Machine Learning Algorithm for Estimating Oil Recovery Factor

Kachalla Aliyuda[1,2] and John Howell[1]
1. School of Geosciences, University of Aberdeen, United Kingdom
2. Department of Geology, Gombe State University, Gombe, Nigeria

Abstract

The methods used to estimate recovery factor change through the life-cycle of a field. During appraisal, prior to development when there are no production data we typically rely on analogue fields and empirical methods. Given the absence of a perfect analogue, these methods are typically associated with a wide range of uncertainty. During plateau, recovery factors are typically associated with simulation and dynamic modelling, while in later field life, once the field drops off the plateau, decline curve analysis is also used. The utilisation of different methods during different stages of the field life leads to uncertainty and potential inconsistencies in recovery estimates. A wide range of interacting, partially related, reservoir and production variables control production and recovery factor. Machine learning allows more complex multivariate analysis that can be used to investigate the roles of these variables using a training dataset and then ultimately predict future performance in fields. To investigate this approach we used a data set consisting of producing reservoirs all of which are at plateau or in decline to train a series of machine learning algorithms that can potentially predict recovery factor with minimal percentage error. The database for this study consists of categorical and numerical properties for 93 reservoirs from the Norwegian Continental Shelf. Of these, 75 are from the Norwegian Sea, the Norwegian North Sea, and the Barents Sea, while the remaining 18 reservoirs are from the Viking Graben in the UK sector of the North Sea. The data set was divided into training and testing sets, the training set comprised about 80% of the total data while the remaining 20% was the testing set. Linear regression models and a support vector machine (SVM) models were trained with all parameters in the dataset (30 parameters), then with 16 most influential parameters in the dataset, the performance of these models was compared from results of 5-fold cross-validation. SVM training using a combination of 16 geological/engineering parameters models with Gaussian kernel function has root mean square error of 0.12, mean square error of 0.01 and R-squared of 0.76. This model was tested on 18 reservoirs from the testing set, the test results are very similar to cross-validation results during models training phase, suggesting that this method can potentially be used to predict future recovery factor.

INTRODUCTION

The recovery factor is the ratio between the volume of hydrocarbon initial in place in a reservoir and the volume that is recovered. It depends on a range of reservoir and engineering properties including reservoir geology (stratigraphy and structure), fluids properties as well as a series of engineering parameters such as number of wells, completion strategy, production mechanism etc. The correct estimation of recovery efficiency (or recovery factor) is a highly important part of the decision-making process at numerous stages in the life of an asset as it determines the overall value of a field.

Recovery factor is a key component of the recoverable reserve's calculation. There are different methods of calculating or estimating recovery factor, these methods include statistical, volumetric, field analogy, use of decline curve, material balance, simulation and empirical methods. The choice of a method depends on the life stage of the field (appraisal, development, production) and the availability of data. During early field life, when there is no production data, estimates are derived from using analog and empirical methods. Empirical equations for solution gas drive reservoirs such as Arps et al. (1967) and Guthrie and Greenberger (1995) for water-drive reservoirs were found to be inadequate to predict recovery factor accurately (Sharma and Srinivasan, 2010). Arps (1955) suggests that recovery increases with oil gravity, Muskat (1946) conclude that a decrease in field's ultimate recovery is related to increasing viscosity. Many empirical and analog methods are based on play or basin specific and do nothing to shed light on the fundamental controlling parameters. Consequently, empirical methods result in large uncertainties ranges. Analog methods are also limited because no two fields are identical and given that recovery factor is a function of a large number of variables it is not possible to predict production from one reservoir by simply looking at another.

Later in field life, once there is historical production data decline curve analysis is commonly used. This is limited because it does not consider future interventions such as infill wells or enhanced oil recovery and also because there is significant uncertainty associated with assigning appropriate decline functions.

Machine learning provides a potential to bridge the gap between empirical and analog methods by considering a much wider range of variables and controls on production. The proposed method can be applied at any stage of the life cycle and can also be used to improve our understanding of what the fundamental controls on recovery are. Machine learning is a computational method of learning from existing data directly without relying on a predetermined equation as a model but using algorithms that look for natural patterns in the data. Artificial Neural Network is one of the common machine learning methods that has been applied in solving similar problems in the oil industry. They are currently used in predicting oil flow rates (Mirzaei-Paiaman and Salavati, 2012), facies identification (Naeini and Exley, 2017), seismic data processing, seismic quantitative interpretation (Liu, 2017), Ahmadi et al (2015) used artificial neural network linked to particle swarm optimization tools to predict the performance of horizontal oil wells under pseudo-steady state conditions, and many others. Support vector machine, neural networks, bagged decision trees, stepwise regression are some of the powerful machine learning algorithms commonly used.

We used a data set of 93 producing reservoirs which are all at plateau or in decline to train a series of machine learning algorithms that attempt to predict recovery factor with minimal percentage error. The analysis includes estimates for the final recovery factor made by the field operators and the state regulator (Norwegian Petroleum Directorate).

DATA AND METHODS

To test the methods and to train an algorithm to estimate recovery factor, a database was built specifically for this study. That database consists of categorical and numerical properties for 93 fields from the Norwegian Continental Shelf (NCS) (Fig. 1). There are 30 variables available for each reservoir entry. These variables include factors such as depositional environment, reservoir depth, average porosity, average permeability, initial pressure, initial temperature, net to gross ratio, number of faults compartments, number of faults population, structural complexity, API gravity, pay area, bulk rock volume, well density, number of production wells, number of injection wells, average gas oil ratio (GOR), reservoir thickness, water saturation, production strategy, trap type, original oil in place, diagenetic impact, paleoclimate, stratigraphic heterogeneity, well spacing, total number of wells, maximum well rate, average monthly depletion rate and cumulative oil produced. (Table 1, 2).

The parameters in Table 1 were compiled for all of the fields from publicly available literature. The data were then pre-processed which included normalization and dealing with missing values. All missing values were replaced with NaN. To normalize the dataset, the following formula was used;

$$Y = \frac{(x - \min(d)) * (\max(n) - \min(n))}{\max(d) - \min(d)} + \min(n) \qquad . \quad . \quad . \quad . \quad . \quad . \qquad 1$$

Where,

Min(d)= Minimum or lowest value in data

Max(d)= Maximum or highest value in data

Min(n)= Minimum value in new range

Max(n)= Maximum value in new range

All entries in the database were normalized to the range 0 to 1. Categorical variables such as gross depositional environment, trap type, stratigraphic heterogeneity, and production strategy were converted to discrete numerical values (Table 2).

The data set was divided into a training and testing subsets, the training set comprised about 80% of the total data (75 reservoirs) while the remaining 20% was the testing set (18 reservoirs).

Given that outliers might affect the accuracy of the regression models, box plots were used to graphically investigate observations that were considerably different from the majority of the data set (Fig. 2), such observations were removed. Parameters in the database with outliers are; original oil in place, permeability, porosity, pressure, and bulk rock volume. The complete workflow used is presented in figure 4.

To quantify stratigraphic heterogeneity, a scale of 0 - 8 was used taking into account the vertical and lateral heterogeneity of a given reservoir's depositional sub-environment following the method of Tyler and Finley (1991), also summarized in Manzocchi et al (2008). In this scheme Zero is a reservoir with low vertical and lateral heterogeneity while 8 (eight) is a reservoir with high vertical and lateral heterogeneity (Fig. 3). All the numbers in the scale and their corresponding depositional sub-environments are defined below;

0 = Low vertical, low horizontal heterogeneity (**wave dominated proximal delta, barrier island**)

1 = Low vertical, medium horizontal heterogeneity (**Eolian deposit, wave modified shoreface**)

2 = Low vertical, high horizontal heterogeneity (**submarine fan, turbidites, mass complexes**)

3 = Medium vertical, low horizontal heterogeneity (**proximal delta front, mouth bar, tidal deposits**)

4 = Medium vertical, medium horizontal heterogeneity (**alluvial fan, distal delta front, lacustrine delta**)
5 = Medium vertical, high horizontal heterogeneity (**Fluvial meander belt, braid delta**)
6 = High vertical, low horizontal heterogeneity (**meander fluvial single unit bar, back barrier, fluvially dominated delta**)
7 = High vertical, medium horizontal heterogeneity (**braided river deposit, tide dominated delta**)
8 = High vertical, high horizontal heterogeneity (**stacked fluvially dominated delta and meander belt, stacked back barrier deposit**)

RESULTS/DISCUSSION

Multivariate linear regression

Multivariate variables linear regression model takes the form

$$y_{ik} = b_{0k} + \sum_{j=1}^{p} b_{jk} X_{ij} + e_{ik} \quad . \quad . \quad . \quad . \quad . \quad . \qquad (2)$$

Where

y_{ik} = is the real-valued response for the *i*th observation

b_{0k} = is the regression intercept for the *k*th response

b_{jk} = is the *j*th predictor's regression slope for *k*th response

X_{ij} = is the *j*th predictor for the *i*th observation

$(e_{i1}, \ldots, e_{im}) \sim^{iid} N(0_m, \Sigma)$ is a Multivariate Gaussian error vector

$i \in \{1, \ldots, n\}$ and $k \in \{1, \ldots, m\}$

$p > 1$ predictors

The model is expressed as a linear combination of the predictors with *p* parameters. Most of the predictors have a weak correlation with recovery factor, however, there are combinations of these predictor variables which may yield a better prediction for recovery factor. Two sets of linear regression models were trained. 1) Using all 30 parameters in the database, and 2) Using 16 selected parameters with strong control on recovery factor. Principal component analysis (PCA) (Fig. 5, 6 and 7) and simple cross plots were used to identifying the 16 selected variables, they were; depositional environment, reservoir depth, average porosity, average permeability, initial pressure, temperature, structural complexity, API gravity, bulk rock volume, well density, stratigraphic heterogeneity, the total number of wells, maximum reservoir rate, cumulative oil produced, average monthly depletion rate and original oil in place.

For the two linear regression models trained, the model trained using all the 30 parameters in the database has a root mean square error (RMSE) of 0.35 and R- Squared of -1.37, while the model trained using only the 16 most influential parameters has RMSE of 0.21 and R-Squared of 0.26 (Table 3). This shows that the model trained with 16 influential parameters has less error and better accuracy. This is because the combination of PCA and cross-plots allowed redundant dimensions to be identified and removed thereby improving the performance of the model.

Support vector regression

Support vector machine (SVM) is a machine learning algorithm that uses hyperplanes in a high dimensional space, which can be used for both regression and classification of models. It is a linear algorithm developed initially to solve pattern recognition, which was extended to solve non-linear problems (Vapnik 1982). To generate a non-linear function input vectors *x* are transformed into vectors *Φ(x)* of a higher dimensional space (Cristianini and Shawe-Taylor 2000). Commonly used kernel functions are listed here with their mathematical expressions;

Linear kernels $K(x, y) = xTy$ . . . . (3)

Polynomial kernels $K(x, y) = (e + xTy)d$ . . . . (4)

Gaussian kernels $Kc(x,y) = e{-}\|x-y\|2/e$ . . . . (5)

Hyperbolic Tangent (Sigmoid) kernels $K(x,y) = \tanh(\alpha xT, y + c)$. (6)

Three sets of SVM models were trained with different sets of variables drawn from the 16 most influential parameters; 1) only geological parameters (depositional environment, reservoir depth, average porosity, average permeability, net to gross, structural complexity, stratigraphic heterogeneity); 2) Only engineering parameters (initial pressure, initial temperature, API gravity, bulk rock volume, well density, number of production wells, number of injection wells, water saturation, production strategy, well spacing, total number of wells, original oil in place, average monthly depletion rate, maximum reservoir rate, cumulative oil produced) and, 3) a combination of both geological and engineering parameters. A comparison among these models shows that the models trained with both geological and engineering parameters control the recovery factor (Fig. 8).

The model was used to predict the recovery factors of 18 reservoirs and the results were compared with recovery factors estimated by the operators and the Norwegian Petroleum Directorate (NPD) (Fig. 9, Table 4). All the SVM models were trained using the Gaussian kernel, with a kernel scale of 5.1 and 5fold cross-validation. Cross-validation is a resampling procedure used in evaluating machine learning models. In 5fold cross-validation, the training dataset was split into 5 equal groups, one of the groups was held back as a test set while the remaining groups were used as a training set to fit the model and evaluate the model on the test set. This process was repeated with each of the five groups being used as a test set, the use of 5fold cross validation was to restrict against model over-fitting.

The two major categories of variables in the dataset (geological and engineering variables) were used to train 3 set of models in order to investigate how the different categories of variables effect recovery factor. Results of these iterations (Fig. 8) suggest that a combination of both geological and engineering parameters yields the best predictive model.

Table 4 shows the results of a blind test conducted using the trained SVM model on 18 test fields from the Norwegian North Sea and Viking graben. There is a good correlation between the recovery factor predicted by SVM and the recovery factor estimated using more detailed analysis on a field by field basis (Fig. 9). Despite the absence of some typical engineering parameters such well head pressure, choke, injection rate and nature of injected liquid which are not publicly available, 80% of the predicted recovery factors matched closely with the volumetrically estimated recovery factor (table 4, Fig. 9).

These results suggest that the SVM model trained with 16 parameters can be used as an alternative method of estimating recovery factor for fields during the appraisal and production stages. This research has highlighted the importance of some geology dependent variables such as depositional environment, porosity, permeability and stratigraphic heterogeneity as influential predictors in controlling predicted recovery factor. The results also illustrate that both engineering and geological parameters effect production. This is significant because whilst geological parameters are “fixed” many engineering parameters can be controlled by the operator and therefore offer scope for improved recovery.

CONCLUSION

In this study a unique database that was developed to evaluate the role of geology in controlling oil field performance, was used to train and test linear regression algorithms and support vector machine algorithms. Once trained the algorithms were used to estimate the recovery factor in a blind test on a subset of the test dataset and could be applied on assets at various stages of field life. The method is applicable for reservoirs developed with different production strategy (Water injection, gas injection and pressure depletion).

Linear regression models and a support vector machine models were trained with all of the 30 parameters and independently with a subset of 16 of the most influential parameters. The performance of these models was compared from results of 5fold cross-validation. This led to 14 redundant parameters being removed from the predictor variable using principal component analysis and cross plots. Models trained without the redundant parameters, include a combination of geological and engineering factors. Removing the redundant parameters resulted in better accuracy and less error. The SVM model with Gaussian kernel function trained using 16 parameters have RMSE of 0.12, mean squared error of 0.01 and R-squared of 0.76. The model was used to predict recovery factor on 18 reservoirs from the test data, the results were similar to the cross-validation results obtained from the training set.

ACKNOWLEDGEMENT

This work has been supported by SAFARI III and the Petroleum Technology Development Fund (PTDF), we are grateful to them for providing funds for this project.

REFERENCE

Ahmadi Ali Mohammed, Reza Soleimani, Moonyong Lee and Tomoaki Kashiwao., 2015., Determination of oil well production performance using artificial neural network (ANN) linked to the particle swarm optimization (PSO) tool., Petroleum., Vol. 1., pp 118-132.

Arps , J. J., 1955., The effect of the relative permeability ratio, the oil gravity, and the solution gas-oil ratio on the primary recovery from a depletion type reservoir. Petroleum transaction, AIME. 204. 120-127.

Arps, J.J., Brons, F., van Everdingen, A.F., Buchwald, R.W. and Smith, A.E. 1967. A statistical study of recovery efficiency. API, Bulletin D14, 1-27.

Craze. R. C and Buckely S. E., 1945., A factual analysis of the effect of well spacing on oil recovery. In drilling and production practice: American Petroleum institute. 144.

Cristianini C. and Shawe-Taylor J. 2000. An Introduction to Support Vector Machines and other Kernel-based learning methods. Cambridge, Uk: Cambridge University press.

Guthrie, R.K. and Greenberger, M. H. 1955. The use of multiple correlation analysis for interpreting petroleum engineering data. American Petroleum Institute, *Drilling and production Practice*, 130.

Sharma A. and Srinivasan S., 2010., Classification of oil and gas reservoirs based on recovery factor: A data-mining Approach. SPE Annual Technical conference and exhibition, Forence, Italy. 2010. SPE 130257. 1- 21.

Liu Yexin., 2017., Application of Machine Learning for Seismic Quantitative Interpretation., Geoconvention, Canada, 2017. 1-5.

Manzocchi, T., Carter, J.N., Skorstad, A. et al. 2008. Sensitivity of the impact of geological uncertainty on production from faulted and unfaulted shallow-marine oil reservoirs: objectives and methods. Petroleum Geoscience, 14, 3–15.

Mirzaei-Paiaman A. and Salavati S., 2012., The Application of Artificial Neural Networks for the Prediction of Oil Production Flow Rate, Energy Sources, Part A: Recovery, Utilization, and Environmental effects., Taylor & Francis., 34:19, 1834-1843.

Muskat, M. and Taylor, M. O., 1946., Effect of reservoir fluids and characteristics on production histories of gas-driven reservoirs., AIME. 165, 78.

Naeini Ehsan Zabihi and Russell Exley., 2017., Quantitative interpretation using facies-based seismic inversion., Interpretation, Vol. 5, No. 3 pp. 1-8.

Tyler, N. & Finlay, R.J. 1991. Architectural controls on the recovery of hydrocarbons from sandstone reservoirs. In: Miall, A.D. & Tyler, N. (eds) The Three Dimensional Facies Architectures of Terrigeneous Clastic Sediments and its Implications for Hydrocarbon Discovery and Recovery. SEPM Concepts in Sedimentology and Palaeontology, 3, 1–5.

Vapnik V. N. 1982. Estimation of Dependences Based on Empirical Data. Berlin, Germany: Springer series in Statistics, Springer.

LIST OF FIGURES AND TABLES

Figure 1: Map of the Norwegian continental shelf showing different wells and field names in the northern North Sea, Norwegian Sea and the Barents Sea. Field names in red have continental reservoir, shallow/parallic reservoirs in yellow and deep marine reservoirs in green.

Figure 2: Box plot of **A**) average porosity and bulk rock volume, **B**) average permeability, pressure and original oil in place showing outliers.

Figure 3: Depositional heterogeneity and flow unit diagram showing heterogeneity scale used for all the different Sub-environments, 0 = means low vertical and horizontal heterogeneity, highly connected sand reservoir with no clay or shale barriers. 8 = extremely heterogeneous,

low net to gross reservoir sand bodies are isolated within clay or shale intervals (Modified after Tyler and Finley, 1991).

Figure 4: Workflow used to train linear regression and SVR models.

Figure 5: Principal components analysis results, the scree plot shows that most variance in the dataset is explained by principle components (PC) 1-7 with an eigenvalue of 1 and above, it explained more than 90% variability of the data set.

Figure 6: The eigenvectors of all the 30 parameters in the first principal component, the most important parameters are highlighted in blue.

Figure 7: The eigenvector plot of the second principal component, also in blue are the most important parameters.

Figure 8: A comparison of the statistic (RMSE, R- squared and MSE) of three models trained with a set of 16 geological, engineering and geological/engineering parameters.

Figure 9: Performance of the SVM model trained using the 16 selected parameters on a testing data set. The results show a fairly good correlation between the predicted recovery factor and the volumetrically estimated recovery factor in the database.

Table 1: Description of all numeric parameters in the database

Table 2: Categorical parameters in the database

Table 3: Comparison between Linear regression models and SVM models both trained with 30 parameters and 16 parameters from the train data set. The best model is SVM models trained 16 selected predictor variables.

Table 4: Recovery factor predicted by the trained SVM model compared to the actual recovery factor calculated using the volumetric method. Errors colored green for less than 0.015 and yellow 0.15-0.29. Recovery factor is both in fraction and percentage.

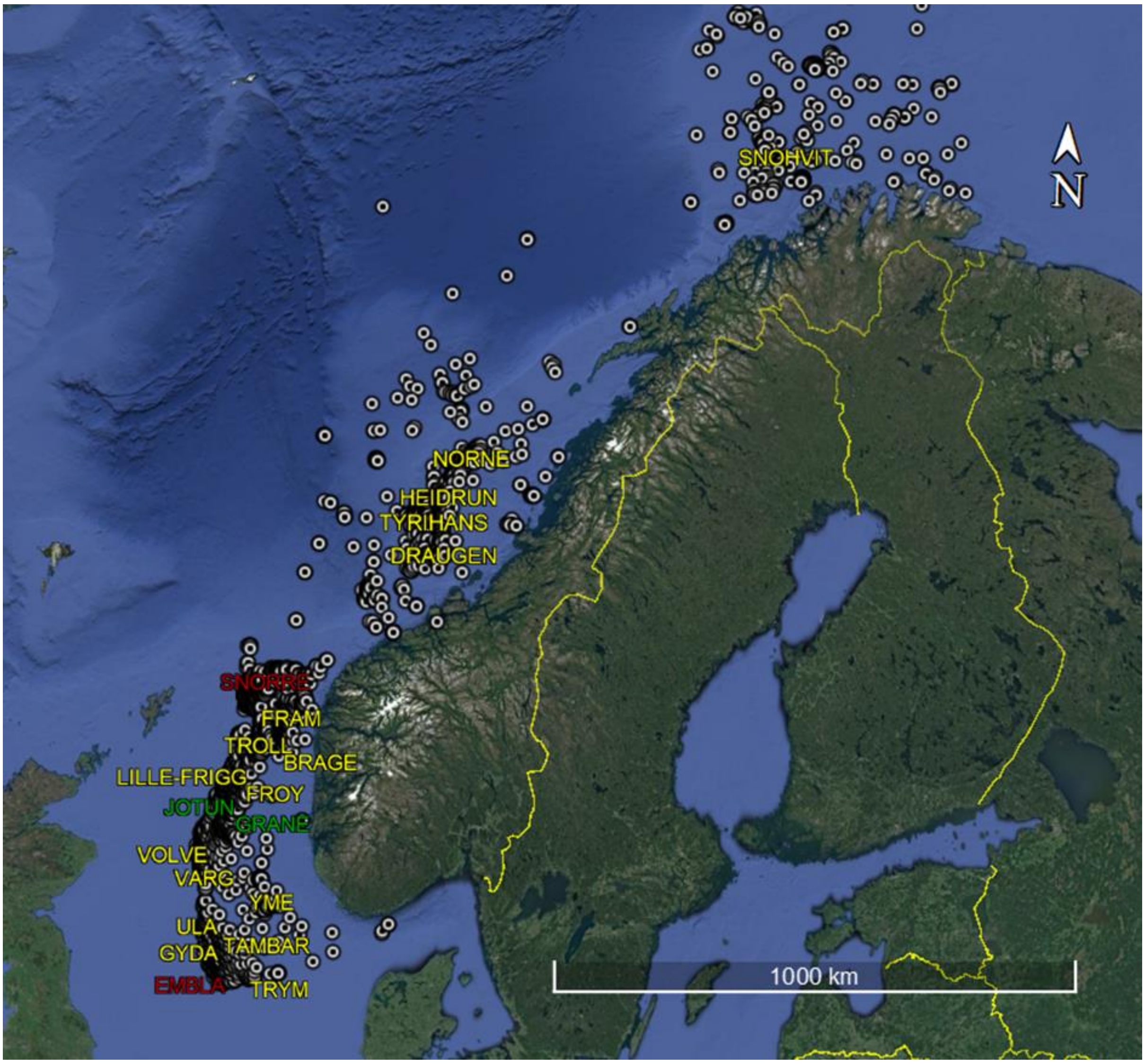

Figure 1: Map of the Norwegian continental shelf showing different wells and field names in the northern North Sea, Norwegian Sea and the Barents Sea. Field names in red have continental reservoir, shallow/parallic reservoirs in yellow and deep marine reservoirs in green.

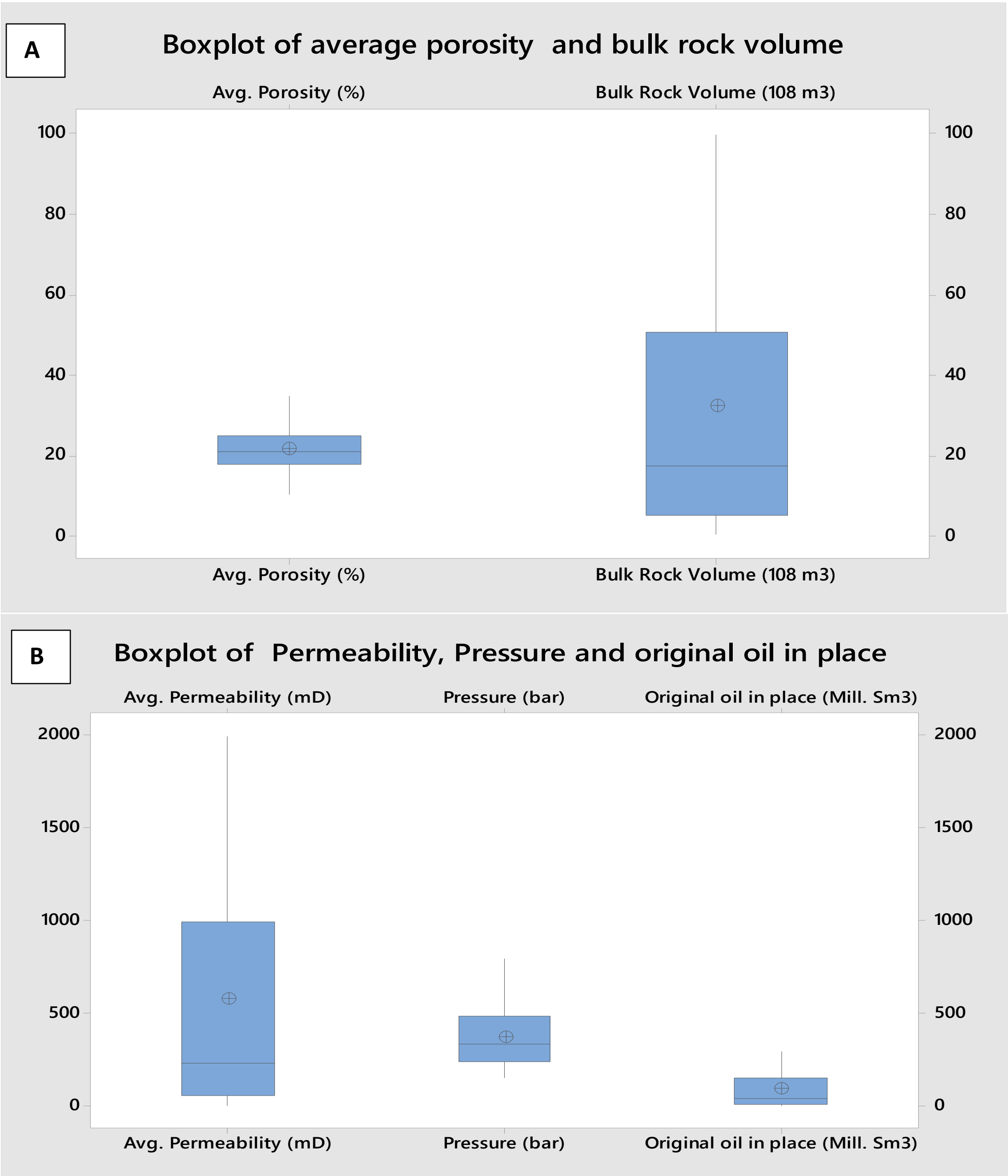

Figure 2: Box plot of **A**) average porosity and bulk rock volume, **B**) average permeability, pressure and original oil in place showing outliers.

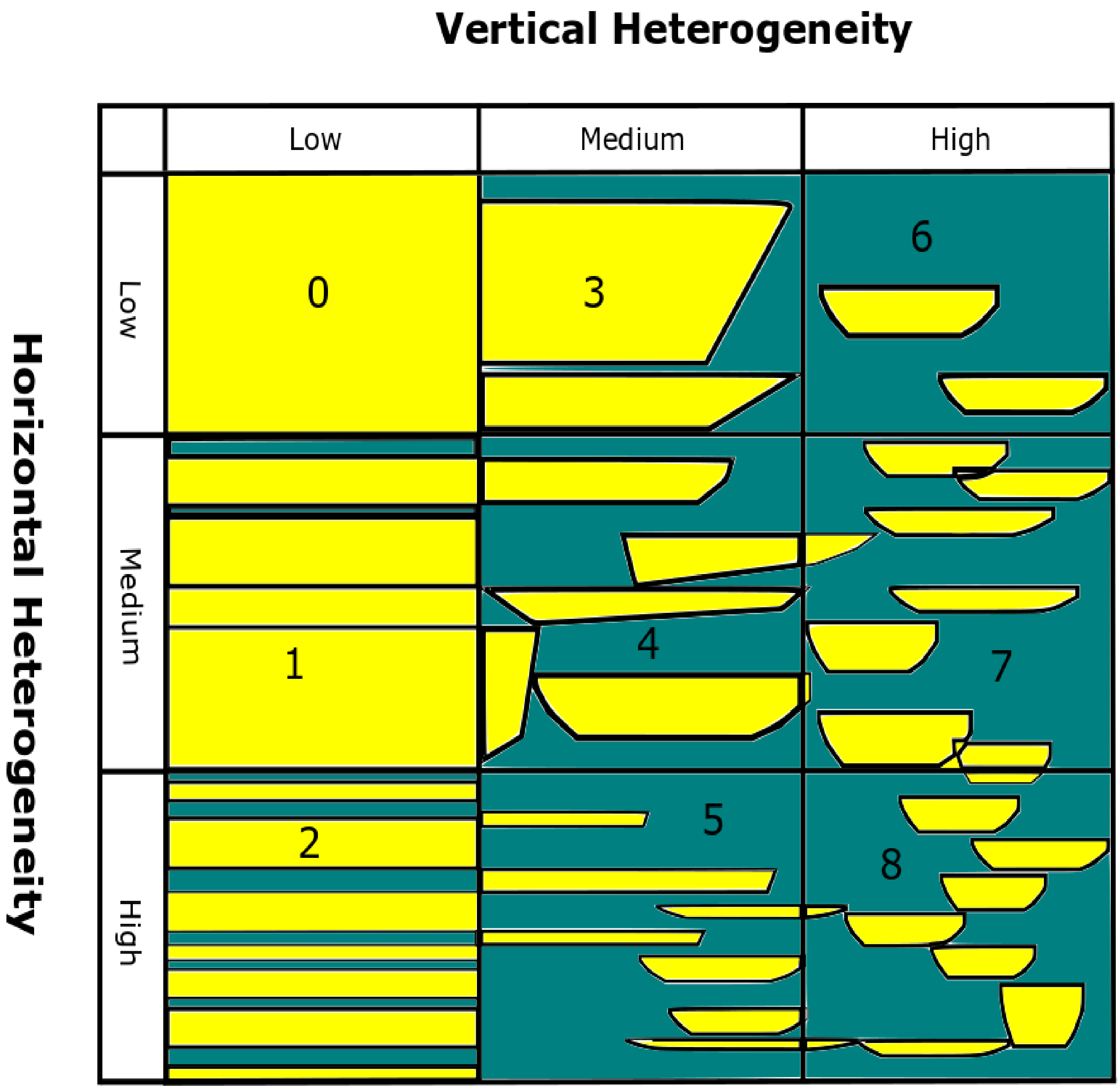

Figure 3: Depositional heterogeneity and flow unit diagram showing heterogeneity scale used for all the different Sub-environments, 0 = means low vertical and horizontal heterogeneity, highly connected sand reservoir with no clay or shale barriers. 8 = extremely heterogeneous, low net to gross reservoir sand bodies are isolated within clay or shale intervals (Modified after Tyler and Finley, 1991).

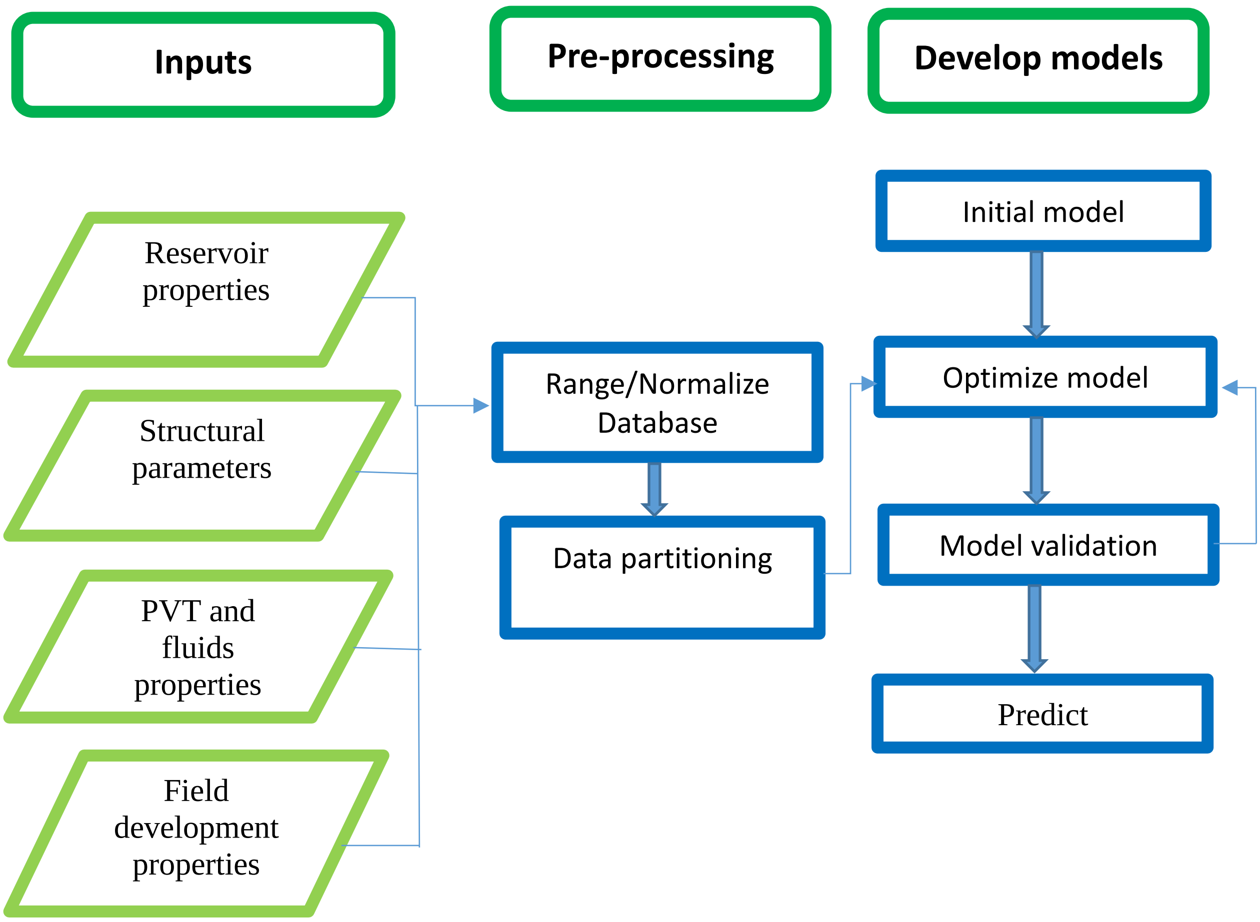

Figure 4: Workflow used to train linear regression and SVR models

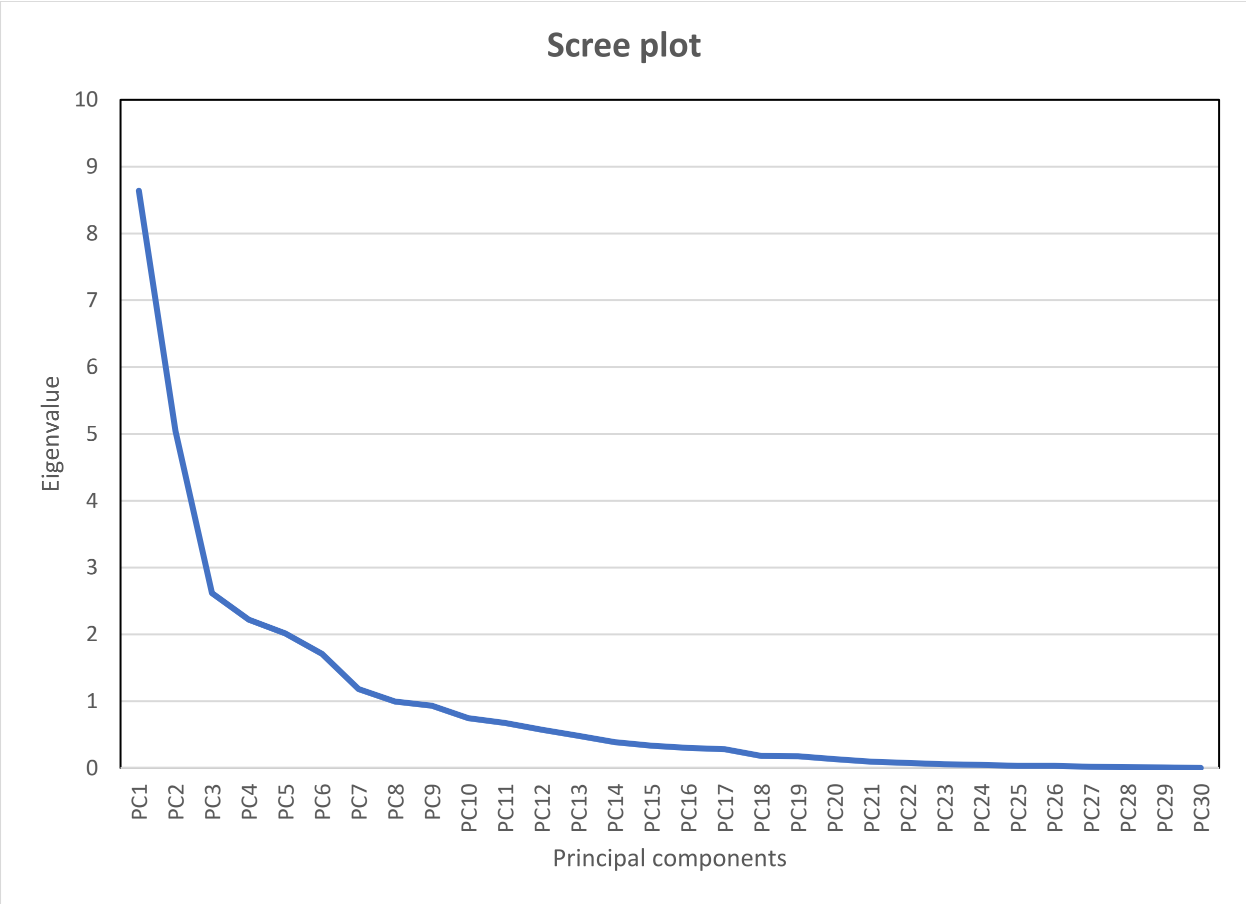

Figure 5: Principal components analysis results, the scree plot shows that most variance in the dataset is explained by principle components (PC) 1-7 with an eigenvalue of 1 and above, it explained more than 90% variability of the data set.

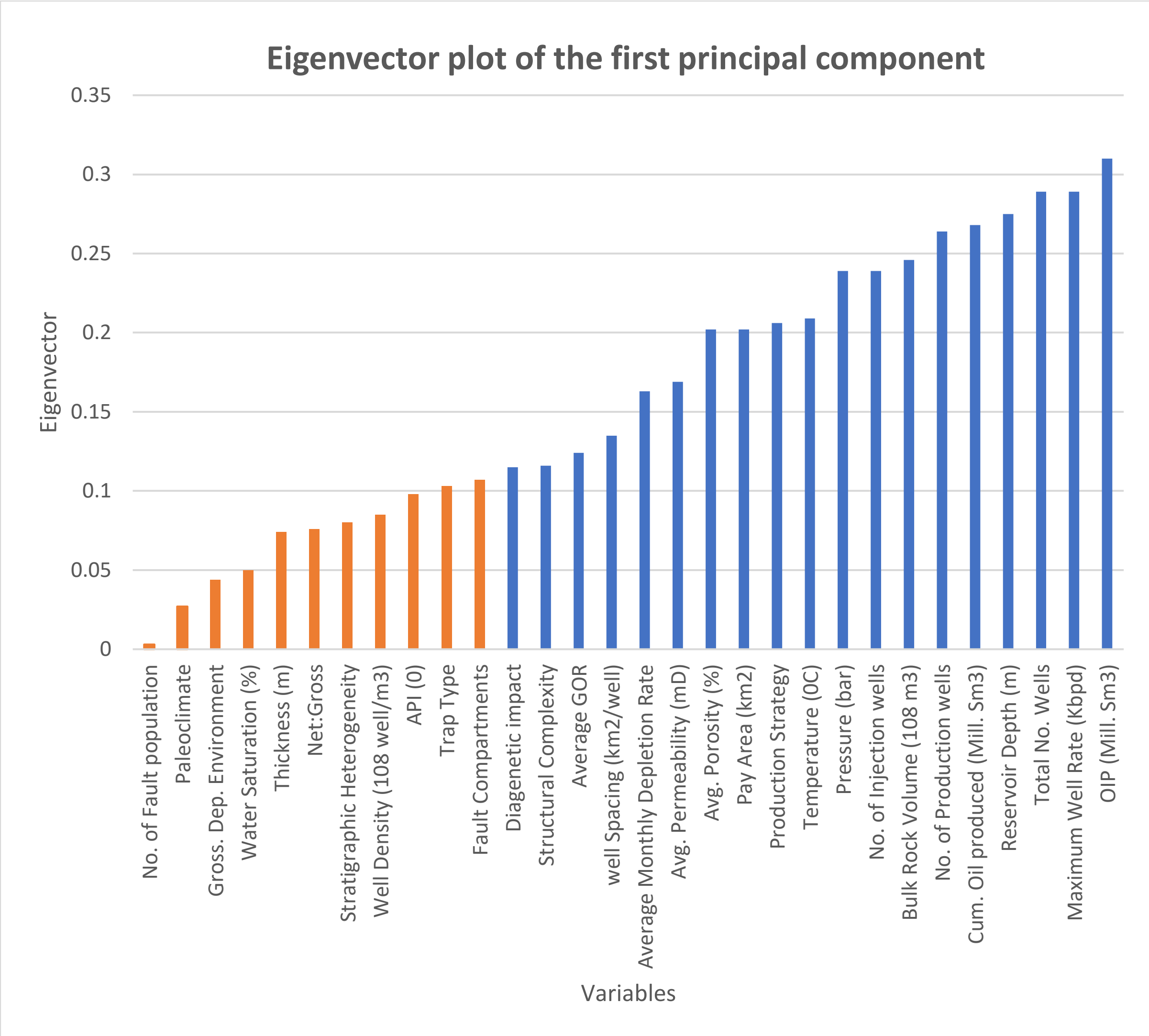

Figure 6: The eigenvectors of all the 30 parameters in the first principal component, the most important parameters are highlighted in blue.

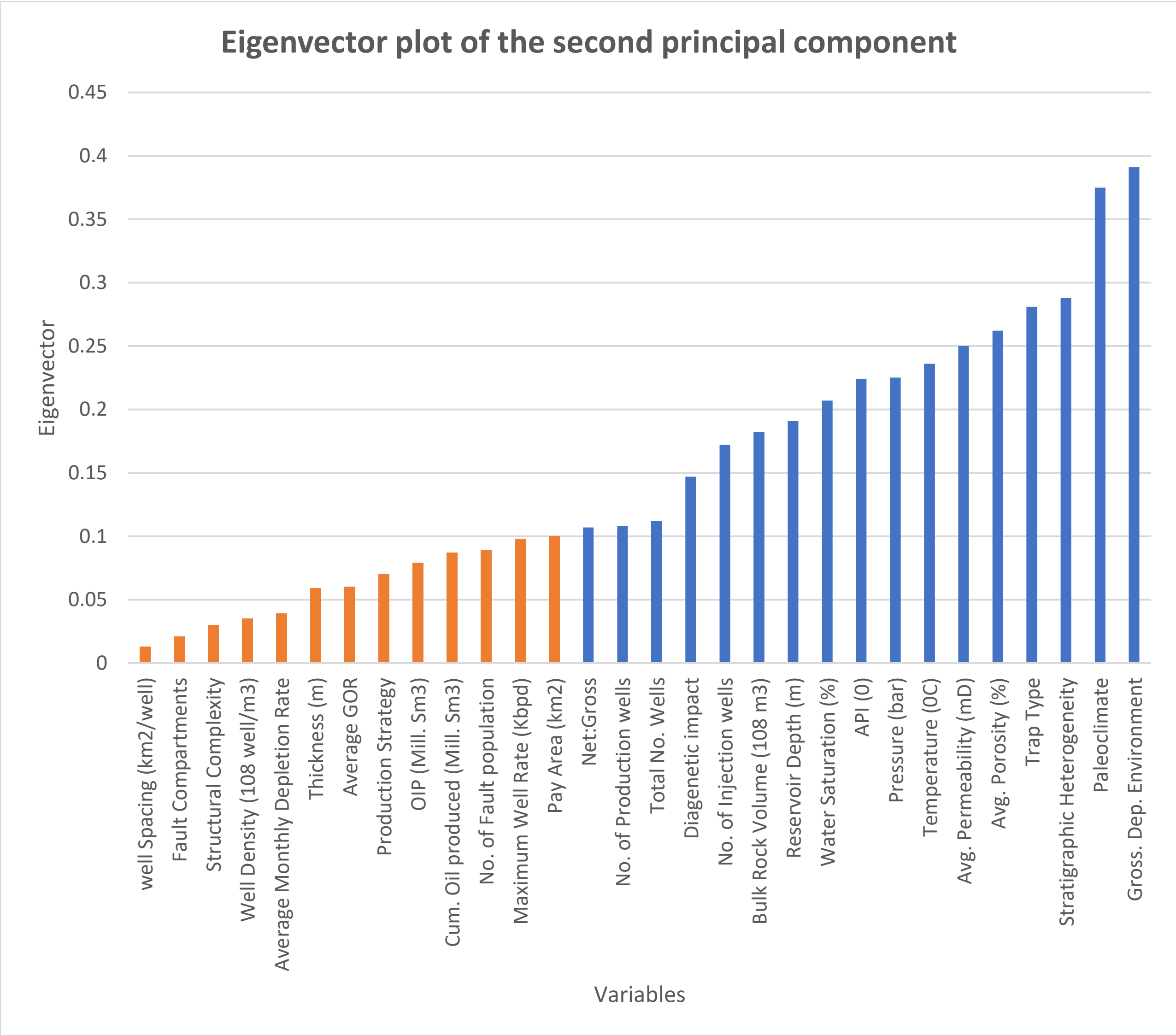

Figure 7: The second principal component also in blue are the most important parameters.

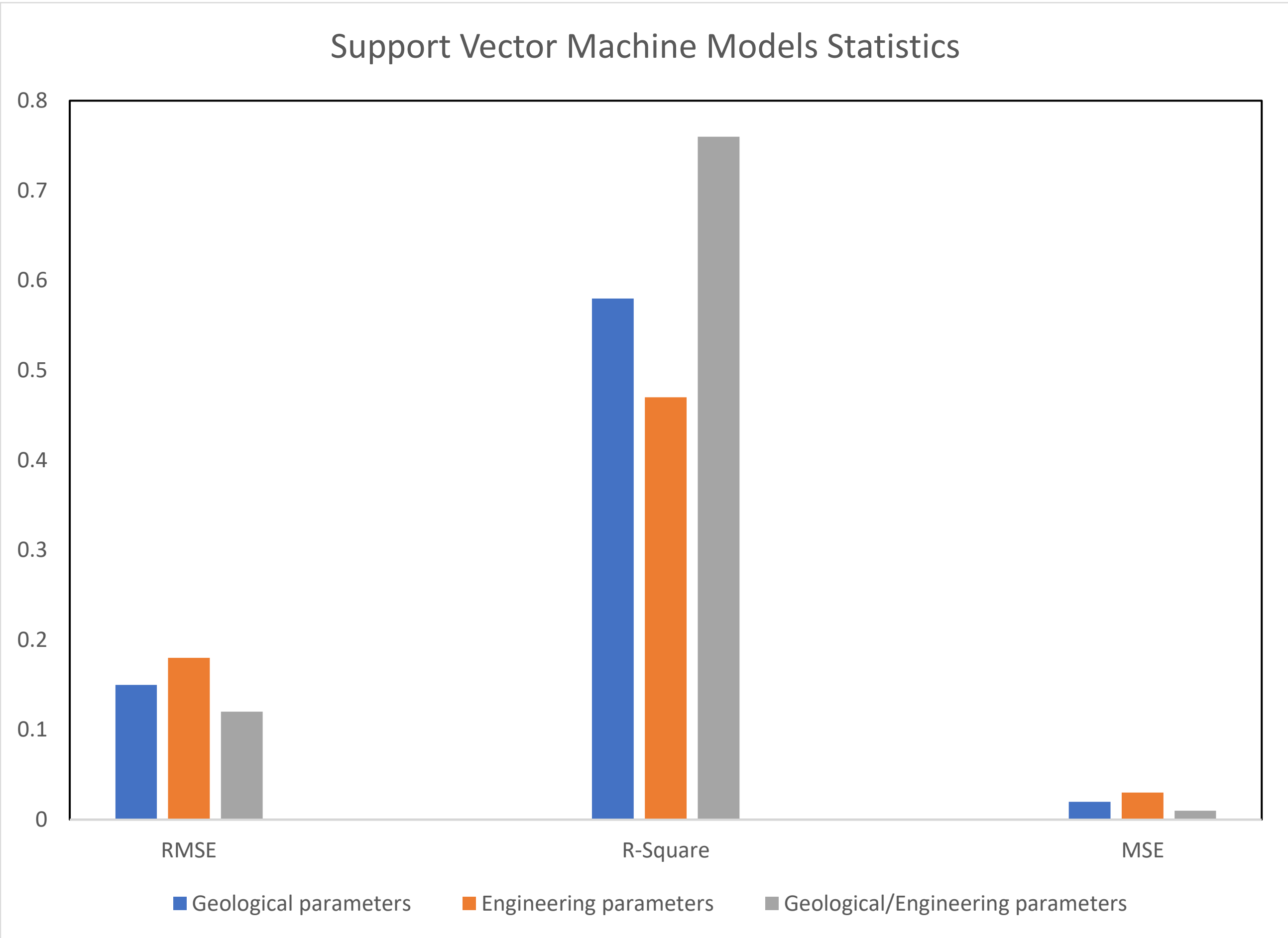

Figure 8: A comparison of the statistic (RMSE, R- squared and MSE) of three models trained with a set of 16 geological, engineering and geological/engineering parameters.

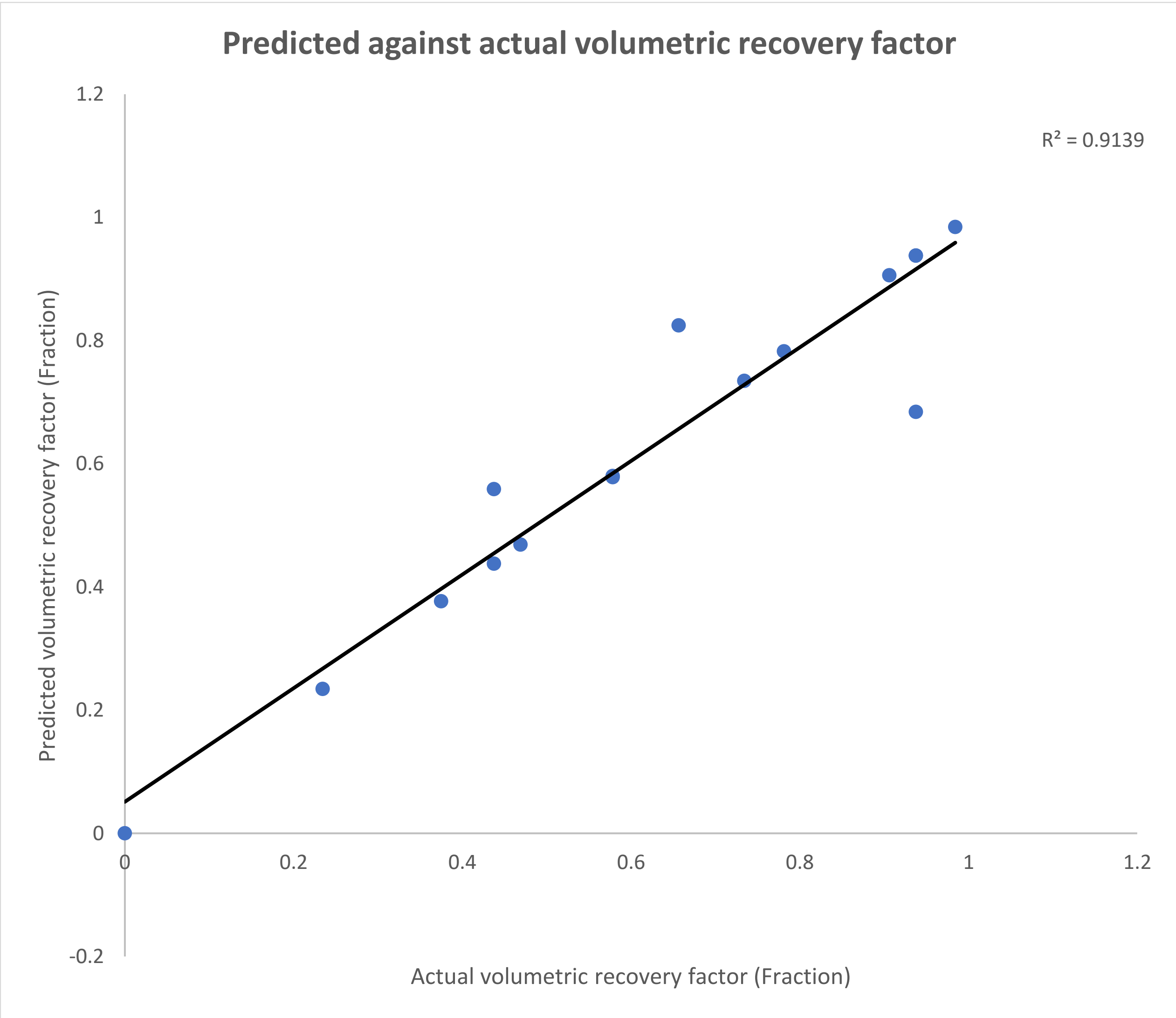

Figure 9: Performance of the SVM model trained using the 16 selected parameters on a testing data set. The results show a fairly good correlation between the predicted recovery factor and the volumetrically estimated recovery factor in the database.

Parameter	Description
Porosity (%)	The volume of void space relative to entire rock volume (from log/core).
Permeability (mD)	A measure of fluid flow within rock pores taken (from core).
Reservoir depth (m)	Reservoir present depth of burial (from completion log).
Pay area (km^2)	Area of a porous and permeable section of the reservoir within hydrocarbon accumulation (from completion log).
API gravity (^{0})	Average value of oil API density (from a well testing report).
Initial temperature (^0C)	Value of absolute reservoir temperature (from well test).
Initial pressure (bar)	Value of fluid pressure within the pores of a reservoir.
Bulk rock volume (BRV)	The total volume of bulk reservoir rock calculated from area and thickness of the reservoir.
Original oil in place (Mill. Sm^3)	The total volume of oil stored in the reservoir prior to production (from discovery field reports).
Number of wells	A total number of both production and injection wells.
Water saturation (%)	Value of initial proportion of water within pay interval.
Well spacing (km^2/well)	Value of field area against a number of producing wells in the field.
Reservoir thickness (m)	This is the gross thickness of the stratigraphic interval considered as the reservoir bed.
Number of production wells	This is the total number of production wells in a given reservoir whose content is oil.
Number of injection wells	The total number of wells injecting fluids into the reservoir (gas, water, etc) in order to optimize production.
Net to gross ratio	This is the proportion or ratio of net reservoir that can store hydrocarbon and with enough porosity and permeability to allow hydrocarbon to flow, over the total rock thickness within a given reservoir interval. Measured in fraction from 0 to 1.
Well density Well/m^3)	The total number of wells in a field or reservoir divides by the bulk rock volume of the reservoir or field. It was measured in 10^{-8}well/m^3
Average gas/oil ratio (GOR)	This is the average of the ratio of gas that comes out of solution, to the volume of oil at standard condition.

Average monthly depletion rate	The fraction of recoverable oil produced at a given month for a reservoir or field. We calculated the depletion rate of each field from the onset of production to the end of 2017 for active fields, for shutdown fields we calculated it for the entire field life.
Maximum reservoir rate (Kbpd)	This is the maximum amount of oil produced in a given field or reservoir per day in thousand barrel of oil per day.
Cumulative oil produced (Mill. Sm^3)	The total amount of oil produced from the reservoir or field up to the end of 2017 for active fields, the entire amount of oil produced for shutdown fields or reservoir (in million-meter cube).

Table 1: Description of all numeric parameters in the database

Parameter	**Description**
Depositional environment	Specific environments of sediment deposition, reservoirs were classified further into Depositional Environments (DE) and Sub-environment (SE) using SAFARI classification Schema. In the database 0.0 = Deep marine, 0.5 = Paralic/shallow marine and 1 = Continental
Stratigraphic heterogeneity	A measure of aerially extensive architecturally bounding surfaces that compartmentalize the reservoirs. A scale of 0-8 was used with 0 = Very low heterogeneity and 8 = Extremely heterogeneous (Fig. 3)
Diagenetic impact	Negative impact of reservoir sediments reconstitution and/or rearrangement resulting in a reduction of porosity and permeability only. It is classified into low, moderate or high impact. 0 = Low, 0.5 = Moderate, and 1 = High.
Trap type	Reservoir trap type: stratigraphic, structural and stratigraphic/structural. 0 = Structural trap, 1 = Structural/Stratigraphic trap, and 2 = Stratigraphic trap. Label and one hot encoding were used to assigned integer numbers for this parameter.
structural complexity	Intricate nature of reservoir due to the prevalence of fault. A scale of 0 - 5 was used with 0 = showing no faults, 1 = Minor fault (less than reservoir throw), 2 =Minor faults (more than reservoir throw), 3 = Several faults with two compartments, 4 = complex (3-4 compartments) and 5 = Very complex (5 or more compartments).
Number of fault compartments	The number of reservoir compartments resulting from compartmentalization due to faults. Ranging from 0 to 8 for the reservoirs in the database.

Production strategy	Drive mechanism, four drive mechanisms are considered; 1) Pressure depletion, 2) Gas Injection, 3) Water Injection and 4) Gas/Water Injection. Label encoding, and one hot encoding were used to assigned integer numbers; 0 = Gas/Water injection, 1 = Gas Injection, 2 = Water Injection, and 3 = Pressure depletion.
Paleoclimate	Climate at the time a reservoir is deposited, three paleoclimates in the database are; arid/semi-arid = 0, polar = 1 and temperate = 2. Label encoding, and one hot encoding were used to assigned integer numbers for this parameter.
Number of fault population	Number of dominant faults pattern or trends in a given field or reservoir. This number is on a scale of 5 ranges from 0.1, 0.4, 0.6 0.8 and 1

Table 2: Description and scale of all categorical parameters in the database

Linear Regression Model			SVM Models	
	With all 30 Predictors	With 16 selected Predictors	With all 30 Predictors	With 16 selected Predictors
R-Square	-12	0.26	0.49	0.76
RMSE	0.83	0.21	0.16	0.12
MSE	0.68	0.04	0.03	0.01

Table 3: Comparison between Linear regression models and SVM models both trained with 30 parameters and 16 parameters from the train data set. The best model is SVM models trained 16 selected predictor variables.

Reservoirs	**Depositional Environment**	**Estimated RF (Fraction)**	**Predicted RF (Fraction)**	**Estimated RF (%)**	**Predicted RF (%)**	**Error**
Heather	Paralic/Shallow marine	0.4375	0.4375	31	31	0
Magnus	Deep Marine	0.578125	0.5782	40	40.0048	-7.5E-05
South Brae	Deep Marine	0.46875	0.4687	33	32.9968	5E-05

Staffa	Paralic/Shallow marine	0.234375	0.2344	18	18.0016	-2.5E-05
Strathspey Brent Group	Paralic/Shallow marine	0.90625	0.9061	61	60.9904	0.00015
Strathspey Bank Group	Continental	NaN	NaN			
Thistle	Paralic/Shallow marine	0.734375	0.7344	50	50.0016	-2.5E-05
Glitne	Deep Marine	0.65625	0.8246	45	55.7744	-0.1683
Grane	Deep Marine	0.984375	0.9845	66	66.008	-0.0001
Gungne	Continental	NaN	NaN			
Gyda	Paralic/Shallow marine	0.578125	0.5803	40	40.1392	-0.0021
Heimdal	Deep Marine	0.9375	0.6841	63	46.7824	0.2534
Jotun	Deep Marine	0.78125	0.7824	53	53.0736	-0.0012
Lille Frigg	Paralic/Shallow marine	0.4375	0.5588	31	38.7632	-0.1213
Mime	Paralic/Shallow marine	0	-0.0001	3	2.9936	0.0001
Oseberg 1	Paralic/Shallow marine	0.9375	0.9379	63	63.0256	-0.0004
Oseberg 2	Paralic/Shallow marine	0.9375	0.9378	63	63.0192	-0.0003
Oseberg Sor 2	Paralic/Shallow marine	0.375	0.3768	27	27.1152	-0.0018

Table 4: Recovery factor predicted by the trained SVM model compared to the actual recovery factor calculated using the volumetric method. Errors colored green for less than 0.015 and yellow 0.15-0.29. Recovery factor is both in fraction and percentage.