**Projet GIT2 : Séminaire Technique**

Projet  :

L’objectif de ce travail est de comprendre au mieux n’importe quelles données et les actus pour le faire à bien.

Développer une première stratégie de modélisation des données.

1. Importer les packages essentielles à l’exploration des données
2. Importer votre dataFrame **« dataset.xlsx »**
3. Copier votre data set et expliquer pourquoi le faire ???
4. **Analyse de Forme :**
5. Trouver la variable Target
6. Donner le nombre de lignes et colonnes de votre dataFrame
7. Donner les types des variables : Afficher pour toutes les variables puis compter le nombre de types de variables

Faire un graphique en secteur

1. Analyser des valeurs manquantes

* Visualiser en entier votre toutes les colonnes
* Visualiser en entier tout le dataFrame en utilisant une image sur **« Matplotlib »** ou par **« Seaborn »**
* Analyser le graphique obtenu dans **« f »**
* Mesurer le pourcentage des valeurs manquantes que nous avons dans le dataFrame « **p**=sum de na de chaque colonnes/nombre de Ligne » où **p** est le pourcentage des valeurs manquantes. Analyser le résultat.
* Ensuite éliminer toutes les variables dont les valeurs manquantes est supérieur à 90%.
* En suite chercher à éliminer la colonne **« Patient ID »**

1. Conclusion sur la forme des données
2. **Analyse de Fond :**
3. Visualisation de la Target

* Compter le nombre de case **« positive »** et **« négative ». Que concluez-vous ???**
* Calculer le pourcentage de ces deux cas.

1. Signification des variables

* Tracer les histogrammes des variables continues : vous pouvez utiliser la boucle « **for**»
* Conclusion
* Tracer la courbe de la variable « **Patient age quantile** », que remarquez-vous ???

1. Visualiser les variables catégorielles de type objet (Variable qualitative)

* Vérifier les différentes catégories qui résident dans chaque variable en définissant la marge entre les variables et leurs catégories.
* En utilisant Matplotlib, tracer le graphe que vous interpréteriez chaque catégorie puis données une conclusion.

1. **Relation Variables/Target**
2. Création de sous-ensembles positifs et négatifs en faisant le booléen indexing
3. Créer en suite les ensembles Blood et Viral en utilisant le pourcentage des valeurs manquantes

* Relation Target et Blood en utilisant un diagramme de distribution
* Relation Target et age en utilisant un diagramme de distribution ou le package « countplot »

1. **Relation Target/Viral**
2. Comparer les deux virus suivant « SARS-Cov-2 exam result » et « Influenza A ». Faite une analyse des resultats.
3. Faite une image de comparaisons des variables.
4. **Analyse un peu plus avancée**