation d'alcool et le tabagisme ?

Étude de l'existant :

- Le domaine de la santé s'appuie de plus en plus sur les données comportementales.
- Les systèmes décisionnels aident les structures médicales à identifier les profils à risque.
- Ce projet s'inscrit dans la tendance de la **Health Data Science** avec un angle BI & ML.

Source de Données :

- Jeu de données : Kaggle – Health and Lifestyle Data for Regression
- Nombre de lignes : 1000
- Format : CSV

Structure des Données :

- **Âge** : Âge de l'individu en années (variable continue).
- **IMC** : Indice de Masse Corporelle de l'individu (variable continue).
- **Fréquence_Exercice** : Nombre de jours par semaine où l'individu fait de l'exercice (catégorique, valeurs de 0 à 7).
- **Qualité_Régime** : Un indice reflétant la qualité du régime alimentaire, avec des valeurs plus élevées indiquant des habitudes alimentaires plus saines (continue, 0100).

- **Heures_Sommeil** : Nombre moyen d'heures de sommeil par nuit (continue).
- **Statut_Fumeur** : Variable binaire où 0 = Non-fumeur, 1 = Fumeur.
- **Consommation_Alcool** : Nombre moyen d'unités d'alcool consommées par semaine (continue).
- **Score_Santé** : Un score de santé calculé reflétant l'état de santé global (continue, 0-100).

Technologies utilisées :

- **ETL** : Talend Open Studio
- **Visualisation** : Power BI
- **Machine Learning** : Python (Pandas, Scikit-learn, Seaborn)
- **Stockage** : PostgreSQL / MySQL (pour le data warehouse)
-

Étapes du Projet :

1. **Exploration et nettoyage des données** (fait ✓)
2. **Normalisation et réduction de la dimensionnalité**
3. **Conception du data warehouse** (schéma en étoile)
4. **Alimentation avec Talend**
5. **Visualisation interactive via Power BI**
6. **Modélisation (régression linéaire + RFE + Cross-Validation)**

Livrables attendus :

- PDF du cahier des charges (ce document)
- Rapports à chaque étape avec captures et interprétations
- Tableau de bord Power BI
- Script Python de modélisation
- Présentation finale