**Charotar University of Science and Technology [CHARUSAT]**

**Chandubhai S. Patel Institute of Technology [CSPIT]**

**U & P U. Patel Department of Computer Engineering**

**Lab Manual**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Subject code | : | CE-350 | Semester | : | 6 | Academic Year | : | 2021 |
| Subject name | : | Data Warehouse and Data Mining | | | | | | |

**Practical – 5**

**Aim:**

**Perform data mining operations like basic operations [viewing data, charts [Histogram, Boxplot Analysis- Visualization] association rule mining, classification, Generating Classification Tree -Decision Trees, Cross Validation, text mining using Orange- An open-source data visualization and analysis tool.**

* **Prerequisite**

S/W: - Orange Tool

Little bit knowledge about the Orange tool and how it works on Data Mining.

* **Step to install Orange Tool:**
* Download and open [Orange3 installation file](http://orange.biolab.si/download/files/Orange3-3.2.dev0+e196459.win32-py3.4-install.exe).
* Follow installation wizard to install Orange 3 and its dependencies (Python 3.4, NumPy, etc.).
* Install Bioinformatics add-on.
* Restart Orange 3

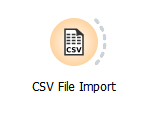
**Process flow:**

Orange is basically a drag and drop GUI application. If u follow along the screenshots, you could end up with the same outputs.

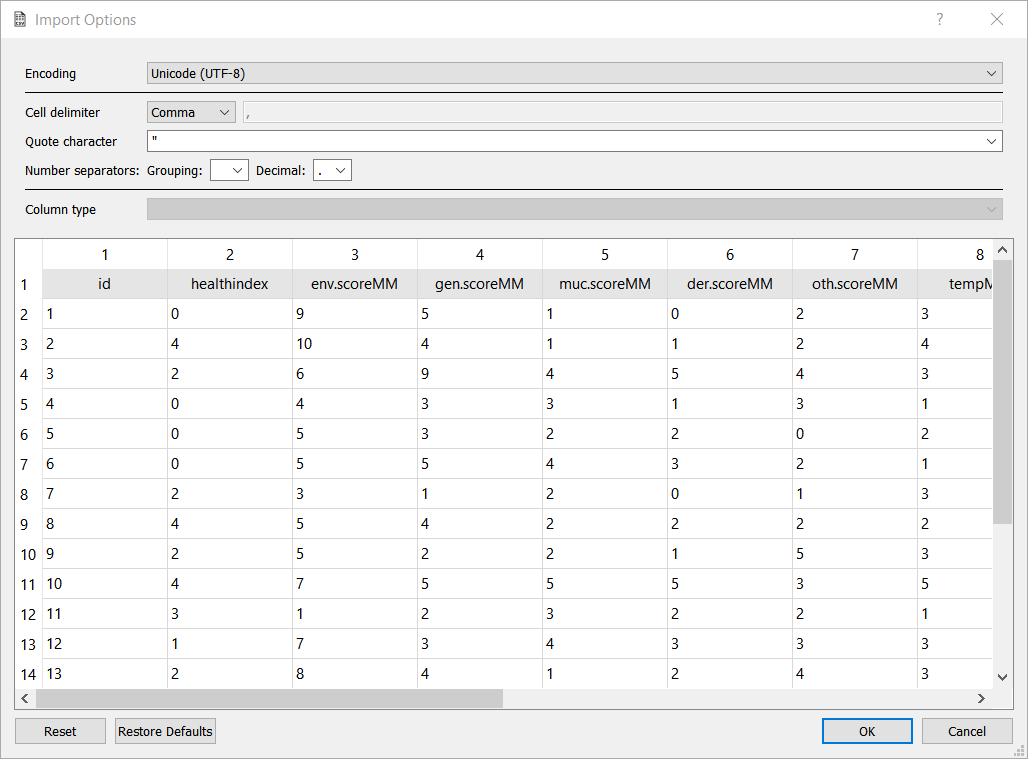
**Step 1: Selecting a dataset:**

For this particular example, I will use a Health Index Dataset.

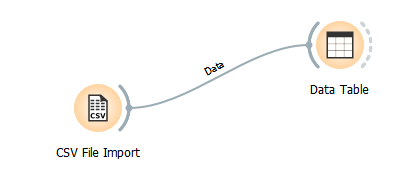
To import a particular dataset, you would drag and drop “CSV File Import” from the tool window.

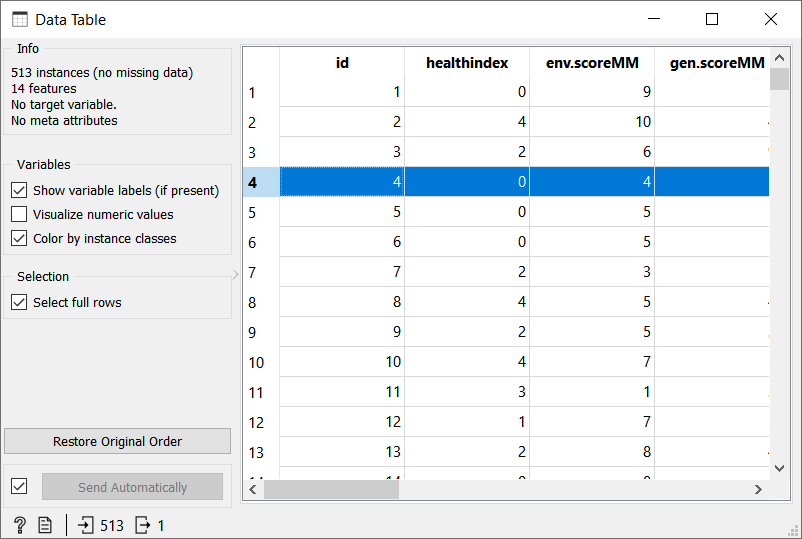


Next you would browse to find the particular dataset you need. Before you click OK you can preview your dataset and/or make changes.



To view your dataset again, connect a “Data Table” tool to your “CSV File Import”.

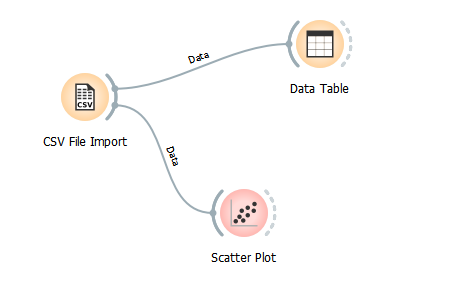




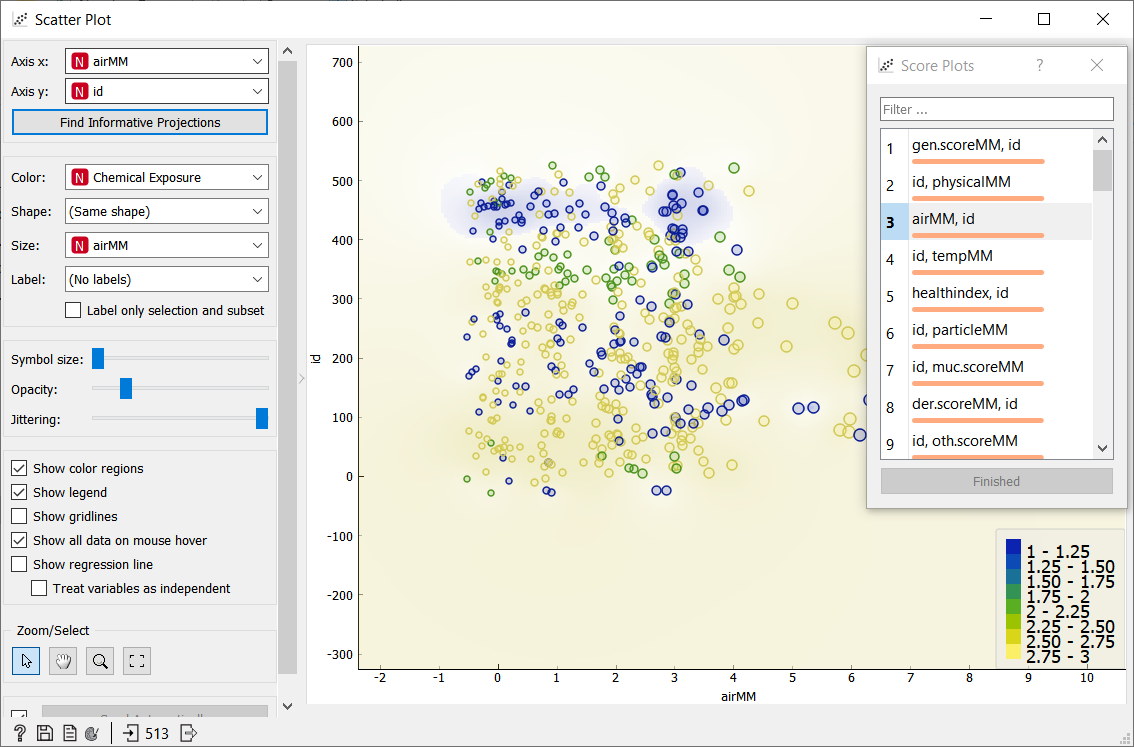
**Step 2: Visualizations:**

* **Scatter Plot**

Connect “CSV File Import” to a scatter plot tool.

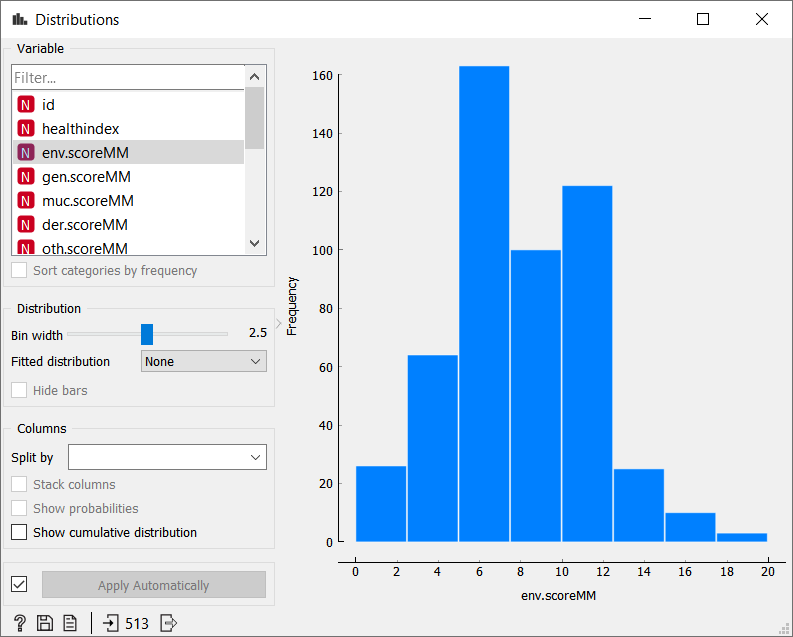
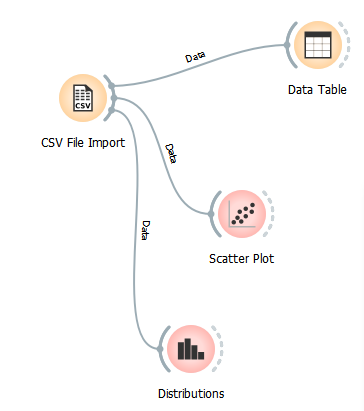


You can either set your own scatter plot or click on the “Find Informative Projections”. Orange Tool is a Robust Application and can find out meaningful relations on its own. Use it make your life easier.



* **Histogram**

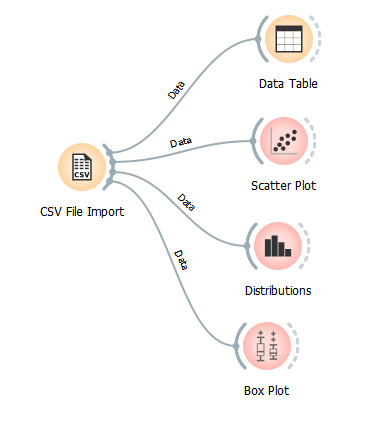
For Histogram make use of “Distributions” tool.

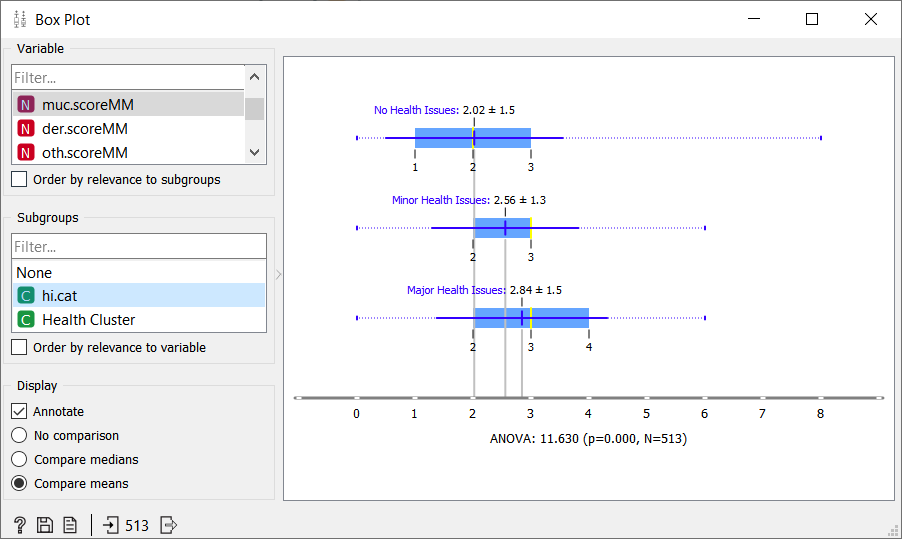


You will be able to set filters as per your need.

* **Box Plot**

For Box-plot analysis, you have to attach “Box Plot” tool to the CSV File import tool.





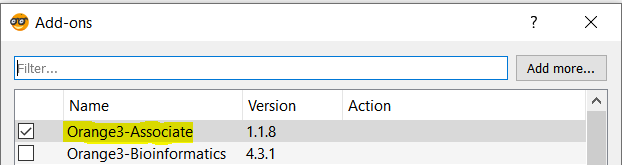
You can choose what variable you would like a box plot analysis. Great thing about Orange is it can automatically identify certain sub-groups. Analyse them if you deem them of any use.

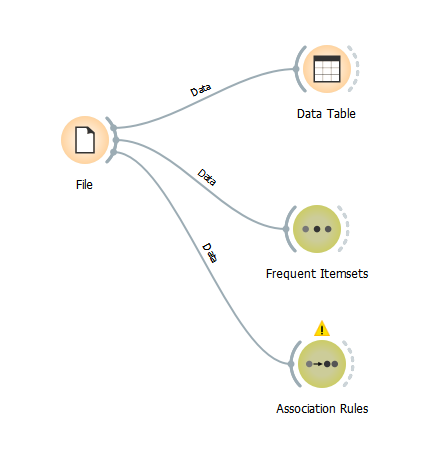
**Step 3: Association Rule Mining:**

Let’s take a **different dataset** this time and perform a Market-Basket Analysis. This is a dataset from a grocery store. Let’s check which items go hand in hand while shopping…

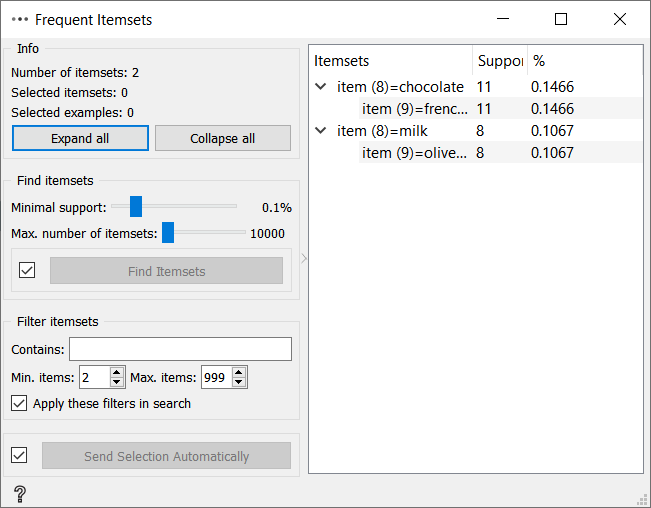
First you will have to install the following:

Options>Add-Ons>Orange3-Associate

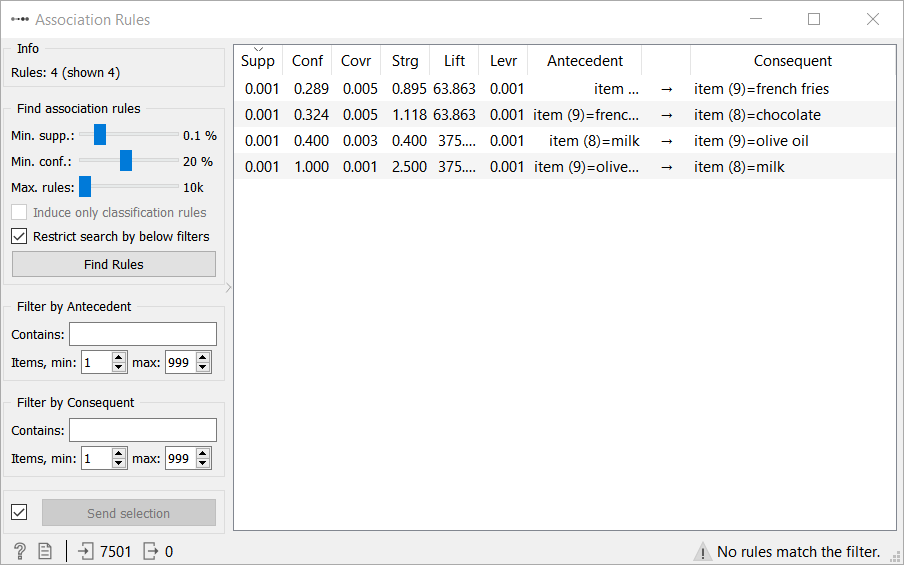




First let’s check our data set. Now we connect Frequent Itemset to our File widget and observe the results. We’re mostly interested in the action-consequence relationship here. In other words, if a person buys one item, what is the most likely second item she will buy? Association Rules will help us discover that.



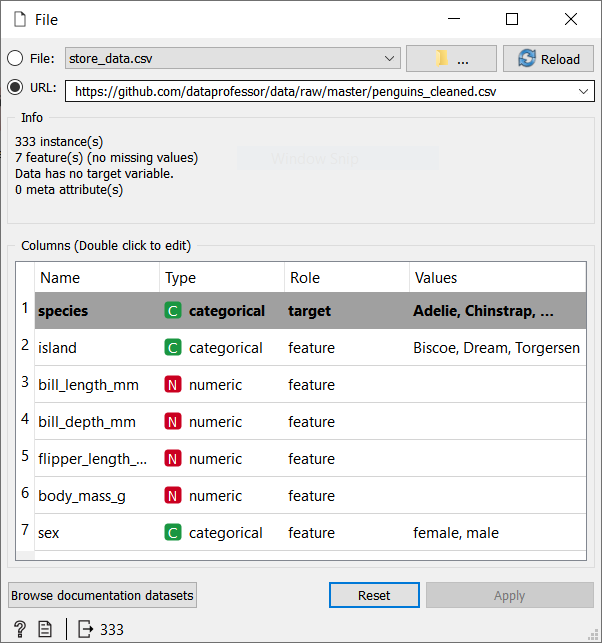
If there’s a little arrow on the left side of the item, you can expand it to see all the other items connected to the selected attribute. So, if a person buy fresh vegetables, it is most likely to buy fresh fruits as an accompanying product group. Now you can explore frequent item sets to understand what really sells in your store.



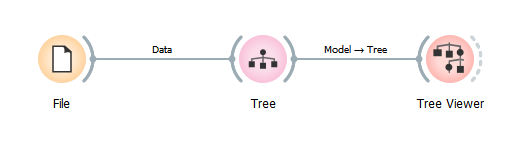
**Step 4: Classification Tree:**

Taking a new dataset for classification. A penguin species classification dataset (<https://github.com/dataprofessor/data/raw/master/penguins_cleaned.csv>)

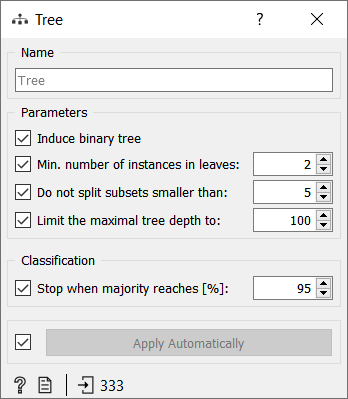
To begin with I will find what column values I need to classify (in this case the species) and change its role to **target**.

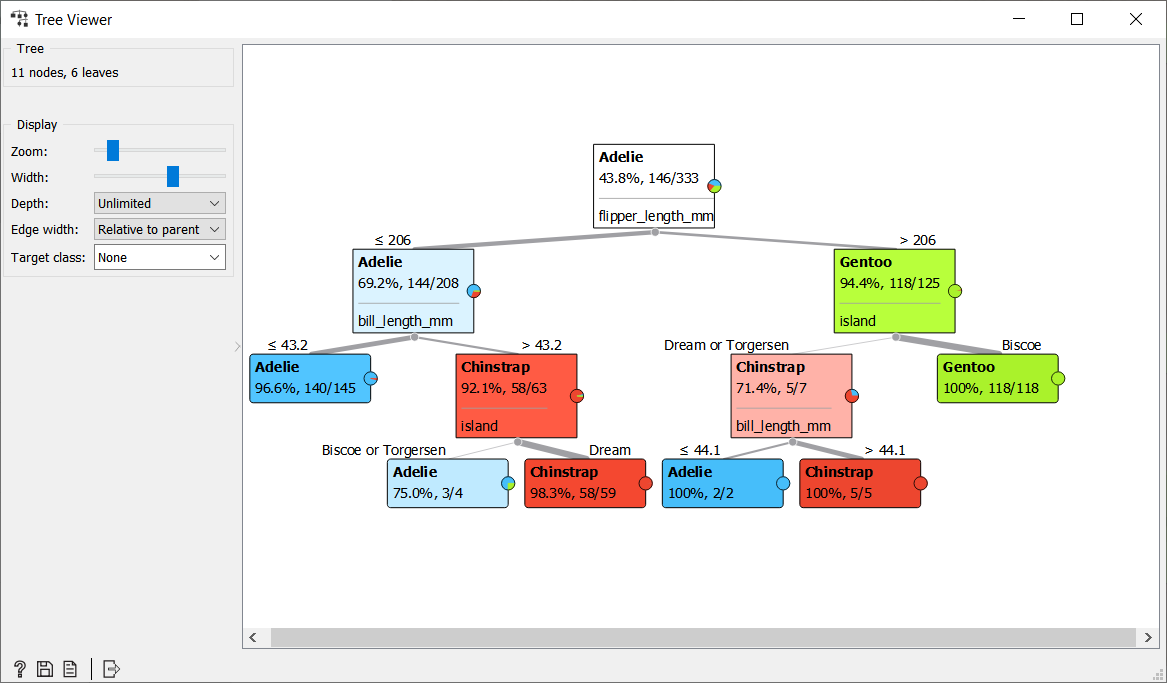


Connect like always:



Now in earlier versions tree was called Classification Tree and Tree Viewer was called Classification Tree Viewer. So, keep an eye out.



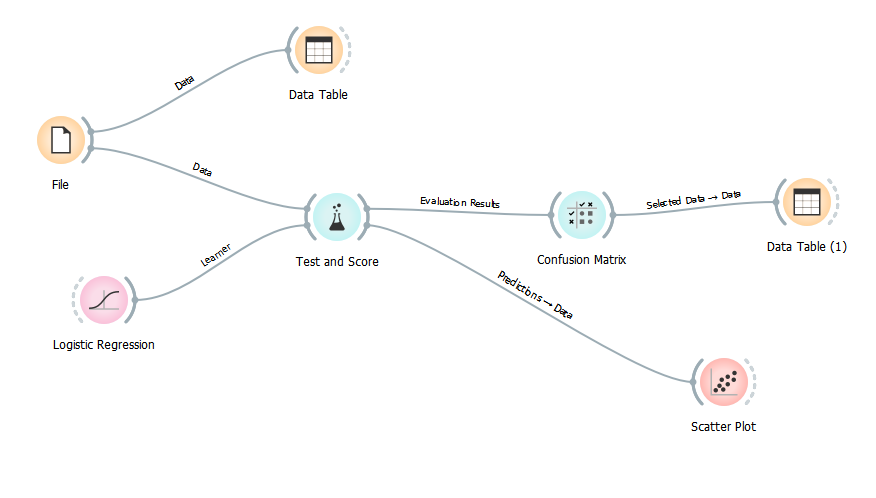


On the left side you are given options to set the depth. Look wise you can change the zoom, width.

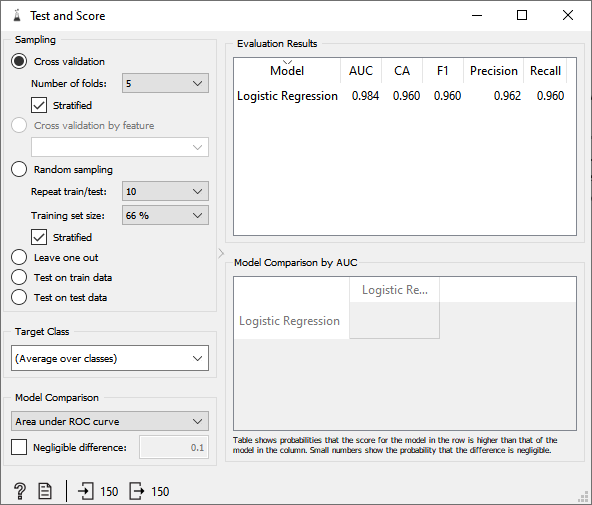
**Step 5: Cross-Validation:**

Once we are done with training our model, we just can’t assume that it is going to work well on data that it has not seen before. In other words, we can’t be sure that the model will have the desired accuracy and variance in production environment. We need some kind of assurance of the accuracy of the predictions that our model is putting out. For this, we need to validate our model. This process of deciding whether the numerical results quantifying hypothesised relationships between variables, are acceptable as descriptions of the data, is known as validation.

Let’s use an Iris dataset that is readily available.



To avoid overfitting, we build the model on training data (Test and Score). We are training a model using one of the simplest classification techniques Logistic Regression.

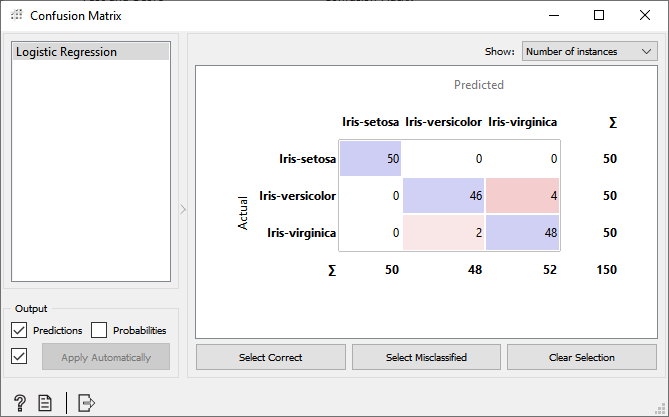


We want to repeat this technique multiple times and return the average. This is what cross validation does.

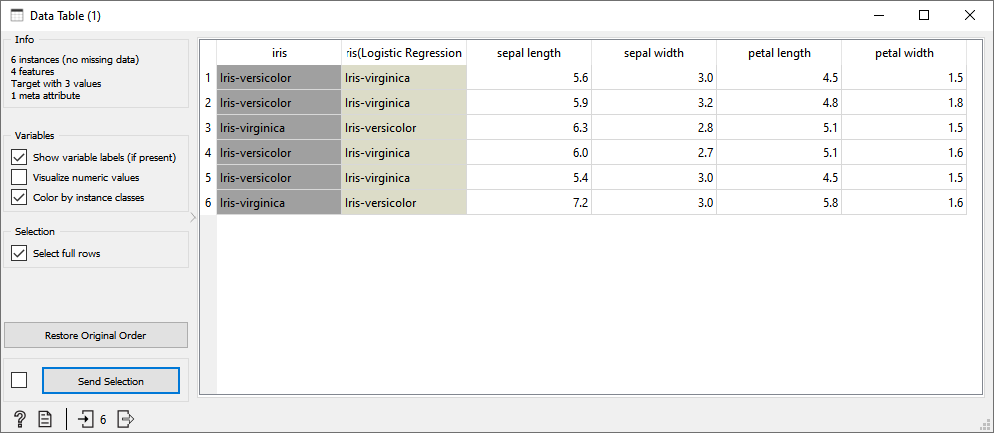
You can set the number of folds or repeats.

CA: Classification Accuracy

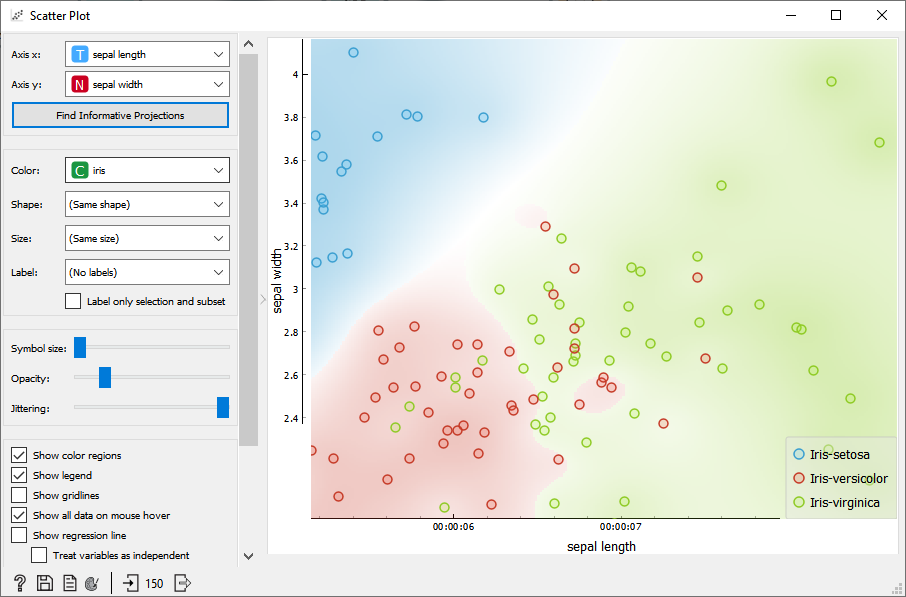
The classification accuracy here is 96%. The rest misclassified data can be checked through confusion matrix.



We can select the misclassified and view them in a separate data table.



Just by looking at this table it is still difficult to interpret the results. Instead let’s visualize them in a scatter plot.

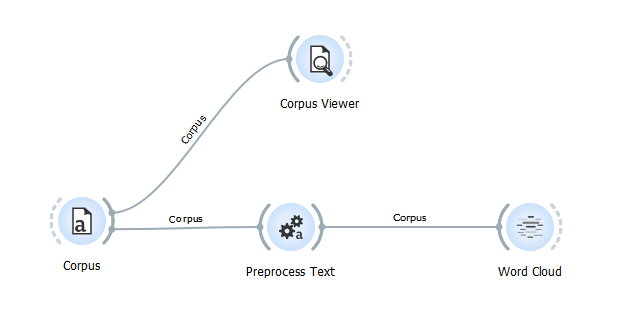


Observing this you can see some green values in the red while some red in the green. This is what gives erroneous values.

**Step 6: Text-Pre-processing:**

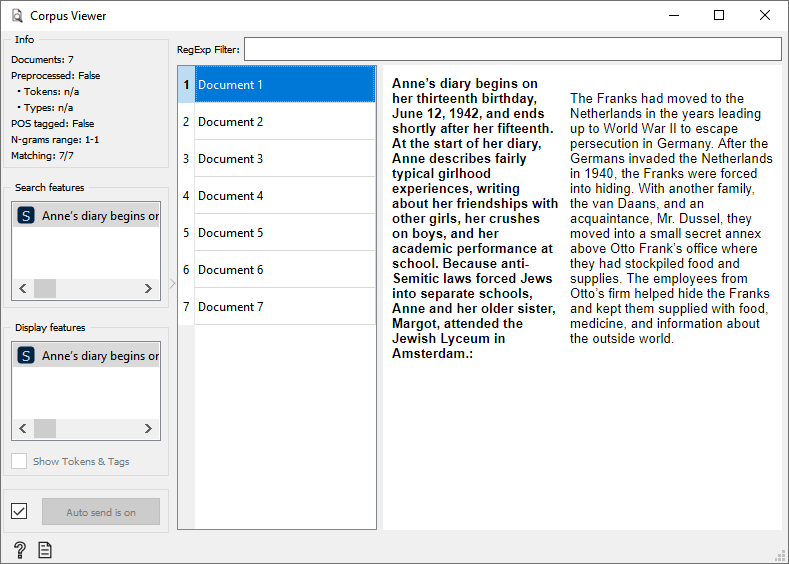
Using Orange to extract meaningful data from documents.

For this example, I will be using a summery of the book “The Diary of a Young Girl”.



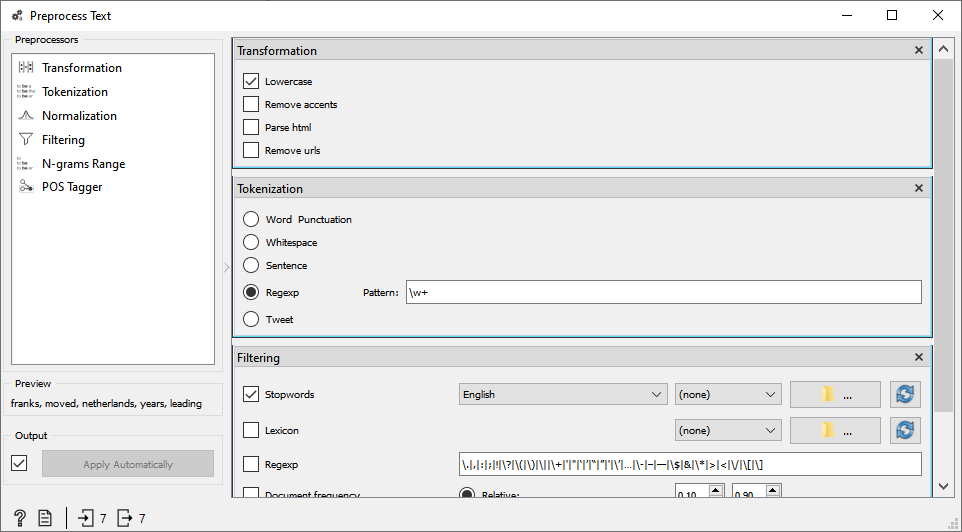
Use Corpus to add you text file. If you have none, just create on in notepad and rename its extension to .tab

Corpus viewer would be used to view your data document.

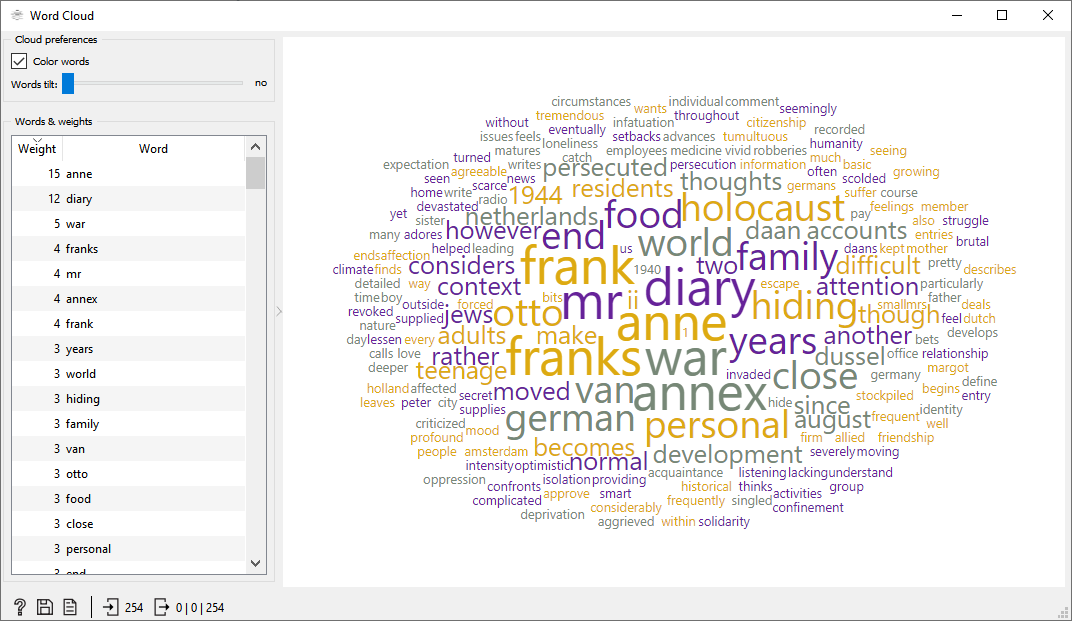


The text in bold is an actual summery in around 50 to 100 words auto generated by Orange. How cool is that?!

Another Widget for visualizing the text is Word Cloud. It displays word Frequencies in a cloud. But directly processing the word cloud through the corpus shows silly things like punctuations. So let us pre-process the data.



You can convert all text to lowercase, separate words by punctations, remove punctions and do much more.



Now we can remove worlds like a, an, the. The words that are not meaningful for finding out much needed information from the data. We can create a file where we write the custom words we need to omit and pass it into Orange. There are endless things you can do.

**PROS AND CONS OF ORANGE:**

**PROS:**

Benefits of Orange include its rich visualizations, interactive models, and speed. Often you can use Orange to get a quick look at what the data looks like or cross validate a model's performance. interactive K-Means and Gradient Descent widgets are great tools for understanding what is happening with these algorithms under the hood.

**CONS:**

One of the drawbacks is that it is not enormously robust for working with large datasets. Datasets which may work well in Python may make Orange crash. This is why it is best suited for smaller projects, pedagogical purposes, or exploratory data analysis.

**QUESTION AND ANSWERS:**

**What is orange tool?**

Orange is an open-source data visualization, machine learning and data mining toolkit. It features a visual programming front-end for explorative rapid qualitative data analysis and interactive data visualization.

**What types of data does Orange support?**

Core Orange supports Excel, comma- and tab-delimited files (.xlsx, .csv, .tab). It also reads online data, such as Google Spreadsheets. SQL widget supports PostgreSQL and MSSQL databases.

Add-ons can load additional formats. For example, Orange3-ImageAnalytics add-on can import images (.jpg, .tiff, .png) and Orange3-Text add-on can import text files (.txt, .docx, .pdf).

**How safe is my data?**

Orange does not store any data. It is a locally installed software that can be used without internet connection. The only exceptions are embedding widgets, which send data to the server, compute and return the result. The data is never stored on the server.

**What are Orange Widgets?**

Orange widgets give us a graphical user interface to orange's data mining and machine learning techniques. They incorporate widgets for data entry and pre-processing, classification, regression, association rules and clustering a set of widgets for model assessment and visualization of assessment results, and widgets for exporting the models into PMML.

**What is Orange Scripting?**

If we want to access Orange objects, then we need to write our components and design our test schemes and machine learning applications through the script. Orange interfaces to Python, a model simple to use a scripting language with clear and powerful syntax and a broad set of additional libraries. Same as any scripting language, Python can be used to test a few ideas mutually or to develop more detailed scripts and programs.

* **Conclusion**

From this practical we have learnt how orange tool works and perform data mining operations like basic operations [viewing data, charts [Histogram, Boxplot Analysis- Visualization] association rule mining, classification, Generating Classification Tree -Decision Trees, Cross Validation, text mining.

|  |  |  |  |
| --- | --- | --- | --- |
| **Prepared By:** | Rajiv Kumar Gupta (18CE137) | **Date:** | 04-08-2021 |