DALARNA UNIVERSITY

Assignment

Author
Daniel Stevens
email: h17daste@du.se

Omar Omran
email: h24omaom@du.se

Course code
AMI22T

Semester and year
Spring semester 2025

Home Exercise 2

OK report. There are issues in the clustering part and the presentation of results. Based on the figures it would seem you have 10 clusters for hclust. You receive a grade of 12 for this report.

1 Introduction

This exercise contained two tasks to be performed on a dataset.

The dataset, "student_performance_large_dataset_new.csv" (Shamim, 2025), contains personal information about the students' age and gender; their weekly hours for studying, sleeping, and spending time on social media; their preferred learning style; their number of online courses and percentage of assignments completed; their attendance rate; whether they participated in discussions or used "educational tech"; their self-reported stress level; and finally their exam score and resulting final grade.

The first task was to use decision trees, SVM models, random forests, and optionally boosting/BART to predict the students' performance based on the dataset. This will be covered under the subheadings labelled "Task 1" in this document.

The second task was to use k-means and hierarchical clustering on the full dataset (less the response) and discuss the results. This will be covered under the subheadings labelled "Task 2".

2 Statistical Methods

The code for these tasks was done in Python, and the source code is provided in separate files.

1.1. Task 1

2.1.1 Data Preprocessing

As we decided to perform a class-based rather than quantitative prediction, the Exam_Score column was dropped in favour of Final_Grade. The Student_ID column, containing the unique ID of each student, was also dropped as it was deemed irrelevant.

To be able to use the categorical columns in our classifiers, we encoded them using label encoding.

The data was then split into a 80:20 train-test split, and Final_Grade was used as the class to be predicted.

2.1.2 Decision Tree

To properly prune the tree, we used cost complexity pruning to determine the point where validation accuracy was at its highest. (Fig. 1)

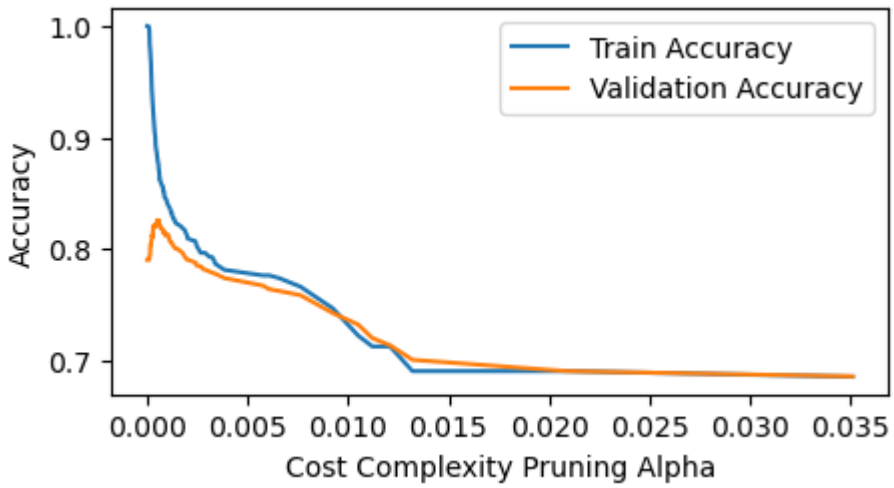

Figure 1 Decision Tree accuracy by CCP alpha

This resulted in a tree of depth 10. A cropped version can be seen in Fig. 2, with the most telling factor for determining the grade appearing to be how much sleep the student gets.

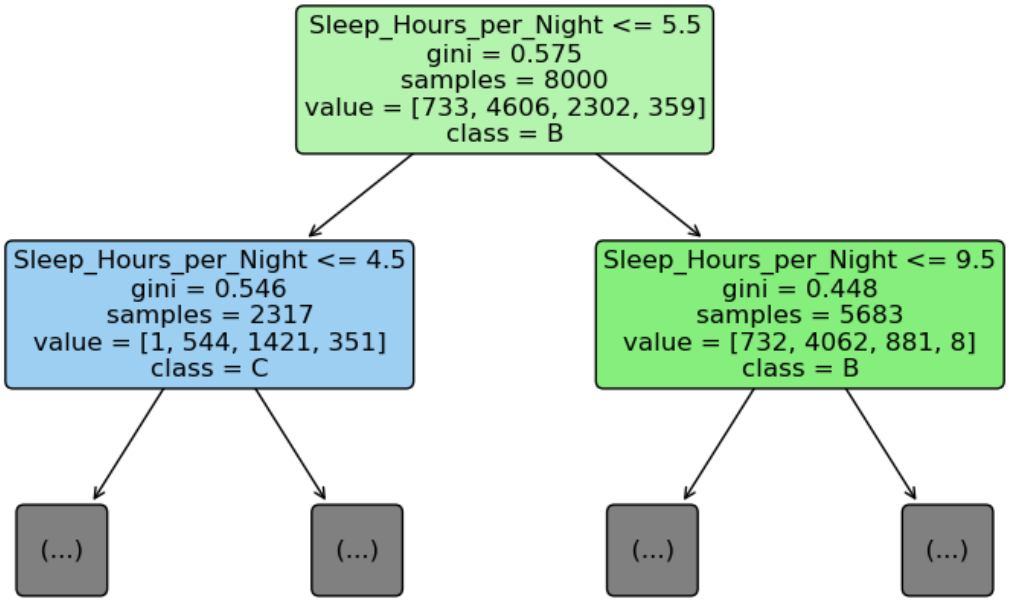

Figure 2 Decision tree plot, trimmed to depth 1

2.1.3 SVM

As we had the time for it, for the SVM we ran a grid-search with five-fold cross-validation. After 75 minutes and 23.9 seconds of computing, it was determined that the best parameters were to have the default scikit-learn settings (radial basis function as the kernel, kernel coefficient set to “scale”), except with a time-consuming regularization parameter of 10.

2.1.4 Random Forest

The random forest function was again done with a five-fold cross-validation grid search, with the best parameters being determined to be to have no limit on the max depth, a limit of sqrt(n) for the feature amount, and 200 trees in the forest.

2.1.5 Gradient Boosting

Gradient boosting, much like with the SVC and random forest, was done using a five-fold cross-validation grid search. The best parameters appear to be a learning rate of 0.1, a max depth of 3 on each tree, and 200 boosting stages.

2.2 Task 2

2.2.1 Data Preprocessing

The dataset was cleaned by removing unnecessary columns, including `Student_ID`, index, and the outcome variables `Final_Grade` and `Exam_Score`. Categorical features such as stress levels and participation indicators were transformed into binary dummy variables to ensure compatibility with clustering algorithms. All numeric features were then standardized to have zero mean and unit variance, allowing them to contribute equally during clustering.

2.2.2 Determining the Number of Clusters (Elbow Method)

To determine an appropriate number of clusters, we applied the elbow method by running k-means clustering for k values ranging from 1 to 10. We plotted the within-cluster sum of squares (WCSS) and identified an "elbow" at k = 4, where the rate of decrease in WCSS began to level off. This indicated that four clusters provided a good balance between compactness and interpretability.

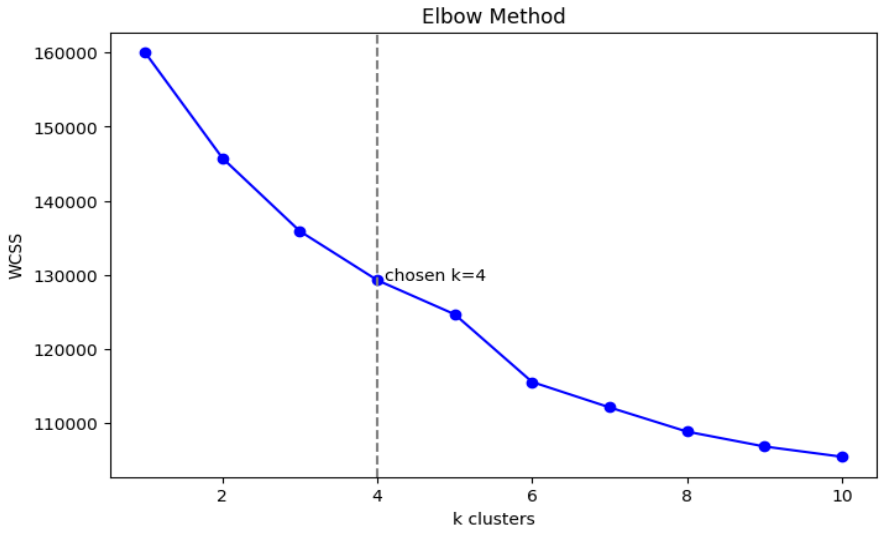

Figure 3 Elbow plot showing optimal number of clusters at k = 4 based on WCSS.

2.2.3 k-Means Clustering

We implemented k-means clustering with four clusters, using 10 random initializations to improve clustering reliability, and a fixed seed for reproducibility. The algorithm converged in under 10 iterations, confirming stable clustering behavior.

2.2.4 Hierarchical Clustering (Average Linkage)

In addition to k-means, hierarchical clustering was applied using average linkage, which merges clusters based on the average pairwise distance between observations. A dendrogram was generated to visualize the hierarchical structure, and the tree was cut at four clusters for consistency with the k-means analysis.

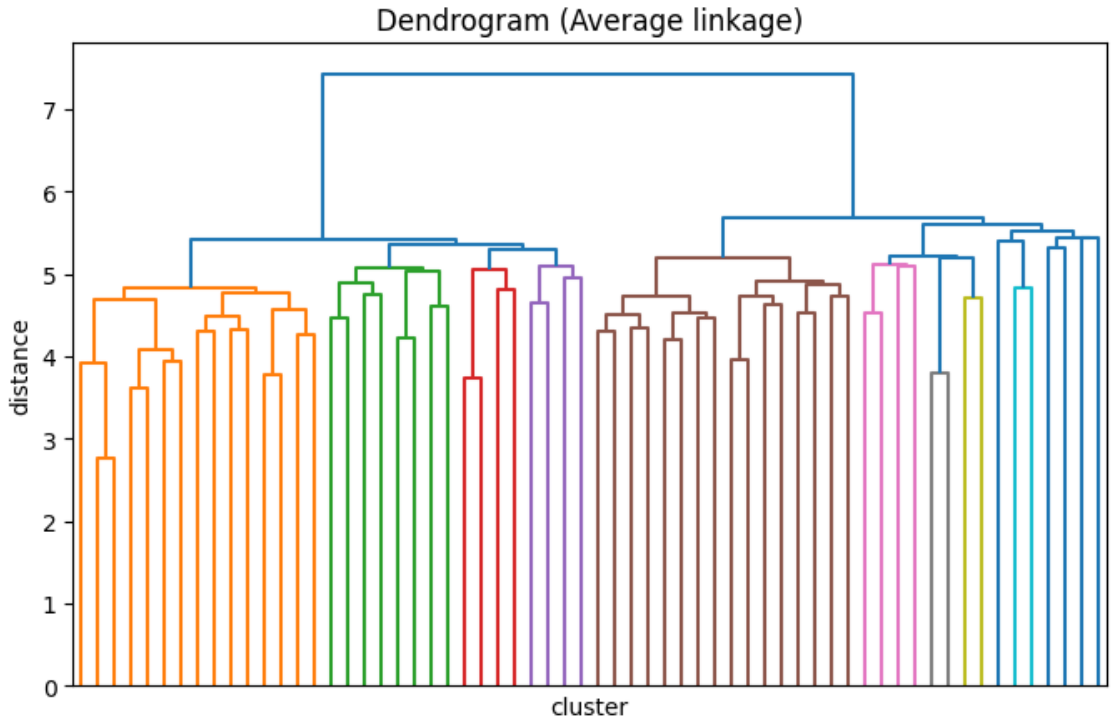

Figure 4. Dendrogram using average linkage, cut at four clusters for comparison with k-means.

2.2.5 Dimensionality Reduction and Visualization (PCA)

To visualize the cluster separations, we performed Principal Component Analysis (PCA) to reduce the dataset to lower dimensions. Scatter plots were generated using the first few principal components, comparing cluster assignments from both k-means and hierarchical clustering. This allowed for direct visual comparison of the groupings and how well each method separated students in 2D space.

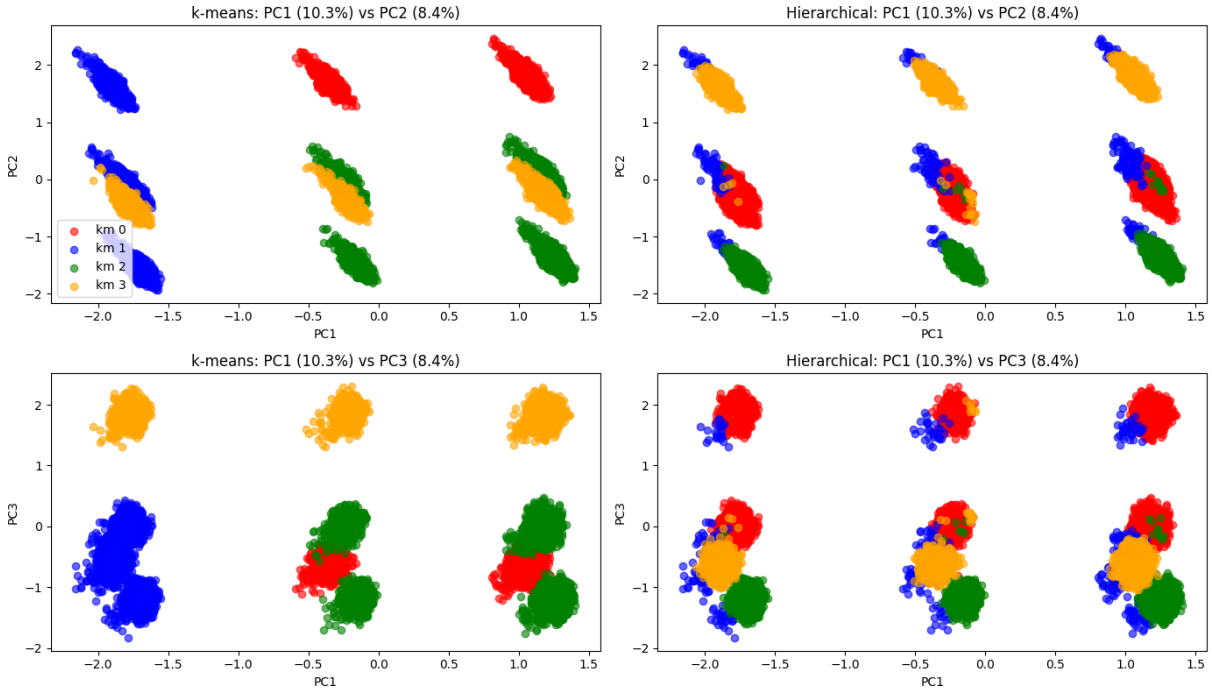

Figure 5 Side-by-side PCA projections comparing k-means (left) and hierarchical clustering (right) using PC1 against PCs 2–5. Variance explained by each component is not

2.2.6 Cluster Comparison and Validation

To evaluate the similarity between the two clustering approaches, we calculated the Adjusted Rand Index (ARI), which was approximately **0.31**. This indicates moderate agreement between the methods. External validation was conducted by comparing cluster assignments with students' actual exam scores and final grades, supporting the practical relevance of the identified clusters. To further examine the characteristics of each cluster, we visualized the average values of selected behavioral features using a heatmap

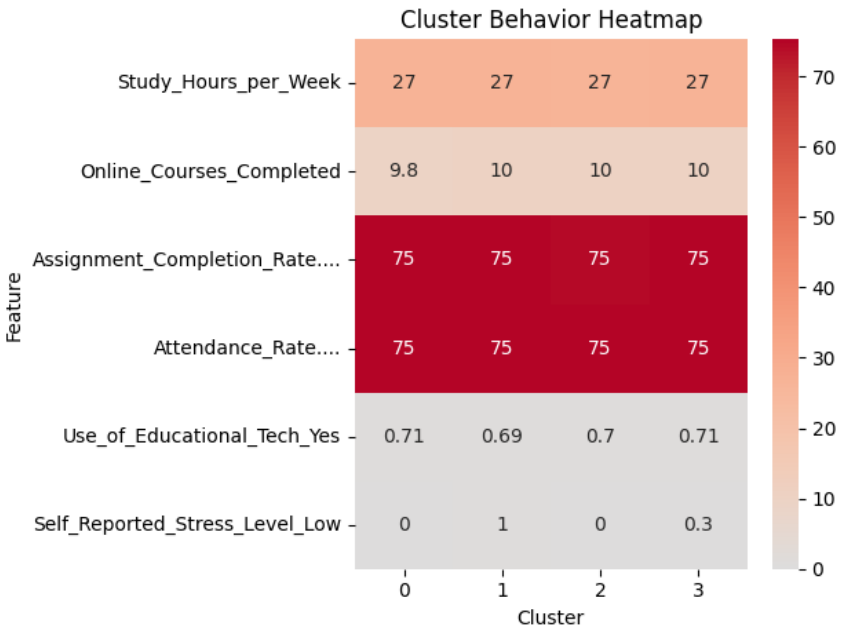

*Figure **6.** Heatmap of average behavioral features across k-means clusters.*

This heatmap reveals how clusters differ in terms of behavior. While most clusters show similar averages for study hours and assignment completion, differences in stress levels and tech usage suggest meaningful distinctions in how students manage their academic routines.

3 Results & Discussion

3.1 Task 1

The overall performance of the models can be seen below, in Fig.6.

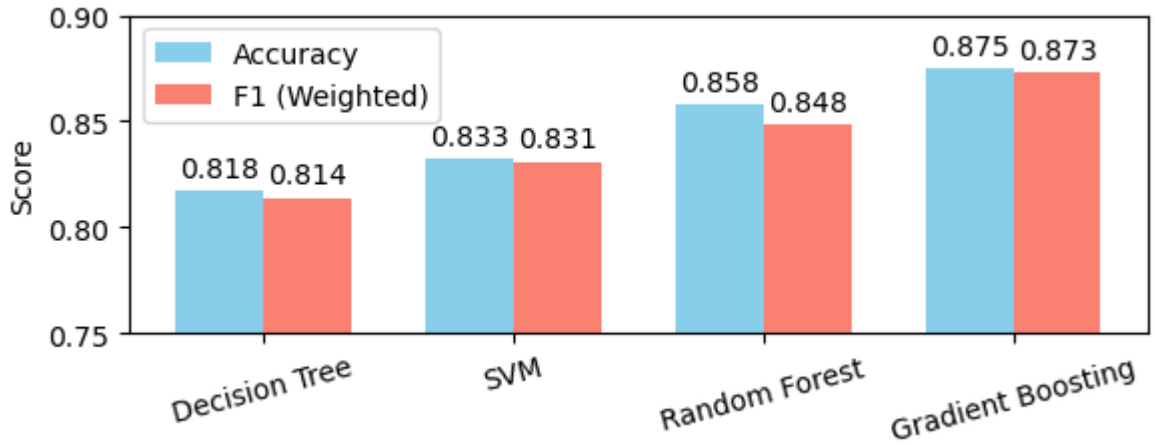

Figure 6 Model performance comparison

As we can see in Fig. 6, gradient boosting has the highest accuracy and F1-score of all the tested models. In Fig.7, below, we can see that a large part of this accuracy comes from the very frequent

“B” grade, with rarer grades like A and D having lower scores. However, note how the predictions are always in the scores adjacent to the true one: an A may be misclassified as B, but not as a C.

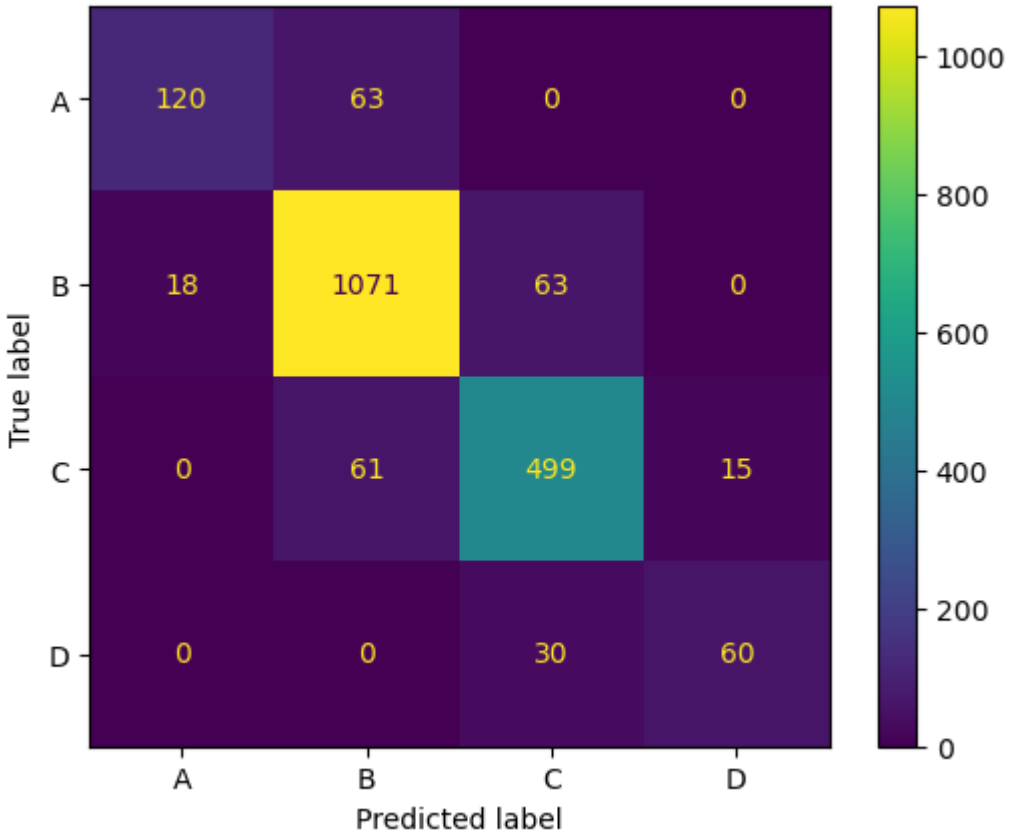

Figure 7 Gradient boosting confusion matrix

The difference in accuracy depending on the grade especially apparent when looking at Table 1, below, where we can see a 0.913 F1-score for A students but only a 0.727 for D students. This is likely because of the dataset already being imbalanced.

Table 1 Classification report of the gradient boosting model

Grade	Precision	Recall	F1-Score	Support
A	0.870	0.656	0.748	183
B	0.896	0.930	0.913	1152
C	0.843	0.868	0.855	575
D	0.800	0.667	0.727	90
Accuracy			0.875	2000
Macro average	0.852	0.780	0.811	2000
Weighted average	0.874	0.875	0.873	2000

3.2 Task 2

3.2.1 Results

1. How many clusters are optimal based on the results?

Using the elbow method (Figure 1), we determined the optimal number of clusters to be **four**. This was based on where the within-cluster sum of squares (WCSS) showed a clear drop-off, indicating that four clusters captured most of the natural structure in the data.

2. Do the grades validate the clusters created?

Yes. Clusters were validated using students' actual exam scores and final grades. The validation showed strong alignment between clusters and grades, confirming that the clusters created accurately reflect real performance differences. The "high achievers" consistently had the highest average exam scores and the greatest percentage of A grades, while "struggling students" showed the lowest scores and higher percentages of D grades. The middle groups "lower steady" and "upper steady" reflected incremental differences in exam scores and grade distributions, accurately capturing subtle performance variations.

Cluster Summary Table
The table below presents key statistics for each cluster, including group size, average exam scores, grade distributions, and assigned performance labels.

Table 2 Summary statistics of hierarchical clusters by average exam score and grade distribution.

size	avg_exam	pct_A	pct_B	pct_C	pct_D	label	
0	4728	73.632026	0.112944	0.559433	0.283841	0.043782	upper_steady
1	406	73.555852	0.088670	0.603448	0.270936	0.036946	lower_steady
2	2437	73.995063	0.098482	0.604842	0.258925	0.037751	high_achievers
3	2429	71.595570	0.043639	0.573899	0.326883	0.055578	struggling_students

These results reinforce the validity of the clustering, highlighting clear performance trends between groups such as high achievers and struggling students.

3. Do k-means and hierarchical clustering provide the same results?

Not exactly. The Adjusted Rand Index (ARI) of approximately **0.31** indicates **moderate agreement** between the two clustering methods. K-means and hierarchical clustering

produced similar groupings but not identical ones. The overlap was substantial, but each method had slightly different interpretations of the middle-performing students.

Clustering with both methods was confirmed visually in the PCA plots (Figure 3), which show side-by-side comparisons of k-means and hierarchical clustering, highlighting both overlap and differences in cluster separation.

Analyzing feature-level summaries revealed that students in the "high achievers" group generally reported more study hours, higher assignment completion rates, lower stress levels, and higher participation rates. Conversely, "struggling students" reported fewer study hours, lower assignment completion, higher stress, and less participation. Among the features considered, those most strongly associated with cluster separation were study hours, stress levels, and participation-related behaviors. Other features such as educational tech usage and online course completion showed smaller differences across clusters, suggesting limited impact on the final groupings.

3.2.2 Discussion

Our analysis confirmed meaningful and practical groupings among students, effectively answering the assignment questions. K-means clustering was clearer in separating struggling students from average performers, while hierarchical clustering highlighted subtle differences, especially within the high-performing groups.

The identified clusters have direct implications for educational intervention. Students labeled as struggling could benefit from targeted academic support, increased monitoring, and stress management programs. Meanwhile, high achievers may be better served by enrichment activities or leadership opportunities.

Overall, combining both clustering methods provided a comprehensive understanding of student performance. Future analyses could further explore additional methods or longitudinal data to enhance these insights and better support diverse student needs.

Limitations and Improvements

One limitation is that the clustering was based only on the available behavioral and demographic features. Some features, such as assignment completion or attendance, showed little variation across clusters, which may have limited their contribution. Including additional data such as motivation levels, time spent on specific subjects, or emotional well being could improve differentiation between groups.

Another limitation is the use of a fixed number of clusters (k=4) based on the elbow method. While reasonable, this choice may not capture all underlying subgroup patterns. Exploring a range of k values or applying other clustering evaluation metrics could help confirm the robustness of the chosen solution.

Finally, the clustering was static and a snapshot at one point in time. In future work, applying clustering over multiple time periods could reveal how students' behaviors and performance change, allowing more dynamic and personalized interventions.

4 References

Shamim, A. (2025). Student Performance & Learning Style. *Kaggle.* https://www.kaggle.com/datasets/adilshamim8/student-performance-and-learning-style