Performance Evaluation of Machine Learning Algorithms in Predicting Dew Point Pressure of Gas Condensate Reservoirs

Princewill Ikpeka*, Johnson Ugwu, Paul Russell and Gobind Pillai

[a] *School of Computing, Engineering and Digital Technologies, Teesside University, Middlesbrough, United Kingdom*

* https://orcid.org/0000-0002-1174-1491

p.ikpeka@tees.ac.uk

j.ugwu@tees.ac.uk

p.russell@tees.ac.uk

g.g.pillai@tees.ac.uk

Performance Evaluation of Machine Learning Algorithms in Predicting Dew Point Pressure of Gas Condensate Reservoirs

Abstract

Accurate knowledge of the dew point pressure for a gas condensate reservoir is necessary for optimizing mitigation operations during field development plan. This study explores the use of machine learning models in predicting the dew point pressure of gas condensate reservoirs. 535 experimental dew point pressure data-points with maximum temperature and pressure of 304^{o}F and 10500psi were used for this analysis. First, a standard multiple linear regression (MLR) was used as a benchmark for comparing the performance of the machine learning models. Multilayer perceptron Neural Networks (MLP) [optimized for the number of neurons and hidden layers], Support Vector Machine (SVM) [using radial basis function kernel] and Decision Tree [Gradient boost Method (GBM) and XG Boost (XGB)] algorithms were used to predict the dew point pressure. The performance of these algorithms was then compared with results obtained from published machine learning models. The input parameters for the model include; gas composition, specific gravity, the molecular weight of the heavier component and compressibility factor. The performances of these algorithms were analysed using root mean square error (RMSE), absolute average relative deviation percentage (AARD %) and coefficient of determination (R^2).

Keywords: Dew Point Pressure Prediction, Machine Learning Algorithms, Gas Condensate Reservoir.

Article Highlights

- Fed with the same training dataset, dewpoint pressure predictions from MLP, SVM, GBM and XGB algorithms exhibited significant disparity
- The choice of algorithm for dewpoint pressure prediction depends on the number of data points and the reliability of each data points for training the model.
- SVM algorithms were most prone to bias in training data while MLP algorithms were more sensitive to their architecture and number of training data

Declarations

Funding: Petroleum Technology Development Fund, under grant number PTDF/ED/PHD/PPI/1028/17

Conflicts of interest: The authors declare that they have no conflicting interest

Availability of data and material (data transparency): See attached

Code availability (software application or custom code): See attached

Authors' contributions: Not applicable

1. Introduction

Gas condensate reservoirs have a temperature range greater than the critical temperature and less than the cricondentherm (maximum temperature for which two phases can exist). They are characterized by producing both gas and condensate liquid at the surface (Louli et al., 2012; Skylogianni et al., 2015). At initial pressure, they exist in a gaseous phase. However, as the pressure declines below the dew-point pressure, condensates are formed (Thomas et al., 2009). The dew point pressure (DPP) is a principal parameter to be evaluated by reservoir engineers during a gas condensate field development. Below the dew point pressure, liquid condenses out of the gaseous phase. This liquid condensate forms a "ring" or "bank" across the production well as shown in Figure 1. This phenomenon is normally referred to as condensate banking (Hassan et al., 2019). Normally the condensate bank will not flow until its saturation increases beyond the critical saturation (S_{cc}) because of capillary pressure and relative permeability in the porous medium. Hence it is very essential to accurately predict the dew point pressure of the reservoir fluid. Condensate banking has the potential to decrease the well productivity significantly especially for low permeability reservoirs.

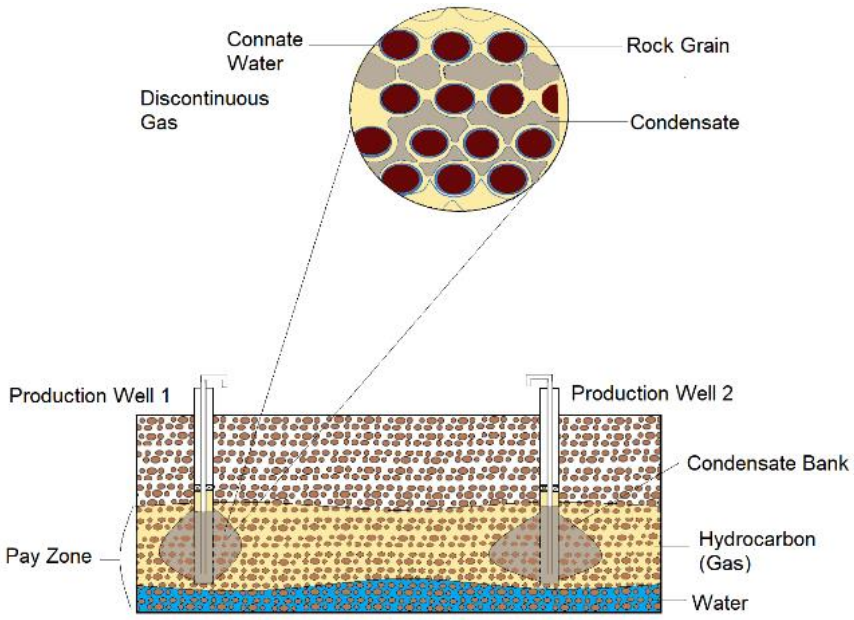

Figure 1: Schematics of Condensate blockage

From literature, there are 3 primary methods of predicting gas condensate dew point pressure: (i) Laboratory measurement (ii) Equations of State (EOS) and (iii) Empirical Correlations (Louli et al., 2012; Nemeth & Kennedy, 1967; Skylogianni et al., 2015). Examples of laboratory test for dewpoint pressure are constant composition expansion (CCE) and constant volume depletion (CVD) tests. CCE test is on gas samples retrieved at the initial stage of a reservoir when its pressure is still above the dew point pressure (Al-Shawaf et al., 2014). CCE experiments require several sequential pressure decreases in little stages. CVD tests are preferentially done on volatile oils and light hydrocarbons because there is considerable

change in the hydrocarbon composition when pressure decreases during the test (Rabiei et al., 2015a). Although both tests generate accurate results, the cost and time required to run the test becomes an important limitation. In additional to this, several possible sources of error can be introduced during Laboratory measurements; from gas sample collection to measurement errors during the test (Arabloo et al., 2014; Hosein & Dawe, 2012). Peng-Robinson and Soave-Redlich-Kwong equation of state (EOS) models have been used to predict PVT properties of hydrocarbons. These EOS models have to be tuned with experimental data for heptane-plus fraction before they can be used to characterize hydrocarbon phase behaviour(González et al., 2003; Rabiei et al., 2015a). The performance of EOS models is generally satisfactory for simple hydrocarbon fluids. However, it gets complicated when applied to volatile oil and gas condensate reservoirs. In addition, the accuracy of EOS models is dependent on several factors: binary interaction coefficient (BIC), volume shift parameter (S), critical pressure (Pc), critical temperature (Tc) and acentric factor (ω) for the last carbon group (Louli et al., 2012; Nasrifar et al., 2005). These parameters need to be estimated for each component in the gas condensate mixture and becomes increasingly complicated when characterizing the plus fractions of condensate mixtures (Aghamiri et al., 2018; Elsharkawy, 2002; Jhaveri & Youngren, 1988; Rabiei et al., 2015a). Empirical correlations are also used to predict dewpoint pressure and it offers a simple, fast estimate of dewpoint pressure for used in planning purposes. Several empirical correlations estimating the dewpoint pressure as a function of reservoir fluid properties such as temperature, gas compressibility, specific gravity and gas composition have been proposed by researchers(Almehaideb et al., 2003; Elsharkawy, 2002; Nemeth & Kennedy, 1967). Empirical correlations, although easy to use, do not adequately capture the non-linear relationship between the dewpoint pressure and dynamic reservoir properties (Zhong et al., 2018). This limits the accuracy of empirical correlations especially when estimating dewpoint pressure at elevated temperature conditions. Many researchers have explored the potential of applying artificially intelligent tools and other robust techniques such as; artificial neural networks (ANNs) (Ahmadi et al., 2014; González et al., 2003; H. Kaydani et al., 2013; Majidi et al., 2014), support vector machine (SVM) (Ahmadi & Ebadi, 2014; Zhong et al., 2018), genetic algorithm (GA) (Najafi-Marghmaleki et al., 2016; Rabiei et al., 2015b), multi-gene genetic programming approach (Hossein Kaydani et al., 2016) and swarm particle optimization (SPO) (Zhong et al., 2018) to predict dewpoint pressure of gas condensate reservoirs.

Gonzales and Startzman (González et al., 2003) used a set of about 641 experimental data points comprising of gas composition, molecular weight, specific gravity of the heavier fraction and reservoir temperature, to train their neural network model. Results from their study reveal that the neural network model was more accurate than EOS model (Peng-Robinson) and other empirical correlations

(Nemeth(Nemeth & Kennedy, 1967) and Elsharkawy(Elsharkawy, 2002)). Nowroozi et al. (Nowroozi et al., 2009) in a bid to optimize the architecture of the neural network combined fuzzy logic with neural network models in a system called Adaptive Neuro-Fuzzy Inference System. The resulting hybrid system was shown to be faster than conventional neural network models. Arabloo et al., (Majidi et al., 2014) applied least square support vector machine (LSSVM) to predict dew point pressure. In their work, coupled simulated annealing (CSA) was used to optimize some parameters of the model. Ahmadi et al. (Ahmadi & Ebadi, 2014) applied a fusion of particle swarm optimization and ANNs to predict the dewpoint pressure. They indicated that their model showed superior performance when compared to other hybrid models. Rabiei et al. (Rabiei et al., 2015b) combined multilayer perceptron (MLP) and genetic algorithm (GA) to predict dewpoint pressure using 308 data point, while Zhong et al. (Zhong et al., 2018) executed a mixed-kernels-function support vector machine based model to predict dewpoint pressure.

The key benefit of using machine learning algorithms is their ability to characterize the non-linearity, uncertainty and ambiguities between dewpoint pressure and its corresponding input parameters (Ahmadi & Ebadi, 2014). Each algorithm has its own unique limitations and strengths; the accuracy of neural network algorithms depend on its architecture (Kubat, 1999), the accuracy of the SVM is tied to the number of reliable data fed to the algorithm (Zhong et al., 2018). The dewpoint predictions obtained using different algorithms have shown considerable disparity. Some algorithms are prone to bias in the data, while others are not and this affects the overall accuracy of their final predictions. The aim of this study is to compare the performance of four major machine learning algorithms: multilayer perceptron neural network (MLP), support vector machine (SVM), gradient boost (GBM) and extreme gradient boost (XGB). The disparities in dewpoint predictions are compared against experimental data using standard error analysis. Then the number of training data is reduced to see the effect on the accuracy of each algorithm's prediction. This paper is organized as follows: Section 2 describes the validation method applied to the experimental data used for training and testing the algorithm. Then the algorithms and their predictions are given in subsequent subsections. Section 3 introduces the error parameters used for comparing the performance of each algorithm. The results of the comparison are discussed in Section 4 while Section 5 is the conclusion.

2. Methodology

535 dew point pressure experimental data points were obtained from published literature. A summary of the key data point obtained is presented in Table 1. Data validation was done using the method outlined in (Bashbush et al., 2004). The validation process is shown in steps 1-3.

(1) A graph of Log Ki vs Pressure was prepared

(2) Data sections of the graph revealing sharp divergence on the curve were removed

(3) Then the graph of Log K_i vs Boiling Temperature (T_b) of each component was adjusted to 100% sum of gas mole percentages

Table 1: Data Ranges used for this study

Parameters	**Max**	**Min**	**Mean**	**Std Dev**
Methane	0.966	0.034	0.8006	0.1260
Ethane	0.151	0.003	0.0569	0.02980
Propane	0.109	0.001	0.0297	0.01992
Butane	0.080	0.001	0.0193	0.01297
Pentane	0.123	0.000	0.0126	0.0130
Hexane	0.087	0.0004	0.0095	0.00911
Nitrogen	0.432	0	0.0103	0.03105
Carbon Dioxide	0.919	0	0.0154	0.05768
Hydrogen Sulfide	0.299	0	0.0069	0.03128
Heptane Plus	0.136	0.002	0.0378	0.02854
Compressibility	1.759	0.644	0.9801	0.16242
Specific Gravity	0.888	0.306	0.7866	0.03429
Molecular Weight of	235	106	148.08	23.6516
Dew Point Pressure	10790	1405	4751.7	1640.49

The distribution of the dew point data used for this analysis is shown in figure 2

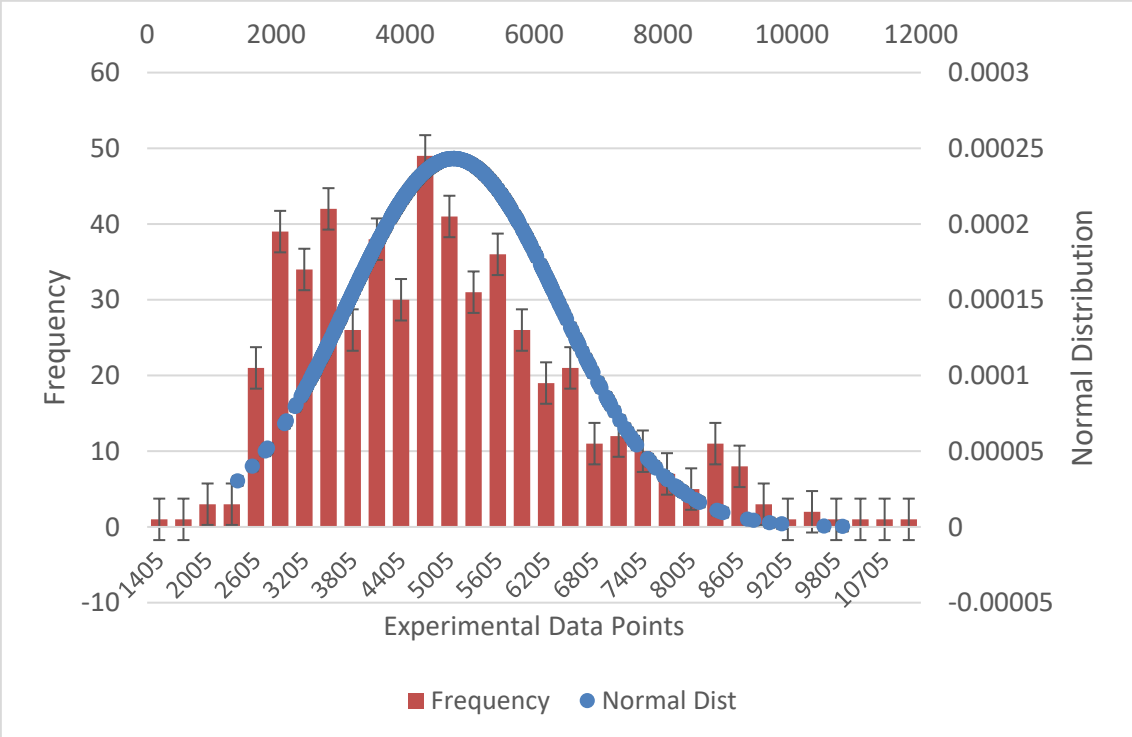

Figure 2: Distribution of Experimental Data points used in this study

2.1 MLP Neural Network

A simple multilayer feed forward neural network model was trained by using 13 input nodes to predict the dew point pressure output node. The algorithm used for predicting the output node is described in Algorithm (1). One main challenge of using the multilayer perceptron network is selecting the appropriate number of

neurons and hidden layers that would yield the least error when compared with the observed data. In this work, the number of hidden units and neurons were varied as shown in Figure 4. Their accuracy and precision were measured, and the results presented in Figure 3. It was observed that Neural Network with single hidden layer and 5 neurons gave the best prediction with least error whereas the neural network with 6 neurons gave the least accuracy. The single-layered 5 neurons architecture was then trained using three different algorithms; Levenberg–Marquardt, Scaled Conjugate Gradient, and Bayesian Regularization

Algorithm 1 – Feed Forward Multilayer Perceptron Network

$\mathbb{N}$: Number of experimental dew point pressure data, $\mathbb{N} = \{\mathbf{1}, \dots, \boldsymbol{n}\}$
$\mathbb{N}_{\boldsymbol{train}}$: Set of data for training
$\mathbb{N}_{\boldsymbol{test}}$: Set of data for testing
$\boldsymbol{\eta}$: Number of Input Neurons
$\boldsymbol{\rho}$: Number of hidden layer
$\emptyset$: Activation function
$\emptyset_{\boldsymbol{d}}$: Derivative of activation function
$\boldsymbol{x_i}$: Input Node
$\boldsymbol{y_i}$: Output Node
$\boldsymbol{W_{ji}}$: Weight
%%Split data set into testing and training dataset.
1: for each $\boldsymbol{\eta}$ and $\boldsymbol{\rho}$ do
2: Compute each input node, $\boldsymbol{i} \in \boldsymbol{\eta}$,
3: $\boldsymbol{x_i} = \boldsymbol{Output_i}$
4: Compute each hidden layer node, $\boldsymbol{j} \in \boldsymbol{\rho}$,
5: $\boldsymbol{Output_j} = \sum_i \emptyset(\boldsymbol{W_{ji}} . \boldsymbol{Output_i}) - \boldsymbol{b_j}$
6: Compute each output layer neuron k,
7: $\boldsymbol{Output_k} = \sum_k \emptyset_{\boldsymbol{d}}(\boldsymbol{W_{kj}} . \boldsymbol{Output_i})$
8: Compute Output Error,
9: $\boldsymbol{\varepsilon_{output}} = \sum_k (\boldsymbol{Target_k} - \boldsymbol{Output_k})^{\mathbf{2}}$
10: end for
11: Return Average($\boldsymbol{\varepsilon_{output}}$, InputSet)
12: Repeat procedure for $\boldsymbol{i} + \mathbf{1} \in \mathbb{N}_{\boldsymbol{train}}$, while adjusting $\boldsymbol{W_{ji}}$ and $\boldsymbol{b_j}$
13: End

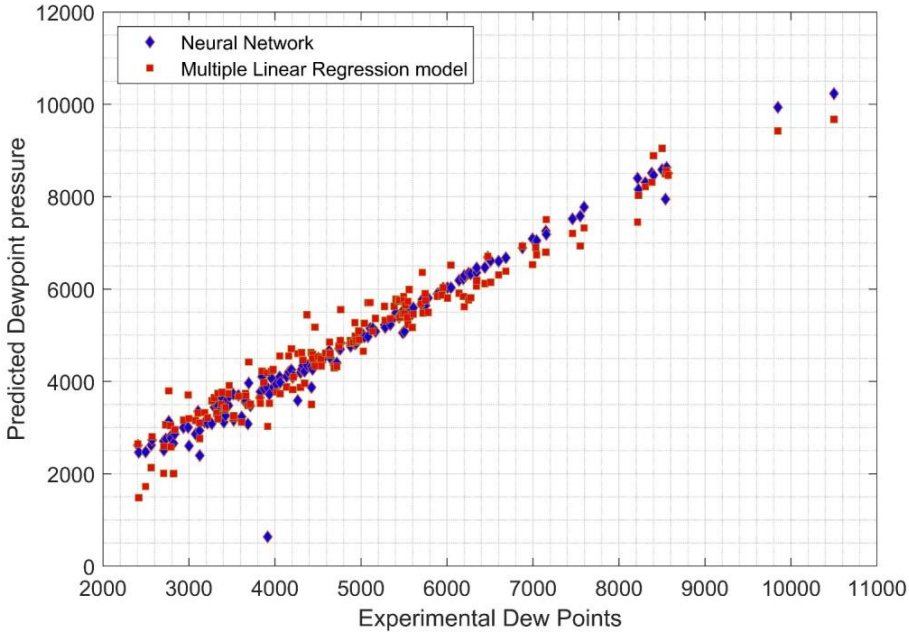

Figure 3: NN model vs MLR model plotted against Experimental data

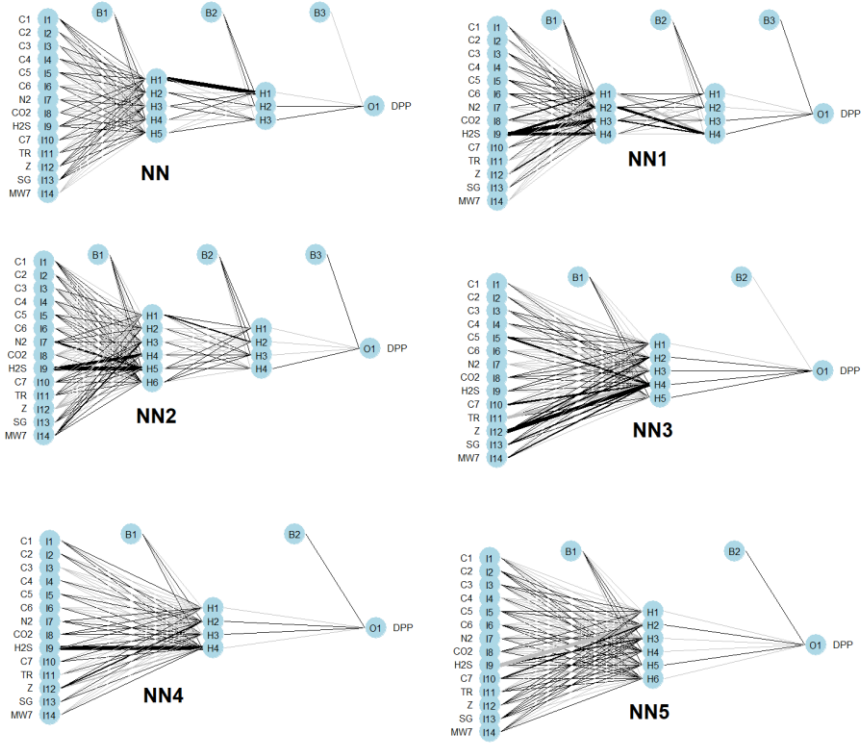

Figure 4: Neural Network Architecture used in this analysis

2.2 Support Vector Machine Models

The SVM model for this analysis utilizes the whole support vectors to produce an output. The decision function for the support vector algorithm is described in equation (1).

$$f\big(x(t)\big) = \sum_{i=1}^{n} \alpha^* y_i K(x^*, x(t)) + b^* \quad (1)$$

x^* – i[th] vector of n support vectors

y_i – Class label

$x(t)$ – t[th] input frame vector

b^* – Optimization bias

α^* – Langrange multiplier

The kernel function used in this analysis is that of the radial basis function represented by equation (2).

$$K(x^*, x(t)) = \exp\left(-\gamma \|x^* - x(t)\|^2\right) \quad (2)$$

γ – Kernel parameter of the RBF

Algorithm 2 – Support Vector Machine Regression		
$\mathbb{N}$:	Number of experimental dew point pressure data,	$\mathbb{N} = \{1, \dots, n\}$
$\mathbb{N}_{sv}$:	Number of Support Vectors	
$\mathbb{N}_{ft}$:	Number of features in support vector	
y:	Decision Output	
%% Split data set		
1: for $i \in \mathbb{N}$ do		
2:	$y = 0$	
3:	for $j \in \mathbb{N}_{sv}$ do	
4:	$\sigma = 0$	
5:	for k $\in \mathbb{N}_{ft}$ do	
6:	$\sigma \mathrel{+}= (SV[j].feature[k] - IN[i].feature[k])^2$	
7:	end for	
8:	$k = \exp(-\gamma \times \sigma)$	
9:	$y \mathrel{+}= SV[j].\alpha^* \times k$	
10:	end for	
11:	$y = y + b^*$	
12:End		

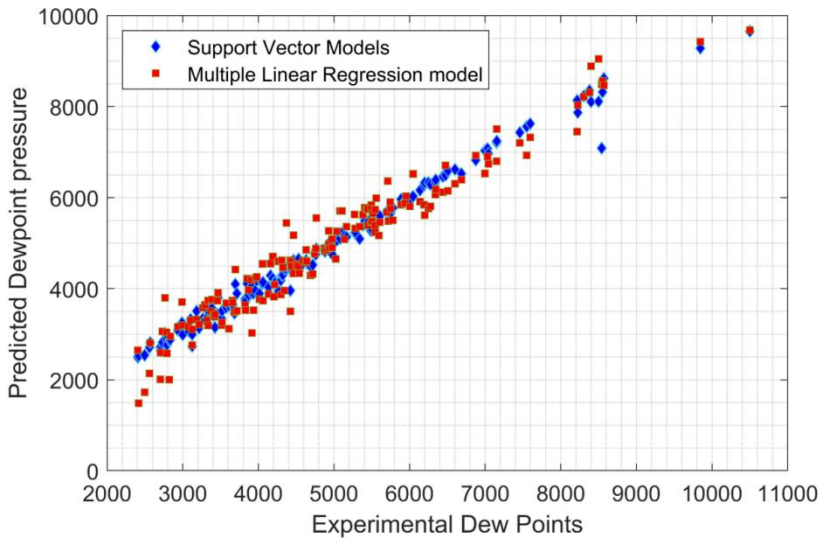

Figure 5: SVM predictions vs Experimental Dew Point

2.3 Stochastic and Extreme Gradient Boosting

In this algorithm 20,000 trees were used to fabricate additive regression models by logically adding a simple function (base learner) to the iteratively obtained residuals using least squares method. At each iteration, the differential of the loss function represents the residuals being reduced according to the model parameter. The algorithm used is adapted from (Friedman, 2002) and is shown in Algorithm (3). The predictions for each algorithm are presented in figures 6 and 7.

Algorithm 3 – Stochastic Gradient Boost Model	
$\boldsymbol{X}$	Vector of input variables $\boldsymbol{X} = (x_i, \dots, x_k)$
Y	Set of response variables Y $= (x_i, \dots, x_k)$
$\Omega(y_i, \gamma)$	Differentiable loss function
M	Number of Iterations
$\{\tau(i)\}_1^N$	Random permutation of integers
1: Initialize model with constant value:	
2	$F_0(x) = \arg min_\gamma \sum_{i-1}^{N} \Omega(y_i, \gamma)$
3: Introduce Randomness into selection of subsample	
4	$\{\tau(i)\}_1^N = rand\_perm\ \{i\}_1^N$
5: For m=1 to M do:	
6:	$\tilde{y}_{\tau(i)m} = -\left[\frac{\partial \Omega\left(y_{\tau(i)}, F(x_{\tau(i)})\right)}{\partial F(x_{\tau(i)})}\right]_{F(x)=F_{m-1}(x)}, i = 1, \tilde{N}$
7:	$\{R_{lm}\}_1^L = L - terminal\ node\ tree(\{\tilde{y}_{\pi(i)m}, x_{\pi(i)}\}_1^{\tilde{N}})$
8:	$\gamma_{lm} = \arg min_\gamma \sum_{x_{\tau(i)} \in R_{lm}} \Omega(y_{\tau(i)}, F_{m-1}(x_{\tau(i)}) + \gamma)$
9:	$F_m(x) = F_{m-1}(x) + v.\gamma_{lm}(x \in R_{lm})$
10: End for	

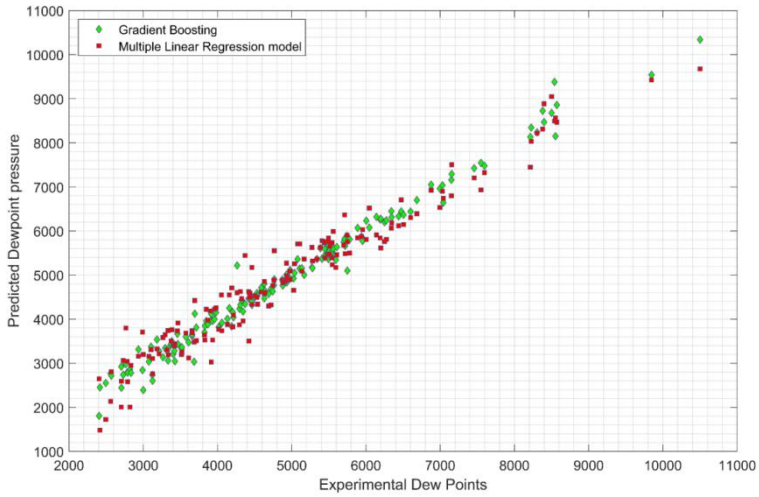

Figure 6: GBM prediction vs Experimental Dew Point

Algorithm 4 – Extreme Gradient Boost Model	
X	Vector of input variables $X = (x_i, \dots, x_k)$
Y	Set of response variables Y $= (y_i, \dots, y_k)$
ϵ	Random error with zero mean
k	Number of input parameters
w	Vector score on leaf
T	Number of leaves.
I_j	Set of data points assigned to p-th leaf $I_p = \{i \mid q(x_i) = p\}$

%% Split data set

1: for a tree structure $q(x)$,

2: $Y = \sum_{k=1}^{k} h_k(x_i)\,,\ h_k \in \Phi$

3: $Obj^{(t)} = \sum_{p=1}^{T}\left[G_p w_p + \frac{1}{2}\left(H_j + \lambda\right){w_p}^2\right] + \gamma T$ %% Objective function

4: $G_p = \sum_{i \in I_p} g_i$

5: $H_p = \sum_{i \in I_p} h_i$

6: $g_i = \frac{\partial}{\partial \hat{y}_i^{(t-1)}}\left[l\left(y_i, \hat{y}_i^{(t-1)}\right)\right]$ 1st order term

7: $f_i = \frac{\partial^2}{\partial \hat{y}_i^{(t-1)}}\left[l\left(y_i, \hat{y}_i^{(t-1)}\right)\right]$ 2nd Order term

8: $\hat{y}_i^{(0)} = 0$

9: $\hat{y}_i^{(1)} = h_1(x_i) = \hat{y}_i^{(0)} + h_1(x_i)$

10: $\hat{y}_i^{(2)} = h_1(x_i) + h_2(x_i) = \hat{y}_i^{(0)} + h_1(x_i)$

…

11: $\hat{y}_i^{(t)} = \sum_{k=1}^{t} h_k(x_i) = \hat{y}_i^{(t-1)} + h_t(x_i)$

12: End

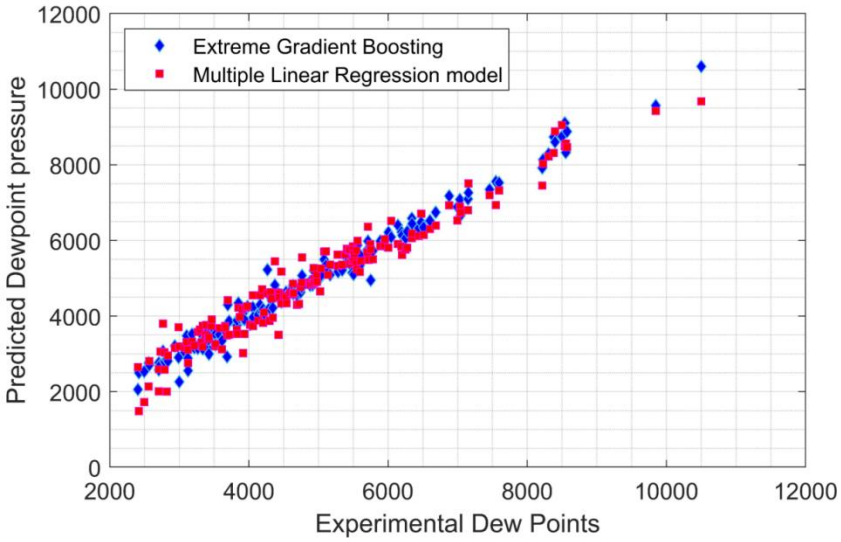

Figure 7: XGB predictions vs Experimental Dew Point

2.4 Published machine learning algorithms used for dewpoint prediction

2.4.1 Committee Machine Intelligent System (Haji-Savameri et al., 2020)

Committee Machine Intelligent System combines several models into a single model, allocating weights to each model based on their accuracy. To predict dewpoint, Haji-Savameri (Haji-Savameri et al., 2020) linearly combined four ANN models into a single model using a weighted average. However, in this study, we utilized three MLP models and 1 SVM model to build the CMIS model as shown in figure 8. The constants a_0 to a_4 are derived from regression as shown in equation (2).

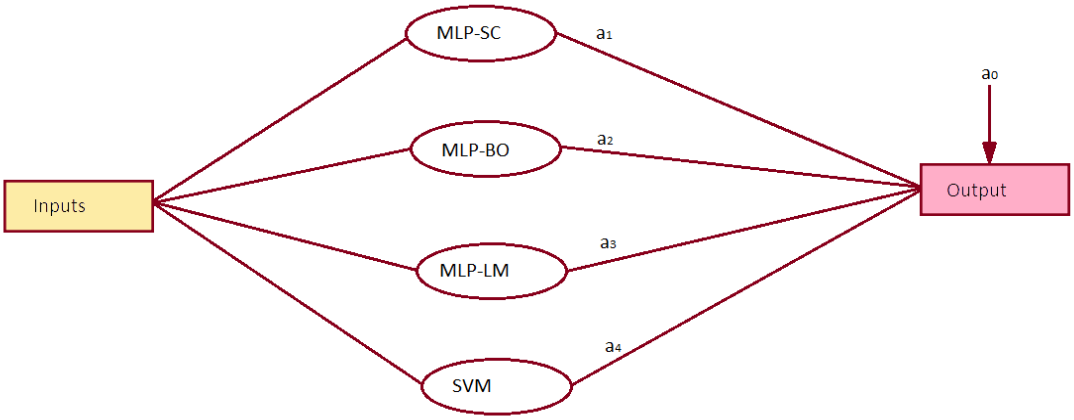

Figure 8: CMIS used in this study

$$P_d = a_0 + a_1 P_{d\,MLP-SC} + a_2 P_{d\,MLP-BO} + a_3 P_{d\,MLP-LM} + a_4 P_{d\,SVM}$$

Where the constants a_0 to a_4 are given as:

$a_0 = -0.047502367$ $a_1 = 0.367736043$ $a_2 = 0.544060545$ $a_3 = 0.146960632$ $a_4 = -0.061931495$

3. Results and Analysis

3.1 Error Analysis

Goodness of fit (R^2) was performed on the various predicted models to describe its proximity with the experimental data. The higher the R^2 value the better the predicted model fits the experimental data. It is important to note however that R^2 value does not indicate the presence of bias in the predicted model.

$$R = 1 - \frac{\sum_{i-1}^{m}\left[(P_d)_{exp} - (P_d)_{pred}\right]^2}{\sum_{i-1}^{m}\left[(P_d)_{exp}\right]_i - \overline{\Delta P_d}} \quad (3)$$

$$\overline{\Delta P_d} = \frac{1}{n}\sum_{i-1}^{N}\left[(P_d)_{exp}\right]_i \quad (4)$$

P_d – Dew point pressure

$(P_d)_{act}$ – Experimental dew point pressure

$(P_d)_{pred}$ – Predicted dew point pressure

Average Absolute Relative Deviation (AARD) and Root Mean Square Error were obtained from Equation (5-7):

$$E_r = \frac{1}{n}\sum_{i-1}^{N}|E_i| \quad (5)$$

$$E_i = \left[\frac{(P_d)_{exp} - (P_d)_{pred}}{(P_d)_{exp}}\right] \times 100 \quad (6)$$

$$RMSE = \sqrt{\left[\frac{1}{N}\sum_{i-1}^{N} E_i{}^2\right]} \quad (7)$$

Case Study 1:

The data set was split into two clusters: 428 data points were used in training the model and 107 data points were used for testing the model. The results of the error analysis performed are presented in Table 2. Using Multiple Linear Regression (MLR) as the base case, the performance of the prediction models is presented in figure 8.

Table 2: Result of Error Analysis on Predictive Models (Case study 1)

	R^2	AARD (%)	RMSE
CMIS	0.9848	3.11%	189.5197
SVM	0.9791	3.43%	222.2439
MLP-LM	0.9783	3.82%	226.3954
GBM	0.9672	4.01%	278.0877
MLR	0.9588	5.67%	311.7844
MLP-BO	0.9502	4.14%	342.79
XGB	0.9473	5.98%	352.5907
MLP-SC	0.9203	6.09%	433.8276

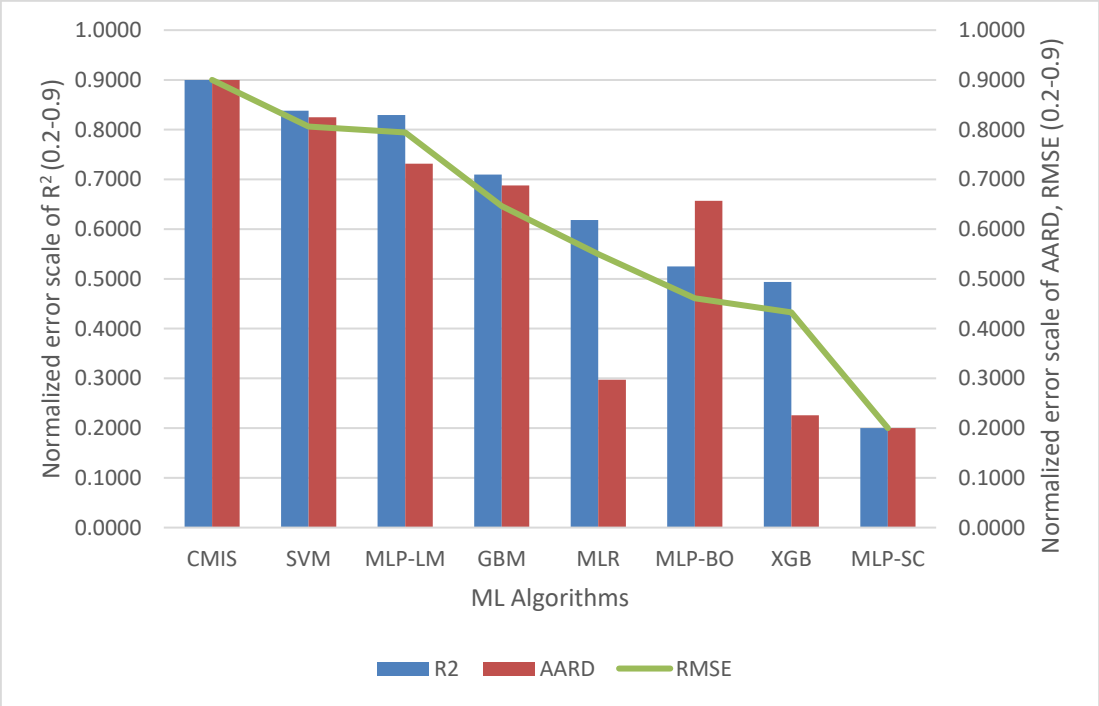

Figure 9: Normalized Error with respect to Multiple Linear Regression (case study 1)

Case Study 2:

In order to understand the effect of the number of training data on the performance of the models, the experimental data set used for the training the models were decreased from 428 to 375. The split ratio represents 70% for training while 30% for testing. Results from the error analysis of each model are presented in Table 3. Figure 9 presents a normalized error analysis.

Table 3: Result of Error Analysis on Predictive Models (Case study 2)

	R^2	AARD	RMSE
CMIS	0.9672	0.0862	381.8288
GBM	0.9603	0.0951	419.8862
XGB	0.9459	0.1149	490.3562
SVM	0.9246	0.1138	578.5537
MLP-LM	0.9084	0.1216	637.9442
MLP-SC	0.8975	0.1570	674.6094
MLP-BO	0.5793	0.1865	1367.079
MLR	-0.3613	0.3191	2458.985

Figure 10: Normalized Error with respect to Multiple Linear Regression (case study 2)

4. Discussion

While developing the MLP algorithm, the following parameters were pre-set; the number of layers in the network, number of neurons in each layer and the activation function of each layer (see algorithm 1). Six pre-set parameters were randomly chosen, and the resulting predictions checked for accuracy against experimental data. Ranking the performances based on the combined error parameters for Case Study 1, we observe a superior performance of the CMIS algorithm over other algorithms. Results from the analysis reveal superior performance of the GBM model when the number of training data set is altered. To prevent overfitting of the tree booster in XGB algorithm, we applied a shrinkage factor of 0.02, maximum depth of 10 and a subsampling ratio of 0.5. The details of each parameter used are found in the attached XGB Code. The XGB algorithm makes use of parallel computing ability to generate faster results. From results obtained in this study, we observe that the accuracy of the XGB algorithm improves with higher number of training

data. While the accuracy of the GBM algorithm improves with less training data. This can be attributed to the use of sparse matrices, overfitting parameters used in the XGB algorithm.

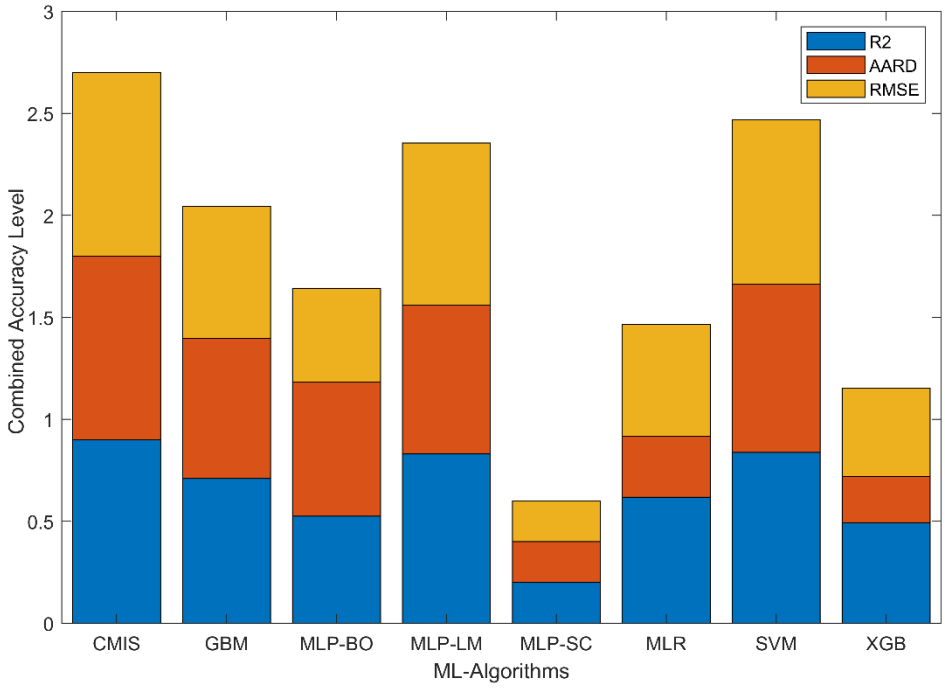

Figure 11: Performance based on combined error analysis (case study 1)

For case study 2 we observed a different ranking among the prediction model based on their combined error analysis as shown in figure 11. Neural network with single hidden layer and 4 neurons performed better than others. The SVM algorithm uses a nonlinear radial basis function kernel to create an optimal hyperplane that separates all the training data sets into classes. This can be seen in Algorithm [2]. The resulting implication is that the SVM algorithm takes more computing time to process large amount of data. The accuracy of prediction is satisfactory for small number of data points as can be seen in figure 11. The decision on the choice of algorithm is dependent on the number of data available and reliability of the data.

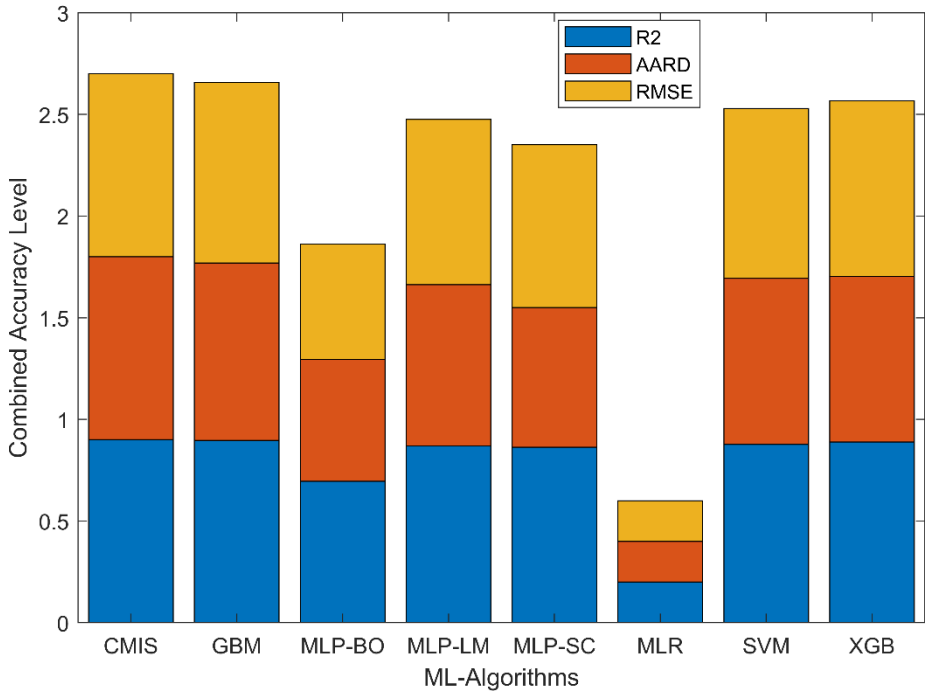

Figure 12: Performance based on combined error analysis (case study 2)

5. Conclusion

In this paper CMIS, SVM, MLP, GBM and XGB algorithms were successfully adapted to predict dewpoint pressure of Gas condensate reservoirs. The algorithms were modelled as a linear function of the input data - Gas composition, specific gravity, compressibility, temperature, and molecular weight of heavier fractions. For evaluating the performance of each algorithm, the predictions were compared with experimental data obtained from published literature(Nemeth & Kennedy, 1967). AARD was used as a measure of precision while RMSE was used as a measure of accuracy. On the measure of precision, SVM significantly outperforms other models. This is because SVM selects support vectors from the training data and uses it to make new predictions; it is able to achieve high precision. The drawback however is for complex data structure, more support vectors are needed to achieve reasonable accuracy and this result in high computational time. Consequently, the accuracy of the SVM model decreases with a decrease in the number of training data. Although CMIS model performs better than other algorithm, it is important to note that the CMIS model is a combination of four other models MLP-SC, MLP-LM, MLP-BO and SVM. Based on the MLP training algorithms used in this study, the MLP-LM algorithm consistently showed superior performance over other MLP algorithms and yielded the least error. This study concludes that the choice of machine learning algorithm for dewpoint prediction depends on the number of training data available and the reliability of such data points. For small and reliable datasets, SVM framework is preferred while for larger data sets, neural network models perform better. The accuracy of predictions of each algorithm used in this study can be improved by incorporating optimization techniques such as genetic algorithm

(Ahmadi & Ebadi, 2014; Najafi-Marghmaleki et al., 2016), swarm algorithm (Haji-Savameri et al., 2020; Zhong et al., 2018) and fuzzy logic(Ahmadi et al., 2014).

Acknowledgements

The authors acknowledge the Petroleum Technology Development Fund (PTDF) for providing funding for this study.

References

Aghamiri, S., Tamtaji, M., & Ghafoori, M. J. (2018). Developing a K-value equation for predict dew point pressure of gas condensate reservoirs at high pressure. *Petroleum*, *4*(4), 437–443. https://doi.org/10.1016/j.petlm.2017.08.002

Ahmadi, M. A., & Ebadi, M. (2014). Evolving smart approach for determination dew point pressure through condensate gas reservoirs. *Fuel*, *117*(PARTB), 1074–1084. https://doi.org/10.1016/j.fuel.2013.10.010

Ahmadi, M. A., Ebadi, M., & Yazdanpanah, A. (2014). Robust intelligent tool for estimating dew point pressure in retrograded condensate gas reservoirs: Application of particle swarm optimization. *Journal of Petroleum Science and Engineering*, *123*, 7–19. https://doi.org/10.1016/j.petrol.2014.05.023

Al-Shawaf, A., Kelkar, M., & Sharifi, M. (2014). A new method to predict the performance of gas-condensate reservoirs. *SPE Reservoir Evaluation and Engineering*, *17*(2), 177–189. https://doi.org/10.2118/161933-PA

Almehaideb, R. A., Ashour, I., & El-Fattah, K. A. (2003). Improved K-value correlation for UAE crude oil components at high pressures using PVT laboratory data. *Fuel*, *82*(9), 1057–1065. https://doi.org/10.1016/S0016-2361(03)00004-8

Arabloo, M., Heidari Sureshjani, M., & Gerami, S. (2014). A new approach for analysis of production data from constant production rate wells in gas condensate reservoirs. *Journal of Natural Gas Science and Engineering*, *21*, 725–731. https://doi.org/10.1016/j.jngse.2014.09.028

Bashbush, B. J. L., León, G. A., Mazariegos, U. C., Corona, B. A., & Unam, C. P. P. F. (2004). *SPE 91505 On the Validation of PVT Compositional Laboratory Experiments*. 1–7.

Elsharkawy, A. M. (2002). Predicting the dew point pressure for gas condensate reservoirs: empirical models and equations of state. *Fluid Phase Equilibria*, *193*(1–2), 147–165. https://doi.org/10.1016/S0378-3812(01)00724-5

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics and Data Analysis*, *38*(4),

367–378. https://doi.org/10.1016/S0167-9473(01)00065-2

González, A., Barrufet, M. A., & Startzman, R. (2003). Improved neural-network model predicts dewpoint pressure of retrograde gases. *Journal of Petroleum Science and Engineering*, *37*(3–4), 183–194. https://doi.org/10.1016/S0920-4105(02)00352-2

Haji-Savameri, M., Menad, N. A., Norouzi-Apourvari, S., & Hemmati-Sarapardeh, A. (2020). Modeling dew point pressure of gas condensate reservoirs: Comparison of hybrid soft computing approaches, correlations, and thermodynamic models. *Journal of Petroleum Science and Engineering*, *184*(June 2019), 106558. https://doi.org/10.1016/j.petrol.2019.106558

Hassan, A., Mahmoud, M., Al-Majed, A., Alawi, M. B., Elkatatny, S., BaTaweel, M., & Al-Nakhli, A. (2019). Gas condensate treatment: A critical review of materials, methods, field applications, and new solutions. *Journal of Petroleum Science and Engineering*, *177*(December 2018), 602–613. https://doi.org/10.1016/j.petrol.2019.02.089

Hosein, R., & Dawe, R. A. (2012). *Tuning of the Peng-Robinson Equation of State for Gas Condensate Simulation Studies*. *D*(1985). https://doi.org/10.2118/158882-ms

Jhaveri, B. S., & Youngren, G. K. (1988). Three-Parameter Modification of the Peng-Robinson Equation of State To Improve Volumetric Predictions. *SPE Reservoir Engineering*, *3*(03), 1033–1040. https://doi.org/10.2118/13118-PA

Kaydani, H., Hagizadeh, A., & Mohebbi, A. (2013). A dew point pressure model for gas condensate reservoirs based on an artificial neural network. *Petroleum Science and Technology*, *31*(12), 1228–1237. https://doi.org/10.1080/10916466.2010.540616

Kaydani, Hossein, Mohebbi, A., & Hajizadeh, A. (2016). Dew point pressure model for gas condensate reservoirs based on multi-gene genetic programming approach. *Applied Soft Computing Journal*, *47*, 168–178. https://doi.org/10.1016/j.asoc.2016.05.049

Kubat, M. (1999). Neural networks: a comprehensive foundation by Simon Haykin, Macmillan, 1994, ISBN 0-02-352781-7. *The Knowledge Engineering Review*, *13*(4), 409–412. https://doi.org/DOI: 10.1017/S0269888998214044

Louli, V., Pappa, G., Boukouvalas, C., Skouras, S., Solbraa, E., Christensen, K. O., & Voutsas, E. (2012). Measurement and prediction of dew point curves of natural gas mixtures. *Fluid Phase Equilibria*, *334*, 1–9. https://doi.org/10.1016/j.fluid.2012.07.028

Majidi, S. M. J., Shokrollahi, A., Arabloo, M., Mahdikhani-Soleymanloo, R., & Masihi, M. (2014). Evolving an accurate model based on machine learning approach for prediction of dew-point pressure in gas condensate reservoirs. *Chemical Engineering Research and Design*, *92*(5), 891–902.

https://doi.org/10.1016/j.cherd.2013.08.014

Najafi-Marghmaleki, A., Tatar, A., Barati-Harooni, A., Choobineh, M. J., & Mohammadi, A. H. (2016). GA-RBF model for prediction of dew point pressure in gas condensate reservoirs. *Journal of Molecular Liquids*, *223*, 979–986. https://doi.org/10.1016/j.molliq.2016.08.087

Nasrifar, K., Bolland, O., & Moshfeghian, M. (2005). Predicting Natural Gas Dew Points from 15 Equations of State. *Energy & Fuels*, *19*(2), 561–572. https://doi.org/10.1021/ef0498465

Nemeth, L. K., & Kennedy, H. T. (1967). A Correlation of Dewpoint Pressure With Fluid Composition and Temperature. *Society of Petroleum Engineers Journal*, *7*(02), 99–104. https://doi.org/10.2118/1477-PA

Nowroozi, S., Ranjbar, M., Hashemipour, H., & Schaffie, M. (2009). Development of a neural fuzzy system for advanced prediction of dew point pressure in gas condensate reservoirs. *Fuel Processing Technology*, *90*(3), 452–457. https://doi.org/10.1016/j.fuproc.2008.11.009

Rabiei, A., Sayyad, H., Riazi, M., & Hashemi, A. (2015a). Determination of dew point pressure in gas condensate reservoirs based on a hybrid neural genetic algorithm. *Fluid Phase Equilibria*, *387*, 38–49. https://doi.org/10.1016/j.fluid.2014.11.027

Rabiei, A., Sayyad, H., Riazi, M., & Hashemi, A. (2015b). Determination of dew point pressure in gas condensate reservoirs based on a hybrid neural genetic algorithm. *Fluid Phase Equilibria*, *387*, 38–49. https://doi.org/10.1016/j.fluid.2014.11.027

Skylogianni, E., Novak, N., Louli, V., Pappa, G., Boukouvalas, C., Skouras, S., Solbraa, E., & Voutsas, E. (2015). Measurement and prediction of dew points of six natural gases. *Fluid Phase Equilibria*, *424*, 8–15. https://doi.org/10.1016/j.fluid.2015.08.025

Thomas, F. B., Bennion, D. B., & Andersen, G. (2009). Gas Condensate Reservoir Performance. *Journal of Canadian Petroleum Technology*, *48*(07), 18–24. https://doi.org/10.2118/09-07-18

Zhong, Z., Liu, S., Kazemi, M., & Carr, T. R. (2018). Dew point pressure prediction based on mixed-kernels-function support vector machine in gas-condensate reservoir. *Fuel*, *232*(May), 600–609. https://doi.org/10.1016/j.fuel.2018.05.168

Appendix

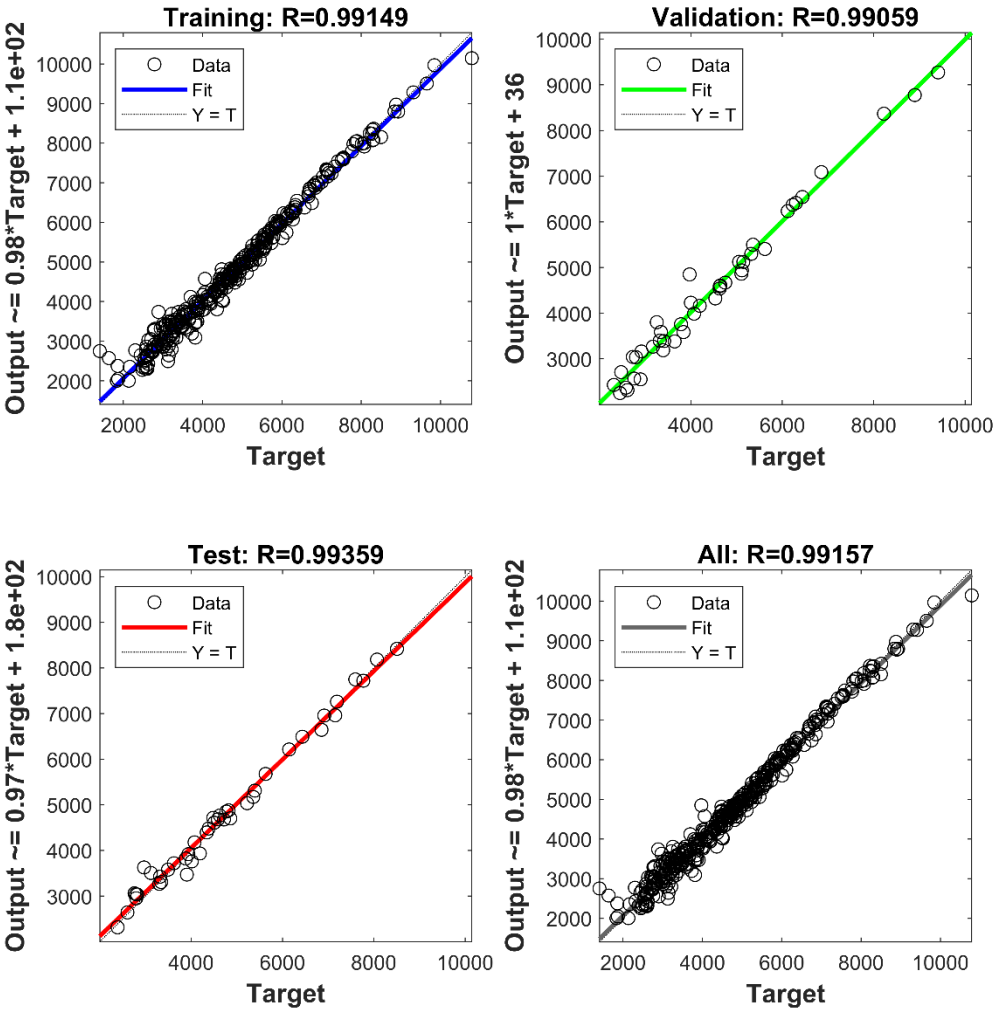

Figure 13: Performance of the MLP-LM model

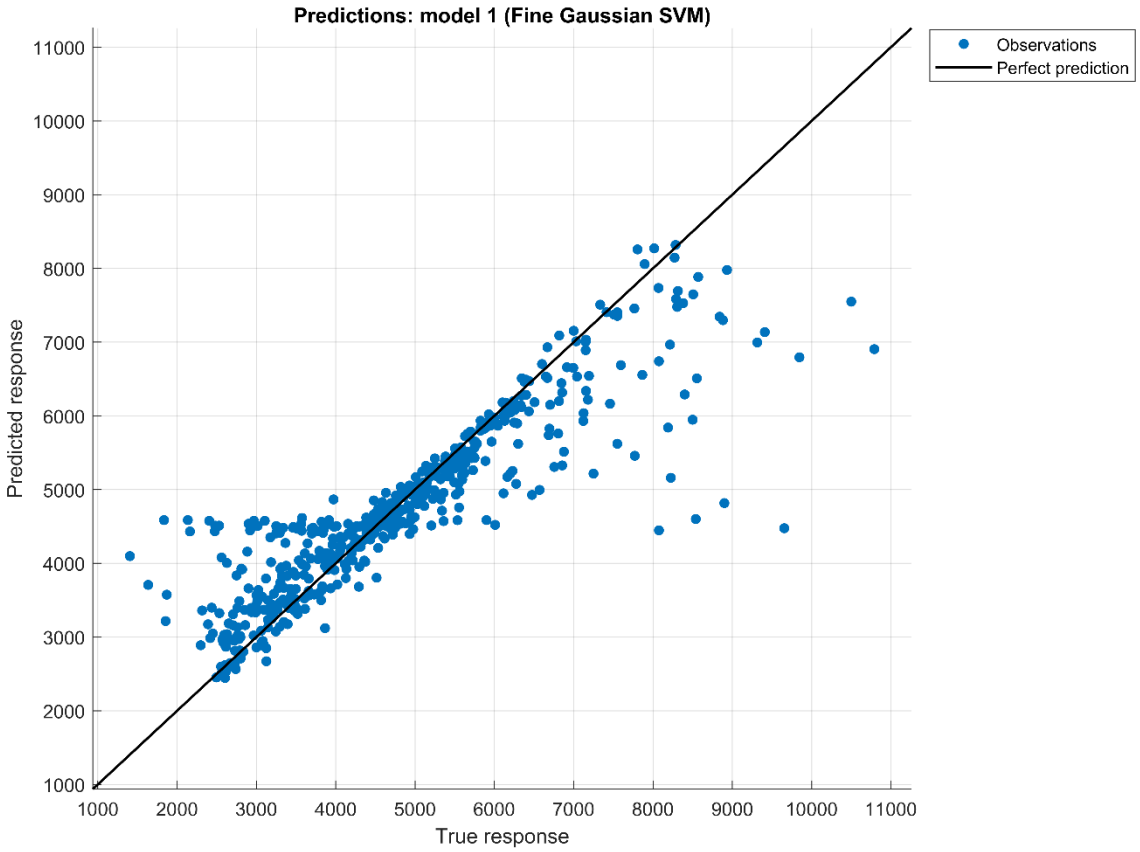

Figure 14: Performance of the SVM model