Léos Coutrot & Colin Coërchon 26 Novembre 2023

Projet : Prédire le décrochage ou la réussite scolaire des élèves

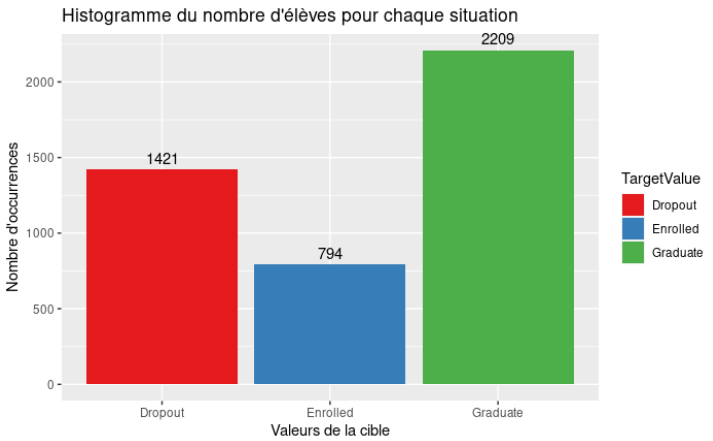
Introduction

La réussite scolaire, au centre des préoccupations depuis plus d'un siècle, notamment en France où l'égalité des chances éducatives demeure un sujet de débat majeur, constitue un domaine d'étude crucial et fascinant.

Présentation du jeu de données et de la variable cible

Nous disposons d'un jeu de données avec 37 variables et 4424 observations obtenues grâce à des élèves de l'Institut polytechnique de Portalegre (Portugal) entre 2009 et 2019. Le jeu de données est complet et ne présente aucune donnée manquante. Elles sont, de plus, relativement riches (puisque nous avons n>>p). Parmi les différentes variables à notre disposition, nous avons notamment le genre, l'âge, le niveau d'études de la mère ou du père (qui permettent d'avoir un aperçu de l'environnement familial des élèves).

Le jeu de données comporte une colonne du nom de "Target". C'est une variable catégorielle qui indique si l'élève en question à réussi ou non à valider son année. Elle sera donc **notre variable cible** dans cette étude. Elle a trois valeur possibles : Graduate", "Dropout" et "Enrolled". Pour simplifier l'étude des données et se rapprocher de notre cours, nous allons réduire cette variable à deux valeurs possibles : 1 pour la valeur "Graduate" et 0 sinon. Ainsi, notre étude va se concentrer uniquement sur le fait de savoir si l'élève a complètement validé son année ou non.



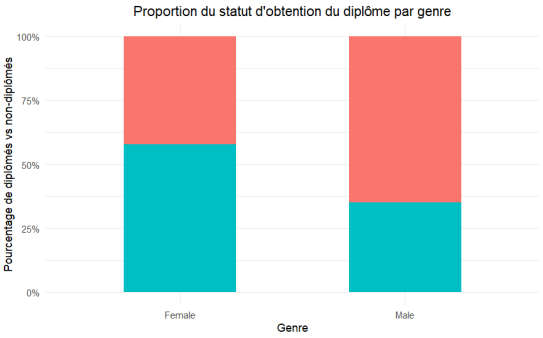
Problématique

L'objectif principal de notre étude est d'identifier les facteurs clés influençant la réussite ou l'échec académique des étudiants, afin de minimiser l'échec dans l'enseignement supérieur.

La question centrale de notre travail est donc la suivante : **Quelles sont les variables ayant un impact significatif sur la performance scolaire des étudiants ?** Pour répondre à cette interrogation, nous utiliserons **la régression logistique**, une méthode statistique robuste et adaptée pour modéliser les relations entre plusieurs variables et d'autant plus pertinente qu'ici notre variable cible est une variable catégorielle. Cette approche nous permettra de déterminer les facteurs les plus pertinents pour la réussite académique et d'établir un modèle prédictif pour identifier les étudiants à risque.

Étude du jeu de données

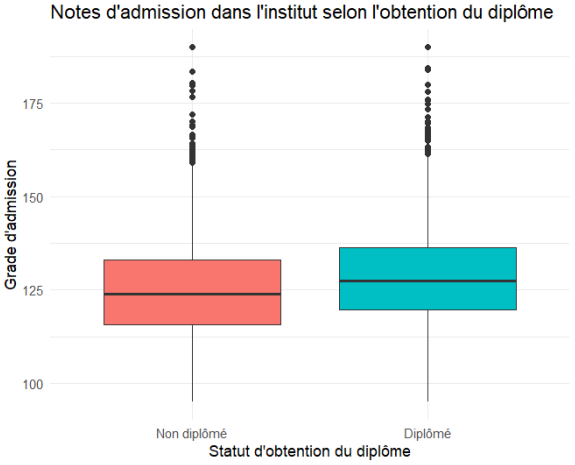
Commençons par étudier quelques variables qui nous ont semblé pertinentes :



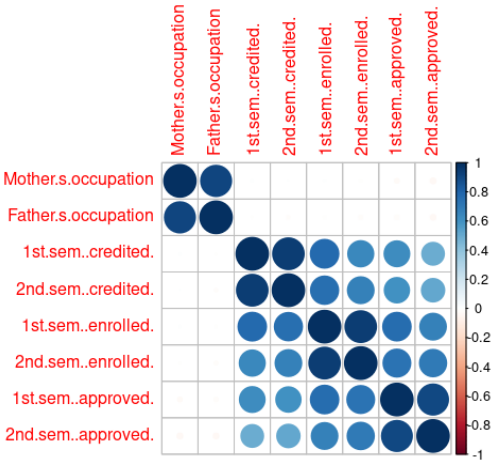
Pour commencer, ce premier graphique montre une différence notable dans les taux de réussite entre **les genres**, avec environ 60% de réussite chez les femmes contre 35% chez les hommes. Cette disparité suggère que le genre pourrait être un prédicteur significatif de la réussite académique.

Ensuite, nous nous sommes posés la question de savoir si les résultats finaux avaient un lien avec **les précédents résultats des élèves**. La première boîte à moustaches présente la corrélation entre les notes du premier semestre et la réussite finale. Et effectivement, la mise en parallèle de ce facteur avec des données récentes sur les élèves (derniers résultats) suggère qu'il joue un rôle notable dans les échecs académiques.

Mais, à l'inverse, la dernière analyse montre que les **notes d'admission initiales** à l'institut ne sont pas un indicateur fiable de la réussite académique. La faible différence de moyenne entre les groupes (~ 5 points sur 200) indique que les performances initiales à l'admission ne prédisent pas nécessairement les résultats finaux.



Enfin, si on calcule la matrice de corrélation, on remarque tout de suite que **certaines variables sont très corrélées**. Par exemple, ce graphique représente la corrélation entre toutes les variables fortement corrélées du jeu de données (chacune ayant plus de 0.9 de corrélation avec une autre variable). Ces variables peuvent ouvrir potentiellement la voie à d'autres modèles (en supprimant par exemple une variables parmi deux très corrélés) qui peuvent s'avérer plus pertinents. Cependant, il est bon de rappeler qu'une forte corrélation ne veut pas forcément dire qu'il est bon de supprimer l'une des deux variables, parfois la différence contient justement des informations très importantes pour le modèle, **il faudra donc étudier attentivement chacune des possibilités**.



Ces analyses préliminaires, bien que révélatrices, ne sont que le début de notre exploration. Elles s'inscrivent clairement dans une logique statistique Bayésienne, où chaque nouvelle information affine notre compréhension et notre modèle prédictif. Des variables telles que le niveau d'éducation, la profession des parents, la fréquentation en journée ou en soirée, etc… restent à explorer.

Notre objectif final est de développer des **modèles de régression logistique pertinents** qui intègrent ces variables pour prédire au mieux la réussite académique des élèves.