Rapport sur la Sélection de Caractéristiques et Évaluation des Modèles

1. Introduction

Ce projet vise à effectuer une sélection de caractéristiques à l'aide du test **Chi2** sur un jeu de données issu de TF-IDF, enrichi d'étiquettes artificielles générées par le clustering **K-Means**. L'objectif est de réduire la dimensionnalité des données et d'évaluer les performances de trois modèles d'apprentissage supervisé :

- **Random Forest**
- **Support Vector Machine (SVM)**
- **Naive Bayes Multinomial**

2. Méthodologie

2.1. Création des étiquettes artificielles

Les étiquettes ont été générées en divisant les données en deux clusters (à l'aide de l'algorithme **K-Means** avec n_clusters=2). Cette approche permet de simuler des classes pour évaluer les méthodes de sélection et les modèles.

2.2. Sélection des caractéristiques (Chi2)

Le test de Chi2 a été utilisé pour mesurer l'indépendance entre chaque caractéristique et les étiquettes générées. Les 10 caractéristiques les plus pertinentes ont été sélectionnées en fonction des scores Chi2 les plus élevés.

2.3. Division du jeu de données

Le jeu de données a été divisé en deux ensembles :

- **Entraînement** : 70%
- **Test** : 30%

2.4. Évaluation des modèles

Les modèles **Random Forest**, **SVM**, et **Naive Bayes Multinomial** ont été entraînés sur les caractéristiques sélectionnées. Les performances ont été évaluées avec des métriques telles que :

- **Précision**
- **Rappel**
- **F1-Score**

3. Résultats

3.1. Caractéristiques sélectionnées

Les 10 meilleures caractéristiques sélectionnées selon les scores Chi2 sont :

- feature_1
- feature_2
- feature_3
- feature_4
- feature_5
- feature_6
- feature_7
- feature_8
- feature_9
- feature_10

Les scores Chi2 pour ces caractéristiques ont été sauvegardés pour une analyse ultérieure.

3.2. Performances des modèles

a) Random Forest

- Précision globale : **99%**
- F1-Score global : **0.99**
- Détails par classe :
    - Classe 0 : Précision = 0.98, Rappel = 1.00, F1-Score = 0.99
    - Classe 1 : Précision = 1.00, Rappel = 0.98, F1-Score = 0.99

b) SVM

- Précision globale : **93%**
- F1-Score global : **0.93**
- Détails par classe :
    - Classe 0 : Précision = 0.91, Rappel = 0.95, F1-Score = 0.93
    - Classe 1 : Précision = 0.95, Rappel = 0.92, F1-Score = 0.94

c) Naive Bayes Multinomial

- Précision globale : **91%**
- F1-Score global : **0.91**
- Détails par classe :
    - Classe 0 : Précision = 0.84, Rappel = 1.00, F1-Score = 0.91
    - Classe 1 : Précision = 1.00, Rappel = 0.83, F1-Score = 0.90

4. Conclusions

1. **Modèle le plus performant :** Le modèle **Random Forest** a montré les meilleures performances avec une précision et un F1-Score globaux de **99%**.
2. **Importance des caractéristiques :** Les caractéristiques telles que feature_3, feature_5, et feature_8 se distinguent comme étant particulièrement importantes pour différencier les classes.
3. **Applications possibles :** Cette méthode peut être utilisée pour d'autres jeux de données similaires, notamment dans des contextes où les labels sont absents ou générés artificiellement.