**MAC-LRN**

**MILESTONE 2**

Arevalo, Mark Anthony A.

Nablo, Juan Paolo J.

Vivo, Roi Emmanuel A.

Analyze your 2 datasets by applying k-means clustering and self-organizing maps techniques. Have a detailed discussion on each dataset and be guided by the following:

**Data set 1: Spambase**

**A. k-Means Clustering**

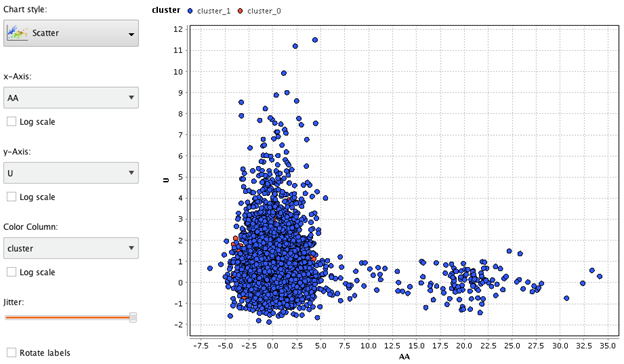
a) K=2

Figure 1: Clustering – k=2 – Scatter Plot

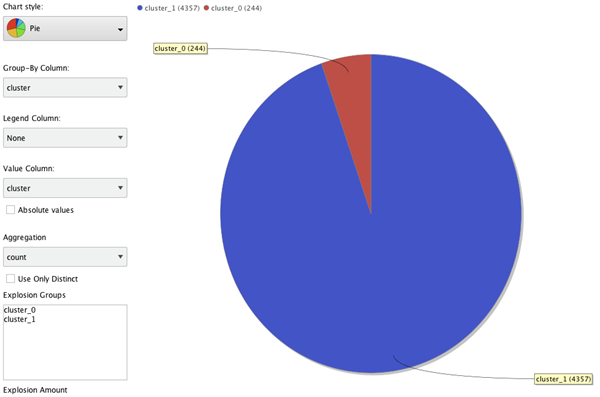


Figure 2: Clustering – k=2 – Pie Chart

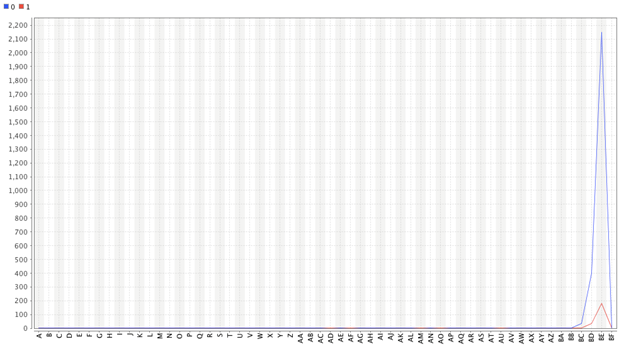


Figure 3: Plot View

Max value for Centroid 0 and 1 is feature BD.

b) K=20

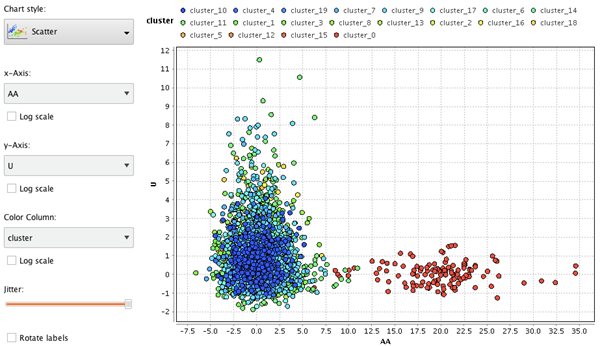


Figure 4: Clustering – k=20 – Scatter Plot

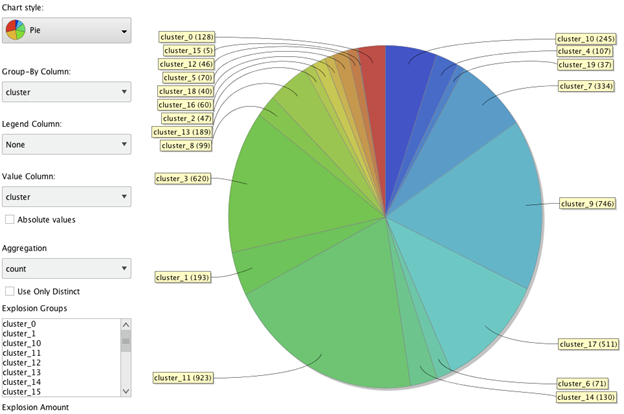


Figure 5: Clustering – k=20 – Pie Chart



Figure 6: K=20 Plot View

**K=20**

1. Find the best set of clusters by performing several experiments. Once the number of clusters have been finalized, describe each cluster according to a) its distinct features and b) the kind of samples (or class distribution) it has. Compare the different clusters.

a. Based on the data, cluster 0 from the spam mail indicator feature dictates that this is supposedly not spam because the contradicts to the set of clusters which are related to cluster 10, and cluster 0 means that one sample is a spam letter. On the other hand, the remaining clusters would indicate that these are not spam letters.

The process was done by testing numerous K values. Though the results were not appropriate if the values were set to 1-10. Once K was set to have 20, the clusters were more noticeable. The features of every cluster are almost the same. The feature BD is dominant for all clusters.

b. On the data, cluster 0 indicates that these samples are spam letters, different with the rest of the samples in the other clusters. This means that the remaining clusters are not spam letters. In one of the features most data for cluster 0 have similar results in the AA feature. One sample indicator is that a feature labeled the AA feature indicates that an email is a spam when the keyword “WORD” appears in the body more than 10 times. Most samples in cluster 0 have values ranging from 10 to 30 in the AA feature.

2. Show the number of clusters as well as the number of samples in each cluster.

20 clusters (cluster\_0 – cluster\_19)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Cluster** | **Number of samples** | **Cluster** | **Number of samples** | **Cluster** | **Number of samples** |
| 0 | 128 | 8 | 99 | 16 | 60 |
| 1 | 193 | 9 | 746 | 17 | 511 |
| 2 | 47 | 10 | 245 | 18 | 40 |
| 3 | 620 | 11 | 923 | 19 | 37 |
| 4 | 107 | 12 | 46 |  |  |
| 5 | 70 | 13 | 189 |  |  |
| 6 | 71 | 14 | 130 |  |  |
| 7 | 334 | 15 | 5 |  |  |

Figure 7: 20 clusters (cluster\_0 – cluster\_19)

3. Explain the reasons for choosing such a number of clusters. You can apply the objective function (mentioned in the slides) to assess the quality of each cluster. You can mention the results of your previous experiments that may help justify for the choice of k clusters.

Starting with a number of 1 to 10 for the number of clusters did not show appropriate results. It only displayed 1 major cluster with a similar color. Choosing the k as 20 displayed several clusters that had different colors from each other. Also, even though there are different clusters, only 1 cluster was different from the others based from color shades. This solution gave a result that could be used for clustering two types of samples easily. The solution may become better if k is greater than 20.

4. Summary and Conclusion of the experiments

The k-means clustering is a process that groups samples together with similar characteristics from other clusters with different qualities. Based on the data recorded for Milestone 1, the top 2 features for PCA was referred to for the x and y axis for the Cluster Plot. The charts above have been generated by doing this process. Based on the data gathered, the clusters are more similar than dissimilar. Because of this issue, the class is difficult to distinguish.

**B. SOM**

1. Train a 10x10 SOM map for as many iterations/cycles as possible in order to generate a good and specialized set of weights for each node. Include the generated map in the document.

1. 1,000 Training Rounds c. 15,000 Training Rounds
2. 10,000 Training Rounds

2. Make sure that the SOM map is labelled or color-coded (based on labels). A map is well-organized if nodes having the same label are close to each other.

a. 1,000 Training Rounds

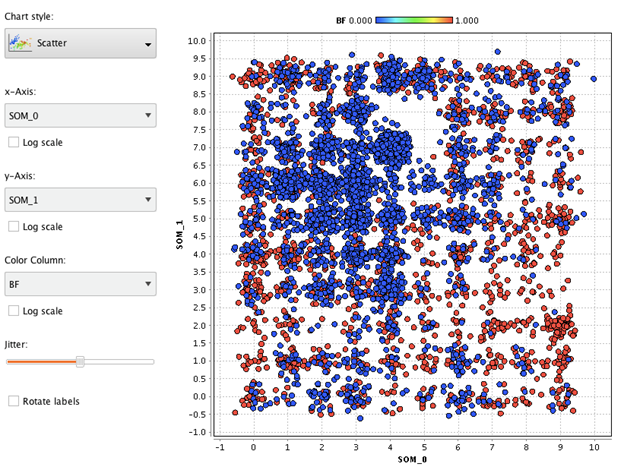


Figure 8: Scatter Plot -1,000 Training Rounds

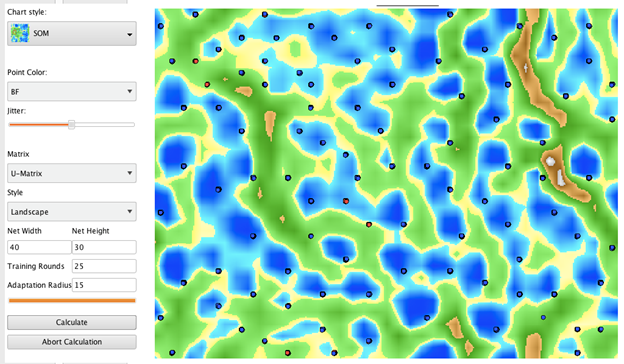


Figure 9: SOM Plot - 1,000 Training Rounds

b. 10,000 Training Rounds

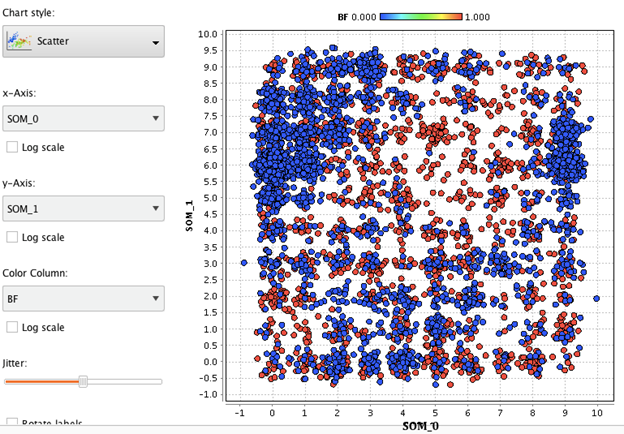


Figure 10: Scatter Plot - 10,000 Training Rounds

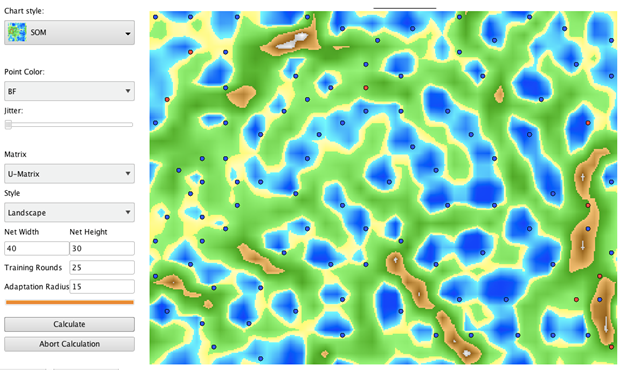


Figure 11: SOM - 10,000 Training Rounds

c. 15,000 Training Rounds

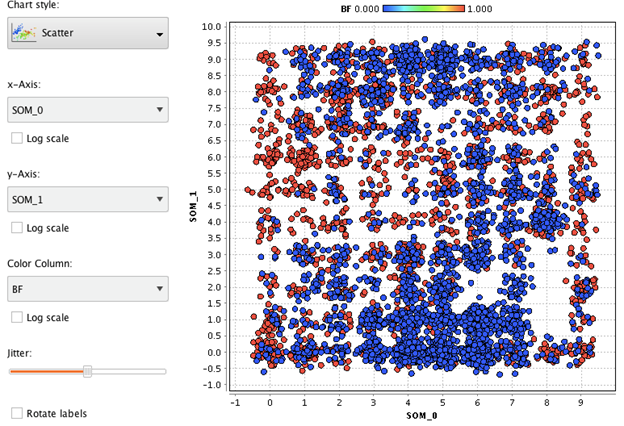


Figure 12: Scatter Plot - 15,000 Training Rounds

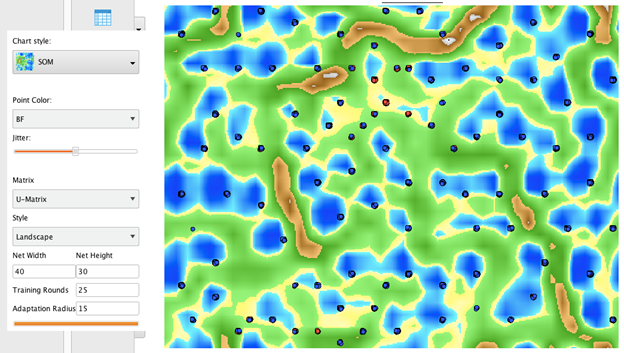


Figure 13: SOM - 15,000 Training Rounds

3. Interpret the results of the map. Analyze the kind of samples that were assigned on nodes that are close in the map. For example, determine the kind of samples that were assigned in the upper left corner nodes of the map. Have some analyses on several areas of the map.

Based on all the maps, it is difficult to distinguish what kind of samples are assigned to the parts of the map especially with no labels on it. Just like the clustering process, the maps are unorganized though the map with 1000 training rounds is more “clean” compared to the others.

4. Summary and Conclusion of the experiments

Based on the results, Using SOM with different training rounds such as 1000, 10000, and 15000 on the dataset did not organize the map. However, when compared to all the outputs, SOM with 1000 training rounds look more organized than the other training rounds.

Even with unsupervised learning, the dataset was still unorganized due to similarities of the data in the features. The data can still be used for analysis, but it will not be as accurate for predictions.

**Dataset #2: Diabetic Retinopathy**

**A. k-Means Clustering**

**a.) K=2**

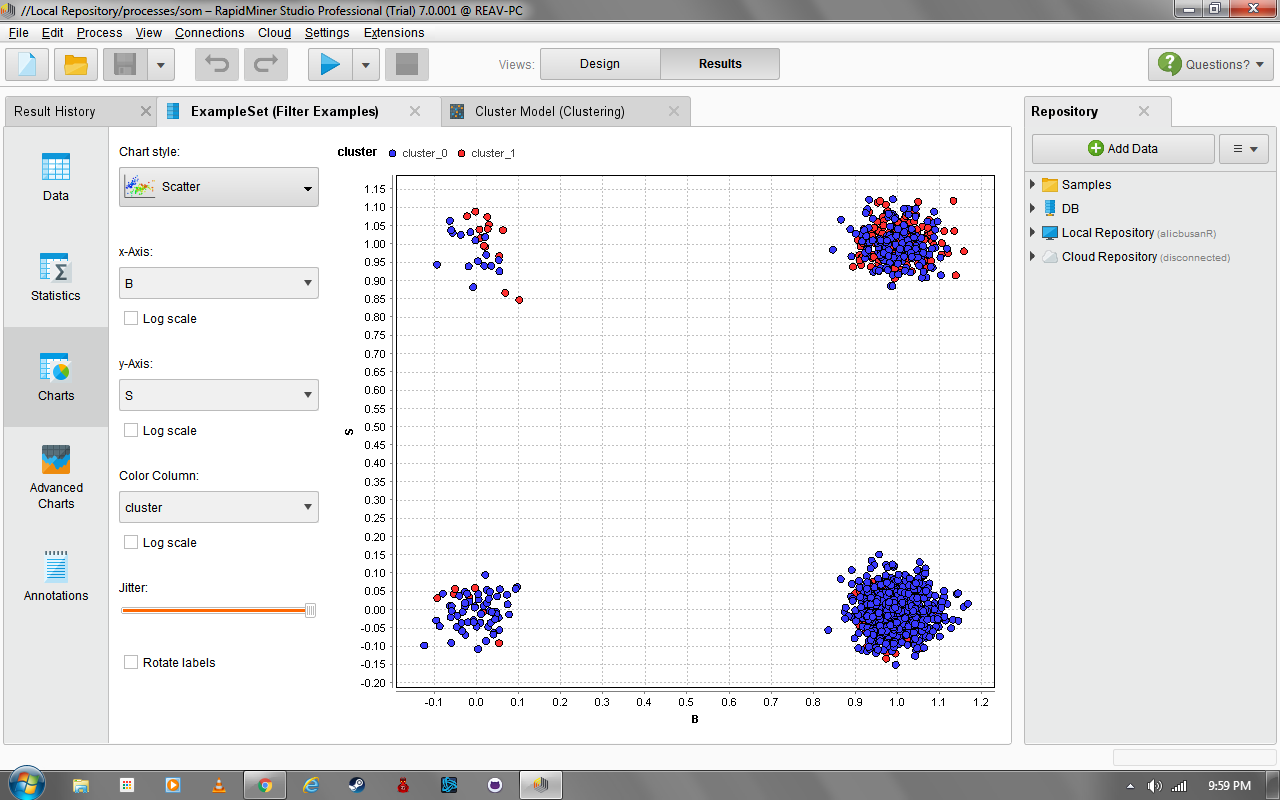


Figure 14: Cluster Plot (x=B,y=S)

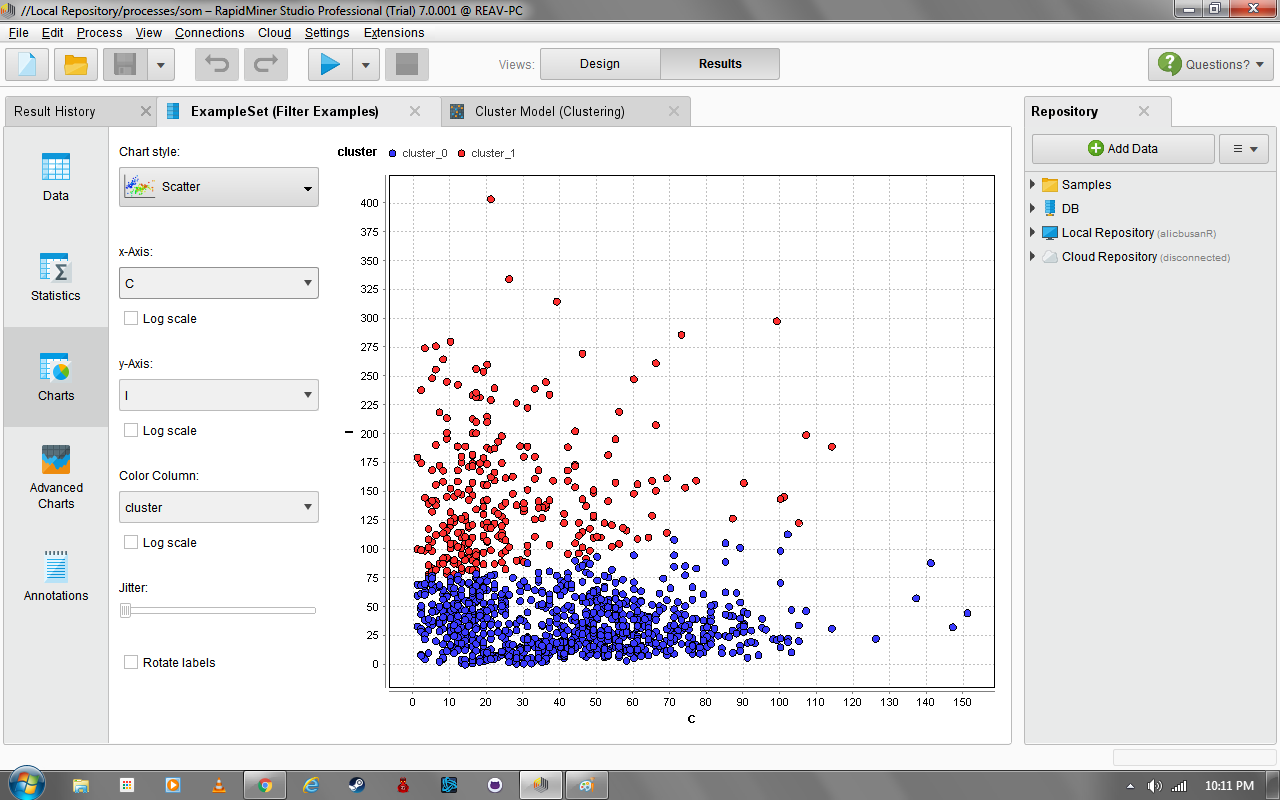


Figure 15: Cluster Plot (x=C, y=I)

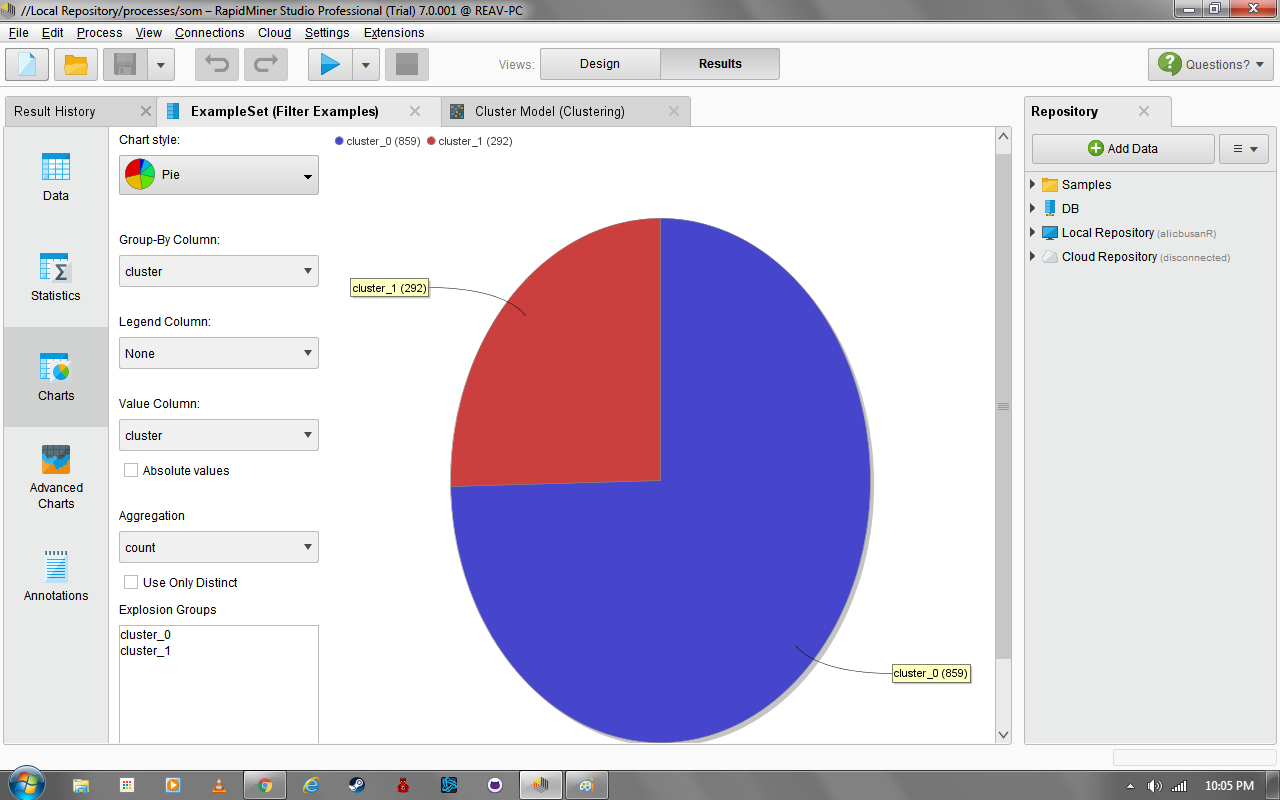


Figure 16: Pie Chart



Figure 17: Plot View

1. Find the best set of clusters by performing several experiments. Once the number of clusters has been finalized, describe each cluster according to a) its distinct features and b) the kind of samples (or class distribution) it has. Compare the different clusters.

a. Based on Figure 17, cluster 1 has the feature I as the most dominant feature in the cluster. As for cluster 0, it has feature C.

b. The clusters are based on the features I and C. Clusters 0 and 1 are a mix of both output classes/labels (diabetes or no diabetes).

2. Show the number of clusters as well as the number of samples in each cluster.

2 clusters (cluster\_0 - cluster 1)

|  |  |
| --- | --- |
| Cluster | Number of samples |
| 0 | 292 |
| 1 | 859 |

3. Explain the reasons for choosing such a number of clusters. You can apply the objective function (mentioned in the slides) to assess the quality of each cluster. You can mention the results of your previous experiments that may help justify for the choice of k clusters.

There are only 2 possible values for the class, which are 1 and 0. These are boolean values to indicate of a patient has diabetes or not. This results to making k=2. Since the clusters are distinguishable from one another, this is a good reason to choose k as 2.

4. Summary and Conclusion of the experiments

The k-means clustering is a process that groups samples together with similar characteristics from other clusters with different qualities. Based on the results, k-means clustering successfully grouped the samples into two clusters. However, the centroids of these 2 clusters are similar that at some point some samples cannot be differentiated from each other.

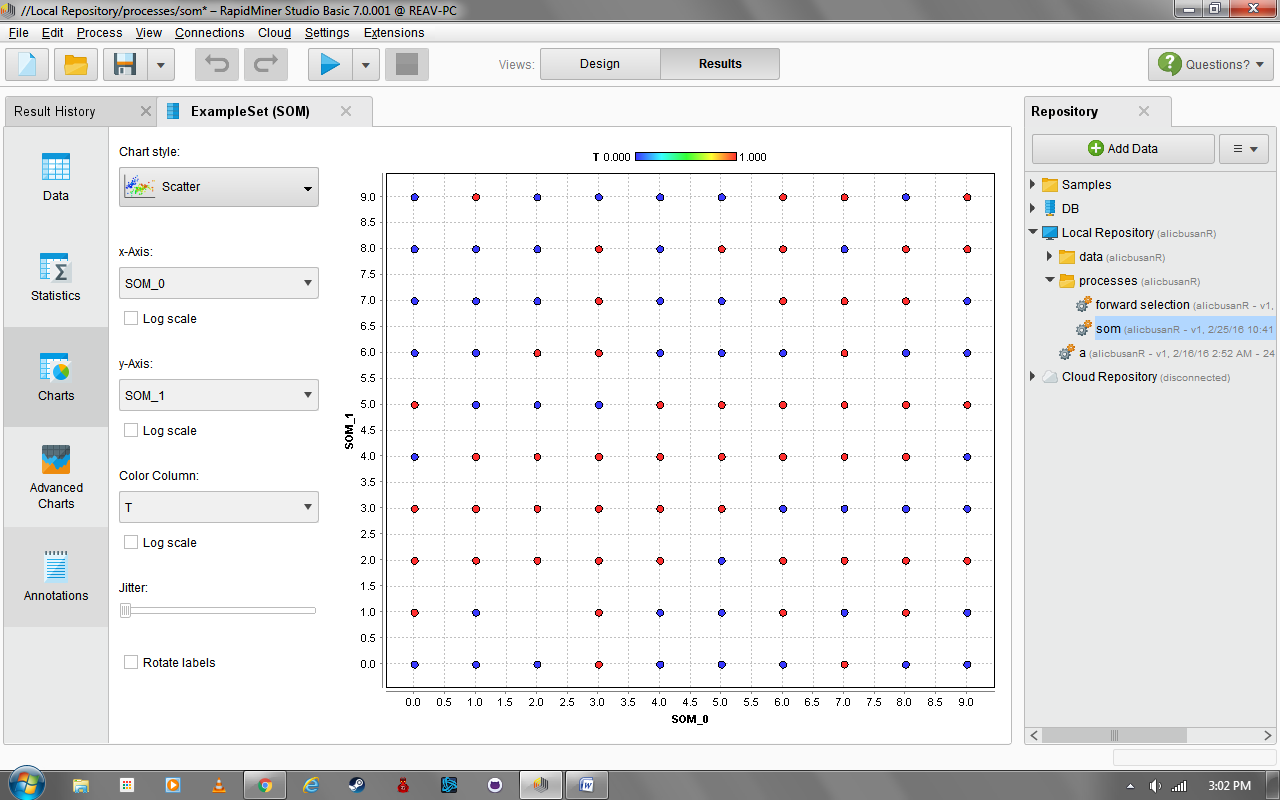
**B. SOM**

1. Train a 10x10 SOM map for as many iterations/cycles as possible in order to generate a good and specialized set of weights for each node. Include the generated map in the document.

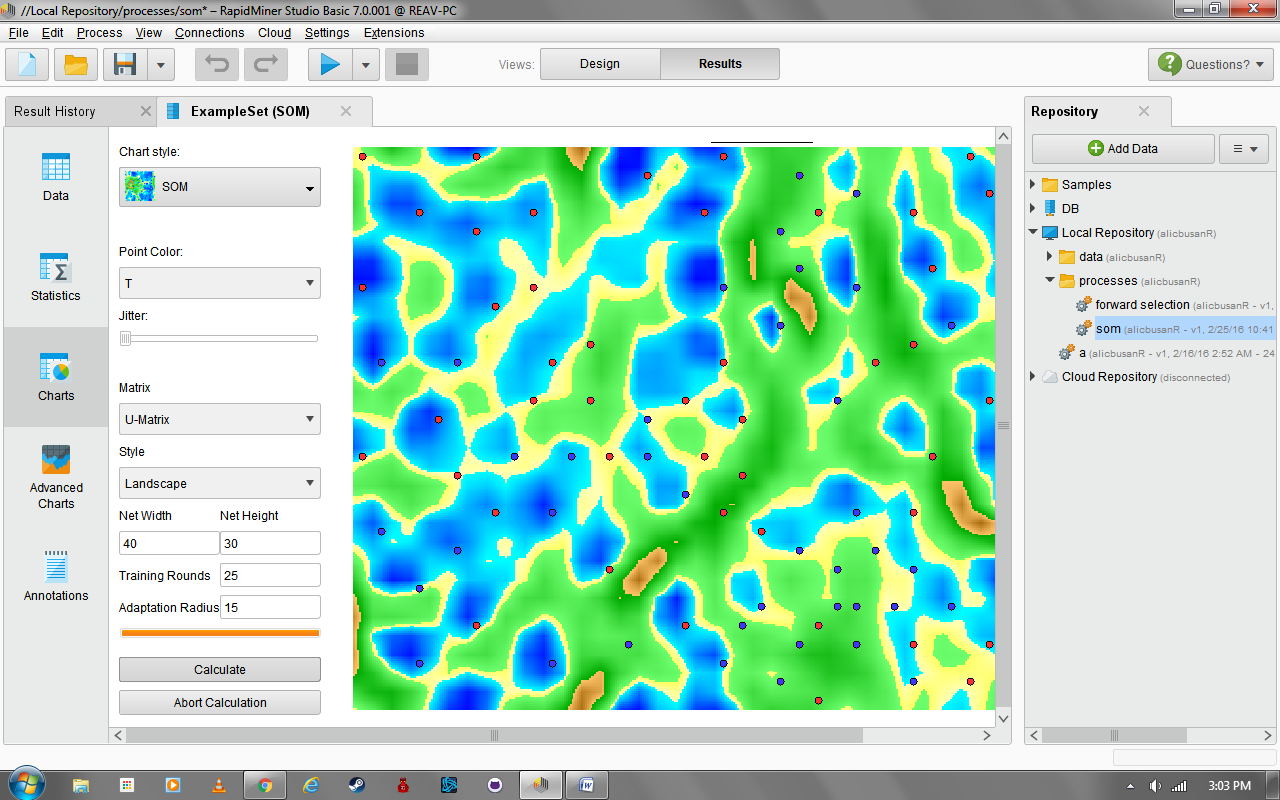
1. 1,000 Training Rounds c. 15,000 Training Rounds
2. 10,000 Training Rounds

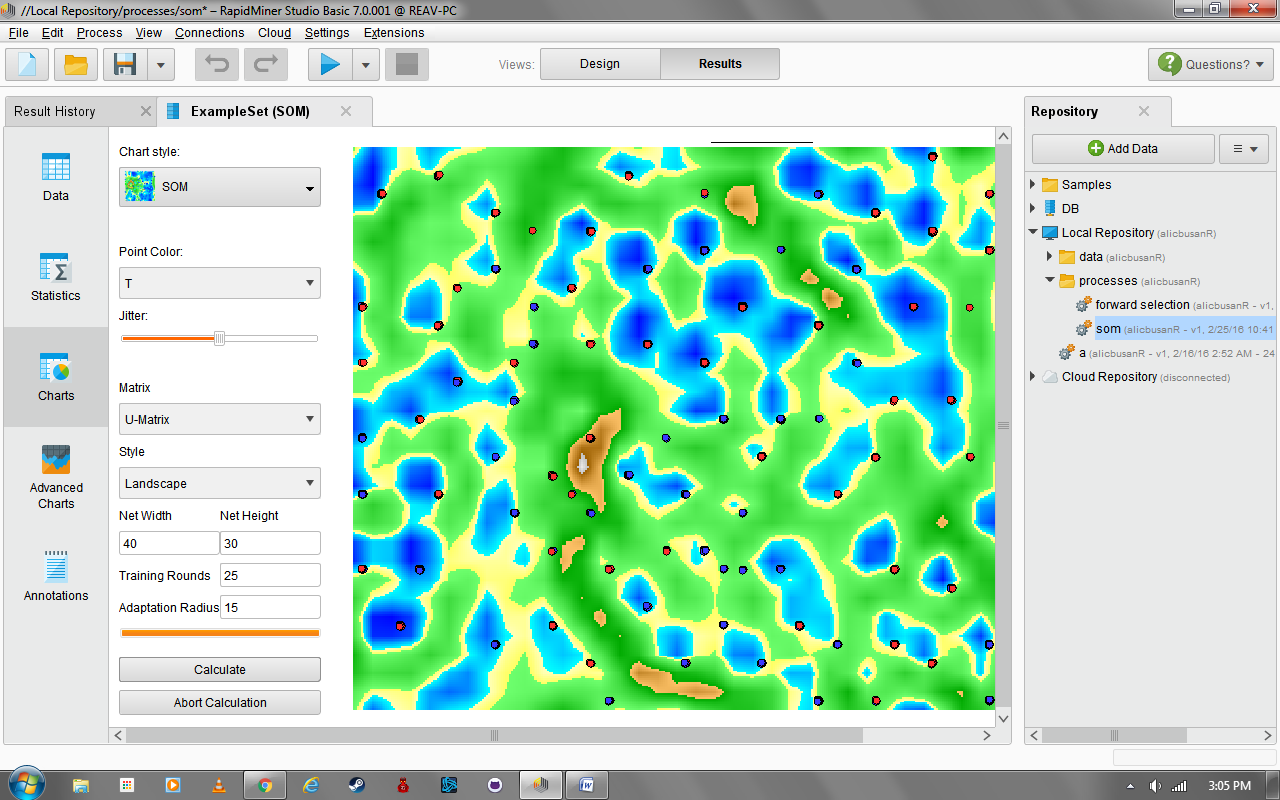
2. Make sure that the SOM map is labelled or color-coded (based on labels). A map is well-organized if nodes having the same label are close to each other.

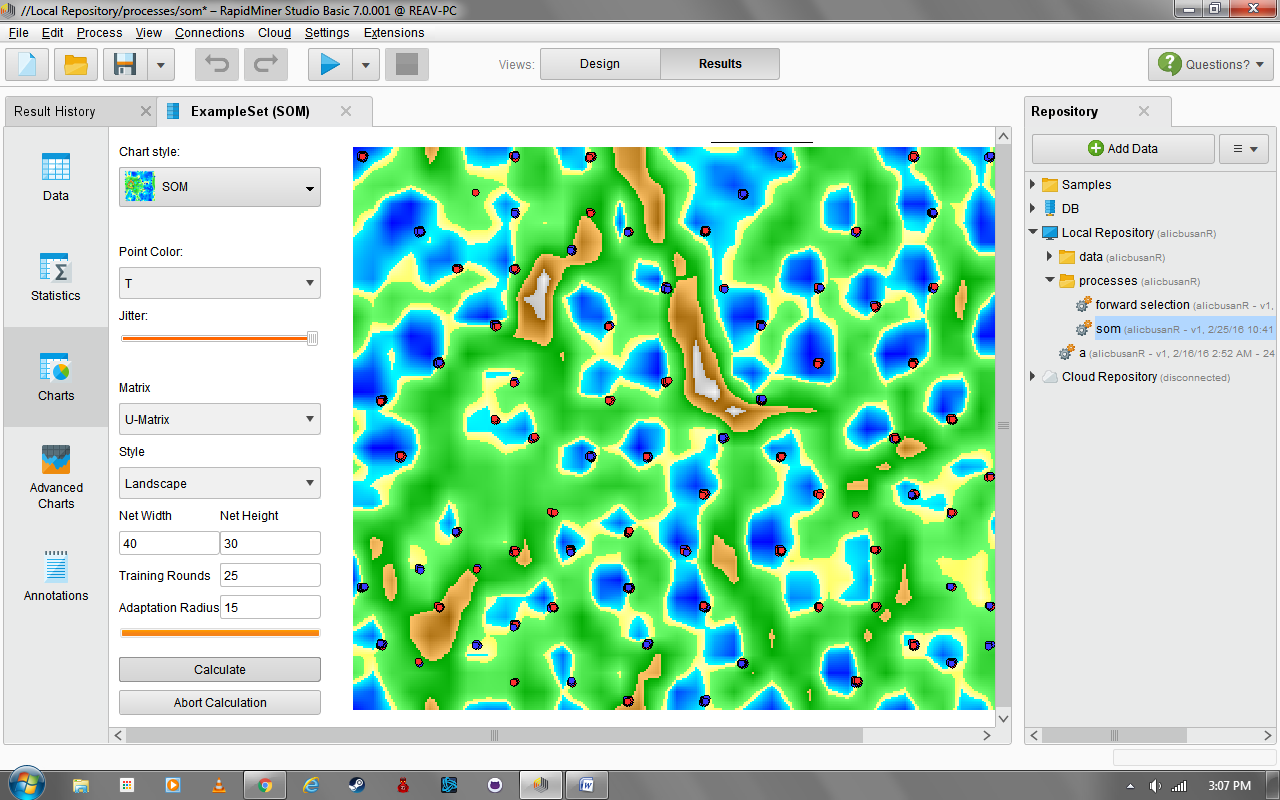
1. 1,000 Training Rounds



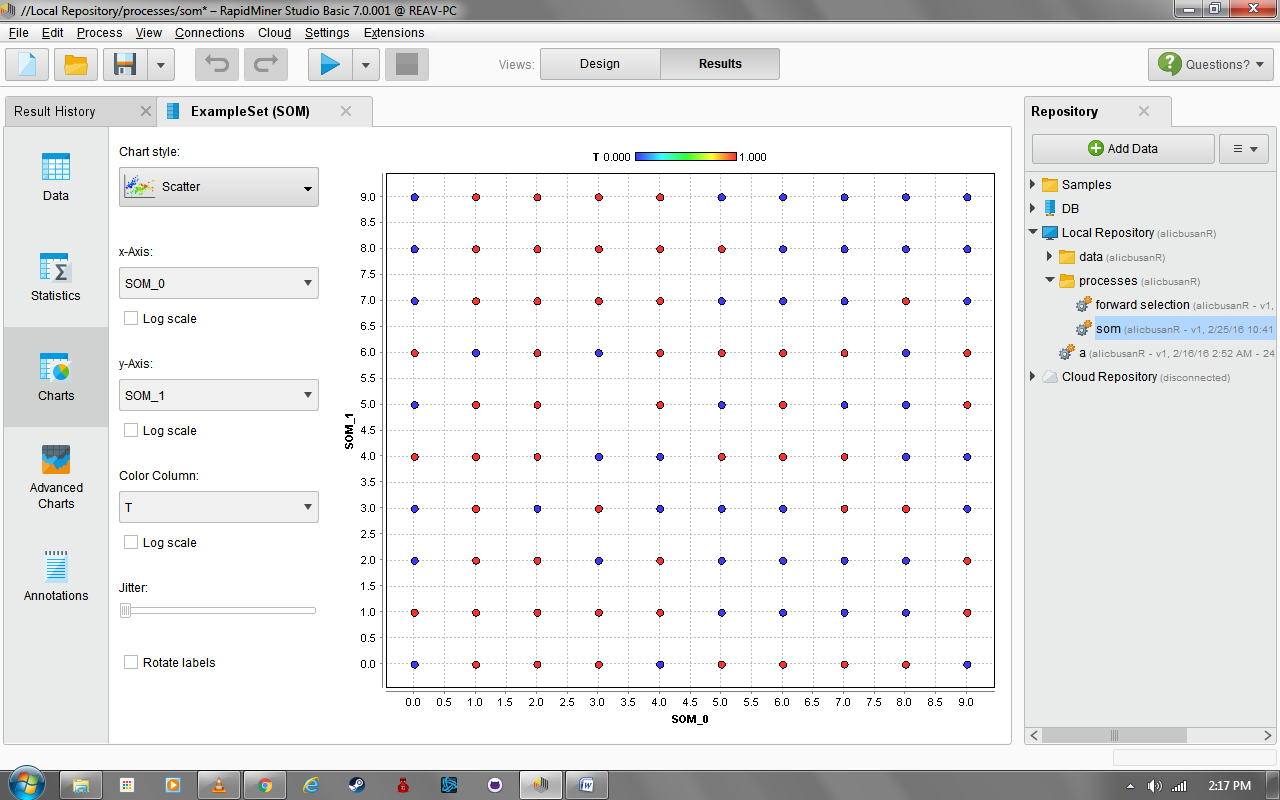
**Figure 18: Scatter Plot - 1,000 Training Rounds - Minimum Jitter**

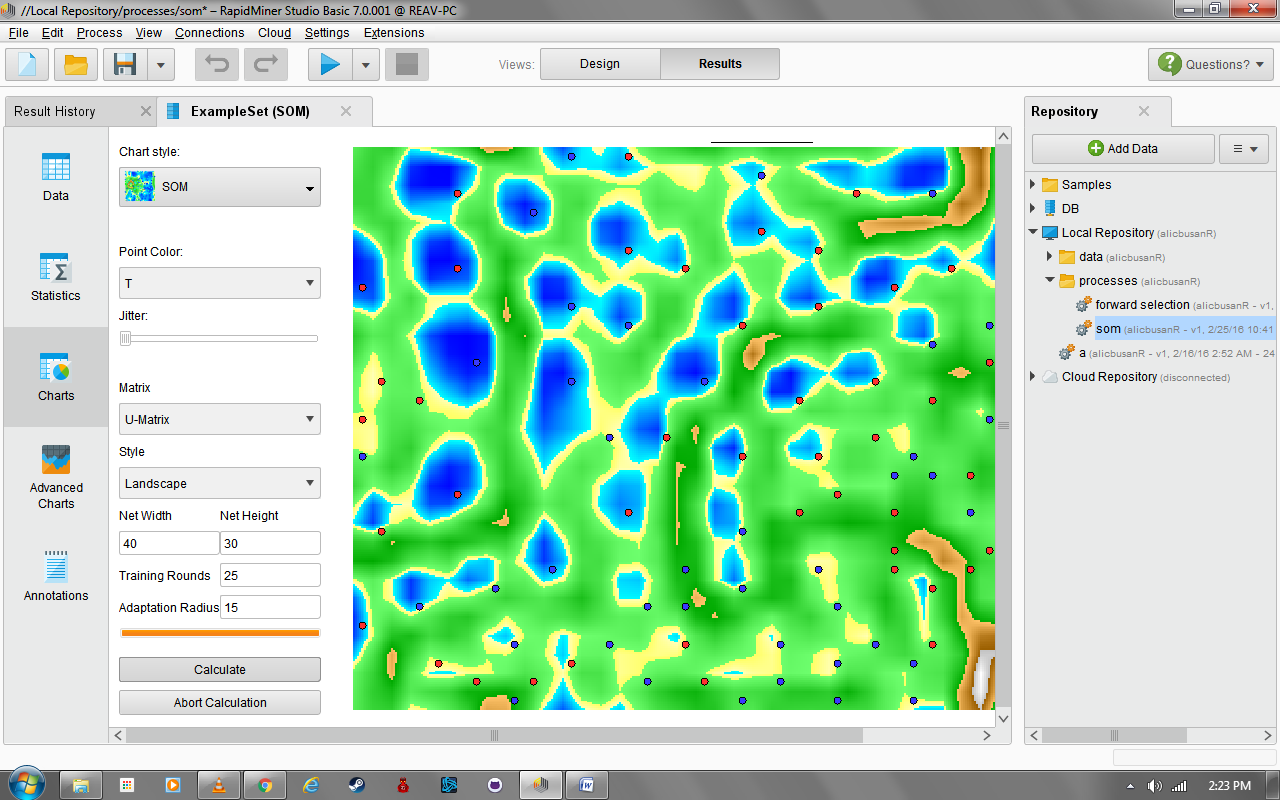
**Figure 19: SOM - 1,000 Training Rounds - Minimum Jitter**

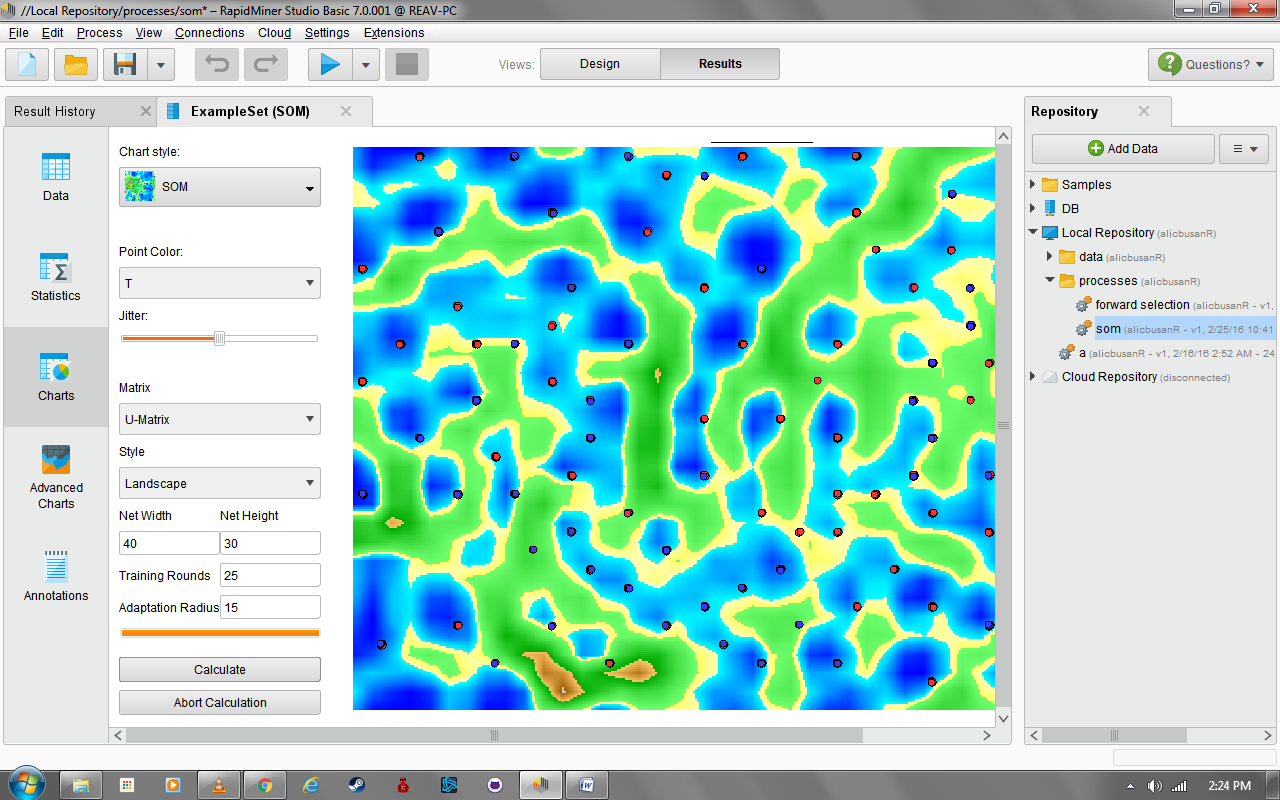
**Figure 20: SOM - 1,000 Training Rounds - Medium Jitter**

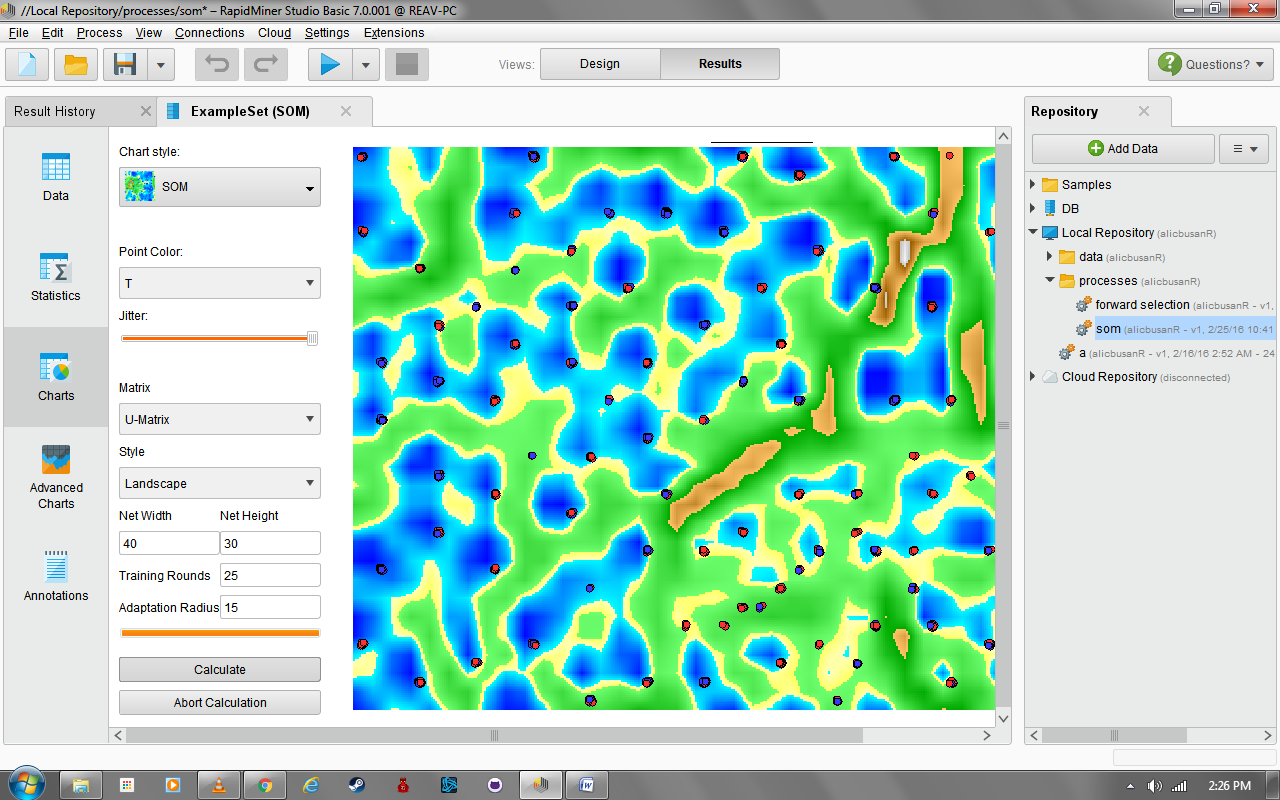
**Figure 21: SOM - 1,000 Training Rounds - Maximum Jitter**

1. 10,000 Training Rounds

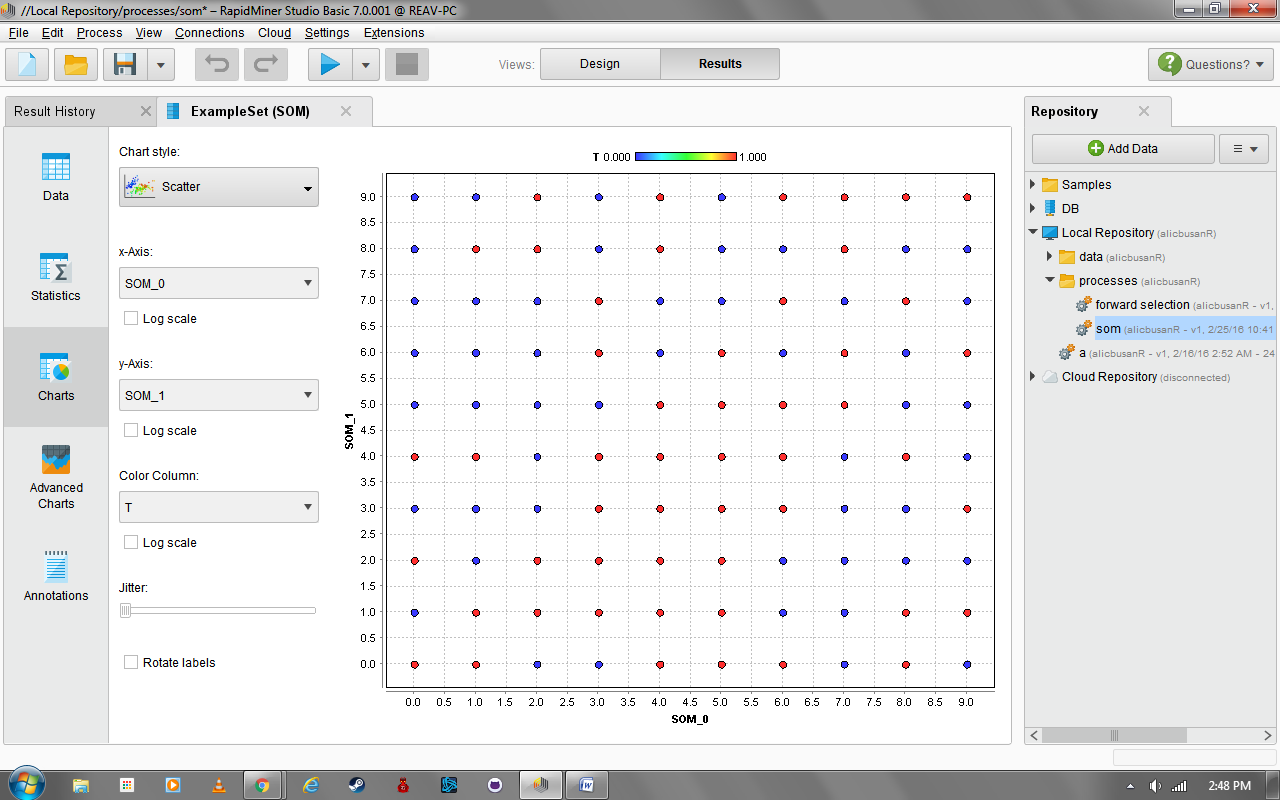
**Figure 22: Scatter Plot - 10,000 Training Rounds - Minimum Jitter**

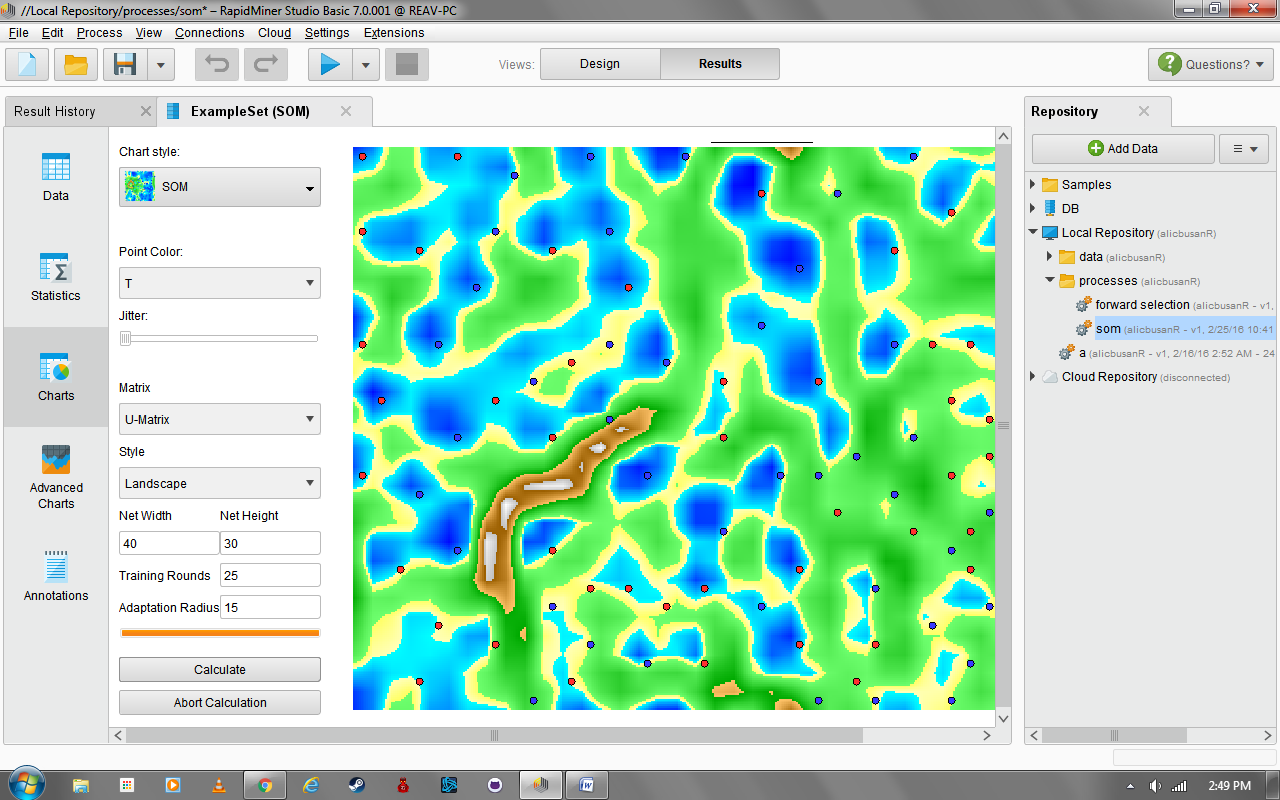
**Figure 23: SOM - 10,000 Training Rounds - Minimum Jitter**

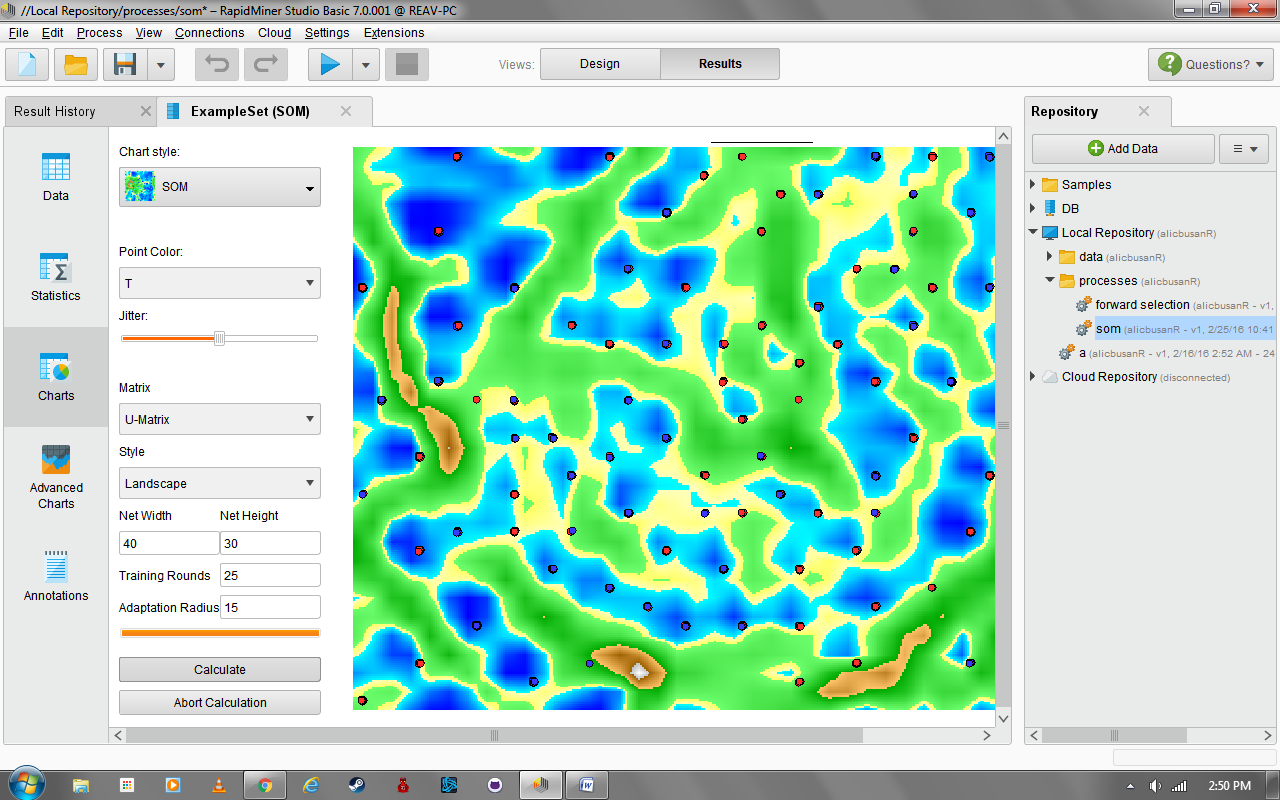
**Figure 24: SOM - 10,000 Training Rounds - Medium Jitter**

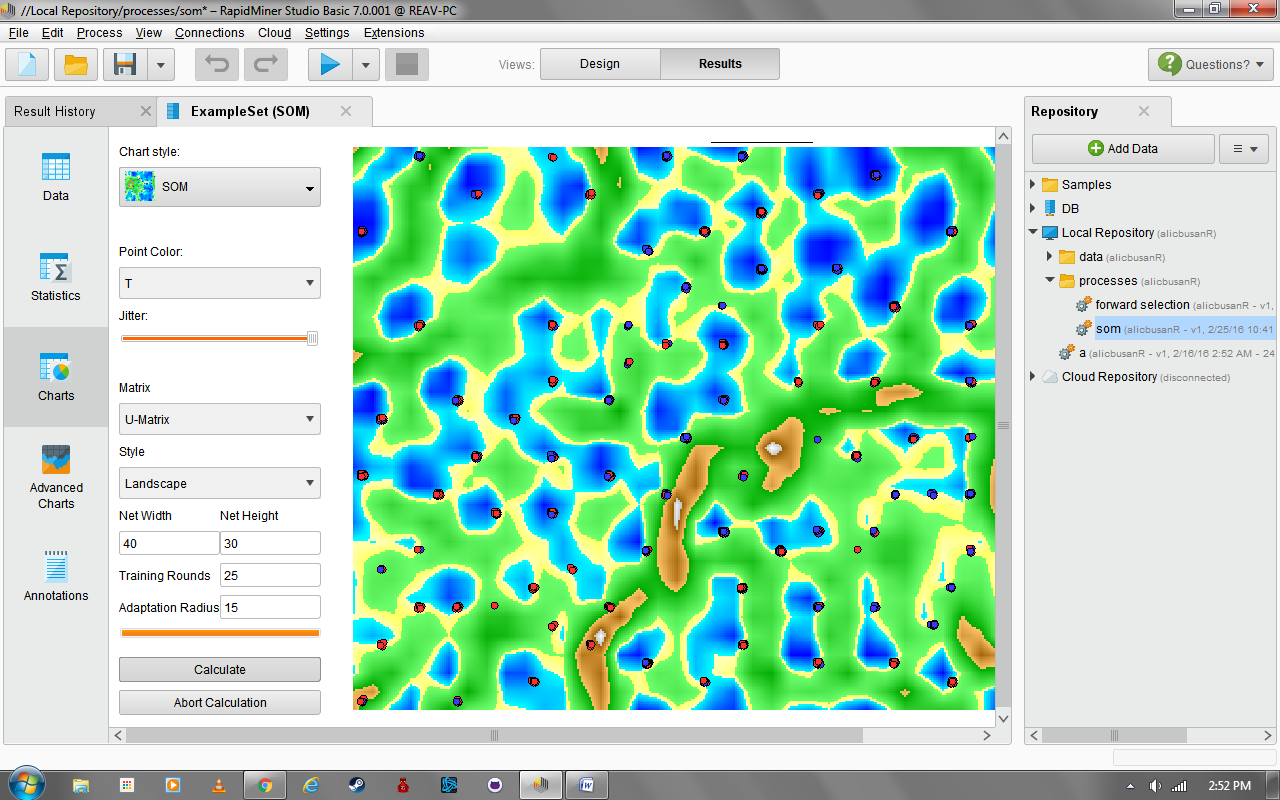
**Figure 25: SOM - 10,000 Training Rounds - Maximum Jitter**

1. 15,000 Training Rounds

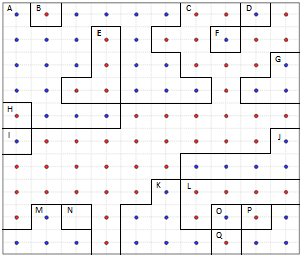
**Figure 26: Scatter Plot - 15,000 Training Rounds - Minimum Jitter**

**Figure 27: SOM - 15,000 Training Rounds - Minimum Jitter**

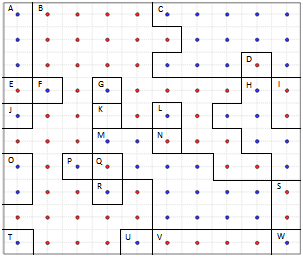
**Figure 28: SOM - 15,000 Training Rounds - Medium Jitter**

**Figure 29: SOM - 15,000 Training Rounds - Maximum Jitter**

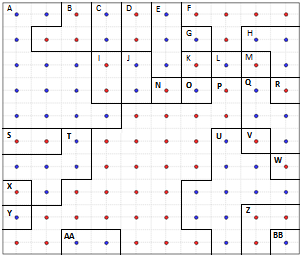
Disclaimer: The scatter plot blocks (figure …) were bordered to simplify the image and emphasize the nodes. Every block grouped together with the same color was marked with letters to easily identify each node from the rest.



**Figure 30: Simplified Scatter Plot - 1,000 Training Rounds Simplified**



**Figure 31: Simplified Scatter Plot - 10,000 Training Rounds Simplified**



**Figure 32: Simplified Scatter Plot - 15,000 Training Rounds Simplified**

3. Interpret the results of the map. Analyze the kind of samples that were assigned on nodes that are close in the map. For example, determine the kind of samples that were assigned in the upper left corner nodes of the map. Have some analyses on several areas of the map.

Based on the results, taking the SOM map of dataset 2 with 1000, 10000, and 15000 training rounds produced a varying number of nodes. With 1000 training rounds, the results displayed 17 distinct nodes. At 10000, 23 nodes were produced. And finally at 15000 training rounds, 28 nodes were retrieved.

4. Summary and Conclusion of the experiments

All of the information here were taken at jitter set to a minimum producing a 10x10 SOM map with clear, specific, and distinct nodes that allows the classification of each sample more precise and accurate. Although the actual relationship between each node were not discussed, this still shows a limited display of the SOM of diabetes.