**Projet analyse SARS-COV-2**

*1. Définir un objectif mesurable*

- Objectif :

Prédire si une personne est infectée en fonction des données cliniques disponibles.

- Métrique :

Précision -> permet de réduire à maximum le taux de faux positif -> Score F1 de 50%

Recall -> permet de réduire à maximum le taux de faux négatif -> Recall de 70%

*2. EDA (Exploratory Data Analysis*

*Objectif :*

Comprendre au maximum les données dont on dispose pour définir une stratégie de modélisation.

Analyse de la forme :

Identification de la Target : SARS-COV-2 exam result

Nombre de lignes et de colonnes : 5644, 111

Types de variables :

Float64 : 70

Object : 37

Int64 : 4

Identification des valeurs manquantes :

Beaucoup de NaN, plus de la moitié des variables possèdent plus de 90% de NaN

On distingue 2 groupes de données : -> 76% Test Viraux -> 80% Taux Sanguins

*Analyse du fond :*

Visualisation de la Target (Histogramme/ Boxplot) :

10% de cas positifs

90% de cas négatifs

Compréhension des différentes variables (Internet) :

. Variables continues standardisées, skewed (asymétriques), test sanguin

. Age quantile : difficile d'interpréter ce graphique, clairement ces données ont été traitées, on pourrait penser 0—5, mais cela pourrait aussi être une transformation mathématique. On ne peut pas savoir car la personne qui a mis ce dataset ne le précise nul part. Mais ça n'est pas très important

. Variable qualitative : binaire (0,1), viral, rhinovirus qui semble très élevée

Visualisation des relations features— Target (Histogramme/ Boxplot) :

. Target / Blood : les taux de Monocytes, Platelets, Leukocytes semblent liés au covid-19 -> hypothèse à tester

. Target/Age : les individus de faible âge sont très peu contaminés ? -> attention on ne connait pas liage, et on ne sait pas de quand date le dataset (s'il s'agit des enfants on sait que les enfants sont touchés autant que les adultes). En revanche cette variable pourra être intéressante pour la comparer avec les résultats de tests sanguins

. Target / Viral : les doubles maladies sont très rares. Rhinovirus/ Entérovirus positif – covid-19 négatif ? -> hypothèse à tester ? mais il est possible que la région soit subie une épidémie de ce virus. De plus on peut très bien avoir 2 virus en même temps. Tout ça n'a aucun lien avec le covid-19

*Conclusions initiales :*

. Beaucoup de données manquantes (au mieux on garde 200/0 du dataset)

. 2 groupes de données intéressantes (viral, sanguin)

. Presque pas de variable "discriminante" pour distinguer les cas positifs/négatifs, ce qui nous permet de dire qu'il n'est pas vraiment approprié de vouloir prédire si un individu est atteint du Covid-19 en se basant sur ces simples tests sanguins. Mais ce n’est pas grave, il faut quand même poursuivre l'analyse pour essayer de voir ce qu'on peut apprendre.

. Donc maintenant quand même quelque chose de positif : on a pu identifier des variables intéressantes qui sont susceptibles de jouer un rôle non négligeable (monocytes etc.)

*Analyse plus détaillée*

• Relation Variables I Variables

• Blood data / Blood data : certaines variables sont très corrélées : +0.9

• Blood data / Age : très faible corrélation

• viral / viral : Influenza rapid test donne de mauvais résultat du a sa sensibilité, il faudra peut-être ne pas la considérer pour plus tard

• relation maladie / Blood data : les taux sanguins entre malades et covid-19 sont différents

• relation hospitalisation / est malade

• relation hospitalisation / Blood : intéressant dans le cas où on voudrait prédire dans quelle service un patient devrait aller

• NaN analyse

• Viral : 1350(92/8), Blood : 600(87/13), Both : 90

• Hypothèses nulle (HO) :

* Les individus atteints du covid-19 ont des taux de Leukocytes, Monocytes, Platelets significativement différents

• HO = Les taux moyens sont ÉGAUX chez les individus positifs et négatifs

Cette hypothèse nulle est rejetée

* Les individus atteints d'une quelconque maladie ont des taux significativement différents

*3. Preprocessing*

Objectif :

Transformer le data pour le mettre dans un format propice au Machine Learning

Preprocessing :

Création du Train set / Test Set

Élimination des NaN : dropna (), imputation, colonnes « vides »

Encodage

Suppression des outliers néfastes au modèle

Features Sélections

Features Engineering

Features Scaling

*4. Modeling*

Objectif :

Développer un modèle de Machine Learning qui réponde à l'objectif final.

Modeling :

Définir une fonction d'évaluation

Entrainement de différents modèles

Optimisation avec GridSearchCV

(Optionnel) Analyse des erreurs et retour au Preprocessing / EDA

Learning Curve et prise de décision