Analyse de l'engagement de l'étudiant lié à la difficulté d'items d'évaluation : Une étude d'EdNet fondée sur l'apprentissage automatique

Mohamed Lamgarraj[1][0009−0007−4708−1740], Céline Joiron[1][0009−0008−5782−8024], Aymeric Parant[2][0000−0003−0993−0661], and Gilles Dequen[1][0000−0001−7497−1182]

[1] Laboratoire de Modélisation, Information et Systèmes (MIS), Picardie Jules Verne university, Amiens 80000, France

[2] Centre de Recherche en Psychologie : Cognition, Psychisme et Organisations (CRP-CPO), Picardie Jules Verne university, Amiens 80000, France

Résumé . L'intégration des plateformes numériques dans l'éducation offre des opportunités pour suivre les activités des étudiants et utiliser ces données dans l'apprentissage automatique afin d'analyser les schémas d'apprentissage et leur impact sur les performances des étudiants. Cette étude s'inscrit dans le cadre de nos travaux de recherche sur l'analyse et l'exploitation des traces pédagogiques pour optimiser les systèmes de formation numérique. S'appuyant sur le jeu de données EdNet, comprenant plus de 100 millions d'interactions issues d'une plateforme d'apprentissage et d'évaluation numérique, nous examinons les corrélations entre les activités des apprenants, les contenus consultés, leurs taux de réussite, et la difficulté des questions proposées. En poursuivant nos travaux antérieurs sur la caractérisation de la difficulté des questions, nous appliquons des modèles d'apprentissage automatique pour prédire la réussite des étudiants, et identifier les principaux facteurs influençant cette réussite, tels que la nature des interactions, les niveaux de difficulté, et l'exposition aux contenus pédagogiques.

Mots-clés. Traces pédagogiques, Évaluation des Apprentissages, Difficulté, Apprentissage automatique, Intelligence artificielle (IA).

Abstract. The integration of digital platforms in education offers opportunities to track student activities and leverage this data in machine learning to analyze learning patterns and their impact on student performance. This study is part of our research on the analysis and use of educational traces to optimize digital training systems. Using EdNet dataset, which includes over 100 million interactions from a digital learning and assessment platform, we examine correlations between learners' activities, the content they engage with, their success rates, and the difficulty of the questions presented. Building on our previous work on characterizing question difficulty, we apply machine learning models to predict student success. These models also help identify the key factors influencing this success, such as the nature of interactions, levels of difficulty, and exposure to educational content.

Keywords: Learning Analytics · Data Mining · Student Performance · Question Difficulty · Artificial Intelligence (AI)

1 Introduction

Nos travaux s'inscrivent dans la cadre d'une recherche sur l'évaluation des apprentissages dans les systèmes numériques, et plus spécifiquement l'apport des techniques d'apprentissage automatiques dans la compréhension et la proposition d'évaluations et d'items d'évaluation dont on peut estimer/prédire la difficulté. En effet, avec l'évolution rapide des technologies et l'essor des plateformes numériques dans le domaine de l'éducation, la collecte et l'analyse des données éducatives offrent des opportunités pour mieux comprendre les facteurs influençant la réussite académique des étudiants, en particulier lors des test d'évaluation en ligne. De nombreuses études se sont concentrées sur les effets de l'engagement envers le contenu (l'ensemble des interactions de l'apprenant avec les ressources pédagogiques, telles que la consultation de contenu explicatifs, les ressources lié au sujet, ou encore le temps passé sur ces éléments.), les niveaux de difficulté des exercices, et leur impact sur les performances.

Les recherches montrent que l'interaction avec des supports éducatifs riches et adaptés peut améliorer significativement la compréhension et les résultats des étudiants. Des travaux tels que ceux de Chi et Wylie, 2014 ont introduit le concept de "tutoring cognitif actif", soulignant que l'engagement actif, mesuré par des interactions fréquentes et prolongées avec les contenus, favorise une meilleure mémorisation et une compréhension approfondie [4]. Pour le Séquençage des contenus, Des études ont également montré que l'ordre dans lequel les étudiants accèdent au contenu (par exemple, regarder des explications avant de répondre à des questions) a un impact direct sur leur réussite [13].

D'un autre côté, la difficulté des questions joue un rôle crucial dans le maintien de l'équilibre entre motivation et apprentissage. Par exemple la théorie de Zone proximale de développement (ZPD), inspirée des travaux de Vygotsky, suggère que les questions doivent être suffisamment difficiles pour stimuler l'apprenant, mais pas au point de décourager ses efforts [11]. Cette approche a été explorée dans des systèmes adaptatifs tels que ASSISTments [9]. Dans la modélisation des performances, des algorithmes tels que Knowledge Tracing (KT) et ses variantes améliorées (Bayesian KT, Deep KT) ont permis de modéliser la probabilité qu'un étudiant réponde correctement à une question donnée en fonction de la difficulté et de son historique d'apprentissage [16] .

Peu de recherches croisent directement l'impact de l'engagement, avec celui de la difficulté des questions, sur les performances globales des étudiants. Nous pouvons mentionner des travaux sur l'interaction entre engagement et adaptativité, qui ont exploré comment les plate-formes adaptatives, telles que ALEKS et Duolingo, modifient dynamiquement la difficulté des questions en fonction du niveau d'engagement de l'utilisateur, obtenant ainsi des gains d'apprentissage significatifs [7]. Enfin, une étude sur les plate-formes EdTech a montré que les

explications interactives, lorsqu'elles sont combinées avec des questions adaptées, augmentent les performances des étudiants de 25 % en moyenne [15].

Bien que ces études aient été menées sur chacun des aspects importants dans le knowledge-tracking, peu de travaux relient directement l'engagement, la difficulté des questions, et le succès des étudiants dans un cadre unifié. Cependant, une analyse intégrée nous semble nécessaire pour comprendre les interactions complexes entre engagement, difficulté des questions et succès des étudiants. Cela pourrait servir de base pour concevoir des systèmes encore plus performants et adaptatifs à l'avenir.

L'objectif principal de nos travaux est d'explorer la relation entre l'engagement des étudiants avec le contenu éducatif, les niveaux de difficulté des questions auxquelles ils sont confrontés, et leur succès académique. Ils visent à analyser comment les étudiants interagissent avec différents types de contenu — que ce soit des questions, des leçons ou des explications — et comment ces interactions influencent leur réussite. Aussi la manière dont la difficulté des questions, en relation avec l'engagement des étudiants, impacte leur capacité à réussir les tests. Cette étude exploite les données issues des interactions des étudiants avec le contenu éducatif sur une plate-forme numérique d'apprentissage, et disponibles dans la littérature (le dataset Ednet).

Cet article présente en premier lieu la constitution d'un jeu de données riche et diversifié en exploitant les traces d'utilisation disponibles dans Ednet [5], et en cherchant à obtenir un maximum d'informations pertinentes concernant les questions, les utilisateurs et le contexte d'apprentissage. Cette approche nous permet de bâtir une base de données capable de refléter les interactions complexes entre les étudiants et le contenu éducatif. Une fois le jeu de données constitué, nous détaillerons la méthodologie adoptée, qui repose sur l'application de techniques de data mining pour analyser les données. Ces méthodes nous permettent d'analyser les comportements des étudiants et d'examiner l'impact de leur engagement sur leur réussite académique. Enfin, nous discuterons des résultats obtenus et de leurs implications pratiques, notamment dans le cadre de l'évaluation numérique des apprentissages. .

2 Les données

Pour étudier la relation entre la consultation des ressources pédagogiques et les taux de réussite des apprenants, il est essentiel de disposer d'un ensemble de données riche et diversifié qui dépasse les indicateurs d'engagement traditionnels. Une compréhension approfondie de cette relation nécessite des données qui capturent non seulement les résultats des tests mais aussi l'accès aux ressources. Enfin, l'ensemble de données doit être suffisamment vaste et diversifié pour garantir que les modèles formés sur celui-ci généralisent bien sur divers sujets, thèmes et niveaux de difficulté. Les données que nous préparerons doivent alors être d'une échelle importante, répondre au critère de variété, pour pouvoir identifier des variables nuancées, et aux critères de qualité, c'est-à-dire d'authenticité, de cohérence et de granularité afin de garantir l'efficacité de modèles prédictifs.

Les données utilisées dans cette étude sont les données d'EdNet [5], une des plus grandes bases des traces pédagogiques ouvertes . Elle contient plus de 100 millions d'interactions étudiant-contenu, et inclut des informations détaillées sur les performances des étudiants. EdNet [5] se présente comme un ensemble de données hiérarchique, rassemblant deux ans de journaux d'interaction des étudiants provenant de Santa [1], une solution d'auto-apprentissage multi-plateforme conçue pour aider les étudiants à se préparer au test TOEIC (Test d'anglais pour la communication internationale). En plus des tests d'évaluation, les étudiants ont accès à une variété de ressources pédagogiques de formation en langue anglaise (vidéos, lectures, explications), permettant de renforcer leur compréhension des concepts et de se préparer de manière optimale. Le jeu de données EdNet est composé de 131 441 538 interactions collectées auprès de 784 309 utilisateurs. Elle comprend 19,4 Go de données réparties en 1,6 million de fichiers individuels au format CSV, structurés en quatre niveaux (KT1 à KT4), chacun fournissant des informations de plus en plus détaillées sur les actions des étudiants. KT1 se concentre sur les réponses aux questions, KT2 inclut aussi les interactions avec les explications et les exercices, KT3 couvre la consommation de vidéos et de supports pédagogiques, et KT4 ajoute des éléments sur la gestion du temps et l'ensemble des actions des étudiants sur la plate-forme. Les interactions des étudiants englobent leurs actions avec le matériel pédagogique (questions, vidéos, explications), leurs réponses et le temps passé sur chaque activité. Ces données permettent d'analyser leurs stratégies d'apprentissage et leur gestion du temps. Les questions et bundles représentent des ensembles de problèmes et de cours (plus de 13 000 problèmes et 1 000 cours), regroupés par thème, passage ou média. Les étudiants doivent résoudre toutes les questions d'un bundle pour le compléter, ce qui permet de suivre leur progression dans un contexte cohérent d'apprentissage. Enfin un fichier nommé "Contents", contient des informations sur les explications et les cours, et intègre les actions liées à la consultation d'explications (commentaires d'experts) et à la visualisation de vidéos. Chaque question ou item d'évaluation est associé aux contenus à l'aide d'étiquettes (tags). Ces supports sont cruciaux pour étudier l'impact de consultation de contenue sur la performance des étudiants et leur compréhension. En résumé, EdNet offre une vue d'ensemble détaillée des actions des étudiants à travers plusieurs niveaux, en analysant leur engagement avec différents types de contenu et leur gestion du temps d'apprentissage.

2.1 Présentation, enrichissement et préparation des données EdNet

Afin de constituer un ensemble de données pertinent, répondant à nos besoins de recherche, nous avons mené un processus rigoureux de regroupement, d'organisation et d'agrégation de l'ensemble des données EdNet pour nos premières tâches de recherche. L'ensemble EdNet, d'une taille de 19,4 Go et composé de 1,6 million de fichiers .csv, a servi de base pour cette analyse. Pour garantir la richesse des données et leur adéquation avec nos objectifs, Plusieurs étapes de préparation ont été effectuées lors de nos premières tâches de recherche, qui visaient à identifier les éléments influençant la difficulté perçue et susceptibles d'aider à prédire

celle-ci [14]. Ces étapes ont permis de structurer et d'optimiser les données de la façon suivante :

- Fusion des traces des apprenants : regroupement des 784,309 fichiers CSV relatifs à chaque utilisateur présents dans Ednet-KT1 en un fichier plat intégrant l'identifiant de l'utilisateur. Le résultat est un fichier composé de 95 293 926 lignes et 6 colonnes.

- Insertion des données relatives aux questions, et au contenu, issues du fichier 'Contents' telles que la réponse correcte à la question, le contenu associé à chaque question, les thèmes ou d'autres méta-données pertinentes.

- Amélioration des données : augmentation des données extraites avec des caractéristiques calculées et déduites. Ces caractéristiques conçues peuvent inclure des ratios, des statistiques, ou d'autres métriques dérivées des données initiales qui fournissent une compréhension plus approfondie des motifs sous-jacents dans l'ensemble de données.

Cette procédure élargit la portée de notre enquête tout en nous fournissant un ensemble de variables plus complet, susceptible de révéler des liens non découverts et d'aider à une meilleure prise de décision. Une série d'expérimentations progressives, consistant à ajouter ou retirer certaines caractéristiques afin d'identifier la combinaison produisant les meilleurs résultats en termes de corrélation avec la réussite des étudiants. Les informations et caractéristiques supplémentaires que nous avons intégrées au jeu de données sont comme suit :

Caractéristiques utilisateur :
mean_user_accuracy : taux de réussite moyen de chaque utilisateur.
question_answered : nombre total de questions répondues par un utilisateur.
AverageTimeperUser : temps moyen passé par l'utilisateur sur une question.
question_explanation : indique si l'étudiant a vu une explication.

Caractéristiques des questions :
answered_correctly : résultat de l'interaction (1 pour une bonne réponse, 0 pour une mauvaise réponse).
tags_community : groupes d'étiquettes associées aux questions, créées par des algorithmes de clustering présenté en section 3.2.
mean_content_accuracy : taux de réponse correcte pour chaque question.
difficulty_level : niveau de difficulté basé sur le taux de bonnes réponses (et notre modèle de prédiction).

Caractéristiques combinées :
performance_gap : Différence entre la précision moyenne d'un utilisateur et le taux de réussite moyen d'une question.
Les données finales sont constituées des éléments initiaux de base de données, plus les éléments suivants que nous avons ajoutés (dérivés des données brutes) :

Table 1: Noms de colonnes de l'ensemble de données

user_id	question_id	correct_answer	user_answer
is_correct	elapsed_time	bundle_cl	explanation_id
part	tags-cl	question_answered	mean_user_accuracy
AverageTimeperUser	AverageTimeforCorrect	averageTimeforFalse	performance_gap
question_explanation	difficulty (GT)	answered_correctly	tags_community

2.2 Exploration et analyse des données

À la suite des étapes de préparation et d'enrichissement, notre ensemble de données final regroupe les traces d'interaction de 393 656 utilisateurs uniques qui ont effectué 101 230 332 interactions. Ainsi, en moyenne, un utilisateur a effectué 257,15 interactions. Ces données vont nous permettre d'identifier des tendances entre certaines caractéristiques des interactions (comme la durée moyenne par réponse ou le taux d'engagement vis à vis des tags de contenus) et la probabilité de répondre correctement à une question ('answer_correctly'). Il convient alors de pré-traiter les données et de créer des caractéristiques permettant aux modèles de prédire la probabilité d'exactitude. Ainsi, la relation entre la variable et les autres caractéristiques doit être étudiée.
Dans cette phase, nous procédons à une Analyse des Données Exploratoire (EDA), qui a pour objectif d'examiner en détail les relations entre les différents éléments de l'ensemble de données. Cette étape permet non seulement d'identifier les connexions et dépendances potentielles entre les variables (comme la difficulté des questions, le temps de réponse, ou les communautés de tags), mais aussi d'évaluer la pertinence des données pour la tâche d'apprentissage menée. En explorant les interactions et les caractéristiques associées, nous cherchons à extraire des statistiques descriptives et des observations clés, qui guideront les étapes suivantes de modélisation.
Parmi les premières observations, nous notons que 65,73% des questions ont été correctement répondues par les utilisateurs (Fig. 1). Ce taux de réussite global reflète une performance relativement positive des apprenants. Il fournit également des indications sur l'équilibrage de notre base de données nécessaire pour analyser l'impact des caractéristiques des questions et des contenus éducatifs sur les résultats obtenus.

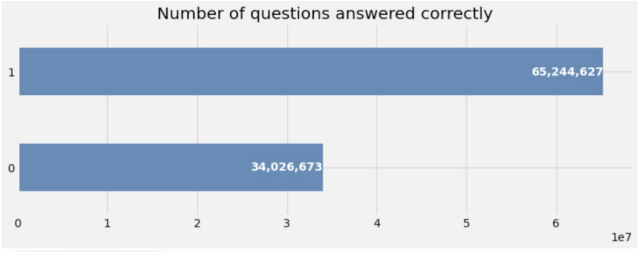


Fig. 1: Nombre des questions correctement répondues '1', avec des reponses fausses '0'

Chaque question peut avoir une ou plusieurs étiquettes associées, représentant des compétences spécifiques évaluées dans la question. Un premier travail consiste donc à encoder ces combinaisons d'étiquettes de façon à pouvoir les exploiter via l'analyse des données. Etant donné le caractère pédagogique des items, nous partons de l'hypothèse que certaines étiquettes sont regroupées plus fréquemment avec d'autres. Ainsi, plutôt que de stocker chaque tag associé à une question, l'idée est d'associer les informations relatives aux groupes de tags auxquels chaque item fait référence. Il s'agit alors de regrouper les tags en communautés en les divisant selon leurs clusters, et d'utiliser ces derniers pour encoder les étiquettes. Un premier graphe a été construit à l'aide du module Networkx. Chaque étiquette y est représentée par un nœud, et les arêtes indiquent les co-occurrences des étiquettes dans les items de la base. Le poids de chaque arête représente le nombre de questions dans lesquelles deux étiquettes adjacentes apparaissent ensemble(Fig. 2).

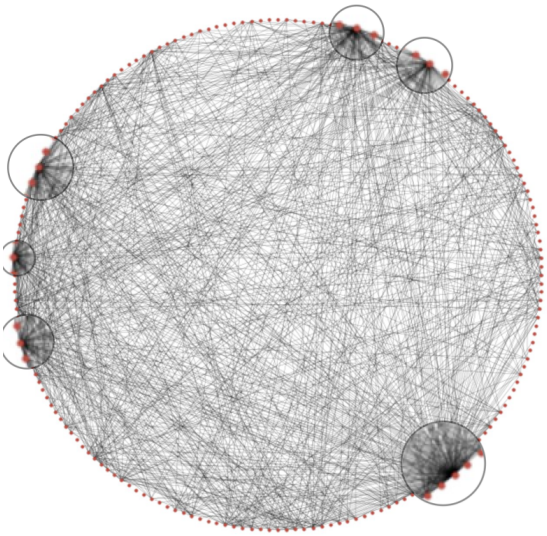


Fig. 2: Graphe des associations d'étiquettes d'items (tags)

Ce graphe révèle que certains nœuds forment des zones plus denses, ce qui indique que les étiquettes associées à ces nœuds apparaissent plus fréquemment que d'autres. Une analyse plus approfondie montre également que certains nœuds ne possèdent aucune arête ce qui indique que certaines étiquettes apparaissent toujours seules dans les items de la base. Ces dernières ont donc été isolées afin de poursuivre l'observation des motifs de regroupement en traçant à nouveau le graphe avec les étiquettes restantes(Fig. 3). Ce second graphe mets en évidence que certains nœuds ne sont connectés qu'au nœud central. Ainsi, ces étiquettes apparaissent soit seules, soit en paire avec l'étiquette représentée par le nœud central, que nous identifions comme étant l'étiquette 162. Ces étiquettes ayant une seule connexion sortante ont été listées, ainsi que le nombre de fois où elles

apparaissent en paire avec l'étiquette 162. Enfin, un troisième graphe permet de représenter les nœuds restants, qui présentent des associations intrinsèques plus complexes (Fig. 4). Ce dernier établit l'existence de deux regroupements. Un algorithme de détection de communautés permettant de séparer les nœuds en deux clusters distincts puis de les colorer en rouge et bleu confirme cette observation.

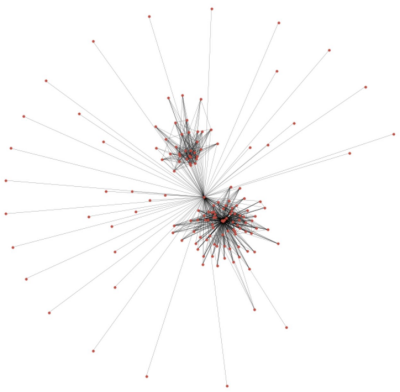


Fig. 3: Etiquettes par rapprochement des nœuds les plus liés

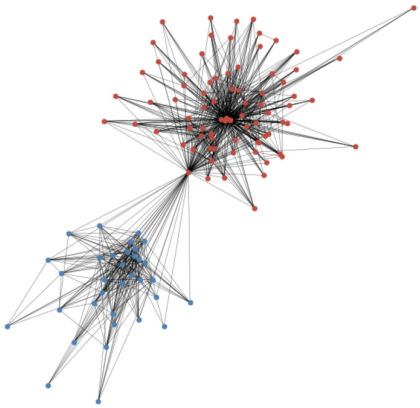


Fig. 4: Groupes d'étiquettes

Les étiquettes de 12 284 questions ont été catégorisées en quatre communautés distinctes : une liste contenant toutes les étiquettes qui apparaissent uniquement de manière isolée (communauté 0) ; une liste regroupant toutes les étiquettes qui apparaissent seules ou en paire avec l'étiquette 162 (communauté 1); deux clusters révélées à partir des étiquettes restantes (communautés 2 et 3). Le tableau (Tab. 2) présente la répartition des items par communauté.

Table 2: la répartition des questions sur les 4 groupes de tages

tags Community	0	1	2	3
number of questions	512 (4.17%)	2,045 (16.64%)	5,972 (48.62%)	3,755 (30.57%)

En plus des étiquettes thématiques pour chaque question, les données EdNet classifient les items en différentes parties, correspondant à des types spécifiques d'activités éducatives réalisées par les apprenants. Ces parties se décomposent en sections telles que Écoute comprend les parties 1 à 4 (incluant des activités comme Photographies, Questions-Réponses, Conversations et Discours) et Lecture dans les parties 5 à 7 (comprenant des activités comme Phrases Incomplètes, Complétion de Texte, et Passages Simples ou Multiples). Ces différentes parties reflètent la diversité des compétences sollicitées dans le processus d'apprentissage et dévaluation, et peuvent nous permettre d'étudier des patterns d'interactions spécifiques. Ainsi, la figure ci-dessous illustre la répartition des réponses correctes

et incorrectes par partie, offrant un aperçu des performances des apprenants en fonction du type d'activité (Fig. 5).

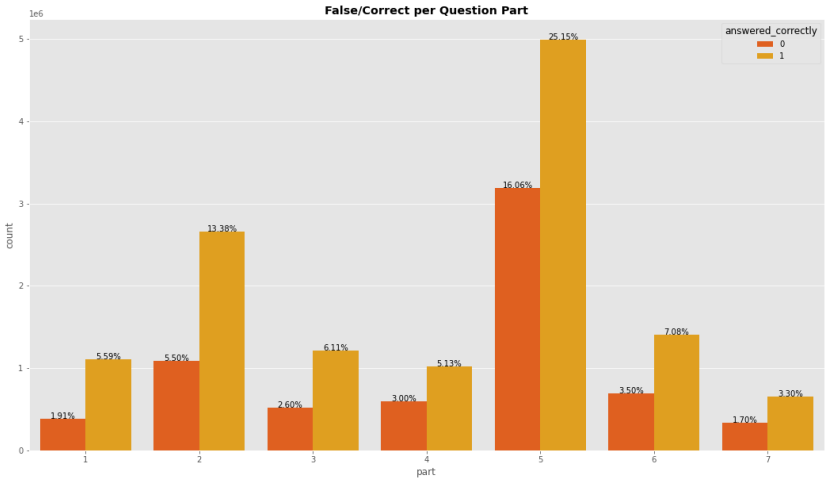


Fig. 5: Répartition des réponses correctes et incorrectes par parties

La section suivante explore les modèles expérimentés sur ces ensembles de données afin d'évaluer leur capacité à prédire la réussite des apprenants, au regard de la difficulté des questions, et à mettre en évidence les relations entre les différents contenus, thématiques ou activités pédagogiques des apprenants. Les résultats obtenus sont ensuite analysés.

3 Modèles et résultats

Après avoir identifié les données significatives pour les tâches d'apprentissage automatique, notre approche consiste à entraîner des modèles capables d'exploiter cet ensemble de données pour fournir des représentations précises de l'engagement de l'apprenant, tout en permettant une estimation robuste de son impact sur la réussite des apprenants aux items. Pour cet entraînement il convient en premier lieu de sélectionner un échantillon représentatif des données issues de la base de données construite comme préconisé dans [2]. Ainsi, Les profils utilisateurs sélectionnés couvrent une diversité de comportements, incluant des apprenants sans interaction et des utilisateurs modérément et fortement actifs, selon le nombre total d'interactions. Ces derniers représentent 40% des profils sélectionnés de façon à analyser avec précision les comportements d'apprentissage les plus significatifs. Les questions ayant un nombre de passages compris entre 10 et 100 utilisateurs ont été sélectionnées afin de garantir une représentativité équilibrée entre popularité et difficulté. De plus, seules les données des interactions enregistrées durant la même période ont été utilisées pour s'assurer de leur fraîcheur et pertinence. Enfin, nous nous sommes assurées que les utilisateurs sélectionnés

soient diversifiés, composés d'utilisateurs sans interactions avec le contenu et d'utilisateurs qui ont interagi avec des types de contenus différents à des fréquences variées, afin de capturer une diversité de comportements et d'optimiser la richesse des analyses.
Une série d'expériences permet d'évaluer l'importance des différents éléments de la base de données dans la prédiction de la variable cible indiquant la réussite (answered_correctly).Plusieurs modèles d'apprentissage automatique ont été utilisés afin de mesurer l'impact de chaque ajout progressif d'éléments sur les performances prédictives. Les modèles sélectionnés incluent LightGBM, XGBoost et un réseau de neurones implémenté avec Keras. Chaque modèle a été entraîné sur 50 % des données et testé sur l'autre moitié, en utilisant l'AUC (Area Under Curve) [8] comme métrique principale pour évaluer la performance (Tab. 3).Une répartition équivalente (50/50) entre les données d'entraînement et de test a été retenue afin de maximiser la représentativité de l'échantillon de test et d'éviter le sur-apprentissage, tout en conservant un volume suffisant pour entraîner des modèles performants. La métrique AUC (Area Under Curve) a été utilisée pour évaluer les performances des modèles de classification, car elle est particulièrement adaptée aux problèmes où les classes sont déséquilibrées [6], comme dans notre cas. L'AUC mesure la capacité d'un modèle à distinguer correctement entre deux classes (par exemple, réussite ou échec). Elle est calculée comme l'aire sous la courbe ROC (Receiver Operating Characteristic), qui représente graphiquement le compromis entre le taux de vrais positifs (True Positive Rate) et le taux de faux positifs (False Positive Rate) à différents seuils de classification.

L'entraînement des modèles a été réalisé sur une infrastructure de calcul dédiée fournie par l'université **MatriCS** [17], plus précisément la partition comprenant les environnements **bigmem** et **risk-bigmem**. Ces environnements sont composées de 12 serveurs bi-processeurs Intel Xeon E5-2680 v4 à 2,40 GHz, chacun équipé de 28 cœurs (14 par processeur) et prenant en charge jusqu'à 56 threads grâce à l'architecture Hyper-Threading. Chaque serveur dispose de 512 Go de mémoire vive pour traiter de grands ensembles de données, et utilise un système de stockage haute performance optimisé pour le traitement intensif des données.

Ces ressources ont permis d'entraîner efficacement les modèles sur des données volumineuses, tout en maintenant des temps d'exécution raisonnables pour les expérimentations et les validations croisées. Cette approche permet de mieux comprendre la contribution de chaque élément de la base de données dans la construction d'un modèle prédictif performant. Les résultats obtenus révèlent les différences de performance entre les modèles et mettent en lumière les relations clés entre les caractéristiques extraites et les prédictions finales.
Les modèles utilisés ainsi que les résultats optimaux obtenus sont les suivants :

LightGBM (Gradient Boosting) [12]: Modèle principal utilisé grâce à sa capacité à traiter de grands ensembles de données avec une faible mémoire. LightGBM est une implémentation des arbres de décision à gradient boost-

ing. Le boosting est une technique d'ensemble qui consiste à ajouter de nouveaux modèles successivement pour corriger les erreurs des modèles existants, jusqu'à ce qu'il ne soit plus possible d'améliorer les performances. De plus, le gradient boosting utilise l'algorithme de descente de gradient pour minimiser la perte lors de l'ajout de nouveaux modèles destinés à prédire les erreurs des modèles existants. Les prédictions finales sont ensuite obtenues en sommant l'ensemble des modèles construits.

Paramètres clés :

- `num_leaves` : Contrôle la complexité de l'arbre.
- `max_depth` : Limite la profondeur des arbres pour éviter le sur-apprentissage.
- `early_stopping_rounds` : Arrêt anticipé pour prévenir le sur-apprentissage.

Avantages : Résultat : AUC (Area Under Curve) de **0.760**.

XGBoost (eXtreme Gradient Boosting) [3]: Utilisé pour comparaison, avec des paramètres similaires à ceux de LightGBM. Résultat : AUC de **0.754**, légèrement inférieur à LightGBM.

Keras Neural Network [10]: Modèle de réseau de neurones basé sur des couches denses et convolutives. Utilisé pour comparer avec les modèles de boosting. Résultat : AUC de **0.746**, moins performant que LightGBM et XGBoost.

L'utilisation de l'AUC dans notre étude s'appuie sur ses propriétés adaptées aux caractéristiques de nos données et sur sa capacité à fournir une évaluation fiable des modèles appliqués. Les résultats montrent que le modèle LightGBM offre la meilleure performance en terme de discrimination, ce qui en fait le choix le plus pertinent pour explorer la relation entre les variables de notre base de données et l'objectif (le succès).

Table 3: Les résultas obtenus utilisant le métrique AUC .

Model	LightGBM	XGBoost	keras NN
AUC	0.760	0.754	0.746

D'après les résultats obtenus, les étudiants passant plus de temps sur les explications (en lectures précisément) ont montré une amélioration de 10% de leur précision par rapport à ceux qui ne les ont pas consultées(Fig. 6). La fréquence des interactions est positivement corrélée avec les performances globales, bien qu'une diminution des bénéfices ait été observée après un grand nombre d'interactions (500 interactions), suggérant des rendements décroissants au-delà d'un certain seuil d'engagement. Les étudiants qui interagissaient avec des explications avant d'aborder des questions difficiles ont montré une amélioration de 30 % de leurs performances, mettant en évidence l'importance d'une préparation adéquate face à des défis plus complexes.

L'analyse des taux de réussite en fonction des tags révèle des disparités significatives, indiquant l'existence de variations de difficulté entre les questions appartenant à des catégories similaires (Fig. 7). Ces résultats suggèrent une corrélation entre les types de contenus éducatifs et les performances des apprenants. Certaines questions, bien qu'associées à des classes identiques, présentent des taux de réussite nettement plus faibles, ce qui pourrait être attribué à des facteurs tels

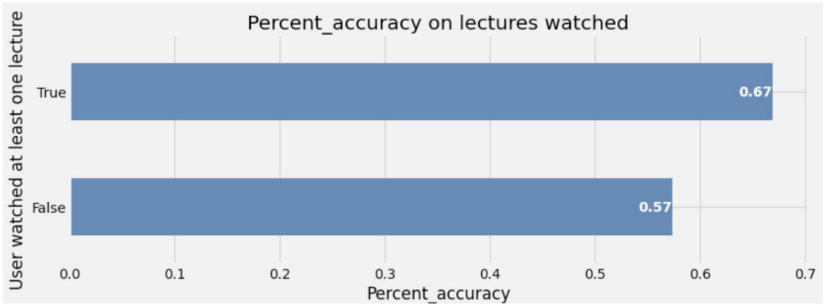


Fig. 6: Impact de l'engagement dans les lectures sur le taux de réussite

que la complexité contextuelle ou l'exigence cognitive. Cette observation met en évidence l'importance d'examiner non seulement le contenu éducatif lui-même, mais également la manière dont il est structuré et présenté, afin de mieux comprendre son influence sur les résultats académiques. Les figures suivantes mon-

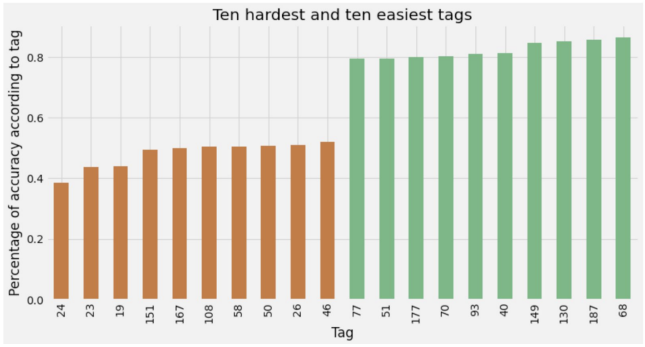
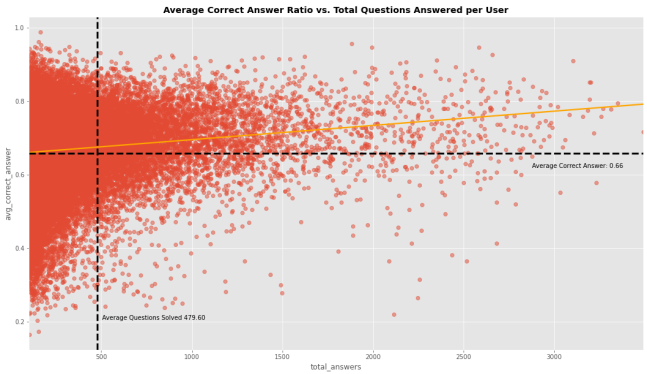


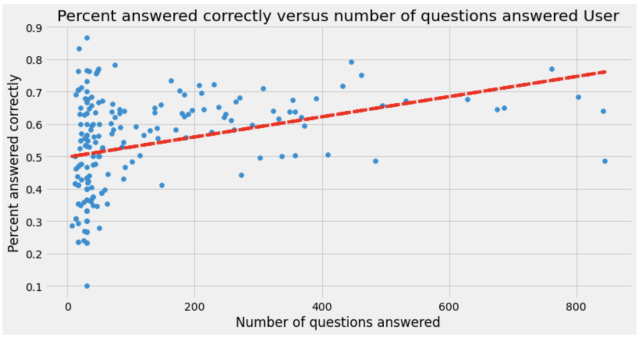
Fig. 7: Taux de réussite pour les 10 tags les plus difficiles et les 10 plus faciles

trent la distribution des réponses correctes en fonction du nombre total de questions répondues par utilisateur (Fig. 8), tant pour les données d'entraînement que pour les prédictions. Cette comparaison permet d'évaluer la performance du modèle en terme de cohérence entre les données réelles et les prédictions effectuées.

La cohérence observée entre les distributions des réponses correctes en fonction du nombre total de questions répondues, tant pour les données d'entraînement que pour les prédictions, suggère une bonne capacité de généralisation du modèle. En effet, les distributions similaires indiquent que les caractéristiques utilisées pour entraîner les modèles sont pertinentes et significatives pour prédire le succès des étudiants. Parmi ces caractéristiques, les colonnes relatives aux statistiques de visites des contenus, telles que la fréquence et le temps passé sur les explications, ainsi que la colonne de difficulté des questions, apparaissent comme des facteurs clés dans la performance prédictive.



(a) Distribution des réponses correctes en fonction du nombre total de questions répondues par utilisateur (données d'entraînement)



(b) Distribution des réponses correctes en fonction du nombre total de questions répondues par utilisateur (Prédictions)

Fig. 8: Comparaison de la distribution des réponses correctes en fonction du nombre total de questions répondues par utilisateur, entre les données d'entraînement et les prédictions.

Les statistiques de visite des contenus, en particulier, montrent une forte corrélation avec les réponses correctes, ce qui confirme l'importance de l'engagement actif des étudiants avec les ressources pédagogiques pour améliorer leur précision dans les réponses. De même, la difficulté des questions, bien que variée, s'est avérée un bon indicateur dans la capacité des étudiants à réussir les questions en fonction de leur niveau de préparation et de leurs interactions avec le contenu.

4 Conclusion

Cette étude met en évidence l'importance des systèmes de formation et d'évaluation numériques dans l'amélioration des performances académiques des étudiants. L'engagement des étudiants avec le contenu pédagogique et l'ajustement des niveaux de difficulté des questions sont des leviers essentiels pour favoriser la réussite. Nous avons identifié que l'utilisation régulière des explications pédagogiques,

telles que les feedbacks, améliore la compréhension et la précision des réponses des étudiants. De plus, l'adéquation entre la difficulté perçue des questions et les capacités des étudiants est cruciale pour optimiser l'apprentissage.
Notre travail s'inscrit dans une démarche visant à exploiter les données issues des traces pédagogiques pour identifier les facteurs influençant la réussite académique et la perception de la difficulté des éléments d'évaluation [14]. En continuité avec nos précédentes recherches sur la caractérisation et la prédiction des difficultés des questions. S'inscrivant dans ce démarche, cette étude a permis de mettre en lumière l'interaction entre divers facteurs clés des traces pédagogiques, notamment l'engagement des étudiants et leurs performances. Nous montrons que l'analyse des données pédagogiques peut concevoir des modèles prédictifs puissants, permettant d'adapter les contenus et évaluations aux besoins individuels des apprenants.
En dépassant les limites des approches théoriques traditionnelles, notre approche basée sur les données et l'apprentissage automatique, appliquée à un large volume de données d'interactions des apprenants sur une plate-forme numérique(EdNet), permet d'identifier des corrélations précises entre les activités des étudiants, la difficulté des questions et leur réussite. Nos analyses révèlent qu'une forte corrélation existe entre les facteurs clés extraits des bases de données et la réussite des étudiants. Par exemple, les étudiants passant plus de temps sur les contenus explicatifs ont un taux de réussite supérieur de 15-20% sur les questions complexes. Un séquencement optimal des tâches (simples avant complexes) améliore également les performances. Pour approfondir ces résultats, il est essentiel d'enrichir les bases de données avec des facteurs psychologiques et comportementaux, tels que la motivation, la gestion du stress et les styles d'apprentissage. Ces éléments, souvent sous-explorés, pourraient interagir de manière significative avec les facteurs académiques. En les intégrant aux traces d'interactions éducatives, nous pourrions non seulement améliorer les modèles prédictifs, mais aussi maximiser l'efficacité des systèmes d'apprentissage en ligne.

5 Remerciements

Ces travaux sont financés dans le cadre du projet ****** porté par ***** et résultant de l'appel à projet Démonstrateur Numérique de l'Enseignement Supérieur de France 2030.

References

1. Aitutor santa. https://www.aitutorsanta.com/vn/
2. Bengio, Y., LeCun, Y.: Scaling learning algorithms toward ai (2007)
3. Chen, T.: Xgboost: extreme gradient boosting. R package version 0.4-2 **1**(4) (2015)
4. Chi, M.T., Adams, J., Bogusch, E.B., Bruchok, C., Kang, S., Lancaster, M., Levy, R., Li, N., McEldoon, K.L., Stump, G.S., et al.: Translating the icap theory of cognitive engagement into practice. Cognitive science **42**(6), 1777–1832 (2018)

5. Choi, Y., Lee, Y., Shin, D., Cho, J., Park, S., Lee, S., Baek, J., Bae, C., Kim, B., Heo, J.: Ednet: A large-scale hierarchical dataset in education. In: Artificial Intelligence in Education: 21st International Conference, AIED 2020, Ifrane, Morocco, July 6–10, 2020, Proceedings, Part II 21. pp. 69–73. Springer (2020)
6. Davis, J., Goadrich, M.: The relationship between precision-recall and roc curves. In: Proceedings of the 23rd international conference on Machine learning. pp. 233–240 (2006)
7. Dong, L., Mallinson, J., Reddy, S., Lapata, M.: Learning to paraphrase for question answering. arXiv preprint arXiv:1708.06022 (2017)
8. Fawcett, T.: An introduction to roc analysis. Pattern recognition letters **27**(8), 861–874 (2006)
9. Feng, M., Heffernan, N., Koedinger, K.: Addressing the assessment challenge with an online system that tutors as it assesses. User modeling and user-adapted interaction **19**, 243–266 (2009)
10. Gulli, A., Pal, S.: Deep learning with Keras. Packt Publishing Ltd (2017)
11. Heffernan, N.T., Heffernan, C.L.: The assistments ecosystem: Building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching. International Journal of Artificial Intelligence in Education **24**, 470–497 (2014)
12. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.Y.: Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems **30** (2017)
13. Koedinger, K.R., Aleven, V.: Exploring the assistance dilemma in experiments with cognitive tutors. Educational Psychology Review **19**, 239–264 (2007)
14. Lamgarraj, M., Joiron, C., Parant, A., Dequen, G.: Exploring item difficulty prediction: Data driven approach for item difficulty estimation. In: International Conference on Intelligent Tutoring Systems. pp. 415–424. Springer (2024)
15. Liu, J., Yuan, H., Lu, X.M., Wang, X.: Quantum fisher information matrix and multiparameter estimation. Journal of Physics A: Mathematical and Theoretical **53**(2), 023001 (2020)
16. Piech, C., Bassen, J., Huang, J., Ganguli, S., Sahami, M., Guibas, L.J., Sohl-Dickstein, J.: Deep knowledge tracing. Advances in neural information processing systems **28** (2015)
17. Plateforme MatriCS: Plateforme matrics – université de picardie jules verne. https://www.matrics.u-picardie.fr/accueil/presentation-generale/ (2025), plateforme cofinancée par l'Union Européenne (FEDER) et le Conseil Régional des Hauts-de-France