Facteurs de réussite étudiante dans un environnement numérique

Compte-rendu final – SY09

Belayachi Ibrahim Tahiri Mohamed Malaterre Sacha

Juin 2025

1 Introduction

Les environnements d'apprentissage hybrides, combinant présentiel et numérique, soulèvent des questions importantes sur l'impact de l'engagement en ligne sur la réussite académique. Ce rapport explore cette problématique à travers l'analyse d'un jeu de données issu d'un cours universitaire hybride, initialement décrit dans *Learning Analytics Methods and Tutorials*. Ce dataset, étant issu d'un cours, a fait l'objet d'analyses. De ce fait, aucun membre du groupe n'a consulté ces dernières pour éviter d'influencer nos méthodes. Nous avons bien eu conscience de ce problème dans le cadre de SY09, et nous avons décidé de changer notre approche en reconfigurant une des variables pour vraiment nous démarquer.

Le dataset comprend **130 étudiants**, caractérisés par trois catégories de variables : **démographiques** (origine, statut professionnel, mode de participation), **d'activité en ligne** (données fournies par le Learning Management System (Moodle) de l'école : consultations de contenus, interactions sociales, soumissions de devoirs) et de **performance académique** (notes par évaluation et note finale, ces notes sont sur 10). Deux variables dérivées, `ActivityGroup` (niveau d'activité en ligne) et `AchievingGroup` (niveau de réussite), ont été adaptées pour créer l'unicité de notre analyse. Afin d'obtenir une classification plus nuancée des étudiants et d'éviter une simple binarisation des résultats, nous avons redéfini la variable `AchievingGroup` à partir des notes finales (`Final_grade`) en trois niveaux de performance :

- **Low achievers** : étudiants ayant une note strictement inférieure à 5, correspondant à une situation d'échec.
- **Medium achievers** : étudiants ayant une note comprise entre 5 (inclus) et 8 (exclus), représentant une validation fragile ou moyenne.
- **High achievers** : étudiants ayant une note supérieure ou égale à 8, indiquant une réussite solide.

Ce découpage est cohérent avec la distribution naturelle des notes observée sur l'ensemble de l'échantillon. Comme le montre la figure ci-dessous, les notes présentent une structure en trois groupes distincts, validée par un clustering non supervisé (**KMeans**).

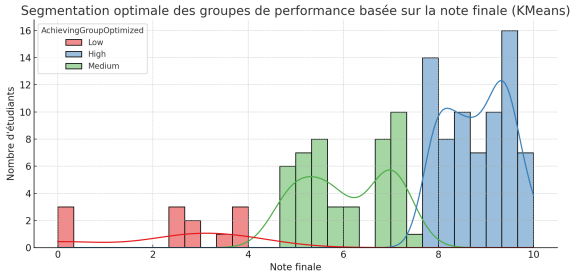

FIGURE 1 – Segmentation optimale des groupes de performance basée sur la note finale (**KMeans**)

Aucun nettoyage préalable n'a été nécessaire, les données étant déjà bien formatées et complètes. Nous explorons les liens entre engagement numérique et réussite académique en mobilisant différentes méthodes statistiques.

Les résultats mettront en lumière les corrélations significatives ou non, les profils types d'étudiants, s'il en existe, et les implications pédagogiques potentielles. Les détails techniques des variables sont synthétisés en annexe 2 pour alléger la présentation.

2 Objectifs

L'objectif principal est de modéliser le lien entre les comportements d'usage du LMS et les résultats finaux, en vue de construire un outil prédictif de détection des risques d'échec.

Plus précisément, nous cherchons à :

- Identifier des profils d'étudiants à partir de leurs comportements (clustering non supervisé) ;

— Reconstituer une variable cible fiable de performance finale ;
— Utiliser des méthodes d'analyse discriminante et de classification supervisée pour prédire si un étudiant risque ou non d'échouer.

Cette démarche vise à fournir des outils d'aide à la décision pédagogique, permettant aux enseignants de détecter en amont les étudiants en difficulté et d'agir en conséquence.

3 Visualisation des données

En analysant les données, nous cherchons à identifier des facteurs susceptibles d'expliquer l'existence potentielle d'inégalités entre les différents étudiants du dataset. Si certaines variables contextuelles comme le statut d'emploi ont montré des tendances intéressantes, nous nous concentrerons ici uniquement sur les données issues de l'utilisation de Moodle.

On peut néanmoins mentionner les visualisations sur les variables du travail étudiant, ou de la localisation, par exemple.

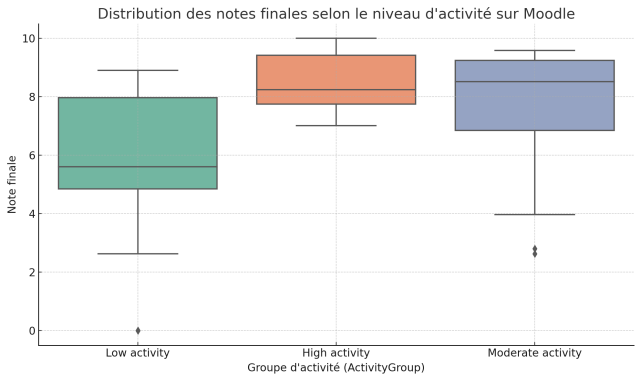

FIGURE 2 – Distribution des notes finales selon Le niveau d'activité Sur Moodle

La Figure 2 illustre une corrélation positive entre l'activité sur Moodle et les notes finales. Les étudiants les plus actifs obtiennent en moyenne de meilleurs résultats, tandis que ceux du groupe **Low Activity** présentent des performances nettement inférieures. Cette tendance soutient l'intégration des variables LMS dans les modèles prédictifs.

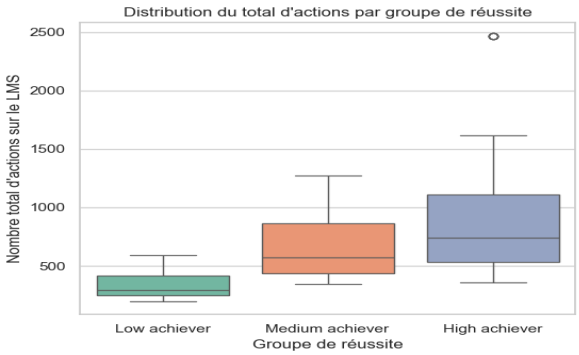

FIGURE 3 – Distribution du total d'actions par groupe de réussite

La Figure 3 présente un diagramme en boîte comparant la distribution du nombre total d'actions sur Moodle selon le groupe de réussite. On observe une tendance claire : les **High achievers** sont globalement plus actifs, avec une médiane nettement supérieure aux deux autres groupes. Les **Medium achievers** se situent dans une position intermédiaire, mais leur distribution est plus étalée, traduisant une variabilité des profils. À l'inverse, les **Low achievers** montrent une activité plus faible et concentrée. Cette répartition suggère une association positive entre le niveau d'engagement numérique et la réussite académique.

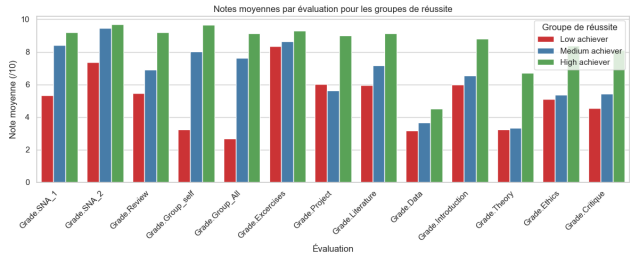

FIGURE 4 – Notes moyennes par évaluation pour les groupe de réussite

La Figure 4 met en évidence des écarts marqués entre les groupes de réussite sur l'ensemble des évaluations du cours. Les **High achievers** présentent systématiquement des moyennes plus élevées que les autres groupes, tandis que les **Low achievers** affichent des performances nettement inférieures. Les **Medium achievers** occupent une position intermédiaire mais plus proche des **Low achievers** sur certaines évaluations, ce qui reflète une hétérogénéité dans leur profil.

La Figure 5 présente une carte thermique des corrélations entre les actions LMS et la note finale. Les travaux de groupe (0,54) et les retours d'information (0,53)

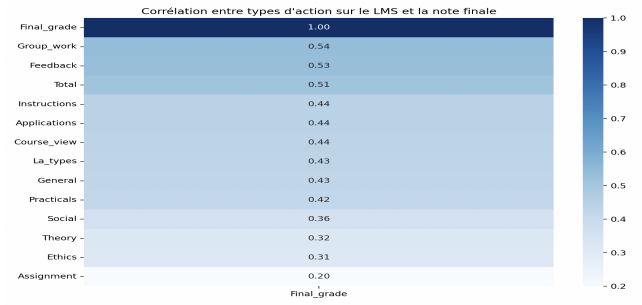

FIGURE 5 – Corrélation entre LMS et note finale

sont les plus corrélés à la réussite. Les instructions, applications pratiques et consultations de cours affichent des corrélations moyennes (0,44), tandis que les interactions sociales (0,36), discussions théoriques (0,32) et assignations (0,2) sont moins liées à la performance. Ces résultats soulignent l'importance de certaines activités dans la réussite étudiante.

4 Exploration non-supervisée des comportements étudiants

4.1 Analyse en composantes principales (ACP)

Grâce à l'analyse en composantes principales (ACP) on va pouvoir comprendre les relations dans des données multi-variées. Les fréquences d'utilisation de Moodle sont séparées en plusieurs variables. On va essayer de réduire la dimension avec l'ACP pour expliquer ces fréquences avec un minimum de dimensions. Les points sont les observations projetées, et les vecteurs sont les variables projetées.

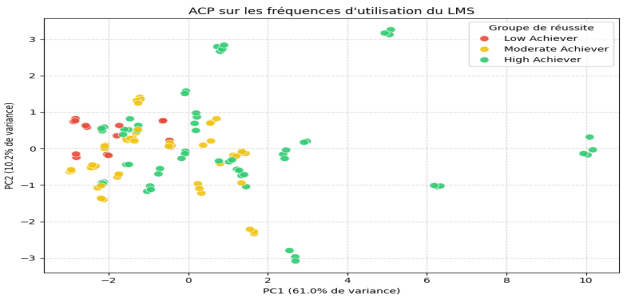

FIGURE 6 – ACP sur les fréquences d'utilisation de Moodle

Cette ACP en 2 composantes principales permet d'avoir 71,2% de variance expliquée. On peut distinguer très nettement certains **High**, les autres se mélangeant avec les **Medium**. Et les **Low** sont assez regroupés aussi.

Variable LMS	Contribution PC1 (%)	Contribution PC2 (%)
Frequency.Applications	5.54	25.63
Frequency.Assignment	6.86	1.42
Frequency.Course_view	12.42	1.48
Frequency.Feedback	7.44	14.61
Frequency.General	10.77	6.43
Frequency.Group_work	9.41	11.73
Frequency.Instructions	10.50	5.93
Frequency.La_types	7.34	6.94
Frequency.Practicals	6.58	2.36
Frequency.Social	10.07	6.33
Frequency.Ethics	6.50	12.50
Frequency.Theory	6.58	4.65

TABLE 1 – Contributions des variables aux deux premières composantes principales (ACP sur les fréquences LMS)

L'analyse des contributions permet de mieux interpréter les dimensions de l'ACP. La première composante (PC1) est influencée par les variables `Course_view`, `General`, `Instructions`, `Social` et `Group_work`, suggérant un axe d'engagement global sur la plateforme. La deuxième composante (PC2) est fortement influencée par `Applications`, `Feedback`, `Ethics` et `Group_work`, davantage axée sur l'implication réflexive et l'apprentissage par retour ou production. Ces résultats confirment la diversité des comportements LMS et justifient leur intégration dans les analyses prédictives ultérieures.

4.2 Clustering des comportements avec K-Means

Pour prolonger l'analyse exploratoire en ACP, nous appliquons l'algorithme **K-Means** sur les données LMS standardisées, afin d'extraire des profils d'engagement *sans utiliser les résultats académiques* comme critère de regroupement, afin de tester dans quelle mesure les comportements uniquement permettent de structurer des groupes cohérents.

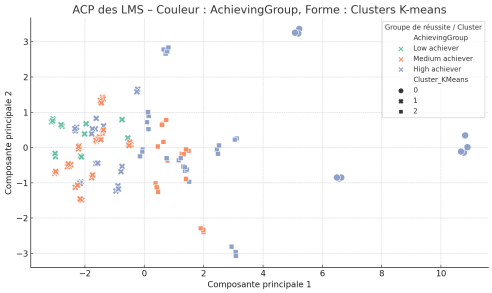

FIGURE 7 – Corrélation visuelle entre profils d'engagement numérique et réussite académique

La Figure 7 présente la projection ACP avec :
— la couleur indiquant le groupe de réussite réel (`AchievingGroup`) ;
— la forme représentant le cluster **K-Means**.

Cette superposition permet d'évaluer a posteriori si les groupes de comportements identifiés par clustering reflètent partiellement les performances académiques. Certains clusters présentent une supériorité de profils appartenant à un même niveau de réussite, en particulier pour les **High achievers**. On constate que certains clusters sont composés majoritairement d'un même type d'étudiants, notamment les **High achievers**. Cela indique que certains profils d'engagement LMS se distinguent et peuvent potentiellement servir à prédire la réussite.

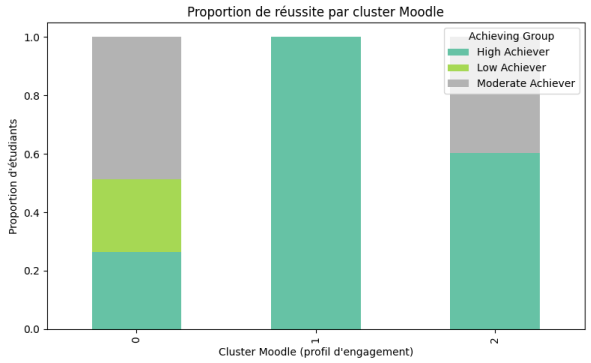

FIGURE 8 – Proportion de réussite par cluster Moodle

La Figure 8 permet de croiser les clusters issus de **K-Means** avec les groupes de réussite académique. On observe que certains clusters sont clairement dominés par un seul type d'étudiant. Par exemple, le **cluster 1** regroupe presque exclusivement des **High achievers**, tandis que le **cluster 0** présente une distribution plus équilibrée entre les trois groupes. Cela suggère que les comportements captés par les variables LMS permettent dans une certaine mesure d'anticiper le niveau de performance, même si certaines zones restent floues.

5 Modélisation supervisée de la réussite académique

5.1 Objectif et démarche

Après avoir mis en évidence des liens forts entre engagement numérique et performance, l'objectif est ici de prédire automatiquement le groupe de performance académique (`AchievingGroup`) d'un étudiant à partir de ses interactions avec la plateforme Moodle. Cette prédiction permettrait, dans une optique d'alerte précoce, de détecter les étudiants à risque.

Nous limitons volontairement les variables explicatives aux seules données d'activité sur Moodle, afin de tester si l'engagement numérique seul suffit à prédire la réussite.

5.2 Préparation des données

Les étapes suivantes ont été réalisées :

— **Sélection des variables** : uniquement les fréquences d'interaction Moodle ;
— **Standardisation** : chaque variable est centrée et réduite ;
— **Découpage** :
  — 80% pour l'**entraînement** (modèle construit),
  — 20% pour le **test** (modèle évalué).

Ce découpage garantit une évaluation fidèle de la capacité de généralisation du modèle.

5.3 Régression logistique : premiers résultats

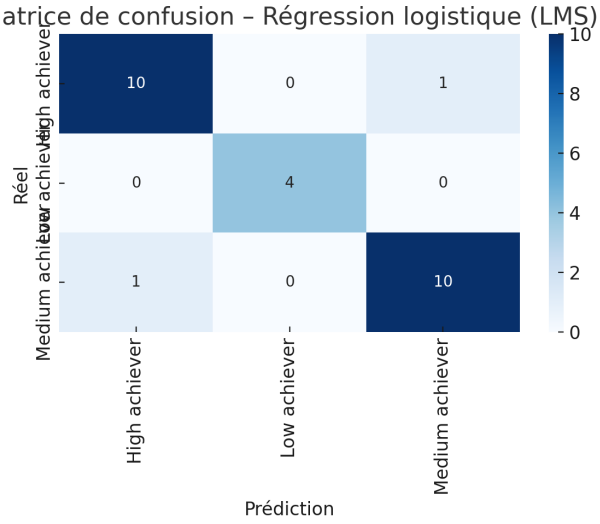

FIGURE 9 – Matrice de confusion – Régression logistique

Par la **régression logistique**, on obtient plus de 90% de prédictions correctes. Les étudiants en situation d'échec (**Low achievers**) sont parfaitement identifiés. Les erreurs concernent surtout la confusion entre **Medium** et **High**, ce qui peut s'expliquer par des profils LMS similaires. Nous avons également testé la méthode par arbre de décision, mais ses performances (80% de bonnes prédictions, voir figure 15) sont inférieures à

celles de la régression logistique, donc elle n'a pas été retenue.

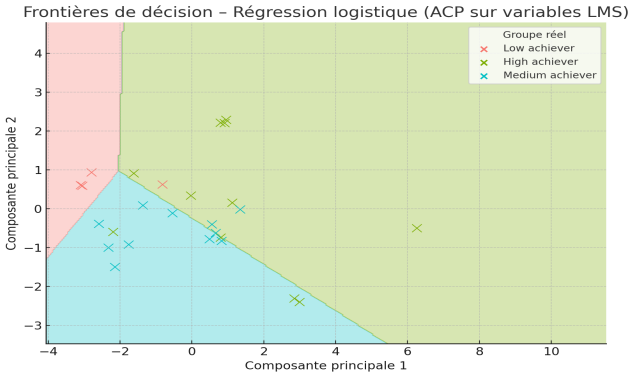

FIGURE 10 – Frontières de décision – Projection ACP des variables LMS

La projection ACP avec frontières de décision révèle une **bonne séparation spatiale des classes**, notamment pour les étudiants à risque. Cette représentation illustre la capacité du modèle à exploiter efficacement les différences de comportement LMS.

5.4 Random Forest pour prédire l'AchievingGroup

Afin d'améliorer la performance prédictive, un modèle de **Random Forest** a été entraîné. Cette méthode, qui repose sur un ensemble d'arbres de décision, est réputée pour sa robustesse et sa capacité à modéliser des relations complexes.

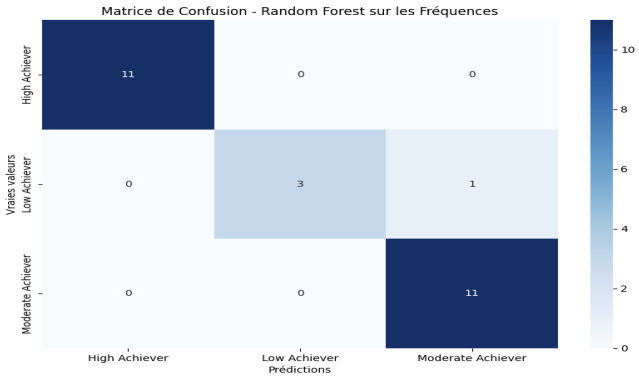

FIGURE 11 – Matrice de confusion – Random Forest

L'analyse de la **matrice de confusion** (Figure 11) révèle une performance de classification très élevée. Le modèle identifie avec une quasi-parfaite exactitude les **Low achievers** et les **High achievers**. Les quelques erreurs de classification restantes concernent principalement la distinction entre les **Medium achievers** et les **High achievers**, ce qui suggère une plus grande similarité dans les profils d'engagement numérique de ces deux groupes.

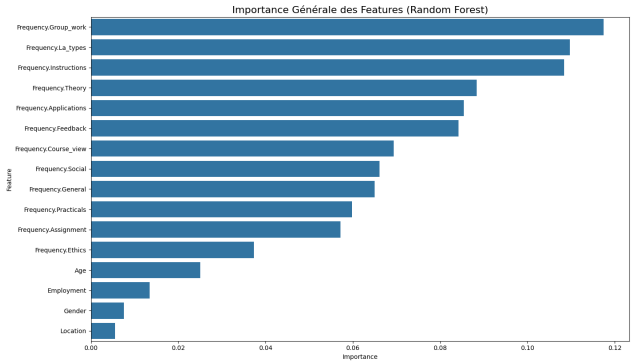

FIGURE 12 – Importance de chaque variable – Random Forest

La figure 12 présente l'**importance des variables** (feature importance), qui quantifie la contribution de chaque type d'action LMS à la performance du modèle. Conformément aux analyses précédentes, les variables telles que `Frequency.Group_work`, `Frequency.Course_view`, et `Frequency.Feedback` émergent comme les prédicteurs les plus influents de la réussite académique. À l'inverse, des activités comme la soumission de devoirs (`Frequency.Assignment`) ont une importance moindre, indiquant que la fréquence de cette action seule est moins discriminante que les comportements d'interaction continue avec les ressources du cours.

On peut notamment apercevoir que les variables qui ne sont pas liées à l'activité en ligne, notamment l'âge, le sexe, la situation d'emploi et la localisation (sur campus ou en ligne) apparaissent comme étant les variables qui contribuent le moins à la performance du modèle. Cela confirme notre hypothèse qui consistait à ne pas les prendre en compte lors de la conception de nos modèles.

5.5 Régression linéaire : prédire la note finale

En complément des approches de classification supervisée, nous avons testé une **régression linéaire** pour modéliser la note finale des étudiants (`Final_grade`) en fonction de leurs variables d'engagement sur Moodle.

Méthodologie : Les données LMS ont été standardisées par centrage-réduction, puis divisées en deux sous-ensembles : 80% des individus ont été utilisés pour

l'**entraînement** afin d'ajuster le modèle, et les 20% restants ont servi en tant que jeu de **test**. Le modèle apprend une fonction linéaire reliant les fréquences d'utilisation des ressources Moodle à la note finale obtenue par l'étudiant.

Résultats : la régression linéaire obtient un $\mathbf{R^2} = \mathbf{0.75}$ sur le jeu de test, indiquant que 75% de la variance de la note finale peut être expliquée par les seules données LMS. Les erreurs de prédiction (MAE : 0.57, RMSE : 0.72) restent faibles, montrant la pertinence de ce modèle.

Interprétation : certaines variables comme `Frequency.GroupWork`, `Frequency.Feedback` et `Frequency.Application` présentent des coefficients fortement positifs, confirmant leur lien avec la réussite. Ce modèle fournit donc une estimation fine et continue de la performance finale, complémentaire à la classification en groupes.

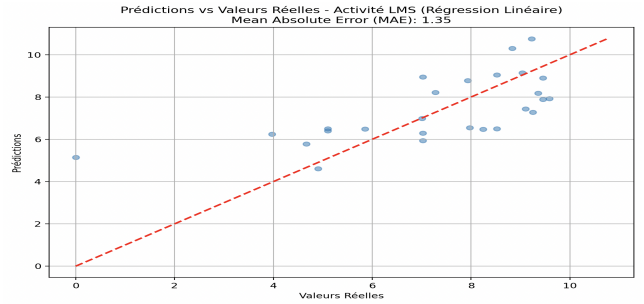

FIGURE 13 – Comparaison des notes réelles et prédites (régression linéaire)

Cette approche enrichit l'analyse en fournissant une estimation quantitative personnalisée de la réussite, là où les méthodes précédentes opéraient par classes.

6 Conclusion

Ce projet visait à évaluer dans quelle mesure l'engagement numérique des étudiants sur la plateforme Moodle peut être associé à leur réussite académique. L'analyse d'un jeu de données issu d'un cours hybride montre des corrélations notables entre la fréquence d'activité en ligne et la performance, sans pour autant établir un lien de causalité direct.

L'analyse exploratoire **(ACP, K-Means)** met en évidence des profils distincts d'engagement, dont certains sont majoritairement associés à un niveau de réussite. Toutefois, cette correspondance n'est pas systématique, ce qui reflète la complexité des comportements étudiants. La modélisation supervisée confirme l'intérêt prédictif des données LMS, en particulier pour identifier les étudiants en situation d'échec.

Cependant, ces résultats doivent être interprétés avec prudence. D'une part, ils sont issus d'un contexte pédagogique spécifique, avec un échantillon limité à un seul cours et environnement. D'autre part, les comportements numériques n'explicitent qu'une partie de l'engagement étudiant. Ce dernier inclut également des dimensions **pédagogiques** (stratégies d'apprentissage), **émotionnelles** (motivation, stress) et **comportementales** (participation, interactions humaines), qui ne sont pas rendues ici.

De plus, les prédicteurs identifiés sont dépendants de l'environnement numérique analysé : si les ressources pédagogiques ou la plateforme LMS évoluent, les modèles devront être réajustés.

Ainsi, bien que les données LMS puissent alimenter des outils d'alerte précoce, elles ne doivent pas être utilisées seules. Une approche qui combine les données numériques, le contexte du cours et les différents aspects de l'engagement des étudiants permettrait mieux de prévoir et d'accompagner leur réussite.

Annexe

A Description des variables

Variable	Description
User	Identifiant de l'étudiant dans le système d'apprentissage en ligne (LMS)
Name	Prénom de l'étudiant
Surname	Nom de famille de l'étudiant
Origin	Pays d'origine de l'étudiant
Gender	Sexe de l'étudiant : F pour féminin, M pour masculin
Birthdate	Date de naissance de l'étudiant
Location	Lieu d'études : indique si l'étudiant suit le cours sur le campus (présentiel) ou à distance
Employment	Statut d'emploi en parallèle des études : aucun emploi, temps partiel ou temps plein
Frequency.Applications	Nombre d'événements liés à la ressource « Applications »
Frequency.Assignment	Nombre d'événements liés aux soumissions de devoirs
Frequency.Course_view	Nombre de visites de la page principale du cours
Frequency.Feedback	Nombre de consultations des retours (feedback) sur les devoirs
Frequency.General	Nombre d'événements liés aux ressources générales du cours
Frequency.Group_work	Nombre d'événements liés au travail de groupe (projet en équipe)
Frequency.Instructions	Nombre de consultations des consignes des devoirs
Frequency.La_types	Nombre d'événements liés à la ressource « LA types » (types d'analytique de l'apprentissage)
Frequency.Practicals	Nombre d'événements liés aux travaux pratiques
Frequency.Social	Nombre d'événements liés aux interactions sociales (forums de discussion)
Frequency.Ethics	Nombre d'événements liés à la ressource « Ethics » (module sur l'éthique)
Frequency.Theory	Nombre d'événements liés à la ressource « Theory » (contenu théorique)
Frequency.Total	Nombre total d'événements enregistrés pour l'étudiant (toutes actions confondues)
Grade.SNA_1	Note du premier devoir d'analyse de réseaux sociaux (SNA) - sur 10
Grade.SNA_2	Note du deuxième devoir SNA - sur 10
Grade.Review	Note du devoir de revue d'études - sur 10
Grade.Group_self	Note individuelle obtenue pour le projet de groupe - sur 10
Grade.Group_All	Note collective (de groupe) pour le projet de groupe - sur 10
Grade.Exercises	Note aux exercices pratiques - sur 10
Grade.Project	Note du projet final - sur 10
Grade.Literature	Note du devoir de revue de littérature - sur 10
Grade.Data	Note du devoir d'analyse de données - sur 5
Grade.Introduction	Note du devoir d'introduction - sur 10
Grade.Theory	Note du devoir sur la théorie - sur 10
Grade.Ethics	Note du devoir sur l'éthique - sur 10
Grade.Critique	Note du devoir de critique - sur 10
Final_grade	Note finale obtenue pour le cours - sur 10
ActivityGroup	Catégorie de niveau d'activité de l'étudiant : activité élevée, modérée ou faible selon les actions totales
AchievingGroup	Catégorie de réussite académique : performances élevées, moyennes ou faibles selon la note finale

TABLE 2 – Description des variables utilisées dans l'étude

B Arbres de décisions

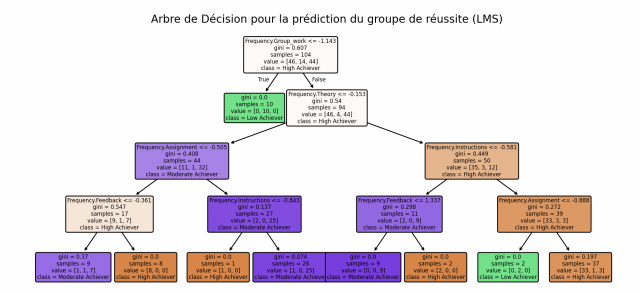

FIGURE 14 – Arbre de décision sur activité LMS

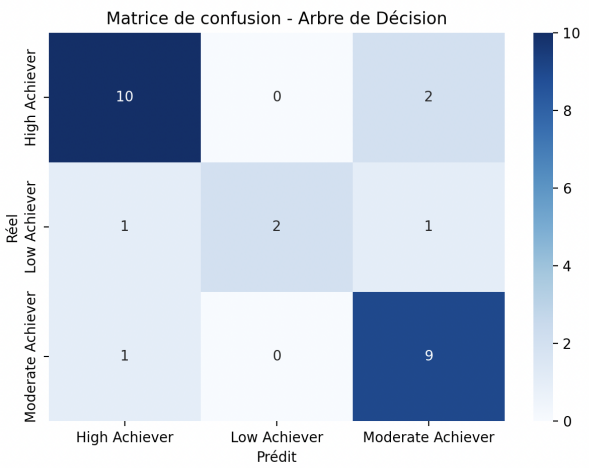

FIGURE 15 – Matrice de confusion, arbre de décision sur activité LMS