**Informatics Institute of Technology**

**Department of Computing (B.Eng.) in Software Engineering**

**Module: 5DATA001C.2 Machine Learning and Data Mining**

**Module Leader: Dr. V.S. Kontogiannis**

* Date of Submission : April 25, 2022
* Student ID : 2019822
* Student UoW ID : w1761405
* Name : D. W. K. S. Gunathilake
* Tutorial Group : Part Time – Group A

Contents

[Objective 01: 3](#_Toc102953387)

[Objective 02 5](#_Toc102953388)

### Objective 01:

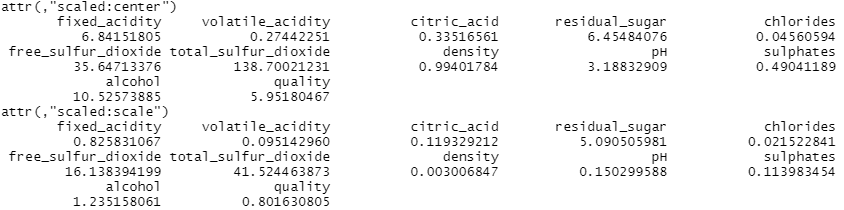
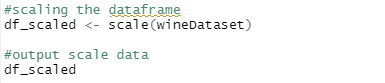
Pre-Processing

1. Scaling and outliers’ removal and briefly justify your answer. (Order of scaling and outliers’ removal)

**Ans**:

Scale is a technique of comparing data that is not measured in the same way. Normalizing a dataset using the mean value and the standard deviation is called scale.

Figure 1: scaled dataset



The normalizing of a dataset using the mean value and standard deviation is known as scaling.

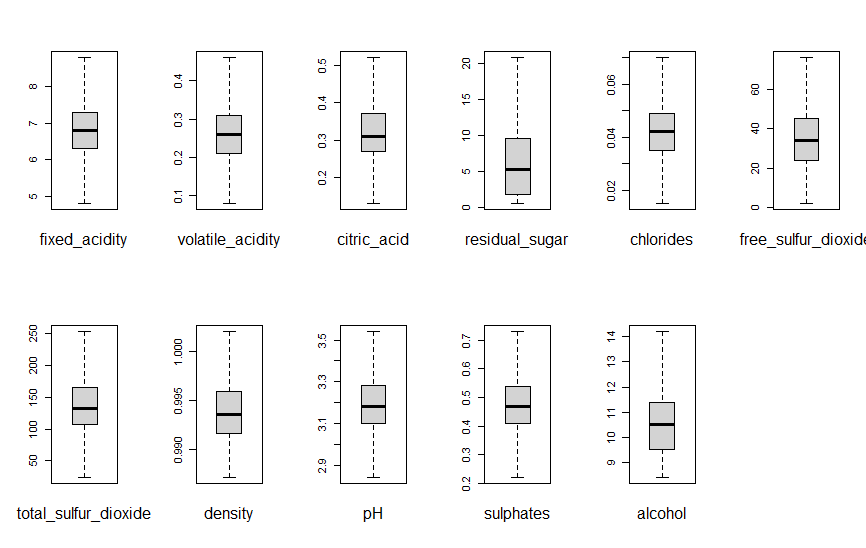
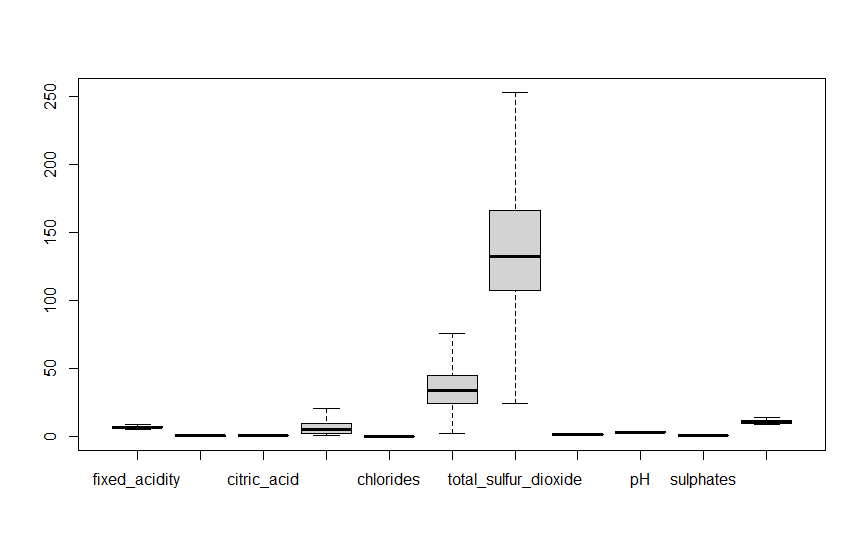
This part is a process of cleaning the data sample given, in order to get a best visualize to the problem.

Graphical user interface, application

Description automatically generatedOutliers – Values in the data set with outside of the range which acts very different from other data in the dataset. In cleaning process, we should remove these outliers for better understanding of the problem.

Figure 2: new dataset without any outliers

Figure 3: Visualized dataset without outliers



1. Define the number of cluster centres.

**Ans**:

First three are the automated ways to get the best fitting number of cluster centers.

**Silhouette method**

silhouette method we can calculate silhouette coefficient and easily find exact number of k.

Logo

Description automatically generated with medium confidence

Chart, line chart

Description automatically generated

**Gap statistic**

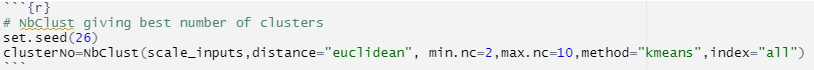
Text

Description automatically generatedThe best is to apply it and then run the Gap-Statistic plot for a couple of times. Taking the average of the Gap-Statistics can be an increased evaluation criterion.

Chart, line chart

Description automatically generated

**NbClust giving best number of clusters**

****The NbClust package determining cluster numbers and suggests using the best cluster resolution scheme with the various results obtained by modifying all combinations of cluster number, distance measurements, and cluster algorithms.

**Text, letter

Description automatically generated**

**Chart, line chart

Description automatically generated**

**Chart, line chart

Description automatically generated**

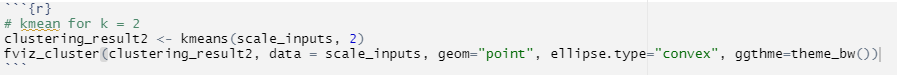
**Elbow method**

**** The elbow method is a heuristic used in determining the number of clusters in a data set.

Chart, line chart

Description automatically generated

1. Using all input variables, perform a kmeans analysis;
   1. Chart, scatter chart

      Description automatically generatedK =2

**BSS:** 9469.04

**TSS:** 41052

**BSS/TSS ratio:** 0.2306596

**Within Cluster sum of squares:** [12311.01 19271.95]

* 1. Chart, scatter chart

     Description automatically generatedK = 3

**BSS:** 12170.07

**TSS:** 41052

**BSS/TSS ratio:** 0.2964551

**Within Cluster sum of squares:** [9722.067 9721.606 9438.253]

* 1. Chart, scatter chart

     Description automatically generatedK = 4

**BSS:** 14214

**TSS:** 41052

**BSS/TSS ratio:** 0.3462437

**Within Cluster sum of squares:** [5983.483 8872.553 6847.724 5134.245]

**Confusion Matrix/ Accuracy/ Recall/ Precision for each K value**

1. k = 2

Table

Description automatically generated with medium confidenceConfusion Matrix :

= **0.6504152**

Graphical user interface, application

Description automatically generated

= **0.595122**

Graphical user interface, application, Word

Description automatically generated

= **0.4066667**

Graphical user interface, text, application, Word

Description automatically generated

1. k = 3

Table

Description automatically generated with low confidenceConfusion matrix:

Graphical user interface, application, Word

Description automatically generated

Accuracy = **0.5309403**

Recall = **0.3621742**

Precision = **0.5207337**

1. **A picture containing text

   Description automatically generated**k = 4

**Graphical user interface, application, Word

Description automatically generated**

Accuracy = **0.6879186**

Recall = **0.5073171**

Precision = **0.440678**

**According to the recall values, k = 2 clustering option gives us the most probability to retrieve the relevant items. So, winner should be k = 2 even though it has less values for the accuracy comparing to others.**

1. Apply a PCA for this white wine dataset. 5. And Create a new transformed dataset with PC as attributes.

A picture containing text

Description automatically generated

Chart, histogram

Description automatically generated

1. Calendar

   Description automatically generated Choose those PCs that provide a cumulative score > 96%.

Graphical user interface, text, application, email

Description automatically generatedReason to remove the 9,10,11 PCs is their cumulative value is greater than 96%. Below graph will shows it clearly.

Chart, line chart, scatter chart

Description automatically generated

1. Chart

   Description automatically generated with medium confidenceApply kmeans analysis to new transformed dataset with the same K as winner for previous section and show r related outputs.

Table

Description automatically generated

A picture containing chart

Description automatically generated

1. Discuss performance of PCA model
   1. Calculating related BSS

**Ans: 9467.792**

* 1. BSS/TSS ratio

**Ans:** between\_SS / total\_SS = **0.245**

1. WSS indices and compare the produced indices against the related ones from the winner model

|  |  |  |
| --- | --- | --- |
|  | Before applying PCA | After applying PCA |
| BSS | 9469 | 9467.792 |
| TSS | 9469 | 38582 |
| BSS/TSS Ration | 0.231 | 0.245 |
| WSS | [12311, 19272] | [11285.74 17828.32] |

### Objective 02

1. Brief discussion of the various methods used for defining the input vector in electricity load forecasting problems

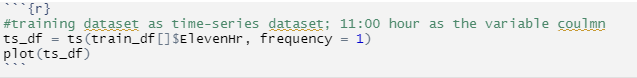
* The time-series models—autoregressive (AR), moving average (MA), and autoregressive moving average (ARMA)
* Multiple linear regression
* Advanced approaches (i.e., neural networks, fuzzy logic, SVM, etc., and its hybrid mode)

1. Input/Output Matrices for (t-x)

Text

Description automatically generatedI have renamed the columns for better usage and easier handling.

Chart

Description automatically generated Extract the data into a Time Series dataset

As shown below each train and test data set processed to create input/output matrices for :

(t-1/t-2/t-3/t-4/t-7)

Graphical user interface, text, application

Description automatically generatedGraphical user interface, text, application

Description automatically generated

So, created I/O matrices are show like below.

**Train**

Table

Description automatically generatedt – 1

t – 2

Table

Description automatically generated

Table

Description automatically generatedt – 3

t – 4

Table

Description automatically generated

Table

Description automatically generatedt – 7

**Test**

Table

Description automatically generatedt – 1

t – 2

Table

Description automatically generated

Table

Description automatically generatedt – 3

Table

Description automatically generatedt – 4

t – 7

Table

Description automatically generated

1. Normalization the matrixes

A picture containing graphical user interface

Description automatically generated

Table

Description automatically generatedAfter normalization matrices will be shown as below. Showing only one matrix after normalized.

1. Implement a number of MLPs for the AR approach, using various structures (layers/nodes)/input parameters/network parameters and show in a table, their performances comparison (based on testing data) through the provided stat. indices.

**AR analysis**

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Text

Description automatically generatedText

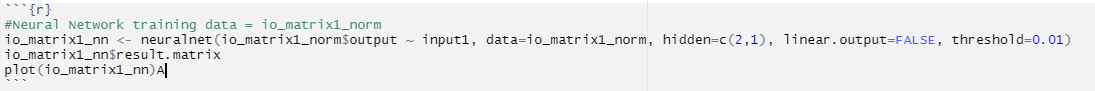
Description automatically generated

Text

Description automatically generated with medium confidence

**Creating neutral networks using these matrixes**

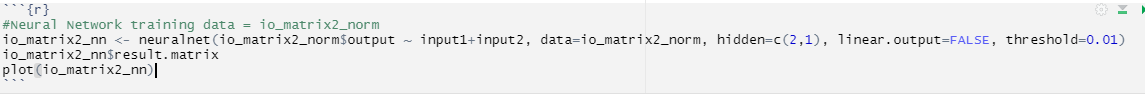
**T – 1**

****

Chart

Description automatically generated with medium confidence

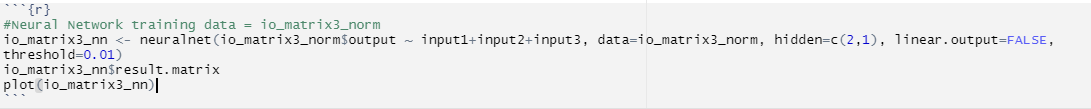
**T – 2**



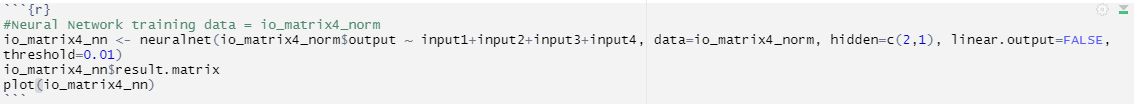
A picture containing chart

Description automatically generated

**Chart

Description automatically generated with low confidence** **T – 3**

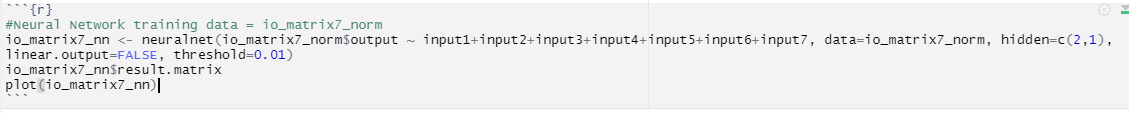
**T – 4**

****

**A picture containing chart

Description automatically generated**

**A picture containing diagram

Description automatically generated T – 7**

**Prediction model using test dataset**

**Graphical user interface, text, application, email

Description automatically generated**

**Graphical user interface, text, application, email

Description automatically generated**

**MAE value**

**Graphical user interface, application

Description automatically generated**

**Graphical user interface, application

Description automatically generatedGraphical user interface, application

Description automatically generatedGraphical user interface, application

Description automatically generated**

**Graphical user interface, application

Description automatically generated**

**RMSE Value**

**Graphical user interface, application, Teams

Description automatically generated**

**Graphical user interface, text, application, email, Teams

Description automatically generated**

**Comparison for performances and testing performances**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Matrix** | **Error** | **MAE** | **RMSE** | **likelihood** |
| T – 1 | 7.958855e+00 | 0.3718972 | 0.4518572 | -1951.25 |
| T – 2 | 7.674473e+00 | 0.3672024 | 0.4478461 | -1951.25 |
| T – 3 | 7.566236e+00 | 0.3418974 | 0.4197478 | -1951.25 |
| T – 4 | 7.184324e+00 | 0.3438052 | 0.4199514 | -1951.25 |
| T – 7 | 12.433971723 | 0.5880558 | 0.6497547 | -1951.25 |

1. Discussion of the meaning of these stat. indices
2. Discuss the issue of “efficiency” with your two best NN structures 4
3. Provide your best results both graphically (your prediction output vs. desired output) and via performance indices (3 marks for the graphical display and 3 marks for showing the requested statistical indices) 6