 Mini Projet

Machine Learnig

| **Covid19 Detection** |
| --- |

Prof : Dr Fadoua Bouafif

Réalisé par : Saghrouni Naziha

Hsouna Hiba

1ére mastère BC

Détection du Covid­19 grâce au Machine Learning

Les progrès technologiques ont un effet profond sur toutes les sphères de la vie, que ce soit dans le domaine médical ou dans tout autre domaine.

L’intelligence artificielle a montré des résultats prometteurs dans le domaine des soins de santé en prenant ses décisions en analysant et en traitant des données. Pour prévenir la propagation et le développement d’une maladie potentiellement mortelle, l’étape la plus importante est son diagnostic précoce.

Le COVID­19 est une maladie hautement contagieuse et est devenue une épidémie mondiale à laquelle il faut lutter le plus rapidement possible.

En raison de sa vitesse de propagation rapide, il est nécessaire de disposer d’un système pouvant être utilisé pour détecter le virus. Avec l’utilisation croissante de la technologie, de nombreuses données sur le COVID­19 sont facilement accessibles du bout des doigts, qui peuvent être utilisées pour obtenir des informations importantes sur le virus.

Dans ce projet, nous avons comparé la précision de différents algorithmes d’apprentissage automatique pour prédire le COVID­19 et avons utilisé le plus précis lors des tests du modèle final.

**INTRODUCTION**

En décembre 2019, le nouveau Coronavirus est apparu dans la ville de Wuhan en Chine et a été signalé à l'Organisation mondiale de la santé (OMS) le 31 décembre 2019.

Le virus représentait une menace mondiale et a été nommé COVID­19 par l'OMS le 11 Février 2020.

L’OMS a déclaré l’épidémie d’urgence de santé publique et a déclaré ce qui suit : "Le virus se propage par les voies respiratoires lorsqu'une personne saine 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 d'affections modérées à sévères sont une toux sèche, de la fatigue et de la fièvre, tandis que dans les cas graves, une dyspnée, de la fièvre et de la fatigue peuvent survenir. Les personnes atteintes d’autres maladies telles que l’asthme, le diabète et les maladies cardiaques courent un plus grand risque de contracter le virus et peuvent tomber gravement malades. Un système permettant 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, les ensembles de données et les logiciels d’apprentissage automatique sont des outils essentiels pour concevoir le modèle prédictif du COVID­19.

Ce projet vise à comparer différents systèmes d'apprentissage automatique des algorithmes tels que Logistic regression , Knn (K­ plus proches voisins) , Random forest et SVM (Support Vector Machine) en ce qui concerne leurs précisions, puis utilisent le meilleur d'entre eux pour développer un système qui prédit si une personne a le COVID ou non en utilisant les données fournies au modèle.

**TRAVAUX CONNEXES**

**Un diagnostic solide et précis du COVID­19 peut sauver des millions de vies et fournir une grande quantité d’informations sur lesquelles les modèles d’apprentissage automatique (ML) peuvent être formés. Le ML peut fournir des informations utiles à cet égard, en particulier pour établir des diagnostics basés sur la littérature clinique, les images radiographiques, etc.**

**Il a été démontré que les études dans différencient efficacement les patients atteints de COVID­19 dans 85 % des cas en utilisant un algorithme de vecteur de support (SVM). L'étude a analysé les résultats des tests COVID­19 effectués à l'hôpital Israelita Albert Einstein (HIAE) à Sao Paulo, au Brésil.**

**C’était l’un des principaux centres de test Covid du pays au cours des premières semaines de l’épidémie. Cette étude a été menée par un groupe de travail dont le but est de répondre à l’urgence du COVID­19. Il a testé les performances de positivité du COVID de certains algorithmes d’apprentissage automatique (réseaux de neurones, arbres boostés par gradient, forêts aléatoires, régression logistique et SVM).**

**L'étude de a utilisé plusieurs diviseurs, notamment la régression logistique, le perceptron multicouche (MLP) et XGBoost dans la même base de données de l'hôpital brésilien .**

**Il a classé les patients atteints du COVID­19 avec une précision supérieure à 91 %. Les travaux ont développé et évalué un algorithme ML pour le diagnostic de Covid­19. L'algorithme a été conçu sur la base de données démographiques et de fonctionnalités de laboratoire. Ils ont collecté des données auprès du système de santé de l'UCLA à Los Angeles, en Californie. Il comprenait toutes les salles d'urgence et les cas de patients hospitalisés ayant reçu un test PCR du SRAS­ CoV­2 avec un ensemble de 1 455 caractéristiques de laboratoire auxiliaires du 1 Mars 2020 au 24 mai 2020. Ils ont testé avec certains modèles ML et ont utilisé une combinaison de ceux­ci pour la classification finale. L'algorithme développé avait une sensibilité de 0,093 et une spécificité de 0,64. La fonction de prédit le COVID­19 avec une précision de 91 % et 89 %, respectivement**

**De plus, une prédiction des besoins en réanimation/semi­réanimation a été faite dans 98 % des cas. Étant donné que très peu de travaux sont effectués sur le diagnostic et la prédiction à l’aide de texte, nous avons utilisé des modèles d’apprentissage automatique pour classer les rapports cliniques en positifs au COVID ou en négatifs au COVID.**

**MÉTHODOLOGIE**

**La méthodologie proposée comprend 4 étapes:**

* **Etape 1: la collecte de données est en cours**
* **Etape 2 : donne un aperçu du prétraitement**
* **Etape 3: une analyse exploratoire est en cours pour comprendre l'ensemble de données et à la dernière étape**
* **Etape 4: inclut le réglage des hyperparamètres par CV de recherche de grille.**

**A. Collecte de données :**

**Alors 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.**

**B. 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.**

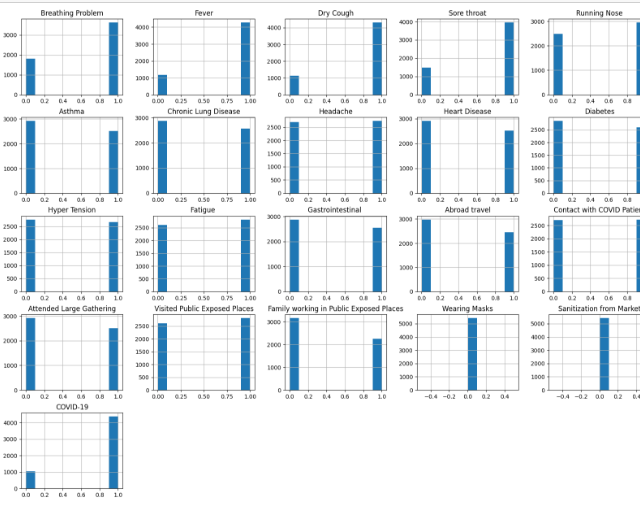
**2) Codage des données catégorielles : le codage d'étiquetage est une forme populaire de gestion de code flexible pour les catégories. Dans ce processus, chaque étiquette reçoit un numéro entier basé sur l'ordre alphabétique. Tous les attributs de notre ensemble de données sont de type « oui » ou « 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.**

**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**

****

***Fig. 1. 1. Nombre de valeurs manquantes et pourcentage manquant de tous les attributs en x et la colonne dépendante « COVID-19 » en y car nous visons à prédire si le patient est positif au COVID ou non.***

****

***Fig. 2. Histogramme de tous les attributs***

**C. Analyse exploratoire des données :**

**L'analyse exploratoire des données est utilisée pour évaluer divers ensembles de données dans le but de les résumer en fonction de leurs caractéristiques clés. Ce résumé peut être visualisé à l'aide de graphiques statistiques et d'autres techniques de visualisation de données. EDA aide les data scientists de différentes manières : ­**

**• Améliorer la compréhension des données**

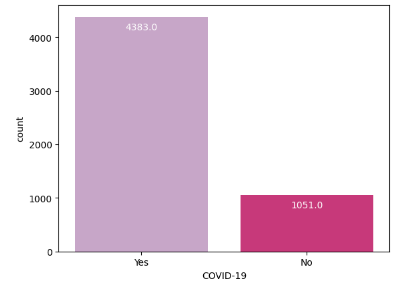
**• Détecter une variété de modèles de données**

**• Meilleure compréhension de l'énoncé du problème**

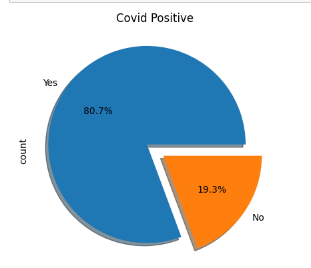
**D. Réglage des hyperparamètres par CV de recherche de grille :**

**Son objectif principal est de découvrir les paramètres optimaux pour lesquels l'efficacité du modèle est la meilleure ou la plus élevée et le taux d'erreur est le minimum.**

**Nous avons utilisé l'outil gridsearchcv pour produire la meilleure combinaison de paramètres, basée sur le score de précision comme mesure de notation lorsque tous les différents paramètres sont introduits dans la grille de paramètres.**

****

***Fig. 3. Histogramme montrant le nombre de patients avec covid positif et négatif***

****

***Fig. 4. Diagramme circulaire montrant le nombre de patients avec covid positif et négatif***

**MISE EN ŒUVRE**

**Avec le développement de la technologie informatique, la modélisation prédictive évolue. Nous sommes désormais en mesure de rendre la modélisation prévisible plus efficace et moins coûteuse qu’auparavant.**

**Dans notre projet, nous utilisons divers algorithmes de classification pour prédire et utiliser un CV de recherche par grille pour trouver la solution la plus avancée pour chaque algorithme. Certains des algorithmes de catégorisation utilisés comprennent :**

**A. Régression logistique : est une technique de catégorisation des données qui utilise l'apprentissage automatique. Cet algorithme modélise les chances des résultats potentiels d’une seule expérience utilisant une fonction logistique. Le moyen le plus simple de comprendre l’influence de nombreux facteurs indépendants sur une seule variable de résultat consiste à utiliser la régression logistique, conçue à cet effet. En général, l'algorithme calcule la probabilité d'appartenir à une classe particulière. Nous avons ici deux classes, y=0,1.**

**B. KNN :**

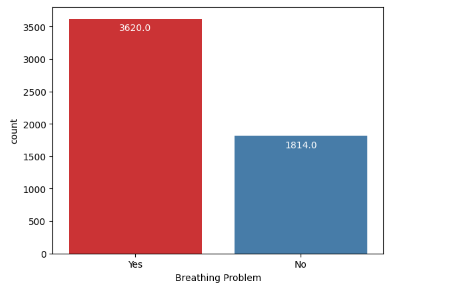
**Le plus ancien algorithme d'apprentissage automatique supervisé pour la classification est KNN, qui classe une instance donnée selon la majorité des catégories parmi ses k voisins les plus proches dans l'ensemble de données. La distance entre l'élément à catégoriser et tous les autres éléments de l'ensemble de données est calculée par l'algorithme.**

**C. Forêt aléatoire :**

**Ce classificateur est un méta­estimateur qui s'adapte aux arbres de décision sur les différents sous­échantillons de l'ensemble de données et utilise la moyenne pour augmenter la précision prévue du modèle et contrôler le surajustement.**

**Dans la plupart des cas, ce classificateur de forêt aléatoire semble être plus précis que les arbres de décision et minimise également le surajustement. Au niveau Random Forest, la moyenne sur tous les arbres est l'importance finale des caractéristiques.**

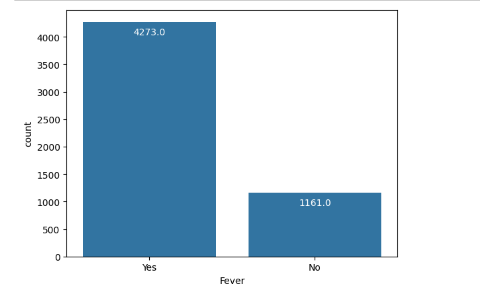
**La valeur de la somme d'importance de l'entité sur chaque arbre est calculée numériquement et divisée par le nombre total d'arbres :**

****

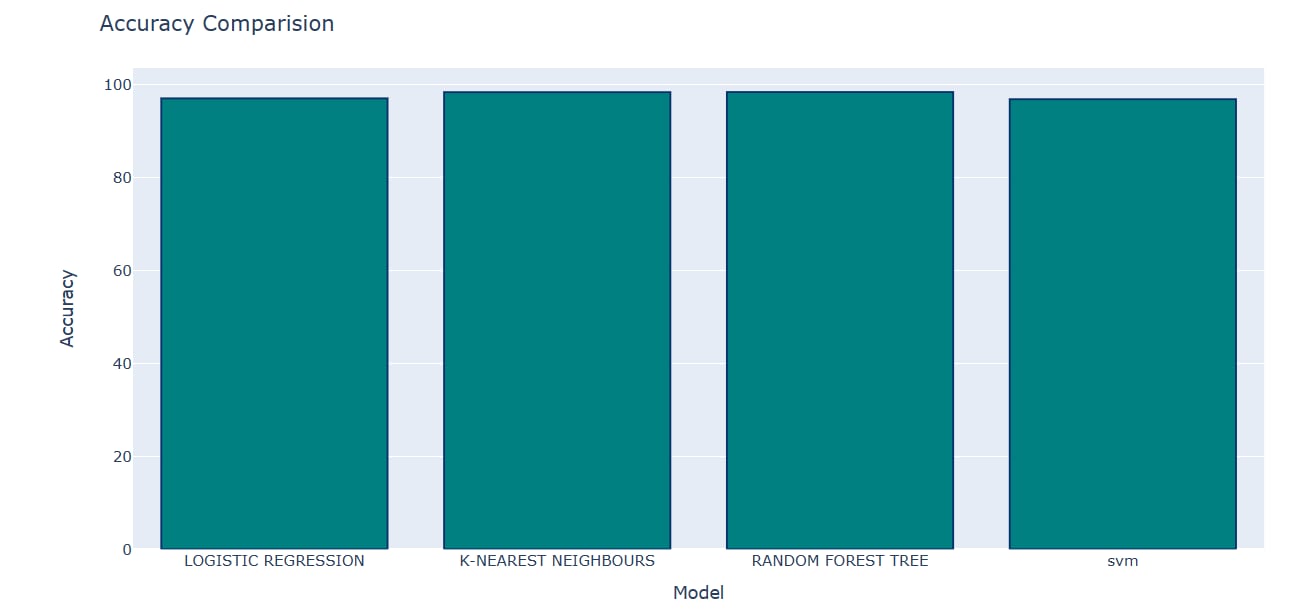
***Fig. 5. Histogramme montrant le nombre de patients ayant des problèmes respiratoires***

**RÉSULTAT ET DISCUSSION**

**Pour évaluer l'efficacité des algorithmes d'apprentissage automatique appliqués dans cette expérience, nous avons décidé d'adopter l'exactitude, l'erreur quadratique moyenne, la précision, le rappel et la mesure F qui sont largement utilisés dans des domaines tels que la recherche d'informations, l'apprentissage automatique et d'autres domaines impliquant classement binaire.**

****

***Fig. 6. Histogramme montrant le nombre de patients ayant de la fièvre***

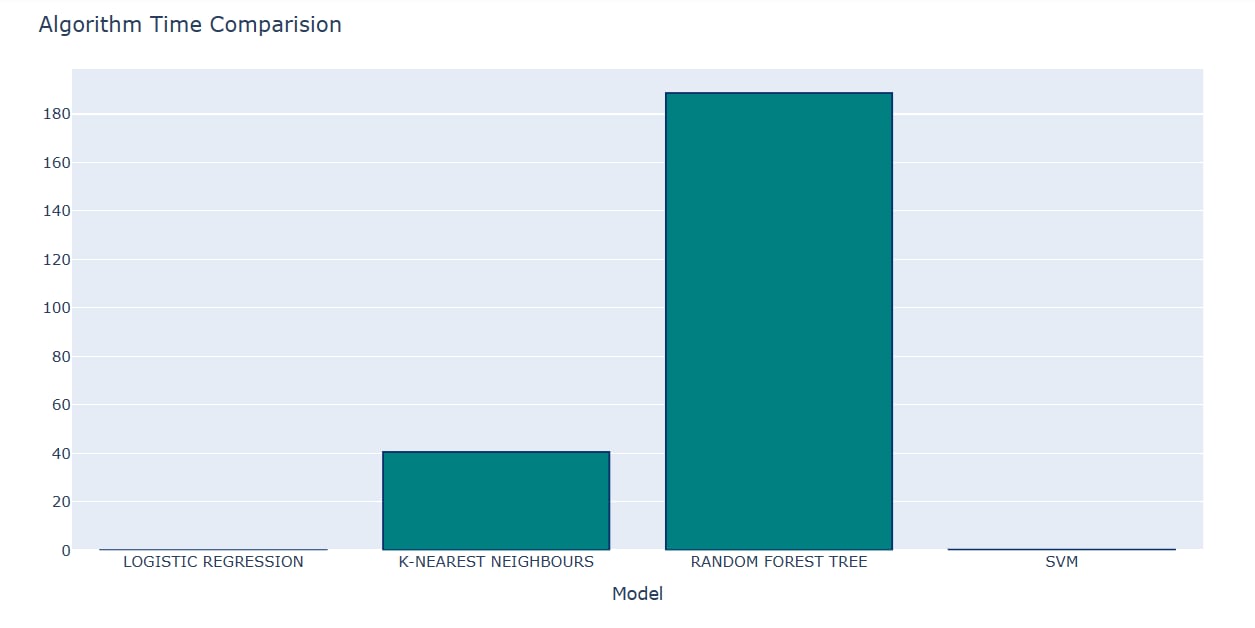
****

***Fig. 7. Analyse des algorithmes par leur précision***

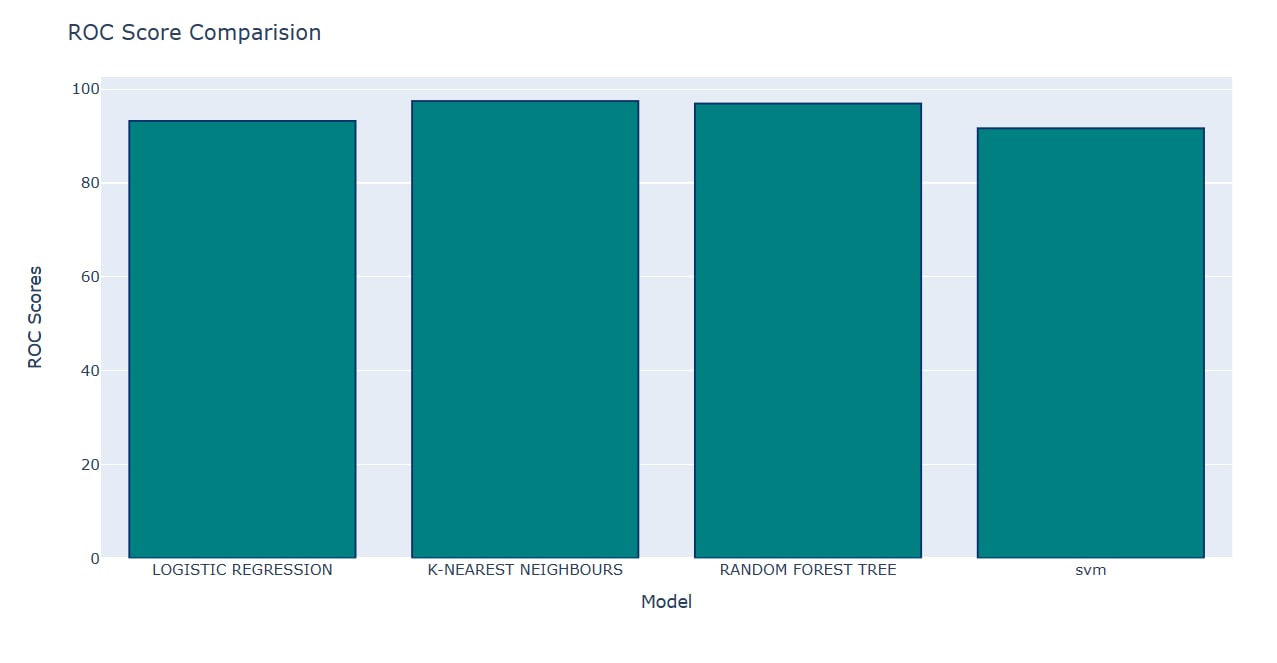
**La figure 7 montre que tous les algorithmes ont bien fonctionné pendant le processus de formation, grâce au fait que les hyperparamètres avaient été affinés. Cependant, il y avait des écarts mineurs dans la précision, comme l'indique la barre bleue.**

**La méthode Random Forest Tree présente la meilleure précision de tous les algorithmes, avec un score de 98,39 %. De plus, les scores R2, les erreurs quadratiques moyennes et les scores ROC sont tracés dans le graphique à barres en rouge, vert et violet respectivement. Avec une précision de 98,37 %, le KNN est le deuxième algorithme le plus acceptable à utiliser.**

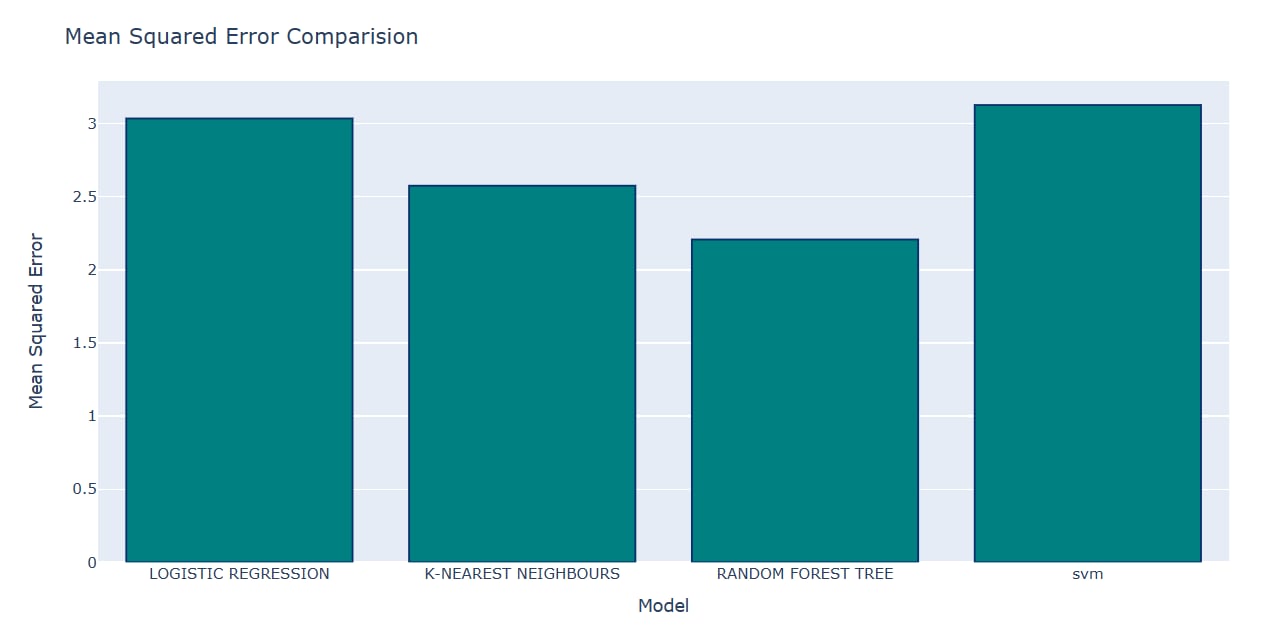
**L'erreur quadratique moyenne donne la différence entre la valeur attendue de la classe et la classe réelle. Cela nous permet de voir dans quelle mesure le modèle construit estime les étiquettes des échantillons par rapport aux valeurs d'origine de l'ensemble de données.**

****

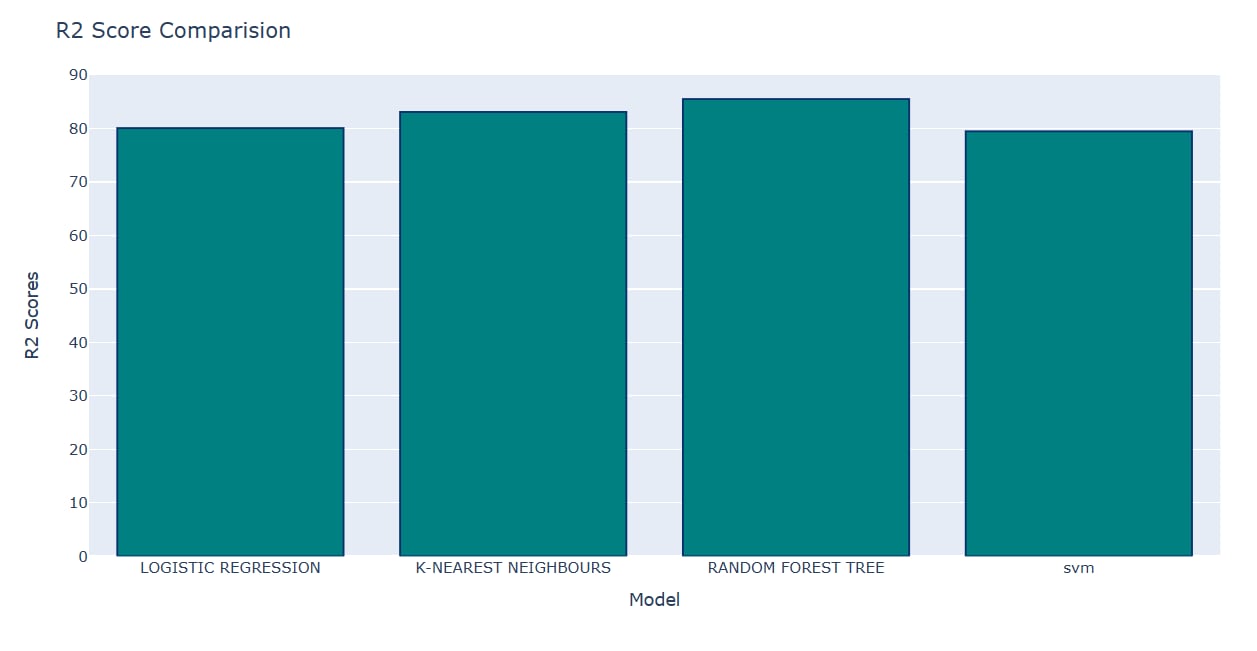
***Fig. 8. Analyse des algorithmes par temps nécessaire pour ajuster le modèle***

****

***Fig. 9. Analyse des algorithmes avec score R2***

****

***Fig. 10. Analyse des algorithmes avec MSE***

****

***Fig. 10. Analyse des algorithmes avec R2 Score***

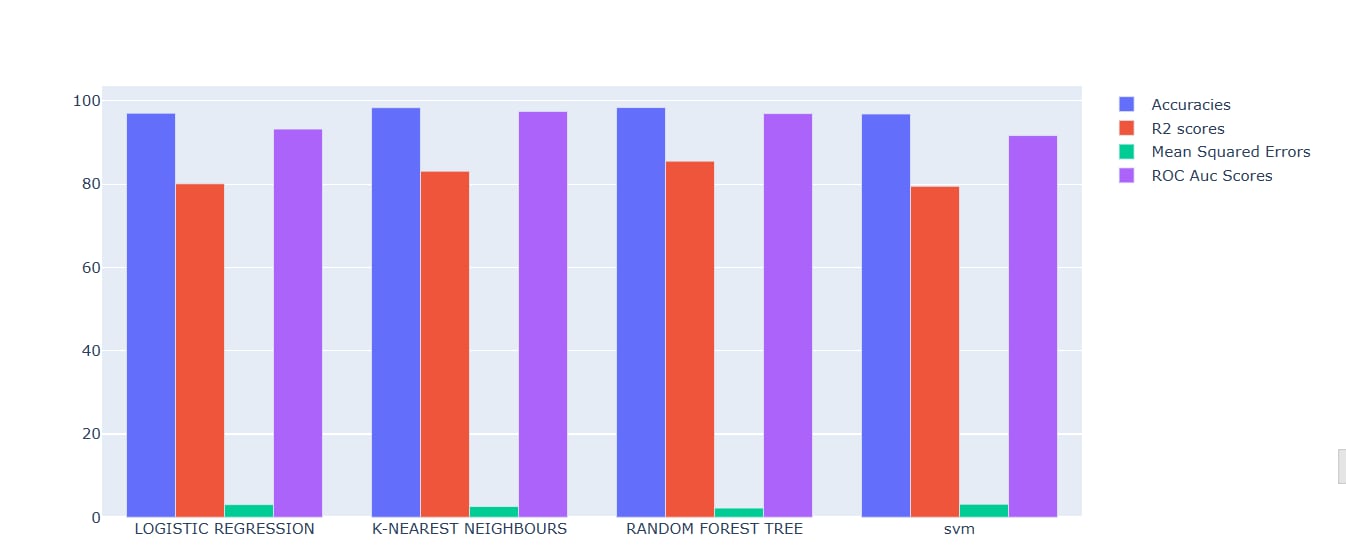
**Un modèle à faible MSE peut être considéré comme plus efficace par rapport à un modèle ayant un MSE élevé. Le MSE le plus bas a été atteint par l'algorithme Random Forest, avec 2,207 %, suivi de KNN avec 2,57 %.**

**Les scores R2 mesurent l’adéquation du modèle aux données.**

**Nous voyons donc ici que Random Forest a le score le plus élevé de 85,51 %, suivi de KNN avec 83,1 %. Enfin, comme le montre la figure 10, le temps nécessaire au développement du modèle a été pris en compte, car il s'agit d'un facteur important pour déterminer le meilleur algorithme pour notre modèle de prédiction. Sur la base des résultats, La régression logistique était l'algorithme le plus rapide avec 0,05 s à former un modèle de classificateur.**

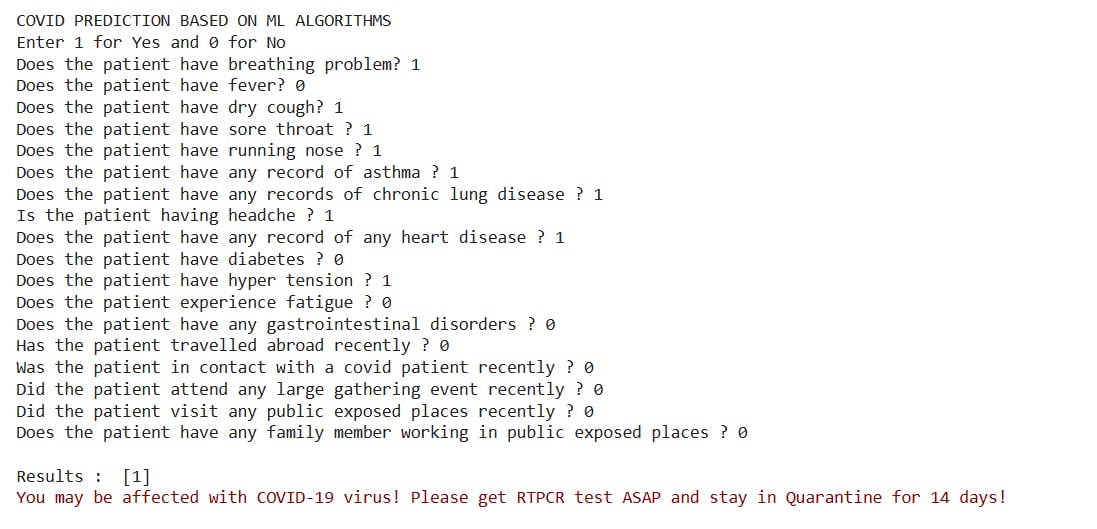
**CONCLUSION**

**Le but de ce travail était d'utiliser les trois techniques d'apprentissage automatique pour créer une présence COVID­19 modèle de prédiction. Les performances du modèle ont été évaluées dans une analyse comparative.**

****

***COMPARAISON DES MÉTRIQUES POUR KNN, RÉGRESSION LOGISTIQUE, FORÊT ALÉATOIRE et SVM***

**Les résultats montrent que le KNN classificateur avec un nombre de voisins à considérer égal à 2 est le meilleur algorithme d'apprentissage automatique, ayant une précision de 98,37 % et une erreur absolue moyenne de 0,026 compte tenu du temps d'exécution pour la formation. Par rapport à d'autres méthodes, le modèle prend un temps moyen mais donne une bonne précision.**

****

**Cette recherche peut être utilisée comme outil d'aide à la décision par les médecins, le modèle établi aidant à reconnaître la présence du COVID­19 chez une personne en fonction de ses symptômes. Les personnes qui souffrent du COVID­19**

**Les symptômes peuvent également l’utiliser pour évaluer s’ils seraient testés positif ou négatif pour le COVID­19. Le modèle qui a été développé ici peut être utilisé pour déployer une application avec les fonctionnalités suivantes :**

* **Les individus peuvent rapidement déterminer s'ils risquent de transmettre la COVID­19 en fonction de leurs symptômes.**
* **Les médecins peuvent utiliser ce test comme test principal évaluation sanitaire pour la détection du COVID.**
* **Aider les entreprises à limiter les interactions pysiques avec les clients susceptibles d'être infectés par la COVID­19**
* **Des informations supplémentaires ou des diagnostics provenant des dossiers hospitaliers, des personnes ayant contracté le virus, des survivants du COVID­19, des patients en cours d'évaluation ou de la direction peuvent tous être inclus pour des recherches futures. Un logiciel capable de prédire la gravité du COVID­19 peut en effet être déployé pour fournir des informations supplémentaires sur les mesures à prendre et les interventions à envisager.**