UCLouvain

UNIVERSITÉ DE NAMUR

**Economics School of Louvain - ESL**

**Economics School of Namur - ESN**

# Quelles sont les caractéristiques des écoles de la chance ?

# Une approche par le Machine Learning


Auteur : Tom Van Zeebroeck

Directeur : Jean Hindriks

Lecteur : Benoît Decerf



Année académique 2019-2020

Master en sciences économiques – 120 crédits – Finalité : spécialisée



Université catholique de Louvain - Economics School of Louvain - Place Montesquieu 3, 1348 Louvain-La-Neuve

Université de Namur – Economics School of Namur – Rempart de la Vierge 8, 5000 Namur

# Remerciements

Tout d'abord, je souhaiterais remercier mon promoteur, Jean Hindriks. Je le remercie pour ses conseils avisés, ses feedbacks, ses encoragements et sa disponibilité durant cette période compliquée.

Ensuite, je voudrais remercier mon lecteur, Benoît Decerf pour le temps qu'il consacrera à lire ce travail.

J'aimerais aussi remercier François Fouss pour ses conseils et ses suggestions pour mon analyse de Machine Learning.

Finalement, je remercie ma famille pour le soutien qu'elle m'a apporté pendant la réalisation de ce travail.

# Table des matières

**1 Introduction** **4**

**2 Revue de la littérature** **6**

**3 Identification des écoles de la chance** **10**

**4 Evolution de la comparaison entre communautés entre 2012 et 2018** **13**
4.1 Données . . . . . 13
4.2 Résultats . . . . . 13
4.3 Robustesse . . . . . 16
4.4 Discussion . . . . . 18
4.5 Conclusion . . . . . 21

**5 Caractéristiques des écoles de la chance au niveau international** **23**
5.1 Le Machine Learning et les différents algorithmes utilisés dans ce mémoire . . . . . 23
5.2 Littérature et méthodologie . . . . . 30
5.3 Construction de la base de données . . . . . 31
5.4 Prétraitement des données . . . . . 35
5.5 Résultats . . . . . 37
5.6 Discussion et interprétation . . . . . 45
5.7 Conclusion . . . . . 52

**6 Conclusion** **55**

**A Annexes** **62**

# 1. Introduction

Ce mémoire traite du lien entre le niveau socioéconomique des élèves et leurs résultats scolaires. Le lien entre niveau socioéconomique et résultats scolaire est très largement documenté dans la littérature scientifique. Les élèves issus de milieux défavorisés évoluent dans un contexte différent de celui des élèves issus de milieux favorisés. Les parents qui ont un statut socioéconomique favorable sont plus susceptibles d'enseigner à leur enfant les comportements et références culturelles considérées à l'école. Ils sont aussi plus susceptibles de fournir à leur enfant un support financier et des ressources matériels pour leur apprentissage. Ils sont aussi plus susceptibles d'offrir un climat d'apprentissage stimulant à domicile (Thomson (2018)). Il existe donc une relation négative importante et persistante entre le niveau socioéconomique des élèves et leurs résultats scolaires (Coleman *et al.* (1966) ou Sirin (2005) par exemple).
Ceci est problématique dans la mesure ou les résultats scolaires ont des implications importantes à long terme. Par exemple, il existe un lien très fort entre éducation et revenus (Card (1999)). D'un point de vue individuel, cela suggère que les individus issus de milieux défavorisés n'ont pas les mêmes opportunités de réussites sociales. D'un point de vue sociétal, l'absence de mobilité intergénérationnel participe à la perpétuation des inégalités dans la société et menace par la même occasion la croissance économique (Cingano (2014)).
Influencée par le très influent rapport de Coleman démontrant le rapport négatif entre niveau socioéconomique des élèves et leurs résultats scolaires (Coleman *et al.* (1966)), le courant de pensée dominant a longtemps été que l'école ne savait rien y changer et avait par conséquent un rôle minime à jouer pour mitiger ces inégalités. Depuis, les choses ont évolué et la recherche empirique à montrée que l'école pouvait bien jouer un rôle important pour mitiger les inégalités sociales à l'école (Downey et Condron (2016)). Les tests PISA(Programme international pour le suivi des acquis des élèves) montrent par exemple que la relation entre niveau social et niveau scolaire, bien qu'elle soit négative partout, varie très fort d'un système scolaire à l'autre (OECD (2019a)).

Dans un tel contexte, il est utile de s'intéresser aux écoles qui jouent le rôle d'ascenseur social. Ces écoles qui parviennent à faire réussir les élèves de milieux défavorisés et permettent d'améliorer l'équité du système scolaire.
Dans cette quête à la mobilité sociale, il serait dangereux d'oublier la dimension d'efficacité, à savoir l'accumulation de compétences. En effet, l'accumulation de compétences reste l'objectif principal d'un système scolaire. A ce sujet, il me paraît utile de mentionner que ces 2 éléments ne doivent pas forcément être opposés. Godin et Hindriks (2018), Parker *et al.* (2018), OCDE (2019a) par exemple montrent qu'il existe un lien positif entre mobilité à l'école et efficacité d'un système scolaire.

Dans ce mémoire, je propose de repartir de l'approche introduite par Hindriks et Godin (2020) pour identifier les écoles qui combinent mobilité sociale et efficacité scolaire, appelées « écoles de la chance ». Cette approche consiste à faire passer un

double test d'efficacité et de mobilité sociale aux écoles. La méthodologie d'identification des écoles de la chance est détaillée plus en profondeur dans la section 3 de ce travail. Cette méthodologie d'identification des écoles de la chance trouve deux applications dans ce mémoire.

Dans un premier temps (section 4), je m'intéresse à la répartition des écoles de la chance au sein des communautés en Belgique. Plus précisément, je regarde comment la répartition des écoles de la chance au sein des communautés à évoluée entre 2012 et 2018. En utilisant les données PISA 2012, Hindriks et Godin (2020) trouvaient un écart important entre les proportions d'écoles de la chance au sein de chaque Communauté. En effet, la proportion d'écoles de la chance était bien plus élevée en communauté flamande qu'en communauté française. Dans ce mémoire, l'analyse est répliquée en utilisant les données PISA 2015 et 2018.

Ensuite, je cherche à identifier les caractéristiques communes des écoles de la chance (section 5). Autrement dit, la question que je me pose est la suivante : « Quelles sont les caractéristiques qui différencient les écoles de la chance des autres écoles ? ».
Pour répondre à cette question, j'utilise les données de la dernière vague d'enquête PISA 2018 pour les pays de l'OCDE (Organisation de Coopération et de Développement Economique). Les méthodes utilisées sont des méthodes de Machine Learning. Ces méthodes me permettent de sélectionner et de hiérarchiser les variables qui permettent de distinguer les écoles de la chance parmi un grand nombre de variables en faisant abstraction de tout idée préconçue. Pour être clair, le Machine Learning ne prétend pas identifier des relations causales. Il permet simplement d'établir des associations et des corrélations un utilisant les données. Les données en coupe transversale comme les données PISA ne permettent de toute façon pas d'identifier de relations causales.
Dans ce mémoire, les résultats de trois algorithmes de machine learning sont comparés pour sélectionner et hiérarchiser les variables importantes : Un algorithme de Support Vector Machine (avec élimination récursive de caractéristiques (RFE))), une régression logistique (avec élimination récursive de caractéristiques (RFE)) et un arbre de décision. A l'issue du processus, les variables les plus importantes pour discrimer les écoles de la chance des autres écoles sont sélectionnées et hiérarchisées.

Pour résumer, ce mémoire se présente de la façon suivante :
Dans la seconde section, je passe en revue la littérature. Une attention particulière est portée aux mesures de l'équité dans le contexte de l'éducation et aux caractéristiques des écoles qui sont liées à la mobilité sociale. Je conclus cette section en positionnant ce mémoire dans la littérature existante.
Dans la troisième section je rappelle la méthode introduite par Hindriks et Godin (2020) et utilisée dans ce mémoire pour identifier les écoles de la chance.
Dans la quatrième section, j'utilise la méthode d'identification présentée à la section 3 à l'analyse de l'évolution de la proportion d'écoles de la chance au sein des communautés en Belgique.
Dans la cinquième section, j'utilise la méthode qui a été présentée à la section 3 pour identifier les caractéristiques des écoles de la chance. Pour cela, j'utilise des méthodes de Machine learning. Finalement, la dernière section conclut.

# 2. Revue de la littérature

Cette revue de la littérature entend donner un aperçu de la littérature relative à l'équité des écoles. Une attention particulière est portée la façon dont l'équité est mesurée et aux différentes caractéristiques auquel elle est reliée dans la littérature.

*Les indicateurs de mobilité sociale à l'école*

Je commence par passer en revue différents indicateurs utilisés pour mesurer la mobilité sociale à l'école. Ensuite, je compare l'indice de mobilité sociale utilisé dans ce travail à ces autres indicateurs. Les indicateurs de mobilité qui sont passés en revue sont essentiellement les indicateurs qui sont développés par l'OCDE.
Trois indicateurs sont fréquemment utilisés par l'OCDE pour mesurer la mobilité sociale dans le contexte de l'éducation : l'intensité et la pente du gradient social et le taux d'élèves résilients.

- L'intensité du gradient social, indique la variance des résultats des élèves qui est expliquée par le niveau socioéconomique de ces élèves.
- La pente du gradient socioéconomique mesure le changement de performance moyen qui est associé à un changement de niveau social.
- Un élève résilient est un élève du bas de l'échelle sociale qui parvient à obtenir des bons résultats scolaires.

Formellement, ces deux premiers indicateurs sont représentés respectivement par le $R^2$ et la pente de la droite de régression du score sur le niveau social. Dans les rapports de l'OCDE, ces deux indicateurs sont principalement utilisés pour comparer entre eux différents systèmes scolaires.
L'intensité du gradient sociale en tant qu'indice repose sur l'hypothèse de linéarité de la relation entre score et niveau social. En effet, l'estimation de la mobilité est biaisée si cette relation n'est pas linéaire. L'indicateur va par exemple surestimer la mobilité sociale d'un pays dont la relation entre résultats scolaires et niveau social est exponentielle (Godin et Hindriks (2018)). Par ailleurs, il est aussi très sensible à la variance des résultats et du niveau social au sein d'un pays. En effet, plus la variation de l'indice socioéconomique est élevée, plus le pouvoir explicatif du modèle linéaire sera élevé. Et au plus les résultats scolaires varient, au plus le pouvoir explicatif du modèle linéaire est faible. Les inégalités sociales et scolaires au sein d'un pays ont donc un rôle mécanique sur l'intensité du gradient sociale (Godin et Hindriks (2018)). Un troisième indicateur utilisé par l'OCDE est le taux d'élèves résilients. Il est déterminé par la proportion d'élèves défavorisés (quartile inférieur) qui se retrouvent parmi les élèves les plus performants (quartile supérieur). Cet indicateur est utilisé pour les comparaisons internationales mais aussi pour comparer les élèves et les écoles au sein d'un même pays. De façon générale, ce principe qui consiste à s'intéresser aux élèves défavorisés qui parviennent à atteindre de bons résultats scolaires est très largement utilisé dans la littérature relative à l'équité dans l'éducation.

L'approche de mobilité sociale utilisée dans ce mémoire se distingue de ces approches. L'indice utilisé ici est la mobilité interdécile. Le principe de cet indice développé par Godin et Hindriks (2018) est de comparer le rang scolaire des élèves à leur rang social. Plus précisément, il s'agit de prendre le rang scolaire sur le rang social. L'indice est présenté en détail dans la section suivante.
Par rapport à l'intensité du gradient social, cette méthode présente l'avantage de ne pas dépendre de la non-linéarité de la relation entre résultats et indice socioéconomique. Elle ne dépend pas non plus des inégalités scolaires ou sociales au sein d'un pays. Par ailleurs, elle permet de prendre en compte les élèves résilients, ces élèves qui ont de bons résultats scolaires en venant d'un milieu social défavorisé (Godin et Hindriks (2018)). Par rapport au taux de résilience, la mobilité sociale présente l'avantage de prendre en compte l'ensemble de la distribution et pas uniquement les élèves issus de milieux défavorisés (Godin et Hindriks (2018)). Une comparaison plus détaillée de ces trois indicateurs avec la mobilité interdécile peut être trouvée dans Godin et Hindriks (2018).

*Les caractéristiques des écoles liées à la mobilité sociale*

Pour évaluer la mobilité sociale au niveau des écoles, l'OCDE utilise le concept d'élèves résilients. Dans les derniers rapports de l'OCDE, les caractéristiques des écoles qui sont mises en lien avec la proportion d'élèves résilients sont notamment : le niveau socioéconomique de l'école (OECD (2018)), le climat disciplinaire qui règne à l'école (OECD (2018)), l'absence aux cours (OECD (2018))) et l'enthousiasme du professeur (OECD (2019a)). Les écoles avec un niveau socioéconomique favorable sont associées à un taux de résilience supérieur de 22 points de pourcentage. Les élèves qui évoluent dans un climat disciplinaire favorable ont un taux de résilience de 7 points de pourcentage supérieur aux autres. Une absence aux cours dans les 14 derniers jours est associée à un taux de résilience supérieur de 7 points de pourcentage. Un taux de résilience supérieur parmi les étudiants issus de l'immigration est également observé avec un professeur plus enthousiaste.

Agasisti *et al.* (2018), Agasisti et Longobardi (2017) et Agasisti et Longobardi (2014) s'intéressent aux caractéristiques des écoles qui ont une influence sur la probabilité d'être un élève résilient. Pour ce faire, ils utilisent une régression logistique. Agasisti et Longobardi (2017) trouvent un effet très important du climat disciplinaire en classe sur la probabilité qu'un élève soit résilient en s'intéressant à 15 pays de l'Union européenne en utilisant les données PISA 2009. Dans une moindre mesure, ils trouvent également un effet positif de l'organisation d'activités créatives à l'école sur la résilience.
Ces résultats sont confirmés par Agasisti *et al.* (2018) sur l'ensemble des pays de l'OCDE entre 2006 et 2015.
Finalement, Agasisti et Longobardi (2014) trouvent un effet positif de la qualité des ressources d'apprentissage sur la probabilité qu'un élève soit résilient sur base des données PISA 2009 en Italie.

Alivernini *et al.* (2016) utilisent un arbre de décision pour identifier les élèves résilients sur base des données PISA 2012. Les élèves considérés comme résilients sont les élèves forts (5%) qui font partie d'un groupe (élèves qui partagent des caractéristiques sociales similaires) au sein duquel la proportion d'élèves ayant un niveau

faible est élevée (95%). Une fois ces groupes identifiés, une régression logistique est utilisée pour voir quelles caractéristiques différencient ces élèves résilients des autres élèves. Ils trouvent un effet positif pour les pratiques d'enseignements (Alivernini *et al.* (2016)).

Dans une revue de la littérature accompagnée d'une étude de cas des écoles performantes, Shaw *et al.* (2017) mettent en évidence différentes caractéristiques des écoles susceptibles d'influencer la mobilité sociale des étudiants. Il ressort qu'il y a un lien entre l'intérêt des parents et leur implication dans la scolarité de leurs enfants et la performance des élèves. Grouper les élèves en groupes de niveau est susceptible d'avoir un impact négatif sur la progression des élèves des groupes d'élèves faibles (Méta annalyse de EEF (2016)). La culture de l'école et notamment du point de vue du climat disciplinaire semblent également être importants tout comme la qualité de l'enseignement.

Finalement, Hindriks et Godin (2020) proposent d'étudier l'équité des écoles en la couplant à l'efficacité. Les deux aspects sont utilisés simultanément en faisant passer un double test aux écoles. L'indicateur de mobilité sociale utilisé est la mobilité interdécile. Dans un premier temps, la méthodologie est appliquée au niveau belge pour comparer entre elles les communautés ainsi que les réseaux d'enseignement. Dans ce mémoire, l'étude de la comparaison des communautés sera reprise et son évolution sera analysée. Ensuite, l'effet de différentes caractéristiques est testé sur la probabilité d'être une école de la chance au moyen d'une régression logistique. La mixité sociale, la proportion de redoublants, le niveau social de l'école sont notamment des facteurs qui sont significativement associés aux écoles de la chance.

Dans ce mémoire, je pars du même principe. Dans un premier temps, je m'intéresse à la comparaison des communautés et à son évolution. Ensuite, je m'intéresse aux caractéristiques communes des écoles de la chance. Dans ce mémoire, un ensemble de caractéristiques plus large sera étudié et les caractéristiques les plus importantes pour discriminer les écoles de la chance des autres seront recherchées et hiérarchisées au moyen d'algorithmes de Machine Learning. Un aperçu de la littérature relative aux applications de Machine Learning sur les données PISA sera présenté dans la partie 2. Cela permettra de positionner les méthodes utilisées dans ce mémoire par rapport à la littérature existante.

Ce mémoire s'inspire donc fort de l'approche utilisée par Hindriks et Godin (2020) avec :

- Deux contributions majeures
  - L'utilisation du Machine Learning pour rationaliser le choix des variables explicatives
  - L'utilisation du Machine Learning pour hiérarchiser ces variables
- Et deux contributions secondaires
  - L'actualisation de la comparaison des communautés réalisée par Hindriks et Godin (2020) pour 2012 en utilisant les données pour 2015 et 2018.

– L'identification d'un rattrapage des écoles flamande par les écoles francophones en termes de proportion d'écoles de la chance

# 3. Identification des écoles de la chance

Dans la première partie, je regarde l'évolution de la proportion d'écoles de la chance dans chaque communauté en Belgique entre 2012 et 2018. Dans la seconde partie, je cherche les caractéristiques communes de ces écoles de la chance. L'étape préalable dans les deux cas est d'identifier les écoles de la chance.
La méthode qui est utilisée dans ce mémoire pour identifier les écoles de la chance est introduite par Hindriks et Godin (2020). L'objectif de cette section est de rappeler cette méthode. La méthode consiste à faire passer un double test aux écoles. Le premier test concerne la mobilité sociale (équité). Le second concerne l'efficacité. Une école est considérée comme une école de la chance si elle réussit ces deux tests. Afin de pouvoir évaluer et comparer les écoles du point de vue de l'équité et de l'efficacité, il y a lieu de mesurer ces deux aspects. Dans la suite de cette section, je présente les indicateurs de mobilité sociale et d'efficacité utilisés pour classifier les écoles tels qu'ils sont définis dans Hindriks et Godin (2020). Je commence par présenter la mesure d'équité puis je présente la mesure d'efficacité.

## L'équité

L'indicateur d'équité utilisé dans Hindriks et Godin (2020) est la mobilité inter-décile [1]. Cet indicateur compare le score d'un élève avec son origine socioéconomique. Concrètement, les élèves sont classés selon leur score au test et selon leur origine socioéconomique. Dans le cadre de ce mémoire, les élèves seront classés en percentiles. La position d'un élève dans le classement scolaire est comparée avec sa position sociale en prenant le rapport de la position scolaire sur la position sociale. Cette relation est formalisée dans l'équation 3.1 [2] qui décrit la mobilité sociale d'une école $s$.

$$Equité = 1/n_s \sum_{i=1}^{n} (y_i/x_i)^r \tag{3.1}$$

où :
$n_s$ = le nombre d'élèves dans l'école s
$y_i$ = la position scolaire de l'étudiant i
$x_i$ = la position social de l'étudiant i

Il s'agit donc de diviser la position scolaire de l'étudiant par sa position sociale. Un étudiant avec une position scolaire supérieure à sa position sociale est en mobilité ascendante et sa mobilité sociale est supérieure à 1. A l'inverse, un élève dont la position scolaire est inférieure à la position sociale est en mobilité descendante et a une mobilité inférieure à 1.
Le paramètre $r$ influence sur le poids qui est donné à la mobilité sociale des élèves du bas de l'échelle sociale et à l'intensité de la mobilité. Une valeur de $r$ qui tend

1. En fait, les auteurs classent les élèves en percentiles sur base de leur indice socioéconomique et de leur score au test PISA. Par abus de langage, ils définissent la mobilité inter percentile en mobilité inter décile

2. issue de Hindriks et Godin (2017)

vers 1 correspond à l'indicateur de mobilité interdécile. Celui-ci donne un poids plus important à la mobilité sociale des élèves du bas de l'échelle sociale et à l'intensité de la mobilité sociale. Un élève qui se situe dans le bas de l'échelle sociale a un potentiel de mobilité plus important qu'un élève qui se situe dans le haut de l'échelle. La mobilité interdécile d'une école est mesurée par la moyenne des mobilités interdéciles individuelles des étudiants.Alternativement, lorsque le paramètre $r$ tend vers 0, l'indicateur de mobilité sociale s'apparente au taux de mobilité ascendante. Un élève est en mobilité ascendante si son rang scolaire est supérieur à son rang social. Le taux de mobilité ascendante est donc caractérisé par la proportion d'élèves dont le rang scolaire est supérieur au rang social au sein d'une école. Comparé à la mobilité interdécile, le taux de mobilité ascendante donne un poids moins important à la mobilité sociale des élèves du bas de l'échelle. De façon général, le poids donné à la mobilité sociale des élèves du bas de l'échelle diminue à mesure que $r$ se rapproche de 0.
L'indicateur de mobilité utilisé dans ce mémoire est l'indicateur de mobilité interdécile (cas ou $r$ vaut 1). Dans la première partie, le taux de mobilité ascendante et différentes spécifications pour $r$ sont utilisées pour tester la robustesse du résultat.

## L'efficacité

Le but ici est d'obtenir une mesure de la performance des élèves. Pour ce faire, il pourrait être tentant de prendre simplement la moyenne du score des élèves d'une école aux tests. On ignorerait alors un élément important. En effet, les résultats d'un élève sont fortement liés à sa position sociale. Si on ne prend pas cela en compte, les écoles qui ont des élèves plus favorisés auront un avantage et seront considérées comme plus performantes. Des écoles avec des élèves favorisés passeraient le test d'efficacité sans que cela ne soit lié au « bon travail » accompli par l'école.
De la même façon, d'autres facteurs comme les compétences innées peuvent influencer les résultats obtenus. Idéalement, il faudrait prendre cela en compte également. En l'absence de contrôle pour les compétences innées, une école peut passer le test d'efficacité en sélectionnant les élèves plus performants. Les écoles qui font plus de sélection académique pourraient donc réussir le double test simplement parce qu'elles parviennent à sélectionner les élèves les plus brillants. Pour pouvoir contrôler ces facteurs, il faudrait avoir de l'information à leur sujet. Malheureusement, les données PISA avec lesquelles je travaille dans ce mémoire ne fournissent pas d'information à ce sujet. Pour contrôler de façon optimale pour des facteurs comme les compétences innées, il faudrait utiliser des données de panel. Ce que ne sont pas les données PISA.

Pour mesurer l'efficacité d'une école, les élèves sont classés selon leur rang scolaire et leur rang social. Dans le cadre de ce mémoire, il s'agira de percentiles. Le rang scolaire est d'abord régressé sur le rang social. La droite de régression ainsi estimée donne une valeur prédite ou valeur attendue pour le score d'un élève en fonction de son rang social. A partir de là, la mesure de performance d'un élève est la différence entre son résultat attendu et son résultat observé. L'idée est qu'un élève qui a un résultat observé supérieur à son résultat prédit par son origine sociale surperforme alors qu'un élève qui obtient un résultat inférieur à ce qui est attendu par rapport à son niveau social sous performe.

Formellement, l'indicateur peut être décrit par l'équation 3.2.

$$Efficacité = 1/n_s \sum_{i=1}^{n} (y_i - \hat{y}_i) \tag{3.2}$$

où :

$n_s$ = le nombre d'élèves dans l'école s
$y_i$ = la position scolaire de l'étudiant i
$\hat{y}_i$ = la position scolaire prédite de l'étudiant i

Pour mesurer l'efficacité au niveau d'une école, la moyenne des scores effectifs de l'ensemble des élèves de l'école est prise.

## Les écoles de la chance

Finalement, ces deux indicateurs (équité et efficacité) sont combinés pour identifier les écoles de la chance. Les écoles de la chance sont celles qui surperforment tant du point de vue de l'équité que de la mobilité sociale. Concrètement, il s'agit donc de regarder au sein d'un groupe de référence (le pays dans ce mémoire) les écoles qui performent au-delà de la médiane nationale du point de vue de ces deux indicateurs.
L'identification des écoles de la chance décrite dans cette section sera utilisée dans les deux sections suivantes.
Dans un premier temps, dans le cadre d'une comparaison des communautés belge (Communauté française et Communauté flamande). Hindriks et Godin (2020) avaient déjà comparé ces deux communautés en 2012. Je propose ici de poursuivre cette démarche en regardant l'évolution de la comparaison des communautés en utilisant les données PISA pour 2012, 2015 et 2018.
Ensuite, je propose d'étudier les caractéristiques communes des écoles de la chance en utilisant l'ensemble des pays de l'OCDE. Des méthodes de Machine learning sont utilisées à cet effet.

# 4. Evolution de la comparaison entre communautés entre 2012 et 2018

Hindriks et Godin (2020) ont comparé la proportion d'écoles de la chance dans chacune des communautés en utilisant les données de PISA 2012. Ils trouvent que la proportion d'écoles flamandes qui sont des écoles de la chance (40%) est nettement supérieure à la proportion d'écoles francophones qui sont des écoles de la chance (15%). Par ailleurs, 58% des écoles flamandes étaient plus efficace que la médiane nationale contre 36% seulement des écoles francophones. 59% des écoles flamandes étaient plus équitables que la moyenne nationale contre 35% des écoles de la communauté française. Dans le cadre de ce mémoire, je regarde comment ces proportions ont évoluées en 2015 puis 2018.

## 4.1 Données

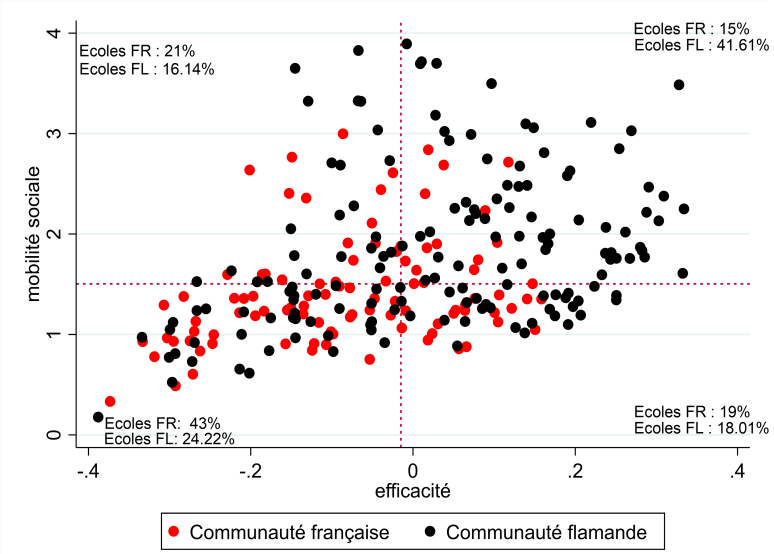
Les données que j'utilise dans cette section sont les données des tests PISA pour les vagues d'enquêtes 2012, 2015 et 2018. Je prends l'ensemble des écoles belge à l'exception des écoles de la Communauté germanophone et des écoles qui contiennent moins de 10 étudiants.
Chaque vague d'enquête PISA évalue les compétences des élèves en sciences, en mathématique et en lecture. Dans les données PISA, les scores au test sont fournis sous forme de valeurs plausibles (p valeurs). Les compétences dans le cadre des tests PISA sont évaluées via un certain nombre de questions dans chaque branche. Malheureusement, par contrainte de temps notamment, chaque élève ne répond qu'à une partie des questions. Ces questions lui sont assignées de façon aléatoire. Pour rendre compte de l'ensemble des compétences dans une branche pour chaque élève, l'OCDE utilise un modèle issu de la théorie de réponse aux items (TRI). Ce modèle permet d'établir la difficulté des différentes questions et de produire des « scores » ou valeurs plausibles pour un étudiant en prenant en compte des caractéristiques des étudiants (Jerrim *et al.* (2017)). En 2012, chaque étudiant avait 5 valeurs plausibles par branche. En 2015 et 2018, il y en a 10. Dans le cadre de ce travail, la moyenne des valeurs plausibles pour chaque élève est prise comme mesure du résultat des élèves.
Pour le niveau social, c'est l'indice de niveau socioéconomique de PISA qui est utilisé (ESCS). Cet indice créé par l'OCDE prend en compte trois composantes : l'occupation la plus élevée des parents, le niveau d'étude le plus élevé des parents et les possessions au domicile (OECD (2019a)).

## 4.2 Résultats

J'applique ici la méthode décrite dans la section précédente pour identifier les écoles de la chance en Belgique en distinguant la communauté dans laquelle elles se situent. Les élèves sont classés en percentiles au niveau national du point de vue

de leur score (moyenne des valeurs plausibles en mathématiques) et de leur origine sociale (indice ESCS). Les élèves sont classés au niveau national. Concrètement, cela signifie que l'ensemble des élèves belges sont classés de 1 à 100 sur base de leur indice socioéconomique et de leur score au test en mathématique. Concernant l'indicateur de mobilité sociale, la valeur attribuée au r est 1. Un poids plus important est donc donné à la mobilité ascendante des élèves du bas de la distribution et à l'intensité de cette mobilité. Les résultats pour d'autres spécifications sont également calculés pour tester la robustesse des résultats à d'autres spécifications.
Je commence par identifier les écoles de la chance pour l'année 2012 en tentant de reproduire les résultats de Hindriks et Godin (2020). La figure 4.1 montre la répartition des écoles de la communauté française et de la communauté flamande selon leur mobilité sociale et leur efficacité en 2012. C'est dans le cadran supérieur droit de ce graphique que se trouvent les écoles de la chance. Ces écoles sont en effet caractérisées par une mobilité et une efficacité supérieure à la médiane nationale.

FIGURE 4.1 – Comparaison des communautés PISA 2012

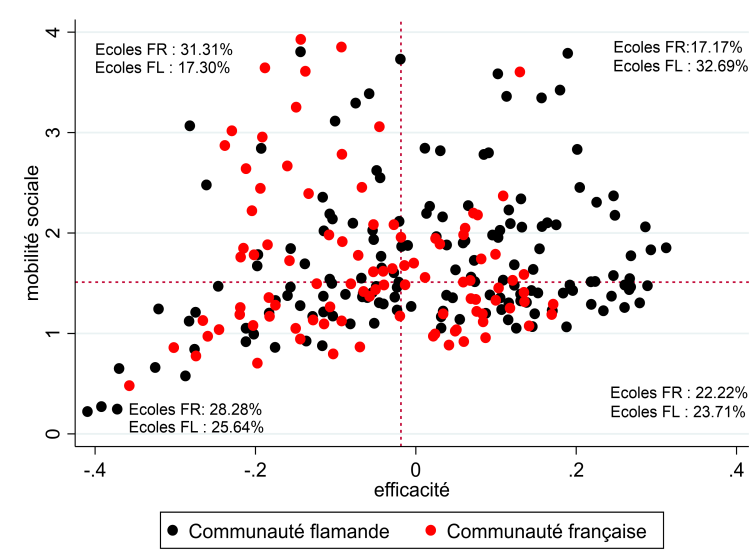



Ce graphique illustre le fait qu'une plus grande proportion des écoles flamandes sont des écoles de la chances comparées aux écoles francophones. La proportion d'écoles avec mobilité sociale supérieur à la médiane nationale est bien plus élevée en Flandre (57%) qu'en Communauté française (36%). Le même constat vaut pour l'efficacité (59% contre 34%).

J'ai ensuite réalisé la même analyse pour l'année 2015. L'écart en termes d'écoles de la chance semble se résorber légèrement. En effet, il y a maintenant 17% (15% en 2012) d'écoles francophones qui sont des écoles de la chance contre 32% (40% en 2012) d'écoles flamandes.
Par ailleurs, 48% des écoles de la Communauté française ont une mobilité sociale qui est supérieure à la médiane nationale contre 50% des écoles flamandes. Du point de vue de l'efficacité, 39% des écoles francophones sont au-dessus de la médiane nationale en termes d'efficacité contre 56% du côté flamand.
La figure 4.2 illustre la répartition des écoles francophones et flamandes du point de

vue de l'équité et de l'efficacité en 2015.

FIGURE 4.2 – Comparaison des communautés PISA 2015




Finalement, l'analyse est répétée pour l'année 2018. Il y a maintenant 24% d'écoles de la chance parmi les écoles francophones et 27% du côté flamand. L'écart a donc continué à se résorber entre 2015 et 2018 et est maintenant très faible. Du point de vue de l'efficacité, 39% des écoles francophones se trouvent au-dessus de la médiane nationale contre 56% des écoles flamandes. Cela ne constitue pas un changement majeur par rapport aux vagues d'enquête précédentes. En revanche, sur l'axe de la mobilité, 42% des écoles flamandes se trouvent désormais au-dessus de la médiane nationale contre 60% pour les écoles francophones. L'évolution est considérable.
La figure 4.3 illustre la répartition des écoles francophones et flamandes selon leur mobilité sociale et leur efficacité.

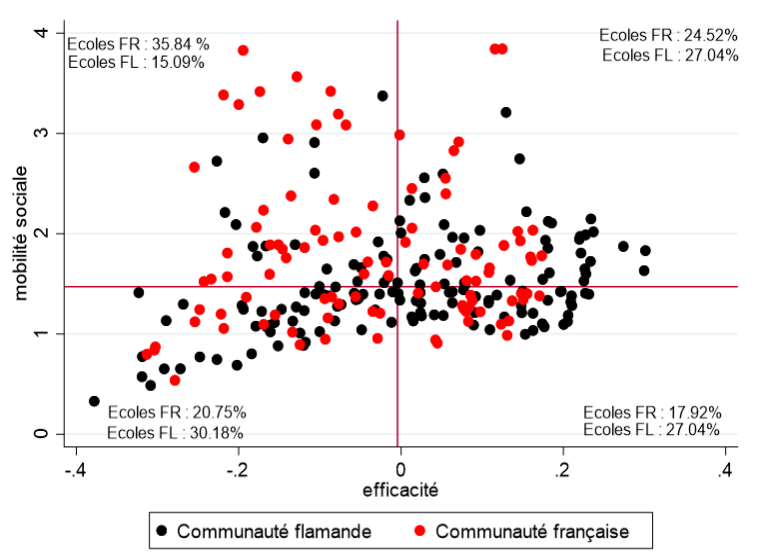
A mon sens, 3 résultats intéressants ressortent de l'analyse de l'évolution de la comparaison entre communautés.

- L'écart en termes de proportion d'écoles de la chance dans chaque communauté semble s'être considérablement réduit entre 2012 et 2018. Le rattrapage semblait avoir déjà commencé en 2015.

- La proportion d'écoles flamandes avec une efficacité supérieure à la médiane nationale était largement supérieure à cette même proportion en communauté française en 2012. Elle l'est toujours en 2018.

- La proportion d'écoles flamandes dont la mobilité était supérieure à la médiane nationale était largement supérieure à cette même proportion en Communauté française en 2012. Ces proportions se sont totalement inversées entre 2012 et 2015. Les écoles francophones se positionnent bien mieux sur l'axe de la mobilité en 2018.

FIGURE 4.3 – Comparaison des communautés PISA 2018




Pour résumer, la Communauté française semble avoir rattrapé la Communauté flamande en termes d'écoles de la chance. Ce changement semble surtout être dû à un changement sur l'axe de la mobilité sociale. Avant de discuter plus en profondeur ce résultat, je commence par vérifier sa robustesse à un indicateur de mobilité différent et à la prise en compte de différentes matières.

## 4.3 Robustesse

*Robustesse du résultat à un indicateur de mobilité différent*

N'étant pas neutre, on pourrait se soucier de la favorisation de l'une ou l'autre communauté par l'un ou l'autre indicateur. Par exemple, si les flamands se situent plus haut sur l'échelle sociale en moyenne, l'indicateur de mobilité interdécile pourrait leur être moins favorable qu'un indicateur accordant moins d'importance à l'origine sociale d'un élève.
Le taux de mobilité ascendante constitue une alternative à la mobilité interdécile. Pour rappel, le taux de mobilité ascendante mesure la proportion d'élèves au sein d'une école qui sont en mobilité ascendante. Cet indicateur est moins sensible à la mobilité des élèves du bas de l'échelle sociale ainsi qu'à l'intensité de cette mobilité. J'ai donc effectué la même analyse que dans le point précédent en utilisant le taux de mobilité ascendante comme indicateur de mobilité interdécile.

Le tableau 4.1 montre les résultats. La première colonne montre l'évolution de la proportion d'écoles de la chance dans chaque communauté. La 2ème colonne montre l'évolution de la proportion d'écoles avec une mobilité supérieure à la médiane nationale dans chaque communauté. La 3ème colonne montre l'évolution de la proportion d'écoles qui ont une efficacité supérieure à la médiane nationale dans chaque communauté.
Entre 2012 et 2018, la proportion d'écoles de la chance en Communauté française est

passé de 18 à 33%. La proportion d'écoles de la chance en Communauté flamande est passée de 36 à 33%. Le rattrapage est toujours présent. Par ailleurs, les proportions d'écoles avec mobilité et efficacité supérieure à la moyenne évoluent dans le même sens qu'avec la mobilité interdécile. Par exemple la proportion d'écoles avec mobilité sociale supérieure à la médiane nationale est supérieure en Communauté française en 2018 alors qu'elle était largement inférieure en 2012.

TABLE 4.1 – Robustesse au taux de mobilité ascendante

| | *Ecoles de la chance* | | *Mobilité supérieure* | | *Efficacité supérieure* | |
|---|---|---|---|---|---|---|
| | FL | FR | FL | FR | FL | FR |
| *2012* | 54 | 18 | 64 | 25 | 58 | 34 |
| *2015* | 46 | 20 | 52 | 37 | 55 | 35 |
| *2018* | 36 | 33 | 43 | 47 | 50 | 41 |

J'ai également vérifié la robustesse du résultat en faisant varier le paramètre $r$. Le rattrapage y est robuste également. Les résultats peuvent être trouvés dans le tableau A.1 en annexe.

*Robustesse du résultat à d'autres matières*

Jusqu'à présent, seuls les résultats en mathématiques avaient été considérés. Cependant, on pourrait imaginer qu'une des communautés soit plus performante dans une des branches. Une des communautés serait alors favorisées.
Par ailleurs, les tests PISA se focalisent sur une branche différente lors de chaque vague d'enquête. En 2012, le test PISA se focalisait sur les mathématiques. En 2015 il s'agissait des sciences et en 2018, de la lecture. La branche principale du test fait l'objet d'une attention particulière. La proportion d'étudiants effectivement testés est par exemple plus importante pour cette matière (Zieger *et al.* (2020)).

Je propose donc de consolider la variable de score en prenant en compte l'ensemble des matières. Concrètement, je prends la moyenne des valeurs plausibles pour l'ensemble des branches et plus uniquement en math.
Le tableau 4.2 présente les résultats. Le rattrapage est encore présent. Les mouvements au niveau de la mobilité et de l'efficacité vont dans le même sens que dans le cas où seules les mathématiques sont considérées.

TABLE 4.2 – Robustesse à d'autre branches

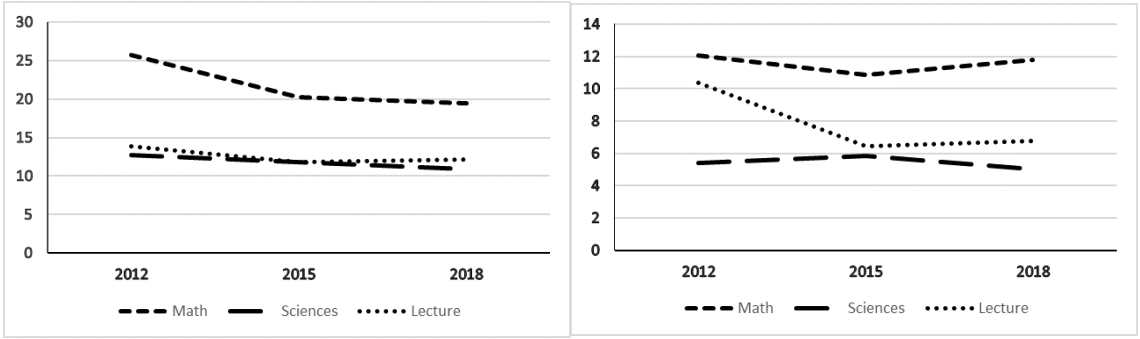
| | *Ecoles de la chance* | | *Mobilité supérieure* | | *Efficacité supérieure* | |
|---|---|---|---|---|---|---|
| | FR | FL | FR | FL | FR | FL |
| *2012* | 20 | 38 | 37 | 57 | 40 | 56 |
| *2015* | 18 | 29 | 52 | 49 | 41 | 56 |
| *2018* | 25 | 28 | 60 | 42 | 45 | 52 |

## 4.4 Discussion

Dans ce point ci, je discute les résultats obtenus. Pour rappel, le rattrapage des écoles flamandes par les écoles francophones semblait venir principalement d'un changement sur l'axe de la mobilité plutôt que sur l'axe de l'efficacité. Ceci en soit est un résultat assez surprenant. Ces dernières années, l'évolution des scores aux tests PISA a occupé l'espace public en Flandre, notamment après publication des résultats des enquêtes. Avec comme argument principal, la baisse généralisée des performances des élèves flamands sur ces 15 dernières années en mathématique notamment (La proportion d'élèves forts en mathématique a fortement baissé depuis 2003) [1] . C'est également un argument utilisé par le ministre de l'enseignement du gouvernement flamand, Ben Weyts (NVA) pour justifier une profonde réforme de l'enseignement flamand [2].

La figure 4.4 rend compte de l'évolution du pourcentage d'élèves forts dans les 3 branches entre 2012 et 2018 dans chaque communauté (Calculs propre selon la méthode de l'OCDE (OECD (2009))).
En Flandre, il y a effectivement une baisse du pourcentage d'élèves forts entre 2012 et 2015 pour les scores en mathématique. Pour le reste, le pourcentage d'élèves forts en sciences et en lecture est resté assez stable. Le pourcentage d'élèves forts en Flandre n'a donc pas baissé de façon impressionnante pendant la période étudiée (La baisse à laquelle il est fait référence dans les articles concerne une période plus longue).
En Communauté française, le pourcentage d'élèves forts est resté assez stable entre 2012 et 2018 hormis en lecture ou il a un peu baissé. Ces observations sont compatibles avec les résultats observés. A savoir, un changement négligeable sur l'axe de l'efficacité.

FIGURE 4.4 – Evolution du pourcentage d'élèves forts par branche




(a) Communauté flamande (b) Communauté française

*Dégradation de la position relative des élèves flamands et amélioration de la position relative des élèves francophones*

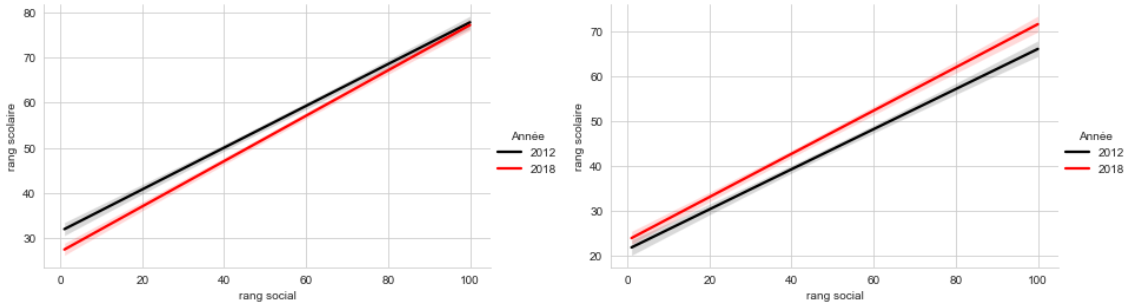
Un changement du point de vue de la mobilité sociale pourrait être expliqué par un affaiblissement de la position des élèves du haut ou du bas de la distribution en

1. Voir DeMorgen 3/12/2019, Knack 3/06/2020 par exemple
2. Knack 3/12/2020

Flandre et/ou une amélioration de la position relative des élèves du haut ou du bas de la distribution en Communauté française. Pour tenter d'identifier ces mouvements, je me suis intéressé à l'évolution de la régression rang à rang nationale au sein de chaque communauté entre 2012 et 2018.

Les figures 4.5a et 4.5b présentent l'évolution des régressions rang national à rang national pour les élèves de chaque communauté entre 2012 et 2018.
Il y a bien une dégradation de la position relative des élèves du bas de l'échelle sociale en Flandre et une amélioration de la position relative des élèves du haut de l'échelle sociale en Communauté française. Ces évolutions pourraient expliquer le changement qu'on observe sur l'axe de la mobilité sociale entre 2012 et 2018. En effet, les élèves flamands du bas de la distribution se retrouveraient plus bas pour un même niveau socioéconomique. Cela diminuerait leur mobilité sociale. Les élèves francophones du haut de la distribution se trouvant plus haut à même niveau social se retrouveraient alors avec une mobilité sociale améliorée.

FIGURE 4.5 – Evolution de la régression rang à rang national dans chaque communauté



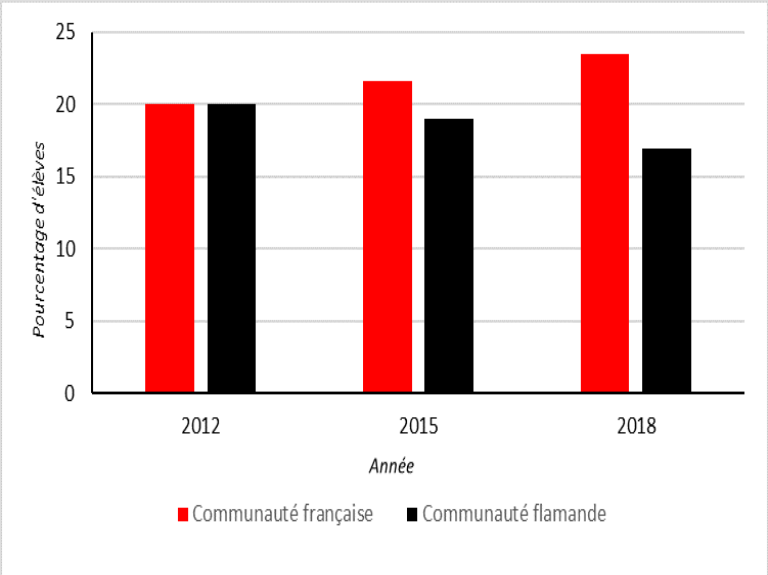

(a) Communauté flamande (b) Communauté française

*Un effet « composition »*

Le changement de mobilité pourrait aussi s'expliquer par un effet « composition » qui pourrait apparaître si la distribution du niveau social change au cours des différentes vagues d'enquêtes. Imaginons par exemple que la proportion d'élèves dans le bas du classement social national soit identique dans les deux communautés en 2012. Mais qu'en 2018, on retrouve une proportion d'élèves francophones dans le bas de la distribution qui est supérieure à celle des élèves flamands. Et dans le haut de la distribution, une proportion d'élèves flamands supérieure.
La mobilité sociale étant moins évidente pour les élèves du haut de la distribution, il serait naturel d'observer une baisse de la mobilité en Flandre et une augmentation de la mobilité en Communauté française. Pour identifier un éventuel effet composition, j'ai regardé la composition des premiers et derniers quintiles pour 2012, 2015 et 2018.

La figure 4.6 montre l'évolution de la proportion d'écoles au sein de chaque communauté qui se retrouvent dans le premier décile socioéconomique.

FIGURE 4.6 – Evolution de la proportion d'élèves de chaque communauté dans le premier quintil social




En 2012, 20% des élèves francophones et 20% des élèves flamandes se trouvaient dans le premier quintile. 20% des élèves flamandes et 18% des élèves francophones dans le dernier quintile. En 2015, la proportion d'élèves flamands dans le premier quintile était déjà plus faible (et plus élevé en Communauté française) et inversement dans le quintile supérieur. En 2018, la proportion d'élèves flamands dans le premier quintile n'était plus que de 17% et celle des élèves francophones de 23%. Le phénomène inverse est visible pour le quintile supérieur. En effet, 23% des élèves flamands s'y trouvent contre seulement 14% des élèves francophones. On ne peut donc pas éliminer la possibilité d'un effet cohorte pour expliquer les résultats obtenus.

*Une augmentation de la proportion des élèves ne parlant pas néerlandais dans le bas de la distribution ?*

Pour expliquer la baisse des résultats scolaires des élèves flamands sur ces 15 dernières années, il a notamment été question d'une augmentation de la proportion d'élèves ne parlant pas néerlandais à la maison. En effet, la NVA (Nieuw-Vlaamse Alliantie) notamment, pointe du doigt l'augmentation du nombre d'élèves ne parlant pas néerlandais à la maison comme responsable de la baisse des résultats scolaires aux tests PISA sur ces 15 dernières années [3].
Imaginons que les élèves ne parlant pas néerlandais performent moins bien que les élèves parlants néerlandais. Dans ce cas, une augmentation de la proportion d'élèves ne parlant pas néerlandais à la maison (particulièrement dans le bas de la distribution) pourrait expliquer la baisse sur l'axe de la mobilité sociale. J'ai donc vérifié s'il s'agit d'une explication possible pour le rattrapage observé.

J'ai commencé par m'intéresser à l'évolution de la proportion d'élèves ne parlant pas la langue de la communauté en communauté flamande et en communauté française entre 2012 et 2018. Les données utilisées sont les données PISA. Je regarde

3. Agirdag (2020)

donc comment la composition de l'échantillon PISA à évoluer entre ces 2 années au sein de chaque communauté.
Le tableau 4.3 montre la proportion d'étudiants ne parlant pas la langue à la maison au sein de chaque communauté pour 2012, 2015 et 2018. La colonne de gauche (total) montre l'évolution de cette proportion dans l'ensemble de la communauté. La colonne de droite (Q1) montre l'évolution de cette proportion parmi les élèves du quartil socioéconomique inférieur. Il y a bien une augmentation de la proportion d'élèves flamands ne parlant pas néerlandais à la maison entre 2015 et 2018. Pour cette même période, la proportion est stable en Communauté française. Entre 2012 et 2015, en revanche, le phénomène inverse est observé. Sur cette période, il y a une augmentation d'élèves ne parlant pas français en communauté française alors qu'il n'y a pas de changement en Flandre. Pourtant le rattrapage était déjà amorcé entre 2012 et 2015.

TABLE 4.3 – Evolution de la proportion d'étudiants ne parlant pas langue dans chacune des communautés dans les données PISA

| | Communauté flamande | | Communauté française | |
|---|---|---|---|---|
| | Total | Q1 | Total | Q1 |
| 2012 | 14,14 | 22,6 | 8,67 | 13,77 |
| 2015 | 14,13 | 20,95 | 15,21 | 25,27 |
| 2018 | 16,38 | 28,23 | 16,01 | 25,27 |

Par ailleurs, j'ai vérifié si le rattrapage disparaissait lorsque les élèves ne parlant pas la langue sont retirés de l'échantillon. Le tableau 4.3 montre l'évolution de la proportion d'écoles de la chance dans chaque communauté lorsque les élèves ne parlant pas la langue à la maison sont retirés de l'échantillon. Le rattrapage semble toujours bien présent.

TABLE 4.4 – Evolution de la proportion d'écoles de la chance au sein de chaque communauté

| | *Ecoles de la chance* | |
|---|---|---|
| | FR | FL |
| *2012* | 14 | 39,13 |
| *2015* | 17,17 | 30,76 |
| *2018* | 23,58 | 29 |

## 4.5 Conclusion

Malgré l'amélioration de la position sociale relative des élèves flamands entre 2012 et 2018, l'efficacité des écoles flamandes change peu. En revanche, leur mobilité sociale se détériore. Cette détérioration est probablement liée en partie à la

détérioration de la position sociale relative des élèves flamands. Ceci tient à l'indicateur de mobilité sociale utilisé qui favorise la mobilité des élèves du bas de la distribution. En revanche l'augmentation de la proportion d'élèves ne parlant pas flamand à l'école ne semble pas expliquer le rattrapage.
Du côté francophone, la détérioration de la position sociale n'a pas diminué l'efficacité des écoles. Par ailleurs, la mobilité sociale relative des écoles francophones s'améliore. Cela peut en partie s'expliquer la détérioration de la position sociale relative des élèves francophones.

# 5. Caractéristiques des écoles de la chance au niveau international

L'objectif de cette section est d'identifier et hiérarchiser les principales caractéristiques communes des écoles de la chance. Pour cela, j'utilise des algorithmes de Machine Learning et les données des questionnaires étudiants et écoles de PISA 2018 pour l'ensemble des pays de l'OCDE. Je m'inspire de plusieurs articles qui ont utilisé des méthodes de Machine Learning avec les données PISA ou PIRLS (Programme international de recherche en lecture scolaire).
Plus précisément, je commence par appliquer trois algorithmes de classification (pour classer les écoles en écoles de la chance et écoles qui ne sont pas des écoles de la chance) : un Support Vector Machine (SVM), une régression logistique et un arbre de décision. J'applique ensuite un algorithme d'élimination récursive des caractéristiques au SVM et à la régression logistique. Cela me permet de classer les caractéristiques selon leur importance dans un premier temps puis de sélectionner un set de variables optimales pour la classification des écoles. Le set de variables optimal est celui permet de classifier les écoles avec le plus de précision.

Cette deuxième partie de mon mémoire se présente de la façon suivante :
Dans la première section, je présente une rapide introduction au Machine Learning ainsi qu'aux différents algorithmes utilisés dans le cadre de ce mémoire.
Dans la deuxième section, je passe en revue les différents articles qui ont utilisé les techniques de Machine Learning sur les données PISA. Ce sont ces articles qui établissent les fondements de mon approche. Je clôture la section en détaillant la méthodologie qui sera utilisée dans ce mémoire.
Dans la troisième section, j'explique la façon dont la base de données a été élaborée avec une définition précise des données et indices synthétiques utilisés.
La quatrième section détaille le prétraitement des données avec une attention particulière pour le traitement des données incomplètes.
La cinquième section présente les résultats de l'application des différents algorithmes de Machine Learning à ma base de données.
Dans la sixième section, les résultats sont interprétés à la lumière de la littérature existante.
Finalement, la septième section conclu la deuxième partie du travail en passant en revue les limites de l'analyse et en résumant les principaux résultats.

## 5.1 Le Machine Learning et les différents algorithmes utilisés dans ce mémoire

Le but de cette section est de donner une courte explication du Machine Learning ainsi qu'une brève description des différents algorithmes qui sont utilisés dans le cadre de ce mémoire.
Athey (2018) définit le Machine Learning comme étant « une branche qui développe

des algorithmes conçus pour être appliqués à des bases de données, avec comme principaux domaines d'intérêts, la prédiction, la classification et le groupement de tâches ou clustering » (Athey (2018)). Le Machine Learning est utilisé dans des domaines comme l'informatique, la médecine, les statistiques ou l'ingénierie par exemple, mais s'étend de plus en plus à des domaines plus larges comme les sciences sociales et notamment l'économie [1] (Athey (2018)).

Le Machine Learning est une technique de traitement des données qui vise à découvrir des patterns et à effectuer des prédictions à partir d'un grand volume de données (big data) en se basant sur des statistiques, sur du forage de données, sur la reconnaissance de patterns et sur les analyses prédictives.
Les outils analytiques traditionnels ne sont souvent pas suffisamment performants pour exploiter pleinement la diversité des données du Big Data. Le volume de données est parfois trop large pour des analyses exhaustives, mais surtout, les corrélations et relations entre ces données sont trop importantes pour que les analystes puissent tester toutes les hypothèses afin de dégager une valeur à partir de ces données. Plus les données injectées à un système Machine Learning sont nombreuses, plus ce système peut apprendre et appliquer les résultats à des insights de qualité supérieure. Le Machine Learning permet ainsi de découvrir les patterns enfouis dans les données avec plus d'efficacité que l'intelligence humaine. Il ne prétend cependant pas établir de causalité mais des associations et patterns.

Les tâches effectuées par le Machine Learning se divisent essentiellement en deux branches. L'apprentissage supervisé et l'apprentissage non supervisé.
L'apprentissage non supervisé a pour but de rassembler des ensembles d'observations qui partagent des caractéristiques similaires.
Les algorithmes d'apprentissage supervisé ont pour but la prédiction (variable à prédire continue) et la classification (variable à prédire discrète).
Les algorithmes utilisés dans ce travail sont essentiellement des algorithmes d'apprentissage supervisé. Concrètement, il s'agit d'utiliser un algorithme qui, sur base des données (observations) qui lui sont fournies, va apprendre à prédire la valeur ou la classe de nouvelles observations qu'il n'a jamais vues.
En apprentissage supervisé, le jeu de données initial est toujours séparé en un jeu d'entraînement et un jeu de test. Le jeu d'entraînement sert à entraîner le modèle. Le jeu de test sert à tester le modèle sur des données qu'il n'a jamais vues. Ceci est primordial et constitue la base du Machine Learning. En effet, un algorithme peut être très performant sur des données sur lesquelles il a été entraîné puisqu'il s'y est bien adapté, mais très mauvais sur des données qu'il n'a jamais vues. On dit d'un tel algorithme qu'il est en surapprentissage [2]. Le but est de maximiser la capacité de prédiction de l'algorithme sur le jeu de test.
Une autre façon d'évaluer un modèle de Machine Learning consiste à faire une validation croisée. Le principe est de diviser le jeu de données en plusieurs parties. Concrètement, l'algorithme est entraîné autant de fois qu'il y a de parties. A chaque entraînement, une partie différente sert de jeu de test et le modèle est entraîné sur les parties restantes qui servent de jeu d'entraînement. Le score de validation croisée

1. Athey (2018)
2. En cas de sur-apprentissage, l'analyse de prédiction ou de classification peut ne pas correspondre à des données supplémentaires ou ne pas prévoir de manière fiable les observations futures.

est donné par la moyenne des scores sur chaque partie lorsqu'elle servait de jeu de test. La validation croisée est souvent utilisée pour comparer différents modèles et choisir le meilleur. Il est parfois aussi utilisé pour évaluer le modèle final. De façon générale, Il est courant de présenter les résultats d'une validation croisée sur le set d'entraînement ainsi que les résultats sur le jeu de test. C'est l'approche que j'utilise ici.

Dans la suite de cette section, je vais présenter les différents algorithmes utilisés dans le cadre de ce mémoire. Je vais commencer par les algorithmes de classification : le Support Vector Machine (SVM), la régression logistique et l'arbre de décision. Ces algorithmes permettront de classifier les écoles sur base des critères définis dans la première partie. Ces différents algorithmes ont des paramètres. Ces paramètres doivent être réglés. Pour trouver les meilleurs paramètres pour chaque modèle, un algorithme de recherche sur grille avec validation croisée est utilisé. Il est décrit également dans la section suivante. Ensuite, j'expliquerai les algorithmes de sélection des variables : L'élimination récursive de caractéristiques/variables (RFE) et l'élimination récursive de caractéristiques/variables avec validation croisée (RFECV). Ces algorithmes seront utilisés pour trouver les variables les plus importantes pour la classification.

## Les algorithmes de classification

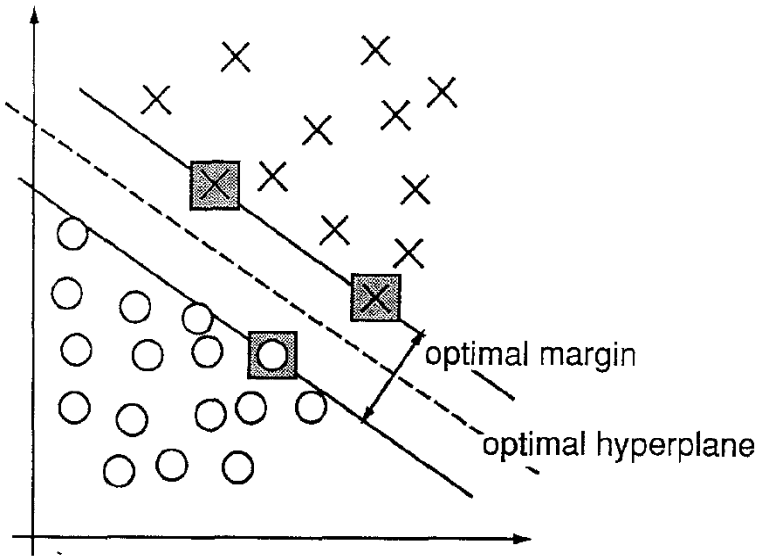
*Support Vector Machine (SVM)*

Le Support Vector Machine est un algorithme de classification qui a été introduit par Cortes et Vapnik en 1995 (Cortes et Vapnik (1995)). Le but de cet algorithme est de classifier les observations (écoles dans ce mémoire) dans la catégorie qui leur correspond. Pour cela, l'idée est de chercher une frontière de décision dans l'espace qui permet de séparer au mieux les deux catégories en utilisant les données observées. C'est cette frontière qui déterminera la classe d'une nouvelle observation (école) dont on observe certaines caractéristiques et dont on veut connaître la catégorie.

Dans la figure 5.1, les axes représentent des caractéristiques observables (le climat disciplinaire et la taille de l'école par exemple dans le cadre de ce mémoire). On cherche à trouver une frontière optimale pour séparer les ronds et les croix (écoles de la chance et écoles qui ne sont pas des écoles de la chance) dans l'espace.
Pour déterminer la frontière de décision, on s'intéresse aux observations de classes différentes qui sont les plus proches les unes des autres dans l'espace. Dans la figure 1, il s'agit des croix et ronds qui sont entourés d'un carré plein. Ces observations sont appelées les vecteurs de support. L'idée est de maximiser la distance (marge) entre ces vecteurs de supports et la frontière de décision.
Pour comprendre cela, imaginons qu'on place la frontière de décision près des ronds. Dans ce cas, une nouvelle observation qui se trouve plus proche des ronds que des croix mais au-dessus de la frontière sera classifiée comme croix. Pourtant, elle présente des caractéristiques plus proches des ronds que des croix. La meilleure frontière est donc celle qui est la plus éloignée des vecteurs de support. Elle est représentée par la droite en pointillés dans la figure 1. Une nouvelle observation sera classifiée comme croix si elle est au-dessus et comme rond si elle est en dessous de la frontière.

La frontière de décision est caractérisée par un vecteur de coefficients (un coefficient par variable). Ce vecteur de coefficients est orthogonal à la frontière de décision. La magnitude du coefficient d'une variable donne de l'information sur l'utilité de cette variable pour la classification. Au plus le coefficient associé à une variable est grand en valeur absolue, au plus la variable a de l'importance pour la classification (Mladenić *et al.* (2004), Chang et Lin (2008)).
Dans la figure 5.1, les lignes parallèles en traits continus sur lesquelles se situent les vecteurs de support donnent de l'information quant à la position des autres points par rapport à la marge (distance entre les vecteurs de support et la frontière).

FIGURE 5.1 – Résolution d'un problème de classification par le SVM




Source: Cortes et Vapnik (1995)

Plusieurs remarques s'imposent :

- Souvent, il n'est pas possible (ou raisonnable) de séparer parfaitement les observations car il existe pour chacune des catégories des valeurs « aberrantes », qui présentent des caractéristiques très similaires à l'autre catégorie. Il faut donc autoriser des erreurs de classifications. On parle alors de marge souple. L'intensité avec laquelle ces erreurs de classification sont autorisées est donnée par un paramètre C. Lorsque l'algorithme est utilisé, le paramètre C doit être réglé. Au plus il est élevé, au plus les erreurs de classifications sont autorisées. Autoriser très peu de mauvaises classifications pourrait résulter en de bonnes prédictions sur le jeu d'entraînement mais avoir de faibles performances sur le jeu de test. On serait alors en surapprentissage.

- Par ailleurs, l'espace considéré ici est en 2 dimensions. Evidemment, la plupart des applications du SVM se font dans un espace avec un ensemble bien plus large de variables. Concrètement au lieu d'être séparées par une droite, les observations seront séparées par un hyperplan de dimension N (nombre de caractéristiques) -1.

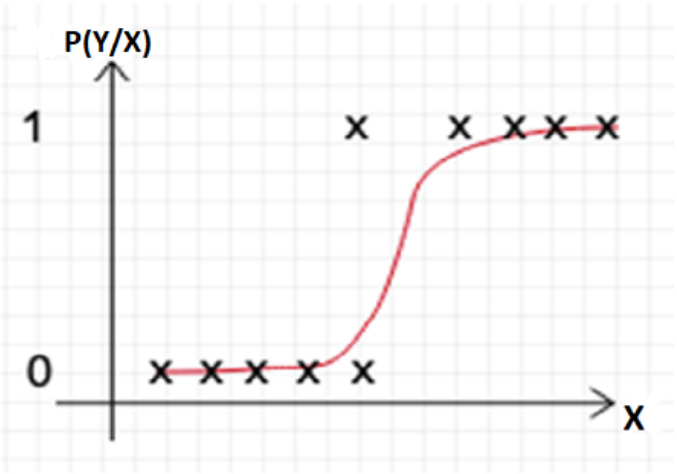
- Une propriété intéressante des SVM est qu'ils peuvent être utilisés pour résoudre des problèmes non linéaires. Cependant cela rendrait l'interprétation des coefficients et l'utilisation des algorithmes de sélection des variables utilisés dans ce mémoire difficile. De plus, Gorostiaga et Rojo-Álvarez (2016) trouvent des résultats semblables pour les SVM linéaires et non linéaires avec les données PISA 2012. Je ne vais donc pas en faire usage dans le cadre de ce mémoire.

*Régression logistique*

La régression logistique est aussi bien utilisée pour des applications statistiques classiques qu'en Machine Learning. La particularité de la régression logistique par rapport à la régression linéaire est qu'elle est utilisée dans le cas d'une variable dépendante discrète (souvent binaire).

Au lieu d'ajuster les données à l'aide d'une droite comme dans la régression linéaire, les données sont ajustées par une fonction logistique (sigmoïde). Cette fonction peut être visualisée dans la figure 5.2. La méthode utilisée pour ajuster cette courbe aux données est le maximum de vraisemblance. Concrètement, il s'agit de trouver la courbe qui maximise la vraisemblance. La courbe pour laquelle le maximum de vraisemblance est le plus élevé est choisie. La courbe (fonction logistique) décrit pour chaque observation une probabilité pour Y en fonction de la valeur de X. Pour ce mémoire par exemple, la courbe décrit pour chaque valeur de X (le climat disciplinaire par exemple) la probabilité qu'il s'agisse d'une école de la chance. Dans la figure 5.2, la probabilité que Y valle 1 est très élevée pour les observations avec une grande valeur pour X et faible pour les observations ayant une faible valeur pour X. Pour utiliser la régression logistique comme algorithme de prédiction, un seuil de probabilité est déterminé (0.5 par exemple). Au-dessus de ce seuil, une observation est considérée comme appartenant à la catégorie 1 et en dessous de ce seuil, à la catégorie 0. Les fonctions logistiques permettent également d'obtenir des coefficients pour les différentes caractéristiques.

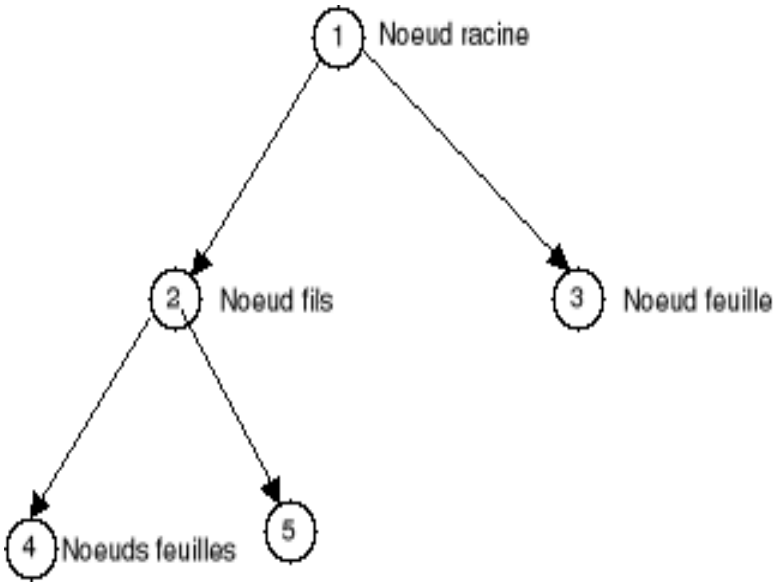
FIGURE 5.2 – Résolution d'un problème de classification par la régression logistique



*L'arbre de décision*

Un arbre de décision est composé d'un ensemble de nœuds qui sont disposés selon un ordre hiérarchique comme sur la figure 5.3. Le premier nœud est appelé nœud racine et les nœuds finaux sont les nœuds feuilles. Chaque nœud, peut-être vu comme une question. Selon la réponse à cette question les observations sont séparées. Le but est que la séparation permette de classer les observations dans leur bonne catégorie (dans le cadre de ce mémoire, en écoles de la chance et en écoles qui ne sont pas des écoles de la chance). Les questions doivent donc discriminer au maximum les deux catégories. Cela doit permettre un maximum d'homogénéité dans chaque nœud. Autrement dit, il s'agit d'avoir le moins d'impureté (hétérogénéité) possible. Un nœud n'a pas d'impureté si toutes les observations qu'il contient font partie de la même catégorie.
Le choix de la question posée à chaque nœud est déterminé par la quantité d'impureté qu'elle génère. La question qui sépare le mieux les données (l'impureté la plus basse) sera située à la racine de l'arbre (nœud racine). Depuis la racine, deux nouveaux nœuds sont créés (selon la réponse, un nœud si vrai et un nœud si faux, dans la figure 5.3, il s'agit des nœuds 2 et 3). Pour chaque nouveau nœud, la question éliminant le plus d'impureté est choisie. Ce processus est répété jusqu'à ce que l'impureté ne puisse plus être réduite. On atteint alors le nœud final (nœud feuille). Précisons que plusieurs méthodes peuvent être employées pour mesurer l'impureté. Il s'agit d'un paramètre du modèle qu'il y a lieu de régler.
Les arbres de décisions sont utilisés pour la classification. Ils offrent aussi un moyen assez facile et intuitif de représenter et visualiser l'importance des variables et leurs interactions.

FIGURE 5.3 – Résolution d'un problème de classification par un arbre de décision



## Algorithme de sélection des meilleurs modèles

*Recherche sur grille avec validation croisée*

Les algorithmes de classification ont des paramètres qu'il faut régler. Le C dans le cadre d'un SVM ou la mesure de l'impureté dans le cadre de l'arbre de décision en sont des exemples. Il existe un grand nombre de valeurs possibles pour les différents paramètres. Il faut donc trouver les valeurs optimales pour ces paramètres. Il est assez difficile de faire des hypothèses quant à la meilleure valeur à donner à un paramètre. Il s'agit donc d'en tester plusieurs et plus précisément une combinaison de plusieurs de ces paramètres. C'est précisément le but de la recherche sur grille avec validation croisée. L'algorithme est entraîné avec différentes combinaisons de paramètres. Pour chaque combinaison, un score de validation croisée est calculé et la combinaison de paramètres qui donne le meilleur résultat (plus haute précision dans la classification) est choisie.

## Algorithmes de sélection des variables

*Elimination récursive des caractéristiques/variables(RFE)*

Cet algorithme a été présenté par Guyon *et al.* (2002) dans le cadre de la sélection de gènes pour la prévision de cancers. Il propose une méthode pour sélectionner les caractéristiques les plus importantes pour la classification sur base de modèles de Machine Learning. Il peut également être appliqué pour ranger les caractéristiques selon leur importance.
L'idée est de choisir un critère pour déterminer l'importance des variables pour les algorithmes de classification. Un critère de sélection courant est le carré des coefficients. Au plus le carré du coefficient est élevé, au plus la variable a de l'importance. Ce critère requiert que les variables aient des ordres de grandeurs comparables (Zhu et Hastie (2004)). Il est important de normaliser les données avant d'appliquer ces algorithmes (Guyon et Elisseeff (2003)). La technique de normalisation utilisée ici est la même que dans la plupart des autres applications d'élimination récursive de caractéristiques. Elle est décrite dans la section 4.
La procédure de sélection fonctionne de la façon suivante : Un premier algorithme (SVM ou régression Logistique par exemple) est entraîné sur l'ensemble des variables. Sur base du critère de sélection (carré des coefficients ici), la variable la moins importante (carré du coefficient le plus faible) est éliminée. L'algorithme est alors réentraîné sur le nouveau jeu amputé de la variable la moins importante. Cette procédure itérative est répétée jusqu'à ce qu'il ne reste plus de variables à éliminer ou qu'un certain seuil de variables qu'on ne veut pas dépasser a été atteint. Les variables peuvent être classées selon leur ordre d'élimination.

Une version légèrement différente de cet algorithme est l'élimination récursive de caractéristiques avec validation croisée (RFECV). Cet algorithme fonctionne selon le même principe que le RFE. Mais il est couplé à une validation croisée. Lors de chaque élimination d'une caractéristique, un score de validation croisée est calculé pour le modèle. Finalement, l'algorithme retourne un ensemble de caractéristiques optimal (celui qui obtient le score de validation croisée (plus haute précision dans

la classification) le plus élevé).

## 5.2 Littérature et méthodologie

Dans cette section, je passe en revue la littérature sur l'application de techniques de Machine Learning dans le domaine de l'éducation. Le but étant de positionner mon analyse par rapport à la littérature existante.

### Littérature

Pour commencer, Chen *et al.* (2019) utilisent un algorithme de Support Vector Machine (SVM) couplé à un algorithme d'élimination récursive de caractéristiques (RFE). Leur but est d'identifier les facteurs contextuels liés à l'école, l'étudiant et l'environnement familial de l'étudiant qui permettent de prédire le niveau de l'étudiant. En d'autres mots, ils cherchent à classer les étudiants selon leur niveau en sciences (faible ou fort) et à identifier les variables contextuelles les plus importantes pour cette classification. Leur analyse est effectuée en utilisant la vague d'enquête de PISA 2015.
Après pré-traitement des données, un SVM est utilisé pour classer les élèves selon leur niveau (faible ou fort). A partir de ce modèle un autre algorithme, le RFE, est utilisé pour classer les variables selon leur importance pour la classification. Une fois le classement établi, les auteurs se basent sur leurs connaissances pour construire différents sets de caractéristiques (15, 20 et 30 caractéristiques) dont ils testent le pouvoir de classification commun.

Une approche similaire est utilisée par Xiao et Hu (2019) sur les données du test PIRLS (Programme international de recherche en lecture scolaire) 2016 et par Dong et Hu (2019) sur les données PISA 2015. Leur but est aussi de classer les élèves selon leur niveau (faibles et forts). Comme dans Chen *et al.* (2019), les auteurs utilisent un SVM RFE pour classer les caractéristiques selon leur importance pour la classification. Leur approche diffère concernant la recherche du set optimal de caractéristiques. Au lieu de se référer à leur connaissance du domaine, les auteurs utilisent une version alternative du RFE pour trouver le set optimal de variables : l'élimination récursive de caractéristiques avec validation croisée (RFECV). Ici, les auteurs comparent plusieurs sets selon le score qu'ils obtiennent et leur taille en prenant soin de considérer le set optimal.

Gorostiaga et Rojo-Álvarez (2016) utilisent une approche différente pour classifier les élèves espagnols selon leur niveau (forts ou faibles) en math avec les données PISA 2015. Dans cet article, les performances de différents algorithmes sont comparées (SVM et régression logistique notamment). Ils utilisent d'abord les différents modèles sur l'ensemble du jeu de données pour évaluer les performances de classification de chacun des algorithmes en utilisant l'exactitude (accuracy) et le rappel (recall). Ils développent ensuite une nouvelle méthode propre pour sélectionner les meilleures caractéristiques et fonder un set de variables optimal.

Finalement, un autre algorithme de Machine Learning utilisé sur les données PISA et dans le domaine de l'éducation est l'arbre de décision. Il est notamment

utilisé par Alivernini (2013) pour explorer les caractéristiques différenciant les élèves forts et les élèves faibles en lecture avec les données PIRLS 2006 sur vingt pays.
Alivernini *et al.* (2016) utilisent un arbre de décision pour séparer les élèves résilients des autres élèves. L'arbre de décision leur permet d'identifier deux groupes d'élèves (caractéristiques socioéconomiques ou culturelles similaires) parmi lesquels il y a très peu d'élèves forts. Ils s'intéressent ensuite aux élèves forts dans ces groupes ou la majorité des élèves sont faibles. Ils recherchent les caractéristiques communes de ces élèves au moyen d'une régression logistique. D'autres applications des arbres de décision dans la littérature sur l'éducation et le Machine Learning peuvent être trouvées dans Liu et Ruiz (2008), Sanzana *et al.* (2015) ou Masci *et al.* (2018).

### Méthodologie utilisée dans ce mémoire

Dans mon mémoire, j'utilise l'approche suivante :
Les résultats de classification pour séparer les écoles de la chance des autres écoles et de sélection des variables contextuelles sur base de trois algorithmes de Machine Learning sont comparés : un arbre de décision, un SVM (noyau linéaire) et une régression logistique.
Dans un premier temps, les données sont prétraitées. Le prétraitement des données consiste à gérer les valeurs manquantes et standardiser les données. J'envisage différentes stratégies de prétraitement et compare leurs performances pour les trois modèles (dont les paramètres sont optimisés). A l'issue de ce processus, la meilleure stratégie est conservée pour l'étape suivante qui consiste à classer et sélectionner les meilleures variables.
Pour l'arbre de décision, visualiser l'arbre suffit puisque les caractéristiques offrant les meilleures prédictions sont automatiquement sélectionnées et hiérarchisées.
Pour le SVM et la régression logistique, un RFE est utilisé pour classer les variables selon leur importance comme dans Chen *et al.* (2019), Xiao et Hu (2019) et Dong et Hu (2019).
Ensuite, un set optimal de variables est recherché. Pour cela, j'utilise un RFECV pour identifier le set optimal et visualiser le score du modèle en fonction du nombre de variables. L'approche est similaire à celle qui est utilisée par Xiao et Hu (2019) et Dong et Hu (2019).
Les deux sections suivantes traitent respectivement des données utilisées et du prétraitement de ces données.
Les trois dernières sections présentent les résultats, les discutent et la dernière section conclut en rappelant les forces et les faiblesses de l'analyse.

## 5.3 Construction de la base de données

Je pars de la base de données de l'OCDE construite à partir du questionnaire adressé aux directeurs d'écoles de PISA 2018 (base de données établissements). J'ai choisi une seule vague d'enquête (2018) et non plusieurs vagues d'enquêtes car les caractéristiques présentes dans la base de données varient d'une vague d'enquête à l'autre. Par ailleurs, certaines questions ont un lien direct avec la branche principale de la vague d'enquête (lecture pour 2018).

Garder toutes les variables de cette base de données telles quelles ne me semblait pas approprié. Il y a beaucoup de variables qui sont semblables ou qui ont peu de sens lorsqu'elles sont prises individuellement. Il y a donc lieu de faire un certain tri. D'ailleurs dans les rapports de l'OCDE, les variables sont souvent rassemblées sous forme d'indices. Certains de ces indices se trouvent dans la base de données établissement de PISA 2018. Ce sont donc les premières variables que j'inclus dans la base de données.
D'autres indices ne se retrouvent pas dans la base de données 2018 mais étaient présents en 2012 ou 2015. Il est possible de retrouver ces indices et la façon dont ils sont créés dans le rapport technique [3] de l'OCDE pour l'année en question. J'ai recréé, de la façon la plus fidèle possible et dans la mesure ou les questions étaient disponibles, ces indices pour l'année 2018.
Ensuite, il y a des caractéristiques directement disponibles dans la base de données qui sont issues des réponses des directeurs et qui me semblaient avoir un sens même en étant prises individuellement. Je les ai donc simplement rajoutées telles quelles.
Il y a certaines variables que j'ai moi-même rassemblées. J'ai fait cela en utilisant les méthodes classiques que l'OCDE utilise pour créer les indices.
Finalement, j'ai rajouté des variables issues du questionnaire étudiant de PISA 2018. Cela permet de prendre directement en compte l'avis des élèves. Les variables sélectionnées dans la base de données étudiants sont des indices créés par l'OCDE et qui sont liés à l'établissement scolaire (le climat disciplinaire qui règne en classe tel que perçu par l'étudiant par exemple). Pour associer ces indices aux écoles, je prends la moyenne de ces indices pour les étudiants au sein d'une école (comme dans Agasisti *et al.* (2018) par exemple). Les variables qui sont créées à partir des indices étudiants sont précédées d'un $m$ (pour moyenne). Le tableau 5.1 à la page suivante fournit une courte description des variables qui sont inclues dans la base de données utilisée ici.

Pour finir, en utilisant l'approche développée dans la section 3, je créé une variable qui prend 1 si l'école est une école de la chance et 0 dans le cas contraire (sous la médiane nationale du point de vue de l'équité et de l'efficacité). Autrement dit, les écoles qui sont au-dessus de la médiane nationale du point de vue de la mobilité sociale et de l'efficacité reçoivent la valeur 1. Les écoles qui sont en-dessous de la médiane nationale pour ces 2 indicateurs prennent la valeur 0. Les écoles qui se situent dans les 2èmes et 3èmes cadrans sont supprimées. On compare donc les meilleures écoles avec les mois bonnes écoles (du point de vue des critères de mobilité et d'efficacité). Cette opération est effectuée au sein de chaque pays. Comme dans Hindriks et Godin (2020), les écoles ghettos riches et pauvres sont retirées. Il s'agit des écoles qui se situent dans les déciles supérieurs et inférieurs de la distribution du niveau socio-économique. Le but étant d'éviter un biais lié à l'origine sociale de l'école puisque l'indicateur de mobilité utilisé ici favorise les écoles riches. Les écoles contenant moins que 10 élèves sont également supprimées de l'échantillon.
Par ailleurs, seuls les pays de l'OCDE sont considérés pour cette analyse. Précisons que l'Espagne est retirée de l'échantillon car ses résultats aux tests de lecture n'ont pas été dévoilés. Je me retrouve donc avec 4081 écoles issues de 36 pays pour entamer l'analyse de Machine Learning. La base de données contient 76 variables.

3. Lors de chaque vague d'enquête, l'OCDE publie un rapport technique qui détaille les aspects techniques liés à l'utilisation des données PISA

TABLE 5.1 – Description des variables inclues dans la base de données

| Variable | Description | Variable | Description |
|---|---|---|---|
| *stratio* | Ratio étudiants - professeurs | *village* | L'école se trouve dans un vlillage |
| *schsize* | Taille de l'école | *town* | L'école se trouve dans une petite ville |
| *ratcmp1* | Ordinateurs par étudiant | *city* | L'école se trouve dans une grande ville |
| *ratcmp2* | Ordinateurs connectés à internet | *externconsult* | Consultant externe pour qualité |
| *proatce* | Enseignants entièrement certifiés | *parpartschgov* | Implication des parents dans l'école |
| *proat5am* | Enseignants avec diplôme de master | *equitypol* | Politiques d'équités langue étrangère |
| *clsize* | Taille de la classe | *techteach* | tableaux interactifs/projecteurs par élève |
| *creactiv* | Activités extrascolaires créatives | *mpercomp* | Climat de compétition entre étudiants |
| *edushort* | Manque de matériel scolaire | *mpercoop* | Coopération entre étudiants |
| *staffshort* | Manque de personnel | *mbelong* | Sentiment d'appartenance à l'école |
| *stubeha* | Comportement perturbateur étudiants | *totat* | Nombre total d'enseignants |
| *teachbeha* | Comportement perturbateur professeurs | *discussoutcvol* | Parents discutent les progrès enfants volontairement |
| *lectactiv* | Activités extrascolaires lecture | *mrepeat* | Proportion d'élèves ayant redoublés |
| *ASSESS* | Indice d'utilisation des évaluations | *propboys* | proportion de garçons dans l'école |
| *Acadselect* | Indice de sélection académique | *propgirls* | proportion de filles dans l'école |
| *Accountability* | Compte rendu à des acteurs externes | *Qualeval* | Evaluations internes ou externes pour qualité école |
| *carguidspeccouns* | Conseillers spéciaux en orientation | *nivsocialecole* | Niveau social de l'école |
| *carguid* | Aide à l'orientation à l'école | *mdisclima* | Climat disciplinaire |
| *carguidteach* | Professeurs pour aide orientation | *mstimread* | Etudiant stimulé par le professeur |

TABLE 5.1 – Description des variables inclues dans la base de données

| Variable | Description | Variable | Description |
|---|---|---|---|
| *softavail* | Disponibilité logiciels adéquats | *mictsch* | Disponibilité de matériel ICT à l'école |
| *teachhelpdigituse* | Aide enseignants utilisation digital | *mtmins* | Temps d'apprentissage total |
| *schoolcompetition* | Itensité compétition avec autres écoles | *mlmins* | Temps d'apprentissage en langue du test |
| *ReliSelec* | Admission critères religieux | *mteachsup* | Soutient du professeur |
| *abilitydiffclasses* | Elèves groupés par classes selon aptitudes | *mdirins* | Enseignement explicite |
| *abilitywithinclass* | Elèves groupés au sein de classe par capacités | *mperfeed* | Feedback reçu par le professeur |
| *Difflang* | Langue différente de langue test | *madaptivity* | Adaptation des cours au niveau des élèves |
| *specneeds* | Elèves avec besoins spéciaux | *mteachint* | Enthousiasme professeur pour donner cours |
| *profdevteach* | Formations profs | *Qualgoals* | Objectifs précis |
| *visitteach* | Acceuil de professeurs étrangers | *Qualdata* | Données pour améliorer la qualité |
| *classred* | Réduction classes pour élèves langue étrangère | *discussoutc* | Parents discutent progrès enfants(obligatoire) |
| *Certificate* | Elèves quittant l'école sans diplôme | *SDESCS* | Ecart type niveau socioéconomique |
| *addlessons* | Cours additionnels durant les heures d'école | *musesch* | Utilisation de l'ICT à l'école |
| *roomsavail* | Aide à l'étude : Salles pour étudier | *mictclass* | Utilisation ICT apprentissage |
| *helpwork* | Professeurs dispo aide aux devoirs | *proat5ab* | Professeurs avec un diplôme de bachelier |
| *peerhelp* | scéances de tutorat entre élèves | *proat6* | Professeurs avec un diplôme supérieur master |
| *areaselect* | Sélection basée sur le lieu de résidence | *mbeingbullied* | Expérience de harcèlement |
| *selectfamily* | Sélection critères familiaux | *Schltype* | public ou privé |
| *stufeed* | Feedback élèves pour qualité | *ecoldelachance* | |

## 5.4 Prétraitement des données

A ce stade, la base de données contient un certain nombre de valeurs manquantes. Il faut donc les supprimer (SUPP) ou les remplacer. Le Machine Learning prévoit des techniques pour remplacer les valeurs manquantes. Les écoles pour lesquelles plus de 40 valeurs étaient manquantes ont été supprimées.
Pour le remplacement des données manquantes, les différentes techniques que j'ai mobilisées dans le cadre de mon analyse sont : le remplacement par la moyenne, le remplacement par la moyenne des voisins les plus proches (NN) et le remplacement par la valeur la plus fréquente (MF, pour les variables qualitatives).
Le principe de l'algorithme des voisins les plus proches est de déterminer la valeur d'une donnée manquante en s'intéressant aux voisins de l'observation (une école de ce mémoire) à laquelle cette donnée manquante appartient dans l'espace. Ces voisins sont des écoles pour lesquelles on a de l'information et qui ont des caractéristiques semblables à l'école pour laquelle la valeur est manquante. Il est important de déterminer le nombre de voisins que l'on veut prendre en compte (à l'aide d'une recherche sur grille avec validation croisée). Finalement la moyenne des valeurs pour les voisins les plus proches est prise pour approximer la valeur manquante.
Le jeu de données qui nous occupe contient différents types de variables. Il y a des variables quantitatives, des variables qualitatives ordinales et des variables nominales (binaires). Tous les types de variables ne peuvent pas être imputés de la même façon.

Par ailleurs, pour certains algorithmes de Machine Learning (particulièrement le SVM et RFE), il est utile de normaliser les données. Pour normaliser les données, je retire la moyenne et divise par l'écart type. Il s'agit simplement de la méthode utilisée dans les papiers décrits plus haut.

Les variables nominales qui n'étaient pas binaires ont été transformées en variables binaires pour représenter chaque catégorie séparément.

A partir de ces différents éléments, différentes stratégies d'imputation ont été envisagées, évaluées et comparées. Pour évaluer ces stratégies, le jeu de données est séparé de façon aléatoire en un jeu de test et un jeu d'entrainement (20% dans le jeu de test et 80% dans le jeu d'entraînement). Pour chaque stratégie, un arbre de décision, un SVM et une régression logistique sont entrainés. A chaque fois la meilleure combinaison de paramètres est recherchée à l'aide d'une recherche sur grille avec validation croisée. Pour évaluer ces modèles, la métrique utilisée est l'exactitude (Accuracy).

$$Accuracy = \frac{Nombre\ de\ classifications\ correctes}{Nombre\ de\ classifications}$$

Il s'agit de la proportion d'écoles qui sont correctement classifiées par le modèle. Pour chacune des différentes stratégies de prétraitement essayées, je regarde les résultats obtenus sur le jeu de test ainsi que les résultats d'une validation croisée (cinq parties) sur le set d'entraînement. Le tableau 5.2 donne un aperçu de différentes stratégies envisagées.

TABLE 5.2 – Les différentes stratégies d'imputation envisagées

| | **Quantiatives** | | **Qualitatives** | | **Binaires** |
|---|---|---|---|---|---|
| | *Imputation* | *Standardisation* | *Imputation* | *Standardisation* | Imputation |
| S1 | SUPP | Non | SUPP | Non | SUPP |
| S2 | Moyenne | Non | MF | Non | MF |
| S3 | Moyenne | Oui | MF | Oui | MF |
| S4 | Moyenne | Oui | MF | Non | MF |
| S5 | Moyenne | Oui | Moyenne | Non | MF |
| S6 | Moyenne | Oui | Moyenne | Oui | MF |
| S7 | KNN | Oui | KNN | Oui | MF |
| S8 | KNN | Oui | KNN | Oui | SUPP |
| S9 | KNN | Oui | MF | Oui | MF |
| S10 | Moyenne | Oui | KNN | Oui | MF |
| S11 | Moyenne | Oui | MF | Non | SUPP |

Je suis parti d'un modèle très simple dans lequel j'ai simplement supprimé l'ensemble des valeurs manquantes et ai essayé différentes stratégies. Parmi ces stratégies, la stratégie S7 est la stratégie qui est utilisée dans la littérature.
Le tableau 5.3 donne les résultats obtenus (taux de classifications correctes) sur le jeu de test et ceux de la validation croisée sur le jeu d'entraînement.

TABLE 5.3 – Scores des différentes stratégies

| | **Arbre de décision** | | **Svm** | | **Regression logistique** | |
|---|---|---|---|---|---|---|
| *Modèle* | *Test* | *cross val* | *test* | *Cross val* | *Test* | *Cross val* |
| S1 | 64,737 | 65,13 | 67,36 | 72,02 | 69,474 | 73,17 |
| S2 | 64,277 | 65,26 | \ | \ | 72,471 | 72,15 |
| S3 | 64,277 | 65,26 | 72,08 | 72,09 | 72,343 | 72,12 |
| S4 | 64,277 | 65,26 | 72,08 | 72,06 | 72,343 | 72,28 |
| S5 | 64,277 | 65,26 | 71,575 | 72,09 | 72,215 | 72,06 |
| S6 | 64,277 | 65,26 | 71,703 | 71,99 | 71,831 | 72,15 |
| S7 | 65,685 | 64,24 | 70,807 | 72,34 | 71,63 | 72,22 |
| S8 | 66,287 | 65,09 | 73,16 | 70,83 | 74,104 | 70,99 |
| S9 | 65,429 | 64,24 | 71,703 | 71,93 | 71,191 | 72,06 |
| S10 | 64,277 | 65,26 | 71,319 | 72,317 | 72,471 | 72,31 |
| S11 | 65,63 | 67,12 | 73,473 | 71,07 | 74,43 | 71,07 |

Les éléments suivants me paraissent intéressants concernant le prétraitement :

- Remplacer les données manquantes semble améliorer les performances de la régression logistique et du SVM.
- Standardiser les données améliore les résultats pour le SVM mais importe assez peu pour la régression logistique et l'arbre de décision. C'est surtout la vitesse d'exécution de l'algorithme pour le SVM qui est fortement diminuée.

En revanche il est important de normaliser les données pour l'étape suivante qui consiste en la sélection des meilleurs variables (Guyon et Elisseeff (2003)).

- Au-delà de ces deux changements, les différentes stratégies donnent des résultats similaires.

- Les meilleurs modèles pour la régression logistique et le SVM donnent des résultats aux alentours de 72% de classifications correctes. A titre de comparaison, les meilleurs modèles dans les articles cités précédemment sont souvent entre 75 et 80 % de classifications correctes pour le niveau de l'étudiant.

- L'arbre de décision semble être clairement moins performant sur ce jeu de données (65% de classifications correctes).

Pour rester cohérent avec la littérature, je choisis de poursuivre avec la stratégie S7 pour le SVM et la régression logistique et la stratégie S2 pour l'arbre de décision pour lequel la normalisation n'a aucune influence. Les résultats pour la stratégie S4 (stratégie de prétraitement très élémentaire) pour le SVM et la régression logistique peuvent être trouvés en annexe. Ils sont très semblables.
Dans la section suivante, on poursuit donc avec la stratégie S7 (imputation par les plus proches voisins et normalisation pour les variables qualitatives et quantitatives ordinales, imputation par la valeur la plus fréquente pour les variables binaires) pour extraire les caractéristiques les plus importantes pour la classification dans le cadre du SVM et de la régression Logistique. La stratégie S2 (imputation par la moyenne pour les variables quantitatives et par la valeur la plus fréquente pour les variables qualitatives et binaires) est utilisée pour visualiser l'arbre de décision.

## 5.5 Résultats

Je commence par décrire l'arbre de décision. Ensuite, pour la régression logistique et le SVM, je classe les caractéristiques en utilisant un RFE et je recherche le set optimal en utilisant un RFECV.

### Arbre de décision

La figure 5.4 représente l'arbre de décision obtenu en utilisant la stratégie S2 [4] . Les variables qui apparaissent dans l'arbre sont les suivantes : Un indice qui mesure le climat disciplinaire à l'école (*mdisclima*) [5], le taux de redoublants dans une école (*mrepeat*), la proportion de filles que contient l'école (*propgirls*), la mixité sociale d'une école (*sdescs*) et un indice mesurant le sentiment d'appartenance des élèves à l'école (*mbelong*). Une description plus détaillée des variables peut être trouvé à le

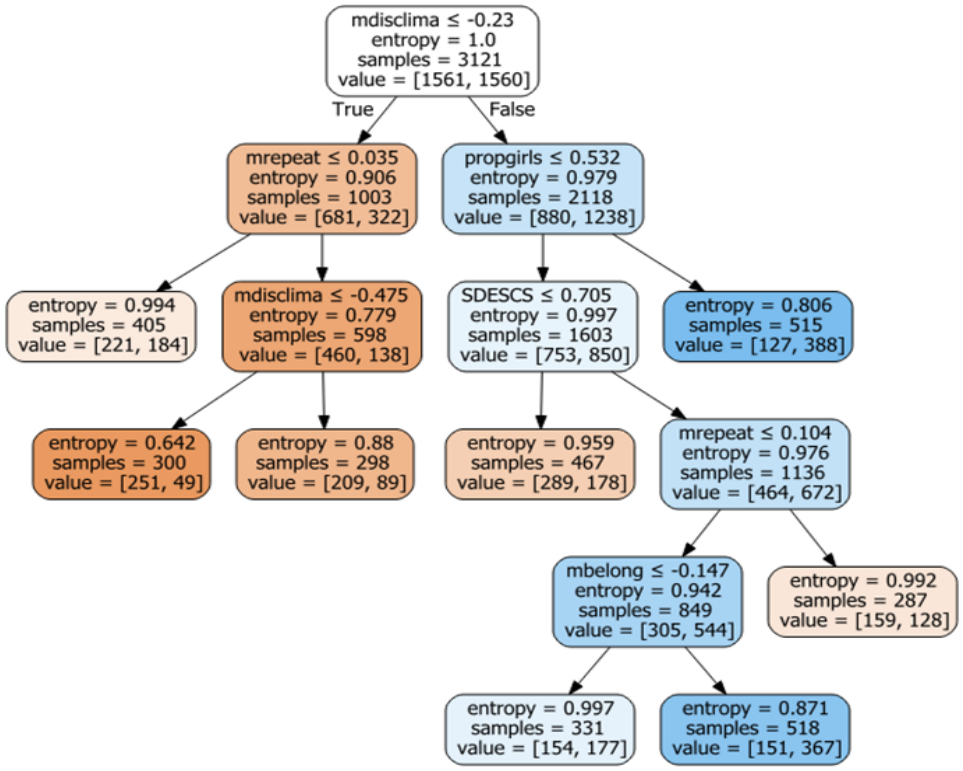
4. Une précision s'impose quant à la représentation donnée ici. Cette représentation ne correspond pas exactement à l'arbre optimal pour cette stratégie. Une contrainte plus forte a en effet été mise sur le nombre d'observations devant se trouver dans les nœuds finaux. Concrètement, les nœuds finaux doivent contenir un plus grand nombre d'écoles dans cette nouvelle spécification. Le but étant d'éviter d'avoir un grand nombre de nœuds finaux ne contenant que quelques écoles. Ceci dans le but de faciliter l'interprétation de l'arbre. Le pouvoir de prédiction ne s'en trouve pas grandement modifié (65.26 vs 63.28). Les variables importantes sont les mêmes.

5. Voir section suivante pour une description plus détaillée des variable

section discussion de cette 2ème partie.

Avant de se lancer dans l'interprétation de cet arbre, il est utile de rappeler que la capacité de classification de l'arbre de décision était moins élevée que celle du SVM et de la régression logistique.

FIGURE 5.4 – Arbre de décision pour la stratégie S2




Sur la figure 5.4, la variable qui se trouve au sommet de l'arbre de décision (dans le nœud racine) est la variable relative au climat disciplinaire d'une école (*mdisclima*) . La question posée dans ce nœud est : « Le climat disciplinaire de l'école est-il supérieur à -0.23 ? ». Ce critère est choisi car c'est celui qui permet de minimiser l'impureté. Pour rappel, l'impureté mesure l'hétérogénéité d'un nœud. Au plus un nœud est homogène (contient des écoles du même type) au moins il est impur. L'impureté est nulle lorsqu'un nœud ne contient qu'un type d'écoles (écoles de la chance ou écoles qui ne sont pas des écoles de la chance). Les 1003 écoles pour lesquelles l'indice de climat disciplinaire moyen des élèves inférieur à -0.23 sont rassemblées dans le nœud inférieur gauche du nœud racine. Celles qui ont un climat disciplinaire supérieur à ce seuil se retrouvent dans le nœud inférieur droit. A chaque nouveau nœud, le processus est répété jusqu'à ce que l'impureté ne puisse plus être réduite.
Par exemple, pour les 1003 écoles qui ont un climat disciplinaire inférieur à -0.23 (nœud en bas à gauche du nœud racine), la question qui permet de minimiser l'impureté est la suivante : « le taux de redoublants dans l'école (*mrepeat*) est-il inférieur à 0.035 ? ». Les écoles pour lesquelles c'est le cas sont rassemblées dans le nœud à gauche de ce nœud et les autres pour lesquelles ce n'est pas le cas vont dans le nœud à droite de ce nœud. Le nœud de gauche est un nœud feuille (nœud final) puisqu'à

ce stade aucun critère ne permet de baisser encore plus l'impureté.
Les nœuds qui contiennent une majorité d'écoles de la chance sont en bleu. Les nœuds qui contiennent majoritairement des écoles qui ne sont pas des écoles de la chance sont en orange. Au plus un nœud est foncé, au plus le nœud est homogène.

Ce processus permet à mon sens d'identifier trois critères de classifications intéressants :

- Parmi les 300 écoles qui ont un très mauvais climat disciplinaire (15%des écoles ont un si mauvais climat disciplinaire) et plus de redoublants que les autres (parmi les 45 % d'écoles qui en ont le plus), on retrouve très peu d'écoles de la chance (16.3%).
- Parmi les 515 écoles avec un climat disciplinaire supérieur à -0.23 (67% des écoles ont un climat disciplinaire aussi élevé) et une proportion de filles supérieure à 53% (22% des écoles ont une proportion de filles plus élevée que 53%), on retrouve 75,3% d'écoles de la chance.
- On retrouve 70%d'écoles de la chance parmi les 518 écoles qui ont un climat disciplinaire favorable (67% supérieurs de la distribution), un bon niveau de mixité sociale (dans les 71% supérieurs de la distribution), peu de redoublants (75% inférieurs de la distribution) et dont l'indice du sentiment d'appartenance à l'école moyen des élèves est bon (parmi les 60 d'écoles qui performent le mieux de ce point de vue).

Pour le reste, les groupes identifiés par l'arbre de décision ne semblent pas faire de distinction tranchée entre les écoles de la chance ou non.

## Support Vector Machine

Je repars du meilleur modèle pour le SVM avec la stratégie de prétraitement S7 (meilleurs paramètres). Je commence par appliquer un SVM avec un algorithme de sélection récursive de caractéristiques/variables (RFE) pour classer les variables selon leur ordre d'élimination. Ensuite, j'applique un RFECV pour trouver un set de variables optimal (du point de vue de la précision de la classification).
Le graphique 5.5 à la page suivante, montre l'évolution du score de validation croisée (proportion d'écoles correctement classifiées) du modèle en fonction du nombre de variables sur lequel il est entraîné (résultat du RFECV).
Deux éléments me semblent intéressants :

- Le pic est à 73. C'est donc avec 73 variables que le score de validation croisée est le plus élevé.
- Avec une quinzaine de variables le modèle prédit déjà si une école est une école de la chance avec une précision aux alentours de 71.35%.

Le tableau 5.4 à la page suivante donne les résultats de validation croisée (5 split) sur le jeu d'entraînement et les résultats sur le jeu de test pour différents sets de variables.

En ne considérant que les 10 variables les plus importantes, on arrive déjà à une prédiction supérieure à 70%. Les sets de 15 et 20 variables améliorent la prédiction

FIGURE 5.5 – Evolution du score en fonction du nombre de variables inclues dans le modèle

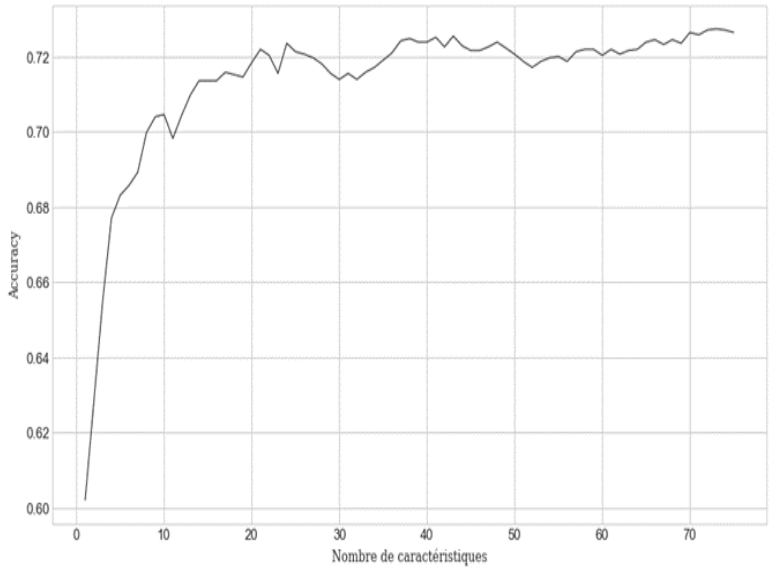
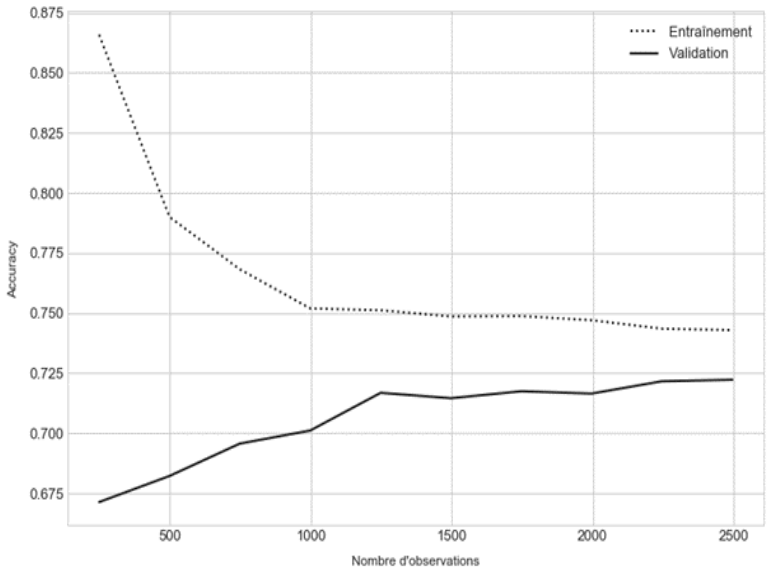



TABLE 5.4 – Scores des différents sets de variables sur le jeu de test et sur la validation croisée

| SCORE TEST | | | | SCORE CROSS VAL | | | | |
|---|---|---|---|---|---|---|---|---|
| *10* | *15* | *20* | *Optimal* | *10* | *15* | *20* | *Optimal* | *NOptimal* |
| 71,19 | 72,09 | 72.09 | 70.09 | 70,45 | 71,35 | 71,83 | 72,73 | 73 |

également. La meilleure prédiction est obtenue avec un set de 73 variables et un score de validation croisée de 72.5 %. Cependant, on voit assez clairement qu'au-delà de 20 variables, la précision de la prédiction ne s'améliore plus beaucoup et fluctue autour de 72%.
La figure 5.6 représente les courbes d'apprentissage pour le modèle optimal (73 variables). Les courbes d'apprentissages montrent comment la capacité de classification du modèle évolue sur le jeu d'entraînement et pour la validation croisée en fonction du nombre d'observations. La courbe en pointillés dans la figure 5 montre comment la précision de la classification sur le jeu d'entraînement évolue avec le nombre d'observations utilisé sur le jeu d'entraînement. La courbe en trait continu montre comment le score de validation croisée évolue avec le nombre d'observations considéré. Concrètement, cela permet de voir comment l'algorithme apprend en fonction du nombre d'observations utilisées. Par ailleurs, les courbes d'apprentissage nous permettent de détecter un éventuel surapprentissage. Dans le cas qui nous occupe, le surapprentissage ne semble pas constituer un problème majeur puisque les deux courbes convergent et finissent assez proche. Précisons que les courbes d'apprentissage pour les modèles contenant moins de variables sont tout aussi proches voir même plus proches.
Finalement, le tableau 5.6 dans la section suivante donne les variables qui sont inclues dans le set de 15 variables et leur ordre d'importance pour la classification.

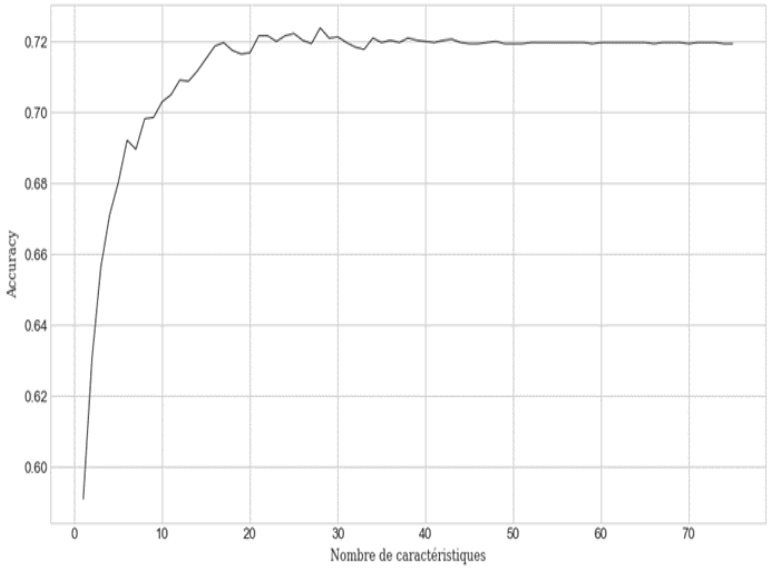
FIGURE 5.6 – Convergence des courbes d'apprentissage pour le modèle optimal




## Régression Logistique

La même approche est utilisée pour identifier un set de caractéristiques optimal pour la régression logistique. La figure 5.7 montre l'évolution du score de la validation croisée en fonction du nombre de caractéristiques.

FIGURE 5.7 – Evolution du score en fonction du nombre de variables dans le cas de la régression logistique



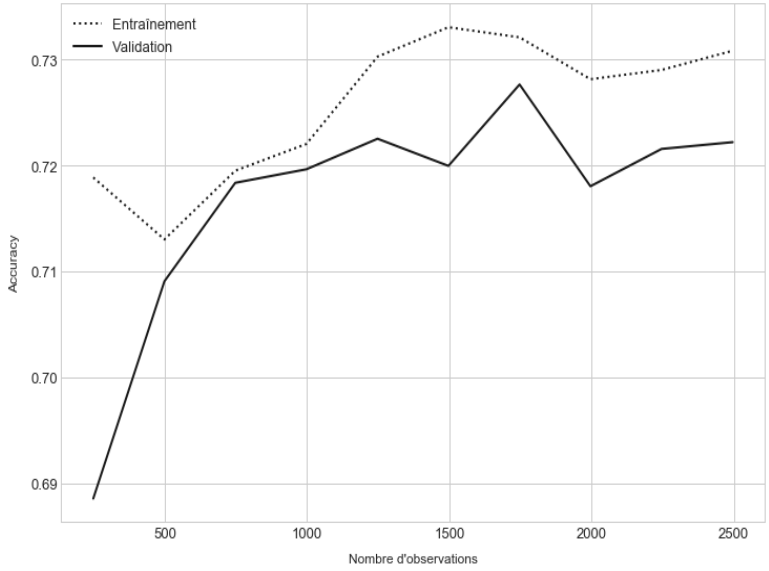

Ici, le set optimal contient 28 variables. Cependant, on voit assez clairement sur la figure 7 qu'après une quinzaine de variables, le modèle ne s'améliore plus sensiblement et tourne autour de 72%. Après une quinzaine de variables, le score de prédiction est déjà aux alentours de 71%.
Le tableau 5.6 donne les scores sur le set de test et la validation croisée.

TABLE 5.5 – Scores pour la régression logistique sur différents sets de variables

| SCORE TEST | | | | SCORE CROSS VAL | | | | |
|---|---|---|---|---|---|---|---|---|
| *10* | *15* | *20* | *Optimal* | *10* | *15* | *20* | *Optimal* | *NOptimal* |
| 69.78 | 71.45 | 70.55 | 70.29 | 70.29 | 71.51 | 71.67 | 72,38 | 28 |

La figure 5.8 montre les courbes d'apprentissage pour le modèle optimal. De nouveau, les scores sur le jeu d'entraînement et le jeu de test sont assez similaires.

FIGURE 5.8 – Convergence des courbes d'apprentissage pour le modèle optimal




Le tableau 5.6 de la page suivante reprend l'importance des différentes variables pour chacun des modèles. Les variables sont classées selon l'importance qu'elles ont pour classifier les écoles. Les 3 premières variables sont le climat disciplinaire, la proportion de redoublants et l'écart type de niveau social pour les 2 algorithmes. Pour le reste, les 2 classements sont assez similaires. Les 15 variables les plus importantes sont les mêmes pour les 2 algorithmes à l'exception de la variable qui décrit la taille de l'école et la variable sur l'organisation d'activités créatives à l'école.

L'ensemble des calculs réalisés dans cette section l'a été avec la libraire Scikit-Learn de Python 3.7. L'inconvénient est qu'il ne s'agit pas d'un logiciel statistique classique et que sa vocation principale est la prédiction et la classification. Il est possible de visualiser les coefficients et leur sens mais obtenir des p valeurs et des intervalles de confiance pour les coefficients est déjà moins évident (bien que ça soit possible).
Pour visualiser de façon plus direct l'importance des différentes variables, j'effectue donc une régression logistique dans Stata avec l'ensemble des variables qui sont présentes dans le tableau 5.7 de la page 46. Les résultats sont présentés dans le tableau 7.
Pour faire simple, les données ne sont pas prétraitées ici. Autrement dit, toutes les

valeurs manquantes sont supprimées et il n'y a pas de standardisation. Ce qui est plus commun dans les applications statistiques en général. Le sens des coefficients est le même que dans l'application de Machine Learning pour toutes les variables à l'exception du type d'école (privé (1) ou public (0)) qui est négatif dans les applications de Machine Learning. Par ailleurs cette variable est la seule qui n'est pas significative.

TABLE 5.6 – Classement des variables pour les 2 modèles

| *Variables* | *SVM* | *Régression Logistique* |
|---|---|---|
| *mdisclima* | 1 | 1 |
| *mrepeat* | 2 | 2 |
| *SDESCS* | 3 | 3 |
| *mperfeed* | 4 | 7 |
| *mteachint* | 5 | 6 |
| *mdirins* | 6 | 4 |
| *propboys* | 8 | 8 |
| *musesch* | 11 | 13 |
| *schsize* | 12 | / |
| *mbeingbullied* | 15 | 10 |
| *nivsocialecole* | 7 | 5 |
| *mpercoop* | 13 | 14 |
| *creactiv* | / | 11 |
| *parpartschgov* | 14 | 15 |
| *schtype* | 9 | 12 |
| *mstimread* | 10 | 9 |

TABLE 5.7 – Régression logistique dans stata

| VARIABLES | (1)<br>ecoldelachance |
|---|---|
| mdisclima | 1.140*** |
| | (0.144) |
| mrepeat | -5.486*** |
| | (0.648) |
| SDESCS | 2.467*** |
| | (0.394) |
| mperfeed | -0.914*** |
| | (0.219) |
| mteachint | 0.867*** |
| | (0.209) |
| mdirins | -1.567*** |
| | (0.214) |
| propboys | -1.650*** |
| | (0.378) |
| musesch | -1.029*** |
| | (0.142) |
| schsize | 0.000541*** |
| | (0.000118) |
| mbeingbullied | -0.432*** |
| | (0.161) |
| nivsocialecole | -0.816*** |
| | (0.133) |
| mpercoop | 0.806*** |
| | (0.169) |
| creactiv | 0.123** |
| | (0.0546) |
| parpartschgov | -0.00848*** |
| | (0.00237) |
| schltype | 0.195 |
| | (0.139) |
| mstimread | 0.676*** |
| | (0.251) |
| Constant | -1.409*** |
| | (0.533) |
| Observations | 2,103 |

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

## 5.6 Discussion et interprétation

Dans cette section, les résultats trouvés dans la section précédente sont discutés. Chacune des variables les plus importante est décrite puis analysée au regard de la littérature. Cela permet notamment d'évaluer la pertinence des résultats au regard de la littérature existante. Les variables discutées sont les 14 variables qui se trouvent dans le top 15 pour les 2 algorithmes ci-dessus.

*Climat disciplinaire (mdisclima)*

Cette variable décrit le climat disciplinaire qui règne au sein de la classe. Il s'agit d'un indice créé par l'OCDE sur base des réponses (à tous les cours, la plupart des cours, certains cours, jamais ou presque jamais) des étudiants aux propositions suivantes :

- Les étudiants n'écoutent pas ce que dit le professeur
- Il y a du bruit et du désordre
- Le professeur doit attendre longtemps avant que les élèves se taisent
- Les étudiants ne peuvent pas travailler de façon optimale
- Les étudiants ne commencent à travailler qu'après un long moment après que le cours a commencé

Pour avoir le climat disciplinaire d'une école, la moyenne de l'indice pour les étudiants de l'école est prise. Une valeur élevée de cette variable indique un bon climat disciplinaire dans l'école. L'impact du climat disciplinaire sur l'efficacité et l'équité d'une école est très largement documenté dans la littérature.

Le climat disciplinaire est associé à un effet positif sur les performances scolaires. Shin *et al.* (2009) trouvent par exemple une forte association entre le climat disciplinaire et résultats en mathématiques pour les étudiants coréens, japonais et américains sur base des données Pisa 2003. Ning *et al.* (2015) trouvent une association positive entre climat disciplinaire et résultats en lecture dans 53 des 65 pays ayant participé à l'enquête PISA 2009.
Par ailleurs,Cheema et Kitsantas (2014) trouvent une association positive du climat disciplinaire sur les résultats en mathématiques pour les élèves américains en utilisant PISA 2003. Ils mettent également en évidence des différences d'effets significatives pour les minorités. Un bon climat disciplinaire tend à être plus important pour les élèves issus des minorités (Afro-Américains et Hispaniques). Cela suggère qu'un bon climat disciplinaire est plus important encore pour les élèves défavorisés. Un bon climat disciplinaire pourrait donc avoir un effet important sur l'équité.
A ce sujet, Agasisti et Longobardi (2017) étudient les caractéristiques des écoles qui sont associées à la résilience des élèves dans les pays de l'union européenne (UE 15) en utilisant les données de PISA 2009. Ils trouvent une association forte entre le climat disciplinaire d'une école et la probabilité qu'un élève soit résilient.
Agasisti *et al.* (2018) confirment ces résultats en utilisant les données PISA 2006 à

2015. Dans leur étude, une augmentation d'une unité dans l'indice de climat disciplinaire double presque la probabilité qu'un élève soit résilient.
Par ailleurs, l'OCDE (OECD (2018)) dans un rapport de 2018 établi que la proportion d'élèves résilients est plus forte dans les écoles qui ont un meilleur climat disciplinaire (OECD (2018)).
Les résultats obtenus dans ce travail sont donc en lien avec la littérature.

*Proportion de redoublants (mrepeat)*

Cette variable indique la proportion d'élèves qui ont redoublés au moins une année parmi les élèves ayant participés au test PISA 2018 au sein de l'école.
L'écrasante majorité des études qui a évalué l'impact du redoublement sur les performances trouve un effet négatif important (Hattie (2009)). En effet, les élèves sont retenus dans des programmes qui ne leur étaient déjà pas bénéficiaires l'année précédente sans que l'école ou les professeurs ne prévoient d'interventions particulières pour ces élèves. De plus, les élèves doivent interagir avec des étudiants d'âges différents. Les élèves issus de milieux défavorisés sont aussi plus susceptibles de redoubler que les autres à résultats égaux (Hattie (2009)).
La proportion d'élèves ayant recommencé une année a un effet négatif sur la probabilité d'être une école de la chance. C'est donc cohérent avec la littérature.

Cependant, l'effet négatif qui est capté ici pour le redoublement tient probablement plus à la sélection des élèves. Les écoles sélectionnant les meilleurs élèves auront évidemment un taux de redoublement inférieur aux autres. Les élèves ayant redoublé pourraient être redirigés vers des écoles qui accueillent plus d'élèves en difficulté après avoir redoublé. Les écoles qui ont le plus de redoublants sont donc probablement des écoles qui ont des élèves dont le niveau initial est plus faible que les autres. Cela ne signifie pas qu'elles ne font pas du bon travail.
C'est une variable qui met en évidence les problèmes de sélection des élèves entre écoles que l'analyse en cross-section ne permet pas de contrôler. L'effet négatif de la proportion d'élèves ayant redoublés était également présent dans Hindriks et Godin (2020).

*Mixité sociale de l'école (sdescs)*

Cette variable caractérise la mixité sociale d'une école. Elle est mesurée par l'écart type de l'indice de niveau socioéconomique des élèves au sein d'une école.
Un écart type de niveau socioéconomique élevé est associé de façon positive aux écoles de la chance. Une mixité sociale importante à l'école est donc associée positivement aux écoles de la chance. Ces effets de la composition pourraient être liés aux effets des pairs. Un élève défavorisé pourrait bénéficier du contact avec un élève issu d'un milieu favorisé. Pour les élèves favorisés, l'effet est moins clair.
Dans un rapport de 2019 (OECD (2019c)), l'OCDE établit un lien négatif entre ségrégation sociale et équité des systèmes scolaires. Les pays dans lesquels la ségrégation dans les écoles est la plus importante ont souvent des systèmes scolaires moins équitables (mesurés par l'intensité et la pente du gradient social). Par ailleurs, l'OCDE (OCDE (2019c)) établi aussi que pour les élèves favorisés, la ségrégation semble avoir un effet légèrement positif sur les résultats. En revanche, cet effet est fortement négatif pour les élèves défavorisés. Selon l'OCDE, le faible effet positif

pour les élèves issus de milieux favorisés est largement insuffisant pour compenser l'effet négatif sur les élèves défavorisés.
L'effet positif de la mixité sociale était également présent dans Hindriks et Godin (2020).

*Feedback du professeur tel que perçu par l'élève (mperfeed)*

Cette variable défini la quantité de feedback donnée par les professeurs tel que perçue par les étudiants. Elle est négativement associée aux écoles de la chance. L'indice de feedback perçu par les étudiants est créé sur base des réponses (jamais ou presque jamais, quelques cours, beaucoup de cours, tous les cours ou presque) sur la fréquence des propositions suivantes aux cours de langue du test :

- Le professeur me dit dans quels domaines je peux encore m'améliorer.
- Le professeur me dit comment je peux améliorer mes performances.
- Le professeur me donne du feedback sur mes forces dans cette matière.

Une valeur élevée pour cet indice indique que l'étudiant perçoit un plus grand feedback de la part des professeurs. L'effet trouvé pour cette variable dans ce mémoire est négatif. Ceci est assez surprenant et va à contre sens de ce que suggère la littérature sur le sujet.
Dans une méta analyse sur les impacts du feedback, Hattie et Timperley (2007) trouvent un effet positif fort du feedback des enseignants sur les performances des élèves. Mais tous les feedbacks ne sont pas efficaces. Il y a une grande variabilité d'effets pour différents types de feedback. Les feedbacks qui comprennent de l'information à propos d'une tâche ou la façon de la réaliser de façon plus efficace ont un effet positif très fort. A l'inverse, des feedbacks qui se basent sur des récompenses extrinsèques, la punition ou qui menacent l'estime de soi, ont tendance à avoir un effet négligeable voir négatif. Le feedback est donc associé à de meilleures performances à condition d'être bien réalisé.
L'effet négatif du feedback qui est trouvé dans ce mémoire vient probablement du fait que les élèves qui reçoivent le plus de feedback sont ceux qui en ont le plus besoin (les élèves en difficulté). C'est en tout cas de cette façon que l'OCDE justifie la corrélation négative trouvée entre fréquence du feedback et performances en sciences après prise en compte du niveau socioéconomique des élèves dans PISA 2015 (OECD (2016)).

*Enthousiasme du professeur (mteachint)*

Cette variable décrit l'enthousiasme du professeur. Elle est positivement associée aux écoles de la chance. Cet indice est basé sur les réponses (pas du tout d'accord, pas d'accord, d'accord, entièrement d'accord) des élèves aux propositions suivantes :

- C'était clair pour moi que le professeur aimait nous donner cours.
- L'enthousiasme du professeur m'a inspiré.
- C'était clair que le professeur aime traiter le sujet du cours.
- Le professeur montrait de l'amusement à donner cours.

L'effet trouvé pour l'enthousiasme du professeur dans ce mémoire est positif. Ceci est cohérent avec la littérature sur le sujet. Beaucoup d'études trouvent un effet positif pour l'enthousiasme du professeur sur les performances scolaires même si certaines études montrent un effet négligeable (Keller *et al.* (2016)). L'effet positif peut s'expliquer de plusieurs façons. Les professeurs pourraient mieux capter l'attention des élèves, une contagion émotionnelle ou encore la transmission de passion par le professeur qui est perçu comme un modèle par les étudiants (Keller *et al.* (2013)).
L'OCDE (OECD (2019a)) trouve aussi un effet positif de l'enthousiasme des enseignants sur les performances des élèves en lecture même après contrôle pour le niveau socioéconomique.
Finalement l'OCDE (OCDE (2019a)) trouve un effet positif entre la résilience académique et l'enthousiasme du professeur en particulier pour les élèves issus de l'immigration.

*Enseignement explicite (mdirins)*

Cette variable évalue dans quelle mesure le professeur donne un enseignement avec des instructions et objectifs clairs. L'indice est créé par l'OCDE sur base des réponses des élèves (tous les cours ou presque, beaucoup de cours, quelques cours, jamais ou presque jamais) aux propositions suivantes :

- Le professeur établi des buts clairs pour notre apprentissage
- Le professeur pose des questions pour vérifier que l'on ait compris ce qui est enseigné
- Au début du cours, le professeur présente un bref résumé de ce qui a été vu lors du dernier cours
- Le professeur nous dit ce qu'on doit apprendre

L'enseignement explicite a pour but d'offrir aux élèves un cours bien structuré, clair et informatif à propos d'un sujet. Cela comprend des explications de la part du professeur et des débats ou des questions des élèves (OECD (2016)). Il est souvent associé à une façon plus passive d'enseigner (OECD (2016)). Cette façon d'enseigner à donc souvent une connotation négative (Hattie (2009)). L'un des principaux avantages de cette méthode est qu'elle permet de voir une grande quantité de matière en peu de temps.
Dans une méta analyse sur l'enseignement explicite, Hattie (2009) trouve un effet positif important de cette façon d'enseigner sur les performances scolaires. Cet effet positif est valable pour tous les étudiants et différentes matières.
Dans un rapport de 2016, l'OCDE (OECD (2016)) fait également état d'un lien positif entre l'enseignement explicite et les résultats scolaires même après prise en compte du niveau socioéconomique. Les coefficients trouvés ici pour l'enseignement explicite sont négatifs. Il s'agit simplement d'un changement dans la façon dont les données sont encodées par l'OCDE en 2018 par rapport à 2015. Pour 2018, une valeur faible pour l'indice signifie plus d'enseignement explicite. Les résultats obtenus sont donc bien cohérents avec la littérature sur le sujet.

*Niveau social de l'école (nivsocialecole)*

Le niveau social de l'école est mesuré par la moyenne du niveau socioéconomique des élèves d'une école. L'effet est négatif. Les écoles de la chance sont donc associées à un niveau social moyen des élèves faible (des écoles socialement défavorisées). A priori, ceci n'est pas cohérent par rapport à ce qu'on retrouve dans la littérature. En effet, dans la littérature, une école favorisée aura tendance à favoriser la mobilité sociale.
L'OCDE (OECD (2018)) montre par exemple que le pourcentage d'élèves défavorisés qui sont résilients est plus élevé dans les écoles favorisées que dans les écoles défavorisées (écart de 22%). Agasisti *et al.* (2018) trouvent également un lien positif entre le niveau socioéconomique d'une école et la résilience académique. Ceci est assez intuitif. A nouveau, ce sont les effets de pairs qui sont supposés jouer. Un élève défavorisé dans une école plus favorisée sera confronté à des élèves de milieux plus favorisés. La résilience académique est donc plus élevée dans ces écoles favorisées.

Le fait que je trouve un résultat opposé ici tient vraisemblablement à l'indicateur de mobilité sociale utilisé dans ce travail. Pour rappel, l'indicateur de mobilité sociale utilisé ici (la mobilité interdécile) donne implicitement un poids plus important à la mobilité sociale des élèves défavorisés. Une école qui en contient plus aura donc plus de chances d'être une école de la chance.
L'effet négatif du niveau socioéconomique de l'école était également présent dans Hindriks et Godin (2020).

*Proportion de garçons (propboys)*

Cette variable décrit la proportion de garçons qu'il y a au sein de l'école. L'effet est négatif. Les écoles contenant moins de garçons semblent avoir plus de chances d'être des écoles de la chance.
Cette relation est compatible avec ce qui est trouvé par Agasisti et Longobardi (2017). Ils trouvent que la probabilité qu'une fille soit résiliente en lecture est bien plus élevée que celle d'un garçon.
En revanche, ce résultat contraste avec d'autres études (OECD (2018)), Agasisti *et al.* (2018)) qui trouvent une probabilité d'être résilient plus élevée pour les garçons en sciences et en mathématiques respectivement.
Ces différences sont dues à la matière considérée (OECD (2018)). En effet, alors que les garçons ont tendance à performer mieux en sciences et en mathématique, les filles performent mieux en lecture.

*Utilisation outils technologiques (musesch)*

Cette variable décrit l'utilisation des outils technologiques à l'école. L'effet est négatif. La variable est créée sur base des réponses (jamais ou presque jamais, une ou deux fois par mois, une ou deux fois par semaine, presque tous les jours, tous les jours) des élèves aux propositions suivantes :

- « Chatter » en ligne
- Utiliser les e-mails à l'école
- Naviguer sur internet pour des travaux scolaires

- Télécharger depuis ou naviguer sur le site de l'école (« intranet »)
- Poster du travail sur le site de l'école
- Jouer à des jeux de rôles sur l'ordinateur à l'école
- La pratique ou l'exercice pour l'apprentissage de langue étrangère ou des mathématiques
- Faire ses devoirs sur un ordinateur de l'école
- Utiliser les ordinateurs de l'école pour des travaux de groupe ou la communication avec d'autres élèves de l'école
- Utilisation d'applications ou sites d'apprentissage

L'analyse effectuée révèle que les écoles de la chance sont moins souvent associées à un usage intensif des TICE (Technologies de l'information et de la communication pour l'enseignement). La littérature au sujet des outils technologiques sur les performances est mitigée. Concernant l'effet spécifique de la variable USESCH avec les données PISA, plusieurs études trouvent un effet négatif (Bulut et Cutumisu (2017), Petko *et al.* (2017), Rodrigues et Biagi (2017).
Pour expliquer cela, plusieurs explications sont avancées (Odell *et al.* (2020)). Par exemple, il se pourrait que les étudiants soient distraits par les outils digitaux et s'engagent dans des activités qui ne soient pas liées à l'apprentissage. Il se pourrait également que les élèves s'attellent moins à la résolution de problèmes complexes dans un contexte où la plupart de leurs questions peuvent trouver réponse facilement (via l'ordinateur) sans que cela ne demande de réflexion excessive.

*Perception de la coopération entre élèves (mpercoop)*

Cette variable est un indice de l'OCDE qui indique la mesure dans laquelle les étudiants perçoivent la coopération entre étudiants au sein de leur école. Son effet est positif. Cet indice est créé sur base des réponses (pas du tout vrai, un peu vrai, très vrai, extrêmement vrai) aux propositions suivantes :

- Les étudiants semblent accorder de l'importance à la coopération
- Il semble que les étudiants coopèrent entre eux
- Les étudiants semblent partager le sentiment que coopérer est important

Dans le cadre de cette analyse, les écoles de la chance sont plus souvent associées à une bonne coopération entre élèves.
Dans un rapport de 2019 sur la vague d'enquête 2018, l'OCDE (OECD (2019b)) trouve une association positive entre la coopération perçue par les étudiants et la mobilité sociale. En effet, la proportion d'élèves résilients est plus élevée parmi les élèves qui disent percevoir un climat de coopération entre élèves. Par ailleurs, il semble y avoir un consensus sur le fait qu'un environnement coopératif améliore les performances (Hattie (2009)). Ceci constitue une illustration supplémentaire de l'importance des effets de pairs dans l'éducation.

*Participation parentale (parpartschgov)*

Cette variable indique la proportion de parents qui participent à la gouvernance au sein de l'école de leurs enfants. L'effet est négatif. Les écoles de la chance sont associées à une implication parentale plus faible.
Dans la littérature, l'effet de l'implication des parents dans la scolarité des élèves sur leurs performances est assez mitigé. Les effets trouvés dépendent du type d'implication et sont parfois faibles voir négatifs (Hattie (2009)).
Dans son rapport sur la vague d'enquête de 2018 (OECD (2019b)), l'OCDE trouve un effet négatif pour la participation des parents dans la gouvernance de l'école. Les scores moyens en lecture sont inférieurs dans les pays ou plus de parents participent à la gouvernance de l'école. L'association persiste même lorsque le niveau socioéconomique et d'autres formes d'implications parentales sont prises en compte.

*Type d'école, public ou privé (schtype)*

Cette variable prend en compte l'effet du type d'école (Public = 0 ou privé = 1). L'effet trouvé ici est négatif. Ceci est surprenant compte tenu du reste de la littérature sur le sujet.
L'OCDE (OECD (2011)) par exemple trouve que les élèves qui fréquentent des écoles privées ont tendance à mieux performer que les autres. Cependant, cette relation tient essentiellement aux caractéristiques des étudiants qui fréquentent ces écoles. En effet, l'effet devient faible lorsque le niveau socioéconomique des élèves est pris en compte. L'OCDE conclut en disant qu'à composition similaire, l'avantage des écoles privées disparait. Trouver un effet négatif pour cette variable n'en reste pas moins étonnant.

*Stimulation de l'engagement en lecture par le professeur (mstimread)*

Cette variable mesure la façon dont l'enseignement stimule les élèves pour les engager en lecture. Elle est positivement associée aux résultats scolaires des élèves. L'indice est créé sur base des réponses (jamais ou presque jamais, dans quelques cours, dans la plupart des cours, dans tous les cours) aux propositions suivantes :

- Le professeur encourage les étudiants à exprimer leur opinion par rapport à un texte
- Le professeur aide les étudiants à lier les histoires qu'ils lisent à leurs vies
- Le professeur montre aux étudiants comment l'information présente dans les textes s'ajoute à ce qu'ils savent déjà
- Le professeur pose des questions qui motivent l'étudiant à participer activement

Dans la littérature, la stimulation de l'engagement en lecture par le professeur est souvent associée à un effet positif important sur le plaisir pris à la lecture (De Naeghel *et al.* (2014), OECD (2019b), Ho et Lau (2018)). La stimulation à l'engagement en lecture a donc un impact important sur la motivation intrinsèque de l'étudiant. Le rôle positif de la motivation intrinsèque sur les résultats en lecture est lui aussi très largement documenté dans la littérature (Retelsdorf *et al.* (2011), Ho et Lau

(2018) par exemple). Il n'est donc pas étonnant de trouver une relation positive entre stimulation de l'engagement en lecture et écoles de la chance.

*Harcèlement à l'école (mbeingbullied)*

Cette variable indique la mesure dans laquelle le problème de harcèlement est récurent dans l'école. Il est associé négativement aux écoles de la chance. Il s'agit d'un indice créé par l'OCDE sur base des réponses des étudiants (jamais ou presque jamais, quelques fois dans l'année, quelques fois par mois, une fois par semaine ou plus) aux propositions suivantes :

- Les étudiants m'ont écarté d'une activité exprès
- Les autres étudiants se sont moqués de moi
- J'ai été menacé par d'autres étudiants

La littérature semble confirmer les effets négatifs du harcèlement sur les performances scolaires. L'OCDE par exemple (OECD (2019b)) trouve un effet négatif du harcèlement sur les résultats scolaires en lecture. En effet, au sein des pays de l'OCDE, les élèves qui rapportaient subir du harcèlement au moins quelques fois par mois avaient en moyenne des résultats inférieurs aux autres élèves après prise en compte du statut socioéconomique.
En outre, les élèves issus de milieux défavorisés semblent être plus souvent touchés par des problèmes de harcèlement. Erberber *et al.* (2015) mettent en évidence un lien négatif significatif entre résilience académique et performances en lecture dans certains pays en utilisant les données du test TIMSS (Trends in International Mathematics and Science Study) 2011.

## 5.7 Conclusion

### Une méthodologie intéressante mais des résultats à considérer avec bon sens

La méthodologie utilisée ici présente l'avantage de considérer un ensemble très important de variables, de sélectionner les plus importantes et de les hiérarchiser en faisant abstraction de toute idée préconçue. A partir d'un ensemble de variables qui à la base était très large, la machine a donc sélectionné les plus importantes et les a classées. L'approche offre donc un cadre d'analyse rigoureux dénué d'idées préconçues ou les seuls biais sont les biais qui sont présents dans les données par opposition à l'intelligence humaine plus susceptible de biais ou idées préconçues.

En revanche, il y a lieu de garder à l'esprit certains éléments concernant le Machine Learning. Les algorithmes ne permettent pas d'établir de causalité. Ils permettent simplement d'identifier des corrélations ou associations dans les données. Cela mène parfois à des résultats surprenants ou inattendus. Dans le cadre de ce mémoire par exemple, les relations négatives entre le type d'école (écoles privées) ou le feedback perçu et les écoles de la chance allaient à l'encontre de ce qui était attendu. De façon général, les résultats des algorithmes de Machine Learning ne

sont pas toujours faciles à interpréter. Il y a peu d'information sur l'intensité de la relation entre les caractéristiques et la variable dépendante. Les mécanismes sous-jacents à cette relation sont également difficiles à mettre en évidence.

Pour résumer, les algorithmes présentent l'avantage de faire un certain « tri » dans les données en faisant abstraction de toute idée préconçue. Mais les résultats doivent être interprétés pour ce qu'ils sont : Une première sélection de variables qui semblent avoir un lien avec les écoles de la chance et qu'il serait intéressant d'étudier plus en profondeur.

## Des résultats intéressants mais des raisons d'être prudent (les principaux enseignements de cette partie)

*Des résultats intéressants*

- La variable qui est en tête de la hiérarchie dans les 3 modèles est le climat disciplinaire. Le taux de redoublants et la mixité sociale d'une école complètent toujours le top 3. Un bon climat disciplinaire, un faible taux de redoublants et une mixité sociale élevée semblent donc être les caractéristiques les plus importantes des écoles de la chance. Les variables qui décrivent les pratiques d'enseignements comme l'enseignement explicite et l'enthousiasme du professeur sont aussi associées de façon récurrente aux écoles de la chance.

- Les variables issues du questionnaire élèves semblent surreprésentées parmi les variables les plus importantes. Ceci pourrait indiquer que les élèves sont plus à même de juger de la situation qu'ils vivent que les directeurs d'écoles. Les variables les plus importantes semblent aussi être liées aux pratiques et politiques d'une école plus qu'aux ressources dont disposent les écoles. Ceci est cohérent avec la littérature (voir Agasisti *et al.* (2018) par exemple).

- Du point de vue de la précision des classifications, la régression logistique et le SVM parviennent à classifier correctement les écoles sur base des caractéristiques observées avec une précision de 72%. Une certaine partie de la variation entre écoles peut donc être expliquée par les variables inclues dans le modèle. A titre de comparaison, la précision des classifications effectuées dans les articles cités plus haut se situe entre 75 et 80% selon les articles. Ces études prédisent le niveau d'un étudiant et ont en général un nombre d'observations et de caractéristiques plus larges.

- Une vingtaine de variables sont suffisantes pour classifier les écoles.

- Les variables considérées comme importantes par les algorithmes et le sens de leurs effets sont cohérentes avec la littérature existante à l'exception de la variable pour le type d'école (public ou privé).

*Des raisons d'être prudent*

Outre la remarque relative à la méthodologie évoquée plus haut, je souhaiterais attirer l'attention sur 3 limites de l'analyse.

- L'analyse ne contrôle pas pour le problème de sélection des élèves au sein des écoles. Les écoles qui sélectionnent les meilleurs élèves par exemple seront plus efficaces que les autres. L'importance du taux de redoublants trouvé dans ce mémoire illustre ces problèmes de sélection.

- Les résultats de prédiction (72% de prédictions correctes) sont intéressants. Mais 28% des écoles sont mal classifiées. Il y a donc une grande partie de la variance entre écoles qu'on ne parvient pas à expliquer. Il y a donc potentiellement d'autres caractéristiques importantes qui n'ont pas été identifiées ici.

- Un certain nombre de variables liées au management et à l'indépendance des écoles notamment, ont montré un effet positif sur l'efficacité et l'équité des écoles par le passé. Elles n'étaient pas présentes dans les données relatives à la vague d'enquête PISA 2018. Aussi, seuls les résultats en lecture ont été considérés ici. Il pourrait être intéressant d'étendre l'analyse à d'autres vagues d'enquêtes. Cela permettrait en outre de s'intéresser à d'autres branches (math et Sciences par exemple).

# 6. Conclusion

Dans ce mémoire travail, je suis reparti de l'approche présentée par Hindriks et Godin (2020) pour identifier les écoles qui combinent excellence scolaire et mobilité sociale (écoles de la chance). Cette approche consistait à faire passer un double test aux écoles. D'une part, un test d'efficacité qui mesure la capacité d'une école à faire réussir les élèves au-delà de ce qui est attendu sur base de leur niveau social. D'autre part, un test d'équité qui mesure la capacité d'une école à favoriser la mobilité sociale de ses élèves. Ce travail présente deux applications de cette approche.

Une première application concerne la comparaison des communautés belges et plus précisément l'évolution de cette comparaison. Pour 2012, Hindriks et Godin (2020) trouvaient une proportion d'écoles de la chance bien plus importante parmi les écoles de la Communauté flamande que parmi les écoles de la Communauté française. Dans ce travail, je me suis intéressé à l'évolution de ces proportions entre 2012 et 2018 avec un résultat intéressant : Alors qu'en 2012, la Communauté flamande avait une proportion d'écoles de la chance bien supérieur à celle de la Communauté française, cet écart était presque inexistant en 2018. Il semble donc qu'il y ait eu un rattrapage des écoles flamandes par les écoles francophones sur cette période. Par ailleurs, ce rattrapage semblait venir d'un changement du point de vue de la mobilité sociale plus que de l'efficacité.
Plusieurs pistes ont été envisagées pour tenter d'expliquer ce rattrapage. J'ai notamment mis en évidence le fait que la composition de l'échantillon n'était pas la même dans les trois vagues d'enquêtes considérées. En effet, la proportion d'élèves flamands dans le bas de l'échelle sociale a baissé entre 2012 et 2018. Puisque l'indicateur de mobilité sociale que j'utilise dans ce mémoire donne plus d'importance à la mobilité sociale des élèves du bas de l'échelle sociale, cela signifie que le potentiel de mobilité sociale des écoles flamandes est inférieur en 2018 qu'en 2012.
Ce travail a donc montré qu'entre 2012 et 2018, l'efficacité des écoles flamandes n'a pas changée malgré l'amélioration du niveau social relatif des élèves flamands. En revanche la mobilité sociale de des écoles flamandes s'est fortement dégradée par rapport aux écoles de la communauté française. En communauté française, l'efficacité n'a pas changée malgré la détérioration du niveau social. Par ailleurs, la mobilité sociale des écoles de la communauté française par rapport aux écoles flamandes s'est considérablement améliorée.

Une deuxième application concerne l'identification des caractéristiques communes des écoles de la chance. Pour cela, j'ai utilisé une méthode encore peu utilisée en économie : le Machine Learning. L'idée du Machine learning est d'apprendre par les données. Dans le contexte de ce mémoire cette méthode a deux atouts majeurs.
Cette méthode m'a permis de sélectionner et hiérarchiser les variables les plus importantes pour discriminer les écoles de la chance et les autres écoles en faisant abstraction de toute idée préconçue. L'approche par le Machine Learning ne prétend pas établir de causalité. (De toute façon il serait difficile d'établir des liens de causalité avec les données en coupe transversale comme le sont les données PISA.

Elle permet simplement de dégager des corrélations et des associations dans les données. Un des défauts du Machine learning est la difficulté à laquelle on fait face lorsqu'on doit interpréter les données.
Les algorithmes de Machine Learning suggèrent que la variable qui différencie le mieux les écoles de la chance des autres écoles est le climat disciplinaire qui règne à l'école. D'autres variables que les algorithmes ont permis d'identifier sont la mixité sociale de l'école, le taux de redoublants et deux variables liées aux pratiques d'enseignement, l'enseignement explicite et enthousiasme du professeur.
Ce mémoire a également montré que sur base d'une vingtaine de caractéristiques, les algorithmes de Machine learning (SVM et régression logistique) parviennent à prédire avec une précision de 72 % la catégorie d'une école (école de la chance ou non).

# Bibliographie


Agasisti, T., Avvisati, F., Borgonovi, F. et Longobardi, S. (2018). Academic resilience. *Documents de travail de l'OCDE sur l'éducation*, (167).

Agasisti, T. et Longobardi, S. (2014). Inequality in education : Can italian disadvantaged students close the gap ? *Journal of Behavioral and Experimental Economics*, 52:8–20.

Agasisti, T. et Longobardi, S. (2017). Equality of educational opportunities, schools' characteristics and resilient students : An empirical study of eu-15 countries using oecd-pisa 2009 data. *Social Indicators Research*, 134(3):917–953.

Agirdag, O. (20/11/2020). Kwaliteitsdaling in het onderwijs. schuld van migratie ? https://urlz.fr/dDd9. *kuleuvenblogt.be*, Consulté le 15/08/2020.

Alivernini, F. (2013). An exploration of the gap between highest and lowest ability readers across 20 countries. *Educational Studies*, 39(4):399–417.

Alivernini, F., Manganelli, S. et Lucidi, F. (2016). The last shall be the first : Competencies, equity and the power of resilience in the italian school system. *Learning Individual Differences*, 51:19–28.

Athey, S. (2018). *The impact of machine learning on economics*, pages 507–547. University of Chicago Press.

Bulut, O. et Cutumisu, M. (2017). When technology does not add up : Ict use negatively predicts mathematics and science achievement for finnish and turkish students in pisa 2012. *In EdMedia and Innovate Learning*, pages 935–945. Association for the Advancement of Computing in Education (AACE).

Card, D. (1999). *The causal effect of education on learnings*, volume 3, pages 1801–1863. Elsevier.

Chang, Y.-W. et Lin, C.-J. (2008). Feature ranking using linear svm. *In Causation and Prediction Challenge*, pages 53–64.

Cheema, J. R. et Kitsantas, A. (2014). Influences of disciplinary classroom climate on high school student self-efficacy and mathematics achievement : A look at gender and racial–ethnic differences. *International Journal of Science Mathematics Education*, 12(5):1261–1279.

Chen, J., Zhang, Y., Wei, Y. et Hu, J. (2019). Discrimination of the contextual features of top performers in scientific literacy using a machine learning approach. *Research in Science Education*.

Chetty, R., Hendren, N., Kline, P. et Saez, E. (2014). Where is the land of opportunity ? the geography of intergenerational mobility in the united states. *The quarterly journal of economics*, 129(4):1553–1623.

Cingano, F. (2014). Trends in income inequality and its impact on economic growth. *Documents de travail de l'OCDE sur les questions sociales, l'emploi et les migrations*, (163).

Coleman, J., Campbell, E., Hobson, C., McPartland, J., Mood, A., Weinfeld, F. et York, R. (1966). Equality of educational opportunity. *US Government Printing Office*.

Cortes, C. et Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3):273–297.

De Naeghel, J., Valcke, M., De Meyer, I., Warlop, N., Van Braak, J. et Van Keer, H. (2014). The role of teacher behavior in adolescents' intrinsic reading motivation. *Reading and Writing*, 27(9):1547–1565.

Dong, X. et Hu, J. (2019). An exploration of impact factors influencing students' reading literacy in singapore with machine learning approaches. *International Journal of English Linguistics*, 9(5).

Downey, D. B. et Condron, D. J. (2016). Fifty years since the coleman report : Rethinking the relationship between schools and inequality. *Sociology of Education*, 89(3):207–220.

EEF (2016). Teaching  learning toolkit : Setting or streaming. *Educational Endowment Foundation*.

Erberber, E., Stephens, M., Mamedova, S., Ferguson, S. et Kroeger, T. (2015). Socioeconomically disadvantaged students who are academically successful : Examining academic resilience cross-nationally. policy brief no. 5. *International Association for the Evaluation of Educational Achievement*.

Godin, M. et Hindriks, J. (2018). An international comparison of school systems based on social mobility. *Economie et Statistique*, 499(1):61–78.

Gords, P. (3/12/2019). Opdoffer voor vlaams onderwijs : nu ook koppositie voor wiskunde kwijt. https://urlz.fr/dDd6. *DeMorgen*, Consulté le 15/08/2020.

Gorostiaga, A. et Rojo-Álvarez, J. L. (2016). On the use of conventional and statistical-learning techniques for the analysis of pisa results in spain. *Neurocomputing*, 171:625–637.

Guyon, I. et Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182.

Guyon, I., Weston, J., Barnhill, S. et Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine learning*, 46(1-3):389–422.

Hattie, J. (2009). *Visible learning : A synthesis of over 800 meta-analyses relating to achievement*. Routledge, Abingdon.

Hattie, J. et Timperley, H. (2007). The power of feedback. *Review of educational research*, 77(1):81–112.

Hindriks, J. et Godin, M. (2017). The school of opportunity. *In* ESL, éditeur : *Equity in Education*. KULeuven.

Hindriks, J. et Godin, M. (2020). Quelles sont les écoles qui combinent excellence et mobilité sociale ? *CORE Discussion Paper*, (14/2020).

Ho, E. S. C. et Lau, K. (2018). Reading engagement and reading literacy performance : Effective policy and practices at home and in school. *Journal of Research in Reading*, 41(4):657–679.

Jerrim, J., Lopez-Agudo, L. A., Marcenaro-Gutierrez, O. D. et Shure, N. (2017). What happens when econometrics and psychometrics collide ? an example using the pisa data. *Economics of Education Review*, 61:51–58.

Keller, M., Neumann, K. et Fischer, H. E. (2013). Teacher enthusiasm and student learning. *International guide to student achievement*, pages 247–250.

Keller, M. M., Hoy, A. W., Goetz, T. et Frenzel, A. C. (2016). Teacher enthusiasm : Reviewing and redefining a complex construct. *Educational Psychology Review*, 28(4):743–769.

Liu, X. et Ruiz, M. E. (2008). Using data mining to predict k–12 students' performance on large-scale assessment items related to energy. *J Journal of Research in Science Teaching*, 45(5):554–573.

Masci, C., Johnes, G. et Agasisti, T. (2018). Student and school performance across countries : A machine learning approach. *European Journal of Operational Research*, 269(3):1072–1085.

Mladenić, D., Brank, J., Grobelnik, M. et Milic-Frayling, N. (2004). Feature selection using linear classifier weights : interaction with classification models. *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*.

Ning, B., Van Damme, J., Van Den Noortgate, W., Yang, X. et Gielen, S. (2015). The influence of classroom disciplinary climate of schools on reading achievement : A cross-country comparative study. *School Effectiveness and School Improvement*, 26(4):586–611.

OCDE. *Balancing School Choice and Equity : An International Perspective Based on Pisa*. Éditions OCDE.

OCDE. *PISA 2018 Results (Volume II) :Where all students succeed*. Editions OCDE.

Odell, B., Galovan, A. M. et Cutumisu, M. (2020). The relation between ict and science in pisa 2015 for bulgarian and finnish students. *EURASIA Journal of Mathematics, Science Technology Education*, 16(6):em1846.

OECD (2009). *PISA Data Analysis Manual : SAS, Second Edition : SAS*. OECD Publishing, FR, 2nd édition.

OECD (2011). Private schools : who benefits ? *PISA in Focus*, (7).

OECD (2016). *PISA 2015 Results (Volume II). Policies and Practices for Successful Schools*. OECD publishing.

OECD (2018). *Equity in education : Breaking down barriers to social mobility*. OECD publishing.

OECD. (2019b). *PISA 2018 Results (Volume III) : What School Life Means for Students' Lives*. OECD publishings.

Parker, P. D., Marsh, H. W., Jerrim, J. P., Guo, J. et Dicke, T. (2018). Inequity and excellence in academic performance : Evidence from 27 countries. *American Educational Research Journal*, 55(4):836–858.

Pauli, W. (3/06/2020). De franstaligen pakten de coronacrisis professioneler aan. *Knack*, 23(23).

Petko, D., Cantieni, A. et Prasse, D. (2017). Perceived quality of educational technology matters : A secondary analysis of students' ict use, ict-related attitudes, and pisa 2012 test scores. *Journal of Educational Computing Research*, 54(8):1070–1091.

Retelsdorf, J., Köller, O. et Möller, J. (2011). On the effects of motivation on reading performance growth in secondary school. *Learning and instruction*, 21(4):550–559.

Rodrigues, M. et Biagi, F. (2017). Digital technologies and learning outcomes of students from low socio-economic background : An analysis of pisa 2015. *Joint Research Centre Science for Policy Report*.

Sanzana, M. B., Garrido, S. S. et Poblete, C. M. (2015). Profiles of chilean students according to academic performance in mathematics : An exploratory study using classification trees and random forests. *Studies in Educational Evaluation*, 44:50–59.

Shaw, B., Baars, S., Menzies, L., Parameshwaran, M. et Allen, R. (2017). Low income pupils' progress at secondary school.

Shin, J., Lee, H. et Kim, Y. (2009). Student and school factors affecting mathematics achievement : International comparisons between korea, japan and the usa. *School Psychology International*, 30(5):520–537.

Sirin, S. R. (2005). Socioeconomic status and academic achievement : A meta-analytic review of research. *Review of educational research*, 75(3):417–453.

Thomson, S. (2018). Achievement at school and socioeconomic background—an educational perspective. *npj Science Learn*, 3(5).

Xiao, Y. et Hu, J. (2019). Assessment of optimal pedagogical factors for canadian esl learners' reading literacy through artificial intelligence algorithms. *International Journal of English Linguistics*, 9(4):1–14.

Zhu, J. et Hastie, T. (2004). Classification of gene microarrays by penalized logistic regression. *Biostatistics*, 5(3):427–443.

Zieger, L., Jerrim, J., Anders, J., Shure, N. *et al.* (2020). Conditioning : How background variables can influence pisa scores. Rapport technique, Centre for Education Policy and Equalising Opportunitites, UCL Institute of Education.

# A. Annexes

## Proportions d'écoles de la chance avec d'autres valeurs pour r

TABLE A.1 – Ecoles de la chance avec d'autres valeurs pour r

| | *r = 0,5* | | *r = 0,25* | |
|---|---|---|---|---|
| | FL | FR | FL | FR |
| 2012 | 47 | 19 | 50 | 20 |
| 2015 | 39 | 17 | 43 | 18 |
| 2018 | 31 | 29 | 36 | 33 |

## Résultats pour la stratégie de prétraitement S4 pour la régression logistique et le SVM

TABLE A.2 – Résultats en utilisant les stratégies de l'article

| | SCORE TEST | | | | SCORE CROSS VAL | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | *10* | *15* | *20* | *Optimal* | *10* | *15* | *20* | *Optimal* | *NOptimal* |
| *SVM* | 71,19 | 72,09 | 72,09 | 70,9 | 70,5 | 71,35 | 71,8 | 72,73 | 73 |
| *RL* | 69,78 | 71,45 | 70,55 | 70,29 | 70,29 | 71,51 | 71,67 | 72,38 | 28 |

TABLE A.3 – Classement en utilisant la stratégie de prétraitement S4

| *Variables* | *SVM* | *Régression Logistique* |
|---|---|---|
| *mdisclima* | 1 | 1 |
| *mrepeat* | 2 | 2 |
| *SDESCS* | 3 | 3 |
| *mperfeed* | 4 | 7 |
| *mteachint* | 5 | 6 |
| *mdirins* | 6 | 4 |
| *propboys* | 8 | 8 |
| *musesch* | 9 | 12 |
| *schsize* | 12 | 11 |
| *mbeingbullied* | / | 15 |
| *nivsocialecole* | 7 | 5 |
| *mpercoop* | 10 | 13 |
| *creactiv* | 15 | / |
| *parpartschgov* | 13 | 14 |
| *schtype* | 11 | 9 |
| *madaptivity* | 14 | / |
| *mstimread* | / | 10 |