Fichier Descriptif du Projet – Analyse Exploratoire et Prédiction des Cas COVID-19 à partir de Données Médicales Réelles

Titre du Projet

Analyse Exploratoire des Données Médicales liées au COVID-19

Membres du Groupe

- Abdeljalil Idalahaj
- Ibtissam El asslouj

Contexte :

Dans ce projet, nous avons mené une analyse exploratoire des données (EDA) **approfondie sur un ensemble de données médicales réelles liées au COVID-19, puis développé un** modèle de machine learning performant **pour prédire le statut COVID-19 des patients à partir de leurs caractéristiques cliniques, biologiques et virales.**

🔬 Partie 1 : Analyse Exploratoire des Données (EDA)

- Nettoyage et préparation des données : traitement des valeurs manquantes, codage des variables catégorielles, standardisation.
- Identification et visualisation des **valeurs aberrantes** (outliers) .
- Étude des variables les plus corrélées à la positivité COVID-19, notamment via des tests statistiques comme le **t-test**.
- Visualisation avec **histogrammes, pairplots et courbes de densité** pour comparer les distributions entre patients malades et non malades.
- Création de nouvelles variables telles que le **statut d'hospitalisation** pour affiner le profil des patients.

🤖 Partie 2 : Modélisation avec le Machine Learning

- Utilisation de modèles supervisés tels que **RandomForestClassifier** et **Pipeline combinant PolynomialFeatures, SelectKBest et RandomForest**.
- Séparation du dataset en ensemble d'entraînement et de test (80/20 stratifié).
- **Sélection des meilleures variables prédictives** avec `SelectKBest` basé sur le test ANOVA (f_classif).
- **Évaluation des performances** via :
  - Matrice de confusion,
  - Rapport de classification (précision, rappel, F1-score),
  - **Courbe d'apprentissage** pour visualiser l'évolution des performances.

📈 Le modèle final (pipeline) a atteint une **précision d'environ 96%**, ce qui montre un fort pouvoir de généralisation et une bonne capacité de détection des cas positifs.

Outils Utilisés

Python (Pandas, Seaborn, Matplotlib, SciPy, Scikit-learn), Jupyter Notebook.

Lien de la Vidéo de Présentation

https://drive.google.com/file/d/17JxTA9rCVGCjPGTM8hsp54O38qRHIK2V/view?usp=drive_link

Conclusion

Ce projet nous a permis de combiner une **analyse exploratoire approfondie des données médicales** avec une approche de **modélisation prédictive en machine learning**, dans un contexte de santé publique particulièrement critique : la détection et la compréhension des cas de COVID-19.

L'analyse exploratoire nous a aidés à repérer les **variables les plus pertinentes** (comme les marqueurs biologiques, les données virales, et les niveaux d'hospitalisation) pour différencier les patients positifs des négatifs. Des visualisations claires et des tests statistiques nous ont permis d'**interpréter les comportements des données**, de détecter les outliers, et de révéler les profils à risque.

Dans un second temps, nous avons développé un **modèle prédictif performant** basé sur un Random Forest, intégré dans un pipeline intelligent avec transformation polynomiale et sélection de variables. Ce modèle a atteint une **précision de plus de 96 %**, validée par des courbes d'apprentissage et un rapport de classification, ce qui démontre sa robustesse. Ce travail met en évidence l'intérêt de l'analyse de données dans le domaine médical, et ouvre la voie à des **applications concrètes** comme la **détection automatisée** de patients à risque, l'assistance au diagnostic, ou encore la **surveillance épidémiologique intelligente**.Ce projet constitue ainsi une base solide pour de futurs travaux, intégrant plus de données, ou visant à déployer un outil prédictif utilisable en milieu hospitalier.