

Projet Data Mining

Croissance des enfants

Présenté par :

Dosse Inza COULIBALY
Lore IGOUWE
Jeanne MEDENOU
Mame Diarra NOBA

Encadré par :

Mme Emilie SOHIER

SOMMAIRE

1. Problématique	2
2. Approches et résultats	3
1. Importation, visualisation et prétraitement des données	3
a. Gestion des données manquantes du sexe.....	3
b. Interpolation des valeurs manquantes de la taille et du poids	3
2. Analyse descriptive	4
3. Conclusion	6

1. Problématique

Le sujet soumis à notre expertise porte sur l'analyse de la croissance des enfants pour diagnostiquer les problèmes hormonaux chez ces derniers.

La base de données dont nous disposons recense les informations et données sur un effectif de 9184 individus distincts à savoir : le sexe, l'âge, le poids et la taille de chacun d'eux.

Afin de résoudre la problématique soulevée par le sujet, il a été question de :

- Mettre en place un outil qui permet de vérifier si la croissance de l'individu correspond à celle de son sexe pour son âge
- Et de vérifier si son Indice de Masse Corporelle (IMC) n'est pas alarmant

2. Approches et résultats

1. Importation, visualisation et prétraitement des données

Dans un premier temps, nous avons importé la base de données croissance.csv mise à disposition. Cette base regroupe un large ensemble d'observations portant sur la taille, le poids et l'âge d'une cohorte représentative d'enfants suivis jusqu'à 18 ans. Chaque individu est identifié par un identifiant unique, et un champ sexe indique s'il s'agit d'un garçon ou d'une fille.

Après importation, deux premières opérations de préparation ont été réalisées :

- Conversion des variables ind et sexe en facteurs, afin d'en faciliter l'analyse.
- Inspection des valeurs manquantes, concernant notamment la variable sexe, poids et taille.

a. Gestion des données manquantes du sexe

Quelques individus présentaient une valeur manquante dans la colonne sexe. Pour résoudre ce problème, nous avons mis en place une stratégie simple et cohérente avec la structure des données.

Si la valeur manquante apparaissait entre deux enregistrements appartenant au même individu, nous avons imputé le sexe en utilisant l'observation précédente ou suivante.

Cette méthode repose sur l'hypothèse raisonnable qu'un même individu conserve évidemment le même sexe d'une observation à l'autre. Cette approche garantit donc la fiabilité des données pour les analyses ultérieures.

b. Interpolation des valeurs manquantes de la taille et du poids

Concernant les observations manquantes des variables taille et poids, nous avons utilisé une interpolation linéaire cubique pour y remédier. Nous avons choisi cette méthode car elle

reconstruit une trajectoire de croissance continue, et est également la mieux adaptée aux phénomènes biologiques.

2. Analyse descriptive

Une analyse descriptive générale a été menée afin de mieux comprendre la structure globale des données.

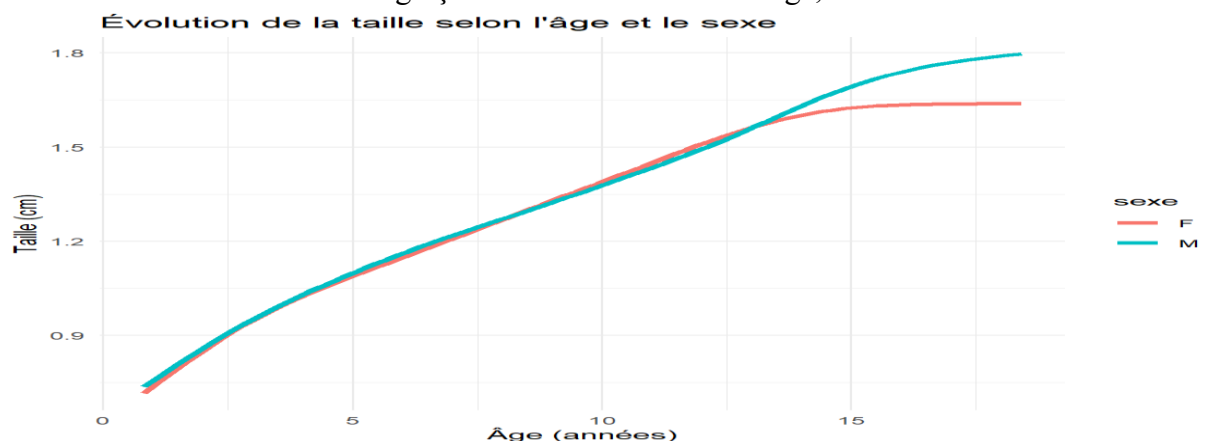
Le tableau ci-dessous nous donne un aperçu global de la base.

	sexe	n_individus	age_min	age_max	taille_moyenne	poids_moyen
1	F	4576	0.8	18.4	131.6970	33.64807
2	M	4608	0.8	18.4	134.6951	35.21579

Ces statistiques montrent que :

- Les deux groupes présentent des effectifs très proches, ce qui permet des comparaisons fiables entre filles et garçons
- L'échantillon couvre l'ensemble de la tranche d'âge allant de 0 à 18 ans.
- Les garçons présentent en moyenne une taille et un poids légèrement supérieurs, ce qui est cohérent avec les observations couramment rapportées dans les travaux scientifiques sur la croissance.

Afin de confirmer cette observation, on peut visualiser à travers ce graphe qui compare l'évolution de la taille chez les garçons et les filles suivant l'âge, une certaine différence.



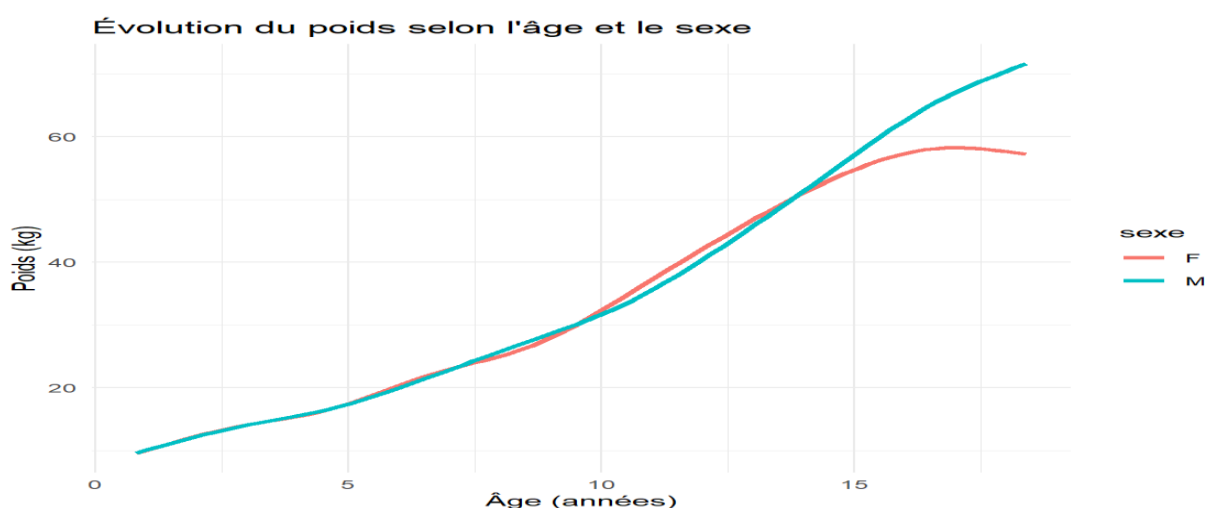
On observe qu'avant l'adolescence, les deux sexes présentent une croissance très proche, leurs courbes étant presque superposées jusqu'à environ 12 ans.

À partir de 12–13 ans, une différence apparaît :

- Chez les filles, la taille semble se stabiliser plus tôt, vers 15–16 ans.
- Chez les garçons, la courbe continue de monter plus longtemps et plus fortement.

Ainsi au-delà de 12 ans les garçons prennent plus de taille que les filles.

En ce qui concerne le poids, on observe dans la même optique que jusqu'à environ 10 ans, les deux courbes suivent pratiquement la même trajectoire, les filles et les garçons prenant du poids au même rythme.



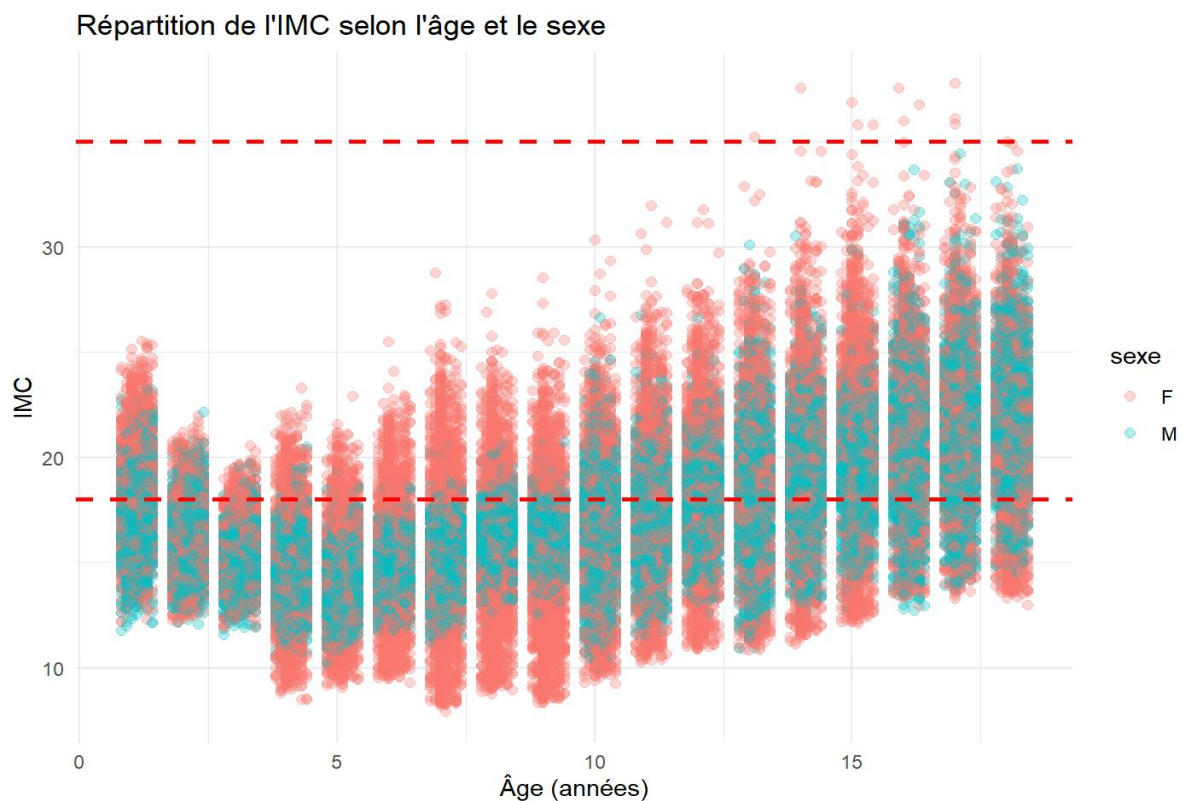
Cependant entre 10 ans et 12, les filles prennent plus de poids que les garçons avant que ces derniers ne les dépassent définitivement vers 14 ans. Les filles continuent toutefois d'augmenter en poids mais plus lentement. Cela correspond aux changements hormonaux de la puberté avec les premières menstrues pour les filles et le développement de la masse musculaire chez les garçons.

Le test de significativité effectué afin de pouvoir évaluer la différence entre les deux sexes permet d'affirmer qu'il existe une différence significative entre les garçons et les filles au fur et à mesure qu'ils grandissent.

Dans la poursuite de la résolution de notre problématique, on étudie la répartition de l'IMC selon l'âge et par sexe. L'IMC s'obtient par le rapport du poids par le carré de la taille.

A travers le graphique obtenu sur les individus de notre base, on observe une augmentation progressive de l'IMC avec l'âge. Les lignes horizontales rouges à 18 et 35 permettent d'identifier rapidement les enfants en sous poids ($IMC < 18$) ou obèses ($IMC > 35$). Les points situés au-dessus ou en dessous de ces lignes représentent des individus qui pourraient nécessiter un suivi médical.

Ainsi, il y a un grand nombre d'individus qui sont en sous poids pendant leur croissance et un faible nombre ayant un IMC supérieur à 35.



A présent que nous savons comment distinguer le sexe d'un individu à partir de sa courbe de croissance, et avons une idée de la distribution de l'IMC par sexe, il s'agira de vérifier si le sexe attribué aux individus de la base est correct et de pouvoir attribuer à un nouvel individu son sexe sans se tromper.

Pour ce fait, nous appliquons une classification supervisée et effectuons par la suite une validation croisée 5-fold afin de confirmer la méthode retenue. Avec notre jeu de données, la méthode de classification retenue est celle du KNN

3. Conclusion

La croissance d'un garçon et d'une fille est différente à partir de la puberté surtout.

En se basant sur l'âge, le poids et la taille d'un individu, on peut donc déterminer son sexe, calculer son IMC et en fonction du fait que ce résultat soit anormalement grand (>35) ou bas (<18), affirmer s'il a des problèmes hormonaux.

Selon notre méthode de détection basée sur la classification du sexe et l'analyse de l'IMC, environ 81 000 enfants pourraient présenter des anomalies nécessitant une vérification clinique supplémentaire. Parmi eux, environ 20 000 correspondent à un sexe prédit différent du sexe déclaré, tandis que les autres présentent un IMC en dehors des seuils définis. Ces résultats

doivent toutefois être interprétés avec prudence, car ils dépendent fortement des performances du modèle et des seuils utilisés.