Cours PCD – Labo 3 : Imputation de données manquantes

Résumé

L'objectif est d'appliquer plusieurs méthodes **d'imputation** à un jeu de données, examiner leur effet sur les performances de **prédiction** d'une colonne, et discuter les résultats dans un court **rapport**, par exemple des cellules *markdown* d'un *notebook* Jupyter.

Déroulement du travail et questions auxquelles vous devez répondre

1. Téléchargez les données fournies par la ville de Boston (USA) sur les salaires de ses employés en 2021 : https://data.boston.gov/dataset/employee-earnings-report – si le site n'est pas accessible, le jeu est également fourni sur Cyberlearn. Chargez les données dans un *Data Frame.*
2. Combien de lignes et de colonnes a le *Data Frame* ? Pour chaque colonne, combien y a-t-il de données manquantes ? Y a-t-il des lignes entièrement vides ? Si oui, veuillez les supprimer.
3. Veuillez convertir les colonnes qui représentent des nombres dans un type numérique Python.
4. Cherchez des données aberrantes (*outliers*) par une méthode univariée simple, à votre choix. Si nécessaire, vous pouvez encore supprimer jusqu'à 6 personnes.
5. La tâche de prédiction est la suivante : pour chaque personne, prédire l'attribut 'DEPARTMENT_NAME' uniquement à partir des informations sur les salaires et les primes (valeurs numériques). Veuillez donc supprimer les colonnes non pertinentes.
6. Existe-t-il des classifieurs dans Scikit-learn qui acceptent des données manquantes ?
7. Testez quatre méthodes d'imputation différentes (par zéro, par la médiane, 'KNN' et 'iterative') et trois classifieurs différents (*k* plus proches voisins (KNN), arbre de décision (Decision Tree), et forêt d'arbres de décision (Random Forest).
    a. Veuillez créer des ensembles d'entraînement et de test aléatoires (avec sklearn. model_selection.train_test_split), en gardant 10% des données pour le test.
    b. Si on attribue à tous les items la classe majoritaire, quel score F1 avec *micro*-moyenne obtient-on ? (On peut répondre par un simple calcul.)
    c. Quelle est la combinaison parmi les 12 qui a le meilleur score F1 avec *micro*-moyenne ?
    d. Comment se comparent les scores F1 avec *micro*-moyenne par rapport à ceux calculés avec une *macro*-moyenne ?
    e. Commentez brièvement les scores observés et leur différences.