UNIVERSITE CLERMONT AUVERGNE

IUT d'Aurillac

Sciences de données

Projet pour la SAE 5-03

Croissance des enfants

Encadrant :

Mme SOHIER Emilie

Membres du groupe :

BELLO Habibath
KONATE Batté Naïmatou
DABIRE Saânbèterfaa Joël
EL GHADHI Sidati

Année Universitaire : 2025–2026

Table des matières

Table des figures

Liste des tableaux

Introduction

L'INSEE a autorisé plusieurs hôpitaux universitaires à accéder à une base de données longitudinale concernant la croissance d'une cohorte d'enfants français. Ces informations sont utilisées par une équipe médicale inter-CHU souhaitant disposer d'un outil automatisé capable d'aider au dépistage précoce de troubles hormonaux pouvant affecter la croissance. Les médecins s'intéressent en particulier à la possibilité qu'un enfant présente une courbe de croissance atypique par rapport à son sexe biologique, ainsi qu'à l'identification des enfants dont l'indice de masse corporelle (IMC) est anormalement faible ou élevé.

Dans ce cadre, nous avons étudié la base croissance.csv, qui contient pour chaque individu des mesures répétées d'âge, de taille, de poids et de sexe. Le premier objectif du projet est d'évaluer si, en fonction de leur évolution de taille et de poids, certains enfants présentent une croissance plus proche du profil moyen du sexe opposé. Le second objectif est de repérer ceux dont l'IMC moyen est inférieur à 18 ou supérieur à 35, valeurs considérées comme potentiellement préoccupantes d'un point de vue clinique. Enfin, l'enjeu final est de proposer une liste prioritaire d'enfants « à surveiller » pour un éventuel contrôle médical complémentaire.

Pour répondre à ces problématiques, nous avons d'abord procédé à un traitement complet des données, incluant la gestion des valeurs manquantes, la détection d'anomalies et l'harmonisation temporelle des mesures par interpolation spline. Nous avons ensuite réalisé une analyse descriptive, une comparaison statistique des profils de croissance selon le sexe, puis ajusté plusieurs modèles de classification supervisée. Les résultats obtenus, ainsi que leur interprétation, sont présentés dans les sections suivantes du rapport.

1 Méthodologie

Cette étude vise à analyser les courbes de croissance des enfants âgés de 1 à 18 ans afin d'identifier d'éventuels retards de croissance, anomalies hormonales ou comportements atypiques. La méthodologie suivie respecte les étapes ci-après.

1.1 Description et traitement des données

L'étude s'appuie sur la base de données *croissance.csv*, composée de **125933 observations** correspondant à **9184 individus** suivis entre 1 et 18 ans. Chaque entrée contient : l'identifiant individuel (*ind*), le sexe, l'âge, la taille et le poids.

La première étape a consisté à identifier les valeurs manquantes. Le sexe était absent pour 19 observations, la taille pour 71 et le poids pour 54. Pour traiter ces manques :

- **Propagation du sexe :** pour chaque individu, si le sexe était renseigné dans au moins une mesure, cette information a été reportée sur l'ensemble de ses observations.
- **Vérification des valeurs aberrantes :** des seuils biologiquement plausibles ont été définis (taille : 40–220 cm, poids : 2–150 kg), et aucune valeur incohérente n'a été détectée.
- **Interpolation pour taille et poids :** afin de disposer de trajectoires complètes pour chaque individu, les mesures manquantes ont été estimées par interpolation spline cubique, appliquée individuellement sur chaque enfant, permettant de reconstituer des valeurs pour chaque âge enregistré.

Cette préparation assure un jeu de données harmonisé, cohérent et exploitable pour l'analyse

statistique.

1.2 Analyse descriptive

L'analyse descriptive vise à caractériser la croissance en fonction du sexe et de l'âge :

- **Structure de la base :** examen du nombre d'individus, du nombre de mesures par individu et de la couverture longitudinale pour confirmer la qualité des trajectoires interpolées.
- **Répartition par sexe :** calcul de la proportion de filles et de garçons afin de contextualiser les différences morphologiques observées.
- **Statistiques descriptives des variables continues :** pour chaque sexe, la moyenne et l'écart-type de la taille et du poids ont été estimés, offrant une première appréciation de la tendance centrale et de la variabilité.
- **Évolution de la croissance avec l'âge :**
  - Courbes lissées de la taille et du poids selon l'âge pour chaque sexe, afin d'observer les tendances et les différences morphologiques.
  - Calcul et représentation de l'IMC (Indice de Masse Corporelle) pour chaque âge et sexe, permettant de compléter la caractérisation de la croissance.

Cette analyse descriptive fournit une vue d'ensemble des trajectoires de croissance, identifie les tendances générales et prépare les analyses comparatives et prédictives.

1.3 Approches statistiques et supervisées

Pour approfondir l'étude des différences entre sexes et prédire le sexe à partir des mesures de croissance, plusieurs méthodes ont été mises en œuvre :

Tests statistiques comparatifs

Des tests t indépendants ont été réalisés à chaque âge pour la taille et le poids, afin de détecter les écarts significatifs entre filles et garçons.

Classification supervisée

Plusieurs modèles de classification ont été utilisés pour prédire le sexe à partir des mesures de croissance (taille, poids, âge) :

- Arbre de décision (AD)
- Analyse discriminante linéaire (LDA)
- Analyse discriminante quadratique (QDA)
- k-plus proches voisins (KNN)
- Classifieur de Bayes naïf
- Régression logistique

Pour le KNN, la prédiction est définie par :

$$\hat{y} = mode(y_{(1)}, \ldots, y_{(k)}),$$

où $y_{(i)}$ sont les étiquettes des k observations les plus proches.

Les performances de chaque modèle ont été estimées via un découpage en ensembles d'entraînement et de test, complété par une validation croisée à 10 volets :

$$Accuracy_{CV} = \frac{1}{10}\sum_{i=1}^{10} Accuracy_i.$$

Détection des trajectoires atypiques

Deux analyses complémentaires ont permis d'identifier des profils de croissance atypiques.

1) Analyse de l'IMC. L'indice de masse corporelle a été calculé pour chaque mesure :

$$IMC = \frac{poids}{(taille/100)^2}.$$

L'IMC moyen par individu a ensuite été utilisé pour repérer les valeurs extrêmes (outliers).

2) Discordance prédictive (KNN). Pour chaque enfant, les prédictions répétées du modèle KNN ont été comparées au sexe réel. Les enfants dont les prédictions majoritaires étaient inverses ont été considérés comme atypiques :

$$\frac{1}{N}\sum_{i=1}^{N} 1(\hat{y}_i \neq y) > 0.5.$$

Ces approches permettent de quantifier les différences de croissance entre sexes, de prédire le sexe à partir des mesures et de détecter les profils atypiques pour un suivi spécifique.

2 Analyse des résultats

2.1 Statistiques drescriptives

Structure de la base de données

La base finale, après nettoyage et interpolation, contient :

— **125 933 observations** ;
— un total de **9 184 individus** différents ;
— en moyenne **13,7 mesures par individu**, avec une médiane de **14 mesures**.

Ces chiffres montrent que la base est relativement dense, chaque enfant étant observé plusieurs fois au cours de sa croissance, ce qui permet des analyses temporelles robustes.

Répartition par sexe

La distribution des individus selon le sexe est équilibrée :

— **62 692 observations féminines**, soit **49,78 %** ;
— **63 241 observations masculines**, soit **50,22 %**.

Cette répartition quasi symétrique garantit que les comparaisons entre filles et garçons ne sont pas biaisées par un déséquilibre de taille d'échantillon.

Statistiques descriptives globales des variables de croissance

TABLE 1 – Statistiques descriptives globales des variables de croissance

Statistique	Âge	Taille (cm)	Poids (kg)	IMC
Min	0.80	57.29	1.58	2.66
Q1	5.00	109.60	17.30	15.16
Moyenne	9.57	133.20	34.44	17.58
Médiane	9.40	135.30	29.60	17.08
Q3	14.00	159.90	49.00	19.76
Max	18.40	218.30	125.50	100.98
Écart-type	5.19	30.20	19.69	3.45

Le tableau met en évidence une cohérence globale des mesures de croissance. L'âge s'étend de 0,8 à 18,4 ans, ce qui confirme que la base couvre l'ensemble du cycle de croissance infantile et adolescente. La taille et le poids présentent une progression structurée, avec des médianes respectivement de 135,3 cm et 29,6 kg, valeurs compatibles avec les normes pédiatriques. Les écarts-types modérés indiquent une variabilité attendue entre enfants, sans dispersion excessive.

L'IMC présente une distribution centrée autour de 17, ce qui correspond majoritairement à des corpulences normales pour cette tranche d'âge, malgré quelques valeurs extrêmes (IMC maximal supérieur à 100), probablement liées à des cas atypiques ou à la forte croissance relative chez certains individus.

Ces statistiques confirment la qualité et la cohérence du jeu de données, et justifient la poursuite d'analyses comparatives et prédictives sur cette base.

Analyse graphique de la croissance

Taille et poids en fonction de l'âge

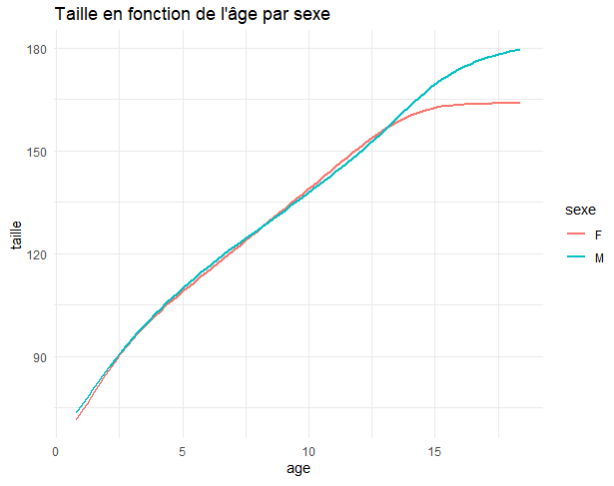

FIGURE 1 – Évolution de la taille en fonction de l'âge selon le sexe

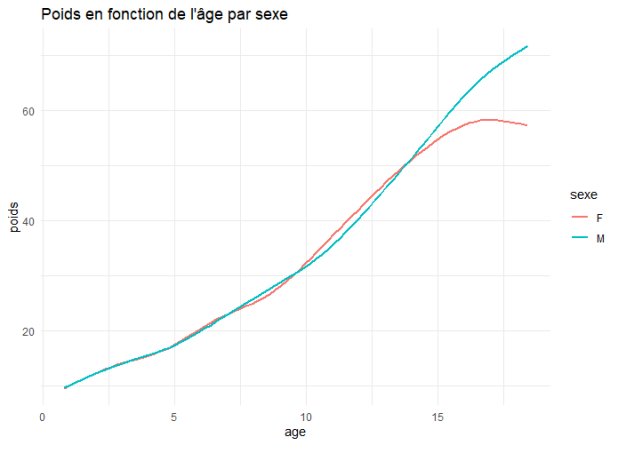

FIGURE 2 – Évolution du poids en fonction de l'âge selon le sexe

Les deux graphiques côte à côte montrent clairement l'évolution des paramètres anthropométriques :

— La croissance est régulière pour les deux sexes de 1 à 10 ans.
— Les filles grandissent plus rapidement entre 10 et 12 ans.
— Après 13 ans, les garçons deviennent progressivement plus grands et plus lourds.

Indice de Masse Corporelle (IMC) en fonction de l'âge

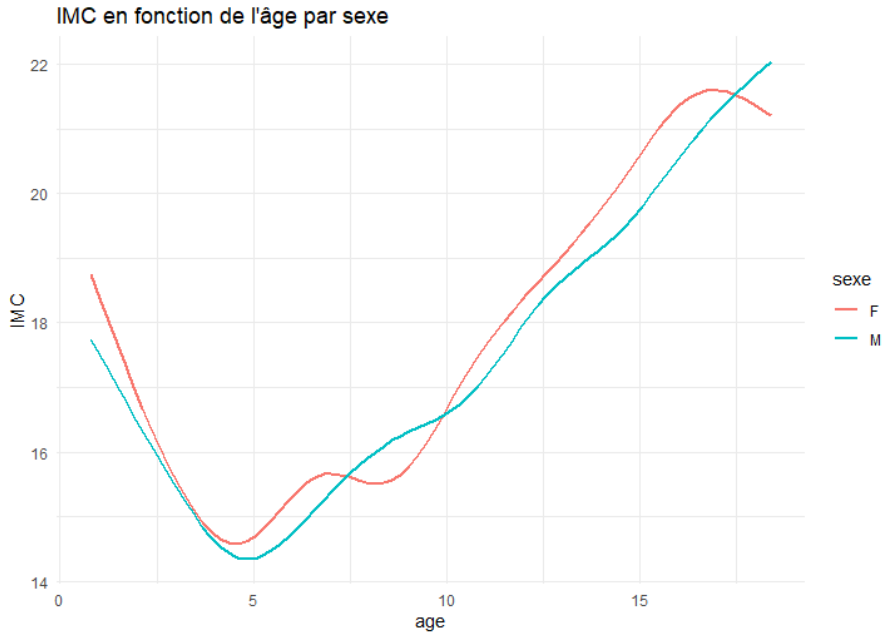

FIGURE 3 – Évolution de l'IMC en fonction de l'âge selon le sexe

L'évolution de l'IMC suit le schéma classique du développement corporel. On observe d'abord une diminution de l'IMC, qui passe d'environ 17–18 à 14–15 entre 1 et 5 ans, traduisant un allongement plus rapide que la prise de poids durant la petite enfance. À partir de 6 ans, l'IMC remonte progressivement pour atteindre environ 17–18 autour de 12 ans, puis continue d'augmenter jusqu'à environ 21–22 chez les filles et 20–21 chez les garçons à 18 ans. Les écarts entre sexes restent globalement faibles (souvent inférieurs à une unité d'IMC), ce qui confirme que cet indicateur est moins discriminant que la taille ou le poids pour différencier les morphologies selon le sexe.

2.2 Comparaison statistique des profils de croissance selon le sexe

Après avoir interpolé les trajectoires de croissance de 1 à 18 ans de manière à disposer de mesures comparables à chaque âge, nous avons entrepris une analyse statistique visant à comparer les profils de croissance entre filles et garçons. Pour cela, des tests t indépendants ont été réalisés séparément pour la taille et pour le poids, à chaque âge, afin de déterminer si les différences observées sont statistiquement significatives. Cette approche permet non seulement d'identifier les âges où les écarts entre sexes sont les plus marqués, mais aussi de mettre en évidence les périodes où les trajectoires deviennent similaires, notamment autour de la transition pubertaire. Les résultats obtenus renseignent ainsi de manière précise l'évolution différentielle de la croissance selon le sexe.

TABLE 2 – Résultats des tests t par âge entre filles et garçons

Âge	p-val Poids	Signif.	p-val Taille	Signif.
1	0.4509	Non	2.7e-10	Oui
2	0.5999	Non	6.5e-06	Oui
3	0.7974	Non	4.7e-26	Oui
4	0.5189	Non	1.8e-08	Oui
5	0.2081	Non	5.5e-57	Oui
6	0.0119	Oui	1.1e-84	Oui
7	0.1561	Non	3.1e-61	Oui
8	1.4e-11	Oui	1.8e-04	Oui
9	1.7e-10	Oui	3.3e-20	Oui
10	1.2e-04	Oui	9.4e-24	Oui
11	6.8e-38	Oui	3.9e-66	Oui
12	4.9e-30	Oui	1.0e-67	Oui
13	8.0e-08	Oui	0.96	Non
14	0.0262	Oui	2.6e-136	Oui
15	2.2e-17	Oui	0.00	Oui
16	1.9e-98	Oui	0.00	Oui
17	1.4e-191	Oui	0.00	Oui
18	1.3e-208	Oui	0.00	Oui

Les résultats montrent que :

— Pour le **poids**, les différences filles/garçons deviennent significatives à partir de 6 ans, puis extrêmement marquées dès 8 ans.
— Pour la **taille**, les différences sont significatives à presque tous les âges, sauf autour de 13 ans où les courbes se rapprochent temporairement.

Ces observations confirment que les trajectoires de croissance deviennent nettement différenciées entre filles et garçons dès le début de l'adolescence.

2.3 Classification supervisée

TABLE 3 – Performances des modèles de classification supervisée pour la prédiction du sexe

Modèle	Accuracy
Arbre de décisison (AD)	0.5684
Analyse Discriminante Linéaire (LDA)	0.5233
Analyse Discriminante Quadratique (QDA)	0.5095
k-Plus Proches Voisins (KNN, k=5)	**0.7249**
Naïve Bayes	0.5027
Régression Logistique	0.5256

Ainsi, parmi l'ensemble des modèles testés, le **KNN apparaît comme la solution la plus adaptée** pour la classification du sexe, grâce à sa capacité à exploiter la structure locale des données de croissance. Son accuracy de 72,49% dépasse largement celles des méthodes linéaires (LDA, RL, Bayes) et non linéaires (QDA), qui peinent à distinguer efficacement les deux groupes.

Validation croisée

Afin de confirmer la robustesse du modèle KNN identifié comme le meilleur classifieur dans la section précédente, une validation croisée en **10 folds** a été réalisée. Cette méthode consiste à diviser aléatoirement l'échantillon en 10 sous-ensembles de tailles équivalentes. À chaque itération, 9 folds sont utilisés pour l'entraînement du modèle et le fold restant sert de jeu de test. L'opération est répétée 10 fois afin de couvrir tous les sous-échantillons.

TABLE 4 – Validation croisée (10-fold) du modèle KNN

Modèle	**Accuracy moyenne**	**Écart-type**
KNN (k = 5)	0.72198	0.00318

2.4 Détection des enfants à surveiller

Dans cette section, nous identifions les enfants présentant des anomalies de croissance pouvant nécessiter un suivi médical. Deux indicateurs ont été utilisés :

— **IMC anormal** : un enfant est considéré à risque si son IMC moyen vérifie

$$IMC < 18 \quad ou \quad IMC > 35.$$

— **Croissance atypique (KNN)** : un enfant est classé à risque si plus de 50% des prédictions de sexe issues du modèle KNN ne correspondent pas à son sexe réel.

Les résultats obtenus sont synthétisés dans le tableau suivant :

TABLE 5 – Détection des enfants présentant des anomalies de croissance

Indicateur	**Nombre**
Enfants analysés	9184
Croissance ressemblant au sexe opposé (KNN)	237
IMC anormal (IMC $<$ 18 ou IMC $>$ 35)	5374
Présentant les deux anomalies	111

Les résultats montrent que sur les 9184 enfants de la base, **5374 présentent un IMC anormal**, indiquant un risque potentiel de sous-poids ou de surpoids. De plus, **237 enfants** possèdent une courbe de croissance atypique, ressemblant davantage à celle du sexe opposé. Enfin, **111 enfants cumulent les deux anomalies**, ce qui en fait des profils particulièrement sensibles et nécessitant une attention clinique renforcée.

3 Discussion

L'ensemble des analyses réalisées met en évidence des dynamiques de croissance conformes aux modèles biomédicaux connus, tout en révélant des profils atypiques pertinents pour le dépistage précoce. La base, particulièrement dense (125 933 mesures pour 9 184 enfants), permet une modélisation longitudinale fiable après harmonisation par interpolation spline.

Les analyses descriptives et les tests statistiques confirment des différences structurées entre sexes : la taille diverge significativement à presque tous les âges, tandis que les écarts de poids

apparaissent principalement à partir de l'adolescence, en cohérence avec les effets pubertaires différenciés. L'IMC montre une variabilité plus modérée et une capacité discriminante limitée.

Sur le plan prédictif, les méthodes linéaires et paramétriques (AD, LDA, QDA, Bayes, régression logistique) s'avèrent globalement peu performantes. À l'inverse, le modèle KNN obtient la meilleure capacité de discrimination (accuracy = 72,49 %), résultat confirmé par une validation croisée robuste. Cela souligne une structure non linéaire des relations entre les paramètres anthropométriques et le sexe.

Enfin, la détection des anomalies indique que les deux critères étudiés — IMC extrême et discordance morphologique prédite — identifient des sous-populations distinctes mais complémentaires. Au total, 111 enfants cumulent les deux signaux, constituant un groupe à haut potentiel de risque clinique nécessitant une évaluation approfondie.

Conclusion

Ce projet met en évidence l'apport d'une approche structurée combinant prétraitement, analyse descriptive, tests statistiques et classification supervisée pour analyser des données longitudinales de croissance. Les résultats montrent qu'il est possible d'extraire des tendances robustes et d'exploiter les trajectoires interpolées pour prédire le sexe avec une efficacité raisonnable. L'utilisation du KNN, particulièrement adapté à la variabilité des courbes individuelles, démontre l'intérêt de méthodes flexibles dans un contexte où les relations entre variables ne sont pas linéaires.

Au-delà de la prédiction, l'étude fournit un outil opérationnel pour repérer des enfants nécessitant une attention particulière. La combinaison de critères morphologiques (IMC extrême) et comportementaux (discordance morphologique avec le sexe) permet de dresser une liste ciblée d'individus potentiellement à risque. Cette approche ne remplace pas un diagnostic médical, mais elle offre un soutien pertinent pour prioriser les cas à examiner. Des travaux futurs pourraient intégrer des données cliniques supplémentaires afin d'améliorer la précision et l'interprétation biologique des résultats.