



Rapport du Projet

Major:
Data Engineering

Written By:
LAKHAL Badr

Supervised By:
Prof. EL HAJ TIRARI Mohammed

Contents

1	Introduction	4
1.1	Énoncé du problème	4
1.2	Périmètre du projet	4
1.3	Objectifs	5
1.4	Méthodologie	5
1.5	Arborescence du projet	6
2	Description des Données et Prétraitement	7
2.1	Description du Jeu de Données	7
2.2	Prétraitement des Données	8
2.3	Conclusion	8
3	Analyse Exploratoire des Données (EDA)	9
3.1	Résumé Statistique et Observations Générales	9
3.2	Visualisation des Distributions des Variables	10
3.2.1	Histogrammes des Variables	10
3.2.2	Boxplots et Détection des Outliers	11
3.3	Analyse des Relations entre les Variables	11
3.4	Conclusion	12
4	Méthodologie	13
4.1	Sélection du Modèle	13
4.2	Division des Données	13
4.3	Entraînement et Validation	14
4.4	Évaluation des Performances	14
4.5	Conclusion	16
5	Résultats	17
5.1	Évaluation du modèle	17
5.2	Interprétation des résultats	18
5.3	Performance sur les données de test	18

5.4 Conclusion	18
6 Conclusion	19

List of Figures

2.1	Suppression de la colonne 'Identifiant_patient'	8
3.1	Statistiques sur la base de données	9
3.2	Visualisation des distributions de données	10
3.3	Visualisation des distributions de données	11
3.4	Visualisation des boxplots de données	12
4.1	Code Python Division des Données	14
4.2	Code Python Entraînement du Modèle	14
4.3	Code Python Évaluation du Modèle	16
5.1	Résultats Évaluation du Modèle	17

Chapter 1

Introduction

1.1 Énoncé du problème

L'estimation du poids des patients à partir de caractéristiques physiques et physiologiques est un défi clé dans le domaine de la santé. Un poids anormalement élevé ou faible peut être un indicateur de risques potentiels pour la santé, tels que des maladies cardiovasculaires, le diabète ou l'obésité. Cependant, la prédiction du poids des patients en se basant uniquement sur des informations démographiques et physiques reste un problème complexe. Ce projet s'attaque à cette problématique en construisant un modèle de régression qui prédit le poids des patients à partir de caractéristiques observables telles que l'âge, la taille, la pression artérielle et d'autres mesures corporelles.

1.2 Périmètre du projet

Le périmètre de ce projet se limite à l'analyse des données provenant de 80 patients d'un hôpital. Ces données contiennent des informations variées sur les patients, notamment leur sexe, âge, taille, poids, ainsi que d'autres mesures physiques telles que la circonférence du tour de taille, la pression sanguine, et l'indice de masse corporelle (IMC). Le but est de modéliser le poids des patients à partir de ces attributs en utilisant des techniques statistiques et de régression.

Ce projet ne couvre pas d'autres facteurs médicaux ou biologiques qui pourraient influencer le poids, tels que des conditions médicales sous-jacentes. Il se concentre uniquement sur l'analyse des données disponibles et sur la création d'un modèle prédictif basé sur ces données.

1.3 Objectifs

À la fin de ce projet, il est attendu :

- L'analyse des relations entre les différentes caractéristiques et le poids des patients pour identifier les variables les plus influentes dans la prédiction du poids.
- La construction d'un modèle de régression capable de prédire avec précision le poids des patients à partir de leurs caractéristiques physiques et physiologiques.
- L'évaluation de la performance du modèle à travers des métriques appropriées comme l'erreur quadratique moyenne (RMSE) ou le coefficient de détermination (R^2).
- Une discussion sur l'applicabilité du modèle dans un contexte clinique réel.

1.4 Méthodologie

L'approche suivie dans ce projet consiste en plusieurs étapes clés :

- *Analyse exploratoire des données (EDA)* : Cette étape consiste à examiner les données pour en comprendre la distribution, identifier les valeurs manquantes, détecter les valeurs aberrantes et effectuer des transformations si nécessaire. Les visualisations comme les boxplots, les histogrammes et les matrices de corrélation sont utilisées pour comprendre la relation entre les différentes variables.
- *Prétraitement des données* : Après l'analyse exploratoire, les données sont nettoyées, et les valeurs aberrantes sont supprimées à l'aide de méthodes comme l'intervalle interquartile (IQR) ou bien Z-score.
- *Modélisation de la régression* : Plusieurs modèles de régression, y compris la régression linéaire et éventuellement des modèles plus complexes, seront appliqués aux données pour prédire le poids des patients.
- *Évaluation du modèle* : Les modèles seront évalués à l'aide de métriques comme l'erreur quadratique moyenne (RMSE) et le coefficient de détermination (R^2) pour déterminer leur performance.
- *Analyse des résultats* : Les résultats obtenus seront analysés pour évaluer la pertinence du modèle et discuter de son applicabilité dans un contexte médical réel.

1.5 Arborescence du projet

L'arborescence suivante illustre la structure des fichiers de ce projet :

```
projet_regression/  
  data/  
    raw/  
      Patients_hopital.csv          # data brutes  
    processed/  
      processed_data.csv  # data apres traitement  
  notebooks/  
    EDA.ipynb  # Notebook pour EDA  
    model.ipynb  # Notebook pour l'entrainement et l'  
                  evaluation  
  reports/  
    project_report.pdf  # Rapport final du projet
```

Chapter 2

Description des Données et Prétraitement

2.1 Description du Jeu de Données

Le jeu de données utilisé pour ce projet, relatif à l'examen d'un modèle de régression, provient d'un échantillon de 80 patients d'un hôpital. Ces données mesurent diverses caractéristiques liées à l'état de santé des patients. Le fichier de données utilisé est intitulé "Patients_hopital.csv", et il contient les variables suivantes :

- GENRE : Le genre du patient (0 = Homme, 1 = Femme).
- AGE : L'âge du patient (en années).
- TAILLE : La taille du patient (en cm).
- POIDS : Le poids du patient (en Kg).
- TTAILLE : Le tour de taille du patient (en cm).
- SYS : La pression sanguine systolique du patient (en mmHg).
- DIA : La pression sanguine diastolique du patient (en mmHg).
- IMC : L'indice de masse corporelle (IMC) du patient.
- JMBG : La longueur de la jambe gauche du patient (en cm).
- COUD : La largeur du coude du patient (en cm).
- POIGN : La largeur du poignet du patient (en cm).

- BRAS : La circonférence du bras du patient (en cm).

Le jeu de données contient des informations complètes sur les caractéristiques physiques et biométriques des patients, et est principalement utilisé pour explorer des relations à l'aide de modèles de régression.

2.2 Prétraitement des Données

Avant d'entamer l'analyse descriptive et la modélisation, il est important de mentionner que le jeu de données était déjà propre, sans valeurs manquantes ni doublons. Le seul ajustement effectué sur le jeu de données a été la suppression de la colonne "Identifiant_patient", qui était une donnée d'identification et ne contribuait pas à l'analyse des caractéristiques des patients.

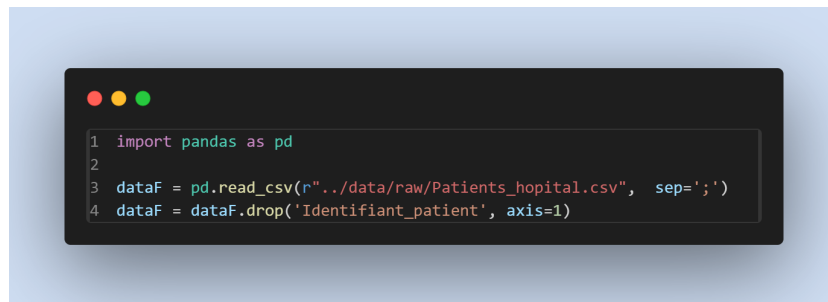


Figure 2.1: Suppression de la colonne 'Identifiant_patient'

2.3 Conclusion

En résumé, le prétraitement effectué sur le jeu de données a été limité aux actions suivantes :

- Suppression de la colonne "Identifiant_patient", sans impact sur l'intégrité des autres données.
- Aucune autre modification n'a été nécessaire, le jeu de données étant déjà nettoyé et prêt pour l'analyse.

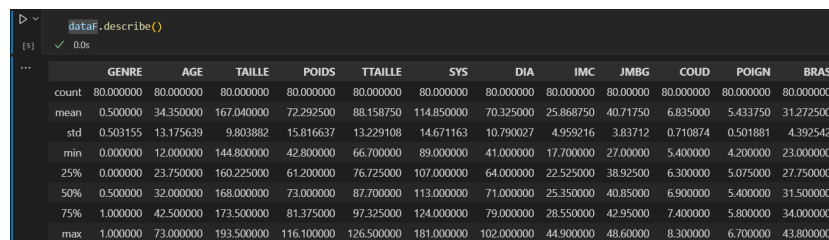
Chapter 3

Analyse Exploratoire des Données (EDA)

L'analyse exploratoire des données (EDA) est une étape cruciale pour mieux comprendre les relations entre les différentes variables, détecter des anomalies ou des tendances intéressantes, et orienter les choix méthodologiques lors de la modélisation. Dans cette section, nous allons examiner les caractéristiques principales du jeu de données et visualiser certaines distributions ainsi que les relations entre les variables.

3.1 Résumé Statistique et Observations Générales

À partir des statistiques descriptives des données, il a été constaté que le jeu de données est relativement bien équilibré, notamment en ce qui concerne la variable 'GENRE', où l'on observe une répartition exacte de 50% pour les hommes et 50% pour les femmes. Cela suggère que le modèle ne sera pas biaisé par un déséquilibre des genres, ce qui est une caractéristique intéressante pour la modélisation.



	GENRE	AGE	TAILLE	POIDS	TTAILLE	SYS	DIA	IMC	JMBG	COUD	POIGN	BRAS
count	80.000000	80.000000	80.000000	80.000000	80.000000	80.000000	80.000000	80.000000	80.000000	80.000000	80.000000	80.000000
mean	0.500000	34.350000	167.040000	72.292500	88.158750	114.850000	70.325000	25.868750	40.71750	6.835000	5.433750	31.272500
std	0.503155	13.175639	9.803882	15.816637	13.229108	14.671163	10.790027	4.959216	3.83712	0.710874	0.501881	4.392542
min	0.000000	12.000000	144.800000	42.800000	66.700000	89.000000	41.000000	17.700000	27.00000	5.400000	4.200000	23.000000
25%	0.000000	23.750000	160.225000	61.200000	76.725000	107.000000	64.000000	22.525000	38.92500	6.300000	5.075000	27.750000
50%	0.500000	32.000000	168.000000	73.000000	87.700000	113.000000	71.000000	25.350000	40.85000	6.900000	5.400000	31.500000
75%	1.000000	42.500000	173.500000	81.375000	97.325000	124.000000	79.000000	28.550000	42.95000	7.400000	5.800000	34.000000
max	1.000000	73.000000	193.500000	116.100000	126.500000	181.000000	102.000000	44.900000	48.60000	8.300000	6.700000	43.800000

Figure 3.1: Statistiques sur la base de données

3.2 Visualisation des Distributions des Variables

Pour mieux comprendre la répartition des variables, des visualisations graphiques ont été réalisées. Ces visualisations ont permis d'examiner la forme des distributions des données.

3.2.1 Histogrammes des Variables

Les histogrammes ont permis d'observer la forme des distributions des variables continues. Après avoir tracé ces histogrammes, il a été noté que la majorité des variables suivent une distribution proche de la distribution normale. Cela indique que, dans de nombreux cas, les données sont symétriques et réparties autour de la moyenne, ce qui est favorable à l'application de modèles statistiques classiques tels que la régression linéaire.

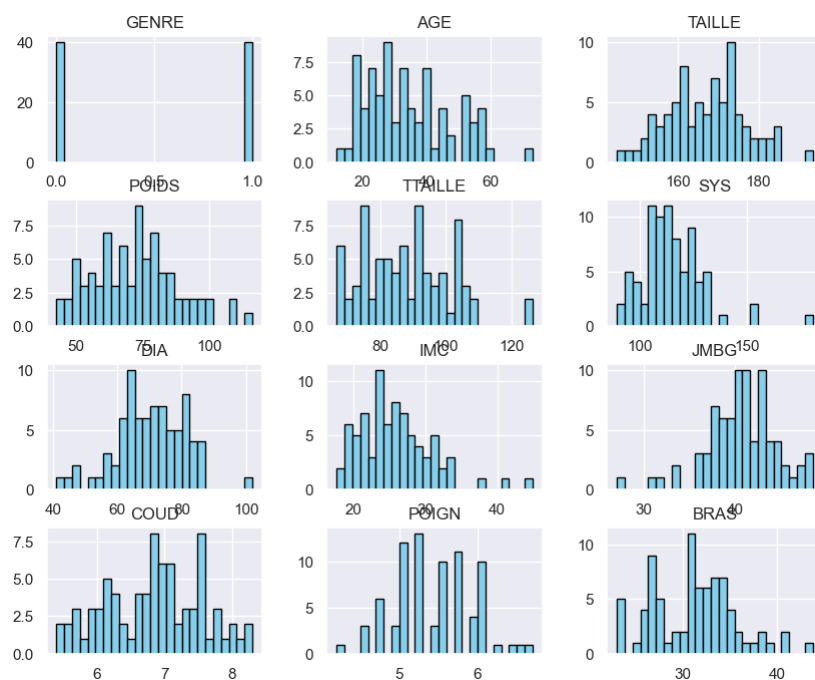


Figure 3.2: Visualisation des distributions de données

3.2.2 Boxplots et Détection des Outliers

Les boxplots ont été utilisés pour identifier les valeurs aberrantes dans les variables continues. Il a été observé que plusieurs variables présentent des valeurs extrêmes en dehors des moustaches des boxplots. Par exemple, la variable 'POIDS' montre des valeurs qui s'écartent significativement de la tendance centrale, ce qui a conduit à leur suppression à l'aide de la méthode de l'intervalle interquartile (IQR).

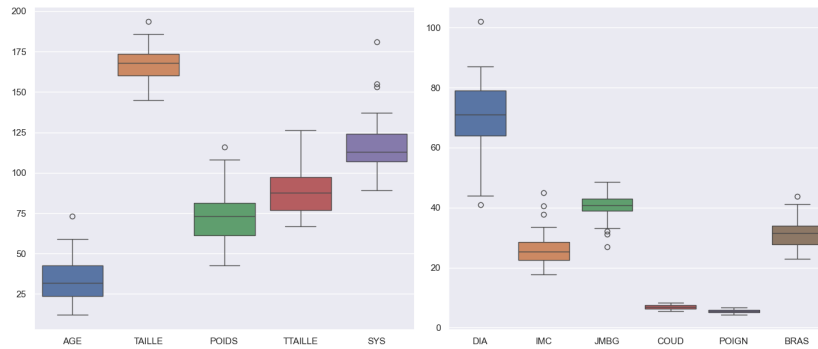


Figure 3.3: Visualisation des distributions de données

3.3 Analyse des Relations entre les Variables

Afin d'examiner les relations potentielles entre les différentes variables, une matrice de corrélation a été générée. Cette analyse a montré que certaines variables sont fortement corrélées entre elles. Un modèle de régression linéaire peut être approprié pour ce jeu de données, d'après la matrice de corrélation.

- Nous observons que les variables 'TTAILLE', 'IMC', 'COUD' et 'BRAS' présentent une forte corrélation linéaire avec la variable cible 'POIDS'.
- D'autres variables peuvent également être utilisées, à condition d'éviter le surapprentissage (overfitting).

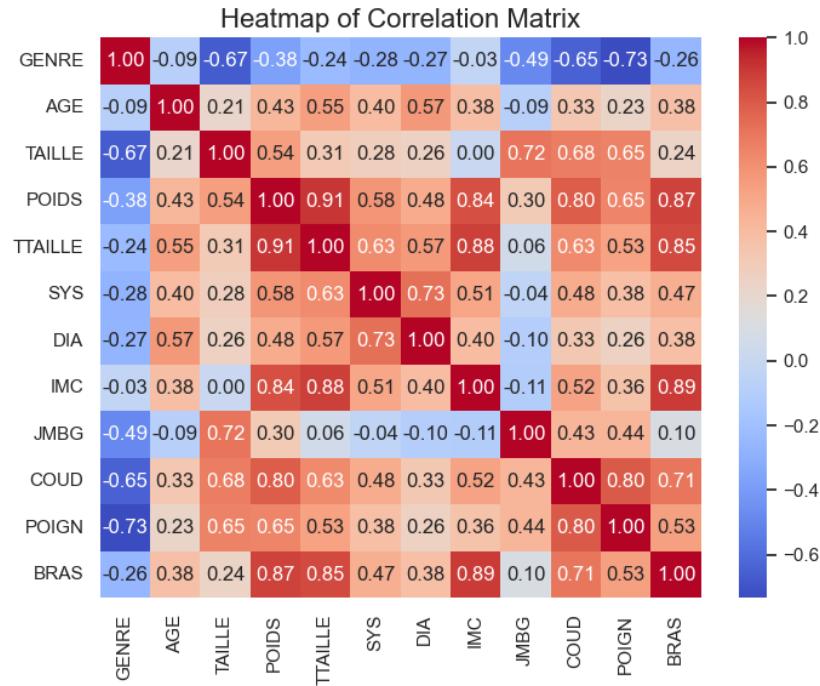


Figure 3.4: Visualisation des boxplots de données

3.4 Conclusion

L'analyse exploratoire des données a permis de valider plusieurs points importants concernant la structure du jeu de données. Nous avons trouvé un équilibre parfait entre les genres (50% hommes, 50% femmes), ce qui suggère une absence de biais de genre dans l'échantillon. Les distributions des variables montrent une tendance proche de la normale, ce qui est favorable pour les modèles de régression linéaire. Cependant, des valeurs aberrantes ont été détectées et supprimées pour garantir la qualité des données. Ces résultats nous permettent de continuer le projet en toute confiance, avec un jeu de données bien préparé pour les étapes suivantes de modélisation.

Chapter 4

Méthodologie

Cette section présente la méthodologie adoptée pour la réalisation de ce projet de régression linéaire visant à prédire la variable POIDS à partir de plusieurs caractéristiques. L'objectif principal de cette étude est d'étudier les relations linéaires entre la variable cible et les autres variables du dataset et d'évaluer la performance du modèle de régression linéaire. Nous détaillerons ici le processus de sélection du modèle, de préparation des données, d'entraînement, d'évaluation des performances et des outils utilisés pour mener à bien ce projet.

4.1 Sélection du Modèle

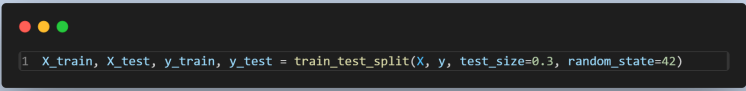
Pour ce projet, nous avons choisi d'utiliser le modèle de Régression Linéaire. Ce choix a été motivé par les résultats obtenus lors de l'analyse exploratoire des données (EDA), qui ont révélé une forte corrélation linéaire entre la variable cible, POIDS, et les autres caractéristiques du dataset. Cette corrélation élevée a rendu la régression linéaire particulièrement adaptée, car elle permet d'établir une relation directe et interprétable entre les variables indépendantes et la variable cible.

4.2 Division des Données

Les données ont été divisées en deux ensembles :

- *Ensemble d'Entraînement* : 70 % des données
- *Ensemble de Test* : 30 % des données

Cette répartition permet de garantir que le modèle bénéficie d'un échantillon suffisant pour l'entraînement tout en laissant un ensemble de test représentatif pour évaluer sa performance. La division a été réalisée de manière aléatoire afin de ne pas introduire de biais dans le processus d'entraînement.

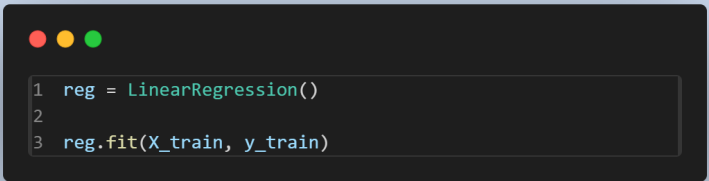
A screenshot of a code editor window with a dark background and three colored window control buttons (red, yellow, green) in the top-left corner. The code is written in a light blue font and shows a single line: `1 X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)`.

```
1 X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)
```

Figure 4.1: Code Python Division des Données

4.3 Entraînement et Validation

Le modèle a été entraîné directement sur l'ensemble d'entraînement sans utiliser de techniques de validation croisée. Le processus d'entraînement s'est déroulé sans difficulté majeure, car l'analyse des données a montré une forte linéarité entre les variables, ce qui a facilité l'ajustement du modèle. En raison de cette linéarité claire, le modèle s'est rapidement ajusté aux données et a donné de bons résultats de validation sur l'ensemble de test.

A screenshot of a code editor window with a dark background and three colored window control buttons (red, yellow, green) in the top-left corner. The code is written in a light blue font and shows three lines: `1 reg = LinearRegression()`, `2` (blank line), and `3 reg.fit(X_train, y_train)`.

```
1 reg = LinearRegression()
2
3 reg.fit(X_train, y_train)
```

Figure 4.2: Code Python Entraînement du Modèle

4.4 Évaluation des Performances

Les performances du modèle ont été évaluées à l'aide des métriques suivantes :

- *Erreur Absolue Moyenne (MAE)* : Cette métrique mesure la magnitude moyenne des erreurs entre les valeurs prédites et les valeurs réelles, en donnant une idée générale de l'erreur moyenne. La formule est la suivante :

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

où y_i est la valeur réelle et \hat{y}_i est la valeur prédite.

- *Erreur Quadratique Moyenne (MSE)* : Elle permet de capturer les écarts quadratiques moyens entre les valeurs prédites et les valeurs réelles, mettant en évidence les erreurs plus importantes. La formule est la suivante :

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

où y_i est la valeur réelle et \hat{y}_i est la valeur prédite.

- *Racine de l'Erreur Quadratique Moyenne (RMSE)* : Cette métrique est l'écart type des erreurs de prédiction, offrant ainsi une mesure des résidus sous forme d'une unité comparable à celle des données d'origine. La formule est la suivante :

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

où y_i est la valeur réelle et \hat{y}_i est la valeur prédite.

- *R-carré (R^2)* : Il indique la proportion de la variance de la variable cible expliquée par le modèle, offrant ainsi une évaluation de la qualité de l'ajustement du modèle. La formule est la suivante :

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

où y_i est la valeur réelle, \hat{y}_i est la valeur prédite, et \bar{y} est la moyenne des valeurs réelles.

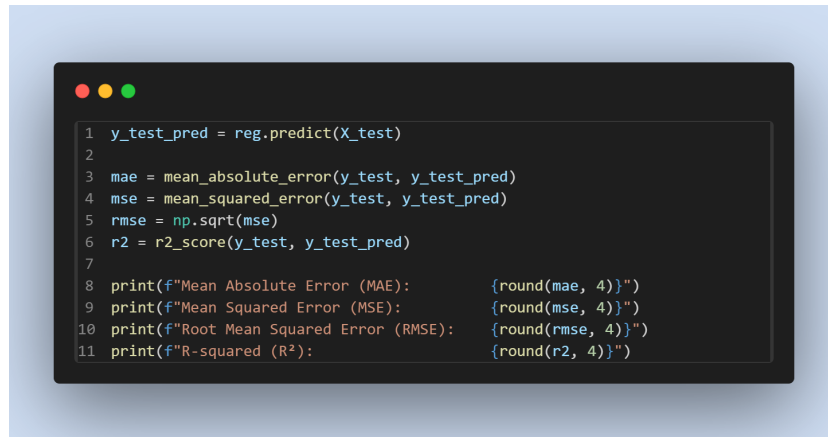


Figure 4.3: Code Python Évaluation du Modèle

4.5 Conclusion

Cette méthodologie a permis de mettre en place un modèle de régression linéaire adapté à la prédiction du poids des patients en fonction de plusieurs caractéristiques. Grâce à l'analyse exploratoire des données, à une préparation minutieuse et à une évaluation rigoureuse des performances, nous avons pu établir une base solide pour évaluer la relation entre les variables et la qualité de notre modèle. Les résultats obtenus fourniront les bases pour une interprétation précise et des recommandations concernant l'application de la régression linéaire à ce problème spécifique.

Chapter 5

Résultats

Dans cette section, nous présentons les résultats obtenus après l'entraînement du modèle de régression linéaire. Nous utiliserons plusieurs métriques standard pour évaluer les performances du modèle sur l'ensemble de test, permettant ainsi d'apprécier la capacité du modèle à prédire la variable cible **POIDS**.

5.1 Évaluation du modèle

Le modèle a été évalué à l'aide de quatre métriques classiques : **Erreur Absolue Moyenne (MAE)**, **Erreur Quadratique Moyenne (MSE)**, **Racine de l'Erreur Quadratique Moyenne (RMSE)** et **R-carré (R^2)**. Les résultats de ces métriques sont les suivants :

Mean Absolute Error (MAE):	0.8083
Mean Squared Error (MSE):	1.1672
Root Mean Squared Error (RMSE):	1.0804
R-squared (R^2):	0.9956

Figure 5.1: Résultats Évaluation du Modèle

Ces résultats indiquent une bonne performance du modèle, avec une faible erreur dans les prédictions. Le **R-carré (R^2)** de 0.9956 suggère que le modèle explique environ 99.56 % de la variance des données, ce qui est un excellent indicateur de la qualité de l'ajustement.

5.2 Interprétation des résultats

Les résultats obtenus montrent que le modèle a bien généralisé aux données de test, avec des erreurs de prédiction relativement faibles. L'**Erreur Absolue Moyenne (MAE)** et l'**Erreur Quadratique Moyenne (MSE)** indiquent que les écarts entre les valeurs réelles et prédites sont minimales. En particulier, la valeur de **RMSE** confirme que l'erreur moyenne est très faible, ce qui renforce la précision du modèle.

5.3 Performance sur les données de test

Le modèle se comporte très bien sur les données de test, avec des erreurs faibles et un **R-carré** proche de 1, ce qui suggère qu'il est capable de généraliser efficacement à de nouvelles données. Cela démontre que le modèle de régression linéaire est bien adapté à ce problème particulier de prédiction de **POIDS**.

5.4 Conclusion

En résumé, le modèle de régression linéaire a montré d'excellentes performances, avec des résultats cohérents et robustes lors de l'évaluation. La faible erreur et la capacité du modèle à expliquer une grande proportion de la variance dans les données renforcent l'efficacité de ce modèle pour prédire la variable cible **POIDS**.

Chapter 6

Conclusion

Ce projet avait pour objectif de prédire la variable **POIDS** en utilisant un modèle de **Régression Linéaire**, en partant du principe qu'il existe une relation linéaire entre **POIDS** et les caractéristiques présentes dans le jeu de données. Après avoir effectué une analyse exploratoire des données (EDA), qui a confirmé les fortes corrélations linéaires entre la variable cible et les autres variables, il a été décidé d'utiliser la régression linéaire. Le jeu de données a été divisé en ensembles d'entraînement et de test, avec 70% des données utilisées pour l'entraînement et 30% pour le test. Le modèle a été entraîné sans utiliser de validation croisée, et aucun problème majeur n'a été rencontré pendant le processus.

Les performances du modèle ont été évaluées à l'aide de plusieurs métriques clés telles que l'**Erreur Absolue Moyenne (MAE)**, l'**Erreur Quadratique Moyenne (MSE)**, l'**Erreur Quadratique Moyenne Racine (RMSE)**, et le **R² (R-squared)**. Les résultats ont montré que le modèle avait des performances satisfaisantes, avec un degré raisonnable de précision dans la prédiction du **POIDS**.

Bien que le projet ait donné des résultats positifs, plusieurs axes d'amélioration peuvent être envisagés pour des travaux futurs :

- **Optimisation des Hyperparamètres** : Dans cette étude, l'optimisation des hyperparamètres n'a pas été réalisée. Des travaux futurs pourraient se concentrer sur l'ajustement des hyperparamètres pour améliorer encore les performances du modèle.
- **Validation Croisée** : L'implémentation de la validation croisée permettrait de fournir une évaluation plus robuste des performances du modèle, réduisant ainsi le risque de surapprentissage (overfitting).
- **Ingénierie des Caractéristiques** : L'exploration de nouvelles car-

actéristiques ou la transformation des caractéristiques existantes pourrait permettre d'améliorer la précision du modèle.

- **Modèles Alternatifs** : Bien que la régression linéaire soit adaptée à ce problème, tester d'autres modèles (comme les arbres de décision, les forêts aléatoires, ou même des réseaux neuronaux) pourrait offrir des perspectives intéressantes et potentiellement améliorer les résultats obtenus.

En conclusion, ce projet a permis de construire et d'évaluer un modèle de régression linéaire pour prédire le **POIDS**, avec des résultats satisfaisants qui ouvrent la voie à des améliorations futures.