



RÉSUMÉ DE THÈSE

MOHAMED MOUAICI

AIDE AU SUIVI DES APPRENANTS DANS LE CONTEXTE DE LA FORMATION
PROFESSIONNELLE EN LIGNE



Résumé

La formation professionnelle est un levier important permettant aux individus d'acquérir le savoir et le savoir-faire nécessaires à l'exercice d'un métier ou d'une activité professionnelle. Avec l'essor des technologies éducatives, notamment les environnements numériques d'apprentissage, la formation professionnelle se détache du modèle pédagogique classique en présentiel et permet d'envisager des formations à grande échelle. Un des avantages de ces environnements est qu'ils permettent de collecter des ensembles de données riches : les traces des activités de l'apprenant et notamment de ses interactions avec les contenus pédagogiques. Ces ensembles de données peuvent être analysés dans le but de comprendre et d'optimiser l'apprentissage et son contexte.

Ce travail de recherche s'intéresse plus particulièrement aux formations professionnelles en ligne en France. Il est réalisé dans le cadre d'un partenariat entre le laboratoire SYMME de l'université Savoie Mont Blanc et l'entreprise Logipro représentée par l'équipe pédagogique de sa filiale Campus26. Cette équipe souhaite disposer d'un tableau de bord d'aide au suivi lui permettant d'appréhender le comportement des apprenants. Une approche centrée sur l'utilisateur (UCD) qui implique l'équipe pédagogique dans toutes les phases du projet, a permis d'identifier trois objectifs de suivi : la détection le plus tôt possible des apprenants en difficulté, le suivi de l'engagement des apprenants et le suivi de leur progression. Ce travail de recherche traite le premier objectif considéré comme prioritaire par l'équipe pédagogique.

Les résultats initiaux des expérimentations de plusieurs modèles prédictifs sur les données d'un scénario d'apprentissage libre caractérisé par l'absence de la notion d'échec et la courte durée des activités *e-learning* ont permis de démontrer que le modèle des forêts aléatoires était le meilleur mais que ses performances prédictives devaient être améliorées pour atteindre un seuil minimal satisfaisant pour l'équipe pédagogique qui correspond à un score $F1 \geq 70\%$.

Ainsi, la contribution de ce travail porte en tout premier lieu sur la proposition d'une approche permettant d'améliorer la détection, le plus tôt possible, des apprenants en difficulté. Cette approche est composée de cinq étapes séquentielles qui ont comme objectif d'identifier les problèmes qui peuvent impacter négativement les prédictions puis de proposer des solutions à ces problèmes. Ces problèmes peuvent être liés aux données étudiées ou au modèle prédictif.

Les cinq étapes de l'approche se résument ainsi :

La première étape de l'approche proposée consiste à analyser les corrélations entre les indicateurs de suivi et les apprenants en difficulté pour s'assurer de la nécessité du modèle prédictif dans le cas étudié et pour éliminer les problèmes de colinéarité.

La deuxième étape se focalise sur le traitement des données déséquilibrées en proposant un mécanisme de perturbation permettant de renforcer le poids des observations les plus difficiles à classer et de réduire le problème de sur-apprentissage qui peut être causé par les différentes techniques de sur-échantillonnage. Cette étape permet également d'identifier automatiquement la technique de sur-échantillonnage la plus adaptée à l'ensemble de données étudié et le taux de perturbation à appliquer sur ces données.

La troisième étape vise l'optimisation des hyper-paramètres du modèle prédictif en proposant une solution pour identifier les hyper-paramètres les plus impactants afin de réduire la complexité des méthodes d'optimisation et par conséquent réduire leur temps d'exécution.

La quatrième étape propose une méthode pour analyser les raisons des classifications erronées du modèle prédictif. Cette méthode se base sur les règles de classification construites avec les indicateurs les plus importants pour le modèle afin de permettre à l'utilisateur de comprendre les raisons principales de ces classifications erronées. Cela peut l'aider à rajouter ou supprimer

des indicateurs de son modèle, à changer la gestion des valeurs manquantes ou encore à normaliser les valeurs des indicateurs.

La dernière étape de l'approche proposée évalue la stabilité du modèle prédictif issu des quatre premières étapes en l'appliquant sur un nouvel ensemble de données inconnu du modèle. Cette étape permet de valider ou d'invalider la pertinence du modèle. Cette étape a aussi servi à valider l'approche elle-même.

L'application de l'approche proposée sur les données du scénario d'apprentissage libre a contribué à une amélioration de 33,42% sur le score F1 le faisant passer de 44,36% (obtenu avec le modèle prédictif initial) à 77,78%. Ensuite, l'approche a été évaluée sur deux autres scénarios d'apprentissage toujours dans le cadre de la formation professionnelle en ligne : un scénario séquentiel avec évaluations obligatoires au niveau de chaque activité *e-learning* et un scénario séquentiel avec évaluations optionnelles au niveau des activités *e-learning*. Les résultats prédictifs obtenus sur ces deux scénarios témoignent de la pertinence de l'approche proposée. En effet, celle-ci a permis de faire passer le score F1 de 66,67% à 74,07% dans le premier cas et de 61,96% à 72,57% dans le deuxième cas.

Pour aller plus loin dans l'évaluation de l'approche proposée, celle-ci a été testée sur les données d'un cours universitaire dispensé dans le cadre d'une pédagogie hybride par un enseignant de l'Université Savoie Mont Blanc (USMB) sur une plate-forme Moodle. Pour ce faire, plusieurs professeurs de l'USMB ont été impliqués dans la démarche d'évaluation. Cela a permis de définir avec eux ce qui permet de dire qu'un étudiant est en difficulté, d'avoir accès aux traces d'apprentissage collectées par Moodle pour ce cours, d'identifier les indicateurs de suivi, de leur présenter les résultats prédictifs obtenus via l'approche proposée et de recueillir leur *feedback*. L'application de l'approche sur les données de ce cours a conduit à une amélioration très remarquable (de 465 %) des résultats prédictifs en aboutissant à un score F1 de 62,07% au lieu de 13,33 % initialement obtenu sans l'application de l'approche.

L'implémentation de cette approche a été réalisée de manière à permettre sa réutilisation dans d'autres contextes tout en donnant la possibilité à l'utilisateur d'intervenir sur chaque étape de celle-ci pour accepter, refuser ou ajuster les préconisations proposées.

Ces résultats prometteurs ont motivé la conception et le développement d'une solution de type tableau de bord d'aide au suivi dans le contexte de la formation professionnelle (le contexte de ce travail de thèse). De ce fait, une autre contribution de ce travail consiste à proposer une architecture à base de composants spécifiques pour le développement de ce tableau de bord. La mise en place de cette architecture s'est concrétisée par le développement d'une première version de tous les composants nécessaires, notamment le composant qui permet de récupérer les traces d'apprentissage et de les transformer en déclarations xAPI (eXperience API), le composant LRS (*Learning Record Store*) responsable du stockage de ces déclarations, et le tableau de bord lui-même qui utilise ces déclarations pour calculer les indicateurs et les afficher à l'équipe pédagogique. Ce tableau de bord permet non seulement d'afficher le plus tôt possible les notifications de la détection des apprenants en difficulté mais aussi d'accompagner ces prédictions des indicateurs les plus importants et des explications textuelles basées sur ces indicateurs pour présenter à l'équipe pédagogique les explications de chaque prédiction. Ce choix d'architecture permet d'utiliser le tableau de bord quel que soit l'environnement numérique d'apprentissage utilisé, sous réserve de développer le composant d'acquisition des traces. Ce tableau de bord a été évalué par l'équipe pédagogique de Campus26. Si les retours montrent indubitablement l'intérêt des aides au suivi proposées, ils donnent aussi des pistes d'amélioration concernant notamment la présentation des explications textuelles des prédictions et la prise en compte des *feedbacks* de l'équipe pédagogique dans la construction du modèle prédictif.

Aujourd'hui, d'autres pistes d'amélioration du modèle prédictif sont en cours d'étude et les expérimentations du tableau de bord à grande échelle se poursuivent.

Mots-clés : analyse de l'apprentissage, apprenants en difficulté, prédiction le plus tôt possible, traces, indicateurs, apprentissage automatique supervisé, perturbation, sur-échantillonnage, forêts aléatoires, hyper-paramètres.