Assignment 1 - Starting with Clustering

* **Due** Apr 8 by 2:29pm

* **Points** 10

* **Submitting** a file upload

* **Available** until Apr 8 at 2:29pm

This assignment was locked Apr 8 at 2:29pm.

Consider the [dataset](https://bits-pilani.instructure.com/courses/341/files/66861/download)comprising of 109 cricket players (attached here with assignment), filename is ‘cricketers.csv’. Each row comprises of following details of a cricketer:

1. matches\_played: number of matches played by the player,
2. innings\_batted: number of times got the chance to bat,
3. runs\_scored: total runs scores across all innings,
4. highest\_runs: highest run scored taking all innings in account,
5. ball\_faced: total balls faced across all innings,
6. average\_runs: average runs scores taking all innings in account,
7. strike\_rate: it is computed as (runs\_scored / balls\_faced) x 100,
8. innings\_bowled: number of times got the chance to bowl,
9. overs: number of overs bowled across all innings,
10. runs\_given: total run given across all innings,
11. wickets\_obtained: total wickets obtained across all innings,
12. average\_runs\_per\_wicket: it is computed as (runs\_given / wickets\_obtained)
13. bowling\_economy: average runs scored per over

Questions:

Notes: Write all code related to this assignment in single jupyter notebook. Implement K-means algorithm from scratch, do not use any predefined ML library function for K-means algorithm. Use single word document to answer questions.

1. Given that K-means depends on distance metric, it is a convention to normalize the data attributes so that attributes are on the same scale.  So, in this first task, normalize all data attributes. [2 points]

2. Write your own code for K-means algorithm using two attributes namely average\_runs and bowling\_economy. Take K=2. Plot clusters on a scatter plot with X and Y being the two attributes namely average\_runs and bowling\_economy, respectively. Color data points belonging to the first cluster with red and the second cluster with blue. Copy the plot diagram in the word document and interpret the output. [3 points]

3. Redo question-2 on different values of K = 2,3,4,5. For each case, draw the plot of clusters as stated above. Visualize these plots, copy the plot diagrams in the word document,  and comment on which is better clustering (and reasons) based on visualization only. [3 points]

4. Write a few lines in a word document about the interpretation of the best clusters obtained. Also write a few statements about how these clusters can be useful. [2 points]

Deliverables:

One jupyter notebook for Q 1,2,3.

One word document for Q 2,3,4.

PS: For any queries, contact rishabh.kaushal@wilp.bits-pilani.ac.in

Rubric

**Rubric - Assignment 1**

| Rubric - Assignment 1 | | |
| --- | --- | --- |
| **Criteria** | **Ratings** | **Pts** |
| This criterion is linked to a Learning OutcomeQ1 - normalize all features  full 2 pts if all features covered else 1 pt for some features | |  |  | | --- | --- | | **2.0 pts**  **Full Marks** | **0.0 pts**  **No Marks** | | 2.0 pts |
| This criterion is linked to a Learning OutcomeQ2 - kmeans & plot  2 pts for K-means algo from scratch 1 pt for plot | |  |  | | --- | --- | | **3.0 pts**  **Full Marks** | **0.0 pts**  **No Marks** | | 3.0 pts |
| This criterion is linked to a Learning OutcomeQ3 - Kmeans plots  1pt each for cluster plots for K=3,4,5 along with explanation | |  |  | | --- | --- | | **3.0 pts**  **Full Marks** | **0.0 pts**  **No Marks** | | 3.0 pts |
| This criterion is linked to a Learning OutcomeQ4 - cluster description  1 pt for cluster interpretation like which is better 1 pt for usefulness in context of problem like it helps in identifying batsmen and bowlers | |  |  | | --- | --- | | **2.0 pts**  **Full Marks** | **0.0 pts**  **No Marks** | | 2.0 pts |
| Total Points: 10.0 | | |