**Experiment No.6**

**Title:** Applying data normalization and discretization

|  |
| --- |
|  |

**Batch: A4 Roll No.: 1914078 Experiment No.: 6**

**Aim:** Applying data normalization and discretization

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Resources needed:** Any programming language, any data source (RDBMS/Excel/CSV)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Theory:**

**Normalization:**

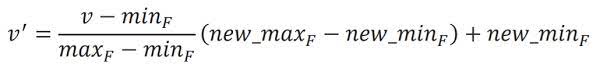
Measurement unit used can affect the data analysis. Hence data are scaled to fall within a smaller range like 0.0 to 1.0. Such transformation or mapping the data to a smaller or common range will help all attributes to gain equal weight. This is known as Normalization. Methods used for normalization:

1. min-max normalization
2. z-score normalization
3. decimal scaling.

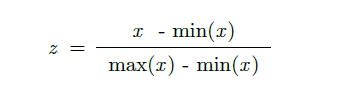
The normalization parameters such as mean, standard deviation, the maximum absolute value must be preserved in order to normalize the future data uniformly. If not normalized, one feature might completely dominate the others. If every data point have the same scale then feature is equally important. It will help to speed up the learning phase while dealing with attributes on a different scale. It will avoid dependence on the choice of measurement units. Comparison can be made easily. The application of data mining algorithms becomes easier, effective and efficient. Once the data is normalized, the extraction of data from databases becomes a lot faster. More specific data analyzing methods can be applied to normalized data. It prevent attributes with initially large ranges (e.g., income) from outweighing attributes with initially smaller ranges (e.g., binary attributes

1. Min Max Normalization

Min-max normalization performs a linear transformation on the original data in range  [0, 1] or [−1, 1]. Selecting the target range depends on the nature of  data. If***minF***and ***maxF***are the **minimum**and **maximum**values of an **attribute F**, Min-max normalization maps a value, ***vi*** of A to ***vi'*** in the range [***new\_minF***,***new\_maxF***] by computing:

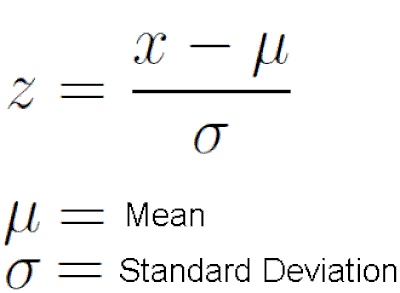


Min-max normalization preserves the relationships among the original data values. It encounter an “out-of-bounds” error if a future input case for normalization falls outside of the original data range for F. In generalized manner the formula can be written as:



Z-Score Normalization

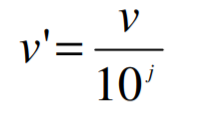
This method normalizes the value for attribute X using the **mean** and **standard deviation**. The formula for the same is:



Data can include multiple dimensions. Feature standardization makes the values of each feature in the data have zero-mean (when subtracting the mean in the numerator) and unit-variance but not normal distribution it can be still skewed.  This method is widely used for normalization in many machine learning algorithms (e.g., support vector machines, logistic regression, and artificial neural networks.

Decimal Scaling

This method normalizes the value of attribute A by moving the decimal point in the value. This movement of a decimal point depends on the maximum absolute value of A.The formula for the decimal scaling is:



J is the The smallest integer j such that Max(| vi/10j|) < 1. Number of digits in data value with largest absolute value.

**Data discretization**

Discretization is defined as a process of converting continuous data attribute values into a finite set of intervals with minimal loss of information and associating with each interval some specific data value. (e.g., 0–10, 11–20, etc.) or conceptual labels (e.g., youth, adult, senior). The goal of discretization is to reduce the number of values a continuous variable assumes by grouping them into a number, b, of intervals or bins. Interval labels can then be used to replace actual data values. Discretization reduce data size. Discretization is considered a data reduction mechanism because it diminishes data from a large domain of numeric values to a subset of categorical values. Following are data discretization techniques:

1. Binning:Binning is a top-down splitting technique based on a specified number of bins. Binning is an unsupervised discretization technique. Main challenge in discretization is to choose the number of intervals or bins and how to decide on their width.
2. Histogram Analysis: It is a Top-down split. Since histogram analysis does not use class information so it is an unsupervised discretization technique. Histograms partition the values for an attribute into disjoint ranges called buckets.
3. Cluster Analysis: Cluster analysis is a unsupervised method which can be top-down split or bottom-up merge. A clustering algorithm can be applied to discrete a numerical attribute of A by partitioning the values of A into clusters or groups. Each initial cluster or partition may be further decomposed into several subcultures, forming a lower level of the hierarchy. Detect and remove outliers
4. Decision-tree analysis (supervised, top-down split)
5. Correlation (e.g. chi merge) analysis (unsupervised, bottom-up merge)

**K Means Clustering Algorithm**

A cluster refers to a collection of data points aggregated together because of certain similarities. K-mean algorithm is one of the centroid based technique.  K means algorithm takes the unlabeled dataset as input, divides the dataset into k-number of clusters, and repeats the process until it does not find the best clusters without the need for any training. K  refers to the number of pre-defined clusters that need to be created in the process, as if K=2, there will be two clusters, and for K=3, there will be three clusters.  Each cluster is associated with a centroid and algorithm aims to minimize the sum of distances between the data point and their corresponding cluster centroids.

Diagram

Description automatically generated

Steps:

1. Select the number K to decide the number of clusters. K can be determined using some techniques like Elbow Method using WCSS(Within Cluster Sum of Squares), Silhouette Method.
2. Compute random seed points as the centroids or K points of the clusters of the current partitioning ( A centroid is the imaginary or real location representing the center of the cluster or mean point of the cluster .)
3. Assign each object to the cluster with the nearest seed point based on eulidean distance. Using a different distance function other than (squared) Euclidean distance may prevent the algorithm from converging. Partition objects into *k* nonempty subsets by seeing the closest centroids for each data points.
4. Calculate the variance and place a new centroid of each cluster based on mean value of the cluster. Better choice is to place them as much as possible far away from each other
5. Go back to Step 3, stop when the assignment does not change

There are essentially three stopping criteria that can be adopted to stop the K-means algorithm:

1. Centroids of newly formed clusters do not change
2. Points remain in the same cluster
3. Maximum number of iterations are reached

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

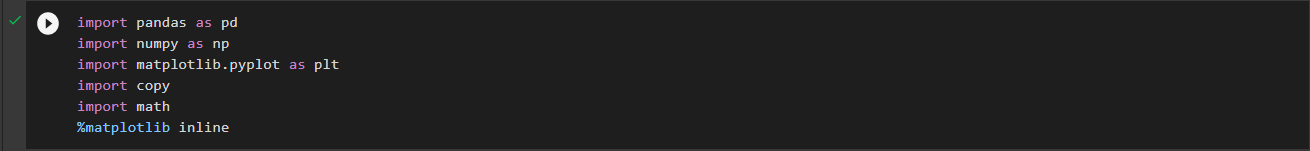
**Procedure / Approach /Algorithm / Activity Diagram:**

1. Identify the suitable attributes to apply the following normalization techniques and display the consolidated output of all 3 normalizations.
   1. Min-Max Normalization
   2. Z-score Normalization
   3. Decimal Scaling
2. Identify the suitable attributes to apply the K-means discretisation techniques and display the output of discretized data

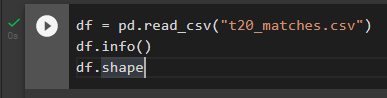
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

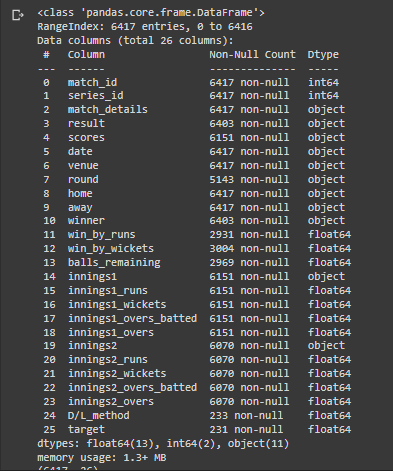
**Results: (Program printout with output / Document printout as per the format)**

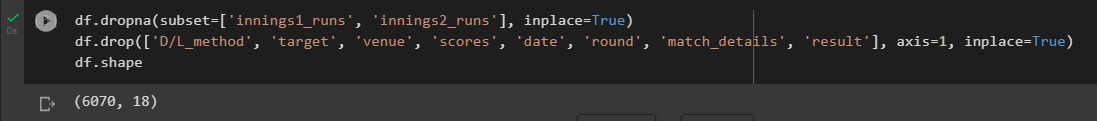
**Importing required libraries:**

****

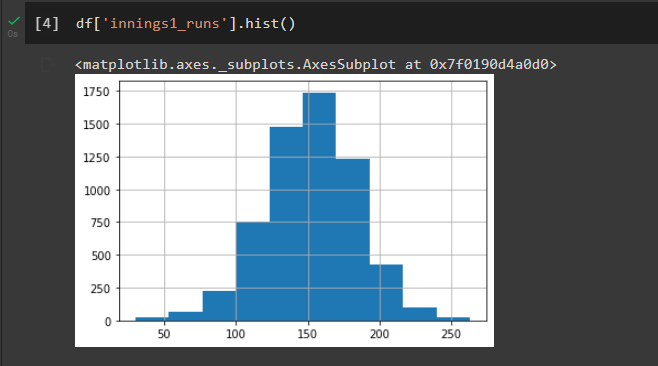
**Reading the dataset:**

****

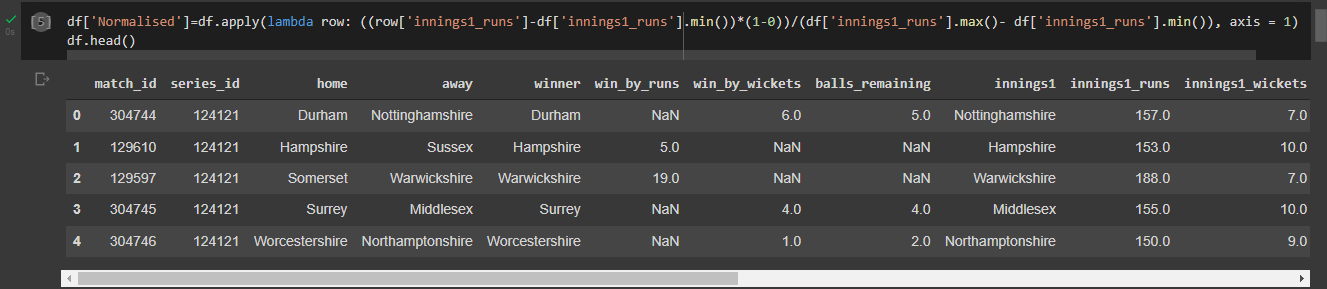
****

****

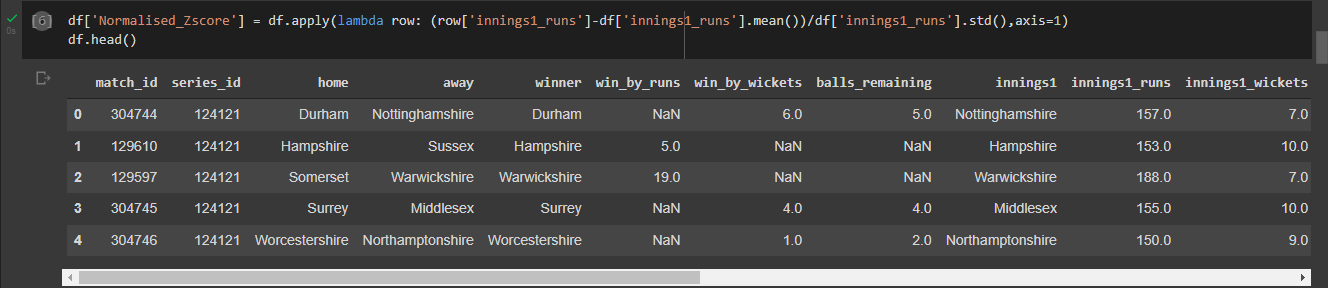
**Plotting histogram for the attribute considered for normalization:**

****

**Min max Normalization:**

****

**Z-Score Normalization:**

****

**Decimal Normalization:**

****

**K-means Clustering (K taken as 3):**

x=df['innings1\_runs'].astype(int).values

df["Cluster"]=np.nan

ct=0

elbow=[]

kvalue=[]

for i in range (3,4):

    pts=[]

    d=[]

    index=[]

    for j in range(i):

        index.append(np.random.choice(x.shape[0], replace=False))

    for j in index:

        pts.append(x[j])

    while True:

        ct+=1

        old=copy.copy(df["Cluster"])

        for k in range (x.shape[0]):

            d=[]

            for pt in pts:

                d.append(abs(x[k]-pt))

            minpos = d.index(min(d))

            df['Cluster'].iloc[k]=minpos

        avg=[]

        for k in range (i):

            cluster=df.query("Cluster==@k")

            if math.isnan(np.nanmean(cluster['innings1\_runs'])):

                avg.append(0)

            else:

                avg.append(np.nanmean(cluster['innings1\_runs']))

        print('Average of the clusters=',avg)

        pts=copy.copy(avg)

        print('New centroids=',pts)

        # if previous cluster allocation was same as new, we break the while loop

        if old.equals(df['Cluster']):

            break

    print('Cluster value counts: ')

    print (df['Cluster'].value\_counts()) # we get the count of points in the 5 clusters

    wcss=0

    for k in range (i):

        cluster=df.query("Cluster==@k")

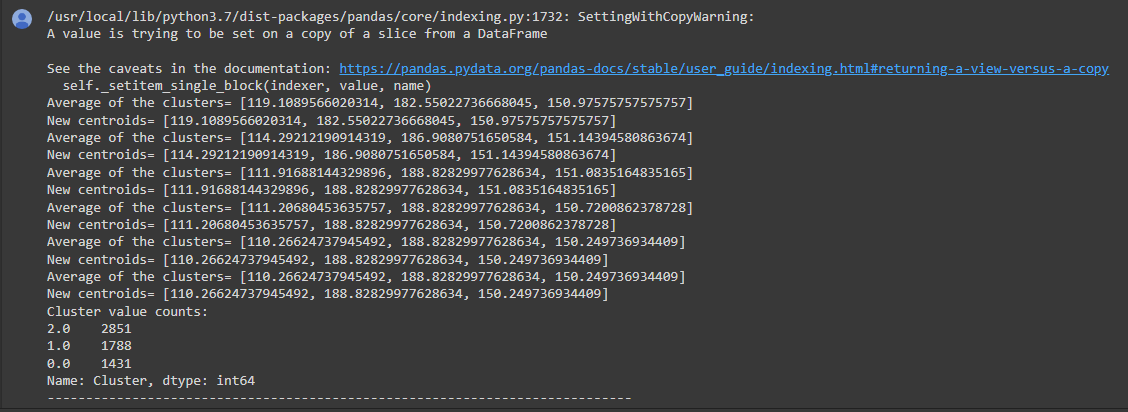
        for m in range (cluster.shape[0]):

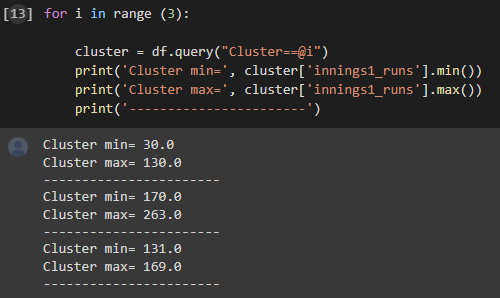
            wcss+=abs(cluster['innings1\_runs'].mean()-cluster['innings1\_runs'].iloc[m])\*\*2

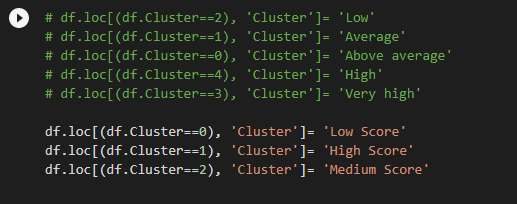
    elbow.append(wcss)

    kvalue.append(i)

    print('----------------------------------------------------------------------------')

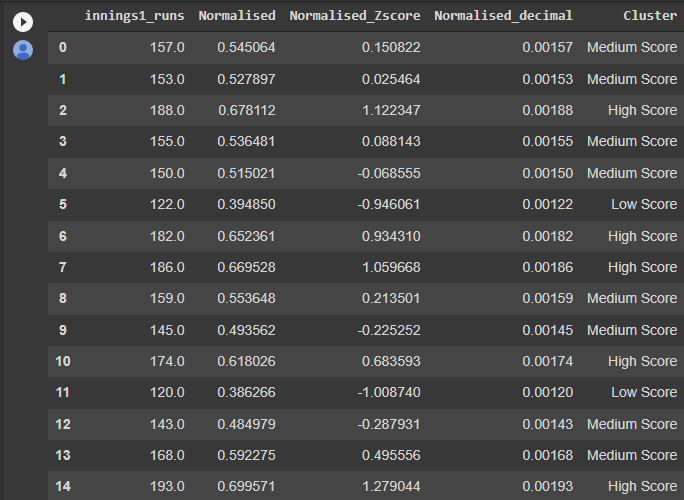
****

****

****

**Final Output:**

****

****

****

**Questions:**

1. What happens if data is not normalised?

**Ans.**

The accuracy of machine learning models that use such data can be dramatically reduced if the data requires normalisation and is not properly standardised. Some machine learning techniques demand that data be regularly distributed. When data is used directly, prediction accuracy suffers, and the models created are inadequate.

1. Is there any criteria for selecting attributes for K-means algorithm. If so discuss.

Selection of the number of clusters

**Ans.**

• Values of K specified within a range or set

• Values of K determined by statistical measures

• Values of K equated to the number of classes

• Values of K determined through visualization

• Values of K determined using a neighbourhood measure

Factors affecting the selection of K

• Approach bias

• Level of detail

• Internal distribution versus global impact

**Outcomes:**

**CO3**- Apply the transformations required on data to make it suitable for mining \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Conclusion: (Conclusion to be based on the objectives and outcomes achieved)**

Thus performed Min-Max Normalization, Z-score Normalization, Decimal Scaling and K means clustering on an appropriate attribute of a dataset using python language.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Grade: AA / AB / BB / BC / CC / CD /DD**

Signature of faculty in-charge with date

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**References:**

Books/ Journals/ Websites:

1. Han, Kamber, "Data Mining Concepts and Techniques", Morgan Kaufmann 3nd Edition