Projet Big Data

Analyse du Baccalauréat

PARIS SCHOOL OF BUSINESS

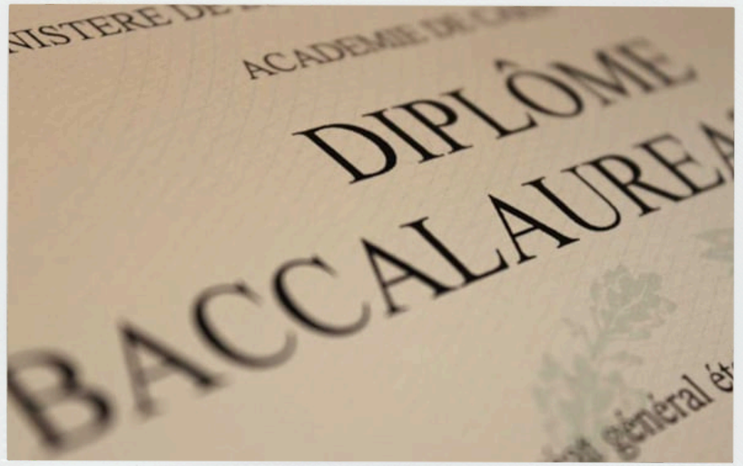
SOMMAIRE

1. Introduction

- Le baccalauréat est une étape clé du système éducatif français.
- Les données permettent d'analyser la performance des académies, des filières et des profils d'élèves.
- L'étude prend en compte le sexe, le statut, la série et la voie de formation des élèves.
- Le projet exploite des outils Big Data.

Objectif ?

Détecter des disparités & prédire l'obtention d'une mention TB.



2. Contexte Général et Problématique

- **Des écarts de performance sont observés entre académies, voies et profils d’élèves.**
- **L’objectif est de détecter les déséquilibres et comprendre les facteurs de réussite ou d’échec.**
- **Le projet cherche à identifier les profils ayant le plus de chances d’obtenir une mention Très Bien.**
- **Questions clés :**
  - Quelles académies ont les meilleurs taux de réussite ?
  - Quel est le rôle du sexe, du statut et de la voie dans les résultats ?
  - Peut-on prédire l'obtention d'une mention TB ?

3. Description du Jeu de Données

- **Source :** Données publiques éducatives
- **Nom du fichier :** fr-en-baccalaureat-par-academie.csv
- **Taille :** 34 371 lignes, 23 colonnes
- **Format :** CSV
- **Période** : 2021 - 2024
- **Variables clés :**
  - *Académie, Session, Sexe, Voie, Série, Statut du candidat*
  - *Nombre d'inscrits, Nombre de présents,*
  - *Nombre d'admis, mentions, refusés*

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 34371 entries, 0 to 34370
Data columns (total 23 columns):
 #   Column                                                Non-Null Count  Dtype
---  ------                                                --------------  -----
 0   num_ligne                                             34371 non-null  float64
 1   Session                                               34371 non-null  int64
 2   Code académie                                         34371 non-null  int64
 3   Académie                                              34371 non-null  object
 4   Sexe                                                  34371 non-null  object
 5   Statut du candidat                                    34371 non-null  object
 6   Voie                                                  34371 non-null  object
 7   Série                                                 34371 non-null  object
 8   Diplôme spécialité                                    34371 non-null  object
 9   Nombre d'inscrits                                     34371 non-null  int64
 10  Nombre de présents                                    34371 non-null  int64
 11  Nombre d'admis au 1er groupe                          34371 non-null  int64
 12  Nombre de refusés au 1er groupe                       34371 non-null  int64
 13  Nombre d'ajournés, passant les épreuves du 2nd groupe 34371 non-null  int64
 14  Nombre d'admis à l'issue du 2nd groupe                34371 non-null  int64
 15  Nombre de refusés à l'issue du 2nd groupe             34371 non-null  int64
 16  Nombre d'admis totaux                                 34371 non-null  int64
 17  Nombre d'admis avec mention TB avec les félicitations du jury  34371 non-null  int64
 18  Nombre d'admis avec mention TB sans les félicitations du jury  34371 non-null  int64
 19  Nombre d'admis avec mention B                         34371 non-null  int64
 20  Nombre d'admis avec mention AB                        34371 non-null  int64
 21  Nombre d'admis sans mention                           34371 non-null  int64
 22  Nombre de refusés totaux                              34371 non-null  int64
dtypes: float64(1), int64(16), object(6)
memory usage: 6.0+ MB
```

4. Analyse Exploratoire des Données (EDA)

- **L'analyse exploratoire a été réalisée à l'aide de Pyhton avec Google Collab :**
  - Statistiques descriptives :

- **Académies avec le plus d'inscrits :**
  VERSAILLES (**281 366**),
  CRETEIL (**215 425**),
  LILLE (**184 836**),
  NANTES (**171 031**)

- **Voies :** GENERALE (**1 544 276**)
  - PROFESSIONNELLE (**839 607**),
    TECHNOLOGIQUE (**598 439**)

- **Sexe** : FEMININ (**1 495 643**),
  MASCULIN (**1 469 727**)

```
Nombre total d'étudiants inscrits par voie :
Voie
BAC GENERAL          1544276
BAC PROFESSIONNEL     839607
BAC TECHNOLOGIQUE     598439
```

Sexe	
FEMININ	1495643
MASCULIN	1469727

Académie	Nombre d'inscrits
VERSAILLES	281616
CRETEIL	215666
LILLE	185012
NANTES	171224
RENNES	151799

5. Préparation des Données

- **Création indicateurs supplémentaires :**
  - **TauxReussite =** Nombre d'admis totaux / Nombre de présents
  - **TauxMentionTB** = (Mentions TB avec + sans félicitations) / Nombre d'admis totaux

Code Python :

```
df["TauxReussite"] = df["Nombre d'admis totaux"] / df["Nombre de présents"]
df["TauxMentionTB"] = (
 (df["Nombre d'admis avec mention TB avec les félicitations du jury"] +
 df["Nombre d'admis avec mention TB sans les félicitations du jury"]) /
 df["Nombre d'admis totaux"]
)

```

- **Nettoyage des valeurs nulles et préparer les données pour la visualisation et le Machine Learning**

6. Visualisation et Interprétation des Données

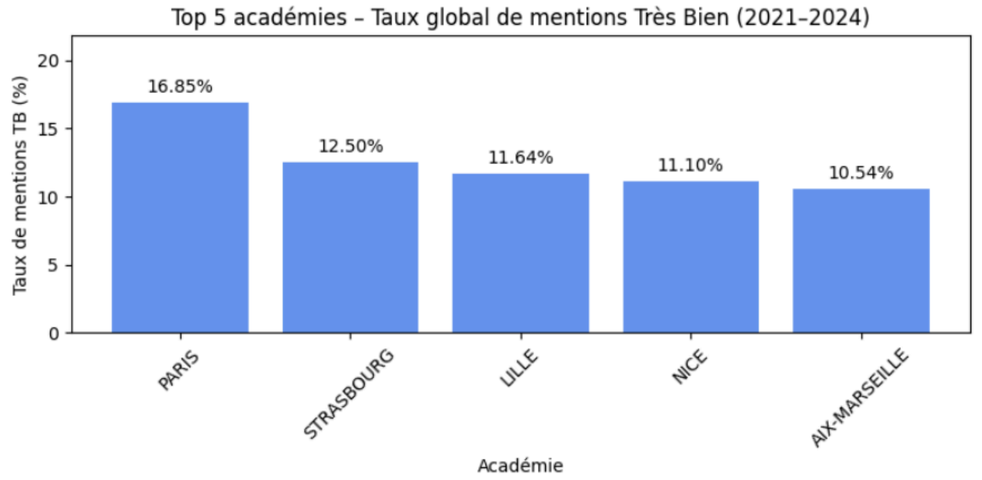
- **Objectif :** Comprendre les différences de performance entre les voies du bac et les académies.

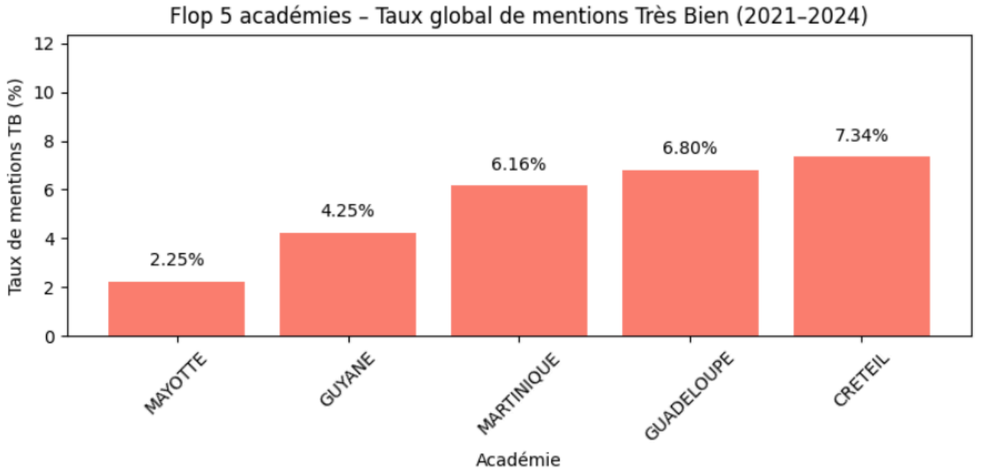
- **Académies les plus performantes :**
  - Paris **(16,85 %)**
  - Strasbourg **(12,5 %)**
  - Lille **(11,64 %)**

- **Ces régions combinent souvent un bon taux de réussite et un fort taux de mentions TB.**

- **Explication possible :** Bon niveau global des élèves ou soutien institutionnel renforcé.

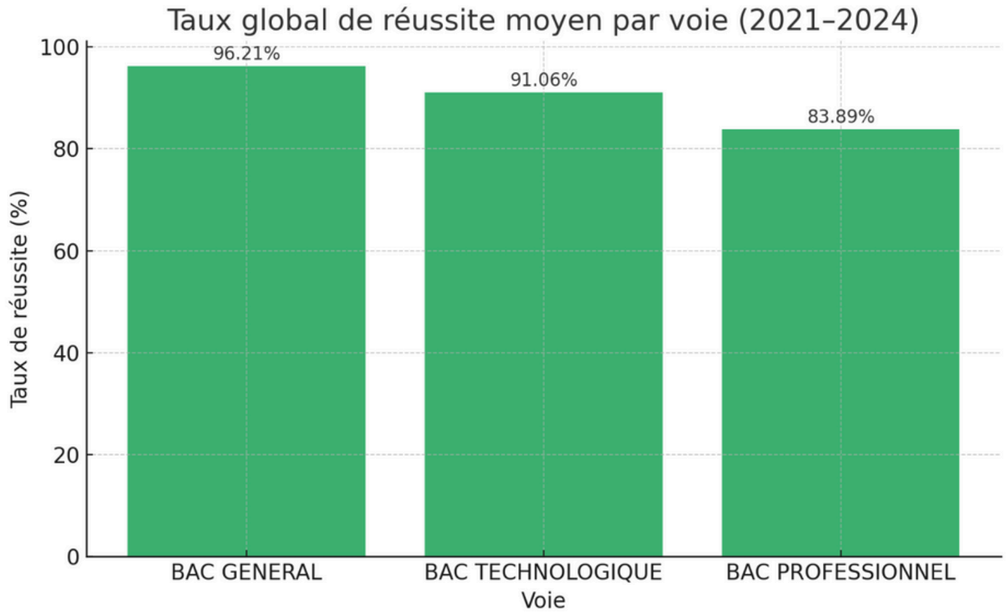
- **Académies moins performantes :**
  - Taux de mentions TB plus faibles.
  - Cela pourrait s'expliquer par des inégalités socio-économiques, d'accès à l'information ou à un enseignement de qualité.





6. Visualisation et Interprétation des Données

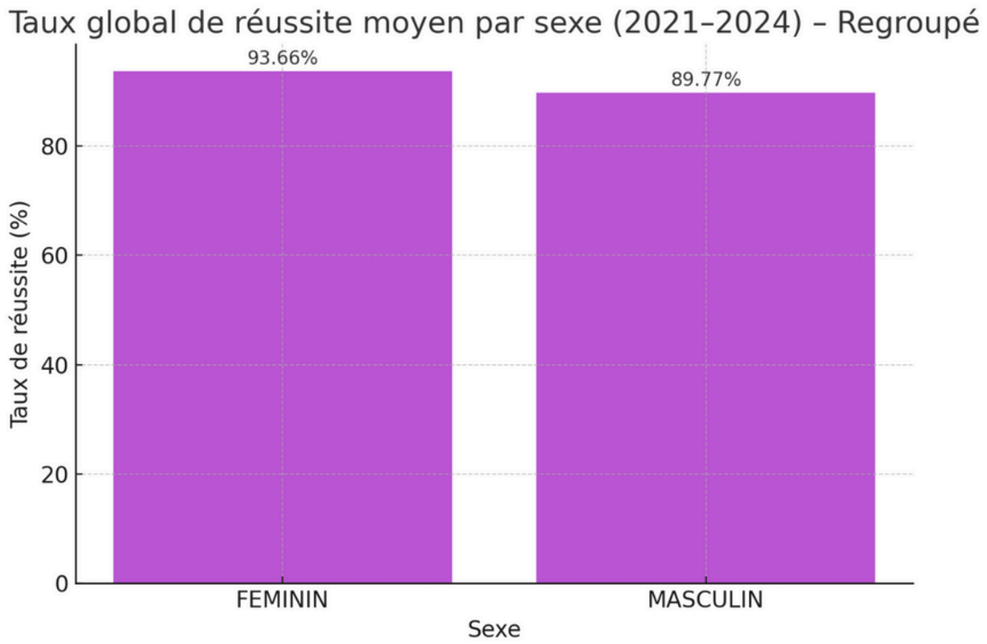
- **Taux de réussite par voie :**
  - **Voie Générale : 96,21 %**, reflétant son objectif académique et préparatoire aux études supérieures.
  - **Voie Technologique : 91,6 %**, un bon taux mais inférieur à celui de la voie générale.
  - **Voie Professionnelle : 83,89 %**, plus bas, en partie à cause de difficultés structurelles (public plus hétérogène, insertion directe, etc.).
- **Conclusion :** Ces écarts montrent qu'il faut mieux accompagner les élèves, surtout dans les filières professionnelles.



6. Visualisation et Interprétation des Données

- **Les filles ont un taux de réussite légèrement supérieur à celui des garçons.**

- **Cette différence, souvent constatée dans les statistiques scolaires, pourrait s'expliquer par :**
  - Un meilleur accompagnement pédagogique.
  - Une implication différente des filles vis-à-vis de l'examen.

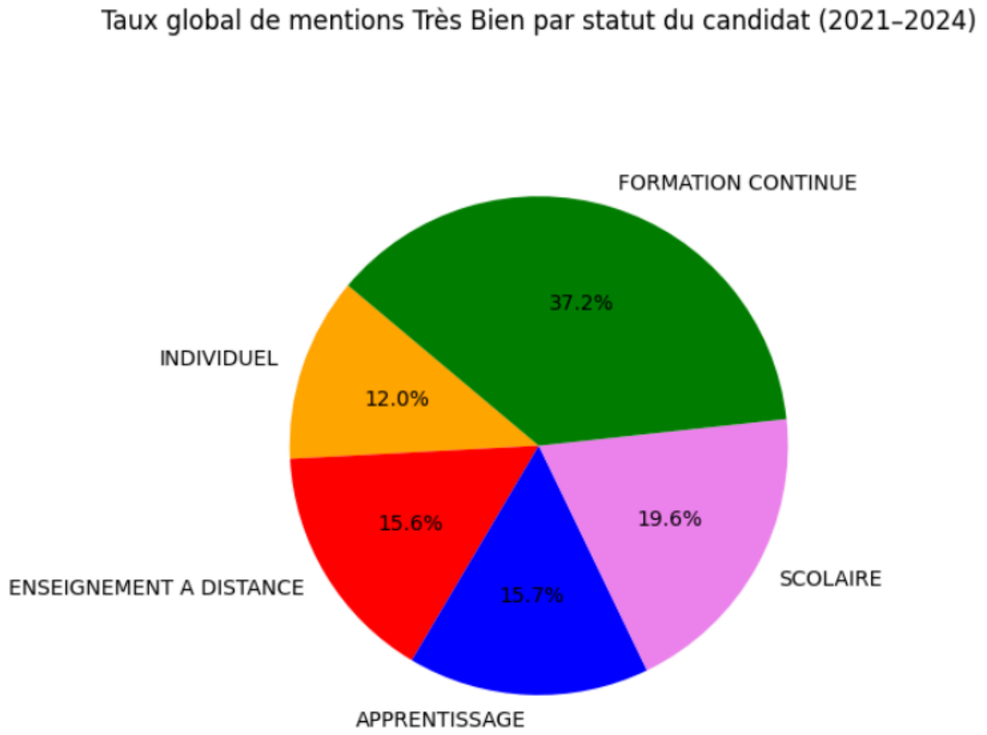


6. Visualisation et Interprétation des Données

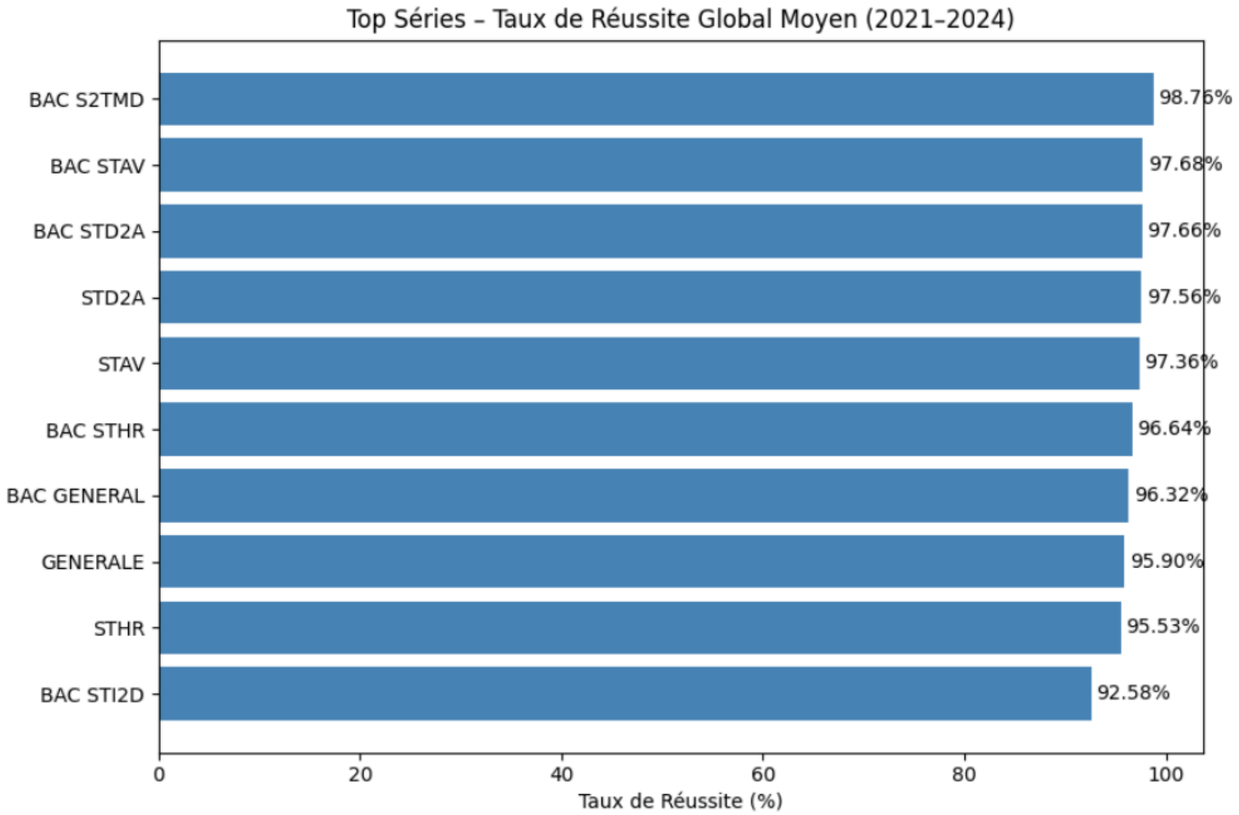
- **Les élèves du parcours scolaire obtiennent plus souvent une mention TB que ceux en apprentissage.**

- **Cela peut s'expliquer par :**
  - Un accès inégal aux ressources.
  - Une préparation académique moins poussée en alternance.
  - Un encadrement différent.

- **Conclusion :** Il est nécessaire de mieux soutenir les filières en alternance pour favoriser l’excellence.



6. Visualisation et Interprétation des Données



- **Les séries générales (ES, S, L) ont des taux de réussite très élevés.**

- **Les séries technologiques et professionnelles présentent une plus grande variabilité.**

- **Le graphique met en lumière :**
  - Les zones de performance.
  - L'impact du type de bac sur les résultats.
  - Le besoin éventuel de renforcer l'accompagnement dans certaines filières.

6. Visualisation et Interprétation des Données

Constats généraux :

- La voie générale est avantagée en matière de résultats et de reconnaissance des performances.
- Les disparités régionales restent importantes.

Intérêt de la suite du projet :

- Utiliser le machine learning pour prédire l'obtention de la mention Très Bien.
- Mieux comprendre les facteurs qui influencent cette réussite.

7. Application des Algorithmes de Machine Learning

Préparation du modèle

- **Objectif :** Prédire si une ligne du dataset correspond à une mention TB (variable cible MentionTB = 1 si >20 % des admis ont eu TB).

- **Variables utilisées :**
  - Voie (encodée avec *StringIndexer*)
  - Nombre de présents
  - Nombre d'admis
  - Données vectorisées avec *VectorAssembler*.

Choix du modèle : Random Forest

- **Modèle choisi :** Random Forest Classifier pour une meilleure généralisation.
- **Paramètres :**
  - ***numTrees* = 100 (100 arbres de décision)**
  - ***maxDepth* = 10 (profondeur maximale de l'arbre)**

- **Avantage :** Réduit le surapprentissage par rapport à un arbre de décision unique.

Résultats

- **Précision (Accuracy) du modèle : 91,1 %**
- **Gain de précision significatif (+9 %) grâce au Random Forest.**
- **Le modèle est jugé plus fiable pour prédire les mentions Très Bien.**

7. Conclusion

L’analyse a mis en lumière les disparités entre voies et académies dans les résultats du baccalauréat.
Pyhton a permis de gérer un grand volume de données et d’appliquer un modèle de machine learning efficace.

- **Perspectives :**
  - Tester des modèles plus avancés (comme Gradient Boosted Trees ou XGBoost).
  - Intégrer des données socio-économiques régionales pour affiner l’analyse.
  - Créer un dashboard dynamique de suivi par académie.

Merci !

BOUDJERIDA NADIR

PARIS SCHOOL OF BUSINESS