Clustering Analysis for Athletes Data

Sharmin Akhter
ID. 201891112
Master of Data Science
Memorial University of Newfoundland

August 11, 2023

Contents

1 Introduction

The provided dataset features information on athletes from various sports disciplines. The primary objective of gathering this data was to investigate potential differences in blood hemoglobin levels between athletes engaged in endurance-based activities and those participating in power-based events. In our analysis, we will categorize the athletes according to their respective events using a clustering algorithm. Additionally, we will assess whether the algorithm effectively distinguishes male and female athletes. Ultimately, we will introduce a new categorical variable to represent each athlete's involvement in endurance or power events and evaluate the clustering algorithm's adherence to this classification.

2 Aim of the Analysis

1. Data contains information for athletes from various sports.
2. Test if blood hemoglobin levels differ between endurance and power athletes.
3. Analysis aims to cluster athletes into their respective events.
4. Check if clustering algorithm separates athletes by gender.
5. Evaluate if the clustering algorithm follows the endurance/power classification.

Cluster the data into ten classes each one representing a sport. If that is not possible then Cluster into female and male or power-related and endurance-related.

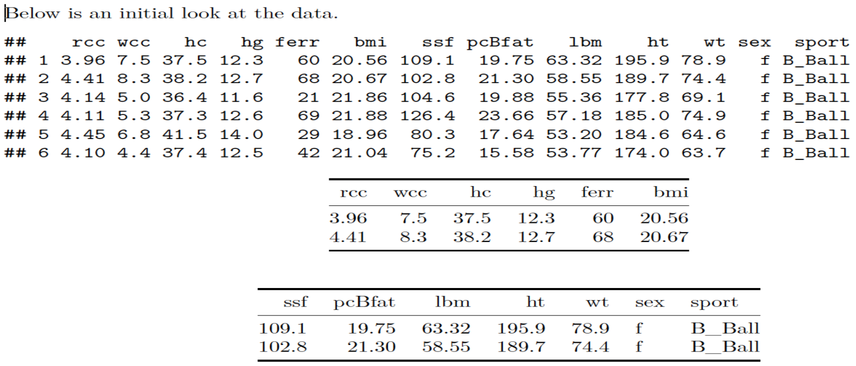
3 Data Information

The dataset contains information on Australian athletes, with 202 observations representing individual athletes and 13 variables, including the class variable indicating the sport each athlete plays. The data is obtained from R.

Summary of Variables:

1. Rcc - Red blood cell count (quantitative)
2. Wcc - White blood cell count per liter (quantitative)
3. Hematocrit - Percent of hematocrit (quantitative)
4. Hg - Hemoglobin concentration in g per decaliter (quantitative)
5. Ferr - Plasma ferritins in ng (quantitative)

6. Bmi - Body mass index in kg/m2 (quantitative)
7. Ssf - Sum of skin folds (quantitative)
8. PcBfat - Percentage of body fat (quantitative)
9. Lbm - Lean body mass in kg (quantitative)
10. Ht - Height in cm (quantitative)
11. Wt - Weight in kg (quantitative)
12. Sex - Athlete's sex (qualitative categorical)
13. Sport - Sport played by the athlete, used as the class variable (qualitative categorical)



Below is an initial look at the data.

```
##     rcc wcc   hc   hg ferr   bmi   ssf pcBfat   lbm    ht   wt sex  sport
## 1 3.96 7.5 37.5 12.3   60 20.56 109.1  19.75 63.32 195.9 78.9   f B_Ball
## 2 4.41 8.3 38.2 12.7   68 20.67 102.8  21.30 58.55 189.7 74.4   f B_Ball
## 3 4.14 5.0 36.4 11.6   21 21.86 104.6  19.88 55.36 177.8 69.1   f B_Ball
## 4 4.11 5.3 37.3 12.6   69 21.88 126.4  23.66 57.18 185.0 74.9   f B_Ball
## 5 4.45 6.8 41.5 14.0   29 18.96  80.3  17.64 53.20 184.6 64.6   f B_Ball
## 6 4.10 4.4 37.4 12.5   42 21.04  75.2  15.58 53.77 174.0 63.7   f B_Ball
```

rcc	wcc	hc	hg	ferr	bmi
3.96	7.5	37.5	12.3	60	20.56
4.41	8.3	38.2	12.7	68	20.67

ssf	pcBfat	lbm	ht	wt	sex	sport
109.1	19.75	63.32	195.9	78.9	f	B_Ball
102.8	21.30	58.55	189.7	74.4	f	B_Ball

Figure 1: Data Image

4 Clustering into Sports

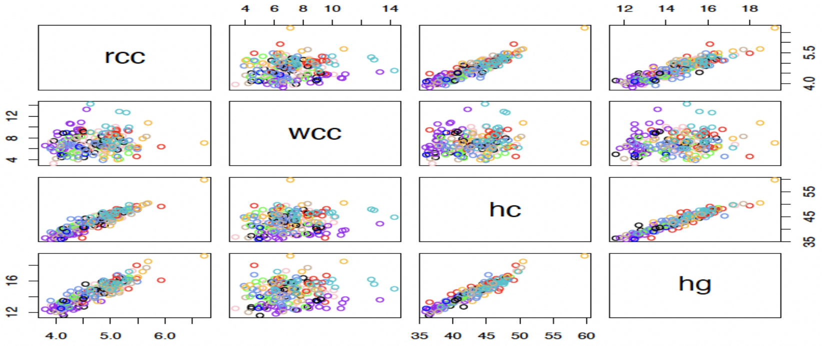


Figure 2: Clustering into Sports

Interpretation: The various sports are colored differently. It is obvious that the sports are not easily distinguishable from each other using two variables.

5 Clustering into Gender

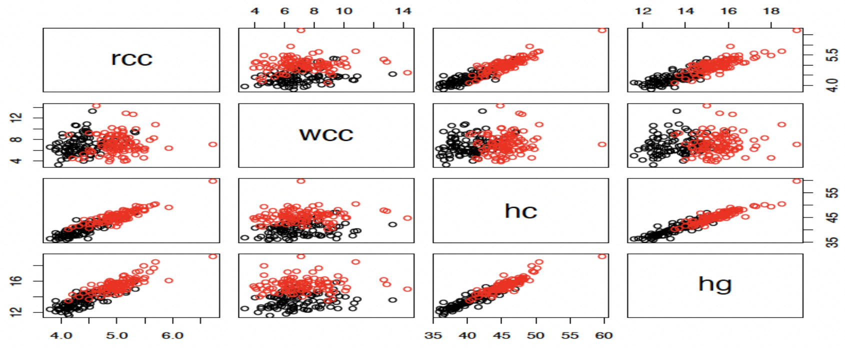


Figure 3: Clustering into Gender

Interpretation: In 3 , we can see that the two clusters are clearly distinguishable. Men are represented by red points and women are represented by black points. Hemoglobin and Hematocrit are the most effective variables for differentiating between Men and Women.

6 Clustering into Two Sports Groups

1. **Endurance sports:** Basketball, Rowing, Sprint (400m), Tennis, and Water Polo.
2. **Power sports:** Gym, Netball, Swimming, Sprints (¡400m), and Field.

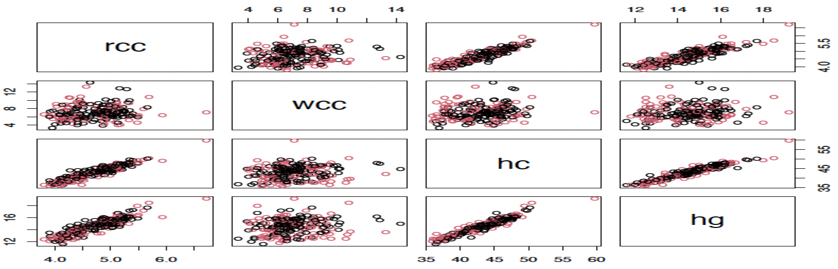


Figure 4: Clustering into two sports group

Interpretation: we can see in 4 that it is not easy to distinguish between the sports in 2D.

7 Hierarchical Clustering k=2, k=10, k=2

Confusion Matrix for Gender, Sports, Power/Endurance Sports

Confusion Matrix (Clusters Represent Gender)

1	2
99	5
1	97

: Confusion Matrix (Clusters Represent Sports)

4	5	3	10	8	7	1	6	2	9
17	2	0	0	0	0	5	1	0	1
2	19	0	0	0	2	5	5	0	1
3	1	4	0	0	7	2	3	0	3
0	1	0	4	0	0	1	0	6	0
1	1	0	0	4	4	0	0	1	2
0	0	0	4	5	15	5	7	4	2
0	11	0	5	1	0	6	4	1	0
0	0	0	1	3	1	0	2	0	1
0	2	0	3	0	0	1	0	7	0
0	0	0	0	2	0	0	0	0	1

Confusion Matrix (Clusters Represent Power/Endurance

2	1
47	57
36	62

Figure 5: Confusion Matrix

Interpretation: From the confusion matrix in 5 we can see that clusters represent Gender perfectly match.

8 Hierarchical Clustering and K-means

1. **Error Rate and R^2 for Hierarchical**

   (a) Cluster Represent Gender 2.9% and 0.3470876

   (b) Cluster Represent Sports 60% and 0.6568497

   (c) Cluster Represent Power/Endurance Sports 46% and 0.3470876

1. **Error Rate and R^2 for K-mean**

   (a) Cluster Represent Gender 3.9% and 0.3551394

   (b) Cluster Represent Sports 61% and 0.9355204

   (c) Cluster Represent Power/Endurance Sports 47% and 0.3551394

9 Conclusion

1. Data is best suited for clustering male and female athletes due to the lowest classification error when k = 2.

2. Clustering based on sports was not successful, but may be possible with more data.

3. Clustering sports into power and endurance categories has potential, but requires better judgment of which sports belong to each category.

4. Evaluating all combinations to find the lowest error rate is possible, but computationally expensive and time-consuming.