## **À propos de l'ensemble de données**

## **Contexte**

La maladie à coronavirus (COVID-19) est une maladie infectieuse causée par un nouveau coronavirus. La plupart des personnes infectées par le virus COVID-19 souffriront d’une maladie respiratoire légère à modérée et se rétabliront sans nécessiter de traitement particulier. Les personnes âgées et celles qui souffrent de problèmes médicaux sous-jacents tels que les maladies cardiovasculaires, le diabète, les maladies respiratoires chroniques et le cancer sont plus susceptibles de développer une maladie grave.  
Tout au long de la pandémie, l’un des principaux problèmes auxquels les prestataires de soins de santé ont été confrontés est le manque de ressources médicales et de plan approprié pour les distribuer efficacement. En ces temps difficiles, être capable de prédire le type de ressources dont une personne pourrait avoir besoin au moment où elle est testée positive ou même avant cela sera d’une aide immense pour les autorités, car elles seront en mesure de se procurer et d’organiser les ressources nécessaires pour sauver la vie de ce patient.

L'objectif principal de ce projet est de construire un modèle d'apprentissage automatique qui, compte tenu des symptômes, de l'état et des antécédents médicaux actuels d'un patient atteint de Covid-19, prédira si le patient présente un risque élevé ou non.

## **contenu**

L'ensemble de données a été fourni par le gouvernement mexicain [(lien)](https://datos.gob.mx/busca/dataset/informacion-referente-a-casos-covid-19-en-mexico) . Cet ensemble de données contient un nombre considérable d'informations anonymisées relatives aux patients, y compris les conditions préalables. L'ensemble de données brutes se compose de 21 caractéristiques uniques et de 1 048 576 patients uniques. ****Dans les caractéristiques booléennes, 1 signifie « oui » et 2 signifie « non ». Les valeurs telles que 97 et 99 sont des données manquantes**** .

* sexe : 1 pour femme et 2 pour homme.
* âge : du patient.
* classification : résultats du test covid. Les valeurs 1 à 3 signifient que le patient a été diagnostiqué avec covid à différents  
  degrés. 4 ou plus signifie que le patient n'est pas porteur du covid ou que le test n'est pas concluant.
* type de patient : type de soins que le patient a reçu dans l'unité. 1 pour le retour à domicile et 2 pour l'hospitalisation.
* pneumonie : que le patient présente ou non une inflammation des sacs aériens.
* grossesse : si la patiente est enceinte ou non.
* diabète : si le patient est diabétique ou non.
* BPCO : indique si le patient souffre ou non d'une bronchopneumopathie chronique obstructive.
* asthme : si le patient souffre d’asthme ou non.
* inmsupr : si le patient est immunodéprimé ou non.
* hypertension : si le patient souffre d’hypertension ou non.
* cardiovasculaire : si le patient souffre d’une maladie cardiaque ou vasculaire.
* maladie rénale chronique : si le patient souffre ou non d’une maladie rénale chronique.
* autre maladie : si le patient a une autre maladie ou non.
* obésité : si le patient est obèse ou non.
* tabac : si le patient est un consommateur de tabac.
* usmr : Indique si le patient a été traité par des unités médicales de premier, deuxième ou troisième niveau.
* unité médicale : type d’établissement du Système National de Santé qui a fourni les soins.
* intubé : si le patient était connecté au respirateur.
* icu : indique si le patient a été admis dans une unité de soins intensifs.
* date du décès : Si le patient est décédé, indiquez la date du décès, et 9999-99-99 sinon.

## **A propos du fichier**

Contient un grand nombre d'informations anonymisées relatives aux patients, y compris les conditions préalables. L'ensemble de données brutes comprend 21 caractéristiques différentes et 1 048 576 patients uniques. Dans les caractéristiques booléennes, 1 signifie « oui » et 2 signifie « non ». Les valeurs telles que 97 et 99 sont des données manquantes.

## **Lien du dataset**

https://www.kaggle.com/datasets/meirnizri/covid19-dataset