Rapport sur la Sélection de Caractéristiques et l'Évaluation des Modèles

1. Introduction

Ce projet vise à identifier les caractéristiques les plus pertinentes d'un jeu de données en utilisant la méthode de sélection de caractéristiques progressive (Forward Feature Selection) et à évaluer les performances de trois modèles d'apprentissage supervisé :

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

Le jeu de données est basé sur des représentations TF-IDF et enrichi de labels artificiels générés à l'aide de l'algorithme de 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) :**
   - Des scores d'importance ont été calculés en utilisant la régularisation L1 d'un modèle **Logistic Regression**.
   - Les 10 meilleures caractéristiques ont été retenues 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 caractéristiques sélectionnées.
   - Les performances ont été mesurées à l'aide de métriques telles que :
     - **Précision**
     - **Rappel**
     - **F1-Score**

3. Résultats

3.1 Caractéristiques sélectionnées

Les 10 meilleures caractéristiques retenues selon les scores d'importance sont :

1. **at_home**
2. **father**
3. **home**
4. **reputation**
5. **yes**

6. **no**
7. **course**
8. **mother**
9. **services**
10. **other**

3.2 Performances des modèles

a) Random Forest

- **Précision globale :** 98%
- **F1-Score global :** 0.98

Détails par classe :

- Classe 0 :
  - Précision = 0.97
  - Rappel = 1.00
  - F1-Score = 0.98
- Classe 1 :
  - Précision = 1.00
  - Rappel = 0.97
  - F1-Score = 0.98

b) SVM (Support Vector Machine)

- **Précision globale :** 97%
- **F1-Score global :** 0.97

Détails par classe :

- Classe 0 :
  - Précision = 0.93
  - Rappel = 1.00
  - F1-Score = 0.97
- Classe 1 :
  - Précision = 1.00
  - Rappel = 0.94
  - F1-Score = 0.97

c) Multinomial Naive Bayes

- **Précision globale :** 98%
- **F1-Score global :** 0.98

Détails par classe :

- Classe 0 :
  - Précision = 0.97
  - Rappel = 1.00
  - F1-Score = 0.98
- Classe 1 :
  - Précision = 1.00
  - Rappel = 0.97
  - F1-Score = 0.98

4. Conclusions

1. Modèle le plus performant

Le modèle **Random Forest** s'est montré le plus performant, avec une précision globale de **98%** et un F1-Score global de **0.98**.

2. Importance des caractéristiques

Les caractéristiques comme **reputation**, **mother**, et **course** se distinguent comme étant particulièrement importantes pour différencier les classes.

3. Applications possibles

Cette méthode peut être appliquée à d'autres jeux de données similaires, notamment dans des contextes :

- Où les labels sont absents ou artificiellement générés.
- Qui nécessitent une identification précise des caractéristiques discriminantes pour des applications prédictives.