Projet Machine Learning

Yen Phi Do | Hugo Alpiste | Sébastien Martel | Morgane Geoffroy

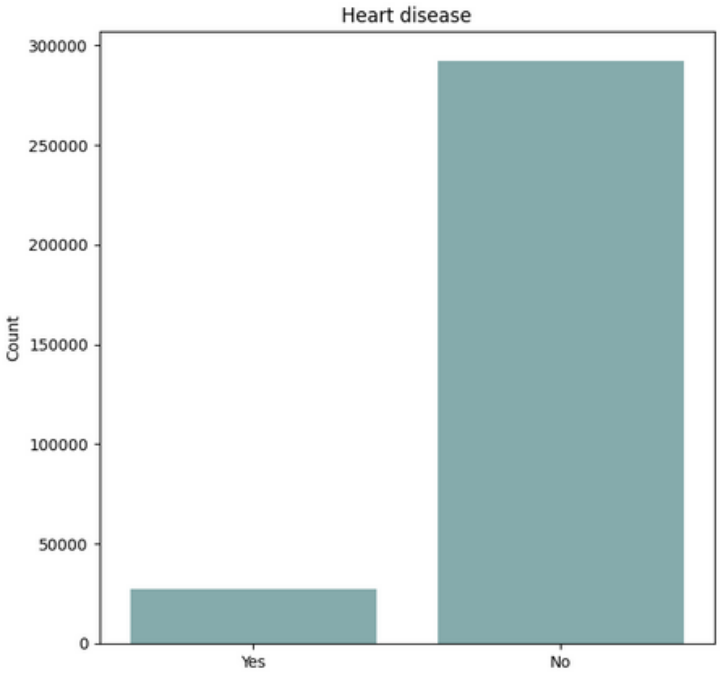
30/08/2023

diginamic FORMATION

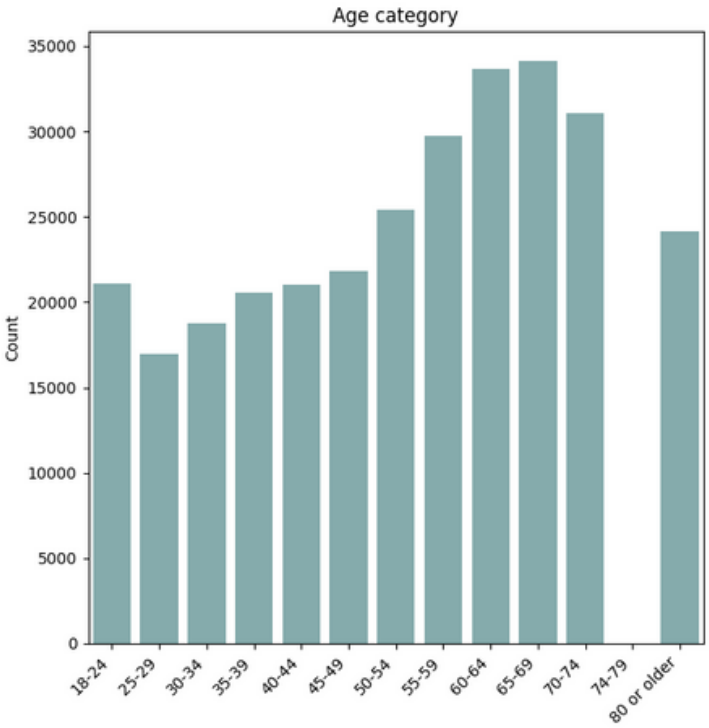
« Votre passeport pour l'emploi numérique »

SOMMAIRE

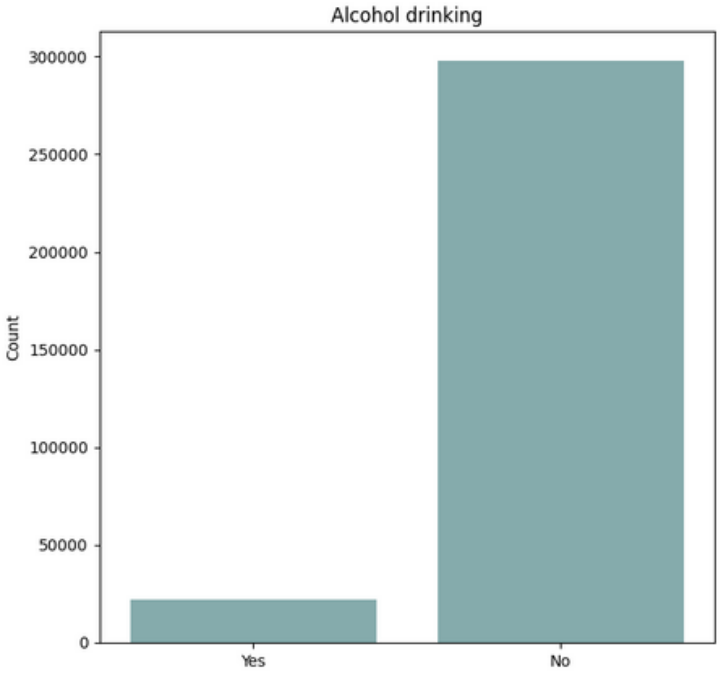
Présentation des données



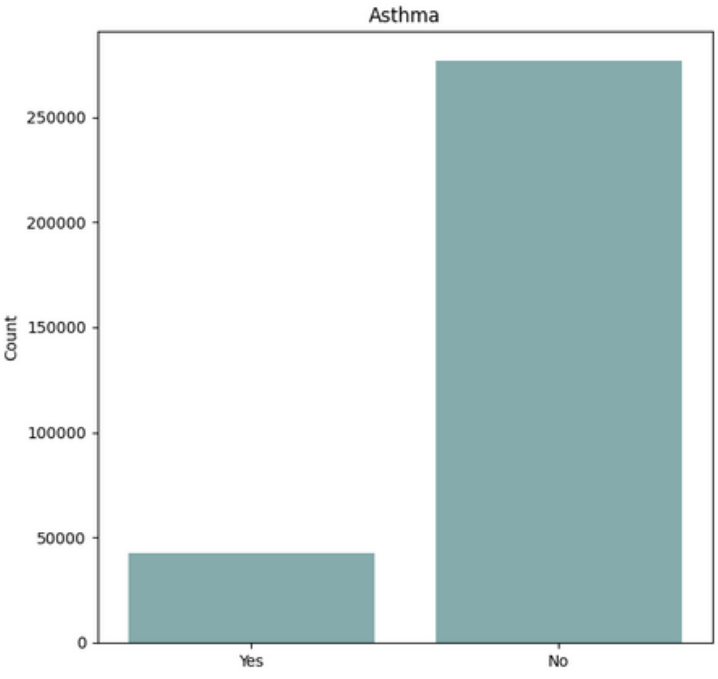
HeartDisease
Indicates whether respondents have ever reported having coronary heart disease (CHD) or myocardial infarction (MI)



AgeCategory
Represents respondents' age categorized into fourteen levels

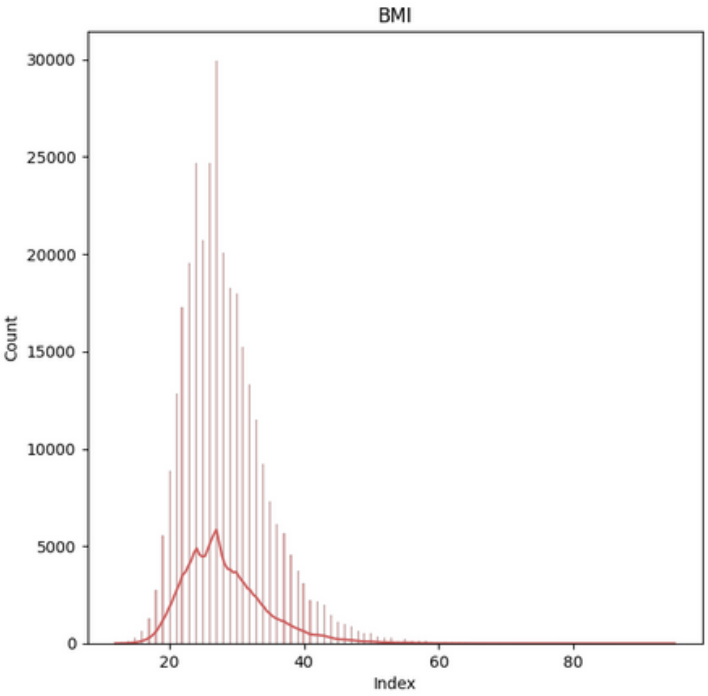


AlcoholDrinking
Indicates whether respondents are considered heavy drinkers. (Note: For adult men, heavy drinking refers to consuming more than 14 drinks per week, while for adult women, it means consuming more than 7 drinks per week)

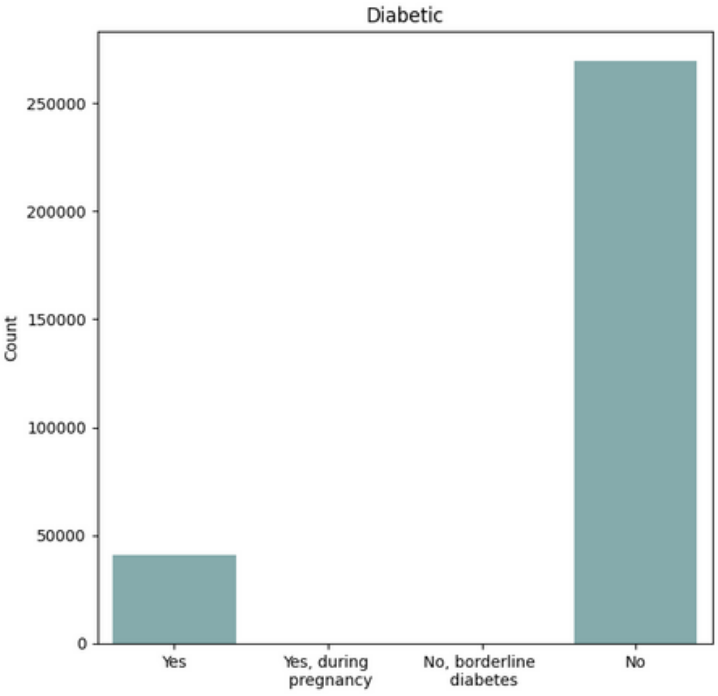


Asthma
Determines if respondents have ever been told they had asthma

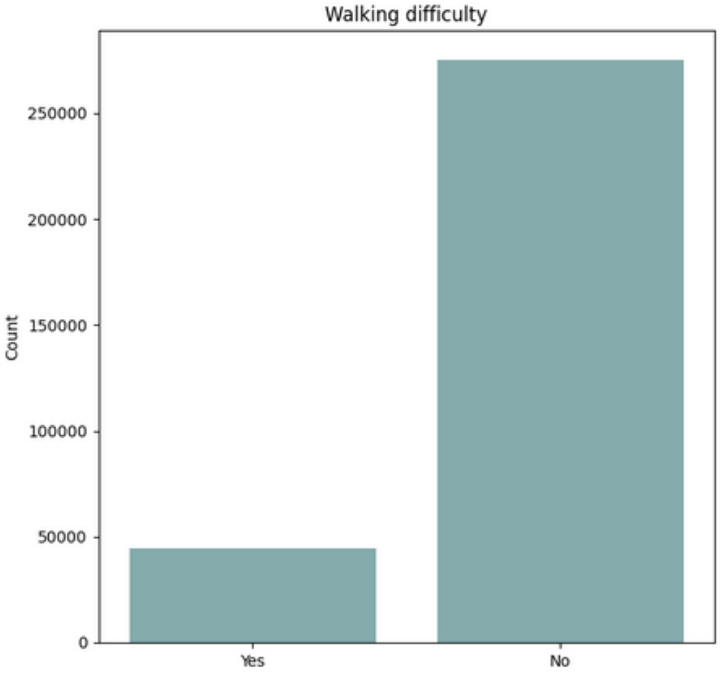
Présentation des données



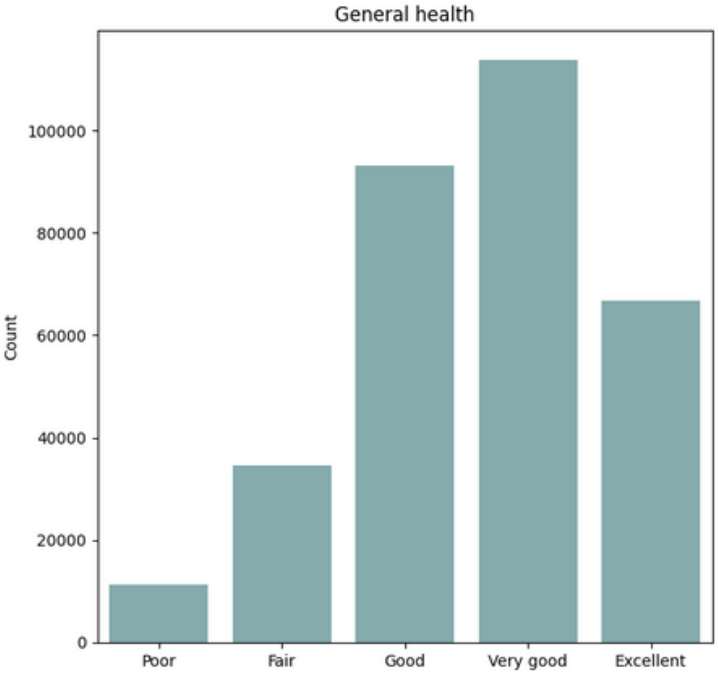
BMI
Body Mass Index (BMI)



Diabetes
Determines if respondents have ever had diabetes

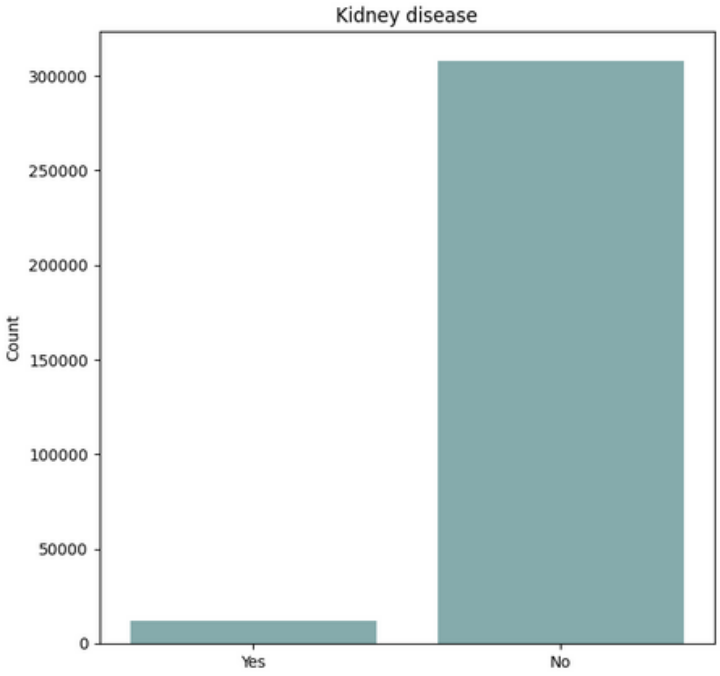


WalkingDifficulty
Determines if respondents have serious difficulty walking or climbing stairs

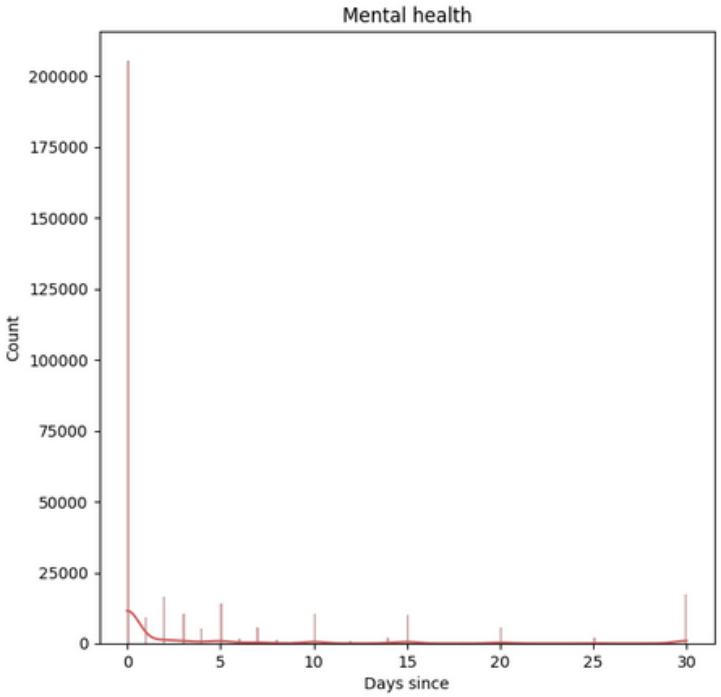


GeneralHealth
Reflects respondents' perception of their overall general health

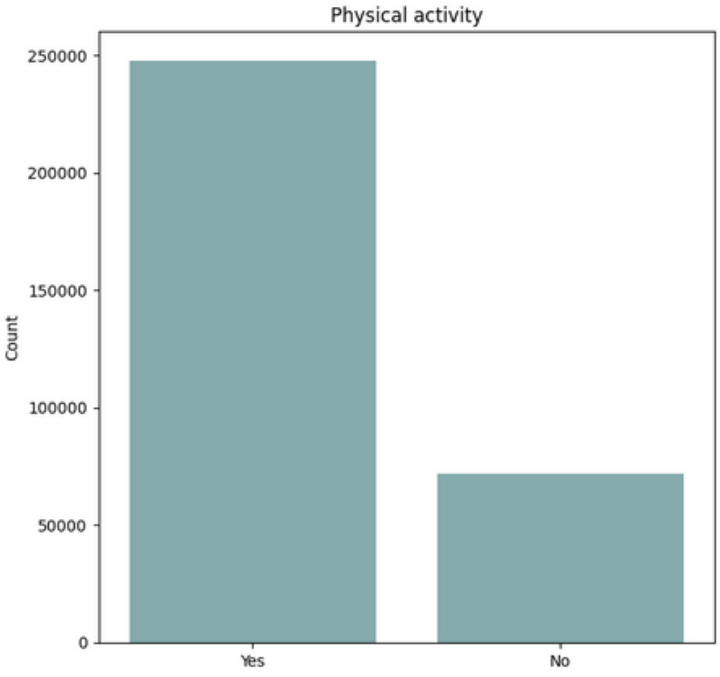
Présentation des données



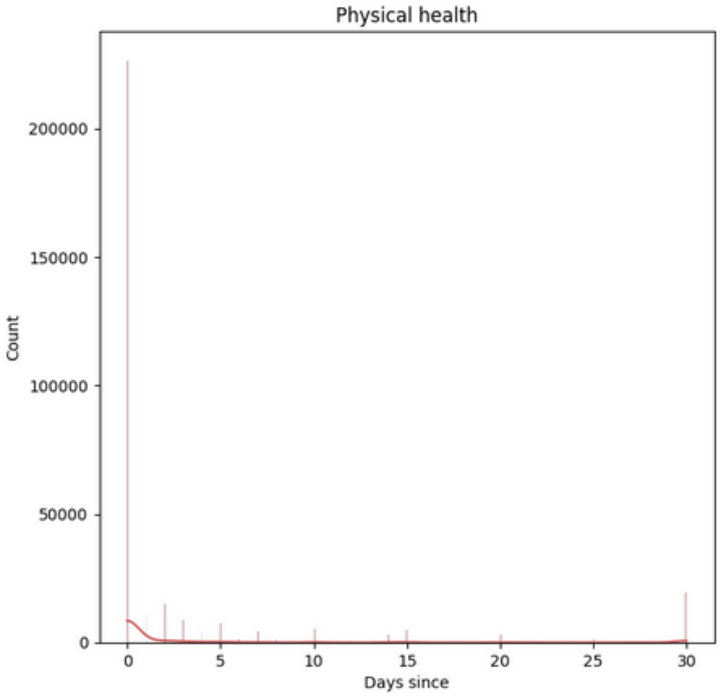
KidneyDisease
Indicates if respondents have ever been diagnosed with kidney disease, excluding kidney stones, bladder infections, or incontinence



MentalHealth
Indicates the number of days during the past 30 days when respondents experienced poor mental health

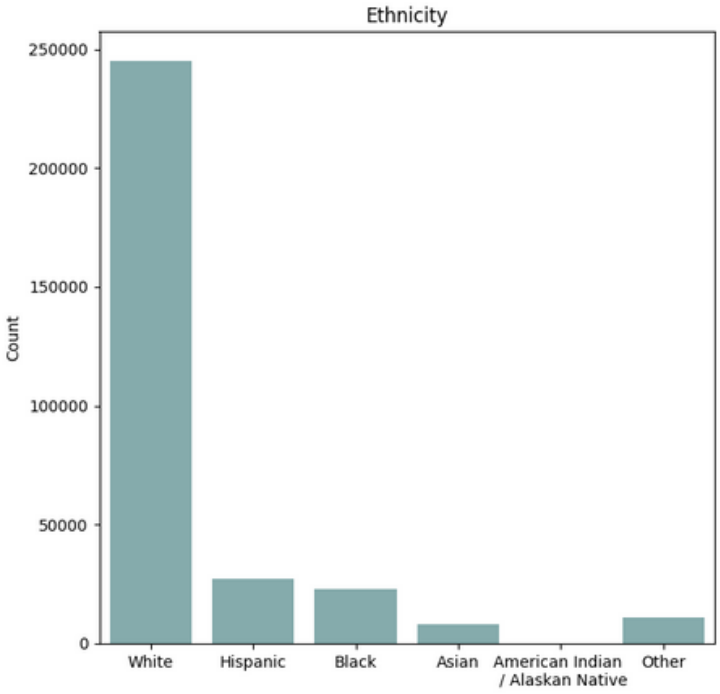


PhysicalActivity
Indicates whether respondents have engaged in physical activity or exercise during the past 30 days, excluding regular job-related activities

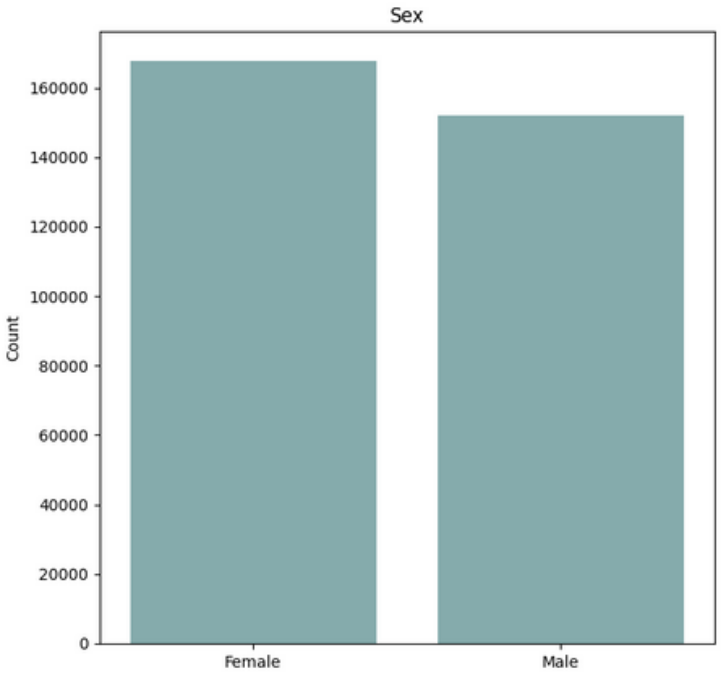


PhysicalHealth
Reflects the number of days during the past 30 days when respondents experienced poor physical health, including illness and injury

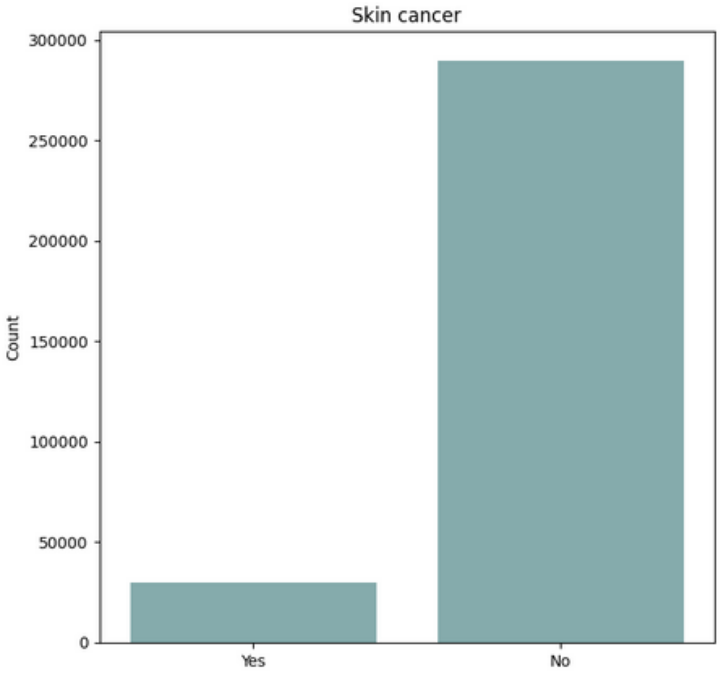
Présentation des données



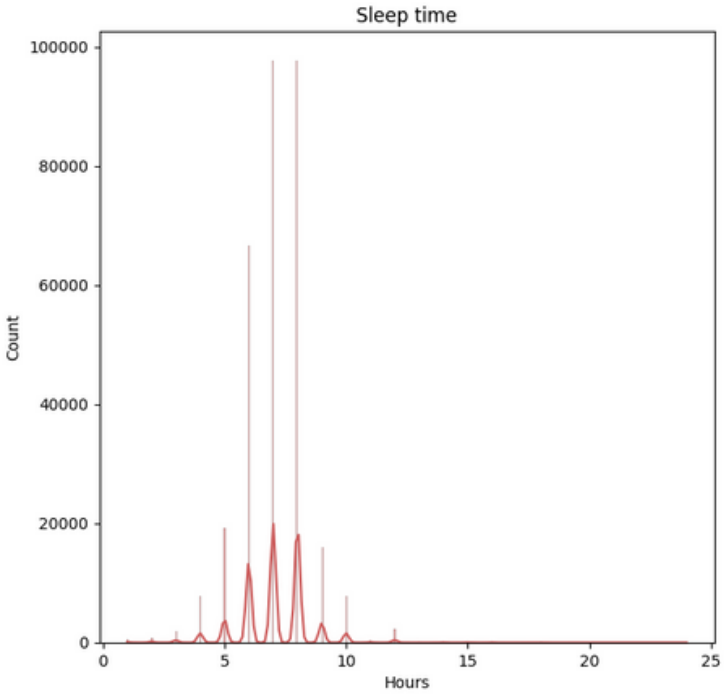
Ethnicity
Indicates the imputed race/ethnicity value



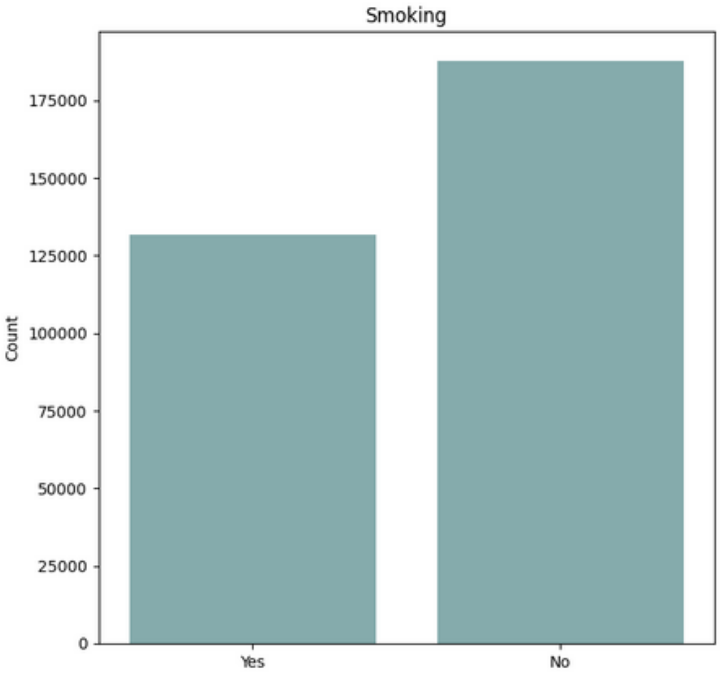
Sex
Identifies respondents' gender as male or female

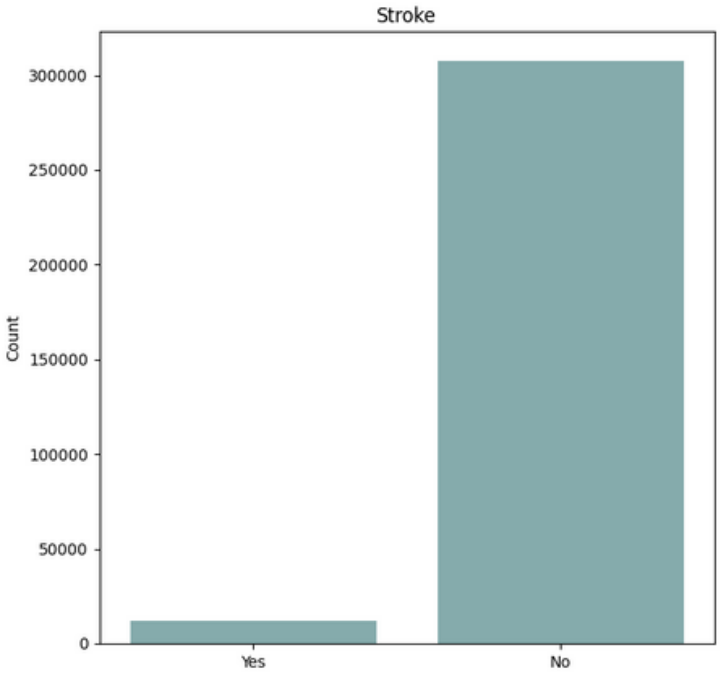


SkinCancer
Determines if respondents have ever been told they had skin cancer



SleepTime
Represents the average number of hours of sleep obtained within a 24-hour period





Smoking
Determines if respondents have smoked at least 100 cigarettes in their lifetime. (Note: 5 packs = 100 cigarettes)

Stroke
Indicates whether respondents have ever suffered a stroke

Analyses univariées

On décide de réaliser des tests pour chacune des variables afin de déterminer s'il existe une différence significative au sein des groupes de patients sains et souffrant de maladie cardiaque.

Pour les variables catégorielles, on décide de réaliser un test du Chi2 adapté aux variables qualitatives. Afin de pouvoir réaliser un test du Chi2 il nous faut au moins 5 individus dans chaque groupe.

Pour les variables numériques, on décide de réaliser un test de Student. Afin de pouvoir le réaliser, plusieurs conditions doivent être respectées :

- Indépendance des deux groupes
- Normalité
- Homoscédasticité (égalité des variances)

Tests préliminaires

On lance une première phase de tests afin de se familiariser avec les données.

		Grouped by HeartDisease				
		Missing	Overall	No	Yes	P-Value
n			319795	292422	27373	
BMI, mean (SD)		0	28.3 (6.4)	28.2 (6.3)	29.4 (6.6)	<0.001
Smoking, n (%)	Yes		131908 (41.2)	115871 (39.6)	16037 (58.6)	<0.001
AlcoholDrinking, n (%)	Yes		21777 (6.8)	20636 (7.1)	1141 (4.2)	<0.001
Stroke, n (%)	Yes		12069 (3.8)	7680 (2.6)	4389 (16.0)	<0.001
PhysicalHealth, mean (SD)		0	3.4 (8.0)	3.0 (7.4)	7.8 (11.5)	<0.001
MentalHealth, mean (SD)		0	3.9 (8.0)	3.8 (7.8)	4.6 (9.2)	<0.001
DiffWalking, n (%)	Yes		44410 (13.9)	34382 (11.8)	10028 (36.6)	<0.001
Sex, n (%)	Female	0	167805 (52.5)	156571 (53.5)	11234 (41.0)	<0.001
	Male		151990 (47.5)	135851 (46.5)	16139 (59.0)	
AgeCategory, n (%)	18-24	0	21064 (6.6)	20934 (7.2)	130 (0.5)	<0.001
	25-29		16955 (5.3)	16822 (5.8)	133 (0.5)	
	30-34		18753 (5.9)	18527 (6.3)	226 (0.8)	
	35-39		20550 (6.4)	20254 (6.9)	296 (1.1)	
	40-44		21006 (6.6)	20520 (7.0)	486 (1.8)	
	45-49		21791 (6.8)	21047 (7.2)	744 (2.7)	
	50-54		25382 (7.9)	23999 (8.2)	1383 (5.1)	
	55-59		29757 (9.3)	27555 (9.4)	2202 (8.0)	
	60-64		33686 (10.5)	30359 (10.4)	3327 (12.2)	
	65-69		34151 (10.7)	30050 (10.3)	4101 (15.0)	
	70-74		31065 (9.7)	26218 (9.0)	4847 (17.7)	
	75-79		21482 (6.7)	17433 (6.0)	4049 (14.8)	
	80 or older		24153 (7.6)	18704 (6.4)	5449 (19.9)	
Race, n (%)	American Indian/Alaskan Native	0	5202 (1.6)	4660 (1.6)	542 (2.0)	<0.001
	Asian		8068 (2.5)	7802 (2.7)	266 (1.0)	
	Black		22939 (7.2)	21210 (7.3)	1729 (6.3)	
	Hispanic		27446 (8.6)	26003 (8.9)	1443 (5.3)	
	Other		10928 (3.4)	10042 (3.4)	886 (3.2)	
	White		245212 (76.7)	222705 (76.2)	22507 (82.2)	
Diabetic, n (%)	No	0	269653 (84.3)	252134 (86.2)	17519 (64.0)	<0.001
	No, borderline diabetes		6781 (2.1)	5992 (2.0)	789 (2.9)	
	Yes		40802 (12.8)	31845 (10.9)	8957 (32.7)	
	Yes (during pregnancy)		2559 (0.8)	2451 (0.8)	108 (0.4)	
PhysicalActivity, n (%)	Yes		247957 (77.5)	230468 (78.8)	17489 (63.9)	<0.001
GenHealth, n (%)	Excellent	0	66842 (20.9)	65342 (22.3)	1500 (5.5)	<0.001
	Fair		34677 (10.8)	27593 (9.4)	7084 (25.9)	
	Good		93129 (29.1)	83571 (28.6)	9558 (34.9)	
	Poor		11289 (3.5)	7439 (2.5)	3850 (14.1)	
	Very good		113858 (35.6)	108477 (37.1)	5381 (19.7)	
SleepTime, mean (SD)		0	7.1 (1.4)	7.1 (1.4)	7.1 (1.8)	<0.001
Asthma, n (%)	Yes		42872 (13.4)	37939 (13.0)	4933 (18.0)	<0.001
KidneyDisease, n (%)	Yes		11779 (3.7)	8324 (2.8)	3455 (12.6)	<0.001
SkinCancer, n (%)	Yes		29819 (9.3)	24839 (8.5)	4980 (18.2)	<0.001

Observations et décisions

Variables catégorielles

- Chacun des échantillons est constitué d'au moins 5 individus, la condition d'application est donc respectée.

- On regroupe les classes d'âge deux à deux afin de d'étoffer les effectifs des plus petits groupes qui comprennent un peu plus de 100 individus sur les 319 000 individus du jeu de données global.

- On recode les variables dichotomiques en binaire

- On "one-hot encode" les variables à strictement plus de deux catégories.

Vérification des conditions d'application du test de Student

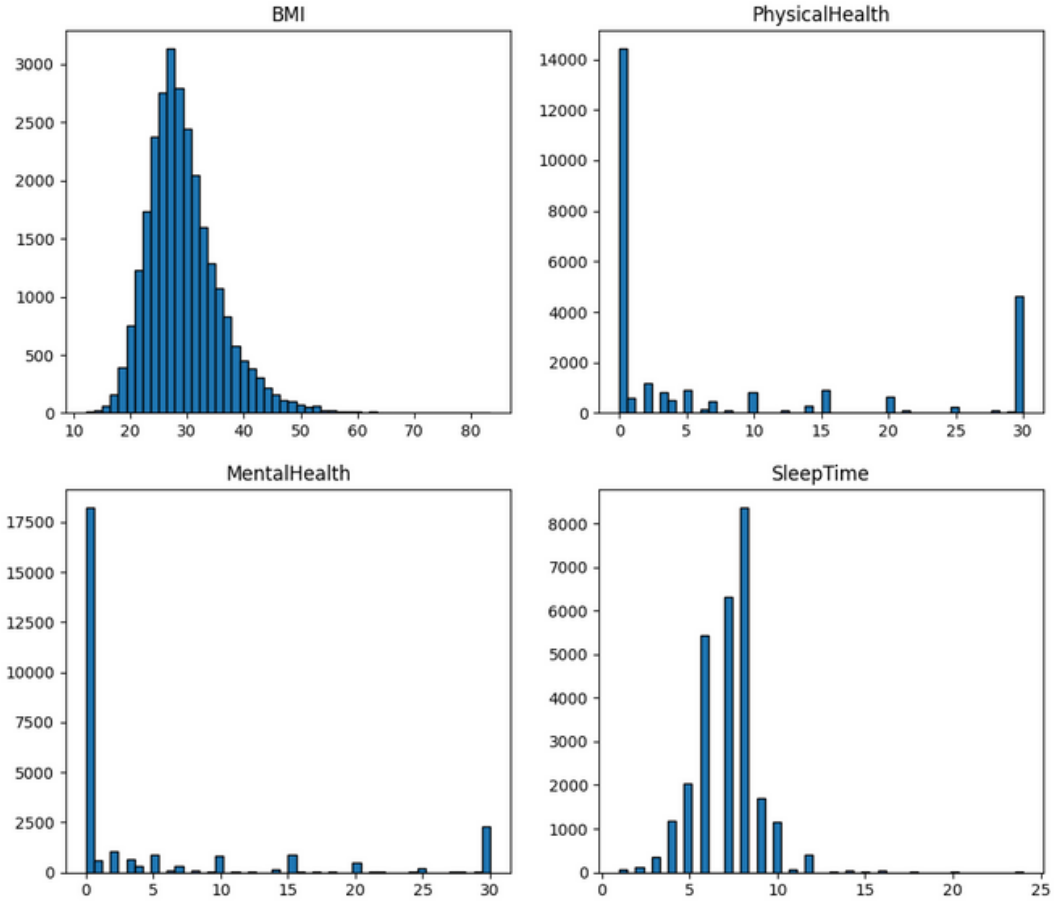
- **Indépendance des échantillons** : les valeurs observées dans les différents groupes sont indépendantes les unes des autres.

- **Normalité** : on décide de la vérifier visuellement car le Théorème Central Limite nous permet de la supposer lorsque l'effectif de chacun des échantillons est suffisamment grand. Elle est vérifiée (p.8) pour les variables BMI et SleepTime. Elle n'est pas vérifiée (p.8) pour les variables PhysicalHealth et MentalHealth. On décide de passer ces variables en variables catégorielles.

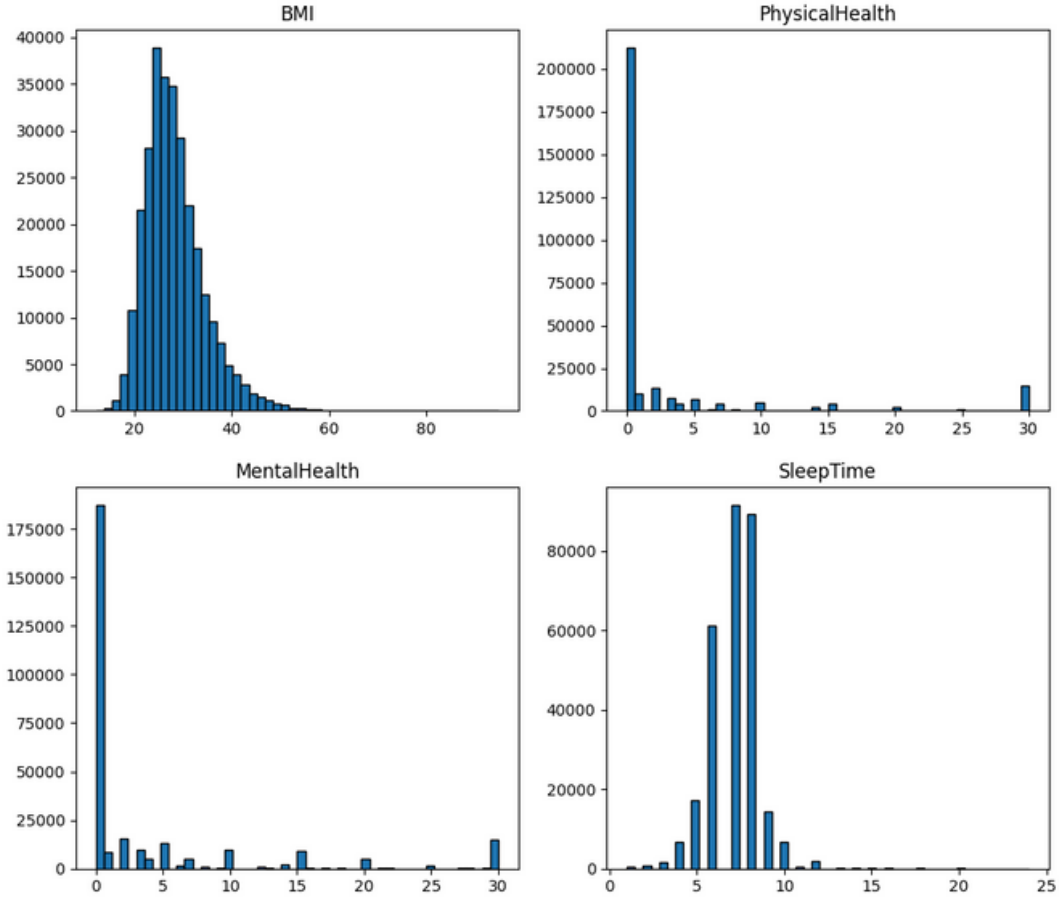
- **Homoscédasticité** : on décide de la vérifier pour les deux variables numériques restantes à l'aide d'un test de Bartlett. Celui-ci, et dans les deux cas, rejette l'hypothèse d'une égalité entre les variances des deux groupes étudiés. Il s'agit d'un résultat attendu du fait de la constitution des groupes. On décide donc de s'en affranchir.

Vérification de la normalité

Groupe sain



Groupe malade



Base de données

Les données ont d'abord été déployées dans une base MongoDB à l'aide d'un script de setup.
Un singleton a été utilisé pour partager la connexion à celle-ci entre les différentes parties du projet.

Analyse exploratoire

Notre première priorité a ensuite été d'examiner nos données afin d'établir une marche à suivre.
Pour ce faire, nous avons réalisé l'analyse préliminaire à l'aide des scripts :

- /scripts/explore/exploration_graphs.py : génération des graphiques de répartition des individus pour chaque attribut (voir plus haut "Présentation des données"). Exportation au format png dans un sous-dossier "img/".
- /scripts/explore/devScript.py : analyse exploratoire et statistique (cf chapitre précédent "Analyse exploratoire").

Conclusions

La compréhension obtenue du jeu de données à mené :

- aux décisions concernant les prétraitements à appliquer pour le machine learning (voir plus haut "Observations et décisions").

Ceux-ci sont appliqués à l'aide du script scripts/explore/dataPreprocessing.py.

- au choix des modèles de ML à tester pour notre prédiction.

Sélection d'un modèle de Machine Learning

Test de trois modèles de classification de la bibliothèque scikit-learn : KNeighborsClassifier, LogisticRegression et RandomForestClassifier

Pour chacun de ces modèles, nous avons suivi le process suivant :

- train_test_split : séparation des données en un jeu d'entraînement et un jeu de test. Utilisation des paramètres "random_state" pour la reproductibilité, et "stratify" pour le respect de la proportion des catégories de la cible (seulement 8% des individus présentant une pathologie cardiaque).
- Création d'un dictionnaire de paramètres à tester, propres au modèle.
- Création et entraînement d'un objet GridSearchCV fournissant le meilleur modèle possible à partir de ces paramètres.
- Récupération du score (voir récapitulatif plus bas).
- Prédiction sur le jeu de test.
- Mesure du score de la prédiction à l'aide de la fonction accuracy_score.
- Comparaison du score d'entraînement avec celui de prédiction afin d'évaluer l'overfitting; aucun de nos modèles n'a eu de problème à ce niveau.
- Sauvegarde du modèle entraîné au format joblib.

Résultats du test

Scores obtenus lors des tests :
KNeighborsClassifier(n_neighbors=45) : score 0.916
LogisticRegression(max_iter = 10000, C=5.963623316594643, penalty='l2') : score 0.916
RandomForestClassifier(n_estimators=700, max_features=1) : score 0.905

Nous avons au final opté pour le modèle KNN, fréquemment rencontré lors d'études de même type, et doté dans notre cas d'un **score de 91.6%.**

Application finale et interface utilisateur

Afin d'obtenir une interface ergonomique, portable et évolutive, nous avons opté pour un affichage web.
Notre choix s'est porté sur la bibliothèque *streamlit*, pour sa facilité de mise en place, ainsi que sa faculté à nativement afficher des graphiques à partir de *dataframes* pandas, ce qui pourrait s'avérer intéressant pour des évolutions ultérieures de l'application.

La commande *python -m streamlit run MEDICAL_ASSISTANT.py* ouvre automatiquement l'application dans une fenêtre de navigateur.

L'utilisateur sélectionne alors les paramètres à évaluer et clique sur le bouton "Submit"; les données sont envoyées au script de prétraitement puis transmises au modèle entraîné, et la prédiction obtenue est renvoyée vers la page web.