Projet 1 : Classification bayésienne

Résumé : *Nous développerons un projet de classification bayésienne en utilisant l'ensemble de données sur les émotions (Kaggle) en plusieurs étapes. Nous allons employer une série de pré-traitement plus complexes et éventuellement étendre l'approche bayésienne pour inclure des ajustements (tuning) ou des probabilité supplémentaires.*

Objectif principal : Développer un classificateur bayésien pour prédire les émotions à partir de données textuelles.

Source des données : Jeu de données sur les émotions.

- **Lien** : Emotion Dataset (kaggle.com)

Aperçu des tâches :

1. **Chargement et exploration des données :**
   - Chargez le jeu de données dans R.
   - Effectuez une analyse exploratoire des données (EDA) pour comprendre la distribution des classes, la longueur des entrées de texte et tout autre modèle.
2. **Prétraitement des données :**
   - Nettoyez les données textuelles en supprimant les caractères spéciaux, les chiffres et les mots vides.
   - Appliquez “tokenization” , “stemming”, or “lemmatization”.
   - Vectorisez le texte à l'aide de la fréquence des termes et de la fréquence inverse du document ou « Term Frequency-Inverse Document Frequency » (TF-IDF).
3. **Entraînement du modèle bayésien :**
   - Utilisez le **package e1071** R pour entraîner un classifieur bayésien naïf.
   - Divisez les données en un ensemble d'apprentissage et un ensemble de test.
   - Entraînez le modèle sur l'ensemble d'apprentissage et réaliser des prédictions sur l'ensemble de test.
4. **Évaluation du modèle :**
   - Calculez l'exactitude, la précision, le rappel et le score F1 (Accuracy, Precision, Recall, and F1-score).
   - Créez une matrice de confusion pour comprendre les performances du modèle à travers différentes émotions.
5. **Amélioration et optimisation :**
   - Expérimentez différentes techniques de prétraitement et d’ingénierie des caractéristiques (feature engineering) pour améliorer le modèle.

- Utilisez la validation croisée k-fold (k-fold cross-validation) pour évaluer la robustesse du modèle.

Rapport et présentation :

- Documentez le processus et les résultats dans un rapport détaillé.
- Préparez une présentation résumant la méthodologie et les résultats.

Livrables:

1. **Code :** script R ou document RMarkdown contenant tout le code pour le prétraitement, l'analyse et la classification.
2. **Rapport :** rapport détaillé (HTML) expliquant la méthodologie, les résultats et les performances du modèle de classification.
3. **Attribution :** 2 étudiants (max.).

Annexe:

Contenu du rapport (RMarkdown ou HTML)

Le rapport doit comprendre :

1. **Introduction:**
   - Décrivez l'ensemble de données et les objectifs du projet.
2. **Méthodologie:**
   - Détaillez les étapes de prétraitement, le raisonnement qui sous-tend l'utilisation de Naive Bayes et toute ingénierie de fonctionnalité (feature engineering) effectuée.
3. **Résultats:**
   - Présentez les résultats de la classification, les mesures d'évaluation du modèle et toute validation croisée effectuée.
4. **Discussion:**
   - Interprétez les résultats, discutez des défis rencontrés et des raisons potentielles de la performance du modèle.
5. **Conclusion et travaux à venir :**
   - Résumez les résultats et suggérez des améliorations possibles ou des orientations futures pour la recherche.