PROJET DE FIN DE FORMATION

Prédiction et Analyse du Risque de Maladie Coronarienne

(Framingham Heart Study)

---

1. Description Générale du Projet

Titre :
Analyse des Facteurs de Risque et Prédiction de la Maladie Coronarienne à 10 ans.

Contexte :
Les maladies cardiovasculaires constituent l'une des principales causes de mortalité dans le monde. Ce projet s'appuie sur les données de l'étude de Framingham afin d'explorer, analyser et modéliser les facteurs de risque associés au développement d'une maladie coronarienne (CHD) sur une période de 10 ans.

Rôle du participant :
En tant que data analyst junior, vous êtes mandaté par un institut de recherche en santé publique pour analyser ces données et produire un rapport détaillé. L'objectif est d'identifier les facteurs les plus prédictifs et de construire un modèle simple de score de risque pour aider à la prévention.

Données :
Le fichier framingham_heart_study.csv contient plus de 4 000 observations et 16 variables (données démographiques, habitudes de vie, antécédents médicaux, mesures biologiques, etc.).

---

2. Objectifs Pédagogiques et Compétences Visées

Ce projet vise à évaluer la capacité de l'apprenant à mener un projet d'analyse de données de bout en bout.

Les compétences évaluées sont :

- Analyse Exploratoire des Données (AED) : nettoyage, gestion des valeurs manquantes et aberrantes, analyses univariées et bivariées.
- Visualisation de données : production de graphiques clairs et pertinents.
- Statistiques descriptives : calcul et interprétation d'indicateurs clés.
- Modélisation statistique simple : mise en œuvre d'une régression logistique (variable cible binaire : risque CHD à 10 ans Oui/Non) et interprétation des coefficients (odds ratios).
- Communication : rédaction d'un rapport structuré et présentation orale synthétique.

---

3. Structure et Découpage du Projet en Phases

Le projet est divisé en quatre phases principales.

Phase 1 : Compréhension du Sujet et Préparation des Données

Objectif :
Prendre en main le dataset et le préparer pour l'analyse.

Tâches :

1. Importation et première inspection :
    - Charger les données avec Pandas.
    - Afficher les premières lignes (.head()).
    - Examiner les informations générales (.info()).
    - Produire les statistiques descriptives de base (.describe()).
2. Nettoyage des données :
    - Identifier et quantifier les valeurs manquantes.
    - Définir et justifier une stratégie de traitement (suppression ou imputation).
    - Vérifier et supprimer les doublons.
    - Détecter et traiter les valeurs aberrantes
3. Ingénierie des caractéristiques :
    - Créer des variables catégorielles à partir de variables continues si pertinent.
    - Vérifier et corriger les types de données (variables catégorielles, numériques, etc.).

Phase 2 : Analyse Exploratoire des Données (AED)

Objectif :
Explorer les données pour identifier tendances et relations.

Tâches :

1. Analyse univariée :
    - Histogrammes pour les variables continues (âge, IMC, pression artérielle, etc.).

- Diagrammes en barres pour les variables catégorielles (sexe, tabagisme, éducation).
- Description et interprétation des distributions.

2. Analyse bivariée :
    - Étudier la relation entre la variable cible (CHD à 10 ans) et chaque variable explicative.
    - Variables catégorielles : tableaux de contingence et diagrammes en barres groupées.
    - Variables continues : boxplots comparant les groupes CHD Oui/Non.
3. Matrice de corrélation :
    - Calcul et visualisation (heatmap).
    - Identification des corrélations fortes et discussion sur la multicolinéarité.

Phase 3 : Modélisation Prédictive Simple

Objectif :
Construire un modèle statistique pour prédire le risque de CHD.

Tâches :

1. Préparation des données :
    - Séparer les données en ensemble d'entraînement (70–80 %) et ensemble de test (20–30 %).
    - Encoder les variables catégorielles (variables indicatrices).
2. Construction du modèle :
    - Implémenter une régression logistique avec la variable cible TenYearCHD.
    - Entraîner le modèle sur l'ensemble d'entraînement.
3. Évaluation du modèle :
    - Prédictions sur l'ensemble de test.
    - Calcul des métriques : matrice de confusion, accuracy, precision, recall, F1-score.
    - Analyse particulière du recall (identifier correctement les personnes à risque).
    - Tracer la courbe ROC et calculer l'AUC.

Phase 4 : Interprétation et Rapport Final

Objectif :
Synthétiser les résultats et formuler des recommandations.

Tâches :

1. Interprétation des coefficients :
    - Analyse des coefficients ou odds ratios.
    - Identification des facteurs de risque les plus influents.
2. Discussion :
    - Limites du modèle.
    - Biais possibles.
    - Généralisation des résultats.
3. Rédaction du rapport final.
4. Préparation de la présentation orale.

---

4. Structure des Livrables

A. Rapport d'Analyse (PDF)

Le document doit contenir :

1. Page de garde (titre, nom, date, formation).
2. Résumé exécutif (1 page maximum).
3. Introduction (contexte médical et objectifs).
4. Méthodologie (données, nettoyage, analyse, modélisation).
5. Résultats :
    - Statistiques descriptives.
    - Visualisations clés.
    - Résultats de la modélisation.
    - Interprétation des facteurs de risque.
6. Discussion et conclusion.
7. Annexes :
    - Code propre et commenté.

- Détails des traitements des valeurs manquantes.

B. Présentation Orale (10–15 minutes)

Structure suggérée :

1. Titre et contexte.
2. Problématique et objectifs.
3. Aperçu des données.
4. Nettoyage et préparation.
5. Résultats exploratoires (2–3 graphiques clés).
6. Modélisation et performance.
7. Facteurs de risque principaux.
8. Conclusion et recommandations.
9. Limites et perspectives.

---

5. Organisation des Fichiers (Structure Recommandée)

projet_framingham/

- data/
  - framingham_heart_study.csv
- notebooks/
  - 01_exploration_nettoyage.ipynb
  - 02_analyse_exploratoire.ipynb
  - 03_modelisation.ipynb
- scripts/
  - fonctions_utils.py (optionnel)
- outputs/
  - figures/
  - modele_logistique.pkl (optionnel)
- rapport/
  - Rapport_Framingham_Nom_Prenom.pdf

- presentation/
    - Presentation_Framingham_Nom_Prenom.pptx
- README.md