



Instituto Superior de Engenharia de Lisboa  
Mestrado em Engenharia Informática e de Computadores  
Mestrado em Engenharia Informática e Multimédia  
Big data mining (MDLE)

Laboratory Class #2 — Dimensionality Reduction and Data Representation  
2<sup>nd</sup> semester, 2023/2024 (March, 20)

Code and Report about [the highlighted blue text questions/comments](#) are due by April, 8

---

---

## PART I. MATERIALS AND METHODS

### 1. Datasets

For this laboratory class, we will consider two datasets, as described in Table 1. More details on the first dataset are available at <https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database>

Table 1: Datasets with  $d$  features,  $c$  classes,  $n$  instances and the corresponding problem/task to solve.

Dataset	$d$	$c$	$n$	Problem/Task
Diabetes (Pima)	8	2	768	Detect if a patient shows signs of diabetes
Lisbon-2023-01-01-2023-01-31.csv	?	?	?	?

### 2. Software Tools

In this laboratory class, we will use the following tools to explore the concepts lectured in this module:

- (i) Part II - Orange data mining, available at <https://orangedatamining.com>
- (ii) Part III - R, using R Studio, which together with Spark scales for high-dimensional (big) data.

### 3. Laboratory Class Setup/Preparation

- Install Orange data mining software, available at <https://orangedatamining.com>, and check for its proper functioning.
- Check the Lisbon-2023-01-01-2023-01-31.csv dataset from the Laboratory Project. Identify the features as well as their meaning. Consider the *severrisk* feature as the class label.

---

---

## PART II. ORANGE

---

---

### (a) Orange environment and analysis of existing examples

- (a) Run the Orange application and select the **Examples** (Example Workflows) option to see some examples and demos of the use of the software, as depicted in Figure 1.

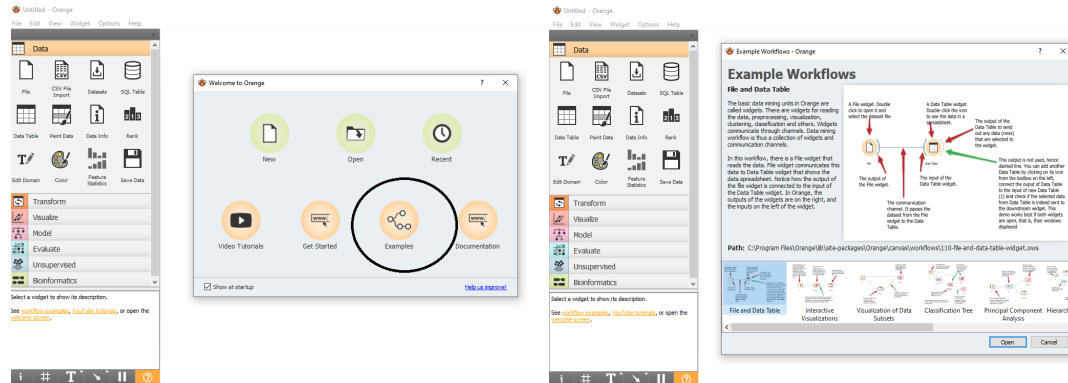


Figure 1: Orange software and its example workflows

- (b) Run the **File and Data Table** example and check its key functionalities, using the (default) Iris dataset. Add a **Feature Statistics** Widget and analyze the four features in the dataset.  
**Identify the list of supported file formats/types for datasets. Show a screen-shot of the statistical analysis.**
- (c) Run the **Interactive Visualizations** example. Add a **Rank** widget and report the Information Gain (IG) and FCBF relevance measures for all the features.  
**Explain the purpose of this example. Identify the information provided by the Data Info widget. What are the most relevant features with the IG and FCBF criteria?**

### (b) Feature ranking and selection

- (a) Run the **Feature Ranking** example (also available on the Web, <https://orangedatamining.com/workflows/Feature-Selection>) with the Iris dataset. On the **Rank** widget, try all the available scoring methods and look for the most relevant feature.  
**What seems to be the most relevant feature?**
- (b) On the **Feature Ranking** example with the Iris dataset, use the **Scatter Plot** widget to identify the most relevant feature.  
**Show some screen-shots that of your analysis to find the most relevant feature and justify your answer.**

### (c) Feature reduction with principal component analysis and discretization

- (a) Run the **Principal Component Analysis** example, <https://orangedatamining.com/widget-catalog/unsupervised/PCA>, with the default Brown-Selected dataset.  
**Explain the key actions of this demo and find an adequate number of reduced dimensions.**
- (b) Modify the example to discretize the data with the EFB technique, in the reduced dimensionality space. Save the discretized data into a file.  
**Show the Orange widget that performs these actions as well as the resulting file.**

(a) **Feature Selection**

For both datasets:

- (a) Compute the (unsupervised) relevance of each feature, using variance and mean-median, as the relevance measures.  
**Plot the sorted relevance values in decreasing order. Comment on the resulting plot. Compare on the smallest and the largest relevance value.**
- (b) For the relevance values found in (a), compute an adequate number of features,  $m$ , by the cumulative sorted relevance criterion, with three different thresholds.  
**State the value of the considered thresholds as well as the corresponding values of  $m$ .**
- (c) Repeat (a) and (b) using the Fisher ratio as the relevance measure instead of the variance/mean-median relevance. **Comment on the results.**

(b) **Feature Reduction**

For both datasets:

- (a) Compute the PCA decomposition.  
**Plot the corresponding eigenvalues sorted in decreasing order. What would be an adequate number of reduced dimensions,  $m$ , for this dataset?**
- (b) Compute the SVD decomposition.  
**Plot the corresponding singular values sorted in decreasing order. What would be an adequate number of reduced dimensions,  $m$ , for this dataset?**
- (c) Using the decomposition results of (a) and (b), compute the dimensionality-reduced versions of both datasets.  
**Explain how you perform the dimensionality reduction. State the number of features of the reduced datasets.**

(c) **Feature Discretization**

For one dataset of your choice, compute a discretized version with one unsupervised technique and with one supervised technique, at your choice.

**State the chosen discretization technique as well as the number of discretization intervals for each feature. Comment on the results.**