Apprentissage Automatique 1

TP N°4 – Régression Logistique

Exercice N°1 : Prédictions des diabètes

Créer un classifieur de la régression logistique pour prédire si une personne aura un diagnostic de diabète positif ou non.

Nous utilisons une base de données des personnes qui ont été diagnostiquées positives avec le diabète et d'autres ont été diagnostiquées négatives.

L'objectif est de classifier les personnes en deux classes (diabètes et non diabètes) à travers d'un ensemble des caractéristiques.

Explication de la base de données :

- ✓ **Pregnancies:** Nombre de grossesses
- ✓ **Glucose:** Concentration de glucose dans le plasma à 2 heures d'un test de tolérance au glucose oral.
- ✓ **BloodPressure:** Pression artérielle diastolique (mm Hg)
- ✓ **SkinThickness:** Épaisseur du pli cutané du triceps (mm)
- ✓ **Insulin:** insuline sérique à 2 heures (mu U/ml)
- ✓ **BMI:** Indice de masse corporelle (poids en kg/(taille en m)^2)
- ✓ **DiabetesPedigreeFunction:** Fonction de pedigree de diabète
- ✓ **Age:** Age (Année)
- ✓ **Outcome:** Variable de classe (0 ou 1)

1. Importer les bibliothèques nécessaires
2. Lire la base de données avec **pandas**.
3. Afficher la longueur (nombre d'enregistrements) de la base de données.
4. Afficher les 10 premières lignes de la base de données.
5. Afficher les informations et une description du dataset
6. Préciser **les entrées X** du classifieur Régression logistique et **la sortie y**
7. Chercher les valeurs IsNan et tracer le graphe nécessaire en utilisant la fonction : heatmap(arguments)
8. Tracer le graphe qui montre la distribution des personnes ceux ayant la maladie diabètes et celles n'ayant pas, en utilisant la fonction : **countplot**(argument)
9. Tracer la distribution de l'Age par la fonction : **distplot**(argument)
10. Afficher la matrice de corrélation sous forme d'un tableau
11. Afficher la distribution de toutes les variables du dataset : sns.**pairplot**(argument)

12. Afficher la distribution de l'âge en fonction de BMI sous d'une boxplot .
13. Diviser la base de données en 2 parties : **apprentissage (80%** du nombre total des enregistrements) et **test (20%).** Utiliser la fonction train_test_split.
14. Appliquer l'apprentissage automatique.
15. Calculer la prediction
16. Afficher le rapport de la classification.
17. Calculer la matrice de confusion.
18. Calculer l'accuracy_score.
19. Afficher la matrice de confusion sous cette forme :