Dossier technique

Objectif

L’objectif du projet est de développer un modèle permettant de prédire les notes d’étudiants à partir des logs d’activité sur la plateforme Arche. Le projet est donc séparé en deux phases : (i) la création d’un modèle permettant de prédire les notes, (ii) la création d’une interface utilisateur permettant à un enseignant de rentrer un fichier de logs et d’obtenir les notes prédites par le modèle sélectionné en (i).

Construction du modèle

Le logiciel de construction du modèle se base sur un ensemble de script python définissant des fonctions réalisant les actions suivantes :

1. Lecture des données (read\_files.py) : lecture des fichier logs et notes à partir d’un PATH défini par le développeur. Les deux fichiers sont harmonisés (fonctions filter\_\*) : les entrées absente de l’un des fichiers sont effacées (pseudo associé à un professeur …). Enfin des variables utilisées pour créer les features sont ajouté au tableau des logs (fonction split\_columns).
2. Fonction de création des features (features\_creation.py) : à partir d’un fichier de logs des features potentiellement explicatives des notes obtenues sont créés. 20 features sont créés chacune avec une fonction dédiée prenant toujours en input le fichier de logs obtenu en sortie de read\_files.py et avec comme output un DataFrame avec le pseudo et la features créée.
3. Etape de preprocessing (preprocessing.py) : la fonction centrale de cette étape est la création d’un DataFrame rassemblant pour chaque pseudo les 20 features creée (fonction creer\_df). Ensuite un ensemble de fonctoin permettent de nettoyer ce jeu de donnée brut : séparation en jeu train et test, drop des colonnes parfaitement corrélées (enlever\_correlation\_complets), OneHot encodage des variables qualitatives (encodage), MinMax scaling des variables (scaling), ajout de tirets aux noms de colonnes avec des espaces (df\_transformer), et enfin une fonction garantissant que le DataFrame obtenu en fin d’étape de préprocessing à les colonnes attendus par le modèle (align\_columns).
4. Etape de construction du modèle

Inputs