

## PRACTICAL 9

### Case Study: To Study Orange tool for Data Mining

#### Why do we need Data Mining?

Nowadays there is a huge amount of data available everywhere, the only concern is how to utilize this data to generate some sort of knowledge that can be used for decision making. Data mining is the answer.

Data mining discovers hidden patterns of the already available data, extracts knowledge from that data and establishes relationships to solve problems through data analysis. Different techniques of data mining can help companies increase their effectiveness and profit.

#### Orange (tool)


Orange is component-based visual programming software for data mining, machine learning, and data analysis. Workflows are created by linking predefined or user-designed components called widgets. They read the data, process it, visualize it, do clustering, build predictive models and so on.


#### Predictions in Orange


Prediction is one of the most used techniques in Data mining. Orange has a predefined widget for making predictions.

We'll show a simple example where we use two files. But first, we will explain some terms here:

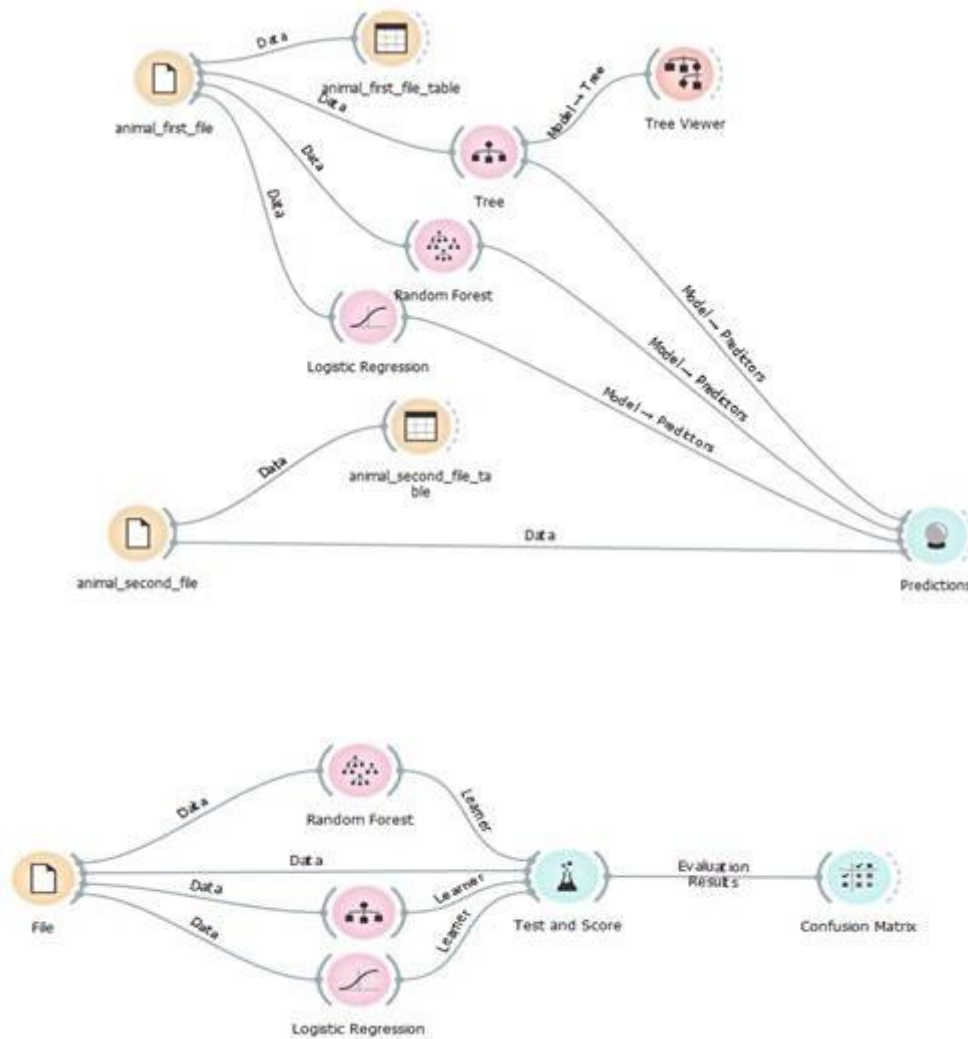
- Meta variables are metadata, data about data, not used for statistical inference.
- Features or variables or attributes are the measured inputs of the problem domain, the independent variables.
- The target variable is the dependent variable or the measure we're trying to model or forecast.

 In the first file, there are 82 instances (animals) with 16 features specified. The type will be the *target* role and the name will be *meta* role. In the second file, there are 4 instances with 16 features specified, with unknown name and type of animal.

 The purpose of this workflow is to make a prediction on the animal types of the second file. To make the prediction we'll need two data sets: the training data, that is loaded from the first Datasets widget (animal\_first\_file) and the data to predict availability in the second Dataset widget (animal\_second\_file).

 This workflow starts with the File widget. The data from the animal\_first\_file is sent to Data table, Tree, Random Forest and Logistic regression widgets. The second file which contains the test set is connected with Predictions and Data table widget.

o Note: when working with machine learning algorithms, we should have in mind that the data should be analyzed, preprocessed and cleaned before we use it in the algorithm. In our example, we have a simple data set and the data is ready for the predictions.



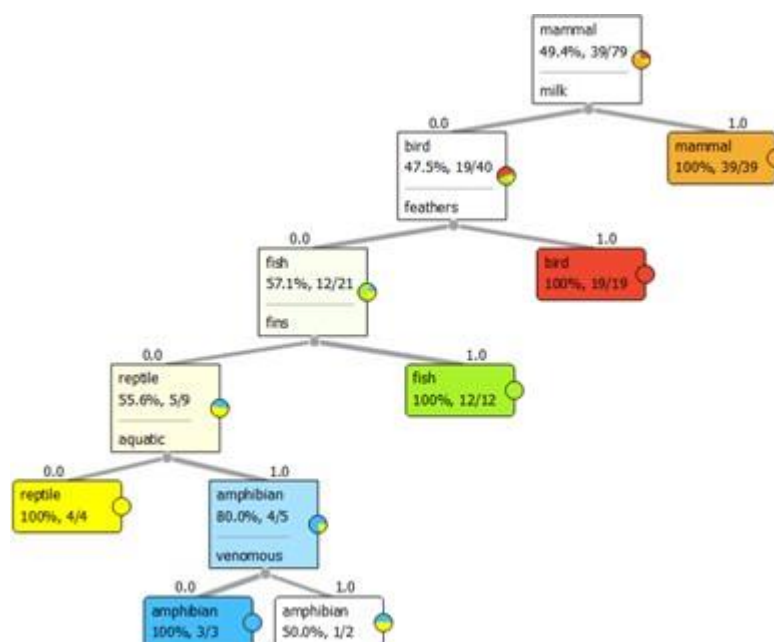
Test and score widget has values for classification accuracy (CA) that give us a proportion of correctly classified data instances. We can conclude that in this case Random Forest is less suitable than the other classification models. Also, we can see that Tree classifier and Logistic regression give us approximate accuracy and precision.

| Model               | AUC   | CA    | F1    | Precision | Recall |
|---------------------|-------|-------|-------|-----------|--------|
| Tree                | 0.980 | 0.975 | 0.974 | 0.976     | 0.975  |
| Random Forest       | 0.995 | 0.949 | 0.946 | 0.951     | 0.949  |
| Logistic Regression | 0.996 | 0.975 | 0.972 | 0.976     | 0.975  |

As we have classification accuracy around 97%, 3% of the data, we can say that the instances are misclassified. Using the Confusion Matrix, we can observe how many instances were misclassified and in which way. Figure 4 shows the test result from the Tree classification model.

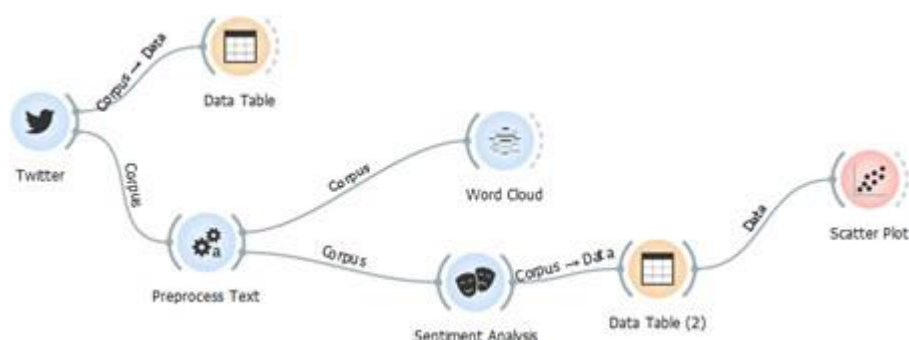
|          |           | Predicted |      |      |        |         |          |
|----------|-----------|-----------|------|------|--------|---------|----------|
|          |           | amphibian | bird | fish | mammal | reptile | $\Sigma$ |
| Actual   | amphibian | 3         | 0    | 0    | 0      | 1       | 4        |
|          | bird      | 0         | 19   | 0    | 0      | 0       | 19       |
|          | fish      | 0         | 0    | 12   | 0      | 0       | 12       |
|          | mammal    | 0         | 0    | 0    | 39     | 0       | 39       |
|          | reptile   | 0         | 0    | 1    | 0      | 4       | 5        |
| $\Sigma$ |           | 3         | 19   | 13   | 39     | 5       | 79       |

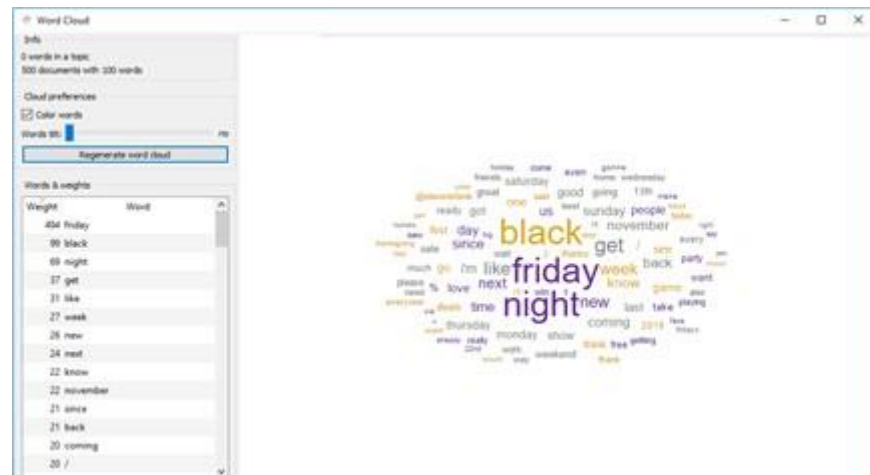
Tree classifier is a simple algorithm that splits the data into nodes by class purity.



## Text mining in Orange

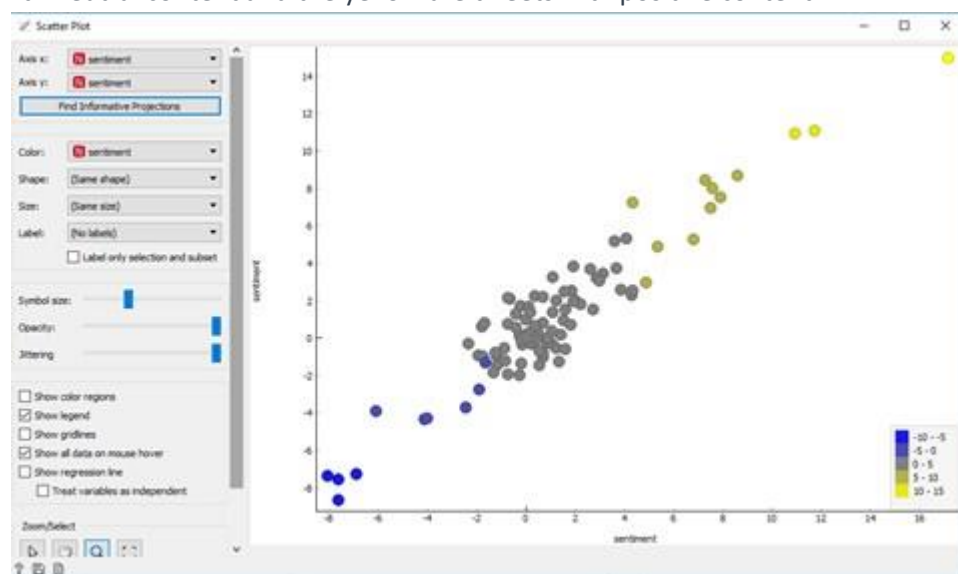
- Orange has an option called add-ons, where several packages are stored. In the following example, we have used the package for Text mining. After installing the new package there is an additional widget set. Twitter widget enables querying tweets through Twitter API and allows queries for up to two weeks back.





One of the most interesting analysis that can be done using Orange is Sentiment Analysis. There are two types of sentiment modules available here: Liu Hu and Vader. Liu Hu computes a single normalized score of sentiment in the text (negative score for negative sentiment, positive for positive, 0 is neutral), while Vader outputs scores for each category (positive, negative, neutral) and appends a total sentiment score called a compound.

We want to know if we have positive or negative emotions in the tweets that we got by using the widget. In this example, we can use Liu Hu module. This analysis represents negative numbers for tweets with negative words, zero for neutral tweets and positive numbers for tweets with a positive sentiment. The scatter plot widget graphically shows the Sentiment Analysis. In the plot, there is a legend where we can see the values from the analysis for all the tweets that were found before. The blue dots are tweets with negative content, the gray dots are with neutral content and the yellow are tweets with positive content.



## Conclusion

- Data mining is used to build prediction models based on historical data. They can help in making decisions and predict future trends. Orange is a very helpful tool for data visualization and analyzing bigdata sets. It is open-source software that allows trying different algorithms and supports visual programming tools for Data mining. Moreover, after performing practical implementation Orange has done everything as its feature said. This tool makes analysis work easier.

## Practical-10

**Aim:** Explore various tools available for Data Mining. Explain any two tools of your choice in detail.

### Data Mining tools

- Data Mining is the set of techniques that utilize specific algorithms, statistical analysis, artificial intelligence, and database systems to analyze data from different dimensions and perspectives.



- Data Mining tools have the objective of discovering patterns/trends/groupings among large sets of data and transforming data into more refined information.
- It is a framework, such as Rstudio or Tableau that allows you to perform different types of data mining analysis.
- We can perform various algorithms such as clustering or classification on your data set and visualize the results itself. It is a framework that provides us better insights for our data and the phenomenon that data represent. Such a framework is called a data mining tool.
- The Market for Data Mining tool is shining: as per the latest report from ReortLinker noted that the market would top **\$1 billion** in sales by **2023**, up from **\$ 591 million** in **2018**

These are the most popular data mining tools:



## 1. SAS Data Mining:



- ✚ SAS stands for Statistical Analysis System. It is a product of the SAS Institute created for analytics and data management. SAS can mine data, change it, manage information from various sources, and analyze statistics. It offers a graphical UI for non-technical users.
- ✚ SAS data miner allows users to analyze big data and provide accurate insight for timely decision-making purposes. SAS has distributed memory processing architecture that is highly scalable. It is suitable for data mining, optimization, and text mining purposes.
- ✚ SAS Enterprise Miner is an advanced analytics data mining tool intended to help users quickly develop descriptive and [predictive models](#) through a streamlined data mining process.
- ✚ Enterprise Miner's graphical interface enables users to logically move through the five-step SAS SEMMA approach: sampling, exploration, modification, modeling and assessment. Users can build a process flow by selecting the appropriate tab from Enterprise Miner's toolbar, and then dragging and dropping step-specific nodes onto a pallet.
- ✚ Enterprise Miner supports several algorithms and techniques, including [decision trees](#), time series, neural networks, linear and [logistic regression](#), sequence and web path analysis, market basket analysis, and link analysis.
- ✚ Enterprise Miner's [client-server architecture](#) enables business users and data analysts to collaborate and share models and other work.



## Enterprise Miner features for the data mining process

- **SAS Rapid Predictive Modeler** is a component of SAS Enterprise Miner that can run as an add-on to Microsoft Excel, enabling business users to perform predictive modeling directly from within their Excel spreadsheets. Models developed in Rapid Predictive Modeler can be customized by data analysts using Enterprise Miner.
- **Integration of R code.** Analysts and developers who develop in [the R language](#) can integrate the models and transformations they write within an Enterprise Miner process flow.
- **Support for in-database and in-Hadoop scoring.** When combined with a SAS Scoring Accelerator, scoring algorithms created in SAS Enterprise Miner can be deployed and executed within a database or [Hadoop environment](#). Scoring Accelerators are available for Hadoop, Pivotal, DB2, IBM Netezza, Oracle, Teradata and SAS Scalable Performance Data Server.
- SAS also offers Factory Miner, an add-on product that provides users with an automated, web-based framework to help them reduce the time needed to develop models. The product enables users to build, run and retrain multiple predictive models across business or customer segments quickly. It can also help [identify a champion model](#).
- Release 14.2 improved upon the previous version of Enterprise Miner by adding new nodes to execute code in a SAS Viya environment -- an open, cloud-ready, in-memory platform -- as part of a process flow diagram. Other enhancements include improvements to the Score node and Score Code Export nodes.
- Enhancements in Factory Miner 14.2 enable users to create batch code that can be used for retraining models with updated data.
- Contact SAS for pricing. Although there is no trial version of Enterprise Miner, SAS offers members of academia free use of its products through web access via its SAS OnDemand for Academics offering.

## 2. Rapid Miner:



- ✚ Rapid Miner is one of the most popular predictive analysis systems created by the company with the same name as the Rapid Miner. It is written in JAVA programming language. It offers an integrated environment for text mining, deep learning, machine learning, and predictive analysis.
- ✚ The instrument can be used for a wide range of applications, including company applications, commercial applications, research, education, training, application development, machine learning.
- ✚ Rapid Miner provides the server on-site as well as in public or private cloud infrastructure. It has a client/server model as its base. A rapid miner comes with template-based frameworks that enable fast delivery with few errors (which are commonly expected in the manual coding writing process).
- ✚ RapidMiner is a free of charge, open source software tool for data and text mining. In addition to Windows operating systems, RapidMiner also supports Macintosh, Linux, and Unix systems. It is available as a stand-alone application for data/text analysis and as a data/text mining engine for the integration into your own products. Thousands of applications of RapidMiner in more than 40 countries are successfully developed to give its users a competitive edge.
- ✚ The RapidMiner software tool, along with its extensions (including text analytics extension) and documentation, can be found and downloaded from [www.rapid-i.com](http://www.rapid-i.com). Once the proper version of the tool is downloaded and installed, it can be used for a variety of data and text mining projects.  
Its graphical user interface is a little different from the ones we often see in other commercial data mining tools, such as IBM SPSS Modeler, SAS Enterprise Miner, and *STATISTICA* Data Miner. Such differences may lead to a longer learning curve, but once understood it is quite logical and informative.
- ✚ RapidMiner provides data mining and machine learning procedures including: **data loading and transformation (ETL)**, data preprocessing and visualization, predictive analytics and statistical modeling, evaluation, and deployment.