King Saud University

College of Computer and Information Sciences

Information Technology Department

IT 326

Data mining project

Final report

G1

77682

Fatmah Alsufaian	444200467
Raseel Alluhaib	444200585
Rana Alotaibi	444201191
Hessah Alotaibi	444200349
Norah Alasmari	444200062

- ## Problem:

In these days schools care about their student’s academic performance but it could be difficult for them to determine who is likely would have low or high grades or even to keeping up with them.

we believe that tracking down student’s performance and other factors to see who is likely to get low or high grades is beneficial to them especially for those who would get low grades, making them feel secured that the school does care about their education and with that they will be encouraged to try their best.

Student final grades determine their understanding of the materials they’ve been getting lectured about in the year, so when we predict that a student will get low grade, we know that he isn’t understanding the materials so why don’t we help him from the beginning?

- ## Data mining task:

In our project, we will employ two data mining tasks to help predict students' final grades based on social factors and academic performance: classification and clustering.

Classification

For the classification task, we will train our model to determine the final grades of students, which can be categorized as low, medium, or high. This classification will be based on a set of features such as age, sex, family size, parental education, study time, failures, activities, and other relevant social and academic factors. The class attribute for this task will be the "final grade" (G3) category.

Clustering

For the clustering task, our model will create groups of students who share similar characteristics without considering their final grade classifications. This will help identify patterns and similarities in the data, potentially leading to a deeper understanding of the factors that influence students' academic performance. By uncovering these groupings, we may gain insights into the social and academic factors that correlate with different levels of student achievement.

Goals of the Data Mining Tasks

- **Classification Goal**: To accurately predict the final grades of students based on various features, enabling educators and institutions to identify students at risk of underperforming and provide necessary interventions.
- **Clustering Goal**: To uncover distinct groups of students based on their social and academic characteristics, which can help in understanding the underlying factors that influence academic performance and guide tailored support strategies.

- ## Data:

  -Source: https://www.kaggle.com/datasets/uciml/student-alcohol-consumption
  -Number of Objects: 648
  -Number of attributes: 33
  -Class name: G3

Attributes' description

ATTRIBUTE NAME	DESCRIPTION	DATA TYPE
SCHOOL	STUDENT'S SCHOOL	BINARY
SEX	STUDENT'S SEX	BINARY
AGE	STUDENT'S AGE	NUMERIC
ADDRESS	STUDENT'S HOME ADDRESS TYPE	BINARY
FAMSIZE	FAMILY SIZE	BINARY
PSTATUS	PARENT'S COHABITATION STATUS	BINARY
MEDU	MOTHER'S EDUCATION	NUMERIC
FEDU	FATHER'S EDUCATION	NUMERIC
MJOB	MOTHER'S JOB	NOMINAL
FJOB	FATHER'S JOB	NOMINAL
REASON	REASON TO CHOOSE THIS SCHOOL	NOMINAL
GUARDIAN	STUDENT'S GUARDIAN	NOMINAL
TRAVELTIME	HOME TO SCHOOL TRAVEL TIME	NUMERIC
STUDYTIME	WEEKLY STUDY TIME	NUMERIC
FAILURES	NUMBER OF PAST CLASS FAILURES	NUMERIC
SCHOOLSUP	EXTRA EDUCATIONAL SUPPORT	BINARY
FAMSUP	FAMILY EDUCATIONAL SUPPORT	BINARY
PAID	EXTRA PAID CLASSES WITHIN THE COURSE SUBJECT MATH OR PORTUGUESE	BINARY
ACTIVITIES	EXTRA-CURRICULAR ACTIVITIES	BINARY
NURSERY	ATTENDED NURSERY SCHOOL	BINARY
HIGHER	WANTS TO TAKE HIGHER EDUCATION	BINARY
INTERNET	INTERNET ACCESS AT HOME	BINARY
ROMANTIC	WITH A ROMANTIC RELATIONSHIP	BINARY
FAMREL	QUALITY OF FAMILY RELATIONSHIPS	NUMERIC

FREETIME	FREE TIME AFTER SCHOOL	NUMERIC
GOOUT	GOING OUT WITH FRIENDS	NUMERIC
DALC	WORKDAY ALCOHOL CONSUMPTION	NUMERIC
WALC	WEEKEND ALCOHOL CONSUMPTION	NUMERIC
HEALTH	CURRENT HEALTH STATUS	NUMERIC
ABSENCES	NUMBER OF SCHOOL ABSENCES	NUMERIC
G1	FIRST PERIOD GRADE	NUMERIC
G2	SECOND PERIOD GRADE	NUMERIC
G3	FINAL GRADE	NUMERIC

Missing value:

```
Missing Values in Each Column:
 school        0
sex           0
age           0
address       0
famsize       0
Pstatus       0
Medu          0
Fedu          0
Mjob          0
Fjob          0
reason        0
guardian      0
traveltime    0
studytime     0
failures      0
schoolsup     0
famsup        0
paid          0
activities    0
nursery       0
higher        0
internet      0
romantic      0
famrel        0
freetime      0
goout         0
Dalc          0
Walc          0
health        0
absences      0
G1            0
G2            0
G3            0
```

Statical measures for each numeric column:

- **Age**: There is moderate variability in ages, with a mean of **16.74** an d a variance of **1.48**. This

  indicates a generally youthful population with a limited age range.

- **Mother's Education (Medu)**: The average level of mother's education is **2.51**, with a variance of **1.29**, suggesting moderate variability in education levels.
- **Father's Education (Fedu)**: The average level of father's education is **2.31**, with a variance of **1.21**, indicating similar variability to Medu.
- **Travel Time to School (traveltime)**: Mean travel time is **1.57**, with a variance of **0.56**, showing that most students live relatively close to school.
- **Study Time (studytime)**: The mean study time is **1.93**, with a variance of **0.69**, indicating low variability in study habits.
- **Failures (failures)**: The average number of past class failures is **0.22**, with a variance of **0.35**, suggesting that most students have few or no failures.
- **Family Relationship Quality (famrel)**: The mean family relationship quality is **3.93**, with a variance of **0.91**, suggesting generally positive family dynamics.
- **Free Time (freetime)**: Students have an average free time score of **3.18**, with a variance of **1.10**, indicating moderate variability.

- **Going Out Frequency (goout)**: The mean frequency of going out is **3.18**, with a variance of **1.38**, suggesting diverse social activity levels.
- **Workday Alcohol Consumption (Dalc)**: The mean is **1.50**, with a variance of **0.86**,
  indicating generally low levels of workday alcohol consumption.
- **Weekend Alcohol Consumption (Walc)**: The mean is **2.28**, with a variance of **1.65**,
  showing slightly higher alcohol consumption during weekends.
- **Health Status (health)**: The mean health score is **3.54**, with a variance of **2.09**,
  suggesting moderate variability in health status
- **Absences (absences)**: The mean number of absences is **3.66**, with a variance of **21.54**,
  indicating extreme outliers or high variability in attendance.
- **First Period Grade (G1)**: The average grade is **11.40**, with a variance of **7.54**, suggesting noticeable variability in performance.
- **Second Period Grade (G2)**: The mean is **11.57**, with a variance of **8.49**, showing similar variability to G1.
- **Final Grade (G3)**: The mean final grade is **11.91**, with a variance of **10.44**, indicating
  slightly higher variability in final performance.

	age	Medu	Fedu	traveltime	studytime	failures
count	649.000000	649.000000	649.000000	649.000000	649.000000	649.000000
mean	16.744222	2.514638	2.306626	1.568567	1.930663	0.221880
std	1.218138	1.134552	1.099931	0.748660	0.829510	0.593235
min	15.000000	0.000000	0.000000	1.000000	1.000000	0.000000
25%	16.000000	2.000000	1.000000	1.000000	1.000000	0.000000
50%	17.000000	2.000000	2.000000	1.000000	2.000000	0.000000
75%	18.000000	4.000000	3.000000	2.000000	2.000000	0.000000
max	22.000000	4.000000	4.000000	4.000000	4.000000	3.000000

	famrel	freetime	goout	Dalc	Walc	health
count	649.000000	649.000000	649.000000	649.000000	649.000000	649.000000
mean	3.930663	3.180277	3.184900	1.502311	2.280431	3.536210
std	0.955717	1.051093	1.175766	0.924834	1.284380	1.446259
min	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000
25%	4.000000	3.000000	2.000000	1.000000	1.000000	2.000000
50%	4.000000	3.000000	3.000000	1.000000	2.000000	4.000000
75%	5.000000	4.000000	4.000000	2.000000	3.000000	5.000000
max	5.000000	5.000000	5.000000	5.000000	5.000000	5.000000

	absences	G1	G2	G3
count	649.000000	649.000000	649.000000	649.000000
mean	3.659476	11.399076	11.570108	11.906009
std	4.640759	2.745265	2.913639	3.230656
min	0.000000	0.000000	0.000000	0.000000
25%	0.000000	10.000000	10.000000	10.000000
50%	2.000000	11.000000	11.000000	12.000000
75%	6.000000	13.000000	13.000000	14.000000
max	32.000000	19.000000	19.000000	19.000000

Show the Variance:

Variance helps understand the extent of dispersion or scatter of values in each column. As the variance increases, it indicates that the values are more spread out and scattered away from the mean, whereas decreasing variance suggests that the values are less scattered and closer to the mean value. Therefore, our variance results:

-Absences (21.54) have the highest variance, meaning the number of student absences is very spread out, with some students having many absences and others having very few.

-Grades (G1: 7.54, G2: 8.49, G3: 10.44) also show high variance, indicating that students' academic performance varies significantly.

-Health (2.09) and alcohol consumption (Walc: 1.65) show moderate variance, meaning there are some differences in students' health and weekend alcohol consumption.

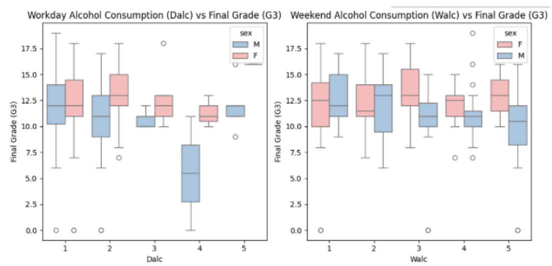
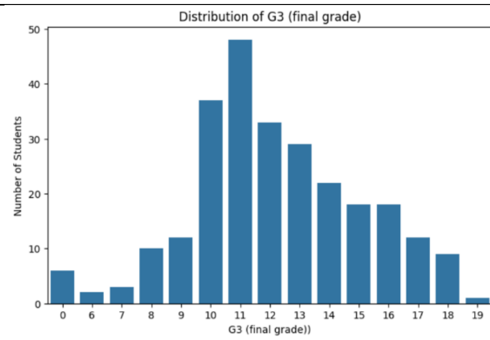
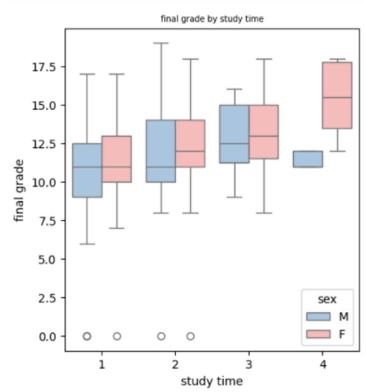
-Failures (0.35) and travel time (0.56) have low variance, meaning most students have similar values in these categories.

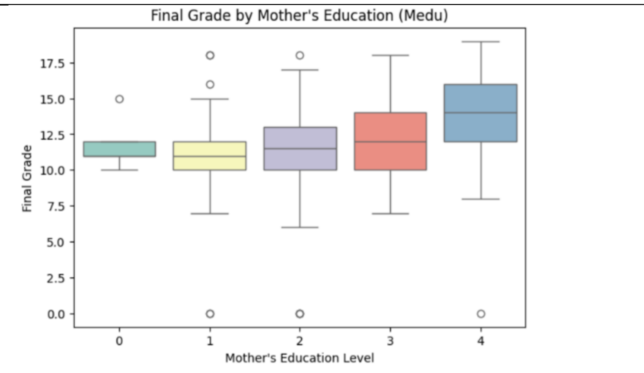
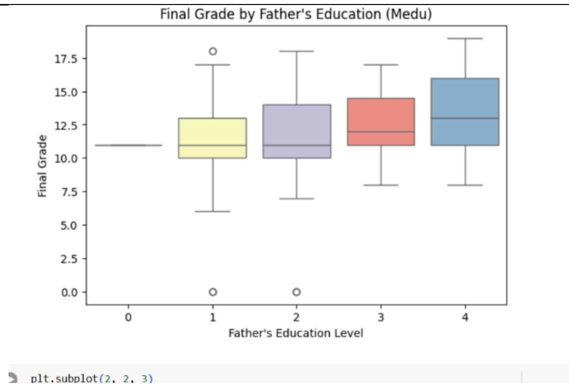
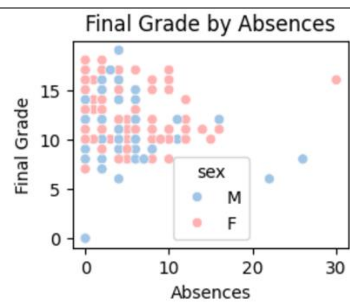
```
Variance Values:
 age            1.483859
Medu           1.287208
Fedu           1.209848
traveltime     0.560492
studytime      0.688086
failures       0.351928
famrel         0.913395
freetime       1.104796
goout          1.382426
Dalc           0.855319
Walc           1.649632
health         2.091665
absences      21.536642
G1             7.536481
G2             8.489290
G3            10.437140
```

Understanding the data through graph representations:

To analyze the relationship between students' academic performance and all the features in the dataset, the "G3" label serves as the class label representing the final grades. This label is linked to all other attributes in the data, allowing us to explore how various factors correlate with students' final performance. This analysis can help identify whether increases in specific attributes—such as study time, family support, or parental education—are associated with higher final grades. Additionally, it allows for the examination of differences in performance between genders and how these differences may relate to age, family size, and socioeconomic status. Understanding these relationships can provide insights into the factors influencing academic achievement, potentially guiding educators in identifying indicators that may support student success and early intervention strategies.

Name of Graph	Picture of Graph	Description

Boxplot		The boxplots show the link between workday (Dalc) and weekend (Walc) alcohol consumption and students' final grades (G3), split by gender. Higher alcohol use is associated with lower grades, especially for boys. This suggests that drinking may negatively affect academic performance.
histogram		This histogram illustrates the distribution of students' final grades (G3). Most students scored around 10 to 12, with a noticeable peak at 11, indicating that this is the most common final grade.
boxplot		This boxplot shows final grades by study time, separated by gender. As study time increases, grades improve, with females generally scoring higher than males. The variability in male grades is wider, indicating more differences in performance.

<table>
<tr>
<td>boxplot</td>
<td>
</td>
<td>This boxplot displays final grades based on the mother’s education level. As the education level increases, the median final grade also rises, indicating a positive correlation.</td>
</tr>
<tr>
<td>boxplot</td>
<td>
</td>
<td>This boxplot represents final grades based on the father's education level. There is a clear correlation between higher education levels (0 to 4) and increased median final grades. Additionally, the spread of grades decreases with higher education levels, indicating greater stability in student performance.</td>
</tr>
<tr>
<td>scatterplot</td>
<td>
</td>
<td>This scatter plot displays final grades in relation to student absences, with points colored by gender, higher absences correlate with lower final grades. The distribution shows a mix of performance levels across both genders, with some high-achieving students despite significant absences</td>
</tr>
</table>

- **Data preprocessing:**

  Data cleaning:

  - Checking missing values:

```
Missing Values in Each Column:
 school        0
sex           0
age           0
address       0
famsize       0
Pstatus       0
Medu          0
Fedu          0
Mjob          0
Fjob          0
reason        0
guardian      0
traveltime    0
studytime     0
failures      0
schoolsup     0
famsup        0
paid          0
activities    0
nursery       0
higher        0
internet      0
romantic      0
famrel        0
freetime      0
goout         0
Dalc          0
Walc          0
health        0
absences      0
G1            0
G2            0
G3            0
dtype: int64
```

Description:

Missing or null values can significantly impact the quality and reliability of a dataset, as well as the insights derived from it. Therefore, we examined our dataset to identify any potential missing or null values. If such values were present, we addressed them by replacing them with the mean value of the target column. Upon review, we found that our dataset is complete and free from missing values.

- Detecting and removing the outliers:

G3 outliers:

	school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	\
163	GP	M	18	U	LE3	T	1	1	other	other	
172	GP	M	16	U	GT3	T	3	3	other	services	
440	MS	M	16	U	GT3	T	1	1	at_home	services	
519	MS	M	16	R	GT3	T	2	1	other	services	
563	MS	M	17	U	GT3	T	2	2	other	other	
567	MS	M	18	R	GT3	T	3	2	services	other	
583	MS	F	18	R	GT3	T	2	2	other	other	
586	MS	F	17	U	GT3	T	4	2	teacher	services	
597	MS	F	18	R	GT3	T	2	2	at_home	other	
603	MS	F	18	R	LE3	A	4	2	teacher	other	
605	MS	F	19	U	GT3	T	1	1	at_home	services	
610	MS	F	19	R	GT3	A	1	1	at_home	at_home	
626	MS	F	18	R	GT3	T	4	4	other	teacher	
637	MS	M	18	R	GT3	T	2	1	other	other	
639	MS	M	19	R	GT3	T	1	1	other	services	
640	MS	M	18	R	GT3	T	4	2	other	other	

	...	famrel	freetime	goout	Dalc	Walc	health	absences	G1	G2	G3
163	...	2	3	5	2	5	4	0	11	9	0
172	...	4	5	5	4	4	5	0	10	10	1
440	...	5	4	5	4	5	3	0	7	0	0
519	...	5	2	1	1	1	2	0	8	7	0
563	...	1	2	1	2	3	5	0	7	0	0
567	...	2	3	1	2	2	5	0	4	0	0
583	...	5	5	5	1	1	3	0	8	6	0
586	...	5	5	5	1	3	5	0	8	8	0
597	...	4	3	3	1	1	4	0	9	0	0
603	...	5	3	1	1	1	5	0	5	0	0
605	...	5	5	5	2	3	2	0	5	0	0
610	...	3	5	4	1	4	1	0	8	0	0
626	...	3	2	2	4	2	5	0	7	5	0
637	...	4	4	3	1	3	5	0	7	7	0
639	...	4	3	2	1	3	5	0	5	8	0
640	...	5	4	3	4	3	3	0	7	7	0

[16 rows x 33 columns]

Dalc outliers:

	school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob
29	GP	M	16	U	GT3	T	4	4	teacher	teacher
54	GP	F	15	U	LE3	A	3	3	other	other
61	GP	F	16	U	GT3	T	1	1	services	services
66	GP	M	15	U	GT3	A	4	4	other	services
100	GP	M	16	U	GT3	T	4	4	services	services
142	GP	M	18	U	LE3	T	3	1	services	services
143	GP	F	18	U	GT3	A	3	2	other	services
172	GP	M	16	U	GT3	T	3	3	other	services
189	GP	M	17	U	LE3	T	4	3	teacher	other
206	GP	M	17	U	GT3	T	1	2	at_home	services
225	GP	M	17	U	LE3	T	4	4	services	other
237	GP	M	18	U	GT3	T	2	2	other	other
242	GP	M	18	U	LE3	T	2	1	at_home	other
250	GP	M	17	U	LE3	T	2	2	other	other
279	GP	M	22	U	GT3	T	3	1	services	services
291	GP	M	18	U	LE3	T	2	1	services	other
322	GP	F	19	R	GT3	T	3	2	services	services
369	GP	F	19	U	GT3	T	1	1	other	other
379	GP	M	17	R	GT3	T	2	2	services	other
413	GP	M	21	R	LE3	T	1	1	at_home	other
415	GP	F	19	U	GT3	T	4	4	teacher	other
418	GP	M	18	R	GT3	T	2	3	other	services
440	MS	M	16	U	GT3	T	1	1	at_home	services
447	MS	M	17	R	GT3	T	2	1	other	other
457	MS	M	17	R	LE3	T	1	2	at_home	services
461	MS	F	16	R	GT3	T	1	1	at_home	other
491	MS	F	19	U	GT3	T	1	1	other	other
500	MS	M	17	U	GT3	T	1	2	other	other
523	MS	M	18	U	LE3	T	4	4	at_home	health
530	MS	M	17	U	GT3	T	3	3	services	services
576	MS	M	18	R	GT3	T	3	2	other	other
598	MS	M	18	U	LE3	T	1	2	at_home	services
626	MS	F	18	R	GT3	T	4	4	other	teacher
640	MS	M	18	R	GT3	T	4	2	other	other

	...	famrel	freetime	goout	Dalc	Walc	health	absences	G1	G2	G3
29	...	4	4	5	5	5	5	4	12	11	12
54	...	5	3	4	4	4	1	0	13	12	13
61	...	5	5	5	5	5	5	0	10	10	16
66	...	1	3	3	5	5	3	0	11	12	12
100	...	4	5	5	5	5	4	12	9	9	8
142	...	3	3	4	4	5	4	2	11	11	12
143	...	4	3	3	5	1	5	10	12	11	11
172	...	4	5	5	4	4	5	0	10	10	1
189	...	4	4	4	4	4	4	0	10	11	11
206	...	4	4	4	4	5	5	16	10	11	12
225	...	5	3	5	4	5	3	15	13	12	12
237	...	3	3	3	5	5	4	9	10	9	10
242	...	4	3	2	4	5	3	2	9	10	11
250	...	4	4	2	5	5	4	0	16	16	16
279	...	5	4	5	5	5	1	12	7	8	5
291	...	4	4	5	4	4	5	4	11	10	11
322	...	3	3	3	4	3	3	0	9	8	10
369	...	1	5	5	4	3	5	12	10	10	11
379	...	4	4	5	5	5	4	2	11	10	10
413	...	5	3	3	5	2	4	21	9	10	10
415	...	3	2	5	4	4	5	5	9	10	11
418	...	3	1	3	4	5	4	13	13	14	14
440	...	5	4	5	4	5	3	0	7	0	0
447	...	5	5	5	5	5	3	8	8	10	9
457	...	5	5	5	5	5	3	4	10	11	11
461	...	4	2	2	4	3	2	0	13	12	14
491	...	1	1	4	4	1	1	12	7	8	9
500	...	5	3	5	5	5	1	12	6	7	7
523	...	5	5	5	5	5	5	2	5	6	6
530	...	4	1	4	5	5	3	8	7	10	9
576	...	2	5	5	5	5	5	8	9	10	11
598	...	4	1	4	5	5	1	8	10	11	11
626	...	3	2	2	4	2	5	0	7	5	0
640	...	5	4	3	4	3	3	0	7	7	0

[34 rows x 33 columns]

```
absences outliers:
    school sex  age address famsize Pstatus  Medu  Fedu      Mjob      Fjob  \
40      GP   F   16       U     LE3       T     2     2     other     other
103     GP   F   15       U     GT3       T     3     2  services     other
150     GP   F   15       U     GT3       A     3     3  services  services
155     GP   M   17       U     GT3       T     2     1     other     other
161     GP   M   16       U     GT3       T     4     4   teacher   teacher
197     GP   F   17       U     LE3       T     3     3     other     other
206     GP   M   17       U     GT3       T     1     2   at_home  services
211     GP   M   16       R     LE3       T     3     3   teacher     other
212     GP   F   17       U     GT3       T     4     4  services   teacher
217     GP   F   17       R     GT3       T     2     2     other     other
230     GP   F   17       U     GT3       T     4     3     other     other
253     GP   F   18       U     LE3       A     2     4  services     other
254     GP   F   18       U     LE3       T     2     2   at_home  services
256     GP   M   18       U     GT3       T     2     2     other   at_home
263     GP   M   18       U     GT3       T     2     2     other  services
311     GP   F   19       U     GT3       T     3     3     other  services
325     GP   M   17       U     LE3       A     4     1  services     other
326     GP   M   17       U     LE3       A     3     2   teacher  services
397     GP   F   17       U     GT3       A     2     2   at_home   at_home
405     GP   F   19       U     LE3       A     2     3   at_home     other
413     GP   M   21       R     LE3       T     1     1   at_home     other

     ... famrel freetime  goout  Dalc  Walc health absences  G1  G2  G3
40   ...      3        3      3     1     2      3       16  11  11  10
103  ...      4        3      5     1     1      2       16  11  10  10
150  ...      1        3      2     2     3      1       24   9   8   9
155  ...      5        4      5     1     2      5       22   9   7   6
161  ...      3        3      2     2     1      5       16   9   9   8
197  ...      5        3      3     2     3      1       32  14  13  14
206  ...      4        4      4     4     5      5       16  10  11  12
211  ...      3        3      4     3     5      3       16  10  11  12
212  ...      4        2      4     2     3      2       30  14  15  16
217  ...      5        3      2     1     2      3       21  13  13  13
230  ...      3        4      5     2     4      1       16  11   9  10
253  ...      4        3      3     1     1      3       18  10  10  10
254  ...      5        3      1     1     1      5       16   9   8  10
256  ...      4        4      3     2     2      1       26   7   8   8
263  ...      5        5      4     3     5      2       16   8   7   8
311  ...      4        3      5     3     3      5       16  11  12  12
325  ...      4        5      4     2     4      5       22  11  11  10
326  ...      4        4      4     3     4      3       18  13  13  13
397  ...      3        3      1     1     2      4       18  10  12  14
405  ...      2        2      3     3     4      5       16  10  11  11
413  ...      5        3      3     5     2      4       21   9  10  10

[21 rows x 33 columns]
```

```
Data after removing outliers:
    school sex  age address famsize Pstatus  Medu  Fedu      Mjob      Fjob
0       GP   F   18       U     GT3       A     4     4   at_home   teacher
1       GP   F   17       U     GT3       T     1     1   at_home     other
2       GP   F   15       U     LE3       T     1     1   at_home     other
3       GP   F   15       U     GT3       T     4     2    health  services
4       GP   F   16       U     GT3       T     3     3     other     other
..     ...  ..  ...     ...     ...     ...   ...   ...       ...       ...
644     MS   F   19       R     GT3       T     2     3  services     other
645     MS   F   18       U     LE3       T     3     1   teacher  services
646     MS   F   18       U     GT3       T     1     1     other     other
647     MS   M   17       U     LE3       T     3     1  services  services
648     MS   M   18       R     LE3       T     3     2  services     other

     ... famrel freetime  goout  Dalc  Walc health absences  G1  G2  G3
0    ...      4        3      4     1     1      3        4   0  11  11
1    ...      5        3      3     1     1      3        2   9  11  11
2    ...      4        3      2     2     3      3        6  12  13  12
3    ...      3        2      2     1     1      5        0  14  14  14
4    ...      4        3      2     1     2      5        0  11  13  13
..   ...    ...      ...    ...   ...   ...    ...      ...  ..  ..  ..
644  ...      5        4      2     1     2      5        4  10  11  10
645  ...      4        3      4     1     1      1        4  15  15  16
646  ...      1        1      1     1     1      5        6  11  12   9
647  ...      2        4      5     3     4      2        6  10  10  10
648  ...      4        4      1     3     4      5        4  10  11  11

[584 rows x 33 columns]
```

Description:

We detected 65 outliers in our dataset out of a total of 650 rows. Since this represents a relatively small proportion of the dataset, these outliers were not considered significantly impactful on the overall analysis. Therefore, we decided to remove them from the dataset to simplify the analysis and ensure more consistent results, while still maintaining the integrity of the remaining data.

- Data transformation:

  - Encoding:

```
     school  sex  age  address  famsize  Pstatus  Medu  Fedu  Mjob  Fjob  ...
0         0    0   18        1        0        0     4     4     0     4  ...
1         0    0   17        1        0        1     1     1     0     2  ...
2         0    0   15        1        1        1     1     1     0     2  ...
3         0    0   15        1        0        1     4     2     1     3  ...
4         0    0   16        1        0        1     3     3     2     2  ...
..      ...  ...  ...      ...      ...      ...   ...   ...   ...   ...  ...
644       1    0   19        0        0        1     2     3     3     2  ...
645       1    0   18        1        1        1     3     1     4     3  ...
646       1    0   18        1        0        1     1     1     2     2  ...
647       1    1   17        1        1        1     3     1     3     3  ...
648       1    1   18        0        1        1     3     2     3     2  ...

     famrel  freetime  goout  Dalc  Walc  health  absences  G1  G2  G3
0         4         3      4     1     1       3         4   0  11  11
1         5         3      3     1     1       3         2   9  11  11
2         4         3      2     2     3       3         6  12  13  12
3         3         2      2     1     1       5         0  14  14  14
4         4         3      2     1     2       5         0  11  13  13
..      ...       ...    ...   ...   ...     ...       ...  ..  ..  ..
644       5         4      2     1     2       5         4  10  11  10
645       4         3      4     1     1       1         4  15  15  16
646       1         1      1     1     1       5         6  11  12   9
647       2         4      5     3     4       2         6  10  10  10
648       4         4      1     3     4       5         4  10  11  11

[584 rows x 33 columns]
```

Description:
Encoding is a method used to convert categorical data into numerical format, which is necessary for most machine learning algorithms. This transformation ensures that the data can be processed effectively by machine learning models, enabling better predictions and analysis. In our case, we converted the following categorical variables: school, sex, address, famsize, pstatus, Mjob, Fjob, reason, guardian, schoolsup, famsup, paid, activities, nursery, higher, internet, and romantic into numerical values. Each category was encoded with values starting from 0 and ranging up to 6, based on the number of distinct categories for each variable.

- Normalization:

	school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	...	famrel	freetime	goout	Dalc	Walc	health	absenc
0	0.0	0.0	0.500000	1.0	0.0	0.0	1.00	1.00	0.00	1.00	...	0.75	0.50	0.75	0.0	0.00	0.5	0.2666
1	0.0	0.0	0.333333	1.0	0.0	1.0	0.25	0.25	0.00	0.50	...	1.00	0.50	0.50	0.0	0.00	0.5	0.1333
2	0.0	0.0	0.000000	1.0	1.0	1.0	0.25	0.25	0.00	0.50	...	0.75	0.50	0.25	0.5	0.50	0.5	0.4000
3	0.0	0.0	0.000000	1.0	0.0	1.0	1.00	0.50	0.25	0.75	...	0.50	0.25	0.25	0.0	0.00	1.0	0.0000
4	0.0	0.0	0.166667	1.0	0.0	1.0	0.75	0.75	0.50	0.50	...	0.75	0.50	0.25	0.0	0.25	1.0	0.0000

ss	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	...	famrel	freetime	goout	Dalc	Walc	health	absences	G1	G2	G3
.0	0.0	0.0	1.00	1.00	0.00	1.00	...	0.75	0.50	0.75	0.0	0.00	0.5	0.266667	0.000000	0.428571	0.384615
.0	0.0	1.0	0.25	0.25	0.00	0.50	...	1.00	0.50	0.50	0.0	0.00	0.5	0.133333	0.473684	0.428571	0.384615
.0	1.0	1.0	0.25	0.25	0.00	0.50	...	0.75	0.50	0.25	0.5	0.50	0.5	0.400000	0.631579	0.571429	0.461538
.0	0.0	1.0	1.00	0.50	0.25	0.75	...	0.50	0.25	0.25	0.0	0.00	1.0	0.000000	0.736842	0.642857	0.615385
.0	0.0	1.0	0.75	0.75	0.50	0.50	...	0.75	0.50	0.25	0.0	0.25	1.0	0.000000	0.578947	0.571429	0.538462

Description:

In the normalization process, we standardized the attributes by unifying their scale [0-1], as the ranges of the attributes varied significantly. This approach ensures that all values in the dataset are formatted consistently, making it easier to analyze the data by preventing attributes with larger ranges from dominating the analysis. Normalization helps streamline the dataset and enhances the reliability of the results

- Discretization:

```
     school  sex  age  address  famsize  Pstatus  Medu  Fedu  Mjob  Fjob  ...  \
0         0    0    1        1        0        0     4     4     0     4  ...
1         0    0    1        1        0        1     1     1     0     2  ...
2         0    0    0        1        1        1     1     1     0     2  ...
3         0    0    0        1        0        1     4     2     1     3  ...
4         0    0    0        1        0        1     3     3     2     2  ...
..      ...  ...  ..      ...      ...      ...   ...   ...   ...   ...  ...
644       1    0    2        0        0        1     2     3     3     2  ...
645       1    0    1        1        1        1     3     1     4     3  ...
646       1    0    1        1        0        1     1     1     2     2  ...
647       1    1    1        1        1        1     3     1     3     3  ...
648       1    1    1        0        1        1     3     2     3     2  ...

     famrel  freetime  goout  Dalc  Walc  health  absences      G1      G2  \
0         4         3      4     1     1       3    Medium     Low     Low
1         5         3      3     1     1       3       Low     Low     Low
2         4         3      2     2     3       3    Medium  Medium  Medium
3         3         2      2     1     1       5       Low    High    High
4         4         3      2     1     2       5       Low  Medium  Medium
..      ...       ...    ...   ...   ...     ...       ...     ...     ...
644       5         4      2     1     2       5    Medium     Low     Low
645       4         3      4     1     1       1    Medium    High    High
646       1         1      1     1     1       5    Medium  Medium  Medium
647       2         4      5     3     4       2    Medium     Low     Low
648       4         4      1     3     4       5    Medium     Low     Low

         G3
0       Low
1       Low
2    Medium
3      High
4    Medium
..      ...
644     Low
645    High
646     Low
647     Low
648     Low

[584 rows x 33 columns]
```

- Description:

In the discretization process, we categorized the attributes (G1, G2, G3) into three categories: low, medium, and high. Additionally, we grouped the absence attribute into two categories (low and medium) and the age attribute into three categories (0, 1, and 2). This transformation simplifies the dataset, making it easier to analyze patterns and relationships within the data. By grouping these attributes, we enhance interpretability while preserving the essential information in the dataset.

- ## Data Mining Technique:

We utilized both supervised and unsupervised learning methods on our data using classification and clustering techniques.

For our classification task, we utilized a decision tree. This algorithm recursively builds a tree structure where each leaf node corresponds to a final decision. Our model aims to predict the final grade (G3) of students, categorizing the results into three classes: '0' (low grade), '1' (medium grade), and '2' (high grade). It makes predictions based on several attributes: absences, first-period grade (G1), and second-period grade (G2).

Since classification is a type of supervised learning, training data is required to train the model. We split our dataset into two subsets: training data and testing data. We experimented with three different training subset sizes: 70%, 60%, and 80%, and used two attribute selection measures: Information Gain (Entropy) and Gini Index.

To evaluate our model and determine the best partitioning, we measured its accuracy and used a confusion matrix to calculate performance metrics such as precision, recall (sensitivity), specificity, and error rate. By comparing these metrics across different partitioning ratios and attribute selection measures, we were able to gain insights into the model's performance and determine the most effective setup for predicting student grades.

In the clustering process, which is a type of unsupervised learning, we excluded the "G3" grade label attribute since clustering does not rely on class labels. Instead, we utilized all other relevant attributes, such as: school, sex, age, address, famsize, Pstatus, Medu, Fedu, Mjob, Fjob, reason, guardian, traveltime, studytime, failures, schoolsup, famsup, paid, activities, nursery, higher, internet, romantic, famrel, freetime, goout, Dalc, Walc, health, and absences. All of these attributes are numeric or binary and required no conversion prior to clustering.

For the clustering task, we employed the K-means algorithm. This algorithm generates K clusters, each represented by the centroid of the cluster. It assigns each student to the closest cluster and then iteratively recalculates the centroids and reassigns the students until the centroids stabilize, indicating the correct cluster assignments.

For cluster validation, we calculated the Average Silhouette Score for each cluster using the Average Silhouette Score method and visualized these scores. Additionally, we applied the Within-Cluster Sum of

Squares (WSS) method to compare three different cluster sizes and determine the optimal number of clusters. This was done by evaluating the separation and compactness of the clusters.

- ## Evaluation and Comparison:
  - Classification [70% training, 30% testing] Information Gain:

Figure (1) (Decision tree):

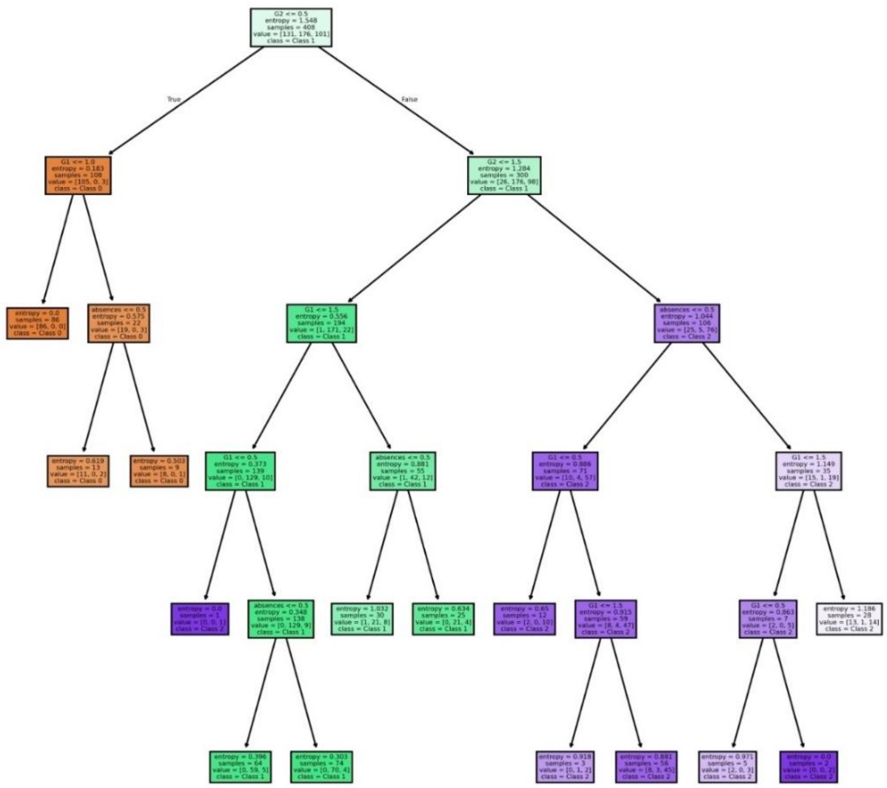
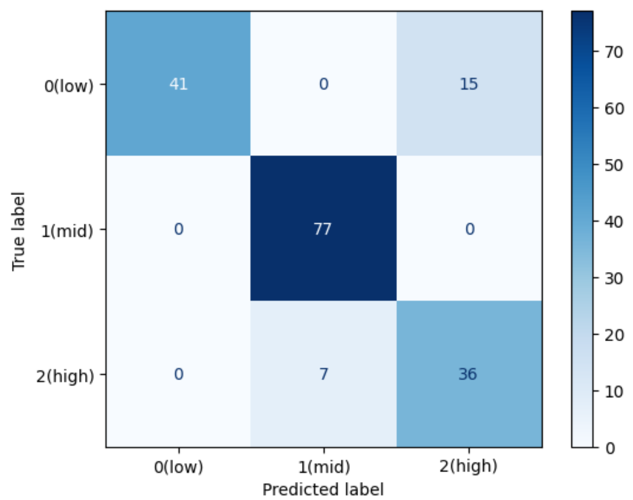


Figure (2) (confusion matrix):



- Classification [80% training, 20% testing] Information Gain:

Figure (1) (Decision tree):

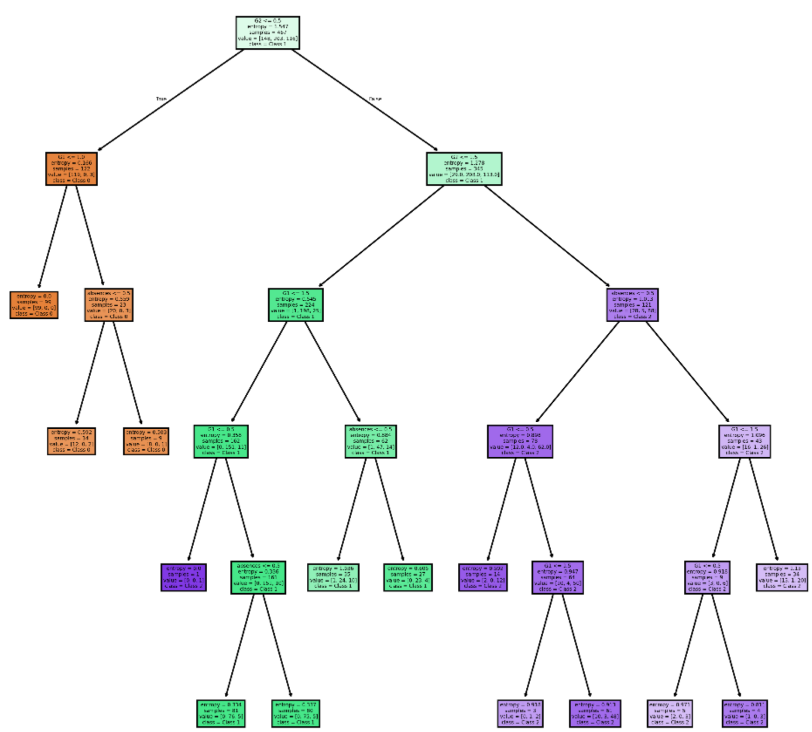
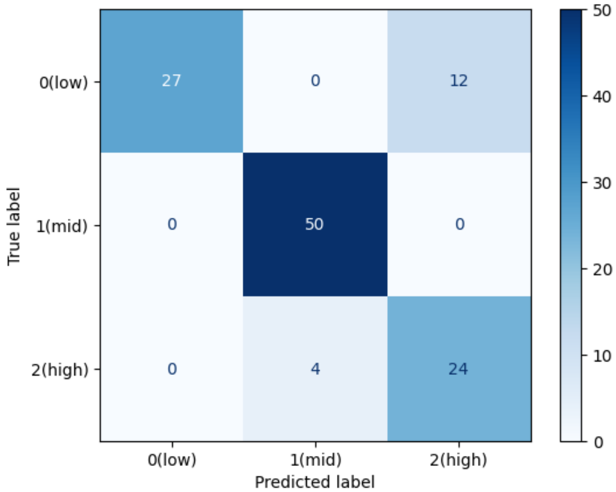


Figure (2) (confusion matrix):



- Classification [60% training, 40% testing] Information Gain:

Figure (1) (Decision tree):

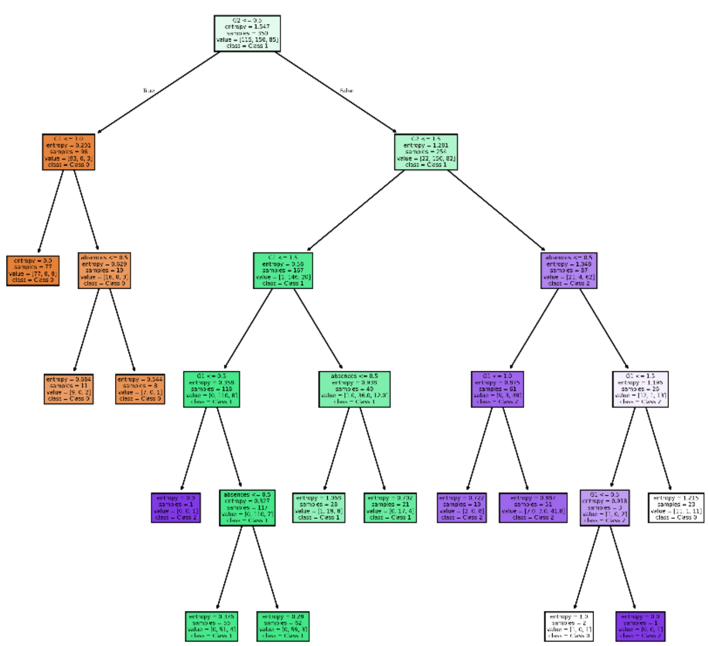
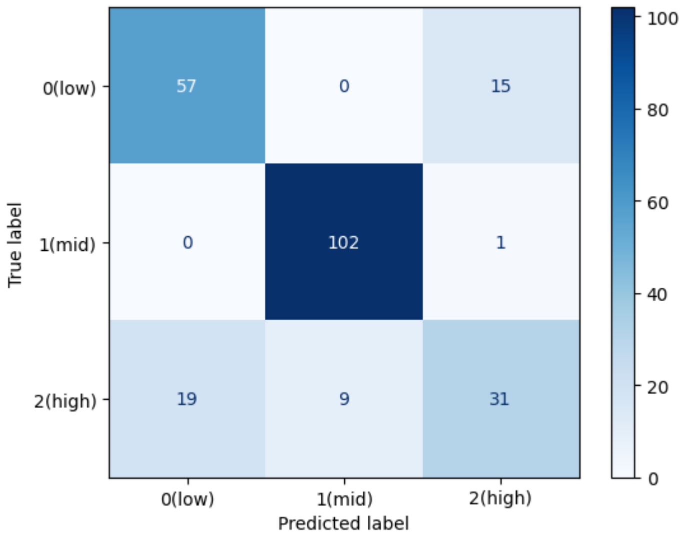


Figure (2) (confusion matrix):



Mining task	Comparison Criteria

<table>
<tr>
<td>Classification for information Gain</td>
<td>
We tried 3 different sizes for dataset splitting to create the decision tree:

70% training data, 30% testing data.
<table>
<tr><td>Accuracy</td><td>87%</td></tr>
<tr><td>recall</td><td>87%</td></tr>
<tr><td>precision</td><td>89%</td></tr>
<tr><td>Error rate</td><td>13%</td></tr>
</table>
80% training data, 20% testing data.
<table>
<tr><td>Accuracy</td><td>~ 86.2%</td></tr>
<tr><td>recall</td><td>~ 86.2%</td></tr>
<tr><td>precision</td><td>~ 88.8%</td></tr>
<tr><td>Error rate</td><td>~13.8%</td></tr>
</table>
60% training data, 40% testing data.
<table>
<tr><td>Accuracy</td><td>81%</td></tr>
<tr><td>recall</td><td>81%</td></tr>
<tr><td>precision</td><td>~ 80%</td></tr>
<tr><td>Error rate</td><td>19%</td></tr>
</table>
</td>
</tr>
</table>

- Classification [70% training, 30% testing] Gini index:

Figure (1) (Decision tree):

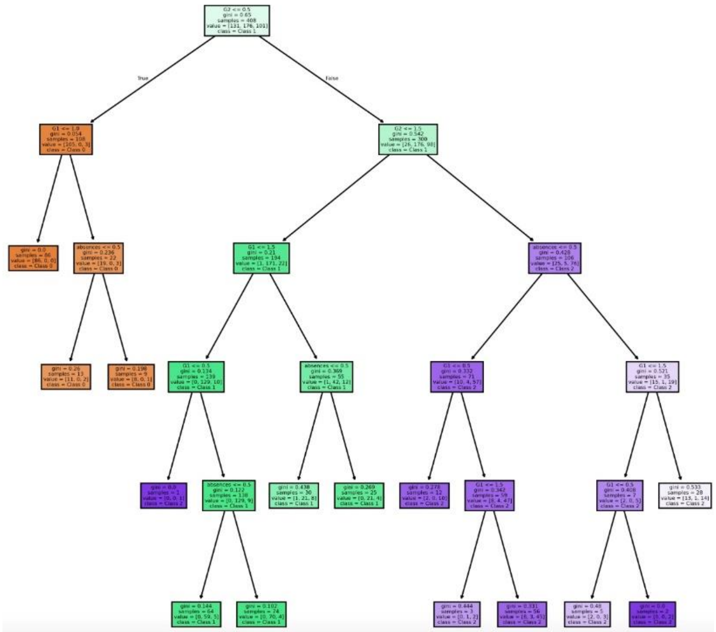
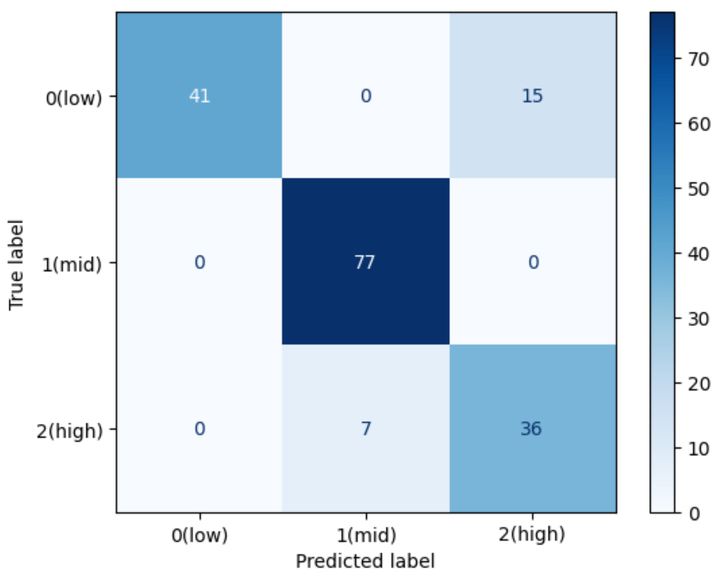


Figure (2) (confusion matrix):



- Classification [80% training, 20% testing] Gini index:

Figure (1) (Decision tree):

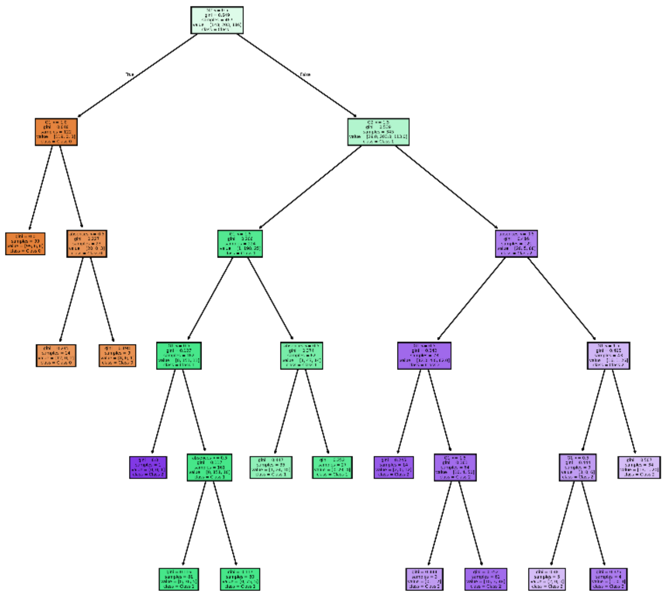
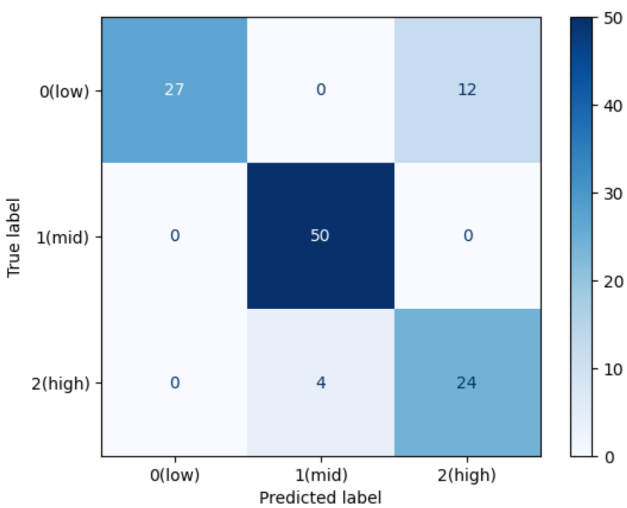


Figure (2) (confusion matrix):



- Classification [60% training, 40% testing] Gini index:

Figure (1) (Decision tree):

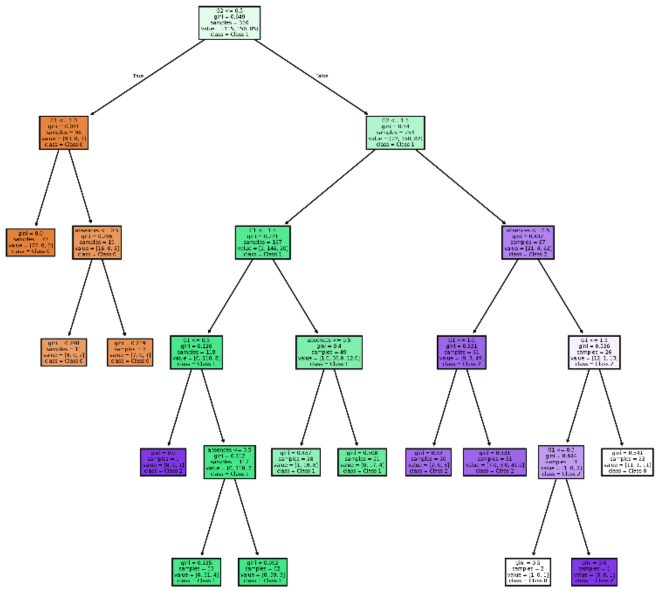
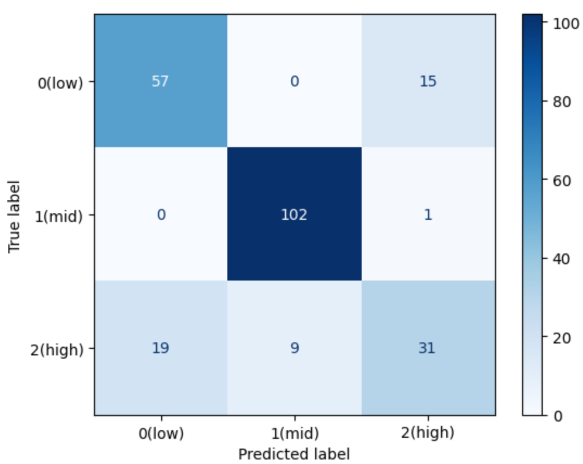
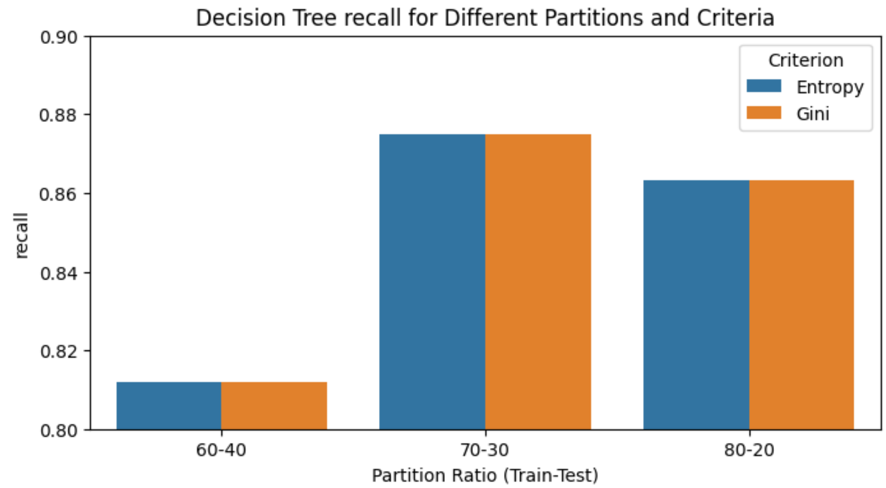


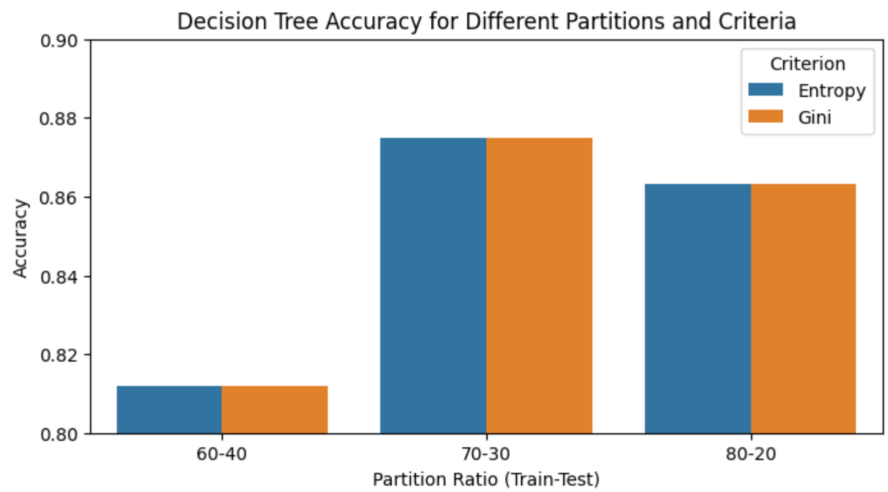
Figure (2) (confusion matrix):

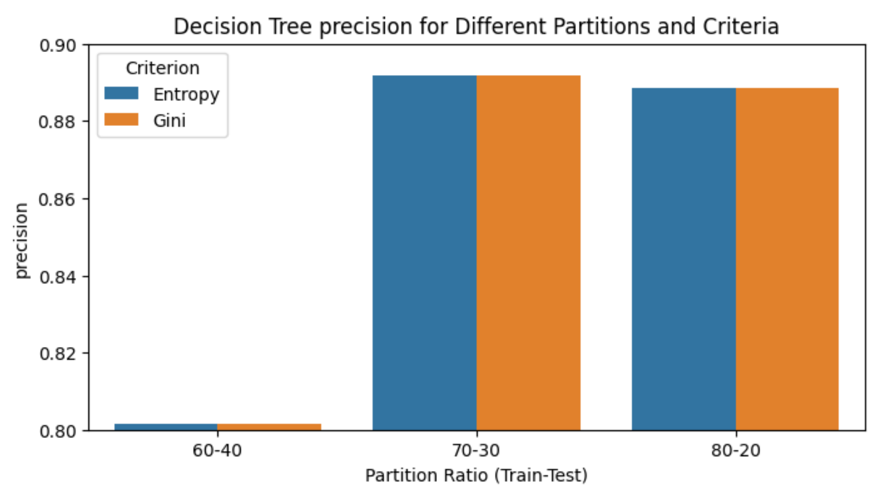


<table>
<tr><th>Mining task</th><th>Comparison Criteria</th></tr>
<tr><td>Classification for Gini index</td><td>We tried 3 different sizes for dataset splitting to create the decision tree:

70% training data, 30% testing data.
<table>
<tr><td>Accuracy</td><td>87%</td></tr>
<tr><td>recall</td><td>87%</td></tr>
<tr><td>precision</td><td>89%</td></tr>
<tr><td>Error rate</td><td>13%</td></tr>
</table>
80% training data, 20% testing data.
<table>
<tr><td>Accuracy</td><td>~ 86.2%</td></tr>
<tr><td>recall</td><td>~ 86.2%</td></tr>
<tr><td>precision</td><td>~ 88.8%</td></tr>
<tr><td>Error rate</td><td>~13.8%</td></tr>
</table>
60% training data, 40% testing data.
<table>
<tr><td>Accuracy</td><td>81%</td></tr>
<tr><td>recall</td><td>81%</td></tr>
<tr><td>precision</td><td>~ 80%</td></tr>
<tr><td>Error rate</td><td>19%</td></tr>
</table>
</td></tr>
</table>







The better partitioning:

The 70% - 30% split using Gini or entropy Index yields better overall performance, with high accuracy, and high values for recall, and precision, which is why it is considered the best based on the provided results.

Clustering:

We tried k to be from 2 to 10 then We took 3 different sizes [2,3,4] based on the result of the validation methods that we will apply then we will use these sizes to perform the. K-means clustering.

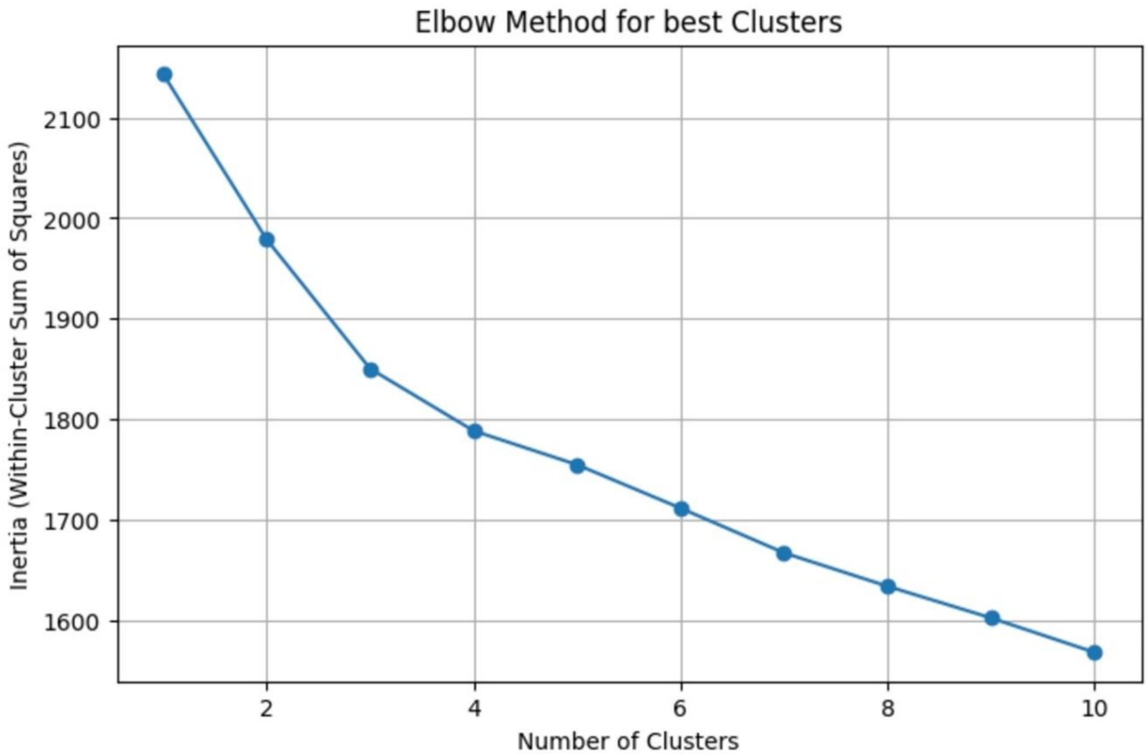
Silhouette method:

The Silhouette method is a technique used to evaluate the quality of clustering results. It measures how well each data point fits within its assigned cluster compared to neighboring clusters.

In silhouette we chose the number of clusters who have the highest score in this case it's 2 scoring 0.080

Elbow method:

The Elbow method is a technique used to determine the optimal number of clusters in a dataset for K-means clustering.



In elbow we chose the number of clusters where it's located as the elbow in this case it's 3 and 4

Mining task	Comparison Criteria

<table>
<tr>
<td>Clustering</td>
<td>We tried 3 different sizes for dataset splitting to create the decision tree:

K=2, k=3, k=4
<table>
<tr><td>No. Of Cluster</td><td>K=2</td><td>K=3</td><td>K=4</td></tr>
<tr><td>Average Silhouette width</td><td>0.08</td><td>0.07</td><td>0.06</td></tr>
<tr><td>Total within-cluster sum of square</td><td>1978.8</td><td>1850.26</td><td>1788.51</td></tr>
</table>
</td>
</tr>
</table>

- ## Findings:

My goal is to have a model that predict student grades as (low – mid – high) so we can identify if there is someone who is struggling and he is likely will get low grade

So, we did two data mining techniques, classification and clustering now let's see the best solutions for both of these techniques

- **Classification Analysis:**

	70%-30%	80%-20%	60%-40%
Accuracy	87%	~86.2%	81%
recall	87%	~86.2%	81%
precision	89%	~ 88.8%	~ 80%

Error rate	13%	~13.8%	19%

For accuracy, recall and precision we want the highest percentage because it's means the model predict in this particular splitting better in our case it goes for when we split the data 70% of it for training and 30% for testing having 87% accuracy and recall

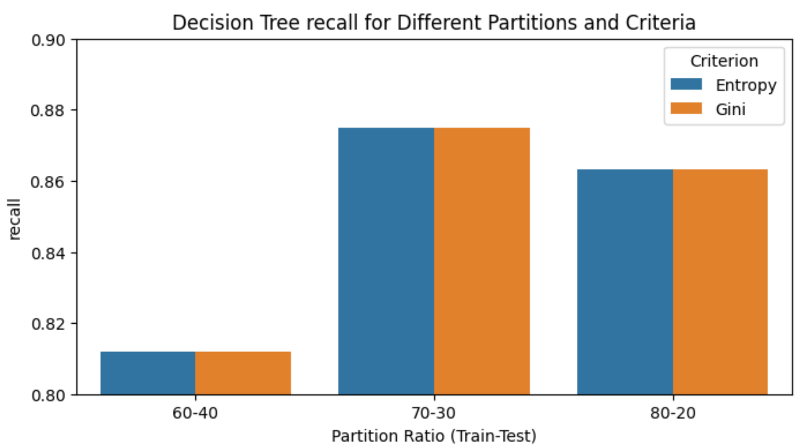
Means the model's predictions 87% of them is correct

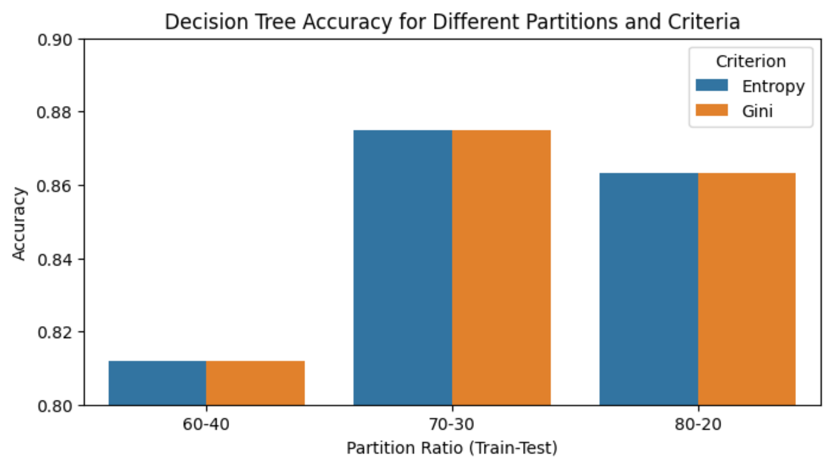
And for error rate the lower the rate the better prediction the model makes, also 70%-30% splitting has the best error rate among the others which is 13% only.

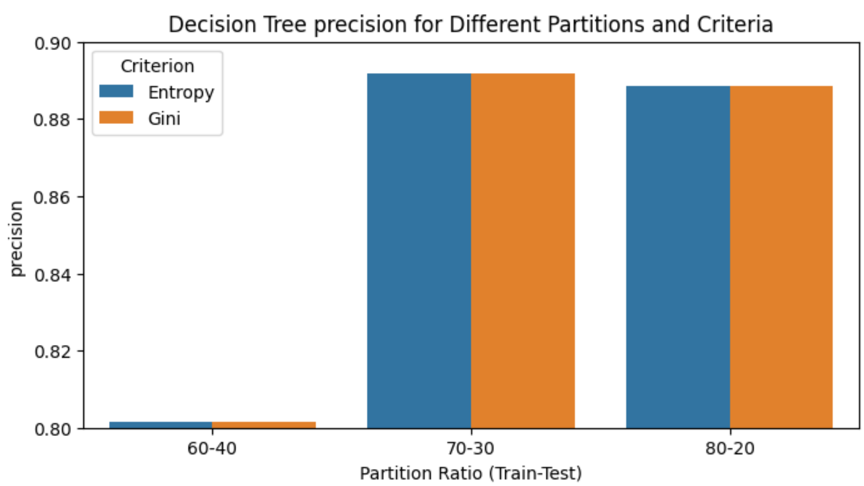
Means the model's prediction 13% of them is incorrect

Now what is the better solution for 70%-30% splitting? Is it entropy or Gini index?

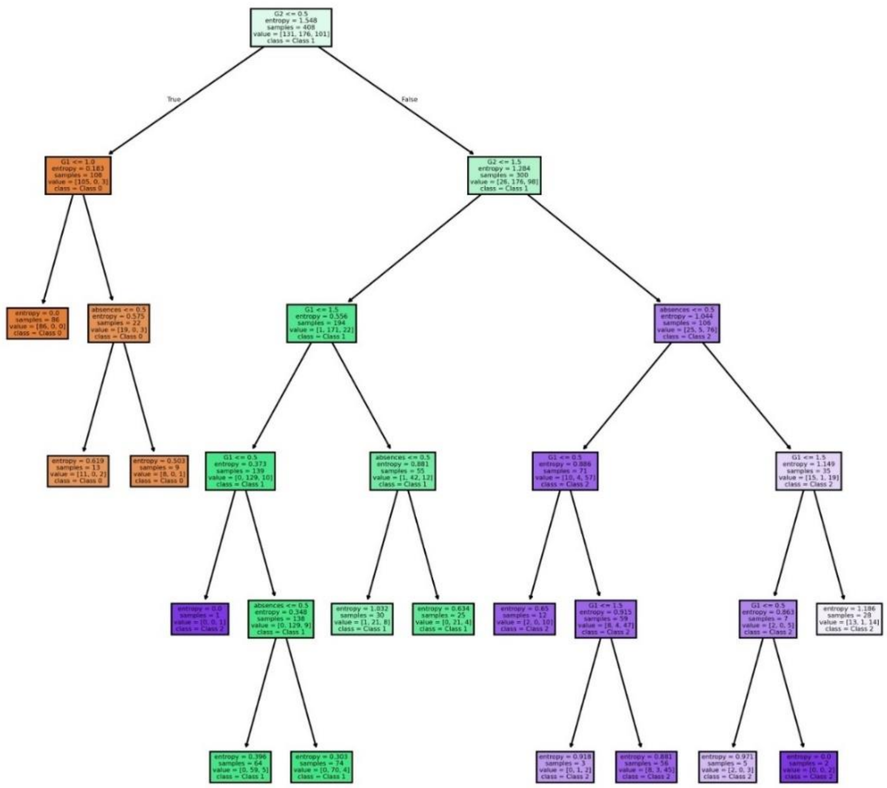
Well, both have the same result as shown in these figures







<u>And this is the tree</u>



- **Clustering**

No. Of Cluster	K=2	K=3	K=4
Average Silhouette width	0.08	0.07	0.06
Total within-cluster sum of square	1978.8	1850.26	1788.51

When we 2 clusters the distance (Average Silhouette width) between them is 0.08 while having 3 or 4 clusters the distance decreases, means the clusters overlaps and become similar to each other so let’s take our best number of clusters which is 2

Also, when we see the Total within-cluster sum of square having 2 clusters gives the highest Total within-cluster sum of square

So, 2 is our best choice indicating that it creates distinct and cohesive clusters.

- ## References:

1. UCI Machine Learning Repository, "Student Alcohol Consumption Dataset," Kaggle, Available: https://www.kaggle.com/datasets/uciml/student-alcohol-consumption. [Accessed: Sep. 8, 2024].

2. “Labs and Lecture Slides," College of Computer Science, Department of Information Technology, King Saud University.