# Contexte et état de l’art du problème

## Analyser le contexte des données

### Contexte

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.

### Quel est la provenance des données?

Urine (test cumulé 12h journée 12h nuit ou 24h - vérifier)

### Contexte de mesure/obtention des données (entrées principalement)

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​ |

### Contexte d’obtention de la valeur de sortie

Pour nos valeurs de sortie, nous avons pris la décision de faire une classification comme la figure ci-dessous

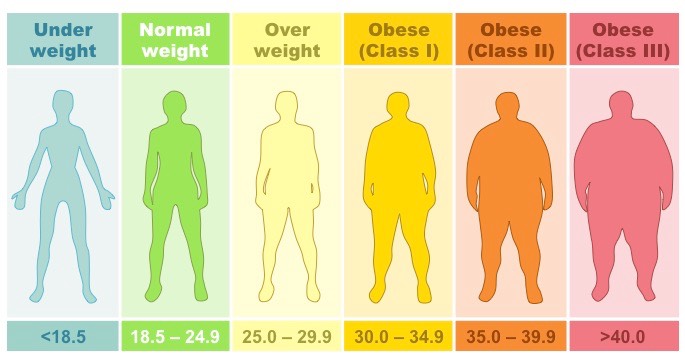


Figure : Classification

## Comprendre/déﬁnir la question à laquelle on cherche à répondre

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 l'algorithme 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.

### Classiﬁcation pure versus décision diagnostique ou autre?

Notre but est de faire un outil qui réalise une classification pure ne nos participant en fonction du BMI prédit grâce aux stéroïdes prélevé chez ces dernier. En prédisant sont BMI, nous serons capable de classer la personne dans une des six catégories présente dans la figure 1.

### Importance relative de l’explication et de la performance

Comprendre les antécédents de modélisation

### Existe-il des articles présentant des résultats sur ce problème (si non: discuter avec l’expert ou la personne de référence (prof, assistant))

### Voici quelques possibles questions à se poser:

### Quelles sont les performances atteignables?

### Quelles sont les approches utilisées/ proposées?

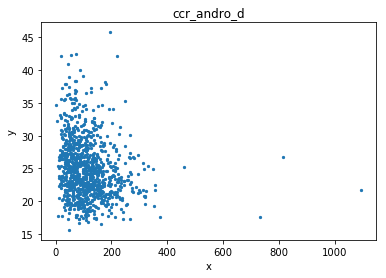
### Existe-t-il des solutions à base de logique ﬂoue?

Nous n’avons pas encore abordé cette question

Apprivoiser” et analyser les données

### Utiliser des méthodes de statistique descriptive classique pour comprendre les données: boites à moustaches, corrélations, etc.

La première approche abordée fût de simplement faire des skaterplot avec le stéroïde sur l’axe des X et le BMI sur l’axe des Y afin de pouvoir voir s’il y a une corrélation entre un stéroïde précis et le bmi, dans la expérience, nous n’avons pas compte des données manquantes (pas d’utilisation de KNN) ni fait de séparation entre les hommes et les femmes.



### Faire quelques graphiques illustrant les distributions des données autres méthodes à choix (recherche et exploration de méthodes existant sur Python)?

# Solution “experte”

## Implémenter une version “experte” du classiﬁcateur en utilisant le code Python

### Une première solution sera fournie par le professeur (et/ou l’assistant)

Analyser le système

### Analyser les performances globales du système: pourcentage de classiﬁcation, erreur numérique, sensibilité/spéciﬁcité, autres?

### Analyser la “connaissance” contenue dans le système: règles, fonctions d’appartenance,

### Analyser quelques cas

## Amélioration des performances du classiﬁcateur par “tuning” manuel d’un ou plusieurs paramètres:

### Adaptation des fonctions d’appartenance

### Ajout/élimination de variables à certaines règles et/ou

### Ajout/élimination de règles complètes

Discussion sur l’intérpretabilité de la solution

# 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.