Rapport sur la Sélection de Caractéristiques et Évaluation des Modèles

1. Introduction

Ce projet vise à identifier les caractéristiques les plus pertinentes d'un jeu de données en utilisant **l'information mutuelle** (Mutual Information) et à évaluer les performances de trois modèles d'apprentissage supervisé :

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

Le jeu de données, construit à partir de TF-IDF, est enrichi de labels artificiels générés par un clustering K-Means.

2. Méthodologie

Étapes principales :

1. **Création des étiquettes artificielles** :
   - Les étiquettes ont été générées en divisant les données en deux clusters (n_clusters=2) à l'aide de l'algorithme **K-Means**.
2. **Sélection des caractéristiques (Feature Selection)** :
   - Les scores d'information mutuelle ont été calculés pour chaque caractéristique afin de mesurer leur pertinence pour discriminer les classes.
   - Les **10 meilleures caractéristiques** ont été sélectionnées sur la base des scores les plus élevés.
3. **Division du jeu de données** :
   - Le jeu de données a été divisé en deux ensembles : **70% pour l'entraînement** et **30% pour les tests**.
4. **Évaluation des modèles** :
   - Les modèles **Random Forest**, **SVM**, et **Multinomial Naive Bayes** ont été entraînés et évalués sur les données sélectionnées.
   - Les performances ont été mesurées à l'aide de métriques telles que **la précision**, **le rappel** et **le F1-score**.

3. Résultats

3.1 Caractéristiques sélectionnées

Les 10 meilleures caractéristiques sélectionnées selon les scores d'information mutuelle sont :

- at_home, father, home, reputation, yes, no, course, mother, services, other.

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) Multinomial Naive Bayes

- **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 reputation, mother, et course 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.