PES Institute of Technology

A data mining tool

Mohit Surana 1PI13CS092
Shiva K Deviah 1PI13CS147
Shrey Agarwal 1PI13CS150

What is Orange?

Orange is a component-based **data mining** and **machine learning** software suite. It contains

- **Visual frontend UI** for data analysis and visualization
- Python bindings and libraries for scripting

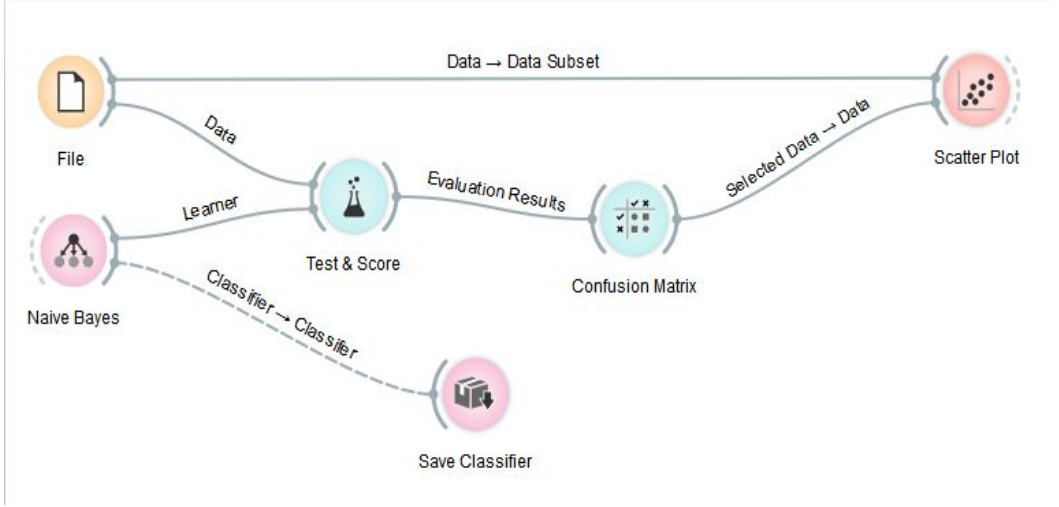
It is implemented in C++ and Python.

Workflows and Widgets

Workflows specify the sequence of actions to be performed on data are created at the visual frontend. The basic unit of the workflow is the **widget**.

Widgets can be used to do

- data preprocessing
- feature scoring and filtering
- modeling
- model evaluation
- exploration techniques



An example workflow

Advantages of Orange

- Open source
- Easy to learn
- Well integrated with Python, Numpy, and C
- Works both as a script and with an ETL (Extract, Transform, Load) workflow GUI
- Great at handling errors and debugging

Pitfalls of Orange

- No inbuilt support for neural networks
- ML algorithms limited, not handled uniformly between the different libraries
- Orange 2 and 3 do not have the exact same set of features, and there are compatibility issues b/w them
- Weak in classical statistics; provides no widgets for statistical testing
- Limited reporting capabilities, constrained to exporting visual representations of data models only
- Additional overhead in terms of performance issues

Plan of Action

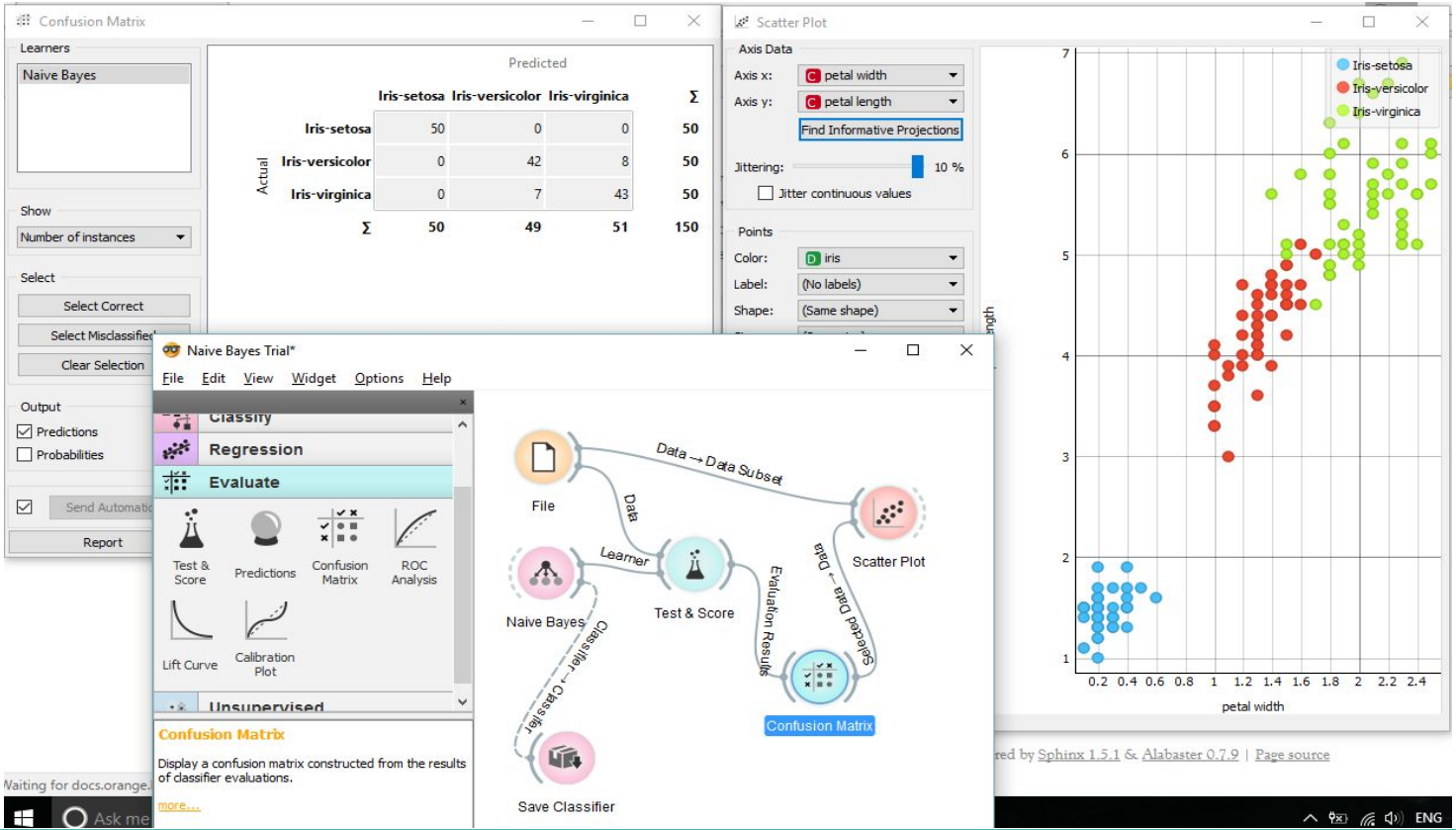
Orange makes it very easy to

- collect data using its web crawling widgets
- preprocess data

In addition, Orange also has add-ons that integrate with scikit-learn and the nltk library that we can use to design NLP related logic

Since we'd like to use neural networks, and since orange does not have widgets to directly support this, we can write python code with the keras framework and then neatly integrate it with Orange using a Python script widget.

DEMO: NBC for the Iris Dataset



Related Work (1)

We referred a research paper for a thorough comparison between the software titled "**Comparative Study of Data Mining Tools**" by Kalpana Rangra and Dr. K. L. Bansal

According to the paper:

- **KNIME** is recommended for novices
- **Weka** is very similar to KNIME and has many built-in features requiring no programming knowledge
- **RapidMiner** and **Orange** are appropriate for advanced users because of the additional programming skills needed, and the visualization support that it provides

Related Work (2)

Table 1 : Technical Overview of best six data mining open source tools

S. N	Tool Name	Release Date	Release date/ Latest version	License	Operating System	Language	Website
1.	RAPID MINER	2006	21November,2013 /6.0	AGPL Proprietary	Cross platform	Language Independent	www.rapidminer.com
2	ORANGE	2009	6 May,2013/2.7	GNU General Public License	Cross Platform	Python C++,C	www.orange.biolab.si
3	KNIME	2004	6December,2013/2.9	GNU General Public License	Linux ,OS X, Windows	Java	www.knime.org
4	WEKA	1993	24 April,2014/3.7.11	GNU General Public License	Cross Platform	Java	www.cs.waikato.ac.nz/~ml/weka
5	KEEL	2004	5 June,2010/2.0	GNU GPL v3	Cross Platform	Java	www.sci2s.ugr.es/keel
6	R	1997	10 April,2014/3.1.0	GNU General Public License	Cross Platform	C, Fortran and R	www.r-project.org

Next Steps

1. Formulate a concrete problem statement for clarity
2. Pitch our idea to the project panel for approval
3. Prepare a project timeline
4. Research past work done in this domain
    - “Application of Bloom’s Taxonomy in day to day Examinations” by Bhargav H S, Gangadhar Akalwadi and Nitin V Pujari
    - Istanbul paper
5. Get a lot of data
6. Figuring out how to implement a POC, what features are in and out, etc
7. Implementation of the POC with good accuracy