**PREDICTION DES MALADIES CARDIAQUES A L'AIDE DE VARIABLES CLINIQUES**

**Introduction**

Les maladies cardiovasculaires représentent l'une des principales causes de mortalité dans le monde. Une détection précoce des patients à risque permettrait de mettre en place des traitements préventifs et de réduire la mortalité. Ce projet vise à construire un modèle prédictif permettant de détecter la présence de maladie cardiaque à partir de variables cliniques recueillies lors d'examens médicaux standards.

**Description du projet**

* Présentation du data pipeline sous forme de schéma
* Présentation du dataSet
* Présentation des traitements effectués sur la data
* Pourquoi le choix d’un tel algo
* Bonne interprétation des résultats
* Synthèse sur le projet

**Objectifs du projet**

* Comprendre et explorer les données cliniques fournies
* Nettoyer et préparer les données pour la modélisation
* Appliquer des modèles de machine learning pour prédire la variable cible "num"
* Évaluer la performance des modèles avec des métriques telles courbe ROC

1. **PRESENTATION DU DATA PIPELINE SOUS FORME DE SCHEMA**

**A diagram of a model

AI-generated content may be incorrect.**

1. **PRESENTATION DU DATASET**

Le dataset utilisé est issu de la base UCI Heart Disease. Il comprend les attributs suivants :

|  |  |
| --- | --- |
| **age** | Âge du patient |
| **sex** | Sexe (0 = femme, 1 = homme) |
| **cp** | Type de douleur thoracique (0 à 3) |
| **trestbps** | Pression artérielle au repos (mm Hg) |
| **chol** | Taux de cholestérol (mg/dl) |
| **fbs** | Glycémie à jeun élevée (1 = oui, 0 = non) |
| **restecg** | Résultat de l’ECG au repos |
| **thalach** | Fréquence cardiaque maximale atteinte |
| **exang** | Présence d’angine de poitrine à l’effort (1 = oui, 0 = non) |
| **oldpeak** | Dépression du segment ST après effort |
| **slope** | Forme de la courbe ST à l'effort |
| **ca** | Nombre de vaisseaux colorés à l’examen |
| **thal** | Résultat du test thallium (fonction cardiaque) |
| **num** | Présence de maladie cardiaque (0 = non, 1 = oui) |

**A screenshot of a computer

AI-generated content may be incorrect.**

1. **PRESENTATION DES TRAITEMENTS EFFECTUES SUR LA DATA**
2. **Nettoyage des données et imputation :**

**A black rectangular object with a black border

AI-generated content may be incorrect.**

Le dataset contient quelques valeurs manquantes :

ca : 4 valeurs manquantes

    thal : 2 valeurs manquantes

**Imputation de la médiane**

On utilisera l’imputation de la médiane, qui consiste à remplacer les valeurs manquantes d’une variable numérique par la médiane ce cette variable

A screenshot of a computer

AI-generated content may be incorrect.

1. **Vérification des types de variables**

**A screenshot of a computer

AI-generated content may be incorrect.**

1. **Encodage et standardisation des variables catégorielles avec One-Hot Encoding**

* L’encodage ici consiste à transformer les colonnes contenant du texte ou des catégories en valeurs numériques,
* Normaliser / standardiser permetra de mettre les variables numériques à la même échelle, ce qui est important pour les modèles sensibles aux distances ou aux gradients

A screenshot of a computer

AI-generated content may be incorrect.

1. **CHOIX DE L’ALGORITME**

Dans ce projet, nous avons choisi la régression logistique comme modèle principal pour sa simplicité, son efficacité dans la classification binaire, et surtout sa capacité à fournir des probabilités interprétables, ce qui est essentiel en contexte médical. Nous avons également testé Random Forest pour améliorer les performances prédictives, notamment en capturant des relations non linéaires entre les variables cliniques.

**régression logistique**

A screenshot of a computer program

AI-generated content may be incorrect.

A screen shot of a computer

AI-generated content may be incorrect.

**Random forest**

A screen shot of a computer

AI-generated content may be incorrect.

A screen shot of a graph

AI-generated content may be incorrect.

Cette courbe montre que la **Random Forest** surpasse légèrement la **régression logistique** et a une meilleure capacité à distinguer les patients à risque

1. **BONNE INTERPRETATION DES RESULTATS**
2. **Matrice de corrélation**

La matrice de corrélation : permet d'identifier les relations entre les différentes variables et leur influence potentielle sur la présence de maladie cardiaque

**A screenshot of a computer

AI-generated content may be incorrect.**

**Corrélation avec la variable cible num (présence de maladie cardiaque)**

* **cp (type de douleur thoracique) : +0.43**
  + Forte corrélation positive. Certains types de douleur (angine typique) sont fortement associés à une maladie cardiaque.
* **thalach (fréquence cardiaque max) : -0.42**
  + Corrélation négative : plus la fréquence cardiaque max est basse, plus le risque de maladie cardiaque est élevé.
* **slope (pente du segment ST) : +0.34**
  + Une pente anormale (flat, descendante) est souvent un signe de pathologie.
* **exang (angine d'effort) : -0.44**
  + Les personnes ayant une angine d'effort sont plus susceptibles d’avoir une maladie cardiaque.
* **oldpeak (dépression ST) : +0.43**
  + Plus la dépression est importante, plus le risque est élevé.
* **ca (nombre de vaisseaux obstrués) : -0.38**
  + Corrélation négative : plus il y a d’obstructions, plus la probabilité de maladie augmente.
* **thal (résultat du test thallium) : +0.51**
  + La plus forte corrélation avec num dans cette matrice. Un test thallium anormal est un indicateur majeur de maladie cardiaque.

**Corrélation entre variables explicatives**

* trestbps et age : +0.28 → les personnes plus âgées ont une pression artérielle légèrement plus élevée.
* chol et age : +0.21 → légère tendance à un cholestérol plus élevé avec l’âge.
* Faibles corrélations entre la majorité des autres variables, donc faible multicolinéarité.

1. **Histogramme des variables numériques**

**A screenshot of a graph

AI-generated content may be incorrect.A screenshot of a graph

AI-generated content may be incorrect.**

1. **Pairplot coloré selon la classe cible NUM qui matérialise la présence des maladie cardiaques**

**A screenshot of a computer screen

AI-generated content may be incorrect.**

**A collage of a graph

AI-generated content may be incorrect.**

1. **Distribution des variables**

**A screenshot of a computer

AI-generated content may be incorrect.**

**A screenshot of a computer

AI-generated content may be incorrect.A screenshot of a computer

AI-generated content may be incorrect.A black rectangle with white text

AI-generated content may be incorrect.**

Histogrammes et pairplots : Ont aidé à visualiser la distribution des variables et à détecter des patterns ou des anomalies

**A screen shot of a computer

AI-generated content may be incorrect.A screenshot of a computer

AI-generated content may be incorrect.**

L’analyse des importances des variables issues du modèle Random Forest met en évidence plusieurs facteurs cliniques déterminants dans la prédiction de la maladie cardiaque :

* **Thalach** : Une fréquence cardiaque maximale élevée lors de l’effort est généralement le signe d’un cœur en bonne santé. À l’inverse, une valeur basse peut refléter une incapacité du cœur à réagir à l’effort, ce qui augmente le risque de maladie cardiaque.
* **ca** : Cette variable est fortement corrélée à la présence de maladies coronariennes. Plus le nombre de vaisseaux touchés est élevé, plus le risque est important.
* **Oldpeak** : Elle indique une altération de l’activité électrique du cœur après un effort. Une dépression significative est souvent associée à un manque d’oxygénation du muscle cardiaque, typique d’une pathologie cardiaque.
* **Cp** :Certains types de douleur (comme la douleur angineuse typique) sont plus évocateurs de maladies cardiaques que d’autres. Cette variable catégorielle permet de classer la douleur selon sa nature et son intensité.
* **Exang** :Un résultat positif (valeur = 1) est fortement suggestif d’un dysfonctionnement cardiaque. Cette variable renforce donc la prédiction du modèle.

**Conclusion**

L’analyse montre que la préparation minutieuse des données et l'utilisation de techniques avancées de machine learning permettent de construire des modèles prédictifs robustes et utiles pour la détection précoce des maladies cardiaques.

Les résultats obtenus confirment la cohérence médicale du modèle, qui identifie des variables cliniques connues comme étant des facteurs de risque majeurs.

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |