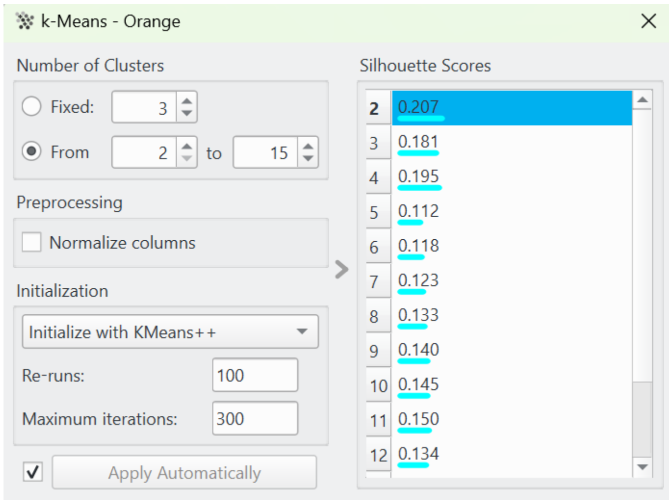
Clustering-Based Identification of Sepsis and Non-Sepsis Patient Groups Using Clinical and Laboratory Data

Aim of Project:

The aim of this study is to assess whether C-reactive protein (CRP) levels and routine hemogram parameters, which include lymphocyte count, platelet count, and others, either individually or in combination, can effectively differentiate sepsis from non-sepsis SIRS in ICU patients at the time of admission. The study focuses upon identifying cost-effective as well as readily available blood markers. Critical care settings see these markers help differentiate sepsis from systemic inflammatory response syndrome.

1. K-Means Clustering

The silhouette score analysis showed that two clusters achieved the highest value (0.207), making it the most suitable choice for grouping patients. While the score indicates moderate separation, it is stronger than for other cluster counts, meaning the data forms two reasonably distinct patient groups.



k-Means - Orange

Number of Clusters

○ Fixed: 3

● From 2 to 15

Preprocessing

□ Normalize columns

Initialization

Initialize with KMeans++

Re-runs: 100

Maximum iterations: 300

☑ Apply Automatically

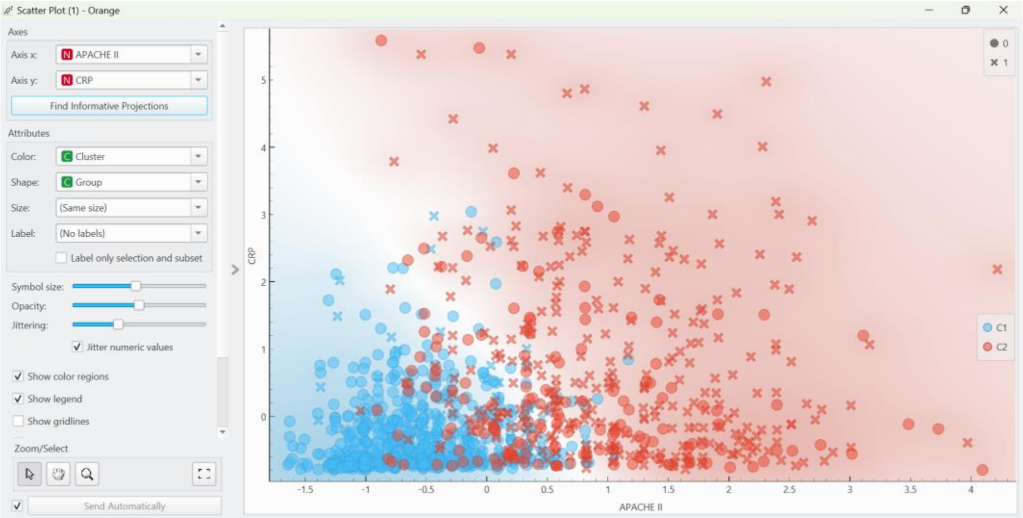
Silhouette Scores

2	0.207
3	0.181
4	0.195
5	0.112
6	0.118
7	0.123
8	0.133
9	0.140
10	0.145
11	0.150
12	0.134

Results Interpretation

APACHE II vs CRP (Scatter Plot):

The higher-risk cluster (C2) is concentrated in the range of APACHE II scores above 13 and CRP levels often above 3 mg/dL, with many values between 3–6 mg/dL. The lower-risk cluster (C1) mainly falls below APACHE II scores of 13 and CRP levels of 0–2 mg/dL. APACHE II shows a stronger influence on separating the groups, indicating that higher disease severity is closely linked to sepsis cases. CRP adds to the distinction, as sepsis cases tend to have elevated CRP, but with some overlap between groups.



Scatter Plot (1) - Orange

Axes

Axis x: APACHE II

Axis y: CRP

Find Informative Projections

Attributes

Color: Cluster

Shape: Group

Size: (Same size)

Label: (No labels)

☐ Label only selection and subset

Symbol size:

Opacity:

Jittering:

☑ Jitter numeric values

☑ Show color regions

☑ Show legend

☐ Show gridlines

Zoom/Select

☑ Send Automatically

● 0
✖ 1

● C1
● C2

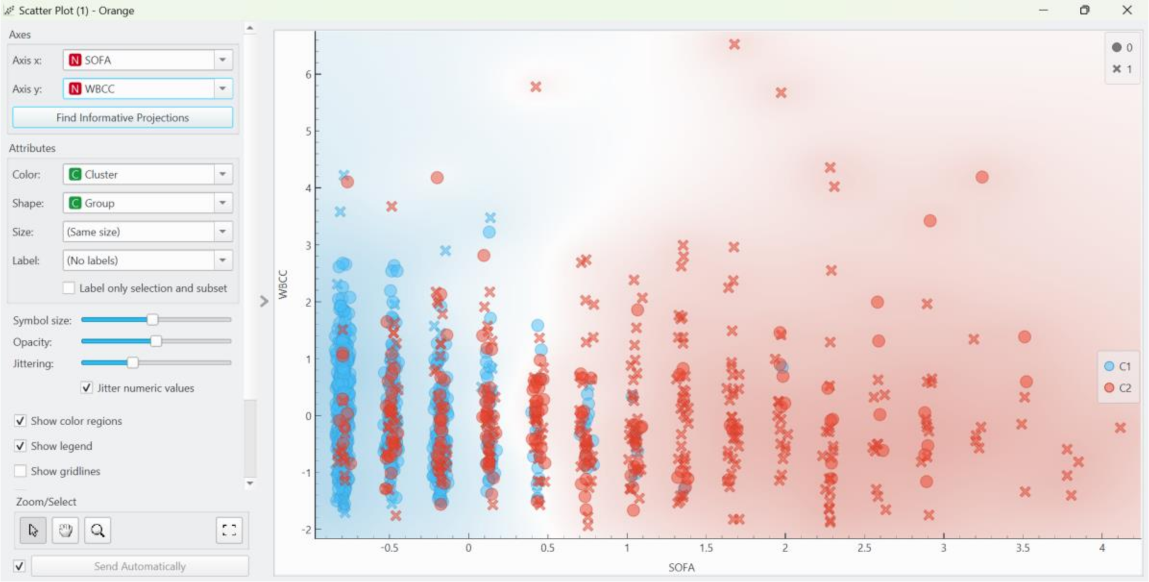
Scatter Plot (SOFA vs WBCC):

The plot shows the relationship between SOFA score and white blood cell count (WBCC) in separating clusters linked to sepsis and non-sepsis cases.

Patients in the higher-risk cluster (C2) generally have SOFA scores above 4 and WBCC values frequently above 12 ×10³/µL, with some cases extending much higher. This pattern reflects more severe organ dysfunction and elevated immune response, both of which are more common in sepsis.

The lower-risk cluster (C1) is concentrated in the range of SOFA scores between 0–3 and WBCC values around 6–10 ×10³/µL, which aligns with less severe illness and a lower likelihood of sepsis.

SOFA score shows a stronger influence on the separation than WBCC, but WBCC still provides additional discriminative value, as very high counts are more frequent in the sepsis-heavy cluster.



Scatter Plot (1) - Orange

Axes

Axis x: SOFA

Axis y: WBCC

Find Informative Projections

Attributes

Color: Cluster

Shape: Group

Size: (Same size)

Label: (No labels)

☐ Label only selection and subset

Symbol size:

Opacity:

Jittering:

☑ Jitter numeric values

☑ Show color regions

☑ Show legend

☐ Show gridlines

Zoom/Select

☑ Send Automatically

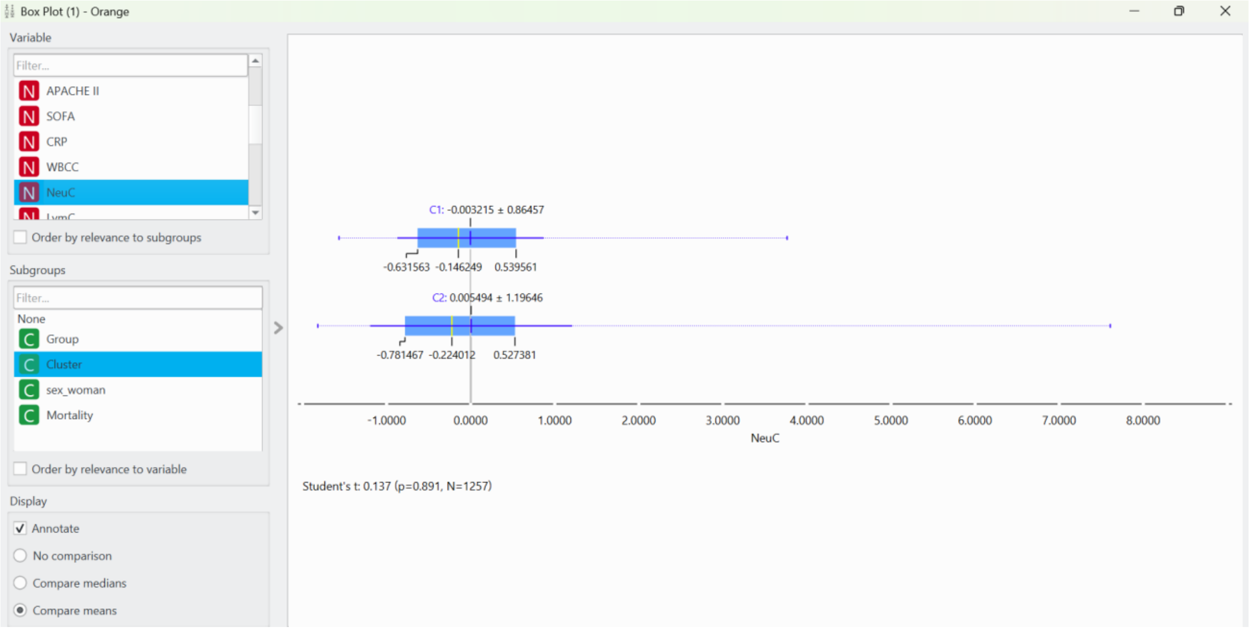
● 0
✖ 1

● C1
● C2

NeuC (Box Plot by Cluster):

The plot shows the distribution of neutrophil count (NeuC) for the two clusters. Both clusters have a very similar range and median NeuC values, with substantial overlap. In Cluster 2 (higher-risk/sepsis-heavy), NeuC values are generally between 5–14 ×10³/µL, while in Cluster 1 (lower-risk/non-sepsis-heavy), values also fall in a similar range.

The statistical test (p = 0.891) confirms that there is no significant difference in neutrophil counts between the clusters. This indicates that NeuC does not have a strong impact on distinguishing sepsis from non-sepsis patients in this dataset.



Box Plot (1) - Orange

Variable

Filter...

N APACHE II
N SOFA
N CRP
N WBCC
N NeuC
N LymC

☐ Order by relevance to subgroups

Subgroups

Filter...

None
C Group
C Cluster
C sex_woman
C Mortality

☐ Order by relevance to variable

Display

☑ Annotate
○ No comparison
○ Compare medians
● Compare means

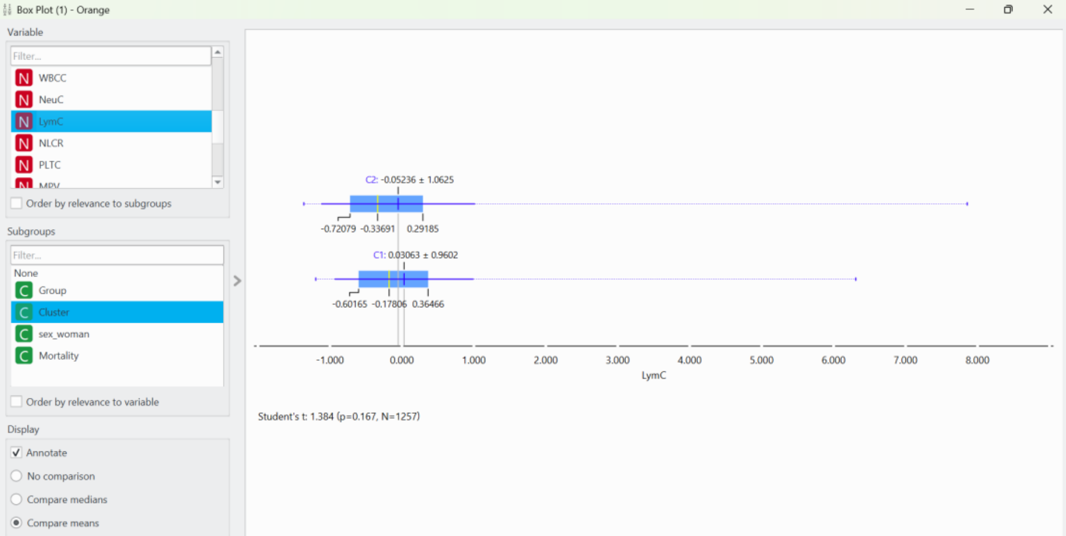
Student's t: 0.137 (p=0.891, N=1257)

LymC (Box Plot by Cluster):

The plot compares lymphocyte count (LymC) across the two clusters. Both clusters show a large degree of overlap, with Cluster 2 (sepsis-heavy) having median lymphocyte counts around 0.3–0.4 ×10³/µL, and Cluster 1 (non-sepsis-heavy) also showing a similar median.

Values in both clusters mostly fall below 1.5 ×10³/µL, and extremely low counts (<0.5 ×10³/µL) occur in both groups. The p-value (0.167) indicates that the difference between clusters is not statistically significant.

This suggests that while low lymphocyte counts are common in sepsis, in this dataset LymC alone does not clearly separate sepsis from non-sepsis patients, and its impact on clustering is minimal.



Box Plot (1) - Orange

Variable: LymC

☐ Order by relevance to subgroups

Subgroups: Cluster

☐ Order by relevance to variable

Display

☑ Annotate

○ No comparison

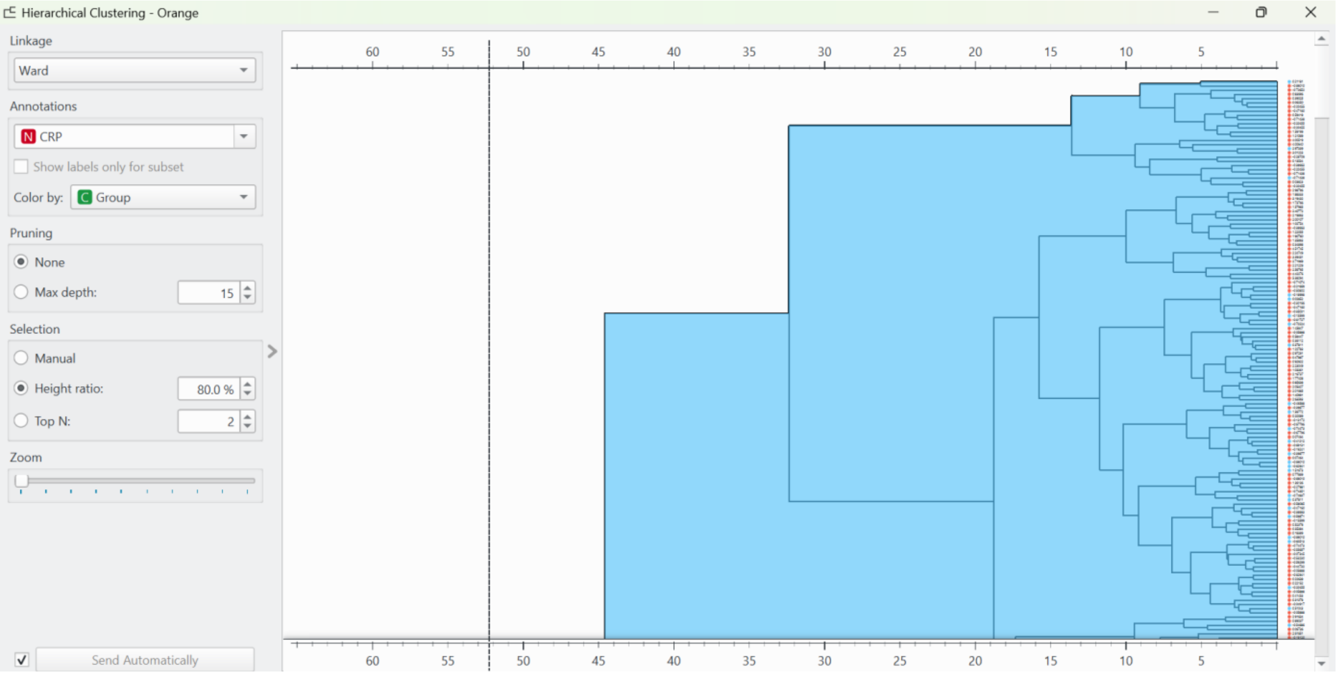
○ Compare medians

● Compare means

Student's t: 1.384 (p=0.167, N=1257)

2. Hierarchical Clustering

The hierarchical clustering, using Ward's linkage method, produced two primary clusters that align with the separation of sepsis and non-sepsis patients. The dendrogram shows clear branching between these groups, suggesting meaningful structural differences in the data based on the included clinical and laboratory variables.



Hierarchical Clustering - Orange

Linkage: Ward

Annotations: CRP

☐ Show labels only for subset

Color by: Group

Pruning

● None

○ Max depth: 15

Selection

○ Manual

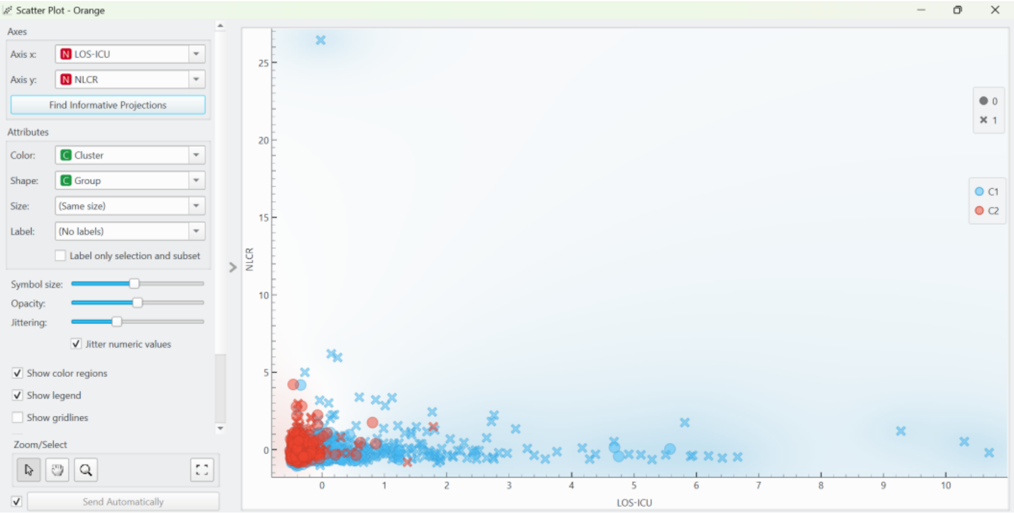
● Height ratio: 80.0 %

○ Top N: 2

☑ Send Automatically

LOS-ICU vs NLCR:

Patients in the higher-risk cluster (C2) tend to have shorter ICU stays (mostly below 3 days) but elevated NLCR values, often above 5, with some extreme values exceeding 20. This pattern indicates that in sepsis cases, an intense inflammatory response can be present early, even before prolonged ICU admission, highlighting NLCR as a potential early marker. In contrast, the lower-risk cluster (C1) has lower NLCR values, generally between 0–5, and shows a broader range of ICU stays.


Scatter Plot - Orange

Axes

Axis x: LOS-ICU

Axis y: NLCR

Attributes

Color: Cluster

Shape: Group

Size: (Same size)

Label: (No labels)

☐ Label only selection and subset

☑ Jitter numeric values

☑ Show color regions

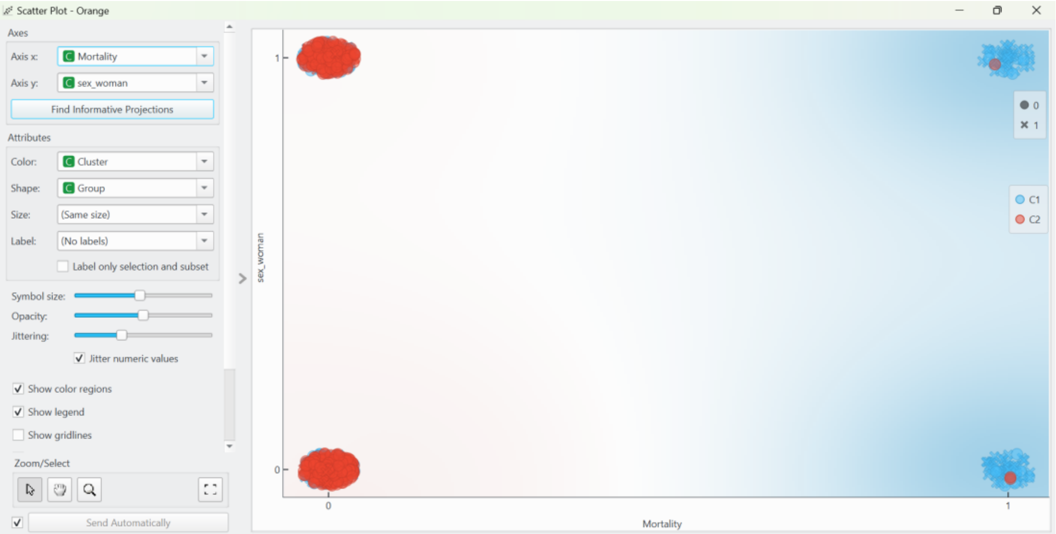
☑ Show legend

☐ Show gridlines

☑ Send Automatically

Mortality vs sex-women:

The mortality pattern shows that most deaths are concentrated in one part of the plot, with males making up a notable proportion of these cases in the higher-risk cluster (C2). The lower-risk cluster (C1) contains most survivors, with both male and female patients. This reinforces the association between the higher-risk cluster and poorer clinical outcomes, further linking it with the sepsis-heavy group.


Scatter Plot - Orange

Axes

Axis x: Mortality

Axis y: sex_woman

Attributes

Color: Cluster

Shape: Group

Size: (Same size)

Label: (No labels)

☐ Label only selection and subset

☑ Jitter numeric values

☑ Show color regions

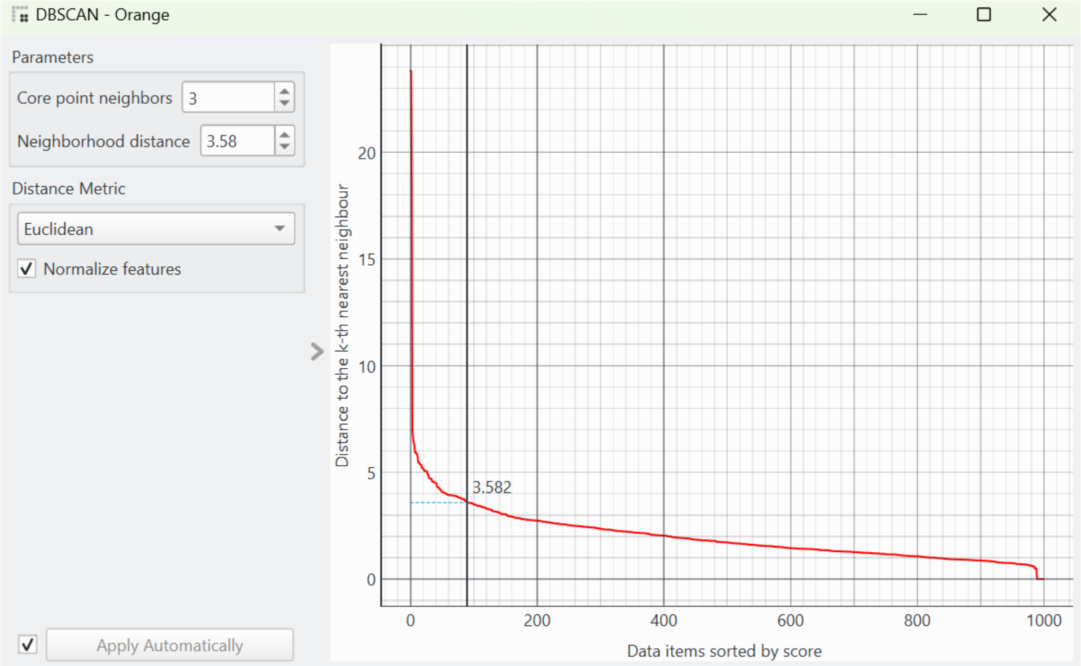
☑ Show legend

☐ Show gridlines

☑ Send Automatically

3. DBSCAN:

DBSCAN parameters (ε = 3.58, minPts = 3) separate patients into two well-defined clusters. C2 forms a dense region in feature space where high severity, high mortality, and worse lab scores overlap. C1 patients cluster around lower severity and lower mortality, meaning less critical conditions.



DBSCAN - Orange

Parameters

Core point neighbors 3

Neighborhood distance 3.58

Distance Metric

Euclidean

☑ Normalize features

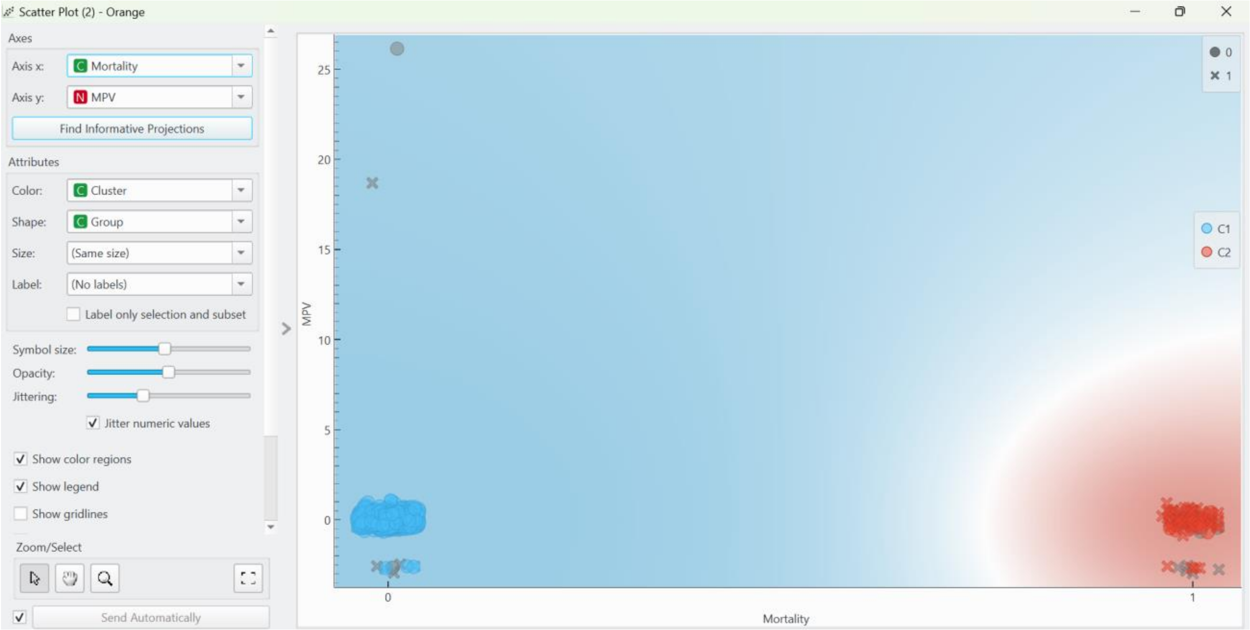
☑ Apply Automatically

Distance to the k-th nearest neighbour
20
15
10
5
0
3.582
0
200
400
600
800
1000
Data items sorted by score

Scatter Plot with Mortality:

In the plot, almost all red points (C2) are on the Mortality = 1 side, meaning these patients died during the study period. Blue points (C1) are concentrated on Mortality = 0, indicating most survived. This direct association between cluster color and mortality rate strongly suggests C2 is higher risk.

MPV values are roughly 0 to 3 fL for most patients, with almost all points densely packed in the low MPV range. Slightly more spread in MPV, with some points reaching slightly higher than C1's range. MPV distribution is tightly clustered at the lower end, around 0–2 fL.



Scatter Plot (2) - Orange

Axes

Axis x: Mortality

Axis y: MPV

Find Informative Projections

Attributes

Color: Cluster

Shape: Group

Size: (Same size)

Label: (No labels)

□ Label only selection and subset

Symbol size:

Opacity:

Jittering:

☑ Jitter numeric values

☑ Show color regions

☑ Show legend

□ Show gridlines

Zoom/Select

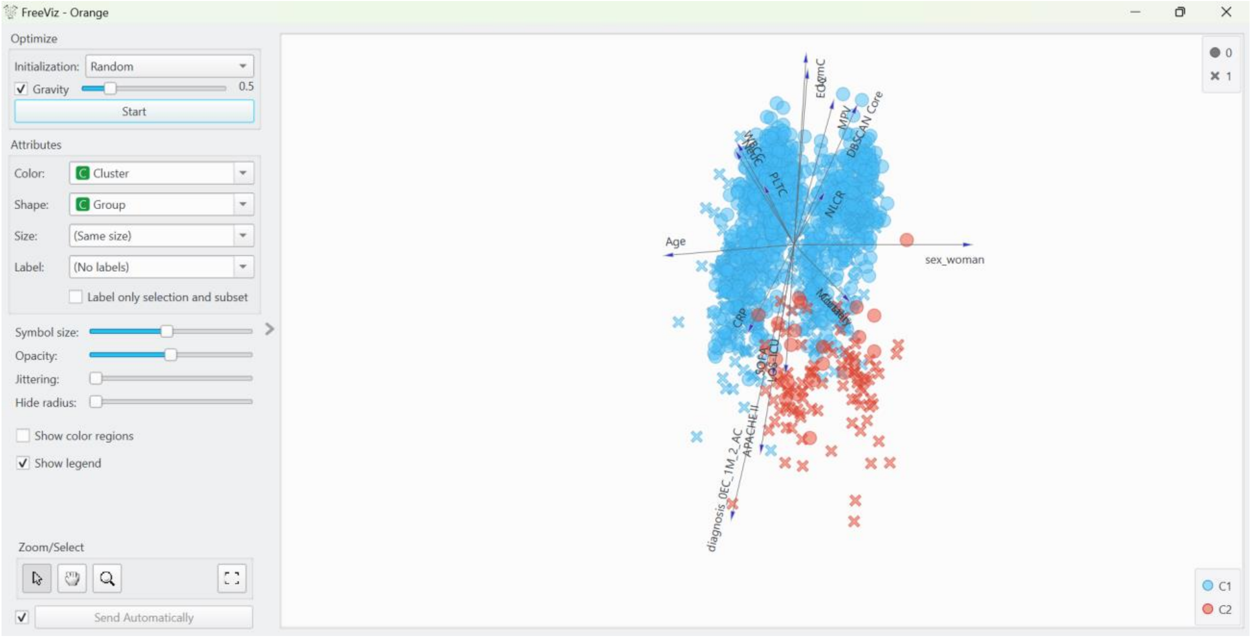
☑ Send Automatically

FreeViz Plot:

C1 (Blue): Densely packed in the **upper-left** region, overlapping mostly with non-sepsis patients.

C2 (Red): Concentrated toward the **lower-right**, containing a larger proportion of sepsis patients and higher mortality cases.

The clear diagonal separation between clusters suggests that the selected features are effective in distinguishing high-risk from low-risk patients.



FreeViz - Orange

Optimize

Initialization: Random

☑ Gravity 0.5

Start

Attributes

Color: Cluster

Shape: Group

Size: (Same size)

Label: (No labels)

□ Label only selection and subset

Symbol size:

Opacity:

Jittering:

Hide radius:

□ Show color regions

☑ Show legend

Zoom/Select

☑ Send Automatically

Summary of Variable Influence (Arrow Directions & Lengths):

Variable	Arrow Direction (Toward Cluster)	Influence Strength	Interpretation
Mortality	C2 (Red, High-Risk)	Very High	Higher mortality rates concentrated in C2, marking it as the severe sepsis group.
SOFA	C2 (Red, High-Risk)	Very High	Higher organ failure scores in C2, strongly linked with sepsis severity.
APACHE II	C2 (Red, High-Risk)	Very High	Indicates greater overall illness severity in C2 patients.
CRP	C2 (Red, High-Risk)	High	Elevated C-reactive protein in C2, indicating strong inflammation.
Diagnosis Codes (DEC1ML, 2AC)	C2 (Red, High-Risk)	Moderate	Certain diagnoses occur more frequently in severe sepsis cases.
sex_woman	C2 (Red, High-Risk)	Moderate-Low	Slightly higher proportion of females in C2, weak influence.
Age	C1 (Blue, Lower-Risk)	Low	C1 tends to have slightly older patients on average.
MPV	C1 (Blue, Lower-Risk)	Low	Mean Platelet Volume is slightly higher in C1, not a major factor.
NLCR	C1 (Blue, Lower-Risk)	Low	Neutrophil-to-Lymphocyte Count Ratio is slightly higher in C1.
PLTc	C1 (Blue, Lower-Risk)	Low	Platelet count a bit higher in C1, minimal influence.