Apprentissage supervisé (Partie 2)

1. *Modèles basés sur la descente de gradient*

- ***Analyse des corrélations*** *: On souhaite maintenant modifier le modèle.*
  1. *En utilisant la fonction heatmap de seaborn affichez la matrice des corrélations. Que constatez-vous.*
  2. *Utilisez maintenant la SelectKBest avec un test d'Anova (f_classif) pour identifier les 4 variables les plus significatives. Les résultats sont-ils en adéquation avec ceux trouvés à la question précédente. La fonction get_support() permet d'identifier les features sélectionnées et la variable pvalues_ de connaitre les valeurs.*
  3. *Utilisez maintenant SelectKBest associée à la métrique (mutual_info_classif) pour identifier les variables qui expliquent le mieux la target. Que constatez-vous.*
  4. *Découpez le DataFrame en deux DataFrame correspondant aux deux classes et pour chaque features affichez via kdeplot de seaborn les distributions de ces classes. Vous découperez la figure d'affichage en une zone 3 lignes et 3 colonnes via la fonction subplots de matplotlib.pyplot et vous afficherez les courbes dans chacune de ces zones.*
- ***Modèles SVC et Gaussien*** *:*
  1. *Si l'on souhaite utiliser un modèle de descente de gradient, il est nécessaire de centrer et réduire les données pour que les différences de valeurs n'impactent pas le résultat. Créer une fonction permettant de centrer et réduire les données, vous utiliserez la méthode StandarScalar pour cela.*
  2. *Utilisez un premier modèle de régression linéaire SGDClassifier pour traiter le data. Quel résultat obtenons-nous ?*
  3. *Compte tenu du peu de corrélation entre les variables (indépendance des features) les classificateurs bayésiens pourraient donner des résultats. Utiliser les modèles BernoulliNB et GaussianNB pour l'estimation. Quels sont les résultats obtenus.*

4. *Utilisez maintenant un modèle KNeighborsClassifier pour l'estimation. Estimez en faisant varier l'hyper-paramètre n_neighbors entre 2 et 50 quel est le meilleur paramètre.*
5. *Quel est le résultat avec un réseau de neurones MLPClassifier.*
6. *Quel est le résultat avec un réseau de neurones SVC.*
7. *Que peut-on en conclure concernant les algorithmes de descente de gradient.*

2. *Finalisation*

- ***Méthodes ensemblistes** : On va commencer par tester si d'autres approches sont efficaces.*
    1. *Testez des modèles de type AdaBoostClassifier, BaggingClassifier, GradientBoostingClassifier. Quels sont les résultats obtenus par ces trois modèles.*
    2. *Comme vu dans la partie 3, les classes sont déséquilibrées. Comment se comportent les modèles ensemblistes si les données sont équilibrés*
    3. *On souhaite maintenant tester si le déséquilibre entre les classes a un impact sur l'apprentissage, et donc sur les résultats obtenus.*
- ***Polynomial expension** : Nous essayons de voir si l'expansion polynomiale de degré 2 peut avoir une influence sur les résultats.*
    1. *A partir d'un dataFrame sur lequel vous devez supprimer les nan et scalariser les données, modifier les noms des colonnes pour ne garder que les deux premiers caractères. La fonction rename permet de renommer les colonnes, pour cela il est nécessaire de lui fournir un dictionnaire de transformation. Le dictionnaire peut être créé en utilisant la fonction dict appliquée au zip entre les columns du DataFrame et la liste des nouveaux noms.*
    2. *Créer un modèle PolynomialFeatures et effectué un fit_transform pour créer les nouvelles colonnes. Vous enregistrerez le résultat dans un dataFrame.*
    3. *Renommer les colonnes en fonction des noms générés par le model que l'on trouve dans get_feature_names_out.*
    4. *Lancer un apprentissage GradientBoostingClassifier avec un n_estimator=300 et RandomForestClassifier avec les paramètres obtenus précédemment. Quels sont les résultats obtenus.*

- ***Traitement des Outliers*** *: Les recommandations internationales précisent que la qualité de l'eau potable ne peut excéder certaines valeurs pour certains paramètres.*
    1. *Il est précisé que le ph doit obligatoirement être compris entre 6.5 et 8.5. Identifiez combien de fois le Data indique que l'eau est potable avec des taux de ph inférieur à 6 et supérieur à 9. Supprimez ces valeurs quels sont les résultats obtenus.*
    2. *Pour le sulfate les valeurs ne doivent pas être supérieure à 250. Supprimez toutes les valeurs supérieures à 400. Quels sont les résultats obtenus.*
    3. *Quels sont les résultats de l'apprentissage si l'on supprime les valeurs extrêmes sur toutes les variables pour les données potables. Rappel : On considèrera qu'une valeur est extrême, pour la potabilité, si elle s'écarte de la moyenne de plus de 3 ou 4 fois la valeur la dispersion (std).*