Projet de Machine Learning

Deadline: 22 Mai 2023

Jeu de données

Les données ont été extraites et mises en forme à partir du jeu de données "QSAR biodegradation" disponibles sur le site UCI Machine Learning Repository.

Référence : Mansouri Kamel, Ringsted Tine, Ballabio Davide, Todeschini Robert, Consonni Viviana. (2013). QSAR biodegradation. UCI Machine Learning Repository.

Le jeu de données comprend 1055 observations et 41 variables, parmi lesquelles nous en avons retenu 20. L'objectif est de classifier 1055 molécules en deux classes (biodégradable ou non biodégradable). Les molécules sont décrites par les variables suivantes, qui correspondent à des descripteurs moléculaires:

- $SpMax_L$: Leading eigenvalue from Laplace matrix
- $J_Dz.e.$: Balaban-like index from Barysz matrix weighted by Sanderson electronegativity
- $C.$: Percentage of C atoms
- $SdssC$: Sum of dssC E-states
- $HyWi_B.m.$: Hyper-Wiener-like index (log function) from Burden matrix weighted by mass
- LOC: Lopping centric index
- $SM6_L$: Spectral moment of order 6 from Laplace matrix
- $F03.C.O.$: Frequency of C - O at topological distance 3 (variable qualitative)
- Me : Mean atomic Sanderson electronegativity (scaled on Carbon atom)
- Mi : Mean first ionization potential (scaled on Carbon atom)
- $SpPosA_B.p.$: Normalized spectral positive sum from Burden matrix weighted by polarizability
- $B01.C.Br.$: Presence/absence of C - Br at topological distance 1 (variable qualitative)
- $B03.C.Cl.$: Presence/absence of C - Cl at topological distance 3 (variable qualitative)
- $SpMax_A$: Leading eigenvalue from adjacency matrix (Lovasz-Pelikan index)
- $Psi_{i1}d$: Intrinsic state pseudoconnectivity index - type 1d
- $SdO.$: Sum of dO E-states
- $TI2_L$: Second Mohar index from Laplace matrix
- $SpMax_B.m.$: Leading eigenvalue from Burden matrix weighted by mass
- Psi_{iA} : Intrinsic state pseudoconnectivity index - type S average
- $SM6_B.m.$: Spectral moment of order 6 from Burden matrix weighted by mass.

La variable à prédire Y est une variable qualitative (binaire) : $Y = 0$ si la molécule n'est pas biodégradable et $Y = 1$ si la molécule est biodégradable.

Questions posées

Analyse exploratoire des données

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 une analyse descriptive unidimensionnelle des données. Voyez-vous des anomalies?
2. Des transformations des variables quantitatives vous semblent-t-elles pertinentes ? Certaines variables qualitatives ont de nombreuses modalités, il peut être pertinent de regrouper certaines modalités entre elles.
3. Les distributions sont-elles comparables entre le jeu d'apprentissage et le jeu de test ?
4. Poursuivez avec une analyse descriptive multidimensionnelle. Utilisez des techniques de visualisation : par exemple scatterplot, correlation plot ... Analysez les dépendances entre les variables.
5. Réalisez une analyse en composantes principales des variables quantitatives et interprétez les résultats.
6. Visualisez la possible dépendance entre les variables qualitatives et la variable à prédire.

Modélisation

Nous considérons maintenant le problème de la prédiction (classification binaire) 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. Tout d'abord, divisez les 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 de régression logistique avec/sans sélection de variables avec/sans pénalisation, d'un 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 SVM), et ajustez soigneusement les paramètres (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 ?
4. Interprétation et retour sur l'analyse des données. Vos résultats sont-ils cohérents avec l'analyse préliminaire des données, par exemple en ce qui concerne l'importance des variables ?

Organisation et rapport à rendre

Vous réaliserez le projet par groupe de 4 étudiant.e.s. **Deadline: 22 Mai 2023**. Comme livrable, vous rendrez un rapport au format pdf ne dépassant pas 30 pages. Il doit comprendre une introduction, une description succincte des algorithmes utilisés, une interprétation des résultats, une conclusion, etc. De plus, vous rendrez deux notebooks Jupyter, un en R et un en Python, incluant les sorties des codes. N'oubliez pas de commenter votre code. Le dépôt se fera sur Moodle : chaque groupe téléchargera un fichier zip contenant le rapport (format pdf) et les deux notebooks Jupyter.

L'évaluation tiendra compte de la présentation du rapport et de la rédaction (clarté, argumentation, etc.), de la cohérence de l'étude, de la qualité de présentation des notebooks, des interprétations des résultats (graphiques et autres).