Lab-5 [Methods of Evaluation]

Out date: Jun 29, 2022
Due date: July 3, 2022, 11:59PM

Submission

1. Prepare your solution in Orange and save the workspace for Problem 1 (e.g., Lab-5_LastName.ows) **[20 points]**
2. Complete the tables given below and save the file (e.g., Lab-5_LastName.docx). **[80 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

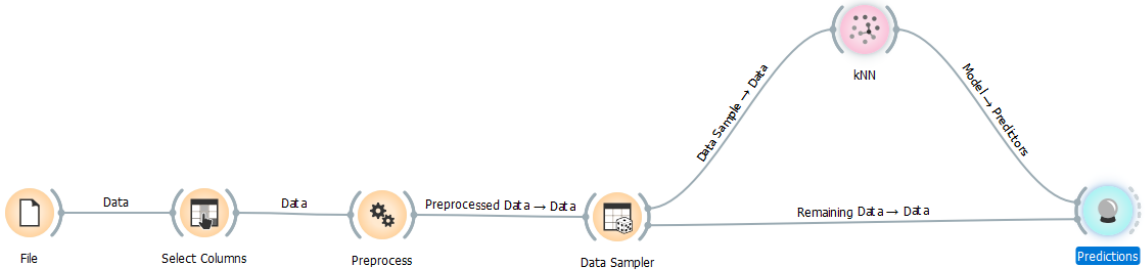
Please refer to http://www.kgs.ku.edu/General/copyright.html regarding use of data / information from Kansas Geological Survey.

Objective: To evaluate Machine Learning models for predict lithology using log measurements in an oil & gas well. This is a classification problem. We will build on the Orange pipeline that was generated as part of Lab-1 in Badge-1.

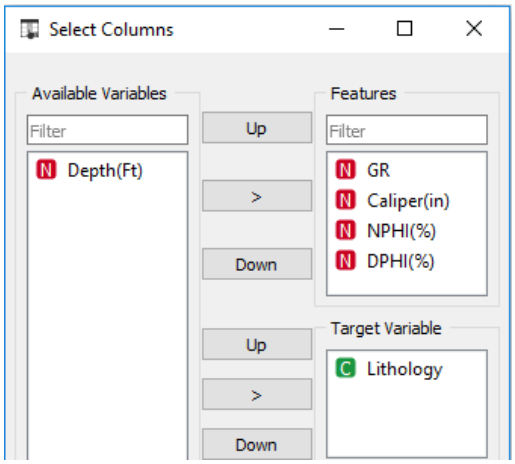
Data: Relevant log data was extracted to an excel file (***Log Lithology classification example.xlsx***) and lithology labels were created to be used for the hands-on lab.

Lab Instructions

1. Download Badge1_Lab5_Start.ows orange pipeline and the Excel data file (*Log Lithology classification example.xlsx*). Your pipeline should look as below:



2. Open **Select Columns** widget and confirm the following selections.

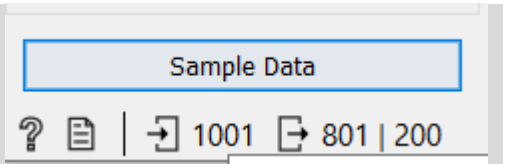


3. Add **Feature Statistics** and **Distributions** widget to Select Columns. Open the widgets, inspect the data and complete the following table: (12 points)

Go to data distribution, click on (green) Lithology

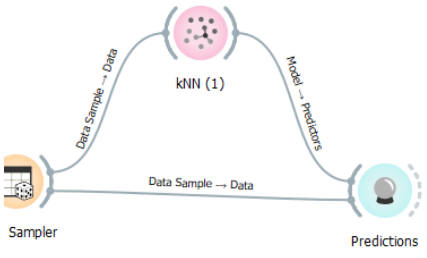
Lithology class label	% data
Dolomite	34.27
Granite	8.19
Granitewash	2.80
Limestone	26.57
Sandstone	8.49
Shale	19.68

4. Open **Data Sampler** and confirm the following: (3 points)
   Fixed proportion of data is selected and at 80%.
   Replicable (deterministic) sampling & Stratify Sampling Options enabled.

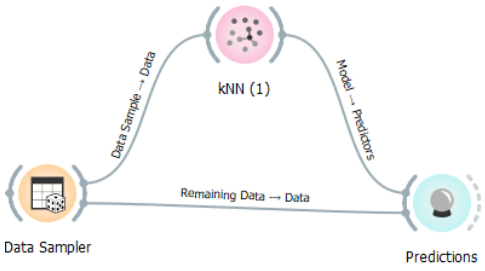
Number of Training Samples	

5. Open **kNN** widget and select k=3, Euclidean and Uniform weight.
6. Open **Predictions** widget and record the observed results in the table below: (10 points)

Dataset	AUC	CA	F1	Precision	Recall
Training (Sample)	1.000	0.986	0.986	0.986	0.986
Test (Remaining)	0.989	0.955	0.955	0.955	0.988

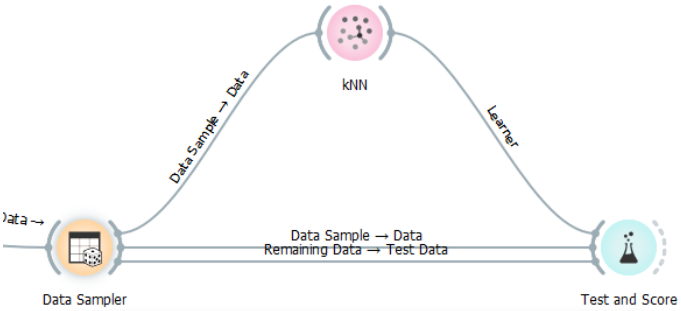


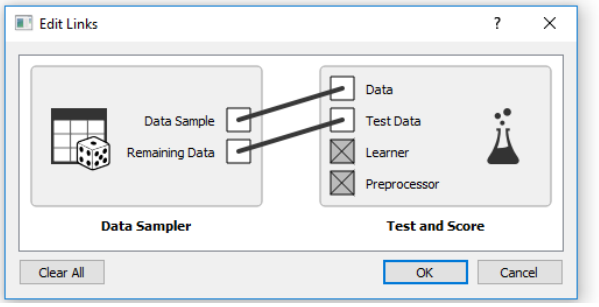
Model	AUC	CA	F1	Precision	Recall	Specificity
kNN	1.000	0.986	0.986	0.986	0.986	0.996



Model	AUC	CA	F1	Precision	Recall	Specificity
kNN	0.989	0.955	0.955	0.957	0.955	0.988

7. Remove **Predictions** widget and add **Test and Score** widget.
8. Connect **kNN** widget to **Test and Score** widget.
9. Connect **Data Sampler** to **Test and Score** as shown below.

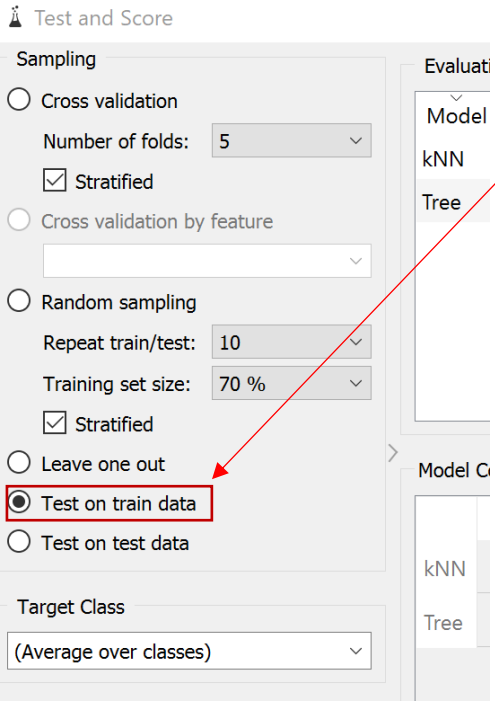




10. Add **Confusion Matrix** to **Test and Score** widget.

11. Open **Test and Score** and **Confusion Matrix** widgets.

12. Double click on **Test and Score** and select **Test on train data** as **Sampling** methods as shown below.



13. We will evaluate model performance on the **Training** dataset for the following scenarios. Complete table below: (20 points)

Model	**AUC**	**CA**	**F1**
kNN (3, Euc, Uni)	1.000	0.986	0.986
kNN (3, Euc, **Dist**)	**1.000**	**1.000**	**1.000**
kNN (3, Man, Uni)	1.000	0.986	0.986
kNN (3, Man, **Dist**)	**1.000**	**1.000**	**1.000**
kNN (5, Euc, Uni)	0.999	0.976	0.976
kNN (5, Euc, **Dist**)	**1.000**	**1.000**	**1.000**
kNN (5, Man, Uni)	0.999	0.979	0.979
kNN (5, Man, **Dist**)	**1.000**	**1.000**	**1.000**

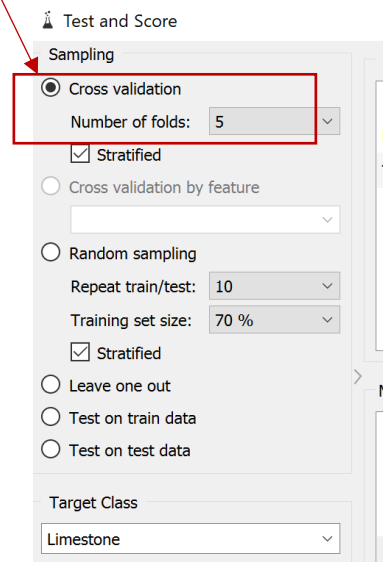
Record your observation(s) from the above table.	Model is doing well on the training data set when

	Whenever you have the distance weighted metric, whether you choose Manhattan or Eucl, you can get a perfect model on the training set, given it is the only training data set.

14. Set the kNN parameters to the best performing models from step 15. Change the **Target Class** in **Test and Score** to **Limestone** (class of interest). Complete the table below for **Test on train data**. (6 points)

Model	AUC	CA	F1
kNN (3, Euc, Dist)	1.000	1.000	1.000

15. Select **Cross validation** as **Sampling** method in **Test and Score** widget. Select 5 folds. Select Target Class as Limestone. Complete table below: (8 points)

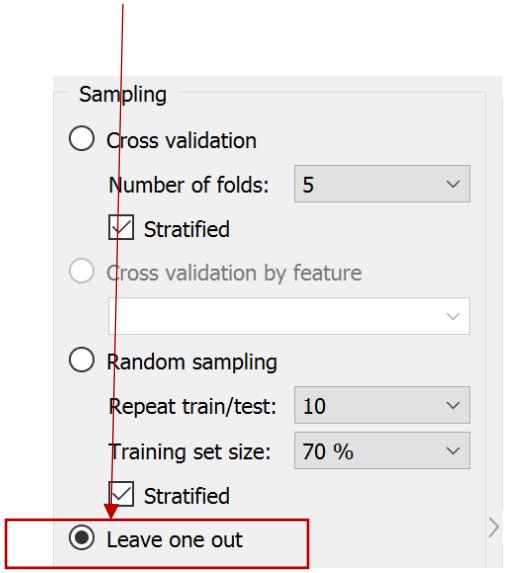


Model	AUC	CA	F1
2.929 kNN (3, Euc, Uni)	0.991	0.979	0.959

2.933 kNN (3, Euc, Dist)	0.991	0.980	0.962
2.932 kNN (3, Man, Uni)	0.987	0.981	0.964
2.940 kNN (3, Man, Dist)	0.987	0.984	0.969

Identify the best performing kNN model	kNN = 3 Man, Dist

16. Change Sampling to **Leave-one-out** in **Test and Score** widget and complete the table below. (12 points)



Model	Train time(s)	Test time (s)	CA	F1
kNN (3, Man, Dist)	4.186	1.600	0.986	0.973

Change Sampling to **Cross Validation** and observe the difference in Train and Test time.

Model	Train time(s)	Test time (s)
kNN (3, Man, Dist)	0.038	0.017

17. Let us test the performance of the above best performing models on test data by selecting **Test on test data** in the Test and Score widget.

Model	AUC	CA	F1
kNN (3, Man, Dist)	0.983	0.970	0.951