

A decorative graphic on the left side of the slide consisting of two overlapping parallelograms. The front one is blue and the back one is a light greenish-blue. They are positioned diagonally, with the blue one partially covering the green one.

Analyse de mesures anthropométriques

David Abulius, Romain Lagarde, Yanis Marmier

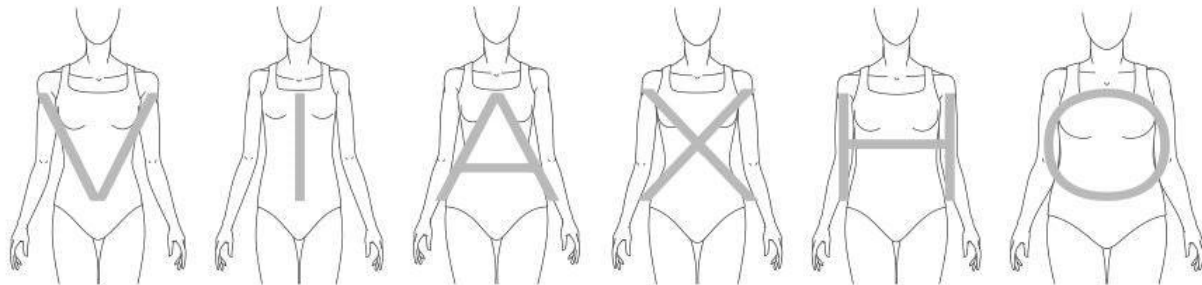


PLAN

1. Présentation globale du projet
 - a. Description
 - b. Objectifs
 - c. Outils
2. Base de données:
 - a. Présenter les bases
 - b. Attester de leur véracité
 - c. Normalisation
3. Objectif; la prédictions de données anthropométriques.
 - a. Corrélation et clustering.
 - i. Corrélation linéaire
 - ii. Motivation derrière le clustering
 - iii. Clustering : PCA et TSNE
 - iv. Modélisations.
 - b. Prédictions
 - i. Régression linéaire
 - ii. Arbre de tri
 - iii. Réseau RNN
4. Conclusion .

Description

Kleep




Objectifs



Cohérence des
données

Recherche de
clusters

Mesure de
variabilité



Recherche de
corrélations

Prédictions de
mesure



Outils

organisation



langage



Les bases de données

| Trouser | Waist_Height_Front | Ankle_Girth_Left | Ankle_Girth_Right | Calf_Height | Ankle_Height | Ankle_Circ | Thigh_Height | Shoulder_Width_ThruTheBody | Hip_Width_ThruTheBody | Overarm_Width_ThruTheBody | photo_names |
|---------|--------------------|------------------|-------------------|-------------|--------------|------------|--------------|----------------------------|-----------------------|---------------------------|-------------|
| 1029.0 | 271.0 | 263.0 | 455 | 706 | 267.0 | 35 | 379.0 | 400.0 | | 484 | male0181 |
| 1164.0 | 292.0 | 274.0 | 460 | 740 | 283.0 | 40 | 409.0 | 450.0 | | 488 | male0195 |
| 1127.0 | 307.0 | 306.0 | 450 | 746 | 306.0 | 40 | 437.0 | 458.0 | | 527 | male0803 |
| 1147.0 | 298.0 | 294.0 | 500 | 784 | 296.0 | 40 | 451.0 | 423.0 | | 542 | male0817 |
| 1069.0 | 266.0 | 270.0 | 410 | 693 | 268.0 | 40 | 368.0 | 402.0 | | 452 | male0142 |
| 1062.0 | 267.0 | 254.0 | 455 | 733 | 261.0 | 40 | 407.0 | 403.0 | | 493 | male0624 |
| 951.0 | 255.0 | 251.0 | 400 | 635 | 253.0 | 40 | 404.0 | 410.0 | | 489 | male0630 |
| 1183.0 | 286.0 | 286.0 | 545 | 877 | 286.0 | 45 | 385.0 | 430.0 | | 464 | male0156 |
| 955.0 | 254.0 | 249.0 | 385 | 626 | 252.0 | 40 | 371.0 | 390.0 | | 476 | male1248 |
| 1015.0 | 271.0 | 279.0 | 445 | 721 | 275.0 | 40 | 394.0 | 395.0 | | 459 | male1260 |
| 1013.0 | 282.0 | 285.0 | 445 | 696 | 283.0 | 40 | 298.0 | 391.0 | | 404 | male0618 |
| 961.0 | 264.0 | 257.0 | 420 | 658 | 261.0 | 40 | 349.0 | 397.0 | | 441 | male1274 |
| 1038.0 | 243.0 | 249.0 | 445 | 706 | 246.0 | 40 | 365.0 | 381.0 | | 446 | male0383 |
| 1133.0 | 269.0 | 266.0 | 475 | 789 | 267.0 | 40 | 372.0 | 406.0 | | 455 | male0397 |
| 1016.0 | 234.0 | 233.0 | 450 | 686 | 233.0 | 40 | 336.0 | 313.0 | | 389 | male1089 |
| 1218.0 | 297.0 | 287.0 | 510 | 812 | 292.0 | 45 | 381.0 | 402.0 | | 461 | male0426 |
| 1195.0 | 282.0 | 305.0 | 480 | 771 | 294.0 | 40 | 399.0 | 402.0 | | 456 | male0340 |
| 956.0 | 255.0 | 247.0 | 415 | 652 | 251.0 | 40 | 323.0 | 327.0 | | 376 | male0354 |
| 1089.0 | 267.0 | 272.0 | 440 | 731 | 269.0 | 40 | 368.0 | 374.0 | | 462 | male0432 |
| 1193.0 | 303.0 | 314.0 | 505 | 772 | 308.0 | 40 | 407.0 | 418.0 | | 522 | male1062 |
| 1090.0 | 271.0 | 272.0 | 470 | 758 | 271.0 | 40 | 425.0 | 430.0 | | 554 | male1076 |
| 1116.0 | 269.0 | 281.0 | 490 | 780 | 275.0 | 40 | 403.0 | 406.0 | | 466 | male0368 |
| 990.0 | 268.0 | 279.0 | 360 | 631 | 273.0 | 40 | 393.0 | 421.0 | | 481 | male0591 |
| 1072.0 | 269.0 | 284.0 | 480 | 757 | 276.0 | 40 | 389.0 | 380.0 | | 458 | male0585 |
| 1031.0 | 245.0 | 245.0 | 450 | 719 | 245.0 | 40 | 384.0 | 387.0 | | 477 | male0552 |
| 1004.0 | 281.0 | 265.0 | 460 | 706 | 273.0 | 40 | 375.0 | 395.0 | | 432 | male0234 |
| 1035.0 | 310.0 | 297.0 | 450 | 733 | 303.0 | 40 | 428.0 | 436.0 | | 506 | male0220 |
| 1107.0 | 275.0 | 294.0 | 460 | 715 | 284.0 | 40 | 386.0 | 411.0 | | 441 | male0546 |
| 1062.0 | 293.0 | 274.0 | 450 | 738 | 283.0 | 40 | 393.0 | 393.0 | | 506 | male0208 |
| 989.0 | 415.0 | 272.0 | 420 | 681 | 343.0 | 40 | 412.0 | 389.0 | | 485 | male1116 |
| 1053.0 | 249.0 | 258.0 | 450 | 724 | 253.0 | 40 | 395.0 | 400.0 | | 465 | male1102 |
| 995.0 | 263.0 | 272.0 | 425 | 696 | 267.0 | 40 | 392.0 | 406.0 | | 481 | male0793 |
| 979.0 | 306.0 | 328.0 | 405 | 676 | 317.0 | 40 | 410.0 | 460.0 | | 539 | male0787 |
| 994.0 | 278.0 | 315.0 | 360 | 641 | 296.0 | 35 | 386.0 | 487.0 | | 497 | male0977 |
| 1098.0 | 303.0 | 291.0 | 485 | 777 | 297.0 | 40 | 447.0 | 415.0 | | 528 | male0963 |
| 1016.0 | 275.0 | 416.0 | 435 | 703 | 346.0 | 35 | 339.0 | 378.0 | | 413 | male1328 |

Environ 9000 individus différents

Pas uniquement des nombres.

Données réelles.

Confidentialité.

Cohérence des données : comparer les données entre elles

Écart de 5% à 95% :

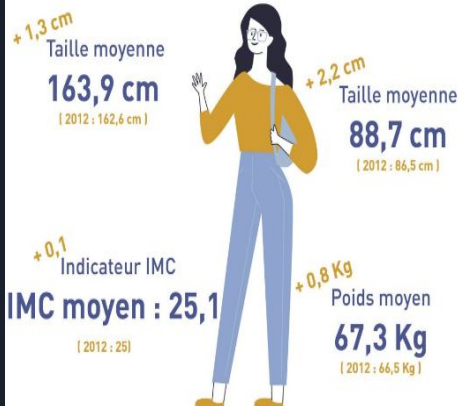
- 'Calf_Left', 'Calf_Right' : Mollet gauche, Mollet droit
- 'Calf_Left', 'Calf_Circ' : Mollet gauche, Circonférence du mollet
- 'Calf_Right', 'Calf_Circ' : Mollet droit, Circonférence du mollet
- 'TrouserWaist_Height_Back', 'Outseam' : Hauteur du pantalon à la taille (arrière), Couture extérieure
- 'TrouserWaist_Height_Back', 'Waist_Height_Back_EZ' : Hauteur du pantalon à la taille (arrière), Hauteur de la taille du pantalon (arrière, facile)
- 'Outseam', 'TrouserWaist_Height_Front' : Couture extérieure, Hauteur du pantalon à la taille (avant)
- 'Outseam', 'Waist_Height_Back_EZ' : Couture extérieure, Hauteur de la taille du pantalon (arrière, facile)

3% à 97% :

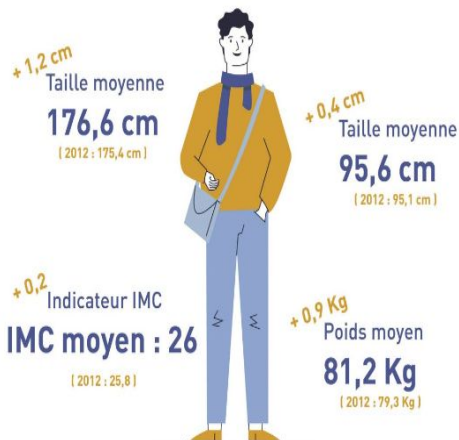
- ('Calf_Left', 'Calf_Circ')
- ('Calf_Right', 'Calf_Circ')
- ('TrouserWaist_Height_Back', 'Waist_Height_Back_EZ')
- ('Shoulder_to_floor_Right', 'Shoulder_to_floor_Left')

Cohérence des données : Comparer les données à des moyennes nationales

Chez la femme



Chez l'homme



Résultat de l'enquête ObePi Roche 2020

```
df[df["gender"] == "male"]["height_cm"].mean()
```

176.23761357429717

```
df[df["gender"] == "female"]["height_cm"].mean()
```

163.67015257628816

```
df[df["gender"] == "female"]["IMC"].mean()
```

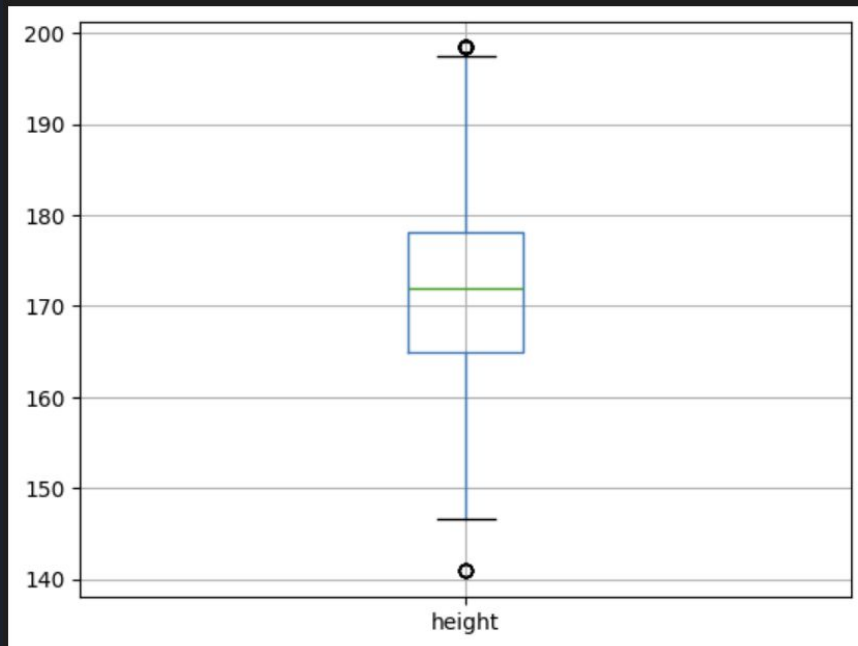
25.127469795622645

```
df[df["gender"] == "male"]["IMC"].mean()
```

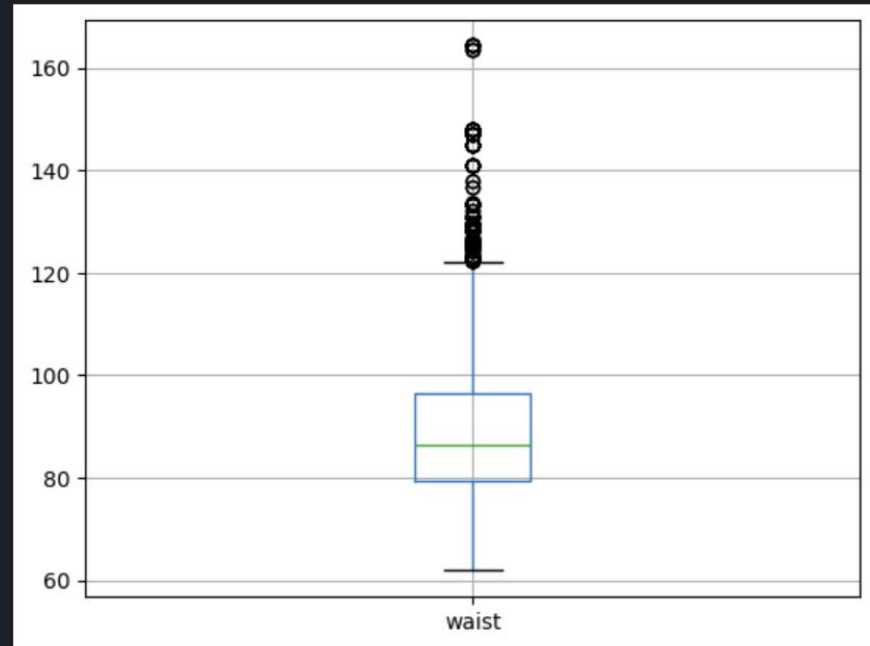
25.7032752491774

Nos données

Cohérence des données : vérifier qu'il n'y a pas de valeurs aberrantes



Répartition des tailles



Répartition des tours de taille

Nettoyage et normalisation

$$X_{standard} = \frac{X - \mu}{\sigma} ,$$


Corrélation linéaire

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Corrélation

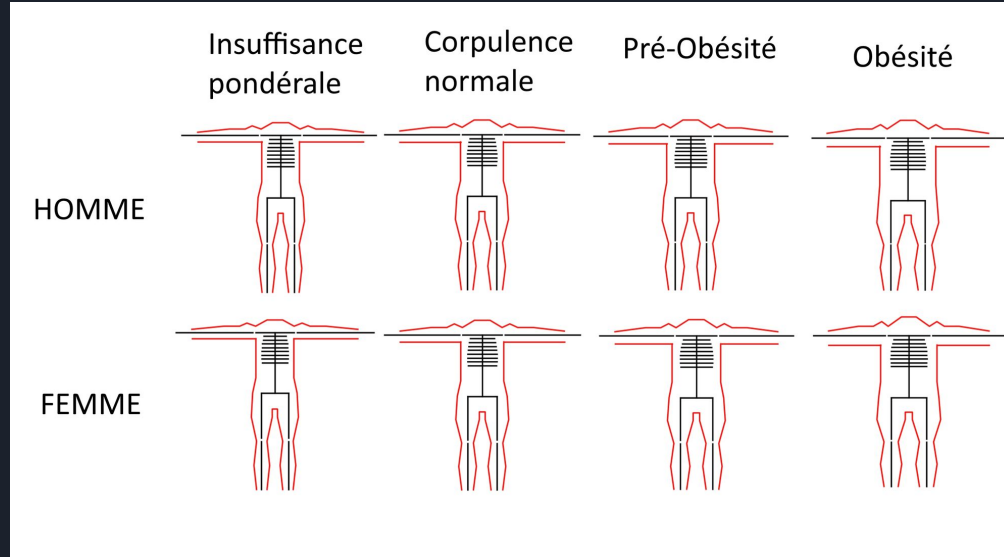
- Longueur du bras - taille : 0,9140225338321994
- Longueur du bras - longueur de la jambe : 0,9301998252048627
- Poitrine - taille de taille : 0,9301641850546507
- Poitrine - poids en kg : 0,9181131886573974
- Taille - longueur du bras : 0,9140225338321994
- Taille - longueur de la jambe : 0,906996751170562

Coefficient de
Pearson



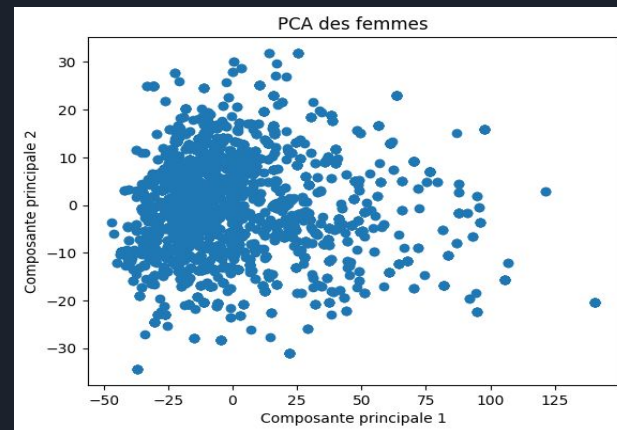
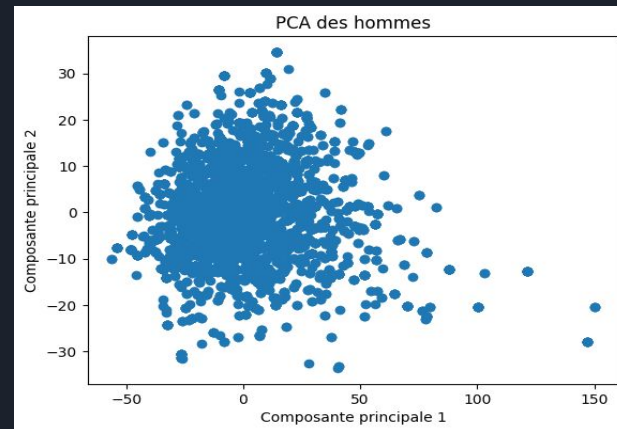
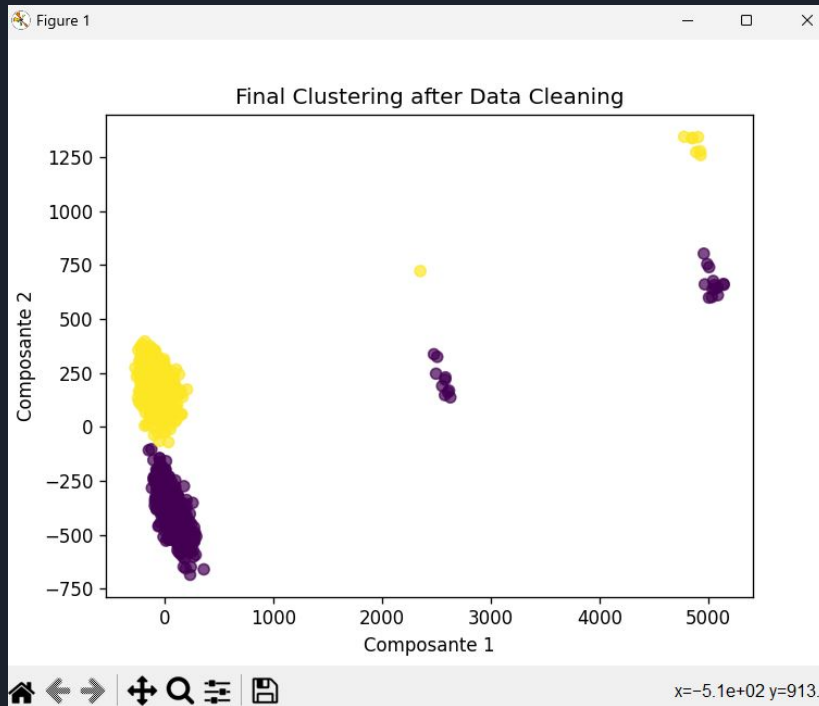
Motivation derrière le clustering

- On souhaite savoir s'il est possible de répartir les corps dans différents clusters
- Chaque cluster se voit attribuer un corps type



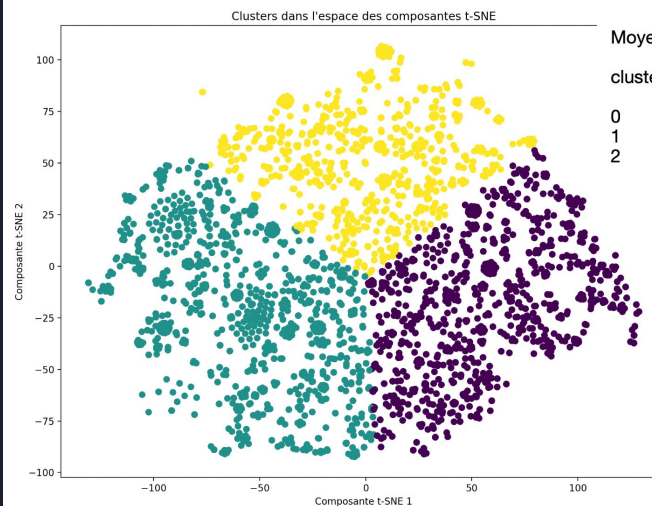
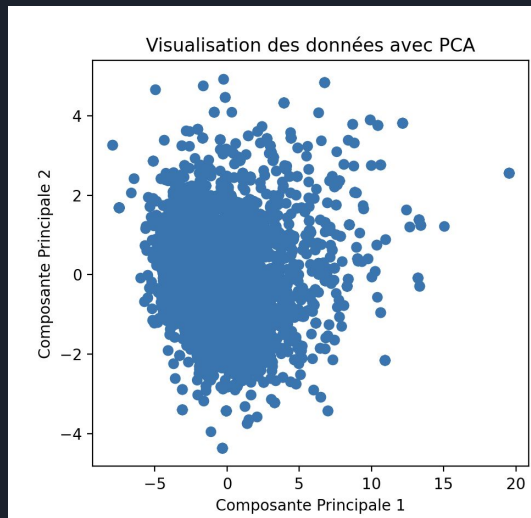
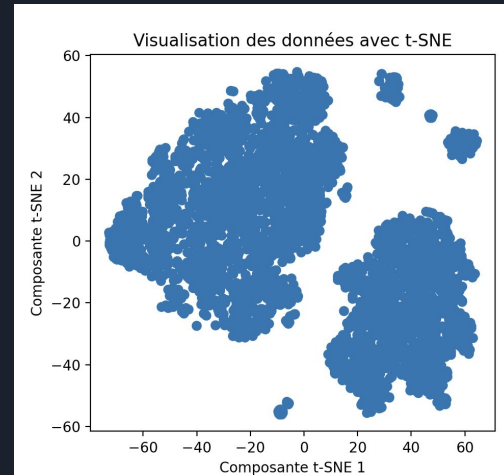
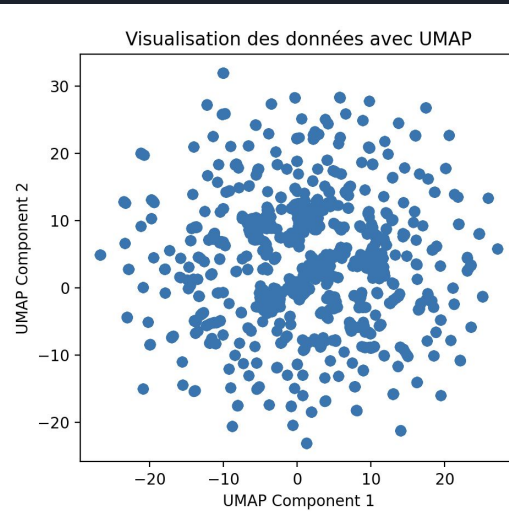
Rendu du corps type pour quelques catégories d'IMC.

Méthode PCA et clustering



Proportion de la variance expliquée par les deux premières composantes principales: 0.94

Méthode PCA, t-SNE, et Clustering



Moyenne et écart-type pour chaque cluster pour la colonne 'waist'

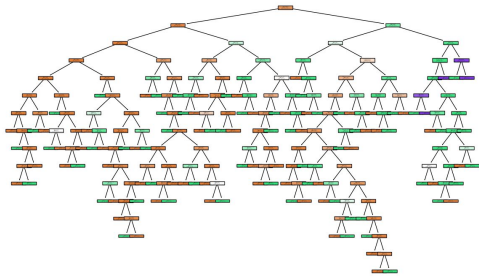
| cluster | mean | std |
|---------|------------|-----------|
| 0 | 115.856823 | 11.080243 |
| 1 | 92.820281 | 7.243426 |
| 2 | 79.271389 | 6.237599 |

Régression linéaire

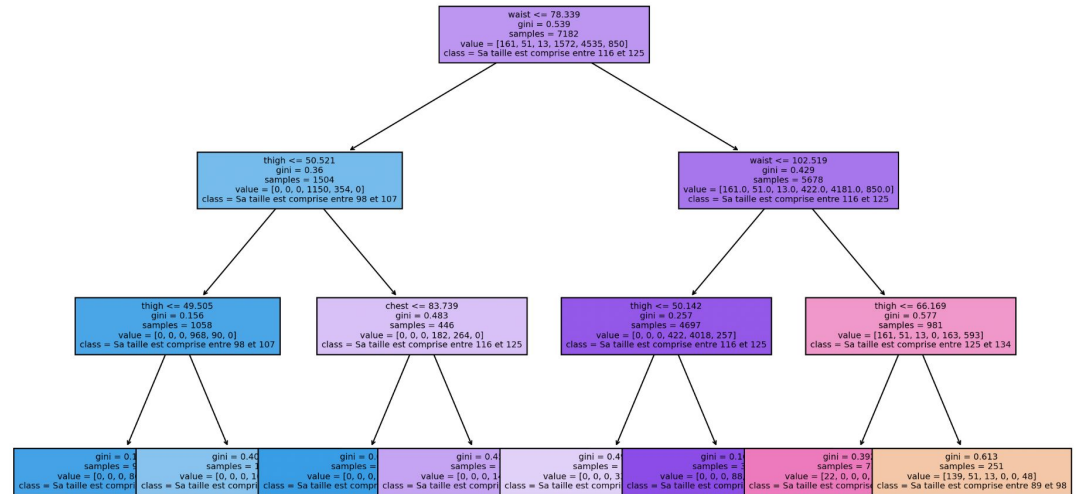
```
def deduce_missing_components(vector_with_missing, covariance_matrix, mean_vector):
    missing_indices = np.isnan(vector_with_missing)
    known_values = vector_with_missing[~missing_indices]
    unknown_indices = np.where(missing_indices)[0]
    cov_known = covariance_matrix[~missing_indices][:, ~missing_indices]
    cov_mixed = covariance_matrix[~missing_indices][:, missing_indices]
    deduced_values = np.dot(np.dot(cov_mixed.T, np.linalg.inv(cov_known)), known_values - mean_vector[~missing_indices])
    + mean_vector[unknown_indices]
    vector_with_missing[missing_indices] = deduced_values
    return vector_with_missing
```

Mean Squared Error (données normalisées): 0.01778173548586269

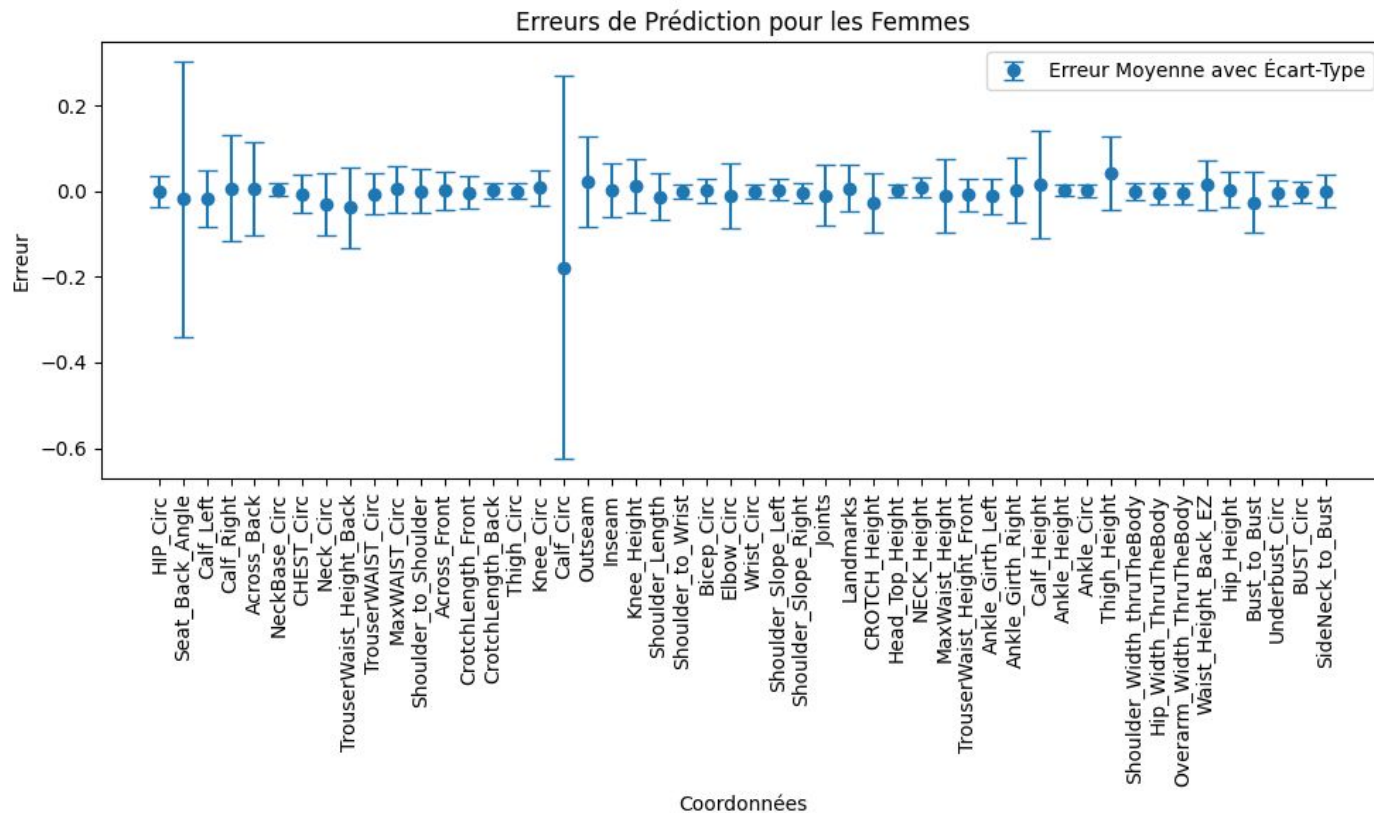
Changer d'échelle : Arbre de tri et RNN



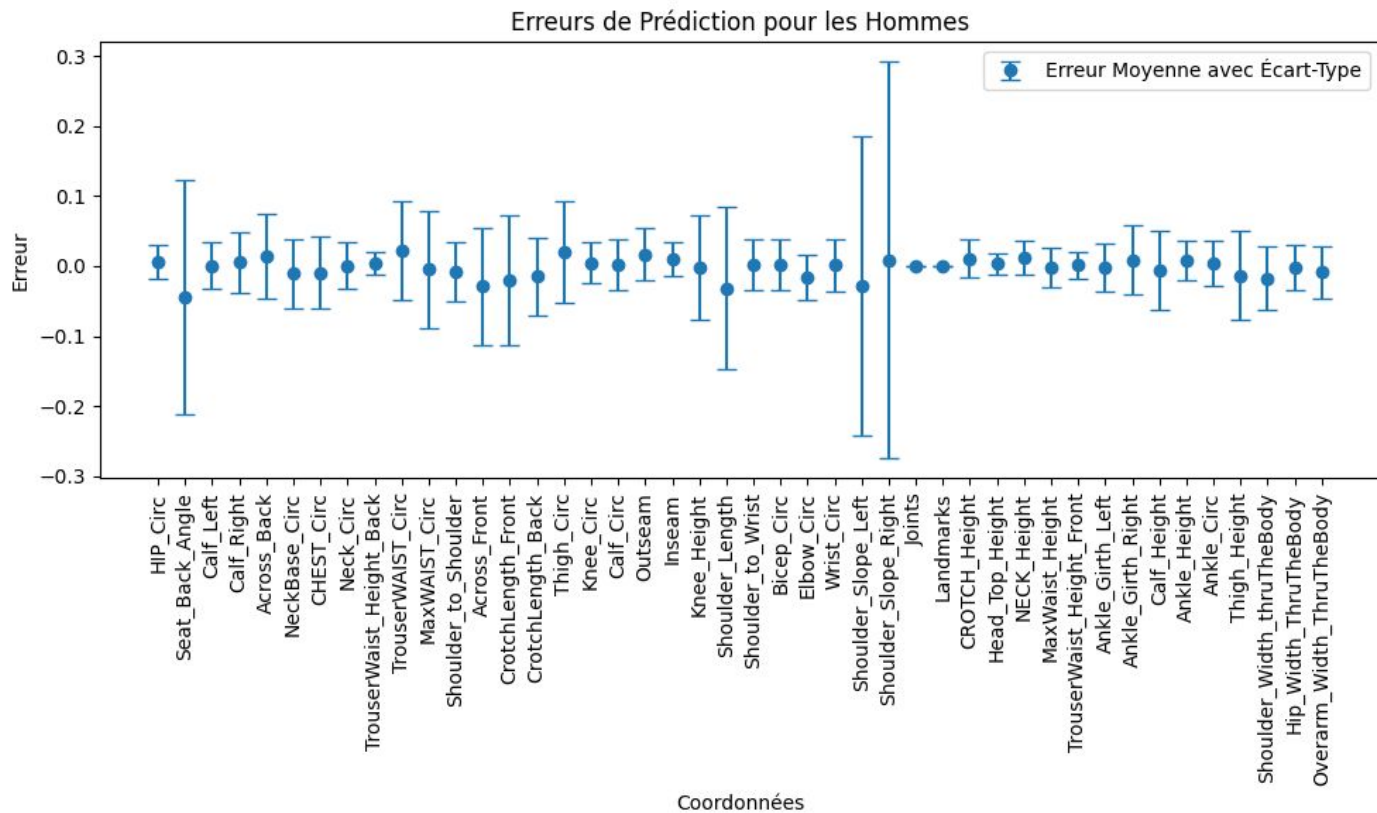
$$Gini = 1 - \sum_{i=1}^C (p_i)^2$$



Arbres de décisions, les résultats



Arbres de décisions, les résultats

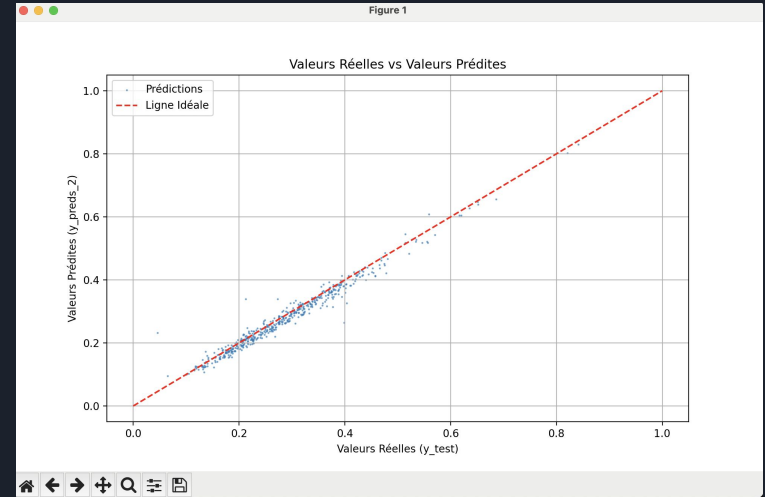


Réseau de Neurones Récurrents

-> Prédire le poids avec le moins d'éléments possibles

-> Coefficient de détermination $R^2 = 0.96$

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$



```
# Replicate model_1 and add an extra layer
model_2 = tf.keras.Sequential([tf.keras.layers.Dense(64, activation='relu', input_shape=(11,)), # Couche avec 64 neurones et input_shape=(16,)
                                tf.keras.layers.Dense(64, activation='relu'),
                                tf.keras.layers.Dense(64, activation='relu'),
                                tf.keras.layers.Dense(32, activation='relu'), # Couche avec 32 neurones
                                tf.keras.layers.Dense(1) # add a second layer
                                ])
```



Conclusion

- Différentes méthodes
- Différents cas d'usage
- À la base de tout modèle prédictif; la qualité des données.



Merci pour votre écoute