Apprentissage SVC

Exercice 1 : Classification des pingouins

- *Lire les données à partir du fichier* Penguins.csv. *Attention le caractère séparateur est le ' ;'.*
- *Rechercher et afficher les individus qui ont au moins une de ses* feature *à NaN. Pour cela vous pourrez combiner la fonction* isna *avec la fonction* any *à appliquer sur les colonnes* axis=1.
- *Parmi les individus ayant au moins une de leur* feature *à* NaN *identifiez ceux qui doivent être absolument supprimés. Récupérer les index des individus à supprimer et supprimez les de la base de données.*
- *Calculez maintenant le poids moyen des individus de sexe mâle et femelle.*
- *En fonction de ce poids moyen nous souhaitons remplacer les valeurs manquantes sur les sexes en fonction de leur poids. Vous affecterez à l'individu le sexe correspondant à son poids, c'est-à-dire aura le sexe femelle si son poids et plus proche du poids moyen des femelles et le sexe mal sinon.*
- *Pour cela sélectionnerez dans un premier temps les individus qui n'ont pas la* feature sex *renseignée, et ensuite grâce à la fonction* iterrows *pour pourrez parcourir les individus en ligne. Pour modifier la case d'un* DataFrame *vous utiliserez l'accesseur* at[index,feature].
- *Remplacez les valeurs de la* feature sex *par les entiers 0 et 1. Vous utiliserez pour cela la fonction* apply.
- *Nous souhaitons maintenant faire de même pour la feature island. Vous utiliserez dans ce cas la fonction replace à laquelle vous passerez un dictionnaire indexé sur les iles et avec comme valeur les codes à appliquer : { 'Torgersen' : 0, 'Biscoe' : 1 … }*
- *Découper maintenant le* DataFrame *en un jeu de test et un jeu d'entrainement. Vous effectuerez une découpe 30/70.*
- *Utilisez un modèle de type SVC pour classer les individus.*
- *Récupérez ensuite la liste des valeurs prédites et utilisez la fonction* heatmap *de* seaborn *pour afficher la matrice de confusion. Que constatez-vous.*

- *Utilisez maintenant le modèle SVC avec un kernel 'linear', les résultats sont-ils meilleurs.*

Exercice 2 : Classification des champignons

- *Lire les données sur les champignons dans le fichier Champignons.csv.*
- *Combiens y-a-t-il des données dans le DataFrame.*
- *Y-a-t-il des données manquantes. Dans ce cas supprimés ces données.*
- *Affichez les pourcentages de champignons toxiques et comestibles.*
- *Effectuez une copie de la base d'origine.*
- *Nous voulons encoder les données catégorielles autres que la target Classe. Pour cela vous utiliserez l'encodeur LabelEncoder de la bibliothèque preprocessing de sklearn. Pour cela vous devez donc créer un modèle de type LabelEncoder et ensuite l'entrainer (fit) sur une colonne et transformer les données (transform) sur cette même colonne. La fonction fit_transform effectue l'opération en un coup.*
- *Vous devrez donc parcourir toutes les colonnes du DataFrame et les transformer. Attention vous ne devez pas transformer la target.*
- *Découpez maintenant le DataFrame en un jeu de teste et un jeu d'entrainement avec une découpe de 0.33.*
- *Cette découpe respecte-t-elle la répartition en pourcentage entre les individus toxiques et comestibles. Vous afficherez donc ces pourcentages pour les y test et train.*
- *Utilisez un modèle de support vector machine pour entrainer et estimer la toxicité des champignons. Quels sont les résultats obtenus.*
- *Grâce à la matrice de confusion affichez les différences entre les valeurs prédites et celles attendues.*
- *Recherchez quels sont les individus qui ont été mal classés et affichez-les.*