Cycle d'Ingénieurs en
Génie Informatique
Faculté des Sciences et Techniques Errachidia

جامعة مولاي إسماعيل
UNIVERSITÉ MOULAY ISMAÏL

كلية العلوم والتقنيات
FACULTÉ DES SCIENCES ET TECHNIQUES

APPRENTISSAGE STATISTIQUE

Prédiction du risque d'infection du covid-19

Encadré par : Prof.Youssef Qarai

Présenté par :

- Boua Ali Sanogo
- Oumar Cissé
- Adamou Moussa Hassane

Sommaire

01 Introduction

02 Problématique

03 Collecte de données

04 Analyse exploratoire des données

05 Prétraitement des données

06 Modélisation

07 Conclusion

Introduction

La pandémie de COVID-19 a mis en évidence les défis liés à la gestion des ressources médicales, en particulier pour les patients à risque élevé de complications graves. Ce projet vise à développer un modèle d'apprentissage automatique capable de prédire, à partir des symptômes et des antécédents médicaux, si un patient est à haut risque. Cette solution permettra d'optimiser l'allocation des ressources et de renforcer la gestion des soins en période de crise.

Problématique

La gestion des ressources médicales a été l’un des principaux défis posés par la pandémie de COVID-19, notamment face à l’augmentation rapide des cas graves nécessitant des soins intensifs. Identifier de manière précoce les patients à haut risque, en s’appuyant sur leurs symptômes, leur état de santé et leurs antécédents médicaux, est crucial pour prioriser les soins et organiser les ressources. Cependant, cette tâche est complexe en raison des données souvent incomplètes, hétérogènes et parfois imprécises, particulièrement dans des contextes où la collecte rapide d'informations fiables est essentielle.

Collecte des données

Pour ce projet, l’ensemble de données utilisé provient du gouvernement mexicain, qui a fourni une ressource précieuse pour l’analyse et le développement du modèle. Cet ensemble de données contient des informations anonymisées sur 1 048 576 patients uniques. Chaque entrée inclut 21 caractéristiques distinctes, permettant d’explorer divers facteurs liés à la santé des patients, notamment leurs antécédents médicaux et leurs conditions préexistantes.

Analyse exploratoire des données

Statistique Descriptive

A

Résument les principales caractéristiques des données (moyenne, médiane, écart-type, minimum et maximum), offrant une vue d’ensemble sur les distributions et les anomalies.

Matrice de corrélation

B

Analyse les relations linéaires entre les variables pour identifier les corrélations positives ou négatives, utiles pour réduire la redondance dans les données.

Valeurs Manquantes

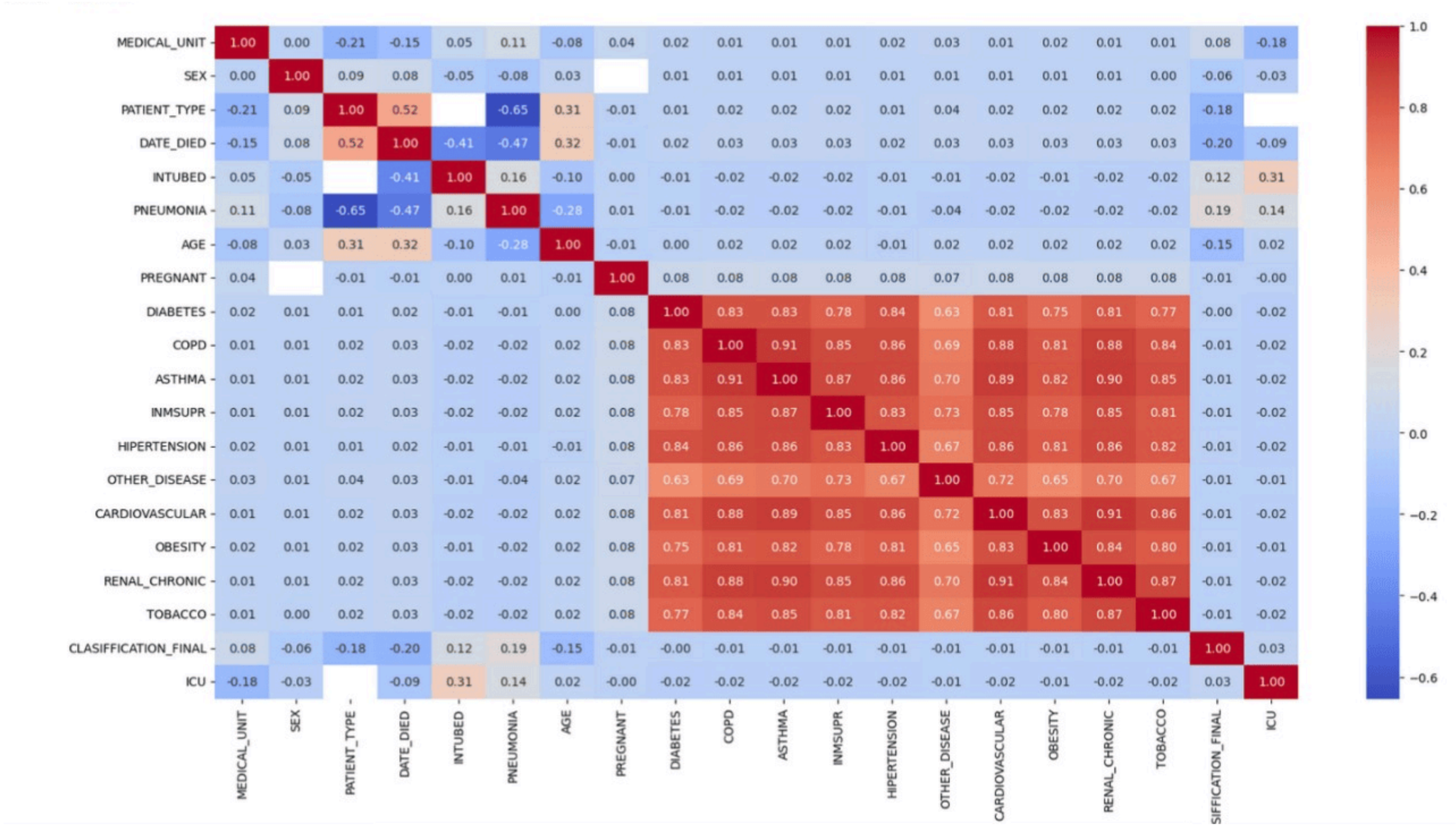
C

Évalue la quantité et la localisation des données absentes pour décider des stratégies de traitement comme l’imputation ou la suppression.

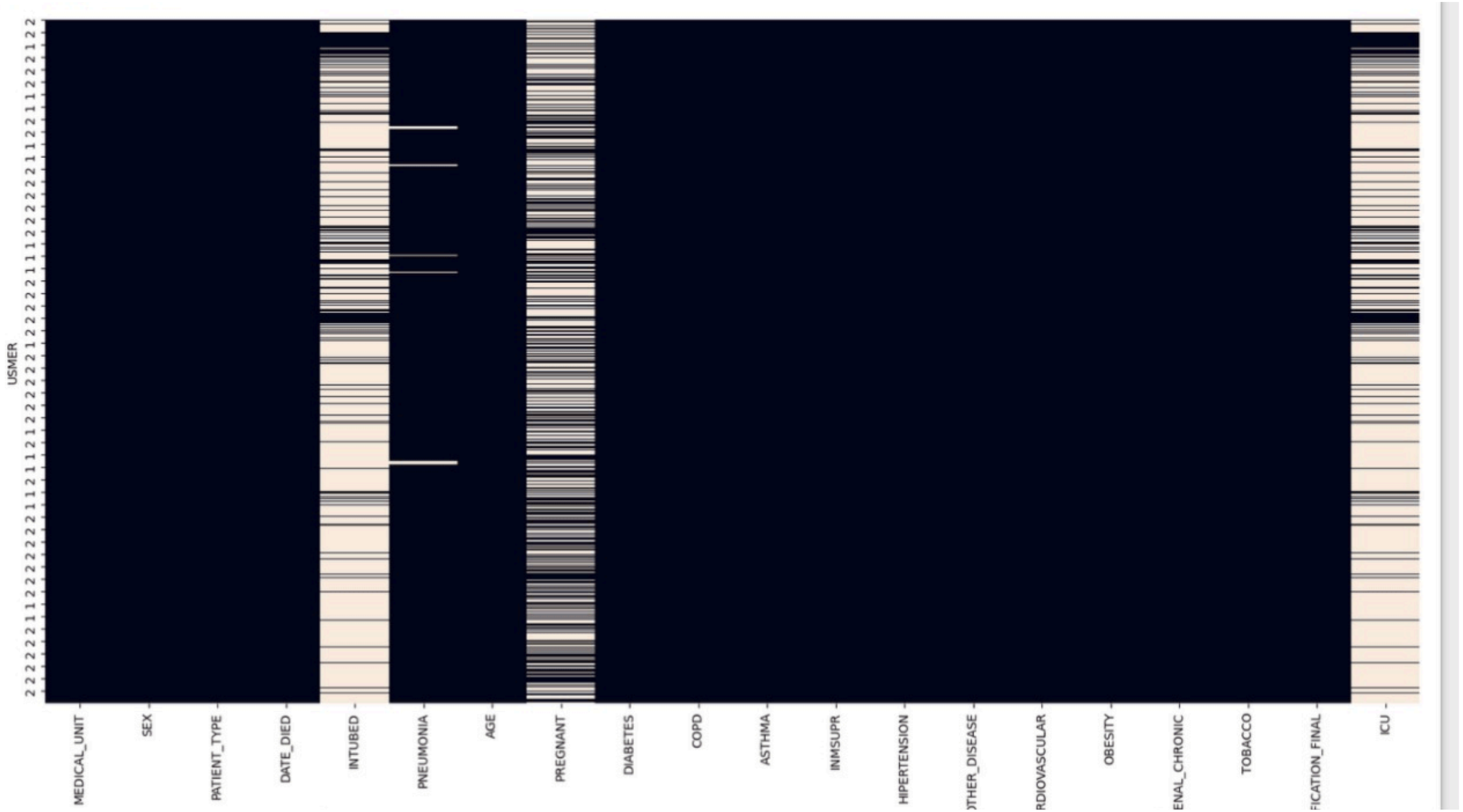
A- Statistiques descriptives

	MEDICAL_UNIT	SEX	PATIENT_TYPE	INTUBED	PNEUMONIA	AGE	PREGNANT	DIABETES	COPD	ASTHMA	INMSUPR	HIPERTENSION	OTHER_DISEASE	CARDIOVASCULAR	OBESITY	RENAL_CHRONIC	TOBACCO	CLASIFFICATION_FINAL	ICU
count	1048575.00	1048575.00	1048575.00	1048575.00	1048575.00	1048575.00	1048575.00	1048575.00	1048575.00	1048575.00	1048575.00	1048575.00	1048575.00	1048575.00	1048575.00	1048575.00	1048575.00	1048575.00	1048575.00
mean	8.98	1.50	1.19	79.52	3.35	41.79	49.77	2.19	2.26	2.24	2.30	2.13	2.44	2.26	2.13	2.26	2.21	5.31	79.55
std	3.72	0.50	0.39	36.87	11.91	16.91	47.51	5.42	5.13	5.11	5.46	5.24	6.65	5.19	5.18	5.14	5.32	1.88	36.82
min	1.00	1.00	1.00	1.00	1.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
25%	4.00	1.00	1.00	97.00	2.00	30.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00	3.00	97.00
50%	12.00	1.00	1.00	97.00	2.00	40.00	97.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00	6.00	97.00
75%	12.00	2.00	1.00	97.00	2.00	53.00	97.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00	7.00	97.00
max	13.00	2.00	2.00	99.00	99.00	121.00	98.00	98.00	98.00	98.00	98.00	98.00	98.00	98.00	98.00	98.00	98.00	7.00	99.00

B- Matrice de corrélation



C- Valeurs Manquantes



Prétraitement des données

1- Encodage initial des données :

- Les valeurs des colonnes ont été uniformisées pour faciliter l'analyse et la modélisation. Les valeurs 1 ont été utilisées pour représenter le "Faux" et les valeurs 0 pour le "Vrai". Ainsi, les valeurs 2 ont été transformées en 1 dans toutes les colonnes.
- Dans la colonne Pregnant, les valeurs 97, qui représentent les cas non applicables (par exemple, les hommes ne pouvant pas être enceintes), ont été transformées en 0.
- Dans la colonne DATE_DIED, les dates 9999-99-99, indiquant des patients vivants, ont été remplacées par 0, tandis que les autres valeurs ont été encodées en 1 pour indiquer le décès.

2- Nettoyage des valeurs manquantes :

Les valeurs 97 et 99, considérées comme manquantes (NaN), ont été supprimées de l'ensemble de données. Ce nettoyage a réduit le nombre de lignes de l'ensemble de données, passant de 1 048 576 à 388 878 tout en maintenant les 21 colonnes initiales.

Prétraitement des données

3-Création de la variable cible AT_RISK :

- Trois colonnes fortement corrélées, DATE_DIED, INTUBED et ICU, ont été fusionnées pour former une nouvelle colonne nommée AT_RISK. Cette variable cible regroupe les définitions du risque, car un patient décédé, intubé ou admis en soins intensifs est directement considéré comme à risque.
- Les valeurs ont été encodées en 0 pour les patients non à risque et 1 pour ceux à risque. Après la création de cette colonne, les colonnes originales utilisées dans la fusion (DATE_DIED, INTUBED, et ICU) ont été supprimées.

4- Suppression de colonnes redondantes :

La colonne Classification_final, qui indiquait déjà si un patient était à risque, a également été supprimée, car elle ne fournissait pas d'information supplémentaire après la création de la variable AT_RISK.

Prétraitement des données

5-Réduction finale des dimensions :

Après ces transformations, l'ensemble de données final se compose de 388 878 lignes et 17 colonnes, prêtes pour l'étape de modélisation. Ces étapes de prétraitement ont permis de simplifier et de rendre les données plus exploitables pour l'analyse et l'apprentissage automatique.

Modélisation

Dans la modélisation nous sommes passées par la création d’un ensemble de données en passant par:

1- Préparation des données :

Les données ont été divisées en trois ensembles distincts : entraînement, validation, et test. L'ensemble d'entraînement a servi à ajuster les paramètres du modèle, tandis que l'ensemble de validation a été utilisé pour évaluer la performance et ajuster les hyperparamètres. Enfin, l'ensemble de test a permis de mesurer l'efficacité du modèle sur des données non vues, afin d'obtenir une évaluation réaliste.

2-Méthode de division :

Nous avons utilisé une méthode de division aléatoire, en veillant à ce que la distribution des classes soit équilibrée dans chaque sous-ensemble. Une technique comme la validation croisée a été appliquée pour éviter les biais liés à une division unique des données. Cette division aléatoires a été effectuée par la méthode train_test_split de la bibliothèque scikit-learn.

Modélisation

Après la création de l'ensemble des données on est passé à l'utilisation des différents modèles:

3-Modèles utilisés :

Les modèles suivants ont été testés sur ces ensembles de données :

- K-Nearest Neighbors (KNN)
- Régression Logistique
- Support Vector Machine (SVM)
- Naïve Bayésienne
- Decision Tree
- Random Forest
- Q-Learning

K-Nearest Neighbors (KNN)

```
Meilleure valeur de k : 11
Précision moyenne sur la validation (cross-validation) : 0.8820
Matrice de confusion
 [[13624  2871]
 [  209  2740]]

Rapport de classification (test) :
              precision    recall  f1-score   support

           0       0.98      0.83      0.90     16495
           1       0.49      0.93      0.64      2949

    accuracy                           0.84     19444
   macro avg       0.74      0.88      0.77     19444
weighted avg       0.91      0.84      0.86     19444

Précision globale (test) : 0.8415963793458137
F1-score (test) : 0.6401869158878505
```

Régression Logistique

```
Meilleurs paramètres : {'logisticregression__C': 0.1, 'logisticregression__max_iter': 100,
er': 'liblinear'}  'logistiqueregressionSolver': 'liblinear'
Rapport de classification :
              precision    recall  f1-score   support

           0       0.98      0.85      0.91     16445
           1       0.53      0.91      0.67      2999

    accuracy                           0.86     19444
   macro avg       0.75      0.88      0.79     19444
weighted avg       0.91      0.86      0.87     19444

Précision globale : 0.8604196667352396
```

Support Vector Machine (SVM)

```
Meilleurs paramètres ( SVM ) :  {'svc__C': 10, 'svc__kernel': 'rbf'}

Matrice de confusion :
 [[27436  5542]
 [  303  5607]]
              precision    recall  f1-score   support

           0       0.99      0.83      0.90     32978
           1       0.50      0.95      0.66      5910

    accuracy                           0.85     38888
   macro avg       0.75      0.89      0.78     38888
weighted avg       0.92      0.85      0.87     38888


F1-Score :  0.8662919472105224
```

Naïve Bayésienne

```
Meilleurs paramètres (Naïve Bayes) :  {'naive_bayes__priors': [0.4, 0.6], 'selectkbest__k': 5}

F1-Score :  0.8747351282742788

Matrice de confusion :
 [[28065  4913]
 [  500  5410]]

Rapport de classification :
              precision    recall  f1-score   support

           0       0.98      0.85      0.91     32978
           1       0.52      0.92      0.67      5910

    accuracy                           0.86     38888
   macro avg       0.75      0.88      0.79     38888
weighted avg       0.91      0.86      0.87     38888
```

Decision Tree

```
Meilleur modèle trouvé avec F-mesure sur validation: 0.871221872212511
Évaluation sur l'ensemble de test:
F1-Score: 0.8674
Accuracy: 0.8529
Precision: 0.9049
Recall: 0.8529
```

Random Forest

```
Meilleur modèle trouvé avec F-mesure sur validation: 0.6636420919974795
Évaluation sur l'ensemble de test:
F1-Score: 0.8723
Accuracy: 0.8582
Precision: 0.9104
Recall: 0.8582
```

Q-Learning

```
Meilleure récompense moyenne sur l'entraînement: 28030.8000
Évaluation sur l'ensemble de test:
F1-Score: 0.7385
Accuracy: 0.7381
Precision: 0.7389
```

Conclusion

Au vu de notre présentation , certes nous avons travailler sur un domaine passé mais assez connu qui est le Covid, nous pouvons sans souci dire que les algorithmes de machine learning sont assez efficace pour déterminer le risque d'être affecter par le covid qu'une personne peut présenter avec un assez bon degré de précision et de recall.

Au terme de notre étude basée sur les différents modèles nous en avons conclus que c'est le model du Random Forest qui donne un meilleur résultat par comparaison du f1-score.

Ce qui est important c'est l'application que nous pouvons en faire. En effet lors d'une nouvelle pandémie où nous nous trouverons avec un déficite de matériel et de ressource il nous sera possible ou il sera possible aux différents hôpitaux d'apdoter cette méthode et determiner qui sont les personnes à risques ou non et que faire dans ce sens.

Merci pour votre attention