



**Sujet: Prédire les jours  
d'hospitalisation des patients**

EL FALLEH Wissam  
EL GHOULI Moussa  
YAAKOUBI Salma

## PLAN

**1.INTRODUCTION GENERALE**

**2.SUJET**

**3.PARTIE BI**

**4.PARTIE ML**

**5.CONCLUSION**



# Table des matières

<b>1.1. Introduction .....</b>	<b>4</b>
2.1. Presentation .....	5
3.1 Partie BI .....	6
4. Partie ML .....	13
4.1. Architecture de la solution: .....	13
4.2 Dataset used: .....	13
4.3. Les algorithms utilisés : .....	14
4.4.Les protocoles expérimentaux : .....	16
4.5.Résultats et discussion: .....	16
5.Conclusion .....	18

Figure 1:création d'une base sur SQLServer7 .....	8
Figure 2:charger la base.....	8
Figure 3:Creation du table de Faites .....	8
Figure 4:Remplir table de Faites .....	9
Figure 5:table de Faites charge .....	9
Figure 6:Cube.....	10
Figure 7:nombre de visiteur par Age.....	11
Figure 8:nombre de visiteur par département & région .....	11
Figure 9:nombre de visiteur par Age & département.....	12
Figure 10:Algorithmes Random Forest.....	14
Figure 11:Algorithmes XGBOOST .....	15
Figure 12:Algorithmes Régression linear multiple .....	15



# 1. INTRODUCTION GENERALE

## 1.1. Introduction

Le terme Business Intelligence (BI), ou informatique décisionnelle, désigne les applications, les infrastructures, les outils et les pratiques offrant l'accès à l'information, et permettant d'analyser l'information pour améliorer et optimiser les décisions et les performances d'une entreprise.



En d'autres termes, la Business Intelligence est le processus d'analyse de données dirigé par la technologie dans le but de déceler des informations utilisables pour aider les dirigeants d'entreprises et autres utilisateurs finaux à prendre des décisions plus informées.

Ainsi, la BI regroupe une large variété d'outils, d'applications et de méthodologies permettant de collecter

des données en provenance de systèmes internes et de sources externes, de les préparer pour l'analyse, de les développer et de lancer des requêtes au sein de ces ensembles de données.

Ces outils permettent ensuite de créer des rapports, des tableaux de bord et des visualisations de données pour rendre les résultats des analyses disponibles pour les preneurs de décisions.

## 2. SUJET

### 2.1. Presentation

L'analyse prédictive est un outil de plus en plus important dans le domaine de la santé, car les méthodes modernes d'apprentissage automatique (ML) peuvent utiliser de grandes quantités de données disponibles pour prédire les résultats individuels des patients.

Par exemple, les prédictions ML peuvent aider les prestataires de soins de santé à déterminer les probabilités de maladie, à faciliter le diagnostic, à recommander un traitement et à prédire le bien-être futur.

Pour ce projet, j'ai choisi de me concentrer sur une métrique plus logistique des soins de santé, la durée d'hospitalisation.



## 3. PARTIE BI

### 3.1 Partie BI

Comme point de départ pour la recherche de données, mon intuition était que l'ensemble de données devrait idéalement inclure des caractéristiques telles que la catégorie du patient (par exemple, maladie cardiaque, accouchement, etc.) L'âge, le sexe.

Après avoir cherché une base de données médicale utile, j'ai fini par choisir la base de données (Prédire les jours d'hospitalisation des patients) de la part de KAGGEL.

En raison de la grande quantité d'informations qu'elle contenait.



La première étape de notre projet c'est de créer la base de données.  
En fait on a créé 2 bases de données SA-Hôpital et DW-Hôpital.

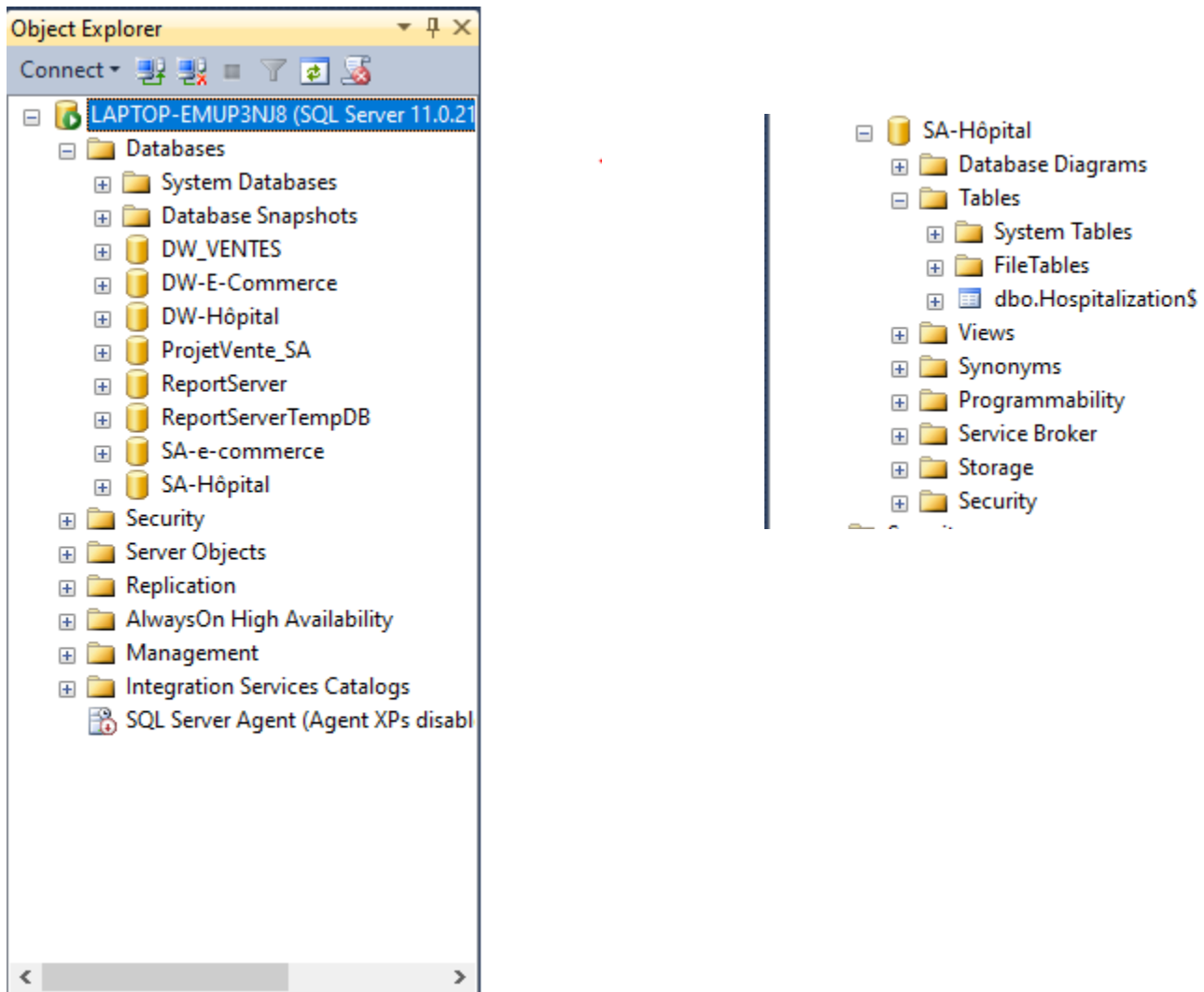


Figure 1: création d'une base sur SQL Server

En suite on a charger la base:

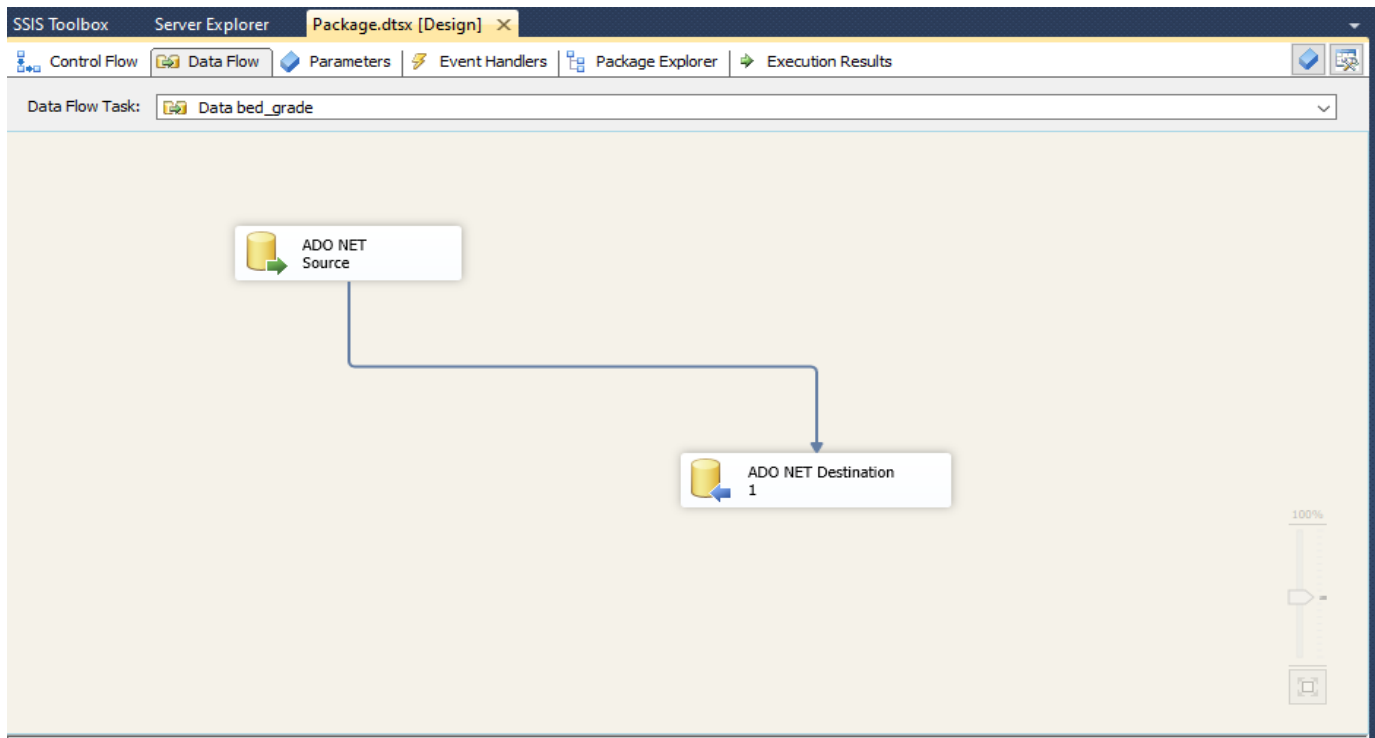


Figure 2:charger la base

Par la suite on a cree la table de faites:

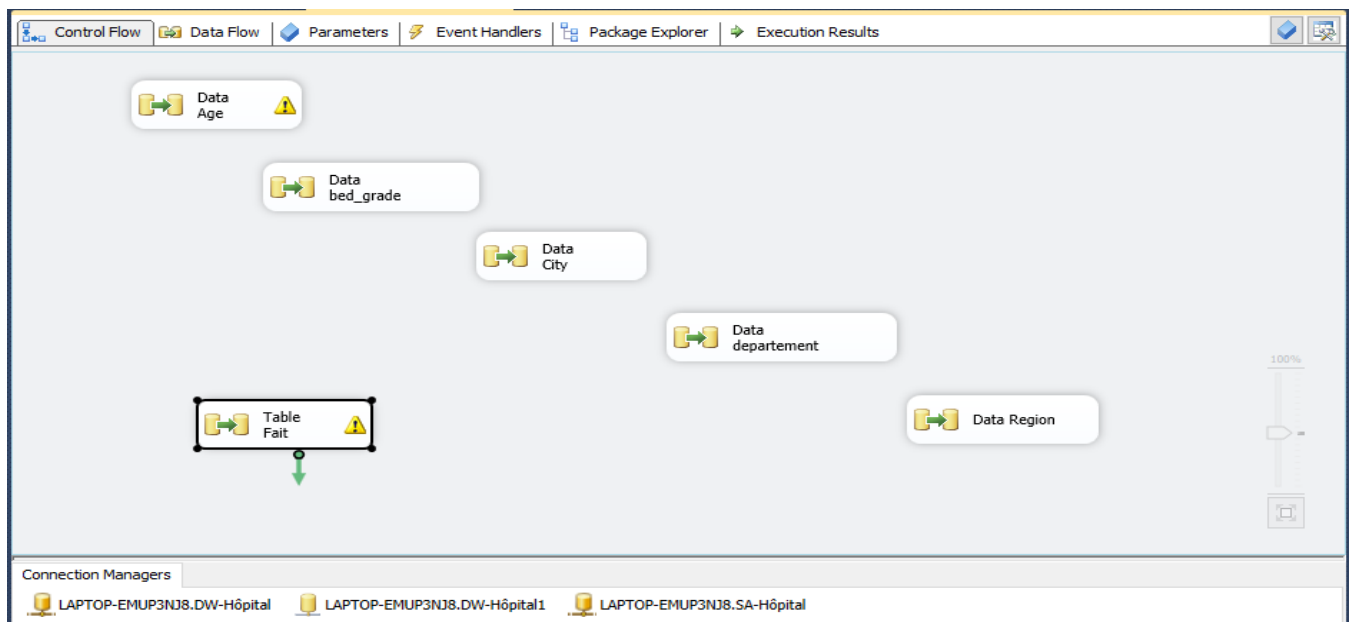


Figure 3:Création du table de Faites



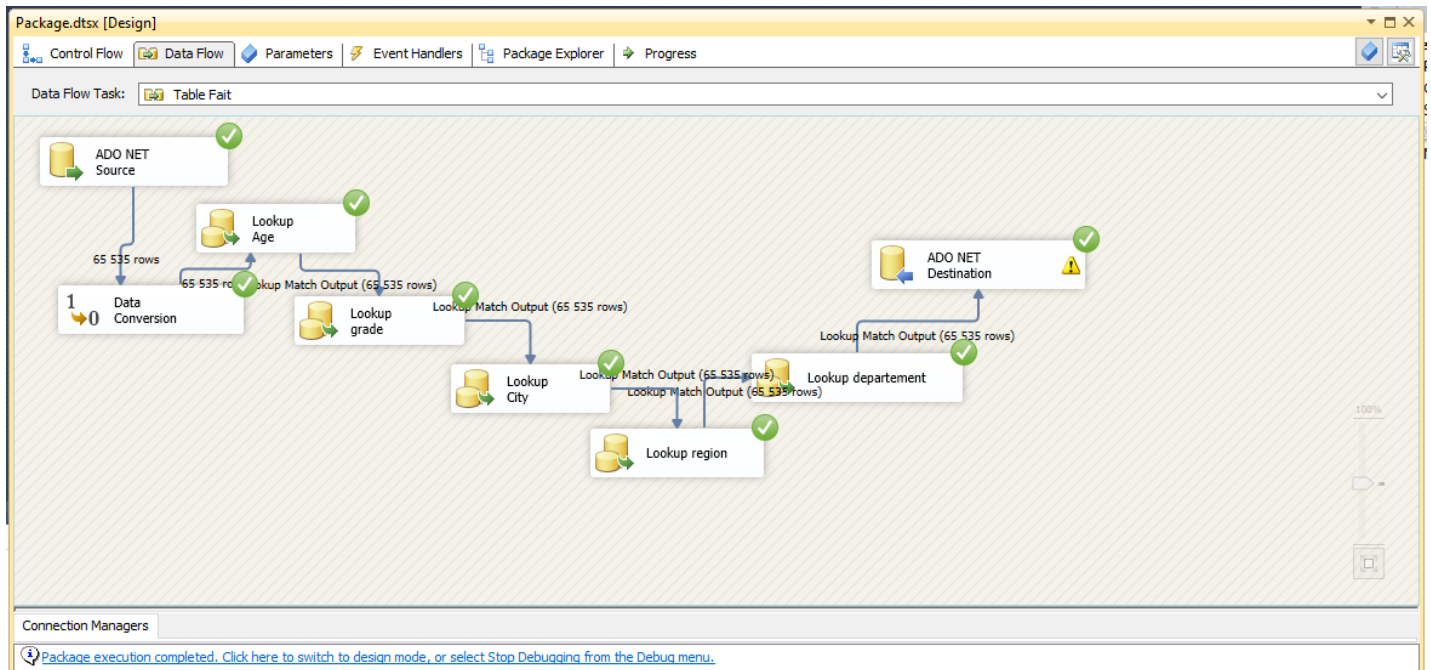


Figure 4: Remplir table de Faits

SQLQuery1.sql - LAPTOP-EMUP3N8.master (LAPTOP-EMUP3N8\HP (51)) - Microsoft SQL Server Management Studio

File Edit View Query Project Debug Tools Window Help

master Execute Debug

Object Explorer

Connect

LAPTOP-EMUP3N8 (SQL Server 11.0)

Databases

System Databases

Database Snapshots

DW-VENTES

DW-E-Commerce

DW-Hôpital

Database Diagrams

Tables

System Tables

FileTables

dbo.Dim\_Age

dbo.Dim\_Bed\_grade

dbo.Dim\_city

dbo.Dim\_region

dbo.Dim-Departement

dbo.fait\_hospital

Views

Synonyms

Programmability

Service Broker

Storage

Security

ProjetVente\_SA

ReportServer

SQLQuery1.sql - LAPTOP-EMUP3N8\HP (51)

```

/***** Script for SelectTopNRows command from SSMS *****/
SELECT TOP 1000 [Id_Age_FK]
, [Id_dept_FK]
, [Id_region_FK]
, [Id_city_FK]
, [Id_Bed_FK]
, [nbr_dispo_champ]
, [nbr_visitor_pas]
, [type_hobital]
FROM [DW-Hôpital].[dbo].[fait_hospital]

```

100 %

Results Messages

	Id_Age_FK	Id_dept_FK	Id_region_FK	Id_city_FK	Id_Bed_FK	nbr_dispo_champ	nbr_visitor_pas	type_hobital
1	10	2	2	13	15	4	3	c
2	10	2	1	12	15	3	5	a
3	10	2	3	16	11	5	3	d
4	10	2	2	13	15	6	3	c
5	10	2	3	14	14	5	4	d
6	10	2	1	17	15	3	3	a
7	10	2	1	17	15	2	3	a
8	9	2	3	22	14	3	3	b
9	9	2	3	18	11	3	8	f

Query executed successfully. LAPTOP-EMUP3N8 (11.0 RTM) LAPTOP-EMUP3N8\HP (51) | master | 00:00:00 | 1000 rows

Output

Show output from:

Properties

Current connection parameters

Aggregate Status

Connection failure

Elapsed time 00:00:00.3111839

Finish time 12/04/2021 09:48:54

Name LAPTOP-EMUP3N8

Rows returned 1000

Start time 12/04/2021 09:48:54

State Open

Connection

Connection name LAPTOP-EMUP3N8 (LA)

Connection Details

Connection elapsed 00:00:00.3111839

Connection finish 12/04/2021 09:48:54

Connection rows 1000

Connection start t 12/04/2021 09:48:54

Connection state Open

Display name LAPTOP-EMUP3N8

Login name LAPTOP-EMUP3N8\HP

Server name LAPTOP-EMUP3N8

Server version 11.0.2100

Session Tracing ID

SPID 51

Name

The name of the connection.

Ready

Ln 1 Col 1 Ch 1 INS

Figure 5: table de Faits charge

Après la creation des tables et les dimension et la charge de la base on a fait le cube

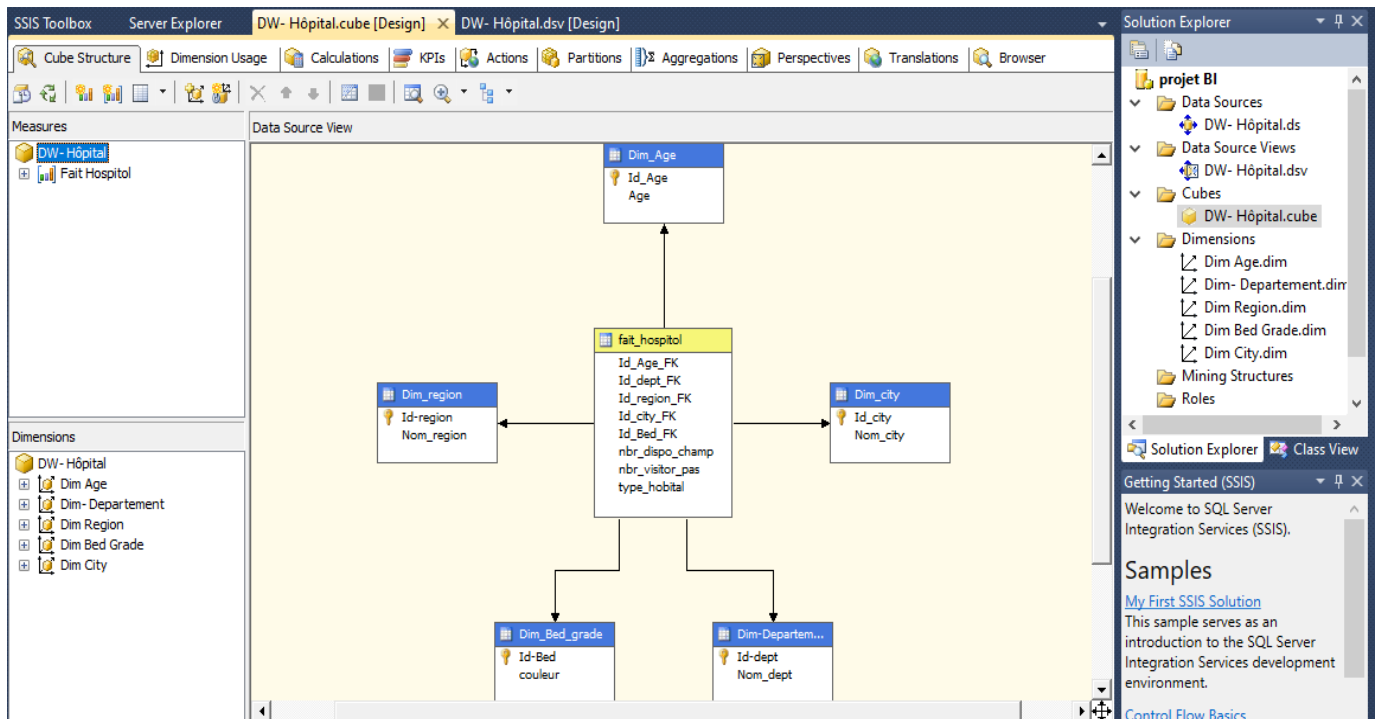
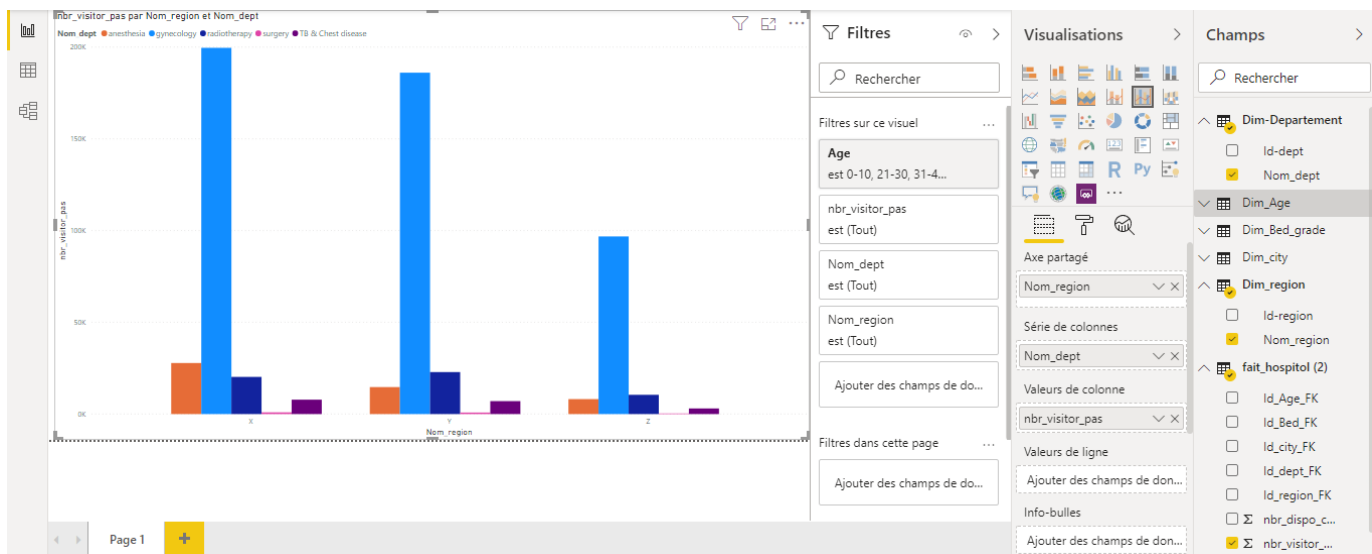
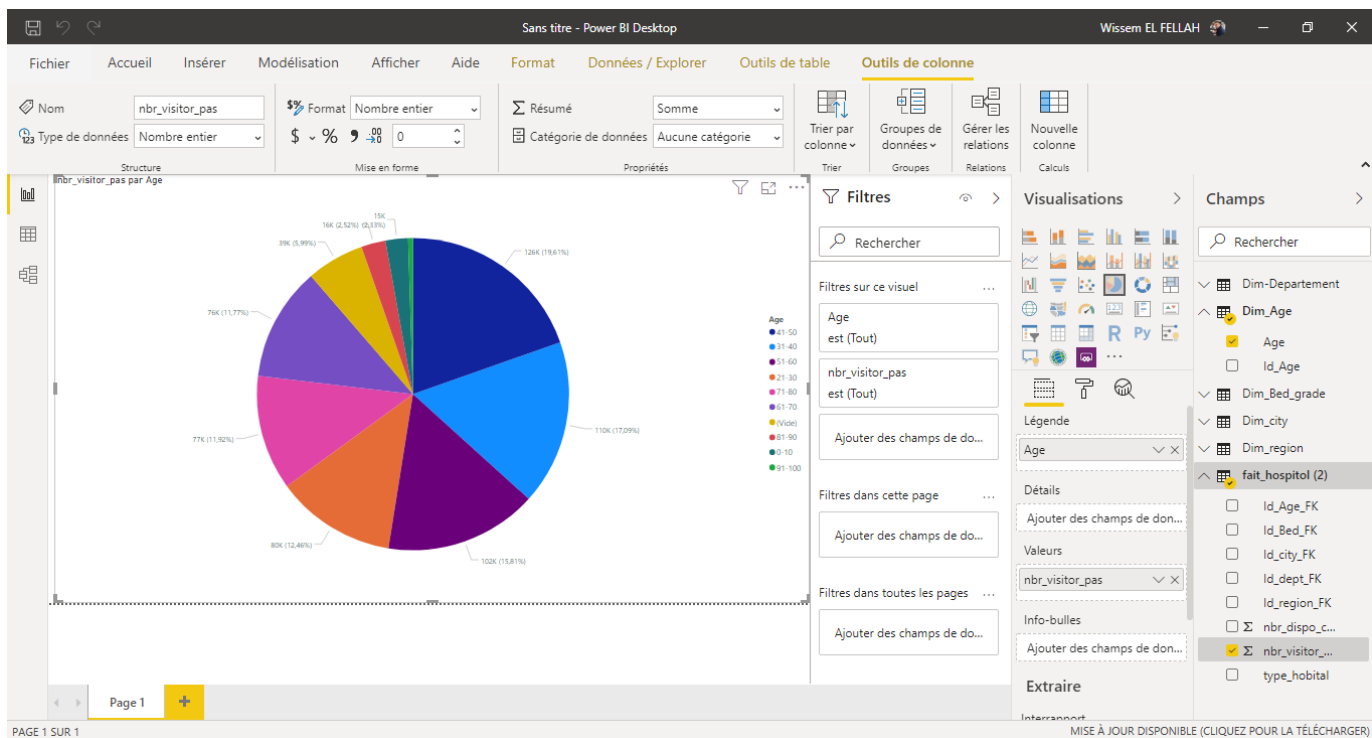


Figure 6: Cube

Et finalement on a construit notre power BI



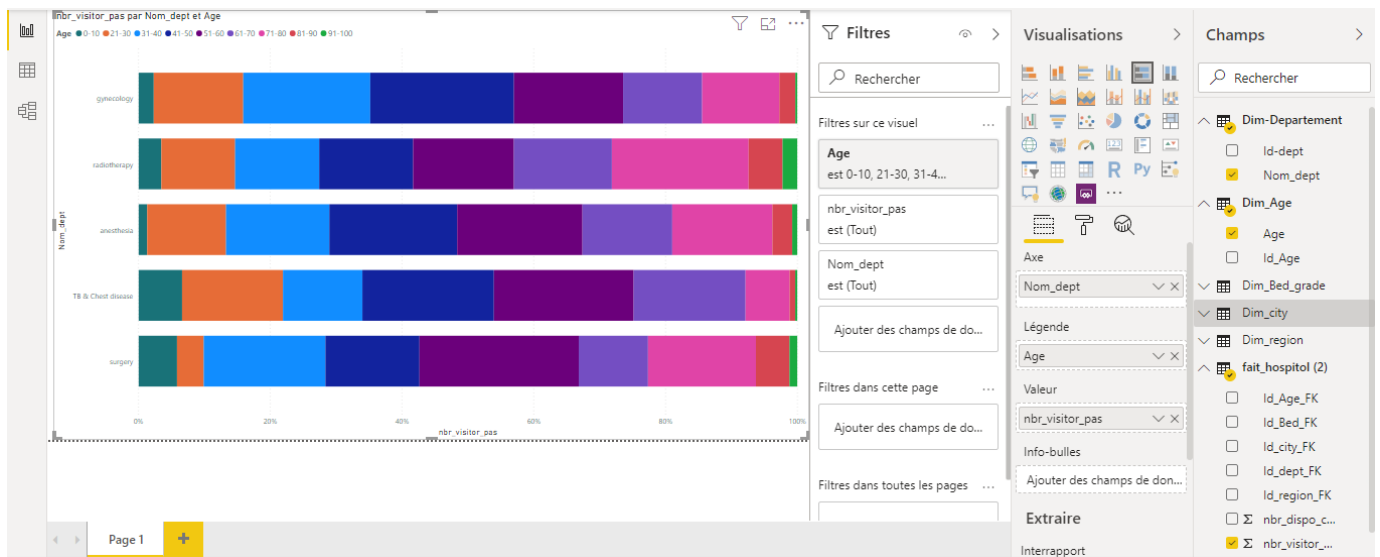


Figure 9:nombre de visiteur par Age & département

## 4. PARTIE ML

### 4. Partie ML

#### 4.1. Architecture de la solution:

Notre système est composé de deux phases principales:

La première étape concerne la partie BI, elle est composée de trois sous-étapes qui sont:

- 1) la collecte de donnée,
- 2) la transformation des données,
- 3) le stockage de données.

La deuxième Cette étape est composée de deux sous étapes, à savoir le prétraitement et le nettoyage de données, et la classification qui est basée sur les algorithmes de l'apprentissage automatique (Machine Learning: ML).

#### 4.2 Dataset used:

Un hôpital dispose de plusieurs types de données. Ces données concernent la ville, le patient, le département, etc. – il y a tellement de chiffres et de statistiques que nous pourrions collecter, c'est époustouflant! Nous réduirons notre champ d'application à certains domaines spécifiques pour ce projet:

**Case\_id:** Identifiant du cas (ce n'est pas crucial pour l'analyse, mais au niveau de la base de données, il sera utile d'avoir cette information)

**Hospital\_code:** indique Code de l'hôpital

**City\_Code\_Hospital:** indique le code d'hôpital par ville

**Hospital\_region\_code :** indique le code d'hôpital par région

**Available Extra Rooms in Hospital:** indique la disponibilité de chambre vide dans l'hôpital

**Department:** c'est le non de département dans l'hôpital

**Ward\_Type:** indique Types de services hospitaliers

**Ward\_Facility\_Code:** Code de installation

**Bed Grade:** Catégorie de lit

**Patient id:** Identifiant du patient



**City\_Code\_Patient:** Code de ville de Patient

**Type of Admission:** Type d'admission (Lors d'une hospitalisation, le régime obligatoire de la Sécurité sociale laisse à la charge du patient des frais pouvant évoluer selon le type d'admission).

**Severity of Illness:** Gravité de la maladie

**Visitors with Patient:** Visiteurs avec patient

**Age:** indique l'âge de patient

**Admission\_Deposit:** Caution d'admission (Pendant l'admission et tout au long de votre séjour

**Stay:** temps de séjour de patient dans l'hôpital.

### 4.3. Les algorithmes utilisés:

Dans cette étape, nous étudions les algorithmes d'apprentissage automatique qu'on a utilisé.

On a commencé par l'algorithme Random Forest :

```
"4)Random forest

model=RandomForestRegressor(random_state=0, n_estimators=100)
model=LinearRegression()

import time
debut=time.time()
model.fit(Data, y)
fin=time.time()-debut
ypred=model.predict(Data)

from sklearn.metrics import mean_squared_error
mse=mean_squared_error(y,ypred)
import math
print('RMSE', math.sqrt(mse))

from sklearn.metrics import explained_variance_score
EV=explained_variance_score(y,ypred)
print("Explained variance : %f" % (EV))
```

Figure 10:Algorithmes Random Forest

En suit on a utilise XGBOOST:

```
"6)XGBoost
import xgboost as xgb
model=xgb.XGBRegressor(objective='reg:linear', learning_rate=0.3, n_estimators=

import time
debut=time.time()
model.fit(Data, y)
fin=time.time()-debut
pred=model.predict(Data)

from sklearn.metrics import mean_squared_error
mse=mean_squared_error(y,pred)
import math
print('RMSE', math.sqrt(mse))

from sklearn.metrics import explained_variance_score
EV=explained_variance_score(y,pred)
print("Explained variance : %f" % (EV))
```

Figure 11:Algorithmes XGBOOST

Et on a terminer avec la Regression linear multiple:

```
"7)Regression linear multiple
from sklearn.linear_model import LinearRegression
model = LinearRegression()
model.fit(Data,y)

a=model.coef_
print(a)

b=model.intercept_
print(b)
X=np.array(Data)

pred=model.predict(Data)

from sklearn.metrics import mean_squared_error
mse=mean_squared_error(y,pred)
import math
print('RMSE', math.sqrt(mse))

import pandas as pd
Y1=pd.DataFrame(y)
Y1.describe()

from sklearn.metrics import explained_variance_score
EV=explained_variance_score(y,pred)
print("Explained variance : %f" % (EV))

from sklearn.metrics import r2_score
r=r2_score(y,pred)
print(r)
```

Figure 12:Algorithmes Régression linear multiple

## 4.4. Les protocoles expérimentaux:

Pour notre première algorithmes on a utilisé RMSE et EV comme des protocoles expérimentaux car avec l'algorithme de Random Forest il est primordial d'utiliser RMSE et surtout EV

Et pour notre deuxième algorithme XGBOOST, on a utilisé aussi RMSE et EV

Et finalement Regression linear multiple aussi on a utilisé RMSE, et EV

## 4.5. Résultats et discussion:

	Reg-Mult	Random-F	XGBOOST
RMSE	1.66932	1.45193	1.39624
EV	0.23545	0.38833	0.43435
CC	0.0984	0.1036	0.2364

En terme de RMSE et EV, nous clairement remarquons que XGBOOST donne meilleur résultat que Reg-Mult et Random-Forest.

➔ XGBOOST donne un meilleur résultat que Random-F car XGBOOST est un algorithme séquentiel (boosting), et Random-F est un algorithme indépendant (bagging)

➔ Random-F donne un meilleur résultat que Reg-Mult car R-F est un ensemble de prédicteurs alors que Reg-Mult est un seul prédicteur

En terme de complexité de calcul XGBOOST a une grande complexité de calcul car il est séquentiel.

En suite Random-F un peu faible car plusieurs predicteur, et finalement Reg-Multi donne Meilleur CC.

## 5. CONCLUSION

Ce rapport est destiné à vous inspirer sur la façon d'utiliser les données.

Il existe de plus en plus d'outils disponibles que vous pouvez utiliser pour tirer des enseignements des données publiques. J'espère que cette procédure pas à pas vous donnera quelques idées sur la façon de faire fonctionner les données pour vous.

Vous pouvez également utiliser cette analyse pour créer des modèles d'apprentissage automatique. Nous avons effectué le nettoyage et l'exploration des données - faites-le avancer et utilisez vos algorithmes préférés pour prédire les jours d'hospitalisation des patients. Les possibilités sont infinies.