NETTOYAGE

|  |  |  |
| --- | --- | --- |
| **Taille du dataset** | Action réalisé sur le jeu de données | Conséquences |
| **(320772, 162)** |  |  |
| **(320772, 54)** | Suppression des colonnes contenant > 80% NaN | -108 colonnes |
| **(262833, 54)** | Suppression des lignes si aucune valeur nutritionnelle n’est renseignée | -57939 lignes |
| **(262817, 54)** | Suppression : 23 lignes erronées | -23 lignes |
| **(259434, 54)** | Suppression de 3383 produits sans nom | -3383 lignes |
| **(241155, 54)** | Suppression des doublons | -79617 lignes |
| **Après cette première phase de nettoyage, il nous reste 75.18% des lignes (produits) et 33.33% des colonnes (variables) du jeu de données.** | | | |
| **Gestion des données aberrantes** | | | |
| **(240818, 54)** | Energy\_100g >3800kj | -337 lignes |
| **(240641, 54)** | les éléments \_100g > 100 | -177 lignes |
| **(240624, 54)** | les éléments \_100g < 100 | -17 lignes |
| **(240297, 54)** | Incohérences entre sodium/salt, fat et satured\_fat | -327 lignes |
| **Il nous reste 74.91% des lignes et 33.33% des colonnes du jeu de données.** | | | |
| **SELECTION DES DONNEES** | | | |
| **(240297, 23)** |  | -31 colonnes |
|  |  |  |
| **IMPUTATIONS des données sélectionnées** | | | |
| **(240297, 23)** |  |  |

|  |  |  |
| --- | --- | --- |
| Démarche | Nb Ligne | Nb var |
| Chargement du jeu de données | 320772 | 162 |
| Suppression des variables avec %NaN > 80% | 320772 | 54 |
| Suppression des lignes si aucune valeur nutritionnelle n’est renseignée | 262833 | 54 |
| Suppression : 23 lignes erronées | 262817 | 54 |
| Suppression de 3383 produits sans nom | 259434 | 54 |
| Suppression des doublons | 241155 | 54 |
| Energy\_100g >3800kj | 240818 | 54 |
| Les éléments \_100g > 100 | 240641 | 54 |
| Les éléments \_100g < 100 | 240624 | 54 |
| SELECTION DES DONNEES | 240297 | 23 |
|  |  |  |

Gestion des données manquantes

|  |  |
| --- | --- |
| **Variables** | **Démarche** |
| **Additive\_n** | Fillna(0) |
| **ingredients\_from\_palm\_oil\_n** | Fillna(0) |
| **country** | Fillna(‘monde’) |
| **brands** | Fillna(‘inconnue’) |
| **Additive\_fr** | Fillna(‘ ‘) |
| **Nutrition\_grade\_fr** | Fillna(0) |
| **fiber\_100g**  **nutrition\_score\_fr\_100g**  **saturated\_fat\_100g**  **carbohydrates\_100g**  **fat\_100g**  **sugars\_100g**  **sodium\_100g**  **salt\_100g**  **proteins\_100g**  **energy\_100g** | knnImputer des 5 plus proches voisins |