

King Saud University

College of Computer and Information Sciences

Information Technology Department

**IT326: Data Mining**

Project final report

Student flexibility in Online Learning

Group#:6

LAB Day-Time: Wed-8

|  |  |  |
| --- | --- | --- |
| Group#: | 6 | |
| Section#: | 74557 | |
| **Group members** | Student name | ID |
| Noha salem aljohani | 443200429 |
| Norah nasser aljedai | 443200841 |
| Ftoon Fawaz Binkhatttaf | 443200499 |
| Batool Alkhuraim | 443200604 |

Contents

[1. Problem 2](#_Toc165923343)

[2. Data Mining Task 2](#_Toc165923344)

[3. Data 2](#_Toc165923345)

[4. Data preprocessing 6](#_Toc165923346)

[5. Data Mining Technique 9](#_Toc165923347)

[6. Evaluation and Comparison 10](#_Toc165923348)

[7. Findings 17](#_Toc165923349)

[8. References 19](#_Toc165923350)

# Problem

Our focus lies in the flexibility level of students during online learning. The prevalence of online education has significantly increased in recent years, particularly due to the COVID-19 pandemic. Our project aims to study and analyze students' data to pinpoint factors influencing flexibility levels, we aim to provide insights that can enhance flexibility in online studying for students to become more flexibility in studying online

# Data Mining Task

In our project we will use two data mining tasks to help us predict the flexibility level, which are classification and clustering.

For classification we will train our model to be able to classify the flexibility level of student, based on location, device, age, internet and financial status. For the clustering our model will create a set of clusters for the students who have similar characteristics, then these clusters will be used to predict new students results

# Data

The Source: <https://www.kaggle.com/datasets/shariful07/student-flexibility-in-online-learning>

Number of objects:1206

Number of attributes:11

characteristics of attributes

|  |  |
| --- | --- |
| attributes | Data type |
| Education level | Categorical (Ordinal) |
| Gender | Binary(symmetric) |
| Age | Numerical |
| Device | Categorical (nominal) |
| Location | String |
| Financial Condition | Categorical (Ordinal) |
| Institution Type | Categorical (nominal) |
| Internet Type | Categorical (nominal) |
| Network Type | Categorical (nominal) |
| Flexibility Level | Categorical (Ordinal) |
| It student | Binary |

|  |  |  |
| --- | --- | --- |
| Missing values |  | To check the missing values in the dataset |
| Distribution |  | The bar plot shows the flexibility levels of the students ,the low and moderate flexibility levels are more common among students, with approximately 80 students, while high flexibility is less frequent. |
| statistical measures |  | We counted the  Statistical measures for the data |
| Graphs and tables show variables distribution |  | The plot histogram illustrate the frequency of the age attribute within our dataset |
| Boxplot : |  | The boxplot for the   * High Flexibility Level shows that there are no outliers in the age attribute. The Interquartile Range (IQR) extends from approximately between 17.5 to 22.5 years, indicating that the median age is around 20 years * Low Flexibility Level there are no outliers for the age attribute. The median age is consistently around 20 years, The Interquartile Range (IQR) approximately from 17.5 to 22.5 years, mirroring the values observed in the- High Flexibility Level-   Moderate Flexibility Level, the median age is also around 20 (IQR) is approximately from 17.5 to 22.5 years, The whiskers extend much less in this category, down to about 10 years, indicating a wider range of ages compared  in the High and Low Flexibility Levels. |
| Bar plots : |  | The bar plots shows an illustration for all our categorical variables  The education level the univrities has more frequnce the school then the college |
|  | **A bar chart with different colored bars  Description automatically generated** | The bar plot demonstrates the frequency of each flexibility level among students, ordered from most to least frequent as follows: Low, Moderate, and High. Both 'Low' and 'Moderate' flexibility levels have nearly the same highest frequency |
|  | **A bar graph with different colored squares  Description automatically generated** | The network type (4G) is more frequent than (3G-2G) |
|  | **A bar graph showing different types of internet data  Description automatically generated** | The bar plot indicates that mobile data is the most frequently used internet type |
|  | **A bar graph with different colored squares  Description automatically generated** | Bar Plot of Financial Condition" illustrates the frequency of financial conditions(mid-poor-rich) |
|  |  | bar plot representing gender shows a higher frequency of males than females |
|  | **A bar chart with blue and orange squares  Description automatically generated** | The (Town)shows higher frequency than rural areas |
|  | **A bar graph with blue and orange squares  Description automatically generated** | the most frequent student were not IT student |
|  | **A bar graph with different colored squares  Description automatically generated** | The Bar Plot of Device illustrates that the most used device is mobile -computer then taps |
|  | **A bar chart with blue and orange squares  Description automatically generated** | The Bar Plot of Institution Type illustrates that private institutions have a higher frequency of attendance compared to public institutions |

# Data preprocessing

* Checking for missing values

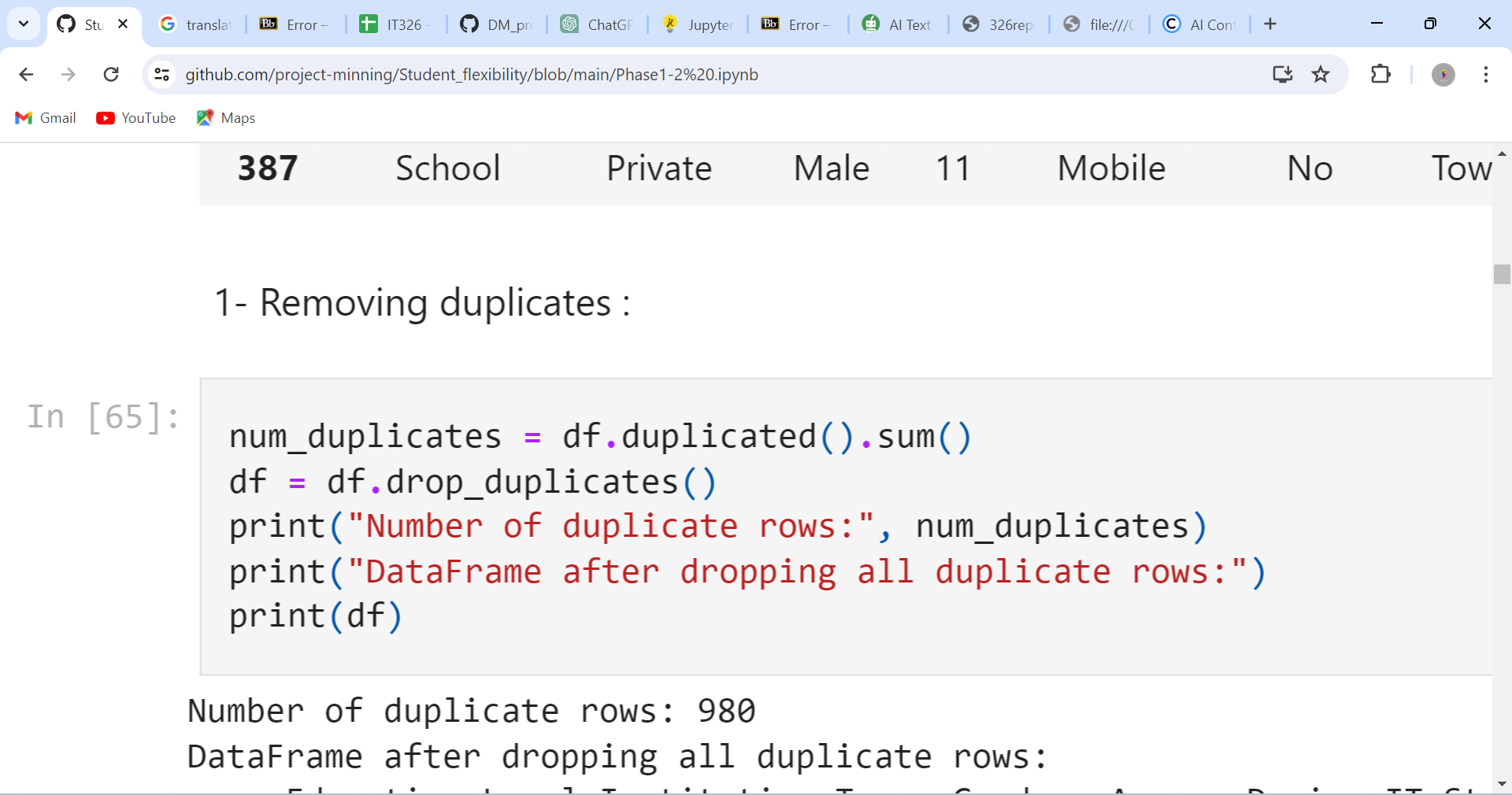
A close-up of a computer screen

Description automatically generated

**Description** :

Identifying and addressing missing values in datasets is crucial for maintaining the integrity and reliability of data analysis. Missing values can compromise statistical estimates and lead to misleading conclusions. Analyzing missing data patterns helps refine data collection strategies, ensuring more accurate and robust analysis outcomes.

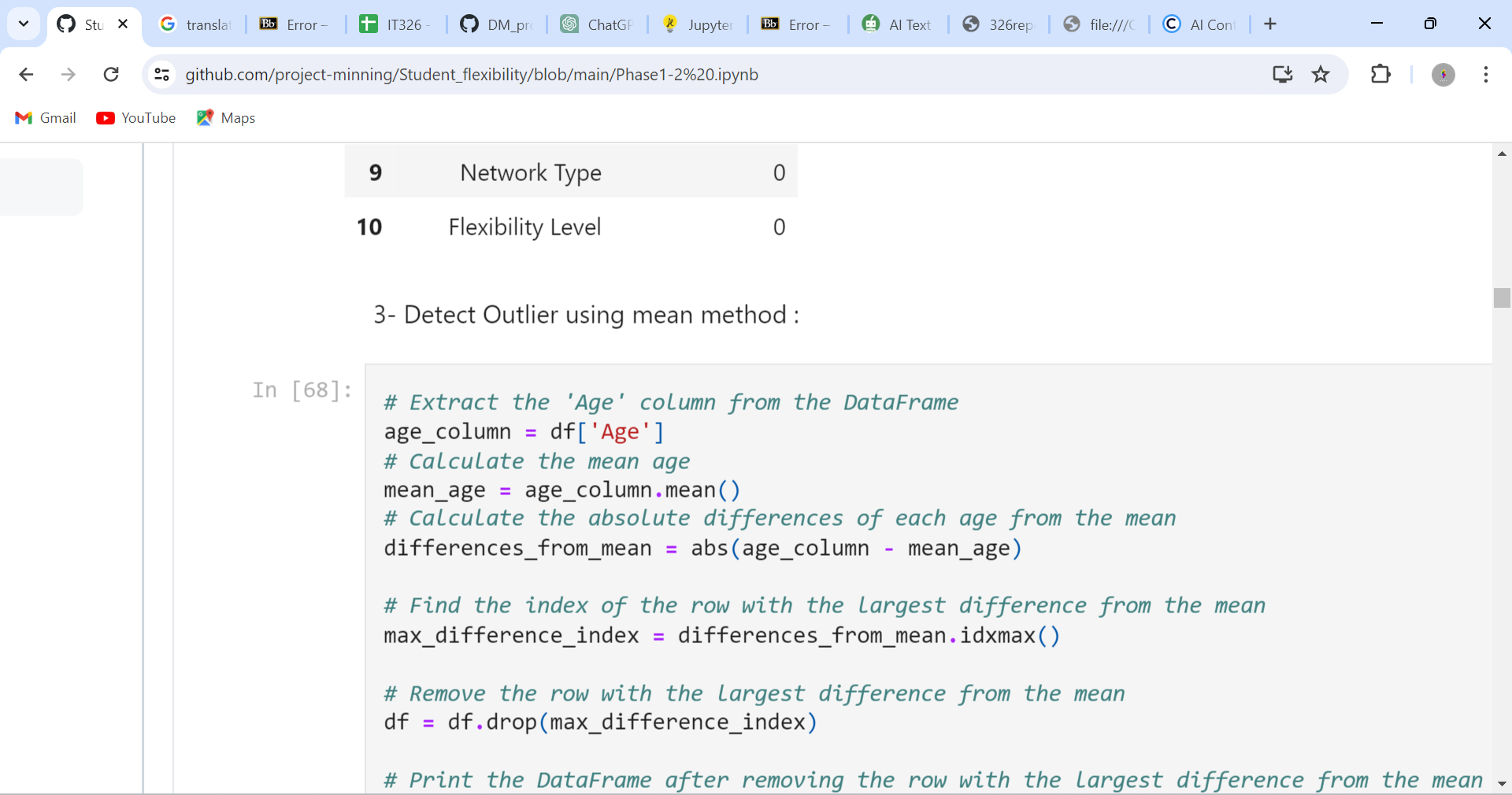
* Removing duplicates



**Description** :

Duplicates can lead to inaccuracies in analysis by artificially inflating certain statistics or biasing results. Removing duplicates helps maintain the integrity of your dataset and to give Accurate Model Training beside Duplicate entries can cause inconsistencies and removing the duplicates ensures the efficient of the data to make reliable decisions

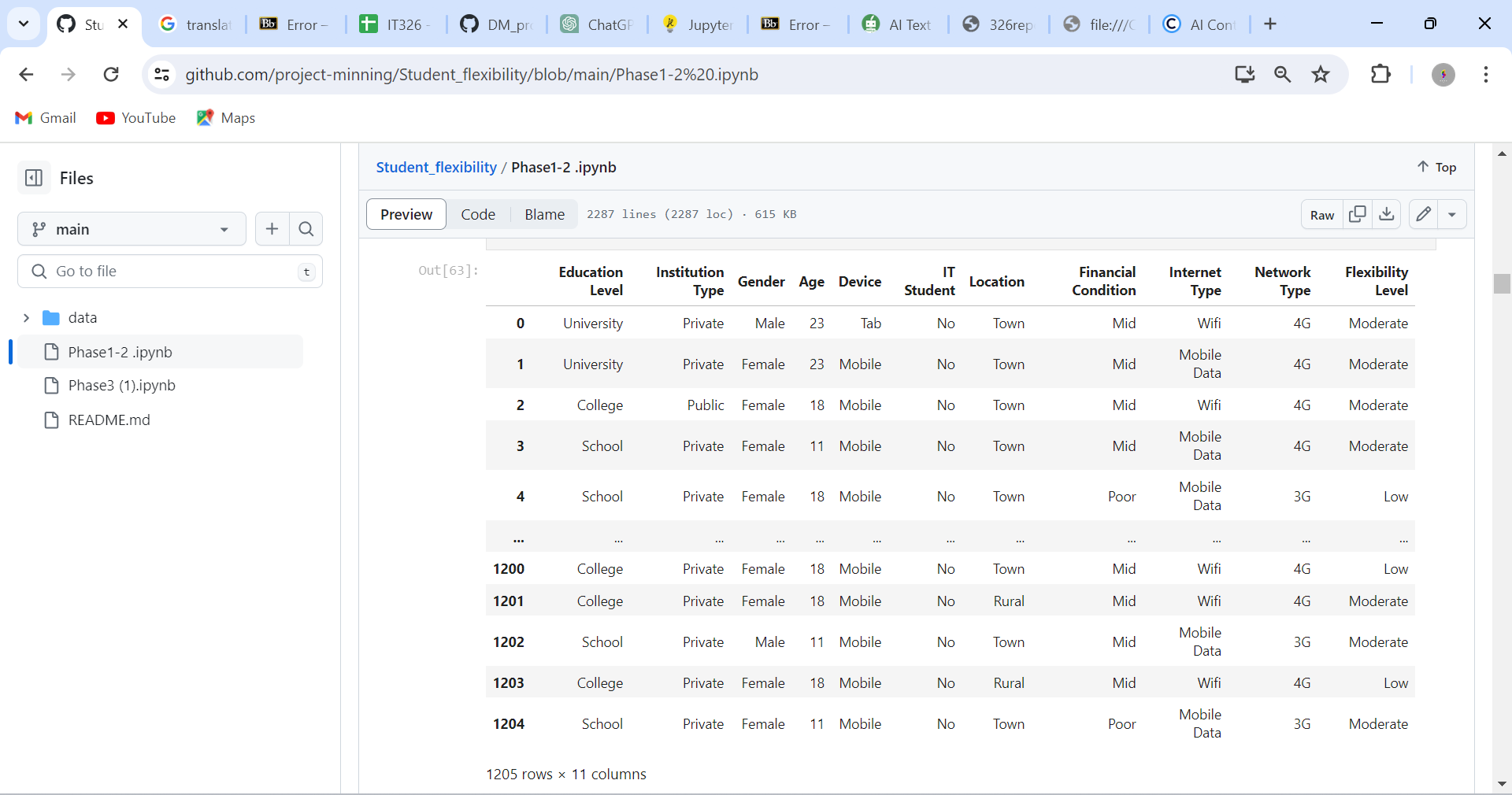
* Detect Outliers



**Description** :

Since there were no outliers, we removed the row with the largest difference from the mean to refine the dataset and enhance the accuracy of our results. This adjustment helps ensure that our data is more representative and reliable for achieving optimal outcomes.

**Raw data** Our raw dataset before Removing duplicates

****

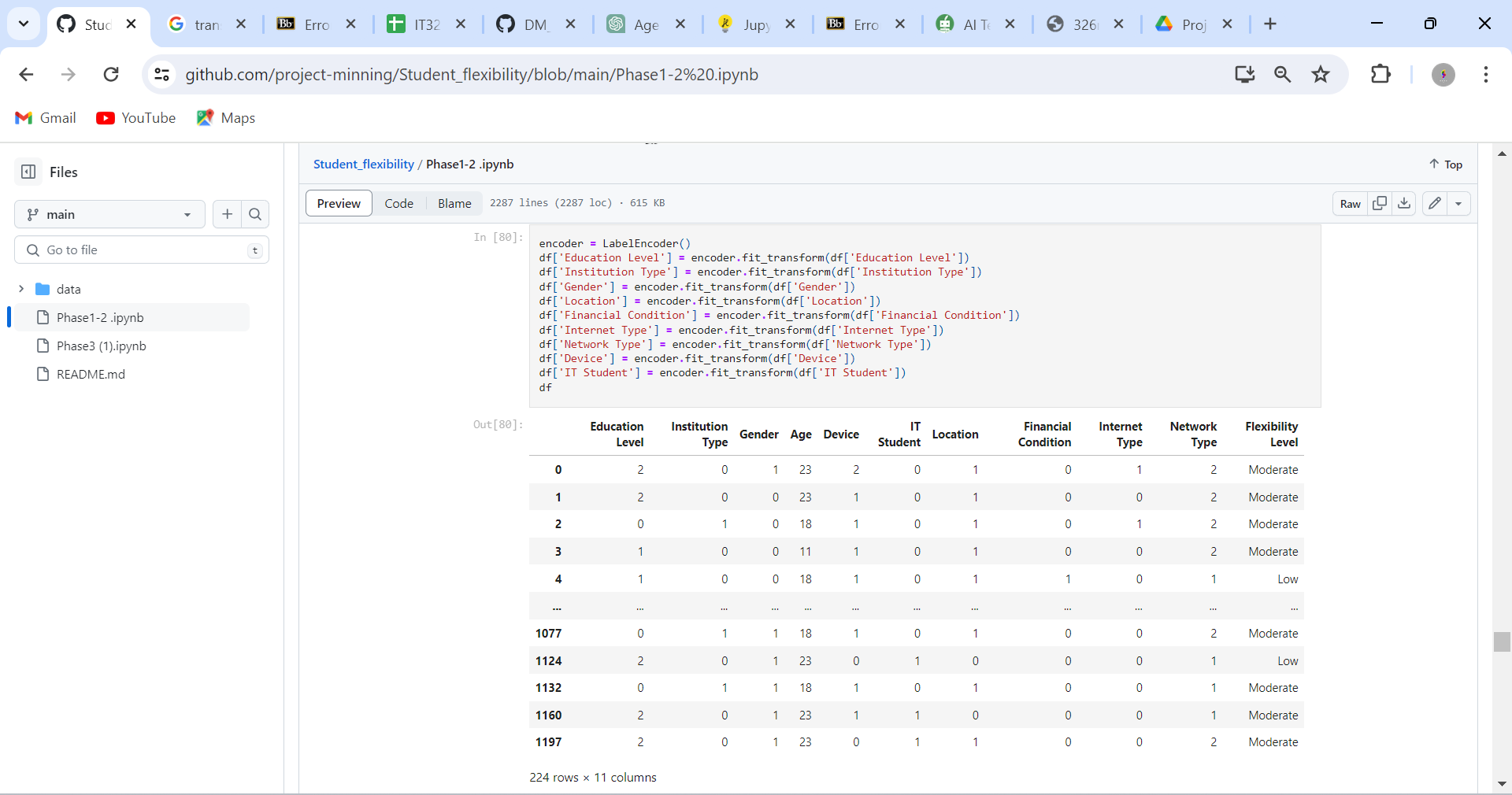
data Our dataset after Removing duplicates

**A screenshot of a computer

Description automatically generated**

**Data transformation:**

**Data encoding**

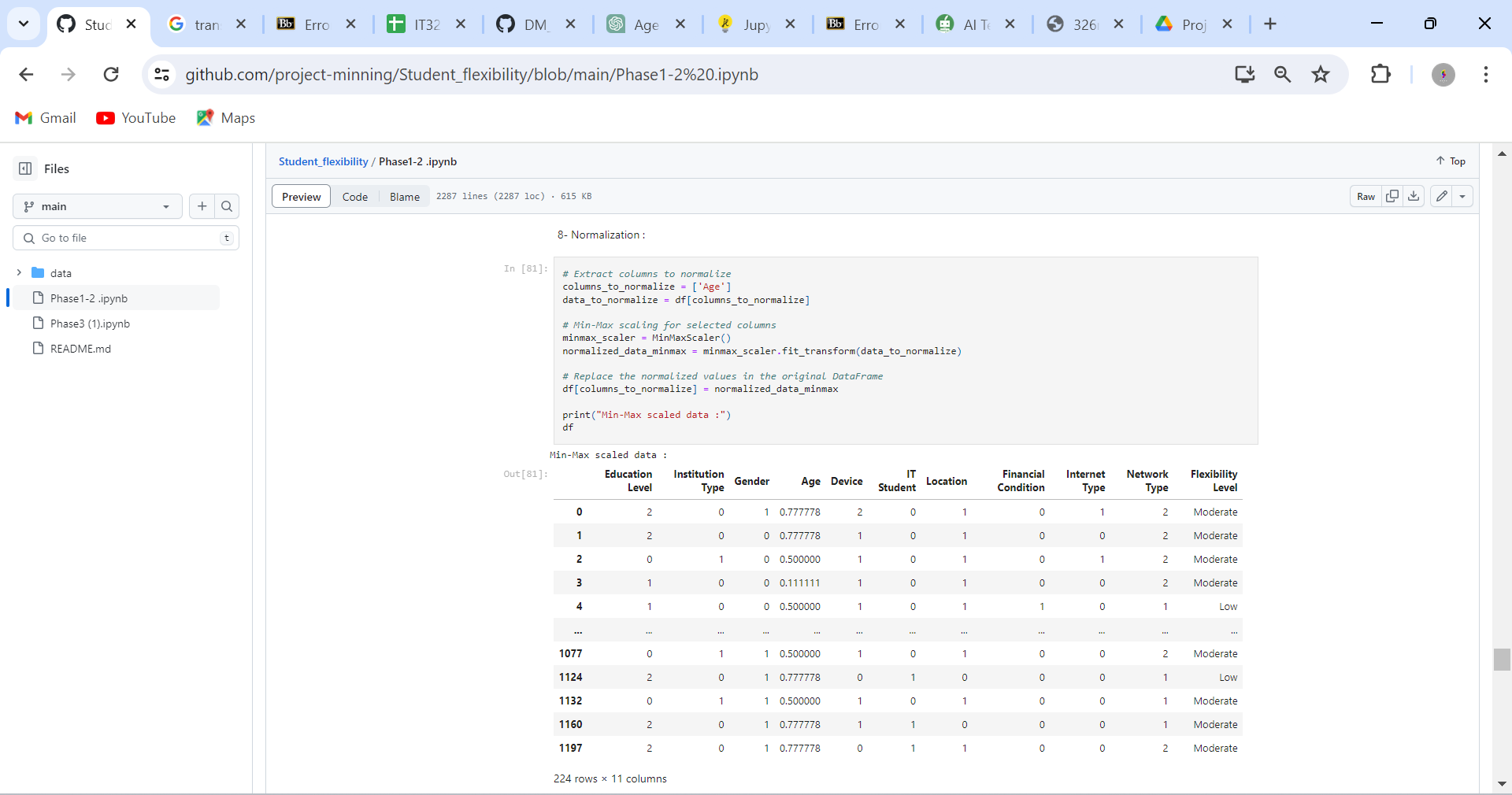
****

**Description**

We encoded variables such as education, institution, gender, location, financial condition, network type, device, and IT student status, Encoding categorical variables into numerical formats is crucial for machine learning models, enhancing predictive accuracy and performance. This simplifies the dataset, making it computationally efficient for analysis, and enhances data handling and modeling.

**Normalization** :

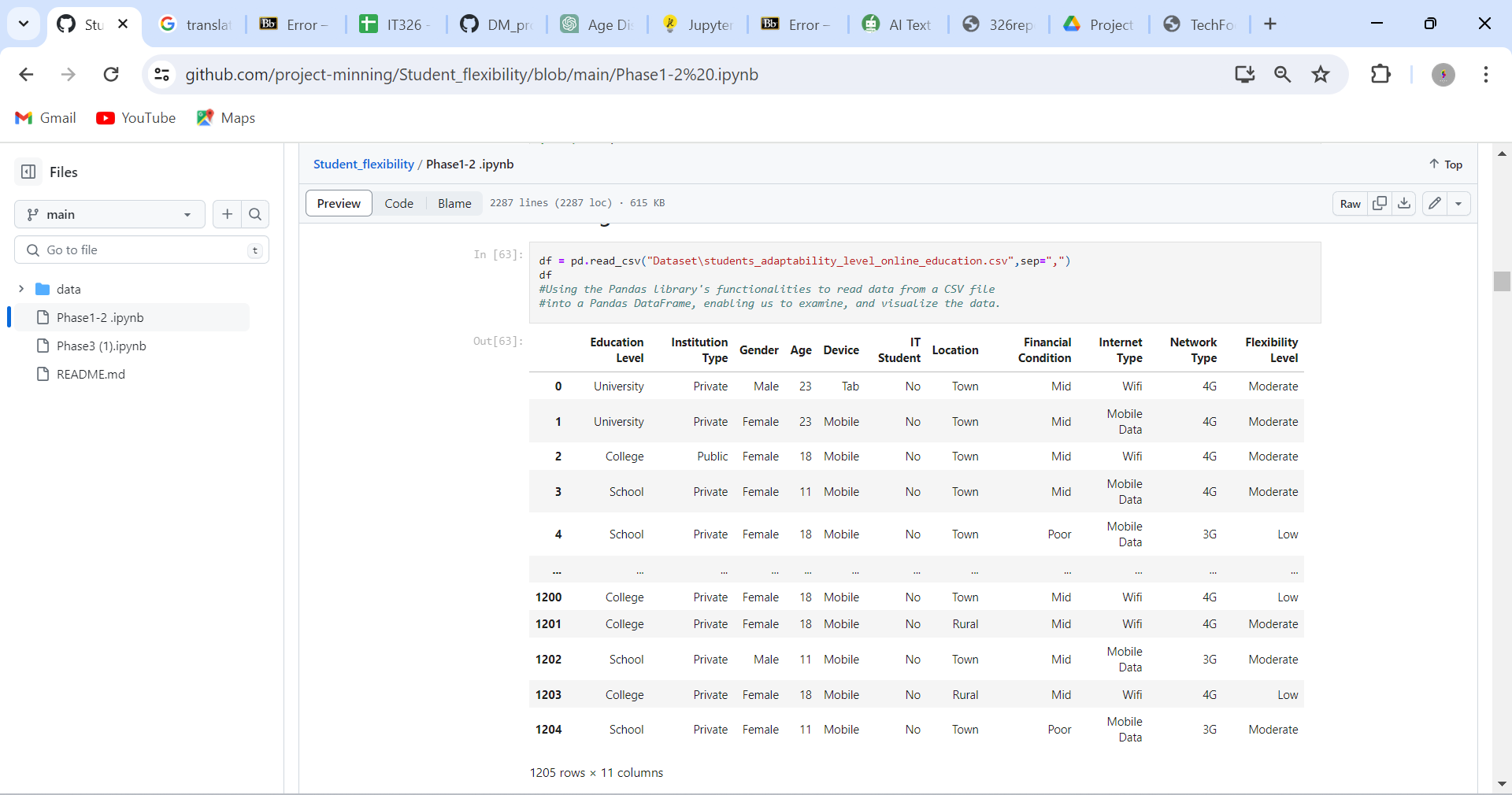
Data after normalization



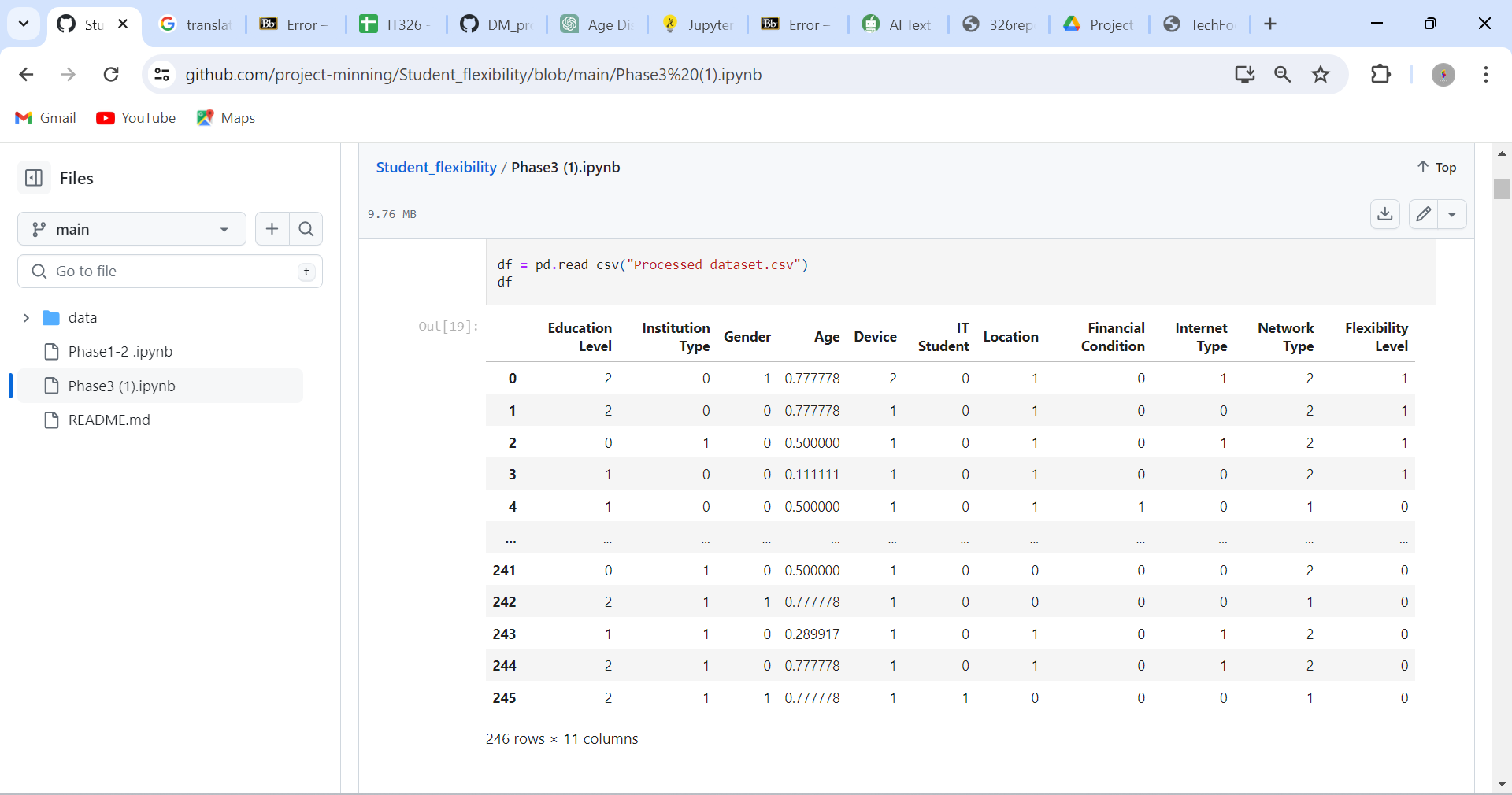
**Description**

We have normalized the age attribute to a uniform range, using Min-Max scaling to help us handle the data easily and to ensures that the age attribute has an equal opportunity to influence the outcomes.

**Raw data :**

****

**Data after processing :**

****

# Data Mining Technique

We utilized both supervised and unsupervised learning on our dataset, employing classification and clustering techniques.

For a classification, we used a decision tree. This recursive algorithm generates a tree with leaf nodes representing the final decisions. Our model will predict the class label (Flexibility Level) which has three classes: high, moderate and low, the prediction is based on the remaining attributes: Education Level, Institution Type, Gender, Age, Device, IT Student, Location, Financial Condition, Internet Type, Network Type. This technique includes dividing the dataset into two sets:

Training dataset: used for building the decision tree

Testing dataset: used to evaluate the constructed model.

Lastly, to assess our model, we evaluate the accuracy and cost-sensitive measures of the dataset using a confusion matrix.

We used (confusion matrix) method for evaluating the method.

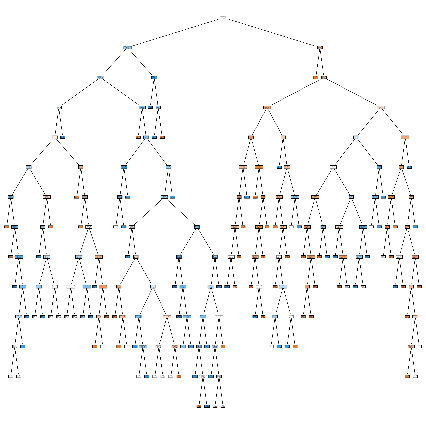
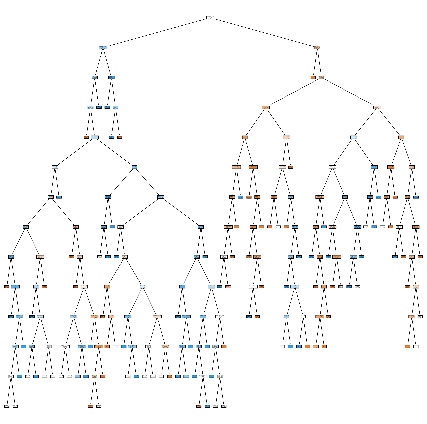
For clustering, since it's unsupervised learning, it doesn’t use a class label for implementing the

cluster thus we deleted the class label attribute “Flexibility Level "and used all other attributes in clustering (Education Level, Institution Type, Gender, Age, Device, IT Student, Location, Financial Condition, Internet Type, Network) we use the K-means clustering algorithm to group the students into clusters with different number of clusters. We evaluate the K-Means algorithms using silhouette coefficient and Total Within-Cluster Sum of Squares and we plot the Elbow curve to determine the optimal number of clusters.

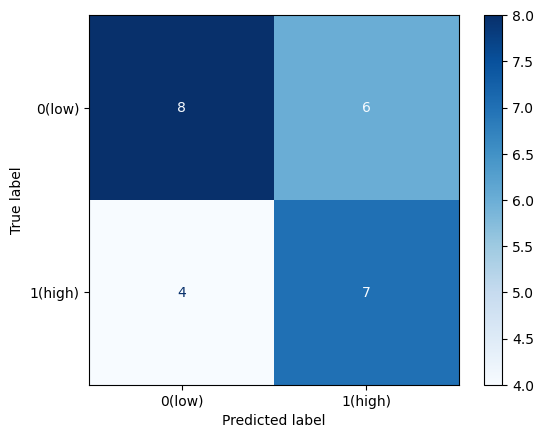
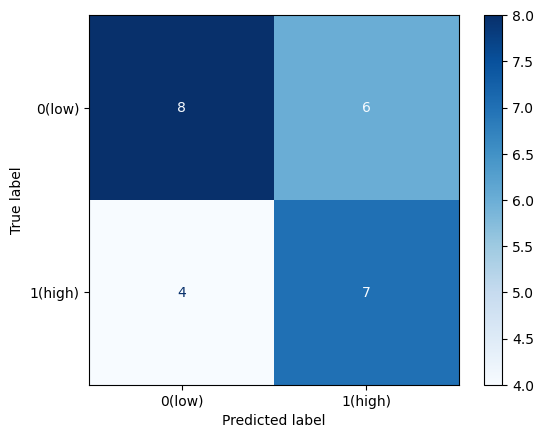
# Evaluation and Comparison

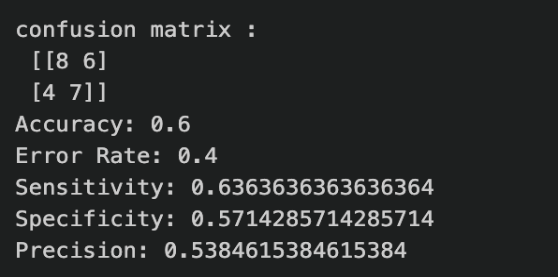
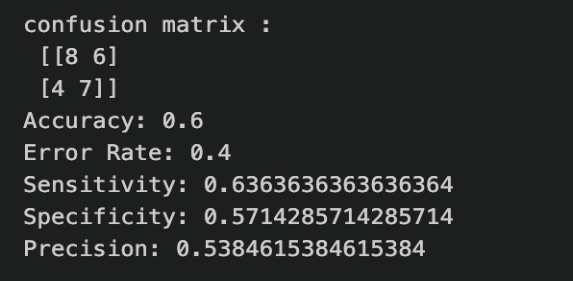
* **Classification**
* **Classification [90% training, 10% test]:**

**Figure (1) (decision tree)(entropy): Figure (2) (decision tree) (Gini Index):**



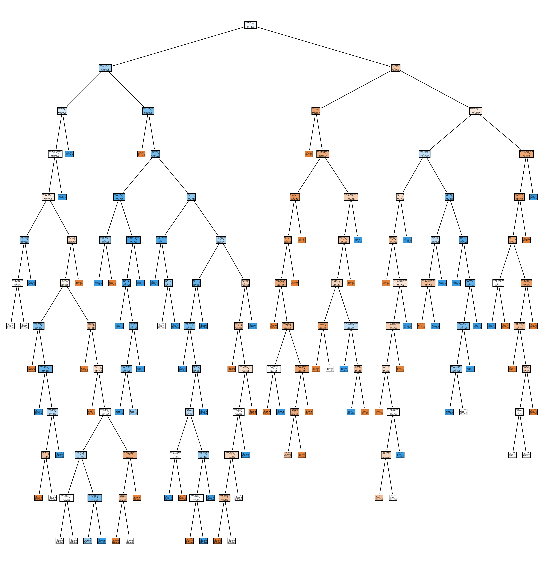
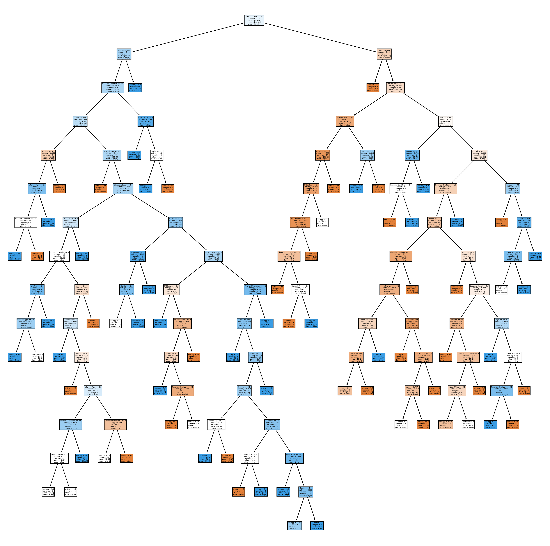
**Figure (1) (matrix confusion) (entropy): Figure (2) (matrix confusion) (Gini Index):**



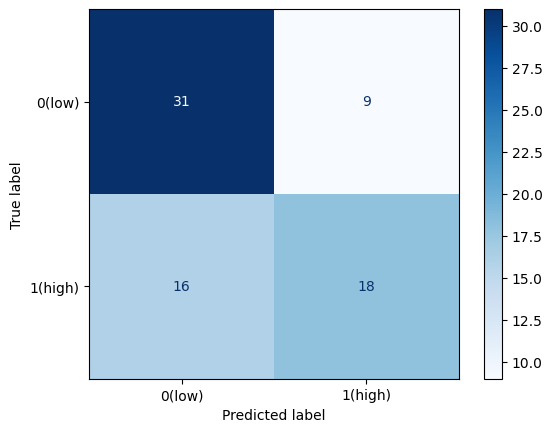
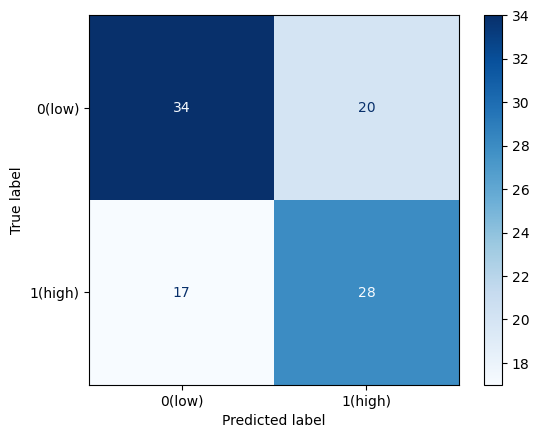


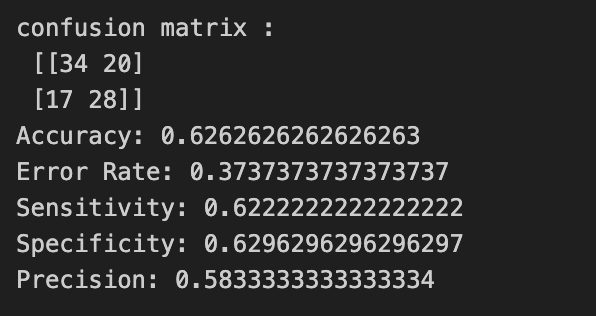
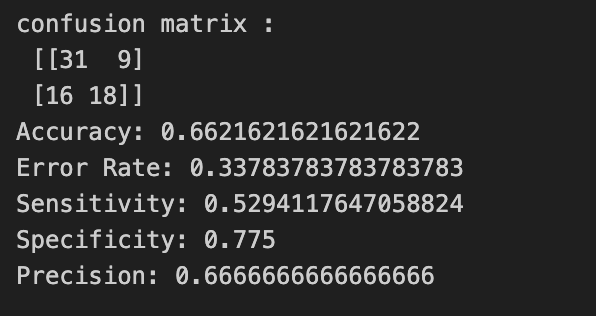
* **Classification [60% training, 40% test]:**

**Figure (1) (decision tree)(entropy): Figure (2) (decision tree) (GiniIndex):**



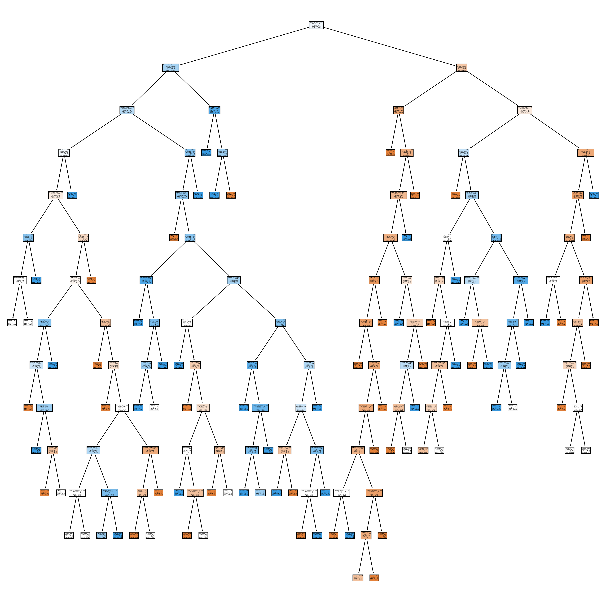
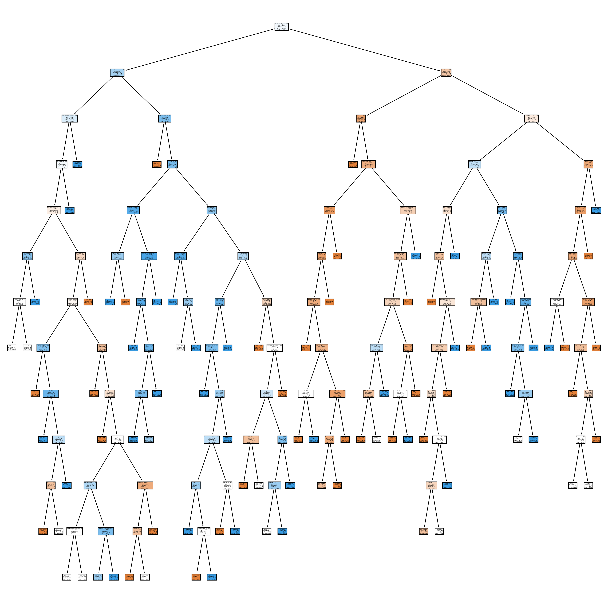
**Figure (1) (matrix confusion)(entropy): Figure (2) (matrix confusion) (Gini Index):**

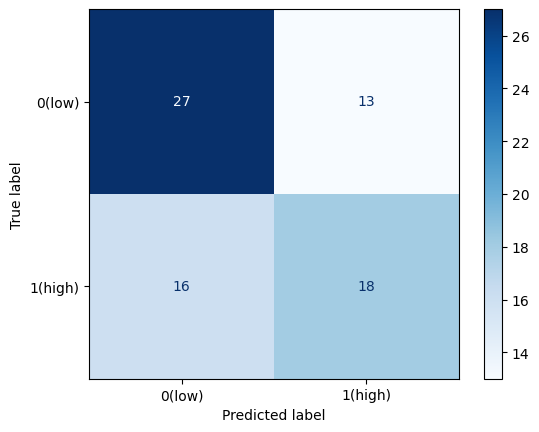
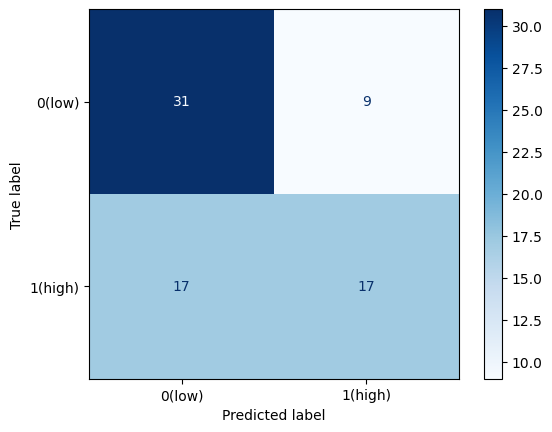


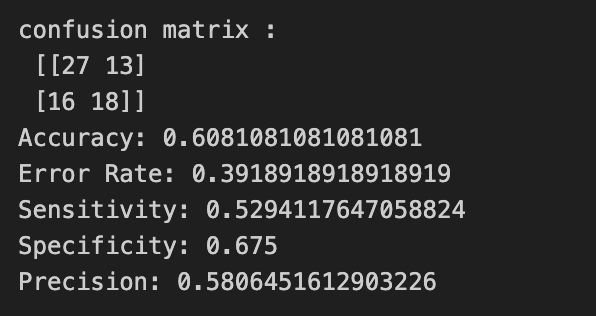
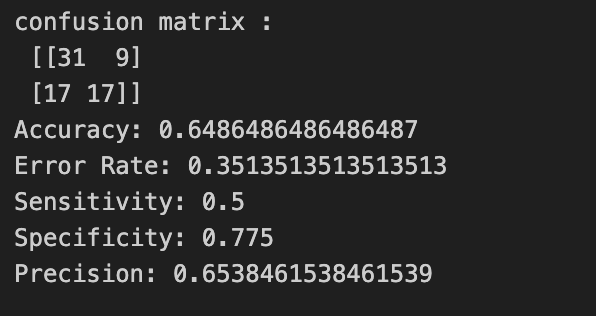


* **Classification [70% training, 30% test]:**

**Figure (1) (decision tree)(entropy): Figure (2) (decision tree) (Gini Index):**

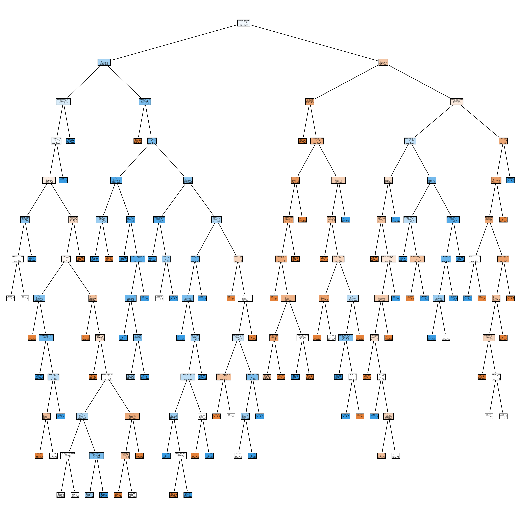
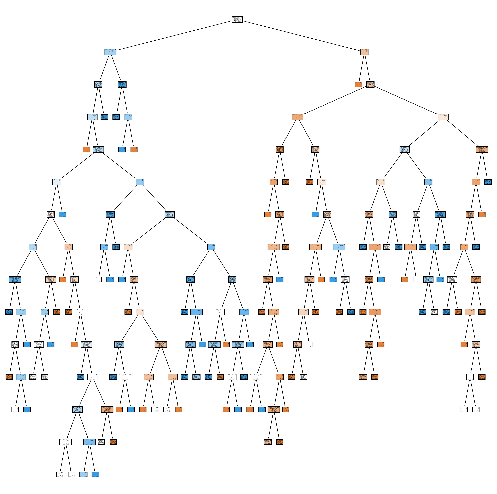
**Figure (1) (matrix confusion) (entropy) Figure (2) (matrix confusion)(Gini Index):** 



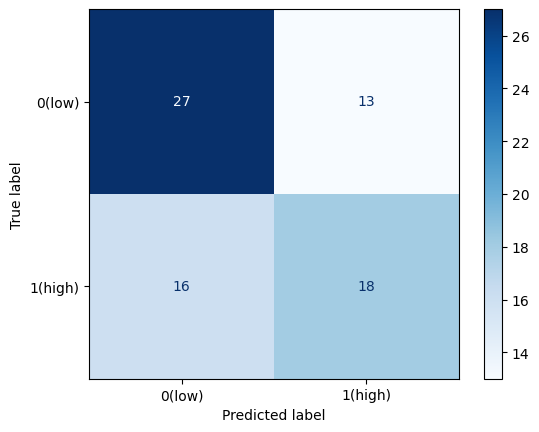
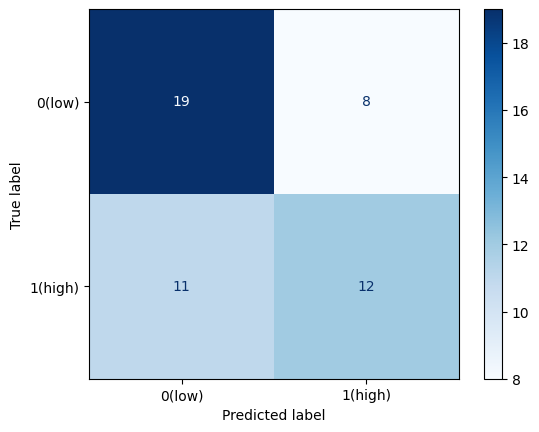


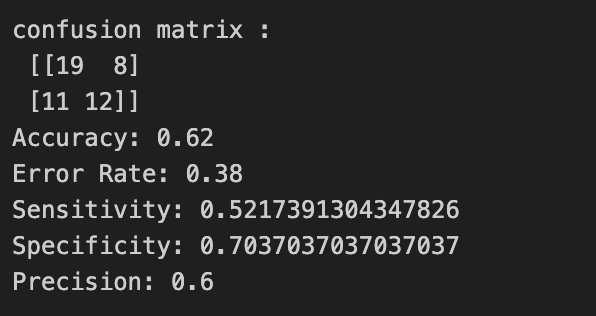
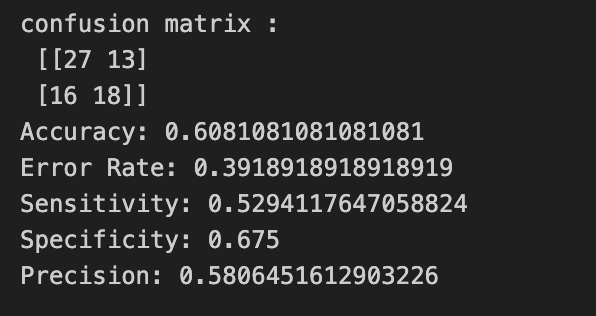
* **Classification [80% training, 20% test]:**

**Figure (1) (decision tree) (entropy): Figure (2) (decision tree) (Gini Index):**

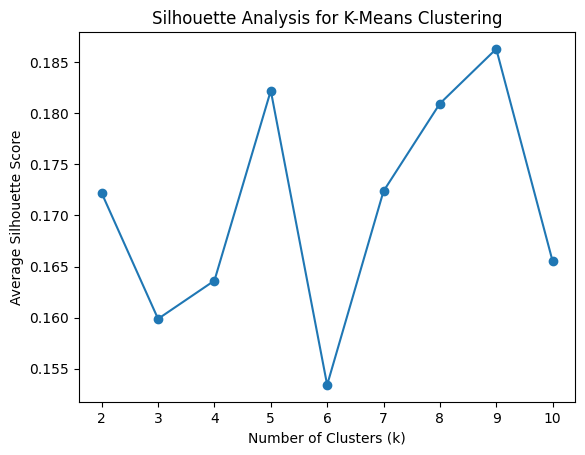


**Figure (1) (matrix confusion)(entropy): Figure (2) (matrix confusion) (GiniIndex):**

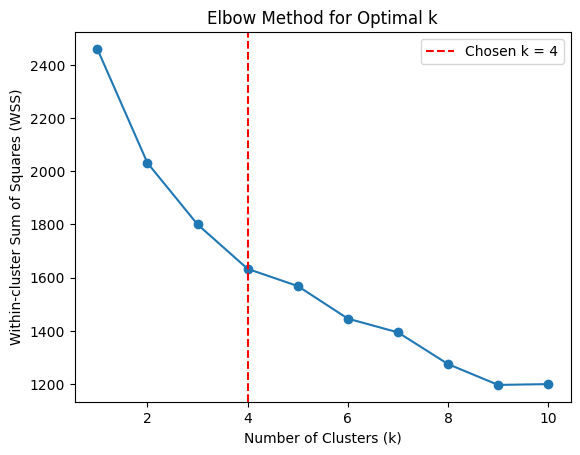




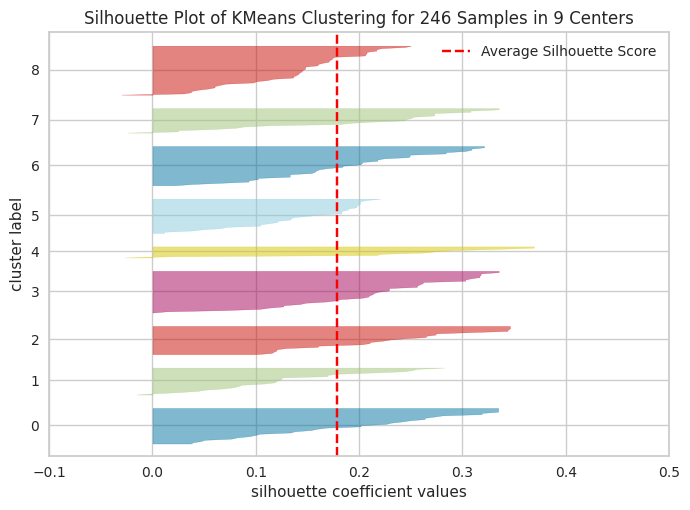
* **Clustering**
* Silhouette method (Silhouette Analysis):



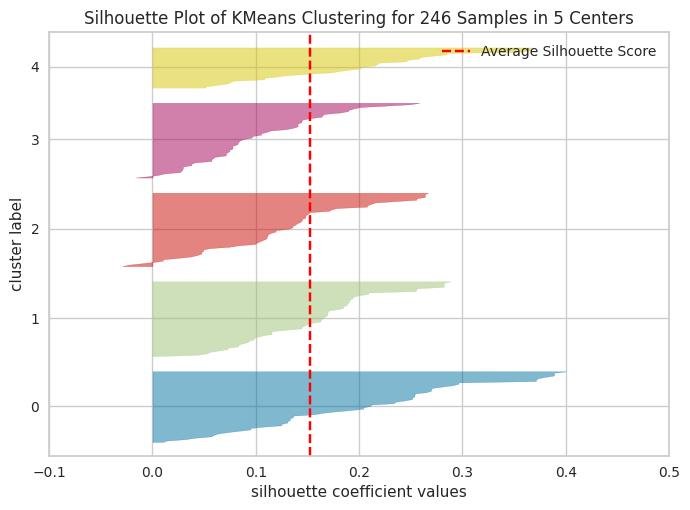
* Elbow method



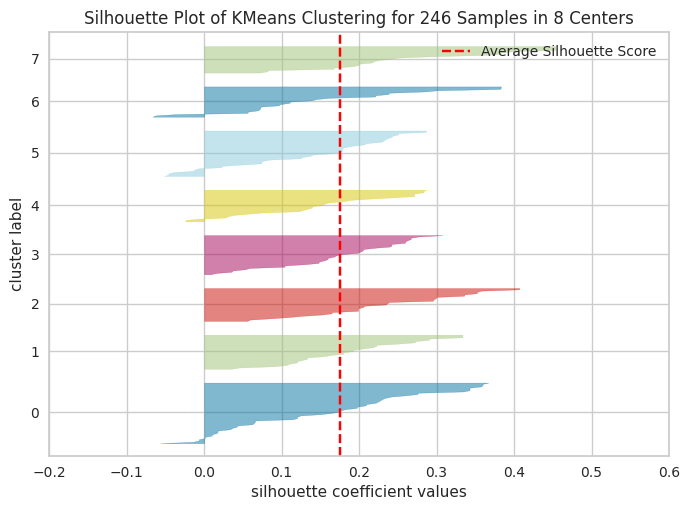
* **Clustering [K=9]:**



* **Clustering [K=5]:**



* **Clustering [K=8]:**



|  |  |  |
| --- | --- | --- |
| Mining task | Comparison Criteria |  |
| Classification |  |
| Clustering | |  |  |  |  | | --- | --- | --- | --- | |  | K=9 | K=5 | K=8 | | Average Silhouette width | 0.1784279148509518 | 0.1519666106128939 | 0.1749348910251044 | | total within-cluster sum of square | 1202.1795760464634 | 1550.551138764653 | 1250.7346543811886 | |

# Findings

**We have examinated the adaptability of students in online learning environments. With the surge in online education, notably accelerated by the COVID-19 pandemic, our project seeks to investigate students' data to identify the key factors impacting their adaptability.**

**By uncovering insights, to ensure our analysis of student flexibility is accurate, we used different methods to prepare the data well. We made plots like box plots and histograms to see the data clearly and decide what to do next. We got rid of any missing or unusual data that could mess up our results. Also, we changed some data to make it easier to work with and give each part of the data the same importance. These steps helped us understand and improve student flexibility in online learning.**

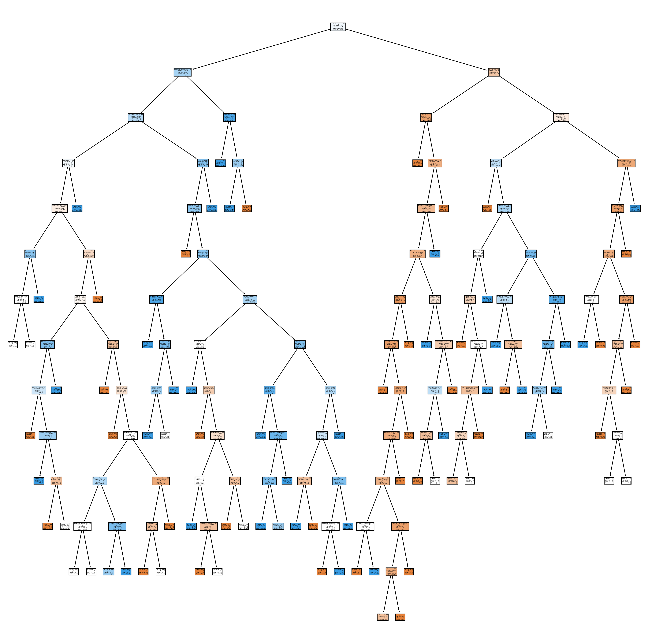
**As a result, we utilized data mining techniques, focusing on classification and clustering. For classification, we employed the decision tree method to build our model. We experimented with four different sizes of training and testing data to find the optimal setup for constructing and evaluating our model. Here are our findings:**

* **70% Training, 30% Testing: Accuracy = 0.64**
* **60% Training, 40% Testing: Accuracy = 0.626**
* **80% Training, 20% Testing: Accuracy = 0.62**
* **90% Training, 10% Testing: Accuracy = 0.6**

**The model achieving the highest accuracy, trained on 70% of the data and tested on 30%, stood out with an accuracy score of 0.64. This outcome suggests that this particular training-testing split yielded the most successful performance among the evaluated scenarios.**

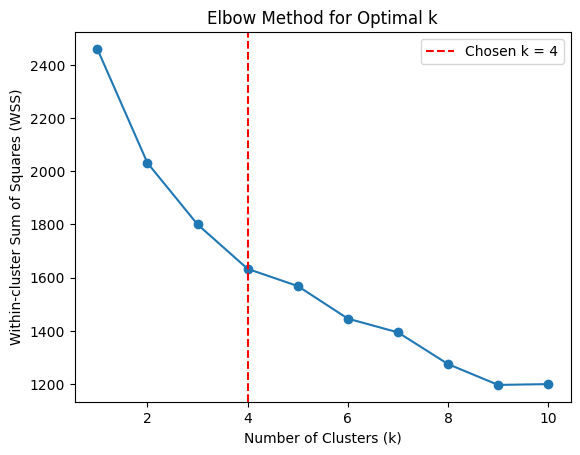
**The rationale behind its effectiveness likely stems from the equilibrium between the training and testing set sizes. By allocating 70% of the data for training, the model had ample information to discern complex patterns and nuances within the dataset. Meanwhile, the 30% reserved for testing facilitated rigorous evaluation without the risk of overfitting.**

**While other factors such as data quality and feature selection could have influenced the results, the 70-30 split emerged as the optimal configuration for maximizing accuracy in this context.**



* **Root Node: The root node (the top node) represents the entire dataset. This node splits the data based on the attribute that provides the highest information gain, which is calculated using entropy.**
* **Internal Nodes: Each internal node represents a decision point where the data is further split based on other attributes. The choice of attribute at each internal node is again determined by the attribute that provides the highest information gain.**
* **Branches: The branches represent the outcome of a decision at an internal node, leading to another internal node or a leaf node.**
* **Leaf Nodes: The leaf nodes (the nodes at the end of the branches) represent the final predictions of the model. In your case, these would be the predicted levels of “student flexibility”.**
* **Path: A path from the root node to a leaf node represents a rule. For example, if a path from the root to a leaf node passes through the decisions A=True, B=False, and C=True, then the rule is “If A is True, B is False, and C is True, then predict the class label at the leaf node”.**

For Clustering, we used K-means algorithm with 3 different K to find the optimal number



of clusters, we calculated the average silhouette width for each K, and we concluded the following results:

* For 𝐾 =9:

Average Silhouette Width: 0.1784

Within-Cluster Sum of Squares (WSS): 1202.18

Interpretation: With K =9, the clusters exhibit a relatively high average silhouette width, indicating well-defined clusters. Additionally, the WSS is relatively low, suggesting that the clusters are compact and tightly packed around their centroids.

* For 𝐾 =5:

Average Silhouette Width: 0.1520

Within-Cluster Sum of Squares (WSS): 1550.55

Interpretation: With K =5, the average silhouette width is lower compared to K=9K=9, indicating less well-defined clusters. The WSS is relatively high, suggesting that the clusters are less compact compared to K=9K=9, with more spread-out data points within each cluster.

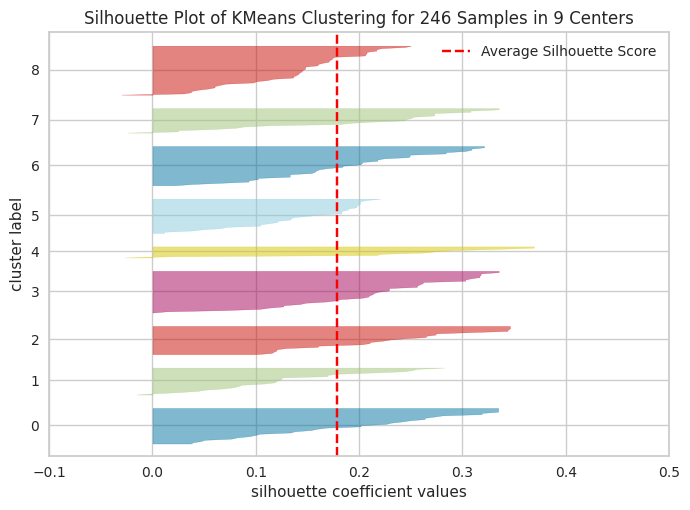
* For 𝐾 =8:

Average Silhouette Width: 0.1749

Within-Cluster Sum of Squares (WSS): 1250.73

Interpretation: With K =8, the average silhouette width is higher compared to K =5 but slightly lower than K =9. The WSS is intermediate between the values for K =5 and K =9, indicating moderately compact clusters.

In summary, each value of K yields different clustering performance metrics. 𝐾 =9 appears to result in the most well-defined and compact clusters based on both average silhouette width and WSS, followed by K =8 and then K =5.



**in conclusion, both models have proven valuable in predicting the level of flexibility exhibited by students, thereby contributing significantly to our overarching goal of assisting individuals in adapting to online learning environments. However, given that our dataset includes a class label "student flexibility," supervised learning models, particularly classification models, are deemed more accurate and suitable for application.**

**Supervised learning approaches are more accurate than unsupervised learning model(clustering), as the expected output is known beforehand this way we make use of the class label attribute. we harness this existing knowledge to refine the accuracy and relevance of our predictive models, empowering students to make informed decisions about their learning strategies and adaptability in online educational settings.**

# References

*-* [*https://www.kaggle.com/datasets/shariful07/student-flexibility-in-online-learning*](https://www.kaggle.com/datasets/shariful07/student-flexibility-in-online-learning)

* *King Saud university - IT326 lab*
* [*https://lms.ksu.edu.sa/bbcswebdav/pid-9410056-dt-content-rid-147775017\_1/courses/Merged\_IT326\_74557\_52846\_11\_452/Lab\_week%232\_Python\_Introduction.pdf*](https://lms.ksu.edu.sa/bbcswebdav/pid-9410056-dt-content-rid-147775017_1/courses/Merged_IT326_74557_52846_11_452/Lab_week%232_Python_Introduction.pdf)
* [*https://lms.ksu.edu.sa/bbcswebdav/pid-9443189-dt-content-rid-148159257\_1/courses/Merged\_IT326\_74557\_52846\_11\_452/Lab\_week%233\_Data%20Exploration%20and%20Visualization%20using%20Python.pdf*](https://lms.ksu.edu.sa/bbcswebdav/pid-9443189-dt-content-rid-148159257_1/courses/Merged_IT326_74557_52846_11_452/Lab_week%233_Data%20Exploration%20and%20Visualization%20using%20Python.pdf)
* [*https://lms.ksu.edu.sa/bbcswebdav/pid-9536829-dt-content-rid-148887652\_1/courses/Merged\_IT326\_74557\_52846\_11\_452/Data%20Preprocessing%20-%20Python%283%29.pdf*](https://lms.ksu.edu.sa/bbcswebdav/pid-9536829-dt-content-rid-148887652_1/courses/Merged_IT326_74557_52846_11_452/Data%20Preprocessing%20-%20Python%283%29.pdf)
* [*https://lms.ksu.edu.sa/bbcswebdav/pid-9577926-dt-content-rid-150572431\_1/courses/Merged\_IT326\_74557\_52846\_11\_452/Lab7%20Classification-%20Python%281%29.pdf*](https://lms.ksu.edu.sa/bbcswebdav/pid-9577926-dt-content-rid-150572431_1/courses/Merged_IT326_74557_52846_11_452/Lab7%20Classification-%20Python%281%29.pdf)
* [*https://lms.ksu.edu.sa/bbcswebdav/pid-9595752-dt-content-rid-151204445\_1/courses/Merged\_IT326\_74557\_52846\_11\_452/Clustering\_Python.pdf*](https://lms.ksu.edu.sa/bbcswebdav/pid-9595752-dt-content-rid-151204445_1/courses/Merged_IT326_74557_52846_11_452/Clustering_Python.pdf)