Enseignants	TP2	Classe : 3ème GLSI
Cours : A. NAJJAR-M. FARHAT-I. BEN OTHMEN **TP : F. JENHANI- S.BIROUZA – I. HAMROUNI**	**Machine Learning**	

Partie 1 : Visualisation des données

Dans cette partie, nous allons travailler sur des données des annonces de vente et location de biens en Tunisie.

1. Placer l'ensemble de données dans votre espace Drive.
2. Charger les données.
3. Afficher les premières lignes de la base
4. Afficher quelques informations autour de la base
5. Utiliser la méthode plot de la bibliothèque "**matplotlib"** pour afficher des bars mesurant le nombre de biens par catégorie
6. Utiliser la méthode "**countplot"** de la bibliothèque "**seaborn**" pour afficher le nombre de biens par ville
7. Afficher un histogramme de biens en vente ou en location par ville en utilisant la méthode "**subplot"** de la bibliothèque "**matplotlib**"
8. Afficher le nombre de biens en vente vs le nombre de biens en location moyennant la méthode "**countplot**" de la bibliothèque "**seaborn**"
9. Faites un affichage du nombre de biens par cité trié en décroissant.
10. Créer une **dataFrame** ne contenant que les appartements avec uniquement des informations autour du prix et de la taille.
11. Faites un affichage en nuage de point de l'évolution des prix des appartements en fonction de leur taille.
12. Qu'est-ce que vous remarquez?
13. Redéfinissez la **dataframe** des appartements en utilisant la colonne "**log_price"** au lieu de la colonne "**price"** et rajouter la colonne "**type"**
14. Refaire l'affichage en nuage de points des prix en fonction des tailles.
15. Utiliser les couleurs pour différencier les appartements en location de ceux en vente.

Partie 2 : Préparation des données

Plusieurs techniques de traitement des données peuvent être utilisées. Le choix de la technique appropriée dépend du contexte.

Dans cette partie, nous utilisons la base "**pima-indians-diabetes.data.csv**" manipulée lors du premier TP.

a. Gérer les données manquantes

Certaines instances peuvent avoir des attributs ayant des valeurs nulles. Ce qui peut dégrader les performances d'un algorithme d'apprentissage. La solution la plus simple est de supprimer les individus dont les valeurs de certains attributs sont manquantes. Mais, ceci peut causer la perte de données importantes. Un seconde alternative et de déterminer, la valeur médiane de chaque attribut, puis de remplacer les valeurs nulle par la médiane.

Questions

1. Lire le contenu du fichier "**pima-indians-diabetes.data.csv**" et en extraire les valeurs des attributs.
2. Filtrer les valeurs de l'attribut '**SkinThick'** pour déterminer les valeurs non nulles.
3. Calculer la médiane de ces valeurs
4. Remplacer les valeurs nulles de l'attribut "**SkinThick"** par la médiane**.**
5. Pourquoi on ne peut pas faire la même chose avec l'attribut **" NumTimesPrg"** ?

b. Uniformisation d'échelle

Le changement d'échelle est une sorte de normalisation. L'intervalle dans lequel vari les variables numériques peut être différent selon l'attribut. Ceci peut influencer les performances de certains algorithmes d'apprentissage automatique, surtout ceux qui se basent sur le calcul de distances. Pour cela, le but de cette étape est de ramener toutes les valeurs dans l'intervalle [0,1]. Une des techniques utilisées pour la normalisation est la suivante :

$$Val_{norm} = \frac{Val - Val_{min}}{Val_{max} - Val_{min}}$$

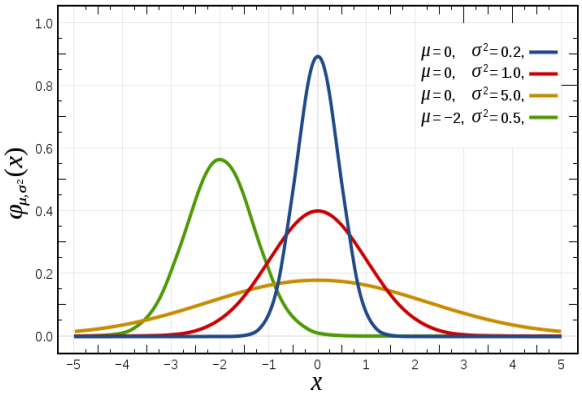
Questions

1. Ramener les valeurs de tous les attributs dans un intervalle [0-1].
2. Réafficher les données.

c. Normalisation

La standardisation des données est aussi une sorte de normalisation. Ramener les valeurs des attributs à l'intervalle [0-1] est parfois insuffisant surtouts dans le cas des bases qui contiennent beaucoup de zéros ou des algorithmes qui multiplient ces valeurs par une certaine pondération. Les valeurs d'un attribut sont transformées pour suivre une loi gaussienne ayant une moyenne **μ=0** et un écart type **σ=1**.

$$\varphi_{\mu,\sigma^2}(x) = \frac{1}{\sigma.\sqrt{2\pi}}.e^{-\frac{1}{2}.\left(\frac{X-\mu}{\sigma}\right)^2}$$



Questions

1. Normaliser les valeurs des attributs.
2. Réafficher les données.