Projet Data Mining pour l'éducation (prédiction de performance d'un apprenant).

Auteur : Mamadou BANGOURA /Ibrahima Youssouf KEITA / Odilon LOUA
Date : 16 Janvier 2025

Table des matières

Introduction

Ce projet a pour but de construire un modèle prédictif des performances(note) des apprenants dans le domaine de l'éducation. Ce model nous permet de prédire la note (Grade score) de l'apprenant en fonction de certains paramètres connus (le score représentant le statut économique de l'apprenant, le nombre moyen d'heure d'étude,

le nombre moyen d’heure de sommeil par jour, le pourcentage de présence en classe).

L'approche suit un processus complet de Data Mining, allant de l'intégration et la préparation des données à la modélisation, l'évaluation et la présentation des connaissances.

Objectifs du projet

- **Prédiction du score (note) de l’apprenant :** Utiliser les données fournies pour la prédiction de la note.
- **Visualisations et analyses :** Fournir une vue à travers des graphiques (distributions, corrélations, distribution des erreurs de prédiction, etc.).

Description du Dataset

Le dataset utilisé se présente sous forme de fichier CSV comprenant les colonnes suivantes (exemple) :

1. **Socioeconomic Score** : Un score représentant le statut socio-économique de l'élève.
2. **Study Hours** : Le nombre moyen d'heures d'étude par jour.
3. **Sleep Hours** : Le nombre moyen d'heures de sommeil par jour.
4. **Attendance (%)** : Le pourcentage de présence en classe.
5. **Grades (score)** : Les notes obtenues (semble représenter la réussite).

NB : il y a 1388 lignes dans le dataset.

- ---

Architecture et Processus

Le notebook se décompose en plusieurs étapes, correspondant aux processus de Data Mining :

1. Intégration des données

- **Chargement du dataset :** Lecture du fichier CSV via `pandas` (`pd.read_csv`).

- **Aperçu initial :** Affichage des premières lignes, informations globales et statistiques descriptives.

2. Sélection et Nettoyage des données

Sélection des features : Choix des colonnes pertinentes grâce à l'apport des informations de la matrice de corrélation entre les données (`Socioeconomic Score', 'Study Hours', 'Attendance (%)', 'Grades'])`

- **Nettoyage :**
  - Suppression des doublons.
  - Gestion des valeurs manquantes.

3. Transformation des données

- **Renommage des colonnes :** les nouveaux noms des colonnes , `(['statut socioeconomique', 'nmheure detude par jour','presence en classe (%)', 'notes obtenues/100'].`
- **Standardisation de toutes les valeurs :** standardisation des valeurs numériques pour éviter d'avoir des valeurs numériques prédominantes

4. Modélisation et Exploration (Data Mining)

Deux algorithmes sont utilisés : LinearRegression et RandomForest.

5. Évaluation et Visualisations

- **Mesures de performance :** F1-score pour chaque algorithme
- **Visualisation des erreurs absolues :** Visualisation sous forme de heatmaps.
- **Autres visualisations :**
  - Distribution des erreurs de prédictions
  - Heatmap de corrélation entre variables numériques.

6. Présentation des connaissances (Rapport)

En conclusion, l'utilisation de l'algorithme Random Forest pour prédire les notes des apprenants a donné des résultats très prometteurs, avec un score R^2 de 0.9776. Cela montre que le modèle est très précis dans ses prédictions. Ce résultat ouvre la voie à des applications futures dans la personnalisation de l'apprentissage et l'anticipation des besoins des étudiants.

Comment exécuter le notebook

1. **Installation des dépendances :**
    - Assure-toi d'avoir Python (3.7+) et les librairies requises installées (voir section Dépendances et Installation).
2. **Ouverture dans un environnement interactif :**
    - Ouvre le notebook dans **Jupyter Notebook**.
3. **Exécution séquentielle :**
    - Exécute les cellules une à une pour suivre l'intégralité du processus, de l'intégration à l'évaluation et au rapport final.

Livrables

Ce projet fournit les éléments suivants :

- **Notebook complet** (.ipynb) intégrant toutes les étapes du projet.
- **Le dataset : Le fichier data.csv**
- **Le fichier Readme** : qui détaille le process et cheminement suivi tout au long de l'exécution du projet.

Perspectives d'amélioration

- Augmenter la taille du dataset pour améliorer pour plus renforcer le modèle.
- Tester d'autres algorithmes de prédiction afin d'avoir le meilleur modèle de prédiction (ex. **Gradient Boosting, le Support Vector Machines)**.

Dépendances et Installation

Les librairies principales utilisées sont :

- Python 3.7 ou supérieur
- `numpy`
- `pandas`
- `matplotlib`
- `seaborn`
- `scikit-learn`