Projet de Santé des Données : Préparation des données pour la prédiction des maladies hépatiques

Enseignant : Yohann Chasseray

18 novembre 2025

Introduction et modalités

Ce projet permet de mettre en application les différents concepts évoqués et travaillés au long du module de santé des données. Il sera effectué en **groupes de 2 étudiants**.

Rendu et attendus

Le travail sera évalué au travers des éléments suivants :

- Un **rapport écrit** permettant de détailler les étapes de traitement réalisées tout en justifiant les choix, les raisons et l'amélioration apportée par chaque traitement. Le rapport comportera une annexe indiquant l'implication en pourcentage de chaque membre du groupe dans la réalisation du projet et ses différentes tâches (développement, méthode, rapport, etc.).
- Un **fichier jupyter** ou un **fichier python** implémentant les fonctions utilisées pour faire le traitement du jeu de données. Vous remettrez également au format CSV, le jeu de données qui résulte du traitement.
- Une **soutenance orale** en présentiel, au cours de laquelle vous présenterez votre méthode et les étapes de nettoyage des données. Un focus sera fait sur une étape que vous souhaitez détailler. La modalité de la présentation est la suivante : **10 minutes de présentation** suivie de **5 minutes de questions**.

Le rendu du rapport et du code associé seront réalisés via un dépôt sur moodle. La date limite de dépôt est fixée à 23h59 la veille de la date de soutenance. Par souci d'équité, chaque heure de retard sur le dépôt coûtera un point de pénalité au groupe.

Utilisation d'outils génératifs

Vous êtes autorisés à utiliser des outils génératifs pour réaliser le projet, aux conditions suivantes :

- Vous devez ajouter la mention *ce travail a été réalisé à l'aide d'une intelligence artificielle*, en début de rapport.
- Vous devez indiquer explicitement chaque élément de votre travail qui a été généré par un outil génératif.

— Vous rédigerez un document annexe au rapport, qui fournit les prompts utilisés pour chacun des éléments générés, et apporte un regard critique sur la réponse obtenue.

Tout projet faisant appel à des outils génératifs et qui ne respecte pas ces contraintes ne peut pas espérer être évalué au-dessus de la moyenne. Les sections générées seront également corrigées avec plus d'exigence en terme de simplicité du code.

Objectif du projet

Le travail demandé consiste à nettoyer des jeux de données pour qu'il puissent être utilisés dans un algorithme de machine learning.

Descriptif des données

Des jeux de données, issus de deux hôpitaux différents, vous sont fournis. L'un (liver.json) est au format JSON, et l'autre (liver.csv) est au format CSV. Les jeux de données fournis décrivent les caractéristiques démographiques et physiques de patients dans le cadre de l'étude des facteurs favorisant l'apparition de maladies du foie.

Le jeu de données fourni par le premier hôpital (liver.json) contient 11 colonnes, dont le contenu est détaillé ci-dessous :

— **Age :** Tranche d'âge (années) dans laquelle se trouve le patient.
— **Gender :** Sexe du patient (Homme ou Femme).
— **Total_Bilirubin :** Taux de bilirubine totale mesuré en mg/dL.
— **Direct_Bilirubin :** Taux de bilirubine directe mesuré en mg/dL.
— **Alkaline_Phosphatase :** Niveau de phosphatase alcaline mesuré en UI/L.
— **Alamine_Aminotransferase :** Niveau de l'enzyme ALT mesuré en UI/L.
— **Aspartate_Aminotransferase :** Niveau de l'enzyme AST mesuré en UI/L.
— **Total_Proteins :** Concentration totale des protéines mesurée en g/dL.
— **Albumin :** Concentration sérique de l'albumine mesurée en g/dL.
— **Albumin_and_Globulin Ratio :** Rapport entre les concentrations d'albumine et de globuline.
— **Outcome :** Indique si le patient est atteint ou non d'une maladie du foie (1 : Patient atteint d'une maladie du foie, 2 : Patient sain).

Le jeu de données fourni par le deuxième hôpital (liver.csv) contient également 11 colonnes, dont le contenu est détaillé ci-dessous :

- **Age :** Tranche d'âge (années) dans laquelle se trouve le patient.
- **Gender :** Sexe du patient (Homme ou Femme).
- **Total_Bilirubin :** Taux de bilirubine totale mesuré en mg/L.
- **Direct_Bilirubin :** Taux de bilirubine directe mesuré en mg/L.
- **Alkaline_Phosphatase :** Niveau de phosphatase alcaline mesuré en UI/L.
- **ALT :** Niveau de l'enzyme ALT mesuré en UI/L.
- **AST :** Niveau de l'enzyme AST mesuré en UI/L.
- **Total_Proteins :** Concentration totale des protéines.
- **Albumin :** Concentration sérique de l'albumine mesurée en g/dL.
- **Albumin_and_Globulin Ratio :** Rapport entre les concentrations d'albumine et de globuline.
- **Result :** Indique si le patient est atteint ou non d'une maladie du foie (1 : Patient atteint d'une maladie du foie, 2 : Patient sain).