

# Badge-3 Kaggle Competition (Lab-6)

**Out date:** Aug 15, 2022

**Due date:** Aug 21, 23:59HRS

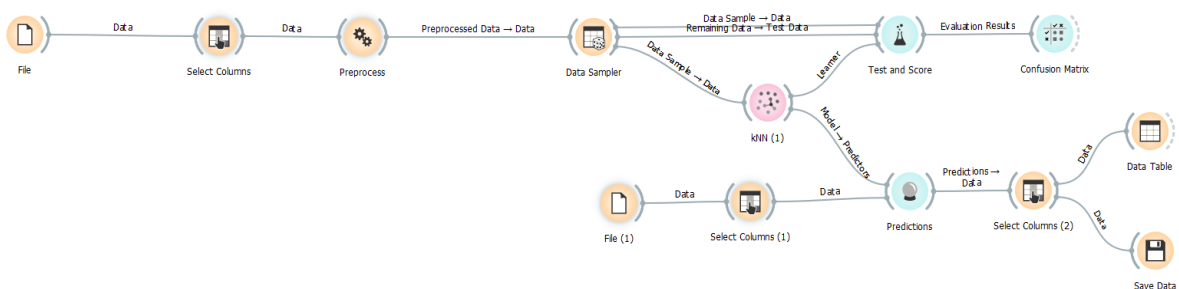
## Submission

1. Prepare your solution in Orange and save the workspace (e.g., Badge3\_IP\_LastName.ows).
2. Complete the tables provided in this document and save the document. (e.g., Badge3\_IP\_LastName.docx)
3. Save your final predictions in a csv file (e.g., Badge2\_IP\_Predictions\_LastName.csv).
4. Upload the following files to Canvas :
  - a. Badge3\_IP\_LastName.ows
  - b. Badge3\_IP\_LastName.docx
  - c. Badge3\_IP\_Predictions\_LastName.csv

**Objective:** Your customer is asking you to use Machine Learning to predict **Facies** in two wells (Well-A and Well-B), contained in the [topredict\\_facies.xlsx](#) file. As a data scientist, your job is to train machine learning models using the [Train.xlsx](#) file provided to you using **Orange**, deploy your best performing model to predict **Facies** for the two wells and submit the predictions to your client.

**Data:** Please download the following files from Canvas to complete your assignment:

1. IP\_start.ows file with the starting pipeline as shown below:



2. [Train.xlsx](#) containing the following features:

	Name	Type	Role	Values
1	Facies	C categorical	target	
2	Well Name	S text	meta	Well-1, Well-2, Well-3, Well-4, Well-5, Well-6, Well-7, Well-8, Well-9, Well-10
3	Depth	N numeric	meta	
4	GR	N numeric	feature	
5	ILD_log10	N numeric	feature	
6	DeltaPHI	N numeric	feature	
7	PHIND	N numeric	feature	
8	PE	N numeric	feature	
9	NM_M	N numeric	feature	1, 2
10	RELPOS	N numeric	feature	

prodFacies is meta, not considered

3. [topredict\\_facies.xlsx](#) which contains the same features as the [Train.xlsx](#) file shown above except the **ID** and **Facies** target variable.

#### Seven predictor variables:

Five wire line log curves include gamma ray (GR), resistivity (ILD\_log10), photoelectric effect (PE), neutron-density porosity difference (DeltaPHI), and average neutron-density porosity (PHIND).

Two geologic constraining variables: nonmarine-marine indicator (NM\_M) and relative position (RELPOS)

Target variable: Facies, contains nine discrete rock classes from 1 to 9.

Reference: <https://library.seg.org/doi/full/10.1190/tle35100906.1>

## Project Instructions

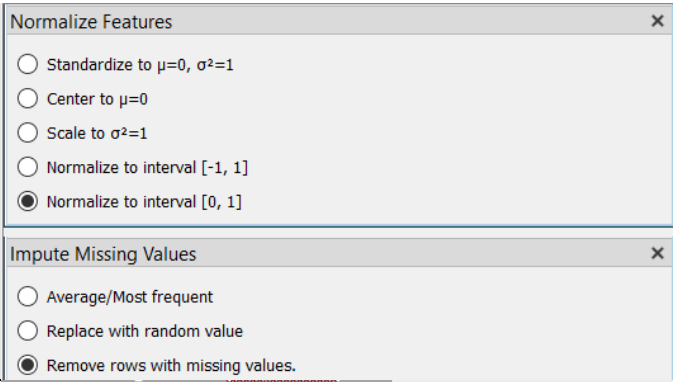
1. Launch Orange and open the [IP\\_start.ows file](#).
2. Add required widgets to the pipeline and answer the questions below: ( 10 points) –  
Training set

How many samples?	4149
How many meta features?	<i>Note: Meta features are not to be used in training your models</i> 3 meta roles, 0 meta attributes
How many features?	11
Which features have missing values and what is the % of missing values?	PE 22%
What are the distribution types and ranges for your features?	Combination of bar and skewed curves; target is bar only
What is your target variable? What are its class labels? What is the % of each class label in the dataset?	Facies 1 (6.46), 2 (22.66%), 3 (18.80), 4 (6.53), 5 (7.13), 6 (14.03), 7 (3.40), 8 (16.53), 9 (4.46)

3. You are free to apply all the techniques you have learnt in the Bronze Belt to train your Machine Learning models on the input dataset that you loaded in the above steps. You are encouraged to consider the following aspects:
- Preprocessing features (Imputing missing values, Standardizing or Normalizing features)
  - Ranking and dimensionality reduction
  - Splitting your training dataset to train and test using stratified sampling. Consider selecting Replicable sampling for repeatable results.
  - Select starting parameters for your machine learning models (kNN, Tree etc.,) using guidelines from class lectures.
  - Using Test and Score widget and Confusion Matrix widget to understand and evaluate the performance of your various models when you tune various parameters.
  - Consider class imbalance issues and over / under sampling to improve model performance.

Complete the table below:

( 20 points)

What preprocessing options did you consider?				
What is your best performing model and model parameters?	kNN (cv 20, stratified, 2 neighbors, man, distance, 90% fixed) –helped score??	0.910	0.753	0.753 0.956
Did you use Ranking to consider impact of different features? If you used Ranking to reduce features for your final model, list features you used in your final model:	No, we didn't explore this			
Did you explore PCA for dimensionality reduction and evaluate its impact on model performance?	Yes, didn't help			
If you used dimensionality reduction using PCA for your final model, provide details on how many components you considered and the % variance explained:	N/A			
Did you consider over / under sampling to address class imbalance and evaluate its impact on model performance?	No			
If your final model used over / under sampling, what were the final sampling choices for your different class labels?	N/A			

Complete the table below based on **cross validation (5 folds)** on your training data. Select *Target Class as Average over classes* and use metrics from Evaluation Results in the **Test and Score** widget. You should try at least five different models:

(30 points)

Model	AUC	CA	F1	Specificity
kNN (5N, Euclidean, Uniform ) (example entry)	0.92	0.653	0.652	0.941
Random forest (30 trees, 3 split, no balance, limit 18, subset 2)	0.943	0.712	0.709	0.945
kNN (20 folds, stratification, number of neighbors 2, Manhattan, distance)	0.935	0.808	0.808	0.966
Logistic Reg (c=700, no balance)	0.901	0.574	0.559	0.919
Logistic Reg (c=700, balance)	0.897	0.514	0.508	0.933
kNN (98% fixed proportion in data sampler, 2 neighbors, Man, distance, 20 folds, remove rows with missing values preprocess)	0.941	0.818	0.818	0.967
kNN (cv 20, stratified, 2 neighbors, man, distance, 90% fixed) –helped score??	0.910	0.753	0.753	0.956

Submission and Description	Public Score	Use for Final Score
<a href="#">Badge2_IP_Predictions_RattanR-3.csv</a> 2 minutes ago by <a href="#">Riyan Rattan</a> <a href="#">add submission details</a>	0.53975	<input type="checkbox"/>
<a href="#">Kaggle-6.csv</a> 2 minutes ago by <a href="#">Priyanka Kumari</a> <a href="#">add submission details</a>	0.46746	<input type="checkbox"/>
<a href="#">Kaggle-6.csv</a> 3 minutes ago by <a href="#">Priyanka Kumari</a> <a href="#">add submission details</a>	Error	<input type="checkbox"/>
<a href="#">Badge2_IP_Predictions_RattanR-2.csv</a> 6 minutes ago by <a href="#">Riyan Rattan</a> <a href="#">add submission details</a>	0.56867	<input type="checkbox"/>
<a href="#">topredict_facies.csv_team4_2.csv</a> 8 minutes ago by <a href="#">Chan Chakrya Menh</a> <a href="#">add submission details</a>	0.02409	<input type="checkbox"/>
<a href="#">Kaggle-5.csv</a> 12 minutes ago by <a href="#">Priyanka Kumari</a>	0.55662	<input type="checkbox"/>

Public Private

This leaderboard is calculated with approximately 50% of the test data. The final results will be based on the other 50%, so the final standings may be different.

#	Team	Members	Score	Entries	Last	Code	Join
1	Team 9		0.61927	5	5m		
2	Team 1		0.59277	9	5m		
3	Team 5		0.57349	7	11m		
4	Team4		0.56867	10	3m		



Your Best Entry!

Your submission scored 0.53975, which is not an improvement of your previous score. Keep trying!

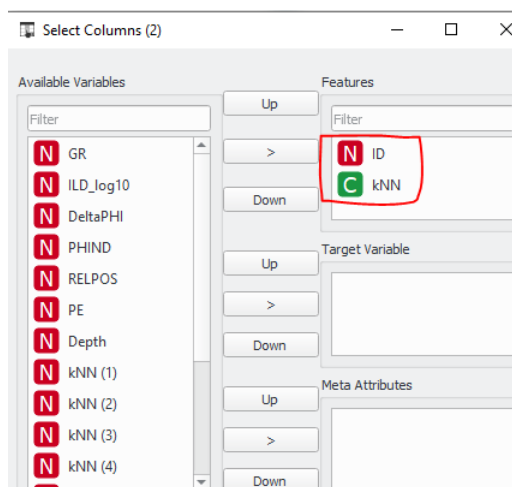
Once you have trained your models and decided on the best model, select **Test on test data** in the **Test and Score** widget and use the Evaluation Results to complete the table below (10 points):

Target Class	AUC	CA	F1	Specificity
Average Class	0.922	0.752	0.751	0.958

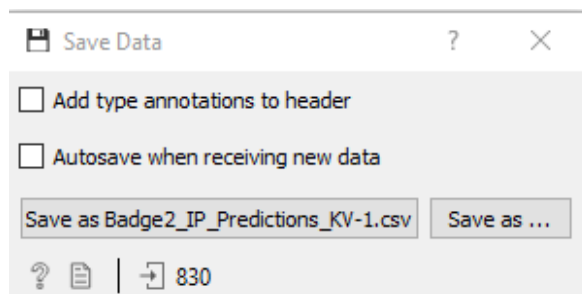
4. Open the **File** widget and load the [topredict\\_facies.xlsx](#) file. ( 5 points)

How many wells and samples are there in this data set?	Well A and B, 830
--	-------------------

- Verify the rest of this pipeline and ensure **Predictions** widget contains the predicted **Facies** values for all the samples, from your best model.
- Use **Select Columns(2)** widget to select only the predicted values and verify using the **Data Table** widget that you have only one column of predicted **Facies** values.



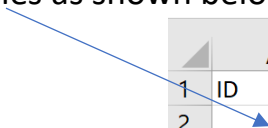
- Use **Save Data** widget to save your predictions as *Badge2\_IP\_Predictions\_LastName.csv*. Uncheck **Add type annotations to header** and save the .csv file to your local folder.



- Predictions csv file should output results in two columns: 1) ID, and 2) the name of ML algorithm (e.g., kNN here) as shown below.

	A	B
1	ID	kNN
2	1	1
3	2	1
4	3	1
5	4	1
6	5	1
7	6	1
8	7	1
9	8	1
10	9	1

9. Rename column B to Facies as shown below.



	A	B
1	ID	Facies
2	1	1
3	2	1
4	3	1
5	4	1
6	5	1
7	6	1
8	7	1
9	8	1
10	9	1
11	10	1

*Note: If Kaggle gives you an error during submission, change the ID column to Text format and submit your results.*

10. Upload the csv file to Kaggle.

<https://www.kaggle.com/t/996806730910450eb3f6cf3d0852606c>

11. Upload your final .ows and final .csv predictions to Canvas. (25 points)

Make sure to select column A (ID), right click to format, change to text, if get errors

Select A column (ID), right click to format, change to text