STA5077Z: UNSUPERVISED LEARNING - PROJECT

Corne Oosthuizen - OSTAND005

Due: 14 August 2017

Table of Contents

[1 - Leukemia Dataset 2](#_Toc490485330)

[Dataset 2](#_Toc490485331)

[Principal Component Analysis (PCA) 2](#_Toc490485332)

[All Genes – Unchanged 3](#_Toc490485333)

[All Genes – Scaled 5](#_Toc490485334)

[All Genes – Log Transformed 7](#_Toc490485335)

[All Genes – Log Transformed and Scaled 9](#_Toc490485336)

[Selectin the Top 100 Gene Expressions 11](#_Toc490485337)

[Top 100 – Unchanged 12](#_Toc490485338)

[Top 100 – Log Transformed and Scaled 14](#_Toc490485339)

[Conclusion 16](#_Toc490485340)

[2 – Lecture Recording 17](#_Toc490485341)

[Dataset 17](#_Toc490485342)

[Conclusion 21](#_Toc490485343)

[3 - Notes 21](#_Toc490485344)

[4 - References 22](#_Toc490485345)

# 1 - Leukemia Dataset

## Dataset

The dataset for this part of the assignment comprises of the leukemia dataset provided as a resource on Vula.

The dataset describes the genes expression (22283) for 16 patients, the first 8 of which has a good prognosis and the last 8 having poor prognosis. The prognosis is measured by the number of leukemic cells present in the patient’s bone marrow after a period of treatment.

## Principal Component Analysis (PCA)

The dataset is then examined by using Principal Component Analysis (PCA), the variations are as follows:

* All Genes:
  + Unchanged
  + Scaled
  + Log transformed
  + Log transformed and scaled
* Top 100 Genes: (selected from the previous PCA information – 1st Principal Component)
  + Unchanged
  + Log Transformed and scaled

Performing these operations on the dataset using PCA, we can see the impact that scaling and centering the data can have to filter off some trivial effects, which could skew our PCA.

### All Genes – Unchanged

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |

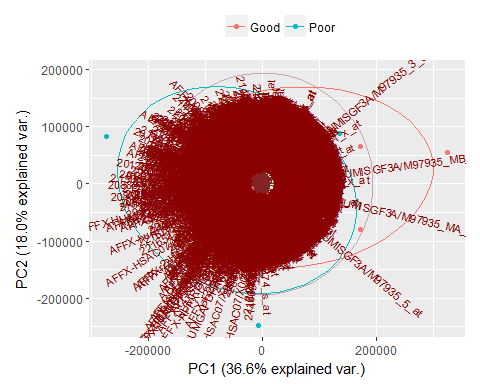
The mapping of the 1st and 2nd principal component does seem to imply that using them we could likely separate the **good** and **poor** patient prognosis by looking at positive PC2 values.

Importance of components:  
 PC1 PC2 PC3 PC4  
 Standard deviation 156238.6335 109660.68045 89020.73562 80953.43395  
 Proportion of Variance 0.3663 0.18045 0.11892 0.09834  
 Cumulative Proportion 0.3663 0.54676 0.66568 0.76402

PC5 PC6 PC7 PC8  
 Standard deviation 56506.15293 46154.70483 44327.27105 40692.64712  
 Proportion of Variance 0.04791 0.03197 0.02949 0.02485  
 Cumulative Proportion 0.81193 0.84390 0.87338 0.89823

PC9 PC10 PC11 PC12  
 Standard deviation 36028.25444 35068.76793 33728.38573 31976.10588  
 Proportion of Variance 0.01948 0.01845 0.01707 0.01534  
 Cumulative Proportion 0.91771 0.93616 0.95323 0.96858

PC13 PC14 PC15 PC16  
 Standard deviation 27846.40656 26133.25873 25210.79476 3.745941e-10  
 Proportion of Variance 0.01164 0.01025 0.00954 0.000000e+00  
 Cumulative Proportion 0.98021 0.99046 1.00000 1.000000e+00



It is quite difficult to determine the overall grouping and direction of the patients in relation to their gene expressions from the complete dataset. This might improve in the top 100 selection graphs.

### All Genes – Scaled

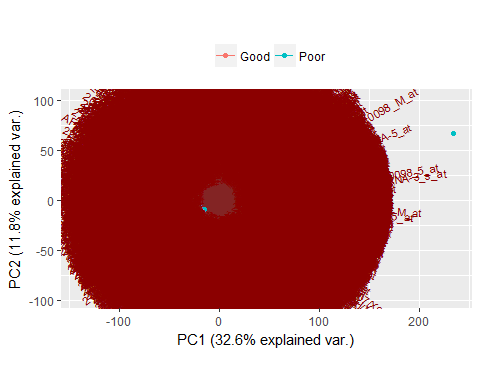
|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |

The scaled PCA provides us with a more comprehensive distribution of patient results and looking at 1st and 2nd PC graph indicates that we might use the top right quadrant to classify patients. There are some overlaps with (*p3\_poor*, *p5\_poor* and *p6\_poor*).

Importance of components:  
 PC1 PC2 PC3 PC4 PC5  
 Standard deviation 85.22634 51.34150 44.34176 37.21473 36.33934  
 Proportion of Variance 0.32597 0.11829 0.08824 0.06215 0.05926  
 Cumulative Proportion 0.32597 0.44426 0.53250 0.59465 0.65391  
  
 PC6 PC7 PC8 PC9 PC10  
 Standard deviation 34.33981 30.23176 29.39799 28.51449 27.80705  
 Proportion of Variance 0.05292 0.04102 0.03878 0.03649 0.03470  
 Cumulative Proportion 0.70683 0.74785 0.78663 0.82312 0.85782

PC11 PC12 PC13 PC14 PC15  
 Standard deviation 27.09373 26.27332 25.18051 24.22583 22.86498  
 Proportion of Variance 0.03294 0.03098 0.02845 0.02634 0.02346  
 Cumulative Proportion 0.89077 0.92175 0.95020 0.97654 1.00000

PC16  
 Standard deviation 1.691711e-13  
 Proportion of Variance 0.000000e+00  
 Cumulative Proportion 1.000000e+00



The mapping is still to cluttered to interpret properly.

### All Genes – Log Transformed

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |

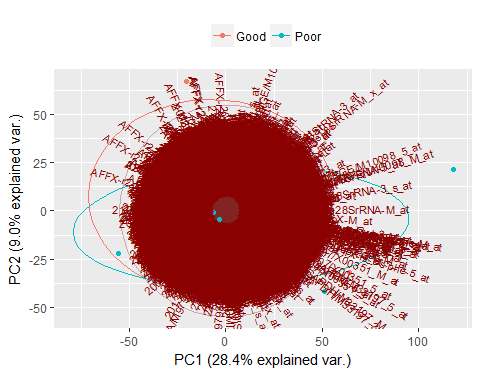
The log transform gives us a very clear distinction between the **good** and **poor** prognosis and could be used to predict the patient’s prognosis.

Importance of components:  
  
 PC1 PC2 PC3 PC4 PC5  
 Standard deviation 49.36871 27.82194 25.52599 22.75810 22.53319  
 Proportion of Variance 0.28385 0.09015 0.07588 0.06032 0.05913  
 Cumulative Proportion 0.28385 0.37399 0.44988 0.51019 0.56933

PC6 PC7 PC8 PC9 PC10  
 Standard deviation 20.90393 20.46430 20.17392 19.64844 19.49222  
 Proportion of Variance 0.05089 0.04877 0.04740 0.04496 0.04425  
 Cumulative Proportion 0.62022 0.66899 0.71639 0.76135 0.80560

PC11 PC12 PC13 PC14 PC15  
 Standard deviation 19.21435 18.93668 18.69074 17.77911 16.61446  
 Proportion of Variance 0.04300 0.04176 0.04068 0.03681 0.03215  
 Cumulative Proportion 0.84859 0.89035 0.93104 0.96785 1.00000

PC16  
 Standard deviation 9.516276e-14  
 Proportion of Variance 0.000000e+00  
 Cumulative Proportion 1.000000e+00



### All Genes – Log Transformed and Scaled

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |

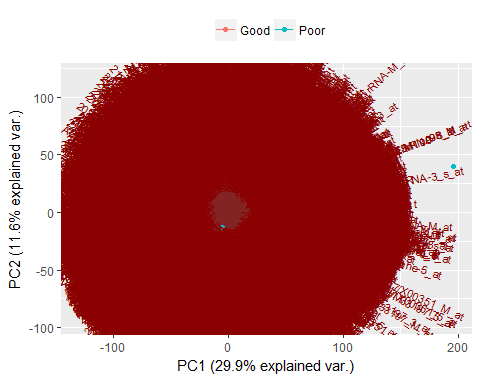
Similar to the log transform in the previous graph group, the log transform and scale shows a distinction between patients’ prognosis.

Importance of components:  
 PC1 PC2 PC3 PC4 PC5  
 Standard deviation 81.63214 50.79477 41.81428 37.86222 36.23344  
 Proportion of Variance 0.29905 0.11579 0.07846 0.06433 0.05892  
 Cumulative Proportion 0.29905 0.41484 0.49331 0.55764 0.61656

PC6 PC7 PC8 PC9 PC10  
 Standard deviation 33.74602 31.46372 30.65406 30.11230 29.10130  
 Proportion of Variance 0.05111 0.04443 0.04217 0.04069 0.03801  
 Cumulative Proportion 0.66766 0.71209 0.75426 0.79495 0.83296

PC11 PC12 PC13 PC14 PC15  
 Standard deviation 28.68284 27.93695 27.29255 26.39643 26.02576  
 Proportion of Variance 0.03692 0.03503 0.03343 0.03127 0.03040  
 Cumulative Proportion 0.86988 0.90491 0.93833 0.96960 1.00000

PC16  
 Standard deviation 1.920718e-13  
 Proportion of Variance 0.000000e+00  
 Cumulative Proportion 1.000000e+00



### 

### Selectin the Top 100 Gene Expressions

Looking at the PCA of the *log\_transform* and scale we can use the 1st Principal Component to select the Top 100 genes that contribute to the variance described for that component.

Table showing the Top 6 of the Top 100

p1\_good p2\_good p3\_good p4\_good p5\_good p6\_good p7\_good  
 204365\_s\_at 381.6 404.7 322.8 326.9 317.1 384.7 330.2  
 206766\_at 421.2 472.4 319.4 322.3 370.8 404.3 407.4  
 221870\_at 21.1 22.9 22.7 18.1 22.8 32.4 24.6  
 204542\_at 64.3 55.1 55.7 49.1 51.4 70.9 51.7  
 215356\_at 19.7 23.6 20.2 15.9 19.7 19.1 16.2  
 211843\_x\_at 27.3 28.6 25.5 23.0 23.4 23.0 24.7  
 p8\_good p1\_poor p2\_poor p3\_poor p4\_poor p5\_poor p6\_poor  
 204365\_s\_at 339.5 794.6 1344.8 243.2 614.6 365.1 228.7  
 206766\_at 397.8 673.9 1144.6 349.8 728.3 436.9 273.7  
 221870\_at 19.8 46.2 106.5 21.5 46.5 25.1 17.8  
 204542\_at 62.4 107.6 174.3 47.7 97.3 67.2 41.6  
 215356\_at 18.7 40.3 102.7 16.6 49.4 18.1 14.7  
 211843\_x\_at 27.5 49.6 137.0 24.5 41.6 27.6 19.5  
 p7\_poor p8\_poor  
 204365\_s\_at 820.6 741.1  
 206766\_at 863.7 714.2  
 221870\_at 53.9 55.8  
 204542\_at 115.2 110.0  
 215356\_at 63.1 57.1  
 211843\_x\_at 57.2 65.4

### Top 100 – Unchanged

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |

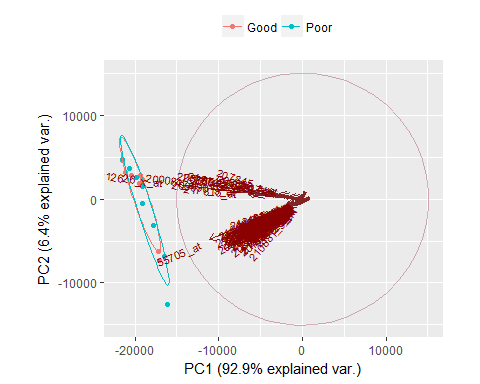
The unchanged top 100 selected genes do not prove to be an improvement on the full set of genes that was run initially.

Importance of components:  
 PC1 PC2 PC3 PC4  
 Standard deviation 19810.47483 5220.99979 998.60488 772.76111  
 Proportion of Variance 0.92859 0.06450 0.00236 0.00141  
 Cumulative Proportion 0.92859 0.99309 0.99545 0.99686

PC5 PC6 PC7 PC8 PC9  
 Standard deviation 646.24559 490.19201 409.49121 322.17051 312.13495  
 Proportion of Variance 0.00099 0.00057 0.00040 0.00025 0.00023  
 Cumulative Proportion 0.99785 0.99842 0.99881 0.99906 0.99929

PC10 PC11 PC12 PC13 PC14  
 Standard deviation 303.05889 272.19239 212.91015 180.07516 147.22423  
 Proportion of Variance 0.00022 0.00018 0.00011 0.00008 0.00005  
 Cumulative Proportion 0.99951 0.99968 0.99979 0.99987 0.99992

PC15 PC16  
 Standard deviation 139.72407 122.73943  
 Proportion of Variance 0.00005 0.00004  
 Cumulative Proportion 0.99996 1.00000



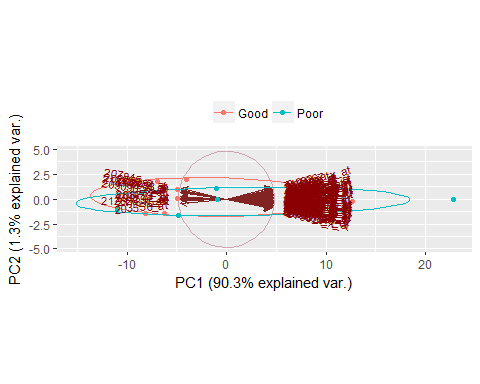
The mapping does show the directionality of the principal components but the grouping is still quite hard to interpret.

### Top 100 – Log Transformed and Scaled

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |

The log transformed and scaled top 100 genes when compared to the full set of genes do not provide a significant advancement on grouping and predictive accuracy.

Importance of components:  
 PC1 PC2 PC3 PC4 PC5  
 Standard deviation 9.501008 1.122253 1.049703 1.002219 0.9346711  
 Proportion of Variance 0.902690 0.012590 0.011020 0.010040 0.0087400  
 Cumulative Proportion 0.902690 0.915290 0.926300 0.936350 0.9450900  
 PC6 PC7 PC8 PC9 PC10  
 Standard deviation 0.9245676 0.8447119 0.8257738 0.7861957 0.7459383  
 Proportion of Variance 0.0085500 0.0071400 0.0068200 0.0061800 0.0055600  
 Cumulative Proportion 0.9536300 0.9607700 0.9675900 0.9737700 0.9793300  
 PC11 PC12 PC13 PC14 PC15  
 Standard deviation 0.7293633 0.6685855 0.6316559 0.6042104 0.5688881  
 Proportion of Variance 0.0053200 0.0044700 0.0039900 0.0036500 0.0032400  
 Cumulative Proportion 0.9846500 0.9891200 0.9931100 0.9967600 1.0000000  
 PC16  
 Standard deviation 6.579005e-15  
 Proportion of Variance 0.000000e+00  
 Cumulative Proportion 1.000000e+00



The directionality of the components is divided into 2 main directions, but it did not improve the grouping of patient prognosis results.

## 

## Conclusion

The variations on the data transformation performed on the leukemia dataset shows that a outlying data points can unduly influence the results of the PCA component results and it should be good practice to scale, center, and in some cases also log transform the data to reduce the influence of these data points to influence the final results.

The assignment prescribed that a cluster analysis should also be performed on the dataset and that it should also be separated into the two groups that was used for the PCA.

Some code for this part of the assignment is included in the markdown file. Unfortunately, the results were not conclusive or complete enough to include in this report.

# 2 – Lecture Recording

## Dataset

The dataset for this part of the assignment comprises of lecture recording workflow processing information for the last month. Due to the University of Cape Town midyear vacation falling in this period the number of scheduled recordings for this period is below the expected production volume. Using this subset of data limited by the date range and the reduction of processed recordings will provide an opportunity to create a flow to be able to analyze the workflow process before the start of term and the volume increase that will push the system to capacity.

Each recording has a workflow operation associated with it on completion of recording and then after review and trimming takes place it is published to the lecture recording tool in Vula.

The workflow data describes the type of operation (edit, publish), the type of track configuration and the queue time before processing and the processing time.

The operational information describes the operations that is performed for each of the workflows.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| tracks | type | count | mean.queue\_time | mean.run\_time | mean.full\_time |
| A | edit | 19 | 7.3684 | 3706.4 | 3713.8 |
| A | publish | 5 | 6.4000 | 3493.4 | 3499.8 |
| B | edit | 24 | 6.9583 | 9838.2 | 9845.2 |
| B | publish | 17 | 7.0000 | 12127.5 | 12134.5 |
| C | edit | 19 | 7.6316 | 14124.6 | 14132.2 |
| C | publish | 18 | 7.5556 | 14228.6 | 14236.1 |

Summary of Workflows

|  |
| --- |
| **type tracks duration queue run operations day hour** |
| edit :62 A:24 Min. : 0 Min. : 4.00 Min. : 32 Min. : 4.0 2:22 13 :19 |
| publish:40 B:41 1st Qu.: 3600 1st Qu.: 6.00 1st Qu.: 3517 1st Qu.:39.0 3:15 14 :15 |
| C:37 Median : 6600 Median : 7.00 Median : 8024 Median :41.0 4:25 10 :14 |
| Mean : 8216 Mean : 7.25 Mean :10340 Mean :39.3 5:18 12 : 9 |
| 3rd Qu.: 9000 3rd Qu.: 8.75 3rd Qu.:13274 3rd Qu.:41.0 6:10 8 : 6 |
| Max. :25201 Max. :10.00 Max. :57596 Max. :43.0 7:12 9 : 6 |
| (Other):33 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| type | count | mean.queue\_time | mean.run\_time | mean.full\_time |
| encode | 111 | 63.4759 | 1799.512432 | 1862.9883 |
| inspect | 122 | 4.8023 | 1836.707426 | 1841.5097 |
| prepare-av | 294 | 3.3084 | 1244.919524 | 1248.2279 |
| editor | 37 | 4.5976 | 830.501351 | 835.0989 |
| normalize-audio | 37 | 4.1483 | 654.284865 | 658.4331 |
| composite | 360 | 4.3615 | 363.622264 | 367.9838 |
| segment-video | 74 | 3.5653 | 304.147203 | 307.7125 |
| compose | 403 | 4.6836 | 51.144799 | 55.8284 |
| publish-engage | 40 | 4.3926 | 35.315300 | 39.7079 |
| waveform | 61 | 4.3897 | 10.748705 | 15.1384 |
| silence | 61 | 4.8200 | 10.298984 | 15.1190 |
| image | 148 | 4.0062 | 10.370709 | 14.3769 |
| segmentpreviews | 74 | 3.8805 | 8.172892 | 12.0534 |
| archive | 196 | 4.6306 | 7.422255 | 12.0529 |
| publish-configure | 98 | 4.6028 | 6.763429 | 11.3663 |
| cleanup | 101 | 4.6648 | 4.575446 | 9.2402 |
| ingest-download | 61 | 4.9244 | 0.082328 | 5.0067 |
| defaults | 40 | 4.6178 | 0.284350 | 4.9022 |
| comment | 98 | 4.6775 | 0.179184 | 4.8567 |
| include | 469 | 4.5313 | 0.167808 | 4.6991 |
| analyze-tracks | 355 | 4.5188 | 0.174623 | 4.6934 |
| tag | 575 | 4.4767 | 0.152317 | 4.6291 |
| publish-youtube | 37 | 4.4282 | 0.000000 | 4.4282 |
| series | 61 | 2.9586 | 0.511984 | 3.4706 |
| export-wf-properties | 61 | 2.8971 | 0.196787 | 3.0939 |

|  |
| --- |
| **type queue run full day hour tracks duration** |
| analyze-tracks :355 Min. : 0.0 Min. : 0.0 Min. : 0.0 2: 845 0 : 9 A: 946 Min. : 300 |
| archive :196 1st Qu.: 4.3 1st Qu.: 0.1 1st Qu.: 4.6 3: 595 1 : 21 B:1596 1st Qu.: 3600 |
| cleanup :101 Median : 4.6 Median : 0.2 Median : 5.1 4:1027 6 : 5 C:1432 Median : 7200 |
| comment : 98 Mean : 6.0 Mean : 258.3 Mean : 264.3 5: 701 7 : 21 Mean : 8400 |
| compose :403 3rd Qu.: 4.9 3rd Qu.: 10.4 3rd Qu.: 15.1 6: 336 8 :112 3rd Qu.: 9000 |
| composite :360 Max. :4035.1 Max. :11566.7 Max. :11569.5 7: 470 9 :205 Max. :25201 |
| defaults : 40 10:418 |
| editor : 37 11:261 |
| encode :111 12:305 |
| export-wf-properties: 61 13:317 |
| image :148 14:597 |
| include :469 15:403 |
| ingest-download : 61 16:297 |
| inspect :122 17:199 |
| normalize-audio : 37 18:209 |
| prepare-av :294 19:248 |
| publish-configure : 98 20:140 |
| publish-engage : 40 21: 77 |
| publish-youtube : 37 22: 63 |
| segment-video : 74 23: 67 |
| segmentpreviews : 74 |
| series : 61 |
| silence : 61 |
| tag :575 |
| waveform : 61 |

|  |  |
| --- | --- |
|  |  |
|  | |
|  |  |

## Conclusion

The initial data analysis and plotting of the relationship between run time and duration, indicated that even without any processing being required for that step (some steps are skipped in the workflow which is dependent on the type of tracks provided for that workflow) it will always queue up for around **6.0** seconds. The actual running time of the operation greatly depends on the duration of the video and the type of tracks to be processed and accordingly is the main indicator of processing time for the workflow.

Using Self Organising Maps (SOM) and Multidimentional Scaling (MDS) on this data might inform us on why certain processes takes much longer than the calculated rate should be and it is likely that this is due to concurrency of workflows running on the same day and in overlapping times.

The code for the SOM models and MDS mapping is included in the code base but could not be included in this report due to data transformation issues and time restraints.

# 3 - Notes

This assignment was helpful in exploring the concepts of unsupervised learning methods and in highlighting the shortcomings in my understanding of some of the concepts; more than that it showed the deficiency in my processing of data to provide the models with the structure that is then used to construct the models and produce the output for interpretation.

All the code, data and related items can be found on GitHub.

GitHub: <https://github.com/TurRil/STA5077Z-UNSUPERVISED-LEARNING>

# 4 - References

* Project Description (assignment.pdf, DataScienceProject.html, leukemia\_array.txt), Author: Juwa Nyirenda. Last Accessed: 24-Jul-2017 17:51. Vula Resources.
* R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: <https://www.R-project.org/>.
* Hadley Wickham (2017). tidyverse: Easily Install and Load 'Tidyverse' Packages. R package version 1.1.1. <https://CRAN.R-project.org/package=tidyverse>
* Kirill Müller and Hadley Wickham (2017). tibble: Simple Data Frames. R package version 1.3.3. <https://CRAN.R-project.org/package=tibble>
* Hadley Wickham (2017). tidyr: Easily Tidy Data with 'spread()' and 'gather()' Functions. R package version 0.6.1. <https://CRAN.R-project.org/package=tidyr>
* Hadley Wickham, Jim Hester and Romain Francois (2016). readr: Read Tabular Data. R package version 1.0.0. <https://CRAN.R-project.org/package=readr>
* Lionel Henry and Hadley Wickham (2017). purrr: Functional Programming Tools. R package version 0.2.2.2. <https://CRAN.R-project.org/package=purrr>
* Hadley Wickham and Romain Francois (2016). dplyr: A Grammar of Data Manipulation. R package version 0.5.0. <https://CRAN.R-project.org/package=dplyr>
* Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K.(2016). cluster: Cluster Analysis Basics and Extensions. R package version 2.0.5.
* Venables, W. N. & Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0
* Jan de Leeuw, Patrick Mair (2009). Multidimensional Scaling Using Majorization: SMACOF in R. Journal of Statistical Software, 31(3), 1-30. URL <http://www.jstatsoft.org/v31/i03/>.
* Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K.(2016). cluster: Cluster Analysis Basics and Extensions. R package version 2.0.5.
* Venables, W. N. & Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0
* Jan de Leeuw, Patrick Mair (2009). Multidimensional Scaling Using Majorization: SMACOF in R. Journal of Statistical Software, 31(3), 1-30. URL <http://www.jstatsoft.org/v31/i03/>
* Max Kuhn. Contributions from Jed Wing, Steve Weston, Andre Williams, Chris Keefer, Allan Engelhardt, Tony Cooper, Zachary Mayer, Brenton Kenkel, the R Core Team, Michael Benesty, Reynald Lescarbeau, Andrew Ziem, Luca Scrucca, Yuan Tang, Can Candan and Tyler Hunt. (2017). caret: Classification and Regression Training. R package version 6.0-76. <https://CRAN.R-project.org/package=caret>