09/05/2017

Simon Baehler & Thibault Schowing

HEIG-VD

Projet Logique Floue

Y a-t-il une corrélation entre le BMI et les stéroïdes présents dans le corps ?

Table des matières

[Contexte et état de l’art du problème 2](#_Toc482109950)

[Contexte des données 2](#_Toc482109951)

[Méthode d’obtention des données 2](#_Toc482109952)

[Obtention de la valeur de sortie 2](#_Toc482109953)

[Question de recherche 3](#_Toc482109954)

[Comprendre les antécédents de modélisation 3](#_Toc482109955)

[Analyse exploratoire des données 4](#_Toc482109956)

[Modélisation par co-évolution (FUGE) 7](#_Toc482109957)

[FUGE sur une instance simple de training/validation 7](#_Toc482109958)

[Analyse du compromis (trade-off) entre performance et taille 7](#_Toc482109959)

[Analyse de la tolérance à l’incertitude 7](#_Toc482109960)

[Une des approches suivantes devra être développée comme une contribution additionnelle des étudiants (au choix) 7](#_Toc482109961)

[Apprentissage semi-supervisé (ci-dessous une approche possible, mais pas la seule alternative) 7](#_Toc482109962)

[Modélisation supervisé par apprentissage (neuro-fuzzy ou similaire) quelques alternatives sont: 7](#_Toc482109963)

[Implémentation de logique ﬂoue de type 2 7](#_Toc482109964)

[Une autre méthode trouvé dans la literature et implémentée et testée par l’équipe. 7](#_Toc482109965)

[Table des illustrations 9](#_Toc482109966)

# Contexte et état de l’art du problème

## Contexte des données

Les données proviennent du projet SKIPOGH. SKIPOGH (Swiss Kidney Project on Genes in Hypertension) est l’acronyme du Projet Suisse sur la génétique de l’hypertension et du rein; son but est d'explorer le rôle de la génétique et de l'environnement (la nutrition, l'activité physique et le mode de vie) sur la régulation de la pression sanguine ainsi que la fonction des reins. SKIPOGH vise également à enrichir les connaissances sur les statuts de santé des habitants des cantons de Lausanne, Genève et Berne. Cette étude permettra aux scientifiques d'évaluer à quel degré les paramètres tels que la tension, la taille des reins et ses fonctions, le poids corporel et le métabolisme sont héréditaires. L'étude SKIPOGH fait partie de EPOGH (Projet européen sur les gènes dans l'hypertension), une étude internationale explorant les déterminants génétiques de la tension artérielle (BP) dans la population générale et suit en grande partie le même protocole standardisé. (http://www.skipogh.ch, Consulté le 09.03.2017)

Ce set de données contient des informations concernant les quantités de divers stéroïdes naturels dans plus de 1’120 personnes. Ce set de données est composé de 40 variables d’entrée qui représentent, chacune, 1 stéroïde différent exprimée par une valeur continue et une variable de sortie représentant la valeur de l’IMC.

## Méthode d’obtention des données

Les données proviennent de prélèvements d’urine (test cumulé 12h journée 12h nuit ou 24h)

Nos données d’entrée sont les 40 stéroïdes se trouvant dans le fichier « *Skipogh Steroids\_original.csv ».* Nous possédons aussi un autre fichier du nom de « Skipogh Steroids\_cat2.csv » qui lui contient des informations relatives au participant, en voici les définitions. Contrairement au premier fichier, ces données sont d’importance moindre, car notre travail est avant tout basé sur la corrélation entre les stéroïdes et le BMI

|  |  |
| --- | --- |
| Waistc1 | Waist circumference in cm |
| Bw | Body weigh |
| Bia\_fmkg | Bio impendance - Fat mass kg |
| Hipcirc1 | Hip circumference in cm |
| ARMCICRC1\_L | Arm circumference left |
| SKINF2\_1 | Left subscapular skin fold in mm, first measure |
| bsa\_dubois | Body surface area using DuBois formula |
| BIA\_FMP | Bio impendance - Fat mass % |
| kid\_vol | Mean of right and left kidney volume |
| Crap | CHOL/HDL |
| URA | Uric acid |
| DBP5 | Fifth DBP (diastolic blood pressure |
| SBP | systolic blood pressure |
| Cuff\_sh | Cuff size |
| ckd\_epi | eGFR using CKD-EPI formula |
| Sex | Sexe de la personne |
| crp | C-reactive protein​ |

## Obtention de la valeur de sortie

La valeur de sortie correspond au BMI du patient. Nous avons pris la décision de faire une classification comme la figure ci-dessous :

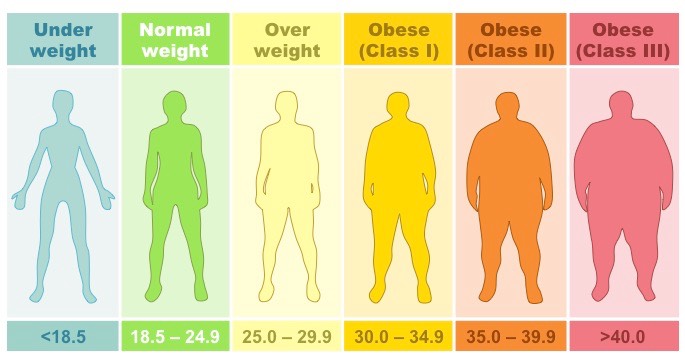


Figure 1 : Classification selon le BMI

## Question de recherche

L'objectif de ce projet est de savoir si l'indice de masse corporelle (IMC ou BMI en anglais) est lié à la quantité de stéroïdes présents dans le corps. Dans le cas où notre outil, ici FUGE, nous permettrait de réaliser une bonne classification, il faudrait trouver quels sont les stéroïdes qui sont les plus affectés par l'IMC. Dans le cas où la classification ne serait pas pertinente, d'autres questions pourraient être posées, par exemple déterminer s'il y a un impact des stéroïdes sur la pression systolique ou diastolique etc.

Notre but est de faire un outil qui réalise une classification de nos participants en fonction du BMI prédit grâce aux stéroïdes prélevés chez ces derniers. En prédisant sont BMI, nous serons capables de classer la personne dans une des six catégories présentes dans la figure 1.

Comprendre les antécédents de modélisation

D’après nos recherches, aucun antécédent ou du moins aucun article en parlant n’est paru sur le sujet précis de la corrélation entre les stéroïdes et le BMI. Nous avons trouvé quelques informations sur une éventuelle relation entre testostérone et BMI[[1]](#footnote-1) mais rien ne comparant le BMI à une grosse quantité de stéroïdes provenant de l’urine.

Analyse exploratoire des données

Dans un premier temps, nous avons réalisé des skaterplots avec le stéroïde sur l’axe des X et le BMI sur l’axe des Y afin de pouvoir détecter une éventuelle corrélation entre un stéroïde précis et le BMI de manière visuelle. Dans l’expérience, nous n’avons pas pris compte des données manquantes (pas d’utilisation de KNN), ni fait de séparation entre les hommes et les femmes.

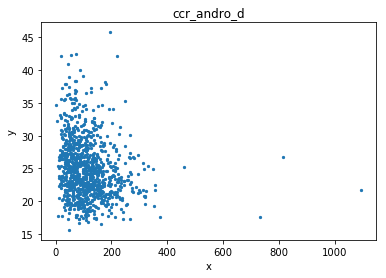


Figure Scatterplot du stéroïde ccr\_andro\_d par rapport au BMI

Après avoir pu visualiser ces données, nous avons entrepris de nettoyer les éléments qui ne contenait pas assez de données. En effet, les données de beaucoup de participants[[2]](#footnote-2) à l’étude étaient incomplètes. Le graphique ci-dessous montre le nombre de participants (y) avec X données manquantes.

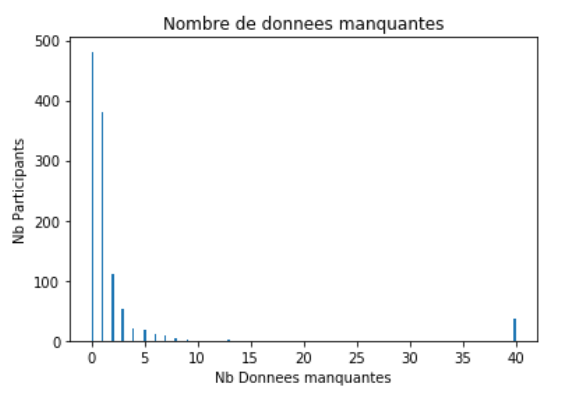


Figure Nombre de données manquantes par participant

On a donc 480 participants qui n’ont aucune donnée manquante et en revanche nous en avons une cinquantaine qui ont 40 données manquantes, donc la totalité des stéroïdes. Après avoir croisé les données avec le second ficher d’information, il s’est avéré que la cinquantaine de participants sans données, avaient déjà été éliminée du dataset. Nous avions donc déjà fait un premier nettoyage, mais il restait à combler les autres trous (voir figure 4).

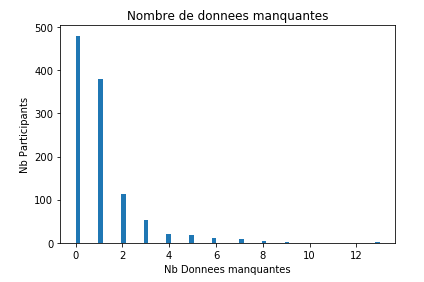


Figure Premier nettoyage

Pour combler les données manquantes, nous avons utilisé la méthode KNN proposée par la librairie PANDAS[[3]](#footnote-3) qui possède des outils prévus à cet effet. Nous avons donc transformé nos données en DataFrame Pandas et avons appliqué KNN. Mais avant, nous avons fixé une limite à 5 données manquantes par participant. Tout ceux ayant plus de 5 données manquantes sont éliminés.

Nous avons donc un dataset propre et complet et nous pouvons poursuivre l’analyse exploratoire.

Nous avons réalisé d’autres graphiques à titre expérimental afin de pouvoir mieux visualiser les données. Par exemple, quelle est la quantité d’un certain stéroïde dans l’urine « habituellement » ? Réponse ci-dessous :

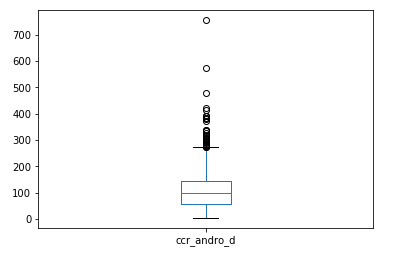
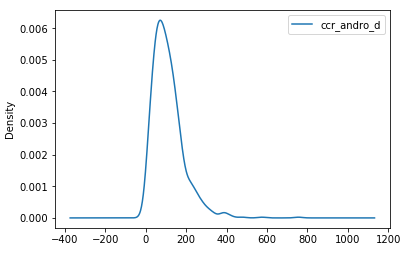


Figure Boxplot ccr\_andro\_d

Il est plus intéressant, surtout concernant les stéroïdes, de comparer les répartition homme/femme lors des mesures. Nous avons réalisé les graphiques suivant afin de se rendre compte de l’importance que peut avoir le sexe dans les mesures réalisées. Voici quelques graphiques évocateurs :

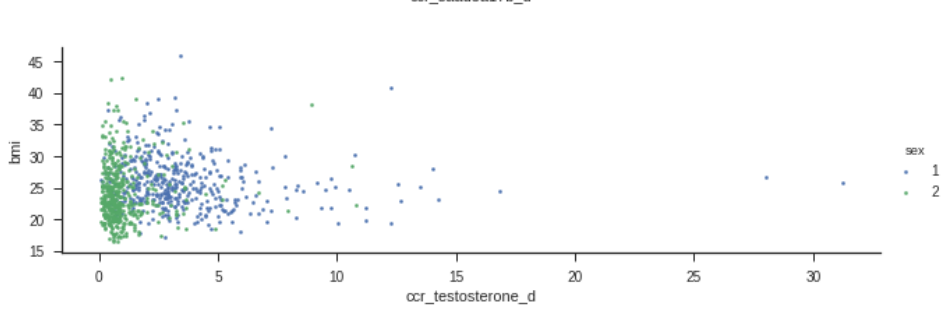


Figure Graphique testostérone femme vs homme vs BMI

Sur cette figure on réalise qu’il y a une grande différence de concentration entre homme et femme. En revanche il est très difficile de voir une corrélation entre concentration et BMI.

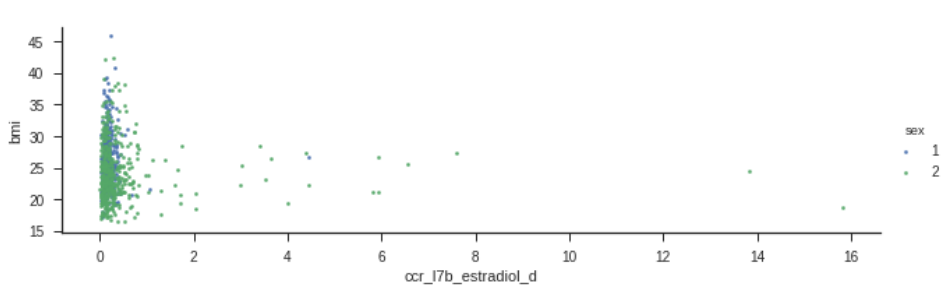


Figure Graphique estradiol\_d femme vs homme vs BMI

Sur la figure ci-dessus, on peut observer une différence entre homme et femme, certes plus faible qu’avec la testostérone mais quand même subtile. L’estradiol\_d est une hormone qui participe chez la femme au développement et au maintien des caractères sexuels. Chez l’homme, ce stéroïde est produit mais il est relativement bas sauf en cas d’affection du foie. Nous avons donc un patient mâle qui présente peut-être des problèmes de foie dans notre set de données.

À venir :

Modélisation par co-évolution (FUGE)

FUGE sur une instance simple de training/validation

Séparer les données en training/validation sets

Déﬁnir une combinaison de critères de performance (ﬁtness)

Déﬁnir et/ou chercher de façon itérative ou séquentielle les paramètres de recherche (évolution) de l’algorithme (via script)

Utiliser des paramètres du système ﬂou fournis par les enseignants

Sélectionner quelques classiﬁcateurs et valider leurs performances sur le set de test

Analyse du compromis (trade-off) entre performance et taille

Explorer les paramètres structuraux du classiﬁcateur ﬂou(règles, variables, fonctions d’appartenance?)

Analyser les chiffres et les graphiques de performance versus taille pour tirer des conclusions

Sélectionner deux classiﬁcateurs: LE meilleur en termes numériques et LE meilleur en termes d’équilibre performance / interpretabilité

Les analyser, les comparer et conclure

Analyse de la tolérance à l’incertitude

Données manquantes

Bruit / variabilité des entrées

Une des approches suivantes devra être développée comme une contribution additionnelle des étudiants (au choix)

Apprentissage semi-supervisé (ci-dessous une approche possible, mais pas la seule alternative)

Extraction de “règles” par clustering ﬂou (fuzzy c-means, adaptation de k-means, ou autre)

Affectation des sorties (conséquents) par statistiques (p.ex.simple majorité, décomposition en sub-clusters (alpha-cut) ou approche hiérarchique)

Application de critères d’Interpretabilité (dès l’extraction de règles, amélioration par post-traitement)

Modélisation supervisé par apprentissage (neuro-fuzzy ou similaire) quelques alternatives sont:

ANFIS + amélioration de l’interpretabilité

NEFCLASS/NEFCON + amélioration de la performance

ANN + extraction de règles ﬂoues interprétables

Implémentation de logique ﬂoue de type 2

Une autre méthode trouvé dans la literature et implémentée et testée par l’équipe.

# Table des illustrations

[Figure 1 : Classification selon le BMI 3](#_Toc482109891)

[Figure 2 Scatterplot du stéroïde ccr\_andro\_d par rapport au BMI 4](#_Toc482109892)

[Figure 3 Nombre de données manquantes par participant 4](#_Toc482109893)

[Figure 4 Premier nettoyage 5](#_Toc482109894)

[Figure 5 Boxplot ccr\_andro\_d 5](#_Toc482109895)

[Figure 6 Graphique testostérone femme vs homme vs BMI 6](#_Toc482109896)

[Figure 7 Graphique estradiol\_d femme vs homme vs BMI 6](#_Toc482109897)

1. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4476341/ [↑](#footnote-ref-1)
2. Les données proviennent de plusieurs ***participants*** à l’étude SKIPOGH [↑](#footnote-ref-2)
3. http://pandas.pydata.org/ [↑](#footnote-ref-3)