![Text

Description automatically generated with low confidence]() Text

Description automatically generated

**Informatics Institute**

**of Technology**

Machine Learning &

Data Mining

5DATA002W

Coursework Report

Name : Rayaan Rilshad

IIT Student ID : 20211315

UoW ID : w1899330

Tutorial Group : R

# Partitioning Clustering Part

## 1st Subtask

### Pre-Processing

##### Outlier detection and removal

Outliers are data points that stand out from the rest of the data. They can happen because of different reasons like mistakes in data entry, measurement errors, or just natural variation in the data. Outliers can mess up machine learning, so it's important to find and remove them before training a model. There are different ways to find outliers. One way is to use z-scores. A z-score tells you how far away a data point is from the average of its group. If a data point has a z-score that's bigger than a certain number, it's an outlier. Another way to find outliers is to use boxplots. A boxplot is a picture that shows the middle, quarters, and outliers of the data.

Scaling and outlier detection/removal are two important techniques used in data precessing.

Scaling helps to normalize the values of different variables in a dataset to a similar scale, preventing one variable from having a more significant impact than others.

Outliers are values that significantly differ from other observations and can introduce bias into a model. Detecting and removing outliers is necessary to ensure the model's accuracy and robustness.

In this code, the data preprocessing steps are performed as follows:

* The "Class" and "Samples" columns are removed from the dataset.
* Null values are checked and removed from the dataset.
* Z-scores are calculated for each variable, which is a measure of how many standard deviations away a data point is from the mean.
* Outliers are identified using a threshold of 3 standard deviations from the mean, and a boxplot is created to visualize the outliers.
* Outliers are removed from the dataset.
* Finally, the dataset is scaled to ensure that all variables have a similar range and contribute equally to the model.

### Defining the number of Cluster Centres

#### Clustering using manual tools

#### Clustering using Automated Tools

##### NbClust Method

##### Elbow Method