



저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원저작자를 표시하여야 합니다.



비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.



변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 [이용허락규약\(Legal Code\)](#)을 이해하기 쉽게 요약한 것입니다.

[Disclaimer](#)

August 2021

Master's Degree Thesis

**A Combined Study of Clustering
Techniques and Artificial Neural
Network on Predictive Models for Gas
Productivity in Shale Gas Wells**

Graduate School of Chosun University

Department of Energy and Resources Engineering

Anderson Maud Takyiwaa

A Combined Study of Clustering Techniques and Artificial Neural Network on Predictive Models for Gas Productivity in Shale Gas Wells

세일 가스정의 생산예측을 위한 군집화기법 및 인공신경망
복합모델 연구

August 27, 2021

Graduate School of Chosun University

Department of Energy and Resources Engineering

Anderson Maud Takyiwaa

A Combined Study of Clustering Techniques and Artificial Neural Network on Predictive Models for Gas Productivity in Shale Gas Wells

Advisor: Il Sik Jang

This thesis is submitted to Graduate School of Chosun University
in partial fulfillment of the requirement for Master's degree


April, 2021


Graduate School of Chosun University


Department of Energy and Resources Engineering

Anderson Maud Takyiwaa

Anderson Maud Takyiwaa의 석사학위논문을 인준함

위원장 조선대학교 교수 강성승 (인) 

위원 조선대학교 부교수 장일식 (인) 

위원 조선대학교 조교수 최태진 (인) 

2021년 5월

조선대학교 대학원

Acknowledgement

This work would have not been possible without the blessings of the most high God for granting me the knowledge, wisdom, and courage to pursue my master's research degree study. I would like to express my sincerest appreciation to the Korean Government and the National Institute for International Education of the Global Korea Scholarship Program (NIIED-GKS) for the financial support given me throughout the years of my study.

My gratitude also goes to Professor Il Sik Jang for his supervision, advice, and support in carrying out this thesis work. He actively provided the academic time and constantly advised me on the means to strengthen my ideas and arguments towards my research. Not forgetting Professor Kang Seong Seung and Professor Cho Tae Jin for their professional guidance in my studies and extensive review of my research work.

To my mom and sister, thank you for the prayers that gave me the strength to push until the very end. I appreciate my lab mates Oh Hyeon Teak Oppa and Lee Deukseon Oppa for helping with my Korean processes at every stage of my studies. And to my friends who supported me and encouraged me in the pursuit of my master's degree, I say a big thank you, most especially to Archibold, Fortune, Peter, and Samuel. As the saying goes, A real friend walks in when the rest of the world walks out. They have been real gems, present and ever willing to help anytime, given me the maximum support, inspiration, and encouragement even in the good and hardest times of my study.

<Table of Contents>

목차	i
List of tables	iii
List of Figures	iv
초록	viii
ABSTRACT	x
1. INTRODUCTION	1
2. BACKGROUND	5
2.1 Artificial Intelligence and Machine Learning	5
2.2 Literature Review	8
3. MACHINE LEARNING (ML) MODELS AND ALGORITHMS	12
3.1 Artificial Neural Network (ANN)	12
3.1.1 Neural Network Training Algorithms	13
3.1.2 Gradient descent backpropagation	16
3.1.3 The Quasi-Newton method	17
3.1.4 Levenberg-Marquardt algorithm	17
3.2 CLUSTERING ALGORITHM	18
3.2.1 K-Means Clustering	18
3.3 Evaluating Clusters	20
3.3.1 Elbow Method	20
3.3.2 Silhouette Value Index Plot	21

3.4 Principal Component Analysis	22
3.5 K-Nearest Neighbor(KNN)	23
4. GEOLOGY OF THE RESEARCH CONCESSION	26
4.1 Barnett Shale Concession	26
5. METHODOLOGY AND RESEARCH WORKFLOW	30
5.1 Data Generation and Preprocessing	32
5.2 Raw Data Extraction	33
5.3 Data Quality and Control Analytic Methods	36
5.3.1 Missing data	36
5.3.2 Duplicate data	37
5.3.3 Irrelevant Data	37
5.4 Descriptive Statistics for Variable Importance Analysis	38
5.4.1 Pearson's Correlation Plot	38
5.4.2 Color Maps / (Heatmaps) Matrix Plots	40
5.5 Processed Data and Statistics	41
5.6 Building Regression Neural Network Models	43
6. RESULTS AND INTERPRETATION	44
6.1 Base Case 1 (Present data Static Only)	44
6.1.1 Ann Model (Base Case 1)	48
6.2 Base Case 2 (Present data Static & IP gas)	53
6.2.1 Ann Model_Base Case 2	57
6.3 Base Case_3 (Present data Static, IP gas & Cum12 gas)	62

6.3.1 Ann Model_Base Case 3	65
6.4 Clustered Models	70
6.4.1 PCA & K-Means Clustering Results	70
6.5 ANN Model Results for Clustered Data Groups	74
6.5.1 Model Improvement of Cluster Groups	79
7. DISCUSSION	90
7.1 Outputs of the Neural Network: Network performance	90
7.2 Outputs of the Neural Network: Error Histogram	91
7.3 Outputs of the Neural Network: Regression Plot	91
7.4 Outputs of the Neural Training State Plot	92
7.5 Improvement in Predictive Models on Clustered groups	92
8. CONCLUSIONS	95
REFERENCES	97
자작물 이용 허용서	106

List of Tables

Table 5.1 Original Raw data extraction	34
Table 5.2 Input and Output data selection for machine learning models	35
Table 6.1 Input parameters for Static Case, base_Case 1	48
Table 6.2 Summary of average values for best 20 validated regression neural models on total present Static data	51
Table 6.3 Summary of future data prediction error in base Case_1	52
Table 6.4 Input parameters for Static&Ipgas, base_Case 2	57
Table 6.5 Summary of average values for best 20 validated regression neural models on total present Static&Ipgas data	60
Table 6.6 Summary of future data prediction error in base Case_2	61
Table 6.7 Input parameters for Static+IPgas+Cum12, base Case_3	65
Table 6.8 Summary of average values for best 20 validated regression neural models on total present Static,IPgas&Cum12 gas data	67
Table 6.9 Summary of future data prediction error in base Case_3	68
Table 6.10 Regression Input data for 2 grouping clusters	75
Table 6.11 Regression Input data for 3 grouping clusters	76
Table 6.12 Regression Input data for 4 grouping clusters	77
Table 6.13 Regression Input data for 5 grouping clusters	78
Table 6.14 Estimation of total improvement of clustered models_Case 1	86
Table 6.15 Estimation of total improvement of clustered models_Case 2	87
Table 6.16 Estimation of total improvement of clustered models_Case 3	88
Table 6.27 Verification of Improvement and Optimal Cluster number with Future data	89

List of Figures

Figure 2.1 The subset of Artificial Intelligence (AI)	6
Figure 2.2 Artificial Intelligence Applications in the Oil and Gas Industry over the years	8
Figure 3.1 Image of a biological neuron	13
Figure 3.2 A feedforward neural network	15
Figure 3.3 A diagram of a Recurrent neural network	15
Figure 3.4 Sigmoidal function plot	16
Figure 3.5 The elbow method for K-means	20
Figure 3.6 Typical silhouette visualization plot	22
Figure 4.1 Generalized stratigraphic column in the Fort Worth Basin	28
Figure 4.2 Newark East (Barnett Shale) field yearly production curve	29
Figure 5.1 Sample illustrations decline curves with their decline rate changes ·	30
Figure 5.2 Summarized research flow process	32
Figure 5.3 Diagram of measured depth and total depth	36
Figure 5.4 Correlation patterns (Pearson correlation coefficient)	40
Figure 5.5 Degrees in strength in coefficient values	40
Figure 5.6 A customized correlation matrix from Python Seaborn Library	41
Figure 6.1 Histograms of data distribution of the Input Measured Depth and Gross Perforated interval data	45
Figure 6.2 Pearson's Correlation plot for Case_1 Static data	46
Figure 6.3 Color Map plot for Case_1 Static data	47
Figure 6.4 The neural network architecture for the generated models in base_Case 1	49
Figure 6.5 Regression fit plot on training data_best validated model Case1	49
Figure 6.6 Regression fit plot on validation data_best validated model Case1 ..	49
Figure 6.7 Regression fit plot on test data_best validated model Case1	50
Figure 6.8 Performance plot on best validated model_Case1	50
Figure 6.9 Training state on best validated model_Case1	50

Figure 6.10 Error Histogram for simulated models on future dataset_Case1 50

Figure 6.11 Histograms of data distribution of Initial peak production gas and
Ground Elevation Interval data 54

Figure 6.12 Pearson’s Correlation plot for base_Case2 55

Figure 6.13 Color Map plot for Case_2, Static&Ipgas data 56

Figure 6.14 The neural network architecture for the generated models in
base_Case 2 58

Figure 6.15 Regression fit plot on training data_best validated model Case2 58

Figure 6.16 Regression fit plot on validation data_best validated model Case2 · 58

Figure 6.17 Regression fit plot on test data_best validated model Case2 59

Figure 6.18 Performance plot on best validated model_Case2 59

Figure 6.19 Training state on best validated model_Case2 59

Figure 6.20 Error Histogram for simulated models on future dataset_Case2 59

Figure 6.21 Pearson’s Correlation plot for base_Case3 63

Figure 6.22 Color Map plot for Case_3, Static&Ipgas&Cum12gas data 64

Figure 6.23 View of neural architecture for generated models in base_Case3 · 66

Figure 6.24 Regression fit plot on training data_best validated model Case3 66

Figure 6.25 Regression fit plot on validation data_best validated model Case3 · 66

Figure 6.26 Regression fit plot on test data_best validated model Case3 67

Figure 6.27 Performance plot on best validated model_Case3 67

Figure 6.28 Training state on best validated model_Case3 67

Figure 6.29 Error Histogram for simulated models on future dataset_Case3 67

Figure 6.30 Two group clustering on static data only 71

Figure 6.31 Silhouette plot on two group clusters 71

Figure 6.32 Variance explained by the principal components of static data only
on two group clusters 71

Figure 6.33 Three group clustering on static&Ipgas data 72

Figure 6.34 Silhouette plot on three group clusters on static&Ipgas data 72

Figure 6.35 Variance explained by the principal components of static&Ipgas data
on three group clusters 72

Figure 6.36 Four group clustering on static&Ipgas&Cum12 gas data 73

Figure 6.37 Silhouette plot on four group clusters on static&Ipgas&Cum12 gas data 73

Figure 6.38 Variance explained by the principal components of static&Ipgas& Cum12 gas data on four group clusters 73

Figure 6.39 Regression fit plot on train data_cluster label 2 (Two group clustering_static&Ipgas data) 79

Figure 6.40 Regression fit plot on validation data_cluster label 2 (Two group clustering_static&Ipgas data) 79

Figure 6.41 Regression fit plot on test data_cluster label 2 (Two group clustering_static&Ipgas data) 80

Figure 6.42 Performance plot on best validated model_cluster label 2 (Two group clustering_static&Ipgas data) 80

Figure 6.43 Training state on best validated model_cluster label 2 (Two group clustering_static&Ipgas data) 80

Figure 6.44 Error Histogram for simulated models on future dataset (Two group clustering_static&Ipgas data) 80

Figure 6.45 Regression fit plot on train data_cluster label 3 (Three group clustering_static data) 81

Figure 6.46 Regression fit plot on validation data_cluster label 3 (Three group clustering_static data) 81

Figure 6.47 Regression fit plot on test data_cluster label 3 (Three group clustering_static data) 81

Figure 6.48 Performance plot on best validated model_cluster label 3 (Three group clustering_static data) 81

Figure 6.49 Training state on best validated model_cluster label 3 (Three group clustering_static data) 82

Figure 6.50 Error Histogram for simulated models on future dataset (Three group clustering_static data) 82

Figure 6.51 Regression fit plot on train data_cluster label 1 (Four group clustering_static,Ipgas&Cum12 gas data) 82

Figure 6.52 Regression fit plot on validation data_cluster label 1 (Four group clustering_static,Ipgas&Cum12 gas data) 82

clustering_static,Ipgas&Cum12 gas data)	82
Figure 6.53 Regression fit plot on test data_cluster label 1 (Four group clustering_static,Ipgas&Cum12 gas data)	83
Figure 6.54 Performance plot on best validated model_cluster label 1 (Four group clustering_static,Ipgas&Cum12 gas data)	83
Figure 6.55 Training state on best validated model_cluster label 1 (Four group clustering_static,Ipgas&Cum12 gas data)	83
Figure 6.56 Error Histogram for simulated models on future dataset (Four group clustering_static,Ipgas&Cum12 gas data)	83
Figure 6.57 Regression fit plot on train data_cluster label 5 (Five group clustering_static&Ipgas data)	84
Figure 6.58 Regression fit plot on validation data_cluster label 5 (Five group clustering_static&Ipgas data)	84
Figure 6.59 Regression fit plot on test data_cluster label 5 (Five group clustering_static&Ipgas data)	84
Figure 6.60 Performance plot on best validated model_cluster label 5 (Five group clustering_static&Ipgas data)	84
Figure 6.61 Training state on best validated model_cluster label 5 (Five group clustering_static&Ipgas data)	85
Figure 6.62 Error Histogram for simulated models on future dataset (Five group clustering_static,Ipgas data)	85
Figure 6.63 Percentages of Improvement on clustered predictive model groups vs cluster numbers_Case 1	86
Figure 6.64 Percentages of Improvement on clustered predictive model groups vs cluster numbers_Case 2	87
Figure 6.65 Percentages of Improvement on clustered predictive model groups vs cluster numbers_Case 3	88
Figure 6.66 Plot of the sum of Rmse/total cluster data objects vs cluster numbers on Future data	89

초록

A Combined Study of Clustering Techniques and Artificial Neural Network on Predictive Models for Gas Production in Shale Gas Wells

Anderson Maud Takyiwaa

Advisor : Prof. Jang, Il Sik, Ph.D.

Department of Energy & Resources Engineering

Graduate School of Chosun University

유가스장에서 생산성을 예측하는 것은 기존 생산장의 최적 운영계획 수립뿐만 아니라 새로운 시추정에 대한 생산 능력을 미리 평가할 수 있게 하여 최적의 저류층 개발에 기여한다. 그러나 치밀한 비전통 셰일 저류층(tight unconventional shale reservoirs)에서 생산성을 추정하는 것은 매우 복잡하고 어려운 일이다. 생산감퇴곡선법(DCA), 생산량천이분석(RTA)과 같은 기존 방법은 셰일 저류층의 생산량 예측을 위해 적용되어 왔지만 이러한 전통적인 방법은 치밀 셰일 저류층의 현상을 완전히 반영하지 못하는 경향이 있다. 최근 석유 산업의 디지털전환 및 활용에 대한 지속적인 노력으로 머신러닝은 셰일 저류층의 생산 성능을 더 정확히 평가할 수 있는 방법으로 평가되고 있다.

이 연구에서는 바넷셰일 분지에 있는 524개의 유정에서 60개월 누적 가스 생산량을 예측하기 위해 머신러닝 군집화 및 회귀분석을 적용하였다. 머신러닝 입력자료로 다음과 같이 3가지 경우를 고려하였다. 즉, 지질자료, 유정완결자료 등 정적 데이터 (Static data)만 있는 경우, 정적 데이터와 초기 피크 가스 생산량(Initial peak gas rate) 데이터의 조합, 그리고 정적자료, 초기 피크가스 생산량과 더불어 12개월 누적 가스 생산량 정보를 사용한 경우에 대해 분석하였다.

불균질성이 강한 셰일저류층에 대한 최적 예측 모델을 구축하기 위해 군집화기법을 적용하여 생산장의 특성에 따라 분류하는 기법을 적용하였다. 즉, 거리기반 분

석법인 K-Means 클러스터링을 사용하여 유사한 특성을 보이는 생산정 그룹을 생성한 후, K-Nearest Neighbor 분류기를 학습하여 미래의 새로운 데이터에 적용할 수 있는 방법을 제안하였다. 미래 생산량을 예측하기 위해 신경망 회귀 모델의 적용 연구를 수행하였다. 특히, 생산정을 분류하는 경우와 분류하지 않는 경우의 생산성 예측의 정확도를 분석하였다. 또한 군집화 과정에서 군집의 수에 따른 예측 성능을 평가하였으며, 신규 시추 생산정의 수에 따른 군집 개수의 최적화를 수행하였다.

분석결과 클러스터된 신경망 학습과 테스트 모델들이 전체 (클러스터되지 않은) 데이터 세트를 사용할 때보다 학습데이터에서는 최소 19%에서 최대 52%, 테스트 데이터에서는 최소 9%에서 최대 24% 더 정확한 것으로 예측되었다. 이 연구에서는 셰일저류층의 생산자료에 대한 예측모델을 구성할 때 군집화 기법의 중요성을 재확인하였으며, 최적화된 군집모델과 회귀모델의 통합연구를 통해 미래 유정에 대한 생산예측 성능을 개선할 수 있었다.

Abstract

Predicting productivity in an oil-gas reservoir contributes to the development of an optimal reservoir by enabling the pre-evaluation of the production capacity for new drilling wells as well as establishing the optimal operation plan of the existing production wells. However, estimating productivity in tight unconventional shale reservoirs is very complex and difficult. Existing methods such as the Decline Curve Analysis (DCA) and Rate Transient Analysis (RTA) have been applied to predict the production volume of the shale reservoir, but these traditional methods tend not to fully reflect the phenomenon of the heterogeneities in shale reservoir. Machine learning is therefore being evaluated as a more accurate way to evaluate the production performance of shale reservoirs in recent efforts to digitally transform and utilize the oil industry.

In this study, machine learning clustering and regression analysis were applied to predict the 60-month cumulative gas production in 524 wells in the Barnett Shale Basin. The following three cases were considered as machine learning input data. First, if there is only static data such as geological data and oil well completion data. Secondly, a combination of static data and initial peak gas rate data, and thirdly static data, 12-months cumulative gas along with initial peak gas production. To construct an optimal prediction model for the heterogeneous shale reservoir, a clustering technique was applied to classify the data according to the characteristics of production wells. In addition, a method that can be applied to new data in the future by learning a K-Nearest Neighbor classifier is proposed after creating a group of production wells showing similar characteristics using K-means clustering, a distance-based analysis method.

To predict future production, a study on the application of the neural

network regression model was analyzed based on a system of model error-dependencies. In particular, the accuracy of the productivity prediction in the case of clustered and non-clustered production wells was analyzed. Furthermore, the prediction performance according to the number of clusters was evaluated during the clustering process, and the optimal number of clusters was estimated from cluster model groups with maximum prediction performance improvement on the future data.

The results showed that clustered neural network training models were predicted to be at least 19% and up to 52% (trained models) and at least 9% up to 24% (test models) more accurate than when using an entire (non-clustered) dataset. The study re-establishes the importance of predictive clustering techniques for building machine learning models utilizing data from shale reservoirs. By integrating the unsupervised clustering method and neural network regression models, it was possible to improve the performance of future production forecasts for future wells by classifying them into an optimal number of clusters.

1. INTRODUCTION

In petroleum engineering production, one of the key objectives is to make profits from the entire stages from exploration, drilling, development, production to the refinery. To make profit, it is required from engineers and operators from the onset to be able to develop and analyze the productivity of new wells or existing shut-in wells.

Over the years, many conventional techniques have been used in forecasting well productivity of “tight” shale reservoirs that indicate that oil and gas are not easily extracted. Operators have therefore learned how to effectively apply means such as hydraulic fracturing modeling, reservoir simulation, Rate Transient Analysis (RTA), and Decline Curve Analysis (DCA) to produce hydrocarbons in commercial quantities and estimate the productivity in formation wells. However, each of these traditional methods has underlying challenges to the physics of flow behavior and hydraulic fracture properties, heterogeneity, tightness and complexities of shale reservoir, absorption-desorption flow, and fracture variation of shale reservoirs.

DCA designed by Arps (1945) makes use of production data versus time series and assumes that flow in the reservoir must be in a pseudo-steady state but unconventional shale type reservoirs have very low permeability and are most of the time hydraulic fractured due to their tightness. Clarkson et al. (2016) stated that one of the biggest challenges in analyzing unconventional shale reservoirs is that their flow regimes in the transient flow take a very long time before entering into the boundary-dominated flow. Consequently using DCA that primarily depends on production data becomes very difficult and less feasible to estimate recoverable resources along with reservoir properties such as fracture half-length, permeability, drainage area, and fracture conductivity in

unconventional shale type reservoirs.

In addition, production analysis using RTA was carried out in the Marcellus Shale wells located in Greene and Washington Counties, Pennsylvania for identification of successful well completions methods and optimum production enhancements but techniques in RTA involve knowledge of normalized pseudo pressure data and material balance information to estimate productivity (Belyadi et al., 2015). Samandarli et al. (2012) performed production analysis from Barnett Shale wells using rate normalized pressure from build-up responses that required wells to shut in for operational reasons to highlight the heterogeneity and directly estimate formation permeability. However, the RTA required regular daily measurements of production data and wellhead flowing pressures. Therefore there are so many limitations to the production analysis when there is unavailable data on well flow rates and pressures and these methods cannot be widely applied to well formations with just completion data or limited availability of data

The petroleum industry has seen ongoing digitalization in different aspects of production analysis, the discovery and recovery of hydrocarbons, locating of drilling points in formations, complex multilateral drilling designs and implementations, forecasting into future production, optimization in production, interpretation of flow regimes and fluid flow behaviors, 3D-4D seismic data interpretation, estimating reservoir properties and reservoir phase behaviors by the use of machine learning. Machine learning can significantly assist real-time decision-making by operators and analysts in various workflow and designs.

Mohaghegh (2005) explained that the industry has realized immense potential offered by intelligent systems and soft computing that can integrate statistical (hard) and intelligent (soft) attributes capable of bringing real-time analysis and

decision-making power in the shortest possible time. Since the 2000s, many studies have utilized data-based supervised and unsupervised machine learning to predict productivity in unconventional shale reservoirs. Some researchers analyzed influential factors on shale gas productivity for predicting cumulative production rates (Bansal et al. 2013, Lolon et al. 2016, Wang and Chen, 2019). Zenko et al. (2008) addressed the two most important tasks of data mining which are predictive modeling and clustering and recommended that even though predictive modeling and clustering are regarded as two distinct techniques, unsupervised clustering technique is extended to regression problems. Han (2020) developed supervised learning models from Support Vector Machine (SVM) and Random Forest (RF) to improve the performance of productivity prediction by variable importance method and cluster analysis. It was discovered that the retraining of models through cluster analysis gave excellent predictive performances. Jung et al. (2003) mentioned that it is impossible to decide which distribution of clusters is best, given certain input patterns without an objective measure for clustering optimality. Dubes (1987) used the method of pattern recognition to answer the question "How many clusters are best" in his paper based on a Monte Carlo analysis of cluster validation. However the above studies have limitations in that, RF algorithm requires much computational power and the ensemble of decision trees suffers clearer interpretations and significance of each variable. The SVM is also inherently a binary classifier because of the way it creates the hyperplane to discriminate between two classes. The method of determining optimal clusters was based on strong assumptions and therefore heuristic. Unsupervised clustering although a useful technique is not common and under-researched in the petroleum engineering literature.

This study therefore, aims to develop a relationship between constraint-based

clustering and neural network regression predictive models. The study provides a straightforward mathematical and intuitive method of determining the measure of an optimal number of clustering with neural network regression models based on error dependencies of cluster models to establish the appropriateness of predictive modeling on homogeneous groups of data in shale reservoirs. By using neural networks, models can recognize hidden patterns and correlations in data, cluster and classify them, and over time continuously learn and improve on new data.

2. BACKGROUND

2.1 Artificial Intelligence and Machine Learning

Artificial Intelligence (AI) has been defined as a wide-ranging branch in computer science concerned with building smart machines capable of performing tasks that typically require human intelligence. Many definitions have been coined for artificial intelligence. One of the popular definitions by Nilsson (2010) describes AI as that activity devoted to making machines intelligent and its intelligence is the quality that enables an entity to function properly with foresight in its environment. To simply put, AI is an approach toward making machines behave like humans to make decisions either through logical reasoning or by experience. AI exists as an umbrella that comprises all computer programs that exhibit a cognitive power as humans do such as image recognition, learning through inferences, and self-improvement abilities. The two major subfields of AI include machine learning and deep learning. Mostly the difference between machine learning and AI are not clearly defined but one certain thing is that by conducting machine learning AI is also being executed as simply shown in Figure 2.1. While machine learning is probabilistic (output can be explained, thereby ruling out the black-box nature of AI), deep learning is deterministic Anirudh (2021) asserts. Machine learning can not be talked about exclusively without the mention of big data because any type of AI technique is largely dependent on its dataset for huge results. Machine learning presents itself as a handy tool by performing various programming techniques and algorithms that can extract useful information from data.

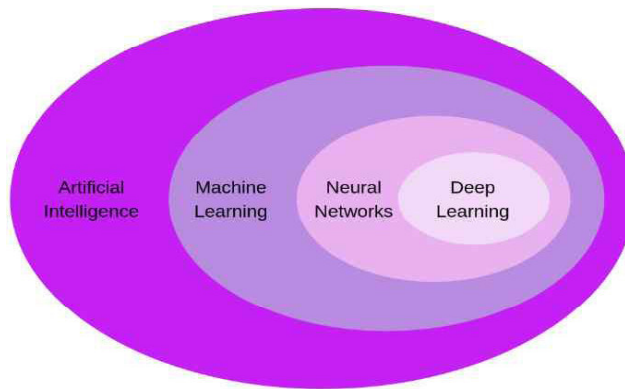


Figure 2.1: The subset of Artificial Intelligence (AI)
 (<https://ople.ai/ai-blog/3-subsets-of-artificial-intelligence/>)

Machine Learning (ML) is a branch of AI that involves systems learning from data to identify patterns within data and making decisions with minimal human operative power. The most important aspect of machine learning to the industry is the iterative process involved as models are exposed to new data and can learn from their previous computations and be independently adapt. ML involves many algorithms that can automatically perform complex mathematical computations on big data generated from field operations and quickly deliver more accurate results, identifying profitable opportunities and avoiding unknown risks that is very important criteria in the industry. Machine Learning comes with different ways of training algorithms with the kind of data that is utilized; labeled data and unlabeled data. The difference between the two is that labeled data just as the name implies are data that comes with labels or tags either a name, value or, type whilst unlabeled data have no tags or labels, just data in its non-adjudged class. From data types, three main families of machine learning are recognized.

Supervised Learning is a type of machine learning in which the algorithm is trained on a labeled data. This learning algorithm works by remembering the data that was initially ingested, formulates a model for the data input, and then

makes predictions based on the label on the data for a new set of data. This gives rise to two types of supervised learning models, classification models where the model predicts a state of data, and regression models predict a value or weight, size or quantity (Serrano 2019). Jason (2019) describes a classification problem as a predictive modeling task of approximating a mapping function from input variables to discrete output variables where the output values are called the categories also known as the labels whereas a regression predictive modeling involves approximating a mapping function from input variables to a continuous output variable. Because a regression predictive model predicts a quantity, the skill of the model must be reported as an error usually the Mean Square Error (MSE) or Root Mean Square Error (RMSE) in those predictions illustrated in the equations 1 and 2 where N represent the total number of data points and y and \bar{y} is the square of difference between the actual and predicted.

$$MSE = \frac{1}{N} \sum_{i=1}^N (y - \bar{y})^2 \quad (1)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^N (y - \bar{y})^2}{N}} \quad (2)$$

Unsupervised learning differs from supervised learning in that the data has no state and the ideal predictive class is unavailable but a lot of very useful information can be extracted from the data. The two main families in unsupervised learning are clustering and dimensionality reduction. Clustering algorithms such as K-means, Gaussian Mixture Models (GMM), hierarchical clustering, etc. split data into similar groups whilst dimensionality reduction algorithms such as principal component analysis, singular value decomposition, linear discriminant analysis, etc. simplify the dataset with the description of fewer features without losing the generality.

2.2 Literature Review

Machine Learning Applications In The Oil & Gas Industry

Applications of artificial intelligence and machine learning techniques have been used on the various aspects of petroleum engineering and many research studies have been carried in both conventional and unconventional reservoirs using ML to predict and forecast into future productions, to analyze and determine fluid phase behaviors in reservoirs, interpret flow regimes and reservoir heterogeneity, estimate reservoir parameters such as permeability and skin, determine drilling location points for new wells in existing formations, quantify and optimize productions, to mention a few. Since the 1990s, the most commonly used techniques of AI in the oil and gas sector have been the Genetic Algorithm (GA), the fuzzy logic, and Artificial Neural Network (ANN) as depicted in Figure 2.2 in the trend in AI over the years.

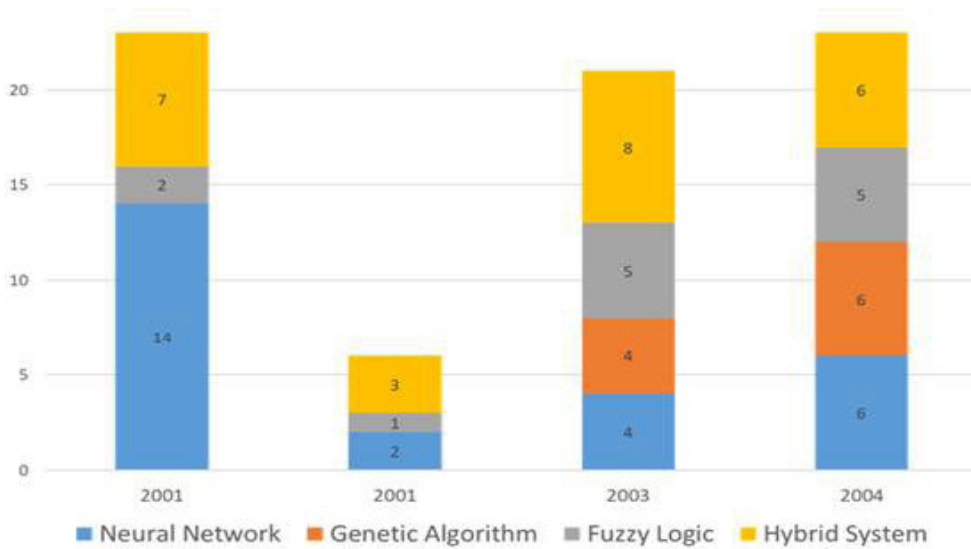


Figure 2.2: Artificial Intelligence Applications in the Oil and Gas Industry over the years (Cocchi and Mazzeo, 2018)

Alkinani et al. (2019) reviewed several good numbers of research studies that have been done by using ANN in the petroleum industry and grouped them into four; applications in explorations, drilling, production, and reservoir engineering. Wang et al. (2009) provided a brief review of machine learning and its applications by describing the structure and variety of machine learning methods including rote learning, inductive learning, analogy learning, and explained learning based on neural networks. In his descriptions, he highlighted rote learning as a memory that stores knowledge directly and calls for it any time when necessary not requiring bulk processing, inductive learning as one which applies an inductive consequence and guided by users, an analog learning describes the similarity between objects clearly and concisely by methods of similar definitions and similar transformations and finally an explained learning as a method that learns new knowledge in the process of acquiring new information. For most applications in the energy sector, the machine learning methods utilized are inductive, analog, and explained learning.

In a major advance in estimating well productivity in horizontal and vertical wells, Hassan et al. (2019) take artificial intelligence techniques further by applying them to fishbone multilateral wells, generating models that estimate the performance of fishbone wells in a heterogeneous and anisotropic gas reservoir. He demonstrated how artificial intelligence models such as artificial neural network, a fuzzy logic system, and radial basis networks were able to predict fishbone wells productivity with artificial neural network giving the most acceptable minimal error value. Another groundbreaking research was investigated using a data-driven approach to derive a root cause of slug flow (flow-type regime) in a subsea by training an ensemble of models with feature importance analysis to predict slug severity (Sandnes et al., 2019). The Model types included the random forest and lasso regression. Through this approach, a measure of the importance of the given variables in prediction was estimated.

Zhang (2015) demonstrated a clustering algorithm and data visualization approach for EOR prediction and steam flooding screening. He applied AI

method and hierarchical clustering to analyze the prediction of EOR methods for unknown reservoir conditions. In addition Krasnov et al. (2018) developed a proxy model using machine learning approach to enhanced oil recovery prediction and proposed that there is a need for a good and astute data processing technique since a change of parameters can occur several times a day. Machine learning becomes the definite choice for the industry as companies deal with a vast amount of data per every operation.

More recent evidence is investigated by Feder (2020) who describes an automated machine learning approach to determine the spatial variation of decline type curves for shale gas production based on existing data for forecast production performance. In his introduction, he explains that DCA is based on the fact that a similar production profile is expected from closest wells or wells with similar properties but DCA is manual and very subjective. Hence he combined clustering technique to particular decline curves and estimated the decline curves for a new set of variables. A related hypothesis for advanced research would be to further research on performance on cluster models generated and estimate an ideal clustering number which in reality is not an easy task.

Cao et al. (2016) utilized machine learning methods to forecast production from existing and new wells in unconventional assets using inputs of production history, pressure data, and operational constraints. Also, Chang et al. (2019) in a recent study developed key elements for well performance surveillance in a deep-water oil in the Gulf of Mexico using automated workflows like K-means clustering to identify shut-in periods, the SVM algorithm combined with kernel methods to identify transient flow-regime recognition and non-linear regression models to estimate reservoir and well properties assessing uncertainty. The result of his work proved to be effective and assisted in continuous well monitoring and link well productivity issues. More approaches using machine learning have been applied to shale formations to quantify production profiles for new target wells and one of them is the research work conducted by Bakay et al. (2019) in the producing wells of the

Duvernay Formation. The study also used machine learning classification and clustering methods, SVM, K-means, and functional classification and regression trees in determining the spatial variation in decline type curves for gas production. Based on the numerous successful studies that have been obtained using machine learning algorithms, a similar fashion is followed and carried out in this research to reach the objectives of the research work. For this research, the application of AI in the Barnett Shale gas formation is analyzed with a particular focus on productivity with clustering analysis and ANN model performance.

3. MACHINE LEARNING (ML) MODELS AND ALGORITHMS

This section describes all the machine learning algorithms that were utilized in the workflow of the research study.

3.1 Artificial Neural Network (ANN)

The neural network emerged from a popular machine learning algorithm named perceptron. The first systematic study performed by McCulloch and Pitts (1943) invented configurations of a single layer perceptron exhibiting inhibitory synapses upon each other under excitement states under the assumption that the time excitation of all neurons is the same. However, they emphasized the central problem of time parameter and the need to find a method of constructing a net of neurons so that specified afferent neurons are activated at a time only when conditions are satisfied. Rosenblatt (1958) also invented a mathematical model of a neural network following the progression of the work done by McCulloch and Pitts (1943) by proving that the memory of a perceptron is distributed in the sense that any association of units makes use of a large proportion of cells in a system and removal of a portion of the linked system would have an appreciable effect on the performance of any one association indicating pitfalls in all learned associations. Mohaghegh (2000) points out the drawback in his perceptron as not having general learning algorithms that can determine the weights of a particular calculation. After some years, a kick in neural network research began once again and Hopfield (1982) proposed new computational algorithms like the backpropagation that is used in carrying out the learning process in a neural network.

In general, the neural network in machine learning adapts the neurobiology concept depicted in Figure 3.1. By mimicking the human brain, a system of neurons is designed to generate the best possible results which are opened to

new changing inputs (Mhatre, 2020).

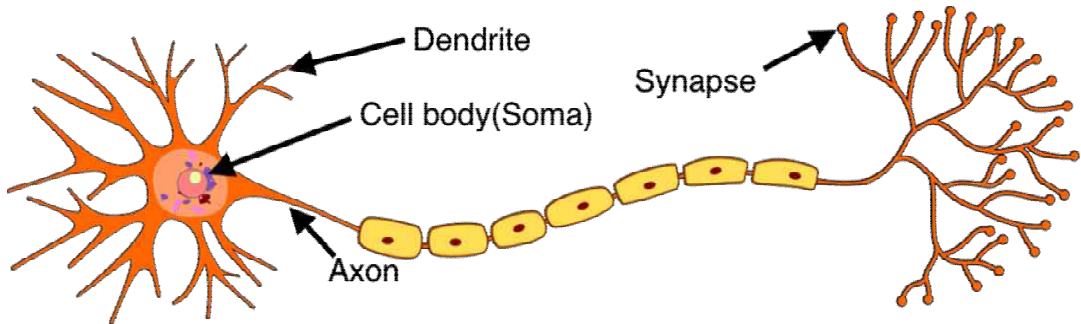


Figure 3.1 Image of a biological neuron (Wikimedia Commons)

There are many different types of algorithms used to train a neural network. These algorithms differ in precision, memory, and speed. All the algorithms in training a neural network are formulated on the minimization of a loss index (error) which measures the performance of a neural network. The loss function depends on adaptive parameters (biases and synaptic weights) in the neural network (Graves, 2011). This research does not seek to dive deeper into the mathematical equations that govern the evaluation of the loss index but the training algorithms in neural networks are briefly explained.

3.1.1 Neural Network Training Algorithms

ANN has been linked to biological neural computations of several units and neurons which are unified with coefficients (weights). Ghaffari et al. (2006) reiterate what was described by Agatonovic-Kustrin and Beresford (2000); Bourquin et al. (1997) concerning ANN described as Processing Elements(PE) that has weighted inputs, transfer functions, and a single output type. The PE simply balances the inputs to outputs by a training algorithm or connection by a formula rule. With the development of the Multilayer Perceptron (MLP) an advanced neural network which was developed to emend the problem of simple

layer perceptron (layers with no associated weight connections) investigated by Rosenblatt (1958) overcomes the limitation in a single layer perceptron by addition of one or more hidden layer to generalize more complex problems that are more quantified in deep learning. The operation of the ANN is usually impacted by the connection formula and learning algorithms. There are two main connection rules; feedback or recurrent and feedforward connections. The difference lies in what data is used as input after every iterative process. With a feedforward connection, the first fully connected layer of the neural network has connections from the network's input (predictor data) and the subsequent layers have connections from the previous layer. Each fully connected layer multiplies the input by a weight matrix and then adds a bias vector. The network learns from the weights and biases so that the output of the network either correctly predicts or classifies a value or label. The output of a layer does not travel from the output back to the input neurons whereas the opposite is described for a recurrent connection as the name implies shown in Figure 3.2 and Figure 3.3 respectively. By applying an activation function that utilizes the weights (w) and bias terms (b) such as the common sigmoid function in Figure 3.4, the network is trained. A small change in the weight would possibly cause a small change in output value. The training of the neural network involves the adjustment of the weight values to achieve the desired outcome of a minimal cost or loss function.

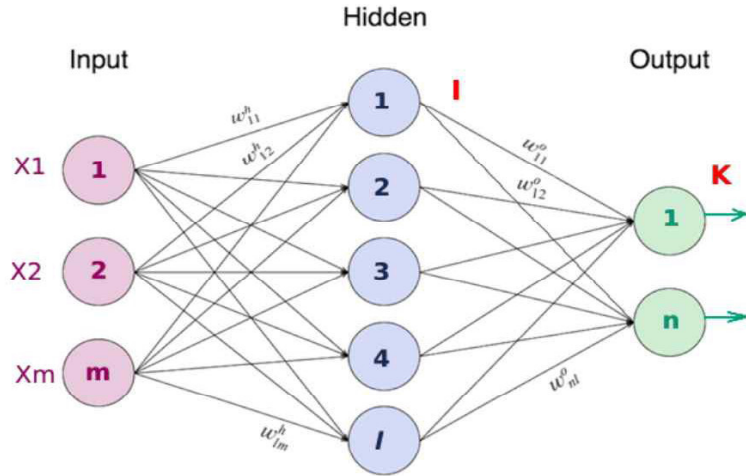


Figure 3.2 A feedforward neural network (Imam, 2020)

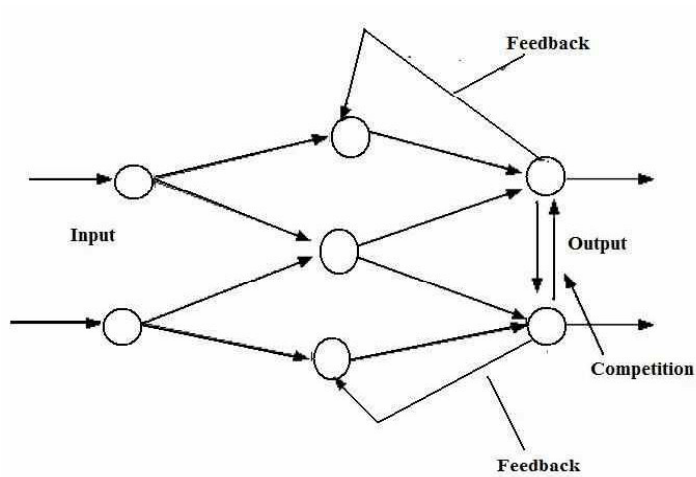


Figure 3.3 A diagram of a Recurrent neural network (Verma and Csed, 2021)

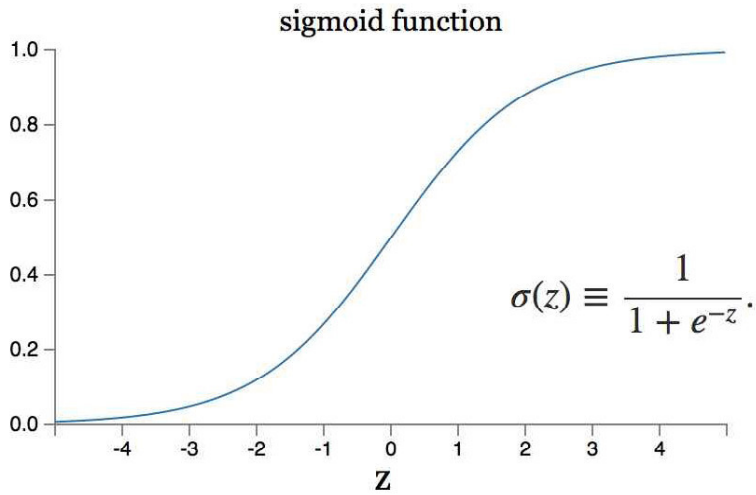


Figure 3.4; sigmoidal function plot

As stated earlier in the literature review, numerous algorithms have been used for training ANN. The most commonly used ones are the algorithms described in following sub-sections.

3.1.2 Gradient descent backpropagation

This algorithm is famous for training a feed-forward neural net. It simply allows updating weights by moving forward and backward until a local minimum is reached. So the network is trained to map a set of input data by iterative adjustment of the weights. Information from inputs is fed forward through the network to optimize the weights between neurons. Optimization of the weights is made by backward propagation of the error during the training or learning phase. The ANN reads the input and output values in the training data set and changes the value of the weighted links to reduce the difference between the predicted and target (observed) values. The prediction error is minimized across many training cycles (iteration or epoch) until the network reaches a specified level of accuracy. A complete round of forward-backward

passes and weight adjustments using all input-output pairs in the data set is called an epoch or iteration. If a network is left to train for too long, however, it will be overtrained and will lose the ability to generalize (Ghaffari et al., 2006).

3.1.3 The Quasi-Newton method

It is known to require many operations to evaluate and compute the index of the Hessian matrix. Szabo (2015) in his book, *Linear Algebra Survival Guide* explains the Hessian matrix as a square matrix whose elements are second-order partial derivatives of a given function. The Newton's method is an iterative paradigm considered for univariate error minimization. The idea is to start from some initial guess and solve a series of easier sub-problems on local models to determine a sequence of steps that lead to increasingly better estimates of the minimizer.

3.1.4 Levenberg-Marquardt algorithm

Levenberg-Marquardt algorithm is also known as the damped least-squares method. It is designed such that the loss functions take the form of the sum of squared errors by making use of the Jacobian matrix instead of the Hessian matrix and approaches the Newton method way of accelerating the convergence to the minimum (Ghaffari et al., 2006; Queseda, 2020.).

3.2 CLUSTERING ALGORITHM

Clustering is an analysis for finding a metaphysical meaningful group among data objects that share common characteristics or differ in attributes. Clustering plays an important role in data mining as it helps to discover useful information that is hidden within data groups. The technique measures the similarity of datasets and classifies target or output groups to identify similar objects within the same group and differences among groups(Han et al., 2019).

The earliest researcher on cluster analysis is traced back to the 1900s, when Pearson(1901) used the moment machine method to determine mixture parameters of two-single variable components and since then more immense research has been devoted to designing new clustering algorithms from the earliest most widely used K-means clustering (Wu, 2012) and now others such as the GMM, the hierarchical clustering, etc. Few research works that have been executed using clustering techniques have proven to be more successful and shown greater interpretation in data analysis.

Han et al. (2019) used clustering technique to improve the performance of productivity prediction with machine learning models for his study to analyze the similarities in his dataset and recreate the machine learning model for each cluster to compare training and test result. Based on clustering results, a machine learning classification was used to draw distinct geographic regions, within which a combination of geological, completion, and production factors were similar (Bakay et al., 2019).

3.2.1 K-Means Clustering

K-means clustering in machine learning is an unsupervised machine learning technique, one of the oldest yet widely used clustering algorithms. The K-means is also a simple iterative method that partitions data into an operator-specified number of clusters k . The use of the term “operator-specified” is because one of the drawbacks in using this algorithm is

that the user needs to make an initial guess towards the cluster number to partition the data in groups. Consequently, two given constraints are described (Wagstaff et al., 2001), “The Must-link” where the constraints specify that two instances have to be linked in the same cluster and “The Cannot-link” where the constraints denote that instances must not be placed in the same cluster.

The algorithm is initiated by defining a (k) centroid, one for each cluster group which is located distinctly because different centroid gives a different result. Hence a good cluster is one in which data objects with a given cluster exhibits good cohesion and cluster groups are rather far away from each other. The next approach in the algorithm is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed and an early group is done. At this point, it is needed to re-calculate k new centroids as centers of the clusters resulting from the previous step. After these k new centroids, a new binding has to be done between the same data points and the nearest new centroid. A loop is then generated. As a result of this loop, it may notice that the k centroids change their location step by step until no more changes are done i.e. centroids do not move anymore. Finally, Kodinariya and Makwana (2013) describes that the algorithm aims at minimizing an objective function, in this case, a squared error function. The objective function is given by the formula below;

$$W(S, C) = \sum_{k=1}^K \sum_{i \in S_k} \| y_i - c_k \|^2 \quad (3)$$

where S is a K-cluster partition of the entity set, represented by vectors $(i \in D)$ in an M-dimensional feature space, consisting of non-empty non-overlapping clusters S_k , each with a centroid c_k ($k = 1, 2, \dots, K$).

3.3 Evaluating Clusters

There are several approaches to evaluating clusters of data but the most common and simplest algorithms are the ones described within this section. By mere visualization plots, easy interpretations can be made about the separations similar to some ground truth that exist in the data.

3.3.1 Elbow Method

To address the issue of selecting the right K number there are different approaches to this, some of which are quantitative or by visualization. The oldest known method in cluster evaluation is the elbow method which is a visualization method. The measure is conceptualized in that the K value begins from 2 and increases by 1 whilst calculating the cost of the training. At some point, the K value drastically drops and by using the elbow and knee of the curve, the point at which the diminishing process no longer generates an additional cost is what is preferred.

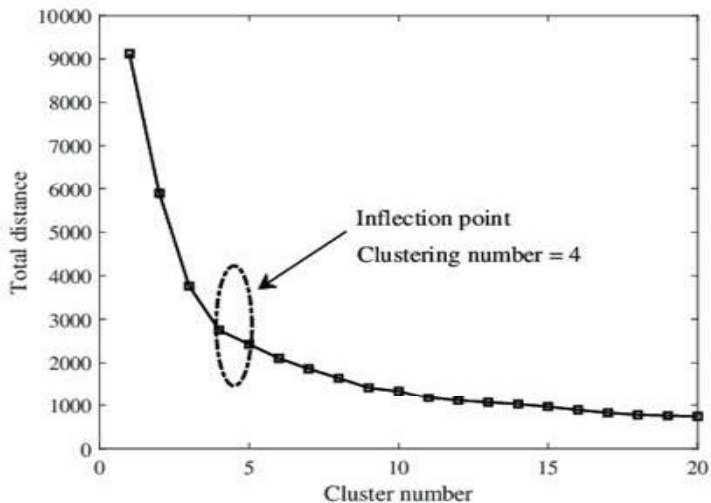


Figure 3.5 The elbow method for K-means (Zhang et al., 2020)

3.3.2 Silhouette Value Index Plot

Another measure of K-means which is estimated both quantitatively and can be interpreted through visualization plots is the silhouette measure. It measures how well an observation is clustered and estimates the average distance between the clusters. In the plot, it is displayed as the measure of how good observations are clustered and estimate the average distance between clusters. Quantitatively, it is estimated as average, over all clusters, of the silhouette width of the cluster data points. If x is a point in the cluster C_k and n_k is the number of points in C_k then the silhouette width of x is defined by the ratio in the equation (Brun et al., 2007);

$$S(x) = \frac{b(x) - a(x)}{\max[b(x) - a(x)]} \quad (4)$$

where $a(x)$ is the average distance between x and all other points in C_k and $b(x)$ is the minimum of the average distances between x and the points in the other clusters.

In choosing K using the silhouette plot, the largest average silhouette width, over different K , indicates the best number of clusters.

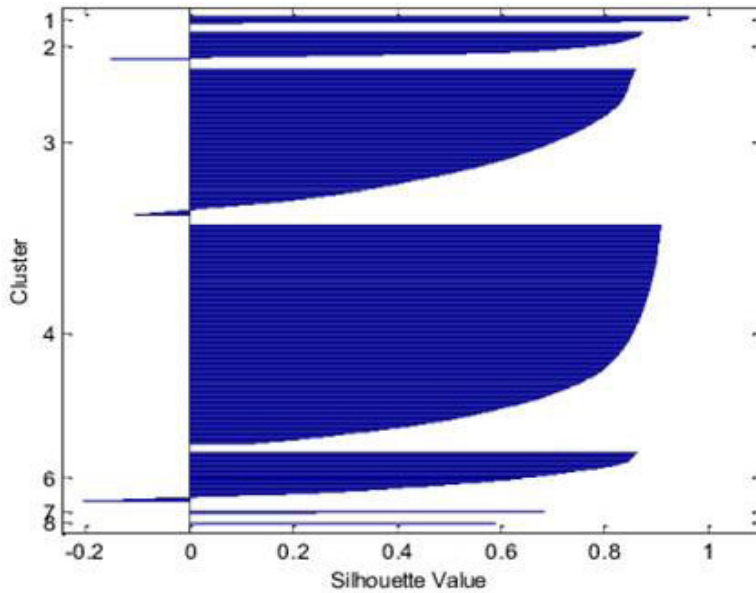


Figure 3.6 Typical silhouette visualization plot taken from (Tache et al., 2013)

3.4 Principal Component Analysis

Principal Component Analysis (PCA) is a machine learning multivariate–statistical dimension reduction technique with a history traced back to Pearson (1901). In his illustrative work, he elaborated on “best-fitting” lines and planes for case samples up to a number of variables in a correlated regression system. Abdi and Williams (2010) described the main goal in using this analysis is to extract the most important information from a table, to represent it as a set of new orthogonal variables which are known as the Principal Components (PC).

PCA depends upon the Eigen-decomposition of positive semi-definite matrices and the Singular Value Decomposition (SVD) of rectangular matrices. It is determined by eigenvectors and eigenvalues. Eigenvectors and eigenvalues are vectors and numbers associated with square matrices. Together they provide the eigendecomposition of a matrix, which analyses the structure of this matrix such as correlation, covariance, or cross-product matrices (Sarkar et al., 2017).

To attain the significant aspect of the data to use, a linear combination of the original is computed, that is, the PCs. Usually, the first principal component is a single axis in space which is expected to have the largest possible variance. The second principal component is another axis in space computed under the constraint of being perpendicular to the first component. Thus by projecting the observations on this axis generates another new variable. The variance of this variable is the maximum among all possible choices of this second axis. All other components are computed in a similarly and the new set of variables are interpreted geometrically as the projections of the observations onto the principal components, still maintaining the structure of the original set of variables. The values of these new variables for the observations are called “scores”. A more common algorithm for the sum of variances of the first few PCs is likely to exceed 80% of the total variance of the original data.

3.5 K-Nearest Neighbor (KNN)

K-nearest neighbor is a supervised machine learning algorithm that assumes that similar things exist in close contiguity based on distance. It is a very simple but unique algorithm that stores all the available cases and classifies new data values based on how its neighbors are classified. It is usually confused with K-means clustering techniques but they differ in that, in a KNN the objects are assigned to predefined classes whereas in K-means clustering the classes are rather to be defined. The KNN algorithm is sometimes described as non-parametric and a form of lazy learning i.e. it makes no assumption on the normality and generalization on the data and there is only a little time involved in training. This makes it very simple to use in both regression and classification problems.

The principle behind the nearest neighbor is to find a predefined number of training samples closest in distance to a new point and predict the label from these. The number of samples can be a user-defined constant (k-nearest neighbor learning) or vary based on the local density of points (radius-based

neighbor learning). The distance can, in general, be any metric measure: standard Euclidean distance, which is the most common choice defines the distance $d(x,y)$ between two instances x and y as estimated in equation 3 below. Other distance metrics such as cosine, Spearman, Hamming, Mahalanobis can be used, etc.

$$d(x,y) = \sqrt{\sum_{i=1}^n (a_i(x)^2 - a_i(y))^2} \quad (5)$$

where (x,y) are points in the Euclidean n -space, a_i is the origin of vectors in space and n is defined as the n -space.

A simple description of how the algorithm work is as follows. During a 10-fold ($k = 10$) for a supposed two groups of items A and B respectively in which a new item is being introduced into the group to know whether it is an “A item” or “B item”, the first aspect is to split the input samples into 10 partitions P_1, P_2, \dots, P_{10} that have equal sample sizes. It then uses the samples in partitions $P_2, P_3, P_4, \dots, P_{10}$ for the first training and the samples in P_1 are used for testing. Next is the samples in groups $P_1, P_3, P_4 \dots P_{10}$ and the samples in P_2 are used for testing. The process is repeated until each partition has been used for testing. A new item is classified by the majority of votes from the 10 neighbors and classified as perhaps B if 8 out of 10 neighbors are B or as “A” if 6 out of 10 neighbors are item A. The logic is not number-based but how much more neighbors share similarities. Just like K-means the user needs to define a k value for the data set. There is no specific defined structure to determine the best value of k however for most training done, the k value is usually preferred to be 10 but not ideal for all workflow. In parallel, choosing smaller values for k can be noisy and will have a higher influence on the result. The algorithm seems simple to use but many researchers have labeled few shortcomings that confront the algorithm. Goldberger et al. (2005) and Jiang et al. (2007) have conducted experiments to address these pitfalls with more

accurate metric distance functions to replace the standard Euclidean distance. They set the best neighborhood size to replace the artificial k input parameter and find some more accurate class probability estimation methods to replace the flat-out voting process. Goldberger et al. (2005) also combined Linear Discriminant Analysis (LDA), Relevant Component Analysis (RCA), and Neighborhood Component Analysis (NCA) transformations against the standard metrics for KNN as a process of optimization.

4. GEOLOGY OF THE RESEARCH CONCESSION

4.1 Barnett Shale Concession

The Barnett Shale is an onshore natural shale gas field in the United States. The field name for most part of the production portion of the formation is designated as the Newark, East (Barnett Shale Field) located in the Bend Arch-Fort Worth and Permian Basins West Texas. The sediments within the basin are traced back to the Mississippian period. It is the most common hydrocarbon source rock formation in the Fort Worth Basin with an area distribution of approximately $1.3 \times 10^4 \text{ km}^2$. The Barnett Shale includes two distinct layers namely the Upper Barnett Shale (thick in lithology) and the thin shale Lower Barnett (Jamshidnezhad, 2015)

Flippin (1982) also designed the generalized stratigraphic column in the Fort Worth basin as shown in Figure 4.1. The Barnett Shale is the primary source for petroleum in the Fort Worth basin sourcing conventional resources systems with both oil and gas (Jarvie et al., 2005). Jarvie et al. (2007) also mention that the age-equivalent of the basin spans from the Ordovician and conformably overlain by Pennsylvanian Marble Falls Limestone. The eastern section of the basin, the upper quarter of the Barnett Shale is separated from the Lower Barnett shale by the Forest-burg Limestone.

The Barnett Shale over the last two decades has been the most productive gas field in Texas in terms of daily production and constantly growing at the rate of more than 10% in Figure 4.2 (Bowker, 2007).

Many researchers have reported on the geometry and geological structure and petroleum engineers have also assessed the productivity of oil and gas and as well assessed the open natural fractural system in the shale asset. The fast progress of hydraulic fracturing technology in the previous years has led to the extraction of natural gas and oils from tens of thousands of wells that have been drilled in the shale asset with great commercial value. However, the

contexts of natural fractures have been a contentious topic critical to the production of Barnett gas. Bowker (2007) believed that these open fractures could lead to the migration of gas out of the shale to the overlying rocks.

Consequently, many research works have been conducted using machine learning and statistical techniques to analyze Barnett Shale. Awoleke and Lane (2011) used mechanisms of Self-Organizing Maps (SOMs) and K-means algorithms to identify clusters within well completions to predict the average water production and represent wells with high water throughput.

Mehana et al. (2021) explained in his research study that due to the ultra-low permeability zones characterized by shale reservoirs, a combination of the extended linear flow with the indeterminate onset of boundary-dominated flow challenges current deterministic analytic approach to forecast the estimated ultimate recovery (EUR). Additionally, they proposed unsupervised learning methodology k-means together with regularization to identify the optimal number of signals and validate the approach by hindcasting of the production data and achieved an excellent agreement.

GENERALIZED SUBSURFACE STRATIGRAPHIC SECTION					
SYSTEM	STAGE	GROUP	FORMATION		
CRETACEOUS	LOWER	COMANCHE	FREDRICKSBURG	GOODLAND PALUXY	
			TRINITY	GLEN ROSE	
PENNSYLVANIAN	UPPER	CISCO	THRIFTY	BRECKENRIDGE KING GUNSIGHTS SWASTIKA BUNGER	
			GRAHAM		
		CANYON	CADDO CREEK	HOME CREEK LIMESTONE COLONY CREEK SHALE RANGER LIMESTONE FLACID SHALE	
			BRAD	WINCHELL CEDAR TOWN ADAMS BRANCH UPPER BROWNWOOD SHALE	
			GRAFORD	PALO PINTO KEECHI CREEK ALESVILLE	
	MIDDLE	STRAWN	LAMPASAS	LONE CAMP	CAPPS LIME MORRIS SANDSTONE GARNER
				MILLSAP LAKE	GRINDSTONE CREEK LAZY BEND
				KICKAPOO CREEK	CADDO I
		ATOKA		CADDO II AND III PREGNANT SHALE BIG SALINE	SMITHWICK ATOKA CLASTICS
	LOWER	MORROW		MORROW MARBLE FALLS	SMITHWICK
				COMYN	
MISSISSIPPIAN	CHESTER		BARNETT SHALE		
	MERAMEC				
	OSAGE				
ORDOVICIAN		VJOLA	CHAPPEL		
		SIMPSON	VIOLA		
	CANADIAN	ELLENBURGER	SIMPSON HONEYCUT GORMAN TANYARD		
CAMBRIAN	OZARKIAN		WILBERNS		
	UPPER		RILEY HICKORY		
PRECAMBRIAN					

Figure 4.1 Generalized stratigraphic column in the Fort Worth Basin (modified by Flippin, 1982)

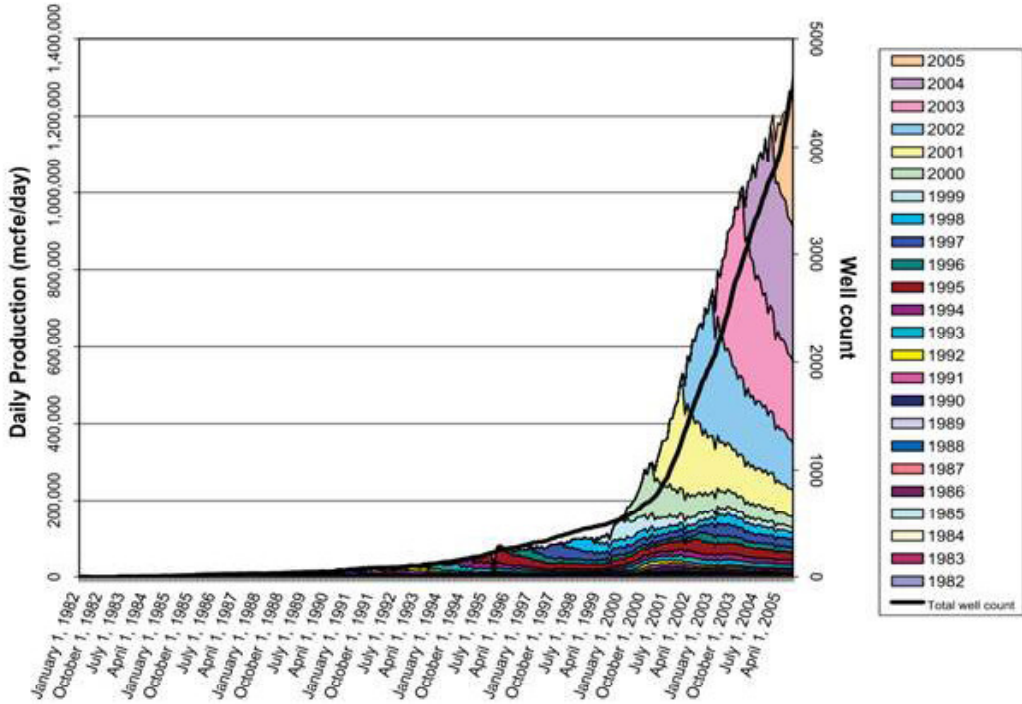
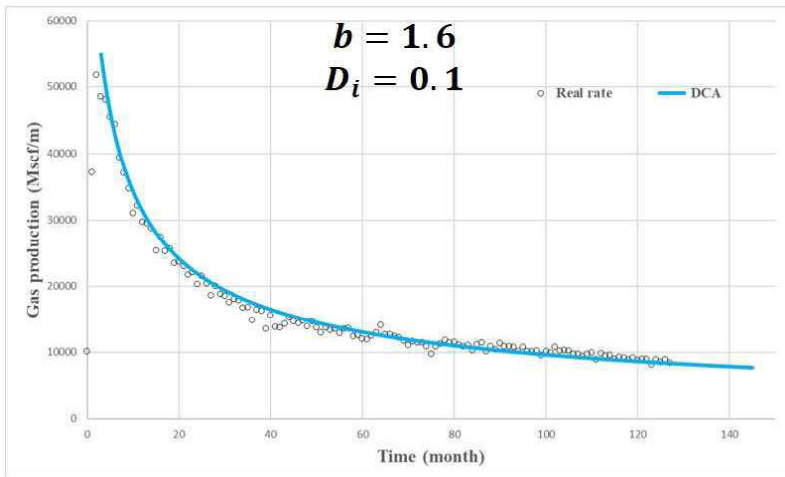


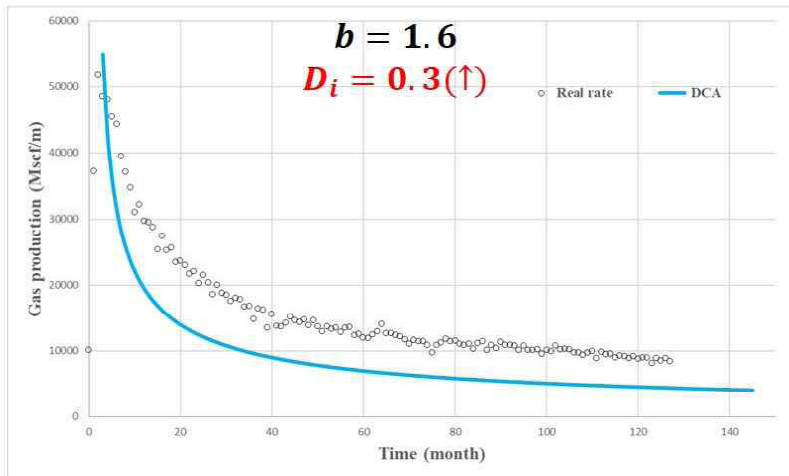
Figure 4.2 Newark East (Barnett Shale) field yearly production curve
(Texas Railroad Commission, 2005)

5. METHODOLOGY AND RESEARCH WORKFLOW

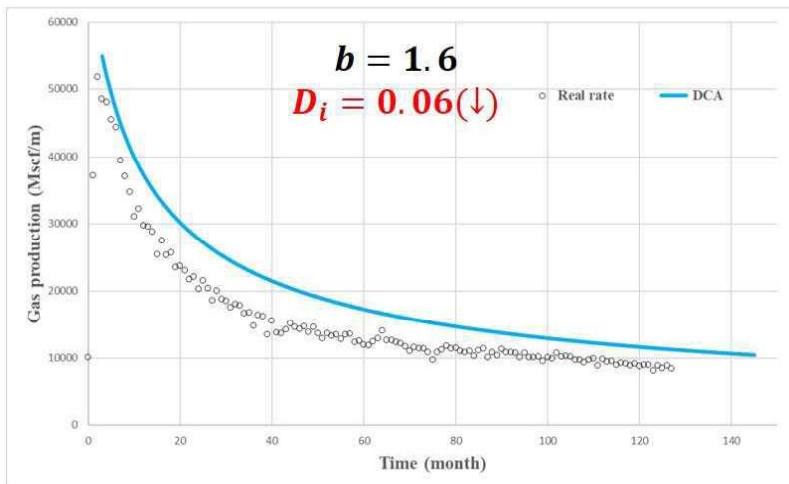
The well data utilized for the research are wells that have predefined decline curves from previous research studies carried out by Oh (2020) using machine learning algorithms to develop the Estimated Ultimate Recovery (EUR) using decline curve analysis (DCA) parameters; decline rate b , decline exponent D_i ; respectively. Principally wells that have defined decline curves (Figure 5.1) from the machine learning models generated to successfully predict the DCA parameters and their changes over 60 months of gas production for the predicted EUR compared to the real EUR was utilized and therefore a considerable level of confidence is preset for the wells production trend.



(a)



(b)



(c)

Figure 5.1 Sample illustrations decline curves with their decline rate changes (a) Normal (b) D_i increase (c) D_i decrease (Oh, 2020)

5.1 Data Generation and Preprocessing

This section summarizes the research workflow of the data acquisition, data preprocessing, machine learning models generation, and evaluation of model.

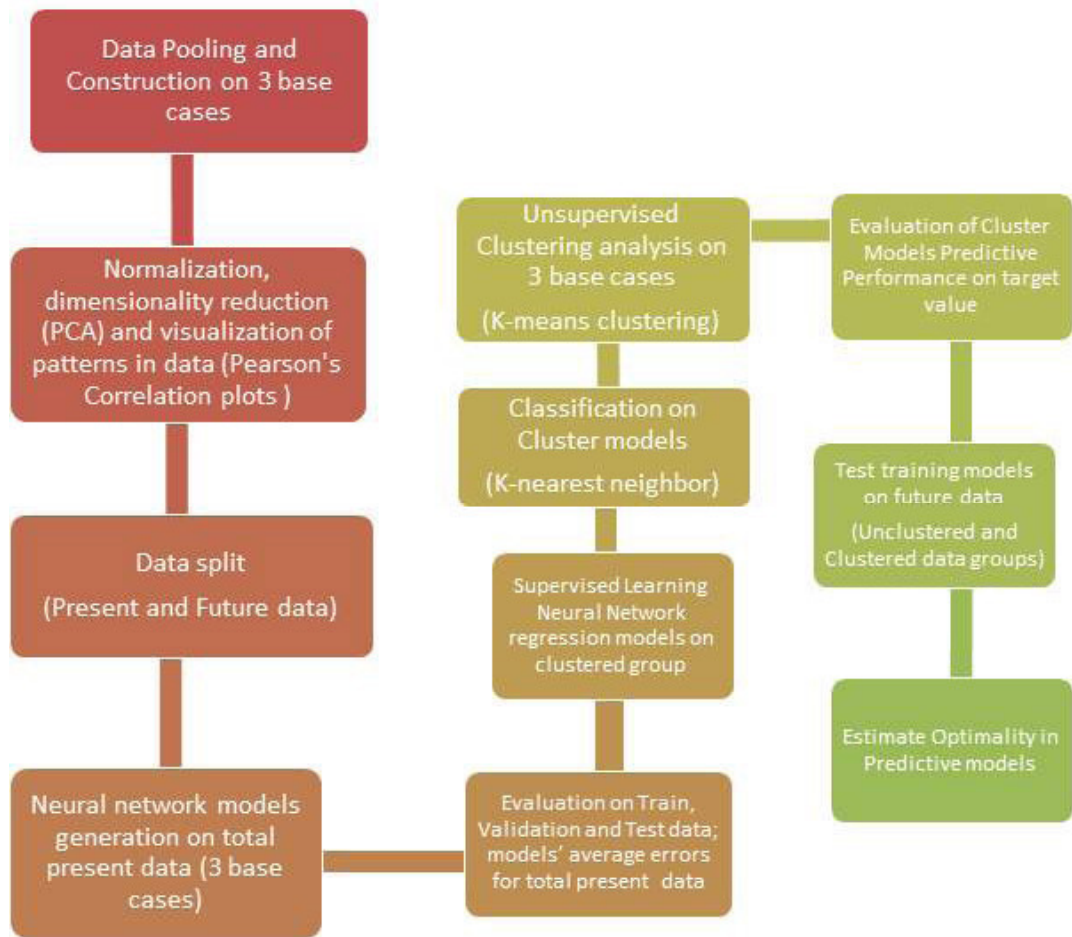


Figure 5.2 Summarized research flow process

5.2 Raw Data Extraction

The first stage of building every good machine learning model deals with preliminary processing of data and structuring the data to align with the objectives of the study. The prior condition set for the well data extraction was to select data only from the lower section of the Barnett Shale in horizontal wells only, to predict 60 months cumulative gas production from the neural network models using static, production rate and cumulative production data.

The static data were obtained from field geological data, well completion, and reservoir and log data. This represented the initial base case for the total set of data used. The second base case developed was the static data and peak production gas information. The same static input variable parameters were used and the third base case was designed for the latter together with 12 months of cumulative gas production as history data

Data features comprised of both numerical and categorical types. Before implementing the artificial neural network, clustering and classification algorithms, the data preprocessing involved removing all missing data, redundant data, senseless data, and visualizing the distribution of the data and their correlations to the target value.

Variable importance analysis was executed as a method for feature selection procedure using Pearson's Correlation plots and feature variables are selected based on their correlation value coefficients.

Table 5.1 Original Raw data extraction

No.	Original Data Selection	Feature Types
1	API index	Numerical
2	Formation (Barnett Lower)	Categorical
3	Well names	Categorical
4	Drill type section (Horizontal)	Categorical
5	Azimuth	Numerical
6	Latitude	Numerical
7	Longitude	Numerical
8	Total Fluids (bbl.)	Numerical
9	True Vertical Depth	Numerical
10	Density log	Numerical
11	Deep Resistivity log	Numerical
12	Gross Perforated Interval	Numerical
13	Gamma Ray log	Numerical
14	Horizontal Length	Numerical
15	Temperature	Numerical
16	Ground Elevation	Numerical
17	Measured Depth	Numerical
18	Neutron log	Numerical
19	Cumulative 60 months gas production	Numerical

Table 5.2 Input and Output data selection for machine learning models

Data Type	Name of Data
Output Data	Cumulative 60 Months Gas production
Input Data Categories	
Well Completion Data	Azimuth Longitude Latitude True Vertical Depth (tvd) Gross Perforated Interval Horizontal Length Ground Elevation Measured Depth
Log data	Density log Gamma Ray log Neutron log Deep Resistivity log
Reservoir data	Temperature Mean
Production data	Total fluids (bbl) Initial peak gas (IPgas) Cumulative 12 months gas (Cum12gas)

Both true vertical depth and measured depth were included as input parameters since the selection was based on horizontal wells only in the lower Barnett section. The measured depth differs from the true vertical depth (tvd) of the well in all but vertical wells. A pictorial illustration of the true vertical depth and measured depth can be seen in Figure 5.3

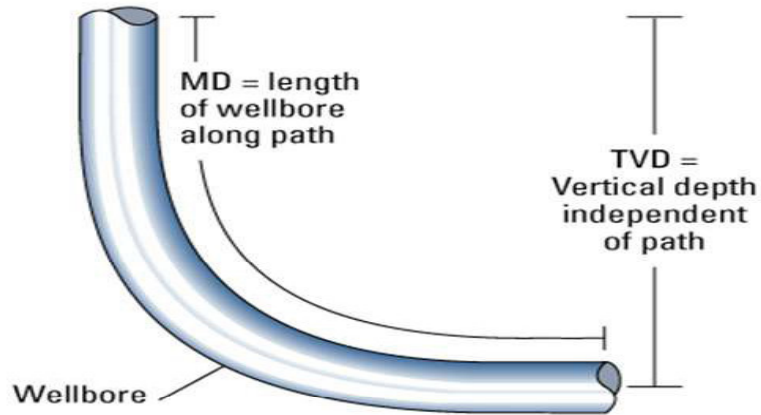


Figure 5.3 Diagram of measured depth and total depth
 (https://www.glossary.oilfield.slb.com/en/terms/m/measured_depth)

5.3 Data Quality and Control Analytic Methods

The data quality used for building models has a great influence on how well the model performs. For most established datasets three issues are involved; missing data, duplicate data, and senseless data. To rectify this issue, the same methods are applied to all 3 base case studies with the entire total data set and clustering group data set.

5.3.1 Missing Data

A very ordinary issue that is seen for most data set is the unavailability of data sections. To solve this, the user has to decide whether to set the missing data as NAN (Not Available Number) or to delete all information with unavailable data section. Concerning this work, all feature variables that had missing portions were deleted, resulting in a reduced number of observatory wells, a total of 524 out of over 1000 initial well data that was available for use for the research. This approach has its pros and cons because erasing data will produce a real continuous and unbiased set of data nevertheless the data may be enormously shortened which was the situation in this study. Also

ignoring missing data and working with the data in its entirety limits the probabilities of having to delete other available inputs although most machine learning algorithms will also fail to work on missing data feature variables.

5.3.2 Duplicate data

Input data selection for the model generation is not an easy task as the likelihood of selecting data types that are the same or data types with slightly different records in the project is high. Again all duplicate variables in this research were eliminated. For example, in the original raw data extraction, there was available information on the Bottom Hole Longitude and Bottom Hole Latitude in the database file. Upon using Pearson's correlation plots to perform the variable importance analysis (VIA), these two parameters exhibited duplicity with Latitude and Longitude despite having slight differences in records and thus they were removed to avoid a more biased model.

5.3.3 Irrelevant Data

Irrelevant data in this study are data variables that had very low or no correlation to the target value. All feature variables between -0.09 and 0.09 were considered irrelevant and were deleted from the input data used in building the models.

5.4 Descriptive Statistics in Normalization and Variable Importance Analysis of Data

Before building the neural network models, analyzing the input parameters by transforming the data and visualizing data from scatter plots, histograms can easily and quickly uncover patterns, get rid of anomalies that make analysis more complicated and reduce a large amount of data that may lower the performance of training models to a subset of interest. The data descriptive statistics involved normalization and performing sensitivity analysis.

Normalization is necessary as a form of data management because it restructures the database by a series of normal forms. Clean data does not mean getting rid of duplicates only. The first goal towards normalizing the data was to eliminate any redundancies that may occur. The other aim is to logically group data together, therefore data that relate to each other are stored together. The normalized data in this work returned the z-score values by scaling data to have a mean of 0 and a standard deviation of 1. The z-score value is simply defined as how far from the mean is a data point.

Correlation plots explain the strength and weakness in the relationship between two or more variable parameters. The correlation coefficients range between -1 to 1. A correlation coefficient of 1 means that for every positive increase in one variable, there is a positive increase of a fixed proportion in the other. In contrast, a correlation coefficient of -1 means that for every positive increase in one variable, there is a negative decrease of a fixed proportion in the other. 0 indicates there is no relation between the parameter values and are uncorrelated. Thus, for every increase, there is no positive or negative increase.

5.4.1 Pearson's Correlation Plot

This is a matrix of plots that serves as a standard measure of the strength of pairwise relationships. As a first step toward model specification, it is very useful to identify any possible dependencies among predictors. A good way to

do this is to visualize the correlation matrix and the most common measure of relations among data pairs is conducted with Pearson's Correlation. In statistics the most commonly used formula for estimating the Pearson's Correlation coefficient r is given by equation (6);

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}} \quad (6)$$

where n is the sample size, x_i, y_i are the individual sample points indexed with i and \bar{x}, \bar{y} denote the sample means

The Pearson's correlation plot from the Econometric toolbox in Matlab was utilized for the feature selection and specification analysis for the model. The matrix plot shows the histograms of the variables which appear along the diagonal of the matrix and scatter plots of the variable pairs which appear in the off-diagonal with their distinct patterns and the respective coefficient values on the scatter. Figure 5.4 and Figure 5.5 describe generalized and common patterns in data relationships and a measure in strength and direction of relationships with coefficient values respectively.

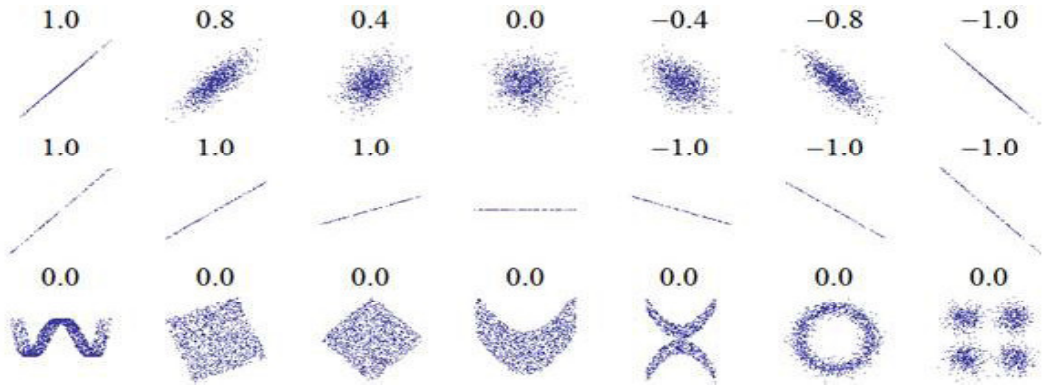


Figure 5.4 Correlation patterns (Pearson correlation coefficient - Wikipedia)

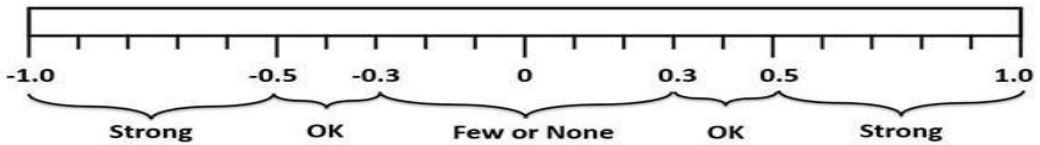


Figure 5.5 Degrees in strength in coefficient values

5.4.2 Color Maps (Heatmaps) Matrix Plots

Another visualization technique for variable importance feature selection before model training application was the use of color maps correlation matrix plot to identify features that affected the target variable the most. This is also known as the heatmap plot in some software applications. They are very useful for data set with many columns because visualizing data is easier to understand than reading from tabular data. With heatmap, colors of weakly correlated pairs visually disappear in plots and strongly correlated pairs are more distinct in their color palette. The color matrix plot maps have the color bar with corresponding color values to match low and high correlated pairs.

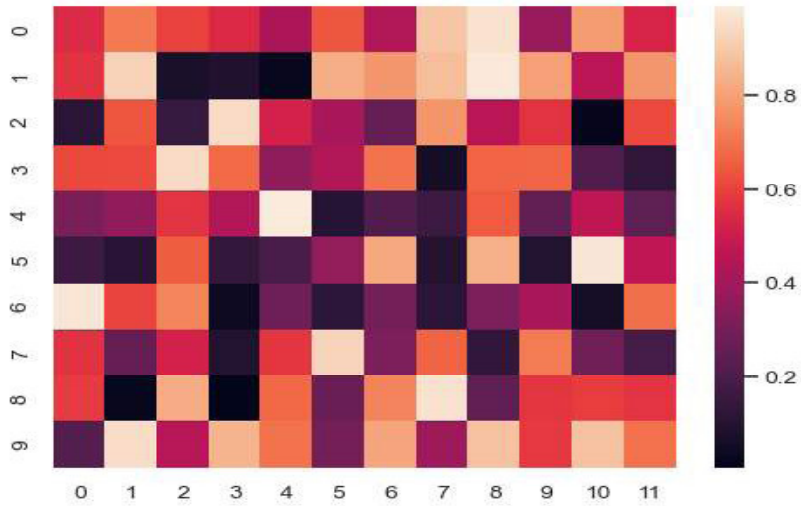


Figure 5.6 A customized correlation matrix from Python Seaborn Library

5.5 Processed Data and Statistics

After preprocessing the data and visualizing the distribution of data inputs, models were generated for a total number of 524 horizontal wells. The data were initially normalized and split into "present" data set which constituted 96% observations). The models were analyzed for the 3 base case studies; Case_1, Case_2, and Case_3 on the present data to predict the target 60 months cumulative gas production. The present data in each case study were clustered into several groups with a group label and the future data points were assigned to these group labels with the same analysis to develop predictive models that can estimate 60 months of cumulative gas production with minimized error. Based on the error output, an optimal number of clustering was then determined for the present and future dataset.

The clustering was done using the K-means clustering algorithm combined with dimensionality reduction on the data using PCA. Clustered groups were given labels and the future set which served as unseen test data was used to

validate the clustered models by applying the K-nearest neighbor classifier algorithm.

- Case_1 (Static data only) derived from the well completion, log and reservoir data.
 - Present data (504 observatory wells data)
 - Future data (20 observatory wells data)
- Case_2 (Static data and Initial peak gas production= (IP)). The same static data used for base Case_1 was used together with the initial available gas production. For the subsequent part of the work, the initial gas production is represented as IPgas.
 - Present data (504 observatory wells data)
 - Future data (20 observatory wells data)
- Case_3 (Static data, IP gas data, and Cumulative 12 months production gas= (Cum12 gas)). An inclusion of history production with the same data used in base case 2. The cumulative 12 months gas production is also shortened as Cum12gas for simplicity.
 - Present data (504 observatory wells data)
 - Future data (20 observatory wells data)

The future data points were set aside as unseen data to validate the clustered models and evaluate the neural network performance also by the measure of their root mean squared errors (RMSE).

To determine the optimal number of clustering and assess how well data points are clustered for each base case models, the silhouette measure of clustering was applied on the low dimensionality clustered data, and finally, neural network models were generated and the best 20 validated models were selected from a loop of 1000 generated models based on their sorted RMSE

values. A measure of clustered model percentages of improvement was estimated by calculating the improvement at each case's total average training validation and test error values to substantiate the purpose of clustering and determine the best cluster number on the dataset. The present data set was clustered from 2 to 5 groups for all base case studies and the 20 best validated ANN models selected and these network models were used in estimating the predictive performances on the future data as well.

The neural network generated in this research is a shallow neural network with sigmoid hidden neurons and linear output neurons to fit the multidimensional matrix. The network is trained on a scaled-conjugate backpropagation algorithm. The data divisions use random divide parameters and compute the mean square errors as the performance in the model using the minimum excludant (MEX) calculations in all models.

5.6 Building Regression Neural Network Models

In all 1000 models that were generated, the neural network was fitted with a single hidden with 20 neurons. The "x-by-504" matrix of the total present data (for each base case study) contained the input feature values represented by x into the regression neural net and the "y-by-504" matrix of the total present data contained the associated target feature values represented by y in the data observations. The data set was divided into 80% training data, 10% validation, and 10% test data for the total and clustered cases based on a random split in data ratios. The fitted neural network was then used to train a feedforward, fully connected neural network for the regression. The sigmoid function was activated in the single hidden layer and the linear regression activation function on the output layer respectively. The goal of minimization was set to $1.0e-6$.

6. RESULTS AND INTERPRETATION

This research was conducted using machine learning algorithms by a combination of unsupervised clustering techniques, supervised artificial neural regression networks and K-nearest neighbor classification techniques. These learning algorithms were deployed to build models that can predict the target value of 60 months cumulative gas production in the Barnett formation and establish a rational concept behind the importance of clustering on predictive performances of models, based on similarities and contrasts among data objects. The results were analyzed for the objective of the research to establish clustered model groups as a better choice in shale gas productivity models than the use total data in estimating the predictive performance of neural network models.

6.1 Base Case 1 (Present data Static Only)

The Case_1 base case had an original 15 variable parameters and 504 observations representing the total present static data set including the target. Before building the ANN models to predict the target value, the data was preprocessed and visualization in patterns was plotted using the correlation matrix and color map plot respectively. The condition set was to remove all feature variables that have coefficient values close to 0.0, hence all coefficient values that range between -0.09 to $+0.09$ with respect to the output variable were removed in all study cases. Distributions of variables that showed positive correlations were visualized in histogram plots to observe their skewness. The Measured Depth and Gross Perforated Interval variables displayed in Figure 6.1 both indicated positive (right) skewed distribution. Also from the Pearson's correlation plot in Figure 6.2, the plot shows that the Longitude variable had

the highest positive linear relationship (coefficient value of 0.43) and the Density log variable term had the highest negative linear relationship with the target value in the static data case. However, no variable term had a perfect positive and negative linear relation with other pairs. For example, the Measured Depth had a high correlation with the Gross Perforated variable term (0.84) but had zero correlation with Density log(0.06) and Ground Elevation(0.05) terms respectively. The azimuth term had little or no correlation with almost all variable terms. The distribution of the data values is represented in the histogram plots. The minimum measured depth of the drilled pipes in the bottomhole assembly ranges from 8,701ft and maximum depth 15,825ft. The gross perforated interval, the section of the wellbore that was prepared for production by creating channels between the reservoir formation and the wellbore ranges from a minimum of 847ft to a maximum of 10,144ft. Also from the color map plot in Figure 6.3, the relationship between the data values is correlated with each other by mapping the values to levels in colors. For this base Case_1, all variables lie within the conditional range of selection, hence all the 14 input features were used to build the regression artificial model.

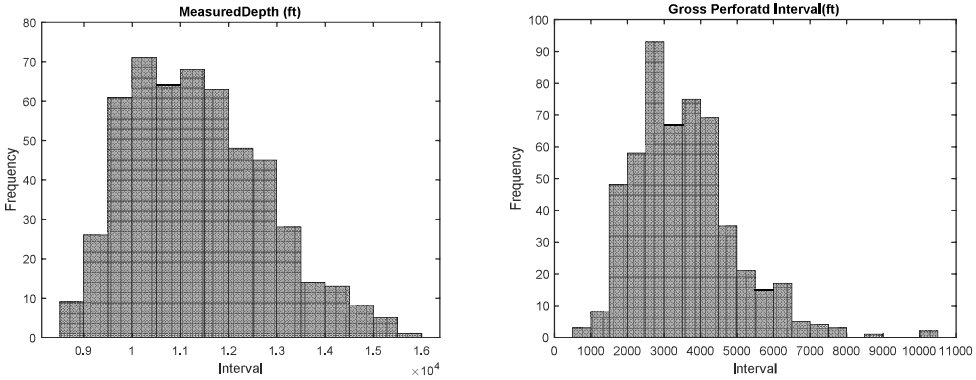


Figure 6.1; Histograms of data distribution of the Input Measured Depth and Gross Perforated interval data.

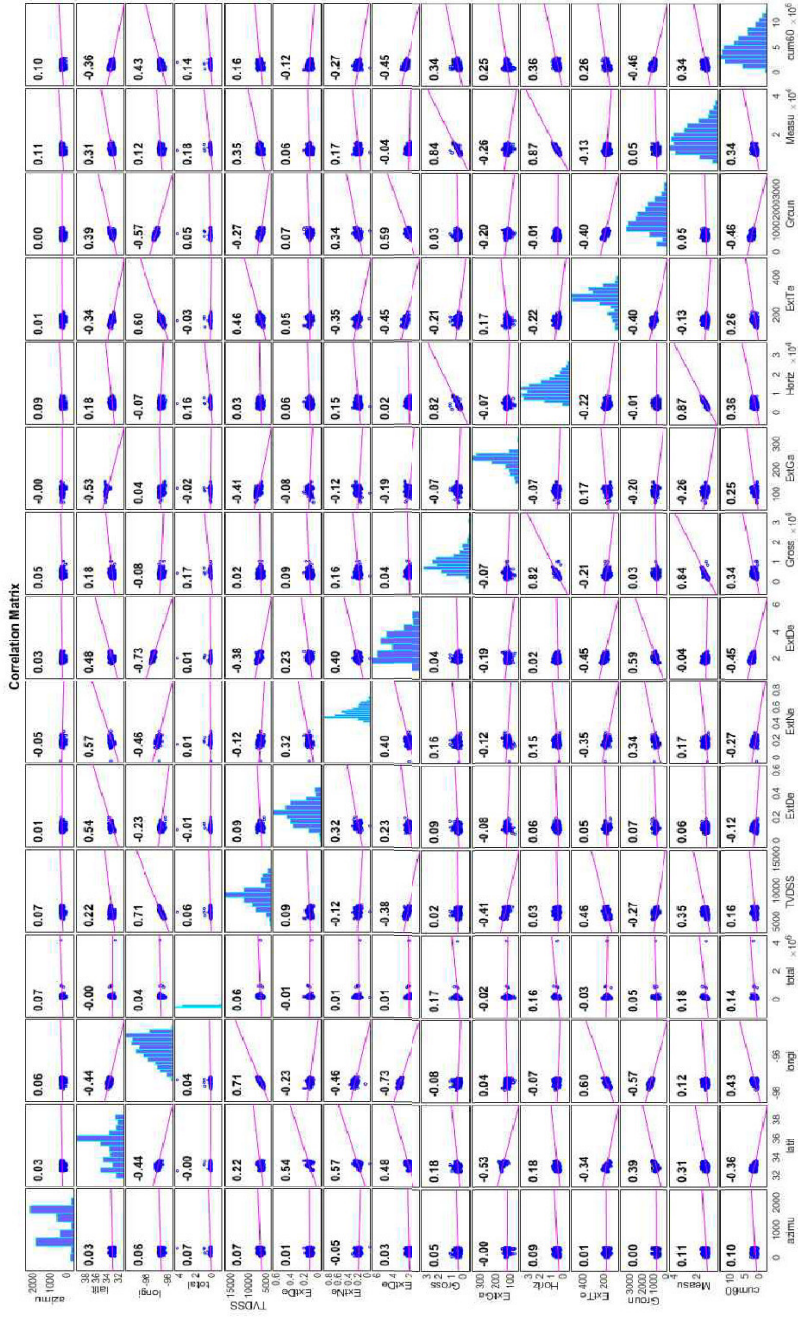


Figure 6.2: Pearson's Correlation plot for Case_1 Static data

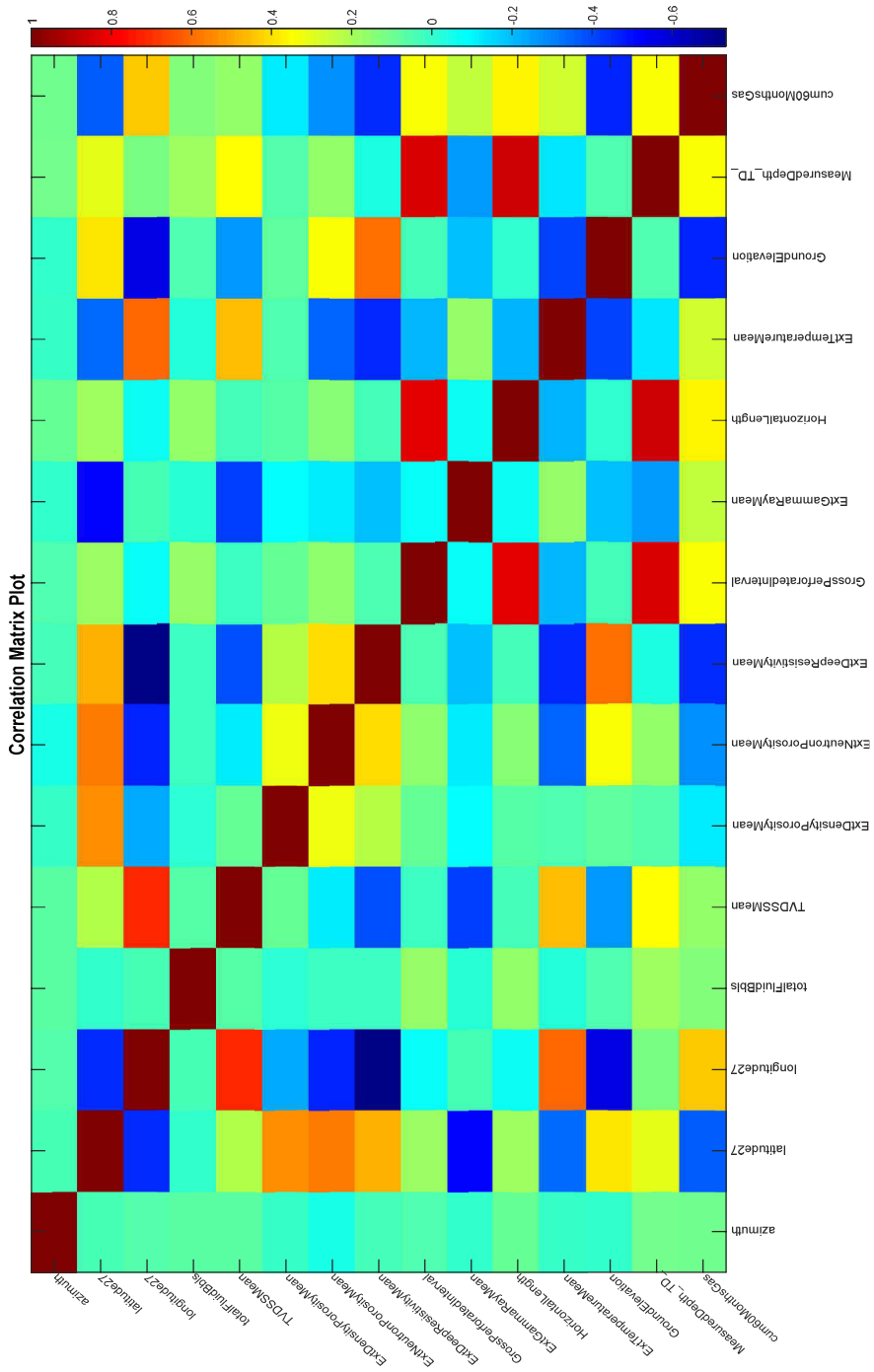


Figure 6.3 Color Map plot for Case_1 Static data

6.1.1 ANN Model (Base Case 1)

From a looping process, 1000 arrays of shallow artificial neural network regression models were generated on the total present data (504) observations with 14 input features and 1 target value in Table 6.1. The average root mean square errors from the training, validation, and test errors were estimated and the best 20 models based on the least validation errors were selected. The issue with the neural network model is that predictive performance changes for every execution. Hence to get an ideal representation of the predictive performance of the model, a seed number was set before building the models to obtain the same results for every execution. The best 20 validated models were used to test the static future data and the average RMSE(s) recorded.

Table 6.1 Input parameters for Static Case, base_Case 1

Output data	Cumulative 60 months gas
Input Data	<ol style="list-style-type: none"> 1. azimuth 2. latitude 3. longitude 4. total fluids (bbl.) 5. true vertical depth_TVD (ft.) 6. Density log 7. Neutron log 8. Deep Resistivity log 9. Gross Perforated Interval (ft) 10. Gamma Ray log 11. Horizontal Length (ft.) 12. Temperature 13. Ground Elevation (ft) 14. Measured Depth

The training data, validation data, and test data for base Case_1 were 404, 50, and 50 respectively which represent 80%, 10%, and 10% each of the total 504 observations. The regression plots, error performance plot, and training state plot for the best model(model 280) out of the 20 selected models based on the sorted validation errors are represented from the Figures 6.5 to 6.10 below. The MSE(s) and RMSE(s) for the selected models and the averages are summarized in Table 6.2.

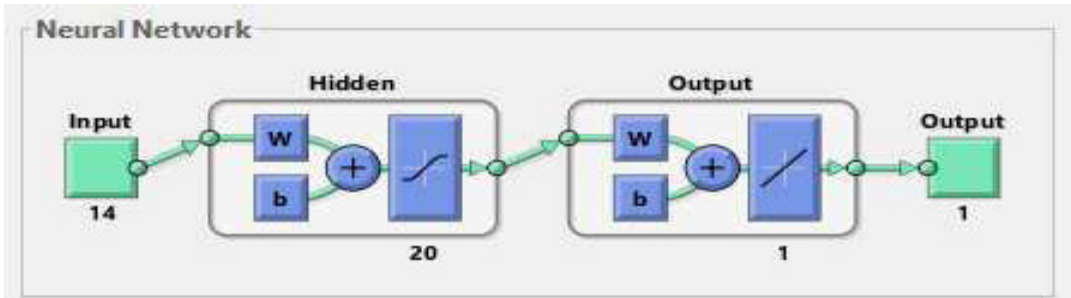


Figure 6.4 the neural network architecture for the generated models in base_Case 1

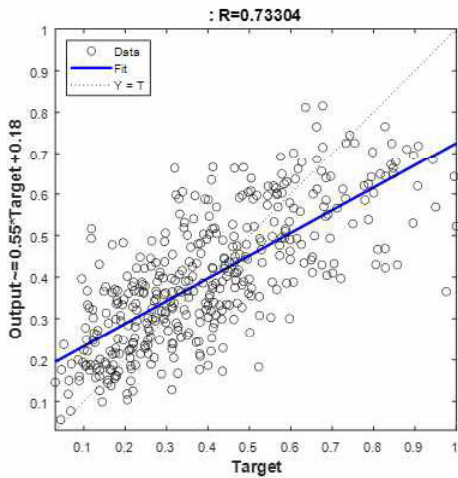


Figure 6.5 Regression fit on training data_best validated model Case 1

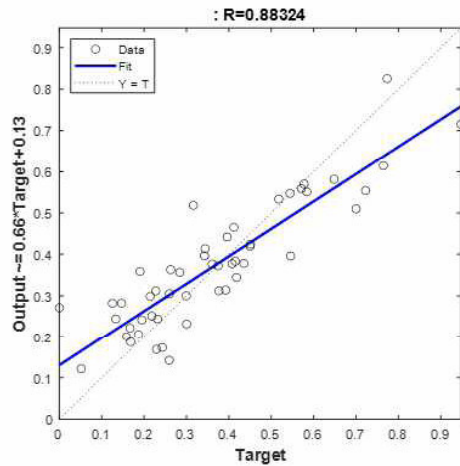


Figure 6.6 Regression fit on validation data_best validated model Case 1

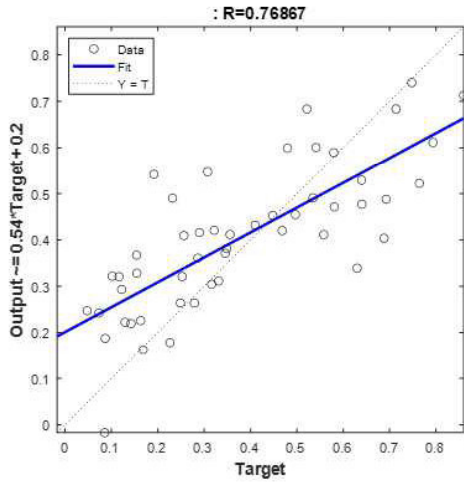


Figure 6.7 Regression fit on test data_best validated model Case 1

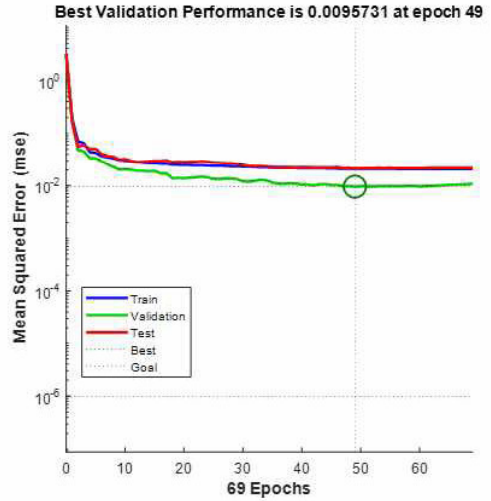


Figure 6.8 Performance plot on best validated model_Case1

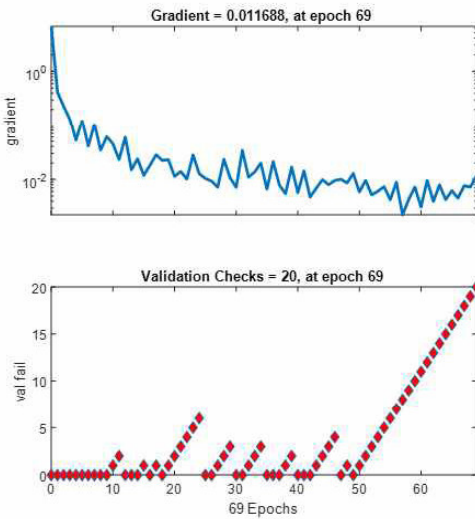


Figure 6.9 Training state on best validated model_Case 1

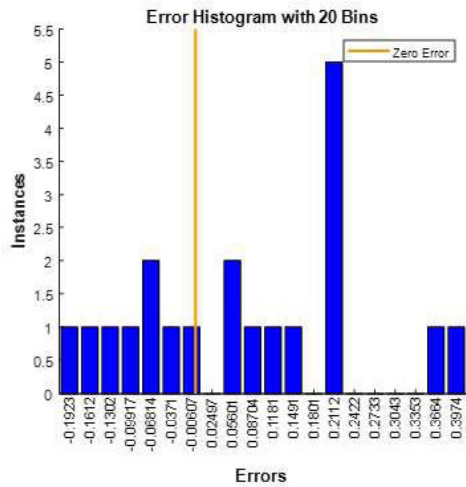


Figure 6.10 Error Histogram for simulated models on future dataset_Case 1

Table 6.2 Summary of average error values for best 20 validated regression neural models on total Static data

PRESENT_STATIC DATA						
ANN Models (static data only)	Error (MSE)			Error (RMSE)		
	Train	Validation	Test	Train	Validation	Test
1	0.01257	0.00957	0.01197	0.11210	0.09784	0.10939
2	0.01460	0.00996	0.01216	0.12085	0.09982	0.11025
3	0.01482	0.01099	0.01275	0.12172	0.10486	0.11290
4	0.01503	0.01203	0.01310	0.12258	0.10968	0.11448
5	0.01516	0.01211	0.01400	0.12314	0.11004	0.11831
6	0.01531	0.01211	0.01442	0.12375	0.11005	0.12007
7	0.01544	0.01233	0.01448	0.12428	0.11105	0.12032
8	0.01567	0.01244	0.01457	0.12520	0.11153	0.12071
9	0.01573	0.01253	0.01471	0.12540	0.11193	0.12130
10	0.01575	0.01255	0.01473	0.12550	0.11201	0.12135
11	0.01577	0.01264	0.01492	0.12557	0.11243	0.12214
12	0.01588	0.01271	0.01493	0.12601	0.11273	0.12220
13	0.01589	0.01297	0.01495	0.12605	0.11386	0.12226
14	0.01606	0.01297	0.01510	0.12673	0.11391	0.12288
15	0.01606	0.01311	0.01513	0.12673	0.11450	0.12301
16	0.01616	0.01314	0.01521	0.12712	0.11461	0.12333
17	0.01636	0.01318	0.01522	0.12790	0.11482	0.12339
18	0.01636	0.01319	0.01538	0.12791	0.11486	0.12400
19	0.01651	0.01323	0.01540	0.12851	0.11503	0.12408
20	0.01657	0.01329	0.01542	0.12874	0.11527	0.12416
Average	0.01559	0.01235	0.01443	0.12479	0.11104	0.12003

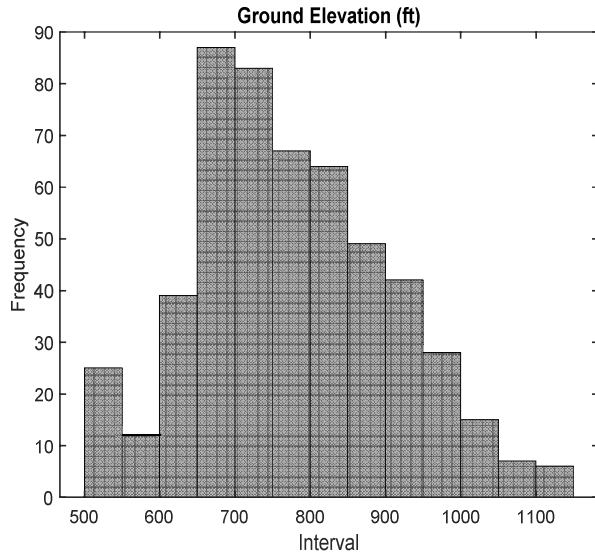
There were no major problems during the training as seen from the performance plot. The maximum fail was set at 20 hence the training stops if there is no improvement in the validation after 20 more iterations at a given epoch as seen in the training state plot. In figure 6.8 above the model weights were updated in each iteration where the least error was estimated at the 49th iteration for a total of 69 epochs. The R-squared value for the training, validation, and test data for the best model out of the 20 were approximately 0.73, 0.88, 0.77 respectively. The network models were simulated on the 20 data set that was set aside as future unseen data and the average RMSE value was as well evaluated. The results are recorded in table 6.3 below;

Table 6.3 Summary of future data prediction error in base Case_1

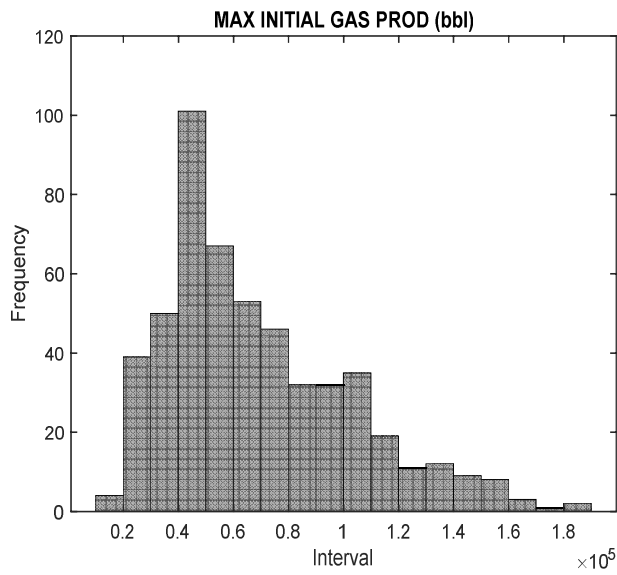
Future static_simulated	MSE	RMSE
1	0.03611	0.19004
2	0.03138	0.17716
3	0.02921	0.17090
4	0.04808	0.21928
5	0.03247	0.18020
6	0.02756	0.16601
7	0.03337	0.18269
8	0.03562	0.18874
9	0.03079	0.17547
10	0.02778	0.16666
11	0.04199	0.20492
12	0.03876	0.19687
13	0.03541	0.18817
14	0.03584	0.18932
15	0.03440	0.18547
16	0.03692	0.19214
17	0.03181	0.17836
18	0.03718	0.19282
19	0.03307	0.18186
20	0.03498	0.18703
Average	0.03464	0.18570

6.2 Base Case 2 (Present data Static & IP gas)

The base Case_2 includes total present static data and initial peak gas production. It comprised of 504-by-16 total data including the target value. Before generating the shallow neural network models, the same procedure was followed by carrying out the variable importance analysis (VIA) and visualizing data patterns by selecting features under the set condition mention in section 6.1 above. From the Pearson's correlation plot in base_Case study 2, the initial peak gas (maxIp) term had the highest positive linear relationship (coefficient value of 0.84) and the Ground Elevation variable had the highest negative(-0.46) linear relationship with respect to the target value for the static&IPgas case. No variable term had a perfect positive and negative linear relation with other pairs indicating that no redundancy exists in the input data selection. The distribution of the highest negative and positive linear variable data are represented in the histogram plots below in Figure 6.11(a & b respectively) which corresponds to plots in their variables terms as seen in the diagonal of the Pearson's correlation matrix plot in Figure 6.12. They both exhibit a positive skewness. The minimum initial peak gas production was 13,649 Mscf/m and the maximum production of gas was 189,144 Mscf/m. The ground elevation depth (the section of any point in well of the distance between the reference elevation and that point) on the other hand, ranges from a minimum of 510ft to a maximum of 1143ft for the total 504 observatory wells. In the base Case_2, all variables lie within the conditional range of selection, hence all the 15 input features were used to build the regression artificial model.



(a)



(b)

Figure 6.11 Histograms of data distribution of the (a) initial peak production gas and (b) Ground Elevation interval data.

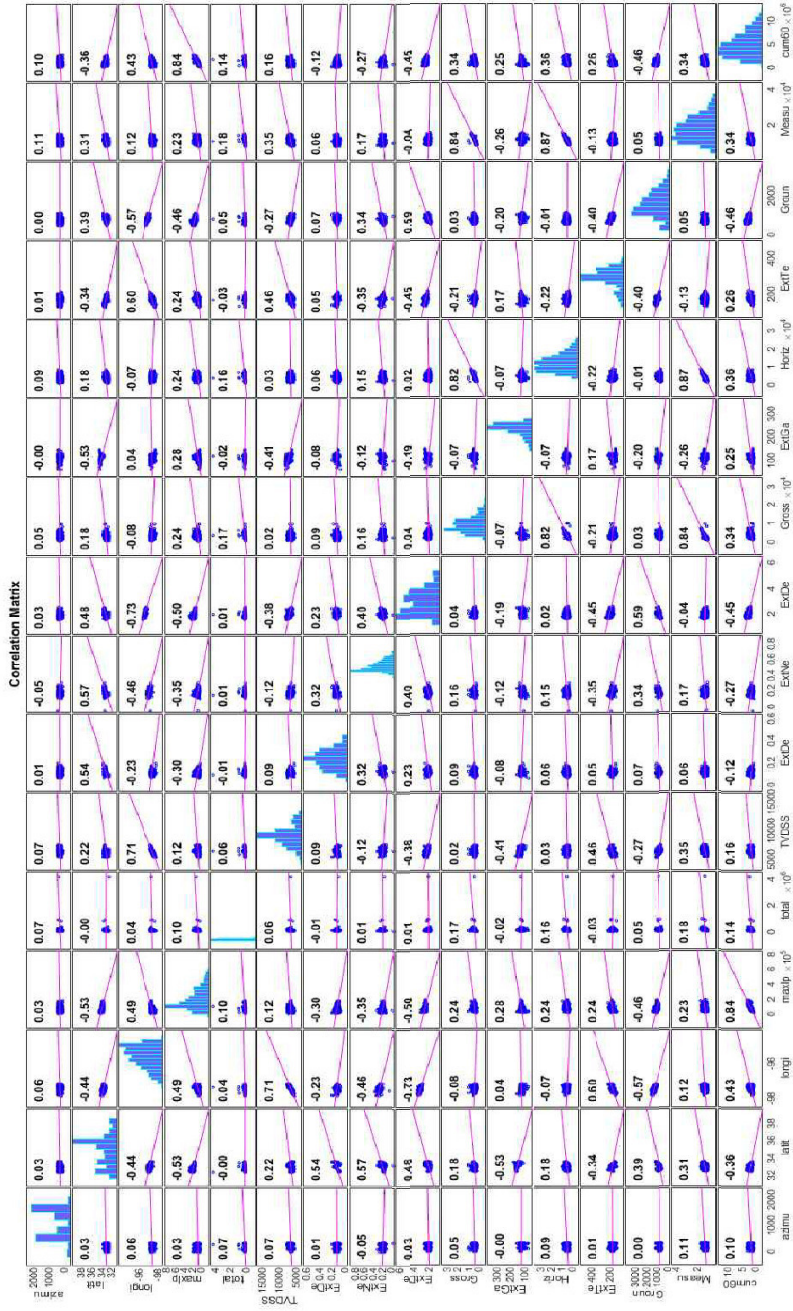


Figure 6.12 Pearson's correlation plot for base_Case 2

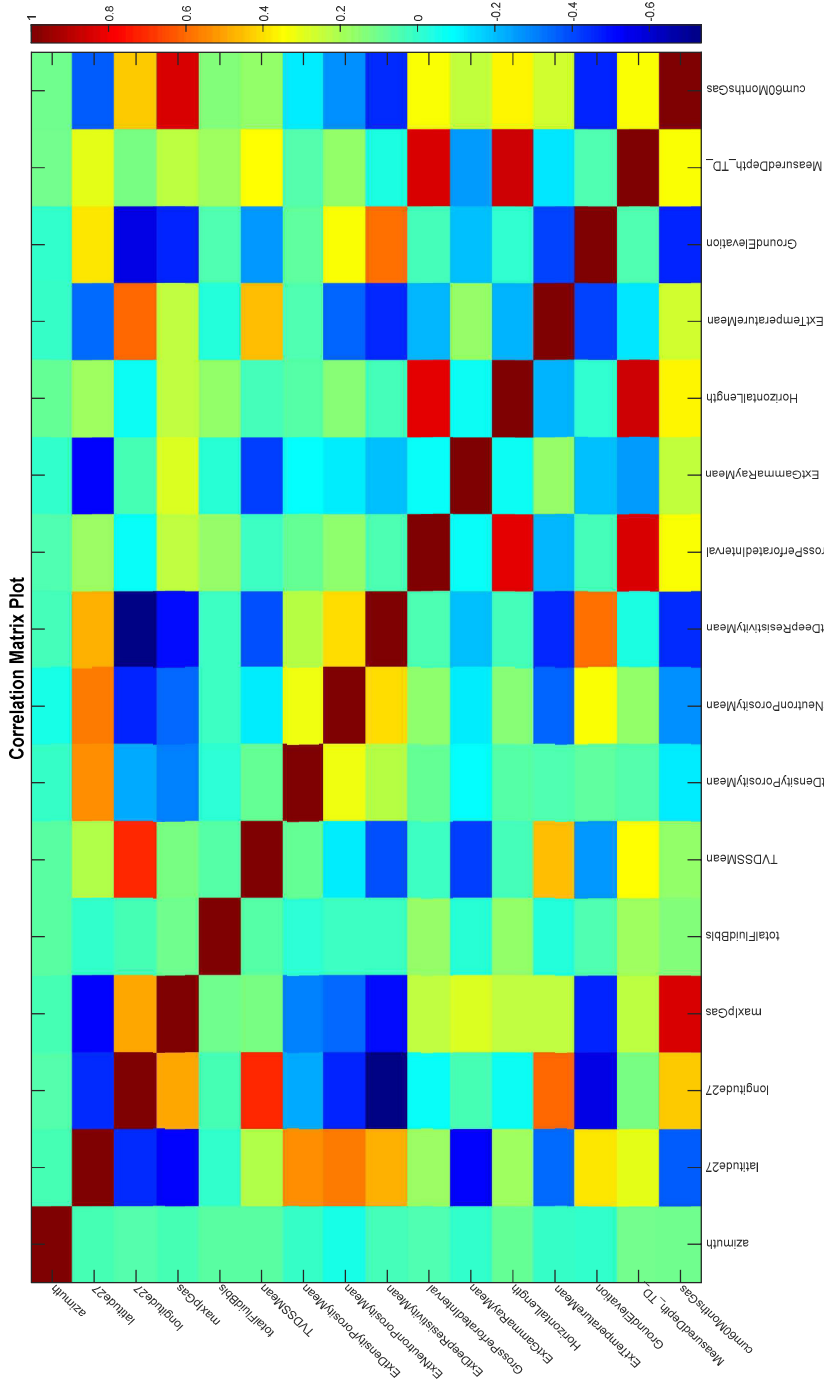


Figure 6.13 Color Map plot for Case_2 Static data & Ipas

6.2.1 ANN Model (Base Case 2)

Similarly in the base_Case 2, shallow artificial network regression models were generated on the total present data (504 observations) with 15 input features and 1 target value given in table 6.4. 1000 neural network regression models were generated and the best 20 models based on the minimum validation error were selected and organized in the result table. The best model's results out of the 20 ANN models are depicted.

Table 6.4 Input parameters for Static+IPgas, base Case_2

Output data	Cumulative 60 months gas
Input Data	<ol style="list-style-type: none"> 1. azimuth 2. latitude 3. longitude 4. total fluids (bbl.) 5. true vertical depth_TVD (ft.) 6. Density log 7. Neutron log 8. Deep Resistivity log 9. Gross Perforated Interval (ft) 10. Gamma Ray log 11. Horizontal Length (ft.) 12. Temperature 13. Ground Elevation (ft) 14. Measured Depth 15. Initial. peak gas production (IPgas)

The training data, validation data and test data for base Case_2 were 404, 50 and 50 respectively which represent 80%, 10%, and 10% each of the total 504 observations. The regression plots, error performance plot, and training state plot for the best model(model 581) out of the 20 selected models based on the sorted validation errors are represented from the Figures 6.15 to 6.20 below. The present and future data MSE(s) and RMSE(s) from the models and the averages are summarized in Tables 6.5 and 6.6 respectively.

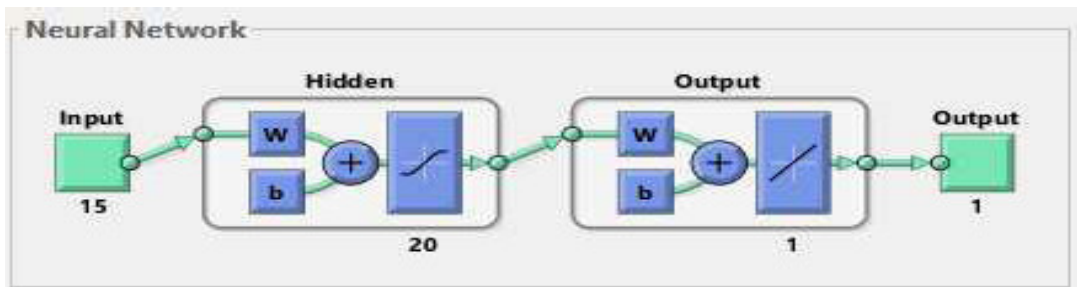


Figure 6.14 Neural network architecture for the generated models in base_Case2

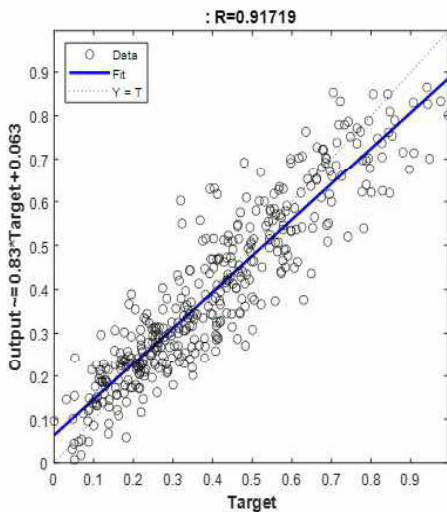


Figure 6.15 Regression fit on training data_best validated model Case 2

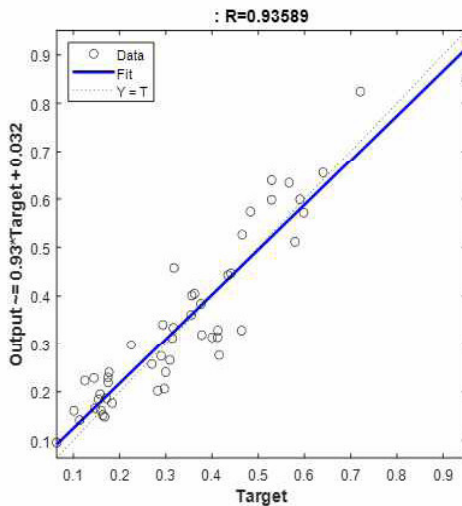


Figure 6.16 Regression fit on validation data_best validated model Case 2

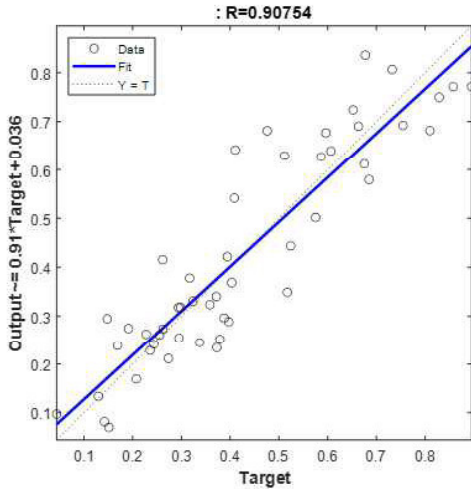


Figure 6.17 Regression fit on test data_best validated model Case 2

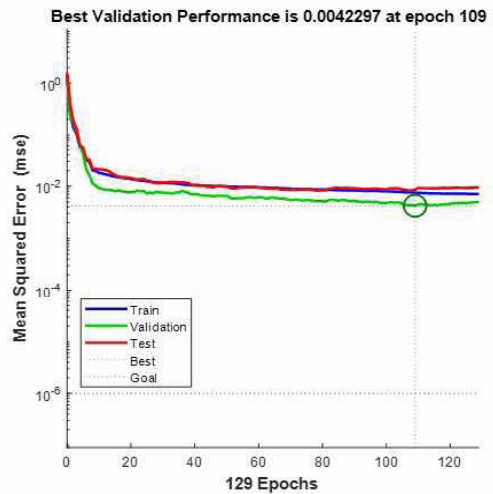


Figure 6.18 Performance plot on best validated model_ Case 2

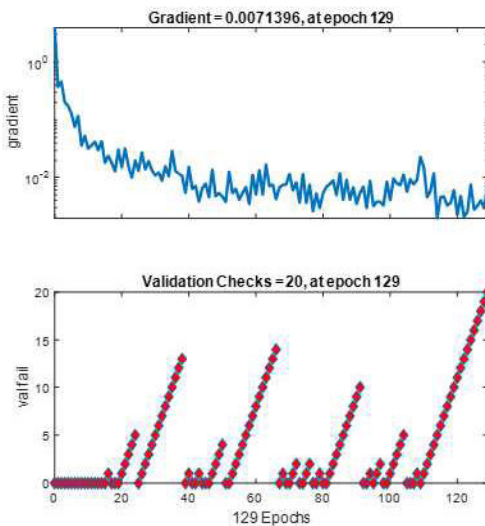


Figure 6.19 Training state on best validated model_best validated model Case 2

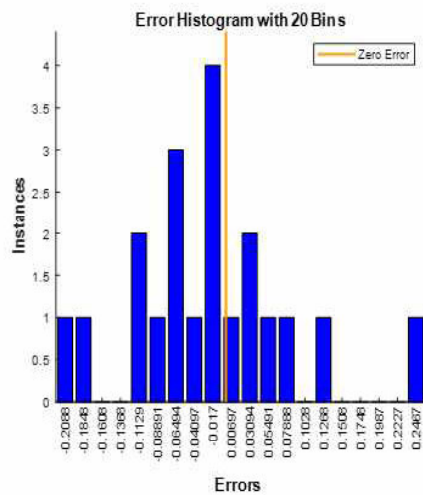


Figure 6.20 Error Histogram for simulated models on future dataset_Case 2

Table 6.5 Summary of average error values for best 20 validated regression neural models on Static&IPgas gas data

Models (static+IPgas data)	PRESENT_STATIC & IPGAS DATA					
	Error (MSE)			Error (RMSE)		
	Train	Validation	Test	Train	Validation	Test
1	0.00513	0.00423	0.00524	0.07161	0.06504	0.07237
2	0.00520	0.00458	0.00597	0.07210	0.06767	0.07724
3	0.00558	0.00469	0.00597	0.07471	0.06846	0.07726
4	0.00561	0.00469	0.00601	0.07487	0.06850	0.07752
5	0.00562	0.00486	0.00604	0.07497	0.06972	0.07775
6	0.00571	0.00505	0.00605	0.07555	0.07108	0.07780
7	0.00573	0.00509	0.00607	0.07567	0.07132	0.07791
8	0.00579	0.00511	0.00626	0.07607	0.07147	0.07911
9	0.00581	0.00515	0.00628	0.07622	0.07180	0.07923
10	0.00593	0.00522	0.00629	0.07701	0.07222	0.07931
11	0.00595	0.00527	0.00634	0.07713	0.07257	0.07961
12	0.00596	0.00527	0.00635	0.07719	0.07261	0.07971
13	0.00597	0.00533	0.00636	0.07727	0.07298	0.07975
14	0.00598	0.00539	0.00643	0.07731	0.07343	0.08020
15	0.00599	0.00542	0.00645	0.07737	0.07364	0.08029
16	0.00603	0.00545	0.00646	0.07763	0.07384	0.08039
17	0.00611	0.00548	0.00652	0.07817	0.07402	0.08074
18	0.00621	0.00550	0.00660	0.07880	0.07417	0.08121
19	0.00623	0.00558	0.00663	0.07891	0.07472	0.08140
20	0.00623	0.00562	0.00665	0.07891	0.07497	0.08152
Average	0.00584	0.00515	0.00625	0.07637	0.07171	0.07902

The performance of the best validated model also signifies no problems during the training and iterations continued until the 129th epoch where the validation error no more improved and the updates in the weights and bias term are halted. The minimum error in this model was therefore estimated at the 109th iteration. The corresponding R-squared value for the training, validation and test data of the best model out of the 20 were approximately 0.92, 0.91, 0.94 respectively and the average RMSE values evaluated. The network models was simulated on the 20 future unseen static&Ipgas data and the average RMSE values evaluated. The results is seen table 6.6 below

Table 6.6 Summary of future data prediction error in base Case_2

Future static&Ipgas simulated data	MSE	RMSE
1	0.01420	0.11915
2	0.00838	0.09152
3	0.00947	0.09731
4	0.01187	0.10895
5	0.00604	0.07774
6	0.01408	0.11865
7	0.00727	0.08528
8	0.01800	0.13417
9	0.01054	0.10266
10	0.00744	0.08627
11	0.00911	0.09546
12	0.01159	0.10763
13	0.01020	0.10099
14	0.01345	0.11596
15	0.01058	0.10287
16	0.00715	0.08456
17	0.02162	0.14703
18	0.01007	0.10034
19	0.00885	0.09406
20	0.01178	0.10854
Average	0.01108	0.10396

6.3 Base Case 3 (Present data Static, IP gas & Cum12 gas)

The base Case_3 comprised of total present static data and initial peak gas production and 12 months history cumulative gas production. The total data set was 514-by-18 matrix data including the target value. Variable importance analysis (VIA) to select features for models were investigated with the Pearson's and Color map plots under the same set condition; elimination of all variables with coefficient values between -0.09 and 0.09 . From the Pearson's correlation plot in base_Case 3 Figure 6.21 as expected the second variable term (cum12 gas) had the highest positive linear relationship (0.93) and the Ground Elevation variable had the highest negative linear relationship(-0.46) with respect to the target value. The cumulative 12 months water production(third term) variable in Figure 6.21 had a relatively low positive linear relationship with the target value. No variable term had a perfect positive and negative linear relation with other pairs indicating that no redundancy exists in the input data selection. Similarly in this base Case_3, all variables lie within the conditional range of selection, hence all the 17 input features were used to build the regression artificial model.

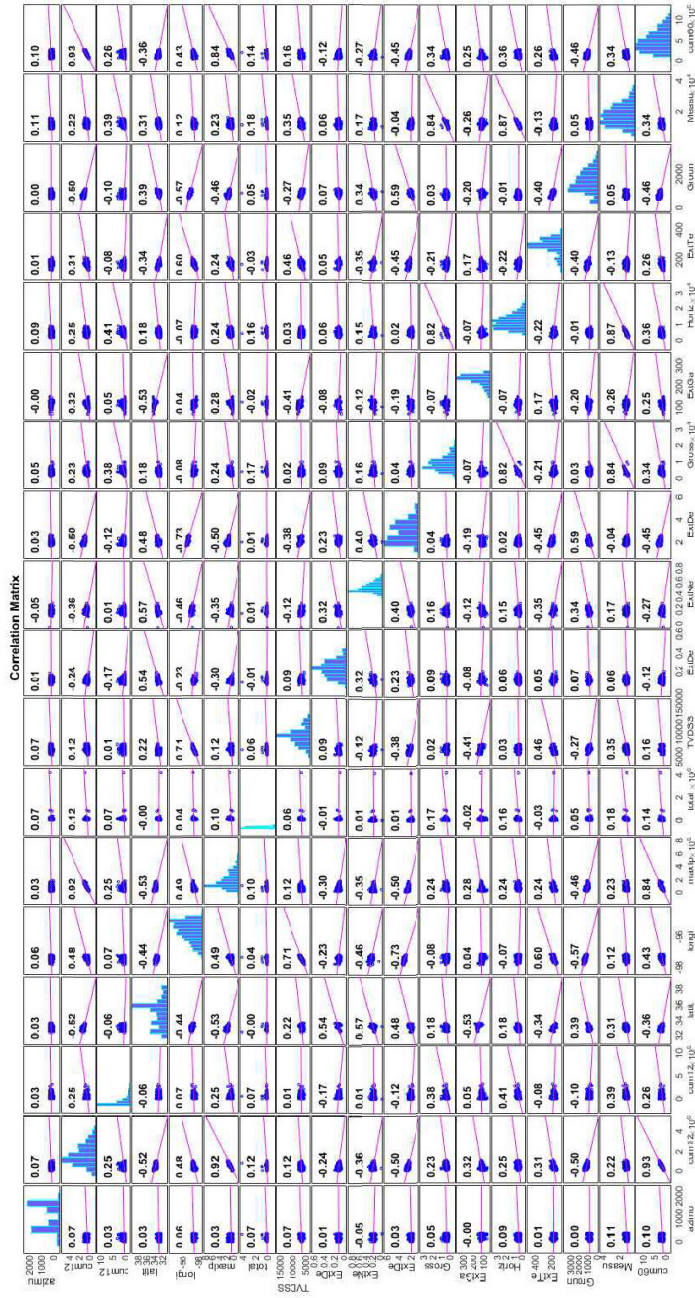


Figure 6.21 Pearson's Correlation plot for base_Case 3

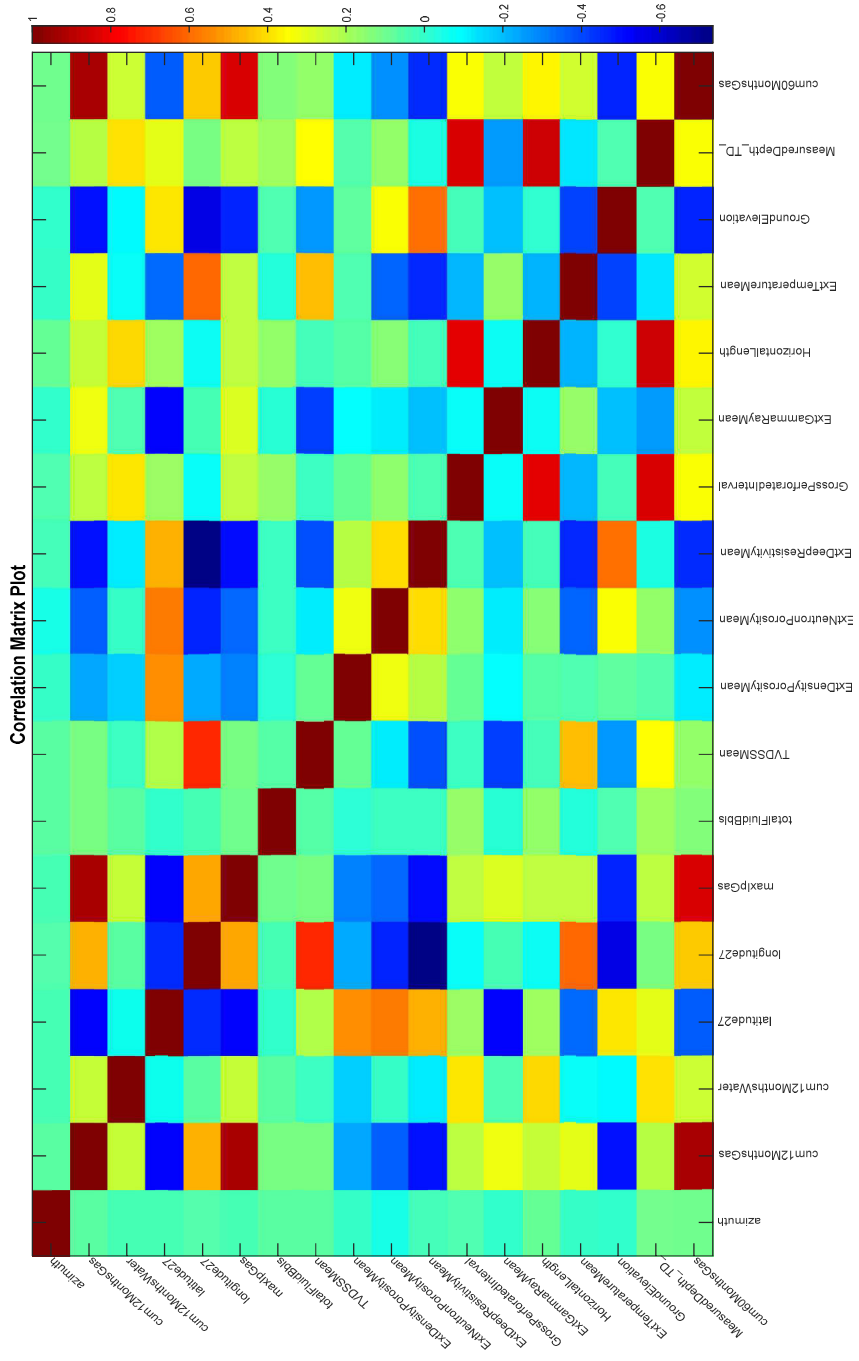


Figure 6.22 Color Map plot for base_Case 3

6.3.1 Ann Model (Base Case 3)

Again 1000 shallow artificial neural network regression models were generated on the total present data (504) with 17 total input features and 1 target value under this case study. Table 6.7 summarizes the input features.

Table 6.7 Input parameters for Static+IPgas+Cum12gas, base Case_3

Output data	Cumulative 60 months gas
Input Data	<ol style="list-style-type: none"> 1. azimuth 2. latitude 3. longitude 4. total fluids (bbl.) 5. true vertical depth_TVD (ft.) 6. Density log 7. Neutron log 8. Deep Resistivity log 9. Gross Perforated Interval (ft) 10. Gamma Ray log 11. Horizontal Length (ft.) 12. Temperature 13. Ground Elevation (ft) 14. Measured Depth 15. Initial. peak gas production (IPgas) 16. cumulative 12 water production 17. cumulative 12 gas production (Cum12gas)

The training data, validation data and test data for base Case_3 were 404, 50 and 50 respectively which represent 80%, 10% and 10% each of the total 504 observations. The regression plots, error performance plot, and training state plot for the best model(model 989) out of the 20 selected models based on the sorted validation errors are represented from the Figures 6.24 to 6.29 below. The present and future data MSE(s) and RMSE(s) from the models and the averages are summarized in the tables 6.8 and 6.9 respectively.

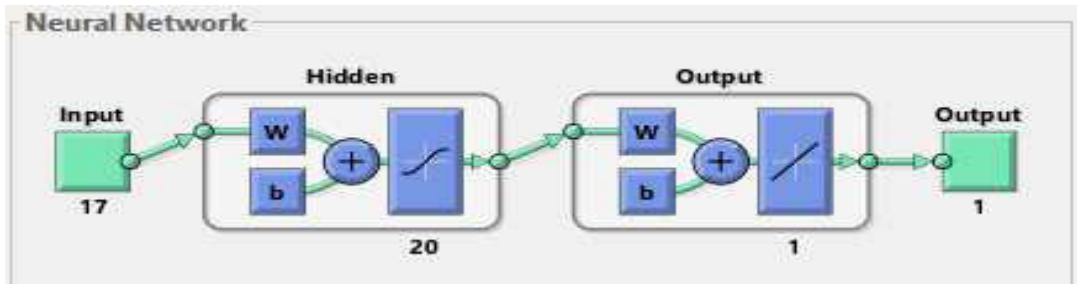


Figure 6.23 View of neural network architecture for the generated models in base_Case 3

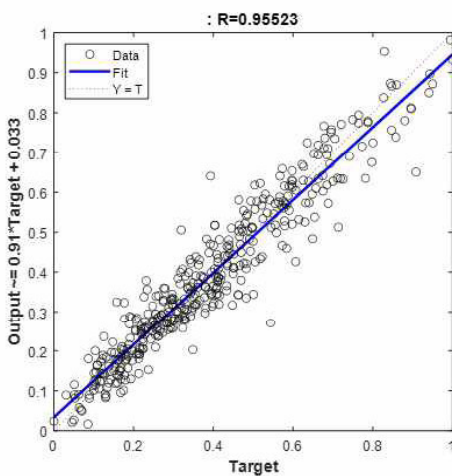


Figure 6.24 Regression fit on training data_best validated model Case 3

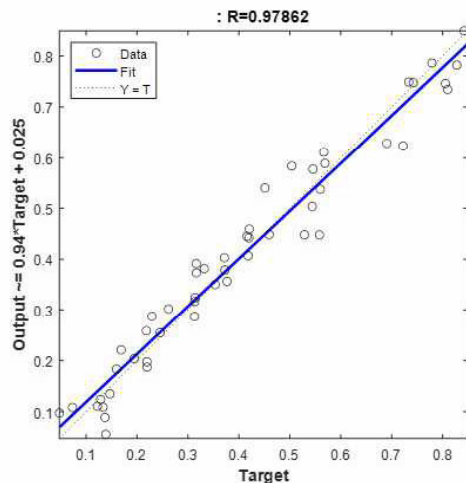


Figure 6.25 Regression fit on validation data_best validated model Case 3

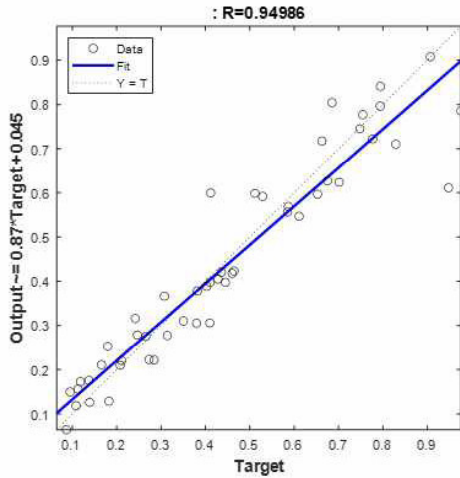


Figure 6.26 Regression fit on test data_best validated model Case 3

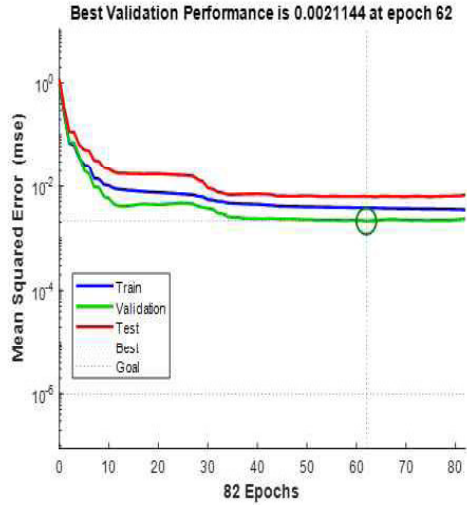


Figure 6.27 Performance plot on best validated model_Case 3

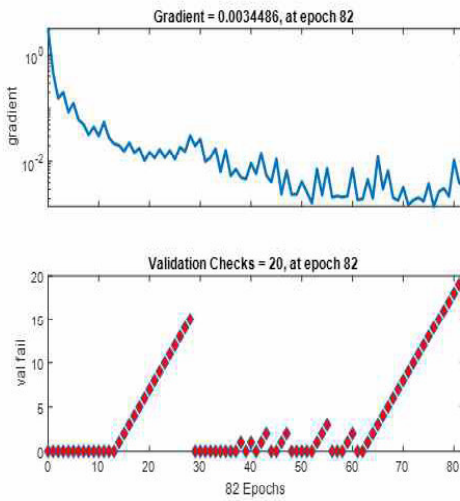


Figure 6.28 Training state on best validated model_Case 3

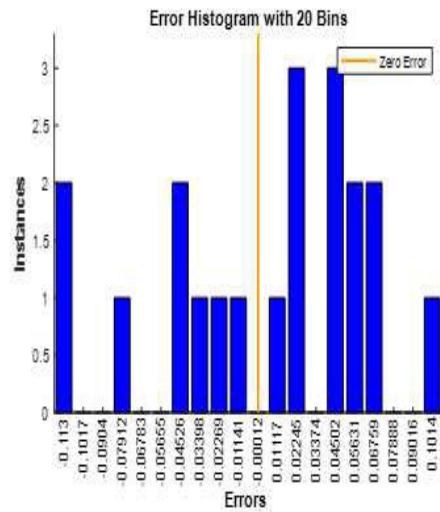


Figure 6.29 Error Histogram for simulated models on future dataset_Case 3

Table 6.8 Summary of average error values for best 20 validated regression neural models_ Static, IP&Cum12 gas

ANN MODEL		Error (MSE)			Error (RMSE)		
PRESENT_STATIC & IPGAS& CUM12 DATA		Train	Validation	Test	Train	Validation	Test
(static+IPgas+cum12gas data)							
1		0.00238	0.00211	0.00211	0.04882	0.04598	0.04596
2		0.00242	0.00212	0.00213	0.04919	0.04602	0.04619
3		0.00258	0.00216	0.00232	0.05082	0.04651	0.04819
4		0.00271	0.00223	0.00238	0.05203	0.04725	0.04876
5		0.00276	0.00227	0.00245	0.05251	0.04769	0.04946
6		0.00280	0.00233	0.00247	0.05288	0.04825	0.04969
7		0.00280	0.00233	0.00260	0.05292	0.04830	0.05100
8		0.00281	0.00241	0.00261	0.05298	0.04904	0.05104
9		0.00281	0.00242	0.00262	0.05304	0.04921	0.05117
10		0.00283	0.00243	0.00268	0.05320	0.04925	0.05178
11		0.00284	0.00248	0.00272	0.05332	0.04979	0.05220
12		0.00285	0.00248	0.00278	0.05341	0.04981	0.05271
13		0.00286	0.00250	0.00284	0.05352	0.04996	0.05333
14		0.00287	0.00252	0.00285	0.05353	0.05024	0.05335
15		0.00287	0.00255	0.00285	0.05356	0.05048	0.05341
16		0.00288	0.00257	0.00286	0.05365	0.05065	0.05351
17		0.00288	0.00257	0.00287	0.05367	0.05066	0.05358
18		0.00288	0.00259	0.00290	0.05369	0.05086	0.05382
19		0.00289	0.00259	0.00292	0.05373	0.05086	0.05407
20		0.00289	0.00260	0.00294	0.05373	0.05096	0.05419
Average		0.00278	0.00241	0.00265	0.05271	0.04909	0.05137

Based on the performance plot the minimal error during the training of the best validated model represented in this case was estimated at the 62nd iteration and the training stopped after 82 epochs due to increasing validation error. The corresponding R-squared value for the training, validation and test data of the best model out of the 20 were approximately 0.96, 0.98, 0.95 respectively. Subsequently, the selected models were simulated on the 20 future unseen static&Ipgas&cum12gas data case and the average rmse values were obtained as shown below.

Table 6.9 Summary of future data prediction error in base Case_3

Future Cumgas case_simulated	MSE	RMSE
1	0.00419	0.06471
2	0.00437	0.06608
3	0.00366	0.06051
4	0.00392	0.06264
5	0.00357	0.05972
6	0.00374	0.06119
7	0.00360	0.06004
8	0.00339	0.05820
9	0.00428	0.06545
10	0.00340	0.05832
11	0.00391	0.06255
12	0.00356	0.05964
13	0.00432	0.06572
14	0.00417	0.06456
15	0.00383	0.06187
16	0.00345	0.05874
17	0.00330	0.05744
18	0.00504	0.07096
19	0.00471	0.06864
20	0.00358	0.05982
Average	0.00390	0.06234

6.4 Clustered Models

This section describes results for the clustered models generated. As previously mentioned in the workflow and methodology, the K-means algorithm was adopted to divide the data into clustered groups using the principal components of the normalized data that reduce the multiple variable input features to a lower dimensionality in space. Also, silhouette plots were visualized to assess each cluster grouping's cohesiveness. The data were initially clustered from 2-4 groups in each case study on the present data. Future data representing the red data points in the clustered design plots in Figures 6.30, 6.33, and 6.36 were introduced into the clustered group system and allocated to their class labels employing the K-nearest neighbor classifier. A comparison is done between the total data predictive models' errors and the clustered data predictive models' errors. As one of the main objectives was also to evaluate the optimal number of clustering for the data used in this research, further investigative study for 5 group clustering cases was conducted on the present data and the future data predictive models. The average errors for the clustered group models were also used to specify and prove a selection of optimal cluster number grouping was reasonable. Again 1000 neural network regression models were developed under each case study for 2-5 cluster grouping cases and on each cluster group label. Similarly, the best 20 validated models were selected.

6.4.1 PCA & K-Means Clustering Results

The principal component matrices of the static data case, static & IPgas data case, and static, IPgas & cum12gas data case after the normalization of the data were obtained for the numeric present and explained variances within the data was determined. The principal components are returned from column variables from the data set. For instance, in the static Case (504-by-15), the principal components returned are 15-by 15 matrix coefficients and each column contains coefficients of one principal components in descending order. The

scores of the principal components which are the representation of the matrix data in the principal component space were used in the K-means clustering algorithms. The clustering is visualized in 3-D plots and the 20 future data points were introduced to determine each data assigned cluster label. Below are figures representing the 2 group clustering, 3 group clustering, and 4 group clustering on static data, static&Ipgas data, and static&Ipgas&Cum12gas cases respectively.

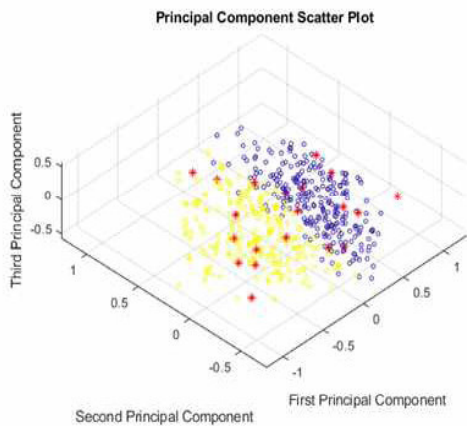


Figure 6.30 Two group clustering on static data only

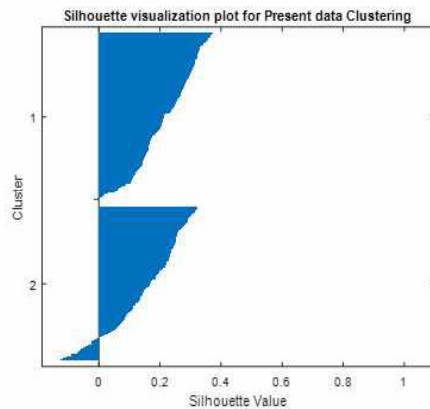


Figure 6.31 Silhouette plot on two group clusters

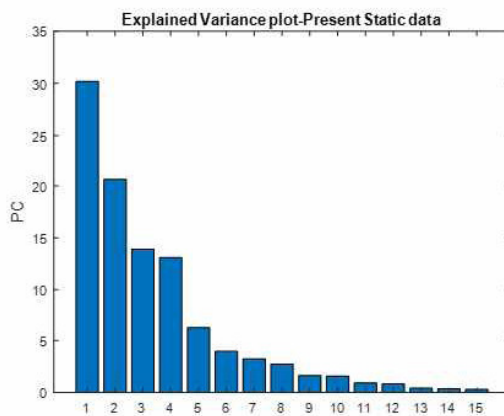


Figure 6.32 Variance explained by the principal components of static data only on two group clusters

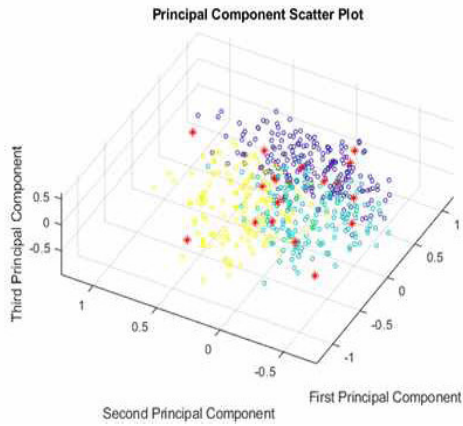


Figure 6.33 Three group clustering on static&Ipgas data

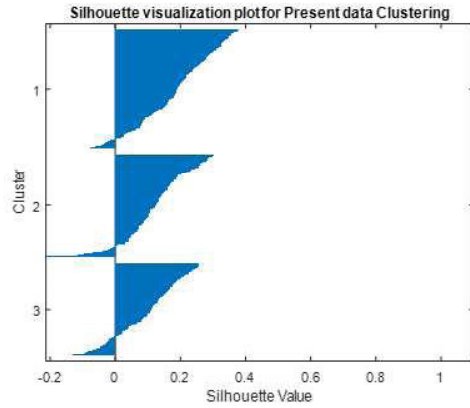


Figure 6.34 Silhouette plot on three group clusters on Static&Ipgas data

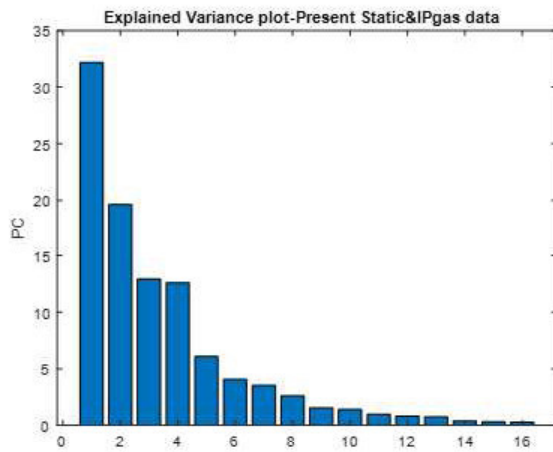


Figure 6.35 Variance explained by the principal components of static&Ipgas data on three group clusters

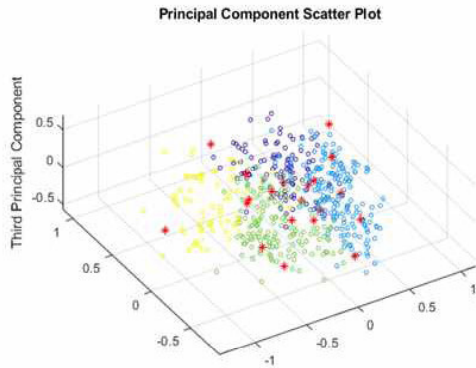


Figure 6.36 Four group clustering on static&Ipgas&Cum12gas data

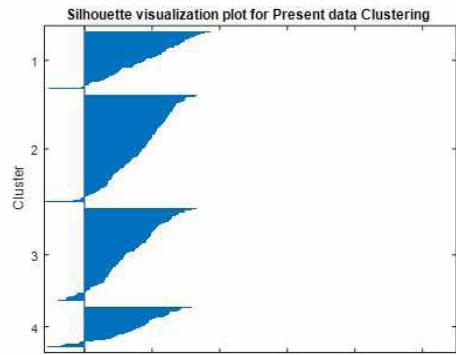


Figure 6.37 Silhouette plot on 4 group clusters static&Ipgas&Cum12gas data

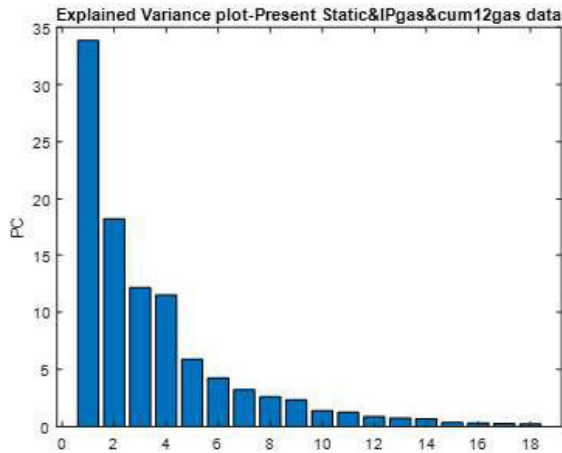


Figure 6.38 Variance explained by the principal components of static&Ipgas&Cum12gas data on four group clusters

6.5 ANN Model Results for Clustered Data Groups

To estimate the predictive performance and estimate the likely possibility in improvement on the similar data groups, same principle is followed in generating models and selecting the best 20 validated models for (static), (static&IPgas), (static&IPgas&cum12gas) data in each cluster group collection for all 5 clustering cases respectively. Once again, variable importance analysis was conducted on each cluster group to select variables of greater influence to predicting the 60 months cumulative gas by removing all irrelevant features in each cluster group. Similarly, the condition was to eliminate all variables coefficients within the conditional range of -0.09 to 0.09 from the Pearson's Correlation plot. It was discovered that each cluster set in all clustering cases have variables unique to the target value in terms of the sensitivity of the variable, hence the input variable features used for each cluster label for all the clustering numbers differed. Again 80%, 10%, 10% for each cluster group data is used for training, validation, and testing. The average values of the best 20 validated models are calculated and their percentages of improvements are estimated. The results of the clustering on the present data and the number of future data objects that were classified on each cluster label groups are summarized in the tables 6.10 to 6.13 below.

Table 6.10 Regression Input data for 2 grouping clusters

2 GROUPING CLUSTERS						
Static Data Case						
	Input data	# Train_data	# Validation_data	# Test_data	# Future data	
Cluster 1	263x10	211	26	26	10	
Cluster 2	241x13	193	24	24	10	
Static&Ipgas Data Case						
	Input data	# Train_data	# Validation_data	# Test_data	# Future data	
Cluster 1	202x9	202	25	25	11	
Cluster 2	252x13	202	25	25	9	
Static&Ipgas&cum12gas Data Case						
	Input data	# Train_data	# Validation_data	# Test_data	# Future data	
Cluster 1	211x10	169	21	21	7	
Cluster 2	293x14	235	29	29	13	

Table 6.11 Regression Input data for 3 grouping clusters

3 GROUPING CLUSTERS						
Static Data Case						
	Input data	# Train_data	# Validation_data	# Test_data	# Future data	
Cluster 1	201x9	161	20	20	10	
Cluster 2	158x11	126	16	16	6	
Cluster 3	145x13	115	15	15	4	
Static&Ipgas Data Case						
	Input data	# Train_data	# Validation_data	# Test_data	# Future data	
Cluster 1	190x6	152	19	19	8	
Cluster 2	166x12	132	17	17	8	
Cluster 3	148x13	118	15	15	4	
Static&Ipgas&cum12gas Data Case						
	Input data	# Train_data	# Validation_data	# Test_data	# Future data	
Cluster 1	219x9	172	22	22	7	
Cluster 2	218x10	174	22	22	12	
Cluster 3	70x14	56	7	7	1	

Table 6.12 Regression Input data for 4 grouping clusters

4 GROUPING CLUSTERS						
Static Data Case						
	Input data	# Train_data	# Validation_data	# Test_data	# Future data	
Cluster 1	97x10	77	10	10	3	
Cluster 2	182x12	146	18	18	8	
Cluster 3	157x10	125	16	16	8	
Cluster 4	68x9	54	7	7	1	
Static&Ipgas Data Case						
	Input data	# Train_data	# Validation_data	# Test_data	# Future data	
Cluster 1	103x11	152	19	19	6	
Cluster 2	168x10	132	17	17	6	
Cluster 3	167x8	118	15	15	7	
Cluster 4	66x11	52	7	7	0	
Static&Ipgas&cum12gas Data Case						
	Input data	# Train_data	# Validation_data	# Test_data	# Future data	
Cluster 1	83x12	172	22	22	4	
Cluster 2	168x12	174	22	22	7	
Cluster 3	184x9	56	7	7	9	
Cluster 4	69x9	55	7	7	0	

Table 6.13 Regression Input data for 5 grouping clusters

5 GROUPING CLUSTERS						
Static Data Case						
	Input data	# Train_data	# Validation_data	# Test_data	# Future data	
Cluster 1	156x8	124	16	16	5	
Cluster 2	117x9	93	12	12	6	
Cluster 3	117x10	93	12	12	5	
Cluster 4	45x9	35	5	5	1	
Cluster 5	69x11	55	7	7	3	
Static&Ipgas Data Case						
	Input data	# Train_data	# Validation_data	# Test_data	# Future data	
Cluster 1	154x12	124	15	15	5	
Cluster 2	108x8	86	11	11	6	
Cluster 3	108x12	86	11	11	5	
Cluster 4	67x11	53	7	7	1	
Cluster 5	67x12	53	7	7	3	
Static&Ipgas&cum12gas Data Case						
	Input data	# Train_data	# Validation_data	# Test_data	# Future data	
Cluster 1	120x13	96	12	12	4	
Cluster 2	110x10	88	11	11	8	
Cluster 3	149x13	119	15	15	5	
Cluster 4	49x14	39	5	5	1	
Cluster 5	76x12	60	8	8	2	

6.5.1 Model Improvement of Cluster Groups

The percentages of improvement among clusters for the 2, 3, 4, and 5 grouping set were investigated on both the present data and unseen future data and the root mean square error per total cluster number was calculated for each cluster label under each study case. For simplicity, the regression fit plots and performances of the best-validated models among few clustering cases are represented below under the study cases. Plots of the percentage of improvement for the training, validation, and test data on the 5 clustering number cases, as well as on the improvement on the future simulated data are also illustrated below. The improvement is calculated and plotted from the difference of the sum of root mean square errors per data of all cluster labels from the total data and dividing by the total sum of errors for train, validation, and test data. The summaries of the improvement tables are represented in this section.

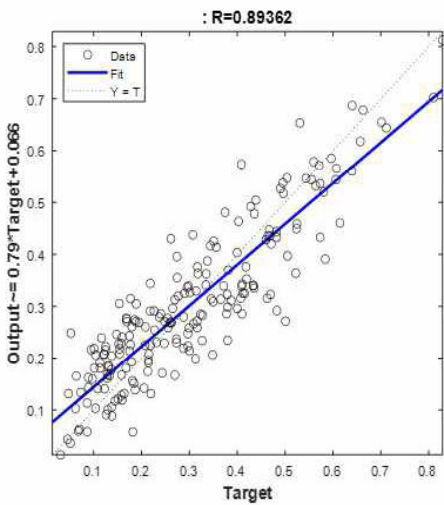


Figure 6.39 Regression fit on train data_cluster label 2 (Two group clustering_static&Ipgas data)

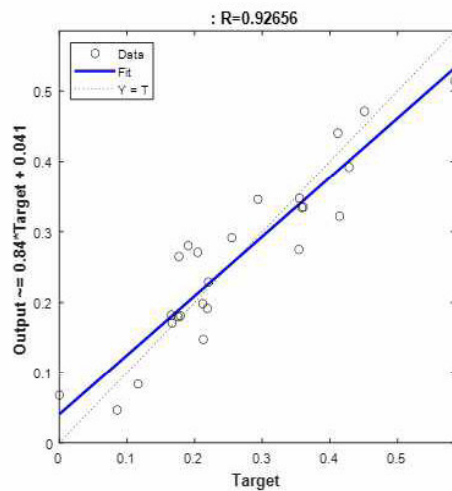


Figure 6.40 Regression fit on validation data_cluster label 2 (Two group clustering_static&Ipgas data)

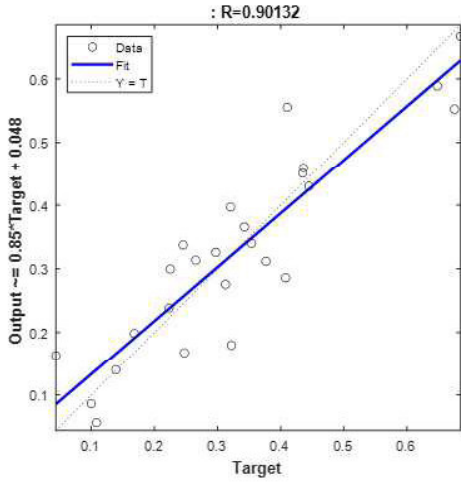


Figure 6.41 Regression fit on test data_cluster label 2 (Two group clustering_static&Ipgas data)

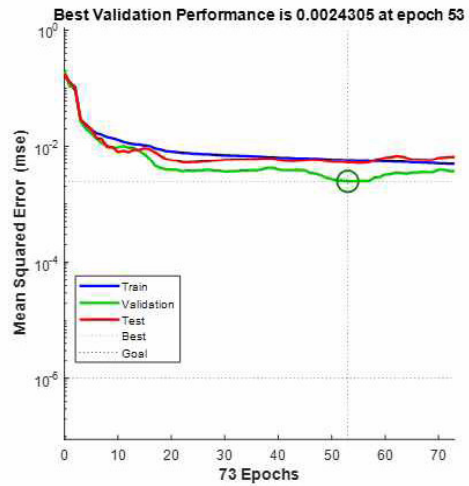


Figure 6.42 Performance plot of best validated model in _cluster label 2 (Two group clustering_static&Ipgas data)

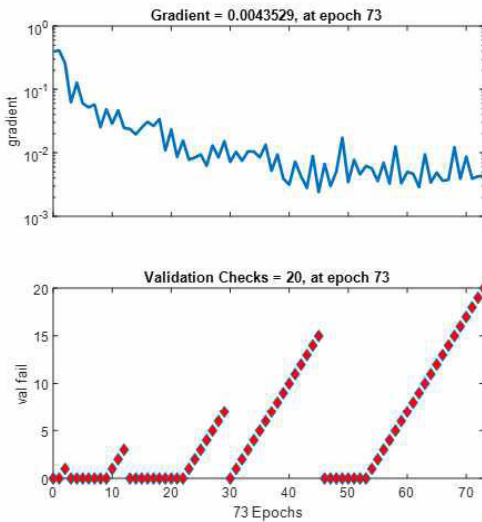


Figure 6.43 Training state of best validated model in_cluster label 2 (Two group clustering_static&Ipgas data)

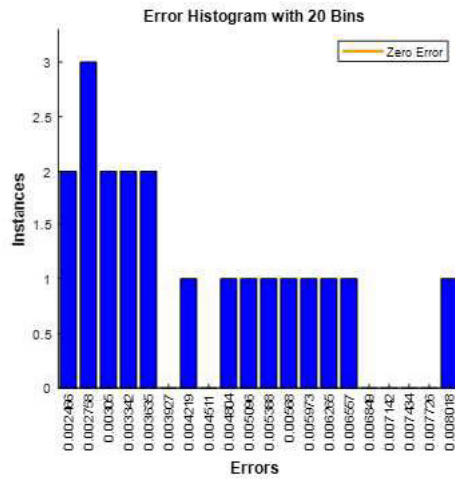


Figure 6.44 Error histogram for simulated models on future dataset (Two group clustering_static&Ipgas data)

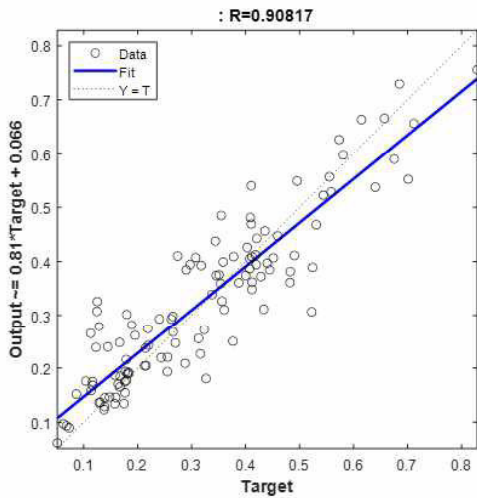


Figure 6.45 Regression fit on train data_cluster label 3 (Three group clustering_static data)

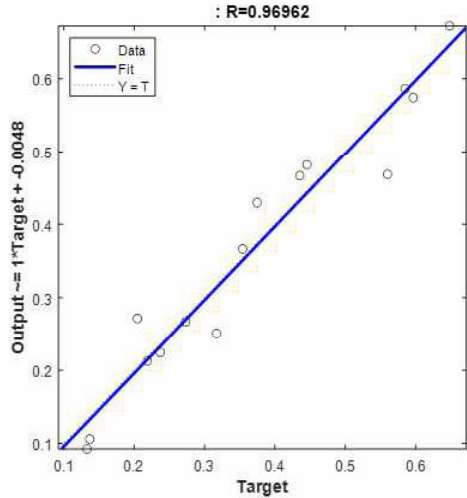


Figure 6.46 Regression fit on validation data_cluster label 3 (Three group clustering_static data)

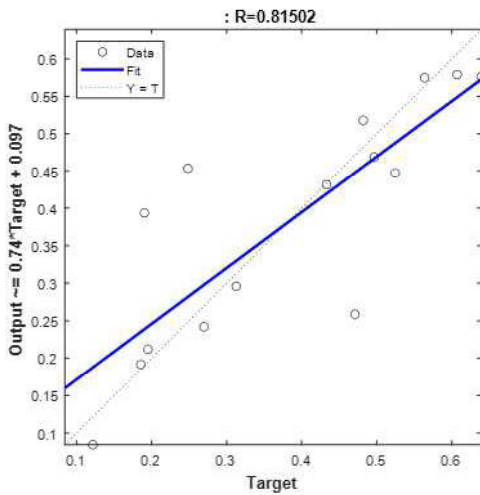


Figure 6.47 Regression fit on test data_cluster label 3 (Three group clustering_static data)

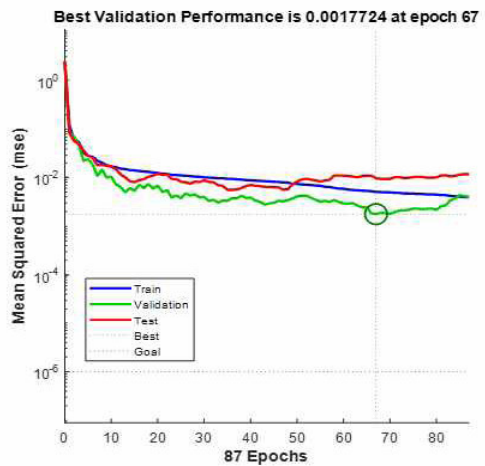


Figure 6.48 Performance plot of best validated model in_cluster label 3 (Three group clustering_static data)

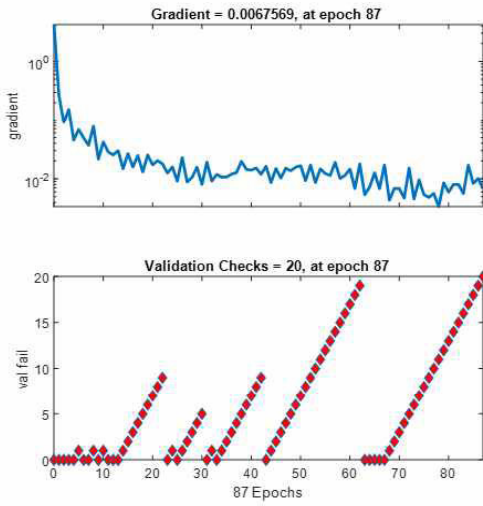


Figure 6.49 Training state of best validated model in_cluster label 3 (Three group clustering_static data)

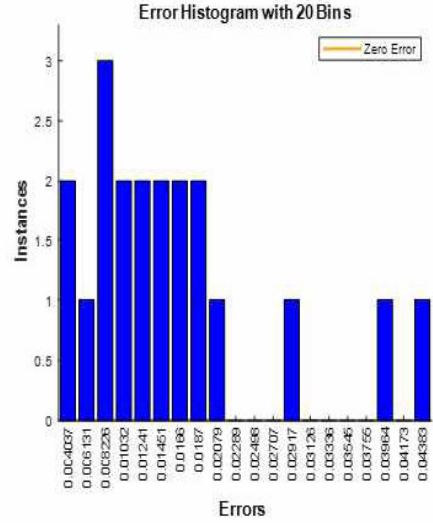


Figure 6.50 Error histogram for simulated models on future dataset (Three group clustering_static data)

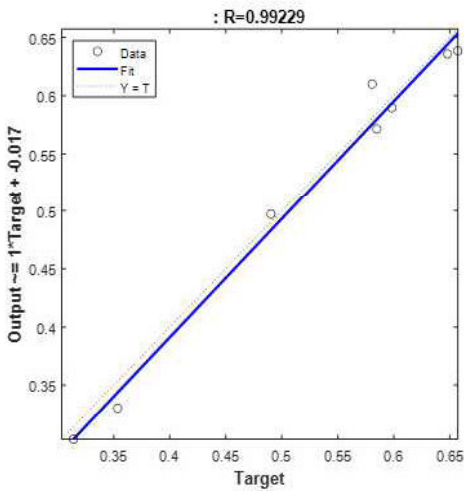


Figure 6.51 Regression fit on training data_cluster label 1 (Four group clustering_static, Ipgas&cum12gas data)

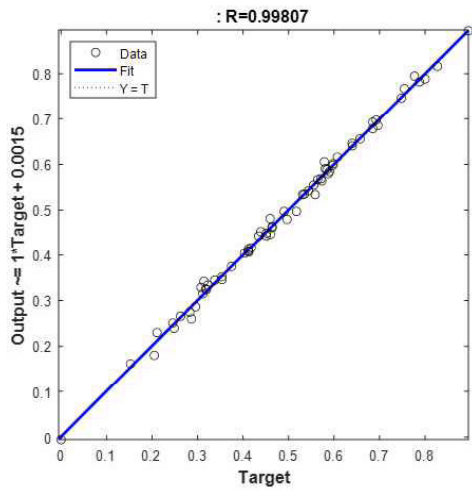


Figure 6.52 Regression fit on validation data_cluster label 1 (Four group clustering_static, Ipgas&cum12gas data)

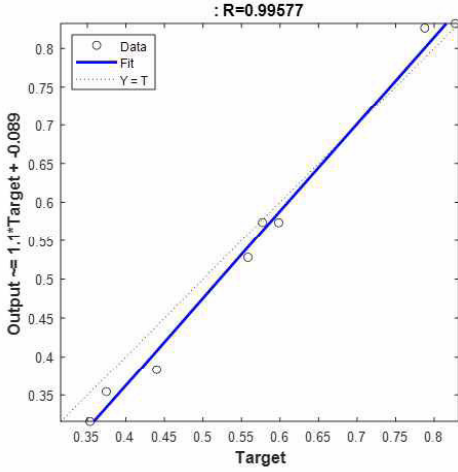


Figure 6.53 Regression fit on test data_cluster label 1 (Four group clustering_ static, Ipgas&cum12gas data)

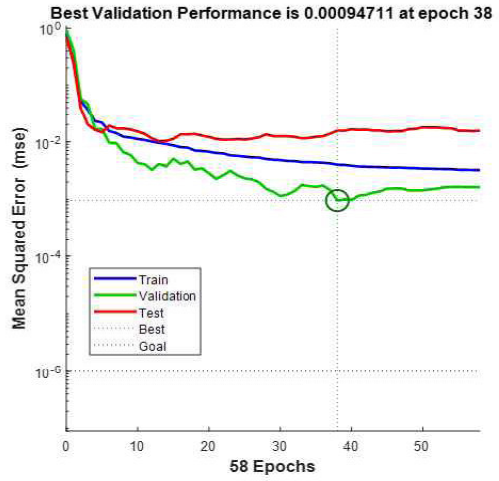


Figure 6.54 Performance plot of best validated model in_cluster label 1 (Four group clustering_ static, Ipgas&cum12gas data)

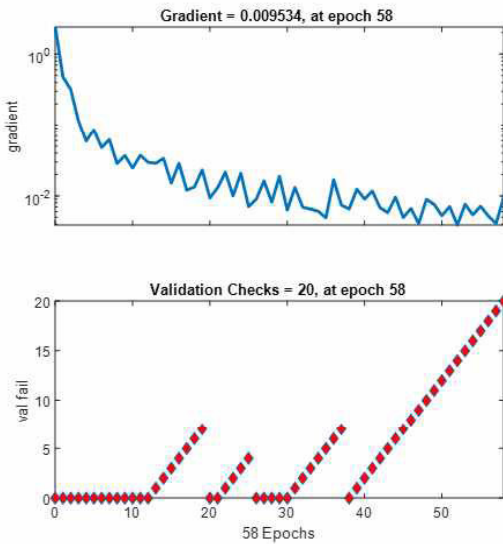


Figure 6.55 Training state of best validated model in_cluster label 1 (Four group clustering_ static, Ipgas&cum12gas data)

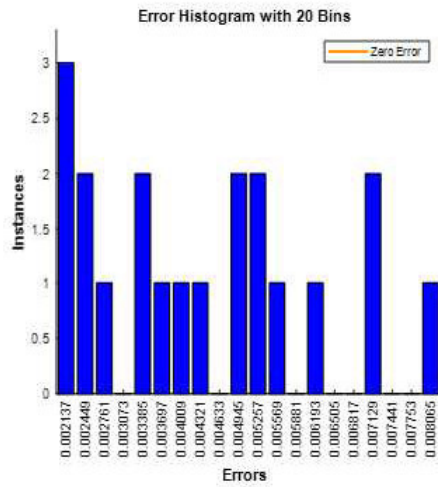


Figure 6.56 Error histogram for simulated models on future dataset (Four group clustering_ static, Ipgas&cum12gas data)

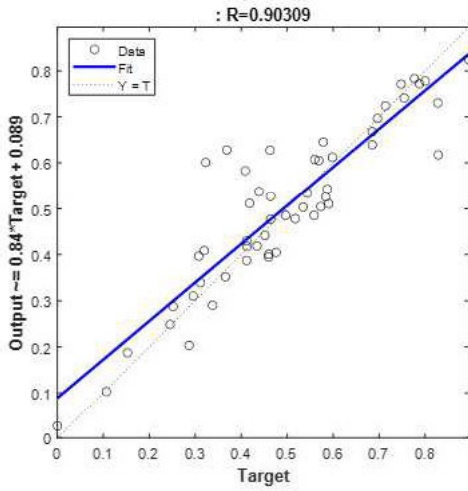


Figure 6.57 Regression fit on training data_cluster label 5 (Five group clustering_ static&Ipgas data)

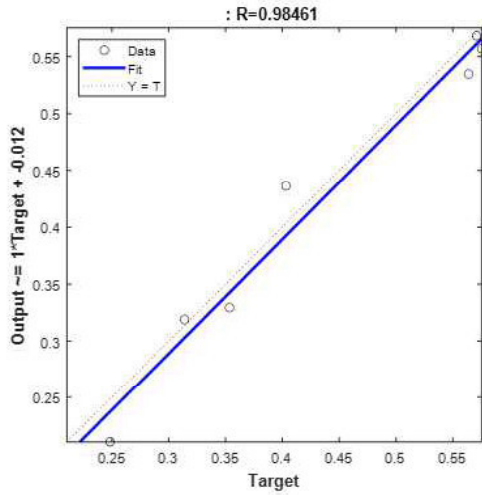


Figure 6.58 Regression fit on validation data_cluster label 5 (Five group clustering_ static&Ipgas data)

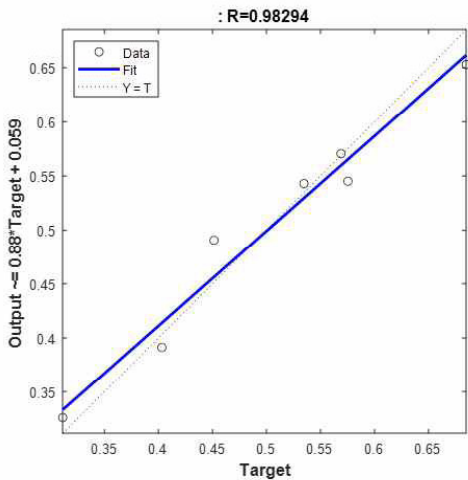


Figure 6.59 Regression fit on test data_cluster label 5 (Five group clustering_ static&Ipgas data)

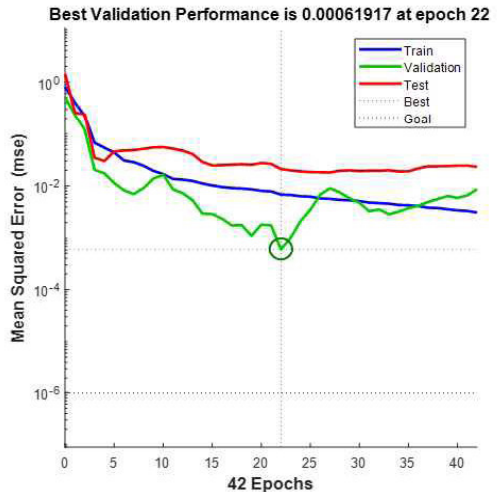


Figure 6.60 Performance plot of best validated model in_cluster label 5 (Five group clustering_ static&Ipgas data)

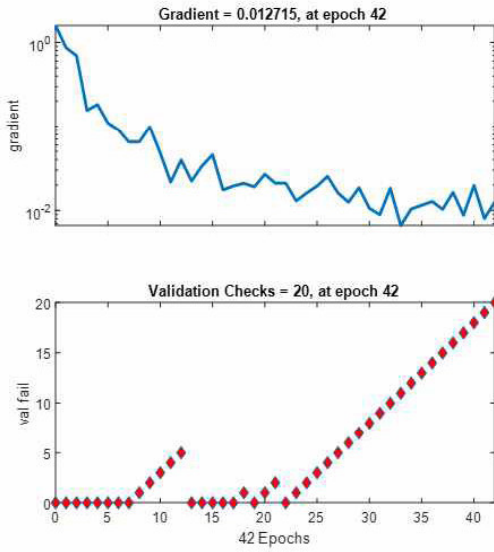


Figure 6.61 Training state of best validated model in_cluster label 5 (Five group clustering_static&Ipgas data)

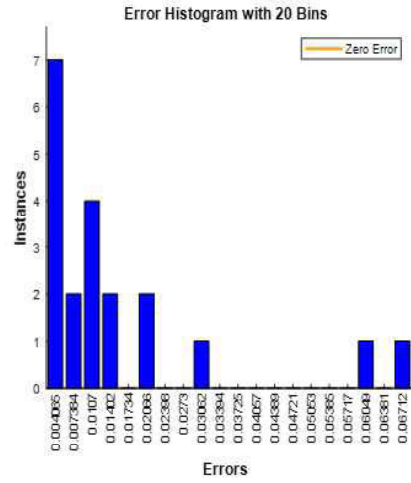


Figure 6.62 Error histogram for simulated models on future dataset (Five group clustering_static&Ipgas data)

Table 6.14 Estimation of total improvement of Clustered models_Case 1

STATIC CASE						
	# Cluster	Train_Rmse/data	Validation_Rmse/data	Test_Rmse/data	Train	Improvement
Total-data	0	0.124789329	0.111041539	0.120025609	19.2%	Test
Cluster_2	2	0.100824028	0.095153222	0.1088467	19.2%	14.3%
Cluster_3	3	0.099815855	0.085050931	0.09845114	20.0%	23.4%
Cluster_4	4	0.076043041	0.074084147	0.091804619	39.1%	33.3%
Cluster_5	5	0.078782168	0.068772364	0.084269113	36.9%	38.1%
						29.8%

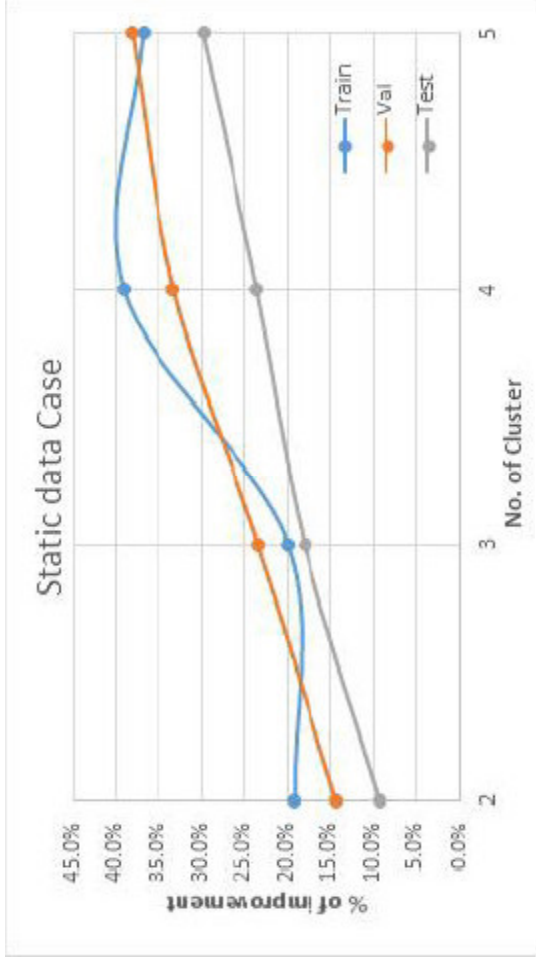


Figure 6.63 Plot of percentages of improvement on clustered predictive model groups vs cluster numbers_Case 1

Table 6.15 Estimation of total improvement of Clustered models_Case 2

STATIC & INITIAL GAS CASE						
	# Cluster	Train_Rmse/data	Validation_Rmse/data	Test_Rmse/data	Train	Improvement
Total-data	0	0.076374292	0.071712415	0.079016021	Test	
Cluster_2	2	0.070554352	0.064771337	0.070819826	7.6%	10.4%
Cluster_3	3	0.060793816	0.055760847	0.063406917	20.4%	19.8%
Cluster_4	4	0.051151968	0.049189733	0.058930116	33.0%	25.4%
Cluster_5	5	0.040114391	0.045779294	0.056496558	47.5%	28.5%

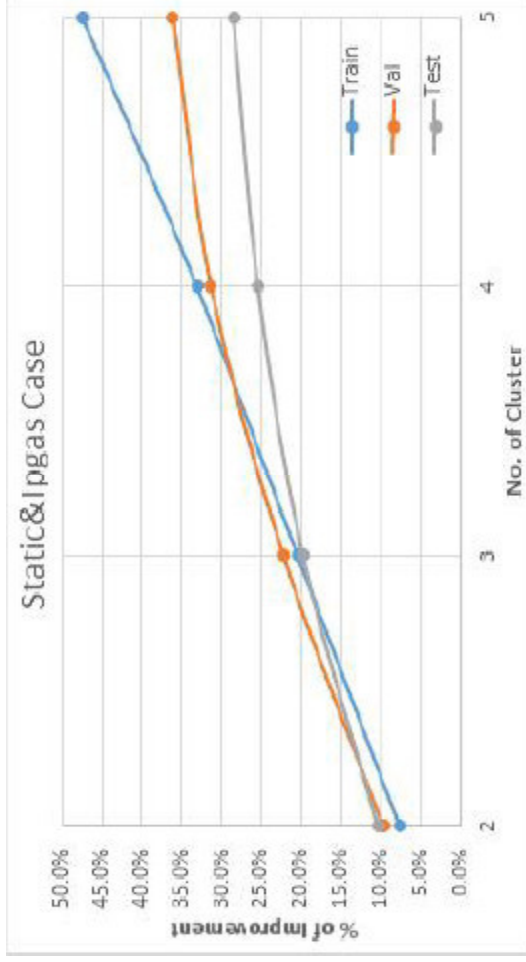


Figure 6.64 Plot of percentages of improvement on clustered predictive model groups vs cluster numbers_Case 2

Table 6.16 Estimation of total improvement of Clustered models_Case 3

STATIC & INITIAL GAS & CUMULATIVE GAS CASE						
	# Cluster	Train_Rmse/data	Validation_Rmse/data	Test_Rmse/data	Train	Improvement
Total-data	0	0.052708931	0.049089515	0.051371051	Train	Test
Cluster_2	2	0.041857108	0.040157402	0.045344323	20.6%	18.2%
Cluster_3	3	0.041256978	0.036643355	0.04192823	21.7%	25.4%
Cluster_4	4	0.032051154	0.033574871	0.039996915	39.2%	31.6%
Cluster_5	5	0.025300144	0.02950026	0.038929498	52.0%	39.9%
						24.2%

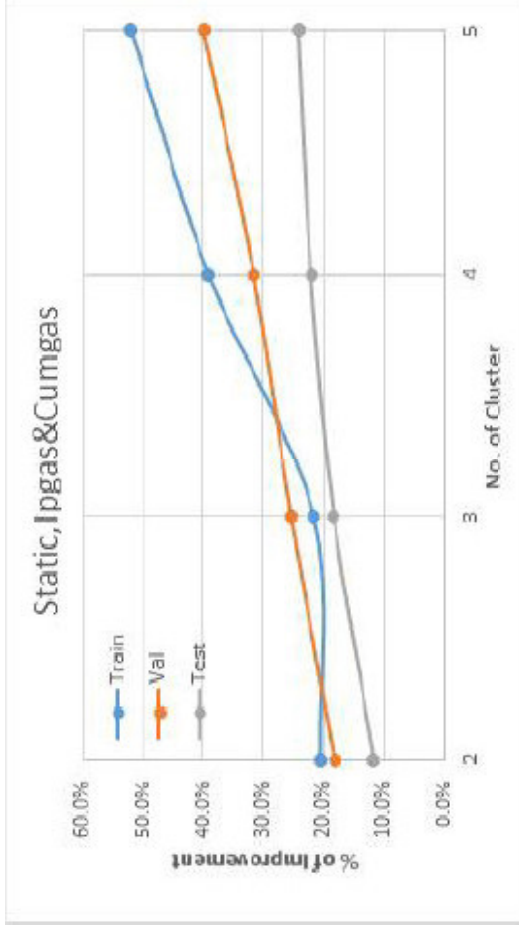


Figure 6.65 Plot of percentages of improvement on clustered predictive model groups vs cluster numbers_Case 3

Table 6.17 Verification of Improvement and Optimal Cluster number with Future data

FUTURE DATA				
Static data		Static&IP gas data		
# Cluster	Sum of Rmse/total cluster #	Sum of Rmse/total cluster #	Sum of Rmse/total cluster #	Static&IP&Cum gas data
1	0.185704484	0.10395825	0.062341119	
2	0.159795627	0.066594323	0.052377968	
3	0.134011745	0.074700683	0.05273278	
4	0.117488913	0.073244877	0.055157971	
5	0.138650314	0.085127199	0.138650314	

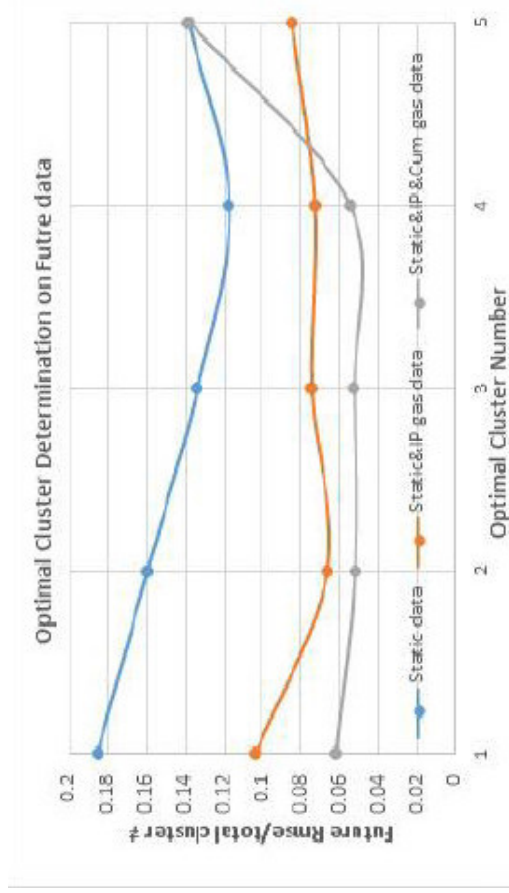


Figure 6.66 Plot of sum of the rmse/total cluster data objects vs cluster numbers on Future data

7. DISCUSSION

7.1 Outputs of the Neural Network: Network Performance

The network performance is designed such that the training reaches to a minimum value at a specific iter value and then continues until a maximum fail is reached where the validation is no more generalizing better in the model. Typically one epoch of training is defined as a single presentation of the input vectors into the network. Generally network performance can be described to undergo cases;

- ◇ Underfitting: Validation and Train errors high
- ◇ Overfitting: Validation error is high and Train error is low
- ◇ Good fit: Validation error is decreasing as Train error improves. Sometimes slightly higher than the train error but is expected and desirable.
- ◇ Unknown fit: Validation error low, training error high.

Ideally, the validation loss should be similar to but can be slightly higher than training loss. As long as validation loss is lower than or even equal to training loss, the learning model would keep on training. If training loss is reducing without increase in validation loss the model still continues with the training process. If the validation loss starts increasing, then the training ceases, a typical example is seen from performance plot in Figure 6.60 in previous section. In Matlab, the default learning is designed such that after six validation checks of no improvement in validation, the training is brought to halt. However in order to induce an increasing iteration, the default was changed to a maximum fail value 20. The number of epochs is related to the number of rounds of optimization applied during the training. With increasing rounds of optimization, the training error reduces further, nevertheless there comes a point where the network may overfit the training data and loses its performance in terms of generalization to non-training or unseen data. To explain and avoid

such instances, selection based on validation for the best 20 models were utilized.

7.2 Outputs of the Neural Network: Error Histogram

In the case of error histograms, the histogram of errors between target values and predicted values after training of the neural network is represented in 20 bins. It indicates how predicted values are differing from the target values, hence error values can be both negative and positive. The error histograms displayed in section 6 are histograms for the optimal validated models from the present data simulated on the future data which also serves as unseen test data. The Y-axis, therefore, represent the number of samples that lies in a particular bin and the X-axis indicates the difference in error of the targets and predicted outputs. Zero error line corresponds to the null value on the error axis (i.e. X-axis). All values to the left of the zero error indicate that the predicted value is greater than the target value and values to the right of the zero error indicate predicted values are lower than the target values. The plots also give a measure of outliers in the data. By comparing the error histogram plots for total future data and clustered group future data, the predicted values from the simulated models were greater than target values in many instances of the total data study cases than the clustered grouping study cases. This can also imply that narrowing down similarities within the data group enhances the model's ability to predict values in a less exaggerated state.

7.3 Outputs of the Neural Network: Regression Plot

The regression plots give the measure of the correlation and fitness between the output and target values which is estimated from the R-Squared values in the train, validation and test data. Generally, R-squared values close to +1 are more desirable and shows closeness to fitness. Since only the best validated model was illustrated, it is duly acknowledged that the best validation model may not necessarily give a higher closeness to fit in training or test case. A

typical case is what is seen in figures 6.45, 6.46 and 6.47 respectively.

7.4 Outputs of the Neural Network: Training State Plot

The train function adopted in building all the regression models with the neural network had activation of the sigmoid function in the hidden layers and the regression $y=f(x)=x$ in the output layers respectively. The training function was the Scaled Conjugate Gradient (SCG). The scaled conjugate gradient training algorithm trains the network as long as its weight, net input and transfer functions have derivative functions and backpropagation is employed in calculating the model's performance with relation to the weight and bias variables. It also requires no critical user-dependent parameters and also avoids time-consuming line search as compared to other algorithms.

The gradient refers to the slope of the square of the error function (known targets-predicted outputs) with respect to the unknown weights and bias terms. The training objective is to optimize the weights and bias by minimizing the sum of squared errors using the strategy of steepest descents.

7.5 Improvement in Predictive Models on Clustered groups

The illustrative diagrams of the k-mean clustered groups showed that the clustering groups were not ideally a "pure" clusters case hence diversion was expected in the training of the data even among cluster groups. The silhouette plots illustrated above indicated portions of negative and positive silhouette values. Despite the lower positive silhouette values indicating lower cohesion within cluster groups, another shift in focus was the separation measures between cluster groups which were necessary for this study. The goal was not to find pure clusters that show enormous discrimination with target variables but to find cluster groups of homogeneous objects with all sections of the data. Therefore the nearest neighbor classifier method was appropriate for the data. A general problem of obtaining clusters in which data are perfectly separated is that models are generally likely to overfit but as seen from the performance

plots, this incidence was avoided with the loop regression models generation step and selection of the best-validated models.

From the ANN model results, the static case study showed approximate improvements of 19%, 20%, 39%, and 37% in the training for 2, 3, 4, and 5 cluster grouping cases respectively. In the static&Ipgas_Case 2, there is improvement in the predictive productivity of models from 8%, 20%, 33%, and 48% approximately from 2 to 5 clustering group each. A similar trend is exhibited in the training of the static, Ipgas&Cum12 gas case with 21%, 22%, 39%, and 52% approximately for the increasing cluster numbers. It is also clear that with the availability of history data there is a massive improvement in (case_study 3) predictive models of the cumulative 60 months gas production.

The results from the present data clustered groups showed very good predictive improvement with the selected models as compared to using unclustered (total) data. The true case in cluster grouping is that an increasing number of clustering generally implies that similarities between data objects are narrowed down to a greater extent but it is quite unclear deciding on the optimal number of cluster groups for the dataset on the present data only. Therefore, further analysis to validate the clustering and ANN models employed was introducing the models to unseen data to establish a linkage with present data cluster groupings based on the elbow point of the minimum estimated error values. From the percentage of improvement plots, it is observed that the optimal number of clusters in the Static data case is realized at number 4 based on the training models but the same could not be said for the validation and testing models. This made it difficult to decide on the optimality in the number of clusters. In the static&Ipgas data case, the improvement increased with an increasing number of cluster groups. The same can be said with the Case 3 study. All three case studies indicated grouping data in 3 groups would be a bad choice on this data as the percentage of improvement plots is elbowed down in the training of the models. With regards to estimating the predictive performance of the clustered data, the results suggested that clustering the data in a distinct group of similar objects will enhance the predictive performances

of the models in all 3 study cases however, the clustering algorithm is not desirable with just a single analysis of splitting data with the present data but generally required another supportive mechanism to verify and validate it. Therefore the approach deployed as previously mentioned to establish and validate the choice of optimal cluster number was to test the performance of the trained models with clusters of the future dataset. This approach was a benchmark for deciding what was the best cluster number.

The plot of optimal cluster determination in Figure 6.66 with the use of the future data showed that the root mean squared errors over the total number of data continued decreasing from the total unclustered data until cluster number 4 where there was a sharp increase in the trend. By using the node of the elbow break in all the 3 study cases, it was determined that the optimal cluster number on the data set was preferably cluster grouping 4. As seen from the plot, the future rmse over total cluster number continued to decrease when the clustering initiated from 2 through to 4 but showed a drastic increase in the error for clustering number 5. The point of inflection in the curves under the 3 case studies indicated the underlying models fit best at that point. This directly relates to the concept of the elbow method of evaluating the value of k .

Although clustering technique is independent of predictive modeling, and estimating optimal cluster numbers on huge multivariate data is quite a daunting task, the approach followed in this research has proven to be a good procedure to combine the two most commonly addressed data mining tasks with error dependencies and to expand on the appropriateness of using conventional clustering in neural network predictive models. It is worth mentioning that building a good predictive model always relies on how good data fed into the model is. If the data indicate good cohesiveness and separation, better performance can be expected. Therefore by generating regression models with the clustered data, the models' predictive performances were enhanced.

8. CONCLUSIONS

In this research study, unsupervised clustering used together with supervised regression prediction although independent of each other was implemented. Both techniques were combined to introduce data clustered groups as an appropriate method for developing predictive models in shale gas assets due to their underlying heterogeneities. The approach introduced in this study tackles the hectic aspect of clustering, which is deciding on the optimal number of cluster groups that was best for predictive models.

The major findings from the study exposed that present data predictive model groups showed continuous improvement with increasing clustering numbers. The productivity of gas from the lower Barnett Shale Formation can easily be predicted on future data with a minimal error by clustered neural network model groups and thus it is possible to improve the predictive performances of models rather than using the total dataset. Specifically to this study, the desired number of the cluster that best fits the input data and produces optimum models was realized at 4. Based on the other algorithms utilized in this research, similar groups in data that were unidentified with the total database were easily recognized, and more accurate predictive models for the cumulative 60 months of gas production were determined with the generated models in each unique cluster group.

It is important to note that constraint-based clustering combined with regression predictive modeling is an under-research topic for shale gas assets and therefore more experiments are necessary to evaluate the proposed paradigm and implementation. The major drawback in this study was the lack of data for a better generalization of the framework on future unseen data. Due to this reason the study could not fully reflect if the optimum cluster number that best fit the model remains constant if the future assigned data is increased. The lack of data was fairly linked to the methods of eliminating data with data preprocessing that hugely shortened the data and therefore further studies on

how to effectively process data whilst keeping the data in its optimum database without restrictions of work in many machine learning models should be considered. Other directions for further research would be further developments and more experiments on a large dataset to establish and evaluate the proposed methods used for the three distinct case studies in this research. The approaches presented in this research can be followed as a methodological guide to validate clustered models in other shale gas assets with the availability of more data, for more intuitive study and deeper interpretations towards clustering analysis with predictive regression rules.

REFERENCES

- Abdi, H., and Williams, L. J., 2010. Principal component analysis. In Wiley Interdisciplinary Reviews: Computational Statistics (Vol. 2, Issue 4, pp. 433 - 459). <https://doi.org/10.1002/wics.101>
- Agatonovic-Kustrin, S., and Beresford, R. 2000. Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research. In Journal of Pharmaceutical and Biomedical Analysis (Vol. 22, Issue 5, pp. 717 - 727). [https://doi.org/10.1016/S0731-7085\(99\)00272-1](https://doi.org/10.1016/S0731-7085(99)00272-1)
- Alkinani, H. H., Al-Hameedi, A. T. T., Dunn-Norman, S., Flori, R. E., Alsaba, M. T., and Amer, A. S. 2019. Applications of artificial neural networks in the petroleum industry: A review. SPE Middle East Oil and Gas Show and Conference, MEOS, Proceedings, March 2019.
- Anirudh V. K., 2020. What is Machine Learning: Definition, Types, Applications and Examples | Toolbox Tech. Potentia Analytics
<https://www.potentiaco.com/what-is-machine-learning-definition-types-application-and-examples>
- Arps, J. J., 1945. Analysis of Decline Curves. Transactions of the AIME, 160(01), 228 - 247. <https://doi.org/10.2118/945228-g>
- Awoleke, O. O., and Lane, R. H., 2011. Analysis of data from the barnett shale using conventional statistical and virtual intelligence techniques. SPE Reservoir Evaluation and Engineering, 14(5), 544 - 556. <https://doi.org/10.2118/127919-PA>
- Bakay, A., Caers, J., Mukerji, T., Miller, P., Cartier, C., and Briceno, A., 2019. Machine learning of spatially varying decline curves for the Duvernay formation. Proceedings - SPE Annual Technical Conference and Exhibition,

September 2019. <https://doi.org/10.2118/196110-ms>

Bansal, Y., Ertekin, T., Karpyn, Z., Ayala, L., Nejad, A., Suleen, F., Balogun, O., and Sun, Q., 2013. Forecasting well performance in a discontinuous tight oil reservoirs using artificial neural networks. Paper SPE 164542 presented at the SPE unconventional Resources Conference held in Woodlands, Texas, 10–12 April

Belyadi, H., Yuyi, S., and Junca-Laplace, J.-P., 2015. Production analysis using rate transient analysis. SPE Eastern Regional Meeting.

Bourquin, J., Schmidli, H., Van Hoogevest, P., and Leuenberger, H., 1997. Basic concepts of Artificial Neural Networks (ANN) modeling in the application to pharmaceutical development. *Pharmaceutical Development and Technology*, 2(2), 95 - 109. <https://doi.org/10.3109/10837459709022615>

Bowker, K. A., 2007. Barnett Shale gas production, Fort Worth Basin: Issues and discussion. *American Association of Petroleum Geologists Bulletin*, 91(4), 523 - 533. <https://doi.org/10.1306/06190606018>

Brun, M., Sima, C., Hua, J., Lowey, J., Carroll, B., Suh, E., and Dougherty, E. R., 2007. Model-based evaluation of clustering validation measures. *Pattern Recognition*, 40(3), 807 - 824. <https://doi.org/10.1016/j.patcog.2006.06.026>

Cao, Q., Banerjee, R., Gupta, S., Li, J., Zhou, W., and Jeyachandra, B., 2016. Data driven production forecasting using machine learning. Society of Petroleum Engineers – SPE Argentina Exploration and Production of Unconventional Resources Symposium. <https://doi.org/10.2118/180984-ms>

Chang, O., Pan, Y., Dastan, A., Teague, D., & Descant, F., 2019. Application of machine learning in transient surveillance in a deep-water oil field. SPE Western Regional Meeting Proceedings, 2019. <https://doi.org/10.2118/195278-ms>

- Clarkson, C. R., Haghshenas, B., Ghanizadeh, A., Qanbari, F., Williams-Kovacs, J. D., Riazi, N., Debuhr, C., and Deglint, H. J., 2016. Nanopores to megafractures: Current challenges and methods for shale gas reservoir and hydraulic fracture characterization. *Journal of Natural Gas Science and Engineering*, 31, 612 - 657.
- Cocchi, M., and Mazzeo, R. L., 2018. Current trends in Artificial Intelligence (AI) Application to Oil and Gas Industry.
- Dubes, R. C., 1987. How many clusters are best? - An experiment. *Pattern Recognition*, 20(6), 645 - 663. [https://doi.org/10.1016/0031-3203\(87\)90034-3](https://doi.org/10.1016/0031-3203(87)90034-3)
- Feder, J., 2020. Machine-Learning Approach Determines Spatial Variation in Shale Decline Curves. *Journal of Petroleum Technology*, 72(10), 65 - 66. <https://doi.org/10.2118/1020-0065-jpt>
- Flippin, J. W. 1982. The stratigraphy, structure, and economic aspects of the Paleozoic strata in Erath County, north-central Texas. *American Association of Petroleum Geologists Bulletin*, 129 - 155.
- Ghaffari, A., Abdollahi, H., Khoshayand, M. R., Bozchalooi, I. S., Dadgar, A., and Rafiee-Tehrani, M., 2006. Performance comparison of neural network training algorithms in modeling of bimodal drug delivery. *International Journal of Pharmaceutics*, 327(1 - 2), 126 - 138.
- Goldberger, J., Roweis, S., Hinton, G., and Salakhutdinov, R., 2005. Neighbourhood components analysis. *Advances in Neural Information Processing Systems*.
- Graves, A., 2011. Practical variational inference for neural networks. *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011, NIPS 2011*.

- Han, D., Jung, J., and Kwon, S., 2020. Comparative study on supervised learning models for productivity forecasting of shale reservoirs based on a data-driven approach. *Applied Sciences (Switzerland)*, 10(4), 1267 - 1276. <https://doi.org/10.3390/app10041267>
- Han, D., Kwon, S., Son, H., and Lee, J., 2019. Production Forecasting for Shale Gas Well in Transient Flow Using Machine Learning and Decline Curve Analysis. *Asia Pacific Unconventional Resources Technology Conference*, Brisbane, Australia, 18-19 November 2019, 1510 - 1527.
- Hassan, A., Elkatatny, S., and Abdulraheem, A., 2019. Application of artificial intelligence techniques to predict the well productivity of fishbone wells. *Sustainability (Switzerland)*, 11(21). <https://doi.org/10.3390/su11216083>
- Holland, J. H., 1992. *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence*. MIT press.
- Hopfield, J. J., 1982. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, 79(8), 2554 - 2558.
- Imam, A., 2020. "Neural Networks, Feed-Forward Neural Networks(FFNN)". *The Startup Medium*. <https://medium.com/swlh/neural-networks-4b6f719f9d75>.
- Jamshidnezhad, M., 2015. *Experimental design in Petroleum Reservoir Studies*. Gulf Professional Publishing. <https://doi.org/10.1016/C2014-0-04184-6>.
- Jarvie, D. M., Hill, R. J., and Pollastro, R. M., 2005. Assessment of the gas potential and yields from shales: the Barnett shale model. *Unconventional Energy Resources in the Southern Midcontinent Symposium Program and Abstracts*, 37 - 50. <https://www.researchgate.net/publication/311569482>

- Jason, B., 2019. "Difference Between Classification and Regression in Machine Learning." Machine Learning Mastery.
<https://machinelearningmastery.com/classification-versus-regression-in-machine-learning/>
- Jiang, L., Cai, Z., Wang, D., and Jiang, S., 2007. Survey of improving K-nearest-neighbor for classification. Proceedings - Fourth International Conference on Fuzzy Systems and Knowledge Discovery, FSKD 2007, 1, 679 - 683. <https://doi.org/10.1109/FSKD.2007.552>
- Jung, Y., Park, H., Du, D. Z., and Drake, B. L., 2003. A decision criterion for the optimal number of clusters in hierarchical clustering. Journal of Global Optimization, 25(1), 91 - 111. <https://doi.org/10.1023/A:1021394316112>
- Kodinariya, T. M., and Makwana, P. R., 2013. Review on determining number of Cluster in K-Means Clustering. International Journal, 1(6), 90 - 95.
- Krasnov, F., Glavnov, N., and Sitnikov, A., 2018. A machine learning approach to enhanced oil recovery prediction. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 10716 LNCS(January), 164 - 171. https://doi.org/10.1007/978-3-319-73013-4_15
- Lolon, E, Hamidieh, K, Weijers, L, Mayhofer, M, Melcher, H and Oduba, O., 2016. Evaluating the Relationship Between Well Parameters and Production Using Multivariate Statistical Models: A Middle Bakken and Three Forks Case History. Paper SPE 179171 presented at the SPE Hydraulic Fracturing Technology Conference, The Woodlands, Texas, USA, 9-11, February.
- Marco, A. C., Mazzeo, L., and Campus, R. 2018. Current Trends in Artificial Intelligence (AI) Application.
- McCulloch, W. S., and Pitts, W., 1943. Review Reviewed Work (s): A Logical

- Calculus of the Ideas Immanent in Nervous Activity by Warren S. McCulloch and Walter Pitts Review by : Frederic B . Fitch Source : The Journal of Symbolic Logic , Vol . 9 , No . 2 (Jun ., 1944), pp . 49-50
Publishe. Journal of Symblic Logic, 9(2), 49 -50.
- Mehana, M., Gultinan, E., Vesselinov, V., Middleton, R., Hyman, J. D., Kang, Q., and Viswanathan, H., 2021. Machine-learning predictions of the shale wells' performance. Journal of Natural Gas Science and Engineering, 88, 103819. <https://doi.org/10.1016/j.jngse.2021.103819>
- Mhatre, S., 2020. What Is The Relation Between Artificial And Biological Neuron? <https://towardsdatascience.com/what-is-the-relation-between-artificial-and-biological-neuron-18b05831036>
- Mohaghegh, S., 2000. Virtual-intelligence applications in petroleum engineering: Part 1—Artificial neural networks. Journal of Petroleum Technology, 52(09), 64 - 73.
- Mohaghegh, S. D., 2005. Recent Developments in Application of Artificial Intelligence in Petroleum Engineering. Journal of Petroleum Technology, 57(4), 86 - 91. <https://doi.org/10.2118/89033-JPT>
- Moore, A. W., and Lee, M. S., 1994. Efficient Algorithms for Minimizing Cross Validation Error. In Machine Learning Proceedings 1994. Morgan Kaufmann Publishers, Inc. <https://doi.org/10.1016/b978-1-55860-335-6.50031-3>
- Nilsson, N. J., 2010. The Quest for Artificial Intelligence - Nils J. Nilsson - Google Books. <https://books.google.co.kr/books>.
- Oh, H. 2020. A Study on the Machine Learning Model for the Production Trend Forecast of Oil and Gas Wells (Master's Thesis). Energy and Resources Engineering Department, Chosun University. Gwangju, South Korea.

- Pearson, K., 1901. LIII. On lines and planes of closest fit to systems of points in space. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 2(11), 559 - 572.
- Peterson, L., 2009. K-nearest neighbor. Scholarpedia, 4(2), 1883.
<https://doi.org/10.4249/scholarpedia.1883>
- Queseda, A., 2020. Five algorithms to train a neural network.
https://www.neuraldesigner.com/blog/5_algorithms_to_train_a_neural_network
- Rosenblatt, F., 1958. The perceptron: A probabilistic model for information storage and organization in the brain. Psychological Review, 65(6), 386 - 408.
<https://doi.org/10.1037/h0042519>
- Samandarli, O., Valbuena, E., and Ehlig-Economides, C., 2012. Production data analysis in unconventional reservoirs with rate-normalized pressure (RNP): Theory, methodology, and applications. Society of Petroleum Engineers - SPE Americas Unconventional Resources Conference 2012, 389 - 407.
<https://doi.org/10.2118/155614-ms>
- Sandnes, A. T., Uglane, V., and Grimstad, B., 2019. Slug flow root cause analysis: A data-driven approach. Offshore Technology Conference Brasil 2019, OTCB 2019, October, 29 - 31. <https://doi.org/10.4043/29925-ms>
- Sarkar, S., Taraphder, U., Datta, S., Swain, S. P., and Saikhom, D., 2017. Multivariate Statistical Data Analysis-Principal Component Analysis (PCA). International Journal of Livestock Research, 7(5), 60. <https://doi.org/10.5455/ijlr.20170415115235>
- Schwab, K., 2016. The Fourth Industrial Revolution: what it means and how to respond. World Economic Forum, 1 - 7.
- Serrano, L. G., 2019. 2.1 What is the difference between labelled and unlabelled data? - Grokking Machine Learning MEAP V13. In Manning publications.

<https://livebook.manning.com/book/grokking-machine-learning/2-1-what-is-the-difference-between-labelled-and-unlabelled-data-/v-4/50>

Szabo, F. E., 2015. Hessian Matrix. Encyclopedia of Computational Chemistry, 1 - 3. <https://doi.org/10.1002/0470845015.chd010>

Tache, I. A., Vasseur, C., Stefanoiu, D., Vermandel, M., and Popescu, D., 2013. Supervised classification of cerebral blood vessels. 2013 2nd International Conference on Systems and Computer Science, ICSCS 38 - 43, August 2013. <https://doi.org/10.1109/IcConSCS.2013.6632020>

Verma, K., and Csed, L., 2021. Use of Fuzzy Set and Neural Network to Extract Fingerprint Minutiae Points and Location.

Wagstaff, K., Cardie, C., Rogers, S., and Schroedl, S., 2001. Constrained K-means Clustering with Background Knowledge. <http://www.cs.cornell.edu/home/wkiri/cop-kmeans/>.

Wang, S and Chen, S. 2019. Insights to fracture stimulation design in unconventional reservoirs based on machine learning modeling. Journal of Petroleum Science and Engineering, 174: 682-695

Wang, H., Ma, C., and Zhou, L., 2009. A brief review of machine learning and its application. Proceedings- International Conference on Information Engineering and Computer Science, ICIECS 2009, 3-6. <https://doi.org/10.1109/ICIECS.2009.5362936>

Wu, J., 2012. Advances in K-means Clustering: a data mining thinking. In Springer Theses: recognizing outstanding Ph.D. Research.

Ženko, B., 2008. Learning predictive clustering rules. Informatica (Ljubljana), 32(1), 95 - 96.

Zhang, J., Liu, L., Fan, Y., Zhuang, L., Zhou, T., and Piao, Z., 2020. Wireless

Channel Propagation Scenarios Identification: A Perspective of Machine Learning. IEEE Access, 8, 47797 - 47806.

<https://doi.org/10.1109/Access.2020.2979220>

Zhang, N., 2015. Scholars ' Mine Steam flooding screening and EOR prediction by using clustering algorithm and data visualization Presented to the Graduate Faculty of the In Partial Fulfillment of the Requirements for the Degree.

저작물 이용 허용서					
학과	에너지자원공학과	학번	20197782	과정	석사
성명	한글 : 앤더슨 마우드 타키와 한문: 영문 : Anderson Maud Takyiwa				
주소	광주광역시 동구 지산동 466-22 단비오피스텔 208호				
연락처	E-MAIL : maudanderson08@chosun.kr /maudanderson08@gmail.com				
논문제목	한글 : 셰일 가스정의 생산예측을 위한 군집화기법 및 인공신경망 복합모델 연구. 영어 : A Combined Study of Clustering Techniques and Artificial Neural Network on Predictive Models for Gas Productivity in Shale Gas Wells.				
본인이 저작한 위의 저작물에 대하여 다음과 같은 조건아래 조선대학교가 저작물을 이용할 수 있도록 허락하고 동의합니다.					
- 다 음 -					
1. 저작물의 DB구축 및 인터넷을 포함한 정보통신망에의 공개를 위한 저작물의 복제, 기억장치에의 저장, 전송 등을 허락함 2. 위의 목적을 위하여 필요한 범위 내에서의 편집·형식상의 변경을 허락함. 다만, 저작물의 내용변경은 금지함. 3. 배포·전송된 저작물의 영리적 목적을 위한 복제, 저장, 전송 등은 금지함. 4. 저작물에 대한 이용기간은 5년으로 하고, 기간종료 3개월 이내에 별도의 의사표시가 없을 경우에는 저작물의 이용기간을 계속 연장함. 5. 해당 저작물의 저작권을 타인에게 양도하거나 또는 출판을 허락을 하였을 경우에는 1개월 이내에 대학에 이를 통보함. 6. 조선대학교는 저작물의 이용허락 이후 해당 저작물로 인하여 발생하는 타인에 의한 권리 침해에 대하여 일체의 법적 책임을 지지 않음 7. 소속대학의 협정기관에 저작물의 제공 및 인터넷 등 정보통신망을 이용한 저작물의 전송·출력을 허락함.					
동의여부 : 동의(O) 반대()					
2021 년 6월 1일					
저작자: 앤더슨 마우드 타키와 (인)					
조선대학교 총장 귀하					