



# PYTHON FOR DATA ANALYSIS

PROJECT : DIABETES | 30-US HOSPITALS FOR YEARS | 1999-2008

# PLAN

## I/ Introduction

## II/ Analyse du dataset

- A/ Data pre-processing
- B/ Data visualisation

## III/ Le problème choisis

- A/ Ins and outs of the problem
- B/ Comment le problème s'inscrit dans le contexte de l'étude?
- C/ Nos variables
- D/ Hypothèses
- E/ Modélisation

## IV/ Conclusion

# I/ INTRODUCTION

**Notre sujet porte le dataset diabetic\_data.csv  
(Diabetes 130 US hospitals for years 1999–2008)**

- **Diabète :** Décrit comme une maladie chronique caractérisée par des taux élevés de glucose dans le sang.
- **Objectif de l'analyse :** Examiner les tendances historiques des soins liés au diabète dans les hôpitaux américains sur les années 1999-2008 à l'aide d'une vaste base de données clinique (101 766 dossiers de patients)
- Comprendre si un patient est susceptible d'être réadmis à l'hôpital est crucial. Cette information est essentielle pour ajuster le plan de traitement afin de prévenir une récurrence de l'hospitalisation.


## II/ ANALYSE DU DATASET

### A/ DATA PRE-PROCESSING

Nettoyage des données : "Nous avons traité les valeurs manquantes et les anomalies, en particulier dans des variables telles que 'poids'. »



Encodage et Normalisation : "Les variables catégorielles ont été encodées pour l'analyse, et les caractéristiques numériques ont été normalisées pour assurer la cohérence."



Techniques d'imputation : "Nous avons appliqué des méthodes d'imputation avancées pour traiter efficacement les données manquantes."

## I / ANALYSE DU DATASET

### B / DATA VISUALISATION



Outils de visualisation : "Nous avons utilisé Matplotlib et Seaborn pour une visualisation approfondie des données."



Visualisations clés : "Les représentations graphiques ont révélé des tendances entre des variables telles que 'âge', 'sexe', 'durée du séjour à l'hôpital' et les taux de réadmission."

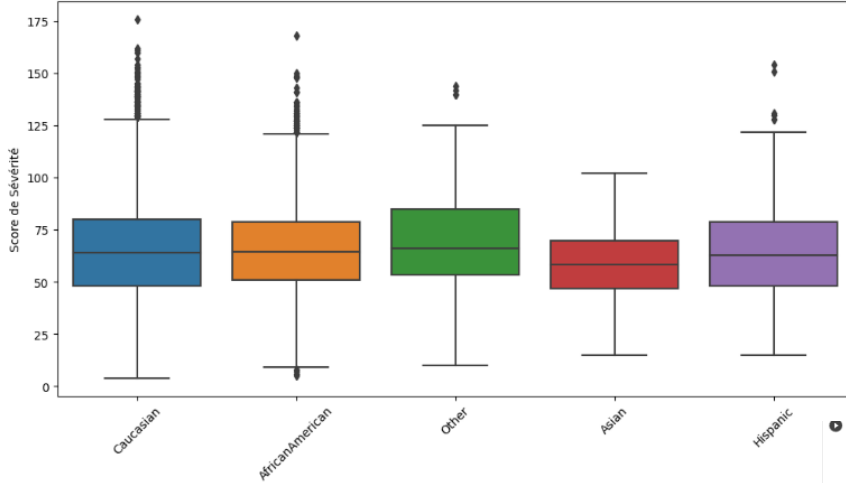


Interprétation : "Ces visualisations offrent des perspectives intuitives sur les données, aidant à identifier les facteurs critiques influençant les réadmissions."

# II/ ANALYSE DU DATASET

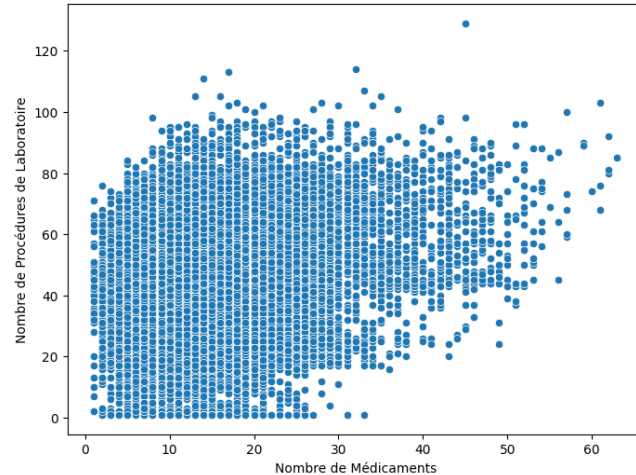
## B/ DATA VISUALISATION

Ethnie / Sévérité de la Maladie

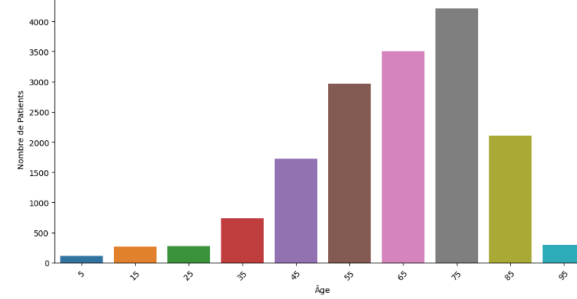


Ethnie

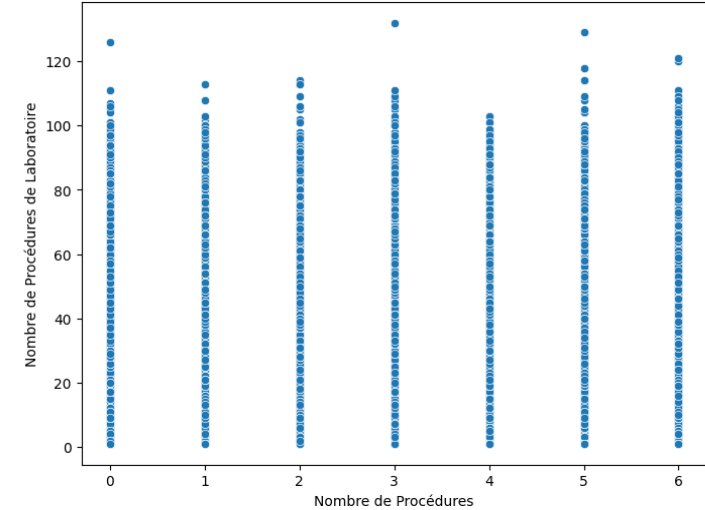
Nombre de Médicaments vs Nombre de Procédures de Laboratoire



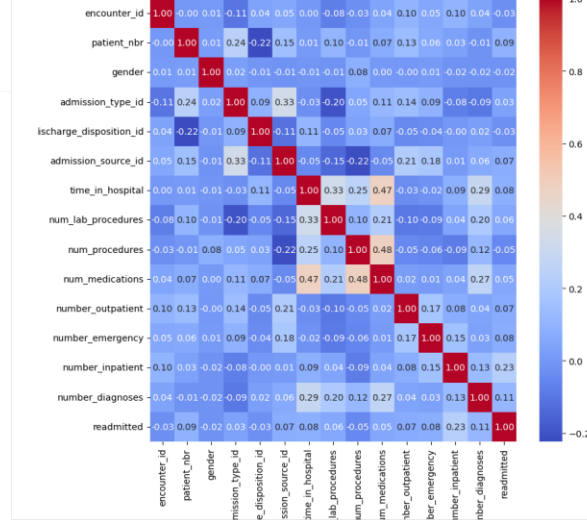
Nombre de Patients / Âge



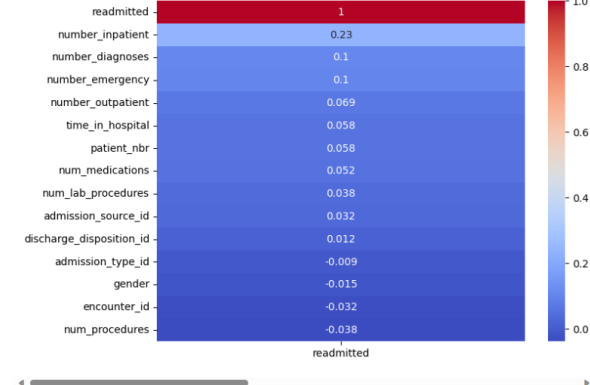
Nombre de Procédures vs Nombre de Procédures de Laboratoire



Heatmap des Corrélations



Corrélations avec la variable Readmitted



### III/ LE PROBLÈME CHOISI

#### A/ INS AND OUTS OF THE PROBLEM



Nous souhaitons prédire si un patient sera réadmis. La variable "réadmis" est définie comme 1 si le patient a été réadmis (que ce soit '<30 jours' ou '>30 jours') et 0 autrement. Nous étudions les corrélations entre "réadmis" et différentes variables précédemment décrites, telles que :

- Est-ce que l'ethnie a un impact sur la fréquence de réadmission ?
- Est-ce que le résultat d'un test influence la réadmission ?
- Les médicaments prescrits indiquent-ils une possible réadmission ?

### III/ LE PROBLÈME CHOISI

#### B/ COMMENT LE PROBLEME S'INSCRIT DANS LE CONTEXTE DE L'ETUDE?

Comprendre ces enjeux peut permettre d'optimiser les différents processus.  
Voici comment le problème s'intègre dans le contexte de l'étude :

##### **Pertinence Médicale :**

- Le diabète est une préoccupation majeure de santé, nécessitant une gestion continue. La réadmission rapide est un défi crucial.

##### **Coûts Importants :**

- Réduire les réadmissions à 30 jours peut significativement baisser les coûts de santé.

##### **Priorité Politique et Qualité des Soins :**

- Réduire les réadmissions est une priorité politique, alignée sur les objectifs d'amélioration de la qualité des soins.

##### **Conformité aux Objectifs de l'Étude :**

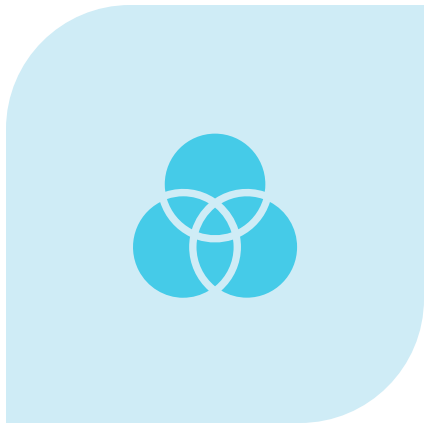
- La prédiction des réadmissions diabétiques en 30 jours correspond parfaitement aux objectifs définis.

##### **Contraintes et Préoccupations :**

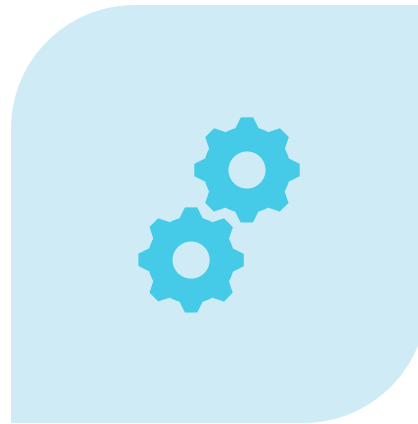
- La contrainte d'interprétabilité et la gestion des coûts des erreurs de classification sont spécifiques au contexte médical.



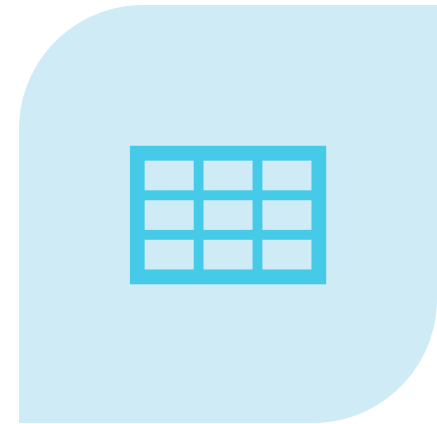
### III/ LE PROBLÈME CHOISI C/ NOS VARIABLES



SÉLECTION DE L'ALGORITHME : "NOUS AVONS EXPLORÉ DIVERS ALGORITHMES, NOTAMMENT LE KNN, LES ARBRES DE DÉCISION ET RANDOM FOREST, POUR PRÉDIRE LES RÉADMISSIONS."



AJUSTEMENT DES HYPERPARAMÈTRES : "NOUS AVONS UTILISÉ LA « GRID SEARCH » POUR OPTIMISER LES PARAMÈTRES DU MODÈLE."



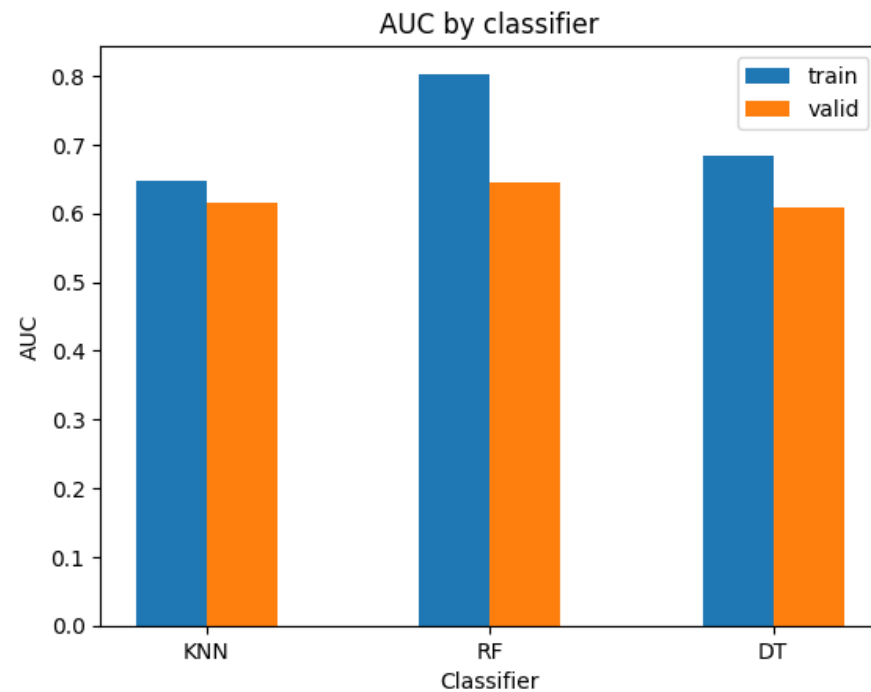
COMPARAISON DES MODÈLES : "LES MODÈLES ONT ÉTÉ COMPARÉS EN FONCTION DES MÉTRIQUES DE PERFORMANCE, VISUALISÉES À TRAVERS UN GRAPHIQUE POUR UNE INTERPRÉTATION AISÉE."

### III/ LE PROBLÈME CHOISI

#### C/ NOS VARIABLES

- race (Caucasian, Others)
- gender (Male, Female)
- age (0-80)
- time\_in\_hospital (integer)
- num\_lab\_procedures (integer)
- num\_procedures (integer)
- num\_medications (integer)
- (integer) number\_diagnoses
- (test, >200, <300, Norm) max\_glu\_serum
- (test, >200, <300, Norm) A1Cresult
- (yes or no) change
- (yes or no if diabete medicament taken)diabetesMed
- (<30, >30, No) readmitted
- (string) admission\_type
- (string) discharge\_disposition
- (string) admission\_source

### III/ LE PROBLÈME CHOISI/ MODÉLISATION



- RF : sur-ajustement, AUC le plus élevé
- KNN : bonne généralisation du modèle, AUC inférieur
- DT : sur-ajustement plus marqué que le KNN pour un AUC similaire

# CONCLUSION

- Principales conclusions : Notre analyse a identifié des prédicteurs significatifs pour les réadmissions tels que l'âge du patient, la durée du séjour à l'hôpital et certains indicateurs médicaux.
- Implications pour les soins de santé : Ces insights peuvent aider les prestataires de soins de santé à élaborer des interventions ciblées pour réduire les taux de réadmission.



# MERCI

DE NOUS AVOIR ÉCOUTÉ

VOUS AVEZ DES QUESTIONS?