Décrochage scolaire : évaluation du potentiel de l'IA par une démarche guidée par les SHS

Noëllie Martin[1] Chahrazed Labba[1] Armelle Brun[1] Guillaume Cleuziou[2]
(1) Université de Lorraine, LORIA, MosAIk, 54506 Vandœuvre-lès-Nancy, France
(2) Université d'Orléans, INSA-CVL, LIFO, UR4022, F45067 Orléans, France
noellie.martin@univ-lorraine.fr, chahrazed.labba@univ-lorraine.fr,
armelle.brun@univ-lorraine.fr, guillaume.cleuziou@univ-orleans.fr

RÉSUMÉ
La prédiction du décrochage scolaire constitue un enjeu humain fort, auquel les chercheurs en SHS s'intéressent depuis longtemps. L'Intelligence Artificielle, en se généralisant, a été exploitée dans des études pour étudier ce phénomène. Nous nous intéressons à savoir, au travers d'un jeu de données contenant les facteurs identifiés par les SHS comme liés au décrochage scolaire, (1) si les modèles classiques de la littérature en IA permettent de prédire de façon fiable quels élèves seront décrocheurs et (2) s'il est possible d'identifier automatiquement des profils type de décrocheurs. Nous montrons que la complexité du phénomène de décrochage limite la capacité des modèles classiques à traiter ces deux tâches de façon satisfaisante.

ABSTRACT
School dropout prediction is a major social challenge that has long been studied in the social sciences and humanities. With the rise of Artificial Intelligence, researchers have sought to leverage it to better understand this phenomenon. In this study, we investigate whether, using a dataset containing factors identified by social scientists as dropout predictors, (1) standard machine learning models can reliably predict at-risk students, and (2) whether it is possible to automatically identify typical dropout profiles. We show that the inherent complexity of school dropout prevents classical models from addressing either task satisfactorily.

MOTS-CLÉS : Décrochage scolaire, apprentissage automatique, K12.

KEYWORDS: School Dropout, Machine Leaning, K12.

1 Introduction

L'éducation nationale définit le décrochage scolaire comme le fait de sortir de l'école sans diplôme. Chaque année en France, des milliers d'élèves quittent ainsi le système scolaire sans avoir obtenu de qualification. Ce phénomène représente un enjeu à la fois humain, social et économique. En effet, l'absence de qualification a des conséquences durables sur la vie personnelle et professionnelle de ces personnes (risque de chômage, de précarité économique, exclusion sociale, impact sur la santé mentale et le bien-être) [OECD, 2021, De Witte et al., 2013].

Le décrochage scolaire est un phénomène complexe à appréhender. De nature longitudinale : il se construit progressivement au cours de la scolarité [Bernard, 2013] et résulte de l'interaction de plusieurs facteurs [Bruno et al., 2017]. Par ailleurs, il n'existe pas de profil unique de décrocheur :

la littérature établit plusieurs typologies de profils [Bernard, 2011, Bonnéry, 2011, Bruno et al., 2017, Janosz, 2000]. Du fait de cette complexité, de nombreux élèves futurs décrocheurs ne sont pas identifiés par leur institution ou leur entourage, ou le sont trop tardivement. Cela limite la mise en œuvre d'actions de prévention.

Dans ce contexte, l'intelligence artificielle (IA) représente une approche prometteuse pour détecter les élèves à risque qui échappent habituellement à l'attention des acteurs éducatifs. En effet, en exploitant des données relatives à la scolarité des élèves et à leur contexte, l'IA peut être utilisée pour prédire le risque de décrochage scolaire à partir d'une analyse multidimensionnelle et/ou longitudinale. Dans cet article, nous proposons d'étudier dans quelle mesure l'IA, et en particulier l'apprentissage automatique (ML), permet de prédire et de comprendre le décrochage scolaire dans le secondaire. Cette étude repose également sur des connaissances en sciences humaines et sociales (SHS). Pour ce faire, nous utilisons 'High School Longitudinal Study' [1] (HSLS), un jeu de données de référence pour cette tâche et dont la particularité est de combiner des données provenant de multiples sources.

Ainsi, à partir des facteurs de HSLS identifiés par les SHS comme explicatifs du décrochage scolaire, cette étude s'articule autour de deux questions : (1) dans quelle mesure les approches classiques d'apprentissage automatique permettent-elles de prédire le risque de décrochage ? (2) est-il possible de retrouver automatiquement une typologie de décrocheurs telle que proposée en SHS ?

Cet article est organisé comme suit. La section 2 présente l'état de l'art sur le décrochage scolaire en science de l'éducation et en IA. La section 3 décrit la méthodologie appliquée, incluant les données et les modèles utilisés. La section 4 présente et discute les résultats obtenus. Enfin, la section 5 conclut ce travail et ouvre des perspectives de recherche.

2 État de l'art

2.1 Le décrochage scolaire

Le décrochage scolaire est défini de deux manières par la littérature [Bernard, 2013]. D'une part, il est considéré comme un processus longitudinal, évoluant tout au long de la scolarité d'un élève. Cette définition prend en compte le phénomène dans son entièreté, allant des premiers signes, qui peuvent survenir très tôt dans la scolarité, jusqu'à l'acte de décrochage [Galand and Hospel, 2015]. D'autre part, le décrochage est considéré comme un état : une scolarité inachevée en dessous d'un certain niveau de qualification [Janosz, 2000]. Cette définition est la plus communément adoptée, car elle offre un moyen simple de mesurer le phénomène.

Le décrochage scolaire est influencé par un ensemble de facteurs [Galand and Hospel, 2015, Vinciguerra et al., 2021a]. Certains sont immuables, comme les caractéristiques personnelles. D'autres reflètent des situations passagères, comme les résultats scolaires [Bonnéry, 2011]. Les travaux de la littérature regroupent ces facteurs en catégories : individuels, sociaux, familiaux, culturels, socio-économiques et institutionnels [Galand and Hospel, 2015, De Witte et al., 2013, Bonnéry, 2011]. Certaines catégories de facteurs exercent une influence plus importante sur le décrochage en raison de leur proximité temporelle avec le phénomène [Janosz, 2000].

L'interaction entre ces facteurs est complexe [Vinciguerra et al., 2021a], rendant l'analyse du dé-

1. https://nces.ed.gov/surveys/hsls09

crochage scolaire difficile à généraliser. Ainsi, il n'existe pas de décrocheur universel. Des études ont regroupé les décrocheurs en se basant sur des caractéristiques communes [Bruno et al., 2017]. Dans les années 2000, Michel Janosz a proposé quatre types de décrocheurs [Janosz, 2000] : les décrocheurs discrets (peu visibles), les décrocheurs désengagés (n'apprécient pas l'école), les décrocheurs sous-performants (difficultés scolaires) et les décrocheurs inadaptés (expérience scolaire négative). Par cette typologie, les décrocheurs se distinguent entre eux par leur nature, l'intensité des difficultés scolaires ainsi que par leur niveau d'adaptation sociale. Cette typologie s'intéresse principalement à des aspects scolaires du décrochage (note, comportement et engagement). Peu de typologies considèrent des aspects liés à l'environnement plus général de l'élève tels que le contexte familial ou bien local (ex. climat scolaire).

2.2 La prédiction du décrochage scolaire appuyée par l'IA

La littérature s'est intéressée à la prédiction du décrochage scolaire appuyée par des modèles d'IA, en particulier les modèles d'apprentissage automatique (ML) : les Arbres de Décisions (DT), la Régression Logistique (LR) ou encore les Machines à Vecteurs de Support (SVM). Par ailleurs, les modèles d'ensemble tels que les Forêts Aléatoires (RF) sont très largement utilisés dans la littérature : ils limitent le surapprentissage et présentent, en général, de meilleurs résultats [Park and Yoo, 2021].

En SHS, environ 10% d'élèves sont décrocheurs. Cette proportion implique un fort déséquilibre entre les deux classes : les élèves décrocheurs (classe minoritaire) et les non-décrocheurs (classe majoritaire). En ML, ce déséquilibre constitue un problème, car il peut nuire à la qualité de la prédiction. En effet, la classe minoritaire ne contient pas assez d'exemples pour permettre aux modèles de correctement la distinguer de la classe majoritaire. Pour cela, deux méthodes classiques visent à équilibrer le nombre de données dans chacune des classes [Mohammed et al., 2020] : le sur-échantillonnage qui réplique la classe minoritaire et le sous-échantillonnage qui écarte des instances de la classe majoritaire.

Il est important de noter que des métriques non dédiées peuvent biaiser l'évaluation, en favorisant la prédiction de la classe majoritaire [Farou et al., 2023, Ferrer, 2023, Gaudreault et al., 2021]. Il est donc pertinent de privilégier des métriques telles que la précision et le rappel, qui permettent une analyse au niveau de chaque classe (dont la classe "décrocheurs" en particulier).

3 Méthodologie

Cette étude s'inscrit dans une approche expérimentale s'inspirant des travaux de la littérature dans un double objectif : tout d'abord, nous étudions la qualité des modèles de ML classiquement utilisés sur la tâche de prédiction du risque de décrochage scolaire, afin d'en comprendre les limites et d'identifier des perspectives d'améliorations. Ensuite, nous étudions si une typologie de décrocheurs peut être trouvée automatiquement au sein du jeu de donnée HSLS. En cohérence avec la littérature en IA, nous considérons le décrochage scolaire comme un état.

Les jeux de données sur le décrochage scolaire sont peu nombreux. Nous avons retenu pour cette étude le jeu de données 'High School Longitudinal Study' (HSLS) qui est l'un des plus utilisés. Il a été produit par IES NCES (*National Center for Education Statistics*)[2] et provient d'une étude américaine longitudinale de 2009, qui vise à analyser des trajectoires d'élèves de la fin du collège

2. https://nces.ed.gov/datalab/onlinecodebook

jusqu'au supérieur. Les données ont été récoltées à partir d'enquêtes, d'évaluations et de dossiers scolaires dans plus de 900 établissements. Sur les 23 503 élèves que compte HSLS, 8% sont identifiés comme décrocheurs. 9 614 variables sont disponibles du fait du nombre important d'items dans les questionnaires. Afin de sélectionner les variables les plus pertinentes, nous nous sommes appuyés sur plusieurs articles portant sur les facteurs du décrochage scolaire. Les variables liées aux facteurs déterminants du décrochage scolaire, selon les SHS ont été sélectionnées. La Table 1 présente les 19 variables retenues portant sur des aspects individuels (sexe), familiaux (niveau d'éducation des parents) et institutionnels (moyenne générale).

TABLE 1 – Variables sélectionnées qui sont associées au risque de décrochage scolaire selon les SHS

Variable	**Lien avec le décrochage scolaire**	**Références**
X1SEX	Les garçons seraient plus sensibles au décrochage scolaire que les filles.	[Bruno et al., 2017, Galand and Hospel, 2015]
X1MOMEDU	Un faible niveau de diplôme parental augmente le risque de décrochage scolaire.	[Bruno et al., 2017, Galand and Hospel, 2015]
X1DADEDU	Un faible niveau de diplôme parental augmente le risque de décrochage scolaire.	[Bruno et al., 2017, Galand and Hospel, 2015]
X1HHNUMBER	La composition du foyer, combiner à d'autres facteurs, peut influencer le risque de décrochage.	[De Witte et al., 2013]
X1SCHOOLBEL	Un faible sentiment d'appartenance à l'école peut favoriser le décrochage scolaire.	[Galand and Hospel, 2015]
X1SCHOOLENG	Un faible engagement scolaire est associé à un risque accru de décrochage.	[De Witte et al., 2013, Galand and Hospel, 2015]
X1STUEDEXPCT	De faibles ambitions scolaires peuvent être liées à un risque plus élevé de décrochage.	[De Witte et al., 2013]
X1PAREDEXPCT	De faibles attentes parentales peuvent mener au décrochage scolaire.	[De Witte et al., 2013]
X1SCHOOLCLI	Un mauvais climat social entre pairs peut favoriser le décrochage scolaire.	[Galand and Hospel, 2015]
X2BEHAVEIN	Un mauvais comportement scolaire est associé à un risque accru de décrochage scolaire.	[Galand and Hospel, 2015]
X2PROBLEM	Un nombre élevé de sanctions augmente le risque de décrochage scolaire lorsqu'il est associé avec d'autres facteurs.	[Bonnéry, 2011, Bernard, 2013]
X3TGPAWGT	Une faible moyenne, associé à d'autres facteurs, augmente le risque de décrochage scolaire.	[Bruno et al., 2017, De Witte et al., 2013]
X3TGPAMAT	Des performances faibles en mathématiques augmentent le risque de décrochage scolaire.	[Vinciguerra et al., 2021b]
X3TGPAENG	Des difficultés dans la langue pratiquée à l'école augmente le risque de décrochage scolaire.	[Vinciguerra et al., 2021b]
X4DISABLED	La présence d'un handicap peut accroître la vulnérabilité face au décrochage scolaire.	[De Witte et al., 2013]
S2ABSENT	L'absentéisme est fortement associé au décrochage scolaire.	[Galand and Hospel, 2015, Bernard, 2013]
S2SKIPCLASS	L'absence volontaire à certains cours constitue un signe de décrochage scolaire.	[Bernard, 2013]
P1REPEATGRD	Le redoublement est fréquemment associé à un risque accru de décrochage scolaire.	[Galand and Hospel, 2015, De Witte et al., 2013]
A2CYBERBULLY	Le harcèlement scolaire ou numérique peut favoriser le désengagement et le décrochage scolaire.	[De Witte et al., 2013]

Les données ont fait l'objet d'un prétraitement : les valeurs manquantes ont été imputées et les variables ont été normalisées pour réduire leur variance. Le jeu de données a été divisé en un ensemble d'entraînement (80%) et un ensemble de test (20%). Face au déséquilibre des classes, deux stratégies d'échantillonnage ont été considérées : le sur-échantillonnage via SMOTE [Chawla et al., 2002] et le sous-échantillonnage par sélection aléatoire [He and Ma, 2013].

Nous avons sélectionné cinq modèles de classification classiquement utilisés pour la tâche de prédiction du décrochage : la régression logistique (LR) [Berkson, 1944], les arbres de décision (DT)

[Quinlan, 1986], les machines à vecteurs de support (SVM) [Cortes and Vapnik, 1995], ainsi que deux modèles d'ensemble : les forêts aléatoires (RF) [Breiman, 2001] et le gradient boosting (XGBoost) [Friedman, 2001]. Ces modèles se distinguent par leur équilibre entre interprétabilité et performance prédictive. La régression logistique et les arbres de décision produisent des résultats facilement explicables, tandis que les SVM, les forêts aléatoires et XGBoost sont plus complexes mais offrent généralement de meilleures performances. Les hyperparamètres de chacun des modèles ont été optimisés par une recherche en grille (grid search), afin de sélectionner la configuration qui offre les meilleures performances. De plus, une validation croisée à 5 plis a été appliquée sur l'ensemble d'entraînement. Leurs performances ont ensuite été comparées au moyen de métriques adaptées au déséquilibre des classes : le rappel, la précision et le F2-score. Le F2-score est une variante du F1-score accordant plus de poids au rappel, de sorte que les faux négatifs - décrocheurs non détectés - soient davantage pénalisés. Dans le contexte du décrochage scolaire, manquer un élève réellement en difficulté (faux négatif) a des conséquences plus lourdes que de détecter à tort un élève qui ne décroche pas (faux positif).

Concernant la typologie des décrocheurs, nous avons sélectionné des modèles de clustering reposant sur des principes différents : K-means [Lloyd, 1982], Modèle de Mélanges Gaussiens (GMM) [Dempster et al., 1977], clustering hiérarchique agglomératif [Ward Jr, 1963] et HDBSCAN [Campello et al., 2013].

Nous avons sélectionné 9 variables scolaires (moyenne générale, moyenne en mathématiques, moyenne en anglais, climat scolaire, échelle de problème au lycée), sociales (comportement, engagement, sentiment d'appartenance), et familiales (nombre de personnes dans le foyer). Elles sont toutes numériques, car les modèles de clustering sélectionnés nécessitent des données quantitatives.

4 Résultats et discussion

Cette section poursuit deux objectifs sur les données HSLS : évaluer la capacité des modèles de classification à prédire le décrochage scolaire, d'une part, et vérifier si les modèles de clustering permettent de retrouver une typologie établie de décrocheurs, d'autre part.

4.1 Prédiction du décrochage scolaire

La Table 2 présente la moyenne des scores obtenus lors de la validation croisée pour la précision, rappel et F2-score pour la classe "décrocheur". Les performances sur la classe "non-décrocheur" ne sont pas reportées : elles sont uniformément élevées quel que soit le modèle et n'ont pas d'intérêt au regard de notre objectif. Un modèle de prédiction performant se caractérise par un F2-score, une précision et un rappel élevés.

Sans échantillonnage, SVM présente une précision élevée (0.73) mais un rappel très faible (0.06). L'échantillonnage inverse cette tendance et améliore sensiblement le F2-score (0.52), ce qui traduit une forte sensibilité du modèle au déséquilibre des classes. LR est plus robuste : un rappel de 0.72–0.73 et un F2-score de 0.52–0.53, ses performances restent stables quel que soit l'échantillonnage, ce qui en fait le meilleur des deux modèles sur cette tâche. DT obtient des résultats comparables à LR sans échantillonnage (rappel : 0.68, F2-score : 0.50), mais l'échantillonnage dégrade ses performances : le rappel et le F2-score diminuent et la précision augmente (0.45 et 0.32 respectivement).

TABLE 2 – Performance des différents modèles en fonction de la méthode d'échantillonnage

Modèle	Sans échantillonnage			Sur-échantillonnage			Sous-échantillonnage		
	Précision	Rappel	F2-score	Précision	Rappel	F2-score	Précision	Rappel	F2-score
SVM	**0.73**	0.06	0.07	0.25	0.68	0.51	0.26	0.70	0.52
LR	0.25	**0.72**	**0.52**	0.25	**0.73**	**0.53**	0.25	**0.73**	**0.53**
DT	0.25	0.68	0.50	**0.45**	0.34	0.36	**0.32**	0.55	0.48
RF	0.33	0.59	0.50	0.41	0.43	0.41	0.27	0.67	0.52
XGB	0.59	0.20	0.23	0.41	0.43	0.42	0.27	0.72	**0.53**

Parmi les modèles d'ensemble, RF sans échantillonnage présente un rappel et F2-score inférieurs à ceux de LR. Le sous-échantillonnage améliore ces deux métriques, sans toutefois atteindre les performances de LR. XGBoost, dont le rappel est très faible sans échantillonnage (0.20), bénéficie davantage du sous-échantillonnage : son rappel atteint 0.72 au détriment de la précision (0.27), et son F2-score devient équivalent à celui de LR. Le sous-échantillonnage a tendance à augmenter le rappel des modèles lorsque celui-ci est initialement faible. Cependant, la précision reste faible pour l'ensemble des modèles. À l'opposé, le sur-échantillonnage a tendance à impacter négativement le rappel, mais augmenter la précision, principalement pour les modèles à base d'arbres.

LR est le modèle le plus performant, au regard du rappel et du F2-score, il surpasse notamment les modèles d'ensemble. Cependant, de notre point de vue les valeurs de rappel et de F2-score ne sont pas suffisamment élevées. En effet, la plus grande valeur de rappel est de 0.73, ce qui signifie qu'un décrocheur sur quatre n'est pas identifié par les modèles classiques. Par ailleurs, malgré un rappel élevé, la précision demeure très faible (0.25) : la plupart des élèves identifiés comme à risque ne le sont pas réellement. En conclusion, aucun des modèles testés ne permet d'identifier de façon fiable les décrocheurs, ce que nous pouvons expliquer doublement : tout d'abord, les données utilisées, bien que représentant les facteurs identifiés par la littérature comme étant déterminants du décrochage scolaire, ne sont possiblement pas assez riches. En ce sens, la quantité d'informations véhiculée par les facteurs choisis n'est pas suffisante. De futures expérimentations intégreront d'autres facteurs pour confirmer cette hypothèse. Ensuite, il est possible que les modèles sélectionnés ne soient pas adaptés à la complexité (notamment le caractère longitudinal) du phénomène de décrochage. Par conséquent, dans des travaux futurs, nous proposerons de nouveaux modèles dédiés.

4.2 Identification de profils de décrocheurs par clustering automatique

La création d'une typologie de décrocheurs correspond à une tâche de clustering. Nous avons choisi plusieurs algorithmes de clustering de la littérature, qui diffèrent par leur façon de procéder au regroupement (partitionnement, probabiliste, hiérarchique, densité). Les modèles sélectionnés ont mené à des résultats similaires en termes de score Silhouette (à maximiser) et de composition de clusters. Nous avons étudié Silhouette entre 2 et 9 clusters et le nombre de clusters optimal a été observé entre 2 et 4 clusters pour tous les modèles. Par conséquent, nous ne présentons pas le détail de leurs performances et faisons le choix de présenter uniquement les résultats obtenus par K-means.

Dans un premier temps, nous avons cherché à établir le nombre de clusters optimal. La Figure 1 présente l'évolution de la valeur de Silhouette en fonction du nombre de clusters. K-means obtient un score de silhouette de 0.18 pour 2 clusters (Figure 1). Au-delà de 2 clusters la courbe ne fait que décroître. Cela suggère que K=2 clusters semble être le nombre optimal. Cependant, 0.18 est un score

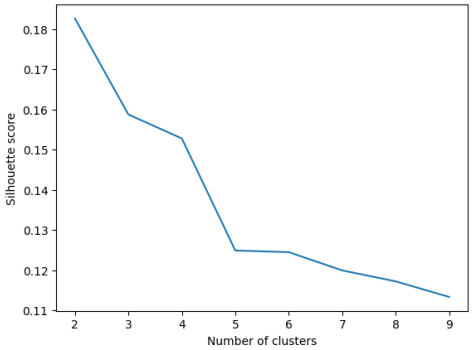

FIGURE 1 – Évolution du score de Silhouette en fonction du nombre de clusters (K-Means)

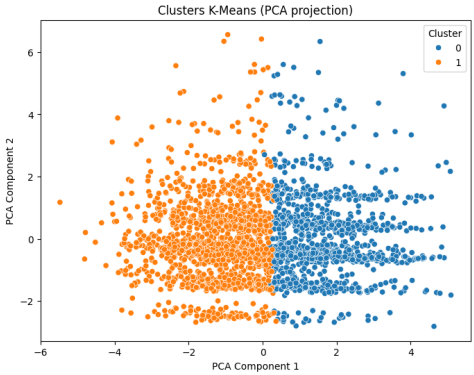

FIGURE 2 – Projection des clusters obtenus par K-Means sur 2 composantes principales

très faible qui suggère que les clusters sont mal délimités, voire qu'ils se chevauchent. Cela semble confirmer l'hypothèse liée à la complexité du phénomène et à la limitation des données utilisées.

Dans un second temps, nous avons visualisé les clusters afin de mieux comprendre les valeurs de Silhouette. Pour cela, nous avons procédé à une ACP avec deux composantes [Maćkiewicz and Ratajczak, 1993], présentée en Figure 2. L'absence visuelle de séparation entre les clusters corrobore le faible score de Silhouette obtenu.

Ici encore, ces résultats suggèrent que les données utilisées, et en particulier les facteurs exploités, ne permettent pas de capturer les spécificités (typologie) des décrocheurs.

5 Conclusion

L'objectif de cet article était de faire le lien entre les connaissances en SHS et les méthodes en IA pour étudier le phénomène de décrochage scolaire. Nous nous sommes appuyés sur un jeu de données pour prédire le décrochage scolaire et analyser la typologie des décrocheurs. Les résultats concernant la prédiction du décrochage scolaire montrent que les méthodes utilisées dans la littérature ne sont pas performantes. Ainsi, il serait intéressant de concevoir des approches différentes afin de mieux capturer ce phénomène. Il serait en particulier envisageable d'adopter une approche longitudinale du décrochage. Cependant, cette approche complexifie la labellisation des données. Pour cela, des modèles semi ou non supervisés seraient pertinents. En particulier, nous envisageons d'intégrer des connaissances en SHS sur le décrochage scolaire pour guider l'utilisation de l'IA. En ce qui concerne l'analyse de la typologie des décrocheurs aucun résultat ne révèle une structuration des données en différents types de décrocheurs. Ces résultats pourraient s'expliquer par une insuffisance informationnelle du jeu de données pour cette tâche ou par un manque d'adéquation des modèles de clustering utilisés. En effet, une approche longitudinale pourrait, ici encore, être plus pertinente. Elle permettrait d'observer des changements de comportements dans les trajectoires scolaires des élèves.

Références

Joseph Berkson. Application of the logistic function to bio-assay. *Journal of the American statistical association*, 39(227) :357–365, 1944.

Pierre-Yves Bernard. Le décrochage des élèves du second degré : Diversité des parcours, pluralité des expériences scolaires. *Les Sciences de l'éducation-Pour l'Ère nouvelle*, 44(4) :75–97, 2011.

Pierre-Yves Bernard. Le décrochage scolaire. *Presses Universitaires de France*, May 2013. ISSN 0768-0066. DOI : 10.3917/puf.berna.2013.01.

Stéphane Bonnéry. Blaya catherine. décrochages scolaires. l'école en difficulté. *Revue française de pédagogie*, 2011.

Leo Breiman. Random forests. *Machine learning*, 45(1) :5–32, 2001.

Françoise Bruno, Christine Félix, and Frédéric Saujat. L'évolution des approches du décrochage scolaire. *Carrefours de l'éducation*, 43(1) :246–271, July 2017. ISSN 1262-3490. DOI : 10.3917/cdle.043.0246.

Ricardo JGB Campello, Davoud Moulavi, and Jörg Sander. Density-based clustering based on hierarchical density estimates. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 160–172. Springer, 2013.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote : synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16 :321–357, 2002.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3) :273–297, 1995.

Kristof De Witte, Sofie Cabus, Geert Thyssen, Wim Groot, and Henriëtte Maassen van den Brink. A critical review of the literature on school dropout. *Educational Research Review*, 10 :13–28, December 2013. ISSN 1747-938X. DOI : 10.1016/j.edurev.2013.05.002.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society : series B (methodological)*, 39(1) : 1–22, 1977.

Zakarya Farou, Mohamed Aharrat, and Tomáš Horváth. A Comparative Study of Assessment Metrics for Imbalanced Learning. In Alberto Abelló, Panos Vassiliadis, Oscar Romero, Robert Wrembel, Francesca Bugiotti, Johann Gamper, Genoveva Vargas Solar, and Ester Zumpano, editors, *New Trends in Database and Information Systems*, pages 119–129, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-42941-5. DOI : 10.1007/978-3-031-42941-5_11.

Luciana Ferrer. Analysis and Comparison of Classification Metrics, September 2023.

Jerome H Friedman. Greedy function approximation : a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

Benoît Galand and Virginie Hospel. Facteurs associés au risque de décrochage scolaire : vers une approche intégrative. *L'orientation scolaire et professionnelle*, (44/3), September 2015. ISSN 0249-6739. DOI : 10.4000/osp.4604.

Jean-Gabriel Gaudreault, Paula Branco, and João Gama. An Analysis of Performance Metrics for Imbalanced Classification. In Carlos Soares and Luis Torgo, editors, *Discovery Science*, pages 67–77, Cham, 2021. Springer International Publishing. ISBN 978-3-030-88942-5. DOI : 10.1007/978-3-030-88942-5_6.

Haibo He and Yunqian Ma. Imbalanced learning : foundations, algorithms, and applications. 2013.

Michel Janosz. L'abandon scolaire chez les adolescents : perspective nord-américaine. *Diversité*, 122(1) :105–127, 2000. DOI : 10.3406/diver.2000.1141.

Stuart Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2) : 129–137, 1982.

Andrzej Maćkiewicz and Waldemar Ratajczak. Principal components analysis (pca). *Computers & Geosciences*, 19(3) :303–342, 1993.

Roweida Mohammed, Jumanah Rawashdeh, and Malak Abdullah. Machine learning with oversampling and undersampling techniques : overview study and experimental results. In *2020 11th international conference on information and communication systems (ICICS)*, pages 243–248. IEEE, 2020.

OECD. OECD Digital Education Outlook 2021. https ://www.oecd.org/en/publications/oecd-digital-education-outlook-2021_589b283f-en.html, June 2021.

Hee Sun Park and Seong Joon Yoo. Early Dropout Prediction in Online Learning of University using Machine Learning. *JOIV : International Journal on Informatics Visualization*, 5(4) :347–353, December 2021. ISSN 2549-9904. DOI : 10.30630/joiv.5.4.732.

J. Ross Quinlan. Induction of decision trees. *Machine learning*, 1(1) :81–106, 1986.

A. Vinciguerra, I. Nanty, C. Guillaumin, E. Rusch, L. Cornu, and R. Courtois. Les déterminants du décrochage dans l'enseignement secondaire : Une revue de littérature. *Psychologie Française*, 66 (1) :15–40, March 2021a. ISSN 0033-2984. DOI : 10.1016/j.psfr.2019.09.003.

Antony Vinciguerra, Isabelle Nanty, Catherine Guillaumin, Emmanuel Rusch, Laurence Cornu, and Robert Courtois. Les déterminants du décrochage dans l'enseignement secondaire : une revue de littérature. *Psychologie française*, 66(1) :15–40, 2021b.

Joe H Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301) :236–244, 1963.