

INSTITUT NATIONAL UNIVERSITAIRE JEAN-FRANÇOIS CHAMPOLLION
École d'ingénieurs ISIS

RAPPORT PROJET D'INNOVATION

Projet KADUCEO

Salomé CHEVAILLER

Camille MARC

Thibault TEISSEIRE

Promotion 2023 - Année universitaire 2022/2023

Remerciements

Nous souhaitons tout d'abord remercier le commanditaire de ce projet, Kaduceo, et tout particulièrement son représentant, M. Jean-Baptiste Excoffier, pour le temps qu'il nous a accordé, pour sa grande disponibilité, et pour son aide et ses conseils dont nous sommes très reconnaissants.

Également, nous souhaitons remercier notre professeur et tuteur école de ce projet, M. Francis Faux, de nous avoir permis de questionner nos recherches et résultats, et pour la relecture de ce rapport.

Glossaire

Apprentissage supervisé : technique de machine learning pour laquelle la machine apprend à prédire la valeur y en fonction des caractéristiques X qu'on lui donne

Cas d'usage : correspond à un ensemble d'actions réalisées par un système en interaction avec des acteurs, en vue d'une finalité

Connected Papers : outil d'aide à la recherche d'articles scientifiques

Data Science (ou science des données) : science interdisciplinaire visant à donner du sens aux données brutes

Dictionnaire (en informatique) : collection associant une clé à une valeur

Donnée qualitative : donnée utilisée pour décrire une information et pouvant être facilement regroupée en catégorie

Ecart-type : mesure de la dispersion des valeurs d'un échantillon statistique ou d'une distribution de probabilité

Encodage : désigne l'action de transcrire des données vers un format donné

Google Scholar : service de Google permettant la recherche d'articles et de publications scientifiques

Hyperparamètre : paramètre dont la valeur est utilisée dans le processus d'apprentissage

Jira : plateforme multifonction visant à faciliter la gestion de projet

Jupyter Notebook : outil permettant aux utilisateurs d'exécuter et partager du code en Python

Kaduceo : start-up toulousaine de santé fondée en 2014 qui développe et déploie des solutions logicielles basées sur de l'analyse de données pour les établissements hospitaliers

Kaggle : plateforme web accueillant la plus grande communauté de Data Science au monde et fournissant des outils et des ressources puissants en science des données

Machine learning : consiste à développer un modèle mathématique à partir de données expérimentales

Matrice de confusion : résumé des résultats de prédiction pour un problème particulier de classification

Matrice de corrélation : utilisée pour évaluer la dépendance entre plusieurs variables en même temps

Méthode agile : cycles de développement courts, très ciblés, impliquant le client et favorisant la collaboration entre des équipes pluridisciplinaires

Microsoft Teams : application de communication collaborative lancée par Microsoft en novembre 2016

Modèle (en machine learning) : fichier qui a été entraîné à partir d'une base d'apprentissage en vue d'automatiser des tâches et capable de générer des résultats à partir de données qu'il n'a jamais vues

No-code : illustre la capacité d'une solution à supprimer la barrière de la programmation informatique pour être accessible à tous

No-show : une personne qui ne se présente pas et qui n'annule pas son rendez-vous

Normalisation : permet de mettre un ensemble de données quantitatives sur une même échelle

Proof of Concept (ou preuve de concept) : réalisation ayant pour vocation de montrer la faisabilité d'un procédé ou d'une innovation

Python : langage de programmation interprété, multiparadigme et multiplateformes

Scikit-learn : bibliothèque Python permettant le machine learning

Sprint : brève période limitée dans le temps dont une équipe a besoin pour effectuer une quantité de travail donnée

Standardisation : méthode de normalisation très utilisée en machine learning

Validation croisée : méthode d'estimation de fiabilité d'un modèle fondée sur une technique d'échantillonnage

Validation croisée : technique permettant d'entraîner puis valider un modèle sur plusieurs découpages possibles du train set

Variable explicative : variable dont dépend une variable expliquée

Zoom : service de conférence à distance qui combine la vidéoconférence, les réunions en ligne, le chat et la collaboration mobile

Résumé

L'entreprise Kaduceo, start-up toulousaine de santé, a développé une solution clé en main d'aide à la décision permettant aux acteurs hospitaliers de construire des analyses rétrospectives et modèles prédictifs.

Cet outil, destiné aux professionnels de santé, comporte plusieurs cas d'usage, notamment celui du “no-show”. Notre étude porte donc sur ce cas d'usage : recherche de variables explicatives du no-show dans la lecture scientifique, manipulation d'un jeu de données pour étudier les variables en question, maquettage d'une solution à proposer aux professionnels.

À ces missions principales, s'ajoute la recherche de nouveaux cas d'usage qui nous semblent intéressants à étudier et à développer pour la plateforme, afin de venir compléter l'offre proposée par Kaduceo.

Sommaire

Remerciements	1
Glossaire	2
Résumé	4
Introduction	6
1- Contexte du projet	7
1.1 - Présentation du commanditaire	7
1.2 - Objectifs et enjeux du projet	7
2 - Choix d'environnement technique	8
2.1 - Outils de recherche utilisés	8
2.2 - Outils techniques utilisés	9
3 - Mise en application	10
3.1 - Prise de connaissance du “no-show”	10
3.2 - Recherche d'articles scientifiques et choix des variables explicatives	12
3.3 - Jeu de données utilisé	13
3.4 - Manipulation du jeu de données	16
3.5 - Elaboration de maquettes	21
4 - Gestion du projet et de l'équipe	25
5 - Conclusion et perspectives	27
5.1 - Conclusion générale	27
5.2 - Perspectives, pistes d'amélioration	27
5.3 - Ouverture (reprise du processus sur d'autres cas d'usage)	27
Bibliographie	29
Annexes	32

Introduction

Les rendez-vous médicaux non honorés sont un problème auquel sont confrontés les établissements de santé du monde entier, et ils ont un impact négatif sur les revenus, les coûts, ou la gestion de ressources. Il est donc essentiel de comprendre les facteurs associés au comportement de non-présentation. Ces comportements portent le nom de “no-show” : il s'agit de patients qui ne se présentent pas à leur rendez-vous, sans l'annuler ou sans en avertir le personnel. L'absentéisme d'un patient à son rendez-vous est une entrave à l'accès au système de soins pour d'autres patients, et ne fait qu'aggraver les délais de prise en charge, déjà trop longs.

Afin d'éviter ces répercussions négatives sur le système de santé, certains établissements sont forcés d'adopter des stratégies telles que le surbooking qui consiste à prévoir plus de patients que de créneaux disponibles, ou le rappel du rendez-vous au patient avant le jour-j.

Bien que les acteurs du système de santé souhaitent comprendre les raisons derrière ces comportements de no-show, l'identification des caractéristiques qui influencent ces derniers reste un problème de taille. Cela est notamment dû à la variabilité des contextes dans les établissements de santé.

Heureusement, ces dernières décennies ont vu l'intelligence artificielle prendre une place très importante dans le domaine de la santé. Les algorithmes de machine learning sont désormais des outils efficaces pour comprendre le comportement des patients, et gérer au mieux les plannings des professionnels, en se basant sur des modèles prédictifs.

Notre rapport s'inscrit dans une démarche de gestion d'un projet de machine-learning, de la définition du problème à résoudre, en passant par l'analyse et l'exploration de données, jusqu'à la présentation des résultats.

Focalisés sur le sujet du no-show dans un contexte de santé, nous faisons des recherches bibliographiques, manipulons un jeu de données pour étudier des variables, maquettions une solution à destination des professionnels, et enfin, recherchons de nouveaux cas d'usages pour le commanditaire de ce projet, l'entreprise Kaduceo. Celle-ci possède une plateforme d'aide à la décision hospitalière qui propose de nombreux cas d'usages, via la construction d'analyses et modèles prédictifs.

1- Contexte du projet

1.1 - Présentation du commanditaire

Kaduceo, fondée en 2014, est une start-up toulousaine de santé qui développe et déploie des solutions logicielles basées sur de l'analyse de données pour les établissements hospitaliers.

En 2022, Kaduceo réinvente son offre et regroupe tout son savoir-faire dans une solution clé en main (no-code) d'aide à la décision permettant aux acteurs hospitaliers de construire des analyses rétrospectives et modèles prédictifs (machine learning) actionnables au quotidien, pour un meilleur pilotage des établissements de santé. Ils peuvent par exemple prédire les flux patients et des durées de séjour pour anticiper des pics de saturation, évaluer un risque d'abandon de parcours ou d'absentéisme patient, optimiser le déroulement de parcours complexes en détectant la prise en charge la plus appropriée, etc.

Notre interlocuteur, dans le cadre de ce projet tutoré, est Jean-Baptiste Excoffier, ingénieur Data Scientist à Kaduceo depuis bientôt 4 ans. Il supervise les sujets et problématiques R&D, au contact des utilisateurs et prospects.

1.2 - Objectifs et enjeux du projet

L'objectif de ce projet est de concevoir un pipeline général de machine learning sur un cas d'usage de la plateforme de Kaduceo, celui de la prédiction du no-show. Il s'agit d'étudier la littérature scientifique sur le sujet, concevoir des modèles prédictifs avec des indicateurs issus de la lecture scientifique, et enfin réaliser des maquettes de présentation visuelles pour les acteurs hospitaliers, des résultats issus des modèles.

Le no-show correspond aux patients qui ont un rendez-vous programmé, mais qui pourtant ne s'y présentent pas. Avoir une prédiction du no-show sur les jours à venir permet aux acteurs hospitaliers de mieux gérer leurs plannings. Ils peuvent par exemple surcharger un planning en sachant qu'il y a un risque qu'un grand nombre de patients ne se présentent pas.

Kaduceo a déjà mis en place plusieurs cas d'usages variés tels que la prévision de l'affluence aux urgences ou la prévision du nombre de lits occupés. Ils en ont également plus d'une dizaine en cours de développement tels que la prédiction de la durée de parcours, ou la prédiction des évolutions financières pour un compte donné. Le cas d'usage du no-show n'étant pas vraiment avancé sur la plateforme Kaduceo, cela constitue un enjeu pour l'entreprise que d'approfondir les recherches et la création d'algorithmes, ainsi que le maquettage sur ce cas d'usage. Notamment, cela permettra de faire un *Proof of Concept (POC)* ou Preuve de Concept afin de montrer la faisabilité de ce procédé ; ainsi Kaduceo pourra par exemple voir s'il y a d'autres variables auxquelles ils n'ont pas pensé, et qui sont à prendre en compte. L'approfondissement de ce cas d'usage viendra compléter l'offre proposée par Kaduceo pour les hôpitaux.

2 - Choix d'environnement technique

2.1 - Outils de recherche utilisés

Pour nous aider dans notre recherche de documents scientifiques, nous avons utilisé 2 outils.

Le premier est Google Scholar. C'est un service de Google permettant la recherche d'articles et de publications scientifiques. C'est un réel inventaire d'articles approuvés ou non, de thèses de type universitaire, de citations ou encore de livres scientifiques. Il nous a permis de trouver des articles précis sur notre sujet d'étude.

Le second est Connected Papers. C'est un outil d'aide à la recherche d'articles scientifiques qui nous propose des articles en rapport avec un sujet ou un autre article en particulier. Il utilise les citations et les références dans les articles pour trouver ceux qui se font référence entre eux. Connected Papers présente ses résultats sous forme de graphique, ce qui permet de représenter rapidement et visuellement les résultats (cf. *Figure 1*). Nous l'avons utilisé dans le but d'élargir notre recherche d'articles qui était composée initialement de ceux proposés par notre commanditaire. Ainsi, nous avons sélectionné dans le graphe les articles scientifiques les plus proches de l'article initial, en termes de distance, tout en vérifiant que le sujet de celui-ci soit en lien avec le no-show de patients, dans un contexte de santé. En effet, plus les articles sont espacés de l'article initial sur le graphe, moins ils sont similaires à celui-ci.

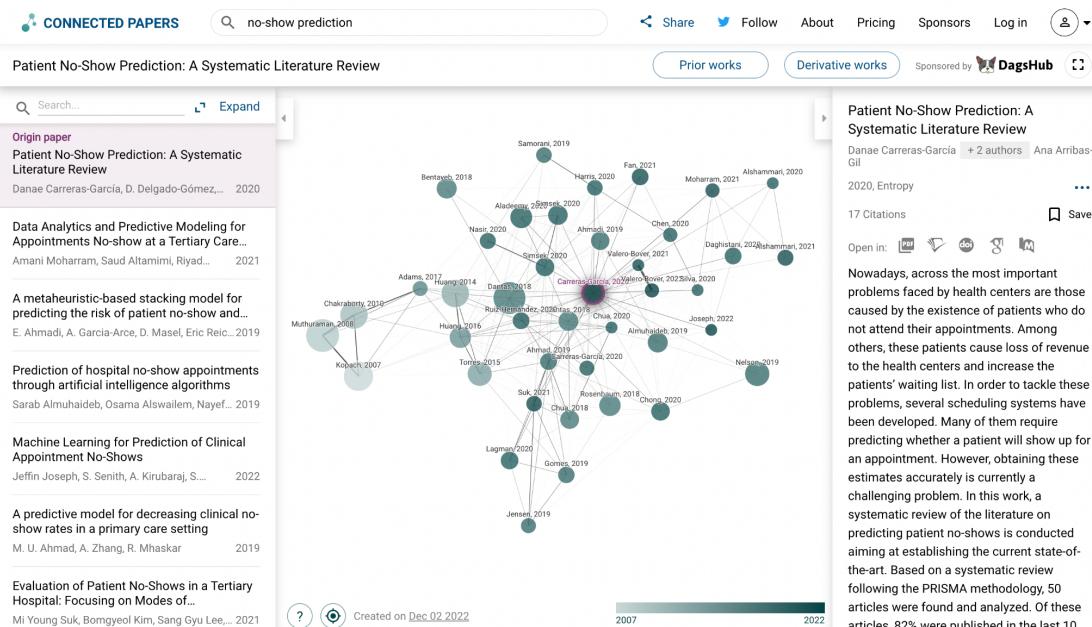


Figure 1 - Résultats d'une recherche avec Connected Papers

2.2 - Outils techniques utilisés

Afin de réaliser le projet, nous avons essentiellement utilisé Jupyter Notebook, qui est un outil puissant permettant aux utilisateurs d'exécuter et partager du code en Python. Pour le code, nous avons utilisé les bibliothèques Python suivantes :

- Pandas : permet la manipulation et l'analyse de données
- Scikit-learn : permet le machine learning
- NumPy : permet la manipulation de matrices ou tableaux multidimensionnels et de fonctions mathématiques
- Matplotlib : permet de tracer et visualiser des données sous forme de graphiques
- Seaborn : permet de créer des graphiques statistiques

Nous avons également utilisé la plateforme de démonstration Kaduceo afin d'avoir un aperçu des fonctionnalités proposées par l'entreprise et utiliser nos données sur le modèle de "no-show" actuellement présent.

Pour les maquettes visuelles de présentation, nous avons d'abord utilisé l'application de présentation collaborative Google Slides pour réaliser des ébauches très rapidement. Cela nous a permis de facilement créer des champs pour indiquer où seraient situés les éléments d'une page web, notamment à l'aide de simples rectangles.

Pour réaliser des maquettes plus abouties, nous avons utilisé l'application de présentation Keynote, qui est l'équivalent de PowerPoint, développée par Apple. Cela nous a permis de construire des maquettes plus propres et plus symétriques qu'avec Google Slides.

Ayant pour ambition, dans nos maquettes finales, d'utiliser la charte graphique de la plateforme de Kaduceo, pour que celles-ci paraissent natives à l'entreprise, nous avons utilisé l'outil "Inspecter l'élément" du navigateur. Bien que servant initialement à voir précisément comment les développeurs ont construit un site web, cela nous a permis de modifier les pages web de la plateforme Kaduceo à notre guise, dans le but d'en faire des captures d'écran. Nous l'avons notamment utilisé pour obtenir des titres avec la même police d'écriture que la plateforme, ou des graphiques. Ce sont ces captures d'écran de portions de la page que nous avons manipulées et agencées dans l'outil Keynote pour donner une impression de page web statique.

3 - Mise en application

3.1 - Prise de connaissance du “no-show”

Un “no-show” est une personne qui ne se présente pas et qui n’annule pas son rendez-vous. D’après l’article *The effect of no-show appointments on patients and Hospitals* [1], pas moins de 42% des patients ne se présenteraient pas à leur rendez-vous aux Etats-Unis. De plus, le no-show n’est pas seulement présent en Amérique du Nord, mais dans tous les systèmes de santé, qu’ils soient en Amérique du Sud, Afrique, Asie, Europe ou Océanie, comme l’indique l’étude portée par Parsons, J. et al en 2021 [2] dans la Table 2.

Bien qu’étant une problématique mondiale, le no-show n'est pas équitablement représenté dans la littérature scientifique. La majorité des informations que l'on trouve à son sujet concerne les États-Unis ou les pays anglophones plus généralement, et les raisons de non-présentation à un rendez-vous pour un patient sont multiples, et très variables.

La première raison pour laquelle les patients ne se présentent pas à leur rendez-vous est financière. En 2012, 49 millions de patients adultes ont renoncé à des soins recommandés en raison de leur coût élevé aux Etats-Unis [1]. Ce n'est toutefois pas une raison que nous nous attendons à retrouver en France, grâce à la protection universelle maladie qui est un droit pour toute personne qui travaille ou réside en France de manière stable et régulière. Cependant, certains soins (comme des soins dentaires, ou des chirurgies par exemple) ne sont pas entièrement remboursés.

Une autre raison pour laquelle un rendez-vous peut être manqué relève simplement de l'oubli de ce rendez-vous. En effet, malgré la présence et l'utilisation de nombreux canaux de communication : mails, sms, appels téléphoniques, applications mobiles, 48% des patients ne se présentent tout de même pas aux rendez-vous [3].

D'autres raisons amènent les patients à manquer des rendez-vous, notamment les problèmes de transport. En effet, de nombreux patients ne disposent pas d'un véhicule personnel et dépendent donc d'un proche ou des transports en commun pour les amener à leur rendez-vous [4].

Enfin, une mauvaise connaissance de la médecine, associée à la peur du soin et du système de santé, et des barrières démographiques (âge, langue parlée etc.) sembleraient être d'autres éléments pouvant influer sur le taux de no-show [5].

Certaines raisons semblent donc facilement être présentes également en France comme l'oubli de rendez-vous ou encore les problèmes de transport, tandis que d'autres comme la raison financière, ne sont sûrement pas transposables avec la France.

Tout comme les raisons, les conséquences d'un rendez-vous non honoré ne concernent pas seulement le patient, mais également les professionnels et le système de santé dans sa globalité.

D'après un article du 17 février 2017 du journal The Connexion [6], la Confédération des Syndicats Médicaux Français (CSMF) stipule qu'une étude a montré que les médecins perdaient en moyenne 6 à 8 rendez-vous pour un total de 2 heures de consultation par semaine. Mis à l'échelle de la France, cela représenterait pas moins de 28 millions de rendez-vous manqués chaque année. D'après les médecins, le temps perdu avec ces consultations non honorées aurait pu être utilisé pour faire des visites à domicile, notamment en zone rurale.

Ainsi, les conséquences ne sont pas seulement financières. Le no-show a un impact négatif sur la santé d'autres patients n'ayant pu être soignés, car il entrave l'accès au système de soins pour d'autres patients, et augmente les délais de prise en charge.

3.2 - Recherche d'articles scientifiques et choix des variables explicatives

Après avoir été familiarisés avec la problématique du "no-show", nous avons recherché des articles et revues littéraires scientifiques en lien avec cette notion afin d'en ressortir les variables explicatives les plus pertinentes à ajouter dans la plateforme Kaduceo, pour la prédiction de l'absence des patients à leur rendez-vous.

Comme mentionné précédemment, la littérature scientifique s'intéresse pour la grande majorité aux pays d'Amérique du Nord, en particulier aux États-Unis. C'est le cas des articles que nous avons étudiés.

Huit articles dont un résumé ont été retenus [7][8][9][10][11][12][13][14]. Toutes les variables explicatives trouvées dans ces articles et revues scientifiques ont été répertoriées dans un tableau (cf. *Figure 2*). Elles ont été triées selon les catégories suivantes pour obtenir un meilleur aperçu : données démographiques du patient, données liées au comportement du patient, données liées au rendez-vous, données liées au docteur, données liées à l'assurance maladie du patient et données liées aux caractéristiques cliniques du patient. Pour chacune des variables, son taux d'apparition dans les documents sélectionnés a été calculé (division du nombre d'apparitions dans les documents par le nombre total de documents) pour savoir quelles sont celles ayant le plus d'incidence dans la prédiction du no-show.

Les variables explicatives du no-show varient non seulement en fonction de la région étudiée, mais également en fonction de la spécialité du rendez-vous. De plus, les données auxquelles nous avons accès représentent une contrainte qui s'impose à l'application des variables, par exemple dans une autre région.

Ainsi, les résultats obtenus dans un hôpital public de campagne ne seront pas les mêmes que ceux d'une clinique de ville, que l'établissement soit en Europe ou en Amérique, ou encore que le rendez-vous concerne un cardiologue ou un chirurgien-dentiste.

Catégories	Données démographiques du patient	Comportement du patient	Rendez-vous
Variables	Age (100%) [7][8][9][10][11][12][13][14]	Taux d'absence (37,5%) [10][12][14]	Jour de la semaine (62,5%) [9][10][11][12][14]
	Genre (75%) [7][9][11][12][13][14]	Absence au RDV précédent (25%) [8][11]	Moment de la journée (75%) [9][10][11][12][13][14]
	Langue parlée (37,5%) [7][11][14]	Délai entre la prise de RDV et le RDV (62,5%) [8][9][11][13][14]	Mois (50%) [8][9][11][14]
	Ethnie (50%) [7][11][12][14]	Nombre de jours depuis le dernier RDV (25%) [10][11]	Saison (25%) [10][11]
	Province d'origine (12,5%) [9]	Statut de la dernière visite (12,5%) [11]	Météo (12,5%) [11]
	Code postal (12,5%) [11]	Expérience précédente (12,5%)	Jour de vacances (12,5%) [11]
	Distance du lieu de RDV (62,5%) [8][9][10][11][14]	Annulation du RDV précédent (12,5%) [11]	Type de l'examen (12,5%) [8]
	Région géographique (12,5%) [12]	Mode de prise du RDV (12,5%) [11]	Spécialité (37,5%) [8][11][12]
	Statut économique (12,5%) [11]	Satisfaction (12,5%) [11]	Statut du patient (37,5%) [9][12][14]
	Statut marital (25%) [10][11]	RDV pris 14 jours avant (12,5%) [10]	SMS de confirmation (25%) [8][13]
	Niveau d'éducation (12,5%) [11]	Heure de la prise du RDV (12,5%) [9]	Lieu du RDV (37,5%) [9][11][12]
	Emploi (12,5%) [11]		
	Accès au téléphone (12,5%) [11]		
	Religion (25%) [11][14]		
	Nombre de personnes dans le foyer (12,5%) [14]		
Catégories	Docteur	Assurance maladie	Caractéristiques cliniques
Variables	Note en ligne (12,5%) [9]	% des coûts couverts (12,5%) [10]	Diagnostic (12,5%) [11]
	Type (12,5%) [9]	En possession (50%) [7][10][11][14]	Durée du suivi (12,5%) [11]
	Titre (12,5%) [9]	Titulaire de l'assurance (12,5%) [14]	Nombre d'admissions à l'hôpital (12,5%) [10]
			Nombre de précédents RDV (25%) [10][11]
			Hypertension (12,5%) [13]
			Handicap (12,5%) [13]
			Index Charlson (12,5%) [10]
			Diabète (25%) [10][13]
			Maladie cardiaque (12,5%) [10]
			Dépression majeure/profonde (12,5%) [10]
			AVC ou démence (12,5%) [10]
			Douleur chronique (12,5%) [10]
			Insuffisance cardiaque (12,5%) [10]
			Maladie pulmonaire chronique (12,5%) [10]
			Dépendance aux drogues (12,5%) [10]
			Consommation de stupéfiants (12,5%) [10]
Légende	50% d'apparition		
	> 50% d'apparition		

Figure 2 - Variables explicatives trouvées dans la documentation

Comme nous pouvons le constater, les variables les plus fréquentes sont l'âge du patient [7][8][9][10][11][12][13][14], son genre [7][9][11][12][13][14], son ethnie [7][11][12][14], la distance entre son domicile et le lieu du rendez-vous [8][9][10][11][14], le délai entre la prise du rendez-vous et le rendez-vous [8][9][11][13][14], le jour de la semaine du rendez-vous [9][10][11][12][14], le moment de la journée du rendez-vous (matin, après-midi, soir) [9][10][11][12][13][14], le mois du rendez-vous [8][9][11][14], et si le patient est en possession ou non d'une assurance maladie [7][10][11][14]. Cette dernière variable n'est toutefois pas forcément pertinente pour le cas de la France car l'assurance maladie concerne toute personne qui travaille ou réside en France de manière stable et régulière. Mais, cela peut être intéressant de prendre pour variable la mutuelle du patient, certains soins n'étant pas ou peu remboursés.

Pour qu'un choix puisse être fait, les variables déjà utilisées dans la plateforme Kaduceo ont été répertoriées, sous la même forme que précédemment (cf. Figure 3).

Catégories	Données démographiques du patient	Comportement du patient	Rendez-vous	Docteur	Assurance maladie	Caractéristiques cliniques
Variables	Age	Venue lors des 3 derniers jours	Humidité		Type d'assurance	Durée de séjour
	Sexe	Venue lors des 7 derniers jours	Jour de la semaine			Hospitalisation lors des 3 derniers jours
		Venue lors du mois précédent	Pluie			Hospitalisation lors des 7 derniers jours
		Nombre de venues les 3 derniers mois	Pression au niveau de la mer			Hospitalisation lors du dernier mois
		Consultation lors des 3 derniers jours	Vitesse du vent			
		Consultation lors des 7 derniers jours	Température			
		Consultation lors du mois précédent				
		Taux d'absentéisme passé				
		Statut de la précédente visite				

Figure 3 - Variables explicatives de la plateforme Kaduceo

Tout d'abord, une comparaison entre les variables explicatives déjà présentes dans la plateforme Kaduceo et celles trouvées dans la littérature scientifique a été faite. Finalement, les variables sélectionnées (cf. *Figure 4*) sont celles ayant un taux d'apparition dans la littérature élevé et pouvant être aisément récoltées d'un point de vue de la faisabilité mais aussi d'un point de vue éthique (il n'est par exemple pas envisageable en France de recueillir l'éthnie du patient).

Catégories	Données démographiques du patient	Comportement du patient	Rendez-vous	Caractéristiques cliniques
Variables	Distance du lieu de RDV	Absence à un RDV précédent	Moment de la journée	Hypertension
		Délai entre la prise de RDV et le RDV	Mois	Diabète
		Expérience précédente	Spécialité	
		Heure de la prise de RDV	SMS de confirmation	
		Statut du patient		

Figure 4 - Variables explicatives retenues

3.3 - Jeu de données utilisé

Afin de pouvoir mettre en œuvre la prédiction du “no-show” à l'aide de machine learning, notre commanditaire, Jean-Baptiste Excoffier, nous a conseillé d'utiliser un jeu de données Open Source [\[15\]](#). Un travail a donc été fourni sur des données du Brésil car Kaduceo ne pouvait pas nous fournir de données d'établissements partenaires français. De plus, très peu de jeux de données pertinents sont à disposition. Ce jeu de données est disponible sur Kaggle, une plateforme web accueillant la plus grande communauté de Data Science au monde et fournissant des outils et des ressources puissants en science des données.

Ce jeu de données recueille des informations sur plus de 110 000 rendez-vous médicaux pris dans plus de 45 établissements médicaux différents du Brésil pour plus de 60 000 patients entre le 29/04/2016 et le 06/06/2016. Chaque rendez-vous médical est composé de 14 variables qui sont :

- PatientId : identifiant du patient
- AppointmentID : identifiant du rendez-vous
- Gender : genre du patient (homme ou femme)
- ScheduledDay : date où le patient a pris le rendez-vous
- AppointmentDay : date du rendez-vous
- Age : âge du patient
- Neighbourhood : lieu du rendez-vous
- Scholarship : aide financière (aux familles brésiliennes pauvres) (0 ou 1)
- Hypertension : patient atteint d'hypertension (0 ou 1)
- Diabetes : patient atteint de diabète (0 ou 1)
- Alcoholism : patient atteint d'alcoolisme (0 ou 1)
- Handcap : patient atteint de handicap (0 ou 1)
- SMS_received : message de confirmation envoyé au patient (0 ou 1)
- No-show : patient absent au rendez-vous (oui ou non)

Une analyse exploratoire des données de ce dataset a tout d'abord été menée pour évaluer la pertinence de son utilisation. Pour cela, il était nécessaire de "préparer" les données pour faciliter leur traitement. Certaines colonnes ont été converties dans un autre format, par exemple les colonnes "ScheduledDay" et "AppointmentDay" représentées par des chaînes de caractère et où il était préférable de les avoir au format date-time. Les lignes comportant une valeur inférieure à zéro pour la colonne "Age" ont également été supprimées car un âge ne peut pas être négatif. De plus, certaines colonnes ont été renommées pour mieux s'y retrouver et coder plus facilement par la suite (exemple : Hipertension → Hypertension). Puis les lignes ayant une date de prise de rendez-vous plus tard que la date de rendez-vous ont été supprimées car cela n'est tout simplement pas possible. Pour finir, il était pertinent de créer les colonnes suivantes à partir des colonnes déjà présentes dans le jeu de données, en se basant sur l'étude faite précédemment des variables explicatives utiles à la prédiction du "no-show" :

- PreviousAppointment : nombre de précédents rendez-vous que le patient a pris
- PreviousNoShow : taux d'absence du patient
- ScheduledHour : heure de la prise du rendez-vous
- ScheduledDate : date de la prise du rendez-vous
- AppointmentDate : date du rendez-vous
- AppointmentWeekDay : jour de la semaine du rendez-vous
- AppointmentMonth : mois de l'année du rendez-vous
- WaitingDays : délai entre la prise du rendez-vous et le rendez-vous (en jours)

Une fois les données prêtées, des statistiques sur les données ont été calculées pour voir celles ayant un impact sur l'absence d'un patient à un rendez-vous. Pour cela, une matrice de corrélation des valeurs les unes avec les autres (cf. *Annexe 1*) a été créée et la corrélation entre les différentes variables et l'absence à un rendez-vous a été affichée. Cette matrice fait ressortir trois corrélations : une corrélation entre l'hypertension et l'âge, une corrélation entre l'hypertension et le diabète et une corrélation entre le diabète et l'âge. Ces corrélations peuvent-être expliquées d'un point de vue médical et ne nous sont pas utiles ici. Il ne semble pas y avoir de corrélation forte entre une caractéristique quelconque et le fait de ne pas se présenter à un rendez-vous.

Ainsi, il était intéressant d'étudier chaque caractéristique une à une par rapport au fait de ne pas se présenter au rendez-vous. Nous avons pu en déduire que le taux d'absence est similaire pour les hommes et les femmes et que l'hypertension est un facteur important pour une fréquence de présence plus élevée (cf. *Figure 5*).

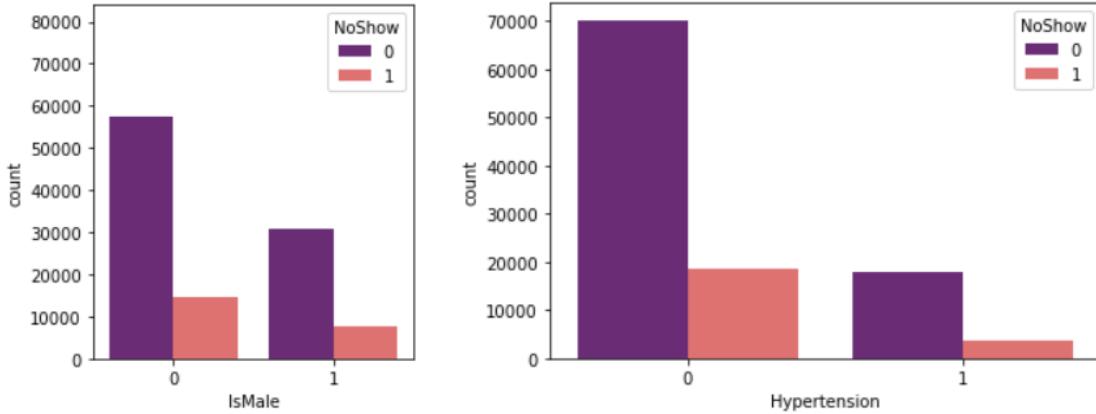


Figure 5 - Répartition des données selon le sexe et l'hypertension

De manière surprenante, nous avons observé que le taux d'absence était plus élevé pour les personnes recevant un SMS de confirmation du rendez-vous que pour les personnes n'en recevant pas, ce qui est contraire à ce que nous avons pu lire dans la littérature. Concernant les jours de la semaine, nous n'avons pas trouvé de différence significative pour ce qui est d'être présent à un rendez-vous (cf. Figure 6). Le taux d'absence semble être proportionnel au nombre de rendez-vous pris par jour de la semaine.

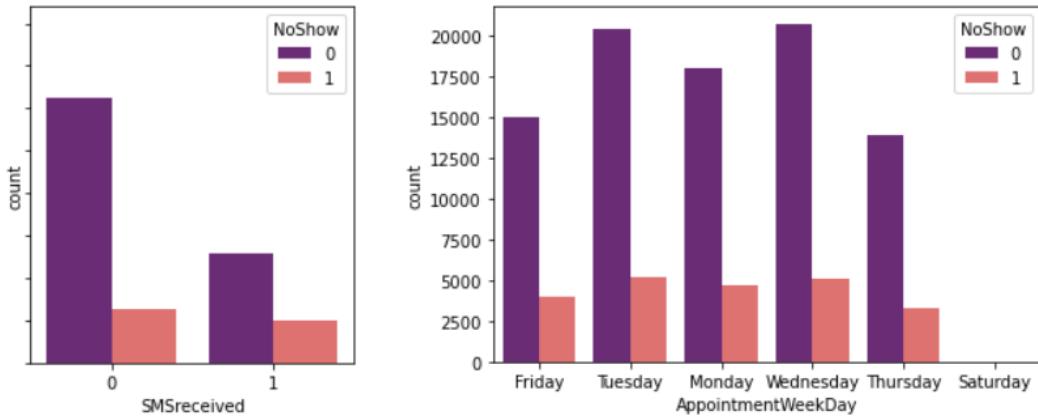


Figure 6 - Répartition des données selon la réception d'un SMS et le jour de la semaine

Suite à cette analyse exploratoire des données, quelques limites à ce jeu de données ont pu être trouvées. Tout d'abord, les données ont été recueillies entre le 29/04/2016 et le 06/06/2016. Si les données avaient été collectées sur une période de plus d'un an, cela refléterait mieux les changements dépendant du temps, notamment la saisonnalité. L'heure du rendez-vous est également manquante. Il serait intéressant de l'avoir car pendant les heures d'affluence, la probabilité de manquer un rendez-vous peut être plus élevée. Pour finir, certaines données cruciales corrélées à l'absence à un rendez-vous du patient telles que la spécialité du rendez-vous, la distance entre le domicile du patient et le lieu du rendez-vous, la situation du patient en matière d'assurance maladie et le niveau d'éducation auraient été utiles pour de meilleures analyses et prédictions.

3.4 - Manipulation du jeu de données

Après avoir été familiarisés avec la problématique du no-show, il était intéressant de manipuler le jeu de données évoqué précédemment à l'aide du machine learning pour pouvoir prédire si un patient sera présent ou non à un rendez-vous. Pour ce faire, nous avons suivi de nombreuses vidéos tutorielles nous présentant et expliquant des notions et techniques de machine learning [16].

Le machine learning consiste à développer un modèle mathématique à partir de données expérimentales. Ici, l'apprentissage supervisé a été utilisé car il était souhaitable de prédire une étiquette en fonction de caractéristiques données.

Le prétraitement des données est une des étapes les plus importantes pour avoir des modèles avec de bonnes performances. Comme évoqué dans la partie précédente, un nettoyage des données a déjà été fait pour les préparer à la prédiction. En effet, les données incohérentes et les valeurs incorrectes ont été supprimées et les données ont été converties dans des formats plus appropriés.

Néanmoins, il était indispensable de convertir les données qualitatives en valeurs numériques. Pour cela, la méthode d'encodage one-hot (cf. *Figure 7*) a été utilisée. Avec cette méthode, chaque catégorie est représentée de façon binaire dans une colonne qui lui est propre. La variable initiale est décomposée en plusieurs sous-variables, créant donc autant de colonnes qu'il y a de catégories dans cette variable.

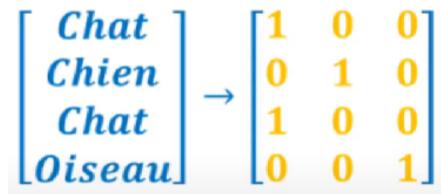


Figure 7 - Schéma illustrant la méthode d'encodage one-hot

Cette méthode d'encodage a donc été utilisée sur les colonnes "Neighbourhood", "Handicap", "AppointmentWeekDay" et "AppointmentMonth" (cf. *Figure 8*).

```
df = pd.get_dummies(df, columns=['Neighbourhood', 'Handicap', 'AppointmentWeekDay', 'AppointmentMonth'], prefix=['Neighbourhood', 'Handicap', 'AppointmentWeekDay', 'AppointmentMonth'])
```

Figure 8 - Code correspondant à l'encodage des données qualitatives

Une normalisation des données quantitatives a également été effectuée, c'est-à-dire que les données ont toutes été mises sur une même échelle. Pour cela, la standardisation a été utilisée. Cette technique consiste à transformer les données de telle sorte à ce que chaque variable ait une moyenne égale à 0 et un écart-type égal à 1. Pour cela il faut soustraire chaque valeur à la moyenne initiale de la variable et diviser le tout par l'écart-type de la variable (cf. *Figure 9*). On obtient ainsi des données très simples à utiliser pour la plupart des modèles statistiques.

$$X_{scaled} = \frac{X - \mu_X}{\sigma_X}$$

Figure 9 - Formule de la standardisation

Cette méthode de normalisation a donc été utilisée sur les colonnes "Age", "WaitingDays", "ScheduledHour", "PreviousAppointment" et "PreviousNoShow" (cf. *Figure 10*).

```
std=StandardScaler()
columns = ['Age', 'WaitingDays', 'ScheduledHour', 'PreviousAppointment', 'PreviousNoShow']
new_columns = ['AgeScaled', 'WaitingDaysScaled', 'ScheduledHourScaled', 'PreviousAppointmentScaled', 'PreviousNoShowscaled']
scaled = std.fit_transform(df[columns])
scaled = pd.DataFrame(scaled, columns=new_columns, index=df.index.values)
df = df.merge(scaled, left_index=True, right_index=True, how = "left")
```

Figure 10 - Code correspondant à la normalisation des données quantitatives

Suite à ce prétraitement des données, il a fallu les répartir en deux parties : les données d'entraînement et les données de test. Ainsi, 80% des données ont été mises dans le train set et 20% dans le test set. On obtient donc 88562 données dans le train set et 21959 données dans le test set (cf. *Figure 11*).

Train set data shape: (88562, 106)
Test set data shape: (21959, 106)

Figure 11 - Tailles du train set et du test set

Pour la prédiction, il a fallu choisir les algorithmes sur lesquels entraîner les données. En effet, tous les algorithmes ne sont pas appropriés à n'importe quelles données, certains sont mieux adaptés à certains types de données et à certains problèmes. Il en existe quatre types : les algorithmes de classification, de régression, de clustering (ou regroupement) et de réduction de la dimensionnalité.

Avec l'aide de la carte de Scikit-learn [17] (cf. *Figure 12*), seuls les algorithmes de classification ont été retenus car il était souhaitable de prédire une catégorie et que les données utilisées sont étiquetées.

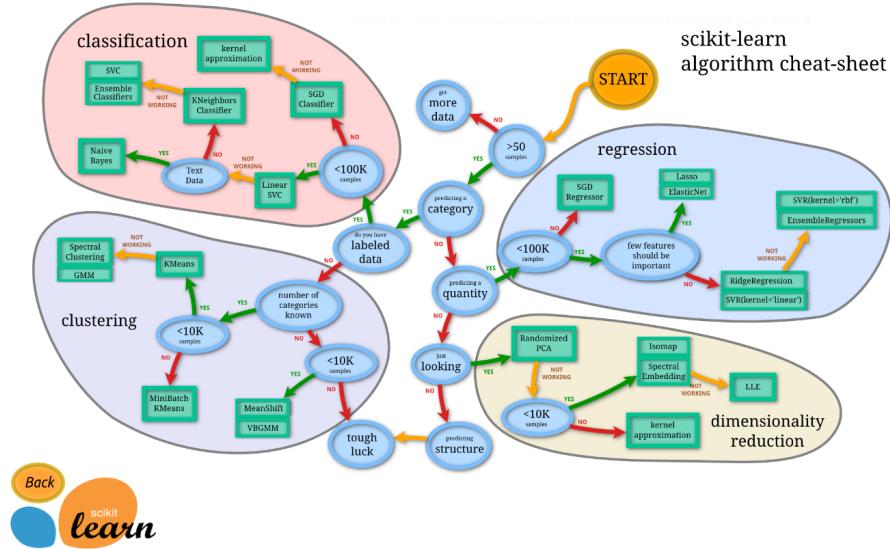


Figure 12 - Carte des algorithmes de Scikit-learn

Les données ont donc été entraînées avec les algorithmes suivants :

- KNeighbors
- Naive Bayes
- Decision Tree
- Random Forest
- XGboost
- AdaBoost
- LinearSVC

Comme évoqué précédemment, chacun des modèles a été entraîné sur les données du train set et évalué sur les données du test set. Concernant l'évaluation des performances du modèle, l'exactitude (ou accuracy) du modèle ainsi qu'un rapport de la classification et la matrice de confusion associée ont été affichés.

Le rapport de classification (cf. *Figure 13*) permet d'évaluer les modèles selon différentes métriques de performance de classification, à savoir la précision, le rappel et le f1-score, pour chaque classe. Les métriques sont calculées en utilisant les vrais et faux positifs, les vrais et faux négatifs. Dans le cas étudié, négatif et positif sont 0 et 1, 0 pour completed et 1 pour no-show. Il existe quatre façons de vérifier si les prédictions sont justes ou fausses :

- Vrai Négatif (VN) : lorsqu'un cas est négatif et qu'il était prévu qu'il le soit.
- Vrai Positif (VP) : lorsqu'un cas est positif et qu'il était prévu qu'il le soit.
- Faux Négatif (FN) : lorsqu'un cas est positif mais qu'il était prévu qu'il soit négatif.
- Faux Positif (FP) : lorsqu'un cas est négatif mais qu'il était prévu qu'il soit positif.

Classification report :					
	precision	recall	f1-score	support	
0	0.86	0.97	0.91	17392	
1	0.77	0.41	0.54	4567	
accuracy			0.85	21959	
macro avg	0.82	0.69	0.72	21959	
weighted avg	0.84	0.85	0.83	21959	

Figure 13 - Exemple de rapport de classification

La précision est la capacité d'un classificateur à ne pas étiqueter comme positive une instance qui est en réalité négative. Le rappel (ou recall), quant à lui, est la capacité d'un classificateur à trouver toutes les instances positives. Le score F1 est une moyenne harmonique pondérée de la précision et du rappel, de sorte à ce que le meilleur score soit 1.0 et le pire 0.0. En règle générale, les scores F1 sont inférieurs aux mesures d'exactitude car ils intègrent la précision et le rappel dans leur calcul. La moyenne pondérée de F1 est utilisée pour comparer les modèles de classificateurs, et non la précision globale. Et enfin, l'exactitude (ou accuracy) fait référence à la proximité d'une mesure par rapport à la valeur réelle ou acceptée.

La matrice de confusion permet de classer les résultats en quatre catégories : les vrais négatifs, les faux négatifs, les faux positifs et les vrais positifs (cf. Figure 14, de gauche à droite et de haut en bas). Pour rappel, dans le cas étudié le positif représente un patient qui ne s'est pas présenté à son rendez-vous et le négatif un patient s'étant rendu à son rendez-vous.

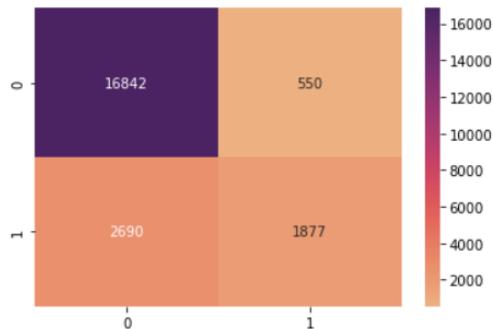


Figure 14 - Exemple de matrice de confusion

Au final, les résultats des modèles ont été reportés dans un tableau (cf. Figure 15) pour en avoir un aperçu global et pouvoir les comparer plus facilement.

Modèles	Score F1	Accuracy	Légende :
KNeighbors	0.51	0.8402	Meilleur score
Naive Bayes	0.39	0.4620	Deuxième
Decision tree	0.54	0.8116	Troisième
Random forest	0.53	0.8517	
XGboost	0.50	0.8541	
AdaBoost	0.50	0.8523	
LinearSVC	0.49	0.8485	

Figure 15 - Résultats des différents modèles utilisés

Nous pouvons en déduire que les trois meilleurs algorithmes pour la prédiction du no-show sont XGboost, AdaBoost et Random forest.

Cependant, l'exactitude (ou accuracy) n'était jamais supérieure à 0.86 et les modèles devaient être améliorés pour obtenir des scores plus élevés et donc obtenir une meilleure prédiction.

Pour cela, il était souhaitable de trouver les meilleurs hyperparamètres pour les modèles. Il existe dans Scikit-learn une méthode permettant de trouver le modèle avec les meilleurs hyperparamètres en comparant les différentes performances de chaque combinaison grâce à la technique de validation croisée : GridSearchCV. Pour cela, un dictionnaire contenant les différents hyperparamètres à régler ainsi que chaque valeur à tester pour ces hyperparamètres a été créé.

Le type de modèle a ensuite été passé dans la fonction GridSearchCV ainsi que le dictionnaire créé et un nombre de découpages pour la validation croisée. Une grille a ainsi été obtenue et entraînée avec les données du train set. Une fois cet entraînement terminé, le modèle qui a eu le meilleur score ainsi que les meilleurs hyperparamètres utilisés a été affiché et les performances du modèle ont été évaluées sur les données de test pour avoir un aperçu de la performance réelle (cf. *Figure 16*).

```
param_grid_dt = {'criterion': ['gini', 'entropy'], 'max_depth': np.arange(1, 30),
                 'min_samples_leaf': np.arange(1, 5), 'ccp_alpha': [0, 0.1, 0.01, 0.001]}
grid_dt = GridSearchCV(DecisionTreeClassifier(), param_grid_dt, cv=5)
grid_dt.fit(X_train2, y_train2)

Meilleure performance obtenue sur les données d'entraînement

grid_dt.best_score_
0.8492407489311514

Meilleurs hyperparamètres

grid_dt.best_params_
{'ccp_alpha': 0.01,
 'criterion': 'gini',
 'max_depth': 25,
 'max_features': 'sqrt',
 'min_samples_leaf': 1}

Performance obtenue sur les données de test

model_dt2 = grid_dt.best_estimator_
model_dt2.score(X_test2, y_test2)

0.858987090367428
```

Figure 16 - Code correspondant à la recherche des meilleurs hyperparamètres du modèle Decision Tree

Les résultats des modèles après optimisation des paramètres ont été répertoriés dans un tableau (cf. *Figure 17*), de la même manière que pour les modèles initiaux.

Modèles	Score F1	Accuracy	
KNeighbors	0.50	0.8512	
Naive Bayes	0.47	0.8494	
Decision tree	0.45	0.8495	
Random forest	0.53	0.8513	
XGboost	0.47	0.8513	
AdaBoost	0.47	0.8524	
LinearSVC	0.44	0.8490	

Figure 17 - Résultats des différents modèles utilisés avec les meilleurs hyperparamètres

Nous pouvons en déduire ici que les trois meilleurs algorithmes pour la prédiction du “no-show” sont Random forest, AdaBoost et KNeighbors, ce qui est plus ou moins les résultats obtenus sans l’optimisation des hyperparamètres.

Nous constatons d’ailleurs que cela n’a pas eu d’incidence sur la performance des modèles : on retrouve presque les mêmes résultats et il n’y a pas d’amélioration significative. Il est possible que cela soit dû au fait que certains hyperparamètres n’ont pas d’influence sur la performance, suggérant ainsi que les modèles initiaux utilisés avec les hyperparamètres de base étaient ceux les plus appropriés à nos données.

3.5 - Elaboration de maquettes

La plateforme Kaduceo permet aux clients qui l’utilisent d’avoir une représentation graphique des résultats des algorithmes de machine learning. Ces clients sont notamment des acteurs du système hospitalier, tels que des médecins, des chefs de service. Ils ont donc besoin d’avoir un affichage d’informations claires et simples, pour pouvoir répondre à leurs questions, qui ne concernent pas le machine-learning en soi, mais les résultats obtenus.

Nous avons donc imaginé des maquettes de pages de la plateforme Kaduceo, pour proposer des présentations et des agencements des résultats des algorithmes qui prédisent le no-show. La consigne du commanditaire du projet, Jean-Baptiste Excoffier, était de réaliser des maquettes simples et efficaces, qu’il pourrait par la suite présenter à l’UX designer de Kaduceo. Ainsi, nos maquettes avaient pour but d’être des pistes ou des idées de présentation des résultats, et qu’elles soient compréhensibles par des médecins. Des planches statiques de pages web répondent donc à la consigne.

Des acteurs du système hospitalier tels que des chefs de service utiliseront, au travers de la plateforme Kaduceo, ces pages de présentation des résultats concernant le no-show. Ils auront connaissance d’indicateurs comme le taux de no-show, et anticiper ce dernier pour les jours à venir leur permettra de mieux réagir, pour une meilleure gestion générale de leur service, et de l’hôpital.

Le fait que cela soit porté à leur connaissance est bénéfique pour plusieurs raisons. Tout d'abord, cela permet une meilleure allocation des ressources, qui pourraient être gâchées si le chef de service ne s'est pas préparé à l'éventualité qu'un patient n'honore pas son rendez-vous. Ces ressources peuvent être du matériel médical, mais également du personnel médico-social. Ainsi, le chef de service pourra éviter le sur-staffing, qui est le fait d'avoir un trop grand nombre d'employés par rapport à une quantité de travail trop faible, en prévoyant par exemple 1 seul médecin au lieu de 3 habituels.

Afin de mieux gérer l'utilisation des ressources de l'hôpital, le chef de service pourra, en fonction des taux de no-show prévisionnels, employer différentes stratégies de gestion de son service. Une de celles-ci pourrait être l'autorisation de surbooker, c'est-à-dire allouer un nombre de places de rendez-vous supérieur à celui dont le service dispose réellement, en clair : prévoir plus de patients que de créneaux. Par conséquent, les patients surnuméraires combleraient les créneaux des rendez-vous non honorés.

Ainsi, le fait que les personnes en charge de l'organisation de l'hôpital puissent avoir accès aux prédictions des patients qui sont les plus susceptibles de manquer leur rendez-vous, leur permet de guider l'établissement vers une meilleure orientation et une meilleure prise en charge.

Malheureusement, le no-show ne touche pas seulement le secteur de la santé, mais plus généralement le secteur du service. Cela inclut donc les restaurants, les centres de bien être, etc.[\[18\]](#). Comme il s'agit d'un phénomène déjà connu et bien présent, nous avons pu nous inspirer de ce qui existe déjà dans ces industries, en plus de voir ce qu'il pouvait déjà exister en termes d'interfaces pour le no-show. C'est à l'issue de cela que nous nous sommes inspirés d'une représentation issue d'une publication de "Journal of Big Data", *Predictors of outpatients' no-show: big data analytics using apache spark*, *Figure 4* [\[19\]](#).



Figure 18 - Maquette 1

Cette première maquette (cf. *Figure 18*) affiche l'ensemble des informations dont un chef de service ou pôle de l'hôpital pourrait se servir pour connaître les indicateurs de no-show.

Le bouton de filtre “Période” est une liste déroulante qui permet de sélectionner des périodes hebdomadaires : la dernière semaine, les 2 dernières, les 3 dernières, etc.

Le bouton de filtre “Pôle” est une liste déroulante permettant de sélectionner le pôle qui nous intéresse. Un pôle contient plusieurs services. Par exemple, le pôle Cardiovasculaire & métabolique regroupe les services de cardiologie, diabétologie, endocrinologie etc. Le bouton permet d'afficher les données groupées relatives au pôle.

Par exemple, un chef de pôle sélectionnera le pôle qu'il dirige, pour voir les indicateurs de ce pôle.

Le bouton de filtre “Service” est une liste déroulante permettant de sélectionner le service qui nous intéresse. Cela permet d'afficher les données propres à un service d'un pôle. Par exemple, un chef de service sélectionnera le service qu'il dirige, pour voir les indicateurs de ce service.

Le bouton de filtre “Professionnel” est une liste déroulante permettant de sélectionner le professionnel qui nous intéresse afin d'afficher les taux de no-show prévisionnels le concernant. Par exemple, un chirurgien orthopédique se sélectionnera dans la liste, pour voir les indicateurs basés sur ses patients programmés.

Ainsi, lorsque l'on souhaite filtrer des résultats, il est possible de renseigner tous les éléments, ou pas. Si l'on choisit seulement le professionnel, le pôle et le service se sélectionnent automatiquement. Mais si seul le pôle chirurgie nous intéresse, nous ne sommes pas obligés de choisir un service et un professionnel afin d'obtenir des résultats.

La catégorie “Nombre total de no-show” représente le nombre total de no-show correspondant aux critères sélectionnés dans les listes déroulantes.

Le graphique “No-show over time” représente graphiquement le nombre de no-shows du jour ainsi que les prédictions des jours suivants.

Enfin, la “Vue d'ensemble des no-show des différents pôles” permet de visualiser les no-show par semaine de tous les pôles de l'établissement. Cette visualisation permet à n'importe quel professionnel de voir l'état de no-show des autres pôles. Cette vue d'ensemble est également prévue pour les no-show des différents services.

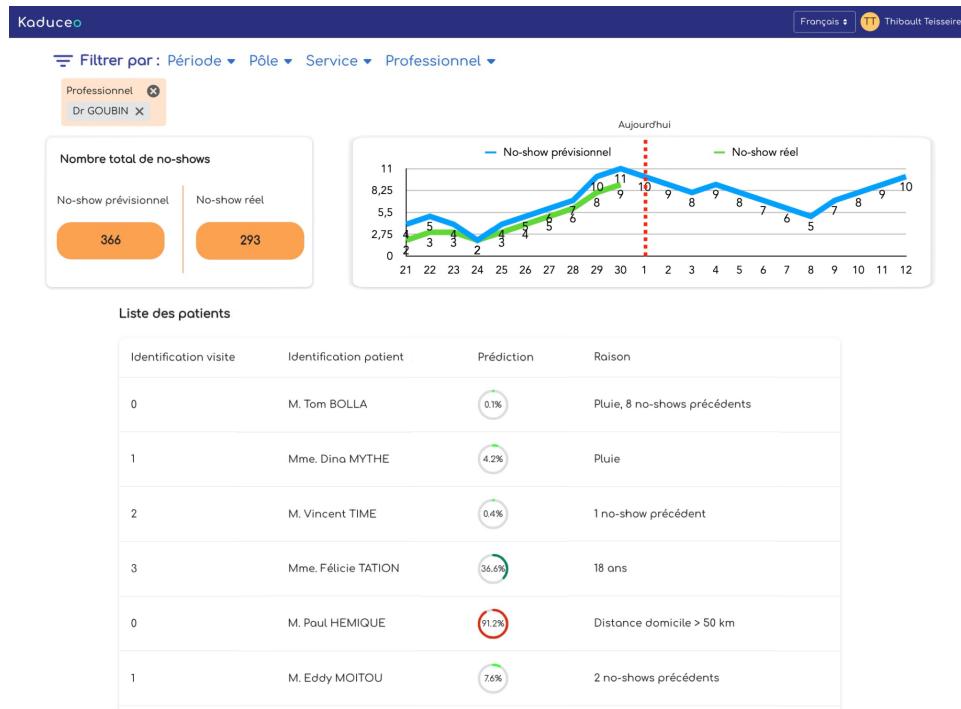


Figure 19 - Maquette 2

Cette seconde maquette (cf. *Figure 19*) affiche l'ensemble des informations lorsque l'on sélectionne un professionnel dans la maquette précédente. En effet, s'affiche alors la liste de ses patients, avec l'identifiant de la visite, leur prénom et leur nom, la prédiction de no-show qui leur est propre, et la raison du no-show.

La raison du no-show est obtenue à partir des variables explicatives ayant le plus influencé la prédiction. Par exemple, le patient, M. Paul HEMIQUE, a une prédiction de no-show de 91.2%, et cela est dû en très grande majorité au fait que son domicile soit situé à plus de 50 km de l'hôpital, et Mme Dina MYTHE a une prédiction de no-show de 4.2%, et cela est dû en très grande majorité au fait qu'il pleuve le jour du rendez-vous.

Il sera donc par exemple à l'appréciation du responsable en charge de l'organisation de l'hôpital, d'interpréter ces prédictions, et de prévoir une stratégie en fonction de cet affichage.

À noter que les maquettes ne se basent pas sur des données réelles, mais fictives, dans le seul but de représenter graphiquement de possibles résultats.

Ces deux maquettes proposées diffèrent, en quelques points, des interfaces proposées par la plateforme de Kaduceo pour le no-show, notamment :

- Un graphique sous forme de barres quantifiant le no-show par pôle et service
- Un graphique sous forme de courbe représentant l'évolution du no-show prévisionnel au cours des jours
- Un affichage des raisons principales du no-show

4 - Gestion du projet et de l'équipe

Nous étions un groupe de 3 étudiants à réaliser ce projet : Salomé Chevailler, Camille Marc, Thibault Teisseire. Ce dernier a été désigné chef de projet. Ainsi, Thibault avait pour mission de notamment piloter notre équipe, communiquer avec les tuteurs école et entreprise, et superviser les différentes étapes du projet.

Nous nous répartissions les tâches de manière interne, et nous faisions entre nous des réunions pour développer et débriefer sur l'état d'avancement du projet et ce qu'il restait à faire.

Afin de nous organiser dans la réalisation de ce projet, nous avons utilisé plusieurs logiciels. Le premier, Jira (cf. *Figure 20*), nous a permis de travailler avec une méthode agile, adaptée pour ce type de projet. En effet, à l'issue de chaque réunion avec Jean-Baptiste Excoffier, notre commanditaire, nous ajustions les missions et la direction que nous donnions au projet. En utilisant Jira, nous avons donc pu découper notre projet en "sprints", qui rythmaient les nouvelles phases du projet. Nous avons choisi de travailler avec des sprints d'une semaine, chaque début et chaque fin de sprint étant marqués par une réunion de projet avec le commanditaire. Chaque sprint comporte plusieurs tâches, certaines données explicitement par le commanditaire, d'autres venant de notre initiative, dans le but de répondre à certaines consignes.

0	0	0	Démarrer un sprint	...
<input checked="" type="checkbox"/> KADUCEO-1 S'informer sur la plateforme de Machine Learning à destination des hôpitaux : https://kaduceo.com/no-code-machine-learning-don...	À FAIRE			
<input checked="" type="checkbox"/> KADUCEO-2 Consulter les onglets Ressources et R&D du site	À FAIRE			
<input checked="" type="checkbox"/> KADUCEO-3 Lire les 5 articles scientifiques pour compréhension générale de la problématique du no-show	À FAIRE			
<input checked="" type="checkbox"/> KADUCEO-4 Rechercher d'autres articles scientifiques connexes aux 5	À FAIRE			
<input checked="" type="checkbox"/> KADUCEO-5 Lister les variables explicatives référencées dans la littérature	À FAIRE			
<input checked="" type="checkbox"/> KADUCEO-6 Tester la plateforme Kaducéo	À FAIRE			

+ Créer un ticket

Figure 20 - Extrait de Jira

Ensuite, ayant réalisé ce projet entièrement en distanciel que ce soit entre étudiants, ou avec le commanditaire situé à Toulouse, nous avons utilisé les logiciels Microsoft Teams et Zoom comme solutions de visioconférence, avec une préférence pour la dernière, car elle permettait un partage d'écran plus aisé que Teams.

Ainsi, chaque semaine, nous avions une réunion en visioconférence avec le commanditaire du projet, Jean-Baptiste Excoffier. Ces réunions permettaient de réorienter et affiner la direction du projet, et de lui faire un retour sur nos recherches. C'est très agréable car cela permet d'être guidés, tout en ayant une liberté de manœuvre chaque semaine pour faire nos recherches. Nous avons pu avancer de manière autonome, en ayant des objectifs et des consignes claires.

Également, nous avons eu deux réunions, une au début du projet et une autre à la fin, avec notre tuteur école, M. Francis Faux. Celles-ci ont permis de lui faire un retour sur l'avancée de notre projet, ainsi que de demander ses conseils et bénéficier de son expertise quant au travail que nous réalisons. En effet, cela a été nécessaire pour que nous prenions du recul et comprenions exactement les enjeux du projet, et questionnions nos recherches et résultats, dans le cadre d'un projet traitant de machine learning.

Pour visualiser le projet dans le temps et faciliter notre organisation, un diagramme de Gantt prévisionnel (cf. *Figure 21*) a été réalisé à l'issue de la réunion de lancement du projet. Celui-ci ne comporte évidemment que les grandes étapes et n'a pas vocation à indiquer toutes les tâches à effectuer mais seulement à avoir une vision d'ensemble sur le travail à réaliser et une estimation des délais.

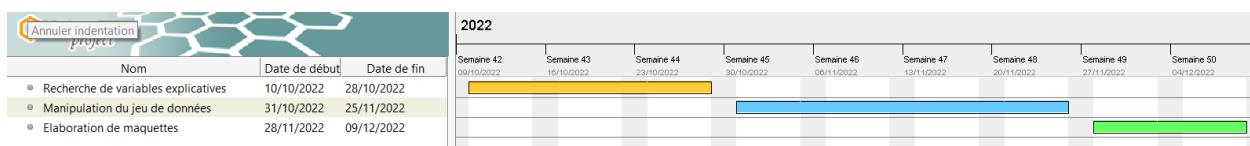


Figure 21 - Diagramme de Gantt prévisionnel

Il est intéressant ici de le comparer au diagramme de Gantt du projet tel qu'il s'est réellement déroulé (cf. *Figure 22*).

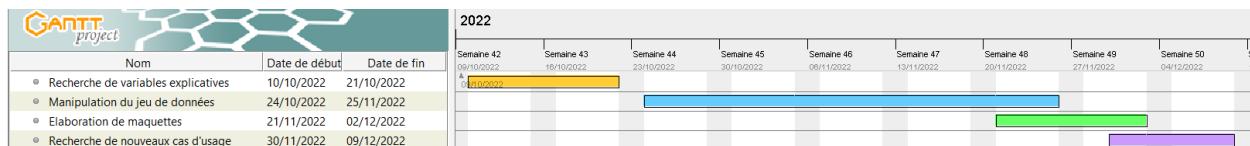


Figure 22 - Diagramme de Gantt du projet

Nous pouvons constater que la phase de recherche de variables explicatives a duré une semaine de moins que prévu, ce qui a permis de débuter la phase de manipulation du jeu de données une semaine plus tôt. Néanmoins, cette dernière a duré plus longtemps que prévu (5 semaines au lieu de 4) car il n'était plus nécessaire à la fin que l'entièreté de l'équipe s'y consacre. La capacité de l'équipe pour réaliser cette tâche a donc été réduite mais cela a permis de débuter la phase d'élaboration de maquettes une semaine plus tôt. Cette phase a duré autant de temps que prévu, ce qui a permis de pouvoir ajouter une phase de recherche de nouveaux cas d'usage à la fin du projet, qui n'était au départ pas prévue.

5 - Conclusion et perspectives

5.1 - Conclusion générale

Pour conclure, nous pensons avoir rempli les différents objectifs de ce projet, à savoir concevoir un pipeline général de machine learning sur le cas d'usage du no-show, à la fois en étudiant la littérature scientifique, en concevant des modèles prédictifs et en réalisant des maquettes de présentation visuelles à destination d'acteurs hospitaliers. Nous espérons que le travail fourni pourra aider Kaduceo à approfondir ce cas d'usage afin de compléter l'offre qu'ils proposent aux hôpitaux.

Ce projet fut très formateur de tous points de vue car nous avons pu appréhender la recherche bibliographique sur un sujet que nous ne connaissions pas auparavant. Il nous aura également permis de mettre à profit de nombreuses connaissances acquises au cours de notre formation d'ingénieur à ISIS et de les améliorer. La communication et la gestion de projet se sont notamment révélées être des compétences essentielles à la bonne réalisation du projet.

5.2 - Perspectives, pistes d'amélioration

Bien que nous ayons conçu un pipeline général sur le cas d'usage du no-show, certains points restent à perfectionner.

Concernant la partie machine learning, la performance des modèles est encore à améliorer pour obtenir de meilleures prédictions du no-show. Il serait également intéressant de tracer la courbe ROC (Receiver Operating Characteristic) représentant la sensibilité et d'évaluer l'aire AUC, ce qui permettrait d'avoir une visualisation des performances des modèles et de pouvoir les comparer les uns avec les autres.

Pour l'élaboration des maquettes, comme évoqué précédemment, il n'était pas nécessaire de réaliser un prototype cliquable puisqu'il s'agissait seulement d'obtenir une première vue de l'interface métier. Cependant, il aurait été intéressant que nous puissions mettre en œuvre les compétences acquises au sein de notre formation ISIS pour réaliser un réel prototype de l'application à l'aide d'outils dédiés (Figma). Cette phase sera néanmoins réalisée par la suite par l'UX designer de l'entreprise.

5.3 - Ouverture (reprise du processus sur d'autres cas d'usage)

Le projet avait également pour portée de cibler de nouveaux cas d'usages potentiels pour la plateforme de Kaduceo, afin de construire d'autres modèles prédictifs que ceux qu'elle a déjà en place. Nous proposons donc le cas d'usage de l'identification précoce des patients atteints de sepsis présentant un risque élevé de décès à l'hôpital.

Le sepsis est défini comme un état aigu de dysrégulation de la réponse de l'organisme à une infection entraînant la perte de fonction des organes et un risque vital pour le patient [20]. Il peut survenir de façon imprévisible, lors de n'importe quelle infection, le plus souvent bactérienne, mais aussi virale, notamment la grippe ou la COVID-19. Le sepsis doit être suspecté dès qu'une infection est présente, que la respiration s'accélère (plus de 22 cycles par

minutes), que la tension artérielle est basse (systolique (maxima) < 10), et que la conscience s'altère (propos incohérents, hallucinations, etc.).

Ce cas d'usage est très intéressant, notamment car le sepsis est presque totalement inconnu du public et encore mal appréhendé par les professionnels de santé. En Europe, on estime que le sepsis est responsable chaque année de près de 700 000 décès, dont près de 57 000 décès en France. Le coût moyen est d'environ 16 000 € par hospitalisation. Ce sont des arguments, ayant un poids sanitaire, économique et social, qui sont en faveur de la recherche et de l'amélioration de la prise en charge des patients.

Un article scientifique d'octobre 2020 intitulé “*Using machine learning methods to predict in-hospital mortality of sepsis patients in the ICU*” [21] s'est penché sur l'identification des patients atteints de cet état, à l'aide du machine-learning. Un total de 86 variables explicatives, telles que des données démographiques, des tests de laboratoires, et des comorbidités de patients, ont été utilisées (cf. Annexe 2). Dans l'étude, trois méthodes d'apprentissage automatique, à savoir LASSO, Random Forest et Gradient Boosting Machine, ainsi que la méthode LR, ont été utilisées pour développer des modèles prédictifs. De plus, l'outil de scoring, SAPS II, qui est un système de classification de la gravité d'une maladie (notamment dans le contexte des soins intensifs) a été utilisé.

Ainsi, les modèles de machine learning dans cette étude avaient de bonnes performances de prédiction. Parmi eux, le modèle GBM a montré les meilleures performances pour prédire le risque de décès à l'hôpital. Il a le potentiel d'aider les médecins à effectuer des interventions cliniques appropriées pour les patients atteints de sepsis et peut donc aider à améliorer leurs pronostics vitaux.

Nous pourrions appliquer un pipeline de recherche bibliographique - machine learning - maquettage, comme nous l'avons fait pour le no-show, mais cette fois sur ce sujet d'identification précoce des patients atteints de sepsis, et ainsi proposer un nouveau cas d'usage à Kaduceo.

Bibliographie

[1] Tšernov, K. (2022, 12 janvier) The Effect of No-Show Appointments on Patients and Hospitals

<https://www.qminder.com/blog/queue-management/no-shows-affect-hospitals/>

[2] Parsons, J., Bryce, C., & Atherton, H. (2021). Which patients miss appointments with general practice and the reasons why: a systematic review. In British Journal of General Practice (Vol. 71, Issue 707, pp. e406–e412). Royal College of General Practitioners.

<https://doi.org/10.3399/bjgp.2020.1017>

[3] Tranthimy, L. (2019, 11 avril) Rendez-vous non honorés : les « lapins » se banalisent et pourrissent l'agenda médical (et les patients le savent !)

[https://www.lequotidiendumedecin.fr/archives/rendez-vous-non-honores-les-lapins-se-banalisen t-et-pourrissent-lagenda-medical-et-les-patients-le](https://www.lequotidiendumedecin.fr/archives/rendez-vous-non-honores-les-lapins-se-banalisent-et-pourrissent-lagenda-medical-et-les-patients-le)

[4] Ofei-Dodoo, S., Kellerman, R., Hartpence, C., Mills, K., & Manlove, E. (2019). Why Patients Miss Scheduled Outpatient Appointments at Urban Academic Residency Clinics: A Qualitative Evaluation. Kansas journal of medicine, 12(3), 57–61. PMC6710029

[5] Patient No-Shows: Everything Practice Managers Need to Know

<https://www.relatient.com/patient-no-shows/>

[6] (2017, 17 février) Doctors want pay for 'no-shows'

<https://www.connexionfrance.com/article/Archive/Doctors-want-pay-for-no-shows>

[7] Kaplan-Lewis, E., & Percac-Lima, S. (2013). No-Show to Primary Care Appointments. In Journal of Primary Care & Community Health (Vol. 4, Issue 4, pp. 251–255). SAGE Publications.

<https://doi.org/10.1177/2150131913498513>

[8] Salazar, L. H. A., Parreira, W. D., Fernandes, A. M. da R., & Leithardt, V. R. Q. (2022). No-Show in Medical Appointments with Machine Learning Techniques: A Systematic Literature Review. In Information (Vol. 13, Issue 11, p. 507). MDPI AG.

<https://doi.org/10.3390/info13110507>

[9] Fan, G., Deng, Z., Ye, Q., & Wang, B. (2021). Machine learning-based prediction models for patients no-show in online outpatient appointments. In Data Science and Management (Vol. 2, pp. 45–52). Elsevier BV.

<https://doi.org/10.1016/j.dsm.2021.06.002>

[10] Daggy, J., Lawley, M., Willis, D., Thayer, D., Suelzer, C., DeLaurentis, P.-C., Turkcan, A., Chakraborty, S., & Sands, L. (2010). Using no-show modeling to improve clinic performance. In Health Informatics Journal (Vol. 16, Issue 4, pp. 246–259). SAGE Publications.

<https://doi.org/10.1177/1460458210380521>

[11] Carreras-García, D., Delgado-Gómez, D., Llorente-Fernández, F., & Arribas-Gil, A. (2020). Patient No-Show Prediction: A Systematic Literature Review. In Entropy (Vol. 22, Issue 6, p. 675). MDPI AG.

<https://doi.org/10.3390/e22060675>

[12] AlMuhaideb, S., Alswailem, O., Alsubaie, N., Ferwana, I., & Alnajem, A. (2019). Prediction of hospital no-show appointments through artificial intelligence algorithms. In Annals of Saudi Medicine (Vol. 39, Issue 6, pp. 373–381). King Faisal Specialist Hospital and Research Centre.

<https://doi.org/10.5144/0256-4947.2019.373>

[13] Alshammari, A., Almalki, R., & Alshammari, R. (2021). Developing a Predictive Model of Predicting Appointment No-Show by Using Machine Learning Algorithms. In Journal of Advances in Information Technology (Vol. 12, Issue 3). Engineering and Technology Publishing.

<https://doi.org/10.12720/jait.12.3.234-239>

[14] Hanauer, D. A., & Huang, Y. (2014). Patient No-Show Predictive Model Development using Multiple Data Sources for an Effective Overbooking Approach. In Applied Clinical Informatics (Vol. 05, Issue 03, pp. 836–860). Georg Thieme Verlag KG.

<https://doi.org/10.4338/aci-2014-04-ra-0026>

[15] Hoppen, J. (2016). Medical Appointment No Shows Dataset
<https://www.kaggle.com/datasets/joniarroba/noshowappointments>

[16] Machine Learnia, (2019) Python spécial Machine Learning, vidéos 20 à 23,
https://www.youtube.com/watch?v=P6kSc3qVph0&list=PLO_fdPEVlfKqMDNmCFzQISI2H_nJcEDJq&index=20

[17] Documentation Scikit-learn, algorithmes
https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html

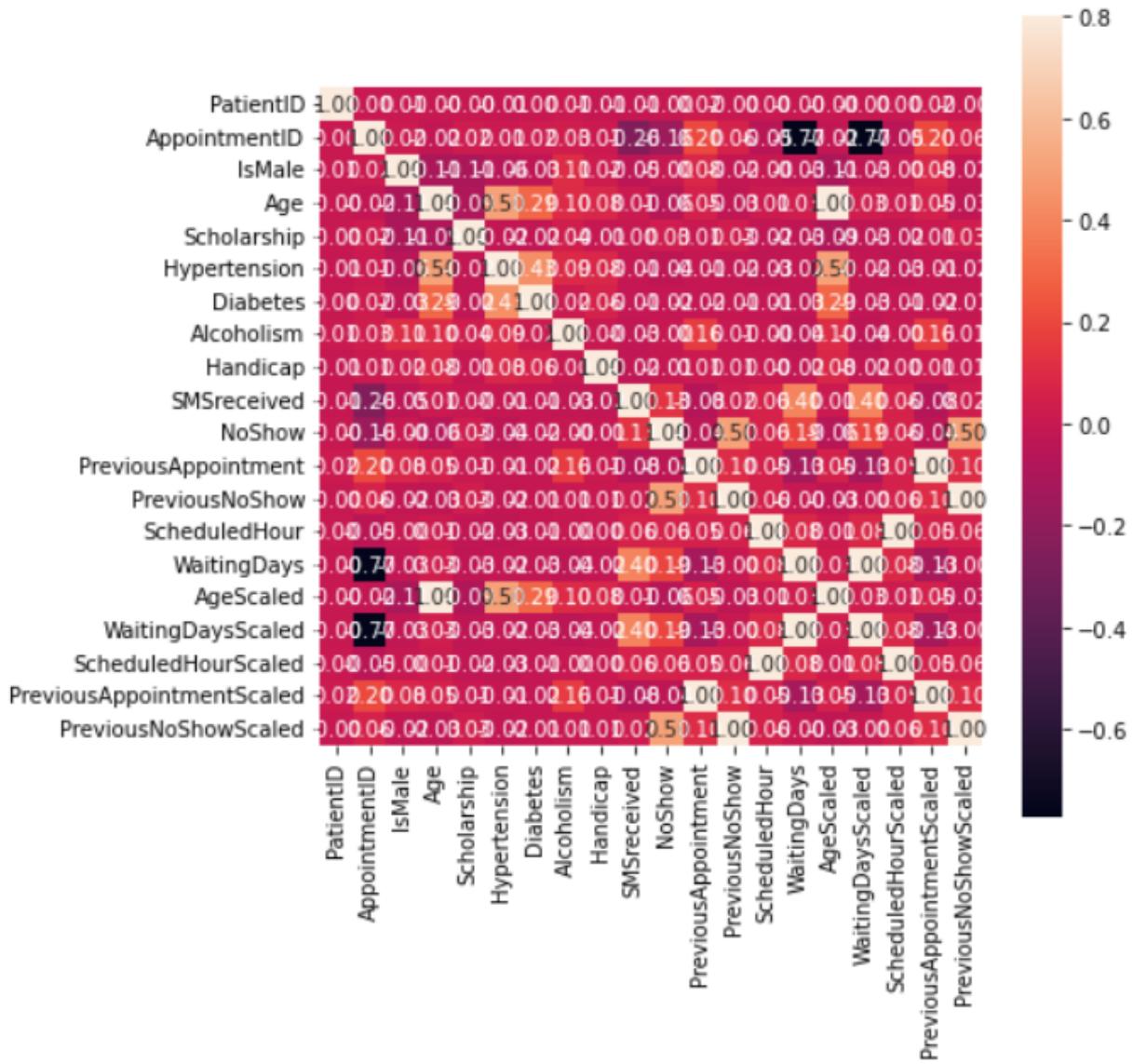
[18] Victoria, (2019, 19 mars), Quels leviers mettre en place face au « no-show » clients ?
<https://www.solocal.com/ressources/articles/prise-rdv-en-ligne-face-no-show-clients>

[19] Daghistani, T., AlGhamdi, H., Alshammari, R., & AlHazme, R. H. (2020). Predictors of outpatients' no-show: big data analytics using apache spark. In Journal of Big Data (Vol. 7, Issue 1). Springer Science and Business Media LLC.
<https://doi.org/10.1186/s40537-020-00384-9>

[20] Ministère de la santé et de la prévention, (2022, 18 juillet), Définition du sepsis
<https://solidarites-sante.gouv.fr/soins-et-maladies/prises-en-charge-specialisees/article/prevention-et-prise-en-charge-du-sepsis#:~:text=Le%20sepsis%20est%20d%C3%A9fini%20comme,risque%20vital%20pour%20le%20patient>

[21] Kong, G., Lin, K., & Hu, Y. (2020). Using machine learning methods to predict in-hospital mortality of sepsis patients in the ICU. In BMC Medical Informatics and Decision Making (Vol. 20, Issue 1). Springer Science and Business Media LLC.
<https://doi.org/10.1186/s12911-020-01271-2>

Annexes



Annexe 1 - Matrice de corrélation des variables explicatives

Predictors	
Acute physiology (first 24 h in the ICU)	Chronic health status
Heart rate*	Elixhauser comorbidity index
Systolic blood pressure*	Congestive heart failure
Diastolic blood pressure*	Cardiac arrhythmias
Mean blood pressure*	Valvular heart disease
Respiratory rate*	Pulmonary circulation
Temperature*	Peripheral vascular
SpO ₂ * (blood oxygen saturation)	Hypertension
Total CO ₂ *	Other neurological diseases
pCO ₂ * (partial pressure of CO ₂)	Chronic obstructive pulmonary disease
pH* (acidity in the blood)	Diabetes without complications
Urine output	Diabetes with complications
Glasgow Coma Score (GCS)	Hypothyroidism
GCS (eye)	Renal failure
GCS (motor)	Liver disease
GCS (verbal)	Metastatic cancer
Anion gap*	Coagulopathy
Bicarbonate*	Obesity
Creatinine*	Fluid electrolyte
Chloride*	Alcohol abuse
Glucose*	Depression
Haematocrit*	Renal replacement therapy
Haemoglobin*	Other
Lactate*	Gender
Platelet*	Weight loss
Potassium*	Ventilation
Partial thromboplastin time*	Age
INR*	Weight
Prothrombin time*	SAPS II score (first 24 h in the ICU)
Sodium*	SOFA score (first 24 h in the ICU)
Blood urea nitrogen (BUN)*	
WBC*	
Acute kidney injury	

Annexe 2 - Variables explicatives utilisées dans l'étude de prédiction du sepsis