+1 sur la note

Travail de groupe Machine Learning

Haryadi Nayla - Dugand Elise - Belmont Sébastien

BASE DE DONNÉE :

La base de données utilisée provient de data gouv et couvre les admissions en 2018 des étudiants.dans l'enseignement supérieur

Nettoyage des données :

- Suppression des colonnes inutiles.
- Traitement des valeurs manquantes par imputation ou suppression des lignes incomplètes.
- Normalisation des catégories comme les noms de régions et de filières.
- Élimination des doublons

Ces étapes nous ont permis d'obtenir une base fiable et prête pour notre première analyse avec un modèle de régression linéaire.

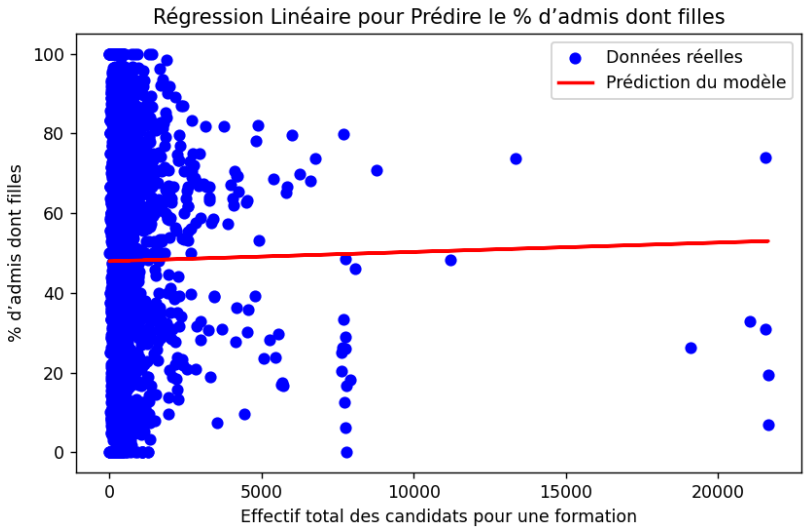
Démarche et colonnes retenues

Pour répondre à notre problématique, nous avons filtré les colonnes de notre base de données pour ne conserver que celles qui étaient pertinentes. Voici les colonnes principales que nous avons décidé de conserver et leur utilité :

- **Filière de formation très agrégée** / **Filière de formation** : pour classifier les données selon le type de formation.
- **Effectif des candidats présents** : pour évaluer le nombre total de candidats ayant participé aux processus d'admission.
- **Effectif des admis** : Pour analyser combien de candidats ont été effectivement admis.
- **% d'admis néo-bacheliers** : Pourcentage de nouveaux bacheliers admis, permettant de cibler les admissions des étudiants sortant directement du lycée.
- **% d'admis néo-bacheliers avec mention Assez Bien/Bien/Très Bien au bac** : Pour examiner comment les performances au baccalauréat influencent les admissions.

+1 sur la note

Modèle 1.1 : Régression linéaire pour voir les tendances d'admission dans les établissements d'enseignement supérieur.



Pour ce cas précis, on a utilisé une régression linéaire pour analyser les variations des taux en fonction des effectifs présents.

Résultat ,1 élève sur 2 est admis quand on ne regarde que la courbe. Cependant si l'on regarde le détail avec l'ensemble des données, on peut réaliser que la réalité est toute autre. La majorité des points de données sont concentrés vers le côté gauche du graphique, où l'effectif total des candidats est relativement faible (moins de 5000). Cela suggère que la plupart des formations ont un petit nombre de candidats.

Modèle 1.2: Impact de la mention au baccalauréat sur le choix des filières → Régression linéaire

Ici, nous avons étudié l'impact des mentions au bac sur le choix des filières via une régression linéaire à nouveau.

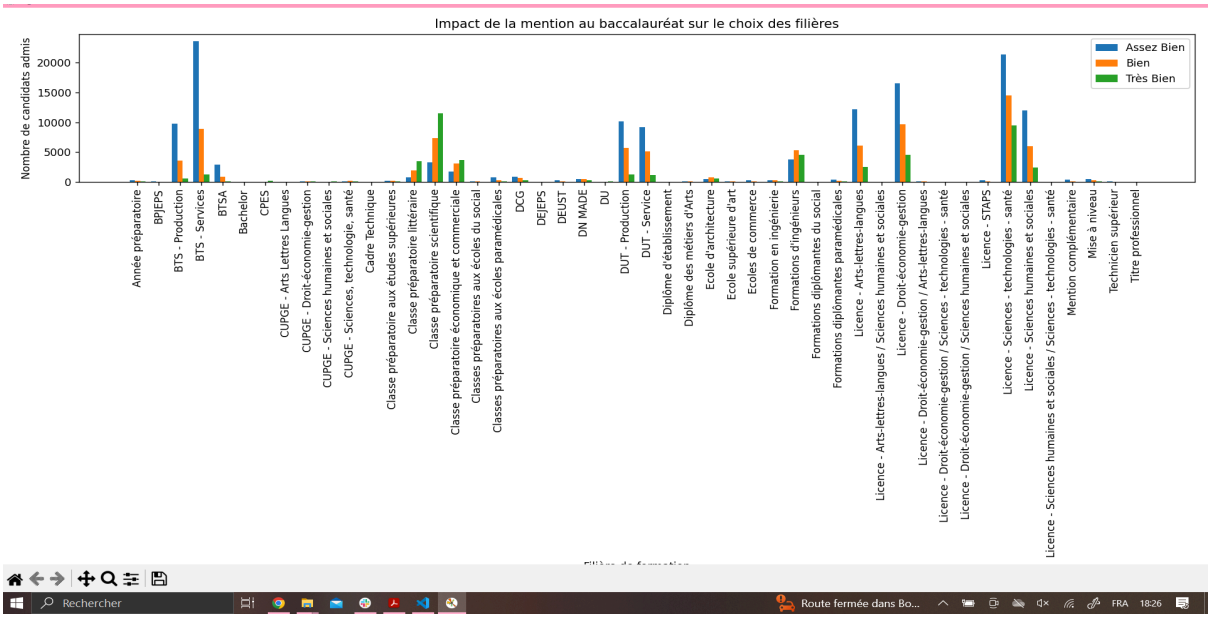
Résultats clés :

- Les filières telles que "Licence - Sciences" et "Licence - Droit" montrent un nombre notable de candidats avec les mentions "Bien" et "Très Bien", ce qui peut indiquer une plus grande sélectivité ou attractivité pour les étudiants ayant obtenu de meilleures performances au baccalauréat

- Certaines filières montrent une grande variabilité dans le type de mention obtenue. Par exemple, les filières comme "Licence - Sciences économiques et gestion" et "Licence - Lettres" accueillent une répartition relativement équilibrée des mentions, suggérant une accessibilité diverse

+1 sur la note

Ainsi ces données suggèrent que les étudiants avec des mentions plus élevées tendent à s’orienter vers des filières perçues comme plus “prestigieuses” ou du moins plus “exigeantes”, telles que les sciences ou le droit.



Modèle 2 : Forêt Aléatoire

Bases de données choisies et source

Base admissions 2018 (la même que pour les précédentes visualisations) :

- Source : data.gouv.fr
- Contenu : Informations sur les admissions dans les établissements d'enseignement supérieur en 2018
- Variables clés : Effectifs de candidats et d'admis, région de l'établissement, et type de formation

Base indicateurs socio-économiques 2018 :

- Source : INSEE.
- Contenu : Indicateurs socio-économiques régionaux tels que le revenu médian, le taux de bas revenus, et l'indice de Gini
- Ces deux bases complémentaires nous permettent d'explorer l'impact des conditions socio-économiques sur les taux d'admission.

+1 sur la note

Problématique choisie

Les conditions socio-économiques régionales représentent-elles un facteur impactant sur les taux d'admission et donc l'accès à l'enseignement supérieur en 2018 ?

Modèle choisi et pourquoi

Nous avons choisi un modèle de forêt aléatoire. Ce modèle est adapté pour :

- Identifier les caractéristiques socio-économiques les plus importantes expliquant les différences de taux d'admission
- Catégoriser les régions en trois groupes basés sur leurs taux d'admission : faible (Low), moyen (Medium), et élevé (High).

Ce choix repose sur :

- La capacité du modèle à gérer des données complexes avec des interactions entre variables.
- L'objectif d'analyser l'importance des variables pour mieux comprendre les facteurs influençant l'accès à l'enseignement supérieur.

Les étapes de nettoyage des bases de données et leur but

Base admissions :

- Sélection des colonnes pertinentes : nous avons conservé les variables relatives aux régions, aux effectifs de candidats et d'admis, et aux types de formations
- Suppression des doublons et des valeurs manquantes : pour garantir la qualité des données et éviter les erreurs
- Ajout d'une colonne de taux d'admission : calculé comme le rapport entre le nombre d'admis et le nombre total de candidats

Base socio-économique :

- Extraction des indicateurs clés : Revenu médian, taux de bas revenus, et indice de Gini
- Conversion des codes IRIS en codes régionaux : Agrégation des données au niveau régional pour correspondre aux données d'admissions
- Suppression des doublons et gestion des valeurs manquantes : Pour garantir la cohérence et la fiabilité des données

+1 sur la note

Fusion des bases

Nous avons fusionné les deux bases sur la clé régionale (Region Code) après avoir harmonisé les formats des deux bases. Les données socio-économiques ont été agrégées au niveau régional pour correspondre aux données d'admissions.

Analyse en fonction des résultats observés

Importance des caractéristiques :

Le graphique montre l'impact relatif des caractéristiques socio-économiques sur la prédiction des catégories de taux d'admission :

Median income (revenu médian) :

- Variable ayant le plus grand impact.
- Les régions avec un revenu médian élevé ont généralement des taux d'admission supérieurs.

Low income rate (taux de bas revenus) :

- Importance modérée.
- Les régions avec un pourcentage élevé de ménages à faibles revenus montrent une corrélation avec des taux d'admission plus faibles.

Gini index (indice de Gini) :

- Importance plus faible comparée aux autres.
- Les inégalités économiques n'expliquent qu'une partie des variations des taux d'admission.*

Tendances observées :

1. **Impact du revenu médian (variable clé) :**

Les régions économiquement favorisées (revenu médian élevé) affichent des taux d'admission plus élevés, soulignant l'avantage des conditions économiques.

2. **Influence modérée du taux de bas revenus :**

Un taux de bas revenus élevé est corrélé à des taux d'admission plus faibles, mais cette influence reste moindre comparée au revenu médian.

3. **Inégalités économiques :**

L'indice de Gini a un impact réduit dans ce modèle, ce qui pourrait indiquer que d'autres facteurs (comme des politiques spécifiques à l'éducation) jouent un rôle dans l'atténuation des inégalités.

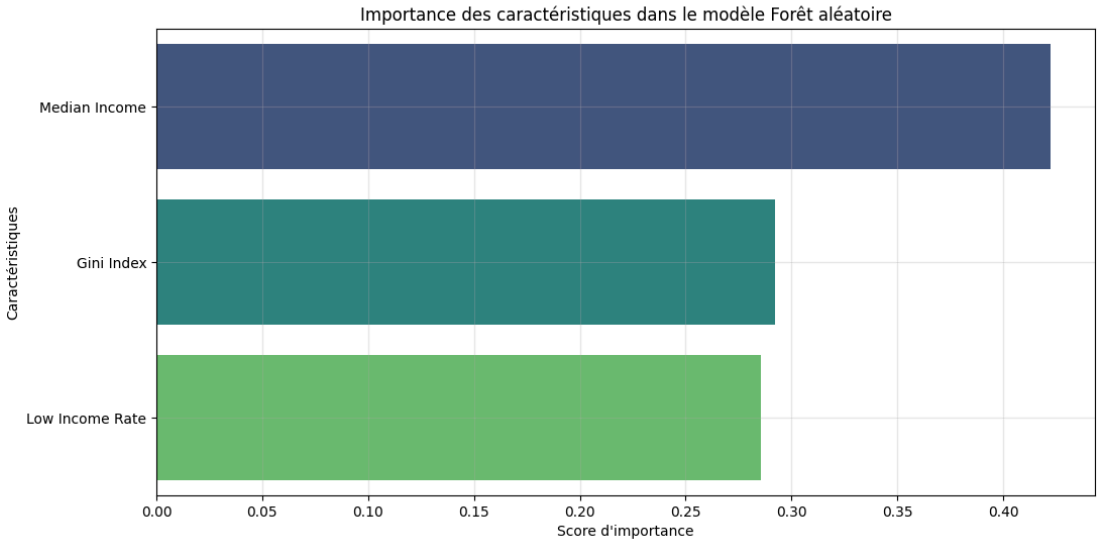
+1 sur la note

Conclusion :

Cette analyse nous a permi de mettre en évidence que le **revenu médian** est le facteur clé expliquant les disparités dans les taux d'admission régionaux.

Pour pousser l'analyse, nous pourrions notamment réaliser l'analyse sur plusieurs années avec le même type de données pour constater ou non une évolution et en tirer des conclusions utiles aux politiques régionales.

La forêt aléatoire s'est avérée utile pour identifier et hiérarchiser les variables influentes.



Explication et lecture du résultat :

En abscisse on retrouve le score d'importance, il indique l'impact relatif de chaque caractéristique sur la prédiction des taux d'admission. Plus le score est élevé, plus la caractéristique influence fortement la catégorisation des régions

En ordonnée on retrouve les variables analysées donc Median income (revenu médian), low income rate (taux de bas revenus) et enfin Gini index qui permet de mesurer les inégalités économiques dans une région.

Chaque barre va représenter l'importance relative d'une caractéristique et donc son impact sur le taux d'admission, les couleurs sont simplement présentes pour faciliter la lecture et la compréhension de la visualisation.

Voici l'interprétation/lecture des résultats par caractéristiques (cités auparavant) :

Revenu médian (median income) :

- Score le plus élevé, indiquant que les régions avec un revenu médian plus élevé ont un impact significatif sur les taux d'admission.

+1 sur la note

Taux de bas revenus (low income rate) :

- Impact modéré, reflétant une corrélation entre les ménages à faibles revenus et des taux d'admission plus faibles, surtout par rapport à la première variable.

Indice de Gini (Gini index) :

- Importance relative plus faible, suggérant que les inégalités économiques influencent moins directement les taux d’admission.