ST!D
Aurillac
Statistique & informatique décisionnelle Cybersécurité

Data Mining

2023

CROISSANCE DES ENFANTS

Préparer par :
Laurian Jamin
Anthony Faizandier
Motchian Jean Kevin
Hadji Iskander

INTRODUCTION

Dans le cadre hospitalier, une avancée significative a récemment été réalisée grâce à l'autorisation accordée par l'INSEE à certains hôpitaux universitaires d'accéder à des informations médicales cruciales sur la croissance d'une cohorte représentative d'enfants français. Au cœur de cette initiative se trouve une équipe de médecins inter-CHU motivée par la volonté d'améliorer les diagnostics des problèmes hormonaux influençant directement la croissance des enfants.

L'objectif principal de cette équipe médicale est de mettre en place un outil automatisé novateur capable d'analyser les courbes de croissance des enfants. Ce dispositif se révélerait essentiel en fournissant des indications spécifiques, telles que la similitude de la croissance entre les sexes et la détection de variations importantes de l'indice de masse corporelle (IMC). Plus précisément, l'outil cherchera à déterminer si la croissance d'une fille ressemble davantage à celle d'un garçon, et vice versa, tout en évaluant si l'enfant présente un score d'IMC supérieur à 35 ou inférieur à 18.

Cette entreprise médicale se situe à l'intersection de la technologie et de la médecine, illustrant la façon dont les données statistiques peuvent être exploitées pour fournir des diagnostics plus rapides et précis des problèmes hormonaux ayant un impact direct sur la croissance des enfants.

Pour répondre à ces problématiques la structure de notre rapport se déroulera en trois étapes. Tout d'abord, il s'agira de présenter la donnée sous forme d'analyse descriptive pour ainsi faire un constat de comment est constitué la donnée. Par la suite, nous allons réaliser le le prétraitement de ces données (nettoyage, arrondir les âges, les rassembler en un seul individu unique, etc…) pour mieux les analyser.
Ensuite, à partir des données nous allons vérifier s'il y a une différence significative entre les filles et les garçons (pour la taille et pour le poids) à l'aide de différents modèles de classification supervisée, et que notre modèle sera tester avec une validation croisée pour vérifier de sa pertinence.
Pour finir afin de pousser davantage sur notre analyse nous allons déterminer grâce à une base de données existante quels sont les enfants pour lesquels il faut vérifier s'ils n'ont pas de problèmes hormonaux.

Analyse descriptives

Dans cette section cruciale de notre étude, nous plongeons dans une analyse descriptive approfondie de notre base de données médicales. Cette première étape est essentielle pour mettre en lumière les informations capitales, offrant ainsi un éclairage significatif sur la problématique sous-jacente à notre recherche. Ces résultats préliminaires jetteront les bases de nos investigations futures, guidant notre compréhension des tendances de croissance chez les d'enfants âgé de

sexe	age_moyen	taille_moyenne	Poids_moyen	Effectif_total
F	9.556484	131.6931	33.6463	4576
M	9.582153	134.7002	35.2178	4608

tableau illustrant les indicateurs statistique globaux des données

Pour débuter, examinons quelques indicateurs statistiques globaux. Nous disposons d'un échantillon robuste de 125 933 enregistrements sur 9184 individus, illustrant la diversité de notre population d'enfants. Ces informations (poids et tailles) ont été pris sur des enfants de à partie de 0 an jusqu'à 18 ans pour d'autre.
En se penchant sur la répartition entre les sexes, la cohorte compte 49.78% garçons et 50.22% de filles. Ces chiffres entraînent une distribution relativement équilibrée garantissant une représentation significative des deux sexes dans notre étude. Cette bonne répartition des deux modalités de la variable sexe est intéressante permettant ainsi à notre algorithme de classification d'apprendre mieux sur qu'est ce qui peut différencier la croissance entre homme et femme.

Lorsque nous analysons les moyennes générales, nous pouvons observer que l'âge moyen est d'environ 9,57 ans, la taille moyenne est de 133,20 cm, et le poids moyen est de 34,44 kg. Ces chiffres fournissent une base descriptive essentielle pour évaluer la croissance globale de cette population d'enfants entre 0 et 18 années.

En examinant plus en détail les moyennes séparées par sexe, nous constatons des variations subtiles. Les garçons présentent une légère augmentation de l'âge moyen (9,58 ans), de la taille moyenne (134,70 cm), et du poids moyen (35,22 kg) par rapport aux filles (âge moyen : 9,56 ans, taille moyenne : 131,69 cm, poids moyen : 33,65 kg).

Ces différences dans les moyennes entre les sexes soulignent l'importance de tenir compte des caractéristiques spécifiques à chaque sexe lors de l'analyse de la croissance des enfants. Une compréhension approfondie de ces différences contribuera à développer un outil automatisé efficace pour l'évaluation si la croissance d'un enfant ressemble davantage à celle d'un garçon ou d'une fille, et si l'enfant présente un score d'IMC significatif (supérieur à 35 ou inférieur à 18).

Analyse descriptives

Le graphique présente l'histogramme de la distribution du poids par enregistrement, Chaque barre représente une plage de poids, et la hauteur de la barre indique le nombre d'enregistrement dont le poids se situe dans cette plage spécifique. L'axe des x représente les plages de poids, et l'axe des y représente la fréquence.

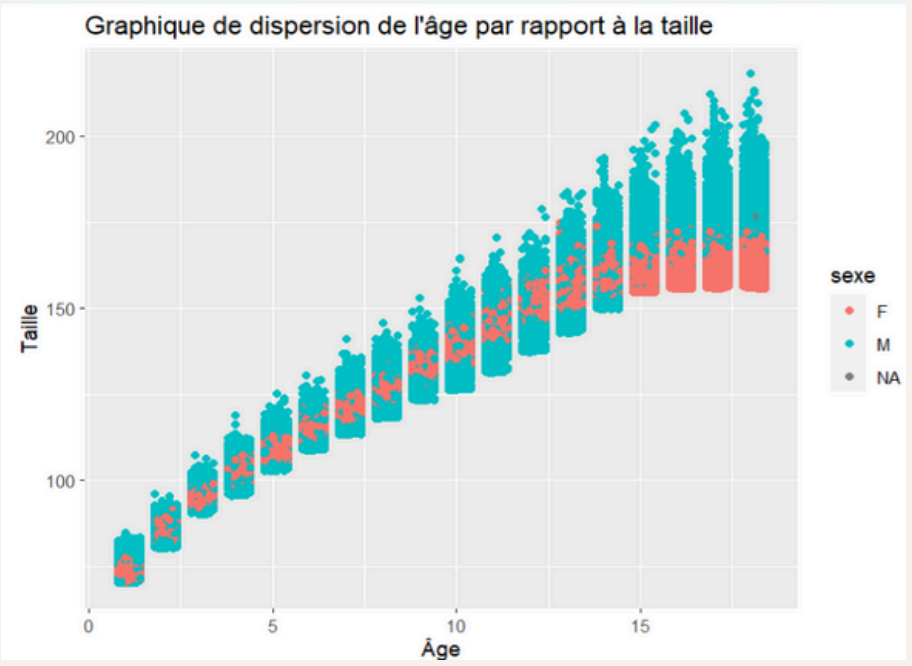
La majorité des enfants semble se situer dans la plage de poids entre 10 et 30 kg. En particulier, la plus grande fréquence se trouve dans la plage de 10 à 20 kg, ce qui suggère que de nombreux enfants ont un poids dans cette fourchette.
Le nombre d'enregistrement diminue progressivement à mesure que le poids augmente, indiquant que moins d'enfants ont un poids plus élevé. Les barres dans les plages de poids supérieures, telles que (90,100] kg et (100,110] kg, montrent que très peu d'enfants de cette cohorte ont un poids dans ces gammes élevées.

Intervalles de Poids	Fréquence
(0,10]	3937
(10,20]	35561
(20,30]	24201
(30,40]	16896
(40,50]	15358
(50,60]	13855
(60,70]	9340
(70,80]	4440
(80,90]	1621
(90,100]	443
(100,110]	114
(110,120]	19
(120,130]	4

tableau montrant ça fréquence pour chaque intervallle de poids

En observant les données plus spécifiques, nous constatons que la fréquence diminue à mesure que le poids augmente. Les plages de poids entre 50 et 60 kg conservent une fréquence relativement stable, mais au-delà de 60 kg, la fréquence diminue considérablement.

Le graphique représente dispersion de l'âge par rapport à la taille pour les garçons, chaque point représente un enregistrement, positionné en fonction de son âge sur l'axe des x et de sa taille sur l'axe des y.
En examinant les tableaux de répartition pour la taille et l'âge des garçons, on remarque que pour chaque tranche d'âge, le nuage de points représentatifs des filles est initialement centré parmi les points des garçons, mais à partir d'environ 15 ans , la taille des filles commence à être généralement inférieure à celle des garçons.



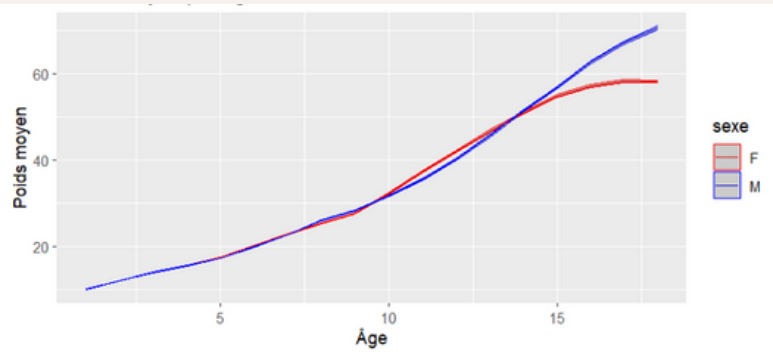
Dans la section pré-traitement, nous détaillons les variations enregistrées dans le nuage de points.

Classification

Prétraitement

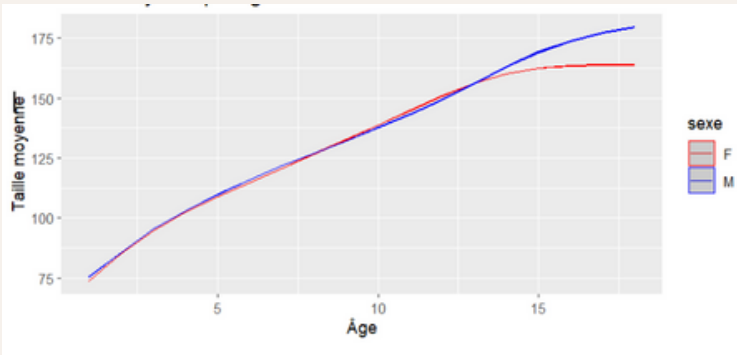
Ce processus de prétraitement a pour objectif de récupérer les valeurs manquantes, telles que le sexe, qui ont pu être omises pour certains individus au fil du temps. Il est important de souligner que le processus d'enregistrement s'est étendu sur 18 ans. En outre, nous avons observé une variation du nombre d'enregistrements par individu, allant de 7 à 17, avec la majorité des individus ayant 15 enregistrements. Nous avons ajouté la variables IMC = Poids/(Taille*Taille) pour voir si l'enfant a un score d'IMC qui est supérieur à 35 ou inférieur à 18.

Afin d'obtenir une courbe de croissance de l'enfance interprétable, il est nécessaire d'aligner les données de chaque individu sur une échelle temporelle commune, de 1 an à 18 ans. Pour atteindre cet objectif, nous utiliserons la méthode de spline cubique. Cette approche offre également la possibilité d'imputer les données manquantes relatives au poids et à la taille pour l'ensemble des individus.



Graphique représentant la variation moyenne du poids en relation avec l'âge.

L'évolution de la taille moyenne en fonction de l'âge, différenciée par sexe, révèle que les femmes et les hommes présentent une croissance similaire jusqu'à l'âge d'environ 10 ans. Après cette période, les hommes continuent leur croissance, tandis que les femmes prennent légèrement l'avantage jusqu'à l'âge de 15 ans, le moment où les deux courbes se croisent à nouveau, marquant le début d'une croissance plus prononcée chez les hommes.



Graphique représentant la variation moyenne de la taille en relation avec l'âge.

De manière similaire, avant l'âge de 15 ans, il n'y a pas de distinction significative dans l'évolution de la taille entre les enfants, que ce soit chez les filles ou les garçons. Cependant, après cet âge, une distinction réelle émerge, avec une courbe de croissance chez les garçons devenant encore plus prononcée.

Les deux graphiques précédents illustrent une corrélation significative entre nos variables distinctes, à savoir la taille, le poids et l'âge. Afin de mieux répondre à notre problématique, il est crucial d'entreprendre une première étape visant à atténuer cette structure de dépendance entre ces variables. Pour cela, nous avons opté pour l'Analyse en Composantes Principales (ACP). Cependant, malgré l'utilisation de la validation croisée par la méthode de k-Fold, les performances obtenues ne sont pas satisfaisantes. Le taux d'exactitude entre les données d'entraînement et de test pour toutes les méthodes employées (arbre de décision, plus proche voisin, analyse quadratique, analyse discriminante linéaire et le bayésien naïf) se situe autour de 54%.

Une deuxième approche consiste à maintenir nos données initiales et à appliquer les divers classifieurs à ces données. Les résultats des deux approches indiquent que la deuxième méthode est préférable, car le taux d'exactitude dépasse largement les 90%.

Validation Croisée

Cette phase implique une analyse répétée des ensembles de données afin d'assurer une évaluation robuste du modèle. Cette démarche, bien que chronophage en raison du volume important de données traitées, est tout à fait normale et conforme aux attentes. Les résultats obtenus sont récapitulés dans le tableau ci-dessous pour nombre d'itération k=10.

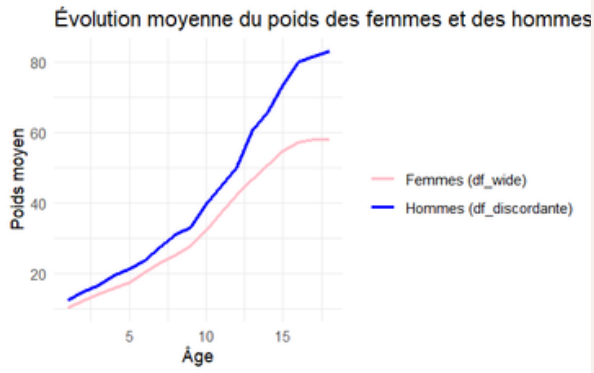
Classifieurs	Taux Exactitude (%)	durée exécution (min)
Plus proche voisin	99.78	1.25
Arbre de décision	97.68	0.63
Régression logistique	99.66	0.55
Analyse discriminante linéaire	99.82	0.1137
Analyse discriminante quadratique	99.78	0.11
Bayésien naïf	95.62	3

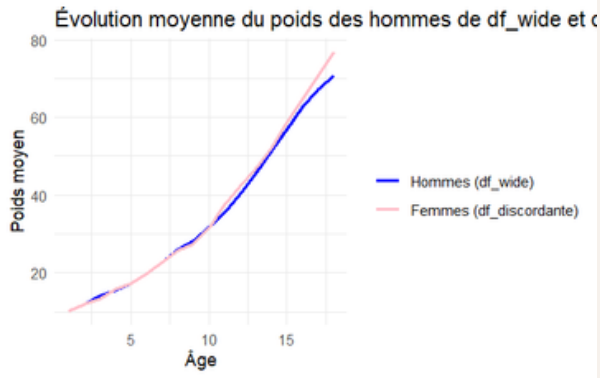
Les performances d'un même classifieur peuvent montrer des variations d'une exécution à l'autre. Cependant, il est important de noter que ces fluctuations restent généralement proches des valeurs moyennes ou typiques, ce qui peut être attribué à des facteurs tels que la variabilité inhérente des données, la sensibilité aux paramètres du modèle et la complexité du modèle. En d'autres termes, bien que des fluctuations puissent être observées, elles ne s'éloignent généralement pas significativement des tendances centrales attendues.

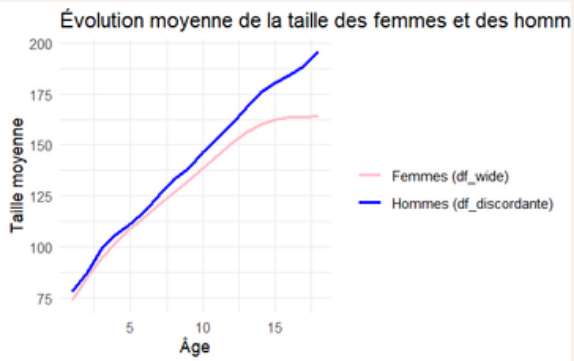
En conclusion sur le choix du meilleur modèle, bien que l'Analyse Discriminante Linéaire ait présenté les performances les plus élevées en termes de résultats, la régression logistique est sélectionnée comme le modèle le plus adapté à notre situation. Cette décision repose sur deux raisons principales : Nos données, étant sur une échelle temporelle, bénéficient davantage de l'utilisation de la régression logistique. En effet, celle-ci offre une meilleure interprétation des relations temporelles présentes dans les données, aspect crucial pour notre analyse.

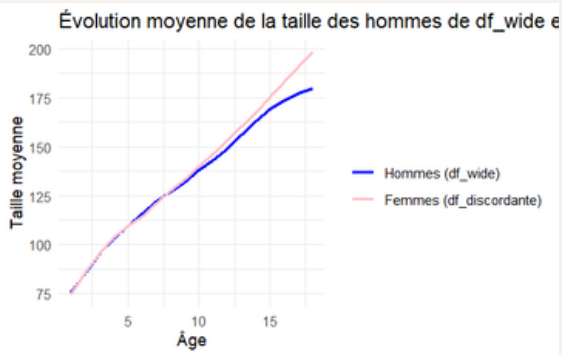
Prédictions incorrectes

Pour les mauvaises prédictions, c'est-à-dire les hommes qui ont été prédits comme étant des femmes et les femmes prédites comme étant des hommes, nous pouvons, dans un premier temps, vous les illustrer à l'aide de quatre graphiques :









Ces quatre graphiques présentent des cas où la prédiction de genre (homme ou femme) a été inversée par rapport au sexe biologique réel. On observe un homme mal classifié et trois femmes dans la même situation. Pour l'homme en question, le graphique concernant le poids montre une nette divergence dans la tendance d'évolution par rapport à la moyenne des femmes. En revanche, pour la taille, la différence est beaucoup moins marquée. Bien que l'évolution semble similaire entre cet homme et les femmes, cela pourrait indiquer une erreur dans mon modèle de prédiction. En ce qui concerne les femmes mal classifiées, la différence est également moins évidente. Leurs courbes de croissance se superposent presque avec la moyenne des hommes, suggérant que la croissance de ces femmes pourrait être similaire à celle d'un homme.

Conclusion

En conclusion notre analyse approfondie des données médicales sur la croissance des enfants a révélé des tendances significatives. Avec un échantillon solide de 125 933 enregistrements provenant de 9 184 individus, la répartition équilibrée entre les sexes a renforcé la capacité de notre algorithme à apprendre les différences de croissance entre garçons et filles. Les indicateurs statistiques globaux ont montré des moyennes subtiles, soulignant l'importance de considérer les caractéristiques spécifiques à chaque sexe.

L'analyse de la répartition du poids a mis en évidence que la majorité des enfants se situent dans la plage de 10 à 30 kg. Le processus de prétraitement des données, y compris l'ajout de la variable IMC, a été essentiel pour obtenir une courbe de croissance interprétable.

En matière de modélisation, malgré des performances initiales insatisfaisantes de l'ACP, la deuxième approche a montré des taux d'exactitude dépassant largement les 90% pour divers classificateurs. Les variations enregistrées dans les performances sont attribuables à des facteurs tels que la variabilité inhérente des données et la sensibilité aux paramètres du modèle.

Une véritable disparité de croissance, que ce soit en termes de taille ou de poids, se manifeste généralement vers l'âge de 15 ans, marquant une distinction significative entre hommes et femmes. Il est donc pertinent d'explorer cette période, car si une femme est prédite comme un homme, cela pourrait suggérer qu'elle a tendance, après l'âge de 15 ans, à adopter un profil similaire à celui des hommes, et vice versa.

Pour résoudre les différentes problématiques, plusieurs options se sont présentées, et toutes nos conclusions peuvent être impactées par le choix du classifieur, la méthode de validation de notre algorithme, le choix de certains paramètre pour minimiser le temps d'exécution de notre algorithme, ainsi que les interprétations qui en découlent. Cependant, l'algorithme permet d'identifier les hommes et femmes présentant une croissance anormale par rapport à leur catégorie respective, afin d'isoler leur profil pour une meilleure compréhension de leur évolution et de détecter d'éventuelles anomalies de croissance.