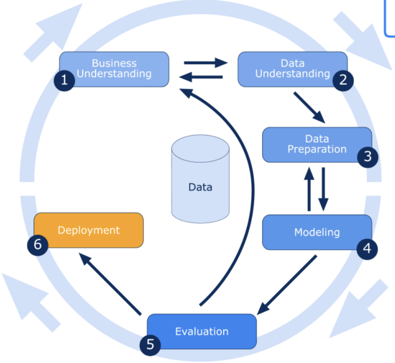
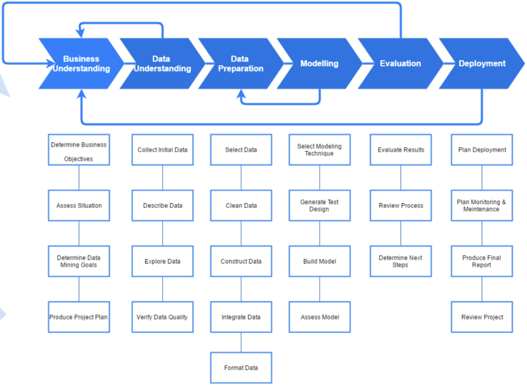
AIML430/COMP309: ML Tools and Techniques
Tutorial Week 4: Data Mining Tools & Orange

Ali Knott
School of Engineering and Computer Science, VUW

VICTORIA UNIVERSITY OF WELLINGTON TE HERENGA WAKA

Recap: CRISP-DM





Data mining tools

Python has a number of modules for data mining.

R is a free software environment for statistical computing and graphics.

- Written by Ross Ihaka and Robert Gentleman, Auckland Uni…
- Again, some good data mining functionality.

Weka is a free machine learning toolkit.

- Produced at Waikato! Written in Java…
- It provides some algorithms for data mining tasks.

Orange is a free, open source data mining software tool.

- It's a visual programming package, but integrates well with Python.
- Good for quick visualisations, quick analyses.

Orange is a good thing to know about!

- The visual interface is neat—a ‘no-programming’ style of working
- It’s easy to try different machine learning algorithms
- There are lots of add-ons. (E.g. for bioinformatics, network analysis, text mining)
- Free, open-source…
- You can download at `https://orangedatamining.com/`
- There’s good documentation there too. (Incl. good intro videos.)

Orange is one of the options for Assignment 2!

Assignment 2: ‘Data Exploration, Manipulation and Modelling’.

- Worth 15%, out tomorrow…
- Due 16th August (Friday Week 6) 23:59.

There’s a focus on these topics:

- Cross Industry Standard Process for Data Mining (CRISP-DM)
- Exploratory Data Analysis (EDA)
- Data Preparation
- Feature Manipulation

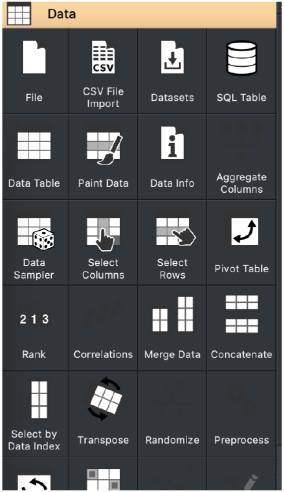
You can use Python for all this, or Orange, or a mixture.

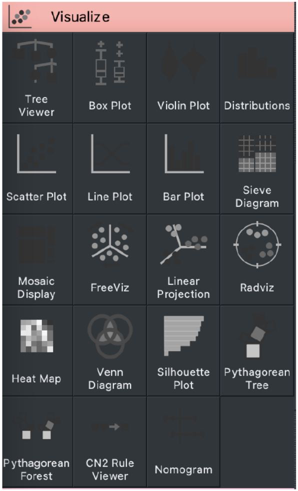
- ‘Research into online resources for AI & ML is encouraged.’

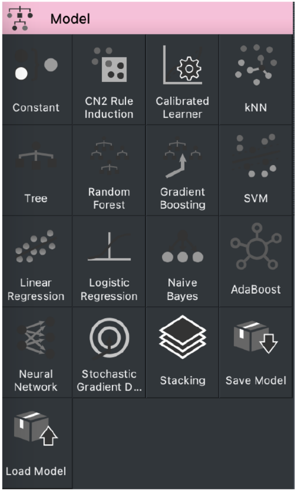
An introduction to Orange

Widgets in Orange

In Orange, workflows are created by dragging widgets onto a canvas, and linking them with connections.



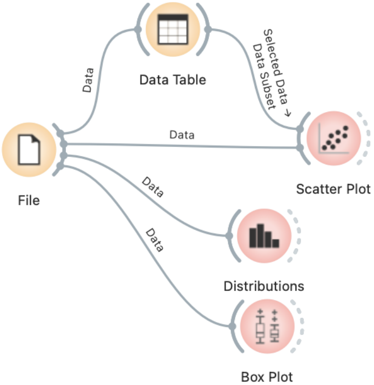




Some visualisation widgets

File → Iris dataset →

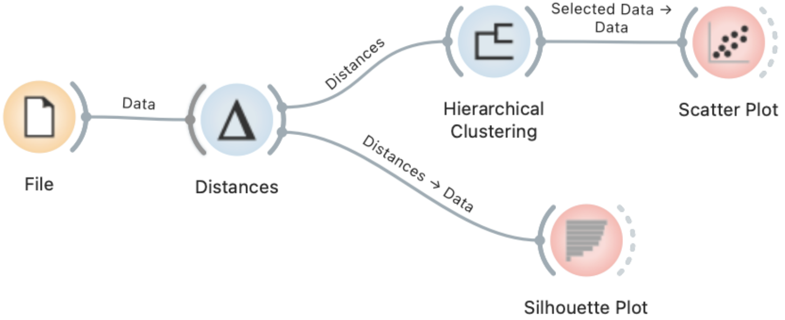
- Data table
- Scatter plot → Informative projections
- Data table → Scatter plot—connection options
- Distributions
- Box plot



Some clustering widgets

File → Iris dataset →

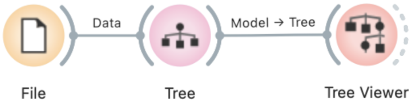
- Distances → Hierarchical clustering (→ Scatter plot)
  → Sillhouette plot



Some classification widgets

File → Iris dataset →

- → Tree (classification tree) → Tree viewer

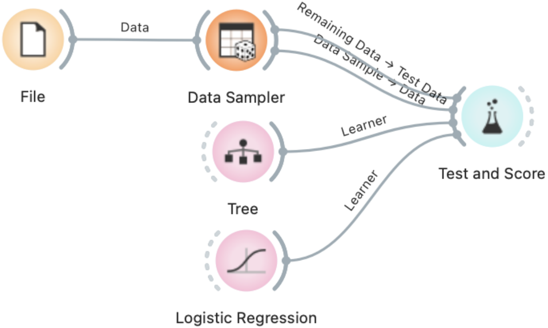


Some classification widgets

For a simple train/test data split:

File → Iris dataset → Test and score

- Tree →
  Logistic regression →

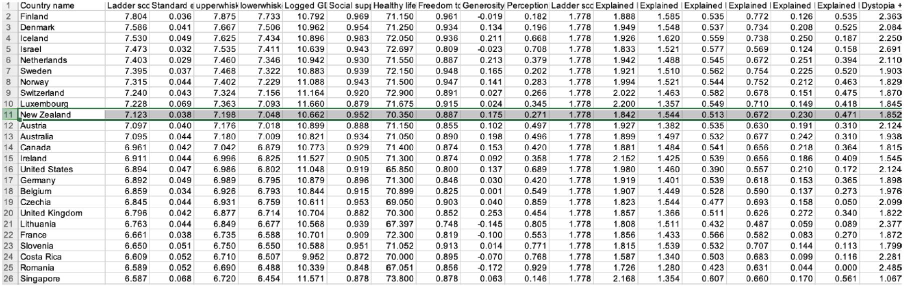


For a more complex train/validation/test set split, you need to create files manually…

Another domain example: World Happiness Report

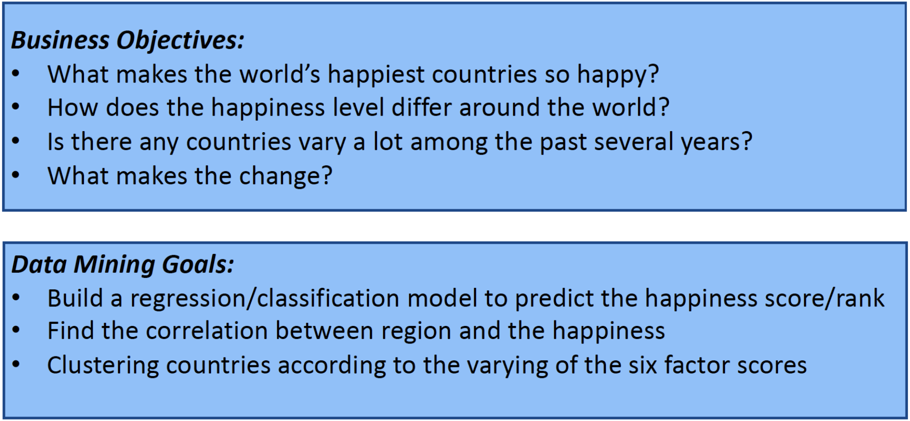
The 2022 World Happiness Report ranks 137 countries by their ‘happiness level’.

- Each country is scored on six criteria: economics, social support, health, freedom, trust, and generosity.



1	Country name	Ladder scc	Standard e	upperwhisk	lowerwhisk	Logged GI	Social sup	Healthy life	Freedom tc	Generosity	Perception	Ladder scc	Explained	Explained	Explained	Explained	Explained	Explained	Dystopia +
2	Finland	7.804	0.036	7.875	7.733	10.792	0.969	71.150	0.961	-0.019	0.182	1.778	1.888	1.585	0.535	0.772	0.126	0.535	2.363
3	Denmark	7.586	0.041	7.667	7.506	10.962	0.954	71.250	0.934	0.134	0.196	1.778	1.949	1.548	0.537	0.734	0.208	0.525	2.084
4	Iceland	7.530	0.049	7.625	7.434	10.896	0.983	72.050	0.936	0.211	0.668	1.778	1.926	1.620	0.559	0.738	0.250	0.187	2.250
5	Israel	7.473	0.032	7.535	7.411	10.639	0.943	72.697	0.809	-0.023	0.708	1.778	1.833	1.521	0.577	0.569	0.124	0.158	2.691
6	Netherlands	7.403	0.029	7.460	7.346	10.942	0.930	71.550	0.887	0.213	0.379	1.778	1.942	1.488	0.545	0.672	0.251	0.394	2.110
7	Sweden	7.395	0.037	7.468	7.322	10.883	0.939	72.150	0.948	0.165	0.202	1.778	1.921	1.510	0.562	0.754	0.225	0.520	1.903
8	Norway	7.315	0.044	7.402	7.229	11.088	0.943	71.500	0.947	0.141	0.283	1.778	1.994	1.521	0.544	0.752	0.212	0.463	1.829
9	Switzerland	7.240	0.043	7.324	7.156	11.164	0.920	72.900	0.891	0.027	0.266	1.778	2.022	1.463	0.582	0.678	0.151	0.475	1.870
10	Luxembourg	7.228	0.069	7.363	7.093	11.660	0.879	71.675	0.915	0.024	0.345	1.778	2.200	1.357	0.549	0.710	0.149	0.418	1.845
11	New Zealand	7.123	0.038	7.198	7.048	10.662	0.952	70.350	0.887	0.175	0.271	1.778	1.842	1.544	0.513	0.672	0.230	0.471	1.852
12	Austria	7.097	0.040	7.176	7.018	10.899	0.888	71.150	0.855	0.102	0.497	1.778	1.927	1.382	0.535	0.630	0.191	0.310	2.124
13	Australia	7.095	0.044	7.180	7.009	10.821	0.934	71.050	0.890	0.198	0.496	1.778	1.899	1.497	0.532	0.677	0.242	0.310	1.938
14	Canada	6.961	0.042	7.042	6.879	10.773	0.929	71.400	0.874	0.153	0.420	1.778	1.881	1.484	0.541	0.656	0.218	0.364	1.815
15	Ireland	6.911	0.044	6.996	6.825	11.527	0.905	71.300	0.874	0.092	0.358	1.778	2.152	1.425	0.539	0.656	0.186	0.409	1.545
16	United States	6.894	0.047	6.986	6.802	11.048	0.919	65.850	0.800	0.137	0.689	1.778	1.980	1.460	0.390	0.557	0.210	0.172	2.124
17	Germany	6.892	0.049	6.989	6.795	10.879	0.896	71.300	0.846	0.030	0.420	1.778	1.919	1.401	0.539	0.618	0.153	0.365	1.898
18	Belgium	6.859	0.034	6.926	6.793	10.844	0.915	70.899	0.825	0.001	0.549	1.778	1.907	1.449	0.528	0.590	0.137	0.273	1.976
19	Czechia	6.845	0.044	6.931	6.759	10.611	0.953	69.050	0.903	0.040	0.859	1.778	1.823	1.544	0.477	0.693	0.158	0.050	2.099
20	United Kingdom	6.796	0.042	6.877	6.714	10.704	0.882	70.300	0.852	0.253	0.454	1.778	1.857	1.366	0.511	0.626	0.272	0.340	1.822
21	Lithuania	6.763	0.044	6.849	6.677	10.568	0.939	67.397	0.748	-0.145	0.805	1.778	1.808	1.511	0.432	0.487	0.059	0.089	2.377
22	France	6.661	0.038	6.735	6.588	10.701	0.909	72.300	0.819	-0.100	0.553	1.778	1.856	1.433	0.566	0.582	0.083	0.270	1.872
23	Slovenia	6.650	0.051	6.750	6.550	10.588	0.951	71.052	0.913	0.014	0.771	1.778	1.815	1.539	0.532	0.707	0.144	0.113	1.799
24	Costa Rica	6.609	0.052	6.710	6.507	9.952	0.872	70.000	0.895	-0.070	0.768	1.778	1.587	1.340	0.503	0.683	0.099	0.116	2.281
25	Romania	6.589	0.052	6.690	6.488	10.339	0.848	67.051	0.856	-0.172	0.929	1.778	1.726	1.280	0.423	0.631	0.044	0.000	2.485
26	Singapore	6.587	0.068	6.720	6.454	11.571	0.878	73.800	0.878	0.063	0.146	1.778	2.168	1.354	0.607	0.660	0.170	0.561	1.067

Business understanding



Business Objectives:

- What makes the world's happiest countries so happy?
- How does the happiness level differ around the world?
- Is there any countries vary a lot among the past several years?
- What makes the change?

Data Mining Goals:

- Build a regression/classification model to predict the happiness score/rank
- Find the correlation between region and the happiness
- Clustering countries according to the varying of the six factor scores

Orange has an ‘add-on’ for ‘World Happiness’. . .

(. . . You can find it under ‘options’.)

The widget lets us choose which countries and years to look at.

The database has lots of features.

- Note how much *missing data* there is!
- We can choose features of interest with the ‘Select Columns’ widget.

	EXP.GDP.PER.CAPITA	HAP.SCORE
Country 1		
Country 2		
Country *n*		

Exploratory data analysis

We might try some clustering…

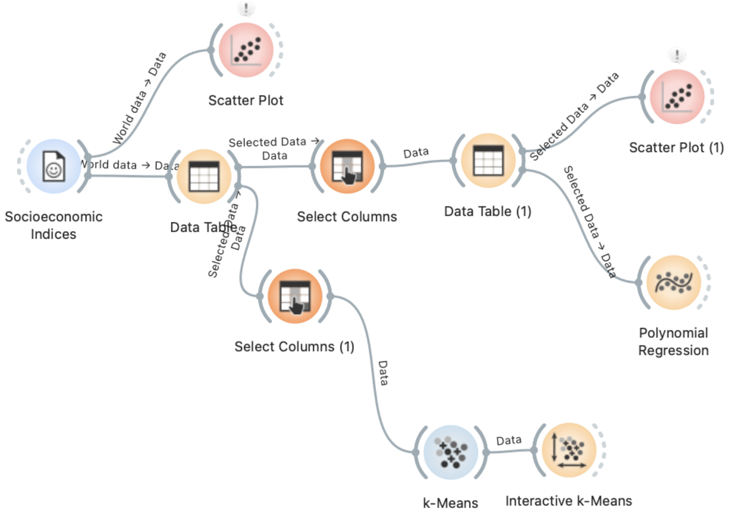
- There's a useful group of 'Educational' add-on widgets…
  - 'Interactive *k*-means' gives you a nice view of the clustering process.

Exploratory data analysis

We might try building some predictive models. . .

- For this, it's natural to set 'HAP.SCORE' to be the target feature. . .
- Then we can build various predictive models. . .
  - 'Polynomial regression' (also from Educational add-ons) is a good one to try.

My Orange happiness analyses. . .



Summary

Orange is a nifty data mining / analytics / ML tool!

Some good resouces for learning:

- Orange's own introductory videos (available from 'Help')
- The Web has a lot of other good material.

Some useful pointers:

- Remember that to create a training/validation/test data split, you need to manually create the relevant files.
- For our course, the 'Educational' add-on widgets are quite helpful.