Department of Computer Science & Engineering

Final Year B. Tech. (CSE) – I: 2022-23 5CS462: PE5 - Data Mining Lab

Assignment No. 2

PRN: 2020BTECS00205

NAME: Monika .V. Chitrakathi

BATCH: B8

**Title:** Design the data analysis tools (GUI) to perform the following pre- processing task.

1. Correlation analysis - Chi-Square Test
   1. User should be able to choose any two attributes.
   2. Display the contingency table
   3. Show the chi-square value and conclusion whether the selected attributes are correlated or not.
2. Correlation analysis – Correlation coefficient (Pearson coefficient) & Covariance a. User should be able to choose any two attributes.

b. Show the calculated values and conclusion whether the selected attributes are correlated or not.

1. Normalization using following techniques:
   1. Min-max normalization
   2. Z-Score normalization
   3. Normalization by decimal scaling
2. User should be able to choose any attributes.
3. Show the calculated values in tabulated form.
4. Show the scatter plot for normalized attributes

# Objective/Aim:

* 1. To implement data analysis tool using python programming language.
  2. To implement correlation analysis for selected attributes using Chi-Square Test and Correlation Coefficient (Pearson Coefficient) & Covariance.
  3. To implement normalization techniques for selected attributes using min- max normalization, Z-Score normalization & normalization by decimal scaling, displaying calculated values and plotting scatter plot for normalized attributes.

# Introduction:

Correlation Analysis in Data mining:

Correlation analysis is a statistical method used to measure the strength of the linear relationship between two variables and compute their association. Correlation analysis calculates the level of change in one variable due to the change in the other. A high correlation points to a strong relationship between the two variables, while a low correlation means that the variables are weakly related.

Normalization:

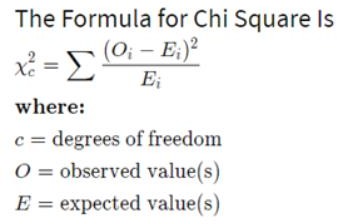
Normalization is generally required when we are dealing with attributes on a different scale, otherwise, it may lead to a dilution in effectiveness of an important equally important attribute (on lower scale) because of other attribute having values on larger scale. In simple words, when multiple attributes are there but attributes have values on different scales, this may lead to poor data models while performing data mining operations. So, they are normalized to bring all the attributes on the same scale.

# Theory:

Correlation Analysis:

* + 1. Chi-Square Test:

A chi-square test is used to help determine if observed results are in line with expected results, and to rule out that observations are due to chance. A chi-square test is appropriate for this when the data being analysed are from a [random sample](https://www.investopedia.com/terms/s/simple-random-sample.asp), and when the variable in question is a categorical variable.



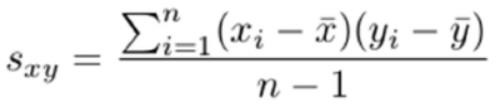
The degrees of freedom for the chi-square are calculated using the following formula: df = (r-1)(c-1) where r is the number of rows and c is the number of columns.

If chi-square value is greater than the degree of freedom then two attributes are strongly correlated and if less than zero then two attributes are not correlated to each other.

* + 1. Correlation Coefficient (Pearson Coefficient) and Covariance:

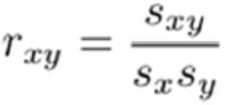
Covariance and Correlation are very helpful in understanding the relationship between two continuous variables. Covariance tells whether both variables vary in the same direction (positive covariance) or in the opposite direction (negative covariance). There is no meaning of covariance numerical value only sign is useful. Whereas Correlation explains the change in one variable leads how much proportion change in the second variable.

Covariance between two variables x and y can be calculated as follows:



Where, x̄ is the sample mean of x, ȳ is sample mean of y, xi and yi are the values of x and y for ith record in the sample and n is the no of records in the sample.

Correlation between x and y can be calculated as follows:



Where sxy is the covariance between x and y, sx and sy is standard deviation of x and y respectively and rxy is correlation or Pearson coefficient.

* + 1. Normalization methods:
       1. Min-max normalization:

For all values vi of attribute, A (i = 1 … n), replace vi as, vi = (vi-min(A))/(max(A)-min(A)).

* + - 1. Z-Score normalization:

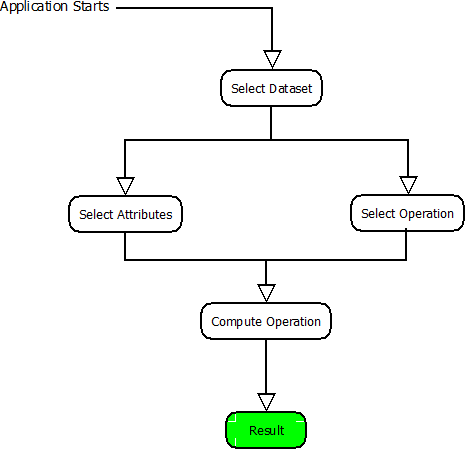
For all values vi of attribute, A (i = 1 … n), replace vi as, vi = (vi-mean(A))/Standard Deviation(A)

* + - 1. Normalization by decimal scaling:

For all values vi of attribute, A (i = 1 … n), replace vi as, vi = vi /10j

where j is the smallest integer value such that max(A)/10j < 1.

# Functional Block Diagram:



**Procedure:**

Given problem statement is solved using python programming language and specifically used tkinter module to implement GUI application and pandas module to load .csv file as dataset.

# Screenshots:

