**Ecole Militaire Polytechnique Chahid Abbderrahmane Taleb**

**Série de TD-TP en fouille de données 2021**

**Exercice 1 :**

1. Quelles sont les hypothèses nécessaires pour que le classifieur LDC (linear discriminant classifieur) donne des meilleurs taux de classification ?
2. Quelle est la forme de la frontière séparant deux classes lorsque nous utilisons le classifieur LDC ? justifier votre réponse par le développement de la fonction de décision.
3. Les deux questions précédentes 1) et 2) pour le classifieur QDC.

**Exercice 2 :**

1. Charger le dataset public iris .
2. Décrire les éléments de ce dataset
3. Afficher la distribution des classes pour les différentes paires des attributs, que constatez-vous ?
4. Subdiviser ce dataset en deux parties égales (train|test)
5. Réaliser l’apprentissage KNN (k=4)
6. Mesure les différentes métriques de test.

**Exercice 3 :**

1. Ecrire une fonction qui implémente la validation croisée et qui retourne le taux d’erreur.
2. Evaluer cette fonction pour la configuration suivante (k=5 ; classifieur =LDC).
3. Utiliser cette fonction et conduire une recherche empirique (écrire une autre fonction) pour l’identification des meilleurs noyaux SVM et les paramètres SVM optimaux.

**Exercice 4: Participer dans une compétition**

1. Pour tous les élèves:
2. Choisir un autre dataset volumineux contenant plusieurs milliers d’instances, soit A ce dataset.
3. Sélectionner aléatoirement 500 instances pour le test final, soit A cet ensemble. Ainsi, B est le même pour tous les élèves.
4. Pour chaque élève :
5. Vous n’avez pas le droit d’utiliser B, utiliser seulement A-B pour la construction d’un classifieur :

-utiliser un classifieur populaire (KNN, LDC, QDC,…etc)

-expliquer la démarche suivie

1. Test final :
2. Chaque élève utilise son classifieur et puis teste l’ensemble B. Les élèves seront classés dans la compétition selon les performances obtenues.