Dans le TP 2, l'étudiant doit

- Respecter le format du TP 1 différent des années passées (2022)
- Créer son propre dataset sous format ARFF et sur des données réelles.
    a. Il ne doit pas ramener un dataset du net pour ce TP spécialement.
    b. Il doit citer les références (pages web et ressources) utilisées pour créer le dataset
    c. Nombre d'attributs min =5 plus une ne classe(min=6).
    d. Nombre min d'instances= 50. (Un nombre petit d'instances peut ne pas être suffisant pour l'apprentissage et le test. Je vous conseille un dataset assez grand)

- Le sujet est laissé au choix de l'étudiant : ça permettra d'avoir des travaux personnels.
    a. Pas deux ne se ressembleront : Les mêmes attributs ou valeurs de la classe ou nombre d'instances impliquent que les étudiants ont copié le TP(d'autant plus que vous allez envoyer la version numérique)

- Appliquer les 5 algorithmes comme dans le TP1 avec les méthodes d'évaluation que vous avez :
    a. Cross-Validation **avec variation de nombre de fold (3 folds différents minimum)**
    b. Pourcentage split **avec variation du pourcentage (3 pourcentages différents minimum)**
    c. Leave one out

- Dois être capable de lire et d'interpréter les résultats du processus data Mining à la fin :
    a. C'est-à-dire analyser la connaissance en sortie sous forme de règles ou d'arbres.
    b. Et même de conclure sur l'intérêt et la bonne marche de son TP ou du contraire (dataset de bonne qualité ou non).

Pour la partie création de dataset : À titre d'exemple et pour clarifier les choses, je donne le sujet suivant:

- Certaines villes dans le monde sont connues pour être **bonnes à y vivre** et d'autres **non**
- On a donc deux valeurs dans la classe.
- L'information de bonne ville ou de mauvaise ville est disponible sur le Net.
- L'étudiant trouvera donc des exemples minimums des deux classes (en tous 30 exemples minimum)
- Il faut maintenant qu'il trouve ou propose les **attributs**. Par exemple : nom, pays, la culture, la météo, la langue, le niveau de vie, la religion, l'éducation, la santé, l'économie, la race, la criminalité, le régime…etc.
- Toutes ses informations doivent être réelles et elles peuvent être facilement ramenées du Net (d'où le besoin d'avoir des références dans le rapport).
- Une fois le fichier constituer, on passera au processus data Mining et à application de multiples algorithmes, méthodes d'évaluation et à l'analyse des découvertes. C'est à dire pourquoi selon l'algorithme et les données, certaines villes sont considérées bonne et d'autres non.

Le sujet des villes n'était qu'un exemple, ça peut être n'importe quoi **smartphone haut de gamme, voiture, livres**...etc.

Soyez originaux, si vous voulez une bonne note. Répéter les sujets des smartphones/villes ne vous donnera qu'une note moyenne et à condition qu'il soit juste (note sur 1,25 si le sujet est ville, note sur 1,75 si sujet est smartphone, livre dans ce cas, et sur 2.0 si le sujet est original).

La partie implémentation :
L'étudiant doit faire appel à une API (weka ou autre) qui lui permet d'exécuter :

- Un algorithme d'apprentissage au choix sur votre dataset
- La méthode d'évaluation Cross-Validation avec variation de nombre de fold sur votre dataset
- La méthode pourcentage split avec variation du pourcentage sur votre dataset
- Plus une interprétation (affichage du fold ou du pourcentage split qui donne les meilleurs résultats)
- La note est sur 01 point (compréhension ligne par ligne requise)