RAPPORT D'ANALYSE DATA

Prédiction d'un Risque d'Abandon Scolaire

Members : OMAR KENGNI

OBJECTIF DU PROJET :

L'objectif principal de ce projet est de prédire l'abandon scolaire des étudiants en utilisant des données historiques. Cela implique :

- Une exploration et une compréhension approfondie des données.
- La préparation des données pour la modélisation.
- L'identification des métriques clés pour évaluer les performances des modèles.
- La mise en place d'un pipeline reproductible pour la production.

Nous avons chargé le dataset `Dataset_Abandon_Scolaire.csv` et effectué une exploration initiale pour comprendre sa structure. Nous allons donc présenter les étapes qui ont suivi ce process après observation et compréhension à l'œil humain des données.

Exploration des données

Résumé des données

- **Dimensions** : 2000 lignes et 7 colonnes.

- **Colonnes principales** :

- `Age` : Âge des étudiants.

- `Sexe` : Sexe des étudiants (Homme/Femme).

- `Taux_presence` : Pourcentage de présence en classe.

- `Nombre_retards` : Nombre de retards enregistrés.

- `Note_moyenne` : Moyenne des notes obtenues.

- `Situation_familiale` : Situation familiale (Célibataire, Marié, etc.).

- `Abandon` : Variable cible (1 = abandon, 0 = non-abandon).

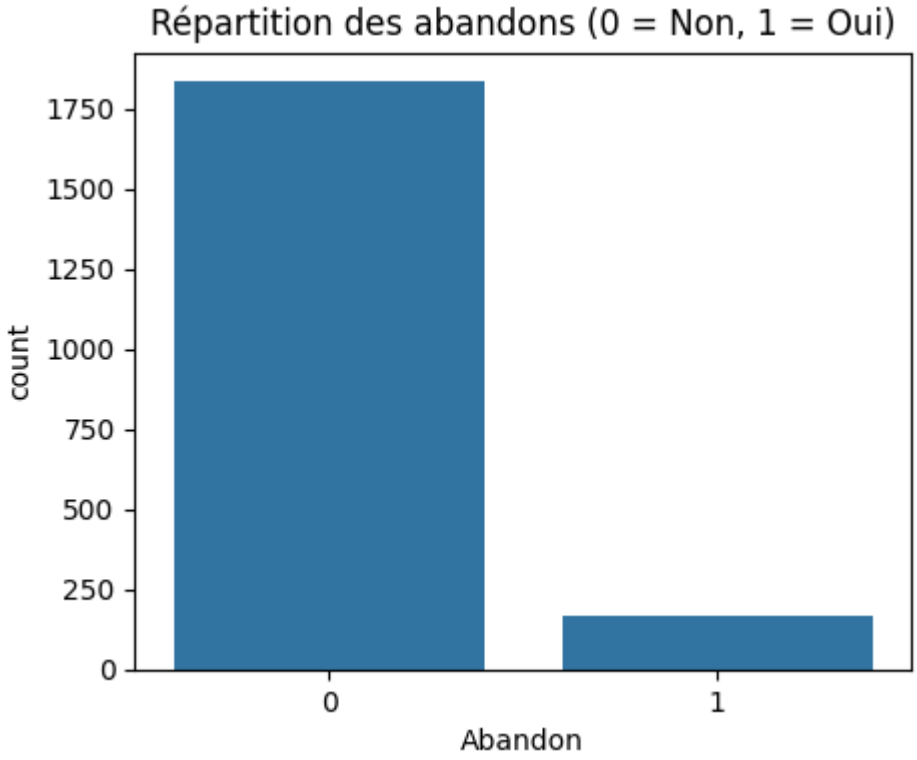
Valeurs manquantes : Aucune valeur manquante trouvé dans les données.

Déséquilibre des classes :

- Le dataset est déséquilibré : les cas d'abandon scolaire (Abandon = 1) sont nettement minoritaires.

- Sur 2000 observations, seule une petite fraction représente des abandons, ce qui risque de complexifier l'apprentissage du modèle.

- De plus, le faible volume total de données limite la robustesse de certaines méthodes d'analyse.



Statistiques descriptives : Les statistiques descriptives montrent des variations importantes dans les variables numériques, notamment :

- `Age` varie entre 18 et 25 ans.

- `Taux_presence` varie entre 50% et 100%.

- `Nombre_retards` varie entre 0 et 10.

- `Note_moyenne` varie entre 0 et 20.

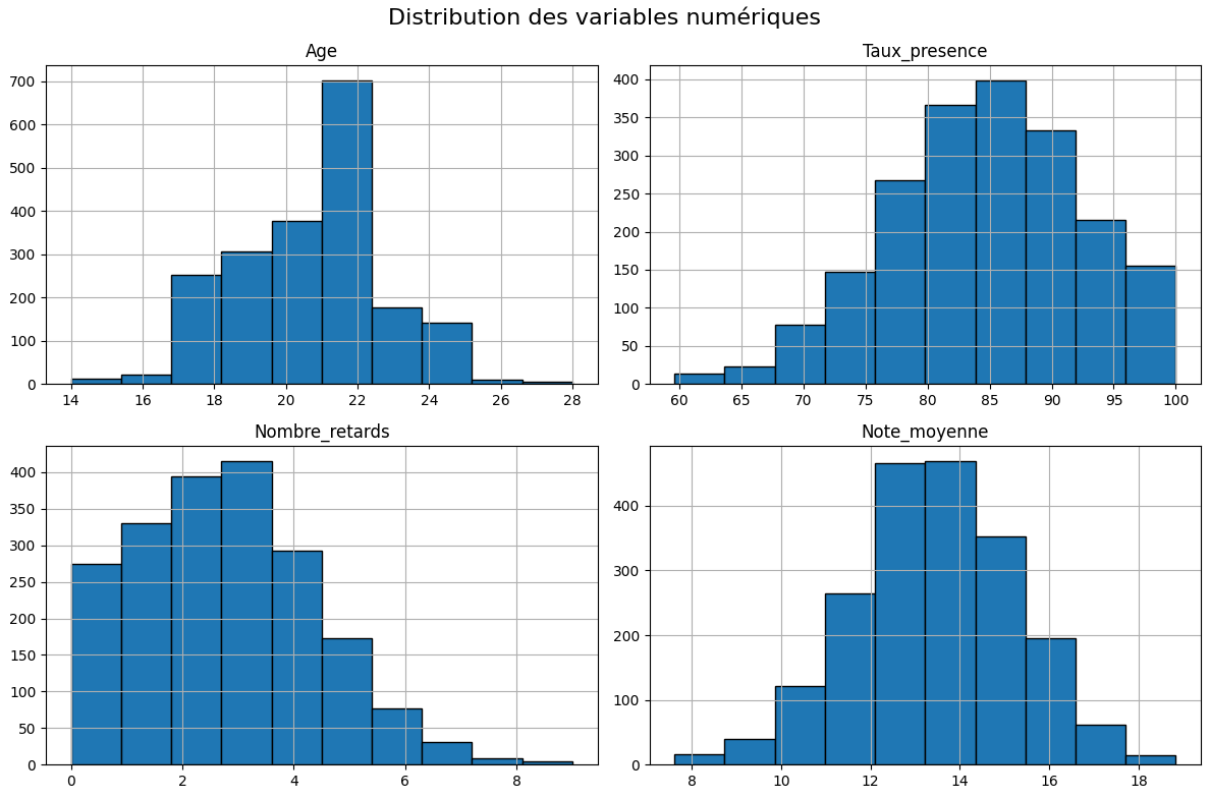
Visualisation des données

Des visualisations ont été conçus pour mieux comprendre la répartition des variables et leurs relations.

Répartition des variables numériques :

Les histogrammes montrent que :

- **Âge :** Majorité concentrée entre 20 et 22 ans, ce qui est cohérent avec une population étudiante.
- **Taux de présence :** Fortement concentré entre 80 et 100 %, ce qui traduit un bon suivi global.
- **Nombre de retards :** Faible pour la majorité des étudiants (0 à 3 retards).
- **Note moyenne :** Centrées autour de 13, avec une répartition relativement normale.

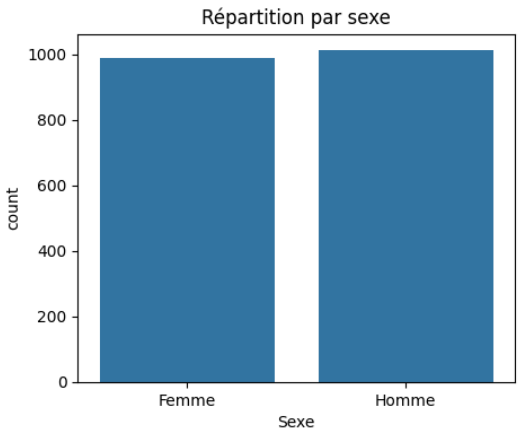


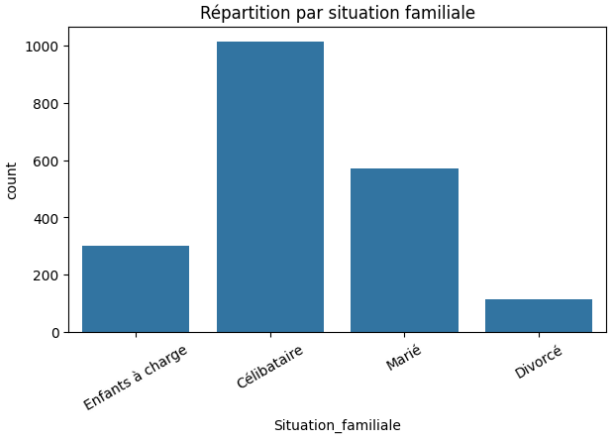
Répartition par sexe et situation familiale :

Sexe : répartition parfaitement équilibrée entre hommes et femmes.

Situation familiale :

- Une majorité d'étudiants sont célibataires.
- Les mariés viennent en second, puis les étudiants avec enfants à charge.
- Très peu de divorcés.



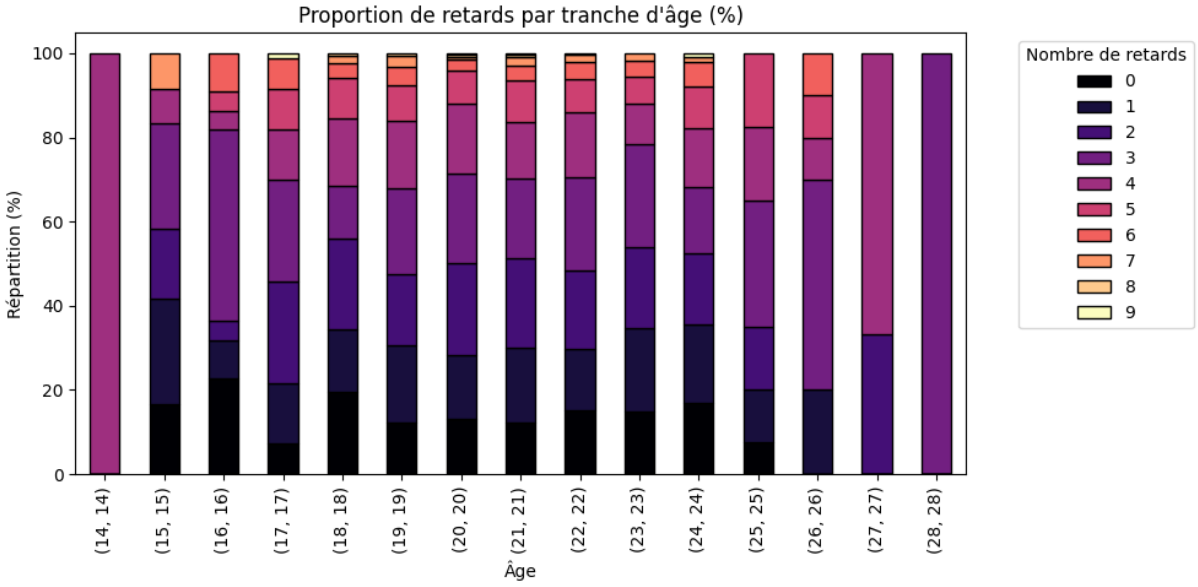


Relations entre variables et abandon scolaire

Les histogrammes croisés avec Abandon apportent des éléments de réponse utiles :

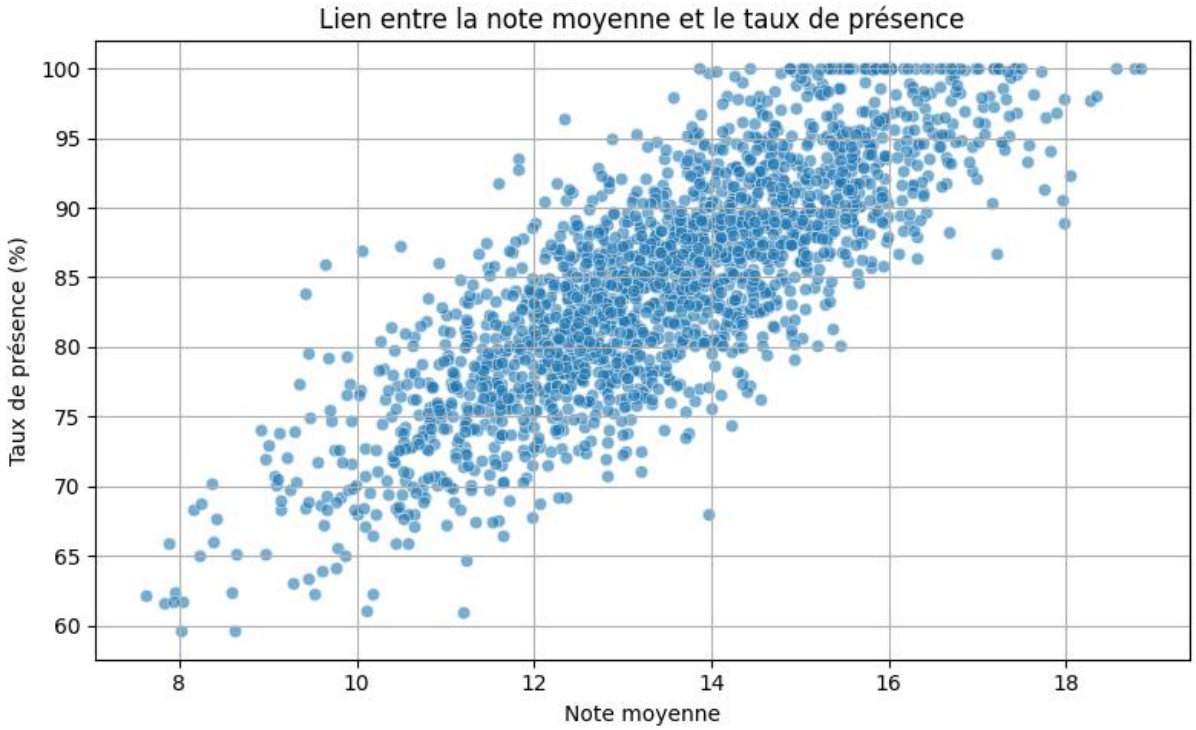
Âge vs Abandon :

L'abandon est légèrement plus fréquent chez les plus jeunes, en particulier autour de 18–19 ans.



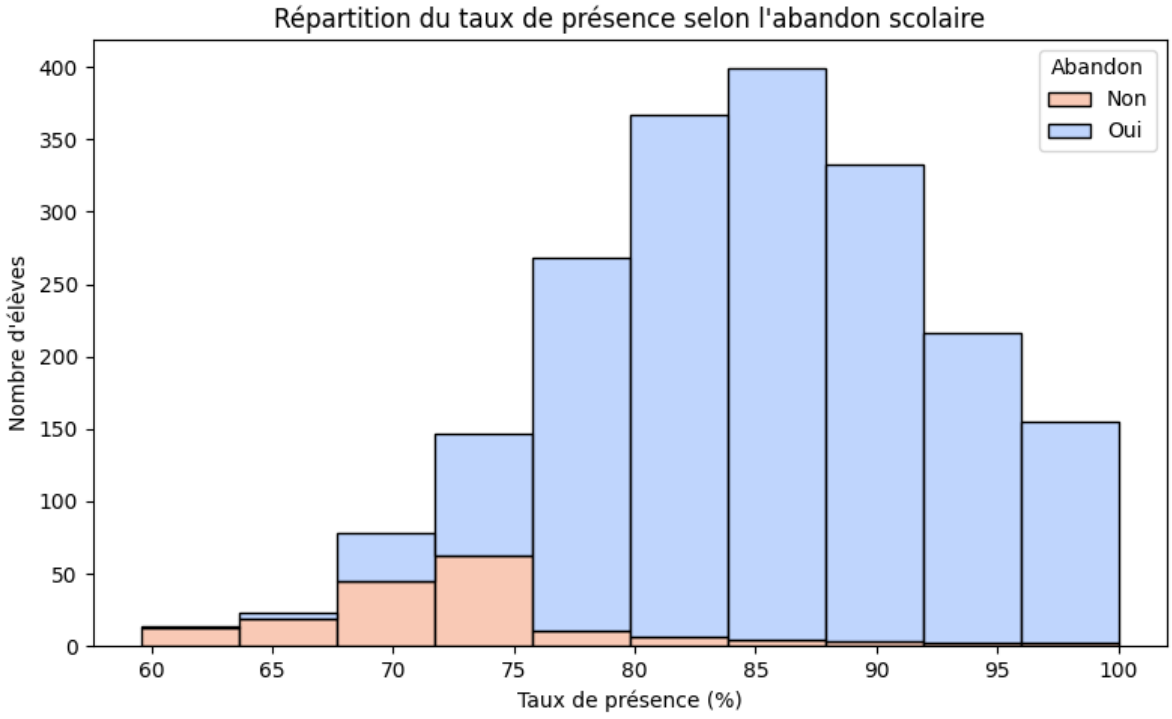
Note moyenne vs Abandon :

Les élèves en échec scolaire (notes < 11) ont une probabilité d'abandon nettement plus élevée.



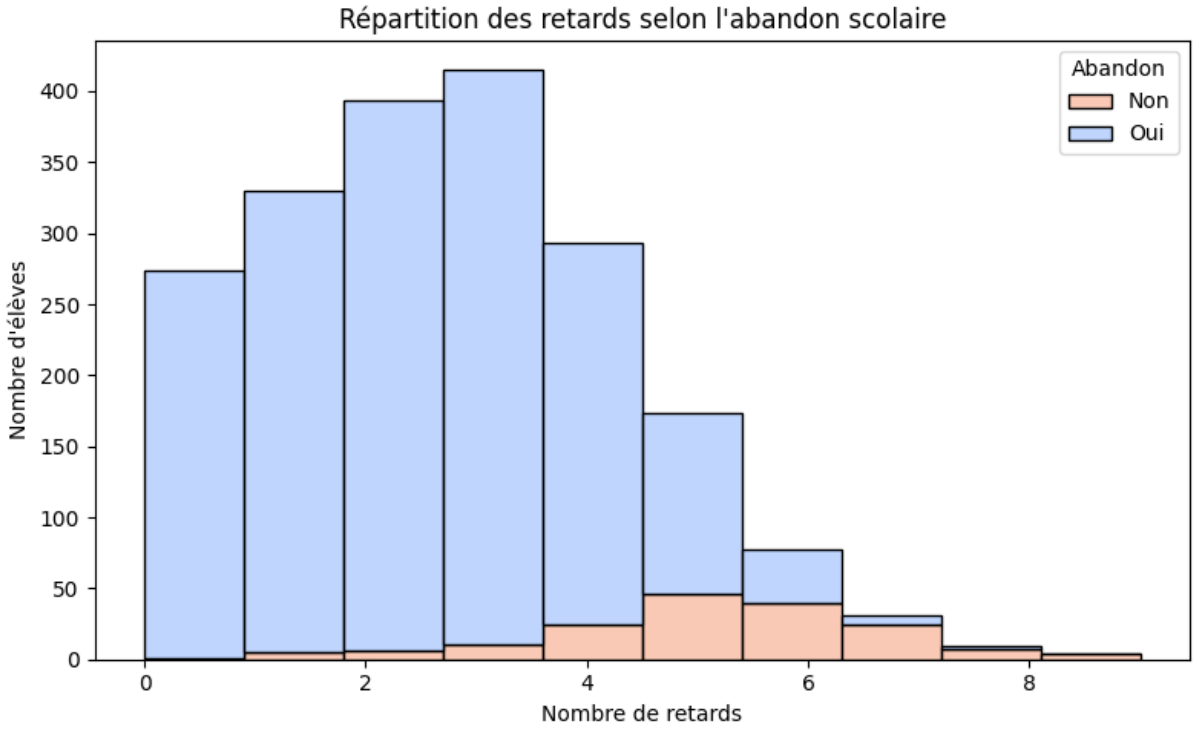
Présence vs Abandon :

Le taux d'abandon augmente fortement en dessous de 80 % de présence.



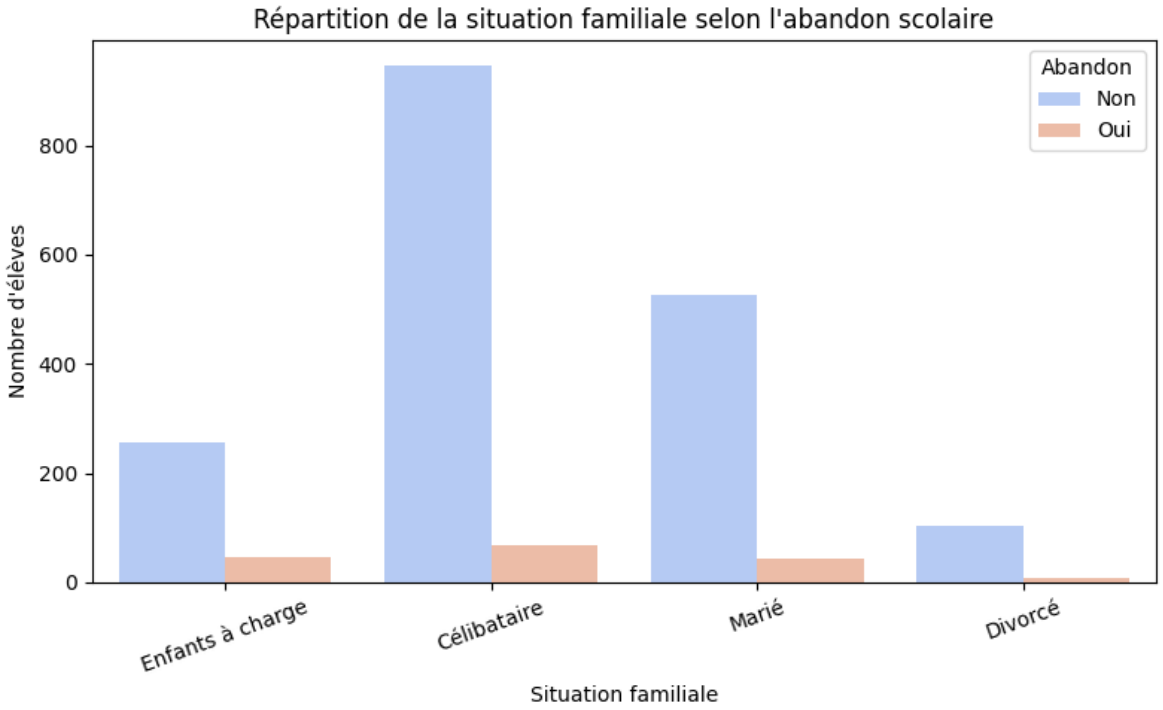
Retards vs Abandon :

Les élèves ayant plus de 3 retards sont largement surreprésentés dans la classe des abandonneurs.



Situation familiale vs Abandon :

Les élèves célibataires sont les plus nombreux parmi les abandonneurs, mais ce constat est à relativiser car ils sont également très majoritaires dans la population globale.



Conclusion

- Les données sont globalement propres et ne demandent pas de traitement particulier.
- On remarque déjà des tendances intéressantes : les élèves qui ont plus de retards, une faible présence ou de mauvaises notes semblent plus à risque d'abandonner.
- Le jeu de données est déséquilibré, avec peu de cas d'abandon. Il faudra donc adapter les modèles pour éviter qu'ils favorisent trop la majorité.
- Les visualisations nous donnent de bonnes pistes, mais on verra plus loin si ces premières impressions se confirment avec les modèles et les matrices de confusion.

Préparation des données

Nettoyage & Preparation

Plusieurs étapes ont été réalisées pour rendre les données exploitables par les algorithmes de Machine Learning :

- Encodage de la variable Sexe :
  - Nettoyage des valeurs (suppression des espaces, capitalisation).
  - Transformation en valeur numérique : Femme = 0, Homme = 1.
- Encodage de la variable Situation_familiale :
  - Application d'un encodage One-Hot sans suppression de catégorie (drop_first=False), ce qui permet de garder toutes les modalités (Célibataire, Marié, Divorcé, Enfants à charge).

Standardisation :

Certaines variables ont des échelles très différentes :

- Taux_presence va de 60 à 100,
- Age est autour de 20,
- Nombre_retards est souvent très bas.

Ces écarts peuvent fausser les modèles comme KNN ou PCA.

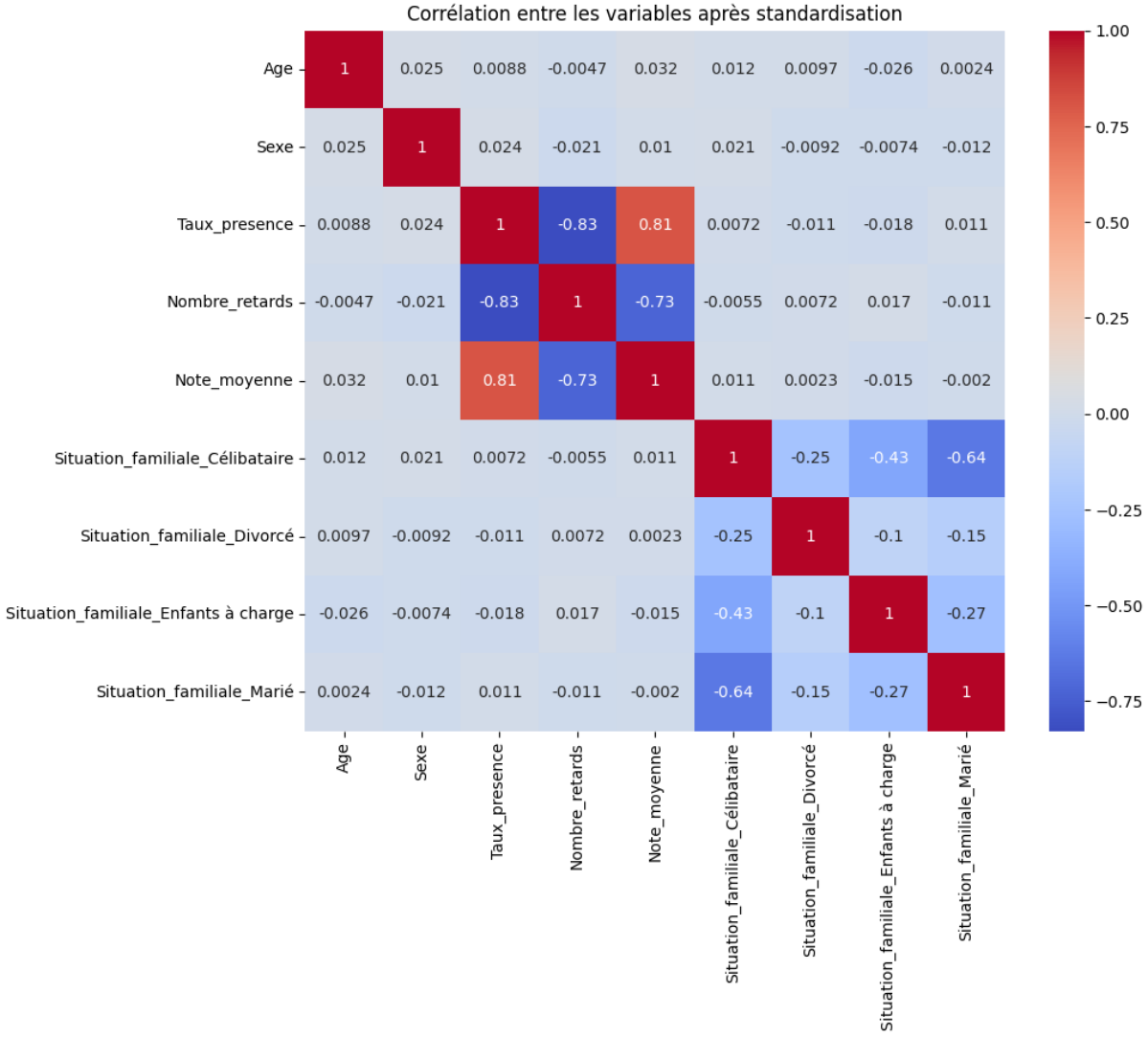
Les données ont donc été standardisées avec StandardScaler, pour obtenir :

- Une moyenne de 0,
- Un écart-type de 1 pour chaque variable.

Corrélation entre variables

Après standardisation, une matrice de corrélation a été calculée. Quelques observations :

- Taux_presence et Note_moyenne sont positivement corrélés (+0.81) : plus un élève est présent, meilleures sont ses notes.
- Nombre_retards est négativement corrélé à Taux_presence (-0.83) et à Note_moyenne (-0.73) : plus on est en retard, moins on est présent, et plus les notes sont faibles.
- Les variables liées à la situation familiale présentent de légères corrélations avec les notes ou la présence, mais rien de très fort.



Séparation des données

Pourquoi séparer les données ?

- Pour évaluer les performances des modèles sur des données "jamais vues".
- Pour respecter la répartition des classes (stratify=y).

Split choisi :

- 70% pour l'entraînement (X_train) → 1400 lignes.
- 30% pour le test (X_test) → 600 lignes.
- Répartition conservée entre abandons (1) et non-abandons (0).

Équilibrage avec SMOTE

Le jeu de données est déséquilibré : seulement ~15% d'élèves ont abandonné.

Pour améliorer l'apprentissage, on a utilisé SMOTE (Synthetic Minority Over-sampling Technique), une technique qui génère artificiellement des exemples de la classe minoritaire (les abandons).

Résultats

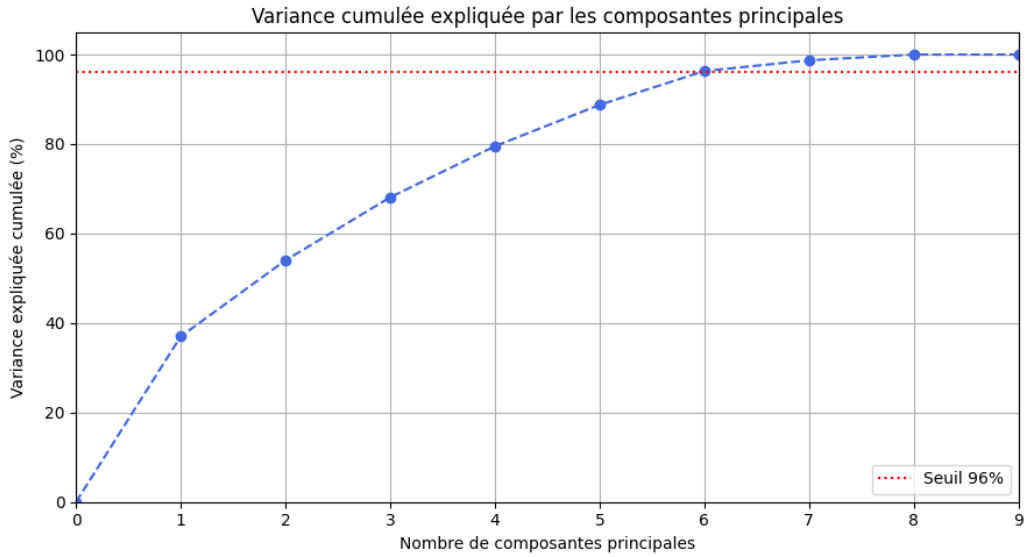
- X_train_balanced : 2564 échantillons équilibrés (50% abandon / 50% non-abandon),
- Les dimensions des données sont conservées (9 variables explicatives).

Analyse en Composantes Principales (PCA)

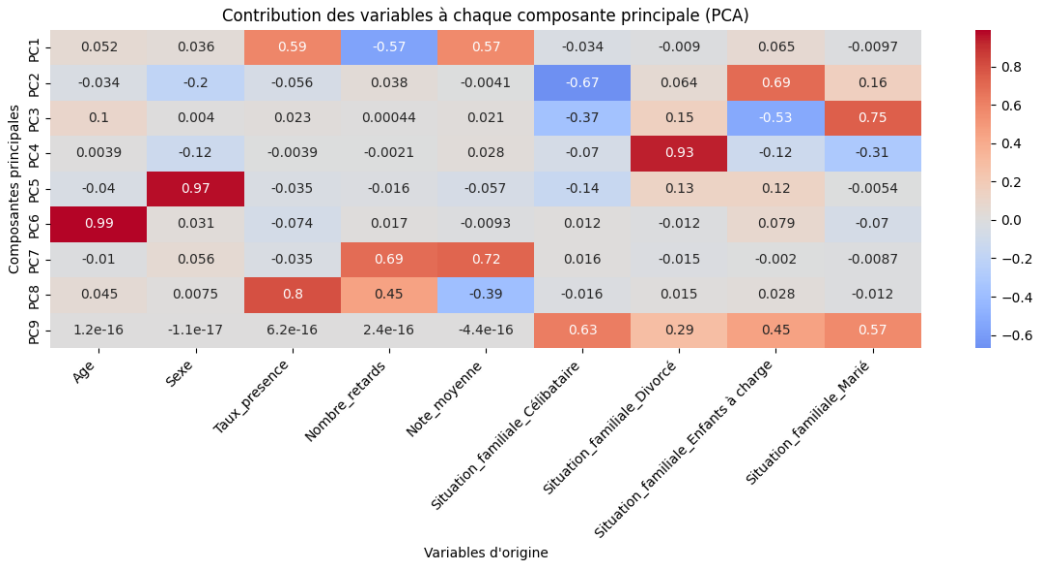
L'objectif de la PCA est de réduire le nombre de calcul tout en gardant un maximum d'information, ce qui facilite la visualisation et le coût utilisé.

Résultats principaux :

- Les 6 premières composantes principales permettent de conserver plus de 96 % de l'information.



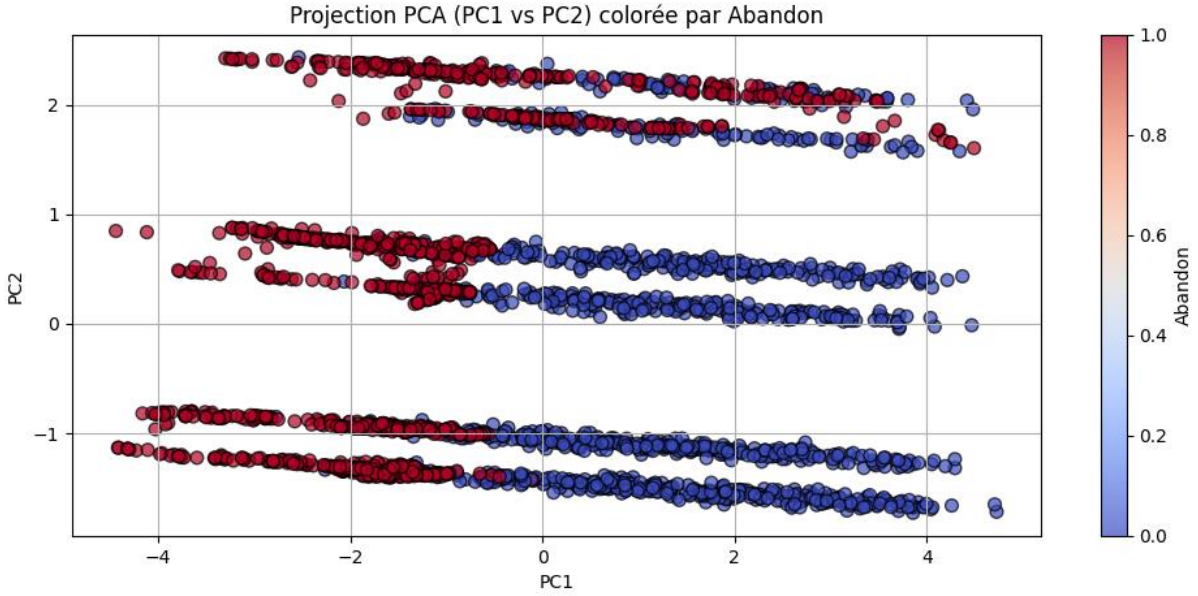
- À partir de la PC6, l'apport en information devient négligeable.
- La PC1 est surtout liée à des variables académiques comme le taux de présence, les retards et les notes moyennes.
- La PC2, elle, semble plutôt influencée par la situation familiale.



Projection dans l'espace réduit :

- La projection des élèves dans le plan PC1 vs PC2 permet de visualiser certains regroupements.
- En colorant les points selon l'abandon, on observe une légère séparation, mais pas de frontières claires entre les classes.

- Cela confirme que les relations sont non linéaires, et que la PCA seule ne suffit pas à prédire les abandons.



Méthode du coude (Elbow Method) et clustering K-Means

Objectif :

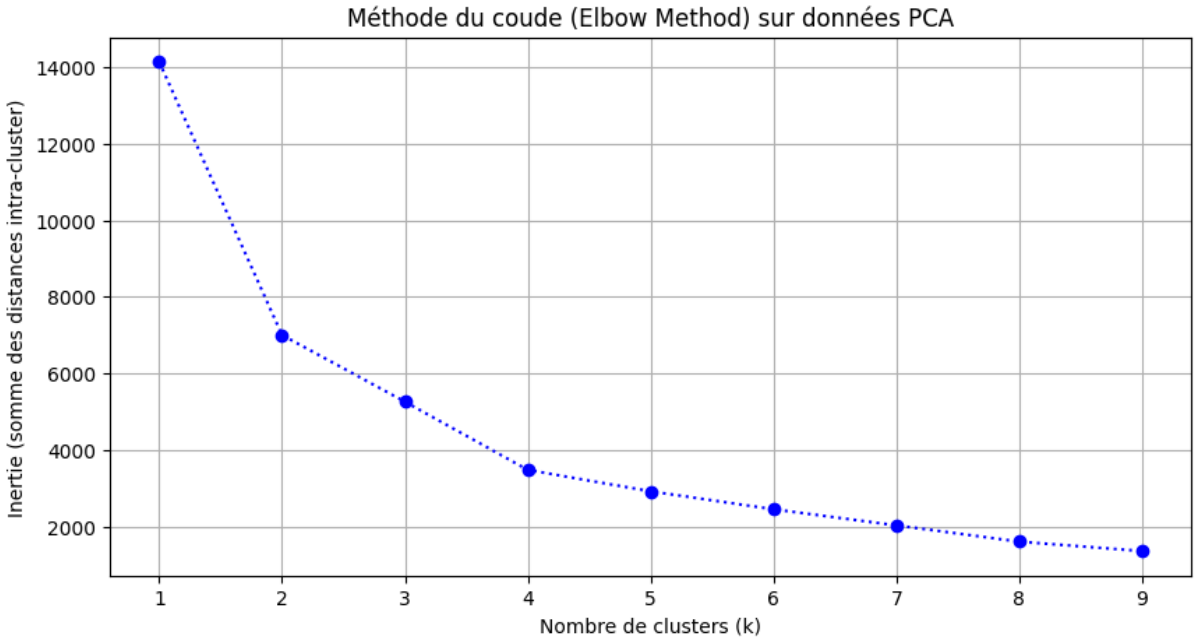
L'objectif de cette étape est de segmenter les étudiants selon leurs caractéristiques (présence, notes, situation, etc.) à l'aide d'un algorithme non supervisé : K-Means.

Pour cela, nous devons d'abord déterminer le nombre optimal de groupes (clusters).

Méthode du coude (Elbow Method) :

Nous avons appliqué la méthode du coude sur deux jeux de données :

- Jeu original (avant SMOTE) : L'inflexion est visible autour de k = 4, ce qui suggère l'existence de 4 groupes naturels.
- Jeu augmenté (après SMOTE) : L'inflexion semble moins marquée et suggère plutôt k = 2, sans pour autant apporter une meilleure lecture des segments.

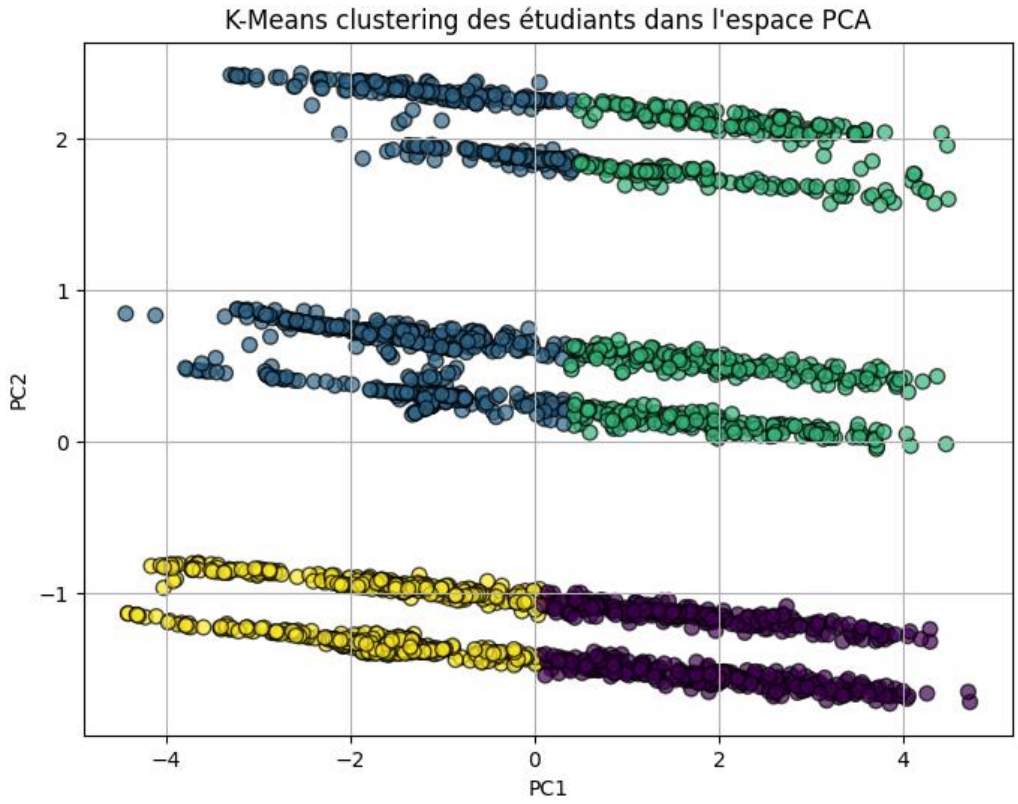


Choix final : Nous avons conservé k = 4 pour rester cohérent avec les observations sur les données d'origine, plus proches de la réalité terrain (non artificiellement équilibrées).

Clustering K-Means :

L'algorithme K-Means a été appliqué sur les données projetées via PCA (PC1 et PC2).

On observe bien une répartition des étudiants en 4 groupes distincts dans l'espace réduit.



Interprétation :

L'analyse dans le plan PCA (PC1 vs PC2) permet de visualiser la répartition des étudiants selon deux axes principaux :

- PC1 (abscisse) : reflète principalement les facteurs académiques → note moyenne, taux de présence et nombre de retards.
- PC2 (ordonnée) : traduit surtout les caractéristiques socio-familiales → notamment la situation familiale.

Que nous apprend cette projection ?

- On observe une séparation verticale (selon l'axe PC2), ce qui indique que la situation familiale joue un rôle important dans la segmentation des profils.
- Horizontalement (axe PC1), les étudiants se différencient surtout par leur implication scolaire (présence, retards, résultats).

En résumé, le clustering fait apparaître des profils types d'élèves où la combinaison de la situation personnelle et des comportements scolaires semble avoir un lien avec le risque d'abandon.

Ces premiers indices seront confrontés à la modélisation supervisée dans les prochaines étapes.

Métriques : recall ou F1-score ?

Pourquoi comparer deux métriques ?

Dans un premier temps, notre objectif était de maximiser le rappel (recall), c'est-à-dire repérer un maximum d'élèves à risque d'abandon. Cette approche avait du sens sur le plan humain, car elle permettait de ne "rater" que très peu d'élèves concernés.

Cependant, cela avait un gros inconvénient : trop d'élèves non concernés par l'abandon étaient faussement identifiés comme à risque. Cela aurait généré des coûts inutiles si une intervention était déclenchée à chaque alerte.

Nous avons donc préféré utiliser le F1-score comme métrique principale. Ce compromis entre précision (éviter les faux positifs coûteux) et rappel (ne pas louper les vrais abandons) permet d'optimiser les ressources tout en maintenant une détection correcte.

À noter : Ce problème vient en grande partie de la taille réduite de notre dataset (2000 observations), qui rend l'apprentissage plus délicat pour des classes déséquilibrées.

Résultats avec Recall comme métrique

Modèle	Recall Classe 1	Précision Classe 1	F1 Classe 1	ROC AUC
KNN (k=17)	0.90	0.33	0.48	0.93
Decision Tree	0.90	0.36	0.51	0.88
Random Forest	0.88	0.34	0.49	0.93
XGBoost	0.92	0.34	0.50	0.94

Très bon rappel, mais précision très faible (~0.33–0.36). En clair on détecte bien les élèves à risque, mais avec beaucoup on se trompe beaucoup également.

Résultats avec F1-score comme métrique

Modèle	Recall Classe 1	Précision Classe 1	F1 Classe 1	ROC AUC
KNN (k=1)	0.70	0.44	0.54	0.81
Decision Tree	0.78	0.48	0.60	0.87
Random Forest	0.76	0.49	0.59	0.93

XGBoost	0.72	0.47	0.57	0.94

La précision est meilleure et onse trompe moins ça réduit les coûts, tout en conservant une bonne capacité à détecter les abandons.

Conclusion

Après comparaison des différents modèles en utilisant le F1-score comme critère principal, nous avons retenu le modèle XGBoost comme le plus adapté à notre problématique.

Pourquoi XGBoost ?

- Il offre le meilleur compromis entre précision et rappel.
- Il génère moins de faux positifs que les autres modèles tout en maintenant un bon niveau de détection des cas réels d'abandon.
- Son F1-score pour la classe "abandon" atteint 0.57, ce qui est le plus haut parmi les modèles testés avec F1 comme objectif.

Ce que ça signifie concrètement

Le modèle XGBoost permet de :

- Identifier environ 72 % des élèves qui vont réellement abandonner (recall).
- Prédire un abandon à bon escient dans 47 % des cas (précision).
- Avec une précision globale de 91 % sur l'ensemble des élèves, il reste fiable sur l'ensemble des classes.

En résumé : le modèle réussit à détecter une bonne partie des élèves en difficulté, tout en limitant les alertes inutiles, ce qui est essentiel pour réduire les coûts d'intervention, sans pour autant compromettre l'objectif principal qui est d'aider le plus de personne possible à ne pas abandonner leur étude.

Limites rencontrées

- La performance est fortement impactée par la taille réduite du jeu de données (2000 élèves).
- Le déséquilibre entre élèves ayant abandonné (15%) et les autres rend l'apprentissage plus instable, même avec du SMOTE.

Même si la data augmentation (SMOTE) a permis d'équilibrer le dataset, elle ne remplace pas des données réelles :

- Elle conçoit des données artificielles basées.
- Elle peut masquer certaines structures naturelles des données, et rendre les modèles moins généralisables.

Pour aller plus loin

- Il serait judicieux de collecter davantage de données réelles, notamment sur les profils à risque.
- Des variables supplémentaires (problèmes financiers, distance domicile-école, etc.) pourraient améliorer la prédiction.

En conclusion, malgré les limites, le modèle retenu permet de poser une première base solide de détection préventive des abandons scolaires. Il faudra cependant renforcer les données pour obtenir une solution vraiment performante et durable.

MERCI !!