# Rapport Final – Prédiction du Risque d’Abandon Scolaire

**Rappel du contexte et de l’objectif**

L’objectif de ce projet était de construire un modèle de **classification supervisée** permettant de prédire si un étudiant risquait d’abandonner ses études. Cette prédiction est essentielle pour les établissements scolaires, car elle permettrait de **mettre en place des actions préventives ciblées**, comme un suivi personnalisé ou des aides pédagogiques.

Le dataset utilisé contient des **données réelles ou simulées** liées à :

* L’âge
* Le sexe
* Le taux de présence (%)
* Le nombre de retards
* La note moyenne sur 20
* La situation familiale
* Et enfin, la variable cible Abandon (1 = a abandonné, 0 = n’a pas abandonné)

**Analyse exploratoire des données**

**📊 Distribution des variables**

L’analyse descriptive a montré que :

* L’âge des étudiants est concentré entre **18 et 23 ans**, ce qui correspond à une population typique d’université ou de fin de lycée.
* Le **taux de présence** est globalement élevé, mais **on observe une baisse significative chez certains élèves abandonnant**.
* Le **nombre de retards** varie beaucoup mais reste généralement entre 0 et 5.
* La **note moyenne** se situe majoritairement entre **11 et 15**, mais les étudiants qui abandonnent ont souvent des moyennes < 11.
* La répartition de la variable Sexe est équilibrée (0 = Homme, 1 = Femme).
* La variable cible Abandon est **fortement déséquilibrée** : environ **97% n’ont pas abandonné**, contre **3% qui ont abandonné**.

Ce déséquilibre rend le problème difficile : un modèle naïf pourrait avoir une accuracy très élevée en prédisant toujours « non-abandon », sans être utile.

Une image contenant texte, capture d’écran, diagramme, Tracé

Le contenu généré par l’IA peut être incorrect.

Une image contenant texte, capture d’écran, diagramme, Tracé

Le contenu généré par l’IA peut être incorrect.

**Corrélation entre les variables**

La **matrice de corrélation (heatmap)** a permis d’identifier :

* Une **corrélation positive forte entre Taux de présence et Note moyenne** (+0.81), ce qui confirme que les étudiants assidus réussissent mieux.
* Une **corrélation négative entre Taux de présence et Abandon** (-0.47), et entre Note moyenne et Abandon (-0.43).
* Une **corrélation positive entre Nombre de retards et Abandon** (+0.44).

💡 **Interprétation :**  
Ce sont donc **trois indicateurs clés** de prédiction de l’abandon :  
**absentéisme**, **baisse de notes**, et **accumulation de retards**.

Une image contenant texte, capture d’écran, diagramme, nombre

Le contenu généré par l’IA peut être incorrect. **Modélisation et évaluation des performances**

Trois modèles classiques de classification supervisée ont été utilisés :

1. **KNN (K Nearest Neighbors)**
2. **Arbre de décision**
3. **Régression logistique**

Après entraînement sur 80% des données et test sur les 20% restantes, les modèles ont été évalués à l’aide des métriques suivantes :

* **Accuracy** (exactitude globale)
* **Recall** (capacité à repérer les abandons)
* **Precision** (fiabilité des prédictions d’abandon)
* **F1-score** (compromis entre recall et précision)

**Validation croisée :**

| **Modèle** | **F1-score moyen (CV 5 folds)** |
| --- | --- |
| KNN | 0.58 |
| Arbre de décision | 0.56 |
| Régression logistique | 0.57 |

**Résultats bruts :**

| **Modèle** | **Accuracy** | **F1-score (classe 1)** | **Recall (classe 1)** | **Précision (classe 1)** |
| --- | --- | --- | --- | --- |
| KNN | 0.95 | 0.63 | 0.53 | 0.76 |
| Arbre de décision | 0.95 | 0.67 | 0.63 | 0.70 |
| Régression logistique | 0.96 | 0.64 | 0.50 | 0.88 |

**Optimisation avec GridSearchCV**

**🔹 KNN :**

* Meilleur K : **11**
* F1 optimisé : **0.63**

**🔹 Arbre :**

* Meilleurs paramètres : max\_depth = 3, min\_samples\_split = 5
* F1 optimisé : **0.66**

👉L’**arbre de décision optimisé est le modèle le plus performant** dans ce contexte.

**Interprétation critique**

**✅ Points forts du modèle :**

* Tous les modèles atteignent une **accuracy supérieure à 0.95**, ce qui montre une bonne généralisation.
* Le modèle **réagit bien à l’optimisation** grâce à GridSearch.
* Les variables les plus influentes sont : Note moyenne, Taux de présence, Nombre de retards.

**⚠️ Limites :**

* **Déséquilibre de classes** : peu d’étudiants abandonnent (≈ 3%), ce qui **baisse le rappel sur la classe minoritaire**.
* Même si l’accuracy est élevée, un **faible rappel** signifie que **plusieurs vrais abandons ne sont pas détectés**.
* Le dataset manque de variables psycho-sociales (stress, isolement, etc.) qui pourraient enrichir la prédiction.

**✅ Recommandations pédagogiques**

1. **Surveiller activement les étudiants ayant :**
   * Un taux de présence < 75%
   * Une note moyenne < 11
   * Plus de 3 retards
2. **Compléter le dataset** avec des variables liées à :
   * Le stress ou la motivation
   * La situation financière
   * L'engagement dans la vie étudiante
3. **Utiliser un modèle équilibré (SMOTE, class\_weight, etc.)** dans les prochaines versions pour améliorer la détection des abandons.

**Conclusion générale**

Ce projet démontre l’intérêt de l’analyse prédictive dans le domaine de l’éducation.  
Grâce à une méthodologie rigoureuse (CRISP-DM), des outils de data science adaptés, et une interprétation critique des résultats, nous avons pu :

* **Identifier les étudiants à risque d’abandon**
* **Construire des modèles fiables et optimisables**
* **Proposer des recommandations concrètes à l’établissement**