Nicolas  
Dumont

**Projet Statistiques**

Contents

[1 - Analyse Descriptive 1](#_Toc94106795)

[2 – Feature Engineering 1](#_Toc94106796)

[3 – Test de Student 1](#_Toc94106797)

[4 – Test de Chi-Square 2](#_Toc94106798)

[5 – Test de Normalité 2](#_Toc94106799)

[6 – Regression Linéaire 2](#_Toc94106800)

[1er modèle : 2](#_Toc94106801)

[2ème modèle : 3](#_Toc94106802)

[7 – Clustering 4](#_Toc94106803)

[Conclusion : 5](#_Toc94106804)

# 1 - Analyse Descriptive

Après avoir importé le tableau de données, nous remarquons qu’il n’y a pas de valeurs manquantes et que chaque colonne contient des variables numériques. Le tableau de données contient 42 lignes pour 16 colonnes. Notre variable cible, le poids à la naissance, comprends une moyenne de 34.6, un écart type de 0.6 avec des valeurs allant de 1.92 à 4.57. Aux premiers abords, la variable contenant le poids à la naissance semble suivre une loi normale. Une analyse de la corrélation de Pearson montre que la variable cible « Birthweight » et le plus hautement (ou négativement) corrélée avec les colonnes « Smoker », « Headcirc », « Length », « Gestation » et « lowbwt ». Nous remarquons aussi que la taille de la mère est relativement hautement (~0.5) corrélée à la longueur du bébé signifiant que plus la mère est grande plus le bébé sera long. De la même façon, plus le bébé restera longtemps en période de gestation, plus sont poids à la naissance sera élevé. Le poids de la mère avant d’être enceinte cependant n’est que faiblement corrélée au poids à la naissance du bébé.

# 2 – Feature Engineering

Nous créons 5 nouvelles variables dans le tableau : IMC de la mère et de l’enfant afin de pouvoir identifier plus facilement le sous/surpoids de l’échantillon. Le ratio de circonférence abdominale de l’enfant, calculé à l’aide de la moyenne calculée en fonction du nombre de semaines de gestation. La circonférence abdominale ainsi que la taille de l’enfant par rapport à la taille de ses parents.

Dans le cas de l’IMC, nous ne pouvons utiliser cette variable car cela reviendrait à de la fuite de données, et dans le cas des autres variables créées, nous remarquons qu’elles ne sont pas aussi bien corrélées au poids à la naissance que les variables qui ont été utilisées pour les créer.

# 3 – Test de Student

Nous comparons tout d’abord le poids des enfants entre ceux dont la mère fume et les autres. Un premier test d’indépendance bilatéral nous donne une valeur p inférieure à 5% : nous rejetons avec un intervalle de confiance de 95% l’hypothèse nulle selon laquelle les bébés des mères fumeuse pèseraient le même poids que les autres. Afin d’aller un peu plus dans le détail, nous effectuons un test d’indépendance unilatéral vers la gauche qui nous donne à nouveau une valeur p inférieure à 5% et nous permet avec un intervalle de confiance de 95% de rejeter l’hypothèse nulle : les bébés des mères fumeuses pèsent donc en effet moins que les autres.

Nous avons aussi effectué un test d’indépendance bilatéral sur le poids en fonction de l’âge de la mère, et avons trouvé que les bébés des mères ayant plus de 35 ans pèsent probablement autant que les autres.

# 4 – Test de Chi-Square

Nous souhaitons vérifier si avoir un faible poids à la naissance et avoir une mère fumeuse sont deux variables indépendantes. Pour cela nous effectuons un test de chi-square. La p-value obtenue étant supérieure à 5% nous ne pouvons pas rejeter l'hypothèse nulle avec un intervalle de confiance de 95%. Avoir une mère fumeuse et avoir un faible poids à la naissance sont probablement deux variables indépendantes.

Nous effectuons aussi le même test avec le même résultat en fonction de l’âge de la mère et du poids de l’enfant à la naissance.

# 5 – Test de Normalité

Nous souhaitons vérifier notre première hypothèse selon laquelle le poids à la naissance suit une loi normale. Pour cela nous utilisons d’abord le test de normalité d’Agostino, puis celui de Shapiro-Wilks avant d’utiliser celui de Kolmogorov-Smirnov. Ces trois tests nous donnent une valeur p supérieure à 5%, nous ne pouvons donc pas rejeter l’hypothèse nulle avec un intervalle de confiance de 95%. Le poids à la naissance suit donc probablement une loi normale.

Nous effectuons aussi un test de normalité d’Agostino sur le nombre de cigarettes fumées par la mère par jour et trouvons que cette variable ne suit pas une distribution normale.

# 6 – Regression Linéaire

## 1er modèle :

Nous utilisons uniquement les variables smoker et lowbwt : deux variables indépendantes et corrélées à Birthweight.

Graphical user interface

Description automatically generated

On remarque que les enfants catégorisés comme ayant un poids de naissance faible ont environ un poids à la naissance inférieur de 1.04 kg aux autres. On remarque aussi que les enfants de femmes fumeuses ont un poids à la naissance inférieur d’environ 0.2 kg.

Graphical user interface, text

Description automatically generated

Ne possédant qu’un échantillon de la population globale, notre coefficient ne peut pas s’appliquer précisément à autre chose que notre échantillon. Cet intervalle de confiance nous montre les valeurs que le véritable coefficient pourrait prendre (avec 95% de sûreté). On remarque notamment que selon cet intervalle, avoir une mère fumeuse pourrait potentiellement augmenter le poids à la naissance de son enfant. Cela semble invraisemblable. Le coefficient de notre échantillon se situe d’ailleurs hors de cet intervalle. Nous voyons ici que l’intervalle de confiance prédit ne contient pas forcément le vrai coefficient.

Text

Description automatically generatedLes valeurs p sont inférieures à 0.5. Nous pensons donc que les variables « smoker » (comme démontré par le test de Student) et « lowbwt » sont corrélées avec « Birthweight ».

Graphical user interface, text, application, website

Description automatically generatedEn outre le R-squared est de 0.45 et le Adjusted R-squared est de 0.421. Nous allons maintenant comparer ces valeurs à celles obtenues avec notre deuxième modèle.

## 2ème modèle :

Cette fois-ci, nous utilisons les valeurs hautement corrélées (ou négativement corrélées) : lowbwt, smoker, Gestation, Length, et Headcirc.

Nous remarquons que certaines variables (principalement lowbwt et Length) ne sont plus significatives puisque rajouter des variables a causé l’augmentation de leurs valeurs p au-dessus du seuil de 5%. Nous décidons donc de supprimer ces 2 variables qui ne sont plus significatives. Nous remarquons aussi l’augmentation de la valeur R-Squared, ce qui signifie que ce modèle est supérieur au 1er. Cela pourrait être expliqué par le nombre de variables explicatives supérieur, mais l’Adjusted R-Squared a lui aussi augmenté. Bien que l’Adjusted R-Squared ne pénalise pas suffisamment l’ajout de variables, n’ayant utilisé qu’une variable de plus que précédemment, nous décidons donc de garder ce modèle.

En utilisant la librairie scikit-learn et à l’aide d’un train\_test\_split afin de réaliser un test conclusif, nous obtenons un score R-Squared d’environ 0.7 sur les données test.

Ci-dessous un graphique des prédictions :

Chart, scatter chart

Description automatically generated

De plus, malgré le faible nombre de données menant à des scores de précision assez aléatoires, nous remarquons qu’il est possible de prédire de façon relativement satisfaisante la longueur du bébé en utilisant uniquement la taille de la mère, ainsi que le poids du bébé en fonction de son temps de gestation.

Finalement, une régression logistique nous permet de prédire précisément (80%+) si un bébé a un poids faible ou non en fonction de son temps de gestation.

# 7 – Clustering

Finalement, nous effectuons une analyse clustering en utilisant un modèle PCA transformant nos variables contenant le plus d’informations en une seule variable. Suite à cela, nous utilisons l’algorithme K-Means afin d’identifier les différents clusters dans le jeu de données. L’heuristique de la méthode du coude nous permet d’identifier qu’il y aurait 3 différents clusters. Cependant, une identification graphique nous permet d’observer que 2 gros clusters existent et qu’ils correspondent aux données en fonction de si la mère est fumeuse ou non.

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

Finalement une analyse CAH nous permet d’observer que les groupes peuvent se former en fonction de la proximité du poids à la naissance des enfants. On remarque aussi que les valeurs extrêmes des deux extrémités sont souvent plus proches l’une de l’autre que d’autres valeurs plus proches numériquement. On peut en déduire que le poids des bébés à la naissance indique généralement la valeur d’autres variables du tableau, et que les valeurs extrêmes sont issues du même type de variables.

# Conclusion :

Le poids à la naissance est principalement affecté par la mère (moins par le père) et peut être déduis le plus facilement à l’aide d’informations directement liées à l’enfant. Cependant cette approche serait uniquement explicative et non prédictive puisqu’elle nécessite que le bébé soit déjà né. Cela est moins intéressant qu’utiliser uniquement des variables issus des parents, ce qui permettrait de prédire les caractéristiques du bébé avant sa naissance – ce qui par exemple pourrait être utile afin de savoir s’il aurait besoin de soins médicaux à la naissance.