République Algérienne Démocratique et Populaire
Université Abderrahmane MIRA de Béjaia
Faculté des Sciences Exactes

Département de Recherche Opérationnelle

Mémoire Présenté pour L'obtention du Diplôme de Master
en Mathématiques Appliquées

Spécialité : Sciences de Données et Aide à la Décision

Le Data Mining pour gérer les admissions dans un CHU :Prédiction des temps d'attente et priorisation des cas urgents

Présenté par :
LAIB ITHRY MAJID JUNIOR

Sous la direction de : M M. SAADI

Défendu le 29/06/2025, devant le jury composé de :

Me.C.Boughani	M.C. classe/ B	President de jury	UAMB - Bejaia.
Me.S.Kendi	M.C. classe/ B	Examinateur	UAMB - Bejaia
Me.K.Krimat	M.C. classe/ B	Examinateur	UAMB - Bejaia.

Année Universitaire 2024 – 2025

Remerciements

À Allah Le Tout Miséricordieux

Qui par Sa sagesse infinie a guidé mes pas tout au long de ce parcours exigeant. C'est par Sa grâce que j'ai pu puiser la force intellectuelle et spirituelle nécessaire pour mener à terme ce travail. Son illumination a été la lumiére qui a éclairé chaque challenge transformé en opportunité d'apprentissage.

À mon respectable encadreur, Monsieur SADI Moustapha

Je vous exprime ma profonde reconnaissance pour votre mentorship exceptionnel. Votre expertise pointue, votre disponibilité sans faille et votre vision prospective ont élevé ce projet bien au-delà de mes espérances initiales. Vos feedbacks constructifs, toujours formulés avec bienveillance, ont été les pierres angulaires de la rigueur scientifique de ce mémoire. Votre confiance en mes capacités a été un moteur déterminant.

Aux éminents membres du jury

Me C.Boughani,Me K. Krimat et Me S. Kendi, recevez mes plus sincères remerciements pour avoir accepté d'évaluer ce travail. Votre temps précieux et votre expertise reconnue donneront toute sa valeur académique à cette recherche. Je suis honoré par votre attention.

Aux héros du quotidien

Du CHU de Béjaïa à l'EPH de Kherrata, à chaque infirmier, médecin, administrateur et agent qui m'a accueilli dans leur quotidien professionnel : votre dévouement sans faille a été ma plus grande leçon d'humilité et d'excellence. Ces mois d'immersion sur le terrain ont ancré en moi la conviction que la data science au service de la santé doit rester profondément humaine. Merci pour votre patience à partager votre expérience.

À mon pére bien-aimé, initiateur et pilier

Tes sacrifices discrets, ta sagesse pratique et ta foi inébranlable en mon potentiel ont été les fondations de cette réussite. Ce projet que tu as su voir avant moi a ouvert mes yeux sur le pouvoir transformateur des données dans l'amélioration concréte des systémes de santé. Tes conseils avisés ont constamment recentré ma recherche sur l'essentiel : l'impact réel pour les patients et soignants.

À ma tendre mére

Ton amour inconditionnel et ton soutien quotidien ont été mon havre de paix dans les moments de doute. Tes priéres et ta fierté silencieuse ont été une source d'énergie constante. Cette réussite est d'abord la tienne, toi qui m'as appris que la persévérance vient du coeur avant de venir de l'esprit.

À ma chére famille

Votre présence bienveillante a été mon filet de sécurité. Merci pour votre patience durant mes absences, pour vos encouragements quand le découragement pointait, pour vos petites attentions qui ont rendu ce parcours plus léger. Vous êtes ma premiére équipe, mes supporters les plus fidèles.

À mes compagnons d'étude et amis

Vos échanges stimulants, vos relectures attentives et même vos simples présences ont enrichi ce travail bien au-delà de ce que vous imaginez. Nos discussions parfois tardives ont souvent été source de clarifications précieuses.

Au lecteur de ces lignes

En prenant le temps de parcourir ce travail, vous lui donnez sa raison d'être. Puisse-t-il contribuer, á sa modeste échelle, á l'avancement des connaissances et inspirer de futures recherches.

«*La gratitude est non seulement la plus grande des vertus, mais la mére de toutes les autres*» (Cicéron). C'est le coeur empli de cette vertu que je conclus ces remerciements, sachant pertinemment que ni les mots ni les pages ne suffiraient à exprimer pleinement ma reconnaissance envers chacun de vous.

Dédicace

À mon pére,

le roc sur lequel j'ai bâti mes rêves.C'est grâce à toi que ce mémoire est né. Ton soutien inébranlable, ta foi en mon avenir, ont été la source de mon courage. Sans toi, cette page ne serait qu'un espace vide. Ce travail te revient autant qu'à moi.

À mes soeurs et fréres,

complices de mon enfance et alliés dans cette aventure.Vous avez partagé mes joies, mes angoisses, mes défis. Vos mots, parfois simples, ont souvent eu le pouvoir de redonner le sourire quand tout semblait trop lourd.

À mes grands-parents,

gardiens des valeurs, des traditions et de la mémoire familiale.Votre sagesse, transmise au fil des ans, a façonné l'homme et la personne que je suis aujourd'hui. Vous êtes les racines de mon arbre,profondes et solides.

Et à toute ma famille proche,

vous qui avez tissé autour de moi un filet d'amour invisible, mais si fort.Merci d'avoir cru en moi, même avant que je ne croie vraiment en moi-même.

Ce mémoire,

c'est aussi le fruit de vos sacrifices, de vos priéres, de vos sourires.Je le dédie à chacun de vous, avec tout mon coeur.

Table des matières

Table des figures

Liste des tableaux

Liste d'abriviations et notations

- **DM** : Data Mining
- **RF** : Random Forest
- **RL** : Régression Linéaire
- **SVM** : Support-Vector Machine
- **RM** : Régression Multiple
- **CS** : Classification Supervisée
- **KNN** : K-Nearest Neighbors
- **CH** : Clustering Hiérarchique
- **DBSCAN** : Density-Based Spatial Clustering of Applications with Noise
- **ARIMA** : Auto Regressive Integrated Moving Average
- **LSTM** : Long Short-Term Memory
- **MAPE** : Mean Absolute Percentage Error
- **RMSE** : Root Mean Square Error

Introduction générale

Contexte générale

Dans de nombreux pays, les patients des Centres Hospitaliers Universitaires (CHU) rencontrent souvent des temps d'attente élevés, en particulier dans les services d'urgence. Cette situation découle principalement du volume important et parfois imprévisible de patients, ce qui peut entraîner une surcharge des ressources disponibles telles que le personnel médical, les salles d'examen et le matériel hospitalier. Ces facteurs influencent négativement la qualité des soins, la satisfaction des patients, ainsi que les résultats cliniques [16].

En Algérie, les centres hospitalo-universitaires (CHU) sont confrontés à des temps d'attente élevés aux urgences, une situation aggravée par plusieurs facteurs spécifiques au contexte national. L'afflux constant et croissant de patients, y compris ceux souffrant d'affections ne relevant pas de l'urgence, résulte du manque de structures de soins primaires accessibles ou opérationnelles, de la centralisation des ressources médicales dans les grands CHU, ainsi que de la hausse des maladies chroniques et des accidents, notamment en milieu urbain. Par ailleurs, la gestion inefficace des ressources hospitaliéres, marquée par un déficit en personnel médical qualifié, des équipements insuffisants et une organisation peu adaptée aux variations d'affluence, contribue à l'engorgement des urgences. Ces délais prolongés ont des répercussions négatives sur la qualité des soins, entraînant une aggravation de l'état des patients critiques, une baisse de la satisfaction et de la confiance des usagers, ainsi qu'un stress accru pour le personnel, exposé à un risque d'épuisement professionnel. Des facteurs externes tels que la saisonnalité des pathologies, les événements locaux et les disparités socio-économiques accentuent encore cette pression, les épidémies hivernales ou les accidents estivaux provoquant des pics d'affluence, tandis que l'inégalité d'accés aux soins en zones rurales pousse les populations vers les CHU des grandes villes.

Nous avons appliqué les techniques de Data Mining afin de modéliser, analyser et prédire les temps d'attente dans les services d'urgence de CHU. Nous nous sommes attachés à identifier les variables clés influençant les flux de patients, telles que l'heure d'arrivée, le type de pathologie et la disponibilité du personnel ; dans le but de développer des modéles prédictifs fiables permettant d'estimer les temps d'attente à partir de données historiques. En paralléle, nous avons proposés un systéme de priorisation des patients fondé sur l'analyse des données médicales et des niveaux de gravité. Nous avons également analysé les données pour détecter les goulots d'étranglement et suggéré des scénarios d'optimisation, ainsi que la conception d'un tableau de bord interactif destiné aux gestionnaires hospitaliers, leur offrant une visualisation en temps réel des indicateurs clés de performance.

Structure du mémoire :

Ce mémoire est organisé en quatre principaux chapitres :

— **Chapitre 1 : Présentation des établissements hospitaliers**
Présentation détaillée des deux structures étudiées : l'Établissement Public Hospitalier (EPH) de Kherrata et le Centre Hospitalier Universitaire (CHU) Khelil-Amrane. Ce chapitre aborde leur localisation, leur organisation, leurs missions ainsi que les défis qu'ils rencontrent dans la gestion des flux de patients.

— **Chapitre 2 : Revue de littérature et Cadre théorique**
Présentation des concepts fondamentaux du Data Mining, revue de la littérature scientifique et cadre théorique relatif à la gestion des flux hospitaliers.

— **Chapitre 3 : Méthodologie**
Détail des étapes de collecte, de nettoyage, d'exploration et de modélisation des données.

— **Chapitre 4 : Résultats et discussion**
Analyse des performances des modéles utilisés, interprétation des tendances identifiées et proposition de solutions concrétes.

— et enfin,le mémoire se termine par une conclusion générale et quelques perspectives.

1

Présentation de l'EPH de Kherrata et du CHU Khelil Amrane de Béjaia

Sommaire

Introduction

Nous posons dans ce chapitre, les bases conceptuelles et scientifiques nécessaires à la compréhension et a la mise en oeuvre du projet. Il se compose d'une revue de littérature sur les travaux académiques existants concernant l'utilisation du Data Mining dans le domaine hospitalier, d'un cadre théorique présentant les concepts clés avec les techniques utilisées (régression, classification, clustering) ainsi que leur application au contexte des admissions aux urgences. En s'appuiant sur des recherches récentes ainsi que sur des outils éprouvés tels que Random Forest , K-Means , Prophet , et LSTM , tous intégrés dans un processus rigoureux suivi dans le cadre du modéle CRISP-DM .

1.1 Etablissement Public Hospitalier (EPH) de Kherrata

L'EPH de Kherrata, créé par décret exécutif n° 07-140 du 19 mai 2007, est un établissement de santé publique de référence situé dans la wilaya de Béja ia. Il assure une couverture sanitaire étendue, incluant des populations locales et limitrophes.

FIGURE 1.1 – EPH de Kherrata

Situation géographique et démographie

— **Localisation :**

- **Zone sud :** sud-est de la wilaya de Béjaia.
- **Axe stratégique :** Positionné entre les pôles urbains de Sétif et Béjaia, prés d'un axe routier majeur(Incluant un tunnel à haut risque).

— **Couverture :**

- **2 Da irates :** Kherrata et Darguina.
- **6 Communes :** Kherrata, Dra El Gaid, Taskriout, Ait Smail, Darguina, Tamricht.
- **Superficie :** 485,54 km^2.
- **Étendue :** Communes des wilayas voisines (Sétif, Jijel, Béjaia).

Organisation et Services Hospitaliers

Arrêté ministériel n° 2738 du 26 janvier 2008.

Capacité d'Accueil

Service	Lits Techniques	Unités	Lits Organisés
Chirurgie générale	24	Hospitalisation (H/F)	12 H / 12 F
Médecine interne	52	Hospitalisation (H/F)	26 H / 26 F
Gynéco-obstétrique	24	Gynécologie / Obstétrique	24
Pédiatrie	16	Pédiatrie / Néonatalogie	10 / 6
Urgences médico-chirurgicales	20	Observation / Réanimation	8 / 10

TABLE 1.1 – Capacités et unités techniques spécifiques de l'EPH de Kherrata

Service	Lits Techniques	Unités	Lits Organisés
Hémodialyse	12		12
Épidémiologie	0	Information sanitaire / Hygiéne hospitaliére	0
Radiologie central	0	Radiologie / Écographie	0
Laboratoire central	0	Microbiologie / Biochimie	0
Pharmacie	0	Gestions des produits pharmaceutiques / Distribution des produits pharmaceutiques	0
Médecine de travail	0	Surveillance médicale des personnels santé / examen périodique de santé de travail	0
Total = 11 services	**148**	**20 unités**	**146**

TABLE 1.2 – Capacités et unités techniques spécifiques de l'EPH de Kherrata

Services Techniques

- Radiologie centrale (radiologie, échographie).
- Laboratoire (microbiologie, biochimie).
- Pharmacie (gestion et distribution).
- Médecine du travail (surveillance du personnel).
- Épidémiologie (hygiéne hospitaliére, information sanitaire).

Ressources Humaines

Personnel Médical

Catégories	Effectif	Ratio
Médecins spécialistes	25	1/10 916 habitants
Médecins généralistes	19	1/4 517 habitants
Infirmiers (ISP)	228	1/1 723 habitants
Sages femmes	23	1/7 277 habitants
AMAR (réanimation)	23	1/18 714 habitants

TABLE 1.3 – Répartition du personnel médical à l'EPH de Kherrata

Spécialités Médicales

Médecins spécialiste	Effectifs	Médecins spécialistes	Effectifs
Chirurgie générale	7	M. légiste	1
Orthopédistes	3	Néphrologues	2
Réanimateurs	1	Cardiologues	2
Pédiatres	4	Médecine interne	3
Gynécologue	1	Biochimie	1

TABLE 1.4 – Répartition des spécialités médicales à l'EPH de Kherrata

Particularités :

- **Service d'hémodialyse :** Opérationnel depuis 2008 (arrêté n° 693/MSPRH/MIN).
- **Rôle régional :** Couverture sanitaire étendue Ã 3 wilayas (Béjaia, Sétif, Jijel).

Service des Urgences de l'EPH de Kherrata

Présentation Générale

Le service des urgences de l'EPH de Kherrata est une unité médico-chirurgicale opérationnelle 24h/24 qui assure la prise en charge initiale des urgences vitales et relatives pour une population de 131 000 habitants (avec extension aux wilayas limitrophes).

Capacité et Organisation

- **Surface totale :** 450 m^2
- **Lits organisés :**20 (dont 10 de réanimation)
- **Zones fonctionnelles :**
  - — Accueil et triage.
  - — Box d'examen (3).
  - — Zone de réanimation (10 lits).
  - — Salle de déchocage.

Délais de prise en charge

Niveau	Couleur	Degrés d'urgence	Temps d'attente
U1-1 !	Bleu	Réanimation	Immédiat
U2-2 !	Rouge	Trés urgent	15 min
U3-3 !	Orange	Urgent	30 min
U4-4 !	Vert	Moins urgent	60 min

TABLE 1.5 – Classification des niveaux d'urgence à l'EPH de Kherrata

Moyens Techniques

- **Équipements principaux :**
  - 2 moniteurs cardiaques
  - 1 défibrillateur
  - 1 respirateur portable
  - 1 échographe d'urgence
  - Laboratoire d'analyses rapides
- **Médicaments essentiels :**
  - Armoire à pharmacie d'urgence
  - Kit de réanimation
  - Antidotes spécifiques

Ressources Humaines

Effectifs permanents

Médecins	Effectifs	Paramédicaux	Effectifs
Urgentistes	2	Infirmiers DE	5
Généralistes	3	Aides-soignants	3
Chirurgien de garde	1	Ambulanciers	2

TABLE 1.6 – Répartition du personnel médical et paramédical affecté au service des urgences

Activité Annuelle

— **Passages annuels :** 25 000 (soit 70/jour)

— **Répartition :**

Service	Pourcentage
Médecine interne	40%
Traumatologie	35%
Pédiatrie	15%
Gynécologie	10%

TABLE 1.7 – Répartition des patients selon les services d'accueil

— **Taux d'hospitalisation :**30%

— **Taux de transfert vers CHU :** 15% (cas complexes)

Particularités

— **Axe routier dangereux :** Prise en charge privilégiée des accidents de la route (tunnel de Kherrata).

— Plateforme de téléconsultation avec le CHU de Béjaia.

— Unité d'observation pour les cas nécessitant ¡24h d'hospitalisation.

Défis et Perspectives

- **Défis actuels :**
  - — Saturation périodique (weekends et été).
  - — Ressources limitées en personnel spécialisé.
  - — Équipements à renouveler.
- **Projets 2024 :**
  - — Extension physique (+5 boxes).
  - — Formation continue du personnel.
  - — Acquisition d'un nouveau défibrillateur.

1.2 Centre Hospitalo-Universitaire (CHU) Khelil Amrane Béjaia

FIGURE 1.2 – CHU de bejaia

Présentation Institutionnelle

Généralités :

- **Statut** : Établissement Public Hospitalier (EPH) à caractére universitaire
- **Année de création :** 2009 (décret exécutif n° 09-94)
- **Superficie :** 62 000 m^2 couverts
- **Population couverte :** Wilaya de Béjaia (1,5 million d'habitants) et wilayas limitrophes

Organisation Structurelle

- **Directions Opérationnelles :**
  - Direction des Soins
  - Direction de la Formation et de la Recherche
  - Direction de l'Administration et des Finances
  - Direction des Plateaux Techniques
- **Pôles Médicaux (selon le site officiel) :**
  - Pôle Médico-Chirurgical
  - Pôle Urgences-Réanimation
  - Pôle Spécialités Médicales
  - Pôle Imagerie et Biologie

Services Cliniques (Détail)

Services	Lits
Urgences Medico-Chr	26
Neurochirurgie	43
Anesthésie réanimation	12
Soin intensif	7
Médecine interne	24
Cardiologie	21
Orthopédie	16
Gastro-entérologie	21
Oncologie	28
Chirurgie générale	26
Laboratoires centrales	0
Imagerie médicale	3

TABLE 1.8 – Répartition des lits par service au CHU Khelil Amrane

Plateaux Techniques

- **Imagerie Médicale :**
  - 1x scanner
  - 3x Radiologies interventionnelle
  - 1x échographie
  - 1x mammographie
  - 1x IRM
- **Laboratoires :**
  - Biochimie automatisée
  - Microbiologie moderne
  - Anatomopathologie
- **Autres :**
  - Pharmacie centrale robotisée
  - Stérilisation centrale
  - Banque du sang

Mission Universitaire

- **Formation :**
  - 150 internes accueillis annuellement
  - 12 spécialitÃ©s médicales enseignées
  - Partenariat avec la Faculté de Médecine de Béjaia
- **Recherche :**
  - 5 laboratoires de recherche agréés
  - 20 publications annuelles en moyenne
  - Projets européens (Horizon 2020)

Ressources Humaines

- **Effectif 2024 :**
  - 377 paramédicaux
- **Ratio :**
  - 1 médecin/6 000 habitants
  - 1 infirmier/250 patients

Projets en Cours

- **Extension :** Nouveau bâtiment de 200 lits (livraison 2025)
- **Numérique :** Dossier patient informatisé complet
- **Équipements :** Acquisition d'un robot chirurgical Da Vinci

Service des Urgences du CHU Khelil Amrane Béjaia

Présentation Générale

Le service des Urgences du CHU Khelil Amrane constitue la porte d'entrée principale de l'etablissement, opérationnel 24h/24 et 7j/7. Il assure la prise en charge de toutes les pathologies aigues selon une organisation rigoureuse conforme aux standards internationaux.

Capacité et Infrastructure

- **Capacité d'accueil :**
  - 9x Boxes de consultation.
  - Salle d'échocage
  - Salle de plâtre
  - 5x salle d'observation
  - Salle de soin

- **Secteurs spécialisés :**
  - 3x Boxes de consultation.
  - Salle de soin
  - Salle d'échocage

Organisation du Trafic

Niveau	Couleur	Degrés d'urgence	Temps d'attente
U1-1 !	Bleu	Réanimation	Immédiat $<5min$
U2-2 !	Rouge	Trés urgent	15 min
U3-3 !	Orange	Urgent	30 min
U4-4 !	Vert	Moins urgent	60 min

TABLE 1.9 – Classification des niveaux d'urgence au CHU Khelil Amrane

Moyens Techniques

- **Équipements de pointe :**
  - 6 moniteurs multi paramétres
  - 2 échographes dédiés
  - 1 scopie mobile
  - 1 laboratoire d'analyses rapides (Troponine, gaz du sang)
- **Médicaments :**
  - Armoire à pharmacie sécurisée
  - Stock d'antidotes et thrombolytiques

Ressources Humaines

- **Effectifs permanents :**
  - 5 médecins généraliste
  - 9 infirmiers
  - 12 aides-soignants
  - 2 psychologues
- **Disponibilité :**
  - Chirurgien de garde
  - Cardiologue référent
  - Neuro-vasculaire

Particularités

— **Unité AVC :** Thrombolyse sur place 24h/24

— **Lien SMUR :** 2 ambulances médicalisées dédiées

— **Plateforme télé-expertise :** Consultation spécialisée à distance

Projets 2025-2026

- **Digitalisation :**
  — Dossier patient informatisé
  — Gestion électronique du trafic
- **Formation :**
  — Certification européenne des infirmiers
  — Simulation haute-fidélité

1.3 Problématique

Pendant notre stage, qui comprenait des enquêtes auprés des patients et du personnel hospitalier aux bureaux de tri des urgences, nous avons travaillé au CHU Khelil-Amrane du 2 mars au 12 mars 2025 et à l'EPH de Kherrata du 8 avril au 27 avril 2025. Nous avons remarqué que les patients font face à de longs temps d'attente aux urgences. Ces attentes prolongées augmentent le stress des patients et nuisent à la relation entre les patients et le personnel médical, ce qui peut affecter la qualité des soins et des diagnostics. Plusieurs facteurs locaux aggravent cette situation,comme le manque de centres de soins primaires accessibles répondant aux attentes des patients, la concentration des ressources médicales dans les grands hôpitaux, et l'augmentation des maladies chroniques et des accidents. Cela entraîne une affluence croissante de patients, y compris ceux qui n'ont pas besoin de soins urgents.

De plus, la perception d'une mauvaise gestion des ressources hospitaliéres, marquée par un manque de personnel médical qualifié, des équipements insuffisants et une organisation inadaptée aux variations de fréquentation, aggrave le sentiment de dysfonctionnement généralisé. Il est donc crucial de trouver des solutions scientifiques pour optimiser la gestion de ces établissements hospitaliers et améliorer l'efficacité des soins, en particulier dans les services d'urgence que nous avons étudiés.Nous avons formulé notre problématique comme suit : Comment utiliser les techniques de Data Mining, largement employées dans les pays développés, pour prédire les temps d'attente et prioriser les admissions dans un centre hospitalier universitaire (CHU), en tenant compte du contexte spécifique de l'EPH de Kherrata et du CHU Khelil-Amrane de Béjaia ?

Conclusion

Nous concluons que la nécessité d'améliorer la fluidité des flux de patients, de réduire les inégalités d'accés aux soins et de renforcer la prise de décision opérationnelle dans les services d'urgence justifie notre problématique. Les données brutes recueillies dans ces deux établissements nous ont permis de développer des modéles prédictifs et classificateurs, ainsi qu'un tableau de bord interactif utilisant des outils comme Random Forest, K-Means et Dash/Plotly. L'objectif final est de concevoir une solution basée sur les données pour améliorer l'organisation hospitaliére et la qualité des soins prodigués.

2

Revue de littérature scientifique et cadre théorique

Sommaire

Introduction

Nous posons dans ce chapitre, les bases conceptuelles et scientifiques nécessaires à la compréhension et a la mise en oeuvre du projet. Il se compose d'une revue de littérature sur les travaux académiques existants concernant l'utilisation du Data Mining dans le domaine hospitalier, d'un cadre théorique présentant les concepts clés avec les techniques utilisées (régression, classification, clustering) ainsi que leur application au contexte des admissions aux urgences. En s'appuiant sur des recherches récentes ainsi que sur des outils éprouvés tels que Random Forest , K-Means , Prophet , et LSTM , tous intégrés dans un processus rigoureux suivi dans le cadre du modéle CRISP-DM .

2.1 Revue la littérature

2.1.1 Utilisation du Data Mining dans le secteur médical

Le Data Mining est une discipline interdisciplinaire qui combine statistiques, informatique et intelligence artificielle pour extraire des informations utiles à partir de grandes quantités de données brutes. Son objectif est de découvrir des motifs, tendances et relations cachées dans les données, afin de faciliter la prise de décision. « Le DM est un processus visant à découvrir des patterns significatifs dans les données à grande échelle. » [13]
Dans le domaine de la santé, le Data Mining trouve des applications variées [13] :

— Prédiction des pathologies.
— Analyse des résultats cliniques.
— Gestion des flux de patients.
— Optimisation des ressources hospitaliéres.

2.1.2 Travaux antérieurs sur la gestion des flux hospitaliers

Modélisation prédictive des temps d'attente

— Sun et al. (2013) ont utilisé des modéles de régression linéaire pour prédire la durée de séjour des patients aux urgences.
— « Les modéles de régression linéaire sont couramment utilisés pour prédire des variables continues comme les temps d'attente dans les services d'urgence. » [23]

Priorisation des cas urgents

— Provost et Fawcett (2013) ont montré l'efficacité des algorithmes tels que Random forest ou SVM pour classer les patients selon leur niveau de gravité.
— « Ces techniques permettent une prise en charge plus rapide des cas critiques grâce à une priorisation automatisée. » [20]

Détection des goulots d'étranglement

— Kumar et Ozdamar (2004) ont appliqué des simulations Monte Carlo pour analyser les flux de patients et proposer des ajustements organisationnels.
— « La simulation offre un outil puissant pour tester différents scénarios opérationnels sans impacter le systéme réel. » [8]

2.1.3 Applications spécifiques en Algérie

Surcharge des urgences dans les CHU algériens

— Boukhalfa et Moussaoui (2017) ont analysé les causes de la surpopulation dans les urgences des CHU en Algérie.
— « L'afflux massif de patients non urgentistes et la centralisation des ressources médicales aggravent la situation dans les CHU algériens. » [9]

Impact de la gestion inefficace des ressources

— Benamrane Belkacem (2020) ont montré que le manque de personnel qualifié et d'infrastructures adaptées conduit à une dégradation globale de la qualité des soins.
— « Un personnel insuffisant ou mal réparti ralentit la prise en charge des patients. » [6]

2.2 Concepts fondamentaux du Data Mining

2.2.1 Définition du Data Mining

Le DM est « un processus visant à découvrir des patterns significatifs dans les données Ã grande échelle » [13].
Le DM consiste donc à extraire des connaissances implicites à partir de grands ensembles de données, souvent complexes et hétérogénes. Cette discipline s'appuie sur des méthodes issues de la statistique, de l'intelligence artificielle et de la gestion des bases de données.

2.2.2 Domaines d'application dans le secteur hospitalier

Le DM est particuliérement pertinent dans le domaine médical, notamment :
— Analyse des dossiers médicaux anonymisés.
— Prédiction des admissions et temps d'attente.
— Priorisation des patients par niveaux de gravité.
— Gestion des ressources (personnel, matériel, salles).
— L'optimisation des plannings et des étapes de triage.

2.2.3 Techniques clés du Data Mining utilisées dans ce projet

Technique, Description, Application dans le projet :

Technique	**Description**	**Application dans le projet**
Régression multiple	- Prédiction d'une variable continue à partir de plusieurs variables explicatives	- Estimation des temps d'attente
Classification (Random Forest, SVM, KNN)	- Attribution d'étiquettes à des observations	- Priorisation des patients selon la gravité
Clustering (K-Means, DBSCAN)	- Regroupement des données similaires sans étiquettes	- Segmentation des patients selon leurs caractéristiques
Séries temporelles	- Modélisation de données ordonnées dans le temps	- Prévision des pics d'affluence
Analyse exploratoire des données (EDA)	- Exploration des données pour identifier tendances et corrélations	- Compréhension des flux de patients

2.3 Techniques de Data Mining et Algorithmes Utilisés

2.3.1 Régression multiple

La RM est une méthode statistique utilisée pour modéliser la relation entre une variable cible continue (comme le temps d'attente) et plusieurs variables explicatives.
Objectif dans le projet : Prédiction des temps d'attente des patients.

- **Variables explicatives possibles :**
  - Heure d'arrivée
  - Type de pathologie
  - Gravité du cas
  - Disponibilité du personnel
  - Données externes (météo, événements locaux)
- **Avantages :**
  - Facile à comprendre et à implémenter.
  - Permet une bonne interprétation des facteurs influenccant le temps d'attente.
- **Limites :**
  - Repose sur l'hypothése de linéarité entre les variables.
  - Sensible au surapprentissage si trop de variables sont incluses.
- **Métrique d'évaluation :**
  - RMSE (Root Mean Squared Error) pour mesurer la précision des prédictions.

2.3.2 Classification supervisée

La classification supervisée est une technique d'apprentissage automatique visant à affecter une étiquette (classe) à chaque observation à partir d'un ensemble de données labellisées.

Principaux algorithmes utilisés

- **Random Forest :** est une technique qui combine plusieurs arbres de décision pour améliorer les performances et réduire le surapprentissage. Il agrége les prédictions de multiples arbres pour obtenir une prédiction finale.parmi ces avantages :
  - Les forêts aléatoires sont robustes au bruit et aux valeurs aberrantes et peuvent gérer des données avec un grand nombre de variables d'entrée. [3]
- **Support Vector Machine (SVM) :**est une technique d'apprentissage automatique supervisé qui utilise des machines à vecteurs de support (SVM) pour classer des données. Elle fonctionne en trouvant le meilleur hyperplan qui sépare les différentes classes de données dans un espace multidimensionnel, en maximisant la marge entre l'hyperplan et les points de données les plus proches de chaque classe (vecteurs de support).
  Les SVM représentent plusieurs avantages, notamment ceux-ci :
  - Elles ont une base théorique solide [2] .
  - Les SVM sont efficaces dans les espaces de grande dimension [15].

— Différentes fonctions noyau peuvent être spécifiées [15]

Malgré leurs performances, les SVM représentent aussi des faiblesses, notamment celles-ci :

— Elles utilisent des fonctions mathématiques complexes pour la classification [15].

— Les SVMs demandent un temps énorme durant les phases de test [2].

- **K-Nearest Neighbors (KNN) :**est un algorithme qui peut être utilisé à la fois pour des problémes de classification et de régression. Cependant, il est plus largement utilisé dans les problémes de classification dans l'industrie. KNN est un algorithme simple qui stocke tous les cas disponibles et classe les nouveaux cas par un vote majoritaire de ses k voisins. Le cas attribué à la classe est le plus courant parmi ses K voisins les plus proches mesurés par une fonction de distance. Ces fonctions de distance peuvent être la distance euclidienne, Manhattan, Minkowski et Hamming. Les trois premiéres fonctions sont utilisées pour la fonction continue et la quatriéme (Hamming) pour les variables catégorielles.

  — Si K = 1, le cas est simplement affecté à la classe de son plus proche voisin. Parfois, choisir K s'avére être un défi lors de la modélisation KNN [1].

  — **Objectif dans le project :**
    - Classer les patients en trois niveaux de priorité (faible, moyenne, élevée), en fonction de leur âge, sexe, pathologie, gravité, antécédents, etc.

  — **Avantages :**
    - Efficace pour des décisions rapides (triage).
    - Bonne précision lorsque les données sont bien étiquetées

  — **Limites :**
    — Besoin de données labellisées (gravité connue)
    — Complexité accrue avec un grand nombre de classes ou de features

  — **Métriques d'évaluation :**
    — Matrice de confusion.
    — Précision, Rappel, F1-score
    — AUC (Area Under Curve)

2.3.3 Clustering (non supervisé)

Le clustering est l'affectation d'un ensemble d'observations en sous-ensembles (appelés clusters) de sorte que les observations au sein d'un même cluster soient similaires selon un ou plusieurs critéres pré désignés, tandis que les observations tirées de clusters différents sont différentes. Différentes techniques de clustering font différentes hypothéses sur la structure des données, souvent définies par une métrique de similitude, tel que la compacité interne, ou la similitude entre les membres d'un même cluster, ou la différence entre les clusters. D'autres méthodes sont basées sur une densité estimée et une connectivité graphique. Les algorithmes de clustering les plus connus sont K-Means et le Mean-Shift Clustering [4].

Principaux algorithmes utilises

- **K-Means :** Il s'agit d'un type d'algorithme non supervisé qui résout le probléme de clustering. Sa procédure suit une méthode simple et facile de classer un ensemble de données à travers un certain nombre de clusters (supposons k clusters). Les points de données à l'intérieur d'un cluster sont homogénes et hétèrogénes aux autres groupes [21]
  - **Parmi ces avantages :**
    On peut citer les avantages suivants : [10, 24]
    - Il est simple, facile à comprendre et à implémenter.
    - Il est applicable à des données de grandes tailles : K-means convient à un grand nombre d'ensembles de données
    - Insensible à l'ordre des données.
    - Un objet peut être affecté à une classe au cours dâ€™une itération puis changer de classe à l'itération suivante, ce qui n'est pas possible avec la classification ascendante hiérarchique pour laquelle une affectation est irréversible.
  - **Parmi ces inconvénient** [24, 17]
    - Les centres des clusters, mis à part des centres initiaux, sont des objets inexistants puisqu'ils correspondent à des moyennes calculées sur un sous-ensemble d'observations à chaque itération.
    - Le résultat final dépend fortement du choix des centroides initiaux.
    - Le nombre de classes est un paramétre de l'algorithme. Un bon choix du nombre k est nécessaire, car un mauvais choix de k produit de mauvais résultats (Une forte influence des valeurs aberrantes sur les résultats).
    - Il n'est pas applicable aux données non numériques
    - Il est difficile de prévoir les valeurs k ou le nombre de clusters. Il est également difficile de comparer la qualité des clusters produits.
- **Clustering hiérarchique :**L'algorithme de clustering hiérarchique est un autre algorithme d'apprentissage automatique non supervisé, qui est utilisé pour regrouper les points de données non étiquetés ayant des caractéristiques similaires dans un cluster [14]. L'avantage de cette méthode est qu'elle n'est soumise à aucune initialisation particuliére de paramétre(s) ce qui la rend déterministe, et en outre, que le nombre de classe n'a pas à être fixé a priori [7]. Dans cet algorithme, nous développons la hiérarchie des clusters sous la forme d'un arbre, et cette structure en forme d'arbre est connue sous le nom de dendrogramme.
  La technique de clustering hiérarchique a deux approches [19] :
  - **Approche agglomérative :**Dans les algorithmes hiérarchiques agglomératifs, chaque point de données est traité comme un seul cluster, puis le processus fusionne ou agglomére successivement les paires de clusters (approche ascendante). La hiérarchie des clusters est représentée sous la forme d'un dendrogramme ou d'une arborescence [7].
  - **Approche divisive :** Dans cette approche, tous les points de données (individus) sont considérés comme une seule classe au début, et le processus de clustering

divise successivement les classes en classes plus raffinées (approche descendante). Le processus marche jusqu'à ce que chaque classe contienne un seul point ou bien si l'on atteint un nombre de classes désiré [19].

- **DBSCAN (Density-Based Spatial Clustering of Applications with Noise) :**
  - L'algorithme de regroupement spatial basé sur la densité d'applications avec bruit (DBSCAN) est un algorithme non supervisé trés connu en matiére de clustering. Il a été proposé par Martin Ester et al. en 1996[11] .
  - C'est un algorithme de clustering basé sur la densité (La densité d'un objet peut être calculée par le nombre d'objets proche de celui-ci) qui fonctionne sur l'hypothése que les clusters sont des régions denses dans l'espace séparées par des régions de densité inférieure. DBSCAN recherche les objets principaux, c'est à dire les objets qui ont des voisinages denses. Il relie les objets centraux et leurs voisins pour former des régions denses appelés clusters. Il regroupe les points de données densément connectés en un seul cluster. Il peut identifier les clusters dans de grands ensembles de données spatiales en examinant la densité locale des points de données. La caractéristique la plus intéressante du clustering DBSCAN est qu'il est robuste aux valeurs aberrantes. Il ne nécessite pas non plus que le nombre de clusters soit indiqué à l'avance [22], contrairement aux K-means, oú nous devons spécifier le nombre de centro ides.
  - L'algorithme DBSCAN ne nécessite que deux paramétres : epsilon et minPoints. Epsilon est le rayon du cercle à créer autour de chaque point de données pour vérifier la densité et minPoints est le nombre minimum de points de données requis à l'intérieur de ce cercle pour que ce point de données soit classé comme point central [22]. Dans les dimensions supérieures, le cercle devient hypersphére, epsilon devient le rayon de cette hypersphére et minPoints est le nombre minimum de points de données requis à l'intérieur de cette hypersphére [22].
  - **Objectif dans le projet :**
    - Segmenter les patients en groupes homogénes pour personnaliser la gestion des flux et améliorer la planification.
  - **Avantages :**
    - Aucun besoin d'étiquettes.
    - idéal quand les données ne sont pas encore segmentées.
    - Détecte des structures inconnues dans les données.
  - **Limites :**
    - Difficulté à valider sans données labellisées.
    - Choix du nombre de clusters peut être subjectif

2.3.4 Séries temporelles

Les modéles de séries temporelles sont utilisés pour analyser des données ordonnées dans le temps. Ils permettent d'identifier :

- Des tendances.

— Des saisonnalités.
— Des cycles répétitifs.

Principales méthodes

- **ARIMA (AutoRegressive Integrated Moving Average) :**ARIMA est un outil d'analyse est essentiellement utilisé à des fins de prévision des valeurs futures, de détermination des valeurs manquantes dans une série de points ou d'identification de la structure de la série temporelle. Un modéle ARIMA s'exprime en fonction de la notation ARIMA (p, d, q) oú p, d, q renvoient au nombre (ou l'ordre) de termes autorégressifs, de différenciation et de moyenne mobile dans le modéle ARIMA final. (Par exemple : un modéle ARIMA (1,0,0) est un modéle qui comprend un terme autorégressif et aucun terme de différenciation ni de moyenne mobile).
  Si une série chronologique a des modéles saisonniers, vous devez ajouter des termes saisonniers et elle devient SARIMA, abréviation de « Seasonal ARIMA ».parmi ces avantages :
  — Adapté aux séries stationnaires.
  — Bonne performance pour les prévisions à court terme.
- **LSTM (Long Short-Term Memory) :**La mémoire à long court terme (LSTM) est une architecture de réseau neuronal récurrent (RNN) spécifique qui a été conccue pour modéliser les séquences temporelles et leurs dépendances à longue portée avec plus de précision que les RNN conventionnels [18]. Les LSTM sont un type spécial de RNN, capable d'apprendre les dépendances à long terme et de mémoriser des informations pendant des périodes prolongées par défaut. Selon Olah, le modéle LSTM est organisé sous la forme d'une structure en chaîne. Cependant, le module répétitif a une structure différente. Au lieu d'un seul réseau neuronal comme un RNN standard, il comporte quatre couches d'interaction avec une méthode de communication unique [18].parmi ces avantages :
  - Réseau de neurones récurrent (Deep Learning).
  - Capable de capturer des dépendances temporelles longues.
- **Prophet (Facebook) :** Prophet est une procédure de prévision des données de séries chronologiques basée sur un modéle additif oú les tendances non linéaires sont adaptés à la saisonnalité annuelle, hebdomadaire et quotidienne, ainsi qu'aux effets des vacances. Cela fonctionne mieux avec des séries chronologiques qui ont de forts effets saisonniers et plusieurs saisons de données historiques. Prophet est robuste aux données manquantes et aux changements de tendance, et gére généralement bien les valeurs aberrantes [5].
  Avantages :
  - Facile à utiliser et interprétable.
  - Idéal pour les données avec saisonnalités fortes.

  Objectif dans le projet :
  - Anticiper les flux quotidiens/semaines/mois dans les urgences en analysant les variations historiques (ex. : pic d'activité pendant les mois d'hiver).

Applications pratiques :

- Planification proactive du personnel.
- Anticipation des besoins en matériel et en salles d'examen.

Limites :

- Nécessite un jeu de données historique fiable.
- Peu adapté en cas de changements imprévus.

Métrique d'évaluation :

— MAPE (Mean Absolute Percentage Error), RMSE

2.3.5 Réseaux en Machine Learning

En machine learning (ML), les réseaux font référence à des structures de données ou d'algorithmes inspirées du fonctionnement biologique du cerveau humain. Ces modéles sont capables de capturer des relations complexes entre des entrées et des sorties, notamment dans des tâches comme la reconnaissance d'image, le traitement du langage naturel ou la prédiction temporelle. Le terme "réseau" est principalement associé aux Réseaux de Neurones Artificiels (RNA), mais il peut aussi englober d'autres types de structures comme les réseaux bayésiens, les réseaux markoviens, etc.

— **Définition des Réseaux de Neurones Artificiels (ANN - Artificial Neural Networks)**
Les réseaux de neurones artificiels sont des modéles algorithmiques composés de couches interconnectées de "neurones", qui imitent le comportement des neurones biologiques. [12]
Structure classique d'un réseau de neurones :
- **Couche d'entrée :** Reccoit les données brutes
- **Couches cachées :** Traitent les informations via des transformations non linéaires
- **Couche de sortie :** Fournit la prédiction finale Chaque connexion entre deux neurones a un **poids**, qui est ajusté pendant l'entraînement pour minimiser l'erreur du modéle.

— **Types de Réseaux de Neurones Populaires :**
  — Réseaux de Neurones Multicouches (MLP - Multi-Layer Perceptron) :
    - Architecture simple avec plusieurs couches cachées
    - Utilisé pour la classification et la régression
    - Exemple : Reconnaissance de chiffres manuscrits (MNIST)
  — Réseaux de Neurones Convolutifs (CNN - Convolutional Neural Networks) :
    - Spécialisés dans le traitement d'images
    - Utilisent des filtres (convolutions) pour extraire des caractéristiques locales
    - Exemples : VGGNet, ResNet, YOLO
  — Réseaux de Neurones Récursifs (RNN - Recurrent Neural Networks) :
    - Conccus pour traiter des séquences (texte, audio, temps)
    - Mémoire intégrée pour conserver l'information précédente

  - Variante populaire : LSTM (Long Short-Term Memory)
- Réseaux Générateurs Adversaires (GAN - Generative Adversarial Networks) :
  - Composés de deux réseaux concurrents : un générateur et un discriminateur
  - Utilisés pour générer des images, textes, voix synthétiques

— **Avantages des Réseaux de Neurones :**

- **Apprentissage automatique :** Capables d'apprendre à partir de données sans programmation explicite
- **Généralisation :** Bonnes performances sur des données inconnues aprés entraînement
- **Flexibilité :** Adaptés à divers domaines (vision, NLP, finance, etc.)

— **Limites et Défis :**

- **Données massives nécessaires :** Beaucoup de données pour bien s'entraîner
- **Temps de calcul élevé :** Entraînement long, surtout avec des architectures profondes
- **Interprétabilité faible :** Difficile d'expliquer comment un réseau arrive à sa décision (« boîte noire »)
- **Surapprentissage (overfitting) :** Risque d'adaptation trop forte aux données d'entraînement

— **Applications Réelles :**

Domaine	**Applications des réseaux de neurones**
Vision par ordinateur	Détection d'objets, reconnaissance faciale, segmentation d'images
Santé	Diagnostic médical assisté, analyse d'imagerie médicale, prédiction de maladies
Finance	Prévision boursiére, détection de fraude, scoring de crédit
Marketing	Recommandation de produits, segmentation de clientéle, personnalisation de contenu
Transport	Conduite autonome, prévision de trafic, gestion intelligente des itinéraires

TABLE 2.1 – Exemples d'applications des réseaux de neurones par domaine

2.4 Cadre théorique du projet

2.4.1 Modélisation prédictive dans les urgences hospitaliéres

Prédiction des temps d'attente

Pour répondre à la question : Comment prédire les temps d'attente des patients en fonction des données historiques ? Nous avons opté pour des modéles de régression multiple et d'analyse des séries temporelles. Ces modéles intégrent des données qu'on appelle :
Variables explicatives :

- Heure d'arrivée
- Type de pathologie
- Gravité du cas
- Disponibilité du personnel
- Données météorologiques ou événements exceptionnels

Priorisation des patients

Pour répondre à la question : Comment prioriser efficacement les patients en fonction de la gravité de leur état ? Nous avons mis en place un modéle de classification supervisée, basé sur des algorithmes comme :

- **Random Forest :** robuste face aux données déséquilibrées.
- **SVM :** efficace pour des frontiéres de décision complexes.
- **KNN :** basé sur la similarité avec des cas passés.
- **EWS Early Warning Score :** Scores médicaux standardisés

Ce modéle classe les patients en trois catégories : faible priorité, moyenne priorité, haute priorité, pour une meilleure gestion du triage.

Détection des goulots d'étranglement

Avec l'aide du clustering (notamment K-Means), nous identifions les segments de patients qui gènérent des retards constants (patients chroniques, affections mineures,etc...).Cette segmentation permet de :

- Comprendre les comportements répétitifs.
- Mettre en place des actions ciblées pour fluidifier les flux.

2.5 Systémes d'aide à la décision en santé

2.5.1 Tableau de bord interactif

Objectifs

- Visualiser les indicateurs clés en temps réel.
- Faciliter la prise de décision par les gestionnaires hospitaliers.

Outils possibles

- Power BI
- Tableau
- Dash/Plotly (en Python)

2.5.2 Indicateurs clés de performance (KPIs)

- Temps d'attente moyen
- Taux d'occupation des salles d'examen
- Nombre de patients traités par heure
- Niveau de gravité moyen par tranche horaire

2.6 Cadre méthodologique CRISP-DM

Le projet s'appuie sur le cadre CRISP-DM (Cross-Industry Standard Process for Data Mining), largement adopté dans les projets de data science dans les domaines médicaux et hospitalier.

2.6.1 Etapes du processus CRISP-DM

— **Compréhension métier :** Identifier les besoins du CHU (réduire les temps d'attente, prioriser les urgences).

— **Compréhension des données :** Collecte des données hospitaliéres, contrôle de la qualité et identification des sources.

— **Préparation des données :** Nettoyage, transformation, enrichissement par fusion avec des données externes (météo, événements).

— **Modélisation :** Entraînement des modéles prédictifs (régression, classification, clustering).

— **Evaluation :** Validation des modéles via RMSE, AUC, matrice de confusion.

— **Déploiement :** Intégration des résultats dans un systéme opérationnel (tableau de bord interactif).

2.7 Apports du Data Mining à la gestion hospitaliére

2.7.1 Avantages

- Anticipation des pics d'affluence.
- Meilleure allocation des ressources (personnel, matériel).
- Amélioration de la qualité des soins grâce à une priorisation efficace.
- Réduction des temps d'attente et amélioration de la satisfaction des patients.

2.7.2 Limites

- Qualité des données (manque de fiabilité, anonymisation).
- Complexité des modéles pour les acteurs non spécialisés.
- Résistance au changement dans certains environnements hospitaliers.

2.8 Synthése et justification des choix méthodologiques :

Ce chapitre a permis de situer notre travail dans un cadre théorique solide, appuyé sur des recherches existantes et des techniques éprouvées de Data Mining. Les choix méthodologiques (modéles, variables, outils) s'appuient sur une revue approfondie de la littérature scientifique et une adaptation aux spécificités du contexte algérien.

Conclusion

La présente revue de littérature a permis d'identifier les techniques et outils pertinents pour répondre à la problématique des temps d'attente élevés dans les CHU. Elle a également mis en lumiére les défis propres au contexte local, tout en fournissant les fondations théoriques nécessaires à la suite du mémoire. Ces éléments serviront de base pour la mise en oeuvre pratique du modéle prédictif présenté dans le chapitre suivant.

3
Méthodologie

Sommaire

Introduction

Nous allons détailler la méthodologie utilisée pour réaliser notre projet intitulé ”Le Data Mining pour gérer les admissions dans un CHU (ou EPH) : Prédiction des temps d’attente et priorisation des cas urgents”. Notre approche s’appuie sur le cadre CRISP-DM (Cross-Industry Standard Process for Data Mining), une méthode largement reconnue et utilisée dans les projets de data science appliquée.

Nous disposons de données provenant de deux établissements distincts. Nous commencerons par analyser les données de **l’EPH de Kherrata**, qui offre un volume de données plus important. Ensuite, nous traiterons les données du **CHU Khelil Amrane** afin de permettre une comparaison entre les deux établissements.

Les étapes principales que nous avons suivi sont : La collecte et la compréhension des données, la préparation et le nettoyage des données, l’analyse exploratoire des données (EDA), la feature engineering, la modélisation prédictive (pour la régression et la classification), la validation croisée des modéles puis leur validation et enfin le déploiement qui sera sous forme de « tableau de bord Dash ».

3.1 EPH de Kherrata

3.1.1 Collecte et compréhension des données

Source des données

Les données que nous avons utilisées proviennent de l'EPH de Kherrata et couvrent la période du 8 avril au 27 avril 2025, de 8h à 18h chaque jour. Elles concernent les patients admis aux urgences. Chaque patient arrivant au bureau de tri médical des urgences a été invité à répondre à un questionnaire. Ce questionnaire visait à faciliter l'évaluation de la gravité de la pathologie pour laquelle le patient consultait.
Le questionnaire comprenait les informations suivantes : âge, sexe, date et heure d'arrivée, date et heure de passage, pathologie, et présence de maladies chroniques. Une fois le questionnaire rempli, une évaluation de la gravité de la pathologie était effectuée et exprimée par la couleur d'un ticket : rouge pour les cas urgents, orange pour les cas intermédiaires, et vert pour les cas non urgents. Toutes ces informations étaient ensuite enregistrées dans un tableau, avec un identifiant unique attribué à chaque patient.Une fois le patient est passé en consultation, on renseigne sur le tableau,la date et l'heure de passage, le médecin consultant, la salle de consultation, l'équipe soignante constituée d'infirmiers et d'aides soignants.

3.1.2 Nettoyage et préparation des données

Nous avons importé plusieurs bibliothéques et modules Python qui sont couramment utilisés pour l'analyse de données, le machine learning, et la création de visualisations interactives :

- **pandas :** Importe la bibliothéque **Pandas**, qui est utilisée pour la manipulation et l'analyse de données. Elle fournit des structures de données comme les DataFrames, qui sont essentielles pour travailler avec des données tabulaires.
- **numpy :** Importe la bibliothéque **NumPy**, qui est utilisée pour les calculs numériques. Elle fournit des outils pour travailler avec des tableaux multidimensionnels et des fonctions mathématiques avancées.
- **matplotlib.pyplot :** Importe le module **Pyplot de Matplotlib**, une bibliothéque utilisée pour créer des graphiques et des visualisations statiques en 2D.
- **seaborn :** : Importe la bibliothéque **Seaborn**, qui est construite sur Matplotlib et offre une interface plus simple pour créer des graphiques statistiques attrayants et informatifs.
- **datetime :** Importe la classe **datetime** du module **datetime**, qui est utilisée pour manipuler les dates et les heures.
- **from sklearn.model_selection import train_test_split :** Importe la fonction **train_test_split** de scikit-learn, qui est utilisée pour diviser un jeu de données en ensembles d'entraînement et de test.
- **from sklearn.ensemble import RandomForestRegressor :** Importe la classe **RandomForestRegressor** de scikit-learn, qui est utilisée pour créer des modéles de régression basés sur des forêts aléatoires.
- **from sklearn.ensemble import RandomForestClassifier :** Importe la classe **RandomForestClassifier** de scikit-learn, qui est utilisée pour créer des modéles de classification basés sur des forêts aléatoires.

- **from sklearn.metrics import mean_squared_error :** Importe la fonction **mean_squared_error** de scikit-learn, qui est utilisée pour calculer l'erreur quadratique moyenne, une métrique courante pour évaluer les modéles de régression.
- **from sklearn.metrics import r2_score :** Importe la fonction **r2_score** de scikit-learn, qui est utilisée pour calculer le coefficient de détermination R2, une métrique pour évaluer la performance des modéles de régression.
- **from sklearn.metrics import classification_report :** Importe la fonction **classification_report** de scikit-learn, qui est utilisée pour générer un rapport texte sur les principales métriques de classification.
- **from sklearn.metrics import confusion_matrix :** Importe la fonction **confusion_matrix** de scikit-learn, qui est utilisée pour calculer la matrice de confusion, une métrique pour évaluer la performance des modéles de classification.
- **from sklearn.cluster import KMeans :** Importe la classe **KMeans** de scikit-learn, qui est utilisée pour effectuer le clustering K-means, une méthode de clustering non supervisé.
- **from sklearn.preprocessing import StandardScaler :** Importe la classe **StandardScaler** de scikit-learn, qui est utilisée pour standardiser les caractéristiques en enlevant la moyenne et en mettant à l'échelle à l'unité de variance.
- **dash :** Importe la classe **Dash** de la bibliothéque Dash, qui est utilisée pour créer une nouvelle application Dash.
- **from dash import dcc :** Importe le module **dcc** (Dash Core Components) de Dash, qui fournit un ensemble de composants pour créer des interfaces utilisateur interactives.
- **from dash import html :** Importe le module **html** de Dash, qui permet d'utiliser des balises HTML dans les applications Dash.
- **plotly.express :** Importe la bibliothéque **Plotly Express**, qui est utilisée pour créer des visualisations interactives et riches en fonctionnalités avec une syntaxe simple et concise.
- **import scipy.stats :**importe le module stats de la bibliothéque scipy, qui est utilisé pour effectuer des tests statistiques.

Chargement des données

On charge les données à partir du tableau Excel « EPH_Kherrata.xlsx », qui contient les informations relatives aux patients admis au service des urgences.

```
dh = pd.read_excel("EPH_Kherrata.xlsx", header=None)
```

L'option header=None est utilisée pour éviter toute erreur liée à l'absence de ligne d'en-tête.

	Patients	Age	Sexe	date_arrivee	heure_arrivee	date_passage	heure_passage	pathologie	maladie_chronique	couleur_ticket	salles_examens	medecin
0	A3080425	27	M	2025-04-08 00:00:00	09:27:00	2025-04-08 00:00:00	09:51:00	angine	NaN	OR	SCA1	MEDJ1
1	A15080425	78	M	2025-04-08 00:00:00	09:30:00	2025-04-08 00:00:00	10:03:00	vertige	NaN	OR	SCA1	MEDJ1
2	A16080425	59	M	2025-04-08 00:00:00	09:43:00	2025-04-08 00:00:00	10:04:00	plaie infectée	NaN	OR	SSA1	MEDJ2
3	E4080425	6	F	2025-04-08 00:00:00	09:44:00	2025-04-08 00:00:00	09:57:00	fievre	NaN	OR	SCE1	MEDJ3
4	E5080425	4	F	2025-04-08 00:00:00	09:45:00	2025-04-08 00:00:00	10:03:00	fievre	NaN	OR	SCE1	MEDJ3

FIGURE 3.1 – Tableau EPH KHERRATA

on a utilisé la fonction **dh.columns** permet d'afficher toutes les colonnes

```
: dh.columns
: Index(['Patients', 'Age', 'Sexe', 'date_arrivee', 'heure_arrivee',
       'date_passage', 'heure_passage', 'pathologie', 'maladie_chronique',
       'couleur_ticket', 'salles_examens', 'medecin', 'equipeS_soignants_ats',
       'evenement'],
      dtype='object')
```

FIGURE 3.2 – Toutes les colonnes du Dataframe EPH KHERRATA

Voici la présentation de chaque colonne :

- **Patients :** identifiant unique du patient.
- **Age :** age du patient au moment de l'admission.
- **Sexe :** sexe du patient (**M** pour masculin, **F** pour féminin).
- **date_arrivee :** date à laquelle le patient est arrivé(e) aux urgences.
- **heure_arrivee :** heure d'arrivée du patient.
- **date_passage :** date à laquelle le patient a été vu par un médecin ou pris en charge.
- **heure_passage :** heure à laquelle le patient a été vu par un médecin.
- **pathologie :** type de pathologie ou symptome rapporté par le patient à l'admission.
- **maladie_chronique :** indique si le patient souffre d'une maladie chronique.
- **couleur_ticket :** couleur du ticket attribué lors du triage, indiquant la priorité médicale
- **salles_examens :** salle d'examen où le patient a été pris en charge.
- **medecin :**médecin qui a pris en charge le patient.
- **equipeS_soignants_ats :** équipe soignante ayant pris en charge le patient.
- **evenement :** informations complémentaires sur l'événement ou l'entrée.

Nettoyage des données

Comme nous avons-nous même manuellement renseigné les données des patients, nous n'avons pas eu de ligne vide.Toutefois,la colonne « maladie_chronique » n'est pas renseigné pour tous les patients, du fait qu'ils ne sont pas tous chroniques.

Préparation des données

- **Préparation et Transformation des Données Temporelles pour l'Analyse des Temps d'Attente :**
  Afin d'obtenir, une variable numérique continue représentant le temps d'attente en minutes, utilisable dans les modéles prédictifs.

```
# Étape 1 : Convertir les colonnes en datetime si nécessaire
dh['date_arrivee'] = pd.to_datetime(dh['date_arrivee']).dt.date
dh['date_passage'] = pd.to_datetime(dh['date_passage']).dt.date

# Étape 2 : Convertir les dates (datetime.date) en str pour pouvoir les concaténer
dh['date_arrivee'] = dh['date_arrivee'].astype(str)
dh['date_passage'] = dh['date_passage'].astype(str)

# Étape 3 : Convertir les heures en str aussi (au cas où)
dh['heure_arrivee'] = dh['heure_arrivee'].astype(str)
dh['heure_passage'] = dh['heure_passage'].astype(str)

# Étape 4 : Combiner date + heure et convertir en datetime complet
dh['datetime_arrivee'] = pd.to_datetime(dh['date_arrivee'] + ' ' + dh['heure_arrivee'])
dh['datetime_passage'] = pd.to_datetime(dh['date_passage'] + ' ' + dh['heure_passage'])

# Étape 5 : Calcul du temps d'attente
dh['temps_attente'] = dh['datetime_passage'] - dh['datetime_arrivee']
dh['temps_attente_minutes'] = dh['temps_attente'].dt.total_seconds() / 60

# Optionnel : Ne garder que les valeurs positives
dh = dh[dh['temps_attente_minutes'] >= 0].reset_index(drop=True)
```

FIGURE 3.3 – Preparation et Transformation des Donnees Temporelles pour l'Analyse des Temps d'Attente

Voici une interprétation détaillé :

— **Étape 1 : On a converti les colonnes « date_arrivee » et « date_passage » en format « date » :** Ces deux lignes convertissent les colonnes « **date_arrivee** » et « **date_passage** » en objets datetime , ce qui nous a permis d'extraire uniquement la partie date (sans l'heure).
Pourquoi faire cela ? Les données brutes peuvent être au format texte ou mal formatées. Ainsi, en les transformant en objet « **datetime** », on peut facilement effectuer des opérations temporelles (comparaisons de dates, calculs, etc.). Et encore, « **.dt.date** » garde seulement la date, sans l'heure, pour simplifier les manipulations futures si nécessaire.

— **Étape 2 : On a converti les dates en chaîne de caractéres (str) pour les concaténer :** Cela a transformé les valeurs de « **date_arrivee** » et « **date_passage** » en chaînes de caractéres (str) pour pouvoir les concaténer avec les heures (« **heure_arrivee** », « **heure_passage** ») plus tard.
Pourquoi faire cela ? Pandas ne permet pas de concaténer directement un objet « **datetime.date** » avec une heure sous forme de texte. Cela a préparé les données à créer un horodatage complet (date + heure) dans la prochaine étape.

— **Étape 3 : On a converti les heures en chaîne de caractéres aussi :** C'est pour s'assurer que les colonnes « **heure_arrivee** » et « **heure_passage** » sont bien au format texte, même si elles le sont déjà.
Pourquoi on a fait cela ? Pour éviter les erreurs lors de la concaténation avec les

dates. Par exemple, **"09 :17"** doit être une chaîne pour être combinée avec **"2025-04-08" devient "2025-04-08 09 :17"**.

— **Étape 4 : On a combiné date + heure, puis on l'a converti en « datetime » complet :** On a combiné les colonnes « **date_arrivee** » + « **heure_arrivee** » pour créer une nouvelle colonne « **datetime_arrivee** ». On a fait de même pour « **datetime_passage** ».

— **Étape 5 : On a calculé du temps d'attente (en minutes) :** Ce qui, calcule la différence entre l'heure de passage et celle d'arrivée donnant « **temps_attente** », convertit cette différence en minutes « **temps_attente_minutes** ».

— **Étape 6 : On n'a gardé que les valeurs positives (en tout cas il n'y avait que des valeurs positives) :** Cela, supprime toutes les lignes où le temps d'attente est négatif (cas impossibles : le patient arrive aprés avoir été vu), réinitialise les index aprés suppression (« **reset_index** » (drop=True)).
Pourquoi faire cela ? Il peut y avoir des erreurs dans les données (ex. mauvaise saisie de l'heure) et un temps d'attente négeatif n'a aucun sens clinique ou logique donc ces cas doivent être supprimés.

- **Normalisation de l'âge en années décimales pour une utilisation numérique dans les modéles :**
Lors de notre enquête, nous avons enregistré l'âge en années pour les adultes, les adolescents et les enfants, tandis que pour les bébés, nous l'avons exprimé en mois ou en jours. À travers notre revue de littérature, nous avons remarqué que la majorité des études utilisent l'âge en années, sauf lorsque l'étude se concentre sur les enfants en dessous d'un certain âge, où les mois ou les jours sont utilisés pour exprimer l'âge des patients. Par conséquent, nous avons choisi d'exprimer les âges en années, étant donné que les bébés ne représentent pas la majorité des patients. Ainsi, nous avons converti les valeurs de la colonne "Âge" en années. Le bloc de code ci-dessous nettoie la colonne "Âge" pour la rendre exploitable dans les analyses statistiques et les modéles prédictifs, tels que la régression, les réseaux de neurones, etc...

```
# Fonction pour convertir l'âge en années décimales
def age_to_years(age):
    if isinstance(age, str):
        if 'ans' in age:
            return float(age.split()[0])
        elif 'mois' in age:
            return float(age.split()[0]) / 12
        elif 'jour' in age or 'jrs' in age:
            return float(age.split()[0]) / 365
    return age

dh['Age'] = dh['Age'].apply(age_to_years)
```

FIGURE 3.4 – Normalisation de l'age en annees decimales pour une utilisation numerique dans les modeles

La fonction « age_to_years(age) » a pour but de convertir n'importe quelle unité d'âge en années décimales , peu importe le format d'entrée.

— Si l'âge contient "ans" résultat ; extraction du nombre et conversion en flottant.
— Si l'âge contient "mois" résultat ; convertit en années (mois / 12).
— Si l'âge contient "jour" ou "jrs" résultat ; convertit en années (jours / 365).
— Si l'âge est déjà un nombre (numérique ou vide) résultat ; il est conservé tel quel.

dh['Age'] = dh['Age'].apply(age_to_years) : Applique la fonction à chaque ligne de la colonne « Age ».

- **Mapping des couleurs de ticket en niveaux de priorité :**
  Nous avons converti les valeurs catégorielles ('RO', 'OR', 'V') en valeurs numériques pondérées (5, 3,1) exprimant les niveaux de gravité afin de les intégrer dans des modéles de Machine Learning.

```
ticket_map = {'RO': 5, 'OR': 3, 'V': 1}
dh['Gravite'] = dh['couleur_ticket'].map(ticket_map)
```

  L'attribution des valeurs numériques pondérées 5, 3, 1 aux couleurs des tickets 'RO', 'OR', 'V' est une décision basée sur une logique de priorisation et de gravité.

  — **Rouge (RO) - Valeur 5 :** La couleur rouge est souvent associée à une situation d'urgence ou de haute priorité. En attribuant la valeur la plus élevée (5), cela indique que les patients avec un ticket rouge nécessitent une attention immédiate et sont considérés comme les plus urgents.
  — **Orange (OR) - Valeur 3 :** La couleur orange est généralement utilisée pour indiquer une situation de priorité moyenne. La valeur 3, qui est intermédiaire, refléte une urgence modérée. Ces patients doivent être vus rapidement, mais leur condition n'est pas aussi critique que ceux avec un ticket rouge.
  — **Vert (V) - Valeur 1 :** La couleur verte est souvent utilisée pour indiquer une situation non urgente ou de faible priorité. La valeur la plus basse (1) signifie que ces patients peuvent attendre plus longtemps pour être vus, car leur condition est considérée comme stable ou non critique.

  Il est à noter que les valeurs 4 et 2 ne sont utilisées dans aucun des services médicaux. Elles sont difficilement attribuables à un patient.
  Cette pondération permet de quantifier le niveau de gravité ou d'urgence associé à chaque couleur de ticket d'une facc on claire ("RO" >"OR" >"V"), facilitant ainsi l'utilisation de ces données dans des modéles analytiques ou prédictifs. En convertissant les catégories en valeurs numériques, il devient plus facile d'effectuer des calculs et des analyses statistiques, comme la régression ou le clustering, qui nécessitent des entrées numériques.
  on a utilisée cette ligne de code :

- **Encodage des Variables Catégorielles via One-Hot Encoding :**
  On a encodé les variables catégorielles (pathologie, equipeS_soignants_ats) par des valeurs numériques binaires (0/1) pour qu'elles puissent être utilisées dans les modéles d'apprentissage automatique. Cette technique transforme une variable catégorielle en plusieurs colonnes binaires (une par catégorie). Chaque nouvelle colonne indique la présence ou l'absence de la valeur correspondante (ex. pathologie_grippe, pathologie_diabete, etc.). Les algorithmes de Machine Learning (comme Random Forest, KNN, réseaux de neurones) ne peuvent pas traiter directement les valeurs textuelles comme "grippe", "fracture", "équipe A", etc.

```
dh = pd.get_dummies(dh, columns=['pathologie','equipeS_soignants_ats'])
```

Ce que fait exactement pd.get_dummies() :
Prend chaque catégorie unique dans les colonnes 'pathologie' et « equipeS_soignants_ats ». Crée une nouvelle colonne binaire pour chaque catégorie. Remplit le DataFrame avec 1 si la catégorie est présente, 0 sinon.

3.1.3 Exploration des données (EDA)

Analyse des Statistiques Descriptives âge et Temps d'Attente

Nous avons montré un apercu de la distribution des âges et des temps d'attente pour mieux comprendre la structure des données et identifier les tendances ou les anomalies potentielles.

	Age	temps_attente_minutes
count	1411.000000	1411.000000
mean	33.252525	16.230333
std	25.207365	19.718446
min	0.008219	0.000000
25%	10.000000	4.000000
50%	30.000000	10.000000
75%	51.000000	20.500000
max	98.000000	245.000000

FIGURE 3.5 – Analyse des Statistiques Descriptives Age et Temps d'Attente a l'EPH de Kherrata

- **Nombre d'entrées :** L'ensemble de données comprend 1411 entrées complètes pour les deux variables analysées, l'âge et le temps d'attente.
- **Âge moyen :** L'âge moyen des patients est de 33.25 ans, ce qui suggére une population relativement jeune. Cependant, avec un écart-type de 25.21 ans, il y a une grande variabilité dans les âges, indiquant la présence de patients de tous âges, des nouveau-nés aux personnes âgées jusqu'à 98 ans.
- **Temps d'attente moyen :** Le temps d'attente moyen est de 16.23 minutes. Bien que cela puisse sembler raisonnable, l'écart-type de 19.72 minutes montre une variabilité significative, avec certains patients attendant jusqu'à 245 minutes, soit plus de 4 heures.
- **Distribution des ages :** 25% des patients ont moins de 10 ans, et 50% ont moins de 30 ans, ce qui indique une forte présence de jeunes patients. Cependant, 75% des patients ont moins de 51 ans, montrant une bonne répartition des âges.
- **Distribution des temps d'attente :** 25% des patients ont un temps d'attente inférieur à 4 minutes, et 50% attendent moins de 10 minutes, ce qui est plutot rapide. Cependant, 25% des patients attendent plus de 20.5 minutes, et certains jusqu'à 245 minutes, ce qui peut indiquer des inefficacités ou des goulots d'étranglement dans le systéme pour certains cas.

En résumé, bien que la majorité des patients aient des temps d'attente courts, il existe une minorité significative de patients confrontés à des attentes beaucoup plus longues, ce qui pourrait nécessiter une attention particuliére pour améliorer l'efficacité globale du systéme de soins.

Histogramme : "Distribution des Temps d'Attente"

Nous avons représenté le nombre de patients ayant attendu un certain nombre de minutes en fonction du temps d'attente

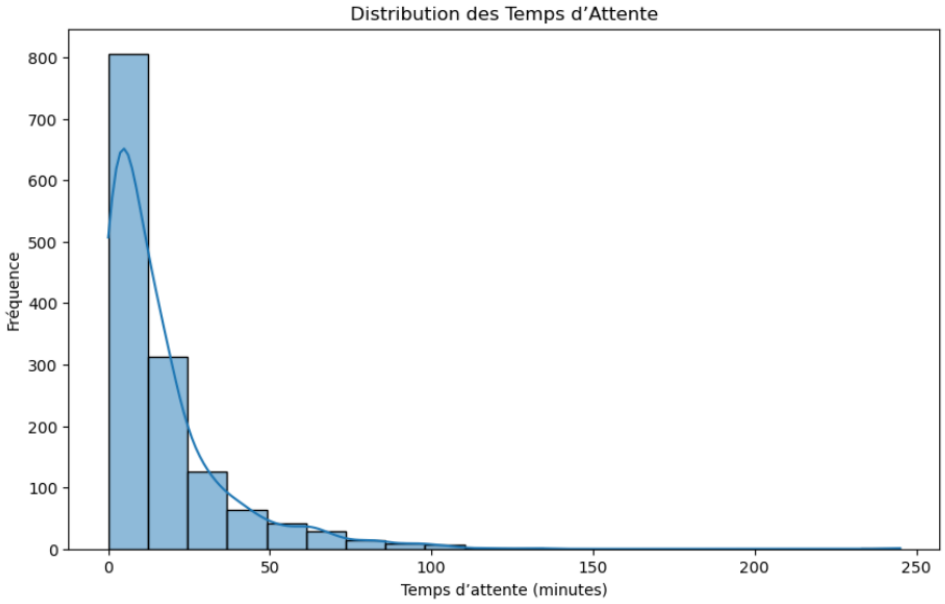


FIGURE 3.6 – Histogramme : "Distribution des Temps d'Attente" a L'EPH de Kherrata

Analyse du graphique :

- **Concentration des temps d'attente courts :** La majorité des patients ont des temps d'attente trés courts. Cela est indiqué par la barre la plus haute à gauche du graphique, qui montre que le plus grand nombre de patients a un temps d'attente proche de 0 minute.
  La fréquence diminue rapidement à mesure que le temps d'attente augmente, ce qui est visible par la décroissance rapide des barres aprés le pic initial.
- **Décroissance rapide :** Aprés le pic initial, la fréquence des temps d'attente diminue rapidement. Cela signifie que moins de patients ont des temps d'attente plus longs. La ligne de densité (courbe bleue) montre une décroissance exponentielle, ce qui est typique des distributions où la plupart des valeurs sont proches de zéro.
- **Longue traîne :** Bien que la majorité des patients aient des temps d'attente courts, il y a une "longue traîne" de patients avec des temps d'attente plus longs, allant jusqu'à 250 minutes.
  Ces temps d'attente plus longs sont moins fréquents, comme le montrent les barres plus courtes et espacées sur la droite du graphique.

interprétation :

- **Efficacité du systéme :** Le graphique suggére que le systéme est relativement efficace pour la majorité des patients, qui ont des temps d'attente courts.
- **Problémes potentiels :** Cependant, il y a un petit nombre de patients qui subissent des temps d'attente beaucoup plus longs, ce qui pourrait indiquer des inefficacités ou des goulots d'étranglement dans le systéme pour certains cas.

Ce type de visualisation est utile pour identifier les tendances générales dans les temps d'attente et pour cibler les domaines nécessitant des améliorations.

Analyse des Temps d'Attente par Niveau de Priorité avec un Boxplot

On commence par :

- **Transformation des Données :**

```
#Création d'une version lisible de la gravité
dh['GraviteLabel'] = dh[ 'couleur_ticket'].map({'RO': 'Élevée', 'OR': 'Moyenne', 'V': 'Faible'})
```

FIGURE 3.7 – Transformation des donnees

ce code transforme les valeurs de la colonne couleur_ticket en libellés de gravité plus lisibles : 'Élevée', 'Moyenne', et 'Faible'. Cela permet de catégoriser les patients en fonction de la gravité de leur condition.

- **Création de Catégories :**

```
# Optionnel : tri des catégories
dh['GraviteLabel'] = pd.Categorical(dh['GraviteLabel'], categories=['Faible', 'Moyenne', 'Élevée'], ordered=True)
```

FIGURE 3.8 – Creation de Categories

cette ligne permet de convertir la colonne GraviteLabel en un type de données catégoriel avec un ordre spécifique. Cela permet de s'assurer que les catégories sont traitées dans un ordre logique lors des analyses et des visualisations.

ensuite on continue avec :

- **La Visualisation des Données :** Ce graphique montre comment les temps d'attente varient selon le niveau de priorité des patients : faible, moyenne, et élevée.

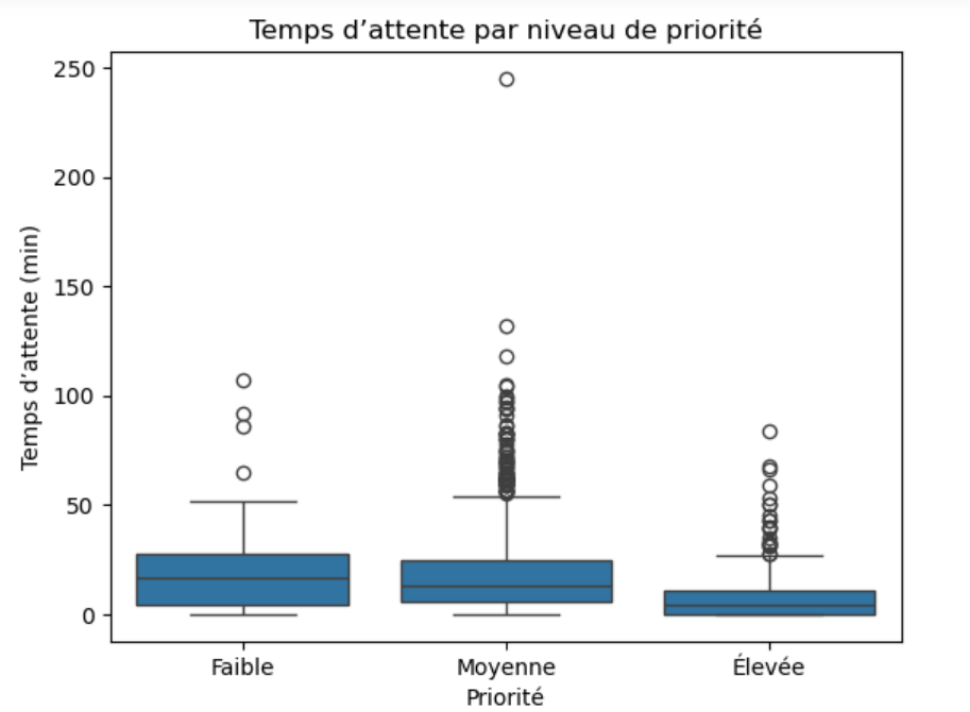


FIGURE 3.9 – Analyse des Temps d'Attente par Niveau de Priorite avec un Boxplot a L'EPH de Kherrata

interpretation :

— **Faible Priorité :** La plupart des patients avec une faible priorité attendent un temps relativement court. Cependant, il y a quelques exceptions où certains patients attendent beaucoup plus longtemps que la majorité.

— **Moyenne Priorité :** Les patients avec une priorité moyenne ont des temps d'attente plus variables. Bien que la plupart attendent un temps modéré, certains doivent attendre beaucoup plus longtemps, ce qui est visible par les points éloignés sur le graphique.

— **Élevée Priorité :** Les patients avec une priorité élevée sont généralement vus plus rapidement, comme on peut s'y attendre. La plupart d'entre eux ont des temps d'attente courts et constants, bien qu'il y ait encore quelques exceptions avec des attentes plus longues.

En conclusion, ce graphique montre que les patients avec une priorité plus élevée sont généralement vus plus rapidement, mais il y a toujours des variations et des exceptions dans chaque catégorie de priorité.

- **Test ANOVA :** Nous avons utilisé un test ANOVA (Analysis of Variance) pour comparer les temps d'attente entre trois groupes de gravité différents : Faible, Moyenne, et Élevée.

```python
# Extraction des groupes
attente_faible = dh[dh['GraviteLabel'] == 'Faible']['temps_attente_minutes']
attente_moyenne = dh[dh['GraviteLabel'] == 'Moyenne']['temps_attente_minutes']
attente_elevee = dh[dh['GraviteLabel'] == 'Élevée']['temps_attente_minutes']

# Test ANOVA à un facteur
f_stat, p_val = stats.f_oneway(attente_faible, attente_moyenne, attente_elevee)

# Résultat
print(f"Statistique F : {f_stat:.4f}")
print(f"Valeur p : {p_val:.4f}")

# Interprétation
if p_val < 0.05:
    print("⇒ Il existe une différence significative entre au moins deux groupes de gravité.")
else:
    print("⇒ Aucune différence significative détectée entre les groupes.")
```

FIGURE 3.10 – Test ANOVA

Voici une explication détaillée du code et une interprétation des résultats :

— **Extraction des groupes** : Ces lignes extraient les temps d'attente pour chaque niveau de gravité (Faible, Moyenne, Élevée) dans des séries distinctes.

— **Test ANOVA à un facteur :** Cette ligne effectue un test ANOVA à un facteur pour comparer les moyennes des temps d'attente entre les trois groupes de gravité.
stats.f_oneway est une fonction qui prend en entrée les trois groupes de données et retourne deux valeurs : la statistique **F (f_stat)** et la valeur **p (p_val)**.

— **Interprétation des résultats :**

- **Statistique F : 51.1568** :La statistique F est une mesure de la variabilité entre les moyennes des groupes par rapport à la variabilité au sein des groupes. Une valeur élevée de la statistique F indique une différence significative entre les moyennes des groupes.
- **Valeur p : 0.0000 :** La valeur p est la probabilité que les différences observées entre les groupes soient dues au hasard. Une valeur p trés faible (inférieure à 0,05) indique que les différences observées sont statistiquement significatives

— **Conclusion :** Puisque la valeur p est inférieure à 0,05, le code conclut qu'il existe une différence significative entre au moins deux des groupes de gravité en termes de temps d'attente. Cela signifie que les temps d'attente différent significativement entre les niveaux de gravité Faible, Moyenne, et Élevée.

Analyse de l'Affluence Horaire

Nous avons présenté un diagramme en barre qui montre le nombre de patients arrivant à différentes heures de la journée

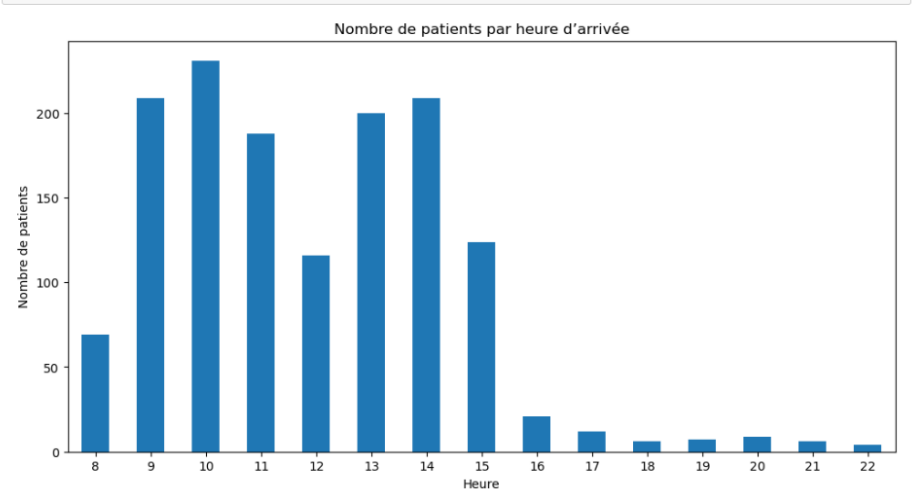


FIGURE 3.11 – Analyse de l'Affluence Horaire Distribution des Patients par Heure d'Arrivee a L'EPH de Kherrata

Analyse du Graphique :aprés analyse du diagramme ; on observe :

- **Heures de Pointe :** Les heures entre 10h et 15h semblent être les plus chargées, avec un pic notable à 10h et 14h où le nombre de patients dépasse 200. Ces pics indiquent des périodes de forte affluence, probablement dues à des facteurs tels que les horaires de travail, les pauses déjeuner, ou d'autres routines quotidiennes.
- **Heures Creuses :** Les heures en dehors de cette plage, notamment aprés 16h, montrent une baisse significative du nombre de patients, avec trés peu de patients arrivant aprés 18h. Cela pourrait être dû à la fermeture des services ou à une diminution générale des activités en fin de journée
- **Variabilité :** Il y a une variabilité notable dans le nombre de patients arrivant à différentes heures, avec des pics marqués et des creux tout aussi prononcés. Par exemple, l'arrivée des patients chute brusquement aprés 15h, ce qui pourrait indiquer un changement dans la disponibilité des services ou des habitudes des patients.

Interprétation : on peut en déduire que ;

- **Pour la Gestion des Ressources :** Les heures de pointe identifiées peuvent aider à la planification des ressources. Par exemple, s'assurer que suffisamment de personnel est disponible pendant ces périodes pour gérer l'afflux de patients.
- **Pour l'Optimisation des Services :** Comprendre les heures de forte et faible affluence peut aider à optimiser les services, par exemple en planifiant des pauses pour le personnel pendant les heures creuses ou en s'assurant que les ressources sont maximisées pendant les heures de pointe.
- **Pour l'Amélioration de l'Expérience Patient :** En anticipant les périodes de forte affluence, les établissements peuvent mettre en place des mesures pour réduire les temps d'attente et améliorer l'expérience globale des patients.

En conclusion, ce graphique est un outil précieux pour comprendre les habitudes d'arrivée des patients et peut être utilisé pour optimiser la gestion des ressources et améliorer les services dans un établissement de santé.

3.1.4 Feature Engineering (Ingénierie des Caractéristiques)

L'ingénierie des caractéristiques est une étape cruciale dans l'analyse des données. Elle consiste à créer de nouvelles variables ou à transformer les variables existantes pour améliorer les performances des modéles de machine learning. Voici ce que fait chaque partie du code :

Heure de Pointe

```
# Heure de pointe
dh['HeurePointe'] = pd.cut(dh['heure_arrivee'],
                           bins=[7, 12, 16, 20],
                           labels=['MATIN', 'APRESMIDI', 'SOIR'])
```

FIGURE 3.12 – Heure de Pointe

Cette ligne crée une nouvelle colonne appelée « HeurePointe » qui catégorise l'heure d'arrivée des patients en trois périodes de la journée : matin (7h-12h), aprés-midi (12h-16h), et soir (16h-20h). Cela permet d'analyser comment les temps d'attente varient selon le moment de la journée.

Patient Chronique

```
dh['EstChronique'] = dh['maladie_chronique'].notnull().astype(int)
```

Cette ligne crée une nouvelle colonne « EstChronique » qui indique si un patient souffre d'une maladie chronique. Si la colonne « maladie_chronique » contient une valeur (c'est-à-dire qu'elle n'est pas nulle), alors « EstChronique » est défini à 1, sinon à 0. Cela permet de facilement identifier les patients chroniques dans les analyses ultérieures.

Disponibilité du médecin

```
dh['MedecinCount']=dh.groupby('medecin'['Patients'].transform('count')
```

Cette ligne crée une nouvelle colonne « MedecinCount » qui compte le nombre de patients vus par chaque médecin. Cela peut aider à évaluer la charge de travail de chaque médecin et à identifier les mèdecins les plus sollicités.

Patients en Attente :

```
# Initialiser les colonnes
dh['patients_RO_devant'] = 0
dh['patients_OR_devant'] = 0
dh['patients_V_devant'] = 0

for i, row in dh.iterrows():
    current_time = row['datetime_arrivee']

    # Patients qui ne sont pas encore passés au moment de l'arrivée
    mask_waiting = (dh['datetime_arrivee'] < current_time) & (dh['datetime_passage'] > current_time)

    # Compter ceux avec une priorité supérieure ou égale
    dh.at[i, 'patients_RO_devant'] = ((dh['couleur_ticket'] == 'RO') & mask_waiting).sum()
    dh.at[i, 'patients_OR_devant'] = ((dh['couleur_ticket'] == 'OR') & mask_waiting).sum()
    dh.at[i, 'patients_V_devant'] = ((dh['couleur_ticket'] == 'V') & mask_waiting).sum()
```

FIGURE 3.13 – Patients en Attente

Ce bloc de code parcourt chaque ligne du DataFrame et compte le nombre de patients en attente avec des priorités différentes (RO, OR, V) au moment de l'arrivée de chaque patient. Cela permet de comprendre la charge de travail et les temps d'attente en fonction de la priorité des patients.

Récupération des Colonnes de Pathologie

```
# Récupère toutes les colonnes qui commencent par 'pathologie_'
patho_cols = [col for col in dh.columns if col.startswith('pathologie_')]

print("Colonnes récupérées :", patho_cols)
```

FIGURE 3.14 – Recuperation des Colonnes de Pathologie

Cette ligne récupère toutes les colonnes du DataFrame qui commencent par 'pathologie_'. Cela permet de facilement accéder á toutes les colonnes liées aux pathologies pour des analyses ultérieures.
En conclusion, nous avons avec ce code préparé et enrichit les données pour une analyse plus approfondie en créant de nouvelles caractéristiques et en transformant les caractéristiques existantes. Cela nous a permis d'améliorer la qualité des modéles de machine learning et de mieux comprendre les données.

3.1.5 Modélisation Prédictive

Nous avons utilisé des techniques statistiques et d'apprentissage automatique pour prévoir des résultats futurs en se basant sur nos données récoltées. Ca nous a permis d'identifier des tendances, de faire des prévisions. En principe, ca doit aider à optimiser les ressources, améliorer les temps d'attente et prioriser les soins des patients. Cette approche est essentielle pour améliorer l'efficacité et la qualité des services médicaux.

Préparation des Données

— **Variables Explicatives et Cibles :**

```
# Variables explicatives et cibles
#Prédiction
X_reg = dh[['Age','Gravite','heure_arrivee','patients_RO_devant','patients_OR_devant','patients_V_devant']]
y_reg = dh['temps_attente_minutes']
#Classification
X_clf = dh[['Age','EstChronique','MedecinCount']+ patho_cols]
y_clf = dh['couleur_ticket']
```

FIGURE 3.15 – Variables Explicatives et Cibles

- **Prédiction :** Nous sélectionnons certaines colonnes de notre jeu de données pour créer un ensemble de variables explicatives (X_reg) qui pourraient influencer le temps d'attente. La variable cible (y_reg) est le temps d'attente en minutes que nous voulons prédire.
- **Classification :** Pour la classification, nous utilisons un ensemble différent de variables explicatives (X_clf) pour prédire la catégorie de la couleur du ticket (y_clf), qui représente la priorité du patient.

— **Division des Données :**

```
# Division des données
#Prédiction
X_train_r, X_test_r, y_train_r, y_test_r = train_test_split(X_reg, y_reg, test_size=0.2, random_state=42)
#Classification
X_train_c, X_test_c, y_train_c, y_test_c = train_test_split(X_clf, y_clf, test_size=0.2, random_state=42)
```

FIGURE 3.16 – Division des Donnees

- **Prédiction :** Nous divisons les données en deux parties : un ensemble d'entraînement (80%) et un ensemble de test (20%). Cela nous permet d'entraîner notre modéle sur une partie des données et de tester sa performance sur une autre partie non vue pendant l'entraînement.
- **Classification :** De même, nous divisons les données pour la classification en ensembles d'entraînement et de test.

Modéle de Régression

— **Random Forest Regressor :**

```python
# Modèle de régression
reg_model =model = RandomForestRegressor(
    n_estimators=100,      # Nombre raisonnable d'arbres
    max_depth=8,           # Limite la complexité des arbres
    min_samples_split=50,  # Évite de trop diviser les nœuds
    min_samples_leaf=20,   # Empêche de créer des feuilles trop spécifiques
    max_features='sqrt',   # Utilise seulement √(n_features) variables à chaque split
    random_state=42,
    oob_score=True         # Active l'estimation OOB (comme validation interne)
)
reg_model.fit(X_train_r, y_train_r)
#Prédiction et Évaluation
preds_r = reg_model.predict(X_test_r)
y_pred_train = reg_model.predict(X_train_r)

rmse_train = np.sqrt(mean_squared_error(y_train_r,y_pred_train))
rmse_test = np.sqrt(mean_squared_error(y_test_r,preds_r))

r2_train = r2_score(y_train_r, y_pred_train)
r2_test = r2_score(y_test_r, preds_r)

print(f"RMSE - Train: {rmse_train:.4f}, Test: {rmse_test:.4f}")
print(f"R²   - Train: {r2_train:.4f}, Test: {r2_test:.4f}")
```

FIGURE 3.17 – Random Forest Regressor

- **Création et Entraînement du Modéle :**Nous utilisons un modéle de forêt aléatoire pour la régression, qui est un ensemble d'arbres de décision. Les paramétres définissent la complexité et la structure des arbres.
  n_estimators=100 : Nous utilisons 100 arbres dans la forêt.
  max_depth=8 : Chaque arbre peut avoir une profondeur maximale de 8 niveaux.
  min_samples_split=50 : Un noeud doit avoir au moins 50 échantillons pour être divisé.
  min_samples_leaf=20 : Une feuille doit avoir au moins 20 échantillons.
  max_features='sqrt' : Le nombre de caractéristiques à considérer pour la meilleure division est la racine carrée du nombre total de caractéristiques.
  random_state=42 : Pour la reproductibilité des résultats.
  oob_score=True : Utilise des échantillons hors du sac pour estimer l'erreur de généralisation.
- **Prédiction et Évaluation :** Nous utilisons le modéle entraîné pour faire des prédictions sur l'ensemble de test.
  Nous calculons l'erreur quadratique moyenne (RMSE) et le coefficient de détermination (R^2) pour évaluer la performance du modéle sur les ensembles d'entraînement et de test.
- **Résultats :**
  RMSE : 16.04 (Train), 16.48 (Test).
  R^2 : 0.24 (Train), 0.25 (Test).
  Ces résultats indiquent que le modéle de forêt aléatoire a une performance modérée pour prédire les temps d'attente. Le RMSE montre l'erreur moyenne des prédictions, tandis que le R^2 indique la proportion de variance expliquée par le modéle.
- **Importance des variables :**

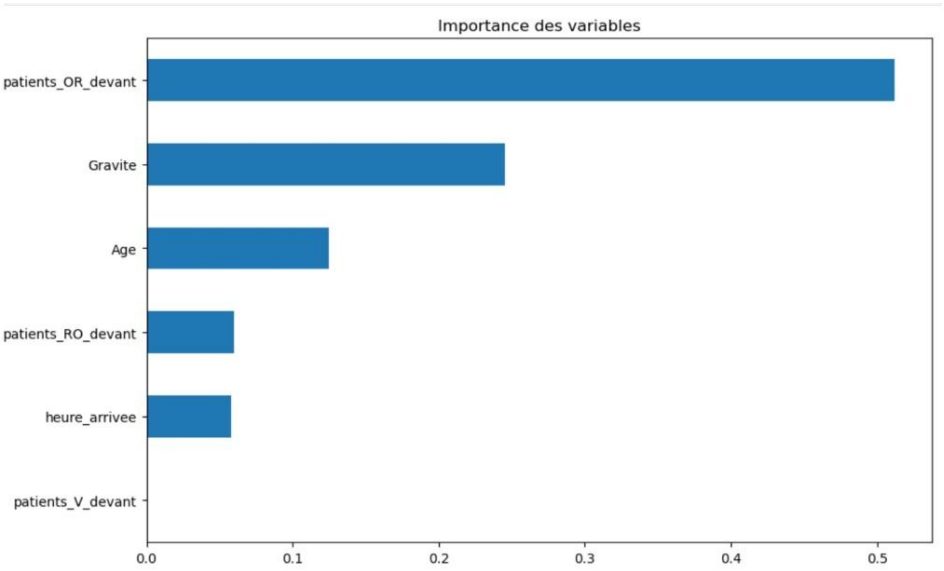


FIGURE 3.18 – Importance des variables expliquatives utilisée dans Random Forest Regressor

Interprétation :
Impact des Variables : Les variables avec une importance plus élevée ont un impact plus significatif sur les prédictions du modéle. Par exemple, le nombre de patients de priorité orange en attente est le facteur le plus influent pour prédire le temps d'attente, tandis que le nombre de patients de priorité verte en attente a le moins d'impact.
Optimisation des Ressources : Comprendre l'importance des variables peut aider à optimiser les ressources et à améliorer les processus dans un environnement de soins de santé. Par exemple, en se concentrant sur la gestion des patients de priorité orange, on pourrait potentiellement réduire les temps d'attente pour tous les patients.
Amélioration du Modéle : En identifiant les variables les plus importantes, on peut également envisager de collecter plus de données ou de créer de nouvelles caractéristiques basées sur ces variables pour améliorer la performance du modéle.
En conclusion, ce graphique montre que le nombre de patients de priorité orange en attente est le facteur le plus influent pour prédire le temps d'attente, suivi de la gravité et de l'age des patients. Les autres variables ont un impact relativement faible sur les prédictions du modéle.

— **Réseaux de neurones :**

```python
# Standardisation des données
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X_reg)

# Division des données en ensembles d'entraînement et de test
X_train_r, X_test_r, y_train_r, y_test_r = train_test_split(X_scaled, y_reg, test_size=0.2, random_state=42)

# Création d'un modèle de réseau de neurones simple
model = Sequential()
model.add(Dense(64, activation='relu', input_shape=(X_train_r.shape[1],)))
model.add(Dense(32, activation='relu'))
model.add(Dense(1))  # Couche de sortie pour la régression

# Compilation du modèle
model.compile(optimizer=Adam(learning_rate=0.001), loss='mse')

# Entraînement du modèle
model.fit(X_train_r, y_train_r, epochs=50, batch_size=16, verbose=1)

# Prédiction sur l'ensemble de test
y_pred = model.predict(X_test_r).flatten()

# Calcul des métriques de performance
rmse = mean_squared_error(y_test_r, y_pred, squared=False)
r2 = r2_score(y_test_r, y_pred)

print(f"RMSE: {rmse}")
print(f"R²: {r2}")
```

FIGURE 3.19 – Reseaux de neurones

- **Standardisation des Données :** Nous standardisons les données pour que chaque caractéristique ait une moyenne de 0 et un écart-type de 1. Cela est important pour les modéles de réseau de neurones.
- **Création et Entraînement du Modéle :** Nous créons un modéle de réseau de neurones simple avec deux couches cachées et une couche de sortie. Nous utilisons l'optimiseur Adam et la fonction de perte de l'erreur quadratique moyenne (MSE) pour entraîner le modéle.
- **Prédiction et Évaluation :** Nous utilisons le modéle entraîné pour faire des prédictions sur l'ensemble de test et calculons le RMSE et le R^2 pour évaluer la performance du modéle.
- **Résultat :**
  RMSE : 16.04.
  R^2 : 0.29.
  Le modéle de réseau de neurones a une performance similaire au modéle de forêt aléatoire, avec un RMSE de 16.04 et un R^2 de 0.29.

— **Régression lineaire :**

```python
# Création et entraînement du modèle
model = LinearRegression()
model.fit(X_train_r, y_train_r)

# Prédiction
y_pred = model.predict(X_test_r)

# Évaluation
rmse = np.sqrt(mean_squared_error(y_test_r, y_pred))
r2 = r2_score(y_test_r, y_pred)

print(f"RMSE : {rmse:.2f} min")
print(f"R²   : {r2:.4f}")
```

FIGURE 3.20 – Regression lineaire

- **Création et Entraînement du Modéle :** Nous utilisons un modéle de régression linéaire simple pour prédire les temps d'attente.
- **Prédiction et Évaluation :** Nous utilisons le modÃ¨le entraîné pour faire des prédictions sur l'ensemble de test et calculons le RMSE et le R^2 pour évaluer la performance du modéle.
- **Résultat :** RMSE : 16.70
  R^2 : 0.23
  Le modéle de régression linéaire a une performance légérement inférieure aux autres modéles, avec un RMSE de 16.70 et un R^2 de 0.23.

Modéle de Classification

— **Random Forest Classifier :**

```
# Modèle de classification
clf_model = RandomForestClassifier(random_state=42)
clf_model.fit(X_train_c, y_train_c)
preds_c = clf_model.predict(X_test_c)

print("\nModèle Classification - Performance")
print(classification_report(y_test_c, preds_c))
```

FIGURE 3.21 – Random Forest Classifier

- **Création et Entraînement du Modéle :** Nous utilisons un modéle de forêt aléatoire pour la classification, qui est similaire au modéle de régression mais utilisé pour prédire des catégories.
- **Prédiction et Évaluation :** Nous utilisons le modéle entraîné pour faire des prédictions sur l'ensemble de test et imprimons un rapport de classification pour évaluer la performance du modéle.
- **résultat et interpretation : Rapport de Classification :**
  OR : Précision = 0.81, recall = 0.88, F1-score = 0.84, Support = 194
  RO : Précision = 0.70, recall = 0.59, F1-score = 0.64, Support = 83
  V : Précision = 0.00, recall = 0.00, F1-score = 0.00, Support = 6
  Accuracy : 0.77
  Macro Avg : Précision = 0.50, recall = 0.49, F1-score = 0.49, Support = 283
  Weighted Avg : Précision = 0.76, recall = 0.77, F1-score = 0.77, Support = 283
  Le modéle de classification a une bonne performance pour les catégories OR et RO, mais une performance nulle pour la catégorie V. Cela pourrait être dû à un désèquilibre dans les données ou à une difficulté à classer correctement cette catégorie.
- **Visualisation de la matrice de confusion :** Nous tracons une matrice de confusion pour visualiser la performance du modéle de classification. Cela montre combien de prédictions sont correctes et où le modéle fait des erreurs.

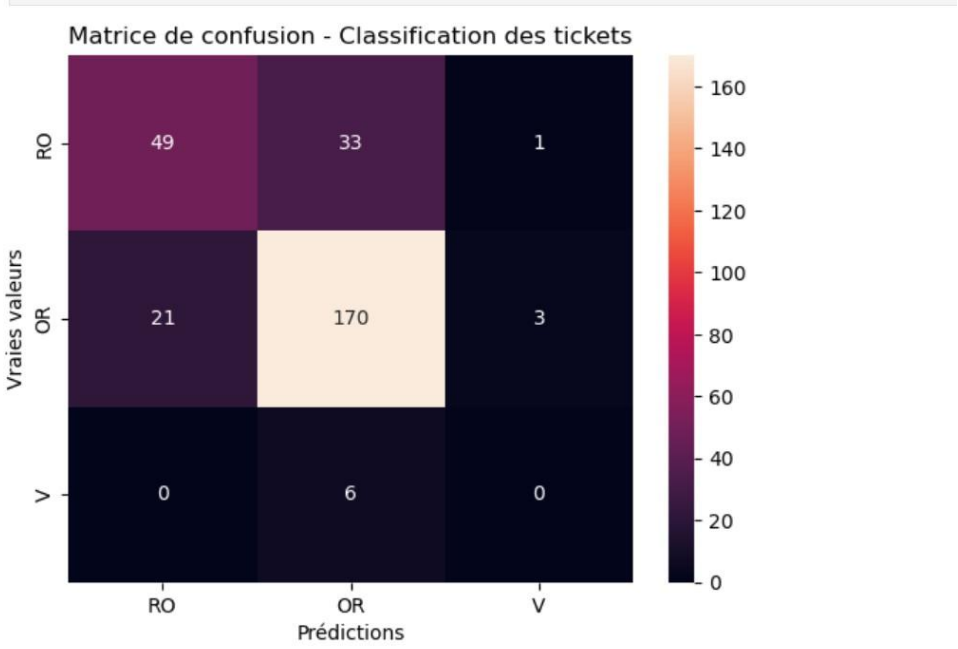


FIGURE 3.22 – matrice de confusion de l'EPH Kherrata

En résumé, ce code prépare et utilise différents modéles de machine learning pour prédire les temps d'attente et classer les patients en fonction de leur priorité. Il évalue également la performance de ces modéles et visualise les résultats pour une meilleure compréhension.

Clustering des patients

Pour mieux comprendre les besoins et les comportements des patients, nous avons fait appel au clustering qui est une technique d'analyse de données consistant à regrouper des objets similaires en ensembles appelés clusters. C'est d'ailleurs le fait qu'il est utilisé dans des études similaires qui nous a conduit à l'utiliser afin d'identifier des groupes de patients qui ont des caractéristiques similaires

```
# Sélection des caractéristiques pour le clustering
X_clust = dh[['Age', 'temps_attente_minutes', 'Gravite']].dropna()
#Standardisation des Données :
X_scaled = StandardScaler().fit_transform(X_clust)
#Application de l'Algorithme K-Means :
kmeans = KMeans(n_clusters=3, random_state=42)
kmeans.fit(X_scaled)
dh['Cluster'] = kmeans.labels_
```

FIGURE 3.23 – Clustering(KMeans)

— **Sélection des Caractéristiques pour le Clustering :** Cette ligne sélectionne les colonnes 'Age', 'temps_attente_minutes', et 'Gravite' du DataFrame dh et supprime les lignes avec des valeurs manquantes. Ces caractéristiques sont utilisées pour le clustering.

— **Standardisation des Données :** La standardisation des données est une étape importante pour s'assurer que toutes les caractéristiques ont la même échelle. Cela permet

d'éviter que certaines caractéristiques dominent le processus de clustering simplement parce qu'elles ont une plus grande échelle.

— **Application de l'Algorithme K-Means :** L'algorithme K-Means est utilisé pour regrouper les données en 3 clusters. n_clusters=3 spécifie le nombre de clusters à former. random_state=42 assure la reproductibilité des résultats.
kmeans.fit(X_scaled) : Cette ligne applique l'algorithme K-Means aux donnés standardisées.
dh['Cluster'] = kmeans.labels_ : Cette ligne ajoute une nouvelle colonne 'Cluster' au DataFrame dh, qui contient les labels des clusters assignés à chaque patient.

— **Visualisation des clusters :**

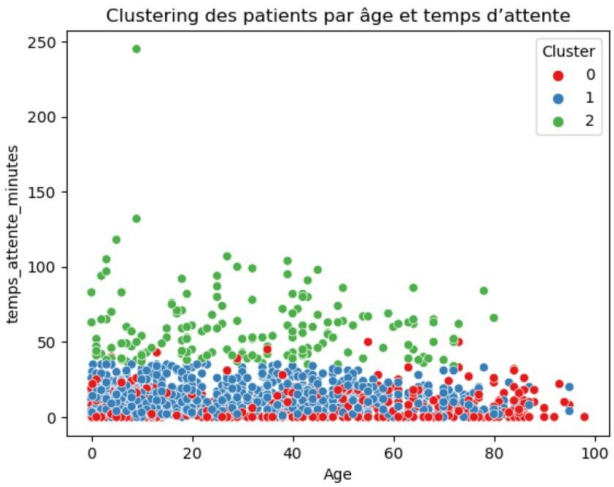


FIGURE 3.24 – Clustering des patients par age et temps d'attente a l'EPH de Kherrata

- **Affichage des Moyennes par Cluster :** pour comprendre les caractéristiques typiques de chaque groupe de patients.
  Les résultats du clustering montrent les moyennes des caractéristiques pour chaque cluster :
  Cluster 0 :
  Age moyen : 33.89 ans
  Temps d'attente moyen : 57.23 minutes
  Gravité moyenne : 3.02
  Cluster 1 :
  Age moyen : 20.50 ans
  Temps d'attente moyen : 11.94 minutes
  Gravité moyenne : 3.09
  Cluster 2 :
  Age moyen : 59.02 ans
  Temps d'attente moyen : 6.87 minutes
  Gravité] moyenne : 4.54

- **Évaluation du Modéle :**

— **Inertie du Modéle :** L'inertie est une mesure de la qualité du clustering. Elle représente la somme des distances au carré de chaque point à son centro ide de cluster. Une inertie plus basse indique un meilleur clustering. Dans ce cas, l'inertie est de 2058.62, ce qui signifie que les points sont relativement bien regroupés autour de leurs centro ides.

— **Silhouette Score :** Le Silhouette Score est une autre mesure de la qualité du clustering. Il varie entre -1 et 1, où une valeur plus élevée indique que les clusters sont bien sṕarś. Un score de 0.4024 indique une sṕaration modéré des clusters, suggérant que les clusters sont raisonnablement bien dfinis mais qu'il pourrait y avoir des chevauchements.

Le clustering des patients en fonction de leur age, de leur temps d'attente et de leur gravité permet d'identifier des groupes distincts de patients. L'inertie de 2058.62 et le Silhouette Score de 0.4024 indiquent que les clusters sont raisonnablement bien formés, bien qu'il y ait des possibilités d'amélioration. Ces mesures, combinées à la visualisation et à la répartition des niveaux de gravité, offrent des insights supplémentaires sur les caractéristiques de chaque cluster, aidant ainsi à mieux comprendre les besoins spécifiques de chaque groupe et à adapter les ressources et les soins en conséquence.

3.1.6 Tableau de bord interactif

Le tableau de bord interactif que nous avons développé avec Dash et Plotly offre une vue en temps réel des temps d'attente des patients aux urgences de l'EPH Kherrata. Ce tableau de bord est concu pour aider le personnel médical et administratif à visualiser et analyser les données de manière efficace, facilitant ainsi la prise de dècision rapide et éclairée. En affichant des graphiques tels que l'histogramme des temps d'attente par niveau de gravité et le boxplot des temps d'attente par priorité, nous pouvons identifier les tendances et les domaines nécessitant une attention particuliére pour améliorer l'efficacité et la qualité des soins.

Histogramme des Temps d'Attente par Niveau de Gravité

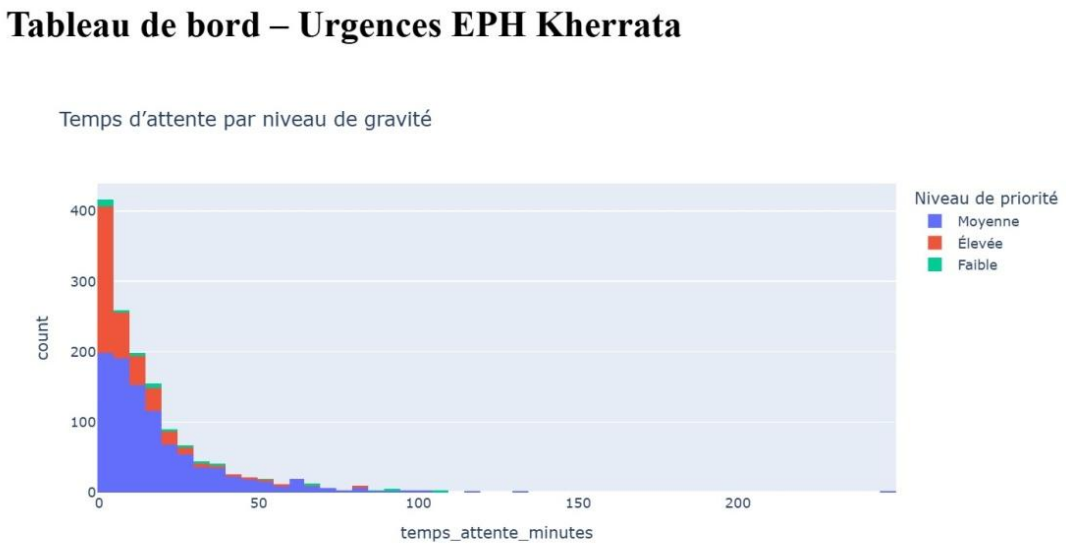


FIGURE 3.25 – temps d'attente par niveau de gravite a L'EPH de Kherrata

- **Description :** Ce graphique montre la distribution des temps d'attente pour les patients, divisée par niveau de gravité (priorité).
- **Interprétation :** Vous pouvez observer comment les temps d'attente varient en fonction de la gravité. Par exemple, les patients avec une priorité élevée peuvent avoir des temps d'attente plus courts, tandis que ceux avec une priorité faible peuvent attendre plus long-temps.
  Les couleurs différentes aident à distinguer visuellement les niveaux de priorité, facilitant l'identification des tendances et des anomalies.

Boxplot des Temps d'Attente par Priorité

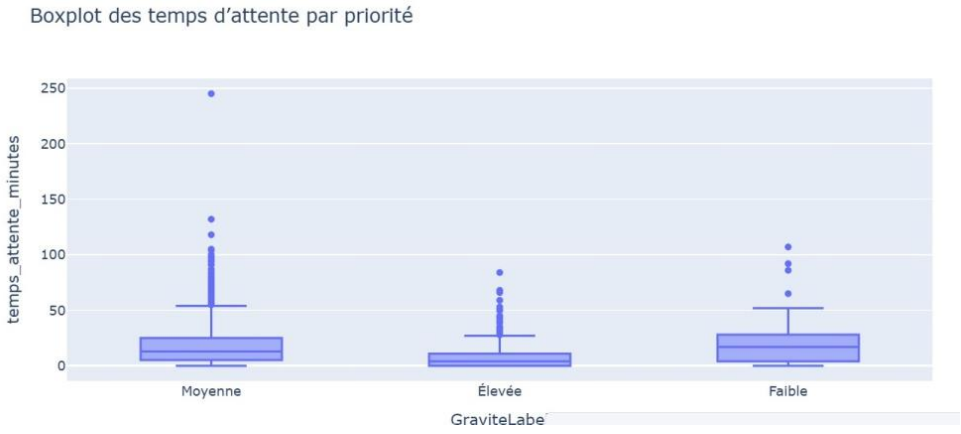


FIGURE 3.26 – Boxplot des temps d'attente par priorite a L'EPH de Kherrata

- **Description :** Ce graphique montre la distribution des temps d'attente á travers un box-plot, qui est un moyen standard de visualiser la distribution d'un ensemble de données.
- **Interprétation :** Chaque boite représente l'intervalle interquartile (IQR), qui contient les 50% centraux des données. La ligne à l'intérieur de chaque boite montre la médiane.
  Les "moustaches" s'étendent aux valeurs minimales et maximales à l'intérieur de 1,5 fois l'IQR. Les points en dehors de cette plage sont considérés comme des valeurs aberrantes et sont affichés individuellement.
  Ce graphique permet de comparer facilement la distribution des temps d'attente entre les différents niveaux de priorité. Par exemple, vous pouvez voir si les patients de haute priorité ont généralement des temps d'attente plus courts et moins variables que ceux de faible priorité.

Ce tableau de bord interactif permet aux utilisateurs de visualiser et d'analyser les temps d'attente des patients en fonction de leur niveau de priorité. Les graphiques aident à identifier les tendances, les anomalies et les domaines nécessitant une attention particuliére pour améliorer l'efficacité et la qualité des soins dans un service d'urgence

3.2 CHU Khelil Amrane de Béjaia

Nous avons détaillé les étapes pour l'EPH de Kherrata, le protocole de collecte de données étant quasiment le même. Nous allons nous limiter à interpréter les résultats.
Nous rappelons que le fichier dans lequel nos données sont inscrites porte le nom de "CHU BEJAIA FINAL.xlsx" et la période de collecte s'était étalée du 2 mars au 12 mars 2025.

Out[7]:

	Patients	Age	Sexe	date_arrivee	heure_arrivee	date_passage	heure_passage	pathologie	maladie_chronique	couleur_ticket	salles_examens	medecin
0	A6210425	48	M	2025-03-02	08:19:00	2025-03-02	09:00:00	traumatisme du cou	NaN	OR	urgence de neuro chirurgie	équipe médecin de neuro chirurgie
1	A63020325	63	M	2025-03-02	08:34:00	2025-03-02	08:40:00	cancer de l'estomac	NaN	RO	centre de trie médicale	équipe médecin trie
2	A8210425	47	M	2025-03-02	08:48:00	2025-03-02	12:00:00	douleur abdominal	NaN	V	centre de trie médicale	équipe médecin trie
3	A9210425	47	M	2025-03-02	08:54:00	2025-03-02	12:05:00	angine	NaN	V	centre de trie médicale	équipe médecin trie
4	A10210425	79	M	2025-03-02	09:03:00	2025-03-02	09:03:00	déshydratation	NaN	RO	centre de trie médicale	équipe médecin trie

FIGURE 3.27 – Tableau CHU Khellil Amrane

3.2.1 Exploration des données (EDA)

Statistiques Descriptives du CHU

Les statistiques descriptives nous donnent un apercu des données concernant l'age et le temps d'attente des patients au CHU Khelil Amrane. Voici une interprétation accessible de ces statistiques :

	Age	temps_attente_minutes
count	502.000000	502.000000
mean	47.784861	50.525896
std	20.135408	47.346878
min	16.000000	0.000000
25%	32.250000	8.000000
50%	45.000000	39.000000
75%	63.000000	70.000000
max	101.000000	248.000000

FIGURE 3.28 – Statistiques Descriptives du CHU

- **Age des Patients :**
  - **Nombre de patients :** Les données couvrent 502 patients.
  - **Age moyen :** En moyenne, les patients ont environ 48 ans. Cela signifie que la plupart des patients sont d'age moyen.
  - **Variabilité de l'age :** Il y a une grande variété d'ages parmi les patients, allant de 16 ans à 101 ans.Cela montre que le CHU Khelil Amrane traite des patients de tous ages.
  - **Répartition des ages :**
    - 25% des patients ont moins de 32 ans, ce qui signifie qu'un quart des patients sont relativement jeunes.
    - La moitié des patients ont moins de 45 ans.
    - 75% des patients ont moins de 63 ans, indiquant que la majorité des patients sont d'age moyen ou plus agés.
- **Temps d'Attente :**
  - **Nombre de patients :** Les données sur le temps d'attente couvrent également ces 502 patients.
  - **Temps d'attente moyen :** En moyenne, les patients attendent environ 50 minutes avant d'etre vus. Cela donne une idée générale du temps que les patients passent en attente.
  - **Variabilité du temps d'attente :** Les temps d'attente varient beaucoup, allant de 0 minute à 248 minutes (soit plus de 4 heures). Certains patients sont vu immédiatement, tandis que d'autres attendent trés longtemps.
  - **Répartition des temps d'attente :** 25% des patients attendent moins de 8 minutes, ce qui montre que certains patients sont vus assez rapidement. La moitié des patients attendent moins de 39 minutes. 75% des patients attendent moins de 70 minutes, ce qui signifie que la majorité des patients attendent un peu plus d'une heure
- **Interprétation Globale :**
  - **Variabilité :** Il y a une grande diversité tant dans les ages que dans les temps d'attente des patients.Cela signifie que le CHU Khelil Amrane doit gérer une population trés variée en termes d'age et de besoins de soins.
  - **Temps d'attente :** Bien que le temps d'attente moyen soit d'environ 50 minutes, la médiane est de 39 minutes. Cela suggére que, bien que la plupart des patients attendent moins d'une heure, quelquespatients ont des temps d'attente trés longs qui augmentent la moyenne.

— **Gestion des ressources :** La grande variabilité dans les temps d'attente pourrait indiquer des inefficacités ou des périodes de forte affluence. Identifier les causes de ces longs temps d'attente pourrait aider à améliorer l'efficacité et la satisfaction des patients.

En conclusion, ces statistiques nous donnent une bonne idée de la diversité des patients et de leurs expériences en termes de temps d'attente au CHU Khelil Amrane. Cela peut aider à mieux comprendre comment améliorer les services et la gestion des ressources pour répondre aux besoins de tous les patients.

L'Histogramme des temps d'attente

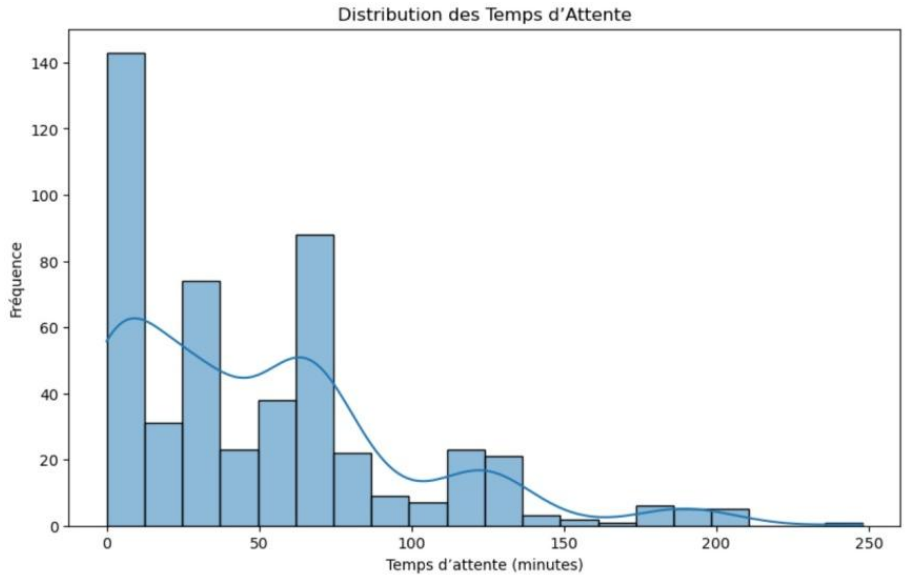


FIGURE 3.29 – L'Histogramme des temps d'attente au CHU

- **Interprétation de l'Histogramme :**
  — **Temps d'Attente Courts :** La majorité des patients ont des temps d'attente relativement courts. La barre la plus haute, située entre 0 et 20 minutes, montre que c'est l'intervalle de temps d'attente le plus fréquent. Ce qui signifie que beaucoup de patients sont vus assez rapidement aprés leur arrivée.
  — **Variabilité des Temps d'Attente :** Il y a une grande variété dans les temps d'attente. Certaines barres montrent des temps d'attente plus longs, allant jusqu'à 250 minutes, bien que ces cas soient moins fréquents. La courbe bleue, qui semble etre une courbe de densité, montre une tendance générale où la majorité des patients ont des temps d'attente courts, mais il y a une traine de patients qui attendent plus longtemps.
  — **Distribution des Temps d'Attente :** La distribution des temps d'attente semble etre asymétrique, avec une majorité de patients ayant des temps d'attente courts et un nombre plus faible de patients ayant des temps d'attente plus longs. Les pics dans l'histogramme montrent des intervalles de temps d'attente plus communs, tandis que les barres plus basses montrent des intervalles moins fréquents.

En conclusion, cet histogramme nous montre que, bien que la plupart des patients au CHU Khelil Amrane soient vus assez rapidement, il existe une variabilité significative dans les temps d'attente.Certains patients attendent beaucoup plus longtemps, ce qui pourrait indiquer des périodes de forte affluence ou des inefficacités dans la gestion des ressources. Comprendre cette distribution peut aider à identifier les moments où les temps d'attente sont les plus longs et à mettre en place des stratégies pour les réduire, améliorant ainsi l'expérience des patients et l'efficacité des services de santé.

Boxplot par niveau de priorité

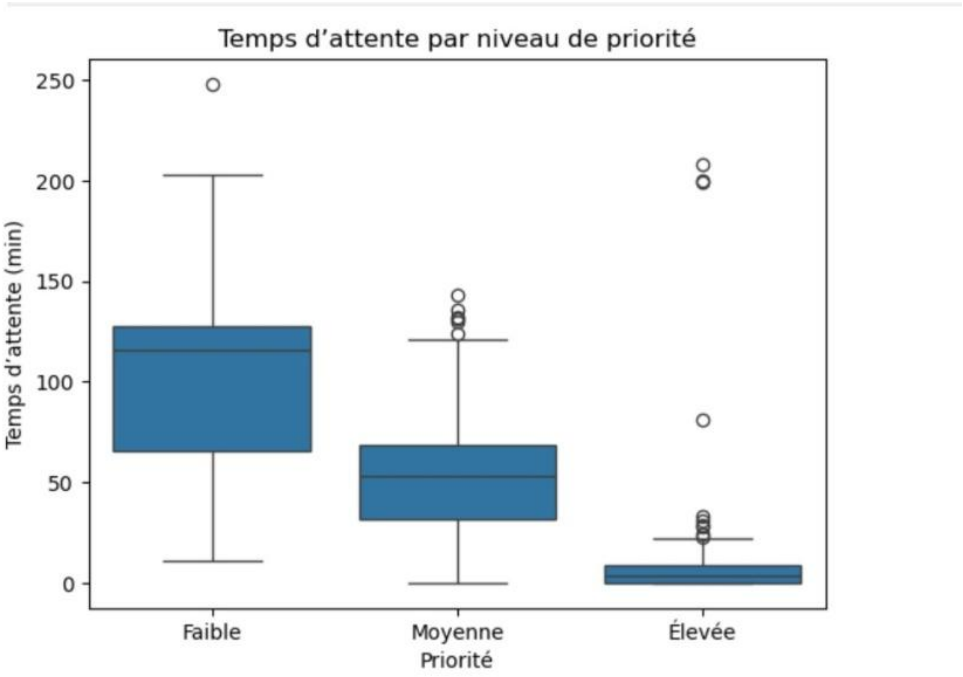


FIGURE 3.30 – Boxplot par niveau de priorite au CHU

- **Priorité Faible :**
  - **Médiane :** La ligne centrale de la boite montre que la médiane des temps d'attente pour les patients de priorité faible est relativement élevée, autour de 100 minutes.
  - **Interquartile Range (IQR) :** La boite elle-même montre l'intervalle interquartile, qui est la plage dans laquelle se situent les 50% centraux des données. Pour la priorité faible, cet intervalle est large,indiquant une grande variabilité des temps d'attente.
  - **Valeurs Aberrantes :**Les points au-dessus de la boite représentent des valeurs aberrantes, c'est-à dire des temps d'attente exceptionnellement longs par rapport à la majorité des données. Cela signifie que certains patients de priorité faible attendent beaucoup plus longtemps que la plupart.
- **Priorité Moyenne :**
  - **Médiane :** La médiane des temps d'attente pour les patients de priorité moyenne est plus basse que celle de la priorité faible, autour de 50 minutes.

— **IQR :** L'intervalle interquartile est plus étroit que pour la priorité faible, indiquant une variabilité modérée des temps d'attente.

— **Valeurs Aberrantes :** Il y a plusieurs valeurs aberrantes, indiquant que certains patients de priorité moyenne ont des temps d'attent significativement plus longs.

- **Priorité Élevée :**
  — **Médiane :** La médiane des temps d'attente pour les patients de priorité élevée est la plus basse parmiles trois catégories, autour de 10 minutes. Cela est logique car ces patients devraient etre vus plus rapidement.
  — **IQR :** L'intervalle interquartile est étroit, similaire à celui de la priorité moyenne, indiquant une faible variabilité.
  — **Valeurs Aberrantes :** Comme pour les autres catégories, il y a des valeurs aberrantes, mais elles sont moins nombreuses et moins extremes.

  En conclusion, ce boxplot montre que les patients avec une priorité plus élevée sont généralementvus plus rapidement, ce qui est conforme à l'attente. Cependant, il y a toujours des variations et desexceptions dans chaque catégorie de priorité. Les valeurs aberrantes indiquent qu'il y a des cas où les patients attendent beaucoup plus longtemps que la majorité, ce qui pourrait nécessiter une attention particuliére pour améliorer l'efficacité globale du systéme de soins. En comprenant ces variations, le CHU Khelil Amrane peut mieux gérer les ressources et améliorer les temps d'attente pour tous les patients.

- **Test Anova :**
  — Statistique F : 304.3927.
  — Valeur p : 0.0000
  — Alors, il existe une différence significative entre au moins deux groupes de gravité.La statistique F est un rapport de variances qui compare la variabilité entre les moyennes des groupes à la variabilité au sein de chaque groupe. Une valeur élevée de la statistique F indique que la variabilité entre les groupes est plus grande que la variabilité au sein des groupes.Dans ce cas, une statistique F de 304.3927 est trés élevée, ce qui suggére qu'il existe une différence significative entre les moyennes des groupes de gravité.

  — La valeur p est une mesure de la probabilité que les différences observées entre les groupes soient dues au hasard. Une valeur p faible indique que cette probabilité est faible, ce qui signifie que les différences observées sont probablement réelles et non dues à la chance. Une valeur p de 0.0000 est extremement faible, bien en dessous du seuil de signification commun de 0.05. Cela indique qu'il existe une différence statistiquement significative entre au moins deux des groupes de gravité comparés.

  En conclusion, les résultats du test ANOVA indiquent qu'il existe une différence significative entre au moins deux des groupes de gravité. Cela signifie que les temps d'attente, ou une autre variable d'intéret, différent significativement entre ces groupes.

Graphique temporel affluence horaire

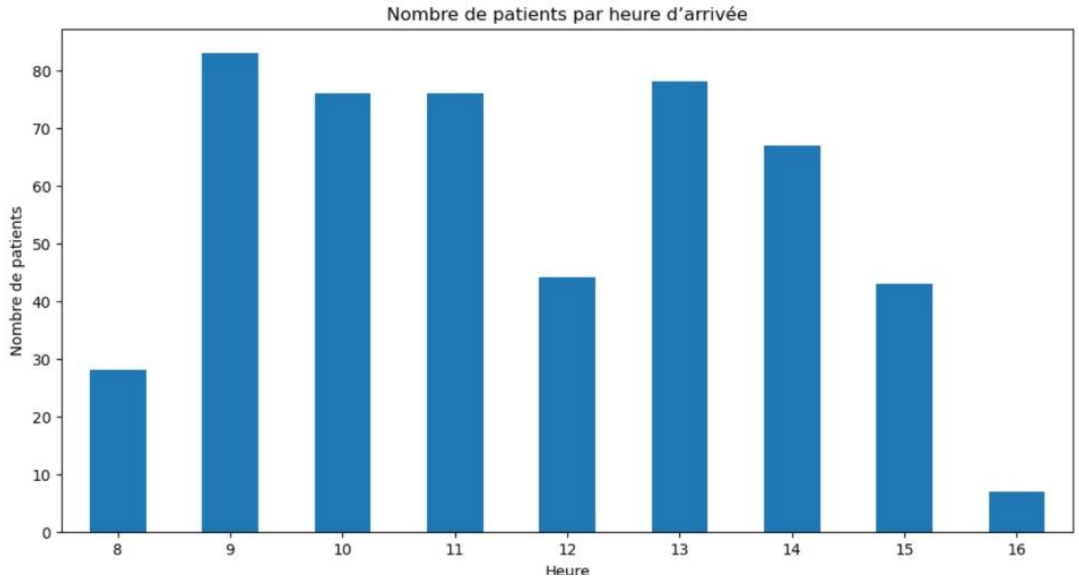


FIGURE 3.31 – Graphique temporel affluence horaire au CHU

- **Interprétation du Graphique :**
  - **Heures de Pointe :** Les heures entre 9h et 13h semblent etre les plus chargées, avec un pic notable à 9h et 13h où le nombre de patients dépasse 70. Cela signifie que ces heures sont les plus fréquentées par les patients.
    Ces pics indiquent des périodes de forte affluence, probablement dues à des facteurs tels que les horaires de travail, les pauses déjeuner, ou d'autres routines quotidiennes.
  - **Heures Creuses :** Les heures en dehors de cette plage, notamment 8h et aprés 15h, montrent une baisse significative du nombre de patients, avec trés peu de patients arrivant aprés 15h.
    Cela pourrait etre dù à la fermeture des services ou Ã une diminution générale des activités en fin de journée.
  - **Variabilité :**Il y a une variabilité notable dans le nombre de patients arrivant à différentes heures, avec des pics marqués et des creux tout aussi prononcés.
    Par exemple, l'arrivée des patients chute brusquement aprés 15h, ce qui pourrait indiquer un changement dans la disponibilité des services ou des habitudes des patients.

  En conclusion, Ce graphique est un outil précieux pour comprendre les habitudes d'arrivée des patients au CHU Khelil Amrane. En identifiant les heures de pointe et les heures creuses, le personnel médical et administratif peut mieux planifier les ressources et les services. Par exemple, s'assurer que suffisamment de personnel est disponible pendant les heures de pointe pour gérer l'afflux de patients, ou planifier des pauses pour le personnel pendant les heures creuses. Cela peut aider à optimiser les services et à améliorer l'expérience des patients.

3.2.2 Modélisation prédictive

Modéle de régression

- **Random Forest Regressor :**
  - **Résultat :**
    - RMSE - Train : 32.8008, Test : 30.5090
    - R^2 - Train : 0.5276, Test : 0.5508
  - **Interprétation :**
    - **RMSE - Train : 32.8008 :** Le RMSE sur l'ensemble d'entrainement est de 32,80. Cela signifie que, en moyenne, les prédictions du modéle sur les données d'entrainement s'écartent de la valeur réelle d'environ 32,80 unités. C'est une mesure de l'erreur moyenne que fait le modéle sur les données qu'il a vues pendant l'entrainement.
    - **RMSE - Test : 30.5090 :** Le RMSE sur l'ensemble de test est de 30,51. Cela signifie que, en moyenne,les prédictions du modéle sur les nouvelles données (non vues pendant l'entrainement) s'écartent de la valeur réelle d'environ 30,51 unités.
    - **Interprétation du RMSE :** Dans ce cas, le RMSE sur l'ensemble de test est légŕement inférieur á celui de l'ensemble d'entrainement, ce qui est une bonne indication que le modéle généralise bien aux nouvelles données et n'est pas surajusté.
    - R^2 **- Train : 0.5276 :** Le R^2 sur l'ensemble d'entrainement est de 0,5276. Cela signifie que le modéle explique environ 52,76% de la variance des données d'entrainement. En d'autres termes, un peu plus de la moitié de la variabilité des données est capturée par le modéle.
    - R^2 **- Test : 0.5508 :** Le R^2 sur l'ensemble de test est de 0,5508. Cela signifie que le modéle explique environ 55,08% de la variance des données de test.
    - **Interprétation du** R^2 **:** Dans ce cas, le R^2 sur l'ensemble de test est légérement supérieur à celui de l'ensemble d'entrainement, ce qui est une bonne indication que le modéle généralise bien aux nouvelles données.

    En conclusion, les valeurs de RMSE et de R^2 indiquent que le modéle a une performance modérée. Il explique un peu plus de la moitié de la variance des données, ce qui est acceptable mais laisse de la place pour des améliorations. Le fait que les valeurs de RMSE et de R^2 soient légérement meilleures sur l'ensemble de test que sur l'ensemble d'entrainement est une bonne indication que le modéle généralise bien aux nouvelles données et n'est pas surajusté. Ainsi, le modéle semble avoir une performance raisonnable et généralise bien aux nouvelles données,mais il pourrait etre amélioré pour expliquer une plus grande partie de la variance des données.
  - **Importance des variables explicatives :**

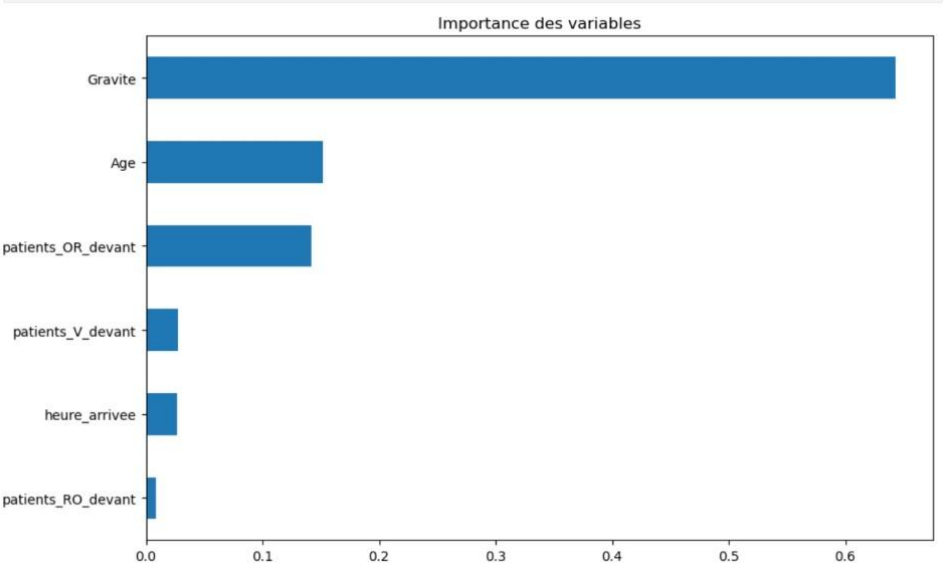


FIGURE 3.32 – Importance des variables explicatives utilise dans Random Forest Regressor

ce graphique montre que la gravité de la condition du patient est de loin le facteur le plus important pour les prédictions du modéle au CHU Khelil Amrane. L'âge et le nombre de patients de priorité orange en attente sont également des facteurs influents, mais dans une moindre mesure.Les autres variables, comme l'heure d'arrivée et le nombre de patients de priorité rouge ou verte en attente, ont un impact relativement faible sur les prédictions. Comprendre l'importance de ces variables peut aider à mieux cibler les ressources et à amÃ©liorer les soins et la gestion des patients.

- **Réseaux de neurones :**
  - **Résultat :**
    - RMSE : 23.178755429765634
    - R^2 : 0.7407301768981003
  - **interpretation :**
    - **RMSE : 23.18 :** Cela signifie qu'en moyenne, les prédictions du modÃ¨le s'écartent de la valeur réelle d'environ 23,18 unités. Un RMSE de 23,18 est relativement faible, ce qui suggére que le modéle fait des prédictions assez précises.
    - **R^2 : 0.7407 :** Un R^2 de 0,7407 signifie que le modéle explique environ 74,07% de la variance des données. Ca signifie également que le modéle capture une grande partie de la variabilité des données et est donc assez performant.

    En conclusion, les valeurs de RMSE et de R^2 indiquent que le modéle de réseau de neurones a une bonne performance. Il explique une grande partie de la variance des données et fait des prédictions assez précises.
    Avec un R^2 de 0,7407, le modéle est capable de capturer une grande partie des variations dans les données, ce qui le rend utile pour faire des prédictions précises. Cela peut etre particuliérement utile dans un contexte hospitalier pour prévoir des

résultats tels que les temps d'attente, les besoins en ressources, ou d'autres indicateurs clés de performance.

En résumé, le modéle de réseau de neurones semble bien performant et capable de faire des prédictions précises, ce qui peut aider à améliorer la gestion et les soins au CHU Khelil Amrane.

Modéle de classification

Ce tableau présente les performances d'un modéle de classification qu'on a utilisé au CHU Khelil Amrane

Modèle Classification - Performance

	precision	recall	f1-score	support
OR	0.87	0.74	0.80	53
RO	0.59	0.82	0.69	28
V	0.71	0.60	0.65	20
accuracy			0.73	101
macro avg	0.72	0.72	0.71	101
weighted avg	0.76	0.73	0.74	101

FIGURE 3.33 – Performance du modele classification applique aux donnee CHU

- **Interprétation et explication :**
  - **Précision (Precision) :**
    - **OR (Orange) : 0.87 :** Cela signifie que lorsque le modéle prédit qu'un patient est de priorité Orange, il a raison 87% du temps.
    - **RO (Rouge) : 0.59 :** Cela signifie que lorsque le modéle prédit qu'un patient est de priorité Rouge, il a raison 59% du temps.
    - **V (Vert) : 0.71 :** Cela signifie que lorsque le modéle prédit qu'un patient est de priorité Vert, il a raison 71% du temps.
  - **Rappel (Recall) :**
    - **OR : 0.74 :** Cela signifie que le modéle identifie correctement 74% des patients qui sont réellement de priorité Orange.
    - **RO : 0.82 :** Cela signifie que le modéle identifie correctement 82% des patients qui sont réellement de priorité Rouge.
    - **V : 0.60 :** Cela signifie que le modéle identifie correctement 60% des patients qui sont réellement de priorité Vert.
  - **F1-Score :**
    - **OR : 0.80 :** Le F1-score est une moyenne harmonique de la précision et du rappel. Un score de 0.80 indique une bonne performance globale pour la classification des patients de priorité Orange.
    - **RO : 0.69 :** Un score de 0.69 indique une performance modérée pour la classification des patients de priorité Rouge.

- **V : 0.65 :** Un score de 0.65 indique une performance modérée pour la classification des patients de priorité Vert.

— **Support :**

- **OR : 53 :** Il y a 53 patients de priorité Orange dans l'ensemble de données.
- **RO : 28 :** Il y a 28 patients de priorité Rouge dans l'ensemble de données.
- **V : 20 :** Il y a 20 patients de priorité Vert dans l'ensemble de données.

— **Mesures Globales :**

- **Accuracy (Exactitude) : 0.73 :** Cela signifie que le modéle a correctement classé 73% de tous les patients, indépendamment de leur priorité.
- **Macro Avg (Moyenne Macro) :** Les moyennes macro pour la précision, le rappel et le F1-score sont toutes autour de 0.72. Cela signifie que le modele a une performance moyenne de 72% pour chaque classe, sans tenir compte de la taille de chaque classe.
- **Weighted Avg (Moyenne Pondérée) :** Les moyennes pondérées pour la précision, le rappel et le F1-score sont autour de 0.74-0.76. Cela signifie que le modele a une performance moyenne de 74-76%en tenant compte de la taille de chaque classe.

En conclusion, le modele de classification a une performance globale assez bonne, avec une exactitude de 73%. Il est particuliérement performant pour la classification des patients de priorité Orange, avec une précision et un rappel élevés. Cependant, il y a des marges d'amélioration pour les classifications des priorités Rouge et Vert, où les performances sont modérées. Comprendre ces résultats peut aider à mieux cibler les améliorations du modéle pour une gestion plus efficace des priorités des patients au CHU Khelil Amrane.

— **Visualisation de la matrice de confusion :**

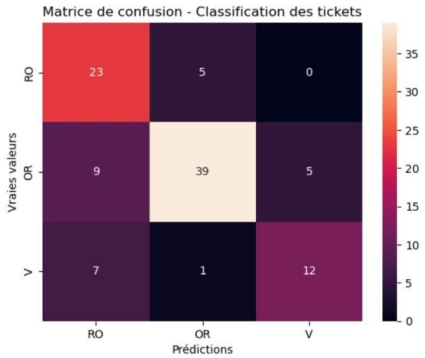


FIGURE 3.34 – Performance du modele classification applique aux donnee CHU

En conclusion, le modéle a bien prédit la majorité des patients de priorité Orange, comme le montre le grand nombre de prédictions correctes (39) pour cette catégorie. Cependant, il y a des erreurs de classification, notamment pour les patients de priorité Rouge et Verte. Les erreurs de classification se produisent principalement entre les catégories Rouge et Orange, et entre Orange et Verte. Cela signifie que le modéle a parfois du mal à distinguer

entre ces catégories.Finalement, on peut dire que cette matrice de confusion montre que le modéle est assez bon pour prédire les priorités des patients, mais il y a encore des erreurs qui pourraient être améliorées pour une meilleure précision.

Clustering des patients

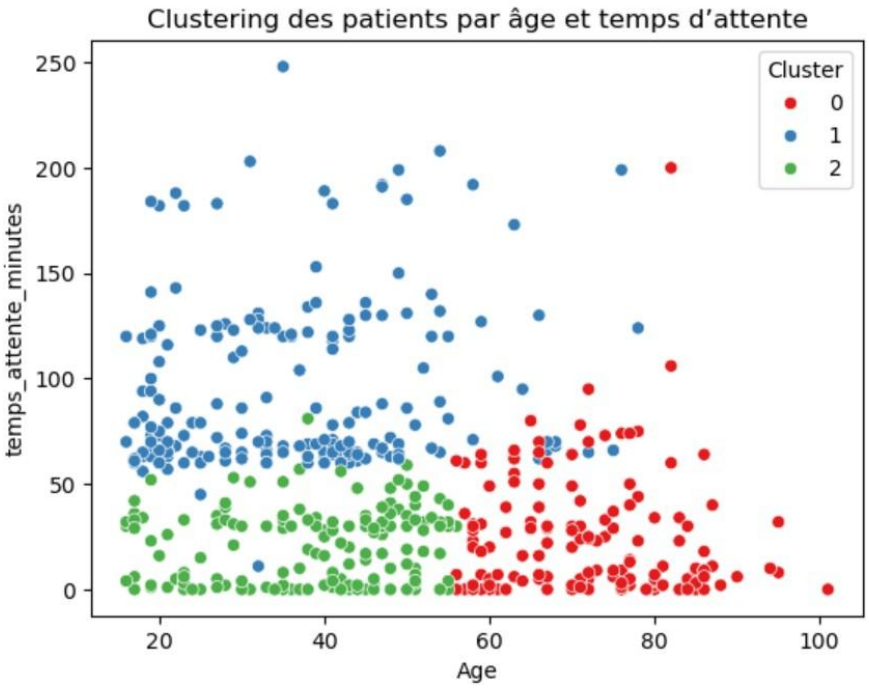


FIGURE 3.35 – Clustering des patients au CHU

— **Affiche les moyennes par cluster :**

GraviteLabel	Faible	Moyenne	Élevée
Cluster			
0	1	59	87
1	106	88	2
2	0	80	79

FIGURE 3.36 – Affiche les moyennes par cluster

— **interpretation des resultats :**

- **Moyennes par Cluster :**
  - **Cluster 0 : Age moyen : 72,5 ans.** Temps d'attente moyen : 24,1 minutes. **Gravité moyenne : 4,2** (sur une échelle où un nombre plus élevé indique une gravité plus élevé) Ce cluster représente des patients plus agés avec des temps d'attente relativement courts et une gravité élevé. Cela pourrait indiquer que les

patients plus agés et plus gravement malades sont prioritaires et donc vus plus rapidement.

— **Cluster 1 : Age moyen : 36,5 ans.** Temps d'attente moyen : 94,9 minutes. **Gravité moyenne : 1,9** Ce cluster représente des patients plus jeunes avec des temps d'attente plus longs et une gravité plus faible. Cela pourrait indiquer que les patients plus jeunes et moins gravement malades attendent plus longtemps.

— **Cluster 2 : Age moyen : 38,8 ans.** Temps d'attente moyen : 20,3 minutes. **Gravité moyenne : 4,0** Ce cluster représente des patients d'age moyen avec des temps d'attente courts et une gravité élevée. Cela pourrait indiquer que ces patients sont également prioritaires en raison de la gravité de leur condition.

- **Mesures de Qualité du Clustering :**
  — **Inertie du modéle :** avec une inertie de 649,57, cela suggére que les clusters sont raisonnablement bien formés, mais il pourrait y avoir des possibilités d'amélioration.
  — **Silhouette Score :** avec un score de 0,31, ca indique une séparation modérée des clusters, suggérant que les clusters sont quelque peu distincts, mais qu'il y a encore des chevauchements.

3.2.3 Tableau de bord interactif pour le CHU Khelil Amrane

Histogramme des Temps d'Attente par Niveau de Gravite

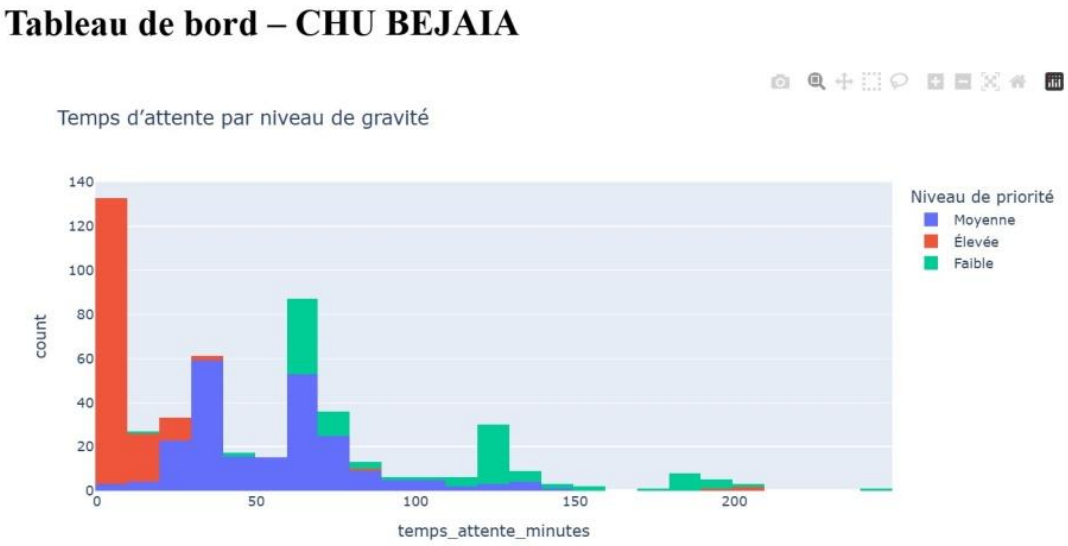


FIGURE 3.37 – Histogramme des Temps d'Attente par Niveau de Gravite au CHU

— **Interpretation :**

- **Gravité Faible (Orange) :**La majorité des patients avec une gravité faible ont des temps d'attente courts, principalement entre 0 et 50 minutes. Il y a un pic notable

autour de 0-20 minutes, indiquant que beaucoup de ces patients sont vus assez rapidement.

- **Gravité Moyenne (Bleu) :** Les patients avec une gravité moyenne ont une distribution plus étalée des temps d'attente, avec des pics autour de 20-50 minutes et 50-80 minutes. Ce qui suggére une variabilité plus grande dans les temps d'attente pour ces patients.
- **Gravité Élevée (Vert) :** Les patients avec une gravité élevée ont généralement des temps d'attente plus courts, principalement entre 0 et 50 minutes. Le pic autour de 0-20 minutes indique que ces patients sont souvent vus rapidement, ce qui est logique étant donné leur condition plus urgente.

Boxplot des Temps d'Attente par Priorité

FIGURE 3.38 – Boxplot des Temps d'Attente par Priorite au CHU

interpretation :

- **Priorité Moyenne :** La boite montre que l'intervalle interquartile (IQR) des temps d'attente pour les patients de priorité moyenne est centré autour de 50-75 minutes. Les "moustaches" s'étendent jusqu'à environ 120 minutes, avec quelques valeurs aberrantes au-delà de 200 minutes. Cela indique que la majorité des patients de priorité moyenne attendent entre 25 et 100 minutes, mais certains attendent beaucoup plus longtemps.
- **Priorité Élevée :** La boite montre que l'IQR des temps d'attente pour les patients de priorité élevée est centré autour de 0-25 minutes. Les moustaches sont courtes, indiquant peu de variabilité, et il y a quelques valeurs aberrantes autour de 50 minutes. Cela suggére que les patients de priorité élevée sont généralement vus trés rapidement, avec peu de variations dans les temps d'attente.
- **Priorité Faible :** La boite montre que l'IQR des temps d'attente pour les patients de priorité faible est centré autour de 75-125 minutes. Les moustaches s'étendent

jusqu'à environ 200 minutes, avec quelques valeurs aberrantes au-delà de 200 minutes. Cela indique que les patients de priorité faible attendent généralement plus longtemps, avec une variabilité significative dans les temps d'attente.

En conclusion, les visualisations montrent que les patients avec une gravité ou une priorité plus élevée sont généralement vus plus rapidement, ce qui est conforme aux attentes dans un environnement de soins de santé. Cependant, il y a une variabilité notable dans les temps d'attente,en particulier pour les patients de priorité moyenne et faible. Comprendre ces distributions peut aider à identifier les domaines où des améliorations peuvent etre apportées pour réduire les temps d'attente et améliorer l'efficacité des soins.

Conclusion

Dans ce chapitre, nous avons décri de maniére détaillée la méthodologie employée dans le projet. À travers la collecte, le nettoyage, l'exploration et la modélisation des données de l'EPH de Kherrata ou du CHU Khelil Amrane, nous avons mis en place un processus rigoureux et adapté au contexte hospitalier local.

4
Résultats et discussion

Sommaire

Introduction

Ce chapitre présente une analyse comparative des résultats obtenus à partir des données des deux établissements hospitaliers, l'EPH de Kherrata et le CHU Khelil Amrane de Béjaia. Nous discuterons des implications de ces résultats et proposerons des recommandations pour améliorer la gestion des flux de patients et optimiser les ressources dans ces établissements.

4.1 Analyse Comparative des Résultats

Nous présentons une comparaison des résultats sous forme de tableaux pour une meilleur compréhension du fonctionnement des deux établissements.

Comparaison des Temps d'Attente

Ce tableau donne les résultats des statistiques descriptives des deux établissements.

Établissement	Temps moyen	Variabilité	Répartition des temps d'attente
EPH de Kherrata	16.23 min	Écart-type de 19.72 min	25% < 4 min 50% < 10 min 75% < 20.5 min
CHU Khelil Amrane	50.53 min	Écart-type de 47.35 min	25% < 8 min 50% < 39 min 75% < 70 min

TABLE 4.1 – Comparaison des temps d'attente entre deux établissements Hospitaliers

On remarque bien que les temps d'attente au CHU Khelil Amrane sont significativement plus longs et plus variables que ceux de l'EPH de Kherrata. Cela peut être attribué à la taille et la complexité des services offerts par le CHU, ainsi qu'à la diversité des cas traités. Les temps d'attente plus courts à l'EPH de Kherrata pourraient être dus à une meilleure gestion des flux de patients ou à une charge de travail moins élevée.

Comparaison des Performances des Modéles de Classification

Ce tableau nous donne les résultats des différents modéles de classification qu'on avait utilisé dans notre des deux établissements.

Établissement	Précision	Rappel	F1-score	Précision globale
EPH de Kherrata	OR : 0.87 RO : 0.59 V : 0.71	OR : 0.74 RO : 0.82 V : 0.60	OR : 0.80 RO : 0.69 V : 0.65	0.73
CHU Khelil Amrane	OR : 0.81 RO : 0.70 V : 0.00	OR : 0.88 RO : 0.59 V : 0.00	OR : 0.84 RO : 0.64 V : 0.00	0.77

TABLE 4.2 – Évaluation des métriques de classification par catégorie et par établissement

On observe bien que les modéles de classification montrent des performances similaires dans les deux établissements, avec une légère amélioration au CHU Khelil Amrane pour la catégorie OR (Orange). Cependant, la performance pour la catégorie V (Vert) est nulle au CHU, ce qui pourrait indiquer un déséquilibre dans les données ou une difficulté à classer correctement cette catégorie.

Comparaison des Clusters

Ce tableau contient les différents clusters des patients pour les deux établissements hospitaliers que l'algorithme Kmeans nous a donné.

Établissement	Cluster	Âge moyen	Temps d'attente moyen	Gravité moyenne
EPH de Kherrata	0	33.89 ans	57.23 min	3.02
	1	20.50 ans	11.94 min	3.09
	2	59.02 ans	6.87 min	4.54
CHU Khelil Amrane	0	72.50 ans	24.08 min	4.17
	1	36.52 ans	94.92 min	1.94
	2	38.82 ans	20.25 min	3.99

TABLE 4.3 – Analyse comparative des clusters patients dans deux établissements hospitaliers

Les clusters montrent des différences significatives entre les deux établissements. Par exemple, le Cluster 1 du CHU Khelil Amrane a un temps d'attente moyen beaucoup plus élevé que celui de l'EPH de Kherrata, ce qui pourrait indiquer des inefficacités spécifiques au CHU.

Comparaison des Résultats des Tests Statistiques

Résultats du test statistique ANOVA

Établissement	Statistique F	Valeur-p
EPH de Kherrata	51.1568	0.0000
CHU Khelil Amrane	304,3927	0.0000

TABLE 4.4 – Résultats des tests statistiques par établissement

Les tests ANOVA effectués sur les données des deux établissements hospitaliers, nous montrent qu'il existe des différences significatives entre les groupes de gravité dans les deux établissements. La statistique F plus élevée au CHU Khelil Amrane indique une plus grande variabilité entre les groupes de gravité, ce qui pourrait être dû à une plus grande diversité des cas traités.

4.2 Recommandations pour l'Amélioration des Services

Optimisation des Ressources

Pour l'EPH de Kherrata, comme les temps d'attente sont relativement courts, la direction pourrait se concentrer sur le maintien de ces performances en optimisant l'allocation des ressources existantes. Cela pourrait inclure la formation continue du personnel pour gérer les pics d'affluence et l'utilisation de techniques de gestion des flux de patients pour améliorer l'efficacité. Par contre, une analyse plus approfondie des patients avec des temps d'attente plus longs pourrait révéler des inefficacités spécifiques ou des besoins particuliers nécessitant une attention supplémentaire. Cela pourrait inclure l'identification des goulots d'étranglement dans les processus et la mise en place de solutions ciblées pour les résoudre.

Pour le CHU Khelil Amrane, la direction, pourra demander à bénéficier d'une augmentation des ressources, notamment en personnel médical et en équipements, pour réduire les temps d'attente. Ce qui inclu, l'embauche de personnel supplémentaire, l'acquisition d'équipements supplémentaires, et l'amélioration des infrastructures pour mieux gérer la charge de travail. L'analyse des clusters pourrait aider à identifier les périodes et les services où les ressources supplémentaires sont le plus nécessaires. Cela pourrait inclure l'identification des groupes de patients avec des temps d'attente particuliérement longs et la mise en place de solutions ciblées pour améliorer leur gestion.

Amélioration des Modéles de Classification

Équilibrage des Données

Pour améliorer la performance des modéles de classification, il est crucial de s'assurer que les données sont équilibrées. Cela pourrait impliquer la collecte de plus de données pour les catégories sous-représentées, notamment pour la catégorie Vert au CHU Khelil Amrane. Cela pourrait inclure l'identification des sources de données supplémentaires et la mise en place de processus pour collecter ces données de maniére systématique.

L'utilisation de techniques d'échantillonnage pour équilibrer les données existantes pourrait également améliorer la performance des modéles. Cela pourrait inclure l'utilisation de techniques telles que le suréchantillonnage des catégories sous-représentées ou le sous-échantillonnage des catégories surreprésentées pour créer un ensemble de données plus équilibré.

Validation des Modéles

Il est important d'utiliser les techniques de validation croisée pour s'assurer que les modéles sont robustes et généralisables. Cela pourrait inclure l'utilisation de différentes métriques d'évaluation, telles que la précision, le rappel, le F1-score, et l'AUC, pour évaluer la performance des modéles sur différents ensembles de données

Utilisation des Résultats du Clustering

Ciblage des Ressources :

Les résultats du clustering peuvent être utilisés pour cibler les ressources là où elles sont le plus nécessaires. Par exemple, si un cluster montre un groupe de patients avec des temps d'attente particuliérement longs, des ressources supplémentaires pourraient être allouées pour ce groupe. Cela pourrait inclure l'identification des périodes de pointe et la mise en place de solutions pour mieux gérer ces périodes.

Les clusters peuvent également aider à personnaliser les soins en fonction des caractéristiques des patients. Par exemple, les patients plus agés avec des temps d'attente plus courts pourraient bénéficier de soins spécialisés pour améliorer leur expérience. Cela pourrait inclure la mise en place de programmes de soins spécifiques pour différents groupes de patients et l'utilisation de techniques de gestion des flux de patients pour améliorer l'efficacité des soins.

Amélioration de la Gestion des Données :

Assurer la qualité et la représentativité des données est crucial pour des analyses précises. Cela pourrait impliquer des audits réguliers des données pour identifier et corriger les erreurs et les incohérences. Cela pourrait inclure la mise en place de processus pour assurer la qualité des données et la formation du personnel sur l'importance de la collecte de donnés précises.

La formation du personnel sur l'importance de la collecte de données précises pourrait également améliorer la qualité des données. Cela pourrait inclure la mise en place de programmes de formation pour le personnel sur les meilleures pratiques de collecte de données et l'utilisation de techniques de nettoyage des données pour corriger les erreurs et les incohérences.

Intégrer des données supplèmentaires, telles que les données météorologiques ou les événements locaux, pourrait améliorer la précision des modéles prédictifs. Cela pourrait aider à mieux comprendre les facteurs externes qui influencent les temps d'attente et à mieux planifier les ressources en conséquence. Cela pourrait inclure l'identification des sources de données supplémentaires et la mise en place de processus pour intégrer ces données de maniÃ¨re systématique.

L'analyse des facteurs externes qui influencent les temps d'attente pourrait également améliorer la précision des modéles prédictifs. Cela pourrait inclure l'identification des tendances saisonniéres, des événements locaux, et d'autres facteurs externes qui pourraient influencer les temps d'attente, et la mise en place de solutions pour mieux gérer ces facteurs.

Formation et Sensibilisation du Personnel

Offrir une formation continue au personnel sur les meilleures pratiques de gestion des flux de patients et l'utilisation des outils d'analyse de données pourrait améliorer l'efficacité et la qualité des soins. Cela pourrait inclure la mise en place de programmes de formation pour le personnel sur les meilleures pratiques de gestion des flux de patients et l'utilisation de techniques d'analyse de données pour améliorer l'efficacité des soins.

La formation du personnel sur l'utilisation des outils d'analyse de données pourrait également améliorer l'efficacité et la qualité des soins. Cela pourrait inclure la mise en place de programmes de formation pour le personnel sur l'utilisation des outils d'analyse de données et l'interprétation des résultats pour améliorer la prise de décision.

Sensibiliser le personnel à l'importance des données et à leur rôle dans l'amélioration des services pourrait encourager une collecte de données plus précise et compléte. Cela pourrait inclure la mise en place de programmes de sensibilisation pour le personnel sur l'importance des données et leur rôle dans l'amélioration des services, et l'utilisation de techniques de collecte de données pour assurer la précision et la complétude des données.

La sensibilisation du personnel à l'importance de la collecte de données précises pourrait également améliorer la qualité des données. Cela pourrait inclure la mise en place de programmes de sensibilisation pour le personnel sur l'importance de la collecte de données précises et l'utilisation de techniques de collecte de données pour assurer la précision et la complétude des données.

Conclusion

En conclusion, ce chapitre, nous a permis de comparer les résultats obtenus à partir des analyses des données des deux établissements hospitaliers et de proposer des recommandations pour améliorer la gestion des flux de patients et optimiser les ressources. Les différences observées entre l'EPH de Kherrata et le CHU Khelil Amrane soulignent l'importance de comprendre les spécificités de chaque établissement pour proposer des solutions adaptées. Les recommandations proposées visent à améliorer la qualité des soins, réduire les temps d'attente, et optimiser l'utilisation des ressources pour répondre aux besoins des patients de maniére plus efficace et efficient. La mise en oeuvre de ces recommandations pourrait conduire à des améliorations significatives dans la gestion des flux de patients et l'optimisation des ressources, ce qui pourrait finalement améliorer la qualité des soins et la satisfaction des patients dans les deux établissements hospitaliers.

Conclusion générale et perspectives

L'étude comparative entre le CHU Khelil Amrane et l'EPH de Kherrata a révélé des différences significatives dans la gestion des flux de patients et des ressources. Les patients avec une priorité plus élevée bénéficient généralement de temps d'attente réduits. Cependant, une variabilité notable a été observée, en particulier pour les patients de priorité moyenne et faible, indiquant des goulots d'étranglement dans les processus actuels.

Les modéles de classification utilisés ont montré une performance globalement satisfaisante,permettant une gestion plus ciblée des ressources. Toutefois, des améliorations sont nécessaires pour mieux distinguer certaines catégories de priorité. L'analyse de clustering a permis d'identifier des groupes distincts de patients, aidant à mieux cibler les ressources et à améliorer les soins en fonction des caractéristiques spécifiques de chaque groupe. Les tests statistiques ont confirmé que les temps d'attente varient significativement en fonction de la gravité des cas.

Les résultats soulignent la nécessité d'une gestion plus efficace des ressources et d'une optimisation des processus pour réduire les temps d'attente et améliorer la qualité des soins. Les différences entre les deux établissements mettent en évidence l'importance de solutions adaptées à chaque contexte spécifique.

Les perspectives futures incluent l'amélioration des modéles de classification, l'intégration de données supplémentaires pour mieux comprendre les facteurs influencant les temps d'attente, et une analyse plus approfondie des clusters pour identifier les périodes de pointe et proposer des solutions ciblées. La formation continue du personnel et l'assurance de la qualité des données sont également essentielles pour améliorer l'efficacité et la qualité des soins.

Pour améliorer davantage l'efficacité et la qualité des soins, j'ai pris l'initiative de concevoir une application visant à optimiser l'accés aux services d'urgence. Cette application combine les temps d'attente en temps réel de chaque établissement hospitalier avec le positionnement GPS de l'usager.Cela permet une orientation optimale vers l'établissement le plus proche en termes de temps global pour accéder aux services d'urgence.

En conclusion, cette étude offre des perspectives précieuses pour optimiser la gestion des patients et des ressources dans les établissements de santé, visant à améliorer la qualité des soins et la satisfaction des patients.

Bibliographie

[1] K-nearest neighbor (knn) algorithm for machine learning javatpoint. https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning. Consulté en juillet 2025.

[2] Scikit-learn svm module. https://scikit-learn.org/stable/modules/svm.html. Consulté en juillet 2025.

[3] Régression logistique. https://www.wikiwand.com/fr/R%C3%A9gression_logistique, 2011. Consulté en juillet 2025.

[4] Clustering — analytics vidhya. https://www.analyticsvidhya.com/blog/2016/11/anintroduction-to-clustering-and-different-methods-of-clustering/, 2016. Consulté en juillet 2025.

[5] Prophet : Forecasting at scale. https://facebook.github.io/prophet/, 2025. Consulté en juillet 2025.

[6] Benamrane, Y., and Belkacem, R. *Un personnel insuffisant ou mal réparti ralentit la prise en charge des patients*. 2020.

[7] Boubou, M. *Méthodes de classification non supervisée via des approches prétopologiques et d'agrégation d'opinions*. Thése de doctorat, Université Claude Bernard Lyon 1, 2007.

[8] Boukhalfa, K., and Moussaoui, A. Analyse des flux de patients dans les services d'urgence des chu algériens. *Revue Algérienne de Santé Publique 15*, 2 (2017), 45–58.

[9] Boukhalfa, K., and Moussaoui, A. Analyse des flux de patients dans les services d'urgence des chu algériens. *Revue Algérienne de Santé' Publique* (2017), 48.

[10] Bousbaci, A. *Algorithmes paralléles de clustering de données*. Thése de doctorat, Université de Béjaia, 2019.

[11] Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD)* (1996), pp. 226–231.

[12] Goodfellow, I., Bengio, Y., and Courville, A. *Deep Learning*. MIT Press, 2016.

[13] Han, J., Kamber, M., and Pei, J. *Data Mining : Concepts and Techniques*, 3rd ed. Morgan Kaufmann, San Francisco, CA, USA, 2011.

[14] Hardy, C. *Contribution au développement de l'apprentissage profond dans les systémes distribués*. Thése de doctorat, Université de Rennes 1, 2019.

[15] Hilali, H. *Application de la classification textuelle pour l'extraction des régles d'association maximales*. Thése de doctorat, Université du Québec Ã Trois-Riviéres, 2009.

[16] Hoot, N. R., and Aronsky, D. Systematic review of emergency department crowding : causes, effects, and solutions. *Annals of Emergency Medicine* (2008).

[17] Khatir, N., and Nait-Bahloul, S. Multi-criteria-based fusion for clustering texts and images : case study on flickr. *Kybernetes* (2018).

[18] Le, X. H., Nguyen, H., Modi, A., and McIntyre, M. G. Application of long short-term memory (lstm) neural network for flood forecasting. *Water 11*, 7 (2019), 1387.

[19] Murtagh, F., and Contreras, P. Algorithms for hierarchical clustering : an overview. *Wiley Interdisciplinary Reviews : Data Mining and Knowledge Discovery 2*, 1 (2012), 86–97.

[20] Provost, F., and Fawcett, T. *Data Science for Business*. O'Reilly Media, 2013.

[21] Schott, M. K-means clustering algorithm for machine learning. `https://medium.com/capital-one-tech/k-means-clustering-algorithmformachine-learning-d1d7dc5de882`, 2020. Medium Capital One Tech. Consulté en juillet 2025.

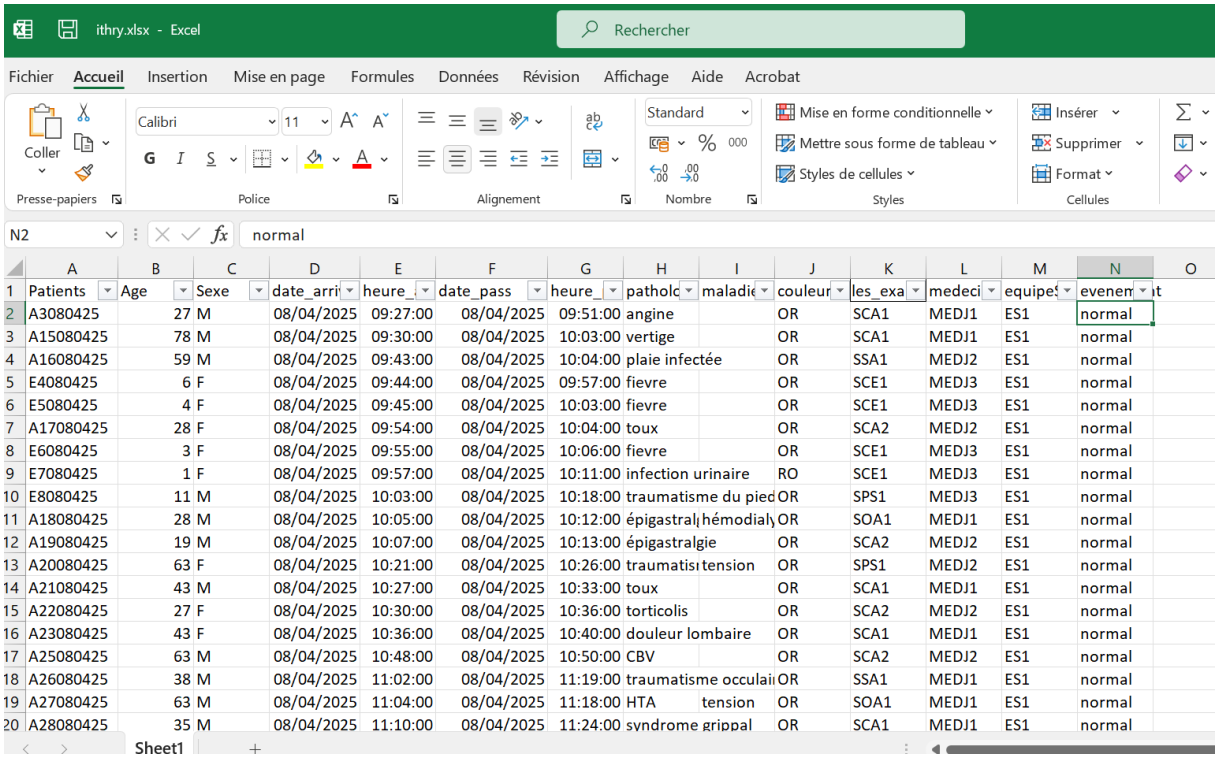
[22] Scicluna, N., and Bouganis, C.-S. Arc 2014 : a multidimensional fpga-based parallel dbscan architecture. *ACM Transactions on Reconfigurable Technology and Systems (TRETS) 9*, 1 (2015), 1–15.

[23] Sun, X., et al. *Some article or book*. Unknown Publisher, 2013.

[24] Yahi, A. *Clustering des données de puces à ADN*. Thése de doctorat, Université' Mohamed Boudiaf M'Sila, 2019.

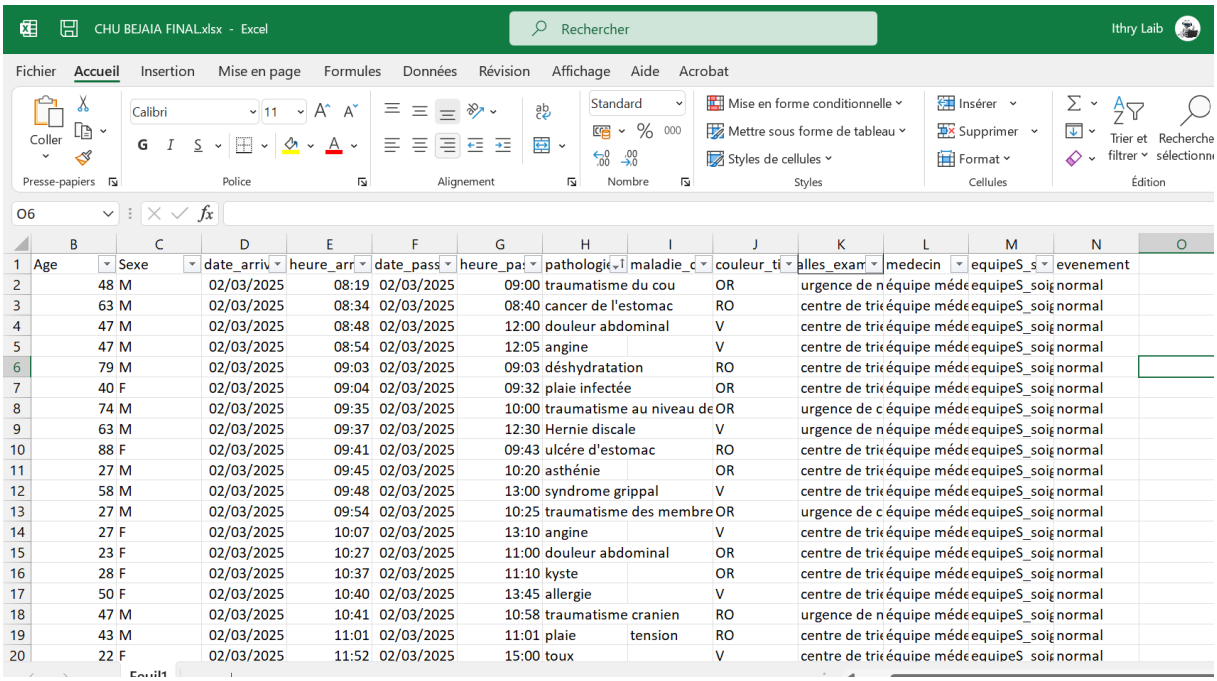
Annexes

4.3 Annexes 1



	Patients	Age	Sexe	date_arri	heure_	date_pass	heure_	patholc	maladie	couleur	les_exa	medeci	equipe	evenem
2	A3080425	27	M	08/04/2025	09:27:00	08/04/2025	09:51:00	angine		OR	SCA1	MEDJ1	ES1	normal
3	A15080425	78	M	08/04/2025	09:30:00	08/04/2025	10:03:00	vertige		OR	SCA1	MEDJ1	ES1	normal
4	A16080425	59	M	08/04/2025	09:43:00	08/04/2025	10:04:00	plaie infectée		OR	SSA1	MEDJ2	ES1	normal
5	E4080425	6	F	08/04/2025	09:44:00	08/04/2025	09:57:00	fievre		OR	SCE1	MEDJ3	ES1	normal
6	E5080425	4	F	08/04/2025	09:45:00	08/04/2025	10:03:00	fievre		OR	SCE1	MEDJ3	ES1	normal
7	A17080425	28	F	08/04/2025	09:54:00	08/04/2025	10:04:00	toux		OR	SCA2	MEDJ2	ES1	normal
8	E6080425	3	F	08/04/2025	09:55:00	08/04/2025	10:06:00	fievre		OR	SCE1	MEDJ3	ES1	normal
9	E7080425	1	F	08/04/2025	09:57:00	08/04/2025	10:11:00	infection urinaire		RO	SCE1	MEDJ3	ES1	normal
10	E8080425	11	M	08/04/2025	10:03:00	08/04/2025	10:18:00	traumatisme du pied		OR	SPS1	MEDJ3	ES1	normal
11	A18080425	28	M	08/04/2025	10:05:00	08/04/2025	10:12:00	épigastral	hémodial	OR	SOA1	MEDJ1	ES1	normal
12	A19080425	19	M	08/04/2025	10:07:00	08/04/2025	10:13:00	épigastralgie		OR	SCA2	MEDJ2	ES1	normal
13	A20080425	63	F	08/04/2025	10:21:00	08/04/2025	10:26:00	traumatis	tension	OR	SPS1	MEDJ2	ES1	normal
14	A21080425	43	M	08/04/2025	10:27:00	08/04/2025	10:33:00	toux		OR	SCA1	MEDJ1	ES1	normal
15	A22080425	27	F	08/04/2025	10:30:00	08/04/2025	10:36:00	torticolis		OR	SCA2	MEDJ2	ES1	normal
16	A23080425	43	F	08/04/2025	10:36:00	08/04/2025	10:40:00	douleur lombaire		OR	SCA1	MEDJ1	ES1	normal
17	A25080425	63	M	08/04/2025	10:48:00	08/04/2025	10:50:00	CBV		OR	SCA2	MEDJ2	ES1	normal
18	A26080425	38	M	08/04/2025	11:02:00	08/04/2025	11:19:00	traumatisme occulai		OR	SSA1	MEDJ1	ES1	normal
19	A27080425	63	M	08/04/2025	11:04:00	08/04/2025	11:18:00	HTA	tension	OR	SOA1	MEDJ1	ES1	normal
20	A28080425	35	M	08/04/2025	11:10:00	08/04/2025	11:24:00	syndrome grippal		OR	SCA1	MEDJ1	ES1	normal

FIGURE 4.1 – Table des donnees EPH



	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Age	Sexe	date_arriv	heure_arr	date_pass	heure_pas	pathologie	maladie_c	couleur_ti	alles_exam	medecin	equipeS_s	evenement	
2	48	M	02/03/2025	08:19	02/03/2025	09:00	traumatisme du cou		OR	urgence de n	équipe méde	equipeS_soi	normal	
3	63	M	02/03/2025	08:34	02/03/2025	08:40	cancer de l'estomac		RO	centre de tri	équipe méde	equipeS_soi	normal	
4	47	M	02/03/2025	08:48	02/03/2025	12:00	douleur abdominal		V	centre de tri	équipe méde	equipeS_soi	normal	
5	47	M	02/03/2025	08:54	02/03/2025	12:05	angine		V	centre de tri	équipe méde	equipeS_soi	normal	
6	79	M	02/03/2025	09:03	02/03/2025	09:03	déshydratation		RO	centre de tri	équipe méde	equipeS_soi	normal	
7	40	F	02/03/2025	09:04	02/03/2025	09:32	plaie infectée		OR	centre de tri	équipe méde	equipeS_soi	normal	
8	74	M	02/03/2025	09:35	02/03/2025	10:00	traumatisme au niveau de		OR	urgence de c	équipe méde	equipeS_soi	normal	
9	63	M	02/03/2025	09:37	02/03/2025	12:30	Hernie discale		V	urgence de n	équipe méde	equipeS_soi	normal	
10	88	F	02/03/2025	09:41	02/03/2025	09:43	ulcére d'estomac		RO	centre de tri	équipe méde	equipeS_soi	normal	
11	27	M	02/03/2025	09:45	02/03/2025	10:20	asthénie		OR	centre de tri	équipe méde	equipeS_soi	normal	
12	58	M	02/03/2025	09:48	02/03/2025	13:00	syndrome grippal		V	centre de tri	équipe méde	equipeS_soi	normal	
13	27	M	02/03/2025	09:54	02/03/2025	10:25	traumatisme des membre		OR	urgence de c	équipe méde	equipeS_soi	normal	
14	27	F	02/03/2025	10:07	02/03/2025	13:10	angine		V	centre de tri	équipe méde	equipeS_soi	normal	
15	23	F	02/03/2025	10:27	02/03/2025	11:00	douleur abdominal		OR	centre de tri	équipe méde	equipeS_soi	normal	
16	28	F	02/03/2025	10:37	02/03/2025	11:10	kyste		OR	centre de tri	équipe méde	equipeS_soi	normal	
17	50	F	02/03/2025	10:40	02/03/2025	13:45	allergie		V	centre de tri	équipe méde	equipeS_soi	normal	
18	47	M	02/03/2025	10:41	02/03/2025	10:58	traumatisme cranien		RO	urgence de n	équipe méde	equipeS_soi	normal	
19	43	M	02/03/2025	11:01	02/03/2025	11:01	plaie	tension	RO	centre de tri	équipe méde	equipeS_soi	normal	
20	22	F	02/03/2025	11:52	02/03/2025	15:00	toux		V	centre de tri	équipe méde	equipeS_soi	normal	

FIGURE 4.2 – Table des donnees CHU

Résumé

Ce mémoire étudie l'utilisation du Data Mining pour améliorer la gestion des urgences dans les hôpitaux algériens, en prédisant les temps d'attente et en priorisant les cas urgents grâce à l'analyse de données. Les données, collectées en avril et en mars 2025, incluent des informations sur les patients et le personnel.

La méthodologie suit le processus CRISP-DM, utilisant des algorithmes comme Random Forest et un tableau de bord interactif pour visualiser les résultats. Les modéles prédisent efficacement les temps d'attente et identifient les cas prioritaires, bien qu'ils aient des difficultés avec les cas non urgents.

Ce travail souligne l'importance de la Data Science dans les services d'urgence et propose des améliorations futures, comme l'intégration de données externes. En conclusion, le mémoire montre comment le Data Mining peut optimiser la gestion hospitalière en Algérie.

Mots clés : Data Mining, temps d'attente, Random Forest, CRISP-DM, gestion hospitalière, tableau de bord interactif, gestion des urgences, priorisation des cas urgents, Data Science.

Abstract

This thesis examines the use of Data Mining to enhance emergency management in Algerian hospitals by predicting patient waiting times and prioritizing urgent cases through data analysis. The data, collected in March and April 2025, includes information about patients and staff.

The methodology follows the CRISP-DM process, employing algorithms like Random Forest and an interactive dashboard to visualize results. The models effectively predict waiting times and identify priority cases, although they struggle with non-urgent cases.

This work highlights the importance of Data Science in emergency services and suggests future improvements, such as integrating external data. In conclusion, the thesis demonstrates how Data Mining can optimize hospital management in Algeria.

Keywords : Data Mining, waiting times, Random Forest, CRISP-DM, hospital management, interactive dashboard, emergency management, prioritization of urgent cases, Data Science.