Lab-1 [kNN]

Out date: Jun 15, 2022
Due date: Jun 19, 2022 (Sunday) at 11:59PM

Submission

1. Prepare your solution in Orange and save the workspace (e.g., Lab-1.ows). **[50 points]**
2. Complete the tables given below and save the file (e.g., Lab-1.docx). **[50 points]**
3. Upload the files to the Canvas.

Background information: Oil and gas reservoirs lie deep beneath the Earth's surface. Geologists and engineers cannot examine the rock formations in situ, so tools called sondes go there for them. Specialists lower these tools into a wellbore and obtain measurements of subsurface properties. The data are displayed as a series of measurements covering a depth range in a display called a well log. Often, several tools are run simultaneously as a logging string, and the combination of results is more informative than each individual measurement
(https://www.slb.com/resource-library/oilfield-review/defining-series/defining-logging).

Link below gives an overview of interpreting lithology using Gamma Ray, Density porosity and Neutron Porosity logs.
http://www.kgs.ku.edu/Publications/Bulletins/LA/05_overlay.html

LAS file (1033440835.las) containing Gamma Ray, Caliper, Density Porosity and Neutron Porosity for well Beck 'A' #1 that is used in the overview link above was downloaded from the link below.
https://chasm.kgs.ku.edu/ords/las.lasd5.SelectWells

Please refer to http://www.kgs.ku.edu/General/copyright.html regarding use of data / information from Kansas Geological Survey.

Objective: To train a Machine Learning model to predict lithology using log measurements in an oil & gas well. This is a classification problem. The algorithm used is kNN.

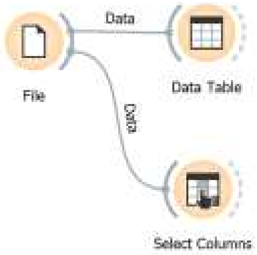
Data: Relevant log data was extracted to an excel file (***Log Lithology classification example.xlsx***) and lithology labels were created to be used for the hands-on lab.

Lab Instructions

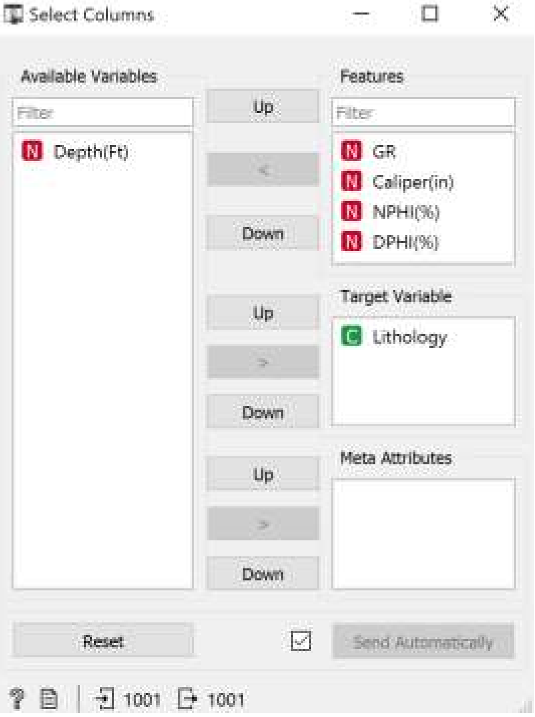
1. Download the 'Log Lithology classification example.xlsx' file to your working folder
2. Launch Orange.
3. Click on the **File** Widget under **Data** to add the widget to your blank Orange canvas.
4. Load the Excel data file by double clicking on the File widget. https://orange-visual-programming.readthedocs.io/loading-your-data/index.html
5. What do you see under Info? Specifically, complete the following table: (3 points)

How many features?	4
How many samples?	6
What is the target feature?	Lithology

6. Add the **Data Table** Widget from the left pane or by right clicking on the canvas.
7. Connect the **File** widget to the **Data Table** widget.
8. Double click on **Data Table** on the canvas and inspect the dataset.
9. Add **Select Columns** Widget to the canvas and connect to the **File** widget. A sample pipeline is shown below.



10. Double click on **Select Columns**. Ensure it is as shown below:



Select Columns

Available Variables

Filter

N Depth(Ft)

Features

Filter

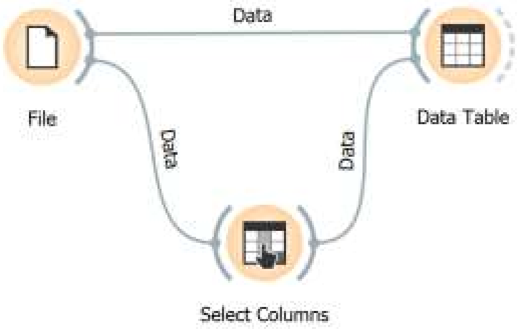
N GR
N Caliper(in)
N NPHI(%)
N DPHI(%)

Target Variable

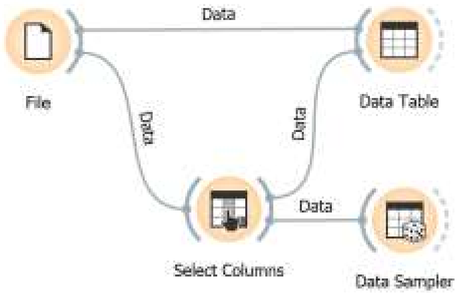
C Lithology

Meta Attributes

11. Connect **Select Columns** to **Data Table** as shown below:



12. Add the **Data Sampler** widget as shown below:

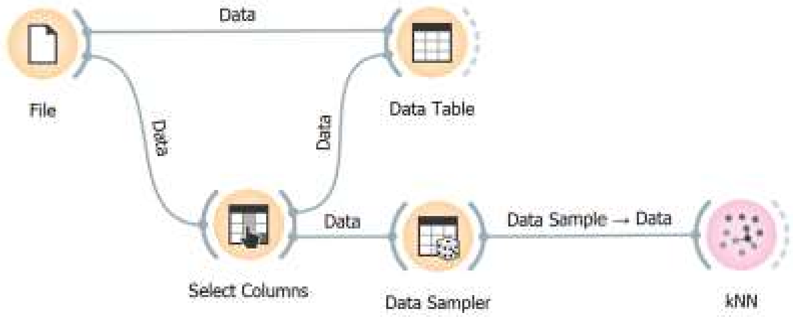


13. Select Fixed Proportion and set its value to 80%. Leave other options as default. What does this imply? Specifically, complete the following table

(2 points)

Training Samples Count	801
Test Samples Count	200

14. Add the **kNN** widget under **Model** or by right clicking on the canvas. Connect to **Data Sampler** as shown below.



15. Double click on **kNN**. (3 points)

a. Change k (**Number of Neighbors**) to sqrt(input samples).

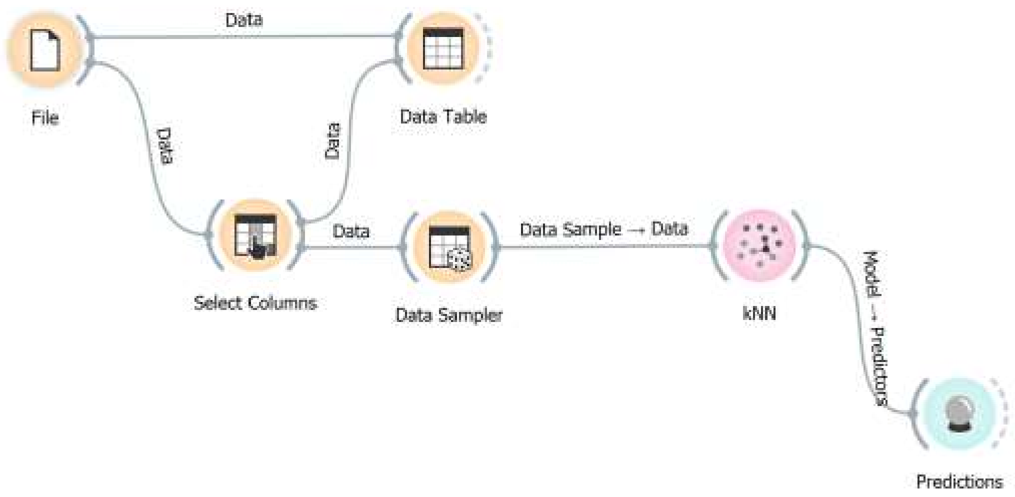
What value of k would you select?	29

Square root #of training samples; classification problems need to be odd

b. Select **Metric** as 'Euclidean' and **Weight** as 'Uniform'.

What is the difference between "Uniform" and "Distance"?	Uniform = equally weighted, equal influence Distance = closest neighbor more priority

16. Add the **Predictions** widget under **Evaluation** to the canvas. Connect to **kNN** as show below.

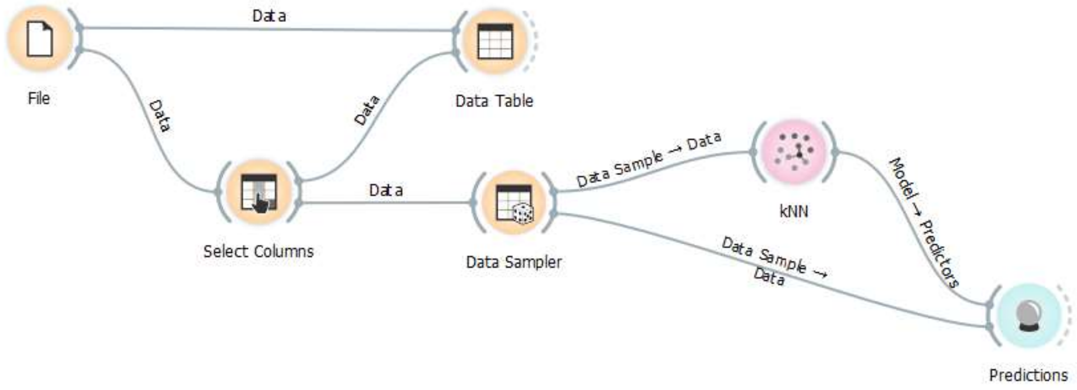


17. Double click on **Predictions**. What do you see? Specifically,

(1 points)

What is the value of CA?	no output, not loaded yet

18. Connect **Data Sampler** to **Predictions** as shown below.

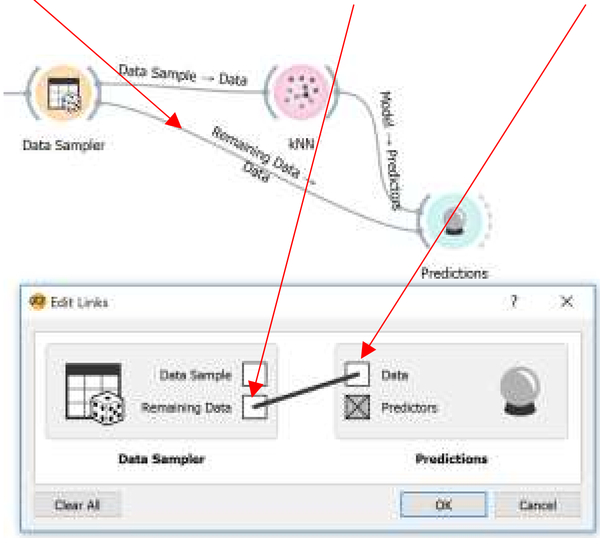


19. Double click on **Predictions**. What do you see? Specifically,

(1 point)

What is the value of CA?	0.918

20. Double click on the line and connect **Remaining Data** to **Data**. Click OK to close the window.



21. Double click on **Predictions**. What do you see? Specifically,

(1 point)

What is the value of CA?	0.920

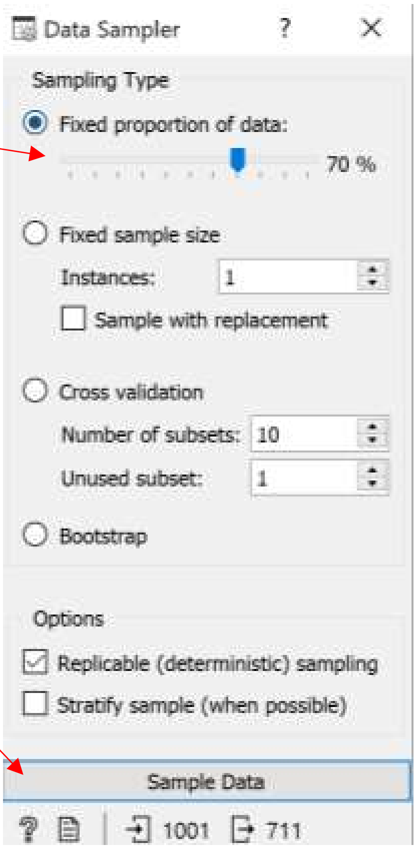
Summary: What is the value of CA? (2 points)

Training Set	0.918
Test Set	0.920

XXXXXXXXXXXXX Model is Ready → Now let's tune the model XXXXXXXXXXXXX

Parameter Tuning: 1/3. Changing Sample Size

22. Double click on **Data Sampler**.
    a. Change **Fixed Proportion** to 70% → Click **Sample Data** button at the bottom of the window → Open **Predictions** and check on **CA**.
    b. Change **Fixed Proportion** to 50% → Click **Sample Data**. → Open Predictions and check on CA.


Data Sampler

Sampling Type

● Fixed proportion of data: 70 %

○ Fixed sample size

Instances: 1

□ Sample with replacement

○ Cross validation

Number of subsets: 10

Unused subset: 1

○ Bootstrap

Options

☑ Replicable (deterministic) sampling

□ Stratify sample (when possible)

Sample Data

1001 711

Summary: What is the value of CA?

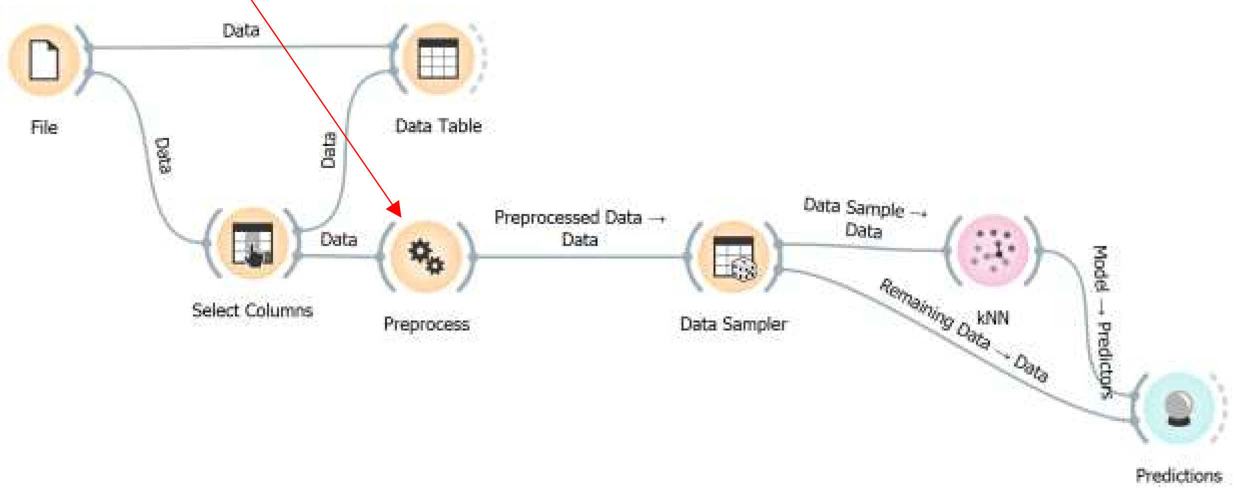
(6 points)

Test Set (20% of the data)	0.920
Test Set (30% of the data)	0.920
Test Set (50% of the data)	0.876

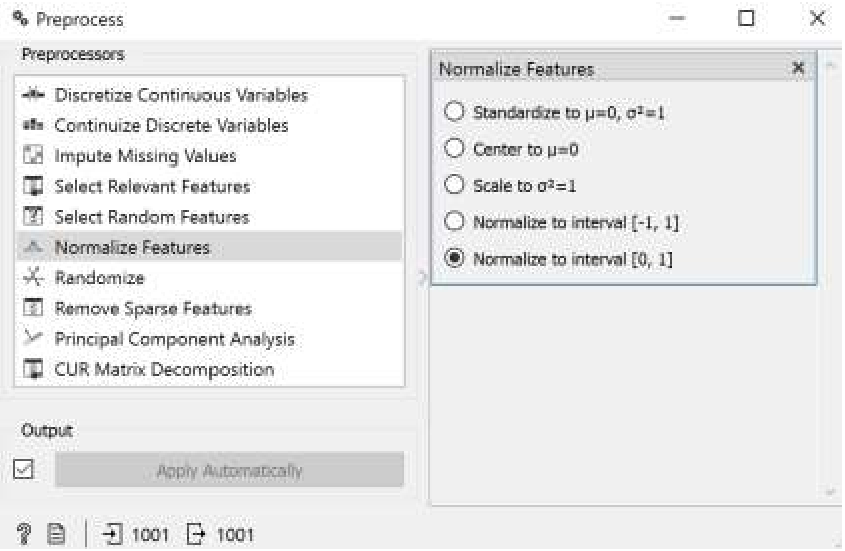
What conclusion would you draw from the above table?	More proportion of data means more accuracy of classification

Parameter Tuning: 2/3. Normalizing Data

23. Disconnect **Select Columns** and **Data Sampler** by right clicking on the line and selecting remove.
24. Add the **Preprocess** widget from **Data** and make connection as shown below**.**



25. Double click on Preprocessor and add the normalization module as shown below:

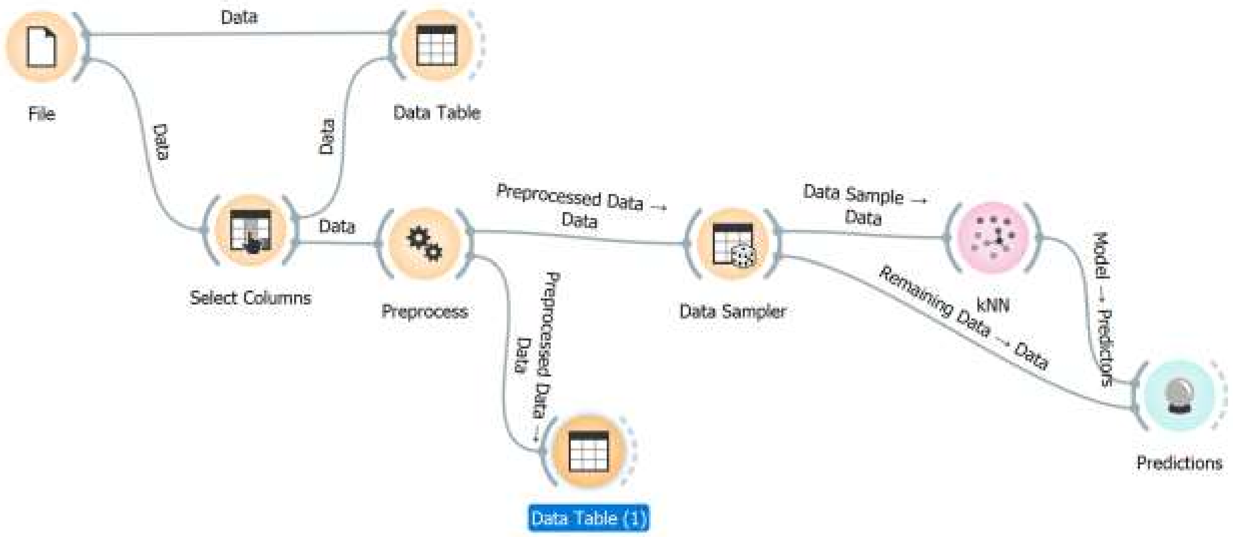


Normalize Features

○ Standardize to μ=0, σ²=1
○ Center to μ=0
○ Scale to σ²=1
○ Normalize to interval [-1, 1]
● Normalize to interval [0, 1]

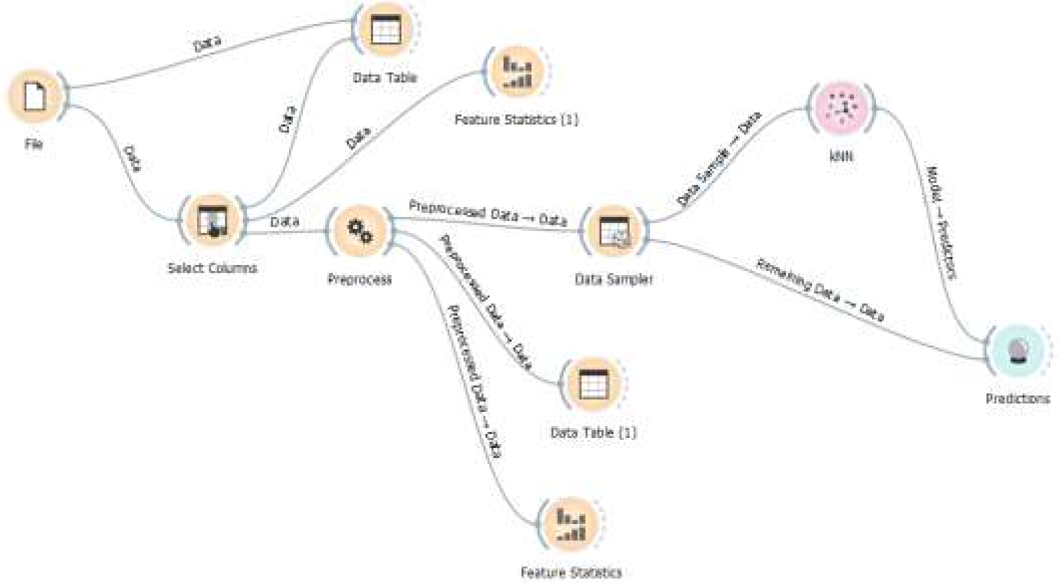
☑ Apply Automatically

26.Add the **Data table** widget. Connect to **Preprocess** as shown below.



27.Add the **Feature Statistics** Widget twice as below. Examine both the results.

(3 points)



What do you observe?	Better visualization of the data and organized

Correct answer: The features are normalized and lie within [0,1] range after preprocessing

28. Select **Data Sampler** → Change the values of **Fixed Proportion** to 80%, 70% and 50% → Click Sample Data → Fill the following table (6 points)

Summary: What is the value of CA?

	Before normalization	After Normalization
Test Set (20% of the data)	0.920	0.925
Test Set (30% of the data)	0.920	0.920
Test Set (50% of the data)	0.876	0.886

Parameter Tuning: 3/3. Tuning k

29. Double click on **kNN** → Change the k (number of neighbors) value to 1, 5, 11, 15, 21, 25, and 35. Keep **Fixed Proportion** of **Data Sampler** to 80%. Record the CA value in **Predictions** for each value. (22 points)

k	CA
1	0.975
5	0.960
11	0.950
15	0.950
21	0.940
25	0.930
35	0.925

Which model are you going to put in production? Why?	The one with k value of 1 since the accuracy seems to get lower when k is increased