ANALYZING THE STUDENT PERFORMANCE USING CLUSTERING TECHNIQUES

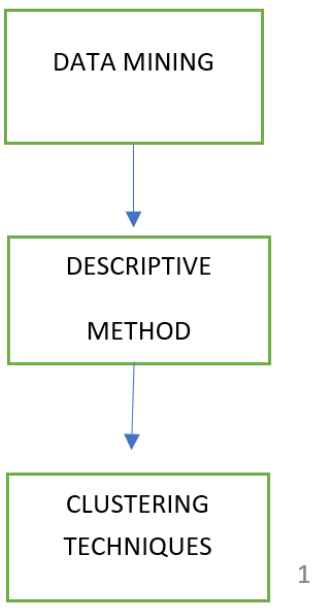
Vasuki Balasubramanian

MS(CSE)

Hood college

ABSTRACT

Data mining is called knowledge discovery in data. Exploration and discovering datasets are data mining. Over two decades, the data mining process involves several steps from data collection to visualization to extract valuable information from large data sets. As mentioned above, data mining techniques are used to generate descriptions and predictions about a target data set. In this project, students performance is performed using clustering techniques. It is very important to explore the factors that affect as well as reason for failures in studies.



1

CLUSTERING:

In clustering techniques involved with so many algorithms like centroid based algorithms, connectivity-based algorithms, distribution-based algorithms, and grid-based algorithms.

DATA CLEANING:

```
1 stu.isna().sum() #Detecting
```

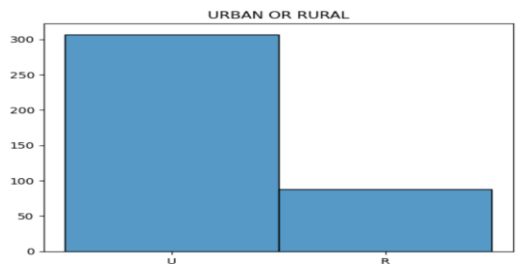
```
school        0
sex           0
age           0
address       0
famsize       0
Pstatus       0
Medu          0
Fedu          0
Mjob          0
Fjob          0
reason        0
guardian      0
traveltime    0
studytime     0
```

```
1 stu.duplicated().sum()
```

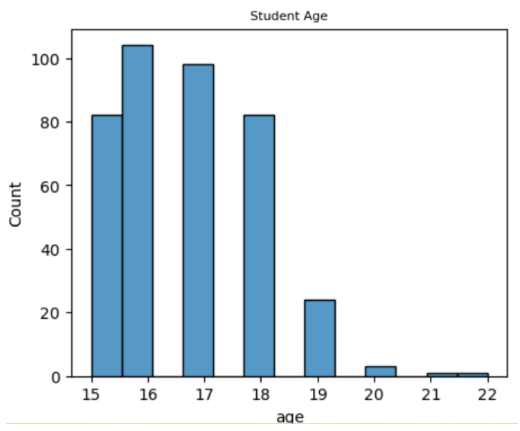
```
0
```

DATA VISUALIZATION:

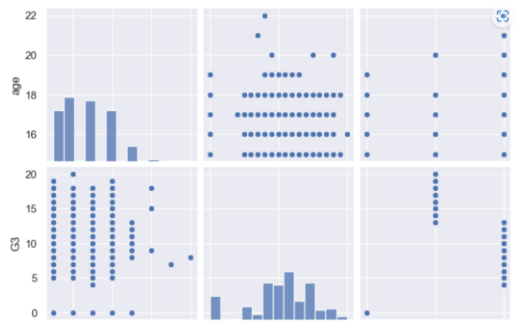
When we talk about data visualization, it is about representation of data using common graphics, plots, charts, heatmap, box plot, line chart, scatter plots002E



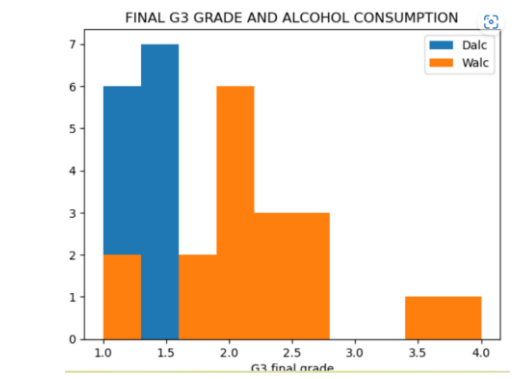

```
#visualization of students age from 17 to 22
plt.figure(figsize=(5, 4))
sns.histplot(stu["age"])
plt.title("Student Age",size=8)
plt.show()
```



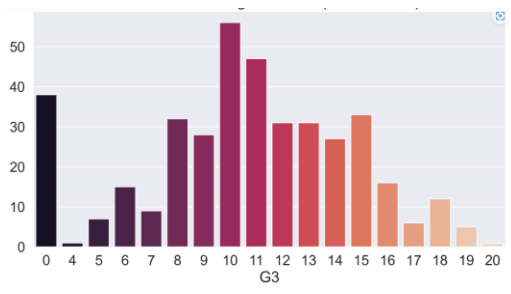
PAIRPLOT FIGURE:



VISUALIZATIONS:

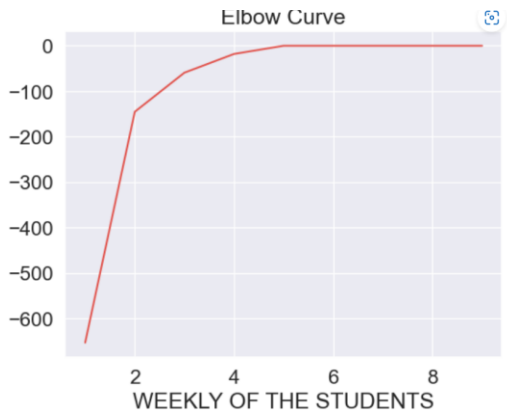


DISTRIBUTION OF TARGET VARIABLE:

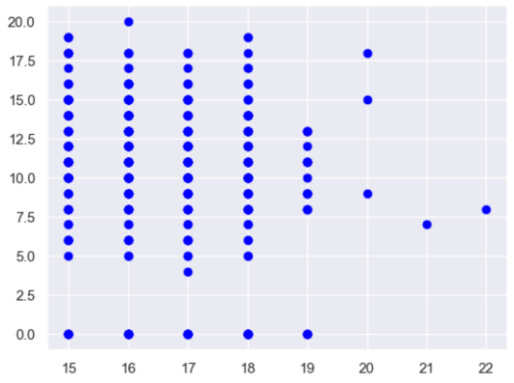


K-Means Clustering:

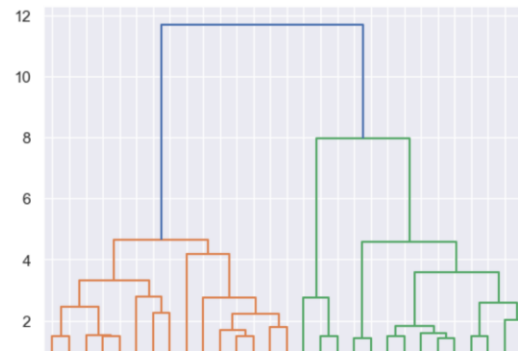
KMeans clustering is one of the lustering models in datamining. Elbow curve is one of the methods used in Kmeans clustering. In the elbow method, the variation changes rapidly until the number of groups in the dataset.



SCATTERPLOT OF THE DATASET:



HIERACHICAL AGGLOMERATIVE CLUSTERING:



In the hierarchy clustering, we have used the bottom up-approach. They are usually represented in dendrogram model. When we speak about hierarchy clustering , it comes along with agglomerative clustering

DBSCAN CLUSTERING:

```
The cophenet coefficient is 0.8424224552530065 for the method average
The cophenet coefficient is 0.8424144237138613 for the method centroid
The cophenet coefficient is 0.7889015988193281 for the method ward
The cophenet coefficient is 0.5642262481314225 for the method weighted
The cophenet coefficient is 0.5369659435168603 for the method median
```

$$d_{L2}(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

The given formulae is to find the Euclidean distance in the algorithm. DBSCAN clustering is Density based spatial clustering with noise.It is one of the famous clustering algorithm, unlike kmeans clustering. Using dbscan we can find the number of clusters and noise points, homogeneity, completeness, V- measure score, metrics adjusted rand score and silhouette coefficient.