Prédiction du Décrochage Scolaire

Hilaire TOUYEM [*] Rakissiwinde BATIONO [†]

Superviseur : Pr. Laurent Charlin

1 Introduction

Le *décrochage scolaire* est un défi majeur pour les systèmes éducatifs à l'échelle mondiale. En affectant gravement les perspectives d'avenir des individus et en générant des coûts économiques considérables pour la société, il devient crucial de détecter précocement les élèves à risque afin de proposer des interventions ciblées. Ce projet s'inscrit dans ce contexte et présente une étude comparative de plusieurs modèles prédictifs basés sur l'apprentissage automatique pour identifier les élèves vulnérables au décrochage scolaire. En exploitant des caractéristiques démographiques, socio-économiques, macroéconomiques et académiques, nous examinons les performances de plusieurs algorithmes, tels que le *Support Vector Machine* , le *k-Nearest Neighbors*, les *Réseaux de Neurones*, et la *Régression Logistique* . Après avoir évalué les performances de ces modèles, nous appliquons la méthode *SHAP* (SHapley Additive exPlanations) sur le "*meilleur*" modèle pour analyser l'importance des caractéristiques et formuler des recommandations concrètes pour prévenir le décrochage scolaire.

2 Revue de Littérature

Les recherches récentes sur la prédiction du décrochage scolaire montrent l'utilisation de divers modèles de machine learning pour identifier les élèves à risque, chacun ayant des avantages spécifiques en fonction des caractéristiques des données. Parmi ces modèles, le Support Vector Machine (SVM) est particulièrement adapté aux ensembles de données déséquilibrés, ce qui est le cas de notre jeu de données. En étendant le SVM pour les classifications multiclasses, il permet de détecter précisément les élèves à risque de décrochage [10]. Le k-Nearest Neighbors (k-NN), qui repose sur la similarité locale des données, est efficace dans les contextes où des élèves présentant des caractéristiques communes (comme les antécédents scolaires ou le contexte socio-économique) montrent des tendances similaires au décrochage [5]. Les Réseaux de Neurones (RN), quant à eux, sont adaptés aux environnements riches en données, comme l'apprentissage en ligne, où ils peuvent identifier des schémas complexes de

[*] hilaire.touyem@hec.ca, MSc en sciences des données et analytique d'affaires, HEC Montréal

[†] rakissiwinde.bationo@hec.ca, MSc en intelligence d'affaires, HEC Montréal

comportement [1]. Enfin, la régression logistique multinomiale, tout en étant un modèle plus simple, est particulièrement utile pour son interprétabilité, ce qui est crucial pour identifier clairement les facteurs de risque de décrochage scolaire [2].

L'interprétabilité des modèles est un aspect clé, notamment pour les décisions ayant des conséquences sociales et éducatives. Selon [9], des techniques comme SHAP permettent de rendre les modèles complexes plus transparents, en identifiant les caractéristiques les plus influentes dans les prédictions. L'interprétabilité est d'autant plus importante dans les contextes éducatifs, où les interventions doivent être justifiables et compréhensibles par les parties prenantes, telles que les enseignants et les administrateurs scolaires [9].

3 Description des Données

Les données utilisées proviennent de [8] et sont disponibles sur Kaggle., Elles concernent des étudiants inscrits dans divers programmes de premier cycle au sein d'un établissement d'enseignement supérieur au Portugal. Le jeu de données comprend 4 424 observations et 34 caractéristiques, regroupées dans 4 catégories. La table ci-dessous présente les 34 caractéristiques ainsi qu'une brève description.

Caractéristiques	Description
Marital status	Le statut matrimonial de l'étudiant (par exemple, célibataire, marié, divorcé)
Nationality	Nationalité de l'étudiant
Displaced	Indicateur d'exil de l'étudiant
Gender	Sexe de l'étudiant
Age at enrollment	Âge de l'étudiant à l'inscription
International	Indicateur d'étudiant international
Mother's qualification	Niveau de qualification de la mère
Father's qualification	Niveau de qualification du père
Mother's occupation	Occupation de la mère
Father's occupation	Occupation du père
Educational special needs	Besoins éducatifs spéciaux
Debtor	Indicateur d'endettement
Tuition fees up to date	Frais de scolarité à jour
Scholarship holder	Boursier
Unemployment rate	Taux de chômage global
Inflation rate	Taux d'inflation global
GDP	Produit intérieur brut global
Application mode	Mode ou type de candidature que l'étudiant a soumis pour s'inscrire au cours
Application order	Indique l'ordre dans lequel l'étudiant a postulé pour le cours.
Course	Le cours ou programme de diplôme dans lequel l'étudiant est inscrit.
Daytime/evening attendance	Type de présence (journée/soirée)
Previous qualification	Qualification antérieure
Curricular units 1st sem (credited)	Unités curriculaires validées au 1er semestre
Curricular units 1st sem (enrolled)	Unités curriculaires inscrites au 1er semestre
Curricular units 1st sem (evaluations)	Unités curriculaires avec évaluations au 1er semestre
Curricular units 1st sem (approved)	Unités curriculaires approuvées au 1er semestre
Curricular units 1st sem (grade)	Note obtenue dans les unités curriculaires du 1er semestre
Curricular units 1st sem (without evaluations)	Unités curriculaires sans évaluations au 1er semestre
Curricular units 2nd sem (credited)	Unités curriculaires validées au 2nd semestre
Curricular units 2nd sem (enrolled)	Unités curriculaires inscrites au 2nd semestre
Curricular units 2nd sem (evaluations)	Unités curriculaires avec évaluations au 2nd semestre
Curricular units 2nd sem (approved)	Unités curriculaires approuvées au 2nd semestre
Curricular units 2nd sem (grade)	Note obtenue dans les unités curriculaires du 2nd semestre
Curricular units 2nd sem (without evaluations)	Unités curriculaires sans évaluations au 2nd semestre

TABLE 1 – Vue d'ensemble des caractéristiques

Le graphique ci-dessus illustre la distribution de la variable cible. On observe que la catégorie **Graduate** est majoritaire avec 2 209 occurrences, suivie par **Dropout** avec 1 421 occurrences, et enfin **Enrolled** avec 794 occurrences. Cela montre une disparité importante dans la répartition des résultats académiques des étudiants.

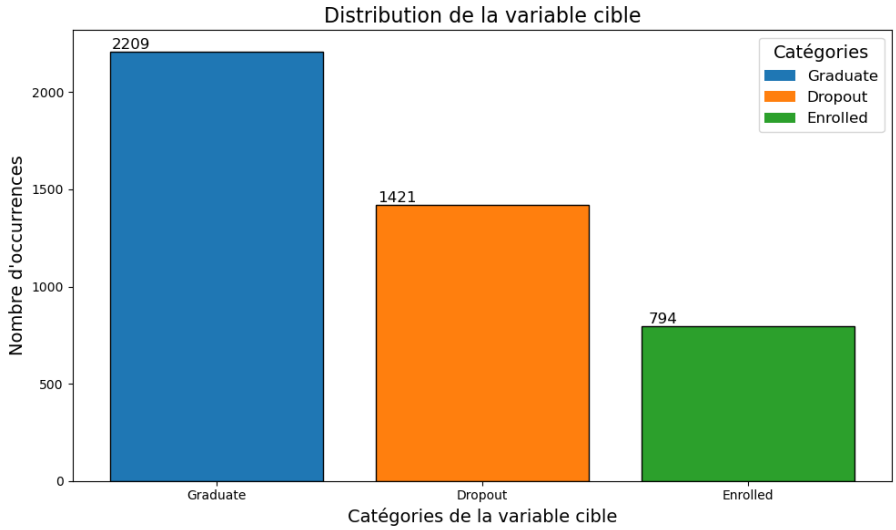


FIGURE 1 – Distribution de la variable cible

4 Méthodologie

4.1 Algorithmes d'apprentissage automatique utilisés

Pour prédire les résultats des étudiants (*Graduate, Dropout, Enrolled*), quatre algorithmes d'apprentissage automatique ont été utilisés. Le **SVM** excelle pour séparer des classes complexes dans des espaces multidimensionnels grâce à des fonctions noyaux, offrant une grande précision dans la classification non linéaire [11]. Le **k-NN** attribue une classe en se basant sur la proximité des données, mais sa performance peut être affectée par la dimensionnalité, ce qui le rend moins performant pour des données volumineuses ou bruitées [12]. La **régression logistique**, rapide et robuste, calcule les probabilités d'appartenance aux classes et fournit des prédictions fiables pour des problèmes linéaires tout en étant simple à interpréter [4]. Enfin, les **réseaux de neurones**, avec leur structure inspirée du cerveau, capturent des relations complexes à travers un apprentissage itératif et flexible, ce qui les rend particulièrement efficaces pour des tâches impliquant de grandes quantités de données et des interactions non linéaires [3].

4.2 Entraînement du modèle et optimisation des hyperparamètres

Pour optimiser les performances des algorithmes tout en évitant le surajustement et le sous-ajustement, deux techniques principales ont été utilisées. La **validation croisée en 5-fold** a permis une évaluation robuste en divisant les données en 5 sous-ensembles et en répétant les entraînements et

validations. Ensuite, le **réglage des hyperparamètres** a été effectué à l'aide de grilles de recherche et de recherches aléatoires, optimisant des paramètres spécifiques à chaque modèle, comme les noyaux pour le SVM, le nombre de voisins pour k-NN, et les architectures pour les réseaux de neurones. Ces ajustements ont permis d'adapter les modèles aux données et d'améliorer leur généralisation. Ces étapes renforcent la fiabilité des résultats tout en minimisant les erreurs liées à une mauvaise configuration des algorithmes.

4.3 Indicateurs de performance

Pour évaluer les performances des modèles, nous avons utilisé quatre métriques principales : l'**accuracy** pour mesurer les prédictions correctes, la **précision** pour éviter les faux positifs, le **rappel** pour identifier les vrais positifs, et le **F-Score**, qui combine précision et rappel de manière équilibrée.

Dans un contexte de **déséquilibre des classes**, l'accuracy peut être trompeuse, car elle favorise la classe majoritaire. Par conséquent, le choix du modèle optimal dans notre projet repose principalement sur des mesures comme la précision, le rappel et le F-Score, qui évaluent mieux les performances sur les classes minoritaires. Toutes les métriques sont calculées en *macro-moyenne*, ce qui signifie que la moyenne est faite pour chaque classe. Par exemple, le rappel en macro-moyenne est donné par :

$$\text{Recall}_{macro} = \frac{1}{N} \sum_{i=1}^{N} \left[\frac{\text{TP}_i}{\text{TP}_i + \text{FN}_i} \right]$$

où N est le nombre total de classes et $\frac{\text{TP}_i}{\text{TP}_i+\text{FN}_i}$ représente le rappel pour la classe i ; avec TP_i : le nombre de vrais positifs pour la classe i ; FN_i : le nombre de faux négatifs pour la classe i.

4.4 Mesure d'importance des caractéristiques

Pour interpréter les résultats des modèles complexes, la méthode **SHAP** a été utilisée pour évaluer l'importance des caractéristiques. Basée sur la théorie des jeux, SHAP décompose les contributions individuelles des variables dans chaque prédiction, permettant une explication précise et transparente [6]. Cette méthode offre des explications à la fois **locales**, pour des observations spécifiques, et **globales**, pour comprendre les tendances générales sur l'ensemble des données [7]. Cette méthode est essentielle dans notre étude car elle permet de formuler des recommandations personnalisées pour les étudiants à risque de décrochage, en fonction de leurs caractéristiques individuelles.

Pour les modèles de classification, SHAP permet d'expliquer les prédictions du modèle en décomposant chaque probabilité prédite pour une classe donnée en contributions des différentes caractéristiques d'entrée. Supposons un modèle de classification avec K classes, et pour une instance x, une probabilité prédite pour chaque classe C_i, notée $p_i(x)$. L'objectif de SHAP est de décomposer $p_i(x)$ en une somme additive [7] :

$$p_i(x) = \phi_0^i + \sum_{j=1}^{n} \phi_j^i \tag{1}$$

où :

– ϕ_0^i : La valeur de base, représentant la probabilité moyenne de la classe C_i,

– ϕ_j^i : La contribution de la caractéristique x_j à la probabilité de la classe C_i,

– n : Le nombre de caractéristiques.

Ainsi, la probabilité de chaque classe est expliquée par les contributions spécifiques de chaque caractéristique.

La contribution de chaque caractéristique x_j à $p_i(x)$ est donnée par :

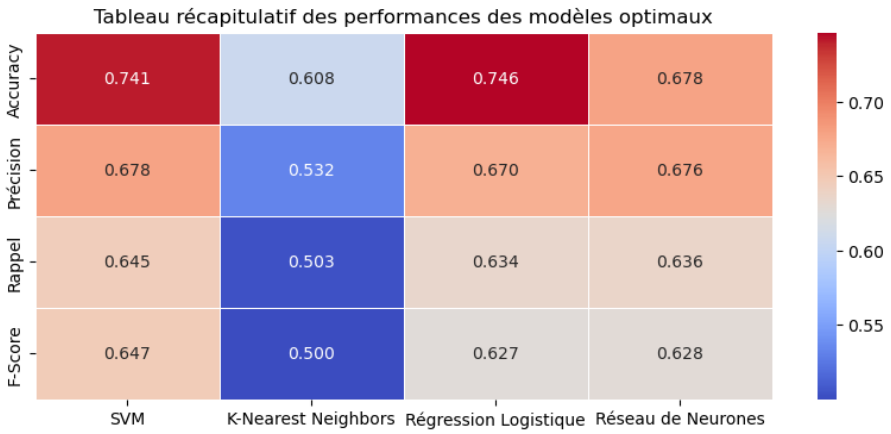
$$\phi_j^i = \sum_{S \subseteq \{1,\ldots,n\} \setminus \{j\}} \frac{|S|!(n - |S| - 1)!}{n!} \left(f_i(S \cup \{x_j\}) - f_i(S)\right)$$

où pour un sous-ensemble T de caractéristiques, et $f_i(T)$ est la probabilité prédite pour la classe C_i avec seulement les caractéristiques du sous-ensemble T.

Cette décomposition permet de comprendre l'impact de chaque caractéristique sur la probabilité de chaque classe, ce qui est crucial pour l'interprétabilité des modèles complexes et formuler des recommandations ciblées.

5 Résultats

5.1 Analyse des performances et sélection du modèle



D'après le tableau récapitulatif des performances, le SVM se distingue avec un F-Score de 0.647, une précision de 0.678 et un rappel de 0.645, offrant ainsi des résultats équilibrés malgré le déséquilibre des classes. La Régression Logistique suit de près avec un F-Score de 0.627, une précision de 0.670 et un rappel de 0.634, bien qu'elle reste légèrement inférieure au SVM. Les Réseaux de Neurones montrent un potentiel intéressant avec un F-Score de 0.628, une précision de 0.676 et un rappel de 0.636, laissant entrevoir des possibilités d'amélioration via un meilleur ajustement des hyperparamètres. En revanche, le k-NN affiche les performances les plus faibles, avec un F-Score de 0.500, une précision de 0.532 et un rappel de 0.503, ce qui souligne ses limites face à des relations complexes dans les données. Par

conséquent, le SVM est recommandé comme modèle principal en raison de ses performances équilibrées, tandis qu'une optimisation des Réseaux de Neurones et un meilleur équilibrage des classes pourraient renforcer les résultats globaux. Nous allons donc analyser l'influence des caractéristiques avec le modèle SVM.

5.2 Importance des caractéristiques avec SHAP

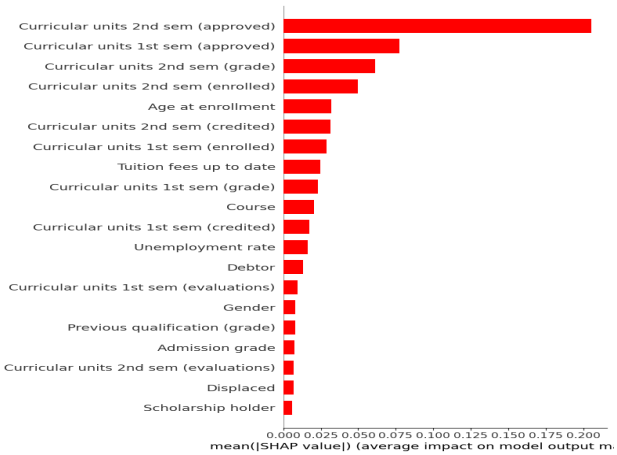


FIGURE 2 – Influence globale des caractéristiques

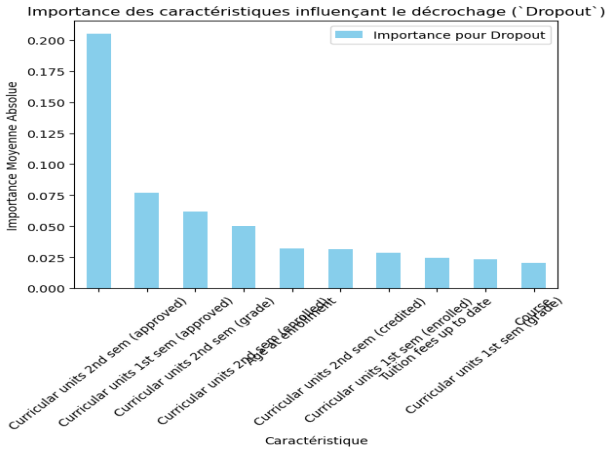


FIGURE 3 – Influence des caractéristiques sur les prédictions Dropout

Les Figures 2 et 3 illustrent respectivement l'importance des caractéristiques pour le modèle global et pour les prédictions spécifiques au décrochage scolaire. Il ressort que les performances académiques influencent fortement les prédictions, notamment pour la prédiction du décrochage. Le tableau ci-dessous présente les cinq caractéristiques les plus influentes, ce qui permet de cibler un ensemble réduit de caractéristiques sur lesquelles il est possible d'intervenir en amont pour atténuer le risque de décrochage. Toutefois, comme nous le verrons dans la section suivante, ces caractéristiques ne capturent pas l'ensemble des profils des étudiants de la classe Dropout, en raison d'une variabilité liée aux spécificités individuelles de chaque élève.

Rang	Caractéristiques	Commentaires
1	Curricular units 2nd sem (approved)	La validation des unités du 2e semestre est le prédicteur global le plus fort, mais son absence augmente significativement les risques de décrochage.
2	Curricular units 1st sem (approved)	La réussite des unités du 1er semestre est un indicateur clé pour la performance globale, et un échec dans ces unités prédit fortement le Dropout.
3	Curricular units 2nd sem (grade)	Des notes faibles au 2e semestre influencent spécifiquement les prédictions de Dropout, soulignant un problème de performance académique.
4	Curricular units 2nd sem (enrolled)	L'absence d'inscription aux unités du 2e semestre est fortement associée au Dropout, reflétant un désengagement académique.
5	Age at enrollment	L'âge à l'inscription influence globalement les prédictions, mais un âge plus avancé augmente la probabilité de décrocher.

TABLE 2 – Top 5 des caractéristiques influentes pour le modèle globale et pour les prédictions Dropout

5.3 Aanalyse locale de deux prédictions Dropout et recommandations

Dans cette section, nous analysons l'influence locale des covariables sur deux observations prédites comme Dropout par notre modèle. Le graphique, interactif force_plot, utilisé permet de visualiser de manière précise comment chaque caractéristique contribue à la probabilité qu'un élève appartienne à la classe Dropout, comme défini par l'Équation 1. Cette analyse locale offre la possibilité de formuler des recommandations personnalisées pour chaque élève. Notez que les valeurs présentées dans les graphiques sont normalisées.

Observation 1

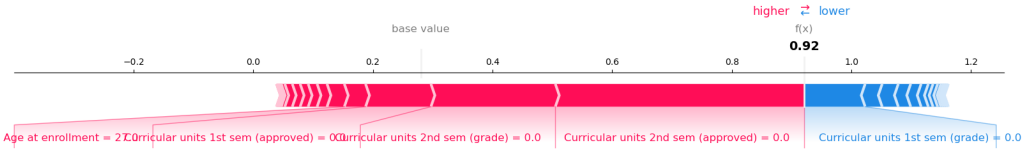


FIGURE 4 – Influence locale des caractéristiques sur l'observation 1

$f(x) = 0.92$ indique une forte probabilité de décrochage pour cet étudiant, car la valeur est significativement supérieure à la base ($\phi_0^{\text{Dropout}} = 1421/4424 \approx 0.32$). Les caractéristiques suivantes contribuent à augmenter ou réduire ce risque, comme précisé dans l'analyse SHAP.

Rang	Caractéristiques	Impacts	Recommandations
1	**Age at enrollment = 27**	**Augmente le risque**	Offrir des programmes adaptés : cours flexibles (en ligne, à temps partiel) pour concilier études et vie personnelle.
2	**Curricular units 1st sem (approved) = 0**	**Augmente le risque**	Proposer un soutien académique : tutorat, séances de rattrapage et stratégies pour réussir les unités du 1er semestre.
3	**Curricular units 2nd sem (grade) = 0.0**	**Augmente le risque**	Identifier les causes des mauvaises performances et proposer des solutions : mentorat, accompagnement académique et planification d'études.
4	**Curricular units 2nd sem (approved) = 0**	**Augmente le risque**	Mettre en place un suivi personnalisé pour éviter les abandons : suivi hebdomadaire et appui sur la charge de travail.
5	**Curricular units 1st sem (grade) = 0.0**	**Réduit le risque**	Renforcer l'engagement académique par des feedbacks positifs et un soutien personnalisé pour maintenir la motivation.

Observation 2

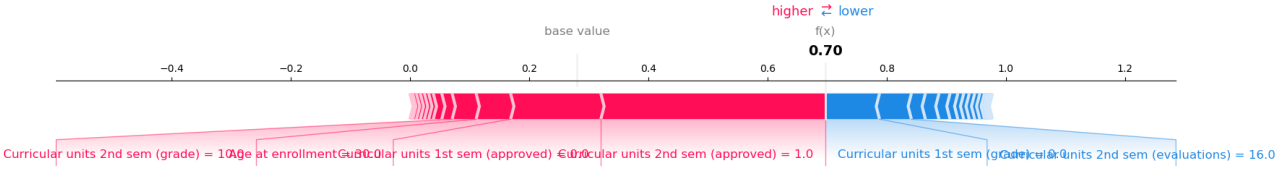


FIGURE 5 – Influence locale des caractéristiques sur l'observation 2

$f(x) = 0.70$ indique une forte probabilité de décrochage pour cet étudiant, car la valeur est significativement supérieure à la base $0,32$. Les caractéristiques suivantes contribuent à augmenter ou réduire ce risque.

Rang	Caractéristiques	Impacts	Recommandations
1	Curricular units 2nd sem (grade) = 10	**Augmente le risque**	Proposer un soutien académique : tutorat, séances de rattrapage pour améliorer les notes et favoriser la réussite.
2	Age at enrollment = 30	**Augmente le risque**	Offrir des programmes adaptés : cours flexibles (en ligne ou temps partiel) pour concilier études et vie personnelle.
3	Curricular units 1st sem (approved) = 0	**Augmente le risque**	Mettre en place un suivi personnalisé pour valider les unités du premier semestre. Identifier les causes des difficultés (manque de préparation, etc.).
4	Curricular units 2nd sem (approved) = 1	**Augmente le risque**	Renforcer l'engagement à travers des plans d'études individualisés et une meilleure gestion de la charge de travail.
5	Curricular units 1st sem (grade) = 14	**Réduit le risque**	Encourager le maintien de bonnes performances académiques à l'aide de feedbacks positifs et du mentorat.
6	Curricular units 2nd sem (evaluations) = 16.0	**Réduit le risque**	Valoriser l'engagement à travers des rappels et des suivis réguliers pour les évaluations. Motiver l'étudiant à poursuivre ses efforts.

Certes, les caractéristiques les plus influentes de la classe Dropout sont présentes dans les deux exemples. Cependant, elles n'influencent pas le risque de décrochage de la même manière. Il est donc fortement recommandé de réaliser une analyse locale de leur influence afin de formuler des recommandations personnalisées, adaptées aux besoins spécifiques de chaque étudiant.

6 Conclusion et Perspectives

En somme, ce projet fournit des bases solides pour la mise en place d'outils prédictifs et d'interventions concrètes destinées à réduire le décrochage scolaire. Il a démontré l'efficacité des algorithmes d'apprentissage automatique pour prédire le **décrochage scolaire** en utilisant des données démographiques, socio-économiques, macroéconomiques et académiques. Le **Support Vector Machine (SVM)** s'est révélé être le modèle le plus performant en raison de sa capacité à gérer les données déséquilibrées, avec un *F-Score* équilibré de **0.647**. L'application de la méthode **SHAP** a permis d'interpréter les prédictions en identifiant les caractéristiques les plus influentes, telles que les performances académiques et l'âge à l'inscription. Ces résultats permettent non seulement de cibler des interventions

spécifiques pour prévenir le décrochage, mais aussi de formuler des recommandations personnalisées pour les étudiants à risque.

Il serait pertinent d'intégrer des données **longitudinales** pour suivre l'évolution des élèves au fil du temps et d'ajouter des variables comportementales telles que l'**engagement en classe** ou la **participation aux activités parascolaires**. L'inclusion de ces variables pourrait améliorer les performances des modèles.

Références

[1] N. Aljohani, T. Ahmad, and M. Hussain. A machine learning based model for student's dropout prediction in distance learning courses. *Journal of Education and e-Learning Research*, 8(3) :245–256, 2021.

[2] J. Arnedo-Moreno, C. Perez-Vidal, and R. Martinez. Exploring statistical approaches for predicting student dropout in higher education. *Higher Education Studies*, 12(1) :41–56, 2023.

[3] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016.

[4] D. et al. Kleinbaum. *Logistic Regression : A Self-Learning Text*. Springer, 2018.

[5] S. Kotsiantis, N. Tselios, A. Filippidi, and V. Komis. Predicting student dropout and academic success using demographic, socio-economic and academic data. *Data in Brief*, 30 :105483, 2020.

[6] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, 2017.

[7] Christoph Molnar. *Interpretable Machine Learning : A Guide for Making Black Box Models Explainable*. Christoph Molnar, 2024. Available online at https://christophm.github.io/interpretable-ml-book/.

[8] Valentim Realinho and et al. Predicting student dropout and academic success. *Data*, 7(11) :146, 2022.

[9] M. T. Ribeiro, S. Singh, and C. Guestrin. Why should i trust you ? explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144, 2016.

[10] F. Silva and M. Neto. Predictive models for imbalanced data : A school dropout perspective. *Education and Information Technologies*, 24 :2765–2780, 2019.

[11] X. et al. Wang. A survey on svm for classification. *Journal of AI Research*, 2019.

[12] L. et al. Yang. k-nn in high-dimensional spaces. *Machine Learning Review*, 2020.