Prédiction de profils étudiants sur une plateforme d'apprentissage en ligne

Pauline Chiquet[1], François Lecellier[1], Philippe Carré[1]

[1] Université de Poitiers, Univ. Limoges, CNRS, XLIM, Poitiers, France

{pauline.chiquet, francois.lecellier, philippe.carre}@univ-poitiers.fr

Résumé

Depuis une vingtaine d'années, les universités utilisent des outils numériques qui génèrent des traces numériques pour améliorer l'accessibilité de leurs cours. L'étude présentée dans cet article a pour objectif de caractériser des profils étudiants de manière automatique à partir de ces données. Cependant, en général, les études dans ce cadre sont confrontées à des classes déséquilibrées. Nous proposons d'analyser l'influence du sur-échantillonnage d'une base de données issue d'une plateforme d'apprentissage en ligne du supérieur de Poitiers. Nos résultats montrent que le sur-échantillonnage permet d'améliorer la précision et le rappel des modèles de prédiction et, par conséquent, de mieux détecter notamment les situations d'abandon.

Mots-clés

Analyse de l'apprentissage, apprentissage supervisé, Moodle, sur-échantillonnage.

Abstract

For the past twenty years, universities have been using digital tools that generate digital traces to improve the accessibility of their courses. The study presented in this article aims to characterize learner profiles automatically from this data. However, in general, studies in this context are confronted with unbalanced classes. We propose to analyze the influence of oversampling on a database from an e-learning platform at Poitiers University. Our results show that oversampling improves the precision and recall of prediction models, and consequently enables better detection of dropout situations in particular.

Keywords

Learning analytics, Moodle, Oversampling, Supervised learning.

1 Introduction

Les Learning Analytics, également appelés analyse de l'apprentissage, consistent à utiliser et analyser des données liées à l'apprentissage et à l'éducation. Les objectifs sont multiples : le suivi de l'acquisition des connaissances, la prédiction de profils ou de résultats, ou encore, la personnalisation de l'enseignement. Les données utilisées sont les traces numériques, les évaluations, les informations démographiques, mais également l'historique académique [1]. Dans cette étude, nous utilisons Motive. Cette plateforme d'autoformation, dédiée aux compétences transversales est construite sous Moodle par l'Université de Poitiers dans le cadre du projet Elans [1]. Elle est composée de 11 chapitres avec 26 tests et 74 modules au total. La figure 1 présente l'organisation de la plateforme Motive. Les chapitres portent sur différentes thématiques telles que la prise de note, la recherche documentaire, la gestion de projet, les compétences numériques, etc.

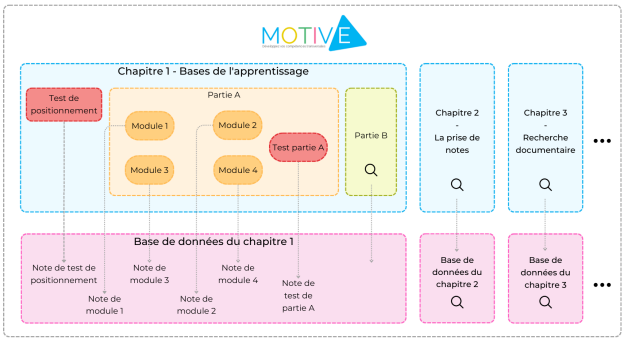

FIGURE 1 – Organisation de la plateforme Motive (zoom sur le chapitre 1). Chaque note obtenue est stockée dans une base de données associée au chapitre.

Chaque chapitre débute par un test de positionnement. Ce test permet de connaître le niveau de l'étudiant. Puis, l'étudiant réalise différents modules à l'issue desquels il obtient des notes de module. Les modules peuvent être réalisés dans n'importe quel ordre. Après avoir effectué tous les modules, l'étudiant réalise un test de partie. Une note de test de partie est alors obtenue. En d'autres termes, les modules correspondent à des exercices d'entraînement portant chacun sur une notion particulière. Les tests de partie correspondent à des évaluations reprenant l'ensemble des notions présentées dans une partie.
Pour chaque chapitre, trois types de fichier sont disponibles : les fichiers de **modules**, de **tests** et de **logs**. Ces derniers contiennent les traces numériques générées par les étudiants sur la plateforme. À partir de ces données, chaque

1. L'Université de Poitiers bénéficie d'une aide de l'État gérée par l'Agence Nationale de la Recherche au titre du programme des Nouveaux Cursus Universitaire (NCU ELANS - réf. ANR-18-NCUN-0026).

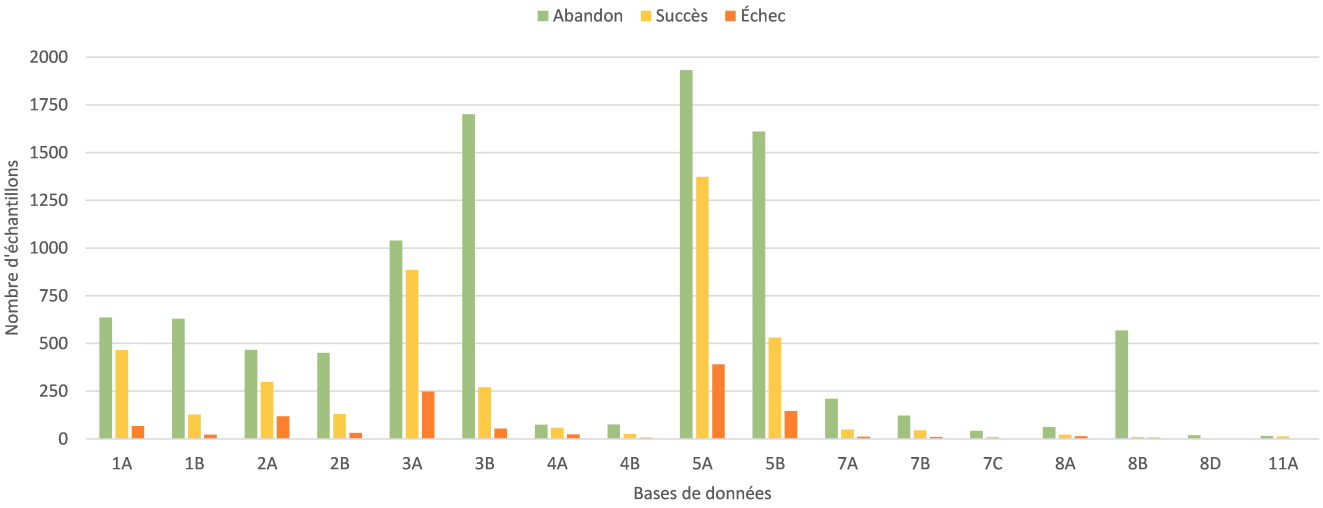

FIGURE 2 – Répartition des profils étudiants au sein des chapitres de Motive

utilisateur a pu être étiqueté, de manière automatique, selon trois catégories : Abandon, Succès ou Échec.
Cet étiquetage permet de connaître la répartition des profils étudiants au sein de Motive (figure 2). Ces données sont très déséquilibrées : la classe Abandon est beaucoup plus représentée que les classes Succès et Échec. Afin de pouvoir travailler sur un ensemble de données statistiquement satisfaisant, nous avons décidé de nous concentrer sur les données issues des parties 2A, 3A, 5A et 5B qui contiennent le plus de données. Il est important de noter que la répartition des classes finales est également fortement déséquilibrée (figure 2).

L'objectif de ces travaux est de prédire les profils étudiants à l'aide d'un modèle de prédiction. Ce modèle permettra de détecter les étudiants en difficulté avant la fin d'un chapitre. Pour cela, nous devons lutter contre le déséquilibre des données. Premièrement, nous présenterons plusieurs méthodes d'analyse de l'apprentissage, mais également des méthodes de sur-échantillonnage des données. Puis, nous expliquerons la démarche de mise en place des modèles de prédiction. Enfin, nous détaillerons l'ensemble des résultats obtenus.

2 État de l'art

2.1 Méthodes de l'analyse de l'apprentissage

L'analyse de l'apprentissage repose sur plusieurs méthodes. Les modèles statistiques sont les plus utilisés (45% des publications selon l'étude de NAMOUN et ALSHANQITI [15]). FOUNG et CHEN [5] se basent sur un modèle de régression pour comprendre comment les étudiants utilisent une plateforme d'apprentissage en ligne, mais également pour prédire la notion de succès en fonction des traces numériques. Afin de pallier la difficulté de prédiction, les auteurs suggèrent de combiner les données liées aux interactions avec la plateforme d'apprentissage à des données externes (par exemple, des informations géographiques ou des antécédents scolaires). De leur côté, MING et MING [13] utilisent l'analyse sémantique latente probabiliste (PLSA) et l'allocation de Dirichlet latente (LDA) pour prédire les notes finales des étudiants à partir des forums de discussion en ligne. Les résultats obtenus par cette combinaison d'approche sont prometteurs.
Les méthodes de Machine Learning sont également très utilisées pour comprendre les profils d'apprentissage des étudiants et les prédire. MORENO-MARCOS et al. [14] ont analysé plusieurs facteurs afin de connaître leur influence sur la prédiction de la performance d'un étudiant. Le jeu de données contient les interactions avec les exercices et les forums, la liste des vidéos qui ont été ouvertes et les suivis des clics. Deux prédictions sont possibles : succès ou échec. Pour analyser l'influence des facteurs, les auteurs utilisent quatre types d'algorithmes standards d'apprentissage supervisé. Les résultats montrent que les données liées aux exercices sont de très bons indices de prédiction. À contrario, les données liées aux forums ou aux clics n'ont généralement que peu d'impacts.
Certains auteurs se sont intéressés au problème de l'apprentissage non supervisé. KUZILEK et al. [11] ont vérifié la corrélation entre les interactions des étudiants avec une plateforme d'apprentissage en ligne et les notes obtenues. Le jeu de données utilisé est OULAD [10]. Les chercheurs utilisent l'algorithme d'espérance-maximisation pour regrouper les données d'interactions des étudiants selon six classes. Les résultats montrent que certains de ces groupes présentent des performances élevées, tandis que d'autres montrent des signes de difficultés dès le début du cours.
FRANCIS et BABU [6] décrivent un modèle de prédiction des résultats des étudiants à partir de données issues de l'enseignement supérieur de l'État Kerala en Inde. Ces données sont organisées selon quatre types de caractéristiques telles que des caractéristiques démographiques, académiques, liées aux interactions avec la plateforme et supplémentaires. Trois prédictions sont possibles : bon résultats, résultats moyens et résultats faibles. La fouille de

données est utilisée via quatre algorithmes de classification standards afin de déterminer les caractéristiques ayant le plus d'impact sur les résultats. Puis, ces caractéristiques sont utilisées comme entrée pour l'algorithme de clustering (K-moyennes [12]). Les résultats montrent une forte corrélation entre les interactions de l'étudiant avec la plateforme d'apprentissage en ligne et ses résultats académiques. Ce type de modèle permet d'obtenir une précision de 0,75.
Suivant le même principe, certains chercheurs combinent l'apprentissage supervisé et l'apprentissage non supervisé. C'est le cas de IATRELLIS et al. [9] qui présentent une méthode permettant de prédire les résultats des étudiants de licence afin de savoir s'ils pourront poursuivre, ou non, leurs études. Les données incluent la moyenne des notes finales, le parcours suivi, des notes de projet, le nombre de redoublement, le rang au sein de la promotion, etc. La première étape consiste à regrouper les étudiants ayant des données similaires à l'aide d'un algorithme de K-moyennes [12] (partie non supervisée). Finalement, trois groupes ont été retenus ($k = 3$). Puis, une forêt d'arbres décisionnels [3] (partie supervisée) est utilisée pour prédire si les étudiants de licence pourront poursuivre leurs études en master ou non. Les résultats montrent qu'un modèle de prédiction précédé d'un clustering obtient de meilleures performances qu'un modèle de prédiction seul.

2.2 Méthodes de sur-échantillonnage

L'ensemble des méthodes d'analyse de l'apprentissage nécessitent une quantité importante de données pour éviter tout risque de sous-apprentissage des modèles. Les données doivent également être suffisamment équilibrées afin de ne pas introduire de biais lors de l'apprentissage. WONGVORACHAN, HE et BULUT [18] présentent trois méthodes pour corriger ce type de déséquilibre : le sur-échantillonnage, le sous-échantillonnage et l'échantillonnage hybride. Le sur-échantillonnage consiste à augmenter le nombre d'échantillons de la ou des classes minoritaires. À l'inverse, le sous-échantillonnage consiste à diminuer le nombre d'échantillons de la ou des classes majoritaires. Enfin, l'échantillonnage hybride consiste à diminuer le nombre d'échantillons de la ou des classes majoritaires et à augmenter le nombre d'échantillons de la ou des classes minoritaires. Du fait du déséquilibre important de nos données, nous avons décidé de nous intéresser plus particulièrement au sur-échantillonnage. La méthode la plus simple est le sur-échantillonnage aléatoire (ou ROS pour Random Oversampling). Celle-ci consiste à dupliquer aléatoirement des échantillons de la classe minoritaire avec remplacement jusqu'à ce que la proportion des deux classes soit équilibrée [18]. La figure 3 illustre cette méthode. Cette méthode est très simple mais occasionne un risque de sur-apprentissage [4].
Il est également possible de créer des échantillons synthétiques à partir des échantillons originaux. Selon CHAWLA et al. [4], le sur-échantillonnage synthétique de la classe minoritaire (ou SMOTE pour Synthetic Minority Oversampling Technique) consiste à augmenter le nombre d'échantillons de la classe minoritaire par la création d'échantillons synthétiques. La classe minoritaire est sur-échantillonnée selon l'algorithme simplifié suivant :

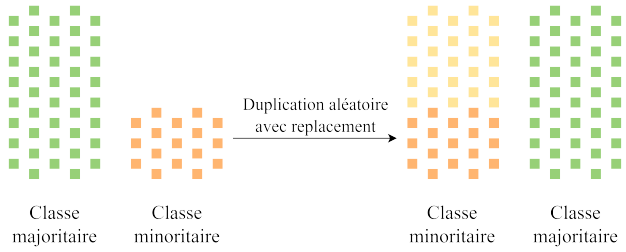

FIGURE 3 – Méthode du sur-échantillonnage aléatoire. En vert, la classe majoritaire. En orange, la classe minoritaire.

1. Sélection aléatoire d'un échantillon de la classe minoritaire ;
2. Identification des k plus proches voisins parmi la classe minoritaire ;
3. Création d'échantillons entre l'échantillon sélectionné et ses k plus proches voisins (par interpolation sur les caractéristiques des échantillons).

Ces étapes sont répétées jusqu'à ce que le nombre d'échantillons de la classe minoritaire soit équivalent à celui de la classe majoritaire. La figure 4 schématise cette méthode.

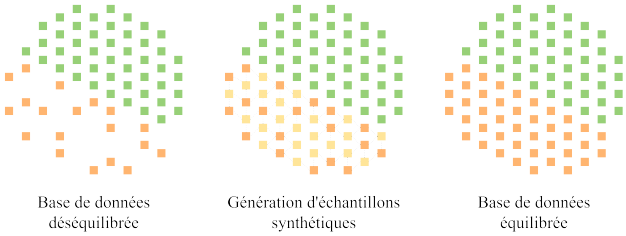

FIGURE 4 – Méthode SMOTE. En vert, la classe majoritaire. En orange, la classe minoritaire. En jaune, les échantillons synthétiques appartenant à la classe minoritaire.

Ces deux méthodes de sur-échantillonnage ont été utilisés par les auteurs HASSAN, AHMAD et ANUAR [8]. L'objectif était de prédire des profils étudiants à partir de données démographiques, académiques et de journaux d'activités. Les profils ont été établis en se basant sur une moyenne pondérée des notes obtenues, classées ensuite en trois catégories : faible, moyenne et excellente. Les données ont été rééquilibrées à l'aide de sur-échantillonnage, de sous-échantillonnage et d'échantillonnage hybride. Les modèles (forêts d'arbres décisionnels) issus du sur-échantillonnage obtiennent un F-score de 0,870 pour le ROS et 0,750 pour le SMOTE. Le meilleur F-score est obtenu avec l'algorithme AdaBoost et le sur-échantillonnage ROS. Les autres algorithmes obtiennent des F-scores inférieurs.
Enfin, nous pouvons également citer deux méthodes de sur-échantillonnage issues du SMOTE. Selon HAN, WANG et MAO [7], la méthode de Borderline-SMOTE consiste à sur-échantillonner uniquement les échantillons minoritaires li-

mites. En d'autres termes, cette méthode génère des échantillons synthétiques de la classe minoritaire uniquement près de la frontière avec la classe majoritaire. La figure 5 illustre cette méthode.

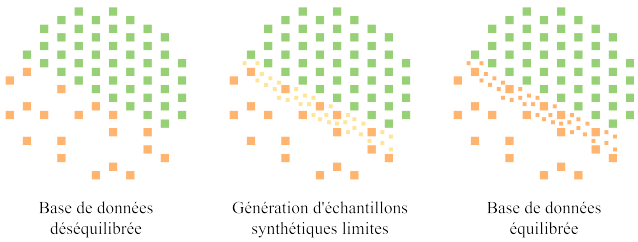

FIGURE 5 – Méthode Borderline-SMOTE. En vert, la classe majoritaire. En orange, la classe minoritaire. En jaune, les échantillons synthétiques limites appartenant à la classe minoritaire.

Il existe également le SVM SMOTE. Il s'agit de la combinaison de la méthode SMOTE et de l'algorithme des machines à vecteurs de support (SVM). Selon NGUYEN, COOPER et KAMEI [16], cette méthode consiste à déterminer les limites des classes à l'aide d'un SVM, puis à générer des échantillons synthétiques de la classe minoritaire. Ces deux dernières méthodes sont intéressantes dans les cas où les échantillons à classifier sont relativement ressemblant. Elles permettent de renforcer les caractéristiques des limites des classes.

3 Modèles de prédiction

3.1 Attributs des nouvelles bases de données

Les données issues de Motive ne sont pas directement utilisables par des modèles de prédiction. Ainsi, des nouvelles bases de données ont été construites à partir des données issues de Motive. Chaque chapitre possède les attributs suivants :

— **Identifiant de l'utilisateur** : `userid` commun aux trois types de fichiers ;

— **Modules** : nombre de tentatives de chaque module pour chaque utilisateur ;

— **Temps moyens des modules** :
- Temps moyen module x : temps moyen pour un module ;
- Temps moyen total : temps moyen pour l'ensemble des modules réalisés (si un ou plusieurs modules ne sont pas réalisés, ils ne sont pas pris en compte dans le calcul du temps moyen total).

— **Notes moyennes des modules** :
- Note moyenne module x : note moyenne pour un module ;
- Note moyenne totale : note moyenne pour l'ensemble des modules réalisés (si un ou plusieurs modules ne sont pas réalisés, ils ne sont pas pris en compte dans le calcul de la note moyenne totale).

— **Test de positionnement** : indique si l'étudiant a réalisé le test de positionnement ou non ;

— **Note du test de positionnement** : note obtenue par l'étudiant au test de positionnement ;

— **Logs** : agrégation des données de logs pour les 101 événements ;

— **Classe de l'échantillon** : Trois classes sont possibles :
- **Abandon** : l'étudiant a commencé le module mais n'a pas réalisé ou n'a pas terminé le test de partie ;
- **Succès** : l'étudiant a commencé le module et a terminé le test de partie avec une note supérieure ou égale à 80% de la note maximale ;
- **Échec** : l'étudiant a commencé le module et a terminé le test de partie avec une note inférieure à 80% de la note maximale.

3.2 Description des pipelines

L'objectif est de prédire le profil des étudiants malgré une base de données déséquilibrée. Trois pipelines seront testés pour déterminer l'impact du sur-échantillonnage sur ces données.

— **Pipeline 1** : Pas de sur-échantillonnage (figure 6).

— **Pipeline 2** : Sur-échantillonnage avant la séparation des données (figure 7).

— **Pipeline 3** : Sur-échantillonnage après la séparation des données (figure 8).

Premièrement, les données sont normalisées à l'aide de la formule suivante :

$$z = \frac{x - \mu}{\sigma}, \tag{1}$$

où z est la valeur centrée réduite, x est la valeur à normaliser, μ est la moyenne de toutes les valeurs à normaliser et σ est l'écart-type de toutes les valeurs à normaliser.
Après cette phase de normalisation, nous séparons les données selon une pondération 70/30 en prenant soin de maintenir la répartition des classes dans chacun des deux échantillons.
Puis, les données sont sur-échantillonnées avant (pipeline 2) ou après (pipeline 3) la séparation des données. Deux méthodes de sur-échantillonnage sont testées : le sur-échantillonnage aléatoire (ROS) et le sur-échantillonnage synthétique de la classe minoritaire (SMOTE). L'état de l'art présente ces deux méthodes de sur-échantillonnage dans le cas de deux classes déséquilibrées (une classe majoritaire et une classe minoritaire). Pour trois classes déséquilibrées, le fonctionnement est le même, sauf que nous sommes en présence de deux classes minoritaires et d'une seule classe majoritaire. Les méthodes de sur-échantillonnage sont appliquées à toutes les classes minoritaires.
Afin de classifier nos données, nous utilisons une forêt d'arbres décisionnels qui est un algorithme proposant les meilleurs résultats. La forêt d'arbres décisionnels offre une meilleure explicabilité que les réseaux de neurones artificiels, facilitant ainsi la compréhension du modèle prédictif.

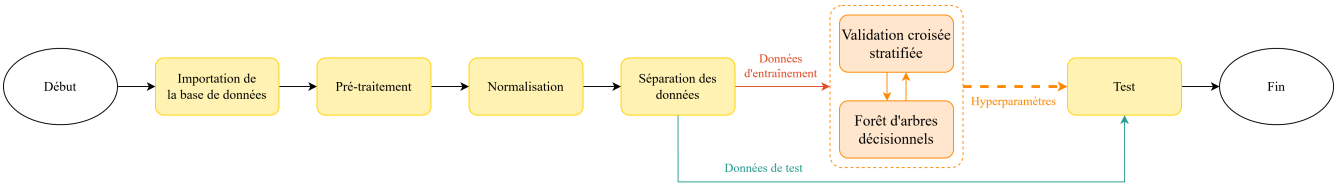

FIGURE 6 – Pipeline 1 (pas de sur-échantillonnage). Le cadre en pointillés orange correspond à l'optimisation des hyperparamètres.

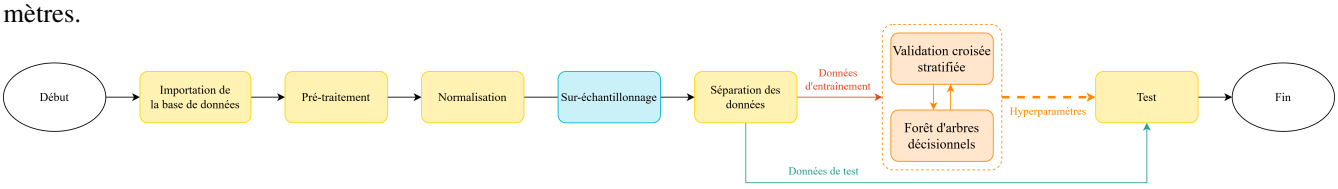

FIGURE 7 – Pipeline 2 (sur-échantillonnage avant la séparation des données). Le cadre en pointillés orange correspond à l'optimisation des hyperparamètres.

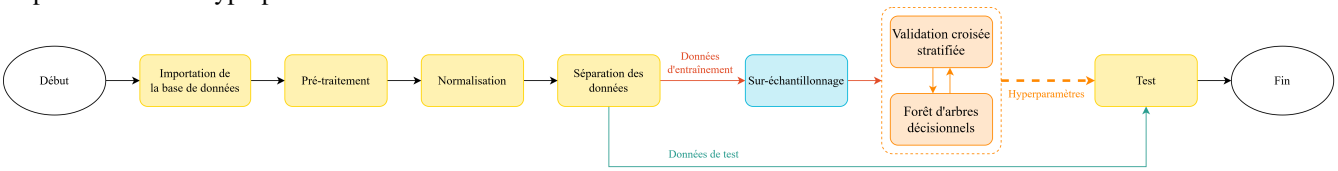

FIGURE 8 – Pipeline 3 (sur-échantillonnage après la séparation des données). Le cadre en pointillés orange correspond à l'optimisation des hyperparamètres.

Différents hyperparamètres sont testés pour déterminer la combinaison optimale :

- **Profondeur maximale d'un arbre** : 2, 3, 4, 5, 6, 7, 8 ou 9.
- **Nombre d'arbres** : 10, 25, 50 ou 100.
- **Critère** : Mesure d'impureté de Gini (BREIMAN [2]) ou mesure de l'entropie (QUINLAN [17]).

Le critère est une mesure utilisée pour évaluer la qualité de la division des nœuds dans un arbre de décision. Les formules 2 et 3 définissent respectivement l'impureté de Gini et l'entropie.

$$gini(Q_m) = \sum_k p_{mk}(1 - p_{mk}) \in [0, 0.5], \qquad (2)$$

où Q_m représente les données au nœud m et p_{mk} est la probabilité de la classe k à un nœud m.

$$entropie(Q_m) = -\sum_k p_{mk} log_2(p_{mk}) \in [0, 1], \qquad (3)$$

où Q_m représente les données au nœud m et p_{mk} est la probabilité de la classe k à un nœud m.

Pour évaluer les performances des modèles de prédiction, nous utilisons la validation croisée stratifiée à k blocs. Cette méthode consiste à tester le modèle sur différentes partitions, appelées blocs, de l'ensemble de données d'entraînement. Nous veillons à ce que chaque bloc contienne la même distribution de classes que les données d'entraînement (stratification).

Enfin, une fois que les hyperparamètres sont optimisés, les données de test sont présentées au modèle entraîné. Les précisions, les rappels et les F-scores des modèles sont calculés pour chaque classe selon les trois équations suivantes :

$$pr\acute{e}cision = \frac{TP}{TP + FP} \in [0, 1], \qquad (4)$$

où TP est le nombre de vrais positifs, FN est le nombre de faux négatifs et FP est le nombre de faux positifs.

$$rappel = \frac{TP}{TP + FN} \in [0, 1], \qquad (5)$$

où TP est le nombre de vrais positifs, FN est le nombre de faux négatifs et FP est le nombre de faux positifs.

$$F_1 = \frac{pr\acute{e}cision \times rappel}{pr\acute{e}cision + rappel} \in [0, 1], \qquad (6)$$

Nous calculons également l'étendue des scores selon l'équation 7. Une faible étendue des scores signifie qu'ils sont semblables entre les trois classes. Ces étendues seront comparées avec les scores maximaux et minimaux. Le modèle idéal possède une faible étendue et des valeurs maximales et minimales proches de 1.

$$\acute{e}tendue = |score_{max} - score_{min}| \qquad (7)$$

En résumé, le pipeline 1 est le modèle de référence. Le pipeline 2 est le modèle permettant de simuler le cas où les données issues de Motive sont équilibrées. Le pipeline 3 est le modèle que nous pourrions appliquer aux données réelles si celles-ci sont déséquilibrées.

4 Résultats

Dans cette partie, nous allons présenter l'ensemble des résultats obtenus avec les données de test. Nous parlerons du pipeline 1, puis du pipeline 2 et, enfin, du pipeline 3. Nous terminerons en comparant les trois pipelines.

4.1 Analyse des résultats du pipeline 1

Dans ce paragraphe, nous présentons les résultats du premier pipeline sans sur-échantillonnage.
La table 1 montre les scores moyens de précision, de rappel et de F-score pour chaque modèle testé sur différentes classes et chapitres. Les scores présentés sont la moyenne des scores pour les trois classes sur chacun des quatre chapitres (2A, 3A, 5A et 5B). Sans l'utilisation du sur-échantillonnage, le modèle a obtenu une précision moyenne de 0,75, un rappel moyen de 0,60 et un F-score moyen de 0,60.

Pipeline	Méthode	Précision	Rappel	F-score
1	-	0,75	0,60	0,60
2	ROS	**0,85**	**0,84**	**0,85**
	SMOTE	0,81	0,80	0,80
3	ROS	0,65	0,65	0,64
	SMOTE	0,64	0,65	0,64

TABLE 1 – Précisions, rappels et F-scores moyens des chapitres 2A, 3A, 5A et 5B. Les scores en gras correspondent aux plus grandes valeurs par colonne.

Les tables 2, 3 et 4 présentent les scores de précision, de rappel et de F-score pour chaque classe et pour les quatre chapitres. Pour le pipeline 1, le modèle a obtenu des précisions de 0,83 pour la classe Abandon, 0,69 pour la classe Succès et 0,73 pour la classe Échec. Cela signifie que le modèle a correctement prédit un nombre relativement élevé d'échantillons positifs par rapport à l'ensemble des échantillons prédits comme positifs.

Pipeline	Méthode	Abandon	Succès	Échec
1	-	0,83	0,69	0,73
2	ROS	**0,92**	**0,78**	**0,86**
	SMOTE	0,90	0,72	0,82
3	ROS	0,88	0,70	0,36
	SMOTE	0,88	0,70	0,33

TABLE 2 – Précisions moyennes des chapitres 2A, 3A, 5A et 5B. Les scores en gras correspondent aux plus grandes valeurs par colonne.

En ce qui concerne le rappel, le modèle a obtenu des scores de 0,84 pour la classe Abandon, 0,78 pour la classe Succès et 0,17 pour la classe Échec. Cela suggère que le modèle est généralement efficace pour détecter les abandons et les succès des étudiants, mais il a du mal à détecter les échecs. Cette difficulté est probablement due au déséquilibre des données.

Pipeline	Méthode	Abandon	Succès	Échec
1	-	**0,84**	0,78	0,17
2	ROS	0,80	**0,84**	**0,89**
	SMOTE	0,78	0,83	0,80
3	ROS	0,78	0,83	0,34
	SMOTE	0,78	0,79	0,39

TABLE 3 – Rappels moyens des chapitres 2A, 3A, 5A et 5B. Les scores en gras correspondent aux plus grandes valeurs par colonne.

Pipeline	Méthode	Abandon	Succès	Échec
1	-	0,83	0,73	0,25
2	ROS	**0,85**	**0,81**	**0,88**
	SMOTE	0,83	0,77	0,81
3	ROS	0,82	0,75	0,35
	SMOTE	0,82	0,73	0,36

TABLE 4 – F-scores moyens des chapitres 2A, 3A, 5A et 5B. Les scores en gras correspondent aux plus grandes valeurs par colonne.

En résumé, les résultats de ce pipeline montrent que le déséquilibre des données impacte les scores et nous allons analyser comment rééquilibrer les données pour améliorer nos résultats avec les pipelines 2 et 3.

4.2 Analyse des résultats du pipeline 2

Nous allons maintenant détailler les résultats obtenus pour le pipeline 2, qui implique l'utilisation du sur-échantillonnage avant la séparation des données. Deux méthodes de sur-échantillonnage sont utilisées : le sur-échantillonnage aléatoire (ROS) et le sur-échantillonnage synthétique de la classe minoritaire (SMOTE).
Avec la méthode de sur-échantillonnage aléatoire (ROS), le modèle obtient une précision moyenne de 0,85, un rappel moyen de 0,84 et un F-score moyen de 0,85. Les scores de précision pour les trois classes varient entre 0,78 et 0,92, les scores de rappels varient entre 0,80 et 0,89 et les scores de F-score varient entre 0,81 et 0,88.
Les résultats obtenus avec le sur-échantillonnage synthétique de la classe minoritaire (SMOTE) sont légèrement inférieurs aux résultats du sur-échantillonnage aléatoire (ROS). Le modèle obtient une précision moyenne de 0,81, un rappel moyen de 0,80 et un F-score moyen de 0,80. Les scores de précision pour les trois classes varient entre 0,72 et 0,90, les scores de rappel varient entre 0,78 et 0,83 et les

scores de F-score varient entre 0,77 et 0,83.
En comparant les deux méthodes de sur-échantillonnage, le sur-échantillonnage aléatoire (ROS) obtient de meilleurs scores moyens que le sur-échantillonnage synthétique de la classe minoritaire (SMOTE) lorsque le sur-échantillonnage s'effectue avant la séparation des données. La méthode SMOTE créée des échantillons synthétiques à partir des échantillons originaux. Cette méthode de sur-échantillonnage semble introduire du bruit dans les données et, par conséquent, diminuer les performances des modèles de prédiction.

4.3 Analyse des résultats du pipeline 3

Enfin, nous présentons les résultats du pipeline 3 qui implique l'utilisation du sur-échantillonnage après la séparation des données.
Avec la méthode de sur-échantillonnage aléatoire (ROS), le modèle obtient une précision moyenne de 0,65, un rappel moyen de 0,65 et un F-score moyen de 0,64. Les scores de précision pour les trois classes varient entre 0,36 et 0,88, le scores de rappel varient entre 0,34 et 0,83 et les scores de F-scores varient entre 0,35 et 0,82.
Les résultats obtenus avec le sur-échantillonnage synthétique de la classe minoritaire (SMOTE) sont encore une fois légèrement inférieurs aux résultats du sur-échantillonnage aléatoire (ROS). Le modèle obtient une précision moyenne de 0,64, un rappel moyen de 0,65 et un F-score moyen de 0,64. Les scores de précision pour les trois classes varient entre 0,33 et 0,88, les scores de rappel varient entre 0,39 et 0,79 et les scores de F-score varient entre 0,36 et 0,82.
En comparant les deux méthodes de sur-échantillonnage, le sur-échantillonnage aléatoire (ROS) obtient des scores moyens comparables à ceux du sur-échantillonnage synthétique de la classe minoritaire (SMOTE) lorsque le sur-échantillonnage s'effectue après la séparation des données. Cependant, les scores obtenus avec le pipeline 3 sont inférieurs aux scores obtenus avec le pipeline 2.

4.4 Comparaison des pipelines

En comparant les trois pipelines, nous constatons que les meilleurs résultats sont obtenus avec le pipeline 2, lorsqu'un sur-échantillonnage aléatoire est effectué avant la séparation des données. Les prédictions sont donc meilleures lorsque les données sont équilibrées artificiellement puis séparées pour entraîner le modèle. Cependant, les données de test de ce pipeline ne sont pas représentatives d'une population réelle. Le pipeline 3 possède des données de test représentatives d'une population réelle. Il obtient des rappels moyens et F-scores moyens supérieurs, mais les précisions moyennes sont inférieures à celle du pipeline 1 (pour les deux méthodes de sur-échantillonnage). La présence du sur-échantillonnage semble donc diminuer les performances du modèle. La table 5 présente les précisions moyennes obtenues par le modèle à l'issue de l'entraînement et à l'issue du test. Le pipeline 1 obtient une précision moyenne d'entraînement de 0,84. Le pipeline 3 obtient une précision moyenne d'entraînement de 0,93 (ROS) et 0,91 (SMOTE). Le sur-échantillonnage des données d'entraînement entraîne donc un phénomène de sur-apprentissage des données. Le modèle du pipeline 3 généralise moins bien et possède donc une précision inférieure à celle du pipeline 1. Ce sur-apprentissage se constate sous une autre forme : le rappel moyen de la classe Abandon du pipeline 1 est supérieur aux rappels moyens de la même classe du pipeline 3 (pour les deux méthodes de sur-échantillonnage). Nous remarquons également que la précision moyenne de la classe Échec est plus élevée avec le pipeline 1 qu'avec le pipeline 3 (pour les deux méthodes de sur-échantillonnage).

Pipeline	Méthode	Entraînement	Test
1	-	0,84	0,75
2	ROS	0,91	0,85
	SMOTE	0,90	0,81
3	ROS	0,93	0,65
	SMOTE	0,91	0,64

TABLE 5 – Précisions moyennes des chapitres 2A, 3A, 5A et 5B lors de l'entraînement et lors du test du modèle.

L'étendue des scores est également à prendre en compte. En observant la table 6, nous constatons que l'étendue des scores est améliorée dans le cas du pipeline 2 par rapport aux deux autres pipelines. L'étendue des scores obtenus dans ce cadre est toujours inférieure à 0,2, ce qui montre une faible dispersion des résultats en fonction des classes et des chapitres et prouve que le sur-échantillonnage avant la séparation des données est le plus efficace pour prédire, de manière fiable, les résultats de étudiants.

Pipeline	Méthode	Précision	Rappel	F-score
1	-	0,15	0,67	0,59
2	ROS	**0,14**	0,09	**0,07**
	SMOTE	0,18	**0,05**	**0,07**
3	ROS	0,51	0,49	0,47
	SMOTE	0,54	0,40	0,46

TABLE 6 – Étendues des scores. Les scores en gras correspondent aux plus petites valeurs par colonne.

Finalement, la figure 7 donne une vision globale des résultats. Le sur-échantillonnage avant la séparation des données (pipeline 2) permet d'améliorer les scores du modèle de prédiction. Il obtient aussi les plus petites étendues des scores. La stratégie de sur-échantillonnage du pipeline 3 n'est pas assez performante pour obtenir des résultats satisfaisants. Notons également que cette étude est généralisable à l'ensemble des chapitres de Motive. En calculant les précisions, rappels et F-scores moyens de tous les chapitres de Motive, nous obtenons des valeurs comparables aux résultats de la table 2. Les étendues des scores sont légèrement supérieures à celles de la figure 9.

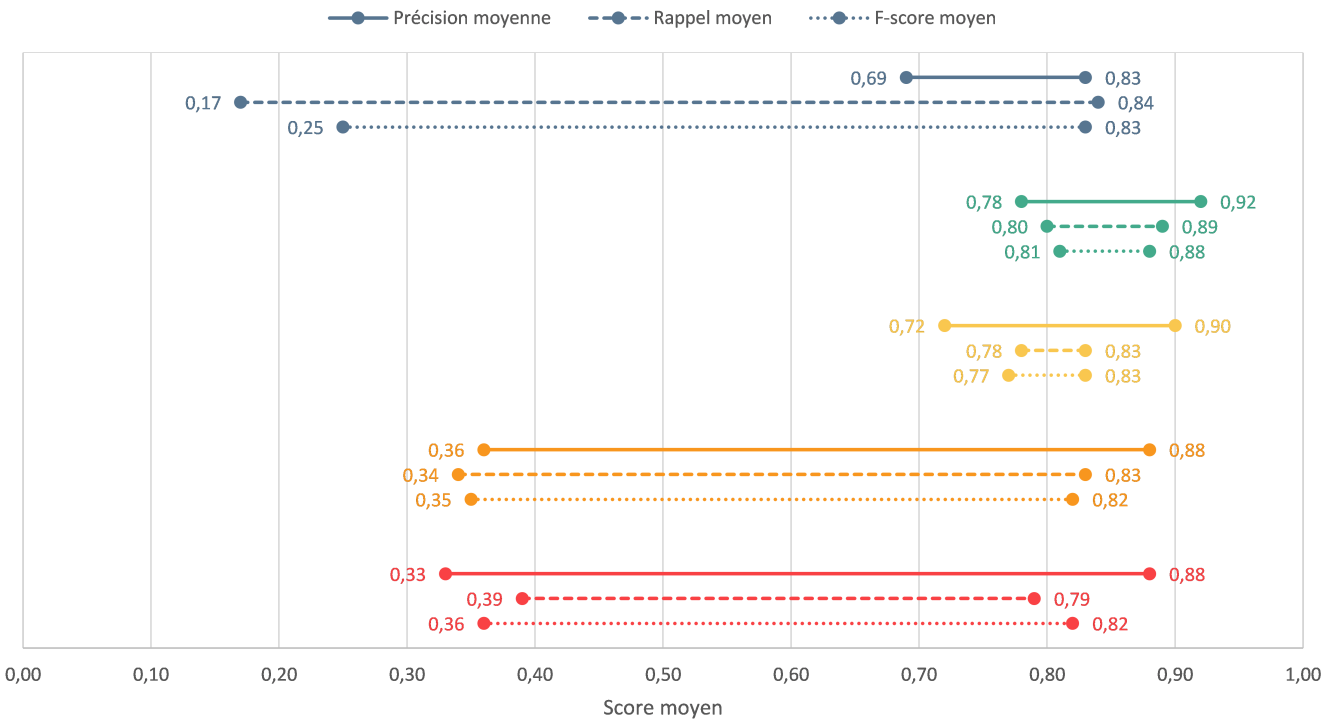

FIGURE 9 – Valeurs minimales et maximales des scores moyens des chapitres 2A, 3A, 5A et 5B. De haut en bas : en bleu, le pipeline 1 ; en vert, le pipeline 2 (ROS) ; en jaune, le pipeline 2 (SMOTE) ; en orange, le pipeline 3 (ROS) et en rouge, le pipeline 3 (SMOTE).

Notre objectif était de prédire des profils étudiants à partir des données issues de la plateforme d'apprentissage en ligne Motive. Ces données présentaient un déséquilibre entre les trois classes. Le sur-échantillonnage a permis d'améliorer les performances des modèles de prédiction de manière significative. Notre contribution réside dans l'utilisation de données (réelles et déséquilibrées) issues uniquement de la plateforme Motive. Aucune information sur les étudiants telles que la provenance géographique, l'historique scolaire et universitaire ou encore le niveau de stress n'ont été utilisé.

5 Conclusion et perspectives

L'objectif de ces travaux était de détecter des profils étudiants à partir des traces numériques des étudiants sur la plateforme Motive. Le principal problème de ces données était le déséquilibre des classes. La classe Abandon est beaucoup plus représentée que les classes Succès et Échec. Pour lutter contre ce déséquilibre, deux méthodes de sur-échantillonnage ont été testées : le sur-échantillonnage aléatoire (ROS) et le sur-échantillonnage synthétique de la classe minoritaire (SMOTE). Les méthodes ont été testées avant et après la séparation des données. Les résultats montrent que le sur-échantillonnage aléatoire avant la séparation des données est la meilleure méthode pour prédire au mieux les profils étudiants. Elle permet d'augmenter la précision, le rappel et F-score. Par ailleurs, le rappel de la classe Échec a considérablement augmenté grâce au sur-échantillonnage.
En utilisant les données d'entraînement des étudiants et en augmentant artificiellement le nombre d'échantillons, il est donc possible de prédire l'abandon, le succès ou l'échec de l'étudiant à un chapitre de Motive. Outre la notion de succès, ces prédictions peuvent permettent aux enseignants et aux professeurs de l'enseignement supérieur de détecter les étudiants en situation de décrochage scolaire, ou ceux en difficultés.
La suite de ces travaux portera sur l'analyse du Moodle complet de l'Université de Poitiers. Le principal défi réside dans l'étiquetage des données et la proposition d'indicateur pertinent dans le suivi de l'étudiant. En parallèle, la publication d'une base de données, similaire à celle de Motive, pourrait être bénéfique afin de permettre à la communauté scientifique de l'explorer, d'effectuer des expérimentations et de comparer les résultats obtenus.

Références

[1] S. K. BANIHASHEM et al. "A systematic review of the role of learning analytics in enhancing feedback practices in higher education". In : *Educational Research Review* 37 (2022), p. 100489.

[2] L. BREIMAN, éd. *Classification and regression trees*. 1998.

[3] L. BREIMAN. "Random forests". In : *Machine Learning* 45.1 (2001), p. 5-32.

[4] N. V. CHAWLA et al. “SMOTE : Synthetic Minority Over-sampling Technique”. In : *Journal of Artificial Intelligence Research* 16 (2002), p. 321-357.

[5] D. FOUNG et J. CHEN. “A Learning Analytics Approach to the Evaluation of an Online Learning Package in a Hong Kong University”. In : *Electronic Journal of e-Learning* 17.1 (2019).

[6] B. K. FRANCIS et S. S. BABU. “Predicting Academic Performance of Students Using a Hybrid Data Mining Approach”. In : *Journal of Medical Systems* 43.6 (2019), p. 162.

[7] H. HAN, W.-Y. WANG et B.-H. MAO. “Borderline-SMOTE : A New Over-Sampling Method in Imbalanced Data Sets Learning”. In : *Advances in Intelligent Computing*. T. 3644. 2005, p. 878-887.

[8] H. HASSAN, N. B. AHMAD et S. ANUAR. “Improved students’ performance prediction for multi-class imbalanced problems using hybrid and ensemble approach in educational data mining”. In : *Journal of Physics : Conference Series* 1529.5 (2020), p. 052041.

[9] O. IATRELLIS et al. “A two-phase machine learning approach for predicting student outcomes”. In : *Education and Information Technologies* 26.1 (2021), p. 69-88.

[10] J. KUZILEK, M. HLOSTA et Z. ZDRAHAL. “Open University Learning Analytics dataset”. In : *Scientific Data* 4.1 (2017), p. 170171.

[11] J. KUZILEK et al. “Analysing Student VLE Behaviour Intensity and Performance”. In : *Transforming Learning with Meaningful Technologies*. 2019, p. 587-590.

[12] J. MACQUEEN. “Some Methods For Classification And Analysis Of Multivariate Observations”. In : *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (1967).

[13] N. C. MING et V. L. MING. “Predicting Student Outcomes from Unstructured Data”. In : (2012).

[14] P. M. MORENO-MARCOS et al. “Analysis of the Factors Influencing Learners’ Performance Prediction With Learning Analytics”. In : *IEEE Access* 8 (2020), p. 5264-5282.

[15] A. NAMOUN et A. ALSHANQITI. “Predicting Student Performance Using Data Mining and Learning Analytics Techniques : A Systematic Literature Review”. In : *Applied Sciences* 11.1 (2020), p. 237.

[16] H. M. NGUYEN, E. W. COOPER et K. KAMEI. “Borderline over-sampling for imbalanced data classification”. In : *International Journal of Knowledge Engineering and Soft Data Paradigms* 3.1 (2011), p. 4.

[17] J. R. QUINLAN. *C4.5 : programs for machine learning*. The Morgan Kaufmann series in machine learning. Morgan Kaufmann Publishers, 1993. 302 p.

[18] T. WONGVORACHAN, S. HE et O. BULUT. “A Comparison of Undersampling, Oversampling, and SMOTE Methods for Dealing with Imbalanced Classification in Educational Data Mining”. In : *Information* 14.1 (2023), p. 54.