Analyse des besoins métiers

Prédiction du statut des rendez-vous médicaux

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

Résumé. Ce document présente le besoin métier lié aux rendez-vous médicaux (no-show, annulations), la preuve de concept en machine learning visant à prédire le statut d'un rendez-vous, et les mesures de conformité inspirées des principes CNIL/RGPD : finalité, minimisation, sécurité, durée de conservation, et prévention des biais.

Table des matières

1 Contexte métier et ancrage dans le réel **2**

2 Objectifs, périmètre et livrables **2**

2.1 Objectifs . . . 2

2.2 Périmètre . . . 3

2.3 Livrables . . . 3

3 Dataset retenu et description **3**

3.1 Source . . . 3

3.2 Fichier utilisé pour l'entraînement . . . 3

3.3 Colonnes principales (extrait) . . . 3

4 Cadre CNIL/RGPD : principes appliqués **4**

4.1 Finalité . . . 4

4.2 Minimisation des données . . . 4

4.3 Base légale (cadre théorique) . . . 4

4.4 Durée de conservation . . . 4

4.5 Sécurité, accès et confidentialité . . . 4

5 Biais, équité et interprétabilité **4**

5.1 Risques de biais 4
5.2 Mesures envisagées 5

6 Démarche de preuve de concept et méthode de test 5

6.1 Baselines et modèle récent 5
6.2 Pipeline et traçabilité 5
6.3 Explicabilité et aide à la décision 5

7 Limites et améliorations possibles 5

7.1 Limites 5
7.2 Améliorations 6

8 Conclusion 6

1. Contexte métier et ancrage dans le réel

La gestion des rendez-vous médicaux est un enjeu opérationnel important pour les cabinets et établissements de santé. Les *no-show* (non-présentations) et les annulations tardives entraînent une sous-utilisation des créneaux, augmentent les délais de prise en charge et dégradent l'organisation des équipes (planification instable, charge administrative).

Dans ce contexte, une approche **data-driven** peut aider à anticiper les rendez-vous à risque afin de déclencher des actions préventives (rappels ciblés, ajustement de planning, priorisation de confirmation) tout en limitant les relances inutiles.

Objectif métier. Réduire la perte de créneaux et améliorer la disponibilité effective des consultations, en anticipant les rendez-vous susceptibles d'être annulés ou manqués.

2. Objectifs, périmètre et livrables

2.1. Objectifs

— Construire une preuve de concept de classification supervisée permettant de prédire la variable cible `status`.
— Comparer des modèles baseline (LogReg, Random Forest) à un modèle récent (XGBoost optimisé).
— Apporter de l'explicabilité via SHAP (globale et locale) pour faciliter la compréhension et l'usage décisionnel.

2.2. Périmètre

Le projet vise une **preuve de concept** (POC) : évaluation technique, reproductible, sur un dataset limité/synthétique. Aucun déploiement en production n'est effectué.

2.3. Livrables

— Code d'entraînement et de suivi d'expériences (MLflow).
— Résultats comparatifs et figures (MLflow, SHAP).
— Documents : note méthodologique et plan prévisionnel.

3. Dataset retenu et description

3.1. Source

Le dataset est disponible en ligne :
https://www.kaggle.com/datasets/carogonzalezgaltier/medical-appointment-scheduling-system.
Il s'agit d'un dataset **synthétique** simulant des rendez-vous médicaux.

3.2. Fichier utilisé pour l'entraînement

Bien que le dataset global puisse contenir plusieurs tables, l'entraînement des modèles a été réalisé uniquement avec le fichier : `appointments.csv`.

La variable cible est : `status` (statut du rendez-vous). Conformément au script d'entraînement, lorsque `status` n'est pas numérique, une transformation en binaire est appliquée (rendez-vous à risque vs honoré), notamment avec des valeurs positives de type `cancelled` / `did not attend`.

3.3. Colonnes principales (extrait)

Colonne	Rôle / signification
scheduling_date	Date de prise de rendez-vous
appointment_date, appointment_time	Date/heure planifiée
scheduling_interval	Délai entre prise et RDV
check_in_time	Heure d'arrivée (si disponible)
appointment_duration, waiting_time	Indicateurs de déroulement
age, age_group, sex	Variables démographiques
status	**Cible à prédire**

Attention variables disponibles. Pour une utilisation en conditions réelles, certaines variables de déroulement (ex. heure réelle de début/fin, durée) pourraient ne pas être disponibles au moment de la prédiction. Ce point est traité dans la section 7.

4. Cadre CNIL/RGPD : principes appliqués

4.1. Finalité

La finalité de traitement, dans le cadre de ce projet, est strictement limitée à une preuve de concept académique : **évaluer la faisabilité et la performance d'un modèle prédictif** du statut des rendez-vous.

4.2. Minimisation des données

Le projet applique le principe de minimisation :

- seules les variables utiles à la modélisation sont conservées ;
- les identifiants techniques (ex. `patient_id`) ne sont utilisés que comme variables de structure, et peuvent être supprimés si non nécessaires ;
- si un champ de type `name` est présent dans le dataset global, il est considéré comme **généré artificiellement** (données fictives) et **n'est pas utilisé** pour l'apprentissage.

4.3. Base légale (cadre théorique)

Dans un contexte réel, un traitement de données de santé est un traitement sensible et nécessite un cadre légal renforcé (consentement explicite, intérêt public, ou cadre réglementaire spécifique). Dans ce projet, les données étant **synthétiques**, l'enjeu principal est de respecter les principes *by design* (minimisation, limitation des accès, transparence).

4.4. Durée de conservation

Les données et artefacts sont conservés uniquement pendant la durée du projet et de l'évaluation, puis supprimés (conservation limitée au besoin académique).

4.5. Sécurité, accès et confidentialité

- Données stockées localement / environnement privé (pas de partage public des fichiers de données).
- Le dépôt GitHub ne contient pas les données brutes (`.csv`), ni les modèles entraînés (`.joblib`), ni les logs MLflow.
- Contrôle des accès : dépôt privé si nécessaire, et partage limité aux encadrants.

5. Biais, équité et interprétabilité

5.1. Risques de biais

Même sur un dataset synthétique, un modèle peut apprendre des associations pouvant mener à des décisions non souhaitées. Des biais peuvent apparaître selon :

- **âge / age_group** (différences de comportement simulées),
- **sexe** (distribution différente dans les données).

5.2. Mesures envisagées

— Analyse des performances par sous-groupes (`sex`, `age_group`).
— Explicabilité via SHAP pour comprendre les facteurs dominants et détecter des usages problématiques.

6. Démarche de preuve de concept et méthode de test

6.1. Baselines et modèle récent

La preuve de concept compare :

— une baseline simple : **régression logistique**,
— une baseline d'ensemble : **Random Forest**,
— un modèle récent : **XGBoost** optimisé par **GridSearchCV**.

6.2. Pipeline et traçabilité

Le pipeline intègre :

— imputation (médiane numériques ; mode catégorielles) ;
— encodage one-hot des variables catégorielles ;
— split train/test stratifié ;
— suivi des métriques et artefacts via **MLflow**.

6.3. Explicabilité et aide à la décision

SHAP est utilisé pour :

— identifier les variables les plus influentes (importance globale) ;
— expliquer une prédiction individuelle (importance locale) ;
— rendre le modèle plus exploitable (justifier une alerte et proposer une action ciblée).

Exemple d'usage POC. Une interface minimale ou une démonstration simple peut interroger le modèle sur un rendez-vous et afficher la prédiction, les raisons principales (SHAP), et l'action recommandée (rappel ciblé, confirmation).

7. Limites et améliorations possibles

7.1. Limites

— Dataset synthétique : généralisation à valider sur données réelles.
— Variables potentiellement indisponibles au moment de la prédiction (déroulement).
— Risque de fuite d'information si l'on utilise des variables observées après l'heure du rendez-vous.

7.2. Améliorations

— Restreindre le modèle aux variables disponibles **avant** le rendez-vous (prédiction réellement actionnable).
— Validation temporelle (train sur passé, test sur période future).
— Définir un coût métier explicite (FN plus coûteux que FP) et ajuster le seuil.
— Reporting d'équité : performances par sous-groupes et surveillance des dérives.

8. Conclusion

Cette analyse formalise le besoin métier et décrit un cadre de preuve de concept en machine learning pour anticiper le statut des rendez-vous médicaux. Les principes CNIL/RGPD sont pris en compte via une finalité limitée, la minimisation, la sécurisation du stockage et l'absence de données brutes dans le dépôt public. L'explicabilité (SHAP) renforce la compréhension des prédictions et ouvre la voie à une utilisation comme outil d'aide à la décision.