Checklist de base : Machine Learning

# Analyse de forme

Variable target : SARS-Cov-2 exam result

Lignes/colonnes : 5644/111

Types des variables : 74 quantitatives, 37 qualitatives

Analyse des valeurs manquantes :

* Deux groupes se distinguent :
  + ~ 76 % : Tests concernant d’autres types de virus
  + ~ 89 % : Analyse de paramètres sanguins

# Analyse de fond

Visualisation de la target :

* 10 % de positifs
* 90% de négatifs (très déséquilibré !)

Très important lors de l’exploration de données : TRACER les histogrammes des variables continues

Signification des variables :

* Variables standardisées, asymétriques, test sanguins
* Age quantile : difficile à interpréter, les données ont été traitées ou transformées mathématiquement
* Variables qualitatives : binaires (0, 1) detected/not detected, viral, Rhinovirus qui semble très élevé

Relation variables/target :

* Target/Blood : Idées à tester :
* Leucocyte
* Monocyte
* Platelets : plaquettes

🡪 Ces taux sont différents entre les personnes testées positives au Covid19 et celles testées négatives. Il faut voir si cela à un sens de considérer que cela est lié.

* Target/Age : Les individus à faible âge sont peu détectés contaminés ?? Attention, âge inconnu en réalité. De plus, les enfants sont autant contaminés que les adultes
* Target/Viral : double maladies rares, mais Rhinovirus/Entérovirus positif – Covid19 négatif 🡪 Hypothèse à tester. Il est cependant juste possible que la région d’où proviennent les données ai subi une épidémie de ce virus simultanément au Covid19. Probablement aucun lie

# Analyse plus détaillée

* Relation Variable/Variable :
  + Blood\_data / Blood\_data : certaine variables sont très corrélées (+0.9 à surveiller plus tard)
  + Blood\_data / Age : très faible corrélation
  + Viral / Viral : influeza rapid test donnee des résultats mauvais, peut-être faudra l’éliminer
  + Relation sickness / Blood\_data : Les taux sanguins entre malades « normaux » et covid19 sont différents (lymphocyte, hémoglobine et hématocrite)
  + Relation hospitalisation / is Sick :
  + Relation hospitalisation / Blood :
* NaN analyse : viral 1350 (92%/8%), blood 600 (87%/13%), précédemment : 90%

**Hypothèse nulle (H0) :**

* Les individus atteints du covid-19 ont des taux de Leucocyte, Monocyte et plaquettes significativement différents de ceux ne l’étant pas
  + H0 = Les taux moyens sont EGAUX chez les individus positifs et négatifs
* Les individus atteints d’une maladie quelconque ont des taux de significativement différents de ceux ne l’étant pas

**Conclusions initiales :**

* Beaucoup de données manquantes : 20% du dataset exploitable
* 2 groupes de données intéressantes : blood/viral
* Les tests sanguins ne permettent pas avec certitude de prédire les cas de Covid19
* Il va falloir remplir les valeurs manquantes, si on les supprime : 99 lignes au lieu de 5644
* Blood\_column : 600 valeurs, viral\_column : 1354 (plus intéressants, mais il faut conserver les deux, car les tests sanguins sont aussi utils)

Si on sélectionne les données virales, il y a le même taux de positifs/négatifs qu’auparavant ; pareillement avec les données sanguines, les taux sont même plus équilibrés !