

Mémoire de fin d'études en vue de l'obtention du Certificat Data Analyst

Intitulé

Analyse et prédition de l'abandon du e-learning dans les plateformes numériques

Présenté le :

30/10/2025

Par :

Mme Héla Zine

Sous la direction de :

M. Samy Wahbi

M. Reda Souni

Promotion

Mai 2025

Table des matières

Introduction.....	3
Objectif	3
Méthodologie.....	4
I. Analyse générale	5
1. Profil démographique	5
2. Niveau académique.....	5
3. Performance académique (score moyen)	5
4. Engagement (crédits moyens)	6
II. Statistiques démographiques pour l'abandon.....	6
1. Sexe et handicap	6
2. Tranche d'âge	7
3. Niveau académique.....	7
4. Répartition géographique	7
III. Statistiques comportementales pour l'abandon	8
1. Nombre d'abandons par état d'inscription.....	8
2. Nombre d'abandons selon les tentatives d'inscription	8
3. Performance académique en fonction des tentatives.....	9
4. Corrélation entre l'activité sur la plateforme et le résultat final	9
IV. Statistiques de performance	10
1. Analyse du score moyen selon la performance académique	10
2. Analyse du crédit moyen selon la performance académique	10
V. Abandon et prédition	11
VI. Recommandations	12
1. Étudiants à haut risque d'abandon	12
2. Étudiants faiblement engagés	12
3. Étudiants en difficulté académique.....	12
4. Étudiants récurrents ou avec tentatives multiples	12
5. Étudiants performants ou très engagés	13
Conclusion	13
Perspectives	14
Amélioration des modèles prédictifs	14
Enrichissement des données	14
Déploiement opérationnel	14
Perspectives pédagogiques.....	14
Référence bibliographique	15
Sources	15
Liste des figures.....	15

Introduction

En 2024, 860 millions de personnes dans le monde étaient inscrites à des cours en ligne¹. Derrière ce chiffre impressionnant se cache un paradoxe : la majorité des apprenants abandonnent avant d’atteindre leurs objectifs. Certaines études montrent que 52 % des inscrits ne consultent même jamais le matériel pédagogique, et que le taux moyen d’abandon atteint 90 % sur cinq ans².

Pourquoi un phénomène si massif ? L’apprentissage en ligne représente un nouvel environnement pour beaucoup, et l’adaptation peut être difficile. Une plateforme perçue comme complexe, des cours mal calibrés — trop longs, trop théoriques, trop faciles ou trop difficiles —, des problèmes techniques, une mauvaise gestion du temps ou encore le manque d’interaction humaine sont autant de facteurs favorisant le décrochage.

Face à ce constat, comprendre les mécanismes de l’abandon et pouvoir le prédire devient un enjeu majeur pour améliorer l’efficacité des formations en ligne.

Objectif

Ce projet complet de data-analyse, allant de la préparation des données à la modélisation prédictive, se concentre sur l’analyse et la prédiction de l’abandon des apprenants sur les plateformes numériques, à partir des données issues de *l’Open University Learning Analytics Dataset (OULAD)*.

L’objectif principal est de mieux comprendre les facteurs qui influencent le décrochage et de développer un modèle prédictif permettant d’identifier les apprenants à risque. Les résultats de cette étude visent à fournir des recommandations concrètes aux plateformes de formation afin de réduire l’abandon, améliorer l’engagement et soutenir la réussite des apprenants.

Le choix de ce sujet est justifié par sa portée transversale :

Santé / Pharma : abandon d’apprenants dans le e-learning médical, en parallèle avec la problématique de non-observance des patients.

E-commerce / Retail : abandon de panier ou d’abonnement.

Télécom / Tech : churn d’abonnés à un forfait ou une application.

Banque / Assurance : désengagement d’un client (fermeture de compte, résiliation).

Énergie : perte d’abonnés à une offre d’électricité ou de gaz.

Ainsi, bien que centré sur l'apprentissage en ligne, ce projet s'inscrit dans une perspective plus large de compréhension et de réduction du désengagement dans divers secteurs.

Méthodologie

La démarche adoptée repose sur plusieurs étapes complémentaires :

1. Choix du dataset
 - Les données proviennent de l'*Open University Learning Analytics Dataset* (OULAD), une base fiable, académique et largement utilisée dans la recherche scientifique en Learning Analytics.
 - Ce dataset, mis à jour en octobre 2024, constitue une référence internationale et garantit la robustesse des analyses.
2. Prétraitement sous SQL
 - Consolidation et croisement des différentes tables de la base relationnelle (étudiants, inscriptions, évaluations, activités en ligne).
 - Création d'une table centralisée regroupant les informations essentielles par étudiant : données sociodémographiques, scores moyens, nombre de clics sur la plateforme, nombre de ressources utilisées, etc.
 - Cette étape a permis de structurer et préparer les données pour l'analyse avancée.
3. Nettoyage et analyse exploratoire sous Python
 - Traduction des données en français, gestion des valeurs manquantes, des doublons et détection des *outliers*.
 - Réalisation d'une analyse exploratoire (EDA) pour comprendre la distribution des variables et mettre en évidence les premiers liens avec l'abandon.
 - Production de statistiques descriptives et de visualisations pour interpréter les données.
4. Modélisation prédictive (Machine Learning)
 - Mise en place d'un modèle de régression logistique pour prédire la probabilité d'abandon.
 - Évaluation des performances à travers des métriques adaptées (précision, matrice de confusion, rapport de classification).
 - Analyse des résultats, interprétation des variables explicatives et discussion des limites.
5. Visualisation et restitution sur Power BI
 - Crédit de tableaux de bord interactifs permettant d'explorer les indicateurs clés : profil des étudiants, taux d'abandon, scores, activité en ligne, prédictions.
 - Élaboration d'une page dédiée à la prédition de l'abandon afin de rendre les résultats accessibles et exploitables pour la prise de décision.

I. Analyse générale

Cette section présente une analyse descriptive des données, en examinant le profil démographique, le niveau académique, la performance et l'engagement des apprenants. L'objectif est d'identifier les tendances et signaux potentiels liés à l'abandon.

1. Profil démographique

- Nombre total d'inscriptions : 1 000
- Nombre d'étudiants : 891
- Répartition par sexe : 54 % hommes, 46 % femmes
- Tranche d'âge : 18 à plus de 50 ans
- 12 % des étudiants déclarent un handicap

2. Niveau académique

- \leq Bac : 792 étudiants
- Diplôme universitaire : 188 étudiants
- Diplôme de 3^e cycle : 12 étudiants
- Sans diplôme : 8 étudiants

3. Performance académique (score moyen)

- Général : 30,45/100
- Diplôme de 3^e cycle : 54/100 (meilleur score)
- Diplôme universitaire : 40/100
- \leq Bac : 28/100
- Sans diplôme : 14/100 (score le plus faible)

4. Engagement (crédits moyens)

- Général : 83 crédits
- Sans diplôme : 111 crédits (plus élevé)
- Diplôme universitaire : 93 crédits
- Autres : proches de la moyenne générale

Observation

Comme illustré à la **Figure 1**, les étudiants plus diplômés performent mieux académiquement, tandis que les apprenants sans diplôme montrent un engagement plus élevé en termes de crédits. Ce contraste suggère que l'engagement seul n'est pas un indicateur suffisant de performance et pourrait constituer un signal pertinent pour la prédiction de l'abandon.

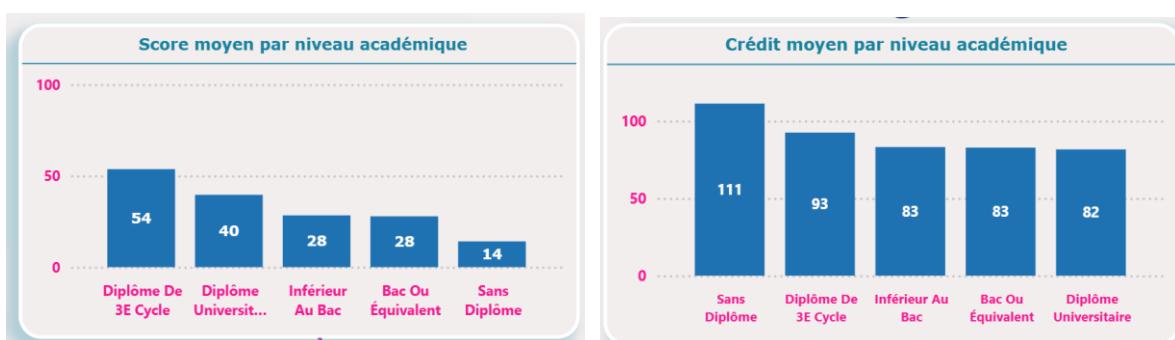


Figure 1 : Répartition des scores moyens et crédits moyens par niveau académique et (capture Power BI)

II. Statistiques démographiques pour l'abandon

Sur 1 000 inscriptions, 319 abandons ont été enregistrés. L'analyse démographique de ces abandons permet d'identifier certains profils plus à risque.

1. Sexe et handicap

- Hommes : 55 % des abandons
- Femmes : 45 % des abandons
- Étudiants ayant déclaré un handicap : 15 %

Les données suggèrent que les hommes présentent une probabilité légèrement plus élevée d'abandon que les femmes.

2. Tranche d'âge

- Moins de 35 ans : 65 % des abandons
- Entre 36 et 54 ans : 34 %
- 55 ans et plus : 1 %

Les jeunes adultes sont donc davantage susceptibles de décrocher, ce qui pourrait être lié à des contraintes personnelles ou professionnelles et à un manque d'expérience avec l'apprentissage en ligne.

3. Niveau académique

- Niveau \leq Bac : taux d'abandon le plus élevé
- Diplôme universitaire : taux intermédiaire
- Diplôme de 3^e cycle : taux le plus faible

Les étudiants avec un niveau académique inférieur ou égal au Bac semblent rencontrer plus de difficultés pour s'adapter au rythme et à l'intensité des cours en ligne.

4. Répartition géographique

Les régions avec le plus fort taux d'abandon sont : West Midlands, Écosse, Londres et North Western Region, avec des taux supérieurs à 10 %. À l'inverse, l'Irlande enregistre le taux le plus faible (1,57 %).

Observation

Les abandons sont plus fréquents chez les hommes, les jeunes adultes et les étudiants ayant un niveau académique inférieur. Ces variables démographiques constituent des signaux importants à prendre en compte pour l'analyse et la prédiction de l'abandon sur les plateformes d'apprentissage en ligne.

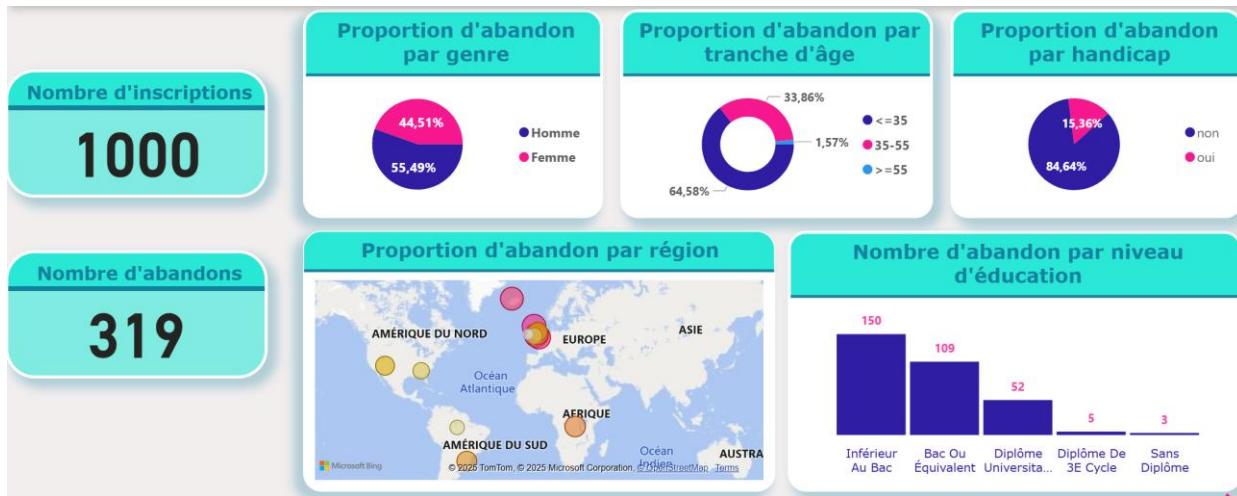


Figure 2 : Statistiques démographiques (capture Power BI)

III. Statistiques comportementales pour l'abandon

L'analyse comportementale des étudiants apporte un éclairage complémentaire sur les mécanismes d'abandon. Le repère temporel utilisé dans les graphiques correspond à la date de début de la formation.

1. Nombre d'abandons par état d'inscription

- 174 abandons enregistrés après 30 jours ou plus du début de la formation
- 141 abandons à moins de 30 jours
- 4 abandons sans désinscription formelle

Observation

Les étudiants qui avancent dans la formation semblent plus susceptibles de décrocher que ceux qui abandonnent dès le début.

2. Nombre d'abandons selon les tentatives d'inscription

- 223 abandons chez les étudiants en première tentative
- 67 abandons chez ceux avec une tentative précédente
- 19 abandons chez ceux ayant 2 tentatives
- Moins de 10 abandons pour les étudiants ayant 3 ou 4 tentatives

Observation

Les étudiants sans expérience préalable en e-learning abandonnent davantage. Ceux qui ont déjà tenté une inscription connaissent mieux la plateforme, ciblent mieux les ressources et gèrent plus efficacement leur temps.

3. Performance académique en fonction des tentatives

- 0 tentative : 329 réussites, 223 abandons, 131 échecs, 75 mentions
- 1 tentative : 54 réussites (dont 4 mentions), 67 abandons, 48 échecs
- 2 tentatives : 11 réussites, 19 abandons, 21 échecs
- 3 tentatives ou plus : 4 réussites, 10 abandons, 8 échecs

Observation

Parmi les 242 étudiants ayant plus d'une tentative antérieure, 181 (soit 75 %) n'atteignent pas leurs objectifs (abandon ou échec). Seuls 69 étudiants parviennent à obtenir leur diplôme, ce qui traduit une persistance des difficultés malgré l'expérience préalable.

4. Corrélation entre l'activité sur la plateforme et le résultat final

- Étudiants avec mention : 79 clics, 1 source utilisée (activité ciblée)
- Étudiants très actifs (28 sources, 394 clics) : réussite plus faible
- Étudiants en échec/abandon : ~6 sources, 319 clics (forte activité mais peu efficace)

Observation

Les étudiants qui réussissent avec mention ont une activité plus ciblée et efficace, tandis que ceux qui échouent ou abandonnent dispersent leurs efforts avec une utilisation désordonnée des ressources.

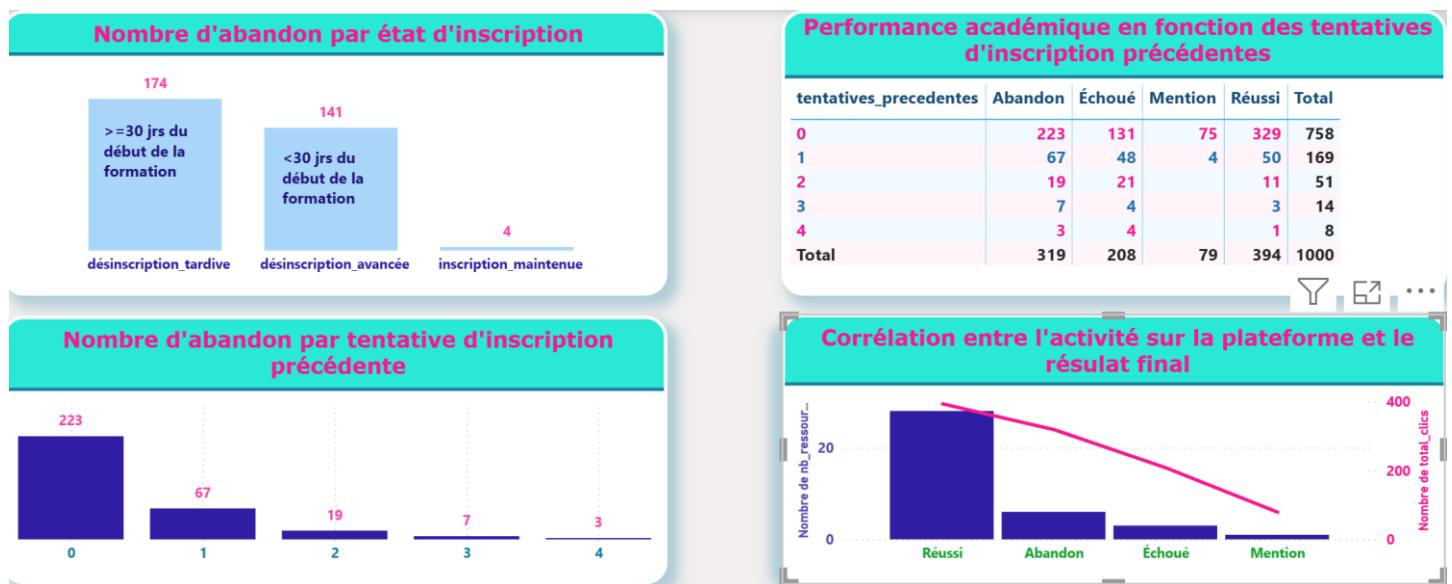


Figure 3 : Statistiques comportementales (capture Power BI)

IV. Statistiques de performance

L'analyse des **1 000 inscriptions** montre que **473 étudiants ont réussi**, dont **79 avec mention**, **208 ont échoué** et **319 ont abandonné**. Ces données servent de base à l'étude des performances académiques et des comportements d'étude.

1. Analyse du score moyen selon la performance académique

Le score moyen obtenu par les étudiants varie significativement selon leur niveau de réussite :

- Les étudiants ayant réussi avec mention présentent le **score moyen le plus élevé (49)** ;
- Ceux ayant réussi sans mention ont un score moyen de **40** ;
- Les étudiants ayant échoué ont obtenu un score moyen de **24** ;
- Les étudiants ayant abandonné ont le **score moyen le plus faible (18)**.

Cette distribution illustre une corrélation positive entre performance académique et score moyen. Les étudiants les mieux notés correspondent à ceux ayant obtenu la mention, tandis que les étudiants ayant abandonné ont des performances globalement faibles.

2. Analyse du crédit moyen selon la performance académique

L'étude du crédit moyen étudié révèle des résultats moins intuitifs :

- Les étudiants ayant abandonné présentent le **crédit moyen le plus élevé (94)** ;
- Les étudiants ayant obtenu une mention ont un crédit moyen de **82** ;
- Les étudiants ayant échoué ont un crédit moyen de **80** ;
- Les étudiants ayant réussi sans mention ont un crédit moyen de **76**.

Contrairement au score moyen, le crédit moyen étudié montre que les étudiants ayant abandonné avaient accumulé le plus grand nombre de crédits. Cette situation peut refléter une surcharge académique ou un désintérêt progressif pour le programme, soulignant l'importance de suivre la charge de travail des étudiants pour prévenir l'abandon.



Figure 4 : Statistiques de performances (capture Power BI)

V. Abandon et prédiction

L'abandon des apprenants dans les formations en ligne représente un enjeu majeur pour les organismes de formation et les étudiants souhaitant suivre une formation continue ou en Bootcamp. Identifier les facteurs contribuant à cet abandon permet de mieux cibler les interventions et d'améliorer l'expérience d'apprentissage.

Dans ce contexte, l'utilisation de modèles de prédiction basés sur les données des apprenants – telles que l'activité sur la plateforme, le crédit étudié ou la réussite aux évaluations – offre la possibilité de détecter de manière proactive les profils à risque. Ces analyses prédictives permettent non seulement de comprendre les comportements des apprenants, mais aussi de mettre en place des actions personnalisées pour réduire l'abandon et maximiser l'efficacité des parcours de formation.

Le modèle a prédit un taux d'abandon de 31.74%, un taux d'étudiants à haut risque d'abandon de 13.60% comparé au taux réel d'abandon qui est de 32%.

VI. Recommandations

L'analyse des données consolidées des apprenants permet d'identifier des profils variés selon le risque d'abandon, le niveau d'engagement et la performance académique. Ces profils offrent une base solide pour formuler des recommandations ciblées afin de réduire l'abandon et d'améliorer la réussite des apprenants.

1. Étudiants à haut risque d'abandon

Ces étudiants présentent un faible score moyen, un nombre limité de clics sur la plateforme et utilisent peu de ressources. L'identification de ce profil grâce au modèle prédictif permet de mettre en place des interventions personnalisées telles que le tutorat en ligne, des entretiens individuels ou en groupe hebdomadaires, des emails de motivation et des notifications ciblées. Ces actions visent à augmenter leur engagement et leur probabilité de réussite.

2. Étudiants faiblement engagés

Certains étudiants consultent peu de contenus ou ont un faible nombre de clics, bien que leur niveau académique ne soit pas forcément critique. Pour ce profil, les recommandations portent sur l'encouragement à utiliser les ressources disponibles. Par exemple, des suggestions automatisées de contenus ou des alertes sur les ressources clés peuvent stimuler leur interaction avec la plateforme.

3. Étudiants en difficulté académique

Ce groupe présente un score moyen inférieur aux seuils critiques, indépendamment de leur engagement. Les recommandations pour ces apprenants incluent des exercices supplémentaires, des sessions de révision et un suivi pédagogique individualisé, des forums de discussions accessibles à tout moment avec des mini quiz à thème. L'objectif est de renforcer leurs connaissances et de réduire les risques d'échec.

4. Étudiants récurrents ou avec tentatives multiples

Les étudiants ayant échoué à plusieurs reprises peuvent bénéficier d'un entretien individuel pour identifier les facteurs d'échec et un plan de soutien personnalisé, intégrant un parcours d'apprentissage adapté à leurs besoins et un suivi rapproché. Cette approche vise à éviter la répétition des échecs et à améliorer leur confiance et performance.

5. Étudiants performants ou très engagés

Enfin, les apprenants ayant un score moyen élevé et un fort engagement doivent être encouragés à poursuivre leurs efforts. Des recommandations telles que l'accès à des contenus avancés, des quiz interactifs ou des badges de reconnaissance contribuent à maintenir leur motivation et à valoriser leurs acquis.

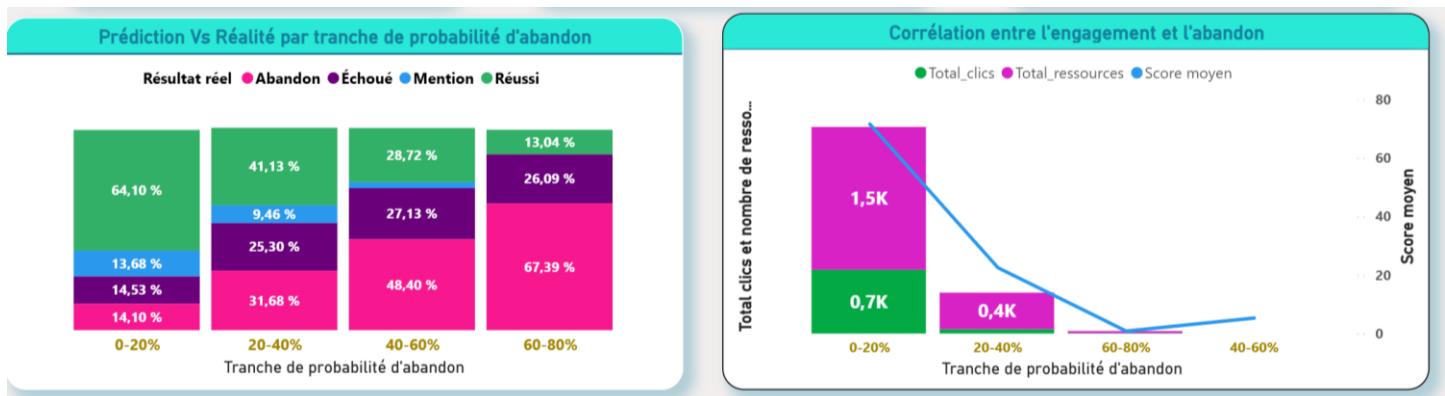


Figure 5 : Prédiction de l'abandon et corrélation avec l'engagement (capture Power BI)

La mise en place de ces recommandations montre l'importance de la **personnalisation de l'accompagnement pédagogique** dans le e-learning. La combinaison d'un modèle prédictif pour identifier les risques d'abandon et de mesures concrètes adaptées aux différents profils permet de :

- Réduire l'abandon
- Améliorer l'engagement des étudiants
- Optimiser l'utilisation des ressources pédagogiques
- Renforcer la réussite globale

Ces recommandations soulignent le rôle central de la **data dans la prise de décision pédagogique**, permettant de passer d'une approche standardisée à une approche proactive et personnalisée.

Conclusion

Ce projet a permis de mieux comprendre les facteurs qui influencent l'abandon des apprenants dans un environnement de formation en ligne.

- La consolidation des données issues de différentes tables (informations démographiques, résultats, interactions avec la plateforme) a permis de créer une base unifiée et riche pour l'analyse.

- Les analyses exploratoires et les visualisations ont montré que l'engagement (mesuré par le nombre de clics et de ressources utilisées) et les performances académiques (score moyen) sont fortement liés au risque d'abandon.
- Le modèle de régression logistique appliqué a permis de prédire avec une précision satisfaisante les étudiants à risque, en confirmant le rôle majeur de variables comme le score moyen, l'âge, les crédits étudiés et l'activité sur la plateforme.
- L'approche adoptée contribue ainsi à fournir un outil de support à la décision, permettant d'identifier en amont les apprenants vulnérables afin de réduire le taux d'abandon.

Perspectives

Plusieurs pistes d'amélioration et de valorisation du projet peuvent être envisagées :

Amélioration des modèles prédictifs

- Tester d'autres algorithmes plus performants (Random Forest, Gradient Boosting, XGBoost) afin d'améliorer la précision.
- Explorer le deep learning pour capturer des relations non linéaires entre variables.

Enrichissement des données

- Intégrer des données qualitatives (feedbacks, enquêtes de satisfaction) pour compléter la vision quantitative.
- Prendre en compte la temporalité des clics et activités (analyse séquentielle ou séries temporelles).

Déploiement opérationnel

- Développer un tableau de bord interactif (via Power BI) pour le suivi en temps réel du risque d'abandon.
- Mettre en place un système d'alerte automatique pour les tuteurs et formateurs.

Perspectives pédagogiques

- Concevoir des stratégies d'accompagnement personnalisées pour les étudiants identifiés comme à risque (tutorat, messages de motivation, ressources supplémentaires).
- Expérimenter des actions correctives et mesurer leur impact sur la réduction du taux d'abandon.

Référence bibliographique

Sources

1 : Colorlib / Statista : <https://fr.statista.com/>

2 : *Class Central — rapport 2023* : www.ClassCentral.com

DATASET : Open University Learning Analytics Dataset (OULAD)

Source : <https://analyse.kmi.open.ac.uk/open-dataset?>

Liste des figures

Figure 1 : Répartition des scores moyens et crédits moyens par niveau académique et (capture Power BI)

Figure 2 : Statistiques démographiques (capture Power BI)

Figure 3 : Statistiques comportementales (capture Power BI)

Figure 4 : Statistiques de performances (capture Power BI)

Figure 5 : Prédiction de l'abandon et corrélation avec l'engagement (capture Power BI)