Projet Support Vector Machine : Analyse et prédiction du risque de crise cardiaque sur 8 763 individus

TRILLAUD Valorys

DUPAU Jasmine

Master 2 Econométrie et Statistiques Appliquées

Encadrées par Monsieur ROUL Benjamin

Table des matières

- Introduction
- I. Analyse exploratoire
  - A. Analyse univariée
    - 1) Variable d'intérêt
    - 2) Prédicteurs quantitatifs
    - 3) Prédicteurs qualitatifs
  - B. Analyse bivariée
    - 1) Variable cible et prédicteurs qualitatifs
    - 2) Etude des corrélations
    - 3) Lien entre les variables qualitatives
- II. Modélisation
  - A. Préparation des données
    - 1) Valeurs atypiques
    - 2) Encodage
    - 3) Split du jeu de données
    - 4) Undersampling
    - 5) Standardisation
  - B. Construction des modèles
    - 1) Modélisation avec les paramètres par défaut
    - 2) Évaluation des modèles (accuracy et F1-score) et choix
      - a) Définitions des métriques
      - b) Présentation du raisonnement pour le choix du modèle final
      - c) Etat des lieux sans la validation croisée
      - d) Etat des lieux avec la validation croisée
      - e) Optimisation des modèles
      - d) Importance des variables
- III. Interprétation du meilleur modèle
  - A. Interprétation globale
    - 1) Partial Dependence Plots - PDP

        - 2) Permutation features importance
    - B. Interprétation locale
        - 1) ICE
        - 2) LIME
- Conclusion

Introduction

Aujourd'hui, les maladies cardiovasculaires figurent parmi les principales causes de mortalité dans le monde. Afin d'anticiper le risque de crise cardiaque, il est possible d'exploiter diverses informations médicales sur les patients, telles que l'âge, le sexe, la présence de diabète, le tabagisme et d'autres facteurs de santé. L'objectif de ce projet est de développer un modèle prédictif capable d'évaluer ce risque à partir de données médicales.

Dans un premier temps, nous réaliserons une analyse exploratoire des données afin de mieux comprendre leur structure et d'identifier les variables les plus pertinentes. Ensuite, nous sélectionnerons et entraînerons le meilleur modèle de prédiction avant d'en analyser les performances. Enfin, nous interpréterons les résultats en utilisant différentes techniques d'explicabilité telles que Partial Dependence Plots (PDP), Permutation Feature Importance, Individual Conditional Expectation (ICE) et LIME. Ces méthodes nous permettront d'analyser l'impact des variables (âge, cholestérol, tension artérielle, etc.) sur la probabilité de survenue d'une crise cardiaque.

L'objectif final est de concevoir un modèle de classification binaire capable de prédire si un individu présente un risque ou non de crise cardiaque, tout en garantissant une interprétabilité claire des prédictions.

I. Analyse exploratoire

> *Avant de procéder à l'analyse exploratoire, nous avons fait un premier nettoyage de la base qui a consisté à :*
>
> *1. Renommer les colonnes*
> *2. Manipuler les colonnes*
> *3. Observer les valeurs manquantes : il n'y en avait pas*
> *4. Mettre les données au bon format*

Pour cette analyse, nous avons utilisé le jeu de données "Heart Attack Risk Prediction Dataset. La variable cible que nous cherchons à prédire est "Heart Attack Risk", qui indique la présence d'un risque de crise cardiaque (1 : Oui, 0 : Non).

Le jeu de données est disponible sur Kaggle.

Ce jeu de données, généré à l'aide de ChatGPT, comprend 8 763 observations et 26 variables, dont la variable cible. Le tableau suivant détaille les variables explicatives utilisées dans notre analyse :

Table N°1 : Tableau descriptif des variables explicatives

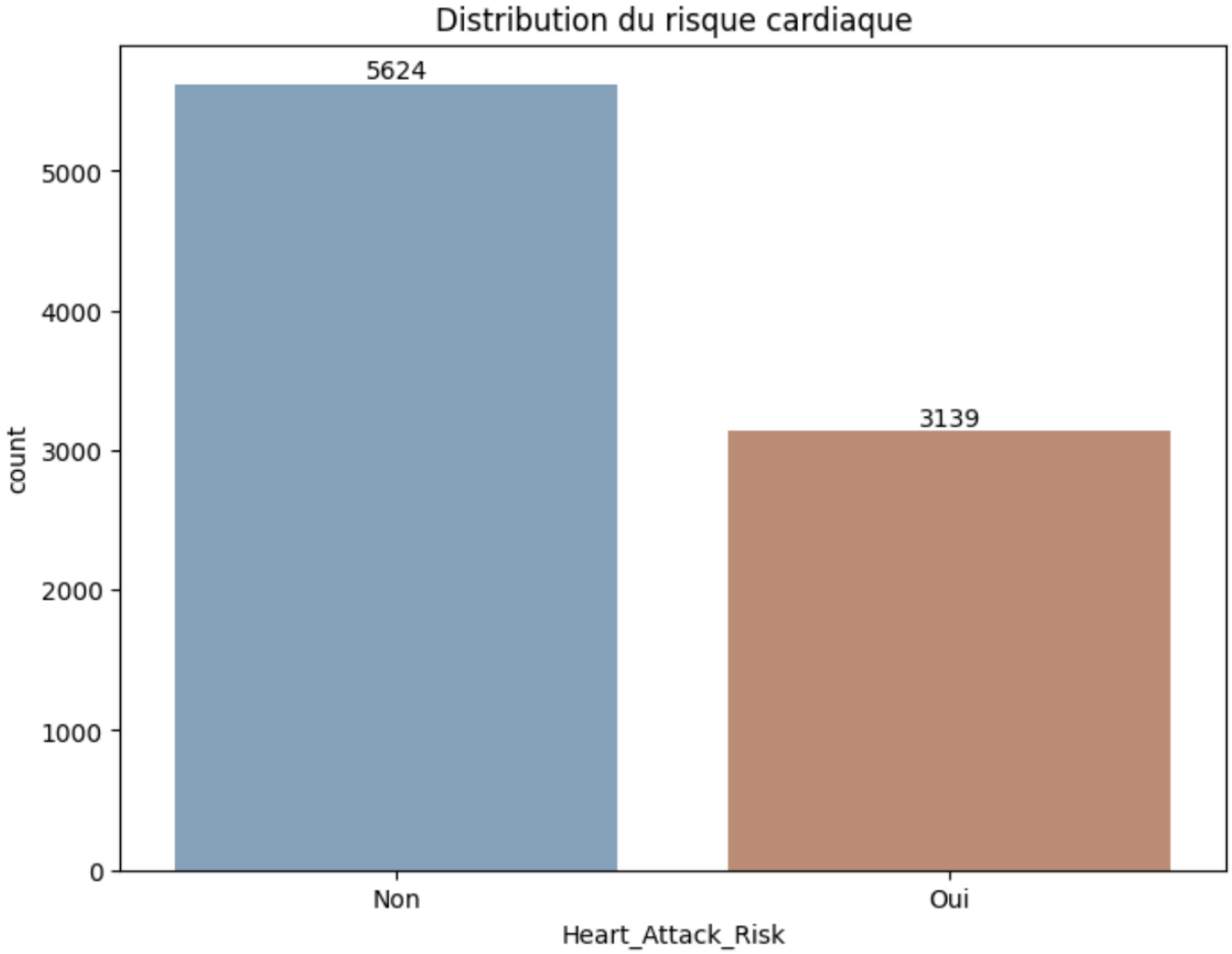
Variables	Description
ID patient	Identifiant unique pour chaque patient
Âge	Âge du patient
Sexe	Genre du patient (Homme/Femme)
Cholestérol	Taux de cholestérol du patient
Pression artérielle	Pression artérielle du patient (systolique/diastolique)
Fréquence cardiaque	Fréquence cardiaque du patient
Diabète	Si le patient est diabétique (Oui/Non)
Antécédents familiaux	Antécédents familiaux de problèmes cardiaques (1 : Oui, 0 : Non)
Tabagisme	Statut tabagique du patient (1 : Fumeur, 0 : Non-fumeur)
Obésité	Statut d'obésité du patient (1 : Obèse, 0 : Non obèse)
Consommation d'alcool	Niveau de consommation d'alcool du patient (Aucun/Léger/Modéré/Fort)
Heures d'exercice par semaine	Nombre d'heures d'exercice par semaine
Régime alimentaire	Habitudes alimentaires du patient (saines/moyennes/malsaines)
Problèmes cardiaques antérieurs	Problèmes cardiaques antérieurs du patient (1 : Oui, 0 : Non)
Utilisation de médicaments	Utilisation de médicaments par le patient (1 : Oui, 0 : Non)
Niveau de stress	Niveau de stress rapporté par le patient (1-10)
Heures sédentaires par jour	Heures d'activité sédentaire par jour
Revenu	Niveau de revenu du patient
IMC	Indice de masse corporelle (IMC) du patient
Triglycérides	Taux de triglycérides du patient
Heures de sommeil par jour	Heures de sommeil par jour
Pays du patient	Pays du patient
Continent	Continent où réside le patient
Hémisphère	Hémisphère où réside le patient

Source : Dossier SVM, Valorys Trillaud et Jasmine Dupau

A. Analyse univariée

1) Variable d'intérêt

Figure n°1 : Répartition de la variable du risque cardiaque

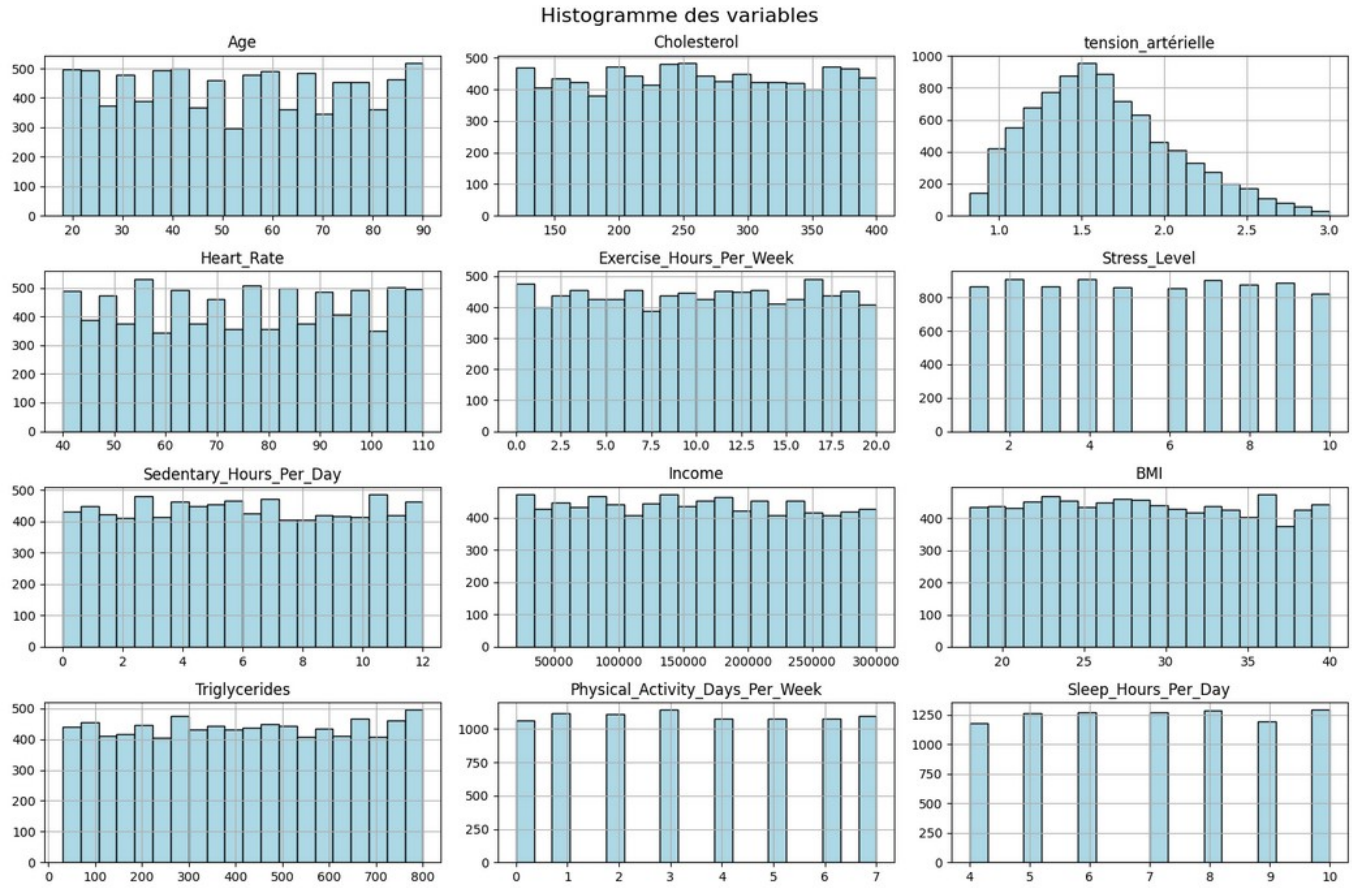


Source : Dossier SVM, Valorys Trillaud et Jasmine Dupau

La représentation graphique du risque de crise cardiaque met en évidence un déséquilibre dans les données : 3 139 individus présentent un risque de crise cardiaque, tandis que 5 624 n'en présentent pas. Ainsi, seuls 35,82 % des observations appartiennent à la classe à risque. Ce déséquilibre peut impacter la performance des modèles prédictifs, en les biaisant en faveur de la classe majoritaire. Pour garantir une meilleure capacité de généralisation, il sera donc nécessaire d'appliquer une technique de rééchantillonnage avant de procéder à la modélisation.

2) Prédicteurs quantitatifs

Figure n°2: Distribution des variables quantitatives

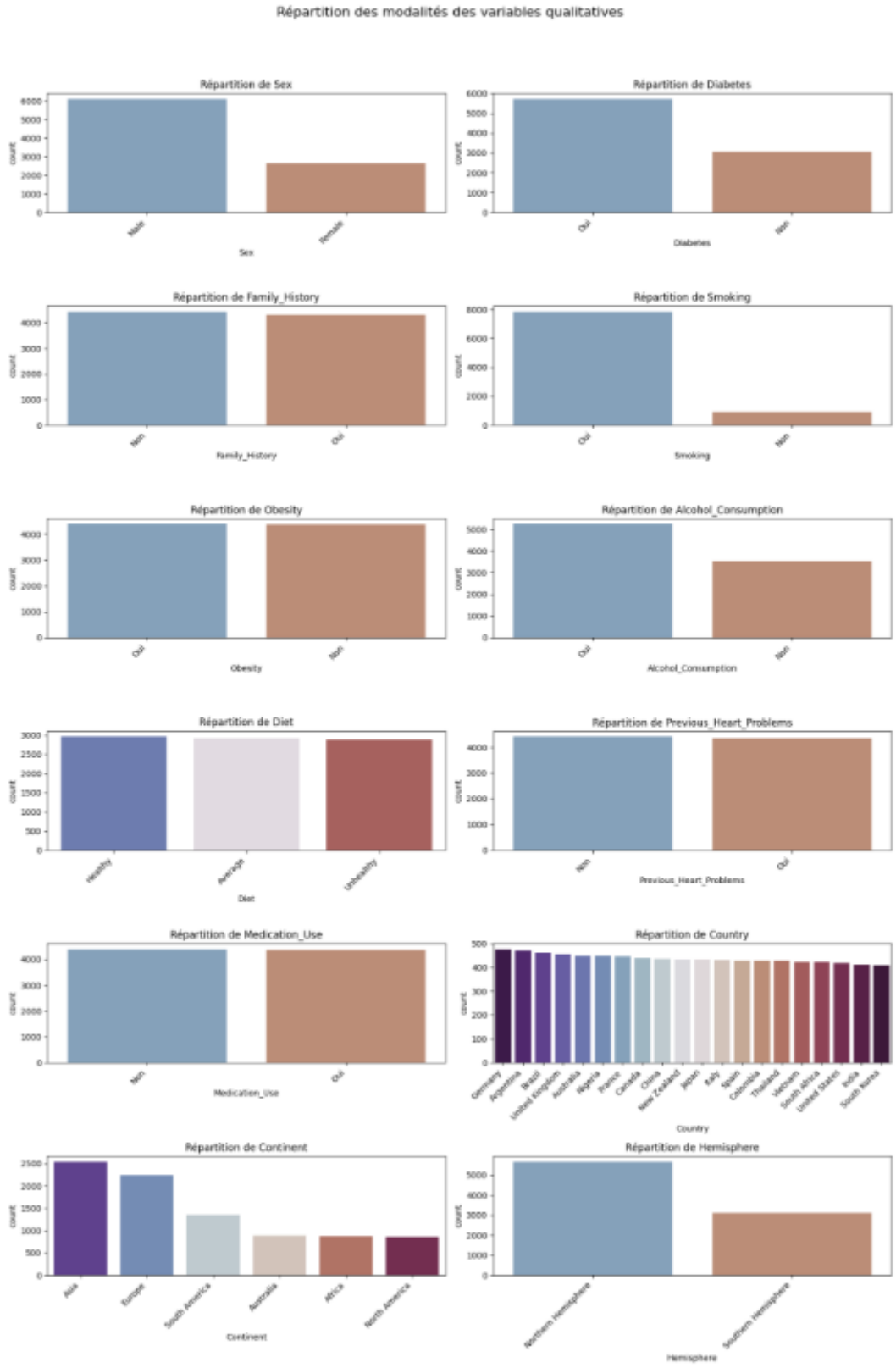


Source : Dossier SVM, Valorys Trillaud et Jasmine Dupau

Les différentes variables présentent globalement une distribution équilibrée. Toutefois, la tension artérielle se distingue par une distribution qui se rapproche davantage d'une loi normale.

3) Prédicteurs qualitatifs

Figure n°3: Distribution des variables qualitatives



Source : Dossier SVM, Valorys Trillaud et Jasmine Dupau

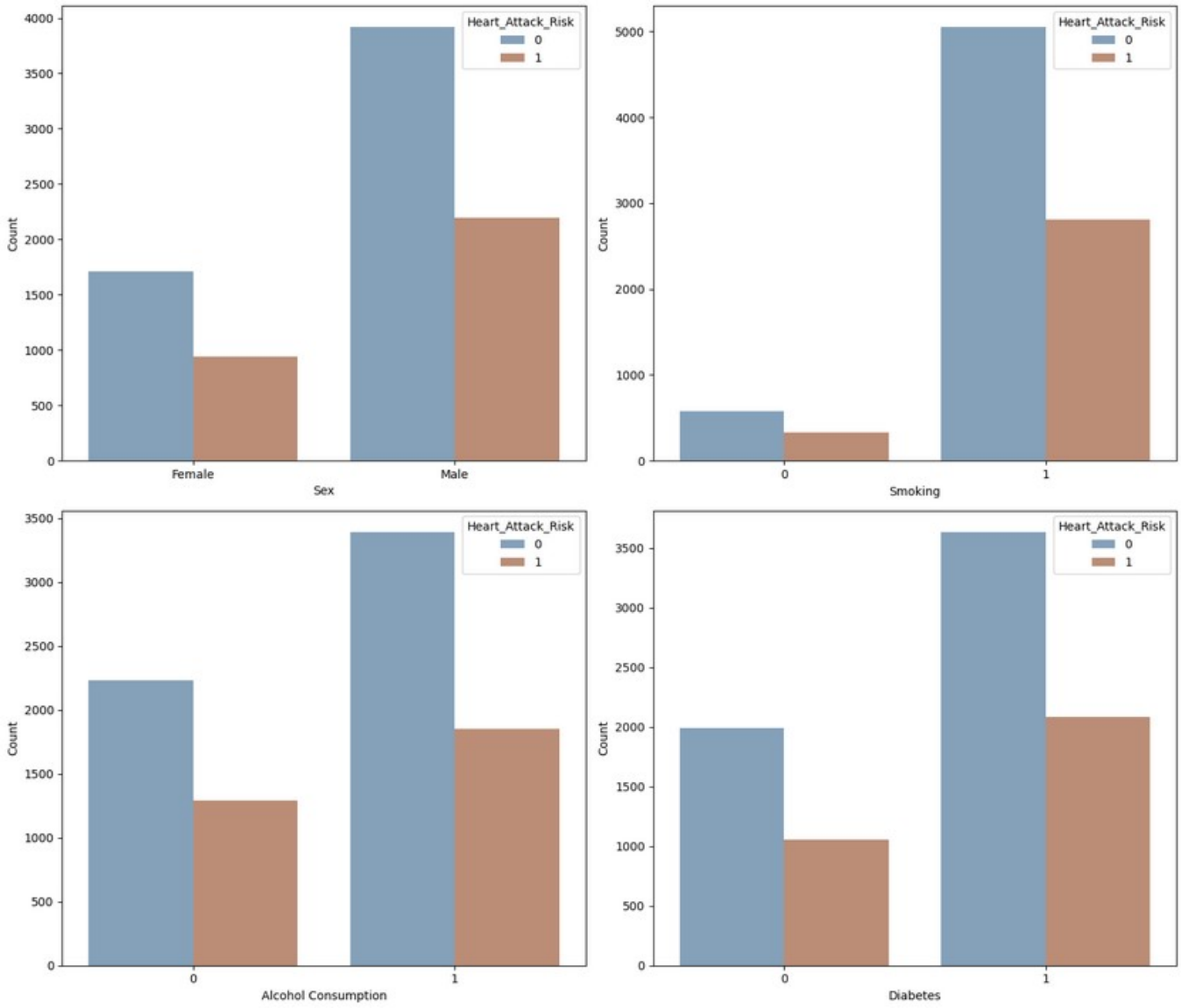
Ces diagrammes en bâtons permettent d'analyser les principales caractéristiques des patients de notre échantillon. Nous observons que la majorité des individus sont des hommes, originaires de pays asiatiques et européens de l'hémisphère nord. De plus, une proportion importante des patients sont fumeurs, diabétiques et consomment de l'alcool. Pour les autres caractéristiques, la répartition semble plus équilibrée, sans déséquilibre majeur.

B. Analyse bivariée

1) Variable cible et prédicteurs qualitatifs

Au regard du grand nombre de prédicteurs qualitatifs, nous avons choisi de croiser notre variable d'intérêt avec les facteurs qui sont les plus susceptibles de causer un risque cardiaque (le genre, le fait de fumer, la consommation d'alcool et le diabète).

Figure n°4: Distribution croisée entre notre variable d'intérêt et des prédicteurs qualitatifs

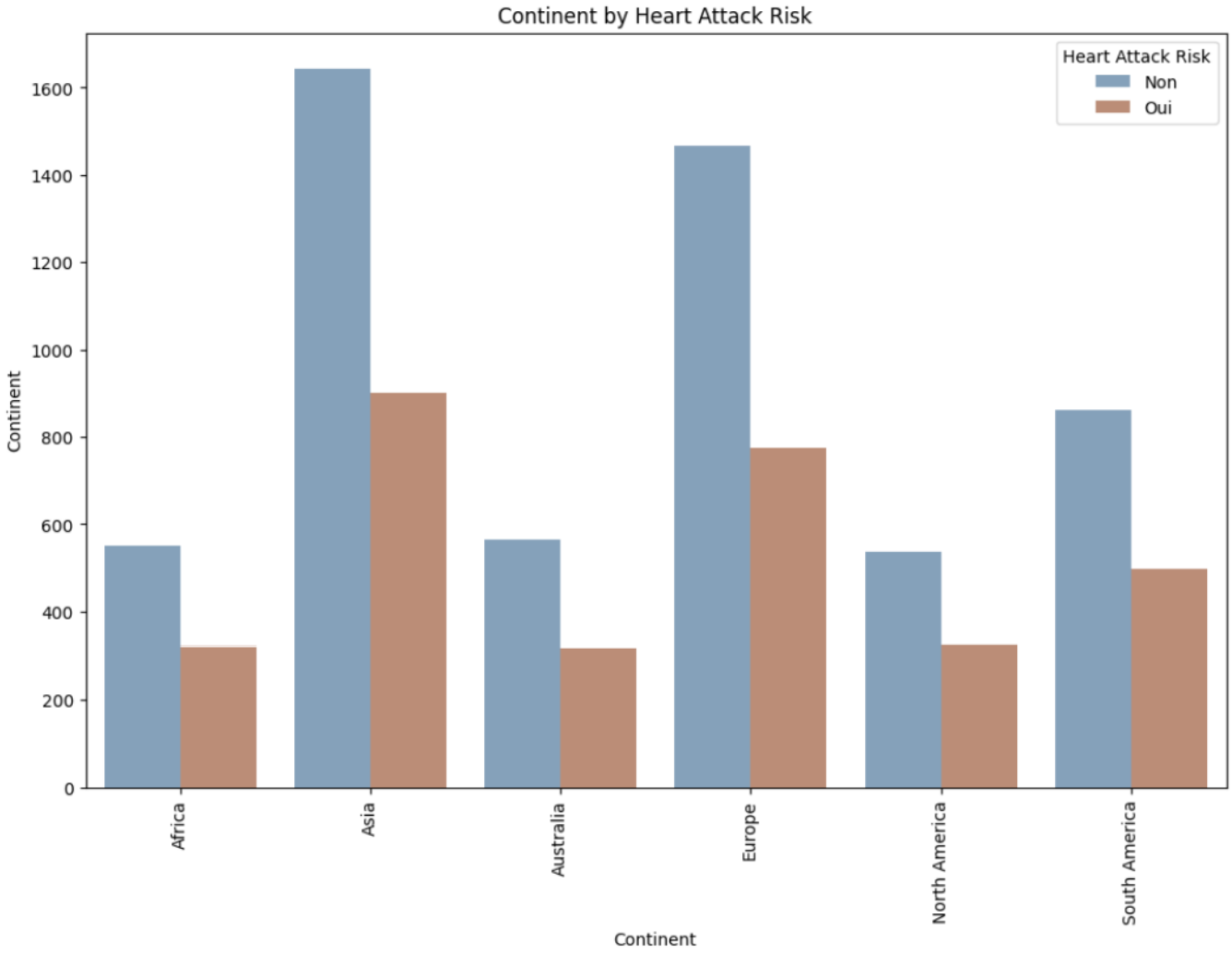


Source : Dossier SVM, Valorys Trillaud et Jasmine Dupau

Les graphiques permettent d'observer la distribution du risque de crise cardiaque en fonction de différentes caractéristiques des patients. Nous pouvons ainsi observer que les hommes diabétiques qui boivent et qui fument augmentent les risques d'être victime d'une crise cardiaque.

En supplément, nous avons également croiser notre variable d'intérêt avec celle du continent. La Figure n°5 nous montre que les individus de notre échantillon proviennent principalement de l'Asie et de l'Europe.

Figure n°5: Distribution croisée entre notre variable d'intérêt et le continent.

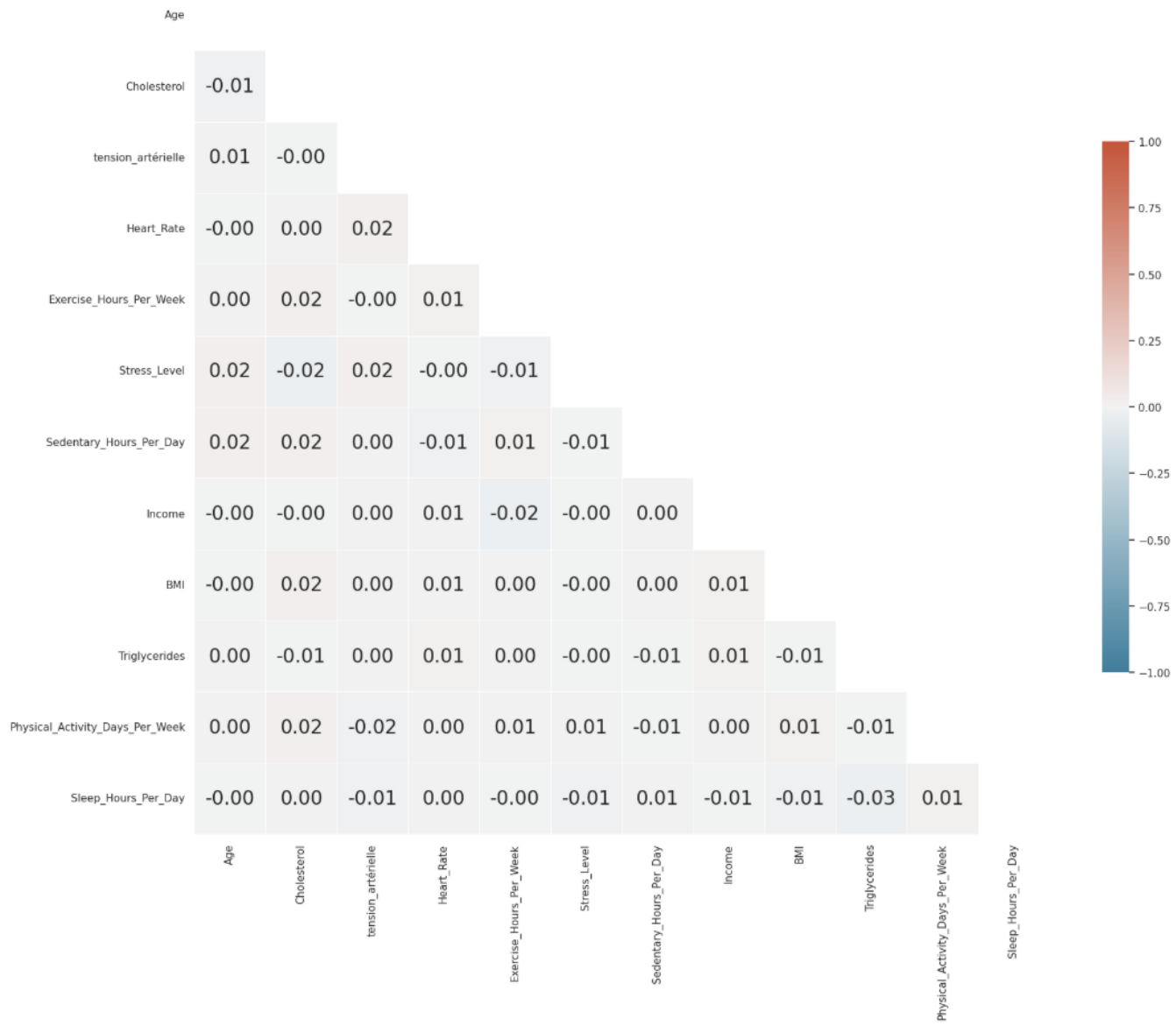


Source : Dossier SVM, Valorys Trillaud et Jasmine Dupau

2) Etude des corrélations

Après avoir calculé la matrice des corrélations de la Figure n°5, nous constatons que les variables quantitatives présentent peu de corrélations entre elles, ce qui indique une faible redondance dans ces données.

Figure n°5: Matrice des corrélations



Source : Dossier SVM, Valorys Trillaud et Jasmine Dupau

3) Lien entre les variables qualitatives

Pour clôturer l'analyse exploratoire, nous allons vérifier si les variables qualitatives sont liées entre elles. Pour ce faire, nous utilisons le test du Chi2 avec la fonction "chi2_contingency".

Figure n°6: Résultats du test du Chi2

Tableau des résultats de dépendance entre les variables (seuil=0.05) :

	Sex	Diabetes	Family_History	Smoking	Obesity	Alcohol_Consumption	Diet	Previous_Heart_Problems	Medication_Use	Country	Continent	Hemisphere	Heart_Attack_Risk
Sex	NaN	Non	Non	Oui	Non	Non	Non	Non	Non	Non	Non	Non	Non
Diabetes	Non	NaN	Non	Non	Non	Non	Non	Non	Non	Non	Non	Non	Non
Family_History	Non	Non	NaN	Non	Non	Non	Non	Non	Non	Non	Non	Non	Non
Smoking	Oui	Non	Non	NaN	Non	Non	Non	Non	Non	Non	Non	Non	Non
Obesity	Non	Non	Non	Non	NaN	Oui	Non	Non	Non	Non	Non	Non	Non
Alcohol_Consumption	Non	Non	Non	Non	Oui	NaN	Non	Non	Non	Non	Non	Non	Non
Diet	Non	Non	Non	Non	Non	Non	NaN	Non	Non	Non	Non	Non	Non
Previous_Heart_Problems	Non	Non	Non	Non	Non	Non	Non	NaN	Non	Non	Non	Non	Non
Medication_Use	Non	Non	Non	Non	Non	Non	Non	Non	NaN	Non	Non	Non	Non
Country	Non	Non	Non	Non	Non	Non	Non	Non	Non	NaN	Oui	Oui	Non
Continent	Non	Non	Non	Non	Non	Non	Non	Non	Non	Oui	NaN	Oui	Non
Hemisphere	Non	Non	Non	Non	Non	Non	Non	Non	Non	Oui	Oui	NaN	Non
Heart_Attack_Risk	Non	Non	Non	Non	Non	Non	Non	Non	Non	Non	Non	Non	NaN

Source : Dossier SVM, Valorys Trillaud et Jasmine Dupau

Les résultats du test montrent que la variable cible n'est pas non plus liée avec les prédicteurs qualitatifs. En revanche, nous voyons que les variables géographiques sont liées entre elles. Afin d'éviter la redondance et pour simplifier le modèle, nous avons décidé de supprimer les variables "pays" et "hémisphère" puisque nous jugeons que l'information fournie par le continent est suffisante pour notre analyse. De plus, la consommation d'alcool est liée à l'obsésité et le fait de fumer est lié au genre. Malgré ces liens, nous décidons de les conserver car elles sont pertinentes pour la prédiction du risque cardiaque.

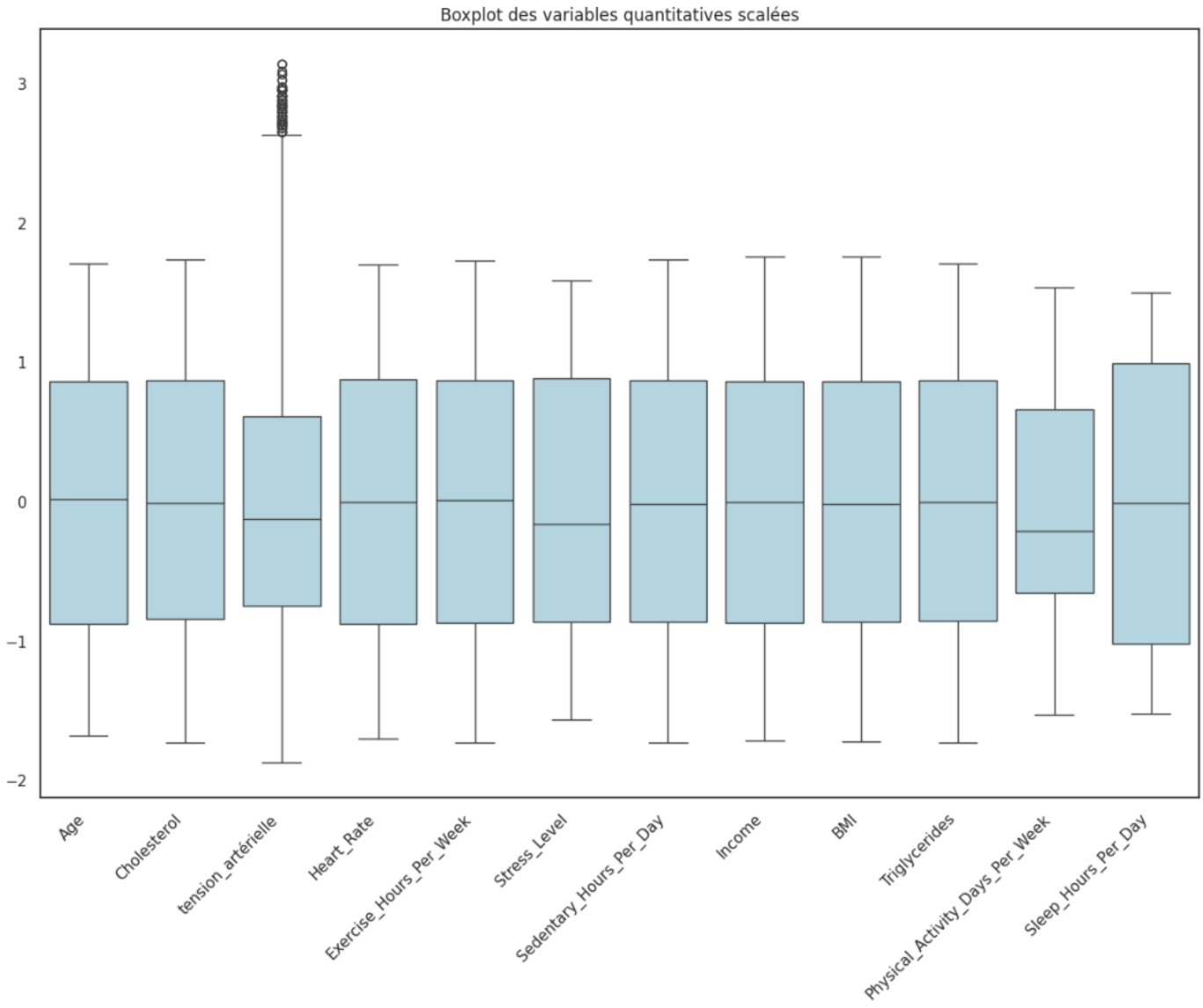
II. Modélisation

A. Préparation des données

1) Valeurs atypiques

Afin de nous assurer qu'il n'y a pas de valeurs atypiques dans nos données, nous avons représenté la distribution des variables à l'aide de boîtes à moustaches.

Figure n°7: Boxplots des variables quantitatives



Source : Dossier SVM, Valorys Trillaud et Jasmine Dupau

Nous constatons que seule la tension artérielle présente des points en dehors de sa boîte. Pour vérifier si ces points sont réellement atypiques, nous avons appliqué le test ESD. Ce test n'a pas détecté de valeurs atypiques significatives. Ainsi, les points observés sur le boxplot ne sont pas des valeurs aberrantes. En conclusion, il n'est donc pas nécessaire de procéder à une winsorisation des données.

2) Encodage

Dans notre analyse, nous avons des variables quantitatives qui ne nécessitent aucune modification. Cependant, nous avons aussi des variables qualitatives qui ne peuvent pas être utilisées telles quelles pour la modélisation. Les variables binaires comme le diabète, le tabagisme ou encore l'obésité, ont été transformées en variables booléennes (0/1). Il restait alors à transformer les variables sexe, régime alimentaire et continent. Pour cela, nous avons utilisé la fonction `LabelEncoder` qui permet de convertir ces variables catégorielles en variables numériques adaptées à notre modèle.

3) Split du jeu de données

Nous procédons au split du jeu de données, une étape essentielle qui consiste à diviser le jeu de données en deux sous-ensembles : un pour l'entraînement du modèle et un autre pour l'évaluation de ses performances. Cette séparation est cruciale pour éviter le surapprentissage et garantir que le modèle ait testé sur des

données qu'il n'a jamais vues auparavant. Cette approche permet de vérifier si le modèle généralise bien, c'est-à-dire s'il peut faire des prédictions précises sur des données inconnues.

4) Undersampling

Nous appliquons la technique d'undersampling pour rééquilibrer nos données, notamment afin de résoudre le problème de déséquilibre des classes. Dans l'undersampling, nous réduisons la taille de la classe majoritaire en supprimant une partie de ses observations, ce qui permet d'atteindre un équilibre entre les classes. Cette approche empêche le modèle de favoriser systématiquement la classe majoritaire et améliore ainsi la précision des prédictions pour la classe minoritaire.

Il est à noter que nous avons également testé la méthode SMOTE, conistant à créer artificiellement de nouveaux cas rares. Mais, celle-ci n'a pas permis d'améliorer nos résultats, voire les a empirés.

5) Standardisation

Enfin, nous procédons à la standardisation des variables quantitatives, à l'exception des variables binaires et des variables encodées avec le `LabelEncoder`. La standardisation permet de rendre les variables comparables en les mettant sur une même échelle.

Une fois la préparation des données réalisée, nous avons pu poursuivre sur la construction des modèles.

B. Construction des modèles

1) Modélisation avec les paramètres par défaut

Pour commencer, nous avons comparé plusieurs modèles en utilisant leurs paramètres par défaut afin d'identifier le modèle le plus performant. Nous avons ainsi testé les algorithmes ensemblistes (Random Forest, Gradient Boosting et XGBoost), la régression logistique ainsi que les algorithmes SVM (SVC linéaire, classifieur SGD, SVM avec noyau linéaire, SVM avec noyau RBF et SDG avec noyau polynomial).

Les modèles SVM consistent à séparer les classes avec la plus grande marge possible et sont classés en 2 catégories :

- **Les linéaires** :

  Lorsque les données sont linéairement séparables, nous utilisons "le SVM linéaire" et le "SGDClassifier" qui se distinguent par l'optimiseur appliqué pour la mise en oeuvre de la classification. En effet, le premier est aussi connu sous le nom de *soft margin classification* avec le paramètre de régularisation C, tandis que le deuxième est basé sur la descente de gradient.

- **Les non linéaires** :

  Lorsque les données ne sont pas linéairement séparables, nous pouvons transformer les features avec la fonction `PolynomialFeatures`. Cependant, cette approche n'est pas recommandée en cas d'un gros volume de données. Par conséquent, l'approche la plus fiable est de recourir aux SVM avec noyau (ou kernel trick). Ces derniers permettent permet d'avoir des modèles complexes adaptés aux grands datasets.

2) Évaluation des modèles (accuracy et F1-score) et choix

a) Définitions des métriques

Pour évaluer la qualité et la performance des modèles, nous nous sommes appuyés sur les deux métriques suivantes : l'accuracy et le F1-score.

• **Mesure d'exactitude (accuracy) et taux d'erreur :**

La précision correspond à la proportion des classifications bien prédites (vrais positifs et vrais négatifs) et répond donc à la question : "À quelle fréquence le modèle est-il correct ?". Le taux d'erreur permet de mesurer à quel point le modèle ne s'adapte pas bien aux données. Ainsi, **un bon modèle a une accuracy proche de 1 et un taux d'erreur proche de 0.**

• **F1-score :**

Il permet d'évaluer la performance d'un modèle de classification et est particulièrement **préféré à l'accuray lorsque la variable cible est déséquilibrée**. Il correspond à la moyenne de la précision (évite les faux positifs) et du recall (évite les faux négatifs). Ainsi, un bon modèle a un score proche de 1.

b) Présentation du raisonnement pour le choix du modèle final

Dans le contexte d'un rééchantillonnage avec la méthode de *l'Undersampling* comme ici, le gros risque auquel tout développeur fait face est **l'overfitting** (ou le sur-ajustement). En effet, la réduction de la classe majoritaire contraint le modèle à apprendre sur moins d'observations et augmente donc le risque d'obtenir des modèles qui s'adaptent très bien voire trop sur les données d'entraînement. Cela aura alors pour conséquence de mauvaises performances sur les données du jeu test.

Ce sur-ajustement est détectable au regard de l'écart entre les métriques des deux bases (jeu train et jeu tes) et principalement, celui du F1-score. Par conséquent, nous choisirons le modèle qui réduira au mieux cet écart pour avoir le moins possible de sur-ajsutement même s'il peut sembler être le moins performant par rapport aux autres modèles. Le sur-ajustement est considéré comme acceptable lorsque l'écart entre les performances sur les jeux d'entraînement et de test reste inférieur à 5-6%.

Pour identifier le modèle qui s'adapte le mieux à nos données, nous allons d'abord comparer les différents modèles sans puis avec la cross-validation. La première analyse va nous permettre d'avoir un premier état des lieux de la présence d'overfitting. Ce dernier nous aiguillera sur les modèles à conserver après la validation croisée. Une fois les modèles sélectionnés, nous les tunerons pour améliorer leurs performances, puis ferons une dernière comparaison afin de choisir le modèle final.

c) Etat des lieux sans la validation croisée

Tableau N°2 : Indicateurs de performances sans la cross-validation

Modèle	Métrique	Train	Test	Écart
Régression logistique	Accuracy	0.53	0.48	5%
	Taux d'erreur	0.47	0.52	-5%
	F1-score	0.53	0.41	12%
Linear SVC	Accuracy	0.53	0.48	5%

Modèle	Métrique	Train	Test	Écart
	Taux d'erreur	0.47	0.52	19%
	F1-score	0.53	0.40	13%
SGD Classifier	Accuracy	0.52	0.54	-2%
	Taux d'erreur	0.48	0.46	2%
	F1-score	0.41	0.33	8%
SVM (Linear)	Accuracy	0.53	0.47	6%
	Taux d'erreur	0.47	0.53	-6%
	F1-score	0.52	0.39	13%
SVM (RBF)	Accuracy	0.65	0.50	15%
	Taux d'erreur	0.35	0.50	-15%
	F1-score	0.66	0.43	**23%**
SVM (Poly)	Accuracy	0.67	0.51	16%
	Taux d'erreur	0.33	0.49	-16%
	F1-score	0.68	0.43	**25%**
Random Forest	Accuracy	1.00	0.50	50%
	Taux d'erreur	0.00	0.50	50%
	F1-score	1.00	0.40	**60%**
Gradient Boosting	Accuracy	0.70	0.49	21%
	Taux d'erreur	0.30	0.51	-21%
	F1-score	0.70	0.43	**27%**
XGBoost	Accuracy	1.00	0.47	53%
	Taux d'erreur	0.00	0.53	-53%
	F1-score	1.00	0.40	**60%**

Source : Dossier SVM, Valorys Trillaud et Jasmine Dupau

Ce tableau récapitule les indicateurs de performance de tous les modèles. Nous observons que les modèles ensemblistes ainsi que les SVM avec noyau RBF et Poly sont très overfittés. Par conséquent, nous les éliminons pour la suite car la cross-validation et le tuning ne vont rien changer car les écarts sont trop élevés.

d) Etat des lieux avec la validation croisée

Tableau N°3 : Classement avec undersampling avec la CV

Rang	Modèle	Accuracy moyenne	Std
1	SVM (linear)	50,80%	0,0158
2	Linear SVC	50,79%	0,0137
3	Reg logistique	50,71%	0,0136
4	SGD Classifier	49,96%	0,0114

Source : Dossier SVM, Valorys Trillaud et Jasmine Dupau

Le tableau ci-dessus nous montre l'accuracy moyenne après une validation croisée de 5 folds des 4 modèless conservés.Le modèle ayant obtenu la meilleure performance selon cette métrique est le SVM avec noyau linéaire, avec une moyenne de 0,51 sur les 5 folds. Cependant, il ressort comme étant le plus instable. Par conséquent, nous allons nous intéresser à l'ajustement des modèles au jeu test après leur optimisation.

e) Optimisation des modèles

Une fois les modèles identifiés, nous avons procédé à leur optimisation à l'aide d'une recherche par grille (GridSearch) dont les paramètres sont affichés dans le tableau suivant :

Tableau N°4 : Combinaisons optimales trouvées des modèles

Modèle	Paramètres
SVM (linear)	`{'C': 1, 'kernel': 'linear'}`
Linear SVC	`{'C': 1, 'class_weight': 'balanced', 'loss': 'squared_hinge', 'max_iter': 1000, 'penalty': 'l2'}`
Régression logistique	`{'C': 0.001, 'class_weight': None, 'max_iter': 100, 'penalty': 'l2', 'solver': 'liblinear'}`
SGD Classifier	`{'alpha': 0.001, 'learning_rate': 'optimal', 'loss': 'hinge', 'max_iter': 1000, 'penalty': 'l2'}`

Source : Dossier SVM, Valorys Trillaud et Jasmine Dupau

Tableau N°5 : Indicateurs de performances après l'optimisation et la cross-validation

Modèle	Métrique	Train	Test	Écart	Écart avant CV
SVM (Linear)	Accuracy	0,53	0,48	5%	
	Taux d'erreur	0,47	0,52	-5%	
	F1-score	0,52	0,39	13%	13%
Linear SVC	Accuracy	0,53	0,49	4%	
	Taux d'erreur	0,47	0,51	-4%	
	F1-score	0,53	0,40	13%	13%

Modèle	Métrique	Train	Test	Écart	Écart avant CV
Régression logistique	Accuracy	0,53	0,51	2%	
	Taux d'erreur	0,47	0,49	-2%	
	F1-score	0,50	0,40	10%	12%
SGD Classifier	Accuracy	0,51	0,56	-5%	
	Taux d'erreur	0,49	0,44	5%	
	F1-score	0,36	0,30	**6%**	**8%**

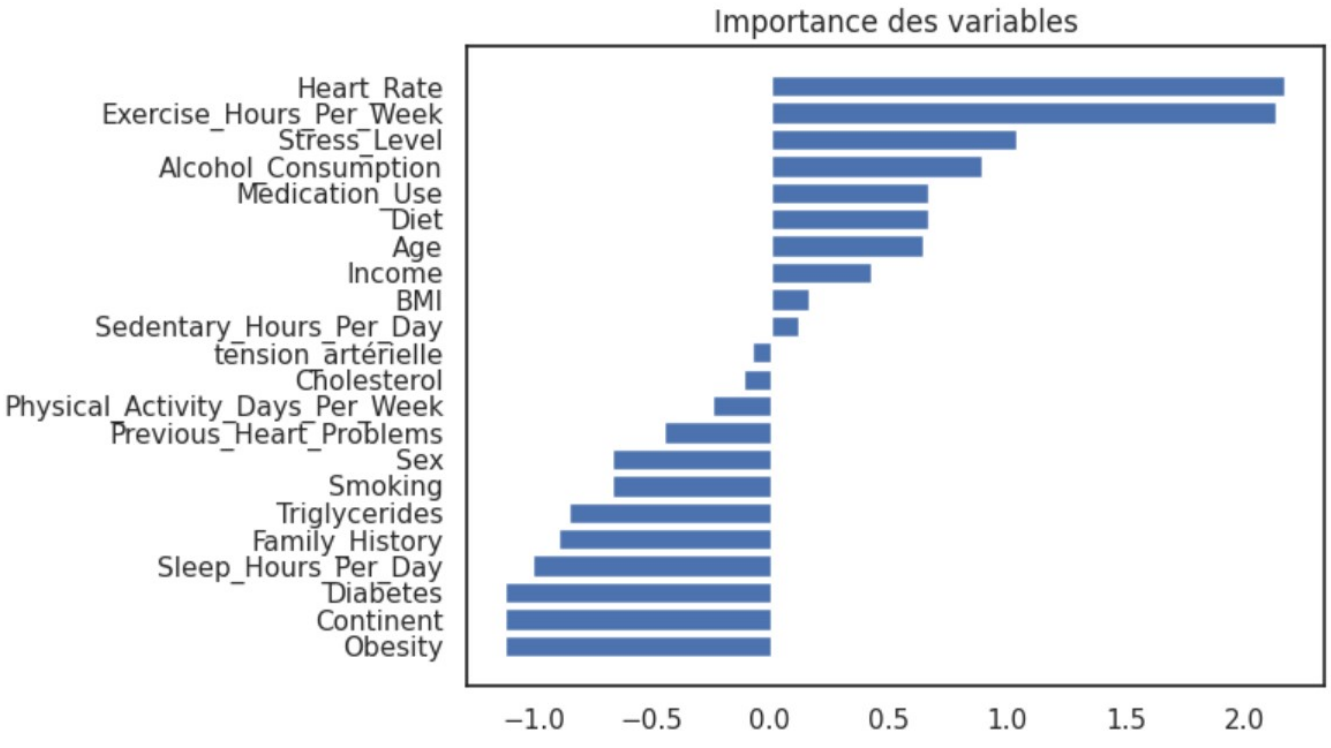
Source : Dossier SVM, Valorys Trillaud et Jasmine Dupau

Ce tableau récapitule les indicateurs de performance de tous les modèles optimisés. Notre choix s'est ainsi porté sur le **SGD Classifier**. En effet, bien que ses performances en termes de F1-score soient légèrement inférieures à celles d'autres modèles, il présente l'un des **écarts les plus faibles entre les jeux d'entraînement et de test**, avec seulement 6% d'écart pour le F1-score. De plus, il ressortait comme le plus stable lors de la validation croisée. Ce choix permet donc d'assurer une prédiction plus fiable sur des données à notre disposition.

d) Importance des variables

Avant d'entamer la dernière partie du rapport qui sera consacrée à l'interprétabilité du modèle, nous trouvions intéressant de représenter graphiquement l'importance des variables.

Figure n°8 : Importance des variables selon le modèle SGD Classifier



Source : Dossier SVM, Valorys Trillaud et Jasmine Dupau

Ce graphique présente l'importance des variables pour prédire le risque cardiaque. Les variables *Heart_rate* ainsi que *Exercise_Hours_Per_Week* sont celles qui discrétisent le mieux le risque cardiaque. Nous observons des résultats contre-intuitifs comme les variables *Obesity*, *Diabetes* et *Diabetes* qui contribuent négativement à la probabilité d'avoir un risque cardiaque. Enfin, les variables *tension_artérielle*, *Sedentary_Hours_Per_Day* et *BMI* ont très peu d'importance dans ce modèle.

Pour aller plus loin dans l'interprétation de ce modèle, nous allons faire appel à des méthodes d'interprétation créées pour rendre les modèles de machine learning, qualifiés de "Black Box", plus compréhensibles. Ces méthodes sont classées en deux catégories :

1. **Approche globale = Interprétabilité :** Compréhension générale du modèle et comment sont réalisées les prédictions
2. **Approche locale = Explicabilité :** Explication de prédiction individuelle

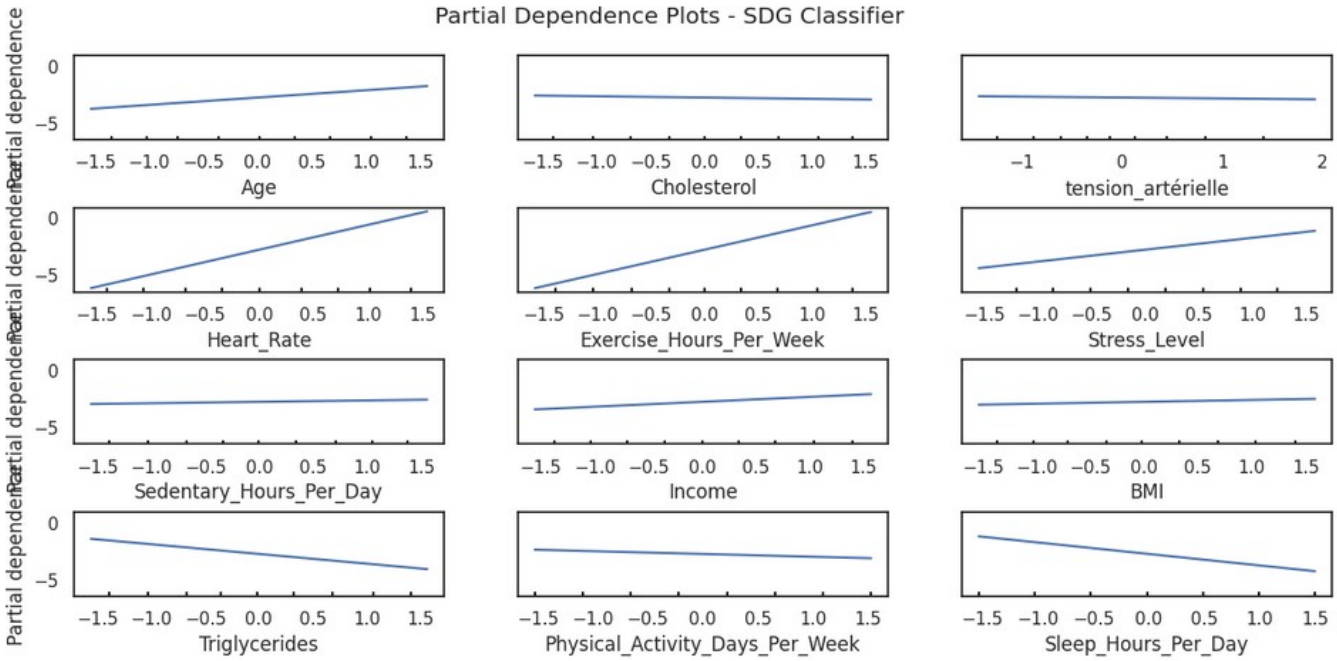
III. Interprétation du meilleur modèle

A. Interprétation globale

1) Partial Dependence Plots - PDP

La Partial Dependence Display (PDP) nous permet d'analyser l'impact individuel de chaque variable sur le risque de crise cardiaque en tenant compte de l'effet moyen des autres variables.

Figure n°9 : Les PDP du modèle SGD Classifier



Source : Dossier SVM, Valorys Trillaud et Jasmine Dupau

D'après cette analyse, nous observons que **l'âge, la fréquence cardiaque, le nombre d'heures d'activité physique par semaine, le niveau de stress et le revenu sont positivement corrélés avec le risque de crise**

cardiaque. Cela signifie que, une augmentation de ces facteurs est généralement associée à une augmentation du risque de crise cardiaque.

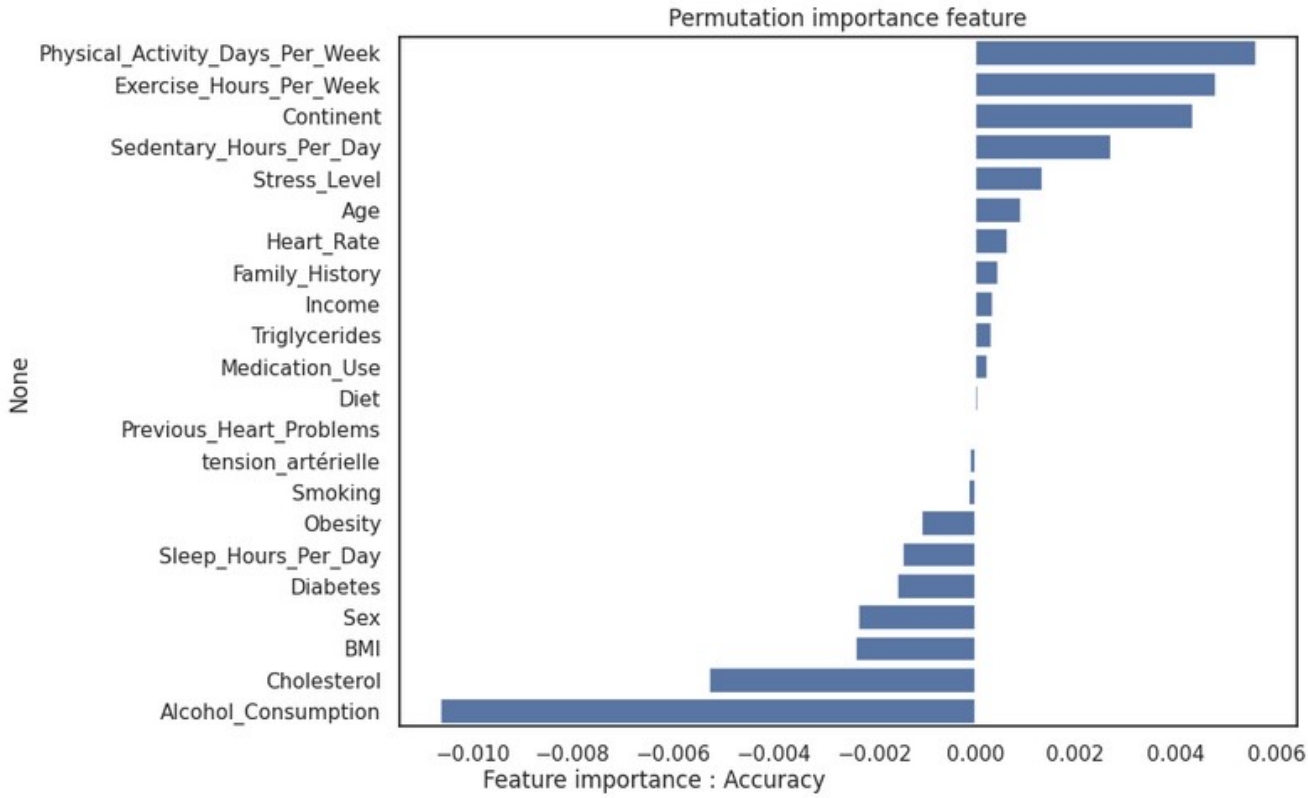
À l'inverse, certaines variables semblent avoir un effet négatif. En particulier, **le taux de triglycérides, le nombre d'heures de sommeil par jour aisni que le nombre de jours d'activité physique par semaine sont négativement corrélés avec le risque de crise cardiaque**.

Nous avons également tenté de faire cette analyse avec les variables qualitatives, mais le code n'a pas fonctionné.

2) Permutation features importance

Par ailleurs, l'analyse de **l'importance des caractéristiques par permutation (Permutation Feature Importance)** permet d'identifier les variables ayant le plus d'influence sur le modèle de prédiction de risque de crises cardiaques. Cette méthode consiste à mesurer l'impact de la perturbation de chaque variable sur les performances du modèle. En résumé, elle permet de distinguer les variables qui participent le plus à l'erreur de prédiction de celles qui n'y participent pas du tout.

Figure n°10 : Les PFI du modèle SGD Classifier



Source : Dossier SVM, Valorys Trillaud et Jasmine Dupau

Les résultats montrent que les facteurs les plus déterminants sont **le nombre de jours d'activité physique par semaine, le nombre d'heures d'exercice par semaine, le nombre d'heures de sédentarité par jour ainsi que le continent** (probablement en raison des différences d'habitudes de vie et d'accès aux soins de santé). Ces variables semblent jouer un rôle clé dans la prévision du risque cardiaque.

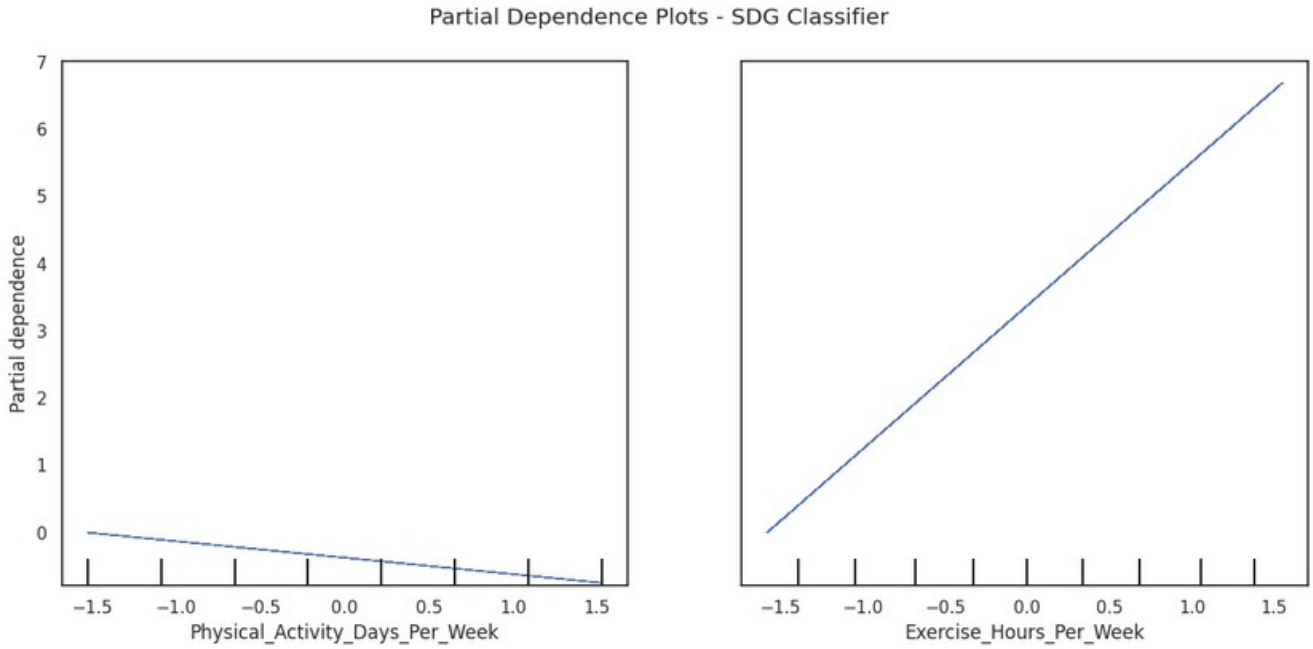
B. Interprétation locale

1) ICE

Les Individual Conditional Expectation (ICE) plots permettent de tracer sur un graphique une ligne par instance qui montre comment la prédiction d'une instance est impactée si on fait varier la valeur d'une feature.

Cette méthode est basée sur les PDP, la PDP représente l'effet moyénniser de toutes les courbes ICE. La méthode est donc exactement la même, on fait varier notre feature d'intérêt tout en fixant à leur valeur moyenne les autres features.

Figure n°11: Les ICE du modèle SGD Classifier



Source : Dossier SVM, Valorys Trillaud et Jasmine Dupau

Nous retrouvons ici les mêmes constats qu'à l'échelle agrégée. En effet, plus le nombre de jours d'activité physique par semaine augmente, plus la probabilité prédite d'avoir une crise cardiaque diminue pour les individus étudiés. Autrement dit, le modèle estime qu'une pratique sportive régulière au fil de la semaine est associée à risque plus faible.

En revanche, lorsque la durée hebdomadaire des séances augmente, la probabilité prédite de crise cardiaque tend à croître. Cela peut sembler contre-intuitif, mais plusieurs hypothèses peuvent l'expliquer : un effet de surentraînement ou un profil particulier d'individus déjà à risque.

2) LIME

Pour clôturer l'interprétabilité du modèle, nous allons faire appel au LIME. **L'algorithme LIME(Local interpretable model-explanation)** est une méthode d'explicabilité conçue pour interpréter des modèles black box comme le nôtre.

L'idée est d'expliquer localement la prédiction d'un modèle en utilisant un modèle plus simple et compréhensible dans un voisinage spécifique de l'exemple à expliquer (Reg linéaire, arbre, Lasso, ...)

Figure n°12: Le LIME du modèle SGD Classifier

Source : Dossier SVM, Valorys Trillaud et Jasmine Dupau

Si nous utilisons l'algorithme Lime pour comprendre comment notre prédiction se réalise pour le 118e individu, nous remarquons que :

- A gauche, le modèle a prédit que l'individu avait 88% de chance d'appartenir à la classe "non risque" contre 12% pour l'autre classe.
- Au centre, nous pouvons lire l'effet des variables discrétisées sur la prédiction. Cela permet de comprendre pourquoi le modèle a prédit ces valeurs pour cet individu.Nous observons que le modèle a sélectionné 2 variables qui contribuent à la baisse de la prédiction (Continent et Obesity) et 3 variables qui tirent la prédiction vers le haut (Diet, Stress_Level et Physical_Activity_Day). Toutefois, l'effet de chacune des variables est faible même si nous pouvons lire à droite que la variable "Continent" a la la valeur la plus élevée positibe. Ainsi, cette prédiction est principalement expliquée par l'origine géographique de l'individu.

Conclusion