American Journal of Sciences and Engineering Research

E-ISSN -2348 – 703X, Volume 5, Issue 6, 2022

Rock Type Classification Using Machine Learning Algorithms with Python in Algerian Triassic Sandstone Reservoir

Dr Abdallah SOKHAL[1], Manel HAMADAS, Sonatrach[2]

Petroleum Engineeringand Development Division, Algiers, Algeria,

Abstract: *In this work, supervised classifiers were tested in a rock type classification issue: Support Vector Machines (SVM), Random Forest Classifier (RFC), Logistic Regression (LR) and K Nearest Neighbors Classifier (KNN). All this implementation was developed on an internal solution codified with Python. The objective was to determine the most effective classifier for rock types prediction in wells without cores in Triassic reservoir from Rhourde of Chegga (RDC) field in Algeria. Study data consist of 550 samples with known rock types from core within each sample having five log measured properties. The Pandas, Numpy, Seaborn and Sklearn libraries from Python were used for loading, preparing, conditioning and mapping the data. The dataset was split into two subsets, one set for training and one set for testing the capability of the trained classifier to accurately classify rock types. Support Vector Machines (SVM) and Random Forest Classifier (RFC) definitely surpassed all other classifiers and are efficient techniques for this specific classification issue.*

Keywords: *Machine learning, Rock type, Classification, SVM, Python, Pandas.*

I. Introduction

Artificial intelligence (AI) is defined as a subfield of computer science that includes the utilization of computers in tasks that typically require skills in reasoning, intelligence,learning, and understanding [7, 8].

Machine Learning (ML) is the spearhead of AI technology which is essentially a category of data analysis algorithms that involve classification, regression, and clusteringapproach [3, 4]. The ML method is mainly split into supervised and unsupervised type. For supervised ML, the key members are input features and target output [6].

Geoscientific challenges meet in oil and gas exploration, development and production environment are now commonly fixed by ML algorithms using wire-line logs and core data [1]. Machine learning approaches are often employed in reservoir characterization dealing with rock type modeling [2, 9]. They are powerful tools in reservoir nonlinearity examination [5, 7]. This ability grades ML among the most used clustering and classification methods [6].

Rock type classification is one of the most relevant step in the reservoir characterization. Each rock type can be defined by similar depositional and diageneticenvironment [2, 9].

In this work, we demonstrated how to train varied machine learning algorithms topredict rock type from cored and well log data. The dataset for this work drives from Triassic reservoir of Rhourde of Chegga (RDC) field which is located in Oued Mya basin, south east Saharan platform in Algeria. It represents the NNE field of the oil discoveries satellites surrounding the super-giant field of Hassi Messaoud (Figure 1).

 Received-03-12-2022, Accepted- 13-12-2022

The RDC field was mainly producing from the Triassic RDC sandstone with three other reservoirs are oil bearing, the Triassic T1 & T2, and Gres Intra volcanics (Figure 2).

The dataset includes core and log data of ten wells. We employed these data to train supervised classifiers such as Support Vector Machines (SVM), Random Forest Classifier (RFC), Logistic Regression (LR) and K Nearest Neighbors Classifier (KNN) inorder to predict discrete rock type groups. All this implementation was developed on an internal application codified with Python.

Comparable to all data sciences methods, we first begin with data examination by reviewing logs and cross plot in order to choose the suitable method to see the variation of the data. Then data requires to be conditioned and delete deficient parts that have incomplete data. Also, to increase the model performance efficiency, the data should be standardized to zero mean and unit variance. The dataset was separated into training, test and blind well data. Than many classifiers were applied to match the model and we explained how to use the cross validation set to do model parameter selection. After that, models are applied to test data.

Finally, once we have developed and tuned the classifiers, we can employ the trained model to classify rock types in wells which do not have core data. Also, we analyzed the model efficiency with a blind well which was not included in the model building process.

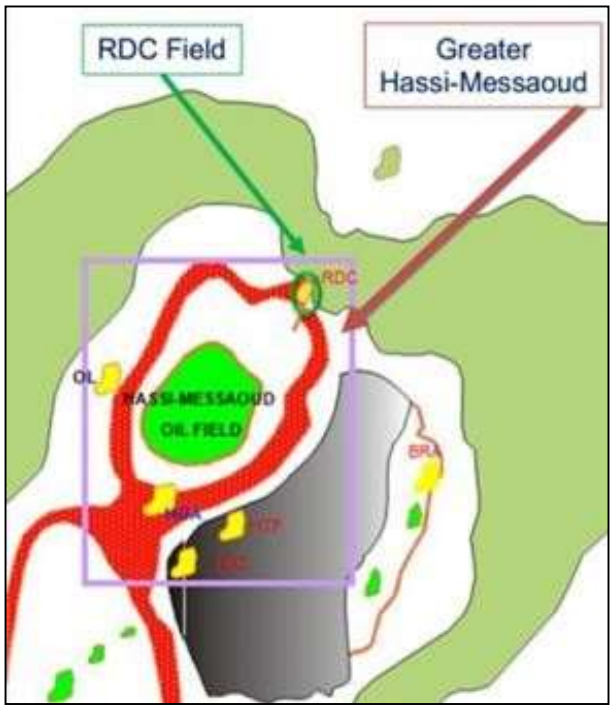

Figure 1. RDC field location.

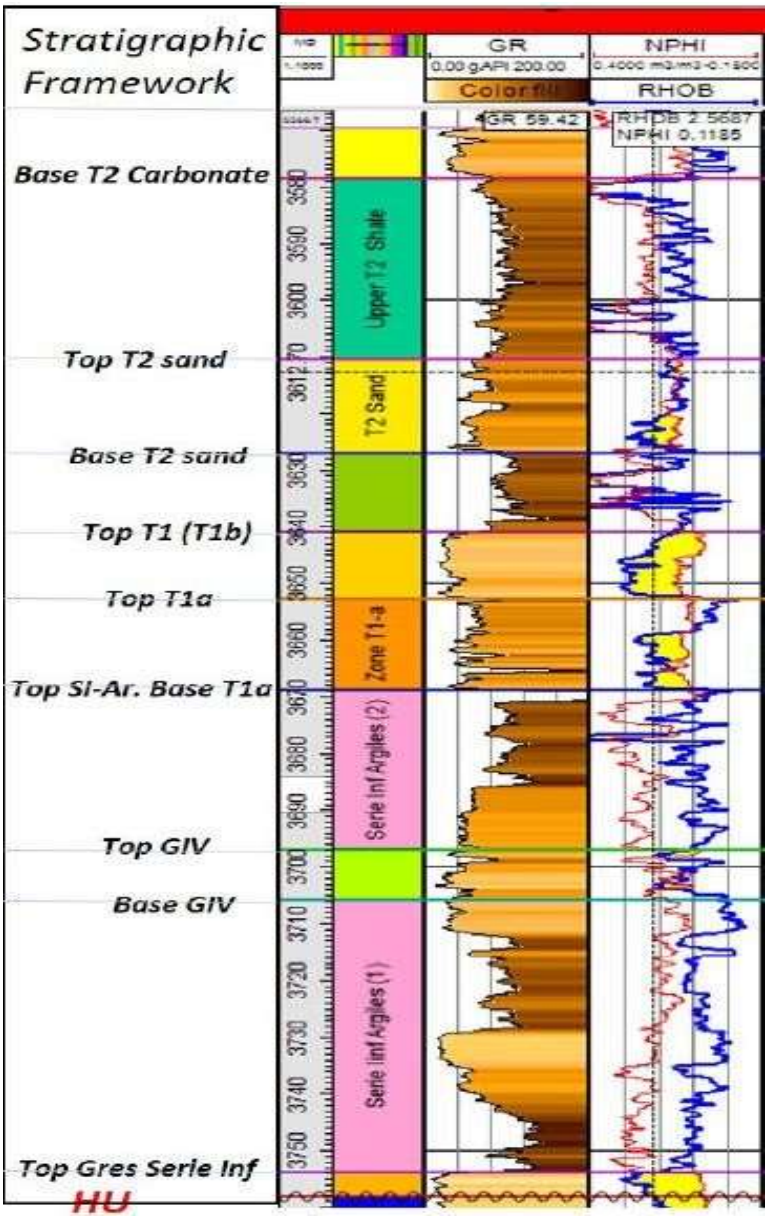

Figure 2. Stratigraphic Framework of RDC field.

II. Exploring the dataset

The training data is stored in CSV file which contains five wireline log measurement and one rock type label. In Machine learning vocabulary, we assume that any log measurement response is a characteristic vector which designs a set of features (log measurements) to a group (rock type) [7]. We employed the Pandas, Numpy and Seaborn libraries from Python to load the data into a data frame that gives an appropriate data architecture to work with well log data [2].

From the Triassic reservoir, ten wells are available. Rock types are studied from core data and fitted with logging data in well location. Feature variables involve six log curves which are gamma ray log (GR), resistivity log (RT), bulk density log (RHOB), neutron porosity log (NPHI), sonic log (DT) and effective porosity log (PHIE).

Based on previous petrographic studies [8], The reservoir T1 is defined by very fine to fine grained sandstones, by detrital mineralogy dominated by quartz with subordinate feldspar, by reservoir quality is likely to be fair at best. The low volume of macrospores in T1 sandstones is due to the relatively fine grain size and common

intergranular clays/cements (Figure 3).

Figure 3. Petrography of T1.

The reservoir T2 is characterized by very fine-grained sandstones, relatively clay-rich, detrital mineralogy dominated by quartz with subordinate feldspar, very rare visible porosity, poor reservoir quality. The low volume of macrospores in T2 sandstones (poor reservoir quality) is due to the very fine grain size; to relatively high volumes of intergranular clay-grade material and/or pervasive pore-filling cements (Figure 4).

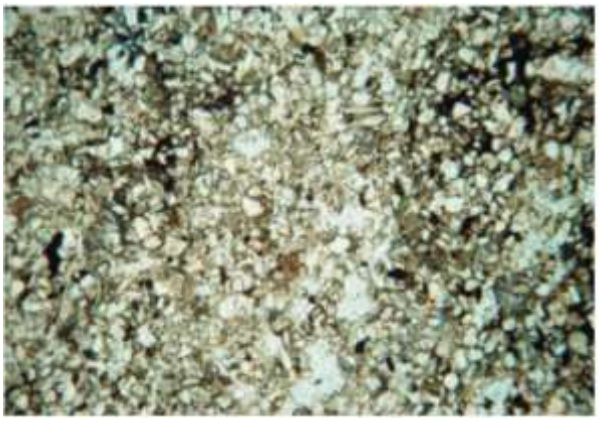

Figure 4. Petrography of T2.

The Deterministic Rock Typing application takes as input high quality data points of porosity and permeability from core. The comparison between the statistical best pore throat and the different methods, shows that the Winland R35 is the best approach to use for T1& T2 reservoir. R35 is based on the relationship between porosity, permeability, and pore throat radius at the point of 35% mercury saturation in capillary pressure measurements and is generally reliable in rocks with only intergranular porosity (such assandstone).
According to the Winand R35 results, the Triassic reservoir is composed of six rock types (RT1 to RT6): RT1 and RT2 with good reservoir quality (average porosity is 9%, average permeability is 7mD), RT3 and RT4 with moderate quality (average porosity is 7%, average permeability is 0.2mD), and RT5 and RT6 with bad quality (average porosity is 2%, average permeability is 0.01mD). The six defined rock types are reported below in table 1.

Porosity	Permeability	R35	Lithology

	Min	Max	Min	Max	Average	
RT1	1.84	14.92	1.57	21.22	3	clean sandstone, well sorted with macro pore throat
RT2	0.46	14.18	0.03	6.96	1.2	clean sandstone, well sorted with meso pore throat
RT3	0.8	12.79	0.01	1.08	0.476	sandstone having low fraction of shale with micro pore throat.
RT4	1.15	11.34	0.003	0.2	0.166	shaly sandstone with very small pore thorat
RT5	3.7	10.23	0.002	0.02	0.078	shaly sandstone with very small pore throat and significant of volume of clay
RT6	0	3	0	0.001	0	shales with high clay volume

Table 1. Summary table of the petrophysical properties of each RT.

"Table 2" depicts a quick view of the statistical distribution of the input variables. All log values have 5063 accurate entries.

	GR	RT	RHOB	NPHI	DT
Count	5063	5063	5063	5063	5063
Mean	62.41	7.28	2.58	0.16	70.37
Std	30.95	9.33	0.08	0.06	7.20
Min	15.54	0.88	2.03	0.02	50.01
25%	40.39	2.38	2.53	0.11	65.86
50%	55.60	3.78	2.60	0.15	69.67
75%	80.43	7.92	2.65	0.21	74.33
Max	318.18	95.30	2.80	0.42	143.32

Table 2. Quick view of the statistical distribution of the input variable.

Before we map out the well data, we employed the Matplotlib library from Python todefine a color map [7], so the rock types are depicted by consistent color in all the plots and we also generated the abbreviated rock types labels and add those to the data frame. "Figure 5" shows an example of plots for the five well log variables and a log for rock types label.

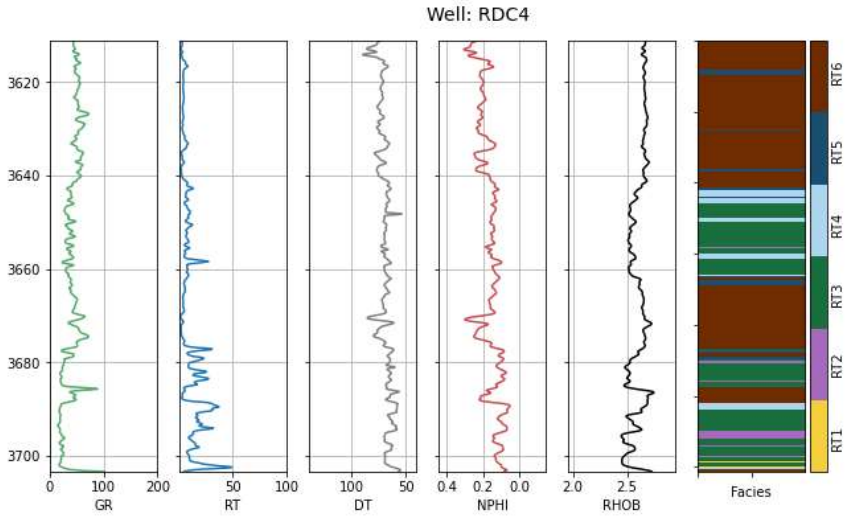

Figure 5. Example of plots for the five well log variables and a log for rock types label.

In addition, we can examine visually how the rock type are showed in the whole training dataset. So, we plotted a histogram of the number of training examples for each rock type class. “Figure 6” displays the distribution of the rock types in the training set. RT1 has the fewest with 85 examples. There are also only 195 RT2 examples.

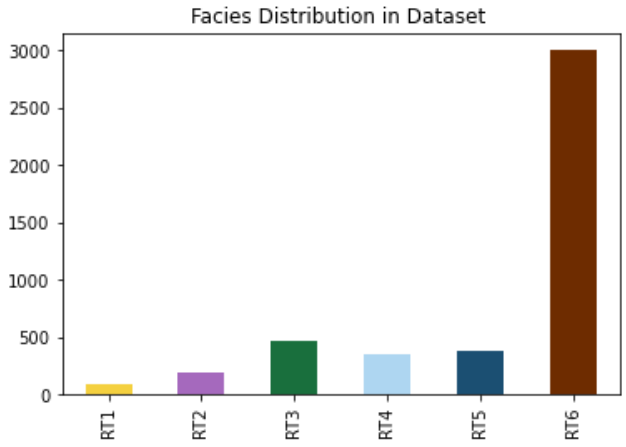

Figure 6. Histogram of the number of training examples for each rock type class.

Cross plots are a common tool used in the rock typing characterization to see how two properties vary with rock type. This data set includes five log variables and a scatter matrix can be benefit to view the variation between the all variables in the dataset.

Each window in the plot displays the relationship between two variables on the X and Y axis within each point is painted according to its rock type label. The same previous color map is used to show the 6 rock types (Figure 7). It is not clear from these crossplots what relationships exist between the measurements and facies labels. This is where machine learning will prove useful.

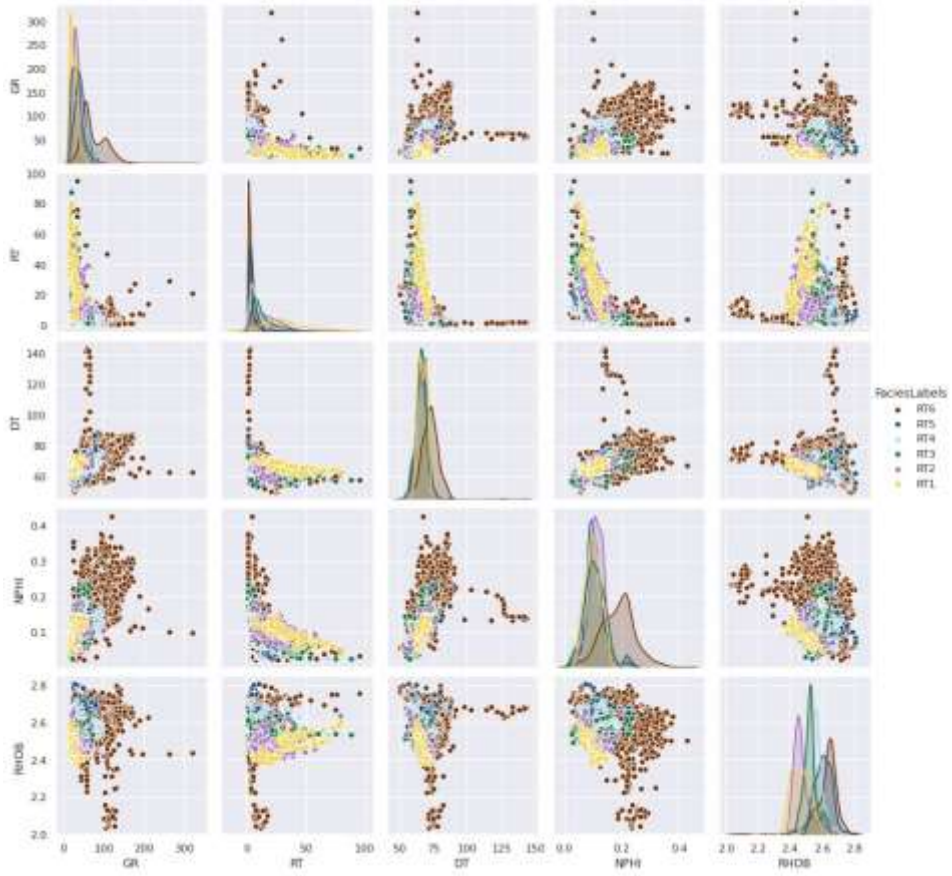

Figure 7. Plot of the relationship between the five log variables within each point is paintedaccording to its rock type label.

We used the famous Seaborn library from Python to generate a good looking scatter matrix [7]. “Figure 8” shows that the best correlation of target label (rock types) is relatedto PHIE with 92% arrangement. RT logging tools also displays 58% agreement.

Figure 8. Scatter matrix of the six log variables.

III. Conditioning the dataset

Practically, complete machine learning algorithms operate accurately when data is scaled for zero mean and unit variance [1]. This work was done by using Scikit preprocessing module from Sklearn library [7]. We used also Scikit data split function tocasually divide the data into training and test sets. The test set includes a small part of feature vectors which are not employed to train the model.

For the reason that we identify the true labels for these examples, we can contrast andcompare the results of the classifiers to the original rock types and evaluate the accuracy of the models. So, we choose 20% of the data for the test set.

IV. Training the classifiers

Presently, we can use the cleaned and conditioned training set to generate multiple rock type classifiers. As mentioned above, there are various kind of model approaches which can be employed for rock type classification. In this work, we used four types ofmachine learning model known as Support Vector Machines (SVM), Random Forest Classifier (RFC), Logistic Regression (LR) and K Nearest Neighbors Classifier (KNN).

The SVM is a map of the feature vectors as points in a multi-dimensional space mapped, so that examples from different rock types are divided by a clear gap that is aswide as possible [1]. The SVM implementation in Scikit learn has a number of relevant parameter, the classifier has been established with the default parameters. Despite, we may able to get enhanced classification results with optimal parameter choices. The SVM learning algorithm deals with two parameters, the parameter C is a regularization factor which informs the classifier how much we want to prevent misclassifying training examples and the kernel function which computes the distance between feature vectors [1]. In this work, we employed the radial basis function, the gamma parameter defines the size of the radial basis function which is based on how faraway two vectors in the feature space need to be considered close. We train a set of classifiers with different values for C and gamma, the best accuracy is reached for gamma=1 and C=10. Now we can generate and train an optimized classifiers based onthese parameters.

V. Model evaluation

There are several ways to evaluate how effectively our classifiers are working and how models perform on dataset prediction. The fundamental for all types of evaluation is based on identical, how far or close isthe predicted data from original data. A confusion matrix is a table that can be employed to define the performance of a classification model [1]. Scikit learn library from Python lets us to quickly generate a confusion matrix by providing the original and the predicted rock type labels. The confusion matrix is a 2D array of predicted and original target label [7]. The entries of confusion matrix Cij are equal to the number of observations predicted to have rock typej, but are known to have rock type i [2].

To make it easier to examine the confusion matrix, a function has been used to show the matrix including rock type labels and several error metrics. The confusion matrix hasrows and columns, the rows display original rock types label and the columns depict model prediction results (Figure 9).

SVM F1-score: 0.939

Pred	RT1	RT2	RT3	RT4	RT5	RT6	Total
True							
RT1	16	4	1				21
RT2	1	28	4	2			35
RT3	1	3	77	9	1		91
RT4		1	10	40	5		56
RT5			2	4	77	4	87
RT6					3	605	608
Precision	0.89	0.78	0.82	0.73	0.90	0.99	0.94
Recall	0.76	0.80	0.85	0.71	0.89	1.00	0.94
F1	0.82	0.79	0.83	0.72	0.89	0.99	0.94

Figure 9. Confusion matrix of SVM classification model.

RT1 is determined with 16 true predictions while one member of RT1 is predicted asRT2 and one member as RT3.

The entries along the diagonal are the rock types that have been accurately classified. As high value as a possible outcome of the model in diagonal of the matrix, as good as model prediction performance.

Precision and recall functions can be calculated efficiently using the confusion matrix.These metrics give more insight into how the classifier operates for individual rock types. The Precision function is the probability that given a classification result for a sample, thesample actually belongs to that class [1]. The recall function is the probability that a sample will be correctly classified for a given class [1].

Let's examine the results, if a sample was labeled RT1, the probability that the sample was correct is 0.89 (precision).

If we know a sample has rock type RT1, the probability that will be correctly labeled by the classifier is 0.81 (recall). The F1 Score combines both to give a single measure ofrelevancy of the classifier results [1].

Confusion matrix of RFC, LR and KNN classification models are displayed below(Figure 10, 11 and 12). As we can see on "Figure 9 and 11", the SVM and KNN classification model show the best performance in the training data.

RFC F1-score: 0.931

Pred	RT1	RT2	RT3	RT4	RT5	RT6	Total
True							
RT1	16	4	1				21
RT2	1	25	6	3			35
RT3	2	5	71	12	1		91
RT4		3	9	40	4		56
RT5				8	76	3	87
RT6					1	607	608
Precision	0.84	0.68	0.82	0.63	0.93	1.00	0.93
Recall	0.76	0.71	0.78	0.71	0.87	1.00	0.93
F1	0.80	0.69	0.80	0.67	0.90	1.00	0.93

Figure 10. Confusion matrix of RFC classification model.

KNN F1-score: 0.936

Pred / True	RT1	RT2	RT3	RT4	RT5	RT6	Total
RT1	17	3	1				21
RT2	2	23	8	2			35
RT3	2	1	79	8	1		91
RT4		3	9	42	2		56
RT5			2	3	74	8	87
RT6					2	606	608
Precision	0.81	0.77	0.80	0.76	0.94	0.99	0.94
Recall	0.81	0.66	0.87	0.75	0.85	1.00	0.94
F1	0.81	0.71	0.83	0.76	0.89	0.99	0.94

Figure 11. Confusion matrix of KNN classification model.

LR F1-score: 0.886

Pred / True	RT1	RT2	RT3	RT4	RT5	RT6	Total
RT1	12	6	3				21
RT2	2	17	13	3			35
RT3	2	2	75	10	2		91
RT4	1	1	17	25	12		56
RT5			1	12	64	10	87
RT6					3	605	608
Precision	0.71	0.65	0.69	0.50	0.79	0.98	0.89
Recall	0.57	0.49	0.82	0.45	0.74	1.00	0.89
F1	0.63	0.56	0.75	0.47	0.76	0.99	0.89

Figure 12. Confusion matrix of LR classification model.

VI. Applying the classification model to the new data

Since we now have trained models for classifying rocks, we may use them to identify different rock types in wells without core data. So, using the same set of well logs as input, we applied the classifiers to one well. We can use the well log plot to view the classification results along with the well logs(Figure 13).

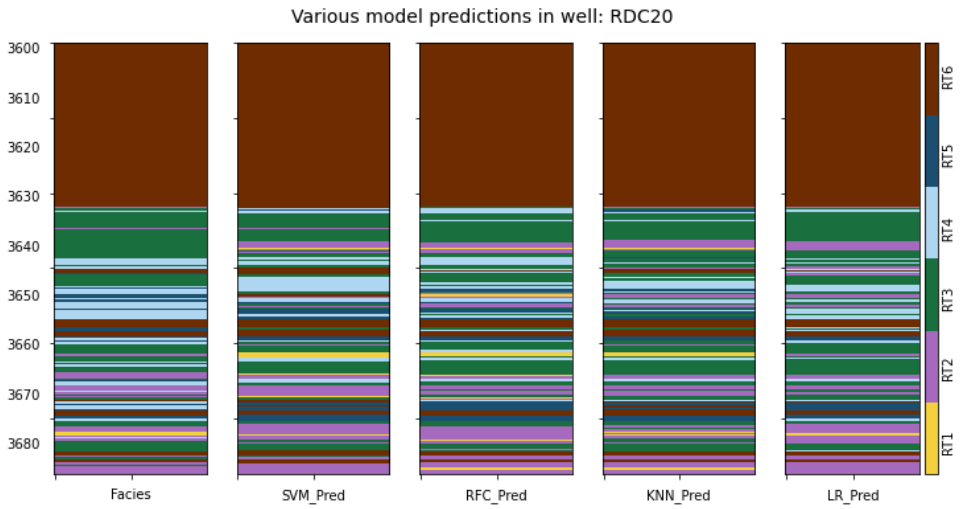

Figure 13. Classification model predictions for one well without core data.

As we can see on “Table 3” the RFC classification model shows the best performance in the test data. In the end, we can write out a CSV file with the well data and the rock types classification results. The obtained rock type classification has a good match with the core and production data. The good rock types are mainly located in the median and bottom part of the Trias which produced 5.13 m^3/h.

Model type	F1 SCORE
SVM	0.71
RFC	0.73
KNN	0.70
LR	0.69

Table 3. Model evaluation performance for one well without core data.

VII. Applying the classification model to the blind well

We choose one well out of the training data from training and model fitting process. So, we can plot the model’s prediction with one blind well data (Figure 14).

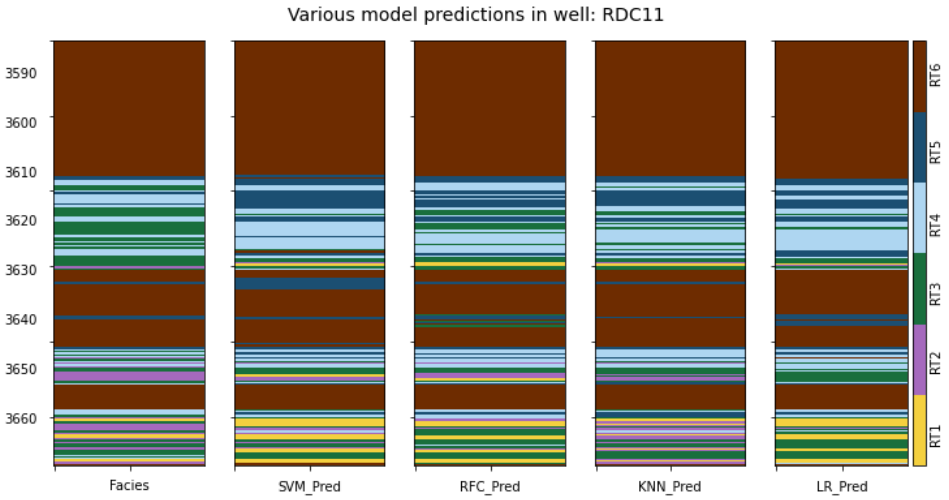

Figure 14. Classification model predictions for one blind well.

As we can see on “Table 4” the RFC classification model also shows the bestperformance in the blind data.

Model type	F1 SCORE
SVM	0.69
RFC	0.74
KNN	0.71
LR	0.68

Table 4. Model evaluation performance for one blind well.

In the end, we can write out a CSV file with the well data and the rock types classification results. The obtained rock type classification has a good match with the production data. The good rock types are mainly located in the bottom part of the Trias(T1a).

VIII. Conclusion

SVM and RFC show the better performance in test data.
All model performance decreases in the blind wells (Table 3 and 4), this can beassociated to a data's lack.
If we analyze the last cross section displaying several model output compared with real rock type distribution, RT3, RT5 and RT6 correlate properly. RT1 has the next bettercorrelation with real data. RT2 and RT4 display the weakest agreement with real Rock type distribution. The tendency cited above is in accord with data sample frequency which displayed inthe histogram. Rock types are predicted better if they are represented enough in the training data set.
To enhance the results, we should insert more data sample to models and enhancemodel parameter.

IX. References

[1] Del Monte A: Seismic Petrophysics: Part 1, The Leading Edge, 2015; 34(4): 440- 442. DOI: 10.1190/tle34040440.1.

[2] Sokhal A, Benaissa, Ouadfeul, S A, Boudella A: Dynamic rock type characterization using artificial neural networks in hamra quartzites reservoir: a multidisciplinary approach. Engineering, Technology and Applied Science Research.2019; 9: 4397-4404. DOI: 10.48084/etasr.2861.

[3] Hall, B: Facies classification using machine learning. The Leading Edge. 2016; 35(10): 906-909. DOI:10.1190/tle35100906.1.

[4] Ouadfeul S A, Aliouane L: Lithofacies classification using the multilayer perceptron and the self-organizing neural networks. In: Huang, T., Zeng, Z., Li, C., Leung, C.S. (eds) Neural Information Processing. Lecture Notes in Computer Science. Springer, Berlin, Heidelberg. 2012; 7667. DOI: 10.1007/978-3-642-34500-5_87.

[5] Moqbel A A, Wang Y: Carbonate reservoir characterization with lithofacies clustering and porosity prediction. Journal of Geophysics and Engineering, 2011; 8(4):592-598. DOI:10.1088/1742-2132/8/4/011.

[6] Hall, B: Facies classification using machine learning. The Leading Edge; 2016; 35(10): 906-909. Doi:10.1190/tle35100906.1.

[7] Ouadfeul S A. Unconventional Hydrocarbon Resources: Exploration and ArtificialIntelligence. 1st Edition. Wiley. 2023; 350p. DOI: under publication

[8] Zoulikha A, Maâmar D, Zenkhri R, Sokhal A, Lounis E, Aguenini A, ZandkarimiG, Mosher P, Xu D, Likrama F, Ahmed E: Conducting integrated reservoir studies in thequartzite hamra reservoir-tight oil, southern periphery of hassi messaoud field, Algeria. Annual convention and Exhibition, 2-5 April, 2017. Houston, Texas, United states.

[9] Sokhal A, Ouadfeul S A, Benmalek A: Rock type and permeability prediction using flow-zone indicator with an application to Berkine Basin (Algerian Sahara); 2016: 3068-3072. DOI: 10.1190/segam2016-13943527.1.