diginamic FORMATION

« Votre passeport pour l'emploi numérique »

Projet - Machine Learning

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

30/08/2023

Plan

Projet

Présentation des objectifs, des étapes de réalisation du projet et des données de départ.

Plan

Projet

Présentation des objectifs, des étapes de réalisation du projet et des données de départ.

Analyse exploratoire

Présentation des différentes étapes de prétraitement du jeu de données initial résultantes de l'analyse exploratoire.

Plan

Projet	Présentation des objectifs, des étapes de réalisation du projet et des données de départ.
Analyse exploratoire	Présentation des différentes étapes de prétraitement du jeu de données initial résultantes de l'analyse exploratoire.
Machine Learning	Présentation des différents modèles de Machine Learning évalués ainsi que des performances du modèle retenu.

Plan

Projet	Présentation des objectifs, des étapes de réalisation du projet et des données de départ.
Analyse exploratoire	Présentation des différentes étapes de prétraitement du jeu de données initial résultantes de l'analyse exploratoire.
Machine Learning	Présentation des différents modèles de Machine Learning évalués ainsi que des performances du modèle retenu.
Application	Présentation des différentes étapes de conception de l'application.

Plan

Projet	Présentation des objectifs, des étapes de réalisation du projet et des données de départ.
Analyse exploratoire	Présentation des différentes étapes de prétraitement du jeu de données initial résultantes de l'analyse exploratoire.
Machine Learning	Présentation des différents modèles de Machine Learning évalués ainsi que des performances du modèle retenu.
Application	Présentation des différentes étapes de conception de l'application.
Démonstration	Réalisation d'une démonstration de l'utilisation de l'application finale.

Projet

Prédire la survenue probable d’un incident cardiaque sur la base de réponses renseignées par le patient dans un questionnaire

Etapes de réalisation du projet

1 - Analyse exploratoire

Exploration des données afin de pouvoir en évaluer la pertinence et résumer leurs principales caractéristiques.

Prédire la survenue probable d'un incident cardiaque sur la base de réponses renseignées par le patient dans un questionnaire

Etapes de réalisation du projet

2 - Pré-traitement

Formatage du jeu de données initial, résultant de l'analyse exploratoire, dans le but de l'adapter à une utilisation dans des algorithmes de Machine Learning

Projet

Prédire la survenue probable d'un incident cardiaque sur la base de réponses renseignées par le patient dans un questionnaire

Etapes de réalisation du projet

3 - Entraînement, optimisation et évaluation des performances de plusieurs modèles

Comparaison de trois modèles adaptés à la classification dont les hyperparamètres auront été optimisés et dont les performances auront été évaluées.

Projet

Prédire la survenue probable d’un incident cardiaque sur la base de réponses renseignées par le patient dans un questionnaire

Etapes de réalisation du projet

4 - Choix d’un modèle

Choix du modèle qui permettra de prédire la possibilité d’une atteinte cardiaque sur la base d’un certain nombre de paramètres renseignés en interface.

Projet

Prédire la survenue probable d'un incident cardiaque sur la base de réponses renseignées par le patient dans un questionnaire

Etapes de réalisation du projet

5 - Application finale

Mise en place d'une interface web fonctionnelle et facile d'utilisation afin de fournir une prédiction à partir des données d'entrée.

Présentation des données

Variables numériques	BMI	PhysicalHealth	MentalHealth	SleepTime
count	319 795	319 795	319 795	319 795
mean	28.33	3.37	3.90	7.10
std	6.36	7.95	7.96	1.44
min	12.02	0.00	0.00	1.00
25%	24.03	0.00	0.00	6.00
50%	27.34	0.00	0.00	7.00
75%	31.42	2.00	3.00	8.00
max	94.85	30.00	30.00	24.00

Projet

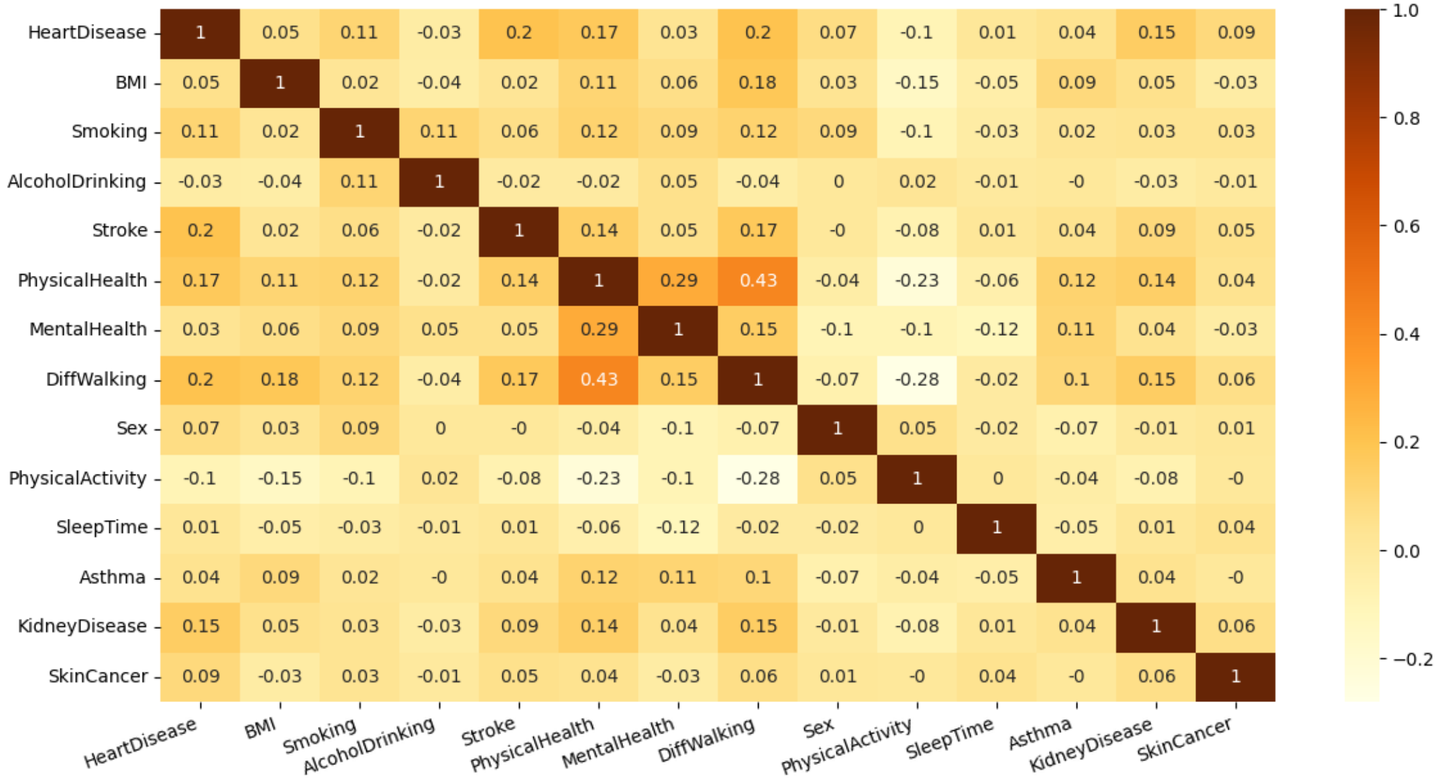
Variables catégorielles	HeartDisease	Smoking	AlcoholDrinking	Stroke	DiffWalking	Sex	AgeCategory
count	319 795	319 795	319 795	319 795	319 795	319 795	319 795
unique	2	2	2	2	2	2	13
top	No	No	No	No	No	Female	65-69
freq	29 422	187 887	298 018	307 726	275 385	167 805	34 151

Variables catégorielles	Race	Diabetic	PhysicalActivity	GenHealth	Asthma	KidneyDisease	SkinCancer
count	319 795	319 795	319 795	319 795	319 795	319 795	319 795
unique	6	4	2	5	2	2	2
top	White	No	Yes	Very good	No	No	No
freq	245 212	269 653	247 957	113 858	276 923	308 016	289 976

Analyse exploratoire

Matrice de corrélations

Des corrélations relativement faibles nous indiquent l'importance de chacune des variables pour la mesure de différentes caractéristiques



Corrélation positive entre **DiffWalking** et **PhysicalHealth**
Lorsque le patient a des difficultés à marcher il va avoir tendance à se considérer en moins bonne santé physique.
De même, lorsque celui-ci considère que sa santé mentale n'est pas bonne.

Analyse exploratoire

Existe-t-il, pour chaque variable, une différence significative entre le groupe de patients sains et celui de patients atteints d'une maladie cardiaque ?

Analyses univariées

Variables catégorielles	Test du Chi2	Au moins 5 individus dans chaque groupe
Variables numériques	Test de Student	Indépendance des deux groupes Normalité Homoscédasticité (égalité des variances)

Analyse exploratoire

		Yes	No	p-value
BMI (mean)		29.40	28.20	< 0.001
PhysicalHealth (mean)		7.80	3.00	< 0.001
MentalHealth (mean)		4.60	3.80	< 0.001
SleepTime (mean)		7.10	7.10	< 0.001
Diabetic (%)	No	64.00	86.20	< 0.001
	No, borderline diabetes	2.90	2.00	
	Yes	32.70	10.90	
	Yes (during pregnancy)	0.40	0.80	
KidneyDisease (%)	Yes	12.60	2.80	< 0.001
SkinCancer (%)	Yes	18.20	8.50	< 0.001

Analyse exploratoire

		Yes	No	p-value
Smoking (%)	Yes	58.60	39.60	< 0.001
AlcoholDrinking (%)	Yes	4.20	7.10	< 0.001
Stroke (%)	Yes	16.00	2.60	< 0.001
DiffWalking (%)	Yes	36.60	11.80	< 0.001
Sex (%)	Male	59.00	46.50	< 0.001
GenHealth (%)	Excellent	5.50	22.30	< 0.001
	Fair	25.90	9.40	
	Good	34.90	28.60	
	Poor	14.10	2.50	
	Very Good	19.70	37.10	

Analyse exploratoire

		Yes	No	p-value
AgeCategory (%)	18-24	0.50	7.20	< 0.001
	25-29	0.50	5.80	
	30-34	0.80	6.30	
	35-39	1.10	6.90	
	40-44	1.80	7.00	
	45-49	2.70	7.20	
	50-54	5.10	8.20	
	55-59	8.00	9.40	
	60-64	12.20	10.40	
	65-69	15.00	10.30	
	70-74	17.70	9.00	
	75-79	14.80	6.00	

Analyse exploratoire

		Yes	No	p-value
AgeCategory (%)	80 or older	19.90	6.40	< 0.001
Race (%)	American Indian/Alaskan Native	2.00	1.60	< 0.001
	Asian	1.00	2.70	
	Black	6.30	7.30	
	Hispanic	5.30	8.90	
	Other	3.20	3.40	
	White	82.20	76.20	
PhysicalActivity (%)	Yes	63.90	78.80	< 0.001
Asthma (%)	Yes	18.00	13.00	< 0.001

Analyse exploratoire

Variables catégorielles

Observations

- Condition d'application respectée : n > 5
- Très (trop ?) petits effectifs dans certaines catégories de AgeCategory
- Formatage des variables non optimisé

Décisions

- Regroupement des classes d'âge deux à deux afin d'en étoffer les effectifs.
- Recodage des variables dichotomiques en binaire
- One-hot encodeage des variables à plus de deux catégories

Analyse exploratoire

Variables numériques

Indépendance des échantillons

Les valeurs observées dans les différents groupes sont indépendantes les unes des autres.

Normalité

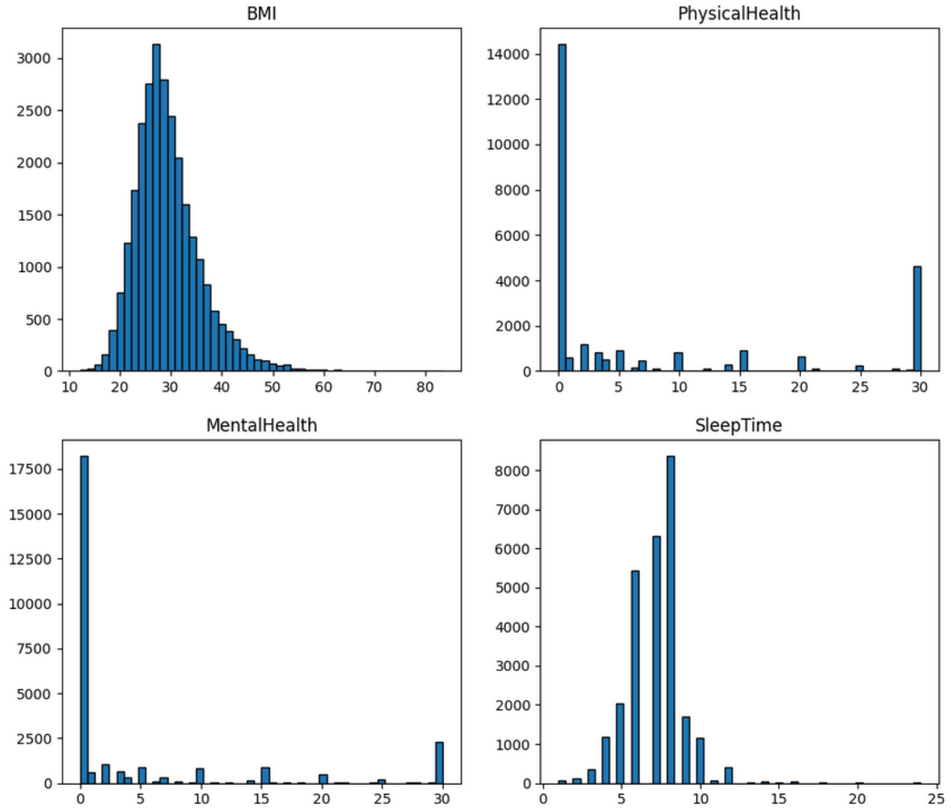
Vérification visuelle car supposée (TCL) pout de grands effectifs.

Homoscédasticité

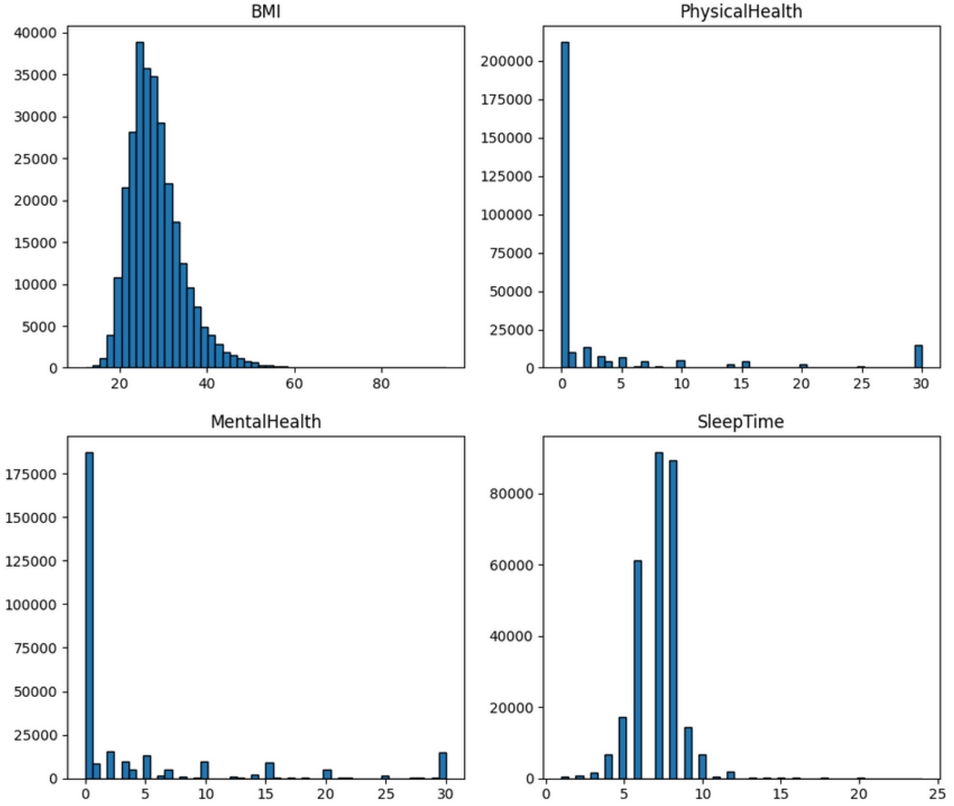
Vérification de l'égalité des variances au sein des deux groupes (sains et malades) pour les variables BMI et SleepTime à l'aide du test de Bartlett. Sans surprise, elle n'est pas vérifiée, on décide donc de s'en affranchir.

Analyse exploratoire

Normalité



Groupe sain

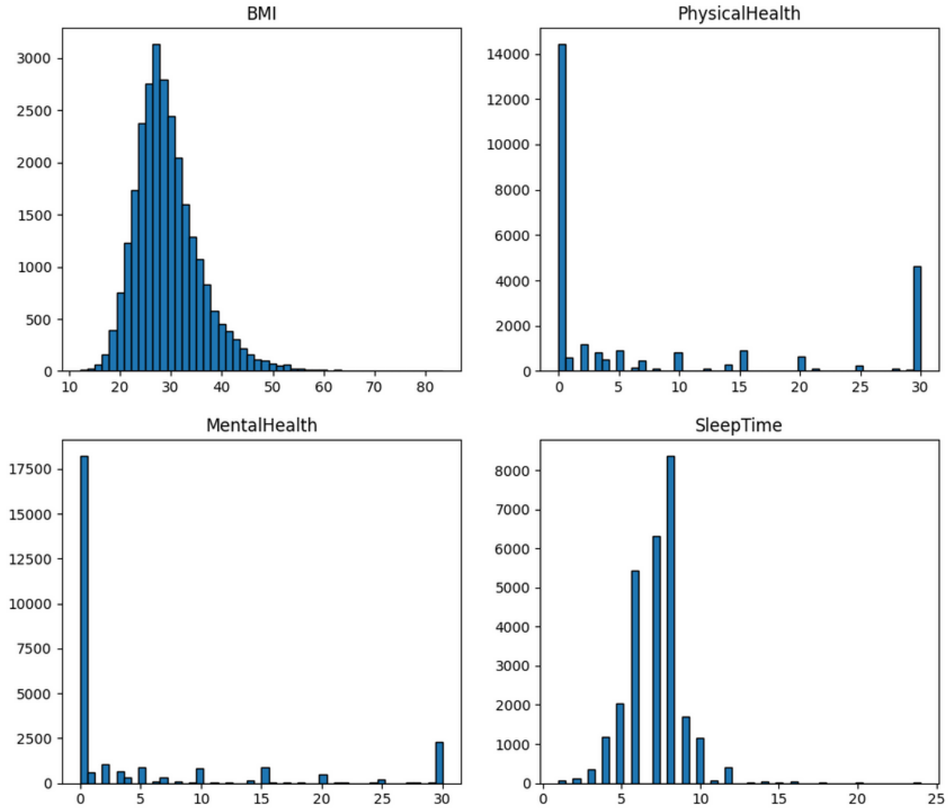


Groupe malade

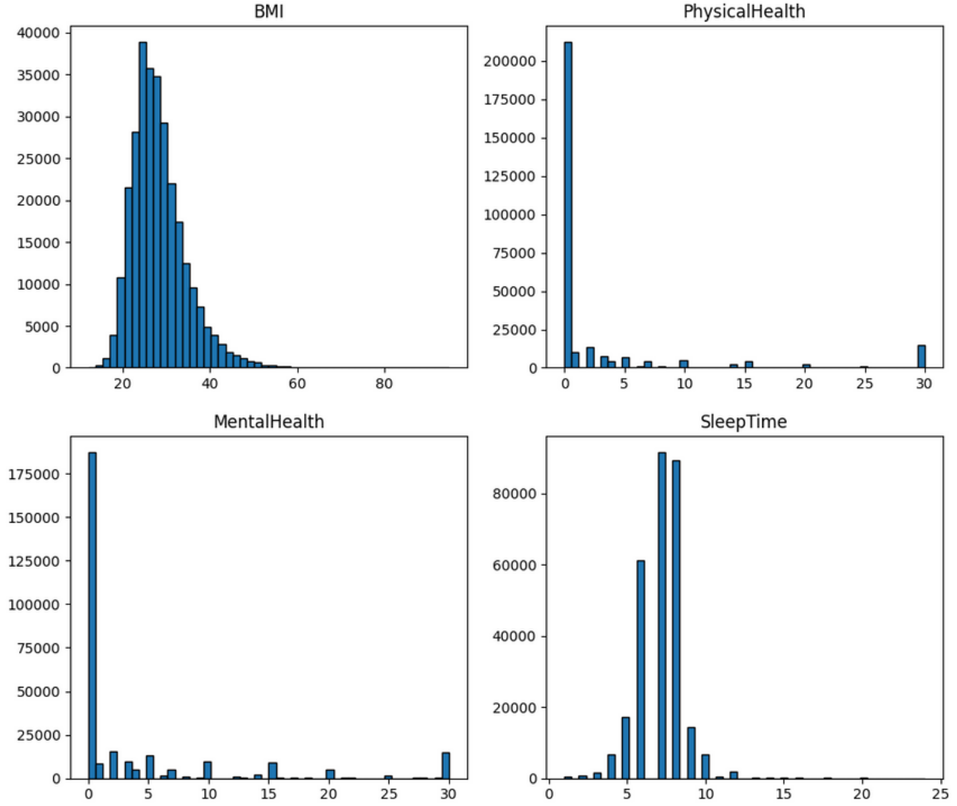
Analyse exploratoire

Normalité

- Vérifiée pour les variables BMI et SleepTime.
- Non vérifiée pour les variables PhysicalHealth et MentalHealth.



Groupe sain

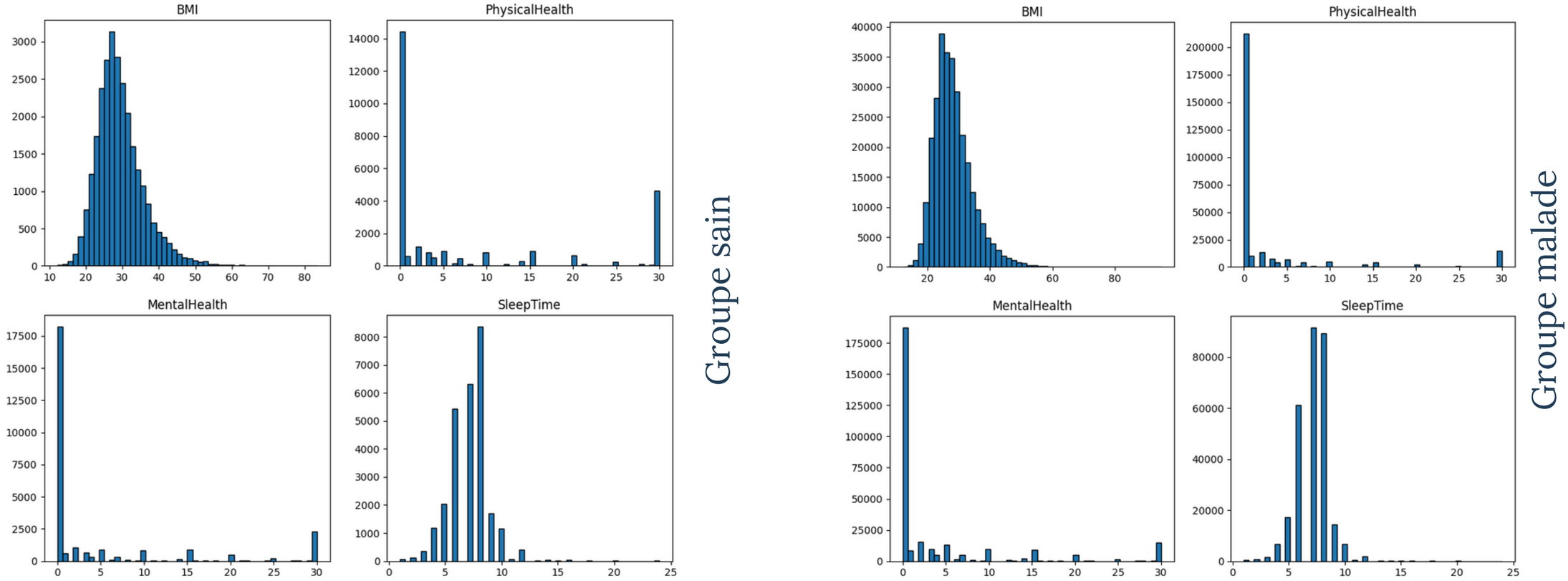


Groupe malade

Analyse exploratoire

Normalité

- Vérifiée pour les variables BMI et SleepTime.
- Non vérifiée pour les variables PhysicalHealth et MentalHealth.



On décide de passer les variables PhysicalHealth et MentalHealth en variables catégorielles.

Machine Learning

Quel algorithme utiliser ?

K-Nearest Neighbors	Modèle qui repose sur le principe que des points similaires peuvent être trouvés à proximité les uns des autres.
Logistic Regression	Modèle statistique de régression qui permet de prédire la probabilité qu'un événement arrive ou non.
Random Forest	Effectue un apprentissage sur de multiples arbres de décision entraînés sur des sous-ensembles de données légèrement différents.

Machine Learning

Quels sont les différentes étapes d'implémentation ?
(librairie scikit_learn)

1 - Séparation des données

train_test_split()

Génération d'un jeu d'entraînement et d'un jeu test.
"random_state" : assure la reproductibilité des résultats
"stratify" : assure le respect de la repartition de la cible
dans les nouveaux jeux de données.

Quels sont les différentes étapes d'implémentation ?
(librairie scikit_learn)

2 - Optimisation des paramètres

GridSearchCV ()

Création d'un dictionnaire de paramètres à tester que l'on passe à la fonction GridSearchCV() qui teste les différentes combinaisons de ces paramètres.

Quels sont les différentes étapes d'implémentation ?
(librairie scikit_learn)

3 - Récupération des informations du modèle

gscv.best_estimator_ et gscv.best_score_

Enregistrement du modèle dans une variable afin de pouvoir le réutiliser et récupération du meilleur score (accuracy) d'apprentissage afin de pouvoir évaluer la performance du modèle.

Quels sont les différentes étapes d'implémentation ?
(librairie scikit_learn)

4 - Prédictions sur le jeu test

predict()

On teste le modèle sur le jeu de données afin de récupérer les prédictions associées à chacun des individus testés.

Quels sont les différentes étapes d'implémentation ?
(librairie scikit_learn)

5 - Mesure de l'adaquation des prédictions

accuracy_score()

Vérification de l'adéquation des prédictions aux étiquettes connues.

Machine Learning

K-Nearest Neighbors | Logistic Regression | Random Forest

Paramètres optimisés

n_neighbors

Résultats de l'optimisation

45

Meilleur score

91.6 %

Accuracy

91.6 %

Machine Learning

K-Nearest Neighbors | Logistic Regression | Random Forest

Paramètres optimisés
n_neighbors

Résultats de l'optimisation
45

Meilleur score
91.6 %

Accuracy
91.6 %

Paramètres optimisés
max_iter | C | penalty

Résultats de l'optimisation
10 000 | 5.96 | L2

Meilleur score
91.6 %

Accuracy
91.6 %

Machine Learning

K-Nearest Neighbors

Paramètres optimisés
n_neighbors

Résultats de l'optimisation
45

Meilleur score
91.6 %

Accuracy
91.6 %

Logistic Regression

Paramètres optimisés
max_iter | C | penalty

Résultats de l'optimisation
10 000 | 5.96 | L2

Meilleur score
91.6 %

Accuracy
91.6 %

Random Forest

Paramètres optimisés
n_estimators | max_features

Résultats de l'optimisation
700 | 1

Meilleur score
90.5 %

Accuracy
90.6 %

Machine Learning

K-Nearest Neighbors	Logistic Regression	Random Forest
Accuracy	**Accuracy**	**Accuracy**
91.6 % (91.6)	91.6 % (91.6)	90.6 % (90.5)

Machine Learning

K-Nearest Neighbors

Accuracy
91.6 % (91.6)

Logistic Regression

Accuracy
91.6 % (91.6)

Random Forest

Accuracy
90.6 % (90.5)

Overfitting : Non ! Underfitting : Non !

Modèles cohérents et valides

Machine Learning

K-Nearest Neighbors	Logistic Regression	Random Forest
Accuracy	**Accuracy**	**Accuracy**
91.6 % (91.6)	91.6 % (91.6)	90.6 % (90.5)

Overfitting : Non ! Underfitting : Non !

Modèles cohérents et valides

Quel modèle choisir ?

Machine Learning

K-Nearest Neighbors	Logistic Regression	Random Forest
Accuracy	**Accuracy**	**Accuracy**
91.6 % (91.6)	91.6 % (91.6)	90.6 % (90.5)

Overfitting : Non ! Underfitting : Non !

Modèles cohérents et valides

Quel modèle choisir ?

K-Nearest Neighbors

Application

Streamlit

1 - Prise en main de l'outil
2 - Test de la fonctionnalité pandas-profiling
3 - Mise en place d'un formulaire adapté
4 - Utilisation de la fonction dataPreprocessing
5 - Utilisation du modèle pour prédire la donnée nouvellement entrée.

Démonstration

MedicalAssistant

diginamic FORMATION

« Votre passeport pour l'emploi numérique »

Merci pour votre attention !

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