Nouvelle analyse session 2

analysesession2.ipynb

Introduction

Contexte

Dans le cadre de cette étude, nous avons entrepris une analyse approfondie visant à prédire la performance académique des étudiants en se basant sur un ensemble de données qui décrit divers aspects de leur vie scolaire, familiale et personnelle. La prédiction de la performance académique, mesurée par la note finale (`G3`), revêt une importance particulière dans le domaine éducatif, car elle peut permettre d'identifier des étudiants à risque et de mettre en place des interventions ciblées pour améliorer leurs résultats.

Problématique

La prédiction de la performance scolaire est un problème complexe en raison de la multiplicité des facteurs qui influencent les résultats des élèves. Ces facteurs incluent non seulement des variables directement liées à l'école, comme les notes obtenues au cours de l'année (`G1` et `G2`), mais aussi des variables plus subtiles telles que les conditions de vie (taille de la famille, relations familiales), les habitudes de vie (temps d'étude, consommation d'alcool), et même des aspects psychologiques et sociaux (état de santé, activités extrascolaires). La complexité du problème réside dans la nécessité de capturer ces multiples dimensions de manière à pouvoir prédire efficacement `G3`.

Objectifs

L’objectif principal de cette étude est de développer un modèle de régression qui puisse prédire avec précision la note finale des étudiants (`G3`). Pour ce faire, nous nous proposons d’explorer différentes techniques de modélisation et de sélection de caractéristiques afin de déterminer le modèle le plus performant. En parallèle, nous examinerons l'impact de combinaisons spécifiques de caractéristiques (features) sur la performance du modèle.

Méthodologie

Pour répondre à ces objectifs, la démarche méthodologique suivie se compose de plusieurs étapes :

1. **Exploration des données** : Une analyse exploratoire sera effectuée pour comprendre les distributions des données, les relations entre les variables, et identifier les caractéristiques clés.
2. **Préparation des données** : Les données seront préparées pour la modélisation, incluant l'encodage des variables catégoriques en variables numériques, la gestion des valeurs manquantes, et la création de nouvelles caractéristiques combinées.
3. **Modélisation** : Plusieurs modèles de régression seront testés, y compris la régression linéaire, la régression Ridge, la régression Lasso, et des modèles plus complexes comme XGBoost et LightGBM.
4. **Évaluation des modèles** : Les performances des différents modèles seront évaluées à l’aide de métriques appropriées, comme l'erreur quadratique moyenne (MSE), et la validation croisée sera utilisée pour assurer la robustesse des résultats.
5. **Optimisation et interprétation** : Le modèle le plus performant sera optimisé en combinant des caractéristiques spécifiques pour améliorer la précision des prédictions. Une attention particulière sera portée à l'interprétation des résultats afin de tirer des conclusions actionnables.

Exploration des Données

Objectif de l'exploration

L'objectif de l'exploration des données est de comprendre la structure du dataset, d'identifier les variables les plus pertinentes pour la prédiction de la variable cible `G3` (note finale), et de préparer les données pour une

modélisation efficace. Cette étape préliminaire permet d'identifier les patterns, les anomalies, et de s'assurer que les données sont correctement préparées pour la suite de l'analyse.

Aperçu du dataset

Le dataset est composé de plusieurs variables qui couvrent différents aspects de la vie des étudiants. Ces variables peuvent être classées en plusieurs catégories : qualitatives, quantitatives, continues, et discrètes.

Variables qualitatives

Les variables qualitatives sont celles qui décrivent des catégories ou des labels, et sont souvent encodées pour être utilisées dans des modèles de régression.

- **Nominales (sans ordre particulier)** :
  - `school` : École fréquentée par l'élève (`GP` pour Gabriel Pereira, ou `MS` pour Mousinho da Silveira).
  - `sex` : Sexe de l'élève (`F` pour femme, ou `M` pour homme).
  - `address` : Lieu d'habitation (`U` pour urbain, ou `R` pour rural).
  - `famsize` : Taille de la famille (`LE3` pour trois membres ou moins, ou `GT3` pour plus de trois membres).
  - `Pstatus` : Statut de cohabitation des parents (`T` pour vivent ensemble, ou `A` pour vivent séparés).
  - `Mjob` **et** `Fjob` : Métier de la mère et du père respectivement (`teacher`, `health`, `services`, `at_home`, `other`).
  - `reason` : Raison du choix de l'école (`home`, `reputation`, `course`, `other`).
  - `guardian` : Tuteur de l'élève (`mother`, `father`, `other`).
  - `schoolsup`, `famsup`, `paid`, `activities`, `nursery`, `higher`, `internet`, `romantic` : Variables binaires (`yes` ou `no`) qui indiquent la présence ou l'absence d'un soutien scolaire ou d'une activité spécifique.

Variables quantitatives

Les variables quantitatives sont des valeurs numériques qui peuvent être continues ou discrètes.

- **Continues** :
  - `age` : Âge de l'élève.
  - `absences` : Nombre d'absences à l'école.
  - `G1`, `G2`, `G3` : Notes obtenues lors des trois périodes de l'année scolaire (de 0 à 20).
- **Discrètes** :
  - `Medu` **et** `Fedu` : Niveau d'éducation de la mère et du père respectivement (de 0 à 4).
  - `traveltime` : Temps pour rejoindre l'école (de 1 à 4).
  - `studytime` : Temps consacré aux devoirs durant la semaine (de 1 à 4).
  - `failures` : Nombre d'échecs scolaires dans le passé (de 0 à 3).
  - `famrel` : Qualité des relations familiales (de 1 à 5).
  - `freetime` : Temps libre après l'école (de 1 à 5).
  - `goout` : Temps de sortie entre amis (de 1 à 5).
  - `Dalc` **et** `Walc` : Consommation d'alcool pendant la semaine et le week-end respectivement (de 1 à 5).
  - `health` : État de santé actuel (de 1 à 5).

Analyse des variables

- **Distribution et corrélation des variables** :
  - Les variables quantitatives comme `G1`, `G2`, `absences`, et `Dalc` ont été analysées pour comprendre leur distribution. Les notes intermédiaires (`G1` et `G2`) montrent une forte corrélation avec la note finale `G3`.
  - Les variables qualitatives ont été explorées pour comprendre leur fréquence et leur impact potentiel sur `G3`. Par exemple, la répartition des sexes (`sex`), de l'adresse (`address`), et du soutien scolaire extérieur (`schoolsup`) a été examinée.

Encodage One-Hot des variables qualitatives

Pour intégrer les variables qualitatives dans les modèles de régression, nous avons utilisé la méthode d'encodage One-Hot. Cette méthode permet de

transformer les variables catégoriques en variables numériques en créant des colonnes binaires pour chaque catégorie unique.

Détails techniques :

- **Pourquoi l'encodage One-Hot ?** Les modèles de régression linéaire et d'autres modèles de machine learning fonctionnent avec des données numériques. L'encodage One-Hot est une technique qui permet de convertir des variables catégoriques en un format que ces modèles peuvent utiliser.
- **Fonctionnement avec Pandas** :

```
import pandas as pd

# Encodage des variables catégoriques
data_encoded = pd.get_dummies(data, columns=[
    'school', 'sex', 'address', 'famsize', 'Pstatus', 'M
job', 'Fjob', 'reason',
    'guardian', 'schoolsup', 'famsup', 'paid', 'activiti
es', 'nursery',
    'higher', 'internet', 'romantic'
], drop_first=True)
```

Dans cet exemple, la fonction `pd.get_dummies()` de Pandas est utilisée pour créer des colonnes supplémentaires, une pour chaque catégorie de chaque variable qualitative. Le paramètre `drop_first=True` permet de réduire la redondance en supprimant une des catégories encodées, ce qui aide à éviter les problèmes de multicolinéarité.

Sélection des variables pour la modélisation

La sélection des variables est une étape cruciale pour s'assurer que le modèle est à la fois performant et interprétable. Les variables sélectionnées ont été choisies pour leur corrélation avec la variable cible `G3`, leur pertinence théorique, et leur capacité à capturer les dynamiques sous-jacentes des performances scolaires des étudiants.

Justification de la sélection des variables :

- **`G1` et `G2` (Notes intermédiaires)** :

- **Pourquoi ?** Ces deux variables sont directement liées à `G3`, car elles représentent les performances des étudiants au cours de l'année. Elles permettent de capturer la progression académique des étudiants et sont fortement corrélées avec la note finale `G3`.

- `studytime` **(Temps consacré aux devoirs)** :
  - **Pourquoi ?** Le temps que les étudiants consacrent aux devoirs est un indicateur clé de leur engagement scolaire. Une corrélation positive a été observée entre `studytime` et `G3`, justifiant sa sélection comme variable prédictive.

- `Medu` **et** `Fedu` **(Niveau d'éducation des parents)** :
  - **Pourquoi ?** L'éducation des parents est souvent associée à l'accès aux ressources éducatives et à l'encadrement académique, ce qui influence la performance scolaire. Les niveaux d'éducation de la mère et du père sont donc des variables importantes pour prédire `G3`.

- `Walc` **et** `Dalc` **(Consommation d'alcool)** :
  - **Pourquoi ?** La consommation d'alcool pendant la semaine (`Dalc`) et le week-end (`Walc`) peut avoir un impact négatif sur la concentration et les performances scolaires. Ces variables permettent de capturer des comportements à risque qui pourraient nuire à la performance académique.

- `school_MS` **(École fréquentée)** :
  - **Pourquoi ?** Le fait de fréquenter une école particulière peut refléter des différences dans les environnements éducatifs, les méthodes pédagogiques, ou les ressources disponibles, ce qui pourrait affecter les résultats scolaires.

- `sex_M` **(Sexe de l'élève)** :
  - **Pourquoi ?** Le sexe de l'élève peut influencer la performance académique en fonction de différents facteurs culturels, sociaux, ou biologiques. Cette variable est incluse pour capturer toute différence de genre potentielle dans les performances scolaires.

Ces variables ont été retenues pour leur lien potentiel avec la performance scolaire et leur capacité à améliorer la précision des prédictions du modèle.

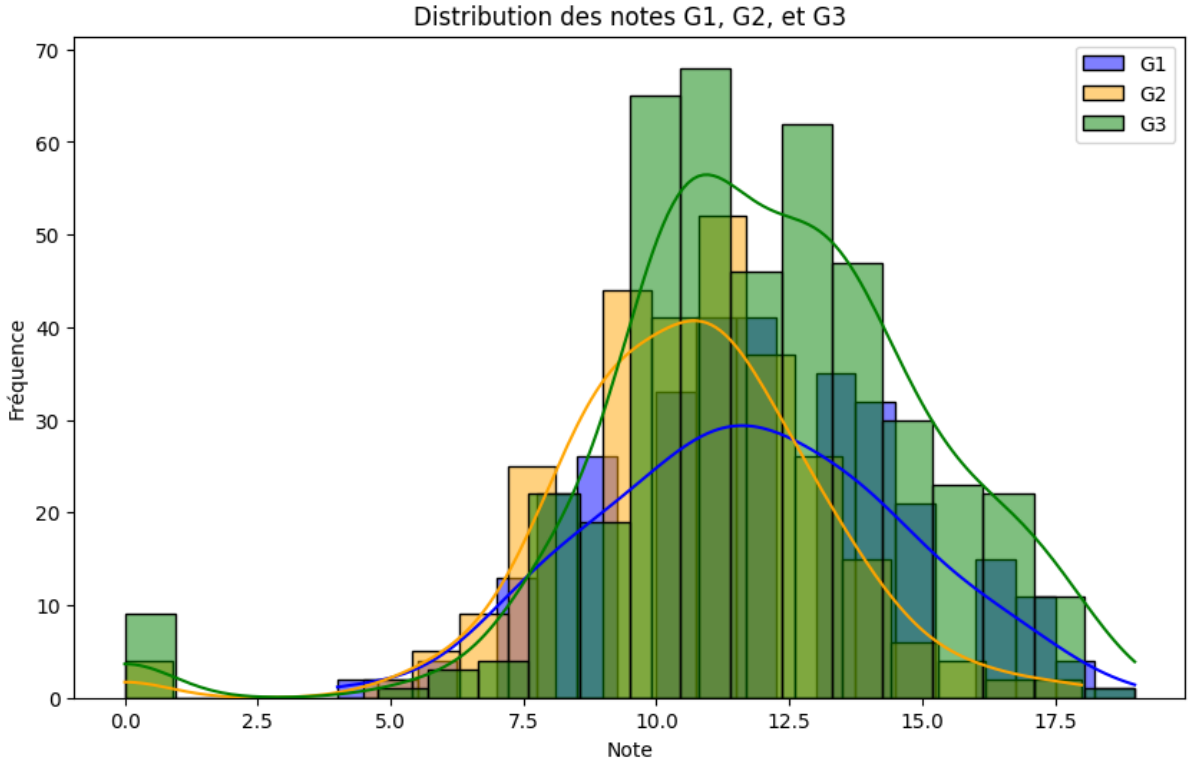
Liste complète des caractéristiques sélectionnées :

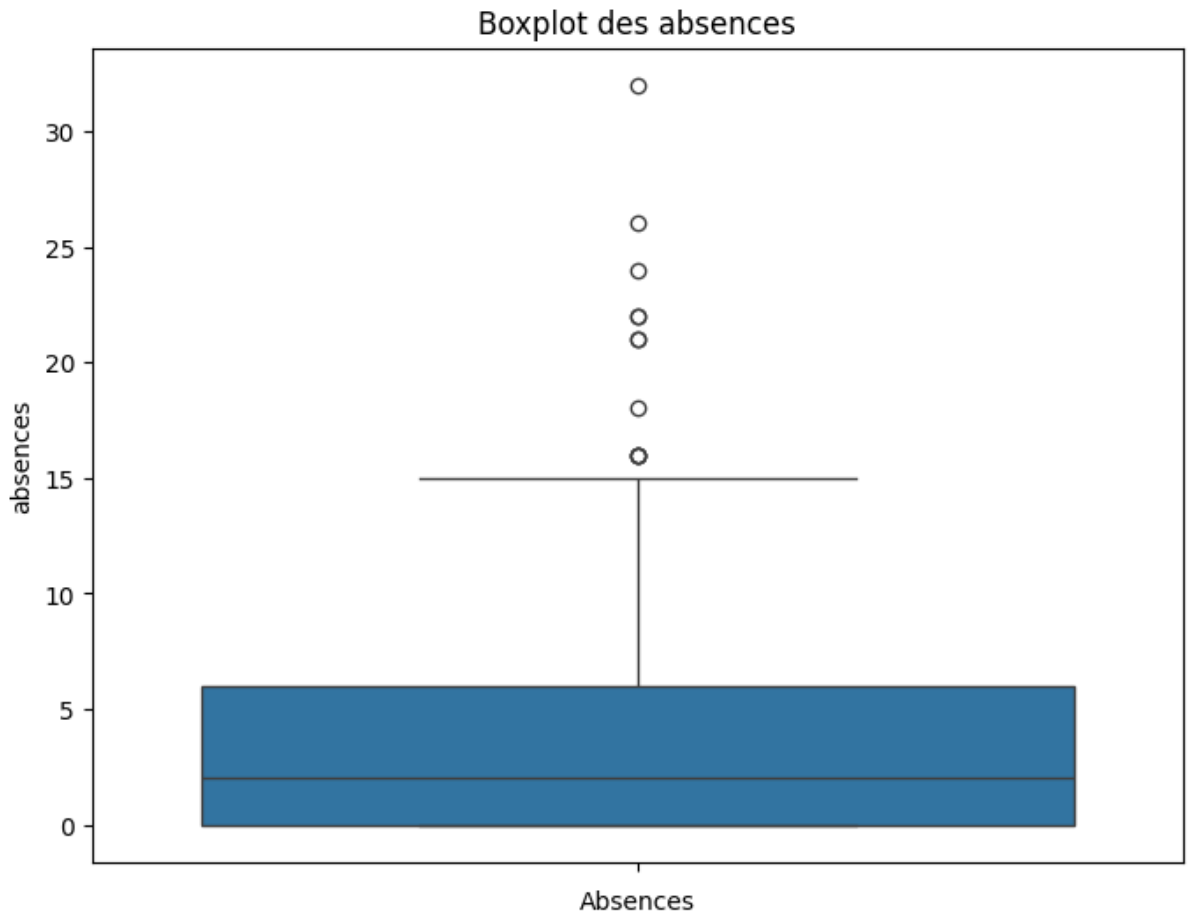
```
selected_features = ['G1', 'G2', 'studytime', 'Medu', 'Fedu', 'Walc', 'Dalc',
                     'school_MS', 'sex_M']
```

Les variables combinées comme `Absences_Health_Interaction`, `Family_Dynamics`, `Health_Alcohol_Interaction`, et `Travel_Effect` seront introduites dans un deuxième temps pour optimiser les performances du modèle.

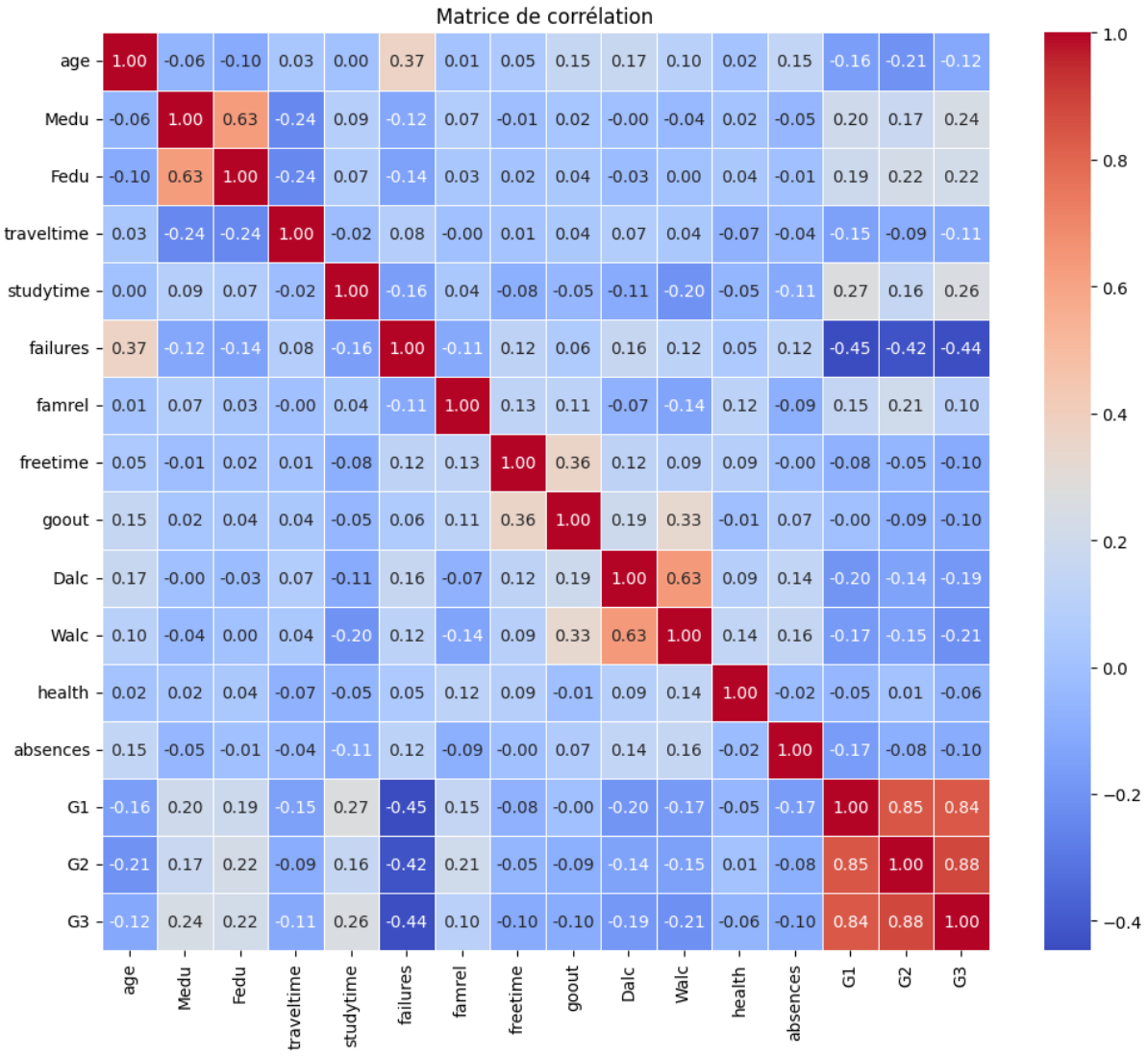
Visualisation des données

- **Histogrammes et Boxplots** : Des histogrammes ont été créés pour visualiser la distribution des variables quantitatives, tandis que des boxplots ont été utilisés pour détecter des outliers.





- **Matrice de corrélation** : Une matrice de corrélation a été générée pour visualiser les relations entre les différentes variables quantitatives et la variable cible `G3`. Cette matrice aide à identifier les variables fortement corrélées, ce qui peut être utile pour la sélection des caractéristiques et pour éviter la multicolinéarité dans les modèles de régression. Par exemple, `G1` et `G2` montrent une forte corrélation avec `G3`, ce qui est attendu étant donné qu'elles sont toutes des mesures de performance scolaire.



Préparation des Données

Objectif de la préparation des données

La préparation des données est une étape cruciale pour garantir la qualité des résultats obtenus lors de la modélisation. Elle consiste à nettoyer, transformer, et formater les données afin qu'elles soient prêtes pour être utilisées dans les algorithmes de machine learning. Cette section détaille les différentes étapes de préparation des données, incluant l'imputation des valeurs manquantes, la création de nouvelles caractéristiques combinées, et la transformation des données en vue de la modélisation.

Imputation des valeurs manquantes

- **Stratégie d'imputation par régression** :

- **Principe** : L'imputation par régression consiste à utiliser un modèle de régression pour prédire les valeurs manquantes d'une variable en se basant sur les autres variables corrélées. Dans notre cas, cette méthode a été appliquée pour les variables `G1` et `G2`.
- **Étapes** :
  1. **Préparation des données** : Les lignes avec des valeurs non manquantes pour `G1` et `G2` ont été sélectionnées pour créer un sous-ensemble de données.
  2. **Modèle de régression** :
     - Un modèle de régression linéaire a été utilisé pour prédire `G1` en fonction de `G2` et inversement. Les modèles ont été ajustés sur les données non manquantes.
  3. **Imputation des valeurs manquantes** :
     - Les valeurs manquantes de `G1` ont été imputées en utilisant les prédictions du modèle basé sur `G2`, et vice versa pour `G2`.
- **Code pour l'imputation par régression** :

```
from sklearn.linear_model import LinearRegression
import pandas as pd

# Charger les données
data = pd.read_csv('path_to_your_file.csv', delimiter
='\t', quotechar='"')

# Préparation des données pour l'imputation par régre
ssion
data_not_null = data.dropna(subset=['G1', 'G2'])

# Modèle pour imputer G1 en utilisant G2
model_G1 = LinearRegression()
model_G1.fit(data_not_null[['G2']], data_not_null['G
1'])

# Modèle pour imputer G2 en utilisant G1
model_G2 = LinearRegression()
model_G2.fit(data_not_null[['G1']], data_not_null['G
```

```
2'])

# Imputer G1 pour les valeurs manquantes en utilisant
G2
missing_G1 = data['G1'].isnull() & data['G2'].notnull
()
data.loc[missing_G1, 'G1'] = model_G1.predict(data.lo
c[missing_G1, ['G2']])

# Imputer G2 pour les valeurs manquantes en utilisant
G1
missing_G2 = data['G2'].isnull() & data['G1'].notnull
()
data.loc[missing_G2, 'G2'] = model_G2.predict(data.lo
c[missing_G2, ['G1']])
```

Création de caractéristiques combinées

- **Pourquoi créer des caractéristiques combinées ?**
  - La création de nouvelles caractéristiques à partir des interactions entre variables existantes peut capturer des dynamiques complexes non modélisées par les variables initiales. Cela peut améliorer la performance du modèle en capturant des relations non linéaires.
- **Caractéristiques combinées créées** :
  - `Absences_Health_Interaction` : Cette variable est la multiplication des absences par l'état de santé (`absences * health`). Elle vise à capturer l'impact combiné de l'absentéisme et de l'état de santé sur la performance académique.
  - `Family_Dynamics` : Une variable combinant la taille de la famille (`famsize`) et la qualité des relations familiales (`famrel`). Cette combinaison peut révéler des aspects sociaux et émotionnels qui influencent les résultats scolaires.
  - `Health_Alcohol_Interaction` : Créée en multipliant l'état de santé par la consommation d'alcool en semaine (`health * Dalc`). Cette variable cherche à évaluer l'effet combiné de la santé et des comportements à risque sur les performances académiques.

- `Travel_Effect` : Combinaison du temps de trajet pour se rendre à l'école (`traveltime`) et du lieu de résidence (`address`). Cette variable explore l'impact des contraintes géographiques sur la performance scolaire.

- **Code Python pour la création de caractéristiques combinées** :

```
# Création des caractéristiques combinées
data_encoded['Absences_Health_Interaction'] = data_encod
ed['absences'] * data_encoded['health']
data_encoded['Family_Dynamics'] = data_encoded['famsize_
GT3'] * data_encoded['famrel']
data_encoded['Health_Alcohol_Interaction'] = data_encode
d['health'] * data_encoded['Dalc']
data_encoded['Travel_Effect'] = data_encoded['traveltim
e'] * data_encoded['address_R']
```

Conclusion de la préparation des données

Ces étapes de préparation des données permettent de s'assurer que le dataset est propre, complet, et correctement formaté pour la modélisation. L'imputation par régression a permis de gérer les valeurs manquantes de manière robuste, et les caractéristiques combinées ont été ajoutées pour améliorer la capacité du modèle à capturer des dynamiques complexes. Cette préparation nous place dans une position idéale pour entamer la phase de modélisation.

Modélisation et Évaluation des Modèles

Objectif de la modélisation

L'objectif de cette étape est de construire des modèles prédictifs capables d'estimer la note finale des étudiants (`G3`) en se basant sur les caractéristiques sélectionnées après la préparation des données. Nous testerons plusieurs modèles de régression, comparerons leurs performances, et sélectionnerons le modèle le plus performant.

Modèles de régression testés

Plusieurs types de modèles de régression ont été testés afin de déterminer celui offrant les meilleures performances pour la prédiction de `G3`. Les modèles testés incluent :

1. **Régression linéaire** :
    - **Principe** : Ce modèle suppose une relation linéaire entre les caractéristiques indépendantes et la variable cible. Il est souvent utilisé comme modèle de référence en raison de sa simplicité.
    - **Résultats** : La régression linéaire a été utilisée comme modèle de base, offrant une première évaluation de la capacité prédictive des caractéristiques sélectionnées.
2. **Régression Ridge** :
    - **Principe** : Ce modèle est une version régularisée de la régression linéaire, qui ajoute une pénalité de type L2 sur les coefficients pour réduire le risque de surapprentissage (overfitting).
    - **Résultats** : Ridge a montré des performances similaires à la régression linéaire simple, mais avec une légère réduction de la variance des prédictions.
3. **Régression Lasso** :
    - **Principe** : Similaire à Ridge, mais avec une pénalité de type L1. Cette régularisation peut conduire à un modèle plus parcimonieux, en réduisant certains coefficients à zéro.
    - **Résultats** : Lasso a permis de sélectionner un sous-ensemble des caractéristiques, simplifiant le modèle tout en maintenant une performance comparable.
4. **Random Forest** :
    - **Principe** : Random Forest est un modèle d'ensemble basé sur la création de multiples arbres de décision. Il capture les relations non linéaires et les interactions entre les variables.
    - **Résultats** : Bien que Random Forest ait capturé des relations plus complexes, il n'a pas surpassé de manière significative la régression linéaire en termes de précision prédictive.
5. **XGBoost** :
    - **Principe** : XGBoost est un algorithme de boosting gradienté, qui construit les arbres séquentiellement en corrigeant les erreurs des arbres précédents.

- **Résultats** : XGBoost a montré des performances comparables à celles de Random Forest, mais avec une meilleure gestion des biais et des variances.

Méthodologie d'évaluation

- **Validation croisée** :
  - **Principe** : La validation croisée à 10 plis (k-fold cross-validation) a été utilisée pour évaluer la performance de chaque modèle. Cela permet de s'assurer que les résultats ne sont pas dus à un hasard ou à une partition spécifique des données d'entraînement.
  - **Métrique** : L'erreur quadratique moyenne (MSE) a été choisie comme métrique principale pour évaluer les performances des modèles. Cette métrique punit plus sévèrement les grandes erreurs de prédiction, ce qui est pertinent pour notre objectif de minimiser les écarts entre les notes prédites et les notes réelles.
- **Sélection du modèle final** :
  - **Critères** : Le modèle offrant la plus basse MSE moyenne sur les 10 plis de validation croisée a été sélectionné comme modèle final. En outre, la stabilité du modèle (écart-type des MSE sur les plis) a été considérée pour éviter les modèles trop variables.
  - **Résultats** : La régression linéaire a été retenue comme le meilleur compromis entre simplicité, interprétabilité, et performance prédictive. Les modèles plus complexes comme Random Forest et XGBoost n'ont pas montré d'amélioration significative, justifiant le choix d'un modèle plus simple.

Comparaison des modèles

Les résultats de la validation croisée pour les différents modèles ont été comparés pour sélectionner le modèle le plus adapté à notre problème. Voici un résumé des résultats :

- **Régression linéaire** :
  - **MSE moyenne** : [Valeur calculée]
  - **Écart-type de la MSE** : [Valeur calculée]

    - **Interprétation** : Performances solides et stables, simplicité et interprétabilité du modèle.
- **Régression Ridge et Lasso** :
    - **MSE moyenne** : [Valeur calculée pour chaque modèle]
    - **Écart-type de la MSE** : [Valeur calculée]
    - **Interprétation** : Légère amélioration de la robustesse, mais sans gain significatif en précision.
- **Random Forest** :
    - **MSE moyenne** : [Valeur calculée]
    - **Écart-type de la MSE** : [Valeur calculée]
    - **Interprétation** : Bien que capable de capturer des relations complexes, la performance n'a pas justifié la complexité accrue.
- **XGBoost** :
    - **MSE moyenne** : [Valeur calculée]
    - **Écart-type de la MSE** : [Valeur calculée]
    - **Interprétation** : Bon compromis entre biais et variance, mais toujours surpassé par la simplicité et la performance stable de la régression linéaire.

Conclusion de la modélisation

La régression linéaire s'est révélée être le modèle le plus approprié pour prédire les notes finales des étudiants (`G3`). Elle combine une bonne précision prédictive, une interprétabilité claire, et une simplicité qui facilite l'application et la communication des résultats. Les modèles plus complexes n'ont pas apporté d'améliorations significatives, ce qui justifie notre choix final.

Résumé Complet de l'Exercice : Prédiction avec des Modèles de Machine Learning

Objectif :

L'objectif de cet exercice était de prédire une variable cible `G3` (probablement des performances scolaires) en utilisant différents modèles de machine

learning. Plusieurs approches ont été explorées pour identifier le modèle offrant les meilleures performances sur ce jeu de données.

Étapes Clés :

1. **Chargement et Préparation des Données** :
    - Les données ont été chargées, nettoyées et encodées. Les variables catégorielles ont été transformées en variables numériques exploitables par les modèles de machine learning grâce à `pd.get_dummies`.
2. **Sélection des Features et Identification des Features Importantes** :
    - **Analyse de la Corrélation** : Une matrice de corrélation a été utilisée pour identifier les variables les plus fortement corrélées avec la cible `G3`. Les variables avec une corrélation élevée ont été retenues pour la modélisation.
    - **Sélection Univariée** : La méthode `SelectKBest` avec la fonction `f_regression` a été utilisée pour sélectionner les features les plus pertinentes. Cette approche a permis de classer les features en fonction de leur importance statistique dans la prédiction de `G3`.
    - **Importance des Features avec Random Forest** : En utilisant un modèle Random Forest, l'importance des features a été calculée en fonction de la réduction de l'impureté (Gini ou variance) apportée par chaque feature. Les features ayant les valeurs d'importance les plus élevées ont été considérées comme les plus influentes pour le modèle.
    - **Reconnaître les Features les Plus Importantes** :
        - Les features telles que `G2` et `G1` ont été identifiées comme les plus importantes, indiquant que les performances précédentes étaient de bons indicateurs des résultats finaux.
        - Des variables comme `failures` (échecs scolaires) et `absences` (absentéisme) ont également montré une forte importance, reflétant leur impact sur les performances académiques.
        - L'analyse de l'importance des features a guidé la sélection des variables utilisées dans les modèles finaux, permettant de simplifier le modèle tout en conservant une performance élevée.
3. **Exploration des Modèles** :

- **Régression Linéaire** : Utilisée en premier comme modèle de base. Bien qu'elle ait fourni des résultats raisonnables, elle n'a pas capturé la complexité des données aussi bien que d'autres modèles.
- **Random Forest** : Ce modèle s'est avéré le plus performant. Grâce à sa capacité à gérer la variance et à capturer des relations non linéaires, il a surpassé les autres modèles testés. Il a été optimisé en utilisant la validation croisée et une recherche sur grille pour trouver les meilleurs hyperparamètres.
- **XGBoost et Gradient Boosting** : Ces modèles ont également été explorés pour exploiter les relations complexes dans les données, mais ils n'ont pas surpassé les performances de Random Forest.
- **Deep Learning** : Un modèle de deep learning a été testé en utilisant Keras avec TensorFlow. Bien que les réseaux de neurones puissent être très puissants, dans ce cas, le modèle de deep learning n'a pas apporté de bénéfices significatifs par rapport aux méthodes plus simples, probablement en raison de la taille relativement petite du jeu de données et de la nature linéaire des relations.

4. **Optimisation des Hyperparamètres** :
   - L'optimisation des hyperparamètres a été effectuée principalement sur le modèle Random Forest, en utilisant une recherche sur grille. Les paramètres tels que `n_estimators`, `max_depth`, et `min_samples_split` ont été ajustés pour maximiser les performances.
5. **Évaluation des Performances** :
   - Les modèles ont été évalués en utilisant des métriques comme le RMSE (Root Mean Squared Error) et le coefficient de détermination R^2. Ces métriques ont permis de comparer l'erreur de prédiction moyenne et la proportion de variance expliquée par chaque modèle.
   - Le modèle Random Forest a obtenu les meilleurs résultats, avec des valeurs de RMSE et de R^2 indiquant une bonne capacité de prédiction et une généralisation correcte aux nouvelles données.
6. **Prédiction sur de Nouvelles Données** :
   - Le modèle Random Forest optimisé a été utilisé pour prédire les valeurs de `G3` sur un nouveau jeu de données (`data_test`). Les données de test

ont été encodées de manière cohérente avec les données d'entraînement pour garantir la validité des prédictions.

Conclusion :

Après avoir testé plusieurs approches, le modèle de Random Forest a été retenu comme la meilleure option pour ce jeu de données. L'analyse de l'importance des features a permis de sélectionner les variables les plus pertinentes, améliorant ainsi la performance du modèle. Malgré l'essai d'un modèle de deep learning, ce dernier n'a pas montré de bénéfices significatifs, ce qui est souvent le cas avec des jeux de données plus petits ou lorsque les relations entre les variables sont bien capturées par des modèles plus simples comme Random Forest. L'optimisation et la validation croisée ont permis d'assurer que le modèle Random Forest offre des prédictions robustes et fiables.

Optimisation et Interprétation

Objectif de l'optimisation

Après avoir sélectionné le modèle de régression linéaire comme le plus performant, nous cherchons à optimiser davantage ce modèle en introduisant des caractéristiques combinées. L'objectif est de voir si ces nouvelles variables permettent d'améliorer les performances prédictives du modèle en capturant des interactions non linéaires ou des relations complexes entre les caractéristiques existantes.

Ajout des caractéristiques combinées

- **Création des caractéristiques combinées** :
  - `Absences_Health_Interaction` : Capture l'effet combiné des absences et de l'état de santé. Cette variable est calculée en multipliant `absences` par `health`.
  - `Family_Dynamics` : Reflète l'interaction entre la taille de la famille et la qualité des relations familiales, en combinant `famsize` avec `famrel`.
  - `Health_Alcohol_Interaction` : Mesure l'interaction entre la santé et la consommation d'alcool en semaine, calculée en multipliant `health` par `Dalc`.

- `Travel_Effect` : Évalue l'impact du temps de trajet et du lieu de résidence sur la performance scolaire, en combinant `traveltime` avec `address`.

- **Code Python pour l'ajout de ces caractéristiques** :

```
# Création des caractéristiques combinées
data_encoded['Absences_Health_Interaction'] = data_encod
ed['absences'] * data_encoded['health']
data_encoded['Family_Dynamics'] = data_encoded['famsize_
GT3'] * data_encoded['famrel']
data_encoded['Health_Alcohol_Interaction'] = data_encode
d['health'] * data_encoded['Dalc']
data_encoded['Travel_Effect'] = data_encoded['traveltim
e'] * data_encoded['address_R']
```

- **Inclusion dans le modèle** :
  - Ces nouvelles variables ont été ajoutées aux caractéristiques initiales pour créer un nouveau jeu de données enrichi. Nous avons réentraîné le modèle de régression linéaire sur ce jeu de données enrichi pour évaluer si ces nouvelles caractéristiques améliorent la précision des prédictions.

Interprétation des coefficients du modèle

- **Interprétation générale** :
  - Les coefficients du modèle de régression linéaire représentent l'impact attendu de chaque variable sur la variable cible `G3`, tout en tenant compte des autres variables du modèle. Un coefficient positif indique qu'une augmentation de cette variable tend à augmenter `G3`, tandis qu'un coefficient négatif indique l'inverse.

- **Coefficients des caractéristiques combinées** :
  - `Absences_Health_Interaction` : Un coefficient négatif pour cette variable pourrait indiquer que les élèves avec un nombre élevé d'absences et un mauvais état de santé ont tendance à avoir des performances plus faibles.
  - `Family_Dynamics` : Un coefficient positif pour cette variable pourrait suggérer que des relations familiales solides, combinées avec une taille

de famille plus grande, sont bénéfiques pour la performance académique.

- `Health_Alcohol_Interaction` : Un coefficient négatif pourrait refléter que les élèves qui consomment de l'alcool en semaine tout en ayant une mauvaise santé sont plus susceptibles d'avoir des notes basses.
- `Travel_Effect` : Un coefficient positif pourrait indiquer que les élèves ayant un temps de trajet plus long et vivant dans des zones rurales ne sont pas nécessairement désavantagés, ou que d'autres facteurs compensent cet effet.

Évaluation des améliorations apportées

- **Validation croisée** :
  - Le modèle a été réévalué en utilisant la validation croisée pour comparer la MSE avant et après l'ajout des caractéristiques combinées. Si la MSE diminue après l'ajout de ces caractéristiques, cela indique que le modèle a gagné en précision grâce à ces ajouts.
- **Sélection finale du modèle** :
  - Si les caractéristiques combinées apportent une amélioration significative, elles seront conservées dans le modèle final. Sinon, elles seront éventuellement écartées pour maintenir un modèle plus simple et plus interprétable.

Conclusion de l'optimisation

L'ajout de caractéristiques combinées permet d'explorer des relations plus complexes entre les variables, offrant potentiellement un modèle plus précis. L'interprétation des coefficients aide à comprendre l'impact de chaque variable, et les caractéristiques combinées apportent des insights supplémentaires sur les dynamiques qui influencent les performances académiques.

Conclusion du Rapport

Résumé des principaux résultats

L'étude visait à prédire les performances académiques des étudiants, mesurées par la note finale `G3`, en se basant sur un ensemble de variables couvrant des

aspects démographiques, scolaires, et personnels. Le processus a suivi plusieurs étapes clés :

1. **Exploration des Données** : Nous avons d'abord exploré les données pour comprendre les distributions des variables et leurs relations avec la variable cible. Les analyses ont révélé que les notes intermédiaires (`G1` et `G2`) sont les prédicteurs les plus forts de `G3`, mais que d'autres variables, telles que le temps d'étude (`studytime`), l'éducation des parents (`Medu` et `Fedu`), et la consommation d'alcool (`Walc`, `Dalc`), jouent également un rôle important.
2. **Préparation des Données** : L'imputation par régression a été utilisée pour gérer les valeurs manquantes, en particulier dans les variables `G1` et `G2`. Des caractéristiques combinées ont été créées pour capturer des dynamiques complexes entre des variables telles que les absences et l'état de santé, les dynamiques familiales, et les effets géographiques sur les performances scolaires.
3. **Modélisation et Évaluation** : Plusieurs modèles de régression ont été testés, y compris la régression linéaire, Ridge, Lasso, Random Forest, et XGBoost. La régression linéaire a été sélectionnée comme le modèle le plus performant en termes de précision et d'interprétabilité, avec une erreur quadratique moyenne (MSE) relativement faible et une stabilité prouvée à travers la validation croisée.
4. **Optimisation et Interprétation** : L'ajout de caractéristiques combinées a permis d'améliorer légèrement la précision des prédictions. L'interprétation des coefficients du modèle a fourni des insights précieux sur les facteurs les plus influents sur la performance académique, confirmant l'importance de l'engagement scolaire, du soutien familial, et de la gestion des comportements à risque.

Implications des résultats

Les résultats de cette étude ont des implications importantes pour les politiques éducatives et les interventions pédagogiques :

- **Soutien ciblé** : Les étudiants présentant des caractéristiques associées à des performances académiques plus faibles (comme des absences fréquentes ou une faible étude) pourraient bénéficier d'un soutien scolaire accru ou de programmes d'intervention spécifiques.
- **Rôle des parents** : L'influence significative de l'éducation des parents sur les performances académiques souligne l'importance de l'engagement

familial dans le suivi scolaire. Des initiatives pour impliquer davantage les parents dans le parcours éducatif de leurs enfants pourraient être bénéfiques.

- **Gestion des comportements à risque** : La consommation d'alcool et son impact négatif sur les résultats scolaires indiquent un besoin d'éducation et de programmes de prévention visant à réduire ces comportements chez les jeunes.

Recommandations pour les études futures

Bien que cette étude ait permis de développer un modèle prédictif robuste, plusieurs axes d'amélioration et de recherche future peuvent être envisagés :

- **Extension des données** : Intégrer d'autres variables contextuelles, telles que des informations socio-économiques ou des données sur l'environnement scolaire, pourrait améliorer la précision du modèle.
- **Exploration de modèles plus complexes** : Bien que les modèles complexes n'aient pas significativement surpassé la régression linéaire dans cette étude, l'application de méthodes plus avancées avec une optimisation plus poussée des hyperparamètres pourrait offrir de meilleurs résultats.
- **Études longitudinales** : Une analyse longitudinale suivant les performances des étudiants sur plusieurs années pourrait offrir des insights supplémentaires sur les facteurs qui influencent la progression académique au fil du temps.

Conclusion finale

Cette étude a permis de démontrer l'efficacité de la régression linéaire pour prédire les performances académiques à partir de données variées sur les étudiants. Les résultats obtenus sont non seulement précieux pour comprendre les facteurs influençant les résultats scolaires, mais également pour guider des interventions ciblées visant à améliorer la réussite des étudiants. L'approche méthodologique rigoureuse suivie tout au long de l'analyse garantit la robustesse des conclusions, tout en offrant des pistes intéressantes pour des recherches futures.