École d'Ingénieurs du Littoral
Côte d'Opale

RAPPORT D'

Analyse De données multidimensionn

Rapport réalisé par :

- EL KHAYATI Mohamed Issam
- ELASRY Haitam

Encadré par :

- Pr. ROUSSEL Gilles

Année

2024

-

2025

1. Présentation de la Base de Données

La base de données utilisée dans ce projet provient d'une source publique accessible sur Internet (GitHub). Ces données étaient initialement uniformes et représentaient des informations réelles issues d'observations dans le domaine sportif. Les variables principales incluent :

- Âge (en années),
- Taille (cm),
- Poids (kg),
- Type de sport (Athlétisme, Natation, Cyclisme),
- Temps d'entraînement (en heures),
- Score de performance (valeurs comprises entre 50 et 100).

Pour répondre aux exigences spécifiques du projet et enrichir l'analyse statistique, des manipulations ont été effectuées :

Ajout de valeurs aberrantes : Ces anomalies visent à explorer leur impact sur les analyses statistiques et les visualisations.

Modification des variables et du nombre d'individus : Certaines variables ont été ajustées, et des individus fictifs ont été introduits pour garantir des scénarios diversifiés.

Analyse des valeurs aberrantes : À partir du fichier CSV, nous avons identifié 3 à 4 valeurs aberrantes qui se distinguent nettement sur certaines variables comme la taille, le poids, ou le temps d'entraînement. Ces valeurs influencent les statistiques globales et les visualisations, notamment les boxplots et les histogrammes.

2. Objectifs de l'Analyse

Le code a été utilisé pour explorer les données à travers plusieurs niveaux d'analyse. Voici un résumé des résultats obtenus :

2.1. Analyse Univariée

2.1.1. Statistiques descriptives

Moyennes, médianes et écarts-types montrent des distributions cohérentes pour la majorité des variables.

Les outliers identifiés augmentent significativement les écarts-types et modifient la position des moyennes pour certaines variables.

2.1.2. Visualisations

Histogrammes : Les distributions des variables, comme l'âge et le score de performance, montrent des pics autour des valeurs centrales. Cependant, certaines distributions, comme le poids ou le temps d'entraînement, révèlent des outliers.

Boxplots (Boîtes à Moustaches) : Les boxplots ont permis d'identifier visuellement les valeurs aberrantes.

Exemple : Pour le temps d'entraînement, les valeurs aberrantes dépassent 20 heures, tandis que la majorité des individus s'entraînent entre 5 et 15 heures.

2.2. Analyse Bivariée

2.2.1. Matrice de corrélation

Une corrélation positive notable a été observée entre le temps d'entraînement et le score de performance ($r \approx 0.6$).

Les outliers modifient légèrement les coefficients de corrélation, mais les relations principales restent intactes.

2.2.2. Tests statistiques

T-test entre Athlétisme et Cyclisme :

- Résultat : $t=2.3$, $p=0.03$, suggérant une différence significative.

ANOVA pour les trois sports :

- Résultat : F=4.5F, p=0.02, Confirmant des variations significatives entre groupes.

Test du Chi² entre le type de sport et le temps d'entraînement :

- Résultat : χ2=6.7, p=0.04, indiquant une association significative.

2.3. Analyse Multivariée (ACP)

2.3.1. Composantes principales

Les deux premières composantes principales expliquent 72% de la variance totale.

Variables les plus influentes : Taille, Poids, et Score de performance.

2.3.2. Impact des outliers

Les valeurs aberrantes apparaissent comme des points isolés sur les projections des composantes principales, modifiant légèrement la structure des groupes.

3. Interprétation Mathématique des Résultats

3.1. Impact des Outliers

À partir des boxplots et des statistiques descriptives, nous avons confirmé la présence de 3 à
4 valeurs aberrantes.

Les outliers augmentent la variance et l'écart-type, rendant certaines analyses sensibles à leur présence. Par exemple :

Pour le poids, les valeurs aberrantes dépassent 120 kg, bien au-delà des autres observations.

Dans le temps d'entraînement, des valeurs supérieures à 25 heures ont été identifiées, contre une médiane d'environ 10 heures.

3.2. ACP

Les variables fortement corrélées (taille, poids) sont regroupées dans la même composante principale.

Les valeurs aberrantes, en modifiant la moyenne et la variance, affectent la direction et la contribution des axes principaux.

3.3. Conclusion mathématique

Les valeurs aberrantes mettent en lumière l'importance du nettoyage des données avant de mener une analyse statistique robuste.

Les méthodes robustes, comme l’ACP ou l’analyse par médiane, peuvent atténuer leur influence sur les résultats globaux.