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

Introduction

Ce rapport présente un projet de prédiction de la performance académique d'étudiants. Les étapes clés incluent la sélection de caractéristiques à l'aide des Fisher's Scores, l'utilisation de SMOTE pour équilibrer les classes, et l'entraînement de modèles de classification. Les résultats obtenus sont analysés en détail pour évaluer la performance des différents modèles.

1 Étapes du Projet

1.1 Chargement et Préparation des Données

Les données TF-IDF sont chargées et nettoyées. La variable cible `G3` (note finale) est transformée en catégories :

- `Low` : notes entre 0 et 10.
- `Medium` : notes entre 11 et 15.
- `High` : notes entre 16 et 20.

Les lignes contenant des valeurs invalides ou manquantes pour `G3` ont été supprimées.

1.2 Sélection des Caractéristiques avec Fisher's Score

Les Fisher's Scores ont été calculés pour chaque caractéristique afin de mesurer leur capacité à discriminer entre les classes (`Low`, `Medium`, `High`). Les 10 meilleures caractéristiques sont :

1. `teacher`
2. `other`
3. `at_home`
4. `services`
5. `health`
6. `reputation`
7. `course`
8. `no`
9. `mother`
10. `home`

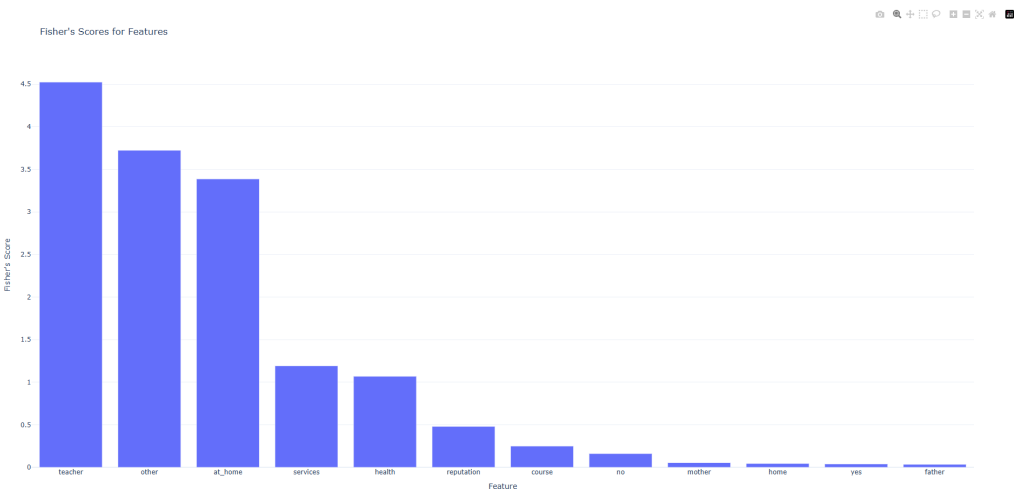

Figure 1: Scores de Fisher pour les caractéristiques.

1.3 Balancement des Classes avec SMOTE

La technique SMOTE a été utilisée pour équilibrer les classes `Low`, `Medium`, et `High`, qui étaient initialement déséquilibrées. Cela a permis de générer des exemples synthétiques pour les classes minoritaires.

1.4 Entraînement et Évaluation des Modèles

Trois modèles ont été entraînés sur les données équilibrées :

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

Chaque modèle a été évalué à l'aide des métriques de précision, rappel, F1-score, et précision globale.

2 Résultats

2.1 Random Forest

- Précision globale : 51%
- Classe `High` : F1-score = 70%, Précision = 73%, Rappel = 67%
- Classe `Low` : F1-score = 43%, Précision = 46%, Rappel = 41%
- Classe `Medium` : F1-score = 38%, Précision = 34%, Rappel = 43%

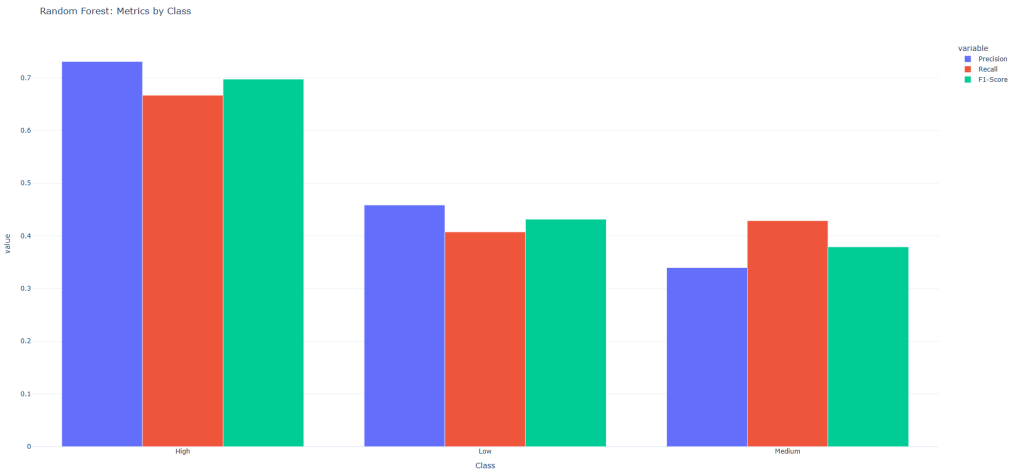

Figure 2: Métriques par classe pour Random Forest.

2.2 SVM (Support Vector Machine)

- Précision globale : 44%
- Classe `High` : F1-score = 47%, Précision = 53%, Rappel = 42%
- Classe `Low` : F1-score = 32%, Précision = 57%, Rappel = 22%
- Classe `Medium` : F1-score = 48%, Précision = 36%, Rappel = 74%

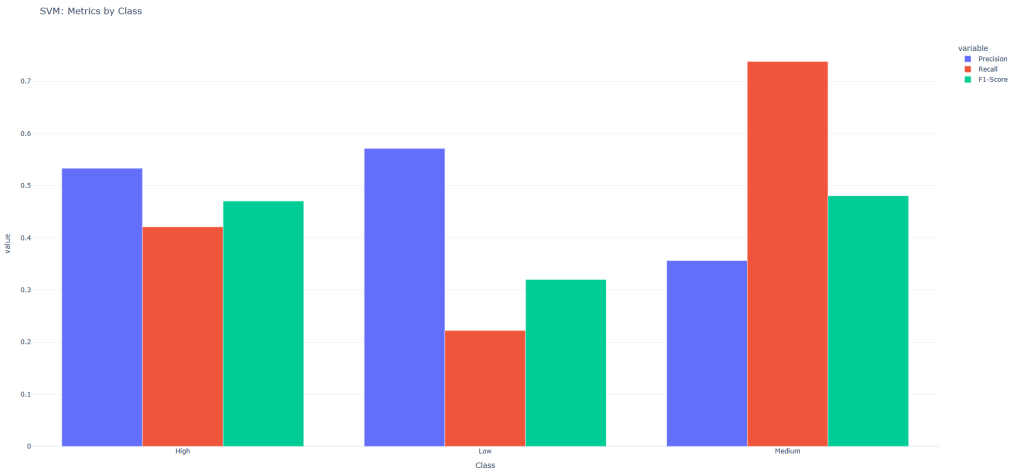

Figure 3: Métriques par classe pour SVM.

2.3 Multinomial Naive Bayes

- Précision globale : 42%
- Classe `High` : F1-score = 47%, Précision = 53%, Rappel = 42%
- Classe `Low` : F1-score = 32%, Précision = 55%, Rappel = 22%
- Classe `Medium` : F1-score = 45%, Précision = 34%, Rappel = 69%

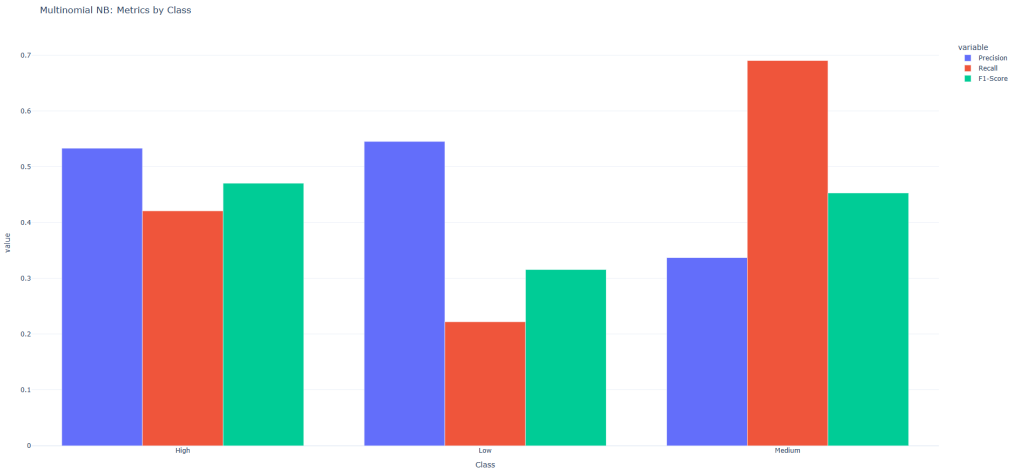

Figure 4: Métriques par classe pour Multinomial Naive Bayes.

2.4 Comparaison des Modèles

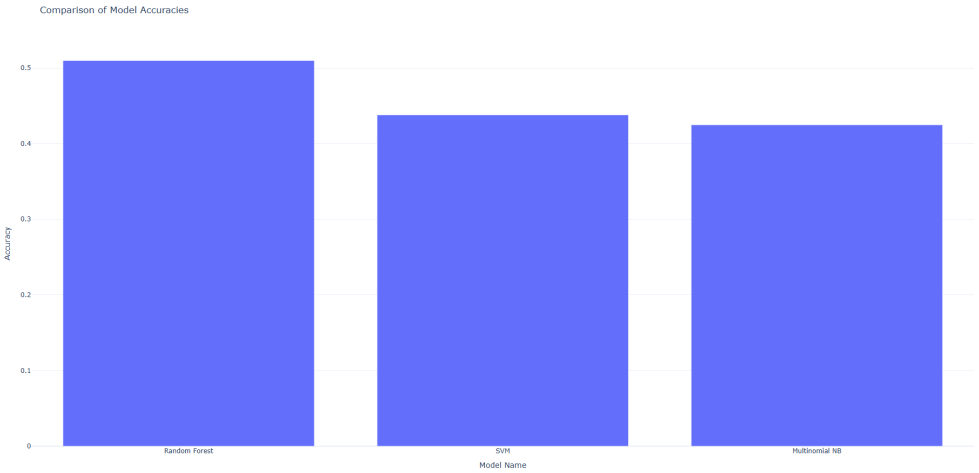

Figure 5: Comparaison des précisions des modèles.

Le modèle Random Forest est le plus performant avec une précision globale de 51%, suivi de SVM (44%) et Multinomial Naive Bayes (42%).

3 Conclusion

Ce projet a montré l'efficacité des Fisher's Scores pour la sélection de caractéristiques et de SMOTE pour équilibrer les classes. Le modèle Random Forest est recommandé comme point de départ pour améliorer les prédictions.