Conseil ontarien de la qualité de l'enseignement supérieur

Un organisme du gouvernement de l'Ontario

Utilisation de la modélisation prédictive afin de guider le dépistage précoce et les interventions d'aide pédagogique perturbatrice et de rehausser le maintien aux études

Ross Finnie, Tim Fricker, Eda Bozkurt,
Wayne Poirier, Dejan Pavlic

Publié par le

Conseil ontarien de la qualité de l'enseignement supérieur

1, rue Yonge, bureau 2402
Toronto (Ont.) Canada, M5E 1E5

Téléphone : 416 212-3893
Télécopieur : 416 212-3899
Site Web : www.heqco.ca
Courriel : info@heqco.ca

Citer ce document comme suit :

Finnie, R., Fricker, T., Bozkurt, E., Poirier, W. et Pavlic, D. (2017) *Utilisation de la modélisation prédictive afin de guider le dépistage précoce et les interventions d'aide pédagogique perturbatrice et de rehausser le maintien aux études*, Toronto, Conseil ontarien de la qualité de l'enseignement supérieur.

Conseil ontarien de la qualité de l'enseignement supérieur

Remerciements

Le présent projet a fait intervenir des dizaines de personnes qui, chacune, ont joué un rôle important, et la concrétisation du présent rapport aurait été impossible sans leur participation et leur appui.

Premièrement, aux membres du personnel du Collège Mohawk :

- Gary Jennings, Hetal Patel et Craig Cook qui ont animé le programme Évaluations de réussite (EDR). Il s'agit d'un programme fabuleux qui comporte des données extraordinaires et procure aux élèves un appui précieux.

- Brian Maloney, George Rombes, Helen Sheridan et Shannon Gould, du service des rapports et de la recherche institutionnelle du Collège. Ils ont chacun consacré de nombreuses heures à la gestion et à l'épuration des données ainsi qu'à l'organisation du Sondage d'accueil des élèves (SAE).

- Les conseillers à la réussite des élèves de même que Barb Russell qui, collectivement, ont effectué un changement colossal à leurs méthodes pour adopter le logiciel ClockWork et faire le suivi des données des services d'aide pédagogique.

- Le Groupe administratif du Collège Mohawk (antérieur et actuel), pour son appui continu envers notre travail.

- Les autres membres de l'équipe de recherche du Collège Mohawk qui ne sont pas répertoriés comme auteurs mais dont le rôle a été important sur le plan opérationnel, notamment Rachel Matthews et Megan Pratt.

Deuxièmement, à l'équipe de l'IRPE :

- Kaveh Afshar, pour ses observations et sa contribution à l'avancement du projet à l'IRPE, tout en jouant le rôle de gestionnaire de projet à l'IRPE.

- Michael Dubois, gestionnaire de projet à l'IRPE, pour avoir donné suite aux efforts de Kaveh de façon à contribuer aux ensembles définitifs de révisions, et pour l'orientation donnée durant les derniers stades de production du document.

- John Sergeant pour ses talents en administration, lesquels ont en outre contribué à maintenir le projet sur la bonne voie.

- Les autres membres de l'équipe de recherche de l'IRPE, lesquels ont contribué aux séances de remue-méninges ou présenté des suggestions relatives à la recherche.

Troisièmement, au COQES pour le financement des travaux et la mise en place du Consortium sur l'accès et la persévérance scolaire. La démarche du COQES axée sur la collaboration et le soutien envers la recherche s'est révélée très utile. Globalement, l'IRPE tient à remercier le COQES de l'appui accordé à une gamme de projets au fil des ans. Sans l'appui du COQES, l'IRPE ne serait pas ce qu'il est maintenant.

Enfin, aux autres membres du consortium pour leur rétroaction et leurs suggestions quant à la conception du projet.

Synthèse

Origines et aperçu du projet

En 2012, le Collège Mohawk a demandé l'appui de l'Initiative de recherche sur les politiques et l'éducation (IRPE) afin de recueillir et d'utiliser les données, notamment administratives, détenues par le Collège Mohawk à propos des élèves, dans le contexte d'une vaste initiative servant à améliorer la réussite des élèves selon le principe de la prise de décisions factuelles.

Le premier projet a fait appel à des analyses en vue de bien concevoir le maintien aux études des élèves au Collège Mohawk à l'aide de méthodes de modélisation descriptive et statistique. Il en a résulté l'élaboration d'un modèle prédictif permettant de discerner les élèves à risque de décrocher des études collégiales.

En 2015, le Collège Mohawk et l'IRPE ont présenté une demande d'admission au Consortium sur l'accès et la persévérance scolaire (CAPS) du Conseil ontarien de la qualité de l'enseignement supérieur (COQES), laquelle a été reçue, en vue de mener un projet qui allait s'appuyer sur le travail réalisé antérieurement. Le but du projet consistait à actualiser le modèle prédictif, à le peaufiner et à en faire des essais poussés, ce qui allait ensuite servir à étoffer et à évaluer un ensemble d'interventions parallèles d'aide pédagogique mises en place pour les élèves qui amorçaient leurs études au Collège Mohawk à la session d'automne 2015.

Dans l'ensemble, l'étude comporte deux phases d'enquête correspondantes. La phase 1, laquelle est l'objet du présent rapport, correspond au travail mené par l'IRPE aux fins suivantes : poursuivre le développement du modèle prédictif du maintien aux études des élèves au Collège Mohawk; mettre à l'essai le modèle prédictif de même que les prédictions qu'il génère; utiliser ce modèle pour générer des prédictions quant à l'éventualité qu'un nouvel élève décroche de ses études au Collège Mohawk; utiliser les prédictions obtenues pour répartir les élèves en trois différents groupes de risque (à fort risque, à risque moyen, à faible risque); analyser la ventilation des élèves dans ces différents groupes en fonction d'une gamme de caractéristiques, notamment à l'échelle individuelle et des programmes; puis calculer les taux de participation aux programmes actuels d'aide pédagogique aux élèves chez les trois groupes de risque.

La phase 2 (et le second rapport) consistera en une analyse des interventions particulières d'aide pédagogique offertes aux élèves de la cohorte arrivée en 2015, notamment en quoi les effets de ces interventions varient entre les groupes de risque susmentionnés, selon ce que le modèle prédictif permet de déceler.

De plus, le présent document ne se contente pas de traiter l'élaboration et la mise en application du modèle prédictif et de l'analyse qui s'y rapporte : il présente également des notions de base sur l'utilisation de la modélisation prédictive dans le contexte de l'analyse prédictive et du maintien aux études des élèves. La motivation liée à ces notions de base découle de l'usage répandu du concept de « modélisation prédictive » s'appliquant au maintien aux études des élèves, ainsi que de l'utilité potentielle du discernement des concepts pertinents.

Le présent document s'amorce donc par une discussion en des termes généraux de ce qui constitue une démarche de modélisation prédictive dans le contexte du maintien aux études des élèves dans l'enseignement postsecondaire (EPS). Nous tentons ensuite d'expliquer ce que cela signifie par une description de chaque étape qui intervient dans la création du modèle prédictif. Les étapes sont les suivantes : la spécification du modèle; l'estimation du modèle; la mise à l'essai initiale du modèle; et l'utilisation du modèle pour générer les taux prédits de décrochage à l'échelle individuelle. Les étapes subséquentes font intervenir une mise à l'essai approfondie du modèle pour déterminer l'exactitude de ces précisions et, au bout du compte, le modèle sert à répartir les élèves du Collège Mohawk en trois groupes selon le niveau de risque.

Principales constatations

Voici quelques-unes des constatations particulières :

- Les principaux facteurs déterminants de décrochage des études au Collège Mohawk avant l'obtention du diplôme, discernés au moyen du modèle prédictif, sont les suivants : les élèves de sexe féminin, ceux de 23 ans et plus[1] et ceux au certificat d'études supérieures présentent des taux de décrochage inférieurs à ceux des autres élèves; les élèves au certificat et ceux dont les notes en début de programme sont basses (tout particulièrement si elles sont égales ou inférieures à D+) montrent des taux supérieurs de décrochage; il existe des différences marquées entre les écoles relativement aux taux de décrochage; la région (en milieu urbain, en milieu rural, à l'étranger) n'est pas statistiquement significative; parmi les catégories en fonction du risque tirées du Sondage d'accueil des élèves (SAE), les seules qui se révèlent significatives sont « clarté sur le plan professionnel » et tout particulièrement « engagement envers l'éducation »; les résultats des évaluations en lecture et en maths sont également importants.
- À l'échelle individuelle, les taux prédits de décrochage chez les élèves (où le décrochage s'entend des élèves qui n'amorcent pas la deuxième session de leur programme) générés par le modèle prédictif, lesquels étaient estimés pour les cohortes arrivées de 2005 à 2012 puis mis à l'essai pour les cohortes arrivées en 2013 et en 2014, suivaient de près les taux réels de décrochage.
- Dans les trois catégories de risque des élèves définies au moyen du modèle prédictif, les taux réels de décrochage se fixent à 24 % chez les élèves à fort risque, à 14 % chez les élèves à risque moyen, et à 9 % chez les élèves à faible risque dans les cohortes arrivées en 2013 et en 2014, ce qui révèle par ailleurs à quel point le modèle prédictif permet de différencier efficacement les élèves selon le niveau de risque.
- Enfin, contrairement aux perceptions répandues chez les intervenants des affaires étudiantes comme quoi les élèves qui prennent part aux programmes d'aide pédagogique et de soutien sont ceux qui n'en ont pas vraiment besoin [Dietsche (2012)], il appert que les élèves des groupes à fort risque et à risque moyen sont ceux qui ont participé le plus aux programmes existants d'aide

[1] Nous avons opté pour 23 ans comme valeur de démarcation du groupe le plus vieux parce que nous voulions obtenir des échantillons d'une taille élargie durant la formation du modèle prédictif.

pédagogique, à environ 22 %, tandis que cette participation plafonnait à 14 % chez les élèves à faible risque (tous les résultats se rapportent aux cohortes d'essai arrivées en 2013 et en 2014).

Retombées sur les politiques du Collège Mohawk

En ce qui concerne le Collège Mohawk, ces constatations ont plusieurs retombées importantes. D'abord, la modification des taux prédits de décrochage à titre de valeur de démarcation servant à définir les différentes catégories de risque des élèves se traduit par des groupes substantiellement différents d'élèves dont la taille varie et qui présentent différentes caractéristiques, ce qui comporte des retombées sur les politiques de soutien aux élèves. Le mode d'utilisation du modèle prédictif révèle donc des ensembles de décisions politiques plutôt que de constituer un exercice strictement statistique.

Bien que toutes les catégories de risque fondées sur le SAE ne soient pas statistiquement significatives, les catégories « clarté sur le plan professionnel » et tout particulièrement « engagement envers l'éducation » constituent deux exceptions notables, et il vaut la peine de se pencher davantage sur celles-ci pour constater ce qu'elles permettent de saisir et comment nous pourrions venir en aide aux élèves répertoriés à risque par ces variables. Les résultats des évaluations en lecture et en maths sont également importants, car ils mettent en relief la nécessité d'aider davantage les élèves à se préparer aux études collégiales en ce qui touche ces compétences essentielles.

Le fait qu'une faible moyenne à l'école secondaire constitue particulièrement un facteur de risque de décrochage (quoique parmi de nombreux autres facteurs de risque répertoriés dans la démarche de modélisation employée ici) porte à croire que le collège voudra peut-être revoir ses exigences minimales ou conditionnelles dans les programmes applicables afin de déterminer si des changements politiques, y compris l'amélioration des initiatives de soutien aux élèves, pourraient se traduire par un rehaussement du maintien aux études.

Conclusion et voies des futures recherches

Dans l'ensemble, les constatations semblent indiquer que le recours aux données administratives et connexes à l'échelle des élèves afin de mettre au point et d'utiliser des modèles prédictifs de maintien aux études des élèves constituent une pratique prometteuse pour aider les établissements d'enseignement à bien concevoir le maintien aux études des élèves, à cibler les élèves qui présentent un risque supérieur de décrochage au moyen de programmes de soutien aux élèves, de même qu'à mettre à l'essai et élaborer davantage des politiques, programmes ou services aux retombées potentiellement positives sur les taux de persévérance scolaire, de maintien aux études et de diplomation des élèves.

Par conséquent, l'une des voies d'approfondissement de la recherche, c'est de compter sur un nombre supérieur d'établissements d'enseignement qui élaborent et utilisent le même type de modèles prédictifs du maintien aux études des élèves en fonction des expériences vécues par les élèves en leur sein. Voilà qui permettrait de brosser un tableau complet du maintien aux études des élèves, d'améliorer le ciblage des

programmes destinés aux élèves, et de diriger une mise à l'essai statistique rigoureuse des effets des initiatives de soutien aux élèves qui sont mises en place.

Parallèlement, il serait possible d'améliorer les modèles employés au moyen des mesures suivantes :

1. Ajouter des variables supplémentaires en fonction des données que possèdent déjà les établissements d'enseignement, comme le besoin en aide financière, ou les données sur les choix de programme tirées des demandes d'admission au collège.
2. Ajouter d'autres variables par la jonction avec d'autres sources de données, comme l'ajout de renseignements socioéconomiques sur l'élève par le recours aux renseignements sur le code postal afin de mettre en lien les données fondées sur le recensement avec le cadre de vie contextuel de l'élève.
3. Ajouter également aux modèles une « déclaration précoce » à propos des élèves comme les présences, les évaluations précoces ou les notes, laquelle témoigne du comportement et des résultats de l'élève après son admission dans l'établissement d'enseignement.
4. Explorer les sources potentiellement colossales de renseignements électroniques qui sont de plus en plus recueillis à propos des élèves, dont ceux en lien avec la participation aux cours, y compris l'interaction en ligne avec les systèmes de gestion d'apprentissage.

Enfin, il y a lieu de tenir compte de l'élaboration et de la mise à l'essai des autres types de modèles prédictifs, comme ceux récemment proposés en fonction d'algorithmes avancés d'apprentissage automatique pour voir si ceux-ci donnent des prédictions améliorées et, dans l'affirmative, si de telles améliorations sont supérieures à leur complexité au chapitre de leur élaboration et de leur mise en application.

Table des matières

Liste des graphiques

Liste des tableaux

1. Introduction

1.1 La signification, l'élaboration et l'utilisation des modèles prédictifs

Dans son sens le plus large, un modèle prédictif peut s'entendre d'une fonction mathématique qui donne des prédictions relatives à un résultat d'intérêt compte tenu des valeurs des variables de prédiction. Dans le contexte de l'enseignement postsecondaire (EPS) et du maintien aux études des élèves, le but d'un tel modèle peut consister à prédire si un élève quittera l'établissement d'enseignement avant d'obtenir son diplôme, d'après les caractéristiques observables de cet élève à un moment donné. Ces prédictions peuvent servir à cibler les élèves au moyen d'interventions et de soutiens qui tiennent compte de leur niveau de risque.

Dans ce contexte, il existe une gamme de méthodes permettant de prédire les élèves qui sont à risque de décrocher de leurs études. D'un côté, de simples indicateurs de risque comme le sexe de l'élève, ses notes en début de programme et ses heures de travail rémunérées peuvent servir d'indicateurs univariés bruts d'élèves « à fort risque » d'après de simples analyses descriptives menées pour d'autres élèves dans d'autres établissements d'enseignement (voire dans d'autres pays), ce qui permet de discerner les liens généraux entre les taux de décrochage et de tels attributs des élèves.

Par exemple, une méthode simple consiste à classer l'élève comme étant à fort risque de décrochage s'il ou si elle présente un certain nombre de ces facteurs (ou caractéristiques) de risque. Toutefois, cette méthode s'appuie sur des méthodes très simples d'analyse statistique (c.-à-d. les indicateurs univariés décrits) de populations d'élèves complètement différentes[2].

Une version quelque peu améliorée de cette méthode consiste à recourir aux données et à l'analyse antérieures en fonction des élèves ou de l'établissement d'enseignement en question afin de déterminer les liens empiriques sous-jacents aux indicateurs mis au point. Cependant, le résultat demeure un ensemble d'indicateurs bruts et, dans une mesure importante, arbitraires zéro-un (univariés), lesquels servent de nouveau à discerner de façon ponctuelle les élèves à risque.

De l'autre côté, il y a l'élaboration de modèles de régression ou davantage perfectionnés (y compris certains des algorithmes avancés d'apprentissage automatique) pour en arriver au modèle qui prédit le mieux le comportement de l'élève (p. ex., le décrochage) à l'établissement d'enseignement en question.

La documentation moderne relativement à la modélisation prédictive est caractérisée par quelques composantes et avantages clés. Premièrement, les modèles sont conçus à partir de rien dans l'optique de maximiser la validité prédictive des nouvelles données. Pour en arriver là, une partie des données

[2] Un exemple particulièrement malheureux (et fréquent) d'erreurs extrêmes qu'une telle démarche peut occasionner en contexte canadien est de prêter attention aux « élèves de première génération » : dans les autres pays, on constate habituellement que le taux de décrochage de ces élèves est supérieur à celui des élèves dont les parents ont l'expérience de l'EPS, mais pour qui des taux de décrochage inférieurs sont souvent constatés au Canada [Finnie et Qiu (2008)].

disponibles sert d'échantillon de formation et, une fois le modèle mis au point, sa validité prédictive est mesurée au moyen de l'échantillon de mise à l'essai. Voilà la différence fondamentale entre, par exemple, un modèle de régression descriptif et un modèle de régression prédictif.

Deuxièmement, les modèles prédictifs peuvent englober beaucoup plus de variables ou de facteurs que ceux pouvant servir dans un modèle descriptif, lequel est habituellement plus parcimonieux afin de raconter en quoi le résultat d'intérêt est lié à un ensemble de variables explicatives clés. La méthode prédictive tire donc parti de tous les renseignements disponibles pour maximiser la validité prédictive.

Troisièmement, les modèles prédictifs produisent une valeur de probabilité propre à chaque élève pour qui il faut faire une prédiction. Les probabilités en question s'étendront sur une échelle continue allant de 0 à 1, où 0 signifie que l'élève ne décrochera certainement pas tandis que 1 signifie qu'il ou elle décrochera assurément. Voilà qui permet aux établissements d'enseignement de distinguer et de régir les élèves, tous niveaux de risque confondus... quoique les probabilités générées par le modèle puissent ultérieurement servir à classer les élèves en fonction des différentes catégories de risque (p. ex., à fort risque, à risque moyen, à faible risque).

Dans le contexte du maintien aux études des élèves, les modèles prédictifs de même que les prédictions sur les élèves à l'échelle individuelle qu'ils génèrent peuvent servir dans une gamme de moyens très commodes[3]. Premièrement, bien que leur finalité première consiste à fournir des prédictions, les modèles prédictifs peuvent également aider un établissement d'enseignement à bien concevoir en quoi le décrochage est lié aux divers facteurs ou caractéristiques, ayant trait notamment aux élèves ou au programme, qui font partie des modèles.

Deuxièmement, les établissements d'enseignement peuvent se servir des prédictions sur le décrochage à l'échelle des élèves générées par les modèles pour canaliser leurs initiatives de réussite des élèves (sinon d'autres programmes ou activités) vers les élèves qui en ont le plus besoin.

Troisièmement, dans le même ordre d'idées, il est possible de procéder à une estimation empirique des effets d'un programme ciblant les élèves au moyen d'un tel ensemble de prédictions et de valeurs de démarcation en lien avec les prédictions (p. ex., les élèves qui se situent au-delà d'une certaine probabilité de risque sont ciblés par le programme, tandis que ceux en deçà de la limite ne le sont pas), précisément en raison de la « discontinuité » qui caractérise cette méthode de ciblage, de même que les méthodes connexes d'estimation[4].

Enfin, l'efficacité d'un programme de réussite des élèves mis en place de façon plus globale (p. ex., un programme auquel tous les élèves ont accès) peut être estimée selon différents niveaux de risque des

[3] Pour obtenir le modèle prédictif dans les travaux relatés ici, nous formons nos données à partir des cohortes de 2005 à 2012 par l'estimation de plusieurs modèles de régression logistique assortis de différents ensembles de paramètres de prédiction, puis nous procédons à une validation externe sur les cohortes de 2013 et de 2014. Ce processus est expliqué en détail à la section 3.3.
[4] Pour en savoir plus sur l'efficacité de la discontinuité de la régression, voir les auteurs Mayhew et al. (2016).

élèves. Par exemple, bien que le fait d'orienter les programmes vers les élèves qui risquent le plus de décrocher prématurément puisse être logique, au moins une partie des programmes pourront se révéler des plus efficaces dans l'amélioration des résultats des élèves à risque moyen ou même à faible risque, pendant que d'autres programmes pourront donner de meilleurs résultats chez les élèves les plus vulnérables. Ce type d'analyse peut donc aider les établissements d'enseignement à concevoir une série de programmes les plus efficaces dans l'amélioration des résultats des élèves en fonction des ressources consacrées à ceux-ci.

1.2 Le projet actuel : Se servir des modèles prédictifs pour mettre à l'essai des interventions auprès des élèves

Le projet dont il est question dans le présent document s'appuie sur l'utilisation de modèles prédictifs pour discerner dans un premier temps les niveaux de risque des élèves, puis mettre à l'essai trois différents programmes d'aide pédagogique aux élèves dans l'ensemble des trois différents niveaux de risque des élèves (à faible risque, à risque moyen, à fort risque) en fonction des prédictions générées par le modèle.

Par conséquent, la recherche se situe non seulement dans le contexte élargi du recours à l'analyse de données et à la modélisation prédictive, mais dans le contexte particulier de la documentation sur le maintien aux études des élèves, où le maintien aux études des élèves (l'inverse du décrochage d'un programme avant l'obtention du diplôme) est généralement conçu comme le résultat d'ensembles complexes de facteurs, ce qui est difficile à mesurer [Wiggers et Arnold (2011)] et à prédire. Parallèlement, chez les intervenants et les chercheurs, l'idée selon laquelle il n'existe pas de solution unique au rehaussement de la réussite des élèves fait également consensus [Kuh, Kinzie, Schuh et Whitt (2005); Reason (2009)].

Les concepts de modélisation prédictive — le dépistage précoce et l'intervention — et d'aide pédagogique dans les écoles afin de soutenir la réussite des élèves constituent donc les fondations de ce projet de recherche. Il s'agit du premier de deux rapports qui traitent du lien entre la modélisation prédictive, l'aide pédagogique dans les écoles et le maintien aux études des élèves au Collège Mohawk.

Ce projet en particulier est ancré dans le plan de réussite des élèves du Collège Mohawk (2013), lequel décrit un réseau de soutien aux élèves à risque de décrocher des études collégiales. Ce réseau englobe une intervention précoce, une aide pédagogique perturbatrice (proactive) complète et la mobilisation des élèves au moyen d'activités parallèles aux cursus et hors cursus. La méthode de soutien du Collège Mohawk repose sur une base d'évaluations suivant l'admission (précédant l'arrivée), de modélisation prédictive et d'intervention précoce facilitée par l'aide pédagogique.

Les buts dans cette phase de la recherche consistent d'abord à actualiser et à évaluer le modèle prédictif préalablement mis au point à l'intention du Collège Mohawk par l'Initiative de recherche sur les politiques de l'éducation (IRPE) à l'Université d'Ottawa, à l'aide des élèves qui ont fait leur entrée au Collège Mohawk de 2005 à 2012. Le modèle sert à prédire l'éventualité de décrochage des études collégiales chez chaque nouvel élève des cohortes de l'automne 2013 et de l'automne 2014. Ces prédictions servent ensuite à vérifier si le modèle prédictif permet de discerner efficacement les niveaux de risque des élèves.

Le modèle prédictif est employé afin de classer les élèves dans l'une des trois classifications du risque des élèves, ou CRE (à faible risque, à risque moyen, à fort risque), de taille égale. Sa mise à l'essai est approfondie par la comparaison des taux réels de décrochage dans les trois catégories de risque des élèves.

Les trois groupes de risque sont analysés pour discerner les différentes caractéristiques des élèves qui ont tendance à figurer dans les groupes de risque des élèves, sans toutefois oublier que c'est l'ensemble complet des caractéristiques des élèves qui détermine les taux prédits de décrochage pour chaque élève et, par conséquent, le groupe de risque auquel ils appartiendront.

Enfin, les tendances de participation aux programmes actuels d'aide pédagogique chez les élèves des différents groupes de risque sont également calculées.

Par la suite, cette modélisation de même que les prédictions et les attributions aux groupes de risque serviront au cours de la seconde phase de la recherche, où sera mise à l'essai l'efficacité des programmes d'action directe et d'aide pédagogique visant à améliorer le maintien aux études des élèves et qui sont mis en place au Collège Mohawk. Les résultats de la seconde phase de recherche seront relatés ultérieurement cette année.

2. Analyse documentaire

2.1 Pratiques servant à améliorer les résultats des élèves

Au sein de la documentation, de nombreuses pratiques qui améliorent la réussite des élèves sont citées. Par exemple, dans l'enquête nationale de l'America College Testing (ACT) intitulée *What Works in Student Retention*, 96 points sont répertoriés comme pratiques potentielles de maintien aux études à envisager, à mettre en œuvre et à évaluer par les intervenants [Habley, Bloom et Robbins (2012)].

L'auteur Kuh (2008) a recensé 10 pratiques à retombées élevées qui rehaussent passablement l'apprentissage des étudiants à l'université, tandis que le Center for Community College Student Engagement (CCCSE) a recensé dans un récent rapport national (2014) 14 pratiques à retombées élevées ayant trait au rehaussement des résultats des élèves dans les collèges. Ce ne sont là que quelques-uns des nombreux exemples.

En outre, les auteurs Tinto (1975; 1993), Braxton et al. (2004), Terenzini et Reason (2005), Reason (2009) et Braxton et al. (2014) ont tous contribué à un cadre à évolution progressive de réussite des élèves. Le cadre théorique des auteurs Braxton et al. (2014) tient compte du rôle joué par les caractéristiques des élèves à leur arrivée, l'engagement de l'établissement d'enseignement au départ, le milieu externe, le milieu interne, les caractéristiques organisationnelles, l'épanouissement intellectuel et à l'école, ainsi que l'engagement de l'établissement d'enseignement par la suite. Chacun de ces éléments influe sur les résultats de la persévérance scolaire des élèves. Ce vaste ensemble de théories et de pratiques témoigne de la complexité de la réussite des élèves.

Toutefois, il y a dans cette documentation — tout particulièrement au sein du secteur des collèges communautaires — une gamme d'activités complémentaires qui sont constamment désignées comme prometteuses pour l'amélioration des résultats des élèves. Parmi ces activités, il y a le recours aux modèles prédictifs [van Barneveld et al. (2012)], aux pratiques de dépistage précoce d'intervention [Center for Community College Student Engagement (2014)] et à l'aide pédagogique dans les écoles [Braxton et al. (2014)].

D'après le Center for Community College Student Engagement (2014), les programmes de dépistage précoce et d'intervention constituent des pratiques à retombées élevées, car ces programmes sont considérés comme ceux ayant l'effet le plus marqué sur le maintien aux études des élèves. Le CCCSE définit les programmes de dépistage précoce et d'intervention comme un processus systématique dans lequel les instructeurs signalent aux intervenants compétents des collèges les cas où les élèves dans leur classe ont de la difficulté à l'école, après quoi l'intervenant entre en contact avec les élèves dans un effort pour leur procurer le soutien dont ils ont besoin.

Cette méthode s'inscrit dans la théorie et la pratique de l'intervention perturbatrice ou proactive d'aide pédagogique dans les écoles [Glennen (1975); Varney (2013)], lesquelles sont des méthodes utiles en ce qui touche les services d'action directe et de soutien. Les interventions proactives contrastent fortement avec la méthode de laissez-faire dans les services de soutien aux élèves, laquelle n'est plus pertinente selon l'auteur Dietsche (2012) pour soutenir les élèves à l'heure actuelle. Après avoir mené une enquête poussée auprès de 60 000 élèves de niveau collégial en Ontario, il en a conclu que les services proactifs d'action directe et d'aide pédagogique jouent un rôle crucial dans la réussite des élèves. L'auteur Poirier (2015) a évoqué le même argument dans son analyse des programmes d'orientation et de transition au sein de trois collèges d'envergure en Ontario.

En s'appuyant sur des exemples des auteurs Habley et al. (2012), Braxton et al. (2014) et du Center for Community College Student Engagement (2014), l'auteur Fricker (2015) a avancé récemment que l'aide pédagogique dans les écoles est fréquemment évoquée en tant que service central qui favorise la réussite des élèves, tout particulièrement celle des élèves des collèges communautaires. L'exemple le plus récent à ce chapitre est une étude des auteurs Braxton et al. (2014), laquelle a permis de trouver des données empiriques pour appuyer une théorie de la réussite des élèves dans les campus en banlieue, mais ce qu'il faut surtout souligner, c'est que l'aide pédagogique dans les écoles y est mentionnée parmi les interventions les plus importantes. Étant donné que la plupart des collèges en Ontario sont des campus en banlieue, cette théorie de même que les recommandations à mettre en pratique peuvent nous intéresser directement.

Hélas, la documentation sur la réussite des élèves, leur maintien aux études ou la pratique d'aide pédagogique dans les écoles au sein des campus collégiaux au Canada est très restreinte [Fricker (2015)]. L'aide pédagogique dans les écoles fait souvent l'objet d'une définition au sens large. L'auteur Grites (1979, p. 1) définit celle-ci comme « un processus décisionnel au cours duquel les élèves prennent connaissance de leur potentiel d'apprentissage maximal au moyen de la communication et d'échanges d'information avec un conseiller ». Les auteurs Braxton et al. (2014) ont évoqué cette définition dans leurs récents travaux. De même, l'Ontario Academic Advising Professionals (s.d.) a affirmé, au sujet de l'aide pédagogique dans les

écoles, qu'il fallait « la concevoir dans son sens le plus large et qu'elle pouvait englober les intervenants dans la prestation de conseils sur le plan scolaire, de conseils sur le plan professionnel, de counseling, de services de liaison ou de possibilités d'apprentissage de compétences pour favoriser la réussite et le maintien aux études des élèves ». Ces définitions vont de pair avec le rôle joué par les conseillers à la réussite des élèves du Collège Mohawk et donnent le contexte du projet de recherche actuel.

Le présent projet donne un aperçu contemporain du rapport entre la participation à l'aide pédagogique dans les écoles et le maintien aux études des élèves au collège.

2.2 Données, analytique des données et modélisation prédictive

Il existe une pléthore de concepts vaguement liés aux pratiques qui se rapportent à la prédiction de la réussite des élèves dans l'enseignement supérieur. Afin d'uniformiser la terminologie relative à l'analytique dans l'enseignement supérieur, les auteurs van Barneveld et al. (2012) ont présenté un cadre utile, lequel permet de définir l'analytique opérationnelle, l'analytique scolaire, l'analytique de l'apprentissage, l'analytique prédictive et l'analytique des mesures.

Dans ce cadre, « l'analytique est un concept global décrit sous l'angle de la prise de décision axée sur les données » [d'après van Barneveld et al. (2012), p. 6] avec l'aide de systèmes informatiques spécialisés, pendant que l'analytique opérationnelle et scolaire « donne à la direction ou aux cadres supérieurs l'accès aux indicateurs — historiques ou en temps réel à l'aide de " tableaux de bord " — du rendement du centre opérationnel (l'établissement d'enseignement supérieur) et de ses unités (les collèges, les écoles ou les départements) ».

À partir de ces concepts, « l'analytique prédictive est un processus qui sert à tous les niveaux de l'enseignement supérieur et opérationnels, et fait fonction de connecteur entre les données recueillies, la mesure intelligente qui peut être prise par suite de l'analyse et, au bout du compte, la prise de décisions éclairées ». Ces auteurs ont également proposé, de façon plus formalisée, que l'analytique prédictive soit définie comme « un secteur de l'analyse statistique qui traite de la saisie des renseignements à l'aide de diverses technologies pour dévoiler les liens et tendances au sein de fortes quantités de données qui pourront servir à prédire le comportement et les événements » [Ibid, p. 8]. Au cœur de ce concept, il y a le recours à l'analytique pour prendre des mesures et instaurer des programmes, services et interventions qui soutiennent la réussite des élèves. Parmi les exemples de projets recensés par les auteurs van Barneveld et al., il y a les plans de réussite des élèves et les répertoires de préparation des élèves.

Les universités et collèges ont accès à divers outils afin d'appuyer ce type de travail. Une méthode particulièrement prometteuse fait intervenir le recours à des modèles prédictifs qui peuvent employer des données antérieures pour prédire les futurs résultats des élèves à l'échelle individuelle. Une méthode répandue consiste à utiliser l'analyse de régression logistique des données historiques pour modéliser les phénomènes de maintien aux études, lesquels peuvent ensuite servir à prédire la réussite des futures cohortes. Dans cette méthode, la première partie des données administratives accessibles au sujet des élèves (ce peut être les données de certaines années ou simplement une sélection aléatoire de l'échantillon complet) est employée pour mettre au point le modèle, tandis que la seconde partie des données (ce peut

être d'autres années — habituellement les plus récentes — ou une sélection aléatoire de l'échantillon complet) servent à mettre à l'essai le modèle et à en mesurer le rendement.

Le modèle ayant la plus faible erreur de prédiction sert ensuite à prédire les probabilités à l'échelle individuelle de décrochage des études collégiales chez les nouveaux élèves. Puisque le modèle prédictif tient compte des diverses caractéristiques des élèves pour lesquelles des données sont accessibles, le modèle révèle dans sa prédiction une exactitude supérieure à celle de l'exercice de prédiction des décrocheurs en fonction des seules statistiques descriptives (p. ex., le fait d'examiner seulement les indicateurs habituels du risque comme les notes à l'école secondaire ou le rendement aux tests d'évaluation des élèves... ou la combinaison de tels indicateurs), de sorte que ce modèle permettra de mieux prédire la réussite des élèves.

Ces types de méthodes de modélisation prédictive servent dans de nombreux différents domaines à prédire les futurs résultats en fonction des tendances historiques. Le recours à de tels modèles peut être le plus fréquent en finances, où la modélisation prédictive est utilisée pour mesurer le risque de faillite chez les particuliers [Foster et Stine (2004)] de même que les entreprises [Atiya (2001)] selon leurs antécédents financiers, ou afin de discerner divers types de fraudes [Phua et al. (2010)].

De même, la modélisation prédictive est employée en médecine pour prédire les maladies non diagnostiquées ou le pronostic d'une maladie diagnostiquée en fonction des caractéristiques du patient [voir, par exemple, Baan et al. (1999) et Federico et al. (2000)]. Les modèles prédictifs servent également de pilier dans certains domaines du génie. À titre d'exemple, ces modèles ont servi à prédire les accidents routiers sur les autoroutes passantes [Hossain et Muromachi (2012)] ou à mettre en œuvre des économies d'énergie dans les véhicules d'après les algorithmes prédictifs du comportement du chauffeur [Murphey et al. (2008)]. De tels algorithmes sont également utilisés quotidiennement par divers sites Web pour présenter des publicités ciblées aux personnes qui consultent ceux-ci en fonction des habitudes particulières de navigation dans Internet [Perlich et al. (2014)].

Par comparaison, les milieux de la recherche en éducation ont adopté tardivement de telles méthodes. Les modèles prédictifs ont servi à prévoir les notes obtenues par les élèves [Kotsiantis (2012)], à discerner les élèves qui risquent de ne pas terminer leurs études à temps [Aguiar et al. (2015); Lakkaraju et al. (2015); Sara et al. (2015)], ainsi qu'aux fins du maintien aux études des élèves [Dekker et al. (2009); Delen (2010); Lin (2012); Nandeshwar et al. (2011); Thammasiri et al. (2014); Yu et al. (2010); Zhang et al. (2010)], pour ne nommer que ces utilisations-là.

Cependant, pour autant que les auteurs le sachent, il n'y a pas de nombreuses études qui mettent à l'essai ou valident un modèle prédictif de maintien aux études des élèves au Canada, hormis celle des auteurs Conrad et Morris (2010) dans laquelle les données administratives des élèves sont analysées au moyen de la technique d'apprentissage automatique de « forêt de survie aléatoire » pour prédire le maintien aux études des étudiants à l'Université York.

Pour leur part, les auteurs Jia et Maloney (2015) ont procédé à une estimation empirique des facteurs déterminants des résultats de l'inachèvement des cours en première année et des résultats du décrochage

scolaire en deuxième année au moyen des données administratives d'une importante université publique en Nouvelle-Zélande.

3. Données et méthodologie

3.1 Définitions des variables

La section 3.1 décrit sommairement les variables employées dans l'analyse dont il est question dans le présent document. La sélection des variables s'est faite en fonction de l'accessibilité des données au Collège Mohawk et s'inscrit dans le modèle théorique instauré par l'auteur Tinto (1975, 1993), lequel est bien connu et abondamment employé dans la documentation sur la persévérance scolaire. Selon ce modèle, les élèves qui amorcent un EPS présentent diverses caractéristiques préalables à l'admission, comme l'âge, la race, le sexe, la structure familiale, le niveau de scolarité des parents et la préparation à l'école secondaire, de même que leurs propres compétences et aptitudes. Ces facteurs contribuent à la formation des buts initiaux des élèves ainsi qu'à leur degré d'engagement dans leurs études. Les données sur la moyenne à l'école secondaire, lesquelles sont expliquées ci-dessous, font également partie du modèle employé dans le présent document.

Variables des élèves et des programmes

L'ensemble des variables des élèves et des programmes englobe dans un premier temps l'année d'admission et le sexe. L'âge, qui est également inclus, est ventilé en six catégories : moins de 18 ans; 18 ans; 19 ans; 20-22 ans; 23-26 ans; et 27 ans et plus. La situation régionale permet de déterminer si l'élève habite en milieu urbain ou rural à l'étape de la présentation de sa demande d'admission ou s'il s'agit d'un élève étranger (et qui ne présente donc pas les catégories indiquées ci-dessus).

La moyenne des notes à l'école secondaire a tendance à figurer parmi les paramètres de prédiction les plus solides quant au maintien aux études des nouveaux élèves [Astin (1997)]. Cette variable est calculée en fonction de la moyenne des six notes les plus élevées dans les cours d'anglais et de mathématiques suivis durant la troisième et la quatrième année de l'école secondaire. Les catégories de cette variable sont les suivantes : A plus; A; A moins; B plus; B; C plus; C; D plus; D; F.

L'école correspond au programme auquel l'élève est inscrit (il y a 17 écoles). La variable des études comporte quatre catégories : le certificat (1 an); le diplôme (2 ans); le diplôme d'études supérieures (3 ans); et le certificat d'études supérieures (1 an).

Variables du Sondage d'accueil des élèves

Au Collège Mohawk, les nouveaux élèves répondent au Sondage d'accueil des élèves (SAE) au début de chaque session. Le SAE a été instauré au Collège Mohawk en 2006 dans le cadre d'un projet financé par le Ministère pour faire passer l'OCSES (sondage sur la participation des élèves dans les collèges de l'Ontario) [Dietsche (2009)]. Le SAE a été mis au point par Peter Dietsche au cours des 20 dernières années, y compris

une version antérieure qui a servi dans le PCSCS (sondage pancanadien sur les élèves de niveau collégial) [Dietsche (2007), (2008)]. Le SAE est constamment utilisé au Collège Mohawk depuis 2006 et il a fait l'objet de certaines corrections et améliorations.

Le SES n'est pas un sondage obligatoire, mais il englobe les réponses d'environ 70 % des nouveaux élèves des cohortes de l'automne 2013 et de l'automne 2014. Il comporte des questions ayant pour objet d'aider à discerner certains des facteurs de risque de décrochage des études collégiales. Les réponses des élèves à des questions ou des groupes de questions en particulier définissent chacune des variables suivantes : faible clarté sur le plan professionnel; faible confiance en ses aptitudes; travailler 15 heures/semaine ou plus pendant les études; effectuer difficilement la transition vers la vie collégiale au Collège Mohawk; et faible engagement envers l'éducation. Toutes ces variables sont binaires (à risque = 1 et pas à risque = 0) et permettent de discerner ce qui est considéré comme des facteurs de risque de décrochage des études collégiales.

Résultats des évaluations

Les nouveaux élèves au Collège Mohawk se livrent à des évaluations en rédaction, en lecture et en mathématiques avant le début de leur première session. Le Projet portant sur les mathématiques au niveau collégial [Orpwood, Schollen, Leek, Marinelli-Henriques et Assiri (2012)] et le Projet portant sur le rendement des étudiantes et étudiants au niveau collégial (2015) ont permis d'étudier et de relater l'importance de telles variables dans la réussite des élèves au sein des collèges de l'Ontario durant la dernière décennie.

Au Collège Mohawk, les évaluations en lecture et en rédaction se déroulent sur la plateforme Accuplacer et font appel au logiciel WritePlacer en ce qui touche la rédaction de la dissertation et le résultat attribué à celle-ci. L'évaluation en mathématiques, conçue par le Collège Mohawk, a lieu sur la plateforme Maple T.A.. Puisque des changements ont été apportés d'une année à l'autre aux échelles de notation de ces évaluations, nous avons attribué de nouveau un résultat à chaque variable d'évaluation pour prendre en compte la position relative de l'élève dans la répartition globale des résultats quant à l'évaluation dont il ou elle a fait l'objet.

Les résultats des évaluations en lecture et en mathématiques sont groupés en huit catégories allant de 1 à 8. La catégorie la plus faible montre que le résultat de l'élève se situe dans l'échelon le plus bas de la répartition des résultats, tandis que la catégorie la plus forte révèle que le résultat de l'élève figure à l'échelon le plus élevé de la répartition des résultats.

Pour ce qui est de l'évaluation en rédaction, il existe deux catégories — 1 et 2 — qui correspondent respectivement à un résultat en deçà et à un résultat au-delà de la médiane.

De plus, puisque tous les élèves ne se livrent pas à ces évaluations, les catégories manquantes font également partie de chaque évaluation. Au total, 34 % des élèves se sont livrés à l'évaluation en mathématiques, tandis que 56 % des élèves ont mené à bien les évaluations en lecture et en rédaction[5].

Les variables des résultats : le décrochage et la participation à un programme d'aide pédagogique

La variable d'intérêt clé des résultats dans la présente analyse est de savoir si l'élève a décroché du programme ou pas. La mesure du décrochage est binaire (n'a pas décroché = 0 et a décroché = 1) et elle correspond au maintien aux études allant de la première à la deuxième session (maintien aux études en première session).

Le 10e jour de chaque session sert de date repère du maintien aux études des élèves, ce qui correspond à la fin de la période d'« ajout ou retrait » des élèves au Collège Mohawk. Autrement dit, il s'agit du dernier jour de la session auquel les élèves peuvent s'inscrire. Les élèves qui sont inscrits à la 10e journée de la première session d'automne font partie de l'analyse et sont réputés être demeurés au Collège Mohawk s'ils ont de nouveau le statut d'inscrit à la 10e journée de la deuxième session, l'hiver (maintien aux études en première session) ou à la 10e journée de la troisième session (maintien aux études durant un an) au cours de la session suivante à l'automne[6].

La participation à l'aide pédagogique aux élèves constitue le second résultat d'intérêt. La variable pertinente est décrite comme suit : si l'élève a cherché à obtenir l'aide des conseillers à la réussite des élèves au moins une fois durant la session. Les données s'y rapportant sont recueillies au moyen de ClockWork, un logiciel d'aide pédagogique dont se servent tous les conseillers à la réussite des élèves au Collège Mohawk. Aux fins de la présente analyse, il s'agit également d'une variable binaire (n'a pas cherché à obtenir de l'aide pédagogique = 0; a cherché à obtenir de l'aide pédagogique = 1). Cette variable ne prend en compte ni la fréquence de l'aide pédagogique fournie à l'élève, ni la durée ou le type d'interaction d'aide pédagogique qui a eu lieu.

3.2 Population et restrictions de l'échantillon

Les données servant à estimer le modèle de maintien aux études sous-jacent aux prédictions sur le décrochage des élèves englobaient les élèves admis au Collège Mohawk de 2005 à 2012.

Par la suite, nous avons généré les taux prédits de décrochage du Collège Mohawk chez les élèves qui y ont amorcé leurs études à la session d'automne 2013 et à la session d'automne 2014.

[5] Tous les nouveaux élèves ne sont pas tenus de se livrer à l'évaluation en maths.

[6] La situation de retrait des élèves de retour à la deuxième session est vérifiée et les élèves qui se retirent avant le 10e jour sont également considérés comme des décrocheurs. Certains élèves qui décrochent (particulièrement ceux qui ne se retirent pas) n'entreront pas dans ce dossier, mais n'apparaîtront tout simplement pas à la session suivante. Notre méthode permet de considérer comme décrocheurs les élèves qui décrochent de leurs études au Collège Mohawk sans passer par le processus officiel de retrait.

L'analyse est restreinte aux élèves du campus (principal) Fennell, parce que c'est à cet endroit que les programmes d'aide pédagogique qui nous intéressent dans le cadre du présent projet ont été mis en place à la session d'automne 2015.

3.3 Le modèle prédictif

Le modèle prédictif employé dans la présente analyse a d'abord été conçu dans le cadre de travaux préalables, mais nous l'avons actualisé pour le projet actuel. Nous nous servons d'une méthode fondée sur un modèle de régression logistique, laquelle est employée fréquemment dans la documentation économique au sens large, afin de modéliser les résultats binaires (0-1) comme aller au collège, être chômeur et prendre des décisions en matière de migration (pour ne citer que quelques exemples).

Dans le modèle de régression logistique qui nous concerne, la probabilité de décrochage est définie comme suit :

$$\boldsymbol{P} = \frac{\boldsymbol{e}^{\boldsymbol{\beta}_0+\boldsymbol{\beta}_1\boldsymbol{X}}}{\mathbf{1}+\boldsymbol{e}^{\boldsymbol{\beta}_0+\boldsymbol{\beta}_1\boldsymbol{X}}}$$

Les termes $\boldsymbol{\beta_0}$ et $\boldsymbol{\beta_1}$ correspondent aux paramètres liés à chaque élément de $\boldsymbol{X}$, l'ensemble des variables des paramètres de prévision qui font partie du modèle (répertoriées ci-dessus) et indiquent l'effet particulier de chaque variable sur le décrochage, compte tenu des autres variables comprises dans le modèle[7].

Le choix définitif du modèle prédictif s'est fondé sur une comparaison entre de nombreuses spécifications différentes de modèles, lesquelles englobaient diverses combinaisons de variables sur le contexte, le risque et les évaluations. Par exemple, une spécification n'englobe que les variables des élèves des programmes (c.-à-d. le sexe, la situation urbaine ou rurale, l'âge, le titre d'études, la moyenne à l'école secondaire et l'école). Nous avons amplifié le modèle choisi pour y intégrer des termes d'interaction entre les différents ensembles de variables. Par la suite, nous avons également amplifié ses modèles pour y intégrer les variables du risque et de l'évaluation, à la fois séparément et ensemble.

Les données des cohortes des sessions d'automne 2005 à 2012 ont servi de données de formation dans l'élaboration de notre modèle prédictif. Nous utilisons les données des cohortes de la session d'automne 2013 et de la session d'automne 2014 pour procéder à une validation externe. Nous avons comparé le rendement de chaque spécification de modèle en fonction de la valeur de la perte de logarithme générée. La perte de logarithme quantifie essentiellement la mesure dans laquelle les probabilités prédites de décrochage (des valeurs entre 0 et 1, d'après les données des sessions d'automne 2005 à 2012) diffèrent des résultats réels en matière de décrochage (sous forme binaire, d'après les données des sessions

[7] La valeur *e* représente la base du logarithme naturel, ce qui correspond à environ 2,718.

d'automne 2013 et 2014)[8]. Nous avons choisi la spécification de modèle ayant la plus basse valeur de perte de logarithme en tant que meilleur modèle prédictif.

3.4 Taux prédits de décrochage

Nous avons ensuite utilisé les estimations de coefficients afin de générer la probabilité prédite de décrochage pour chaque élève des cohortes de la session d'automne 2013 et la session d'automne 2014. Compte tenu des estimations de coefficients du modèle prédictif (c.-à-d. les $\hat{\beta}$), les caractéristiques individuelles et les programmes, les facteurs de risque fondés sur le SAE ainsi que les résultats des évaluations de chaque élève ont été intégrés à la formule axée sur le modèle ci-dessus afin d'obtenir chacune des prédictions quant à la probabilité de décrocher avant l'obtention du diplôme.

La formule mathématique sous-entendue par le modèle de régression logistique pour obtenir chacune des prédictions quant à la probabilité de décrocher des études collégiales ($\hat{P}$) est la suivante :

$$\frac{e^{\hat{\beta}_0+\hat{\beta}_1\times Gender+\hat{\beta}_2\times Age+\hat{\beta}_3\times Cred+\hat{\beta}_4\times Sch+\hat{\beta}_5\times Urb+\hat{\beta}_6\times HSavg+\hat{\beta}_7\times Assess+\hat{\beta}_8\times Risk}}{1+e^{\hat{\beta}_0+\hat{\beta}_1\times Gender+\hat{\beta}_2\times Age+\hat{\beta}_3\times Cred+\hat{\beta}_4\times Sch+\hat{\beta}_5\times Urb+\hat{\beta}_6\times HSavg+\hat{\beta}_7\times Assess+\hat{\beta}_8\times Risk}}$$

À l'échelle individuelle, la probabilité prédite pour chaque élève est exprimée sur une échelle allant de 0 à 1,0 (ou de 0 à 100 %). Les probabilités prédites sont (conformément à l'analyse dans l'introduction) centrées sur le taux global de décrochage chez les différentes cohortes.

Fait important, les élèves dont les caractéristiques individuelles et des programmes, les résultats au SAE et les résultats aux évaluations sont, d'après le modèle, liés à une probabilité accrue de décrochage ont tendance à révéler des probabilités prédites de décrochage supérieures, et vice versa. Toutefois, c'est l'ensemble complet des caractéristiques qui détermine la probabilité prédite de décrochage de chaque élève, ce qui constitue l'attribut essentiel et le point fort d'une méthode de modélisation.

Il convient également de souligner que certains élèves aux probabilités prédites élevées ne décrochent pas concrètement de leurs études, et l'inverse est également vrai. Il s'agit d'une réalité inhérente au caractère prédictif de ce qui consiste essentiellement en un exercice statistique.

[8] La formule relative à la perte de logarithme est la suivante $-\frac{1}{N}\sum_{i=1}^{N} p_i \log(\hat{p}_i) + (1-p_i)\log(1-\hat{p}_i)$, où les valeurs p_i et $\hat{p}_i$ indiquent les résultats réels de décrochage de même que la probabilité prédite de décrochage chez l'élève i, respectivement.

4. Rendement du modèle prédictif

4.1 Résultats des estimations du modèle prédictif

Les résultats du modèle logit utilisé pour générer les taux prédits de décrochage figurent à l'annexe Tableau A.1. Nous ne traiterons pas de ces résultats en détail ici parce que les détails précis du modèle sous-jacent ne sont pas essentiels dans le présent document, lequel est plutôt axé sur les prédictions que le modèle génère.

Cela dit, les prédictions ne seront valables que dans la mesure où les modèles sur lesquels elles s'appuient le sont également, de sorte qu'une analyse sommaire des résultats d'estimation du modèle est justifiée. Le tableau présente les estimations des paramètres du modèle logit. Étant donné le caractère non linéaire du modèle logit, celles-ci n'ont pas de signification intuitive directe, mais l'orientation des effets et la signification statistique des estimations des paramètres sont dignes d'intérêt.

Les principales constatations sont les suivantes :

- Les femmes ont nettement moins tendance à décrocher que les hommes.
- Les élèves âgés (c.-à-d. les 23-26 ans, et tout particulièrement les 27 ans ou plus) montrent des taux de décrochage inférieurs à ceux des jeunes élèves.
- La région, compte tenu de l'adresse de l'élève à l'étape de la demande d'admission au Collège Mohawk (en milieu urbain, en milieu rural, à l'étranger, données manquantes), importe peu.
- Les élèves au certificat montrent les taux de décrochage les plus élevés, pendant que les élèves au certificat d'études supérieures montrent un taux de décrochage inférieur à celui des élèves au diplôme d'études supérieures (le groupe omis/témoin).
- Les élèves dont les notes en début de programme sont faibles, tout particulièrement ceux dont les notes sont égales ou inférieures à D plus, révèlent des taux supérieurs de décrochage.
- Il existe des différences marquées dans les taux de décrochage entre écoles.
- Parmi les catégories de risque axées sur le SAE, « clarté sur le plan professionnel » et tout particulièrement « engagement envers l'éducation » sont des paramètres significatifs de prévision quant aux élèves qui décrochent.
- Les résultats des évaluations en lecture et en maths sont importants.

4.2 Les prédictions : rendement du modèle

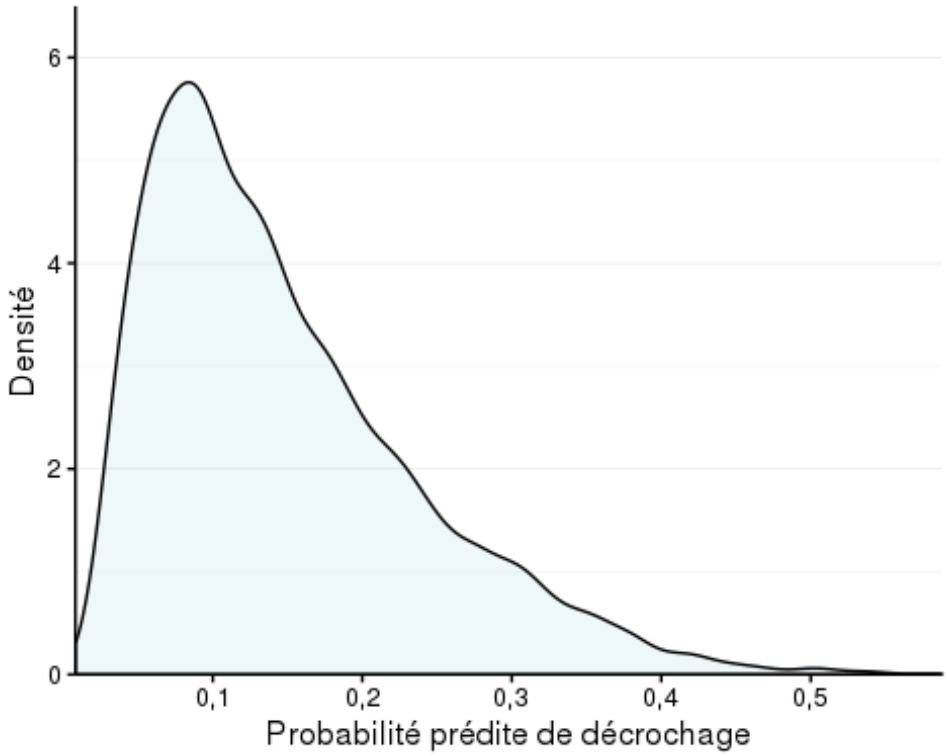
Les taux prédits de décrochage

Nous avons ensuite utilisé le modèle prédictif pour générer la probabilité prédite de décrochage relativement à chaque élève des cohortes de 2013 et de 2014, suivant ce qui est décrit au préalable. Les taux prédits de décrochage qui en ont résulté figurent dans le graphique 1, lequel représente la fonction de densité de probabilité des prédictions. Sur l'axe horizontal du groupe 1 se trouve la gamme de probabilités prédites possibles de décrochage, lesquelles vont de 0 à 1,0. L'axe vertical montre essentiellement la

proportion d'élèves à l'étape où ils sont ventilés dans trois différents niveaux, et le total de ces valeurs correspond à 1 (c.-à-d. il permet donc de saisir l'échantillon au complet).

En fait, les observations vont de probabilités prédites de décrochage très faibles jusqu'à un maximum d'environ 0,5, la plupart étant inférieures à 0,3, ce qui montre que les probabilités prédites de décrochage sont relativement faibles chez la plupart des élèves. Voilà qui n'a rien d'étonnant, dans un contexte où seulement 15,8 % des élèves dans l'échantillon d'estimation (cohortes de 2005 à 2012) et 15,3 % de ceux dans l'échantillon d'essai (cohortes de 2013 et de 2014) décrochent dans les faits. Le sommet de la ventilation se situe à environ 0,10, et la ventilation est désaxée vers la droite.

Graphique 1 : Densité de probabilité de décrochage



Niveaux de risque et taux réels de décrochage

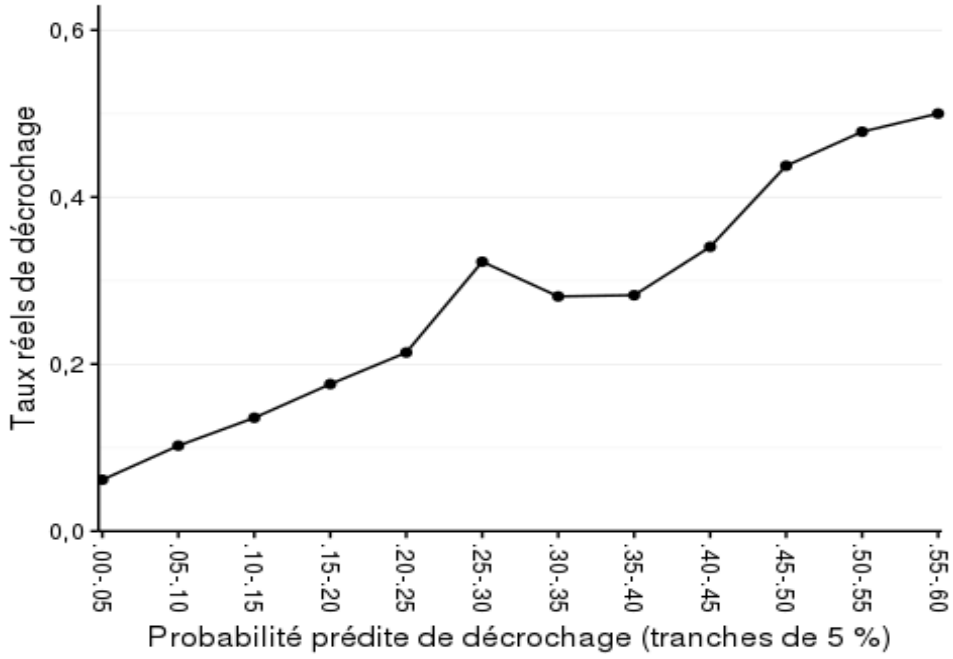
Une façon d'évaluer la capacité d'un modèle à prédire avec exactitude les taux de décrochage consiste à comparer chaque probabilité prédite de décrochage générée par le modèle prédictif avec les taux réels de décrochage. Pour ce faire, nous avons réparti les cohortes de 2013 et de 2014 pour lesquelles des taux de décrochage étaient prédits en 20 groupes définis par leurs niveaux de risque prédits, puis ceux-ci ont été mis en comparaison avec le taux réel de décrochage de chacun de ces groupes (lequel est, bien entendu, connu).

Le graphique 2 révèle que les valeurs de risque prédites sont en forte corrélation avec les taux réels de décrochage. Chaque valeur sur l'axe horizontal représente le groupe d'élèves d'après leur valeur de risque conformément aux attributions par le modèle prédictif, pendant que les valeurs correspondantes sur l'axe vertical représentent les taux réels de décrochage chez ces groupes. Autrement dit, les élèves auxquels le modèle attribue une valeur de faible risque présentent des taux inférieurs de décrochage, tandis que ceux auxquels une valeur de fort risque est attribuée montrent des taux supérieurs de décrochage.

Une telle corrélation étroite et positive serait attendue si les prédictions étaient dans l'échantillon (c.-à-d. des prédictions calculées d'après l'échantillon d'estimation, les données de 2005 à 2012 dans notre cas), puisque le modèle est estimé pour correspondre aux taux réels de décrochage. Toutefois, pareille corrélation étroite positive n'est pas systématiquement garantie dans les prédictions hors échantillon (c.-à-d. les prédictions calculées en fonction d'un nouvel échantillon, les données de 2013 et de 2014 dans le cas présent). Le graphique 2 : Ventilation des probabilités prédites de décrochage par rapport aux taux réels de décrochage montre donc que le modèle prédictif permet très bien de classer les élèves d'après leur tendance à décrocher des études collégiales.

Ces résultats semblent également indiquer que le modèle devrait convenir pour le classement des nouveaux élèves selon leurs niveaux de risque, de telle sorte que les programmes de soutien aux élèves peuvent, à titre d'exemple, être efficacement ciblés, ce qui compte parmi les buts principaux de l'élaboration d'un tel modèle prédictif.

Graphique 2 : Ventilation des probabilités prédites de décrochage par rapport aux taux réels de décrochage



Prédire les décrocheurs et les élèves assidus

Une autre façon de mesurer le rendement d'un modèle prédictif consiste à vérifier le pourcentage de prédictions exactes au total (qu'on appelle également exactitude), de prédictions positives erronées et de prédictions négatives erronées. Le calcul de ces statistiques dépend toutefois du choix d'un seuil qui permet d'attribuer à chaque élève un résultat prédit binaire de décrochage (c.-à-d. on prédit que les élèves décrocheront ou seront assidus).

À titre d'exemple, si le seuil est établi à 50 %, les élèves pour qui les probabilités prédites de décrochage à l'échelle individuelle sont supérieures à 0,50 seront des décrocheurs d'après les prédictions, tandis que ceux dont les mêmes probabilités sont égales ou inférieures à 0,50 seront des élèves assidus selon les prédictions. Les résultats prédits de décrochage qui en découlent sont ensuite compilés par rapport aux résultats réels de décrochage (encore en ce qui concerne les cohortes d'essai de 2013 et de 2014 pour lesquelles nous générons des prédictions et savons au bout du compte si l'élève a décroché ou est demeuré assidu).

Il n'existe toutefois pas un seuil exact unique à employer pour classer les élèves par catégorie de cette façon. De plus, une valeur de démarcation aura pour effet d'attribuer de façon arbitraire à l'une ou l'autre des catégories (c.-à-d. les décrocheurs ou les élèves assidus) ceux qui se situent près de la valeur de démarcation et dont les probabilités prédites de décrochage sont presque identiques (disons 0,29 ou 0,31, si la limite est établie à 0,30), alors qu'en réalité la différence quant à la probabilité de décrochage dans leur cas n'est que minime. En résumé, ni les valeurs de démarcation, ni les prédictions qu'elles génèrent ne sont en soi extraordinaires, et n'importe quel seuil choisi peut poser problème.

Bien qu'une valeur de démarcation de 50 % puisse par intuition sembler attirante pour certains, il ne s'agit pas en règle générale d'un seuil valable afin de répartir les particuliers selon les prédictions : ceux qui décrocheront et ceux qui seront assidus. Compte tenu qu'il existe un ensemble restreint de paramètres de prédiction et que les prédictions qui en découlent ont tendance à former des grappes vers la gauche, ce qui témoigne des taux globaux de décrochage relativement faibles qui sont expliqués par le modèle, une valeur de démarcation de 50 % occasionnerait le discernement d'un nombre très restreint de décrocheurs parmi les élèves.

En règle générale, lorsque vient le temps de choisir une valeur de démarcation (ou un seuil), il y a une corrélation négative entre, d'une part, les prédictions exactes au total et, d'autre part, les prédictions positives et négatives erronées. Le choix d'un seuil (ou d'une valeur de démarcation) élevé quant à la probabilité de décrochage afin de discerner les décrocheurs et les élèves assidus selon les prédictions se traduira habituellement par (suivant ce qui est mentionné au préalable) une sous-estimation du nombre global de décrocheurs (puisqu'un nombre restreint d'élèves présentera des probabilités prédites au-delà de ce seuil) et aura tout particulièrement pour effet de classer de nombreux décrocheurs réels dans la catégorie des élèves assidus selon les prédictions. L'inverse se produira si le seuil choisi est trop bas.

Il est donc coutumier de recourir à différents seuils pour mettre à l'essai un modèle et répondre à la question suivante : « Dans quelle mesure le modèle permet-il de prédire avec exactitude les élèves qui seront assidus et ceux qui décrocheront? » Le tableau A.2 en annexe montre : i) les prédictions exactes au

total; ii) les prédictions positives erronées; et iii) les prédictions négatives erronées, en fonction de différents seuils. Les prédictions exactes sont évidentes : les décrocheurs selon les prédictions décrochent et les élèves assidus selon les prédictions font preuve d'assiduité. Les prédictions positives erronées représentent les cas où les élèves devaient décrocher selon les prédictions mais se sont révélés assidus, tandis que les prédictions négatives erronées représentent les cas où les élèves étaient assidus selon les prédictions mais ont décroché dans les faits.

Pour ce qui est du modèle choisi, l'établissement d'un seuil de 10 % génère de nombreuses prédictions négatives erronées ainsi qu'un nombre relativement restreint de prédictions positives erronées. Comme le montre le tableau A.2, dans les cas où le seuil est supérieur, comme à 20 % ou à 30 %, il y a alors beaucoup plus de prédictions négatives erronées, mais nettement moins de prédictions positives erronées parce que, selon les prédictions, les élèves sont moins nombreux à décrocher lorsque les seuils sont élevés.

Dans un tel exercice, le seuil à adopter dépendra au bout du compte de l'utilisation qui sera faite des prédictions obtenues. De fait, le choix d'une valeur de démarcation privilégiée devient à la fois une décision politique et un exercice statistique et dépendra des coûts estimatifs des erreurs d'un côté ou de l'autre : les prédictions positives erronées par rapport aux prédictions négatives erronées, ou la surestimation du nombre de décrocheurs par rapport à celui des élèves assidus.

Dans les deux cas, l'établissement d'enseignement doit assumer des coûts : une prédiction positive erronée s'apparente à une fausse alarme et risque d'occasionner une intervention superflue et coûteuse auprès des élèves qui n'ont pas vraiment besoin d'aide. Quant à la prédiction négative erronée, elle représente l'élève à risque qui passe inaperçu(e), n'obtient pas de l'aide au moment où il ou elle en aurait besoin, et risque par conséquent de décrocher. La valeur de démarcation idéale permet d'équilibrer ces coûts et dépend au bout du compte de nombreux facteurs, notamment l'efficacité et les coûts des interventions offertes.

5. Classifications du risque des élèves (CRE)

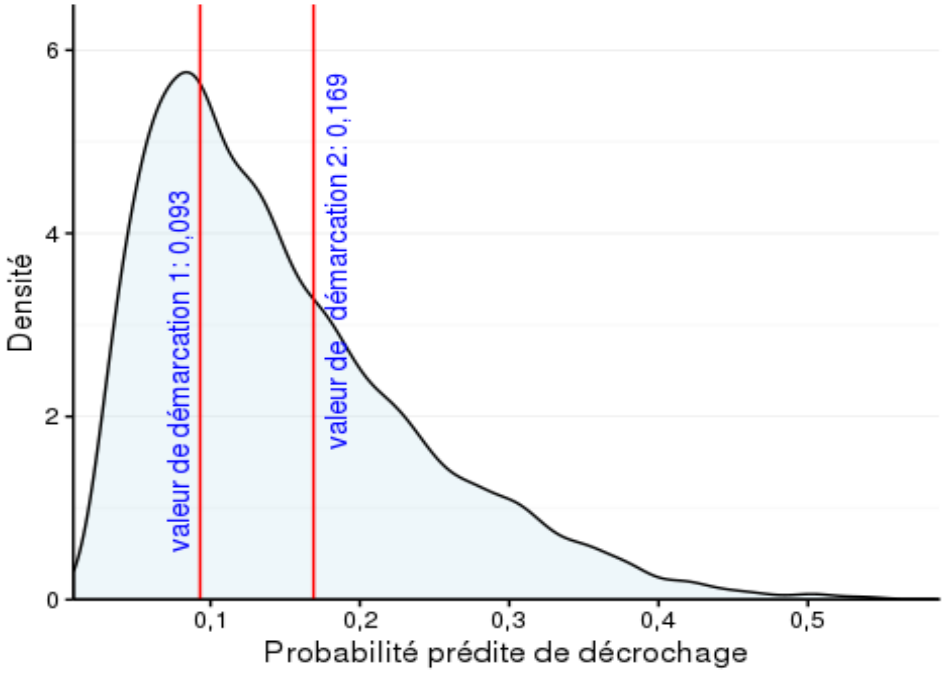
L'un des objectifs d'un projet de recherche de portée générale consiste à mettre à l'essai les différentes stratégies d'aide pédagogique aux élèves instaurées à l'intention de la nouvelle cohorte d'élèves de 2015 dans les différents niveaux de risque des élèves. Nous avons donc réparti la ventilation complète des niveaux de risque (ou des taux prédits de décrochage) en trois tranches, au moyen de deux valeurs de démarcation (valeur de démarcation 1< valeur de démarcation 2).

Le graphique 3 montre de nouveau la ventilation des probabilités prédites de décrochage des études collégiales chez les cohortes arrivées en 2013 et en 2014 mises ensemble. Nous avons choisi les valeurs de démarcation de façon à ce que les élèves soient répartis équitablement dans les trois classifications du risque des élèves (CRE); autrement dit, chaque groupe compte pour 33,3 % de la population d'élèves.

Les élèves dont les probabilités prédites de décrochage sont inférieures ou égales à la valeur de démarcation 1 entrent dans le groupe à faible risque. Ceux dont les probabilités prédites sont supérieures à la valeur de démarcation 1, mais inférieures à la valeur de démarcation 2, entrent dans le groupe à risque

moyen. Enfin, ceux dont les probabilités prédites sont supérieures ou égales à la valeur de démarcation 2 entrent dans le groupe à fort risque.

Graphique 3 : Ventilation des probabilités prédites de décrochage



Puisque la ventilation des probabilités prédites de décrochage est désaxée vers la droite, et que la valeur maximale de la probabilité prédite individuelle de décrochage telle qu'elle est générée par le modèle prédictif correspond à 58,7 %, les valeurs de démarcation qui donnent cette ventilation égale d'élèves ne sont pas très élevées. La valeur de démarcation 1, laquelle permet de discerner le groupe à faible risque, se situe à 0,093 (ou une probabilité prédite de décrochage de 9,3 %) et la valeur de démarcation 2, laquelle permet de distinguer les groupes à risque moyen et à fort risque, se fixe à 0,169 (ou 16,9 %).

Le tableau 1 présente les taux de décrochage des trois CRE pour chacune des cohortes d'élèves d'essai (2013 et 2014). Il révèle des taux croissants de décrochage, à partir du groupe à faible risque jusqu'au groupe à fort risque, ce qui était à prévoir.

Une fois les deux cohortes mises en commun, le pourcentage de décrocheurs dans le groupe à fort risque (24 %) est passablement supérieur au taux global de décrochage (16 %) et, bien entendu, supérieur à celui du groupe à risque moyen (14 %) et, tout particulièrement, à celui du groupe à faible risque (seulement 9 %). De cette façon, le tableau montre également un autre point de vue quant à l'exactitude du modèle prédictif et à son efficacité dans le classement des élèves par niveau de risque.

Tableau 1 : Taux de décrochage (en %) et nombre de décrocheurs (N) par classification du risque des élèves

	2013		2014		Ensemble	
Niveau de risque	Taux	N	Taux	N	Taux	N
Faible	9	163	8	145	9	308
Moyen	13	250	15	246	14	496
Fort	23	431	26	412	24	843
Total	15	844	16	803	16	1647

Note : Valeur de démarcation 1 = 9,3 %, valeur de démarcation 2 = 16,9 %

5.1 Analyse des CRE : Quels élèves figurent dans quels groupes de risque?

Ce que ces résultats représentent

Du tableau 2 au tableau 15, le présent document montre la ventilation des élèves dans les trois CRE provenant des cohortes de 2013 et de 2014 combinées (les résultats des deux cohortes sont très semblables) parmi la gamme de variables représentant les caractéristiques des élèves et des programmes, les indicateurs de risque du SAE et les résultats des évaluations.

Pendant qu'un bon nombre des variables prises en compte englobent certaines catégories qui, chacune, tendent à être en corrélation avec le décrochage (être de sexe masculin, avoir de faibles notes, etc.), le modèle prédictif permet de trier les liens entre le décrochage et chaque facteur en particulier, compte tenu également de l'ensemble des autres facteurs. Par conséquent, certains facteurs qui peuvent sembler importants s'ils sont pris isolément (comme ci-dessous) peuvent ne pas être significatifs une fois inclus dans le modèle général (l'inverse peut également se produire)[9].

Le fait de générer les niveaux prédits de risque à l'échelle individuelle fait alors intervenir la prise en compte de tous les renseignements sur les élèves et l'attribution à ces derniers d'une valeur globale de risque fondée sur le modèle prédictif.

Enfin, les CRE permettent ensuite de classer les élèves par niveau de risque dans les catégories à faible risque, à risque moyen et à fort risque telles qu'elles sont décrites au préalable.

[9] Ce phénomène est simplement attribuable aux corrélations entre les paramètres de prédiction. À titre d'exemple, dans un modèle de régression simple qui n'englobe que le sexe, il est constaté que le taux de décrochage des hommes est supérieur à celui des femmes. Si nous y ajoutons le domaine d'études, cet écart entre les sexes sera réduit si les hommes ont tendance à se trouver dans des domaines d'études où les taux de décrochage sont supérieurs, pendant que l'effet initial du paramètre de prédiction relatif au sexe est attribuable au moins en partie à la corrélation de celui-ci avec l'autre facteur de risque.

Variables des élèves et des programmes

Le tableau 2 révèle que la ventilation selon le sexe diffère entre les CRE. Les femmes comptent pour 47 % de la population globale d'élèves des cohortes de 2013 et de 2014 mises en commun, tandis que le pourcentage de femmes s'établit à seulement environ 36 % dans le groupe à fort risque et à 56 % dans le groupe à faible risque. Autrement dit, les femmes sont enclines à afficher une tendance moindre en matière de décrochage.

Tableau 2 : Ventilations selon le sexe (en %) par classification du risque des élèves

	Faible	Moyen	Fort	Ensemble
Hommes	44	50	64	53
Femmes	56	50	36	47
Total	100	100	100	100

Le tableau 3 révèle que, comparativement aux groupes à risque moyen et à fort risque, le groupe à faible risque compte une proportion supérieure d'élèves âgés. Environ 34 % des élèves de 23 ans et plus se trouvent dans le groupe à faible risque, tandis que 21 % de ces élèves sont dans le groupe à risque moyen et 15 %, dans le groupe à fort risque. C'est donc dire que les élèves âgés ont moins tendance à décrocher des études collégiales.

Tableau 3 : Ventilations selon l'âge (en %) par classification du risque des élèves

	Faible	Moyen	Fort	Ensemble
17 ans et moins	5	6	4	5
18 ans	19	26	21	22
19 ans	13	22	27	21
20-22 ans	28	25	33	29
23-26 ans	19	11	10	13
27 ans et plus	15	9	5	10
Total	100	100	100	100

Comme en témoigne le tableau 4, la ventilation des élèves quant à la situation à l'étranger, en milieu urbain et en milieu rural ne diffère pas considérablement entre les CRE. Le groupe à risque moyen compte des pourcentages légèrement inférieurs d'élèves étrangers et des pourcentages supérieurs d'élèves en milieu urbain comparativement aux deux autres groupes. En ce qui concerne les cohortes de 2013 et de 2014 mises en commun, environ 3 % du groupe à risque moyen se compose d'élèves étrangers, pendant qu'environ 6 % des groupes à faible risque et à fort risque sont composés de ces mêmes élèves. Quant au groupe à faible risque, il compte un pourcentage inférieur (82 %) d'élèves locaux qui habitent en milieu urbain

comparativement aux groupes à risque moyen et à fort risque qui en comptent 86 % et 85 %, respectivement.

Tableau 4 : Ventilations selon la situation régionale (en %) par classification du risque des élèves

	Faible	Moyen	Fort	Ensemble
Milieu rural	11	11	10	10
Milieu urbain	82	86	85	84
Étranger	6	3	6	5
Total	100	100	100	100

Comme le montre le tableau 5, une tendance se manifeste clairement quant à la mesure dans laquelle les ventilations selon la moyenne des notes à l'école secondaire diffèrent par CRE. Après examen des cohortes de 2013 et de 2014 mises en commun, environ 13 % des élèves à faible risque atteignaient une moyenne égale ou supérieure à A moins à l'école secondaire, tandis que seulement 4 % des élèves à fort risque ont maintenu une telle moyenne. À l'inverse, environ 29 % des élèves à fort risque avaient une moyenne égale ou inférieure à D plus à l'école secondaire, tandis que seulement 6 % des élèves à faible risque atteignaient une moyenne en deçà de C. Quant au groupe à risque moyen, sa ventilation selon la moyenne à l'école secondaire s'apparente beaucoup à la ventilation globale, la moyenne de la majorité des élèves s'établissant entre C et B plus.

Tableau 5 : Ventilations selon la moyenne à l'école secondaire (en %) par classification du risque des élèves

	Faible	Moyen	Fort	Ensemble
F	1	2	6	3
D	1	3	10	4
D plus	4	9	13	9
C	9	15	15	13
C plus	14	19	14	16
B	17	17	12	15
B plus	14	12	8	11
A moins	8	6	3	6
A	4	2	1	2
A plus	1	1	0	1
Données manquantes	26	15	18	19
Total	100	100	100	100

Le tableau 6 montre les ventilations de l'échantillon par école. Environ 31 % des élèves à faible risque sont inscrits à des programmes donnés à l'école de justice et de mieux-être, tandis qu'environ 2 % des élèves à fort risque sont inscrits à ces programmes. Environ 9 % des élèves à faible risque sont inscrits aux programmes menant à un certificat d'études supérieures à l'école d'administration des affaires, où il n'y a aucun élève des groupes à risque moyen ou à fort risque. De plus, environ 22 % des élèves à fort risque sont inscrits à l'école des études interdisciplinaires et 27 % font des études en gestion, tandis que seulement 1 % et 8 % des élèves du groupe à faible risque sont inscrits aux programmes de ces écoles.

Tableau 6 : Ventilations selon l'établissement scolaire (en %) par classification du risque des élèves

	Faible	Moyen	Fort	Ensemble
Sciences du bâtiment et de la construction	7	6	5	6
Technologie chimique et environnementale	4	5	5	5
Informatique et TI	7	5	3	5
Électrotechnologie	2	3	6	4
Cours préparatoire de technologie en génie	0	1	4	2
Études supérieures (administration des affaires)	9	0	0	3
Services à la personne	9	19	10	13
Études interdisciplinaires	1	5	22	9
Études en justice et en bien-être	31	14	2	16
Études en gestion	8	13	27	16
Technologie mécanique et industrielle	4	4	6	5
Médias et divertissement	10	16	7	11
Études en administration de bureau	6	6	1	4
Métiers spécialisés	2	2	2	2
Total	100	100	100	100

Comme le montre le tableau 7, la ventilation des élèves au chapitre des titres d'études diffère également d'une CRE à l'autre. Chez le groupe à fort risque, environ 33 % des élèves sont au certificat tandis qu'il n'y a pas d'élèves à fort risque qui sont au certificat d'études supérieures. Chez le groupe à faible risque, environ 17 % des élèves sont au certificat d'études supérieures, tandis qu'il n'y a qu'environ 3 % des élèves qui sont au certificat. Autrement dit, les élèves au certificat d'études supérieures sont, d'après les prédictions, à risque nettement moindre de décrocher que ceux dans un programme menant à un certificat ordinaire.

Tableau 7 : Ventilations selon le titre d'études (en %) par classification du risque des élèves

	Faible	Moyen	Fort	Ensemble
Certificat	3	12	33	16
Diplôme	52	58	43	51
Diplôme d'études supérieures	28	30	23	27
Certificat d'études supérieures	17	0	0	6
Total	100	100	100	100

Variables du Sondage d'accueil des élèves (SAE)

Suivant ce qui est mentionné au préalable, les variables de risque fondées sur les questions du Sondage d'accueil des élèves (SAE) du Collège Mohawk représentent les indicateurs individuels hypothétiques du risque qu'un élève décroche des études collégiales, d'après ces variables prises en compte de façon indépendante et une à la fois. Par contre, le modèle prédictif tient compte de tous les renseignements sur les élèves, dont non seulement les variables de risque du SAE, mais l'ensemble des autres facteurs compris dans le modèle, de sorte qu'il est le meilleur indicateur des variables, notamment les indicateurs de risque du SAE, qui constituent les meilleurs paramètres de prédiction du décrochage.

Après examen des marqueurs de risque du SAE — « clarté sur le plan professionnel »; « confiance »; « engagement envers l'éducation » et « transition » — des tableaux 8 à 11, chez les élèves qui ont passé le SAE, ceux qui entrent dans la catégorie d'élèves à risque moyen ou à fort risque dans la CRE sont également plus susceptibles d'être considérés comme à risque en fonction de la variable pertinente du SAE.

Tableau 8 : Ventilations selon le marqueur de risque (clarté sur le plan professionnel) (en %) par classification du risque des élèves

	Faible	Moyen	Fort	Ensemble
Non	69	68	37	58
Oui	8	11	10	10
Données manquantes	23	21	52	32
Total	100	100	100	100

Tableau 9 : Ventilations selon le marqueur de risque (confiance) (en %) par classification du risque des élèves

	Faible	Moyen	Fort	Ensemble
Non	59	53	29	47
Oui	18	26	19	21
Données manquantes	23	21	52	32
Total	100	100	100	100

Tableau 10 : Ventilations selon le marqueur de risque (engagement envers l'éducation) (en %) par classification du risque des élèves

	Faible	Moyen	Fort	Ensemble
Non	77	77	44	66
Oui	1	2	4	2
Données manquantes	23	21	52	32
Total	100	100	100	100

Tableau 11 : Ventilations selon le marqueur de risque (transition) (en %) par classification du risque des élèves

	Faible	Moyen	Fort	Ensemble
Non	66	66	38	57
Oui	11	13	10	11
Données manquantes	23	21	52	32
Total	100	100	100	100

Contrairement aux autres variables du SAE, chez les élèves qui ont passé le SAE, le pourcentage d'élèves qui travaillent 15 heures ou plus (ce qui est considéré comme une indication du risque de décrocher des études collégiales) ne diffère pas considérablement d'une CRE à l'autre (tableau 12).

Tableau 12 : Ventilations selon le marqueur de risque (travailler 15 heures et +) (en %) par classification du risque des élèves

	Faible	Moyen	Fort	Ensemble
Non	43	46	28	39
Oui	34	33	20	29
Données manquantes	23	21	52	32
Total	100	100	100	100

Résultats des évaluations

Il y a également des différences dans la ventilation des résultats des évaluations en mathématiques, en lecture et en rédaction entre les différentes CRE. Le groupe à fort risque a tendance à compter un pourcentage supérieur d'élèves ayant obtenu de piètres résultats comparativement aux autres groupes, tandis que le groupe à faible risque a tendance à compter un pourcentage supérieur d'élèves qui réussissent bien à ces évaluations. De plus, le groupe à fort risque comporte un pourcentage passablement supérieur d'élèves sans résultat d'évaluation en lecture ou en rédaction. La section suivante montre les résultats en détail quant à la mesure dans laquelle les ventilations varient d'une CRE à l'autre.

Comme le montre le tableau 13, le groupe à faible risque compte un pourcentage supérieur (environ 28 %) d'élèves ayant obtenu un résultat dans les trois meilleures catégories (6, 7 ou 8) comparativement au groupe à fort risque (moins de 1 %). De plus, le groupe à fort risque montre une proportion supérieure (environ 24 %) d'élèves ayant obtenu un résultat dans les trois pires catégories (1, 2 ou 3) comparativement au groupe à faible risque (moins de 2 %).

Tableau 13 : Ventilations selon le résultat de l'évaluation en maths (en %) par classification du risque des élèves

	Faible	Moyen	Fort	Ensemble
1	0	2	10	4
2	0	4	8	4
3	1	5	6	4
4	2	6	4	4
5	4	6	2	4
6	8	4	0	4
7	9	3	0	4
8	11	1	0	4
Données manquantes	63	68	69	67
Total	100	100	100	100

Le tableau 14 montre les ventilations selon le résultat de l'évaluation en lecture par CRE. Environ 36 % des élèves de chacun des groupes à risque moyen et à faible risque n'ont pas de résultats en lecture, pendant que 60 % des élèves à fort risque n'ont pas de résultats d'évaluation. Le groupe à faible risque compte le pourcentage le plus élevé d'élèves ayant obtenu un résultat dans les trois meilleures catégories comparativement au groupe à fort risque. Les groupes à fort risque et à risque moyen comportent un pourcentage supérieur d'élèves (environ 26 %) ayant obtenu un résultat dans les trois pires catégories comparativement au groupe à faible risque (environ 11 %).

Tableau 14 : Ventilations selon le résultat de l'évaluation en lecture (en %) par classification du risque des élèves

	Faible	Moyen	Fort	Ensemble
1	2	8	11	7
2	4	9	8	7
3	5	9	7	7
4	7	9	5	7
5	8	9	4	7
6	10	8	2	7
7	13	7	2	7
8	14	5	1	7
Données manquantes	36	36	60	44
Total	100	100	100	100

Le tableau 15 affiche les ventilations selon le résultat de l'évaluation en rédaction par CRE. Fait qui s'apparente aux résultats en lecture, environ 36 % des élèves dans les groupes à risque moyen et à faible risque n'ont pas de résultat en lecture, contre environ 60 % des élèves à fort risque. Chez le groupe à fort risque, la majorité (27 %) du pourcentage restant (40 %) des élèves ont obtenu un résultat dans la pire catégorie. Chez le groupe à faible risque, environ 23 % du pourcentage restant des élèves dont le résultat n'est pas manquant ont obtenu un résultat dans la pire catégorie.

Tableau 15 : Ventilations selon le résultat de l'évaluation en rédaction (en %) par classification du risque des élèves

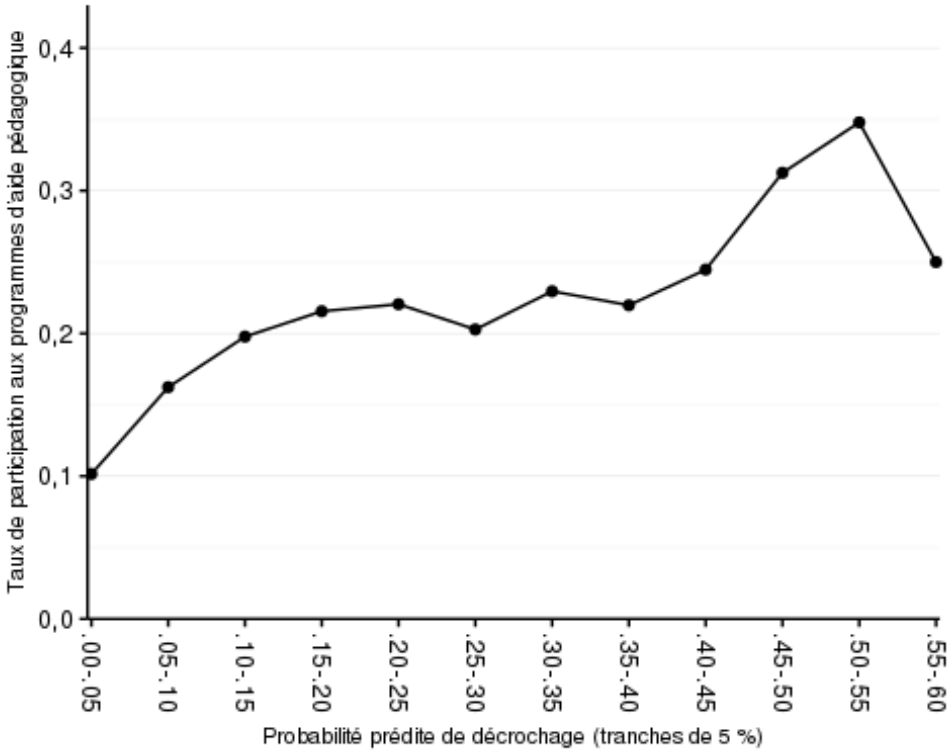
	Faible	Moyen	Fort	Ensemble
1	23	33	27	28
2	41	31	13	28
Données manquantes	36	36	60	44
Total	100	100	100	100

5.2 Taux de participation aux programmes d'aide pédagogique

Le graphique 4 montre le lien général entre les niveaux de risque et la participation aux programmes d'aide pédagogique des élèves. L'aide pédagogique est une valeur binaire qui permet tout simplement de répertorier les élèves qui ont cherché à obtenir l'aide d'un conseiller et ceux qui ne l'ont pas fait. Par conséquent, cette analyse ne représente ni la fréquence de l'aide pédagogique fournie à un élève, ni la

durée ou le type d'interaction d'aide pédagogique qui a eu lieu. Cependant, il est à espérer que les élèves ayant besoin des programmes d'aide pédagogique (c.-à-d. les élèves à risque) sont ceux qui y participent dans les faits. Voilà ce que les résultats semblent montrer : une corrélation générale entre la probabilité prédite de décrochage et les taux de participation aux programmes d'aide pédagogique. Autrement dit, plus le niveau de risque est fort, plus le pourcentage d'élèves qui participent à des programmes d'aide pédagogique est prononcé. Il existe quelques raisons qui peuvent expliquer ce lien. La première, c'est que les élèves à fort risque croyaient avoir besoin le plus des services d'aide pédagogique et ont saisi l'occasion offerte[10].

Graphique 4 : Les probabilités prédites de décrochage par rapport aux taux de participation aux programmes d'aide pédagogique



[10] Dans un sondage auprès de plus de 60 000 élèves de niveau collégial au Canada, l'auteur Dietsche (2012) a constaté que les élèves qui disent avoir besoin de services de soutien n'y ont souvent pas accès dans les faits. Ce type de données est un élément moteur sous-jacent à la théorie et à la pratique de l'aide pédagogique perturbatrice (ou proactive). Il s'agit d'une méthode répandue parmi les conseillers à la réussite des élèves du Collège Mohawk, et il est possible que ce programme d'action directe auprès des élèves jugés à risque de décrochage soit la raison du lien observé dans les données. Ce phénomène sera approfondi à la phase 2 du projet.

Le tableau 16 révèle les élèves qui cherchent à obtenir une aide pédagogique dans l'ensemble des CRE. Les taux de participation à l'aide pédagogique sont les plus élevés chez les groupes à risque moyen et à fort risque, tandis qu'ils sont les plus bas chez le groupe à faible risque.

Tableau 16 : Taux de participation aux programmes d'aide pédagogique et taille des échantillons par classification du risque des élèves*

	2013		2014		Ensemble	
Niveau de risque	Taux	N	Taux	N	Taux	N
Faible	16	279	12	208	14	487
Moyen	21	390	20	323	21	713
Fort	22	424	21	324	22	748
Total	20	1093	17	855	19	1948

Remarque : Le taux de participation aux programmes d'aide pédagogique est calculé sous forme de ratio du nombre d'élèves ayant cherché à obtenir des services d'aide pédagogique à l'école au moins une fois par rapport au nombre total d'élèves au cours de la session.
**Valeur de démarcation 1 = 9,3 %; valeur de démarcation 2 = 16,9 %*

Le taux global de participation aux programmes d'aide pédagogique est inférieur d'environ 3 points de pourcentage chez la cohorte de l'automne 2014 (17%) comparativement à celle de l'automne 2013 (20 %). Le taux global de participation des cohortes de 2013 et de 2014 mises ensemble se fixe à environ 20 %.

La différence dans les taux de participation aux programmes d'aide pédagogique entre les groupes à risque moyen et à fort risque n'est que d'environ 1 point de pourcentage, tandis que le groupe à faible risque montre un taux de participation qui est inférieur d'environ 5 à 7 points de pourcentage à celui du groupe à risque moyen, et d'environ 6 à 9 points de pourcentage à celui du groupe à fort risque.

6. Analyse

Élaborer et utiliser un modèle prédictif pour discerner les élèves à risque de décrochage

Le présent document correspond à la première phase d'un vaste projet de recherche qui tente d'estimer les effets de trois méthodes différentes d'aide pédagogique aux élèves attribuées aléatoirement à tous les nouveaux élèves avant leur arrivée au Collège Mohawk à l'automne 2015. Parmi les différences dans ces méthodes, il y a les divers degrés de perturbation, la façon de joindre les élèves de même que la nature des services d'aide pédagogique offerts, dont les séances individuelles par rapport aux séances en groupe.

L'un des buts de ce vaste projet consiste à estimer les effets des différentes initiatives d'aide pédagogique non seulement à l'échelle globale, mais entre les élèves à différents niveaux de risque quant à la probabilité de décrochage des études au Collège Mohawk (sans obtenir de diplôme). De telles évaluations permettront

ensuite au Collège Mohawk de bien concevoir et cibler les initiatives à ce chapitre à l'avenir pour obtenir au bout du compte les meilleures améliorations possibles dans les résultats des élèves par rapport aux coûts de ces initiatives différentes.

Afin de discerner les taux de risque de décrochage chez les élèves, une méthode de modélisation prédictive a été adoptée. Le travail à ce chapitre s'appuie sur des projets collaboratifs menés au préalable par le Collège Mohawk et l'Initiative de recherche sur les politiques de l'éducation (IRPE) à propos du maintien aux études des élèves.

Une spécification de modèle de régression logistique est employée, dans laquelle le risque qu'un élève décroche est lié à une gamme de facteurs, dont les caractéristiques démographiques et du programme de l'élève (le sexe, l'âge, la région, les notes à l'école secondaire, l'école et le titre d'études); un ensemble d'indicateurs de risque élaborés par le Collège Mohawk à partir de son Sondage d'accueil des élèves (SAE) relativement à des concepts comme la clarté sur le plan professionnel, la préparation émotive, ainsi que le nombre d'heures travaillées dans des emplois externes; de même que les évaluations en lecture, en rédaction et en maths également tenues avant l'admission des élèves au Collège Mohawk.

Dans le cadre du présent projet, les modèles élaborés au cours des travaux préalables ont été actualisés, respécifiés et mis à l'essai de façon approfondie. Parmi ces essais, il y a eu en premier lieu la comparaison d'une vaste gamme de spécifications de modèles au moyen de l'échantillon d'estimation (c.-à-d. les cohortes d'élèves arrivés au Collège Mohawk de 2005 à 2012) et le recours à des techniques économétriques habituelles (c.-à-d. les fonctions de perte de logarithme) afin d'en arriver au modèle qui convient le mieux aux données.

D'autres essais ont été menés afin d'évaluer l'exactitude du modèle prédictif en ce qui touche la prédiction des taux de décrochage à l'aide d'un ensemble distinct de données d'essai (c.-à-d. les cohortes arrivées en 2013 et en 2014). Dans un ensemble d'essais, les probabilités prédites de décrochage (sur une échelle de 0 à 1,0 pour n'importe quel élève) ont été comparées aux taux réels de décrochage pour constater combien les probabilités prédites permettaient un bon suivi des taux réels de décrochage. D'autres essais ont fait intervenir la conversion des taux prédits de décrochage des élèves en des prédictions à l'échelle individuelle quant aux élèves qui persévèreront et à ceux qui décrocheront pour ensuite comparer de nouveau ces valeurs à ceux qui, dans les faits, poursuivent leurs études ou décrochent.

Le modèle a bien fonctionné dans l'un et l'autre des dénombrements, bien qu'il soit strictement axé sur les variables facilement accessibles dans les dossiers administratifs du Collège Mohawk, telles qu'elles sont répertoriées au préalable.

Les taux prédits de décrochage à l'échelle individuelle ont ensuite servi à classer les élèves — une fois de plus, ceux des cohortes d'essai de 2013 et de 2014 — en trois catégories de risque. Par souci de commodité, il a été déterminé arbitrairement que ces catégories seraient de taille égale (c.-à-d. chaque catégorie englobait le tiers des nouveaux élèves). Elles ont d'abord servi à mettre à l'essai de façon approfondie le potentiel de prédiction du modèle prédictif par la comparaison des taux réels de décrochage chez les

catégories à faible risque, à risque moyen et à fort risque avec les taux réels de décrochage; une fois de plus, le suivi obtenu était bon.

Une fois les élèves attribués à l'un des groupes de risque, une analyse descriptive a été menée pour discerner la ventilation des élèves dans ces groupes selon la panoplie des caractéristiques des élèves et des programmes, des indicateurs de risque découlant du SAE et des résultats des évaluations susmentionnés.

Bien entendu, les élèves présentant à l'échelle individuelle les facteurs de risque discernés par le modèle prédictif (p. ex., être un homme, avoir de faibles notes à l'école secondaire, présenter n'importe quel des facteurs de risque axés sur le SAE et définis par le Collège Mohawk, avoir de faibles résultats dans les évaluations) ont été trouvés chez les groupes à fort risque déterminés par le modèle prédictif. Mais tel ne fut pas le cas de nombreux élèves.

Constatations particulières

Voici certaines des constatations particulières :

- Les principaux facteurs déterminants de décrochage du Collège Mohawk avant l'obtention du diplôme tels qu'ils sont discernés par le modèle prédictif sont les suivants : les femmes, les élèves de 23 ans ou plus, ainsi que les élèves au certificat d'études supérieures présentent des taux de décrochage inférieurs à ceux des autres; les élèves au certificat et ceux dont les notes en début de programme sont faibles (tout particulièrement si elles sont égales ou inférieures à D plus) présentent des taux de décrochage supérieurs; il existe des différences appréciables dans les taux de décrochage entre écoles; la région (en milieu urbain, en milieu rural, à l'étranger) n'est pas statistiquement significative; parmi les catégories de risque axées sur le SAE, seules les catégories « clarté sur le plan professionnel » et tout particulièrement « engagement envers l'éducation » sont significatives; les résultats des évaluations en lecture et en maths sont également importants.
- À l'échelle individuelle, les taux prédits de décrochage des élèves (les décrocheurs étant les élèves qui n'ont pas amorcé la deuxième session de leur programme) générés par le modèle prédictif, lesquels étaient estimés pour les cohortes arrivées de 2005 et de 2012, puis mis à l'essai pour les cohortes arrivées en 2013 et en 2014, a permis un suivi étroit des taux réels de décrochage.
- Dans les trois catégories de risque des élèves, les taux réels de décrochage des élèves définis au moyen du modèle productif s'établissent à 24 % chez les élèves à fort risque, à 14 % chez les élèves à risque moyen et à 9 % chez les élèves à faible risque dans les cohortes d'essai de 2013 et de 2014, ce qui révèle la mesure dans laquelle le modèle prédictif distingue bien les élèves selon leur niveau de risque.
- Enfin, contrairement aux perceptions répandues chez les intervenants des affaires étudiantes comme quoi les élèves qui participent aux programmes d'aide pédagogique et de soutien sont ceux qui n'en ont pas réellement besoin [Dietsche (2012)], il appert que les groupes à fort risque et moyen ont participé le plus aux programmes actuels d'aide pédagogique aux élèves, à environ 22 %, tandis que seulement 14 % des élèves à faible risque y ont pris part (tous les résultats se rapportent de nouveau aux cohortes d'essai de 2013 et de 2014).

Tout ce travail permet la mise en place de la phase 2 du projet de recherche, où les autres types d'interventions d'aide pédagogique instaurés à l'intention de la cohorte arrivée en 2015 font l'objet d'une évaluation et leurs effets, d'une mise à l'essai en fonction des différents niveaux de risque des élèves.

Retombées et leçons tirées au Collège Mohawk

Parmi les constatations globales d'importance au Collège Mohawk, il y a celle comme quoi la méthode du modèle prédictif crée un continuum de taux prédits de décrochage des élèves, la majorité d'entre eux étant groupés dans un champ relativement étroit de probabilités prédites axées sur les taux de décrochage globaux réels des différentes cohortes : à environ 16 % relativement au taux de décrochage à la première session, lequel constitue le point de mire du présent document.

Par conséquent, une modification de la valeur de démarcation appliquée à la ventilation complète des probabilités prédites employées pour répartir les élèves en groupes de risque se traduira par des groupes dont l'ampleur et les caractéristiques différeront. Cet élément comporte des retombées sur la planification opérationnelle, comme le fait de déterminer le nombre et les caractéristiques des élèves à cibler, quelle que soit l'initiative de réussite des élèves. D'autres valeurs de démarcation pourront servir et d'autres groupes pourront être déterminés, car les trois groupes déterminés dans le présent document n'ont rien de sacré (ou d'exceptionnel).

La tenue du projet de recherche a également entraîné l'abandon par le Collège Mohawk de sa terminologie de départ, qualifiant l'élève de « prêt pour les études collégiales », de « sous-préparé » et de « à risque » pour adopter plutôt les concepts d'élèves à fort risque, à risque moyen et à faible risque[11]. D'autres travaux et prises en considération sont nécessaires afin de déterminer les termes optimaux pour décrire les différentes catégories à employer, qu'il y en ait trois, quatre, plus ou moins, et quelles que soient les valeurs de démarcation du décrochage prédit employées pour les définir.

Une mesure qui pourrait être mise en place consiste en un programme d'action directe davantage intentionnel et intense pour les élèves afin de rehausser les résultats de ces derniers, ce qui s'inscrirait dans l'importance que le collège accorde à la consultation proactive.

Une autre mesure potentielle, axée sur une mesure améliorée des niveaux de risque des élèves, pourrait être d'exiger la participation au Sondage d'accueil des élèves par divers leviers politiques dans les établissements d'enseignement, comme son intégration au processus d'acceptation ou d'inscription.

La mise à l'essai, l'évaluation et la déclaration de l'une ou l'autre des nouvelles méthodes pourra se traduire par des données significatives sur les pratiques et politiques efficaces d'action directe. De même, les autres

[11] Bien que la description de départ ait aidé le collège à en arriver à comprendre le concept de base, les descripteurs du risque fort, moyen et faible sont meilleurs. Fait à souligner, le concept de « prêt pour les études collégiales » risque de sous-entendre à tort qu'il n'y a aucun risque, pendant que le concept « à risque » peut sous-entendre qu'il n'y a pas de risque lié aux autres classifications.

données fournies par les élèves permettraient de brosser un tableau complet des besoins des élèves, de raffermir le modèle prédictif et d'éclairer les services de soutien aux interventions précoces.

En outre, le SAE sera probablement plus utile s'il fait appel à des échelles pour représenter les facteurs de risque pertinents plutôt que la méthode binaire actuellement employée, puisqu'une échelle fournira davantage de renseignements. De fait, il est peu sensé de créer dans un premier temps des ensembles d'« indicateurs de risque des élèves » binaires qui s'appuient sur les renseignements sous-jacents davantage détaillés et qui sont accessibles pour ensuite intégrer ces indicateurs aux modèles prédictifs. Une meilleure stratégie consisterait à insérer les variables brutes (échelonnées) du SAE directement aux modèles. En ce qui touche les indicateurs de risque, il conviendrait également de les vérifier de façon empirique plutôt que de les fonder sur les attentes *a priori* de ce qui peut témoigner de la situation à risque d'un élève.

Les constatations ont une autre retombée : le collège doit examiner les exigences d'accueil pour faire en sorte qu'elles aient l'effet souhaité. En particulier, le fait d'avoir une moyenne égale ou inférieure à D plus à l'école secondaire correspond à un important ensemble de variables dans les modèles prédictifs. Les élèves dans cette situation étaient davantage susceptibles de se trouver dans le groupe à fort risque, une fois toutes leurs caractéristiques prises en compte. Par conséquent, il y a lieu de se pencher sur des exigences d'accueil minimales ou l'accueil conditionnel afin de déterminer si de tels changements politiques pourraient entraîner une amélioration des taux de maintien aux études en contribuant à la conception d'interventions ou de programmes qui sont propices au soutien des élèves ayant de faibles notes à l'école secondaire. D'autres programmes d'accès pourront également convenir davantage aux élèves de cette catégorie.

Il faut également mener d'autres recherches pour bien concevoir les intentions et buts scolaires des élèves. Bien que les taux de décrochage soient substantiels, il se peut que certains élèves n'aient pas d'emblée l'intention de mener à bien leurs études. Par exemple, certains élèves pourront se servir de l'admission à un programme donné au Collège Mohawk comme tremplin vers une autre expérience d'enseignement postsecondaire. Le fait de bien concevoir les buts des élèves permettrait donc d'améliorer la planification du Collège Mohawk et pourra fournir des renseignements utiles à propos de la mise à niveau des programmes, de l'avancement et des voies d'accès aux autres établissements d'EPS.

Enfin, être capable de concevoir les raisons pour lesquelles les élèves décrochent de leurs études collégiales (et en quoi consistent leurs plans) permettrait de brosser un tableau complet de la réussite et du maintien aux études des élèves. Dans l'un ou l'autre de ces cas, le décrochage des études collégiales pourra être considéré comme une réussite par l'élève, ce qui mérite également d'être relaté.

Limites du modèle et autres possibilités

L'analyse présentée ici comporte quelques limites générales qu'il y a lieu de mentionner. Premièrement, dans notre cas, les résultats du modèle prédictif étaient fonction du comportement des élèves au cours de la période relative à l'estimation (2005 à 2012), puis mis à l'essai auprès des cohortes de 2013 et de 2014 (d'autres méthodes de création, de formation et de mise à l'essai des échantillons peuvent être employées). Si le comportement de l'élève a changé depuis ce temps (p. ex., le décrochage des particuliers se situe à des

taux généralement différents ou à des taux relativement différents entre les groupes ou les caractéristiques particulières des élèves par des façons que ne peut saisir le modèle), le modèle ne reflétera plus le comportement actuel. Cette réalité est inévitable dans tous les modèles prédictifs formés à partir de données antérieures puis mis à l'essai sur de futures données.

Deuxièmement, le modèle prédictif donne les moyens de cibler les élèves dans le contexte d'initiatives de réussite des élèves à l'échelle individuelle puis d'évaluer l'efficacité des programmes mis en place. De tels exercices statistiques peuvent éclairer les futures politiques de façon très importante (permettant aux décisions d'être factuelles), mais ils ne permettent pas, ni ne peuvent permettre, de déterminer les politiques réelles de l'établissement d'enseignement en soi. Autrement dit, les politiques d'un établissement d'enseignement seront au bout du compte alimentées par ces objectifs globaux particuliers, les ressources dont il dispose, ainsi que d'autres facteurs potentiels.

La troisième limite à l'analyse entreprise ici, c'est que le rendement d'un modèle prédictif ne peut être aussi valable que les données fournies. Fait particulier à la présente analyse, elle est restreinte à l'utilisation des renseignements obtenus couramment par le Collège Mohawk avant l'admission des élèves au collège. Cet état de choses découle d'une très bonne raison : le Collège Mohawk souhaite cibler ces élèves d'emblée, avant le début de la première session.

Au chapitre de la collecte des données, une première étape consisterait à faire en sorte que toutes les données pertinentes qui sont actuellement accessibles au Collège Mohawk soient employées dans le modèle prédictif. Les données sur l'aide financière (les prêts) consentie aux élèves ou les données sur les choix de programme tirées des demandes d'admission au collège pourraient constituer des exemples de données qui sont au moins potentiellement accessibles et qui pourraient servir. Les renseignements en provenance du Sondage d'accueil des élèves, utilisés dans le modèle, pourraient également être intégrés sous leur forme d'origine davantage détaillée plutôt que sous leur forme binaire représentée par les facteurs de risque et créée par le collège Mohawk.

Une deuxième étape pourrait consister à recourir aux données accessibles pour créer d'autres variables. Par exemple, il serait possible de créer un indicateur du contexte socioéconomique par l'utilisation des renseignements du code postal que fournissent les élèves dans l'optique d'une mise en lien avec d'autres sources de données (comme le recensement) afin de discerner les caractéristiques du milieu où l'élève a vécu avant son admission au collège (p. ex., revenu moyen, niveaux de scolarité, logement, langue).

Troisièmement, si des services d'aide pédagogique (ou d'autres stratégies de réussite des élèves) étaient offerts après le début de la session (ou au cours d'une session ultérieure), le rendement des élèves au départ pourrait également s'intégrer aux modèles, ce qui permettrait indubitablement d'améliorer en grande partie leur rendement. Parmi les exemples à ce chapitre, il y a la déclaration précoce des cours (même simplement les présences seraient vraisemblablement utiles), les notes intérimaires et les notes définitives pour prédire le maintien aux études durant les sessions subséquentes.

Enfin, de plus en plus de renseignements à propos des élèves sont accessibles par voie électronique et pourraient faire partie d'une analyse sur le maintien aux études des élèves pour ensuite servir à prédire le

maintien aux études des élèves et contribuer à cibler et à mesurer les interventions. La participation des élèves aux cours en ligne constitue un bon exemple. La plupart des cours comportent désormais un important volet en ligne, ou exige du moins une certaine participation en ligne de la part de l'élève, et ces renseignements pourraient être extraits afin de bien concevoir les résultats des élèves et de bien prédire, parmi les élèves, ceux qui sont le plus à risque de décrocher ou menacent de le faire dans très peu de temps.

Phase 2 du projet

Dans la seconde phase du présent projet, nous chercherons à évaluer une nouvelle initiative d'action directe et d'aide pédagogique au Collège Mohawk, conçue pour appuyer avantage l'ensemble des nouveaux élèves qui fréquentent le campus Fennell à la session d'automne 2015. Tous les élèves ont reçu aléatoirement l'une des trois trousses d'action directe et d'aide pédagogique précoce avant le début des cours.

Voies de recherche supplémentaires

Les modèles prédictifs du maintien aux études des élèves servent principalement à trois fins :

- Aider un établissement d'enseignement à bien concevoir dans quelle mesure le décrochage est lié aux divers facteurs ou caractéristiques, notamment à propos des élèves et des programmes, qui font partie des modèles.
- Utiliser les prédictions générées par le modèle à propos du décrochage au niveau individuel et des élèves dans l'optique de cibler les activités ou programmes axés sur les élèves, notamment les initiatives de réussite des élèves.
- Estimer les effets des initiatives de réussite des élèves (ou autres) ciblées à l'aide d'une méthode fondée sur un modèle prédictif par le recours à l'analyse de discontinuité, laquelle est une méthode d'estimation fondée précisément sur les types de valeurs de démarcation qui peuvent servir à cibler les élèves, ou à estimer les effets des autres initiatives dans l'ensemble des différents niveaux de risque des élèves.

Le premier point est, suivant ce qui est indiqué, davantage orienté vers une conception améliorée du maintien aux études des élèves, et les modèles purement « descriptifs », lesquels peuvent et devraient généralement être différents des « modèles prédictifs », consisteront habituellement en la méthode privilégiée si tel est l'objectif unique. Cependant, la méthode générale est essentiellement la même : recourir à des modèles statistiques pour mettre en lien le décrochage (ou le maintien aux études) avec divers facteurs d'intérêt. Les modèles descriptifs se distinguent par leur spécification en mode narratif ou (comme le concept le sous-entend) qui décrit le maintien aux études des élèves... sans aller plus loin.

Par ailleurs, les modèles prédictifs sont moins axés sur la narration et portent plutôt sur l'élaboration de spécifications qui s'agencent le mieux aux données et génèrent les prédictions les plus exactes des résultats des élèves. Par exemple, les modèles prédictifs peuvent se révéler plus flous que les modèles purement descriptifs, notamment par l'inclusion d'un nombre accru d'interactions entre variables. Si le tableau brossé

du maintien aux études des élèves risque de ne pas être aussi limpide, il peut toutefois générer des prédictions améliorées.

Nous avançons qu'une modélisation accrue de ces deux types joue un rôle fondamental pour faire progresser notre conception du maintien aux études des élèves dans son état actuel, ainsi que cibler et mettre à l'essai de nouvelles initiatives visant l'amélioration des résultats des élèves.

Nous espérons donc voir, avant tout, un nombre accru d'établissements d'EPS s'investir dans ces exercices de modélisation comme base en vue d'une conception améliorée de leurs élèves et des résultats qu'ils obtiennent, puis mettre au point de meilleures politiques axées sur les élèves au moyen d'une méthode factuelle.

En deuxième lieu, il est possible d'améliorer les modèles actuels dans presque chaque cas, tout particulièrement par l'ajout de renseignements ou de données aux modèles; nous avons esquissé quelques orientations pour ce faire dans la sous-section précédente, notamment :

1. Ajouter des variables supplémentaires en fonction des renseignements déjà accessibles aux établissements d'enseignement, dont ceux relatifs à l'aide financière, ou les données sur les choix de programme tirées des demandes d'admission au collège, de même qu'utiliser de façon plus rigoureuse certains des renseignements déjà inclus dans les modèles, comme les données contenues dans le Sondage d'accueil des élèves.
2. Créer des variables supplémentaires par la jonction avec d'autres sources de données (p. ex., le recensement afin de saisir les antécédents socioéconomiques) à partir des données actuelles (c.-à-d. le code postal).
3. Ajouter une « déclaration précoce » à propos des élèves, afin de bien concevoir et prédire le comportement et les résultats de l'élève à la suite de son admission à l'établissement d'enseignement.
4. Explorer les sources potentiellement colossales de renseignements électroniques qui sont de plus en plus recueillis à propos des élèves, dont ceux en lien avec la participation aux cours.

En troisième lieu, les modèles prédictifs se présentent sous diverses formes. Celle utilisée dans le cadre du présent projet s'appuie sur une méthode d'un modèle de régression logistique relativement simple; cependant, d'autres méthodes peuvent être mises à l'essai et comparées comme dans certains des travaux de recherche en éducation, essentiellement aux États-Unis, cités à la section Analyse documentaire dans les pages précédentes. À notre sens, les algorithmes d'apprentissage automatique avancés constituent une voie particulièrement importante en vue des nouveaux travaux, bien qu'il soit nécessaire d'élaborer et de mettre à l'essai ces méthodes avec une rigueur accrue avant de déterminer leur efficacité. De plus, ils présentent un inconvénient net : ils sont beaucoup plus complexes que les simples méthodes de modélisation du genre employées ici, ce qui peut se révéler coûteux à plusieurs niveaux différents, notamment en ce qui touche leur utilisation élargie par les intervenants des établissements d'enseignement qui souhaiteront peut-être élaborer des modèles prédictifs, les actualiser et les mettre en application continuellement.

Enfin, il est possible de faire une utilisation nettement accrue des modèles prédictifs à l'étape du ciblage de la mise à l'essai des initiatives d'appui aux élèves, selon ce qui est également mentionné dans d'autres études citées dans le présent document, comme celles de l'auteur Delen (2010) et des auteurs Zhang et al. (2010).

Nous percevons des possibilités extraordinaires en ce qui touche l'ensemble de ces orientations dans le contexte de l'EPS. Puisse le présent document contribuer à une évolution en la matière.

Définitions

L'**analytique des données** consiste en des processus d'évaluation et d'analyse des données afin d'éclairer les décisions prises à tous les niveaux, comme les établissements d'enseignement, les organisations et les entreprises [van Barneveld et al. (2012)].

La **catégorie à fort risque** correspond aux élèves qui composent le tiers supérieur de la ventilation des probabilités prédites de décrochage. Ce groupe d'élèves est le moins susceptible de réussir et le plus susceptible de décrocher des études collégiales après la première session, de sorte qu'il nécessite le plus un soutien et une intervention. Au début du projet, ce groupe était appelé « à risque ».

L'**aide pédagogique perturbatrice ou proactive** correspond à la méthode perturbatrice d'aide pédagogique lancée par l'auteur Glennen (1975), laquelle préconise davantage d'interventions délibérées et de liens coopératifs en matière d'aide pédagogique afin d'accroître la motivation des élèves. Récemment, cette méthode a été qualifiée d'aide pédagogique proactive [Varney (2013)], laquelle est propice à une action directe proactive, à la prestation d'un soutien avant que l'élève en ait besoin, de même qu'à la promotion de liens solides entre le conseiller et l'élève.

L'**aide pédagogique durant le cycle de vie** correspond aux différents services d'aide pédagogique proposés aux nombreux éléments typiques de l'expérience vécue par l'élève au collège. D'une session ou d'une année à l'autre, des activités, échéanciers, défis et expériences prévisibles ont lieu régulièrement (c.-à-d. l'inscription, les échéanciers de paiement, les examens de mi-session et les examens finaux) et composent le cycle de vie d'un élève de niveau collégial. Les diverses activités qui permettent de reconnaître l'élève, d'en tenir compte et de l'appuyer tout au long de ces expériences fréquemment vécues composent l'aide pédagogique du cycle de vie.

La **catégorie à faible risque** englobe les élèves qui constituent le tiers inférieur de la ventilation des probabilités prédites de décrochage. Ce groupe d'élèves est le plus susceptible de réussir et le moins susceptible de décrocher des études collégiales après la première session, de sorte qu'il nécessite le moins de soutien. Au début du projet, ce groupe était appelé « prêt pour les études collégiales ».

La **catégorie à risque moyen** englobe les élèves qui constituent le tiers intermédiaire de la ventilation des probabilités prédites de décrochage. Ce groupe d'élèves occupe une place intermédiaire quant aux possibilités de réussir leurs études collégiales ou d'en décrocher après une session. Au début du projet, ce groupe était appelé « sous-préparé ».

L'**analytique prédictive**, un type d'analytique des données, consiste en un ensemble de technologies employées pour divulguer les liens et phénomènes qui, dans les grands volumes de données, peuvent servir à prédire le comportement et les événements [van Barneveld et al. (2012)].

Le **modèle prédictif** est un extrant du processus de modélisation prédictive qui sert à prédire un résultat d'intérêt, compte tenu des valeurs attribuées aux variables des paramètres de prédiction. Dans le contexte

du présent projet, le modèle prédictif désigne la spécification du modèle de régression employé pour prédire la probabilité de décrochage des études collégiales par un élève après une session. De telles prédictions ont également servi à déterminer les valeurs de démarcation pour discerner les catégories (groupes) d'élèves à faible risque, à risque moyen et à fort risque.

La **modélisation prédictive**, qui s'inscrit dans l'analyse prédictive, consiste en un ensemble de techniques mathématiques employées pour trouver un lien entre un résultat ou une variable dépendante, et un paramètre de prédiction ou une variable indépendante afin de prédire les valeurs inconnues ou nouvelles de la variable dépendante [Dickey (2012)].

Le **Sondage d'accueil des élèves (SAE)** correspond au sondage suivant l'admission mais précédant l'inscription auquel la majorité des nouveaux élèves répondent dans le cadre de leur transition vers le Collège Mohawk. Il se déroule sur le campus au même moment que les évaluations de réussite (EDR). Les EDR correspondent aux tests de placement en lecture, en rédaction et en mathématique que les nouveaux élèves passent suivant l'admission mais précédant l'inscription, durant leur transition vers le collège. D'après les résultats de leurs évaluations, les élèves sont placés dans des cours réguliers ou de rattrapage en communication ou en mathématiques. Ils obtiennent également les ressources et le soutien relativement à toute mise à jour nécessaire avant le début des cours.

Les **classifications du risque des élèves (CRE)** correspondent aux catégories à faible risque, à risque moyen et à fort risque des élèves dans l'ensemble de la ventilation des niveaux de risque (ou taux prédits de décrochage) répertoriés dans le cadre du présent projet de recherche. Les valeurs de démarcation sont choisies de façon à ce que les élèves soient répartis équitablement à l'échelle de la ventilation et que chaque groupe comporte 33,3 % de la population d'élèves.

Bibliographie

Aguiar, E., Lakkaraju, H., Bhanpuri, N., Miller, D., Yuhas, B. et K. L. Addison (mars 2015), « Who, when, and why: A machine learning approach to prioritizing students at risk of not graduating high school on time », dans *Proceedings of the Fifth International Conference on Learning Analytics and Knowledge* (p. 93-102), ACM.

Astin, A. W. (1997), « How "good" is your institution's retention rate? », dans *Research in Higher Education*, vol. 38 n° 6, p. 647 658.

Atiya, A. F. (2001), « Bankruptcy prediction for credit risk using neural networks: A survey and new results », dans *IEEE Transactions on neural networks*, vol. 12 n° 4, p. 929-935.

Baan, C. A., Ruige, J. B., Stolk, R. P., Witteman, J. C., Dekker, J. M., Heine, R. J. et E. J. Feskens (1999), « Performance of a predictive model to identify undiagnosed diabetes in a health care setting », dans *Diabetes Care*, vol 22 n° 2, p. 213-219.

Braxton, J. M., Doyle, W. R., Hartley III, H. V., Hirschy, A. S., Jones, W. A. et M. K. McClendon (2014), *Rethinking college student retention*, San Francisco (Californie), Jossey-Bass.

Braxton, J. M., Hirschy, A. S. et M. K. McClendon (2004), « Understanding and reducing college student departure », dans *ASHE-ERIC Higher Education Report*, vol. 30 n° 3.

Center for Community College Student Engagement (2014), *A matter of degrees: Practices to pathways (High-impact practices for community college student success)*, Austin (Texas), Université du Texas à Austin, Program in Higher Education Leadership.

College Student Achievement Project Team (2015), *College student achievement project: Final report 2015*, Toronto (Ontario), Collège Seneca d'arts appliqués et de technologie. Extrait de : http://csap.senecacollege.ca/docs/CSAP%20Cycle%202%20final%20report%2011Jun15.pdf

Conrad, M. et K. Morris (2010), *Voir la rétention en fonction des risques et non des taux : Une autre façon pour les établissements de gérer le maintien des effectifs*, Toronto (Ontario), Conseil ontarien de la qualité de l'enseignement supérieur.

Dekker, G., Pechenizkiy, M. et J. Vleeshouwers (juillet 2009), « Predicting students drop out: A case study », dans *Educational Data Mining*.

Delen, D. (2010), « A comparative analysis of machine learning techniques for student retention management », dans *Decision Support Systems*, vol. 49 n° 4, p. 498-506.

Dickey, D. A. (2012), Introduction to Predictive Modeling with Examples, SAS Global Forum 2012, article n° 337.

Dietsche, P. (2007), Étude pancanadienne sur les étudiants collégiaux de première année : Les caractéristiques des étudiants et l'expérience collégiale, Rapport 1, Gatineau (Québec), Association des collèges communautaires du Canada, Ressources humaines et Développement social Canada. Extrait de : https://pseinfosys.com/wp-content/uploads/2014/02/Dietsche-ACCC-2008.-Pan-Canadian-Study-of-First-Year-College-Students-1.pdf

Dietsche, P. (2008), Étude pancanadienne sur les étudiants collégiaux de première année : Les caractéristiques et l'expérience des étudiants autochtones, handicapés, immigrants et de minorités visibles, Rapport 2, Gatineau (Québec), Association des collèges communautaires du Canada, Ressources humaines et Développement des compétences Canada. Extrait de : https://pseinfosys.com/wp-content/uploads/2014/02/Dietsche-ACCC-2008.-Pan-Canadian-Study-of-First-Year-College-Students-2.pdf

Dietsche, P. (2009), The Ontario College Student Engagement Survey 2006-2009: Final report-project results, data modelling, tests of reliability and validity and future directions, Toronto (Ontario), préparé pour le ministère de la Formation et des Collèges et Universités de l'Ontario.

Dietsche, P. H. J. (2012), « Use of Campus Support Services by Ontario College Students », dans *La Revue canadienne d'enseignement supérieur*, vol. 42 n° 3, p. 65-92.

Federico, M., Vitolo, U., Zinzani, P. L., Chisesi, T., Clò, V., Bellesi, G. et V. Pavone (2000), « Prognosis of follicular lymphoma: a predictive model based on a retrospective analysis of 987 cases », dans *Blood*, vol. 95 n° 3, p. 783-789.

Finnie, R. et T. Qiu (2008), The patterns of persistence in post-secondary education in Canada: Evidence from the YITS-B dataset, Educational Policy Institute.

Foster, D. et R. Stine (2004), « Variable Selection in Data Mining: Building a predictive model for bankruptcy », dans *Journal of the American Statistical Association*, vol. 99 n° 466, p. 303-313. Extrait de : https://doi.org/10.1198/016214504000000287.

Fricker, T. (2015), « The relationship between academic advising and student success in Canadian colleges: A review of the literature », dans *College Quarterly*, vol. 18 n° 4.

Glennen, R. E. (1975), « Intrusive college counseling », dans *College Student Journal*, vol. 9 n° 1, p. 2-4.

Grites, T. J. (1979), *Academic Advising: Getting Us Through the Eighties* (AAHE-ERIC/Higher Education Research Report No. 7), Washington (District de Columbia), American Association of Higher Education.

Habley, W. R., Bloom, J. L. et S. Robbins (2012), *Increasing Persistence: Research-based strategies for college student success*, San Francisco (Californie), Jossey-Bass.

Hossain, M. et Y. Muromachi (2012), « A Bayesian network based framework for real-time crash prediction on the basic freeway segments of urban expressways », dans *Accident Analysis & Prevention*, vol. 45, p. 373-381.

Jia, P. et T. Maloney (2015), « Using predictive modelling to identify students at risk of poor university outcomes », dans *Higher Education*, vol. 70 n° 1, p. 127-149.

Kotsiantis, S. B. (2012), « Use of machine learning techniques for educational proposes: A decision support system for forecasting students' grades », dans *Artificial Intelligence Review*, vol. 37 n° 4, p. 331-344.

Kuh, G. D. (2008), *High-impact educational practices: What they are, who has access to them, and why they matter*, Washington (District de Columbia), Association of American Colleges & Universities.

Kuh, G. D., Kinzie, J., Schuh, J. H. et E. J. Whitt (2005), *Student success in college: Creating conditions that matter*, San Francisco (Californie), Jossey-Bass. Extrait de http://www.loc.gov/catdir/toc/ecip054/2004027912.html

Lakkaraju, H., Aguiar, E., Shan, C., Miller, D., Bhanpuri, N., Ghani, R. et K. L. Addison (août 2015), « A machine learning framework to identify students at risk of adverse academic outcomes », dans *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (p. 1909-1918). ACM.

Lin, S. H. (2012), « Data mining for student retention management », dans *Journal of Computing Sciences in Colleges*, vol. 27 n° 4, p. 92-99.

Mayhew, M., Rockenbach, A., Bowman, N., Seifert, T., Wolniak, G., Pascarella, E. et P. Terenzini (2016), *How College Affects Students: 21st Century Evidence that Higher Education Works, Volume 3*, San Francisco (Californie), Jossey-Bass, p. 523-574.

Collège Mohawk (2013), *Mohawk's 5 Point Student Success Plan*, Hamilton (Ontario).

Murphey, Y. L., Chen, Z., Kiliaris, L., Park, J., Kuang, M., Masrur, A. et A. Phillips (juin 2008), « Neural learning of driving environment predictions for vehicle power management », dans *IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, (p. 3755-3761), IEEE.

Nandeshwar, A., Menzies, T. et A. Nelson (2011). « Learning patterns of university student retention », dans *Expert Systems with Applications*, vol. 38 n° 12, p. 14984-14996.

Ontario Academic Advising Professionals (s.d.), Terms of Reference, extrait de http://oaap.ca/terms-of-reference/

Orpwood, G., Schollen, L., Leek, G., Marinelli-Henriques, P. et H. Assiri (2012), *Projet de 2011 portant sur les mathématiques au niveau collégial : rapport final*, Toronto (Ontario), Collège Seneca d'arts appliqués et de technologie. Extrait de : http://collegemathproject.senecac.on.ca/cmp/en/pdf/FinalReport/2011/CMP_2011_Final_Report%20-%2002Apr12%20pmh.pdf

Perlich, C., Dalessandro, B., Raeder, T., Stitelman, O. et F. Provost (2014), « Machine learning for targeted display advertising: Transfer learning in action », dans *Machine learning*, vol. 95 n° 1, p. 103-127.

Phua, C., Lee, V., Smith, K. et R. Gayler (2010), A comprehensive survey of data mining-based fraud detection research. Extrait de : arXiv preprint arXiv:1009.6119.

Poirier, W. J. (juin 2015), *Deliberate orientation and transition practices as part of a broader student success strategy: A comparative study of three large urban Ontario colleges* (thèse), Institut d'études pédagogiques de l'Ontario, Université de Toronto, Toronto (Ontario). Extrait de : http://hdl.handle.net/1807/69456

Reason, R. D. (2009), « An Examination of Persistence Research Through the Lens of a Comprehensive Conceptual Framework », dans *Journal of College Student Development*, vol. 50 n° 6, p. 659-682.

Sara, N. B., Halland, R., Igel, C. et S. Alstrup (2015), « High-school dropout prediction using machine learning: A Danish large-scale study », dans *Proceedings* (p. 319), Presses universitaires de Louvain.

Terenzini, P. T. et R. D. Reason (2005), Parsing the first year of college: Rethinking the effects of college on students, présenté à l'assemblée annuelle de l'Association for the Study of Higher Education, Philadelphie (Pennsylvanie).

Thammasiri, D., Delen, D., Meesad, P. et N. Kasap (2014), « A critical assessment of imbalanced class distribution problem: The case of predicting freshmen student attrition », dans *Expert Systems with Applications*, vol. 41 n° 2, p. 321-330.

Tinto, V. (1975), « Dropout from higher education: A theoretical synthesis of recent research », dans *Review of Educational Research*, vol. 45, p. 89-125.

Tinto, V. (1993), *Leaving college: Rethinking the causes and cures of student attrition* (2e éd.), Chicago (Illinois), University of Chicago Press.

van Barneveld, A., Arnold, K. E. et J. P. Campbell (2012), *Anaytics in higher education: Establishing a common language* (Educause Learning Initiative No. 1) (p. 11), EDUCAUSE. Extrait de :

https://library.educause.edu/resources/2012/1/analytics-in-higher-education-establishing-a-common-language.

Varney, J. (2013), « Proactive Advising », dans J. K. Drake, P. Jordan et M. A. Miller (éd.), *Academic Advising Approaches: Strategies that teach students to make the most of college*, San Francisco (Californie), Jossey-Bass.

Wiggers, R. et C. Arnold (2011), *Définir, mesurer et assurer la « réussite des étudiants » des collèges et universités de l'Ontario* (Rapport En question n^{o} 10), Toronto (Ontario), Conseil ontarien de la qualité de l'enseignement supérieur. Extrait de : http://www.heqco.ca/SiteCollectionDocuments/AtIssueStudent%20Success%20ENG.pdf

Yu, C. H., DiGangi, S., Jannasch-Pennell, A. et C. Kaprolet (2010), « A data mining approach for identifying predictors of student retention from sophomore to junior year », dans *Journal of Data Science*, vol. 8 n^{o} 2, p. 307-325.

Zhang, Y., Oussena, S., Clark, T. et H. Kim (2010), « Using data mining to improve student retention in higher education: A case study », dans *International Conference on Enterprise Information Systems*.

Conseil ontarien
de la qualité de
l'enseignement
supérieur

Un organisme du gouvernement de l'Ontario