**Data Analysis and Machine Learning for Beginners and Non-Programmers**

**Introduction and Exploring the Interface**

1. Download and install Orange (orange.biolab.si).
   1. Explore the interface while others download the software
2. Explore & Add widgets:
   1. Click from widget box; drag and drop; right click in working environment

**Importing and Viewing Data**

1. File widget:
   1. Load the Iris dataset
2. Data Table:
   1. Explore the data in the data table
   2. Data can be excel, **tab** (Orange native dataset), csv, etc.
   3. These have 4 features (sepal W&L; Petal W&L), Class is 3 species
3. Select Rows/Columns:
   1. Change variables or create subsets
   2. Save new data format/subset using “save data”
4. Distributions (Bar charts/Histograms)
   1. Look at the different variables and how the separate the three species of Iris
5. Scatter Plot:
   1. The default is a little messy
   2. Explore the “rank projections” to find the best scatterplot to separate the three species
      1. Still there is some overlapping regions
      2. Select an overlapping region and lets look at this in a new data table (new widget)
   3. Create a data browser using subset from selecting some data from the “data table” then send this to “scatter plot” to explore data in the scatter plot window

**Distances and Clustering (Hierarchical Clustering and MDS)**

1. Using the “Distances” for measuring similarity (Euclidian distance only)
   1. Demonstrate how to calculate distance
      1. Two points
      2. Two series (one variable)
      3. Multiple variables
   2. Multi-Dimensional Scaling
      1. Select data same as from scatter plot example
   3. Hierarchical Clustering
      1. Default parameterization (Linkage = average; pruning = none)
         1. Selection
            1. Manual: Look for Versicolor and Virginica overlap

Connect to scatter plot and data table to look at selection

* + - * 1. Height Ratio (100%; 50%; 10%)
        2. Top number of groups (3; for 3 species)

Save data as top 3 clusters (Iris Clusters)

Load Iris Clusters in new “File”

Use “Color” then connect to “distributions”

Explore how the clusters and species differ

**Tree Model and Make Predictions**

1. Creating a Tree Model
   1. New “File” for Iris Dataset
   2. Add and connect data to “Tree”
      1. Keep defaults for today
   3. Connect “Tree” to Tree Viewer
   4. This allows us to see how the algorithm divided the data
      1. The model goes through one variable at a time separating into homogenous classes
2. Making predictions (Using 3 unknown Iris flowers)
   1. We will use our Tree classification model to predict the species of each
   2. Lets connect our unknown Iris to a “file” and look at the “data table”
      1. Can you guess which is which? What if there were 30? Or 30,000?
   3. Connect the unknown file to the “predictions” widget
      1. Open the predictions widget. Is there anything there?
         1. No, because we haven’t attached our model
      2. Let’s attach the Tree model to our predictions
      3. Look at the Predicted Class and Probabilities
         1. Our model is pretty great!
   4. Lets compare our Tree model with a Logistic Regression
      1. Add “Logistic Regression” and connect it to the original data
      2. Connect the “Logistic Regression” to our “Predictions” and compare
      3. Which is more confident?