Plateforme de Prédiction du Risque d'Échec Étudiant

Learning Analytics & Machine Learning

Réalisé par :

El Guelta Mohamed-Saber
El Hadifi Soukaina
Ibnchakroune Houssam
Kamal Salma

Filière : Ingénierie des Données — ENSAH

Encadré par : **Pr. Ouald Chaib Sara**

Année universitaire 2025/2026

Résumé

Résumé — Ce projet développe un système intelligent de prédiction du risque d'échec étudiant dans l'enseignement à distance, exploitant les données du dataset OULAD (32,593 étudiants, 10M+ interactions). La solution comprend un pipeline ETL automatisé, un modèle Random Forest optimisé (26 features), et un dashboard Streamlit pour l'aide à la décision pédagogique. Les performances obtenues sont exceptionnelles : 94.7% d'accuracy et 100% de recall sur la détection des échecs, permettant une intervention précoce dès le 180ème jour du cours. Ce travail démontre l'applicabilité du Machine Learning dans le contexte éducatif et la valeur de l'approche data-driven pour améliorer la réussite étudiante.

Table des matières

Chapitre 1

Introduction

1.1 Contexte

L'enseignement à distance génère quotidiennement des volumes massifs de données d'interaction (connexions, clics, évaluations) qui restent sous-exploitées. Entre 15% et 25% des étudiants échouent dans ce contexte, un taux significativement plus élevé que dans l'enseignement traditionnel. Le problème majeur réside dans la détection tardive : lorsqu'un enseignant constate l'échec, il est généralement trop tard pour intervenir.

1.2 Problématique

Comment exploiter les données d'interaction des étudiants avec les plateformes d'apprentissage en ligne pour prédire précocement le risque d'échec et permettre une intervention pédagogique ciblée ?

1.3 Objectifs

Ce projet vise à développer un pipeline Machine Learning complet comprenant :

— Extraction et transformation des données (ETL)
— Engineering de 26 features prédictives
— Modèle Random Forest avec validation rigoureuse
— Dashboard Streamlit interactif pour visualisation et aide à la décision
— Automatisation et industrialisation (modes dev/prod)

Chapitre 2

Données et Méthodologie

2.1 Dataset OULAD

Le dataset Open University Learning Analytics contient les données de 32,593 étudiants inscrits à 22 modules entre 2013-2014, comprenant :

— **10,655,280 interactions VLE** (clics sur ressources pédagogiques)
— **173,912 soumissions** d'évaluations avec scores
— **Données démographiques** : âge, genre, région, niveau d'éducation
— **Historique académique** : crédits étudiés, tentatives précédentes

Les données sont réparties en 7 fichiers CSV inter-reliés nécessitant fusion et agrégation.

2.2 Architecture du système

Le système adopte une architecture en couches :

1. **Couche Données** : PostgreSQL + NumPy Arrays
2. **Couche Métier** : ETL, Feature Engineering, ML Engine
3. **Couche Orchestration** : Pipeline CLI (modes dev/prod)
4. **Couche Présentation** : Dashboard Streamlit

2.3 Stack technologique

— **Python 3.12** : pandas 2.2, NumPy 1.26, scikit-learn 1.4
— **PostgreSQL 14** : Stockage features et prédictions
— **Streamlit** : Dashboard web interactif
— **Packaging moderne** : pyproject.toml (PEP 621)

Chapitre 3

Feature Engineering

3.1 Catégories de features (26 au total)

Features Académiques (7) : Score moyen, écart-type des scores, taux de soumission à temps, délai moyen de soumission, nombre d'évaluations.

Features Comportementales VLE (12) : Total clics, jours actifs, clics par type de ressource (quiz, forum, documents, etc.), moyenne clics/jour.

Features Temporelles (3) : Engagement précoce (jours 0-30), engagement mi-parcours (jours 31-90), tendance d'évolution (pente régression linéaire).

Features Démographiques (4) : Âge encodé, niveau d'éducation, nombre de tentatives précédentes, crédits étudiés.

3.2 Insights clés

L'analyse de corrélation révèle que :

— Le **score moyen** est le prédicteur dominant (25.6% d'importance)
— L'**engagement quiz** compte pour 8.9% (auto-évaluation)
— Les **activités forum** (7.5%) et **homepage** (7.5%) sont significatives
— Les **ressources pédagogiques** contribuent à 7.1%

Chapitre 4

Modélisation Machine Learning

4.1 Choix de l'algorithme

Random Forest a été sélectionné après comparaison avec Logistic Regression, SVM et Gradient Boosting, pour ses avantages :

— Performance supérieure (94.7% vs 88% pour Logistic Regression)
— Robustesse aux outliers et class imbalance
— Interprétabilité via importance des features
— Pas de preprocessing complexe requis

4.2 Hyperparamètres optimisés

Grid Search avec cross-validation 5-fold a identifié les paramètres optimaux :

— `n_estimators=200` : Nombre d'arbres
— `max_depth=15` : Profondeur maximale (évite surapprentissage)
— `min_samples_leaf=5` : Régularisation
— `class_weight='balanced'` : Ajustement du déséquilibre Pass/Fail (75%/25%)

4.3 Validation

Stratégie : Train/Test split stratifié + Cross-validation 5-fold
Résultats CV :

— Accuracy CV moyenne : 78.65% ± 0.66%
— Accuracy test final : **94.7%**
— Recall Fail : **100%** (objectif prioritaire atteint)
— Precision Fail : 93%
— AUC-ROC : 0.996

Chapitre 5

Résultats et Évaluation

5.1 Performance sur test set (1,000 étudiants)

Métrique	Classe Pass	Classe Fail
Precision	99%	93%
Recall	82%	**100%**
F1-Score	90%	96%
Support	281	719

TABLE 5.1 – Performances détaillées par classe

Accuracy globale : 94.7% (947/1,000 prédictions correctes)
AUC-ROC : 0.996 (discrimination quasi-parfaite)

5.2 Matrice de confusion

	Prédit Pass	Prédit Fail
Réel Pass	230	51
Réel Fail	0	719

TABLE 5.2 – Matrice de confusion - Test Set

Interprétation : **Aucun échec manqué** (0% de faux négatifs), permettant de détecter 100% des étudiants en difficulté. Les 51 faux positifs (18%) représentent des étudiants identifiés à tort comme à risque, mais recevoir un accompagnement supplémentaire n'est jamais nuisible.

5.3 Impact du cutoff day

Le modèle est entraîné sur données jusqu'au jour 180 (fin de cours), représentant une prédiction finale robuste :

— **Performance maximale** : Recall 100% sur la détection des échecs
— **Fiabilité** : AUC 99.6%, precision 93%
— **Utilisation** : Validation finale avant décision administrative

Chapitre 6

Dashboard Streamlit et Aide à la Décision

6.1 Architecture de visualisation

Le pipeline exporte les prédictions vers PostgreSQL, accessible via une application web Streamlit interactive comprenant 4 sections :

Section 1 - Vue d'ensemble : KPIs globaux (nombre d'étudiants, taux de risque), distribution Pass/Fail, graphiques interactifs Plotly.

Section 2 - Liste étudiants à risque : Table filtrable dynamique (par module, niveau de risque, score), export CSV, navigation vers profil détaillé.

Section 3 - Profil individuel : Scores vs moyenne classe, engagement VLE détaillé, recommandations d'actions personnalisées.

Section 4 - Analyse modèle : Matrice de confusion interactive, feature importance, courbes ROC/Precision-Recall.

6.2 Catégorisation du risque

Probabilités d'échec converties en 3 niveaux :

— **Low risk** : $<30\%$ (pas d'action immédiate)
— **Medium risk** : 30-70% (monitoring renforcé)
— **High risk** : $>70\%$ (intervention urgente)

6.3 Cas d'usage

Enseignant : Accède au dashboard Streamlit, filtre les étudiants High risk de son module, exporte la liste pour suivi personnalisé.

Tutrice : Utilise les profils individuels pour identifier les lacunes spécifiques (faible engagement quiz, absence forum) et adapter l'accompagnement.

Administratrice : Analyse l'évolution des métriques via graphiques interactifs, identifie les modules à fort taux de risque pour allocation de ressources.

Chapitre 7

Conclusion

7.1 Contributions

Ce projet démontre la faisabilité d'un système intelligent de prédiction du risque d'échec avec performance exceptionnelle (100% recall) dépassant largement les standards académiques (85-90%). L'approche holistique intègre ETL automatisé, feature engineering méthodique (26 features), modélisation rigoureuse Random Forest, et dashboard Streamlit interactif.

7.2 Impact attendu

Le déploiement institutionnel permettrait de :

— Détecter 100% des étudiants en échec (aucun faux négatif)
— Optimiser le temps enseignant via dashboard web accessible 24/7
— Intervenir avant la finalisation des résultats
— Piloter la stratégie pédagogique de manière data-driven via analytics Streamlit

7.3 Limites

Vigilance nécessaire sur :

— Corrélation vs causalité (engagement faible corrèle avec échec, mais n'est pas forcément la cause)
— Risque de stigmatisation (étudiant étiqueté "High risk")
— Généralisation à d'autres contextes (validation locale requise)
— Conformité RGPD (consentement, droit d'accès, sécurisation)

7.4 Perspectives

— **Prédiction précoce** : Modèle entraîné au jour 90 pour intervention mi-semestre
— **Modèles avancés** : LSTM pour séquences temporelles, Gradient Boosting
— **Features enrichies** : NLP sur forums, analyse sentiment, données textuelles
— **Dashboard mobile** : Application Streamlit responsive pour accès smartphone
— **API REST** : Intégration avec systèmes LMS (Moodle, Canvas)

La transformation digitale de l'éducation doit rester au service d'une vision pédagogique humaniste : l'algorithme assiste l'humain mais ne le remplace jamais.