Data Mining Report: Preprocessing and TF-IDF Feature Extraction

1. Introduction

Ce rapport résume le travail de prétraitement et d'extraction de caractéristiques effectué pour le projet de data mining.

L'objectif principal était de nettoyer le jeu de données et d'appliquer TF-IDF pour extraire des caractéristiques numériques significatives à partir des données textuelles pertinentes.

2. Étapes de Prétraitement

- Chargement du jeu de données contenant 395 lignes et 33 colonnes.
- Sélection des colonnes textuelles pertinentes : 'Mjob', 'Fjob', 'reason', 'guardian' et 'activities'.
- Combinaison de ces colonnes en une seule colonne de texte pour l'extraction de caractéristiques.

3. Extraction de Caractéristiques avec TF-IDF

- Application de TF-IDF (Term Frequency-Inverse Document Frequency) à la colonne combinée de texte.
- Génération d'une matrice numérique avec des termes (colonnes) et leurs scores TF-IDF respectifs pour chaque ligne (document).
- La matrice TF-IDF contient des données creuses (beaucoup de cellules à zéro), ce qui est attendu car la plupart des termes n'apparaissent que dans certains documents.

4. Détails des Résultats

- La matrice TF-IDF résultante contient 395 lignes (documents) et 12 colonnes (termes uniques).
- Les termes extraits incluent : 'at_home', 'course', 'teacher', 'health', 'reputation', 'yes', et 'no'.
- Les valeurs TF-IDF reflètent l'importance relative des termes pour chaque document.

5. Pertinence des Termes

- Les termes tels que 'at_home', 'teacher', 'health', et 'reputation' sont pertinents pour analyser la performance scolaire, car ils donnent un contexte sur l'environnement de l'étudiant.
- Les termes binaires ('yes', 'no') proviennent de colonnes comme 'activities' ou 'guardian'. Leur contribution à l'analyse devra être évaluée plus en détail.

6. Étapes Suivantes

- Passer cette matrice TF-IDF à l'étape de sélection des caractéristiques pour identifier les termes les plus influents.
- Évaluer si certains termes, tels que 'yes' et 'no', doivent être filtrés avant d'utiliser la matrice dans l'algorithme de data mining.
- Partager ce rapport et le fichier CSV mis à jour ('tfidf_output_updated.csv') avec l'équipe pour faciliter l'intégration dans le pipeline complet.