# Plan prévisionnel — Preuve de concept

Prédiction du statut des rendez-vous médicaux

**Étudiants :** BENAOUDIA Leticia - LAMARI Azzeddine
**Formation :** Master Informatique
**Date :** 19 janvier 2026

## 1. Contexte

La gestion des rendez-vous médicaux constitue un enjeu majeur pour les établissements de santé. Les non-présentations et les annulations tardives entraînent une sous-utilisation des créneaux, une perte de ressources et une dégradation de la qualité de service.

Dans un contexte de transformation numérique du secteur de la santé, les approches **data-driven** et les techniques de **machine learning** offrent des perspectives intéressantes pour anticiper le comportement des patients. Ce travail s'inscrit dans ce cadre et vise à évaluer la faisabilité d'un modèle prédictif capable d'anticiper le statut final d'un rendez-vous médical.

L'objectif est de proposer une preuve de concept réaliste, fondée sur des données structurées, permettant d'améliorer la planification et la prise de décision opérationnelle.

## 2. Dataset retenu

Le dataset retenu est un jeu de données synthétique simulant le fonctionnement d'un cabinet médical sur une période étendue. Il décrit les rendez-vous médicaux à travers des variables temporelles, organisationnelles et démographiques.

Le fichier principal utilisé pour l'étude est `appointments.csv`, qui regroupe les informations relatives à la planification, au déroulement et à l'issue des rendez-vous. La variable cible est le **statut du rendez-vous**, indiquant s'il a été honoré, annulé ou manqué.

Ce dataset est particulièrement adapté à une approche de modélisation prédictive, car il contient des informations exploitables avant et autour du rendez-vous, directement liées au risque de non-présentation.

## 3. Modèle envisagé

Le modèle principal envisagé pour ce projet est un algorithme d'ensemble de type **Gradient Boosting**, et plus précisément XGBoost.

Ce choix est motivé par plusieurs arguments :

— XGBoost est reconnu pour ses performances élevées sur des données tabulaires ;
— il est capable de modéliser des relations non linéaires et des interactions complexes entre variables ;
— des études récentes montrent son efficacité dans des problématiques de classification déséquilibrée, proches du contexte des non-présentations médicales.

L'objectif de l'algorithme est de prédire le statut final d'un rendez-vous à partir des informations disponibles, afin d'identifier en amont les rendez-vous à risque. Ce type de modèle peut être utilisé comme outil d'aide à la décision pour déclencher des actions préventives telles que des rappels ciblés ou des ajustements de planning.

## 4. Références bibliographiques

Le travail s'appuiera sur un ensemble de ressources combinant articles de recherche et contenus de vulgarisation de qualité, afin de disposer à la fois d'un cadre théorique solide et d'un retour d'expérience pratique.

Parmi les références envisagées :

— Chen, T. & Guestrin, C. (2016). *XGBoost : A Scalable Tree Boosting System*. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.
— Article de vulgarisation associé présentant XGBoost et ses usages pratiques, par exemple sur *Machine Learning Mastery* ou *KDnuggets*.
— Articles et analyses issus de plateformes spécialisées telles que *MIT Technology Review* ou des newsletters comme *Data Elixir*.

## 5. Démarche de test du nouvel algorithme (preuve de concept)

La démarche envisagée repose sur :

— la mise en place d'un **premier modèle baseline simple** (*régression logistique*), servant de référence minimale, facilement interprétable, afin d'établir un point de comparaison initial
— l'entraînement d'un **second modèle baseline plus expressif** (*Random Forest*), capable de capturer des relations non linéaires entre les variables, permettant d'évaluer le gain apporté par une méthode d'ensemble par rapport à un modèle linéaire
— l'entraînement d'un **modèle récent et plus avancé** (*XGBoost*), optimisé à l'aide de techniques de validation croisée et comparé aux deux baselines à l'aide de métriques adaptées au contexte métier
— l'analyse comparative des performances (accuracy, precision, recall, score métier) afin de quantifier l'apport réel du nouvel algorithme par rapport aux techniques utilisées précédemment
— l'intégration de **méthodes d'explicabilité** basées sur SHAP, permettant d'analyser à la fois l'importance globale des variables et l'explication locale de prédictions individuelles.