DATA JOB

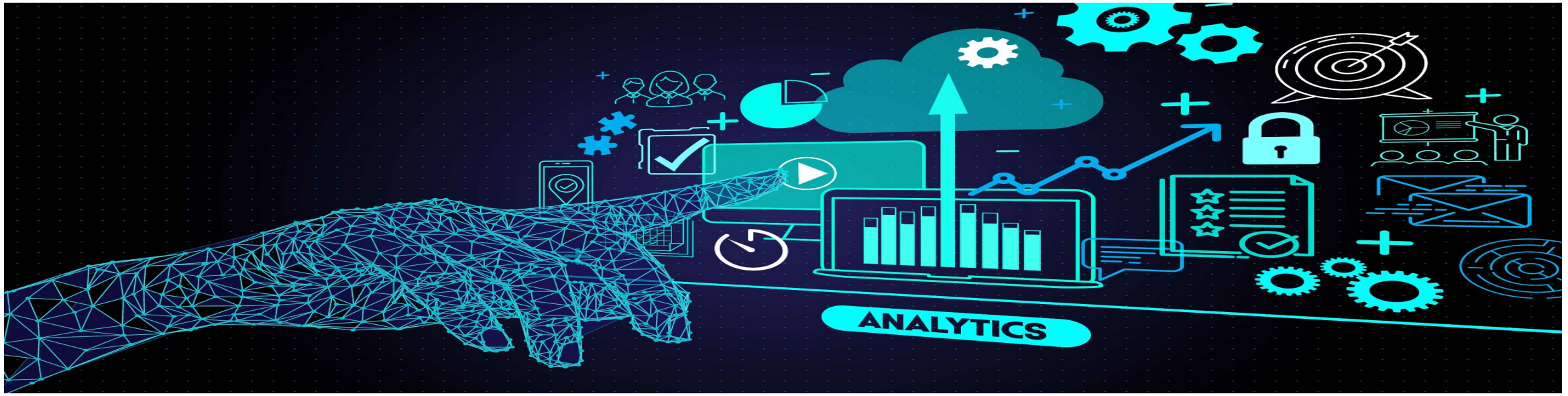
Analyse de données, afin de recommander un métier de la data.

Annabelle et Lydie le 15 juillet 2025

MINES PARIS
EXECUTIVE EDUCATION | PSL

DataScientest • com

Les objectifs



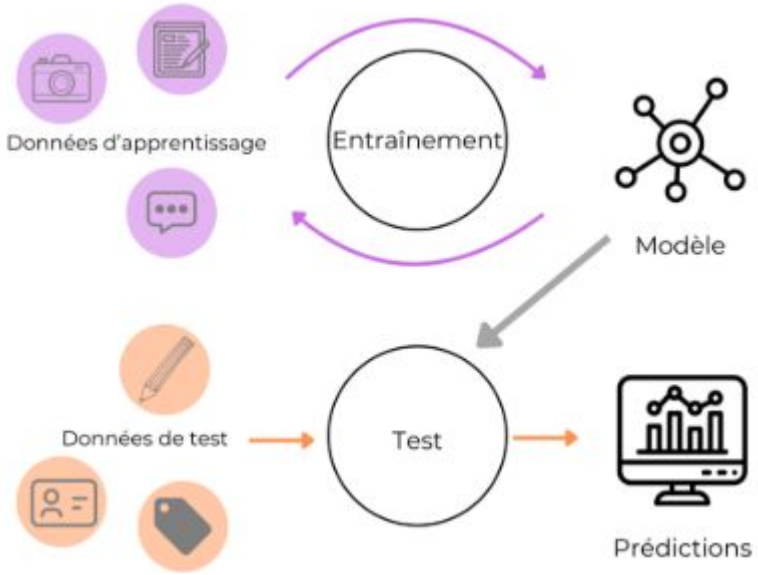
MINES PARIS
EXECUTIVE EDUCATION | PSL

DataScientest • com

Les étapes du projet

Exploratory Data Analysis

EDA : principes généraux



Annabelle et Lydie le 15 juillet 2025

Données utilisées

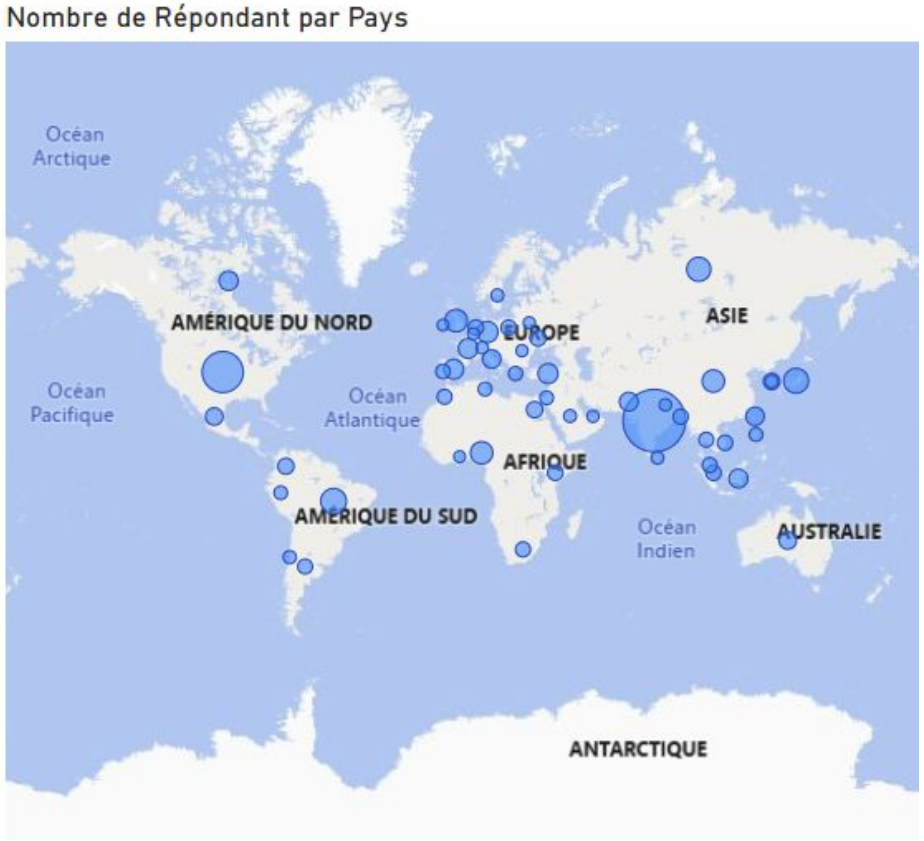
Source:

- Sondage public Kaggle 2020

Contenu:

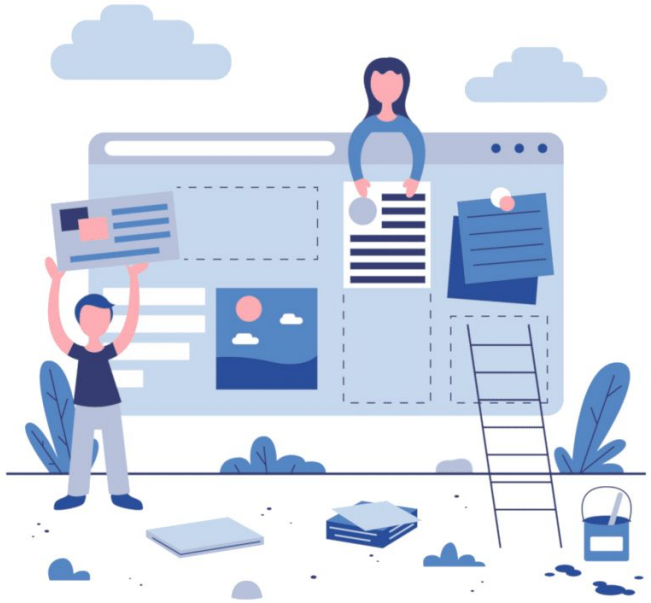
- 355 variables
- Profils variés
- Richesse des thèmes

Jeu de données à l'échelle mondiale



Annabelle et Lydie le 15 juillet 2025

Exploration initiale



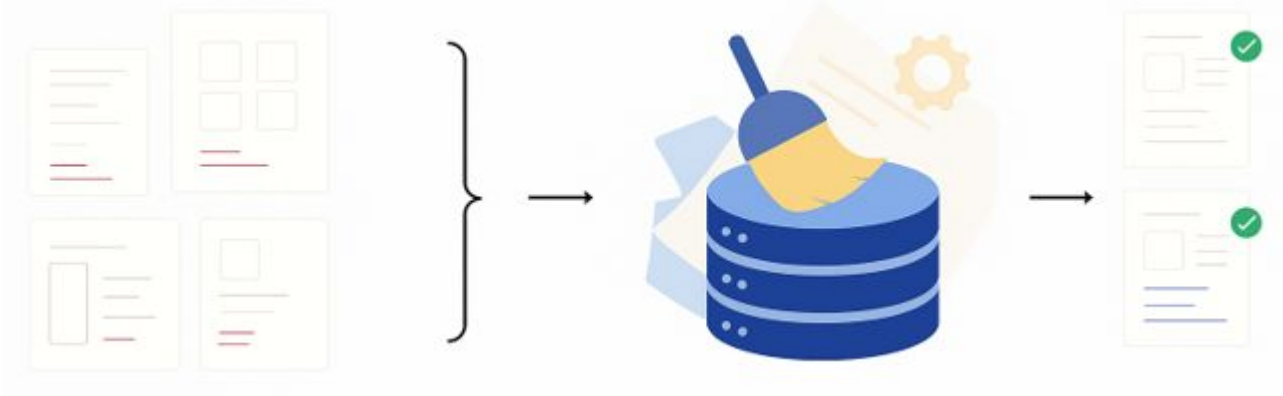
Annabelle et Lydie le 15 juillet 2025

Nettoyage des données

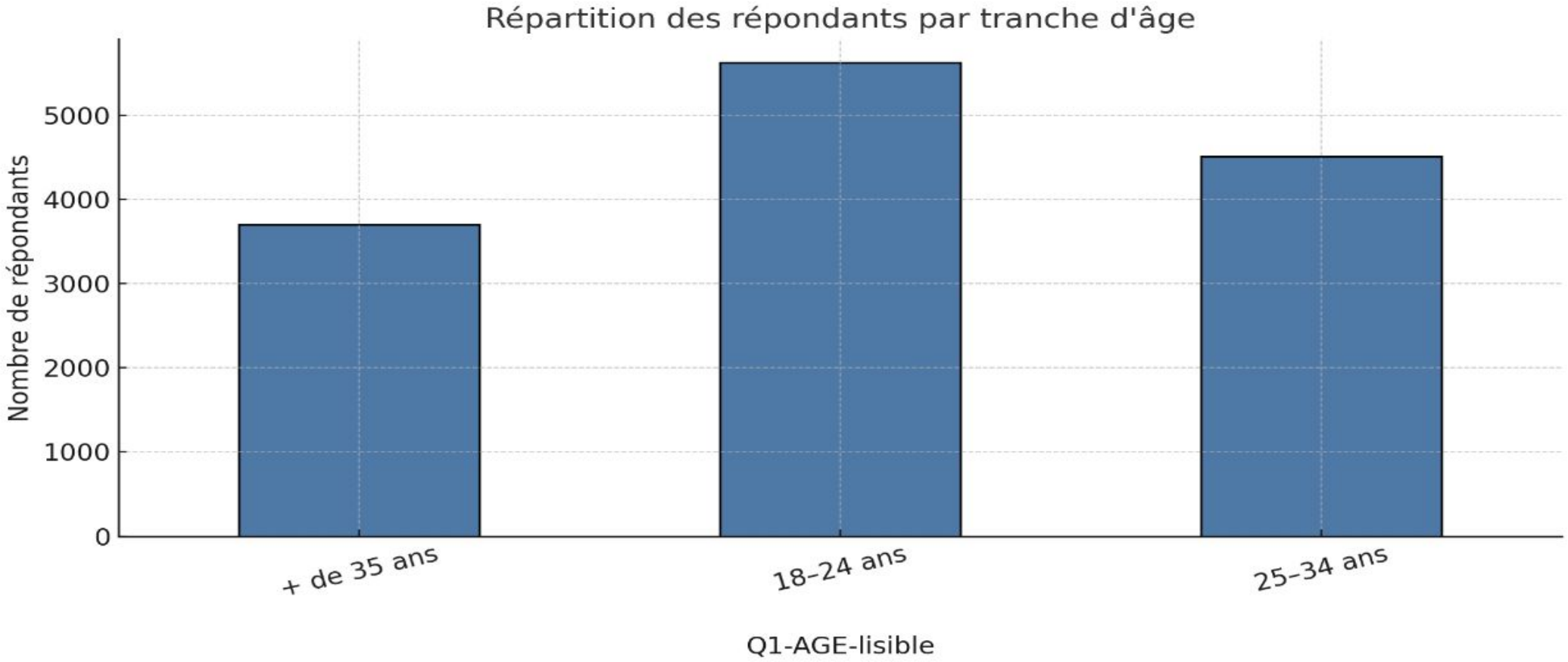
- Suppression des lignes incomplètes
- Nettoyage adapté au type de chaque question
- Suppression des colonnes sans réponses chez les étudiants

Nettoyage des données

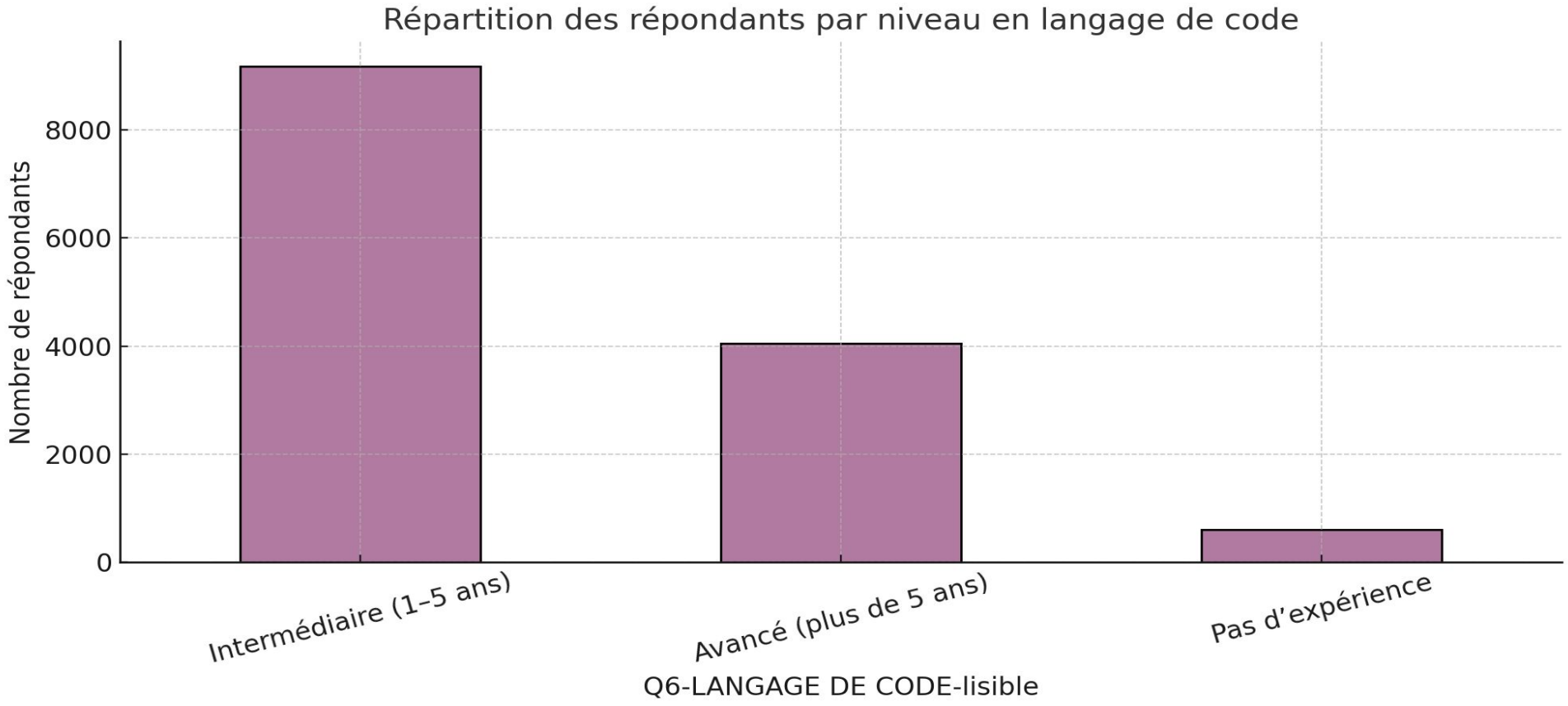
- Suppression des doublons
- Remplacement des valeurs manquantes
- Encodage des variables



Analyse univariée

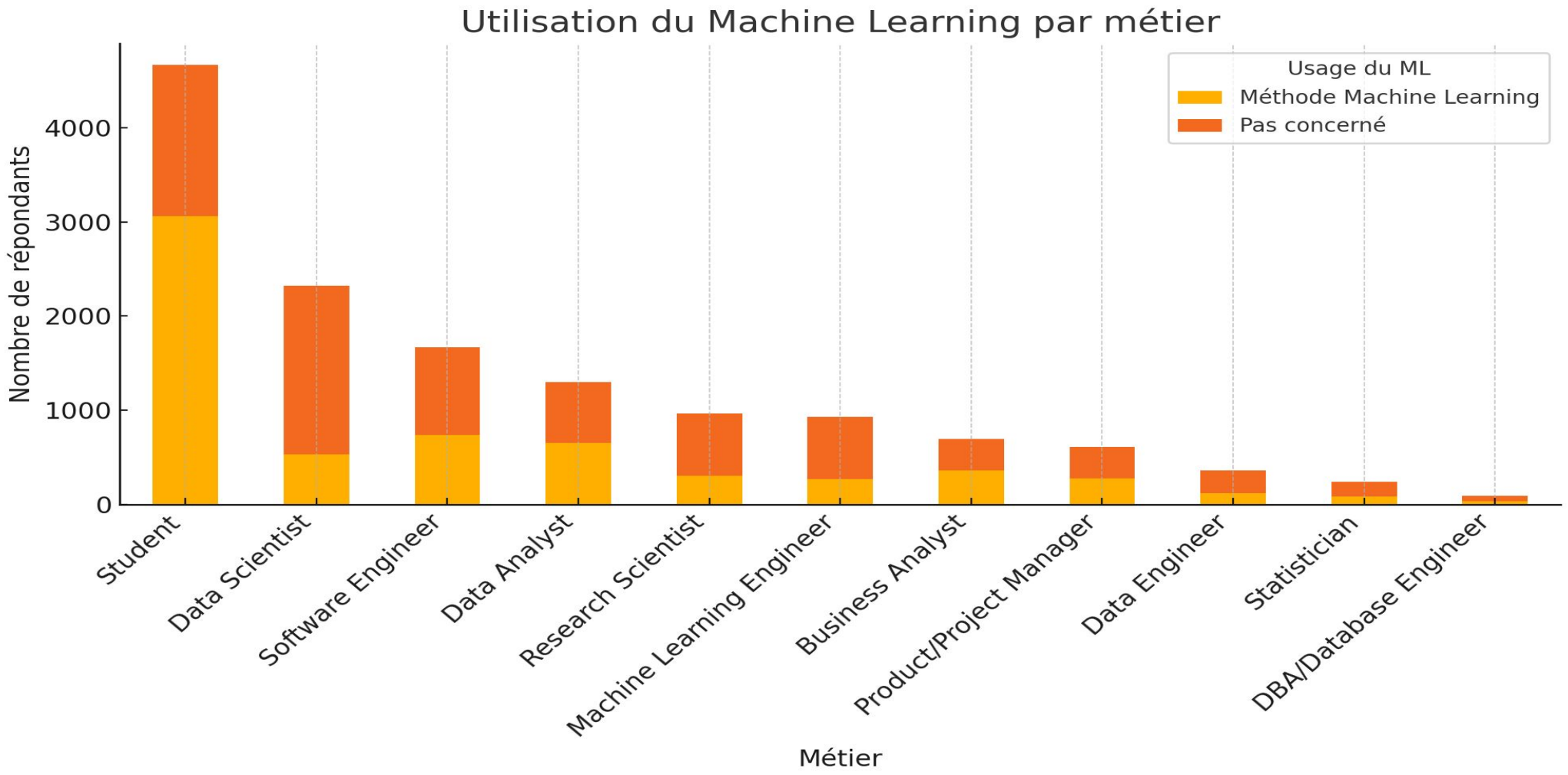


Analyse univariée

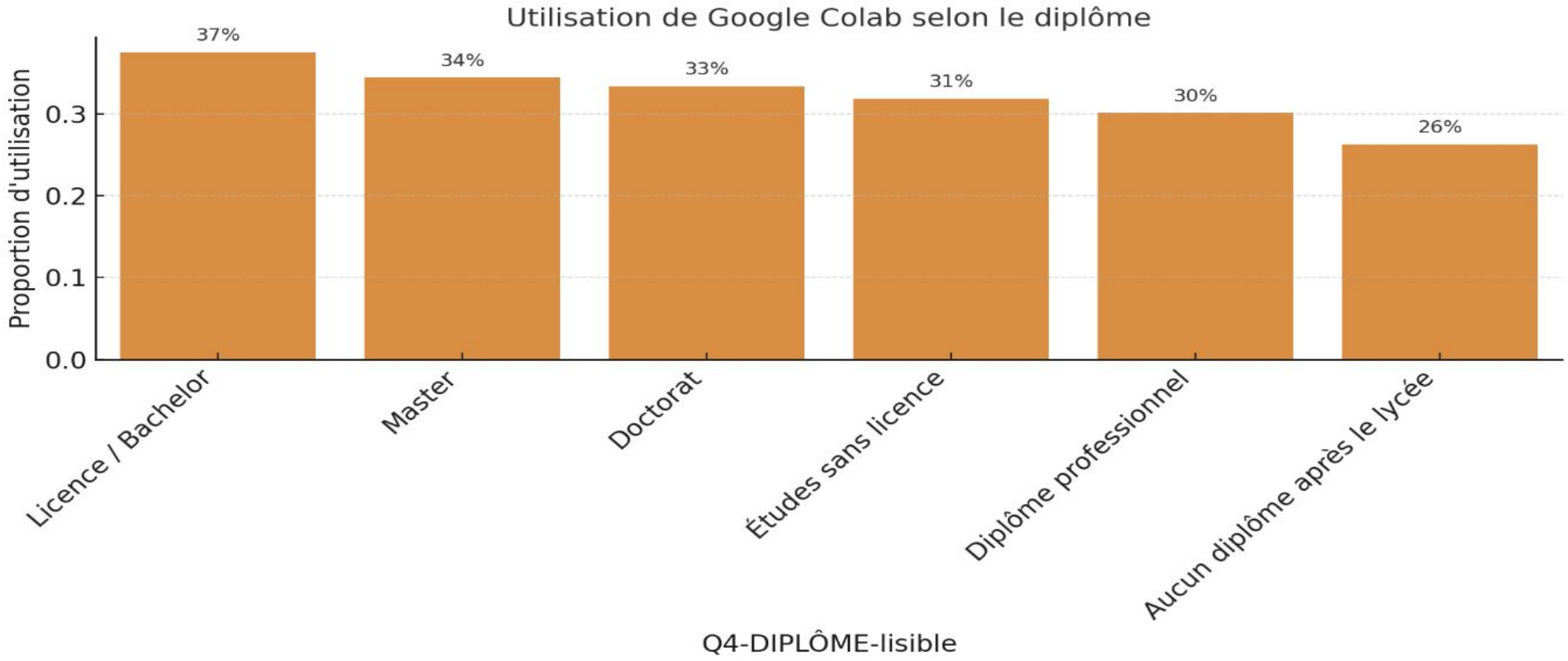


Profil majoritairement intermédiaire en programmation

Analyse bivariée

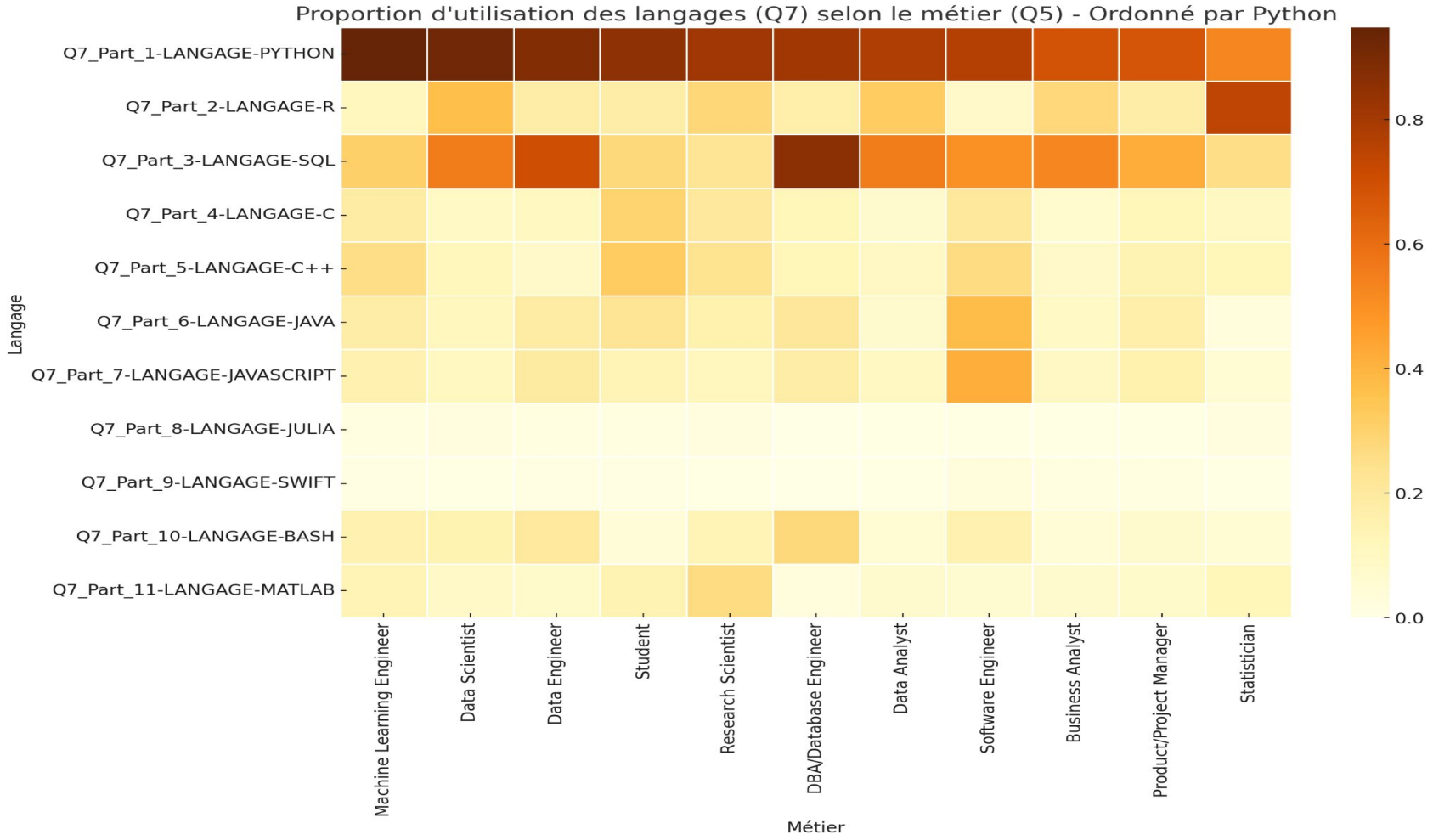


Analyse bivariée



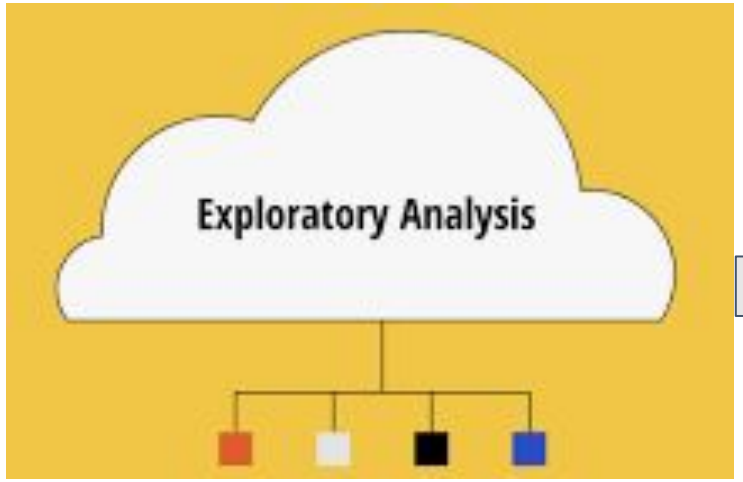
Utilisation de Google Colab homogène selon le niveau d’étude

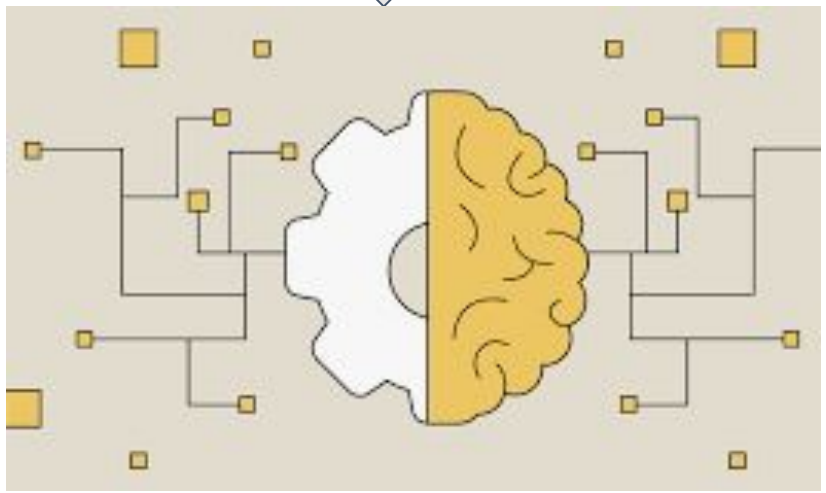
Corrélation



Annabelle et Lydie le 15 juillet 2025

Modélisation





Annabelle et Lydie le 15 juillet 2025

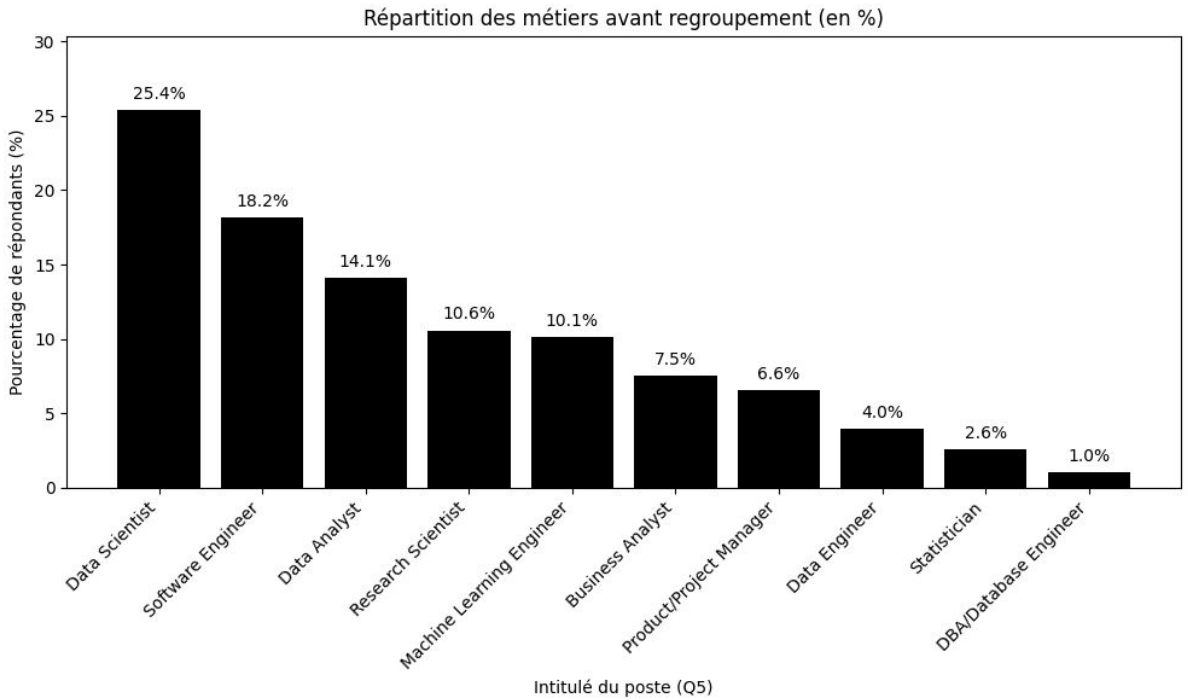
Modélisation

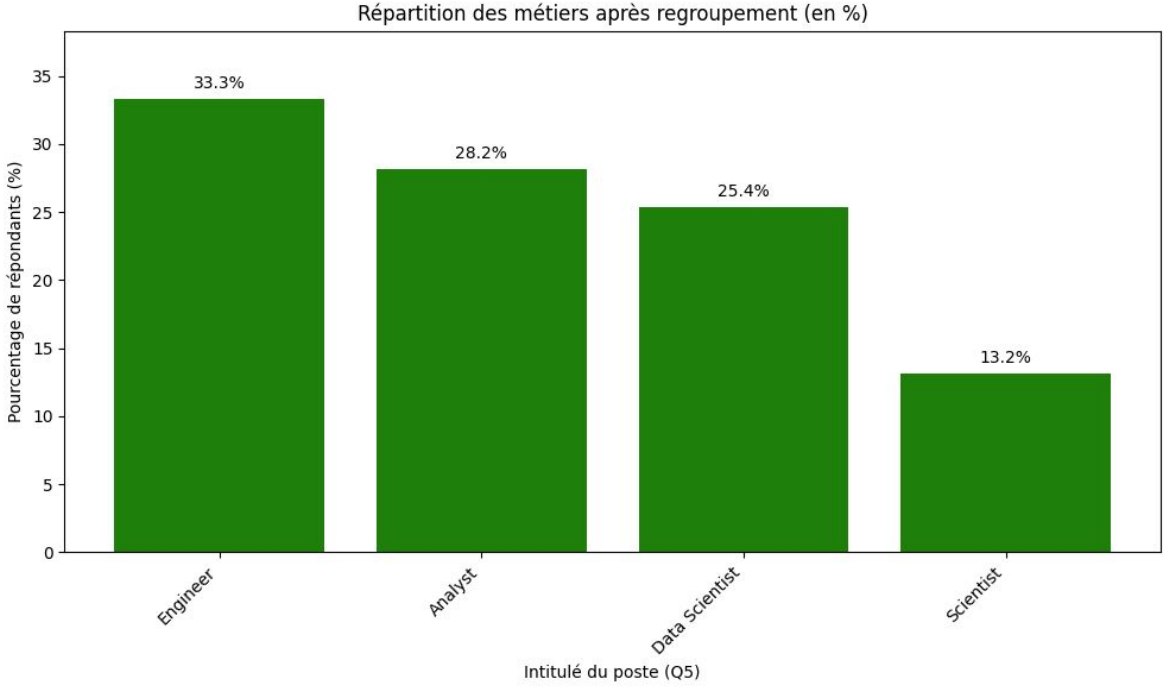
Regroupement et encodage des variables

Variable	Regroupement	Encodage
Q1-Age	18-24/25-34/ et +35 ans	Ordinal encoding (0-2)
Q2-Genre	Man/Woman/Other Gender	One-hot-encoding
Q3-Pays	Continent	One-hot-encoding
Q4-Niveau d'étude	Secondaire au doctorat	Ordinal encoding (0-5)
Q6-Expérience en programmation	Pas d'expérience/1-5ans/+5ans	Ordinal encoding (0-2)
Q15-Expérience en ML	Débutant/Intermédiaire/Avancé	Ordinal encoding (0-2)

Modélisation

Regroupement des métiers





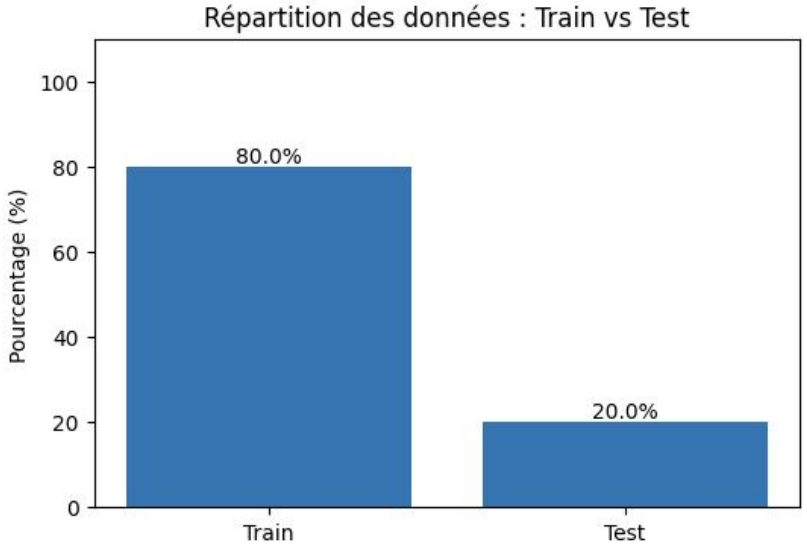
Annabelle et Lydie le 15 juillet 2025

Modélisation

Variable	Encodage
Q5-Métier regroupé	Label encoding

```
Dictionnaire métier regroupé :
Analyst → 0
Data Scientist → 1
Engineer → 2
Scientist → 3
```

Séparation du jeu de données



- Méthode: train_test_split
- Taille: 80% entrainement et 20 % test
- Stratification: basée sur la variable cible
- Random state: 42

Modélisation

Normalisation

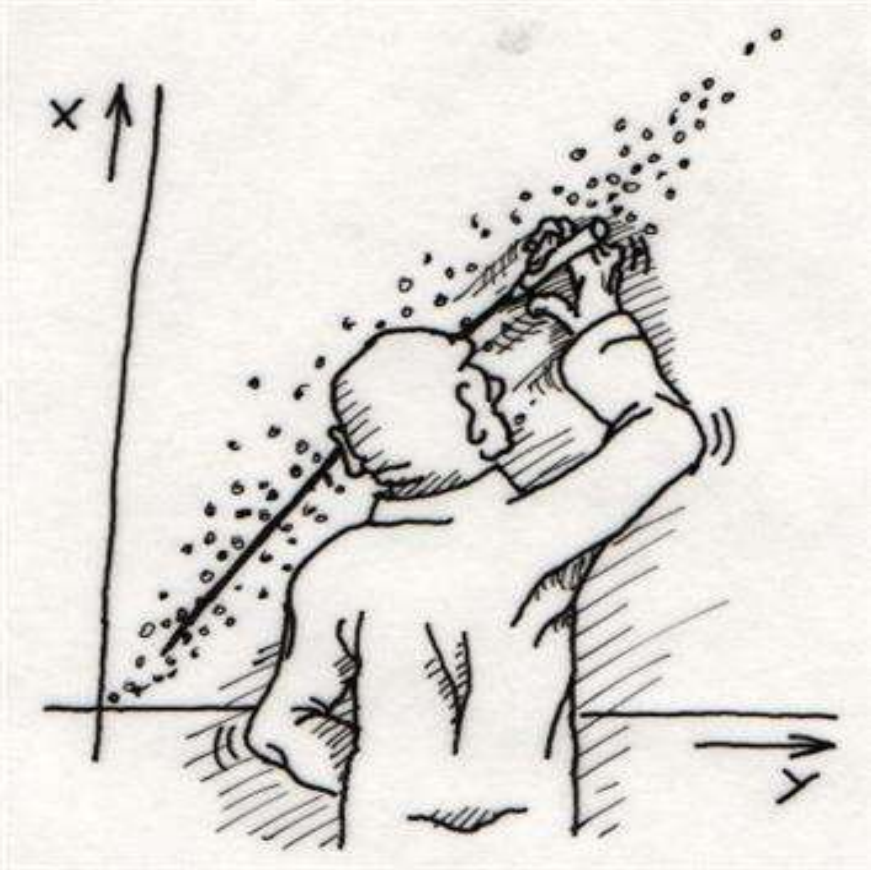
Variable	Type d'encodage
Q1-Age	Ordinal
Q4-Niveau d'étude	Ordinal
Q6-Expérience en programmation	Ordinal
Q15-Expérience en ML	Ordinal

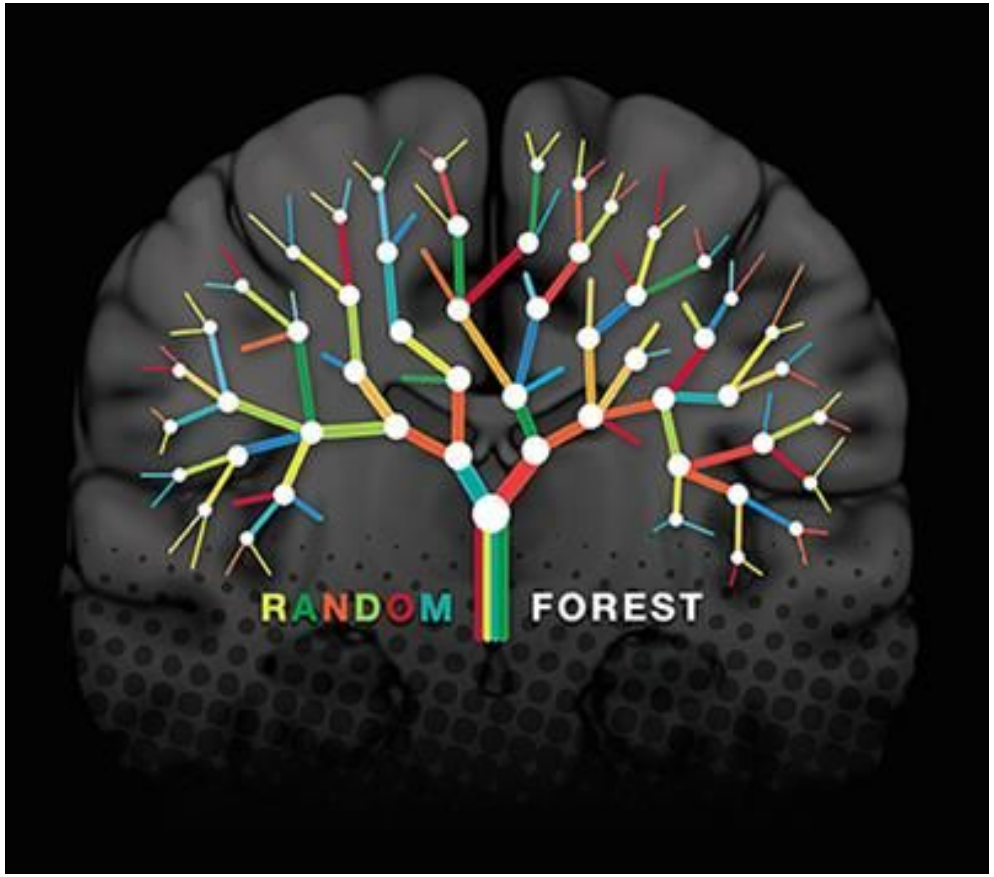
- Méthode: MinMaxScaler (0-1)
- Objectif: harmoniser les échelles pour éviter que certaines variables dominent le modèle.

Q1	Q4	Q6	Q15
1.0	0.6	1.0	1.000000
1.0	1.0	1.0	0.666667
1.0	0.6	1.0	1.000000
0.5	0.4	1.0	0.666667
1.0	0.6	1.0	0.333333

Modélisation

Nous avons testé notre jeu de données avec les algorithmes:

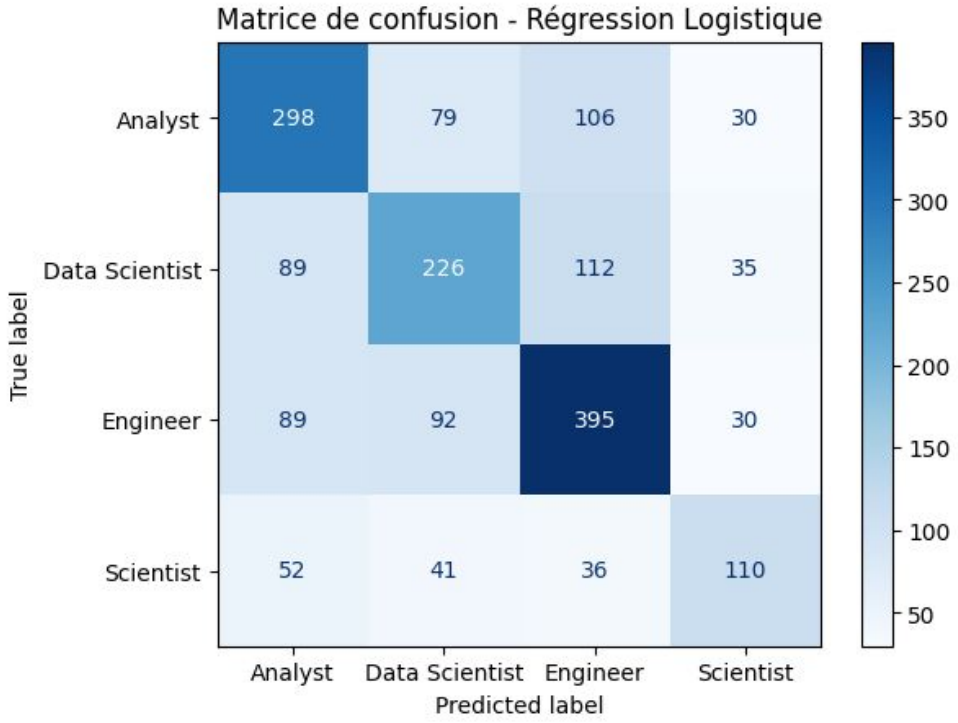


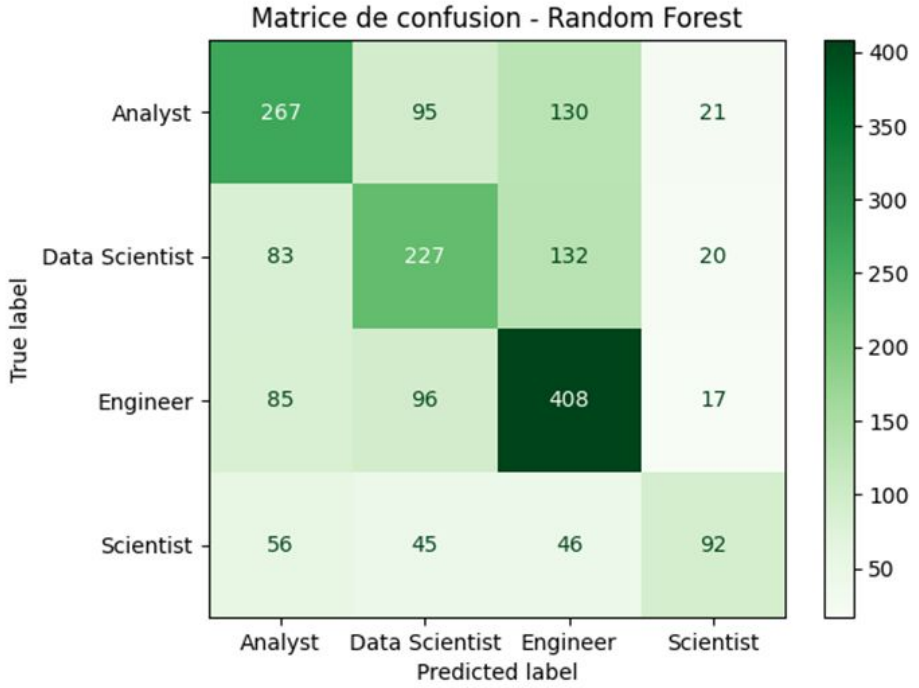


Annabelle et Lydie le 15 juillet 2025

Modélisation

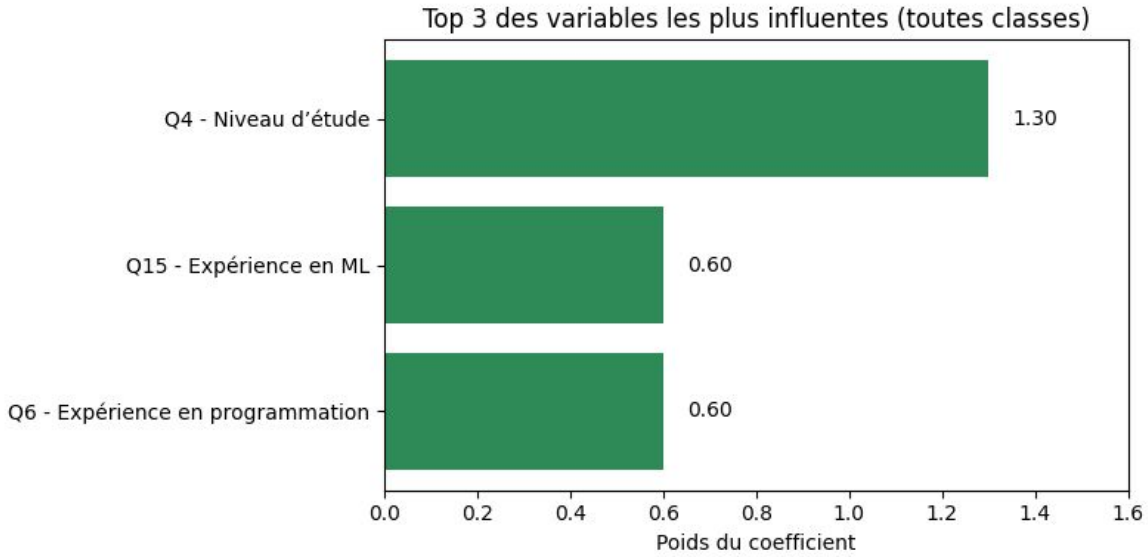
Résultat des modèles





Modélisation

Top 3 des variables les plus influentes



Rang	Variable	Interprétation
1	Q4-Niveau d'étude	Variable la plus déterminante, plus il est élevé plus le répondant sera orienté vers certains rôles
2	Q6-Expérience en programmation	Joue un rôle central surtout pour les rôles techniques
3	Q15-Expérience en ML	Constitue un critère différenciateur important, reflète un niveau de technicité

Interprétation des données

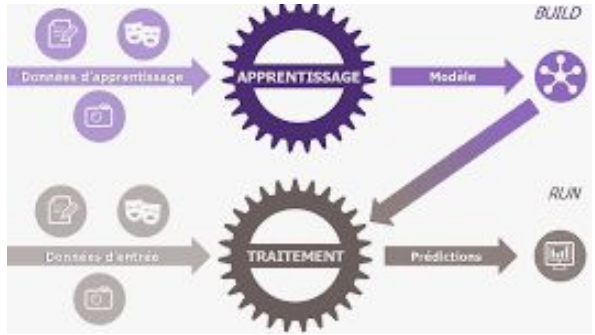
- Profils orientés DATA
- Diversité des niveaux d'expérience
- Utilisation variée des outils et langages de programmation
- Surreprésentation des profils très diplômés
- Auto-évaluation peut-être subjective des compétences

MINES PARIS EXECUTIVE EDUCATION | PSL

DataScientest • com

Conclusion

Étape clé dans notre reconversion

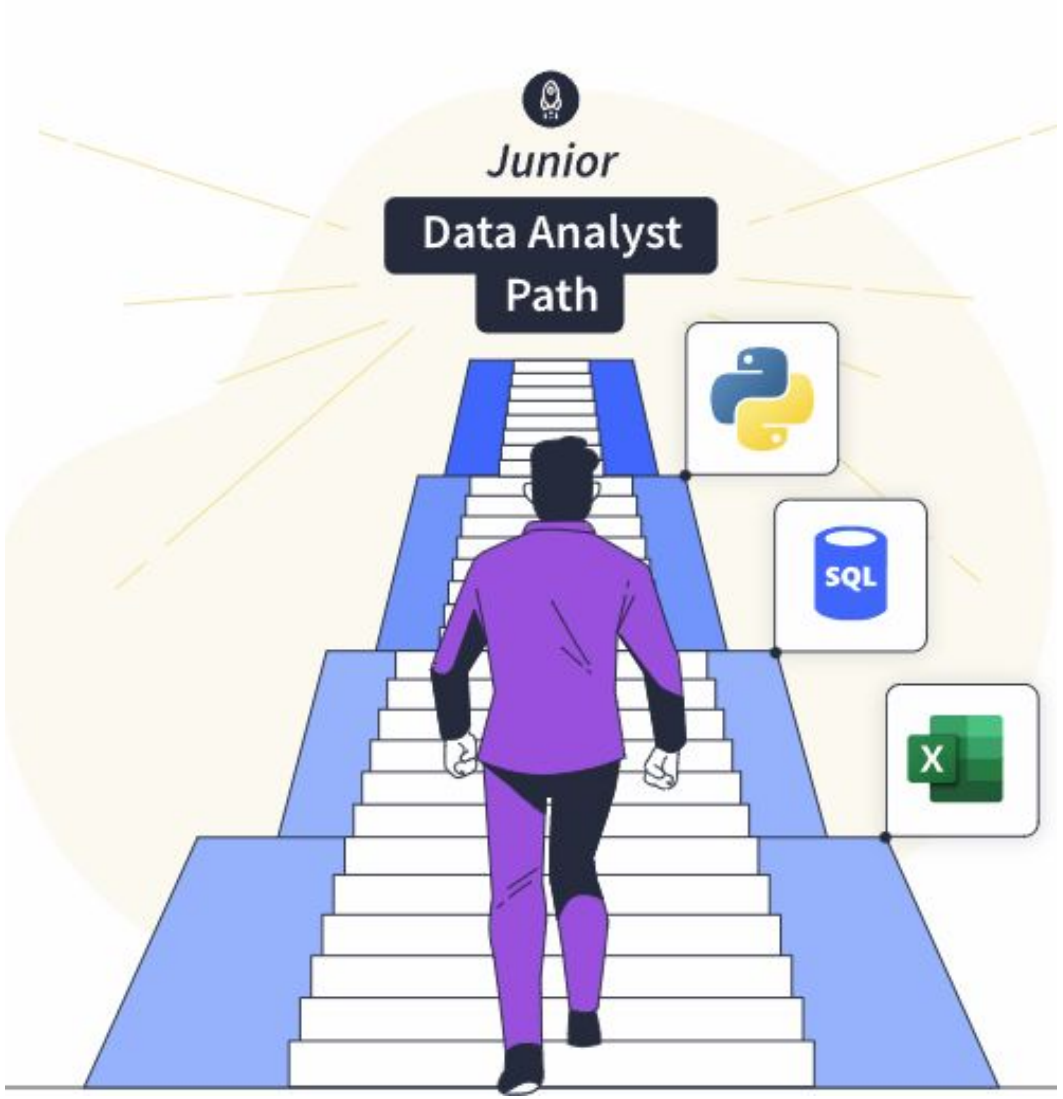


Annabelle et Lydie le 15 juillet 2025

MINES PARIS
EXECUTIVE EDUCATION | PSL

DataScientest • com

MERCI POUR VOTRE ATTENTION



Annabelle et Lydie le 15 juillet 2025