Évaluation des apprentissages, vers une approche sensible aux données pour estimer la difficulté perçue des items d'évaluation

Université de Picardie Jules Verne- Laboratoire de Modélisation, Information & Systèmes (MIS)

Doctorant : LAMGARRAJ Mohamed mohamed.lamgarraj@u-picardie.fr

Contexte

- Utilisation d'outils numériques pour l'évaluation des étudiants en enseignement superieur :
  - Automatisation
  - Objectivité
- Critères à prendre en compte dans la génération des tests d'évaluation :
  - Echantillons différenciés
  - La difficulté des tests générés

Problématique

- Comment peut-on modéliser et intégrer la caractérisation de la difficulté d'une activité dans les algorithmes de génération de tests d'évaluation ?
- Est-il possible de construire un modèle rétrospectif, basé sur l'apprentissage artificiel, qui caractérise la difficulté à partir de corpus de traces d'évaluations ?
- Comment caractériser la difficulté perçue et la distinguer de la difficulté prédite ?

Objectifs

- Construire un modèle adaptatif pour la génération des échantillons de sujets différenciés, mais de difficultés équivalentes.
- Caractériser la difficulté perçue d'une tache ou activité dans la machine, prenant en compte la dimension psychologique de la difficulté.
- Construire des algorithmes de génération sous contraintes prenant en compte la difficulté des questions générées.

Méthodologie

Acquisition des données

- Etude et sélection des bases de données de la littérature utilisées dans le domaine des EIAH
- Collection des traces à partir des LMS utilisé à UPJV
- Mesure de la perception (Eye-Tracking, mouse-tracking ….)

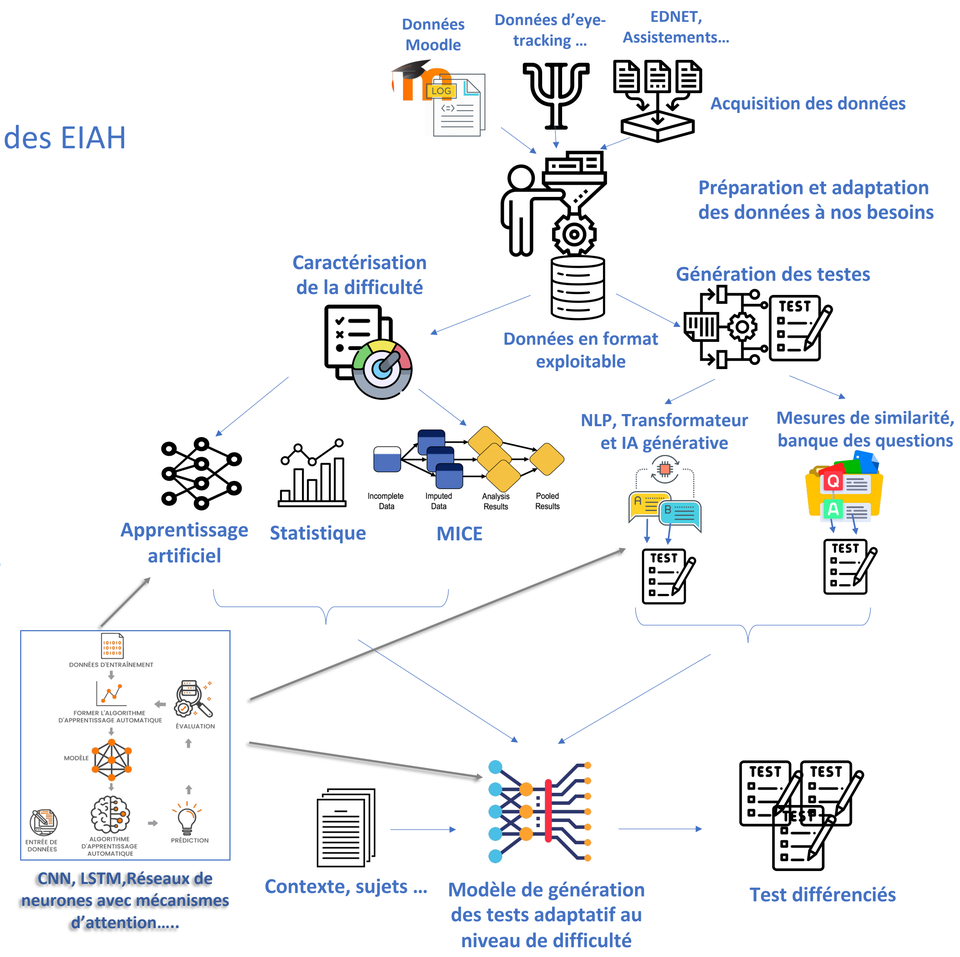
Modélisation

- Caractérisation de la difficulté:
  - Construction d'un model pour la prédiction de la difficulté
  - Identification des facteurs psychiques qui caractérisent la difficulté
- Proposition d'algorithme de génération automatique des tests
  - Algorithme de génèration des tests différenciés de même niveaux de difficulté.
  - Un modèle Adaptatif pour générer des tests d'une niveau de difficulté personalisé.

Configuration d'un pipeline expérimental

Test et validation des résultats

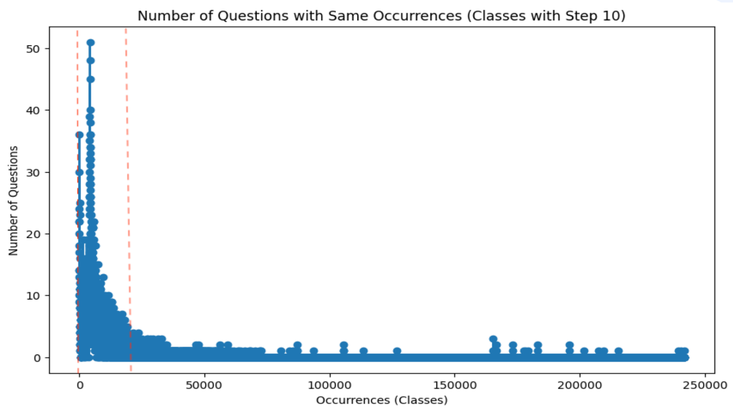
Hébergement de la solution dans le LMS utilisée par l'UPJV



Données

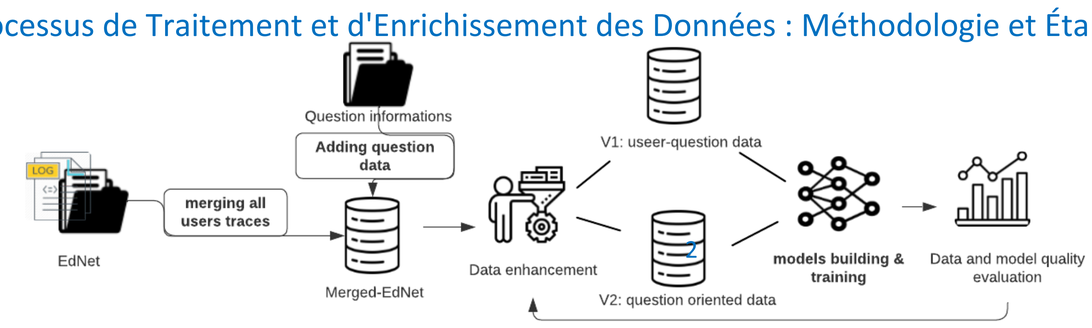
EdNet

- 297,915 → 784,309 Etudiants
- 13,169 Questions
- 131,441,538 Interactions



Répartition des questions par occurrence dans EdNet (les lignes rouges délimitent la zone d'occurrences des données utilisées)

Processus de Traitement et d'Enrichissement des Données : Méthodologie et Étapes



1: Jeu de données utilisateur-question: 35 colonnes ; 79 929 968 entrées. Intègre des attributs liés à l'apprenant

2: Jeu de données orienté questions: 13 170 entrées représentant chaque question des tests ; 27 colonnes issues de l'analyse de 8 687 observations en moyenne par question.

Expérimentations

- 4 modèles d'apprentissage automatique (régression logistique, arbre de décision, forêt aléatoire et XGBoost) sur 80% des données, les 20% restants pour validation. Taux d'exactitude des réponses fournies pour établir des étiquettes de vérité terrain dans nos
- tâches d'apprentissage supervisé

Classes de difficulté utilisé pour l'entraînement des modèles

Difficulty	< 0.3	$0.3 < \text{diff} < 0.6$	$0.6 <$
Class	Easy (0)	Medium(1)	Difficult(2)

Performance des modèles sur les données orientées questions

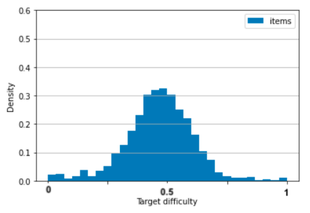
model	LR	DT	RF	XGB
Accuracy	0.54	0.91	0.91	0.93
f1 score	0.50	0.85	0.81	0.82

- Données orienté utilisateurs-questions : 40 % des données choisies autour de la moyenne de 8687 réponses par question, est utilisé.

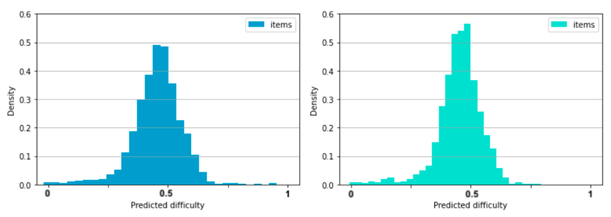
La performance des modèles sur les données orientées questions- utilisateurs.

model	DT	RF	XGB
Accuracy	0.80	0.89	0.74
f1 score	0.65	0.70	0.71

- Impact de l'enrichissement : L'ajout de nouvelles fonctionnalités améliore significativement la précision et les scores F1 des modèles.
- Composition des données : modèles plus performants avec des données orientées questions qu'avec des données orientées utilisateur-question.
- Analyse des performances : valeurs de difficulté prédites correspondent bien aux niveaux réels, montrant l'efficacité de l'étiquetage par classes



(a) Densité des questions par valeurs de difficulté de vérité terrain



(b) Densité des questions par valeurs de difficulté prédit pour les données orientées question

(c) Densité des questions par valeurs de difficulté prédit pour les données question-utilisateur

Futurs Développements:

- Identification expérimentale de facteurs humains influençant la perception de la difficulté auprès d'étudiants en informatique et psychologie.
- Création d'un estimateur de difficulté et intégrer cette prédiction dans les algorithmes de génération automatique de tests à difficulté ciblée.

Références

- Choi, Youngduck, et al. "Ednet: A large-scale hierarchical dataset in education." Artificial Intelligence in Education: 21st International Conference, AIED 2020, Ifrane, Morocco, July 6–10, 2020, Proceedings, Part II 21. Springer International Publishing, 2020.
- Attali, Y., Saldivia, L., Jackson, C., Schuppan, F., Wanamaker, W. (2014). Esti- mating item difficulty with comparative judgments. ETS Research Report Series, 2014(2), 1-8.
- Beck, J., Stern, M., Woolf, B. P. (1997). Using the student model to control problem difficulty. In User Modeling : Proceedings of the Sixth International Conference UM97 Chia Laguna, 1997 (pp. 277-288). Springer.
- Le An Ha, Victoria Yaneva, Peter Baldwin, and Janet Mee. 2019. Predicting the Difficulty of Multiple Choice Questions in a High-stakes Medical Exam. In Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications, pages 11–20, Florence, Italy. Association for Computational Linguistics.
- Pelánek, R., Effenberger, T., & Čechák, J. (2021). Complexity and Difficulty of Items in Learning Systems. International Journal of Artificial Intelligence in Education,32,196-232.

ANR-21-DMES-0002

RJC-EIAH-2024-LAVAL