جامعة الملك سعود
King Saud University

Students Performance predicting

Supervised by:
Dr.Ameera Almasoud

Table of contents

1-
INTRODUCTION

Problem:

Our project focuses on analyzing the Students Performance Dataset, aiming to classify students into performance categories (A, B, C, D, F) and cluster them based on similar characteristics. By applying data mining techniques, we aim to uncover key factors influencing academic success or challenges.

Goal:

The goal is to identify patterns and predictors that distinguish students with high, medium, and low performance. These insights will help educators design strategies to improve learning outcomes and promote equity in education.

2-
DATASET

DATASET

General Information about the dataset

Number of objects in original dataset: 1025

Number of attributes: 15

Class labels:

Performance (A, B, C, D, F)

Missing values: there is no missing values

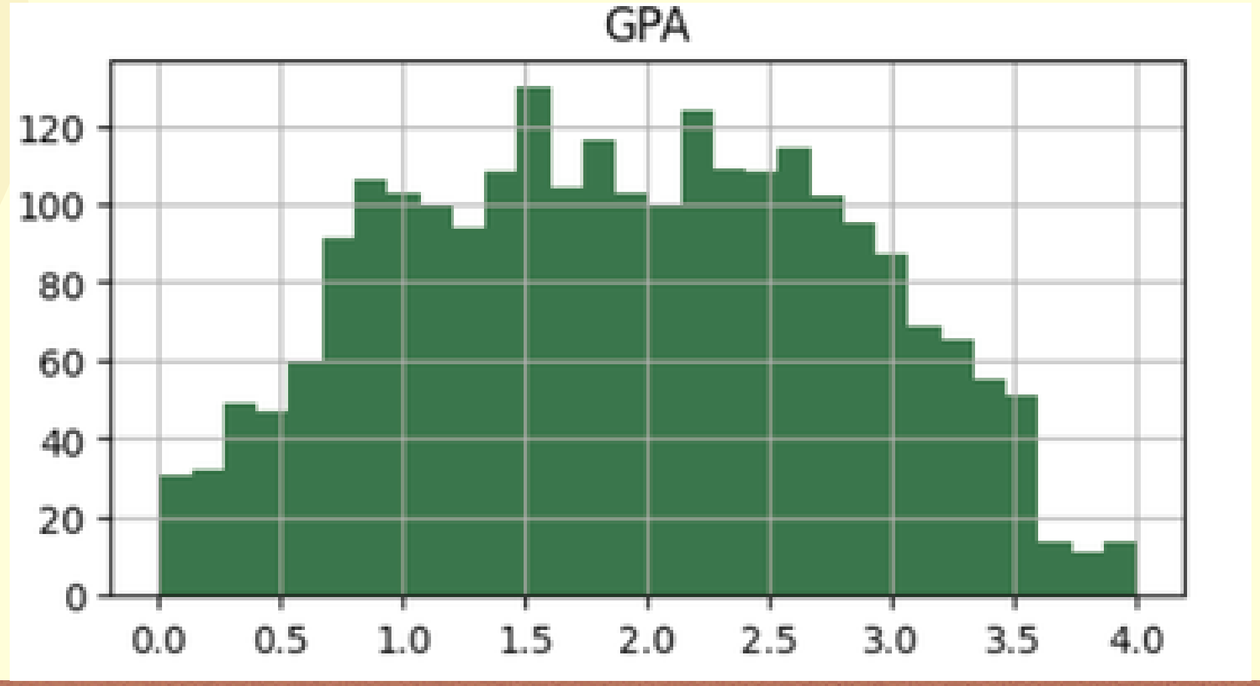
	Missing Values
StudentID	0
Age	0
Gender	0
Ethnicity	0
ParentalEducation	0
StudyTimeWeekly	0
Absences	0
Tutoring	0
ParentalSupport	0
Extracurricular	0
Sports	0
Music	0
Volunteering	0
GPA	0
GradeClass	0

DATASET

General Information about the dataset

Histograms for GPA, Absences, and StudyTimeWeekly

- GPA: Mostly between 1.0 and 3.0, indicating moderate performance.

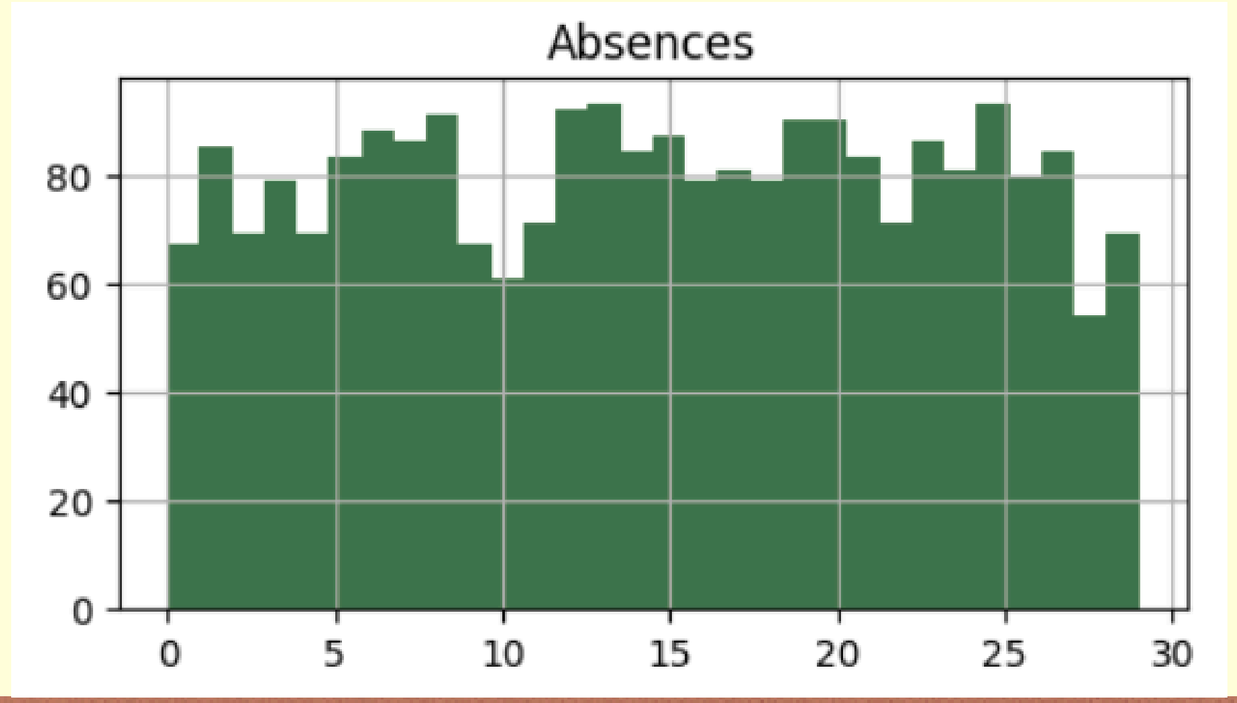


DATASET

General Information about the dataset

Histograms for GPA, Absences, and StudyTimeWeekly

- **Absences: Ranges from 0 to 29, with some higher values suggesting irregular attendance.**

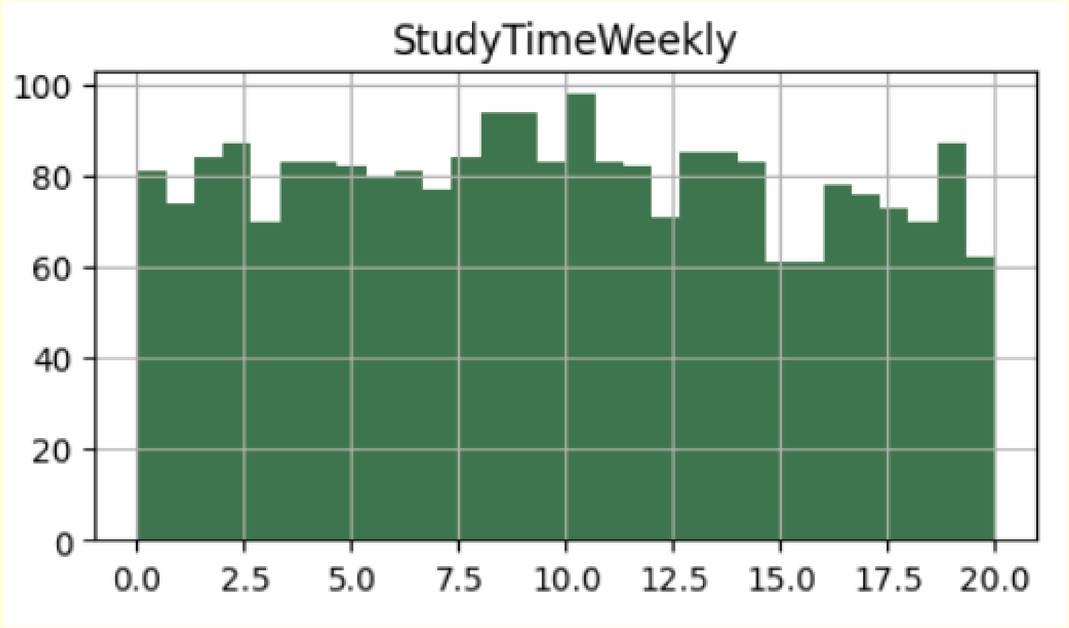


DATASET

General Information about the dataset

Histograms for GPA, Absences, and StudyTimeWeekly

- **StudyTimeWeekly: Ranges from 0 to 20 hours, with a slight concentration around 10 hours, showing diverse study habits.**

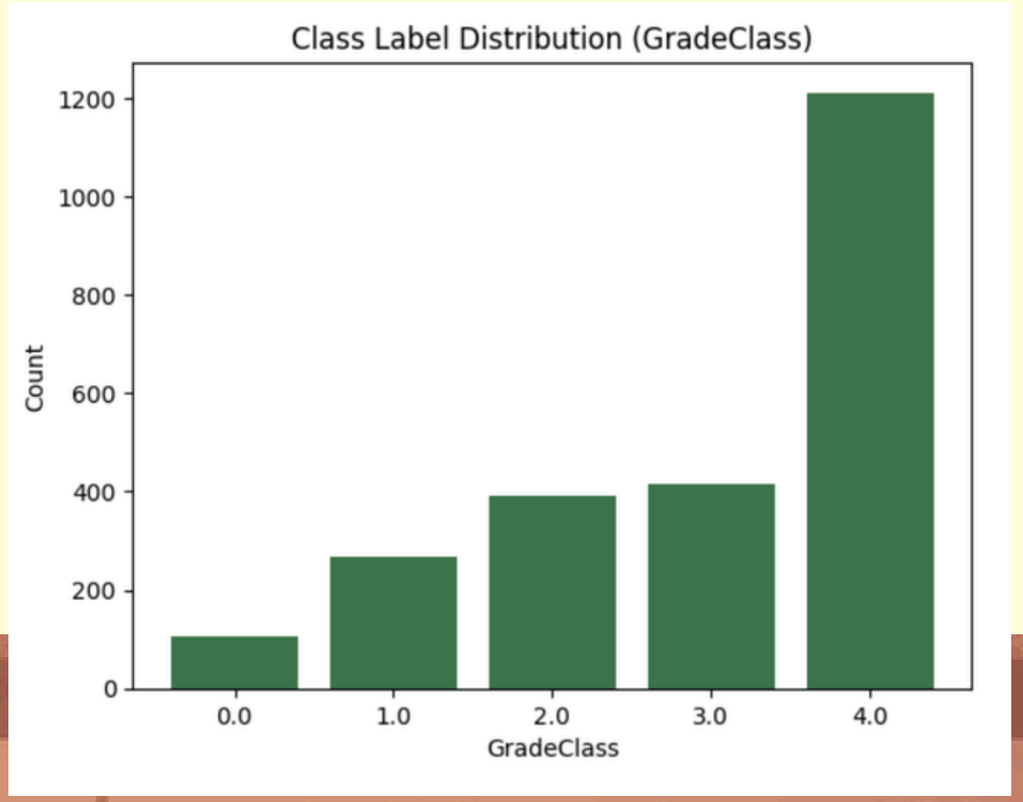


DATASET

General Information about the dataset

Bar Plot for Class Label Distribution:

- The dataset is imbalanced, with most students in GradeClass 4 (high achievers) and very few in GradeClass0.Intermediate categories (1-3) have moderate representation.



3-
CLASSIFICATION

What is Classification

Supervised learning is a technique used to classify data into predefined categories. It plays a crucial role in improving decision-making

In our case, we trained our model to predict students' academic performance based on features such as gender, parental education, study time, exam scores, and extracurricular activities. This classification allows us to categorize students into performance levels (A, B, C, D, F) and identify those at risk of underperforming. By doing so, we can provide targeted interventions and support to improve their academic outcomes.

Evaluation of Classification

Attribute selection measures (splitting criteria):

- **Gini index (criterion="gini")**
- **Information Gain (Entropy) (criterion="entropy")**

Train-test partitions:

We evaluate three train-test splits:

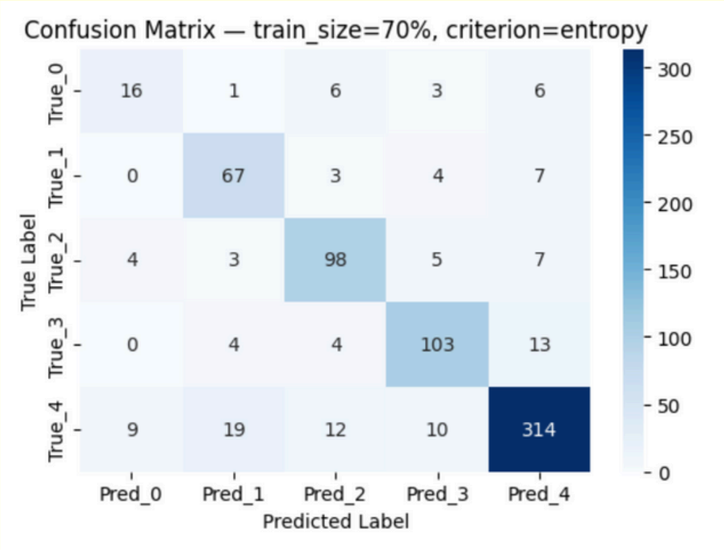
- **60% training – 40% testing**
- **70% training – 30% testing**
- **80% training – 20% testing**

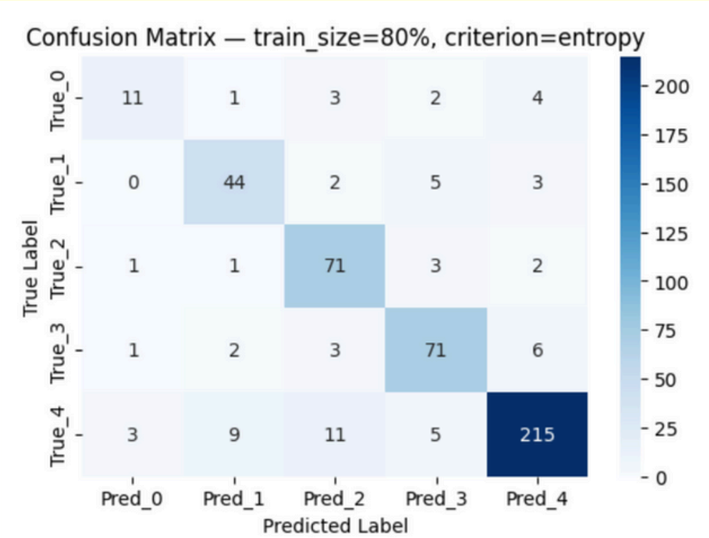
Out[3]:

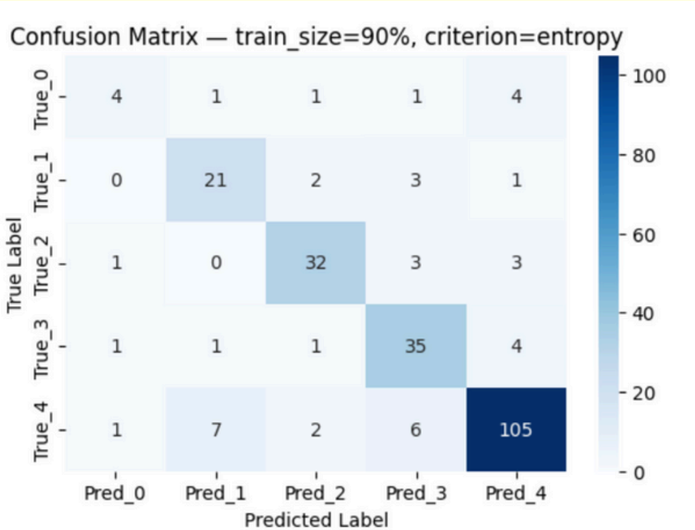
	train_size	criterion	accuracy	precision_macro	recall_macro	f1_macro
0	60%	gini	0.830721	0.759268	0.754334	0.753084
1	60%	entropy	0.840125	0.748562	0.766762	0.756705
2	70%	gini	0.855153	0.772046	0.775849	0.772576
3	70%	entropy	0.832869	0.758027	0.771610	0.763509
4	80%	gini	0.868476	0.808631	0.777610	0.783296
5	80%	entropy	0.860125	0.801737	0.797815	0.796391

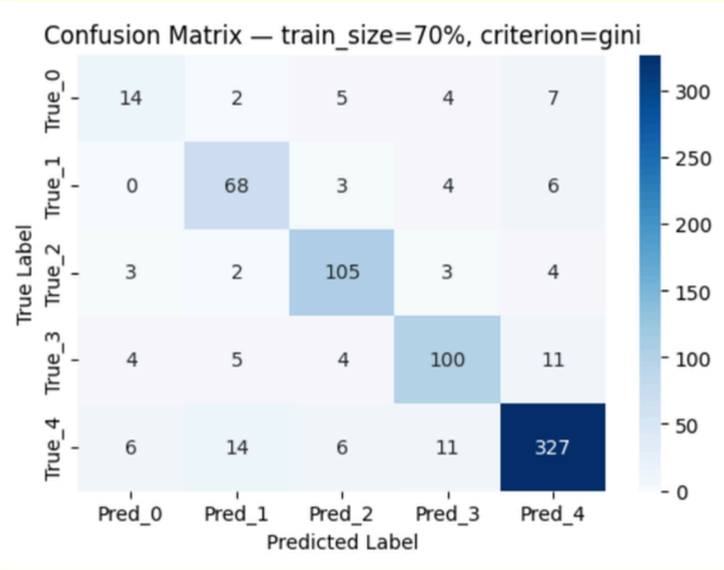
Out[4]:

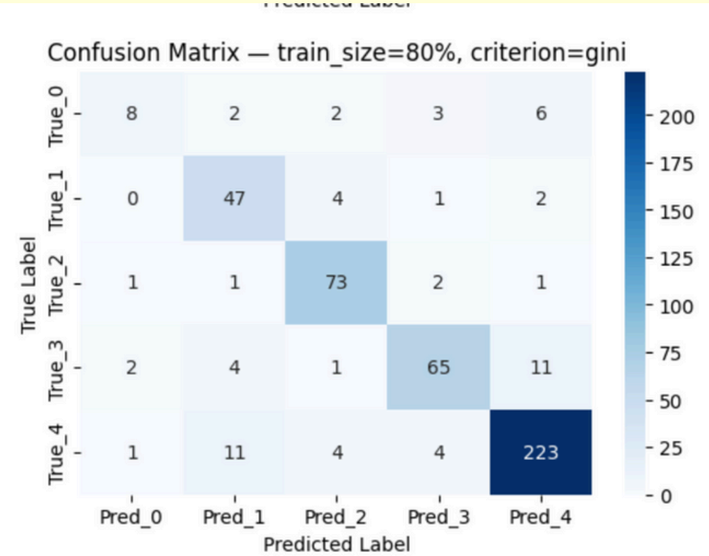
criterion	entropy	gini
train_size		
60%	0.840125	0.830721
70%	0.832869	0.855153
80%	0.860125	0.868476

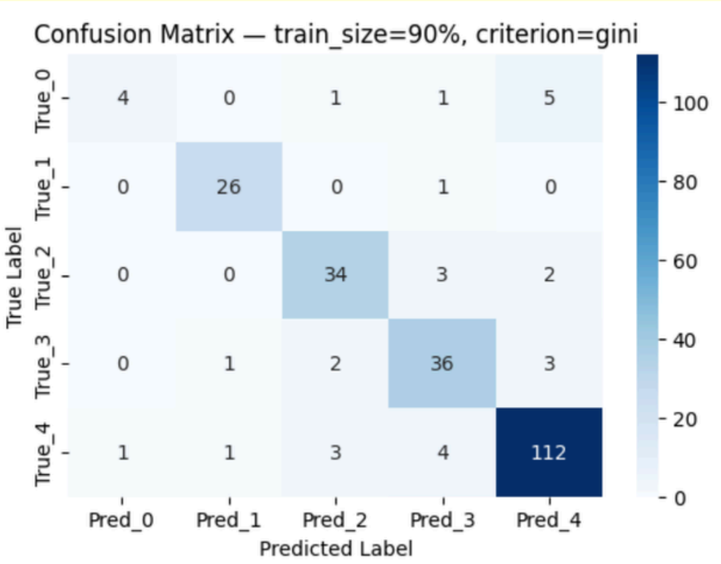

Confusion Matrix — train_size=70%, criterion=entropy
True_0 16 1 6 3 6
True_1 0 67 3 4 7
True_2 4 3 98 5 7
True_3 0 4 4 103 13
True_4 9 19 12 10 314
True Label
Pred_0 Pred_1 Pred_2 Pred_3 Pred_4
Predicted Label
300 250 200 150 100 50 0


Confusion Matrix — train_size=80%, criterion=entropy
True_0 11 1 3 2 4
True_1 0 44 2 5 3
True_2 1 1 71 3 2
True_3 1 2 3 71 6
True_4 3 9 11 5 215
True Label
Pred_0 Pred_1 Pred_2 Pred_3 Pred_4
Predicted Label
200 175 150 125 100 75 50 25 0


Confusion Matrix — train_size=90%, criterion=entropy
True_0 4 1 1 1 4
True_1 0 21 2 3 1
True_2 1 0 32 3 3
True_3 1 1 1 35 4
True_4 1 7 2 6 105
True Label
Pred_0 Pred_1 Pred_2 Pred_3 Pred_4
Predicted Label
100 80 60 40 20 0


Confusion Matrix — train_size=70%, criterion=gini
True_0 14 2 5 4 7
True_1 0 68 3 4 6
True_2 3 2 105 3 4
True_3 4 5 4 100 11
True_4 6 14 6 11 327
True Label
Pred_0 Pred_1 Pred_2 Pred_3 Pred_4
Predicted Label
300 250 200 150 100 50 0


Confusion Matrix — train_size=80%, criterion=gini
True_0 8 2 2 3 6
True_1 0 47 4 1 2
True_2 1 1 73 2 1
True_3 2 4 1 65 11
True_4 1 11 4 4 223
True Label
Pred_0 Pred_1 Pred_2 Pred_3 Pred_4
Predicted Label
200 175 150 125 100 75 50 25 0


Confusion Matrix — train_size=90%, criterion=gini
True_0 4 0 1 1 5
True_1 0 26 0 1 0
True_2 0 0 34 3 2
True_3 0 1 2 36 3
True_4 1 1 3 4 112
True Label
Pred_0 Pred_1 Pred_2 Pred_3 Pred_4
Predicted Label
100 80 60 40 20 0

result

1) Accuracy Comparison

The accuracy results show clear differences across the train–test partitions and attribute selection measures:

- **Best overall accuracy:**
  ⭐ Gini @ 80% training → **0.868476**
- **Second-best:**
  ⭐ Entropy @ 80% training → **0.860125**
- **Lowest accuracy:**
  - Gini @ 60% training → **0.830721**

Conclusions:

- The classifier performs **better with larger training data** (80% > 70% > 60%).
- **Gini consistently outperforms entropy** at 70% and 80% training sizes.
- The performance difference is small, but **Gini is overall the stronger criterion** for this dataset.

4-
CLUSTERING

What is Clustering

Clustering is unsupervised learning; it will group objects in a cluster based on similarity and dissimilarity.

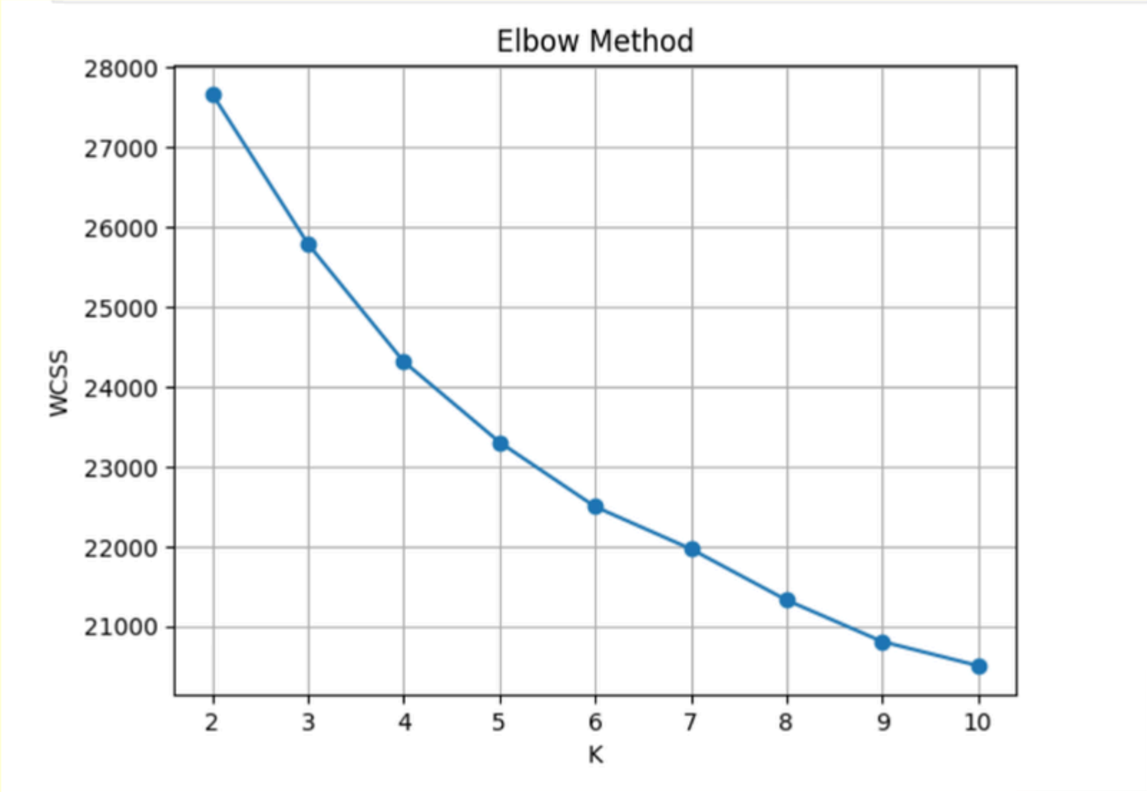
Goal:

Our model uses K-means clustering to group students with similar academic and lifestyle characteristics.

These clusters help reveal hidden patterns in the data and provide valuable insights into factors associated with academic performance and success.

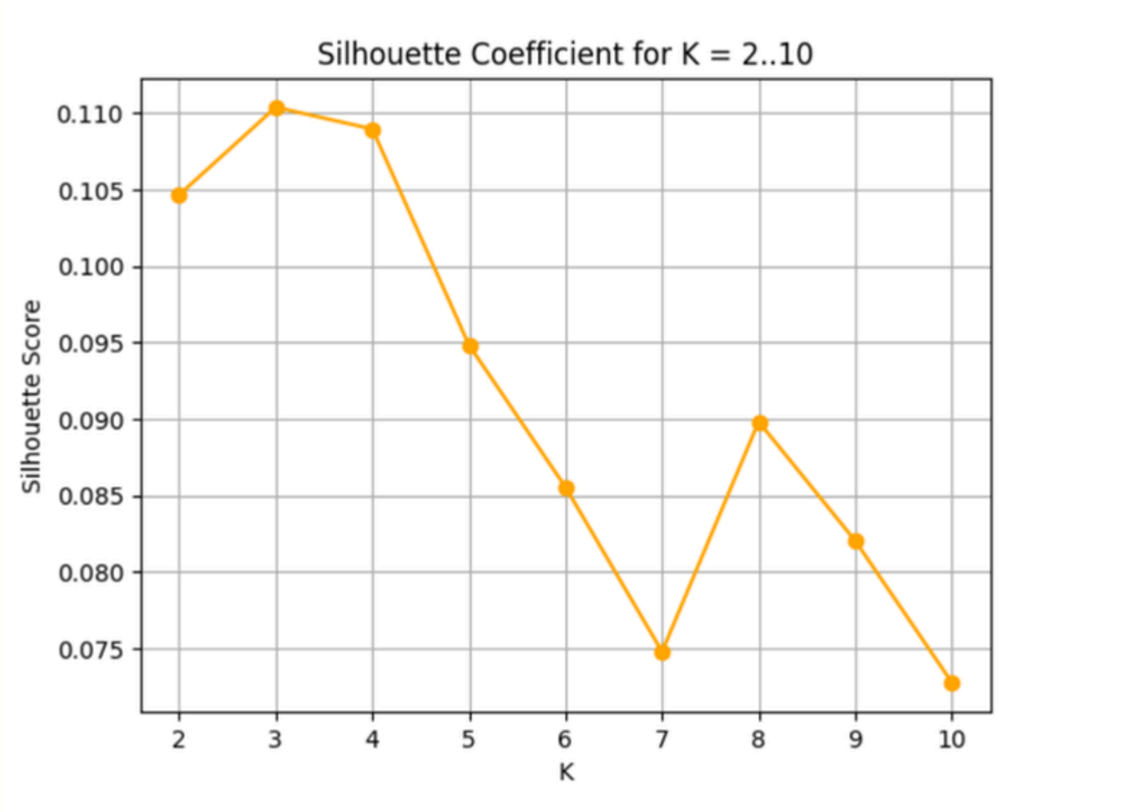
Elbow method Clustering

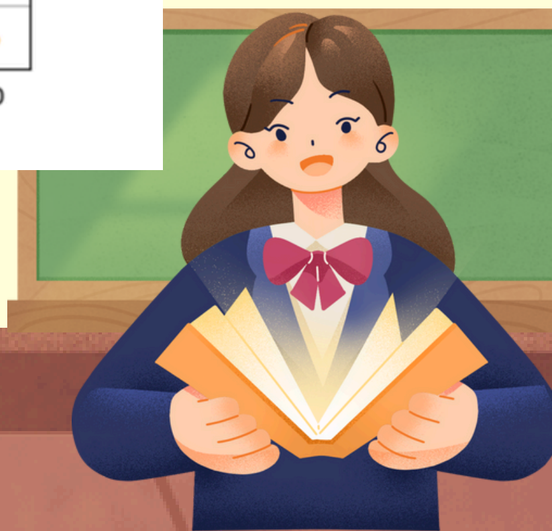
- The curve shows a sharp decrease in WCSS between K = 2 and K = 4.
- After K = 4, the line begins to flatten, meaning additional clusters do not significantly reduce the within-cluster variance.
- This suggests that K = 4 is a reasonable candidate according to the Elbow Method.



Elbow method Clustering

- The highest Silhouette score appears at K = 3, indicating the best separation between clusters at this value.
- K = 4 also shows a relatively high silhouette score but slightly lower than K = 3.





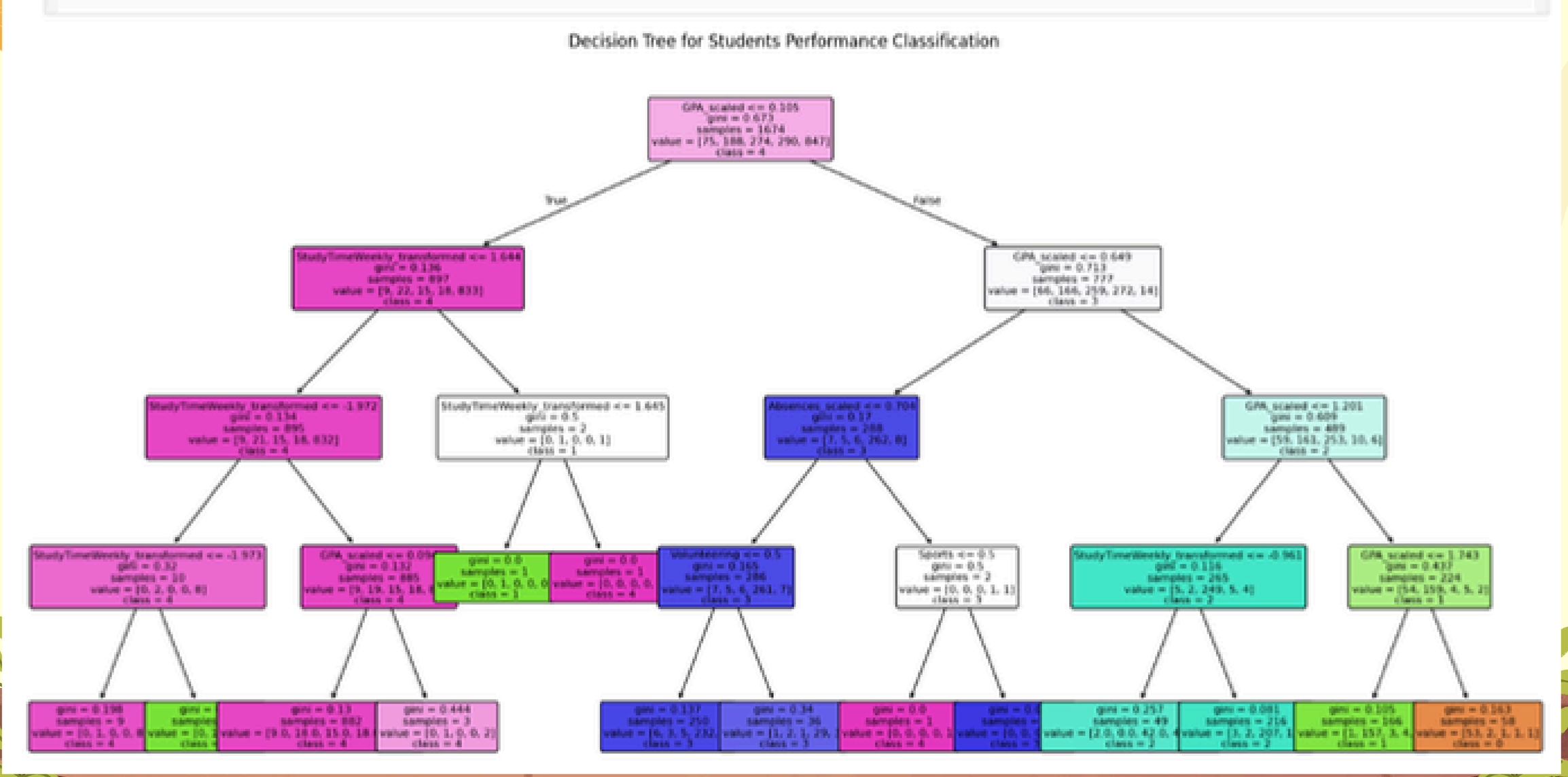
5- FINDINGS AND INSIGHTS

1-Classification

The Decision Tree classifier was evaluated using three train-test splits (60-40, 70-30, 80-20) and two attribute selection measures (Gini and Entropy).

- Results: The highest accuracy (0.868476) was achieved with 80% training data and the Gini criterion, closely followed by 80% training with Entropy (0.860125). The 70% training partition with Gini performed well too (0.855153).
- Insight: The differences in accuracy between Gini and Entropy are minimal, with slight variation depending on the partition.
- Confusion Matrices: Minority GradeClass categories are harder to predict.
- Key Factors: Absences, GPA, and Parental Support are key predictors. Students with more absences and lower GPAs are likely classified into lower GradeClass categories, while students with fewer absences, higher GPAs, and parental support are classified into higher GradeClass categories.

1-Classification



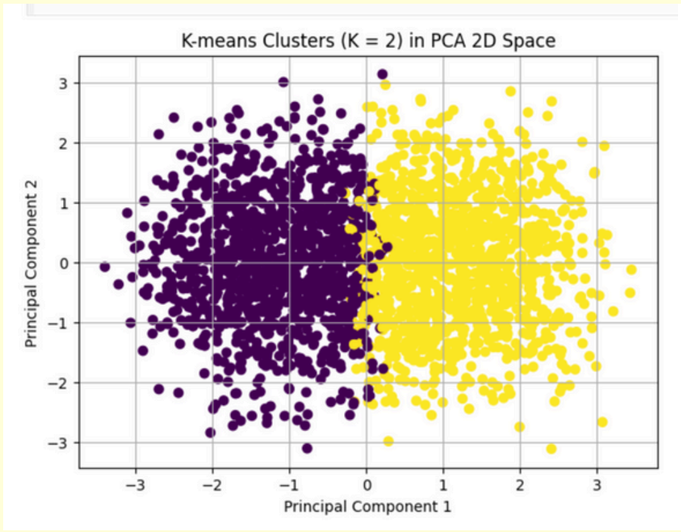
2-Clustering

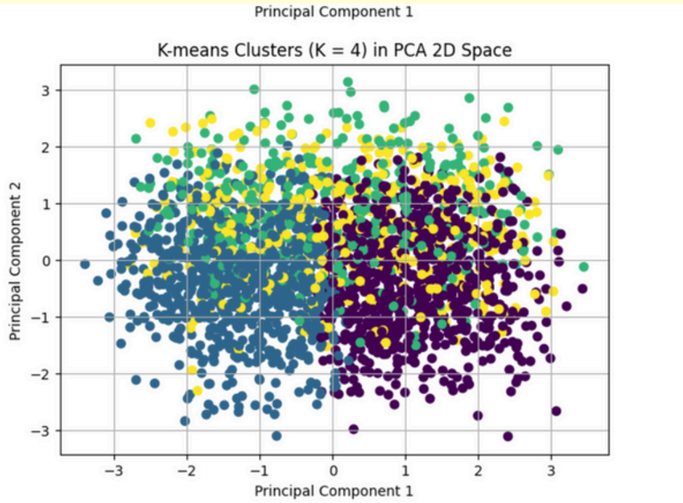
K-means clustering was applied with K values from 2 to 10. Using both WCSS (Elbow Method) and Silhouette scores:

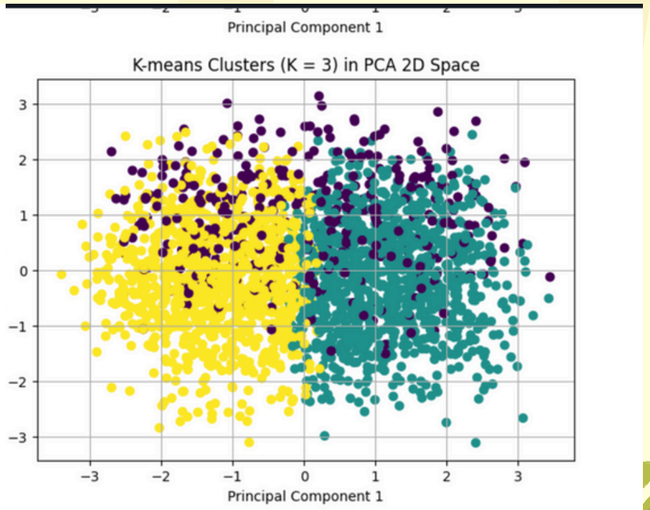
- The Elbow plot suggests K ≈ 4 as a good balance between model complexity and within-cluster variance.
- The Silhouette scores indicate that K = 3 provides the best cluster separation.

The final clusters reveal distinct student groups:

- A high-performance group: high GPA, high study time, low absences, good parental support.
- A low-performance group: low GPA, low study time, many absences, weak parental support.
- One or more intermediate groups with mixed or average characteristics.







THANKS!

Done by:

Remas almutairi

Jood alkhneen

Lujain almajyul

Jwana alothman