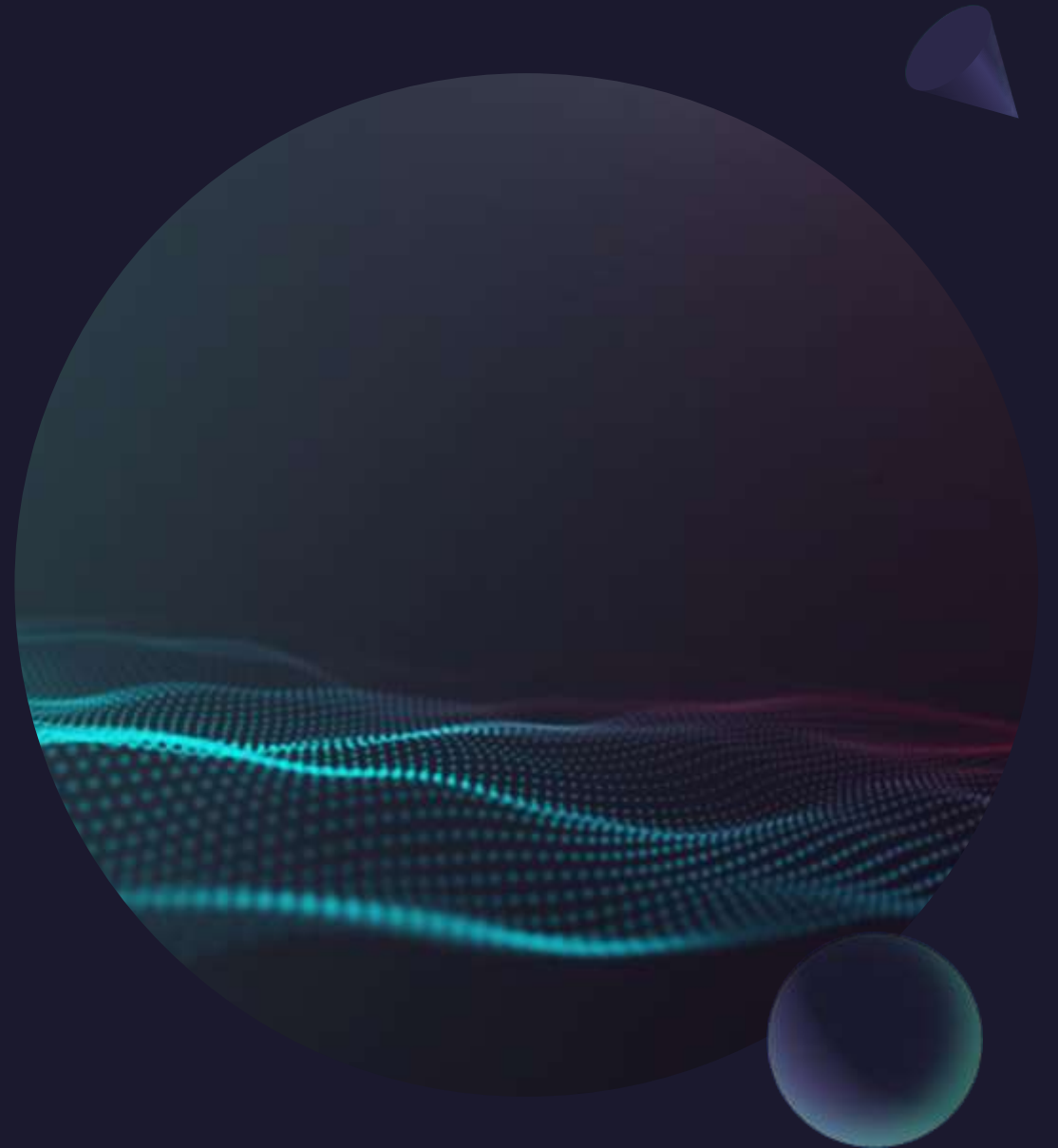


Estimation des niveaux d'obésité

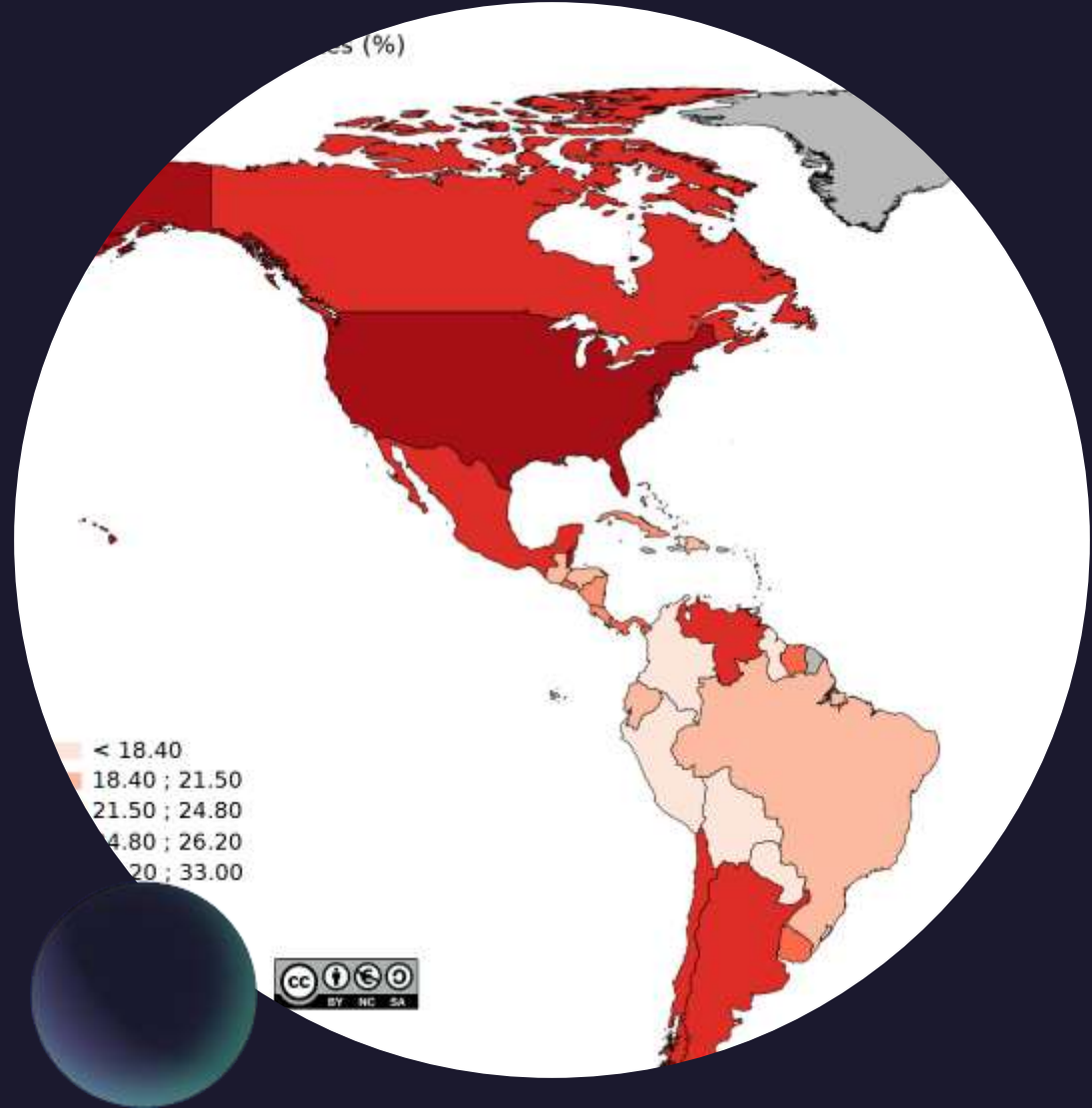
Par Victor DULEBA & Albéric
DUFAURE

Github du projet :



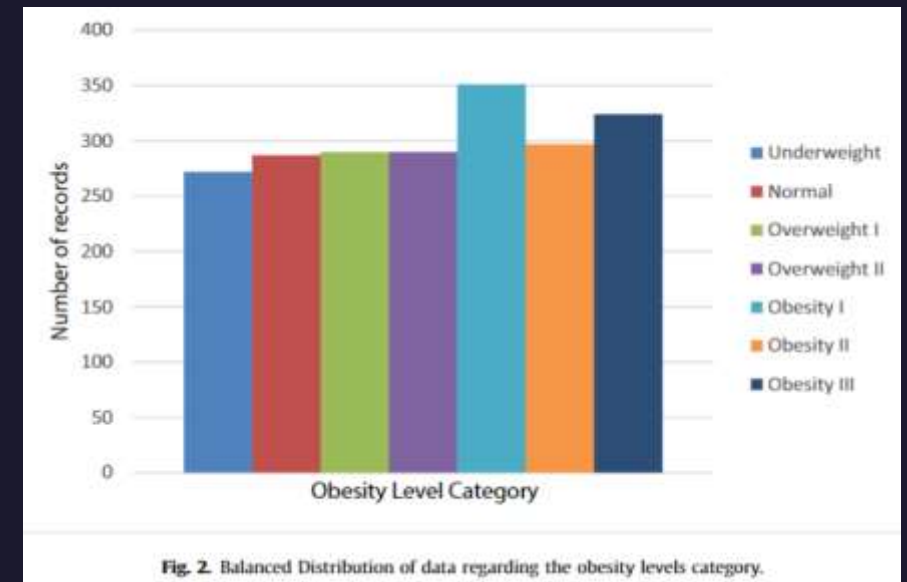
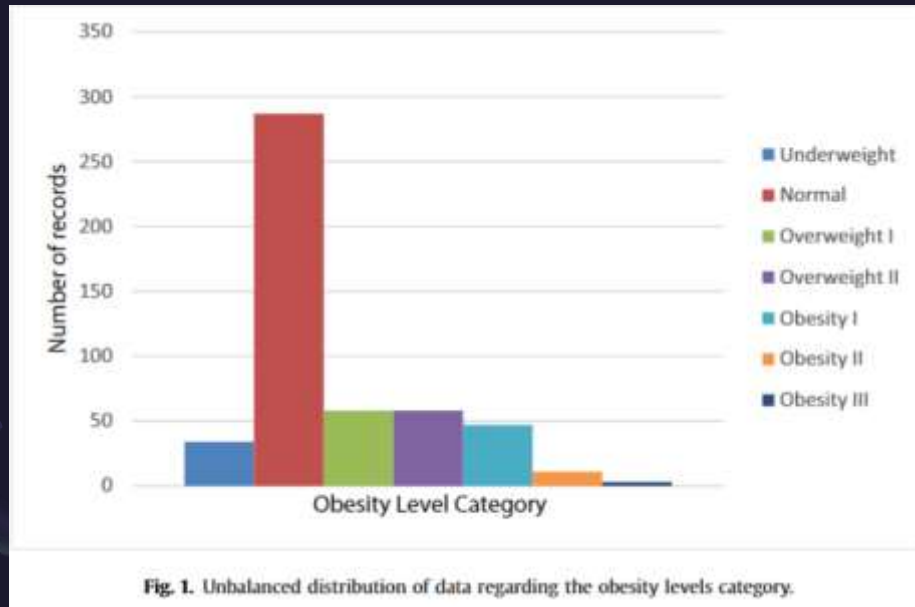
L'obésité est un problème majeur en Amérique

- L'obésité cause de nombreux problèmes de santé majeurs.
- Cette maladie est de plus en plus présente dans les pays sur lesquels porte notre étude.
- Ces pays sont:
- Le Mexique (28% des adultes en **obésité**)
- Le Pérou (21% des adultes en **obésité**)
- La Colombie (21% des adultes en **obésité**)



Le dataset

- Créé par Fabio Mendoza Palechor et Alexis de La Hoz Manotas.
- Données recueillies grâce à un sondage en ligne.
- 77% de données générées en + grâce à Weka et SMOTE dans le but d'égaliser la répartition du nombre d'individu dans chaque niveau de corpulence.



Comment générer un outil intelligent pour identifier les niveaux d'obésité ?

Pour répondre au problème posé nous avons suivi ces étapes :

1. Importation des données
2. Exploration des données
3. Visualisation des données
4. Pre-processing des données
5. Création des modèles
6. Optimisation des hyper-paramètres
7. Exportation du modèle retenu
8. Déploiement d'une API basée sur le modèle retenu




Exploration des données

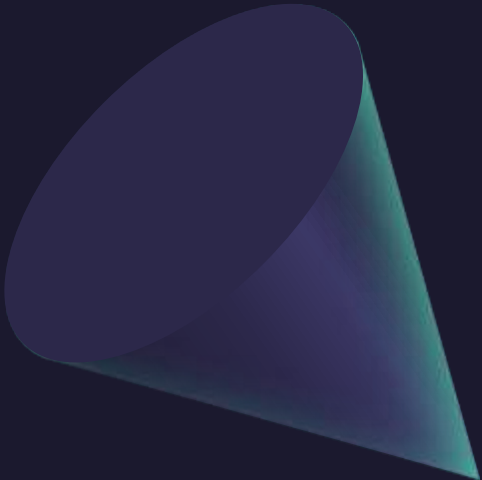

- 17 paramètres sur la fréquence de consommation de certains types d'aliments , d'exercice physique, ainsi que sur le type de moyen de transport utilisé, le genre, etc. (La taille et le poids sont présents mais nous les avons retirés car ils rendaient la tâche de prédiction trop aisée).
- 2111 individus, soit relativement peu mais le grand nombre de paramètres a permis de bons résultats
- 8 classes à prédire: Insufficient Weight, Normal Weight, Overweight Level I, Overweight Level II, Obesity Type I, Obesity Type II and Obesity Type III. Ces classes sont basés sur la valeur d'IMC.

$$\text{Mass body index} = \frac{\text{Weight}}{\text{height} \cdot \text{height}}$$





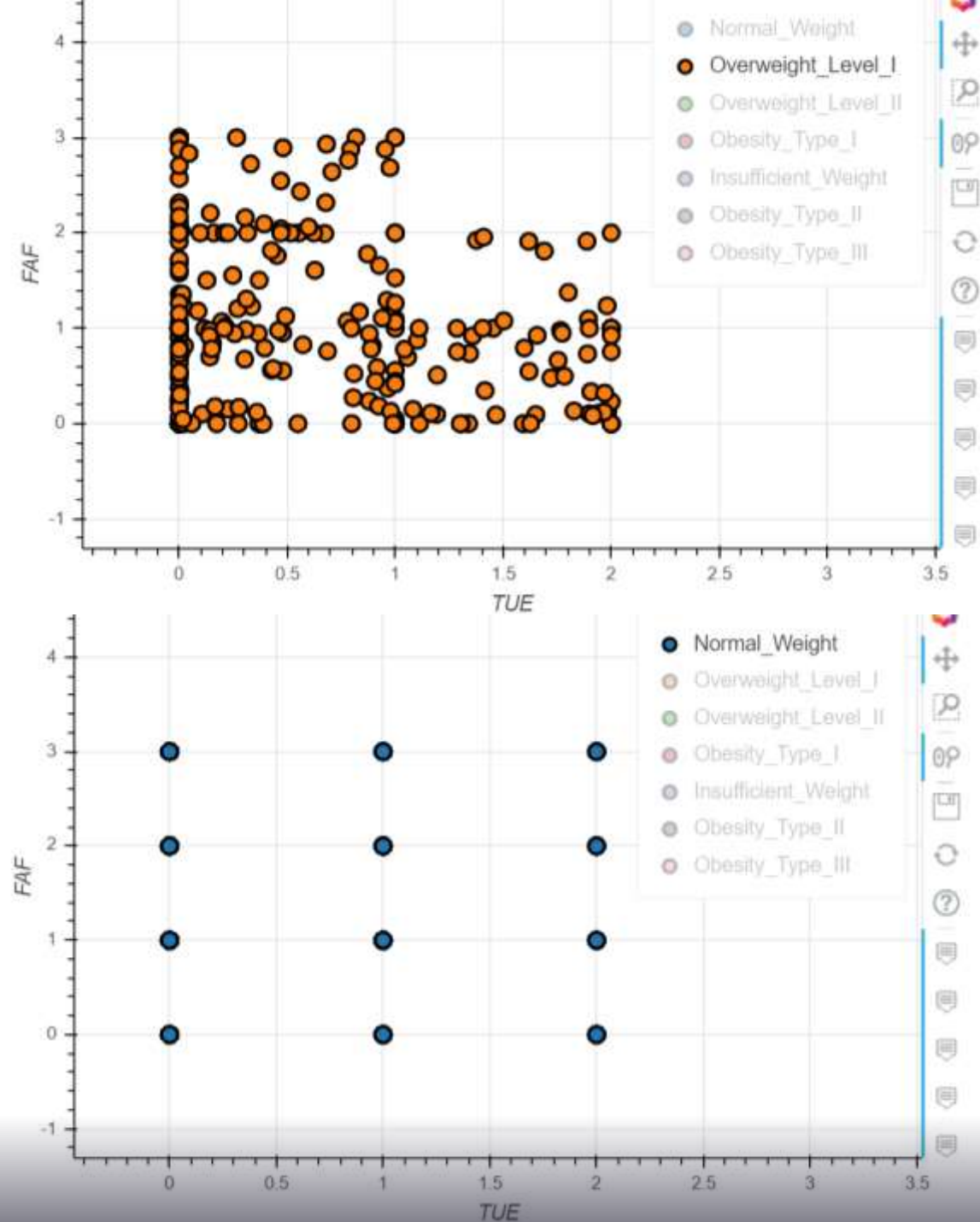
Nous allons maintenant nous
intéresser aux 6 étapes les
plus intéressantes ...



Visualisation des données

Pour voir toutes les data visualisations et reflexion à propos de ceux-ci je vous invite à consulter le notebook...

En dehors des visualisations dans le notebook, nous avons remarqués quelques valeurs étranges dans nos données (voir figure). Nous avons tracé la fréquence d'activité sportive en fonction du temps passé devant les écrans et nous observons des valeurs réparties parfaitement régulièrement pour les gens de poids normaux. Nous pensons que cela est dû au fait que lorsque l'on remplit un sondage, nous allons naturellement rentrer un chiffre rond. En revanche lorsque les données sont générées (pour la catégorie surpoids type I par ex), la machine va générer des chiffres à virgule ce qui donnera une répartition moins régulière.



Pre processing des données

- Mise les données à l'échelle grâce à la fonction StandardScaler car l'ordre de grandeur des données numériques n'était pas toujours le même.
- Catégorisation des valeurs de type string (yes = True, no = False, Sometimes = 1, Frequently=2 ...)
- One hot encoding des colonnes "Public_Transportation" et "Genre"

Nous n'avons pas eu + de traitement à faire que ceux cité au-dessus car le dataset était déjà relativement "propre" (aucune valeur manquante).



Création des modèles

- Fonction `run_models` pour comparer 5 modèles de Machine Learning.
 - Random Forest
 - Decision Trees
 - K nearest neighbours
 - Support vector machines
 - Gradient Boosting
- Nous avons effectué 5 validations croisées pour chacun de ces modèles, puis nous avons évalué leurs performances à l'aide de la fonction `classification_report` qui donne la précision (proportion d'attribution correctes pour la classe sur son nombre d'attribution total), le rappel (proportion d'attribution correctes pour la classe sur son nombre d'individus réels), le score F1 (moyenne harmonique des 2 précédents) et le support (nombre total d'éléments par classe).
- Les meilleurs résultats étaient systématiquement obtenus avec l'algorithme Random Forest, nous l'avons donc retenu pour la suite.

Optimisation des hyper-paramètres

- Résultats de la fonction `hyper_tune` appliquant l'algorithme `GridSearch` pour trouver les meilleurs hyper-paramètres de l'algorithme `Random Forest`.

Accuracy Score = 0.87

```
{'criterion': 'entropy', 'max_depth': 90, 'max_features': 'log2', 'min_samples_split': 3, 'n_estimators': 90}
```

Classification Report:

	precision	recall	f1-score	support
Insufficient_Weight	0.95	0.90	0.92	80
Normal_Weight	0.62	0.86	0.72	65
Obesity_Type_I	0.91	0.89	0.90	91
Obesity_Type_II	0.86	0.97	0.92	72
Obesity_Type_III	0.96	1.00	0.98	73
Overweight_Level_I	0.93	0.73	0.82	70
Overweight_Level_II	0.89	0.70	0.78	77
accuracy			0.87	528
macro avg	0.87	0.86	0.86	528
weighted avg	0.88	0.87	0.87	528

Déploiement de l'API

L'API fait office de rendu final. Vous pouvez rentrer les paramètres et obtenir votre prédiction.

Obesity Form

Legend

FHVC	FCVC	NCP	CABC	CH2O	SCC	FAI	TUE	CALC
Frequent consumption of high caloric food	Frequency of consumption of vegetable - input : float	Number of main meals - input : integer	Consumption of food between meals - input : integer	Consumption of water daily (l) - input : float	Calories consumption monitoring - input : float	Physical activity frequency - input : float	Time using technology devices - input : float	Consumption of alcohol - input : no = 0, sometimes = 1, frequently = 2, always = 3

For Male/Female, Walking, Motorbike, Public Transportation, Automobile and Bike fields, fill with **0** or **1**

Age:

family_history_with_overweight: ☐

FAVC: ☐

FCVC:

NCP:

CABC:

SMOKE: ☐

CH2O:

SCC: ☐

FAI:

TUE:

CALC:

Male:

Welcome to my new API !

Here you can have predictions of corpulence based on eating habits and physical activities

[Predict now](#) [Predict API](#) [Test API Visualization](#)

RESULT

BASED ON THE PREVIOUS FORM, WE MADE THE PREDICTION FOLLOWING :

NORMAL_WEIGHT

Tableau des résultats

RESULTS

Legend

FAVC	FCVC	NCP	CAEC	CH2O	SCC	FAF	TUE	CALC
Frequent consumption of high caloric food	Frequency of consumption of vegetable	Number of main meals	Consumption of food between meals	Consumption of water daily	Calories consumption monitoring	Physical activity frequency	Time using technology devices	Consumption of alcohol

Age	family_history_with_overweight	FAVC	FCVC	NCP	CAEC	SMOKE	CH2O	SCC	FAF	TUE	CALC	Male	Female	Walking	Motorbike	Public Transports	Automobile	Prediction
19.0	True	True	3.0	1.0	0	False	1.0	True	0.0	0.0	0	1	0	0	1	0	0	Normal_Weight
30.686701	True	True	2.96405	2.123138	0	False	1.0	False	1.699181	0.620465	0	1	0	0	0	1	0	Obesity_Type_II
22.307413	True	True	2.049112	2.622955	0	False	2.280555	False	2.052896	0.896185	0	0	1	0	0	1	0	Obesity_Type_I
40.501722	True	True	2.294259	2.850948	0	False	1.87029	False	0.917014	0.0	0	1	0	0	0	0	1	Obesity_Type_II
22.970655	True	True	2.478891	3.371832	0	False	2.004126	False	2.0	0.327376	0	0	1	0	0	1	0	Insufficient_Weight
33.049121	True	True	2.0	1.171027	0	False	2.405172	False	2.300292	0.0	0	1	0	0	0	0	1	Overweight_Level_II
23.384374	True	True	2.712747	2.853676	1	False	1.526313	False	1.0	0.173665	1	1	0	0	0	1	0	Overweight_Level_II