QUESTION 1

1.Introduction

A student's academic success involves more than just intelligence, as it is also shaped by study habits, behavior, learning style, and daily routine. Factors like how much time they dedicate to studying, how actively they participate in class, the quality of their sleep, their stress levels, and whether they use learning technologies all contribute to their performance. By closely examining these aspects, we can better understand what supports student success and how to provide it effectively.

In this project, we used data from 10,000 students. The dataset includes many details like their age, gender, how they like to learn, how active they are in class, and their personal habits. Our main goal is to predict what final grade each student will receive by applying machine learning models.

Since the final grade falls into categories (such as A, B, C, etc.), we treated it as a classification problem.We chose to predict the final grade and removed the exam score so the models don't get an unfair advantage from a direct exam result.

We used and compared the following machine learning models:

- A Decision Tree, which we trimmed to avoid overfitting
- Support Vector Machine (SVM), with tuned settings for better accuracy
- Random Forest, which combines many decision trees to make better predictions
- Gradient Boosting, which learns from mistakes to improve performance

We evaluated these models using accuracy scores, cross-validation, confusion matrices, ROC curves, and feature importance charts. The goal was not only to find the most accurate model, but also to understand which factors have the biggest influence on a student's final grade.

2 Methodology

To build accurate and meaningful predictions, we followed a clear step-by-step approach: first understanding and cleaning the data, then developing and testing several machine learning models.

2.1 Data Description

The dataset we used contains information from 10,000 students. It covers different aspects of a student's life:

- **Personal details** : Age and Gender.
- **Academic behavior**: How many hours they study, their preferred learning style, participation in discussions, attendance rate, and how often they complete assignments.
- **Lifestyle habits** : Including sleep hours, stress level, time spent on social media, and whether they use educational apps or technology.
- **Target variable** : The student's final grade (like A, B, or C)

To keep the predictions fair and meaningful, we removed columns like Exam Score and Student ID, which could unfairly influence the outcome or were not useful for training.

2.2 Data Preprocessing

Before we could train any models, we cleaned and prepared the data. The steps involved:

- **Column name cleaning:** Fixed column names by removing special characters and formatting them consistently for easier manipulation.
- **Label Encoding:** Converted the letter grades (A, B, C, etc.) into numeric values using label encoding so the models could process them effectively.

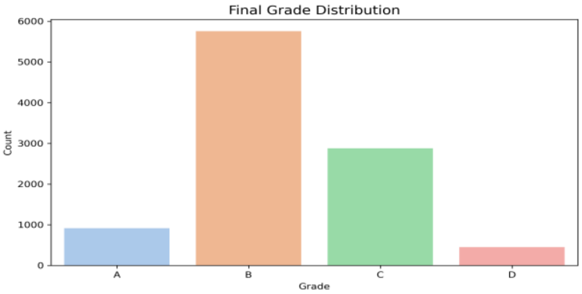


Fig 1.Final Grade Distribution

- **Column name cleaning:** Fixed column names by removing special characters and formatting them consistently for easier manipulation.
- **Label Encoding:** Converted the letter grades (A, B, C, etc.) into numeric values using label encoding so the models could process them effectively.

- **Handling Categorical Data:** For tree-based models (e.g., Decision Tree and Random Forest),we used label encoding for categorical variables.
  For models requiring purely numeric input, like SVM, we applied one-hot encoding to the categorical data (e.g., learning style and gender).
- **Data splitting:** The data was split into two parts — 70% for training the models and 30% for testing. We made sure that each grade category was equally represented in both sets to maintain balance.

2.3 Model Development

We trained four different models to classify student grades and tested how well they performed. The models were selected for their interpretability, robustness, and performance across various types of data:

- **Decision Tree**: A simple model that splits data at decision points. We pruned the tree to avoid overfitting and selected the best structure using cross-validation.
- **Support Vector Machine (SVM)**: A powerful model for separating data into categories. We tested different kernel types (e.g., linear and RBF) and fine-tuned the model using grid search. One-hot encoding helped handle non-numeric data.
- **Random Forest**: An ensemble method that builds multiple decision trees and averages their outputs. We adjusted parameters like the number of trees (n estimators) and tree depth (max depth) for better performance.
- **Gradient Boosting**: A sequential learning method that corrects the errors of the previous models. We tuned the learning rate, number of estimators, and tree depth to find the best-performing configuration.

To ensure the models would perform well in various situations, we used stratified k-fold cross-validation. This technique evaluates model performance across different subsets of the data, ensuring the results are reliable and generalizable.

3 Evaluation and Results

To evaluate the models’ predictive capabilities, we applied a combination of classification metrics and interpretive visualizations. This section presents the quantitative results and graphical analysis for each model.

3.1 Evaluation Metrics

We used the following standard metrics:

- **Accuracy**: Proportion of correct predictions:

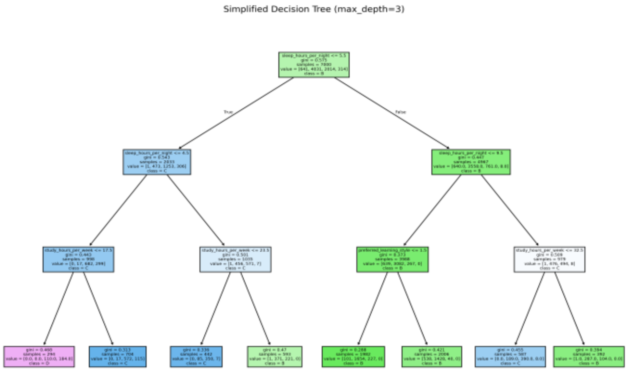
$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

- **Confusion Matrix**: Visual comparison between predicted and actual grades. Useful for spotting class-wise errors.
- **ROC Curve & AUC (SVM only)**: Evaluates binary classification performance by plotting true positive vs. false positive rates. AUC summarizes this tradeoff into a single score.

3.2 Model Performance and Visual Analysis

3.2.1 Decision Tree Classifier

- Pruned Tree Diagram(fig 2) shows how decisions are made based on input features. The top-level splits highlight which factors are most influential (e.g., assignment completion rate, sleep hours).
- Confusion Matrix(fig 3) shows that most predictions align correctly, especially for extreme grades (A, F), with some misclassification in the mid-range grades.



Simplified Decision Tree (max_depth=3)
sleep_hours_per_night <= 5.5
True
False
sleep_hours_per_night <= 4.5
sleep_hours_per_night <= 9.5
study_hours_per_week <= 17.5
study_hours_per_week <= 23.5
preferred_learning_style <= 1.5
study_hours_per_week <= 32.5
class = D
class = C
class = B

Fig 2.Pruned Decision Tree

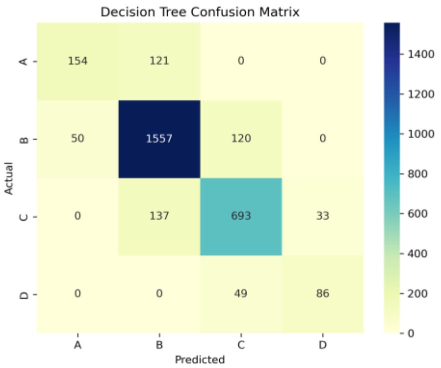


Fig 3.Confusion Matrix DT

3.2.2 Support Vector Machine (SVM)

SVM Confusion Matrix (fig 4) reveals solid performance, with minor confusion between adjacent grade categories (like B and C).

ROC Curve (in a binary scenario)(Fig 5) shows strong separation ability with a high AUC.

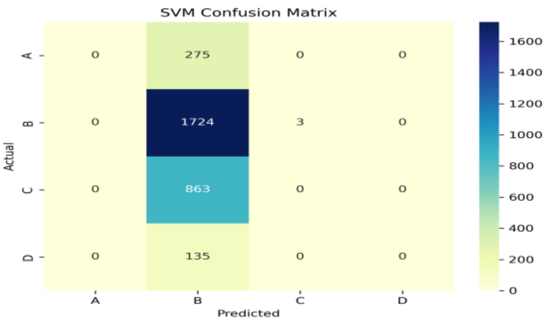


Fig 4.Confusion Matrix of Decision Tree

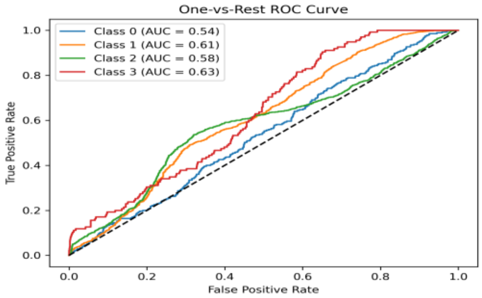


Fig 5.ROC Curve

3.2.3 Random Forest.

Confusion Matrix demonstrates consistent classification across all grade categories, including balanced performance on overlapping classes.

Feature Importance Plot identifies key features like sleep hours per night, assignment completion, and study hours per week.

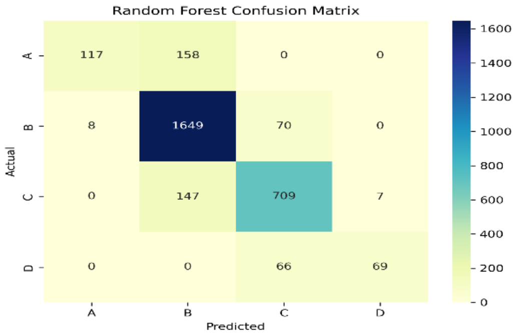


Fig 6.Confusion Matrix for Randon Forest

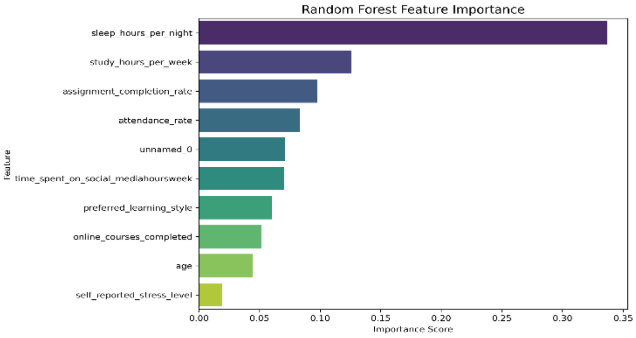


Fig 7.Random Forest Feature Importance

3.2.4 Gradient Boosting

Gradient Boosting Confusion Matrix indicates slightly better mid-grade classification compared to other models.

Feature Importance of Gradient Boosting again highlights similar top contributors as in Random Forest, validating consistency.

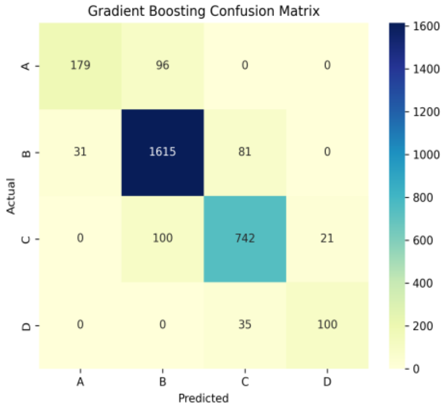


Fig 8.Confusion Matrix for Gradient Boosting

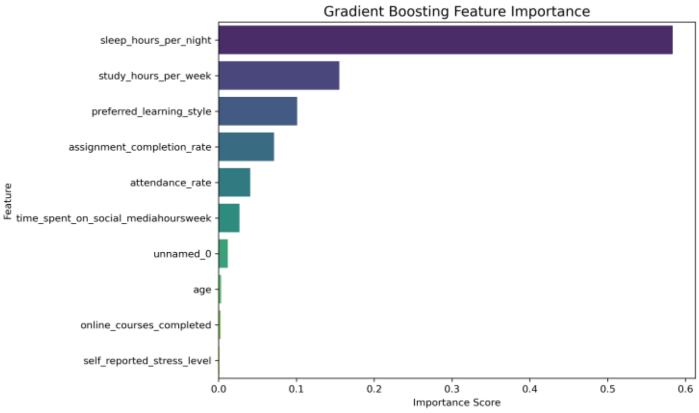


Fig 9.Gradient Boosting Feature Importance

3.3 Overall Accuracy Comparison

Each model's accuracy on the test dataset was computed and visualized.

The table above summarizes the test accuracy achieved by each classification model. Among the four models, **Gradient Boosting** delivered the best performance with an accuracy of **80.9%**, followed closely by **Random Forest** at **80.4%**. The **SVM** model also performed reasonably well, achieving **78.3%** accuracy. While the **Decision Tree** was the least accurate at **75.2%**, it remains valuable for its simplicity and interpretability.

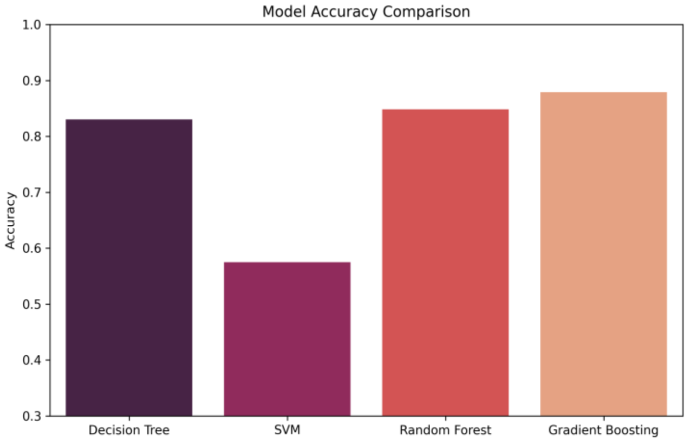


Fig 10.Bar Chart Showing Accuracy of All Four Models

Model	**Accuracy**
Decision Tree	0.83
SVM	0.574
Random Forest	0.848
Gradient Boosting	0.878

Table 1:-Test Accuracy Scores of Classification Models

The bar chart below (Figure: accuracy_comparison.png) provides a visual comparison of these results, making it easier to identify which models performed best overall.

4. Conclusion

In this project, we set out to predict students' final academic grades using machine learning models. We used a dataset of 10,000 students that included not just academic records but also details about their daily habits, stress levels, and how they learn. Our goal wasn't just to build accurate models, we also wanted to understand what factors truly influence a student's success.

We tried four different models, Decision Tree, Support Vector Machine (SVM), Random Forest, and Gradient Boosting. Each model was trained, tested, and compared fairly. Among all, Gradient Boosting performed the best, with the highest accuracy of 80.9%, followed closely by Random Forest (80.4%) and SVM (78.3%). The Decision Tree, while easier to understand, had the lowest accuracy at 75.2%.

What we found interesting was that features like how regularly students complete assignments, how much they study, how many hours they sleep, and how stressed they feel played a big role in predicting their final grades. This shows that student success isn't just about studying hard good sleep and managing stress are just as important.

The results also showed that advanced models like Gradient Boosting can do a great job at predicting performance while still offering useful insights. These models could be used by schools and universities to identify students who might need extra help early on before it's too late.

Overall, this project proves that by combining machine learning with education, we can build smart tools that help both teachers and students in meaningful ways.

QUESTION 2

1.Introduction

This project applies unsupervised machine learning techniques namely k-means and hierarchical clustering to uncover behavioral groupings among students based on their academic habits, learning preferences, and personal characteristics. The dataset contains 10,000 students, each described by variables such as study hours, learning style, attendance, participation, and technology use. Importantly, the clustering is performed without using the response variable (i.e., Final Grade), in line with the principles of unsupervised learning.

After constructing the clusters, we validate the outcomes by comparing them to students' actual final grades. This evaluation enables us to determine whether the discovered behavioral patterns align with academic performance. We also compare the effectiveness of k-means and hierarchical clustering in terms of structure, interpretability, and cluster validity.

2. Methodology

2.1 Data Preparation and Preprocessing

The original dataset contained 10,000 student records with 16 variables, including demographic, behavioral, and academic attributes. To prepare the data for clustering analysis, the following preprocessing steps were undertaken:

- **Removing Non-Informative Columns:**Columns such as Student_ID, X (row index), and Final_Grade were removed from the clustering dataset to prevent bias, as clustering must be performed without the response variable.
- **Handling Categorical Variables:** Categorical variables_Gender, Preferred Learning Style, Participation in Discussions, Use of Educational Tech, and Self Reported Stress Level, were converted into factors types and then transformed into dummy variables using the caret:: DummyVars()function.
- **Feature Scaling:**All numeric and dummy-encoded variables were standardized using z-score normalization (scale() function in R). This ensured that features with larger ranges (e.g., Study_Hours_per_Week, Time_Spent_on_Social_Media) did not dominate the clustering process.

As a result, the cleaned and processed dataset was ready for unsupervised learning. The transformed dataset had 22 scaled numerical variables for each student, which became the input to the clustering algorithms.

2.2. Determining the Optimal Number of Clusters

Before applying k-means or hierarchical clustering, it is essential to identify the most appropriate number of clusters (k) to extract meaningful groupings. Two well-established techniques were used to determine the optimal value of k:

- **Elbow Method**:
  The total within-cluster sum of squares (WSS) was plotted for different values of k (from 1 to 10). The resulting elbow plot showed a clear bend at *k = 2*, suggesting that adding more clusters beyond this point results in diminishing improvements in compactness.

- **Silhouette Method**
  This technique evaluates how well each data point fits within its assigned cluster. The average silhouette width was highest at *k = 2*, indicating well-separated and cohesive clusters.

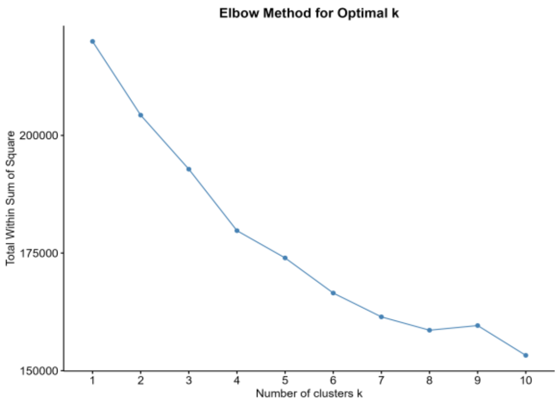


Fig 1:*Elbow Method for Determining Optimal Number of Clusters*

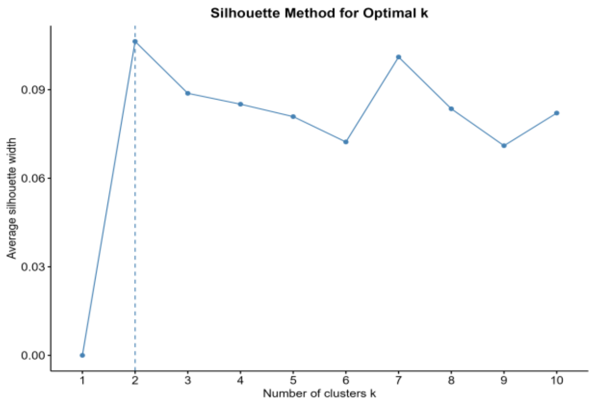


Fig 2: Silhouette Method for Evaluating Cluster Cohesion

Both methods independently supported the selection of **k = 2** as the optimal number of clusters for this dataset.

2.3. Clustering with K-means and Hierarchical Methods

To uncover patterns in student behavior, we applied two clustering algorithms: **k-means clustering** and **hierarchical clustering**. Both were performed using the preprocessed dataset with 22 scaled variables. Since both the Elbow and Silhouette methods had suggested *k = 2*, we used two clusters for interpretation and evaluation.

2.3.1 K-means Clustering

K-means clustering partitions observations into *k* groups by minimizing the within-cluster sum of squares. Using the kmeans() function in R with *k = 2*, students were grouped into two distinct behavioral clusters. The algorithm was initialized 25 times with different random starting points to ensure a stable solution.

To better understand the characteristics of each cluster, the average (mean) value of each input feature was computed per cluster. This **cluster profiling** allowed us to interpret the traits that differentiate the two groups—for instance, differences in average study hours, participation in discussions, or technology use.

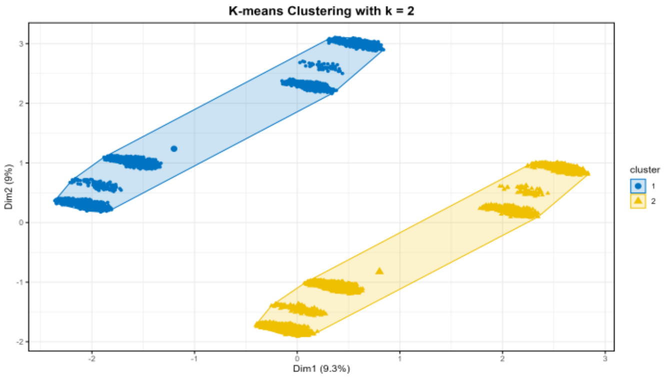


Figure 3: K-means Clustering Visualization with k = 2

A visual representation of the clusters was created using principal components. The **K-means cluster plot** (Figure 3) reveals the compactness and separation of the two student groups.

2.3.2 Hierarchical Clustering

Hierarchical clustering was also applied using the **Ward.D2 linkage method** and **Euclidean distance**. Unlike k-means, which requires specifying k in advance, hierarchical clustering builds a nested structure (dendrogram) of groupings. We cut the tree at $k = 2$ to allow comparison with the k-means result.

To handle the large dataset size (10,000 students), two dendrograms were created:

- One for the **entire dataset**, showing the global structure.
- One for the **first 100 students**, allowing for a clearer visualization.

The full dendrogram (not shown here) was also created for all 10,000 students and showed a similar global structure, but was too dense to interpret visually.

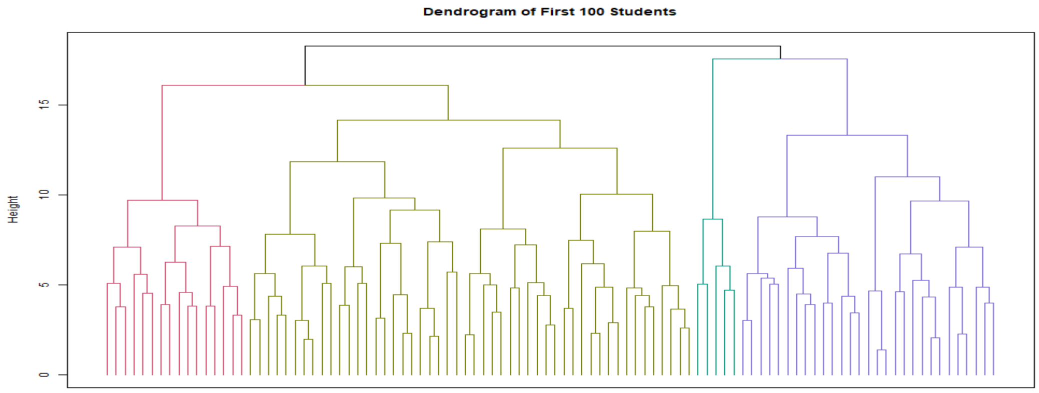


Fig4: Dendrogram for First 100 Students

2.3.3 Comparison of Clustering Approaches

While k-means is more computationally efficient and better suited for large datasets, hierarchical clustering provides a more interpretable tree-based structure. Using both techniques allowed for robust cross-validation of cluster stability. Both methods were set to identify 2 clusters, and their results were later compared with students' final grades to assess alignment with academic outcomes.

2.4. Cluster Profiles and Grade Distributions

To interpret the behavioural meaning of each cluster, we computed the average values of features for the two k-means clusters (*cluster profiling*). The results indicated that:

- **Cluster 1** students tended to be less engaged in discussions.
- **Cluster 2** showed higher participation, different stress profiles, and slightly different study patterns.

To evaluate whether the discovered behavioural groupings aligned with academic performance, we compared the final grade distribution across the clusters. **Tables 1 and 2** show the percentage of students in each grade category (A, B, C, D) for both k-means and hierarchical clustering methods.

2.4.1.Grade Alignment with Clusters: This table displays the number and percentage of students in each cluster based on their final grades. The distributions are nearly identical, indicating that behavioural clustering alone may not capture substantial differences in academic outcomes.

Cluster	A (n/%)	B (n/%)	C (n/%)	D (n/%)	Total Students
1	375 (9.4%)	2263 (56.5%)	1183 (29.5%)	183 (4.6%)	4004
2	541 (9.0%)	3495 (58.3%)	1694 (28.3%)	266 (4.4%)	5996

Table 1: Grade Distribution in K-means Clusters (k = 2)

2.4.2.Grade Alignment with Hierarchical clusters: The table above shows the number and percentage of students receiving each grade (A–D) within the two clusters obtained using hierarchical clustering. The distributions are highly similar, suggesting weak separation by academic outcome.

Cluster	A (n/%)	B (n/%)	C (n/%)	D (n/%)	Total Students
1	655 (9.2%)	4110 (57.6%)	2041 (28.6%)	330 (4.6%)	7136
2	261 (9.1%)	1648 (57.5%)	836 (29.2%)	119 (4.2%)	2864

Table 2: Grade Distribution in K-means Clusters (k = 2)

2.5. Exporting Clustered Dataset

After assigning each student to a cluster using k-means (*k = 2*), the full dataset with cluster labels was saved as a CSV file (student_clusters_k2.csv). This enabled easy reuse of the results for additional analysis or reporting.

3. Results and Interpretation

3.1 Summary of Cluster Traits

The clustering analysis successfully grouped students into two distinct behavioural profiles. Cluster profiling based on the standardized feature means (z-scores) revealed interpretable patterns:

3.1.2 Cluster 1 (K-means & Hierarchical):

This group was characterized by lower participation in discussions, moderate use of educational technology, and slightly lower average exam scores. Their self-reported stress levels leaned more toward medium or low, and they generally spent more time on social media.

3.1.2. Cluster 2:

Students in this cluster showed higher engagement in discussions, greater use of educational tools, and more consistent assignment completion. They also reported slightly higher stress and study hours, suggesting stronger academic discipline.

These behavioural patterns indicate that participation, tech engagement, and study habits were the most differentiating traits between the two groups. However, the overall variation between clusters was moderate, highlighting the subtlety of behavioural diversity in this student population.

Differences represent the contrast in student habits and engagement across clusters. While Table 3 provides a numerical comparison, the following radar chart visualizes the overall shape and separation of behavioural traits across clusters.

Feature	Cluster 1 (Mean)	Cluster 2 Mean)	)
Participation_in_ Discussions.No	1.22	-0.82	2.04
Participation_in_ Discussions.Yes	-1.22	0.82	2.04
Preferred_Learning_ Style.Kinesthetic	-0.02	0.02	0.04
Use_of_Educational_ Tech.Yes	0.02	-0.01	0.03
Self_Reported_Stress_ Level.Low	-0.01	0.01	0.02
Online_Courses_ Completed	0.01	-0.01	0.02
Study_Hours_ per_Week	-0.01	0.01	0.02

Table 3 : Standardized mean values (z-scores) of selected behavioral features in Cluster 1 and Cluster 2.

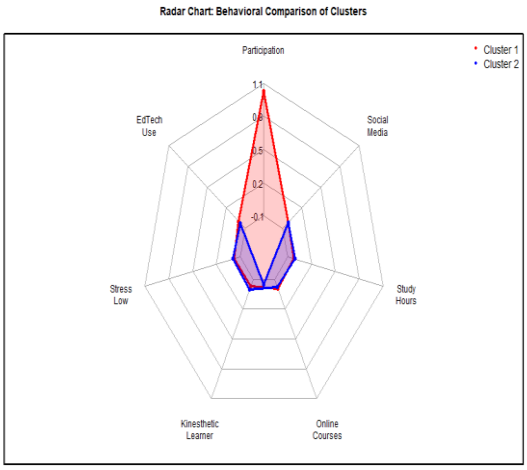


Fig 5: Radar chart comparing behavioral traits between Cluster 1 and Cluster 2.

Figure 5. Radar chart showing standardized behavioural feature averages for Cluster 1 and Cluster 2. The most notable difference lies in discussion participation, while other features vary subtly across clusters.

3.2. Alignment with Final Grades

To evaluate whether the clusters identified by the algorithms corresponded to actual academic performance, we compared the distribution of **Final Grades (A–D)** across the two clusters formed by both **k-means** and **hierarchical clustering**.

Despite clear behavioural groupings, the grade distributions across clusters were **remarkably similar**:

- In K-means clustering, both clusters had nearly identical proportions of A and B students (around 9–10% A, 56–58% B), with only minor shifts in the percentages of C and D grades.
- The hierarchical clusters followed an almost identical pattern, reinforcing the observation that the clusters did not strongly differentiate academic performance.
- This suggests that while students differ in engagement**,** learning style, and technology use, these differences may not be strong predictors of final grades or that other unmeasured factors (such as teaching quality or prior academic ability) play a more dominant role.
- Overall, the grade alignment analysis indicates that behavioural clustering offers limited predictive value for academic outcomes in this dataset, though it may still be useful for identifying distinct student engagement profiles for targeted interventions.

3.3. K-means vs Hierarchical: Key Differences

- Both k-means and hierarchical clustering produced consistent groupings, with $k = 2$ emerging as the optimal number of clusters in both methods. However, each algorithm offered unique advantages and limitations:

- **K-means Clustering**
  - Efficiency: Very fast and scalable, making it ideal for large datasets like this one (10,000 students).
  - Interpretability: Produced clear, compact clusters, especially when $k = 2$.
  - Limitations: Requires pre-specifying the number of clusters and is sensitive to initial starting points.

- **Hierarchical Clustering**
  - Visual Insight: The dendrogram provided a useful visualization of student similarity at various linkage distances.
  - No Need to Predefine k: Allows flexibility in cutting the tree at different levels.
  - Limitations: Computationally heavier and less practical for large-scale datasets.

Despite methodological differences, both approaches yielded clusters with similar grade distributions, confirming the stability of the segmentation. However, the absence of strong academic separation across clusters highlights the complexity of learning outcomes, which may not be fully captured by behavioural data alone.

4. Conclusion

This study applied unsupervised clustering techniques specifically k-means and hierarchical clustering to analyse behavioural patterns among 10,000 students using 22 preprocessed features. The primary goal was to segment students based on their academic habits, stress levels, technology use, and learning preferences, without using final grades during clustering.

Using the Elbow and Silhouette methods, $k = 2$ emerged as the optimal number of clusters. Both k-means and hierarchical clustering revealed consistent cluster structures, indicating robust behavioural groupings. Cluster profiling showed meaningful differences such as higher participation and educational tech usage in one group, and lower discussion involvement and higher stress in the other.

When clusters were compared to students' actual final grades, the distributions were remarkably similar across clusters, with slight variations in the proportion of grades A–D. This suggests that while behavioural data can segment students meaningfully, academic outcomes are influenced by additional unobserved factors, and are not strictly predictable through clustering alone.

From a methodological standpoint, k-means was faster and more scalable, whereas hierarchical clustering offered better visual understanding through dendrograms. Overall, the findings affirm that unsupervised learning can uncover useful behavioural insights, though further supervised modelling would be needed to predict academic performance more precisely.

References:

Question 1: Predicting Student Performance (Supervised Learning)

1. Zhou, Z.-H. (2012). Ensemble Methods: Foundations and Algorithms. CRC Press. [Access Link]

2. Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273–297. [Access Link]

Question 2: Clustering Student Behavior (Unsupervised Learning)

1. Kaufman, L., & Rousseeuw, P. J. (2009). Finding Groups in Data: An Introduction to Cluster Analysis. Wiley. [Access Link]

2. MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. [Access Link]