Le raccrochage scolaire à distance : un projet innovant pour un enjeu de société

Meriem Zerkouk[1], Belkacem Chikhaoui[1], Richard Hotte[1], Shengrui Wang[2]

[1] **Université TÉLUQ:** 5800, rue Saint-Denis, bureau 1105, Montréal (Québec) H2S 3L5

[2] **Université de Sherbrooke:** 2500, boul. de l'Université, Sherbrooke (Québec) J1K 2R1

Résumé

Le décrochage et la réussite scolaire sont des enjeux sociétaux importants dans toutes les sociétés et en particulier au Québec. Pour faire face à ce problème, l'apprentissage à distance apparaît comme une solution adaptée à une partie importante de la population de décrocheurs scolaires. ChallengeU s'est donné comme mission de soutenir les centres de services scolaires (CSS) en leur proposant des services et une plateforme d'apprentissage à distance qui répond à leurs enjeux et de permettre aux raccrocheurs scolaires de réaliser leur plein potentiel. Grâce à cette plateforme d'apprentissage en ligne, ChallengeU a réussi à convaincre des décrocheurs à raccrocher et inciter des adultes ayant vécu de multiples échecs scolaires à reprendre leurs études autrement. Cependant, il reste beaucoup à faire pour augmenter le taux de diplomation de cette population. Pour ce faire, une collaboration entre ChallengeU, l'Université TÉLUQ et les centres de services scolaires du Lac-Saint-Jean et du Lac-Témiscamingue a vu le jour. Ainsi, un projet de recherche alliant intelligence artificielle et accompagnement psychosocial auprès des élèves raccrocheurs. Les résultats de ce projet de recherche sont probants. Le temps passé sur la plateforme, la progression dans les cours et la réussite de cours des élèves qui ont bénéficié des services d'accompagnement psychosocial à distance, développés et expérimentés par ChallengeU, ont été trois fois plus élevés que les métriques observées dans le groupe contrôle. Les résultats de cette recherche nous permettront de développer de nouveaux algorithmes efficaces d'analyse de données hétérogènes, multimodales et de prédiction pour supporter les services d'accompagnement des élèves et, aussi, pour augmenter les taux de diplomation au Québec.

Introduction

Le décrochage scolaire est un problème mondial qui a des conséquences négatives sur les individus, les familles et les communautés. Chaque année, environ 10 000 jeunes quittent les établissements scolaires

sans l'obtention d'un premier diplôme au Québec [1]. Le coût d'un décrocheur scolaire, au cours de sa vie, pour la société est de 120 000 $, soit 120 millions $ par cohorte annuelle. Ceci a des répercussions économiques importantes, car les personnes qui décrochent de l'école ont des difficultés à trouver un emploi et gagnent des revenus plus faibles tout au long de leur vie. Il est donc important de mettre en place des stratégies pour prévenir le décrochage scolaire et pour aider les raccrocheurs à persévérer dans leur retour aux études.

Les facteurs majeurs de décrochage scolaire peuvent inclure des facteurs internes tels que des difficultés d'apprentissage, des problèmes de santé mentale ou de comportement, ainsi que des facteurs externes tels que des difficultés financières, l'intimidation, des problèmes de logement ou de soutien familial insuffisant. L'enseignement à distance est devenu une solution importante et adaptée pour aider les décrocheurs à reprendre leurs études. En effet, il permet l'accès à l'éducation pour les élèves qui ne peuvent pas se rendre dans une école traditionnelle pour diverses raisons (distance, santé, situation familiale, travail, etc.). Cependant, il est important de noter que l'enseignement à distance ne résout pas de façon définitive le décrochage scolaire. D'ailleurs, les élèves suivant une formation à distance asynchrone autoportante sont susceptibles de manquer de motivation à un certain point au cours de leur projet d'études. C'est pourquoi il est important d'adapter les stratégies pédagogiques, motivationnelles et de soutien à l'enseignement à distance pour éviter les risques de décrochage dans ce contexte d'apprentissage.

Dans le cadre de ce projet de recherche, ChallengeU a offert des services d'accompagnement psychosocial à distance pour aider les élèves à surmonter les obstacles pouvant causer leur décrochage scolaire (accompagner, encourager, motiver et donner la confiance en soi). Le but est de suivre l'élève jusqu'à l'obtention de son diplôme.

Pour ce faire, notre vision pour prévenir le décrochage scolaire en formation à distance est basée sur une approche préventive qui repose sur deux axes principaux : 1) l'identification des élèves à risque de décrocher en utilisant des algorithmes d'intelligence artificielle; et 2) l'intégration d'un(e) intervenant(e) en persévérance scolaire auprès des élèves pour les assister.

Pour prévenir efficacement le décrochage scolaire et favoriser la persévérance, il est nécessaire d'identifier les facteurs clés du décrochage en se basant sur les techniques de l'intelligence artificielle, cibler les causes modifiables, connaître les interventions les plus efficaces pour fournir une aide pertinente aux élèves, évaluer la mise en œuvre de nos solutions (prédiction des décrocheurs) et sélectionner les élèves qui ont besoin d'accompagnement.

Nos recherches offrent une solution de prévention et de lutte contre le décrochage scolaire en utilisant des techniques d'analyse appliquées aux données de ChallengeU, et d'apprentissage automatique pour identifier les élèves à risque de décrochage et mettre en place des interventions ciblées pour les aider à persévérer dans leur parcours scolaire. En utilisant des données provenant de différentes sources, ChallengeU peut également évaluer l'efficacité de ses interventions et continuer à améliorer les résultats des élèves.

Cet article se divise en quatre sections. La première décrit les travaux connexes à la thématique du décrochage scolaire. La deuxième expose la méthodologie suivie pour l'analyse de données de ChallengeU. La troisième section porte sur la méthodologie suivie pour le développement des algorithmes d'intelligence artificielle. Pour finir, la quatrième présente et discute les résultats obtenus.

Travaux connexes

L'éducation en ligne, également connue sous le nom de formation à distance, offre une alternative flexible et accessible à l'apprentissage traditionnel en classe. Toutefois, il peut y avoir un taux élevé d'abandons scolaires chez les élèves en formation à distance. La prédiction du décrochage scolaire rallie les objectifs de recherche des milieux de l'éducation et de l'apprentissage machine. L'objectif spécifique de cette étude est d'analyser et étudier les décrocheurs dans leur environnement d'apprentissage en ligne en modélisant leur profil et comportements lorsqu'ils interagissent avec leur plateforme. Afin de montrer l'apport de notre contribution dans ce domaine, nous passons en revue un ensemble de travaux dont l'analyse est faite en se basant sur trois aspects principaux : la plateforme d'apprentissages et les données collectées, la modélisation des élèves (profil et comportement) et la méthode de prédiction. Ainsi, à l'issue de notre étude, l'objectif est de pouvoir identifier les facteurs qui contribuent au décrochage scolaire. Cela permettra de développer des algorithmes de prédiction efficaces pour aider les étudiants à risque de décrochage et de mettre en place des stratégies d'intervention mieux ciblées et plus pertinentes.

Plateforme d'apprentissages et données collectées

La collecte de données comportementales diffère d'un domaine à un autre. Les travaux [2,3] ont collecté des données en utilisant des questionnaires sur leurs plateformes d'apprentissage. Dans d'autres projets, une partie importante de la collecte de données s'est faite de façon automatique à partir des données comportementales tirées des bases de données de la plateforme d'apprentissages des élèves. En effet, ces outils permettent la collecte automatique de plusieurs types de données, notamment les cliques souris, le temps passé sur la plateforme, le nombre de connexions, la progression des élèves dans leur cours, le visionnement des vidéos d'apprentissage, etc.

Modélisation des profils des élèves

Plusieurs travaux se sont intéressés à la modélisation du profil des élèves afin d'identifier les différentes catégories d'élèves. Cela permet 1) de connaître le profil type de la catégorie des élèves qui réussissent, et 2) de faire des interventions personnalisées pour la catégorie des élèves à risque de décrochage. Pour modéliser le profil des élèves, les chercheurs [4,5] ont combiné plusieurs types de données, à savoir les données sociodémographiques, le cursus scolaire et l'historique du travail sur la plateforme d'apprentissages. D'autres chercheurs [6,7] ont utilisé des données comme le sexe, la race, la géolocalisation et l'occupation. Toutes ces données sont regroupées en informations sociodémographiques.

Méthodes d'identification des élèves à risque de décrochage

Les méthodes d'identification des élèves à risque de décrochage se divisent en deux catégories : l'analyse statistique de données et l'apprentissage automatique, qui est une branche de l'intelligence artificielle. D'abord, il existe les méthodes d'analyse statistique des données [6,7], où la plupart des travaux consistent à étudier la corrélation entre un attribut (ou une caractéristique) spécifique et le résultat du décrochage. Ces

méthodes ne permettent pas de généraliser les causes du décrochage scolaire. Ensuite, l'apprentissage automatique [7,8] requiert l'utilisation d'algorithmes d'intelligence artificielle, qui sont disponibles et appropriés aux plateformes d'apprentissages en ligne.

Notre contribution dans ce projet réside dans le fait de combiner des données sociodémographiques et comportementales afin de comprendre en profondeur les causes du décrochage. Cela facilite le développement des techniques d'apprentissages automatiques pour permettre de prédire les élèves à risque de décrochage. L'intervenante aura ensuite un rôle dans la mise en place d'une stratégie de suivi et de soutien aux élèves pour stimuler la persévérance scolaire.

Méthodologie - Analyse de données

Dans cette section, nous allons introduire la méthodologie utilisée pour répondre aux objectifs du projet.

Description des données de ChallengeU

ChallengeU dispose de données empiriques très riches s'étendant de 2016 à 2022. La collecte de données est possible grâce à l'application des meilleures pratiques de protection des données personnelles et sur le respect des politiques de vie privée.

Le serveur de base de données de ChallengeU stocke des données empiriques anonymisées collectées qui couvrent la période de 2016 à 2022. Ces données permettent d'effectuer des analyses statistiques et de mettre en application des techniques d'apprentissage machine pour la prédiction.

Ces données proviennent de diverses sources, telles que les dossiers d'inscription des élèves, des fichiers historiques et les réponses aux activités. Les données peuvent être organisées en données sociodémographiques et comportementales.

Données sociodémographiques : ces données sont fondamentales et déterminent la situation sociale de différents types d'élèves. Ce type de données peut influencer les comportements et les attitudes des élèves sur leur persévérance. Ces données permettent d'avoir une meilleure compréhension de la situation sociale de chaque élève.

Données comportementales : dans le cadre d'apprentissage en ligne sur la plateforme de ChallengeU, les données comportementales incluent, entre autres, des informations sur les interactions de l'élève avec la plateforme telles que la fréquence de connexion, la durée de chaque session, les cours assignés, les activités pédagogiques complétées, le type d'appareil utilisé (téléphone, ordinateur ou tablette). Ces données permettent de fournir un portrait complet de la façon dont l'élève utilise la plateforme.

Nous avons combiné toutes les données présentées ci-dessus pour avoir un seul ensemble de données contenant toutes les informations qui caractérisent les élèves et qui sont nécessaires à la fois à l'analyse

exploratoire des données, au profilage des élèves et au développement des algorithmes de prédiction. Ces données seront utilisées pour identifier les facteurs liés à la persévérance des élèves et au décrochage scolaire.

Analyse exploratoire des données

L'analyse exploratoire des données est une condition préalable importante avant la modélisation et la prédiction. Le but de cette première exploration est de comprendre les tendances, les corrélation et les relations entre les différents attributs dans les données. L'exploration des données consiste à utiliser des techniques de statistiques descriptives pour comprendre les caractéristiques et les relations telle que 1) l'analyse univariée qui consiste à étudier chaque variable (attribut) séparément; 2) l'analyse bivariée qui consiste à étudier l'effet de deux variables sur une troisième; et 3) l'analyse multivariée qui repose sur l'étude simultanée de l'effet de plusieurs variables.

Nous avons utilisé une matrice de corrélation qui permet de mesurer la relation linéaire entre les différentes variables, ce qui permet de déterminer quelles variables sont les plus corrélées et, donc, les plus importantes pour la modélisation et la prédiction. Les valeurs dans la matrice de corrélation varient de -1 à 1, avec une valeur proche de 1 indiquant une forte corrélation positive entre les variables. Une valeur proche de -1 indique une corrélation négative entre les variables, et une valeur nulle ou proche de 0 indique l'absence de relation entre les variables.

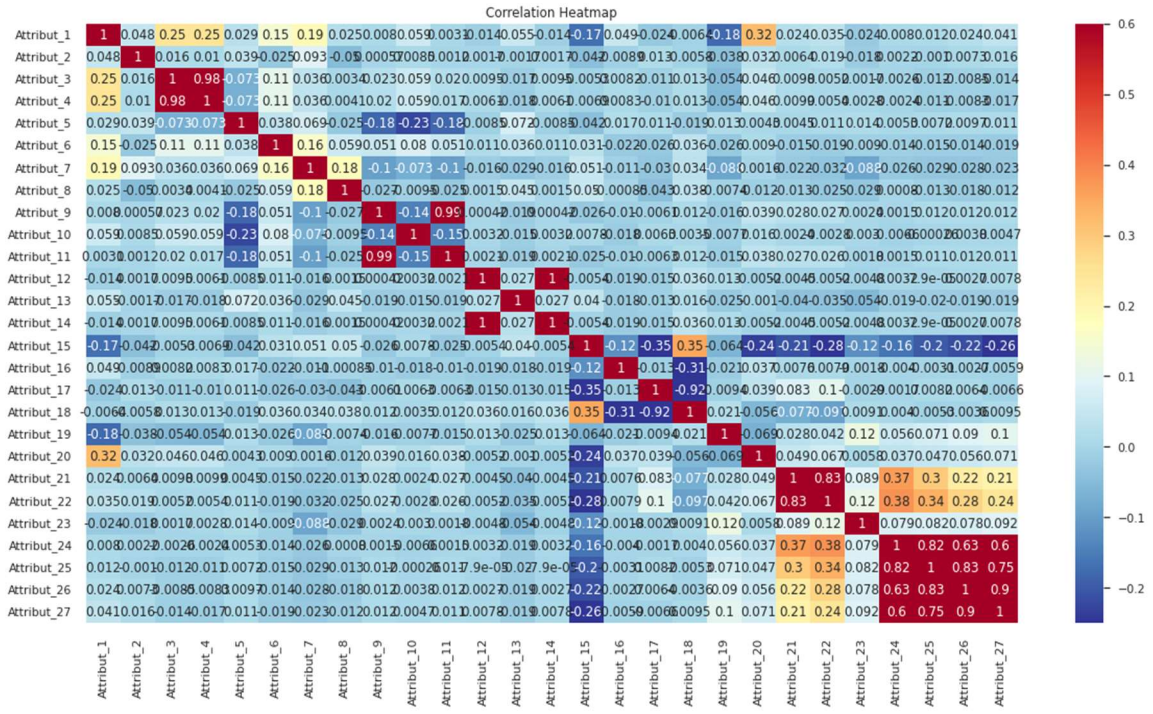

Figure 1. Matrice de corrélation avec des variables (attributs) anonymes.

Cette matrice de corrélation nous permet de déterminer quels attributs sont hautement dépendants et ceux qui ne le sont pas.

Expérience de pairage des élèves avec l'intervenante

Identification de la population d'élèves impliqués dans l'étude

D'abord, afin de déterminer la population des élèves qui serait à l'étude, ChallengeU a consulté l'ensemble des centres de services scolaires partenaires pour identifier ceux qui étaient intéressés à mettre en place cette phase expérimentale de service psychosocial à offrir aux élèves. Ainsi, le premier critère de sélection était que les élèves soient issus des centres de services scolaires du Lac-Saint-Jean ou du Lac-Témiscamingue.

Ensuite, puisque les données empiriques de ChallengeU démontrent que les élèves inscrits à des cours de français étaient surreprésentés par rapport aux élèves qui montrent des difficultés de persévérance scolaire rapidement dans leur parcours, nous avons décidé de sélectionner des élèves suivant des cours de français sur la plateforme (Français 4e ou 5e secondaire).

Méthodologie de recrutement

Une fois la population à l'étude définie, il fallait commencer à former le groupe d'élèves qui allait participer à l'expérience. L'équipe de ChallengeU a préparé une vidéo de présentation du projet, accompagnée d'un formulaire de demande de participation et d'un formulaire de consentement à la collecte et l'utilisation de leurs données personnelles. Ces éléments ont été envoyés aux élèves par vagues de 500 invitations afin de former le groupe d'élèves pour l'expérience. Après trois vagues de 500 invitations, les invitations ont été réduites à des vagues de 100, qui, jusqu'à ce point dans le recrutement, étaient sous-représentées dans l'échantillon de l'expérience. Les variables considérées pour la représentativité du groupe étaient les suivantes :

- Centre de services scolaire;
- Genre;
- Âge;
- Sigle assigné à l'élève;
- Objectif à court terme de l'élève avec ses études;
- Progression de l'élève dans son sigle actif.

Nous pouvons voir ci-dessous, à gauche, la composition de l'échantillon retenu pour l'expérimentation et, à droite, les données de la population à l'étude.

PROJET		
Recrutés TOTAL	**78**	
Attribut_1	**Nombre**	**%**
Attribut_1.valeur_1	38	47,5%
Attribut_1.valeur_2	42	52,5%
Attribut_2	33,60 ans	
Attribut_2.valeur_1	26,00 ans	
Attribut_3		
Attribut_3.valeur_1	47	60,3%
Attribut_3.valeur_2	31	39,7%
Attribut_4		
Attribut_4.valeur_1	41	52,6%
Attribut_4.valeur_2	18	23,1%
Attribut_4		
0%	1	1,3%
1-10%	3	3,8%
10-30%	4	5,1%
30-50%	20	25,6%
50-80%	26	33,3%
80% et plus	24	30,8%
Attribut_5		
Attribut_6.valeur_1	21	26,9%
Attribut_6.valeur_2	2	15,4%
Attribut_6.valeur_3	14	17,9%

DATASET	
Candidats TOTAL	**2469**
Attribut_1	**%**
Attribut_1.valeur_1	45%
Attribut_1.valeur_2	55%
Attribut_2	29,13 ans
Attribut_2.valeur_1	27 ans
Attribut_3	
Attribut_3.valeur_1	56%
Attribut_3.valeur_2	44%
Attribut_4	
Attribut_4.valeur_1	54%
Attribut_4.valeur_2	23%
Attribut_5	
0%	9%
1-10%	10%
10-30%	10%
30-50%	23%
50-80%	28%
80% et plus	21%
Attribut_6	
Attribut_6.valeur_1	27%
Attribut_6.valeur_2	19%
Attribut_6.valeur_3	18%

Figure 2. Population d'élèves impliqués dans l'expérimentation.

Ainsi, le recrutement des élèves a évolué sur quelques semaines pour assurer une représentativité de la population à l'étude. Par exemple, les derniers élèves recrutés ont été surreprésentés parmi les hommes et les élèves de moins de 25 ans pour compenser le biais de la méthode de recrutement qui semblait naturellement surreprésenter les femmes et les élèves de plus de 35 ans.

Finalement, il est important de préciser que, bien que 95 élèves aient rempli les demandes d'adhésion à l'expérience, seulement 80 élèves ont été retenus parce que les 15 autres élèves ont échoué de remplir la condition obligatoire de se présenter à une séance d'information en début de projet, et ce malgré la possibilité de participer à l'une de cinq séances offertes avec des plages horaires flexibles comme le matin, à midi, le soir, et la fin de semaine.

Expérimentation et travail de l'intervenante

Au cours du projet, la structure du travail de l'intervenante en persévérance scolaire se divisait en deux phases : la phase d'exploration et la phase d'accompagnement des élèves.

Phase d'exploration : dans la phase d'exploration, l'objectif de l'intervenante était d'apprendre à connaître l'élève, à créer un lien et, surtout, à identifier son ou ses besoins prioritaires.

Pour ce faire, l'intervenante a conçu un modèle de fiche d'évaluation des besoins prioritaires des élèves, qui était remplie par l'intervenante en fonction :

- Des informations partagées dans le cadre d'entrevues individuelles;

- Des résultats des élèves à un sondage en ligne (obligatoire) sur l'historique scolaire de l'élève;
- Des résultats des élèves à un sondage en ligne (obligatoire) sur la personnalité de l'élève.

Ensuite, les élèves ont complété le test psychométrique ÉMÉ-S afin de mettre en lumière le type de motivation prédominante de l'élève (intrinsèque ou extrinsèque) liée à la scolarisation. Ces tests ont permis d'établir des stratégies d'interventions personnalisées avec les élèves. Le fait de pouvoir comparer les résultats des élèves de l'expérience permettait de relativiser les résultats auprès de la clientèle particulière à cette étude.

Phase d'accompagnement : après la première rencontre individuelle des élèves avec l'intervenante, ces derniers ont complété leur test psychométrique sur la motivation liée à la scolarisation. À cette étape, il importait de dresser un plan d'accompagnement adapté au besoin de l'élève.

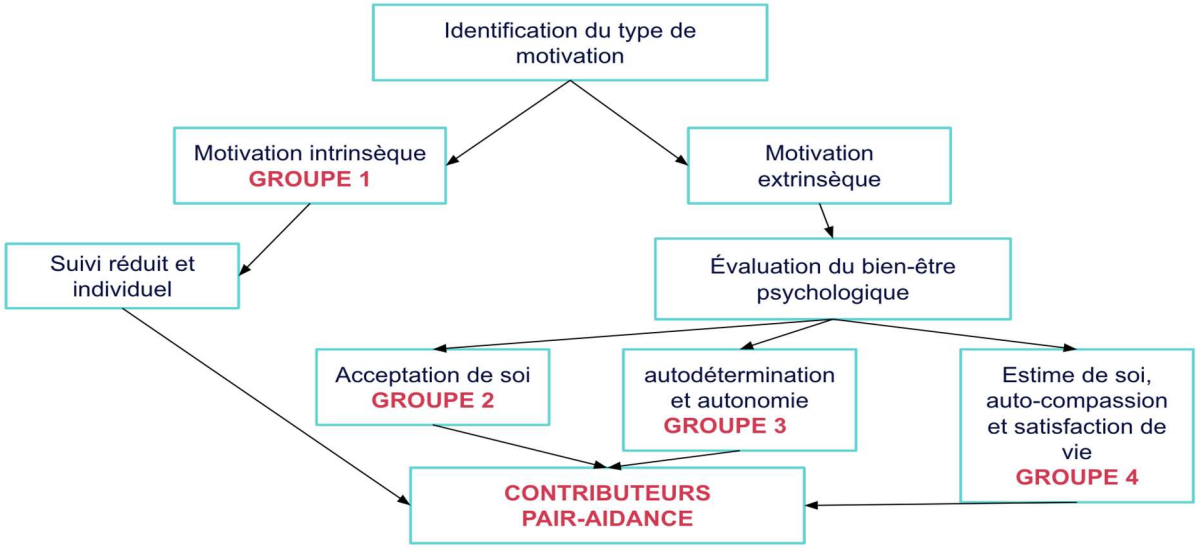

Figure 3. *Structure d'accompagnement offert aux élèves*

Le schéma ci-dessus présente la structure d'accompagnement offerte aux élèves du projet. Nous avons statué que :

- Les élèves avec une motivation principalement intrinsèque sont ceux qui nécessitent le moins de soutien;
- Les élèves avec une motivation principalement extrinsèque nécessitent une évaluation plus approfondie pour comprendre l'ampleur de leur problématique. Ainsi, une série de tests psychométriques additionnels ont été utilisés par l'intervenante :
    - L'échelle d'Estime de soi de Rosenberg;
    - L'échelle de satisfaction de vie (ÉSV);

    - L'échelle du bien-être psychologique (RYFF);
    - Test sur l'autocompassion (NEFF).
- En fonction des résultats aux différents tests, l'intervenante prenait le soin de classer les élèves dans l'un des 4 groupes d'élèves qui représentent chacun la problématique principale reliée à la scolarisation de l'élève.

Les élèves de chaque groupe avaient donc des particularités. Alors que les élèves du Groupe 1 avaient généralement une rencontre par mois avec l'intervenante, les élèves du Groupe 4 avaient systématiquement au moins une rencontre par semaine. En plus de ces rencontres planifiées à des fréquences différentes selon les groupes, l'intervenante contactait régulièrement les élèves tout au long de l'expérimentation par des modes de communication différents : conversations Messenger, SMS, appels téléphoniques, courriels, selon les préférences des élèves.

Finalement, en plus des rencontres et des communications individuelles entre l'intervenante et les élèves, plusieurs autres services ont été mis en place au bénéfice des élèves :

- Un groupe Facebook et des café-causeries pour permettre aux élèves participant au projet de partager leurs expériences;
- Des ateliers sur des sujets connexes à leur projet d'études identifiés avec le test RYFF (1 sur l'estime de soi, 1 sur le stress et l'anxiété);
- Une proposition de pair-aidance pour construire un réseau d'entraide en dyade;
- Une rencontre en personne avec l'intervenante, qui a fait le tour du Québec pour aller à la rencontre des élèves intéressés.

Méthodologie - Développement des algorithmes

Profilage des élèves

Pour identifier les différents profils des élèves, le regroupement (*clustering*) est utilisé pour regrouper les élèves en fonction des caractéristiques ou de comportements similaires. Notre but est de découvrir les profils complexes des élèves, en regroupant les élèves en fonction de leurs activités et de leurs engagements sur la plateforme.

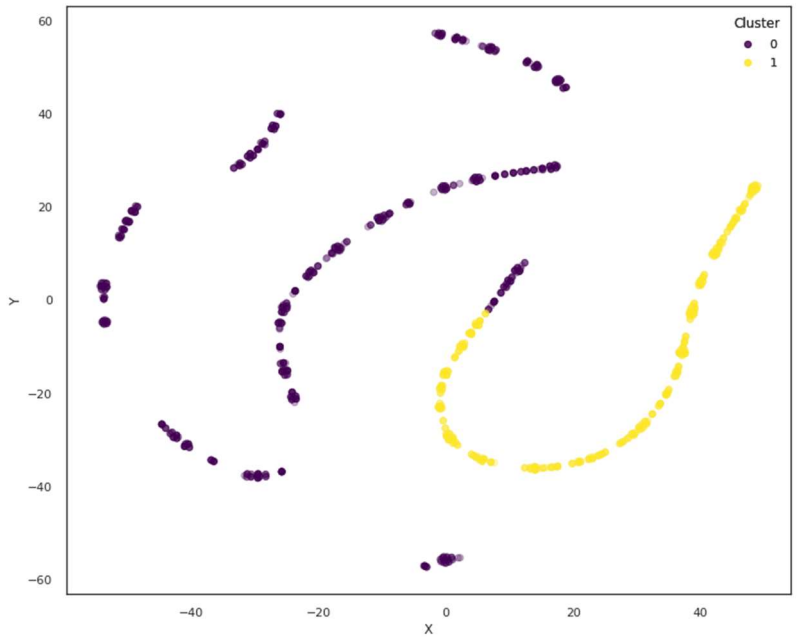

Figure 4. *Les regroupements (clusters) obtenus.*

Le profilage des élèves nous a permis d'identifier les deux profils principaux d'élèves qui existent dans le jeu de données (décrocheurs et actifs). Le groupe (cluster) en jaune indique les élèves non décrocheurs et l'autre représente les décrocheurs. Par conséquent, l'élève à ChallengeU peut avoir le profil décrocheur ou actif suivant son statut.

Développement des algorithmes de prédiction

La prédiction désigne l'utilisation des algorithmes d'apprentissage automatique pour prévoir les comportements futurs des élèves à partir de données sociodémographiques et comportementales observées. Nous avons développé 16 algorithmes de prédiction, à savoir le classificateur factice (dummy classifier), la régression logistique, les forêts aléatoires (random forests), les machines à vecteurs de support (svm), etc. Dans ce qui suit, et par des contraintes d'espace, nous allons présenter uniquement trois de ces algorithmes.

Pour développer ces techniques d'apprentissage machine avec les données de ChallengeU, il fallait d'abord définir l'attribut cible, aussi appelé classe, par exemple : décrocheur. Ensuite, il importait de diviser les données en un ensemble "d'apprentissage" et un ensemble de "tests" pour l'évaluation des algorithmes. Ensuite, les algorithmes ont été entraînés sur l'ensemble d'apprentissage afin d'apprendre les caractéristiques de chaque classe (décrocheur vs non-décrocheur). Puisque la classe des décrocheurs représente une minorité dans notre ensemble de données, nous avons opté pour des techniques d'augmentation des données telle que la méthode SMOTE (Synthetic Minority Over-sampling Technique) pour un suréchantillonnage synthétique de la classe minoritaire. L'augmentation des données vise à assurer que les algorithmes ne seront pas biaisés par la classe majoritaire. Les performances des 3 algorithmes sont

ensuite évaluées sur les données de test à l'aide des mesures de performances telles que l'exactitude, la précision, le rappel et le score F1.

Dummy classifier : peut servir comme un modèle d'analyse des données pour identifier les élèves susceptibles de décrocher. Les résultats obtenus avec ce modèle peuvent être utilisés pour comparer les performances des autres modèles d'apprentissage automatique plus complexes pour prédire le décrochage scolaire.

Régression logistique *:* peut servir pour prévoir notre variable cible binaire (décrocheur/non décrocheur) à partir de plusieurs attributs indépendants. En utilisant les données de ChallengeU et créant des profils d'élèves, il est possible d'utiliser une régression logistique pour prédire les risques de décrochage scolaire chez les élèves en se basant sur des attributs tels que le profil sociodémographique, et le comportement sur la plateforme.

Random forest : il est considéré comme l'un des modèles d'apprentissage automatique les plus puissants et robustes. On l'appelle "forêt" car il s'agit de créer plusieurs arbres de décision (d'où le nom de forêt) et de les utiliser pour faire des prédictions. L'idée principale d'une forêt aléatoire est de combiner les prédictions de plusieurs arbres de décision afin d'améliorer les performances globales du modèle. L'une des façons d'y parvenir est de faire la moyenne des prédictions de chaque arbre, mais d'autres techniques peuvent également être utilisées.

Le choix entre les modèles exploités était basé sur les spécificités des données de ChallengeU. Notre ensemble de données dispose d'un nombre important d'attributs.

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
dt	Decision Tree Classifier	0.7981	0.7991	0.7583	0.7482	0.7508	0.5814	0.5839	0.0030
lightgbm	Light Gradient Boosting Machine	0.7981	0.8625	0.7743	0.7456	0.7548	0.5840	0.5894	0.0110
rf	Random Forest Classifier	0.7918	0.8620	0.7705	0.7315	0.7484	0.5713	0.5742	0.0430
et	Extra Trees Classifier	0.7917	0.8426	0.7783	0.7258	0.7497	0.5718	0.5746	0.0380
xgboost	Extreme Gradient Boosting	0.7870	0.8614	0.7425	0.7352	0.7360	0.5578	0.5606	0.0640
gbc	Gradient Boosting Classifier	0.7695	0.8451	0.7672	0.6919	0.7267	0.5284	0.5317	0.0220
catboost	CatBoost Classifier	0.7664	0.8425	0.7666	0.6884	0.7239	0.5225	0.5269	1.8560
ridge	Ridge Classifier	0.6805	0.0000	0.7243	0.5825	0.6433	0.3601	0.3709	0.0020
lda	Linear Discriminant Analysis	0.6741	0.7394	0.7165	0.5769	0.6370	0.3478	0.3575	0.0040
lr	Logistic Regression	0.6709	0.7482	0.6728	0.5783	0.6196	0.3328	0.3386	0.2840
ada	Ada Boost Classifier	0.6582	0.7407	0.6125	0.5740	0.5895	0.2979	0.3008	0.0160
qda	Quadratic Discriminant Analysis	0.6215	0.6088	0.5454	0.5334	0.5176	0.2151	0.2245	0.0030
nb	Naive Bayes	0.6170	0.6849	0.2229	0.5769	0.2684	0.1093	0.1390	0.0030
svm	SVM - Linear Kernel	0.6056	0.0000	0.2400	0.4815	0.2698	0.0944	0.1292	0.0030
dummy	Dummy Classifier	0.5978	0.5000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0080
knn	K Neighbors Classifier	0.5644	0.5253	0.3988	0.4562	0.4231	0.0761	0.0778	0.1550

Tableau 1. *Comparaisons des résultats obtenus*.

Les performances des modèles utilisés sont ensuite évaluées sur les données de test à l'aide de mesures telles que la précision, le rappel et le score F1. On note bien que l'algorithme LightGBM donne le meilleur score F1.

Évaluation de l'importance des attributs

Après avoir entraîné les modèles de prédiction, nous avons choisi les méthodes suivantes : SHAP (SHapley Additive exPlanations) et LIME (Local Interpretable Model-Agnostic Explanations) pour visualiser les contributions de chaque attribut à la prédiction des élèves qui sont à risque de décrochage, et à identifier les attributs les plus importants pour le modèle. L'importance de chaque attribut peut être déterminée en utilisant des métriques telles que la somme des contributions absolues ou la moyenne des contributions pour tous les exemples dans les données.

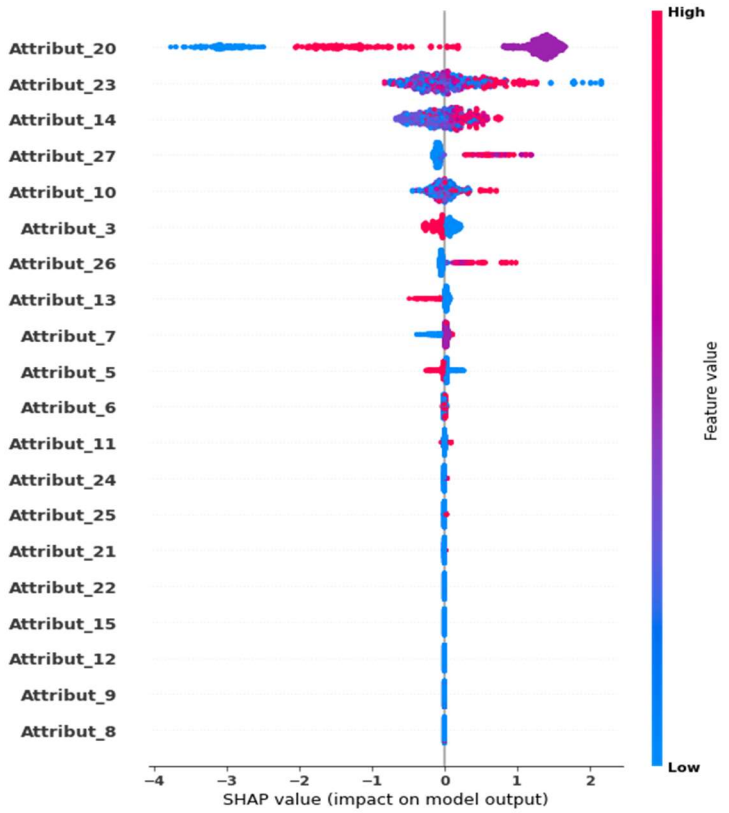

Figure 5. *Classement des attributs importants dans la prédiction (attributs anonymes)*.

Chaque valeur de SHAP représente l'impact attendu d'un attribut sur la sortie du modèle, en supposant que tous les autres attributs ont leurs valeurs moyennes. La figure 5 ci-dessus montre les variables qui ont une contribution positive très élevée lorsque ses valeurs sont élevées, et une faible contribution négative lorsque ses valeurs sont faibles. Toutes les variables sont présentées dans l'ordre d'importance de la caractéristique globale, la première étant la plus importante et la dernière la moins importante.

Résultats et évaluation du rôle de l'intervenante

Le tableau suivant présente les résultats d'intervention auprès des élèves ayant participé au projet par rapport aux élèves du groupe contrôle (qui n'ont pas participé au projet). Les résultats obtenus sont très probants. Le temps passé sur la plateforme, la progression dans les cours et la réussite de cours des élèves qui ont bénéficié des services d'accompagnement psychosocial à distance développés et expérimentés par ChallengeU ont tous augmenté de façon significative comparativement au groupe contrôle. De la même façon, le taux d'abandon a diminué de plus de 50%. Cela démontre clairement le rôle de l'intervenante en persévérance scolaire et l'impact de ses interventions dans le maintien des activités des élèves et l'amélioration de leur parcours académique.

	Groupe Contrôle	**Élèves admis au projet**	**Élèves du projet**	**Analyse**
% Abandon scolaire	66.60%	38.75%	32.81%	La proportion d'élèves ayant terminé leur parcours scolaire avec ChallengeU a été 3x plus élevée auprès des élèves du projet que du groupe contrôle (7.81% c. 2,50%). Alors que la majorité des élèves qui ont participé au groupe sont toujours actifs dans leurs études (59.38%), on peut voir que les élèves qui n'ont pas participé au projet ont majoritairement quitté leur projet d'études (66.60%). On peut donc voir que, dans un contexte où les élèves ont plusieurs cours à compléter dans leur parcours, l'implication de l'intervenante en persévérance scolaire a un impact majeur pour maintenir en vie l'espoir sur une plus longue durée.
% Aux études au 31 décembre 2022	30.90%	53.75%	59.38%	
% études complétées	2.50%	7.50%	7.81%	
% élèves ayant réussi au moins 1 sigle	9.20%	23.75%	28.13%	Un peu plus de 3x plus d'élèves ayant bénéficié d'un accompagnement ont réussi au moins un sigle dans leur parcours scolaire entre le début du projet et la fin de l'année 2022. Puisqu'encore 59.38% des élèves sont toujours actifs dans leur projet d'études (comparativement à 30.90% pour les élèves du groupe contrôle), on peut s'attendre à ce que ce ratio augmente drastiquement au cours de l'année 2023.
Ratio sigles réussis par élève	0.155	0.425	0.515	Le ratio du nombre de sigles réussis par les élèves ayant participé au projet est plus de 3x plus élevé que celui des élèves du groupe contrôle. Puisqu'encore 59.38% des élèves sont toujours actifs dans leur projet d'études (comparativement à 30.90% pour les élèves du groupe contrôle), on peut s'attendre à ce que ce ratio augmente drastiquement au cours de l'année 2023.

Conclusion

Dans ce projet, nous avons montré l'apport de l'intelligence artificielle pour identifier les élèves à risque de décrochage, et le rôle de l'intervenante en persévérance scolaire dans la réussite des élèves raccrocheurs. L'importance du suivi par une intervenante a été soulevée afin d'aider les élèves à persévérer et d'améliorer les résultats scolaires. Une compréhension des causes du décrochage scolaire et l'élaboration de stratégies plus efficaces de prévention peut contribuer considérablement à la réussite scolaire des élèves utilisant des systèmes d'apprentissage en ligne. Les résultats obtenus dans le cadre de ce projet sont probants, et démontrent le grand potentiel de ce projet à augmenter le taux de diplomation des élèves au Québec.

Maintenant qu'on comprend que l'intelligence artificielle a un rôle à jouer dans le dépistage du risque de décrochage scolaire et qu'on sait que le suivi d'un(e) intervenant(e) en persévérance scolaire a un impact important sur la persévérance et la réussite scolaire des élèves, il serait intéressant de mener des expériences où les intervenant(e)s en persévérance scolaire utiliseraient les outils d'intelligence artificielle développés pour intervenir plus rapidement lorsque des élèves montreraient des risques de décrochage.

Références

1. Plus de 10 000 décrocheurs scolaires au Québec/journal le devoir

2. Bardh Prenkaj, Paola Velardi, Giovanni Stilo, Damiano Distante, and Stefano Faralli. 2020. A Survey of Machine Learning Approaches for Student Dropout Prediction in Online Courses. ACM Comput. Surv. 53, 3, Article 57 (May 2021), 34 pages. https://doi.org/10.1145/3388792

3. Pereira, F.D. et al. (2019). Early Dropout Prediction for Programming Courses Supported by Online Judges. In: Isotani, S., Millán, E., Ogan, A., Hastings, P., McLaren, B., Luckin, R. (eds) Artificial Intelligence in Education. AIED 2019. Lecture Notes in Computer Science(), vol 11626. Springer, Cham. https://doi.org/10.1007/978-3-030-23207-8_13

4. Del Bonifro, F., Gabbrielli, M., Lisanti, G., Zingaro, S.P. (2020). Student Dropout Prediction. In: Bittencourt, I., Cukurova, M., Muldner, K., Luckin, R., Millán, E. (eds) Artificial Intelligence in Education. AIED 2020. Lecture Notes in Computer Science(), vol 12163. Springer, Cham. https://doi.org/10.1007/978-3-030-52237-7_11

5. Kemper, Lorenz et al. "Predicting student dropout: A machine learning approach." European Journal of Higher Education 10 (2020): 28 - 47.

6. Solís, M., Moreira, T.M., Gonzalez, R., Fernandez, T., & Hernandez, M. (2018). Perspectives to Predict Dropout in University Students with Machine Learning. 2018 IEEE International Work Conference on Bioinspired Intelligence (IWOBI), 1-6.

7. Gardner, J., & Brooks, C. (2018). Student success prediction in MOOCs. User Modeling and User-Adapted Interaction, 28(2), 127–203. doi:10.1007/s11257-018-9203-z (modeling)

8. Willging, Pedro A. and Scott D. Johnson. “FACTORS THAT INFLUENCE STUDENTS’ DECISION TO DROPOUT OF ONLINE COURSES.” Online Learning (2019).