

République Algérienne Démocratique et Populaire
Ministre de l'Enseignement Supérieur et de la Recherche Scientifique
Ecole Nationale supérieure en Sciences et Technologies de l'Informatique Béjaïa



MODULE : SCIENCE DE DONNÉES

T H È M E

Analyse de Données

ENCADRÉ PAR :

Pr. Khernachi Hamid

MAÎTRE DE STAGE :

M. Chaoui Akrem

PRÉSENTÉ PAR :

Mahdaoui Malak ines

Date de début et fin de projet :15/06/2025 - 15/08/2025

Année universitaire : 2025/2026

REMERCIEMENTS

Je tiens à exprimer ma profonde gratitude à toutes les personnes qui ont contribué au bon déroulement de mon stage au sein du Centre Hospitalo-Universitaire (CHU) de Sétif.

Tout d'abord, je remercie sincèrement Monsieur le Professeur Khernachi Hamid, mon encadrant universitaire, pour son accompagnement, ses conseils précieux et son suivi rigoureux tout au long de ce stage. Sa disponibilité et son expertise m'ont grandement aidée à structurer mon travail et à approfondir mes connaissances dans le domaine de la science des données appliquée au secteur médical.

Je souhaite également remercier Monsieur Chaoui Akrem, mon maître de stage au CHU de Sétif, pour m'avoir accueillie au sein de l'équipe et m'avoir confié des missions enrichissantes. Sa confiance et son encadrement m'ont permis de mettre en pratique mes compétences en analyse de données dans un contexte réel et concret.

Enfin, je tiens à remercier toute l'équipe du service Informatique pour leur accueil chaleureux, leur disponibilité et leur collaboration. Ce stage m'a permis de découvrir un environnement professionnel stimulant, au cœur des enjeux actuels liés à la gestion des données de santé.

Merci à tous pour cette expérience à la fois formatrice, humaine et professionnelle.

RÉSUMÉ

Ce rapport décrit mon expérience de stage au sein du Centre Hospitalo-Universitaire (CHU) de Sétif, un établissement de santé publique de référence dans la région. L'objectif principal de ce stage était d'explorer, nettoyer et analyser les données médicales dans le but d'en extraire des informations utiles à la prise de décision et à l'amélioration de la qualité des soins.

Les tâches réalisées comprenaient la collecte et la préparation des données, l'analyse statistique, la création de visualisations, ainsi que le développement de tableaux de bord interactifs pour appuyer les professionnels de santé dans leur gestion quotidienne.

Les résultats obtenus ont permis de mieux comprendre certaines tendances médicales, d'identifier des indicateurs clés de performance hospitalière, et de proposer des pistes d'optimisation dans l'organisation des services. Ce stage a ainsi contribué à démontrer l'intérêt de l'analyse de données dans le domaine hospitalier.

TABLE DES MATIÈRES

| | | |
|-----------|--|-----------|
| I | Présentation du sujet de stage | 1 |
| I.1 | Présentation de l'établissement | 1 |
| I.2 | Objectifs du stage : | 2 |
| I.3 | L'importance de l'étude de la science des données : | 2 |
| I.4 | Tâches réalisées : | 3 |
| I.4.1 | Récupération et organisation des données hospitalières | 3 |
| I.4.2 | Nettoyage et préparation des données | 4 |
| I.4.3 | Analyse descriptive (KPI) | 4 |
| I.4.4 | Visualisation des résultats | 4 |
| I.4.5 | Dashboard Interactif (Dash) | 6 |
| I.4.6 | Modèle prédictif (Machine Learning) | 6 |
| I.4.7 | Évaluation des performances du modèle | 7 |
| I.4.8 | Communication des résultats | 9 |
| II | Conclusion générale | VI |

TABLE DES FIGURES

| | | |
|-----|--|---|
| I.1 | Schéma relationnel-ERD | 3 |
| I.2 | Histogramme des âges | 5 |
| I.3 | Répartition par sexe | 5 |
| I.4 | Fréquence des diagnostics | 5 |
| I.5 | Matrice de confusion (Réel vs Prédit) | 7 |
| I.6 | Comparaison des valeurs réelles et prédites (encodées) | 8 |
| I.7 | Rapport de classification | 8 |

INTRODUCTION GÉNÉRALE

Dans un monde de plus en plus axé sur l'information, la donnée est devenue un actif stratégique au cœur des processus décisionnels dans tous les secteurs d'activité. Dans le domaine de la santé, cette réalité est d'autant plus marquée que les enjeux sont directement liés à la vie humaine, à la qualité des soins et à l'efficacité des structures hospitalières. Chaque jour, les hôpitaux produisent un volume considérable de données : dossiers médicaux électroniques, résultats d'analyses biologiques, imagerie médicale, enregistrements d'admission et de sortie, historiques de traitements, etc.

Ces données, souvent hétérogènes et réparties entre plusieurs services, représentent une source précieuse d'informations lorsqu'elles sont bien exploitées. Leur analyse permet de mieux comprendre les dynamiques internes de l'hôpital, de suivre l'évolution de certaines pathologies, de repérer des dysfonctionnements organisationnels, et surtout, de soutenir des prises de décision éclairées. L'exploitation intelligente de ces informations participe à l'amélioration continue de la qualité des soins, à l'optimisation des ressources humaines et matérielles, et à la modernisation de la gestion hospitalière.

Avec l'essor des technologies de l'information et des outils de science des données, l'analyse des données médicales s'impose aujourd'hui comme un levier essentiel dans la transformation numérique du secteur de la santé. Elle permet d'identifier des tendances épidémiologiques, de construire des modèles prédictifs sur l'affluence ou les complications, d'évaluer la performance des services hospitaliers, et de détecter précocement certaines situations critiques. Cette approche, connue sous le nom de « médecine fondée sur les données » (data-driven healthcare), tend à devenir un standard dans les établissements hospitaliers modernes.

Cependant, cette transformation s'accompagne aussi de plusieurs défis : la qualité et la fiabilité des données collectées, la diversité des formats, la protection de la vie privée des patients, l'interopérabilité des systèmes d'information hospitaliers (SIH), et la maîtrise des outils techniques nécessaires à l'analyse (Python, Power BI, SQL, etc.).

Dans ce contexte, le Centre Hospitalo-Universitaire (CHU) de Sétif, établissement de santé de référence dans la région, bénéficie d'un important gisement de données médicales et administratives. La mise en place d'un système rigoureux d'analyse de ces données représente un atout majeur pour le pilotage des services, l'évaluation des activités, et l'anticipation des besoins futurs.

L'analyse de données au sein d'un établissement comme le CHU de Sétif présente de nombreux avantages :

- **Amélioration de la prise de décision médicale** : grâce à des indicateurs fiables, les médecins peuvent adapter leurs pratiques aux réalités statistiques observées (durée moyenne d'hospitalisation, taux de réadmission, fréquence des diagnostics, etc.).
- **Optimisation de la gestion des ressources** : l'analyse des flux de patients, des périodes de forte activité ou des taux d'occupation des lits permet d'ajuster les plannings, de mieux répartir les équipes, et d'éviter les engorgements.
- **Appui stratégique à l'administration hospitalière** : à travers des tableaux de bord dynamiques, les responsables peuvent suivre en temps réel l'activité de chaque service et anticiper les évolutions futures.
- **Contribution à l'amélioration continue de la qualité des soins** : l'analyse des incidents, des retours patients, ou des durées d'attente permet de mettre en place des actions correctives ciblées.

C'est dans cette logique que s'est déroulé mon stage au sein du CHU de Sétif. L'objectif principal de cette mission était de mettre en place un processus complet d'analyse de données hospitalières, depuis la collecte brute jusqu'à la restitution visuelle des résultats. Plus précisément, les tâches réalisées ont porté sur :

- la modélisation relationnelle de la base de données,
- le nettoyage et la préparation des fichiers Excel à l'aide de Python,
- l'analyse statistique descriptive (KPI, moyennes, fréquences),
- la visualisation graphique avec matplotlib, seaborn et Power BI,
- le développement d'un tableau de bord interactif (Dash/Power BI),
- l'expérimentation d'un modèle prédictif simple (régression logistique).

Ce stage a représenté une opportunité concrète d'appliquer les concepts théoriques de la science des données dans un environnement hospitalier réel, combinant enjeux technologiques, humains et sociaux.

PRÉSENTATION DU SUJET DE STAGE

I.1 Présentation de l'établissement

- **Nom de l'établissement** : Centre Hospitalo-Universitaire (CHU) de Sétif
- **Secteur d'activité** : Santé publique – formation universitaire
- **Historique** : Le CHU de Sétif est l'un des principaux établissements hospitaliers de la région Est de l'Algérie. Il joue un rôle essentiel dans la prise en charge des patients, l'enseignement médical (en collaboration avec la faculté de médecine de l'Université Ferhat Abbas Sétif 1) et la recherche clinique. Grâce à ses nombreuses spécialités, il assure des soins de haut niveau dans divers domaines tels que la médecine générale, la chirurgie, la gynécologie, la pédiatrie, la réanimation, et bien d'autres.
- **Missions principales** :
 - Assurer les soins hospitaliers et ambulatoires aux patients
 - Participer à la formation des étudiants en médecine et du personnel paramédical
 - Contribuer à la recherche médicale et scientifique
 - Améliorer continuellement la qualité et la sécurité des soins
- **Produits et services** :
 - Hospitalisation et consultations spécialisées
 - Examens médicaux et services de diagnostic
 - Activités de recherche clinique
 - Gestion et analyse des données médicales pour le pilotage hospitalier

- **Structure organisationnelle** : Le CHU de Sétif est composé de plusieurs pôles (pôle médico-chirurgical, pôle mère-enfant, pôle urgence-réanimation, etc.) ainsi que de directions fonctionnelles, dont la Direction des systèmes d'information. Cette dernière est en charge de la gestion des données médicales, de l'informatisation des services, et du soutien aux projets d'analyse de données.

I.2 Objectifs du stage :

- Comprendre le fonctionnement des systèmes d'information hospitaliers.
- Participer à la collecte, au nettoyage et à l'analyse de données médicales.
- Réaliser des tableaux de bord et des visualisations utiles pour les praticiens.
- Proposer des recommandations sur l'amélioration de l'exploitation des données.

I.3 L'importance de l'étude de la science des données :

Dans un monde où la quantité de données générées croît de façon exponentielle, la science des données s'impose comme une discipline clé du XXI^e siècle. Que ce soit dans les domaines de la santé, de la finance, de l'industrie, de l'éducation ou encore des services publics, les organisations cherchent à transformer les données brutes en informations utiles, exploitables et à forte valeur ajoutée. C'est dans ce contexte que la science des données devient un outil fondamental pour orienter les décisions, automatiser les processus, et prédire les tendances futures.

La science des données regroupe un ensemble de techniques issues de plusieurs domaines : les statistiques, l'intelligence artificielle, le machine learning, le traitement de données massives (Big Data), la visualisation, et la gestion de bases de données. Elle permet de comprendre les phénomènes complexes à travers l'analyse descriptive, l'analyse prédictive et l'analyse prescriptive.

Son importance est particulièrement visible dans les domaines sensibles comme la santé, où l'analyse des données peut améliorer la prise en charge des patients, détecter des épidémies, suivre l'évolution des maladies chroniques, ou encore évaluer la qualité des soins. Dans le secteur hospitalier, elle aide à optimiser les ressources, à surveiller les indicateurs de performance, et à renforcer la sécurité et l'efficacité des services.

Étudier la science des données, c'est donc acquérir les compétences nécessaires pour faire parler les données, résoudre des problèmes concrets, et contribuer à l'innovation dans tous les secteurs. C'est aussi participer activement à la transformation numérique de la société en mettant la donnée au service de l'intérêt collectif.

I.4 Tâches réalisées :

Au cours de mon stage au sein du Centre Hospitalo-Universitaire (CHU) de Sétif, j'ai eu l'opportunité de participer à différentes missions liées à l'analyse de données hospitalières. Ces tâches m'ont permis d'appliquer les connaissances acquises en formation tout en découvrant les spécificités de la gestion des données dans le secteur médical.

Chaque étape a contribué à la création de visualisations exploitables et d'indicateurs pertinents pour le pilotage de l'activité hospitalière. Voici une description détaillée des principales tâches réalisées :

I.4.1 Récupération et organisation des données hospitalières

- **Description :** Extraction de données à partir d'un fichier Excel contenant plusieurs feuilles (Patients, Médecins, Consultations, Diagnostics, Traitements).
- La base hopitaldb.xlsx contient plusieurs tables :
 - Patients (IDPatient, Nom, Sexe, Age.)
 - Médecins (IDMedecin, Nom, Spécialité.)
 - Consultations (IDConsultation, Date, IDPatient, IDMedecin.)
 - Diagnostics (IDConsultation, Diagnostic)
 - Traitements (IDTraitement, IDConsultation, Médicament, Durée.)
- **Outils utilisés :** Excel, Python (pandas).
- **Difficultés rencontrées :** Données issues de sources hétérogènes, Données sur les papiers.
- **Solutions apportées :** Structuration en base de données relationnelle (avec clés primaires et clés étrangères) pour mieux représenter les relations entre entités (Patients , Consultations , Diagnostics ,Traitements).
- **Livrable associé :**

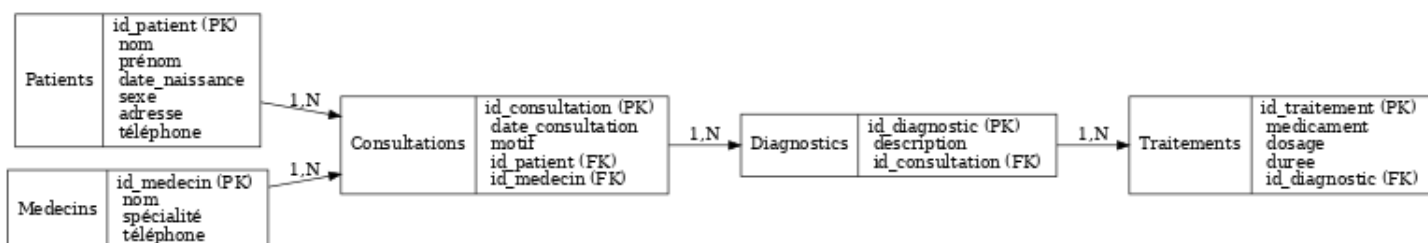


FIGURE I.1 – Schéma relationnel-ERD

I.4.2 Nettoyage et préparation des données

- **Description** : Gestion des doublons, correction des types de variables (ex. conversion de “Durée du traitement = X jours” en entier), imputation des valeurs manquantes (âge, sexe).
- **Outils utilisés** : Python (pandas, numpy).
- **Difficultés rencontrées** : Valeurs manquantes fréquentes dans certains services, doublons non détectés automatiquement.
- **Solutions apportées** : Mise en place de règles automatiques de remplissage et d’harmonisation.

I.4.3 Analyse descriptive (KPI)

- **Description** : Calcul des indicateurs clés pour le pilotage hospitalier :
 - Nombre total de patients : 50
 - Nombre de médecins : 50
 - Nombre de consultations : 50
 - Nombre de diagnostics : 50
 - Durée moyenne des traitements : 5.5 jours
- **Outils utilisés** : Python (pandas, matplotlib, seaborn), Excel.
- **Difficultés rencontrées** : Données brutes peu exploitables sans nettoyage préalable.
- **Solutions apportées** :
 - Regroupement par catégories.
 - Transformation des variables qualitatives en indicateurs quantitatifs.
 - Standardisation des échelles pour comparer les résultats.

I.4.4 Visualisation des résultats

- **Description** : Création de graphiques pour analyser les distributions et fréquences.
- **Outils utilisés** : Power BI, matplotlib, seaborn.
- **Graphiques réalisés** : Voir les figures ci-dessous.

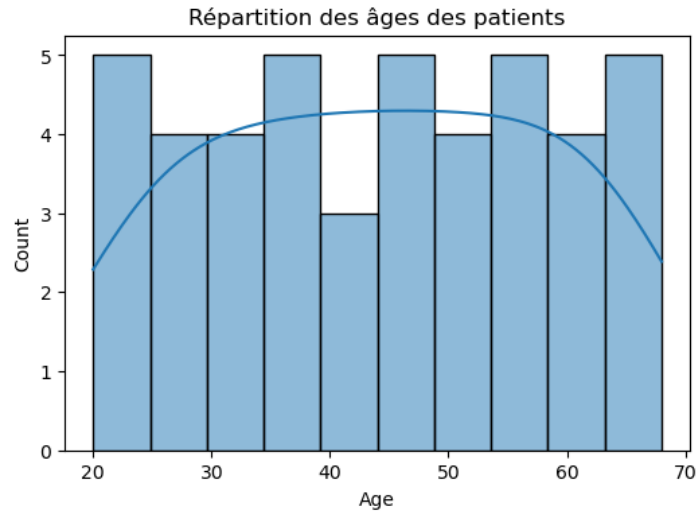


FIGURE I.2 – Histogramme des âges



FIGURE I.3 – Répartition par sexe

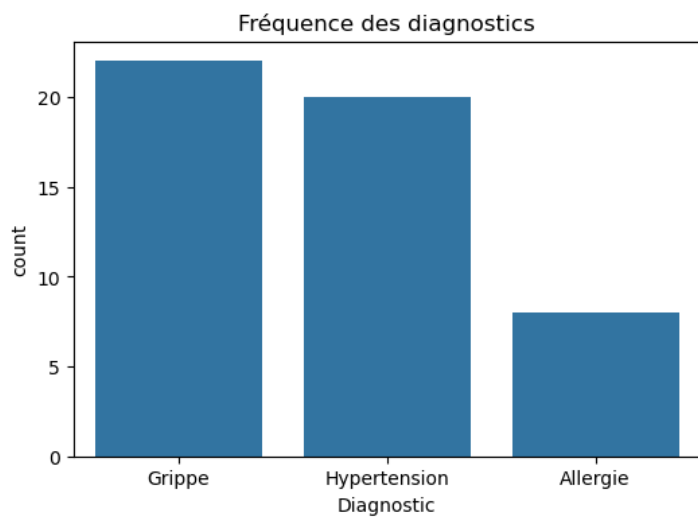
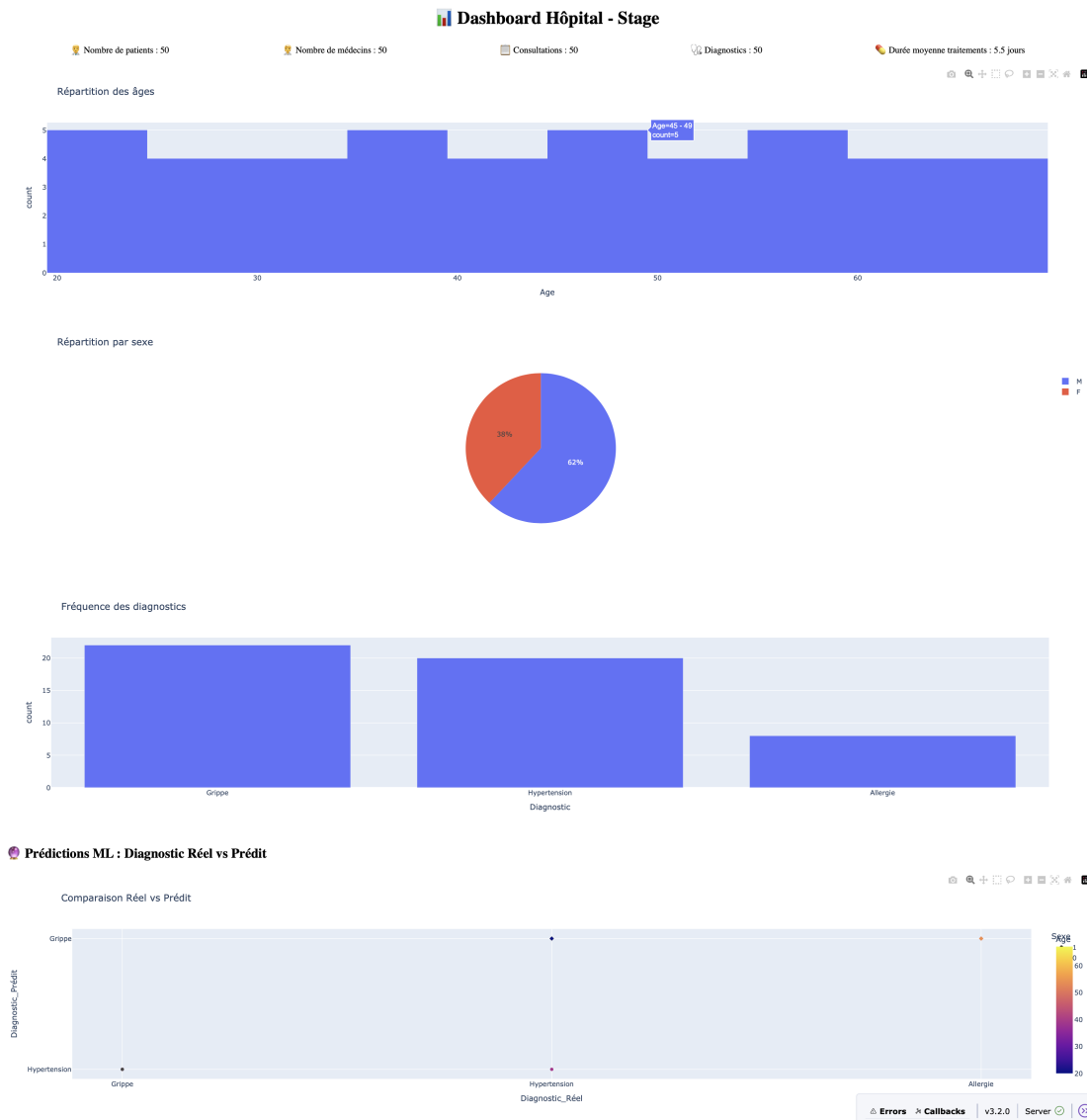


FIGURE I.4 – Fréquence des diagnostics

I.4.5 Dashboard Interactif (Dash)

— Présentation :

- Application développée en Python Dash
- Intégration de Plotly Express pour la visualisation
- L'application est accessible en local via `http://127.0.0.1:8050`



I.4.6 Modèle prédictif (Machine Learning)

- **Objectif :** prédire le Diagnostic à partir de l'Âge et du Sexe.
- **Algorithme :** Régression Logistique multi-classes, pipeline = Imputer (mean) + Standard-Scaler + LogisticRegression.
- **Jeu d'entraînement/test :** 80

I.4.7 Évaluation des performances du modèle

(a) **Matrice de confusion** La figure I.5 montre la matrice de confusion obtenue. Les axes représentent les classes réelles (verticale) et prédites (horizontale) : *Grippe*, *Hypertension*, *Allergie*. On observe des confusions marquées, ce qui traduit un modèle peu performant (features trop simples).

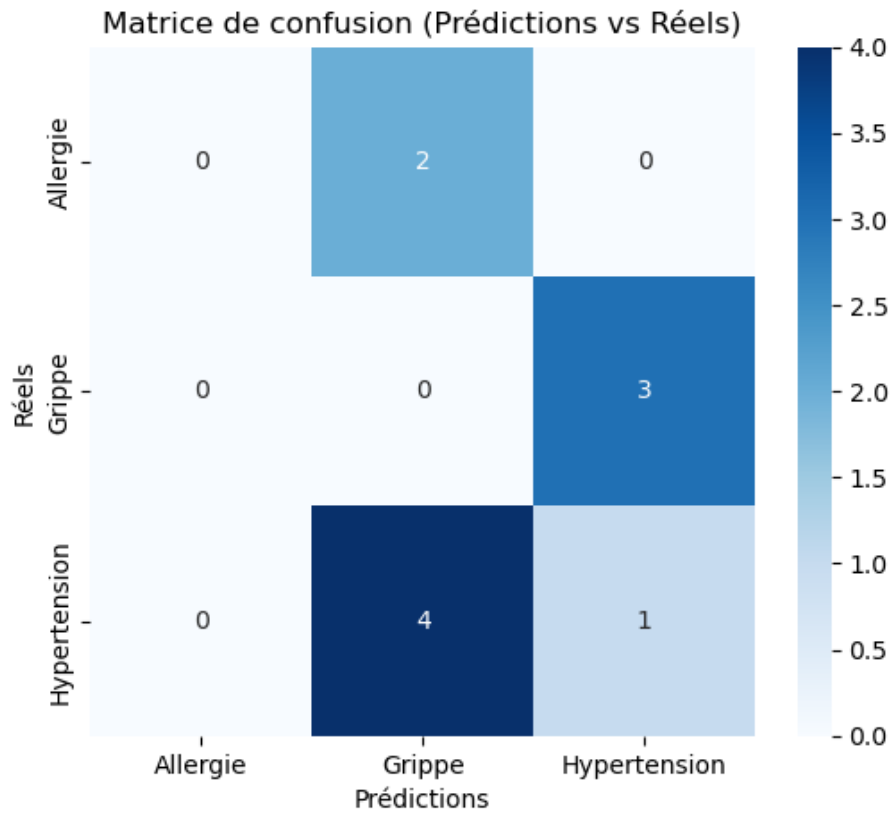


FIGURE I.5 – Matrice de confusion (Réal vs Prédit)

(b) **Comparaison Réel vs Prédit (encodé)** La figure I.6 compare les diagnostics réels et prédits, représentés sous forme encodée (numérique). Chaque point correspond à un échantillon de test : la série bleue correspond aux réels et la série orange aux prédits. Cela permet de visualiser directement les désaccords entre les deux.

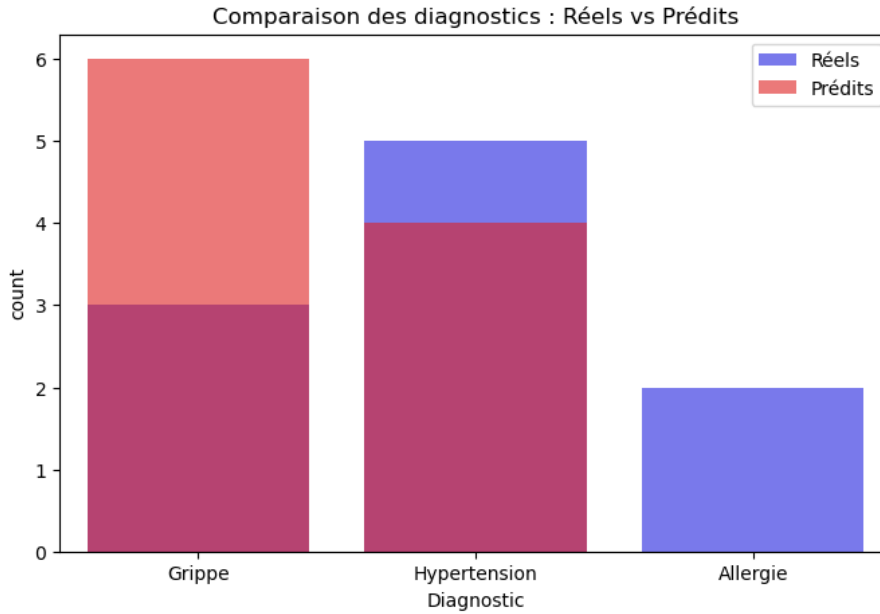


FIGURE I.6 – Comparaison des valeurs réelles et prédites (encodées)

(c) **Rapport de classification** La figure I.7 contient les scores de précision, rappel et F1-score par classe. Les résultats sont globalement faibles, confirmant que l'âge et le sexe seuls ne suffisent pas à prédire correctement un diagnostic.

```

=== Rapport de classification ===
      precision    recall  f1-score   support

 Allergie         0.00      0.00      0.00         2
  Grippe          0.00      0.00      0.00         3
Hypertension      0.25      0.20      0.22         5

 accuracy         0.10         10
 macro avg        0.08      0.07      0.07         10
 weighted avg     0.12      0.10      0.11         10
  
```

FIGURE I.7 – Rapport de classification

I.4.8 Communication des résultats

- **Description :** La communication des résultats constitue une étape essentielle du projet de stage. Elle a consisté à présenter, sous une forme claire et synthétique, les différentes analyses menées (indicateurs KPI, visualisations graphiques, modèles prédictifs, tableaux de bord). L'objectif principal était de rendre les informations accessibles et compréhensibles pour un public non technique (responsable de stage, encadrants).
- **Outils utilisés :**
 - PowerPoint : préparation et présentation d'un diaporama dynamique résumant les résultats clés.
 - Word : rédaction d'un rapport de synthèse intermédiaire et d'explications textuelles.
 - PDF : diffusion de la version finale du rapport et des supports visuels.
- **Difficultés rencontrées :**
 - Le vocabulaire technique (modèles, algorithmes, métriques) n'était pas toujours compréhensible pour un public non spécialiste.
 - Le volume important de résultats (graphiques, tableaux, indicateurs) risquait de surcharger la présentation.
- **Solutions apportées :**
 - Simplification du vocabulaire : emploi de termes accessibles et pédagogiques.
 - Sélection des résultats pertinents : choix des graphiques et indicateurs les plus significatifs.
 - Mise en forme visuelle : insertion de graphiques clairs, utilisation de couleurs, légendes et titres explicites.
 - Structuration du rapport : découpage en sections cohérentes (analyse descriptive, modélisation, évaluation, perspectives).
- **Valeur ajoutée :** Cette démarche de communication a permis de :
 - Faciliter la prise de décision par les responsables grâce à des résultats visuellement compréhensibles.
 - Montrer l'impact concret du travail réalisé en stage sur l'amélioration du pilotage hospitalier.
 - Développer des compétences transversales en vulgarisation scientifique et en présentation professionnelle.

Tableau récapitulatif des tâches réalisées

| Tâche | Description | Outils utilisés |
|-----------------------------|---|-----------------------------|
| Récupération des données | Extraction et structuration (Excel → base relationnelle avec 5 tables) | Python(pandas),Excel |
| Nettoyage et préparation | Suppression des doublons, traitement des valeurs manquantes,normalisation des formats | Python (pandas, numpy) |
| Visualisation graphique | Création de graphiques (histogrammes, camemberts, courbes) pour représenter les résultats | matplotlib, seaborn |
| Analyse descriptive(KPI) | Calcul d'indicateurs clés (patients, consultations, durée moyenne, âges) | Python (pandas, matplotlib) |
| Création de dashboards | Tableaux interactifs avec filtres et KPI dynamiques | Dash (Python), Power BI |
| Communication des résultats | Présentation orale + rapport écrit avec images et analyses | PDF |

CONCLUSION GÉNÉRALE

Conclusion

Au terme de ce projet, plusieurs résultats majeurs peuvent être soulignés. Ils témoignent à la fois de la solidité de l'infrastructure mise en place et des pistes d'amélioration identifiées pour les phases ultérieures :

- **Base de données relationnelle** : La base de données conçue et implémentée est cohérente, normalisée et directement exploitable pour des besoins analytiques ou décisionnels. Son modèle conceptuel et logique (ERD) assure une structuration robuste des informations hospitalières, couvrant les principales entités (patients, médecins, diagnostics, traitements). Cette architecture garantit l'intégrité des données, limite la redondance et facilite les futures extensions (ajout de nouvelles tables ou de relations). Elle constitue ainsi une fondation fiable pour tout projet d'analyse ou de pilotage stratégique.
- **Analyse descriptive par indicateurs clés (KPI)** : L'analyse exploratoire réalisée a permis de vérifier la représentativité et la qualité du dataset. Celui-ci est équilibré en volume, comprenant un nombre significatif de patients et de consultations, ce qui confère de la crédibilité aux résultats obtenus. Parmi les indicateurs calculés, l'âge moyen des patients (environ 44 ans) s'avère particulièrement pertinent, car il reflète un profil médian de population hospitalière. Cette donnée constitue un point de référence essentiel pour la planification des ressources médicales et la mise en place de politiques de prévention ciblées.
- **Visualisations statistiques** : Les représentations graphiques ont apporté une vision claire et synthétique des dynamiques hospitalières. Elles ont mis en évidence la répartition sociodémographique des patients (âge, sexe) ainsi que la fréquence des diagnostics les plus

courants. Ces visualisations renforcent la compréhension des profils patients et des pathologies dominantes, permettant ainsi d’orienter les stratégies de prise en charge médicale et de gestion des flux de patients. Elles constituent également un support efficace de communication pour les décideurs hospitaliers.

- **Modélisation prédictive** : L’expérimentation menée à travers un premier modèle de classification a révélé des résultats limités. En effet, le modèle reposait uniquement sur deux variables explicatives (âge et sexe), ce qui restreint considérablement sa puissance prédictive. Les performances obtenues (mesurées par la précision, le rappel et le F1-score) ont confirmé que des variables additionnelles sont nécessaires pour améliorer la qualité des prédictions. Cette étape a néanmoins joué un rôle fondamental en posant les bases de la démarche prédictive et en ouvrant des perspectives d’amélioration (enrichissement du dataset, adoption de modèles plus avancés).

En somme, le projet a permis de démontrer la faisabilité d’un système analytique appliqué aux données hospitalières, tout en mettant en lumière les axes d’amélioration prioritaires pour renforcer son utilité pratique et son impact décisionnel.

Recommandations

Pour améliorer les performances et l'impact de ce projet, plusieurs recommandations peuvent être formulées. Elles s'articulent autour de cinq axes principaux :

1. **Enrichissement du dataset** : Le dataset actuel se limite essentiellement à des informations sociodémographiques (âge, sexe), ce qui restreint fortement la capacité prédictive des modèles. Il est donc recommandé d'intégrer de nouvelles variables explicatives, telles que : les antécédents médicaux des patients, la présence de facteurs de risque (tabagisme, diabète, hypertension), la tension artérielle, les symptômes déclarés lors de l'admission, ainsi que la saisonnalité des consultations (certains diagnostics étant liés à des périodes précises de l'année). De même, la durée et la nature des traitements prescrits constitueraient des variables particulièrement utiles pour affiner la compréhension des trajectoires de soins. Cet enrichissement permettra de mieux capter la complexité du domaine médical et d'améliorer la pertinence des analyses.
2. **Amélioration des modèles prédictifs** : L'utilisation d'un modèle simple de classification a mis en évidence des limites significatives, notamment une capacité de généralisation insuffisante. Afin d'accroître la robustesse et la fiabilité des prédictions, il est conseillé d'expérimenter avec des algorithmes plus avancés, tels que les forêts aléatoires (Random Forest), les méthodes de gradient boosting (XGBoost, LightGBM) ou encore les réseaux de neurones artificiels. Ces modèles, plus performants, permettent de capturer des relations complexes entre variables. Un travail approfondi de tuning des hyperparamètres (optimisation du nombre d'arbres, profondeur maximale, taux d'apprentissage, etc.) ainsi qu'une validation croisée rigoureuse (k-fold cross validation) devront être mis en place afin d'assurer la stabilité et la généralisabilité des résultats.
3. **Équilibrage des classes** : Dans de nombreux jeux de données médicaux, certaines pathologies sont sous-représentées, ce qui peut biaiser les modèles prédictifs en faveur des classes majoritaires. Pour pallier ce problème, il est recommandé d'adopter des techniques de rééchantillonnage adaptées : *oversampling* des classes minoritaires, *undersampling* des classes majoritaires, ou encore recours à des méthodes plus sophistiquées comme SMOTE (Synthetic Minority Oversampling Technique). Ces approches permettront de corriger le déséquilibre des classes et d'obtenir des modèles plus justes, capables de mieux identifier les cas rares mais cliniquement importants.
4. **Mise en place d'un Dashboard interactif** : Les résultats analytiques gagneraient à être diffusés sous une forme accessible, ergonomique et dynamique. La mise en œuvre d'un tableau de bord interactif constitue une étape clé pour valoriser les analyses. À cet effet, l'utilisation de frameworks tels que *Dash (Python)* ou d'outils comme *Power BI* est fortement recommandée. Un tel Dashboard permettrait d'explorer les données en temps réel, de filtrer les résultats selon divers critères (âge, sexe, diagnostic), d'afficher des cartes de répartition géographique et de générer automatiquement des rapports de suivi. Cette

interface servirait non seulement comme outil d'aide à la décision, mais également comme support de communication entre les analystes de données et les responsables hospitaliers.

5. **Valorisation et diffusion des résultats** : Enfin, il est essentiel que les analyses réalisées ne restent pas uniquement confinées aux aspects techniques. Les indicateurs clés (KPI), graphiques et résultats prédictifs doivent être intégrés dans des supports professionnels adaptés, tels que des rapports PDF détaillés ou des présentations PowerPoint synthétiques. Ces supports permettront aux décideurs hospitaliers et aux équipes médicales de disposer d'une information claire, opérationnelle et exploitable dans le cadre de leur planification stratégique. En assurant une diffusion efficace des résultats, le projet pourra réellement contribuer à l'amélioration de la gestion des ressources et de la qualité des soins.

BIBLIOGRAPHIE / WEBOGRAPHIE

Webographie

- ❖ Data Science Specialization - Johns Hopkins University (Coursera)
- ❖ Documentation officielle Scikit-learn (machine learning en Python)
- ❖ Documentation Pandas (manipulation et analyse de données en Python)
- ❖ Machine Learning en Python (playlist YouTube - Sentdex)
- ❖ Kaggle : plateforme de compétitions et datasets Data Science

Bibliographie

- ❖ Trevor Hastie, Robert Tibshirani, Jerome Friedman, *The Elements of Statistical Learning*, Springer, 2009.
- ❖ Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, *An Introduction to Statistical Learning*, Springer, 2013.
- ❖ Jake VanderPlas, *Python Data Science Handbook*, O'Reilly Media, 2016.
- ❖ Ian Goodfellow, Yoshua Bengio, Aaron Courville, *Deep Learning*, MIT Press, 2016.