الجمهورية الجزائرية الديمقراطية الشعبية

République Algérienne Démocratique et Populaire

Ministère de l'enseignement supérieur et de la recherche scientifique

Université de 8 Mai 1945 - Guelma -

Faculté des Mathématiques, d'Informatique et des Sciences de la matière

Département d'Informatique

Mémoire de fin d'étude en master

Filière : Informatique

Option : STIC

Réalisation d'un système de classification des apprenants à partir d'indicateurs d'évaluation de leur apprentissage

Présenté par :

BOUNEMRA Randa

Devant le jury composé de :

Dr. BENDJEBAR Safia
Pr. LAFIFI Yacine
GOUASMI Noureddine

Juin 2023

Résumé

Pendant la pandémie de COVID 19, les établissements universitaires ont été contraints d'intégrer le e-learning dans leurs stratégies d'apprentissage. Mais l'absence de contact entre pairs lors des activités pédagogiques est un frein à un bon enseignement, la collaboration dans l'apprentissage en ligne, permettant d'augmenter le niveau d'échange d'idées dans un groupe et la stimulation mutuelle, encourageant ainsi l'interaction entre les apprenants.

Dans un système d'apprentissage collaboratif, l'évaluation de l'apprentissage d'un apprenant se fait à travers des tests individuels ou/et des projets collaboratifs. Si les tests permettent de valider l'acquisition de connaissances de l'apprenant, les projets évaluent le résultat de l'application des connaissances acquises par un groupe d'apprenants. Mais pour pouvoir augmenter les interactions entre apprenant, et leur qualité, il est nécessaire de regrouper ensemble des apprenants aux profils complémentaires.

Notre projet de fin d'étude porte sur la conception et la réalisation d'un système de classification des apprenants à partir d'indicateurs d'évaluation de leur apprentissage, pour pouvoir créer des groupes collaboratifs.

Le but à atteindre est la formation de groupes hétérogènes pour favoriser la collaboration et l'obtention de meilleurs résultats, ainsi que la proposition et la comparaison de deux méthodes de regroupement automatique l'une basées sur la classification multi-label et et l'autre, à travers un algorithme génétique.

Mots-clés : Learning analytics, educational data mining, classification multi-label, méthode ML-kNN, algorithme génétique.

Remerciements

Tout d'abord, je remercie le bon " Dieu " puissant de la bonne santé, la volonté et de la patience qu'il nous a donnée tout au long de notre étude de mener ce travail durant toute cette année.

Tout d'abord, ce travail ne serait pas aussi riche et n'aurait pas pu avoir le jour sans l'aide et l'encadrement de monsieur Gouasmi Noureddine, qui a accepté de suivre ce travail. on le remercie pour la qualité de son encadrement exceptionnel, pour son patience, son rigueur et son disponibilité durant notre préparation de ce mémoire.

Mes remerciements vont aussi à tous les membres de jury qui malgré leurs hautes fonctions et ses lourdes responsabilités avoir accepté d'examiné notre mémoire.

Merci beaucoup ma chère maman, la lumière de ma vie, le bonheur de mon existence qui m'a toujours soutenu en toutes circonstances et qui me donnent de la force et la volonté d'avancer et mon cher papa qui a sacrifie toute sa vie à fin de me voir devenir ce que je suis, je vous dis infiniment merci que dieu vous garde et vous accorde longue vie.

Ma deuxième mère qui m'avoir soutenue, encouragée et conseillée.

Mes sœurs Nawel et sara et mes frères Mohamed, Djamel et Houssem qui je souhaite réussitent et persévérance dans tous ce qu'ils empreindront.

Mes familles Bounemra et Salhi qui m'ont donnée l'espoir de continuer ce travail.

A mes amis qui m'ont courage, surtout Rami, doua et amina.

A tous ceux qui m'ont aidé à réaliser ce travail, A tous ceux que j'ai oubliés... Excusez-moi.

Table des matières

Table des figures

Liste des tableaux

Introduction générale

L'apprentissage en ligne, également connu sous le nom d'e-learning, a connu une popularité croissante ces dernières années. Il offre aux apprenants la flexibilité de suivre des cours à distance, à leur propre rythme et selon leur emploi du temps. Dans ce contexte, l'apprentissage collaboratif a émergé comme une approche pédagogique efficace, favorisant l'interaction entre les apprenants et encourageant l'échange de connaissances et d'idées.

L'apprentissage collaboratif permet aux apprenants de travailler en équipe sur des projets et des tâches communes, ce qui leur permet d'acquérir des compétences de résolution de problèmes, de communication et de travail d'équipe. Cependant, il devient essentiel de bien choisir les membres composant un groupe pour bénéficier des avantages offert par un groupe coopérant et collaboratif.

Notre mémoire de fin d'études porte sur la réalisation d'un système de classification des apprenants à partir d'indicateurs d'évaluation de leur apprentissage, dans le but de regrouper les apprenants automatiquement dans des groupes d'apprenants aux caractéristiques hétérogènes.

Dans ce mémoire, nous explorons une méthode de classification multi-label pour pouvoir classer les étudiants dans plusieurs catégories, selon des indicateurs d'activité sur un système d'apprentissage, catégories à partir desquelles les apprenants sont choisis pour constituer des groupes hétérogènes. Ensuite, les groupes obtenus par cette méthode de regroupement sont comparés avec ceux obtenus par une méthode de regroupement utilisant un algorithme génétique.

Notre mémoire est organisé comme suit :

- Dans le premier chapitre, nous avons présenté deux aspects de l'analyse des données appliqué au domaine de l'apprentissage : les *Learning Analytics* et l'*Educational Data Mining*

- Dans le deuxième chapitre, nous explorons les diverses techniques de classification, en analysant leurs bénéfices et leurs limitations. Nous passons en revue les approches variées de la classification multi-label, et nous présentons également des travaux de recherche dans ce domaine.

- Le troisième chapitre est dédié à la description de la conception globale de notre système, en commençant par énoncer les objectifs du projet. Ensuite, nous présentons les deux méthodes de regroupement intégrées dans notre système.

- Finalement, dans le dernier chapitre, nous présentons une expérimentation du système et un comparatif entre les deux méthodes de regroupement.

Chapitre 1

Learning Analytics et Educational Data Mining

1.1 Introduction

L'analyse des données est la branche des statistiques qui traite de la description des données en général. Ces méthodes visent à établir des liens possibles entre différentes données et à en déduire des informations statistiques décrivant plus précisément les principaux rapports entre ces données [1].

Ainsi, En Business Intelligence, l'analyse de données peut aider les entreprises à prendre des décisions stratégiques en examinant les données sur les ventes, les finances, les ressources humaines et d'autres domaines-clés [10], alors que dans le domaine de l'apprentissage, cela peut aider à mieux comprendre son efficacité, en examinant les données sur les étudiants, les cours et les interactions en ligne.

Dans ce qui suit, nous allons présenter deux aspects de l'analyse de données appliqués au domaine de l'apprentissage : les *Learning Analytics (LA)* et l'*Educational Data Mining (EDM)*

1.2 Learning Analytics

1.2.1 Définition

Les learning analytics apparaissent comme un lien entre les données éducatives et l'apprentissage. Bien qu'il n'y ait pas d'accord sur une définition standard des LA, une conceptualisation large permet de le définir comme :

"la mesure, la collecte, l'analyse et la communication de données sur les apprenants et leurs contextes, dans le but de comprendre et d'optimiser l'apprentissage et les environnements dans lesquels il se produit." [15]

LA est un domaine interdisciplinaire qui se concentre sur l'analyse de données provenant d'environnements d'apprentissage numériques pour comprendre les processus d'apprentissage et améliorer la qualité de l'enseignement et l'efficacité de l'apprentissage. Il s'appuie sur des données sur les étudiants, les enseignants, les cours et les environnements d'apprentissage pour prendre des décisions dans les domaines de l'éducation et de la formation [15].

LA inclut un grand nombre de méthodes, telles que : l'intelligence artificielle (IA), l'analyse statistique, l'apprentissage automatique, l'intelligence économique [39], et les systèmes d'aide à la décision [15].

1.2.2 Historique

Les principales étapes dont l'histoire des LA sont [25] :

- Années 1990, en Angleterre, naissance de l'EBM (Evidence-Based Education), approche pédagogique basée sur le profilage des étudiants et une analyse des profil par des méthodes de l'IA.
- Dès 1994, première utilisation du Data Mining dans le monde de la recherche en éducation, en exploitant les méthodes issus du marketing et du Business Analytics.
- En 2011, création de la *Society for Learning Analytics Research (SoLAR)* et le *Journal of Learning Analytics (JLA)*, ainsi que la conférence internationale *Learning Analytic for Knowledge (LAK)*.

1.2.3 Objectif des LA

L'objectif principal de cette discipline est d'améliorer la qualité de l'enseignement et l'efficacité de l'apprentissage en utilisant des méthodes quantitatives et informatiques pour analyser les données provenant d'environnements d'apprentissage numériques [25].

Il s'agit de pouvoir identifier les possibilités d'amélioration de l'enseignement, la personnalisation de l'apprentissage, la prédiction des performances futures des étudiants, et principalement les étudiants à risque, et la détermination des facteurs qui contribuent à l'efficacité de l'apprentissage, d'optimiser l'enseignement des enseignants, d'évaluer l'efficacité des programmes pédagogiques, de soutenir la prise de décisions éclairées et de favoriser la recherche et l'innovation dans le domaine de l'éducation et de la formation [13, 25].

Parmi les objectifs, on peut citer [13, 6, 25] :

- Collecter et analyser les traces numériques laissées par les apprenants afin de comprendre et d'améliorer les processus d'apprentissage.
- Personnaliser l'expérience d'apprentissage pour chaque individu.
- Détecter les difficultés d'apprentissage et fournir un soutien adapté.
- Prévoir les performances des apprenants et intervenir de manière proactive.

1.2.4 Méthodes utilisées

Il existe plusieurs méthodes utilisées pour les LA, chacune ayant ses avantages et ses inconvénients en fonction du domaine d'application et des objectifs de l'analyse.

Quelques-unes des méthodes les plus fréquemment utilisées sont les suivantes :

1. **Analyse statistique :** Les méthodes statistiques sont largement utilisées pour analyser les données d'apprentissage. Cela peut inclure des techniques telles que l'analyse descriptive (moyennes, écarts-types, etc.), l'analyse de corrélation, l'analyse de variance (ANOVA), les régressions linéaires, les modèles de régression logistique, etc. [18].

2. **Visualisation de données :** cette méthode utilise des représentations graphiques pour représenter les données d'apprentissage et faciliter leur compréhension. Les méthodes de visualisation de données couramment utilisées incluent les graphiques en barres, les histogrammes et les nuages de points [18].

3. **Analyse des réseaux sociaux :** Dans les environnements d'apprentissage en ligne, les interactions entre les apprenants peuvent être analysées à l'aide de méthodes d'analyse des réseaux sociaux. Cela permet de comprendre les schémas de collaboration, l'influence sociale, la diffusion de l'information, etc. [18].

1.2.5 Avantages et inconvénients

Parmi les avantages des LA, on peut citer [13, 25] :

- Amélioration de la qualité de l'enseignement en utilisant des données pour prendre des décisions éducatives.
- Personnalisation de l'apprentissage en utilisant des données pour comprendre les besoins individuels des étudiants et adapter l'enseignement en conséquence.
- Optimisation de l'efficacité de l'apprentissage en identifiant les points forts et les opportunités d'amélioration dans les processus d'apprentissage.
- Détection précoce des problèmes d'apprentissage en utilisant des données pour surveiller les progrès et anticiper les problèmes potentiels.

On peut également présenter quelques inconvénients des LA [13, 25] :

- Manque de transparence dans la collecte et l'utilisation des données, ce qui peut entraîner une perte de confiance des étudiants et des enseignants.
- Coût élevé de la mise en place et de la maintenance des systèmes de collecte de données.
- Nécessité d'une expertise en analyse de données pour interpréter correctement les résultats.
- Risque de biais dans les décisions éducatives si les données sont mal interprétées ou incorrectement utilisées.

1.2.6 Exemples d'utilisation des LA

Les cas d'utilisation des LA sont nombreux et variés. Il s'agit d'une approche qui peut être utilisée dans différents domaines pour améliorer la prise de décision et la compréhension des données.

Ci-dessous quelques exemples courants :

Modélisation d'apprenants et détection des profils

Les interactions des apprenants, connues sous le nom de *traces d'apprentissage*, fournissent des informations précieuses sur les caractéristiques et les modèles de l'apprenant. Les métriques telles que la motivation et l'engagement de l'apprenant peuvent être déterminées en utilisant des indicateurs tels que le temps passé sur le parcours d'apprentissage, le nombre de contenus d'apprentissage visualisés, etc.

Grâce à ces modèles, il est possible de déterminer des profils d'apprenants permettant aux enseignants de mieux comprendre leurs étudiants et mettre en place des stratégies d'enseignement adaptées [13].

Meilleur suivi de l'apprenant et mise en place d'actions de remédiation

Le feedback sur les activités des étudiants peut être obtenu grâce à l'analyse des données d'apprentissage, permettent à l'enseignant d'avoir des informations en temps réel sur ces activités individuels pour chaque apprenant. Cela permet un suivi plus efficace et une aide personnalisée plus rapide [13].

Tableaux de bord

Les tableaux de bord en LA sont largement utilisés et ont été démontrés comme ayant un impact positif sur les résultats des apprenants. Ils permettent de visualiser les acquisitions de compétences des étudiants [6].

Prédiction d'abandon dans les MOOC

Ces dernières années, les MOOC (Massive Open Online Courses) ont gagné en popularité et ont suscité un vif intérêt de la part des institutions universitaires du monde entier. Outre leur capacité à offrir de nouvelles méthodes d'apprentissage, les MOOC permettent également de recueillir d'importantes données d'utilisation, qui peuvent être utilisées pour améliorer la qualité de l'enseignement.

L'un des aspects distinctifs des MOOC par rapport aux outils pédagogiques traditionnels est leur accès libre. Cependant, cela entraîne généralement un taux d'abandon élevé, avec plus de 90% des inscrits qui abandonnent en cours de route. Ainsi, l'un des principaux défis de cet outil consiste à prévenir les décrochages [6].

Détection d'élèves à risque

Les systèmes de notification de situations d'échec potentiel des élèves, destinés aux enseignants, sont de plus en plus populaires au sein des universités. Ces outils

visent à améliorer l'apprentissage en identifiant les élèves qui risquent de ne pas réussir et en informant les enseignants de ces situations [6].

1.3 Educational Data Mining

1.3.1 Définition

L'EDM est un sous-domaine de l'analyse de données qui se concentre sur l'analyse de grandes quantités de données relatives aux processus d'apprentissage et d'éducation.

Il applique des techniques d'exploration de données et d'analyse statistique aux données éducatives afin d'extraire des connaissances et des informations utiles pour améliorer la compréhension des processus d'apprentissage et à optimiser l'enseignement en utilisant des méthodes statistiques, d'apprentissage automatique et d'analyse de données pour extraire des connaissances utiles sur les étudiants, les cours, les enseignants et les environnements d'apprentissage [2].

1.3.2 Historique

Les étapes-clés de l'évolution des EDM sont les suivantes :

- À partir de 2000, naissance de l'Educational Data Mining à la conférence ITS (Intelligent Tutor Systems) de Montréal en 2000 [25].
- En 2005, la première conférence internationale sur l'EDM a été organisée, mettant l'accent sur les méthodes statistiques et d'apprentissage automatique pour l'analyse de données éducatives [33].
- En 2006, la première école d'été sur l'EDM a eu lieu, offrant aux participants une formation pratique sur les techniques et les outils de l'EDM [3].
- En 2009, naît la première revue consacrée à ce champ. Le Journal of Educational Data Mining (JEDM) [25].

1.3.3 Objectifs de l'EDM

L'objectif principal de l'EDM est de tirer parti des données éducatives pour améliorer la compréhension des processus d'apprentissage et optimiser l'enseignement. Cela peut inclure des tâches telles que la prédiction de la performance des étudiants, la détection de difficultés d'apprentissage, la recommandation de contenu d'apprentissage personnalisé, l'analyse de la qualité de l'enseignement, la prédiction de l'abandon scolaire etc. [2].

Parmi les objectifs, on peut citer [2, 31] :

- Utiliser les données collectées à partir de diverses sources, telles que les plateformes d'apprentissage en ligne, les systèmes de gestion de l'apprentissage et les dispositifs portables, pour découvrir des modèles, des tendances et des

relations cachées qui peuvent être utilisés pour prendre des décisions éclairées en matière d'éducation.

- Fournir des recommandations personnalisées aux étudiants.
- Évaluer et améliorer le matériel d'apprentissage en ligne.
- Fournir un retour d'information aux enseignants et aux étudiants.
- Détecter les comportements d'apprentissage atypiques.

1.3.4 Méthodes utilisées

Les méthodes utilisées en EDM incluent :

1. **Classification :** utilisation de techniques de classification pour prédire les performances des élèves ou leur appartenance à certaines catégories [31].
2. **Statistiques :** utilisation de méthodes statistiques pour analyser les données éducatives et extraire des informations significatives [31].
3. **Clustering :** utilisation de techniques de clustering pour regrouper les élèves en fonction de leurs caractéristiques ou de leur comportement d'apprentissage [4].
4. **Exploration de règles d'association :** utilisation de méthodes d'exploration de règles d'association pour identifier des schémas et des relations entre les variables dans les données éducatives [4].

1.3.5 Avantages et inconvénients

Parmi les avantages de l'EDM :

- **Amélioration de la prise de décision pédagogique :** L'EDM permet aux éducateurs de prendre des décisions éclairées en se basant sur des données objectives concernant les performances des élèves et les processus d'apprentissage [30].
- **Détection précoce des difficultés d'apprentissage :** L'EDM permet d'identifier rapidement les élèves qui rencontrent des difficultés d'apprentissage et de mettre en place des interventions précoces pour les aider à surmonter ces difficultés [33].
- **Optimisation des ressources éducatives :** L'EDM permet d'optimiser l'utilisation des ressources éducatives en identifiant les contenus et les méthodes pédagogiques les plus efficaces pour favoriser l'apprentissage des élèves [32].

- **Identification des facteurs d'influence :** L'EDM permet d'identifier les facteurs qui influencent les performances des élèves, tels que l'engagement, la motivation, les interactions sociales, et d'ajuster en conséquence les stratégies pédagogiques [32].
- **Développement de modèles prédictifs :** L'EDM permet de développer des modèles prédictifs qui peuvent anticiper les performances des élèves, ce qui peut aider les éducateurs à prendre des décisions éclairées sur les interventions nécessaires [30].

Parmi les inconvénients de l'EDM :

- **Éthique de l'utilisation des données :** L'utilisation des données d'apprentissage des élèves peut soulever des questions éthiques, telles que la transparence dans la collecte et l'utilisation des données, la consentement éclairé des participants et le partage des données entre les différentes parties prenantes [12].
- **Complexité technique :** L'implémentation de l'EDM peut être complexe sur le plan technique, nécessitant des compétences en analyse de données et en informatique, ce qui peut représenter un obstacle pour certains établissements scolaires ou enseignants [37].
- **Protection de la vie privée :** L'EDM implique la collecte et l'analyse de grandes quantités de données personnelles des élèves, ce qui soulève des préoccupations en matière de protection de la vie privée et de sécurité des données [37].
- **Effets néfastes sur l'autonomie de l'apprenant :** L'utilisation intensive de l'EDM peut entraîner une dépendance excessive aux données et à l'analyse, ce qui peut réduire l'autonomie de l'apprenant et limiter sa capacité à prendre des décisions éducatives informées [37].

1.3.6 Exemples d'utilisation de l'EDM

Dans l'EDM, des méthodes et des outils sont utilisés pour analyser les données d'apprentissage pour améliorer la compréhension de l'apprentissage des étudiants et l'efficacité des programmes d'enseignement.

Voici quelques exemples courants d'utilisation de l'EDM [32] :

Prédiction des performances scolaires :

L'EDM peut être utilisée pour prédire les performances scolaires des étudiants en se basant sur des indicateurs tels que les performances antérieures, les comportements d'apprentissage et les caractéristiques démographiques. Cela peut aider à identifier les étudiants à risque et à mettre en place des mesures de soutien appropriées.

Recommandation de ressources d'apprentissage personnalisées :

En analysant les données sur les préférences d'apprentissage, les performances antérieures et les caractéristiques des apprenants, l'EDM peut générer des recommandations personnalisées de ressources d'apprentissage, telles que des vidéos, des exercices ou des lectures, pour optimiser l'engagement et la réussite des étudiants.

Détection du désengagement des étudiants :

L'EDM peut être utilisée pour détecter les signes de désengagement des étudiants, tels que l'absentéisme fréquent, les performances médiocres ou les interactions limitées avec les ressources d'apprentissage. Cela permet aux enseignants d'intervenir rapidement et de fournir un soutien approprié pour prévenir l'abandon scolaire.

Personnalisation de l'apprentissage adaptatif :

L'EDM peut être utilisée pour personnaliser l'apprentissage en ligne en adaptant les ressources, les activités et les évaluations aux besoins individuels des apprenants. Cela permet de fournir un apprentissage plus efficace et engageant, en tenant compte des préférences, des capacités et du rythme d'apprentissage de chaque apprenant [32].

1.4 Comparaison entre LA et EDM

Les deux communautés ont des objectifs communs mais utilisent des approches différentes :

LA	**EDM**
Modélisation des données issus du système d'apprentissage	Automatisation de la prédiction des résultats d'un apprenant
Visualisation des données par les acteurs de l'apprentissage	Personnalisation de la stratégie d'apprentissage

Deux principes sont en action :

LA	**EDM**
Renvoyer les résultats de l'analyse aux acteurs du système (visualisation, tableaux de bord, graphiques, etc.) pour leur donner plus d'autonomie	Exploiter les résultats par le système d'apprentissage en automatisant la gestion de chacune de ses composantes principales (apprenants, enseignants, domaines enseignés)

1.5 Quelques travaux sur LA et EDM

Nous présentons, dans ce qui suit, quelques travaux sur LA et EDM :

1. L'article [29], présente une méthode pour mesurer la qualité de la collaboration en examinant les caractéristiques linguistiques du discours produit par les participants.

   Ils montrent que certaines caractéristiques du discours, telles que la richesse du vocabulaire et la complexité syntaxique, peuvent être utilisées pour prédire avec précision la qualité de la collaboration.

   Les résultats de l'étude suggèrent que l'analyse linguistique du discours peut être utilisée pour évaluer de manière objective la qualité de la collaboration et peut être utilisée pour améliorer les pratiques de collaboration.

2. Dans [14], on examine l'utilisation de Twitter en tant qu'outil pédagogique et développent un tableau de bord d'analyse pour mesurer l'engagement des étudiants sur la plateforme.

   Le tableau de bord permet de suivre les activités des étudiants sur Twitter, telles que le nombre de tweets publiés, les mentions, les hashtags utilisés, etc.

   Ils affirment que cet outil peut aider les enseignants à mieux comprendre comment les étudiants utilisent Twitter dans un contexte d'apprentissage et à améliorer la qualité de l'enseignement.

3. Dans [24], Alejandro *et al.* étudient les différentes techniques d'analyse des données utilisées dans l'éducation, telles que l'analyse des réseaux sociaux.

   Ils montrent comment les techniques d'analyse des données peuvent être utilisées pour améliorer les processus d'enseignement et d'apprentissage, en pointant les défis à relever, notamment en matière de qualité des données et de protection de la vie privée.

4. Dans [11], on fournit un aperçu détaillé des travaux de recherche dans le domaine de l'EDM et met en évidence les tendances et les défis associés à son utilisation. Il offre une base solide pour de futures recherches dans ce domaine et souligne l'importance croissante de l'EDM dans l'amélioration des processus éducatifs.

On présente dans le tableau suivant un récapitulatif des études présentées :

Article	Objectif	Méthode	Résultat	LA / EDM
[29]	Mesurer la qualité de la collaboration dans la résolution de problèmes en utilisant l'analyse linguistique du discours produit par les participants.	Analyser les transcriptions verbales d'interactions pour extraire des caractéristiques telles que la richesse du vocabulaire, la complexité syntaxique, et le nombre de mots prononcés par participant.	Les résultats de l'étude montrent que certaines caractéristiques du discours produit par les participants, telles que la richesse du vocabulaire et la complexité syntaxique, peuvent être utilisées pour prédire avec précision la qualité de la collaboration.	EDM
[14]	Développer un tableau de bord d'analyse pour mesurer l'engagement des étudiants sur Twitter en tant qu'outil pédagogique.	Collecter les données de Twitter des étudiants et les analyser en utilisant des algorithmes de traitement du langage naturel pour produire des indicateurs tels que nombre de tweets publiés, les mentions, les hashtags utilisés, etc.	Les résultats montrent que le tableau de bord peut être un moyen efficace pour aider les enseignants à mieux comprendre l'utilisation de Twitter par les étudiants et à mesurer l'impact de l'utilisation de Twitter sur l'engagement des étudiants	LA
[24]	Méta-analyse de travaux sur les LA et l'EDM afin de mettre en évidence les tendances et les défis liés à l'utilisation de ces approches pour améliorer les processus d'enseignement et d'apprentissage.	Examiner les différentes techniques d'analyse des données utilisées dans l'éducation	Les résultats montrent que l'analyse des données en éducation est un domaine en croissance qui peut offrir de nombreux avantages pour améliorer les processus d'enseignement et d'apprentissage.	LA et EDM

[11]	Fournir une vue d'ensemble complète des travaux de recherche dans le domaine de l'EDM et de mettre en évidence les tendances et les défis associés à son utilisation pour améliorer les processus d'enseignement et d'apprentissage.	Examiner et analyser les travaux en termes de thèmes, de méthodes utilisées, de types de données, de domaines d'application et de résultats obtenus.	Les études ont montré que l'EDM permet de détecter les difficultés des apprenants, de prédire leur réussite et de proposer des interventions ciblées pour améliorer leur parcours éducatif.	EDM

TABLE 1.3 – Tableau récapitulatif de quelques travaux sur LA et EDM

1.6 Conclusion

les Learning Analytics et l'Educational Data Mining offrent des opportunités prometteuses pour améliorer les processus d'enseignement et d'apprentissage. Ils permettent de collecter et d'analyser des données pertinentes, d'extraire des connaissances utiles et de prendre des décisions éclairées en matière d'éducation.

Dans le chapitre suivant, nous allons présenter la classification en tant qu'outil essentiel utilisé dans les domaines des Learning Analytics et de l'Educational Data Mining pour analyser les données éducatives.

Chapitre 2

Classification Multi-label

2.1 Introduction

L'apprentissage automatique est une branche de l'informatique qui permet à des systèmes informatiques d'apprendre à partir de données, sans être explicitement programmés.

La classification est l'une des tâches les plus courantes en apprentissage automatique, qui consiste à prédire la classe d'un objet en fonction de ses caractéristiques. Il existe différents types de classifications, chacun avec ses propres avantages et inconvénients. Parmi ces différents types, on distingue la classification multi-label, qui permet de prédire plusieurs classes pour chaque objet.

Dans ce chapitre, nous étudierons les différentes techniques de classification, leurs avantages et leurs inconvénients. Nous examinerons les différentes approches de classification multi-label et enfin nous présenterons des travaux de recherche dans le domaine.

2.2 Définition

La classification est un processus permettant de regrouper des éléments en catégories ou classes en fonction de leurs caractéristiques communes. Elle est largement utilisée dans de nombreux domaines tels que la reconnaissance de formes, la détection d'anomalies, la segmentation d'images, et la prédiction de résultats. En utilisant des techniques de classification, il est possible de prédire la classe d'un élément donné en fonction de ses caractéristiques [7].

2.3 Les différents types de classification

Il existe différents types de classification :

2.3.1 La classification binaire

La classification binaire est un type de problème d'apprentissage automatique supervisé qui nécessite de classer les données en deux groupes ou catégories mu-

tuellement exclusifs. Les deux groupes peuvent être représentés par 0 et 1, positif et négatif, ou vrai et faux. Un modèle de classification binaire est formé sur un ensemble de données étiqueté avec le résultat souhaité. Le modèle apprend ensuite à prédire les étiquettes pour les nouveaux points de données. La classification binaire peut être utilisée pour diverses applications telles que : détection de spam, détection de fraude et diagnostic médical. Par exemple, vous pouvez former un modèle de classification binaire pour reconnaître si un e-mail est un spam.

Le filtrage des spams est un problème typique de classification binaire. Le classificateur apprend à partir du texte du message et prédit s'il s'agit d'un spam ou non [7].

L'image suivante représente un classificateur de filtrage des spams selon la classification binaire.

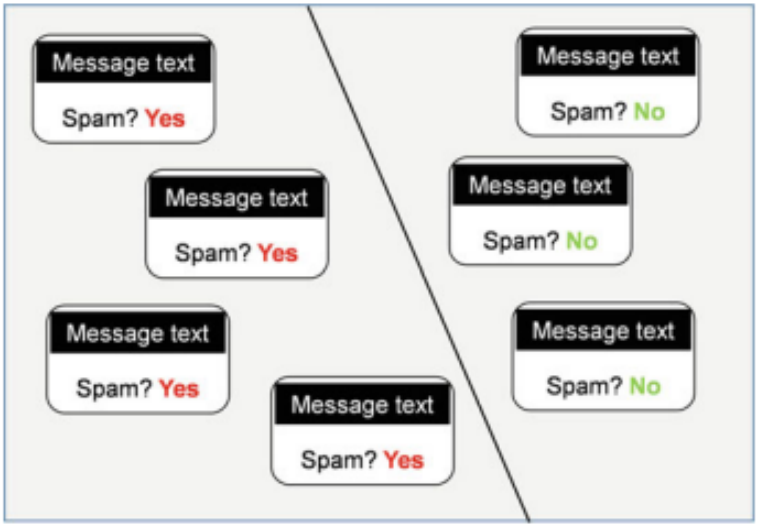

FIGURE 2.1 – La classification binaire [7]

Un autre exemple de classificateur binaire prédit si des images représentent des chiens ou des chats [16]. L'image suivante représente un classificateur de réseau neuronal qui classe une image comme représentant un chien ou un chat.

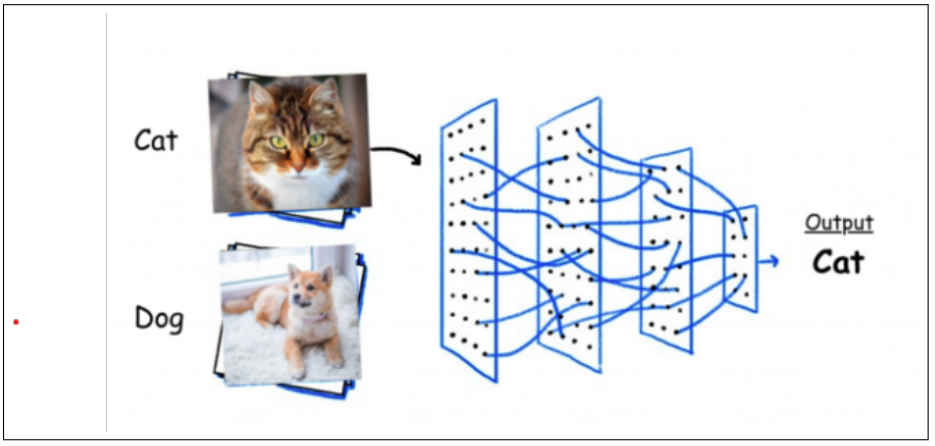

FIGURE 2.2 – La classification binaire [16]

Les applications de la classification binaire sont : le filtrage du courrier électronique, qui permet d'éliminer les messages non sollicités, l'analyse des prêts, pour déterminer si le client est économiquement fiable ou non, l'évaluation médicale, pour déterminer si un patient est atteint d'une certaine maladie ou non, et la reconnaissance de toutes sortes de motifs binaires [7].

2.3.2 La Classification multi-classe

La classification multi-classe est un type de problème d'apprentissage automatique supervisé qui nécessite de classer les données en trois catégories ou plus.

La classification multi-classe peut être considérée comme une généralisation de la classification binaire. Il n'y a qu'une seule sortie, mais elle peut prendre n'importe quelle valeur, alors que dans le cas binaire, elle est limitée à un sous-ensemble de deux valeurs [7].

Par exemple, on peut utiliser un classificateur multi-classe pour classer les images des fruits dans différentes catégories telles que les bananes, les oranges et les pommes. Le modèle apprend à identifier les traits spécifiques associés à chaque catégorie de fruit. Une fois le modèle formé, il peut être utilisé pour classer de nouvelles images dans la bonne catégorie des fruits.

Les algorithmes d'apprentissage automatique qui peuvent être utilisés pour la classification multi-classe comprennent la régression logistique multinomiale et les réseaux de neurones [16].

L'image suivante représente un classificateur de fruits selon la classification multi-classe.

FIGURE 2.3 – La classification multi-classe [16]

La catégorisation des espèces d'iris est un problème classique de classification multi-classe. Le classificateur apprend à partir de la longueur et de la largeur du pétale et du sépale, et prédit à laquelle des trois espèces la fleur appartient [7].

L'image suivante représente un classificateur de catégorisation des espèces d'iris selon la classification multi-classe.

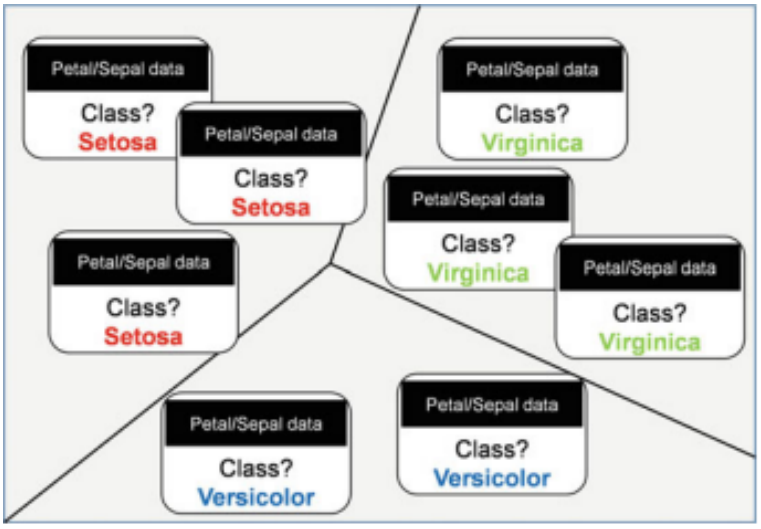

FIGURE 2.4 – La classification multi-classe [7]

2.3.3 La classification multi-label

La classification multi-label est une extension du problème de classification mono-label. Il consiste à associer un objet décrit par un vecteur de variables avec un sous-ensemble restreint de concepts d'intérêt commun, connu sous le nom d'*étiquettes*. Chaque élément du vecteur sera une valeur binaire, indiquant si l'étiquette correspondante est pertinente pour l'échantillon ou non. Plusieurs étiquettes peuvent être actives simultanément.

La classification multi-label est actuellement appliquée dans de nombreux domaines, la plupart d'entre eux étant liés à l'étiquetage automatique des ressources de médias sociaux, telles que les images, la musique, les vidéos, les actualités et les billets de blog, la vision par ordinateur, l'analyse de texte, la biologie, les systèmes de recommandation, etc. [38, 7].

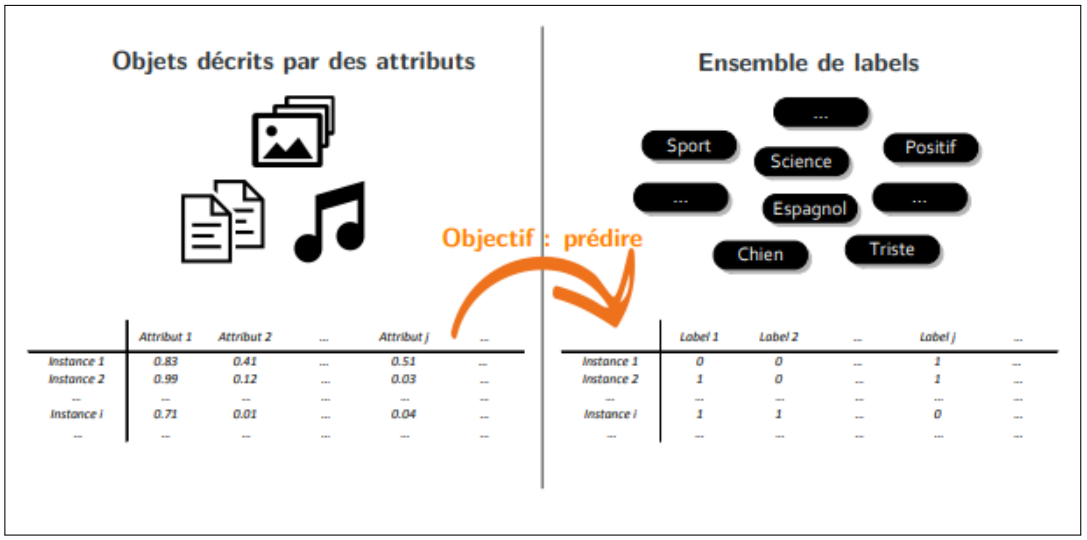

FIGURE 2.5 – Classification multi-label [38]

La figure suivantes illustre l'une des applications multi-label classiques, l'étiquetage d'images. L'ensemble de données comporte quatre étiquettes au total et chaque image peut se voir attribuer l'une ou plusieurs d'entre elles [7].

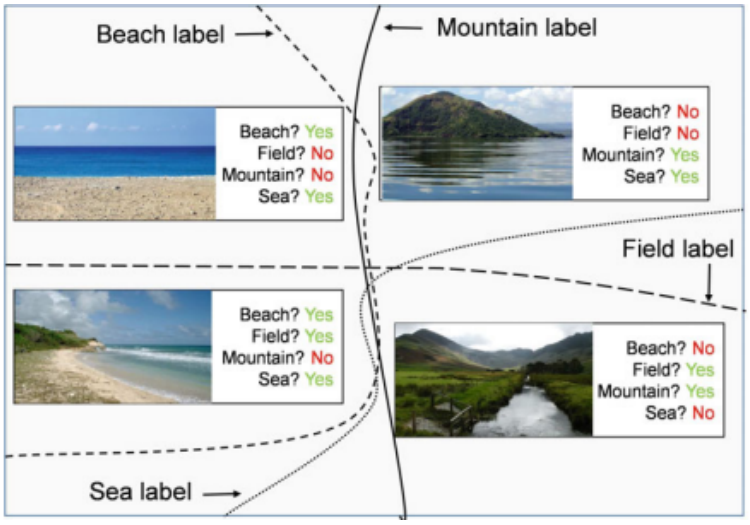

FIGURE 2.6 – Exemple de classification multi-label [7]

La classification est une technique essentielle en apprentissage automatique, où les éléments sont regroupés en différentes catégories en fonction de leurs caractéristiques. la figure suivante représentant les principaux types de classification :

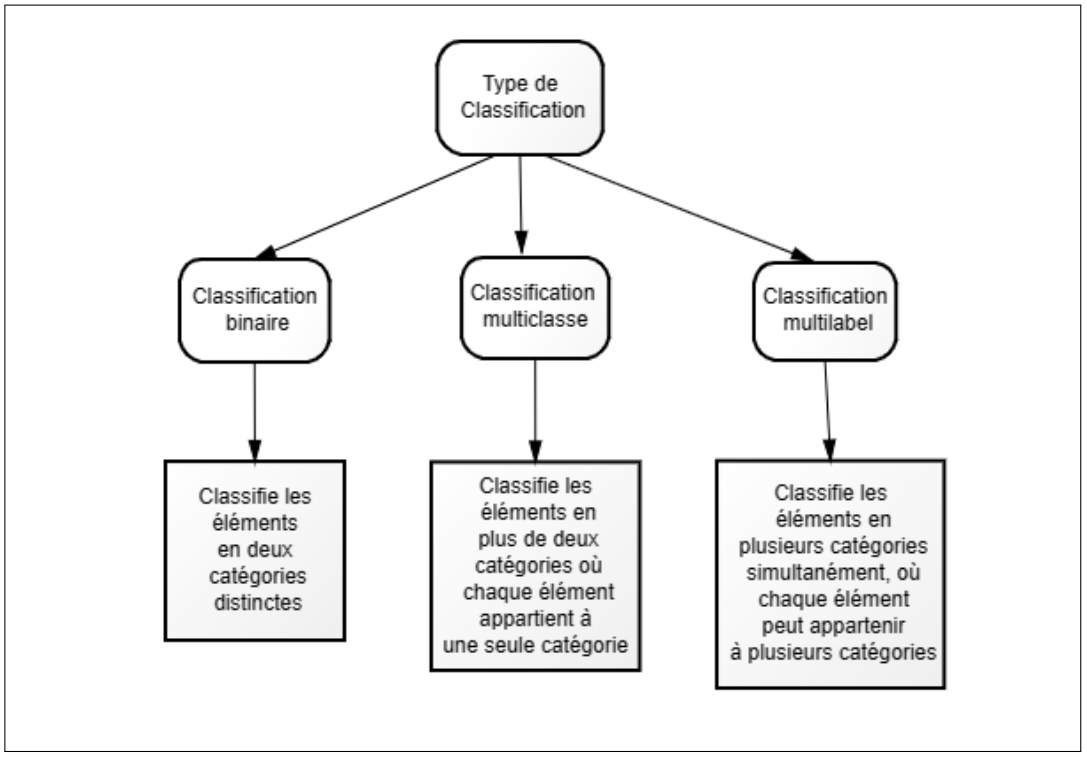

FIGURE 2.7 – Types de classification

2.4 La classification multi-label

La classification multi-label est un type d'algorithme d'apprentissage automatique supervisé qui peut être utilisé pour attribuer zéro ou plusieurs étiquettes à n'importe quel échantillon de données [7].

Par exemple, un document sur la classification automatique peut être associé à plusieurs rubriques, tel que : "Apprentissage automatique", "Science des données", "Statistiques", "Langages de programmation", "Python". Ces concepts sont interdépendants, de sorte qu'un article typique peut être associé à 5 à 6 rubriques [16].

2.4.1 Les avantages et les inconvénients de la classification multi-label

La classification multi-label a plusieurs avantages et quelques inconvénients.

a. Avantages :

Les avantages de la classification multi-label incluent [41, 45, 44] :

- Elle permet de traiter des données qui ont plusieurs caractéristiques ou dimensions simultanément.
- Elle est utile pour les tâches de reconnaissance d'images ou de vidéos, où une seule image ou vidéo peut contenir plusieurs objets ou scènes qui doivent être reconnus.
- Elle permet d'utiliser des données qui sont riches en étiquettes, car elle ne nécessite pas de choisir une seule étiquette pour chaque observation.
- Elle peut gérer une large gamme de problèmes de classification, y compris ceux à plusieurs classes.
- Elle peut gérer les relations entre les classes, ce qui est important dans certaines applications, telles que la classification d'images.
- Elle peut améliorer la précision de la classification en fournissant plus d'informations sur les classes associées à chaque exemple.

b. Inconvénient :

Les inconvénients de la classification multi-label incluent [41, 45, 44] :

- Il est plus difficile de trouver des modèles efficaces pour résoudre des problèmes de classification multi-label, car ils ont tendance à être plus complexes que les modèles de classification mono-label.
- Il est plus difficile d'évaluer les performances des modèles de classification multi-label, car il existe plusieurs façons de mesurer la précision, le rappel et le score F1. Cette diversité d'évaluations peut rendre l'interprétation des

résultats plus complexe et pose un défi pour la comparaison des performances entre différents modèles.

- Il est plus difficile de généraliser les résultats des modèles de classification multi-label à de nouvelles données, car ils ont tendance à être plus sensibles aux variations des données d'entraînement.
- La complexité du modèle peut augmenter considérablement avec un nombre élevé d'étiquettes.
- Les algorithmes de classification multi-label peuvent être plus difficiles à évaluer et à optimiser.
- Les algorithmes de classification multi-label peuvent être plus coûteux en termes de temps de calcul et de mémoire.

2.4.2 Méthodes de classification multi-label

Les méthodes de classification multi-label peuvent être classées en trois grandes familles.

- **Méthodes par transformation du problème :**

  Elles consistent à transformer le problème de classification multi-label en un ou plusieurs problèmes de classification ou de régression mono-label. Cette méthode vise à diviser le problème initial en sous-problèmes plus simples, où chaque sous-problème est traité comme une tâche de classification ou de régression distincte. Cette approche de transformation des problèmes permet de simplifier la tâche du classificateur et d'exploiter des modèles plus simples et mieux adaptés aux problèmes mono-label [7, 9].

- **Méthodes par adaptation du problème :**

  La classification multi-label est une tâche complexe où le classificateur doit prédire plusieurs résultats simultanément. Les algorithmes de classification mono-label sont alors adaptés aux données multi-label. Différents modèles de classification, tels que les SVM, les arbres et les réseaux neuronaux, ont été adaptés pour traiter les problèmes multi-label. Les modèles doivent être capables de prédire plusieurs résultats à la fois, ce qui présente des défis spécifiques pour chaque approche [7, 9].

- **Méthodes hybrides :**

  Elles consistent à regrouper un ensemble de classificateurs dont les résultats sont combinés généralement par une moyenne pondérée ou non pondérée. L'objectif est de tirer parti des forces de différents classificateurs pour obtenir de meilleures performances qu'avec un seul classificateur. L'un des aspects-clés de la construction d'ensembles est la diversité, car des classificateurs individuels plus diversifiés ont tendance à avoir des biais différents. Les ensembles de classificateurs ont été largement utilisés pour la classification multi-classe et sont

également couramment appliqués à la classification multi-label pour relever les défis tels que le déséquilibre des classes [7, 9].

La figure 2.8 présente un récapitulatif des méthodes de classification multi-label.

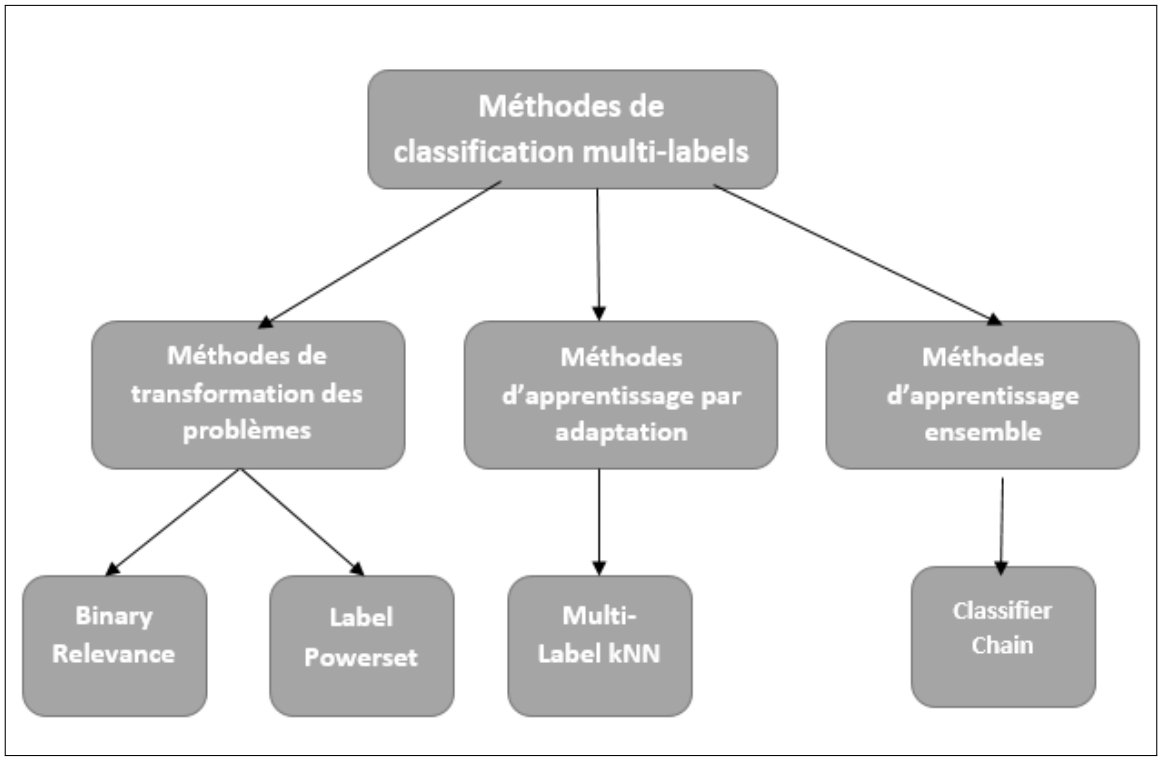

FIGURE 2.8 – Les méthodes de classification multi-label

a. Méthodes par transformation du problème

Binary Relevance : Pour la pertinence binaire, un ensemble de classificateurs binaires simplement étiquetés est formé séparément sur l'ensemble de données d'origine pour prédire l'appartenance à chaque classe [9] (*cf.* figure 2.9).

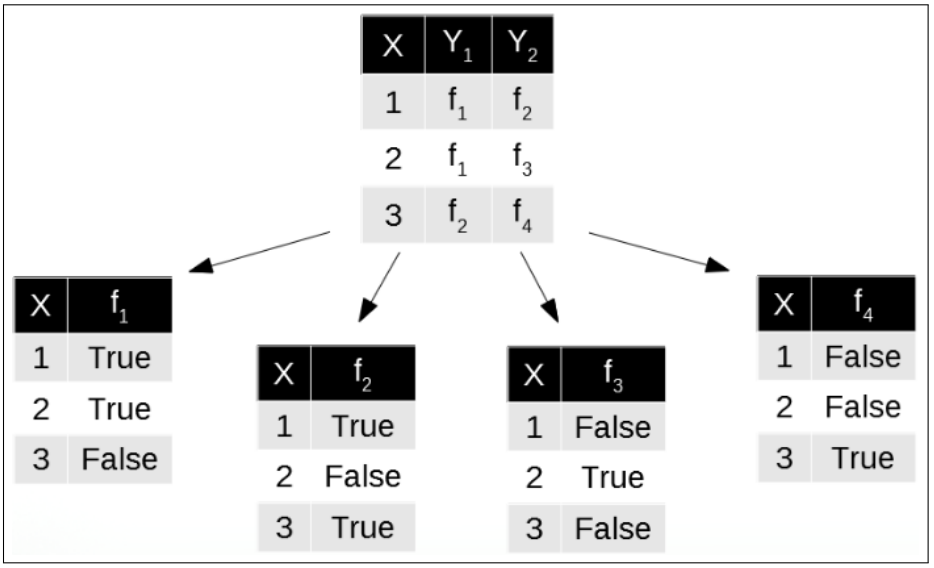

FIGURE 2.9 – Binary Relevance [9]

Label Powerset (LP) : Elle convertit un problème d'apprentissage multi-label en un seul problème d'apprentissage multi-classe à étiquette unique. Toute combinaison d'étiquettes présentes dans l'ensemble d'apprentissage sont transformés en classes, puis on forme un classificateur multi-classe [9, 21] (*cf.* figure 2.10).

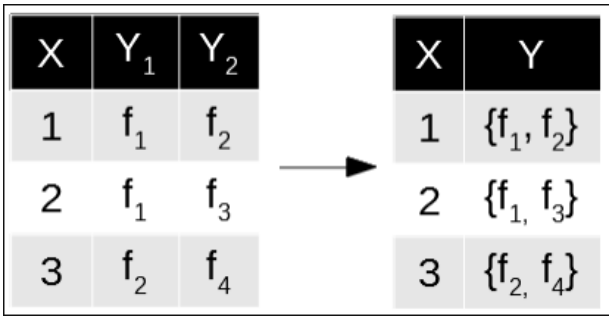

FIGURE 2.10 – Label Powerset [9]

b. Méthodes par adaptation du problème

Multi-Label kNN (ML-kNN) : L'algorithme ML-kNN est une méthode de classification multi-label qui se base sur l'algorithme des k plus proches voisins (kNN). Son fonctionnement repose sur le calcul d'un ensemble de probabilités *a priori* et conditionnelles pour chaque étiquette. Grâce à sa simplicité et à sa faible complexité informatique, ML-kNN est fréquemment utilisé dans les études expérimentales. Il sert également de base pour le développement d'autres algorithmes plus avancés en classification multi-label, tels que IBLR-ML. [7].

Dans ML-kNN, pour chaque instance de l'ensemble de test, ses k plus proches voisins dans l'ensemble de formation sont identifiés. Pour chaque classe y dans Y, le nombre d'instances voisines appartenant à y est utilisé pour calculer les probabilités postérieures que l'instance de test appartient ou n'appartient pas à y. En fonction

de la plus grande de ces probabilités, on décide d'attribuer ou non la classe y à une instance de test [9].

Pour une meilleure compréhension, nous présentons un exemple tiré de [9]. Supposons que les échantillons ci-dessous sont des voisins trouvés dans l'ensemble d'apprentissage pour l'instance de test $X = 1$ (figure 2.11).

X	Y_1	Y_2
1	f_1	f_2
2	f_1	f_3
3	f_2	f_4
1	?	?

FIGURE 2.11 – Les voisins trouvés dans l'ensemble d'apprentissage [9]

Nous devons maintenant calculer les probabilités *a priori* et *a posteriori* pour chaque classe de l'instance de test $X = 1$ (figure 2.12).

$$P(f_1|X = 1) = P(f_1) \cdot P(X=1|f_1) = 2/3 \cdot ½ = 1/3$$
$$P(f_2|X = 1) = P(f_2) \cdot P(X=1|f_2) = 2/3 \cdot ½ = 1/3$$
$$P(f_3|X = 1) = P(f_3) \cdot P(X=1|f_3) = 1/3 \cdot 0 = 0$$
$$P(f_4|X = 1) = P(f_4) \cdot P(X=1|f_4) = 1/3 \cdot 0 = 0$$

FIGURE 2.12 – Calculer les probabilités *a priori* et *a posteriori* pour chaque classe [9]

Des valeurs maximales ont été obtenues pour f1 et f2. Par conséquent, nous assignons l'instance de test à ces deux classes.

L'algorithme de la méthode ML-kNN sera présenté en détail dans la section 3.6.1.

c. Méthodes hybrides

Classifier Chain (CC) : Le modèle de chaîne de classificateurs apprend des classificateurs comme dans la méthode de pertinence binaire. Cependant, tous les classificateurs sont liés dans une chaîne, comme illustré dans la figure [9].

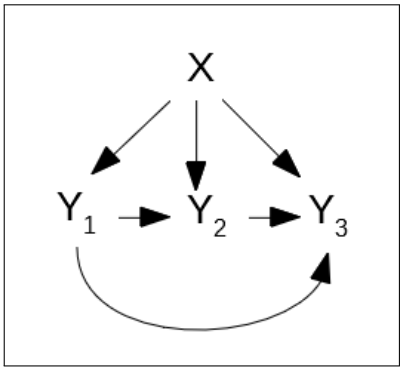

FIGURE 2.13 – Classifier Chain [9]

- Tout d'abord, toutes les caractéristiques $(X1, X2, ..., Xm)$ sont utilisées pour prédire $y1$.
- Ensuite, toutes les caractéristiques $(X1, X2, ..., Xm, y1)$ sont utilisées pour prédire $y2$.
- Enfin, $(X1, X2, ..., Xm, y1, y2)$ sont utilisées pour prédire $y3$.

L'ordre dans lequel les étiquettes sont prédites a un impact important sur les résultats [9].

les avantages et les inconvénients de chaque approche

Le tableau ci-dessous présente un récapitulatif des avantages et inconvénients de chaque approche [9] :

Modèle de classification	**Approche**	**Avantages**	**Inconvénients**
Méthodes de transformation	Binary Relevance (BR)	- complexité linéaire - simplicité	- ignore la corrélation entre les étiquettes en traitant chaque étiquette de manière indépendante
	Label Powerset	- prise en compte des corrélations entre les étiquettes	- complexité de calcul élevée - peut conduire à un ensemble de données déséquilibrées en tenant compte de plusieurs classes associées à peu d'exemples

Méthodes d'apprentissage par adaptation	Multi-Label kNN (ML-kNN)	- une meilleure précision que les méthodes précédentes - les corrélations entre les étiquettes sont considérées	- avec la stratégie de pertinence binaire, l'estimation du ratio peut ne pas être précise - les distributions de données pour certains labels sont déséquilibrées
Méthodes d'apprentissage ensemble	Classifier Chain (CC)	- prise en compte de la corrélation des étiquettes - une complexité de calcul acceptable	- la précision dépend fortement de l'ordre - pour n étiquettes, il y a $n!$ ordres possibles

TABLE 2.1 – Récapitulatif des avantages et des inconvénients des approches de classification multi-label

2.4.3 Applications de la classification multi-label

Cette section décrit certains cas d'utilisation de la classification multi-label [7].

Catégorisation du texte

La classification multi-label trouve son origine dans la nécessité de classer les documents textuels en plusieurs catégories non exclusives. En raison de la large gamme de documents textuels disponibles, tels que les accords, les rapports, les factures, les courriels, les livres, les magazines, les articles en ligne, etc., il est souvent nécessaire de les classer dans plusieurs catégories. C'est pourquoi il existe de nombreuses publications sur la classification multi-label.

Pour transformer un ensemble de documents textuels en un ensemble de données multi-étiquettes (MLD), il est courant d'utiliser des techniques d'exploration de texte. Les documents sources sont analysés, les mots non informatifs sont éliminés, et des vecteurs représentant l'occurrence de chaque mot parmi les documents sont calculés. Ainsi, chaque document est représenté par une ligne dans le MLD, où les colonnes correspondent aux mots et à leurs fréquences.

La classification multi-label offre une approche efficace pour organiser et classer ces documents textuels qui peuvent appartenir à plusieurs catégories simultanément. Cette méthode est largement utilisée dans divers domaines où la classification précise des documents est essentielle pour faciliter la recherche, l'organisation et l'analyse des informations contenues dans ces textes.

Étiquetage des ressources multimédias

Les documents en texte brut étaient les plus courants dans le passé, mais aujourd'hui, avec la croissance fulgurante des technologies de stockage et de com-

munication, les images, les vidéos et la musique sont des ressources courantes. La classification multi-label est utilisée pour étiqueter tous ces types de ressources afin d'identifier les objets qui apparaissent dans une séquence d'images, les émotions générées par une vidéo musicale ou les concepts pouvant être dérivés d'un clip vidéo. De cette manière, un grand nombre de nouvelles ressources peuvent être correctement triées, mais cela est très coûteux.

Génétique/Biologie

La protéomique et la génomique sont des domaines de recherche qui ont connu un essor important ces dernières années. La technologie de Data Mining peut accélérer les processus et réduire les coûts.

L'application de la classification multi-label dans ce domaine vise à prédire les fonctions biologiques des gènes. Les propriétés utilisées comme prédicteurs sont généralement des modèles de protéines et des caractéristiques structurelles internes.

Systèmes de recommandation

Les systèmes de recommandation conscients du contexte sont des extensions des systèmes de recommandation traditionnels qui prennent également en compte les conditions contextuelles de l'utilisateur auquel une recommandation est faite [46].

Les systèmes de recommandation constituent un domaine d'application direct pour la classification multi-label. Les données collectées par les systèmes de recommandation se caractérisent par une augmentation constante et certaines valeurs manquantes [17].

2.4.4 Quelques travaux sur la classification multi-label

Nous présentons, dans ce qui suit, quelques travaux sur la classification multi-label.

1. Dans [5], Bolaño *et al.* ont proposé une technique de reconnaissance d'ingrédients alimentaires basée sur l'apprentissage multi-label.

   Cette technique se base sur l'utilisation de modèles d'apprentissage automatique formés à reconnaître les différents ingrédients présents dans une image. Ces modèles sont alimentés avec de grandes quantités de données d'images étiquetées contenant des exemples d'ingrédients alimentaires clairement identifiables.

   Une fois formés, ces modèles peuvent être utilisés pour analyser de nouvelles images et détecter les ingrédients présents. La technique de reconnaissance d'ingrédients alimentaires par l'apprentissage multi-label convient particulièrement aux cas où une image peut contenir plusieurs ingrédients simultanément, ce qui est souvent le cas dans les aliments préparés.

2. Dans [43], Hong-Xing Yu *et al.* ont proposé une technique de reconnaissance de personnes basée sur l'apprentissage automatique non supervisé.

Cela vise à identifier les personnes dans les images ou les vidéos à l'aide de modèles d'apprentissage automatique formés sur des données non-étiquetées. Ces modèles peuvent apprendre les caractéristiques distinctives des personnes dans les images, telles que leurs visages, leurs corps ou leurs vêtements.

3. Dans [8], Elijah Cole *et al.* proposent une méthode d'apprentissage automatique qui permet de prédire plusieurs étiquettes pour une entrée donnée en utilisant uniquement une étiquette positive.

   La méthode repose sur le principe de traiter les étiquettes manquantes, c'est-à-dire les étiquettes qui ne sont pas présentes dans les données d'entraînement. Pour cela, des techniques de complétion de données sont utilisées, où les informations des étiquettes présentes sont exploitées pour prédire les étiquettes manquantes.

   Plus précisément, la méthode consiste à construire un modèle qui apprend à représenter les relations entre les étiquettes positives et les caractéristiques de l'entrée. En utilisant ces relations, le modèle peut prédire les étiquettes manquantes pour de nouvelles entrées.

   Cette approche est particulièrement utile dans les situations où il est difficile ou coûteux d'obtenir des étiquettes négatives, et où les données d'entraînement sont limitées en termes de couverture des différentes étiquettes possibles.

4. Dans [47], Zhou *et al.* ont examiné l'utilisation de modèles de classification à étiquettes multiples pour le diagnostic des complications diabétiques et présentent l'état actuel du domaine, mettant en lumière les défis et les limites des méthodes existantes et suggèrent des voies possibles pour la recherche future. L'article est pertinent pour les personnes étudiant ou travaillant dans le domaine du diagnostic médical et de l'apprentissage automatique.

Article	Objectif	Méthode	Résultat
[5]	Développer une méthode pour identifier les ingrédients dans les images d'aliments.	Utiliser l'apprentissage multi-label pour résoudre ce problème, en s'appuyant sur des algorithmes de reconnaissance d'images et de traitement du langage naturel.	Les résultats de cette approche sont comparés aux méthodes existantes, montrant que l'utilisation de l'apprentissage multi-label peut améliorer la précision de la reconnaissance des ingrédients.

[43]	Développer une méthode pour ré-identifier les personnes dans des images sans utiliser d'informations d'étiquetage.	Utiliser un apprentissage non-supervisé et des techniques de traitement d'images.	Les résultats de cette approche sont comparés aux méthodes existantes, montrant que l'utilisation de l'apprentissage multi-label améliore la précision de la ré-identification des personnes.
[8]	Développer une méthode pour effectuer de l'apprentissage à étiquettes multiples dans des environnements où seules quelques étiquettes sont disponibles.	Utiliser des algorithmes d'apprentissage semi-supervisé pour résoudre ce problème, en s'appuyant sur une étiquette positive unique pour chaque élément.	Les résultats de cette approche sont comparés aux méthodes existantes, montrant que l'utilisation d'étiquettes positives uniques peut améliorer les performances de l'apprentissage à étiquettes multiples dans des environnements limités.
[47]	Développer une méthode pour diagnostiquer les complications diabétiques à partir de données médicales.	Utiliser des modèles de classification à étiquettes multiples pour aborder ce problème, en s'appuyant sur des algorithmes de traitement de données médicales.	Les résultats de cette approche sont comparés aux méthodes existantes, montrant que l'utilisation de modèles de classification à étiquettes multiples peut améliorer la précision du diagnostic des complications diabétiques.

TABLE 2.2 – Récapitulatifs de quelques travaux sur la classification multi-label

2.5 Conclusion

La classification est une tâche cruciale dans l'apprentissage automatique, qui peut être appliquée dans de nombreux domaines, tels que la reconnaissance d'images, la détection de fraudes, la catégorisation de textes, et bien plus encore. Selon la nature des données et le nombre de classes à prédire, il existe différents types de classifications, y compris la classification binaire, la classification multi-classe et la classification multi-label. Chaque type de classification a ses propres avantages et inconvénients, ainsi que des techniques et des métriques d'évaluation appropriées.

Nous avons exploré différentes techniques et méthodes de classification multi-label, en mettant en évidence leurs avantages et leurs inconvénients. De plus, nous avons examiné quelques une de ces applications courantes. Enfin, nous avons présenté quelques travaux de recherche récents dans ce domaine.

Nous présentons dans le chapitre suivant la modélisation que nous avons proposée pour le problème de la classification des apprenants.

Chapitre 3

Conception

3.1 Introduction

Dans ce chapitre, nous présentons la conception générale de notre système. Nous commençons par définir les objectifs du projet, avant d'exposer les caractéristiques-clés que nous avons identifiées pour la classification des apprenants. Nous poursuivons en présentant les deux méthodes de regroupement intégrées dans notre système.

3.2 Objectifs

L'objectif principal de ce projet est de développer un système basé sur la classification des apprenants, pour leur regroupement automatique, à partir d'indicateurs d'évaluation de leur apprentissage.

Les buts de notre travail sont résumés comme suit :

- Pouvoir classer les apprenants lors d'un travail collaboratif selon un certain nombre d'indicateurs. Il s'agit de suivre les interactions des apprenants pendant leurs tâches d'apprentissage ou d'évaluation pour pouvoir extraire des données sur leur implication et leurs apports au groupe. Ce qui nous permettra de les classer en vue d'un regroupement automatique des apprenants.
- Former des groupes hétérogènes pour améliorer la qualité de la collaboration entre les membres du groupe et ainsi obtenir de meilleurs résultats.
- Proposer deux méthodes de regroupement automatique d'apprenants basées sur la classification multi-label et un algorithme génétique et comparer ces deux méthodes.

Ces objectifs doivent être réalisé dans le cadre d'un réseau social d'apprentissage, conçu et réalisé dans [19].

3.3 Conception du système

Dans le contexte d'un réseau social d'apprentissage (Social Learning Network - SLN), le système est alimenté par des enseignants et des étudiants en tant qu'entrées.

Les informations fournies par ces utilisateurs, telles que les identifiants d'accès, les notes, les sentiments, etc., sont utilisées comme données d'entrée pour les algorithmes ML-kNN et AG. Ces algorithmes analysent ces données et génèrent des groupes en sortie. Les groupes résultants sont basés sur des indicateurs d'activité des étudiants. L'objectif de ce processus est de regrouper les utilisateurs de manière efficace et pertinente, en facilitant la communication, la collaboration et l'interaction entre les membres des différents groupes formés.

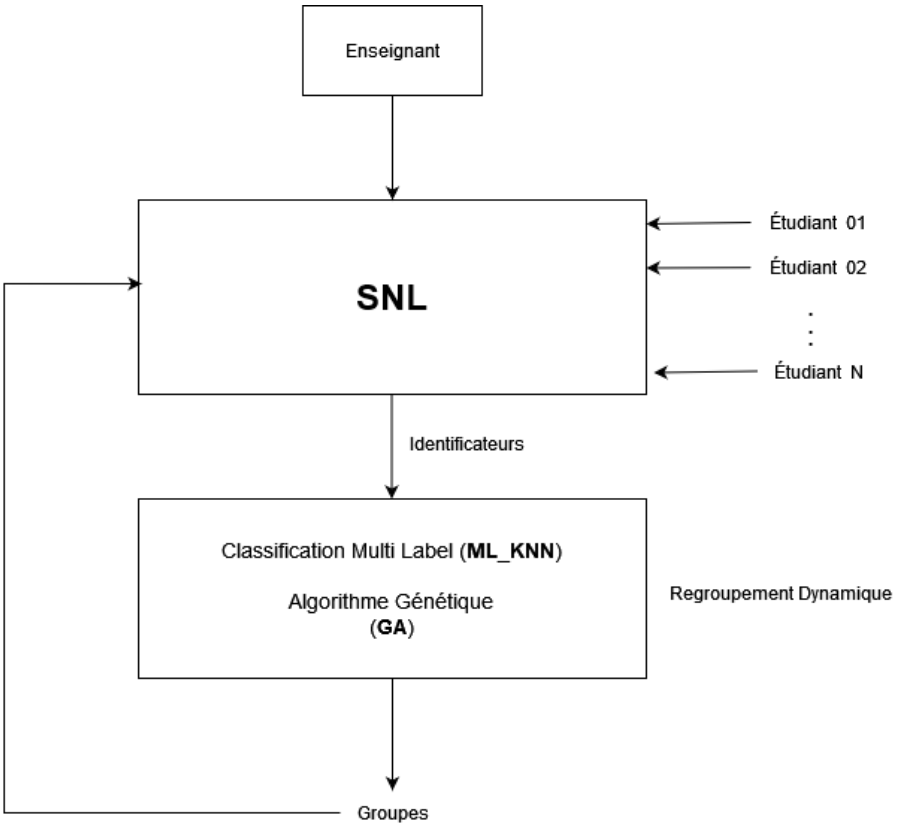

FIGURE 3.1 – Conception générale du système

3.4 Le réseau social d'apprentissage

Le réseau social d'apprentissage est conçu pour les apprenants afin qu'ils puissent partager, aimer et commenter des publications. Ils peuvent également publier, supprimer et partager leurs propres publications. Les étudiants peuvent également communiquer entre eux en utilisant des messages privés [19].

La figure ci-dessous présente la conception générale du système [19] :

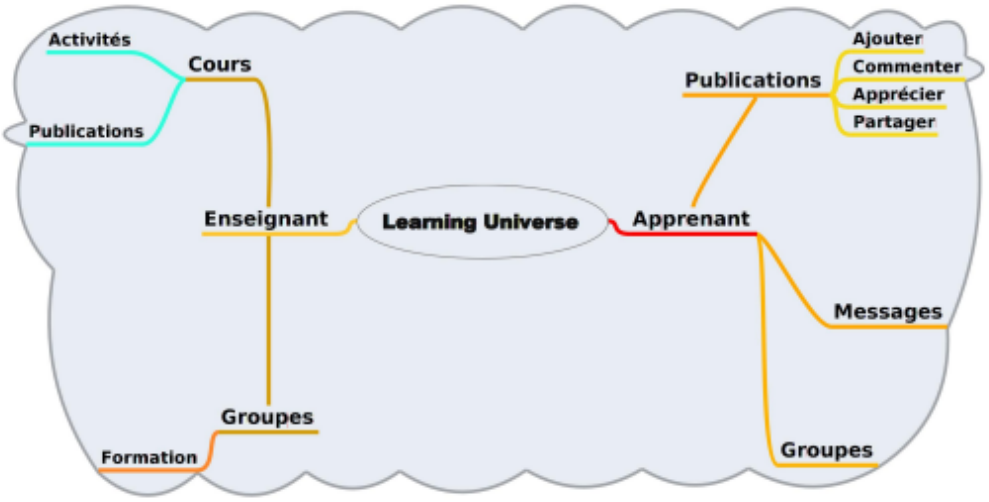

FIGURE 3.2 – Conception générale du SLN [19]

3.4.1 Fonctionnalités

Le système a deux acteurs principaux :

- **L'enseignant :** est responsable de la réalisation des cours et des activités. Il a également un rôle d'animateur sur le système, en publiant, commentant, précisant et partageant des publications pour encourager l'interaction et la collaboration entre les apprenants.
- **L'apprenant :** suit le cours et réalise les activités proposées par l'enseignant. Il doit résoudre ces activités en travaillant de manière collaborative avec les autres apprenants. En plus de cela, l'apprenant peut publier, commenter, préciser et partager des publications avec les autres apprenants.

3.4.2 Diagramme de cas d'utilisation

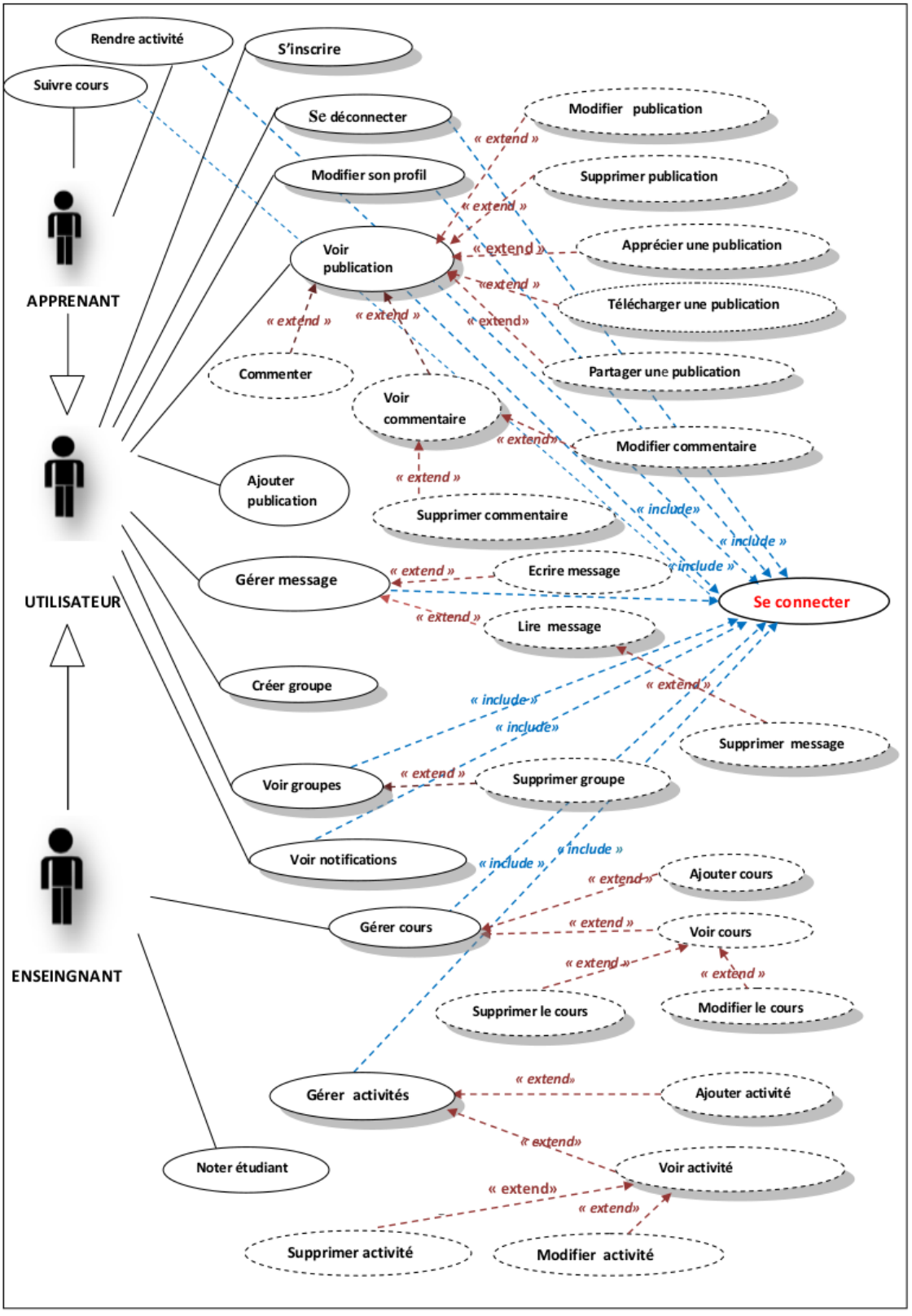

FIGURE 3.3 – Diagramme de cas d'utilisation [19]

3.4.3 Diagramme de classe

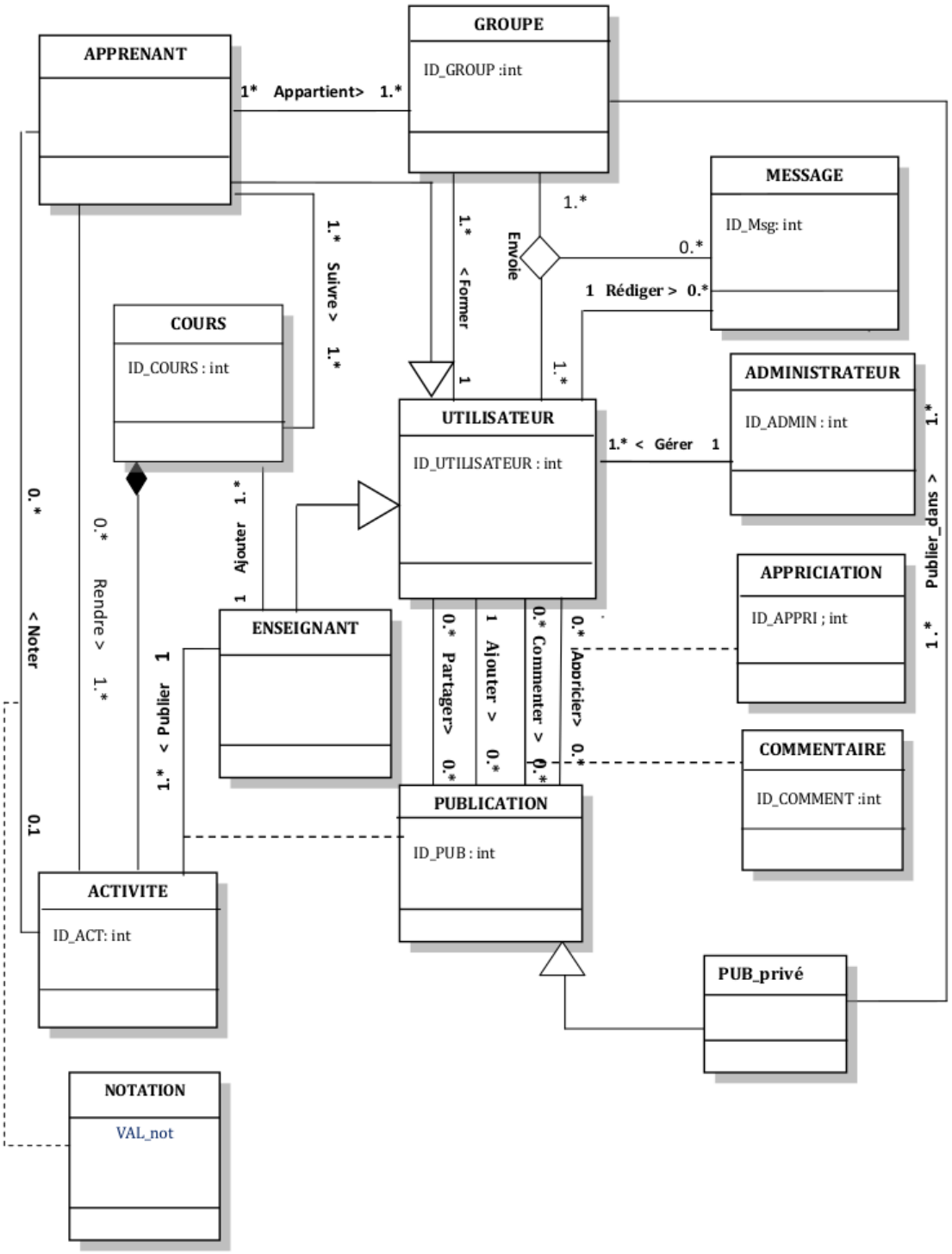

FIGURE 3.4 – Diagramme de classe [19]

3.5 Les caractéristiques des apprenants

Lors de la conception d'un système de regroupement dynamique des apprenants en groupes hétérogènes, il est important de prendre en compte les caractéristiques individuelles de chaque apprenant. Pour ce faire, nous avons utilisé deux critères de collaboration : la **collaboration lors de l'apprentissage (du cours)** et la **collaboration lors de l'examen** [34].

- **La collaboration lors de l'apprentissage :** où les interactions de l'apprenant avec ses collègues et son enseignant, sur la plateforme d'apprentissage, sont mesurées.
- **La collaboration lors de l'examen :** où les interactions lors de d'un examen qui consiste en la résolution d'un problème, collectivement, par un groupe, sont mesurées.

Pour évaluer les critères ci-dessus, des métriques, proposés dans [34], ont été utilisés.

3.5.1 Les métriques pour évaluer la collaboration lors de l'examen :

- **Le niveau de communication :** Le nombre de messages échangés au sein du groupe pendant la période de test est utilisé pour calculer un indice qui évalue la coopération des étudiants dans l'utilisation des réseaux sociaux pour l'apprentissage. Sa valeur varie entre 0 et 100.
- **Le sentiment :** Le sentiment de chaque étudiant est calculé à partir des mots utilisés dans leurs interactions dans le réseau d'apprentissage social. Cette valeur est normalisée à une valeur comprise entre 0 et 100, les valeurs supérieures indiquant un sentiment positif et les valeurs inférieures indiquant un sentiment négatif.
- **La note :** L'évaluation des performances de l'apprenant est basée sur les notes d'examen attribuées par le formateur, en fonction de la qualité de la solution pour chaque activité. Les valeurs vont de 0 à 100.
- **La présence :** La participation de l'apprenant est basée sur le nombre de tests passés, qu'il s'agisse de devoirs ou de tests. Cette mesure est obtenue en divisant le nombre de tests auxquels l'apprenant a répondu par le nombre total de tests. Les valeurs vont de 0 à 100.

3.5.2 Les métriques pour évaluer la collaboration lors de l'apprentissage :

- **Les mentions :** Une mesure de l'appréciation donnée par l'apprenant est effectuée en comptant le nombre de mentions (*likes*) qu'il a données au différentes publications.Les valeurs vont de 0 à 100.

- **Les commentaires :** Une mesure de l'engagement des apprenants à commenter les publications est évaluée en comptant le nombre de commentaires qu'ils ont écrits et en divisant par le nombre total de commentaires publiés. Cette mesure permet de quantifier la contribution de l'apprenant à l'échange sur le contenu proposé. Les valeurs vont de 0 à 100.

3.6 Regroupement automatique des apprenants

Dans notre étude, nous avons expérimenté deux méthodes différentes pour le regroupement automatique des apprenants.

La première méthode que nous avons utilisée est ML-kNN (Multi-Label k-Nearest Neighbors), une approche couramment utilisée pour la classification multi-label. ML-kNN se base sur la recherche des voisins les plus proches d'une instance donnée et prédit les étiquettes en se basant sur les étiquettes des voisins.

La deuxième méthode que nous avons expérimentée est l'AG (Algorithme Génétique), une technique d'optimisation basée sur les principes de la sélection naturelle et de l'évolution.

3.6.1 ML-kNN

ML-kNN est un algorithme de Machine Learning qui appartient à la classe des algorithmes d'apprentissage supervisé, simple et facile à mettre en œuvre qui peut être utilisé pour résoudre les problèmes de régression et de classification multi-label.

La figure 3.5 présente les principales étapes de ML-kNN.

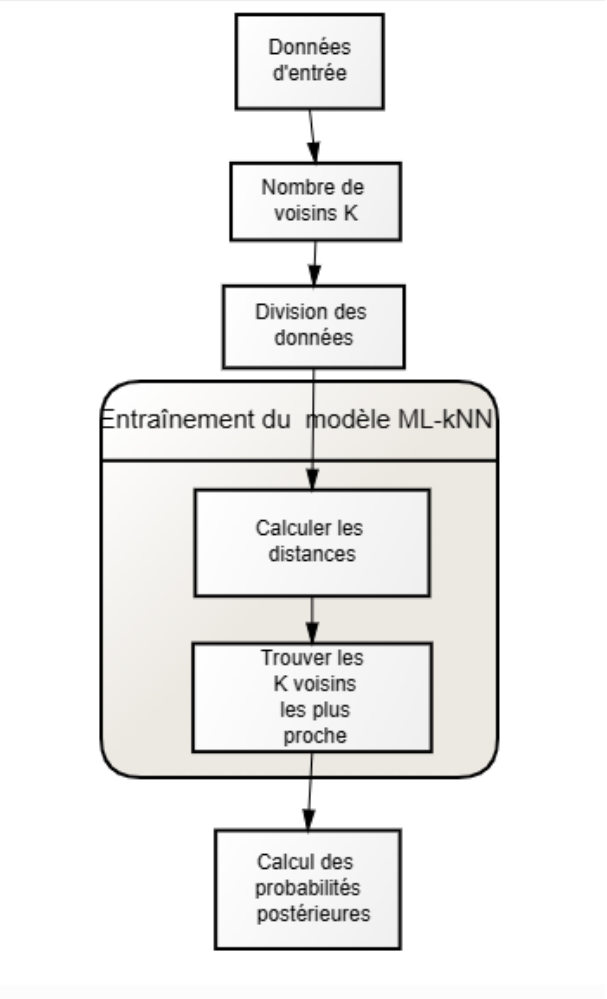

FIGURE 3.5 – Algorithme ML-KNN

L'algorithme complet est présenté par le listing ci-dessous :

Algorithm 1 Algorithme ML-KNN

Entrée :
X : Matrice des caractéristiques d'entraînement (taille : n_samples x n_features)
y : Matrice des étiquettes d'entraînement (taille : n_samples x n_labels)
k : Nombre de voisins à considérer
Sortie :
y_pred : Matrice des étiquettes prédites (taille : n_samples x n_labels)
Algorithme :
Normaliser les caractéristiques de X (si nécessaire)
Calculer la matrice des distances entre les échantillons de X (utiliser une mesure de distance appropriée, comme la distance euclidienne)
Pour chaque échantillon i de X
    Trouver les k voisins les plus proches de l'échantillon i en utilisant les distances calculées
    Pour chaque étiquette l
        Initialiser le compteur de voisins ayant l'étiquette l à 0
        Pour chaque voisin v parmi les k voisins les plus proches
            Si l'étiquette l est présente dans le voisin v, incrémenter le compteur de voisins ayant l'étiquette l
        Calculer la probabilité conditionnelle P(y_l=1 | voisin) en divisant le compteur de voisins ayant l'étiquette l par k
        Si P(y_l=1 | voisin) $>$ 0.5, prédire y_l=1 pour l'échantillon i, sinon prédire y_l=0
Retourner la matrice des étiquettes prédites y_pred

L'algorithme de ML-kNN, appliqué à notre problème, est le suivant :

Étape 1 - Charger les données d'entrée. Tout d'abord, une matrice est créée avec les identifiants des étudiants dans la première colonne et les indicateurs dans les autres colonnes.

Étape 2 - Effectuer une classification pour déterminer si chaque étudiant est collaboratif, assidu ou motivé, en utilisant des seuils prédéfinis appliqués sur les indicateurs des étudiants.

Étape 3 - Insérer les étiquettes de classification dans la matrice de données..

Étape 4 - Diviser les données en ensembles de formation et de test. Cela implique de séparer les données en deux ensembles distincts, l'un pour l'entraînement et l'autre pour la validation, en utilisant la fonction *train-test-split* de la bibliothèque *scikit-learn*.

Étape 5 - Générer des étiquettes de classification aléatoires et Transformer les étiquettes en vecteurs binaires à l'aide de *OneHotEncoder* de la bibliothèque *sklearn.preprocessing*. Cette étape est importante pour créer des étiquettes binaires pour chaque classe de classification, afin que chaque étudiant soit dans une classe ou dans une autre. Ces étiquettes binaires sont ensuite utilisées pour entraîner le modèle de classification ML-kNN.

Étape 6 - Entraîner un modèle de classification ML-kNN en utilisant la méthode de la régression logistique. Nous utilisons *GridSearchCV* de la bibliothèque *sklearn.model-selection* pour entraîner un modèle ML-kNN et pour ajuster les hyper-paramètres. en utilisant différentes valeurs de k (nombre de voisins à considérer pour la prédiction).

Étape 7 - Prédire les étiquettes des étudiants dans l'ensemble de test à l'aide de *predict* sur l'ensemble de données complète. Les étudiants sont ensuite divisés en fonction de leur classification à l'aide de masques booléens (*y-pred-all[:,0] == 1*, etc.).

Étape 8 - Afficher les classes des étudiants *collaboratifs*, *assidus* et *engagé*.

3.6.2 Modèle de regroupement automatique à base de ML-kNN

Notre système forme des groupes hétérogènes d'apprenants. Les groupes sont formés en fonction de l'activité de l'apprenant sur le système d'apprentissage.

Les activités sont représentées par six indicateurs :

- Les **mentions** sur les publications, lors de l'apprentissage.
- Les **commentaires** sur les publications, lors de l'apprentissage.
- Le **niveau de communication** lors de l'examen.
- Le **sentiment** lors de l'examen.
- La **note** lors de l'examen.
- La **présence** lors de l'examen.

Ces paramètres nous permettent de regrouper les apprenants selon 3 étiquettes :

- *Apprenant collaboratif* qui communique activement pendant le cours et échange des messages dans la messagerie.
- *Apprenant assidu* assiste régulièrement et participe activement aux activités pédagogiques.
- *Apprenant engagé* qui participe activement aux activités et exprime souvent des sentiments positifs.

Des groupes sont ensuite constitués en essayant de respecter la règle suivante : *un groupe doit contenir, au moins, un membre ayant une des trois étiquettes.*

Le schéma ci-dessous représente la structure conceptuelle du module qui permet de rassembler les apprenants de manière automatique.

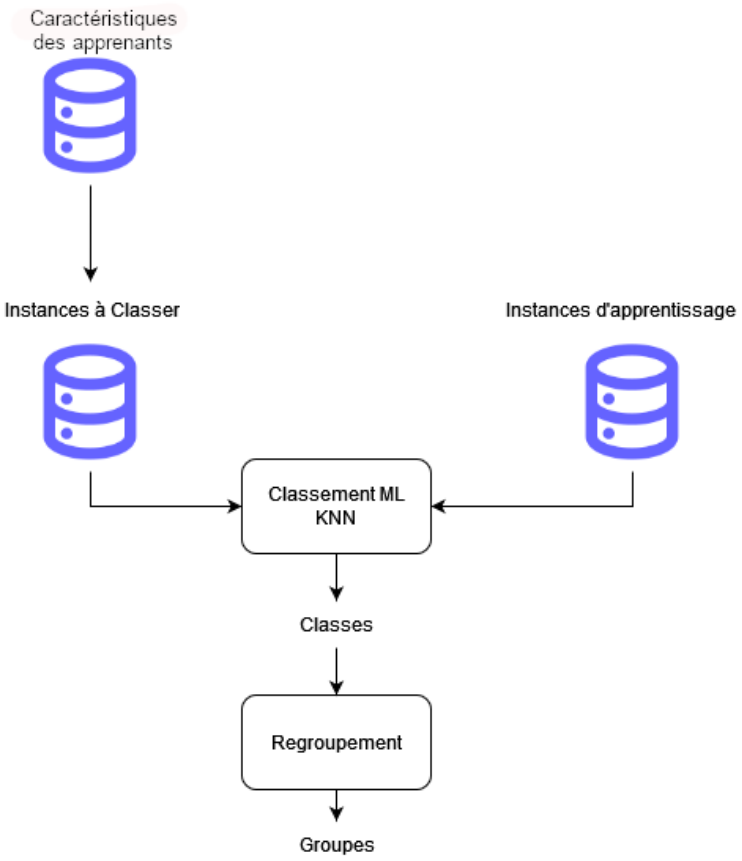

FIGURE 3.6 – Module de regroupement

3.6.3 Algorithme génétique

Un algorithme génétique (AG) est une heuristique de recherche basée sur la théorie de l'évolution de Darwin, qui simule le processus de sélection naturelle. Dans cet algorithme, les individus les plus adaptés sont choisis pour se reproduire, créant ainsi une nouvelle génération. Cela permet d'explorer efficacement l'espace des solutions à un problème donné et d'atteindre des solutions optimales [36].

Ils sont utilisés pour résoudre des problèmes d'optimisation et de recherche de solutions et peuvent être utilisés dans le contexte de la formation de groupes d'étudiants afin de favoriser la constitution de groupes hétérogènes.

Le principe des AG :

Le processus débute par une population composée d'un ensemble d'individus, où chaque individu représente une solution potentielle au problème à résoudre.

Chaque individu est caractérisé par un ensemble de paramètres ou de variables appelés *gènes*. Ces gènes sont regroupés ensemble pour former un *chromosome*, qui représente une solution complète. Dans le contexte des algorithmes génétiques, les gènes d'un individu sont généralement représentés sous forme de chaînes, où chaque élément de la chaîne correspond à une valeur spécifique du gène [36].

Cinq phases sont considérées dans un algorithme génétique [36] :

1. **La génération de la population initiale :** est un processus crucial pour assurer la diversité et la représentativité de la population. Il est essentiel que cette population soit non homogène et couvre l'ensemble du domaine de recherche, surtout lorsque les informations sur le problème à résoudre sont limitées.

2. **Une fonction à optimiser :** appelée *fitness* ou fonction d'évaluation de l'individu. Celle-ci est utilisée pour sélectionner et reproduire les meilleurs individus de la population. Chaque individu de la population se voit attribuer un score de fitness qui évalue sa qualité ou sa performance par rapport au problème à résoudre. La probabilité qu'un individu soit sélectionné pour participer au processus de reproduction est déterminée en fonction de son score de fitness.

3. **La croisement :** également appelée *crossover*, est l'une des étapes les plus importantes dans un algorithme génétique. Lorsqu'il est temps d'accoupler une paire de parents, un point de croisement est sélectionné au hasard à l'intérieur de leurs gènes. Cela permet de mélanger les informations génétiques des parents et de créer une nouvelle progéniture avec une combinaison unique de caractéristiques héritées. Le choix aléatoire du point de croisement garantit une diversité génétique et favorise l'exploration de l'espace des solutions pour trouver de nouvelles combinaisons prometteuses.

4. **La mutation :** La mutation est un processus où certains gènes d'une partie des nouveaux descendants sont soumis à des modifications aléatoires avec une faible probabilité. La mutation se produit dans le but de maintenir la diversité au sein de la population et d'éviter une convergence prématurée vers un optimum local. En introduisant des changements aléatoires dans les gènes, la mutation permet d'explorer de nouvelles régions de l'espace des solutions et d'empêcher que la population ne soit bloquée dans un seul mode de solution. Cela favorise l'exploration continue et offre une chance de découvrir de meilleures solutions.

5. **Les paramètres de dimensionnement :** définissent la configuration de l'algorithme : taille de la population, nombre total de générations ou critère d'arrêt, probabilités d'application des opérateurs de croisement et de mutation.

Algorithme général d'un AG

Les étapes principales d'un AG sont présentées par l'algorithme suivant :

Algorithm 2 Algorithme génétique

Entrée :
Caractéristiques des apprenants
Sortie :
Groupes d'apprenants
Initialiser la population initiale ;
Calculer le fitness ;
repeat
  Sélectionner des individus ;
  Effectuer le croisement ;
  Effectuer la mutation ;
  Calculer le fitness ;
until *La population a convergé*;

Notre objectif est d'explorer comment l'algorithme génétique peut être appliqué pour créer des groupes qui maximisent la diversité des compétences, les interactions positives et l'équilibre des connaissances parmi les étudiants.

Quelques travaux sur les algorithme génétique dans le e-learning

- Dans [40], Anon Sukstrienwong présente une approche utilisant un algorithme génétique pour équilibrer les styles d'apprentissage et les attributs académiques dans la formation hétérogène de groupes d'étudiants. L'objectif de la recherche est de trouver un moyen efficace de regrouper les étudiants en tenant compte de leurs styles d'apprentissage préférentiels et de leurs attributs académiques, afin d'améliorer les résultats et la satisfaction des étudiants dans un environnement d'apprentissage collaboratif.

- Dans [26], Pinninghof *et al.* ont présenté une approche de formation de groupes collaboratifs utilisant des algorithmes génétiques. L'objectif de la recherche est de proposer une méthode efficace pour former des groupes en maximisant la complémentarité des compétences et en favorisant la cohésion et l'interaction positive entre les membres du groupe.

- Dans [27], Angelica *et al.* ont mis en évidence l'utilisation des algorithmes génétiques comme un outil puissant pour la structuration des groupes collaboratifs. L'objectif de cette recherche est de proposer une approche pour la formation de groupes de travail, en optimisant la composition des membres et en favorisant la collaboration et la productivité.

Modèle de regroupement à base d'algorithme génétique

Dans le contexte de la formation de groupes d'étudiants, l'algorithme génétique peut être utilisé pour trouver la combinaison optimale d'étudiants qui maximisera la cohésion et la productivité du groupe. Pour cela, l'algorithme génétique peut aider à créer des groupes *intra-hétérogènes*, qui favorisent la diversité des compétences et l'interaction positive entre les membres d'un même groupe, et *inter-homogène*, en constituant des groupes les plus semblables possibles.

L'algorithme génétique est composé de plusieurs étapes clés qui sont répétées de manière itérative pour trouver la meilleure solution possible. Voici les étapes principales de l'algorithme génétique :

1. **Initialisation de la population :** Une population initiale de chromosomes est générée. Chaque chromosome représente une solution potentielle pour la formation des groupes, où chaque valeur dans le chromosome correspond à un groupe auquel un étudiant est assigné.

2. **Évaluation de la population :** Chaque individu de la population est évalué en termes de fitness. Cela implique l'utilisation de fonctions de fitness pour mesurer à la fois la cohésion intra-groupe et l'interaction inter-groupes. La fonction de fitness intra-groupe évalue la similarité des étudiants au sein d'un groupe, tandis que la fonction de fitness inter-groupes mesure la différence entre les profils moyens des groupes.

3. **Sélection des individus :** Il s'agit de sélectionner les chromosomes ayant les meilleures valeurs de fitness, c'est-à-dire ceux qui correspondent aux groupes les plus homogènes en termes de compétences intra-groupe et les plus hétérogènes en termes de compétences inter-groupes. Les individus sélectionnés ont une plus grande probabilité d'être utilisés pour la prochaine génération.

4. **Opérateur de croisement :** Les individus sélectionnés sont soumis à des opérations de croisement. L'objectif du croisement est de combiner les caractéristiques des parents pour créer de nouveaux individus (enfants).

5. **Mutation :** Un certain pourcentage de la population initiale est sélectionné, et une opération de mutation est appliquée à ces individus sélectionnés. L'objectif de cette étape est d'introduire une exploration plus diversifiée de l'espace de recherche et d'éviter la convergence prématurée vers un optimum local.

6. **Remplacement de la population :** Après avoir effectué la sélection, le croisement et la mutation, on remplace la population initiale par la nouvelle population résultante. Cette nouvelle population comprend les individus sélectionnés, les descendants issus du croisement et les individus mutés. En effectuant ce remplacement, on veut assurer que chaque génération de l'algorithme évolue vers une meilleure adaptation au problème.

7. **Répétition des étapes :** les étapes 2 à 6 sont généralement répétées jusqu'à ce qu'un critère d'arrêt soit atteint. dans notre algorithme, le critère d'arrêt est le nombre d'itérations.

Le schéma ci-dessous montre la représentation graphiques des étapes suivies par l'algorithme génétique pour rassembler les apprenants de manière dynamique.

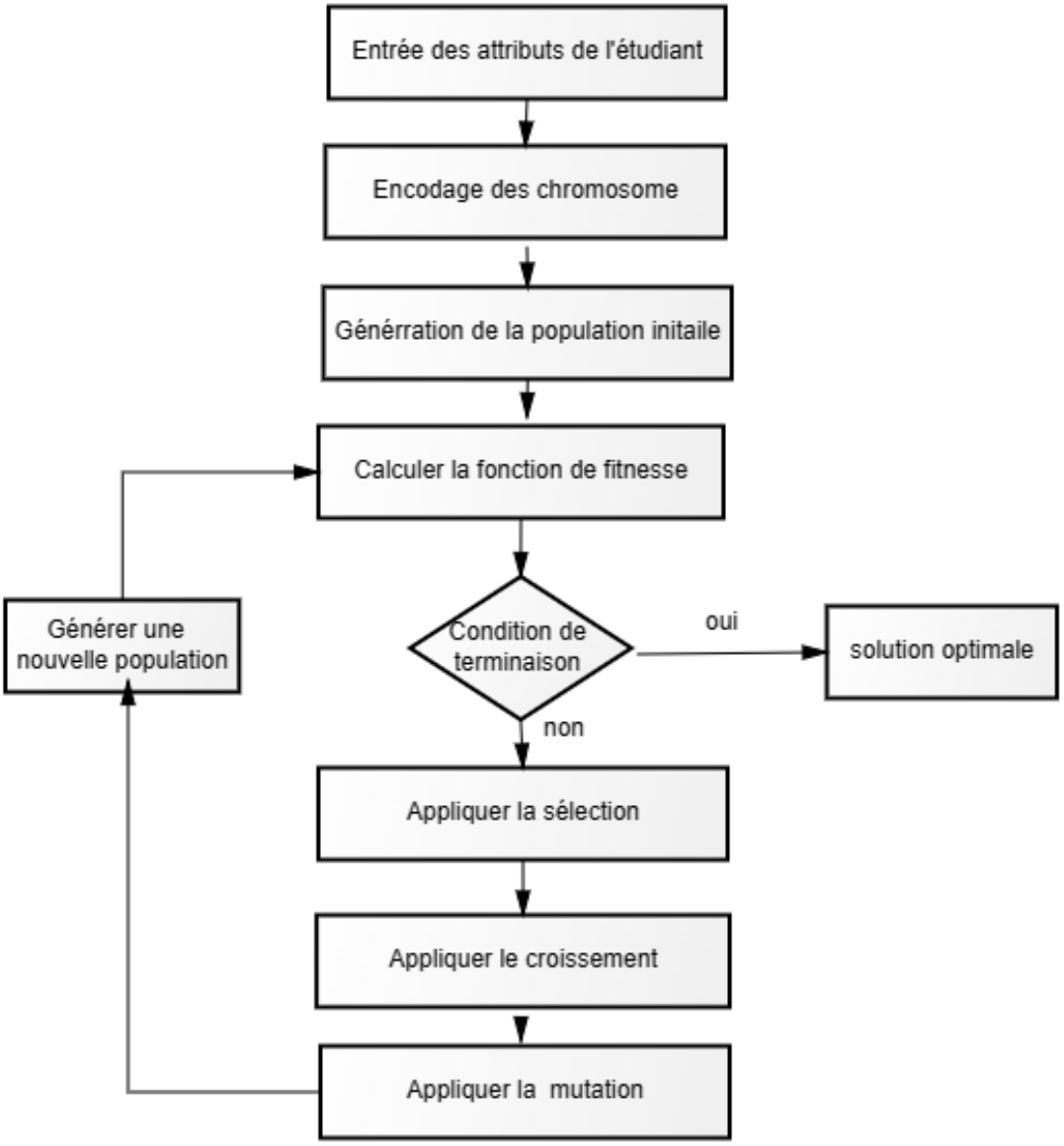

FIGURE 3.7 – Algorithme génétique proposé pour le regroupement des étudiants

3.7 Conclusion

Dans ce chapitre nous avons présenté la conception d'un système de regroupement dynamique d'apprenants utilisant deux méthodes de regroupement, à savoir ML-kNN et un algorithme génétique. Nous avons défini les objectifs de notre projet et identifié les caractéristiques clés des étudiants pour permettre une meilleure adéquation entre les membres de chaque groupe.

Dans le chapitre suivant, nous allons présenter l'implémentation de notre système.

Chapitre 4

Implémentation et résultats

4.1 Introduction

Dans ce chapitre, nous présenterons les langages de programmation, les outils et les environnements matériels et logiciels que nous avons utilisés pour développer notre système. Ensuite, nous procéderons à la description d'une expérimentation de notre approche pour mettre en évidence les points forts et les limitations de notre approche.

4.2 Environnement de développement

4.2.1 Environnement matériel

La machine sur laquelle a été développé notre système a la configuration suivante :

Matériel	Caractéristiques
PC	Processeur : Intel(R) Core(TM) i3-4005U CPU @ 1.70GHZ 1.70 GHZ. Memoire Vive(Ram) : 4.00 Go. Disque Dur : 500Go. Système d'exploitation : Windows 10 Professionnel.

TABLE 4.1 – Caractéristiques du matériel

Cette machine a été utilisé à la fois pour l'implémenation et pour le test du système développé.

4.2.2 Environnement logiciel

Le système a été implémenté, et la documentation rédigée, en utilisant les outils suivants :

Outil de développement

Visual Studio Code

Visual Studio Code est un puissant éditeur de code source qui fonctionne sur les ordinateurs et est disponible en téléchargement gratuit pour Windows, macOS et Linux. Il offre une prise en charge intégrée du JavaScript, du TypeScript et du Node.js, ainsi qu'un écosystème riche en extensions pour d'autres langages (comme C++, C#, Java, Python, PHP, Go) et des environnements d'exécution (comme .NET et Unity) [42].

Langage utilisé

Python

Python est un langage de programmation de haut niveau qui se distingue par sa lisibilité et sa syntaxe concise. Il est conçu pour permettre aux développeurs d'exprimer des concepts et de résoudre des problèmes avec un code concis. Python est distribué sous une licence open source approuvée par l'OSI, ce qui signifie qu'il peut être utilisé et distribué librement, y compris à des fins commerciales.

Python est largement utilisé dans de nombreuses applications professionnelles à travers le monde, et il est également utilisé dans des systèmes critiques et importants [28].

Dans ce travail, nous avons utilisé la version 3.10 de Python pour développer notre projet.

FIGURE 4.1 – Logo de Python

Bibliothèques utilisés

Pandas :

Pandas est une bibliothèque écrite pour le langage de programmation Python permettant la manipulation de données tabulaires. Elle offre des fonctionnalités pour importer et exporter des données à partir de différents formats de fichiers tels que CSV, JSON, Excel, etc. Pandas permet de réaliser des opérations de fusion, de sélection, de nettoyage et de transformation des données. Elle est largement utilisée dans le domaine de l'analyse de données pour son efficacité et sa facilité d'utilisation [23].

NumPy :

NumPy est une bibliothèque pour le langage de programmation Python, permettant de manipuler des tableaux multidimensionnels et d'effectuer des opérations mathématiques avancées sur ces tableaux. Elle offre une alternative efficace aux

listes Python pour le calcul numérique et est largement utilisée dans les domaines scientifiques et d'ingénierie [22].

Scikit-learn :

Scikit-learn est une bibliothèque pour le langage de programmation Python dédiée à l'apprentissage automatique (machine learning) offrant une large gamme d'algorithmes de classification, de régression et de regroupement. Elle s'intègre parfaitement avec les bibliothèques numériques et scientifiques de Python telles que NumPy et SciPy. Scikit-learn est largement utilisée pour développer des modèles prédictifs et analyser des données dans divers domaines [35].

Matplotlib :

Matplotlib est une bibliothèque complète pour la création de visualisations statiques, animées et interactives en Python. Elle permet de créer des graphiques de qualité professionnelle, d'en personnaliser le style visuel et la mise en page, et de les exporter dans de nombreux formats de fichiers. Matplotlib offre également la possibilité de créer des figures interactives avec des fonctionnalités de zoom, de panoramique et de mise à jour. Elle est largement utilisée et bénéficie d'un écosystème riche de packages tiers construits sur sa base [20].

4.3 Expérimentation

Dans cette section d'expérimentation, nous aborderons deux approches différentes pour résoudre notre problème de classification des étudiants. Tout d'abord, nous examinerons l'algorithme ML-kNN, ensuite, nous nous pencherons sur l'algorithme génétique (AG), avec le même objectif : créer des groupes d'apprentissages collaboratifs.

4.3.1 Dataset Utilisé

Les données de l'expérimentation seront générées de manière aléatoire pour représenter une série d'utilisateur (étudiants). Chaque individu est identifié par un ID unique et possède différentes caractéristiques, parmi lesquelles :

- *acces :* Niveau d'accès de l'utilisateur.
- *nb-appreciations :* Nombre d'appréciations données par l'utilisateur.
- *nb-commentaires :* Nombre de commentaires écrits par l'utilisateur.
- *nb-messages :* Nombre de messages envoyés par l'utilisateur.
- *nb-publications :* Nombre de publications postées par l'utilisateur.
- *note :* Notes reçues par l'utilisateur.
- *presence :* Présence de l'utilisateur aux examens.
- *sentiment :* Sentiment exprimé par l'utilisateur lors de ces échanges.

Chaque ligne du dataset représente un utilisateur spécifique avec ses caractéristiques associées.

Il est important de noter que ce dataset est présenté sous forme de valeurs séparées par des virgules (CSV). Pour une utilisation plus pratique et une analyse plus approfondie, nous pouvons importer ces données dans un environnement de programmation, tel que Python, en utilisant des bibliothèques de manipulation de données comme Pandas.

	A	B	C	D	E	F	G	H	I	J	K
1	id	acces	nb_apprecia	nb_commen	nb_message	nb_publicati	note	presence	sentiment	nom	prenom
2	1	83	61	69	12	67	43	8	70	Nom1	Prenom1
3	2	40	94	74	5	9	64	13	95	Nom2	Prenom2
4	3	31	84	24	30	36	19	12	98	Nom3	Prenom3
5	4	65	2	49	18	66	60	60	97	Nom4	Prenom4
6	5	84	56	100	14	41	35	98	91	Nom5	Prenom5
7	6	63	90	17	22	16	8	46	15	Nom6	Prenom6
8	7	5	32	63	72	47	93	78	85	Nom7	Prenom7
9	8	32	39	61	47	30	33	55	42	Nom8	Prenom8
10	9	66	61	32	48	13	90	15	72	Nom9	Prenom9
11	10	85	35	98	61	77	1	42	22	Nom10	Prenom10
12	11	25	40	88	34	84	66	63	65	Nom11	Prenom11
13	12	3	5	51	69	60	79	12	23	Nom12	Prenom12
14	13	52	51	35	26	100	39	51	70	Nom13	Prenom13
15	14	49	77	61	5	4	59	54	4	Nom14	Prenom14
16	15	88	79	61	52	34	88	81	82	Nom15	Prenom15
17	16	51	81	55	70	23	46	31	57	Nom16	Prenom16
18	17	68	92	83	92	82	12	57	66	Nom17	Prenom17
19	18	79	52	55	28	29	74	58	84	Nom18	Prenom18
20	19	32	43	92	71	29	6	95	78	Nom19	Prenom19
21	20	45	6	26	6	51	50	72	67	Nom20	Prenom20

FIGURE 4.2 – Dataset Utilisé

4.3.2 Regroupement avec ML-kNN

Dans cette phase d'expérimentation, notre objectif est d'obtenir des groupes homogènes regroupant des étudiants ayant des profils hétérogènes.

Après la création des données, nous avons procédé aux étapes suivantes :

- **Normalisation des données :** pour éliminer des écarts d'échelle et l'assurance d'une comparaison équitable entre les différentes caractéristiques.
- **Attribution des labels aux étudiants :**
  1. **Étudiants collaboratifs :** si $nb{-}appreciations >= 30$, $nb{-}commentaires >= 40$, $accs >= 50$ et $nb - messages >= 30$.
  2. **Étudiants assidus :** si $accs >= 20$, $note >= 40$ et $presence >= 30$.
  3. **Étudiants engagés :** si $nb - appreciations >= 40$, $accs >= 50$ et $sentiment >= 70$.
- **Division des données :** la matrice a été divisée en deux parties : un set d'entraînement (80% des étudiants) et un set de test (20% des étudiants).

- **Classification des données du set de test :** par application de l'algorithme ML-kNN, du package Scikit-learn.
- **Résultats de la classification des étudiants :** Après l'application de notre modèle de classification, nous avons obtenu des labels pour chaque étudiant en fonction de leur niveau de collaboration, d'assiduité et de motivation. Les étudiants ont été classés en trois catégories distinctes : *collaboratif*, *assidu* et *engagé*.
- **Création de groupes homogènes avec hétérogénéité des profils :** À partir des labels obtenus, nous avons procédé à la création de groupes d'étudiants en combinant les différents critères de collaboration, d'assiduité et de motivation. L'objectif était de former des groupes homogènes en termes de niveaux de ces caractéristiques, tout en maintenant une certaine hétérogénéité à l'intérieur de chaque groupe.

  Chaque groupe a été conçu de manière à inclure au moins un étudiant qui présente un comportement collaboratif, un étudiant assidu et un étudiant engagé. Cela permet d'encourager la diversité et la complémentarité des profils au sein de chaque groupe, favorisant ainsi les interactions et les échanges entre les étudiants.

Le tableaux 4.2 présente les résultats de la classification ML-kNN appliquée au dataset.

Classe	Id des étudiants	Total
Étudiants collaboratifs	1.0, 2.0, 3.0, 4.0, 5.0, 6.0, 7.0, 8.0, 9.0, 11.0, 12.0, 13.0, 14.0, 16.0, 19.0, 20.0	16
Étudiants assidus	10.0, 15.0, 17.0, 18.0	4
Étudiants engagés	2.0, 3.0, 5.0, 6.0, 7.0, 8.0, 9.0, 10.0, 12.0, 16.0, 17.0, 19.0	12

TABLE 4.2 – Classification des étudiants

Dans le cas des étudiants multi-étiquetés, nous pouvons trouver qu'un étudiant peut appartenir à une ou plusieurs classes simultanément. Par exemple, l'étudiant 10 peut être étiqueté comme appartenant aux classes "Assidu" et "Motivé", tandis que l'étudiant 19 peut être étiqueté comme appartenant aux classes "Collaboratif" et "Motivé".

Après regroupement des étudiants issus de chaque classe en groupe de 4 membres, nous obtenons les groupes suivants du tableau 4.3.

Les groupes	les étudiants
Groupe 1	[16.0, 11.0, 1.0, 5.0]
Groupe 2	[18.0, 20.0, 14.0, 4.0]
Groupe 3	[3.0, 17.0, 7.0, 10.0]
Groupe 4	[12.0, 2.0, 15.0, 19.0]
Groupe 5	[9.0, 6.0, 8.0, 13.0]

TABLE 4.3 – Regroupement des étudiants

Validation des résultats

Pour pouvoir comparer les résultats obtenus par le ML-kNN avec ceux obtenus avec l'algorithme génétique, nous avons calculé les fonctions de fitness intra et inter-groupe sur les groupes du tableau 4.3 :

- *fitness intra-groupe :* est la moyenne de la distance euclidienne entre les caractéristiques des apprenants d'un même groupe.
- *fitness inter-groupes :* est la distance euclidienne entre les moyennes des caractéristiques des apprenants de chaque groupe, calculée entre tous les groupes.

Les tableaux 4.4 et 4.5 présentent les valeurs des fitness intra et inter-groupes calculées sur les groupes formé à partir de la classification ML-kNN.

Groupe	Valeur de f-intra
Groupe 1	793.54
Groupe 2	660.23
Groupe 3	1092.80
Groupe 4	1049.74
Groupe 5	586.56

TABLE 4.4 – Valeurs de fitness intra-groupe (ML-kNN)

Groupes	1	2	3	4	5
1	0	86.33	66.41	57.64	63.67
2	86.33	0	123.30	111.31	74.18
3	66.41	123.30	0	33.30	76.83
4	57.64	111.31	33.30	0	75.71
5	63.67	74.18	76.83	75.71	0

TABLE 4.5 – Valeurs de fitness inter-groupes (ML-kNN)

Les figures 4.3 et 4.4 présentent les valeurs des fonctions de fitness sous forme graphique.

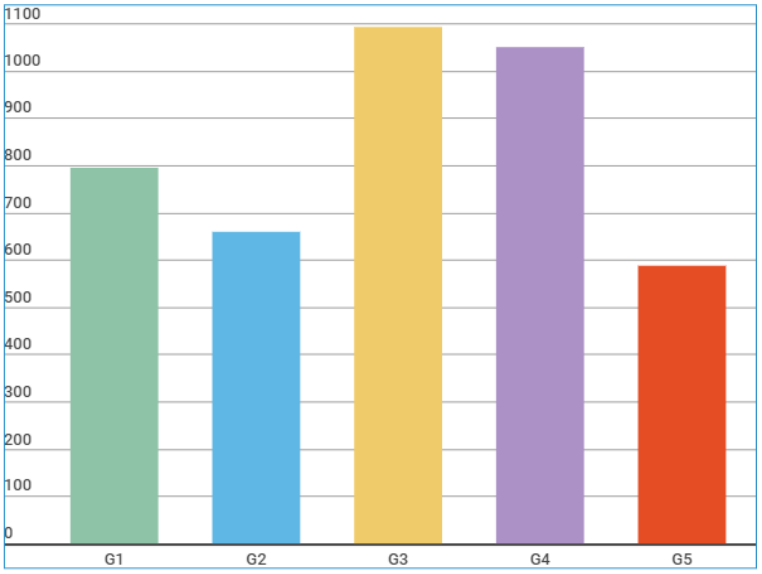

FIGURE 4.3 – Histogramme de la fonction f-intra pour ML-kNN

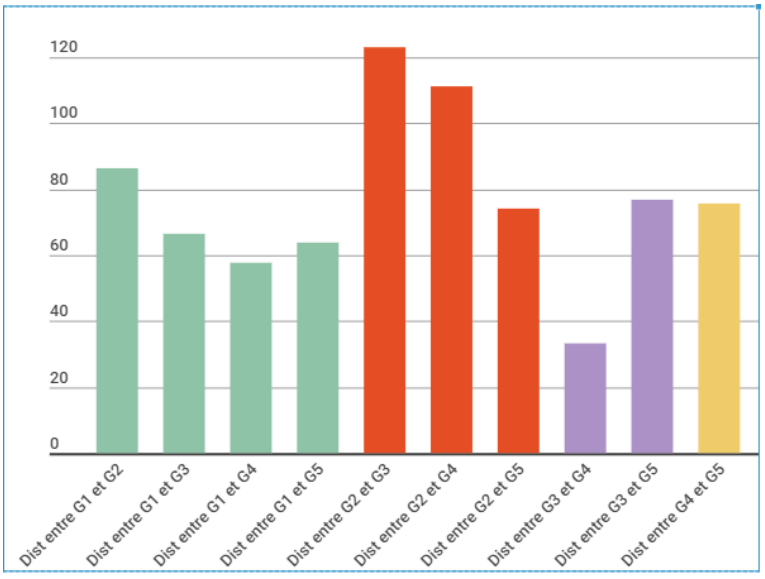

FIGURE 4.4 – Histogramme de la fonction f-inter pour ML-kNN

Trois autres indicateurs ont été calculés pour comparer les méthodes de regroupement ML-kNN et AG :

- Le temps d'exécution : 1.70 seconde.
- La variance de la fonction f-intra : 7172.82
- La variance de la fonction f-inter : 282.49.

4.3.3 Regroupement avec l'algorithme génétique

Pour obtenir des groupes homogènes regroupant des étudiants ayant des profils hétérogènes, nous avons également expérimenté un algorithme génétiques paramétré ainsi :

- Une population de 20 chromosomes, qui représentent des solutions potentielles.
- Un critère d'arrêt fixé à une valeur de 7 itérations.
- Un ratio de sélection de 30% des meilleurs individus.
- Un ratio de croisement de 50% .
- Un ratio de mutation à 20% de la population initiale.

Le tableau 4.6 représente les groupes créés à partir l'application de l'algorithme génétique.

Les groupes	les étudiants
Groupe 1	[11.0, 17.0, 18.0, 20.0]
Groupe 2	[1.0, 6.0, 10.0, 16.0]
Groupe 3	[4.0, 5.0, 7.0, 9.0]
Groupe 4	[8.0, 13.0, 14.0, 19.0]
Groupe 5	[2.0, 3.0, 12.0, 15.0]

TABLE 4.6 – Regroupement des étudiants obtenu par l'algorithme génétique

Validation des résultats

Les tableaux 4.7 et 4.8 présentent les valeurs f-intra et f-inter du regroupement obtenu par l'algorithme génétique.

Groupe	Valeur de f-intra
Groupe 1	1072.40
Groupe 2	899.08
Groupe 3	995.61
Groupe 4	680.94
Groupe 5	1044.98

TABLE 4.7 – Valeurs de fitness intra-groupe (AG)

Groupes	1	2	3	4	5
1	0	45.39	16.18	24.79	49.54
2	45.39	0	54.99	57.51	47.97
3	16.18	54.99	0	34.57	58.61
4	24.79	57.51	34.57	0	28.77
5	49.54	47.97	58.61	28.77	0

TABLE 4.8 – Valeurs de fitness inter-groupe (AG)

Les figures 4.5 et 4.6 présentent les valeurs des fonctions de fitness sous forme graphique.

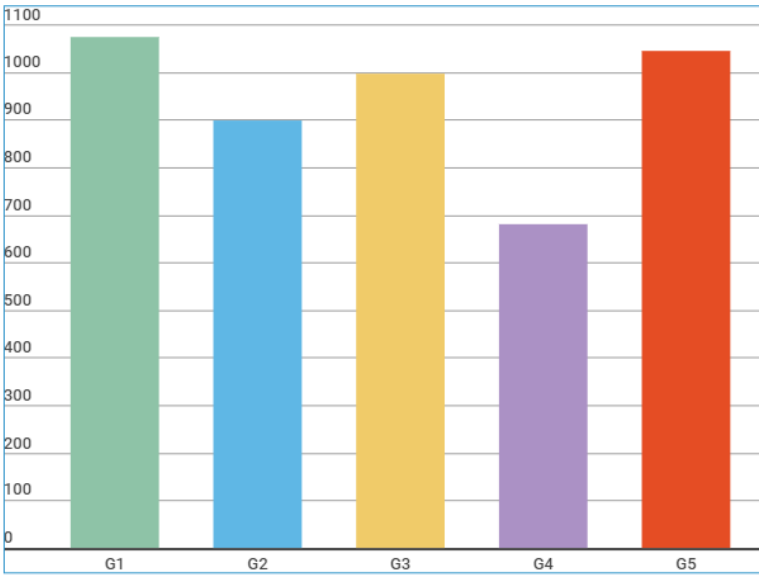

FIGURE 4.5 – Histogramme de la fonction f-intra pour l'algorithme génétique

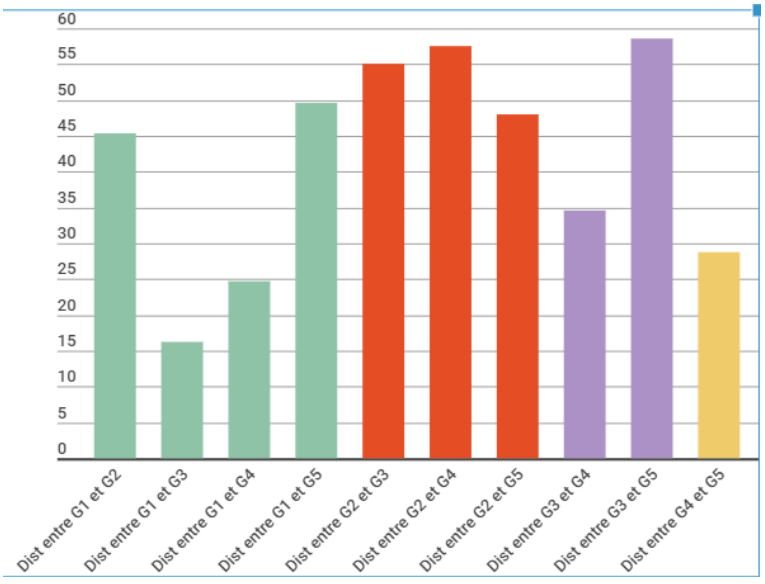

FIGURE 4.6 – Histogramme de la fonction f-inter pour l'algorithme génétique

Trois autres indicateurs ont été calculés pour comparer les méthodes de regroupement ML-kNN et AG :

- Le temps d'exécution : 4.02 secondes.
- La variance de la fonction f-intra : 3700.32.
- La variance de la fonction f-inter :64.26 .

4.4 Discussion des résultats

À partir des résultats obtenus par la simulation, nous avons fait les observations suivantes :

1. On peut observer que les valeurs f-intra de l'AG sont généralement plus élevées que celles obtenu avec un regroupement basé sur ML-kNN, indiquant une meilleure précision des classifications dans les groupes. De plus, l'AG a également montré des valeurs F-inter inférieures, ce qui suggère une meilleure ressemblance entre les groupes.

2. D'après les temps d'exécution de ML-kNN et de l'AG, on peut constater que ML-kNN présente un temps d'exécution plus court, ce qui peut être avantageux dans certaines situations où la rapidité est un critère important.

3. Dans notre comparaison, nous avons constaté que l'AG présente une variance de f-intra de 3700.33, qui est plus faible que celle de ML-kNN (7172.82). Par conséquent, l'AG est plus performant pour former des groupes ou l'hétérogénéité entre les membres est bien distribuée sur l'ensemble des groupes.

4. Quant à la variance de f-inter, qui mesure la séparation ou la différence entre les groupes formés, on constate que l'AG présente une variance de 64.27, tandis que ML-kNN présente une variance légèrement plus élevée de 282.49. Une variance plus faible pour f-inter suggère une meilleure ressemblance entre les groupes formés. Ainsi, l'AG semble également être plus performant dans la formation de groupes homogènes entre eux.

En conclusion, nous avons cherché à former des groupes homogènes où les étudiants au sein d'un même groupe ont des profils hétérogènes, et pour cela, l'AG serait la méthode recommandée, car elle présente une variance de f-inter plus faible par rapport à ML-kNN. Cela indique une meilleure similitude des groupe. De plus, les valeurs de f-intra obtenues avec l'AG sont généralement plus élevées que celles obtenues avec un regroupement basé sur ML-kNN, ce qui suggère une meilleure disparité des profils des étudiants formant les groupes.

Par contre, la force du regroupement basée sur une classification ML-kNN présente comme avantage de présenter à l'enseignant une classification des apprenants en plusieurs catégorie (apprenant collaboratif, assidu et/ou engagé), lui permettant de mieux connaître ses apprenant et d'offrir des formations plus personnalisés.

4.5 Conclusion

Dans ce chapitre, nous avons passé en revue les outils utilisés pour développer notre système, puis une expérimentation nous a permis de comparer entre le regroupement basé sur une classification multi-label et un algorithme génétique.

Conclusion générale

L'importance croissante de l'apprentissage en ligne et de l'apprentissage collaboratif dans le domaine de l'éducation est principalement due à la flexibilité que ces approches offrent aux apprenants, tout en favorisant l'interaction et la collaboration entre pairs. Cependant, pour garantir l'efficacité de ces méthodes d'apprentissage, il est nécessaire de mettre en place des systèmes permettant de constituer des groupes d'apprentissage collaboratif efficaces.

Notre recherche visait à développer un système de classification des apprenants basé sur l'évaluation de leur apprentissage collaboratif, pour pouvoir constituer des groupes collaboratifs.

Nous avons utilisé deux méthodes de regroupement, l'une basé sur une méthode de classification multi-label (ML-kNN) et l'autre sur un algorithme génétique. Le but était de former des groupes homogènes d'apprenants tout en favorisant une certaine hétérogénéité au sein de chaque groupe.

Les résultats de l'expérimentation des deux méthodes de regroupement ont démontré l'efficacité de notre système dans la formation de groupes composés de membres aux profils hétérogènes en fonction de leurs caractéristiques individuelles.

En comparant les deux méthodes de regroupement, nous avons pu mettre en évidence leurs avantages et leurs limites respectives. Le ML-kNN offre une approche basée sur la similarité des caractéristiques, tandis que les algorithmes génétiques permettent d'optimiser la composition des groupes. Le choix entre ces méthodes dépendra des objectifs spécifiques et des contraintes du contexte d'apprentissage.

Pour valider les méthodes de regroupement proposées, il serait utiles de les tester dans un environnement d'apprentissage réel, pour pouvoir évaluer l'impact des groupes formés sur la performance académique et l'engagement des apprenants. Il serait également important d'identifier d'autres indicateurs pour pouvoir mesurer le degré d'engagement des apprenants dans leurs activités et leurs groupes.

Bibliographie

[1] Analyse de données. http://www.statelem.com/analyse_des_donnees.php. Consulté le 08 décembre 2022.

[2] RSJD Baker *et al.* : Data mining for education. *International encyclopedia of education*, 7(3):112–118, 2010.

[3] Ryan S Baker, Albert T Corbett et Vincent Aleven : More accurate student modeling through contextual estimation of slip and guess probabilities in bayesian knowledge tracing. *In International Conference on Artificial Intelligence in Education*, pages 87–94. Springer, 2008.

[4] Ryan SJD Baker, Kalina Yacef *et al.* : The state of educational data mining in 2009 : A review and future visions. *Journal of educational data mining*, 1(1):3–17, 2009.

[5] Marc Bolaños, Aina Ferrà et Petia Radeva : Food Ingredients Recognition Through Multi-label Learning. *New Trends in Image Analysis and Processing – ICIAP 2017*, pages 394–402, 2017.

[6] Geoffray Bonnin et Anne Boyer : Apport des learning analytics. *Administration et Éducation*, (2):125–130, 2015.

[7] Francisco Charte, Antonio J Rivera et María J Del Jesus : *Multilabel classification : problem analysis, metrics and techniques*. Springer International Publishing, 2016.

[8] Elijah Cole, Oisin Mac Aodha, Titouan Lorieul, Pietro Perona, Dan Morris et Nebojsa Jojic : Multi-Label Learning from Single Positive Labels. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6 2021.

[9] Zuzanna Deutschman : Multi-label text classification. https://towardsdatascience.com/multi-label-text-classification-5c505fdedca8, 3 décembre 2019. Consulté le 12 novembre 2022.

[10] John Doe : A study on data analysis. *Journal of Data Analysis*, 15(2):100–120, 2021.

[11] Ashish Dutt, Maizatul Akmar Ismail et Tutut Herawan : A systematic review on educational data mining. *IEEE Access*, 5:15991–16005, 2017.

[12] Rebecca Ferguson : Learning analytics : drivers, developments and challenges. *International Journal of Technology Enhanced Learning*, 4(5-6):304–317, 2012.

[13] Ikram Gagaoua et Pascal Lim : Les learning analytics : un outil pédagogique primordial pour les enseignants ! https://domoscio.com/fr/blog/les-learning-analytics-un-outil-pedagogique-primordial-pour-les-enseignants/, 20 janvier 2022. Consulté le 02 février 2022.

[14] Anatoliy Gruzd et Nadia Conroy : Learning analytics dashboard for teaching with twitter. *In Proceedings of the 53rd Hawaii International Conference on System Sciences*, 2020.

[15] Ángel Hernández-García, Inés González-González, Ana Isabel Jiménez-Zarco et Julián Chaparro-Peláez : Applying social learning analytics to message boards in online distance learning : A case study. *Computers in Human Behavior*, 47:68–80, 2015.

[16] Ajitesh Kumar : Difference : Binary, multiclass & multi-label classification. https://vitalflux.com/difference-binary-multi-class-multi-label-classification, 16 mai 2022. Consulté le 12 décembre 2022.

[17] Khalil Laghmari : *Classification multi-labels graduée : découverte des relations entre les labels, et adaptation à la reconnaissance des odeurs et au contexte big data des systèmes de recommandation*. Thèse de doctorat, Sorbonne Université ; Université Hassan II (Mohammedia, Maroc)., 2018.

[18] Marie Lefevre, Sébastien Iksal, Julien Broisin, Olivier Champalle, Valérie Fontanieu, Christine Michel et Amel Yessad : Learning analytics - etat de l'art sur les outils et méthodes issus de la recherche française. Rapport technique, Direction du Numérique pour l'Éducation, ministère de l'Enseignement supérieur de la Recherche et de l'Innovation (France), 2018.

[19] Sara Mahiedine : conception et realisation d'un modèle de regroupement d'apprenant sur un réseau sociel. Mémoire de master, Université 8 mai 1945, 2018.

[20] Matplotlib : https://scikit-learn.org/stable/index.html. Consulté le 18 juin 2023.

[21] Noureddine-yassine Nair-Benrekia : *Classification interactive multi-label pour l'aide à l'organisation personnalisée des données*. Thèse de doctorat, Université de Nantes, 2015.

[22] Numpy : https://numpy.org/. consulté le 18 juin 2023.

[23] Pandas : https://pandas.pydata.org/. consulé le 18 juin 2023.

[24] Alejandro Peña-Ayala : Educational data mining : A survey and a data mining-based analysis of recent works. *Expert systems with applications*, 41(4): 1432–1462, 2014.

[25] Daniel Peraya : Les learning analytics en question. panorama, limites, enjeux et visions d'avenir. *Distances et médiations des savoirs. Distance and Mediation of Knowledge*, (25), 2019.

[26] M Angélica Pinninghoff, Miguel Ramírez, Ricardo Contreras Arriagada et Pedro Salcedo Lagos : Collaborative group formation using genetic algorithms. *In Bioinspired Computation in Artificial Systems : International Work-Conference on the Interplay Between Natural and Artificial Computation, IWINAC 2015, Elche, Spain, June 1-5, 2015, Proceedings, Part II 6*, pages 330–338. Springer, 2015.

[27] M Angélica Pinninghoff J, Ricardo Contreras A, Pedro Salcedo L et Ricardo Contreras A : Genetic algorithms as a tool for structuring collaborative groups. *Natural Computing*, 16(2):231–239, 6 2017.

[28] Python : https://www.python.org/doc/. Consulté le 10 juin 2023.

[29] Joseph M Reilly et Bertrand Schneider : Predicting the quality of collaborative problem solving through linguistic analysis of discourse. *International Educational Data Mining Society*, 2019.

[30] Cristobal Romero et Salvador Ventura : Educational data mining. *International Journal of Information Technology and Decision Making*, 9(4):569–587, 2010.

[31] Cristobal Romero et Sebastian Ventura : Educational data mining : A survey from 1995 to 2005. *Expert systems with applications*, 33(1):135–146, 2007.

[32] Cristobal Romero et Sebastian Ventura : Data mining in education. *Wiley Interdisciplinary Reviews : Data mining and knowledge discovery*, 3(1):12–27, 2013.

[33] Cristóbal Romero, Sebastián Ventura et Eduardo García : Data mining in course management systems : Moodle case study and tutorial. *Computers & Education*, 51(1):368–384, 2008.

[34] Khouloud Saidia : Conception et réalisation d'un système d'évaluation de l'apprentissage par des exercices d'évaluation collectifs (collaboratifs et coopératifs). Mémoire de master, Université 8 mai 1945 guelma, 2022.

[35] scikit learn : https://scikit-learn.org. consulté le 18 juin 2023.

[36] Clinton Sheppard : *Genetic algorithms with Python*. Smashwords Edition S. l, 2017.

[37] Simon Buckingham Shum et Rebecca Ferguson : Social learning analytics. *Journal of educational technology & society*, 15(3):3–26, 2012.

[38] Wissam Siblini : *Apprentissage multi label extrême : comparaisons d'approches et nouvelles propositions*. Thèse de doctorat, Université de Nantes, 2018.

[39] George Siemens : Learning analytics : The emergence of a discipline. *American Behavioral Scientist*, 57(10):1380–1400, 2013.

[40] Anon Sukstrienwong : A Genetic-algorithm Approach for Balancing Learning Styles and Academic Attributes in Heterogeneous Grouping of Students. *International Journal of Emerging Technologies in Learning (ijet)*, 12(03):4, 3 2017.

[41] G. Tsoumakas et I. Vlahavas : Multi-label classification : An overview. *International Journal of Data Warehousing and Mining*, 3(3):1–13, 2007.

[42] VSCode : https://code.visualstudio.com/. Consulté le 10 juin 2023.

[43] Hong-Xing Yu, Wei-Shi Zheng, Ancong Wu, Xiaowei Guo, Shaogang Gong et Jian-Huang Lai : Unsupervised Person Re-Identification by Soft Multilabel Learning. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6 2019.

[44] Min-Ling Zhang et Zhi-Hua Zhou : ML-KNN : A lazy learning approach to multi-label learning. *Pattern recognition*, 40(7):2038–2048, 2007.

[45] Min-Ling Zhang et Zhi-Hua Zhou : A review on multi-label learning algorithms. *IEEE transactions on knowledge and data engineering*, 26(8):1819–1837, 2013.

[46] Yong Zheng, Bamshad Mobasher et Robin Burke : Context recommendation using multi-label classification. *In 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, volume 2, pages 288–295. IEEE, 2014.

[47] Liang Zhou, Xiaoyuan Zheng, Di Yang, Ying Wang, Xuesong Bai et Xinhua Ye : Application of multi-label classification models for the diagnosis of diabetic complications. *BMC medical informatics and decision making*, 21(1):182, 2021.