**Case study – Clustering of students into sections A, B & C**

**Introduction**

This case study demonstrates the use of cluster analysis to divide 200 students into three sections: A, B, and C. The analysis utilizes a dataset containing student information, including their marks in reading, writing, mathematics, science, and social science, to group students based on academic performance.

**Objective**

To perform cluster analysis to divide 200 students into three sections: A, B, and C.

**Dataset link**

<https://drive.google.com/file/d/1_O-bgbAaVYvHOkkrx_x8anj_o5aNN21Q/view?usp=sharing>

**Understanding the data**

The dataset contains the information of 200 students. Following are the description of columns:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Column |  |  |  | Description | Values |
| id |  |  |  | Roll number of the student | Unique integer identifier for each student |
| Gender |  |  |  | Gender of the student | 0 = Male, 1 = Female |
| race |  |  |  | Racial background of the student | 1 = North, 2 = South, 3 = East, 4 = West |
| ses |  |  |  | Socio-economic status | 1 = Low, 2 = Medium, 3 = High |
| schtyp |  |  |  | School type | 1 = Public School, 2 = Private School |
| prog |  |  |  | Academic program enrolled in | 1 = BBA, 2 = MBA, 3 = B.Tech |
| read |  |  |  | Reading marks obtained | Numerical value representing marks |
| write |  |  |  | Writing marks obtained | Numerical value representing marks |
| math |  |  |  | Mathematics marks obtained | Numerical value representing marks |
| science |  |  |  | Science marks obtained | Numerical value representing marks |
| socst |  |  |  | Social Science marks obtained | Numerical value representing marks |

**Procedure for coding**

* Import necessary libraries
* Load dataset
* Check the number of rows and columns
* Check for missing values
* Define independent (X) and dependent (y) variables
* Standardize independent variables
* Compute distance matrices (euclidean, cityblock, minkowski having p value 3 & 4) for each pair
* Sort distances in ascending order
* Plot dendrogram
* Assign clusters
* Label sections as A, B, C
* Display the number of students in each class
* Create separate data frames for each section

**Code File Link**

<https://github.com/Ishita2003M/Clustering-of-students-into-section-A-B-or-C/blob/main/stud_clust.ipynb>

**Interpretation and conclusion**

1. The analysis aimed to classify students into meaningful academic sections based on their intellectual performance across key subject areas using clustering techniques.
2. The dataset includes scores of 200 students across five academically relevant subjects that serve as proxies for IQ and academic proficiency:
   * Reading
   * Writing
   * Mathematics
   * Science
   * Social Science
3. These features were standardized to bring them to a common scale, ensuring that no individual subject disproportionately influenced the clustering outcome.
4. Distance Metrics and Combination:  
   Multiple pairwise distance matrices were calculated using diverse distance metrics, including Euclidean, Cityblock (also known as Manhattan), and higher-order Minkowski distances (p = 3 and p = 4). A robust combined distance matrix was formed by selecting the minimum distance value across all metrics for each student pair, emphasizing the strongest possible similarity.
5. Hierarchical clustering was applied using Ward’s linkage method, which focuses on minimizing intra-cluster variance.
6. A dendrogram was plotted to visualize how students were grouped based on academic profiles.
7. The optimal number of clusters was set to three, representing different academic sections. These clusters were subsequently labeled as:
   * Sec A
   * Sec B
   * Sec C
8. The number of students classified into each section was as follows:
   * Section C: 77 students
   * Section B: 65 students
   * Section A: 58 students
9. Insights:
   * Section C comprises the largest group, possibly indicating a broad mid-range of academic performers.
   * Section A, being the smallest, may represent students with either high academic performance or unique learning profiles, depending on how clustering aligned.
   * This segmentation allows educators to tailor instruction, support, and enrichment strategies for each group more effectively.