Editions Ishango - UAC
Revue Internationale Multidisciplinaire Etincelle

Ed. ishango

Article

Prédiction des notes finales des étudiants en fin du premier cycle : *Utilisation de Data Mining* éducatif

Sédar Paluku Mutsotsya[1*], Héritier Nsenge Mpia[2], Manassé Munga Nzanzu[1], Nephtali Inipaivudu Baelani[1], Moise Katembo Kasolene[2]

1 Faculté des Sciences économiques et de gestion, Université de l'Assomption au Congo, Butembo, B.P. 104, République Démocratique du Congo
2 Faculté des Sciences Appliquées, Université de l'Assomption au Congo, Butembo, B.P. 104, République Démocratique du Congo

* Corresponding author: sedareug@uaconline.edu.cd

Abstract: L'analyse des performances académiques est basée sur un certain nombre de facteurs qui mesurent les capacités intellectuelles. Cette mesure de la performance est basée sur un certain nombre de variables qui peuvent être utilisées pour prédire la performance académique et les notes finales. Les outils statistiques permettent avant tout de mesurer l'impact des données les unes sur les autres. À cette fin, l'étude des corrélations est le meilleur compromis pour identifier les vrais et les faux prédicteurs entre les variables. Aussi, par le biais des statistiques descriptives, la distribution des données est d'ordre important pour mesurer les différentes dispersions des données dans leur exploration, ces analyses minutieuses amènent le chercheur à devoir cibler les éléments nécessaires à la prédiction des outputs. Pour prédire les performances académiques des étudiants, il est important d'avoir une idée de l'historique de leurs études, afin de pouvoir identifier les facteurs qui peuvent être déterminants pour prédire le résultat final à la fin du cycle. Les modèles d'apprentissage automatique ont joué un rôle clé dans la prédiction des performances. Après avoir exploré les données, nous avons pu comparer les performances des différents algorithmes. Nous avons comparé six algorithmes, dont deux ont atteint une performance acceptable de 80% pour le coefficient de détermination : la régression linéaire et le SVM. L'algorithme le plus performant est la régression linéaire, étant donné que le problème abordé est basé sur la régression. La régression est un type d'algorithme souvent adapté à l'analyse quantitative où la variable cible supporte des valeurs continues. Une fois les modèles entraînés, ils ont été déployés sur la plateforme web à l'aide du Framework Python Flask, en vue de mettre les résultats à la disposition du grand public. Pour illustrer notre modèle, nous avons utilisé 6 variables susceptibles de déterminer les meilleurs prédicteurs. Une fois le modèle entraîné, il peut être utilisé et sauvegardé pour un déploiement ultérieur.

Mots clés : Performance académique, analyse quantitative, données éducatives, université de l'Assomption au Congo, Machine Learning

Citation: Mutsotsya, S.P., Mpia, H.N., Nzanzu, M.M., Baelani, I.N., Kasolene, M.K. Prédiction des notes finales des étudiants en fin du premier cycle : Utilisation de Data Mining éducatif. *Etincelle*, 2024, Vol. 25, no. 2. https://doi.org/10.61532/rime252114

Reçu : 15/10/2023
Accepté : 01/09/2024
Publié : 06/09/2024

Note de l'éditeur: Ishango-uac reste neutre en ce qui concerne les revendications juridictionnelles dans les cartes géographiques publiées et les affiliations institutionnelles des auteurs.

1. Introduction

La vérité scientifique et les compétences intellectuelles avérées constituent un facteur important pour sortir de l'ignorance des valeurs. Ainsi, l'échec scolaire est un problème qui peut freiner cette réussite (Hoffait & Schyns, 2017). En plus, les institutions supérieures ne se lassent plus en cherchant à contribuer à l'amélioration de la qualité de l'enseignement supérieur et du coup celle des performances des étudiants (Osmanbegović & Suljić, 2012). Afin de fournir une éducation de qualité aux étudiants, une analyse approfondie des résultats académiques des étudiants peut jouer un rôle capital.

Une façon d'améliorer ainsi les conditions d'encadrement des étudiants est de décrire les différents niveaux de performance des étudiants. Cette analyse peut se focaliser sur les différents grades obtenus par les étudiants en fin de chaque année du premier cycle. Notons que les données qui nous serviront seront extraites dans les bases des données éducatives se trouvant dans des universités, plus particulièrement, les données provenant du système de l'Université de l'Assomption au Congo (UAC). Ces connaissances et cet aperçu peuvent profiter non seulement aux étudiants mais aussi aux institutions voire aux enseignants (Romero & Ventura, 2010). L'application de Data Mining (DM) dans l'éducation est communément appelée Educationnel Data Mining (EDM) (Davies et al., 2021). L'EDM a suscité un intérêt scientifique en raison de son potentiel pour l'éducation (Tair & El-Halees, 2012). L'EDM fait allusion à l'analyse et à l'amélioration des méthodes de description des performances des apprenants.

Depuis 2021, l'UAC fonctionne avec un système intégré de gestion des activités académiques. Ce système permet tout d'abord la gestion des cotes des étudiants, ensuite la publication des résultats dans les comptes des étudiants, puis la gestion des présences au cours, la gestion des profils des enseignants, il intègre également un système de reconnaissance faciale lors des examens, la gestion des inscriptions etc. Le constat est qu'en République Démocratique du Congo (RDC), la question de la performance académique au premier cycle semble très cruciale étant donné que les étudiants venant fraichement de l'école secondaire et/ou du monde professionnel se soumettent incontestablement à une migration des conditions estudiantines différentes. Communément, la performance académique d'un étudiant est mesurée par ses grades antérieurs, en anglais on parle de Cumulative Grade Point Average (CGPA), ce qui signifie Moyenne Pondérée des Notes (Ashrafa, 2018).

Par ailleurs, il existe d'autres métriques pour déterminer les performances académiques des étudiants, certains facteurs ont été retenus dans les études antérieures dont les facteurs biologiques (différences d'Age), socio-psychologiques (étudiants introvertis et étudiants extravertis), culturels (étudiants issus des familles instruites et étudiants issus des familles non instruites), économiques (familles aisées et familles pauvres), mais aussi géographique (étudiants natifs de la ville où se trouve l'université et étudiants étrangers de cette ville) (Han & Rideout, 2022). Ces facteurs prédisent les performances académiques des étudiants. Curieusement, à l'UAC, les informations sur les performances académiques restent encore moins connues, c'est-à-dire l'on n'a pas la capacité de percevoir d'avance ce que pourrait être le grade d'un étudiant à la fin de son cycle. Vue que les variables susceptibles de mener à une bonne prédiction ne sont pas répertoriées dans le système de l'Université, et compte tenu de l'agencement du nouveau système Licence Master Doctorat (LMD), nous nous sommes contenté, pour cette étude, d'analyser les performances académiques en utilisant les techniques supervisées de Data Mining, en se servant des données secondaires issues des archives et du système de Gestion des activités académiques de l'UAC.

Bien qu'il existe nombreuses techniques heuristiques de prédiction des performances académiques des étudiants, il s'avère qu'il est difficile de prédire avec précision ce phénomène académique vu la diversité des facteurs qui peuvent intervenir pour capturer le contexte qui favorise les succès académiques des étudiants. Par conséquent, cet article a réalisé une étude comparative des algorithmes (individuels) déjà existants dans la revue de littérature empirique utilisés dans l'analyse des performances académiques tout en les comparant avec les méthodes ensemblistes et les algorithmes de régression. Cette analyse comparative s'est basée sur les données concrètes des résultats des étudiants de l'UAC.

A l'UAC, il se constate un problème majeur des mauvais résultats des étudiants, surtout ceux des nouvelles promotions du système LMD. En raison des mauvais résultats de

fin de l'année, nombreux étudiants arrivent à abandonner les études et même la délibération avec des échecs lamentables ne permettant pas le jury de délibération de réaliser sa mission avec rapidité. Certains étudiants de l'ancien système éducatif congolais en échouant sont obligés de reprendre la première année de Licence du système LMD vu que l'UAC organise ce dernier temps les deux systèmes en son sein et qu'elle veut finir progressivement avec l'ancien système. Ce va-et-vient des étudiants dans ces deux systèmes dus aux échecs observés rend la tâche académique complexe au niveau des facultés.

On voit là que les performances académiques des étudiants jouent un rôle important dans la gestion entière d'une institution supérieure (Han & Rideout, 2022) et que l'échec de l'étudiant ne pénalise pas seulement l'étudiant et ses parents mais aussi le bon fonctionnement d'une institution académique. La décision des départements par rapport aux performances observées aura une nécessité dans la définition des nouvelles stratégies d'enseignement. Sur ce, les questions et objectifs de cette recherche sont repris dans le tableau 1 :

Tableau 1. Objectifs Spécifiques et questions de recherche

Objectifs Spécifiques	Questions de recherche
1. Collecter les données académiques sur les étudiants de l'UAC, données se trouvant dans le système de gestion académique de ladite université	1. Sont-ce les données collectées au sein de l'UAC capables d'aider à construire un modèle de prédiction du grade final d'un étudiant ?
2. Capturer les variables qui prédisent le mieux la réussite d'un étudiant en fin de son premier cycle en usant de la technique de feature selection.	2. Quelles sont les variables qui prédisent contextuellement les notes finales des étudiants de l'UAC ?
3. Construire des modèles ML et les comparer afin d'en valider le plus performant pour prédire le grade final d'un étudiant.	3. Quel algorithme ML est mieux pour pré visualiser les notes finales des étudiants de l'UAC ?
4. Déployer le modèle validé comme prototype basé web	4. La technique web est-elle favorable pour déployer un modèle ML comme solution ?

2. Revue de littérature

2.1. Revue de littérature théorique

Cette section élucide des notions sur la performance académique, le ML à titre illustratif les algorithmes utilisés pour l'apprentissage automatique comme l'arbre de décision, apprentissage supervisé et non-supervisé, l'apprentissage par renforcement, les réseaux de neurone, la régression linéaire, l'apprentissage automatique, les méthodes ensemblistes comme le starking, le Boosting et le bagging, le random forest ou foret aléatoire et enfin une idée sur les indices statistiques qui seront mis en relief.

2.1.1. Performance académique

La formation intégrale de tout homme et de tout l'homme que les universités cherchent à promouvoir au sein de la société s'avère une priorité pour le leadership. La qualité des résultats des apprenants fait la réputation des établissements d'enseignement et constitue l'un de leurs indicateurs de qualité (Abdullah & Mirza, 2019). Ainsi la performance académique est le niveau potentiel du succès d'un étudiant conformément à sa situation individuelle. Elle renvoie au niveau de maitrises des savoirs à chaque étape du cursus scolaire. Pour définir une performance académique, on doit étudier plusieurs paramètres d'influence. La performance académique est consécutive à la méthode d'évaluation adoptée par le système du milieu académique. On peut parfois appliquer comme méthodes d'évaluation, la méthode d'évaluation individuelle et la méthode d'évaluation collective.

Il est vrai que, pour réussir à l'université certains indices doivent être au rendez-vous. Ici je peux faire allusion au sens du devoir, l'autodiscipline et l'autodétermination (Benoit-Chabot & Denis, 2018).

La performance, dans notre contexte est liée au rendement intellectuel de l'étudiant en fin du premier cycle. Elle n'est liée à aucun contexte matériel. Plus pratiquement le verbe « performer » en anglais « to perform », revêt beaucoup plus de signification qu'en français (Saoudi & Chroqui, 2017). En anglais, il fait allusion, selon les origines, à l'accomplissement, à la réalisation et aux résultats réels. Ce terme est apparu en anglais vers le 15ème siècle. Du latin « Parformer », il signifie « accomplir, exécuter ». Cette étymologie date du 13ème siècle. Au 19ème siècle, il sera beaucoup plus appliqué aux exploits sportifs. Au 20ème siècle, par ailleurs, il va être appliqué à la capacité de production des machines ou des véhicules (Saoudi & Chroqui, 2017). Depuis toutes ces années, l'on peut comprendre donc que le concept « performance » a connu beaucoup de mutations sémantiques et au cours du 20ème siècle, il sera beaucoup plus appliqué dans le domaine des sciences économiques et de gestion ; il sera donc utilisé pour désigner les résultats satisfaisants des politiques macro-économiques et financières des entreprises (Saoudi & Chroqui, 2017).

Au-delà de cette compréhension générique du concept de la performance, l'attention a été principalement orientée vers la performance académique ou il nous sera intéressant de définir ce qui est en rapport avec le rendement scolaire. La définition de la performance scolaire diffère d'un chercheur à l'autre. Certains chercheurs se sont concentrés sur les conditions scolaires, les résultats scolaires des étudiants, la persistance des rendement des élèves au fil du temps, acquisition des compétences minimales, acquisition des compétences de développement personnel de l'élève, progrès au-delà des attentes, l'équité entre tous les élèves, l'efficacité scolaire et l'efficience (Saoudi & Chroqui, 2017). Vue la définition multidimensionnelle du concept performance scolaire, nous avons considéré ce concept comme étant le niveau d'autoréalisation académique d'un étudiant en fin du premier cycle. Cela s'exprime mieux par le sentiment d'auto-efficacité c'est-à-dire de l'efficacité personnelle où l'individu se met à se mouvoir complètement pour atteindre les résultats souhaités en fin de période. Bandura dira que le sentiment d'auto-efficacité est considérée comme étant la croyance d'un individu de savoir respecter le règlement en vue d'atteindre des résultats satisfaisants (Bandura, 2007)

2.1.2. Grades

Dans le contexte où l'UAC fonctionne, les grades sont des qualifications attribuées à chaque résultat obtenu à la fin d'un exercice académique. A l'UAC, on distingue sept types des grades représentés chacun par une lettre Alphabétique majuscule. Il s'agit de A,B,C,D,E,F,G. A représente *Excellent* et concerne le résultat qui varie entre 90 et 99%, B=*très Bien et* se situe entre 80 et 89%, C=*Bien* pour le résultat entre 70 et 79%, D=*assez-bien* avec comme intervalle 60 à 69 %, E=*Passable* et se trouve entre 50 et 59%, F=*insuffisant* et se situe entre 40 et 49%, G : *insatisfaction* pour tous les résultats en dessous de 40.

2.1.3. Apprentissage automatique

L'apprentissage automatique, en anglais Machine Learning, est un sous domaine de l'intelligence artificielle qui a comme objectif de comprendre la structure des données en vue de les intégrer dans des modèles susceptibles d'être utilisés et standardisés dans les différents contextes de la vie courante. L'apprentissage automatique diffère des approches traditionnelles où les algorithmes constituent des ensembles d'instructions qui sont programmées de façon explicite et dans le but d'opérer des calculs pour résoudre des problèmes spécifiques (Barakati et al., 2020). L'apprentissage automatique, cherche en outre,

à utiliser des algorithmes qui permettent aux ordinateurs d'apprendre, à partir des données et des expériences, utilisant parfois des méthodes statistiques pour automatiser des processus et pour aider à la prise des décisions (Barakati et al., 2020). L'apprentissage automatique se caractérise surtout par l'acquisition et l'intégration des connaissances pour avoir une expérience pour un certain nombre des tâches et de performance (Barakati et al., 2020). Dans cette perspective, plusieurs techniques sont au rendez-vous. Il s'agit des techniques permettant ainsi de construire des modèles correspondant à la problématique d'étude. L'apprentissage automatique peut donc être décrit comme suit :

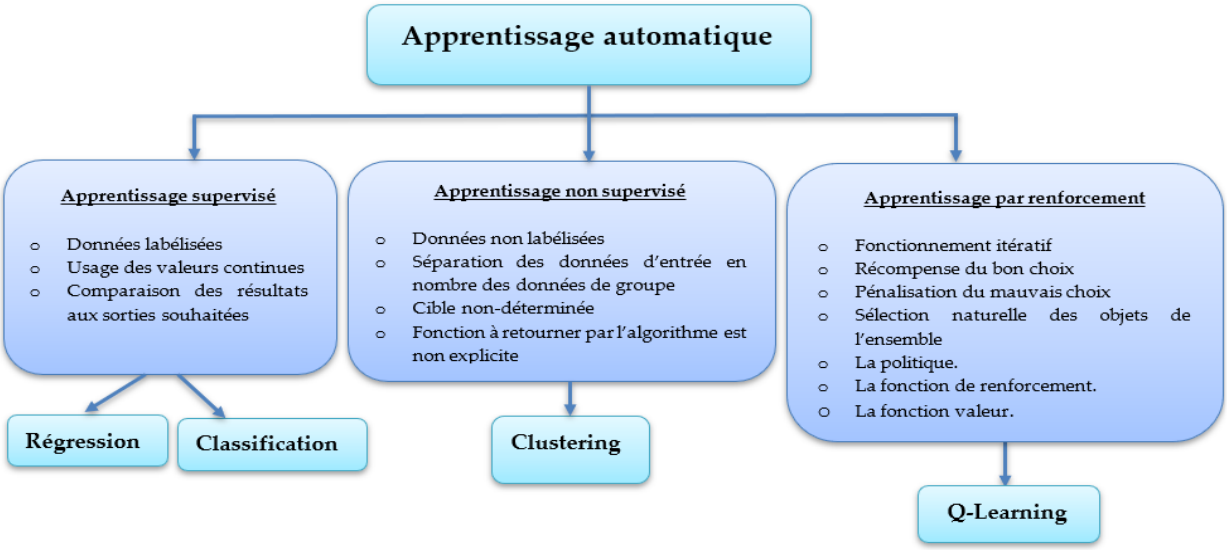

Figure 1. Schéma retraçant les différentes techniques de machine Learning

2.1.4. Apprentissage Supervisé

Dans le contexte classique de l'intelligence artificielle, l'on prend la perspective de modéliser les aspects du monde humain pour doter à la machine des capacités à se promouvoir l'impression d'être intelligente au moment de résoudre un problème qui a trait avec les attitudes humaines (Mathivet, 2014). Par ailleurs, en Machine Learning, la notion de l'apprentissage supervisé fait référence à l'entrainement de la machine à partir de certaines expériences c'est-à-dire on cherche à fournir à la machine des données historiques pour qu'il essaie d'anticiper tant soit peu les phénomènes futurs. Dans ce cas, la machine est dotée de beaucoup d'exemples qui peuvent être aussi appelés des données d'entrainement comportant des données d'entrée (features) et des données de sortie envisagées (target). Ces dernières peuvent être aussi appelées des signaux de surveillance (de Saint-Cirgue, 2019).

2.1.5. Apprentissage non supervisé

Un des types des techniques de ML. Dans ce type d'apprentissage, les variables ne sont pas catégorisées en termes de dépendance ou d'indépendance. Ici la finalité est de regrouper les éléments selon le niveau de rapprochement et la cohérence de leurs caractéristiques. Dans ce type, il n'y a pas de features (variables indépendantes) ni de target (Variable dépendante). On parle ainsi du Clustering ou d'autres méthodes de regroupements (Mathivet, 2014).

2.1.6. L'apprentissage par renforcement

Dans le contexte de l'apprentissage par renforcement, on indique à l'algorithme d'évaluer si la décision prise est bonne pas de manière autonome. Dans ce type d'apprentissage, l'on cherche à initier à la machine d'analyser les comportements de l'environnement avant de prendre des suites des décisions suivant les différents faits (Mathivet, 2014). Ce type d'apprentissage relève moins d'un contexte où l'on a besoin de fournir à l'algorithme des exemples ou des expériences pour avoir une solution mais plutôt d'entrainer le modèle à des expériences de choix stratégiques pour opérer des bonnes décisions et de rejeter les mauvaises. L'apprentissage par renforcement fonctionne de manière itérative et à chaque itération le modèle réalise une tâche et un score lui est attribué en fonction de ses performances. Les paramètres du modèle sont ensuite légèrement modifiés et la tâche est répétée. Un meilleur score oriente le choix des paramètres vers certaines valeurs, assurant que le modèle se rapproche à terme de résultats satisfaisants. Le choix des récompenses et pénalités est donc très important car il conditionne les résultats du modèle final en lui permettant de réaliser et d'améliorer sa prise de décision (Naeem, 2020).

2.1.7. Modèles Machine Learning

Un modèle Machine Learning consiste à construire des raisonnements sous forme d'algorithmes en vue de rendre possible des prédictions en utilisant des variables dites indépendantes pour déterminer ainsi une ou des variables cibles. Une prédiction correspond donc à des fonctions mathématiques sur des variables prédictives d'une observation X (Lemberger, 2015). Le Machine Learning est donc une démarche fondamentalement pilotée par les données. Il sera utile pour construire des systèmes d'aide à la décision qui s'adaptent aux données et pour lesquels aucun algorithme de traitement n'est répertorié (Lemberger, 2015). Plusieurs algorithmes servent ainsi de créer des modèles Machine Learning. Dans cette recherche, nous avons revu seulement quelques-uns des algorithmes Machine Learning de type supervisé que nous avons utilisés pour construire nos différents modèles.

A) La régression linéaire

Pour le problème de régression, la préoccupation majeure est celle de trouver, lors de l'entrainement du modèle des meilleurs paramètres. La régression linéaire se manifeste ainsi par une droite représentant des valeurs observées et permettant de minimiser les écarts entre ses dernières et les valeurs attendues (Kang & Zhao, 2020). La régression linéaire est multi typée. On trouve notamment: (i) la régression multiple standard qui permet de fournir simultanément les variables indépendantes dans l'équation de régression sans aucune distinction les unes des autres, (ii) la régression multiple séquentielle : dans ce type, les variables indépendantes sont fournies de manière séquentielle dans l'équation de régression. Cette régression est aussi dite hiérarchique. Bref, les variables sont ajoutés progressivement au modèle en vue de voir l'évolution progressive de l'amélioration du modèle. Quant à la régression multiple statistique : celle-ci permet d'entrer ou de sortir séquentiellement des variables indépendantes une à une de l'équation de régression tout en se focalisant sur un critère statistique (Maulud & Abdulazeez, 2020). Dans cette même optique, l'estimation des meilleurs paramètres reste l'option majeure pour trouver les valeurs optimales. La régression s'adresse à un type de problème où les 2 variables quantitatives continues X et Y ont un rôle asymétrique : la variable Y dépend de la variable X. La liaison entre la variable Y dépendante et la variable X indépendante peut être modélisée par une fonction de type $Y = \alpha + \beta X$, représentée graphiquement par une droite (Labarere, 2012).

B) L'Arbre de décision

On appelle arbre de décision un modèle de prédiction qui peut être représenté sous la forme d'un arbre. Chaque nœud de l'arbre teste une condition sur une variable et chaque subalterne de ce nœud correspond à une réponse possible à cette condition. Les feuilles de l'arbre correspondent à une étiquette. Pour prédire l'étiquette d'une observation, on suit les réponses aux tests depuis la racine de l'arbre, et on retourne l'étiquette de la feuille à laquelle on arrive (Azencoth, 2018).

L'arbre de décision est un type d'algorithme qui respecte la norme de l'apprentissage hiérarchique. L'arbre de décision se comporte avec une succession des tests conditionnels et chaque tests a une dépendance significative sur le précédant ; il sert aussi à étiqueter une observation (Azencoth, 2018). Il peut avoir comme avantages : (i) traiter des attributs discrets (comme ici la forme, la taille et la couleur), sans nécessiter une notion de similarité ou d'ordre sur ces attributs (on parle d'apprentissage non métrique) ; (ii) permettre de traiter un problème de classification multi-classe sans passer par des classifications binaires ; (iii) permettre de traiter des classes multimodales (comme ici pour l'étiquette « pomme », qui est affectée à un fruit grand et rouge ou à un fruit jaune et rond) (Azencoth, 2018). En outre, il est beaucoup plus important pour le *data scientist* de savoir structurer de façon arborescente un modèle en vue de s'assurer d'un meilleur choix de possibilité. Les arbres de décision sont faciles d'interprétation et de compréhension et permettent une prédiction plus transparente. Ils sont capables de prédire des valeurs numériques tout comme des valeurs catégorielles. L'arbre de décision reflète un caractère déterministe (Gruz, 2015).

C) Random Forest

En pratique, les forêts aléatoires (RF) sont un type d'algorithmes les plus performants et les plus simples à mettre en place. Le RF trouve sa pertinence dans la faible dépendance a ses hyper paramètres (Azencoth, 2018). Le RF sont un remède pour les arbres de décision en ce sens que ces derniers présentent un sur ajustement des données d'apprentissage (Gruz, 2015). Les Forets aléatoires constituent donc un ensemble de plusieurs arbres de décision construits et qui cherchent à trouver par eux-mêmes le mode de classements des entrées (Gruz, 2015). Pour y arriver, certaines techniques sont toujours au rendez-vous pour fusionner les différents tests opérés sur les données entrainées sur chaque arbre des décisions dont les données sont souvent différentes. Nous pouvons illustrer ici le Boostrap_sample pour l'entrainement de chaque arbre. Le Boostrap_aggreging, quant à lui permet d'entrainer les différents arbres de décision sur des données non échantillonnées (Gruz, 2015).

D) Machines à vecteurs de support

Les machines à vecteurs de support, en anglais Support Vector Machines (SVM), sont de puissants algorithmes d'apprentissage automatique. Ces algorithmes sont focalisés sur un algorithme linéaire proposé par Vladimir Vapnik et Aleksandr Lerner en 1963, mais sont bien appropriés pour apprendre bien plus que des modèles linéaires (Chervonenkis, 2013). En effet, au début des années 1990, plusieurs chercheurs ont trouvé comment étendre les SVM efficacement à l'apprentissage de modèles non linéaires grâce à l'astuce du noyau. Cette approche permet de (i) définir un classifieur à large marge dans le cas séparable ; (ii) écrire les problèmes d'optimisation primale et duale correspondants ; et (iii) réécrire ces problèmes dans le cas non séparable. Le SVM utilise l'astuce du noyau pour appliquer le traitement à marge souple dans le cas non linéaire. Définir des noyaux pour des données à valeurs réelles ou des chaînes de caractères est une des étapes essentielles de cet algorithme (Azencoth, 2018).

E) Les méthodes ensemblistes

Les méthodes ensemblistes sont des méthodes ML très puissantes en pratique, qui reposent sur l'idée que combiner de nombreux algorithmes qui, de fois peuvent présenter certaines imperfections (Mpia et al., 2022). Les méthodes ensemblistes adoptent vraisemblablement la politique qu'a l'algorithme RF qui combine des arbres de décisions particulières pour aboutir à un résultat plus performant. Les méthodes ensemblistes, quant à elles vont associer certains algorithmes dans leurs totalités substantielles pour rechercher une performance largement supérieure aux performances individuelles de chaque algorithme (Natras, 2022).

Cette méthode ensembliste permet aux erreurs des uns et des autres algorithmes de se compenser mutuellement pour ainsi les minimiser considérablement. Cette idée est similaire au concept de sagesse des foules. A titre illustratif, si on demandait à cinq étudiants la date officielle du basculement de l'ancien système éducatif Congolais vers le LMD, il est probable que la moyenne de leurs réponses soit assez proche de la bonne; cependant, si l'on demandait cette même date à une seule personne au hasard, on n'aurait aucun moyen de savoir a priori si cette personne connaît la date ou répond au hasard. Pour les méthodes ensemblistes l'idée c'est que tous les algorithmes utilisés pour prédire un fait, sont pris chacun individuellement et mis en commun pour prédire la même réalité(Gruz, 2015).

2.2. Revue de littérature empirique

Dans cette partie, les auteurs se sont focalisés sur des recherches ayant trait à notre étude afin d'y identifier des lacunes conceptuelles et de situer notre recherche dans une perspective nouvelle. Ainsi, la revue de littérature empirique a révélé que Kouakou et Kablan (2015), ont fait l'analyse les facteurs déterminants de la performance scolaire des établissements. Ils ont utilisé des données allant du secondaire jusqu'au baccalauréat (bac). Pour fixer une idée sur leur recherche, ils ont réussi à réunir quelques facteurs de la performance scolaire en identifiant les facteurs notamment l'origine sociale des élèves, le type d'école fréquenté, des enseignants mieux formés, l'effet-classe (Groupe Homogène, Groupe hétérogène, Environnement, Composition du public accueilli), de l'effet-établissements (Groupe Homogène Groupe hétérogène Environnement Composition du public accueilli) (Kouakou & Kablan, 2015). La variable dépendante, dans leur recherche était le taux de réussite au bac pour étudier la disparité des performances entre établissement. Les variables indépendantes ont été regroupées en deux sous-groupes notamment le premier qui est lié à l'effectif des élèves au second cycle, le nombre de groupes pédagogiques, la proportion de redoublants, le taux d'encadrement et le nombre de candidats inscrits à l'examen du bac. Le second groupe réunissait les variables extérieures à l'école. Il s'agit notamment de l'environnement de l'école (urbain vs rural) et le régime de l'établissement (mixte, garçon ou fille) (Kouakou & Kablan, 2015).

Par contre, Lardy et al. (2015), dans leur recherche ont effectué une étude sur la performance individuelle des étudiants au sein des universités françaises. Dans leur conviction, le type de baccalauréat et le taux de réussite sont deux variables très dépendantes. Dans le cadre conceptuel de leur recherche, ils ont distingué trois types des variables, notamment les variables motivationnelles (le sentiment d'auto-efficacité, buts et valeur, les variables comportementales (habileté technique et bonne pratiques estudiantine), les variables indiquant les trajectoires sociodémographiques et académiques (milieu familial et académique) (Lardy et al., 2015). L'objectif poursuivi de ces chercheurs est de vouloir comprendre le soubassement de la réussite des étudiants dans le domaine des technologies.

Ce soucis fait que jusqu'à présent, la préoccupation majeure d'une institution, soucieuse de se qualifier, reste la recherche quotidienne des vraies variables prédicteurs de la performance académique (Abdullah & Mirza, 2019). Leur article s'est focalisé sur certaines

variables notamment l'auto-efficacité académique, la motivation des étudiants, le statut socio-économique, les évaluations du noyau dur de soi, l'approche de la maîtrise, l'accomplissement de la démarche, l'orientation vers un but, l'assiduité, la personnalité, le quotient intellectuel (QI), les habitudes d'étude, l'attitude et les aptitudes, les qualifications universitaires et professionnelles des enseignants, ainsi que d'autres variables démographiques (Abdullah & Mirza, 2019). Du point de vue méthodologique, leur étude a été menée en utilisant la corrélation de Pearson et l'analyse de régression multiple. Par rapport aux résultats obtenus, les corrélations dans tous les types de programmes, notamment le programme de quatre ans et celui de deux ans, étaient significatives et positives et positives, les valeurs *r* avaient montré la corrélation la plus élevée pour les programmes de quatre ans (r= 0.413) par rapport aux programmes de deux ans (r= 0.300). Le même schéma a été observé dans les prédictions à l'aide d'analyses de régression multiple et les résultats des examens cumulatifs précédents comme prédicteurs. Les notes d'examen cumulatives antérieures ont permis de prédire (17,2 %) dans les programmes de baccalauréat en quatre ans avec des valeurs bêta normalisées et avec des valeurs bêta standardisées de (.156 et .320, p< 0.01) pour les notes du baccalauréat et les notes intermédiaires, respectivement. Une prédiction de 8,9 % a été observée pour les programmes de deux ans avec les valeurs bêta standardisées de (.160 et .190, p< 0.01) pour les scores matriculaires et intermédiaires, respectivement (Abdullah & Mirza, 2019).

3. Méthodes et matériels

Cette recherche a procédé par la collecte des données secondaires évidement relatives aux différentes notes obtenues par les étudiants ayant parcouru un cursus allant de la première année de graduat jusqu'à troisième. Cela étant, cette recherche se veut quantitative. Cette technique va nous permettre d'estimer les statistiques de réussite des étudiants au premier cycle, en les quantifiant selon les différents parcours des étudiants. Bien plus, dans le contexte du Data Mining, la méthode prédictive a été utilisée.

3.1. Conception de la recherche

Dans le choix de notre méthode, tenant compte de la nature de notre recherche, nous avons retenu la méthode quantitative. La méthode peut être considérée comme un prérequis de la collecte des données. Pour notre étude, nous avons pris en considération des données de type secondaire. Il s'agit pour nous donc des données secondaires vue qu'elles n'ont pas été obtenues à partir d'un quelconque questionnaire. Au-delà de cette collecte, d'autres données récentes ont été recueillies à partir du site web de l'UAC. A l'issue de cette démarche, nous avons procédé par le prétraitement desdites données, ensuite l'analyse exploratoire de ces données, puis la modélisation, puis encore l'optimisation des modèles, enfin le déploiement du modèle dans une technologie web. Ainsi avons-nous utilisé certains algorithmes pour arriver à nos résultats. Toutes ces étapes peuvent être structurées selon le plan suivant :

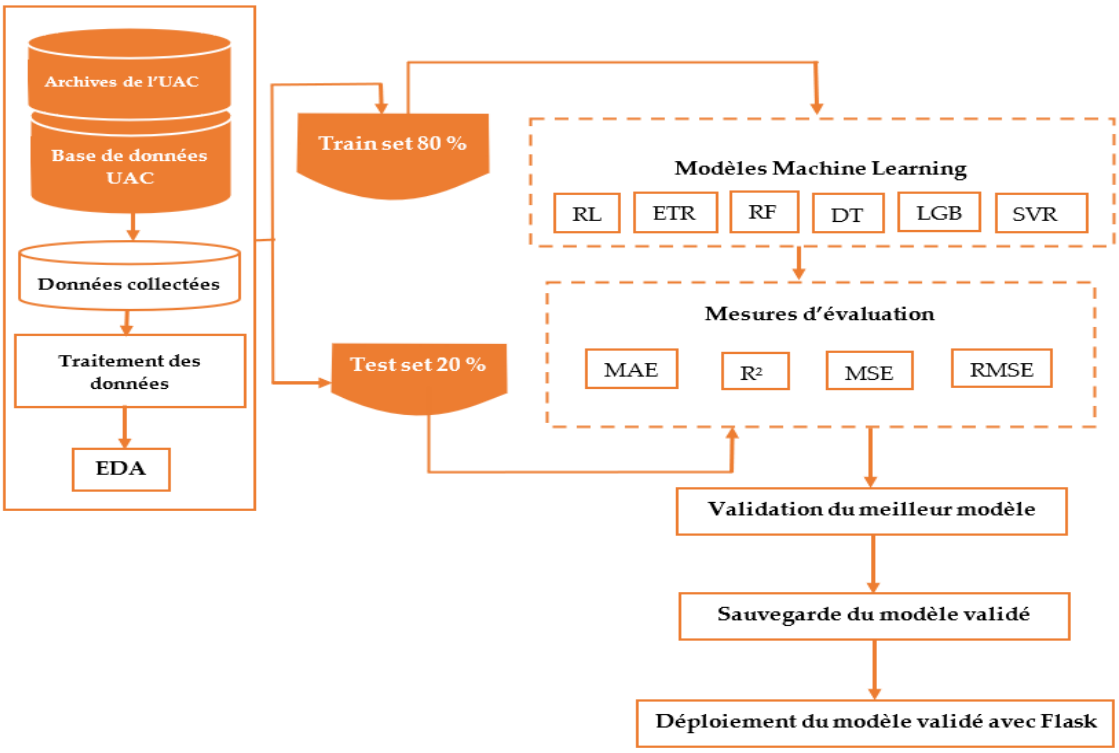

Figure 2. Schémas de la succession des étapes de réalisation de cette recherche

La représentation ci-haut reprend la suite logique des étapes à franchir pour réaliser les objectifs de notre étude. Ceci commence par la récolte des données présentes d'une part dans les archives de l'UAC et d'autre part dans le système de gestion des cotes initiées toujours dans la même institution. Les données sont traditionnellement perçues comme les prémisses des théories. Les chercheurs recherchent et rassemblent des données dont le traitement se réalise par une instrumentation méthodique va produire des résultats et améliorer, ou renouveler, les théories existantes (Thiétart, 2014)

Après cette extraction des données, est venu, ensuite, le prétraitement ou *pre processing* des données qui permet aux analystes de pratiquer le nettoyage des données qui est le processus d'identification des parties incorrectes, incomplètes, inexactes, non pertinentes ou manquantes des données, puis leur modification, leur remplacement ou leur suppression en fonction de la nécessité, à l'harmonisation de leurs natures en vue d'une bonne analyse préalable. Cette étape a été la plus cruciale dans notre recherche en ce sens qu'elle s'était basée sur des données dont les origines étaient hétérogènes à notre environnement de travail. Bien que cette activité prenne beaucoup de temps, elle a des retombées importantes, car une bonne compréhension des données permet d'obtenir des modèles plus performants et de comprendre pourquoi les prédictions sont faites (Harrison & Petrou, 2020) Dans ce contexte, l'analyse exploratoire des données (EDA) nous a permis de répertorier toutes les valeurs manquantes qui existent dans notre base des données qui n'est rien d'autre que le dataset que nous avons constitué lors de l'extraction de nos data. Nous précisons que cette étape nous offre en amont, les possibilités de se rassurer de l'état général des données en procession, de leur compréhension sémantique et statistique, avant de s'impliquer dans l'aventure de leur jeu d'entraînement.

Après avoir rendu prêt toutes les données, nous sommes passés maintenant à l'étape d'entrainement des données dans la phase, bien attendu, de la modélisation proprement dite. Dans notre cas concret, nous nous sommes donné la peine de comparer des algorithmes reposant sur la régression et les méthodes ensembles. Notons la problématique de notre recherche est resté centré sur un problème de régression parce que les données qui doivent être prédites correspondent à la logique des variables continues et les

valeurs de sortie constituent des données numériques. Dans ce processus, nous avons essayé de planifier la manipulation de nos données. La poursuite du processus a migré vers la subdivision du dataset en données d'entrainement et de test. Les auteurs ont entrainé nos modèles avec 80% des données tout en réservant 20% des données pour le test. Cette subdivision nous a paru conforme pour permettre à notre modèle de bien apprendre en vue d'une bonne prédiction. De ce fait, la tâche du Machine Learning s'est focalisée sur six algorithmes qui ont été expérimentés pour prédire la note finale d'un étudiant à la fin de son premier cycle. Il s'agit notamment *Régressions Linéaire (RL), Random Forest(RF), Support Vector Machine (SVR), Light LGB (LGB), Extra Tree Regressor (ETR) et l'arbre de décision (DT).* En guise d'évaluation de ces modèles, les auteurs ont utilisé comme mesures d'évaluation, l'erreur moyenne Absolue (MAE), l'erreur Moyenne Quadratique (MSE), le Root Mean Square Error (RMSE), le Coefficient de détermination (R^2).

3.2. Population Cible

La population cible constitue l'ensemble d'individu concernés par une recherche. La population cible peut être prise comme étant l'ensemble d'individus visés par une recherche dont on voudrait recueillir des informations et extrapoler ou généraliser les résultats. C'est la population étudiée ou le champ de l'enquête (Dussaix, 2009). Une recherche doit se baser logiquement sur une population qui peut être humaine ou non humaine. En statistique, une population est celle qui intéresse la recherche.

Ainsi, pour cette recherche, les auteurs ont choisi comme population, l'ensemble des cotes des étudiants au cours de leur cursus pendant trois ans. Ceci est la conséquence du choix des processus d'implémentation basé sur le problème de Régression. Comme indiqué au niveau de la littérature empirique, notre réflexion s'est beaucoup plus bornée sur les facteurs purement académiques pour prédire le grade final de l'étudiant à la fin de son cycle. Les auteurs ont compris que les cotes antérieures des étudiants constituent des variables les plus fiables pour prédire leurs performances postérieures. C'est cela même qui justifie le choix des cotes antérieures en tant que variables promotrices des vraies prédictions des grades finaux des étudiants.

3.3. Procédure de collecte des données

Dans le cadre de la recherche, les données peuvent être classées en deux catégories notamment les données primaires et les données secondaires. Ceci signifie que les données recueillies sont brutes en ce sens qu'elles sont directement le résultat d'une enquête. Les données secondaires quant à elles, sont extraites à partir des systèmes d'information divers pouvant exister dans les entreprises et qui permette l'exploitation facile des données existantes.

De ce fait, les données secondaires font l'objet d'un certain nombre d'idées reçues quant à leur authenticité, leur impact sur la validité interne ou externe, leur accessibilité et leur flexibilité. La plus tenace d'entre elles concerne sans doute leur statut authentique. Parce qu'elles sont formalisées et publiées, les données secondaires se voient attribuer un statut de « vérité » souvent élevé. Ainsi, on accorde une intégrité plus grande à une information institutionnelle qu'à une information privée de source discrétionnaire, sans même s'interroger sur les conditions de production de ces différentes données. Ce phénomène est accentué par l'utilisation des média électroniques qui fournissent les données dans des formats directement exploitables. La formalisation des données dans un format prêt à l'exploitation peut amener le chercheur à considérer pour acquis le caractère valide des données qu'il manipule (Thiétart, 2014). Bref, ces données sont généralement plus sûres et plus fiables dans la recherche comme telle, car elles sont souvent de grande quan-

tité et permettent une bonne étude de la réalité concrète des faits en se basant bien évidemment sur leur historique. Ces données secondaires sont de type interne parce que reposant sur les documentations exclusivement internes à l'entreprise.

3.4. Traitement et analyse des données

Dans la logique de data science, la donnée constitue l'élément le plus primordial dans la conception d'un phénomène. La donnée a comme soubassement la société qui est l'ensemble d'individus qui produisent des actions. Les données sont construites au sens où elles résultent d'un travail d'élaboration théorique du sociologue (et du statisticien) pour recueillir des échantillons des populations par rapport à un phénomène. Le chargé de ce processus doit chercher à définir des dimensions plus perspicaces du point de vue social, en se basant toujours sur la problématique de départ, les concepts permettant de se représenter la réalité étudiée, les catégories servant à coder les faits observés ainsi que les modalités de protocoles d'interview ou d'observation (Martin, 2009). Pour traiter nos données, nous nous sommes servis de certains modules python notamment *numpy*, *matplotlib*, *pandas*, *seaborn*, *Scikit-learn*.

Le langage python a été utilisé pour effectuer toutes les analyses et implémentations des modèles. Python fournit un mélange unique d'élégance, de simplicité et de puissance. Nous nous sommes servis des librairies de python dont Numpy, Matplotlib, Pandas, Seaborn et Scikit learn. Cette dernière librairie est un package d'apprentissage automatique dans le langage de programmation Python qui est largement utilisé dans la science des données. Le package Scikit-learn comprend la mise en œuvre d'une liste complète de méthodes d'apprentissage automatique sous des conventions de données et de procédures de modélisation unifiées, ce qui en fait une boîte à outils pratique pour les statisticiens de l'éducation et du comportement (Kam, 2023). En se servant de ces quelques librairies python, nous avons réalisé nos différents traitements et analyses pour rendre compréhensibles et manipulables nos données. Tous les détails sont repris avec illustrations dans les paragraphes suivants. L'objectif de l'encodage des données a été de rendre toutes les données numériques afin de bien effectuer la modélisation.

4. Résultats et discussion

4.1. Résultats de la recherche

Les résultats des prédictions ont montré que pour les données utilisées dans cette étude, l'algorithme de régression linéaire est le meilleur. Comme le montre le tableau 2 ci-dessous, les auteurs ont conclu que la régression linéaire était plus performante avec une MAE de 1,12, une MSE de 7,04, une RMSE de 2,64 et un coefficient de détermination de 0,80. Les résultats du tableau 2 montrent clairement que la valeur de l'erreur quadratique moyenne pour le modèle de régression linéaire est inférieure à celle des autres modèles. Cela signifie que l'algorithme a appris avec précision à partir de l'ensemble de données utilisé et qu'il a obtenu de meilleurs résultats. De plus, le coefficient de détermination R^2 de notre modèle est de 0,80. Cela signifie que 80% de la variabilité observée de la note finale d'un étudiant est capturée et apprise par le modèle et que les 20% restants sont dus à d'autres facteurs, qui peuvent être par exemple la santé de l'étudiant.

Tableau 2. Performances des différents modèles utilisés

N°	Modèles	MAE	MSE	RMSE	R2
1.	Régression linéaire	1.12	7.04	2.64	0.80
2.	Forêt aléatoire	2.25	8.74	2.96	0.77
3.	Arbre de décision	2.94	15.54	3.94	0.60
4.	Support Vector Regressor	1.95	7.60	3.94	0.80
5.	Light LGB	2.25	8.66	2.94	0.78
6.	Extra Trees Regressor	2.25	8.91	2.99	0.77

Du point de vue de la MAE, le modèle de régression linéaire a gagné en présentant une valeur de 1,12, suivi par SVM (1,95), Extra Trees Regressor, Light LGB et forêt aléatoire ont donné un même score de 2,25 et l'arbre de décision a été le mieu perforamant de tous. Concernant l'erreur quadratique moyenne (MSE), le modèle de régression linéaire est en première position, avec le MSE le plus bas, soit 7,04. Elle est suivie par le SVM avec une MSE de 7,60, puis Light LGB 8,66 et forêt aléatoire avec 8,74. L'arbre de décision a un MSE supérieure à 15. En général, l'abre de décision a eu des mauvais scores en tout. L'image ci-après illustre la synthèse du tableau 2 sous forme de graphique :

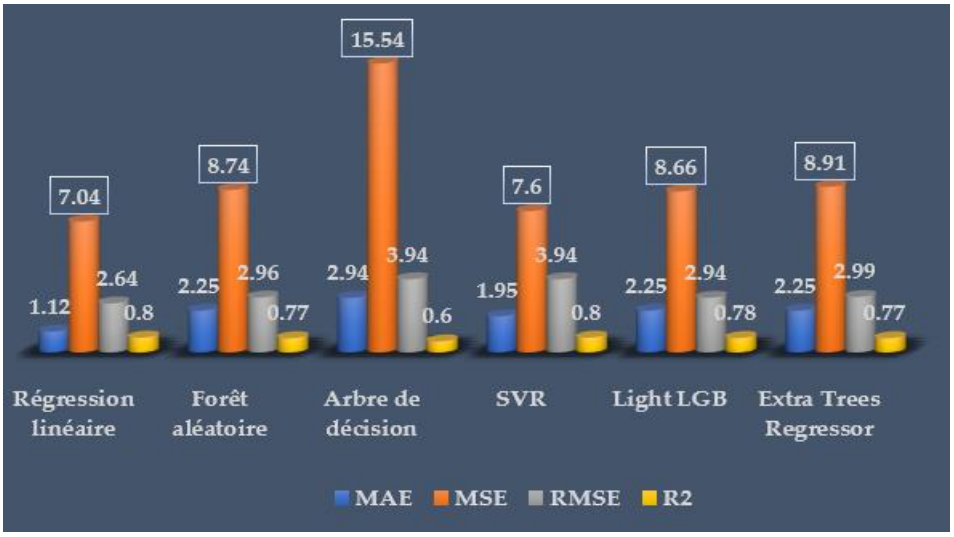

Figure 3. Graphique des performances des modèles

Après validation et sauvegare du modèle de régression linéaire, les auteurs sont passés au niveau du déploiement en utilisant la technologie web. Notons que l'interaction entre le langage python et la technologie web se réalise au moyen de plusieurs Framework. Pour cette application, le Framework Flask a été utilisé qui est beaucoup plus basé sur l'API Representational State Transfer (REST). Flask a deux composants principaux : Werkzeug et Jinja2. Alors que Werkzeug est responsable du routage, du débogage et de la passerelle de serveur Web Interface (WSGI), Flask utilise Jinja2 comme moteur de modèle. Nativement Flask ne prend pas en charge l'accès à la base de données, l'authentification des utilisateurs ou tout autre utilitaire de haut niveau, mais il prend en charge l'intégration d'extensions à ajouter toutes ces fonctionnalités, faisant de Flask un micro mais prêt pour la production (Relan, 2019). REST est un style architectural logiciel. Pour les services Web qui fournissent une norme pour la communication de données entre différents types de systèmes (Relan, 2019). Ainsi, il nous a permis de créer un contrôleur qui permettra à notre code python d'interagir avec les fichiers (.html) qui constituant la couche des vues. Notons que cette programmation respecte l'architecture MVC (Model

View Controller). Le fichier contenant le code Flask est de type (.py). la capture ci-dessous reprend la syntaxe Flask.

Scénario de cas d'utilisation

Le prototypage étant une des techniques souples de production d'artéfact en ingénierie, les auteurs l'ont pu utiliser afin d'évaluer leur étude. Ainsi, les lignes qui suivent présentent le scénario d'utilisation du prototype développé. En fait, Sédar est un étudiant à l'UAC. Il est en troisième année de Graduat (G3). Il veut, savoir quelle note finale il obtiendra en fin de sa dernière année. Sédar saisit l'url de l'application web proposée dans cette recherche dans son navigateur pour accéder à l'application. L'application ouvre la page d'accueil comme le montre la figure 4, et Sédar remplit le formulaire en fonction de son passé académique (résultats de G1 et de G2).

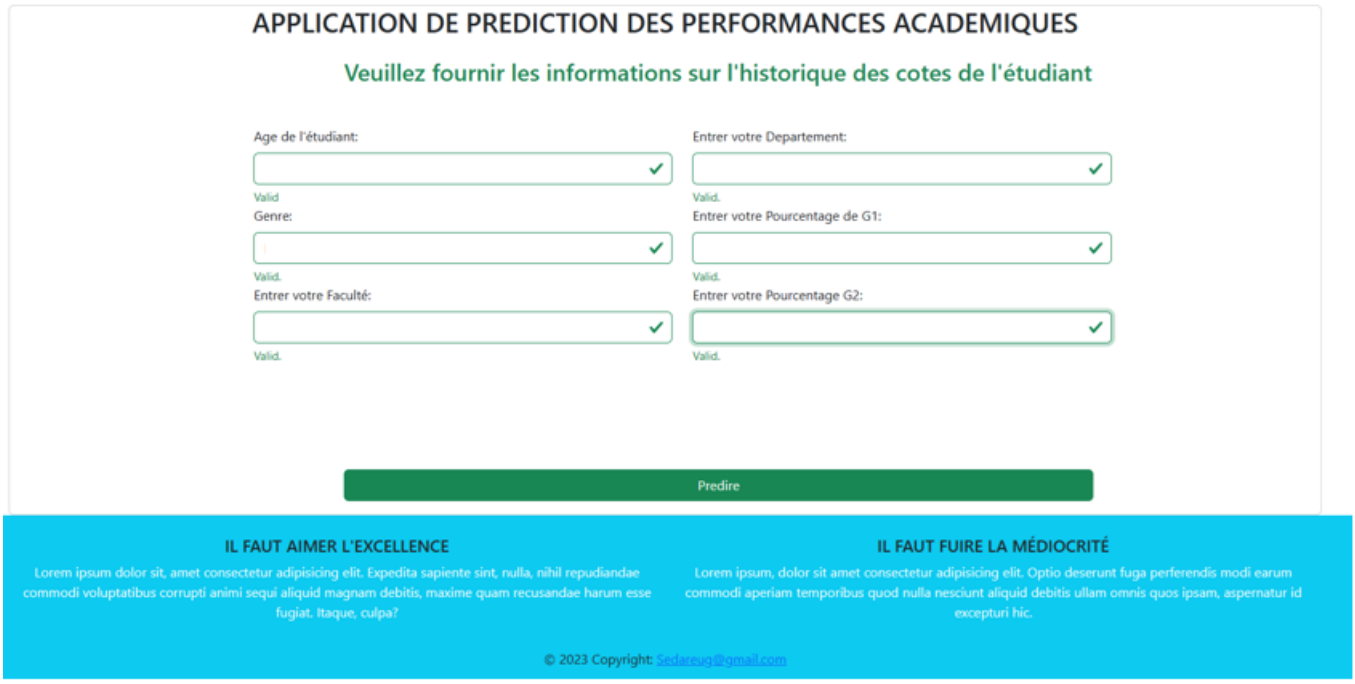

Figure 4. Page d'accueil du prototype

Dans la figure 5, on voit que Sédar a fourni son Age de 23 ans, son genre Masculin (M), sa faculté qui est celle des sciences économiques et de gestion, son département qui est l'informatique de gestion, sa note finale de G1 de 78,54% et celle de G2 de 75,68%. Après avoir inséré toutes ces informations sans laisser aucun champ vide, Sédar clique sur le bouton « Prédire » tel qu'illustrer dans figure 5 et l'application a généré le pourcentage qu'il pourra obtenir en fin de son premier cycle. Certes, cet étudiant pourra finir son cycle avec une note de 76.80%. Le système fait correspondre à ce pourcentage le grade approprié qui a été, pour ce cas, (C) qui signifie Bien.

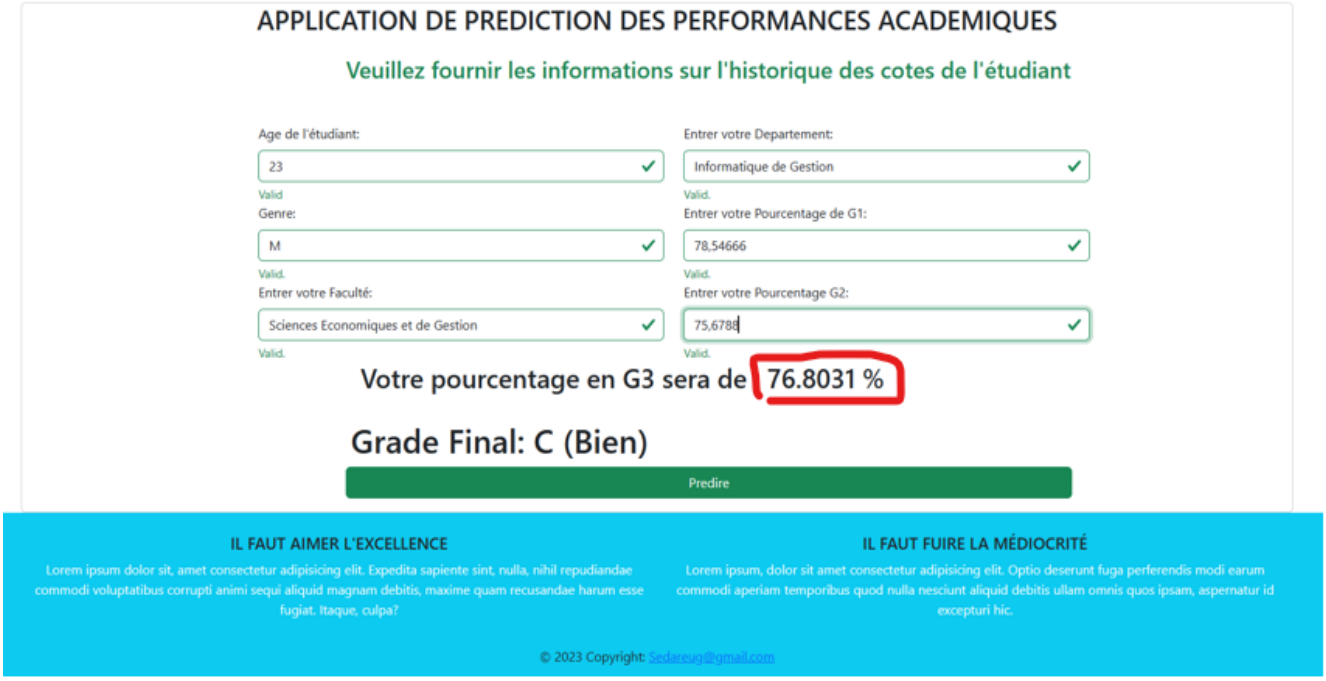

Figure 5. Résultat de la prédiction

4.2. Discussion des résultats

Les travaux de Kouakou et Kablan (2015) ont été analysés. Ces auteurs sont partis des enquêtes statistiques pour analyser le problème de performance académique. Ils se sont plus basés sur l'analyse descriptives des données sans utiliser aucun modèle machine Learning. Leur technique de récolte des données a été la fusion des données provenant de l'évaluation certificative du BAC et de la base des données du ministère, gérées par la Direction des Stratégies, de la Planification et des Statistiques (DSPS). Pour ce qui est des données de la DSPS, ils se sont basés sur des questionnaires contextuels (questionnaires élève, enseignant et établissement). Dans cet échantillon, la variable « Taux de réussite » varie d'un minimum de 0 % à un maximum de 100 %, pour une moyenne (Average) de 49,97 % et un écart-type (Std-dev) de 17,19 %. Ce qui signifie que nous avons environ 68 % des établissements dont le taux de réussite au BAC est compris dans un intervalle de 49,97 % + ou – 17,19 % (soit [32,78 % - 67,16 %]) (Kouakou & Kablan, 2015). Notons leur travail de recherche est beaucoup plus resté focalisé sur l'analyse descriptive et statistique des performances académiques.

Au sujet de étude recherche, les auteurs ont modélisé le problème et entrainer nos modèles moyennant huit algorithmes que nous avons essayé de comparer mutuellement en se basant sur leurs performances respectives. Il s'agit notamment de la régression linéaire, l'arbre de décision, le SVM, le Light LGB, la Forêt aléatoire et l'Extra Trees Regressor. Tous ces algorithmes ont été appréciés avec des bons scores, et peuvent donc être affectés pour le même travail. Toutefois, les auteurs se sont contentés de celui qui a eu le score le plus élevé et parmi eux, nous avons retenu deux qui ont abouti au même résultat de score notamment la régression linéaire et le SVM. Ils ont eu un score de 0,80 de R^2, qui est la métrique mieux située pour évaluer un modèle. Les performances de cette présente recherche semblent plus ou moins meilleures par rapport à ceux de nos prédécesseurs.

5. Conclusion

Cette étude s'est beaucoup focalisée sur la réalité concrète des grades/notes des étudiants de l'UAC. L'analyse étant basée sur la recherche quantitative, les auteurs ont présenté la recherche en tant que problème de régression. Alors que les travaux antérieurs sont plus orientés dans le problème de la classification où des résultats sont de type catégoriel en utilisant les données primaires, qu'ils récoltent au moyen des questionnaires d'enquête, cette recherche a fait usage des données secondaires se trouvant dans la base de données de l'UAC. Ces données servant de données de Data Mining éducatif ont permis d'obtenir la vraie réalité des cotes des étudiants. Aussi, la technique de feature selection a été utilisé pour séléctionné avec exactitude les variables qu prédisent mieux les performances académiques dans ce contexte. Des six algorithmes utilisés dans cette recherche, la régression linéaire a pu avoir les meilleures performances atteignant le MAE de 1,12, le MSE de 7,04, le RMSE de 2,64 et le R^2 de 0,80.

Sachant bien que l'univers de la recherche scientifique est trop large, l'on ne peut prétendre avoir épuisé tous les champs d'application de cette étude. Ainsi, les auteurs recommande au futur analyser d'autres variables supplémentaires qui peuvent aider aux meilleures performances des modèles développés. Pour parfaire cet artéfact, un système de recommandation serait d'une grande envergure de sorte qu'en prédisant chaque résultat, le système peut proposer des stratégies, qui, une fois appliquées, peuvent aider à l'amélioration des notes finales d'un certain étudiant.

Contributions: Conceptualisation, S.P.M. & H.N.M.; méthodologie, H.N.M.; validation, H.N.M. & N.I.B.; investigation, M.M.N. & M.K.K.; ressources, M.M.N.; traitement des données, H.N.M.; écrire le manuscrit, S.P.M. & H.N.M.; visualisation, M.K.K.; supervision, H.N.M. & M.M.N.; correction du manuscrit, H.N.M. Les auteurs ont lu et approuvé la version publiée de ce manuscrit.

Sponsor financier: Cette recherche n'a reçu aucun soutien financier.

Disponibilité des données: Les données ne sont pas disponibles.

Remerciement: Non applicable.

Conflits d'intérêt: Les auteurs déclarent aucun conflit d'intérêt.

Références

1. Ashrafa, S. A. (2018,). A Comparative Study of Predicting Student's Performance by use of Data Mining Techniques, American Scientific Research Journal for Engineering, Technology, and Sciences (ASRJETS), 44(1).
2. Azencoth, C.A. (2018). Introduction au machine Learning, Dunod, Paris.
3. Bandura, A. (2007). Auto-efficacité : Le sentiment d'efficacité personnelle, De Boeck, Paris.
4. Barakati, Y., et al. (2020). Réseaux de Neurones et Apprentissage Automatique des Systèmes Informatiques Intelligents,. International Journal of Innovation and Modern Applied Science, 3.
5. Benoit-Chabot, G., & Denis, P.L. (2018). Accroitre sa performance académique : le rôle des préférences d'apprentissage, Revue des sciences de l'éducation, 44(2). https://doi.org/10.7202/1058115ar.
6. Chervonenkis, A.Y. (2013). Early History of Support Vector Machines, Empirical Inference. https://doi.org/10.1007/978-3-642-41136-6_3.
7. de Saint-Cirgue, G. (2019). Apprendre le machine Learning en une semaine, Machinelearnia.com.
8. Davies, R., Allen, G., Albrecht, C., Bakir, N., & Ball, N. (2021). Using Educational Data Mining to Identify and Analyze Student Learning Strategies in an Online Flipped Classroom. *Educ. Sci.*, *11*, 668. https://doi.org/10.3390/educsci11110668.
9. Dussaix, A.M. (2009). La qualité dans les enquêtes, Revue Modulad (39).
10. Greene, W.H. (2005). Econométrie, 5e édition, Pearson Education, Paris.
11. Gruz, J. (2015,). Data Science par la pratique. Fondamentaux avec python, Eyrolles, Paris.
12. Han, B., & Rideout, C. (2022). Factors Associated with University Students' Development and Success: Insights from Senior Undergraduates . *The Canadian Journal for the Scholarship of Teaching and Learning*, *13*(1). https://doi.org/10.5206/cjsotlrcacea.2022.1.10801.

13. Harrison, M., & Petrou, T. (2020). Pandas 1.x Cookbook. Pratical recipes for scientific computing, time series analysis, and exploratory data analysis using Python, Packt, Birmingham-Mumbai.
14. Hoffait, A.S., & Schyns, M. (2017). Early detection of university students with potential difficulties, Decision Support Systems, 101, 1-11. https://doi.org/10.1016/j.dss.2017.05.003.
15. Kam, H.J. (2023). Machine Learning Made Easy: A Review of Scikit-learn Package in Python Programming Language, Journal of Educational and Behavioral Statistics, 44. https://doi.org/10.3102/1076998619832248.
16. Kang, H, & Zhao, H. (2020). Description and Application Research of Multiple Regression Model Optimization Algorithm Based on Data Set Denoising, Journal of Physics: Conference Series, 1631. https://doi.org/10.1088/1742-6596/1631/1/012063.
17. Kouakou, N., & Bouah Kablan, F.B. (2015). Analyse des déterminants de la performance scolaire des établissements du secondaire public au baccalauréat, Revue Universitaire des Sciences de l'Éducation, 5.
18. Labarere, J. (2012). Corrélation et régression linéaire simple, Université Joseph Fourier de Grenoble, Paris.
19. Lardy, L., Bressoux, P., & Lima, L. (2015). Les facteurs qui influencent la réussite des étudiants dans une filière universitaire technologique : le cas de la première année d'études en DUT GEA, *L'orientation scolaire et professionnelle*. https://doi.org/10.4000/osp.4671.
20. Maulud, D.H., & Abdulazeez, A.M. (2020). A Review on Linear Regression Comprehensive in Machine Learning, Journal of Applied Science and Technology Trends, 1(2), 140-147. https://doi.org/10.38094/jastt1457.
21. Mathivet, V. (2014). L'Intelligence Artificielle pour les développeurs. Concepts et implémentations en C#, ENI, Paris.
22. Mpia, H.N., Mwendia, S.N, & Mburu, L.W. (2022). Predicting Employability of Congolese Information Technology Graduates Using Contextual Factors: Towards Sustainable Employability, Sustainability, 14(20), 13001. https://doi.org/10.3390/su142013001.
23. Naeem, M., Rizvi, S.T.H., & Coronato, A. (2020). A Gentle Introduction to Reinforcement Learning and its Application in Different Fields, *IEEE Access*, 8, 209320-209344. https://doi.org/10.1109/ACCESS.2020.3038605.
24. Natras, R., Soja, B., & Schmidt, M. (2022). Ensemble Machine Learning of Random Forest, AdaBoost and XGBoost for Vertical Total Electron Content Forecasting, *Remote Sensing*, 14(15), 3547. https://doi.org/10.3390/rs14153547.
25. Osmanbegović, E., & Suljić, M. (2012). Data Mining Approach for Predicting student performance, Economic Review: Journal of Economics and Business, X(1). https://hdl.handle.net/10419/193806.
26. Relan, K. (2019). Building REST APIs with Flask, Create Python Web Services with MySQL, New Delhi. https://doi.org/10.1007/978-1-4842-5022-8_1.
27. Romero, C., & Ventura, S. (2010). Educational Data Mininig: A reviwe of the state of the art. IEEE Trans. Syst. Man. Cybern, 40(6), 601-602. https://doi.org/10.1109/TSMCC.2010.205353.
28. Saoudi, K., & Chroqui, R. (2017). Performance de l'école : Définitions et analyse dans le cas de l'enseignement secondaire marocain, Revue Organisation et Territoires, (3). https://doi.org/10.48407/IMIST.PRSM/ot-n3.10088.
29. Tair, M.M.A., & El-Halees, A.M. (2012). Mining Educational Data to Improve Students' Performance: A Case Study, Int. J. Inf. Commun. Technol. Res., 2(2), 140–141.
30. Thiétart, R.A. (2014). Méthodes de recherche en management « Management Sup », Dunod, Paris.
31. Tuysuzoglu, G., & Birant, D. (2020). Enhanced Bagging (eBagging): A Novel Approach for Ensemble Learning",. The International Arab Journal of Information Technology, 17(4). https://doi.org/10.34028/iajit/17/4/10.