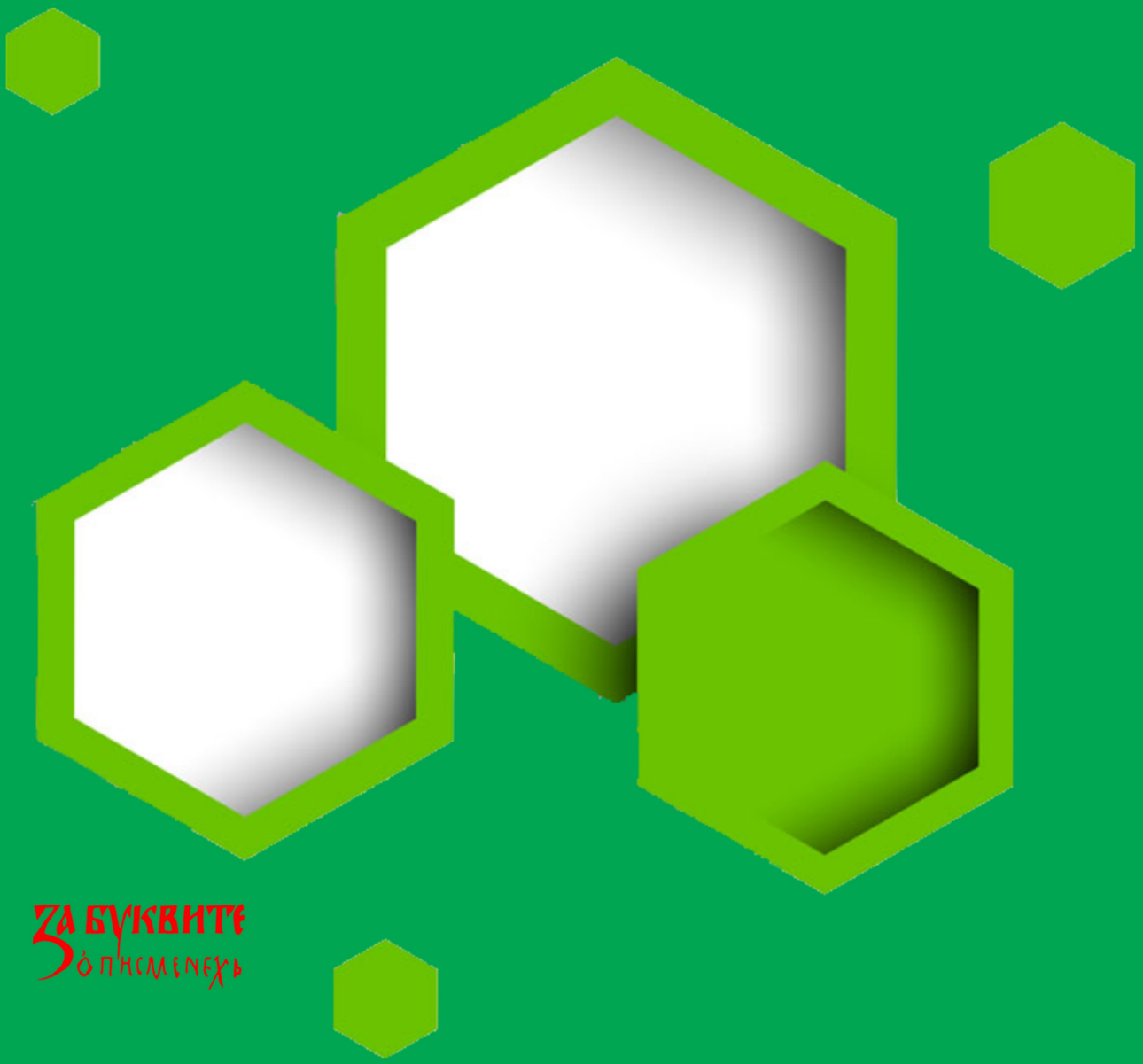
Lubomir Gotsev
Iva Kostadinova

Big Data Analitic Tools



За буквите
о писмености

Lubomir Gotsev Iva Kostadinova

Big Data Analitic Tools

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Big Data Analitic Tools

Study Guide

Academic Publisher "Za bukvite – O pismeneh"
Sofia, 2022

Co-funded by the Erasmus+ Programme of the European Union

This study is published as a result of research carried out within the project "Innovations for Big Data in a Real World" – iBigWorld – 2020 – 2022, financed by Erasmus+ "KA203 – Strategic Partnerships for higher education".

ISBN 978-619-185-572-8

Unlocking new insights from complex, vast, diverse, and massive quantities of data to accelerate digital transformation, innovation, and social sustainability is the primary objective of Big Data Analytics. It uses a palette of advanced techniques and approaches from emerging fields such as Data Mining, Machine Learning and Deep Learning for extracting valuable information. Various tools support and optimize knowledge discovery engineering.

In such a context, the short-term training aims to broaden the students' interest in data-driven projects as a novel paradigm in Industry 4.0 and Society 5.0, by answering a series of questions:

- What acts in the Big Data value-chain paradigm?
- What are the key technologies behind Big Data Analytics (BDA)?
- What tools and software enable meaningful insight from big data?
- What tools and software can be used for BDA with no or minimal coding skills?

Besides these questions, the training provides practical guides to particular tools that support and automate end-to-end data analytics, from data collection, cleansing, and transformation through model building, evaluation, and tuning, to visualization and communication of results.

The learning is based on real use-cases from various application domains to provide a comprehensive and clear understanding of which techniques and approaches are useful for particular data-driven problem and when and how to apply them through analytics tools.

The training is competences oriented utilizing learning-by-doing and case-based methods.

The course is offered to computer science students from University of Nis (Serbia), University of Library Studies and Information Technologies (Bulgaria) and University of Bielsko-Biala (Poland).

It is organized into sessions, including labs and workshop. The last is planned in two stages, starting with team building and followed by the development of data-driven projects in different application domains. Trainers with trainees choose the topics for the team projects, taking into account the academic background.

Certification of attendance was provided to all the participants.

Co-funded by the Erasmus+ Programme of the European Union

BIG DATA ANALYTICS TOOLS

STUDENTS TRAINING

PART OF BIG DATA ACADEMIC CLASS & WORKSHOP, 16-20 MAY 2022, SERBIA

PROJECT: INNOVAIONS FOR BIG DATA IN A REAL WORLD 2020-1-PL01-KA203-082197

PARTNERS

- University of Bielsko-Biala (UBB), Poland
- University of Library Studies and Information Technologies (ULSIT), Bulgaria
- University of Nis (UNi), Serbia
- Taras Shevchenko University of Kyiv (TSNUK), Ukraine

PARTICIPANTS

Students

- Four students from UBB, Poland
- Four students from ULSIT, Bulgaria
- Four students from UNi, Serbia

Trainers

- Four trainers from UBB, Poland
- Four trainers from ULSIT, Bulgaria
- Four trainers from Uni, Serbia
- Three trainers from TSUNK, Ukraine (online)

COUNTRY

Serbia

EVENT HOST

University of Nis

DATA

16 – 20 May 2022

ACADEMIC CLASSES & WORKSHOP SCOPE

Big Data Knowledge and Skills Development

ANALYTICS TOOLS WORKSHOP SCOPE

Skills-building in Data Mining and Machine Learning for Big Data Analytics

FORMAT

Co-funded by the Erasmus+ Programme of the European Union

BIG DATA ANALITICS TOOLS

TRAINERS

GEORGI DIMITROV

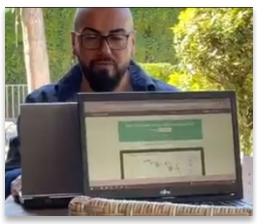
FULL PROF, PhD (ULSIT)

VASIL TOTEV

ASSOC PROF, PhD (ULSIT)

IVA KOSTADINOVA

CHIEF ASSIST, PhD (ULSIT)



LYUBO GOTSEV

ASSIST, PhD Candidate (ULSIT)

ORGANIZATION

- Theoretical Background: Lectures, Individual Research Tasks, Quizzes
- Practical Sessions: Live and Video Demonstrations, Guidelines, Labs, Individual and Group Tasks
- Assessment: Final Group Project

LEARNING METHODS

- Learning-by-doing
- Real-Case and Project-based Learning

LECTURES

1. **Data Analytics Overview (part one)**
- Introduction
- Multi-Disciplinary Nature
- Actors & Processes
- Categories
- Methodologies
- Applications
- Trends

2. **2. Data Analytics Overview (part two)**
- Data Analytics Terminology
- Exploration Data Analysis (EDA) through Summary Statistics
- Exploration Data Analysis (EDA) through Visualization Techniques
- Data Quality Strategy
- Machine Learning for Big Data Analytics: Approaches, Techniques & Algorithms

3. **Analytics Tools Overview (part one)**
- Big Data & AI Tools Landscape
- Popular BDA Solutions: Cloudera, SAP HANA, SAS Viya, Alteryx, Apache Ecosystem, Azure ML
- Languages: R & Python

Co-funded by the Erasmus+ Programme of the European Union

4. Analytics Tools Overview (part two)

- IBM Watson
- KNIME
- Orange
- Tableau

PRACTICAL SESSIONS (BASED ON REAL USE CASES)

Orange

Lab Sessions

- Installing the software
- Workspace (canvas) and components (widgets) familiarization
- Creating a workflow
- Work with built-in datasets
- Basic Data Exploration with Orange
- Feature Statistics
- Data Preparation
- Classification
- Regression
- Cluster Analysis

Tableau

Lab Sessions

- Installing the software – Tableau Public
- Data workspace and loading data
- Using limited preprocessing functionality
- Familiarization with visualization and analysis workspaces
- First visual analysis
- Exploring different visualization techniques
- Forecasting
- Clustering
- Dashboards
- Story

Workshop

Morning Session: Team building & final projects objective and tasks

Afternoon Session: Work in teams and preparing the project in informal envoironment

Final Projects

Morning Session: Teams Projects Presentations

Resources:

- Especially designed videos or live demonstrations illustrating all tasks for the Lab Sessions
- Primary steps in training (for learners)
- Lab Session Notes for trainers
- Workflows of tasks
- Prebuild datasets and Links to data

PURPOSE

The training aims to achieve two primary goals in the learning path of Big Data.

Deepening the interdisciplinarity in the Big Data domain where Data Mining, Machine Learning, Data Science, and Advanced Analytics play a role as an approach palette to knowledge discovery. The lectures provide an overview of the Knowledge Discovery Paradigm based on Big Data, interdisciplinary links between fields, actors, and processes involved in Analytics, and potential applications, impact, and importance for business digital transformation, Industry 4.0, and Society 5.0.

Co-funded by the Erasmus+ Programme of the European Union

Accelerating skill-building in Big Data Analytics by applying supervised and unsupervised approaches for regression, classification, clustering, and feature engineering through particular software tools (Orange, Tableau) following the learning-by-doing and project-based methods.

The training is competences oriented.

Competences

- Ability to select an efficient algorithm(s) for Big Data problem, which takes into consideration the scale.
- Ability to model, analyze, and evaluate an organization's business processes.
- Capability to choose the best sampling and filtering method(s) for a given big data analysis case.
- Effectively use a variety of data analytics techniques (Machine Learning, Data Mining, Prescriptive and Predictive Analytics).
- Apply quantitative techniques (statistics, time series analysis, optimization, and prediction).
- Using a wide range of Big Data analytics platforms.

Skills

- Capable of quickly adapting activities to new technologies.
- Able to perform an objective analysis of a data-driven problem and take appropriate actions to solve it through analytics tools.
- Compare analytics tools and specify differences between them by purpose, features, application domain, limitations and training.
- Identify, compare, and apply open-source and automated machine learning data analytics tool(s).
- Select and apply the most appropriate analytics tool(s) for a specific data-driven problem.
- Critically assess the data source, usefulness, and potential problems associated with the data.
- Upload, edit, save, and export data using analytics tools.
- Assure data quality through analytics tools.
- Apply and fit ML techniques to the analytical problem using the appropriate tool (s).
- Apply adequate model evaluation metrics and accurately interpret the analytics output.
- Use analytics tools for data visualization to present concepts/ideas/phenomena from a new perspective to decision-makers.

Disposition

Accurate in the selection and use of data analytics tools considering domain, defined data-driven problem and proposed solution design.

TOOLS

The following analytics tools are selected in order to accomplish the stated objectives while adhering to the logic of the learning path where the theoretical underpinnings of data mining have been covered.

ORANGE

Orange is a **component-based visual programming open-source** tool utilized for **data mining, machine learning, data analysis and visualization**.

Co-funded by the Erasmus+ Programme of the European Union

Components of Orange range from basic operations such as **data visualization, subset selection, and pre-processing** to more complex tasks such as the **evaluation of learning algorithms** in practice and the development of **predictive models**. It supports **bioinformatics, text, image, and signal processing add-ons,** as well as advanced analytics features.

Benefits

- Free open-source
- Visual programming
- Both no-coding and coding (Python)

Features

- Python 3 data mining library
- Interactive Data Visualization
- Add-ons Extended Functionality
- Manual parameter optimization

Limitations

- Not always reliable support
- No automatic parameter optimization
- Has error measurement but must rebuild model each time

Training

- Blog, docs & online community support
- Classroom training
- Online tutorial and training videos

TABLEAU

Data visualization platform that can perform **big data analytics**.

It can leverage well-known frameworks such as **Apache Hadoop, Spark and NoSQL databases** to meet their data needs.

The vendor offers the following products: Tableau Online, Tableau Desktop, Tableau Prep, Tableau CRM and **Tableau Public (free).**

Benefits

- Responsive dashboards and reports – all without writing a code
- Leverage a fast, in-memory processing engine
- Combine and analyze large data sets

Features

- One Data Interface
- Big Data Integrations
- VizQL (a visual query language for databases)
- Data Catalog

Limitations

- Organizations are dependent on Tableau to maintain servers
- Challenges when interpreting complex business rules

Training

- a library of free self-service training videos
- on-demand and live webinars
- e-learning and classroom training courses

Co-funded by the Erasmus+ Programme of the European Union

ORANGE TRAINING

LAB SESSIONS

Objectives:

- Installing the software
- Familiarization with workspace (canvas) & components (widgets)
- Creating a workflow
- Work with built-in datasets
- Data Exploration

Objectives:

- Feature Statistics
- Data Preparation
- Prediction
- Cluster Analysis
- Text Analysis

INSTALLATION & FAMILIARIZATION OF ORANGE3

INSTALLATION STEPS

Step 1: Download and install Orange

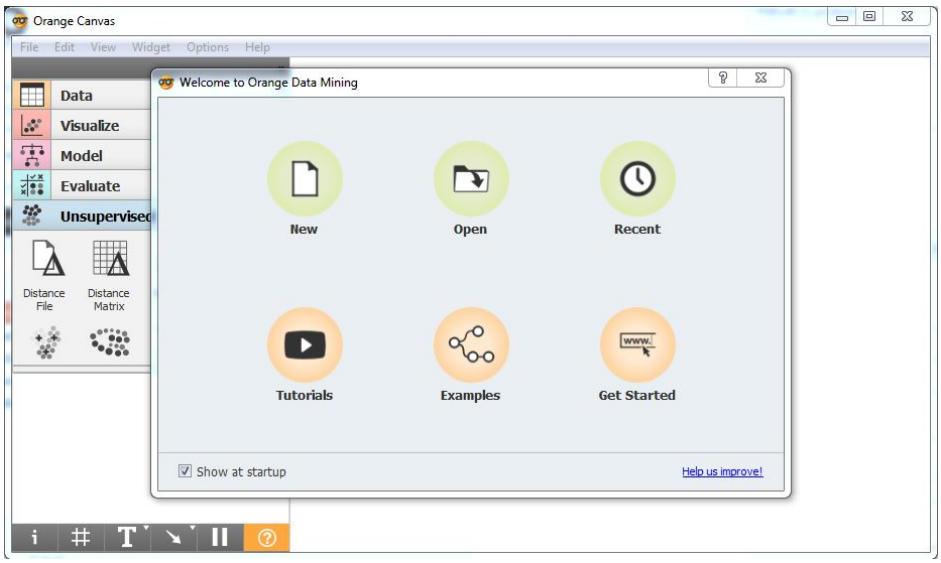
https://orange.biolab.si/download/

Step 2: Run the installer

Step 3: After installation, the Orange icon appears on your desktop; click it to open the Orange tool.

Step 4: You are greeted with an Orange welcome screen.

Co-funded by the Erasmus+ Programme of the European Union

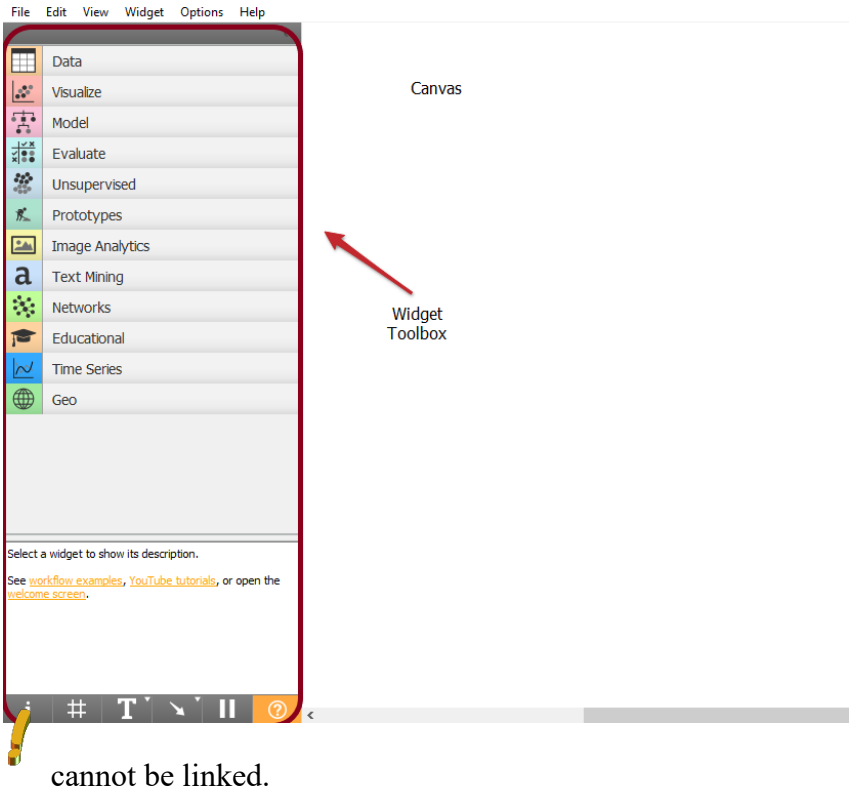


From here, it is possible to start a data analysis workflow, open a recent one, or explore tutorials. The first project can be initiated by choosing the "New" icon. Orange starts with a blank canvas.

WIDGETS AND CANVAS

According to their function, widgets are the computational building units of Orange's visual programming environment. They read, process, and visualize the data; utilize clustering; build predictive models, and otherwise help to explore the data. The widget pane is located on the left side of the screen.

Adding widgets to a workflow can be accomplished in several ways in Orange.



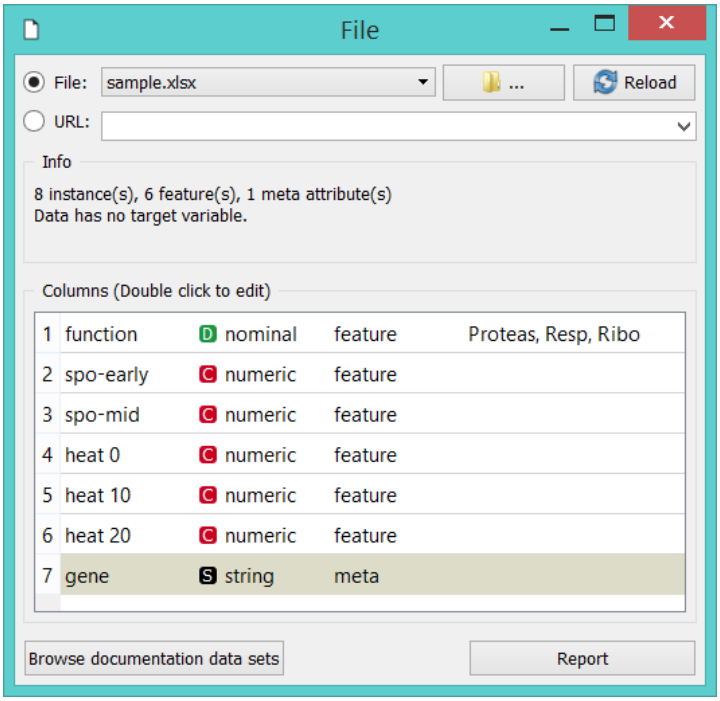
a) The widget will appear in the canvas if you click on it in the widget pane;
b) If you click and drag it to the desired location on the canvas, the widget menu will appear;
c) If you right-click on the canvas, the widget menu will appear. Start typing the widget's name into the filter to find it. Press Enter once you've selected the widget.
d) When you drag a communication channel from the file widget's output, Orange will suggest widgets to which you can connect your original one.

Remember that incompatible widgets cannot be linked.

Co-funded by the Erasmus+ Programme of the European Union

LOADING DATA

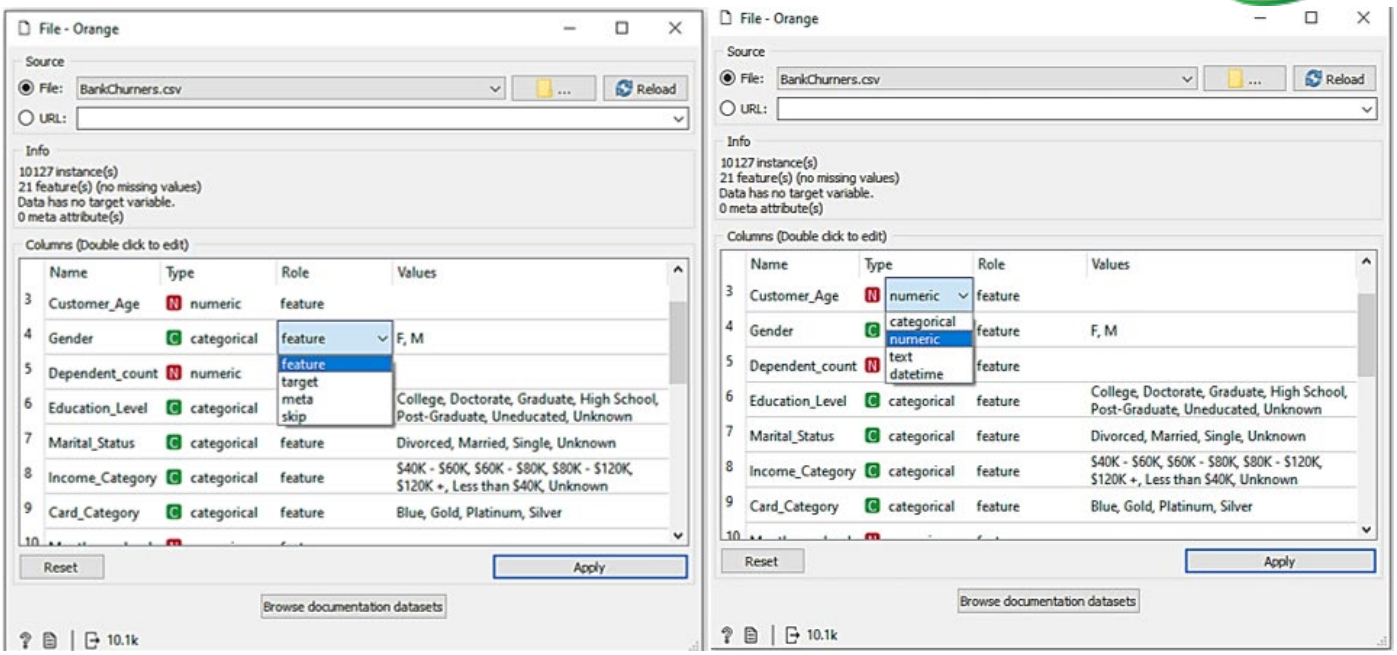
As soon as we open a File widget, we can load our data. It will show up in the canvas when you click on File. Double-click it to open the widget. Orange comes with many data files, and load one of them. You can, naturally, load your own data in simple steps.



So, double click the File widget icon to open it, then click the file browser icon ("...") to locate the downloaded (in the current case, called sample.xlsx) file on your disk.

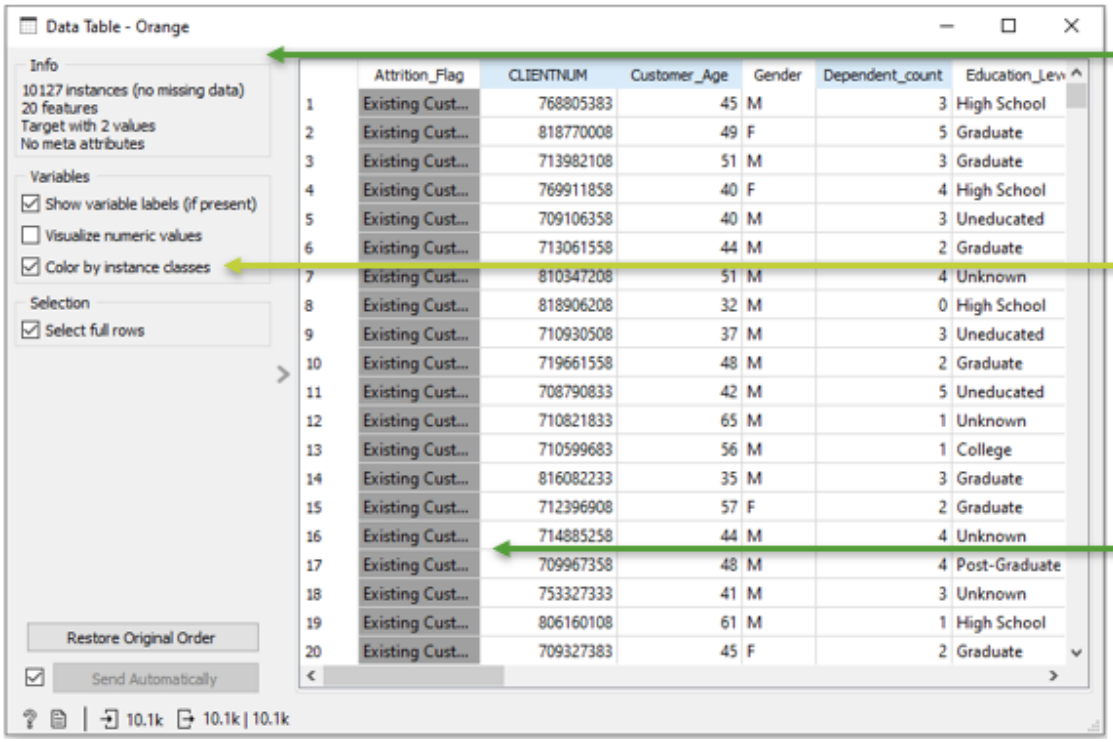
The File widget allows for the configuration of file types and roles. Attributes have roles (input features, meta attributes, and target/class) and can be numeric, categorical, date/time, or textual. Additionally, they can be modified via the File widget.

Co-funded by the Erasmus+ Programme of the European Union



The data contents can be seen through the Data Table widget. For that purpose, we are connecting both widgets.

To see the contents of the Data Table, double-click it:



The data is presented in table format with information on the panel's left side, including the percent of missing values and data types.

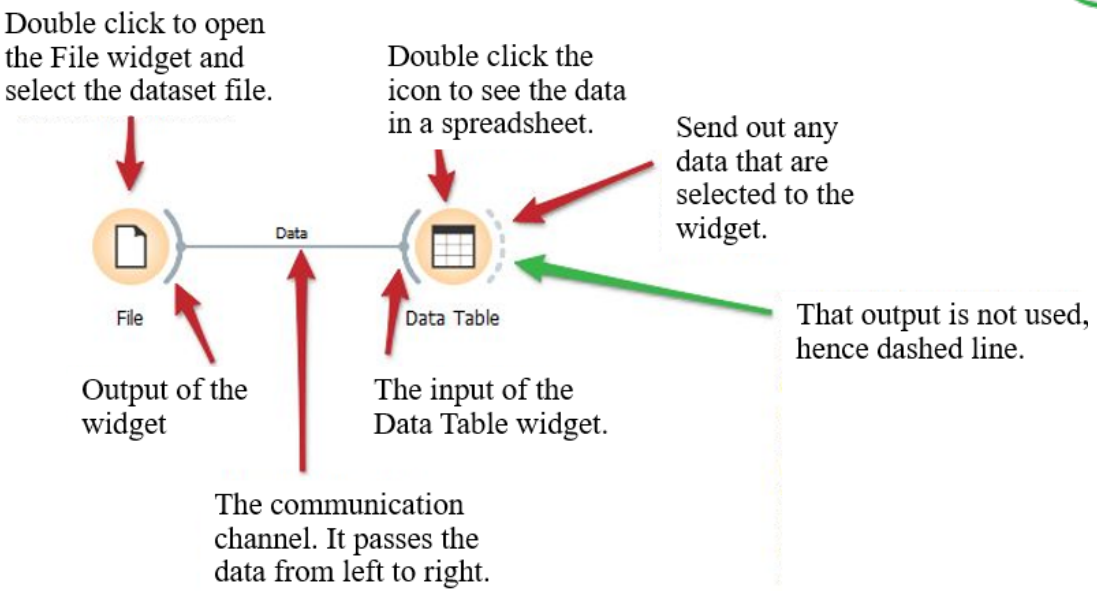
Also, we can color the data by instance classes.

The target is in dark grey, and the meta-attribute is in a lighter tone.

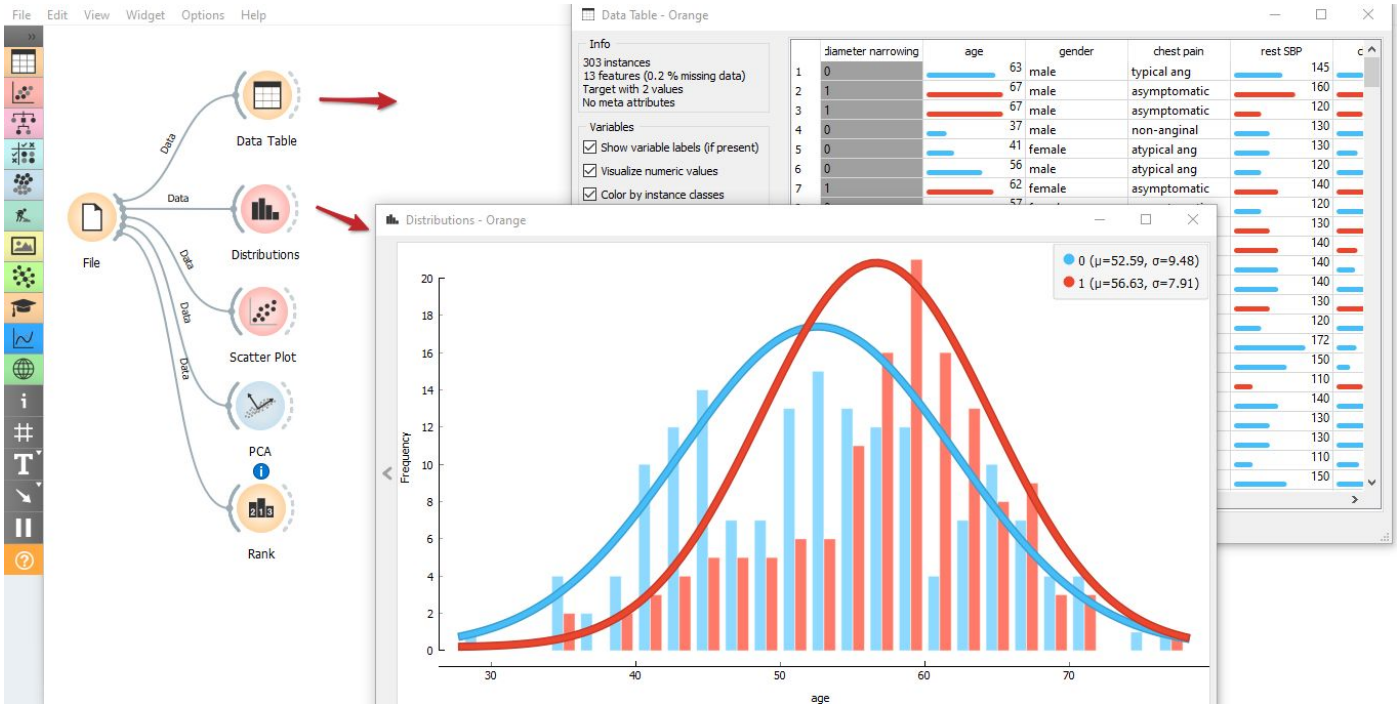
BUILDING WORKFLOWS

Analytical workflows are executed from left to right by placing and connecting widgets on the canvas. In Orange, data does not flow backward.

Co-funded by the Erasmus+ Programme of the European Union



BASIC DATA EXPLORATION

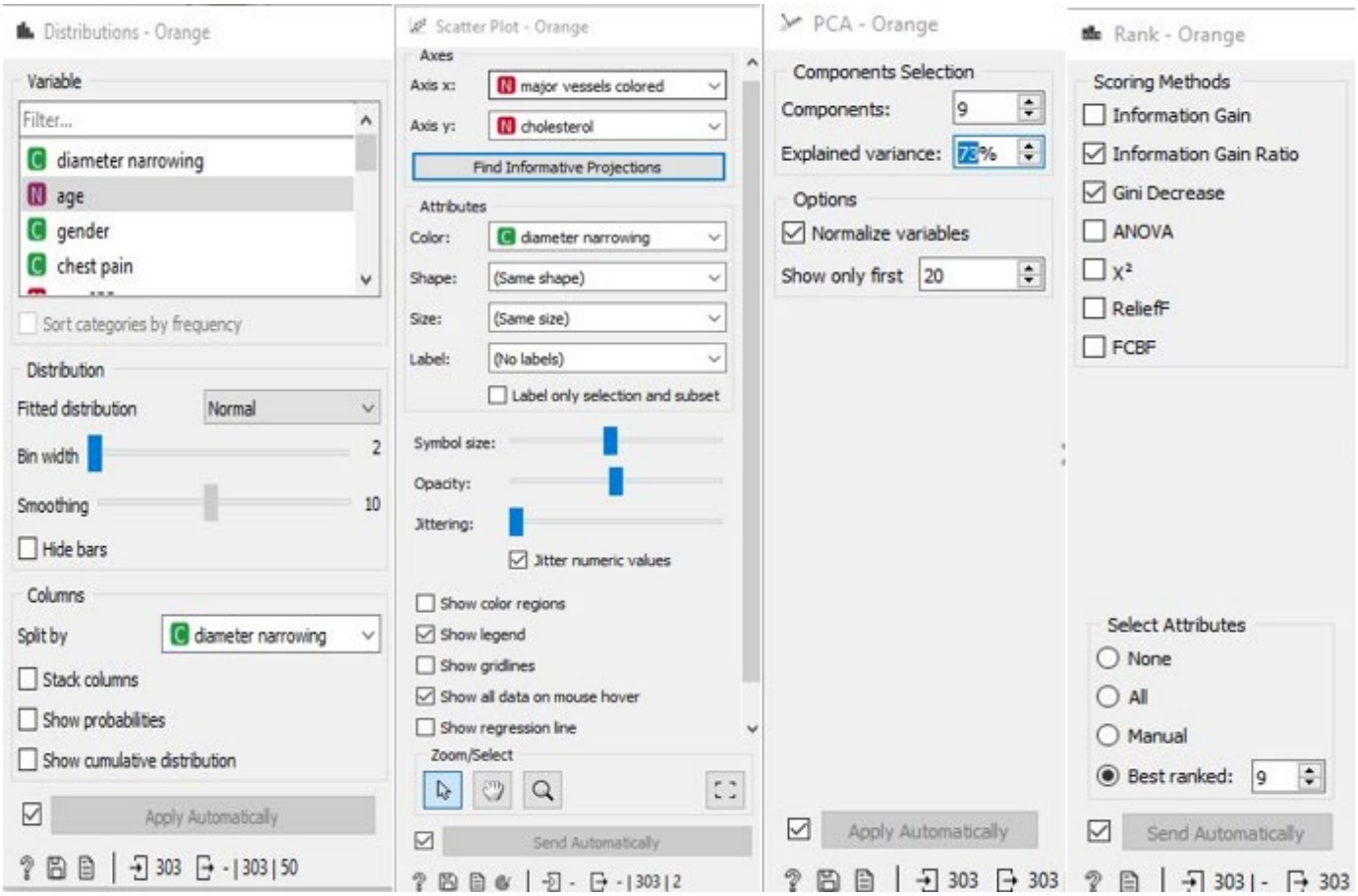


Repeat after the trainer or simultaneously the following steps of basic data exploration over the prebuild "heart_desease" dataset:

- Load the data (widget File)
- Present the data into tabular format (connect to the widget: Data Table)
- Color the values and organize in descending order the first non-target attribute in Data Table
- Explore data related to the target class through widget Data Distribution
- Provide a 2-dimensional scatter plot visualization and find informative projections (widget Scatter Plot)

- Provide a PCA linear transformation of input data (widget PCA)
- Score variables according to their correlation with discrete or numeric target variables (widget Rank)

Here are some of the widget pane settings suitable for the pointed task.

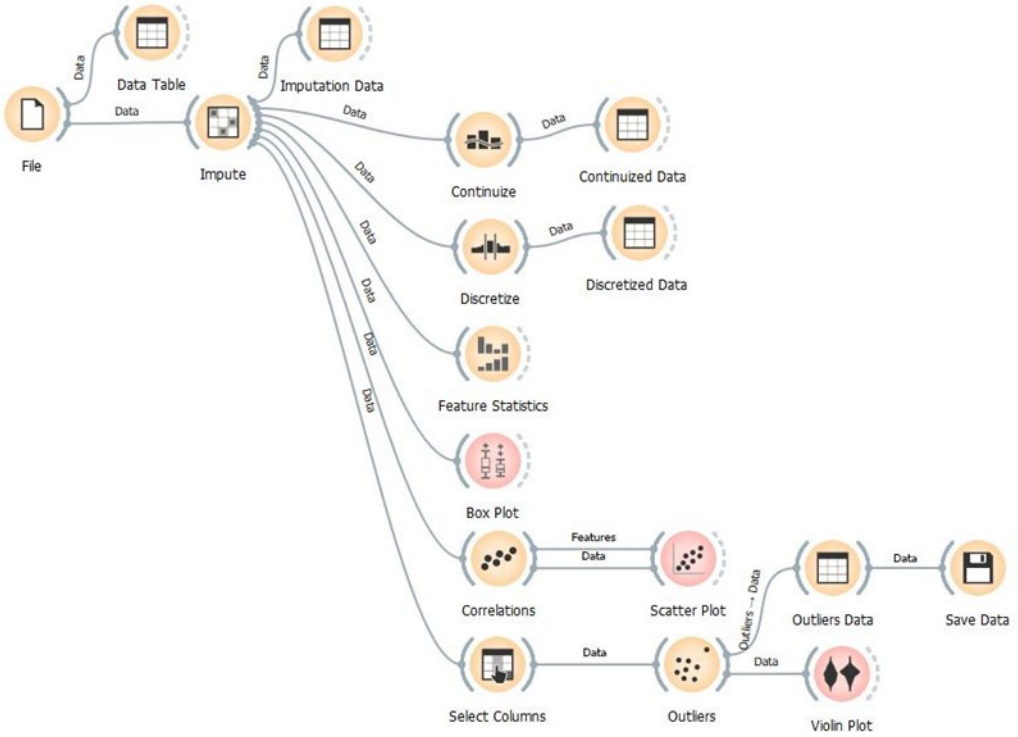


FEATURE STATISTICS AND DATA PREPARATION

Follow the instructions in the **demonstration**. It shows several steps through applying appropriate widgets to prepare the data for analysis and ensure data quality.

For this purpose, a prebuilt **dataset called "heart_desease"** is used. It has a subset of 12 attributes from the Cleveland database. The "goal" field refers to the presence of heart disease in the patient. It is integer–valued from 0 (no presence) to 1 (presence). The associated attributes are age, gender, four types of chest pain (typical Angina, atypical Angina, non-Anginal pain, Asymptomatic), values of measurement: serum cholesterol, resting systolic blood pressure, maximum heart rate. Also, is the fasting blood sugar higher than 120 or not; are the resting electrocardiogram results normal, have left ventricular hypertrophy, or have an ST-T wave abnormality; is thalassemia described as a fixed defect (no blood flow in some parts of the heart), normal blood flow, or reversible defect (blood flow is observed but it is not normal). The slope of the peak exercise ST segment is presented in the data as upsloping, down, and downsloping. Also, the number of major vessels colored is counted from 0 to 3, and the presence or absence of exercise-induced Angina.

Co-funded by the Erasmus+ Programme of the European Union



- Load the data.
- Present the data in a tabular format (Table widget).

The question mark signals a missing value. We guess that such values are less than 0.1% of all the data. Otherwise, the information about that would be given in the File widget. Such a percent is not significant and would not provide a bias. But we have enough available values for all features to deal with the missing data.

- Impute the missing values (Impute widget) and apply the "Average or Most-frequent" method.

In the top-most box, "Default method", the user can specify a general imputation technique for all attributes. It is possible to specify individual treatment for each attribute, which overrides the default treatment set. The imputation methods for individual attributes are the same as the default methods. As almost all features have a few missing data of a different type (numerical and categorical), apply the Average or Most-frequent method. It uses the average value (for continuous attributes) or the most common value (for discrete attributes).

- Check for missing values connecting the imputation output to the Table widget.
- Transform categorical attributes into numeric (Continuize widget)

It receives a data set in the input and outputs the same data set in which the discrete variables (including binary variables) are replaced with continuous ones.

- Apply "Treat as ordinal" and check the result with the Table widget. It converts the variable into a single numeric variable enumerating the original values.
- Normalize the numeric values (to the interval [0,1]) and check the output.

- Discretize the numeric data features (try Entropy-MDL, Equal-frequency, and Equal-width methods).
- Get helpful statistical information for features (Feature Statistics) and make conclusions.
- Visualize the imputed data with the box plot and make conclusions.

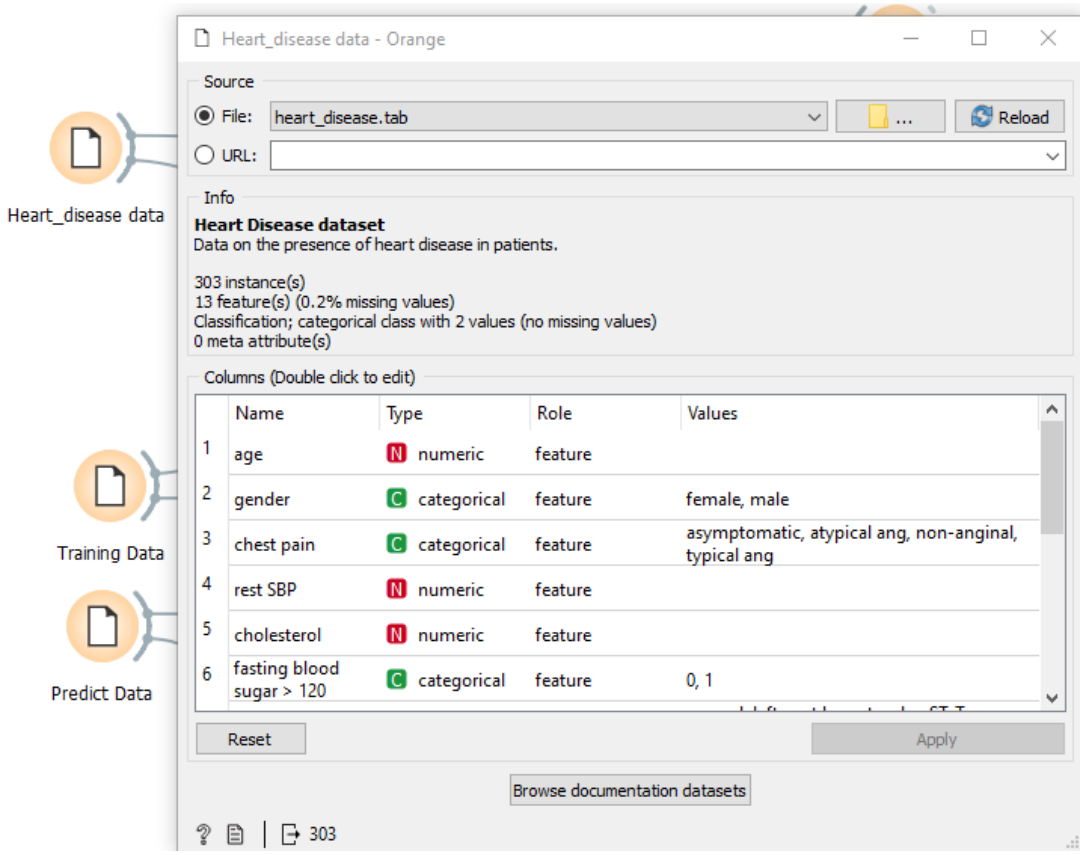
Co-funded by the Erasmus+ Programme of the European Union

- Find the pairwise attribute correlations (Correlations widget). The widget computes Pearson or Spearman correlation scores for all pairs of features in a dataset. These methods can only detect monotonic relationships.
  - Visualize the most correlated pair of attributes with the scatter plot.
- Remove outliers from the age column (Outliers widget), save and visualize the data.

Repeat appropriate steps over the other prebuild dataset to prepare it for analysis.

PREPROCESSING IN ONE WIDGET

1. Load the prebuilt data: heart_disease dataset



2. Connect to the Data Table widget.

The file and table widgets show missing values that we have to deal with. Only 12 attributes describe the data; therefore, additional inspection for dimensionality reduction is not needed so far.

3. Connect to the Distribution and Feature Statistics widgets.

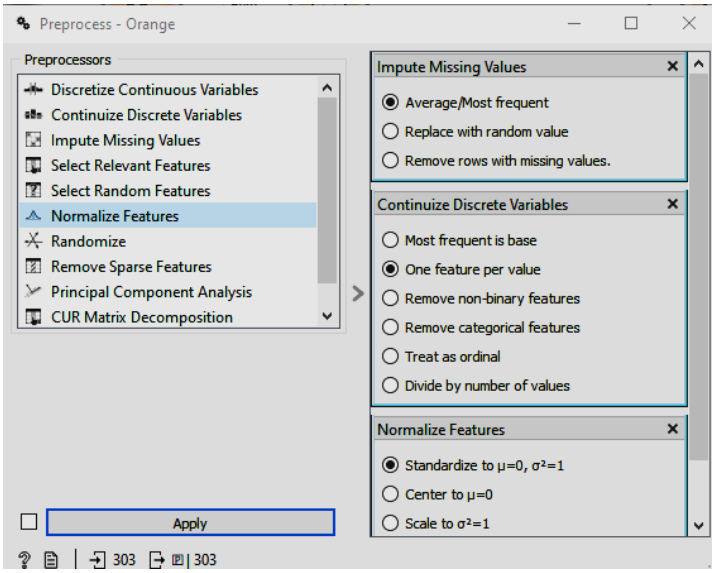
The target binary data are relatively balanced, with 54% instances in class null and 46% in class 1.

Co-funded by the Erasmus+ Programme of the European Union

4. Connect to Preprocess widget

4.1. Apply the “Impute Missing Values” method

4.2. Apply the “Continuize Discrete Variables” method



Try “Treat as ordinal” and “One feature per value” and check the outputs in data tables. What is the difference?

Pay attention to the one-hot encoding technique (Orange equivalent is “One feature per value”). It means each value in the feature gets a new column with value 0 (the instance does not have this feature value) or 1 (the instance has this feature value) for each instance. In those cases categorical features will be labeled with the format feature-name=feature-value = 0/1 – e.g. chest pain=asymptomatic = 1. It means that the feature chest pain has value asymptomatic. Model, in this case, made more columns (attributes) – from 13 to 25.

Co-funded by the Erasmus+ Programme of the European Union

4.3. Apply the "Normalize Features" method

4.4. Apply "PCA", Rank and "CUR Matrix Decomposition" methods) to inspect the Dimensionality reduction approach. Then disconnect methods from the widget window.

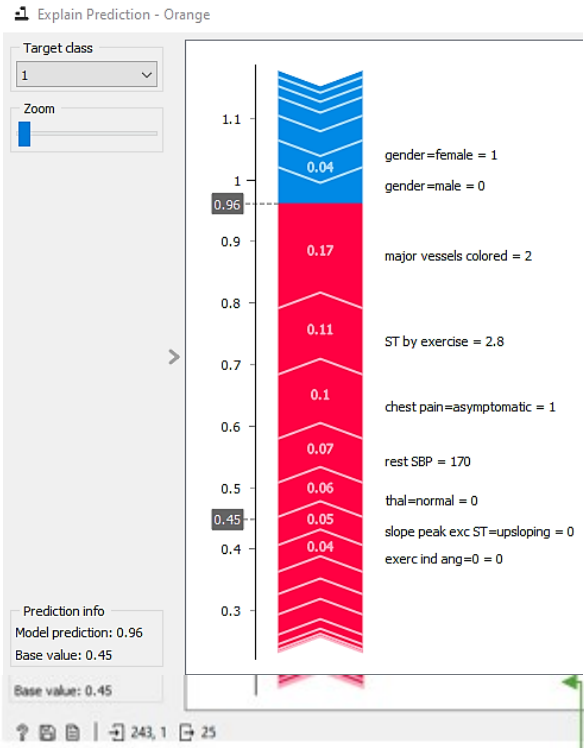
We can manually point the desired number of attributes for the PCA, Rank, or CUR Decomposition. The usage of these methods is illustrative as we needn't such a dimensionality reduction.

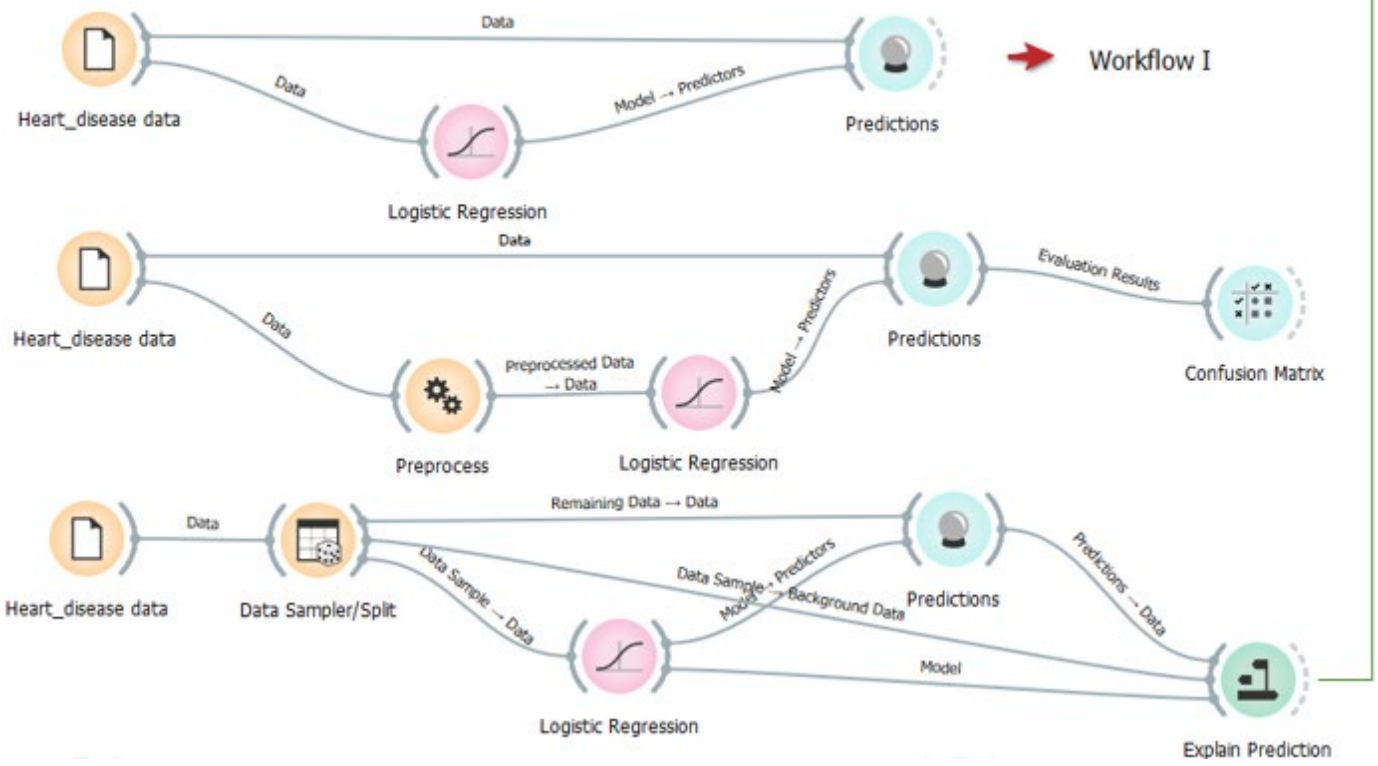
PREDICTION

Knowing how to use the preprocessing pipeline, it's time to know how to apply it to evaluate the models and predict.

The widget "Predictions" shows the probabilities and final decisions of predictive models. The widget's output is another dataset, where predictions are appended as new meta attributes. You can select which features you wish to output (original data, predictions, probabilities). The result can be observed in a Confusion Matrix (Workflow II) and even see Explain Prediction widget (Workflow III) showing what features affect the prediction of selected class the most and how they contribute (towards or against the prediction).

The students tried the following schemas (workflows) with the heart_disease data set.





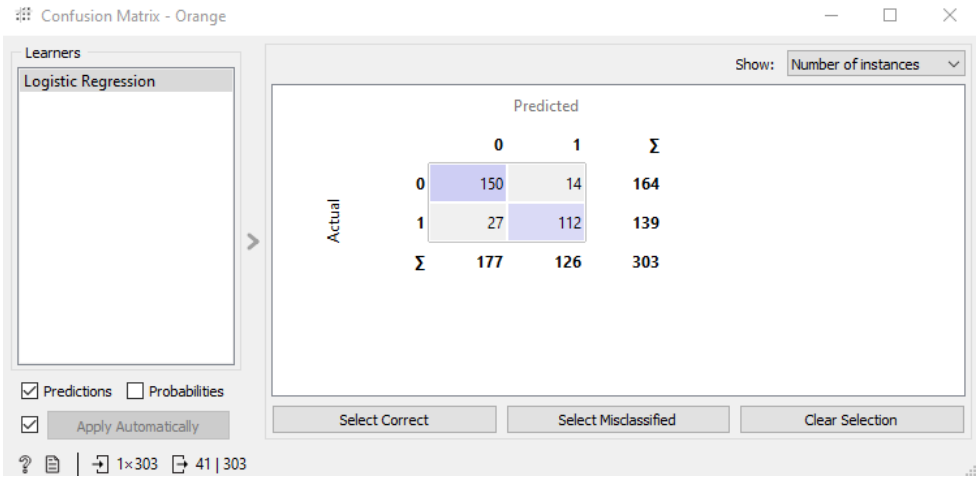
Co-funded by the Erasmus+ Programme of the European Union

The workflow I: using the learner's default preprocessing. Logistic Regression uses default preprocessing when no other preprocessors are given. It executes them in the following order:

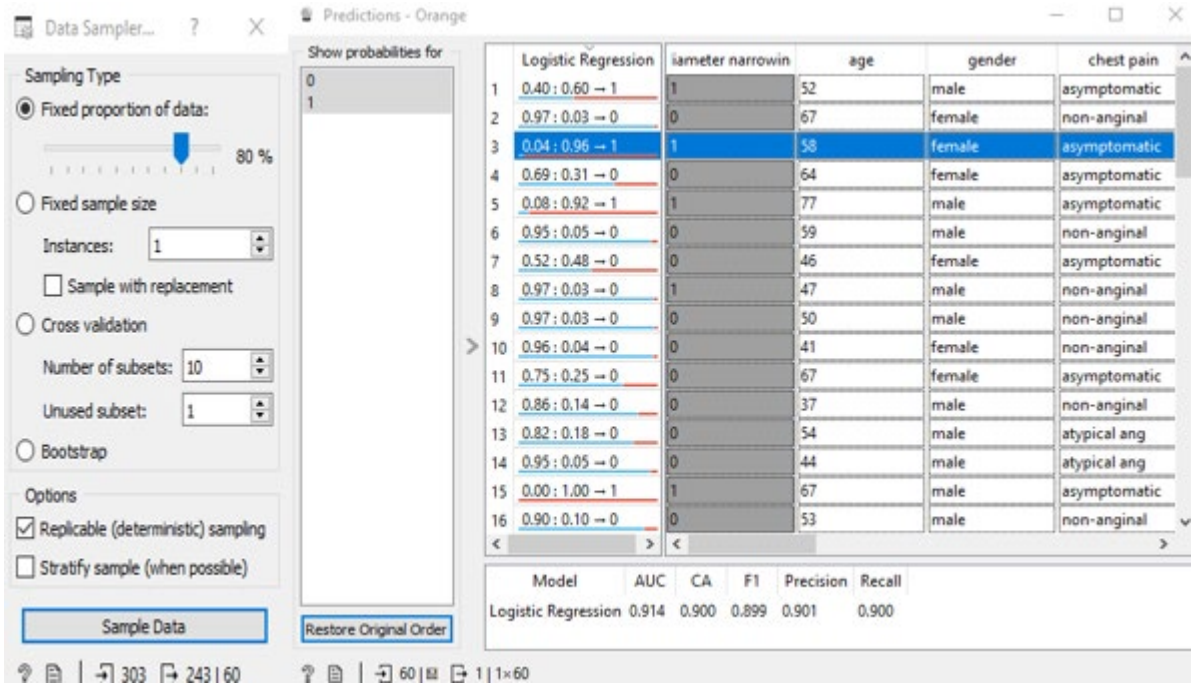
- removes instances with unknown target values
- continuizes categorical variables (with one-hot-encoding)
- removes empty columns
- imputes missing values with mean values

Most learners come with prebuild by default preprocessing. Others like Constant or Decision Tree haven't got. So, you have to check that in the documentation or point the learner in the canvas + F1.

Workflow II: remove default preprocessing by connecting a Preprocess widget to the learner. Confusion Matrix can be used to inspect more detail the outputs.



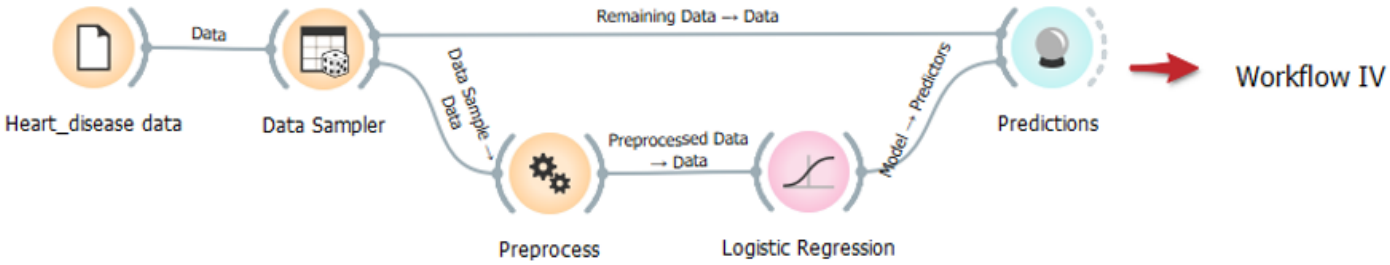
Workflow III demonstrates the Data Sampler widget properties to split the data into training and testing sets. Additionally, Explain Prediction widget shows what features affect the prediction of the selected class the most and



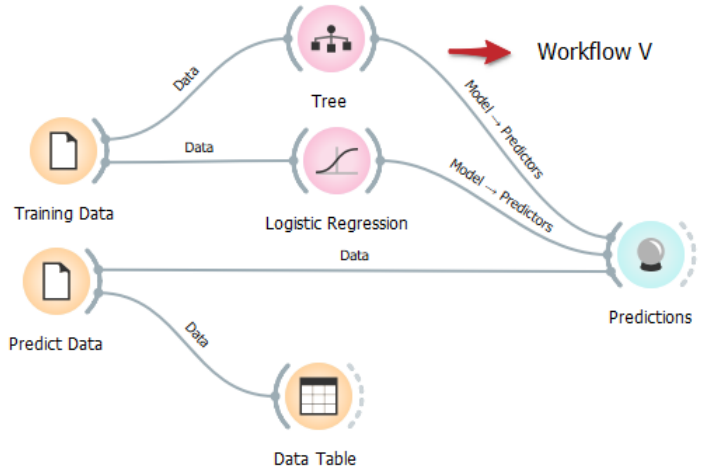
Co-funded by the Erasmus+ Programme of the European Union

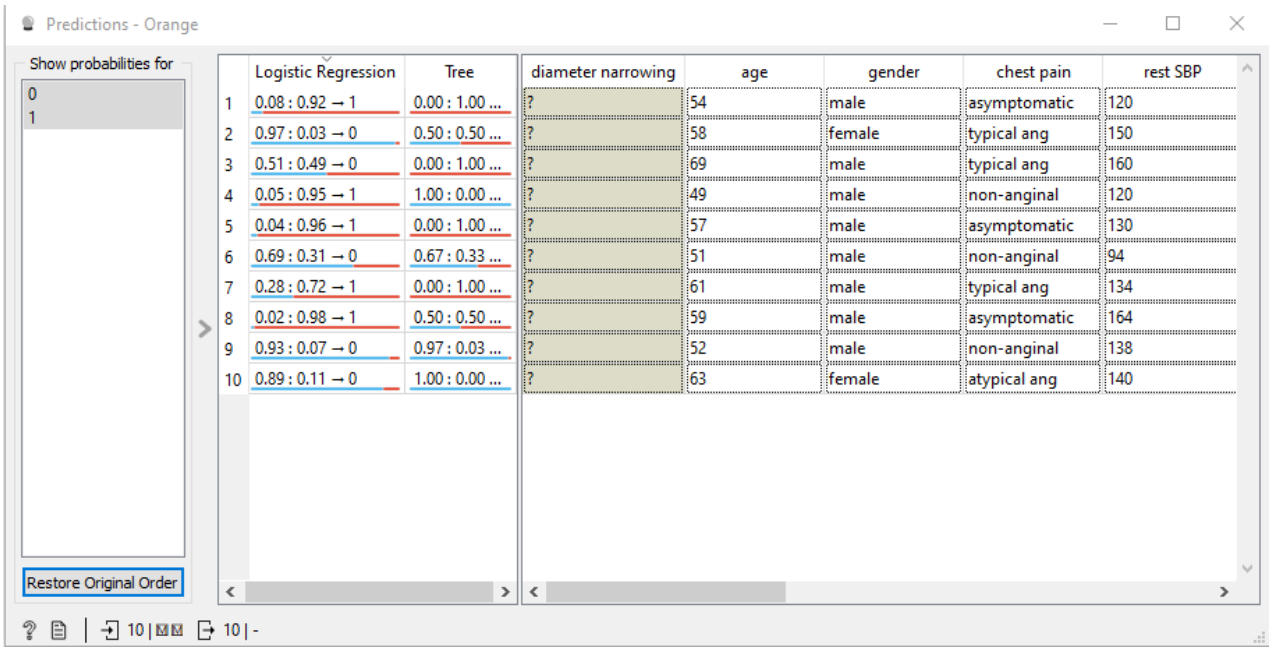
how they contribute (towards or against the prediction).

Workflow IV: combines II (preprocessing) and III (splitting) in one schema.



We can use separate training and predict data applying the following simple workflow.

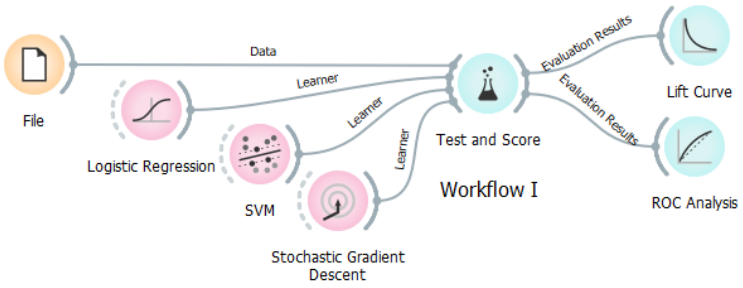




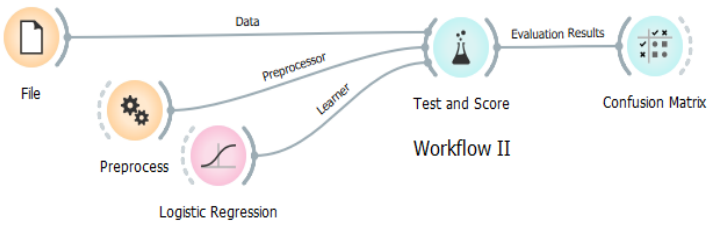
Co-funded by the Erasmus+ Programme of the European Union

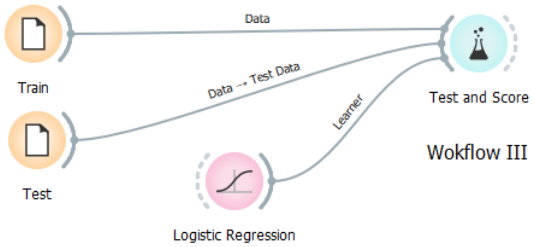
MODEL TESTING AND EVALUATION

The following schema shows the typical test and score workflow.

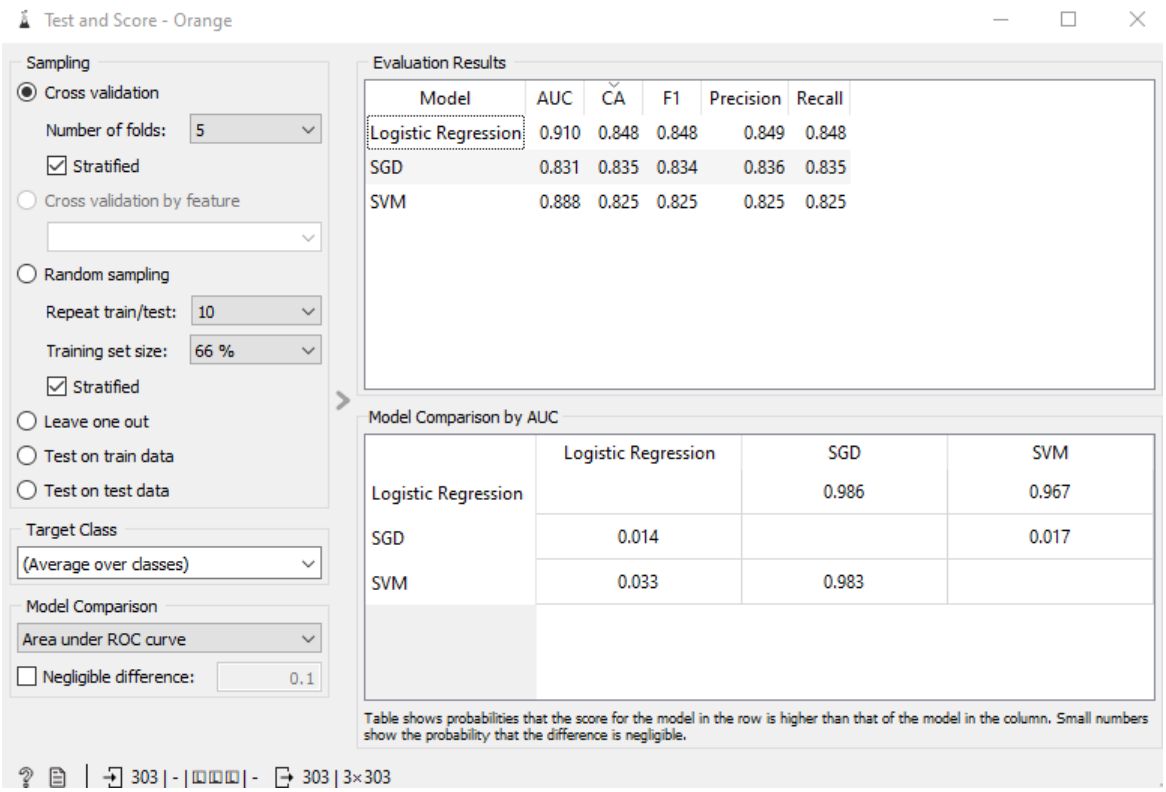


Some extended versions use preprocessing (II) and separate training and testing sets (III).

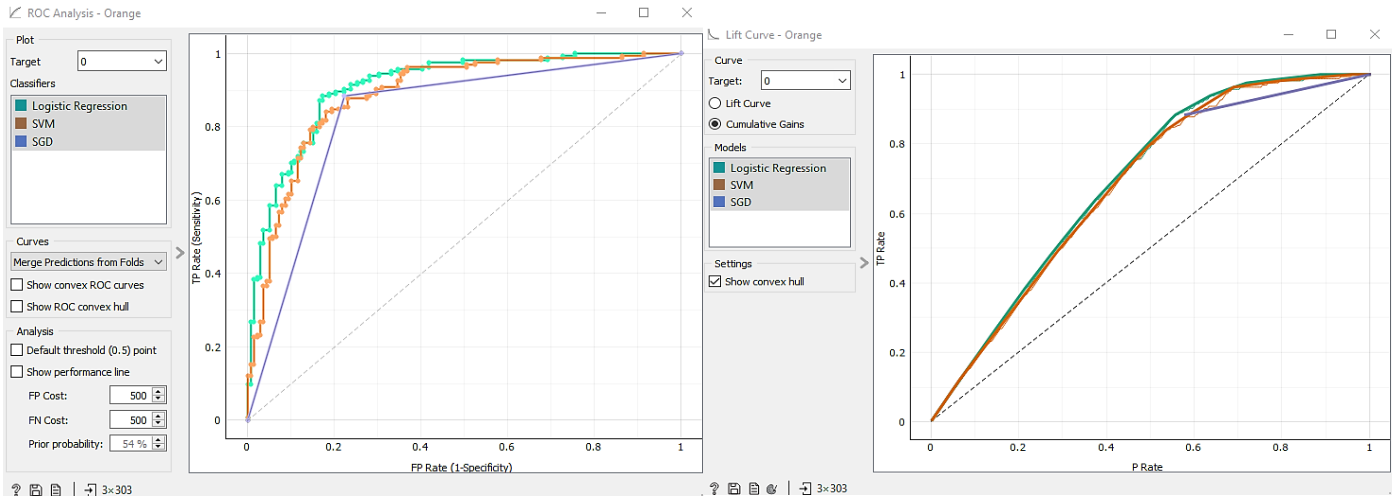




The results can be visualized using various widgets. The confusion matrix and scatter plot, ROC Analysis, and Lift Curve are the most informative.



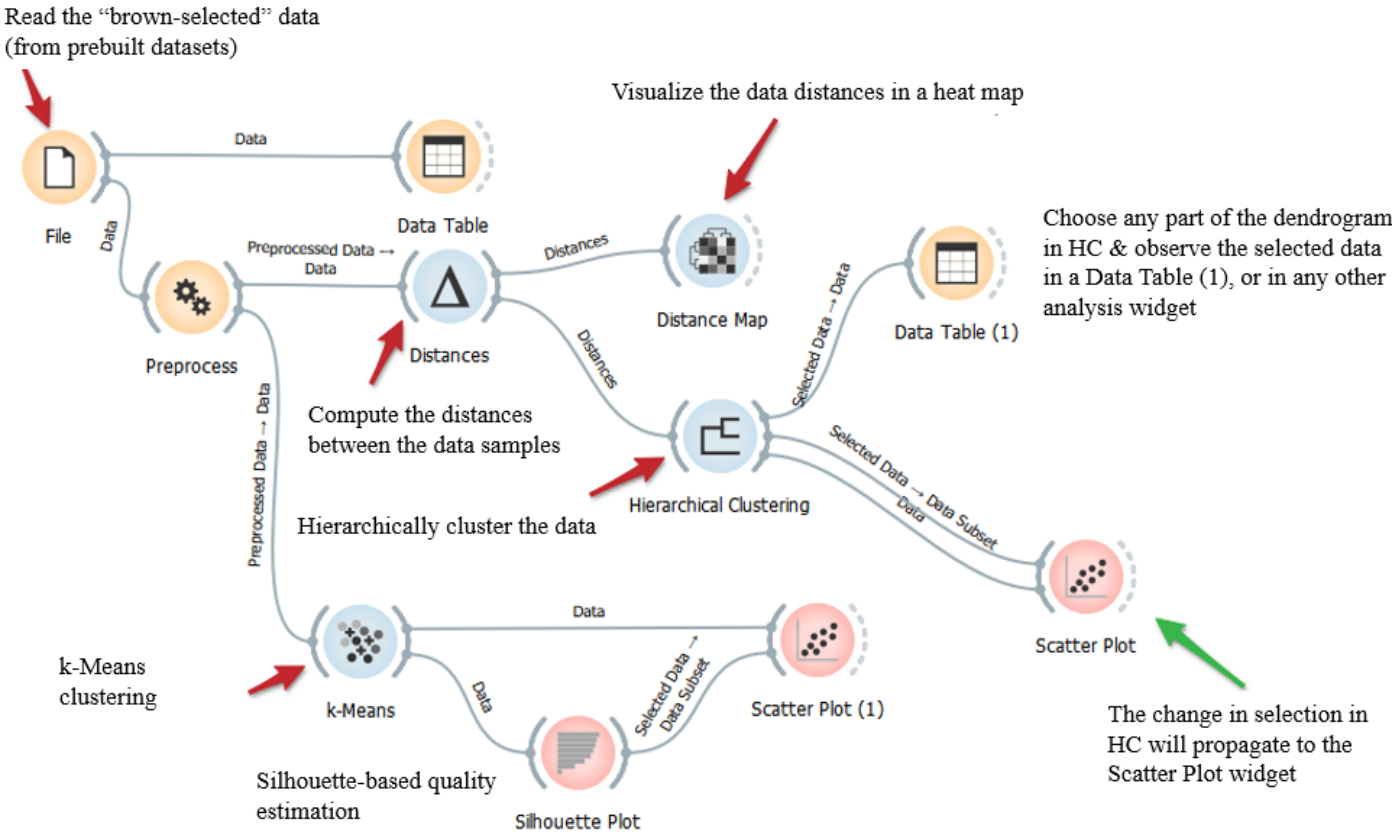
Co-funded by the Erasmus+ Programme of the European Union



CLUSTERING

The brown selected data set comprises 186 rows (genes) and 81 columns. Out of the 81 columns, 79 contain gene expressions of baker's yeast under various conditions, one column (marked as a "meta attribute") provides gene names, and one column contains the "class" value or gene function.

Follow the schema.



Co-funded by the Erasmus+ Programme of the European Union

TABLEAU

Objectives:

- Installing the software – Tableau Public
- Data workspace and loading data
- Using limited preprocessing functionality
- Visualization and analysis workspaces
- First visual analysis
- Exploring different visualization techniques

INSTALLATION & FAMILIARIZATION OF TABLEAU

INSTALLATION STEPS

The free version of the software (Tableau Public) is utilized in lab sessions.

Step 1: Visit the Tableau Public Website

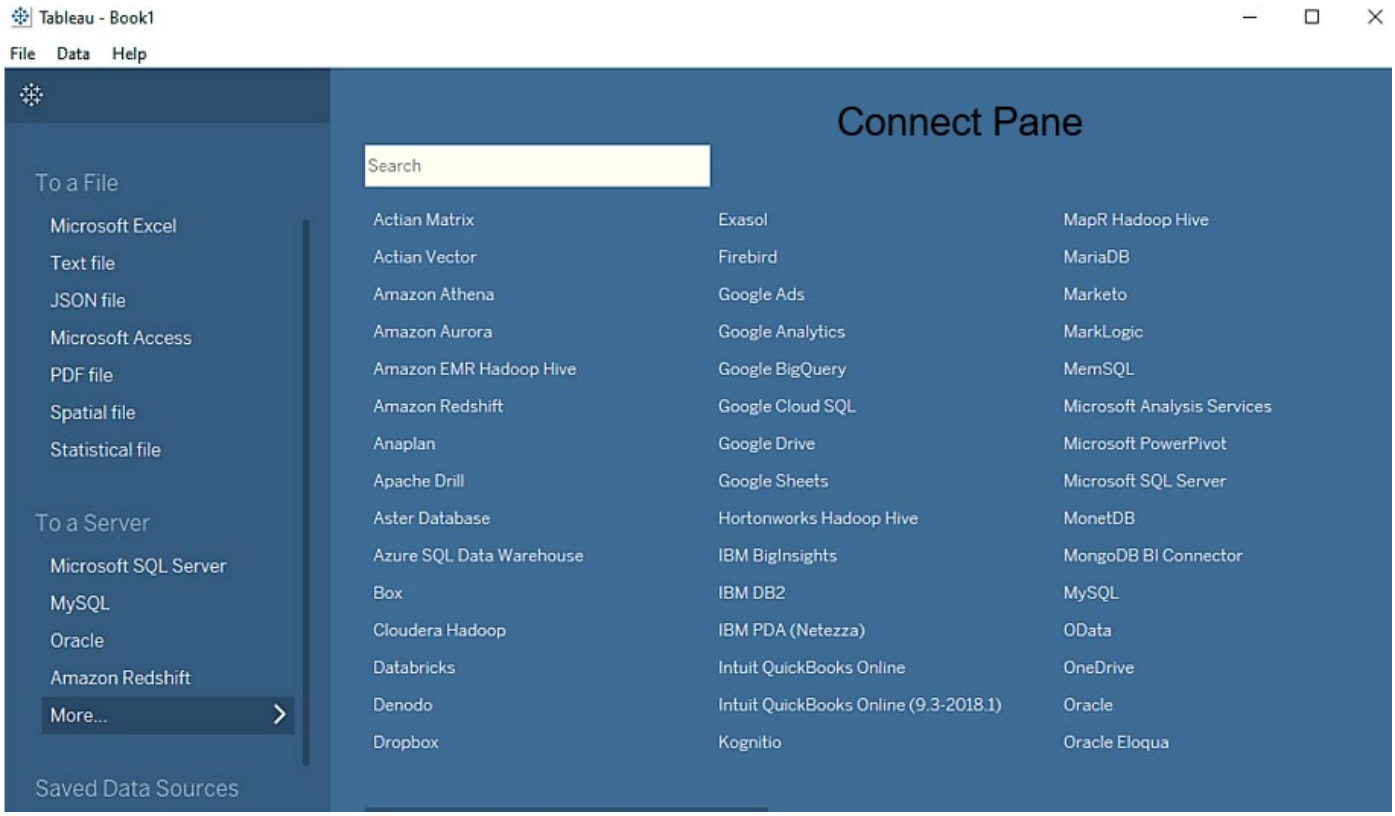
Step 2: Enter your e-mail address to download it.

Step 3: Run the downloaded installer.

LOADING DATA AND PANES

Students followed all the described below steps.

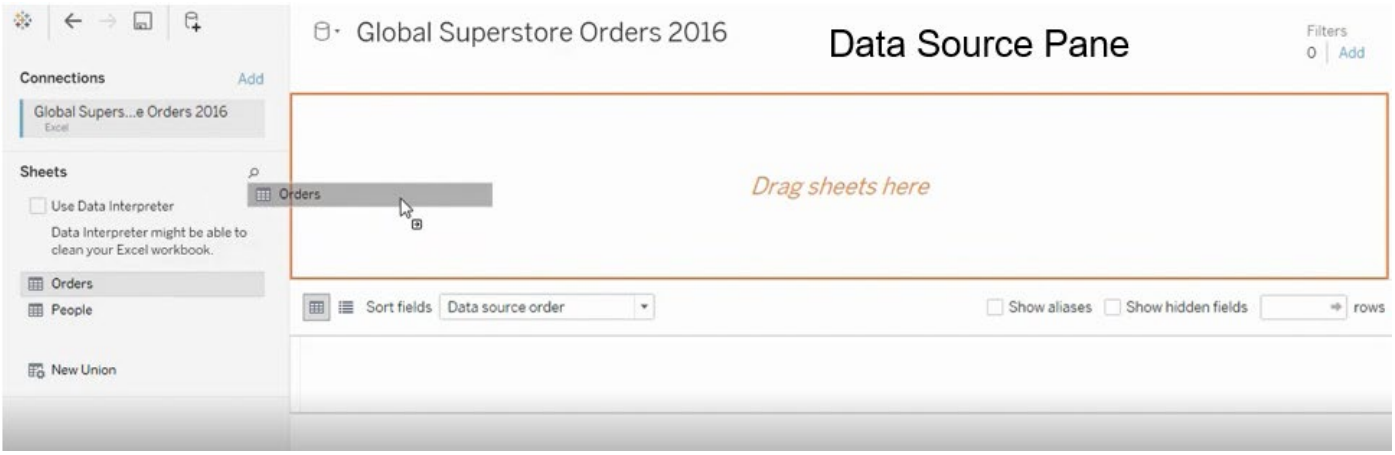
The first screen is called Connect pane. Notably, the exceptional variety of data sources by structure and the ability to connect with various repositories such as IBM, Microsoft, Teradata, Spark is available.



1. Download the Superstore data from the following link. It opens a web storage of various sample data.

Co-funded by the Erasmus+ Programme of the European Union

2. Load the data to Tableau by using the Excell option in the Connect pane.
3. Look at the data in the next pane called Data Source.



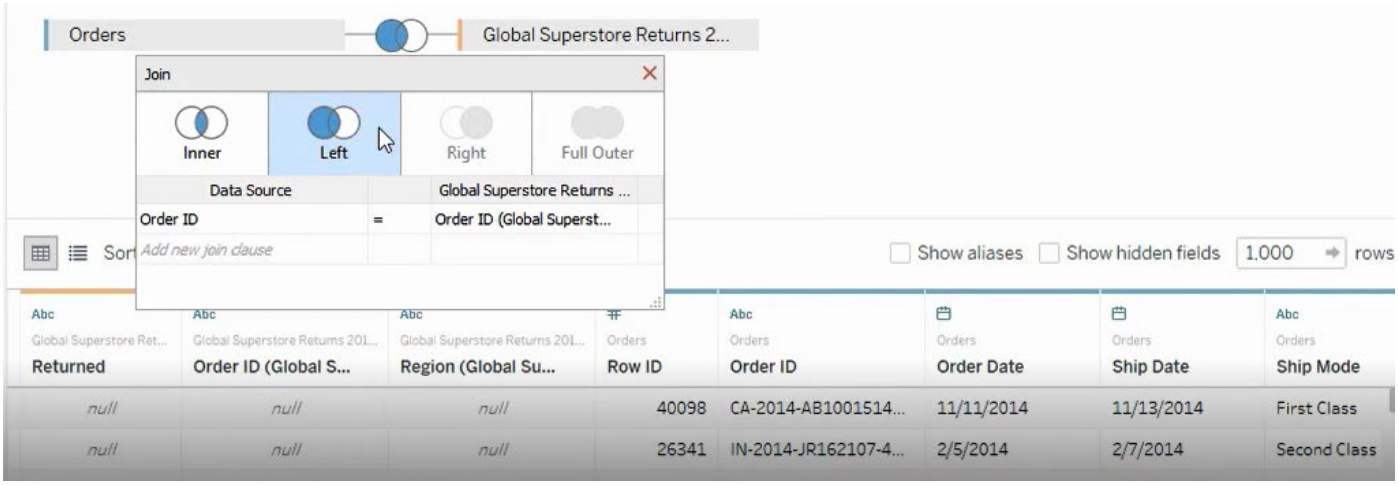
4. Drag the order table to the canvas

The primary set used for this training contains a list of worldwide company transactions described with 24 attributes: Row ID, Order Priority, Discount, Unit Price, Shipping Cost, Customer ID, Customer Name. Ship Mode, Customer Segment, Product Category, Product Subcategory, Product Container, Product Name, Product Base Margin, Country, Region, State or Province, City, Postal Code, Order Date, Ship Date, Profit, Quantity ordered new, Sales, Order ID.

PREPROCESSING FUNCTIONALITY

5. Extract more information from the same data source. Drag the other table onto the workspace.
6. Integrate the data by adding a connection to the other source. A text file of returned orders is saved in a CSV format.
7. Edit the join in the appropriate icon.

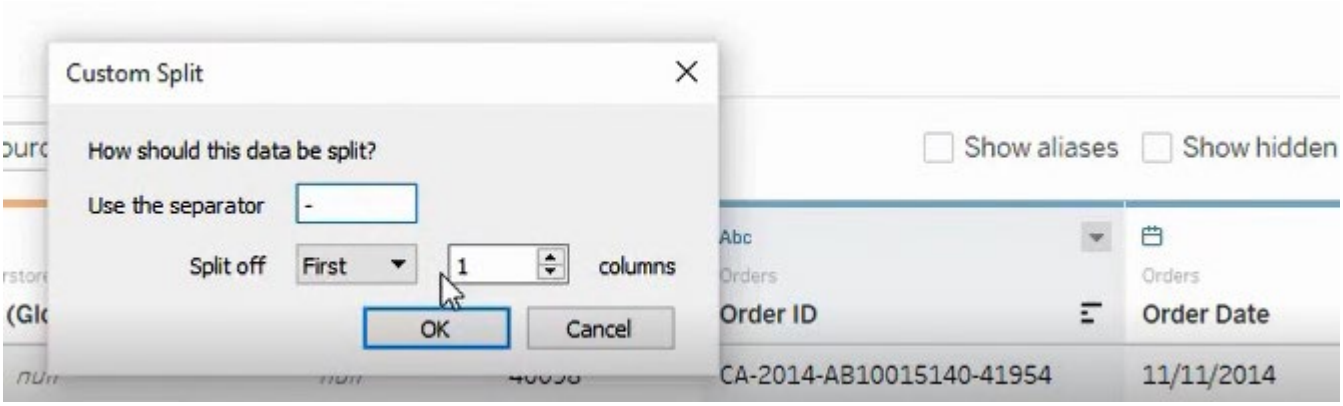
We choose a left join, so we get all the information from the orders table and only relevant returns information. It's already based on order id as the join clause, but we could change this if desired.



Co-funded by the Erasmus+ Programme of the European Union

In the integrated table, the join parts are coded in color. The order return data is in the orange line, and the information on all transactions is marked with a blue line. In this grid view, we can do some essential metadata management.

8. Split the order ID field. It has multiple parts, such as the code of distribution center, the year and two additional codes. Use custom split and a split on a hyphen. Rename the field to a distribution center.

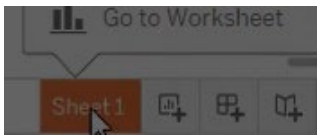


9. Connect to Live and click on the sheet tab down at the bottom.

Another option in the Data Source Pane is the kind of connection to the data we desire: live or extract the data. Connecting live is excellent when we have constantly changed data or want to leverage the high-performance database we are connected to. Alternatively, we may choose to import data into Tableau's fast engine with an extract that takes the data offline and minimizes performance impact while still allowing regularly scheduled refreshes to keep the data up to date.

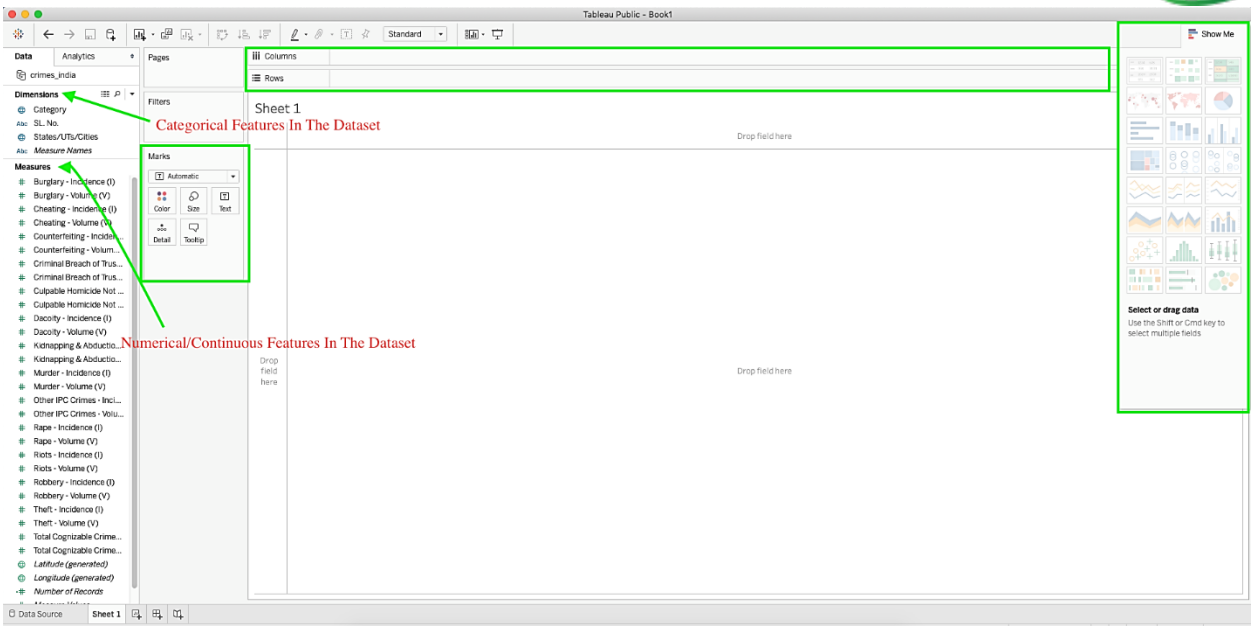
Dimensions and Measures are Tableau's way of distinguishing Categorical and Numerical features from the dataset.

- **First go to the worksheet**



- Dimensions are categorical features responsible for a graph's different dimensions or axes.
- Measures are the continuous values representing a datapoint plotted along an axis.

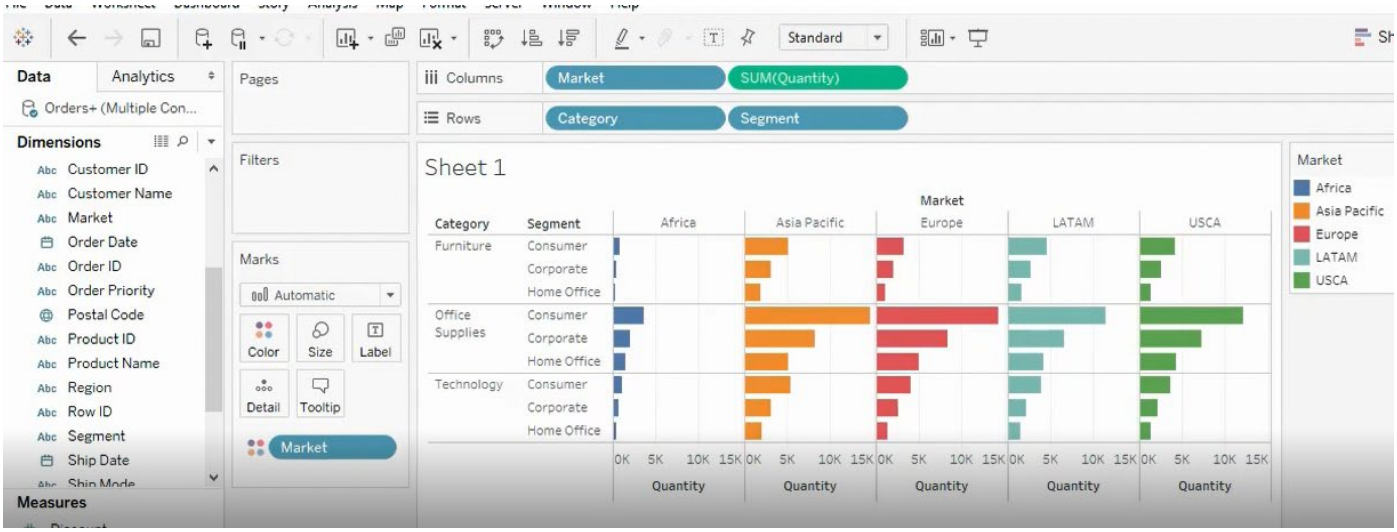
Co-funded by the Erasmus+ Programme of the European Union



10. Bring category to rows, segment to rows, quantity to columns, market to columns. Bring market to color.

VISUAL ANALYSIS & TECHNIQUES

It's that easy to visualize how the sales look per category, customer segment and market. We can quickly see that Africa is an emerging market.



The left pane is broken up into dimensions and measures that represent the column headers in the excel sheet

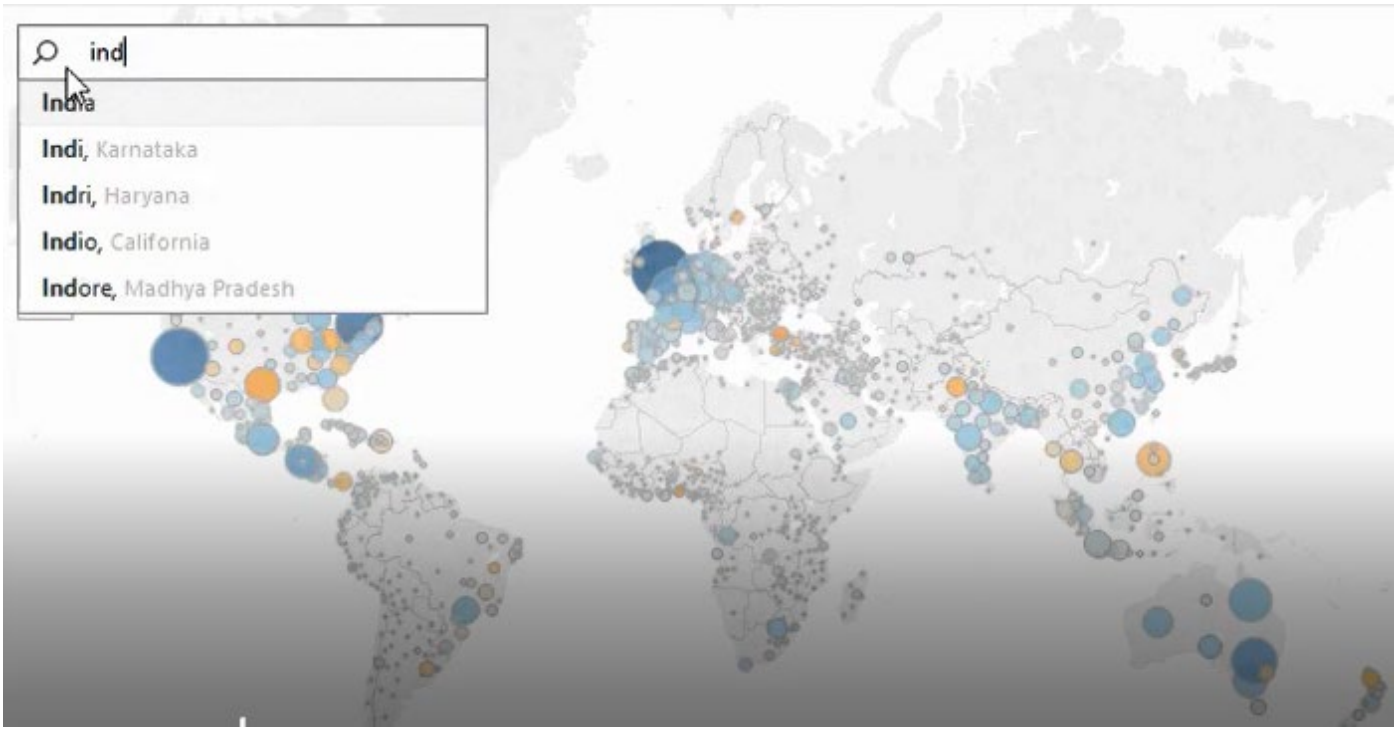
In this case, the dimensions are categorical fields such as date, customer and category. They are often discrete fields and create labels in the chart also are colored in blue. The measures, on the other hand, are our metrics. They are the numbers we want to analyze. Measures are often continuous fields that create axes in the chart and are colored green.

11. Compare what quarterly growth looks like over the years

Co-funded by the Erasmus+ Programme of the European Union

12. Present as a cross-table the following visualization of Sales Seasonality.

13. Export the table in an Excel spreadsheet.
14. Activate the map window from Show Me Pane and load the data on profit and location where the sale took place in the canvas. Indicate the area to which the settlement belongs. Use the size shadow and color settings to present the data more understandably. Different colors should be according to the sales profit.



The map is interactive, and we can choose the desired area or location.

FORECASTING

Co-funded by the Erasmus+ Programme of the European Union

15. Click on forecasting, choosing it from the left Analysis pane.

It is effortless to make a forecast in Tableau, although there are some limitations, one of which is at least five data points, and if the data is seasonal, at least two seasons' worth of data is needed. From the analysis panel, we select forecast. The view has both the forecast and prediction intervals. The last ones are presented as a shaded band. If we change the mark type to a circle the prediction interval becomes whiskers around the point.

The model selection is automatic, but there are some things we, as the user, can adjust. If we want to customize the default forecast, right-click in the view, select forecast, forecast options.

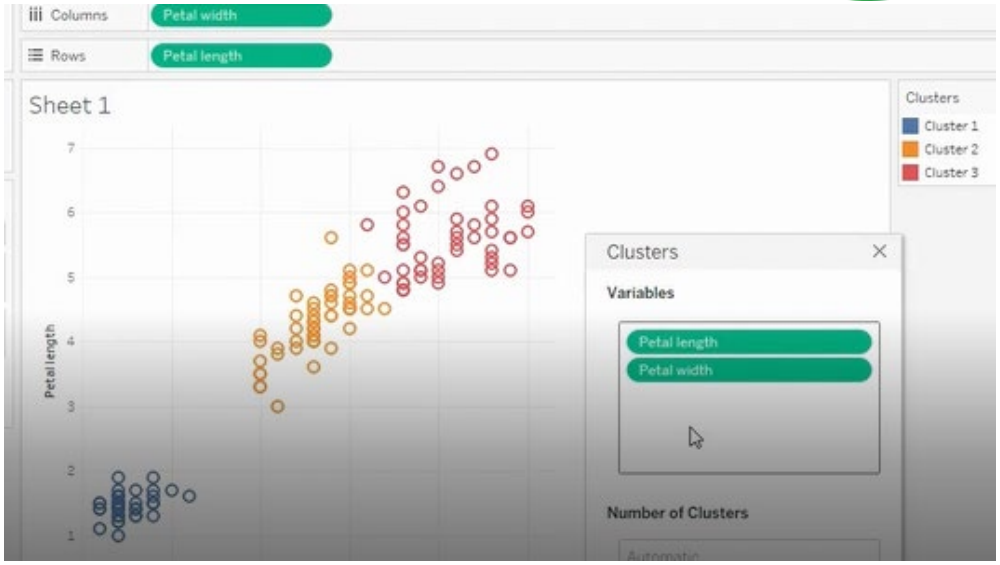
CLUSTERING

The cluster analysis demonstration uses the familiar set of data about the flower iris and its varieties based on the length and width of the sepals and petals.

16. Load the data.
17. Select Clustering from the Analysis Pane.

Co-funded by the Erasmus+ Programme of the European Union

Tableau uses only the k-means method for clustering. We can control the number of clusters and variables to calculate the cluster but not the algorithm itself.



Co-funded by the Erasmus+ Programme of the European Union

WORKSHOP

The session started with team building. The twelve students from ULSIT (Bulgaria), UNi (Serbia), and UBB (Poland) participating in the classes form four teams by blending trainees from the consortium's countries. Each team had to perform its own data-driven project appling Orange, or/and Tableau. The results have to be reported by making an appropriate presentation.

The compulsory components are:

- Objective
- Data
- Methodology
- Experiment set-up
- Results
- Conclusion

The typical workflow for the project may vary, but the following steps have to be considered:

- Data collection
- Data understanding (EDA)
- Data preparation:
  - data cleaning (missing data, duplicated data, irrelevant data, outliers)
  - data transformation (wrangling and feature engineering)
- Model building
- Evaluation
- Communicate Results

Trainers with trainees choose the topics for the projects taking into account the academic background.

The workshop passed through two stages. The first (team building, topic and objective identification, data detection, and brainstorming) occurred in the formal setting of the university hall.Thus, the second stage has been organized at the coffee house vanue of the Nis amphiteathre.

Co-funded by the Erasmus+ Programme of the European Union

TEAM 1

MEMBERS

Djordje Antic (UNi) Kacper Palka (UBB) Damian Grygierczyk (UBB)

TEAM NAME:

"Team 1"

TOPIC

Business Case: Bank Churners

OBJECTIVE

- Predict bank customer churn
- Comparison of the Python implementation and the Orange widget capabilities

TOOL(S)

- Python
- Orange

METHODS

Machine Learning

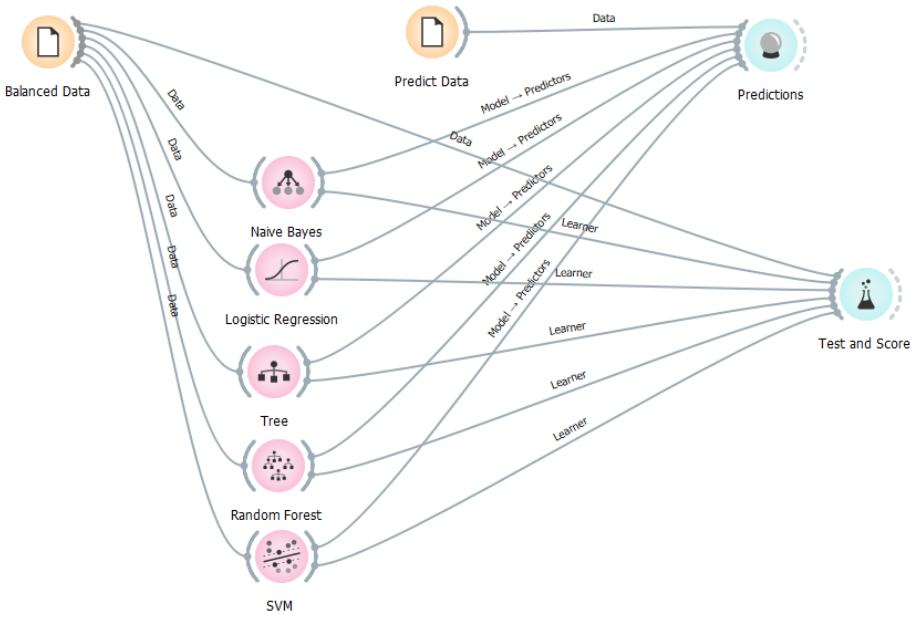
Algorithms: Decision Tree, Random Forest, Support Vector Machines (SVM), Logistic Regression, and Naïve Bayes.

DATA

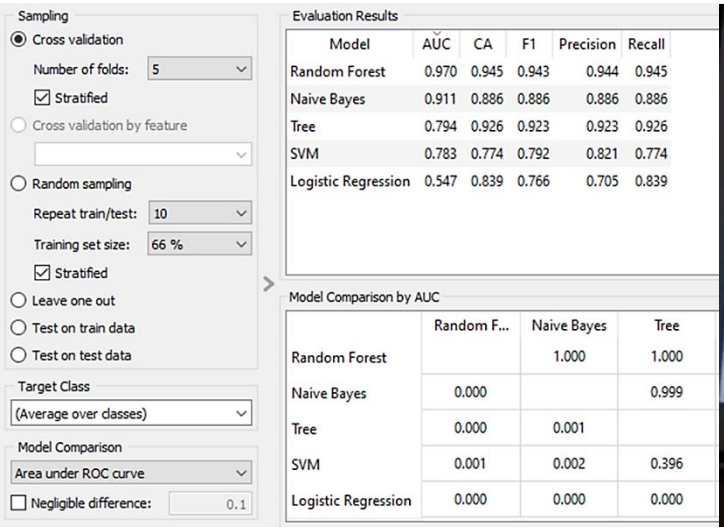
The set consists of 10,000 customers' metadata such as age, salary, marital status, credit card limit, credit card category described in 21 features. The target variable is attrition flag: account is closed 1, else 0.

Co-funded by the Erasmus+ Programme of the European Union

WORKFLOW



RESULTS



Sampling
- Cross validation
  - Number of folds: 5
  - Stratified
- Cross validation by feature
- Random sampling
  - Repeat train/test: 10
  - Training set size: 66 %
  - Stratified
- Leave one out
- Test on train data
- Test on test data

Target Class: (Average over classes)

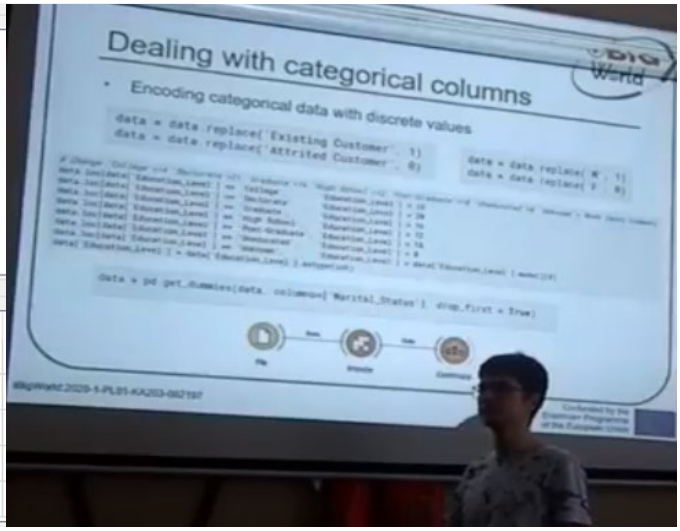
Model Comparison: Area under ROC curve; Negligible difference: 0.1

Evaluation Results

Model	AUC	CA	F1	Precision	Recall
Random Forest	0.970	0.945	0.943	0.944	0.945
Naive Bayes	0.911	0.886	0.886	0.886	0.886
Tree	0.794	0.926	0.923	0.923	0.926
SVM	0.783	0.774	0.792	0.821	0.774
Logistic Regression	0.547	0.839	0.766	0.705	0.839

Model Comparison by AUC

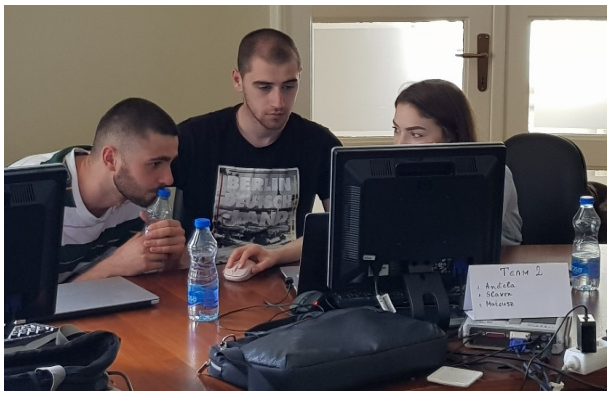
	Random F...	Naive Bayes	Tree
Random Forest		1.000	1.000
Naive Bayes	0.000		0.999
Tree	0.000	0.001	
SVM	0.001	0.002	0.396
Logistic Regression	0.000	0.000	0.000



TEAM 2

MEMBERS

- Mateusz Damek (UBB) on the left side
- Slaven Panov (ULSIT) in the middle
- Andjela Kostic (UNi) on the right side



Co-funded by the Erasmus+ Programme of the European Union

TEAM NAME:

"Investigators"

TOPIC

Medical Case:

Classification of COVID-19 and non-COVID-19 from computed tomography imagies

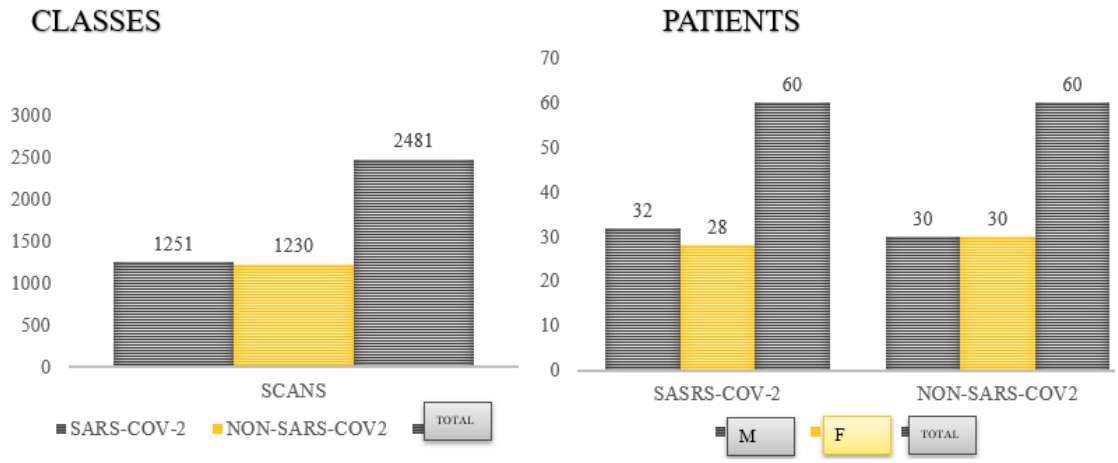
OBJECTIVE

Big Data workflow development for COVID-19 CT scan binary classification: COVID-19 pneumonia or non-COVID 19.

DATA

Open data from Public Hospital of the Government Employees, Metropolitan Hospital of Lapa, Sao Paulo, Brazil

Available at: Kaggle and GitHub



REFERENCE

Soares E, Angelov P, Biaso S, Froes MH, Abe DK. Sars-cov-2 ct-scan dataset: a large dataset of real patients ct scans for sars-cov-2 identification. medRxiv; 2020.

TOOLS

Orange extended by Image Analytics add-on.

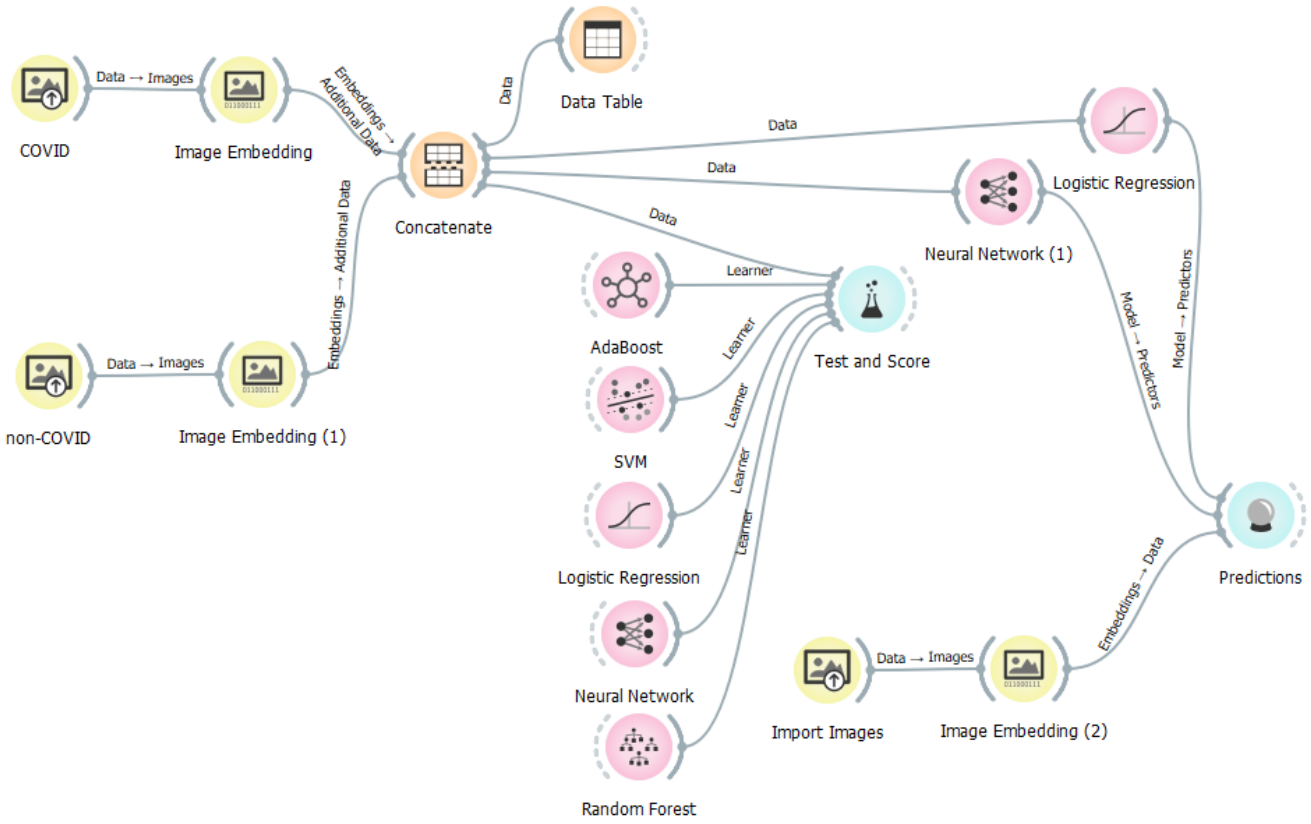
METHODS

Transfer Learning with Inception v3 using image embedding widget for features extraction and applying the following algorithms: a multi-layer perception (MLP) algorithm with backpropagation and

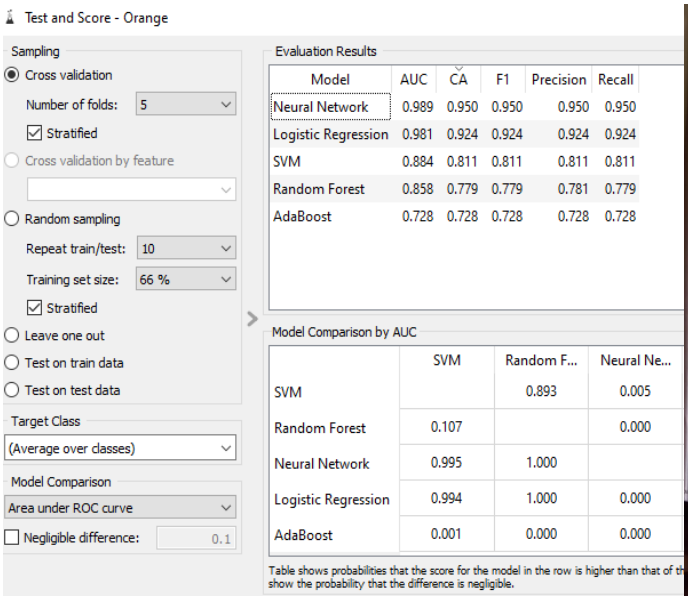
Co-funded by the Erasmus+ Programme of the European Union

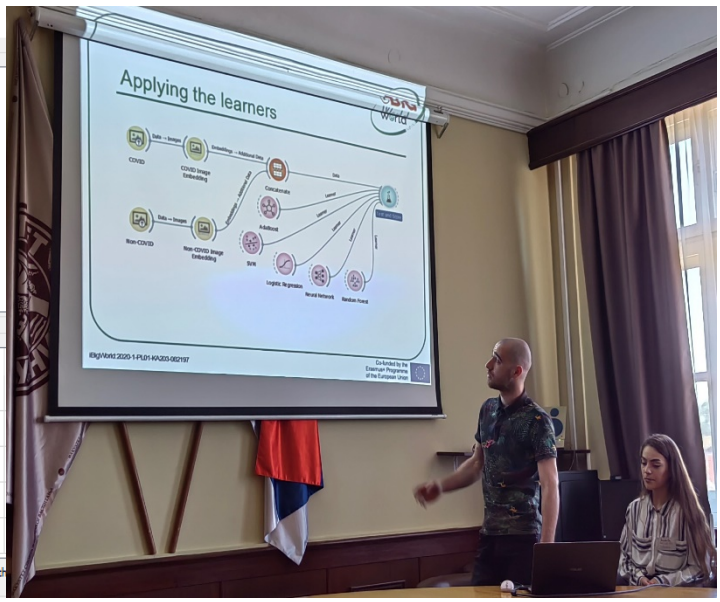
logistic regression. The last two are used as the best approaches after comparing the results of MLP, Logistic Regression, Ada Boost, Random Forest and SVM.

WORKFLOW



RESULTS



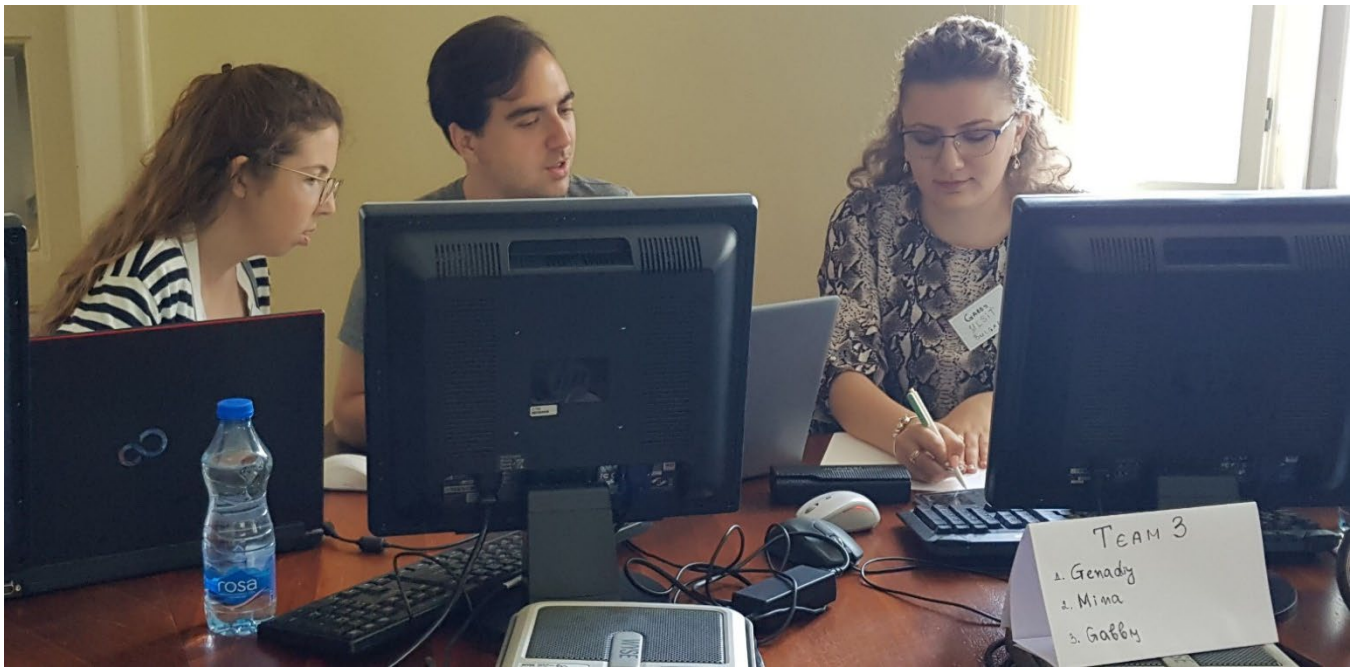


Co-funded by the Erasmus+ Programme of the European Union

TEAM 3

MEMBERS

Mina Krstic (UNi), Genadiy Gospodinov (ULSIT), Gabriela Naydenova (ULSIT)



TEAM NAME

"MGG"

TOPIC

Business Case: Interactive Data-Driven Dashboard for Business Data Analysis (Business Intelligence)

OBJECTIVE

Creating an interactive dashboard and storyline for: analysis of quantity of sales, profit of sales, and shipping costs per product category and subcategory for the four USA regions where the company operates (East, West, South, and North).Creating an interactive dashboard and storyline for: analysis of quantity of sales, profit of sales, and shipping costs per product category and subcategory for the four USA regions where the company operates (East, West, South, and North).

DATA

Tableau Superstore dataset with 10 000 sales transactions organized in 20 features for the period 2014 – 2018.

TOOLS

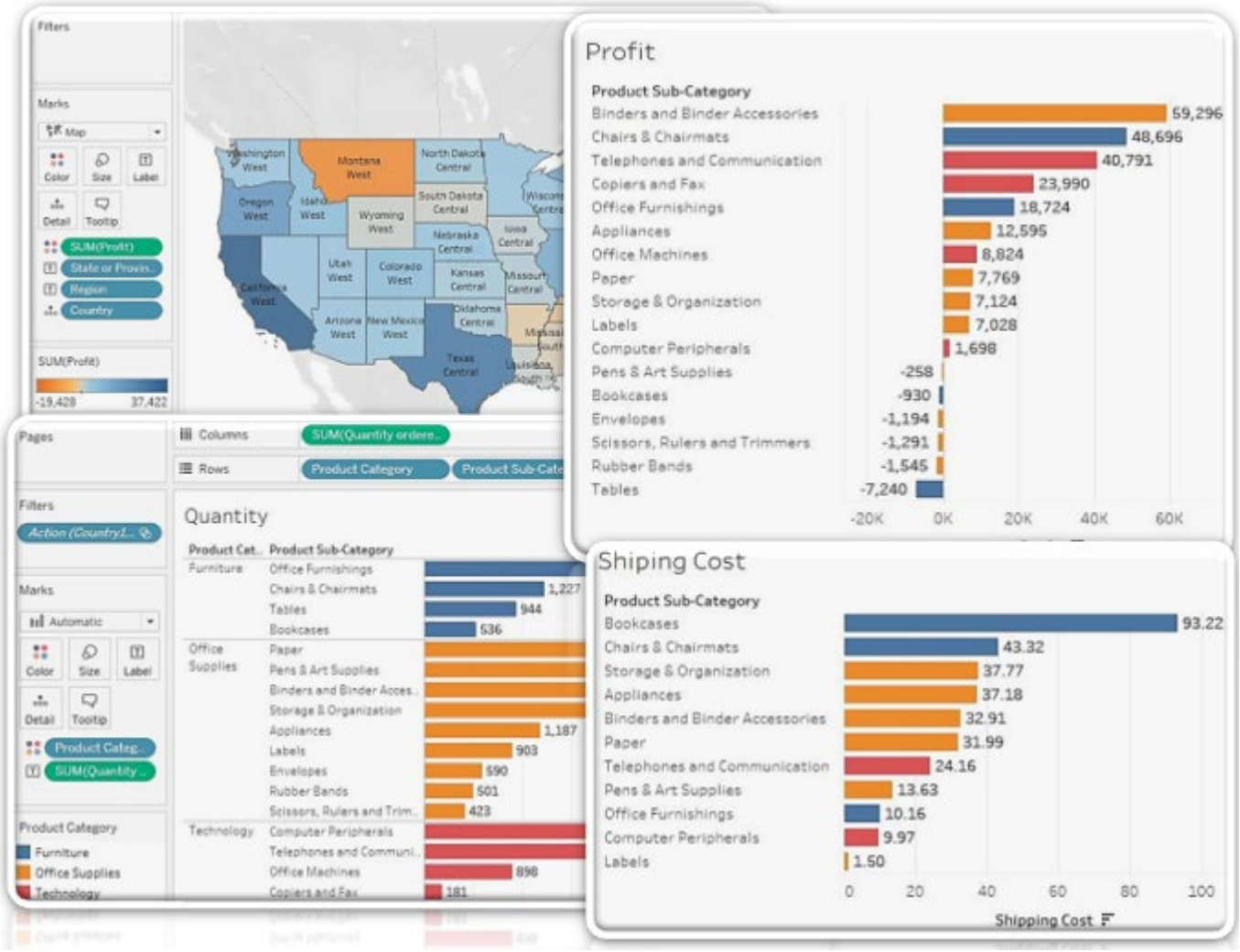
Co-funded by the Erasmus+ Programme of the European Union

Tableau Public

METHODS

Summary statistics and visualization techniques for exploratory data analysis (EDA).

EXAMPLE PLOTS AND DASHBOARDS



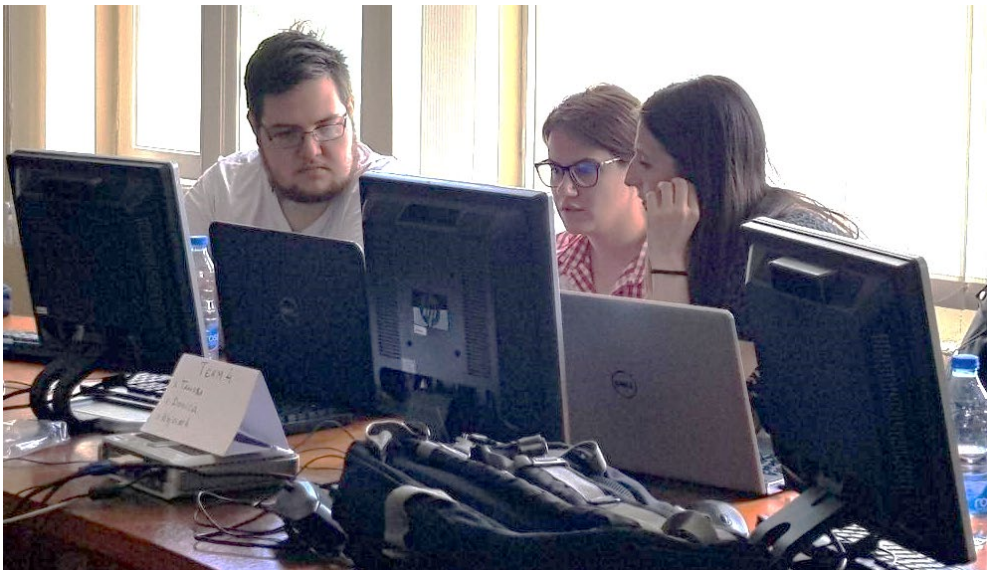
Co-funded by the Erasmus+ Programme of the European Union

RESULTS: STORYLINE

TEAM 4

MEMBERS

Wojciech Kłósko (UBB) Danica Pejcic (UNi) Tamara Ristovska (ULSIT)



TOPIC

Administrative case: Human Development Index

OBJECTIVE

Regression task: Predict HDI (Human Development Index)

Cluster analysis: Grouping by HDI

Co-funded by the Erasmus+ Programme of the European Union

DATA

Data were taken from UNITED NATIONS DEVELOPMENT PROGRAMME, Human Development Reports, Data Center. An adapted version is available as prebuild dataset named HDI. It contains information on more than 180 countries, which are described with more then 60 features. Indeed, the Human Development Index (HDI) is a summary measure of average achievement in key dimensions of human development: a long and healthy life, being knowledgeable and having a decent standard of living. The HDI is the geometric mean of normalized indices for each of the three dimensions.

TOOLS

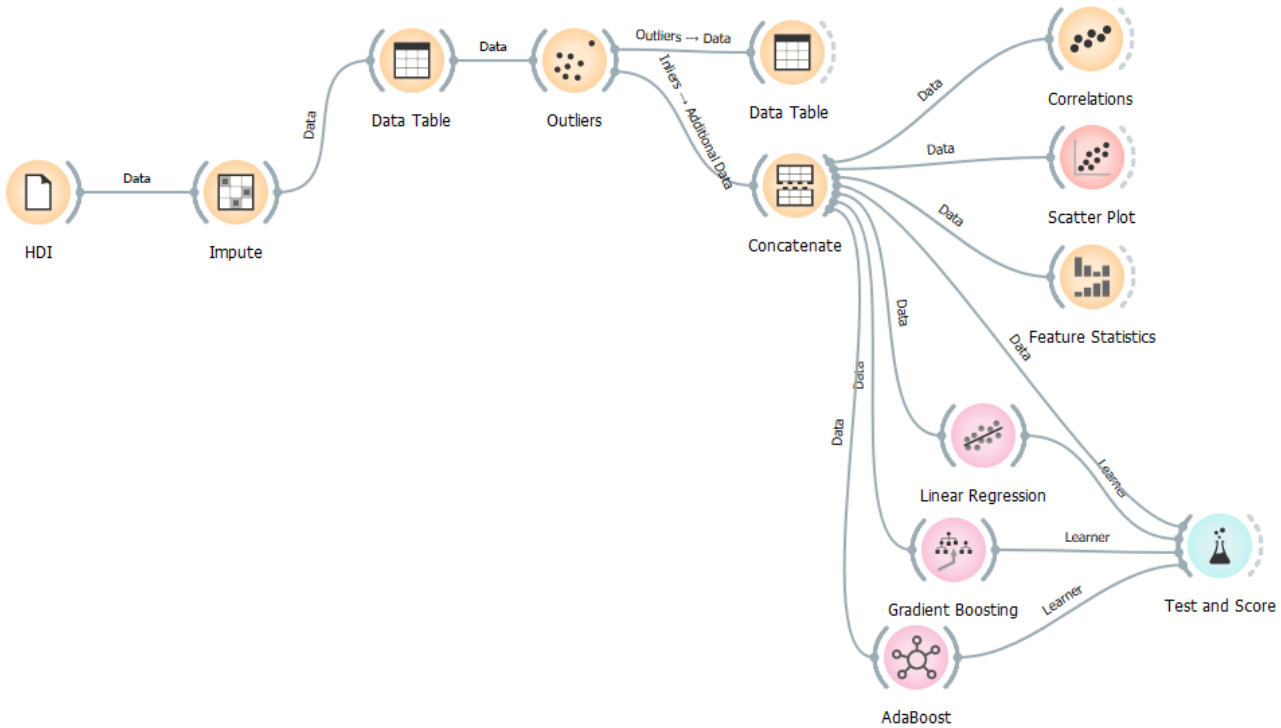
Orange

METHODS

Regression analysis through Linear Regression, AdaBoost, Gradient Boosting and cluster analysis through k-Means, hierarchical clustering and Louvain clustering.

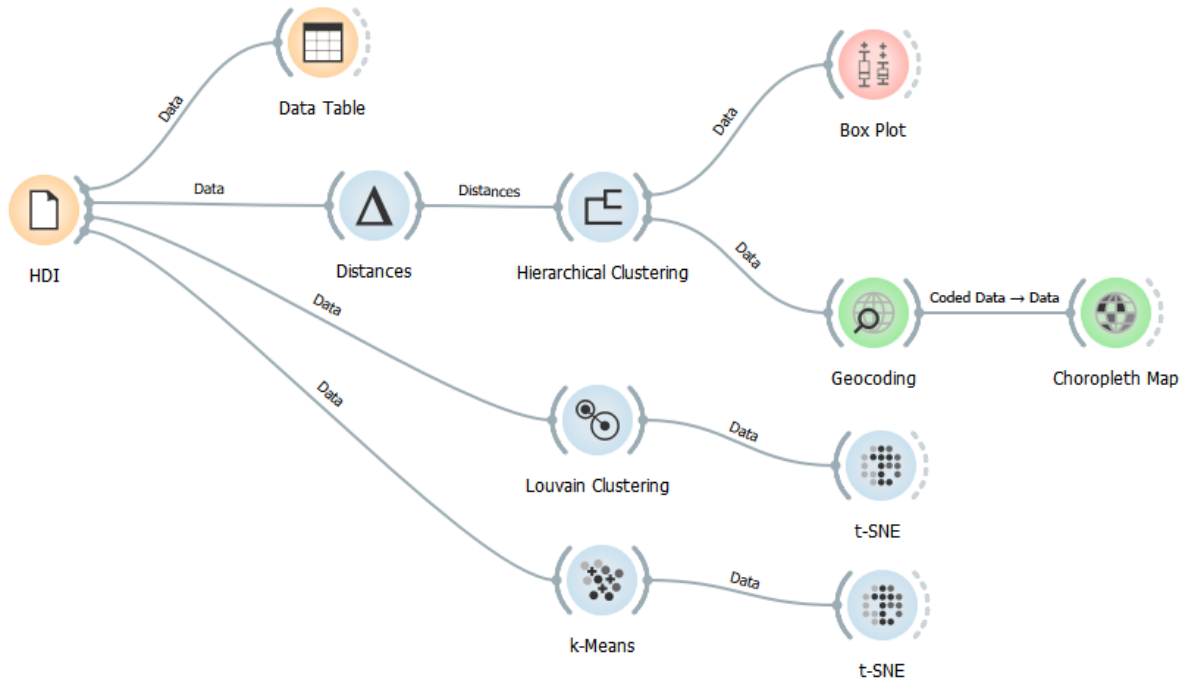
WORKFLOWS

- Regression



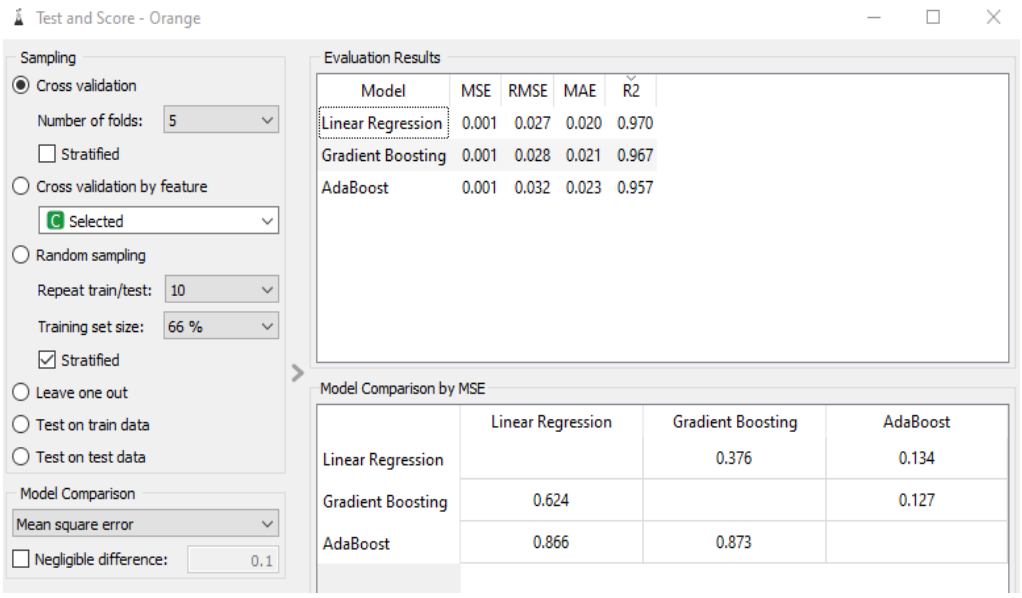
- Clustering

Co-funded by the Erasmus+ Programme of the European Union



RESULTS

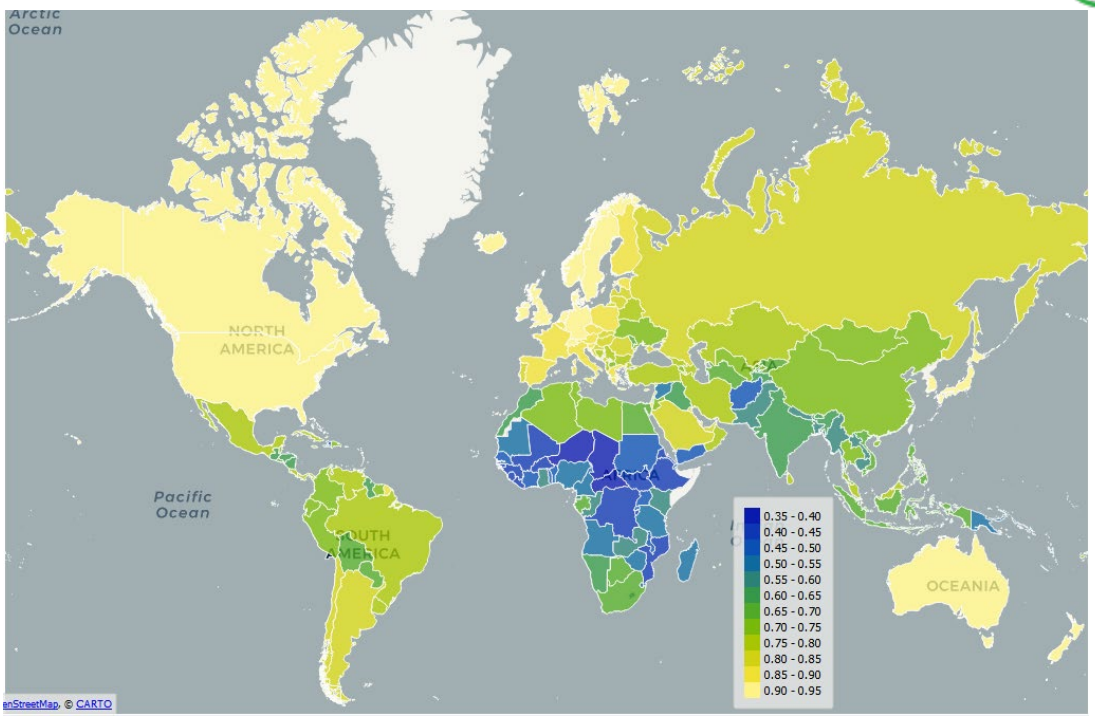
- Regression



R square value is about 0.96 – 0.97. The model is explainable up to 97 %, or the independent variables can explain up to 97% dependent variables.

- Clustering

Co-funded by the Erasmus+ Programme of the European Union



CONCLUSION

The short–term training aims to broaden the students' interest in data-driven projects as a novel paradigm for knowledge discovery and business insights. In such a context, the training answers a series of questions that have arisen:

- What acts in the novel value-chain paradigm?
- What are the key technologies behind BDA?
- What tools and software enable meaningful insight from big data?
- What tools and software can be used for BDA with no or minimal coding skills?

Practical sessions make it possible to examine primary tasks such as regression, classification, clustering, and feature engineering by developing models using particular software tools. That deepens the knowledge of Big Data, Data Mining, and Machine Learning, opening horizons for uncovering new insights.

The learning results confirm that applied methods such as learning-by-doing and real-case-based training are appropriate when dealing with the multidisciplinary nature of a problem. Combining visual programming tools that require no or little coding knowledge in Python or R allows instructors to work with students who have no prior experience or basic knowledge of the subject. That also helps to pay attention to the principles, concepts, and applications of Big Data algorithms in various domains, illustrating the use cases, the need for team collaboration, and understanding what models are suitable for the particular problem and why. Then, we can keep going with advanced parameter optimizations to tune the model and achieve more accurate results.

Co-funded by the Erasmus+ Programme of the European Union

Trainees have some difficulties with communicating the results and making detailed conclusions and recommendations, which we are addressing by changing the learning environment to be more informal, working in teams and discussions.

The students shared that after the workshop they have more clear ideas about data-driven applications and the methods and tools to uncover new knowledge from big data.

REFERENCES

Pedamkar, P. Orange Data Mining. [Online]. Available: https://www.educba.com/orange-data-mining/.
Erden, C. Orange Data Mining Tool and Association Rules, 11 May 2020. [Online]. Available: https://towardsdatascience.com/orange-data-mining-tool-and-association-rules-caa3c728613d.
Demšar, B. Z. J. Orange: Data mining fruitful and fun – A historical perspective, March 2013. [Online]. https://www.researchgate.net/publication/289842192_Orange_Data_mining_fruitful_and_fun_-_A_historical_perspective.
Orange Data Mining, 18 Feb 2020. [Online]. Available: https://towardsdatascience.com/tagged/orange-data-mining.
Orange Data Mining, [Online]. Available: https://www.javatpoint.com/orange-data-mining.
Orange Visual Programming Documentation, Release 3, [Online], Available: https://buildmedia.readthedocs.org/media/, https://orange3.readthedocs.io/projects/orange-visual-programming/en/latest/index.html
Tableau Getting Started, [Online]. Available: https://github.com/arpitran/Tableau-Workshop-Getting-Started
Nair, A. A Hands-On Guide For Beginners. [Online], Available: https://analyticsindiamag.com/tableau-101-a-hands-on-guide-for-beginners/
Menasalvas, E. et al. (2021). Recognition of Formal and Non-formal Training in Data Science. In: Curry, E., Metzger, A., Zillner, S., Pazzaglia, JC., García Robles, A. (eds) The Elements of Big Data Value. Springer, Cham. https://doi.org/10.1007/978-3-030-68176-0_13
Competency Framework for big data acquisition and processing. [Online], Available: https://unstats.un.org/bigdata/task-teams/training/UNGWG_Competency_Framework.pdf
Select Hub. [Online], https://www.selecthub.com
Heart Disease Prediction, https://github.com/yash2189/Heart-Disease-Prediction-ML
Human Development Reports, https://hdr.undp.org/data-center/human-development-index#/indicies/HDI

Co-funded by the Erasmus+ Programme of the European Union

Big Data Analitic Tools

Study Guide

English
First edition

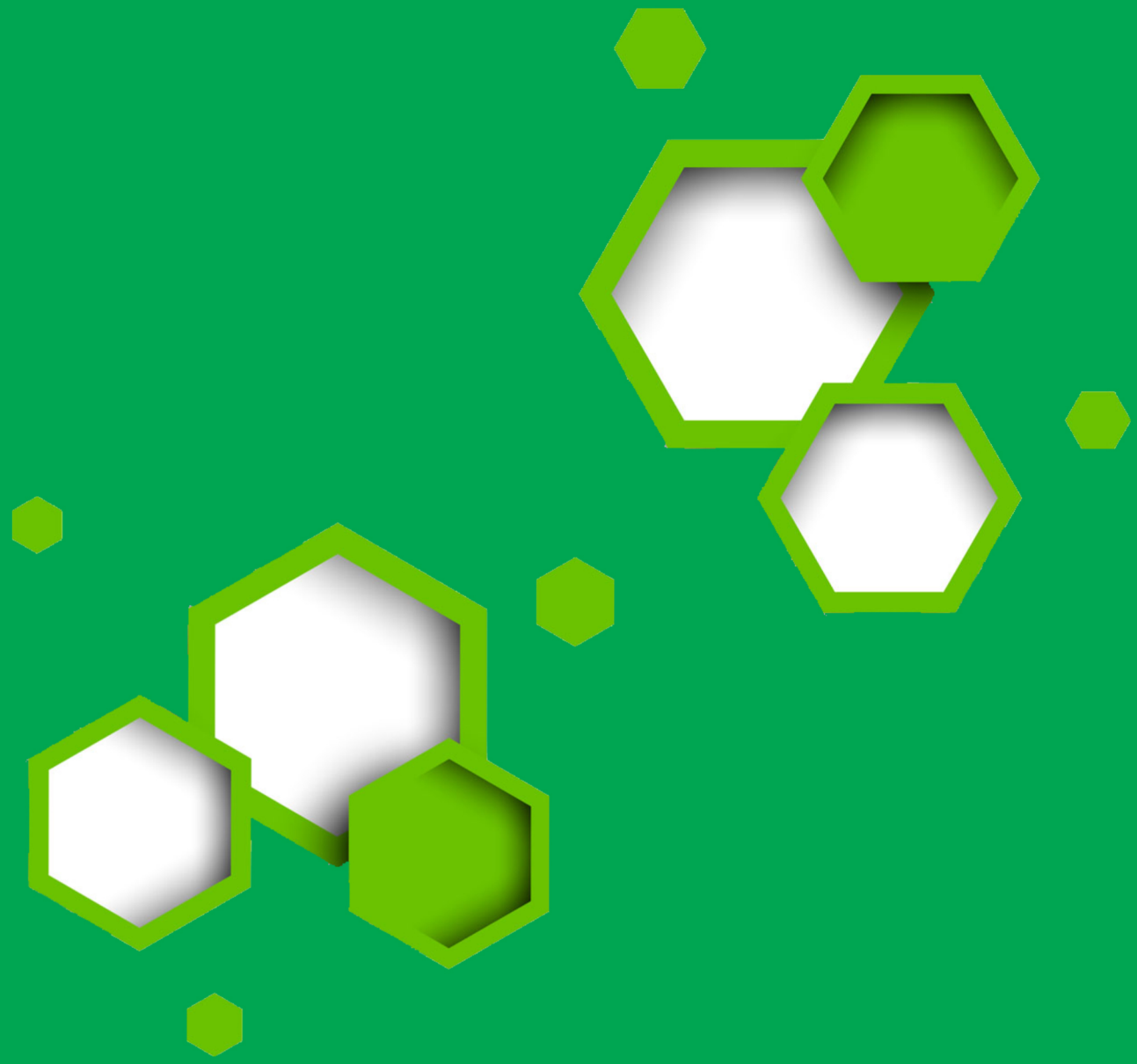
Authors
Lubomir Gotsev, Iva Kostadinova

Scientific Editor
Vasyl Martsenyuk, Georgi Dimitrov

Graphic Design and Cover
Lubomir Gotsev, Diana Stoyanova

Academic Publisher "Za bukvite – O pismeneh"
ISBN 978-619-185-572-8
Sofia, 2022

Co-funded by the Erasmus+ Programme of the European Union



ISBN 978-619-185-572-8