Mini Project due May 1 (5 pm)

**Objective: Predict well deliverables using completion data and identify the key variables that affect them (well deliverables)**

Develop regression model that effectively predicts key well deliverables (IP90, EUR, Costs, etc.) from available data (well data). Use the following regression techniques:

1. Elastic Net
2. Support Vector Regressor
3. Random forest
4. Gradient Boosting Regressor
5. Neural Network

**Dataset and Tasks**

Dataset is an excel file with two sheets:

* Data1\_Completion

The dataset comprises of well completion variables (features) such as: permeability, pay thickness, fluid Intensity, sand Intensity, number of stages, perforation clusters, fracture conductivity etc.

* Data2

The dataset comprises of different well deliverables and economic metrics, such as IP90 (average daily production rate after 2160 hours [90 days] of production), EUR (estimated ultimate recovery).

**Rubric for Grading the Final Project**

**Notebook #1**

* Identify and import features and targets. Clearly state the objectives of your work. Try to predict several well deliverables and economic metrics. – **2**
* Perform outlier detection and fix missing data if any. Perform data scaling/standardization – **2**
* Ensure testing data is not used for developing preprocessing strategy. Avoid any mixing of training and testing data. After the train-test split, the testing data should be treated like a new data that the data preprocessing and model training does not have access to.
* Perform dimensionality reduction (feature selection/feature extraction) – **2**
* Use all the above-mentioned methods for developing the regression models for the various tasks/objectives requiring predictions of various well deliverables and economic metrics. Perform cross validation with hyperparameter optimization to ensure the model is well generalizable. – **5**
* Perform good evaluation using good metrics during the training and testing stages. Which are the best metrics for which targets – **2**
* Visualize the predictions – 2
* Identify the key features that affects each deliverable/target. Visualize the importance – **2**
* Share your opinion and insights based on your petroleum engineering knowledge – **2**

**Notebook #2**

* Save the 3 best performing models developed in Notebook #1 for any objective/task. Deploy them on new, unseen data in separate notebook #2. You cannot evaluate the performance on unseen data because there are no targets. – **3**

**Notebook #3**

* Show/explain something new and interesting. Use visualizations. – **2**

**Notebook #4: Use of clustering for feature extraction**

* Train clustering on training samples. Clustering models should be fitted only on the training data. After training, the trained clustering models can be used to generate cluster labels for both training and testing samples. Do the new clustering-based categorical features improve the performance of regression models? You will need to redo few of the tasks done in Notebook #1 with new clustering-derived features. – **4**