SchoolDrop'app

Application de prédiction du décrochage scolaire

Sous la direction de

Morgan COUSIN

Arona DIENE

Amor KEZIOU

Réalisé par

Jawhara CHAFI

Mathys GENET

Rémi GOMES MOREIRA

Séverine NOR

Julien PHAM

UNIVERSITÉ
DE REIMS
CHAMPAGNE-ARDENNE

Remerciements

Nous exprimons notre plus profonde gratitude envers Morgan COUSIN, Amor KEZIOU et Arona DIENE pour leurs conseils, leurs disponibilités ainsi que leurs expertises. Leurs contributions ont été des plus précieuses durant cet exercice, leur capacité à stimuler notre réflexion a enrichi de manière significative ce travail.
Merci pour leur accompagnement tout au long de ce parcours, cela a permis l'aboutissement de cette application. Nous leur sommes profondément reconnaissants pour leurs soutiens, leurs patiences et leur confiance, qui ont été des moteurs essentiels à l'aboutissement de ce projet.

Nous tenons aussi à remercier nos camarades de promotion pour leur soutien et les discussions qui nous auront grandement aidés à aiguiser notre pensée et à élargir notre vision sur certains points.

Nous voudrions, par la même occasion, exprimer notre gratitude à l'ensemble du personnel encadrant du Master, dont l'engagement et les conseils auront permis de dynamiser la réalisation de ce projet.

Table des matières

Tables des figures

Table des tableaux

Introduction

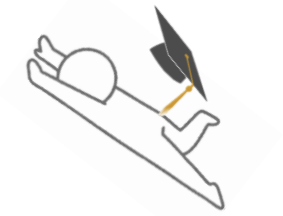
Le décrochage scolaire, défini comme l'abandon prématuré du système éducatif formel avant l'obtention d'un diplôme reconnu, constitue un enjeu majeur pour les sociétés. Ce phénomène complexe, aux causes multiples et aux conséquences variées, a des répercussions importantes sur la vie des individus concernés, mais aussi sur l'ensemble de la société.

En effet, à échelle humaine, les jeunes qui quittent prématurément l'école sont plus susceptibles d'être en situation de précarité que ce soit en termes d'insertion professionnelle, d'exclusion sociale ou encore de pauvreté. Ce phénomène n'est pas qu'un problème à l'échelle micro, à l'échelle macro, le décrochage scolaire peut entraîner des pertes économiques liées au chômage, à la diminution de la productivité lié à une masse salariale moins formé et donc à une baisse de la compétitivité.

Dans l'étude sur laquelle se base notre travail, l'abandon est défini d'un point de vue micro, nous considérons ainsi les changements de domaine d'études et d'institution comme des abandons, indépendamment du moment où ils se produisent. Cette vision a pour effet d'engendrer des taux d'abandon beaucoup plus élevés que la définition plus commune que nous avons cité précédemment.

Le décrochage scolaire est un phénomène multifactoriel qui implique de nombreux acteurs et leviers d'action. Les élèves sont les principaux concerné, les raisons du décrochage sont souvent personnelles et liées à des difficultés scolaires, des problèmes de santé, des difficultés familiales ou sociales. Ce caractère multifactoriel en fait un événement difficilement prédictible par les institutions luttant contre ce phénomène. Dû à l'implication des étudiants et aux enjeux liés à leur étude, il est parfois difficile pour les principaux concernés de savoir si oui ou non, ils sont en phase de décrochage scolaire, mais il est aussi complexe pour les enseignants d'identifier précisément les cas de décrochage dans une classe composé de multiples individus unique.

« SchoolDrop'app » se propose de répondre à cette problématique d'identification des cas de décrochage scolaire en fonction des caractéristiques personnelles des individus.



I. Quelques éléments de contexte

A. Présentation de l'application et fonctionnalités

« SchoolDrop'app » est une application visant à prédire si un étudiant est en situation de décrochage scolaire. Cet outil est basé sur un formulaire Excel, permettant de recueillir les informations de l'utilisateur. Il fonctionne grâce à un modèle de prédiction efficace, exécuté sous python et entrainé sur une base de données conséquente et fiable, boostant ainsi la crédibilité du modèle.

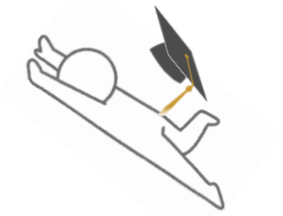
L'application commence par collecter les données de l'utilisateur via un formulaire Excel VBA. Ce formulaire comprend des questions permettant de collecter des données qui expliquerait le décrochage scolaire et alimenterait donc la prédiction. Les informations saisies dans le formulaire sont contrôlées de façon à s'assurer que les informations soient correctement renseignées, réduisant ainsi les erreurs d'entrée et garantissant une collecte de données de qualité.

Une fois le formulaire compléter, il est possible de lancer la prédiction, une infrastructure de scripts Python va ainsi se mettre en marche. Le modèle statistique utilisé est entraîné sur des données représentatives dans le but d'évaluer la probabilité de décrochage scolaire en fonction des variables explicatives fournies par l'utilisateur.

Le retour y est instantané, a la fin des calculs prédictifs, les résultats sont restitués directement à l'utilisateur via l'interface Excel, affichant le risque de décrochage et la crédibilité de la prédiction.

« SchoolDrop'App » vise ainsi à fournir un retour visuel simple indiquant la situation prédite de l'étudiant avec un indicateur de fiabilité de la prédiction, permettant à l'utilisateur de comprendre facilement le niveau de risque associé.

Enfin, les résultats peuvent être enregistrés ou exportés pour une analyse ultérieure.



B. Identification du besoin client

Le projet SchoolDrop'App répond au besoin croissant des établissements éducatifs de prévenir le décrochage scolaire, un problème qui a des impacts significatifs sur les parcours académiques et professionnels des étudiants. De nombreux enseignants, conseillers pédagogiques et responsables éducatifs rencontrent des difficultés pour identifier de manière précoce les étudiants à risque de décrochage, faute d'outils d'analyse adaptés et prédictifs. Ces professionnels ont exprimé le besoin d'une solution capable de fournir une évaluation rapide, objective et personnalisée du risque de décrochage, afin d'optimiser les interventions et de renforcer le soutien aux étudiants en difficulté.

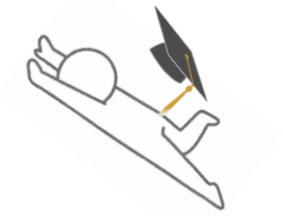
« SchoolDrop'App » a été conçue pour répondre spécifiquement à ces attentes. Cette application permet de collecter des données pertinentes via un formulaire interactif, puis de les analyser grâce à des modèles d'apprentissage automatique, offrant ainsi une prédiction fiable et exploitable du risque de décrochage. L'objectif principal est de fournir un outil accessible et intuitif qui aide les éducateurs à mieux comprendre les situations individuelles des étudiants et à cibler leurs efforts de manière proactive.

Ce besoin est également renforcé par les politiques éducatives actuelles qui valorisent l'inclusion et la réussite pour tous les étudiants. « SchoolDrop'App » vise à fournir une aide décisionnelle aux établissements pour répondre aux exigences d'accompagnement renforcé des élèves, tout en contribuant à réduire le taux de décrochage scolaire et à favoriser un environnement éducatif plus inclusif.

C. Cadre applicatif

L'application SchoolDrop'App est destinée à identifier les risques de décrochage scolaire parmi les étudiants, afin de permettre des interventions préventives et ciblées. Elle est particulièrement adaptée aux enseignants, conseillers pédagogiques, et responsables éducatifs exerçant dans des établissements scolaires ou universitaires. Ces professionnels peuvent l'utiliser à des moments clés de l'année académique, par exemple lorsqu'un élève présente des signes de désengagement.

Conçue pour être intuitive et accessible, l'application s'inscrit dans une démarche de soutien éducatif inclusif, alignée sur les politiques éducatives européennes visant à réduire les taux de décrochage. « SchoolDrop'App » est proposée gratuitement, ce qui garantit son accessibilité à une large audience et renforce son impact sociétal.



L'application fonctionne cependant uniquement sur ordinateur, via une interface Excel et un réseau de scripts Python. Elle peut être utilisée à la fois en ligne et hors ligne. Cependant, l'accès aux mises à jour nécessitent une connexion Internet. L'application est optimisée pour les systèmes d'exploitation majoritaires et ne nécessite pas de compétences avancées en informatique de la part de ses utilisateurs.

« SchoolDrop'App » a pour ambition une portée Européenne, basé sur des données issus d'étudiant portugais, nous assumons ici que les étudiants des pays de l'Union Européenne (UE) sont similaires entre eux dans leur cursus et structure scolaire ou universitaire. Pour étayer cette lourde affirmation, nous nous appuyons sur plusieurs organisme tel que l'Espace européen de l'enseignement supérieur (EEES) qui vise à harmoniser les systèmes d'enseignement en Europe, ce qui peut conduire à des similitudes dans les parcours, les méthodes d'enseignement et les objectifs éducatifs des étudiants européens.

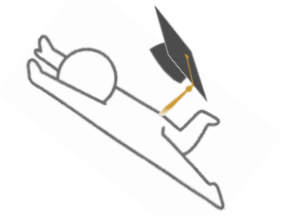
De plus, les politiques éducatives communes, comme le programme Erasmus+ mis en place par l'UE renforce ces similitudes dans les comportements des étudiants à travers les différents pays européens. Par exemple, les étudiants portugais, comme les étudiants d'autres pays, bénéficient des mêmes opportunités de mobilité et des mêmes structures d'aide financière, ce qui pourrait influencer leurs décisions et expériences de manière comparable à celles d'autres étudiants européens.

D. L'utilisateur : établissement d'un persona

L'utilisateur principal de l'application est un enseignant ou une enseignante attentif et engagé, exerçant dans un établissement scolaire en Europe, au sein d'un environnement éducatif marqué par la diversité des parcours d'apprentissage et des contextes socio-économiques. Ce persona est un professionnel passionné, dédié à la réussite de ses élèves, mais confronté à la complexité sociologique que représente le décrochage scolaire.

Cet enseignant ou cette enseignante ne possède potentiellement pas d'expertise spécifique en analyse de données ou en outils technologiques avancés. Ce professionnel se sent parfois démuni face aux indicateurs subtils ou complexes de décrochage scolaire. Malgré ces limites, ce persona fait preuve de curiosité et de motivation pour adopter des solutions pratiques permettant de renforcer son action pédagogique.

Le principal objectif de ce persona est de vérifier ses intuitions concernant les risques de décrochage chez certains élèves de sa classe. Il ou elle souhaite bénéficier d'un outil simple et



accessible pour recueillir des informations, les analyser et obtenir des résultats compréhensibles qui facilitent la prise de décision. En fonction des résultats fournis par l'application, l'enseignant peut mettre en place des actions ciblées, comme proposer un accompagnement individualisé, alerter les familles ou solliciter les services spécialisés au sein de l'établissement.

Ce persona a besoin d'un outil qui soit :

- Intuitif : L'application doit proposer une interface simple, claire et adaptée à un utilisateur non-expert en technologie ou en analyse de données.
- Fiable et précis : Les résultats doivent être pertinents, basés sur des données fiables et des modèles éprouvés.
- Respectueux de la confidentialité : Étant donné que l'utilisateur emploie des informations sensibles sur les élèves, l'application doit garantir la sécurité et la confidentialité des données.

Dans son quotidien, ce persona utilise l'application pour remplir un formulaire spécifique pour chaque élève. Après avoir soumis ces informations, il ou elle obtient un rapport synthétique, sans information superflu, indiquant le niveau de risque pour chaque élève et le taux de crédibilité de ce risque. En fonction des résultats, l'utilisateur peut planifier des actions adaptées, telles que :

- Organiser un entretien avec l'élève concerné.
- Proposer des activités de soutien éducatif.
- Informer les proches pour engager une collaboration.

II. Présentation des données

A. Compréhension des objectifs et de la base de données

L'analyse a pour objectif de profiler les individus ainsi que de fournir des statistiques descriptives globales. La base de données provient de différentes sources de données disjointes de l'Institut Polytechnique de Portalegre, notamment par :

- Système de gestion académique (AMS) de l'institution ;
- Système de soutien à l'activité d'enseignement de l'institution (PAE) ;
- Direction générale de l'enseignement supérieur (DGES) : données annuelles sur les admissions via le Concours national d'accès à l'enseignement supérieur (CNAES) ;
- Base de données Portugal Contemporain (PORDATA) : données macro-économiques.

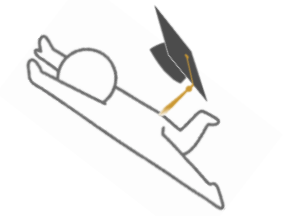
Elle a été utilisée dans le cadre d'un outil d'analyse de l'apprentissage de l'Institut Polytechnique de Portalegre, pour prédire le risque de décrochage et d'échec des étudiants et ainsi, leur apporter une aide. Cette base de données est destinée à des chercheurs souhaitant réaliser une étude sur les performances académiques des étudiants (décrochage ou réussite), mais aussi pour former à l'apprentissage automatique.

Chaque ligne de la base de données correspond à un étudiant. L'ensemble de nos données comprend les étudiants inscrits entre 2008-2009 et 2018-2019, soit un total de 4 424 étudiants.

La constitution de la base de données a été faite en quatre étapes.

- Étape 1

Les premières données collectées sont celles disponibles à l'inscription, comme le cours auquel l'étudiant s'est inscrit au Concours National d'Accès à l'Enseignement Supérieur (CNAES). Chaque année, une base de données sous Microsoft Access est créée après les résultats du concours. Un programme VBA collecte les données nécessaires et les exporte sous un fichier CSV.



❖ Étape 2

Les dossiers des étudiants (dans un fichier CSV) par l'AMS sont préparés afin d'être traités. Ce fichier contient 13 992 lignes et 398 colonnes. Cette étape est le nettoyage de cette base de données en supprimant les lignes et colonnes dupliquées ou non pertinentes pour l'étude. À l'issue de cette étape, on retrouve toutes les données relatives aux données démographiques et socio-économiques.

❖ Étape 3

Le prochain fichier CSV à être traité contient les données sur les évaluations des étudiants. Chaque ligne d'étudiant obtenu par l'étape précédente, se voit attribué toutes ses informations sur ses données académiques à la fin du 1er et 2nd semestre.

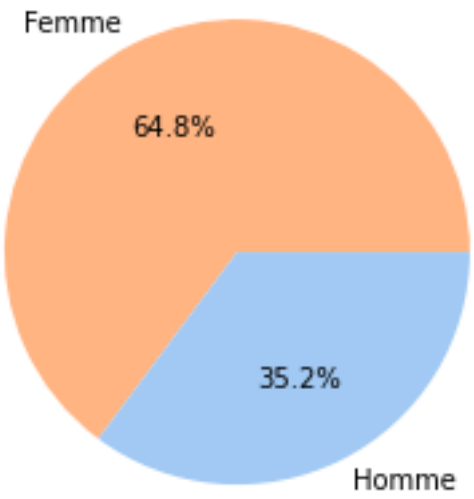
❖ Étape 4

La dernière étape consiste à assembler toutes les bases de données des étapes précédentes en une seule, auquel est ajouté les données macro-économiques. Enfin, un dernier nettoyage a été effectué pour les anomalies, les valeurs aberrantes ou manquantes.

B. Portrait des étudiants

L'analyse débute par une observation générale des effectifs étudiants. L'échantillon étudié comprend un total de 4 424 étudiants, répartis entre 2 868 femmes et 1 556 hommes. Cette répartition met en évidence une forte prédominance de femmes dans la population étudiante. En effet, ces dernières représentent une part de 64,8% contre 35,2% d'hommes (cf. Figure 1 : Représentation des effectifs par genreFigure 1). Notre étude révèle une forte majorité de célibataires, représentant 88,6 % des étudiants, ce qui peut être expliqué par le profil jeune des inscrits, avec une moyenne d'âge de 23 ans. 75 % des étudiants ont moins de 25 ans. À l'inverse, seulement 8,6 % des étudiants (soit 379 individus) sont mariés.

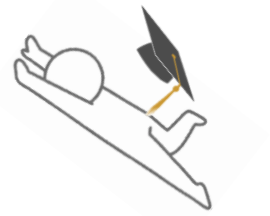
Figure 1 : Représentation des effectifs par genre



Source : Réalisée par les auteurs

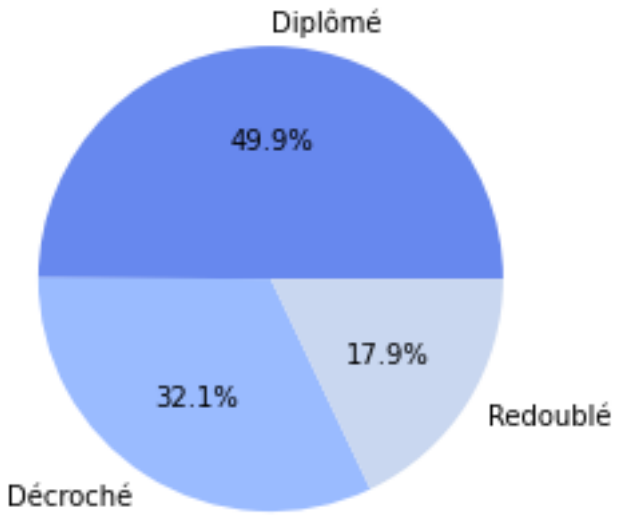
Les célibataires sont majoritairement représentés par des étudiants âgés de 21 ans en moyenne, tandis que les étudiants séparés ou divorcés sont souvent beaucoup plus âgés. Cette observation reflète les différentes étapes de vie et les raisons variées qui poussent les individus à entreprendre des études.

L'analyse de la situation géographique des étudiants permet d'observer que la moitié d'entre eux vivent loin de leur domicile. En effet, 54,8 % des étudiants, soit un total de 2 426 étudiants, ont quitté leur domicile pour poursuivre leurs études, mettant ainsi en évidence la mobilité géographique des étudiants.



Près de la moitié d'entre eux (49,9%) ont obtenu leur diplôme et un tiers des effectifs ont décroché (32,1%) (cf. Figure 2). Pour rappel, le taux de décrochage scolaire est particulièrement élevé dû au fait qu'il est mesuré au niveau micro-économique. En outre, 17,9 % des étudiants sont encore inscrits, mais n'ont pas encore obtenu leur diplôme, ce qui peut indiquer des parcours prolongés ou qui n'ont pas encore réussi à valider tous leurs crédits, sans pour autant avoir décroché.

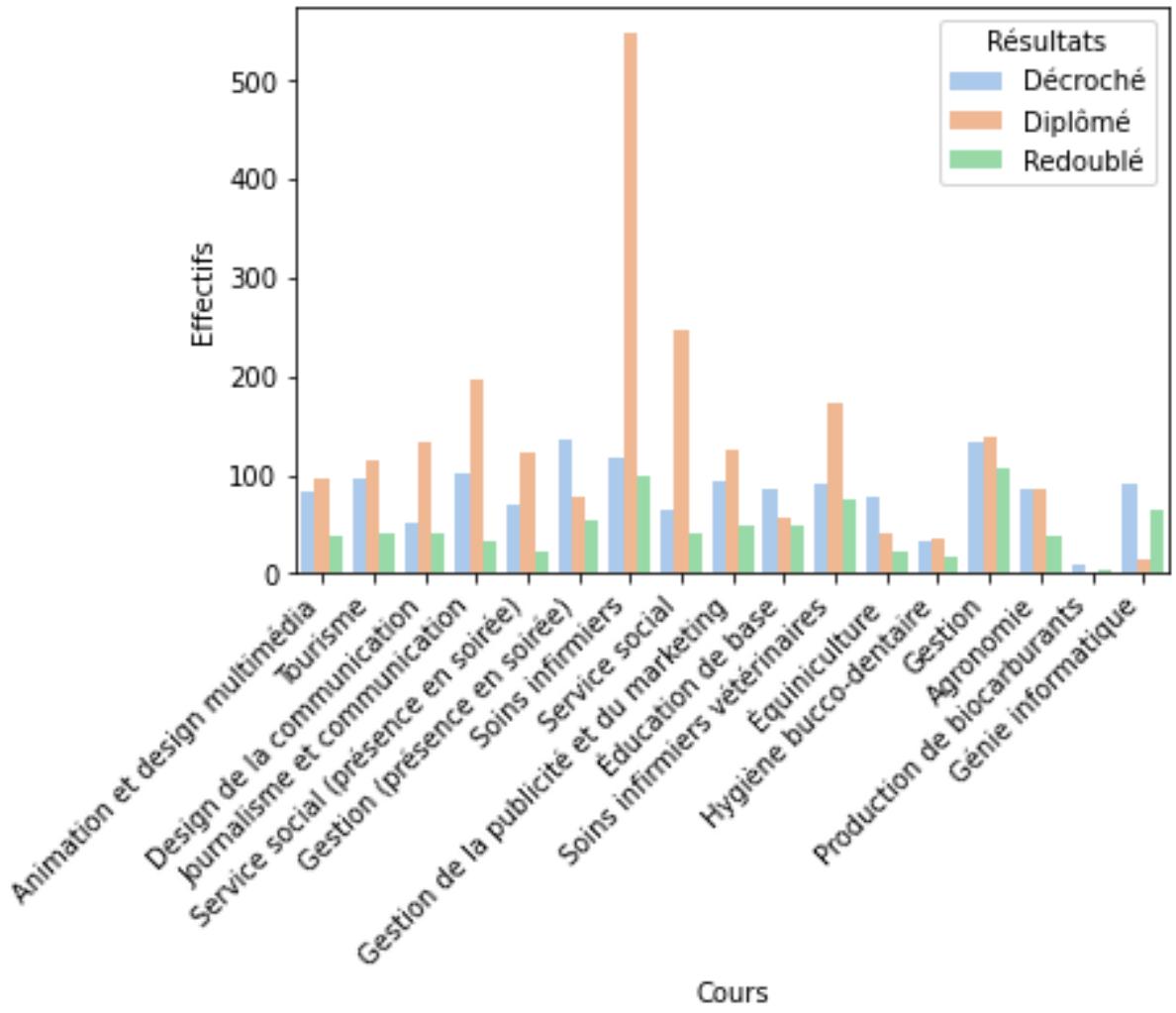
Figure 2 : Répartition des résultats à la fin du programme



Source : Réalisée par les auteurs

On observe un taux de réussite nettement supérieur dans deux enseignements : soins infirmiers et service social (cf. Figure 3) contre un fort taux de décrochage scolaire en génie informatique. On constate une casi-égalité dans le cours de gestion et agronomie entre les étudiants diplômés et ceux qui ont décrochés.

Figure 3 : Répartition des cours selon les résultats



Source : Réalisée par les auteurs

Il est également important d'observer qu'une infime partie des étudiants ont des besoins éducatifs spéciaux (1,2%), cela peut concerner l'accompagnement des étudiants en situation de handicap ou d'étudiants présentant des difficultés d'apprentissage. Bien qu'ils soient très peu représentés, il ne faut pas négliger un accompagnement spécialisé pour ces étudiants afin d'éviter le risque de décrochage scolaire.

C. Parcours académiques

Les hommes s'inscrivent généralement plus tard que les femmes, avec une moyenne d'âge de 24 ans contre 22 ans pour les femmes. Cette différence pourrait être liée à des parcours éducatifs ou professionnels plus diversifiés, donnant aux hommes un éventail d'âges plus large au moment de l'inscription. En outre, les données montrent que les hommes ont tendance à choisir des formations dans des disciplines variées, allant des domaines spécifiques comme le marketing, l'informatique et la gestion, à des cursus plus généralistes ou spécialisés.

Nous pouvons observer que l'enseignement de soins infirmiers est le plus prisé (cf. Figure 4), attirant ainsi 17,3 % d'étudiants. D'autres formations comme la gestion (8,6%) et les services sociaux (8%) suscitent un intérêt marqué parmi ces étudiants. À l'opposé, certaines formations sont nettement moins attractives, notamment le cours de technologie de production de biocarburants (0,3 %, soit 12 étudiants) et le cours d'hygiène bucco-dentaire (1,9 %, soit 86 étudiants).

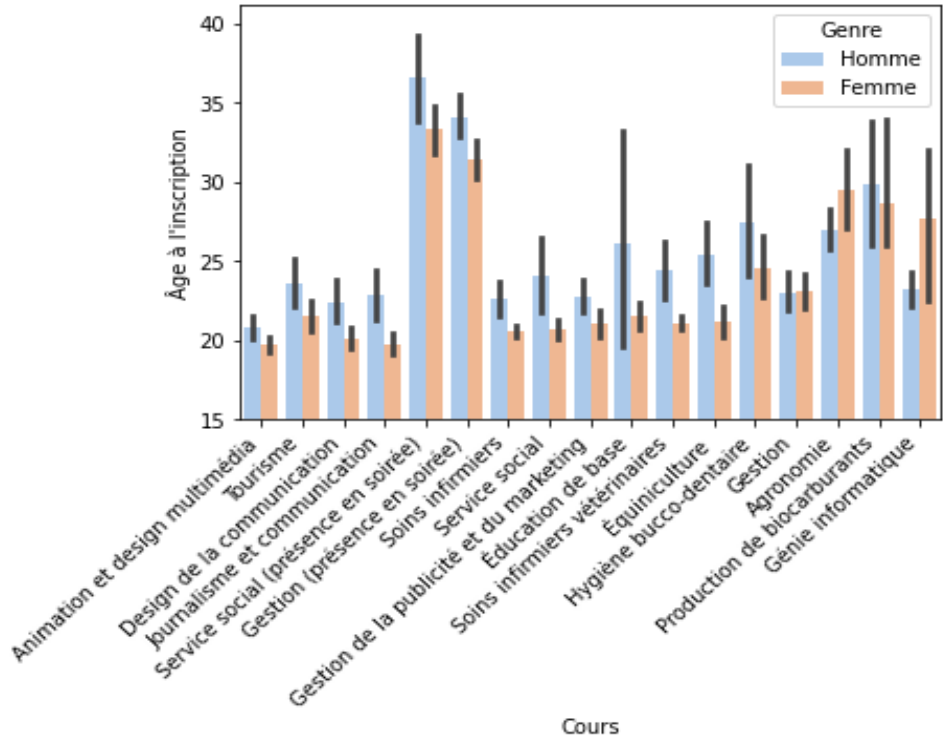
Figure 4 : Nuage de mots des cours choisis par les étudiants

Source : Réalisée par les auteurs

Certains cursus attirent des étudiants plus âgés (cf. Figure 5), aux alentours de 32 ans en moyenne, comme les programmes en gestion et service social, particulièrement ceux en horaires du soir. Ces formations pourraient convenir à des personnes en reconversion professionnelle ou ayant d'autres responsabilités. D'autres cursus, comme ceux en animation ou journalisme,

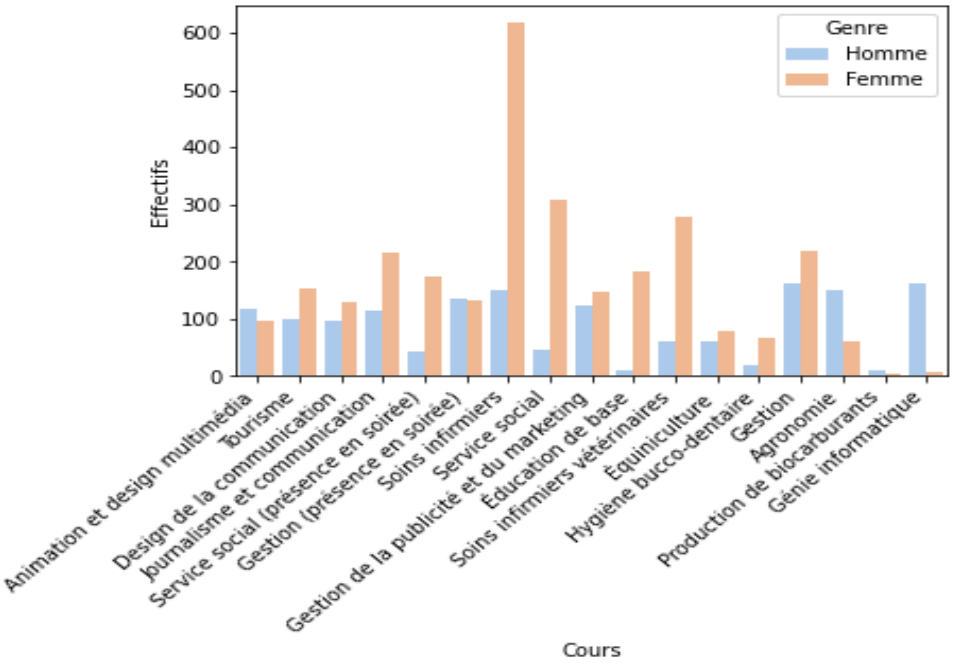
regroupent des étudiants plus jeunes, 20 ans en moyenne, suggérant qu'ils sont souvent choisis dès la sortie du lycée.

Figure 5 : Répartition du genre selon l'enseignement et l'âge



Source : Réalisée par les auteurs

Figure 6 : Répartition du genre selon l'enseignement inscrit

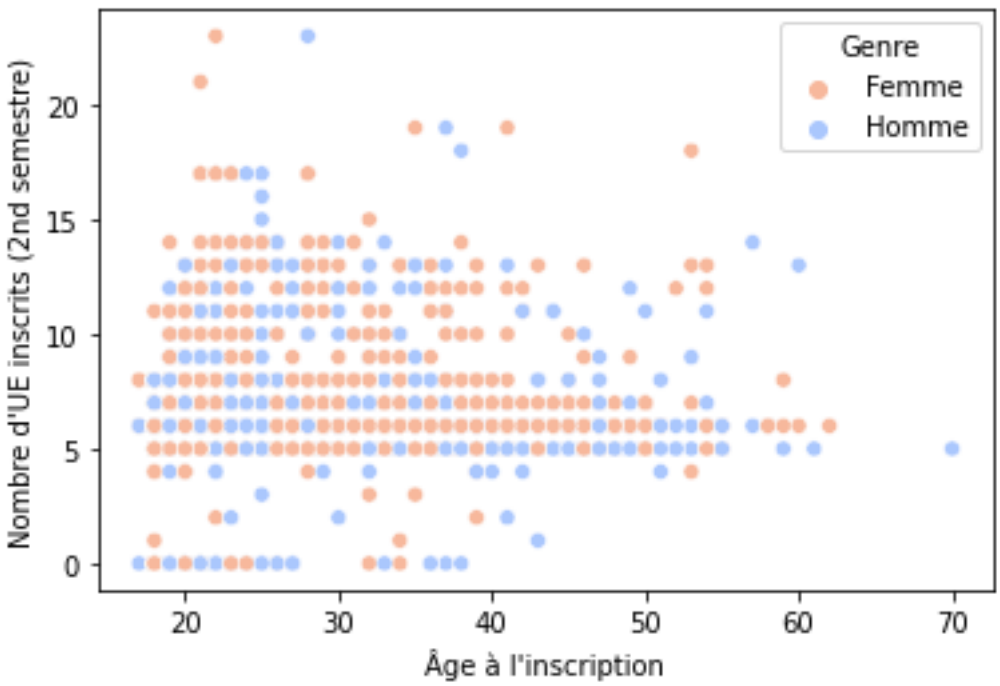


Source : Réalisés par les auteurs

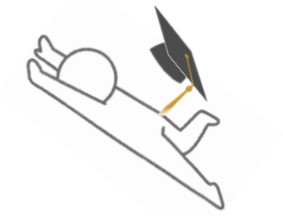
La répartition des étudiants dans les différents programmes de cours révèle des tendances en fonction du genre. Tout d'abord, nous pouvons constater que certains programmes attirent majoritairement un genre. Par exemple, dans des filières comme Soins infirmiers, Service social et Éducation, les femmes représentent une proportion beaucoup plus élevée que les hommes (cf. Figure 6). Cela pourrait refléter des tendances sociétales ou historiques, où des métiers comme les soins infirmiers ou l'enseignement sont traditionnellement perçus comme des professions féminines. En revanche, d'autres programmes, comme Informatique, Technologies de la production de biocarburants et Agronomie attirent une proportion plus élevée d'hommes. Cela peut être lié à des facteurs tels que la perception des métiers techniques et scientifiques comme étant davantage associés aux hommes, ou encore des barrières culturelles et sociales qui influencent les choix de parcours.

De plus, des programmes comme Gestion ou Communication et design présentent une distribution plus équilibrée entre hommes et femmes. Cela pourrait indiquer une tendance croissante vers la diversification des genres dans ces domaines plus neutres, à la fois sur le plan académique et professionnel. Enfin, nous pouvons remarquer un grand écart dans le programme Social Service. La majorité des étudiants dans ce domaine sont des femmes, ce qui pourrait être un indicateur des spécificités sociales et culturelles qui influencent les choix d'orientation professionnelle dans des secteurs liés à l'aide sociale et aux professions de soin.

Figure 7 : Répartition du genre selon l'âge et le nombre d'unités d'enseignements inscrit



Source : Réalisée par les auteurs

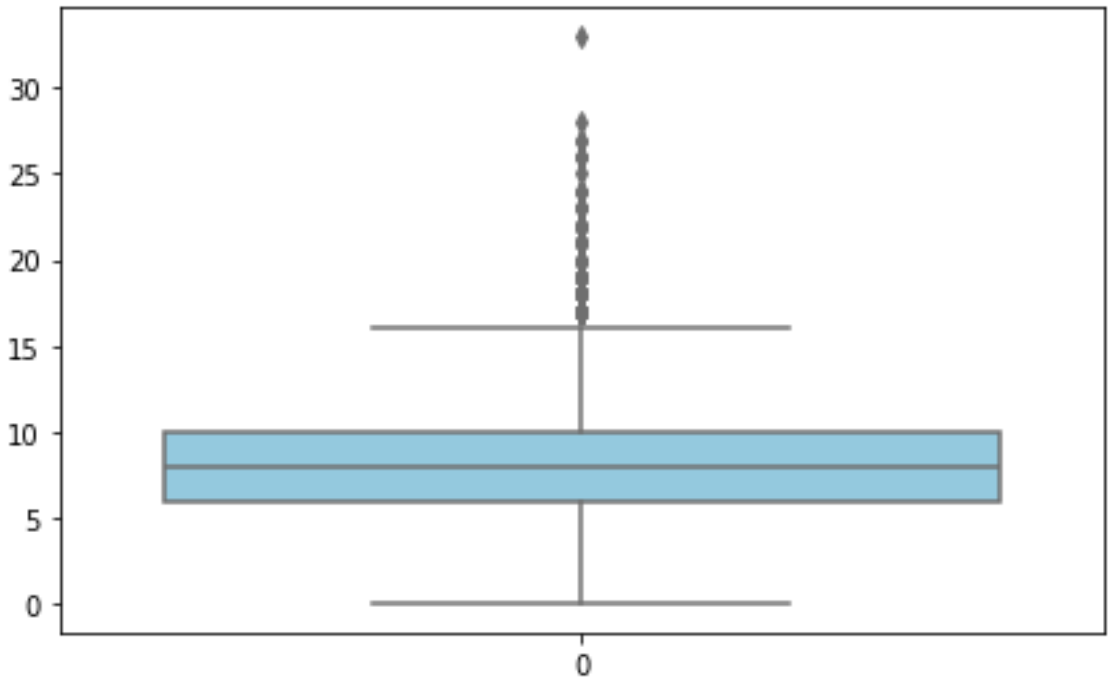


Nous constatons une concentration importante entre 5 et 15 unités d'enseignement inscrits pour la plupart des âges (cf. Figure 7). Peu d'étudiants s'inscrivent à plus de 20 enseignements du fait que cela représente une charge de travail élevée. La distribution est équilibrée entre les genres.

Nos analyses montrent que les femmes s'inscrivent à un nombre légèrement plus élevé d'unités d'enseignement au second semestre par rapport aux hommes. En effet, en moyenne, elles s'inscrivent à 6 enseignements contre 5 pour les hommes. Cependant, les deux groupes suivent des tendances globalement similaires, bien que les hommes présentent une plus grande diversité dans leur nombre d'inscriptions.

Quant au premier semestre, les étudiants s'inscrivent en moyenne à 6 unités d'enseignement avec 75% des étudiants qui prennent maximum 7 unités d'enseignement. Quelques profils atypiques sont observés avec un nombre de 26 unités d'enseignement inscrites (cf. Figure 7). Les étudiants passent en moyenne 8 évaluations au cours du semestre. 75% de ces étudiants passent jusqu'à 10 évaluations. Ces résultats montrent une concentration autour de 8 à 10 évaluations, bien que certains cas extrêmes, avec un nombre d'évaluations très élevé, soient également présents (cf. Figure 8).

Figure 8 : Représentation du nombre d'évaluations passées par les étudiants



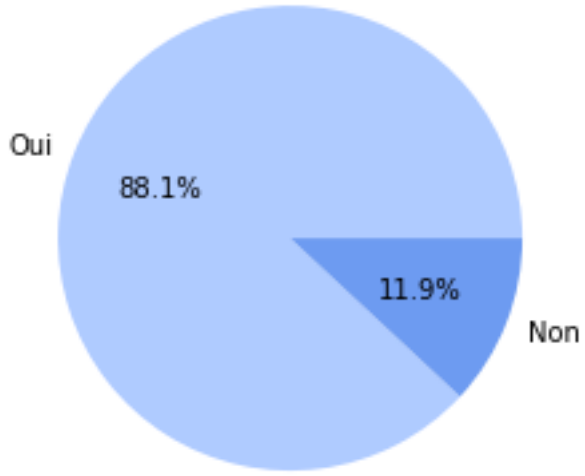
Source : Réalisée par les auteurs

D. Facteurs sociaux et financiers

En ce qui concerne la situation socio-économique des étudiants, nous observons qu'une proportion majoritaire d'entre eux (75,2%) ne bénéficie pas de bourse, tandis que seulement 24,8 % reçoivent un soutien financier. Cette répartition indique que la majorité des étudiants doivent financer leurs études par d'autres moyens, ce qui peut influencer leur engagement et leur réussite académique. Elles sont plus fréquentes chez les étudiants plus jeunes (20 ans en moyenne), ce qui peut s'expliquer par des politiques de financement qui favorisent ceux qui débutent leur parcours académique.

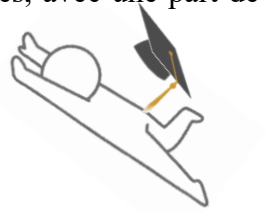
La question des frais de scolarité est également à prendre en compte, puisque 11,9 % des étudiants ne sont pas à jour dans leur règlement (cf. Figure 9). Ces étudiants tendent à être plus âgés de 26 ans en moyenne. Bien que la majorité des étudiants de 20 ans en moyenne (88,1%) soit à jour dans le paiement des frais de scolarité, la part d'étudiants n'étant pas à jour peut être vue comme un signe de fragilité économique.

Figure 9 : Représentation des effectifs ayant réglé ou non leur frais de scolarité



Source : Réalisée par les auteurs

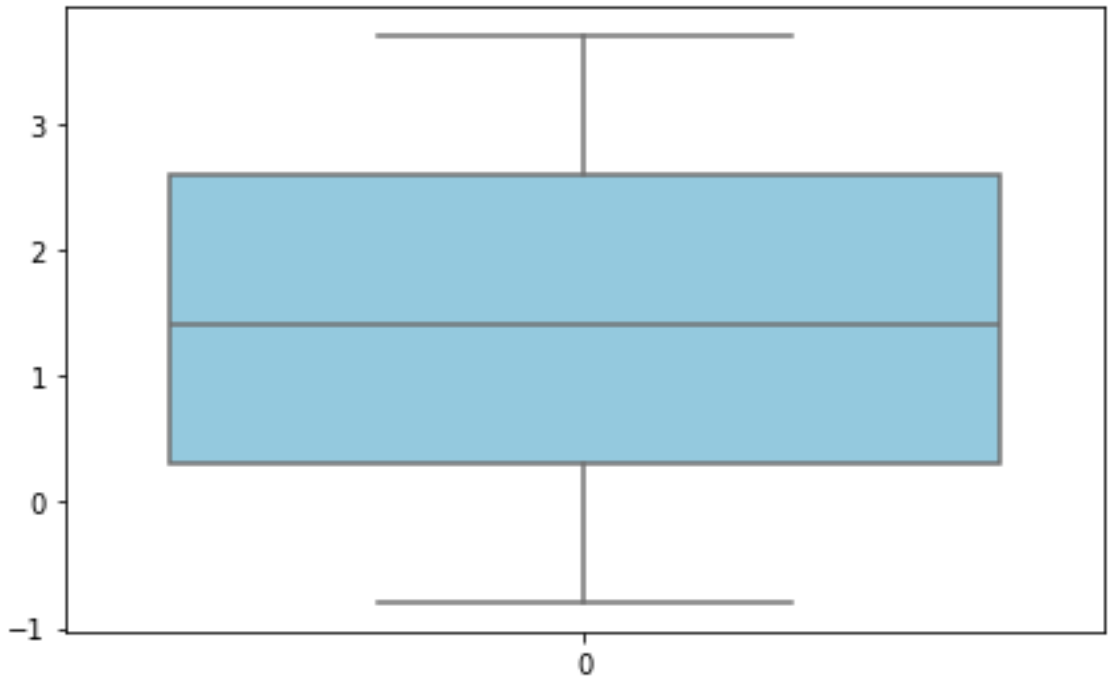
En termes de professions exercées par les parents des étudiants, les trois professions les plus courantes parmi les mères sont les métiers non-qualifiées (35,65%), le personnel administratif (18,47%) et le personnel des services professionnels de la sécurité, de la sûreté et les vendeuses (11,98%). En revanche, les métiers les moins présents chez les mères sont les techniciennes en technologies de l'information et de la communication (2%), les directrices des services administratifs et commerciaux, et autres personnels des forces armées, avec une part de 2%.



Pour les pères, les professions les plus fréquentes sont les travailleurs non-qualifiées également avec 22,83%. En seconde position se trouvent les métiers qualifiés dans l'industrie, la construction et les artisans (15,25%). Enfin, avec une part de 11,66%, nous avons le personnel des services professionnels de la sécurité, de la sûreté et les vendeuses. Concernant les professions les moins fréquentes, nous retrouvons le personnel des services de protection et de sécurité (2%), des spécialistes en finance, comptabilité, organisation administrative et relations publiques (2%) et enfin des spécialistes en sciences physiques, mathématiques et ingénierie (2%). Nous observons une concentration significative dans certaines catégories professionnelles.

Le taux d'inflation présente une moyenne de 1,23% (cf. Figure 10), ce qui signifie qu'en moyenne, l'inflation dans les régions ou pays d'origine des étudiants est modérée. Ce taux varie entre -0,8% et 3,7%. Le taux négatif peut refléter une situation particulière comme une déflation. 75% des étudiants viennent de zones où l'inflation ne dépasse pas 2,6%.

Figure 10 : Représentation du taux d'inflation en fonction du lieu de domicile (régions ou pays) de l'étudiant



Source : Réalisée par les auteurs

III. Frontend : interface Excel VBA

Cette partie se consacre à l'explication du travail effectué sous Excel et VBA, elle présente en détail le fonctionnement de l'application ainsi que les techniques utilisées.

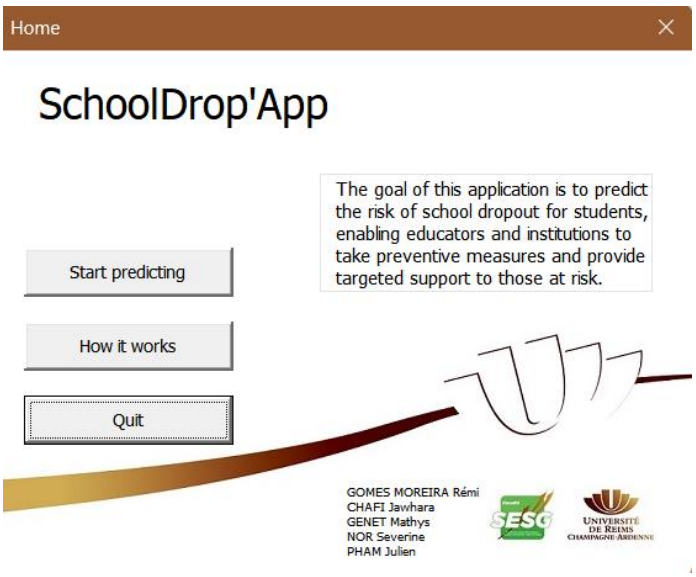
A. L'interface utilisateur

La majeure contrainte dans le développement de l'interface était de garantir une expérience utilisateur simple, intuitive et accessible, même pour des personnes sans compétences techniques. De plus, il fallait rendre l'application fonctionnelle tout en reflétant le thème scolaire grâce à des choix visuels adaptés. Ainsi, les éléments de navigation ont été volontairement réduits à l'essentiel afin d'éviter toute confusion, et chaque composant, qu'il s'agisse des boutons ou des champs à remplir, a été étiqueté pour indiquer sa fonction.

1. L'interface d'accueil et page de renseignement

Lors du lancement de l'application, l'utilisateur arrive sur le menu principal (cf. Figure 11), qui arbore le logo et les couleurs de l'université. Trois boutons sont proposés : *Start Predicting*, *How it works*, et *Quit*. Le nom de l'application et son objectif y sont également affichés.

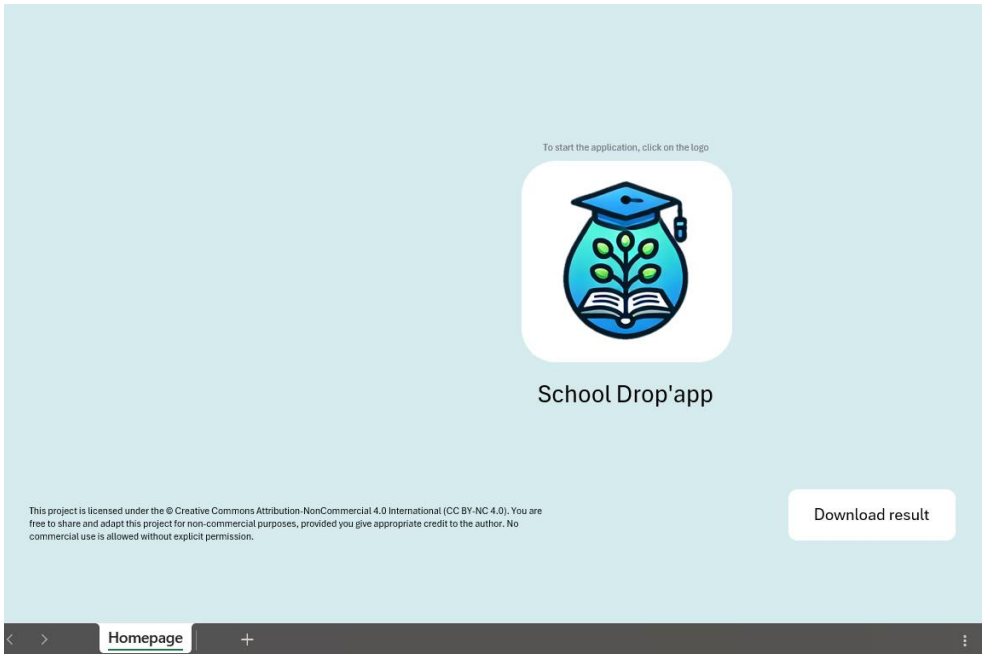
Figure 11 : Page d'accueil (Home) de l'application



Source : Réalisée par les auteurs

Si l'utilisateur sélectionne le bouton *Quit*, l'application se ferme et l'utilisateur est redirigé vers la feuille Excel intitulée *Homepage* (cf. Figure 12), où figurent la licence du projet (Creative Commons Non Commercial 4.0) ainsi que le logo de l'application. En cliquant sur ce logo, l'utilisateur peut rouvrir l'application[1]. Il y figure aussi un bouton qui permet à l'utilisateur de télécharger les résultats des prédictions de la session actuelle, plus de détails dans la partie concernant l'interface de sortie.

Figure 12 : Feuille Excel (Homepage) de l'application

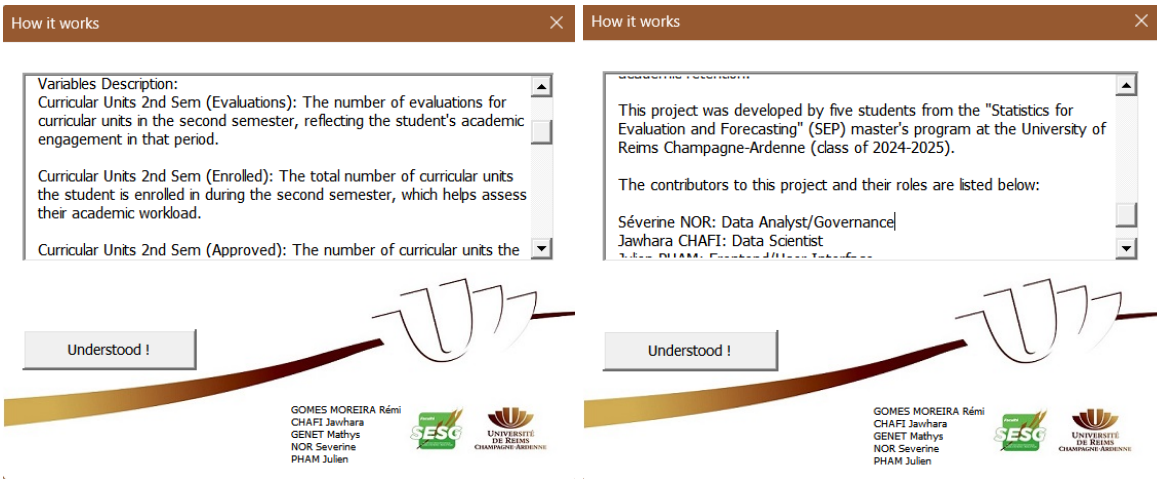


Source : Réalisée par les auteurs

En cliquant sur *How it works*, l'utilisateur accède à une section d'information (cf. Figure 13). Cette dernière comprend une description concise de l'objectif de l'application, une explication des variables utilisées dans le questionnaire de prédiction, et une présentation des rôles des auteurs.

[1] Un petit texte en gris est là pour indiquer que le logo est interactif.

Figure 13 : Section d'information de l'application



Source : Réalisée par les auteurs

L'utilisateur peut naviguer librement à l'aide de la barre de défilement. Des ajustements ont été effectués pour empêcher toute saisie dans les zones de texte (TextBox), tout en permettant leur sélection pour consultation.

2. Le questionnaire

En cliquant sur le bouton « *Start Predicting* », l'utilisateur accède au questionnaire de prédiction (cf. Figure 14). Cette section propose trois boutons : *Back to Homepage*, *Reset Choices*, et *Predict*.

Chaque champ de saisie est accompagné d'un *placeholder*[2], développé manuellement dans un module VBA. Ces *placeholder* fournissent des exemples de valeurs attendues, facilitant ainsi la compréhension des informations à renseigner.

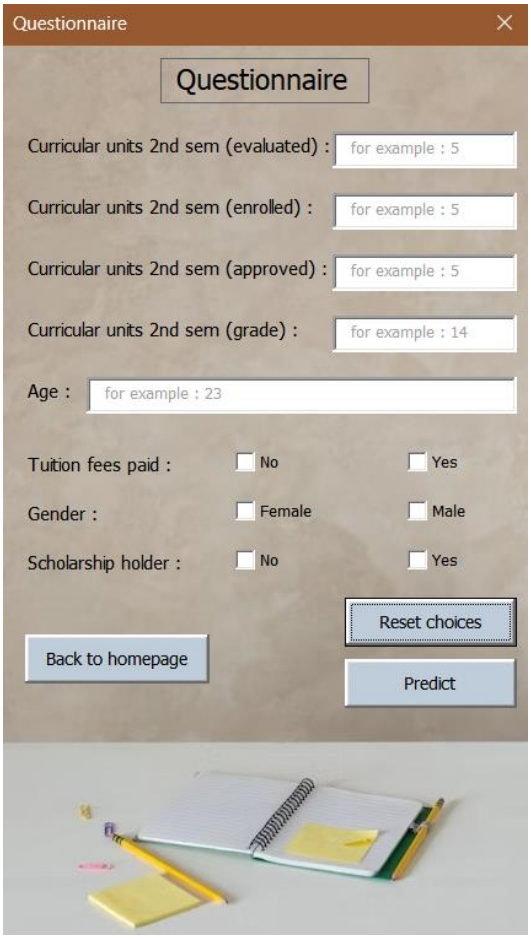
Les cases à cocher, comme celles destinées aux champs *Tuition fees paid*, *Gender*, et *Scholarship holder*, ont été sécurisées grâce à des scripts VBA. Cette configuration garantit qu'une seule case peut être cochée par groupe, empêchant des sélections conflictuelles ou incohérentes.

[2] *placeholder* : texte indicatif octroyant une aide à la saisie

L'agencement des boutons au bas du formulaire offre une navigation fluide :

- En sélectionnant *Back to Homepage*, l'utilisateur est redirigé vers la page d'accueil. Les réponses déjà saisies dans le questionnaire sont sauvegardées en mémoire, permettant de les retrouver intactes lors d'un retour au formulaire. Cela est utile si l'utilisateur veut retourner voir les informations concernant les variables dans la section *How it works*.
- En cliquant sur le bouton *Reset Choices*, une boîte de dialogue s'affiche pour confirmer que les réponses ont bien été réinitialisées.

Figure 14 : Questionnaire de la prédiction



Source : Réalisée par les auteurs

Pour les cinq premières questions l'utilisateur bénéficie d'un retour visuel immédiat grâce à un mécanisme de validation intégré :

Si une réponse est incorrecte, par exemple un nombre négatif ou une valeur non numérique, le fond du champ concerné devient rouge, signalant une erreur à corriger.

Une réponse valide (un nombre positif ou conforme aux contraintes définies) rétablit le fond blanc, indiquant que la saisie est correcte.

De plus, pour la question *Curricular units 2nd sem (grade)*, seules les valeurs numériques comprises entre 0 et 20 sont acceptées, conformément au système de notation. Toute valeur hors de cet intervalle déclenche un retour visuel rouge.

Pour les questions comportant des cases à cocher (*Tuition fees paid*, *Gender*, et *Scholarship holder*), des mécanismes de sécurisation ont été mis en place afin de garantir qu'une seule case peut être cochée à la fois par catégorie.

Enfin, lorsque l'utilisateur appuie sur le bouton *Predict*, des contrôles supplémentaires sont effectués pour vérifier les éventuelles erreurs (cf. Figure 15) :

Champs non valides (en rouge) :

Si l'un des champs contient une erreur (fond rouge), un message d'erreur s'affiche dans une boîte de dialogue (MsgBox) :

"Please correct the highlighted fields before proceeding."

L'utilisateur est alors invité à corriger les champs concernés avant de soumettre le formulaire.

Champs non renseignés :

Si un champ est laissé vide, une boîte de dialogue informe l'utilisateur du champ manquant. Par exemple si *Curricular Units 2nd sem (evaluated*) n'est pas complété, le message suivant s'affichera :

"Please enter the number of Curricular Units 2nd sem (evaluated)."

Cases à cocher non sélectionnées :

Si aucune case n'est cochée pour une question, la même chose s'applique par exemple avec la variable *Gender* :

"Please select a gender (Male or Female)."

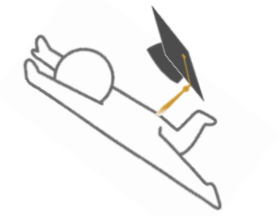
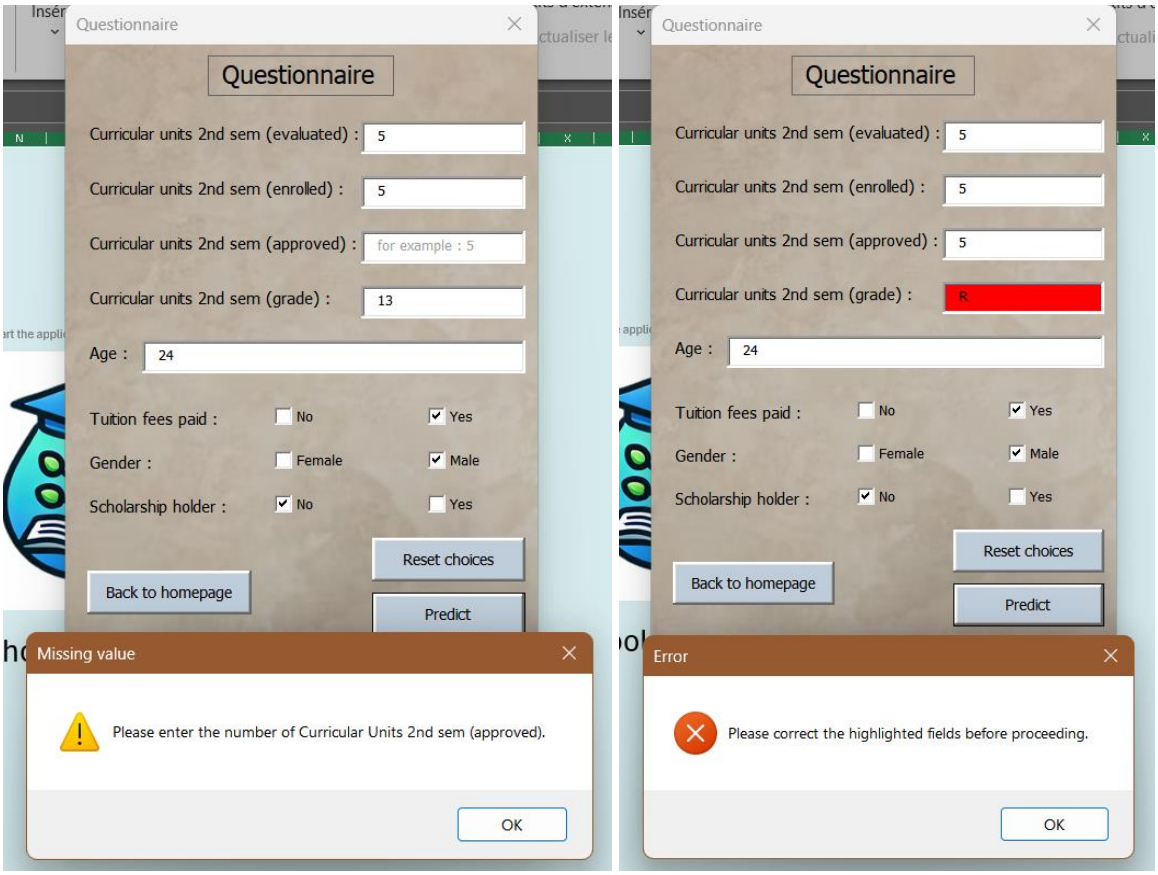


Figure 15 : Boites de dialogues d'avertissements



Source : Réalisée par les auteurs

Si aucune erreur n'est détectée, la connexion avec Python est établie, et l'utilisateur est redirigé vers l'interface de sortie.

3. L'interface de sortie

Sur l'interface de sortie (cf. Figure 16), le résultat de la prédiction est affiché, avec une couleur indiquant le statut :

- **Vert** pour "Graduate" (réussite),
- **Rouge** pour "Dropout" (décrochage).

Le score de la prédiction est également affiché, et sa couleur varie selon sa valeur :

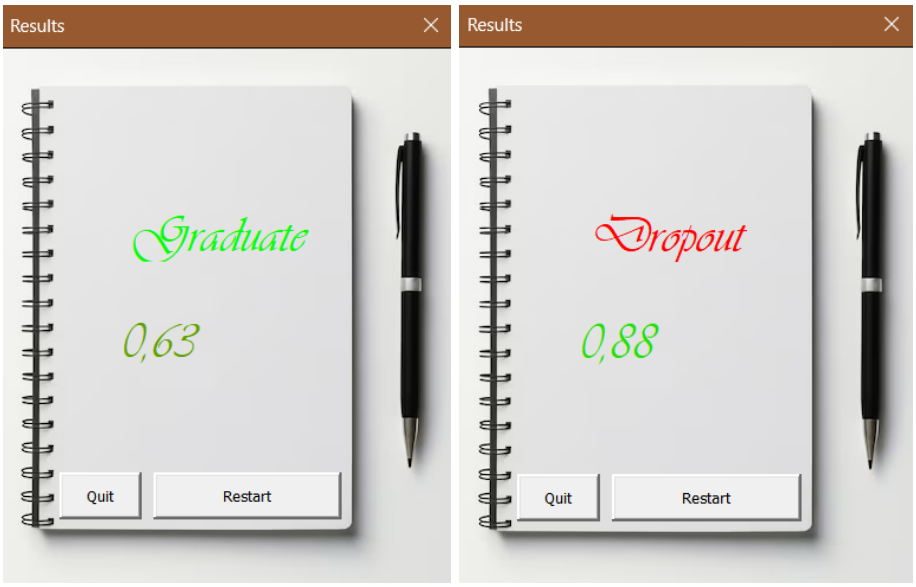
- **Rouge** si le score est proche de 0,
- **Vert** si le score est proche de 1.

Ce dégradé de couleur est calculé en fonction du score, où la composante rouge diminue à mesure que le score augmente, et la composante verte augmente en conséquence.

Deux boutons sont également présents :

- **Quit** pour fermer l'application,
- **Restart** pour relancer la prédiction et rediriger l'utilisateur vers le questionnaire.

Figure 16 : Interfaces de sortie



Source : Réalisée par les auteurs

Si l'utilisateur choisit de recommencer le questionnaire, les résultats des prédictions sont sauvegardés en mémoire. Il pourra ensuite télécharger ces résultats depuis la page d'accueil grâce au bouton *Download result* (cf. Figure 11). La base sauvegardée sera sous forme d'un fichier Excel enregistré à la racine du projet. Ce fichier contiendra les ID des prédictions dans le but de conserver l'anonymat des étudiants analysés, ainsi que leurs résultats et scores associés (cf. Figure 17).

Figure 17 : Résultats enregistrés

ID	Result	Score
1	Graduate	0.63
2	Dropout	0.88

Source : Réalisée par les auteurs

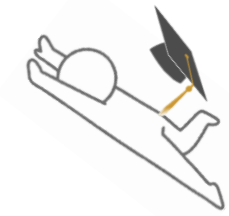
A la fermeture de l'application, toutes les données enregistrées sont supprimées pour un souci de sécurité des données sensibles que représentent les résultats des prédictions.

B. La liaison VBA/Python

Pour relier l'interface développée en VBA aux scripts Python, une procédure spécifique a été mise en place afin d'assurer une compatibilité optimale avec tous les ordinateurs où Python est installé. Cette procédure VBA récupère automatiquement le chemin absolu de l'exécutable *python.exe*. De cette manière, notre application peut fonctionner indépendamment de l'environnement d'installation de Python. En complément, nous avons également intégré une méthode pour identifier le chemin absolu du script Python *__main__.py*. Cette démarche garantit que l'application reste opérationnelle sur un large éventail de configurations système. Une fois ces deux chemins obtenus (celui de l'exécutable Python et celui du script principal), l'application collecte toutes les informations saisies par l'utilisateur dans les différentes boîtes de dialogue. Ces données sont concaténées sous forme d'une chaîne de caractères, où chaque valeur est séparée par un tiret (-). Cette structure simplifiée facilite leur traitement par le script Python. La commande finale d'appel au script se présente sous la forme suivante :

```
/chemin/de/python.exe /chemin/de/__main__.py boite1-boite2-boite3
```

Cette commande est exécutée dans le terminal PowerShell de l'ordinateur, permettant ainsi de lancer le script Python. Après exécution, le script transmet, via une commande print, les résultats sous forme d'une prédiction : la classe assignée à l'utilisateur ainsi que le pourcentage de probabilité associé. Ces informations sont ensuite récupérées par le code VBA, qui les stocke dans deux variables distinctes. Elles sont ensuite affichées sur une interface graphique dédiée, offrant une visualisation claire et conviviale des résultats pour l'utilisateur.



IV. Backend : les rouages de la machine

Le backend d'une application est la partie qui gère le bon fonctionnement de l'application. Contrairement au frontend, qui est l'interface utilisateur visible, le backend fonctionne en coulisses pour traiter les requêtes, exécuter des scripts et gérer les données. Le backend assure que les données sont correctement stockées, sécurisées et accessibles de manière efficace, permettant ainsi au frontend de fonctionner de manière fluide et réactive.

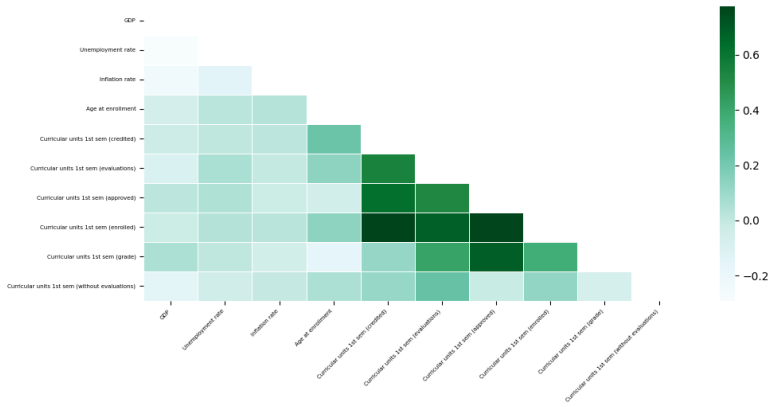
A. La prédiction

Nous allons prédire le statut de l'étudiant à la fin de la durée normale d'un cursus. Cela se décline en trois classes : la réussite, l'abandon ou le redoublement de l'étudiant. C'est donc un problème de classification. Avant de procéder à un modèle de prédiction, il faut choisir convenablement les variables.

1. Sélections de variables

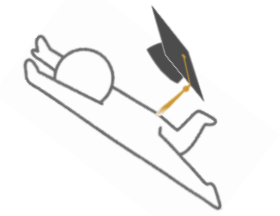
Les variables que nous avons à disposition sont présentes dans l'Annexes 1. Etant donné que nous avons beaucoup de variables, nous avons procédé à une sélection de variables.
Pour commencer, nous avons calculé la matrice de corrélation entre les variables quantitatives.

Figure 18 : Matrice de corrélation entre les variables quantitatives



Source : Réalisée par les auteurs

Concernant le taux d'inflation, l'âge à l'inscription, le taux de variation du PIB et le taux de chômage il existe une faible corrélation (cf. Figure 18). Néanmoins entre les autres variables il y a une forte corrélation (cf. Figure 18).



Par ailleurs, nous avons remarqué que généralement les variables concernant le premier semestre sont fortement corrélées avec celles du second semestre. Donc, nous pouvons prendre l'initiative de supprimer les variables du second semestre et garder celles du premier semestre pour permettre une évaluation le plus tôt possible.

En ce qui concerne les variables qualitatives, il n'existe pas de corrélation forte entre les variables.

Aucune différence significative des moyennes n'est observée pour le taux d'inflation, la variable indiquant si l'étudiant reçoit des besoins éducatifs spéciaux et la nationalité entre les groupes de la variable cible[3].

Pour faciliter les renseignements mis par les enseignements, nous avons décidé de ne pas inclure les variables macroéconomiques.

Pour résumer, nous avons enlevé les variables concernant le premier semestre, les variables macroéconomiques, la variable des besoins éducatifs spéciaux, la nationalité ainsi que la variable indiquant le statut créditeur de l'étudiant qui nous semble être une donnée sensible que peu de professeurs possèdent.

Comme nous avons pu le voir avec l'analyse de nos données, nous avons un grand nombre de variables explicatives qui peuvent contenir jusqu'à plus de 30 modalités ce qui augmente considérablement la dimension. Aux premiers abords, nous commençons par une régression logistique.

[3] Des tests d'indépendance (khi-deux et anova) ont été réalisés avec la variable à prédire et sont disponibles dans l'Annexe 2 (cf. Tableau 4 et Tableau 5).

2. Régression logistique

La régression logistique multinomiale est une généralisation de la régression logistique qui permet de traiter des problèmes de classification avec plus de deux catégories (ou classes) possibles.

L'objectif est de prédire une variable dépendante Y qui peut prendre K catégories différentes.

Soit $Y \in \{1, 2, \dots, K\}$, et une variable indépendante $X = (X_1, X_2, \dots, X_p)$ qui représente les prédicteurs (les variables explicatives).

Nous modélisons la probabilité de chaque catégorie $j \in \{1, 2, \dots, K\}$ à l'aide d'une fonction logistique (logit), mais pour chaque classe. Pour K classes, nous pouvons choisir généralement une classe de référence K et comparer les autres classes à cette référence.

La probabilité de la classe j avec $j \in \{1, 2, \dots, K-1\}$ est donnée par :

$$P(Y = j \mid X) = \frac{e^{X^T \beta_j}}{1 + \sum_{k=1}^{K-1} e^{X^T \beta_k}}$$

où :

$X^T \beta_j$ est le produit scalaire entre les prédicteurs X et les coefficients associés à la classe j,

β_j est un vecteur de coefficients pour la classe j

Cela permet d'exprimer les autres probabilités comme des fonctions des autres classes en termes de la classe de référence.

Pour la classe K, la probabilité est donnée par :

$$P(Y = K \mid X) = \frac{1}{1 + \sum_{k=1}^{K-1} e^{X^T \beta_k}}$$

Les coefficients β_j pour chaque classe sont estimés à l'aide de la méthode des moindres carrés généralisés (ou méthode de maximum de vraisemblance), où l'on cherche à maximiser la log-vraisemblance des observations données.

La log-vraisemblance est :

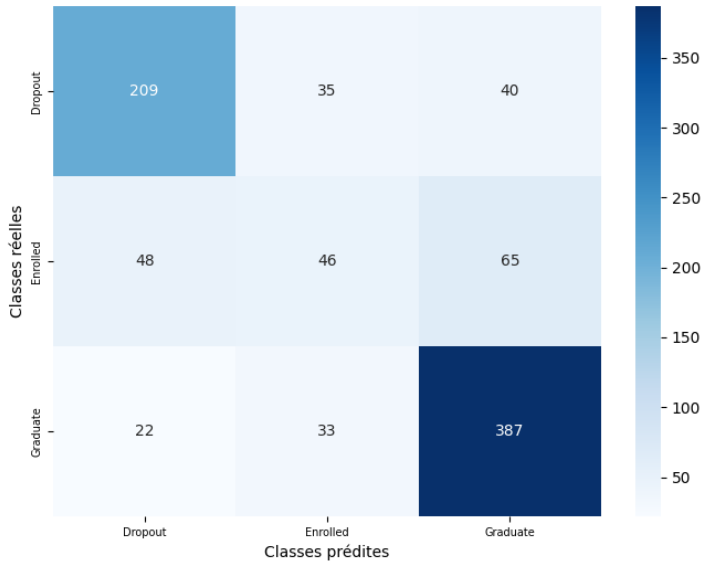
$$\ell(\beta_1, \dots, \beta_{K-1}) = \sum_{i=1}^{n} \log P(Y = y_i \mid X_i)$$

Nous maximisons cette fonction de vraisemblance pour estimer les paramètres (β_1 , … , β_{K-1}).

Lorsque nous effectuons une régression logistique la matrice de design doit être inversible. La matrice de design est une matrice où chaque ligne correspond à une observation et chaque colonne correspond à une variable explicative (plus, éventuellement, une colonne de 1 pour le terme constant). Par exemple, pour n observations et p prédicteurs, X est une matrice $n \times p$. Ainsi pour avoir l'inversibilité, les variables explicatives ne doivent pas être corrélés entre elles. Nous prendrons donc soin à ce qu'il n'y ait pas de multi colinéarité avant de procéder à une régression logistique.

Nous avons appliqué la régression logistique à toutes les variables que nous avons conservées à la suite de la sélection de variables et avons testé le modèle avec des données tests qui n'ont pas servi à l'entraînement du modèle. Nous obtenons ainsi la matrice de confusion ci-dessous.

Figure 19 : Matrice de confusion et rappel de la régression logistique



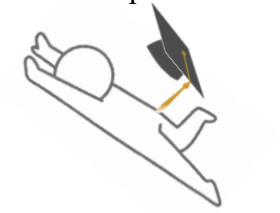
Target	Rappel	Support
Dropout	0.74	284
Enrolled	0.29	159
Graduate	0.88	442

Source : Réalisé par les auteurs

De nombreux critères existent pour la validation d'un modèle tel que la précision, le rappel ou le f1-score. Nous nous décidons à garder le critère « rappel ».

$$Rappel = \frac{Vrais\ positifs}{Vrais\ positifs + Faux\ négatifs}$$

Nous obtenons 88% pour la modalité Graduate, 29% pour celle de Enrolled et 74% pour Dropout ce qui nous fait une moyenne de 66%. Nous pouvons remarquer que la modalité Enrolled est moins bien prédite que les autres, cela réduit considérablement la précision de



notre modèle. Comme nous avons pu le voir dans l'analyse des données, la variable à prédire dispose d'un déséquilibre entre les classes avec la classe minoritaire Enrolled ce qui pourrait expliquer la mauvaise précision de cette classe. Par conséquent, nous avons pris la décision de transformer cette modalité. Soit en supprimant la modalité Enrolled de la base de données, mais cela ne résoudra pas le problème de déséquilibre des classes, soit en transformant la modalité Enrolled en Dropout, ce qui permettrait d'obtenir une variable cible équilibrée. Cela nous entraine à calculer la précision du modèle selon les deux cas comme ci-dessous.

Tableau 1 : Rappel des données tests selon les cas (remplacement de Enrolled ou suppression)

Target	Rappel	Support
Dropout	0.78	443
Graduate	0.86	442

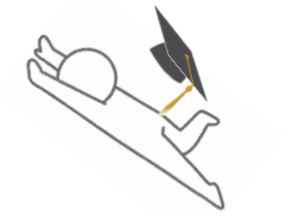
Target	Rappel	Support
Dropout/Enrolled	0.82	284
Graduate	0.95	442

Source : Réalisé par les auteurs

Nous avons préféré remplacer la modalité Enrolled par Dropout car la variable à prédire est plus équilibrée. De plus, la modalité Dropout, celle qui nous intéresse le plus, change très légèrement au niveau du rappel. Nous pouvons désormais définir la modalité Dropout comme étant la non-validation du cursus. Cela pourrait donc être le décrochage scolaire, la réorientation de l'étudiant ou le redoublement, ce qui vient rejoindre notre définition donnée plus tôt de décrochage.

Nous possédons une très grande dimension de 212 variables comprenant l'encodage des variables qualitatives, cela pourrait augmenter considérablement le risque de surapprentissage ainsi que la multicolinéarité. Il est donc essentiel de parvenir à sélectionner des variables.

Pour remédier à cela, nous décidons de tester la régression logistique avec une pénalisation lasso.



3. Régression logistique pénalisée avec lasso

La régression logistique Lasso ajoute un terme de pénalité (L_1) à la fonction de coût de la régression logistique classique. Ce terme de pénalité est la somme des valeurs absolues des coefficients, ce qui a pour effet de forcer certains coefficients à devenir exactement nuls, éliminant ainsi certaines variables du modèle.

La fonction de coût pour la régression logistique Lasso est donnée par :

$$\widehat{w}_{pen}(\lambda) = argmax_{w \in \mathcal{W}}(\frac{1}{n} l_n(w) - \lambda \|\overline{w}\|_1)$$

Où $\lambda > 0$

La présence du terme de pénalité force certains coefficients à devenir nuls, ce qui simplifie le modèle et peut améliorer sa capacité de généralisation.

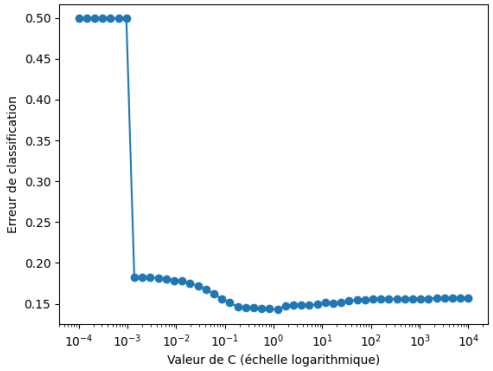
Le λ est généralement choisi de manière à diminuer l'erreur de prévision (par la méthode de validation croisée).

Tout d'abord, nous avons fait une régression logistique avec pénalisation lasso avec toutes les variables que nous avons gardé.

Nous testons 50 valeurs de C et 50 partitions de validation croisée (cf. Figure 20). Pour le critère de sélection, nous avons pris la précision du modèle nommé accuracy.

$$Accuracy = \frac{Nombre\ de\ prédictions\ correctes}{Nombre\ total\ de\ prédictions}$$

Figure 20 : Erreur de classification en fonction de C



Source : Réalisé par les auteurs

Le meilleur estimateur qui en résulte est $C = \frac{1}{\lambda} = 1.2$.

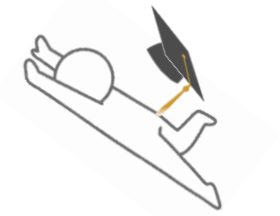
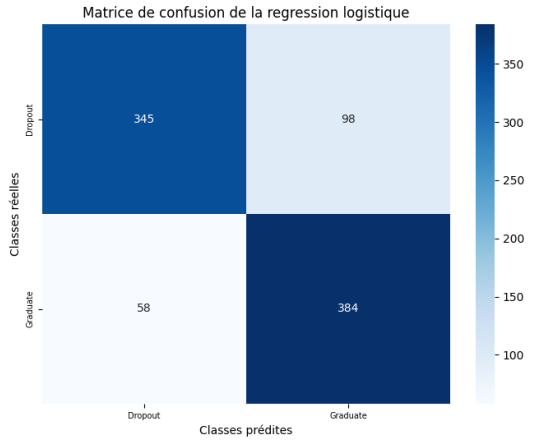


Figure 21 : Matrice de confusion de la régression logistique avec pénalisation lasso



Target	Rappel	Support
Dropout	0.78	443
Graduate	0.87	442

Source : Réalisé par les auteurs

Le rappel de ce modèle atteint 87 % pour la catégorie Graduate et 78% pour Dropout (cf. Figure 21) avec une moyenne de 82%.

Parmi les variables dont les coefficients sont mis à zéro, nous pouvons trouver la variable exprimant le statut international de l'étudiant, plusieurs modalités de la variable exprimant la précédente qualification de l'étudiant, de la variable exprimant la qualification du père, de la mère, le travail du père et de la mère, de la méthode d'application utilisée par l'étudiant ainsi que l'ordre dans lequel l'étudiant l'a appliqué, le statut matrimonial et la variable exprimant le type de formation. Au total, 128/212 variables ont été ignorées par la régression logistique. Comme nous l'avons mentionné, les variables qui possèdent plusieurs modalités n'ont pas été totalement supprimées puisque certaines modalités n'ont pas été mises à zéro par la régression logistique avec pénalisation lasso. Nous avons enlevé complètement ces variables et avons refait une régression logistique sans pénalisation. Les variables restantes sont : le nombre de crédits sans évaluations et avec évaluations, validés, crédités, inscrits, l'âge à l'inscription, le statut boursier, le statut exprimant si l'étudiant s'est déplacé de son domicile, le statut des frais de scolarité, si le cours a lieu en journée ou en soirée et le genre. Néanmoins, cette méthode de sélection de variables a supprimé les variables catégorielles possédant plus de 2 modalités ; au total, nous avons seulement 12 variables.

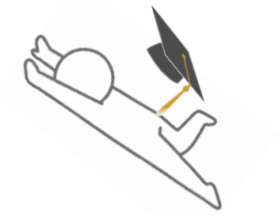
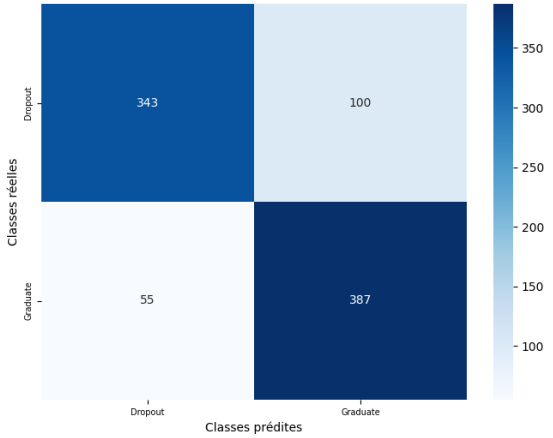


Figure 22 : Matrice de confusion et rappel de la régression logistique



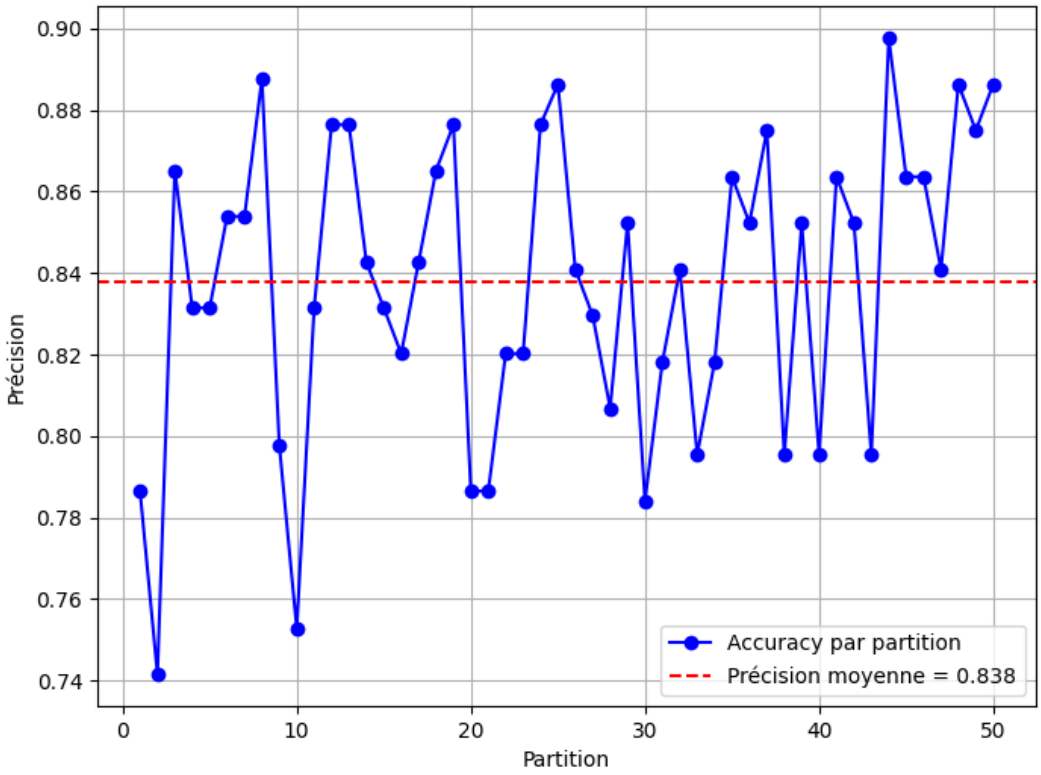
Target	Rappel	Support
Dropout	0.77	443
Graduate	0.88	442

Source : Réalisé par les auteurs

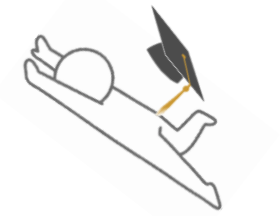
Par ailleurs, nous remarquons que le rappel a légèrement baissé (cf. Figure 22) ce qui nous rassure quant à la sélection de nos variables. La moyenne du rappel des deux classes vaut 82% ce qui équivaut à précédemment.

Par la suite, nous vérifions l'erreur de classification en fonction de l'accuracy par validation croisée (cf. Figure 23).

Figure 23 : Scores de précision pour chaque partition de la validation croisée



Source : Réalisé par les auteurs



Nous obtenons ainsi une précision (accuracy) moyenne de 83% (cf. Figure 23), ce qui nous indique que la performance du modèle est stable.

Pour explorer davantage les interactions potentielles entre les variables et améliorer les performances prédictives, nous optons pour une approche plus flexible en appliquant le modèle Random Forest sous le prisme de notre sujet.

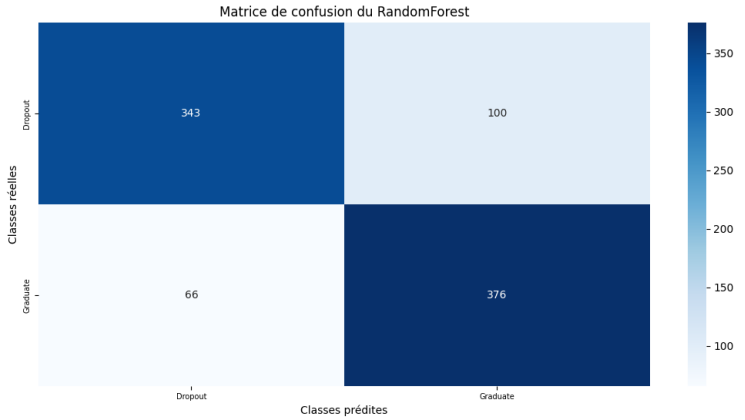
4. RandomForest

Chaque arbre de décision dans la forêt est construit à partir d'un sous-échantillon bootstrap. Lors de la construction de chaque arbre, à chaque nœud, un sous-ensemble aléatoire de caractéristiques est sélectionné. Cette sélection aléatoire de caractéristiques est importante car elle introduit de la diversité entre les arbres, ce qui réduit la corrélation entre eux et améliore la performance globale du modèle.

La sélection de la meilleure division à chaque nœud se fait en utilisant des critères comme l'indice de Gini.

Pour une nouvelle observation, chaque arbre de la forêt fait une prédiction. Pour les tâches de classification, la prédiction finale est obtenue par vote majoritaire. Cela signifie que la classe prédite le plus souvent par les arbres est choisie comme prédiction finale. Pour les tâches de régression, la prédiction finale est la moyenne des prédictions des arbres.

Figure 24 : Matrice de confusion et rappel du RandomForest



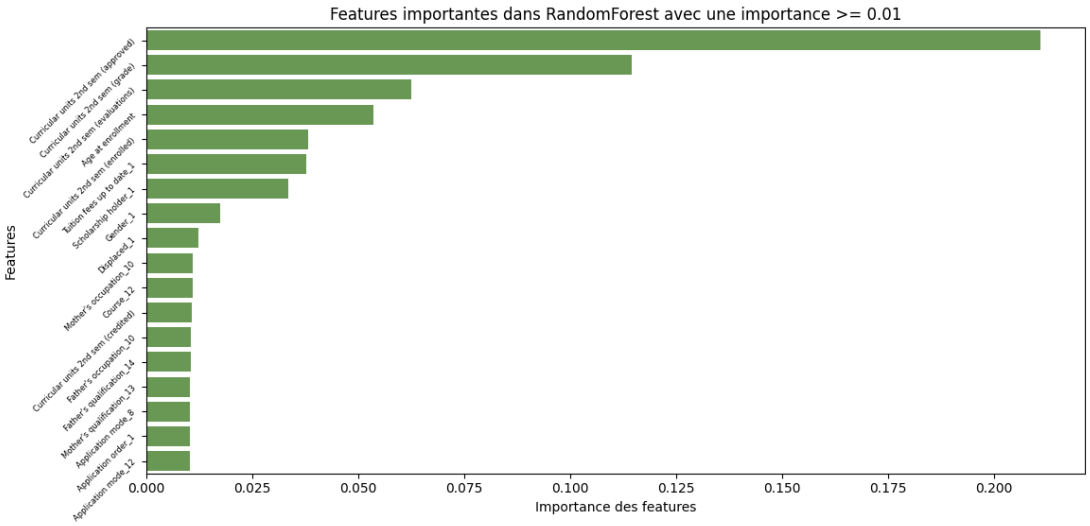
Target	Rappel	Support
Dropout	0.77	443
Graduate	0.85	442

Source : Réalisé par les auteurs

Ainsi, le modèle obtient un rappel moyen de 81% (cf. Figure 24).

Le Random Forest fournit des mesures d'importance des variables, ce qui permet d'identifier les caractéristiques les plus influentes dans les prédictions (cf. Figure 25).

Figure 25 : Classement des variables les plus importantes selon le RandomForest



Source : Réalisé par les auteurs

Nous remarquons, grâce au graphique ci-dessus, que le nombre de crédits validés, ainsi que la moyenne du second semestre, influence considérablement le modèle et que certaines variables n'ont pas beaucoup d'influence dans le modèle, telles que le travail de la mère et du père. Nous observons que les variables qui pèsent le plus sont parmi celles que nous avons choisies durant la régression logistique. Nous décidons donc de sélectionner les variables les plus importantes et de ne garder que les 8 premières (en incluant le genre). Les variables étant sélectionnées, nous pouvons estimer le nombre d'arbres nécessaires pour avoir la meilleure précision (cf. Figure 26).

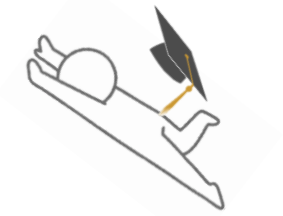
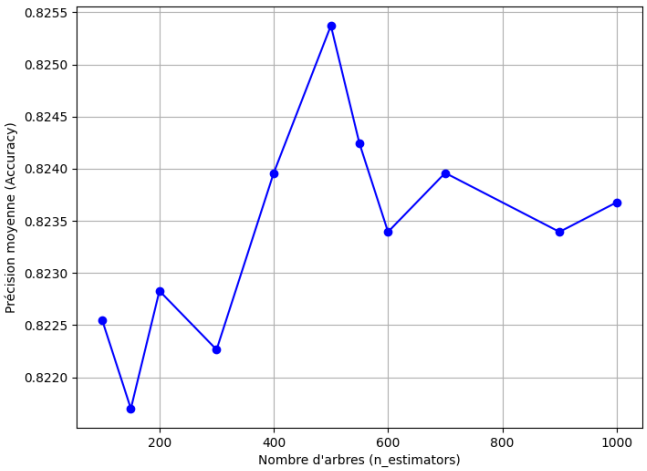


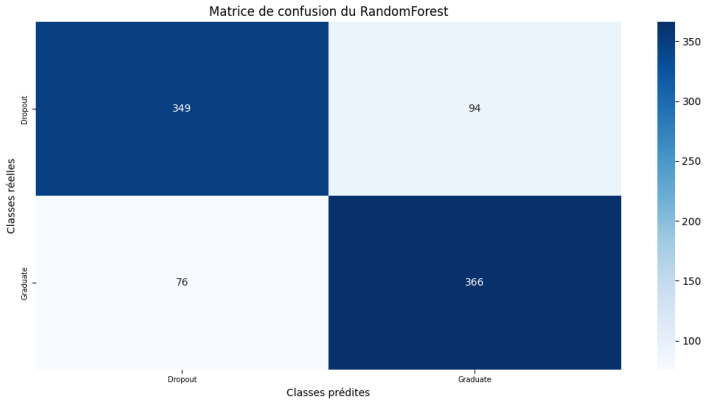
Figure 26 : Performance du modèle RandomForest en fonction du nombre d'arbres



Source : Réalisé par les auteurs

La précision ne varie que très légèrement, mais nous distinguons, à l'aide du graphique ci-dessus, que le nombre d'arbres optimal est de 500.

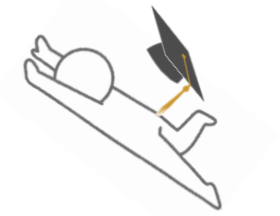
Figure 27 : Matrice de confusion et rappel du RandomForest



Target	Rappel	Support
Dropout	0.79	443
Graduate	0.83	442

Source : Réalisé par les auteurs

La sélection de variables a permis une augmentation du rappel de la variable Dropout malgré une baisse sur celui de Graduate (cf. Figure 27). Nous obtenons un rappel moyen de 81%.



5. Comparaison des deux modèles (Régression Logistique et Random Forest)

Chaque modèle ayant été entraîné sur des ensembles de variables différents, évaluons leurs performances en conséquence.

Tableau 2 : Rappels des modèles comprenant les variables sélectionnées par régression logistique

Target	Rappel (Random Forest)	Rappel (Régression Logistique)	Support
Dropout	0.78	0.77	443
Graduate	0.83	0.88	442
Moyenne	0.81	0.82	

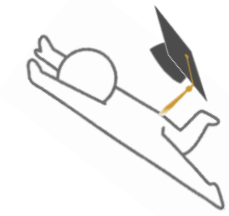
Source : Réalisé par les auteurs

Tableau 3 : Rappels des modèles comprenant les variables sélectionnées par le Random Forest

Target	Rappel (Random Forest)	Rappel (Régression Logistique)	Support
Dropout	0.79	0.76	443
Graduate	0.83	0.88	442
Moyenne	0.81	0.82	

Source : Réalisé par les auteurs

Nous décidons de garder le modèle du Random Forest comprenant 8 variables car ce modèle maximise le rappel de Dropout, modalité qui nous intéresse le plus, et est celui qui comprend le moins de variables à remplir pour le formulaire. Par conséquent, le formulaire à remplir sera davantage précis pour identifier les élèves qui n'auront pas leur diplôme et plus facile à remplir pour les enseignants.



Conclusion

« SchoolDrop'App » se positionne comme une réponse concrète à une problématique complexe : l'identification précoce et fiable des risques de décrochage scolaire. Ce projet, conçu dans un souci de simplicité et d'efficacité, allie une approche technologique avancée à une expérience utilisateur intuitive. L'interface de l'application a été soigneusement pensée pour ne présenter que les informations essentielles, permettant ainsi aux utilisateurs de se concentrer sur les résultats et les actions à entreprendre.

En adoptant le principe KISS ("Keep It Simple, Stupid"), nous avons optimisé la lisibilité et la convivialité de l'interface, garantissant ainsi que même des enseignants peu familiers avec les outils technologiques puissent l'utiliser sans difficulté. Grâce à une interface claire et minimaliste, associée à des modèles prédictifs performants, « SchoolDrop'App » offre une solution pragmatique, respectueuse des besoins des éducateurs et des contraintes des établissements scolaires.

Soucieuse des enjeux liés à la donnée, l'application veille à respecter les principes du RGPD en offrant une sécurité supprimant les données à chaque fermeture de l'application. Impossible ainsi de détourner les données confidentielles des étudiants via l'application.

Bibliographie

Présentation de l'EEES. Dans : *enseignementsup-recherche.gouv.fr* [en ligne]. [s. d.]. [Consulté le 30 décembre 2024]. Disponible à l'adresse : https://www.enseignementsup-recherche.gouv.fr/fr/presentation-de-l-eees-46573

Décrochage scolaire [en ligne]. [S. l.] : [s. n.], 25 septembre 2024. [Consulté le 20 octobre 2024]. Disponible à l'adresse : https://fr.wikipedia.org/w/index.php?title=D%C3%A9crochage_scolaire&oldid=218921306. Page Version ID: 218921306.

REALINHO, Valentim, MACHADO, Jorge, BAPTISTA, Luís et MARTINS, Mónica V. Predict students' dropout and academic success [en ligne]. Zenodo, 13 décembre 2021. [Consulté le 4 novembre 2024]. https://doi.org/10.5281/zenodo.5777340.

Annexes

Dans cette partie, nous retrouverons les annexes qui permettrons d'étayer le rapport.

Annexe 1. Metadata

L'annexe suivant décrit les caractéristiques principales d'un dataset ou d'une donnée, aussi appelé métas donnés.

Nom du jeu de donnes: Aucun
Auteurs : Valentim Realnho, Jorge Machado, Luis Baptista, Monica V. Martins
Langage : Anglais
Date de creation : 11 octobre 2022
Date de publication : 28 octobre 2022

Marital status : (entier) Categorie d'etat marital de l'etudiant

Application mode : (entier) La methode d'application utilise par l'etudiant

Application order : (entier) L'ordre dans lequel l'etudiant l'a applique

Course : (entier) Le cours pris par l'etudiant

Daytime/evening attendance : (entier) Si le cours a lieu le (1) en journée ou (2) en soirée

Previous qualification : (entier) Le diplome obtenu avant de s'inscrire dans l'enseignement supérieur

Nacionality : (entier) La nationalite de l'etudiant

Mother's qualification : (entier) Le diplome de la mere

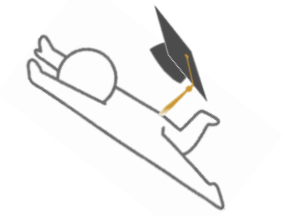
Father's qualification : (entier) Le diplome du pere

Mother's occupation : (entier) La profession de la mere

Father's occupation : (entier) La profession du pere

Displaced : (entier) Si l'etudiant a quitté de son domicile pour aller étudier

Educational special needs : (entier) Si l'etudiant a des besoins educatifs particuliers



Debtor : (entier) Si l'etudiant est debiteur

Tuition fees up to date : (entier) Si les frais de scolarite de l'etudiant sont a jour

Gender : (entier) Le genre de l'etudiant

Scholarship holder : (entier) Si l'etudiant est boursier

Age at enrollment : (entier) L'age de l'etudiant lors de l'inscription

International : (entier) Si c'est un etudiant international

Curricular units 1st sem (credited) : (entier) Le nombre d'unites curriculaires créditees par l'etudiant au cours du premier semestre

Curricular units 1st sem (enrolled) : (entier) Le nombre d'unites curriculaires inscrit par l'etudiant au cours du premier semestre

Curricular units 1st sem (evaluations) : (entier) Le nombre d'unites curriculaires evalue par l'etudiant au cours du premier semestre

Curricular units 1st sem (approved) : (entier) Le nombre d'unites curriculaires approuve par l'etudiant au cours du premier semestre

Curricular units 1st sem (grade) : (float) Moyenne obtenue par l'etudiant au cours du premier semestre

Curricular units 1st sem (without evaluations) : (entier) Le nombre d'unites curriculaires non note par l'etudiant au cours du premier semestre

Curricular units 2nd sem (credited) : (entier) Le nombre d'unites curriculaires créditees par l'etudiant au cours du second semestre

Curricular units 2nd sem (enrolled) : (entier) Le nombre d'unites curriculaires inscrit par l'etudiant au cours du second semestre

Curricular units 2nd sem (evaluations) : (entier) Le nombre d'unites curriculaires evalue par l'etudiant au cours du second semestre

Curricular units 2nd sem (approved) : (entier) Le nombre d'unites curriculaires approuve par l'etudiant au cours du second semestre

Curricular units 2nd sem (grade) : (float) Moyenne obtenue par l'etudiant au cours du second semestre

Curricular units 2nd sem (without evaluations) : (entier) Le nombre d'unites curriculaires non note par l'etudiant au cours du second semestre

Unemployment rate : (float) Taux de chomage

Inflation rate : (float) Taux d'inflation

GDP : (float) PIB de la région

Target : (character) Si l'etudiant a obtenu son diplôme, s'il a decroche ou s'il a redouble

Annexe 2. Résultats modèle de prédiction

Tableau 4 : Tests du khi-deux entre la variable à prédire et les autres variables explicatives

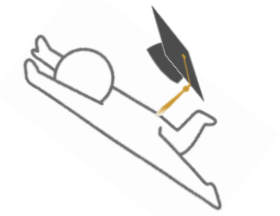
Variable	P-value	Différence significative ?
Debtor	4.9e-57	Oui
Educational special needs	7.3e-01	Non
Displaced	2.9e-13	Oui
Daytime/evening attendance	5.7e-07	Oui
Previous qualification	7.2e-30	Oui
Nacionality	2.4e-01	Non
Father's qualification	3.2e-19	Oui
Mother's qualification	5.8e-21	Oui
Father's occupation	4.5e-19	Oui
Application order	2.3e-09	Oui
Application mode	2.0e-77	Oui
Tuition fees up to date	1.5e-179	Oui
Marital status	8.1e-10	Oui
Mother's occupation	1.6e-31	Oui
Gender	2.2e-51	Oui
Course	2.3e-97	Oui

Source : Réalisée par les auteurs

Tableau 5 : Tests anova entre la variable à prédire et les autres variables quantitatives

Variable	P-value	Différence significative ?
GDP	8.28e-03	Oui
Unemployment rate	2.70e-03	Oui
Inflation rate	1.75e-01	Non
Age at enrollment	1.14e-65	Oui
Curricular units 1st sem (credited)	3.47e-04	Oui
Curricular units 1st sem (evaluations)	6.90e-17	Oui
Curricular units 1st sem (approved)	3.65e-316	Oui
Curricular units 1st sem (enrolled)	3.27e-26	Oui
Curricular units 1st sem (grade)	2.80e-269	Oui
Curricular units 1st sem (without eval)	1.11e-05	Oui

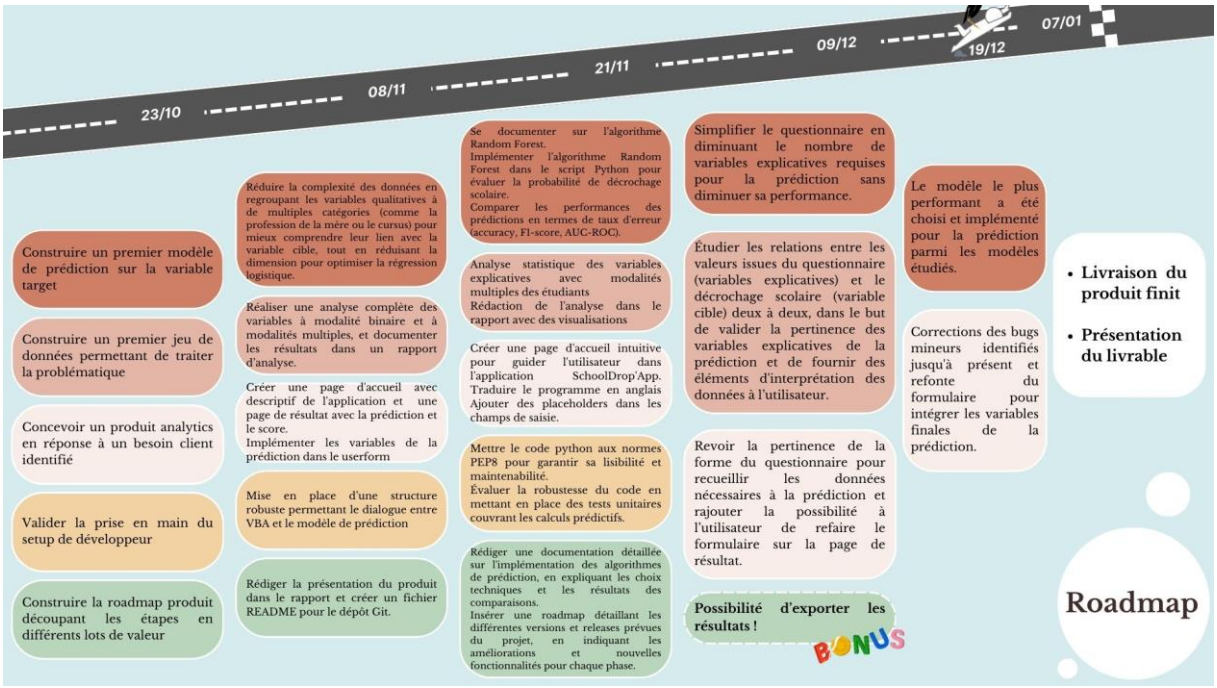
Source : Réalisée par les auteurs



Annexe 3. Roadmap

Dans cette annexe, nous pouvons retrouver la feuille de route (roadmap) (cf. Figure 28), détaillant les différentes versions et releases prévues du projet, en indiquant les améliorations et nouvelles fonctionnalités pour chaque phase.

Figure 28 : Roadmap du projet SchoolDrop'App



Source : Réalisée par les auteurs

- Première release : V0 (23/10/2024)

À ce stade, nous souhaitions réaliser les tâches suivantes : un modèle de prédiction a été choisi et testé sur la variable cible. Le choix a été argumenté et présenté au sein du rapport. La variable d'intérêt et les variables explicatives sont présentées de façon explicite et documentées au sein du rapport. Des maquettes d'interface sont construites et permettent de comprendre le parcours de l'utilisateur au sein de l'application. Un premier script Python exécutable est disponible au sein du projet GitLab, validant l'autonomie de l'équipe dans l'utilisation des outils de développement. Une roadmap produit détaillée par compétences est construite & découpée dans le temps en différents lots de valeur.

Nous sommes au prémisse de l'application, aucun modèle prédictif n'est encore implémenté, simplement en cours de production. Par ailleurs, via l'exploration du jeu de données, les variables explicatives sont identifiées et présentées ; il en va de même pour la variable d'intérêt. Un premier formulaire d'accueil est opérationnel pour la navigation dans l'interface utilisateur,

celui-ci permet de quitter l'interface ou de lancer le second formulaire. Dans ce deuxième formulaire, des menus déroulants sont opérationnels, ceux-ci n'ont pas de liens avec la prédiction. Il est possible, sur ce formulaire, d'interagir avec un bouton fonctionnel permettant de réinitialiser les objets présents dans le formulaire actif, mais aussi un bouton fonctionnel permettant de retourner au premier formulaire. Un dernier bouton pour confirmer les choix est disponible et renvoie à une boîte de dialogue indiquant les variables sélectionnées précédemment. Le formulaire est doté d'une protection, empêchant de laisser des valeurs demandées vides.

- Seconde release : V1 (08/11/2024)

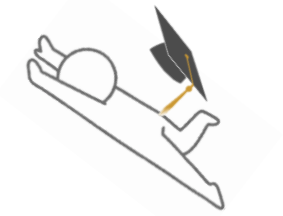
Nos objectifs pour cette version nous amenaient à accomplir les faits suivants : analyser les variables qualitatives à plus de 15 modalités avec la variable à prédire à l'aide d'une AFC, permettant de réduire la dimensionnalité. Les modalités similaires ont été regroupées de manière cohérente, et les associations significatives entre les variables ont été identifiées et documentées. Les résultats de l'AFC ont été présentés dans un rapport, comprenant des graphiques illustrant les relations détectées. L'analyse statistique est réalisée pour chaque variable binaire et est décrite en détail dans le rapport. Les variables à plusieurs modalités sont analysées par des méthodes statistiques descriptives et des représentations graphiques. Les résultats de l'analyse des variables à modalités multiples sont entièrement documentés dans le rapport. Une page d'accueil est disponible dans le fichier Excel à l'exécution, celle-ci contiendra les informations contextuelles de l'application. La page de résultat est opérationnelle et affiche la prédiction ainsi que le score de celle-ci. Les variables que l'utilisateur doit insérer dans le questionnaire sont désormais en lien avec les variables explicatives nécessaires à la prédiction. Une structure pour dialoguer entre VBA et le modèle de prédiction a été mise en place et est opérationnelle. Tous les tests unitaires associés au code de cette structure sont exécutés avec succès pour valider sa robustesse. Le fichier README est complet et disponible dans le dépôt Git, décrivant les principales fonctionnalités du produit et les instructions d'utilisation. Toutes les fonctionnalités du produit sont présentées et documentées de manière détaillée dans le rapport.

Pour cette version, le rapport se précise par l'ajout d'une analyse univariée et bivariée, et le répertoire git du projet est alimenté d'un README.md, permettant ainsi de continuer la dynamique de documentation du projet. Des fonctionnalités viennent s'implémenter dans

l'application, notamment un onglet d'informations contextuelles disponible dans l'application, mais aussi un onglet de sortie permettant d'afficher le résultat de la prédiction, qui n'est pas encore implémenté dans l'application. Une seconde sécurité a été ajoutée, empêchant l'utilisateur de rentrer dans les champs des valeurs qui ne sont pas appropriés. La liaison entre l'interface VBA et les codes de prédictions réalisés en Python est mise en place, il est cependant nécessaire de donner à l'application le chemin permettant de trouver l'exécuteur Python.

- Troisième release : V2 (21/11/2024)

Pour atteindre nos objectifs à ce moment, nous devions atteindre les exigences qui suivent : produire une documentation synthétique (notes ou résumé) sur le fonctionnement de l'algorithme Random Forest. Le script Python contient une implémentation fonctionnelle de Random Forest. L'algorithme est intégré au flux général de l'application et peut être exécuté via l'interface Excel. Les performances de Random Forest et de la régression logistique sont comparées à l'aide de plusieurs métriques (Accuracy, F1-score, AUC-ROC). Une conclusion sur la méthode la plus performante est formulée. Les statistiques descriptives de nos variables à modalités multiples permettent d'établir une vue d'ensemble de nos données concernant l'âge, la profession des parents, les cours et le nombre d'unités d'enseignement auquel l'étudiant s'est inscrit. Les résultats statistiques des variables à modalités multiples sont inclus dans le rapport avec des graphiques pertinents (diagramme en barre, camembert, ...) pour faciliter l'interprétation et ainsi rédiger le profil complet des étudiants de notre base de données. Une page d'accueil dédiée est créée sur une nouvelle feuille Excel. La page inclut des instructions claires pour guider l'utilisateur, des boutons de navigation et un design cohérent avec l'application. Tous les éléments du code sont traduits en anglais. Chaque champ TextBox dispose d'un placeholder contenant un exemple ou des instructions claires. Les placeholders sont visibles lorsque les champs sont vides, mais disparaissent lors de la saisie. Une documentation technique sur l'implémentation des algorithmes est rédigée, avec des détails sur les paramètres, les performances et les limitations. Celle-ci est intégrée dans le rapport. Les utilisateurs et développeurs peuvent comprendre comment et pourquoi chaque algorithme est utilisé dans le projet. Une roadmap claire et détaillée est incluse dans la documentation, listant les versions prévues avec leurs améliorations. Chaque phase de la roadmap est accompagnée de critères d'achèvement et de dates cibles.



Dans les faits, l'algorithme a été intégré dans l'application, ainsi elle est capable de prédire, en fonction du profil de l'individu, créé grâce aux variables explicatives demandées dans le formulaire, s'il est en situation de décrochage scolaire. L'interface Excel contient désormais une page d'accueil avec des instructions simples permettant de relancer le formulaire en fonction des besoins de l'utilisateur. La liaison entre l'interface VBA et Python est désormais bien plus confortable pour l'utilisateur, étant donné qu'il n'a plus d'interaction avec l'application, tout est automatique. La robustesse de l'application est assurée par ses tests unitaires et sa maintenabilité l'est tout autant depuis que son code est conforme aux normes PEP8. La documentation du projet est complétée par cette Roadmap, permettant de suivre en détail l'avancée du projet et le chemin emprunté par l'application au fil des versions.

- Quatrième release : V3 (09/12/2024)

Nos objectifs nous menaient à diminuer le nombre de variables explicatives selon leur degré d'importance et mesurer le degré de précision perdu. Implémenter un modèle simplifié utilisant moins de variables. Réaliser l'analyse bivariée sur toutes les variables explicatives du modèle de prédiction. Produire des visuels pertinents sur l'analyse. Interpréter les résultats de façon cohérente avec le contexte. Implémenter des CheckBox pour les choix binaires : genre (homme/femme), frais d'inscriptions payés (oui/non). Mettre en place des PlaceHolder pour indiquer à l'utilisateur les valeurs attendues. Ajouter un bouton sur la page de résultat pour permettre à l'utilisateur de refaire le formulaire.

Lors de ce sprint, nous avons revu complètement notre organisation et notre méthodologie de mise en place d'objectifs. Ceci nous a permis de gagner en efficacité et en clarté organisationnelle et d'ainsi compléter tous nos objectifs, et même de concrétiser une fonctionnalité qui initialement était prévue en bonus.

- Cinquième releases V4 (19/12/2024)

Pour concrétiser notre projet, les objectifs de ce dernier sprint étaient de résoudre les derniers bugs connus de notre application et de peaufiner notre algorithme de prédiction. Pour ce faire, nous souhaitions choisir et implémenter dans l'application le modèle le plus performant entre la régression logistique, la régression logistique lasso et RandomForest. Concernant l'interface, il était nécessaire que le bouton "reset choice" ne supprime plus l'aide à la saisie et ne modifie

plus la couleur de la police des valeurs saisies après la suppression. L'ID est mis à jour correctement lorsqu'une nouvelle valeur est enregistrée dans la page de résultats. Les questionnaires recueillent les bonnes variables, celle liée à la prédiction la plus optimale.

Annexe 4. Dépôt du projet SchoolDrop'App sur GitLab

« SchoolDrop'App » est disponible en open source sur GitLab, une plateforme de gestion de code et de collaboration pour les développeurs. GitLab permet d'héberger, de versionner et de suivre les évolutions du projet, tout en facilitant la contribution de la communauté. Pour télécharger ou mettre à jour l'application, une connexion Internet est requise afin d'accéder au dépôt GitLab et récupérer la dernière version du projet. Une fois téléchargée, l'application peut être utilisée hors ligne, garantissant ainsi une accessibilité continue, même sans connexion Internet. Le projet est disponible sur le lien suivant :

https://gitlab-mi.univ-reims.fr/gome0039/gpd_m2_sep