In the Jupyter notebook presented, we are introducing an approach to determine lithology from labeled core and well log data for one oil and gas field with multiple horizons. The lithology labels from core data are used to train a supervised machine learning (ML) algorithm to predict lithology from well log response. This algorithm can be extended to any other reservoir to improve rock type prediction.

Coring is an expensive operation, leading to only a small number of cored wells in specific intervals. Opposite to coring, well logging is relatively inexpensive and it is a standard operation for all wells. The relationship between lithology and well log data is often overwhelming, complex and nonlinear. Recently, many studies have shown that machine learning algorithms could improve lithology prediction. Main benefits from improved lithology prediction accuracy would be cost and time efficiency, especially for decision-making in reservoir evaluation and management.

As a baseline solution for this hackathon, we have given a KNN classification ML algorithm. Such data-driven analysis may optimize logging service and laboratory core measurements planning, as the aim in supervised learning is to label small portions of a dataset and allow the algorithm to automate the rest.

Available to participants is a dataset (`Train-dataset.csv`) containing following well log curves from 11 wells across one field:

- Gamma ray log (`**GR**`)

  Gamma ray log measures naturally occurring radioactivity of a formation. GR is mainly used for determination of the shaliness of a formation. Values of gamma ray log will be low in shale-free sandstones (composed of nonradioactive quartz) and high in shales because of high concentrations of radioactive material. However, clean sandstone (i.e., with low shale content) might also produce a high gamma ray response if the sandstone contains potassium feldspars, micas or uranium-rich waters.

- Compensated neutron log (`**CN**`)

  Neutron log is counting the amount of hydrogen atoms in a formation. The tool operates by bombarding the formation with high energy neutrons. These neutrons collide with hydrogen atoms and undergo scattering in the formation, losing energy and producing high energy gamma rays. Hydrogen is mostly concentrated in fluids the pores and hence neutron log is an indicator of porosity. Neutron count is inverse to porosity (Hydrogen Index), but CN is given in porosity units, i.e. percentage after being calibrated to become a direct indicator of porosity magnitude.

- Density (`**DEN**`)

Density log is related to the bulk density of formation. The tool principle is based on a radioactive source and counter. As the tool is run against the borehole wall it emits radioactive particles that interact with electrons in formation via Compton scattering and photoelectric adsorption, resulting gamma ray is counted and analyzed to determine the bulk density of a rock. The bulk density is the overall density of the matrix and the fluids (water, oil, gas) within the pores. Hence the traditional use of density log is determination of porosity. In formation with low porosity, measured density values will be closer to the mineralogical (matrix) density, and with increase of porosity density values will get lower. Measured density curve is used for lithology differentiation, especially in combination with CN.

- Resistivity (`**RT**`)

Resistivity log is indicator of capability of the formation to prevent current flow through it. Hydrocarbons and non-porous rocks do not conduct electric current, but when rock is porous, and pores are interconnected and saturated with formation water, current can flow. Therefore, a difference in resistivity of non-porous rocks, hydrocarbon saturated formation and saline water saturated formation exists and it is essential in identification of hydrocarbon bearing zones. Given RT is a true formation resistivity, reading resistivity values furthest in the formation from the borehole. There are other factors also affecting resistivity values. Clay and other minerals, such as pyrite, are conductive and reduce the resistivity values.

Dataset contains metadata columns:
- `**WELL**` - name of a well
- ‘**X**’ - coordinate
- ‘**Y**’ - coordinate
- `**MD**` - measured depth of a well log response
- `**DEPOSITIONAL_ENVIRONMENT**`` - Describing the conditions in which the sediment is

deposited. Different depositional environments influence on trend of some well log curves and presence of certain lithology classes.
- `**LITH_NAME**` is a lithology class label
- `**LITH_CODE**` is a lithology code

The validation of your prediction will be performed on a hidden dataset (`Test-dataset.csv`) with a similar lithology class distribution as the training set.

We encourage participants to try additional features that will increase the F1 score of the model. For example, look for the spatial correlation of the features, explore the relation between classes (top and bottom from some interval of interest) or try to find some lateral trend in the data.

Your task as a participant is to make the best lithology prediction you can.