Projet de Machine Learning

Soutenances orales: 28 Mai 2025

Jeu de données

Les données sont issues du site du concours KAGGLE; il s'agit du jeu de données " Gym Members Exercise Dataset" disponible ici: https://www.kaggle.com/datasets/valakhorasani/gym-members-exercise-dataset.

Ce jeu de données fournit un aperçu détaillé des routines d'exercice, des attributs physiques et des mesures de la condition physique des membres d'une salle de sport. Il contient 15 variables observées chez 973 individus fréquentant une salle de sport:

- `Age` : âge du membre de la salle de sport.
- `Gender` : Sexe du membre de la salle de sport (qualitative à deux modalités : homme ou femme).
- `Weight..kg.` : Poids du membre en kilogrammes.
- `Height..m.` : Taille du membre en mètres.
- `Max_BPM` : Fréquence cardiaque maximale (battements par minute) pendant les séances d'entraînement.
- `Avg_BPM` : Fréquence cardiaque moyenne pendant les séances d'entraînement.
- `Resting_BPM` : Fréquence cardiaque au repos avant l'entraînement.
- `Session_Duration..hours.` : Durée de chaque séance d'entraînement en heures.
- `Calories_Burned` : Total des calories brûlées au cours de chaque séance.
- `Workout_Type` : Type d'entraînement effectué (qualitative à 4 modalités : cardio, musculation, yoga, HIIT).
- `Fat_Percentage` (Pourcentage de graisse) : Pourcentage de graisse corporelle du membre.
- `Water_Intake..liters.` : Consommation quotidienne d'eau pendant les séances d'entraînement.
- `Workout_Frequency..days.week.` : Nombre de séances d'entraînement par semaine (qualitative à 4 modalités : 2 à 5).
- `Experience_Level` : Niveau d'expérience (qualitative à 3 modalités : 1 pour débutant à 3 pour expert).
- `BMI` : Indice de masse corporelle (IMC), calculé à partir de la taille et du poids.

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

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 correlations, 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 des calories brûlées

Nous considérons maintenant le problème de la prédiction la variable `Calories_Burned` 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), et ajustez soigneusement les hyperparamètres de chaque modèle (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 du niveau d'expérience

Reprenez les étapes précédentes pour la prédiction la variable `Experience_Level` à partir de toutes les autres variables.

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 25 Mai à 23H59**, un fichier zip contenant les deux notebooks Jupyter (R et Python) compilés,
- **le 27 Mai à 18H30**, les slides de l'exposé **au format pdf**.

Soutenances orales les 28 Mai 2025: 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 juste 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).