Institution: Akademi (Bootcamp)

Programme : Data Science & AI

Projet Capstone

Analyse et prédiction des maladies cardiovasculaires

Instructor name: JEROME Wedter

& LAGUERRE Geovany

Students:

- ✓ **DEBREUS Monitès** (spécialiste en Génie Rural & Obs. Météo)
- ✓ **PIERRE Cassio** (spécialiste en Economie et Développement Rural & MEAL)

Scheduled project review date/time: Octobre 2025

Table des matières

Liste des Figures

Liste des Tableaux

1. Introduction

Les maladies cardiovasculaires (MCV) figurent parmi les affections non transmissibles les plus meurtrières à l'échelle mondiale. Selon l'organisation mondiale de la santé, en 2019, elles ont causé près de 17,9 millions de décès. A ce jour, elles sont responsables d'environ 32 % de tous les décès dans le monde. Le fardeau des MCV ne cesse de croître : de 1990 à 2021, le nombre de décès cardiovasculaires est passé d'environ 12,1 millions à 20,5 millions selon les estimations de ***World Heart Federation***, soit une hausse d'environ 60 %.

Globalement, l'institution "***American heart association***" estime qu'en 2021 environ 612 millions de personnes vivaient avec une maladie cardiovasculaire. En Amérique, certains pays sont particulièrement affectés. Haïti se distingue par un taux de mortalité par MCV très élevé : 427,7 décès pour 100 000 habitants, le plus élevé de la région selon les estimations de la ***Pan American Health Organization.***

Dans le contexte haïtien, les études disponibles montrent déjà des signes alarmants : une enquête urbaine à Port-au-Prince a estimé une prévalence ajustée des MCV à ~ 14,7 % dans la population étudiée (incluant l'insuffisance cardiaque, AVC, infarctus) ; l'hypertension est le facteur de risque le plus répandu, touchant 49 % des femmes et 38 % des hommes âgés de 35 à 64 ans selon l'enquête ***DHS de 2017.***

De plus, même parmi les jeunes (18-30 ans), près de 23,5 % présentent une pression artérielle élevée, ce qui suggère une progression précoce du risque cardiovasculaire selon ***World Heart Federation.***

Ces données soulignent l'importance d'une analyse sur les maladies cardiovasculaires pour comprendre les facteurs de risque, identifier les populations les plus vulnérables et proposer des modèles prédictifs adaptés afin de guider les politiques de prévention et d'intervention.

1.1. Objectif général

Analyser les facteurs de risque associés aux maladies cardiovasculaires à partir de données cliniques et comportementales issues du jeu de données ***Cardiovascular Disease Dataset de Kaggle,*** afin de développer des outils prédictifs et des visualisations permettant d'améliorer la prévention et la prise de décision en santé publique en Haïti.

1.2. Objectifs spécifiques

- Décrire les caractéristiques cliniques et comportementales des individus du jeu de données afin d'identifier les tendances générales liées à la santé cardiovasculaire
- Évaluer la relation entre certaines variables (âge, cholestérol, pression systolique, pression diastolique, tabagisme, activité physique, etc.) et la présence d'une maladie cardiovasculaire ;

- Construire et comparer des modèles (logistic regression, random forest, etc.) pour prédire la probabilité d'apparition d'une maladie cardiovasculaire ;
- Identifier les variables les plus significatives influençant le risque cardiovasculaire à travers des techniques de **feature importance** et de corrélation ;
- Développer un tableau de bord interactif sur Power BI pour visualiser les résultats et faciliter l'interprétation des tendances par les acteurs du secteur de la santé ;
- Implémenter le modèle le plus performant .

1.3. Hypothèses

- L'âge, la pression artérielle et le cholestérol sont des facteurs fortement associés à la probabilité de développer une maladie cardiovasculaire ;
- Les modèles de Machine Learning peuvent prédire efficacement la probabilité d'une maladie cardiovasculaire à partir de données cliniques.

2. Limites de l'Étude :

Les variables disponibles se concentrent sur des indicateurs physiques de base et n'intègrent pas de facteurs contextuels (stress, antécédents familiaux, régime alimentaire, traitements médicaux, etc.) qui influencent également le risque cardiovasculaire. Le déséquilibre entre les classes et la présence de cas limites dans les données peuvent avoir réduit la capacité des modèles à généraliser.

3. Public cible

Ce projet s'adresse aux acteurs impliqués dans le diagnostic, la prévention et la gestion des maladies cardiovasculaires. Il vise à fournir des informations exploitables, issues de l'analyse de données, pour appuyer la prise de décision en santé publique et améliorer la gestion clinique des patients à risque. D'une part les institutions de santé comme le MSPP et l'OPS/OMS pourront s'en servir pour cibler les campagnes de sensibilisation et de prévention. D'autre part, les professionnels de santé pourront repérer plus tôt les profils de patients à risque et adapter leurs protocoles de suivi.

En outre, les chercheurs et étudiants en santé ou en data science y trouveront un cadre d'application concret des méthodes analytiques. Les décideurs politiques disposeront d'éléments factuels pour élaborer des politiques de santé préventive efficaces. Enfin, la population générale et les ONG pourront utiliser les visualisations et les indicateurs produits pour renforcer la sensibilisation aux risques cardiovasculaires. Ce projet établit ainsi un lien entre science des données et santé publique au service de la prévention et de la planification sanitaire en Haïti.

4. Compréhension des données

Le jeu de données utilisé dans cette étude provient de la plateforme ***Kaggle***, sous le nom de "***Cardiovascular Disease Dataset***". Il regroupe des informations cliniques et comportementales collectées sur plus de 70 000 individus, utilisées pour prédire la probabilité de survenue d'une maladie cardiovasculaire. Chaque observation représente un patient et est décrite par 12 variables explicatives et une variable cible (cardio) qui indique la présence (1) ou l'absence (0) d'une maladie cardiovasculaire diagnostiquée. Chaque enregistrement correspond à un individu et comprend plusieurs indicateurs médicaux, biologiques et liés au mode de vie.

La variable cible, nommée cardio, prend deux valeurs : 0 pour les personnes considérées comme saines, 1 pour celles présentant une maladie cardiovasculaire diagnostiquée. Les autres variables servent à expliquer ou prédire cette cible. Elles se répartissent en deux grandes catégories : les variables catégorielles encodées, ce qui rend le travail beaucoup plus abordable et les variables numériques (continue et discontinue).

4.1. Pour les variables catégorielles, on trouve :

- Les variables **cholestérol** et **gluc** codées sous forme de chiffres pour représenter des niveaux croissants : 1 = normal, 2 = élevé, 3 = très élevé.
- La variable **gender** est encodée numériquement :1 = femme, 2 = homme.
- La variable **alco** pour la consommation d'alcool :0 pour non-buveur, 1 pour buveur).
- La variable **active** pour le niveau d'activité physique des patients (0 pour inactif, 1 pour actif).
- La variable **smoke** pour le statut de fumeur du patient (0 pour non-fumeur, 1 pour fumeur).

4.2. Pour les variables numériques (continu ou discontinu) :

- La variable **age** du patient en jours.
- La variable **height** donne la taille du patient en centimètre.
- La variaable **weight** pour le poids du patient en kilogrammes.
- La variable **ap_hi** pour la pression artérielle systolique.
- La variable **ap_lo** pour la pression artérielle diastolique.

Ainsi, bien que toutes les valeurs apparaissent sous forme d'entiers ou de nombres réels, plusieurs représentent en réalité des catégories ou des ordres logiques qu'il faudra interpréter correctement au moment de l'analyse

5. Méthodologie

Pour la réalisation du travail, les méthodes suivantes ont été appliquées :

- Compréhension des Données

La compréhension du jeu de données a été d'une importance capitale pour ce travail et passe par l'importation du fichier sur l'extension Jupyter notebook pour sa lecture après avoir importé toutes les bibliothèques jugées nécessaires au préalable. Ainsi on a eu une première idée du contenu du jeu de données et construit par la suite un cadre de travail.

- Data Cleaning (Nettoyage de Données)

Cette étape a été indispensable dans la mesure qu'elle a permis de modifier les noms des colonnes par des noms facilement interprétables par le large public. Elle a également permis de corriger certaines valeurs jugées aberrantes pouvant fausser les résultats, nuire aux analyses et erroner les conclusions.

- Analyse descriptive

Cette étape a permis d'avoir une idée claire de la distribution des variables dans le jeu de données. Ainsi, on a procédé à des :

  - ✓ Mesures de tendance centrale : moyenne, médiane, etc. ;
  - ✓ Mesures de dispersion : écart-type ;
  - ✓ Mesures de position : quartiles, centiles, minimum et maximum ;
  - ✓ Visualisations de données (histogrammes, graphiques à barres, nuages de points).

- Analyse diagnostique

Après avoir visualisé les données, on a procédé à l'analyse diagnostique, qui nous a permis de :

  - ✓ Identifier les relations entre variables (analyses de corrélation) ;
  - ✓ Faire des comparaisons entre différents résultats.

- Analyse prédictive

L'analyse prédictive du jeu de données comprend :

  - ✓ Modélisation statistique ;
  - ✓ Techniques de Machine Learning ;
  - ✓ Validation des modèles (ensemble de test pour évaluer la performance) ;
  - ✓ Prédictions basées sur des ensembles de données.

- Analyse prescriptive :
  - ✓ Modèles d'optimisation ;
  - ✓ Analyse des recommandations pour la prise de décision.

- Implémentation du modèle :

L'implémentation du modèle repose sur Streamlit pour offrir une interface intuitive de prédiction du risque cardiovasculaire. Le modèle XGBoost le plus performant par rapport à sa précision globale,a été sauvegardé via joblib et chargé en mémoire avec gestion des exceptions pour éviter les erreurs de chargement.

Les données saisies par l'utilisateur (âge, tension, cholestérol, etc.) sont validées à l'aide de bornes minimales et maximales pour respecter les limites du contexte de l'étude.

Les variables qualitatives sont encodées en valeurs numériques et l'IMC est automatiquement calculé puis classé.

Ces informations sont rassemblées dans un DataFrame conforme au format d'entraînement du modèle.

Une fois la prédiction effectuée, le résultat et la probabilité de risque sont affichés dynamiquement.

Les prédictions sont sauvegardées dans un fichier CSV pour assurer un suivi historique des résultats.

6. Présentation et Analyse des Résultats

6.1. Analyse descriptive

L'analyse du jeu de données montrent que les informations ont été tirées des personnes âgées de 30 à 65 ans, pesant au moins 10 kg et 55 cm de taille. L'âge, la hauteur et le poids les plus récurrents sont : {**âge** : (56, 6005), **height** : (165, 5827), **weight** : (65.0, 3837)}.

Ces données ont montré que certaines personnes vivent dans des conditions difficiles et d'autres dans des conditions normales. En ce qu'il y a trait avec la pression artérielle, communément appelée tension au niveau des communautés haitiennes, des cas d'hypotension et d'hypertension sévère ont été détectés : la pression exercée par le sang dans les artères lors que le cœur est au repos peut aller de 40 à 180 mmHg et des cas d'hypertension allant de 60 jusqu'à 240mmHg.

En vue d'avoir une idée plus poussée de la situation clinique et comportementale de la population enquêtée, les graphiques 1, 2 et 3 suivantes présentent la distribution des différentes variables du jeu de données.

1. Distribution des variables numériques

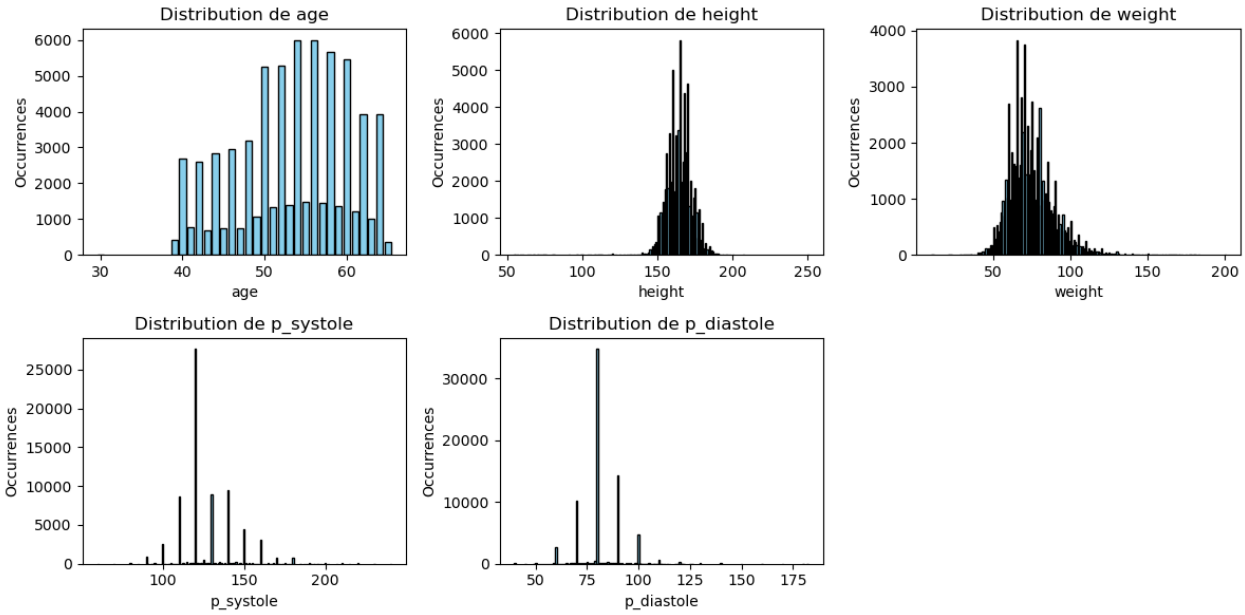


Figure 1: Distribution de certaines variables numériques

Observation :

Ces graphiques montrent qu'une grande majorité de personnes sont âgées de 50 à 60 ans au niveau de l'échantillon analysé et leurs hauteurs sont concentrées entre 160 à 175 cm. La majorité des personnes ont un poids compris entre 60 à 80 kg avec quelques cas isolés pouvant aller jusqu'à 200kg. La pression systolique des personnes de l'échantillon est concentrée autour de 100 à 150

mmHg et autour de 60 à 100 mmHg pour la pression diastolique. Ces données ont montré que dans certains cas, la pression est éstimée à plus de 200 mmHg et plus de 125 mmHg, ces chiffres signalent des cas critiques de pression artérielle pouvant compromettre la santé des patients et conduire au développement des maladies cardiovasculaires.

1. Distribution des variables catégorielles :

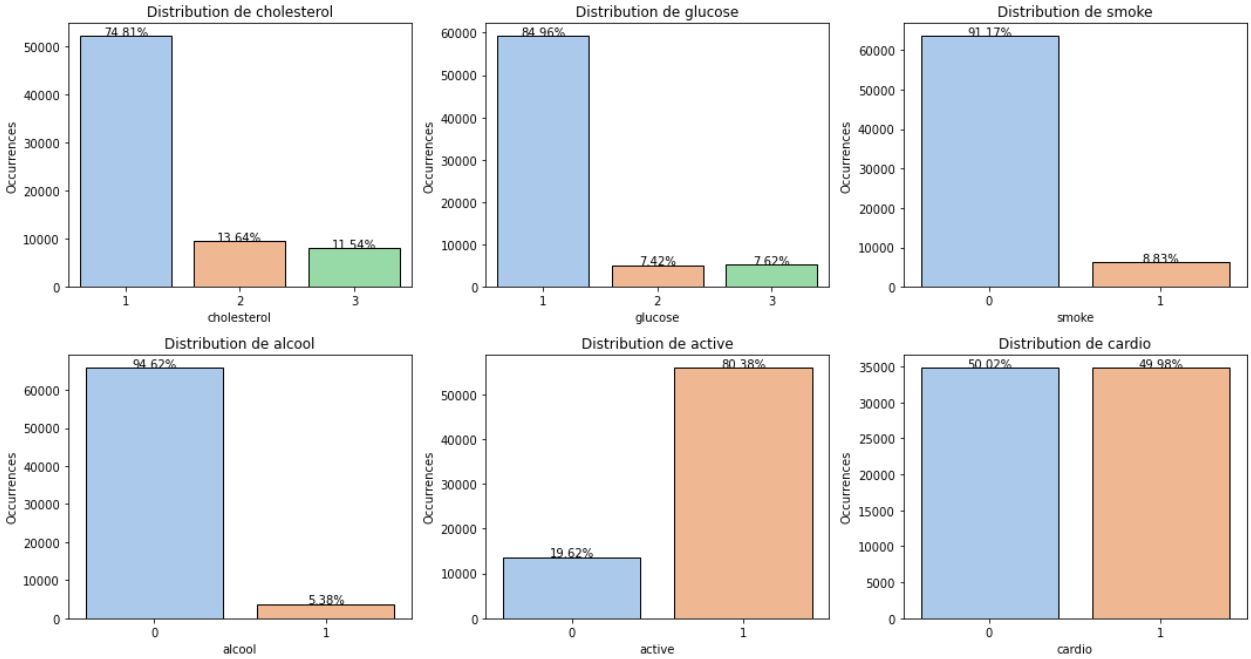


Figure 2: Distribution de certaines variables catégorielles

Observation :

- Pour la variable cholestérol qui se répartit en trois niveaux : 1 pour niveau normal, 2 pour niveau élevé et 3 pour niveau très élevé, la grande majorité des personnes ont un niveau de cholestérol normal, soit 74.81% contre respectivement 13.64% et 11.54% pour le niveau élevé et très élevé ;
- La grande majorité des personnes ont un niveau de glucose normal, soit 84.96% contre 7.42% et 7.62% pour le niveau élevé et très élevé ;
- Les non-fumeurs représentent 91.17% et les fumeurs sont : 8.83% ;
- 94.62% des personnes ne consomment pas d'alcool contre 5.38% ;
- 80.38% des personnes exercent des activités physiques contre 19.62% ;
- 49.98% des personnes ont des maladies cardiovasculaires contre 50.02% saines .

2. Répartition des sexes au niveau de l'échantillon :

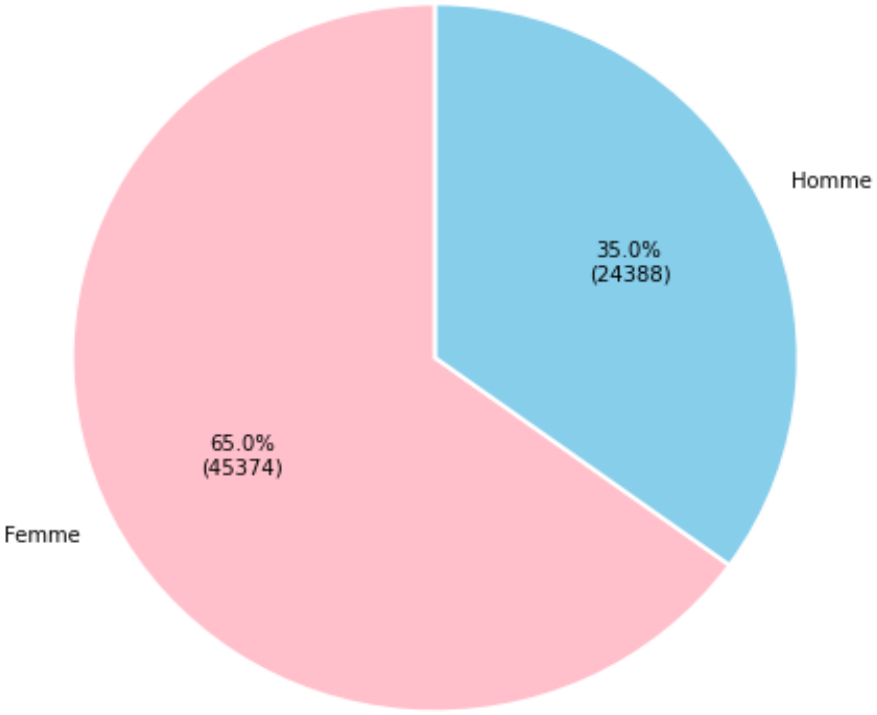


Figure 3: Répartition des sexes au niveau de l'échantillon

Observation :

Ce graphique démontre que ces données ont été collectées majoritairement sur des femmes, qui représente 65% de l'échantillon contre 35% pour les hommes.

6.2. Analyse diagnostique

1. Tendance de maladies cardiovasculaires selon l'âge

Les résultats de l'analyse de ce jeu de données ont démontré que les maladies cardiovasculaires augmentent avec l'âge. Ces maladies surviennent après les 30 ans, et qu'après 45 ans, plus de 40 % des personnes sont atteintes de maladies cardiovasculaires avec un pic autour de 63-64 ans (plus de 70%), comme c'est indiqué sur le graphique 4 suivant. C'est une période à haut risque qui mérite une attention médicale accrue et une prévention ciblée.

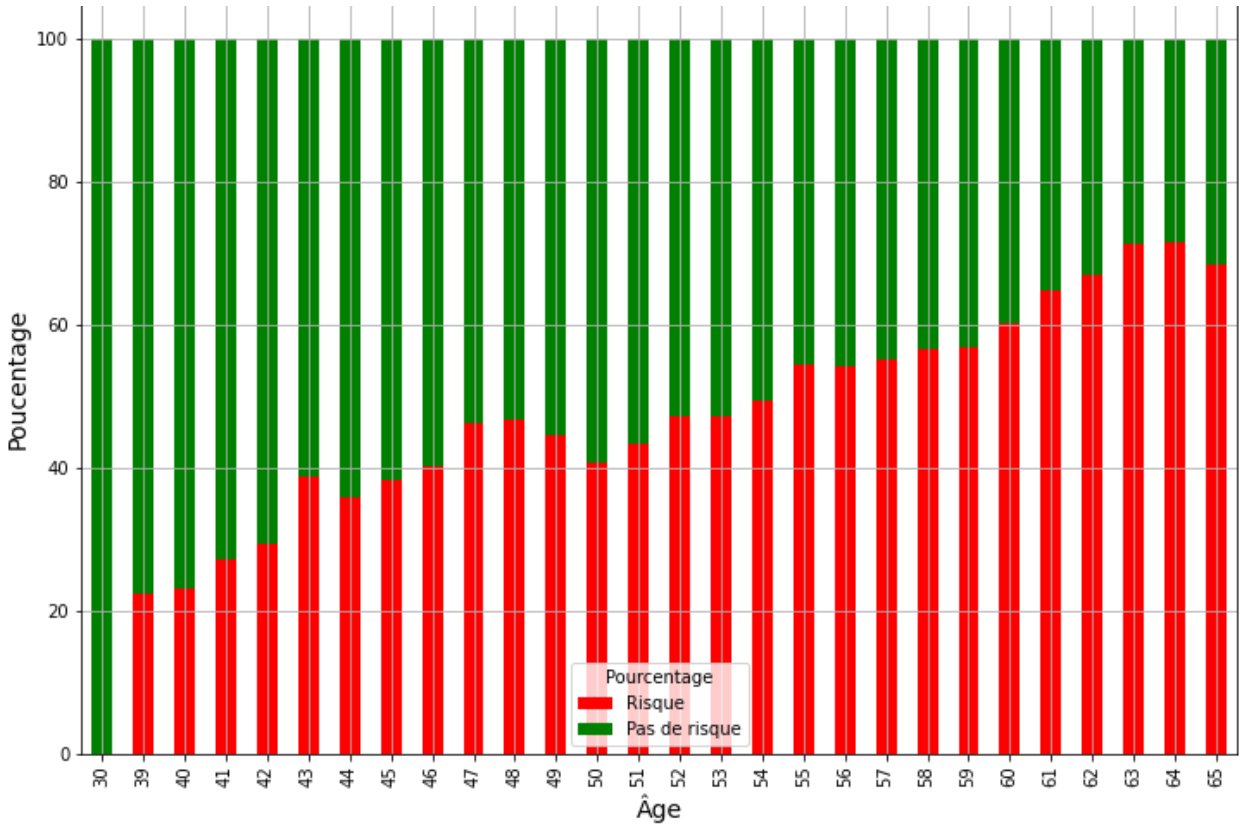


Figure 4: Tendance de maladies cardiovasculaires selon l'âge

2. Tendance des maladies cardiovasculaire selon le genre

Pour les deux groupes (hommes et femmes), la différence est relativement faible. En ce qui concerne les femmes, 50.3% sont saines contre 49.7% malades. Chez les hommes, c'est 49.5% sains contre 50.5% malades. Étant donné que l'échantillon analysé est majoritairement composé de femmes, on s'en suit pour dire que sur la base de ces résultats, il y a beaucoup plus de femmes qui souffrent de maladies cardiovasculaires. Cependant du point de vue de proportionnalité, les maladies cardiovasculaires sont légèrement plus élevées chez les hommes (50.5% contre 49.7%) comme illustré sur le graphique 5 suivant. Ce qui nous rapproche de la littérature médicale qui affirme qu'il y aurait une prédominance pour les maladies cardiaques chez les hommes par rapport aux femmes.

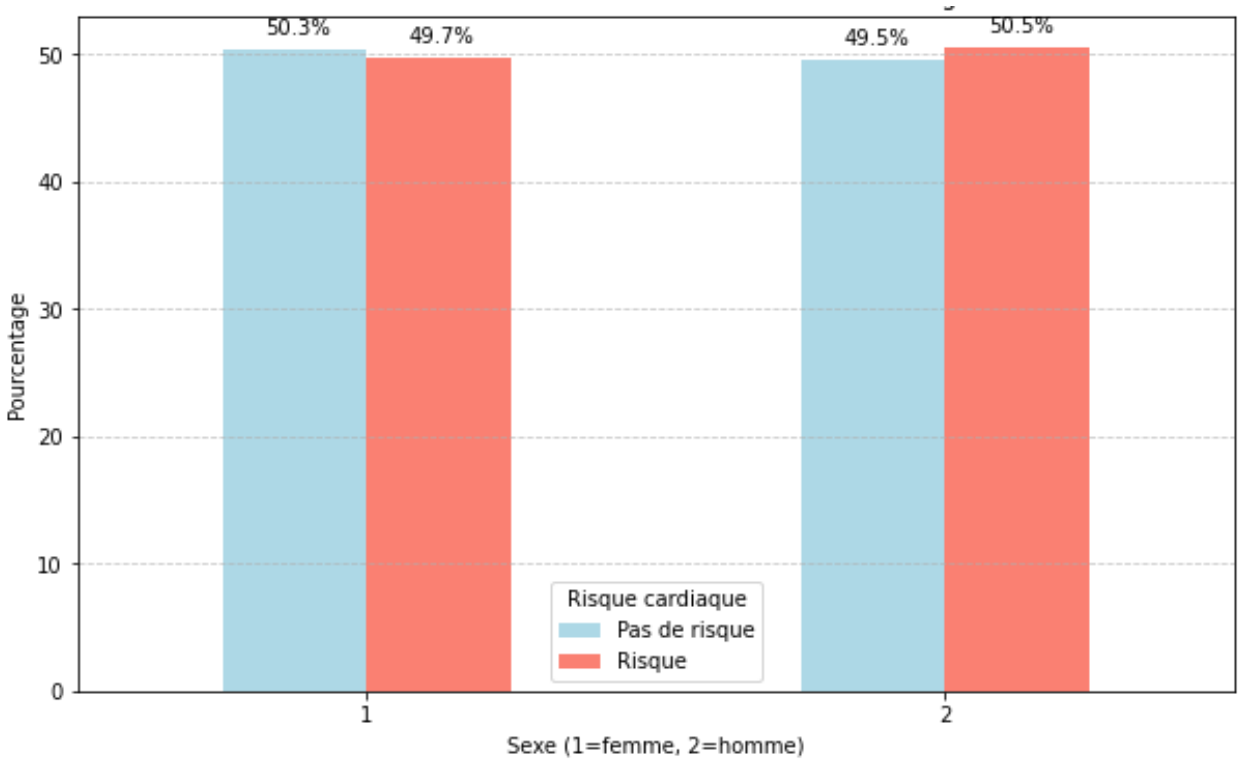


Figure 5: Tendance des maladies cardiovasculaire selon le genre

3. Tendance des maladies cardiovasculaire selon le niveau de cholestérol

L'analyse diagnostique montre qu'il y a de liens entre le niveau de cholestérol des personnes enquêtés (faible, moyen ou élevé) et les maladies cardiovasculaires, comme indiqué sur le graphique 6 suivant. C'est à dire que le risque augmenterait tant que le niveau de cholestérol augmente. Pour les personnes qui ont un taux de cholestérol normal, 56% sont saines contre 44% à risque, alors que parmi celles qui ont des taux de cholestérol très élevé, 76.5% sont malades contre 23.5% sains. ce qui voudrait dire qu'un taux de cholestérol normal ne vous écarterait pas tout à fait des maladies cardiovasculaires, mais il diminuerait le risque d'être atteint.

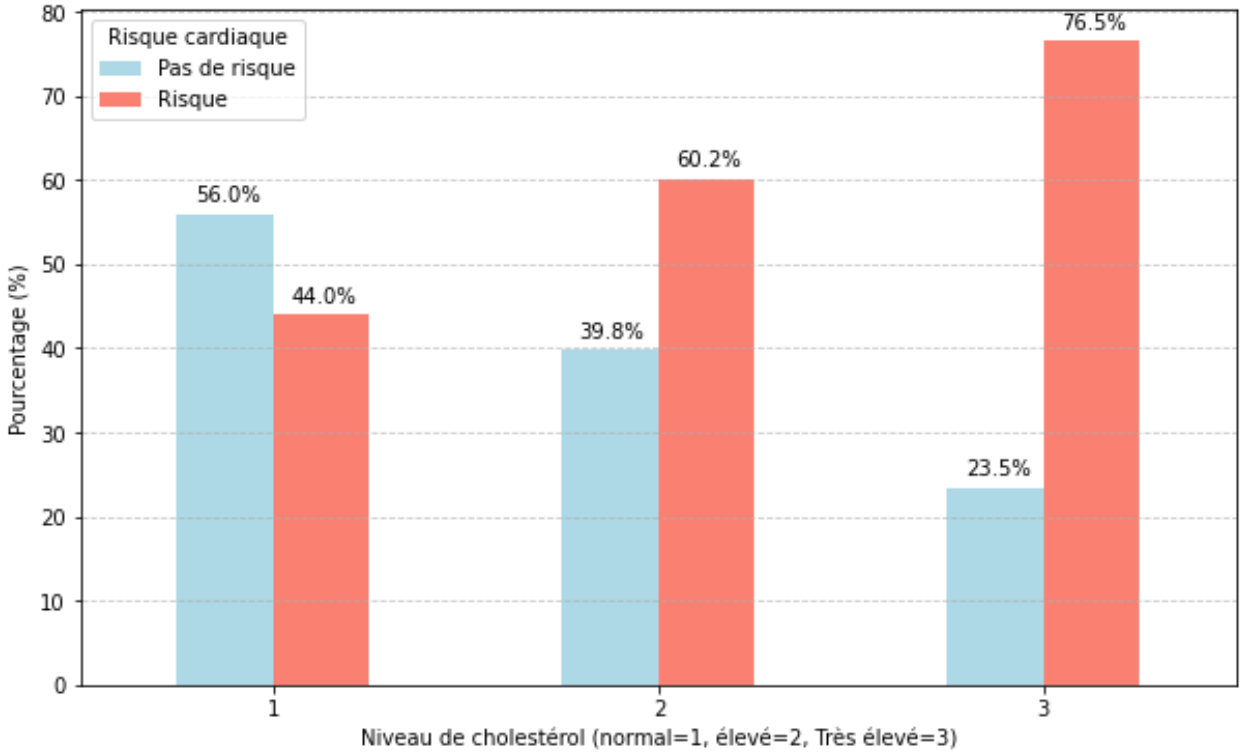


Figure 6: Tendance des maladies cardiovasculaire selon le niveau de cholestérol

4. Tendances des maladies cardiovasculaires selon le niveau d'activités

Comparativement aux personnes inactives, les personnes qui effectuent des exercices physiques sont légèrement moins sujettes aux maladies cardiovasculaires. Le graphique 7 suivant montre que chez les personnes inactives, 53.6% sont à malades contre 46.4% qui ne le sont pas. Or, chez les personnes actives, 49.1% sont à risques contre 50,9% sans risques. Quoique la différence est légère, mais plus vous pratiquez des exercices physiques modérées moins vous êtes exposés aux risques des maladies cardiocasculaires, semble-t-il.

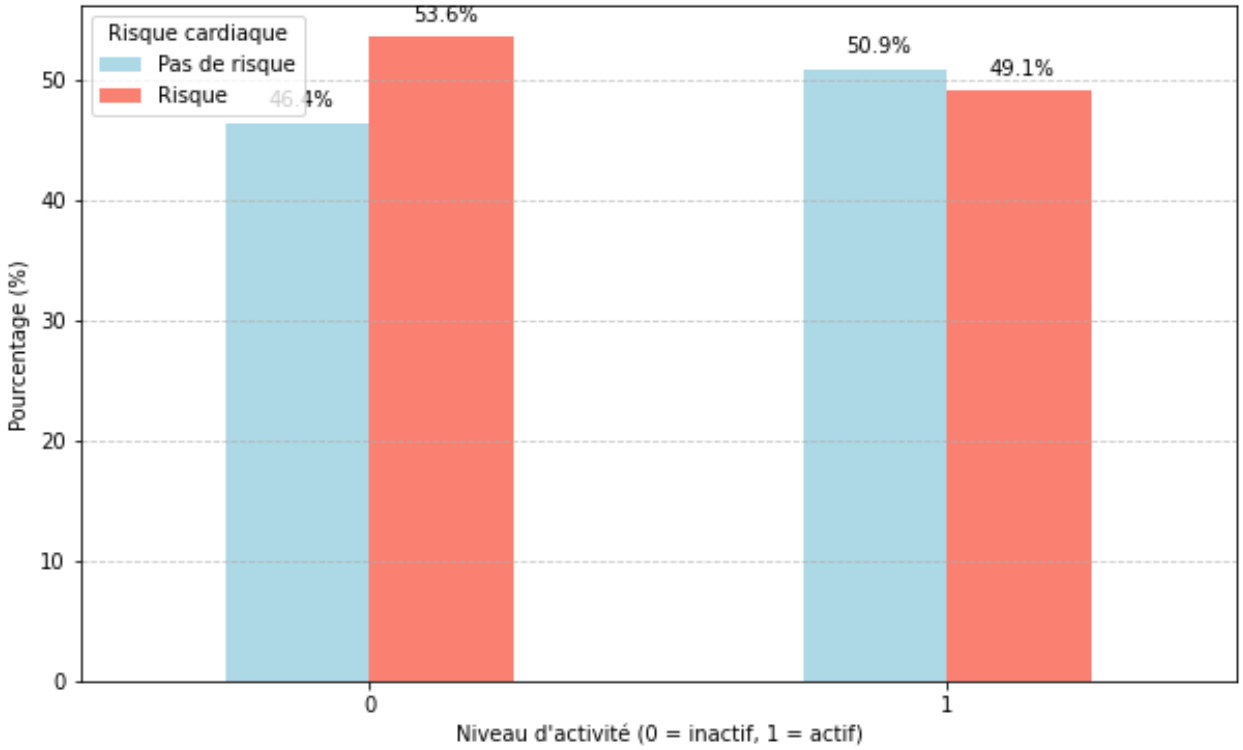


Figure 7: Tendances des maladies cardiovasculaires selon le niveau d'activités

5. Effets de l'alcool et du tabagisme sur les maladies cardiovasculaires

Les données ont montré que le tabagisme à lui seul et l'alcool ne sont pas des facteurs discriminant pour les maladies cardiovasculaires. Les graphiques suivants ont montré que chez les personnes qui ne fument pas 50,2% ont des maladies cardiovasculaires contre 49.8% Cette difference est encore plus marquée chez les fumeurs, car 47.5% ont des maladies cardiovasculaires contre 52.5% sains. La tendance est la même pour l'alcool, parmi ceux qui boivent il y a 48.4% qui en souffrent contre 51.6%.

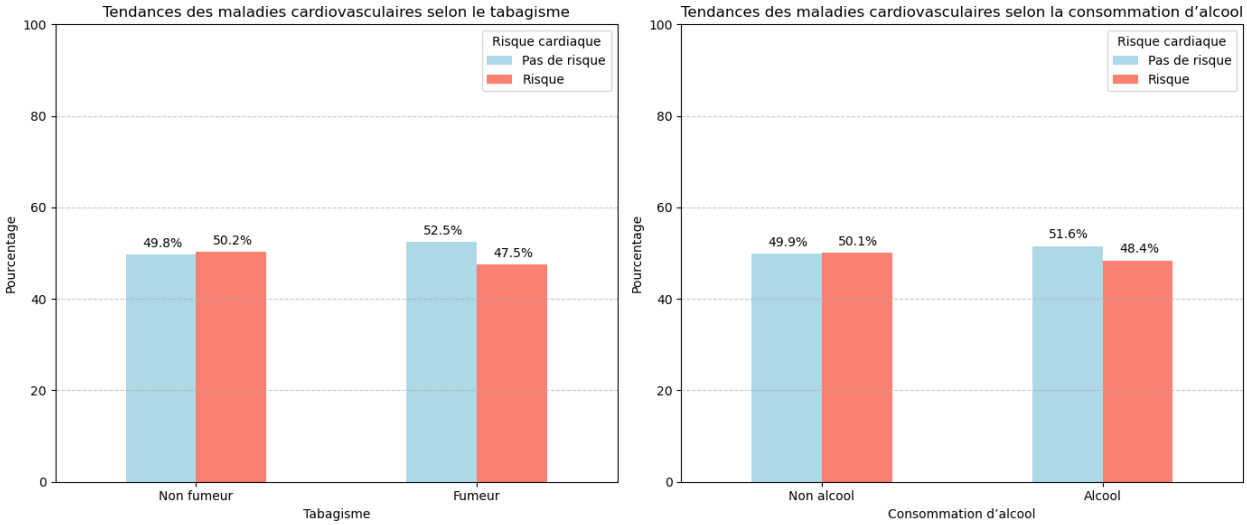


Figure 8: Effets de l'alcool et du tabagisme sur les risques cardio

6. Répartition des personnes selon la consommation d'alcool et le tabagisme

Dans ce jeu de données, la majorité des personnes ne boivent pas et ne fument pas (88.5%) à la fois, Il n'y a qu'une très faible portion qui font les deux à la fois (2.7%).

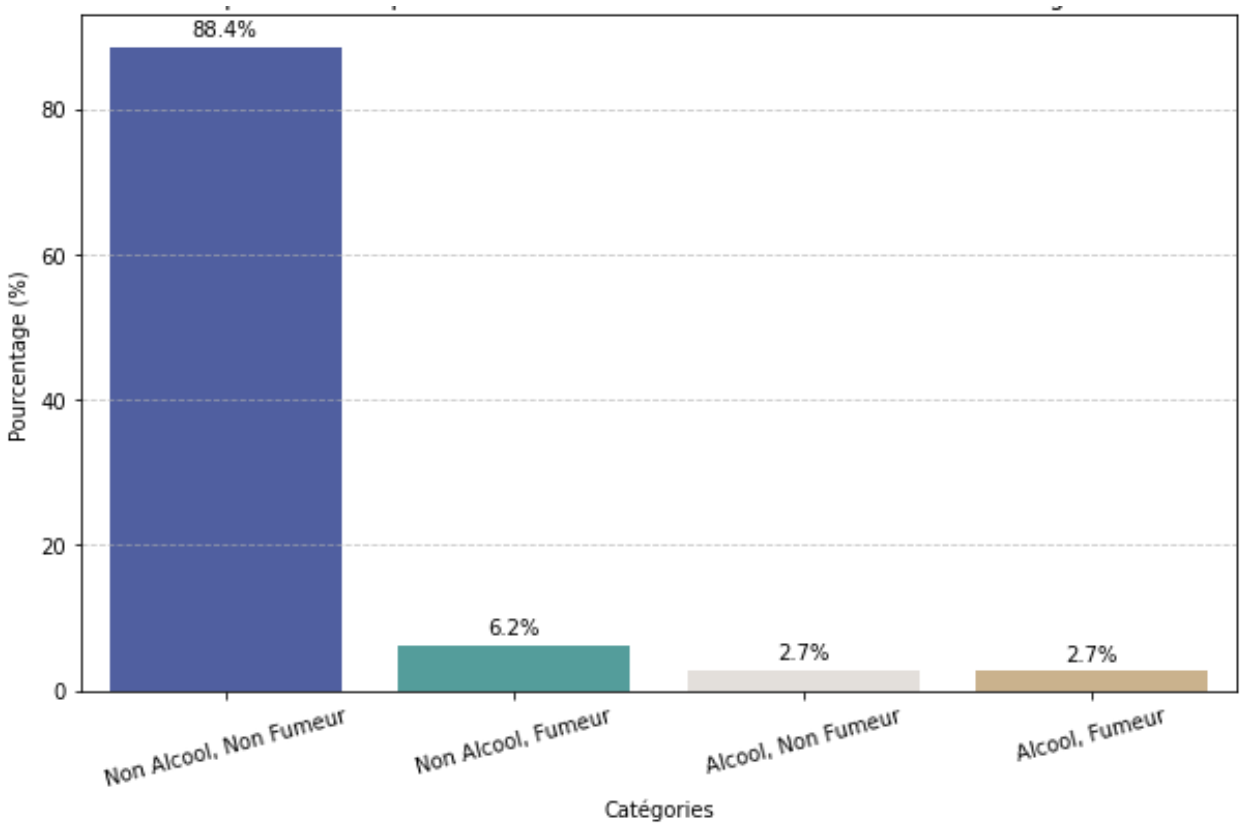


Figure 9 : Répartition des personnes selon la consommation d'alcool et le tabagisme

7. Tendance des maladies cardiovasculaires selon la pression artérielle

Les personnes qui ont des pressions systoliques supérieures à 120 mmHg et des pressions diastoliques supérieures à 90 mmHg sont celles qui souffrent le plus des maladies cardio-vasculaires. Il y a une relation positive entre la pression artérielle (systolique et diastolique) et les maladies cardiovasculaires. À mesure que la pression systolique augmente, le risque augmente, passant de 31.6% à 83.3% pour une variation de pression 60 à 250 mmHg. C'est aussi la même situation pour la pression diastolique qui passe de 28.3% à 84.0% pour une variation de pression 40 à 200 mmHg.

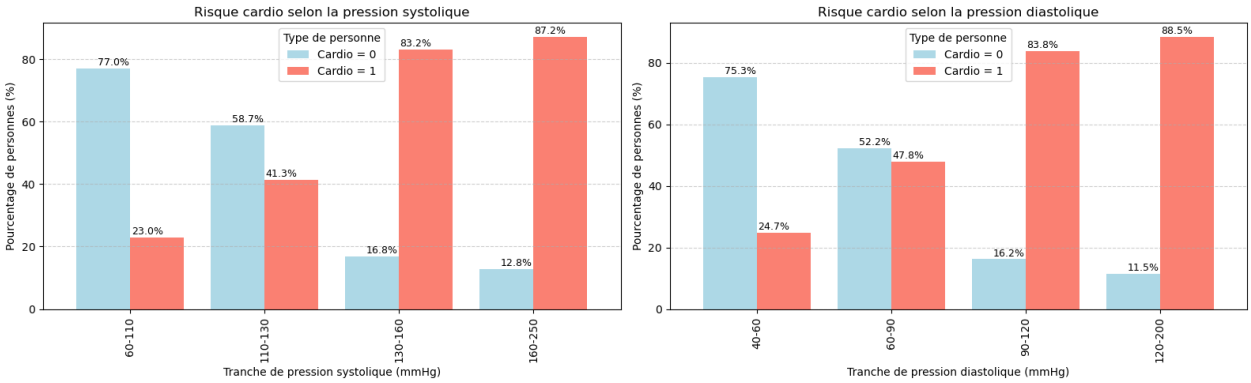


Figure 10 : Tendance des maladies cardiovasculaires selon la pression artérielle

8. Tendance des maladies cardiovasculaires selon le taux de glucose

Les personnes qui ont des taux de glucose élevée sont celles qui souffent le plus des maladies cardiovasculaires.

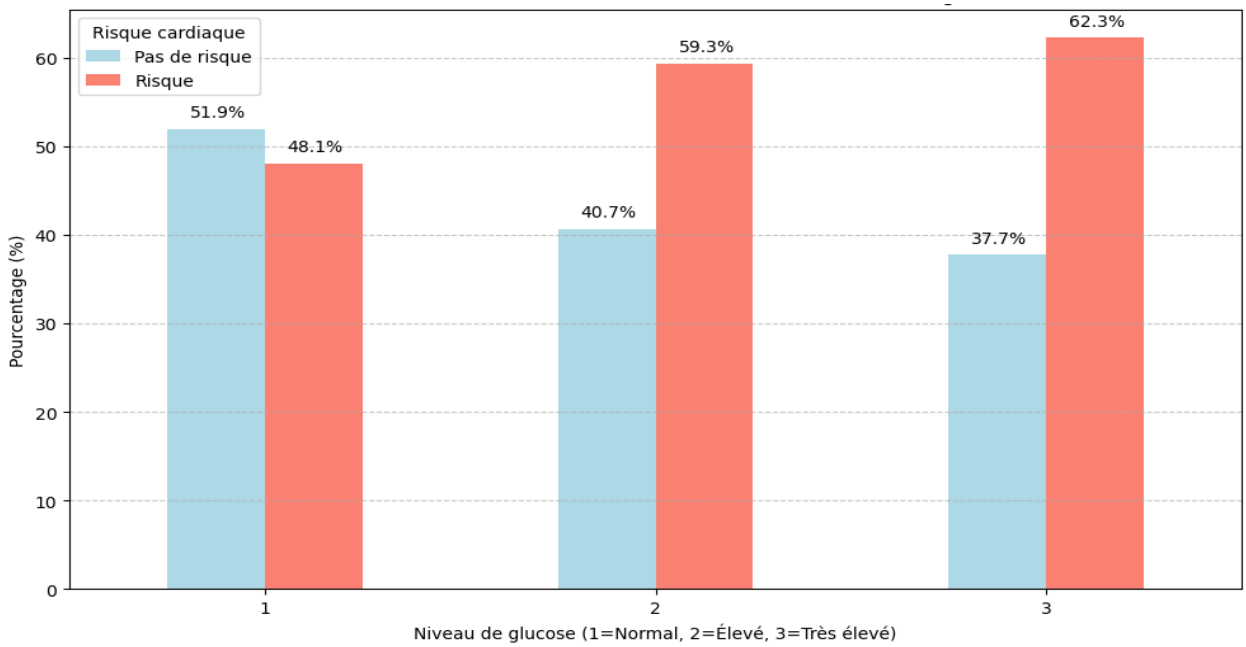


Figure 9: Tendances des maladies cardiaques selon le taux de glucose

9. Matrice de corrélation (a)

Selon la matrice que présente la figure 11, les maladies cardiovasculaires sont beaucoup plus corrélées à la tension artérielle systolique (0.43), diastolique (0.35), l'âge (0.24), le cholestérol (0.22) et le poids (0.18). Inversement, les risques de maladies cardiovasculaires sont négativement corrélés avec les personnes actives (-0.04).

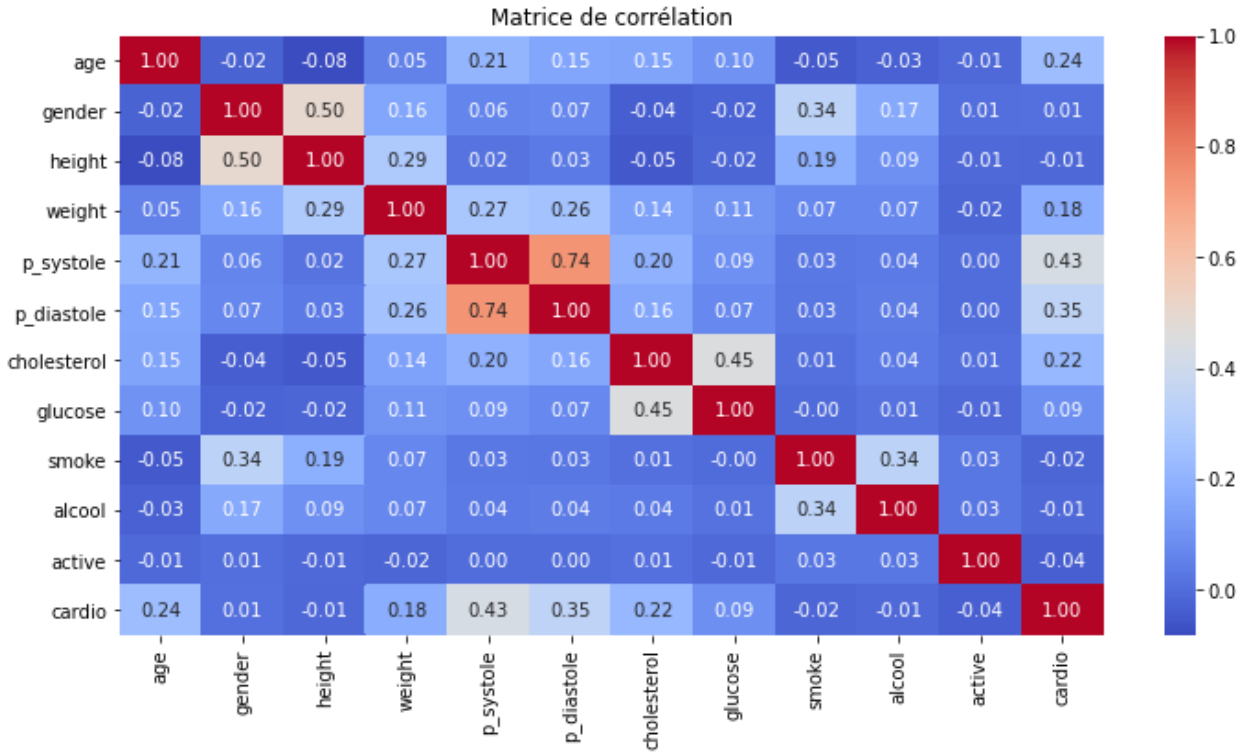


Figure 11 : Corrélation entre les variables

10. Matrice de corrélation (b)

Après avoir introduit l'indice de masse corporelle pour substituer le poids et la hauteur, on voit qu'il n'y a pas de changement dans la corrélation du cardio avec le poids. Puisque cet indice tient compte du poids et de la hauteur à la fois, c'est un bon indicateur pour la suite du travail.

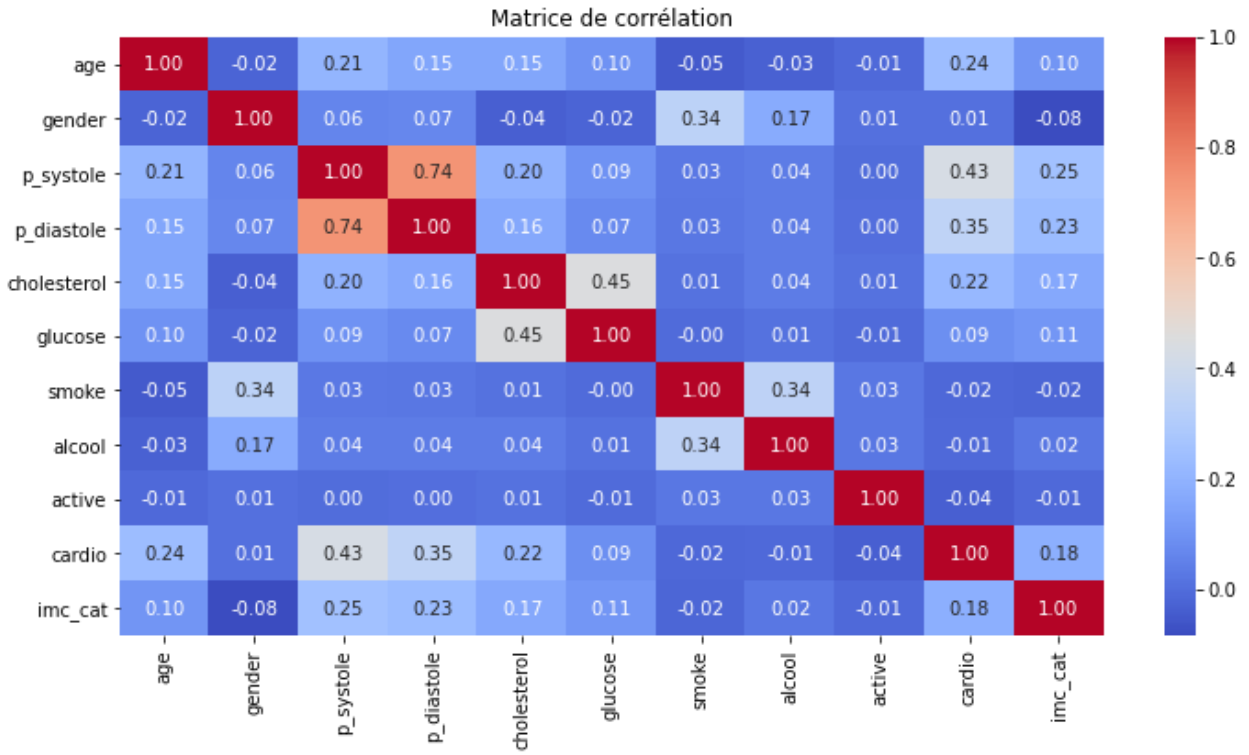


Figure 12 : corrélation entre les variables

6.3. Discussion sur l'analyse exploratoire

L'analyse exploratoire du jeu de données révèle qu'il s'agit d'un échantillon composé de personnes âgées d'au moins 30 ans, majoritairement féminine, avec un âge maximal de 65 ans. Les individus mesurent au minimum 55 cm pour un poids d'au moins 10 kg. Les valeurs les plus récurrentes observées sont l'âge de 56 ans, la taille de 165 cm et le poids de 65 kg.

La plupart des participants présentent des niveaux de cholestérol et de glucose normaux. Globalement, les individus ne fument pas, consomment peu ou pas d'alcool et sont généralement actifs physiquement. Malgré ce mode de vie, les données montrent un équilibre presque parfait entre les personnes atteintes de maladies cardiovasculaires et celles qui ne le sont pas (49.98% de cas de maladies). C'est pourquoi on a été intrigué par l'idée d'identifier les variables qui expliquent mieux la présence des maladies cardiovasculaires.

Sur le plan biologique, les graphiques mettent en évidence un lien entre les niveaux de cholestérol et de glucose et le risque cardiovasculaire. De plus, les personnes ayant une pression artérielle systolique supérieure à 120 mmHg, une pression diastolique dépassant 80 mmHg sont les plus exposées aux maladies cardiovasculaires. À l'inverse, les personnes physiquement actives semblent légèrement mieux protégées.

De plus, l'âge apparaît comme un facteur déterminant : le risque cardiovasculaire augmente nettement à partir de 45 ans pour dépasser les 40 %, atteignant un pic critique autour de 63 à 64 ans, soit un pourcentage de risque de plus de 70 %. Cette tranche d'âge constitue donc une période à haut risque nécessitant une attention médicale accrue.

L'analyse de corrélation confirme ces observations : les maladies cardiovasculaires sont principalement associées à la pression systolique (r = 0,43), à la pression diastolique (r = 0,34), à l'âge (r = 0,24), au cholestérol (r = 0,22) et à la corpulence de la personne (r = 0,18). Après ces facteurs, vient le taux de glucose (r=0.09) qui influence les maladies cardiovasculaires de façon très modérée et le genre qui n'influence presque pas les maladies cardiovasculaires dans le contexte de cet échantillon (0.01). À l'opposé, l'activité physique montre une corrélation légèrement négative (r = -0,04), indiquant un effet protecteur bien que minime.

6.4. Analyse prédictive

Pour l'analyse prédictive, elle est divisée en deux parties : Modélisation initiale et modélisation optimisée. En effet, six (6) modèles ont été testés afin de prédire les risques de maladies cardiovasculaires.

6.4.1. Modélisation initiale :

Les résultats de la modélisation initiale sont présentés dans le tableau 1 suivant afin de mieux comparer les modèles.

Tableau 1: Résumé des modèles initiales

Model / Performances	Accuracy	precision		recall		f-score		support	
		0	1	0	1	0	1	0	1
Regression Logistique	0.7308	0.71	0.75	0.78	0.68	0.74	0.72	7014	6939
Decision Tree	0.6804	0.66	0.71	0.76	0.60	0.70	0.65	7014	6930
Random Forest	0.7061	0.70	0.72	0.74	0.67	0.72	0.69	7014	6930
XGBoost	0.7321	0.72	0.75	0.78	0.69	0.74	0.72	7014	6930
KNN	0.6962	0.70	0.70	0.70	0.69	0.70	0.70	7014	6930
SVM (Support Vector Machine	0.7341	0.71	0.76	0.79	0.67	0.75	0.72	7014	6930

a) Régression Logistique :

Le modèle de régression logistique présente une performance globale de 73,1 %. Ce qui veut dire que le modèle a raison environ 73 fois sur 100.

La matrice de confusion révèle 5472 vrais négatifs et 4730 vrais positifs, mais aussi 1542 faux positifs et 2209 faux négatifs, montrant que le modèle reconnaît mieux les individus sains que les malades.

Avec une précision de 0,71 et un rappel de 0,78 pour les non malades contre respectivement 0,75 et 0,68 pour les malades, le modèle tend à sous-détecter certains cas de maladie cardiovasculaire, ce qui reste une limite dans le contexte médical. Néanmoins, l'équilibre entre les classes et les scores moyens de recall et de f1-score (0,73) démontrent une bonne cohérence interne.

Ce pourcentage de **recall 0.68** pour la classe 1, met l'accent sur le fait que sur 100 malades réels, notre modèle en détecte environ 68 et en rate 32. C'est un résultat moyen, pas tout à fait satisfaisant en médecine, car rater un malade peut avoir de graves conséquences. Il est préférable de rater un non malade que de rater un malade. Donc, on peut maximiser le recall pour ne rater aucun malade.

b) Decision Tree

Le modèle Decision Tree atteint une exactitude globale de 68 %, ce qui traduit une performance moyenne, inférieure à celle de la régression logistique.

Il présente une meilleure aptitude à identifier les individus sans maladie cardiovasculaire que ceux atteints. En effet, la matrice de confusion indique 5298 vrais négatifs contre 4196 vrais positifs, mais aussi 1716 faux positifs et 2743 faux négatifs, soulignant une tendance à manquer plusieurs cas de maladie.

Le rappel de 0,60 pour la classe 1 montre que le modèle détecte seulement 60 % des patients réellement malades, la précision dans ce cas est de 0,71. Ce faible score de rappel met en évidence la faiblesse de ce modèle initial à identifier les cas de maladies.

L'analyse des importances de variables révèle que la pression systolique est de loin le facteur le plus influent (0,45), suivie de l'âge (0,20) et de la pression diastolique (0,09). Les variables liées au mode de vie (tabac, alcool, activité physique) ont un impact plus faible, traduisant un poids moindre dans la décision du modèle.

c) Random Forest :

Le modèle Random Forest atteint une exactitude globale de 71 %, marquant une amélioration notable par rapport au Decision Tree précédent.

La matrice de confusion indique 5187 vrais négatifs et 4666 vrais positifs, contre 1827 faux positifs et 2273 faux négatifs, montrant un certain équilibre entre les deux classes.

Le rappel pour la classe "cardio" est de 0,67, tandis que la précision atteint 0,72, traduisant une performance moyenne pour la détection des patients réellement atteints.

L'analyse des importances de variables met en évidence que la pression systolique (0,33), l'âge (0,27) et la pression diastolique (0,16) sont les facteurs les plus influents dans la prédiction, confirmant leur rôle clé dans le risque cardiovasculaire. Le cholestérol, l'IMC et le glucose conservent une contribution secondaire, tandis que les facteurs comportementaux (tabac, alcool, activité physique) demeurent faiblement explicatifs.

d) XGBoost

Le modèle XGBoost affiche une exactitude globale de 73,2 %, confirmant une nette amélioration des performances par rapport aux modèles précédents.

La matrice de confusion indique que sur 7014 individus sans maladie cardiovasculaire, 5448 sont correctement identifiés (TN) tandis que 1566 sont faussement classés comme atteints (FP). Du côté des 6939 patients réellement cardio, 4768 sont correctement détectés (TP) contre 2171 faux négatifs (FN).

Avec une précision de 0,72 et un rappel de 0,78 pour les non malades contre 0,75 et 0,69 pour les malades, le modèle tend à sous-détecter certains cas de maladie cardiovasculaire, ce qui reste une limite dans le contexte médical.

L'analyse des importances de variables révèle que la pression systolique domine largement l'explication du modèle avec un poids de 0,49, suivie du cholestérol (0,16) et de l'âge (0,09). Les autres facteurs : tels que l'activité physique, le tabac, l'IMC et le glucose contribuent de façon plus modérée, tandis que la pression diastolique et le genre ont un impact marginal.

e) KNN

Le modèle K-Nearest Neighbors (KNN) présente une exactitude globale de 69,6 %, indiquant une performance inférieure à celle des modèles plus complexes comme le Random Forest ou le XGBoost.

La matrice de confusion montre que sur 7014 individus sans maladie cardiovasculaire, 4929 sont correctement identifiés (TN) tandis que 2085 sont faussement classés comme atteints (FP). Parmi les 6939 patients réellement cardio, 4786 sont correctement détectés (TP) contre 2153 faux négatifs (FN).

Les valeurs de précision (environ 0,70 chacune) et les recalls (0.70, 069), suggère que le modèle ne favorise prèsque ni les positifs ni les négatifs, mais qu'il manque légèrement de finesse dans la séparation des classes.

f) SVM (Support Vector Machine)

Le modèle SVM (Support Vector Machine) atteint une exactitude globale de 73,4 %, soit une performance comparable à celle du XGBoost, legèrement meilleure.

La matrice de confusion montre que sur 7014 individus sans maladie cardiovasculaire, 5574 sont correctement identifiés (TN) tandis que 1440 sont faussement classés comme atteints (FP). Parmi les 6939 patients réellement cardio, 4670 sont correctement détectés (TP) contre 2269 faux négatifs (FN), ce qui correspond à un rappel de 0,67 pour la classe 1.

La précision atteint 0,76 pour les prédictions cardio et 0,71 pour les non-cardio, indiquant que le modèle est plus fiable lorsqu'il prédit la présence d'une maladie.

6.4.2. Optimisation

Les résultats de la modélisation initiale sont présentés dans le tableau 2 suivant afin de mieux comparer les modèles.

Tableau 2: Tableau des modèles optimisés

Model / Performances	Accuracy	precision		recall		f-score		support	
		0	1	0	1	0	1	0	1
Regression Logistique	0.7303	0.71	0.75	0.78	0.68	0.74	0.72	7014	6939
Decision Tree	0.7267	0.69	0.78	0.82	0.63	0.75	0.70	7014	6930
Random Forest	0.7339	0.71	0.76	0.79	0.68	0.75	0.72	7014	6930
XGBoost	0.7355	0.72	0.76	0.78	0.69	0.75	0.72	7014	6930
KNN	0.7146	0.72	0.72	0.73	0.71	0.72	0.71	7014	6930
SVM (Support Vector Machine)	0.7320	0.71	0.76	0.79	0.67	0.75	0.71	7014	6930

a) Régression logistique :

Après optimisation avec GridSearchCV et normalisation via un pipeline, la régression logistique n'est pas significativement améliorée par rapport au modèle initial. L'accuracy sur le jeu de test reste à 73,03 %, avec une matrice de confusion qui montre que sur 7014 individus sans maladie cardiovasculaire, 5466 sont correctement identifiés (TN) et 1548 sont faussement classés comme atteints (FP). Parmi les 6939 patients réellement cardio, 4725 sont correctement détectés (TP) et 2214 sont manqués (FN), donnant un rappel de 0,68 pour la classe 1.

La précision pour la classe 1 est de 0,75 et le recall de 0,68, tandis que pour la classe 0 ("pas cardio") precision = 0,71 et recall = 0,78.

Dans ce cas, le modèle a été déjà proche de sa limite pour cette configuration de données dans la modélisation initiale, et l'optimisation hyperparamétrique n'a permis aucun gain supplémentaire.

b) Decision Tree

Après limitation de la profondeur maximale à 5 et réglage du critère à Gini, le modèle Decision Tree optimisé atteint une exactitude globale de 72,7 %, montrant une nette amélioration par rapport au modèle initial.

La matrice de confusion indique que sur 7014 individus sans maladie cardiovasculaire, 5763 sont correctement identifiés (TN) et 1251 faussement classés comme atteints (FP). Parmi les 6939 patients réellement cardio, 4377 sont correctement détectés (TP) et 2562 manquent à l'appel (FN),

Le rappel est de 0,82 pour la classe 0 et 0,63 pour la classe 1, reflétant une meilleure capacité à identifier les vrais négatifs et une amélioration légère dans la détection des cas de maladies par rapport au précédent.

Les importances des variables montrent que la pression systolique est la plus déterminante, suivie de l'âge et de la pression diastolique, tandis que le mode de vie et la catégorie IMC ont un impact plus modeste.

c) Random Forest

L'optimisation du Random Forest a permis d'atteindre une accuracy test de 73,4 %, montrant une amélioration considérable par rapport aux modèles non optimisés à l'exception du SVM.

La matrice de confusion indique que sur 7014 individus sans cardio, 5545 sont correctement identifiés (TN) et 1469 faussement classés (FP), tandis que parmi les 6939 patients réellement cardio, 4696 sont correctement détectés (TP) et 2243 manquent à l'appel (FN), traduisant un rappel de 0,79 pour la classe 0 et 0,68 pour la classe 1.

Les importances des variables confirment la prépondérance de la pression systolique (49 %) et diastolique (20 %), suivies de l'âge et du cholestérol, tandis que le mode de vie et l'IMC ont un impact plus modeste.

Globalement, ce modèle surpassant la régression logistique et le Decision Tree en termes de performances.

d) XGBoost

L'XGBoost optimisé atteint une accuracy test de 73,7 %, confirmant une amélioration par rapport à la version initiale et aux autres modèles.

La matrice de confusion montre que sur 7014 individus sans maladies cardiovasculaires, 5461 sont correctement identifiés (TN) et 1553 faussement classés (FP), tandis que parmi les 6939 patients réellement cardio, 4821 sont correctement détectés (TP) et 2118 manquent à l'appel (FN), donnant un rappel de 0,78 pour la classe 0 et 0,69 pour la classe 1.

L'analyse des importances révèle que la pression systolique reste la variable la plus déterminante (~39 %), suivie de la pression diastolique, du cholestérol et de l'âge, tandis que le mode de vie et la catégorie IMC jouent un rôle secondaire.

e) KNN

Le KNN optimisé affiche une accuracy test de 71,46 %, légèrement améliorée par rapport au modèle initial.

La matrice de confusion indique que sur 7014 individus sans cardio, 5069 sont correctement identifiés (TN) et 1945 faussement classés (FP), tandis que parmi les 6939 patients réellement cardio, 4902 sont correctement détectés (TP) et 2037 manquent à l'appel (FN), donnant un rappel de 0,73 pour la classe 0 et 0,71 pour la classe 1.

La capacité de recall de ce modèle pour la classe des malades est plus ou moins intéressante par rapport aux autres, mais globalement sa précision fait n'est pas la meilleure.

f) SVM

Le SVM optimisé atteint une accuracy test de 73,2 %, proche de son score moyen en validation croisée (73,5 %).

La matrice de confusion montre que sur 7014 individus sans cardio, 5542 sont correctement identifiés (TN) et 1472 faussement classés (FP), tandis que parmi les 6939 patients réellement cardio, 4675 sont correctement détectés (TP) et 2264 manquent à l'appel (FN).

Le rappel est de 0,79 pour la classe 0 et 0,67 pour la classe 1, tandis que la précision est de 0,71 pour les prédictions pas cardio et 0,76 pour les prédictions cardio.

6.4.3. Récapitulatifs des modèles :

Le tableau 3 suivant présente une synthèse de la performance globale des modèles testés dans le cadre de ce travail.

Tableau 3: Récapitulatifs de tous les modèles testés

	Model	Accuracy (test)
1	Logistic Regression (initial)	0.7309
	Logistic Regression (optimized)	0.7304
2	Decision Tree (initial)	0.6804
	Decision Tree (optimized)	0.7267
3	Random Forest (initial)	0.7062
	Random Forest (optimized)	0.7340
4	XGBoost (initial)	0.7320
	XGBoost (optimized)	0.7355
5	KNN (initial)	0.6963
	KNN (optimized)	0.7146
6	SVM (initial)	0.7342
	SVM (optimized)	0.7322

Le meilleur modèle sur le test est : **XGBoost (optimized)** avec un **Accuracy (test) : 0.7369**

6.5. Interprétation Générale :

Analyse descriptive :

Les résultats montrent que les maladies cardiovasculaires touchent une proportion presque équivalente entre les cas positifs et négatifs. L'échantillon analysé est majoritairement féminin avec 65% de femmes et 35% d'hommes âgé de 30 à 65 ans avec une concentration d'âge autour de 50 à 60 ans. L'âge, la taille et le poids le plus fréquent est : 56 ans, 156 cm et 65 kg. La plupart des personnes présentent un niveau normal de cholestérol et de glucose, ne fument pas, consomment peu ou pas d'alcool et exercent une activité physique régulière. Malgré tout, la proportion des maladies cardiovasculaire reste élevée, ce qui reste à croire que d'autres facteurs biologiques

comme la pression artérielle et le niveau de cholestérol peuvent prendre le dessus et entrainent de nombreux cas cardiovasculaires.

Analyse diagnostique :

Les individus ayant une pression systolique supérieure à 120 mmHg et une pression diastolique supérieure à 90 mmHg présentent un risque nettement accru. De même, les personnes avec un cholestérol élevé ou très élevé sont proportionnellement plus touchées. L'âge constitue également un facteur déterminant : à partir de 45 ans, le risque augmente fortement pour atteindre un pic critique autour de 63 à 64 ans. Ces résultats confirment les observations cliniques selon lesquelles la combinaison de l'âge, de l'hypertension et de l'hypercholestérolémie constitue un terrain favorable au développement des maladies cardiovasculaires.

Performance des modèles de prédiction :

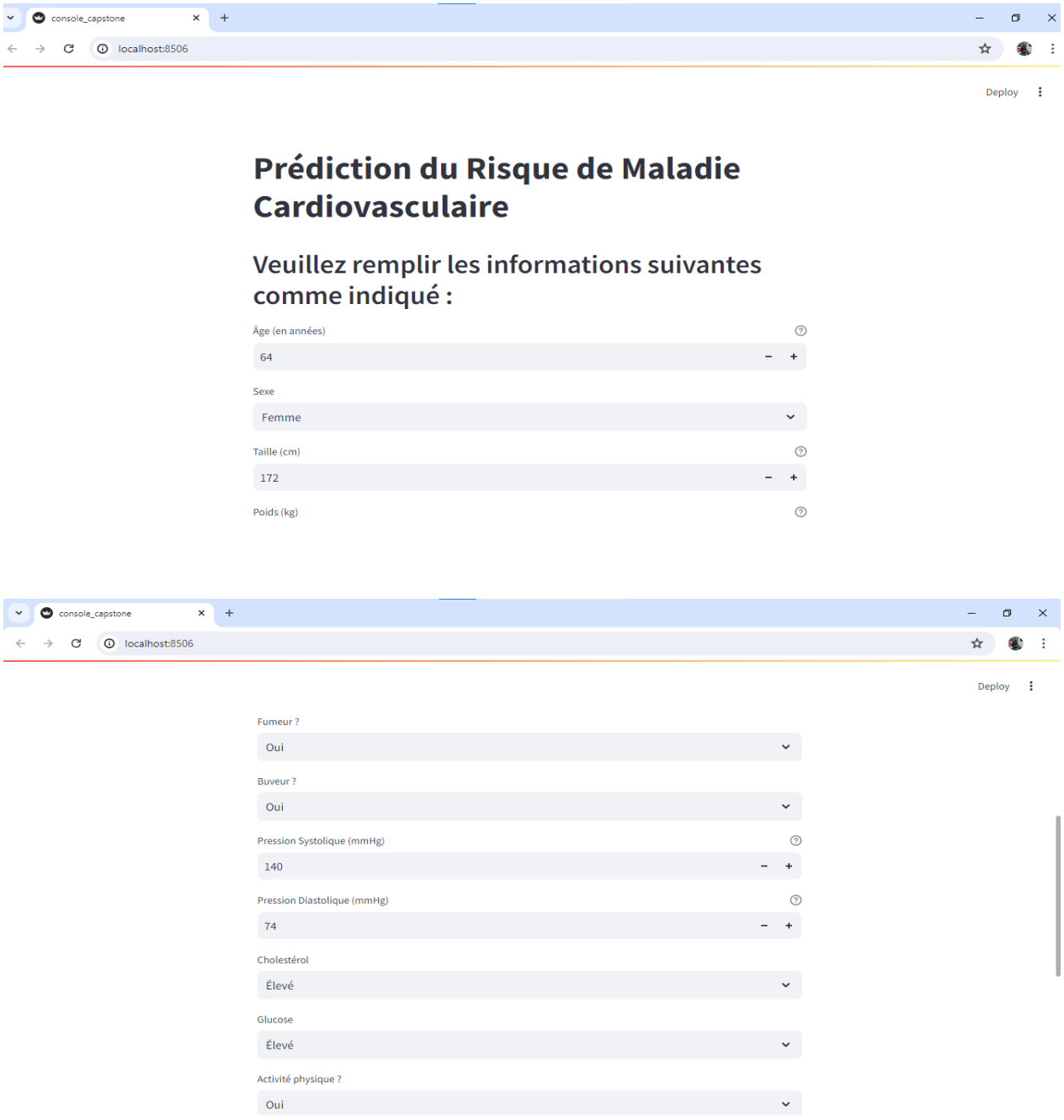
Chaque modèle a d'abord été évalué dans sa version initiale, puis optimisé à l'aide d'une recherche d'hyperparamètres (GridSearchCV). Les résultats de l'ensemble des modèles indiquent des performances globalement moyennes, avec des taux d'exactitude (accuracy) variant entre 68 % et 73.6 %. Le modèle XGBoost optimisé obtient la meilleure performance avec une accuracy sensiblement égale à 73,6 %, suivi du SVM initial (73,4 %) et du Random Forest optimisé (73,4 %). Ces scores traduisent une capacité prédictive modérée. Les valeurs de rappel (recall) pour la classe des malades se situent entre 0,67 et 0,71, ce qui signifie que près de 30 % des cas réels ne sont pas identifiés.

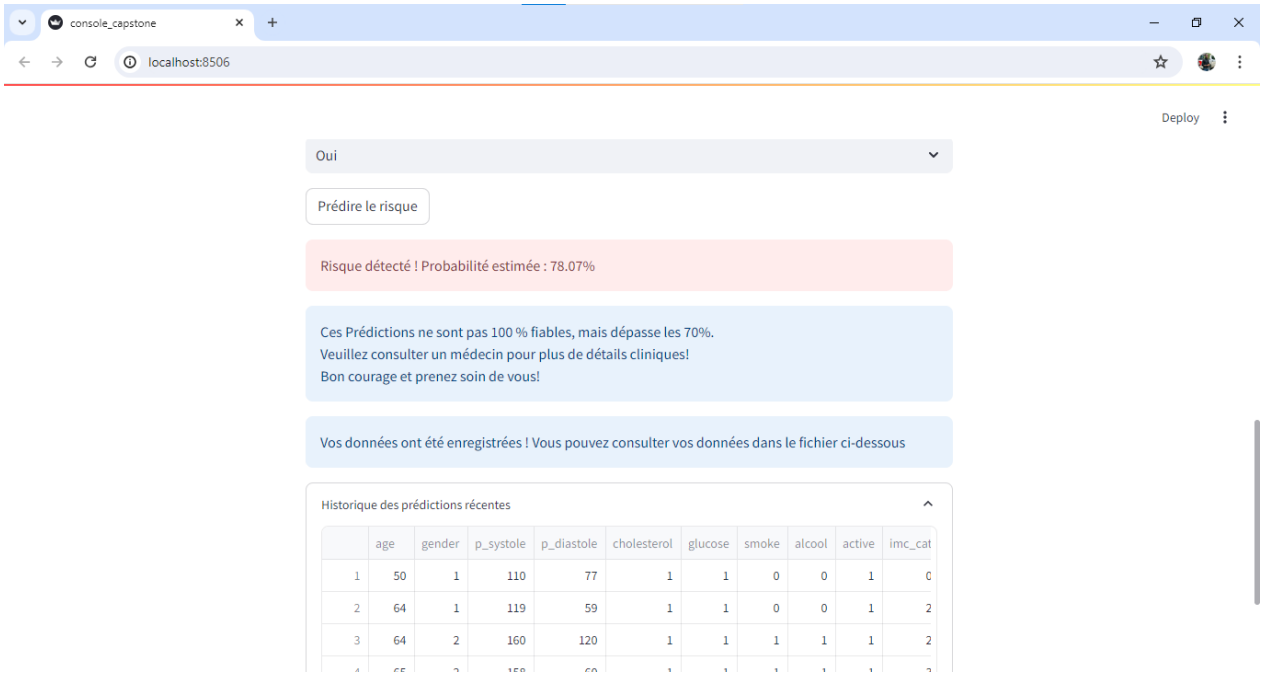
6.6. Implémentation du modèle XGBoost

Après la phase de sélection, le modèle XGBoost optimisé a été implémenté dans une application interactive développée avec Streamlit. Cette interface permet à l'utilisateur de saisir ses informations personnelles (âge, sexe, pression artérielle, cholestérol, glucose, IMC, etc.) et d'obtenir en retour une prédiction du risque de maladie cardiovasculaire, accompagnée d'une probabilité estimée. Les données de chaque prédiction sont automatiquement enregistrées dans un fichier historique afin d'assurer un suivi continu et d'évaluer les tendances.

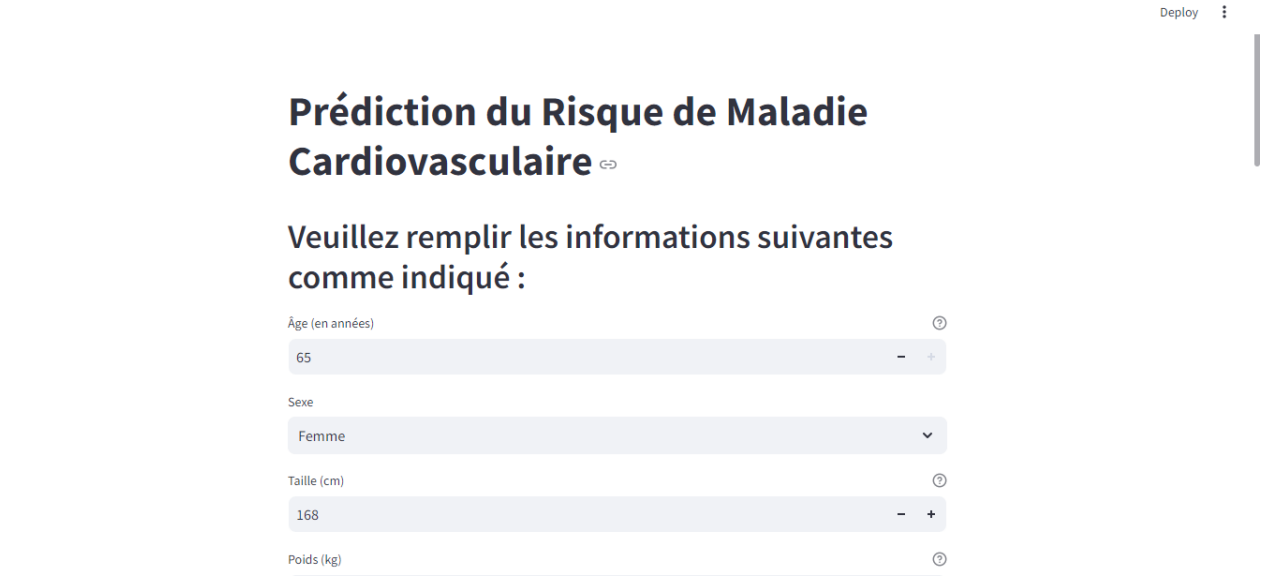
En voici deux cas de résultats présentés par le modèle :

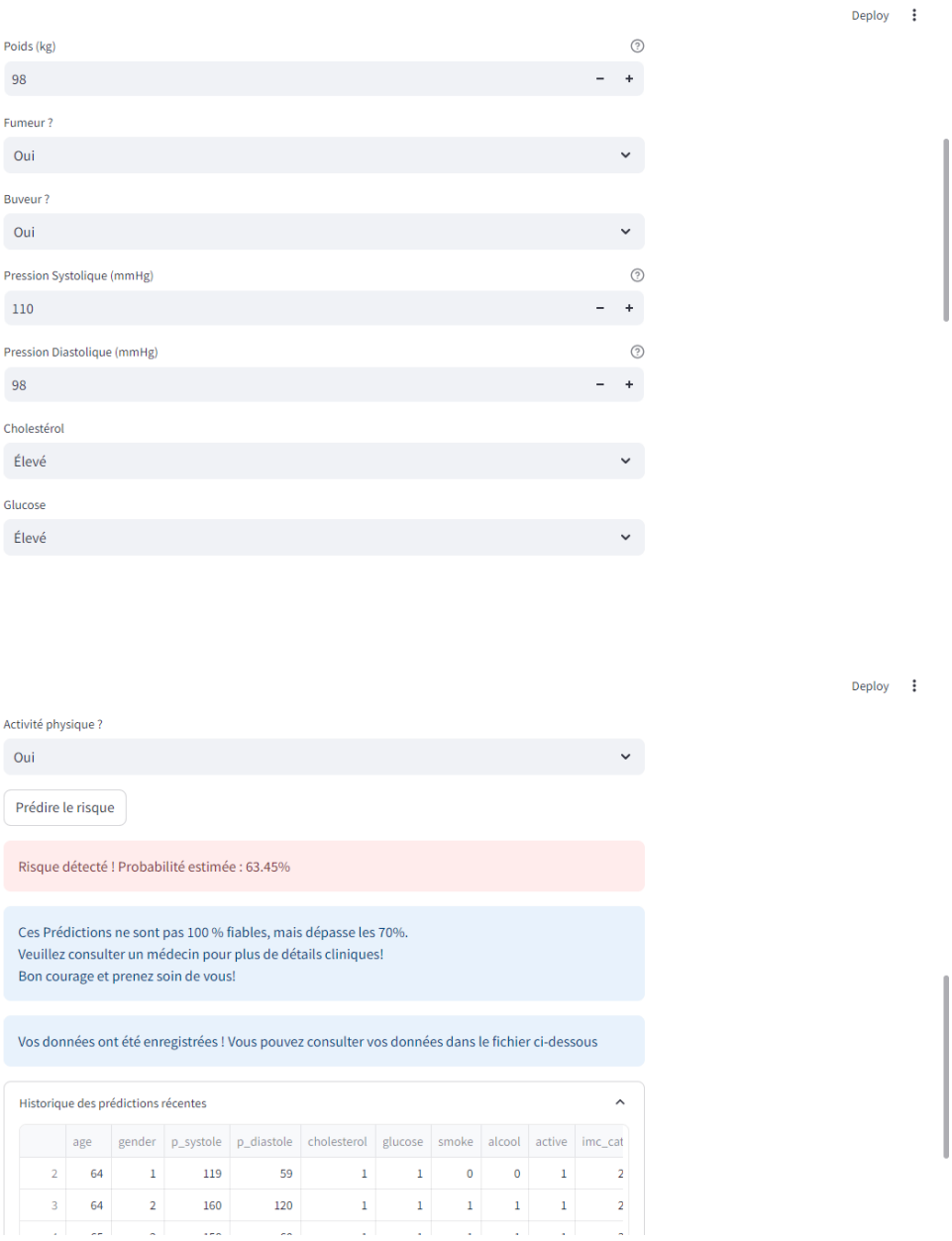
1- Cas de risque de maladies cardiovasculaires détectés


console_capstone
localhost:8506
Deploy
Prédiction du Risque de Maladie Cardiovasculaire
Veuillez remplir les informations suivantes comme indiqué :
Âge (en années)
64
Sexe
Femme
Taille (cm)
172
Poids (kg)
console_capstone
localhost:8506
Deploy
Fumeur ?
Oui
Buveur ?
Oui
Pression Systolique (mmHg)
140
Pression Diastolique (mmHg)
74
Cholestérol
Élevé
Glucose
Élevé
Activité physique ?
Oui



2- Cas de risque de maladies non-détectés





En définitive, bien que le modèle présente certaines limites, notamment l'absence de variables psychosociales telles que le stress, l'alimentation ou encore l'hérédité familiale, et qu'il affiche un rappel moyen autour de 0,69, il demeure un outil prometteur. Cette initiative, si elle est renforcée par davantage de données locales et cliniques, pourrait contribuer de manière significative à la prévention et à la gestion des maladies cardiovasculaires en Haïti, en appuyant les efforts de santé publique par une approche numérique et prédictive.

7. Conclusion et Recommandations

Cette étude a comparé six (6) modèles de classification pour prédire l'événement cible sur notre jeu de données : Logistic Regression, Decision Tree, Random Forest, XGBoost, KNN et SVM. Pour chaque modèle, nous avons évalué la performance initiale, puis testé des optimisations d'hyperparamètres via GridSearchCV afin d'améliorer la précision et la robustesse.

Les résultats obtenus indiquent des niveaux de précision plus ou moins satisfaisantes, mais démontrent que les modèles de Machine Learning testés sont capables de prédire efficacement la probabilité d'une maladie cardiovasculaire à partir de données, à condition qu'on ait pris les variables prépondérantes des maladies cardiovasculaires.

L'analyse des importances des features dans les modèles basés sur les arbres et des corrélations entre les variables indépendantes et la variable cible révèlent que la pression systolique, l'âge, la pression diastolique et le cholestérol sont les variables les plus déterminantes, en cohérence avec les connaissances médicales et la première hypothèse.

En somme, bien que certaines limites persistent , notamment l'absence de variables comme le stress, l'alimentation ou l'hérédité familiale, les hypothèses de départ sont acceptées. L'approche proposée possède un potentiel réel pour appuyer la prévention et la détection précoce des maladies cardiovasculaires. En renforçant la base de données et en intégrant de nouveaux paramètres médicaux, une telle initiative pourrait contribuer efficacement à la surveillance épidémiologique et à la santé publique en Haïti.

En conséquence, pour réduire le risque cardiovasculaire, il est conseillé aux patients de :

- Surveiller régulièrement leur tension artérielle : pression systolique et diastolique, et consulter un médecin en cas de valeurs élevées.
- Maintenir un mode de vie actif : pratiquer une activité physique régulière adaptée à leur condition.
- Adopter une alimentation équilibrée, faible en sel, en sucres et en graisses saturées, pour contrôler le cholestérol et la glycémie et le surpoids.
- Éviter ou réduire la consommation de tabac et d'alcool, car ces comportements augmentent le risque cardiovasculaire.

ENSEMBLE, LUTTONS CONTRE LES MALADIES CARDIOVASCULAIRES