Integrated data analysis approach to select artificial lift method using machine learning.

MAHDI, M.A.A.

2023

Integrated Data Analysis Approach to Select Artificial Lift Method Using Machine Learning

Mohaned Alhaj Abdalla Mahdi

PhD **2024**

Integrated Data Analysis Approach to Select Artificial Lift Method Using Machine Learning

Mohaned Alhaj Abdalla Mahdi

A thesis submitted in partial fulfilment of the

requirements of

Robert Gordon University

for the degree of Doctor of Philosophy.

March 2023

ABSTRACT

The optimisation of artificial lift (AL) selection in the oil and gas industry stands as a critical endeavour, directly impacting production efficiency, cost-effectiveness, and overall operational success. Traditional AL selection methods rely on engineers' time-consuming field data analysis, which is hindered by data heterogeneity and the complexities of finding meaningful correlations among various parameters, resulting in a universal AL selection gap. This gap has led to AL selection inconstancy, uncertainty in AL parameters screening, production loss due to frequent AL replacement following the installation, and extra expenses.

This thesis presents a comprehensive investigation into AL selection, employing innovative machine learning (ML) techniques to upgrade the process by analysing 486,271 data samples, ranging from 2004 to 2021, from 100 wells in a Sudanese oilfield experiencing excessive production loss because of suboptimal AL selection. The study demonstrates the profound impact of ML applications in AL selection, utilizing both supervised learning and clustering techniques. Five supervised ML algorithms are utilised: logistic regression (LR), support vector machines (SVM), K nearest neighbours (KNN), decision tree (DT), and random forest (RF), in addition to K means for clustering. The methodology is applied by developing three distinct ML models, each catering to a unique dataset encompassing production, operation, and environmental/economic parameters. The wells are split into three categories in each model: training, validation, and testing, instead of randomly splitting the datasets. This novel methodology streamlines AL selection by expediting data analysis and affording precise results.

The outcomes of this research are marked by the remarkable improvements in AL selection accuracy and production performance. Validation of the model using actual field data demonstrated its ability to predict AL and optimal size based solely on production data with over 93% and 92% accuracy. Moreover, the model achieved an accuracy of 91% in predicting the optimal AL using only operational data. Economic and environmental data yielded even higher prediction accuracies, surpassing 99%. Key findings indicate that the predicted MTMPCP and GL outperform the current BPU and NF in terms of production and revenue. Well XFE26 is projected to produce 269 STB/D (equating to over 3 million USD yearly

revenue), while Well XJS9 is expected to yield 1878 STB/D (resulting in 11 million USD annual revenue), compared to their current production rates of 97 and 1260 STB/D, respectively.

This thesis delves further into identifying the most influential factors affecting AL and size selection. These factors, namely, gas, cumulative produced fluid, wellhead pressure, well depth, AL setting depth, and AL price, are unravelled through a thorough analysis of the ML models, providing valuable insights into their critical considerations for AL selection in different operational contexts.

In conclusion, this thesis serves as a pioneering exploration of ML applications in AL selection, offering tangible solutions to the challenges faced by the industry. The research concludes in a set of robust recommendations. As the oil and gas sector continues to evolve, this research provides a timely and invaluable contribution, pointing the way towards more efficient, cost-effective, and data-driven AL selection practices.

DECLARATION

The content presented in this thesis originates from research conducted at the School of Engineering, Robert Gordon University, Aberdeen, UK. I affirm that this thesis has not been previously submitted for any other degree or qualification, and all the content is my original work, except where explicitly referenced otherwise within the thesis. This document is a unique contribution to the field and represents my independent scholarly effort.

LIST OF PUBLICATIONS

Two journal papers have been successfully published, and one additional paper is currently undergoing review under the guidance of my supervisory team, with the intention of submitting it for publication.

- An Artificial Lift Selection Approach Using Machine Learning: A Case Study in Sudan, published in *energies.* https://www.mdpi.com/1996-1073/16/6/2853
- A Summary of Artificial Lift Failure, Remedies and Run Life Improvements in Conventional and Unconventional Wells, published in International Journal of Innovative Science and Research Technology (IJISRT). https://ijisrt.com/a-summary-of-artificial-lift-failure-remedies-and-run-life-improvements-in-conventional-and-unconventional-wells

DEDICATION

I dedicate this work to ALLAH and the unwavering pillars of my life—my beloved parents, whose boundless support and love have been the guiding light of my journey. To my siblings, whose camaraderie has made every step of the way a shared adventure.

To my wonderful wife, whose patience, understanding, and encouragement have been the foundation upon which I built this endeavour.

To my precious children, whose smiles and laughter have been my daily inspiration.

With heartfelt gratitude, I dedicate this thesis to my family and friends, for their enduring belief in me and their constant encouragement to reach for the stars. You have been my greatest source of strength and motivation, and I am profoundly thankful for your love and unwavering support.

ACKNOWLEDGMENTS

I extend my deepest gratitude to the numerous individuals whose support and guidance have been instrumental in the completion of this thesis. First and foremost, I would like to express my appreciation to my esteemed Robert Gordon University for providing me with the invaluable resources and opportunities to pursue this research.

I am profoundly indebted to my dedicated supervisory team, Dr Mohamed Amish and Dr Gbenga Oluyemi, for their unwavering encouragement, insightful feedback, and relentless commitment to excellence. Your guidance has been the compass that steered me through the intricacies of this academic journey.

I am also thankful to my family. To my parents (Alhaj and Aljamal), whose boundless love and unwavering support have been the bedrock upon which I built this achievement. To my siblings (Omaima, Wijdan, Mohamed, Samar), whose belief in my potential has been a constant source of inspiration. To my cherished wife (Hadia) and children (Feras, Miral, Mohamed), whose laughter and shared moments provided respite from the demands of academia.

Last but certainly not least, I want to express my thanks to Petro-Energy E&P for supporting me with data. Special thanks to Dr Ahmed Bashir, Rami Saleem, Muhanad Khairy, whose unwavering belief in my abilities provided me with the motivation to persevere through the most challenging of times. Your camaraderie has made this endeavour far more enjoyable.

This thesis is a testament to the collective efforts of all those who have touched my life, and I am deeply grateful for your presence along this journey.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVATIONS AND ACRONYMS

BHP - Bottom hole pressure

P_{wf} – Bottom hole pressure

P_{wh} - Wellhead pressure

ρ_o – Fluid density

g – Gravitational acceleration

h - True vertical depth

ΔP_f – Friction loss

AL – Artificial lift

OGI – Oil and gas industry

BPU – Beam pumping unit

SRP – Sucker rod pump

PCP – Progressive cavity pump

MTMPCP – Metal to metal progressive cavity pump

ESP – Electrical submersible pump

GL – Gas lift

CGL – Continuous gas lift

IGL – Intermittent gas lift

PL – Plunger lift

HJP – Hydraulic jet pump

HPP – Hydraulic piston pump

NF – Natural flow

B/D or bbl/D – Barrel per day

WC% - water cut

IOR - Improved oil recovery

EOR - Enhanced oil recovery

WI – Water injection

N2 – Nitrogen injection

CSS – Cyclic steam stimulation

SF – Steam flooding

CHOP – Cold heavy oil production

CHOPS – Cold heavy oil production with sand

ML – Machine learning

LR – Logistic regression

SVM – Support vector machines

KNN – K nearest neighbour

DT – Decision trees

RF – Random forest

OHE – One hot encoder

GOR – Gas oil ratio

scf – Standard cubic feet

STB – Stock tank barrel

BWPD – Barrel water per day

BLPD – Barrel fluid per day

CAPEX – Capital expenditures

OPEX – Operational expenditures

OOIP – Original oil in place

OGIP – Original gas in place

MM - million

TVD – True vertical depth

MD – Measured depth

PBTD - Plug back total depth

RTUs - Remote transmission units

Dt – Training dataset

σ (z) – Sigmoid function

k – Kernel function

n – Number of classes

X – Input feature

$\hat{Y}$ – Predicted class of feature X

PCA – Principal component analysis

CHAPTER 1
INTRODUCTION

1.1 Background

Some wells can naturally produce oil at the start of production by using reservoir primary drive mechanisms such as solution gas drive, gas expansion, and strong water drive. However, most reservoir energies are finite, will deplete over time, and cannot naturally lift hydrocarbons to the surface (Temizel et al., 2020). The energy can be achieved by installing a downhole pump to reduce the bottom hole pressure (BHP) or injecting gas to minimise the fluid density (Lea, 2007). Natural flow means that the bottom hole pressure can overcome the total pressure loss of the fluid flow from the wellbore upwards to the surface separator. The well is dead when the natural flow stops for two main reasons (Takacs, 2015: p.1-3):

- Bottom hole pressure drops below the total pressure loss due to fluid decline in the reservoir.
- The sum of pressure loss in the fluid column becomes larger than the bottom hole pressure, which results in flow resistance. This is because gas production decreases and results in fluid density increment. Another reason is downhole restrictions such as small tubing size.

The flowing BHP is expressed as:

$$P_{wf} = P_{wh} + \rho_o g h + \Delta P_f \qquad (1.1)$$

Where P_{wf} is the bottom hole pressure, P_{wh} is the wellhead pressure, $\rho_o g h$ is the fluid column (hydrostatic pressure) which is a format of ρ_o the fluid density, h the true vertical depth (TVD), and g the gravitational acceleration, ΔP_f is the friction loss (Nguyen, 2020: p.31). A lifting method is required if any part on the right side of the equation makes the BHP larger.

Artificial lift (AL) is a production system unit that lifts the hydrocarbons from the reservoir to the surface to support insufficient reservoir energy (Lea, 2007). AL is used to produce from the dead wells or to increase the production of the flowing well by confronting the sum of pressure loss. AL produces the hydrocarbons by setting a pump below the fluid level to support the BHP or injecting a pressurised gas to reduce the fluid density to remove the flow restrictions (Takacs, 2015: p.1-

3). AL sometimes is used to obtain flow rates higher than the natural flow rates (Nguyen, 2020: p.111). AL techniques were initially employed in water production prior to their adoption in the the oil and gas industry (OGI), where they have been utilised for over a century. Serving as a cornerstone in the OGI, AL techniques constitute approximately 95% of global oil production (Beckwith, 2014). It is used in conventional and unconventional reservoirs in vertical, deviated, and horizontal wells.

There are several types of AL (**Fig. 1.1**), sucker rod pumping (SRP) or beam pumping unit (BPU), electrical submersible pump (ESP), progressive cavity pump (PCP), gas lift (GL), plunger lift (PL), hydraulic jet pump (HJP), and hydraulic piston pump (HPP). There are no proven estimates of the number of installed AL worldwide; however, there is approximately 350,000 AL deployed across the oil fields in the US (Takacs, 2015: p.1-3).

The inaugural implementation of AL for oil extraction, employing compressed gas in an abundant well declared as unproductive, dates back to 1865. This method involved pumping gas down a high-pressure pipe, resulting in the commencement of oil production at a rate of approximately 30 to 40 B/D, marking a significant breakthrough for the era. Subsequently, ESPs were employed in water production following the invention of the electric motor in 1911 by Arutunoff, who later founded the REDA company (Nguyen, 2020: p.107-108).

The first use of Electric Submersible Pumps (ESP) in oil extraction occurred in 1894, employing a downhole rotary electric motor to operate a plunger pump. A subsequent ESP implementation in 1918 utilized a progressive engine to power a reciprocating plunger pump. It wasn't until 1930 that the Reda Company developed the first commercial ESP, tailored for both onshore and offshore operations, particularly suited for medium to high oil production rates with gas-related constraints.

The inaugural PCP was designed by Moineau in the 1930s for fluid transfer in wineries, later finding application in the petroleum industry in 1936, primarily for heavy oil extraction at rates of up to 6400 B/D. HPP was utilized from the 1950s to the 1970s until supplanted by HJP in the 1970s, offering improved efficiency. The modern hydraulic pump, prevalent in the 1970s, was instrumental in high-production wells, particularly for deviated wells and those with high temperatures,

mixed fluids, solids, and gas, requiring surface-powered fluid facilities for operation (Beckwith, 2014; Fraga et al., 2020).

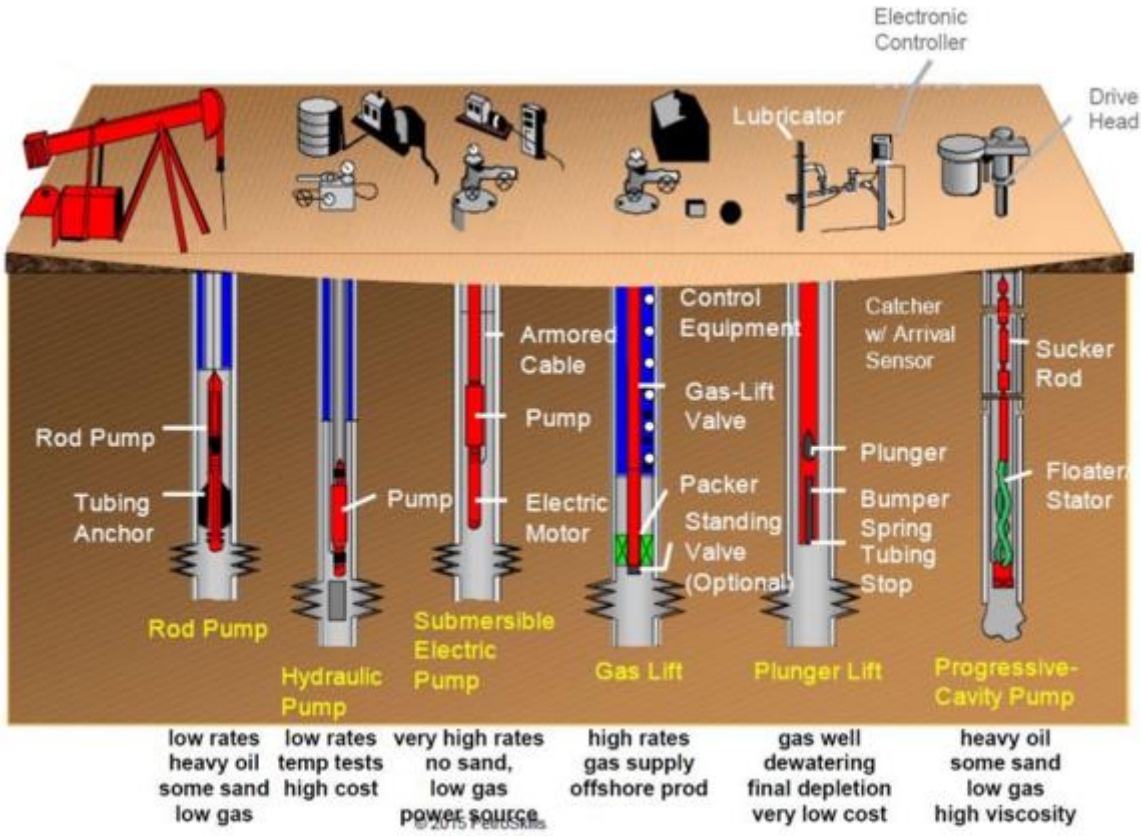


Fig 1.1: AL methods (image source petroskills.com)

1.2 Challenges and Problem Statement

AL selection has been a challenge in the OGI for decades. Optimum AL selection is critical since it determines the daily fluid production (daily revenue) that the oil corporations will gain. The AL selection techniques in the literature followed the same approach by studying the advantages and disadvantages of each lifting method considering field conditions, well and reservoir parameters (Clegg et al., 1993; Syed et al., 2020). The critical issue arises from the interdependence of field parameters, which exhibit variability over successive production years, resulting in inconsistency in selection procedures and necessitating extensive analytical efforts. Consequently, increased expenditures are incurred due to frequent replacement of AL systems within short production intervals. Certain AL specialists contend that conventional selection methodologies have largely fallen out of favour and are, to some degree, outdated (Noonan, 2008). It is due to

technological advances in AL designs and manufacturing, as well as smart metering, which results in big data that requires new selection techniques.

AL selection is complex and critical because the oil field parameters (categoric and numeric) analysed before selection are neither theoretically nor numerically correlated. The conventional selection techniques use qualitative methods, primarily relying on engineers' personal experience (Shi et al., 2019). The conventional selection techniques, especially the selection tables, are the extraction of operation summary and experimental results of AL application in the oil fields. These tables have been used for decades to screen out lifting methods based on fluid properties, reservoir parameters, and field conditions. It is worth mentioning that some parameters have no specific values and vary in most literature. For instance, the flow rate, depth, and temperature limitation of lifting methods fluctuate in selection tables in the literature. **Table 1.1** shows some examples of these variations from the literature.

Table 1.1: AL selection parameters variation examples from literature

Author	BPU	PCP	ESP
Neely et al. (1981)	1000 B/D in shallow wells up to 7000 ft and 200 B/D in deep wells of 14000 ft	-	Applies to any rate above 150 B/D
Brown (1982)	Limited to 12000 ft	-	-
Clegg et al. (1993)	Max 400 B/D at a limited depth of 7500 ft	Limited to 5000 ft. Temperature up to 212°F and 350°F with special elastomer	recommended for flow rates above 1000 B/D. Temperatures below 200°F
Heinze et al. (1995)	-	Recommended 3000 - 4000 ft	Recommended for flow rates above 500 B/D, can produce 100 B/D. Limited to 10000 ft
Lea and Nickens (1999)	-	-	Limited to 10000 ft. Recommended for 20000 B/D
Matthews et. al (2007)	-	Limited to 5040 B/D. Max depth to 9840 ft	-
Takacs (2015)	Up to 16000 ft	Max 6000 B/D. Depth 6000 ft up to 12000 ft. Max operating temperature is 250°F	Up to 400°F
Nguyen (2020)	-	-	Up to 15000 ft

Other parameters, such as AL capital and operating cost, and run life fluctuate according to field conditions. The uncertainty in AL screening and parameter

variations resulted in many AL selection gaps. This pushed many companies to establish their own selection systems or develop computer programs according to their field conditions (Naderi et al., 2014; Espin et al., 1994; Alemi et al., 2010).

1.3 Gaps in Artificial Lift Selection Methods

The uncertainty in field data and screening criteria resulted in the following gaps:

- A universal selection criterion that can be applied to any field condition as the old selection criteria resulted in various selection methods.
- The literature neglected AL optimal size selection, which is essential to consider since it determines the flow rate and affects well production performance.
- AL selection when there is insufficient or missing data. Some data is not recorded due to remote field areas, surface and downhole measuring tools malfunctioning and no calibration, or uninstalled measurement tools. The modelling dataset will be considered as the only available data for analysis rather than the qualitative process.
- Another gap in the AL selection methods is that the crucial field factors have not been critically identified. There should be one or more factors in each data category (production/reservoir, operation, economic, or environmental) that predominantly affect the selection in either adequacy or elimination of the nominated AL.

1.4 Research Aim

The research aims to develop a new data analysis integrated approach to select artificial lift methods.

1.5 Research Objectives

In order to accomplish the proposed research aim and fill in the research gaps, the research objectives will be structured into the following:

- Develop an AL and size selection model based on production and reservoir data.
- Develop an AL selection model based on operation data.
- Develop an AL selection model based on economic aspects, environmental aspects, and safety measures. In addition, the field operator knowledge to AL.
- Develop AL data clustering model.
- Identify the critical field features that chiefly impact the AL selection.
- Validate modelling results with the actual AL field data.

1.6 Conceptual plan

The research strategy is to model numerous field data, production, operation, environmental, and economic, from conventional and unconventional reservoirs. The data was used to predict the most suitable lifting method concerning each field condition. The parameters listed in **Fig. 1.2** are from a conventional sandstone reservoir in a field located in the Muglad basin in Sudan. The model was built using machine learning (ML), and the results were validated with actual field data. Commercial software was used to compare the production performance of the current and newly selected AL. Other lifting methods could have been modelled if more data and unconventional reservoirs data were available.

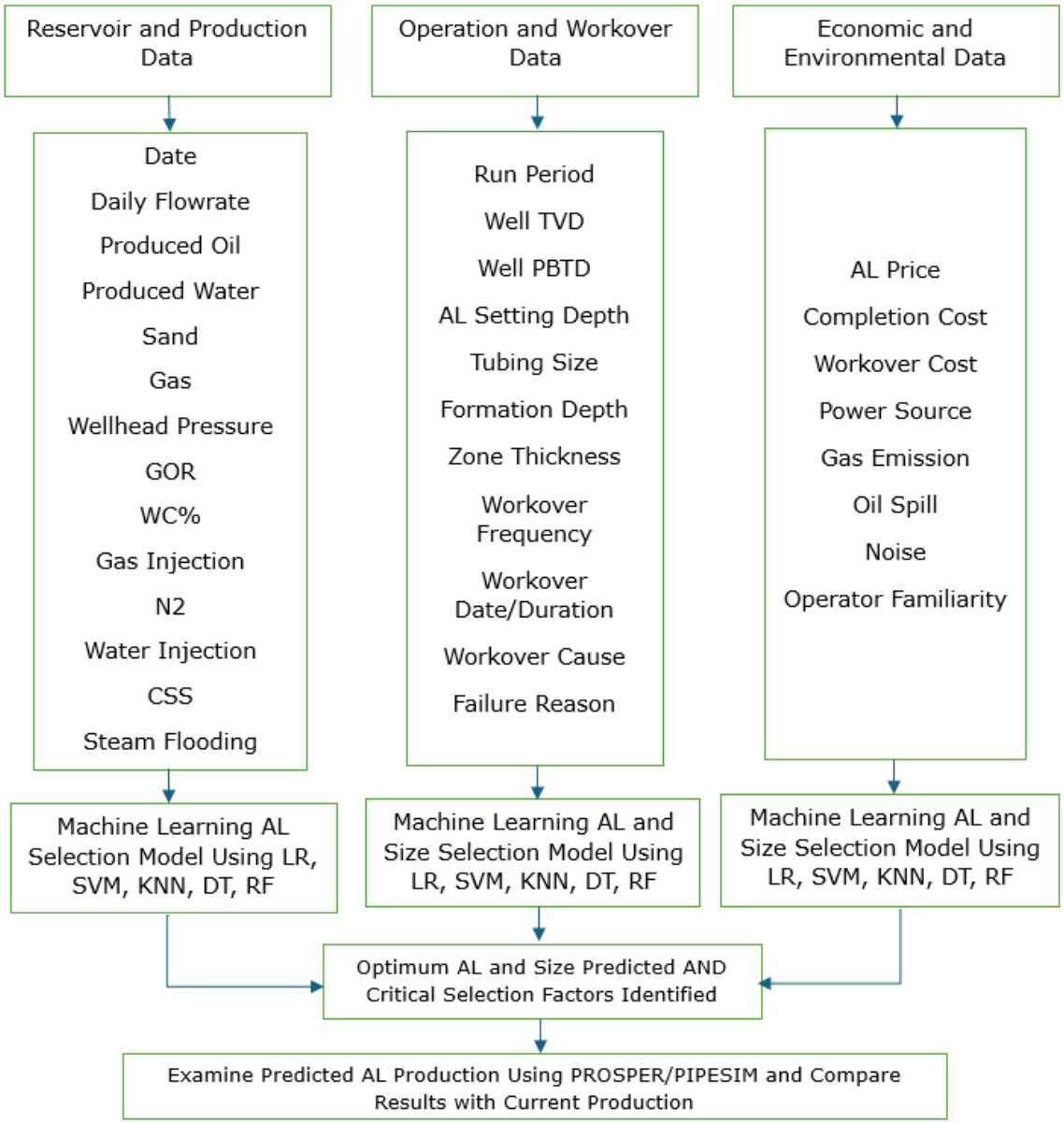


Fig. 1.2: Research work plan

1.7 Research Contribution to Knowledge

This section involves the outcomes from this research that have contributed to the existing knowledge.

- Integrated system using ML algorithms to select optimum AL in conventional reservoirs. The system can be used for current and future oil wells according to their current field conditions. Also, according to AL capability with different enhanced and improved recovery methods.
- The Developed model facilitates selection from a set of available data; meanwhile, there is insufficient or missing field data. The model result validation with actual field data showed that it could predict AL from only production data with an accuracy above 93%. The model also predicted the optimum AL with an accuracy of 91% from only operation data. The model prediction error of the optimum lifting method from economic and environmental data is 1% (above 99% accuracy).
- The developed model provides the optimum AL size while the lifting method is used as an input feature. The model has results above 90% accuracy when applied to production data.
- The developed model predicts the critical factors that primarily impact AL selection concerning the available field data in each field data category. Gas, wellhead pressure, and applied recovery methods are crucial factors in the production and reservoir data. Depth of, well, formation, and setting AL are the most important operation factors to consider when selecting a lifting method. Ultimately, the essential environmental and economic factors in selecting the AL are the amount of oil spill, price of AL, gas emission and noise.
- The developed model nominates the optimum lifting method to improve the current producing well's production performance resulting in extra obtained revenues. In addition, the model serves as a field exploratory data analysis tool for field parameters.

1.8 Thesis Layout and Structure

The thesis comprises seven chapters, below paragraphs are a general overview of the underlying concepts in each chapter.

Chapter one introduces a general background and introduction of the common AL methods used in OGI. In addition, the challenges that drive this work, the research aim, objectives, and the contribution to existing knowledge.

Chapter two presents in-depth introduction to the four targets lifting methods and an extensive literature review to the AL selection in conventional and unconventional reservoirs along with AL failure issues. It also provides a review of the application of ML in OGI and AL, as well as the selected algorithms used in modelling in this research.

Chapter three outlines the proposed methodology, data acquisition, preparation, wrangling and pre-processing. Features selection and the application of ML algorithms in data analysis and visualisation. The chapter also gives an overview of the specific field, a description of the sandstone reservoir, and the selected production wells used in modelling.

Chapter four provides the modelling results of ML application in AL selection, through the three field categories models. It also gives the validation with the actual field data.

Chapter five assesses the application of ML in AL size selection and the validation with the actual field data.

Chapter six highlights the critical field parameters that mostly affect the selection in each field data category. It also provides a sensitivity analysis of the performance of the selection models in addition to ML results comparison to commercial software.

Chapter seven summarises the main findings and the recommendations for future work.

CHAPTER 2
ARTIFICIAL LIFT METHODS AND SELECTION REVIEW, AND MACHINE LEARNING ALGORITHMS REVIEW

2.1 Introduction

This chapter presents an extensive literature review of both old and recent AL selection techniques, failure causes, and remedies for each AL. The chapter also provides a brief introduction to the five supervised learning algorithms used for modelling, namely the most commonly utilised ML techniques in current OGI applications SVM and DT, LR, KNN, and RF. K-means was used for unsupervised learning clustering model. Furthermore, the section provides a comprehensive review of the applications of ML in the OGI, particularly in the selection and failure analysis of AL.

2.2 Artificial Lift Methods

2.2.1 Beam Pumping Unit

The beam pump, reciprocating pump, pump jack, sucker rod pump, and rod pump are different names of the most broadly used and oldest lifting method in the OGI. The beam pump working principle is the same as the other lifting method by reducing the BHP to increase hydrocarbon production (Nguyen, 2020). **Fig. 2.1** illustrates the conventional BPU components. It consists of a surface unit that operates and carries the weight of the downhole unit. The rods connect the surface unit to the downhole plunger pump. The rotation of the prime movers is transmitted to the gear reducer to reduce the speed. Then, the rotation from the prime movers is converted into a reciprocating motion via a mechanical system consisting of the counterbalance, crank, pitman, walking beam, and horsehead. The polished rod connects the horsehead upside and the rod string downside. The last rod is connected to the pump barrel, which contains a standing valve, a travelling valve, and a plunger (Nguyen, 2020).

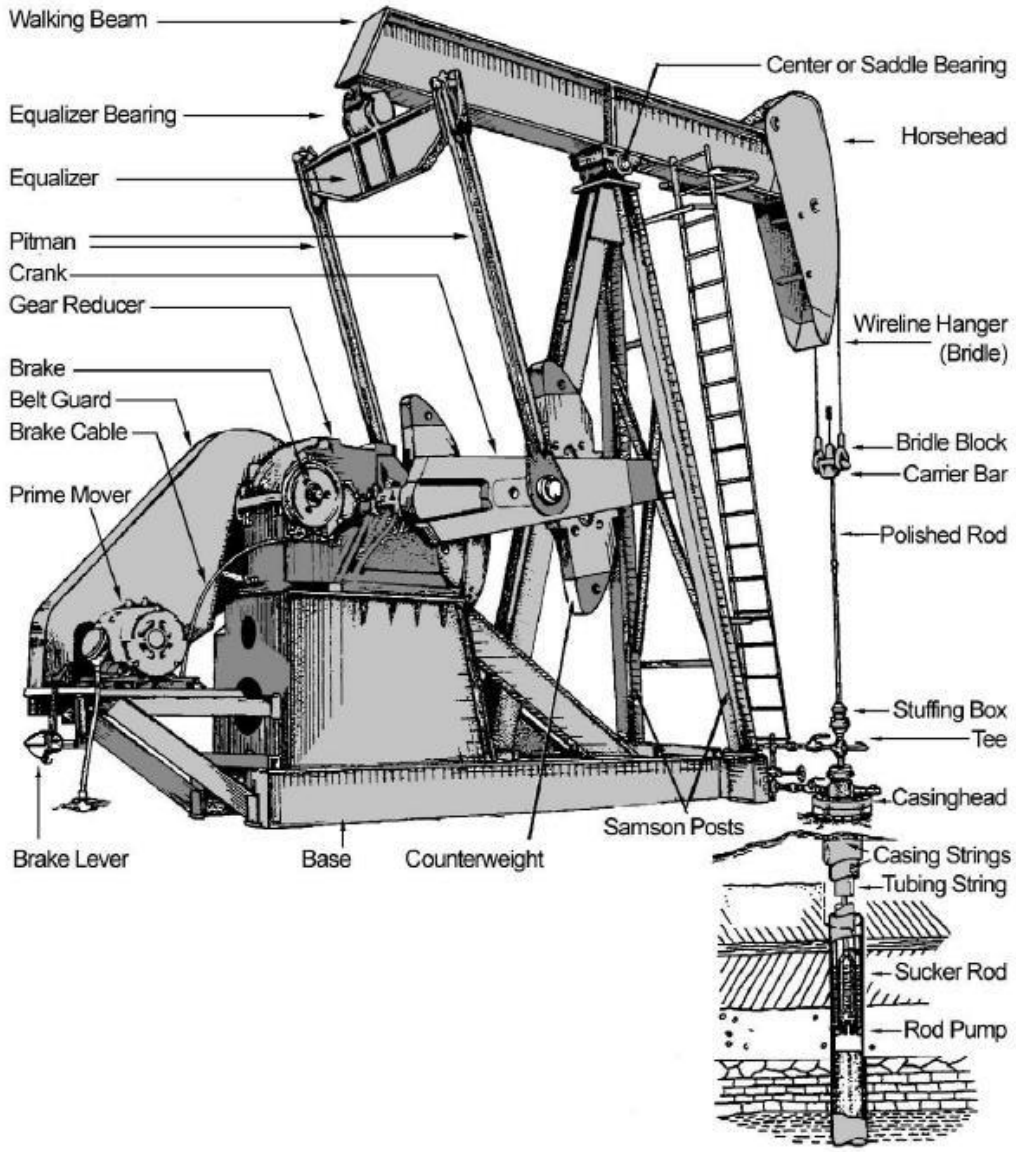


Fig. 2.1: Conventional BPU surface and downhole components (Hein, 2007)

Fig. 2.2 shows the principle of BPU downhole pump operation. During the upstroke (upward movement), the travelling valve closes; meanwhile, the standing valve opens to let the fluid enters the pump barrel, and the fluid in the annulus between the rod string and the tubing is lifted. During the downstroke (downward movement), the standing valve closes, and the travelling valve opens to let the fluid pass to the pump barrel's upper part and be stored (Nguyen, 2020). This cycle is continuously repeated as long as the surface unit operates. In the end, oil production results from transforming the surface unit parts rotation into downhole piston-like displacement. There are other BPU types that differ in the surface unit, such as air-balanced rod pump and vertical rod pump.

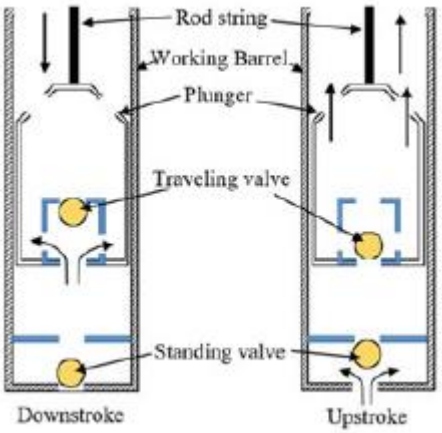


Fig. 2.2: Upstroke and downstroke principle (Nguyen, 2020)

BPU is used in wells with low reservoir pressure to produce medium to low rates of less than 10 B/D, which are called stripper wells (Nguyen, 2020). It can produce from 150 B/D at a depth of 14000 ft to 3000 B/D at less than 2000 ft (Hein, 2007). **Fig. 2.3** illustrates the working principle of the BPU system to reduce the BHP, which relies on ΔP_{pump} the differential pressure existing between the pump intake and discharge. The BPU principle is a fundamental mechanism shared by most pump systems. As shown in the figure, the intersection between the in-flow (IPR) and out-flow (OPR) relationships gives the operating pressure and flow rate points. P_{wf}^{nf} is the naturally flowing well BHP that is higher than P_{wf}^{pump}, the BHP after installing an AL. Q_e^{nf} represents the well natural flow rate which is lower than Q_e^{pump}, the rate obtained by a pump.

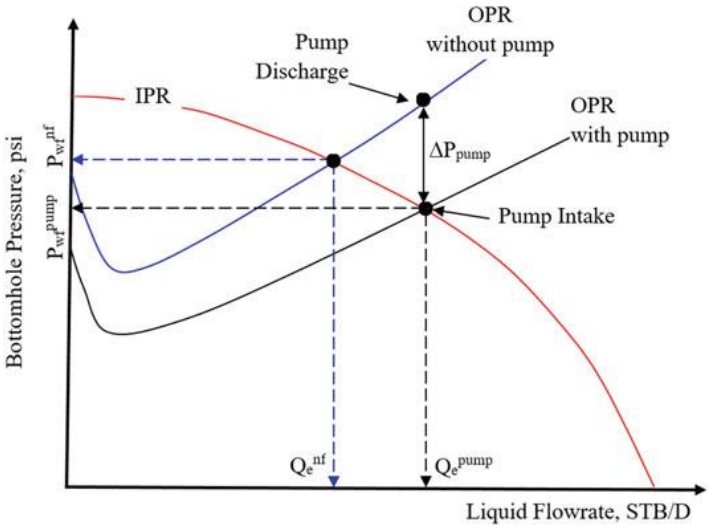


Fig. 2.3: Working principle of BPU (Nguyen, 2020)

BPU is used in unconventional reservoirs; nonetheless, some parameters limit its application in gas hydrate reservoirs, for instance, relatively high-power demands and gas and sand plugging issues. If the pump is used in gas hydrates, it is installed below the perforations to let the gas freely flows through the annulus while the water is pumped through the production tubing. **Table 2.1** summarises the advantages and disadvantages of BPU.

Table 2.1: BPU advantages and disadvantages (Brown, 1982; Clegg et al., 1993; Hein, 2007; Lea and Nickens, 1999; Neely et al., 1981)

Advantages	**Disadvantages**
Simple and easy to operate by field operators	Downhole equipment maintenance requires pulling out
Can pump with low well pressure	Probable solids and paraffin deposition
Low cost of surface unit replacement	Crooked holes cause friction problem
System parts easily transferred to other wells at low cost	Low volumetric efficiency in gassy wells
Applicable with several well completions	Inadequate in offshore and urban areas
Power source can be gas or electricity	Downhole pump may become gas locked
Corrosion and scale treatments are easy to implement through annulus	Depth limitation due to rod weight
Lift high temperature and viscous oil	Downhole pump limitations in small size casing
Analysable and has wide range of knowledge	Special requirements needed for irrigated fields installation
Various sizes available	Environmental concerns of stuffing box leaks
Flexibility to match displacement rate to well capability as well declines	

Double valving pumps can pump in both upstroke and downstroke Automation applicable	

2.2.2 Gas Lift

Some oil pioneers consider GL as the oldest notable AL. They argued that a naturally flowing well is a gas lifted well, and the gas is compressed by nature (Beckwith, 2014). All GL methods use pressurised gas, typically natural gas, and in some cases, N2 and CO2 (Takacs, 2015). There are two GL types: the widely used continuous GL (CGL) and intermittent GL (IGL). In CGL, continuous pressurised gas flown from surface compressors is injected into the bottom of the formation fluid through the annulus as shown in **Fig. 2.4**. The gas enters the tubing through pressure-controlled gas valves installed inside the GL mandrels. The gas then mixes with the fluid in the tubing string to reduce the flow resistance from hydrostatic pressure and friction loss. In other words, after mixing, the fluid density decreased as well as the total pressure loss; therefore, the BHP can lift the hydrocarbon to the surface (Nguyen, 2020).

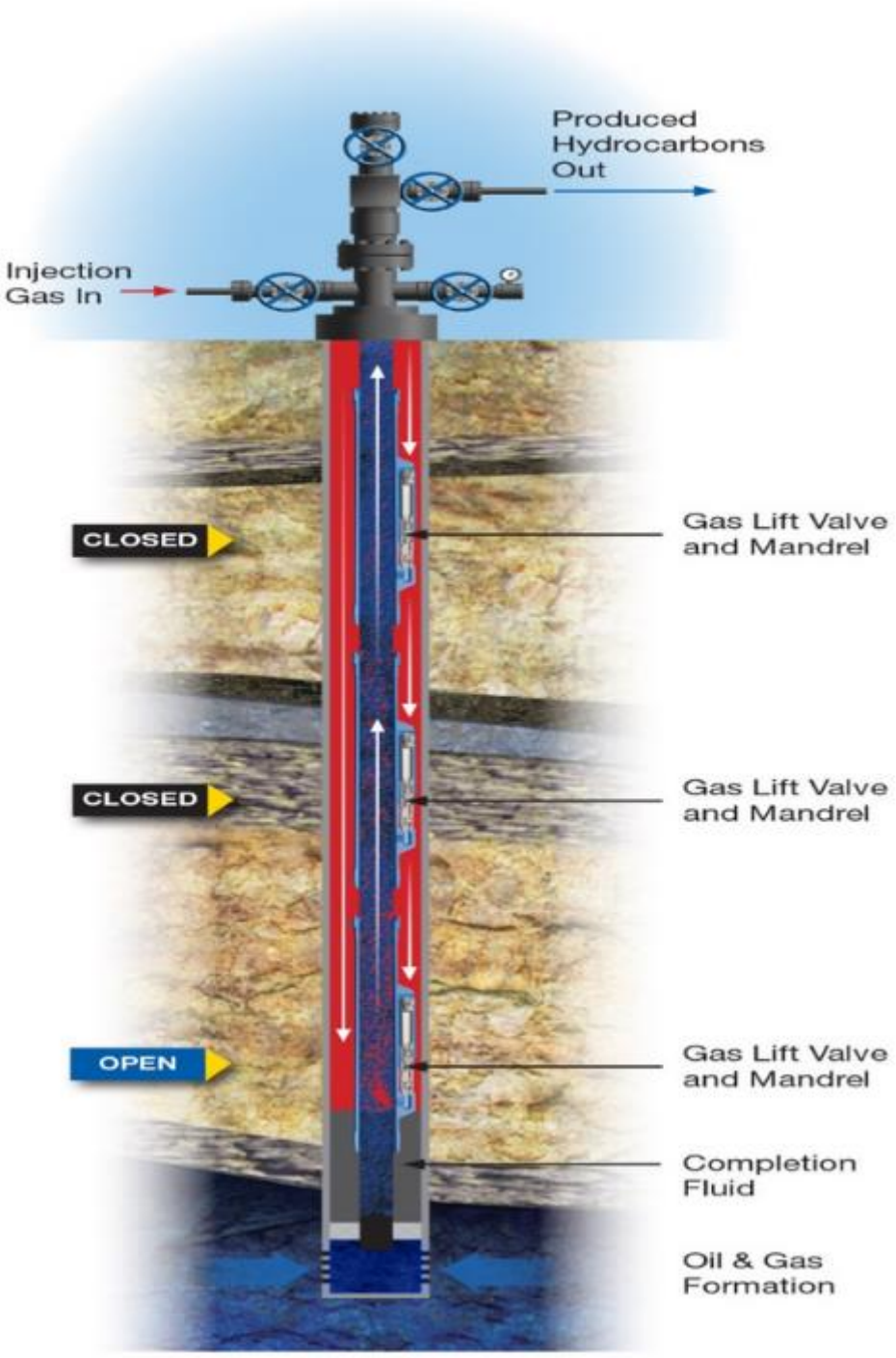


Fig. 2.4: GL system (image source oilfieldbasics.com)

Fig. 2.5 shows the pressure gradient of BHP P_{wf}, wellhead pressure P_{wh}, and saturation pressure P_{so} as well as IPR principle of CGL. The injected gas into the annulus decreases the flowing BHP P_{wf1} to P_{wf2}, which lets the fluid flows into the well resulting in production increase from Q_1 -the well natural flow (NF) rate- to Q_2 -CGL flow rate- (Nguyen, 2020). The CGL is considered a resumption of NF (Takacs, 2015).

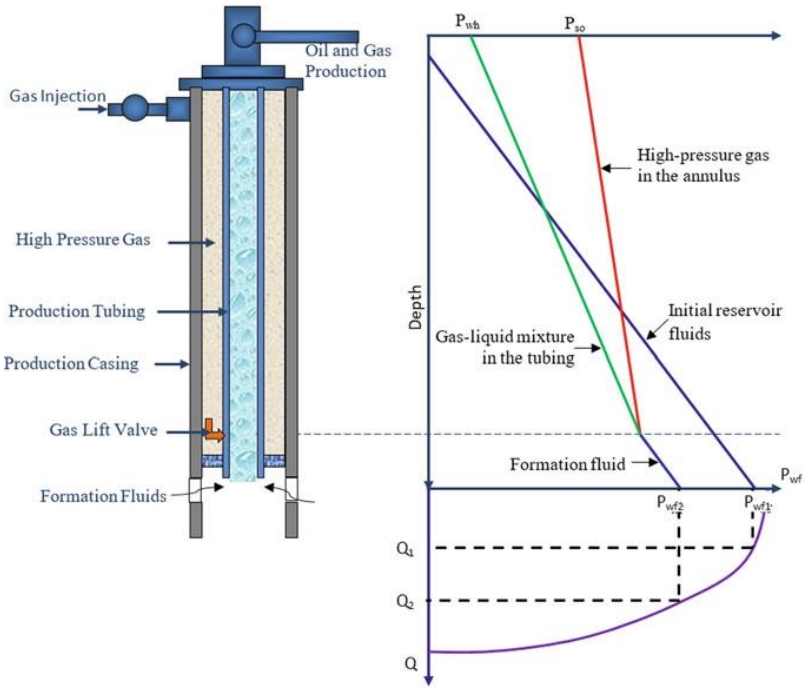


Fig. 2.5: Working principle of CGL (Nguyen, 2020)

The idea of IGL is not to reduce the fluid density as in CGL; its primary purpose is to displace the accumulated slug to the surface. When a slug periodically occurs in the fluid column, a high volume of gas is injected below the slug. As far as the slug is produced, the gas injection is interrupted for fluid volume build-up (Takacs, 2015). **Fig. 2.6** shows the stages of a complete IGL cycle.

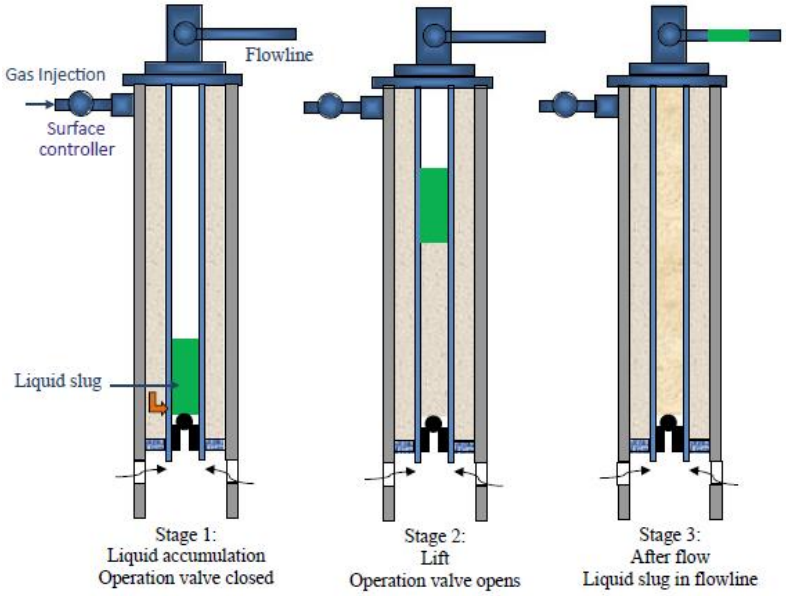


Fig. 2.6: Stages of IGL cycle (Nguyen, 2020)

GL produces high production rates, nearly up to 50000 B/D. It requires a high volume of associated gas in the reservoir and a high-pressure supply source at the surface (Winkler and Blann, 2007). In unconventional wells (unconventionals), gas is injected intermittently to remove water, reduce fluid column density, and produce liquid slugs. Also, intermittent injection is considered advantageous in removing solids that could affect the later-installed AL (Sahu et. al, 2021). The common GL advantages and disadvantages are presented in **Table 2.2**.

Table 2.2: GL advantages and disadvantages (Winkler and Blann, 2007; Brown, 1982; Clegg et al., 1993; Lea and Nickens, 1999; Neely et al., 1981)

Advantages	**Disadvantages**
Can handle high amount of solids	Availability of gas source and low GOR
Can handle high volumes in high productivity index well (50000 B/D)	Limited space for compressors in offshore platforms
Flexible with different rates and depths	Unfeasible with viscous fluid and emulsions
Easy to convert from CGL to IGL or to other AL methods such as plunger lift	Capital cost of lines and compressors is high
Remote power source locations	Inefficient in small fields or one well in term of cost
High deviated wells with high gas oil ratio (GOR)	Require engineering supervision for proper analysis
Urban locations friendly	Casing endurance to high pressure
Offshore compatible	Not reliable with wet gas if not dehydrated
No need to kill the well or pull out the tubing to replace wire-line retrievable valves	Cannot effectively produce if deep wells decline
Pressure gradients easily obtained	Safety concerns due to high pressure
Corrosion is not a concern	
No problems from Crooked hole	

Surface equipment easy to operate and maintain Small installation space Low capital cost of well equipment	

2.2.3 Electrical Submersible Pump

ESP is a type of AL known as rod-less pumps, with no rod string operated from the surface units (Takacs, 2015). An ESP system (**Fig. 2.7**) consists of a surface unit that includes the transformer that transforms the voltage from the electricity source to the downhole motor. The switchboard and junction box control the motor speed and connect the 3-phase electric cables to the well. Another surface component is the wellhead which is commonly an x-tree with chokes and bleeding valves. The downhole unit consists of an electric motor with multiple speeds reaching 3500 rpm at 60 Hz, including sometimes a shroud installed for motor cooling. Another element is the protector to prevent the fluids from entering the motor. The main downhole ESP component is the multistage centrifugal pump designed according to the desired rate, head, wellhead pressure, and friction loss (Nguyen, 2020).

Fig. 2.8 illustrates the working principle of the ESP system in comparison to the NF system. Similar to BPU, the ESP system depends on the differential pressure between the pump intake and discharge. The intake pressure is the lowest BHP equal to the outflow pressure (OPR), while the discharge pressure equals the inflow pressure (IPR).

ESP is recommended for high volume production rates and can produce from 200 to 60000 B/D to a max depth of 15000 ft (Nguyen, 2020). Bearden (2007) mentioned that the ESP can produce from 150 up to 1500000 B/D. The theoretical pump rates are always higher than the actual pump flow rates for many reasons, such as fluid compressibility change, gas presence, and fluid leaks inside the pump (Nguyen, 2020). ESP is designed to produce liquids, and the increased free gas negatively affects pump efficiency and fluid flow, although a gas separator is installed (Bearden, 2007).

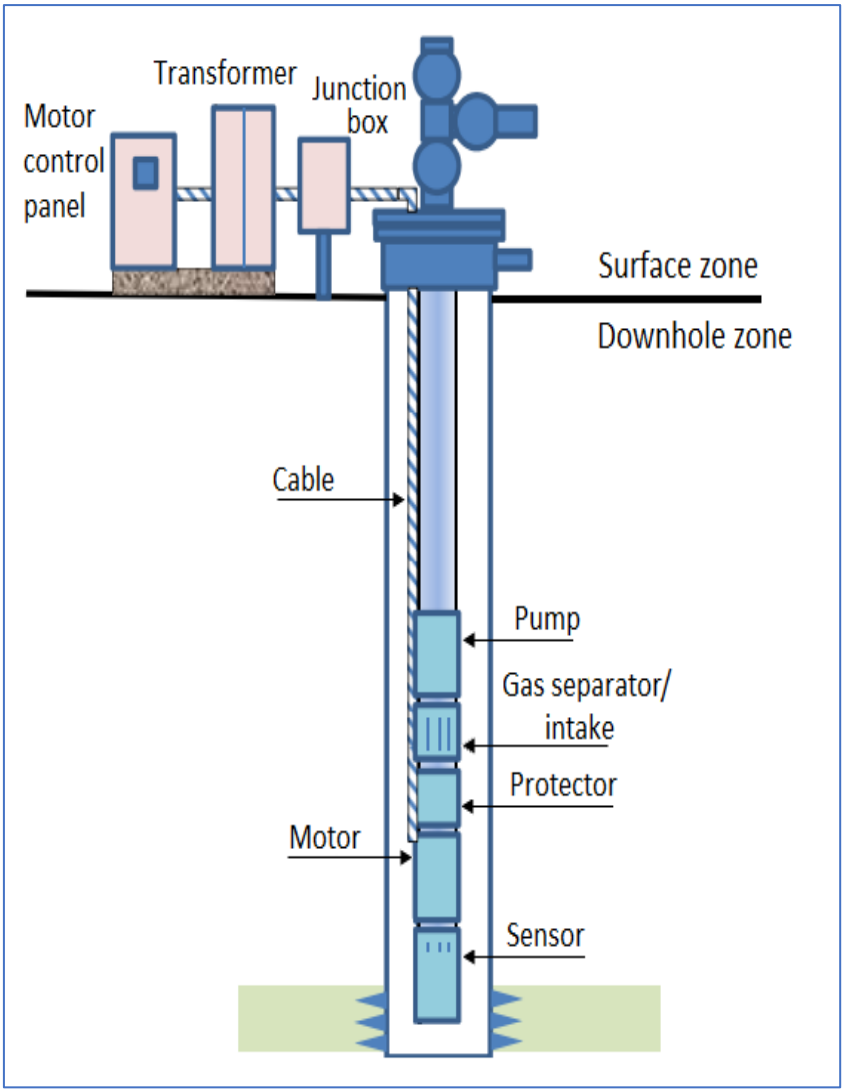


Fig. 2.7: ESP components (Fonsêca et al., 2019)

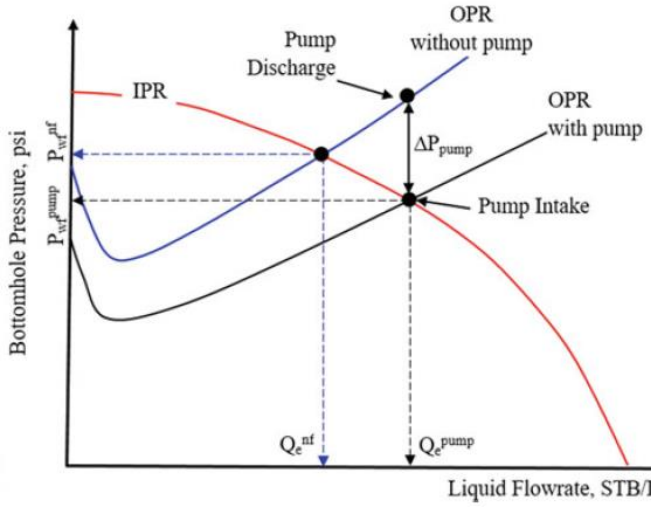


Fig. 2.8: Working principle of ESP (Nguyen, 2020)

The main advantages and disadvantages of ESP are summarised in **Table 2.3**

Table 2.3: ESP advantages and disadvantages (Bearden, 2007; Brown, 1982; Clegg et al., 1993; Lea and Nickens, 1999; Neely et al., 1981).

Advantages	**Disadvantages**
Simple and easy to operate by field operators Produce high volumes up to 150000 B/D Unhindered in urban areas Offshore compatible Downhole pressure and temperature sensors easily installed via cable No problems from Crooked hole Corrosion and scale treatments are easy to implement through annulus Various sizes available Low cost for high volumes lifting	Efficiency decreases < 40 % for rates < 1000 B/D Multiphase flow problems Electric power is a must with high voltages (1000 V) Impractical in low volume wells High temperature damages cables Equipment change is expensive Gas and solids problems Requires proper engineering knowledge for analysis Casing size and depth (10000 ft) limitations Shroud is needed to route fluid by the motor More downtime if pump fails as all parts are downhole Extra power lead to extra cost

2.2.4 Progressive Cavity Pump

PCP, also known as the Moineau pump, is primarily used for medium flow rate capacity. PCP has recently become more popular than the BPU in terms of installation and production cost (Takacs, 2015, Nguyen, 2020). PCP is a positive displacement pump commonly used to lift heavy crude, highly viscous, solid contents hydrocarbons, medium to light oil, cold heavy oil production with sand (CHOPS), and extra heavy and bitumen thermal production. It is also used to lift coal bed methane in unconventional reservoirs (Nguyen, 2020). In addition, PCP is preferable in gas hydrates for its capability to produce from deviated and horizontal wells, efficient sand handling capacity and low power demands. PCP is considered a modern lifting method since the pump was invented in 1930 by Rene Moineau. It consists of two parts; (1) a downhole pump (stator) with rubber or metal cavity elastomer and spiral steel rotor, and (2) a surface drive head system which rotates the rod string and so the rotor to lift fluids to the surface as presented in **Fig. 2.9** (Matthews et al., 2007).

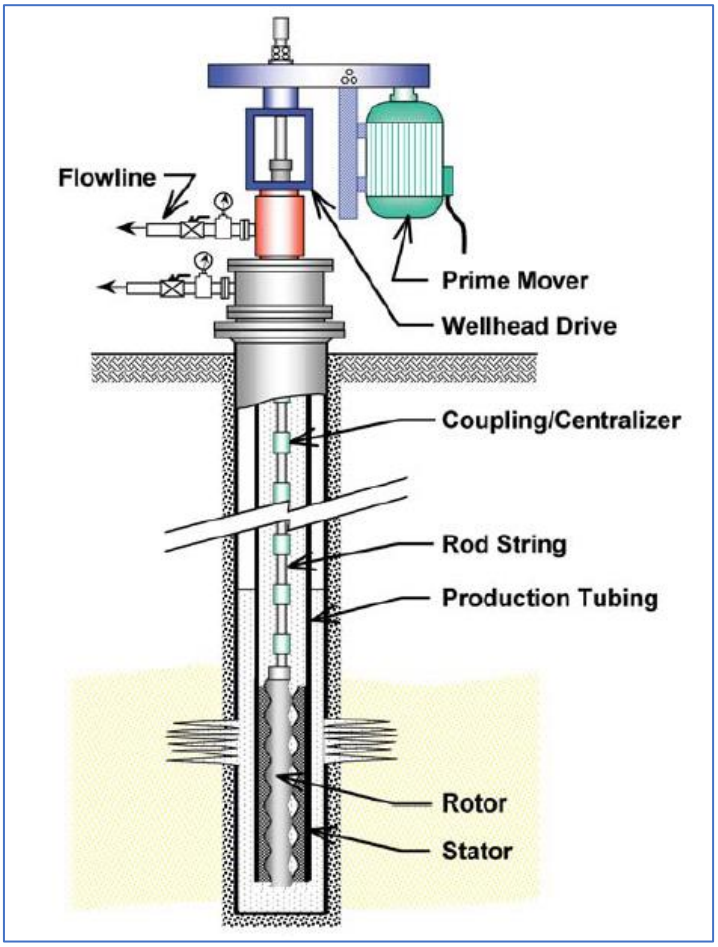


Fig. 2.9: PCP components (Matthews et al., 2007)

The PCP downhole composition is a single helical steel rotor inside the double helical elastomer or steel stator to create multiple cavities as shown in **Fig. 2.10**. The rotor is designed so that all its teeth are in constant contact with the stator. As the drive head rotates and hence the rotor turns, the fluid accumulates inside the cavities and discretely displaces through the tubing upwards to the surface. This movement is similar to the mechanism of the positive displacement pumps. PCP outweighs the conventional BPU in corrosive, viscous and sandy wells because of the common problems that confront the couplings abrasion, rod strings disconnection, and pump travelling valves blocking due to reciprocating movement. The rotating movement of PCP parts gives it the advantage of encountering the problems above (Nguyen, 2020).

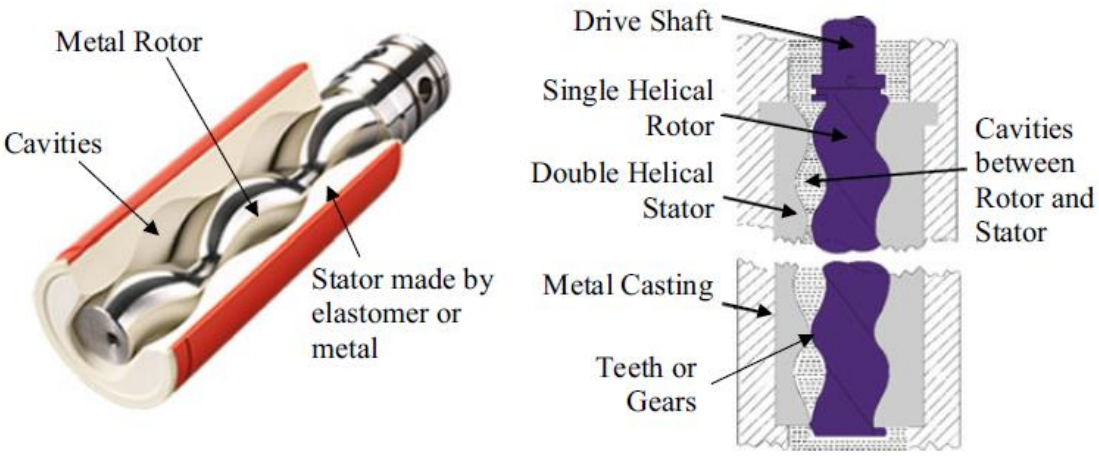


Fig. 2.10: PCP stator and rotor design (Nguyen, 2020)

PCP is designed to have single-lobe 1:2, meaning the rotor has one gear or one tooth, and the stator has two gears or teeth. Some PCPs have multi-lobe, which means that the rotor and the stator are designed with multiple gears, as shown in **Fig. 2.11** (Nguyen, 2020)

Fig. 2.11: Single-lobe and multi-lobe PCPs (Nguyen, 2020)

Fig. 2.12 demonstrates the working principle of PCP inside an oil well, and its concept is the same as in BPU and ESP. The PCP converts the surface electric energy to hydraulic energy. It changes the flow inside the tubing by reducing the fluid column hydrostatic pressure and friction loss, similar to most submersible pumps (Nguyen, 2020).

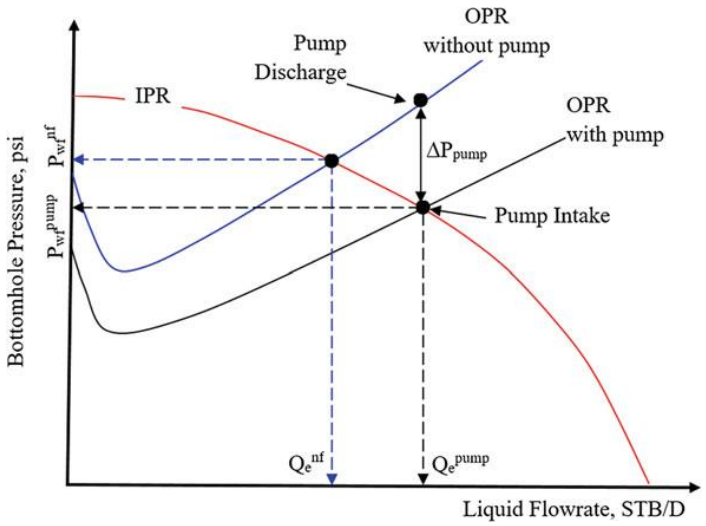


Fig. 2.12: Working principle of PCP (Nguyen, 2020)

Efficiency refers to converting mechanical energy to hydraulic work with respect to the prerequisite power and pressure losses. PCP is considered the best lifting method in terms of efficiency, which can be above 70% because of low energy loss and the simplicity of connecting the surface and downhole pump components. The BPU and ESP have a maximum efficiency of 60%, HPP is around 50%, and GL and HJP are between 10 and 30%. PCP is undoubtedly the desired AL substitute if the oil well conditions deteriorate below the operating ranges (Takacs, 2015).

PCP advantages and disadvantages are shown in **Table 2.4** below.

Table 2.4: PCP advantages and advantages (Brown, 1982; Clegg et al., 1993; Lea and Nickens, 1999; Matthews et al., 2007; Neely et al., 1981)

Advantages	**Disadvantages**
Temperate cost Low profile Easy installation and operation	Elastomer swelling with some fluids Low efficiency in deep wells

Can handle high sand and solid production Ability to handle high amount of free gas No valves thus no gas lock, clog Relatively high pump efficiency between 55-75% Low noise at surface Abrasion resistant Inexpensive power	Rotating rods cause tubing wear in directional and horizontal wells Rod fatigue and disconnection Maximum rate 5000 B/D Limited to shallow wells to depth 9840 ft (lower for large pumps) Low volumetric efficiency with high GOR Waxy crude paraffin problems Pump replacement requires tubing pull out High speed vibration problems Stator damage if pump run dry for short period Limited service and lack of experience in some areas

2.3 Artificial Lift Selection in Conventional and Unconventional wells

Several factors determine AL selection process: depth, rates, reservoir/fluid properties, initial and operating cost, and geographical and environmental aspects. Special AL selection techniques are required to cope with different reservoir, well and field conditions, for instance: high-viscosity oil, high water cut (WC%), sand, gas, low reservoir pressures, high temperatures, low-productivity wells, surface facilities, as well as human interference. Historically, the AL selection process is done qualitatively. It generally begins by studying the advantages and disadvantages of each method. Then the elimination depends on the engineers' decision based on their analysis of the AL record, field data availability, and failure history. Since these factors change over time, the AL

designing for current production conditions without considering future production results in high inconstancy rates and fluctuations in lifting selection (JPT staff, 2014; Lea and Nickens, 1999). The following sections show old and recent selection criteria from the literature, failure issues, and run life improvement in conventional and unconventional wells (conventionals and unconventionals), and different techniques used by the engineers and factors considered in each method.

2.3.1 Artificial Lift Selection in Conventional Wells

At the beginning of the 1980s, a panel of 4 members from oil companies summarized the selection criteria of four lifting methods, GL, BPU, ESP, and HP, by studying the significant merits and demerits concerning reservoir and well properties. They stated that BPUs are appropriate with low volumes but not in offshore, residential areas or wells with sand production history. CGL is suitable for high volumes, high BHP, solids and sand handling; however, back pressure and high cost are significant limitations. IGL cost is lower than CGL; on the other hand, it produces low volumes. ESP is used with high volumes and minimum spaces, such as offshore platforms, and can also handle a deviation of 80°. The major ESP drawbacks are sand production, cost of workover, and inefficiency with rates below 150 B/D. HPs (reciprocating and jet) are compatible with deep wells and can deliver up to 17000 B/D. Jet pumps are adequate with sand production due to no moving parts, whereas reciprocating pumps are efficient with highly viscous fluid; however, they have a shorter life than jets and submersibles because of the maintenance. Jet pumps cannot operate at BHP below 1000 psi, whereas reciprocating pumps can operate at 0 psi (Neely et al., 1981). The above criteria and results are approximately similar to the AL selection decision tree made by (Heinze et al., 1995), where 50% of the tree selection is PI and IPR dependant. Brown (1982) introduced a selection method according to advantages and disadvantages to assist engineers in AL selection. Blais (1986) drew selection charts to determine the operating ranges for AL methods. The charts had been used for a while as a selection reference during that time in addition to simple computer programs used as auxiliary tools.

It is worth mentioning that improper AL selection results in numerous replacements within a short period, decreasing the profit and maximising the operation cost. Clegg et al. (1993) introduced comprehensive reference selection

tables, evaluating eight methods: BPU, PCP, ESP, HP (reciprocating and jet), GL (CGL and IGL), and plunger lift, across 31 parameters. These tables serve as a widespread selection framework, permanently utilised by numerous researchers with minor adjustments and simulation applications, forming the basis for various modified AL selection methodologies. Bucaram and Patterson (1994) selection criterion considered wells location, capital expenditures (CAPEX) and operating expenditures (OPEX), production rates, run life and failure besides essential well and reservoir characteristics to be considered, depth, BHP, gas, sand, and solids. Another essential factor they considered was the latterly drilled wells in developed fields. The lifting method for any new well should match existing surface production facilities to avoid additional costs, such as installing new flowlines and wellhead fittings. Moreover, they provided an example of the selection process for BPUs and factors to consider, as shown in **Table 2.5**. It was apparent that BPUs were eliminated in gassy and deep wells. The critical selection factor is balancing AL reliability with desired production rate and present constraints to keep the pump running smoothly for an extra-long period.

Table 2.5: BPU selection guide (Bucaram and Patterson, 1994)

	Sand	Scale	Depth>7,000 ft	Intermittent Pumping	Corrosion	Large Volume	Low Fluid Level	Gas	Low Speed	Paraffin
Rod pump, traveling barrel, bottom hold down	Good	Good	Good	Better	No	Good	No	No	Good	Good
Rod pump, stationary barrel, bottom hold down	No	Good	Better	Good	No	Good	Good	Good	Good	Good

Rod pump, stationary barrel, top hold down	Better	Good	No	Good	Better	Good	Better	Better	Good	Good
Rod pump, three tubes	Better	Good	No	Better	No	No	No	No	No	Good
Stroke through	Better	Better		No		Better				
Tubing pump	Good	Good	No	Not applicable	Better	Better	Not applicable	No	Good	Good
Casing pump	No	No	No	Not applicable	No	Better	Not applicable	No	Good	No

2.3.1.1 Computer Programming and Nodal Analysis Application

Computer programming and simulators for AL selection started in the early 90s. Espin et al. (1994) developed a coding program to help engineers in selecting the adequate AL from 10 lifting methods by analysing field data which was divided into three categories: (1) quantitative data (well and reservoir props), (2) qualitative data (engineer experience and well geographic location), (3) production problems (corrosion, paraffin, sand and gases) and economic evaluation. The program ranked lifting methods from 1 (least recommended) to 5 (most recommended). Although some lifting methods had a high score, they were eliminated because they were not economically feasible, and lower-ranked ones were used instead. Lanier and Mahoney (2009) likewise applied a ranking matrix to evaluate six lifting methods: GL, ESP, BPU, PCP, Jet, and long-stroke pump. They illustrated technical and operational AL constraints, CAPEX and OPEX for thermally recovered heavy oil reservoir in Oman to obtain higher production rates.

It was evident that high temperature was a significant constraint for both GL and ESP. The costly gas supply and low GOR were not satisfactory with GL. High OPEX and CAPEX eliminated Jet and ESP. Metal-to-metal PCP (MTMPCP) had inexpensive operational costs; nonetheless, it was not a candidate for designed rate limitation (the same for the long-stroke pump), sand production, and failure history. Other reasons for eliminating the Jet were the high power required to lift fluid density of less than 14 API and the scarcity of pump history data in the field. Finally, all trials for new AL options were unsuccessful, and the primary used BPU continued production with attached sinker bars to reduce rod buckling. Williams et al. (2008) applied the same matrix screening to optimize five lifting methods; ESP, PCP, GL, Jet, and BPU used in a field in Colombia to confront common challenges; depth, gases, and solids, which impact each method. The selection criterion was narrowed by flow rates ranging from 0-750 B/D as an eliminating factor. They found that GL was suitable for all flow rates. PCP could produce up to 300 B/D, BPU and ESP were for rates between 300-750 B/D, while GL and ESP were for flowrates exceeding 750 B/D. Naguib et al. (2000) introduced a study in the Egyptian field to compare four AL methods; BPU, ESP, GL, and Jet. Reservoir simulation and well performance analysis were conducted to select the optimum method. BPU and Jet were eliminated because of high reservoir volume and wax content. The selection ended with GL for the availability of gas supply from a nearby company and ESP to control the flow rate with concern on high associated gas, although a downhole gas separator would be installed. Afterwards, further screening was carried out among the two candidates. GL CAPEX and OPEX were lower than the ESP and had a high recovery factor. In terms of high rates, WC% increment and insufficient gas supply, ESP outweighed GL.

Matondang et al. (2011) applied a different AL selection technique using combinations of GL mandrels and ESP, firstly to release the gas that caused many problems to the pump through the casing and secondly, to reduce the amount of WC%. The process was successful. The gas released through the mandrels passed with the gas bled off from the ESP gas separator, resulting in a production increment from 350 to 500 B/D and decreased water production. The technique opened the door to using ESP in high GOR wells; however, well completion determines this hybrid applicability. Zulkapli et al. (2014) evaluated ESP production in Bokor offshore field in Malaysia after replacing dual string GL due to

increased water production and insufficient gas supply. They applied nodal analysis using PIPESIM commercial simulator to simulate the performance. Despite the feasibility of GL, they found that the shortness in gas supply from a nearby field and compressors problems affected GL's efficiency; besides, the faults in real-time measurements impacted the optimisation process and led to miss-monitoring, which is a prevalent worldwide issue. ESP was selected based on the wells' low GOR and no sand production history. Alshmakhy et al. (2019, 2020) applied a new technology to optimise GL. They introduced digital optimisation for single and dual string GL in an onshore field in UAE to avoid common challenges: casing pressure instability, temperature fluctuations, and injection rate control. Digital intelligent AL (DIAL) system was implemented consisting of up to six injection orifices and an electric cable connected to the mandrels to control the opening and closing of GL orifices from the surface. In addition, the system provides real-time measurements of pressure and temperature. The estimation was an increment of 20% in oil production. This technology will probably be promising if implemented offshore, where the cost of workovers is much higher. Caicedo et al. (2015) performed Nodal Analysis screening to select AL for high uncertainty and large reserves field in Abu Dhabi in case of no NF. The primary issue was that the field contained H_2S and was near a residential area; therefore, the selection process was primarily determined by safety factors to avoid any leakage that could endanger human lives. With different values of GOR, WC% and reservoir pressure, the analysis showed that AL was required if the reservoir pressure was below 2500 psi, WC% reached 90%, and GOR below 3000 scf/STB. BPU and PCP were eliminated for the possibility of the stuffing box leakage. GL also was not an option for insufficient gas supply. Finally, ESP was selected with concerns of GOR not exceeding 1500 scf/STB. Valbuena et al. (2016) presented a methodology to select appropriate AL in horizontal gas wells by screening technical and economic factors. The technical screening studied the limitations of lifting methods regarding production rate, depth vs rate, reservoir/fluid properties, and gas handling by referring to standard selection tables and charts. After that, the net present value (NPV) calculation evaluated the feasibility of the lifting methods. In addition to the standard criteria applied in most oil and gas fields, they divided the selection factors into three categories: the weighting factor that represents its importance in the selection process rated from 1 to 10, the suitability factor is calculated by a mathematical equation and

the economic factor which is NPV. The procedure was applied in a field as a case study. Ultimately, they concluded that economic evaluation was the most crucial factor in determining the selection process. Kefford and Gaurav (2016) evaluated well performance for several lifting methods using adjusted correlations and iterative calculation (**Fig. 2.13**). They studied specific reservoir characteristics and operation factors of three fields, including unconventional reservoir, to estimate production rates and AL's capability to handle associated gas. They aimed to widen the selection method and provide a new criterion instead of the standard Blais method. They used modelling software to calculate well performance and then decided which AL could deliver both max and targeted rate considering gas handling, head, and power required. Alferov et al. (2015) and Khabibullin and Krasnov (2015) studied the effect of change of the following parameters: reservoir pressure, BHP, WC%, PI, GOR and flowline pressure on the CAPEX and OPEX of AL methods in Russian fields by developing new selection algorithms (computer coding).

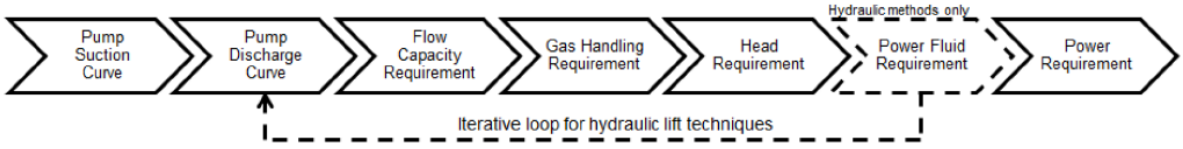


Fig. 2.13: Iterative calculation process (Kefford and Gaurav, 2016)

Alferov et al. (2015) argued that conventional selection technical tables are impractical because they are chiefly based on AL operation history, and the problems count to each method. The case studied field implementing simultaneous water alternating gas (WAG) in a low permeable and heterogeneous reservoir with paraffin, salt, and corrosion. BPU, PCP and Jet were eliminated due to insufficient equipment supply. The best candidates and cost-effective AL for the field development plan (FDP) were ESP and GL because of their ability to handle the change in WC% and GOR, respectively. Khabibullin and Krasnov (2015) AL selection map for a new field showed that ESP and GL have the same results with 100 ATM BHP that lowered their applicability where 40 ATM was preferred.

2.3.1.2 Other Applications in Artificial Lift Selection

Fraga et al. (2020) presented a newly combined pump of PCP and ESP systems known as progressive vortex pump (PVP). The PVP was developed by Petrobras to optimise production and confront the high temperature of cyclic steam stimulation (CSS) and steam flooding (SF) as well as producing different flow rates. The pump consists of a rotor, stator, and diffuser with many stages to convert the kinetic energy into potential energy. PVP efficiency was 50% lower than ESP and can reach 33%; nevertheless, it handled extra head than the ESP. The performance test of PVP found that one stage at 60Hz could give a head of 75.5, equivalent to 32.7 psi. The pump was installed for a pilot test onshore. After 4 months of operation, the pump had a positive efficiency of 6-8%, measured by the difference between consumed and performed power.

A different selection strategy was introduced by (Kaplan and Duygu, 2014) in a Turkish heavy oil field where CO_2 injection was implemented to increase oil recovery. They analysed the axial and radial shear stress and the torque on the power required for two AL methods, BPU and PCP. Due to the high temperature, ESP was not an option. BPU had been producing, though, emulsion and high viscosity caused rod failure and limited the amount of produced oil. The power handling radial shear stress and torque to move the oil for PCP was less than the power required for the axial shear stress for BPU. This power reduction could be achieved by reducing the rpm and using a large pump size. Thus, PCP was selected and replaced the BPU in the field. Mali and Al-Jasmi (2014) applied a selection screening for CHOPS and CSS thermal recovery in a Kuwaiti oil field. The FDP targeted a maximum of 300 B/D cold oil and 1000 B/D hot oil for 12 API density, low GOR, and a depth reaching 3000 ft. Five AL candidates BPU, ESP, electrical submersible PCP (ESPCP), PCP and MTMPCP, were selected for the screening. **Table 2.6** illustrates the selection criteria, and in the end, the decision was for both PCP and MTMPCP for their lower CAPEX and OPEX.

Table 2.6: AL comparison for heavy oil production (Mali and Al-Jasmi, 2014)

Parameters	BPU	ESP	PCP	Jet	ESPCP	Gas Lift
Capital Cost	Low	High	Low	High	Moderate	High
Operating Cost	Low	Moderate	Low	High	Moderate	Moderate
Run life in vertical wells	Average	Average	Average	High	Average	High
Run life in horizontal wells	Low	Average	Low	High	Average	High
Ability to handle sand content	Average	Low	Average	Good	Average	Average
Efficiency	Average	Low	Average	Low	Average	Average
Suitability for thermal production	Applicable	Applicable	Applicable	Applicable	Not Applicable	Applicable
Operational Flexibility	Average	Good	Good	Low	Average	Good
Ability to handle gas content	Average	Good	Good	Good	Good	Good
Production Handling Capacity	Good	Average	Good	Average	Average	Good

Hoy et al. (2020) assessed current used lifting methods to select new AL to a polymer EOR application in an Austrian oil field. They studied the effect of change in viscosity and head column on GL, BPU, ESP, and PCP to check their reliability in delivering the desired flow rate. Their results showed that ESP and BPU were the optimum lifting method. Nonetheless, ESP could handle fluid head; it could not cope with 500 ppm polymer concentration, while BPU showed some friction problems. Zein El Din Shoukry et al. (2020) provided a series of parameters to consider for optimum AL selection, as shown in **Table 2.7**, to gain prolonged run life and high revenue.

Table 2.7: AL selection parameters (Zein El Din Shoukry et al., 2020)

AL	Gas Lift	Foam Lift	Plunger	BPU	PCP	ESP	HJP	HPP
Max Depth	18,000 ft	22,000 ft	19,000 ft	16,000 ft	<9,000 ft	15,000 ft	20,000 ft	17,000 ft
Max Volume	75,000 B/D	500 B/D	200 B/D	6,000 B/D	5,000 B/D	60,000 B/D	35000 B/D	8,000 B/D
Max Temp	450˚F	400˚F	550˚F	550˚F	302˚F	482˚F	550˚F	550˚F
Corrosion Handling	Good to excellent	Excellent	Excellent	Good to excellent	Good	Good	Excellent	Good
Gas Handling	Excellent	Excellent	Excellent	Fair to good	Good	Fair	Good	Fair
Solids Handling	Good	Good	Fair	Fair to good	Excellent	Sand<40ppm	Good	Fair
Fluid Gravity (°API)	>15°	>8°	>15°	>8°	8°<API<45°	Viscosity< 400 cp	≥6°	>8°
Servicing	Wireline or workover rig	Capillary unit	Wellhead catcher or wireline	Workover or pulling rig	Wireline or workover rig	Wireline or workover rig	Hydraulic or wireline	Hydraulic or wireline
Prime Mover	Compressor	Well natural energy	Well natural energy	Gas or electric	Gas or electric	Electric	Gas or electric	Gas or electric
Offshore	Excellent	Good	N/A	Limited	Good	Excellent	Excellent	Good
System Efficiency	10% to 30%	N/A	N/A	45% to 60%	55% to 75%	35% to 60%	10% to 30%	45% to 55%

Another interesting study was presented by (Crnogorac et al. 2020) to select the optimum AL using fuzzy logic and mathematical models. The model is conditioned to an enclosed data inventory of 5 lifting methods and might not be applicable if different input parameters or other ALs are used instead. The AL of a new well is selected based on the one that best matched to AL database. Adam et al. (2022) introduced a recent selection method for Sudanese oil fields. Their decision-making model modified the methodology of (Alemi et al. 2010) using TOPSIS (Technique for Order Preference by Similarity to Ideal Solution) by adding AHP (Analytic Hierarchy Process) for parameter weighting to obtain the best decision. According to the designed flow rate and other parameters, the model ranked the appropriate AL. The promising results could be more robust if economic evaluations were considered.

2.3.2 Artificial Lift Selection in Unconventional Wells

In unconventional wells (unconventionals), the principal used ALs are ESP, GL, BPU, Jets, and plunger lift (**Table 2.8**). Usually, AL is installed either after the well's NF drops or directly at the beginning of production (Chow et al., 2020). The average run life of ESP is 6-9 months. BPU's main issue is that the side load does not exceed 200 lbf/25 ft; if so, another AL should be considered. GL is used at a deviation up to 75°. Jets can handle solids due to no moving parts. Plunger lift is used for low volumes around 200 STB/D (Kolawole et al., 2019; Pankaj et al., 2018).

Table 2.8: AL used in unconventionals (Kolawole et al., 2019)

AL	Percentage of application
GL	40%
ESP	36%
BPU	13%
Jets	4%
Plunger	7%

The rapid production decline in unconventionals in a few years and sometimes a few months is a massive challenge that requires AL replacement and added OPEX. Casing size is a crucial factor in AL design and selection in unconventionals. The larger the casing, the higher the gas produced through the annulus to the surface, which affects AL performance (Parshall, 2013). Recently, engineers developed new improvements in AL for unconventionals. ESP permanent magnet motor and stages design, GL controlled valves, tailpipe design to handle slug, and the new geared centrifugal pump (GCP) that is the same as the ESP, despite the rod being driven from the surface by hydraulic and electric power which is considered adequate with gas than the conventional ESP (Parshall, 2013; Stephenson, 2020). A field case study to select an AL capable of coping with the rapid decline in production considering reservoir/fluid properties and well performance analysis was presented by (Oyewole, 2016). They divided the selection factors into four categories: (1) technical, which contains production rates and associated produced gas (used to determine depletion period), (2) reservoir/fluid properties and drilling conditions, (3) surface facilities and (4) economic evaluation. Similar to (Valbuena et al., 2016), the economic aspect is the critical selection factor in

unconventionals regardless of any recommended methods. Liu and Zerpa (2016) calculated the CAPEX of AL (**Table 2.9**) to find a suitable method for a hydrate reservoir in Alaska with low pressure, low GLR, low reservoir and surface temperature, and sand production. According to the author, PCP was the suitable candidate; nevertheless, it could not handle sand production, and failure occurred shortly. High CAPEX was an issue for ESP, and low GLR might eliminate GL. In addition to CAPEX, (Khan et al., 2014) selection strategy included workover cost, OPEX, oil price, oil treatment and transportation along with maximum NPV to evaluate four AL methods; GL, ESP, ESPCP, BPU for Shale play horizontal wells. They also studied NF conditions and the interval required to move to another lifting method. Their results showed that using ESP followed by BPU after two years was more profitable than using single or three lifting methods. One lifting method had lower efficiency, while three methods increased the CAPEX of production.

Table 2.9: Summary of AL method feasibility for Hydrate reservoir (Liu and Zerpa, 2016)

	ESP	**PCP**	**Beam**	**Hydraulic**	**GL**	**Plunger**	**Compress**	**Foam**	**Vel String**
Shallow depth	Well suited	Well suited	Well suited	Well suited	Well suited	V.well suited	Well suited	Well suited	V.well suited
Offshore	Maybe	Maybe	Poorly suited	Poorly suited	V.well suited	Well suited	Maybe	Maybe	V.well suited
Permafrost	Well suited	Well suited	Well suited	Well suited	Well suited	V.well suited	Well suited	Well suited	Well suited
Low production	Poorly suited	Maybe	Well suited	Maybe	Maybe	V.well suited	V.well suited	V.well suited	Maybe
Low GLR	Well suited	Well suited	Well suited	Well suited	Poorly suited	Poorly suited	V.poorly suited	V.Poorly suited	Poorly suited
Low BHP	Maybe	Maybe	V.well suited	Maybe	Poorly suited	Well suited	Well suited	Well suited	Maybe
Viscous production	Poorly suited	Maybe	Poorly suited	Poorly suited	Poorly suited	V.Poorly suited	Poorly suited	V.Poorly suited	V.Poorly suited
Sandy production	Poorly suited	Maybe	Poorly suited	Poorly suited	Maybe	V.poorly suited	Poorly suited	Well suited	Well suited
Secondary hydrate	Poorly suited	Maybe	Poorly suited	Poorly suited	Maybe	V.poorly suited	Poorly suited	Well suited	Well suited
Ice	Poorly suited	Maybe	Poorly suited	Poorly suited	Maybe	V.poorly suited	Poorly suited	Well suited	Well suited
Slow pressure build up	Well suited	Well suited	Well suited	Well suited	Maybe	Poorly suited	Maybe	Maybe	Poorly suited
Low reservoir temperature	Well suited	V.well suited	Well suited	Well suited	Well suited	Well suited	Well suited	Well suited	Well suited
CAPEX	115,000	35,000	45,000	45,000	25,000	10,000	20,000	7,500	10,000

Pankaj et al. (2018) analysed reservoir properties such as porosity, permeability and saturation, geological structures, GORs, and various production rates 2500-500 STB/D using simulators to find the appropriate ALs that meet the requirements in deep horizontal shale wells. Their results showed that GL and Jets were the best candidates, though Jets could not operate with low rates. Similarly, a simulator was used by (Escobar Patron et al., 2018) to select the optimum ALs in unconventionals in the USA to cope with production decline challenges. The software analysed input parameters such as depth, reservoir and fluid properties and solids contents to eliminate AL methods based on known constraints. Then Nodal Analysis and NPV calculations for forecasting to showcase which AL delivered the required rate at the lowest expenses. Three production periods, 1, 3 and 6 years, were simulated. The screening results showed that the well could naturally flow for 3 months in the three scenarios, then ESP and Jet were nominated for higher flow rates, followed by BPU after production declined. Chow et al. (2020) developed a selection tool for the most feasible lifting method for offshore unconventionals. Likewise, well and fluid properties analysis was performed, and well performance analysis was validated separately for the selected method. Pump feasibility was determined in 3 steps; (1) the ability of the pump to handle GVF, (2) pump and casing size fitting each other, and (3) reservoir deliverability to pump size. A series of calculations and plots were then used to validate the above three steps, followed by company specifications for more filtration regarding temperature, pressure and rates. Following that, several questions named well-integrity, a comparison between qualitative and quantitative well parameters that limit AL application. Lastly, the applicability of each lifting method was ranked. Lane and Chokshi (2016) and Temizel et al. (2020) summarised the use of AL in unconventionals into four production stages (**Fig. 2.14**):

1. High rates Jet pump is used for cleaning operations to remove hydraulic fracturing fluids and continues until production declines.
2. GL, plunger, and foam lift are to handle gas slug flow.
3. GL, Jet, and ESP are used in early production, and an amalgamation of GL and Jet/foam could be applied depending on completion.
4. BPU is used in the later production period after the decline occurs.

Additional selection considerations recommended by the authors are:

1. Integrated planning for well completion, considering several future AL.
2. AL life cycle estimation to reduce workover cost.
3. Continuous well parameters surveillance for production optimisation.

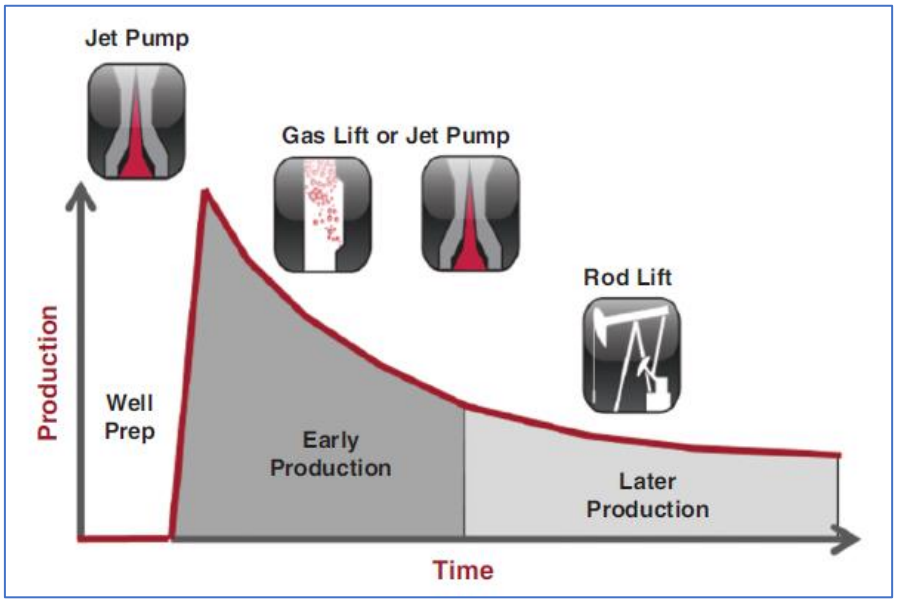


Fig. 2.14: AL life stages in unconventionals (Lane and Chokshi, 2016)

Here are some worldwide AL solution examples of overcoming the challenges in unconventionals (Kolawole et al., 2019):

In a Canadian oil field, new BPU and rod designs increased the run life of 25 pumps by 75% and the production rate of 14 wells up to 90%. Moreover, Cunningham Modified Model and Total Well Management simulators were used to investigate Plunger and Jets problems. Results recommended pump resizing to confront the effect of the gas. The plunger failed above 74° deviation.

In the USA, dual ESP stage and packer-modified gas separators were installed to solve gas slugging, resulting in more than 100% production increment and reduced failure rates. Another developed gas separator increased oil production to 224 B/D in 47 wells in a Texan field. Artificial sump pumping (ASP) was introduced to replace ESP, reduce pump failure, and increase production. The new AL obtained a rate of 220 B/D with an operation period of more than 332 days compared to ESP production of 130 B/D and 157 days lifecycle. A new ESPCP was installed to reduce sand production and pump failure in California, resulting in

fewer failures, low operation cost and a 50 B/D production rate for a more extended period.

In Oman, a new PCP design consists of an anchor that can handle two strings: production and other for intervention. The pump reduced WC% from 100 to 65%, increased oil production from 1 to 32 m^3 and lowered workover expenditures.

In China, a new ESP with a different pump and motor, known as electrical submersible reciprocating pumping (ESRP), increased the rate from 35 to 66 B/D with an efficiency of 65.1%.

In Kuwait, a permanent magnet motor-hydraulically regulated-progressive cavity pump (PMM-HR-PCP) lately designed to solve conventional BPU and PCP problems in a sandy heavy oil field. The application results showed a 20% increase in oil rate with extra run life.

2.4 Artificial Lift Failure and Run Life in Conventional and Unconventional Wells

Pump failure and run life are significant factors in the AL selection process. They are determined from the record of occurrence and running cycles concerning field conditions, fluid, and reservoir properties. Many researchers studied pump failure to find the root cause and provide remedies for extra AL life cycles. Here are examples from the literature.

2.4.1 Failure Prevention and New Designs

Bucaram and Patterson (1994) studied the history of AL failure to prevent future occurrence. They built a system for failure trailing consisting of (1) failure type (tubing, rod, pump), (2) failure location (barrel, plunger), and (3) failure cause (corrosion, sand, rod cut). The data from this system was gathered to analyse the production system and performance of AL. Yang et al. (2011) deployed anti-scaling AL techniques in a Chinese field to prevent PCPs and BPUs failure and extend pumps' lives. Alkaline surfactant polymer EOR was implemented, which led to scale problems. The scaling accumulation caused pump and rod failure resulting in pump-stuck and rod disconnection. To mitigate this issue, they evolved PCP and BPU designs. For PCPs, they used ceramic coating on rod string and intensified

elastomer hardness to lower the effect of scaling. For BPUs, they shortened the piston's length; therefore, pump length will be two times longer than the usual design. In addition, they added chemicals to remove any scale accumulation. The application extended pumps' lives from 47 days PCPs and less than 30 days BPUs to one year, saving huge workover costs. Another study by (Ghareeb et al., 2012) assessed the failures of five AL methods producing oil in the Egyptian field; ESP, BPU, PCP, GL, and Jet. For ESPs, scale inhibitors, sand screens and paraffin solvents were deployed to prevent pump plugging. For BPUs, the manufacturer was contacted to improve the pump design to reduce rod bending (buckling), and the number of strokes per minute (SPM) was also reduced to defy fluid pounding to avoid dry pump running. Regarding PCPs, rod string and elastomer were replaced by newly designed materials along with downhole real-time measurements and frequency control. The process reduced failures and increased running life from 90 days up to 2 years, saving 1 million USD in workover cost. GL was used offshore, and its main issue was high water cut, the paper provided no solution. Due to difficulties in monitoring BHP and high fluid surface requirements, Jets are often replaced by BPUs. Zhongxian et al. (2015) presented a case study of BPU and PCP in a Chinese field with polymer flooding. The high viscous oil and polymer resulted in high axial and radial shear stress, which shortened the AL life cycle to 2-3 months. Higher rod buckling rates, rod-tubing friction, rod disconnection, and stator damage pushed the engineers to find a solution to increase the pumps running life. A new BPU design was introduced, named low-friction pump, by adding many grooves inside the barrel to reduce the friction between the barrel and the plunger. The solution was successful; the run life of 235 wells increased to approximately one year, and pump efficiency jumped to 60%. Regarding PCP, a new rod design and elastomer alignment to reduce break and friction and a small pump were used, resulting in 2 years life cycle and low power and torque required. According to Dave and Mustafa (2017), a solution for rod buckling and fluid pound is using a small pump, long stroke and reducing the SPM, which will provide a longer pump life and better performance. A field in Oman suffered 40% BPU and PCP failure after switching from water to polymer flooding to increase oil recovery in a heavy oil sandstone reservoir. Al-Sidairi et al. (2018) applied series of trials to cope with pump failure caused by sand, scale, wear, and corrosion resulting from CO_2 and H_2S. Low frequency and long stroke pump were applied to solve erosion problems. Continuous rod string with

centralisers was applied to reduce buckling, providing extra five months of the pump running. Regarding corrosion, pump coating was successful in some wells; moreover, sand control mesh was used, which drove pumps to produce for more than 12 months. Alsiemat and Gambier (2016) applied a new ESP completion design to extend the production period and trim workovers. Dual ESP completion, known as rigless-deployed ESP, was used to keep production through if one pump fails, the second continues until a workover takes place. As the author mentioned, the process reduced the downtime from months to hours. On the contrary, Scarsdale et al. (2019) thought that through-tubing ESP (TTESP) is more reliable than dual ESP. Because the backup ESP faces the same conditions that cause the failure of the primary ESP, it will probably not provide the required efficiency. Thus, its purpose is to minimise production loss until a workover occurs. Mesbah et al. (2018) implemented AL alternating strategy for a heavy oil field with CSS recovery in Kuwait to solve pump failure and save 500-750 BBL. Two AL methods were used, BPU and MTMPCP. The presence of sand caused many problems to both pumps that required workover operations to wash the sand and replace the failed pump within a few months after each CSS cycle. Also, it resulted in a further decrease in the temperature and an increase in the viscosity. Moreover, installing MTMPCP after the steam cycle was unsuccessful because of the dry pump run. Thus, a new strategy was presented to reduce workovers cost and downtime by starting production using BPU following the CSS cycle, and after the viscosity increased, PCP was installed until the next cycle. Although the process was cost-effective, it resulted in more prolonged shutdowns because of rig arrangements. Early ESP failure in unconventionals due to gas slugging, erosion, and high temperature drove (Chachula et al., 2019) to design a new rotary gear pump (positive displacement pump speed dependant, not pressure dependant). It was an ESP body without a pump that could overcome those challenges. The pump had low speed and high production capacity, with discharge pressure reaching 4000 psi for one stage. The pump was installed as a field trial and found to have lower efficiency than standard ESP due to backpressure resulting from gears lubrication; however, it can be higher by increasing operating frequency.

2.4.2 Failure Analysis

Lapi et al. (2014) applied root cause failure analysis (RCFA), as shown in **Fig. 2.15**, for ESP and PCP oil production in Chad. The criterion started by collecting data for each lifting method. Reservoir, well, and operation parameters were input into a failure analysis workbook. Then the failed pump was dismantled in a specific workshop for further inspection to determine the root cause of failure. Ultimately, company staff meetings were scheduled to discuss and review the outcome of the RCFA. The application of RCFA from 2007 to 2013 resulted in 70% and 50% failure reduction in 615 ESP and 210 PCP, respectively, as shown in **Fig 2.16**.

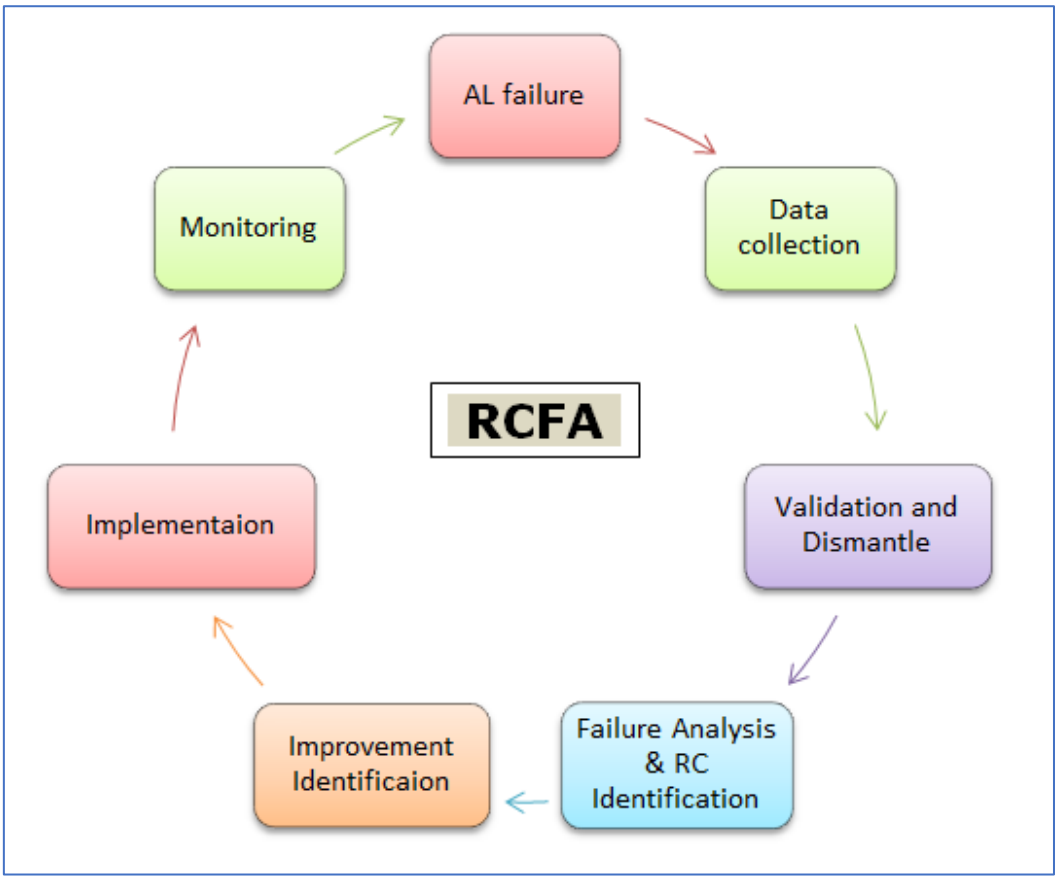


Fig. 2.15: RCFA process (Lapi et al., 2014)

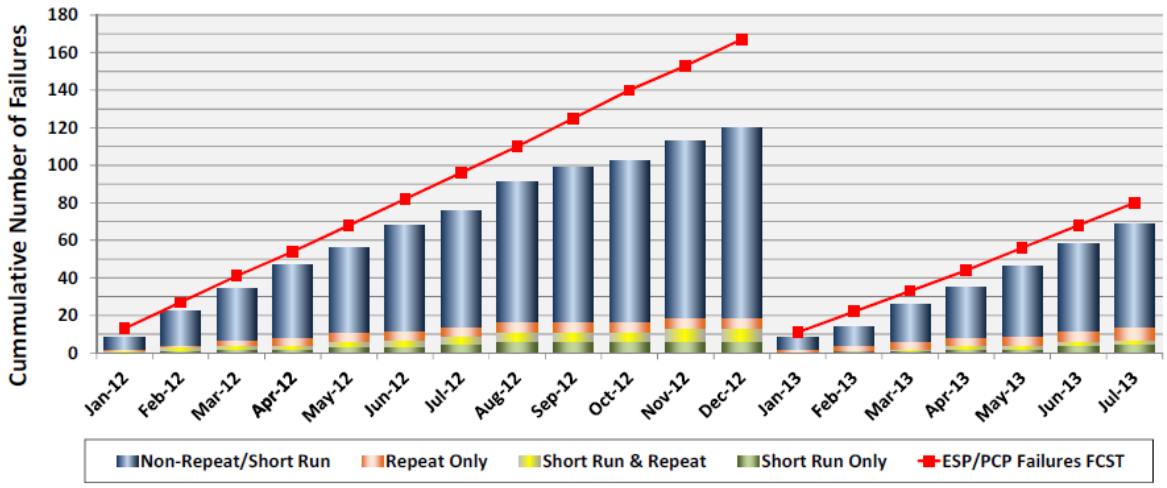


Fig. 2.16: AL failures before and after RCFA (Lapi et al., 2014)

Similarly, RCFA was applied in the La Caira field in Colombia for four lifting methods; BPU, ESP, PCP and ESPCP (Rubiano et al., 2015). After AL failed and pulled out of the well, the data was gathered for failure analysis, as demonstrated in **Fig. 2.17**. AL was then dismantled for inspection, and a detailed report was prepared for discussion by clients and vendors to find AL failure root cause, solutions, and future management plan for performance evaluation. The failure was classified into three categories: (1) AL failure, failure in AL components, (2) failure non-AL, failure in tubing due to sand or paraffin plugging, and (3) no failure; AL removed for high water cut, well abandon or low productivity. Rubiano et al. (2015) also applied key performance indicator formulas (KPI) for performance evaluation by calculating the failure index, pulling index, recurrence index, average run time and average run life. The implementation of RCFA in the field for two years resulted in a total common (called controllable) failure reduction from 161 in 2012 to 92 in 2014 and remarkable improvements in critical wells that had more than one failure per year.

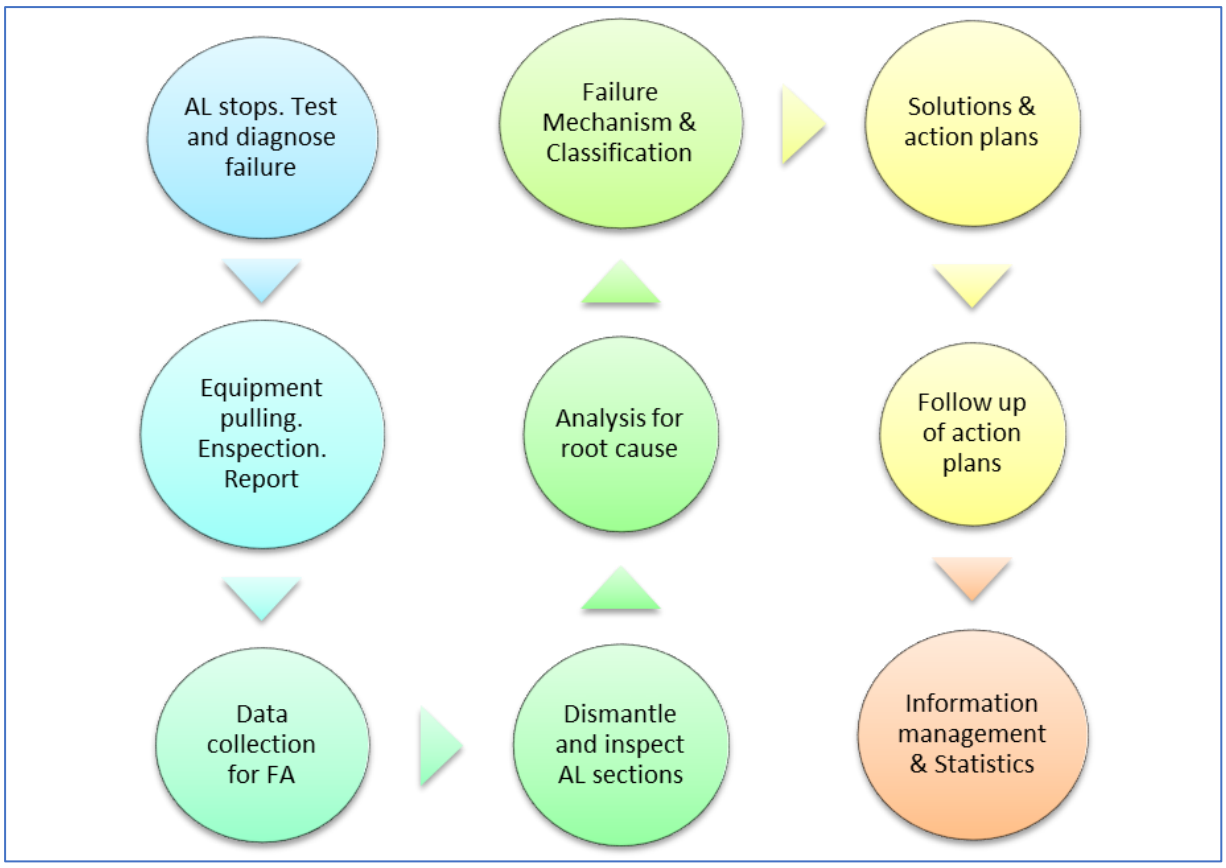


Fig. 2.17: Failure analysis process (Rubiano et al., 2015)

2.5 Artificial Lift Run Life Improvements

Stephenson (2019), a senior engineering advisor, said that extending AL run life should start before the installation by discussing the AL requirements and specifications with the manufacturers for proper designs and following international standards such as the API and ISO. Following the proper installation and operation, the key factors are monitoring, training personnel, and applying RCFA. Phelps (2015) introduced Baker Hughes criterion to reduce ESP damage due to chemical injection in the CO2 EOR field in North America. The approach was to balance the impact of chemicals on AL and the desired production rate considering environmental and safety aspects. Capillary tubes were designed to inject asphaltene and scale inhibitors downhole. The application in 5 wells resulted in a 100,000 B/D annual production increment, 133% extra ESP running time and a reduction of 80% in workover cost over five years. Caballero et al. (2014) introduced a new designed pump to handle the free gas in the Orinoco field in South America. Hydraulically regulated PCP (HRPCP), a special PCP which can

handle up to 40% of free gas, was used to replace the conventional PCP. The gas increased the temperature, which damaged the elastomer and caused pump failure. The HRPCP (**Fig. 2.18**) was designed with a new rod and elastomer cavities to reduce gas compression and temperature effect on the elastomer. The HRPCP was installed in 2 wells with high GOR and viscosity. The run life of the first well increased from 28 to 622 days, with a mean time between failure (MTBF) of 166 days. The second well went from 60 to more than 800 days with an MTBF of 96 days and more than 3 million USD OPEX saving for both wells, plus increased pump efficiency from 30 to 40%.

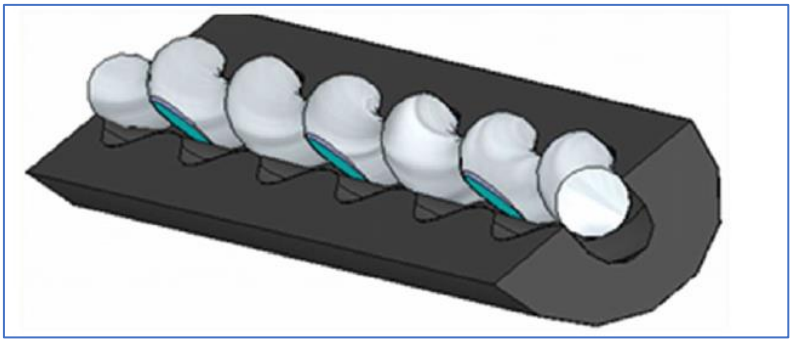


Fig. 2.18: HRPCP rod and elastomer cavities (Caballero et al., 2014)

Ramdé et al. (2014) applied a series of experiments and numerical simulations on pump materials and fluids to design MTMPCP can last longer with thermal recovery and solid contents. Experiments were fatigue and corrosion tests. The corrosion test was carried out on H2S and CO2 solution at 200°C. The numerical simulation analysis was applied to measure stress and strain using Computational Fluid Dynamics (CFD) and Fluid-Structure Interaction (FSI). Then PCP tracking software was used to calculate MTBF for previous and new designs. The modified design's run life was nearly 300 days, whereas the old design's run life was 160 days which was a remarkable result. Lastra (2017) presented valuable discussion for extending ESP life up to 10 years by understanding and improving three concepts reliability, maintainability, and availability. Reliability refers to pump performance, maintainability means restoring system efficiency after failure, and availability is a ratio of reliability to maintainability. The author mentioned that the pump's life could be improved using a dual ESP system, reducing premature pump failure caused by humans and applying preventive and predictive maintenance such as condition-based maintenance, which is real-time failure assessment monitoring, as well as developing new designs and technologies. Skoczylas et al. (2018)

highlighted optimum run-life measurements for AL reliability as: mean time to failure (MTTF), which is the actual running period, and mean time to pull (MTTP), which is the time between workovers. Kadio-Morokro et al. (2017) published a case study for extending ESP lifespan that suffered from high GOR, low production and pump intake pressure (PIP) in the unconventional Permian basin. Three gas handling techniques were applied in newly drilled wells; (1) tapered system with gas handling stages, (2) tapered with multi-vane pump, and (3) encapsulated production system. The result was a more extended ESP running period than the average of other pumps. Castillo et al. (2018) and Khadav et al. (2018) performed several applications to reduce PCP failures in deviated wells in the Bhagyam field in India and the Yaguara field in Colombia. RCFA and predictive analysis found that rod tubing wear is the primary cause of lifting failure. Khadav et al. (2018) used rod centralisers to mitigate the problem along with hollowed rod string for hot water flushing, which increased the average pump life by 27%. Castillo et al. (2018) used hollowed rods for axial load distribution, which reduced the stress by 80%, reduced OPEX and increased run life. Khadav et al. (2018) new completions with/without packer have some benefits and drawbacks regarding the gas and cost. The produced gas through the annulus reduced pump efficiency, while inside (insert) tubing PCP completion decreased workover cost and downtime by refraining from pulling the tubing out. A simulation of freshly designed tubing (Boronized tubing) to handle wear, friction, and large volume PCP for high flow rate and pressure showed remarkable results for future development plans. To increase BPU's run life in the Matzen field in Austria, OMV Company carried out RCFA to reduce the impact of high WC%, corrosion, wear, gas and sand on production and OPEX. Firstly, the equipment was pulled out of the hole for inspection, followed by a detailed failure analysis report, as demonstrated in **Fig. 2.19**. For further mitigation, the company decided to increase the quality of pump equipment and tubing by modifying the material design. Two KPIs equations were also used for monitoring the performance of modified and conventional BPU: (1) MTBF and (2) failure recurrence index (OMV FRI). Due to the low oil price, break-even economic analysis showed that newly designed pumps would result in a 10% increased cost and 50% extra run life (Hoy et al., 2018).

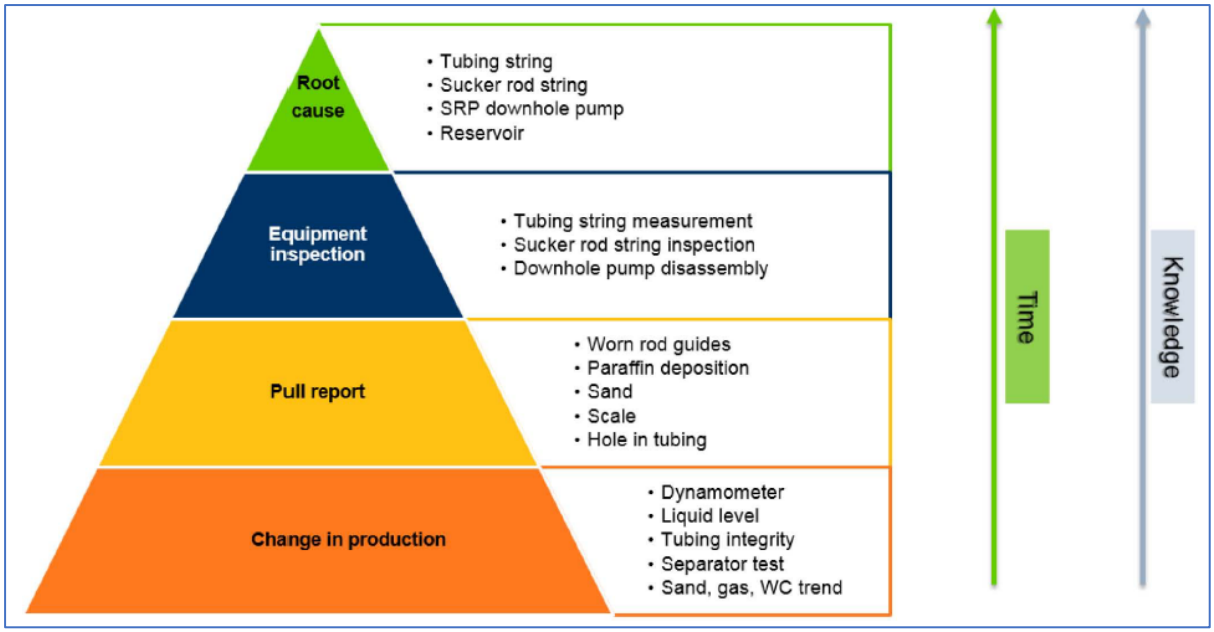


Fig. 2.19: Failure analysis based on available information (Hoy et al., 2018)

$$\text{MTBF} = \frac{\text{Operating wells} * \text{reporting periods (days)}}{\text{No. of failures}} \qquad \text{Eq (2.1) (Hoy et al., 2018)}$$

$$\text{FRI} = \frac{\sum \text{Equipment failures}}{\sum \text{ALS subsystem failures}} \qquad \text{Eq (2.2) (Hoy et al., 2018)}$$

Almajid et al. (2019) introduced an optimisation model for ESP performance and life cycle enhancement consisted of 5 steps:

1. Data gathering and analysis for surface and downhole parameters.
2. ESP design and robust manufacturing to chiefly handle the gas and sand.
3. Operation performance and production enhancements by widening the operating ranges with variable speeds and voltages.
4. Real-time monitoring by transmitting sensors data to SCADA and then using software to alarm any faults.
5. RCFA by dismantling and inspection.

The proposed model was applied in 2 wells and resulted in pump performance increment, fluid production with minimum gas and solved gas locking issue, as well as improvements in ESP running period. Harris et al. (2019) mentioned that ESP cable accounts for over 20% of failures. Their tests of newly designed cable to increase ESP lifespan showed that it could operate for up to 20 years; however, it is more expensive.

2.6 Machine Learning Algorithms

ML is a form of artificial intelligence (AI) that mimics human behaviour and has been employed to address problems. ML excels in dealing with big data analysis, even outperforming humans in some cases (Syed et al., 2020; Osisanwo et al., 2017). The primary function of ML is to support petroleum engineers in interpreting data and making informed decisions in a timely manner. Supervised and unsupervised learning are the two primary categories of ML. Supervised learning entails the algorithm learning from labelled input and output data, but obtaining such data can be challenging. On the other hand, unsupervised learning is employed when the data are not labelled, and the algorithm attempts to extract patterns from the dataset (Mohamed, 2017; Ahmed et al., 2020). Five supervised learning algorithms, including LR, the commonly used in OGI SVM and DT (Noshi and Schubert 2018), KNN, and the powerful RF, as well as K-means for unsupervised learning, were employed in the modelling of AL selection. Algorithms classification criteria are presented in more details in Chapter 3.

2.6.1 Supervised Learning Algorithms

2.6.1.1 Logistic Regression

LR is a discriminative, probabilistic classifier. It learns to find the difference between classes by focusing on a few signs, even without knowing more information. LR is much more robust in feature correlation, especially in the larger dataset than the smaller one (Jurafsky and Martin, 2021). The LR classifier function determines the boundary between the classes and then measures the distance of each class to the boundary for classification (Osisanwo et al., 2017). Logistic regression uses the sigmoid function (a mathematical function used to add non-linearity to ML model for classification) to map the values between 0 (false class) and 1 (true class) (Gianey and Choudhary, 2018).

2.6.1.2 Support Vector Machines

SVM classifies linear and non-linear separable data (Fletcher, 2009) and efficiently performs model training. In linear, SVM uses hyperplanes to separate the data and then calculates the margins between the hyperplanes and the nearest point (Mohamed, 2017). Also, SVM maximises the margins between the hyperplanes to reduce the generalisation error (Osisanwo et al., 2017). In non-linear, SVM uses a set of kernel functions that can be recast into a space with a higher dimension.

2.6.1.3 K Nearest Neighbours

KNN is a simple nonparametric supervised learning algorithm used primarily for classification. It uses the Euclidean distance (distance of each class to all data points in the dataset) to find the nearest class according to a given number of classes (K). For example, if K=3, the selected target will be the most frequent amongst the three neighbours. The algorithm can be run several times using different K values until an accurate output is obtained. KNN is fast, has low computational time, and can handle noisy data (Taunk, 2019; Mohamed, 2017; Patrick and Fischer, 1970).

2.6.1.4 Decision Tree

The DT classifier is a hierarchical algorithm that decides the branch to which each class should belong (Mohamed, 2017). It is easy to use and can handle outliers and missing values. DT classification starts from the tree's roots upwards according to the values of the features (Osisanwo et al., 2017). Furthermore, it uses impurity measurements such as entropy and Gini index to establish data classes. If the impurity value is close to zero, the samples are homogeneous to be classified (Gianey and Choudhary, 2018).

2.6.1.5 Random Forest

The ensemble RF is an upgraded DT or multi-DT that uses random sample features to split the data. In contrast, DT selects the optimum point to split the data, resulting in many similar tree structures. The forest correlates the subtrees' results and then minimises the trees' similar structure error for accurate results (Brownlee, 2016). RF can effectively handle overfitting and provide excellent results within a short time compared to other algorithms such as DT and SVM (Parmar et al., 2018).

2.6.2 Unsupervised Learning Algorithm K-Means

K-means is a commonly used clustering algorithm in unsupervised classification, which involves partitioning a dataset with n items into K clusters determined by the user. The algorithm performs classification in two steps: assigning each parameter to the nearest cluster centre and adjusting the cluster centroids to be the mean of its features. Although K-means is considered a powerful clustering algorithm, its main drawback is the random establishment of centroids, which can

result in unexpected convergence. Furthermore, the algorithm requires the pre-definition of the number of clusters, which can lead to the effect of outliers. Clustering algorithms utilize data distribution to define guidelines for partitioning data into groups with similar attributes. The ideal clustering process is when each cluster contains similar data that is distinct from the data in other clusters. The K-means algorithm is dependent on the value of k, which must be specified for clustering, and different k values can produce different outputs (Ahmed et al., 2020; Kharrat et al., 2009). The algorithm can also be used to cluster categorical features. Some weaknesses of the k-means algorithm are that the number of k must be determined, the real clusters number is difficult to obtain, and features importance is challenging to determine (Teknomo, 2006).

2.7 Machine Learning Applications in Oil and Gas and Artificial Lift

2.7.1 Machine Learning Application in Oil and Gas

Since the last decade, ML application in the OGI has continuously risen in many sectors. **Fig. 2.20** demonstrates the surge of using ML in oil and gas research in recent years from Google Scholar (Pandey et al., 2020).

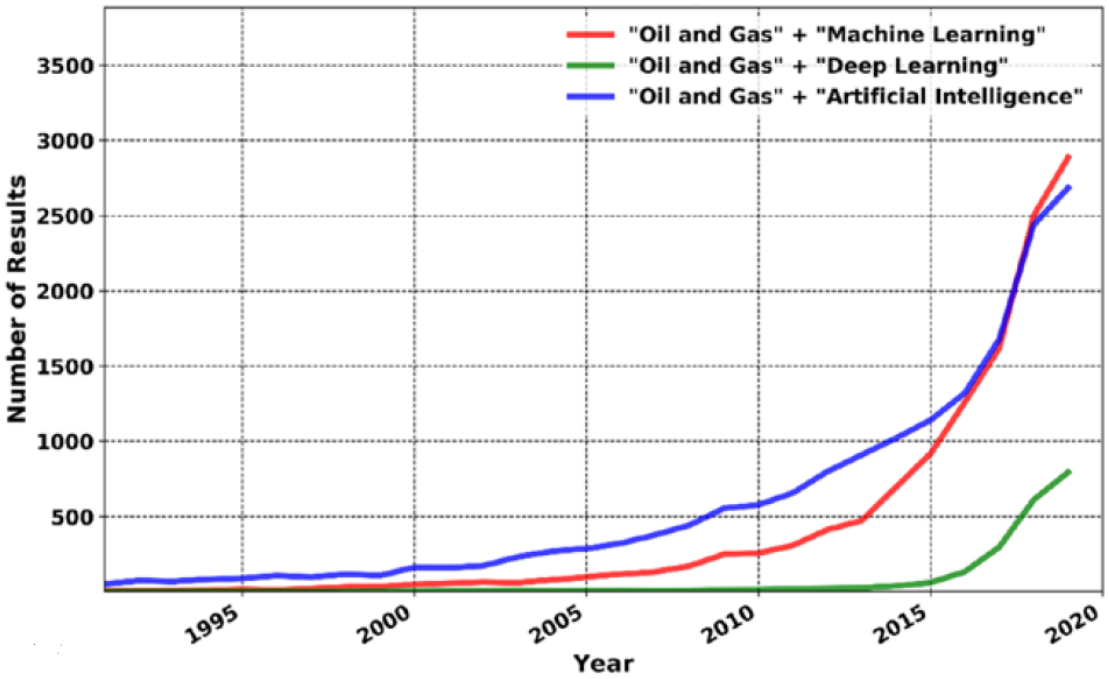


Fig. 2.20: Keywords search on Google Scholar (Pandey et al., 2020).

ML has been applied in the OGI to achieve either remarkable or excellent results in big-data analysis and compare the results with old correlations and commercial software. Theoretical and empirical correlations sometimes are impractical and restrained to specific properties and data (Khan et al., 2019). **Table 2.10** provides examples of ML applications in OGI from the literature. More than 500 papers on OnePetro regarding ML application in OGI (Hajizadeh, 2019).

Table 2.10: Examples of ML applications in OGI from the literature

Study	ML Application
Andrianova et al. (2018)	PVT analysis
Anifowose et al. (2017)	Reservoir characterisation uncertainty
Ahmadi and Chen (2019), Elichev et al. (2019), Alakbari et al. (2017), Onwuchekwa (2018), Ramirez et al. (2017)	Reservoir and fluid properties
Al-Alwani et al. (2019), Al Selaiti et al. (2020), Boguslawski et al. (2019), Bowie (2018), Cao et al. (2016), Han et al. (2019), Herve et al. (2020), Khan et al. (2019), Luo et al. (2018), Pennel et al. (2018), Saghir et al. (2020)	Well performance, production optimisation and forecast
Noshi and Schubert (2018), Pollock et al. (2018)	Drilling and directional drilling optimisation
Pankaj et al. (2018), Prosper and West (2018)	Completion design in unconventionals
Chiroma et al. (2016)	Determining oil prices
Hajizadeh (2019)	Strategic planning and development projects for oil companies

Elichev et al. (2019) Argued that ML algorithms could not be accurate with insufficient data or data that contain much uncertainty, and noise of errors should be ousted to expect rigorous output. On the other hand, many studies (Andrianova et al., 2018; Daigle and Griffith, 2018; Shoeibi Omrani et al., 2019) contradicted their assumptions by proving ML accuracy and adaptability with data aberration in petroleum engineering. Undoubtedly, there is no error-free data since it's either recorded by sensors which are not 100% accurate or collected by operators and exposed to human error. The idea is to train the model to handle those anomalies and learn how to derive approximate or semi-corrected data problem-solving.

2.7.2 Machine Learning Application in Artificial Lift

As mentioned earlier, AL's selection depends primarily on expert engineers' decisions and the history of mature wells. Before selection, engineers conduct well performance and nodal analysis to study well deliverability and production forecast to find a lifting method capable of delivering the designed flow rate. Usually, the

plummet in oil prices postpones several drilling and production operations. As a result, a surge in mature field development arise to increase production from developed wells throughout studying the current AL method. For decades, commercial simulators have been used to design lifting methods which are sometimes repetitive and weary (Kefford and Gaurav, 2016). Many researchers applied ML to determine well performance, reservoir and fluid properties, but few applied ML on the area of AL. Ounsakul et al. (2019) applied supervised ML and data mining to determine an optimum lifting method from ESP, PCP, GL, and BPU. They simplified the selection model to {Lift Selection Model = Algorithm (Field Data)}. The equation illustrates that the model is a tremendous amount of data that algorithms are trained to analyse for AL selection. They aimed to improve the selection criterion by reducing human mistakes. Three algorithms, Naive Bayes, DT, and neural network were used to evaluate 30000 samples from more than 50 wells, reservoir, fluid, and economic factors. Their results showed the ability of ML to select optimum pumps and reduce the cost of the producing well life cycle compared to human decision. Furthermore, the algorithms highlighted the main characteristics that are 80% affecting the selection. Mahdi et al. (2023) the author of this thesis has a recent AL selection work in a Sudanese field using ML. Several production parameters were analysed reflecting four lifting methods installed in 24 wells over 16 years. The top critical factors affecting AL selection are gas and cumulatively produced fluid. Production performance and economic analysis were studied to compare the results of the actual AL in the field and the predicted AL from ML. The results showed that the selection could be done from specific dataset and the predicted AL from the ML had a better production performance than the actual ones at the field. Syed et al. (2020) performed AL system optimisation using ML to select and monitor AL in shale gas fields. Unlike other studies using ML, they added replacing the current AL time to avoid pump failure and increase profits. They also studied monitoring and maintenance practices, which are highly required in the OGI. Ranjan et al. (2015) applied artificial neural network (ANN) to optimise GL in an offshore field in India. A simple model of 10 neurons (reservoir and wells parameters) and one hidden layer was used as an input to obtain the optimum gas injection rate that would be used to reach the maximum oil rate. The ML model was used to substantiate nodal analysis and help the engineers in saving the time of enormous calculations.

2.7.3 Machine Learning Application in Artificial Lift Failure and Run Life

ML also found its way into pump failure and run life estimations. Ounsakul et al. (2020) applied ML algorithms attribute forward selection (AFS) for ESP and BPU failure diagnostics. The parameters of each pump were classified into four categories: (1) failure information (depth, service days), (2) pump configuration (type, size), (3) wellbore geometry (DLS), and (4) subsurface/production information (oil rate, sand production, fluid level and API). According to the effect of each parameter, the results of 1450 failures from AFS analysis were summarised in **Tables 2.11** and **2.12**. Sand production was measured from hang-up depth (HUD) by the slick-line job. The data gathered and bias removed, then statistical analysis showed that the average run life of BPU and ESP was 202 and 728 days, respectively, less than other compared fields. Neural network and square root (R^2) were used to validate the results, which showed concerns about the accuracy of the BPU application model. Rod and pump equipment failure related to manufacturer faults classified as mechanical failures, whereas other failures occurred because of corrosion classified as chemical failures.

Table 2.11: BPU failure parameters (Ounsakul et al., 2020)

Ranking No.	Parameter	% Weight
1	Tortuosity at Pump (deg)	17%
2	Pump Depth (mAH)	17%
3	Fluid Level (m)	16%
4	Sand Produce (pptb)	13%
5	Max Inc. above Pump (deg.)	8%
6	No. of Turn above Pump, DLS > 5 (#)	7%
7	API Gravity (deg)	6%
8	Max DLS above Pump (deg./30m)	5%
9	DLS at Pump (deg./30m)	4%
10	Pump Size (inches)	4%
11	Inc. at Pump (deg.)	3%
12	No. of each Rod Taper (#)	0%

Table 2.12: ESP failure parameters (Ounsakul et al., 2020)

Ranking No.	Parameter	% Weight
1	Sand Production in Terms of Pump Distance from HUD (m)	25.20%
2	Pump Depth (mTVD)	23. 4 %
3	Pump Running Time per Cycle (days)	20.90%
4	Gas Production (MSCF)	11.50%
5	Flowing Temperature (C)	8.30%
6	Gross Production (BBL)	6.40%
7	Inclination at Pump (deg.)	4.30%

Liu and Patel (2013) and Liu et al. 2013 (2010) applied data mining to detect BPU failure by analysing the history of wells. Liu et al. (2010) used SVM, Bayesian Network, and semi-supervised learning, while (Liu and Patel, 2013) used pattern recognition. The process of (Liu and Patel, 2013) was applied through the following steps: (1) data collected from sensors, (2) information extraction from the data, and (3) classification. Liu and Patel (2013) argued that many researchers, including (Liu et al., 2010), ignored the importance of feature extraction and focused on classification, which affected the power of domain knowledge. Supervised learning was used to train the model, and dynamometer card readings were used as input data. Then the model provided either detection or false alarm. The pattern was tested on 100 BPU wells for one year and six months. The results were above 85% correct detection which was higher than (Liu et al., 2010). Liu et al. (2013) applied SVM to develop a prediction model to be applied in universal fields. They mentioned that their first model (Liu et al., 2010) using semi-supervised failure prediction for oil production was valid for a specific field, did not meet the requirements, and consumed much time for labelling. They used approximately 2000 wells data collected from pump-off controllers (POCs) and the life of well information software (LOWIS) database in addition to labelling and clustering enhancement for better results. Moreover, they used two evaluation methods: (1) precision, the ratio of true predictions to all predictions, and (2) recall, which is the percentage of instances truly predicted. The results confirmed that the global model was 1.5% higher in recall and 11.5% in precision compared to their old method, and they aim for applications in other fields.

Sneed (2017) attempted to estimate ESP run-life using the data mining approach SEMMA (sample the data, explore the data, modify the data, model the data, and assess the model) and ML algorithms to lower workovers cost. They studied the history of 51 failures in 37 wells over one year to determine the reasons behind pump failure to mitigate future occurrences. Prosper and West (2018) applied ML in PCP completion design for coal bed methane (CBM). They studied reservoir fluid, history of PCP production, and failure for 1499 samples using Gaussian process regression to predict and extend pumps' life cycle. Their model proved that ML application effectively increased PCP run life which was considered promising for future designs. Another pump failure estimation was done (Bangert, 2019) for BPU using 35292 dynamometer card charts from 299 BPUs. The charts were revised for failure diagnosis and detection using feature engineering. The dataset was divided into two parts, 85% for training and 15% for testing. Four ML algorithms were used; single-layer perceptron neural network, multiple-layer perceptron neural network, extreme learning, and DT that revealed more than 99% accuracy in detecting 11 failures in advance. Boguslawski et al. (2019); Pennel et al. (2018); Saghir et al. (2020) applied ensemble ML algorithms and internet of things (IoT) to interpret real-time measurements to optimise rod pump operation. The model divided a series of dynamometer card readings (images) into clusters and then performed diagnostics to provide an interpretation. A suggested solution was then provided to help the operators figure out the problems at the early stages. The model of (Pennel et al., 2018) could further detect tubing failure from dynamometer card measurements.

2.8 Experts' Opinion on Artificial Lift Selection Methods

Shauna Noonan, an AL expert, pointed out excellent questions regarding AL "do we understand the current technology well enough to know what improvements are needed? Are we advancing technology or just providing" band-aid" solutions because the root cause of failure is not understood? Where does the industry need to focus effort—technology development or reduction of failures caused by poor design, installation, and operating practices?" (Noonan, 2010). Those issues are still significant problems in the OGI. Also, since we are dealing with multi-component fluids with high uncertainty parameters, it becomes challenging to understand downhole conditions; therefore, new data analysis technology and AI

tools are highly required for AL applications. Moreover, achieving the highest profit does not mean the highest oil rate. Most oil and gas companies misunderstand it, and they operate the pumps at higher rates resulting in damaged pumps and reservoirs (Berry, 2016; Bucaram and Patterson, 1994; Ghareeb et al., 2012; Kefford and Gaurav, 2016; Noonan, 2010).

2.9 Summary

The chapter provides a comprehensive literature review of AL selection and failure issues. Despite some modernizations, the same lifting methods and selection techniques have been used for decades. The literature review shows that most AL selection criteria have focused on studying and analysing reservoir parameters, fluid properties, well productivity, surface facilities, power requirements, environmental aspects, corrosion, solids, paraffin handling, gas handling, well completion and design, and economic factors such as workover and maintenance. The conventional selection methods described in existing literature predominantly hinge on engineers' decision-making through a trial-and-error approach, which is notably time-consuming. Furthermore, these methods tend to overlook key considerations such as data volume, data heterogeneity, well conditions change, and the limited historical data in the newer fields. The review highlights the importance of exploring the potential of ML and AI in AL selection to improve its efficiency and effectiveness in the OGI.

CHAPTER 3

PROPOSED METHODS, DATA ACQUISITION AND PREPARATION

3.1 Introduction

This chapter explores the methodologies and means taken to accomplish this research. First, an overview of the case study field in Sudan and the 100 selected pilot wells, followed by a description of the three field data categories used for AL selection, production/reservoir, operation, and environmental/economic. Each data category has been used solely to model the AL selection and identify each parameter's importance. The justification behind the selected criteria has also been provided. The chapter also provides how the raw data is processed and prepared for ML modelling. And a description of how the bias and outliers in the dataset can affect the modelling performance referencing the literature. Many techniques have been implemented to the dataset before modelling, including data cleaning to remove outliers and duplicates, categorical features encoding using one hot encoder (OHE) since the algorithms only deal with digits. Moreover, data normalisation is used to balance the weights of the field parameters, and features correlation to showcase the relation between the parameters and the targeted AL. Some visualisations of the raw data are also included. Lastly, a section presents the algorithms classification mathematics and the accuracy metrics used to measure the performance of the models.

3.2 Data Gathering

Different types of data have been collected from 2003 to 2021. Well by well daily, monthly, and yearly production reports contain surface recorded parameters are disclosed in **Table 3.1**.

Table 3.1: Field data collected for the research

Data Type	Parameters
Production daily, monthly, annually recorded data	Running period, wellhead, casing and flowline pressures, flowrates of oil, water and gas, WC%, produced sand, tubing head and flowline temperatures, pumps frequency, pumps speed and electric current, GOR, dynamic fluid levels
Improved and enhanced oil recoveries (IOR/EOR)	IOR/EOR type, injected gas and water volumes, N2 injection, CSS and SF injected volumes
Reservoir data	OOIP and OGIP, fluids API, initial reservoir pressure and temperatures, fluids density, fluids viscosity, oil formation volume factors, zone thickness, and pour points

Completion and workover data	Operation dates, duration, cause and failure, tubing size, AL type and size of the surface and downhole equipment, wells depth (TVD, MD, PBTD), formation depth, zone thickness and AL setting depth
Environmental and economic data	CAPEX and OPEX, power source and consumption, GHG emissions, oil spill, and field personnel knowledge

Like other kinds of data, raw oil and gas field data cannot be used directly for ML modelling and requires cleaning, filtering, and restructuring. The traditional analysis of the above field parameters by engineers takes a prolonged time, resulting in AL selection inconstancy.

3.3 Field Overview

The selected field is in the Muglad basin in Western Sudan (**Fig. 3.1**), has approximately 700 oil wells, and comprises seven subfields (blocks), XF, XFE, XM, XK, XJ, XH and XS, spread over approximately 17800 square km in seven remote locations (CNPC in Sudan, 2009; Liu et al., 2010).

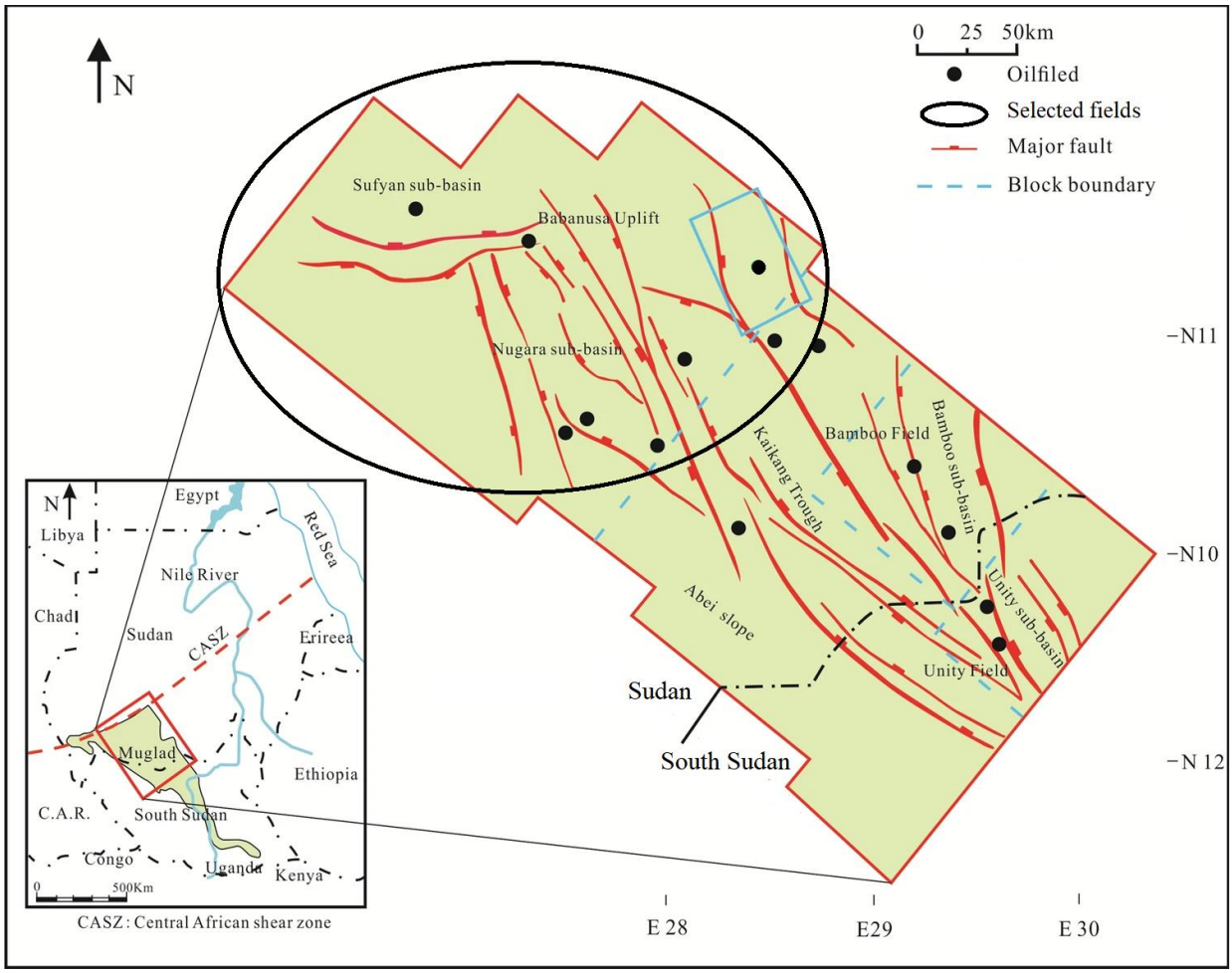


Fig. 3.1: The selected field in Muglad basin in Sudan

The reservoir lithology is sandstone interbedded with shale and has three main formations:

(1) formation A is a deep formation that produces light oil and gas.

(2) formation B is a shallow formation that produces heavy and extra heavy oil.

(3) formation C is a deep formation that produces light oil.

Two more tight formations, D and E, have low oil reserves.

The field's estimated oil reserve is more than 500 million barrels. The field produce light, medium, heavy, and extra-heavy oil. Most production wells are completed with 9⅝ inch casing, 2⅞ inch, 3½ inch, and 4½ inch production tubing. Wells' depths range from 500 m in shallow reservoirs to more than 3000 m in deep reservoirs.

Most light oil wells started production by NF at the beginning of their production life before moving towards AL. Four primary artificial lift methods are installed: PCP, SRP, GL and ESP. PCP and SRP have been the primary lifting methods since the field started production in 2003. The implemented recovery methods are WI, N2, gas injection, and thermal recovery CSS and SF. PCP is installed for cold heavy oil production (CHOP) and cold heavy oil production with sand (CHOPS). BPU and MTM_PCP are used with thermal recovery, while GL and ESP produce light oil. **Table 3.2** summarises the numbers of wells and AL in the field.

Table 3.2: Wells and AL summary

AL	Number of wells	Production
PCP	503	CHOP and CHOPS
BPU	292	Thermal
ESP	33	Light oil low GOR
GL	13	Light oil high GOR
MTMPCP	2	Thermal
NF	15	Light oil high GOR, Gas
Idle wells	20	No potential
Injection wells	39	WI, N2, Gas, CSS, SF

3.4 Pilot Wells

The distribution of the drilled well varies across the subfields according to oil reserves. XFN is the largest block with over 200 wells followed by XFE with 120

wells. Thus, the wells were selected concerning each block capacity to reflect the actual field conditions. The research was conducted on 100 oil producing wells. These wells have an excellent recording of the data compared to other wells in the field. In addition, they construct a perfect combination of dataset in terms of production performance, AL replacements and failure issues, as well as workovers operation records. **Table 3.3** summarise the distribution of the selected pilot wells for the research.

Table 3.3: Selected wells distribution

Subfield	Number of well	Installed AL
XF	16	GL, PCP, MTMPCP, NF
XFE	15	PCP, MTMPCP, BPU
XM	15	PCP, NF
XJ	15	GL, PCP, NF
XK	15	GL, PCP
XH	12	PCP, ESP, NF
XS	12	PCP, ESP

3.5 Field Parameters Screening and Selection (Features Selection)

ML application requires an essential determination of the proper features (input parameters) to predict the accurate targets (targeted AL). Feature selection is the process of determining the relevant and irrelevant input factors. Identifying nonessential features is recommended for dimensionality reduction (reducing number of features to improve algorithm performance) and efficient and faster algorithm performance (Kotsiantis et al., 2006, 2007). Moreover, deriving new features from existing features can effectively influence the model's accuracy. The process is known as feature engineering, resulting in high-performance classifiers and accurate outputs (Kotsiantis et al., 2007).

On the one hand, the deficiency in some measurements and metering tools, such as downhole real-time pressure and temperature gauges, flowmeter calibration, and instrument malfunction, have restricted the features and feature engineering. On the other hand, it is part of the research's aim to select the AL from the insufficient data. There is a poor recording of reservoir parameters and well-testing operations. Most of the data is recorded at the beginning of drilling and

the start of production. Also, some reservoir measurements are intermittently obtained during workovers.

Features have been selected according to their impact on AL and production performance in the specific field concerning the current field conditions. The features are a categoric and numeric mixture. The chosen parameters from the available field data were divided into three categories:

a) Production features:

These are mainly production parameters with some recorded reservoir parameters at the surface, including wellhead pressure (psi), wellhead temperature (°C), flowline pressure (psi), flowline temperature (°C), casing head pressure (psi), daily produced fluid (BLPD), GOR (scf/STB), daily oil production (STB/D), daily gas production (Mscf/D), daily water production (BWPD), daily sand production (RB/D), WC%, pumps frequency (Hz), pump speed (RPM), pump current (amp), and the implemented secondary and tertiary recovery methods to increase oil production. The secondary recoveries, known as IOR and used for reservoir pressure maintenance, are gas injection, N2 injection, and water injection (WI). CSS and SF are the tertiary recoveries or EOR that are used to reduce heavy oil viscosity and increase mobility towards the well-bore.

b) Operation features

These are completion and workover operation parameters and well and formation characteristics which include well true vertical depth TVD (ft), well measured-depth MD (ft), plug back total depth PBTD (ft), AL setting depth (ft), AL running period (days), wells completions (cased-hole, open-hole, commingled), perforation depth (ft), formation type (A, B, C, D, E), production zone thickness (ft), tubing size (inch), AL completion and workover duration (days), workovers frequency over production years, workover cause, AL failure cause, AL replacement cause.

c) Environmental and economic features

The parameters are a combination of AL purchasing cost (surface and downhole equipment), completion cost, workover cost, power source (gas, electricity, natural), gas emission levels, oil spill levels, noise levels, and operator knowledge in AL maintenance and operation.

Some of the selected parameters in each category are not used for modelling as they have low significance in the selection. More details are enclosed in chapter 4.

3.6 Data Pre-Processing (Data Wrangling)

ML acquires knowledge by feeding information to provide accurate outputs; therefore, the quality of data determines the success of the ML model. Here comes the importance of data wrangling for inappropriate and contaminated data (Kotsiantis et al., 2006). Data wrangling, or data preparation and pre-processing, is cleaning and modifying the data for ML modelling (Brownlee, 2020: p.4). It is considered the most challenging step in ML application, and the process is commonly performed through the following steps (Brownlee, 2020: p.72):

- Data collection.
- Features selection.
- Data structuring into rows and columns.
- Outliers' removing and duplicate cleaning.
- Data restructure, remove the rows of missing values or fill them with average values or means as some algorithms cannot deal with missing data, for example, SVM and neural network.

3.6.1 Data Cleaning and Visualisation

Following data collection and feature selection, the data is restructured into rows and columns so the algorithms can easily model it. Although data wrangling is crucial and significantly impacts model performance, it is arduous to detect outliers and noise in the data (Kotsiantis et al., 2006).

Data are being gathered daily by field operators and remote transmission units (RTUs). The raw data has anomalies to be removed, as well as unrecorded data due to wells shut in for maintenance, workover, or power failure issues. In addition, some data was not recorded due to malfunction of measurement tools, for instance, flow meters and pressure/temperature gauge failure issues. Also, some outliers are from field operators recording errors. Thus, the data contains hundreds of duplicates, outliers, and missing values, which require preparation and cleaning before modelling.

The raw data has many outliers, for example, well XF3, XF144 and XF161 have unreasonable values of 252, 785 and 734 barrels of sand, respectively, as shown in **Fig 3.2**. In addition, they have gas measurements that are unlikely to be recorded in heavy oil producers. Fewer anomalies will affect the model efficacy

and result in poor model performance and inaccurate outputs (Brownlee, 2020: p.12).

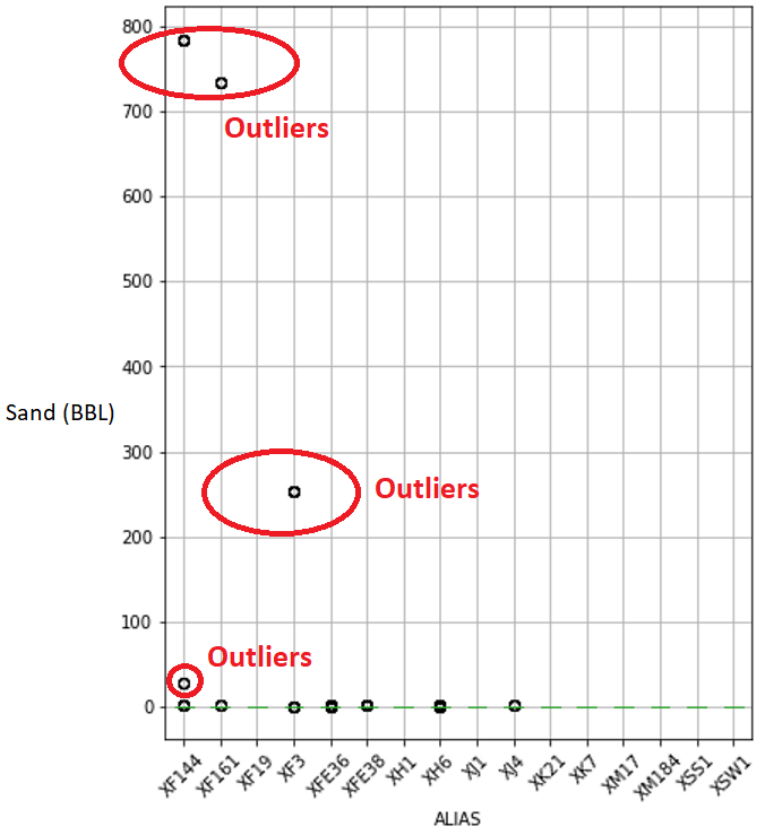


Fig. 3.2: Sand anomalies

There were 26568 missing values in the raw dataset. The missing values were in oil flow rates, wellhead pressures, produced gas, and GOR in GL wells. The unrecorded oil production rates probably indicate that the production well was shut-in or under workover. These missing values were removed as they are not required for modelling. Other critical parameters, namely GOR, gas, and wellhead pressure while the well was running, were unrecorded due to measurement tools malfunction. These measurements are recorded by gas separators that require continuous maintenance, resulting in data noise. Missing data is an inevitable issue in the OGI. A robust solution to handle missing data obstacles is the average value of the specific feature (Kotsiantis et al., 2006). Therefore, the critical values were substituted by the mean value since they significantly impact the AL selection, as demonstrated in chapter 6. The following **Figs** (**3.3**, **3.4**, and **3.5**) demonstrate the missing data on GOR, gas and wellhead pressure.

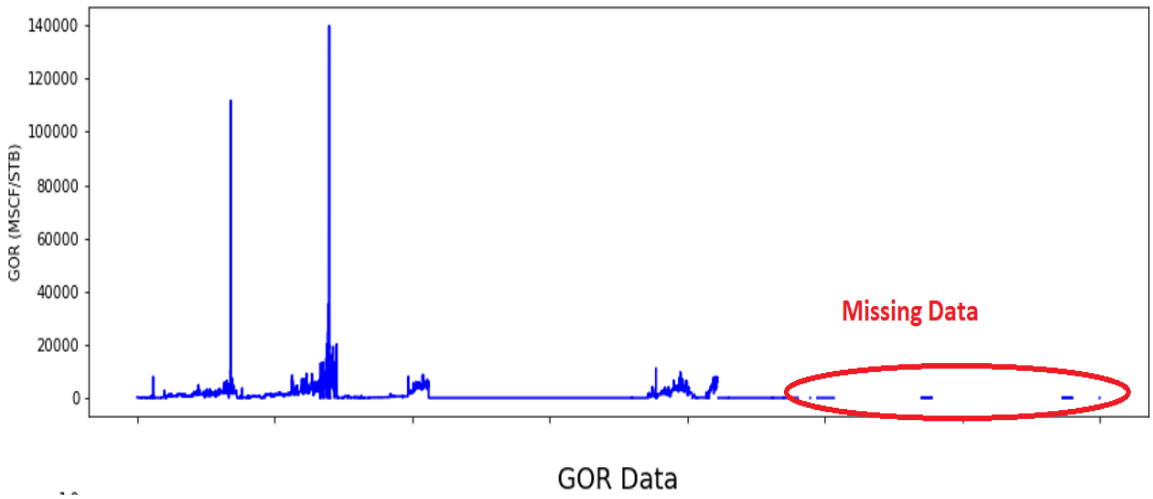


Fig. 3.3: GOR missing data

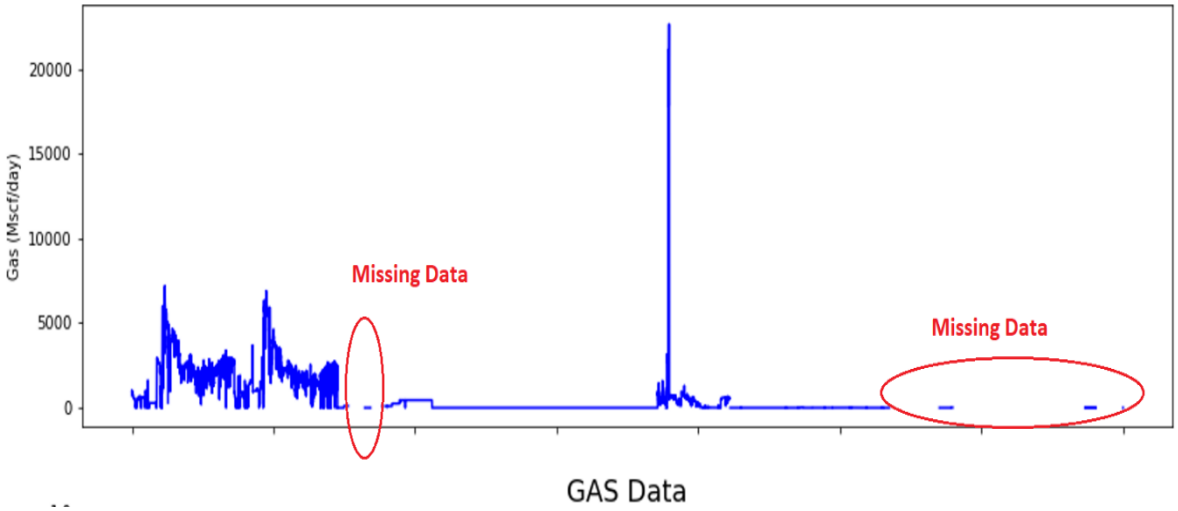


Fig. 3.4: Gas missing data

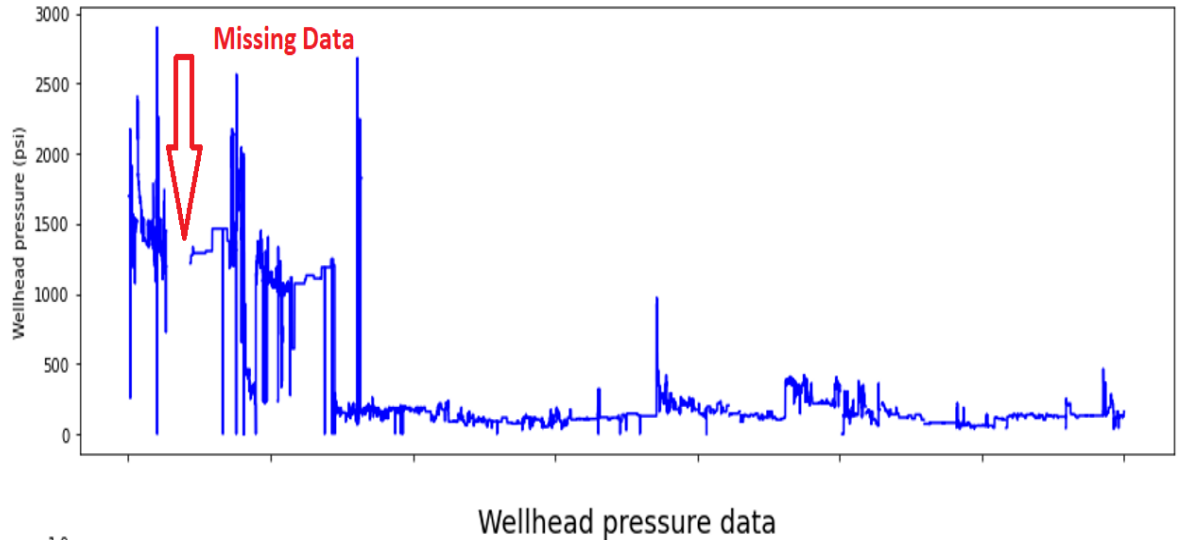


Fig. 3.5: Wellhead pressure missing data

Duplicates are considered one of the worst antagonists for ML modelling. They generally mean that some values appear several times with the same sequence in the dataset. If not cleaned, they will negatively affect the algorithm learning and reduce model accuracy (Brownlee, 2020: p.50). About 4578 duplicates have been removed from the raw dataset.

3.6.2 Categorical Features Encoding

The dataset has many categoric features, such as IOR and EOR recovery methods, workover cause, failure cause, power source, the amount of gas emission, oil spill, noise, and field personnel knowledge of AL. ML only deals with numeric values, so the categorical values should be converted to numeric before modelling. One Hot Encoder (OHE) was applied to code the variables. OHE defines the present values in each sample by 1 and the absence by 0 (Potdar et al., 2017). For example, if WI presents during production, it will equal 1, while other IORs and EORs will be given a 0 value.

3.6.3 Data Normalisation

Normalisation is scaling the data to be in a small range. It is always used to reduce the difference between the maximum and minimum values in the dataset (Kotsiantis et al., 2006). It is essential as algorithms such as KNN and SVM are susceptible to the distance between data samples. The feature elements cannot be directly fed into the algorithm for modelling because the model will focus on larger values to learn rather than others. Thus, the data should be normalised between 0 and 1. Unnormalised data will reduce model performance and provide an unstable model (Brownlee, 2020: p.214-215).

Input values were normalised before modelling in a range of [0,1] by calculating the difference between feature elements and minimum value and then dividing by the max and min difference, equation 3.1.

$$Normalised\ x = \frac{x\ - x\ min}{x\ max - x\ min} \qquad (3.1)$$

3.6.4 Data Correlation

Data correlation is applied through the correlation matrix to find the relation between the input parameters and the target variables. It also shows how input parameters correlate to each other and target variables as well. Positive

coefficients indicate that the two variables have a positive correlation and vice versa. The closer the coefficient to 1, the stronger the correlation is (Kumar and Chong, 2018). For instance, if an AL strongly correlates to the features compared to other AL methods, the algorithm will find no barriers in classifying it, as we will see in later modelling results. The correlation matrix holds a pivotal role in multivariate analysis, serving as a fundamental component for assessing relationships among variables in a dataset. It is instrumental in uncovering possible multicollinearity, which pertains to the extent of linear association between two or more variables. Furthermore, the correlation matrix aids in gauging the magnitude and robustness of these associations (Pham-Gia and Choulakian, 2014). The correlation matrix applies the Pearson correlation coefficient to quantify the linear relationship between the data points X and the target Y. It's computed by dividing their covariance by their standard deviations (stdv), normalizing the covariance to provide an interpretable measure.

$$Pearson's\ correlation\ coefficient = \frac{covariance(X,Y)}{stdv(X)\ *\ stdv(Y)} \tag{3.2}$$

$$covariance(X,Y) = \frac{\sum[(X\ -\ mean(X)\ *\ (Y\ -\ mean(Y)]}{n\ -\ 1} \tag{3.3}$$

Where n is the number of datapoints.

3.7 Machine Learning Algorithms Classification Criteria

3.7.1 Supervised Learning Algorithms

3.7.1.1 Logistic Regression

LR solves the probability *P(Y|X)* task by learning from a training dataset *Dt* written as a vector of weights and a bias *(WiXi +b)* where the weight *Wi* is the importance of the feature *Xi* in the dataset. To make a classification on a new test dataset, LR calculates the weighted sum class evidence *z* and then passes the result down to the sigmoid function *σ (z)*, which narrows the results between 1 true or 0 negative class. The algorithm uses the decision boundary 0.5 to predict the classes (Jurafsky and Martin, 2021).

$$z = \left(\sum_{i=1}^{n} WiXi\right) + b \tag{3.4}$$

$$\sigma(z) = \frac{1}{1 + exp^{-z}} \tag{3.5}$$

$$\mathrm{P(Y|X)} = \begin{cases} 1 \; if \; \sigma(z) > 0.5 \\ 0 \; otherwise \end{cases} \tag{3.6}$$

3.7.1.2 Support Vector Machines

SVM is a binary classifier; however, it can be used for multiclass classification by breaking down the problem into a series of binary classification cases. To do this, we applied the OVO (one vs one) approach. The concept is that each classifier separates the points of two classes, including all OVO classifiers, to establish a multiclass classifier. The number of classifiers needed for this is calculated using *n(n−1)/2* (*n* = no of classes). Eventually, the most common class in each binary classification is then selected by voting (Mathur and Foody, 2008).
We assume that we have a training dataset *Dt* {Xi, Yi} that is binary and linearly separable where each Xi has dimensionality q and is either one of Yi=-1 or Yi=+1 (here q=2). The training data points can be described as follows:

$$Xi \cdot W + b \geq +1 \; for \; Yi = +1 \tag{3.7}$$

$$Xi \cdot W + b \leq -1 \; for \; Yi = -1 \tag{3.8}$$

The SVM draws hyperplanes to separate the classes of each training data point. The hyperplane is expressed as *W·X+b = 0.* The support vectors separate the hyperplanes, while the machines keep the distance as far as possible between the hyperplanes that separate the classes. In our multiclassification problem, if the data points are not linearly separable, then the SVM applies the kernel function *k(Xi, Xj)* to recast the data points into a higher dimensional space $X \mapsto \emptyset(X)$ to be

separable (Fletcher, 2009). Thus, the training data points recasting into the higher space is written as:

$$(Wi)^T \emptyset(X) + b_j \quad where\ j = 1, \dots, n \tag{3.9}$$

We applied the radial based kernel (Gaussian kernel) in our model which is expressed as:

$$k(Xi, Xj) = exp^{-\gamma(\|Xi - Xj\|^2)} \tag{3.10}$$

Where γ controls the width of the kernel function. The decision function of voting that provides the *n* class label of the *mth* function is written as (Mathur and Foody, 2008):

$$majority\ voting_{m=1,\dots,n}\ (W_m^T \emptyset(X_i) + b_m) \tag{3.11}$$

3.7.1.3 K Nearest Neighbours

In KNN, we have a given training dataset *Dt* {(X_1, Y_1), (X_2, Y_2),…,(Xn, Yn)}, the algorithm calculates the Euclidean distance of each class to all training data points using the below formula. Then construct a boundary to each class by determining the K nearest neighbour.

$$\text{Euclidean distance} = \sqrt{\sum_{i=1}^{n} (Xi - Yi)^2} \tag{3.12}$$

The algorithm performance is susceptible to the K value, which is difficult to estimate. A small K will result in overfitting, while a large K leads to class boundary intersections and training data scattering in many neighbourhoods. The best option is to try different K values until the highest accuracy is achieved (Taunk, 2019).

3.7.1.4 Decision Tree

The DT consists of a root node, internal nodes, and leaf nodes that assign the class labels. The concept of DT is that it identifies the informative features regarding each class label. To obtain that, as we use the CART (classification and regression tree) because some features are categoric as well as the outputs; the

DT applies the Gini index in each node to split the data. Gini is defined as a measure of the probability of incorrect predictions when the features are selected randomly (Tan et al., 2006). Assume we have *Dt* training dataset; the Gini index is calculated using the following equation:

$$Gini = 1 - \sum_{i=0}^{n} (Pi)^2 \quad (3.13)$$

Where *Pi* is the probability of partitioned data of class *i* in *Dt*, and *n* is the total number of classes of *Dt*. The feature with a lower Gini value is used to split the data.

3.7.1.5 Random Forest

In RF, let *Dt* be a training dataset {(X_1, Y_1), (X_2, Y_2),…,(Xn, Yn)}, since the RF is a combination of trees, the RF applies the bagging (bootstrap aggregating), an aggregation of multi-trees results. The idea is that the RF randomly splits the training dataset to bootstrap samples to create multi trees and repeat the process. The class Y that has the majority among the results of the trees is then selected by voting (Breiman, 1996). The RF classifies through the following steps (Hastie et al., 2009):

1. For b=1 to B (no of trees)
   a) Create bootstrap sample Z of size N on the Dt
   b) Grow a tree $\boldsymbol{T_b}$ on the bootstrapped data and recursively repeat these steps until minimum node size is reached:
      i. selecting random *m* variable from the *p* variables.
      ii. find the optimum split point among the *m* variable using Gini index to create the daughter nodes.
2. Get the output ensemble trees $\{\{T_b\}_1^B\}$

Let $\widehat{Y_b}(X)$ be the class prediction of the *bth* tree, then the conclusive predicted class is given by:

$$\widehat{Y_{RF}}(X) = majority\ voting\ \left\{\widehat{Y}_b(X)\right\}_1^B \quad (3.14)$$

3.7.2 Unsupervised Learning Algorithm K Means

The k-means clusters the data through the following steps (Kharrat et al., 2009):

1. Let D be the dataset, K be the number of clusters.
2. Clusters C = $\{c_j$, j = 1…k$\}$
3. Let n = |D|
4. Initialise K clusters with randomly chosen d∈D
5. Allocate each input parameter to cluster 0, a_i = 0, i = 1…n.
6. **Repeat** the steps.
7. Allocate each d∈D to its nearest cluster in C.
8. Update each cluster c_j as mean of $\{d_i \setminus a_i$ = j$\}$
9. **Until** A converge

The clusters centroids m_i is defined as:

$$m_i = \frac{\sum_{j=1}^{n} w_{ij} x_j}{\sum_{j=1}^{n} w_{ij}}, \forall_i \tag{3.15}$$

Where w_{ij} is the membership function indicating whether the datapoint x_j belongs to cluster w_i and satisfies the following conditions:

$$\begin{gathered} \sum_{j=1}^{k} w_{ij} = 1, \forall_j \\ 0 < \sum_{i=1}^{n} w_{ij} < n, \ \forall_i \end{gathered} \tag{3.16}$$

Where n is the number of datapoints in the sample x_j,…, x_n, and K is the number of clusters. The clusters classification depends on the Euclidean distance. That is done by minimising the sum of squares of distances between the datapoints and the cluster centroids (Teknomo, 2006) and is written as:

$$J = \sum_{i=1}^{n} \sum_{j=1}^{k} w_{ij} \left\| x_j - m_i \right\|^2 \tag{3.17}$$

To improve the quality of clusters, it is important to consider the distance measurements used. Normalizing the data can enhance the Euclidean distance and, in turn, improve the quality of the clusters (Bansal et al., 2017).

3.7.3 Accuracy Scores

Accuracy is a model performance evaluation tool. It is the ratio of the correct predicted labels to the total number of labels (Mohamed, 2017). It is written as the following formula:

$$Accuracy = \frac{tp + tn}{tp + tn + fp + fn} \tag{3.18}$$

$$Error = 1 - Accuracy = \frac{fp + fn}{tp + tn + fp + fn} \tag{3.19}$$

Where:
tp: is the class of interest that correctly classified (true positive).

tn: is not the class of interest that correctly classified (true negative).

fp: is the class of interest that incorrectly classified (false positive).

fn: is not the class of interest that incorrectly classified (false negative).

(Note: fp, tp not f_p, t_p)

Another precise evaluation tool is the confusion matrix (Choudhary and Gianey, 2017), it elaborates the truly classified targets and those falsely classified. and its detailed report is shown in **Fig. 3.6** below:

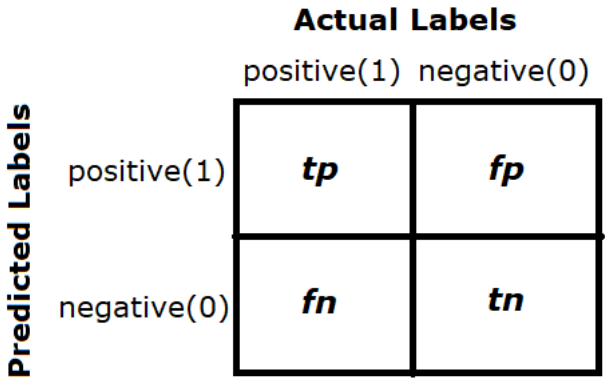


Fig. 3.6: Confusion matrix report

Recall is the percentage of instances truly classified $^{tp}/_{fn + tp}$ (3.20)

Precision is the positive rate $^{tp}/_{fp+tp}$ (3.21)

F1 score is the mean of precision and recall (Lipton et al., 2014)

$$^{2tp}/_{(2tp+fp+fn)} \quad (3.22)$$

3.8 Methodology Workflow and Python Libraries

Fig. 3.7 provides a graphical illustration of the methodological workflow of supervised ML employed in this research. Scikit-learn, a Python library for ML, provides a diverse array of both supervised and unsupervised learning algorithms, as well as tools for activities such as model selection, evaluation, and preprocessing. Designed with a focus on user-friendliness and computational efficiency, Scikit-learn features a unified interface that streamlines the process of transitioning between various algorithms and models. This not only enhances predictive accuracy but also reduces computational time (Pedregosa et al., 2011). Scikit-learn also offers interoperability with different libraries, including NumPy arrays, Pandas data frames, and Matplotlib for data visualization, which are integral in our AL selection predictions. Pandas specialises in data manipulation and analysis, encompassing an extensive set of functions and techniques for data cleansing, manipulation, and visualization. It proves to be a versatile tool throughout the data analysis workflow. Pandas facilitates a range of data operations, such as filtering, selection, grouping, and data aggregation. Moreover, it excels in handling missing data, categorical data, and data encoding/decoding. Pandas is also well-regarded for its capability to compute diverse statistical metrics and correlations among input features (Pedregosa et al., 2011). Within our methodology, Pandas plays a pivotal role in the preprocessing and preparation of data before feeding it into Scikit-learn models for building predictive models and making optimum AL selection predictions. Python codes used for modelling and pre-processing are presented in **Appendices A1-A3.**

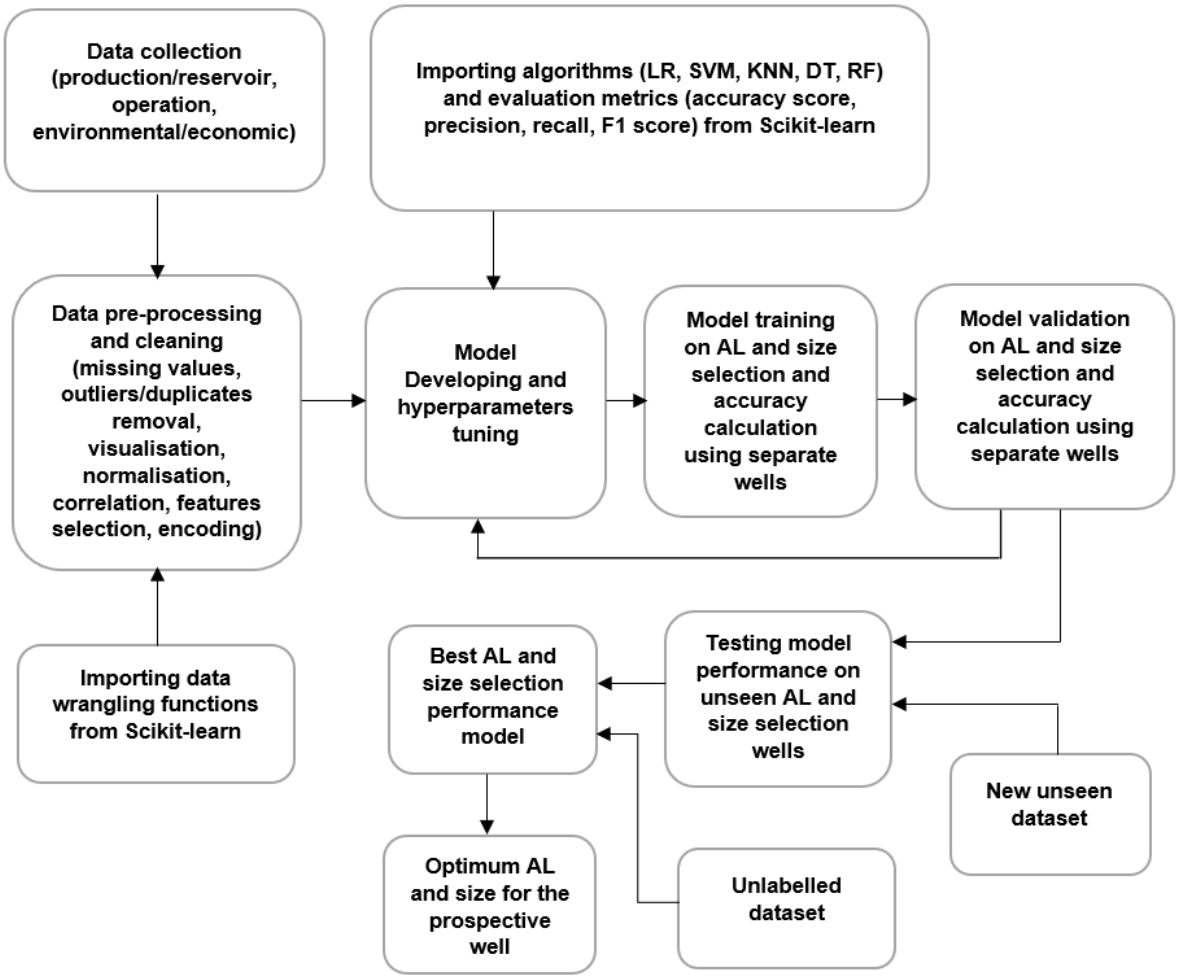


Fig. 3.7 AL selection workflow

3.9 Summary

In this chapter, the methodologies and framework used to conduct the research study and ensure accurate modelling results are presented. The study utilised real recorded field data for developing, testing, and validating the model. To ensure the robustness of the developed model, the oil wells were randomly selected.

CHAPTER 4

MACHINE LEARNING APPLICATION AND ARTIFICIAL LIFT SELECTION MODELS

4.1 Introduction

In this chapter, novel AL selection models employing ML techniques are presented. These ML models offer a remedy for the qualitative conventional AL selection methods, which have historically relied on subjective, experience-based, and rule-based approaches. These traditional methods, result in suboptimal decisions that directly affect well productivity and profitability. The adoption of ML represents an avenue for enhancing the precision of AL selection through the analysis of historical field data to discern patterns and forecast the most suitable AL method. The chapter provides a comprehensive analysis of three ML-based AL selection models, which were developed and tested on three different field data categories, reservoir/production, operation, and environmental/economic. In addition, the chapter investigates the optimum clustering of the field data that would help in optimum AL selection. The models' performance is evaluated based on various metrics, including accuracy, precision, and recall, to demonstrate their effectiveness in selecting the optimal AL method for a given well. Overall, this chapter contributes to the body of knowledge on the application of ML in the petroleum industry and provides practical insights for improving AL selection processes.

4.2 Developed Artificial Lift Selection Models Using Supervised Learning

4.2.1 Selection Model Based on Production Data

4.2.1.1 Input Parameters and Data Visualisation

This model utilises production and some recorded reservoir parameters at the surface. Particularly those mostly correlated to and measured from the flow rate, which measures oil well performance. We used the daily cumulative flow rate and production parameters throughout the AL years of production to thoroughly analyse the AL production performance at the oil field rather than using only flow rate limitations. Also, some parameters were selected according to their effect on

AL and well productivity, such as IOR and EOR recovery methods. Since excessively many features negatively affect model performance, lead to model overfitting, and increase computational cost (Brownlee, 2020: p.111); only nine features were selected for better modelling performance. The parameters listed in **Table 4.1** were selected according to field data availability. The implemented secondary and tertiary recovery methods to increase oil production is the only categorical parameter among the chosen features.

Table 4.1: Production model features

Feature	Unit
Wellhead pressure	psi
Daily produced fluid	BLPD (bbl/D)
gas-oil ratio (GOR)	scf/STB
Daily oil production	STB/D
Daily water production	BWPD (bbl/D)
Water cut	%
Daily gas production	Mscf/D
Daily sand production	bbl/D
IOR/EOR methods	Categories (Gas injection, NI, WI, CSS, and SF)

The distribution of AL methods in the dataset is illustrated in **Fig. 4.1**. Although the GL method has a higher cumulative fluid production (**Fig. 4.2**) than the PCP, ESP, and BPU, PCP dominates the lifting methods used in the field. Additionally, the NF has high cumulative production nonetheless drops rapidly after a short production period due to insufficient reservoir energy. The imbalance in the dataset may lead to an accuracy paradox if not addressed through upsampling or downsampling to achieve approximately equal class distribution in the dataset. However, in this model, the data was kept slightly imbalanced to reflect the actual field state and assess the model's robustness in AL selection. Recent studies have shown advancements in modelling imbalanced data and learning from it (Krawczyk, 2016).

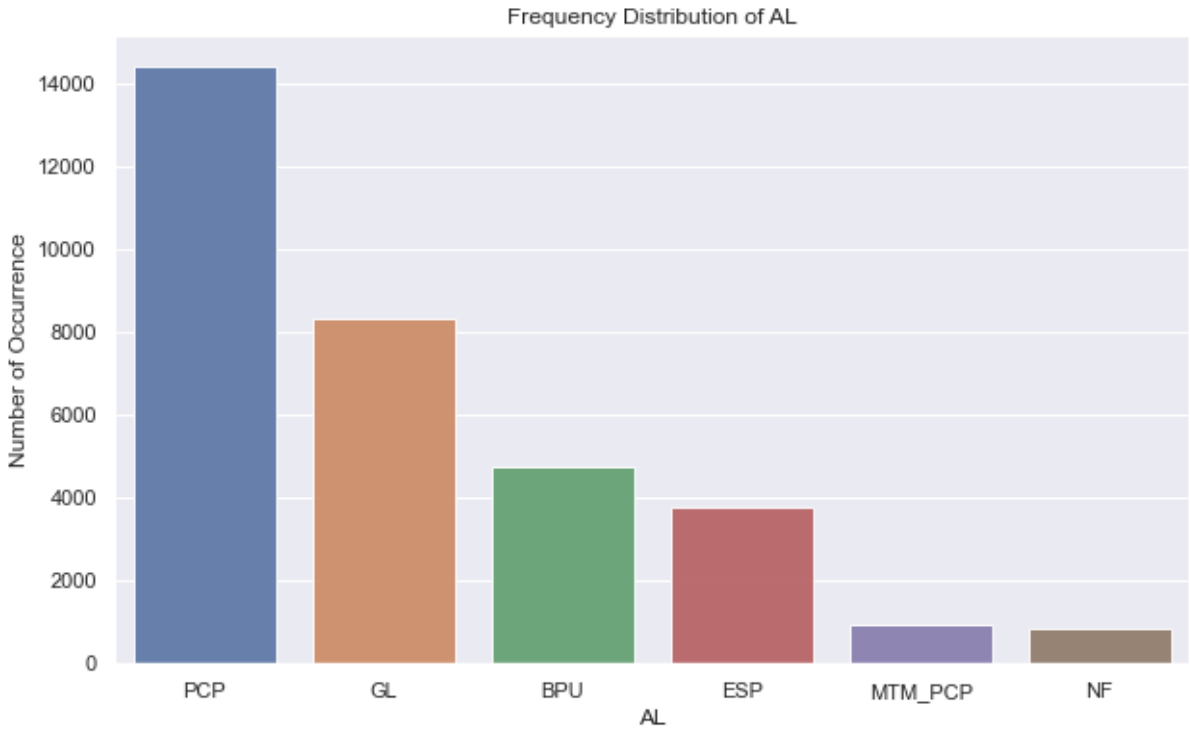


Fig. 4.1: AL distribution in production dataset

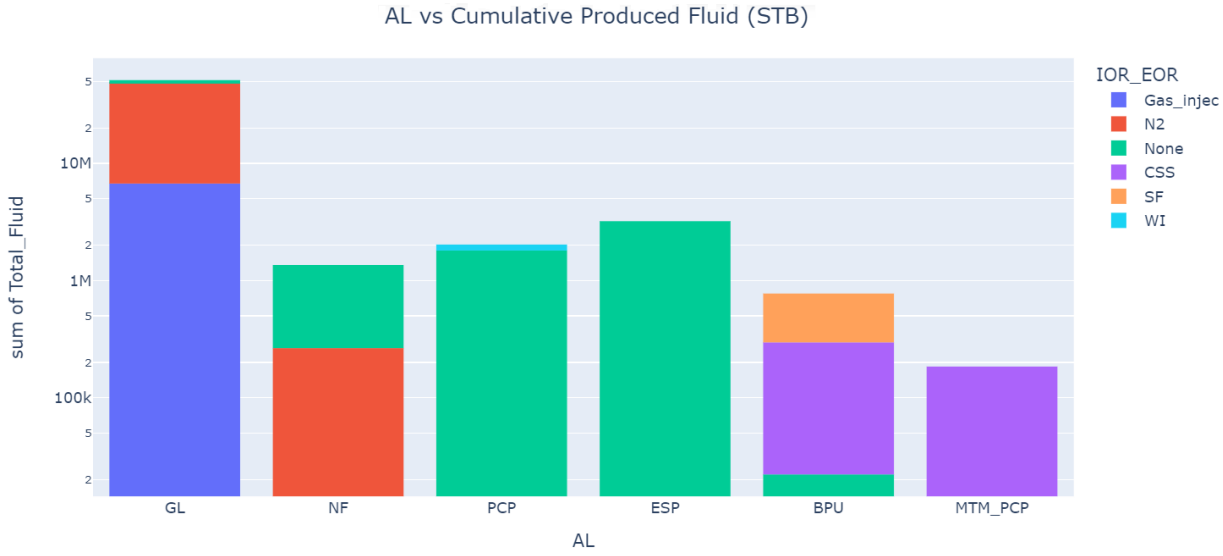


Fig. 4.2: Cumulative Fluid produced by each AL since 2005

Fig. 4.2 also depicts how the IOR and EOR affect oil production. The results show that gas and N2 injection significantly increase light oil production. Interestingly, the amount of oil produced with no IOR/EOR is greater than the production achieved through thermal recovery methods (CSS and SF). This is due to the

longer CHOP and non-IOR/EOR production periods compared to thermal production, chiefly CSS. Wells under CSS undergo multiple workovers to retrieve the injection string and insert the production string into the hole, leading to extended shut-in periods and production loss. GL has a cumulative fluid production of more than 50 million STB. Followed by ESP and PCP of 3.2 and 2 million STB, respectively. NF, BPU, and MTMPCP cumulative production is below 1 million STB.

4.2.1.2 Data Correlation

Fig. 4.3 exemplifies the interrelation between the input parameters and both the target parameters and their mutual interactions. As mentioned in Chapter 3, positive coefficients indicate that the two variables have a positive correlation while negative values reflect the weak correlation between the features. The closer the coefficient to 1, the stronger the correlation and the darker the colour is. The correlation numbers in the figure are calculated using Pearson coefficient represented in *section 3.3.4* (equations 3.2 and 3.3). As shown in the figure, GL strongly correlates to utmost features compared to other AL methods with values ranging between 0.49 to 0.84. Notably, a positive correlation emerges between PCP and BPU with respect to the parameter of sand, while ESP exhibits such an association with GOR and WC%, by 0.05 and 0.14 respectively. The feature run-period correlates positively to all lifting methods with the exception of BPU marking -0.26. This discernible pattern can be attributed to the recurrent CSS cycles, causing shorter operational spans for BPU in contrast to its counterparts, thereby rendering this divergence in correlation patterns. The correlation matrix imparts insights into pivotal attributes that considerably shape the selection procedure, underscoring essential aspects to be taken into account for forthcoming well operations. As outlined in Chapter 3, the analysis of the correlation matrix allows for the exploration of interrelationships between various parameters within the dataset and the target variables, rendering it an invaluable instrument for conducting multifaceted inquiries.

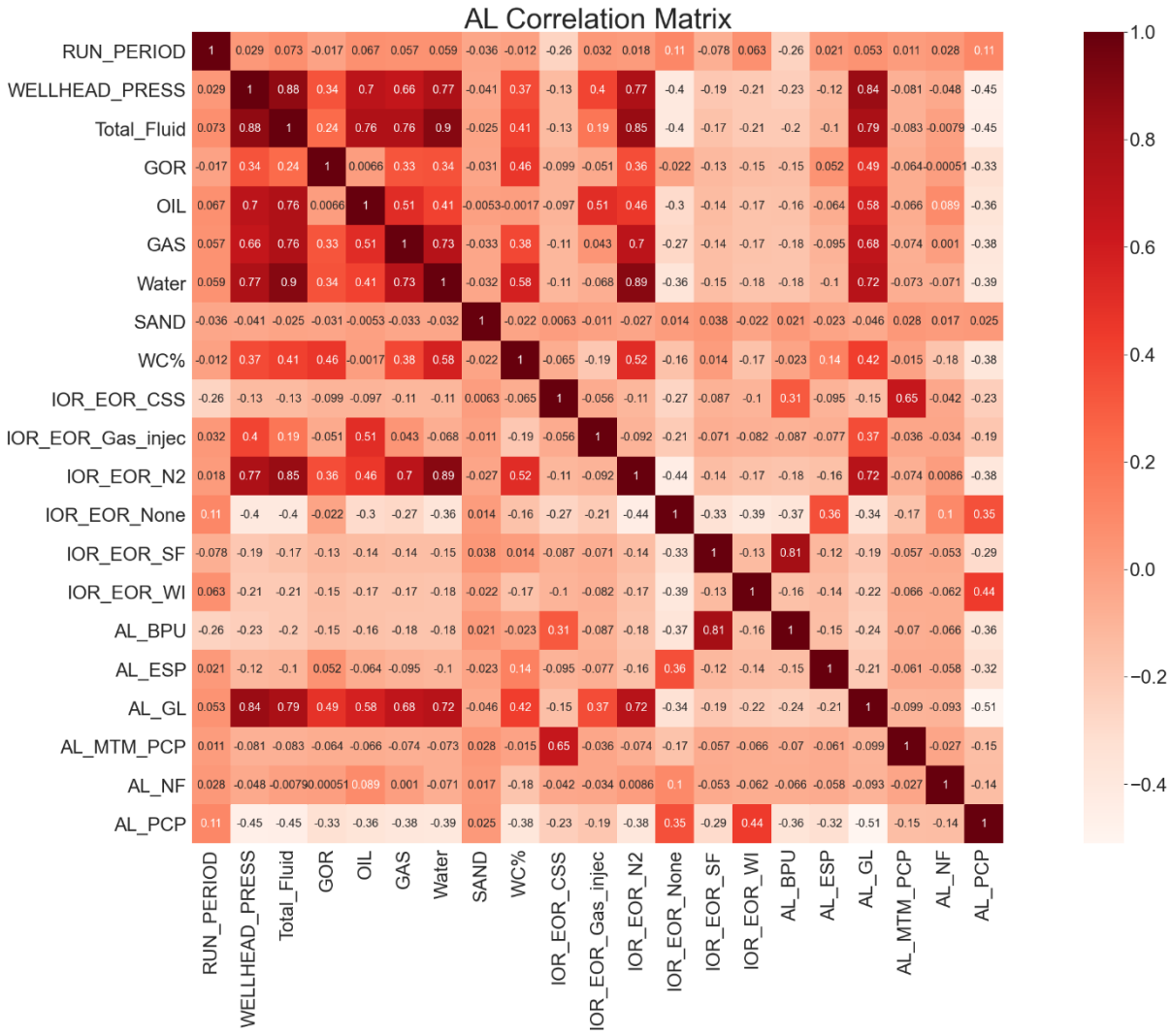


Fig. 4.3: Production features correlation matrix

4.2.1.3 Statistical Data Analysis

The following **Tables**, **4.2** and **4.3**, present an extensive statistical assessment of production attributes prior to and post the application of encoding and normalisation. The significance of normalization and encoding lies in mitigating the influence of larger values and enhancing the stability of statistical analysis to improve accuracy, as detailed in Chapter 3. The tables show the ranges of flow rates, wellhead pressures and GORs used in the model. Furthermore, the tabulated data encapsulates the means and standard deviations affiliated with the input features, furnishing a comprehensive overview of their distributional characteristics. Statistical analysis plays a pivotal role in ML classification models by providing insights into the relationships between input features and the target variable. It helps in identifying significant variables, understanding their

contributions, and assessing multicollinearity, thus aiding in feature selection and model interpretability (Hastie, 2009). Furthermore, statistical techniques offer means to validate the model's generalization ability, enhancing its robustness and reliability in real-world scenarios.

Table 4.2: Statistical analysis of production features before encoding and normalisation

	RUN_ PERIOD	WELLHEAD_PRESS	Total_ Fluid	GOR	OIL	GAS	Water	SAND	WC%
mean	23.595	366.711	1782.041	674.162	716.764	519.511	1065.260	0.005	36.697
std	1.651	464.292	3245.345	1790.038	1532.433	1194.550	2304.702	0.0557	30.875
min	10	27.55	7.438	0	0	0	0	0	0
25%	24	116	79.11	0	53.878	0	6.881	0	6
50%	24	145	207.283	0	113.582	0	46.490	0	32
75%	24	261	1220.89	560.384	520	450	404.786	0	63
max	24	2900	16214	118279.570	15180.5	22656	11783.66	1	100

Table 4.3: Statistical analysis of production features after encoding and normalisation

	Year	Month	Day	RUN_PERIOD	WELLHEAD_ PRESS	Total_Fluid	GOR	OIL	GAS	Water	SAND	WC%	IOR_EOR_CSS	IOR_EOR_ Gas_injec	IOR_EOR_N2	IOR_EOR_ None	IOR_EOR_SF	IOR_EOR_WI
mean	0.665	0.511	0.493	0.971	0.118	0.109	0.006	0.047	0.023	0.090	0.005	0.367	0.065	0.043	0.158	0.504	0.099	0.130
std	0.204	0.314	0.292	0.118	0.162	0.200	0.015	0.101	0.053	0.196	0.056	0.309	0.246	0.204	0.365	0.500	0.299	0.336
min	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
25%	0.533	0.273	0.233	1.000	0.031	0.004	0.000	0.004	0.000	0.001	0.000	0.060	0.000	0.000	0.000	0.000	0.000	0.000
50%	0.733	0.545	0.500	1.000	0.041	0.012	0.000	0.007	0.000	0.004	0.000	0.320	0.000	0.000	0.000	1.000	0.000	0.000
75%	0.800	0.818	0.733	1.000	0.081	0.075	0.005	0.034	0.020	0.034	0.000	0.630	0.000	0.000	0.000	1.000	0.000	0.000
max	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

4.2.1.4 Baseline Model

The baseline model is typically used to evaluate the accuracy of the AL selection model concerning a single target variable from the dataset. This score is considered the minimum accuracy that the model must attain across all target variables (Ameisen, 2020). The PCP method was used to test the model with an obtained score of 53%, indicating that the model would be deemed inadequate if its actual accuracy fell below 53%.

4.2.1.5 Model Training and Validation

The modelling procedure was applied across 24 production wells, encompassing a collective aggregate of 474,656 samples following an extensive phase of data cleaning and pre-processing. A subset of the original data ranging from 2006 to 2019 exclusively drawn from 16 wells was allocated for the dual purpose of training and validation, accounting for a substantial 429,000 samples. The selection of wells was thoughtfully designed to encapsulate the diversification present across subfields (XF, XFE, XM, XJ, XK, XH, XS), thus ensuring the model's exposure to a representative range of geologic and operational variances. The training phase of the model was developed through a comprehensive dataset comprising 12 wells, each reflecting the details of the installed lifting method within the context of prevailing input parameters and field-specific conditions. On the other hand, the remaining four wells were used to validate the model, playing a pivotal role in gauging the model's efficacy. It is noteworthy that this particular approach diverges modestly from the conventional train-test-split conventions in ML, wherein data partitioning typically adheres to random stratification strategies.

4.2.1.5.1 Training and Validation Dataset

Wells XF161, XF3, XF19 and XF144 belong to the XF block, the largest subfield. It produces light and heavy oil in addition to natural gas. The reservoir has a strong water drive aquifer. Water injection (IOR) is implemented to maintain reservoir pressure. XF19 produce oil using thermal recovery. XF block API ranges from 20 to 36 with initial reservoir pressure of 1600 and 3200 psi in formations B and A, respectively. Oil is being produced naturally and artificially using PCPs.
Wells XFE36 and XFE38 belong to the XFE block, the second-largest subfield. The subfield estimated reserve is 180MM original oil in place (OOIP) of heavy and extra

heavy oil with API ranges from 12-18. The initial reservoir pressure is 530 psi. Oil viscosity is between 700-1500 mPa.s. Thermal recovery is primarily used to reduce fluid viscosity to increase oil production. Thermal EOR, chiefly CSS and steam SF have been implemented in this subfield since 2010. BPUs dominate the lifting methods, with few numbers of PCPs producing CHOPS.

Wells XJ1 and XJ4 are from block XJ with an OOIP of 200 MMSTB and 46 BCF original gas in place (OGIP). Gas injection and N2 injection are the implemented secondary and tertiary recovery methods. Both wells started production naturally before being replaced by AL. The field produces light oil and natural gas from A and B formations. Oil API is 31-37 with an average oil formation volume factor of 1.2 bbl/STB. The initial reservoir pressure is 1700 and 3500 psi in formations B and A, respectively. GL and PCP are the main lifting methods.

Wells XM17 and XM184 were selected from the XM subfield, which produces medium and light oil. The subfield produces approximately 653 Mscf/day of associated gas. The wells produce oil naturally or by GL and then PCP if the reservoir pressure is depleted.

Wells XH1 and XH6 were completed and started production in 2012. N2 injection was applied to their block from 2013 to 2014 due to rapid reservoir energy depletion after one year of production. The oil and reservoir properties are 34 API light oil, initial reservoir pressure and temperature of 2500psi and 170°C, respectively. Both wells started production naturally, then PCP was installed, followed by ESP in both wells. ESPs and PCPs are the primary lifting methods in the XH block.

Wells XK7 and XK21 produce light oil from XK, a relatively small subfield. GL and PCPs are the principal AL.

Wells XSS1 and XSW1 belong to the recently developed XS block that was spud production in 2014. The subfield has high pour point crude (40°C), which usually affects the downhole pump operation if the wells undergo surface equipment maintenance or power failure occurs.

4.2.1.5.2 Model Runs and Hyperparameters Tuning

Hyperparameter tuning is an essential process for optimizing the performance of supervised learning models. One popular method for tuning hyperparameters is through grid search, which involves systematically testing all possible

combinations of hyperparameters within a given range. For instance, in LR, hyperparameters such as regularization parameter and solver can be tuned to improve the model's performance. Similarly, in SVM, the kernel type and gamma value can be tuned to improve accuracy (Hastie, 2009). In KNN, hyperparameters such as the number of neighbours and distance metric can be tuned to improve accuracy. Another approach for hyperparameter tuning is randomised search, which randomly samples hyperparameters within a defined range. DT models have several hyperparameters that can be tuned to improve their accuracy, including the maximum depth of the tree, the minimum number of samples required to split a node, and the minimum number of samples required to be at a leaf node. Similarly, RF models also have a number of hyperparameters that can be tuned, including the number of trees in the forest, the maximum depth of each tree, and the minimum number of samples required to be at a leaf node (Breiman, 2001). This approach shows its effectiveness in improving the prediction accuracy of the models.

In the three models, production, operation, and environmental/economic, hyperparameter tuning was performed on DT and RF models to improve the prediction accuracy. The reason is their highest training and validation scores among the other algorithms using the default hyperparameters. The study found that tuning the hyperparameters of both models significantly improved their accuracy compared to their default settings. For the DT model, the best set of hyperparameters was found to be a maximum depth of 8, while keeping others as default. For the RF model, the best set of hyperparameters was found to be 500 trees, a maximum depth of 7, and a maximum number of features required for the finest split of 7.

4.2.1.5.3 Training and Validation Results

Table 4.4 offers an overview of the attained training and validation accuracies, meticulously evaluated for each algorithm within the AL selection model. Impressively, all algorithms showcased commendable performance both in terms of training and validation phases. The zenith of validation accuracy is observed to be held by LR and RF, demonstrating validation scores of 90% and 92% respectively. Following in this trajectory is DT, establishing its validation accuracy at an appreciable 88.7%. However, in contrast, KNN and SVM were marked by

relatively lower accuracies, each achieving a level beneath 85%. Notably, the assessment of the algorithms' performance extends to subsequent testing outcomes. In view of their superior performance in terms of testing accuracy, RF and DT were elected for validation over new wells data, aligning with the deliberate choice guided by their prominent test accuracy records.

Table 4.4: AL selection model training and validation accuracies using production dataset

Algorithm	Training Accuracy (%)	Validation Accuracy (%)
LR	88.61	90.34
SVM	95.46	84.09
KNN	98.09	72.27
DT	99.80	88.69
RF	99.65	92.86

4.2.1.6 Model Test on New Dataset

The model's efficacy underwent rigorous testing through the utilization of a fresh unseen dataset encompassing eight wells, constituting a substantial assemblage of around 11,600 samples. This new dataset, ranging from 2020 to 2021, exhibits an inherent correspondence with the features incorporated in the training and validation phases. The rationale behind the employment of this contemporary data resides in its ability to inspect the model's adaptability to prevailing and forthcoming field production conditions. Notably, this evaluation framework serves a dual purpose, not only illuminating the model's real-time performance but also foreseeing its efficacy when deployed on future unlabelled datasets. The use of a separate test dataset is imperative to evaluate the performance of the models and prevent overfitting, which could occur should the identical dataset be utilized for both training and validation.

4.2.1.6.1 Test Dataset

Below are the selected wells used to validate test model performance in predicting the target variables (AL). Each well belongs to one of the previously mentioned subfields.

- XFE26 produces heavy crude with thermal recovery (CSS) using BPU.
- XH7 produces medium oil by PCP with neither IOR nor EOR.

- XJ14 produces light oil with GL (no N2 or gas injection).
- XM334 is producing light oil by PCP.
- XF66 is naturally producing light crude with gas injection.
- XF18 produces heavy crude with CSS using MTM_PCP.
- XK9 is producing medium crude with neither IOR nor EOR using PCP.
- XSE2 is producing light crude by ESP.

4.2.1.6.2 Model Test Results

Table 4.5 summarises the modelling accuracies obtained by RF and DT, including precision and recall. The RF exhibited a commendable accuracy of 92.42%, while the DT boasted an accuracy of 93.02%. It is important to acknowledge that the discrepancy in these figures arose from the categorization of BPU and MTMPCP, as well as GL and NF, which will be elucidated further through the ensuing clarification provided by the accompanying confusion matrix visualizations. It is imperative to underscore that the ascertained accuracies are underpinned by a conscientious consideration of data imbalance and the intricate actualities of field conditions.

Notably, both RF and DT demonstrate a test accuracy that emulates their validation accuracy to a significant degree. In comparison, the testing accuracies of SVM, LR, and KNN fluctuate within the range of 58-64%. It is salient to observe that each algorithm, without exception, managed to secure accuracies surpassing the baseline model's threshold of 52%.

Table 4.5: AL selection model test accuracies using production dateset

Algorithm	Accuracy (%)	Recall (%) * **	Precision (%) * **	F1 Score (%) * **
LR	62.21	62.21	62.21	62.21
		61.21	67.41	55.77
SVM	58	58	58	58
		49.8	54.36	47.73
KNN	63.3	63.3	63.3	63.3
		48.16	52.71	49.19
DT	93.02	93.02	93.02	93.02
		92.37	84.6	85.25
RF	92.42	92.42	92.42	92.42
		75.38	77.72	75.32
**Micro **Macro average values*				

4.2.1.7 Validation with Field Data Results and Discussion

The results were validated with the actual lifting methods used in the field, considering that they are the optimum lifting methods according to the designated field selection screening and the production performance simulation.

The accuracy of both RF and DT in training and testing is demonstrated in **Fig. 4.4** and **4.5**. DT and RF obtained their highest test accuracy at a maximum tree depth of 8 and 7, respectively. It is important to note that the obtained accuracy considers data imbalance and actual field conditions. **Fig. 4.6** is the classification report (confusion matrix), illustrating the predicted and actual AL used in the field. We can see that both algorithms effectively predicted BPU, PCP, and ESP. The prediction error of 7.5% resulted in thermal recovery pumps, namely BPU and MTM_PCP, while the 7% error is in GL and NF classification. The wrong prediction of thermal AL is because MTM_PCP has the lowest weight in the dataset; nevertheless, both algorithms predicted it for CSS. The BPU was used in both SF and CSS, while MTM_PCP was only used in CSS wells in the field; thus, the model selected MTM_PCP as more appropriate than the BPU considering input parameters. The RF classifier predicted GL instead of the actual NF label due to the similarity of gas and GOR features. Additionally, NF has approximately the same amount of cumulative GL oil before reservoir energy is depleted and replaced by another lifting method. The DT classifier predicted GL for the actual ESP actual because some ESP wells produce a small amount of gas, which the algorithm principally uses to classify the AL.

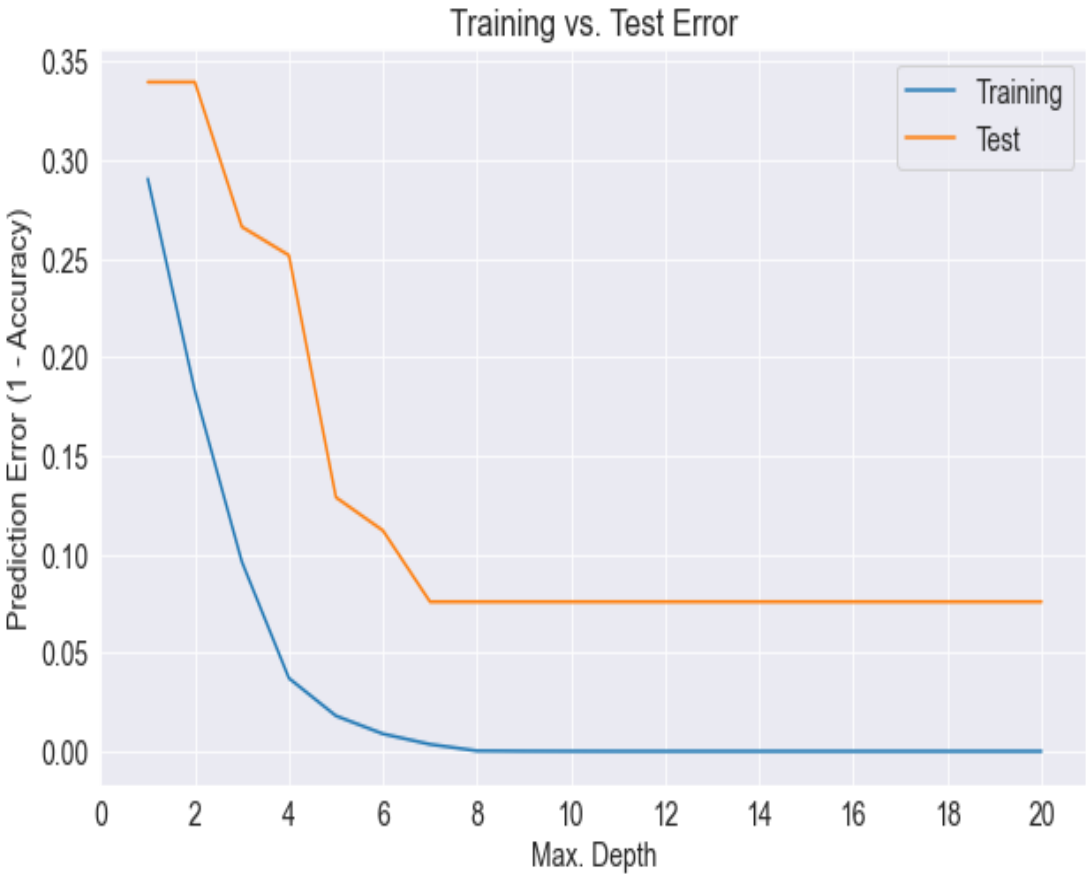


Fig. 4.4: RF training vs. test prediction error

Fig. 4.5: DT training vs. prediction error using production dataset

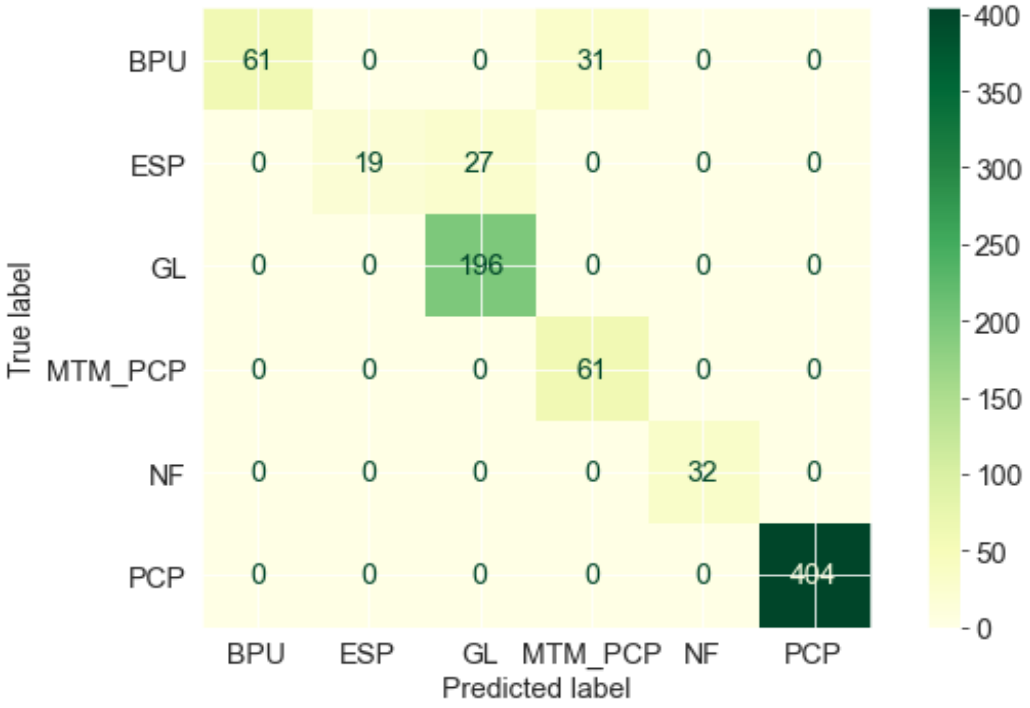


Fig. 4.6: DT confusion matrix (True label vs Predicted label)

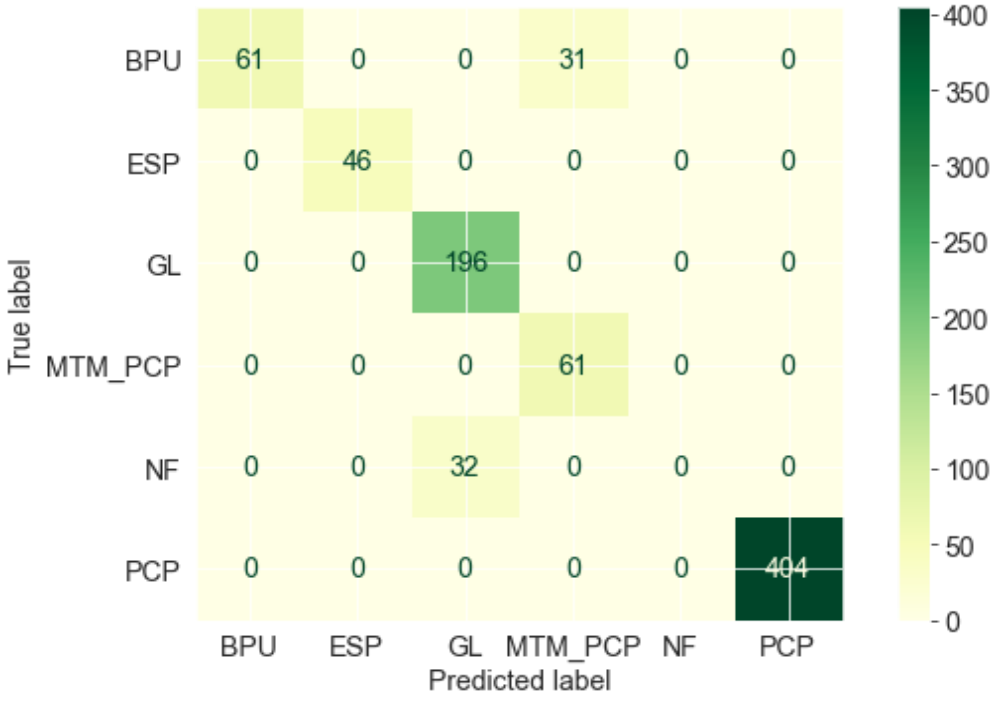


Fig. 4.7: RF confusion matrix (True label vs Predicted label)

4.2.1.8 Model Test on Unlabelled Dataset

The model was tested an unlabelled dataset devoid of any specific AL attributes. Sample datasets, replete with identical features, were utilised for the purpose of AL prediction. The model's predictions regarding the most optimal AL were subsequently cross-validated against the authentic AL methods currently in practice within the field. Notably, this validation process yielded an outstanding accuracy rate of 100% in the prediction of 10 sample points pertaining to BPU. This remarkable degree of accuracy holds promise for prospective AL selection activities in the oil and gas industry.

4.2.2 Selection Model Based on Operation Data

4.2.2.1 Input Parameters and Data Visualisation

The model's inputs encompass the parameters related to completion and workover operations, in addition to the distinct attributes of the well and the geological formation. These facets stand in contrast to the nature of production data, which is recorded on a daily basis. The operation parameters, serving as input variables, are meticulously documented across distinct temporal intervals that encompass multiple stages such as drilling, completion, workover, well testing, stimulation, and the implementation of IOR and EOR techniques. The recording frequency of operation data varies from each subfield to another contingent upon field development plan (FDP). The model encapsulates a total of eleven distinctive features, each of which plays a role in influencing the AL selection process. These features are outlined in **Table 4.6**.

Table 4.6: Operation model features

Feature	Unit
Completion/Workover dates	Year
Well true vertical depth (TVD)	Feet
Plug back total depth (PBTD)	Feet
AL setting depth	Feet
AL running period	Days
Mid of perforation (formation) depth	Feet
Production zone thickness	Feet

Tubing size	Inch
Workovers frequency over production years	Number
Workover cause	Category
AL failure and replacement cause	Category

Fig. 4.8 shows AL distribution of operation dataset which is dominated by PCP. BPU, GL and NF have the same distribution followed by ESP and MTMPCP to have the lowest weight of operation in the specific field. ESP has a higher average running period above 250 days than the other ALs, as shown in **Fig. 4.9**. The reason is that ESP wells produce with no EOR/EOR implementation that requires frequent shutdowns for pull-out-of-hole (POOH) and run-in-hole (RIH) strings. Other lifting methods have an average run period of 100-150 days.

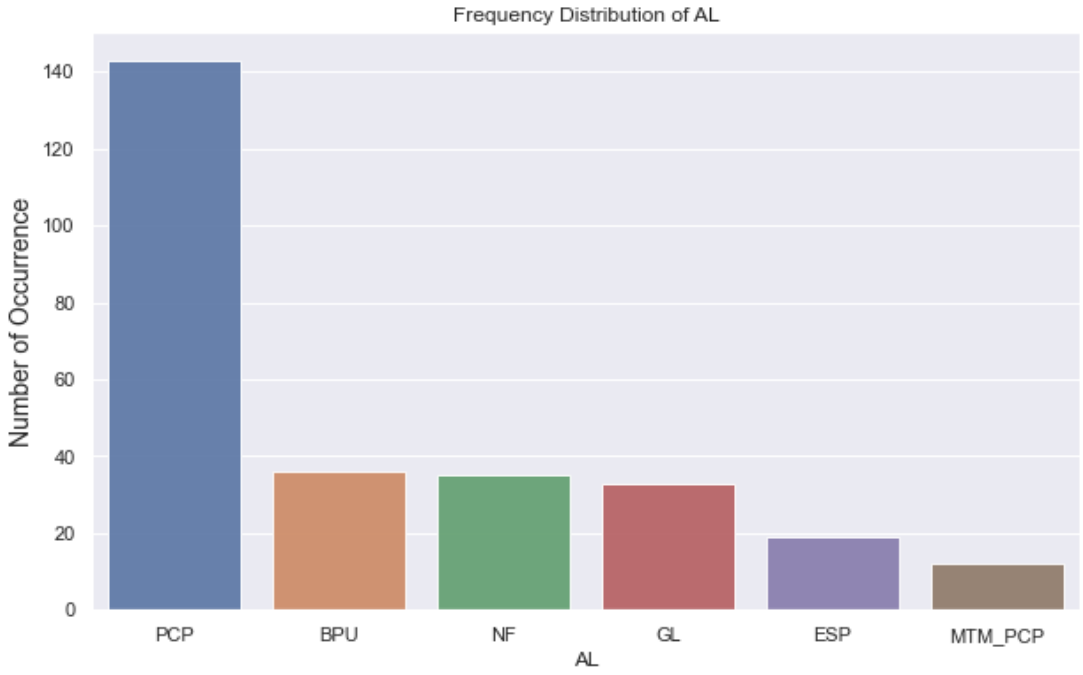


Fig. 4.8: Distribution of AL in operation dataset

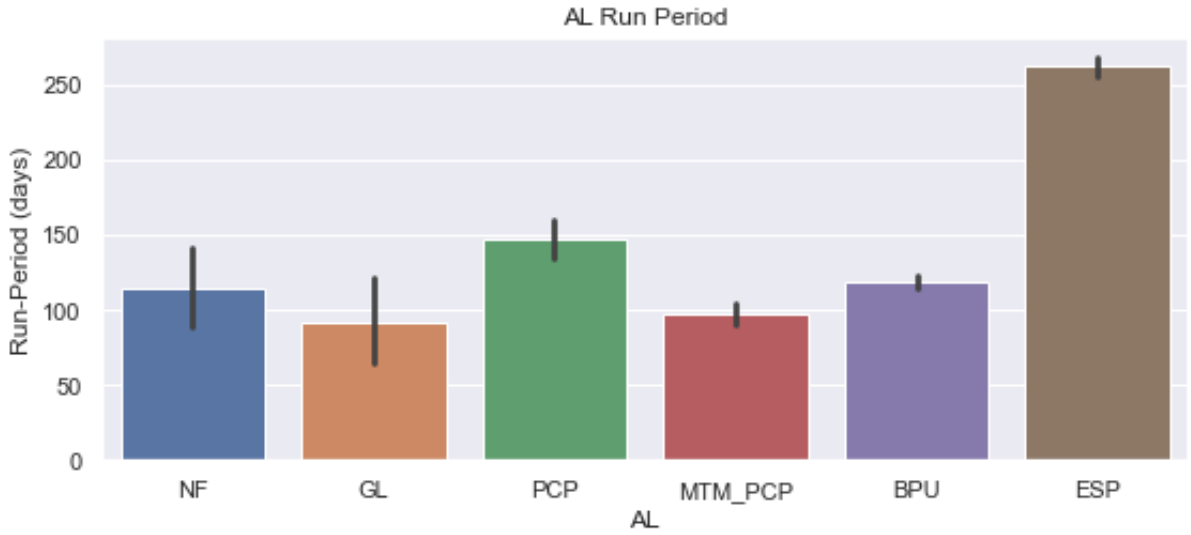


Fig. 4.9: Average AL in run life

Fig 4.10 illustrates the workovers occurred in each production year from 2006 to 2021. This information is essential to evaluate the effectiveness and efficiency of the AL methods used in the field. As we can see in the figure, the year 2008 had the ideal production with zero workover. The failure and workovers then gradually raised until 2017 which experienced approximately 79 total workovers dominated by PCP and PBU. The number then started to decrease after the company implemented DIFA (dismantle, inspection & failure analysis) which resulted in less failures in 2020 and 2021 (workovers reduced to 35 in 2020 and 4 in 2021).

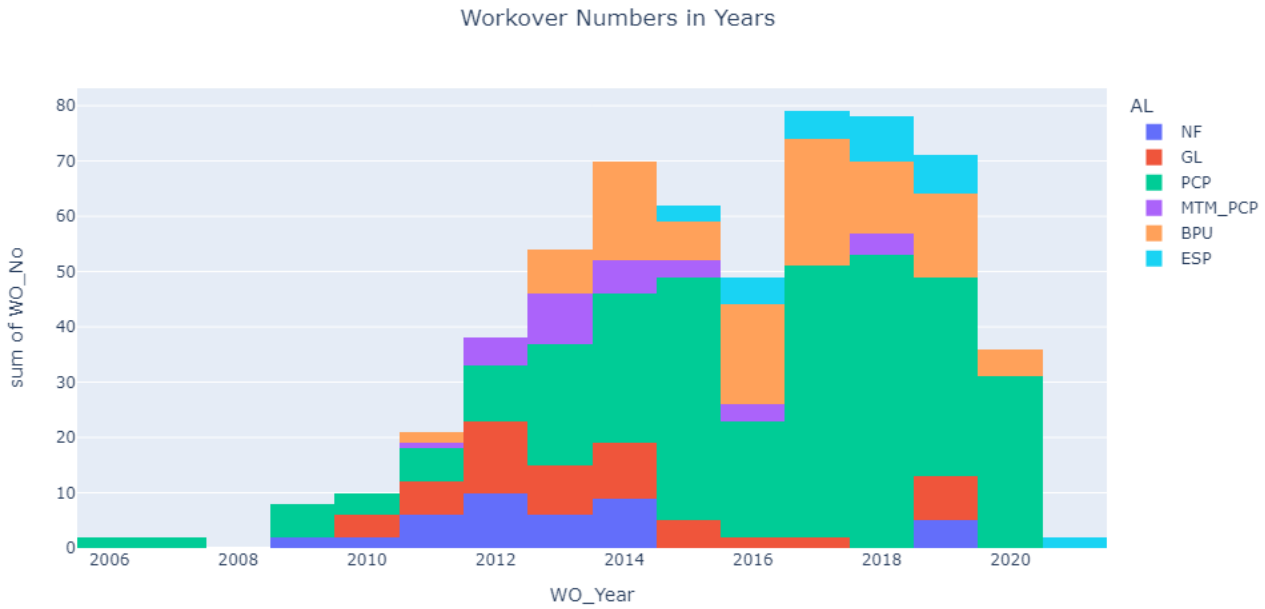


Fig. 4.10: Workover occurrence for each AL from 2006 to 2021

In addition to DIFA, a series of actions were carried out before the well underwent workover to reduce the cost as much as possible. Some examples of these actions include performing a pressure build-up test to ensure there is no fluid flow to the surface, checking if the hex shaft of PCP is broken, interpreting the BPU dyna-card reading in terms of fluid weight and wire tension, interpreting ESP amp chart readings, injecting hot water to flush the downhole pump in sandy wells, injecting diesel to remove plugging in viscous oil. Lastly, ensure no back pressure at surface flowlines by measuring the pressure difference between the wellheads and flowlines. Workovers are also performed to reperforate, change, or add new production zones, test low-productive zones, and shut off high-water and sand production formations. The retesting is usually applied at a sudden increment in the water cut and low production. The pump is also deepened due to low reservoir deliverability in some wells to avoid dry AL running, which will damage the lifting systems. **Table 4.7** summarises the common workover and failure causes across the dataset in the specific field from 2005 to 2021.

Table 4.7: Common workover and failure causes recorded in the dataset

Common Workover Causes	No of Occurrence	Common Failure Causes	No of Occurrence
No Prod	108	None (CSS cycle)	88
Completion	63	Low prod	72
Reperforate (new zone, water shutoff, low production)	44	Stator damage	55
CSS cycle	25	Rod stuck	15
Change pump	24	Tubing damage	13
Retesting (low prod, high WC%)	10	Rod disconnection	12
Pump check	2	High WC%	10
Pump deepen	2	Pump plugged (by sand)	8
-	-	Motor failure	3
-	-	Pump failure	2

4.2.2.2 Data Correlation

Fig. 4.11 illustrates interconnection within the operation dataset, revealing the associations between input and output parameters. Notably, each of the lifting

methods exhibits a favourable positive correlation with depth attributes, except for MTMPCP and BPU with -0.2 and -0.5, which instead demonstrate correlations ranging from 0.1 to 0.6 with tubing size, workover occurrence, and the duration of workover activities. Intriguingly, PCP displays a positive correlation of 0.48 with stator failure. Additionally, GL and NF demonstrate a constructive correlation of 0.33 and 0.12 respectively with zone thickness, a key determinant of the volume of oil to be produced. Furthermore, the feature denoting setting depth emerges as strongly linked with ESP and PCP, signifying its noticeable role in the classification process, as elucidated further in Chapter 6.

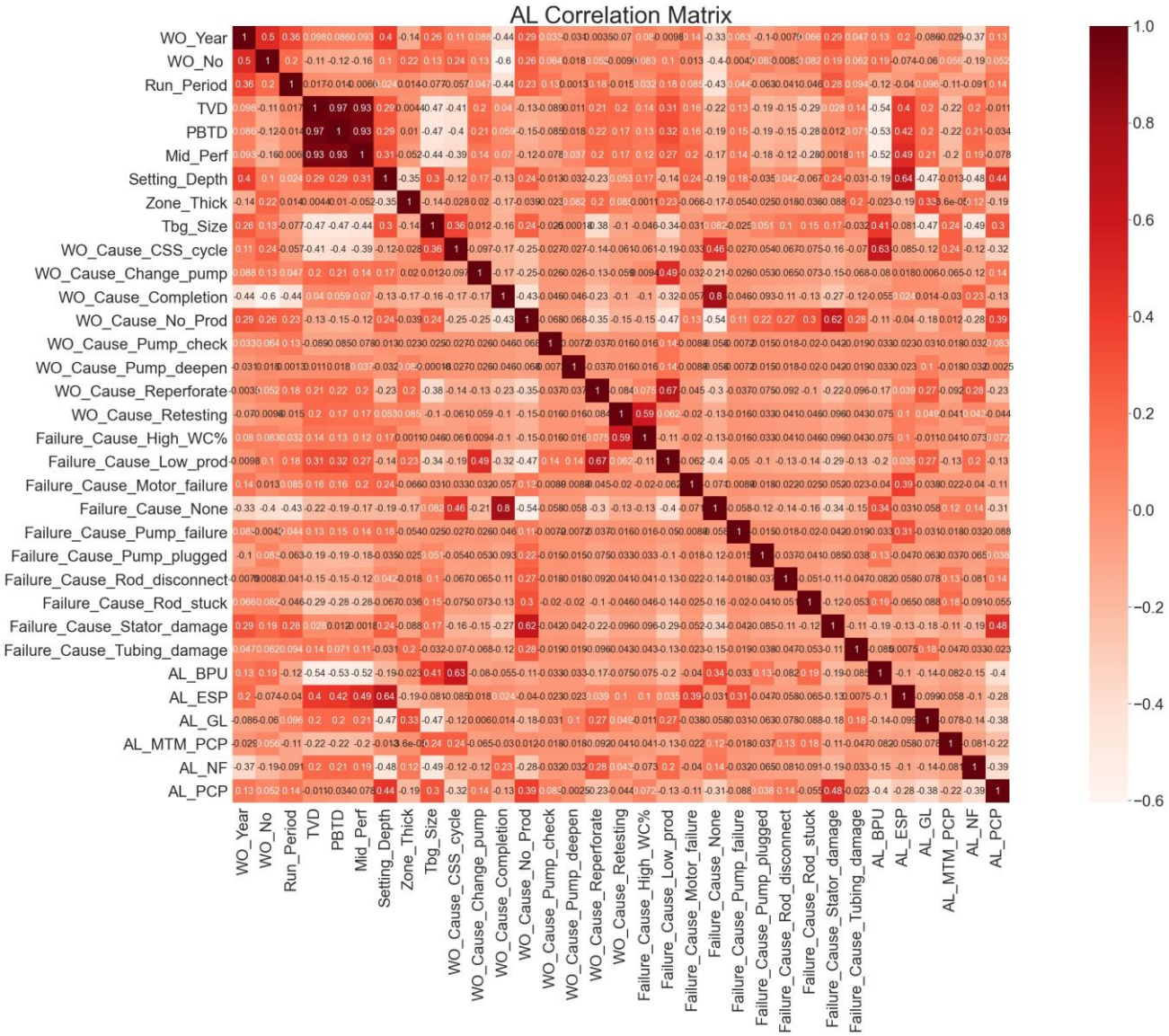


Fig. 4.11: Operation features correlation matrix

4.2.2.3 Statistical Data Analysis

Tables 4.8 and **4.9** provide a comprehensive overview of the depth ranges, operational durations of AL, occurrences of workover interventions across production years, and the dimensions of the tubing employed within the model. Furthermore, these tables also show the statistical details encompassing the mean values and standard deviations of the operation-related attributes.

Table 4.8: Operation features statistical analysis before encoding and normalisation

	WO_Year	WO_No	Run_Period	TVD	PBTD	Mid_Perf	Setting _Depth	Zone_ Thick	Tbg_Size
mean	2014.209	2.0936	476.295	2142.372	2014.299	1690.621	813.275	16.253	3.689
std	3.362513	1.877	586.489	1023.451	969.743	862.518	639.083	14.927	0.635
min	2005	0	0	700	628.5	522	0	2	2.875
25%	2012	1	43.75	1453.5	1355.953	1172	460.333	8	3.5
50%	2014	2	270.5	2200	2065.485	1578	864.835	12.75	3.5
75%	2017	3	627.25	3000	2680.89	2110	1156.65	17.875	4.5
max	2021	8	3910	3800	3728.86	3676	2606.28	101	4.5

Table 4.9: Operation features statistical analysis after encoding and normalisation

	WO_Year	WO_No	Run_Period	TVD	PBTD	Mid_Perf	Setting_Depth	Zone_Thick	Tbg_Size	WO_Cause_CSS_cycle	WO_Cause_Change_pump	WO_Cause_Completion	WO_Cause_No_Prod	WO_Cause_Pump_check	WO_Cause_Pump_deepen	WO_Cause_Reperforate	WO_Cause_Retesting	Failure_Cause_High_WC%	Failure_Cause_Low_prod	Failure_Cause_Motor_failure	Failure_Cause_None	Failure_Cause_Pump_failure	Failure_Cause_Pump_plugged	Failure_Cause_Rod_disconnect	Failure_Cause_Rod_stuck	Failure_Cause_Stator_damage	Failure_Cause_Tubing_damage
mean	0.576	0.262	0.122	0.465	0.447	0.371	0.312	0.144	0.501	0.090	0.086	0.227	0.388	0.007	0.007	0.158	0.036	0.036	0.259	0.011	0.317	0.007	0.029	0.043	0.054	0.198	0.047
std	0.210	0.235	0.150	0.330	0.313	0.273	0.245	0.151	0.391	0.287	0.281	0.419	0.488	0.085	0.085	0.366	0.187	0.187	0.439	0.104	0.466	0.085	0.167	0.204	0.226	0.399	0.212
min	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

25%	0.438	0.125	0.011	0.243	0.235	0.206	0.177	0.061	0.385	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
50%	0.563	0.250	0.069	0.484	0.463	0.335	0.332	0.109	0.385	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
75%	0.750	0.375	0.160	0.742	0.662	0.503	0.444	0.160	1.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000
max	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

4.2.2.4 Baseline Model

The baseline model assessment was also utilized to anticipate the score that could be achieved on the operation dataset by the model. When using only PCP, the model attained a score of 66.6%. This means that the model would be regarded as unsatisfactory if its true accuracy is below 66.6%.

4.2.2.5 Model Training and Validation

This model utilized a total of 100 wells encompassing 6700 datapoints, out of which 64 wells were chosen for model training and validation, while the remaining 36 wells were used for testing the model's performance. The training and validation wells were divided into 50 wells for training, comprising 4464 samples, and 14 wells with 1080 samples for validation. This method demonstrated its effectiveness in the production data model results.

4.2.2.5.1 Training and Validation Dataset

The 64 wells used to train and validate this model were randomly selected according to the oilfield capacity. To enhance the modelling results, additional wells were included to the wells used in production model to augment the dataset size since the operation data is not recorded daily. The overall distribution of wells are 10 wells each from the great XF field, the second largest XFE, and XJ field, and then 5 from XM, XK, XH, XS fields. The wells are listed as follows:

- HJ wells: XJS1, XJS4, XJC1, XJC2, XJS2, XJS7, XJS24, XJS29, XJS10, XJS38.

- HFN wells: XFN1, XFN4-2, XFN23, XFN49, XFN53, XFN80, XFN144, XFN161, XFC19, XF3.
- XFE wells: XFE4, XFE8, XFE12, XFE18, XFE36, XFE38, XFE39, XFE48, XFE81, XFE105, XFE20.
- XM wells: XM13, XM1-4, XM1-8, XM1-12, XM9, XM10, XM17, XM18-4, XM33-3, XM33-5.
- XK wells: XK1, XK2, XK3, XK4, XK6, XK11, XK21, XK23, XKS2, XKN9.
- XH wells: XH1, XH6, XH9, XHN1, XHN8, XHN9, XH5.
- XS wells: XSS1, XSE1, XSW1, XSFN1, XSE2, XS3, XSW6.

The validation wells selected from the above list are XJS10, XJS38, XFN49, XF3, XFE12, XFE81, XM1-3, XM9, XK2, XK23, XH5, XH9, XSFN1, and XSS1. The reservoir characteristics and fluid properties of each block where the wells belong are mentioned earlier in Section (4.2.1.3.1).

4.2.2.5.2 Training and validation Results

Table 4.10 provides a concise overview of the training and validation outcomes. Remarkably, all algorithms demonstrated an accuracy surpassing 80%. Particularly noteworthy is the exceptional performance of DT and RF, which achieved the highest accuracy levels in training, reaching 99.2%. During validation, DT exhibited an accuracy of 91.7%, while RF achieved 94.4%. Given the superior performance of RF and DT, they were chosen to undergo testing with the new dataset, a selection based on their optimal results.

Table 4.10: AL selection model training and validation accuracies using operation dataset

Algorithm	Training Accuracy %	Validation Accuracy %
LR	81.4	80.5
SVM	81	80.6
KNN	84.3	86.1
DT	99.2	91.7
RF	99.2	94.4

4.2.2.6 Model Test on New Dataset

The subsequent phase involves evaluating the model's performance on an unseen dataset comprising 36 wells, which consists of 2450 samples and has similar characteristics to those used in the training and validation stages, spanning the period between 2020 and 2021.

4.2.2.6.1 Test Dataset

To ensure a comprehensive comparison and evaluation of the results, a minimum of five wells from each block were carefully selected. These wells were specifically chosen to match those used in the production model, as well as to be used in environmental/economic model as this approach allows for a more reliable assessment of the performance of the model. By using the same wells, the sensitivity and importance of input features can be measured, which in turn provides more insightful information for future model development and improvement. The list of selected wells is provided below:

- XJS9, XJS14, XJS13, XJS24, XJS34, produce using GL, NF, and PCP with N2 and gas injection.
- XFN148, XFC18, XF1, XF18, XFN17, XFN166, produce using PCP, MTMPCP, and NF with WI and gas injection.
- XFE22, XFE26, XFE40, XFE46, XFE84, produce using PCP and BPU with CHOPS, CSS, and SF.
- XM1-6, XM7, XM21, XM10-2, XM33-4, produce using PCP and NF
- XK9, XK20, XK22, XKS4, XKN12, produce using GL and PCP with gas and N2 injection.
- XHN7, XHN2, XH3-1, XH3-2, XHN6, produce light oil using ESP and PCP
- XSFE1, XSW2, XSFE2, XHG1, XSW7, produce using ESP with no IOR/EOR

4.2.2.6.2 Model Test Results

RF achieved a higher accuracy score of 91.5% compared to DT with a score of 89.5%. These results suggest that RF was able to identify the lifting method for a larger number of wells with greater accuracy. However, it is important to note that both models were able to predict the majority of the data accurately, indicating that they are both effective methods for AL prediction. SVM, LR, and KNN testing accuracies are between (78-82%), higher than those scored from production data.

Nevertheless, one algorithm has an accuracy above 90% in operation model. **Table 4.11** presents the accuracies obtained by each algorithm, including precision, recall, and F1 score.

Table 4.11: AL selection model test accuracies using operation dataset

Algorithm	Accuracy (%)	Recall (%) * **	Precision (%) * **	F1 Score (%) * **
LR	81.7	81.7	81.7	81.7
		67.9	64.5	65.5
SVM	79.7	79.7	79.7	79.7
		66.3	63.2	63
KNN	78.4	78.4	78.4	78.4
		64.2	60.5	62
DT	89.5	89.5	89.5	89.5
		83.5	83.3	83.4
RF	91.5	91.5	91.5	91.5
		91.25	85.2	87.5
**Micro **Macro average values*				

4.2.2.7 Validation with Field Data Results and Discussion

The accuracy scores and model predictions of RF and DT used for AL selection are shown in the performance charts and confusion matrices, **Figs. 4.12-4.15**. RF obtained the highest score at a max depth of 5 (**Fig. 4.12**), while DT required a max depth of 7 to score its highest accuracy (**Fig. 4.13**). Further analysis of the confusion matrix for the two models reveals interesting insights into the model predictions. For both models, they correctly predicted the use of BPU, ESP and PCP, while underpredicting the use of MTMPCP, GL and NF. This suggests that the model may be biased towards these two methods, which is due to the similarity of specific features such as well and formation depth and tubing size that are more indicative of these methods. The RF model appears to be more balanced in its predictions, with a more even distribution of correct and incorrect predictions across all lifting methods. This suggests that the RF model is more robust and less prone to overfitting or bias towards specific features.

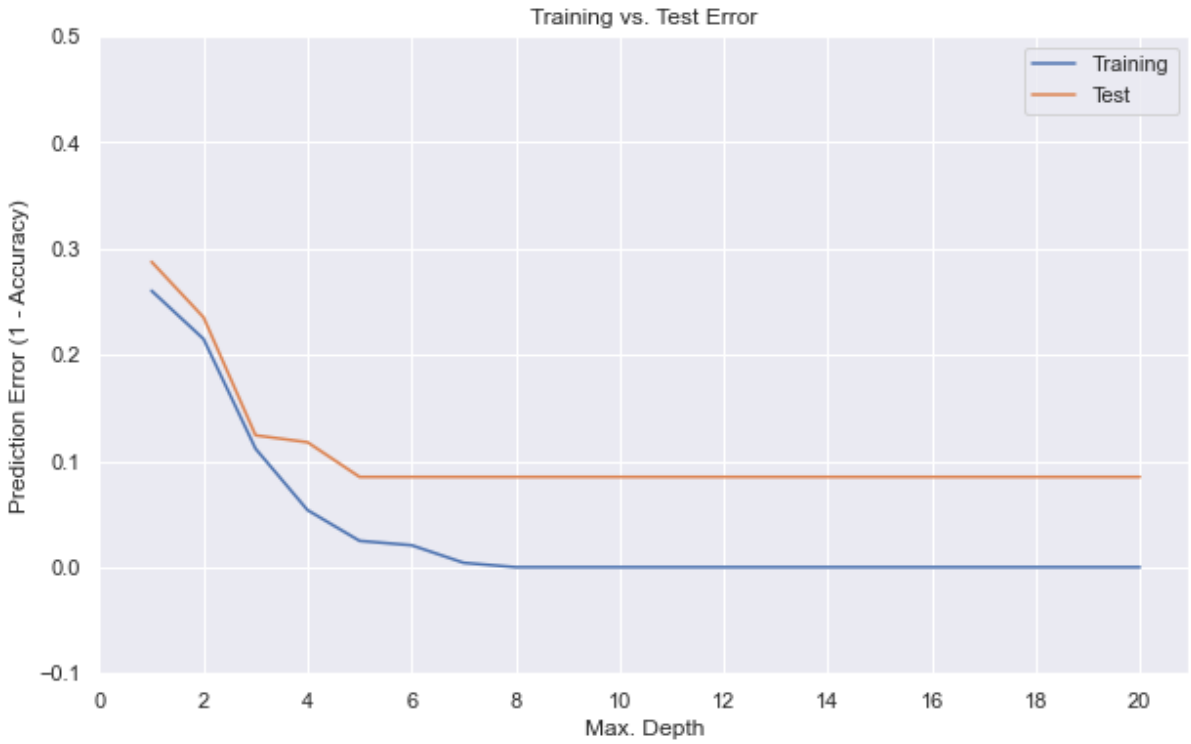


Fig. 4.12: RF AL selection training vs. test Accuracy using operation dataset

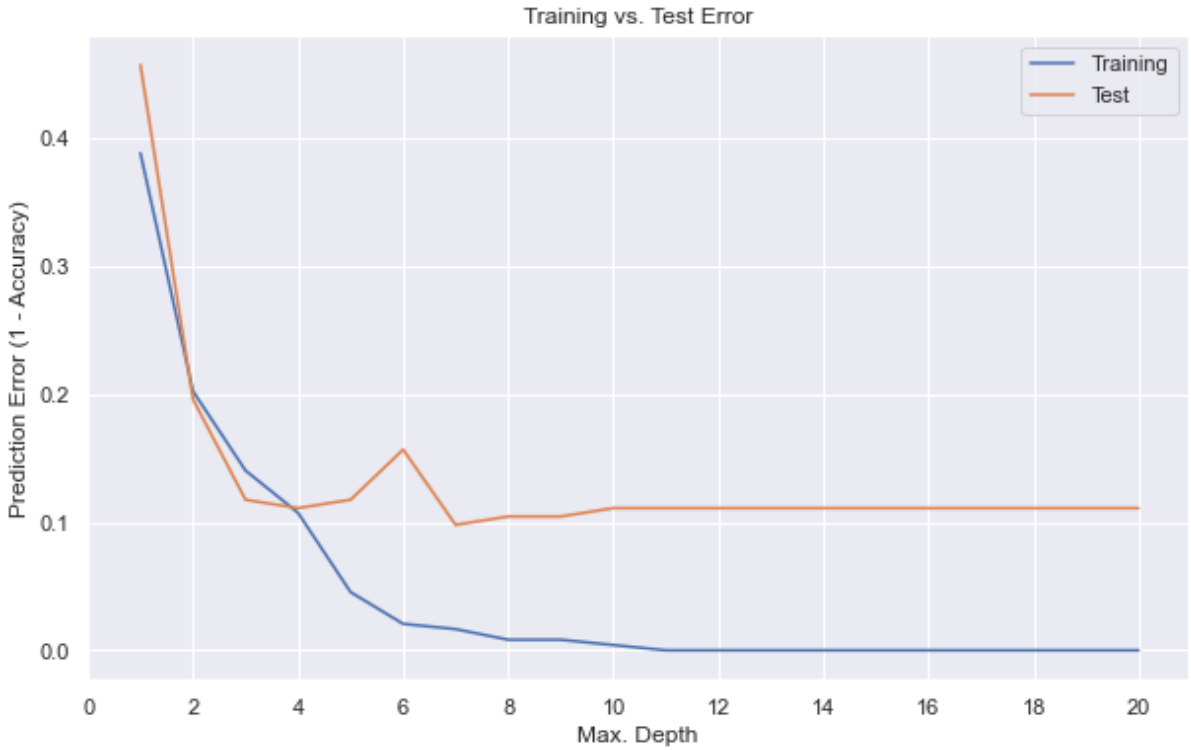


Fig. 4.13: DT AL selection training vs. test Accuracy using operation dataset

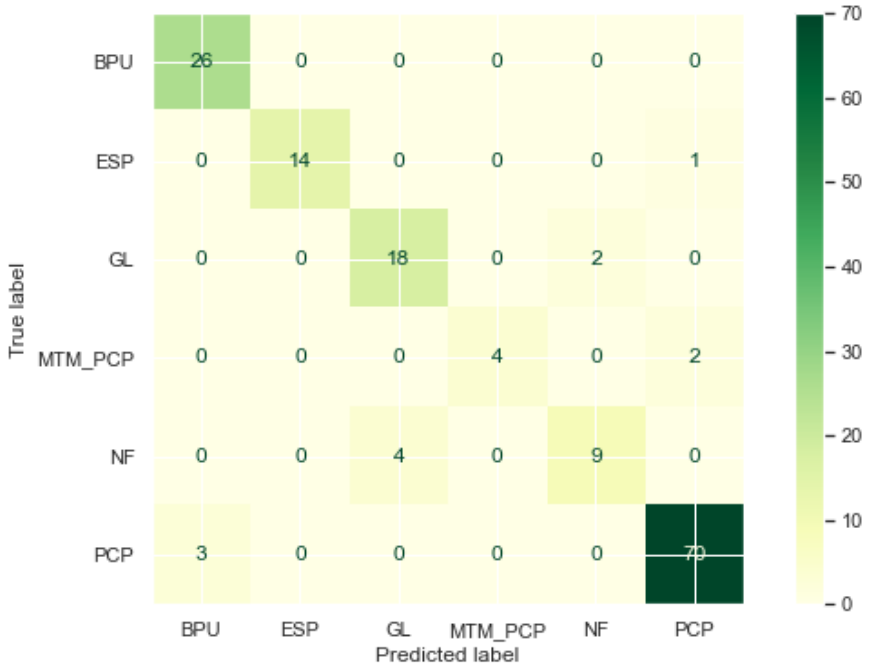


Fig. 4.14: RF AL selection test classification report using operation dataset

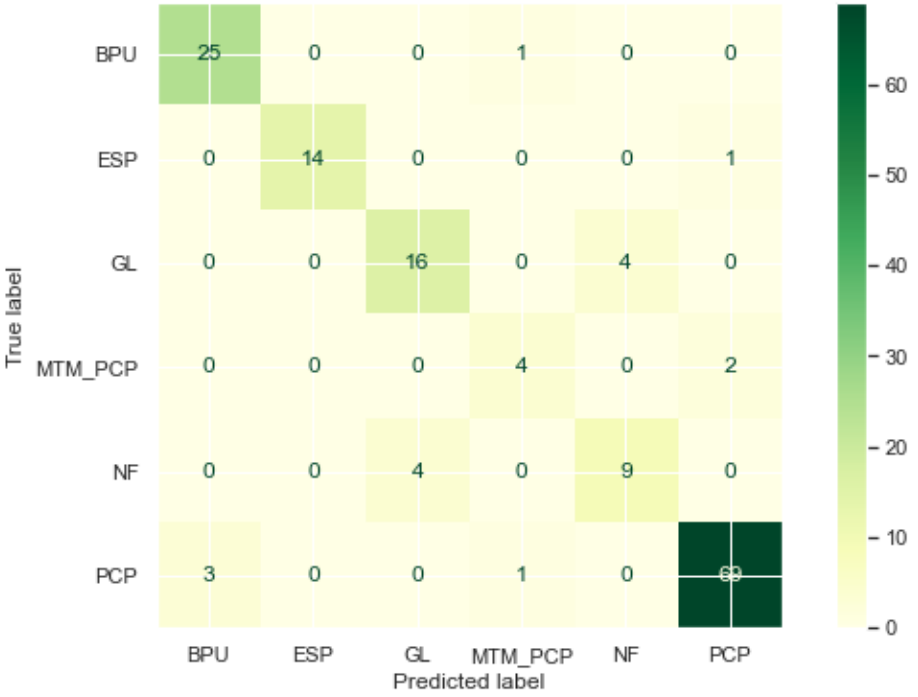


Fig. 4.15: DT AL selection test classification report using operation dataset

4.2.2.8 Model Test on Unlabelled Dataset

After training the model to predict AL on labelled data, it was applied to unlabelled datapoints with 10 samples of PCP and BPU. Specifically, the model predicted AL for the unlabelled datapoints with 70% accuracy, lower than that obtained from production model. This is an indicator of production factors potential for use in AL selection in the oil and gas industry.

4.2.3 Selection Model Based on Economic and Environmental Data

4.2.3.1 Input Parameter and Data Visualisation

The parameters used in this model are a combination of economic and environmental data. The dataset contains eight features presented in **Table 4.12**.

Table 4.12: Environmental and economic model features

Feature	Unit-Type-Level
AL purchasing cost (surface and downhole equipment)	USD
Completion cost	USD
Workover cost	USD
Power source	Gas, Electricity, Natural
Gas emission	Low, Medium, High
Oil spill	Low, Medium, High
Noise	None, Low, Medium, High
Operator knowledge of AL maintenance and operation	No-act, Poor, Good, Excellent

The consideration of AL procurement holds substantial importance in the process of AL selection. This cost factor exhibits variability across different vendors, as depicted in **Fig. 4.16**. Notably, the pricing of PCPs and BPUs falls within the range of 60,000 to 80,000 USD. In contrast, ESPs emerge as the costliest AL option, commanding a price as high as 200,000 USD. It is apparent that the cost associated with GL is comparatively lower, quoting around 33,000 USD, notwithstanding the necessity for a gas source. X-mass trees, on the other hand, present a cost-effective choice, priced at approximately 2,500 USD. Notably, these trees find relevance exclusively in scenarios involving naturally flowing wells.

Fig. 4.16: AL surface and downhole units purchase price in USD

Each lifting method's completion and workover cost depends on many factors, from well location to downhole conditions. Meanwhile, some significant factors such as well depth, formation depth, and fluid properties determine the amount of money to be spent and the time required to resume production.

The power sources available in the specific field are gas for gas-lifted wells and electricity for PCP, BPU, and ESP. The naturally flowing wells do not require a power source since the wells produce using reservoir drive mechanisms.

Environmental aspects consideration is inevitable in AL selection due to the strict policies against the oil and gas industry. Moreover, the world is reducing the GHG effect and heading towards zero-emission. In order to avoid emission tax payments and contamination charges, AL gas leaks and oil spills must be considered because they directly impact humans, fauna, and flora. These emissions and spills endanger the life of plants and livestock and directly impact humans by inhaling the toxic gases and indirectly through the food digestion of the infected livestock. Three environmental aspects parameters have been selected, gas emission, oil spill, and noise. The environmental features were categorised into three levels (low, medium, and high), in addition to none for noise, according to the amount that each lifting method encounters.

As illustrated in **Figs 4.17-4.19**, GL exhibits the highest levels of gas emissions, followed by NF. BPU records notable oil spill and noise levels, a consequence

attributed to stuffing box leakage brought forth by the alternating up and down stroke movements and the activity of surface unit prime movers. PCPs exhibit a moderate performance across various levels. Comparatively, ESP emerges as an environmentally favourable AL option, yet its operation necessitates a substantial electricity supply, thus leading to escalated energy consumption.

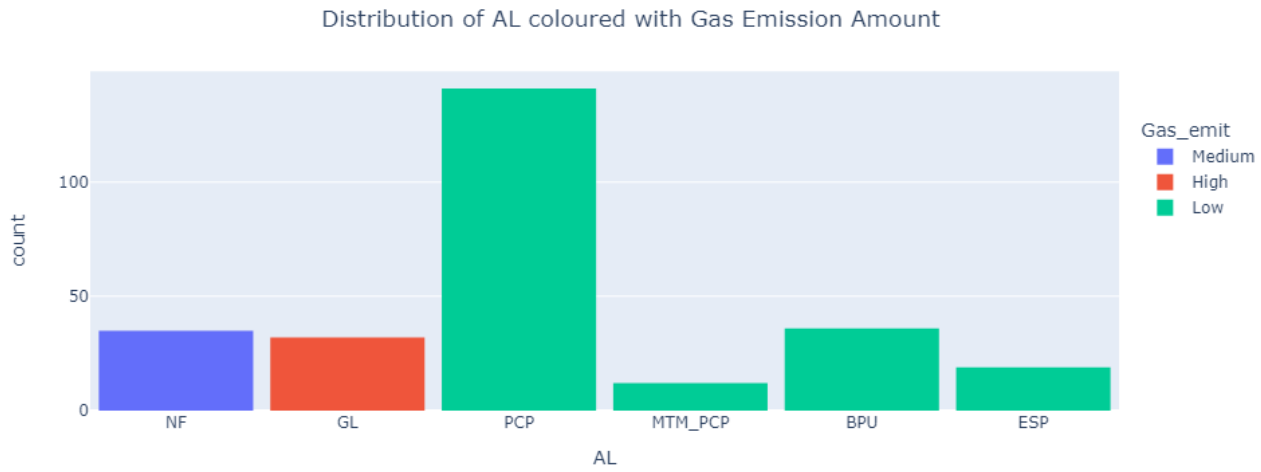


Fig. 4.17: AL gas emission levels

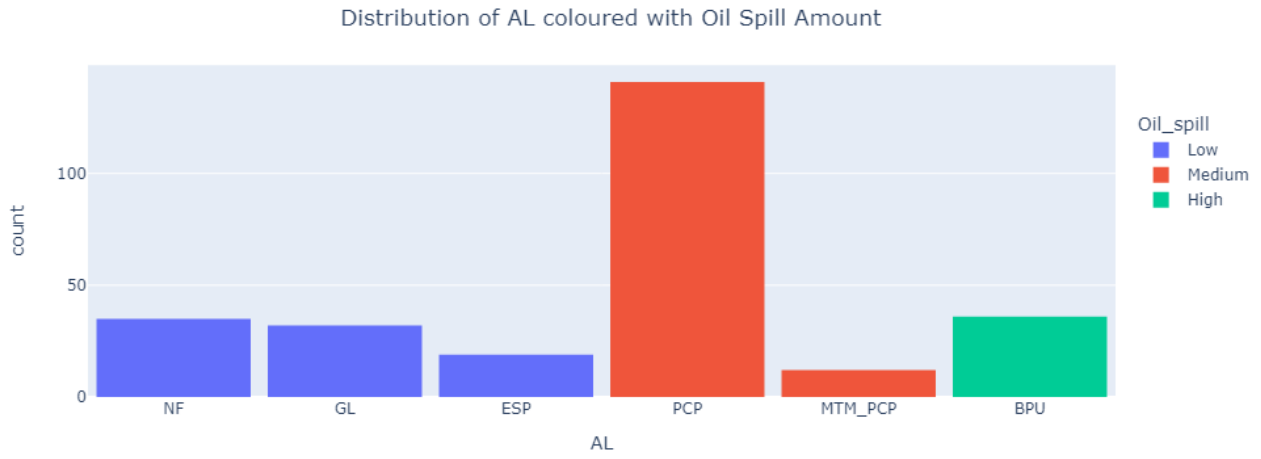


Fig. 4.18: AL oil spill levels

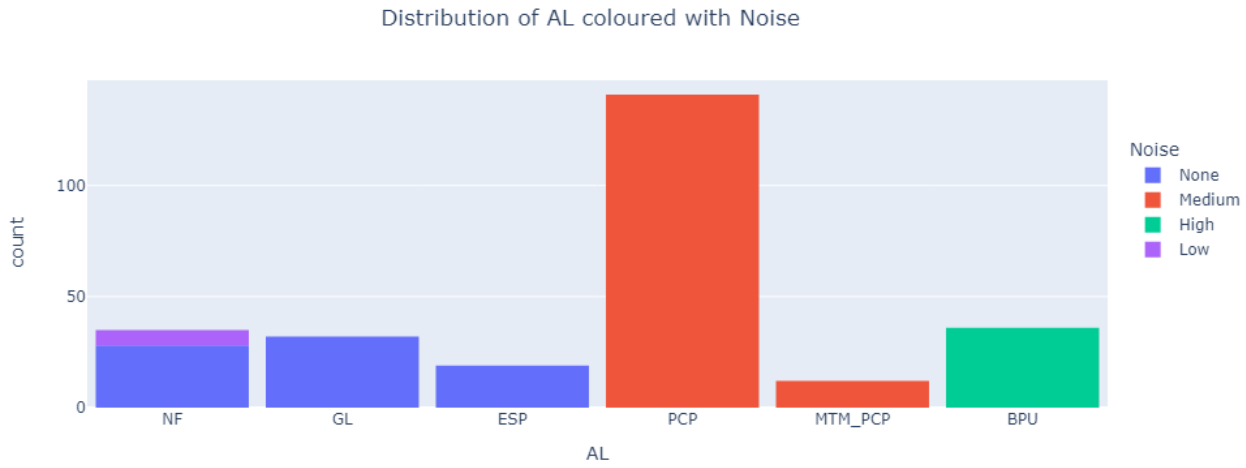


Fig. 4.19: AL noise levels

Operator knowledge is an essential factor in AL selection, principally in remote areas. The familiarity of field operators with any AL aids in early failure diagnosis when the well stops production. For example, broad failures such as electrical and surface equipment problems like motor faults, belt cuts, VSDs and dashboards malfunction, lubricant leaks, overheating, and stuffing box leaks can be resolved by field operators regardless of any need for further workovers. In addition, the operators perform pressure build-up tests for NF and GL wells the moment the pressure drops. The knowledge of field personnel in dealing with such issues can save thousands of oil barrels and sustain the production. In the dataset, the operator knowledge to each AL is categorised into four levels: no-act, which refers to the zero acquaintance of AL operation and failure complications, and the vendor takes complete responsibility for AL maintenance. Poor means the field operator has some knowledge of the selected AL. Good means the field operator is familiar with AL operation and common failure issues. Excellent means that the operator can operate the AL, fix surface problems, and restart it with no workover intervention unless the issue occurs downhole. Further demonstration is shown in **Fig. 4.20**.

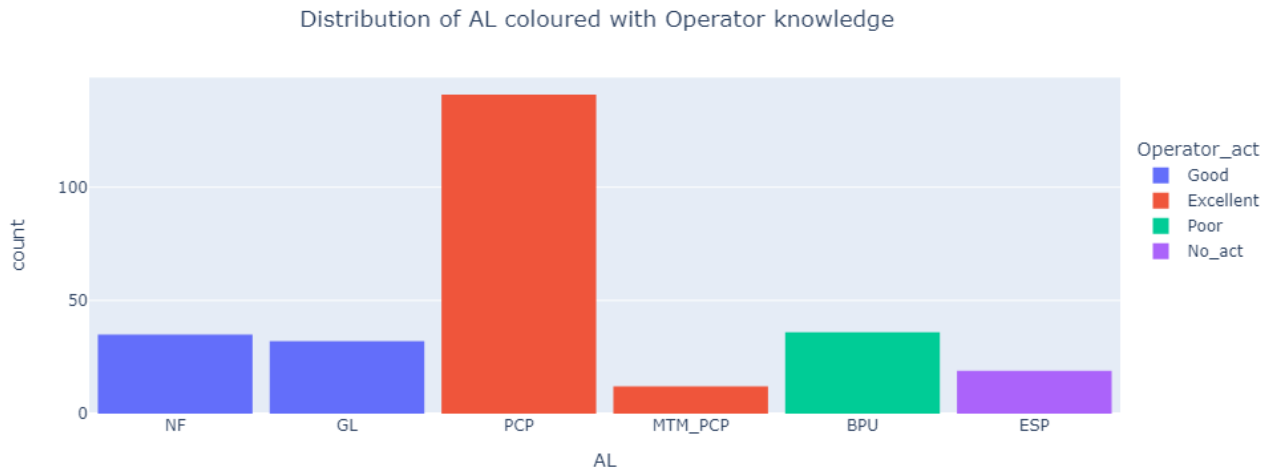


Fig. 4.20: Operator familiarity to AL

4.2.3.2 Data Correlation

Fig. 4.21 illustrates the interrelation between the input variables and the target AL. Notably, the environmental and economic factors exhibit pronounced correlation, surpassing that observed in the production and operation parameters. PCP distinctly aligns with the operator's expertise, as well as oil spill and noise indicators. In contrast, GL and NF impeccably correlate with power sourcing, gas emissions, and operator familiarity marking 0.92. Additionally, BPU showcases a strong connection of 1 with elevated oil spill and noise levels. Conversely, ESP displays a positive correlation of 0.81 with pricing dynamics and a correlation of 1 the field operator's familiarity gaps.

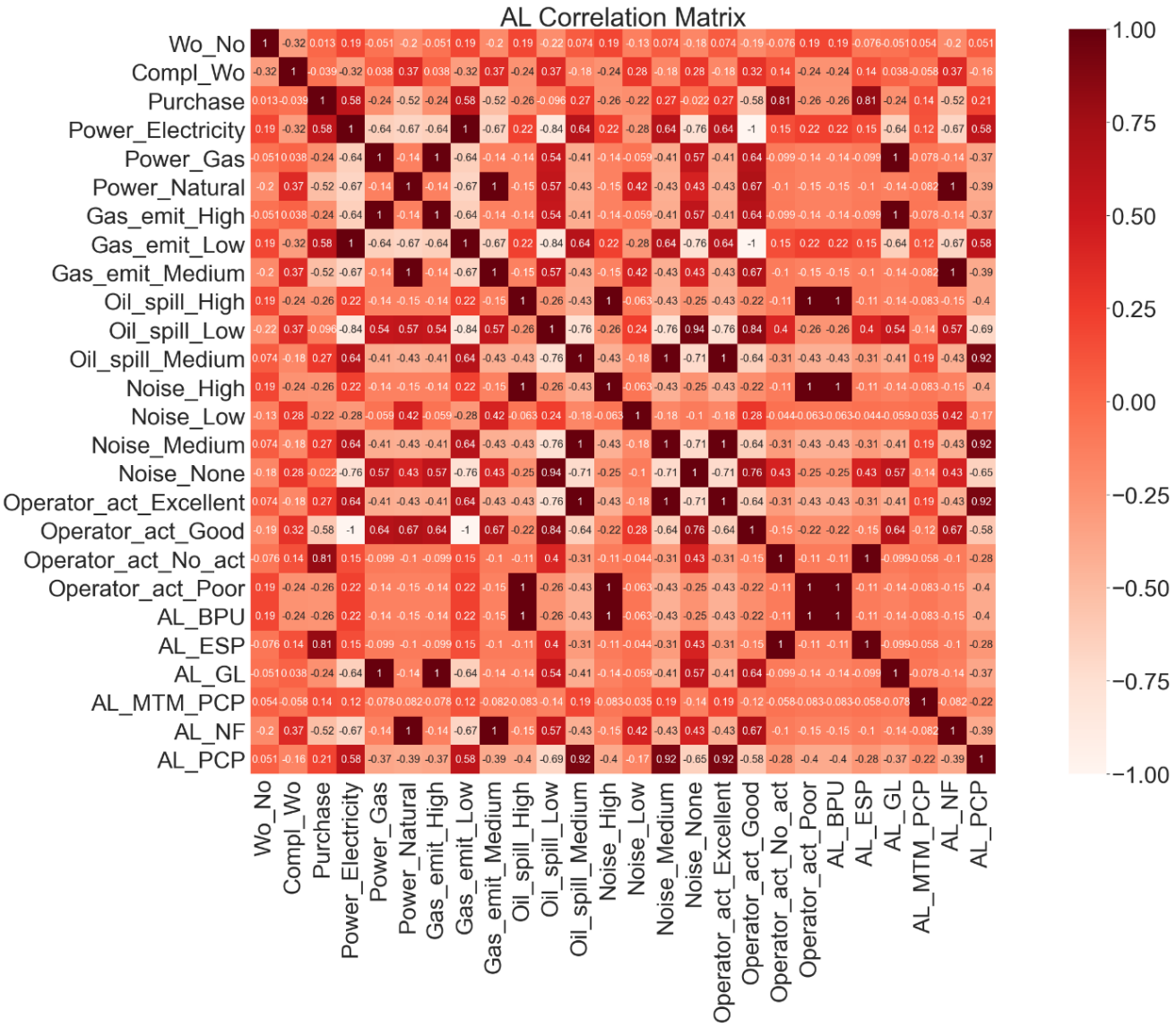


Fig. 4.21: Economic and environmental features correlation matrix

4.2.3.3 Statistical Data Analysis

The selected environmental features are categoric thus, no statistical measurements could be obtained prior to encoding. **Table 4.13** shows the statistical analysis of economic data while both environmental and economic features are shown in **Table 4.14** after features encoding.

Table 4.13: Economic features statistical analysis before encoding and normalisation

	Wo_No	Compl_Wo	Purchase
mean	2.105	152686.219	62458.895
std	1.881	119213.667	44284.963
min	0	21584.45	2500
25%	1	68792.455	32585
50%	2	119007.33	68420
75%	3	206358.72	72460
max	8	899676.1	196500

Table 4.14: Economic and environmental features statistical analysis after encoding and normalisation

	Wo_No	Compl_Wo	Purchase	Power_Electricity	Power_Gas	Power_Natural	Gas_emit_High	Gas_emit_Medium	Gas_emit_Low	Oil_spill_High	Oil_spill_Medium	Oil_spill_Low	Noise_High	Noise_Medium	Noise_Low	Noise_None	Operator_act_Excellent	Operator_act_Good	Operator_act_Poor	Operator_act_No_act
mean	0.263	0.149	0.309	0.756	0.116	0.127	0.116	0.127	0.756	0.131	0.556	0.313	0.131	0.556	0.025	0.287	0.556	0.244	0.131	0.069
std	0.235	0.136	0.228	0.430	0.321	0.334	0.321	0.334	0.430	0.338	0.498	0.464	0.338	0.498	0.158	0.453	0.498	0.430	0.338	0.254
min	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
25%	0.125	0.054	0.155	1.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
50%	0.250	0.111	0.340	1.000	0.000	0.000	0.000	0.000	1.000	0.000	1.000	0.000	0.000	1.000	0.000	0.000	1.000	0.000	0.000	0.000
75%	0.375	0.210	0.361	1.000	0.000	0.000	0.000	0.000	1.000	0.000	1.000	1.000	0.000	1.000	0.000	1.000	1.000	0.000	0.000	0.000
max	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

4.2.3.4 Baseline Model

Exclusively employing the prominent AL method of PCP, the model evaluation yielded an accuracy of 66.7% on the dataset encompassing economic and environmental aspects. This score closely aligns with the accuracy obtained through random guessing in the context of the operational dataset, where here the majority of features are categorical.

4.2.3.5 Model Training and Validation

In this model, the same wells were used as in the operation model, making a total of 100 wells. Out of these, 64 wells were selected for training and validation (3025 samples), while the remaining 36 were kept aside to test the model's performance on unseen data. The training and validation dataset was split into 50 wells for training and 14 wells for validation, with approximately 2563 and 462 data points, respectively. This approach has shown its effectiveness in the initial and subsequent models. We believe that the use of the same wells in both the operation model and this model will provide a fair comparison and allow the measurement of the sensitivity and importance of input features.

4.2.3.5.1 Training and Validation Dataset

The same wells distribution for model training and validation is listed and described as follows:

- XJS1, XJS4, XJC1, XJC2, XJS2, XJS7, XJS24, XJS29, XJS10, XJS38.
- XFN1, XFN4-2, XFN23, XFN49, XFN53, XFN80, XFN144, XFN161, XFC19, XF3.
- XFE4, XFE8, XFE12, XFE18, XFE36, XFE38, XFE39, XFE48, XFE81, XFE105, XFE20.
- XM13, XM1-4, XM1-8, XM1-12, XM9, XM10, XM17, XM18-4, XM33-3, XM33-5.
- XK1, XK2, XK3, XK4, XK6, XK11, XK21, XK23, XKS2, XKN9.
- XH1, XH6, XH9, XHN1, XHN8, XHN9, XH5.
- XSS1, XSE1, XSW1, XSFN1, XSE2, XS3, XSW6.

The validation wells selected are XJS10, XJS38, XFN49, XF3, XFE12, XFE81, XM1-3, XM9, XK2, XK23, XH5, XH9, XSFN1, and XSS1.

4.2.3.5.2 Training and Validation Results

The five algorithms exhibited perfect performance and attained impressive accuracy levels when utilizing default hyperparameters. Consequently, no fine-tuning was necessitated within this model to enhance its performance. As evidenced in **Table 4.15**, the algorithms proficiently predicted the AL using the economic and environmental dataset, with both DT and RF achieving perfect scores of 100% in both training and validation stages. This clear efficacy can be attributed to the predominance of categorical data within the dataset, resulting in a homogeneous and consistent classification of data categories.

Table 4.15: AL selection training and validation accuracy scores using environmental and economic dataset

Algorithm	Training Accuracy (%)	Validation Accuracy (%)
LR	94.85	100
SVM	94.85	100
KNN	94.85	100
DT	100	100
RF	100	100

4.2.3.6 Model Test on New Dataset

The test dataset used in this model consisted of 36 wells, which were selected to test the accuracy and generalizability of the trained models. The dataset contained 1694 data points that had similar features to those used in the training and validation sets. The time range of the test dataset was between 2020 to 2021, and the wells were randomly selected from the same field as the training and validation sets to ensure that the models could generalize well to new data.

4.2.3.6.1 Test Dataset

The same wells used in operation model were also used as a test dataset. However, the features in the dataset are different. Wells list is below:

- XJS9, XJS14, XJS13, XJS24, XJS34, produce using GL, NF, and PCP with N2 and gas injection.
- XFN148, XFC18, XF1, XF18, XFN17, XFN166, produce using PCP, MTMPCP, and NF with WI and gas injection.

- XFE22, XFE26, XFE40, XFE46, XFE84, produce using PCP and BPU with CHOPS, CSS, and SF.
- XM1-6, XM7, XM21, XM10-2, XM33-4, produce using PCP and NF
- XK9, XK20, XK22, XKS4, XKN12, produce using GL and PCP with gas and N2 injection.
- XHN7, XHN2, XH3-1, XH3-2, XHN6, produce light oil using ESP and PCP
- XSFE1, XSW2, XSFE2, XHG1, XSW7, produce using ESP with no IOR/EOR.

4.2.3.6.2 Model Test Results

The five algorithms demonstrated excellent accuracy scores in the AL prediction problem. RF and DT achieved the highest accuracy score of 99.35%, followed by SVM and LR with 96.1% accuracy and KNN with 95.45% accuracy. These results indicate that the models can accurately predict the optimal lifting method for a given well, based on the economic and environmental features. The high accuracy scores also suggest that the models can be deployed in the oil and gas industry to aid in the decision-making process, which can lead to more efficient and cost-effective production. A summary of obtained accuracy scores of each algorithm is given in Table 4.16.

Table 4.16: AL selection test accuracy scores using environmental and economic dataset

Algorithm	**Accuracy (%)**	**Recall (%)** ***** ******	**Precision (%)** ***** ******	**F1 Score (%)** ***** ******
LR	96.1	96.1	96.1	96.1
		82.1	83.33	82.69
SVM	96.1	96.1	96.1	96.1
		82.1	83.33	82.69
KNN	95.45	95.45	95.45	95.45
		82.08	83.11	82.58
DT	99.35	99.35	99.35	99.35
		97.62	99.78	98.61
RF	99.35	99.35	99.35	99.35
		97.62	99.78	98.61

**Micro **Macro average values*

4.2.3.7 Validation with Field Data Results and Discussion

The top performers, RF and DT, were again chosen to predict AL. **Figs. 4.22** and **4.23** illustrate RF and DT prediction performance and accuracy scores. The highest selection accuracy was obtained at a max depth of 3 and 4, for RF and DT, respectively. Both models accurately predicted all lifting methods with only single incorrect prediction in a PCP well (XFE36 CSS well), as shown in **Figs. 4.24** and **4.25**. The algorithms predicted MTMPCP instead of the actual PCP for a CSS well that had recently stopped the CSS project and replaced the MTMPCP with conventional PCP for cost considerations. The well had been producing oil using MTMPCP for years during the undergoing CSS, so the model prediction considered the operation history and nominated MTMPCP as the suitable AL. Moreover, the results indicate that data uncertainty of production and operation parameters affect the first and second model's performance. However, in this model, where most features are categorical and time-independent, the model found no obstacles in AL selection and achieved its highest accuracy scores. Therefore, the model can be relied upon for accurate AL predictions.

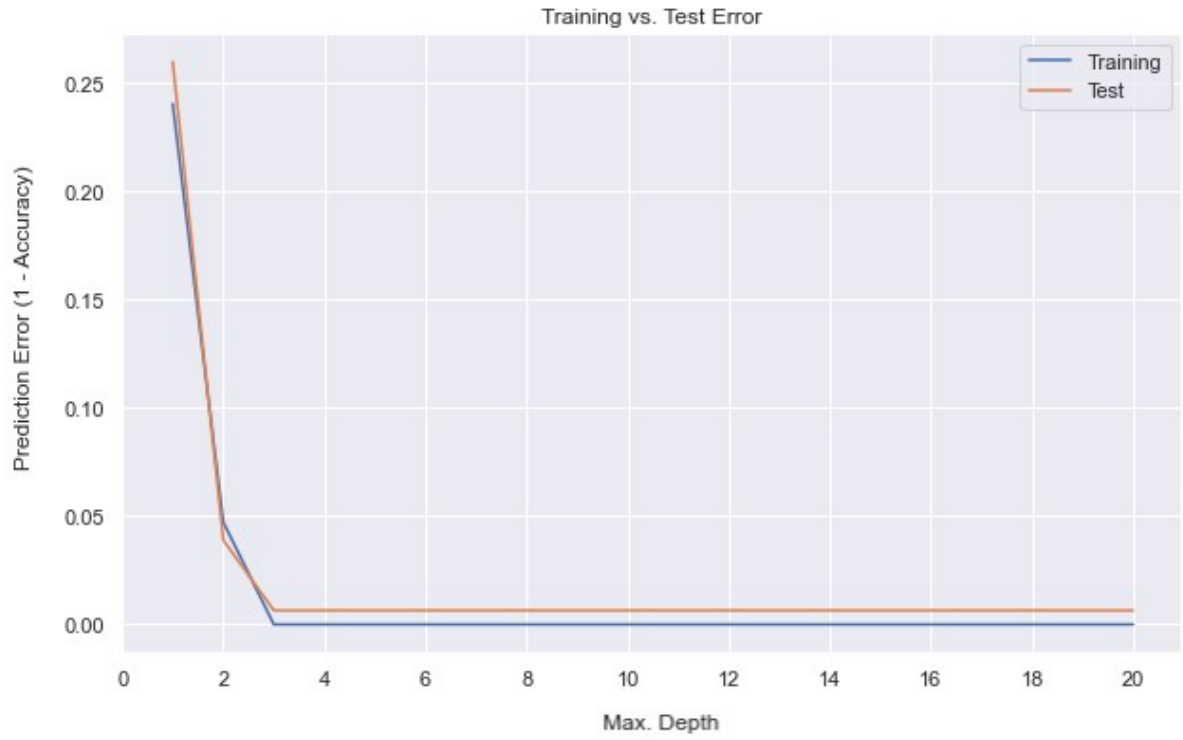


Fig. 4.22: RF training vs. test error using economic and environmental dataset

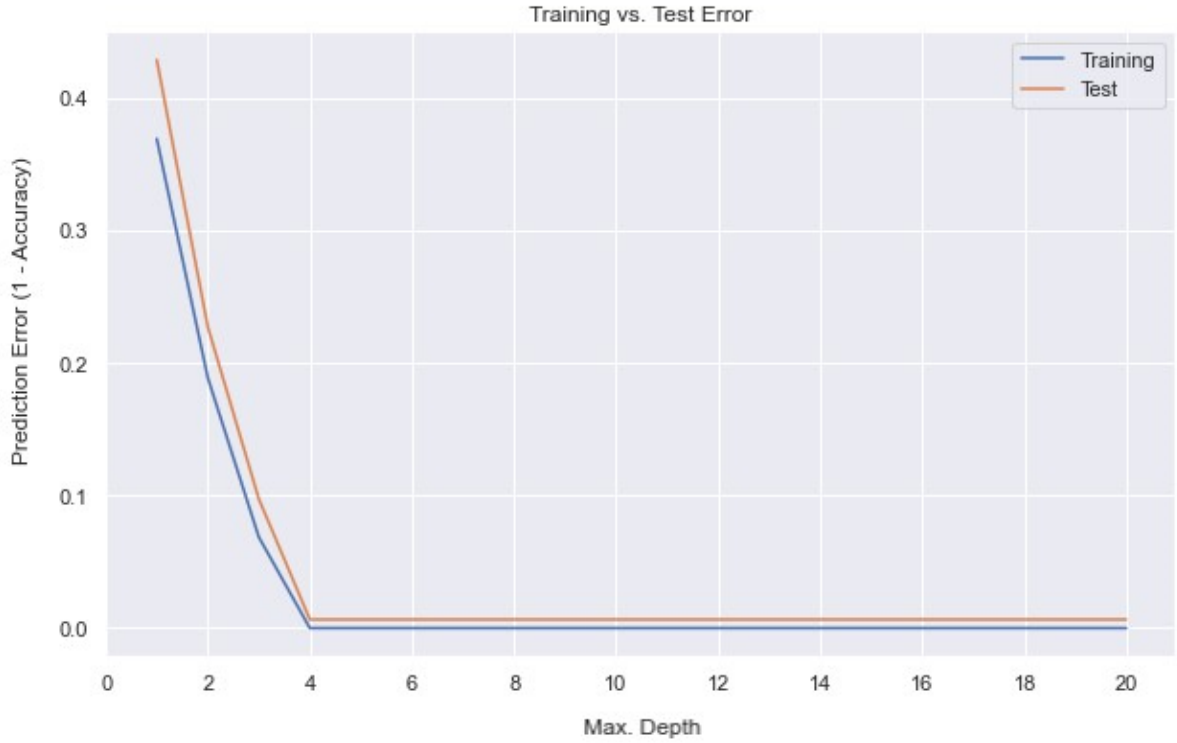


Fig. 4.23: DT training vs. test error using economic and environmental dataset

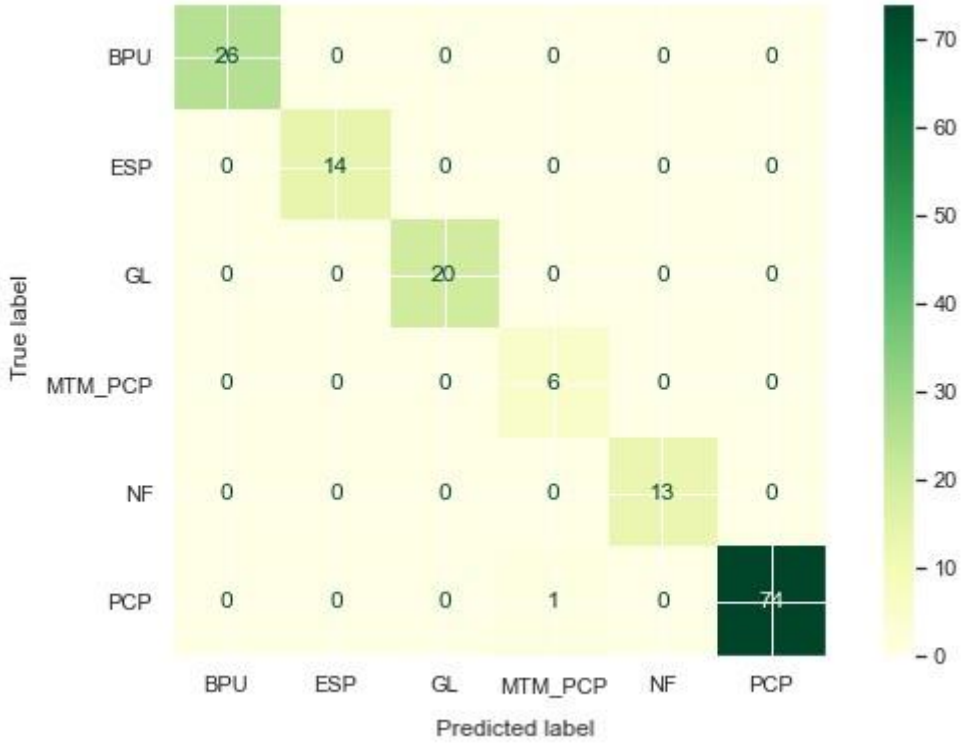


Fig. 4.24: RF AL selection test classification report using environmental and economic dataset

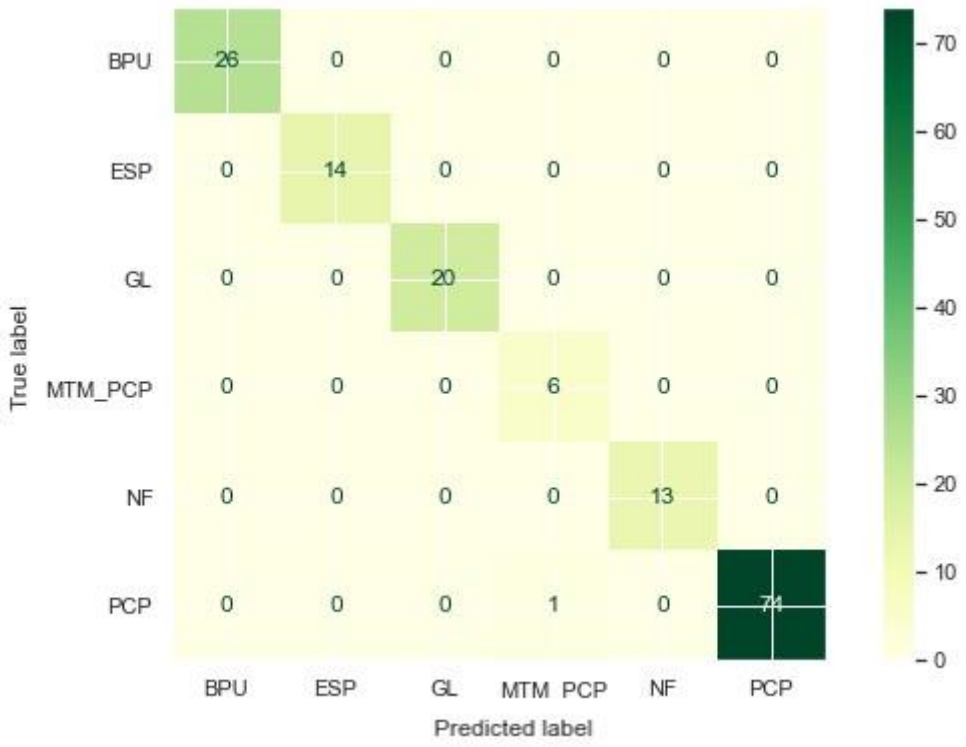


Fig. 4.25: DT AL selection test classification report using environmental and economic dataset

4.2.3.8 Model Test on Unlabelled Dataset

The model accurately predicted BPU and PCP from 10 unlabelled sample points with 100% accuracy. The model was trained using a dataset of wells with similar features and operational conditions, and it was able to generalize its predictions to new unlabelled data.

4.3 Developed Artificial Lift Clustering Model Using Unsupervised Learning

4.3.1 Clustering Process

The clustering was conducted on the production, operation, and environmental/economic datasets. In each model, the distinct clusters were determined to showcase and allow for identifying and grouping similar data points based on their production, operational, and environmental/economic features. The input features were separated from the target variable, allowing us to focus solely on clustering input data. The clustering uses the K-means algorithm, which is a

centroid-based clustering algorithm. It operates by iteratively assigning the data points to the nearest cluster centroid and updating the centroids based on the mean of the data points within each cluster. The resulting clusters are represented by their respective centroids, and each data point in the dataset is assigned to the cluster with the nearest centroid.

To ensure the compatibility of the categorical variables with the K-means algorithm, we performed label encoding for the categorical features, transforming them into numerical values, and scaling of numeric features for dimensionality reduction.

4.3.2 Determining K Using Inertia (elbow) and Silhouette Method

Figs. 4.26-4.28 display the inertia values (within-cluster sum of squares) for different k values. As the number of clusters increases, the inertia decreases, indicating better clustering performance. The elbow point on the graph, where the inertia begins to level off, is used to determine the optimal number of clusters. In this case, **Figs. 4.26-4.28** suggest that k=6 is a suitable choice for production and operation datasets, while **Fig. 4.28** recommends k=5 for the environmental/economic data set as the inertia shows a significant decrease up to this point, after which the decrease becomes less pronounced.

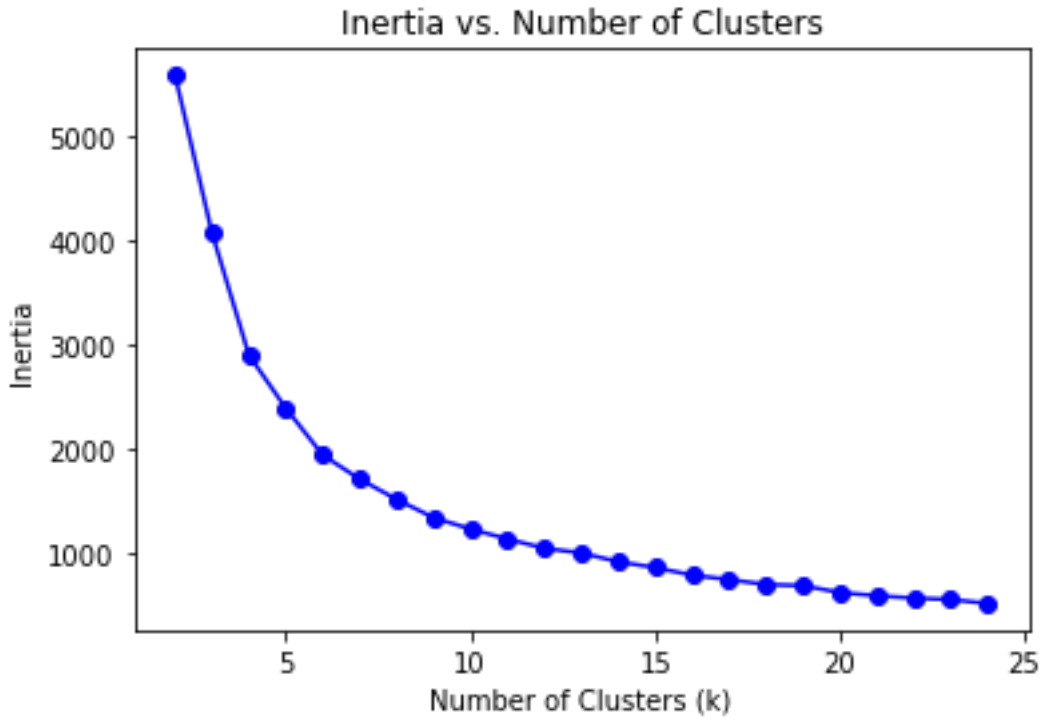


Fig. 4.26: Number of clusters for production parameters using inertia

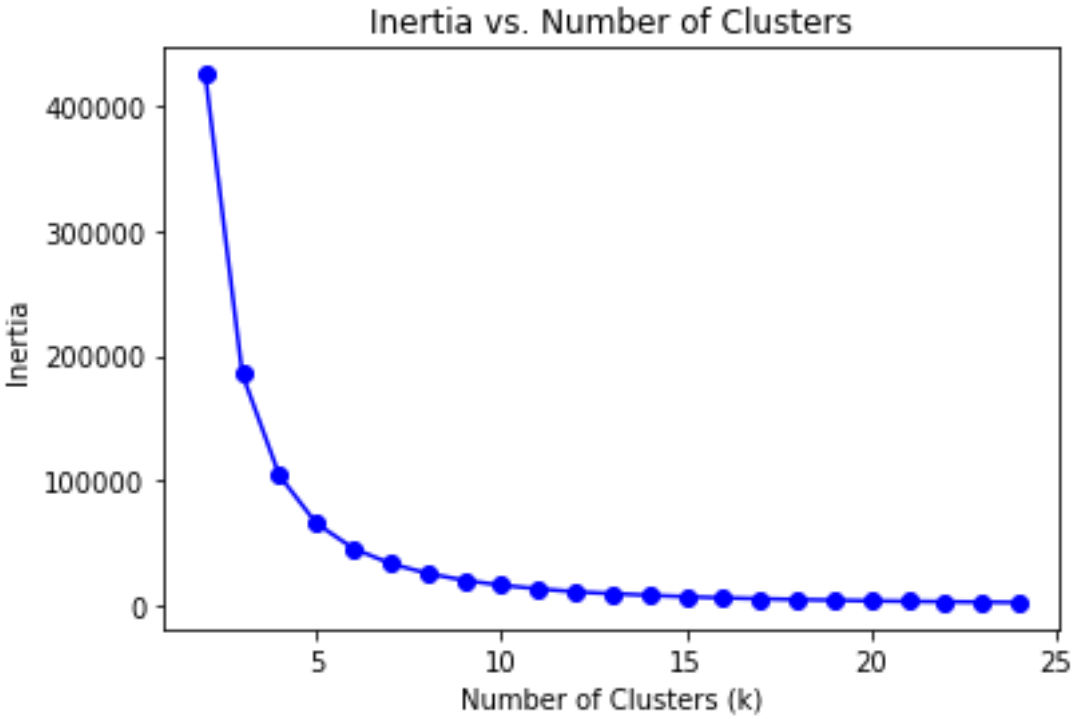


Fig. 4.27: Number of clusters for operation parameters using inertia

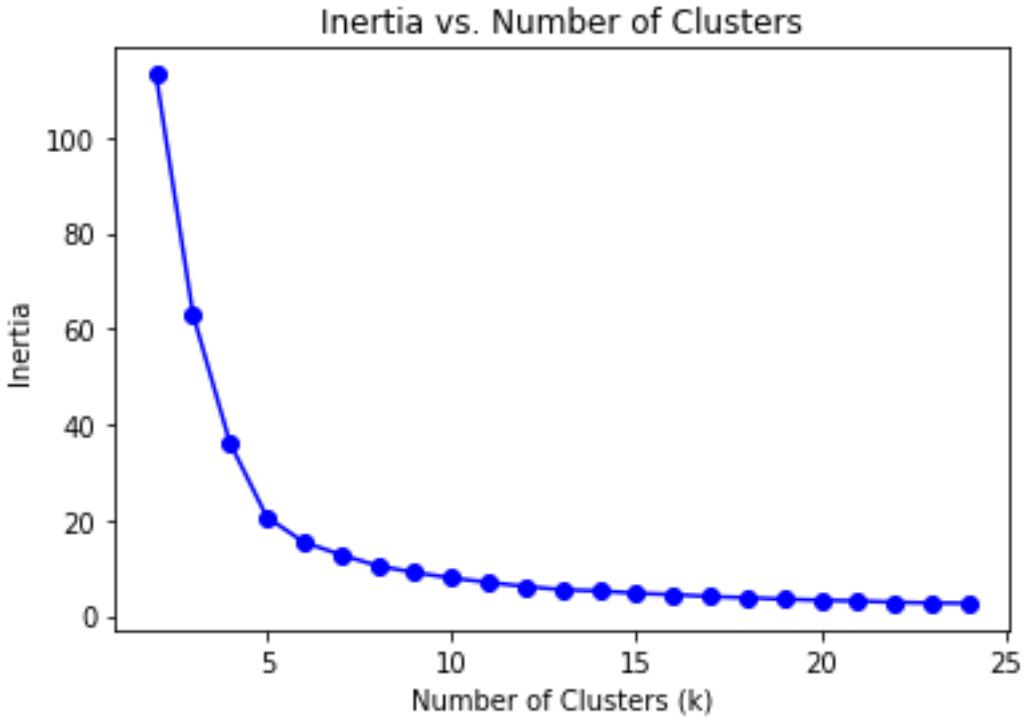


Fig. 4.28: Number of clusters for environmental and economic parameters using inertia

The silhouette score measures the cohesion and separation of clusters. Higher silhouette scores indicate better-defined clusters. **Figs. 4.29-4.31** show the silhouette scores for various k values. The silhouette score for production in **Fig. 4.29** was high at k=6, indicating an optimum clustering and supporting the elbow results. The result for operation parameters, shown in **Fig. 4.30**, did not provide a clear vision as the score gradually decreased. The highest silhouette score for the environmental/economic dataset is achieved at k=5 (**Fig. 4.31**), supporting the elbow choice of 5 clusters as the optimal number.

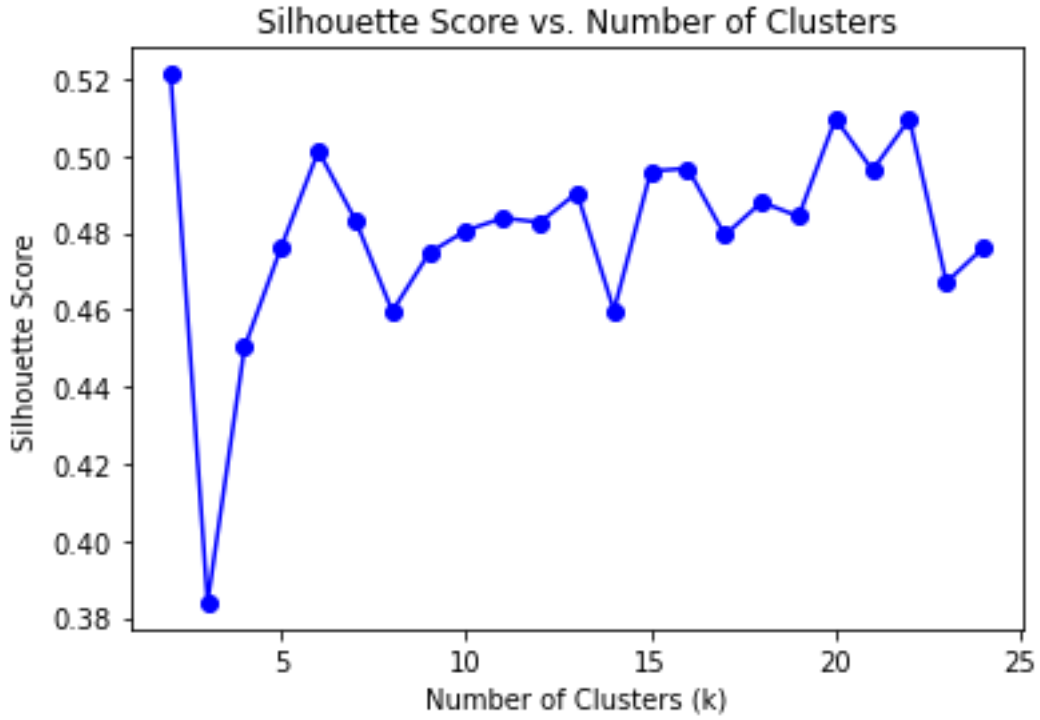


Fig. 4.29: Number of clusters for production parameters using silhouette

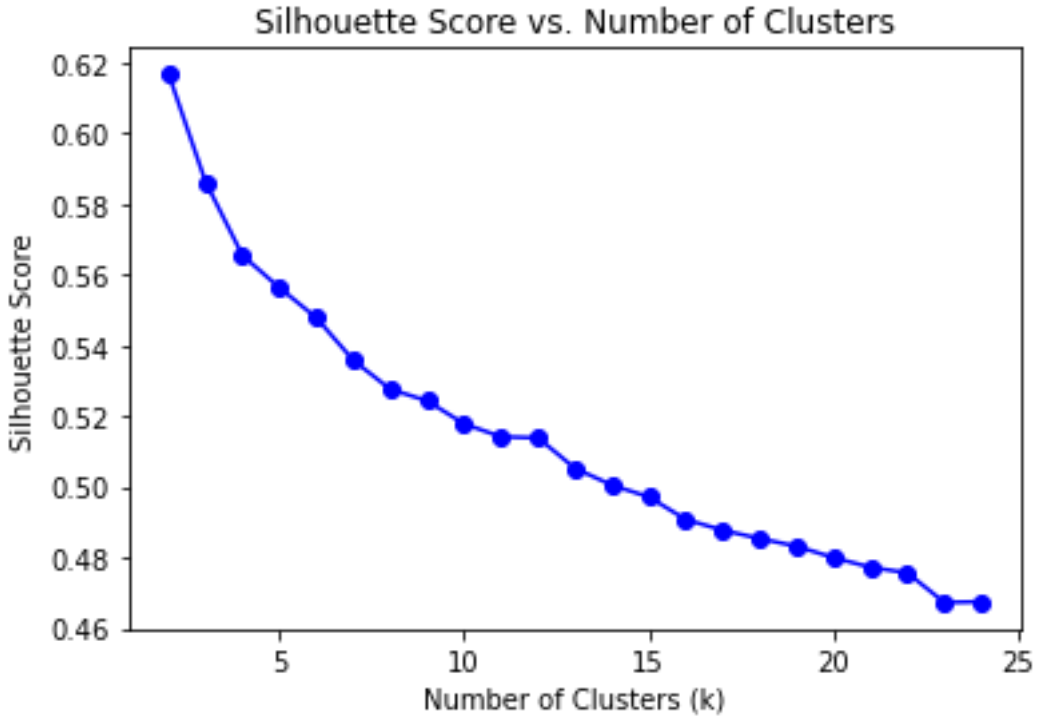


Fig. 4.30: Number of clusters for operation parameters using silhouette

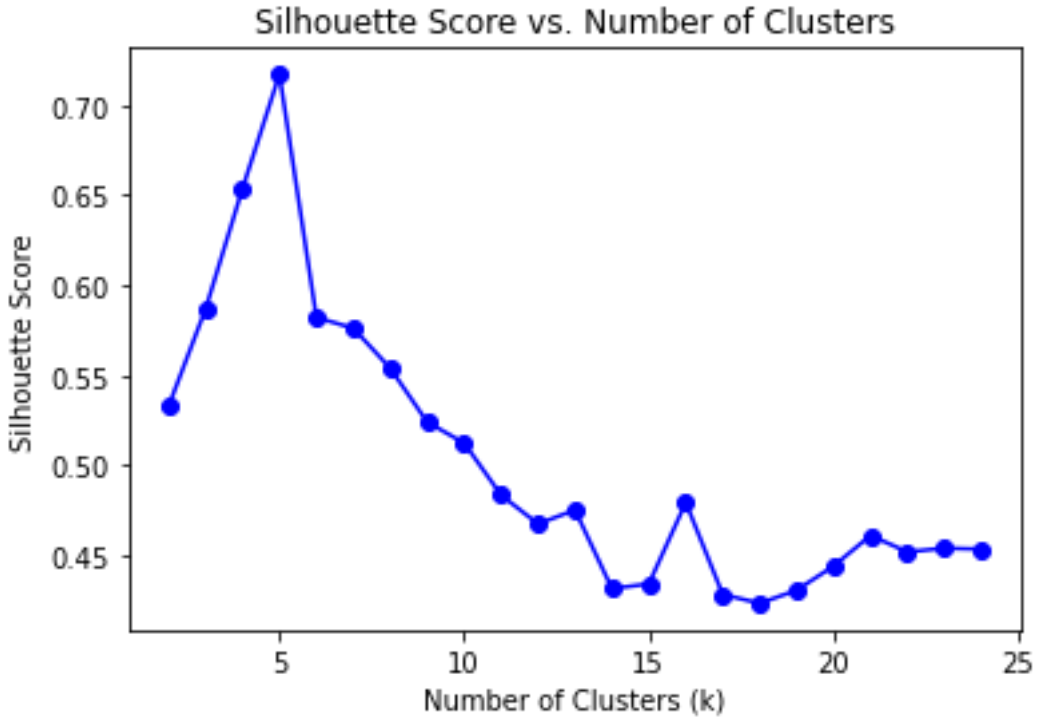


Fig. 4.31: Number of clusters for environmental and economic parameters using silhouette

4.3.3 Clustering Results and Discussion

Figs. 4.32-4.34 show the clusters obtained from applying K-means clustering to the input features of the three datasets. The clustering process aimed to group similar data points together based on their feature values. Each data point is represented in a two-dimensional space (Principal Component 1 and Principal Component 2) after dimensionality reduction using the Principal Component Analysis (PCA). PCA is a dimensionality reduction technique used to transform a high-dimensional dataset into a lower-dimensional space (2D) while preserving the maximum variance in the data (Jolliffe, 2002). When applying PCA to the original input features, it finds the two orthogonal axes (PC1 and PC2) in the new feature space that capture the most variance in the data. PC1 represents the axis with the highest variance, and PC2 represents the second highest variance, orthogonal to PC1. The clusters are denoted by different colours in the plot, indicating their separations in the reduced feature space. The distribution of data points within each cluster suggests the presence of distinct patterns and similarities among the input features.

As evident from **Figs. 4.32** and **4.34**, the clusters are effectively structured and distinctly illustrate the arrangement of production as well as environmental/economic attributes. However, **Fig. 4.33** depicts the dispersion of operational characteristics within each cluster, potentially challenging the identification of homogenous parameter groups as depicted in the silhouette plot. This is noteworthy considering that the algorithm established clusters based on the specified value of k=6.

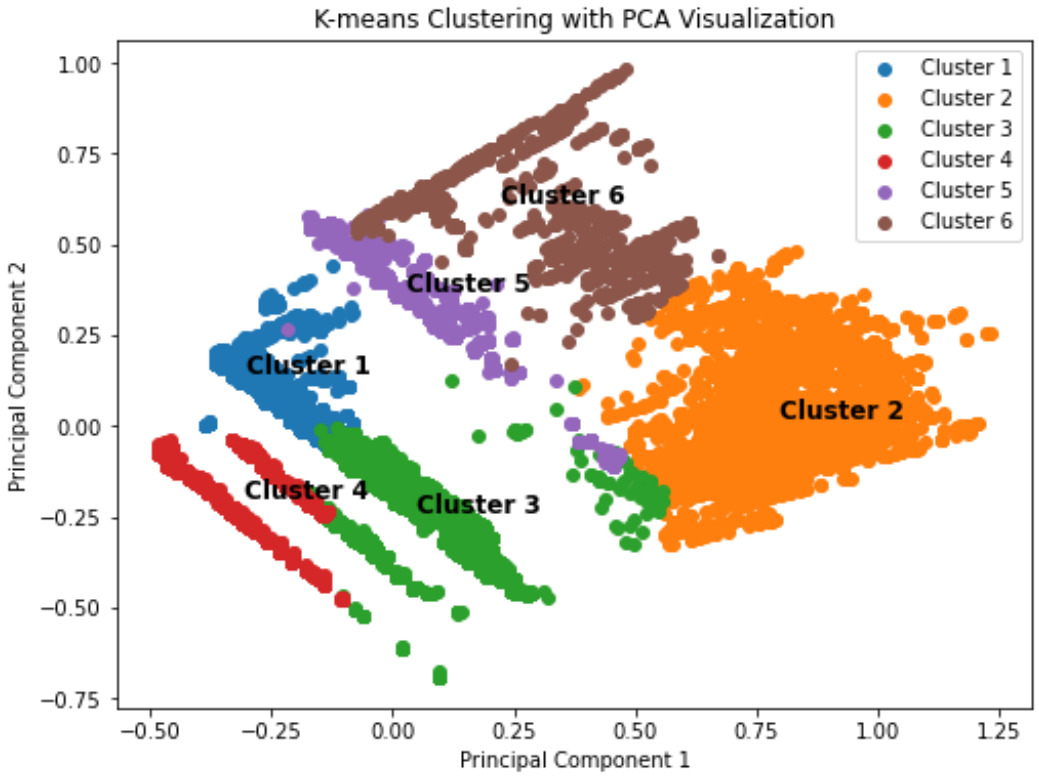


Fig. 4.32: Production data clusters

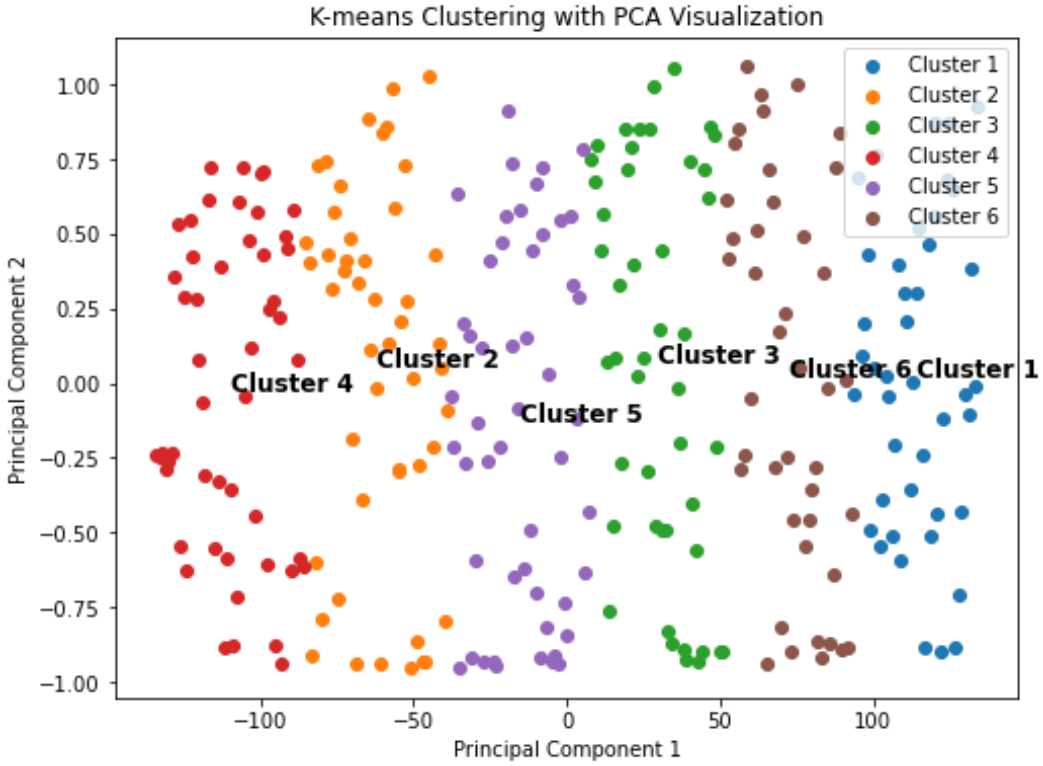


Fig. 4.33: Operation data clusters

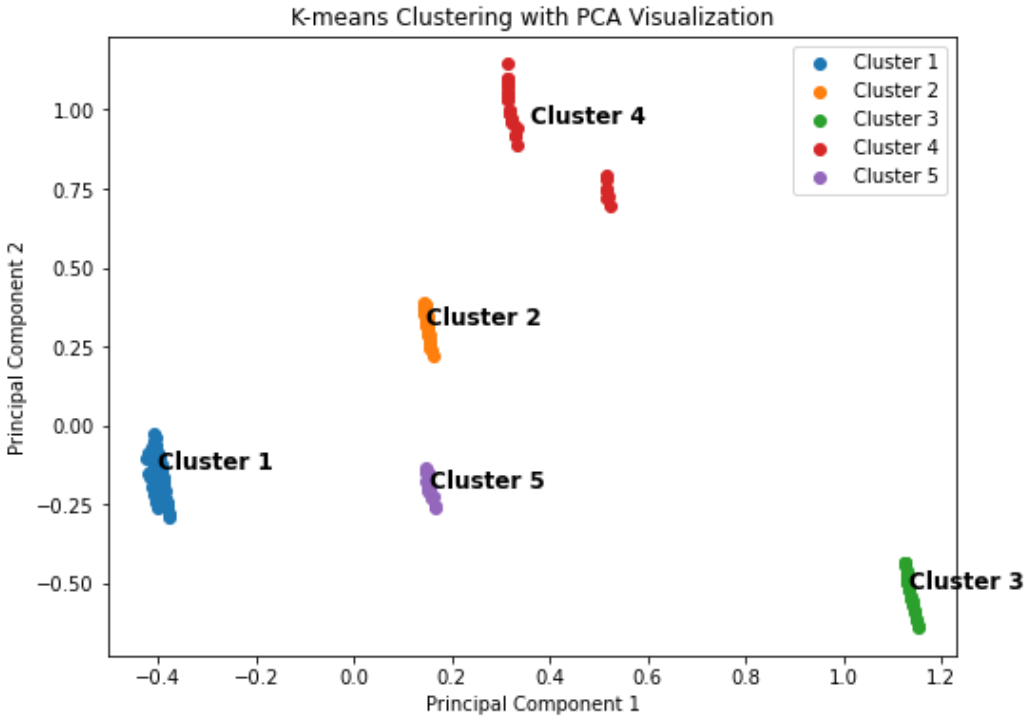


Fig. 4.34: Environmental and economic data clusters

The application of clustering to the classification problem of AL selection holds significant importance in the OGI. Clustering plays a crucial role in the AL selection process as it allows for the identification and grouping of similar data points based on their production, operational, environmental, and economic features. By utilizing K-means clustering, we can segregate wells into distinct clusters based on their performance characteristics, reservoir parameters, and other relevant factors. The clustering results enable engineers and decision-makers to gain valuable insights into the patterns and trends present within the production dataset. It provides a comprehensive understanding of how wells with similar attributes behave and perform under various AL methods. This knowledge is instrumental in making informed decisions about which AL method to apply to a particular well to optimize production efficiency and enhance oil recovery. Moreover, clustering aids in the identification of similarities and differences among different AL methods. It allows for a comparative analysis of the performance of various lifting methods within each cluster, helping operators determine which methods are most suitable for specific groups of wells. Additionally, the clusters aid in detecting outliers and anomalies, which could indicate production issues or the need for AL size adjustments. By leveraging clustering to group oil wells into meaningful clusters, the AL selection process becomes more targeted and

effective. It streamlines the decision-making process, reduces trial and error, and enhances the overall efficiency of AL implementation in oilfield operations. Ultimately, the integration of clustering into the classification problem of AL selection can lead to improved field productivity, reduced operational costs, and better management of the oil and gas reservoirs, making it an indispensable tool for the industry.

4.4 Summary

The chapter delved into the three AL selection models underpinned by ML, each tailored to a distinct dataset encompassing production, operation, and environmental/economic factors, underscores the pivotal role of ML in streamlining the selection process. This approach circumvents the qualitative and time-consuming process of combining the heterogeneous data types in the field. The ML models have exemplified the ability to harness specific datasets to draw precise inferences, avoiding the complexities to correlate multifarious field data. The crucial contribution of these models lies in their ability to streamline the AL selection process. The high accuracy of the model's predictions can be attributed to the effective AL selection and tuning of the ML algorithm's hyperparameters. The model's performance highlights the potential of ML models in predicting AL methods for oil production and can greatly aid in the decision-making process for oil operators. Furthermore, the introduction of clustering techniques within this framework offers a novel perspective on data groupings, serving as a catalyst in optimizing the process of AL selection.

CHAPTER 5

ARTIFICIAL LIFT SIZE SELECTION MODEL

5.1 Introduction

The selection of the appropriate AL method is crucial for maximizing production rates and optimizing oil and gas field operations. However, advancements in AL selection ignored the optimum AL size that has a significant impact in production performance. AL size refers to the specification or dimension of the AL equipment, such as pumps or lift systems. It typically includes parameters such as pump and tubing diameter, number of valves, number of stages, theoretical flow rates, or other relevant parameters that are crucial for optimising the performance and efficiency of AL operations. This chapter introduces a novel approach to AL size selection using production data. The main objective is to develop a ML model that can accurately predict the appropriate size of the AL method based on production data. To achieve this objective, three different ML models were applied. Each model was trained and evaluated using production data from various AL sizes from the four commonly used AL methods: PCP, BPU, ESP, and GL including NF. The obtained accuracies of the models ranged between 60% and 93%, indicating their potential for accurate AL size selection. It is important to highlight that the AL size selection model proposed in this research is unique and has not been previously explored in the literature (Mahdi et. al, 2023). Previous studies have primarily focused on the selection of AL methods without considering the appropriate size. By incorporating the concept of AL size into the models, this research aims to provide a more comprehensive and precise approach to AL selection.

5.2 Developed AL Size Selection Model Using Supervised Learning

5.2.1 Input Parameter and Data Visualisation

The production rates are dependent to the size of the AL. Thus, in this model only production data were used in size selection model. In addition to production parameters used in AL model, the AL itself was used as an input parameter resulting in a total of 10 features as shown in **Table 5.1**.

Table 5.1: Size selection parameters

Feature	Unit
Wellhead pressure	psi

Daily produced fluid	BLPD (bbl/D)
gas-oil ratio (GOR)	scf/STB
Daily oil production	STB/D
Daily water production	BWPD (bbl/D)
Water cut	%
Daily gas production	Mscf/D
Daily sand production	bbl/D
IOR/EOR methods	Categories (Gas injection, NI, WI, CSS, and SF)
AL	PCP, BPU, GL, ESP, and NF

Figs. 5.1-5.3 shows the distribution of 16, 9, and 6 AL sizes in the dataset according to field history. The sizes ranging from small, medium, large, and number of stages for pumps, in addition to mandrels and x-mass trees for non-pumping lifting methods. All sizes are also presented in **Table 5.2**.

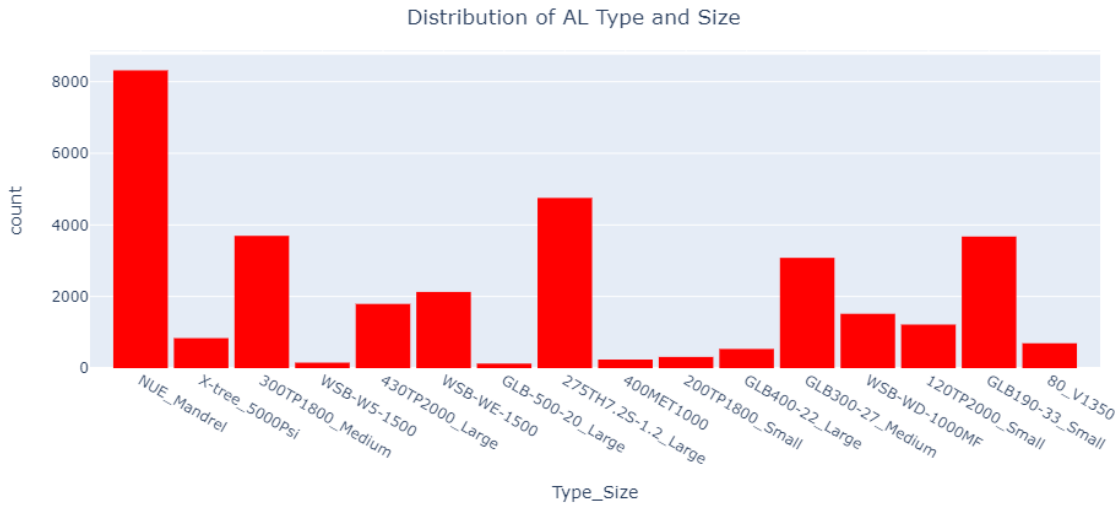


Fig. 5.1: Distribution of 16 AL sizes in the dataset

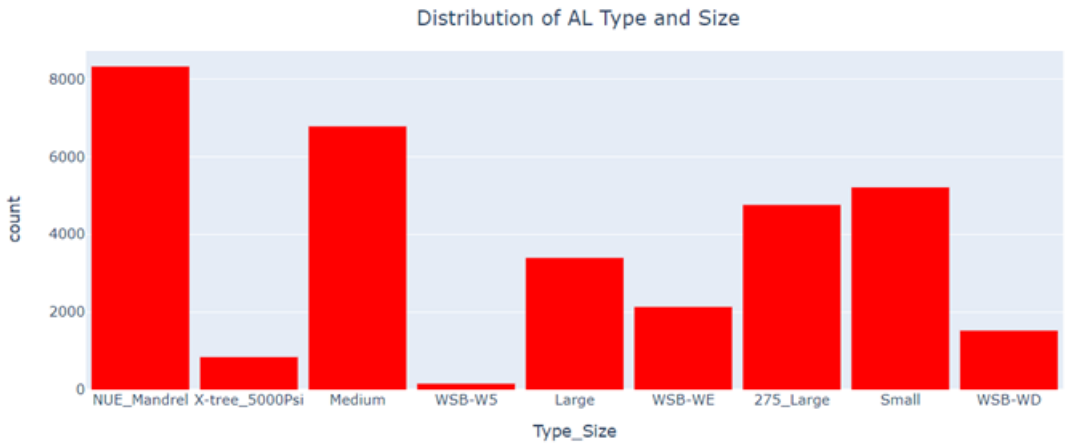


Fig. 5.2: Distribution of 9 AL sizes in the dataset

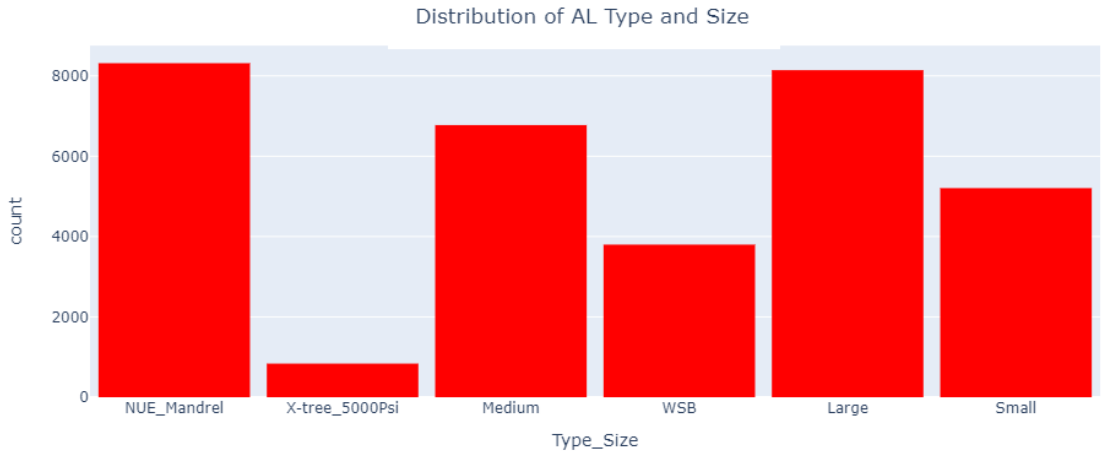


Fig. 5.3: Distribution of 6 AL sizes in the dataset

Table 5.2: AL sizes in the dataset including NF X-trees

PCP		BPU		ESP		GL	NF
GLB500-20	Large	275TH7.2S-1.2	Large	WSB-W5-1500	WSB	NUE Mandrel 3 Valves	X-tree_5000Psi
430TP2000		225TH7.2S-1.2	Small	WSB-WE-1500			X-tree_3000Psi
GLB400-22				WSB-WD-1000MF			
400MET1000 (MTMPCP)							
80_V1350 (MTMPCP)							
300TP1800	Medium						
GLB300-27							
200TP1800	Small						
120TP2000							
GLB190-33							

Figs. 5.4 and **5.5** present the cumulative oil production (sum of oil and total fluid, Y axis) achieved by each size of AL methods. The use of mandrels and X-mass trees in GL and NF operations respectively resulted in the highest aggregate oil and total fluid production of 13 and 16.5 million barrels, respectively. Following is ESP with various sizes and stages, followed by medium-sized PCP and large-sized

BPU. It is important to note that the lower production obtained by large AL sizes such as PCP GLB-500 and MTMPCP 400MET1000 does not indicate their inefficiency. Instead, it reflects the relatively short duration for which these AL sizes were installed before being replaced due to damage or other reservoir/production-related issues.

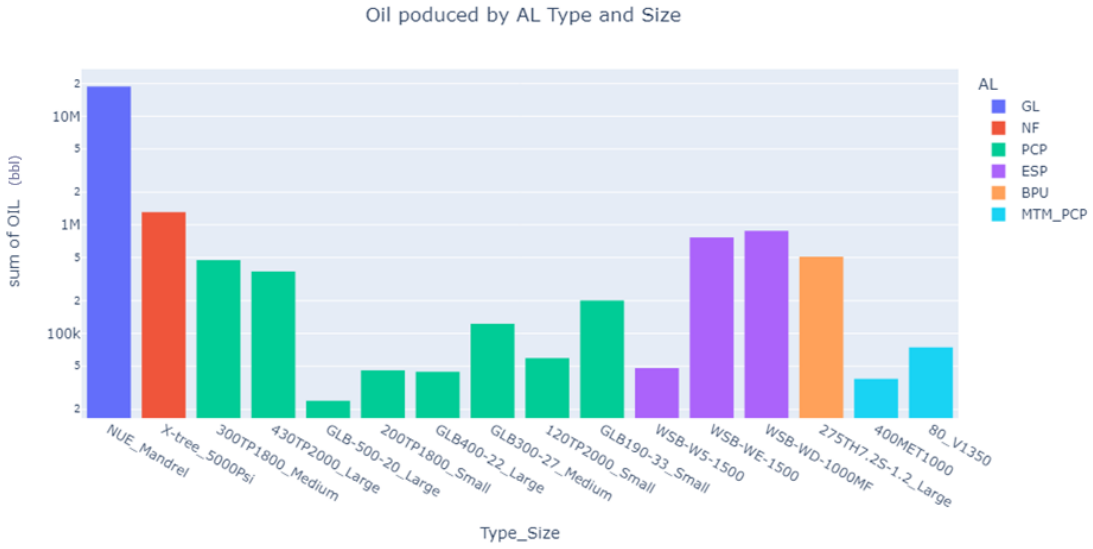


Fig. 5.4: Cumulative oil production by each AL size

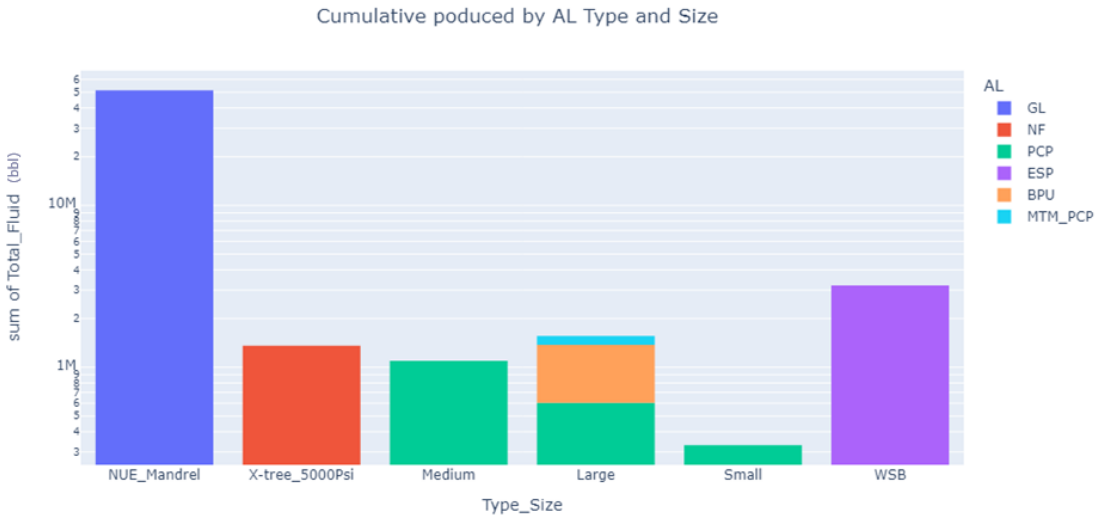


Fig. 5.5: Cumulative fluid production by each AL size

Fig. 5.6 also underscores the impact of AL size on cumulative oil production. The results indicate that the choice of AL size, particularly in the case of GL and NF

methods, can significantly influence the overall production volume. Furthermore, the relatively higher production observed with ESP, BPU, and PCP, which offers a range of size options, suggests the importance of selecting the appropriate size for these AL methods. The lower production associated with large AL sizes highlights the need for careful consideration of installation duration and potential operational issues that may lead to premature replacement. The size has a major effect on the amount of produced water. As shown in **Fig. 5.5**, the larger the PCP and BPU size, the higher the barrels of produced water that affects the well performance. However, the highest water cut recorded was in both GL and ESP in addition to some naturally flowing well.

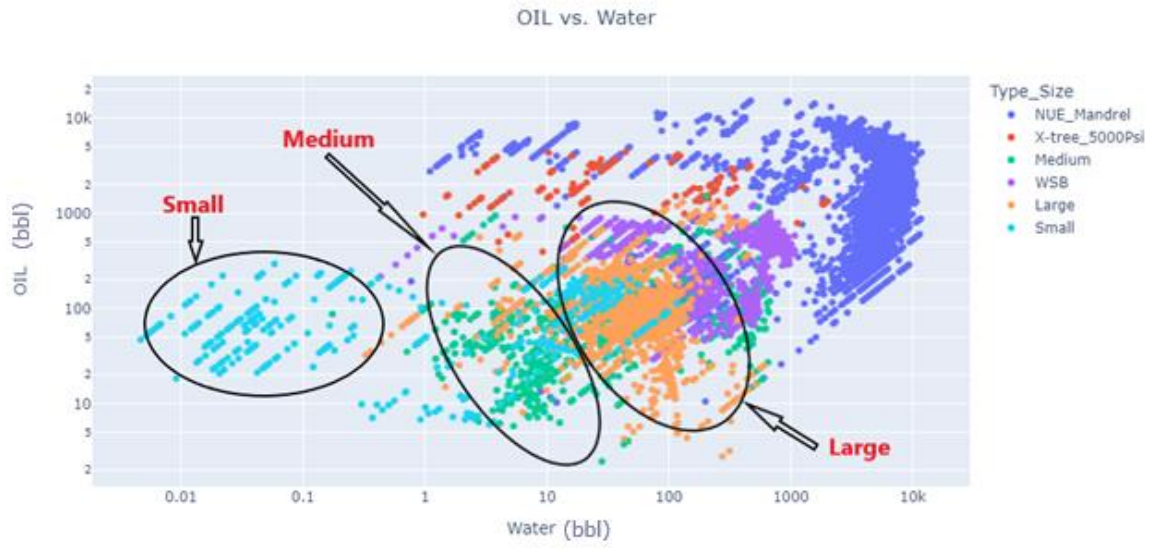


Fig. 5.6: The effect of AL size on water production

5.2.2 Data Correlation

Here we have the data correlation figures of the 3 models. **Fig. 5.7** illustrates the correlation between the input variables and the 16 size classes. **Fig. 5.8** shows the correlation of the reduced size classes to 9 while **Fig. 5.9** demonstrates how the input features correlated to 6 sizes of AL. The GL mandrels have strong correlation in the three models with most input parameters excluding sand production water flooding and thermal recovery methods that shows negative correlation. The two sizes – 'Medium and Small' - positively correlate to water flooding IOR which reflects the implementation of PCP with both sizes. The size

'Large' strongly correlates to thermal recovery methods with a slight positive correlation to sand production which is associated to BPU and MTMPCP wells.

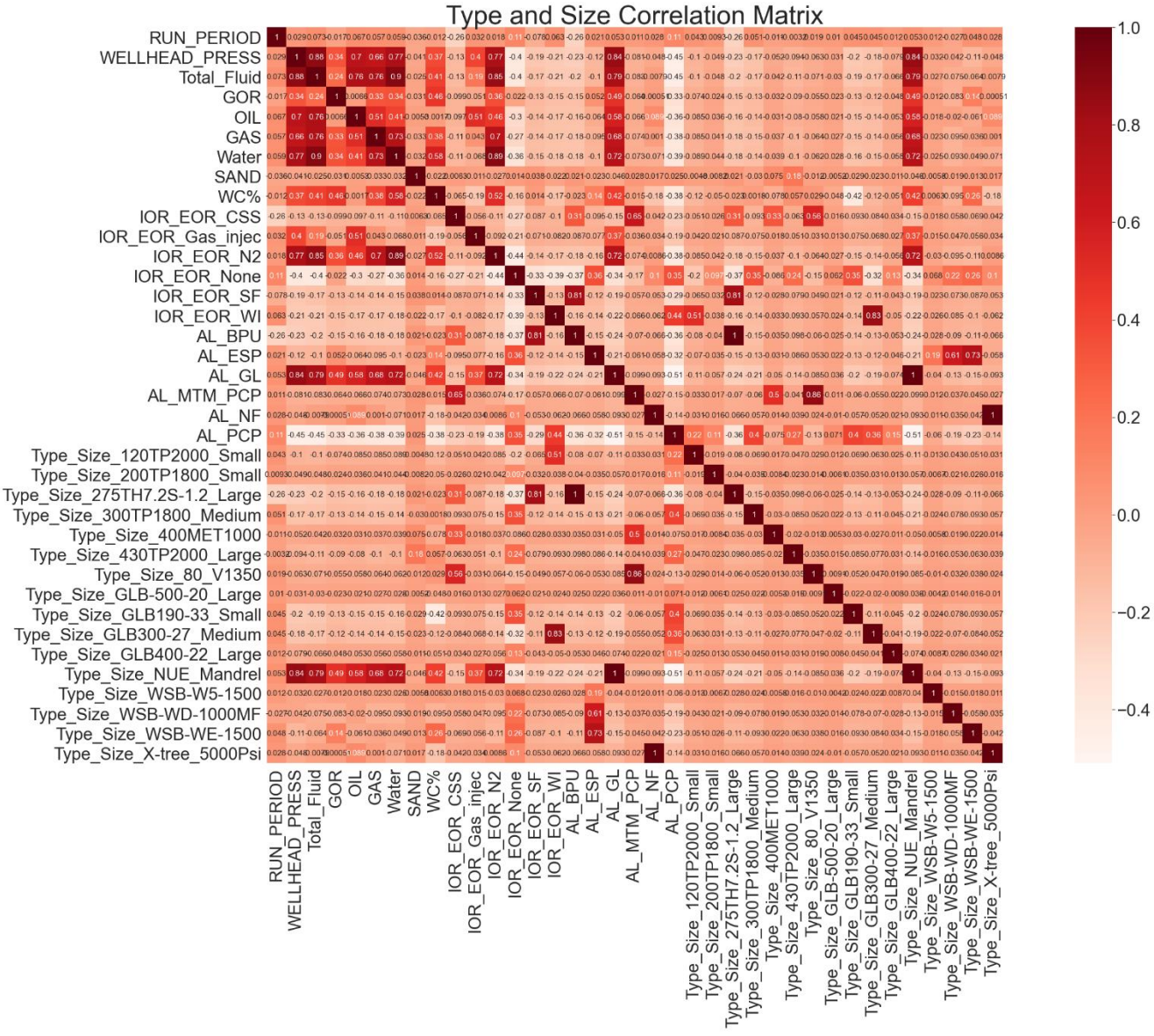


Fig. 5.7: 16 size classes correlation matrix

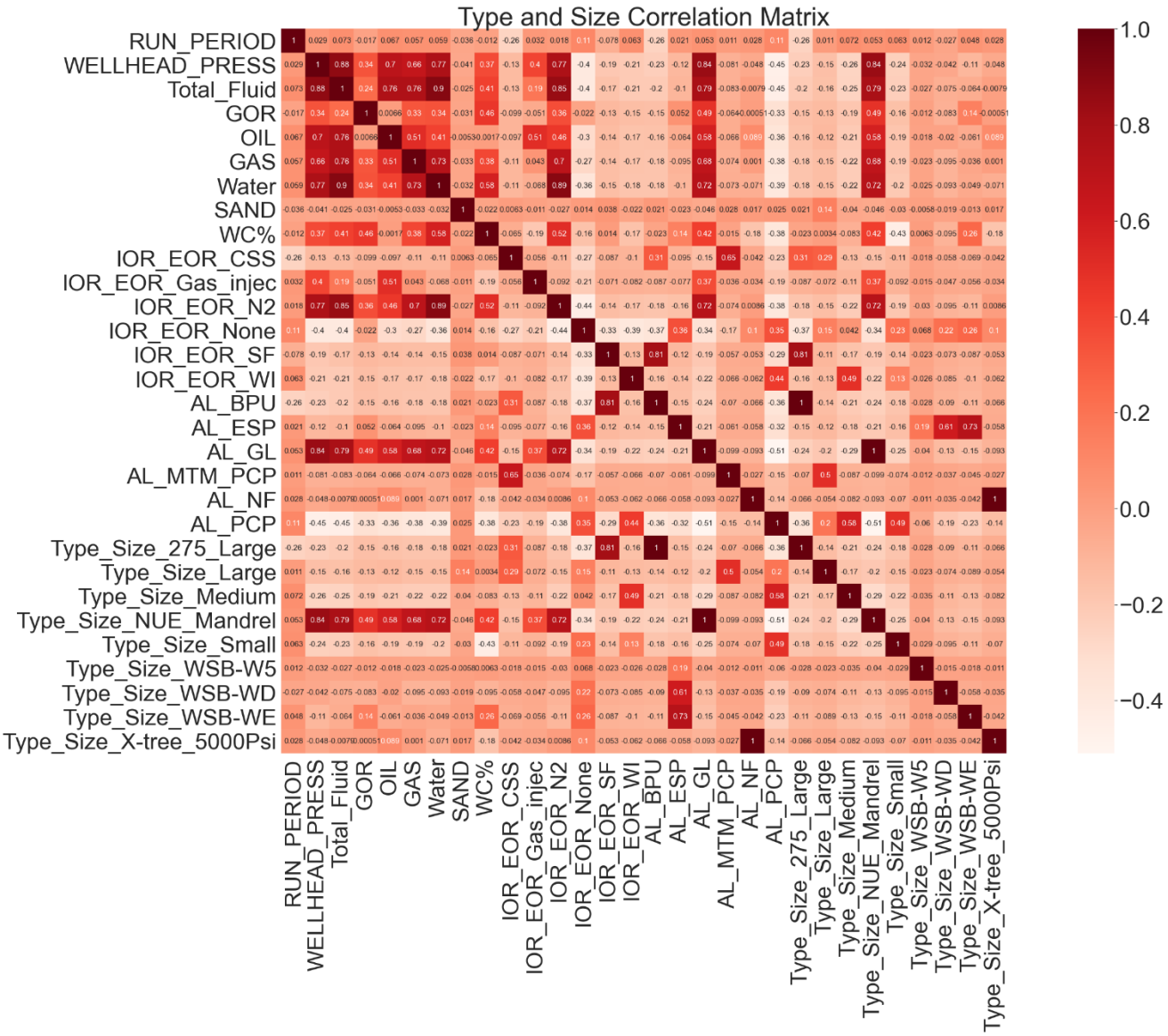


Fig. 5.8: 9 size classes correlation matrix

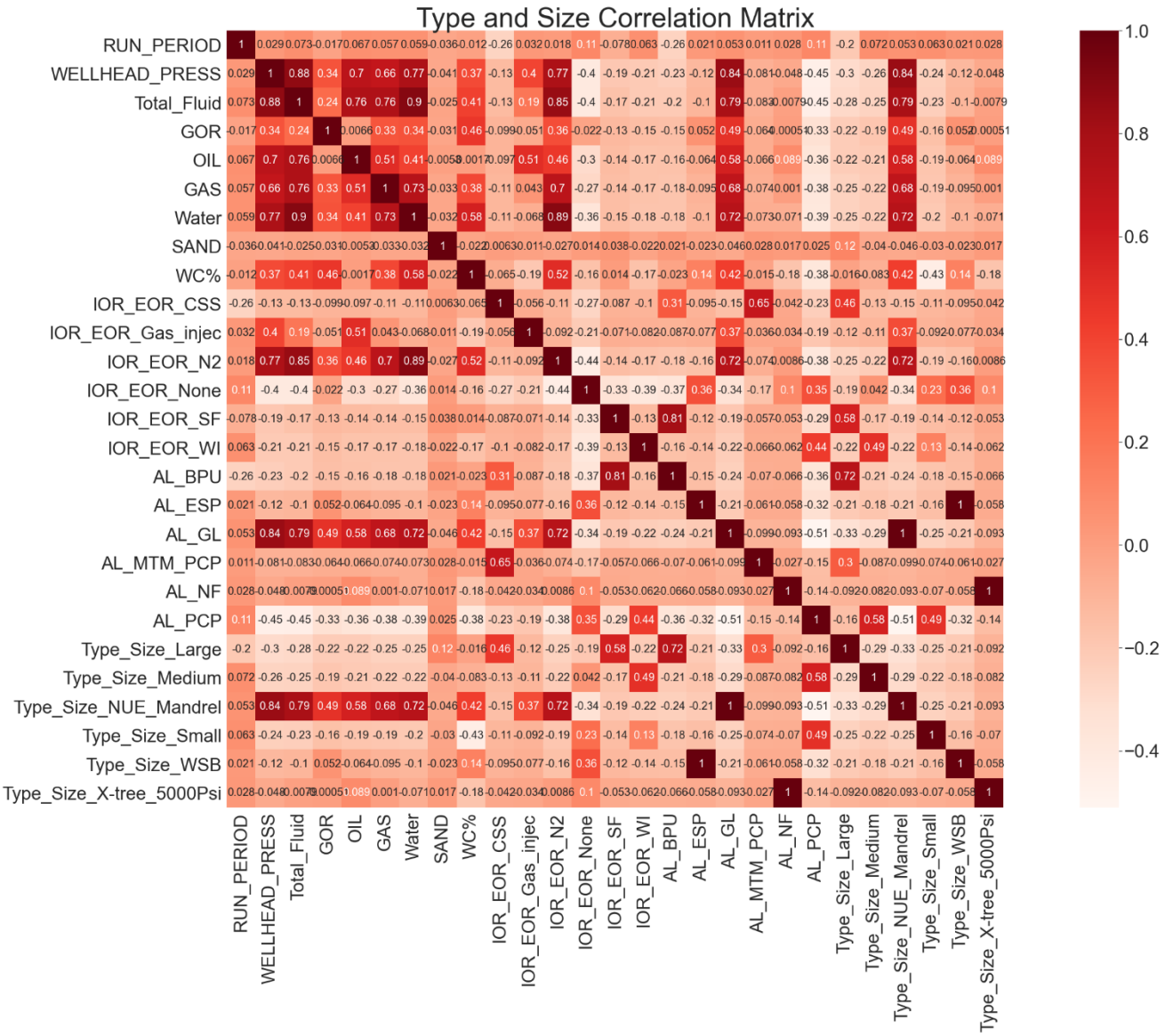


Fig. 5.9: 6 size classes correlation matrix

5.2.3 Statistical Data Analysis

Table 5.3 demonstrates the statistical analysis of production dataset. In addition to production features, the table shows the mean, standard deviation, min, and max values of AL type as an input feature used in size selection. Incorporating AL method as an input parameter for size selection is crucial in statistical analysis because it allows for the investigation of how different AL techniques interact with a wide range of production variables. This approach provides valuable insights into the factors that most significantly influence the selection of AL size, contributing to a more robust and data-driven decision-making process.

Table 5.3: Statistical data of input parameters after encoding and normalisation

	mean	std	min	25%	50%	75%	max
Year	0.665	0.204	0.000	0.533	0.733	0.800	1
Month	0.511	0.314	0.000	0.273	0.545	0.818	1
Day	0.493	0.292	0.000	0.233	0.500	0.733	1
RUN_PERIOD	0.971	0.118	0.000	1.000	1.000	1.000	1
WELLHEAD_PRESS	0.118	0.162	0.000	0.031	0.041	0.081	1
Total_Fluid	0.109	0.200	0.000	0.004	0.012	0.075	1
GOR	0.006	0.015	0.000	0.000	0.000	0.005	1
OIL	0.047	0.101	0.000	0.004	0.007	0.034	1
GAS	0.023	0.053	0.000	0.000	0.000	0.020	1
Water	0.090	0.196	0.000	0.001	0.004	0.034	1
SAND	0.005	0.056	0.000	0.000	0.000	0.000	1
WC%	0.367	0.309	0.000	0.060	0.320	0.630	1
IOR_EOR_CSS	0.065	0.246	0.000	0.000	0.000	0.000	1
IOR_EOR_Gas_injec	0.043	0.204	0.000	0.000	0.000	0.000	1
IOR_EOR_N2	0.158	0.365	0.000	0.000	0.000	0.000	1
IOR_EOR_None	0.504	0.500	0.000	0.000	1.000	1.000	1
IOR_EOR_SF	0.099	0.299	0.000	0.000	0.000	0.000	1
IOR_EOR_WI	0.130	0.336	0.000	0.000	0.000	0.000	1
AL_BPU	0.144	0.351	0.000	0.000	0.000	0.000	1
AL_ESP	0.115	0.319	0.000	0.000	0.000	0.000	1
AL_GL	0.251	0.434	0.000	0.000	0.000	1.000	1
AL_MTM_PCP	0.028	0.166	0.000	0.000	0.000	0.000	1
AL_NF	0.025	0.157	0.000	0.000	0.000	0.000	1
AL_PCP	0.436	0.496	0.000	0.000	0.000	1.000	1

5.2.4 Model Training and Validation

The model that used the production data category was the model used to select the optimum size because the flow rate is related to AL size and production performance as well as well deliverability. Three model runs were conducted using 16, 9, and 6 size classes for training and validation. Different sizes were deployed in modelling to thoroughly study production performance history of each size in order to select the optimum one for prospective well with the highest accuracy. The criterion also examines the effect of number of target variables on the model prediction performance.

5.2.4.1 Training and Validation Dataset

The same wells used in the production dataset (*chapter 4 section 4.2.1.3.1*) were used in this model. The divergence here is the use of AL as an input feature and the target variables are the numerous AL sizes. The wells are: XF161, XF3, XF19 and XF144, XFE36 and XFE38, XJ1 and XJ4, XM17 and XM184, XH1 and XH6, XK7 and XK21, XSS1 and XSW1

5.2.4.2 Training and Validation Results

Table 5.4 presents the training and validation accuracies obtained from 16, 9, and 6 AL sizes. It's evident that many classes reduce model performance while fewer size classes result in high accuracy scores. LR achieved a training accuracy of 91.13% with a validation accuracy of 68.87% using 16 classes. SVM demonstrated higher accuracy with a training accuracy of 92.62% and a validation accuracy of 68.67% using 16 sizes, which further improved to 92.61% and 92.97% for training and 85.38% and 88.85% for validation in 9 and 6 classes modelling, respectively. KNN achieved a training accuracy of 98.63% and a validation accuracy of 66.97%, which increased to 98.75% for training and 86.10% for validation with 6 classes. DT performed well with a training accuracy of 97.50% and a validation accuracy of 60.64%, which improved to 99.06% for training and 70.96% for validation in 9 classes and kept increasing to 99.3% and 74.97% in 6 size classes. RF exhibited the highest training accuracy of 99.83% but had a lower validation accuracy of 59.81% in 16 sizes. However, the validation accuracy increased to 73.45% for training and 77.92% for validation with 9 and 6 classes, respectively.

Table 5.4: Size selection model training and validation accuracies

Algorithm	16 Sizes		9 Sizes		6 Sizes	
	Training Accuracy	Validation Accuracy	Training Accuracy	Validation Accuracy	Training Accuracy	Validation Accuracy
LR	91.13	68.87	90.12	78.67	90.65	82.14
SVM	92.62	68.67	92.61	85.38	92.97	88.85
KNN	98.63	66.97	98.63	82.63	98.75	86.10
DT	97.50	60.64	99.06	70.96	99.30	74.97
RF	99.83	59.81	99.07	73.45	99.48	77.92

5.2.5 Model Test on New Dataset

The same wells used in production data model were utilised in this model (*section 4.2.1.5.1*). Data of the seven wells, XFE26, XH7, XJ14, XM334, XF66, XK9, XSE2, ranging from 2020 to 2021 was used with different AL sizes deployed during that period. **Table 5.5** summarises the test accuracies obtained by each algorithm in the classification from the unseen dataset. The RF model achieved the highest accuracy among the three model runs, scoring 69.68%, 75.45%, and an impressive 92.42%. LR and SVM followed closely with the second highest accuracy of 75.45%, while KNN obtained a score of 69.43%. On the other hand, DT exhibited the lowest accuracy and prediction performance, achieving only 51.87% compared to its superlative results in AL selection from the production dataset.

Table 5.5: Size selection model test accuracies

Algorithm	13 Classes Accuracy %	8 Classes Accuracy %	6 Classes Accuracy %
LR	60.53	66.31	75.45
SVM	64.14	69.92	75.45
KNN	60.29	63.90	69.43
DT	65.94	51.87	57.40
RF	69.68	75.45	92.42

5.2.6 Validation with Field Data Results and Discussion

It is evident that the increased number of classes had a negative impact on prediction accuracy, leading to a decline in model performance. The reasoning behind this lies in the dynamic nature of AL sizes throughout an oil well's lifespan. As the well continues to produce over the years, the distribution of AL sizes changes due to replacements made for various reasons, such as low productivity, failure, or the implementation of IOR or EOR methods. Consequently, certain AL sizes may appear in the training dataset during a specific production period but may disappear in the test dataset during another production period. This variation makes it challenging for the model to identify the optimal size to select, even though the same AL type exists. This can be clearly seen in classification reports in **Figs. 5.10** and **5.11** with 13 and 8 classes respectively.

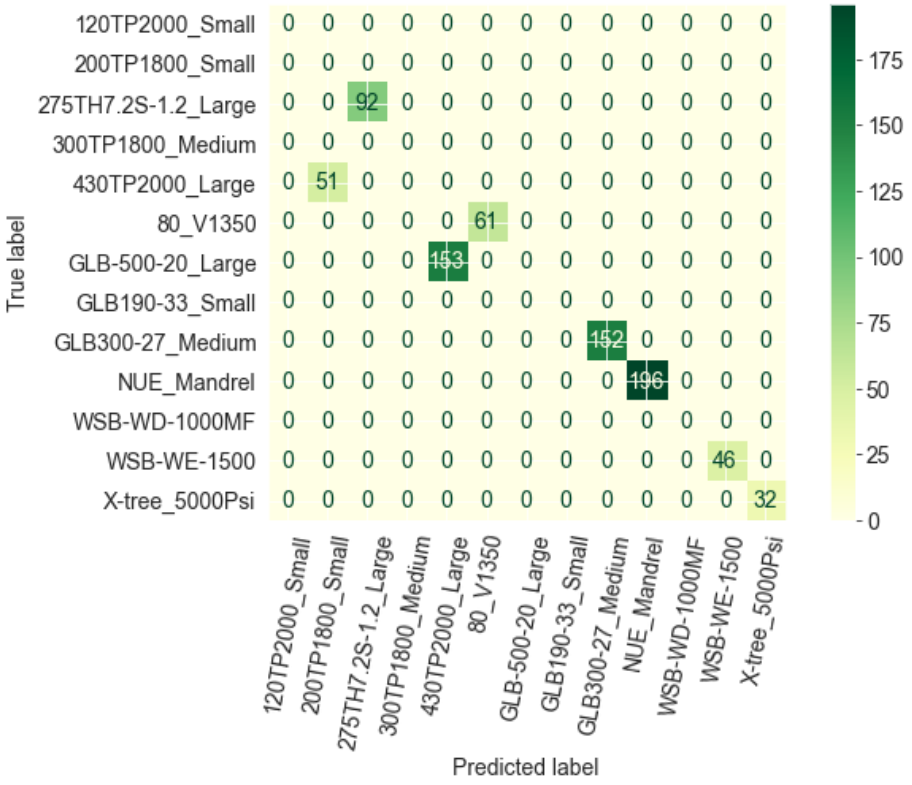


Fig. 5.10: 13 sizes confusion matrix

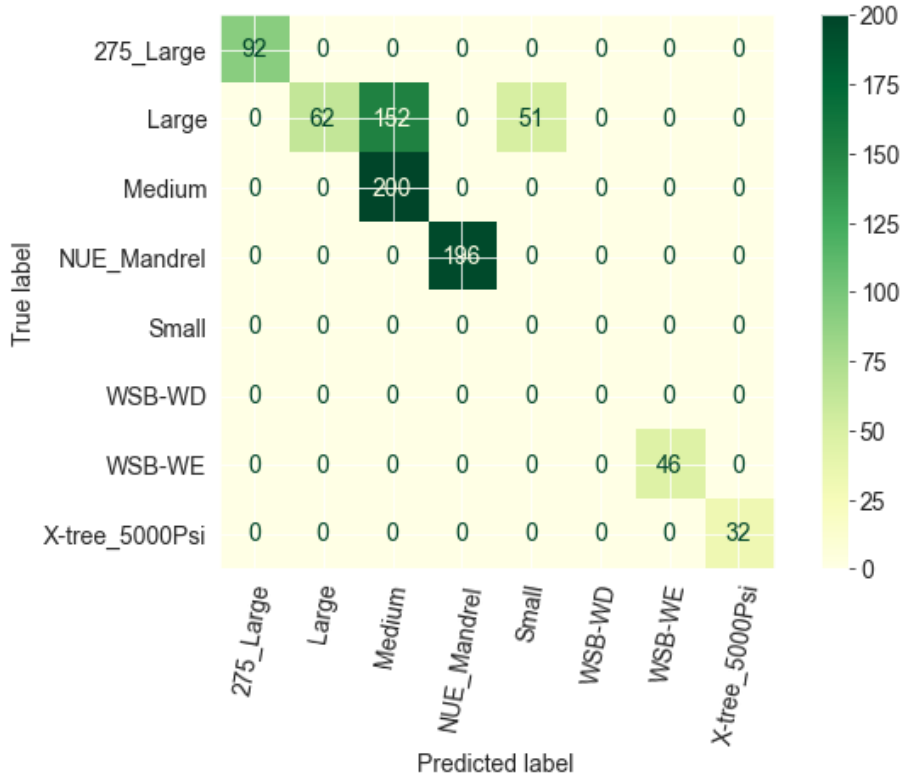


Fig. 5.11: 8 sizes confusion matrix

The model's highest accuracy, achieved by RF at 92.42%, was based on six AL size classes: Large, Medium, and Small for PCP, BPU, and MTM_PCP, WSB for ESP, NUE-Mandrel for GL, and X-tree_5000psi used for naturally flowing wells. The 7.5% prediction error observed in RF was attributed to two wrongly predicted sizes, Small and Medium, whereas the actual size was Large, as indicated in **Fig. 5.12**. The occurrence of such errors can be attributed to the similarity in certain parameters, such as flow rates, pressures, production years, and WC%, recorded for different AL sizes after years of production. The findings highlight the complexities associated with predicting AL sizes accurately over the long duration of an oil well's production life. The varying well and reservoir properties that affect well deliverability over time, coupled with the need for AL size replacements, present challenges for the model. Despite this, RF demonstrated promising accuracy, paving the way for further investigation into refining the model to account for the dynamic changes in AL sizes during oil well operations.

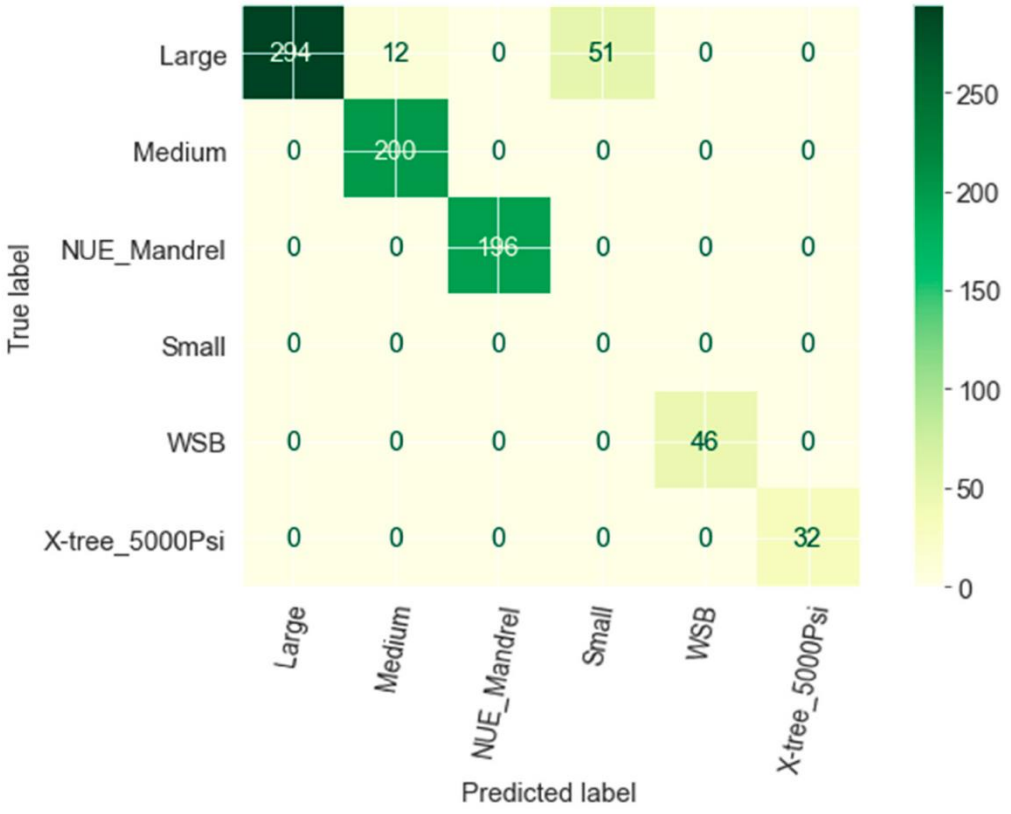


Fig. 5.12: 6 sizes confusion matrix

5.2.7 Model Test on Unlabelled Dataset

The size selection model was subjected to rigorous testing on an unlabelled sample dataset, and the results proved to be highly promising. The model achieved an impressive accuracy of 100% when compared to the actual size currently in use at the field. This outstanding level of accuracy demonstrates the model's capability to accurately predict the appropriate size of AL methods based on the given production data. The successful test results indicate that the size selection model has the potential to revolutionize size selection processes in the OGI, leading to enhanced production efficiency and optimized field performance. The high accuracy score provides strong evidence of the model's reliability and effectiveness, making it a valuable tool for industry practitioners seeking to make informed and precise decisions in AL size selection for their oil wells.

5.3 Summary

The chapter concludes by emphasizing the unique contribution of the research to the existing body of knowledge on AL selection. The findings provide valuable insights for industry practitioners and researchers, offering a more comprehensive and precise approach to AL size selection. The significance of this research lies in its potential to improve the efficiency and effectiveness of AL size selection processes in the oil and gas industry. By leveraging production data and ML algorithms, operators can make informed decisions regarding the size of the AL method to be deployed, leading to enhanced production rates, reduced operational costs, and improved overall field performance.

CHAPTER 6

ANALYSIS AND DISCUSSION OF THE IMPORTANT SELECTION FEATURES, ARTIFICIAL LIFT SELECTION MODELS AND SIMULATION RESULTS

6.1 Introduction

This chapter delves into a comprehensive exploration of the crucial features that guide AL selection within the context of three distinct classification models. The chapter illuminates the factors driving the selection models and contributes to a deeper understanding of the intricate interaction between field parameters and AL methods (feature importance tables are in **Appendices B1-B5**). A comparison to recent studies is also presented. Moreover, this chapter extends its investigation to assess the production performance and economic evaluation of the AL methods predicted by these ML models. A comparative analysis is conducted by comparing the simulation outcomes of the predicted AL methods with the results of the actual AL methods currently deployed in the field. By evaluating the simulation outputs of predicted AL methods in contrast to their real-world counterparts, an assessment of the predictive capabilities of the classification models is achieved.

6.2 Critical Field Parameters of Artificial Lift and Size Selection

6.2.1 Critical Production Parameters

The AL selection models were employed to underscore the pivotal variables influencing both AL and size selection within the field. **Fig. 6.1** presents an illustrative picture of the fundamental features, primarily utilised by the RF model, which profoundly influence AL classification. The cumulative significance score attributed to the assorted features corresponds to 1. Notably, the proximity of the score to unity emphasises the highest importance of the respective feature. The discernible factors are gas and produced fluid, exerting substantial influence on the classification, with coefficients of 0.23 and 0.14, respectively. Wellhead pressure follows closely behind in importance with a coefficient of 0.13. Conversely, oil exhibits the lowest significance, with a coefficient of merely 0.04. It is evident that the algorithm relies on these features, particularly gas, produced fluid, and wellhead pressure, as cornerstones in classifying lifting methods. This underscores their indispensability in the AL selection process. The results thus

emphasise the imperative necessity of comprehensive analysis of these parameters in both ongoing and potential oil well projects, given their direct implications on production performance.

Examining the historical treatment of the gas feature within extant selection methodologies in both literature and practical implementation reveals its adversarial nature to most lifting methods, excluding GL and NF. Notably, the model adeptly discerned the substantive influence of the gas feature, elucidating its profound impact on AL selection and consequential production performance. Similarly, other salient factors illuminated by the study encompass produced fluid, GOR, produced water, thermal recovery implementation, and the years of production. While flow rate maintains its criticality across the selection methodologies outlined in existing literature (Clegg et al., 1993; Neely et al., 1981; Brown, 1982; Heinze and Winkler, 1995; Adam et al., 2022), it is essential to acknowledge that prevalent studies predominantly focus on the flow rate constraints unique to each AL. These prescribed operating thresholds present variable extents across the literature. Our model, conversely, engages in a comprehensive assessment of both daily and cumulative fluid production over the entire operational lifespan of the AL. The outcome is an optimal selection that not only extends the AL's operational longevity but also enhances production performance, an elucidation that will be expounded upon in the subsequent section dedicated to production performance simulations.

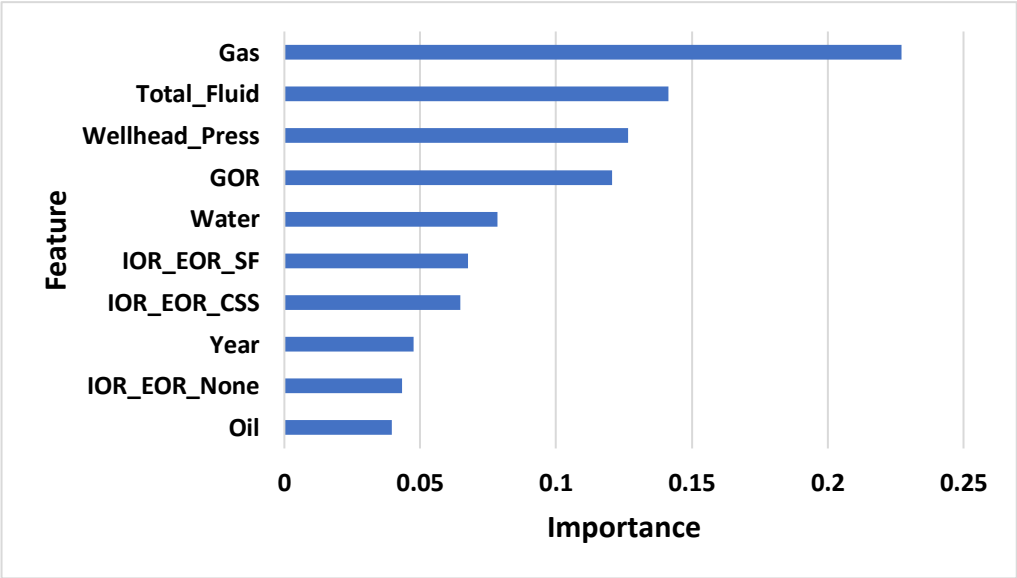


Fig. 6.1: Important AL selection production features

Fig. 6.2 shows the crucial factors underpinning the process of AL size selection. In contrast to the critical determinants governing AL selection, a notable distinction in this context is the significance attributed to the lifting methods used as inputs. It is pertinent to acknowledge that similar to AL selection, the remaining parameters retain an equivalent degree of significance. However, cumulative production holds greater importance than gas, with a coefficient of 0.12 compared to 0.11 for gas. Wellhead pressure ranks lower on the list, with a coefficient of 0.05, followed by WC% with the least value of 0.03.

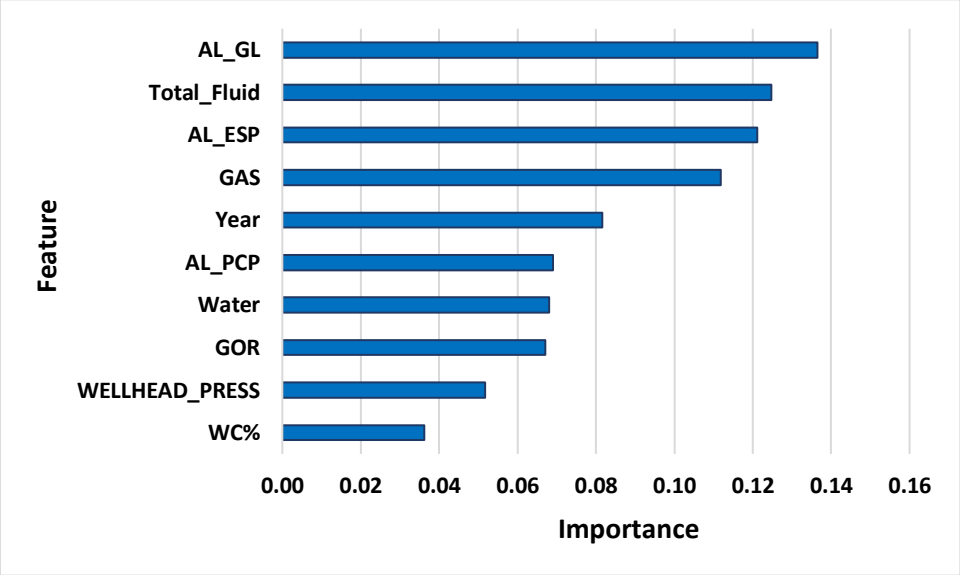


Fig. 6.2: Important size selection features

6.2.2 Critical Operation Parameters

Operational parameters equally assume a critical role in the process of AL selection; however, it is notable that the model's accuracy pertaining to these parameters is comparatively lower than that derived from production parameters. As shown in **Fig. 6.3**, the depth dominates the classification, in particular, the setting depth with the highest significance of 0.26. The determination of the AL setting depth constitutes a pivotal feature in AL selection, as its determination is contingent upon fluid level, reflecting the distance from the surface to the fluid face in a production well. This distance, subject to continuous fluctuations during production (Shedid, 2009). Thus, if the setting depth is not accurately determined will lead to dry pump run and total damage of the lifting method. Key factors such as workover frequency and underlying causes, formation thickness (affording a representation of the reservoir's volumetric capacity), and AL operational period

contribute significantly to AL selection, with coefficients ranging from 0.06 to 0.02, and 0.05 respectively. These variables play pivotal roles in shaping the AL selection process by providing insights into the operational dynamics of the reservoir and the AL's longevity.

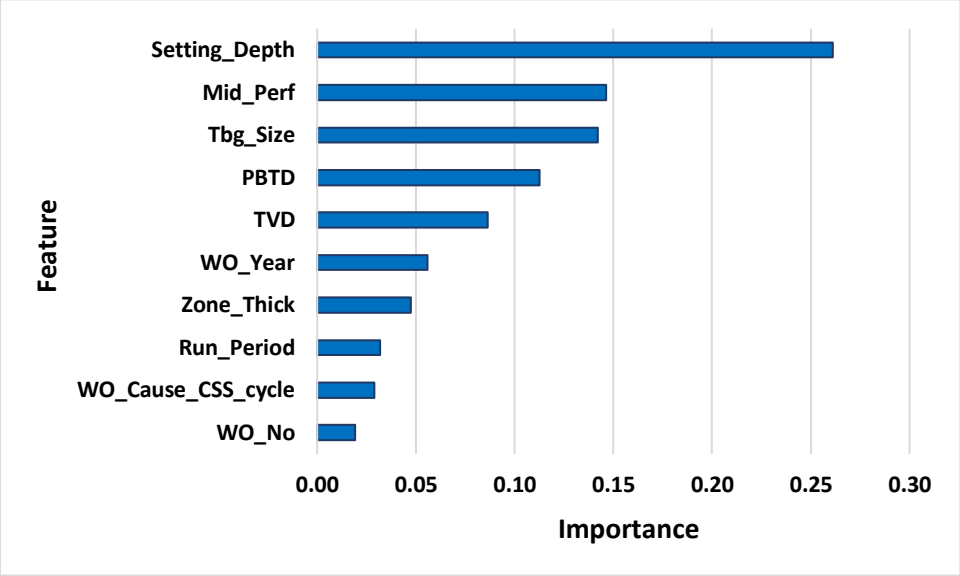


Fig. 6.3: Important AL selection operation features

6.2.3 Critical Environmental and Economic Parameters

Our exploration extended to assessing the importance of environmental and economic parameters within the context of AL selection. Environmental features were used as encoded categorical features as introduced in Chapter 4. The levels were ranged according to the selected oil field company's HSE reports. **Fig 6.4** demonstrates a comprehensive investigation of the gradation of each factor. Evidently, the pivotal determinant lies within the domain of AL purchasing, wielding a comprehensive influence over the overall selection process with 0.21 importance. Subsequently, the knowledge of field personnel exerts a tangible influence, substantiating its consequential role with 0.12. The medium amount of noise and oil spill share an equivalent significance of 0.1. The no-requisite for power comes next as naturally flowing wells reflects the lowest expenses amongst other lifting methods. It is noteworthy that while the AL price feature attains the summit of importance, the prevailing prominence within the feature importance hierarchy is appropriated by the array of environmental attributes. This assertion can be seen in the absence of the other economic parameters integral to the selection model, namely, workover and completion costs. Such parameters,

distinctly variable for each AL based on diverse field surface and downhole conditions, merge to highlight the complex interrelationship between economic aspects and environmental considerations.

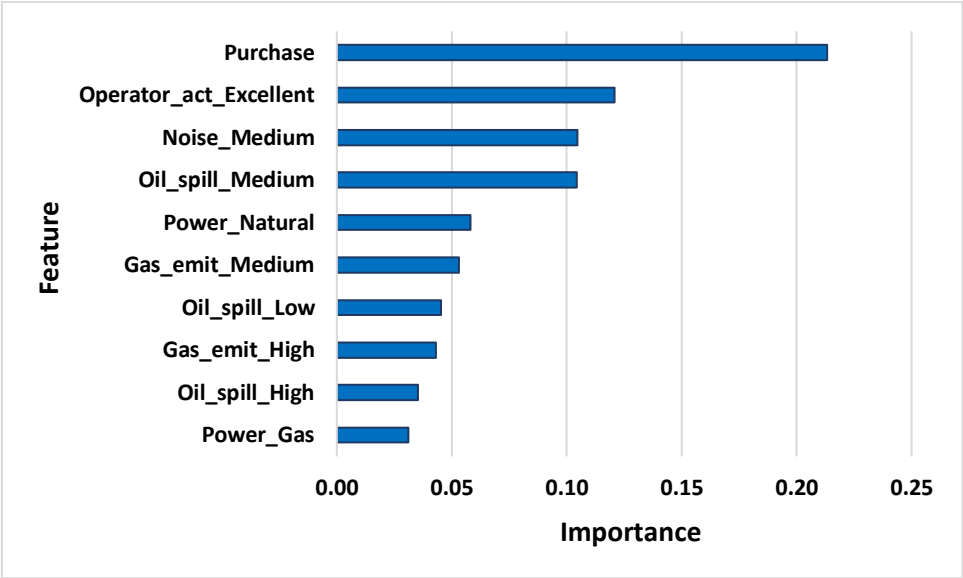


Fig. 6.4: Important AL selection environmental and economic features

6.3 Sensitivity Analysis of the Best Selection Models

Here we conduct a comparative analysis, evaluating the highest accuracies achieved by each algorithm. It is essential to acknowledge the inherent complexity and uncertainty associated with field parameters, which often pose challenges for engineers and algorithms in the quest to analyse and determine the optimal lifting method. However, it is noteworthy that certain algorithms, such as RF and DT, exhibit a remarkable capability to establish meaningful connections among field parameters and discern their interrelations with the most suitable AL method.

The effect of field data heterogeneity on AL selection modelling can be evidently seen in **Fig. 6.5**, particularly production and operation dataset. In which, AL selection accuracies fluctuate compared to those obtained from environmental and economic dataset. On the other hand, the AL selection process using DT and RF shows a considerable harmonisation across the three types of data, where RF outweighs DT in size selection using production data, as shown in **Fig 6.6**. The superiority of RF among the other algorithms is an outcome of utilising cumulative production rate, AL setting depth, and AL price as the determinants of the

optimum AL method. These features are the considered the core of AL selection in the literature, however, the literature uses those significant parameters qualitatively for AL elimination and ignores the important analysis part that performed by the algorithm.

The heterogeneity of field data, characterized by its diverse and sometimes elusive nature, can indeed hinder the analytical process. Nonetheless, the robustness of RF and DT in uncovering these complicated connections between field parameters and AL selection underscores their potential utility as valuable tools in the decision-making process for AL deployment. This study's findings emphasise the significance of employing advanced algorithms like RF and DT to navigate the intricacies of field data effectively, ultimately contributing to more informed and optimized AL selection strategies within the oil and gas industry.

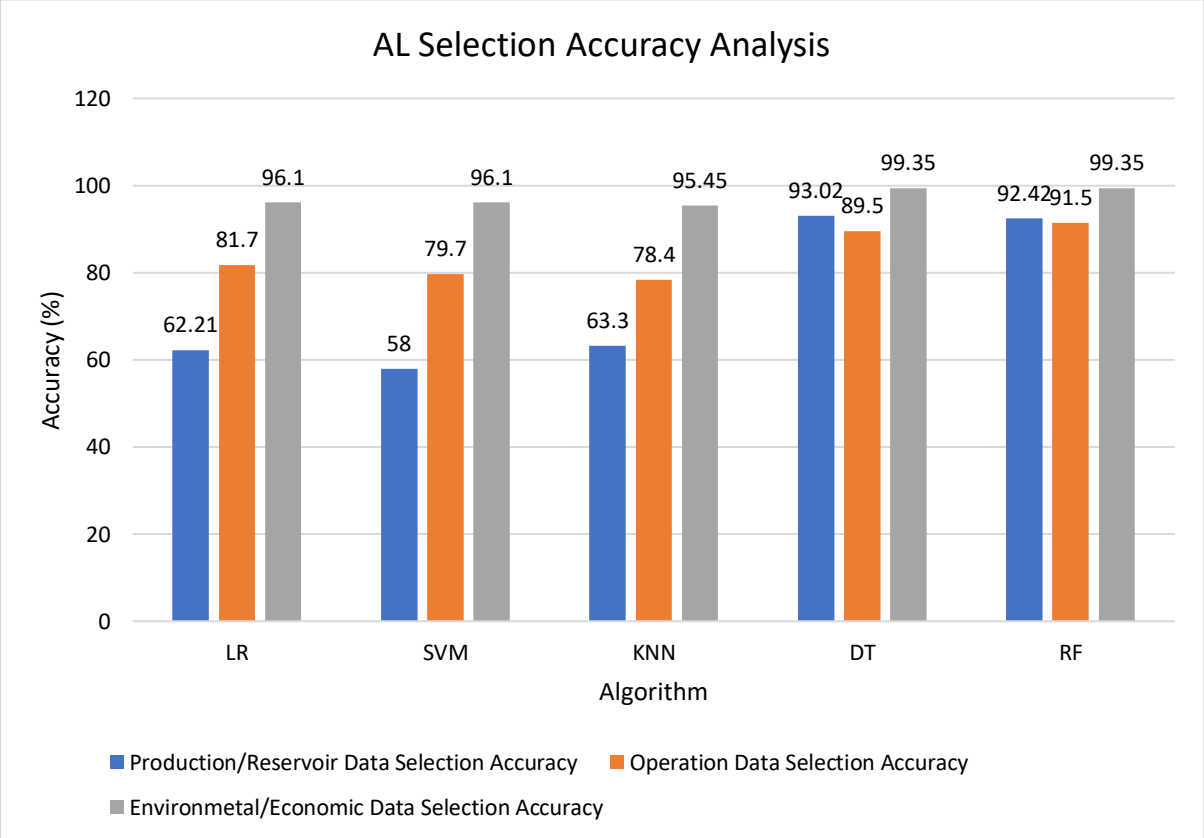


Fig. 6.5: AL Selection Accuracy Sensitivity Analysis

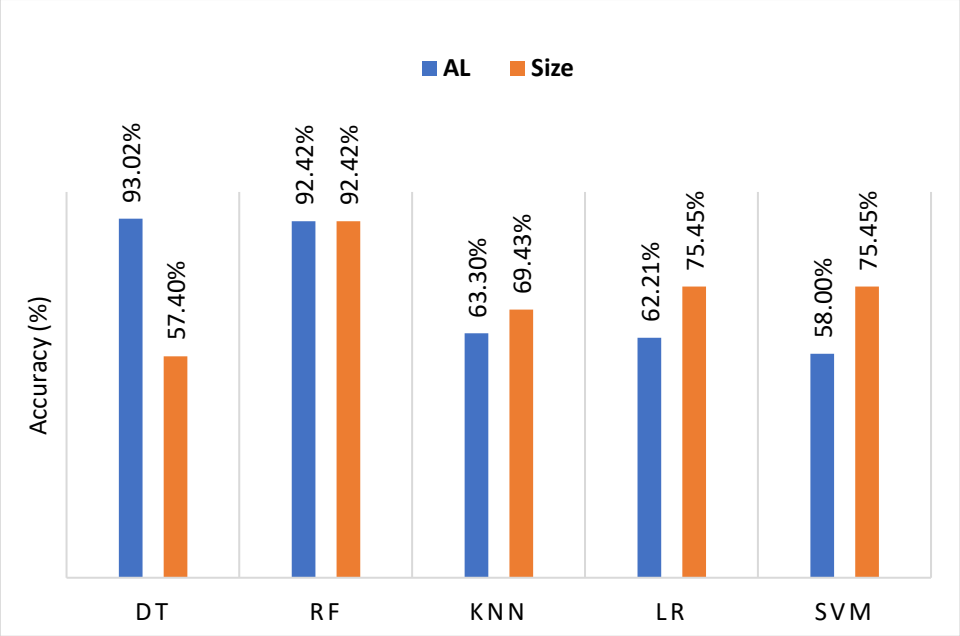


Fig. 6.6: Size Selection Accuracy Sensitivity Analysis

6.4 Comparison to Recent Studies

The ML application in AL selection is still a premature process, as one recent study conducted by (Ounsakul et al., 2019) shows, in addition to this study recent published work (Mahdi et al. 2023). **Table 6.1** compares this study production model to (Ounsakul et al., 2019) selection results. The highest model training accuracy obtained by (Ounsakul et al., 2019) was 94%; no testing scores were mentioned. This study model's testing accuracy is 93%, indicating satisfactory performance in predicting the optimum lifting methods. The study of this thesis proves that the selection could be accomplished using a specific dataset and obtaining the highest accuracy. Another recent study was presented by (Crnogorac et al., 2020) to select the optimum AL using fuzzy logic and mathematical models. However, their model is conditioned on an enclosed data inventory of five lifting methods and would not be applicable if different input parameters or other ALs are used instead. Moreover, (Crnogorac et al., 2020) tested their model on one well which did not reflect the model robustness. In contrast, the model developed in this thesis is unrestricted to specific datasets (any other data and AL can be modelled) and tested on eight wells to substantiate the results.

Table 6.1: AL selection results in comparison to a recent study using ML

Category	This Study	(Ounsakul et al., 2019)
Number of used ML algorithms	5 (LR, SVM, KNN, RF, DT)	3 (DT, Naïve Bayes, Neural Network)
Max training accuracy (%)	99.8 (scored by DT)	94 (scored by DT)
Max test accuracy (%)	93.02	N/A
Number of wells	24	9
Number of samples	474,656	30,000
Number of AL	4 + NF	4

6.5 Comparison to Commercial Software Results and Discussion

The methodology employed involves a comparative analysis of the production performance between the currently deployed AL method in the field and the AL method predicted by the ML model. This comparative evaluation seeks to ascertain the most suitable AL method for field deployment, considering factors such as flow rate, running period, and operational costs.

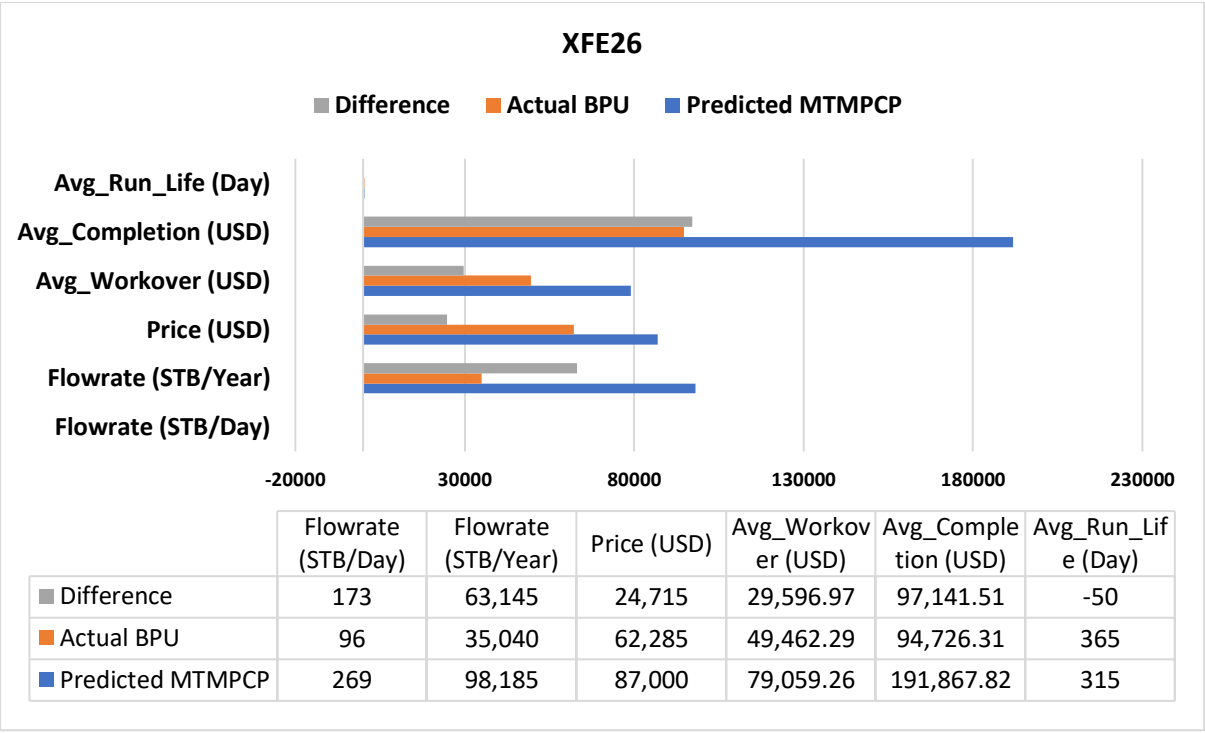
This validation approach is pivotal in addressing the challenge of AL selection, as it leverages the ML model's predictive capabilities to assess the performance of different AL methods under real-world conditions. By systematically comparing actual field performance with the model's predictions, we can determine which AL method is most appropriate for deployment in the field, offering valuable insights into the potential enhancement of production rates and operational cost-efficiency. This process serves as a practical and data-driven means of optimizing AL selection for enhanced oil and gas production operations.

A simulation was conducted using *PROSPER* and *PIPESIM* simulators to compare the production performance of the predicted AL for producing wells XFE26 and XJS9, respectively. These wells have been selected for simulation because the two models using production and operation dataset, predicted the same AL for both wells. XFE26 undergoes CSS and oil is produced by BPU using a thermal sand control pump (275TH7.2S-1.2 Grade-III, Large). The ML model predicted MTM_PCP to employ instead of the current method BPU. In the simulation (PROSPER), XFE26's production was modelled using a MET-80V1000 medium-sized thermal pump. On the other hand, XJS9 is naturally flowing and employs an

X-tree_5000psi, with the simulation matching the predicted AL by employing three GL valves. **Appendices C1** and **C2** contain the field parameters used in the simulation of XFE26 and XJS9.

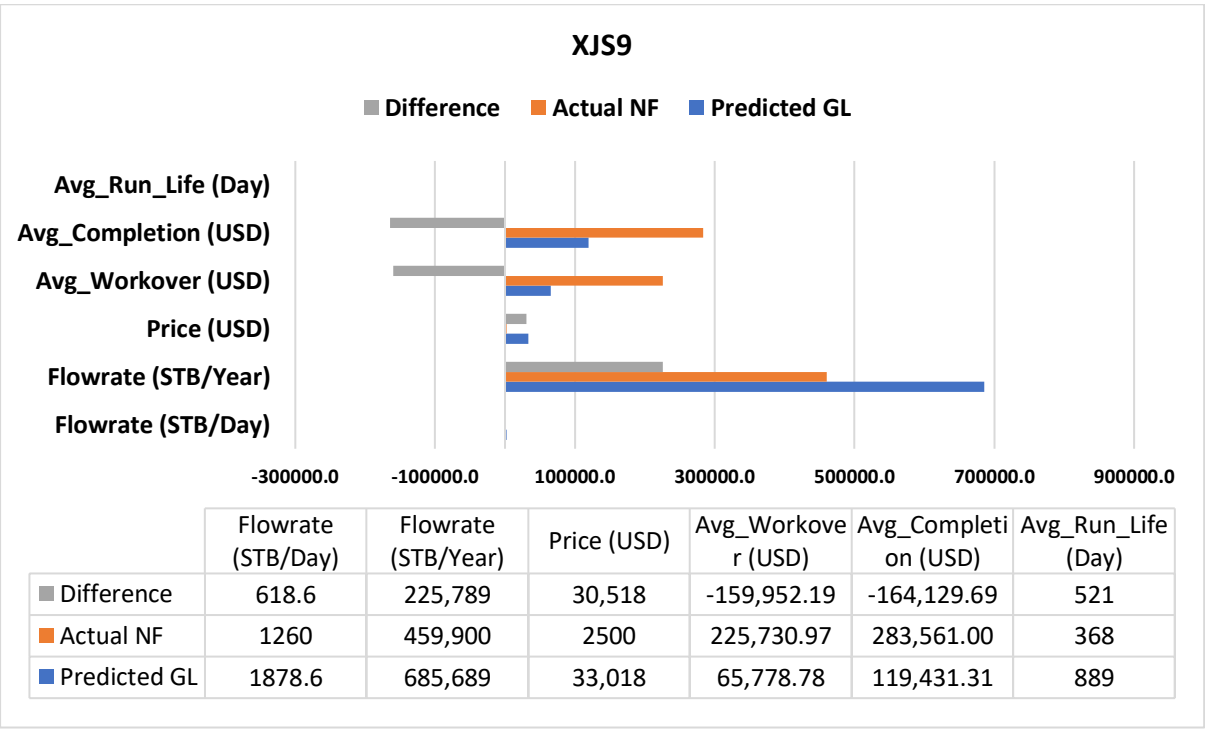
The simulation results demonstrated that the ML-predicted ALs outperformed the currently installed AL methods. Specifically, XFE26, when utilizing MTM_PCP, exhibited a production rate of 269 STB/D, surpassing the current 97 STB/D achieved with BPU. Similarly, XJS9, when operated with GL, exhibited a production rate of 1878 STB/D, a notable improvement over its current natural flow production rate of 1260 STB/D.

Furthermore, **Figs 6.7** and **6.8** present sensitivity analysis of both the actual and predicted lifting methods, comparing their flow rates, average operational costs, and run life. It is noteworthy that the anticipated AL methods may entail higher operational costs; however, the substantial increase in production revenues, exceeding 3 million USD for XFE26 and 11 million USD for XJS9, is anticipated to more than offset these operating expenses. These case study findings hold great promise for broader applications in AL selection for prospective wells, showcasing the potential for improved production performance and profitability.



	Flowrate (STB/Day)	Flowrate (STB/Year)	Price (USD)	Avg_Workover (USD)	Avg_Completion (USD)	Avg_Run_Life (Day)
Difference	173	63,145	24,715	29,596.97	97,141.51	-50
Actual BPU	96	35,040	62,285	49,462.29	94,726.31	365
Predicted MTMPCP	269	98,185	87,000	79,059.26	191,867.82	315

Fig. 6.7: XFE26 sensitivity analysis of actual BPU and predicted MTMPCP



	Flowrate (STB/Day)	Flowrate (STB/Year)	Price (USD)	Avg_Workover (USD)	Avg_Completion (USD)	Avg_Run_Life (Day)
Difference	618.6	225,789	30,518	-159,952.19	-164,129.69	521
Actual NF	1260	459,900	2500	225,730.97	283,561.00	368
Predicted GL	1878.6	685,689	33,018	65,778.78	119,431.31	889

Fig. 6.8: XJS-9 sensitivity analysis of actual NF and predicted GL

6.6 Summary

This exploration bridges the gap between theoretical predictions and practical outcomes, enabling stakeholders to make well-informed decisions regarding AL implementation. The integration of advanced ML techniques not only empowers the selection process but also lays the foundation for future advancements in operational strategies. The chapter unravels the intricate relationship between AL selection crucial features, data heterogeneity effect on AL selection models, and production performance simulation, offering a comprehensive perspective on the dynamic landscape of field management in the oil and gas sector.

CHAPTER 7

CONCLUSIONS AND RECOMMENDATIONS

In essence, this thesis has sought to present a detailed and novel AL selection technique using ML. The present chapter is divided into two main parts. In the first part, general conclusions address the aim, objectives, and outcomes of the research study. In the second part, recommendations are also made for further work on this fascinating field of study.

7.1 Conclusions

The primary accomplishment of this investigation lies in the creation of a versatile model for AL selection, which consistently predicts the optimum AL with accuracy exceeding 90%. Remarkably, the predicted AL methods demonstrated superior performance compared to those actually deployed in the field, delivering enhanced production rates and revenues. Furthermore, the model has effectively identified and underscored the critical factors influencing AL selection. The ensuing section outlines the key conclusions derived from the research undertaken in this thesis.

- The AL selection process in the OGI plays a pivotal role in enhancing well productivity and optimizing field operations. Over the years, this process has followed the same selection approach, which is considered outdated. This research set out to contribute to and evolve this field by employing ML techniques to improve AL selection and subsequently advance our understanding of this crucial task.
- The literature review conducted for this research provided a comprehensive overview of the existing methodologies and challenges in AL selection. Historically, the process relied heavily on qualitative methods, engineering expertise, and simplified decision trees. These traditional approaches often resulted in suboptimal selections and limited the potential for improving production and revenue. The literature review emphasized the importance of optimal AL in optimizing hydrocarbon production and reducing operational expenses, highlighting the ongoing challenges associated with data heterogeneity, data analysis complexity, and the need for precise feature selection. This research examined these challenges and paved the way for the introduction of ML to address them.

- One of the standout aspects of this research was the novel methodology workflow adopted to address AL selection. Rather than combining all types of field data for analysis, three separate models were developed, each tailored to a specific dataset: production, operation, environmental and economic. This criterion significantly expedited the AL selection process by eliminating the time-consuming task of data integration.
- The results of these models were nothing short of impressive. The predicted AL methods exhibited exceptional production performance, surpassing the capabilities of the current ALs in many instances. This research showcased that AL selection can be performed effectively based on the analysis of specific data types, simplifying the process, and making it more efficient.
- The application of supervised learning models, including LR, SVM, KNN, DT, and RF, demonstrated remarkable success in AL selection. By using historical data, these models could accurately predict the most suitable lifting method for a given well. The excellent outcomes of these models have contributed significantly to the field's knowledge, showcasing that data-driven approaches can surpass traditional qualitative methods. In particular, RF model emerged as a standout performer, achieving accuracy scores above 90% across the three models. This underscores the potential of ML in addressing the complexities of AL selection and maximizing production efficiency.
- Clustering techniques were employed to group wells with similar characteristics, offering a streamlined approach to AL selection. The clustering process, via K-Means, facilitated the identification of patterns within production, operation, and environmental/economic datasets, ultimately leading to more informed decisions. The unsupervised nature of clustering allowed the model to uncover hidden patterns and understand feature distribution and their significance for AL selection.
- A significant contribution of this research was the identification of critical factors that influence AL and size selection. These factors, which include gas and GOR, daily and cumulative produced fluid, wellhead pressure, depths, AL running period, and AL price, were found to be central in determining the most suitable lifting method. Recognizing the importance of these factors will enable engineers and field operators to make more informed decisions, leading to improved production performance.

- Through simulation and sensitivity analyses, the research showcased that the ML-predicted AL methods outperformed the current AL methods in terms of flow rates, revenue generation, and operational costs. Well-A is projected to produce 269 STB/D (equating to over 3 million USD), while Well-B is expected to yield 1878 STB/D (resulting in 11 million USD), compared to their current production rates of 97 and 1260 STB/D, respectively. These findings underscore the potential financial benefits of adopting ML-based AL selection.
- In conclusion, this research represents a significant step forward in the field of AL selection within the OGI. The integration of ML techniques offers a data-driven approach that optimizes AL selection, improves production performance, and enhances operational efficiency. The novel methodology developed in this study, focusing on production, operation, environmental and economic datasets, streamlines the selection process and provides critical insights into the factors that influence AL choice. The results have broad implications for the industry, demonstrating that ML-based AL selection can yield substantial financial benefits and operational improvements. As the oil and gas sector continues to evolve, embracing innovative approaches like ML will be essential for staying competitive and maximizing returns in an ever-changing landscape.

7.2 Recommendations for Further Work

As this research concludes, several recommendations and suggestions for future work emerge. This section offers an evaluation of aspects within the study that offer potential for enhancement, in addition to delineating directions for future research.

- Firstly, continued data collection and analysis are essential to keep models up-to-date and reflective of evolving field conditions. Establishing a connection between the ML model and RTUs as well as real-time measurement instruments proves instrumental in the continuous monitoring of AL system performance. This linkage aids in promptly identifying the need for AL replacement based on the historical lifespan and production performance of the AL system. Regular model retraining and validation will ensure that they remain reliable tools for AL selection.
- While this thesis has successfully focused on the modelling of four common AL methods - PCP, BPU, ESP, and GL, there is significant potential in applying these

models to other lifting methods such as Jets and Plunger lift. Investigating these less common methods can further diversify the understanding of AL selection and provide valuable insights for a wider range of oil well scenarios.

- This thesis has primarily examined the application of the model on conventional oil wells. A promising recommendation is to extend the application of the model to unconventional wells. With the increasing extraction of unconventional oil sources in response to global energy demands, analysing how this model performs in such contexts can contribute to more comprehensive and adaptable decision-making in the OGI.
- In addition to the aforementioned recommendations, it is advisable to explore the development of a regression model that complements the AL selection classification model. This regression model would be designed to predict the production performance of the AL method selected by the classification model. By combining these two models, a more comprehensive understanding of the production history can be achieved. Such an approach would not only contribute to the field's knowledge but also expand the application of ML in AL selection methods within the OGI for more simplicity.

REFERENCES

Adam, A.M., Mohamed Ali, A.A., Elsadig, A.A. and Ahmed, A.A., 2022, March. An Intelligent Selection Model for Optimum Artificial Lift Method Using Multiple Criteria Decision-Making Approach. In Offshore Technology Conference Asia. OnePetro.

Ahmadi, M.A. and Chen, Z., 2019. Machine learning models to predict bottom hole pressure in multi-phase flow in vertical oil production wells. The Canadian Journal of Chemical Engineering, 97(11), pp.2928-2940.

Ahmed, M., Seraj, R. and Islam, S.M.S., 2020. The k-means algorithm: A comprehensive survey and performance evaluation. Electronics, 9(8), p.1295.

Alakbari, F.S., Elkatatny, S. and Baarimah, S.O., 2016, November. Prediction of bubble point pressure using artificial intelligence AI techniques. In SPE middle east artificial lift conference and exhibition. OnePetro.

Al-Alwani, M.A., Britt, L., Dunn-Norman, S., Alkinani, H.H., Al-Hameedi, A.T. and Al-Attar, A., 2019, June. Production performance estimation from stimulation and completion parameters using machine learning approach in the Marcellus Shale. In 53rd US Rock Mechanics/Geomechanics Symposium. OnePetro.

Al-Sidairi, N., Osman, M., Oritola, N., Ahmed, F., Zuhaimi, K., Amri, M., Hinai, G. and Al-Mahrooqi, M., 2018, March. Overcoming Artificial Lift Failures in Polymer Flooding Oil Field in the South of Oman. In SPE EOR Conference at Oil and Gas West Asia. OnePetro.

Al Selaiti, I., Mata, C., Saputelli, L., Badmaev, D., Alatrach, Y., Rubio, E., Mohan, R. and Quijada, D., 2020, October. Robust Data Driven Well Performance Optimization Assisted by Machine Learning Techniques for Natural Flowing and Gas-Lift Wells in Abu Dhabi. In SPE Annual Technical Conference and Exhibition. OnePetro.

Alemi, M., Jalalifar, H., Kamali, G. and Kalbasi, M., 2010. A prediction to the best artificial lift method selection on the basis of TOPSIS model. Journal of Petroleum and Gas Engineering, 1(1), pp.009-015.

Alemi, M., Jalalifar, H., Kamali, G.R., Kalbasi, M. and Research, P.E.D.E.C., 2011. A mathematical estimation for artificial lift systems selection based on ELECTRE model. Journal of Petroleum Science and Engineering, 78(1), pp.193-200

Alferov, A.V., Lutfurakhmanov, A.G., Litvinenko, K.V. and Zdolnik, S.E., 2015, October. Artificial lift strategy selection within field development planning. In SPE Russian Petroleum Technology Conference. OnePetro.

Almajid, H., Al Gamber, S., Abou Zeid, S. and Ramos, M., 2019, November. An Integrated Approach Utilizing ESP Design Improvements and Real Time Monitoring to Ensure Optimum Performance and Maximize Run Life. In Abu Dhabi International Petroleum Exhibition & Conference. OnePetro.

Alshmakhy, A., Al Daghar, K., Punnapala, S., AlShehhi, S., Ben Amara, A., Makin, G. and Faux, S., 2019, September. First Digital Intelligent Artificial Lift Production Optimization Technology in UAE Dual-String Gas Lift Well-Business Case and Implementation Plan. In SPE Annual Technical Conference and Exhibition. OnePetro.

Alshmakhy, A., Punnapala, S., AlShehhi, S., Ben Amara, A., Makin, G. and Faux, S., 2020, January. First Digital Intelligent Artificial Lift Production Optimization Technology in UAE Dual-String Gas Lift Well–Completion and Installation Considerations. In International Petroleum Technology Conference. OnePetro.

Alsiemat, M. and Gambier, P., 2016, November. New Artificial Lift Technology Changes the Game for Completion Design. In SPE Middle East Artificial Lift Conference and Exhibition. OnePetro.

Ameisen, E., 2020. Building Machine Learning Powered Applications: Going from Idea to Product. " O'Reilly Media, Inc.".

Andrianova, A., Simonov, M., Perets, D., Margarit, A., Serebryakova, D., Bogdanov, Y., Budennyy, S., Volkov, N., Tsanda, A. and Bukharev, A., 2018, October. Application of machine learning for oilfield data quality improvement. In SPE Russian Petroleum Technology Conference. OnePetro.

Anifowose, F.A., Labadin, J. and Abdulraheem, A., 2017. Ensemble machine learning: An untapped modeling paradigm for petroleum reservoir characterization. Journal of Petroleum Science and Engineering, 151, pp.480-487.

Bangert, P., 2019, March. Diagnosing and predicting problems with rod pumps using machine learning. In SPE Middle East Oil and Gas Show and Conference. OnePetro.

Bansal, A., Sharma, M. and Goel, S., 2017. Improved k-mean clustering algorithm for prediction analysis using classification technique in data mining. International Journal of Computer Applications, 157(6), pp.0975-8887.

Bearden, J., 2007. Electrical Submersible Pumps. In Petroleum Engineering Handbook, Vol. 4: Production Operations Engineering, ed. J. D. Clegg. Richardson, Texas, USA: Society of Petroleum Engineers.

Beckwith, R., 2014. Pumping oil: 155 years of artificial lift. Journal of Petroleum Technology, 66(10), pp.101-107

Berry, M., 2016. Technology Focus: Artificial Lift. Journal of Petroleum Technology, 68(07), pp.67-67.

Blais, R., 1986. Artificial Lift Methods. PennWell Publishing, Tulsa, Oklahoma.

Boguslawski, B., Boujonnier, M., Bissuel-Beauvais, L. and Saghir, F., 2018, November. Edge Analytics at the Wellhead: Designing Robust Machine Learning Models for Artificial Lift Failure Detection. In Abu Dhabi International Petroleum Exhibition & Conference. OnePetro.

Bowie, B., 2018. Machine learning applied to optimize Duvernay well performance. In SPE Canada Unconventional Resources Conference. OnePetro.

Breiman, L., 1996. Bagging predictors. Machine learning, 24(2), pp.123-140.

Brown, K.E., 1982. Overview of artificial lift systems. Journal of Petroleum Technology, 34(10), pp.2384-2396.

Brownlee, J., 2020. Data preparation for machine learning: data cleaning, feature selection, and data transforms in Python. Machine Learning Mastery.

Bucaram, S.M., 1994. Managing artificial lift. Journal of Petroleum Technology, 46(04), pp.335-340.

Caballeroa, D., Hurtado, Y., Gomez, A. and Zimmer, L., 2014, May. PCP run life improvement in Orinoco Belt with new PCP technology. In SPE Latin America and Caribbean Petroleum Engineering Conference. OnePetro.

Caicedo, S., Montoya, C., Abboud, J. and Tiar, S., 2015, November. A Systematic Integrated Approach to Evaluate Artificial Lift Requirements While Dealing With High Uncertainty. In Abu Dhabi International Petroleum Exhibition and Conference. OnePetro.

Cao, Q., Banerjee, R., Gupta, S., Li, J., Zhou, W. and Jeyachandra, B., 2016, June. Data driven production forecasting using machine learning. In SPE Argentina Exploration and Production of unconventional resources symposium. OnePetro.

Castillo, I.C., Valderrama, J. and Godoy, F., 2018, November. Improvement of PCP's Run Life in Highly Deviated and Tortuous Wells Using Novel Design in Rod Strings: A Case of Study. In SPE Middle East Artificial Lift Conference and Exhibition. OnePetro.

Chachula, R., Serafinchan, D. and Dietz, J., 2019, May. Need a Lift? A Rotary Gear Pump, an ESP System without the ESP Pump. In SPE Gulf Coast Section Electric Submersible Pumps Symposium. OnePetro.

Chiroma, H., Abdul-kareem, S., Shukri Mohd Noor, A., Abubakar, A.I., Sohrabi Safa, N., Shuib, L., Fatihu Hamza, M., Ya'u Gital, A. and Herawan, T., 2016. A review on artificial intelligence methodologies for the forecasting of crude oil price. Intelligent Automation & Soft Computing, 22(3), pp.449-462.

Choudhary, R. and Gianey, H.K., 2017, December. Comprehensive review on supervised machine learning algorithms. In 2017 International Conference on Machine Learning and Data Science (MLDS) (pp. 37-43). IEEE.

Chow, J., Gamboa, J., Garcia, G.A., Price, T. and Hall, C., 2020, November. Verifying feasibility of artificial lift methods in rapid selection tool. In SPE Artificial Lift Conference and Exhibition-Americas. OnePetro.

Clegg, J.D., Bucaram, S.M. and Hein Jr, N.W., 1993. Recommendations and Comparisons for Selecting Artificial-Lift Methods (includes associated papers 28645 and 29092). Journal of Petroleum Technology, 45(12), pp.1128-1167.

CNPC country reports, CNPC in Sudan 2009, https://www.cnpc.com.cn/en/crsinSudan/AnnualReport_list.shtml (accessed 15 January 2022).

Crnogorac, M., Tanasijević, M., Danilović, D., Karović Maričić, V. and Leković, B., 2020. Selection of Artificial Lift Methods: A Brief Review and New Model Based on Fuzzy Logic. Energies, 13(7), p.1758.

Daigle, H. and Griffith, N., 2018, September. Optimizing Nanoparticle-Stabilized Emulsion Behavior in Porous Media Through Electrostatic Interactions. In SPE Annual Technical Conference and Exhibition. OnePetro.

Dave, M.K. and Mustafa, G., 2017, November. Performance evaluations of the different sucker rod artificial lift systems. In SPE Symposium: Production Enhancement and Cost Optimisation. OnePetro.

Elichev, V., Bilogan, A., Litvinenko, K., Khabibullin, R., Alferov, A. and Vodopyan, A., 2019, October. Understanding well events with machine learning. In SPE Russian Petroleum Technology Conference. OnePetro.

Escobar Patron, K., Zhang, K., Xu, T., Lu, H. and Cui, S., 2018, September. Case study of artificial lift strategy selection and optimization for unconventional oil wells in the Williston Basin. In SPE Liquids-Rich Basins Conference-North America. OnePetro.

Espin, D.A., Gasbarri, S. and Chacin, J.E., 1994, April. Expert system for selection of optimum Artificial Lift method. In SPE Latin America/Caribbean Petroleum Engineering Conference. OnePetro.

Fatahi, E., Jalalifar, H., Pourafshari, P. and Moradi, B., 2012. Selection of the best artificial lift method for one of the Iranian oil field using multiple attribute decision making methods. International Journal of Engineering and Technology, 2(2), pp.188-193.

Fatahi, E., Jalalifar, H., Pourafshari, P. and Rostami, A.J., 2011. Selection of the best artificial lift method in one of the iranian oil field by the employment of ELECTRE model. British Journal of Applied Science & Technology, 1(4), p.172.

Fletcher, T., 2009. Support vector machines explained. Tutorial paper, pp.1-19.

Fonseca, D., Salazar, A., Gonçalves, T. and Villarreal, E., 2019. Electrical modelling of an electrical submersible pump system three-phase power cable used in power line communication. Przeglad Elektrotechniczny, 10(3), pp.22-26.

Fraga, R.S., Castellões, O.G., Assmann, B.W., Estevam, V., de Moura, G.T., Schröer, I.N. and do Amaral, L.G., 2020. Progressive Vortex Pump: A New Artificial Lift Pumped Method. SPE Production & Operations, 35(02), pp.454-463.

Ghareeb, M., Ellaithy, W.F. and Zahran, I.F., 2012, October. Assessment of Artificial Lifts for Oil Wells in Egypt. In SPE Annual Technical Conference and Exhibition. OnePetro.

Hajizadeh, Y., 2019. Machine learning in oil and gas; a SWOT analysis approach. Journal of Petroleum Science and Engineering, 176, pp.661-663.

Han, D., Kwon, S., Son, H. and Lee, J., 2020, February. Production forecasting for shale gas well in transient flow using machine learning and decline curve analysis. In Asia Pacific Unconventional Resources Technology Conference, Brisbane, Australia, 18-19 November 2019 (pp. 1510-1527). Unconventional Resources Technology Conference.

Harris, D., Banman, M. and Malone, D., 2019, May. Design and Qualification Testing of ESP Cable to Improve ESP System Run Life. In SPE Gulf Coast Section Electric Submersible Pumps Symposium. OnePetro.

Hastie, T., Tibshirani, R. and Friedman, J., 2009. Random forests. In The elements of statistical learning (pp. 587-604). Springer, New York, NY.

Hein, N.W., 2007. Sucker-Rod Lift. In Petroleum Engineering Handbook, Vol. 4: Production Operations Engineering, ed. J. D. Clegg. Richardson, Texas, USA: Society of Petroleum Engineers.

Heinze, L.R., Thornsberry, K. and Witt, L.D., 1989, March. AL: an expert system for selecting the optimal pumping method. In SPE Production Operations Symposium. OnePetro.

Heinze, L.R., Winkler, H.W. and Lea, J.F., 1995, April. Decision Tree for selection of Artificial Lift method. In SPE Production Operations Symposium. OnePetro.

Herve, P., Prado, G. and Rosner, M., 2020, May. How Machine Learning is Improving Production on Offshore Platforms. In Offshore Technology Conference. OnePetro.

Hoy, M., Knauhs, P., Langbauer, C., Pratscher, H.P., Cimitoglu, T., Marschall, C., Puls, C. and Hurch, S., 2020, November. Artificial Lift Selection and Testing for an EOR Redevelopment Project–Lessons Learned from Field Pilots, Laboratory and

Pump Test Facilities. In SPE Artificial Lift Conference and Exhibition-Americas. OnePetro.

Hoy, M., Kometer, B., Bürßner, P., Puscalau, G. and Eder, S., 2018, August. SRP equipment customization creating value by increasing run life in a low oil price environment. In SPE Artificial Lift Conference and Exhibition-Americas. OnePetro.

Jolliffe, I.T., 2002. Principal component analysis for special types of data (pp. 338-372). Springer New York.

JPT staff, _., 2014. Techbits: Artificial Lift Selection Discussed at Workshop. Journal of Petroleum Technology, 66(03), pp.38-40.

Jurafsky, D. and Martin, J.H., 2021. Speech and Language Processing, (draft) edition, chapter 5.

Kadio-Morokro, B., Curay, F., Fernandez, J. and Salazar, V., 2017, April. Extending ESP run life in gassy wells application. In SPE electric submersible pump symposium. OnePetro.

Kaplan, V. and Duygu, E., 2014, May. Selection and Optimization of Artificial Lift System in Heavy Oil Fields. In SPE Latin America and Caribbean Petroleum Engineering Conference. OnePetro.

Kefford, P.A. and Gaurav, M., 2016, September. Well performance calculations for artificial lift screening. In SPE Annual Technical Conference and Exhibition. OnePetro.

Khabibullin, R.A. and Krasnov, V.A., 2015, October. An approach for artificial lift applicability maps construction. In SPE Russian Petroleum Technology Conference. OnePetro.

Khadav, S., Agarwal, S., Kumar, P., Pandey, N., Parasher, A., Kumar, S., Agarwal, V. and Tiwari, S., 2018, August. System run life improvement for rod driven PCP in high deviation well. In SPE Artificial Lift Conference and Exhibition-Americas. OnePetro.

Khan, M.R., Alnuaim, S., Tariq, Z. and Abdulraheem, A., 2019, March. Machine learning application for oil rate prediction in artificial gas lift wells. In SPE middle east oil and gas show and conference. OnePetro.

Khan, N., Ganzer, L., Elichev, V. and Ali, N., 2014, May. An integrated life-time artificial lift selection approach for tight/shale oil production. In SPE Hydrocarbon Economics and Evaluation Symposium. OnePetro.

Kharrat, A., Benamrane, N., Messaoud, M.B. and Abid, M., 2009, November. Detection of brain tumor in medical images. In 2009 3rd International conference on signals, circuits and systems (SCS) (pp. 1-6). IEEE.

Kolawole, O., Gamadi, T.D. and Bullard, D., 2020. Artificial lift system applications in tight formations: The state of knowledge. SPE Production & Operations, 35(02), pp.422-434.

Kotsiantis, S.B., Kanellopoulos, D. and Pintelas, P.E., 2006. Data preprocessing for supervised leaning. International journal of computer science, 1(2), pp.111-117.

Kotsiantis, S.B., Zaharakis, I. and Pintelas, P., 2007. Supervised machine learning: A review of classification techniques. Emerging artificial intelligence applications in computer engineering, 160(1), pp.3-24.

Krawczyk, B., 2016. Learning from imbalanced data: open challenges and future directions. Progress in Artificial Intelligence, 5(4), pp.221-232.

Kumar, S. and Chong, I., 2018. Correlation analysis to identify the effective data in machine learning: Prediction of depressive disorder and emotion states. International journal of environmental research and public health, 15(12), p.2907.

Lane, W. and Chokshi, R., 2014, August. Considerations for optimizing artificial lift in unconventionals. In SPE/AAPG/SEG Unconventional Resources Technology Conference. OnePetro.

Lanier, G.H. and Mahoney, M., 2009. Pushing the Limit: High-Rate-Artificial-Lift Evaluation for a Sour, Heavy-Oil, Thermal EOR Project in Oman. SPE Production & Operations, 24(04), pp.579-589.

Lapi, S.G., Arisman, B. and Johnson, M.E., 2014, January. Artificial Lift Performance Enhancements by applying Root Cause Failure Analysis. In International Petroleum Technology Conference. OnePetro.

Lastra, R., 2017, April. Achieving a 10-Year ESP Run Life. In SPE Electric Submersible Pump Symposium. OnePetro.

Lea, 2007. Artificial Lift Selection. In Petroleum Engineering Handbook, Vol. 4: Production Operations Engineering, ed. J. D. Clegg. Richardson, Texas, USA: Society of Petroleum Engineers.

Lea, J.F. and Nickens, H.V., 1999, March. Selection of artificial lift. In SPE Mid-Continent Operations Symposium. OnePetro.

Lipton, Z.C., Elkan, C. and Narayanaswamy, B., 2014. Thresholding classifiers to maximize F1 score. arXiv preprint arXiv:1402.1892.

Liu, B., Zhou, S. and Zhang, S., 2010, June. The Application of Shallow Horizontal Wells in Sudan. In International Oil and Gas Conference and Exhibition in China.

Liu, F. and Patel, A., 2013, March. Well failure detection for rod pump artificial lift system through pattern recognition. In International Petroleum Technology Conference. OnePetro.

Liu, Y., Yao, K., Liu, S., Raghavendra, C.S., Lenz, T.L., Olabinjo, L., Seren, B., Seddighrad, S. and Dinesh Babu, C.G., 2010, May. Failure prediction for rod pump artificial lift systems. In SPE Western Regional Meeting. OnePetro.

Liu, Y., Yao, K.T., Raghavenda, C.S., Wu, A., Guo, D., Zheng, J., Olabinjo, L., Balogun, O. and Ershaghi, I., 2013, April. Global model for failure prediction for rod pump artificial lift systems. In SPE Western Regional & AAPG Pacific Section Meeting 2013 Joint Technical Conference. OnePetro.

Liu, Z. and Zerpa, L.E., 2016, May. Preliminary study of liquid loading problems for gas hydrate wells and selection of artificial lift methods. In SPE Western Regional Meeting. OnePetro.

Luo, G., Tian, Y., Bychina, M. and Ehlig-Economides, C., 2018, September. Production optimization using machine learning in Bakken shale. In Unconventional Resources Technology Conference, Houston, Texas, 23-25 July 2018 (pp. 2174-2197). Society of Exploration Geophysicists, American Association of Petroleum Geologists, Society of Petroleum Engineers.

Mahdi, M.A.A., Amish, M. and Oluyemi, G., 2023. An Artificial Lift Selection Approach Using Machine Learning: A Case Study in Sudan. Energies, 16(6), p.2853.

Mali, P. and Al-Jasmi, A., 2014, June. Evaluation of artificial lift modes for heavy oil reservoirs. In SPE Heavy Oil Conference-Canada. OnePetro.

Matondang, A.N., Ibnu, A. and Subiantoro, E., 2011, June. Application of hybrid artificial lift to produce multizone with high GOR, contrast PI and contrast water cut. In Brasil Offshore. OnePetro.

Mathur, A. and Foody, G.M., 2008. Multiclass and binary SVM classification: Implications for training and classification users. IEEE Geoscience and remote sensing letters, 5(2), pp.241-245.

Matthews, C.M., Zahacy, T.A., Alhanati, F.J.S., Skoczylas, P., Technologies, C., Dunn, L.J., 2007. Progressing Cavity Pumping Systems. In Petroleum Engineering Handbook, Vol. 4: Production Operations Engineering, ed. J. D. Clegg. Richardson, Texas, USA: Society of Petroleum Engineers.

Mesbah, H., Dange, A. and Tanasescu, I., 2018, November. Hybrid Artificial Lift System-Alternate SRP/PCP in Cyclic Steam Injection for Heavy Oil Wells. In SPE Middle East Artificial Lift Conference and Exhibition. OnePetro.

Mohamed, A.E., 2017. Comparative study of four supervised machine learning techniques for classification. International Journal of Applied, 7(2), pp.1-15.

Naderi, A., Ghayyem, M.A. and Ashrafi, M., 2014. Artificial Lift Selection in the Khesht Field. Petroleum science and technology, 32(15), pp.1791-1799

Naguib, M.A., El Battrawy, A., Bayoumi, A. and El-Emam, N., 2000, October. Guideline of Artificial Lift Selection for Mature Field. In SPE Asia Pacific Oil and Gas Conference and Exhibition. OnePetro.

Neely, B., Gipson, F., Clegg, J., Capps, B. and Wilson, P., 1981, October. Selection of artificial lift method. In SPE Annual Technical Conference and Exhibition. OnePetro.

Nguyen, T., 2020. Artificial lift methods: design, practices, and applications. Springer Nature.

Noonan, S.G., 2008, October. The Progressing Cavity Pump Operating Envelope: You Cannot Expand What You Don't Understand. In International Thermal Operations and Heavy Oil Symposium.

Noonan, S., 2010. Technology Focus: Artificial Lift (July 2010). Journal of Petroleum Technology, 62(07), pp.50-50.

Noshi, C.I. and Schubert, J.J., 2018, October. The role of machine learning in drilling operations; a review. In SPE/AAPG Eastern regional meeting. OnePetro.

Onwuchekwa, C., 2018, August. Application of machine learning ideas to reservoir fluid properties estimation. In SPE Nigeria Annual International Conference and Exhibition. OnePetro.

Ounsakul, T., Rittirong, A., Kreethapon, T., Toempromraj, W., Wejwittayaklung, K. and Rangsriwong, P., 2020, October. Data-driven diagnosis for artificial lift pump's failures. In SPE/IATMI Asia Pacific Oil & Gas Conference and Exhibition. OnePetro.

Ounsakul, T., Sirirattanachatchawan, T., Pattarachupong, W., Yokrat, Y. and Ekkawong, P., 2019, March. Artificial lift selection using machine learning. In International petroleum technology conference. OnePetro.

Oyewole, P., 2016, October. Artificial lift selection strategy to maximize unconventional oil and gas assets value. In SPE North America Artificial Lift Conference and Exhibition. OnePetro.

Pandey, Y.N., Rastogi, A., Kainkaryam, S., Bhattacharya, S., Saputelli, L., Pandey, Y.N., Rastogi, A., Kainkaryam, S., Bhattacharya, S. and Saputelli, L., 2020. Toward Oil and Gas 4.0. Machine Learning in the Oil and Gas Industry: Including Geosciences, Reservoir Engineering, and Production Engineering with Python, pp.1-40.

Pankaj, P., Geetan, S., MacDonald, R., Shukla, P., Sharma, A., Menasria, S., Xue, H. and Judd, T., 2018, April. Application of Data Science and machine Learning for well completion Optimization. In Offshore Technology Conference. OnePetro.

Pankaj, P., Patron, K.E. and Lu, H., 2018, August. Wellbore Modeling and Reservoir Characterization for the Application of Artificial Lift in Deep Horizontal Wells in the Unconventional Reservoirs. In SPE Artificial Lift Conference and Exhibition-Americas. OnePetro.

Parshall, J., 2013. Challenges, Opportunities Abound for Artificial Lift. Journal of Petroleum Technology, 65(03), pp.70-75.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. and Vanderplas, J., 2011. Scikit-learn: Machine learning in Python. the Journal of machine Learning research, 12, pp.2825-2830.

Pennel, M., Hsiung, J. and Putcha, V.B., 2018, April. Detecting failures and optimizing performance in artificial lift using machine learning models. In SPE Western Regional Meeting. OnePetro.

Pham-Gia, T. and Choulakian, V., 2014. Distribution of the sample correlation matrix and applications. Open Journal of Statistics, 2014.

Phelps, R., 2015. Balancing Artificial Lift, Chemical Injection To Achieve Production Goals Over Life of Well. Journal of Petroleum Technology, 67(10), pp.28-30.

Pollock, J., Stoecker-Sylvia, Z., Veedu, V., Panchal, N. and Elshahawi, H., 2018, April. Machine learning for improved directional drilling. In Offshore Technology Conference. OnePetro.

Potdar, K., Pardawala, T.S. and Pai, C.D., 2017. A comparative study of categorical variable encoding techniques for neural network classifiers. International journal of computer applications, 175(4), pp.7-9.

Prosper, C. and West, D., 2018, October. Case study applied machine learning to optimise PCP completion design in a CBM field. In SPE Asia Pacific Oil and Gas Conference and Exhibition. OnePetro.

Ramde, S., Beauquin, J., Bellett, D., Duret-Thual, C., Olivier, G. and Pierchon, A., 2014, October. Innovative solutions in pcp technologies, run life improvement with experimental and numerical works. In SPE Artificial Lift Conference & Exhibition-North America. OnePetro.

Ramirez, A.M., Valle, G.A., Romero, F. and Jaimes, M., 2017, May. Prediction of PVT properties in crude oil using machine learning techniques MLT. In SPE Latin America and Caribbean Petroleum Engineering Conference. OnePetro.

Ranjan, A., Verma, S. and Singh, Y., 2015, March. Gas lift optimization using artificial neural network. In SPE Middle East Oil & Gas Show and Conference. OnePetro.

Rubiano, E., Martin, J.L., Prada, J., Monroy, M., Labrador, L., Celis, J., Gutierrez, J. and Bohorquez, M., 2015, May. Run Life Improvement by Implementation of Artificial Lift Systems Failure Classification and Root Cause Failure Classification. In SPE Artificial Lift Conference—Latin America and Caribbean. OnePetro.

Saghir, F., Gilabert, H. and Mancuso, B.M., 2020, October. Application of Augmented Intelligence and Edge Analytics In Upstream Production Operations: An Innovative Approach for Optimizing Artificial Lift Systems Performance. In SPE Annual Technical Conference and Exhibition. OnePetro.

Sahu, C., Kumar, R. and Sangwai, J.S., 2021. A comprehensive review on well completion operations and artificial lift techniques for methane gas production from natural gas hydrate reservoirs. Energy & Fuels, 35(15), pp.11740-11760.

Scarsdale, K., de Pieri Pereira, D., Garber, M., Goh, K.H., Kee, B., Gastaud, N., Simon, A., Joseph, J. and Cook, W., 2019, May. Alternatively Deployed Artificial Lift System for Deepwater Subsea Operations. In SPE Gulf Coast Section Electric Submersible Pumps Symposium. OnePetro.

Shedid, S.A., 2009, April. Effects of subsurface pump size and setting depth on performance of sucker-rod artificial lift: A simulation approach. In SPE Production and Operations Symposium. OnePetro.

Shi, J., Chen, S., Zhang, X., Zhao, R., Liu, Z., Liu, M., Zhang, N. and Sun, D., 2019, March. Artificial lift methods optimising and selecting based on big data analysis technology. In International Petroleum Technology Conference.

Shoeibi Omrani, P., Dobrovolschi, I., Belfroid, S., Kronberger, P. and Munoz, E., 2018, November. Improving the accuracy of virtual flow metering and back-allocation through machine learning. In Abu Dhabi International Petroleum Exhibition & Conference. OnePetro.

Skoczylas, P., Alhanati, F., Sheldon, J. and Trevisan, F., 2018, August. Use of run-life measures in estimating artificial lift system reliability. In SPE Artificial Lift Conference and Exhibition-Americas. OnePetro.

Sneed, J., 2017, September. Predicting ESP lifespan with machine learning. In Unconventional Resources Technology Conference, Austin, Texas, 24-26 July 2017

(pp. 863-869). Society of Exploration Geophysicists, American Association of Petroleum Geologists, Society of Petroleum Engineers.

Stephenson, G., 2019. Technology Focus: Artificial Lift (March 2019). Journal of Petroleum Technology, 71(03), pp.64-64.

Stephenson, G., 2020. Technology Focus: Artificial Lift (March 2020). Journal of Petroleum Technology, 72(03), pp.47-47.

Syed, F.I., Alshamsi, M., Dahaghi, A.K. and Neghabhan, S., 2022. Artificial lift system optimization using machine learning applications. Petroleum, 8(2), pp.219-226.

Takacs, G., 2015. Sucker-rod pumping handbook: production engineering fundamentals and long-stroke rod pumping. Gulf Professional Publishing.

Tan, P.N., Steinbach, M. and Kumar, V., 2006. Classification: basic concepts, decision trees, and model evaluation. Introduction to data mining, 1, pp.145-205.

Taunk, K., De, S., Verma, S. and Swetapadma, A., 2019, May. A brief review of nearest neighbor algorithm for learning and classification. In 2019 International Conference on Intelligent Computing and Control Systems (ICCS) (pp. 1255-1260). IEEE.

Teknomo, K., 2006. K-means clustering tutorial. Medicine, 100(4), p.3.

Temizel, C., Canbaz, C.H., Betancourt, D., Ozesen, A., Acar, C., Krishna, S. and Saputelli, L., 2020, October. A comprehensive review and optimization of artificial lift methods in unconventionals. In SPE Annual Technical Conference and Exhibition. OnePetro.

Valbuena, J., Pereyra, E. and Sarica, C., 2016, October. Defining the artificial lift system selection guidelines for horizontal wells. In SPE North America Artificial Lift Conference and Exhibition. OnePetro.

Valentin, E.P. and Hoffmann, F.C., 1988, October. OPUS: An Expert Advisor for Artificial Lift. In SPE Annual Technical Conference and Exhibition. OnePetro.

Williams, S., Rozo, R.E., Perez Aya, F. and Salazar Hernandez, J.I., 2008, September. Artificial-Lift Optimisation In The Orito Field. In SPE Annual Technical Conference and Exhibition. OnePetro.

Winkler, H. W. and Blann, J. R. 2007. Gas Lift. In Petroleum Engineering Handbook, Vol. 4: Production Operations Engineering, ed. J. D. Clegg. Richardson, Texas, USA: Society of Petroleum Engineers.

Yang, Y., Zhou, W., Shi, G., Gang, C., Wang, G., Sun, C., Qiang, L., Zhao, Y., Zhao, C., Bai, W. and Wu, M., 2011, July. 17 Years Development of Artificial Lift Technology in ASP Flooding in Daqing Oilfield. In SPE Enhanced Oil Recovery Conference. OnePetro.

Zein El Din Shoukry, A., Soltys, T.W., Bettenson, J. and Ariza, G., 2020, January. First Successful Installation of Progressing Cavity Pump System in an Oil Well at the Kingdom of Saudi Arabia. In International Petroleum Technology Conference. OnePetro.

Zhongxian, H., Gang, C., Lianyu, W., Mingzhan, C., Yuan, F. and He, L., 2015, August. Problem and solution: Artificial lift technology in polymer flooding. In SPE Asia Pacific Enhanced Oil Recovery Conference. OnePetro.

Zulkapli, M.H., Salim, M.M., Zaini, M.Z., Rivero Colmenares, M.E., Curteis, C. and Sepulveda, W., 2014, December. The Evolution of Artificial Lift Completions in an Offshore Brownfield in Malaysia. In International Petroleum Technology Conference. OnePetro.

APPENDICES

Appendix A1 Python Code for AL selection model

AL and size selection code, this is the production data code, the same code was used in all models with different features (operation, environmental, and economic).

```
pip install opendatasets --upgrade
```

```
! pip install plotly --upgrade
```

```
import os

import opendatasets as od
import matplotlib

import matplotlib.pyplot as plt

%matplotlib inline

import plotly.express as px
import seaborn as sns
import pandas as pd

import numpy as np
import warnings

from sklearn.metrics import accuracy_score, confusion_matrix

from sklearn.metrics import f1_score

from sklearn.metrics import recall_score

from sklearn.metrics import precision_score
```

```
raw_df = pd.read_csv(r'C:\Users\moh-m\Desktop\dataset\Last 1st model data files\02. Cleaned data for training and validation\Training and Validation Data.csv')
```

```
raw_df
```

```
raw_df.isna().sum()
```

```
raw_df.columns
```

Exploratory Data Analysis

```
data_df = pd.read_csv(r'C:\Users\moh-m\Desktop\dataset\Last 1st model data files\02. Cleaned data for training and validation\Training and Validation Data.csv')
```

```
data_df['AL'].value_counts()
```

```
data_df.describe()
```

```
px.violin(data_df, y='WELLHEAD_PRESS', x='AL', color='IOR_EOR', title='AL vs IOR & EOR', log_y=True)
```

```
fig = px.histogram(
                    data_df,
                    x="IOR_EOR",
                    y="OIL",
                    log_y=True, #log_x=True,
                    color='AL',
                    title='IOR and EOR vs Oil (STB)')
fig.show()
```

```
fig = px.histogram(
                    data_df,
                    x="AL",
                    y="OIL",
                    log_y=True, #log_x=True,
                    color='IOR_EOR', title='AL vs Oil (STB)')
fig.show()
```

```
fig = px.histogram(
                    data_df,
                    x="AL",
                    y="Total_Fluid",
                    log_y=True, #log_x=True,
                    color='IOR_EOR',
                    title='AL vs Cumulative Produced Fluid (STB)'
   )
fig.show()
```

```
# let's see the distribution of IOR and EOR in the specific field with regards to the selected wells
recovery_method = data_df['IOR_EOR'].value_counts()
sns.set(style="darkgrid")
sns.barplot(recovery_method.index, recovery_method.values, alpha=0.9)
plt.title('Frequency Distribution of Secondary and Tertiary Recovery')
plt.ylabel('Number of Occurrence', fontsize=12)
plt.xlabel('IOR_EOR', fontsize=12)
plt.show()
```

```
fig = px.histogram(
                    data_df,
                    x="IOR_EOR",
                    y="OIL",
                    log_y=True,
                    #log_x=True,
                    color='Type_Size',
                    title='Oil vs AL Type and Size' )
fig.show()
```

```
fig = px.histogram(
                    data_df,
                    x="AL",
                    y="OIL",
                    log_y=True, #log_x=True,
                    color='Type_Size',
                    title='Oil poduced by AL Type and Size' )
fig.show()
```

```
fig = px.histogram(
                    data_df,
                    x="AL",
                    y="SAND",
                    log_y=True,
                    #log_x=True, color='Type_Size',
                  title='Sand Production by AL' )
fig.show()
```

```
fig=plt.figure(figsize=(20,8))
ax1=fig.add_subplot(221)

ax1.set(xlabel='IOR and EOR', ylabel='AL', title='AL vs EOR')
```

```
sns.barplot(y=data_df['AL'].index,   x=data_df['IOR_EOR'],   ax=ax1);
```

```
fig=plt.figure(figsize=(20,8))
ax1=fig.add_subplot(221)

ax1.set(xlabel='AL', ylabel='OIL (BBL)', title='AL vs OIL') sns.barplot(y=data_df['OIL'].index,
x=data_df['AL'],   ax=ax1);
```

```
fig=plt.figure(figsize=(20,8))
ax1=fig.add_subplot(221)

ax1.set(xlabel='AL', ylabel='OIL (BBL)', title='AL vs Produced Fluid')
sns.barplot(y=data_df['Total_Fluid'].index, x=data_df['AL'], ax=ax1);
```

```
# let's see the distribution of AL in the specific field with regards to the
selected wells

artificial_lift = data_df['AL'].value_counts() sns.set(style="darkgrid")

sns.barplot(artificial_lift.index, artificial_lift.values, alpha=0.9) plt.title('Frequency
Distribution of AL')

plt.ylabel('Number of Occurrence', fontsize=12)

plt.xlabel('AL', fontsize=12)

plt.show()
```

```
# let's see the distribution of IOR and EOR in the specific field with regards to
the selected wells



recovery_method = data_df['Type_Size'].value_counts()
sns.set(style="darkgrid")

sns.barplot(recovery_method.index, recovery_method.values, alpha=0.9)
plt.title('Frequency Distribution of AL Type and Size') plt.ylabel('Number of
Occurrence', fontsize=12) plt.xlabel('Type_Size', fontsize=12)

plt.xticks(rotation=80)

plt.show()
```

```
fig = px.histogram(data_df,

                   x='AL',

                   marginal='box', nbins=47,

                   title='Distribution of AL')
fig.update_layout(bargap=0.1)

fig.show()
```

```
fig = px.histogram(data_df,

                   x='Type_Size',
                   marginal='box',

                   color_discrete_sequence=['red'],
```

```
title='Distribution of AL Type and Size') fig.update_layout(bargap=0.1)

fig.show()
```

```
fig = px.histogram(data_df.sample(500),
                           x='AL',
                           marginal='box',
                           color='Total_Fluid',
                            color_discrete_sequence=['green', 'blue', 'pink'],
                           title='Distribution of AL coloured with the Total Fluid')
fig.update_layout(bargap=0.1)
```

```
px.scatter(data_df,
                title='Wells vs. Oil produced',
                x='ALIAS',
                y='OIL',
                log_y=True,
#                 log_x=True,
                color='AL')
```

```
px.scatter(data_df.sample(2000),
                title='WELLHEAD_PRESS vs. Total_Fluid', x='WELLHEAD_PRESS',
                y='Total_Fluid',
                log_y=True,
#                 log_x=True,
                color='AL')
```

```
px.scatter(data_df,
                title='WELLHEAD_PRESS vs. GOR',
                x='WELLHEAD_PRESS',
                y='GOR',
                log_y=True,
                log_x=True,
                color='AL')
```

```
px.scatter(data_df.sample(2000),
                title='Total_Fluid vs. WC%',
                x='Total_Fluid',
                y='WC%',
                #log_y=True,
                log_x=True,
                color='AL')
```

```
px.scatter(raw_df,
                title='OIL vs. SAND', x='OIL',
                y='SAND',
                log_y=True,
                log_x=True,
                color='Type_Size')
```

```
px.scatter(data_df,
                title='OIL vs. Water', x='Water',
                y='OIL',
                log_y=True,
                log_x=True,
                color='AL')
```

```
px.scatter(data_df,
                title='OIL vs. Water', x='Water',
                y='OIL',
                log_y=True,
                log_x=True,
                color='Type_Size')
```

```
px.scatter(data_df,
                title='OIL vs. WELLHEAD_PRESS',
                x='WELLHEAD_PRESS',
                y='OIL',
                log_y=True,
                log_x=True,
                color='AL')
```

```
px.box(data_df,
                title='AL vs. IOR_EOR', x='AL',
                y='IOR_EOR')
```

```
px.scatter(data_df,
                title='AL vs. IOR_EOR',
                x='IOR_EOR',
                y='AL',
 #                log_y=True,
#                 log_x=True,
                color='Type_Size')
```

```
px.scatter(data_df,
                title='AL vs. WELLHEAD_PRESS',
                x='AL', y='WELLHEAD_PRESS',
                log_y=True,
                # log_x=True,
                color='Type_Size')
```

Data Preprocessing - Encoding

```
from sklearn.preprocessing import OneHotEncoder
```

```
cat_cols =      ['IOR_EOR']
encoder = OneHotEncoder(sparse=False, handle_unknown='ignore').fit(raw_df[cat_cols])
```

```
encoded_cols  =  list(encoder.get_feature_names(cat_cols))
```

```
raw_df[encoded_cols] = encoder.transform(raw_df[cat_cols])
```

```
raw_df[encoded_cols]
```

```
raw_df
```

```
raw_df = raw_df.drop(columns = ['IOR_EOR']).copy()
```

```
raw_df
```

```
year = pd.to_datetime(raw_df['Date']).dt.year
month =pd.to_datetime(raw_df['Date']).dt.month
day  =pd.to_datetime(raw_df['Date']).dt.day
```

```
raw_df['Year'] =pd.DataFrame(year)
raw_df['Month'] =pd.DataFrame(month)
raw_df['Day'] =pd.DataFrame(day)
```

```
order_ls = raw_df.columns.tolist()
```

```
order_ls
```

```
raw_df = raw_df.reindex(['ALIAS',
                         'Year',
                         'Month',
                         'Day',
                         'Date',
'RUN_PERIOD',
                         'WELLHEAD_PRESS',
                         'Total_Fluid',
'GOR',
                         'OIL',
                         'GAS',
                         'Water',
                         'SAND',
                         'WC%',
'IOR_EOR_CSS',
                         'IOR_EOR_Gas_injec', 'IOR_EOR_N2',
                         'IOR_EOR_None',
'IOR_EOR_SF',
'IOR_EOR_WI',
'AL',
                         'Type_Size'], axis=1)
```

```
raw_df
```

```
raw_df = raw_df.drop(columns=['Date'])
```

```
raw_df
```

Scaling

```
num_cols = ['Year',
            'Month',
            'Day', 'RUN_PERIOD',
'WELLHEAD_PRESS',
             'Total_Fluid',
             'GOR',
             'OIL',
             'GAS',
             'Water',
             'SAND',
             'WC%']
```

```
num_cols
```

```
raw_df[num_cols]
```

```
from sklearn.preprocessing import MinMaxScaler
```

```
scaler = MinMaxScaler().fit(raw_df[num_cols])
```

```
raw_df[num_cols] = scaler.transform(raw_df[num_cols])
```

```
raw_df.describe()
```

Training & Validation Data

```
raw_df['ALIAS'].value_counts()
```

```
val_wells_ls = ['XK7',
                'XFE36',
                'XM184',
                'XH1']
```

```
val_wells_ls
```

```
raw_df['ALIAS'].isin(val_wells_ls)
```

```
~raw_df.ALIAS.isin(val_wells_ls)
```

```
train_df = raw_df[~raw_df.ALIAS.isin(val_wells_ls)]
```

```
train_df
```

```
val_df = raw_df[raw_df.ALIAS.isin(val_wells_ls)]
```

```
val_df['ALIAS'].unique()
```

```
val_df['ALIAS'].value_counts()
```

Input vs Targets Split

```
[ ]:
```

```
targets_train = train_df[['AL']].copy()
```

```
targets_train
```

```
targets_val = val_df[['AL']].copy()
```

```
targets_val
```

[ ]:

```
inputs_train = train_df.drop(columns = ['ALIAS', 'AL', 'Type_Size'])
inputs_train
```

Validation

```
val_df
```

```
inputs_val = val_df.drop(columns = ['ALIAS', 'AL', 'Type_Size']).copy()
```

```
inputs_val
```

```
inputs_train.shape,  targets_train.shape
```

```
inputs_val.shape,  targets_val.shape
```

Inputs and Targets Correlation

```
data_AL_df = data_df.copy()
```

```
data_AL_df
```

```
categorical_AL_cols    =    data_df[['IOR_EOR','AL']].columns.tolist()
```

```
encoder_AL        =        OneHotEncoder(sparse=False).fit(data_AL_df[categorical_AL_cols])
```

```
encoder_AL_cols    =    encoder_AL.get_feature_names(categorical_AL_cols).tolist()
```

```
data_AL_df[encoder_AL_cols] = encoder_AL.transform(data_AL_df[categorical_AL_cols])
```

```
data_AL_df.corr()
```

```
plt.figure(figsize=(50,  26))
sns.heatmap(data_AL_df.corr(), cmap='Reds',square=True, annot=True,annot_kws={'size':20})
plt.title('AL Correlation Matrix',fontsize=50); sns.set(font_scale=4)
```

```
px.bar(data_AL_df.corr())
```

Base Model

Random Guess

```
targets_train.value_counts()
```

```
targets_cat_AL = targets_train.AL.unique().tolist()
```

```
targets_cat_AL
```

```
targets_cat_AL
```

```
def random_guess(inputs_train):
    return np.random.choice(targets_cat_AL, len(inputs_train))
```

```
accuracy_score(targets_val['AL'], random_guess(inputs_val))
```

Random Guess from 2

```
targets_cat_AL2 = ['PCP','GL']
```

```
def random_guess2(inputs_train):
    return np.random.choice(targets_cat_AL2, len(inputs_train))
```

```
accuracy_score(targets_val, random_guess2(inputs_val))
```

All PCP

```
def all_PCP(inputs_train):
    return np.full(len(inputs_train), "PCP")
```

```
accuracy_score(targets_val, all_PCP(inputs_val))
```

Modelling

Logreg model

```
from sklearn.linear_model import LogisticRegression
```

```
model_logreg_AL = LogisticRegression(solver='liblinear')
```

```
model_logreg_AL.fit(inputs_train, targets_train)
```

```
model_logreg_AL.classes_
```

```
train_preds_AL = model_logreg_AL.predict(inputs_train)
```

```
train_preds_AL
```

```
prob = model_logreg_AL.predict_proba(inputs_train)[0]
```

[ ]:

```
prob
```

```
accuracy_score(targets_train, train_preds_AL)
```

```
val_preds_AL = model_logreg_AL.predict(inputs_val)
```

```
val_preds_AL
```

```
accuracy_score(targets_val, val_preds_AL)
```

```
from sklearn.multioutput import MultiOutputClassifier
```

```
model_logreg = LogisticRegression(solver='liblinear')
```

```
model_logreg.fit(inputs_train, targets_train)
```

```
model_logreg.score(inputs_train, targets_train)
```

```
model_logreg.score(inputs_val, targets_val)
```

```
train_preds_logreg = model_logreg.predict(inputs_train)
```

```
accuracy_score(targets_train, train_preds_logreg)
```

```
val_preds_logreg = model_logreg.predict(inputs_val)
```

```
accuracy_score(targets_val, val_preds_logreg)
```

```
f1_score(train_preds_logreg, targets_train, average='macro')
```

```
f1_score(val_preds_logreg, targets_val, average='macro')
```

Decision Tree Classifier

```
from sklearn.tree import DecisionTreeClassifier
```

```
model_tree = DecisionTreeClassifier(random_state=42)
```

```
%%time

model_tree.fit(inputs_train, targets_train)
```

Training Data

```
train_preds_tree = model_tree.predict(inputs_train)
```

```
train_preds_tree
```

```
accuracy_score(train_preds_tree, targets_train)
```

```
accuracy_score(train_preds_tree, targets_train)
```

```
train_props = model_tree.predict_proba(inputs_train)
```

```
train_props
```

```
from sklearn.metrics import jaccard_score
```

```
jaccard_score(train_preds_tree, targets_train, average='micro')
```

```
f1_score(train_preds_tree, targets_train, average='micro')
```

Validation Data

```
val_preds_tree = model_tree.predict(inputs_val)
```

```
val_preds_tree
```

```
accuracy_score(val_preds_tree, targets_val)
```

```
jaccard_score(val_preds_tree, targets_val, average='micro')
```

```
f1_score(val_preds_tree, targets_val, average='micro')
```

Decision Tree Model-2

```
model_tree_2 = DecisionTreeClassifier(max_depth=8, random_state=42)
```

```
%%time

model_tree_2.fit(inputs_train,  targets_train)
```

```
train_preds_2 = model_tree_2.predict(inputs_train)
```

```
train_preds_2
```

```
accuracy_score(train_preds_2,  targets_train)
```

```
accuracy_score(train_preds_2,  targets_train)
```

```
val_preds_2 = model_tree_2.predict(inputs_val)
```

```
val_preds_2
```

```
accuracy_score(val_preds_2,  targets_val)
```

```
accuracy_score(val_preds_2,  targets_val)
```

```
jaccard_score(val_preds_2,  targets_val,  average='macro')
```

```
f1_score(val_preds_2,  targets_val,  average='macro')
```

```
model_tree_2.score(inputs_val,  targets_val)
```

Random Forest Classifier

```
from sklearn.ensemble import RandomForestClassifier
```

```
model_randomforest = RandomForestClassifier(n_jobs=-1,
                                            random_state=42,
                                            n_estimators=500,
                                            max_features=7,
                                            max_depth=7)
```

```
%%time

model_randomforest.fit(inputs_train, targets_train)
```

```
train_preds_ranfor = model_randomforest.predict(inputs_train)
train_preds_ranfor
```

```
accuracy_score(train_preds_ranfor, targets_train)
```

```
val_preds_ranfor = model_randomforest.predict(inputs_val)
```

```
val_preds_ranfor
```

```
accuracy_score(val_preds_ranfor, targets_val)
```

```
model_randomforest.score(inputs_train, targets_train)
```

```
model_randomforest.score(inputs_val, targets_val)
```

```
f1_score(val_preds_ranfor, targets_val, average='macro')
```

Support Vector Machines

```
from sklearn.multioutput import MultiOutputClassifier

from sklearn import svm
```

```
svm.SVC(decision_function_shape='ovo')
```

```
model_svm = svm.SVC(decision_function_shape='ovo')
```

```
model_svm
```

```
model_svm.fit(inputs_train, targets_train)
```

```
train_preds_svm = model_svm.predict(inputs_train)
```

```
train_preds_svm
```

```
val_preds_svm = model_svm.predict(inputs_val)
```

```
model_svm.score(inputs_train, targets_train)
```

```
model_svm.score(inputs_val, targets_val)
```

```
accuracy_score(val_preds_svm, targets_val)
```

K Nearest Neighbors Classifier

```
from sklearn.neighbors import KNeighborsClassifier
```

```
model_knn = KNeighborsClassifier()
```

```
model_knn.fit(inputs_train, targets_train)
```

```
train_preds_knn = model_knn.predict(inputs_train)
```

```
val_preds_knn = model_knn.predict(inputs_val)
```

```
model_knn.score(inputs_train, targets_train)
```

```
model_knn.score(inputs_val, targets_val)
```

```
accuracy_score(train_preds_knn, targets_train)
```

```
accuracy_score(val_preds_knn, targets_val)
```

Confusion Matrix

```
from sklearn.metrics import confusion_matrix, ConfusionMatrixDisplay
```

```
cm = confusion_matrix(targets_val, val_preds_ranfor, labels=model_randomforest.classes_)
```

```
disp = ConfusionMatrixDisplay(confusion_matrix=cm, display_labels=model_randomforest.classes_)

disp.plot(cmap='YlGn');sns.set(font_scale=4);plt.xticks(rotation='45')
```

```
cm = confusion_matrix(targets_val, val_preds_2, labels=model_tree_2.classes_)
disp = ConfusionMatrixDisplay(confusion_matrix=cm, display_labels=model_tree_2.classes_)

disp.plot(cmap='YlGn');
```

Random Forest

```
def max_depth_error_ranfor(md):

        model = RandomForestClassifier(n_jobs=-1,
                                       random_state=42,
                                       n_estimators=700,
                                       max_features=7,
                                       max_depth=md)

        model.fit(inputs_train, targets_train)
        train_acc = 1 - model.score(inputs_train, targets_train)
        val_acc = 1 - model.score(inputs_val, targets_val)
        return {'Max Depth': md, 'Training Error': train_acc, 'Validation Error': val_acc}
```

```
%%time

errors_ranfor_df = pd.DataFrame([max_depth_error_ranfor(md) for md in range (1,21)])
```

```
plt.figure()

plt.plot(errors_ranfor_df['Max Depth'], errors_ranfor_df['Training Error'])
plt.plot(errors_ranfor_df['Max Depth'], errors_ranfor_df['Validation Error'])

plt.title('Training vs. Validation Error')

plt.xticks(range(0,21, 2))
plt.xlabel('Max. Depth')

plt.ylabel('Prediction Error (1 - Accuracy)')
plt.legend(['Training', 'Validation']);
```

Decision Tree Classifier

```
model_tree = DecisionTreeClassifier(max_depth=8, random_state=42)
```

```
model_tree.fit(inputs_train, targets_train)
```

```
train_preds = model_tree.predict(inputs_train)
```

```
train_preds
```

```
model_tree.score(inputs_train, targets_train)
```

```
val_preds = model_tree.predict(inputs_val)
```

```
val_preds
```

```
model_tree.score(inputs_val, targets_val)
```

```
def max_depth_error(md):
    model = DecisionTreeClassifier(max_depth=md, random_state=42)
    model.fit(inputs_train, targets_train)
    train_acc = 1 - model.score(inputs_train, targets_train)
    val_acc = 1 - model.score(inputs_val, targets_val)
    return {'Max Depth': md, 'Training Error': train_acc, 'Validation Error':val_acc}
```

```
max_depth_error(3)
```

```
max_depth_error(7)
```

```
max_depth_error(9)
```

```
max_depth_error(10)
```

[ ]:

```
%%time

errors_df = pd.DataFrame([max_depth_error(md) for md in range(1, 21)])
```

```
errors_df
```

```
plt.figure()
plt.plot(errors_df['Max Depth'], errors_df['Training Error'])
plt.plot(errors_df['Max Depth'], errors_df['Validation Error'])
plt.title('Training vs. Validation Error') plt.xticks(range(0,21, 2))
plt.xlabel('Max. Depth')
plt.ylabel('Prediction Error (1 - Accuracy)')
plt.legend(['Training', 'Validation']);
```

```
model_tree.tree_.max_depth
```

```
from sklearn.tree import plot_tree, export_text
```

```
model_text = export_text(model_tree, feature_names=list(inputs_train.columns))
print(model_text[:3000])
```

```
plt.figure(figsize=(80,20))

plot_tree(model_tree, feature_names=inputs_train.columns, max_depth=2, filled=True);
```

Features Importance

```
importance_df = pd.DataFrame({
    'Feature':inputs_train.columns,

    'Importance': model_tree.feature_importances_

}).sort_values('Importance', ascending=False)
```

```
importance_df.head(10)
```

```
plt.title('Feature Importance') sns.barplot(data=importance_df.head(10),
x='Importance', y='Feature');
```

```
importance_df = pd.DataFrame({
    'feature':inputs_train.columns,

    'importance': model_randomforest.feature_importances_

}).sort_values('importance', ascending=False)
```

```
importance_df.head(10)
```

```
plt.title('Feature Importance') sns.barplot(data=importance_df.head(10),
x='importance', y='feature');
```

Algorithms Performance

```
#tree_score     =     model_tree.score(inputs_val,     targets_val)
tree_score_2 = model_tree_2.score(inputs_val, targets_val)

knn_score = model_knn.score(inputs_val, targets_val)

svm_score = model_svm.score(inputs_val, targets_val)

rf_score = model_randomforest.score(inputs_val, targets_val)
logreg_score = model_logreg.score(inputs_val, targets_val)
```

```
results = pd.DataFrame({
    'Model':[#'Tree',

             'Tree_2',

            'Knn',


            'SVM',

             'RanFor',

             'Logreg',
             #'XGboost',
             #'Catboost'

            ],

    'Score':[#tree_score,

             tree_score_2,
            knn_score,
            svm_score,
            rf_score,
            logreg_score,
            #XGboost_score,
            #catboost_score

            ]})

sorted_result = results.sort_values(by='Score', ascending=False).reset_index(drop=True)

sorted_result
```

```
f, ax = plt.subplots(figsize=(14,8))
plt.xticks(rotation='90')

sns.barplot(x=sorted_result['Model'],     y=sorted_result['Score'])
plt.xlabel('Model', fontsize=15)

plt.ylabel('Performance', fontsize=15)
#plt.ylim(0.10,   0.12)

plt.title('Score', fontsize=15)

plt.show()
```

Testing model on out of sample data

```
test_df = pd.read_csv(r'C:\Users\moh-m\Desktop\Type and size\Type and size original\Test  Data Labelled.csv')
```

```
test_df
```

Encoding

```
cat_cols_test  =       ['IOR_EOR']
```

```
encoder_test = OneHotEncoder(sparse=False, handle_unknown='ignore').fit(test_df[cat_cols_test])
```

```
encoded_cols_test = list(encoder_test.get_feature_names(cat_cols_test))
```

```
test_df[encoded_cols_test]  =  encoder_test.transform(test_df[cat_cols_test])
```

```
test_df[encoded_cols_test]
```

```
test_df.drop(columns=['IOR_EOR'])
```

```
year   =   pd.to_datetime(test_df['Date']).dt.year  month  =
pd.to_datetime(test_df['Date']).dt.month day   =
pd.to_datetime(test_df['Date']).dt.day
```

```
test_df['Year'] =pd.DataFrame(year)
test_df['Month']  =pd.DataFrame(month)
test_df['Day']  =pd.DataFrame(day)
```

```
order_ls = test_df.columns.tolist()
```

```
order_ls
```

[ ]:

```
test_df = test_df.reindex(['ALIAS',
                           'Date',
                           'Year',
                           'Month',
                           'Day',
                           'RUN_PERIOD',
                           'WELLHEAD_PRESS',
                           'Total_Fluid',
                           'GOR',
                           'OIL',
                           'GAS',
                           'Water',
                           'SAND',
                           'WC%',
                           'IOR_EOR',
                           'IOR_EOR_CSS',
                           'IOR_EOR_Gas_injec', 'IOR_EOR_N2',
                           'IOR_EOR_None',
                           'IOR_EOR_SF',
                           'IOR_EOR_WI',
                           'AL',
                           'Type_Size'], axis=1)
```

```
test_df = test_df.drop(columns=['Date'])
```

```
test_df.isna().sum()
```

```
num_cols_test = ['Year',
                 'Month',
                 'Day',
                 'RUN_PERIOD',
                 'WELLHEAD_PRESS',
                 'Total_Fluid', 'GOR',
                 'OIL',
                 'GAS',
                 'Water',
                 'SAND',
                 'WC%']
```

```
scaler_test = MinMaxScaler().fit(test_df[num_cols_test])
```

```
test_df[num_cols_test] = scaler_test.transform(test_df[num_cols_test])
```

```
test_df[num_cols_test]
```

```
test_targets = test_df['AL']
```

```
test_targets.shape
```

```
test_inputs = test_df.drop(columns=['ALIAS', 'AL', 'Type_Size', 'IOR_EOR']).copy()
```

```
test_inputs
```

```
test_inputs.shape
```

```
inputs_val.shape
```

Logistic Regression

```
test_preds_logreg = model_logreg.predict(test_inputs)
```

```
model_logreg.score(test_inputs, test_targets)
```

```
accuracy_score(test_preds_logreg, test_targets)
```

```
print(f1_score(test_preds_logreg, test_targets, average='macro'))
print(f1_score(test_preds_logreg, test_targets, average='micro'))
```

```
print(recall_score(test_preds_logreg, test_targets, average='macro'))
print(recall_score(test_preds_logreg, test_targets, average='micro'))
```

```
print(precision_score(test_preds_logreg, test_targets, average='macro'))
print(precision_score(test_preds_logreg, test_targets, average='micro'))
```

SVM

```
test_preds_svm = model_svm.predict(test_inputs)
```

```
model_svm.score(test_inputs, test_targets)
```

```
accuracy_score(test_preds_svm, test_targets)
```

```
print(f1_score(test_preds_svm, test_targets, average='macro')) print(f1_score(test_preds_svm,
test_targets, average='micro'))
```

```
print(recall_score(test_preds_svm, test_targets, average='macro'))
print(recall_score(test_preds_svm, test_targets, average='micro'))
```

```
print(precision_score(test_preds_svm, test_targets, average='macro'))
print(precision_score(test_preds_svm, test_targets, average='micro'))
```

KNN

```
test_preds_knn = model_knn.predict(test_inputs)
```

```
model_knn.score(test_inputs, test_targets)
```

```
accuracy_score(test_preds_knn, test_targets)
```

```
print(f1_score(test_preds_knn, test_targets, average='macro')) print(f1_score(test_preds_knn, test_targets, average='micro'))
```

```
print(recall_score(test_preds_knn, test_targets, average='macro'))
print(recall_score(test_preds_knn, test_targets, average='micro'))
```

```
print(precision_score(test_preds_knn, test_targets, average='macro'))
print(precision_score(test_preds_knn, test_targets, average='micro'))
```

Decision Tree

```
test_preds_tree2 = model_tree_2.predict(test_inputs)
```

```
test_preds_tree2[0:10].tolist()
```

```
test_targets[:10].tolist()
```

```
model_tree_2.score(test_inputs, test_targets)
```

```
accuracy_score(test_preds_tree2, test_targets)
```

```
jaccard_score(test_preds_tree2, test_targets, average='macro')
```

```
f1_score(test_preds_tree2, test_targets, average='macro')
```

```
f1_score(test_preds_tree2, test_targets, average='micro')
```

```
recall_score(test_preds_tree2, test_targets, average='macro')
```

```
recall_score(test_preds_tree2, test_targets, average='micro')
```

```
precision_score(test_preds_tree2, test_targets, average='macro')
```

```
precision_score(test_preds_tree2, test_targets, average='micro')
```

Random Forest

```
test_preds_ranfro = model_randomforest.predict(test_inputs)
```

```
test_preds_ranfro[:10]
```

```
test_preds_df = pd.DataFrame(test_preds_ranfro)
```

```
model_randomforest.score(test_inputs, test_targets)
```

```
accuracy_score(test_preds_ranfro, test_targets)
```

```
jaccard_score(test_preds_ranfro, test_targets, average='macro')
```

```
f1_score(test_preds_ranfro, test_targets, average='macro')
```

```
f1_score(test_preds_ranfro, test_targets, average='micro')
```

```
recall_score(test_preds_ranfro, test_targets, average='macro')
```

```
recall_score(test_preds_ranfro, test_targets, average='micro')
```

```
precision_score(test_preds_ranfro, test_targets, average='macro')
```

```
precision_score(test_preds_ranfro, test_targets, average='micro')
```

```
cm = confusion_matrix(test_targets, test_preds_ranfro, labels=model_randomforest.classes_)
```

```
disp = ConfusionMatrixDisplay(confusion_matrix=cm, display_labels=model_randomforest.classes_)

disp.plot(cmap='YlGn');
```

```
cm = confusion_matrix(test_targets, test_preds_tree2, labels=model_tree_2.classes_)

disp = ConfusionMatrixDisplay(confusion_matrix=cm, display_labels=model_tree_2.classes_)

disp.plot(cmap='YlGn');
```

Algorithms performance on test dataset

```
tree_score_2 = model_tree_2.score(test_inputs, test_targets)

knn_score = model_knn.score(test_inputs, test_targets)

svm_score = model_svm.score(test_inputs, test_targets)

rf_score = model_randomforest.score(test_inputs, test_targets)
logreg_score = model_logreg.score(test_inputs, test_targets)
```

```
results = pd.DataFrame({
    'Model':[#'Tree',
             'Tree_2',
            'Knn',
            'SVM',
             'RanFor',
             'Logreg',
             #'XGboost',
             #'Catboost'
            ],
    'Score':[#tree_score,
             tree_score_2,
            knn_score,
            svm_score,
            rf_score,
            logreg_score,
            #XGboost_score,
            #catboost_score
            ]})

sorted_result = results.sort_values(by='Score', ascending=False).reset_index(drop=True)

sorted_result
```

```
f, ax = plt.subplots(figsize=(14,8))
plt.xticks(rotation='90')
sns.barplot(x=sorted_result['Model'],     y=sorted_result['Score'])
plt.xlabel('Model', fontsize=15)
plt.ylabel('Performance', fontsize=15)
#plt.ylim(0.10,   0.12)
plt.title('Score', fontsize=15)
plt.show()
```

Making Predictions on New Unlabelled Inputs

```
def predict_input(model, sample_df):

    pred = list(model.predict(sample_df))

    return pred
```

```
sample_test = test_inputs[750:760] sample_test
```

```
predict_input(model_randomforest, sample_test)
```

```
test_targets[750:760].tolist()
```

```
predict_input(model_tree_2,  sample_test)
```

```
test_targets[750:760].tolist()
```

Saving and Loading Trained Model

```
import joblib
```

```
AL_model = {
    'model': model_randomforest,
    'scaler_train_val': scaler, 'scaler_test':
    scaler_test, 'encoder_train_val': encoder,
    'encoder_test': encoder_test,
    'numeric_cols': num_cols,
    'numeric_cols_test': num_cols_test,
    'categorical_cols': cat_cols,
    'categorical_cols': cat_cols_test,
    'encoded_cols': encoded_cols,
    'encoded_cols_test':  encoded_cols_test}
```

```
joblib.dump(AL_model,  'AL_model.joblib')
```

Reload the Model

```
AL_model_reload = joblib.load('AL_model.joblib')
```

```
test_preds2 = AL_model_reload['model'].predict(test_inputs) accuracy_score(test_targets,
test_preds2)
```

```
test_preds3 = AL_model_reload['model'].predict(test_inputs) accuracy_score(test_targets,
test_preds3)
```

Hyperparameters trials

```
# Forest

def max_depth_error_ranfor2(md2):

        model = RandomForestClassifier(n_jobs=-1,
                                       random_state=42,
                                       n_estimators=700,
                                       max_features=7,
                                       max_depth=md2)
        model.fit(inputs_train, targets_train)
        train_acc = 1 - model.score(inputs_train, targets_train)
        test_acc = 1 - model.score(test_inputs, test_targets)
        return {'Max Depth': md2, 'Training Error': train_acc, 'Test Error': test_acc}
```

```
max_depth_error_ranfor2(3)
```

```
max_depth_error_ranfor2(7)
```

```
max_depth_error_ranfor2(9)
```

```
max_depth_error_ranfor2(10)
```

```
%%time

errors_ranfor_df2 = pd.DataFrame([max_depth_error_ranfor2(md2) for md2 in range(1, 21)])
```

```
errors_ranfor_df2
```

```
plt.figure()

plt.plot(errors_ranfor_df2['Max Depth'], errors_ranfor_df2['Training Error'])
plt.plot(errors_ranfor_df2['Max Depth'], errors_ranfor_df2['Test Error'])

plt.title('Training vs. Test Error')

plt.xticks(range(0,21, 2))
plt.xlabel('Max. Depth')

plt.ylabel('Prediction Error (1 - Accuracy)')
plt.legend(['Training', 'Test']);

#plt.ylim((0,1))
```

```
# Forest

def max_depth_error_ranfor6(md6):
```

```
    model = RandomForestClassifier(n_jobs=-1,
                                   random_state=42,
                                   n_estimators=700,
                                   max_features=7,
                                   max_depth=md6)
    model.fit(inputs_train, targets_train)
    val_acc = 1 - model.score(inputs_val, targets_val)
    test_acc = 1 - model.score(test_inputs, test_targets)
    return {'Max Depth': md6, 'Validation Error': val_acc, 'Test Error': test_acc}
```

```
%%time

errors_ranfor_df6 = pd.DataFrame([max_depth_error_ranfor6(md6) for md6 in range(1, 21)])
```

```
plt.figure()
#plt.plot(errors_ranfor_df2['Max Depth'], errors_ranfor_df2['Training Error'])
plt.plot(errors_ranfor_df6['Max Depth'], errors_ranfor_df6['Validation Error'])
plt.plot(errors_ranfor_df6['Max Depth'], errors_ranfor_df6['Test Error'])

plt.title('Validation vs. Test Error')

plt.xticks(range(0,21, 2))
plt.xlabel('Max. Depth')

plt.ylabel('Prediction Error (1 - Accuracy)')
plt.legend(['Validation','Test']); #plt.ylim((0,1))
```

```
# Tree

def max_depth_error4(md4):
    model = DecisionTreeClassifier(max_depth=md4, random_state=42)
    model.fit(inputs_train,     targets_train)
    train_acc = 1 - model.score(inputs_train, targets_train)
    test_acc = 1 - model.score(test_inputs, test_targets)
    return {'Max Depth': md4, 'Training Error': train_acc, 'Test Error': test_acc}
```

```
%%time

errors_df4 = pd.DataFrame([max_depth_error4(md4) for md4 in range(1, 21)])
```

```
errors_df4
```

```
plt.figure()

plt.plot(errors_df4['Max Depth'], errors_df4['Training Error'])
plt.plot(errors_df4['Max Depth'], errors_df4['Test Error'])
plt.title('Training vs. Test Error')

plt.xticks(range(0,21, 2))
```

```
plt.xlabel('Max. Depth')
plt.ylabel('Prediction Error (1 - Accuracy)')
plt.legend(['Training', 'Test']); #plt.ylim((0,1))
```

```
# Forest

def max_depth_acc_ranfor3(md3):

        model = RandomForestClassifier(n_jobs=-1,
                                       random_state=100,
                                       n_estimators=500,
                                       max_features=7,
                                       max_depth=md3)
        model.fit(inputs_train, targets_train)
        train_acc = model.score(inputs_train, targets_train)
        test_acc = model.score(test_inputs, test_targets)
        return {'Max Depth': md3, 'Training Accuracy': train_acc, 'Test Accuracy': test_acc}
```

```
%%time

acc_ranfor_df3 = pd.DataFrame([max_depth_acc_ranfor3(md3) for md3 in range(1, 22)])
```

```
acc_ranfor_df3
```

```
plt.figure()
plt.plot(acc_ranfor_df3['Max Depth'], acc_ranfor_df3['Training Accuracy'])
plt.plot(acc_ranfor_df3['Max Depth'], acc_ranfor_df3['Test Accuracy'])
plt.title('Training vs. Test Accuracy')
plt.xticks(range(0,21, 2))
plt.xlabel('Max. Depth')
plt.ylabel('Prediction Accuracy %')
plt.legend(['Training', 'Test']);
#plt.ylim((0.7,1.1))
```

```
# Tree test accuracy

def max_depth_acc5(md5):
    model = DecisionTreeClassifier(max_depth=md5, random_state=42)
    model.fit(inputs_train,    targets_train)
    train_acc = model.score(inputs_train, targets_train)
    test_acc = model.score(test_inputs, test_targets)
    return {'Max Depth': md5, 'Training Accuracy': train_acc, 'Test Accuracy': test_acc}
```

```
%%time

acc_tree_df5 = pd.DataFrame([max_depth_acc5(md5) for md5 in range(1, 15)])
```

```
plt.figure()
plt.plot(acc_tree_df5['Max Depth'], acc_tree_df5['Training Accuracy'])
plt.plot(acc_tree_df5['Max Depth'], acc_tree_df5['Test Accuracy'])
plt.title('Training vs. Test Accuracy')
plt.xticks(range(0,21, 2))
plt.xlabel('Max. Depth')
plt.ylabel('Prediction Accuracy)')
plt.legend(['Training', 'Test']);
#plt.ylim((0,1))
```

```
# Tree validation vs test

def max_depth_error7(md7):
    model = DecisionTreeClassifier(max_depth=md7, random_state=42)
    model.fit(inputs_train,    targets_train)
    val_acc = 1 - model.score(inputs_val, targets_val)
    test_acc = 1 - model.score(test_inputs, test_targets)
    return {'Max Depth': md7, 'Validation Error': val_acc, 'Test Error': test_acc}
```

```
%%time

errors_df7 = pd.DataFrame([max_depth_error7(md7) for md7 in range(1, 21)])
```

```
plt.figure()
plt.plot(errors_df7['Max Depth'], errors_df7['Validation Error'])
plt.plot(errors_df7['Max Depth'], errors_df4['Test Error'])
plt.title('Validation vs. Test Error')
plt.xticks(range(0,21, 2))
plt.xlabel('Max. Depth')
plt.ylabel('Prediction Error (1 - Accuracy)')
plt.legend(['Validaion', 'Test']);
#plt.ylim((0,1))
```

```

```

Appendix A2 Python code for K-Means clustering model

Production data clustering

```
import pandas as pd
import numpy as np

from sklearn.cluster import KMeans
import matplotlib.pyplot as plt
import seaborn as sns

from sklearn.preprocessing import LabelEncoder

from sklearn.metrics import silhouette_score
```

```
# Load the data into a pandas DataFrame
data = pd.read_csv(r'C:\Users\moh-m\Desktop\dataset\Last 1st model data files\02. Cleaned data for training and validation\Training and Validation Data.csv')
```

```
# Separate the input features from the target variable

input_data = data.drop(columns =['ALIAS','Date', 'AL', 'Type_Size'])#, axis=1)
```

```
# Perform label encoding for categorical variables

le = LabelEncoder()

for col in input_data.select_dtypes(include='object').columns:
    input_data[col] = le.fit_transform(input_data[col])
```

```
num_cols = ['RUN_PERIOD',
            'WELLHEAD_PRESS',
             'Total_Fluid', 'GOR',
             'OIL',
             'GAS',
             'Water',
             'SAND',
             'WC%',
             'IOR_EOR']
```

```
from sklearn.preprocessing import MinMaxScaler
scaler = MinMaxScaler().fit(input_data[num_cols])
```

```
input_data[num_cols] = scaler.transform(input_data[num_cols])
```

```
input_data.describe()
```

[ ]:

```
# Determine the optimal number of clusters using the elbow method
inertia = []
silhouette = []
k_values = range(2, 25)
for k in k_values:
    kmeans = KMeans(n_clusters=k, random_state=42)
    kmeans.fit(input_data)
    inertia.append(kmeans.inertia_)
    silhouette.append(silhouette_score(input_data, kmeans.labels_))
```

```
# Plot the inertia values
plt.plot(k_values, inertia, 'bo-')
plt.xlabel('Number of Clusters (k)')
plt.ylabel('Inertia')
plt.title('Inertia vs. Number of Clusters')
plt.show()
```

```
# Plot the silhouette scores
plt.plot(k_values, silhouette, 'bo-')
plt.xlabel('Number of Clusters (k)')
plt.ylabel('Silhouette Score')
plt.title('Silhouette Score vs. Number of Clusters')
plt.show()
```

```
# Choose the optimal number of clusters based on the plots or domain knowledge
k = 6

# Perform K-means clustering with the chosen number of clusters
kmeans = KMeans(n_clusters=k, random_state=42)
kmeans.fit(input_data)

# Add the cluster labels to the original data
data['cluster'] = kmeans.labels_
```

```
# Visualize the clusters
sns.scatterplot(x='Total_Fluid', y='AL', hue='cluster', data=data)
plt.xlabel('Total_Fluid')
plt.ylabel('AL')
plt.title('K-means Clustering')
plt.show()
```

```
# Visualize the clusters

sns.scatterplot(x='IOR_EOR', y='AL', hue='cluster', data=data)
```

```
plt.xlabel('IOR_EOR')
plt.ylabel('AL')

plt.title('K-means Clustering') plt.show()
```

```
kmeans.inertia_
```

```
#Perform dimensionality reduction using PCA

from sklearn.decomposition import PCA

pca = PCA(n_components=2, random_state=42) input_data_pca =
pca.fit_transform(input_data)

# Plot the clusters with PCA representation

plt.figure(figsize=(8, 6))

plt.scatter(input_data_pca[:, 0], input_data_pca[:, 1], c=kmeans.labels_,␣
↪cmap='rainbow')

plt.xlabel('Principal Component 1')

plt.ylabel('Principal Component 2')

plt.title('K-means Clustering with PCA Visualization')

#for i, txt in enumerate(data.index):
    # plt.annotate(txt, (input_data_pca[i, 0], input_data_pca[i, 1]), fontsize=8)

plt.show()
```

```
# Plot the clusters with PCA representation

plt.figure(figsize=(8, 6))
```

```
cluster_center = pca.transform([kmeans.cluster_centers_[i]])
plt.scatter(input_data_pca[kmeans.labels_ == i, 0], input_data_pca[kmeans.labels_ == i, 1],
label=f'Cluster {i+1}')

plt.text(cluster_center[0, 0], cluster_center[0, 1], f'Cluster {i+1}', fontsize=12,
fontweight='bold')

plt.xlabel('Principal Component 1')

plt.ylabel('Principal Component 2')

plt.title('K-means Clustering with PCA Visualization') plt.legend()

plt.show()
```

```
# Plot the clusters with PCA representation and feature names

plt.figure(figsize=(8, 6))

for cluster_num in range(k): cluster_features =
    input_data.columns

    cluster_centroid = input_data_pca[kmeans.labels_ == cluster_num].mean(axis=0)

    plt.scatter(input_data_pca[kmeans.labels_ == cluster_num, 0],
↪input_data_pca[kmeans.labels_ == cluster_num, 1], label=f'Cluster{cluster_num}', alpha=0.7)

        plt.text(cluster_centroid[0], cluster_centroid[1], '\n'.

↪join(cluster_features), fontsize=8, ha='center', va='center')

plt.xlabel('Principal Component 1')

plt.ylabel('Principal Component 2')

plt.title('K-means Clustering with PCA Visualization') plt.legend()

plt.show()
```

Operation Data Clustering

```
# Load the data into a pandas DataFrame

data2 = pd.read_csv(r'C:\Users\moh-m\Desktop\2nd Model AL selection objective-2\Training
dataset\Training_dataset_operation_objective_2.csv')
```

```
data2
```

[ ]:

```
# Separate the input features from the target variable

input_data2 = data2.drop(columns =['ALIAS','WO_Start','WO_Start', 'AL', 'Type_Size'])
```

```
# Perform label encoding for categorical variables

le = LabelEncoder()

for col in input_data2.select_dtypes(include='object').columns: input_data2[col]
    = le.fit_transform(input_data2[col])
```

[ ]

```
num_cols = ['WO_Year',

            'WO_No ',

            'Run_Period',
            'TVD',

            'PBTD',

            'Mid_Perf',
            'Setting_Depth',
            'Zone_Thick', 'Tbg_Size',
            'WO_Cause',
            'Failure_Cause']
```

```
from sklearn.preprocessing import MinMaxScaler

scaler2 = MinMaxScaler().fit(input_data2[num_cols])

input_data2[num_cols] = scaler2.transform(input_data2[num_cols])
```

```
input_data2.describe()
```

```
# Determine the optimal number of clusters using the elbow method

inertia = []

silhouette = []

k_values = range(2, 25)

for k in k_values:

    kmeans = KMeans(n_clusters=k, random_state=42)
    kmeans.fit(input_data2)

    inertia.append(kmeans.inertia_)

    silhouette.append(silhouette_score(input_data2, kmeans.labels_))
```

```
# Plot the inertia values
plt.plot(k_values, inertia, 'bo-')
plt.xlabel('Number of Clusters (k)')
plt.ylabel('Inertia')

plt.title('Inertia vs. Number of Clusters')

plt.show()
```

```
# Plot the silhouette scores
plt.plot(k_values, silhouette, 'bo-')
plt.xlabel('Number of Clusters (k)')
plt.ylabel('Silhouette Score')

plt.title('Silhouette Score vs. Number of Clusters') plt.show()
```

```
# Choose the optimal number of clusters based on the plots or domain knowledge

k = 6

# Perform K-means clustering with the chosen number of clusters
```

```
kmeans.fit(input_data2)

# Add the cluster labels to the original data

data2['cluster'] = kmeans.labels_
```

```
kmeans.inertia_
```

```
#Perform dimensionality reduction using PCA

from sklearn.decomposition import PCA

pca = PCA(n_components=2, random_state=42)

input_data2_pca = pca.fit_transform(input_data2)

# Plot the clusters with PCA representation

plt.figure(figsize=(8, 6))

plt.scatter(input_data2_pca[:, 0], input_data2_pca[:, 1], c=kmeans.labels_,cmap='rainbow')

# Annotate the data points with feature labels
#for i, (x, y) in enumerate(input_data_pca):

    #plt.annotate(', '.join([f'{col}: {val}' for col, val in input_data.iloc[i].
→items()]), (x, y), textcoords="offset points", xytext=(0,10), ha='center')

plt.xlabel('Principal Component 1')

plt.ylabel('Principal Component 2')

plt.title('K-means Clustering with PCA Visualization')

plt.show()
```

```
# Plot the clusters with PCA representation

plt.figure(figsize=(8, 6))

for i in range(k):

    cluster_center = pca.transform([kmeans.cluster_centers_[i]])
    plt.scatter(input_data2_pca[kmeans.labels_ == i, 0], input_data2_pca[kmeans.labels_ == i, 1],
    label=f'Cluster {i+1}')

    plt.text(cluster_center[0, 0], cluster_center[0, 1], f'Cluster {i+1}',fontsize=12,
    fontweight='bold')


plt.xlabel('Principal Component 1')

plt.ylabel('Principal Component 2')

plt.title('K-means Clustering with PCA Visualization')

plt.legend()

plt.show()
```

```
# Plot the clusters with PCA representation and feature names

plt.figure(figsize=(8, 6))

for cluster_num in range(k):

    cluster_features = input_data2.columns

    cluster_centroid = input_data2_pca[kmeans.labels_ == cluster_num].mean(axis=0)

    plt.scatter(input_data2_pca[kmeans.labels_ == cluster_num, 0], ␣
↪input_data2_pca[kmeans.labels_ == cluster_num, 1], label=f'Cluster{cluster_num}', alpha=0.7)

        plt.text(cluster_centroid[0], cluster_centroid[1], '\n'.

↪join(cluster_features), fontsize=8, ha='center', va='center')


plt.xlabel('Principal Component 1')

plt.ylabel('Principal Component 2')

plt.title('K-means Clustering with PCA Visualization') plt.legend()

plt.show()
```

```
# Create a new DataFrame to store the cluster information and feature names
cluster_df = pd.DataFrame(input_data2_pca, columns=['PC1', 'PC2']) cluster_df['cluster'] = clusters

# Print the feature names for each cluster

for cluster_num in range(k):

    cluster_mask = cluster_df['cluster'] == cluster_num cluster_features =
    input_data2.columns[cluster_mask]

    print(f"Cluster {cluster_num + 1} features: {', '.join(cluster_features)}")
```

Environmental and Economic Data Clustering

```
# Load the data into a pandas DataFrame

data3 = pd.read_csv(r'C:\Users\moh-m\Desktop\3rd Model AL selection obj-3\Training and Validation dataset\Training_Validation_dataset_obj_3.csv')
```

```
data3
```

[ ]:

```
# Separate the input features from the target variable

input_data3 = data3.drop(columns =['ALIAS', 'AL', 'Type_Size'])
```

```
# Perform label encoding for categorical variables

le = LabelEncoder()

for col in input_data3.select_dtypes(include='object').columns: input_data3[col]
    = le.fit_transform(input_data3[col])
```

```
num_cols = ['Wo_No',
                'Compl_Wo',
                'Purchase', 'Power',
                'Gas_emit',
                'Oil_spill', 'Noise',
                'Operator_act']
```

```
from sklearn.preprocessing import MinMaxScaler scaler3
= MinMaxScaler().fit(input_data3[num_cols])

input_data3[num_cols] = scaler3.transform(input_data3[num_cols])
```

```
input_data3.describe()
```

```
# Determine the optimal number of clusters using the elbow method
inertia = []
silhouette = []
k_values = range(2, 25)
for k in k_values:
    kmeans = KMeans(n_clusters=k, random_state=42)
    kmeans.fit(input_data3)
    inertia.append(kmeans.inertia_)
    silhouette.append(silhouette_score(input_data3, kmeans.labels_))
```

```
# Plot the inertia values
plt.plot(k_values, inertia, 'bo-')
plt.xlabel('Number of Clusters (k)')
plt.ylabel('Inertia')

plt.title('Inertia vs. Number of Clusters')

plt.show()
```

```
# Plot the silhouette scores
plt.plot(k_values, silhouette, 'bo-')
plt.xlabel('Number of Clusters (k)')
plt.ylabel('Silhouette Score')

plt.title('Silhouette Score vs. Number of Clusters') plt.show()
```

```
# Choose the optimal number of clusters based on the plots or domain knowledge

k = 5

# Perform K-means clustering with the chosen number of clusters kmeans
= KMeans(n_clusters=k, random_state=42) kmeans.fit(input_data3)

# Add the cluster labels to the original data

data3['cluster'] = kmeans.labels_
```

```
kmeans.inertia_
```

```
#Perform dimensionality reduction using PCA

from sklearn.decomposition import PCA

pca = PCA(n_components=2, random_state=42) input_data3_pca
= pca.fit_transform(input_data3)

# Plot the clusters with PCA representation

plt.figure(figsize=(8, 6))

plt.scatter(input_data3_pca[:, 0], input_data3_pca[:, 1], c=kmeans.labels_, cmap='rainbow')
```

```
# Annotate the data points with feature labels
#for i, (x, y) in enumerate(input_data_pca):
    #plt.annotate(', '.join([f'{col}: {val}' for col, val in input_data.iloc[i].
↪items()]), (x, y), textcoords="offset points", xytext=(0,10), ha='center')

plt.xlabel('Principal Component 1')
plt.ylabel('Principal Component 2')
plt.title('K-means Clustering with PCA Visualization')
plt.show()
```

```
# Plot the clusters with PCA representation
plt.figure(figsize=(8, 6))
for i in range(k):
    cluster_center = pca.transform([kmeans.cluster_centers_[i]])
    plt.scatter(input_data3_pca[kmeans.labels_ == i, 0], input_data3_pca[kmeans.
↪labels_ == i, 1], label=f'Cluster {i+1}')
    plt.text(cluster_center[0, 0], cluster_center[0, 1], f'Cluster {i+1}', fontsize=12,
    fontweight='bold')

plt.xlabel('Principal Component 1')
plt.ylabel('Principal Component 2')
plt.title('K-means Clustering with PCA Visualization') plt.legend()
plt.show()
```

```
# Plot the clusters with PCA representation and feature names
plt.figure(figsize=(8, 6))
for cluster_num in range(k): cluster_features =
    input_data3.columns
    cluster_centroid = input_data3_pca[kmeans.labels_ == cluster_num].
↪mean(axis=0)
    plt.scatter(input_data3_pca[kmeans.labels_ == cluster_num, 0], ␣
↪input_data3_pca[kmeans.labels_ == cluster_num, 1], label=f'Cluster{cluster_num}', alpha=0.7)
        plt.text(cluster_centroid[0], cluster_centroid[1], '\n'.
↪join(cluster_features), fontsize=8, ha='center', va='center')

plt.xlabel('Principal Component 1')
plt.ylabel('Principal Component 2')
plt.title('K-means Clustering with PCA Visualization')
plt.legend()
plt.show()
```

```
# Create a new DataFrame to store the cluster information and feature names
cluster_df = pd.DataFrame(input_data3_pca, columns=['PC1', 'PC2']) cluster_df['cluster'] = clusters
```

```
# Print the feature names for each cluster

for cluster_num in range(k):

    cluster_mask = cluster_df['cluster'] == cluster_num cluster_features =
    input_data3.columns[cluster_mask]
```

```
clusters
```

```
cluster_df['clusters']
```

```

```

Appendix A3 Python Code for Data Pre-processing model

This model was used to clean the data of the three models

```
import pandas as pd import
numpy as np

from sklearn import preprocessing

import matplotlib.pyplot as plt
```

```
df = pd.read_csv(r'C:\Users\moh-m\Desktop\dataset\last model files\02. Cleaned data for trainig and validation\new training wells nickname added.csv') df.head(10)
```

```
df.shape
```

```
df.dtypes
```

```
#Check the duplicates

# calculate duplicates

dups = df.duplicated()

# report if there are any duplicates

print(dups.any())
```

```
# delete duplicate rows
df.drop_duplicates(inplace=True) print(df.shape)
```

```
df.describe()
```

```
[ ]:
# we need to delete the rows with shut-in and workover periods as there are no production data available.

#Also, any running period less than 10 hours might result in outliers as any recorded data will be lower than the

# normal operation time. thus it would be better to delete them.

df.drop(df[df['RUN_PERIOD'] < 10].index, inplace = True) print(df.shape)
```

```
# as we can see now the samples have decreased from 16140 to 15329 since we removed 556 rows those have no data
```

```
df.describe()
```

```
df.drop(df[df['WELLHEAD_PRESS'] < 10].index, inplace = True) print(df.shape)
```

```
df.describe()
```

```
df.drop(df[df['Total_Fluid'] < 1].index, inplace = True) print(df.shape)
```

```
df.describe()
```

```
# we see there are outliers within the data for instance the sand column have values of 785 and 250 BBL which is unreasonable

# and would definitely result in errors to the model and reduce the accuracy. so let's delete any column has a value greater than 10 BBL

df.drop(df[df['SAND'] > 10].index, inplace = True) print(df.shape)


# now the dataframe have been reduced from 18568 to 17804
```

```
df.describe()
```

```
df.info()
```

```
df.to_excel('New added well GOR and GAS missing.xlsx')
```

```
# now we need to calculate the mean GOR in wells datafrmae in order to substitute nan values in each ALIAS

#JS-01, JS-04, H-1 and H-6

# let's first calculate the mean

avg_gor = df['GOR'].groupby(df['ALIAS']) # let's

print the mean values print(avg_gor.mean())
```

```
# now let's replace nan values in each column with the mean of the ALIAS groups
df['GOR'] = df.groupby(['ALIAS'])['GOR'].transform(lambda x: x.fillna(x.mean()))
#count the number of nan
print(df.isna().sum().sum())
```

```
df.info()
```

```
df.to_excel('New added well cleaned.xlsx')
```

Appendix B1 Feature Importance Criteria

RF has a built-in function that uses the weights of each feature to calculate the importance coefficients. Feature importance mechanism (https://scikit-learn.org) is illustrated below.

In Scikit-Learn, the Gini index metric (*equation 3.13 section 3.7.1.4*) serves to evaluate node impurity. Feature importance, in essence, indicates the decrease in node impurity, weighted by the proportion of samples reaching the node relative to the total sample count, known as node probability. Consider a tree structure with two child nodes, as expressed by the equation:

$$ni_j = w_j C_j - w_{left(j)} C_{left(j)} - w_{right(j)} C_{right(j)}$$

Where:

ni_j is node j importance

w_j is weighted number of samples reaching node j

C_j is impurity value of node j

$left(j)$ child node on left of node j

$right(j)$ child node on right of node j

The above formula provides the significance of a node j, utilized to compute the importance of features for each tree. A specific feature may be employed across various branches of the tree. Feature importance fi_i is calculated as follows:

$$fi_i = \frac{\sum_{j:node\ j\ splits\ on\ feature\ i} ni_l}{\sum_{j \in all\ nodes} ni_l}$$

Appendix B2 Important AL selection production features values

feature	importance
GAS	0.227180
Total_Fluid	0.141392
WELLHEAD_PRESS	0.126647
GOR	0.120649
Water	0.078471
IOR_EOR_SF	0.067489
IOR_EOR_CSS	0.064826
Year	0.047490
IOR_EOR_None	0.043438
OIL	0.039649

Appendix B3 Important AL size selection production features values

Feature	Importance
AL_GL	0.136551
Total_Fluid	0.124671
AL_ESP	0.121222
GAS	0.111911
Year	0.081689
AL_PCP	0.068973
Water	0.068075
GOR	0.067088
WELLHEAD_PRESS	0.051775
WC%	0.036188

Appendix B4 Important AL selection operation features values

Feature	Importance
Setting_Depth	0.261042
Mid_Perf	0.146319
Tbg_Size	0.142204
PBTD	0.112724
TVD	0.086493
WO_Year	0.055906
Zone_Thick	0.047520
Run_Period	0.031914
WO_Cause_CSS_cycle	0.028936
WO_No	0.019237

Appendix B5 Important AL selection environmental and economic features values

Feature	Importance
Purchase	0.213343
Operator_act_Excellent	0.120842
Noise_Medium	0.104681
Oil_spill_Medium	0.104407
Power_Natural	0.058127
Gas_emit_Medium	0.053163
Oil_spill_Low	0.045222
Gas_emit_High	0.043062
Oil_spill_High	0.035220
Power_Gas	0.031121

Appendix C1 Well XFE26 data

Parameter	**Value**
Well name	XFE26
Oil gravity (API)	18
GOR (scf/STB)	0
Water salinity (ppm)	10000
Moles of H2S, CO2, N2 (%)	0
Reservoir pressure (psi)	1500
Reservoir temperature (°F)	518
Bubble point pressure (psig)	1500
WC%	8
PI (STB/d/psi)	0.5
Well TVD (ft)	2300
Casing ID (in)	9.625
Tubing depth/OD (ft/in)	1578, 4.5
Pump setting depth (ft)	1569
Formation depth (mid perforation) (ft)	1689

Appendix C2 Well XJS9 data

Parameter	**Value**
Well name	XJS9
Oil gravity (API)	34
GOR (scf/STB)	1000
Gas specific gravity	0.72
Water salinity (ppm)	10000
Moles of H2S, CO2, N2 (%)	0, 0.42, 0.36
Reservoir pressure (psi)	3500
Reservoir temperature (°F)	167
Bubble point pressure (psig)	2000
WC%	19
PI (STB/d/psi)	1
Well TVD (ft)	2620
Casing ID (in)	9.625
Choke size (in)	50
Tubing depth/OD (ft/in)	2.875
GL valves setting depth (ft)	1300,1500,1800
Formation depth (mid perforation) (ft)	2090
Zone thickness (ft)	26
GL injection rate (MMscf/day)	0.49
Casing pressure (psi)	1500