Détection du Covid-19

grâce au Machine Learning



**Hassouna Hiba  
Naziha saghrouni**

Faculté des sciences économique et gestion

Departement de l’informatique

Master de Recherche BC-1 ère année

PLAN

1. Introduction
2. Travaux liés
3. Description des données
4. Méthodologique
5. Conclusion
6. Références

# I. Introduction

En décembre 2019, le nouveau coronavirus est apparu dans la ville de Wuhan en Chine 1][ et a été signalé à l'Organisation mondiale de la santé (OMS) le 31 décembre 2019. Le virus a représenté une menace mondiale et a été nommé COVID-19 par l'OMS le 11 février 2020. L'OMS a déclaré l'épidémie une urgence de santé publique [2] et a déclaré ce qui suit : "le virus se propage par les voies respiratoires lorsqu'une personne en bonne santé entre en contact avec une personne infectée". Une personne infectée présente des symptômes dans un délai de 2 à 14 jours. Selon l'OMS, les symptômes et signes des affections modérées à graves sont la toux sèche, la fatigue et la fièvre, tandis que dans les cas graves, la dyspnée, la fièvre et la fatigue peuvent survenir. Les personnes atteintes d'autres maladies telles que l'asthme, le diabète et les maladies cardiaques ont un risque accru de contracter le virus et peuvent devenir gravement malades. Un système capable de détecter le virus est devenu nécessaire en raison de la propagation rapide du virus, tuant des centaines de milliers de personnes. Les algorithmes de classification d'apprentissage automatique, les ensembles de données et les logiciels d'apprentissage automatique sont des outils essentiels pour concevoir le modèle prédictif de la COVID-19. Ce projet vise à comparer différents algorithmes d'apprentissage automatique tels que les voisins les plus proches, la forêt aléatoire et le Bayes naïf en ce qui concerne leurs précisions, puis à utiliser le meilleur parmi eux pour développer un système qui prédit si une personne a ou non la COVID en utilisant les données fournies au modèle.

# II . Travaux liés

Un diagnostic précis et fiable de la COVID-19 peut sauver des millions de vies et fournir une grande quantité d'informations sur lesquelles les modèles d'apprentissage automatique (ML) peuvent être entraînés. L'apprentissage automatique peut fournir des contributions utiles à cet égard, notamment pour établir des diagnostics à partir de la littérature clinique, d'images radiographiques, etc. Des études [3] ont montré qu'il était possible de différencier efficacement les patients atteints de la COVID-19 dans 85 % des cas en utilisant un algorithme de vecteurs de support (SVM). L'étude a analysé les résultats des tests COVID-19 réalisés à l'Hôpital Israelita Albert Einstein (HIAE) à São Paulo, au Brésil. C'était l'un des principaux centres de dépistage de la Covid dans le pays au cours des premières semaines de l'épidémie. Cette étude a été menée par une équipe chargée de répondre à l'urgence de la COVID-19. Elle a testé les performances de certains algorithmes d'apprentissage automatique (réseaux neuronaux, arbres de décision boostés par gradient, forêts aléatoires, régression logistique et SVM) pour détecter la positivité de la COVID. L'étude [4] a utilisé plusieurs séparateurs, dont la régression logistique, le perceptron multicouche (MLP) et XGBoost dans la même base de données de l'hôpital brésilien [9]. Elle a classé les patients atteints de la COVID-19 avec une précision supérieure à 91 %. Le travail [5] a développé et évalué un algorithme d'apprentissage automatique pour le diagnostic de la Covid-19. L'algorithme a été conçu sur la base de données démographiques et de caractéristiques de laboratoire. Ils ont collecté des données du UCLA Health System à Los Angeles, en Californie. Cela comprenait tous les services d'urgence et les cas hospitalisés recevant des tests PCR SARS-CoV-2 avec un ensemble de 1 455 caractéristiques de laboratoire auxiliaires du 1er mars 2020 au 24 mai 2020. Ils ont testé avec certains modèles d'apprentissage automatique et ont utilisé une combinaison de ceux-ci pour la classification finale. L'algorithme développé avait une sensibilité de 0,093 et une spécificité de 0,64. La fonction [7] prédit la COVID-19 avec une précision de 91 % et 89 %, respectivement. De plus, une prédiction des besoins en soins intensifs / semi-soins intensifs a été réalisée dans 98 % des cas [6]. Comme très peu de travaux sont réalisés sur les diagnostics et les prédictions à l'aide de textes, nous avons utilisé des modèles d'apprentissage automatique pour classer les rapports cliniques en COVID positif ou COVID négatif.

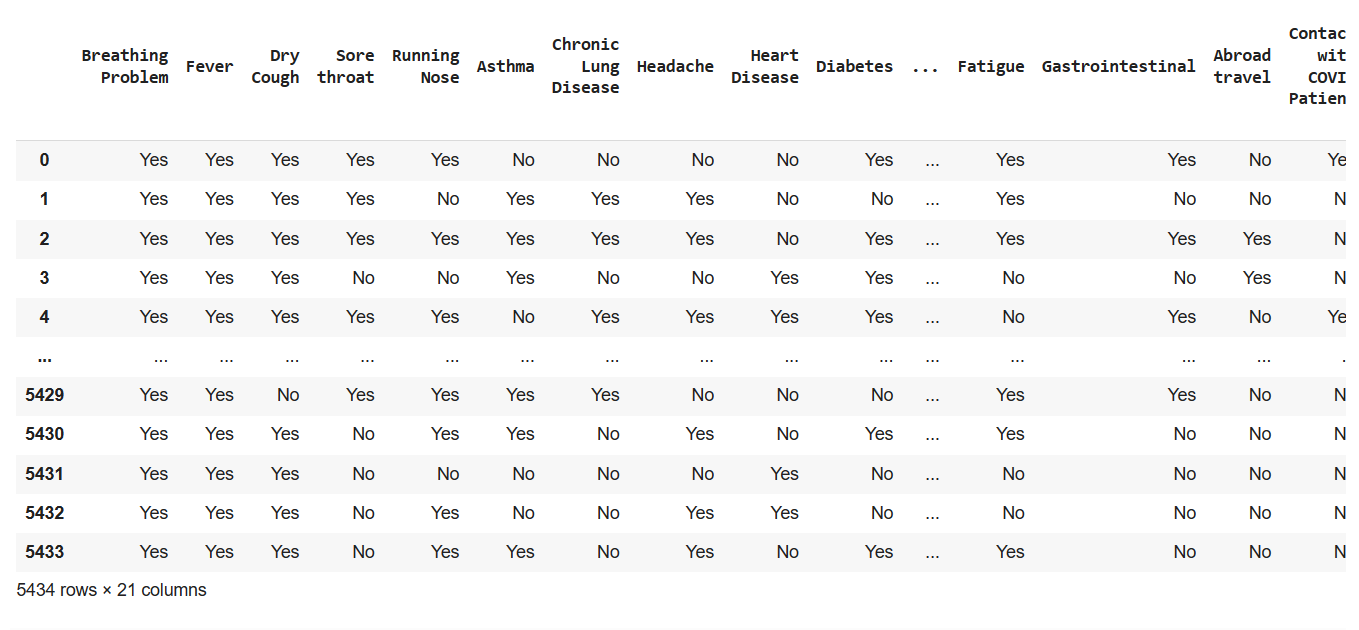
# III . Description des Données

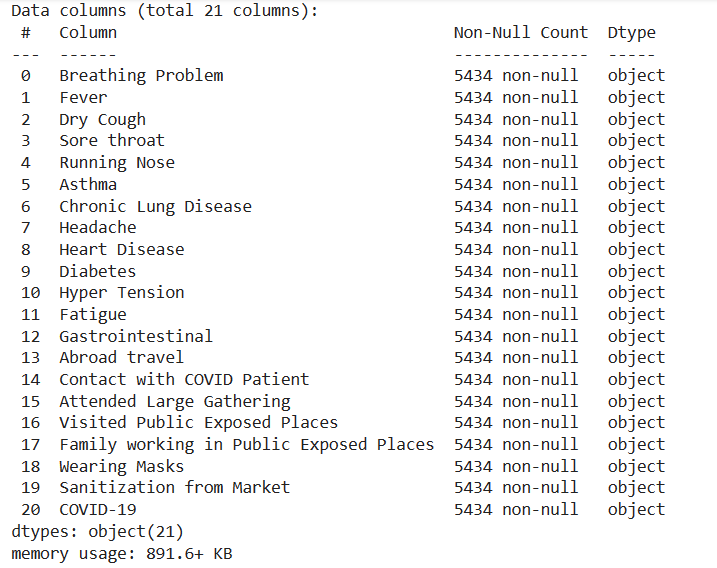
Dès que l’OMS a déclaré la pandémie de coronavirus comme urgence sanitaire, les chercheurs et les hôpitaux ont ouvert l’accès aux données liées à l’épidémie. Nous avons obtenu un ensemble de données auprès de kaggle.com et il comporte :

5434 × 21 lignes de colonnes. Cet ensemble de données contient 20 variables qui pourraient être déterminantes dans la prédiction de la COVID­19, ainsi qu'un attribut de classe qui définit si la COVID­19 est détectée.

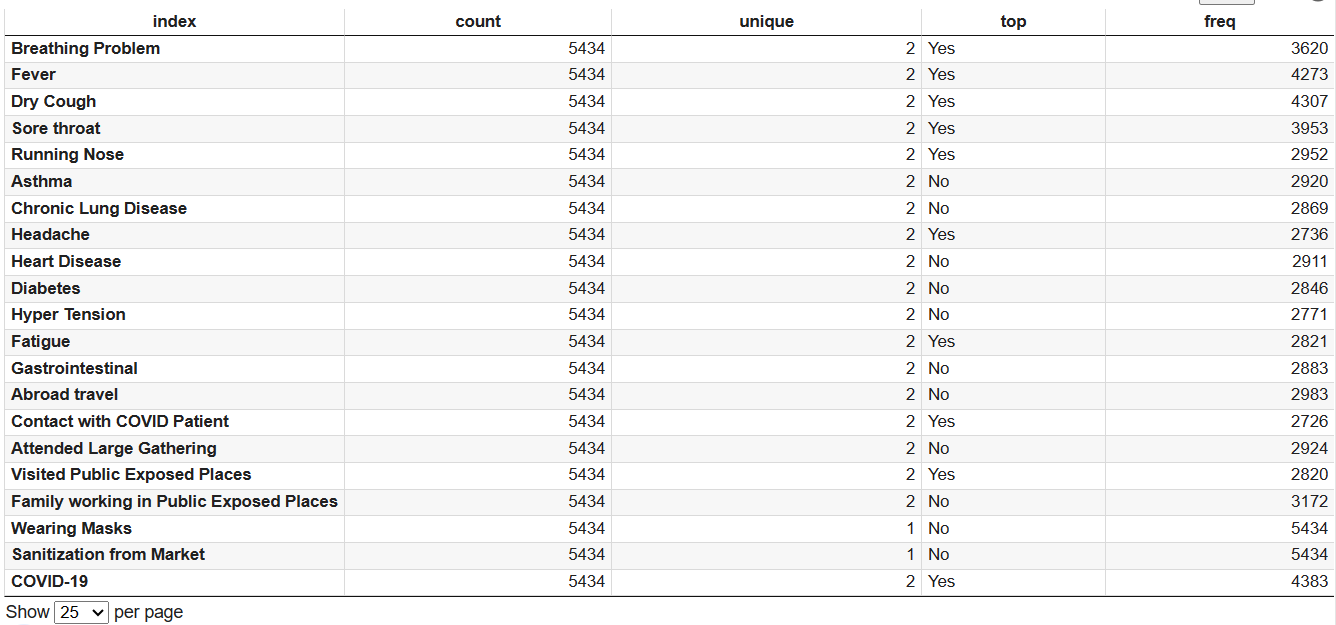
L'ensemble de données utilisé, nommé COVID Dataset, est une collection de données médicales comprenant des informations sur les patients, telles que les symptômes et les diagnostics de COVID-19. Les données sont étiquetées pour indiquer si un patient est positif ou négatif pour la COVID-19.

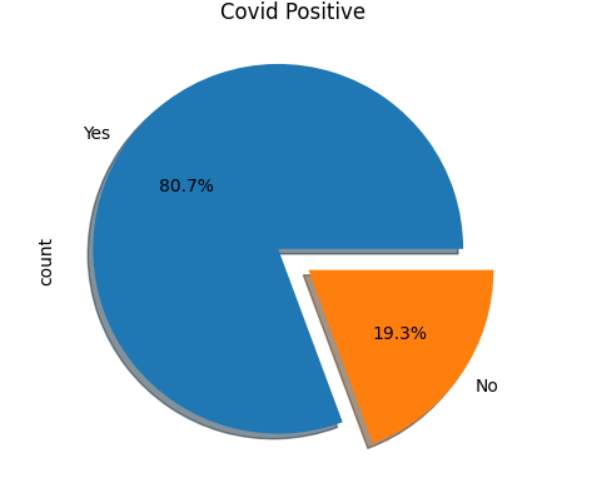
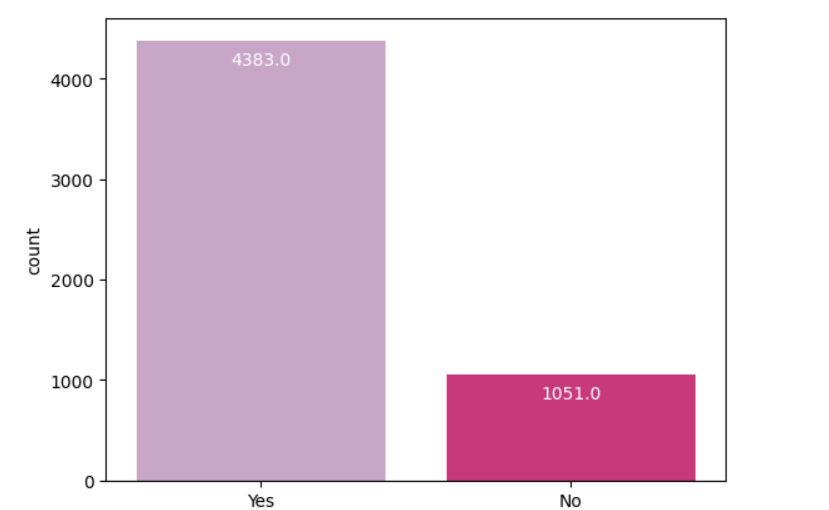
patient = [[Breathing\_Problem,Fever,Dry\_Cough,Sore\_throat,Running\_Nose,Asthma,Chronic\_Lung\_Disease,Headache,Heart\_Disease,Diabetes,Hyper\_Tension,Fatigue,Gastrointestinal,Abroad\_travel,Contact\_with\_COVID\_Patient,Attended\_Large\_Gathering,Visited\_Public\_Exposed\_Places,Family\_working\_in\_Public\_Exposed\_Places]]



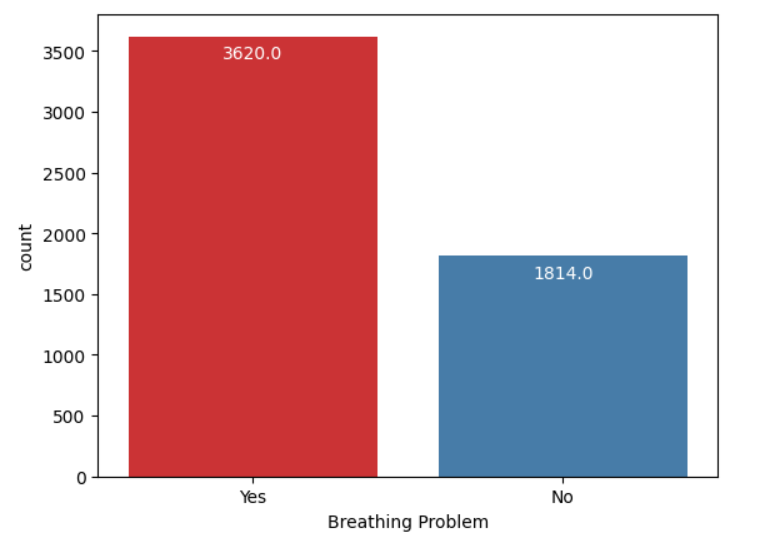


Covid dataset ne contient pas des valeurs non nulles.





Notre Dataset contient 4383 (80.7%) patients qui sont affectés par le virus et 1051 personnes (19.3%) qui ne portent pas le virus.



Parmi ces personnes qui ont porté le virus on a 3620 ont des problèmes respiratoires.

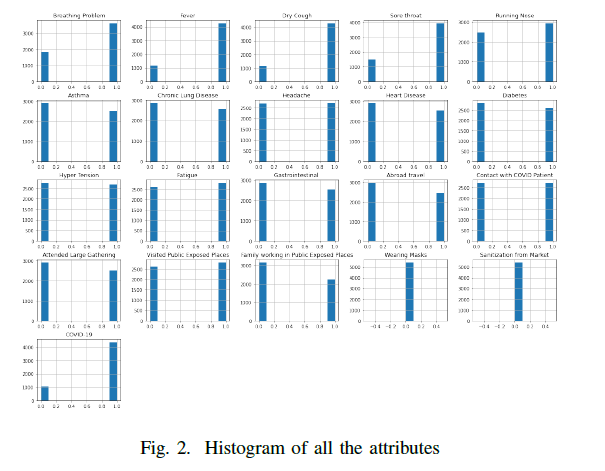
# Iv. Méthodologie

1. Prétraitement des Données: Les données sont nettoyées et préparées pour l'analyse, y compris le traitement des valeurs manquantes et l'encodage des variables catégorielles.
2. Division des Données: L'ensemble de données est divisé en ensembles d'entraînement et de test pour évaluer les performances des modèles.
3. Modélisation: Les quatre algorithmes de machine learning (régression logistique, k-NN, forêts aléatoires et SVM) sont entraînés sur l'ensemble d'entraînement.
4. Évaluation: Les modèles sont évalués sur l'ensemble de tests en utilisant des mesures de performance telles que la précision, le rappel et le score F1.

**1. Prétraitement des données**

Le processus de conversion des données brutes dans un format compréhensible est appelé prétraitement des données. Les données du monde réel peuvent comporter du bruit, des valeurs manquantes ou être dans un format incompatible qui empêche leur utilisation directe dans les modèles d'apprentissage automatique. Le prétraitement des données est une étape essentielle dans laquelle nous nettoyons les données et les rendons compatibles, c'est ­à­ dire aptes à être utilisées dans un modèle d'apprentissage automatique. Cela améliore également la précision et l’efficacité du modèle. Les principales étapes du prétraitement des données sont les suivantes :

1) Suppression de fonctionnalités : à partir de la figure 2, nous pouvons conclure que le port de masques et la désinfection du marché sont deux fonctionnalités qui n'ont qu'une seule valeur qui est « non » car elles n'affectent pas nos prédictions. Nous pouvons simplement supprimer ces colonnes de notre base de données.



2) Codage des données catégorielles : le codage d'étiquetage est une forme populaire de gestion de code flexible pour les catégories. Dans ce processus, chaque étiquette reçoit un numéro entier basé sur l'ordre alphabétique. Tous les attributs de notre ensemble de données sont de type « oui » ou « non ». Nous avons donc utilisé le codage d'étiquettes pour le convertir en 0 et 1 afin que le modèle puisse mieux comprendre l'ensemble de données. Le tableau 4 montre l'ensemble de données après application du codage d'étiquette.

3) Fractionnement de l'ensemble de données : l'étape suivante du prétraitement des données d'apprentissage automatique consiste à diviser l'ensemble de données. L'ensemble de données d'un modèle d'apprentissage automatique doit être divisé en deux parties : la formation et les tests**.**

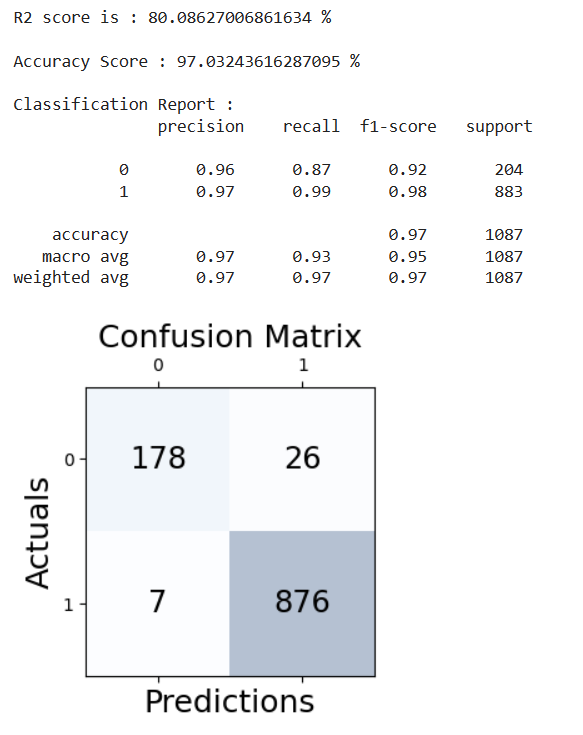
**2. Division des données**

Nous avons divisé les données selon une répartition de 80 :20. Cela signifie que nous utilisons 80 % des données pour entraîner le modèle tout en conservant les 20 % restants pour les tests. Nous prenons les 20 indépendants

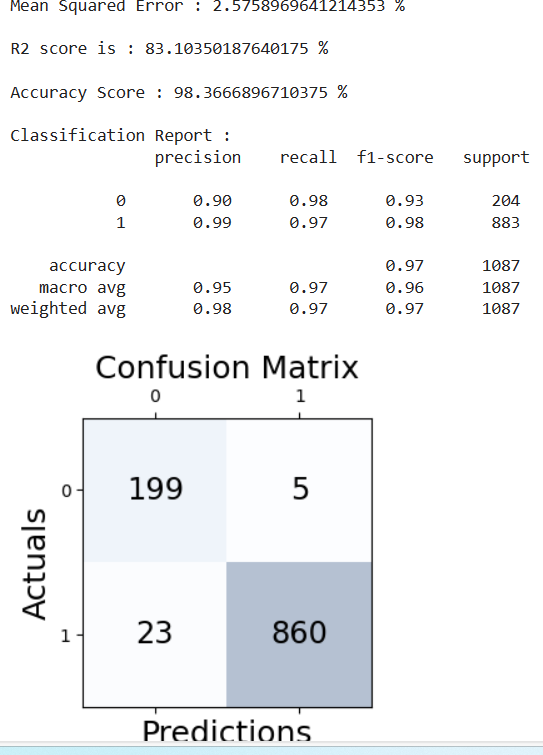
**3. Modélisation**

Dans cette section, nous détaillerons les étapes spécifiques de développement des modèles pour la détection de la COVID-19 en utilisant la régression logistique, k-NN, les forêts aléatoires et les SVM.

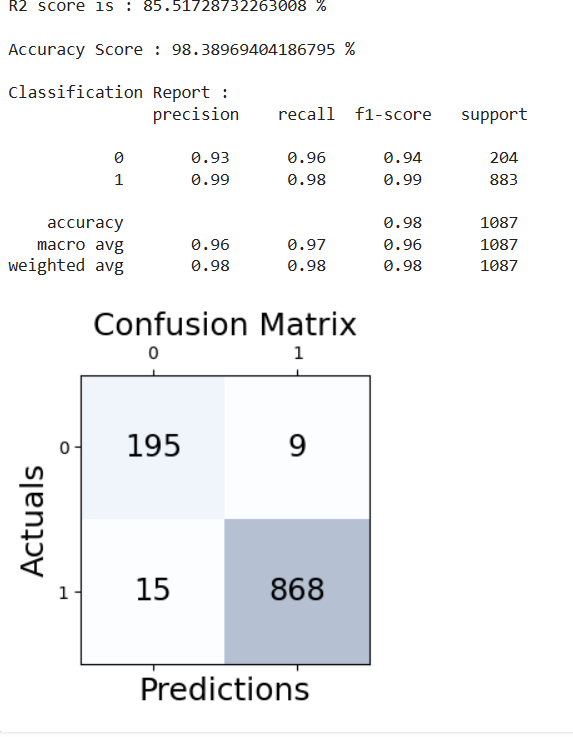
1. Régression Logistique:
   * Entraînement du Modèle: La régression logistique est un algorithme de classification qui modélise la probabilité qu'un échantillon appartient à une classe spécifique. Nous ajustons les poids des caractéristiques par une fonction logistique pour minimiser l'erreur.
   * Réglage des Hyperparamètres: Pour la régression logistique, nous pouvons ajuster des hyper paramètres.



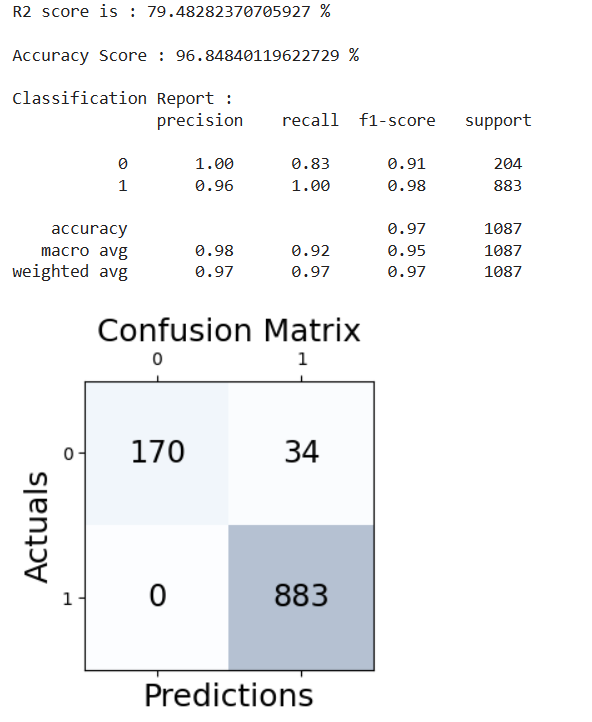
1. k-NN (k plus proches voisins):
   * Entraînement du Modèle: k-NN est une méthode non paramétrique utilisée pour la classification et la régression. Pour chaque échantillon de test, les k échantillons d'entraînement les plus proches sont sélectionnés pour déterminer la classe.
   * Choix de k: Nous devons choisir le nombre optimal de voisins (k) pour notre modèle, ce qui peut être déterminé par validation croisée ou d'autres méthodes d'optimisation.(On utilise dans notre projet k=2)



1. Forêts Aléatoires:
   * Entraînement du Modèle: Les forêts aléatoires sont une méthode d'apprentissage ensembliste basée sur la création de multiples arbres de décision. Chaque arbre est entraîné sur un sous-ensemble aléatoire des données et des caractéristiques.
   * Tuning des Hyperparamètres: Des hyper paramètres tels que le nombre d'arbres, la profondeur maximale des arbres et le nombre minimum d'échantillons par feuille peuvent être ajustés pour optimiser les performances du modèle.



1. SVM (Machines à Vecteurs de Support):
   * Entraînement du Modèle: Les SVM sont des modèles d'apprentissage supervisé utilisés pour la classification et la régression. Ils cherchent à trouver l'hyperplan optimal qui sépare les données en deux classes.
   * Choix du noyau: Les SVM peuvent utiliser différents types de noyaux, tels que linéaire, polynomial ou gaussien (RBF). Le choix du noyau approprié dépend de la structure des données.



Une fois que les modèles ont été entraînés sur l'ensemble de données d'entraînement, nous évaluons leurs performances sur l'ensemble de test en utilisant des métriques telles que la précision, le rappel et le score F1. Ces métriques nous permettent de mesurer la capacité des modèles à détecter correctement les cas positifs et négatifs de la COVID-19.

En résumé, la modélisation implique l'entraînement de plusieurs algorithmes de machine learning sur les données d'entraînement, le réglage des hyperparamètres pour optimiser les performances, et enfin, l'évaluation des modèles sur un ensemble de test pour déterminer leur efficacité dans la détection de la COVID-19.

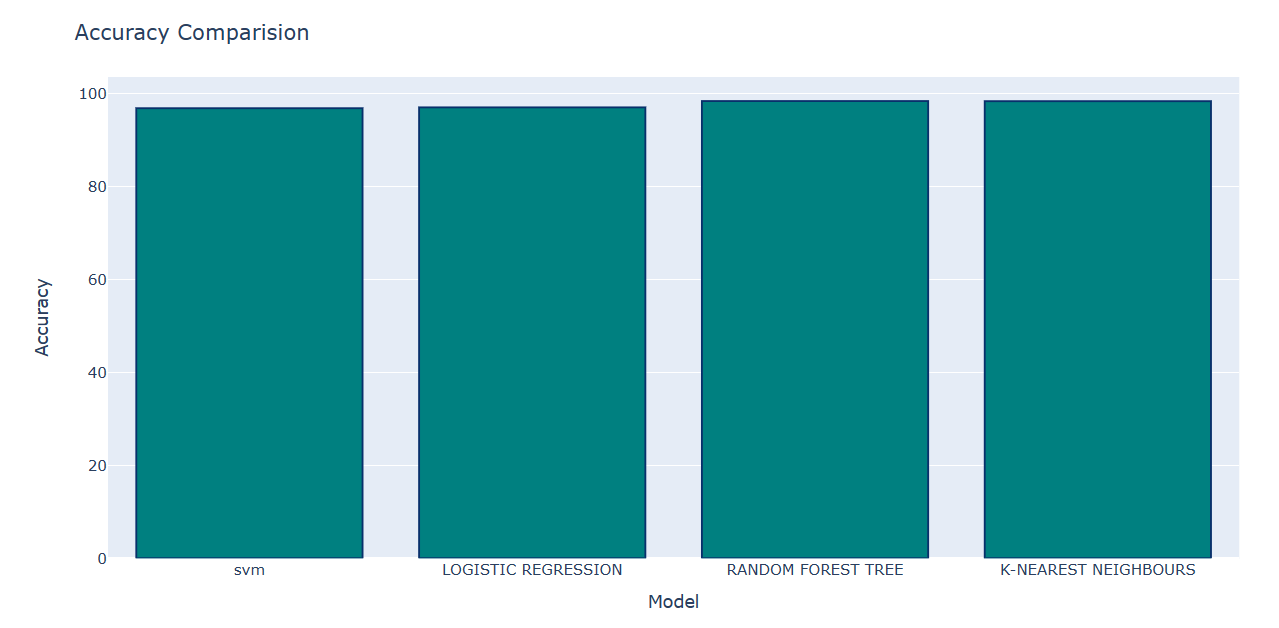
**3. Evaluation**

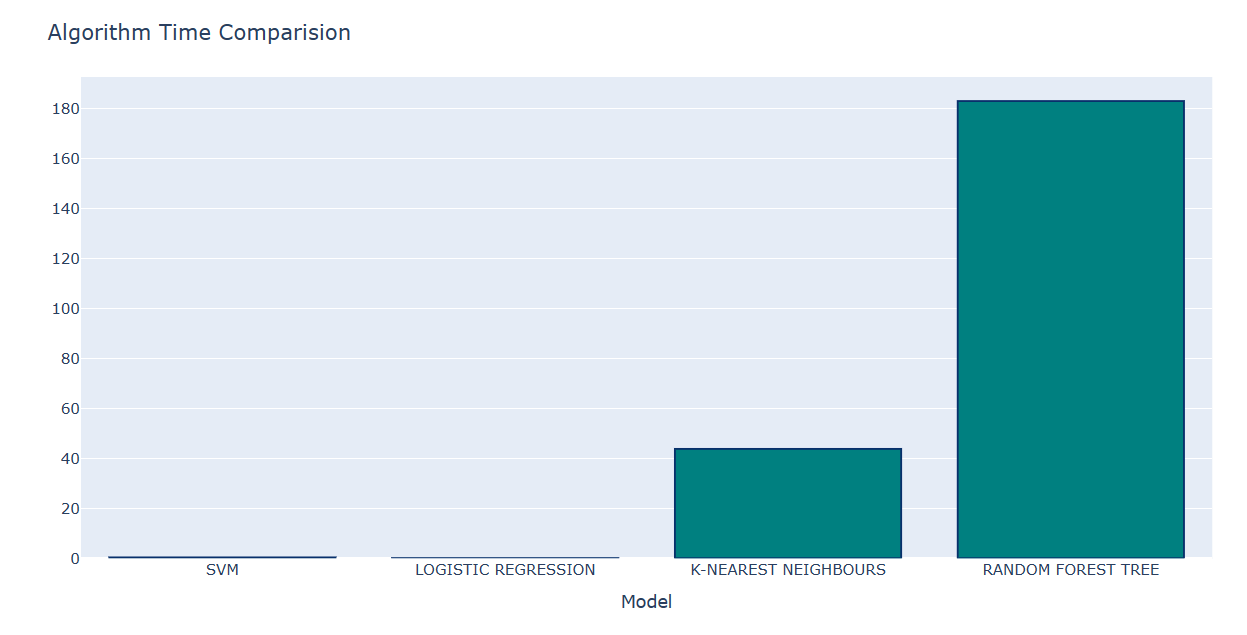
# 

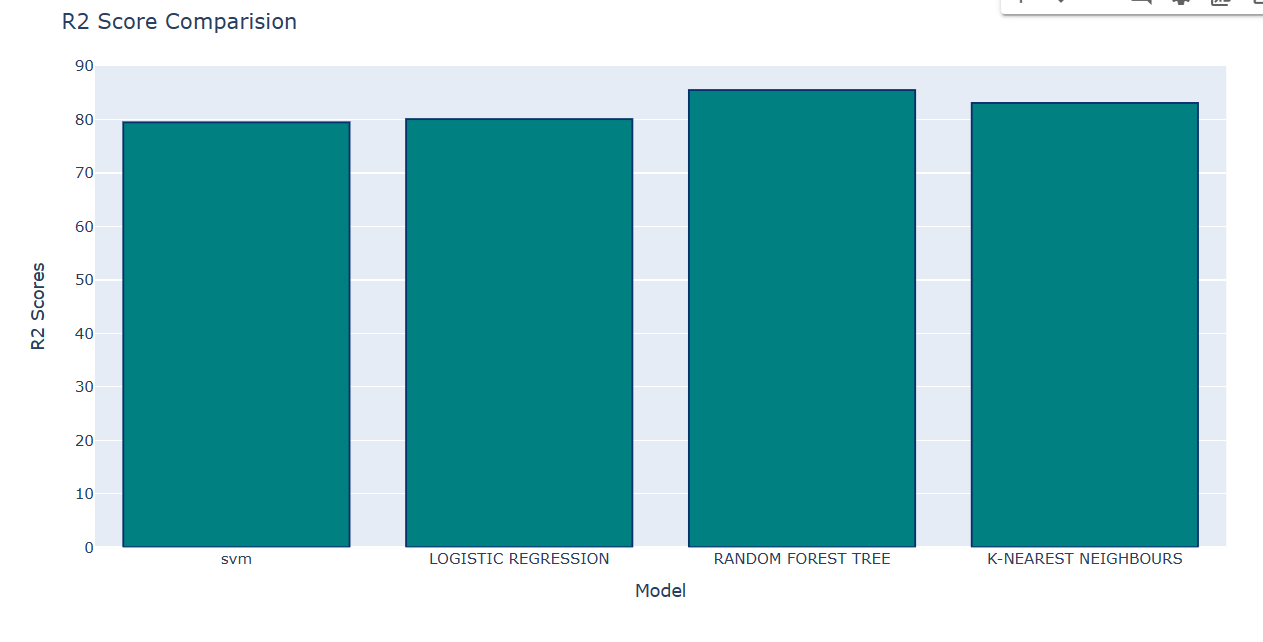
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Modèle | Temps d'exécution(secondes) | Accuracy | F2 SCore | Mean squared error | Roc score |
| Régression Logistique | 0.038 | 97.03% | 80.086 | 93.23% | 3.04% |
| k-Nearest Neighbors (k-NN) | 43.771 | 98.37% | 83.103 | 97.47% | 2.58% |
| Forêts Aléatoires (Random Forest) | 183.028 | 98.39% | 85.517 | 96.94% | 2.21% |
| SVM (Support Vector Machine) | 0.333 | 96.85% | 79.482 | 91.67% | 3.13% |

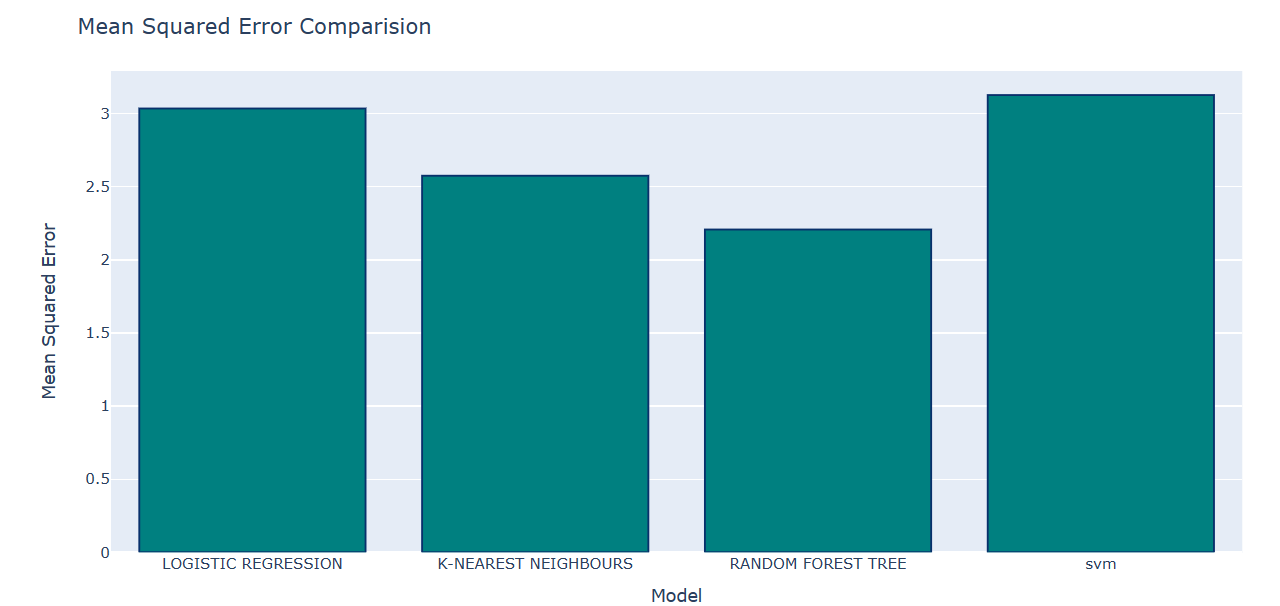
# 

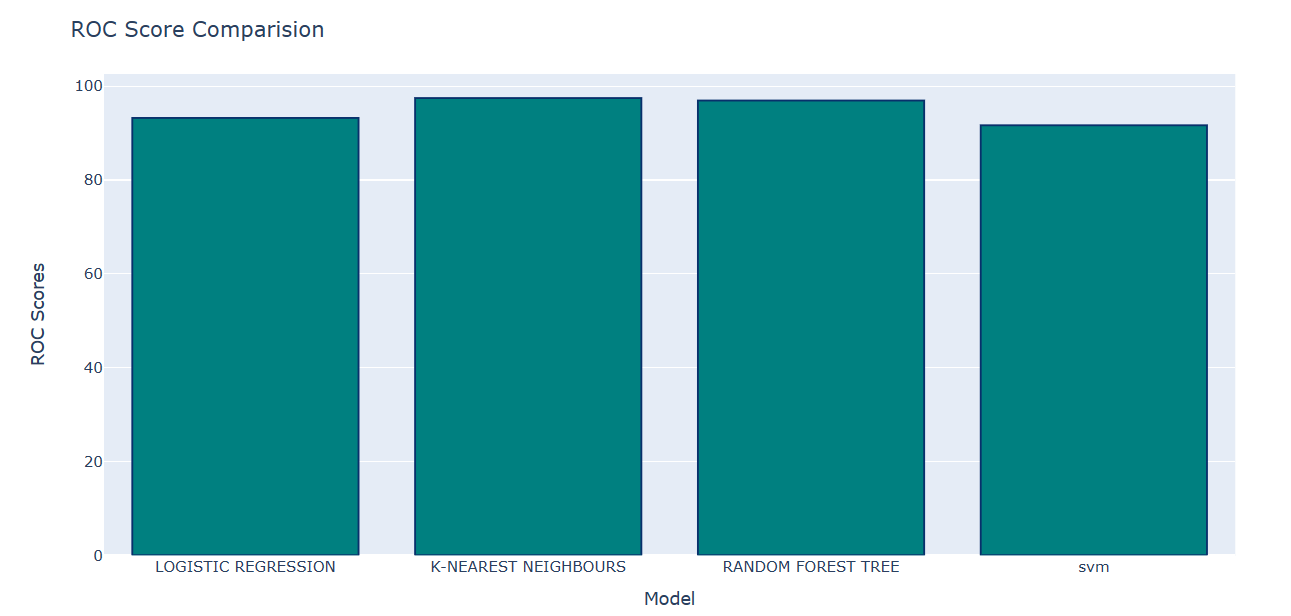
# 











# En comparant ces résultats :

# Accuracy : Les forêts aléatoires et k-NN obtiennent les meilleures performances en termes d'accuracy, suivies de la régression logistique et des SVM.

# F2-score : Les forêts aléatoires ont le meilleur F2-score, suivi de k-NN, de la régression logistique et des SVM.

# ROC score : k-NN obtient le meilleur score ROC, suivi des forêts aléatoires, de la régression logistique et des SVM.

# Mean Squared Error : Les forêts aléatoires ont la plus faible erreur quadratique moyenne, ce qui indique une meilleure précision de prédiction, suivies de k-NN, de la régression logistique et des SVM.

* Temps d'exécution: Régression logistique et SVM sont les meilleurs.

# 

# Dans une analyse comparative, les résultats montrent que le classificateur KNN avec un nombre de voisins égal à 2 est le meilleur algorithme d'apprentissage automatique, avec une précision de 98,37 % et une erreur absolue moyenne de 0,026 en considérant le temps d'exécution pour l'apprentissage. En comparaison avec d'autres méthodes, le modèle prend un temps moyen mais offre une bonne précision.

# Cette recherche peut être utilisée comme un outil de soutien pour la prise de décision par les médecins, le modèle établi aidant à reconnaître la présence du COVID-19 chez une personne en fonction de ses symptômes. Les individus présentant des symptômes liés au COVID-19 peuvent également l'utiliser pour évaluer s'ils seront testés positifs ou négatifs au COVID-19. Le modèle développé ici peut être utilisé pour déployer une application avec les fonctionnalités suivantes:

# Prediction model takes input from the user and gives a result - COVID Positive

# 

# 

# V . CONCLUSION

L'objectif de ce travail était d'utiliser quatres techniques d'apprentissage automatique supervisé pour créer un modèle de prédiction de la présence du COVID-19. Les performances du modèle ont été évaluées en termes d'exactitude, d'erreur quadratique moyenne (MSE), de score R2 et de score ROC, ainsi que du temps d'exécution.

Les résultats montrent que le classificateur KNN avec un nombre de voisins égal à 2 est le meilleur algorithme d'apprentissage automatique, avec une exactitude de 98,37 % et une erreur absolue moyenne de 0,026 en considérant le temps d'exécution pour l'apprentissage. En comparaison avec d'autres méthodes, le modèle prend un temps moyen mais offre une bonne précision.

Cette recherche peut être utilisée comme un outil de soutien pour la prise de décision par les médecins, le modèle établi aidant à reconnaître la présence du COVID-19 chez une personne en fonction de ses symptômes. Les individus présentant des symptômes liés au COVID-19 peuvent également l'utiliser pour évaluer s'ils seront testés positifs ou négatifs au COVID-19.

Le modèle développé ici peut être utilisé pour déployer une application avec les fonctionnalités suivantes :

* Les individus peuvent rapidement déterminer s'ils sont susceptibles de transmettre le COVID-19 en fonction de leurs symptômes.
* Les praticiens médicaux peuvent utiliser ce test comme évaluation primaire de la santé pour la détection du COVID.
* Aider les entreprises à limiter l'interaction physique avec les clients qui pourraient être infectés par le COVID-19.

Des informations supplémentaires ou des diagnostics provenant des dossiers hospitaliers, des personnes ayant contracté le virus, des survivants du COVID-19, des patients en évaluation ou en gestion peuvent tous être inclus pour de futures recherches. Un logiciel capable de prédire la gravité du COVID-19 peut en effet être déployé pour fournir des informations supplémentaires sur les mesures à prendre et les interventions à considérer.

# VI . RÉFÉRENCES

[1] Wu F, Zhao S, Yu B, Chen YM, Wang W, Song ZG, Hu Y, Tao ZW, Tian

JH, Pei YY, Yuan ML, Zhang YL, Dai FH, Liu Y, Wang QM, Zheng JJ,

Xu L, Holmes EC, Zhang YZ (2020) A new coronavirus associated with

human respiratory disease in china. Nature 44(59):265–269

https://www.nature.com/articles/s41586-020-2008-3[1]

[2] Medscape Medical News, The WHO declares public health emergency

for novel coronavirus (2020)

https://www.medscape.com/viewarticle/924596[2]

[3] A. F. M. Batista, J. L. Miraglia, T. H. R. Donato and A. D. P. C. Filho,

”COVID-19 diagnosis prediction in emergency care patients: a machine

learning approach”, medRxiv, 2020

https://www.medrxiv.org/content/10.1101/2020.04.04.20052092v2.full.

pdf[3]

[4] M. R. H. Mondal, S. Bharati, P. Podder and P. Podder, ”Data analytics

for novel coronavirus disease”, Informatics in Medicine Unlocked

Elsevier, vol. 20, pp. 100374, 2020

https://www.researchgate.net/publication/342195015 Data analytics

for novel coronavirus disease[4]

[5] D. Goodman-Meza, A. Rudas, J. N. Chiang, P. C. Adamson, J. Ebinger

et al., ”A machine learning algorithm to increase COVID-19 inpatient

diagnostic capacity”, PLOS ONE, vol. 15, no. 9, pp. e0239474, 2020.

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7508387/[5]

[6] P. Schwab, A. D. Sch¨utte, B. Dietz and S. Bauer, ”Clinical predictive

models for COVID-19: systematic study”, J Med Internet Res, vol. 22,

no. 10, pp. e21439, 2020.https://www.jmir.org/2020/10/e21439/[6]

[7] Y. Sun, V. Koh, K. Marimuthu, O. T. Ng, B. Young, S. Vasoo, M. Chan et

al., ”Epidemiological and clinical predictors of COVID-19”, Clin Infect

Dis, vol. 71, no. 15, pp. 786-792, Jul 2020.

https://academic.oup.com/cid/article/71/15/786/5811426[7]