Projet de Machine Learning

Soutenances orales: 27 et 29 Mai 2026

Jeu de données

Les données sont issues du site du concours KAGGLE; il s'agit du jeu de données synthétiques "Cardivascular Disease Risk Prediction Dataset" disponible ici: https://www.kaggle.com/datasets/bertnardomariouskono/cardiovascular-disease-risk-prediction-dataset.

Ce jeu de données contient 15 000 dossiers médicaux synthétiques de patients spécialement conçus pour prédire le risque de maladie cardiovasculaire. Bien que synthétiques, les données sont générées à l'aide d'heuristiques médicales afin de garantir des corrélations réalistes entre les variables, telles que la relation entre l'âge, l'IMC et la pression artérielle. Le jeu de données comporte 19 variables pour 15000 patients :

- `Patient_ID` : Identifiant unique pour chaque patient
- `Age` : Age du patient
- `Gender` : Genre du patient (qualitative à deux modalités)
- `Height_cm` : Taille du patient en centimètres
- `Weight_kg` : Poids du patient en kilogrammes
- `BMI` : Indice de masse corporelle (IMC), calculé à partir de la taille et du poids en kg/m2
- `Systolic_BP` : Pression artérielle systolique (mmHg)
- `Diastolic_BP` : Pression artérielle diastolique (mmHg)
- `Cholesterol_Total` : Cholestérol sérique total (mg/dL)
- `Cholesterol_LDL` : Lipoprotéines de basse densité / « mauvais » cholestérol (mg/dL)
- `Cholesterol_HDL` : Lipoprotéines de haute densité / « bon » cholestérol (mg/dL)
- `Fasting_Blood_Sugar` : Taux de glycémie à jeûn (mg/dL)
- `Smoking_Status` : Statut tabagique (qualitative à deux modalités)
- `Alcohol_Consumption` : niveau de consommation d'alcool (qualitative à trois modalités)
- `Physical_Activity_Level` : Niveau d'activité physique (qualitative à quatre modalités)
- `Family_History` : Antécédents familiaux de maladie cardiaque (qualitative à deux modalités).
- `Stress_Level` : Niveau de stress auto-déclaré (échelle de 1 à 10)
- `Sleep_Hours` : Nombre moyen d'heures de sommeil par nuit
- `Heart_Disease_Risk` : Risque d'accident cardiaque (qualitative à deux modalités) 0 : risque faible, 1 : risque élevé.

Dans ce projet, on souhaite dans un premier temps, prédire la variable `Heart_Disease_Risk` à partir de toutes les autres variables, et dans un second temps, prédire la variable `Cholesterol_LDL` à partir de toutes les autres variables (excepté `Heart_Disease_Risk`).

Questions posées

Analyse exploratoire des données (langage R ou Python au choix)

L'objectif dans un premier temps est d'explorer les différentes variables, étape préliminaire indispensable à l'analyse. Ci-dessous sont précisées quelques questions basiques. Vous pouvez compléter l'analyse selon vos propres idées.

1. Commencez par vérifier la nature des différentes variables et leur encodage. N'oubliez pas de convertir toutes les variables qualitatives.
2. Commencez l'exploration par une analyse descriptive unidimensionnelle des données. Des transformations des variables quantitatives vous semblent-t-elles pertinentes ?
3. Poursuivez avec une analyse descriptive bidimensionnelle. Utilisez des techniques de visualisation: par exemple les nuages de points (*scatterplot*), des graphes des corrélations, des boîtes à moustaches parallèles, *mosaicplot*... Quelles variables semblent liées ?
4. Réalisez une analyse en composantes principales des variables explicatives quantitatives et interprétez les résultats. Visualisez les dépendances éventuelles entre les variables à prédire et les variables explicatives.

Modélisation (langages R et Python)

Avant de commencer cette partie, pensez à bien faire les mêmes transformations de variables éventuelles pour la suite dans les deux langages.

Prédiction du risque d'accident cardiaque

Nous considérons maintenant le problème de la prédiction la variable `Heart_Disease_Risk` à partir des autres variables du point de vue de l'apprentissage automatique, c'est-à-dire en nous concentrant sur les performances du modèle. L'objectif est de déterminer les meilleures performances que nous pouvons attendre, et les modèles qui les atteignent. Voici quelques questions pour vous guider.

1. Divisez le jeu de données en un échantillon d'apprentissage et un échantillon test. Vous prendrez un pourcentage de 20% pour l'échantillon test. Pourquoi cette étape est-elle nécessaire lorsque nous nous concentrons sur les performances des algorithmes ?
2. Comparez les performances d'un modèle linéaire (éventuellement généralisé) avec/sans sélection de variables, avec/sans pénalisation, d'un SVR/SVM, d'un arbre optimal, d'une forêt aléatoire, du boosting, et de réseaux de neurones. Justifiez vos choix (par exemple le noyau pour le SVR/SVM), identifiez les hyperparamètres de chaque modèle et ajustez-les soigneusement (par validation croisée). Interprétez les résultats et quantifiez l'amélioration éventuelle apportée par les modèles non linéaires.
3. Comparez les différents modèles optimisés sur votre échantillon test. Quels sont les modèles les plus performants ? Quel est le niveau de précision obtenu ? Quels modèles retenir si l'on ajoute une contrainte d'interprétabilité ?
4. Interprétation et retour sur l'analyse des données: vos résultats sont-ils cohérents avec l'analyse exploratoire des données, par exemple en ce qui concerne l'importance des variables ?

Prédiction de la variable `Cholesterol_LDL`

Reprenez les étapes précédentes pour la prédiction la variable `Cholesterol_LDL` (« mauvais » cholestérol) à partir de toutes les autres variables (excepté `Heart_Disease_Risk`).

Modalités et évaluation

Vous réaliserez le projet par groupe de 4 étudiant.e.s. L'évaluation portera sur une soutenance orale et deux notebook Jupyter (un en R et un en Python).

Travail à rendre : Comme livrable, chaque groupe déposera **au plus tard** sur Moodle :

- **le 22 Mai à 18H30**, un fichier zip contenant les deux notebooks Jupyter (R et Python) compilés,
- **le 22 Mai à 18H30**, les slides de l'exposé **au format pdf**.

Soutenances orales les 27 et 29 Mai 2026 : 20 minutes de présentation, puis 5 à 10 minutes de questions. L'exposé doit comprendre une introduction présentant les données ainsi que toutes les transformations que vous avez effectuées, une description succincte des algorithmes utilisés (en précisant bien quels hyperparamètres vous avez optimisés et comment), une interprétation des résultats, et une conclusion. Les questions pourront porter sur votre code (donc pensez à ouvrir vos notebooks et si possible les compiler avant la soutenance).

Critères d'évaluation : L'évaluation tiendra compte de la qualité de présentation orale (clarté, argumentation, interprétation des résultats etc.), de la cohérence de l'étude, de la qualité de présentation des notebooks (n'oubliez pas de commenter votre code), des interprétations des résultats (graphiques et autres).
L'IA générative ne peut être éventuellement utilisée que pour améliorer le travail que vous avez préalablement effectué sans l'utiliser (et non pour réaliser le travail à votre place). Ceci s'applique pour l'ensemble du travail : pour le code et pour la rédaction des commentaires et des slides. Afin de vérifier que cette consigne est respectée, certaines questions porteront sur votre code : toute incompréhension de votre propre code sera fortement pénalisée.

Analyse descriptive unidimensionnelle :
- Variable qualitatives : Histogrammes des effectifs : sur/sous exprimé

- Variables quantitatives : Centrer Réduire ? Pq ?,
Boxplots (différence de hauteur, outliers, dispersion autour des valeurs centrales
écart interquarticale (Q3-Q1), répartition symétrique (médiane située au centre de la box), adjacente inférieure (Q1−1.5×IQR) : moustache)
Histogrammes

Analyse bi-dimensionelle : scatterplot, mosaicplot, matrice de correlation, boxpplot
- Entre variables qualitatives : matrice de correlation -> correlation positive/ négative ?
- Entre variables qualitatives et qualitative : boxplot qualitative contre qualitative,