MATHEMATIQUES POUR LE BIG DATA

Sous la direction de M. Henri LAUDE

Par AUFRERE Thuy, MAZZUCATO Claire, REN Claude

10/01/2021

Contents

Introduction

Le XXIème siècle peut être défini comme le siècle du numérique. Le monde contemporain est un monde ultra digitalisé et connecté. Cela est d'autant vrai avec la crise sanitaire qui a frappé un grand nombre de pays en 2020. Cette crise sanitaire a fortement accentué ce monde numérique, notamment dû aux stratégies de confinement menées par certains pays qui ont obligé certains secteurs tels que l'Éducation à ne fonctionner que via le numérique. Les élèves et les étudiants ont passé une grande partie de leur année 2020 à n'apprendre qu'à travers un écran.

L'apprentissage dispensé par le prisme du numérique est-il aussi optimal que l'apprentissage dans une salle de classe ? L'apprentissage par le prisme du numérique favorise-t-il le décrochage scolaire ? Il est encore tôt pour savoir si l'année 2020 est l'année qui a vu augmenter le nombre de décrochages scolaires. Mais ces questions sont légitimes et en deviennent même une question sociétale. En effet, l'Union Européenne a même adopté une « Stratégie Europe 2020 » qui vise à lutter contre le décrochage scolaire.

C'est pourquoi nous avons voulu travailler sur 3 papiers ayant pour thème la prédiction, plus particulièrement sur la prédiction dans le milieu scolaire avec le décrochage scolaire. Est-il possible de prédire si un élève va abandonner son cursus scolaire grâce à son comportement numérique ? Si nous sommes capables de le prédire, alors des actions peuvent être définies en amont pour empêcher les élèves de décrocher du monde scolaire.

Les 3 papiers tentent de répondre aux questions suivantes :

- Qu'est-ce qu'une bonne prédiction ? (papier 1)
- Peut-on prédire le décrochage des personnes qui suivent les MOOCs ? (papier 2)
- Peut-on prédire les étudiants à risque avec leur comportement d'apprentissage online et offline ? (papier 3)

Il nous a semblé important de comprendre ce qu'est une « bonne prédiction » (d'où la sélection du papier 1) avant de parler de papiers qui proposent des formules mathématiques permettant de faire des prédictions. Avant de discuter sur les 3 papiers, il est pertinent de bien distinguer la prédiction de la prévision. Si la prévision peut sembler similaire à l'analyse prédictive ou à la simple prédiction, elle est en fait différente.

La prévision est un processus qui consiste à prévoir ou à estimer des événements futurs sur la base de données passées et présentes et, le plus souvent, par l'analyse de tendances ou de modèles de données. Les prévisions se basent sur des informations temporelles et se fondent sur des chiffres, des tendances et la saisonnalité pour prédire un résultat alors que la prédiction vise plutôt à comprendre et de prédire le comportement des individus. Grâce à des modèles de prévision basés sur l'intelligence artificielle, on peut de plus en plus déterminer les tendances qui domineront le marché. Grâce à des modèles de prédictions basés sur l'intelligence artificielle, on peut de plus en plus déterminer le comportement des individus.

1. Papier 1 : What is a Good Prediction- Issues in Evaluating General Value Functions Through Error

1.1 Finalité du papier 1

Ce papier tente de démontrer l'importance de mettre en place des approches d'évaluation sur les prédictions. Pour comparer les prédictions, il est nécessaire de disposer de mesure ou d'un moyen d'évaluation. La prédiction est la capacité pour une machine d'adopter un comportement orienté vers un but : capacité à apprendre, à planifier et à agir afin d'accomplir une tâche grâce à l'acquisition d'une connaissance continue. L'intelligence artificielle se base sur des approches prédictives partant de la construction de la connaissance d'un agent qui va être transformée en connaissance machines pour faire de la prédiction. Avec les évaluations actuelles sous-développées, il est pertinent d'avoir des évaluations d'une prédiction avant de l'implémenter dans une machine.

L'approche General Value Functions (GVFs – en français Fonctions de Valeur Générale) existe pour formuler des prédictions, cette approche s'appelle « predictive knowledge » (la connaissance prédictive). Le postulat d'aujourd'hui est que la connaissance (et donc la possibilité de faire des prédictions) peut être construite en ligne, en temps réel, au fur et à mesure qu'un agent interagit avec son environnement. On peut penser que l'évaluation d'une prédiction repose sur la capacité de la machine à apprendre de ses erreurs. Cependant, la machine est-elle capable de choisir ce qu'elle doit prévoir, sans l'intervention d'un ingénieur ?

Selon les auteurs du papier, il existe une approche pour évaluer si une prédiction est bonne ou non : c'est l'approche RUPEE (Recent Unsigned Projected Error Estimate - l'erreur du retour). RUPEE estime l'erreur moyenne projetée d'un GVF. Ce papier cherche à démontrer que RUPEE n'est pas une approche adaptée pour évaluer une prédiction. A première vue, l'utilisation de l'erreur pour différencier si une prédiction est bonne ou mauvaise peut sembler efficace. Ce n'est pas le cas. Ce n'est pas parce que l'erreur de retour d'une prédiction est faible que la GVFs est bonne et inversement. Selon les auteurs, « ce n'est pas parce qu'une prédiction est exacte qu'elle est utile pour informer le comportement ». En effet, le degré d'erreur d'une prédiction n'informe pas sur l'utilité d'une prédiction pour informer d'autres prédictions. Pour évaluer une prédiction, il est nécessaire de mettre l'accent sur la pertinence des caractéristiques. Pour les auteurs, une bonne prévision est celle dont les caractéristiques sont bien alignées avec le problème de prédiction.

Ainsi ce papier cherche à montrer que les méthodes actuelles communes d'évaluations des connaissances prédictives ne permettent pas de faire la différence entre les prédictions utiles et les prédictions ordinaires : c'est-à-dire entre les prédictions utiles pour la prise de décision et l'apprentissage et celles qui ne le sont pas. Les auteurs proposent d'améliorer les méthodes actuelles d'évaluations en suggérant aussi d'évaluer les prédictions non seulement en fonction de leur propre erreur de prédiction mais aussi de l'erreur des autres prédictions qui en dépendent.

1.2 Compréhension des formules mathématiques du papier 1

Les auteurs de ce papier font référence à la formule du GVF pour déterminer une prédiction. Les fonctions de valeur stockent la récompense cumulative prévue d'un agent à partir d'un état. Chaque fonction de valeur peut être décrite comme une réponse à une question de départ.

GENERAL VALUE FUNCTIONS (GVF)

$$G_t = E_\Pi(\sum_{k=0}^{\infty}(\prod_{j=1}^{k}(\gamma_{t+j})C_{t+k+1}) \tag{1}$$

Explication de la formule :

- GVF estime la somme (ou rendement) actualisée de certains signaux C, ici les actions, sur des pas de temps discrets t = 1,2,3,. . . ,n
- L'objet de la prédiction du GVF est déterminé par les paramètres de la question, notamment le signal d'intérêt C
- Un paramètre $0 \leq \gamma \leq 1$ qui est un facteur qu'on applique à l'action C
- Un paramètre Π qui décrit le comportement sur lequel porte les prédictions, tourner à droite ou gauche par exemple

NB : A noter que dans le papier ils ont mis "$0 \leq \gamma \geq 1$", ce qui semble évidemment être étrange, étant donné l'exemple, l'action de toucher est soit fausse donc 0 soit vrai donc 1.

Les GVF peuvent être interprétées comme un rendement attendu à une action C. En ce qui concerne les prédictions, à chaque action C, on obtient un vecteur d'observations qui décrit l'environnement et prend une mesure. Les observations sont utilisées pour construire la fonction ϕ.

Cette fonction à laquelle on applique un poids va être l'estimation de la prédiction grâce à laquelle nous pouvons calculer la fameuse erreur.

EQUATION SUR L'ERREUR DE RETOUR

$$G_t^e = E_\Pi(\sum_{k=0}^{b}(\prod_{j=1}^{k}(\gamma_{t+j})C_{t+k+1}) - V_t(\phi(o_t)) \tag{2}$$

Cette équation représente tout simplement l'équation (1) à laquelle on soustrait la fonction qui représente l'estimation, le nombre b représentant le nombre de pas l'on veut chercher à estimer. C'est notamment cette équation que les auteurs remettent en cause dans leur papier.

1.3 Commentaire sur le papier 1

Il aurait été intéressant de la part des auteurs de ce papier de donner un exemple concret de la vie réelle et non que des exemples abstraits. Il aurait été pertinent, et peut-être plus compréhensible pour tous, que les auteurs réalisent une réelle prédiction et qu'ils en fassent eux-mêmes une évaluation.

Par ailleurs, certaines notions du papier sont expliquées rapidement, rendant difficile certains passages à comprendre.

2. Papier 2 : Dropout Prediction over Weeks in MOOCs via Interpretable Multi-Layer Representation Learning

2.1 Finalité du papier 2

Les cours ouverts en ligne (MOOC) sont devenus des plateformes populaires d'apprentissage en ligne. Si ces cours permettent aux étudiants d'étudier à leur propre rythme, cette flexibilité peut amener les étudiants à plus facilement abandonner un cours. L'objectif de ce papier est de prédire la probabilité qu'un élève abandonne dès la première semaine le cours qu'il suivait, en fonction des données de clickstream (en français flux de clics).

L'hypothèse posée par les auteurs est que le clickstream d'un élève sous-entend certains modèles comportementaux d'abandon. Cependant la prédiction d'un comportement d'abandon est complexe. Tout d'abord, il est difficile d'extraire une représentation comportementale significative d'un utilisateur à partir de données de faible niveau sur le flux de clics, car l'ensemble de données dépend de nombreux facteurs différents et peut-être inconnus tels que le style d'apprentissage de l'utilisateur, le programme ou encore le contenu de la semaine. En outre, il est difficile d'utiliser les approches existantes pour analyser les données de flux de clics. Alors que de nombreuses approches supposent que l'ensemble des données d'entrée arrive à intervalles réguliers, ou est de la même longueur pour chaque observation, le clickstream pour chaque utilisateur peut être différent en termes d'intervalle et de longueur. Afin d'obtenir des résultats interprétables, il est pertinent de s'appuyer sur l'algorithme des branches et des liaisons (BB) via l'exploitation de données de la plateforme Coursera.

La Méthode proposée par les auteurs est de construire une représentation significative et efficace du modèle de comportement d'un élève à partir de son flux de clics sur la plateforme. L'objectif est d'obtenir ce modèle de manière non supervisée afin que des modèles de classification simples comme le perceptron multicouche (MLP) puissent prédire l'abandon de manière comparable à des modèles de classification plus complexes. Compte tenu des flux de clics répétitifs, la méthode utilisée par les auteurs est un algorithme de branche et de liaison modifié et un algorithme MLP.

Dans ce papier de recherche, les auteurs utilisent un algorithme basé sur la séparation et l'évaluation (branch and bound). Il consiste à effectuer une énumération systématique de solutions du problème, du calcul de coût pour chacune, puis de donner le minimum : l'ensemble des solutions candidates est considéré comme formant un arbre enraciné avec l'ensemble complet à la racine. L'algorithme explore les branches de cet arbre, qui représentent des sous-ensembles de l'ensemble de solutions. Avant d'énumérer les solutions candidates d'une branche, la branche est comparée aux limites supérieures et inférieures estimées de la solution optimale, et est rejetée si elle ne peut pas produire une meilleure solution que la meilleure trouvée jusqu'à présent par l'algorithme.

Les auteurs caractérisent les semaines où un apprenant abandonne et celles où il n'abandonne pas en tirant profit des scores d'action interprétables calculés par l'algorithme BB. Certaines actions ont présenté des scores significativement différents entre les semaines de non-abandon et d'abandon. Pour chaque semaine, le score a été calculé pour chaque action et comparé à la moyenne entre les deux groupes (semaines de non-abandon et d'abandon) en utilisant le test t des deux échantillons. Parmi toutes les actions, 10 actions ayant le plus petit t-score ont été caractérisées comme le groupe de « non-décrocheurs » tandis que les actions ayant le plus grand t-score ont été caractérisées comme étant le groupe de « décrocheurs ». Le groupe des non-décrocheurs avait comme des actions qui incluent la réalisation de "Quiz". Cela nous permet de pressentir que les apprenants qui réussissent sont susceptibles de répondre à des quiz qui peuvent les motiver et booster leur intérêt sur le cours concerné. Le groupe des décrocheurs se caractérise par des actions telles

que “SeekBw”, “Pause” et “SeekFw”, ce qui peut être interprété comme signifiant que les apprenants sont susceptibles de décrocher lorsqu'ils ont des difficultés avec des concepts complexes.

Les résultats obtenus à l'aide de l'algorithme Branch and Bound contiennent deux types d'informations : les informations séquentielles entre les semaines et les informations sur la cooccurrence au niveau des actions. Pour compléter ce modèle, les auteurs utilisent le perceptron multicouche (appelé MLP) pour générer une représentation des caractéristiques qui saisit les deux informations.

2.2 Compréhension des formules mathématiques du papier 2

$$\underset{W_{a,b},b_{a,o}}{min} \frac{1}{2wT} \sum_{t=1}^{T} \sum_{-w \leq i \leq w, i \neq 0} (\hat{y}_t - x_{t+1})^2 \tag{3}$$

Le but de l'apprentissange est de minimiser cette fonction ce qui est logique. $\hat{y}_t$ est la représentation de l'action finale à la semaine t, T est le nombre total de semaines, x_{t+i} est la représentation de l'action préliminaire à la semaine t+i obtenue à partir de l'algorithme Branch and Bound.

L'équation est une double somme sur l'ensemble des semaines t et sur un intervalle $2w$ autour de la semaine t. La fonction au carré $(\hat{y}_t - x_{t+1})^2$ est utilisée de manière à prendre en compte des écarts positifs. Le facteur $\frac{1}{2wT}$ est un facteur de normalisation.

2.3 Résultat obtenu du papier 2

A l'issu des modèles utilisés par les auteurs dans ce papier de recherche, l'algorithme BB a montré des performances comparables à des bases de référence spécifiques à des tâches assez complexes. Le MLP-LFR, en revanche, ne parvient pas à réaliser d'interprétations significatives. Les auteurs identifient certaines limites à l'expérience de prédiction réalisée. Tout d'abord, le modèle ne prend pas en compte le progrès de l'étudiant. Le modèle adopte une approche assez simple consistant à travailler avec une séquence de représentations d'actions au niveau de la semaine. Comme les étudiants progressent à travers les cours à un rythme différent selon le contenu du cours et leur niveau de compréhension, un étudiant peut prendre beaucoup plus de temps que les autres pour terminer le cours, et par conséquent, la variance du flux de clics de la semaine peut être très élevée, ce qui entrave la formation de MLP-LFR. Par conséquent, les auteurs supposent qu'il serait plus significatif et efficace d'apprendre des représentations d'actions basées sur les progrès réels d'apprentissage de chaque étudiant.

Enfin, le modèle n'utilise pas d'informations personnelles (par exemple, l'âge et le niveau d'éducation le plus élevé de chaque étudiant), ni d'informations auxiliaires de niveau hebdomadaire qui peuvent fournir des informations non triviales sur le comportement des utilisateurs.

3. Papier 3 : EPARS Early Prediction of At-risk Students with Online and Offline Learning Behaviors

3.1 Finalité du papier 3

Le sujet de ce papier de recherche porte sur la prédiction du comportement d'apprentissage des étudiants à risque (STAR) afin d'intervenir à temps en cas d'abandon scolaire. Les auteurs de ce papier proposent l'algorithme EPARS afin de prédire le risque d'abandon au cours d'un semestre en modélisant les comportements d'apprentissage, qu'ils soient en ligne ou hors ligne (via les registres d'enregistrement de la bibliothèque).

Dans ce papier, on note deux principales observations. Contrairement aux bons étudiants, les étudiants à risque ne disposent pas d'une routine d'étude régulière et claire. Les auteurs ont mis au point une méthode multi-échelle de sac de régularité pour extraire la régularité des comportements d'apprentissage. Deuxièmement, les amis d'étudiants à risque sont plus susceptibles d'être également à risque. Pour cette raison, les auteurs ont également construit un réseau de cooccurrence pour approcher le réseau social sous-jacent et encoder l'homophilie sociale comme des caractéristiques par l'intégration du réseau.

Les auteurs ont mené une expérience approfondie sur un ensemble de données à grande échelle couvrant 15 503 étudiants. Les traces d'apprentissage en ligne proviennent de la façon dont les étudiants utilisent le Blackboard, une plateforme en ligne d'apprentissage. Ainsi, des données de parcours ont été recueillies avec les horodatages de certains des modules les plus populaires, y compris la connexion, la déconnexion, l'accès au matériel de cours, les devoirs, le forum de discussion, etc.

A travers de cette expérience, les auteurs ont été confrontés à trois grands défis: D'une part, le nombre d'étudiants à risque est nettement inférieur à celui des élèves normaux, ce qui crée un déséquilibre. Le classificateur sera facilement dominé par la classe majoritaire (les élèves normaux). D'autre part, il existe un déséquilibre de la densité des données. Les enregistrements de la bibliothèque sont beaucoup plus rares que les traces de clics sur la plate-forme d'apprentissage en ligne, de sorte qu'il est difficile de les fusionner. Enfin, les auteurs notent une insuffisance des données. Les étudiants, en particulier à risque, sont généralement inactifs au début d'un semestre. Par conséquent, les traces de comportement sont loin d'être suffisantes pour une prédiction d'abandon.

Pour découvrir des caractéristiques statistiques significatives, les auteurs effectuent un test ANOVA afin de déterminer quels comportements sont statistiquement significatifs pour distinguer les élèves à risque des élèves normaux.

3.2. Compréhension des formules mathématiques du papier 3

$$\max_{f} \sum_{u \in V} log(\prod_{v_i \in N_s(u)} \frac{exp(f(u).f(v_i))}{\sum_{v \in V} exp(f(u).f(v))}) \tag{4}$$

Cette équation est utilisée pour caractériser l'homophilie, pour se faire, ils commencent par parcourir le voisinage de chaque noeud qui est censé représenté les personnes avec qui le noeud est le plus proche. Ce parcours de noeud est caractérisé par la fonction $p(c_i = u|c_{i-1} = v) = \frac{\alpha_{pq} w_{uv}}{Z}$ lorsque les deux étudiants sont allés à la bibliotèque en même temps et 0 sinon.Dans cette formule Z est une constante de normalisation et α_{pq} est un coefficient qui varie selon la distance la plus courte entre les deux noeuds.

Si l'on revient sur l'équation du début, le fait d'avoir parcouru le voisinage des noeuds permettent donner un ensemble qui est appelé $N_s(u)$. Les fonctions f(.) sont des fonctions de représentation des noeuds en question.

La méthode d'apprentissage utilisée se base sur la méthode maximum de ressemblance (loglikelihood) que celle utilisée le papier "Dropout Prediction over Weeks in MOOCs via Interpretable Multi-Layer Representation Learning" pour l'apprentissage de cooccurrence.

3.3 Résultat obtenu du papier 3

Dans cet article, les auteurs réalisent l'expérience EPARS, un algorithme permettant d'extraire les modèles de régularité d'apprentissage et d'homophilie sociale des élèves à partir des comportements d'apprentissage en ligne et hors ligne pour prédire le décrochage. Les résultats expérimentaux indiquent que cet algorithme améliore la précision des bases de référence de 14,62% $\backsim$ 38,22% et 5,77% $\backsim$ 34,14% de prédire le risque d'abandon.

3.4 Commentaire sur le papier 3

L'illustration de cas concrets et de graphiques, ainsi que l'explication du déroulement de l'expérience ont permis une bonne compréhension de ce papier de recherche.

4. Ressemblances et différences entre les papiers

Premier papier : What is a Good Prediction- Issues in Evaluating General Value Functions Through Error

Sur la forme du papier, il n'y a pas vraiment de raisons de la scinder en 2 colonnes. Certaines notions ont été expliquées un peu tardivement dans le papier comme le RUPEE, même si on arrive à avoir l'idée de ce que ça peut représenter, il est dommage de l'introduire si tôt dans le papier sans une légère explication.

Le papier se veut didactique avec son exemple, qui est plutôt adapté au sujet du papier, et on arrive à comprendre la plupart des problématiques que les auteurs ont sur certains sujets liés à la prédiction. Dans un sens ce papier est l'un des plus original car il a plus pour but de créer un débat autour des problématiques liées à la prédiction plutôt qu'à montrer une solution à travers un ou des exemples. Cette tendance est bien marquée car seule la dernière partie du papier traite de leur solution.

Le papier n'a pas été facile à lire, comme dit précédemment, certaines notions ne sont pas claires et ce manque d'informations rend la recherche plus difficile à reproduire.

L'intérêt du papier lui ne fait pas de doutes, il est important de comprendre les limites des vérifications, et cela devrait être connu des personnes travaillant sur des sujets similaires dans le but d'améliorer la méthodologie autour de la prédiction.

Deuxième papier : Dropout Prediction over Weeks in MOOCs via Interpretable Multi-Layer Representation Learning

Le second papier a été simple à lire, ce qui peut traduire la qualité et un aspect didactique plus présent. La pertinence du choix de la méthode utilisée dans le papier est bien amenée et expliquée, notamment avec la multitude de travaux similaires sur le sujet.

Toutes les notions sont explicitées, et surtout les formules et les données utilisées sont détaillées ce qui facilite la reproductibilité de la recherche.

Les résultats sont intéressants mais comme dit par les auteurs, la méthodologie sur laquelle ils ont choisie de partir n'a pas eu le succès attendu. De notre point de vue, et cette analyse est partagée dans leur critique de leurs résultats, il est difficile de généraliser les méthodes d'apprentissages de tous les utilisateurs à travers leur suite de cliques. L'analyse et leurs points de vue sur leurs résultats est à relever, et les améliorations qu'ils proposent à leur méthode montrent la qualité de leur travail sur le sujet. On peut facilement imaginer que leur méthode peut être très intéressante pour les plateformes de formation en ligne dans la rétention des utilisateurs.

Troisième papier : EPARS Early Prediction of At-risk Students with Online and Offline Learning Behaviors

Sur la forme, le dernier papier est différent des deux premiers en optant de ne pas scinder le papier en deux colonnes. Cela rend le papier plus aéré, et peut faciliter la lecture pour certains.

Sur le fond, il a la particularité d'annoncer le résultat du papier à la suite de l'introduction et de procéder dans un second temps à l'explication et la méthodologie appliquée. Sur ce point, il se démarque par son originalité des autres, qui avait plus un enchaînement que nous pouvons qualifier de classique. Les avis peuvent diverger sur le format mais il semble intéressant d'avoir fait ce choix, car les lecteurs peuvent dans un sens mieux comprendre la suite du papier en ayant le résultat en tête.

Comme pour le second papier, la pertinence de la méthode est montrée et on imagine facilement les problématiques rencontrées lors de l'application de leur méthode. En effet, il n'y a pas que le format qui est bien structuré, les explications le sont aussi, toutes les difficultés sont abordées avec leur solution, et la critique de la solution proposée.

A travers les résultats on voit bien que la méthode arrive bien à reconnaître les élèves en difficulté et qu'elle arrive bien à les reconnaître plus rapidement. Mais tout n'a pas fonctionné comme ils le pensaient, par exemple l'hypothèse disant que les personnes restent plus souvent avec des gens de leur niveau a été plus efficace pour trouver des étudiants « normaux » plutôt que ceux en difficulté. Cependant des travaux comme celui-ci peuvent être très intéressants dans le domaine de l'éducation. Le nombre de personnes arrêtant leur formation postbac en cours d'année, ou avant l'obtention d'un diplôme est très élevé. L'importance de les aider à rester à flot est essentielle pour tous les établissements.

Bilan:

En ce qui concerne l'aspect visuel d'un papier de recherche, il n'y a pas énormément de possibilité, dans les deux premiers papiers, les auteurs ont préféré avoir un aspect compact alors que le dernier a opté sur quelque chose de plus aéré. L'aspect visuel n'a pas réellement d'impact sur l'agréabilité de la lecture, en effet, les deux derniers papiers ont été tout autant agréables à lire.

En termes de qualité et pertinence des méthodes employées, le premier est moins qualitatif que les deux autres. Cependant, cela peut être dû à l'objectif du papier qui est de montrer qu'une méthode de vérification de prédiction peut être inefficace, et non pas de mettre en avant une méthode pour y remédier. Dans ce sens, on pourrait dire que le papier est original. A contrario les deux autres papiers sont « classiques » et ont pour but de montrer l'efficacité de leur solution à une problématique. On voit que le choix de leur méthode s'appuie sur des travaux similaires. De plus, la méthodologie est bien expliquée, les étapes et les problématiques sont abordés ce qui rend leur recherche reproductible. Un point important, les auteurs du dernier papier ont choisis de mettre les résultats de leur méthode juste après l'introduction, ce qui n'est pas le cas des deux autres.

Le fait d'avoir une méthodologie sur laquelle s'appuyer pour la compréhension du papier est importante, c'est pourquoi le premier apparaît en deçà des autres. Et cela va de même pour la lisibilité, et son aspect didactique moins marqué, en effet les formules sont moins bien expliquées, il n'y aucune grandeur sur laquelle s'appuyer pour reproduire les exemples et ce qui est dommage c'est que le papier passe assez vite sur leur algorithme qu'il propose pour améliorer celui existant. Même si on peut comprendre les auteurs de faire le choix de montrer ce qu'il ne va pas pour mettre en avant leur solution.

Pour ce qui est des résultats, les deux derniers papiers peuvent être très intéressants dans deux sous domaine de l'éducation, les formations en ligne et les établissements physiques. Ces résultats sont d'autant plus intéressants car ils concernent le taux d'abandons des élèves. Le premier papier lui met en avant des résultats qui peuvent être utiles pour la recherche afin de développer de meilleure solution pour vérifier la qualité d'une prédiction.

Même si d'apparence on voit beaucoup de différences, les 3 papiers restent similaires les uns des autres dans leur contenu. Il y a toujours le contexte qui prend une part plus ou moins importante avec les problématiques rencontrées et leurs solutions si existantes. La présentation des formules ou outils utilisés, le domaine d'étude et l'explication de son contenu. Les résultats de l'expérience avec leur critique et axe d'amélioration. En prenant en compte tous ces facteurs, le second papier sort du lot notamment pour son autocritique en conclusion accompagné de plusieurs axes d'améliorations.

5. Conclusion

Évoquée en introduction, la prédiction joue et va jouer un grand rôle dans le futur que ce soit pour les travaux de recherches ou pour les entreprises directement. La lecture de ces papiers ont permis de montrer la complexité de prédire un événement et comment il est facile de biaiser une prédiction. A travers le premier papier, les auteurs veulent nous montrer les limites des méthodes de vérification des prédictions, et dans les deux suivants nous avons pu voir des méthodes de prédictions appliquées à des cas concrets. La tendance de ces papiers est de dire que les méthodes de prédictions sont encore loin d'être idéales, mais l'importance de pouvoir prédire des actions permettraient d'énormes progrès dans des multitudes de domaines. Elle est d'ailleurs visible dans les deux derniers papiers dans le domaine de l'éducation, domaine dans lequel nous avons très peu fait évoluer nos méthodologies d'enseignement au cours des années, et où l'on commence à avoir un réel retard par rapport à toutes les nouvelles technologies telle que l'IA.

Evidemment, prédire des actions comme le décrochage scolaire, ou la réussite des étudiants dans la vie professionnelle n'est pas une chose aisée. En effet, pour l'instant les algorithmes de prédictions en place ne peuvent pas prendre une infinité de paramètres, ce qui implique que l'ingénieur fasse des choix qui peuvent biaiser les résultats. Pour arriver à des prédictions de meilleures qualités il faudra donc pallier cette problématique afin de pouvoir prendre en compte des aspects comme le milieu social et économique des étudiants. Ici, nous avons beaucoup parlé de l'éducation mais la prédiction s'applique à tous les domaines ce qui montre à quel point la recherche à ce sujet est vaste et pleine de possibilités.

Auto-critique de notre travail sur l'analyse de 3 papiers :

Nous avons tenté au mieux de comprendre et de retranscrire ce que les chercheurs ont voulu exprimer dans leur papier. Du fait de notre parcours scolaire et expériences professionnelles, nous avons eu quelques difficultés à comprendre les formules mathématiques (notamment pour le papier 1 - What is a Good Prediction- Issues in Evaluating General Value Functions Through Error), c'est pour cela que les explications mathématiques peuvent sembler légères pour le lecteur. Cependant nous avons essayé le mieux possible de faire comprendre le message des chercheurs.

6. Annexes

Papier 1 : What is a Good Prediction- Issues in Evaluating General Value Functions Through Error (https://arxiv.org/pdf/2001.08823.pdf)

Papier 2 : Dropout Prediction over Weeks in MOOCs via Interpretable Multi-Layer Representation Learning (https://arxiv.org/pdf/2002.01598.pdf)

Papier 3 : EPARS Early Prediction of At-risk Students with Online and Offline Learning Behaviors (https://arxiv.org/pdf/2006.03857.pdf)