**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**

**GRAPHIC ERA DEEMED UNIVERSITY**

**DEHRADUN**

**2021-2022**



PROJECT REPORT

ON

ANOMALY DETECTION TECHNIQUES AND ALOGRITHMS

*SUBMITTED BY: RIYA*

*ROLL NO.: 2016961*

*SECTION : H*

DECLARATION

I, Riya student of B-Tech(CSE),semester 4,Department of computer science and engineering,Graphic Era Deemed University,Dehradun,declare that the technical project work entitled “ANOMALY DETECTION TECHQUNES AND ALOGRITHMS”

Has been carried out by me and submitted n partial fulfilment of the course requirements for the award of the degree during the academic year 2021-2022.

The matter embodied in this synopsis has not been submitted to any other university or institution for the award of any other degree or diploma.

DATE : 14 July 2022

PROBLEM STATEMENT

To detect the anomalies from the given datasets .

MOTIVATION

It is critical for network admis to be able to identify and react to changing operational conditions. Any nuances in the operational conditions of Data centers or cloud applications can signal unacceptable levels of business risk.On the other hand ,some divergences may point to positive growth.

Therefore,anomaly detection is central to extracting essential business insights and maintaining core operations.all of the demand the ability to discern between normal and abnormal behaviour precisely and correctly.

**What are Anomalies?**

Anomalies are data points that stand out amongst other data points in the dataset and do not confirm the normal behavior in the data. These data points or observations deviate from the dataset’s normal behavioral patterns.

Anomaly detection is an unsupervised data processing technique to detect anomalies from the dataset. An anomaly can be broadly classified into different categories:

* **Outliers:** Short/small anomalous patterns that appear in a non-systematic way in data collection.
* **Change in Events:** Systematic or sudden change from the previous normal behavior.
* **Drifts:** Slow, undirectional, long-term change in the data.

Anomalies detection are very useful to detect fraudulent transactions, disease detection, or handle any case studies with high-class imbalance.

**THE PROJECT**

GETTING THE DATASETS:-

Datasets used in the projects are the three different csv files.

1: weight-height.csv

2:heights.csv

3:heights.csv

Let’s start with importing the required modules.

**Install Requirements :** Install **Python3**

Install **pandas** as pd

Install **numpy** as np

Install **matplotlib.pyplot** as plt libraries.

1.import four modules as shown above .

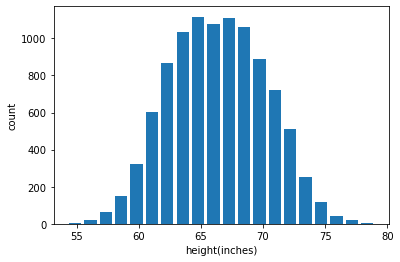
**TOOLS USED:**

1.**JUPYTER NOTEBOOK**

1.**READ THE CSV FILE**

2.predict the outliers using three different techq.

i) Outliers detection and removal using z-score and standard deviation in python pandas (plot the hist.graph)



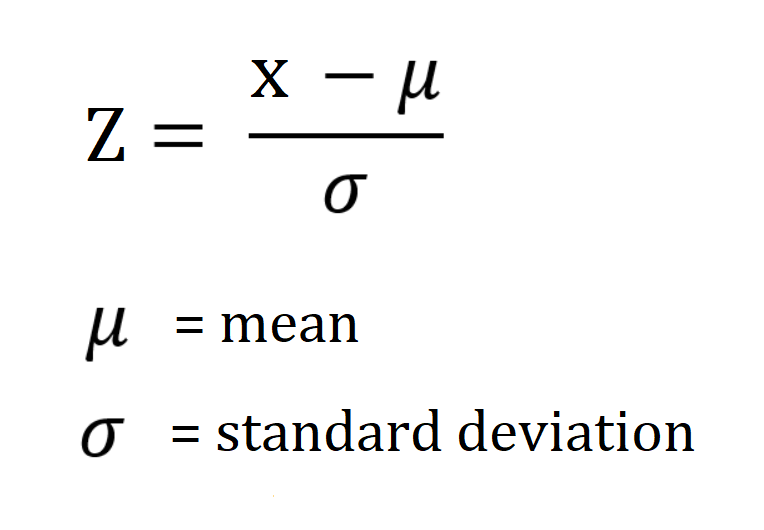
Outlier detection and removal using **Z score**

z score indicates how many standard deviation away a data point is.

(for example: in our case mean is 66.37 and standard deviation is 3.84.

If the value of a data point is 77.91 the Z score for that is 3 because it is 3 standard deviation away (77.91=66.37+3\*3.84))

**CALCULATE THE Z SCORE**



II) **OUTLIER USING IQR(INTERQUARTILE RANGE**)

The interquartile range defines the difference between the third and the first quartile. Quartiles are the partitioned values that divide the whole series into 4 equal parts. So, there are 3 quartiles. First Quartile is denoted by Q1known as the lower quartile, the second Quartile is denoted by Q2 and the third Quartile is denoted by Q3 known as the upper quartile. Therefore, the interquartile range is equal to the upper quartile minus lower quartile.

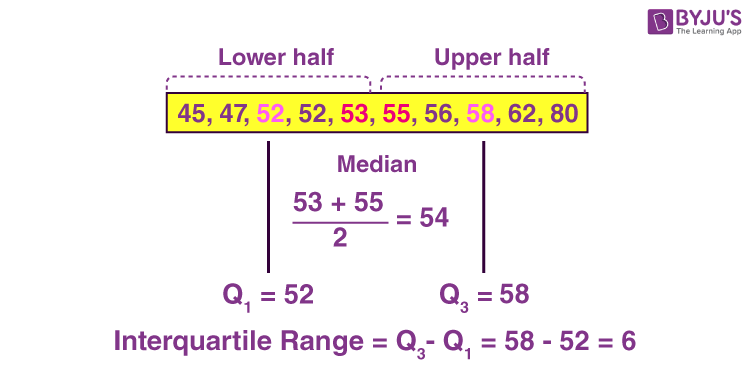
## **Interquartile Range Formula**

The difference between the upper and lower quartile is known as the interquartile range. The formula for the interquartile range is given below

**Interquartile range = Upper Quartile – Lower Quartile = Q­3 – Q­1**

where Q1 is the first quartile and Q3 is the third quartile of the series.

The below figure shows the occurrence of median and interquartile range for the data set.



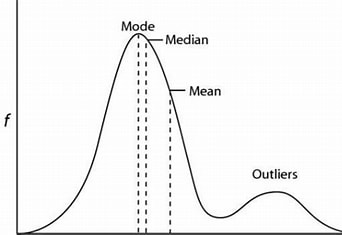
### **How to Calculate the Interquartile Range?**

The procedure to calculate the interquartile range is given as follows:

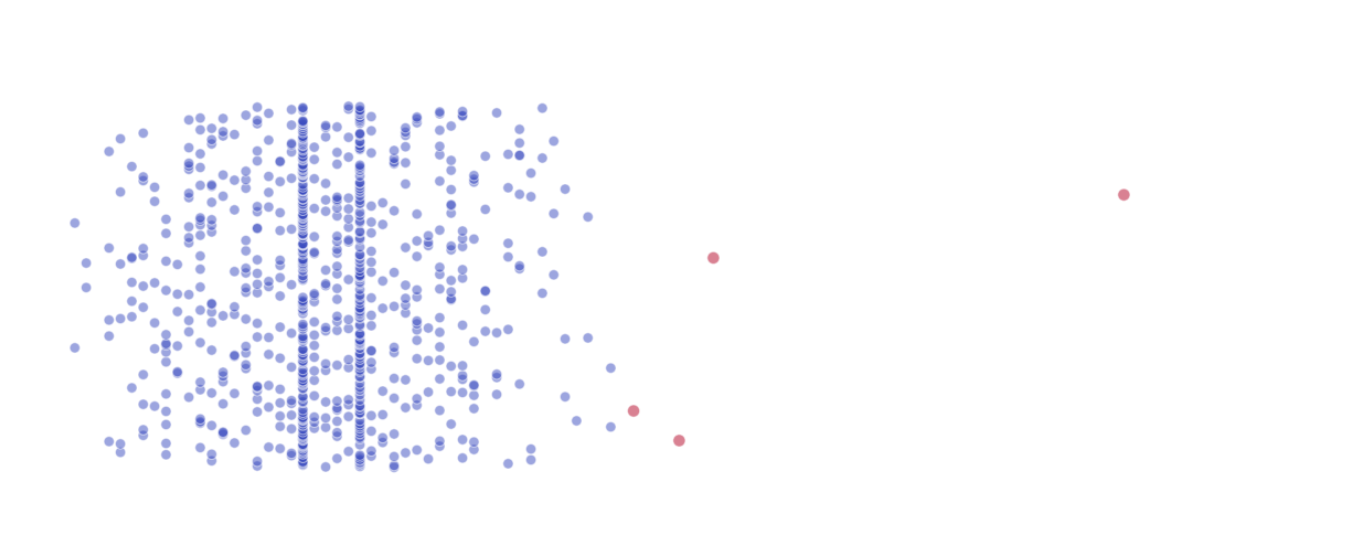
* Arrange the given set of numbers into increasing or decreasing order.
* Then count the given values. If it is odd, then the center value is median otherwise obtain the mean value for two center values. This is known as Q2value. If there are even number of values, the median will be the average of the middle two values.
* Median equally cuts the given values into two equal parts. They are described as Q1 and Q3 parts.
* The median of data values below the median represents Q1.
* The median of data values above the median value represents Q3.
* Finally, we can subtract the median values of Q1 and Q3.
* The resulting value is the interquartile range.

III)**OUTLIERS USING PERCENTILE**

You can simply fix a percentile for the upper limit and lower limit.



A percentile indicates the**value below which a given percentage of observations in a group of observations fall**. Think of sorting data set containing 100 values and dividing it in 100 equal parts, now the value at any place, say at 10th is our 10th percentile, i.e. value at index 10 indicates below which 10% of values fall.



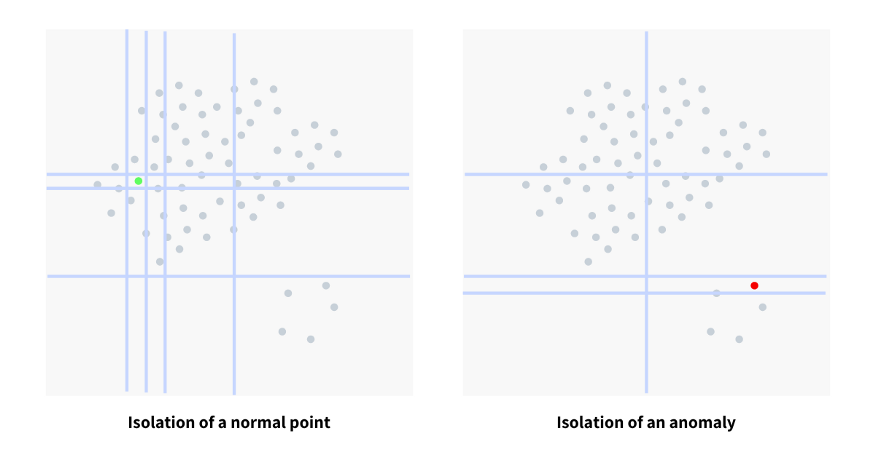
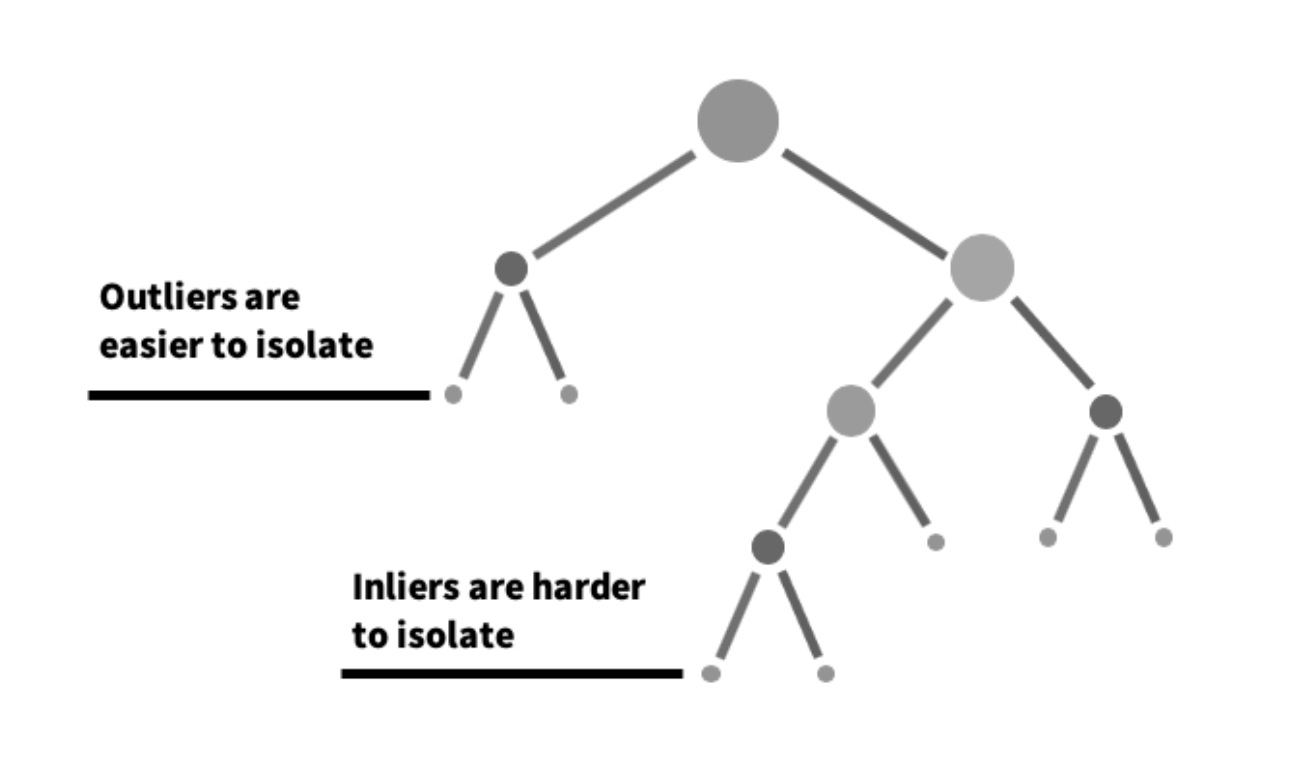
**OUTLIERS ARE SEEN AS SCATTER POINTS**

**ALORITHMS USED :-**

# Isolation Forest:

Isolation Forest is an unsupervised anomaly detection algorithm that uses a random forest algorithm (decision trees) under the hood to detect outliers in the dataset. The algorithm tries to split or divide the data points such that each observation gets isolated from the others.

Usually, the anomalies lie away from the cluster of data points, so it's easier to isolate the anomalies compare to the regular data points.

From the above-mentioned images, it can be observed that the regular data points require a comparatively larger number of partitions than an anomaly data point.

The anomaly score is computed for all the data points and the points anomaly score > threshold value can be considered as anomalies.

# One Class SVM:

A regular SVM algorithm tries to find a hyperplane that best separates the two classes of data points. For one-class SVM where we have one class of data points, and the task is to predict a hypersphere that separates the cluster of data points from the anomalies.

# Conclusion:

Anomaly detection algorithms are very useful for fraud detection or disease detection case studies where the distribution of the target class is highly imbalanced. Anomaly detection algorithms are also to further improve the performance of the model by removing the anomalies from the training sample.

Apart from the above-discussed machine learning algorithms, the data scientist can always employ advanced statistical techniques to handle the anomalies.

**REFERENCES**

1.Youtube

2.Google