Rapport : Analyse des Performances des Athlètes du dataset Décathlon via l'ACP

KINDO Harouna

October 1, 2024

Professeure: Mme BANOUAR Oumayma

Contents

1 Introduction **3**

2 Description des données **3**

3 Standardisation des données **4**

4 Analyse en Composantes Principales (ACP) **5**

5 Diagramme des Valeurs Propres **5**

6 Projection des Athlètes et des Épreuves **6**

7 Interprétation des Profils **7**

8 Partie Autonomie: Analyse des Profils Gagnants **7**
8.1 Visualisation des variables *rank*, *points* et *competition* 8
8.2 Profils des athlètes gagnants 9

9 Conclusion **9**

1 Introduction

Dans le cadre de cette activité pratique, une analyse en composantes principales (ACP) a été réalisée sur les performances des athlètes de décathlon. Le jeu de données utilisé contient des informations sur 10 épreuves de décathlon pour plusieurs athlètes. L'objectif est de réduire la dimensionnalité des données et de dégager des informations d'exploration des données.

2 Description des données

[2]:
```
data=pd.read_csv("/kaggle/input/decathlon/decathlon.csv")
data.head(10)
```

[2]:

	Unnamed: 0	100m	Long.jump	Shot.put	High.jump	400m	110m.hurdle	Discus	Pole.vault	Javeline	1500m	Rank	Points	Competition
0	SEBRLE	11.04	7.58	14.83	2.07	49.81	14.69	43.75	5.02	63.19	291.7	1	8217	Decastar
1	CLAY	10.76	7.40	14.26	1.86	49.37	14.05	50.72	4.92	60.15	301.5	2	8122	Decastar
2	KARPOV	11.02	7.30	14.77	2.04	48.37	14.09	48.95	4.92	50.31	300.2	3	8099	Decastar
3	BERNARD	11.02	7.23	14.25	1.92	48.93	14.99	40.87	5.32	62.77	280.1	4	8067	Decastar
4	YURKOV	11.34	7.09	15.19	2.10	50.42	15.31	46.26	4.72	63.44	276.4	5	8036	Decastar
5	WARNERS	11.11	7.60	14.31	1.98	48.68	14.23	41.10	4.92	51.77	278.1	6	8030	Decastar

Figure 1: chargement des données

Le jeu de données comprend les performances dans les 10 épreuves du décathlon : 100m, saut en longueur, lancer du poids, saut en hauteur, 400m, 110m haies, lancer de disque, saut à la perche, lancer de javelot, et 1500m. Les variables additionnelles incluent le classement (*rank*), les points cumulés (*points*), et la compétition à laquelle l'athlète participe (*competition*).

[3]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 41 entries, 0 to 40
Data columns (total 14 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   Unnamed: 0   41 non-null     object
 1   100m         41 non-null     float64
 2   Long.jump    41 non-null     float64
 3   Shot.put     41 non-null     float64
 4   High.jump    41 non-null     float64
 5   400m         41 non-null     float64
 6   110m.hurdle  41 non-null     float64
 7   Discus       41 non-null     float64
 8   Pole.vault   41 non-null     float64
 9   Javeline     41 non-null     float64
 10  1500m        41 non-null     float64
 11  Rank         41 non-null     int64
 12  Points       41 non-null     int64
 13  Competition  41 non-null     object
dtypes: float64(10), int64(2), object(2)
memory usage: 4.6+ KB
```

Figure 2: informations sur les données

Nous notons que notre dataset est composé de 41 observations. Il n'ya pas de données manquantes, ce qui est un point positif et aussi les données sur les différents sport sont toutes numériques.

3 Standardisation des données

[7]:
```
scaler=StandardScaler()
```

[8]:
```
sport=data.iloc[:, 1:-3]
data_scaled=scaler.fit_transform(sport)
```

Figure 3: standardisation

L'ACP nécessite que les données soient centrées et réduites, c'est-à-dire que chaque variable ait une moyenne de 0 et un écart-type de 1. Cela permet d'éviter que des variables avec de grandes échelles ne dominent l'analyse. Cette étape garantit que toutes les variables contribuent de manière équitable. Par exemple dans notre cas, les valeurs du 100m sont entre 10 et 11 pourtant les valeurs du 1500m atteignent les 300. La standardisation permet d'éviter que cette variance des données n'impacte pas notre analyse.
Vu que l'ACP s'applique sur des données quantitatives, nous avons utilisé uniquement les données sur les 10 épreuves pour la standardisation.

4 Analyse en Composantes Principales (ACP)

Nous avons effectué une ACP sur les 10 variables relatives aux épreuves. En fait, nous pouvons avoir autant de composantes que de variables. Plus il y aura de composantes, plus le modèle pourra capturer la variance des données. Les variables rank, compétition et points serviront à analyser les profils gagnant parmi nos athlètes.

4. Application de l'ACP sur les 10 colonnes

[9]:
```
# Application du PCA
pca = PCA(n_components=10)  # On choisit de réduire à 10 dimensions
data_pca = pca.fit_transform(data_scaled)
```

[10]:
```
explained_variance = pca.explained_variance_ratio_
```

Figure 4: Implémentation de l'ACP

5 Diagramme des Valeurs Propres

Les résultats montrent que les 10 composantes principales expliquent la variance totale de mes données. En effet, la variance cumulée du nous donne 1. Les deux premières composantes expliquent ensemble environ 50% de la variance totale des données, ce qui justifie leur utilisation pour la visualisation et l'interprétation.

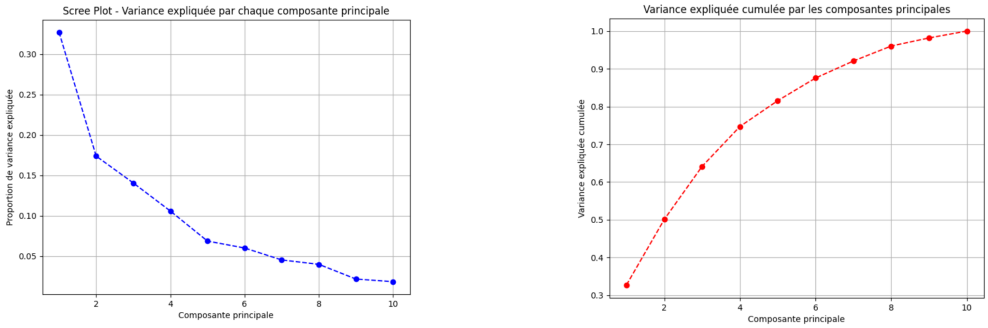

Figure 5: Diagramme des valeurs propres

Ce diagramme montre l'importance relative de chaque composante principale dans l'explication de la variance totale. Nous pouvons bien rémarqué

que les deux premières composantes capturent l'essentiel de la structure des données.
Cependant si nous souhaitons effectuer un processus d'apprentissage (exemple kmeans) sur la projection des performances des athlètes sur un espace réduit, il est necessaire d'utliser au moins 04 composantes principales afin de pour capturer 75% de la variance des données.

6 Projection des Athlètes et des Épreuves

En projetant les épreuves sur les deux premières dimensions, nous obtenons une représentation visuelle des corrélations entre les épreuves.

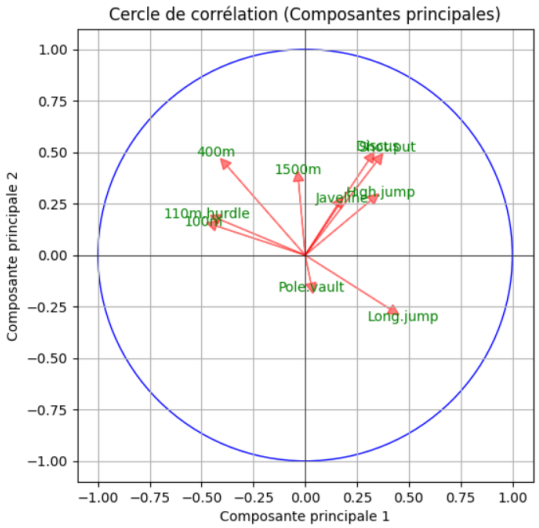

Figure 6: Corrélations entre les épreuves

Ce graphique nous montre 03 types de corrélations entre les épreuves:

1. **Corrélations fortes positives** : Certaines épreuves comme le 100m et le 110m ou encore le lancer de disque et le lancer de poids sont fortement corrélées, ce qui signifie que les athlètes performants dans une épreuve tendent à l'être dans l'autre.

2. **Corrélations faibles ou inexistantes** : Certaines épreuves, comme le lancer de poids et le saut en hauteur, présentent des corrélations faibles ou presque inexistantes. Les performances dans l'une ne prédisent pas nécessairement les performances dans l'autre.

3. **Corrélations négatives** : Dans certains cas, les athlètes qui excellent dans des épreuves de vitesse (comme le 100m) peuvent avoir des performances moins bonnes dans des épreuves de résistance (comme le 1500m), ce qui suggère une corrélation négative.

7 Interprétation des Profils

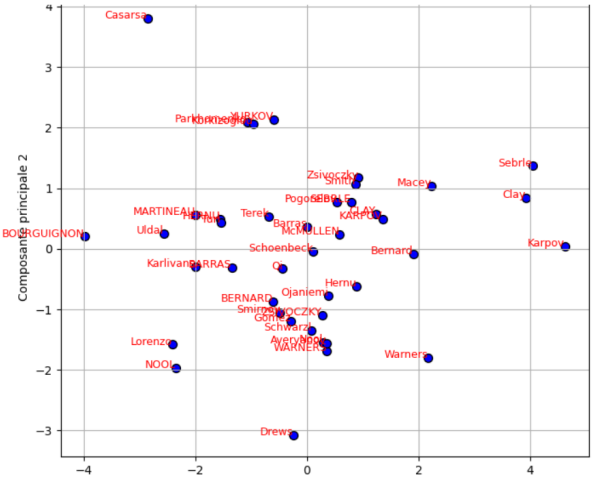

Figure 7: Visualisation des profilses.

Les athlètes peuvent être segmentés en plusieurs groupes selon leurs performances dans les différentes épreuves. Par exemple, certains athlètes se distinguent dans les épreuves de vitesse (100m, 400m), tandis que d'autres sont plus performants dans les épreuves de force (lancer du poids, lancer de javelot). Par nous pouvons observer que Warners est un bon profil en saut en longueur alors que Martineau et Uldal ont plus des profils de vitesse(100m et 110m)

8 Partie Autonomie: Analyse des Profils Gagnants

Dans cette partie, nous nous intéressons à l'identification des profils gagnants, c'est-à-dire ceux des athlètes ayant remporté chaque compétition dans notre jeu de données. Les variables principales utilisées dans cette analyse sont le *rank* (classement), les *points* obtenus par chaque athlète, et la *competition* à laquelle ils ont participé.

8.1 Visualisation des variables *rank*, *points* et *competition*

Pour mieux comprendre les performances des athlètes, nous avons visualisé les variables *rank* (classement), *points*, et *competition*. Ces visualisations permettent de faire ressortir les profils des athlètes gagnants dans chaque compétition.

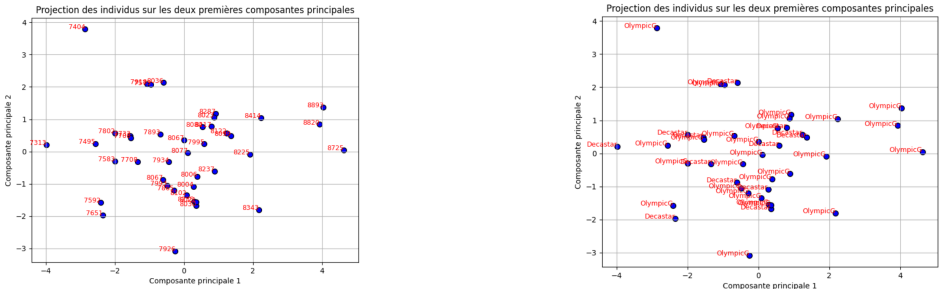

Figure 8: Visualisation des profils en fonction du points et de la compéttion

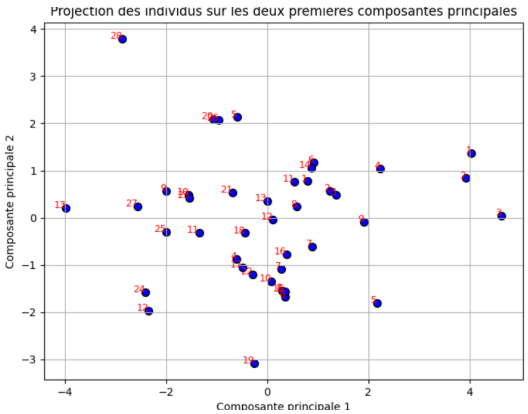

Figure 9: Visualisation des profils en fonction du rang.

Ces graphiques montre la distribution des points en fonction du classement des athlètes pour chaque compétition. Les athlètes en tête (ceux avec les classements les plus bas) obtiennent généralement un nombre de points plus élevé, ce qui reflète leurs performances globales. Cela nous permet d'identifier les athlètes dominants dans chaque compétition.

8.2 Profils des athlètes gagnants

À partir de ces visualisations, nous pouvons identifier les profils suivants :

- **Athlètes de la compétition Décastar** : Les athlètes de cette compétition ont des performances homogènes, avec des points cumulés élevés et une bonne répartition des compétences à travers les différentes épreuves. Les athlètes classés 1er, 2e et 3è dans cette compétition ont obtenu plus de 8000 points chacun, se distinguant par leur performance oriéntée javelot, saut en hauteur, lancer de poids et de disque.

- **Athlètes des Jeux Olympiques** : Les gagnants de cette compétition ont également des performances excellentes, mais leur profil est davantage marqué par des spécialités dans certaines épreuves spécifiques. Par exemple, le 1er, le 2è et le 3è ont montré une certaine dominance dans les épreuves de lancer (poids, disque), ce qui leur a permis de se distinguer des autres athlètes.

Ces analyses montrent que les profils gagnant dans chacune des 02 compétitions ont tendance à obtenir des meilleures performances dans les épreuves de force (lancer du poids et de disque, lancer de javelot et saut en hauteur)

9 Conclusion

L'ACP nous a permis de réduire la dimensionnalité des données du décathlon et d'identifier les relations entre les épreuves et les athlètes. En visualisant les données dans un espace de dimension réduite, il est possible de mieux comprendre les profils de performance et de segmenter les athlètes selon leurs points forts.