SAES5 BUT3 (Aide à la décision et Modélisation Mathématiques)

Le but de cette SAE est de mettre en place des modèles pour réaliser des prédictions (classification). Vous allez travailler en binôme (vous êtes 29 étudiant(e)s donc 13 binômes et 1 trinôme).

Vous devez choisir un jeu de données parmi ceux disponibles ici : datasets.html#amazon_reviews et qui respecte les contraintes suivantes:

- il ne faut pas que 2 binômes différents travaillent sur le même jeu de données
- il faut au moins qu'une variable caractéristique corresponde à des avis d'utilisateur qui sont rédigés en langage naturel
- la variable cible correspondra à une évaluation (une note) qui souvent sera un nombre (mais pas nécessairement). Il faudra que cette évaluation ne soit pas binaire, mais concerne au moins 3 valeurs différentes. C'est cette variable cible qui sera prédite.
- les jeux de données peuvent être volumineux (de l'ordre du giga ou plus pour certains). Vos traitement doivent pouvoir tourner sur les machines de l'IUT. Ce qui signifie que vous pouvez tout à fait récupérer des données volumineuses hors de l'IUT mais il faudra en choisir un extrait pour que les traitements puissent s'effectuer dans des temps convenables.

Vous utiliserez exclusivement ***scikit-learn*** comme bibliothèque pour réaliser de l'apprentissage automatique et ***spacy*** pour réaliser des traitements linguistiques. Une classe peut être utile pour représenter le texte qui est en langage naturel sous une forme exploitable par les systèmes de classification (tableau de nombre) : ***CountVectorizer*** de la bibliothèque ***scikit-learn***. Cette classe permet de représenter un texte sous forme d'une matrice de nombre.

Travail à réaliser

1) Pré-traitement des données

Les données ne pourront pas en général être utilisées directement. Il faudra mettre en place des pré-traitements (suppression des lignes en doublon éventuelles, traitement des cellules vides, …)

2) Mise en place de la classification avec un méthode vue en cours (deux classes à prédire)

Il y aura deux classes à prédire. Elles correspondront à ***avis favorable*** *et* ***avis défavorable***

Vous utiliserez un des algorithmes de classification qui a été vu durant les ressources.
Il faudra :
- réaliser plusieurs expérimentations avec des variables caractéristiques différentes
- optimiser les hyperparamètres du classifieur pour chaque expérience

Le but est de déterminer à l'aide de plusieurs expérimentations différentes la configuration pour obtenir le meilleur score sur un ensemble de test.

3) Mise en place de la classification avec un méthode vue en cours (multi-classe)

Le but est de réaliser la même chose qu'à la question précédente mais vous prédirez maintenant toutes les classes possibles (différentes notes)

4) Pour aller plus loin (bonus)

Utilisation d'une autre méthode de classification de votre choix (qui tourne sur les PC de l'IUT) pour éventuellement améliorer les résultats de la classification. Il faudra par contre avoir compris cette nouvelle méthode et être capable d'en expliquer le principe.
Cette partie est obligatoire pour le trinôme.

Rendu du travail

Le rendu sera réalisé sur **madoc** avant le **mardi 19 décembre 23h59.**

Vous devez rendre le travail dans une archive au format **zip** qui sera nommée avec le nom de chaque étudiant(e). L'archive comprendra :

- le code (notebooks) de vos expérimentations (qui doivent fonctionner à l'IUT)
- les données utilisées dans un répertoire **data** (si elles sont trop volumineuses pour être stockées sur madoc, un lien vers un cloud)
- un rapport en **pdf** de votre travail.

Voici les différentes parties attendues pour le rapport:
- introduction
- présentation du jeu de données choisi
- explicitation des pré-traitements réalisés sur les données
- présentation des différentes expérimentations (les caractéristiques choisies, comment vous avez optimisé les hyperparamètres, un tableau récapitulatif des résultats globaux obtenus...)
- partie pour aller plus loin si elle a été réalisée
- une conclusion

Soutenance finale

La soutenance de votre travail aura lieu en fin de semaine 51. Chaque binôme présentera le travail réalisé. Chaque membre du binôme devra pouvoir répondre aux questions. Le jury sera vigilant à ce que tout ce qui a été codé ou écrit ait été compris par les étudiants.