



# **ANOMALY DETECTION**

<b>Date</b>	<b>02/12/15</b>
<b>Version</b>	<b>1.0</b>
<b>Author</b>	<b>Ranjana Raveesh</b>
<b>Matriculation Number</b>	<b>1135398</b>
<b>Project Name</b>	<b>Anomaly Detection Web Portal (Project 7 – Code: ML-KMEANWEB-00)</b>
<b>File Location</b>	<b>\$/SE-2015-2016/StudentProjects/ANOMALY- DETECTION/Ranjana/AnoDetWebApp</b>

<b>ANOMALY DETECTION.....</b>	<b>1</b>
PURPOSE.....	4
DOCUMENTATION .....	4
BACKGROUND .....	5
<i>Anomaly Detection:</i> .....	5
<i>k-means clustering</i> .....	5
REQUIREMENT DESCRIPTION .....	6
<i>Display Cluster Statistics</i> .....	6
<i>Graphical representation of cluster</i> .....	6
FUNCTIONAL AND DESIGN DESCRIPTION.....	7
<i>Home</i> .....	7
<i>Scatter 2D Graph</i> .....	7
<i>Scatter 3D Graph</i> .....	8
<i>Cluster Detail</i> .....	9
<i>Outliers Detail</i> .....	9
<i>Check Samples</i> .....	10
<i>Clustering Algorithms for k-means</i> .....	10
FLOWCHART.....	12
<i>Scenario</i> .....	13
TECHNICAL DETAILS .....	13

## Purpose

The purpose of this Project is to give user a web interface which can be used to present the result of The Anomaly Detection API in a very user friendly and interesting way using Graphical representation and different tabular formats to display the Cluster statistics. It is also gives an to perform different tasks related to cluster, like uploading new data, updating existing data and changing the number of clusters and running clustering over them.

## Documentation

Title	Location	Vsn	Author
AnomalyDetectionSRS.doc	\$/SE-2015-2016/StudentProjects/ANOMALY-DETECTION/Ranjana/AnoDetWebApp/WebApplication/Document	1.0	Ranjana

## Background

### Anomaly Detection:

#### Ideal Prediction Foresees Failures Before They Occur

We are all witnessing the current explosion of data: social media data, clinical data, system data, CRM data, web data, and lately tons of sensor data! With the advent of the Internet of Things, systems and monitoring applications are producing humongous amounts of data which undergo evaluation to optimize costs and benefits, predict future events, classify behaviors, implement quality control, and more. All these use cases are relatively well established by now: a goal is defined, a target class is selected, a model is trained to recognize/predict the target, and the same model is applied to new never-seen-before productive data.

The newest challenge now lies in predicting the “unknown”. The “unknown” is an event that is not part of the system past, an event that cannot be found in the system historical data. In the case of network data the “unknown” event can be an intrusion, in medicine a sudden pathological status, in sales or credit card businesses a fraudulent payment, and finally, in machinery, a mechanical piece breakdown. A high value, in terms of money, life expectancy, and/or time, is usually associated with the early discovery, warning, prediction, and/or prevention of the “unknown” and, most likely, undesirable event.

Specifically, prediction of “unknown” disruptive events in the field of mechanical maintenance takes the name of “**anomaly detection**”.

#### k-means clustering

In this project K-means clustering algorithm is used. We are using *k*-means clustering in unsupervised learning. The basic approach is first to train a *k*-means clustering representation, using the input training data. The training of data would form the *n* number of clusters. The K-means is the most important flat clustering algorithm. Its objective is to minimize the average squared Euclidean distance of the 2D, 3D or *n*D dataset from their cluster center also called centroid. In 3 dimensions we can imagine the ideal cluster in K-means as a sphere with the centroid as its center of gravity. Ideally, the clusters should not overlap. *k*-means clustering aims to partition *n* observations into *k* clusters. The number *k* is selected based on various methods. I am selecting value of *k* = 3, in 2 Dimensional data for the purpose of Market segmentation based on clusters on Height/Weight for the sizing purpose. In the 3 – Dimensional data of Age/weight/height to monitor the child's growth I am using *k*=3.

## Requirement Description

- ❖ Implement a simple portal which provides interesting cluster information.

Scope of the Project

### Display Cluster Statistics

- Distance to farther sample from cluster centroid.
- Distance to next nearest cluster.
- Distance between nearest samples of nearest clusters.

### Graphical representation of cluster

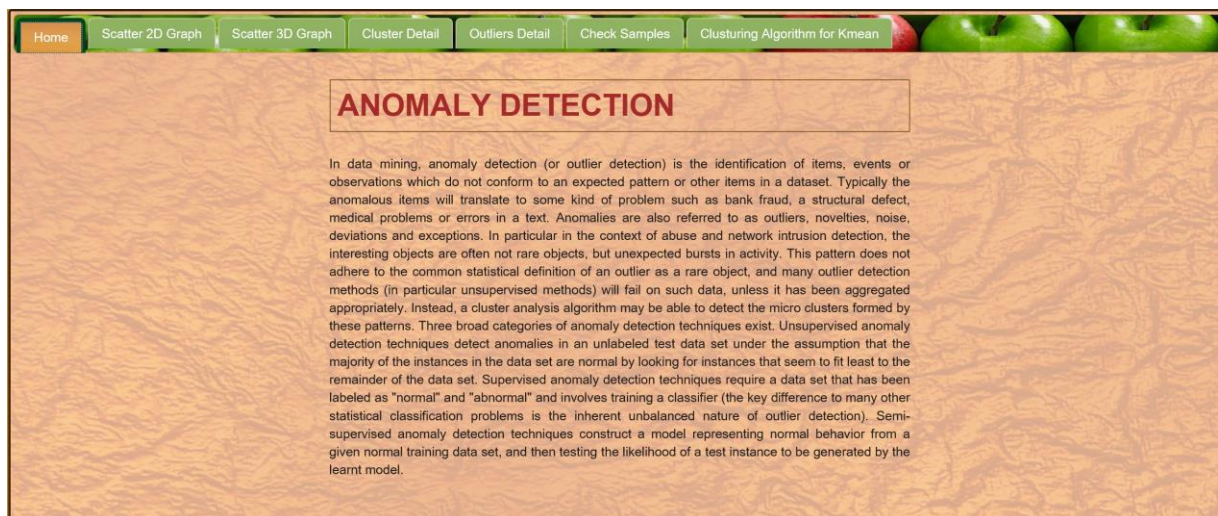
- Cluster-Centroids
- Farthest Sample from Centroid
- Nearest Clusters
- Nearest samples from two clusters.
- Show data in 2D and 3D by freezing remaining scalars.

## Functional and Design Description

There are total 7 Tabs in this Web Application.

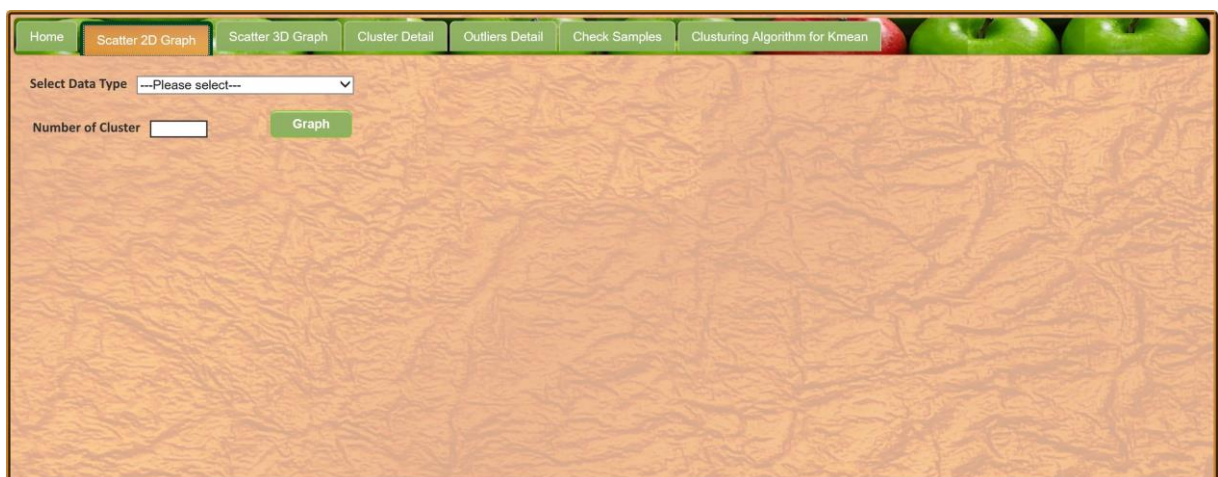


### Home



This Page gives the short Introduction to the Anomaly Detection.

### Scatter 2D Graph



On this page we can select different Data Set Type from the dropdown

Height/Weight
Age/Weight
Age in Month/Head Circumference
Age in Month/Length



and enter number of clusters and then should be able to see the two dimensional representation showing the different Clusters with their Centroids and outlier details.



On the above graph different clusters are shown in different colours. The centroids and outliers are also shown in different colours. We can view only selected data by clicking on desired Legends on the Graph. When we focus on the points on the graph the scalar values (in the above case – Height and Weight) values for that point.

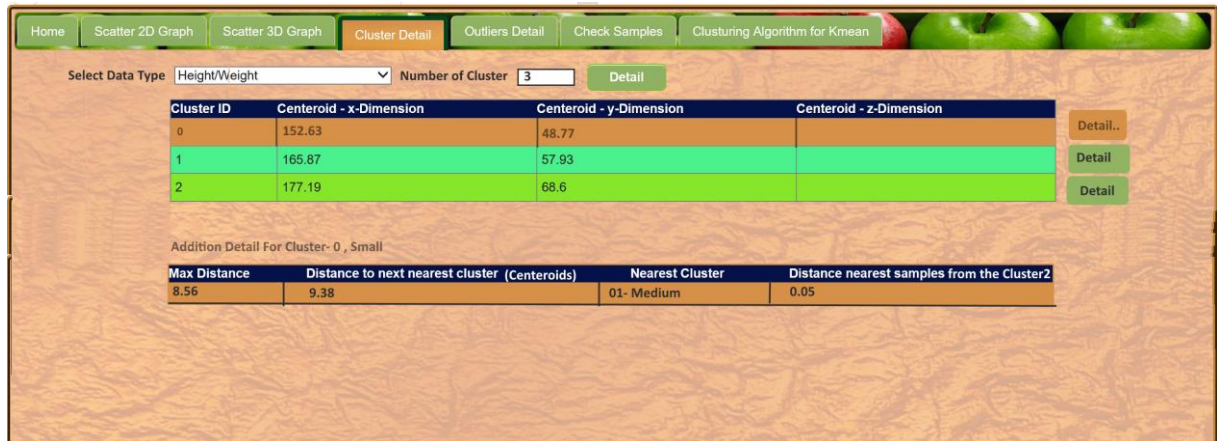
## Scatter 3D Graph



This is the three dimensional representation of data. The functioning of this page is very similar to the two dimensional case.



## Cluster Detail



On this Tab the interesting details like Cluster Details like Centroids, Cluster-Id/Name, Distance to farther sample from cluster centroid, Distance to next nearest cluster, Distance between nearest samples of nearest clusters shown in a Tabular Format.

## Outliers Detail



This Tab shows the details of Outliers Anomalous data which does not belong to any existing cluster. The outliers can be calculated by the API by setting up threshold distance for the clusters or marking any data as outlier which is furthest from the maximum distance from the centroids.

## Check Samples

X- Dimension	Y- Dimension	Z- Dimension	Cluster
147.58	52.58		0 - Small
180.56	72.34		2 - Large
160.01	57.02		1 - Medium
182.25	73		2 - Large
140.50	70		
185.25	50		

This for checking samples for the Data set type if they are belonging to pre existing or precalculated clusters or are outliers. In the above figure the red colour data does not belong to any cluster and are outliers.

## Clustering Algorithms for k-means

We can run Clustering Algorithm K-means on a totally new data set or on pre-existing data by changing number of cluster or uploading more data to preexisting data and running clustering over them.

### Upload new raw data for Clustering

Number of Clusters K:  Valid: false \*required

Type of Data for Training

Upload new data ☒

Pre Existing Data ☐

Upload data to existing Anomaly Data ☐

Browse... \*required

Start

### Clustering Pre Existing Data

Number of Clusters K:  Valid: true \*required

Type of Data for Training

Upload new data ☐

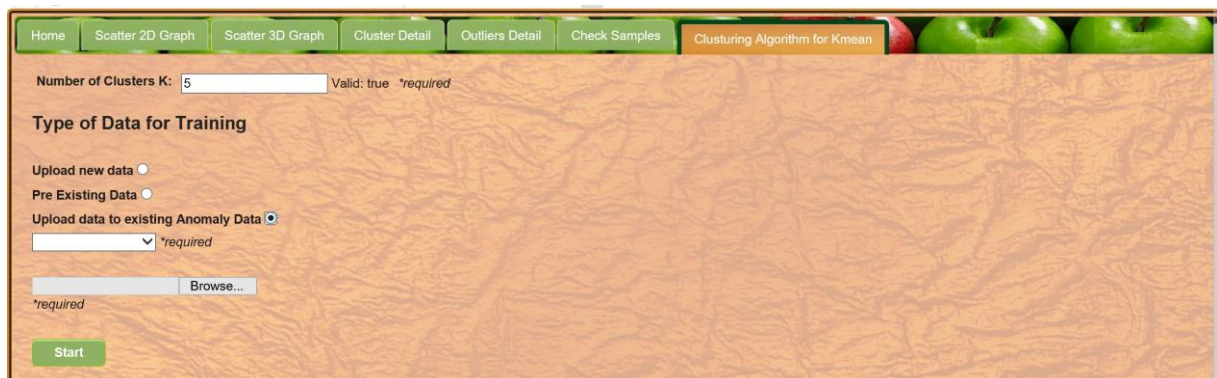
Pre Existing Data ☒

Upload data to existing Anomaly Data ☐

\*required

Start

## Upload new data to Existing Data



Home Scatter 2D Graph Scatter 3D Graph Cluster Detail Outliers Detail Check Samples Clustering Algorithm for Kmean

Number of Clusters K:  Valid: true \*required

**Type of Data for Training**

Upload new data ☐

Pre Existing Data ☐

Upload data to existing Anomaly Data ☒

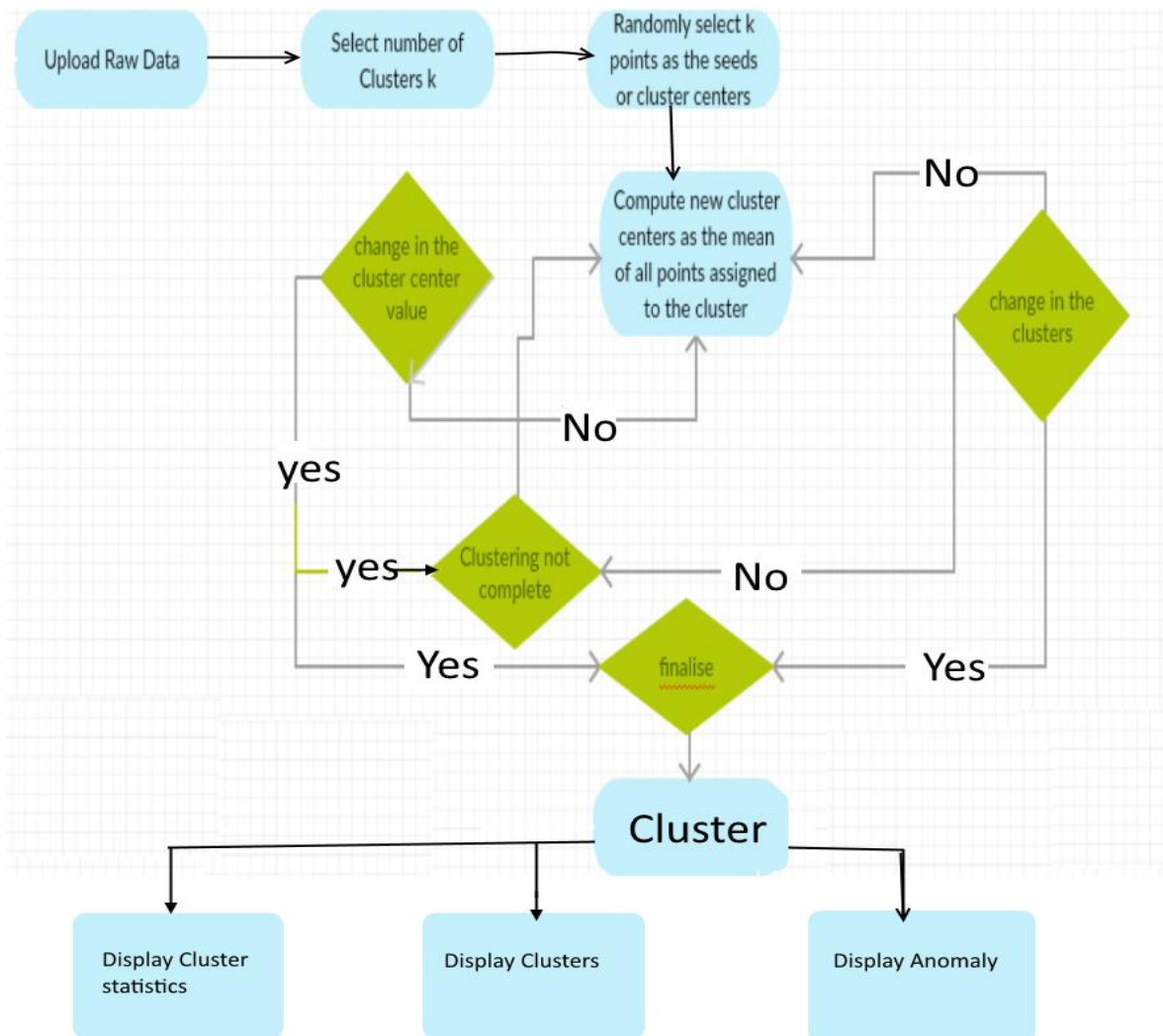
\*required

Browse...

\*required

Start

## Flowchart





## Scenario

In this project the Height/Weight data is represented and clustering is done for Market Segmentation of different cloth sizes – Small, Medium and Large.

The Age, Weight , Length data is represented and clustered for analysing the growth pattern of Babies.

This can also be used for identifying abnormal data items in a very large data set, for example, identifying potentially fraudulent credit-card transactions, risky loan applications and so on.

## Technical Details

For UI HTML5/angular.JS is used.

For Mock REST service ASP.NET WebApi is used

Azure SQL Database is used to store Data related to project

LinQ to SQL is used to connect Mock API to Database

## References Used

<http://www.ke.tu-darmstadt.de/lehre/archiv/ss10/web-mining/wm-cluster.pdf>

<http://amid.fish/anomaly-detection-with-k-means-clustering/>

<https://msdn.microsoft.com/en-us/magazine/jj891054.aspx>

<http://campus.codeschool.com/courses/shaping-up-with-angular->

[js/intro?utm\\_source=google&utm\\_medium=cpc&utm\\_campaign=course&qclid=Cj0KEQIAycCyBRDss-D2ylWd\\_tgBEiQAL-9Rkha3V9h-4FXs\\_VQooGU4\\_tC6p49LW776vNe0xmVpKSgaAqEp8P8HAQ](http://campus.codeschool.com/courses/shaping-up-with-angular-)

Class notes and guidance from Professor **Damir Dobric**.