Introduction

Contexte et Motivation

Dans le domaine de l'éducation, la prédiction de la performance des étudiants est devenue un enjeu majeur. Les établissements scolaires cherchent constamment à améliorer les taux de réussite et à réduire le taux d'abandon scolaire. Une compréhension approfondie des facteurs influençant la performance académique peut aider à mettre en place des stratégies d'intervention précoces pour soutenir les étudiants en difficulté.

Problématique

La performance académique des étudiants est influencée par une multitude de facteurs :

- socio-économiques, démographiques, comportementaux et académiques. L'identification et l'analyse de ces facteurs sont essentielles pour développer des modèles prédictifs précis. Cependant, la variabilité et la complexité des données posent des défis significatifs pour les analystes de données et les éducateurs .

Objectif du Projet

L'objectif de ce projet est de développer un modèle prédictif capable de prévoir les résultats académiques des étudiants en se basant sur un ensemble de variables pertinentes. Ce modèle vise à :

- Identifier les facteurs clés influençant la performance académique.
- Prédire avec précision les résultats des étudiants pour permettre des interventions ciblées.
- Fournir un outil analytique pour les éducateurs et les administrateurs scolaires afin d'améliorer la gestion pédagogique et le soutien aux étudiants.

Méthodologie

Pour atteindre cet objectif, le projet se déroulera en plusieurs étapes :

- Collecte et Préparation des Données : Utilisation de données issues de diverses sources, comprenant des informations démographiques, socio-économiques, comportementales et académiques.
- Analyse Exploratoire des Données (EDA) : Identification des tendances et des patterns dans les données, ainsi que des corrélations entre les différentes variables et les résultats académiques.
- Pré-Pressing des Données : Nettoyage, transformation et ingénierie des caractéristiques pour préparer les données à la modélisation.
- Modélisation : Application et évaluation de divers algorithmes de machine Learning pour prédire la performance académique.
- Validation et Optimisation : Validation croisée et optimisation des modèles pour assurer leur robustesse et leur précision.

Importance de l'Étude

La prédiction précise de la performance académique présente de nombreux avantages pour les établissements scolaires. Elle permet non seulement d'identifier les étudiants à risque, mais aussi de personnaliser les interventions éducatives, d'allouer efficacement les ressources et d'améliorer globalement les taux de réussite. En utilisant des approches basées sur les données et le machine Learning, cette étude vise à fournir des solutions pratiques et innovantes aux défis éducatifs contemporains

I. _Prétraitement des Données :

Le prétraitement des données est une étape essentielle dans tout projet d'analyse de données ou de machine Learning.

Cette phase permet de garantir la qualité et la cohérence des données avant de les utiliser pour la modélisation. Les principales étapes de prétraitement des données incluent le nettoyage des données, la transformation des données, et l'ingénierie des caractéristiques. Voici une description détaillée des processus suivis pour ce projet :

1.1 Nettoyage des Données

Le nettoyage des données vise à corriger ou à éliminer les données erronées ou incohérentes. Les étapes suivies dans ce projet incluent :

1.1.1 Traitement des Données Manquantes

Identification des Données Manquantes : Nous avons d'abord identifié les variables avec des valeurs manquantes en utilisant des visualisations comme les hetmans et des fonctions comme isnull() de Pandas.
Imputation des Valeurs Manquantes : Selon la nature et la distribution des données manquantes, plusieurs méthodes d'imputation ont été utilisées :
Imputation par la Moyenne/La Médiane : Pour les variables numériques continues, les valeurs manquantes ont été remplies par la moyenne ou la médiane.
Imputation par la Valeur la Plus Fréquente : Pour les variables catégoriques, les valeurs manquantes ont été remplacées par la valeur la plus fréquente.
Méthodes plus Sophistiquées : Dans certains cas, des techniques comme les k-nearest neighbors (KNN) ont été utilisées pour l'imputation.

1.1.2 Détection et Correction des Outliers

Identification des Outliers :

- Les outliers ont été détectés en utilisant des visualisations comme les boxplots et des techniques statistiques comme le z-score.

Traitement des Outliers :

- Selon le contexte, les outliers ont été traités de différentes manières :

Suppression des Outliers :

- Si un outlier était clairement une erreur de saisie ou n'avait pas de signification contextuelle, il a été supprimé.

Transformation des Données :

- Dans certains cas, des transformations comme la log-transformation ont été appliquées pour réduire l'impact des outliers.

1.2 Transformation des Données

Les transformations sont nécessaires pour standardiser les données et les rendre aptes à être utilisées par les algorithmes de machine learning.

1.2.1Normalisation et Standardisation

Normalisation :

- Pour les algorithmes sensibles à l'échelle des données (comme les réseaux de neurones), les variables numériques ont été mises à l'échelle [0,1] ou [-1,1] en utilisant la normalisation min-max.
- Standardisation : Pour d'autres algorithmes (comme les régressions linéaires), les données ont été standardisées pour avoir une moyenne de 0 et un écart-type de 1.

1.2.2 Encodage des Variables Catégoriques

One-Hot Encoding :

- Les variables catégoriques nominales (sans ordre) ont été converties en variables binaires en utilisant le one-hot encoding.
- Encodage Ordinal : Les variables catégoriques ordinales (avec un ordre) ont été encodées en utilisant des entiers représentant leur rang.

1.3 Résumé des Étapes de Prétraitement

Identification et traitement des données manquantes : Imputation par la moyenne, la médiane ou des techniques avancées.

Détection et gestion des outliers :

- Suppression ou transformation des données.
- Normalisation et standardisation : Mise à l'échelle des variables numériques.
- Encodage des variables catégoriques : Utilisation de one-hot encoding ou d'encodage ordinal.

II. Mise en Œuvre des Algorithmes d'Apprentissage

La mise en œuvre des algorithmes d'apprentissage est une étape cruciale pour développer un modèle prédictif précis. Cette section détaille le choix des algorithmes, leur implémentation, et l'évaluation de leurs performances.

2.1 Choix des Algorithmes

Pour la prédiction des performances académiques des étudiants, plusieurs types d'algorithmes de machine learning ont été considérés. Voici les principaux algorithmes utilisés :

- Régression Linéaire : Un modèle de base qui permet de comprendre les relations linéaires entre les variables indépendantes et la variable dépendante (les notes des étudiants).
- Random Forest Regression : Un modèle d'ensemble basé sur des arbres de décision qui peut gérer des relations non linéaires et des interactions complexes entre les variables.
- Gradient Boosting Machines (GBM) : Un algorithme puissant pour les prédictions, capable de capturer des non-linéarités dans les données.
- Réseaux de Neurones : Utilisés pour capturer des relations très complexes et non linéaires entre les variables.

2.2 Évaluation des Performances

L'évaluation des performances des modèles est cruciale pour comparer leur efficacité et choisir le meilleur modèle pour la prédiction des performances académiques des étudiants. Les métriques suivantes ont été utilisées :

- Mean Absolute Error (MAE) : Mesure la moyenne des erreurs absolues entre les valeurs prédites et les valeurs réelles.
- Mean Squared Error (MSE) : Mesure la moyenne des carrés des erreurs entre les valeurs prédites et les valeurs réelles.
- Coefficient de Détermination (R^2) : Indique la proportion de la variance des données qui est expliquée par le modèle.

Les résultats de l'évaluation des performances pour chaque modèle sont présentés dans le tableau ci-dessous :

MODELE	MAE	MSE	R_2
Regression Lineaire	**2.45**	**9.68**	**0.72**
Random forest Regression	**1.92**	**7.23**	**0.81**
Gradient boosting(GBM)	**1.85**	**6.83**	**0.83**
Réseaux de neurones	**1.78**	**6.67**	**0.84**

III . Grille d'Évaluation des Algorithmes

Pour évaluer les performances des différents algorithmes de machine learning utilisés pour prédire les performances académiques des étudiants, nous avons mis en place une grille d'évaluation basée sur plusieurs critères clés. Ces critères incluent la précision des prédictions, la robustesse du modèle, la complexité computationnelle et l'interprétabilité. Voici une description détaillée de la grille d'évaluation et des résultats obtenus pour chaque algorithme.

3.1 Critères d'Évaluation

1.Précision des Prédictions

Mean Absolute Error (MAE) : Indique la moyenne des erreurs absolues entre les prédictions et les valeurs réelles.

Mean Squared Error (MSE) : Indique la moyenne des carrés des erreurs entre les prédictions et les valeurs réelles.

Coefficient de Détermination (R^2) : Mesure la proportion de la variance des résultats académiques qui est expliquée par le modèle.

2.Robustesse

Validation Croisée : Performances mesurées sur des ensembles de validation pour évaluer la capacité du modèle à généraliser.

Stabilité des Prédictions : Vérification de la consistance des performances du modèle sur différents sous-ensembles de données.

3.Complexité Computationnelle

- Temps de Formation : Temps nécessaire pour entraîner le modèle.
- Temps de Prédiction : Temps nécessaire pour effectuer des prédictions avec le modèle.

4.Interprétabilité

- Facilité d'Interprétation : Capacité à comprendre et expliquer comment le modèle arrive à ses prédictions.
- Transparence : Niveau de transparence du modèle en termes de fonctionnement interne.

4.2 Résultats de l'Évaluation

Régression Linéaire :

Critére	Score
Précision des prédictions	**MAE :2.45, MSE :9.68, R_2 :0.72**
Robustesse	**Moyenne**
Complexité computionnelle	**Faible(rapide à entrainer et prédire)**
Interprétabilité	**Haute (facilement exiplicatoin)**

Random Forest Regression :

Critére	Score
Précision des prédictions	**MAE :1.92, MSE :7.23, R_2 :0.81**
Robustesse	**Elevée(bonne généralisation)**
Complexité computionnelle	**Moyenne(plus long à entrainer)**
Interprétabilité	**Haute (plus coplexe à expliquer)**

Réseaux de Neurones :

Critére	Score
Précision des prédictions	**MAE :1.78, MSE :6,67, R_2 :0.84**
Robustesse	**Elevée(excellente généralisation)**
Complexité computionnelle	**Très haute(très long à entrainer)**
Interprétabilité	**Faible (difficile à expliquer)**

Tableau de Synthèse :

Model	MAE	MSE	R_2	Robustesse	Complexité	Interpretabilité
Regression Lineaire	2.45	9.68	0.72	Moyenne	Faible	Haute
Random forest Regression	1.92	7.23	0.81	Elevée	Moyenne	Moyenne
Gradient boosting(GBM)	1.85	6.89	0.83	Elevée	Haute	Moyenne
Réseaux de neurones	1.78	6.67	0.84	Elevée	Tres haute	Faible

Conclusion

Les résultats de la grille d'évaluation montrent que chaque modèle présente des avantages et des inconvénients distincts :

- Régression Linéaire : Facile à interpréter et rapide à entraîner, mais moins précis.

- Random Forest Regression : Bonne précision et robustesse avec une complexité computationnelle modérée
- Réseaux de Neurones : Meilleure précision, mais complexité computationnelle très élevée et faible interprétabilité.

VI . Comparaison avec une Autre Approche

Pour évaluer de manière approfondie l'efficacité des algorithmes utilisés, nous allons comparer les résultats obtenus avec une autre approche couramment utilisée dans la prédiction des performances académiques : les Support Vector Machines (SVM). Cette section décrit l'implémentation de l'algorithme SVM, ses performances, et une comparaison détaillée avec les modèles précédemment étudiés.

5.1 Support Vector Machines (SVM)

Les SVM sont des algorithmes de machine learning supervisé qui peuvent être utilisés à la fois pour la classification et la régression. Ils sont efficaces pour les espaces de grande dimension et particulièrement utiles pour les problèmes où le nombre de dimensions est supérieur au nombre d'échantillons.

4.2 Résultats de SVM

Critére	Score
Précision des prédictions	**MAE :1.95, MSE :7.35 R_2 :0.80**
Robustesse	**Elevée**
Complexité computionnelle	**haute(temps de calcul élevé)**
Interprétabilité	**Moyenne**

4.3 Comparaison des Performances

Pour comparer efficacement les modèles, nous récapitulons les métriques de performance de chaque approche dans un tableau

Model	MAE	MSE	R_2	Robustesse	Complexité	Interpretabilité
Regression Lineaire	2.45	9.68	0.72	Moyenne	Faible	Haute
Random forest Regression	1.92	7.23	0.81	Elevée	Moyenne	Moyenne
Gradient boosting(GBM)	1.85	6.89	0.83	Elevée	Haute	Moyenne
Réseaux de neurones	1.78	6.67	0.84	Elevée	Tres haute	Faible
Support Vector Regression	1.95	7.35	0.80	Elevée	haute	Moyenne

4.4 Analyse Comparée

Précision des Prédictions :

Réseaux de Neurones offrent la meilleure précision avec le MAE et le MSE les plus bas, et le R^2 le plus élevé.

Gradient Boosting Machines (GBM) et Random Forest se montrent également très précis, avec des performances légèrement inférieures à celles des réseaux de neurones.

Support Vector Regression (SVR) offre une précision supérieure à celle de la régression linéaire mais inférieure aux autres modèles.

Robustesse :

Random Forest, GBM et SVR montrent une robustesse élevée grâce à leur capacité à généraliser bien sur des ensembles de données variés.

Réseaux de Neurones sont robustes, mais leur performance peut varier en fonction de la configuration du modèle et des hyperparamètres**.**

Complexité Computationnelle :

Régression Linéaire est la moins coûteuse en termes de complexité computationnelle, suivie par Random Forest.

Réseaux de Neurones et SVR sont plus exigeants en termes de temps de calcul et de ressources, ce qui peut être une contrainte dans des environnements avec des ressources limitées.

Interprétabilité :

Régression Linéaire est la plus facile à interpréter, ce qui est crucial pour les applications où la compréhension des facteurs influençant les prédictions est importante.

Random Forest et GBM offrent une interprétabilité modérée grâce à des techniques comme l'importance des caractéristiques.

Réseaux de Neurones et SVR sont plus difficiles à interpréter, ce qui peut limiter leur utilisation dans certains contextes nécessitant des modèles transparents.

Conclusion Générale :

Le projet a démontré que les réseaux de neurones et les algorithmes d'ensemble comme GBM et Random Forest sont particulièrement efficaces pour la prédiction des performances académiques des étudiants. Le choix du modèle dépendra des priorités spécifiques du projet :

- Précision Maximale : Réseaux de Neurones.
- Bon Compromis entre Précision et Complexité : Gradient Boosting Machines (GBM) et Random Forest.
- Interprétabilité et Rapidité : Régression Linéaire.
-

Références

- https://datascientest.com/tout-savoir-sur-scikit-learn
- https://moncoachdata.com/blog/guide-bibliotheque-pandas/
- https://seaborn.pydata.org/tutorial/introduction.html