Introduction

Dans les dix dernières années, les maladies cardiovasculaires (MCV) sont la principale cause de décès au niveau mondiale. 19.8 millions de personnes sont mortes de MCV représentant environ 32% de tous les décès dans le monde (OMS lien). Plus de ¾ de ces décès surviennent dans les pays à revenu faible ou intermédiaire. En Haïti par exemple, le MSPP classe ces maladies comme 6^{e} cause la plus fréquente pouvant causer la mort. Est-il possible de prédire et d'éviter ces morts via des solutions en data science et IA ? C'est cette question que nous proposons de répondre via ce projet.

Dans ce projet nous voulons construire un modèle pouvant prédire la variable *y* : heart-disease.

Si nous arrivons à prédire ce variable grâce à des données concernant l'état de santé d'un individu, nous pouvons alors déployer cela dans des centres de soins primaires pour diminuer drastiquement les morts pour cause de maladie cardiaque. Sachant que les politiques de prévention ont un cout moins élevé que celles de la prise en charge cela aiderait les décideurs en santé à proposer des solutions ayant un impact réel sur la vie des gens.

Ce rapport est divisé en plusieurs sections :

- Revue des données (Analyse exploratoire des données) ;
- Méthodologie et Modélisation ;
- Résultat et Evaluation ;
- Conclusions et Perspectives ;

Revue des données et Préparation

Les données utilisées dans ce projet proviennent de UC Irvine Machine Learning Repository. Il se base sur le secteur de la santé en particulier la prédiction de maladie cardiaque. Ce jeu de données à 14 variables et 270 observations.

Les données, leurs types et explication sur celles qui sont plus particulières :

1. age (real)

2. sex (binary)

3. chest pain type (4 values) (Nominal)

4. resting blood pressure (real)

Tension arterielle au repos. Considere comme normal lorsqu'elle est inferieur a 120 par 80 (Pression systolique par la tension diastolique respectivement) Dans ce cas la tension systolique est rapportée.

5. serum cholestoral in mg/dl (real)

Quantite de cholesterol presente dans le sang. Lorsque cette quantite est superieur a 240 mg/dl, elle est considere comme eleve et represente un risque pour le développement de maladies cardiovasculaire

6. fasting blood sugar > 120 mg/dl

7. resting electrocardiographic results (values 0,1,2)

8. maximum heart rate achieved

9. exercise induced angina

10. oldpeak = ST depression induced by exercise relative to rest

11. the slope of the peak exercise ST segment

12. number of major vessels (0-3) colored by flourosopy

13. thal: 3 = normal; 6 = fixed defect; 7 = reversable defect

14. heart-disease: 1=absence; 2=presence

EDA

Pour l'exploration des données on a combiné un ensemble d'analyse univariée et bivariée afin de déceler les tendances de notre jeu de données et d'entrevoir les problèmes potentiels.

Présentation des analyses pertinentes :

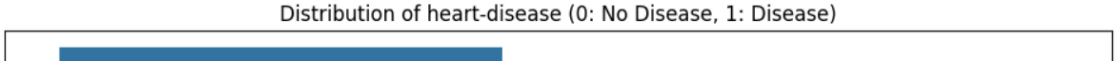
Avant tout on s'est assuré de l'existence de données manquantes et d'un visuel sur les 10 premières lignes de de la base de données (*Tableau 1*). Il faut toutefois mentionner que des traitements ont été faits sur la variable cible en transformant l'absence de maladies par 0 et la présence de cette dernière par 1.

Tableau 1. Affichage des 10 premières lignes du dataset

```
df.head(10)
```

	age	sex	chest-pain	rest-bp	serum-chol	fasting-blood-sugar	electrocardiographic	max-heart-rate	angina	oldpeak	slope	major-vessels	thal	heart-disease
0	70.0	1.0	4.0	130.0	322.0	0.0	2.0	109.0	0.0	2.4	2.0	3.0	3.0	1
1	67.0	0.0	3.0	115.0	564.0	0.0	2.0	160.0	0.0	1.6	2.0	0.0	7.0	0
2	57.0	1.0	2.0	124.0	261.0	0.0	0.0	141.0	0.0	0.3	1.0	0.0	7.0	1
3	64.0	1.0	4.0	128.0	263.0	0.0	0.0	105.0	1.0	0.2	2.0	1.0	7.0	0
4	74.0	0.0	2.0	120.0	269.0	0.0	2.0	121.0	1.0	0.2	1.0	1.0	3.0	0
5	65.0	1.0	4.0	120.0	177.0	0.0	0.0	140.0	0.0	0.4	1.0	0.0	7.0	0
6	56.0	1.0	3.0	130.0	256.0	1.0	2.0	142.0	1.0	0.6	2.0	1.0	6.0	1
7	59.0	1.0	4.0	110.0	239.0	0.0	2.0	142.0	1.0	1.2	2.0	1.0	7.0	1
8	60.0	1.0	4.0	140.0	293.0	0.0	2.0	170.0	0.0	1.2	2.0	2.0	7.0	1
9	63.0	0.0	4.0	150.0	407.0	0.0	2.0	154.0	0.0	4.0	2.0	3.0	7.0	1

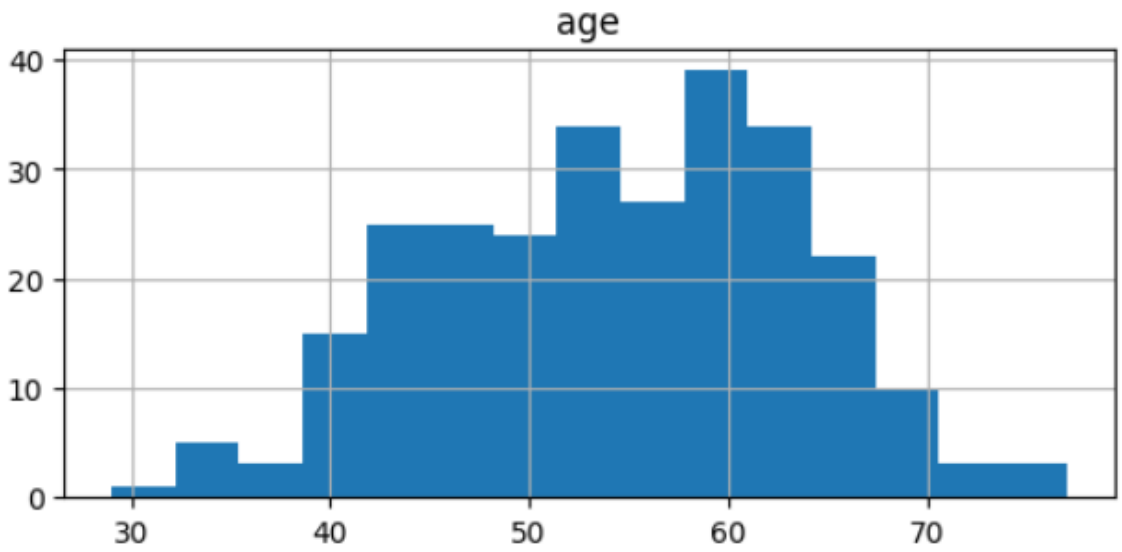
Source: Fichier jupyter notebook, Capstone Project



- Distribution de l'âge

Les individus de notre jeu de données se répartissent entre 29 et 77 ans. On constate que le plus grand nombre d'individus sont dans leur soixantaine. (*Graphique 1*)

Graphique 1. Distribution de l'âge des individus

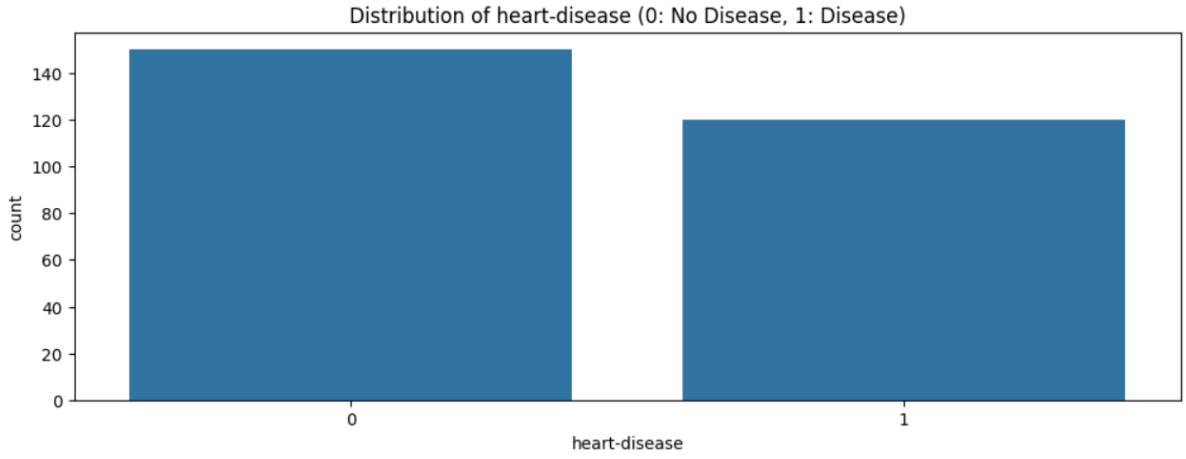


Source: Fichier jupyter notebook, Capstone Project

- Distribution des personnes malades

Dans le jeu de données nous pouvons identifier que plus 55% des individus n'ont pas de maladies cardiaques. (*Graphique 2*)

Graphique 2. Distribution des cas de malades et de non malades

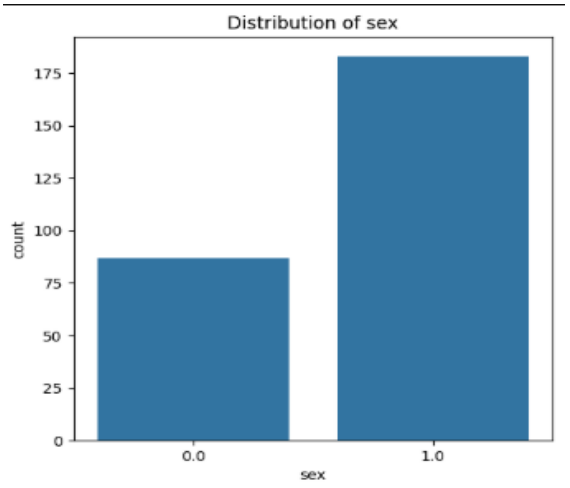


Source: Fichier jupyter notebook, Capstone Project

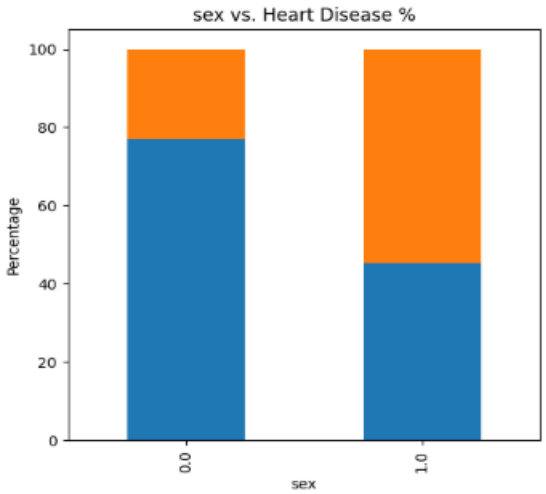
- Distribution des individus par sexe

Nous constatons que la majorité des personnes présentent dans ce jeu de données correspond au sexe 1 dépassant ainsi plus de 175 unités. *(Graphique 4*) Cette variable sera importante afin de déterminer quel genre est le plus affecté par la maladie. En ce sens nous pouvons voir que les personnes de sexe 1 sont les plus affectées (*Graphique 5*)

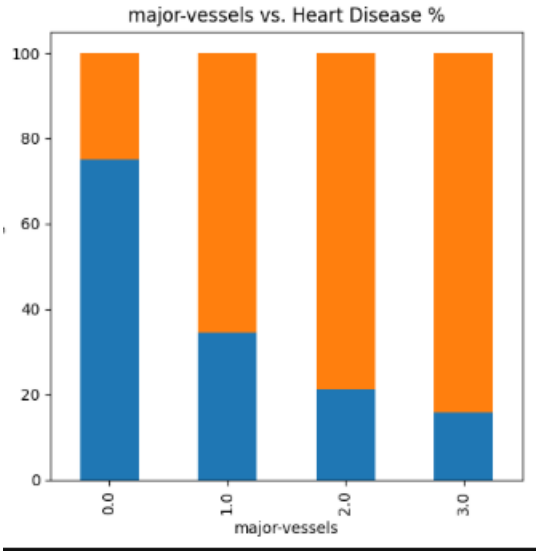
Grap. 4. Distribution des individus par sexe

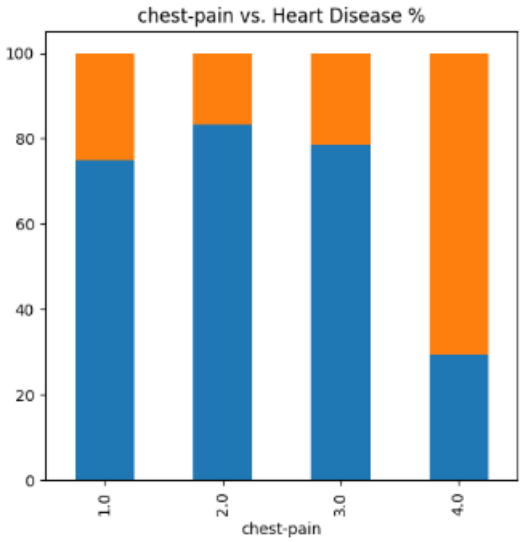


Grap. 5. Répartition sexe en fonction du statut



Source: Fichier jupyter notebook, Capstone Project





Méthodologie et Modélisation

Ce projet a pour but de prédire la variables *heart disease* en fonction des autres variables du jeu de données. Pour cela, nous allons utiliser plusieurs modèles permettant de prédire cette variable. Après la prédiction nous aurons à vérifier les métriques les plus pertinents (recall surtout) pour décider lequel des modèles retenir.

Les deux modèles proposés sont la Régression Logistique et le Random Forest.

Architecture du Modèle 1 (Régression Logistique)

Tout d'abord nos données sont divisées en deux classes, les données numériques et les données catégorielles. On a utilisé les variables *categorical_features* et *numerical_features* pour stocker dans des listes le nom des différentes variables.

Pour la transformation et le traitement de l'ensemble de ces données, l'utilisation d'un pipeline a été fait. Deux éléments intègrent notre preprocessor un *StandardScaler* pour la normalisation de nos données numériques et un *OneHotEncoder* pour la transformation des données catégorielles pour la suite

Architecture du Modèle 2 (Random Forest)

De même que la Régression Logistique, Le modèle de Random Forest utilise les données disponibles sous deux formats numériques et catégorielles.

Résultats et Évaluation

L'examen des différents résultats associe aux différents modèles est la suivante : Pour le 1er Modèle (Régression Logistique), la matrice de confusion indique la répartition des prédictions sur les 81 échantillons de test.

Vrais Positifs (TP) : 34 (La classe 1 a été correctement prédite 34 fois).

Vrais Négatifs (TN) : 39 (La classe 0 a été correctement prédite 39 fois).

Faux Positifs (FP) : 6 (La classe 0 a été prédite à tort pour 6 échantillons de la classe 1).

Faux Négatifs (FN) : 2 (La classe 1 a été prédite à tort pour 2 échantillons de la classe 0).

Le modèle présente une très bonne performance globale : Exactitude de 90% et AUC Score de 0.9364. Il est particulièrement précis pour identifier la Classe 0 (0.95) et affiche un excellent rappel pour la Classe 1 (0.94), ce qui signifie qu'il est très performant pour identifier correctement les cas de maladies.

X. Résultat lie au modèle de Régression Logistique

```
 Modele Initail : Performance de la Regression Logistique
Matrice de Confusion:
[[39  6]
 [ 2 34]]

 Rapport de classification:
              precision    recall  f1-score   support

           0       0.95      0.87      0.91        45
           1       0.85      0.94      0.89        36

    accuracy                           0.90        81
   macro avg       0.90      0.91      0.90        81
weighted avg       0.91      0.90      0.90        81

ROC AUC Score: 0.9364
```

Le second Modèle évalué est Random Forest. La matrice de confusion indique la répartition des prédictions sur les 81 échantillons de test. Vu que le Random Forest est sensible aux hyperparamètres on a pris le soin d'optimiser le modèle afin d'améliorer sa performance globale.

Vrais Positifs (TP) : 31 (La classe 1 a été correctement prédite 31 fois).

Vrais Négatifs (TN) : 38 (La classe 0 a été correctement prédite 38 fois).

Faux Positifs (FP) : 7 (La classe 0 a été prédite à tort pour 7 échantillons de la classe 1).

Faux Négatifs (FN) : 5 (La classe 1 a été prédite à tort pour 5 échantillons de la classe 0).

```
Matrice de Confusion:
[[38  7]
 [ 5 31]]

 Rapport de Classification:
              precision    recall  f1-score   support

           0       0.88      0.84      0.86        45
           1       0.82      0.86      0.84        36

    accuracy                           0.85        81
   macro avg       0.85      0.85      0.85        81
weighted avg       0.85      0.85      0.85        81

ROC AUC Score: 0.9299
```

En se basant sur les performances statistiques globales, la Régression Logistique l'emporte sur tous les indicateurs clés :

Meilleure Exactitude (Accuracy) : Avec 90% contre 85%, le modèle Logistique fait 5% de prédictions correctes en plus sur l'ensemble de test.

Le rappel (recall) est aussi du côté du modèle de Régression Logistique. Pour la classe 0(Absence de maladie) il s'élève à 87 et la classe 1 à 94% contre 84% et 86% pour le modèle de Random Forest

Le score AUC ROC qui mesure la capacité du modèle à distinguer les classes est aussi en faveur de la Régression Logistique (0.9364 vs 0.9170). Avec un AUC plus proche de 1 le modèle Logistique indique une meilleure performance, confirmant qu'il est supérieur pour séparer les observations.

En raison des résultats de l'ensemble des modèles testes, nous avons fait le choix du modèle de la Régression Logistique pour la prédiction de la variable heart-disease.

Conclusion

Ce projet a atteint son objectif de développer un modèle prédictif pour la présence ou l'absence de maladie cardiaque (*heart-disease*) chez les individus, dans le but d'assister les centres de soins primaires et de réduire les décès liés aux maladies cardiovasculaires.

Les résultats de l'évaluation ont montré que la Régression Logistique surpasse le Random Forest sur tous les indicateurs clés. Le modèle de Régression Logistique a atteint une Exactitude (Accuracy) de 90% et un score AUC ROC de 0.9364, ce qui témoigne de sa capacité supérieure à distinguer les classes. Plus important encore pour un problème de santé, son Rappel (Recall) pour la Classe 1 (maladie) s'élève à 94%, indiquant une excellente performance pour identifier correctement les cas de maladies. En conséquence, le modèle de Régression Logistique a été retenu pour la prédiction de la variable *heart-disease.*

Ces maladies cardiovasculaires qui constituent la principale cause de décès au niveau mondial, avec une incidence particulièrement élevée dans les pays à faible ou moyen revenu peuvent être contrées par des solutions en Data Science et IA pour la prédiction et la prise en charge avant toute aggravation de la situation des patients.

Recommandations

1- Mettre en place des programmes de sensibilisation sur la prévention nutritionnelle et les habitudes de vie

La visibilité des artères coronaires peut être affecté par les habitudes de consommation et de vie des individus. Une alimentation riche en sel, en graisses et le manque d'activité physique peuvent augmenter le développement de l'athérosclérose.

Donc en lançant des programmes d'éducation communautaire sur la nutrition cardiosaine (consommation locale : fruits, légumes, réduction du sel et des fritures) et l'activité physique, nous réduirons la prévalence des facteurs de risque, donc le nombre de cas détectés tardivement.

2- Renforcer le dépistage communautaire à bas coût

Les données montrent que plusieurs facteurs simples (pression artérielle, cholestérol, fréquence cardiaque, âge) sont de bons indicateurs précoces.

Mettre en place des campagnes de dépistage itinérantes de proximité pour mesurer ces paramètres de base gratuitement ou à faible coût. Cela permettra de détecter tôt les personnes à risque dans les zones rurales où les hôpitaux sont rares.

3- Créer une base de données nationale des maladies cardiovasculaires

L'absence de données locales limite les politiques publiques. La création d'un système national de collecte de données (âges, tension, cholestérol) dans les hôpitaux et centres communautaires nous

aidera à ajuster les interventions, d’entraîner de meilleurs modèles prédictifs haïtiens et d’orienter le financement vers les zones les plus touchées.