Introduction
The dew point pressure (DPP) is a principal parameter to be evaluated by Reservoir Engineers in the course of a gas condensate reservoir 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 (Barker & Total, 2005; Hassan et al., 2019; Rahimzadeh, Bazargan, Darvishi, & Mohammadi, 2016). Normally the condensate bank will not flow until its saturation increases beyond the critical saturation ($S_{cc}$) due to the effect 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. Several authors have proposed various machine learning algorithms (Perceptron, Support Vectors, Fuzzy Logic and Decision trees) for predicting the dew point pressure with varying levels of accuracy (Ahmadi and Ebadi 2014; Alarfaj et al. 2012; González et al. 2003; Najafi-Marghmaleki et al. 2016). This work compares the performance of these Machine learning algorithms against the number of input data.

Method and/or Theory
In this work, Artificial Neural Networks (ANN) [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 in predicting the dew point pressure using gas composition, specific gravity, the molecular weight of the heavier component and compressibility factor as input parameters.

Data Preparation: 535 experimental dew point pressure data points were obtained from published literature (Nemeth & Kennedy, 1967). A summary of the key data points are presented in table 1. The data was validated using the method outlined in (Bashbush, León, Mazariegos, Corona, & Unam, 2004). A graph of Log Ki vs Pressure was first prepared, then sections of the data indicating sharp divergence on the curve were removed. Thereafter, the graph of Log Ki vs Boiling Temperature (Tb) of each component was adjusted to 100% sum of gas mole percentages.

Model Development: First, a standard multiple linear regression (MLR) was used as a benchmark for comparing the performance of the machine learning models. The ANN model built was trained by using 13 input nodes to predict the dew point pressure output node. The architecture of the ANN model was optimized against Mean Squared Error. For the stochastic GBM, 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. The SVM model was designed to use all support vectors to produce an output. The decision function for the support vector algorithm is described in equation (1).

$$f(x(t)) = \sum_{i=1}^{n} \alpha_i y_i k(x', x(t)) + b$$

(1)
\( x^* \) – \( i^{th} \) vector of \( n \) support vectors, \( y_i \) – Class label, \( x(t) \) – \( t^{th} \) input frame vector, \( b^* \) – Optimization bias, \( \alpha^* \) – LaGrange multiplier

The algorithms for the each model was built using the R-framework and prediction results are presented in figures 4-7.

**Analysis:** Goodness of fit (\( R^2 \)) was performed on the various predicted models to describe its proximity with the experimental data (equation 2-3). The higher the \( R^2 \) value the better the predicted model fits the experimental data.

\[
R = 1 - \frac{\sum_{i=1}^{n} [(P_d)^{exp} - (P_d)^{pred}]}{\sum_{i=1}^{n} (P_d)^{exp} - \Delta P_d}
\]

\[
\Delta P_d = \frac{1}{n} \sum_{i=1}^{n} [(P_d)^{exp} - \Delta P_d]
\]

\( P_d \) – Dew point pressure, \( (P_d)^{exp} \) – 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 (4-6):

\[
E_i = \frac{1}{2} \sum_{i=1}^{n} |E_i| \quad (4)
\]

\[
E_i = \left( \frac{|(P_d)^{exp} - (P_d)^{pred}|}{(P_d)^{exp}} \right) \times 100 \quad (5)
\]

\[
RMSE = \sqrt{\frac{1}{2} \sum_{i=1}^{n} E_i^2} \quad (6)
\]

**Results and Discussions:**

**Model performance vs Number of Training Data:** For the first case, 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 result obtained were then scaled against results obtained for Multiple Linear Regression (MLR). The performance of the prediction models are presented in figure 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 while the testing data was increased to 160. Results from the error analysis of each model is presented in Figure 3 presents a normalized error analysis. Results from the analysis reveal that that neural network with single hidden layer and 5 neurons provided the best prediction whereas the performance of the SVM model for case 2 was almost equivalent to the best neural network model (NN4). The implication of this is that as training data decreases the accuracy of predictions from the SVM model increases.

**Model Performance vs Bias:** The 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 as seen in figures 9 and 10. This study hypothesizes 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 that the more complicated the data structure is, the more support vectors have to be selected to achieve reasonable accuracy.

**Model performance vs Data reliability:** The reliability of data used for the analysis is directly affects the accuracy of result obtained. Observing AARD prediction trend from each model (table 2-3), the study observes that SVM model exhibited the least error percentage consistently in the two cases. This implies that for datasets containing unverifiable points, the SVM model would be the most affected. This challenge is particularly not seen in the tree based models (GBM and XGB).
Table 1: Data Ranges used for this study

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

Figure 2: Normalized Error with respect to Multiple Linear Regression (case study 1)

Figure 3: Normalized Error with respect to Multiple Linear Regression (case study 2)

Figure 4: SVM predictions vs Experimental Dew Point

Figure 5: XGB predictions vs Experimental Dew Point
Conclusions

- The number of data points available for training affects the choice of machine learning algorithm to be used. For large data set neural networks shows better performance whereas SVM is given performs better for smaller datasets.
- Using the AARD as a measure of precision and RMSE was used as a measure of accuracy, this study reveals that SVM models significantly outperforms other models measure of precision. The implication is that SVM is more prone to bias in the presence of large outliers. However when the reliability of the data is not certain, SVM should not be used.

References