User Manual

The web frontend is designed to provide an intuitive exploration of the different datasets, clustering algorithms and distances. In an interactive interface multiple clustering settings can be chosen and a visualisation of the results is directly generated. A cluster-table is used to store previous calculated results, which can be plotted within the evaluation module.
The checkbox "Use precalculated results (with random seed for reproduction)" at the top of the page (see Figure 1 A), which is set by default, allows the use of precalculated clustering results, which have been computed beforehand (with a random seed value) and are stored in the github repository. This was done for reproduction and runtime purposes. The second checkbox "use interactive charts" (see Figure 1 B), also set by default, allows the option for interactive projection plots for pca, t-SNE, and the evaluation module.

The user can choose between four datasets (see Figure 1 C), four distance measures (see Figure 1 E) and four different algorithms (see Figure 1 D) via a drop-down menu. If kmeans, kmedian or kmedoids is chosen a value for the parameter k between 1 and 10 has to be set (see Figure 1 F) The default value for k is 3. If DBSCAN is chosen the user has to define a value for epsilon and the minimal number of nearest points (see Figure 2 A). Additionally a link to a webpage implementing the DBSCAN heuristic helping with estimating the values for epsilon and minPts is shown. A short manual for this page is given in section 0.1. The parameter settings can be adjusted with an interactive slider widget.

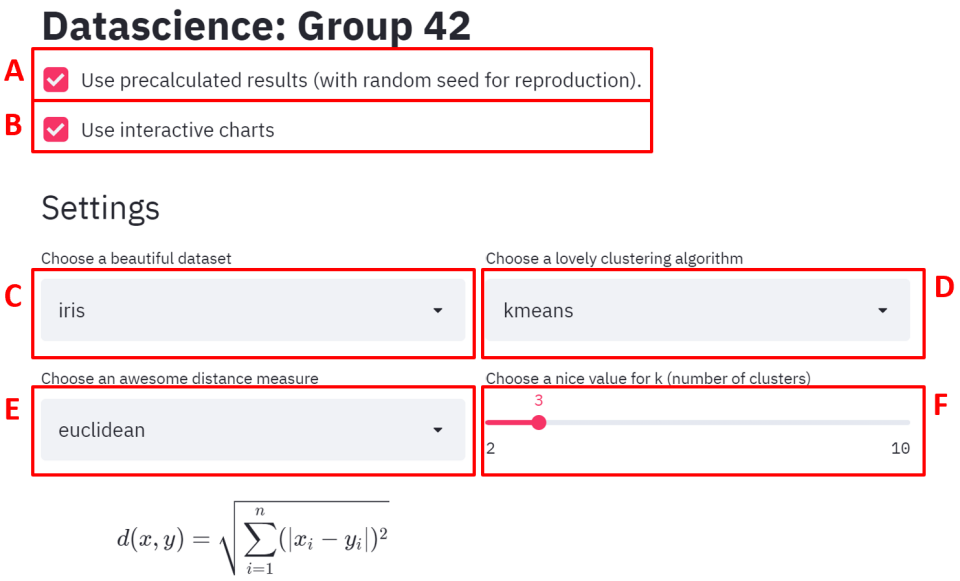


Figure 1: First part of the web frontend. Setting options for clustering parameters for K-Means, K-Medians and K-Medoids.

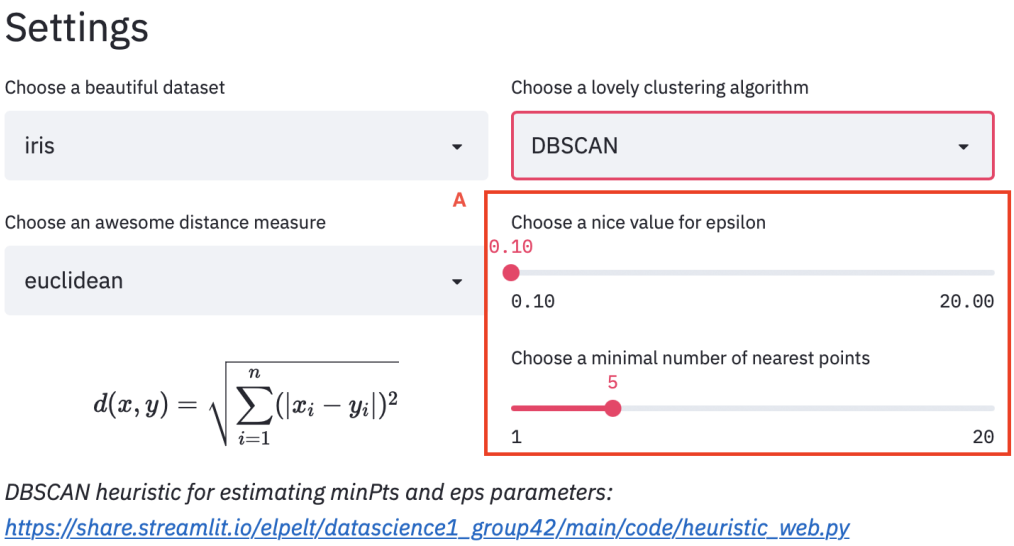
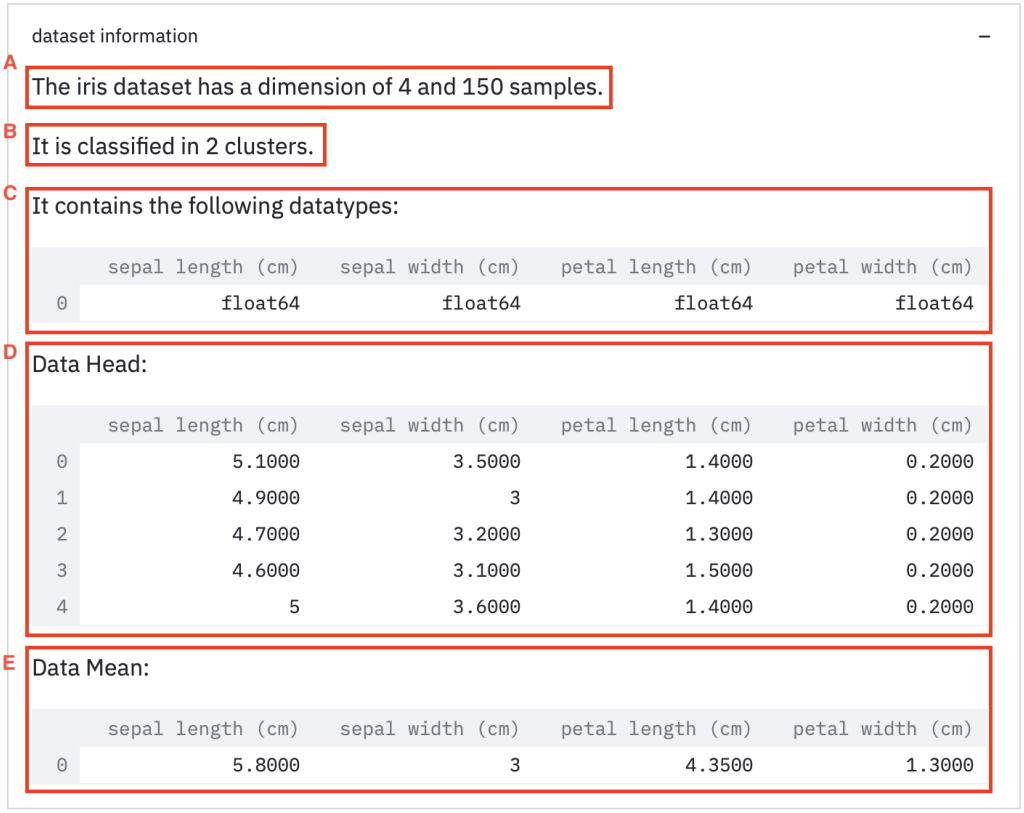


Figure 2: Epsilon and minimal number of nearest points setting options for DBSCAN.

For every dataset some main information can be retrieved via an expander. The dataset dimension (see Figure 3 A), pre classified cluster of the data (not for diabetes dataset) (see Figure 3 B), datatypes per column (see Figure 3 C), dataset preview (see Figure 3 D), mean per column (see Figure 3 E), and performed changes on the dataset are accessible (if performed).



	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)
0	float64	float64	float64	float64

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)
0	5.1000	3.5000	1.4000	0.2000
1	4.9000	3	1.4000	0.2000
2	4.7000	3.2000	1.3000	0.2000
3	4.6000	3.1000	1.5000	0.2000
4	5	3.6000	1.4000	0.2000

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)
0	5.8000	3	4.3500	1.3000

Figure 3: Displayed dataset information by expanding the box by clicking on the plus. Iris dataset as example.

Moreover the perplexity value for the t-SNE projection can be set individually between 5 and 50 with a slider (see Figure 4 A). The default value is 25. The t-SNE and pca plots are shown next to each other to allow a direct comparison of the lower-dimensional projections. For K-Medoids the medoid, for K-Means the mean, and for K-Median the median for each cluster is marked as diamond (see Figure 4 D). For the pca plot both axis are labeled with the percentage of the first two components. To virtually interact with the plots, the button shown in Figure 4 C can be clicked. Please note that this option is only available when the checkmark shown in Figure 1 B is set. Furthermore, this button provides the ability to save the plot in different formats. The calculated runtime is displayed right under the plots (see Figure 4 B). This info is only accessible if precalculated results are not used (see Figure 1 A).

The frontend is reloaded entirely if a parameter value is changed, a different setting is made or a button is clicked.

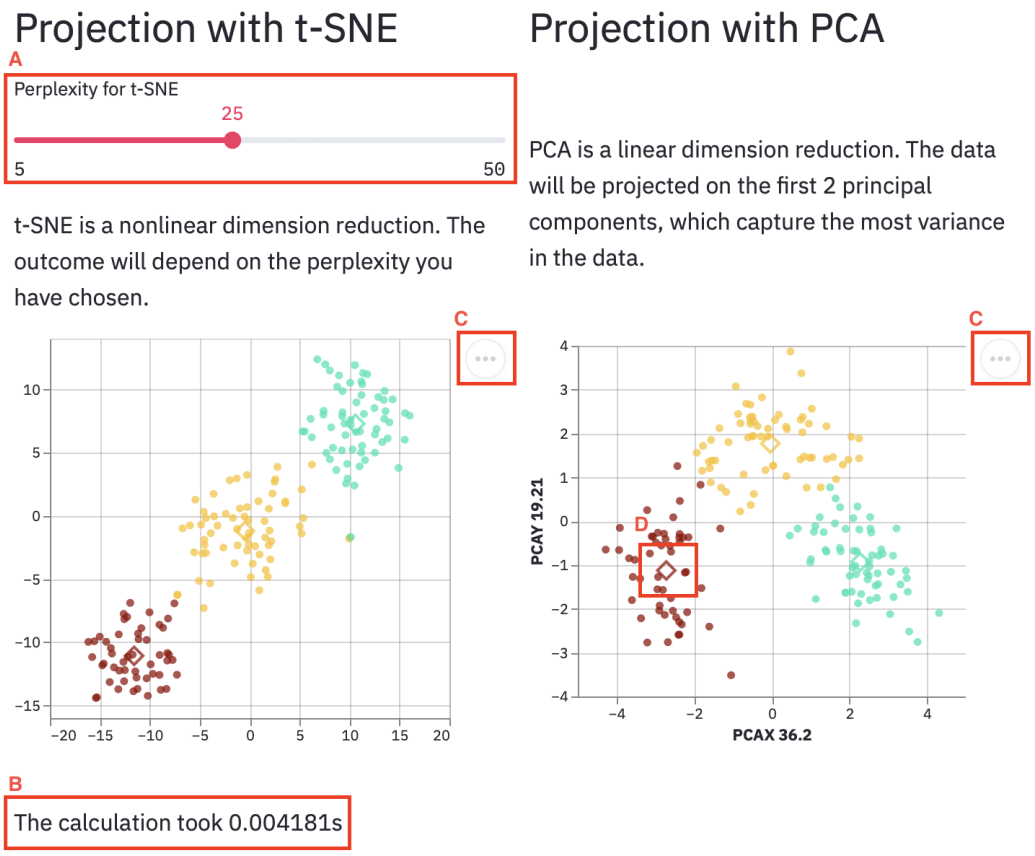
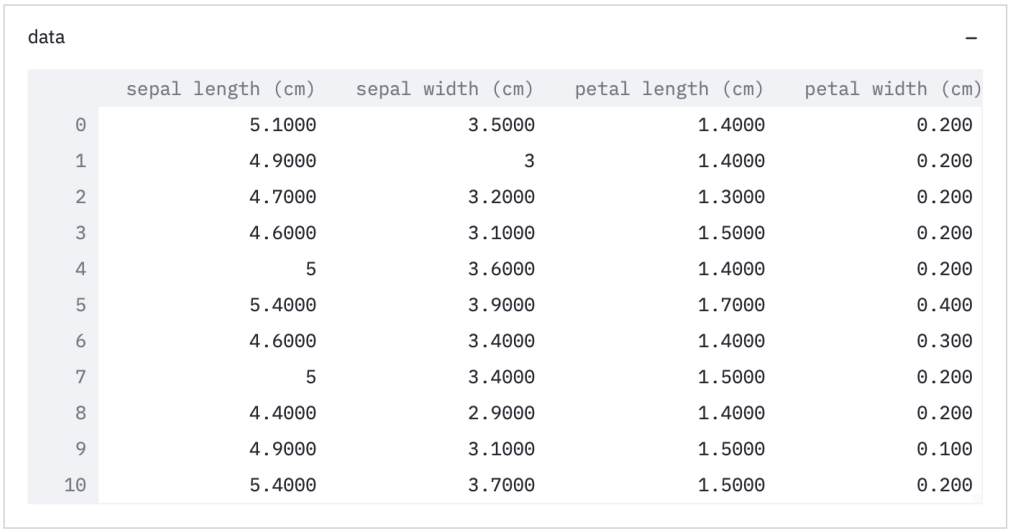


Figure 4: Second part of the web frontend. Projection results of a clustering.

The selected dataset can be viewed in a table format below the plots (see Figure 5).


data

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)
0	5.1000	3.5000	1.4000	0.200
1	4.9000	3	1.4000	0.200
2	4.7000	3.2000	1.3000	0.200
3	4.6000	3.1000	1.5000	0.200
4	5	3.6000	1.4000	0.200
5	5.4000	3.9000	1.7000	0.400
6	4.6000	3.4000	1.4000	0.300
7	5	3.4000	1.5000	0.200
8	4.4000	2.9000	1.4000	0.200
9	4.9000	3.1000	1.5000	0.100
10	5.4000	3.7000	1.5000	0.200

Figure 5: Projection of dataset as expander. Iris dataset as example.

All clustering results can be saved in a cluster table for comparison via the *Add*-button button (see Figure 6 A). To compare this result to clusterings with other settings, the *Add*-button can be clicked repeatedly after desired adjustment of the clustering settings. To start over and compare further clustering indices, the *Reset*-button (see Figure 6 B) can be clicked. The previous display of resulting indices will be cleared as well as the clustering table. For the evaluation of the clustering results, one of five clustering indices can be chosen via a drop-down menu as shown in Figure 6 C. An interactive barplot is displayed immediately, after adding a result to the cluster-table (see Figure 7). A maximum reference value (1) is also always displayed (see Figure 7 A).

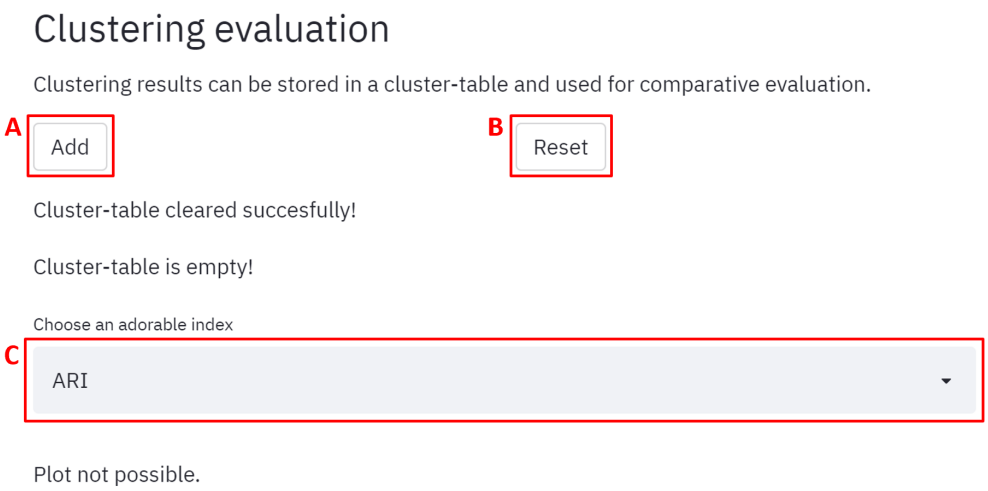


Figure 6: Third part of the web frontend. Evaluation of clustering results.

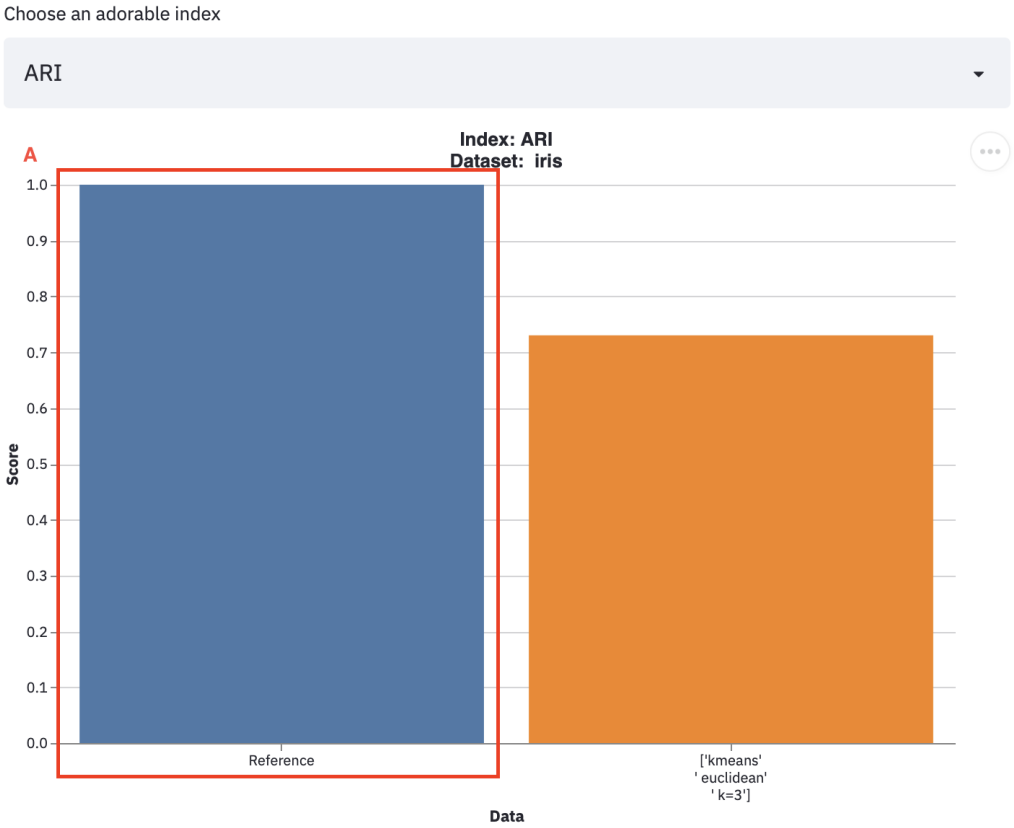


Figure 7: Barplot for evaluation comparison. Kmeans, euclidean, k=3, iris, and ari as example.

Ancillary streamlit settings can be found at the upper right corner. Additional explanatory texts provide help and more information.

0.1 DBSCAN Heuristic Page

The webpage for calculating the sorted k-dist graph for the DBSCAN heuristic is build similarly to the main apps page described above. A simple form is presented were dataset, distance measure and the k value can be chosen using drop-down menus or sliders (see Figure 6 A, B, C). (Note: k in this case is not the number of clusters. k stands for the k-nearest neighbour of any point)

Only when clicking the *Calculate kdist Graph*-button (see Figure 6 D) the sorted k-dist graph is calculated and plotted (see Figure 6 D).

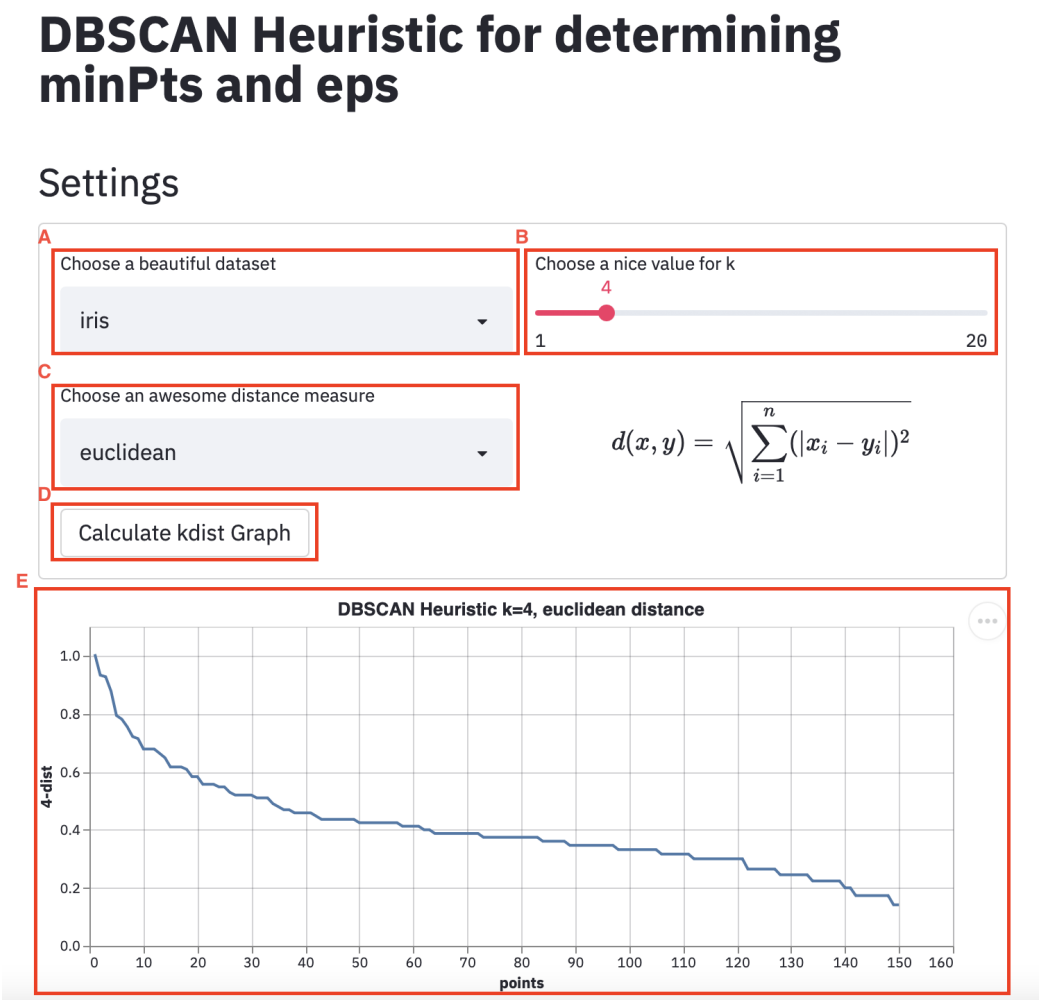


Figure 8: DBSCAN heuristic settings