THE REPUBLIC OF TURKEY

BAHCESEHIR UNIVERSITY

BIG DATA ANALYTICS IN OIL & GAS INDUSTRY A CASE STUDY: PREDICTING ROP IN DRILLING OPERATIONS WITH BIG DATA AND MACHINE LEARNING FOR RUMAILA OILFIELD

Master's Thesis

DUHA ALSAHLANEE

ISTANBUL, 2020

THE REPUBLIC OF TURKEY

BAHCESEHIR UNIVERSITY

GRADUATED SCHOOL OF NATURAL AND APPLIED SCIENCE

INDUSTRIAL ENGINEERING DEPARTMENT

BIG DATA ANALYTICS IN OIL & GAS INDUSTRY A CASE STUDY: PREDICTING ROP IN DRILLING OPERATIONS WITH BIG DATA AND MACHINE LEARNING FOR RUMAILA OILFIELD

Master's Thesis

DUHA ALSAHLANEE

Thesis Supervisor: PROF. DR. MUSTAFA ÖZBAYRAK

ISTANBUL, 2020

THE REPUBLIC OF TURKEY
BAHCESEHIR UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
INDUSTRIAL ENGINEERING

Name of the thesis: Big Data Analytics In Oil & Gas Industry A Case Study: Predicting Rop In Drilling Operations With Big Data And Machine Learning For Rumaila Oilfield
Name/Last Name of the Student: Duha ALSAHLANEE
Date of the Defense of Thesis: 29.06.2020
The thesis has been approved by the Graduate School of Natural and Applied Sciences.

Assoc. Prof. Burak KÜNTAY
Graduate School Director

I certify that this thesis meets all the requirements as a thesis for the degree of Master of Science.

This is to certify that we have read this thesis and we find it fully adequate in scope, quality and content, as a thesis for the degree of Master of Science.

Examining Comittee Members	Signature
Thesis Supervisor Prof. Dr. Mustafa ÖZBAYRAK	----------
Member Assist. Prof. Dr. Adnan ÇORUM	----------
Member Assoc. Prof. Mehmet Fatih HOCAOĞLU	----------

ACKNOWLEDGEMENS

It has been a great opportunity for me to study a master in industrial engineering, I learned many new things during those two years of my life, I grew and developed new skills which I believe would enhance my career life in the future. My big thank goes to my supervisor PROF. Mustafa Özbayrak, he has been my guide through this journey with his precious advice and support.

My dear friend the drilling supervisor Mohammed Al-Rashed has been a great supporter for me throughout the project. He helped me a lot with drilling part of this thesis and provided me with all the required data to build this project form Haliburton company that he works in, also I would I like to thank the company for permitting me to use their data and for giving me a chance to visit one of their drilling sites and staying there to see the real life working environment.

Also, my friend Emirhan Oruç who contribute to this research with his experience in neural networks. He provided me with many useful comments that help to make this work successful.

Finally, I would like to give very sincere thanks to my father, mother, all the family members, and all my friends who always have been next to me and provided the support that increased the quality of this research.

Duha Al-Sahlanee

ABSTRACT

BIG DATA ANALYTICS IN OIL & GAS INDUSTRY A CASE STUDY: PREDICTING ROP IN DRILLING OPERATIONS WITH BIG DATA AND MACHINE LEARNING FOR RUMAILA OILFIELD

Duha ALSAHLANEE

Industrial Engineering Department

Thesis Supervisor: Prof. DR. Mustafa ÖZBAYRAK

June 2020, 50 pages

Big data concept is well known for its potentials in various industries, unlike the other industries oil and gas industry is still lagging behind in this area. The purposes of this thesis are, first to describe the benefits of big data to the oil and gas sector and drilling operations sector specifically with providing examples of companies starting their own big data initiatives. Second, use big data and machine learning algorithms to predict the rate of penetration in drilling operations. Predicting the rate of penetration (ROP) has a great impact on lowering the cost of drilling activities and increase the efficiency of drilling programs. The neural network approach used as the analytical tool to analyze the data of 45 different wells drilled in North and South Rumaila oilfield in the south of Iraq. The ANN used to build a model that can predict ROP using seven drilling parameters (TVD, WOB, RPM, Torque, SPP, Flow in, Mud density). Three layers model with two hidden layers and the output layer has built. 70% of the dataset used to train the ANN while 30% used for both validation and testing sets (20% for validation set and 10% for testing set) to evaluate the performance of the developed model. The results showed quite good predictions of ROP values with MSE of 0.01 and R^2 of 0.70. The model used for real applications with new well and showed a good matching between predicting values of ROP and the raw data of the well. The results ensured the ability of big data and machine learning to enhance drilling operations.

Keywords: big data analytics, oil and gas industry, drilling operations, rate of penetration, artificial neural network (ANN)

ÖZET

PETROL VE GAZ ENDÜSTRİSİNDE BÜYÜK VERİ ANALİTİKLERİ BİR VAKA ÇALIŞMASI: RUMAILA OILFIELD İÇİN BÜYÜK VERİ VE MAKİNA ÖĞRENME İLE SONDAJ İŞLEMLERİNDE TAHMİN HALATI

Duha ALSAHLANEE

Endüstri Mühendisliği Bölümü

Tez danışmanı: Prof. DR. Mustafa ÖZBAYRAK

Haziran 2020, 50 sayfa

Büyük veri farklı endüstriyel alanlardaki potansiyeli ile zaten çok iyi bilinmektedir ancak diğer endüstri kollarından farklı olarak Gaz ve Petrol endüstrisindeki kullanımı henüz diğer endüstriyel alanlar seviyesinde değildir. Bu tez'in amacı, öncelikle büyük verinin petrol ve gaz endüstrisinde kullanılmasının faydalarını tanımlamak, ama özellikle de petrol ve gaz arama işlemlerinde kullanılmaya başlanmasıyla elde edilecek avantajları örneklerle göstermektir. İkinci olarak, büyük veri'yi makina öğrenmesi algoritması ile birlikte kullanarak arama işlemlerinde delme oranının oranını tahmin etmede kullanmaktır. Delme oranının tahmin edilmesi delme işleminin maliyetlerinin azaltılması ve delme programlarının etkinliğinin artırılması konusunda çok önemlidir.

Tahmin işlemleri için yapay sinir ağları yaklaşımı kullanılmıştır ve Güney Irak'taki Kuzey ve Güney Rumalia bölgesinde yer alan 45 farklı kuyunun delinmesi konusundaki verilerin analizi için kullanılmıştır. Model'in kurulumunda yapay sinir ağları (Artificial Neural Networks - ANN) kullanılmıştır ve yedi farklı parametre (TVD, WOB, RPM, Tork, SPP, İçeriye Akış, ve Çamur Yoğunluğu) kullanılarak ROP tahmin edilmiştir. İki gizli, bir çıktı katmanı şeklinde üç katmanlı bir model kurulmuştur. Veri setinin %70'i yapay sinir ağını (ANN) eğitmek için kullanılmıştır, geriye kalan %30 veri seti ise kurulan modelin performansını ölçmek için setlerin testi ve gerçekleme için (%20 set'in gerçeklenmesi ve %20 ise set'in test edilmesi için) kullanılmıştır. Sonuçlar ROP değerinin 0.01 MSE ve 0.70 R^2 değerleri ile gayet iyi tahmin edildiğini göstermektedir. Model, yeni kuyu açılması işlemi ile gerçek bir uygulama için kullanılmış ve ROP değerlerinin tahmini ve kuyuya ait verilerin iyi bir eşleşmesini göstermiştir. Sonuçlar Büyük Veri'nin, Makina Öğrenmesi ile beraber kullanıldığı takdirde delme işlemlerinin iyileştirilebileceğinin bir ispatı olmuştur.

Anahtar Kelimeler: Büyük veri analitiği, petrol ve gaz endüstrisi, delme işlemleri, delme oranı, yapay sinir ağları (ANN).

CONTENTS

FIGURES

ABBREVIATIONS

AAPE	:	Average Absolute Percentage Error
ACEs	:	Advance Collaboration Environment Center
ALOS	:	All Assets Have Appropriate Level Of Smartness
SCADA	:	Supervisory Control And Acquisition
COLAB	:	Google Colaboratory
CORE	:	Collaborative Real Time Environment
DT	:	Travel Transit Time
HER	:	Electronic Health Records
HIS	:	Hydraulics
IoT	:	Internet Of Things
IQR	:	Interquartile Range
KPI	:	Key Performance Indicators
LWD	:	Logging While Drilling
MAE	:	Mean Absolute Error
MSE	:	Mean Squared Error
MW	:	Mud Weight
MWD	:	Measurement While Drilling
NaN	:	Not A Number
NPT	:	Nonproductive Time
POC	:	Proof Of Concept

R	:	Correlation Coefficient
R^2	:	Coefficient Of Determination
RPM	:	Rotation Per Minute
SOA	:	Service oriented Approach
SPP	:	Standpipe Pressure
TDS	:	Top Drive System
TVD	:	True Vertical Depth
WOB	:	Weight On Bit

1. INTRODUCTION

The inspiration of this study came with the advancement in technologies that the world witnessing these days in different areas, the new technologies can generate a large amount of data comparing with the old technologies of previous decades. The increasing volume of data is a key success when adopting and finding the appropriate ways of extracting the values within the data which can lead to extremely big changes and improvement of operations for companies seeking long term competitive advantage.

In the same direction, the oil and gas industry starts to utilize new technologies to increase the efficiency of their processes such as automated rigs, new sensor technologies, drones, etc. However, the industry failed in the term of using the best analytical tools that can handle this large amount of data generated by its operations in different phases of the industry. Against this uncertain environment with the ongoing searching of methods to renewal natural resources, and fluctuation of demand and price, the oil and gas companies need to start adopting solutions to deal with the data that they have in hand (Baaziz & Quoniam 2014, p. 2). Like the other industries, finding effective tools to handle and analyze this data could be a step forward in making improvements in this industry.

The most costly and complex phase of this industry is the drilling operations. In the drilling operations, the large volume of data came with the use of sensors to monitor and assess different drilling conditions and parameters. The data generated daily needs new approaches that could interpret the value of the data and help in fastening the decision-making process leading to safely reduce costs and environmental impacts with increasing efficiency of drilling programs.

This study describes the big data concept and its impact on various industries with examples of how some companies already start their big data initiatives. It also explores the potential of big data for the oil and gas industry and drilling operation specifically with steps taken by companies in this sector towards adopting the solutions of big data. A literature review of the researches in this area was done to examine the nature of works and what have done with this subject by others. The rest of the thesis detects the use of machine learning algorithms as analyzing tools for handling big data. The neural

network approach shows effectiveness when dealing with large datasets and complex relationships between the data; therefore, it was used with python language to design a model that can predict the rate of penetration in drilling operation and providing optimized parameters to increase the efficiency of the drilling process. Some limitations while developing the model were described to give an idea about the problems that could be faced for future works. The main conclusion said it's possible to handle the big data with machine learning techniques to analyze the data and produce models that can solve problems in the drilling operation.

2.BIG DATA AND OIL & GAS INDUSTRY

2.1 INTRODUCTION

The oil and gas industry is an extremely competitive and uncertain environment with the eternal need to renewal natural resources, as well as dealing with unstable prices and demands of the products. To solve these issues oil and gas companies need to increase production, optimize costs, and reduce the impact of environmental risks (Baaziz & Quoniam 2014, p. 2). The oil and gas industry produces Petabytes of data and the size is only increasing (Holdaway 2014, p.3). This huge amount of data is what we now call big data. Oil and gas companies can utilize big data technologies to extract value from the data itself which can lead to increase operations performance, optimize business operations, reduce cost, and increase their competitive advantage.

like the other industries, oil and gas industry became a data-driven as a result of collecting a massive amount of data generated by a multitude of the internet of things IOT sensors from operations such as exploration, drilling, and production to provide continuous data-acquisition, real-time monitoring of assets and environmental conditions in all phases of the industry. This digital data enters the data management center and decision making on the hierarchical levels is depended on generating models based on different situations and processes on the basis of this data (Aliguliyev & Imamverdiyev 2017, p.31). The structure of the data comes in different shapes, it could be “structured”, “unstructured” and “semi-structured” making it complex to store them in traditional data warehouses and analyze them. Moreover, the velocity and complexity of data growth have put immense on traditional data systems making it fundamental to change the way data are collected, stored, analyzed, and accessed to support the real-time and decision-making processes.

The emergence of big data technologies in oil and gas sector is the driving force behind ongoing waves of what known as “digital oilfield” which can be described as a category comprises technologies, services, and related business models focused on both the tools and processes for data and information management in upstream oil & gas activities. Different leading companies working in oil and gas sector have started to apply digital oilfield concepts, such as Chevron's i-Fields, BP's Field of the Future, and Shell's Smart

Fields, and they continue to seek new technologies to increase production output and reduce operating costs (cleantech group 2015).

2.2 DEFINITION OF BIG DATA

Big data defined as an enormous volume of datasets that cannot be handled and managed using traditional data processing tools (Ohlhorst 2013, p.1). The ideology of big data describes the situation when data volumes have grown to an enormous size that the traditional information technology systems can no longer manage the size and growth of data making it even harder to extract knowledge out of it (Ohlhorst 2013, p.1).

Big Data is defined by six characteristics called "6V": Volume, Velocity, Variety, Variability, Veracity, and Value. Volume refers to the huge amount of data, variety means different formats of data from various sources, value means the useful insight extracted from the data, veracity is the inconsistencies and uncertainty in data, velocity is the high speed of accumulation of data, and varaiability is the different sources of data.

The data can be classified into 3 types; "structured" refers to datasets that come in the form of rows and columns, "unstructured" such as Audio, Video files, images, etc, and "semi-structured" such as XML, Weblogs, etc.

The past decades witnessed an unexpectable increase in the volume of data generated around us as a result companies around the world started to capture and store these terabytes of data about different processes such as users' interactions, social media, sensors, and business (Ohlhorst 2013, p.2). The challenge that comes with these issues is how to exactly make sense of this data which refers to the analytical process of the data. The process of analyzing this big data to discover the hidden patterns and extract value from it called big data analytics.

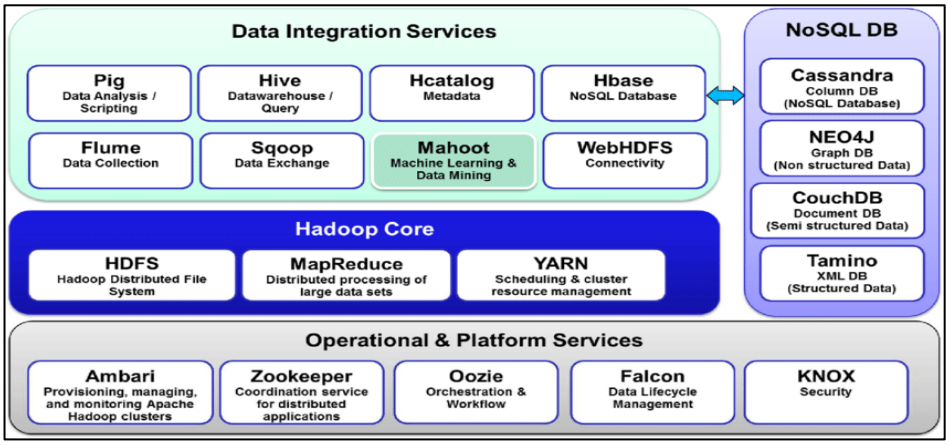
In the term of big data more data means more analysis and more results, so the amount of data that one has at hand is very important as it leads to more understanding of the case, however, the most important step is to determine what big data analytics is going to be used for as the resulting insights can lead to enhance the business and decision making processes.

The increasing number of data generated through different operations created some problems for traditional data systems either because those systems are not designed to handle the variety of today's data or the data systems can not scale quickly and affordably.

One of the famous platforms that can deal with big data, and has the capability of handling problems resulted from massive amounts of data is the Hadoop platform which also offers a benefit of dealing with data in different formats whether it was structure, unstructured or semi-structure (Ohlhorst 2013, p.7). It has the ability to store and process large datasets in parallel and distributed fashion with low cost and simple hardware clusters. It is also scalable and fault-tolerant, so when one node goes down other nodes can process the data, moreover data can be stored in different formats and that makes it more flexible.

Hadoop can break down the data into multiple pieces and then distribute it into various servers. The distributed nature of data makes it possible to access the data from different places. Hadoop tracks the place at which the data is stored and protects it by creating copies of the data stored in multiple servers which means if a server goes down, the data can be replicated from another copy (Ohlhorst 2013, p.8). The Hadoop ecosystem is shown in figure 1.1 below.

Figure 2.1: Hadoop ecosystem



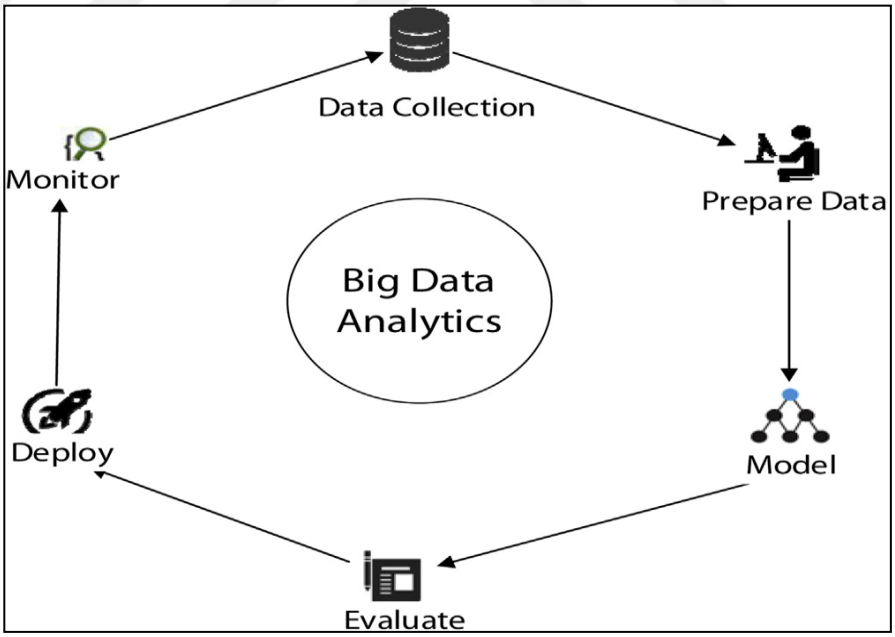
Source: Baaziz, A. & Quoniam, L. (2014). How to use big data technologies to optimize operations in the upstream petroleum industry. *21st World Petroleum Congress.* 15-19 June 2014, Moscow, Russia, pp. 1-9.

2.3 THE WORKING MECHANISM OF BIG DATA ANALYTICS

big data analytics enable enterprises to gain knowledge about their processes from the different data they collected which will be useful to optimize their business and support smart decisions. The amount of data plays a key role in the analytics process providing more data means more accuracy coming with the new insight and predictions (Tankimovich 2018, p.7).

The big data analytics process is divided into six steps as shown in figure 2.2, and includes: (1) big data acquisition from multiple sources, (2) applying data preprocessing methods to increase the quality of raw data, (3) building learning models using statistical methods and machine learning techniques, (4) evaluate the model on the basis on testing dataset, (5) deploying the generated model in real applications, and (6) monitoring the performance of the model in terms of prediction accuracies (Rehman et al. 2016, p.919).

Figure 2.2: Big data analytics process



Source: Rehman, M., Chang, V., Batool, A. & Wah, T. (2016). Big data reduction framework for value creation in sustainable enterprises. *International journal of information management*. 36, pp. 917-928.

Considering optimal data collecting from different sources and be aware of the irrelevant one through the collecting process can help the organization to achieve

optimal extraction of the value. Data preparation is the most important stage in big data analytics, it consists of preprocessing and integration of the data to ensure the quality of big data. the Preprocessing process can bring some advantages to the big datasets such as noise reduction to remove irrelevant data, detecting outliers and removing anomalies to produce high-quality datasets, data elimination, sketching and imputation based methods to handle missing values in big datasets, and many other benefits that secure the quality of big datasets (Rehman et al. 2016, p.919). in other words, the quality of the data and generating knowledge depend on this stage.

Generating the learning models step is based on statistical and machine learning theories to train the data and extract efficient patterns. Machine learning is the science of creating algorithms that learn on their own. It is about automating the data, so no need for human interaction to become better. Big data is the raw materials that feed to the machine learning process, the machine learning algorithms will process the data and identify patterns. Then those patterns can be used on new datasets.

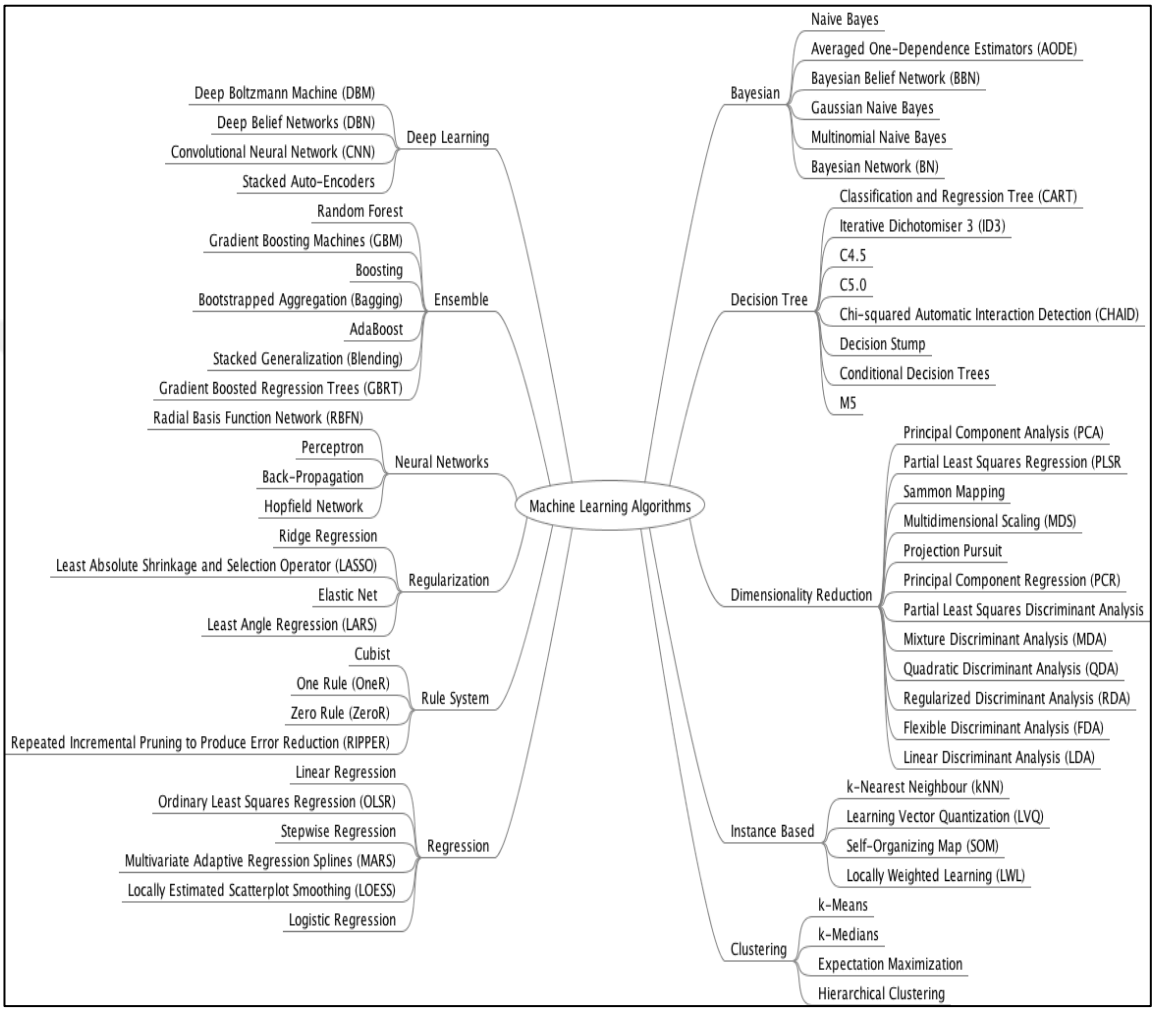
Machine learning models can be classified into supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning. Supervised learning means trained datasets are readily available which consist of a set of data inputs and desired outputs with given roles applied to the datasets to get output. Supervised learning can be classified into a classification which means classifying labeled data, and regression which means predicting trends using previously labeled data.

Unsupervised learning is the opposite, in this situation no trained datasets are available here. The outputs are based on prediction analysis. İt helps to explain the hidden structure from untrained datasets. It includes the clustering process which means finding patterns and groups from unlabeled data.

Semi-supervised learning means that the training data fed to the algorithm is partially labeled. Reinforcement learning, is a little different, the model here is called agent, and its job is to perform actions and it gets rewards, the agent then learns as it is trying to maximize the rewards (towards data science 2019).

There are many other machine learning algorithms as shown in figure 2.3, using any one of them depends on different criteria such as the problem at hand, nature of the data and software available (towards data science 2019).

Figure 2.3: Machine learning algorithms



Source: Towards data science

After creating the model, evaluation of the model using different evaluation methods will take place. Deployment of the learning models through enterprise applications is the next step to capture insight from future big datasets. Finally, to ensure that data can be handled and newly knowledge patterns can be uncovered continuously, monitoring of the learning models through business intelligence (BI) should be taken into consideration.

The analytics process varies in terms of descriptive, predictive, prescriptive, and diagnostic analysis. the Descriptive analysis uses data aggregation and big data mining techniques to provide insight into the past, then it answers what is happening now based on incoming data. In predictive analysis, statistical models and forecast techniques could be used to understand the future and provide answers for what could happen. the Prescriptive analysis uses optimization and simulations algorithms to advise on the possible outcomes, and answer the question of what action should be taken. Diagnostic analytics used to determine why something happened in the past. It helps determine what kind of factors and events contributed to a particular outcome, so mostly it uses probabilities, likelihood, and the distribution of the data for the analysis.

2.4 EXAMPLES OF BIG DATA ANALYTICS IN VARIOUS INDUSTRIES

Today many industries including healthcare, banking, education, manufacturing, retail, and the weather sector can benefit from big data analytics as it can provide solutions for problems leading to cost reductions, time reductions, new product development, and smart decision making.

The rapidly growing amount of data motivates companies around the world to start adopting big data by collecting, storing, and analyzing this data. Big data analytics means a lot to many companies around the world like Amazon, Facebook, and Google who applied this concept in a variety of issues such as marketing and enhancing services to their customers. Other examples are New York Times and Walt Disney Company, both of them used big data analytics tools, the first one used it for text analysis and web mining, while the second one used it for correlating and understanding their customers' behavior (Ohlhorst 2013, p.2).

In education a large number of data related to students, teachers, institutions, courses, and results are being generated every day, this information could be helpful in customized learning programs and schemes for students leading to improve the overall results. Reframing the course material by analyzing the data collected and deciding the proper components that are beneficial for students. Predict student enrollment after a particular course and career prediction to understand which career could be most suitable for a student in the future. The University of Alabama has a large number of

students with a large volume of data. In the past, the data seemed useless because there were no proper methods to analyze it. Nowadays, the university is able to use analytics and data visualizations for this data to draw out patterns of students revolutionizing the university's operations, recruitment, and retention efforts (intellipaat 2016).

In banking, data such as financial reports, stock news, and customer information is used to mitigate risks and fraud, bring insight to customer care, detect money laundering, and misuse of credit cards. National accounting and audit firm (BDO) use big data analytics to identify risk and fraud during audits (intellipaat 2016).

Big data analytics has contributed to healthcare in different ways such as support improved health monitorin and, avoid preventable diseases by detecting them in the early stages. Moreover, it helps in predicting outbreaks of epidemics and decide which action to take in order to minimize the effects. One of the most trends in healthcare is electronic health records (EHR), it stores the patient's entire data then it can be used for analyzing process for various purposes.

In the weather sector, utilizing sensors and satellites have increased the amount of data being collected and analyzing, this data has been used to monitor the weather and environmental conditions. Weather forecasting, predicting, responding to natural disasters, and study global warming are some other benefits of adopting big data in this sector. IBM developed a project called Deep Thunder with the aim of weather forecasting with high-performance computing of big data. This project helped Tokyo to enhance its weather forecasting and predicting the probability of damaged power lines (intellipaat 2016).

One of the main objectives of big data analysis in smart manufacturing is finding new associations, influencing factors and patterns in the data, and observing such findings through Big Data stream observation. The exploitation of this big data analytics will potentially innovate business fields through improved maintenance services (e.g. anomaly/failure, detection/prediction, system observation); pattern observation e.g. for hacker detections; extended manufacturing system reports; KPI improvements/monitoring; customer demands identification based on Big Data analysis (Negorny et al. 2017, p.32). An example of using big data analytics in manufacturing is

in designing a product of higher quality by the Rolls-Royce which manufactures massive jet engines. These engines are used by airlines and armed forces across the world. The company uses big data analytics to analyze how good the engine design is and if there has to be any more improvement.

Retailers use big data analytics mainly for predicting customer purchases, making personalized recommendations of offerings, optimizing supply chains, identify new sources of revenue to increase their growth, and implement customer relationship management strategies.

2.5 OIL AND GAS INDUSTRY OVERVIEW

The oil and gas industry can be classified into three categories: upstream, which includes exploration and discovery, drilling and production; midstream, which includes transportation, wholesale markets, manufacturing, and crude refinement; Downstream which is delivery of the refined products to the consumer. The upstream sector is actively engaging with big data to achieve efficiency gains, while midstream and downstream sectors are lagging behind. The main big sources of data in the oil and gas industry are discovery, drilling, and production.

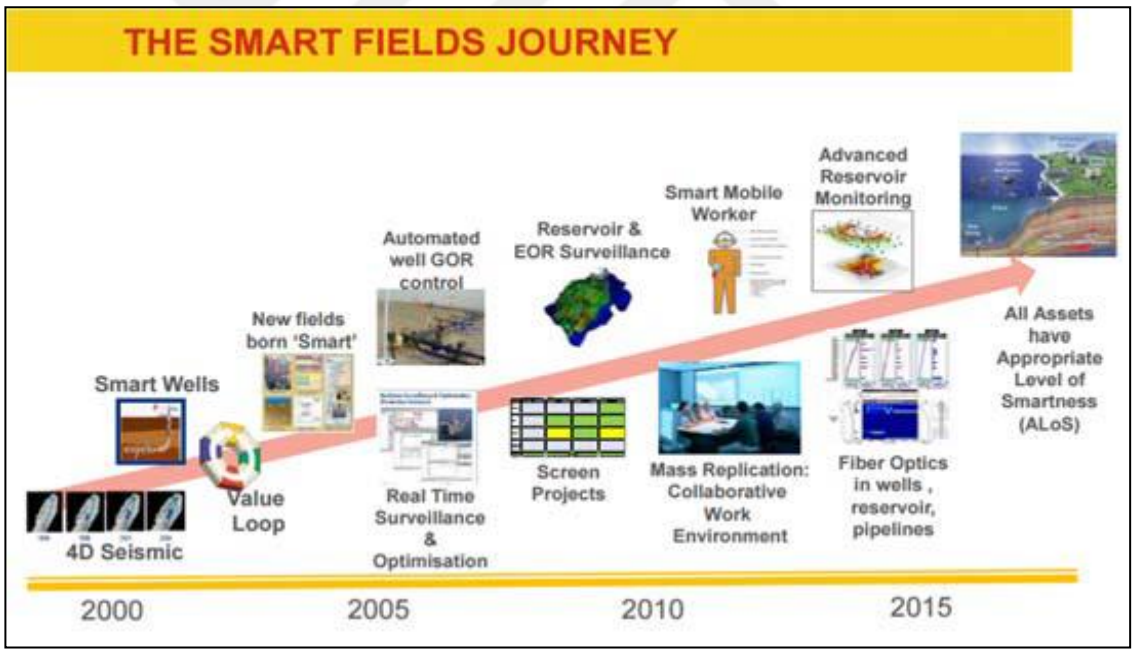
The rapidly changing and competitive market of the oil and gas industry with the problems industry witnesses made it necessary for oil and gas companies to take a step forward and make sense of every bit of data they collect (Microsoft 2014). The increasing amount of data, different sources of the data, and the velocity at which the data is generated create an opportunity for oil and gas firms to capture and manage them to gain measurable value by using the appropriate technology, tools, and expertise to unlock this value. The business analysts take advantage of big data solutions as they want more data at higher rates, stored longer, and analyzed faster.

The digitalization of the oil field, whether the term is smart oil, digital oilfield, i-fields, or smart fields which represented by the instrumentation and automation of equipment and processes, as well as industry-wide focus on the internet of things (IoT) has come to the availability of more data even faster. The innovation in sensor technologies, data analysis, and networking and communications technologies have enabled oil and gas companies to collect critical data in a real-time manner leading to what is now knowing

as the digital oil field. Every oilfield asset is wirelessly connected to allow remote control, asset tracking, and well monitoring. Also, data management solutions allow rapid cross-functional team collaborations across the world which is another application in the digital oilfield (cleantech group 2015).

One of the global companies that start to adopt the digital oil field of the future project is shell by developing a 'smart fields' program. Shell's project consists of three stages, the first stage called 'smart wells' which involved the application of smart analytics to only one well. In the second stage, shell achieves real-time monitoring and optimization of assets based on the first stage. İn the last stage, shell increase its smart wells concept and applied it to a smart mobile terminal calling it 'all assets have an appropriate level of smartness'. The program helped shell to increase its benefits and efficiency of operations (Song 2018, p.288). The 'smart fields' program is shown in figure 2.4.

Figure 2.4: The smart field journey

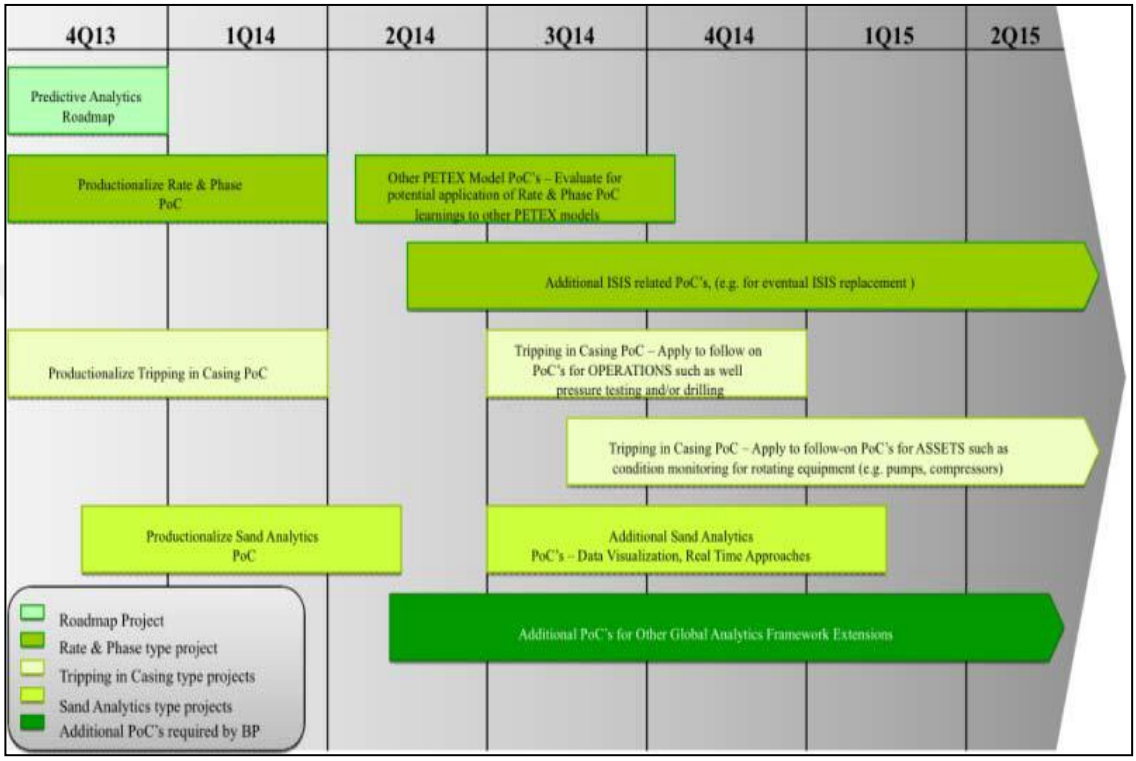


Source: Song, M. (2018). Research on the application of big data in oil & gas ındustry. *International Conference on Computational, Modeling, Simulation, and Mathematical Statistics.* 2018

BP launched a strategic program called "Field Of The Future" that started in 2003. The program consists of big data applications framework, CoRE (Collaborative Real-Time Environment), ACEs (Advanced Collaboration Environment Center). The big data framework allowed BP to accomplish improvement in the efficiency of drilling

operations. Through KPI BP determined big data benefits to its various operations such as increasing drilling program efficiency, decrease the time of equipment failure, optimizing operations schedule, etc (Song 2018, p.288). The roadmap for BP's program is seen in figure 2.5.

Figure 2.5: Roadmap for BP's big data application



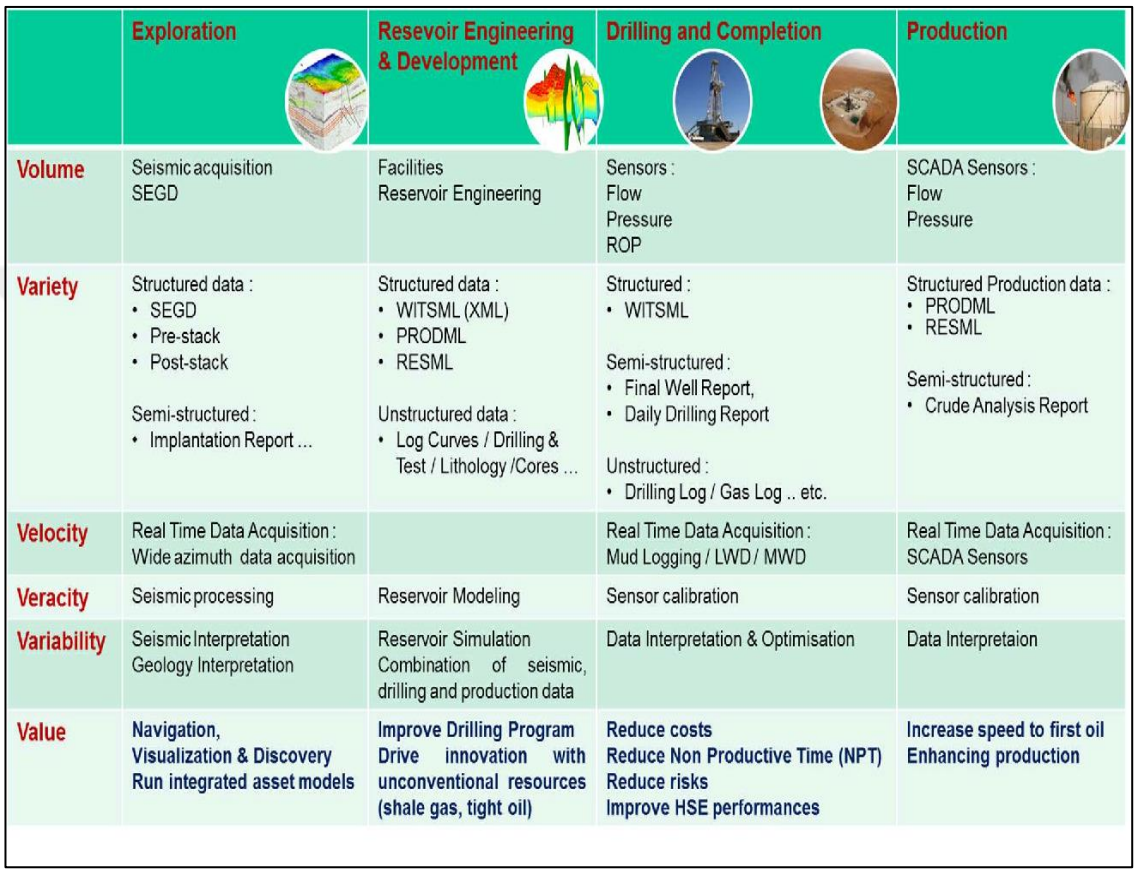
Source: Song, M. (2018). Research on the application of big data in oil & gas ındustry. *International Conference on Computational, Modeling, Simulation, and Mathematical Statistics.* 2018

BP chose "tripping in casing" as their PoC (Proof of Concept) task, BP collected real-time data about drilling from sensors installed on drilling machines to analyze it and use it in building a model that predicts the stuck pipe during drilling operations then compare the predicted model with the historical data to measure the accuracy of the model and predict the potential risk. Testing of the POC ensures that this method can enhance drilling performance (Song 2018, p.289).

In the oil and gas industry, the data volume is coming from sensors, spatial and GPS coordinates, weather services, seismic data, and various measuring devices. Much of this data is "unstructured" or "semi-structured" which means it's difficult or costly to either store in traditional data warehouses or routinely analyzes it (Microsoft 2014, p.4).

The sources of unstructured data in the oil and gas sector are CAD drawings, specifications, seismic, well log, or drilling reports in paper or PDF, web traffic, also social media. The big data characteristics and sources in the upstream sector are shown in figure 2.6 below.

Figure 2.6: Upstream big data



	Exploration	Resevoir Engineering & Development	Drilling and Completion	Production
Volume	Seismic acquisition SEGD	Facilities Reservoir Engineering	Sensors : Flow Pressure ROP	SCADA Sensors : Flow Pressure
Variety	Structured data : • SEGD • Pre-stack • Post-stack Semi-structured : • Implantation Report ...	Structured data : • WITSML (XML) • PRODML • RESML Unstructured data : • Log Curves / Drilling & Test / Lithology /Cores ...	Structured : • WITSML Semi-structured : • Final Well Report, • Daily Drilling Report Unstructured : • Drilling Log / Gas Log .. etc.	Structured Production data : • PRODML • RESML Semi-structured : • Crude Analysis Report
Velocity	Real Time Data Acquisition : Wide azimuth data acquisition		Real Time Data Acquisition : Mud Logging / LWD / MWD	Real Time Data Acquisition : SCADA Sensors
Veracity	Seismic processing	Reservoir Modeling	Sensor calibration	Sensor calibration
Variability	Seismic Interpretation Geology Interpretation	Reservoir Simulation Combination of seismic, drilling and production data	Data Interpretation & Optimisation	Data Interpretaion
Value	**Navigation, Visualization & Discovery Run integrated asset models**	**Improve Drilling Program Drive innovation with unconventional resources (shale gas, tight oil)**	**Reduce costs Reduce Non Productive Time (NPT) Reduce risks Improve HSE performances**	**Increase speed to first oil Enhancing production**

Source: Baaziz, A. & Quoniam, L. (2014). How to use big data technologies to optimize operations in the upstream petroleum industry. *21st World Petroleum Congress.* 15-19 June, 2014, Moscow, Russia, pp. 1-9.

The extensive use of sensor technology is the main source of huge amounts of data generated in the oil and gas industry. Sensors have been used in a large scale of operations in all phases of the industry such as exploration, well drilling, well completion, fracking, refining, production, and decommissioning. Sensors have been used to monitor and sense different parameters such as temperature, pressure, vibrations, etc through different processes. 4C sensors in oil and gas seismic exploration, 4D in the field of geophysical surveys, as well as optical fiber sensors in the wells, oil and gas extraction, processing, and transportation system, are widely used

and they receive scaled data (EMC Corporation 2013). Recent advances in technologies such as the IoT and big data combining with advances in sensing technologies will definitely facilitate better monitoring, security, and management of the oil and gas industry with higher productivity and reduced cost and casualties.

Dealing with a large amount of data is not new to the oil and gas industry. In the exploration phase, high performance computing (HPC) with parallel processing capabilities have been used to perform analysis on massive volumes of data. Also, 3D and 4D visualization have been used to discover new resources and predict changes in reservoir modeling.

In the past searching for potential sites that contain oil and gas were done by monitoring low-frequency seismic wave that moves through the earth below us due to tectonic activity. The pattern of the waves will be distorted when moving through the oil or gas layer. In the past, this would involve collecting thousands of data, but in the past few years with the advancement of technologies, millions of data could be generated from such a process that could help in detecting potential drilling sites. The change that the industry witnesses is the development of digital technologies which vastly increased the amount of data gathering through the industry life cycle.

Royal Dutch Shell is one of the largest oil and gas companies that use fiber optic cables for sensors, and data is transformed to its servers. This data will give a more accurate image of what lies beneath, and enable the geologist to make more accurate recommendations about where to drill. Shell has been developing the idea of a "data-driven field" in order to decrease the drilling cost of its wells. It uses big data to decrease breakdowns and failures of its machines by fitting them with sensors collecting data about its performance and comparing them with aggregated data, this will lead to downtime minimization and replacement of parts in an efficient manner (Forbes 2015).

Big data analytics is able to bring value for the whole lifecycle of the oil and gas industry. In exploration, applying big data and advanced analytics can lead to performing operational decision making, enhancing exploration efforts by using historical drilling and production data to help geologist and geophysicists to verify their assumptions, identify seismic traces using advanced analytics based on Hadoop for

storage, quick visualization and comprehensive processing and imaging of seismic data (EMC Corporation 2013). For drilling operation using big data analytics could bring some benefits such as build assessment of drilling models based on existing well data and refresh the models by utilizing the data coming from sensors in drill rig to optimize drilling parameters, early identifying of anomalies and negative impact factors that would affect drilling operation, using real-time drilling data to for predictive modeling to perform real-time decision making, also predict downtime of drilling equipments (Baaziz & Quoniam 2014, p.5). Big data is of great interest to production and operation work, being able to predict future performance based on historical results, or to identify sub-par production zones can be used to shift assets to more productive areas. Oil recovery rates can be improved, as well, by integrating and analyzing seismic, drilling, and production data to provide self-service business intelligence to reservoir engineers (Baaziz & Quoniam 2014, p.6).

3. LITERATURE REVIEW

Since the emergence of big data analytics concept and its potential to different sectors, the oil and gas sector took the same wave, various review papers and researches have been conducted to study the effect of adopting this idea in oil and gas industry. Potentials, strategies, steps, frameworks, and challenges have been illustrated to support oil and gas companies making their business more valuable.

Azzedin & Ghaleb (2019) have investigated an architecture for handling big data in Oil and Gas industries: Service-Oriented Approach to enables the petroleum industry to select the necessary services from the SOA-based ecosystem and create viable big data solutions. A big data service-oriented architecture for oil & gas industries have proposed. The SOA enables finding, managing, visualizing, and understanding all traditional and big data to be represented as one entity to enhance decision making through many exercises. The study found that oil & gas companies can choose the best suitable service for their needs since service providers are loosely-coupled. Since each organization is unique, solutions are tailored for individual organizations by providing the architecture as services. This study shows that SOA realizes many advantages for oil & gas companies including increased agility, improved workflows, extensible architecture, enhanced reuse, and a longer application life cycle.

Tankimovich (2018) has investigated Big data in the oil and gas industry: a promising courtship to show the impact of big data in the oil and gas industry, as well as other industries that already start to apply big data in their career. The study implemented a supervised learning model to show the high level of accuracy for predictions of well's oil production, based on machine learning through linear regression using Python. The research found that the predicted results trained by Python were close to the real data of well productions and that lead to the understanding that machine learning can be an effective tool for handling big data in future related works. Also, oil and gas companies must encourage big data analytics as a part of their business to help in the processing process and make use of the data to find and produce more oil and gas with less cost and in environmentally friendly ways. This study examines big data as an opportunity for oil and gas companies to increase their benefits through technological advancement.

Rawat (2014) has investigated Big Data analytics in the oil and gas industry to review and suggest some ways that enable oil and gas companies to gather, understand, and analyze their data. The study discussed the various technologies used to capture data in the oil and gas industry and its benefits such as smarter sensors, optical fibers for wells and pipelines, Permanent Reservoir Monitoring (PRM), microseismic technology, and SCADA. Moreover, the study examined the purpose of Big Data in different parts of the oil and gas industry followed by the challenges coming with big data. The study found that big data can really lead to a competitive advantage in this sector if the stakeholders invest in the appropriate technologies that support their big data initiatives. This study showed some of the technologies used to capture the data and their benefits along with the big data initiative to increase the competitive advantage in this sector.

Sofi and Perez (2014) have investigated how innovative oil and gas companies are using big data to outmaneuver the competition and published as Microsoft white paper. This paper has two goals, first help the leading companies in Oil and Gas industry to understand why big data is important to their business, second to show how some of these companies are already involving in big data projects to increase their market advantage. Also, it provided guidance on how to use big data to gain valuable insight and enhance the decision making process. The study discussed some of the problems facing this industry, as well as the need to use big data technology to drive insight from this large amount of data by using appropriate technology, tools, and expertise. A Microsoft Upstream Reference Architecture (MURA) framework was developed which has been created to meet the needs of the O&G industry. İt works as a guide for IT and management teams to plan business systems and deploy software and solutions that can best realize the value of big data. Some recommendation steps have been outlined in this paper to show how oil and gas firms can plan and execute big data initiatives. The study found that by using big data, oil and gas companies can enhance their business value and gain competitive advantage and in order to make this true, they must create a new strategy that makes them manipulate this data to support smarter decision making.

Perrons and Jensen (2015) have investigated Data as an asset: What the oil and gas sector can learn from other industries about "Big Data" to show how oil and gas companies could extract more value from data. They examined the important role

behind big data in shaping the new era of technology and what value it could add to the upstream oil and gas industry and compared them to the practices done by leading companies using big data in other sectors. The study found that the digital revolution will be completed when the sector discovers how to find knowledge out of the data being collected. This study outlined the high level capabilities and new technologies that have given rise to big data, and briefly examined the benefits of these changes to the oil and gas sector.

Aliguliyev and Imamverdiyev (2017) have investigated big data strategy for the oil and gas industry: general directions to develop Big Data strategy for the oil and gas industry and to analyze the potential benefits of big data technology in oil and gas industry, big data application in oil and gas industry and the potential issues in data management. Also, provide the general principles for implementation of Big Data strategy. The study analyzed the potentials, challenges, and trends associated with big data in the oil and gas industry, as well as provided recommendations for the implementations of big data strategy. The study found that big data can be important when implementing new strategies for oil and gas companies. This study provides some trends to address problems regarding the adoption of big data, and the potentials and applications of big data oil and gas industry.

Poor & Torabi (2018) have investigated Big Data analytics in the oil and gas industry as an emerging trend. This study reviews the needs of the oil and gas industry for big data analytics in both the upstream and downstream operations. The study was done by an extensive review of the recent papers about the application of Big Data analytics in both the upstream and downstream oil and gas industry. In the first part of the study, big data is defined and the processing tools are introduced. Secondly, the potentials of big data analytics in the oil and gas sector have reviewed. Finally, the challenges facing big data in the oil and gas industry have been addressed. The research found that big data concept attracted to oil and gas companies because of the necessity to improve efficiency through the industry operations and lifecycles. This study examines the impact of applying big data technologıes in both the upstream and downstream oil and gas industry.

Hassani and Silva (2018) have investigated Big Data as a big opportunity for the petroleum and petrochemical industry to study the influence of the industry's engagement in utilizing big data analytics in upstream, midstream, and downstream sectors. Summarizing the various applications and advantages of using big data in the upstream sector (exploration and discovery, drilling, and production), midstream and downstream sectors have done. The study found that the oil and gas upstream sector is highly engaging with big data to boost the efficiency of its operations while midstream and downstream sectors still lagging behind. The upstream sector is considered as a positive impact not only to midstream and downstream sectors but the entire industry. The big data now encourages innovations in sensors and data-related technologies which motivate the vast growth of this concept in the future. This study examines, in particular, the influence of big data on the P&P industry.

Song (2018) has investigated research on the application of Big Data in oil and gas industry to help oil and gas companies provide the suggestion of big data blueprint and construction for oil & gas companies which can help the industry to recognize the value of the unexploited data and to transform decision-making from reactive to proactive. The study reviewed some potentials of big data applications in the oil & gas industry, and involved the key phases of exploration and development, production, engineering, and management. Also, the study analyzed the best practice of big data in some leading oil & gas companies, including Shell, BP, and Statoil. Finally based on this study, advice for deploying big data projects has stated. The study found that the oil and gas industry must improve not only technological, but production and information and communication processes as well to meet the long term growth in demand for energy resources. This study examines the deploying of a framework for oil and gas companies to apply big data step by step.

Baaziz and Quoniam (2014) have investigated how to use Big Data technologies to optimize operations in the upstream petroleum industry. The purpose of this research is to show how the oil and gas industry can benefit from big data to gain valuable insight and support the decision making process in different activities of the industry. The study reviewed the advantages of using big data technologıes in different stages of the oil and gas industry upstream process (exploration & development, drilling & completion,

production, Equipment maintenance, Reservoir Engineering, Research & Development, Data Management, Security, health, safety& Environment). The study found that oil and gas companies must mark their requirements of technology and data management expert staff by establishing gap analysis to their business allowing for investment in both proven technologies and those who will face growing volumes of data. Oil and Gas companies must adopt new technologies and strategies that help both experts and managers in their business and decision making process. This study examines the impact of the adoption of big data technologies that helps oil & gas companies to track new business opportunities, reduce costs, and reorganize operations.

4. DRILLING OPERATIONS IN OIL AND GAS INDUSTRY

4.1 INTRODUCTION

Drilling processes considered as complex and costly operations in the oil and gas industry, the cost of drilling forms half of well expenditure (Holdaway 2014). However, only 42 percent of drilling time is assigned to drilling itself, while the rest of the time is divided to deal with problems, rig movement, defects, and latency periods. Therefore, in oil and gas well drilling, service providers have been continuously working to increase the efficiency of their drilling programs and reduce drilling costs to operating companies. Also, predicting future problems such as equipment failure and stuck pipe which can lead to unexpected shutdowns and boosting nonproductive time (NPT), as a result, this can have a great impact on enhancing the drilling performance. The analytical models came as a solution for different aspects for example it plays a role as an early warning system, leading to avoiding events that cause NPT during the drilling process and can help in the elimination of other problems.

The digitalization of the oil and gas industry representing by utilizing embedded sensors and IoT has increased the amount of data collected during the drilling process. Despite the improvement of this industry, most of the data collected still underutilized. Nowadays, as the development of the data science approach in the energy sector and other sectors, service companies start to see it's significant value in processing the data in the rig site. However, the data is not the problem anymore because it's available, what matters is the ability to extract wisdom from it. Structured and unstructured data could be used together to optimize decision making in the oil and gas industry. For example, by comparing the real time data inputs with patterns from historical database through analytics or visualization techniques, companies can identify issues and root causes in massive volumes of information, and then identify and implement actions that will treat the cause upon detecting the pattern (Holdaway 2014).

4.2 DATA SOURCES İN DRİLLİNG OPERATİONS

Recent innovations in drilling systems and rig instrumentation have pushed oil rigs to turn into massive data sources which are useful for engineers to optimize operations. Oil and gas companies utilize many sensors to provide continuous collecting of drilling parameters from every equipment installed in the drilling rig.

In the upstream sector, supervisory control and data acquisition (SCADA) used to capture oilfield data such as drilling information, production monitoring, surface, and subsurface facilities, another benefit offered by SCADA is providing this data from remote oil and gas sites without the need for personnel visits.

Drilling information comes from downhole sensors that are placed near the bit, bottom hole assembly, or drill string with the sensors recording data in memory mode or transmitting the data in real time through various forms of telemetry (Evans 2014). For example, tools such as logging while drilling (LWD) and measurement while drilling (MWD) are able to gather data in real time and transfer it to the surface. There are other different sensors installed in the rig site for sensing and monitoring different drilling parameters such as weight on bit(WOB), rotation per minute (RPM), flowrate, temperature, pressure and so on, they are providing a wide range of information every day during the drilling process. Companies started to store this information in their databases and analyzed them to discover the relationship between them and take appropriate actions. Another source of information about drilling is the daily reports which contain different datasets such as activity breakdown, bit information, geology information, drill string, mud information, etc. This information will be sent from the drilling site to the office where it will be stored there in the company’s database and used for different purposes.

4.3 WHY BIG DATA ANALYTICS IS USEFUL?

Data and analytics start to play a key role in different industries, the oil and gas industry in general and drilling industry in specific is taking the same road. For oil and gas companies data and analytics provide a better understanding of operations, so that they can predict and solve a wide range of problems such as improve operational efficiency, develop new strategies, and planning and performance.

Analytics ideal for decision making in the drilling process, it enables engineers making real time decision depending on formation geology and drilling parameters for conducting predictive modeling. Analytics also improve drilling accuracy by identifying issues that have a negative impact on operations and early detection of equipment failure, so experts will be able to predict when maintenance will happen and whether there will be shut down or not to prevent large environmental risks. Unexpected equipment failure can be very expensive and Big Data from sensors in equipment when combined with geological data can enable oil and gas companies to predict failure and understand which equipment works best in which environment (ESDS 2016). Moreover, analyzing of 2D, 3D, and 4D Seismic data imaging by parallel processing of big data analytics platforms help in predicting drilling operation success (Baaziz & Quoniam 2014, p.1). Real-time data on weather can be combined with drilling operation data to avoid dangerous conditions for workers and mitigate environmental risk (ESDS 2016).

Devon Energy in the USA is utilizing big data analytics concept by combining Hadoop, SAS, and text data, they achieved 30 percent of nonproductive time reduction by determining the cause of NPT and address the issue (Hassani & Silva 2018, p.81). Also, IBM combined engineers' expertise and physical based models of wells to build a predictive drilling model to predict the likelihood of stuck pipe occurrence.

4.4 DRİLLİNG OPTİMİZATİON AND DRİLLİNG PARAMETERS

As known most of the well cost is related to drilling operations; therefore, companies are working to avoid drilling difficulties and improve drilling processes to minimize the overall costs. Drilling costs depend on drilling speed, so raising the speed at which the well can be drilled accomplish costs reduction, this referred to as the increasing rate of penetration(ROP). ROP means how fast the drill bit is drilling through the formations, it represents the speed of drilling bit when breaking the rocks (Bourgoyne et al. 1991).

ROP has complex nature as it affected by large numbers of interrelated parameters and it is hard to predict the influence of single parameter on the other because they are depending on each other and because of that until now no certain mathematical relationship between ROP and drilling parameters have found (Eren 2010, p.2). In order to achieve the goal of lowering the operational cost of drilling, those parameters should be optimally controlled with no drilling problems.

Factors affecting ROP are classified into controllable and uncontrollable factors. Controllable factors are the factors that can be changed manually such as WOB, RPM, and mud Flowrate. Uncontrollable factors also called environmental factors that can not be changed such as drilling fluids and formation properties, nevertheless, it is important to know that formation properties are considered as a critical factor in determining drilling performance (Eren 2010, pp. 2-11). Some factors such as fluid properties and bit types although they are controllable are hard to change in ordinary bit runs (Eren 2010, p.2).

A brief description of the factors that most affect ROP and considered in this study will be provided :

WOB: it's the amount of weight applied on the bit in kilo-pound (klb) which is transformed into the formation causing to break it during the drilling phase.

RPM: this means revolutions per minute which is the rotational speed of the drill string.

Mud weight (MW): it is the density of mud in specific gravity. It serves as the primary control of the well preventing formation fluids from entering the wellbore. Any increase in density leads to a decrease in the rate of penetration.

Flow rate: it is the volume of fluid that will be pumped through drilling bit.

Bit wear: it's a term of bit age during the drilling phase, normally it's decreasing while going deeper because of cutter erosion. Drilling torque is a good indication for a bit wearing.

Torque: torque is generated when applying a load and rotating the drill pipe, and can be measured through Top Drive System (TDS). Increasing Torque causes ROP to increase.

Standpipe pressure (SPP): it is the pressure drop due to fluid friction.

4.5 CASE IN IMPROVING DRILLING OPERATION

As an example of a data driven analytics approach to enhance drilling program efficiency, the case of SAS institute was considered. SAS developed a platform for predictive analytics solutions, so companies can take advantage of near real time predictive models.

SAS software can extract the relationship between the data and use them to improve drilling decisions. It can offer different benefits to enhance drilling operations such as identifying the efficient KPI by finding the optimal relationship between data from drilling incidents and key performance indicators. Applying analytics to data from whole drilling operations can reduce NPT, also it provides visualization of the data for the evaluation of drilling performance.

SAS software tested on a customer wanted to find the best combination of drilling parameters to increase the efficiency of the drilling process and ROP, also minimize the drilling costs. The analytics solution can analyze some controllable drilling parameters such as WOB, RPM, slow pump pressure, drill bit selection, downhole assembly, rig type, LWD, MWD, and driller profiles and experience. After the analysis, a recommendation can be provided for the optimal range of values for drilling parameters leading to enhanced ROP value. The results showed improvement in ROP, NPT was reduced by 18 percent, and the foot per day was increased (SAS Institute 2015).

There many other companies try to develop and use the analytics approach and advanced technologies to solve their business problems, enhance their operations, and competitive advantage. In the next chapter, a simple personal case has done to show how to apply the big data analytics technologies to improve the drilling efficiency by building a model to predict ROP values.

5. DEVELOPING ANN TO PREDICT ROP

5.1 INTRODUCTION

As mentioned before optimization of ROP has a great impact in decreasing drilling cost and efficiently enhance the drilling operation, however this process still considered as a big challenge for oil and gas industry because of the high level of uncertainty regarding the geological structures, a large number of uncontrolled factors and their interrelated relationship. The use of big data from the drilling process to build a predictive model can achieve this purpose. The artificial neural network (ANN) approach considered one of the effective analytical tools used in drilling operations to develop learning algorithms and extract the value of the data used in the process. ANNs are essential tools used in modeling complex systems that seek to simulate human brain behavior by processing data during a trial-and-error basis (Wang & Salehi 2015, p.3). From input parameters, they can generate a predictive model that correlates to the output parameter.

A number of researches have been conducted using ANN to optimize ROP and drilling efficiency. Ahmed et al (2019) have investigated new artificial neural networks model for predicting ROP in deep shale formation, the research aimed to use the ANN technique to build a new model that can predict ROP in the shale formation. Both drilling parameters such as RPM, WOB, Torque, SPP and flow pump, and mud properties such as MW, funnel and plastic viscosities, solid and yield point were used to build the model using a data from one well and one formation. The model achieved a correlation coefficient (R) of 0.996 and an average absolute percentage error (AAPE) of 5.776 percent.

Mnati and Hadi (2018) have conducted prediction of ROP and cost with artificial neural network for Alhafaya oil field which aims to develop a model using ANN to predict ROP and cost of drilling operations. The drilling parameters were used in the project were WOB, RPM, hydraulics (HIS), and travel transit time (DT) from mud logging unit and wireline log for Alhalfaya oilfield to design ANN model that consist of two hidden layers and one output layer. Five datasets of five formations were involved in the

model. The model showed good accuracy of ROP when comparing with the raw data of the five formations and a good fitness of drilling cost comparing with the actual cost.

Shadizadeh et al (2010) used ANN models to predict the stuck pipe occurrence in Iranian oil fields. Stuck pipe instances were divided into dynamic and static with different parameters included in the model for the analysis. The final network consists of three layers with 80% of the data used for training, 10% for validation, and 10% for testing. The model achieved more than 90% of the stuck pipe prediction.

Wang and Salehi (2016) used ANN to optimize drilling hydraulics using real time field data. Three layers feedforward network with backpropagation was developed and the forward regression was used for the sensitivity analysis of the input parameters. A combination of 12 parameters included in the model to predict the pump pressure using data from three wells in the same geological area. The result showed good agreement between the developed model and actual field data.

Those researches showed the effectivity of adopting the ANN approach to solve different problems in the drilling industry. İn this chapter ANN was developed to use the concept of big data generated in drilling operation to produce a model that can predict ROP efficiently using data from different wells and larger than the data used in the previous researches.

5.2 AIM OF THIS RESEARCH

The similarity of drilling circumstances for wells drilled at the same locations and have the same geological structures made it possible to collect the past data and utilize it in the optimization process of ROP and providing critical drilling parameters for different formations.

The main intention of this research is to show the capability of big data generated in the drilling industry to shape the future of this industry, analyzing this data makes it possible to solve and optimize different issues regarding drilling sector. As mentioned in the previous sections about the potentials of adopting data analytics techniques and how useful it could be, this section provides dealing with large data to solve one of the real problems in the drilling operation. The historical data from offset wells located in

the same area could be used in predicting an optimized value of ROP and serve in preparing optimized drilling plans for future wells of the same area. High ROP can lead to different problems such as stuck pipe and poor hole cleaning; therefore, it is critical to find the best combination of drilling parameters that cause no drilling problems.

In this section, ANN technique used as the analytical tool to analyze the drilling data that have been collected, ANN is known to deal with the complex relationships between the data and since the relationship between drilling parameters is complicated as mentioned before, it was useful to build the model with ANN. The generated model should be able to predict ROP accurately when comparing with actual data, and it will be compared with a real drilled well to show the accuracy of the resulted model. Python language and different python libraries were used to design and write the code of the artificial neural network. Seven parameters considered as input parameters that feed the ANN and correlate to the output parameter ROP.

The goal that was tried to serve in this work is using the six steps of applying big data analytics techniques in a real situation of increasing ROP by using data as much as possible to build a neural network model that can optimize ROP predictions based on actual data provided. İt serves as a framework for further future works in the same area using even more data.

5.3 THE RESEARCH FRAMEWORK

What has been accomplished in this research is reviewing the big data analytics concept and its potential to shape the future of the oil and gas industry by increasing the efficiency of operations in multiple areas of the industry. This wasn't the only part, a further approach was using with the drilling operations to detect the possibility of using the big data generated by this operation to enhance what is called the rate of penetration which is one of the important parameters to improve the drilling process as mentioned before.

All the amount of data collected is from 45 wells which are approximately 117 K for eight parameters including the ROP parameter, which means the data volume was a limitation for this project and it could affect the analysis process. For big data projects providing data as much as possible could result in more accurate results while analyzing

the data to find new patterns. Hence, Hadoop software has not been used during this project because the volume of the data was not too big and could be handled by only using Python and ANN.

A model that can predict ROP values based on the actual values from the actual wells using python coding language and its libraries (Numpy, Pandas, Scikit Learn, Matplotlib, Seaborn and Keras) with the Artificial Neural Network as a tool for analyzing the data was developed. Before building the model, preprocessing methods were included which are known to increase the quality of the data used. The preprocessing steps included the elimination of missing values, detecting and removing outliers, normalization or scaling of the data, and splitting the data for training the network into training and testing sets. However, any feature selection method did not include in the project because of the number of limited features used in this research coming from the data at hand. Also, any other tools for processing and analyzing the data were not used except for Python and ANN. Even though corrections of features used for building the ANN can provide better predictions and accurate results such as correction of weight on bit applied considering the inclination and additional bit rotation generated by the motor in deviated wells, but the additional information used to do the correction process was not available with the data used in this research.

5.4 THE DATASET AND DRILLING CONDITIONS

In this project, a dataset of drilling parameters from 45 actual wells located in North and South Rumaila oilfield in Basra in the South of Iraq was collected and used to train the model, information about Rumaila oilfield is shown in Appendix A. All the wells were drilled in similar geological areas. The sensors introduced at different equipment in the rig site used to capture and transmit a wide range of drilling data. The Mud Logging unit (MLU) in the site is responsible for collecting and transmitting the drilling data by means of sensors placed at various equipment. MLU measures drilling parameters and records mud cuttings properties, as well as on site and remote locations drilling monitoring services (Eren 2010, p.8). All the gathered drilling parameters have a great impact on the optimization process, the main thing to consider is the accuracy of this data which mainly affects the results of the optimization process. Recording drilling parameters from sensors should be calibrated and measured correctly to ensure the

success of drilling optimization. The data contains seven input parameters (TVD, WOB, RPM, Torque, SPP, Flow In, Density of Mud) and ROP as the only output parameter as shown in figure 5.1 below.

Figure 5.1: Drilling Dataset

	A	B	C	D	E	F	G	H	I
1	TVD(m)	WOB Avg(klb)	RPM Total Avg(	Torque Abs Avg	SPP Avg(psi)	Flow In Pum	Dens Mud In	ROP Avg(m/hr)	
2	42	2	50	1088	413	1947	1.05	11.1	
3	43	2	50	1088	413	1947	1.05	10.1	
4	44	2	50	1088	413	1947	1.05	11.1	
5	45	3	50	876	415	1950	1.05	10.3	
6	46	3	50	872	418	1950	1.05	10.4	
7	47	3	50	863	417	1952	1.05	12.5	
8	48	4	50	832	417	1953	1.05	8.7	
9	49	3	50	845	417	1954	1.05	10.3	
10	50	4	50	848	417	1954	1.05	10.3	
11	51	4	50	841	419	1954	1.05	10.6	
12	52	4	50	865	481	2081	1.05	10.8	
13	53	4	50	929	498	2133	1.05	10.8	
14	54	4	50	942	531	2207	1.05	10.1	
15	55	4	50	957	528	2207	1.05	10.1	
16	56	4	50	987	572	2275	1.05	10.2	
17	57	1	52	1163	721	2489	1.05	10.1	
18	58	1	52	1907	730	2506	1.05	24.9	
19	59	1	52	1700	733	2511	1.05	25.1	
20	60	1	52	1267	733	2511	1.05	25.3	
21	61	2	52	1788	736	2509	1.05	26.1	
22	62	2	52	1583	736	2512	1.05	25.8	
23	63	1	52	1173	733	2512	1.05	26.1	
24	64	1	52	1548	761	2499	1.05	25.9	
25	65	1	52	1348	755	2502	1.05	25.8	

Sayfa1 Sayfa2 Sayfa3

During the project, the formations from the surface to downhole were treated with similar lithology and a model was developed for this case using a total number of 117591 datasets. The formations included in this research are shown in Appendix A-figure-3. Some drilling conditions during the activity of drilling were assumed for the relations of the drilling parameters to be effective :

a- All formations from the surface to drilling target have the same lithology.
b- No bit wearing while drilling.
c- No downhole problems while drilling.
d- The bottom hole is clean.

5.4 DATA ANALYSIS METHODOLOGY

Countless analytical techniques and coding languages can be used on big data to achieve the ultimate goal of processing and handling big data effectively. An important tool used in Data Science is Python which is a programming language contain multiple libraries that provide effective features for data analysis purposes. A description will be provided about some libraries used during the project in the next sections.

5.4.1 Python

Numpy which stands for numerical python is a package for numerical computations, deals with multi dimensional arrays used for processing multidimensional data, and supports for different data types. Since the first step to make the data ready for the analysis process is changing it into arrays, Numpy considers as the specialized package used to deal with those numerical arrays and offer high effective manipulation and storage capability of data. There are several problems with Numpy such as its lack of flexibility when working with missing data, attach labels to data, etc (Tankimovich 2018, p.21).

Pandas is a library that offers functions for data wrangling and manipulation. Unlike Numpy, rows and columns are labeled, so it considered an enhanced version of Numpy. It offers multiple data structures such as ‘DataFrame’, 'Series', and 'Index'(Tankimovich 2018, p.21).

Matplotlib and Seaborn, both of them are plotting libraries used for visualization of data to identify trends and relationships between the data.

Scikit learn is a library used with Python for machine learning and contains different machine learning algorithms for supervised and unsupervised learning. It focuses on modeling the data and contains a range of models for different purposes such as clustering, cross validation, feature extraction, feature selection, etc.

Keras is a python library on the top of TensorFlow used for developing and evaluating neural networks for both convolutional networks and recurrent networks. It used to create and train deep learning models with a simple code.

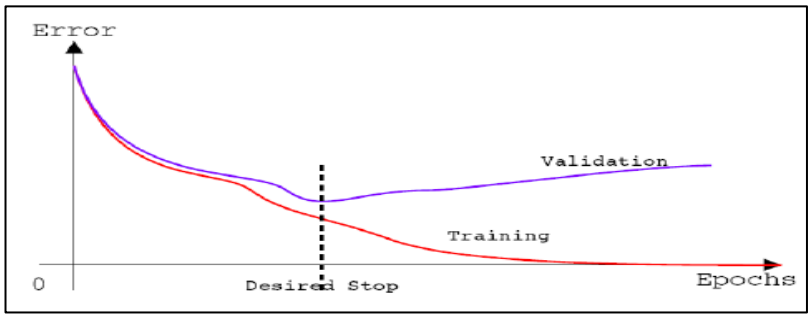
All of the mentioned libraries were used in the project for preprocessing and analyzing of the data. The technique used for analyzing the data in this project is the Neural Network which will be described below.

5.4.2 Neural Network

Neural Network is a machine learning algorithm that works in parallel and distributive fashion to analyze data efficiently with the ability to discover complex relationships between variables that introduced to the network. Neural Network is made up of many artificial neurons which considered as the processing units and their connections which called weights. Weight is simply a floating point number and it's these adjusted when eventually coming to train the network. Each input to the neuron has it’s own weight associated with it, so when the input enters the neuron it will multiplies by its weight, then the neuron sum up all these new input values which gives the activation. All the logical components of NN and it’s working mechanism is shown in appendix B.

ANN consists of an input layer, one or multiple hidden layers, and the output layer. The input layer is for receiving the data, and the number of neurons in this layer corresponds to the number of input parameters presented to the network. The hidden layers are for developing relationships between the input parameters and finally, the output layer forms the results. The number of neurons used depends on the task at hand and should be optimized to find the appropriate number of layers and neurons to avoid overfitting and underfitting problems that could happen during the training phase of the network. A large number of layers and neurons result in increasing the processing time and decrease the generalization of the model by decreasing the training error while the testing error remains high and this case called memorization which resulted in overfitting of the model (Mohaghegh 1994, p.2). Regularization can be used to solve this issue and a common method of regularization is the early stopping which stops the model at a certain point when the validation or testing error starts to increase, whereas the training error decrease. Early stopping provide the model with the number of iterations needed to be run without overfitting, figure 5.2 shows the meaning of this technique. The iterations called epochs which is the number of times that all the data has been cycled through.

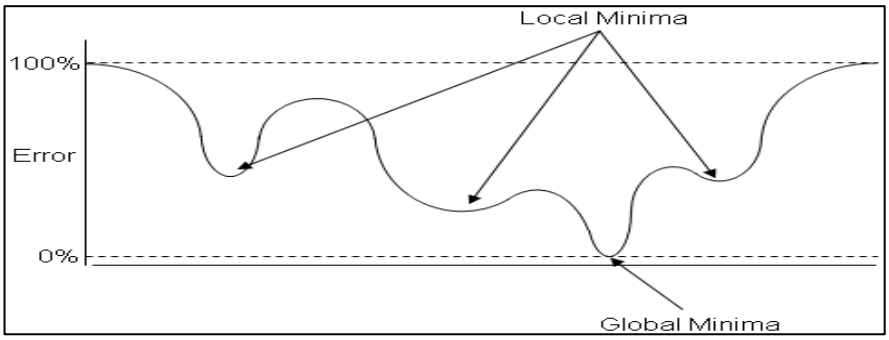
Figure 5.2: Early stopping technique



Source: researchgate.com

There are several kinds of ANN; Radian Basis Function NN (RBFNN), Self Organized Map NN (SOM), Recurrent NN (RNN), Convolutional NN (CNN), Modular NN (MNN), and the most famous one which is used in this study is the Feed Forward NN with Back Propagation (FNN). In the Forward Propagation, initial values of the weights and biases will be considered, then those weights feed to the activation function to calculate the output of each layer and finally the expected value of output layer which in turn will be used to calculate the error of neural network by comparing with actual value. Whereas, the backpropagation used to adjust the weights and biases to find the appropriate values of them that cause minimum global error value. The global minimum is the lowest possible error achieved through the training phase. Another term that could be faced when training the model is the local minima which happens when the network finds an error which is less than the surrounding errors but actually, it's not the ultimate possible error, figure 5.3 below shows this concept.

Figure 5.3: Local and global minima



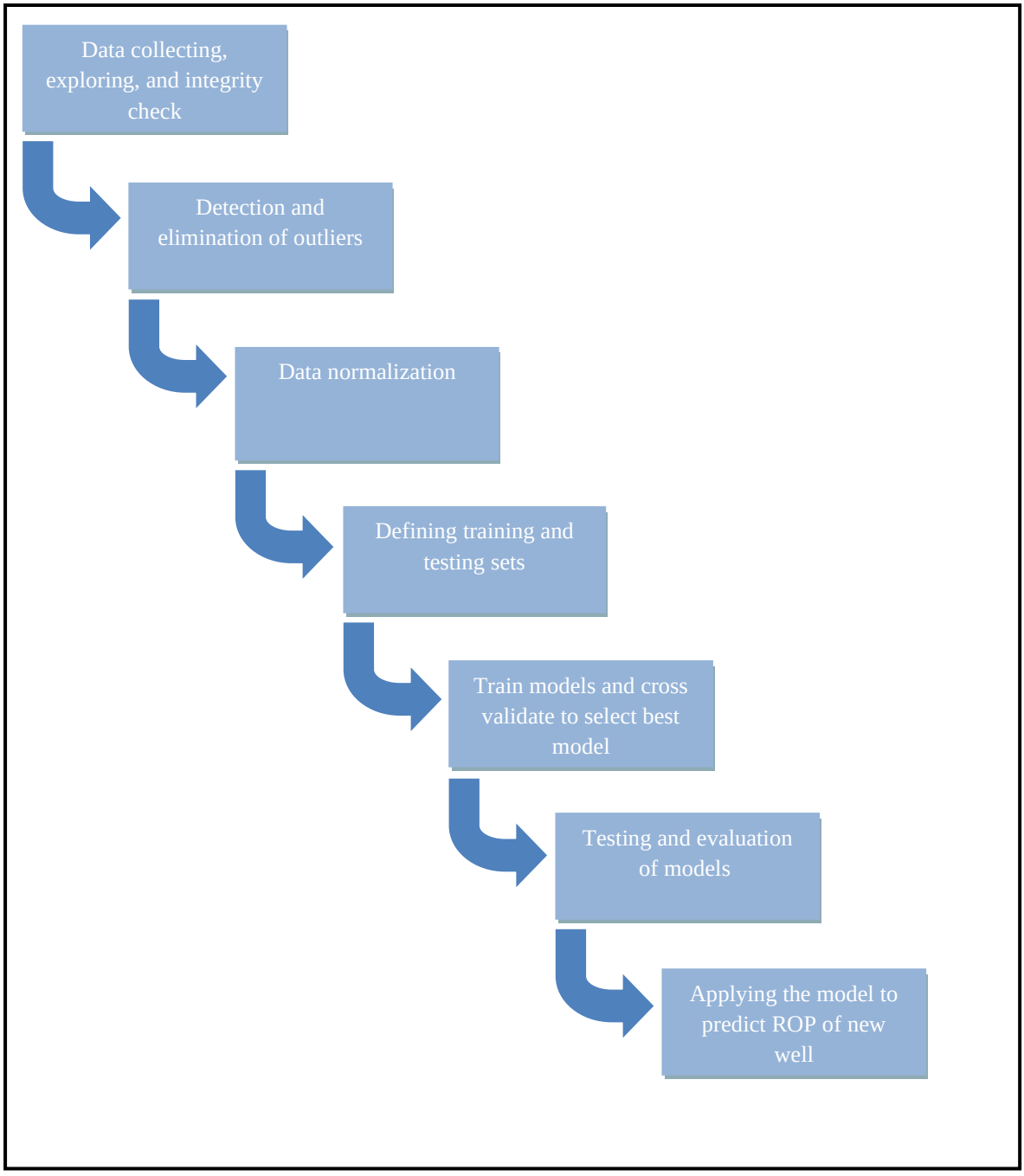
Source: mnemstudio.org

In the Neural Network, the data is divided into three sets; training, validation, and testing sets. The training set is used to set the values of weights and biases to calibrate the model. In the validation set the hyperparameters such as the number of neurons and layers, learning rate, numbers of iterations, activation function and learning alpha rate will be applied to the model to make sure of the generalization of the implemented model in the training phase (Ahmed et al. 2019, p.3). The testing set used to check the performance of the developed Neural Network with the adjusted weights, biases, and hyperparameters coming from training and validation sets. There is no accurate approach to choose the right amount of data for each of the three sets, but it is recommended to increase the amount of data in the training set when working with big data. Providing the Neural Network with data as much as possible increases the accuracy of the generated model as the ANN will be trained on different and more examples of the situation to be solved, hence increasing the performance of ANN.

5.5 NEURAL NETWORK DESIGN

As mentioned before the model developed using python and its libraries. Before data enters the neural network, there are preprocessing steps that applied to ensure the accuracy of the model generated. All the steps of work are described below step by step until reaching the final goal of getting acceptable predicted values of ROP when compared with new unseen values of ROP from a new well. Google Colaboratory (Colab) was used as an integrated development environment to write the code. Colab is a free cloud platform that can be accessed easily on the web and start to write the code in a notebook. Colab provides a GPU accelerated virtual machine to accelerate computational processes when developing a project. The process flow chart is shown below to describe the steps used to generate the model, then in the following section, the steps will be explained.

Figure 5.4: Developing ANN model workflow



5.5.1 Data Preprocessing

Raw data contains some errors and could be incomplete, inconsistent as well as having missing values and outliers which cause the low quality of data, hence affecting the performance of neural networks if not eliminated or corrected. Therefore, data preprocessing is crucial to solving those issues. Missing values could be happened through the data collection process, or due to errors in measurement devices. They could be eliminated directly or handled using some methods such as filling missing data using the mean, median, or mode value of the respective feature.

The data used in the project contains some negative values (less than zero) and Zero values which is not suitable to include in the model because they could decrease the quality of the model; therefore, changing them into null values and eliminate them with the missing data together was done. Figure 5.5 shows a simple statistical analysis for the data. Figure 5.6 shows the detection and elimination of negative and zero values after changing them into not a number (NaN) values.

Figure 5.5: Description of dataset

```
[ ] dataset.describe()
```

	TVD(m)	WOB Avg(klb)	RPM Total Avg(rpm)	Torque Abs Avg(f-p)	SPP Avg(psi)	Flow In Pum Avg(lpm)	Dens Mud In Avg(sg)	ROP Avg(m/hr)
count	117114.000000	117590.000000	117590.000000	117590.000000	117590.000000	117590.000000	117590.000000	117590.000000
mean	1622.773324	10.976193	160.841305	8271.863755	2241.244876	2381.092899	1.148266	24.932608
std	917.724929	5.661867	63.905074	4744.521583	789.800070	429.100477	0.069286	21.637636
min	0.000000	-2.410000	0.000000	0.000000	0.000000	0.000000	1.010000	-9.550000
25%	847.800000	6.710000	99.000000	4718.000000	1667.000000	2051.000000	1.080000	12.600000
50%	1607.480000	10.700000	160.000000	8553.500000	2335.000000	2266.000000	1.130000	21.200000
75%	2346.205000	14.970000	225.000000	11500.750000	2872.000000	2812.000000	1.220000	29.830000
max	3411.700000	42.000000	335.000000	24245.000000	4149.000000	4132.000000	1.350000	425.200000

Figure 5.6: Elimination of negative and zero values with NaN

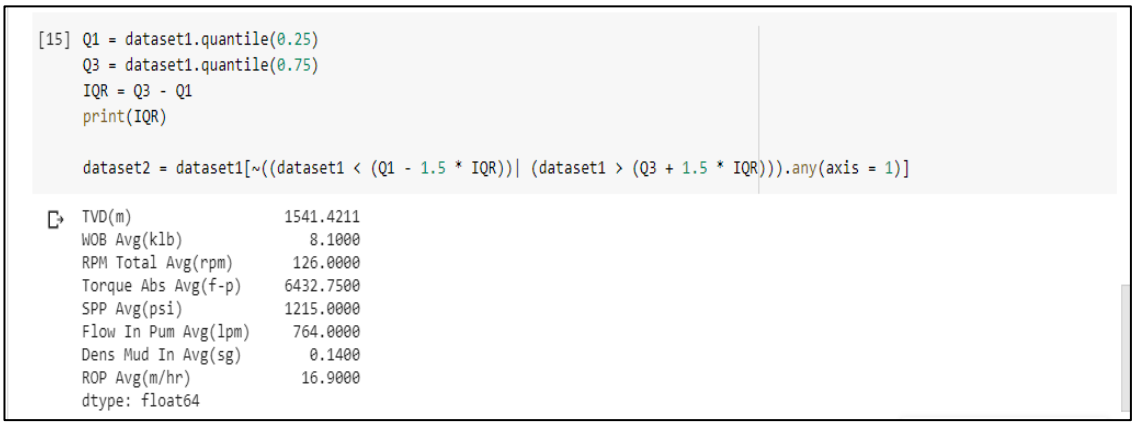
```
[5]  dataset[dataset<0] = np.nan

[6]  dataset['TVD(m)'].replace(0, np.nan, inplace= True)
     dataset['WOB Avg(klb)'].replace(0, np.nan, inplace= True)
     dataset['RPM Total Avg(rpm)'].replace(0, np.nan, inplace= True)
     dataset['Torque Abs Avg(f-p)'].replace(0, np.nan, inplace= True)
     dataset['SPP Avg(psi)'].replace(0, np.nan, inplace= True)
     dataset['Flow In Pum Avg(lpm)'].replace(0, np.nan, inplace= True)
     dataset['Dens Mud In Avg(sg)'].replace(0, np.nan, inplace= True)
     dataset['ROP Avg(m/hr)'].replace(0, np.nan, inplace= True)

[7]  dataset.isnull().sum()

     TVD(m)                  478
     WOB Avg(klb)            123
     RPM Total Avg(rpm)      230
     Torque Abs Avg(f-p)    3507
     SPP Avg(psi)              1
     Flow In Pum Avg(lpm)      3
     Dens Mud In Avg(sg)       0
     ROP Avg(m/hr)             6
     dtype: int64

[8]  dataset1 = dataset.dropna()

[9]  dataset1.isnull().sum()

     TVD(m)                  0
     WOB Avg(klb)            0
     RPM Total Avg(rpm)      0
     Torque Abs Avg(f-p)     0
     SPP Avg(psi)            0
     Flow In Pum Avg(lpm)    0
     Dens Mud In Avg(sg)     0
     ROP Avg(m/hr)           0
     dtype: int64
```

Another concept that should be handled before building the model is the outliers because they are the main source of errors in prediction models (Ahmed et al. 2019, p.6). An outlier is a data point that lies far from the rest of the other data points of the same feature. The common methods to identify outliers are Scatter plots, Box plot, Z-Score, and Interquartile range IQR. Figure 5.7 shows the code used to detect the presence of outliers, and figure 5.8 shows the code used to eliminate the outliers.

Figure 5.7: Outliers detection

```
[12] def find_outliers_tukey(dataset1) :
         q1 = np.percentile(dataset1,25)
         q3 = np.percentile(dataset1,75)
         iqr = q3 - q1
         lower_quartile = q1 - 1.5 * iqr
         upper_quartile = q3 + 1.5*iqr
         outlier_indicies = list(dataset1.index[(dataset1 < lower_quartile) | (dataset1 > upper_quartile)])
         outlier_values = list(dataset1[outlier_indicies])

         return outlier_indicies, outlier_values

[13] tukey_indicies, tukey_values = find_outliers_tukey(dataset1['ROP Avg(m/hr)'])
     print(np.sort(tukey_values))

     [ 55.5  55.5  55.5 ... 387.9 388.9 389.4]
```

Figure 5.8: Elimination of outliers with IQR



```
[15] Q1 = dataset1.quantile(0.25)
     Q3 = dataset1.quantile(0.75)
     IQR = Q3 - Q1
     print(IQR)

     dataset2 = dataset1[~((dataset1 < (Q1 - 1.5 * IQR))| (dataset1 > (Q3 + 1.5 * IQR))).any(axis = 1)]

     TVD(m)                1541.4211
     WOB Avg(klb)             8.1000
     RPM Total Avg(rpm)     126.0000
     Torque Abs Avg(f-p)   6432.7500
     SPP Avg(psi)          1215.0000
     Flow In Pum Avg(lpm)   764.0000
     Dens Mud In Avg(sg)      0.1400
     ROP Avg(m/hr)           16.9000
     dtype: float64
```

Data normalization consider as an important step for scaling the data into a specified range before starting the training process. This step helps to speed the learning process by finding the optimal weights faster. MinMax Scaler is the common method to scale the input parameters and output into the range of (0 to 1). The MinMax Scaler used the following equation to normalize the data :

$$X_{min\text{-}max} = X - X_{min} / (X_{max} - X_{min}) \quad (5.1)$$

$X_{min\text{-}max}$: the normalized value of a parameter

X: the original value of the parameter to be normalized

X_{max}: maximum of the original value

X_{min}: minimum of the original value

Applying the MinMax Scaler to the data using Scikit learn library is shown in figure 5.9 below.

Figure 5.9: Normalization of the dataset

```
[16] scaler = MinMaxScaler()

[17] dataset3 = scaler.fit_transform(dataset2)

[19] pd.DataFrame(dataset3).describe()
```

	0	1	2	3	4	5	6	7
count	106176.000000	106176.000000	106176.000000	106176.000000	106176.000000	106176.000000	106176.000000	106176.000000
mean	0.492748	0.410515	0.498570	0.409843	0.581736	0.523301	0.422004	0.385847
std	0.266056	0.204747	0.186251	0.213606	0.208410	0.151591	0.200565	0.193164
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.271625	0.257553	0.347305	0.254467	0.429730	0.411677	0.264706	0.226065
50%	0.493668	0.404937	0.500000	0.423312	0.615624	0.471551	0.352941	0.371714
75%	0.702122	0.552321	0.679641	0.552708	0.747989	0.659353	0.617647	0.513327
max	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000

Scikit learn library used to split the data into training and testing sets to use in the analytical process. The code of this method is seen in figure 5.10 below.

Figure 5.10: Splitting of the dataset

```
[25] x = dataset3[:, 0:7]
     y = dataset3[:, 7]

[26] from sklearn.model_selection import train_test_split

[27] x_train, x_test, y_train, y_test = train_test_split(x, y, test_size = 0.10, random_state = 0, shuffle = True)

[28] x_train.shape

     (95558, 7)

[29] y_train.shape

     (95558,)

[30] x_test.shape

     (10618, 7)

[31] y_test.shape

     (10618,)
```

No feature selection method was done through the project because the data used does not contain too many features, so it is preferred to use them all.

5.5.2 BUILDING AND TRAINING THE MODEL

Keras library and its tools used to build the model developed through the project. A model type in Keras called ‘Sequential' used to build the network layer by layer. Each layer is added using 'add()' function until finding the optimal number of layers and neurons. The process of finding the best structure of the neural network depends on trial and error, testing multiple structures, and watching the performance and error each time, then deciding on the best structure to use. ‘Dense’ is the layer type used to allow each neuron in the previous layer to connect with neurons in the current layer. The activation function has different kinds, Relu is found to be effective with input and hidden layers, while Sigmoid is preferred for the output layer because it keeps the output between 0 and 1.

Optimizer and loss function are other properties that should be considered when training the network. This is called compiling the model which takes the two parameters optimizer and loss function. The optimizer used to set the learning rate of the network during training, the learning rate means how fast the optimal values of weights could be found. A small learning rate means a longer time to calculate weights but it may lead to more accurate values of weights. The optimizer used in designing the network is 'Adam' which is a version of gradient descent, it is a good optimizer for different cases and shows good results for different problems. The loss function used is the mean squared error (MSE) which computes the average squared difference between predicted and actual values. The loss function is used to monitor the performance of the network while training for both validation and training sets.

After setting the parameters that needed to train the model, fitting the model to training and validation sets, as well as considering the number of epochs and callbacks of the model have done. The ‘fit()’ function allows to set the model to X and Y train sets, also it contains a validation split to decide on the percentage of the validation set to consider when training the model. EarlyStopping used to stop the epochs at a certain point when there is no improvement in the model.

‘Predict()’ function used to make predictions on test data after finishing the training process of the model.

To evaluate the accuracy of the developed model a statistical quality analysis called the Coefficient of Determination ($R^{2)}$ was used. R^2 used to measure the goodness of fit of predicted values from ANN with actual values and shows the strength of the relationship between them using a range of values of -1 to 1. When the value of R^2 is close to 1 this means there is a strong relationship, 0 means no relationship, while -1 means a reversed relationship between actual observations and predicted values by the model. all Keras tools used while developing the model is shown in figure 5.11.

Figure 5.11: Keras tools

```
[ ] from tensorflow import keras
    from keras.models import Sequential
    from keras.layers import Dense
    from keras.optimizers import Adam
    from keras.callbacks import EarlyStopping
    from sklearn.metrics import r2_score
```

A percentage of 70 % training set, 20% validation set, and 10% testing set (7:2:1) showed the best results. Multiple structures for each model run in order to find the optimal structure of ANN for the model which is shown in Appendix C – table-1 and table-2.

6. RESULTS AND DISCUSSION

6.1 RESULTS

While developing the model, different structures were tested the three layers feed forward neural network with backpropagation was selected as this network showed the minimum error. Also, the multiple numbers of neurons within each layer were tested to find the optimal number of neurons. With the selected distribution of data (7:2:1) for training, validation and testing set respectively, a comparison between different structures for each layer was tested and the neural network with 3 layers and 43 neurons for each hidden layer found to give the lowest error of the network. Choosing the activation function 'sigmoid' for hidden layers resulted in decreasing the error of the model; therefore, it was selected instead of Relu which is well known to give the best results in hidden layers. For the output layer, the sigmoid function also selected to scale the results between zero and one. 'Adam' method selected as the optimizer with the learning rate adjusted to 0.0055 which decreased the time needed to analyze the data and show good results compared with other values of learning rate. The number of epochs set to 7000 initially, but as mentioned before the early stopping used to stop the learning process at a certain number of epochs when the network stop to improve. Figure 6.1 below describes the final structure of the model used.

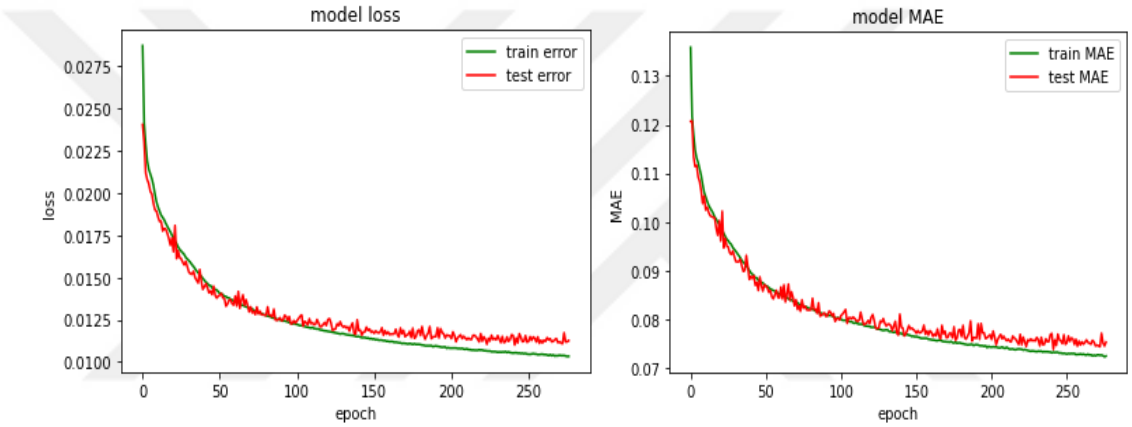
Figure 6.1: Final structure of the model

```
[ ]  #Define the model
     model = Sequential()
     model.add(Dense(43, input_shape = (7,), activation = 'sigmoid'))
     model.add(Dense(43, activation = 'sigmoid'))
     model.add(Dense(43, activation = 'sigmoid'))
     model.add(Dense(1, activation = 'sigmoid'))

     optimizer= keras.optimizers.Adam(learning_rate=0.0055)
     #Compiles model
     model.compile(optimizer= optimizer , loss = 'mean_squared_error', metrics = ['mean_absolute_error'])

     #pass several parameters to 'EarlyStopping'function and assign it to 'earlystopper'
     earlystopper = EarlyStopping(patience = 35)

     #Fits model
     history = model.fit(x_train, y_train, epochs = 7000, validation_split = 0.20, verbose = 1, callbacks = [earlystopper])
     history_dict = history.history

     76446/76446 [==============================] - 3s 45us/step - loss: 0.0105 - mean_absolute_error: 0.0730 - val_loss: 0.0
     Epoch 250/7000
     76446/76446 [==============================] - 3s 45us/step - loss: 0.0105 - mean_absolute_error: 0.0729 - val_loss: 0.0
     Epoch 251/7000
     76446/76446 [==============================] - 3s 45us/step - loss: 0.0105 - mean_absolute_error: 0.0728 - val_loss: 0.0
     Epoch 252/7000
     76446/76446 [==============================] - 3s 45us/step - loss: 0.0105 - mean_absolute_error: 0.0729 - val_loss: 0.0
     Epoch 253/7000
     76446/76446 [==============================] - 3s 45us/step - loss: 0.0105 - mean_absolute_error: 0.0729 - val_loss: 0.0
     Epoch 254/7000
     76446/76446 [==============================] - 3s 45us/step - loss: 0.0105 - mean_absolute_error: 0.0730 - val_loss: 0.0
```

After running the model the best performance found after 277 epochs. The mean squared error which used to calculate the error of training and validation sets while training the model found to be 0.010 for the training set and 0.011 for the validation set. The values of MSE for each training and validation sets were relatively small and close to each other which means that no problem of overfitting or underfitting found with this developed ANN.

The mean absolute error used for measuring the performance of the ANN was 0.072 for the training set and 0.075 for the validation set as shown in figure 6.2 below.

Figure 6.2: MSE And MAE of the model



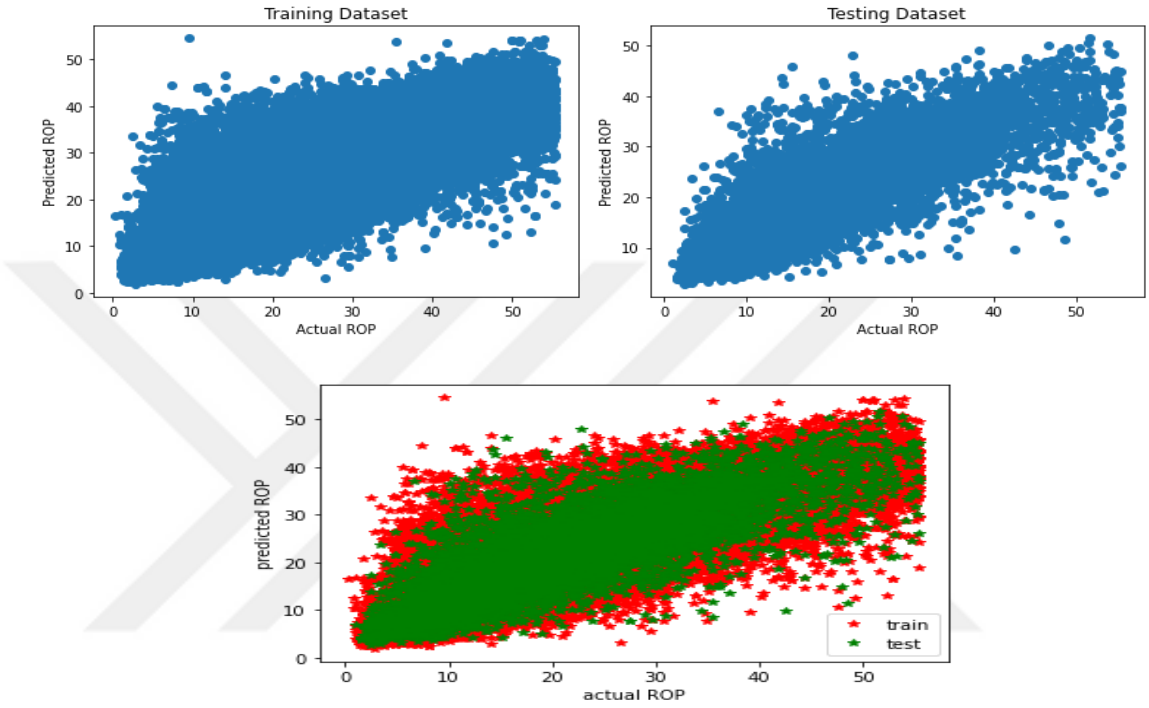
When comparing the actual values of ROP and the predicted values by the generated ANN, the R^2 which used to measure the goodness of fitting of the predicted values found to be 0.74 for the training set and 0.70 for the testing set. The ratio for both training and testing is acceptable and consider to be effective as the ratio is quite close to 1 which is the best ratio of the network. The R^2 value for the implemented model is shown in figure 6.3.

Figure 6.3: R^2 of the model

```
[ ] #calculating and print R2 of training and testing data
    print('the R2 on training set is :\{:0.3f}'.format(r2_score(y_train, y_train_pred)))
    print('the R2 on testing set is :\{:0.3f}'.format(r2_score(y_test, y_test_pred)))

    the R2 on training set is :\0.743
    the R2 on testing set is :\0.700
```

The comparison between actual and predicted values of ROP in training and testing datasets will be shown in figure 6.4 below. The figure showed that the data is correlated and close to each other, as well as centered around the regression line with only some points distant from data aggregation points.

Figure 6.4: Predicted Vs actual ROP in training and testing datasets



With the limitation of data used to develop this ANN in this project, the results seemed to be acceptable and showed the effectivity of the large data concept when using to analyze data and building effective models. The Python language and ANN trained the model well and produce predicted value quite close to the actual values of the wells.

6.2 APPLYING THE MODEL FOR NEW WELL

Finally, to make sure that the model can work properly when introducing to new wells with homogenous data and drilled in the same geological area, the model was tested on a new actual well called (Ru – 484) which has been drilled in the south of Rumaila Oil Field. The following dataset with the same features that the model trained on was introduced to the model and is shown in figure 6.5.

Figure 6.5: New data used to test the model

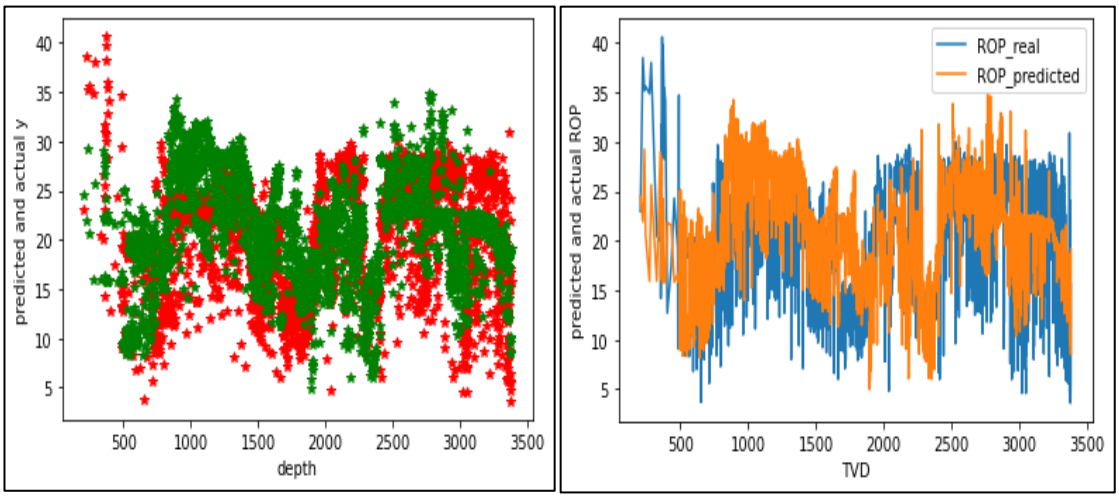
```
[ ] data
```

	TVD(m)	WOB Avg(klb)	RPM Total Avg(rpm)	Torque Abs Avg(f-p)	SPP Avg(psi)	Flow In Pum Avg(lpm)	Dens Mud In Avg(sg)	ROP Avg(m/hr)
0	44.0	0	68	1415	229	1622	1.05	240.8
1	45.0	1	68	1097	227	1622	1.05	16.0
2	46.0	1	68	1051	221	1625	1.05	16.1
3	47.0	1	68	1120	221	1628	1.05	16.1
4	48.0	1	68	1229	227	1628	1.05	16.3
...	...	...	...	...	...	...	...	...
3334	3377.6	13	240	7707	3235	1990	1.32	15.1
3335	3378.6	14	243	6315	3126	2017	1.32	3.6
3336	3379.6	16	242	6516	3141	2014	1.32	5.5
3337	3380.6	16	242	6592	3147	2013	1.32	5.6
3338	3381.6	15	241	7732	3232	1994	1.32	15.7

3339 rows × 8 columns

When the model run with the new above data, it produced quite well predictions for most of the data points except for some points. The red points in the figure below described the actual well ROP ranges while the green ones refer to the predicted values by the ANN model. As shown in figure 6.6 the predicted values lie in the range of actual ones in most cases, but the model was not able to predict the data in some ranges which can be solved when considering more data to train the model on more examples of the situation to be improved. More figures and details about data and results are shown in Appendix C.

Figure 6.6: Applying the model to the new well



6.2 DISCUSSION

Although the data volume in hand was not big enough to call it real big data, the results showed a good performance of the developed model. The big data projects required the supplement of more than 45 wells data and its always useful to provide data as much as possible to find strong patterns and models that can produce highly accurate performance.

According to the results of the developed model, the minimum error of a value approximately 1 percent and a performance of 70 percent when working with testing data was acceptable to say that the model is working properly. The generalization of the model to new data was checked by testing the model with new unseen data of well (RU-484) which as shown can predict a good number of actual data, while in some cases it predicts them distant from actual values. The predictions would be better if more wells were included in the research to train the model effectively. An issue while the developing is that some of the data were eliminated as null values (NaN) and outliers, finding more effective ways to handle them instead of removing them could provide better results as the data volume would increase because considering more data could increase the generalization of the ANN model. Also, including more features that affect ROP directly such as differential pressure, compaction strength of each formation, and other formation properties that consider as uncontrollable parameters and affect the ROP value strongly. The formations properties were including indirectly with this thesis by considering the WOB and Torque as an indicator for the strength of the formation. Moreover, introducing some corrections to drilling parameters used as stated before can enhance the quality of data entering the ANN. The data preprocessing also is an important step to consider before the training because it helps in the removal of noisy data that could mislead the model.

The results showed that big data combining with machine learning techniques as the analytical tools can predict values quite close to what the estimates of real life ROP values could be. Machine learning algorithms in the term of neural network and deep learning are fast and effective tools to deal with big data and make near actual estimations of detected situations in drilling operations. The study considered as a simple example of the possibility of applying big data and machine learning to handle

problems in drilling operations and discussed the methodology and problems while building models to provide motivations for further works in this area.

6.4 CONCLUSION

ANN model with the methodology used, developed with a quite well accuracy to predict ROP values and enhance drilling parameters. This work was part of detecting the use of large datasets with machine learning algorithms to analyze the data and produce models that can work effectively when applying to a real time environment. The data used in this thesis was belonging to wells drilled South and North Rumalia oilfield in the south of Iraq. The ANN rate of penetration model was based on seven drilling parameters (TVD, WOB, RPM, Torque, SPP, Flow in, and mud density). Different preprocessing steps applied to data to ensure the quality of data introducing to the ANN, hence provide more accurate results. Based on multiple trials with different structures of ANN, the optimal structure was feed forward with backpropagation of one input layer with 2 hidden layers and the output layer. The training was exposed to the dataset to obtain high performance of ANN by reducing the mean squared error and increasing the coefficient of determination value.

The proposed model showed good accuracy when comparing the predicted values with raw data. Also, applying the model on data from the new well showed a good matching of predicted and raw data. The neural network approach is not new to the drilling operations as mentioned before, different studies have been conducted to implement models that can optimize ROP. The difference with this thesis is the amount of data used to implement the model, multiple numbers of wells were included during the project. However, the amount was not enough considering more and more data in future researches could enhance the performance of the model. Moreover, future researches could detect the effect of optimal drilling parameters of the generated model to decrease the cost of drilling operations. The optimal parameters are those who will produce the minimum drilled cost. Also, the model with the parameters that gave the less cost could be implemented to work in a real time environment in the rig site to help engineers enhance the decision making process and efficiency of drilling programs.

The object of this work was to show a simple case about the benefits of big data technology and machine learning algorithms in handling large datasets with effective predictions of new situations. Most of the researches on big data for the oil and gas industry were done without trying to apply a real case to show the potentials behind applying this technology in real life. This research serves as an example of detecting the use of a large volume of data more than the usual ones combining with the analytical capability of machine learning to solve the problem of increasing the ROP while drilling wells by building models that can provide the optimal parameters with less cost for drilling activities.

It seems that adopting big data solutions with the analyzing ability of machine learning can significantly change the energy industry just like the other industries. Oil and gas companies need to invest in the modern analytical power provided by new technologies, as well as invest in the right people who can understand both data and business to achieve the best results of finding more hydrocarbon resources and produce them with lower cost and environmentally friendly way.

REFERENCES

Books

Bourgoyne, A., Milheim, K., Chenevert, M., & Young, F. (1991). *Applied Drilling Engineering vol 1*. Richarsdon, TX: Society of Petroleum Engineers, Inc.

Holdaway, K. (2014). *Harness oil and gas big data with analytics: optimize exploration and production with data-driven models.* New Jersy: John Wiley & Sons, Inc.

Ohlhorst, F. (2013). *Big Data Analytics: Turning Big data into Big Money.* New Jersy: John Wiley & Sons, Inc.

Moolayil, J (2019). *Learn Keras for Deep Neural Networks.* New York: Springer Science+Business Media.

Periodicals

Ahmed, A, Ali A., Elkatatny S. & Abdulraheem A. (2019). New artificial neural networks model for predicting rate of penetration in deep shale formation. *Sustainability.*

Aliguliyev, R. & Imamverdiyev, Y (2017). Big data strategy for the oil and gas industry: general directions. *Problems of information technology*. (2), pp. 31-42.

Azzedin, F. & Ghaleb, M. (2019). Towards an rchitecture for handling big data in oil and gas industries: service-oriented approach. *International Journal of Advanced Computer Science and Applications*. **10,** (2), pp. 554-562.

Evans, J. (2019). The drilling industry reaps the rewards of data analytic.*Oilman*.

Hassani, H., & Silva, E. (2018). Big data: a big opportunity for the petroleum and petrochemical industry. *Organization of the Petroleum Exporting Countries.* pp. 74-89.

Mnati, K. & Hadi, H. (2018). Prediction of penetration rate and cost with artificial neural network for alhafaya oil field. *Iraqi Journal of Chemical and Petroleum Engineering*. (4), pp. 21-27.

Mohammadpoor, M. & Torabi, F. (2018). Big Data analytics in oil and gas industry: An emerging trend. *KeAi Advanced Research Evolving Science*.

Nagorny, K., Li-ma-Monteiro, P., Barata, J. and Colombo, A.W. (2017). Big data analysis in smart manufacturing: A Review. *Int. J. Communications, Network, and System Sciences*. (10), pp. 31-58.

Perrons, R. & Jensen, J. (2015). Data as an asset: What the oil and gas sector can learn from other industries about "Big Data". *Energy Policy*. **81**, pp. 117-121.

Rawat, N. (2014). Big data analytics in oil & gas industry. *International Journal of Scientific & Engineering Research*. **5,** (5), pp. 1-6.

Rehman, M., Chang, V., Batool, A. & Wah, T. (2016). Big data reduction framework for value creation in sustainable enterprises. *International journal of information management*. (36), pp. 917-928.

Shadizadeh, S., Karimi F. & Zoveidavianpoor M. (2010). ANN models to predict the stuck pipe occurrence in Iranian oil fields. *Iranian Journal of Chemical Engineering.* **7**, (4).

Wang, Y. & Salehi, S. (2015). Application of real-time field data to optimize drilling hydraulics using neural network approach. *Journal of Energy Resources Technology*. **137**, pp. 1-9.

Other Publications

Baaziz, A. & Quoniam, L. (2014). How to use big data technologies to optimize operations in upstream petroleum industry. *21st World Petroleum Congress.* 15-19 June, 2014, Moscow, Russia, pp. 1-9.

cleantech group. 2015. *Insights into digital oilfield: transforming upstream oil & gas with big data and cloud.* https://www.cleantech.com/wp-content/uploads/2015/01/i3WhitePaper_Digital-Oilfield.pdf [accessed 27 February 2020].

EMC Corporation. 2013. *Big data science challenging the oil industry*. http://web.idg.no/app/web/online/event/energyworld/2013/emc.pdf [accessed 5 January 2020].

Eren, T. (2010). *Real-time optimization of drilling parameters during drilling operations.* Phd Thesis, Middle East Technical University.

ESDS. 2016. *Big value for Big Data in Oil and Gas Industry*. https://www.esds.co.in/blog/big-value-big-data-oil-gas-industry/#sthash.MFmZFOVP.kCQZqfLa.dpbs [accessed 18 February 2020].

Forbes. 2015. *Big Data In Big Oil: How Shell Uses Analytics To Drive Business Success.* https://www.forbes.com/sites/bernardmarr/2015/05/26/big-data-in-big-oil-how-shell-uses-analytics-to-drive-business-success/#619f7d34229e [accessed 18 February 2020]

intellipaat. 2016. *7 Big Data Examples: Applications of Big Data in Real Life*. https://intellipaat.com/blog/7-big-data-examples-application-of-big-data-in-real-life/ [accessed 2 January 2020].

Microsoft. 2014. *Drilling for new business value: how innovative oil and gas companies are using big data to outmaneuver the competition.* http://download.microsoft.com/documents/en-us/Drilling_for_New_Business_Value_April2014_Web.pdf [accessed 27 December 2019].

Mohaghegh, S., Arefi, R., Ameri, S., and Rose, D. (1994). Design and Development of an Artificial Neural Network for Estimation of Formation Permeability. *SPE Petroleum Computer Conference.* 1 July-3 August, 1994, Dallas, USA.

SAS institute, solution brief, 2015, https://www.sas.com/content/dam/SAS/en_us/doc/solutionbrief/oil-and-gas-improve-drilling-efficiency-106477.pdf [retrieval data 20 December 2019].

Song, M. (2018). Research on the application of big data in oil & gas ındustry. *International Conference on Computational, Modeling, Simulation, and Mathematical Statistics.* 2018.

Tankimovich, M. (2018). *Big data in the oil and gas industry: a promising courtship.* Engineering Honors Thesis, The University of Texas at Austin.

towards data science. 2019. *Introduction to Machine Learning Top-Down Approach.* https://towardsdatascience.com/introduction-to-machine-learning-top-down-approach-8f40d3afa6d7 [accessed 10 February 2020].

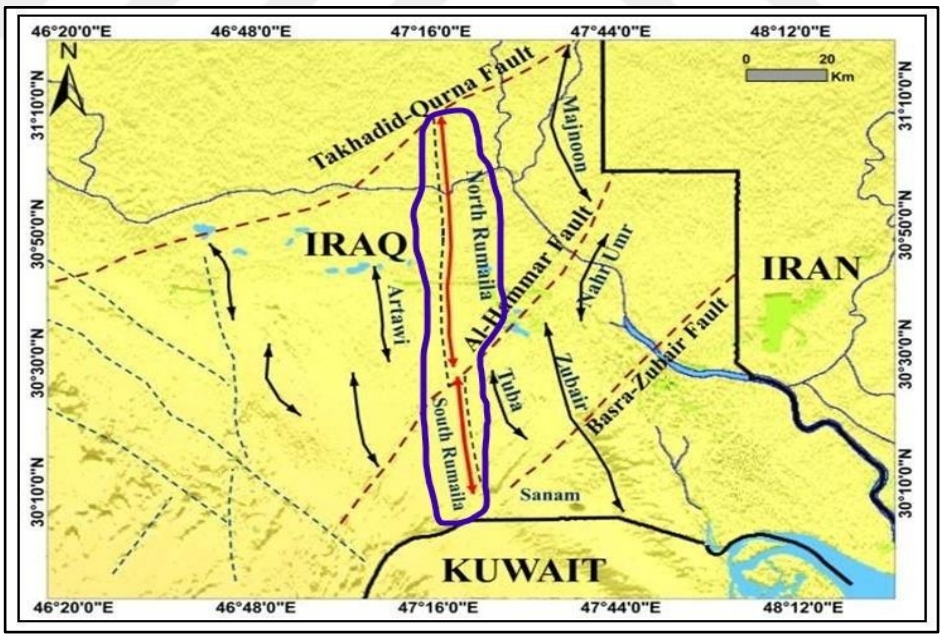
APPENDICES

APPENDIX A.1 RUMAILA OILFIELD AND DATA SPECIFIC INFORMATION

The data used in the study belongs to wells drilled in the Rumail Oilfield in the South of Iraq. The Rumaila oilfield map is shown in the figure-1 and figure-2.

Figure-1: Rumaila oilfield map

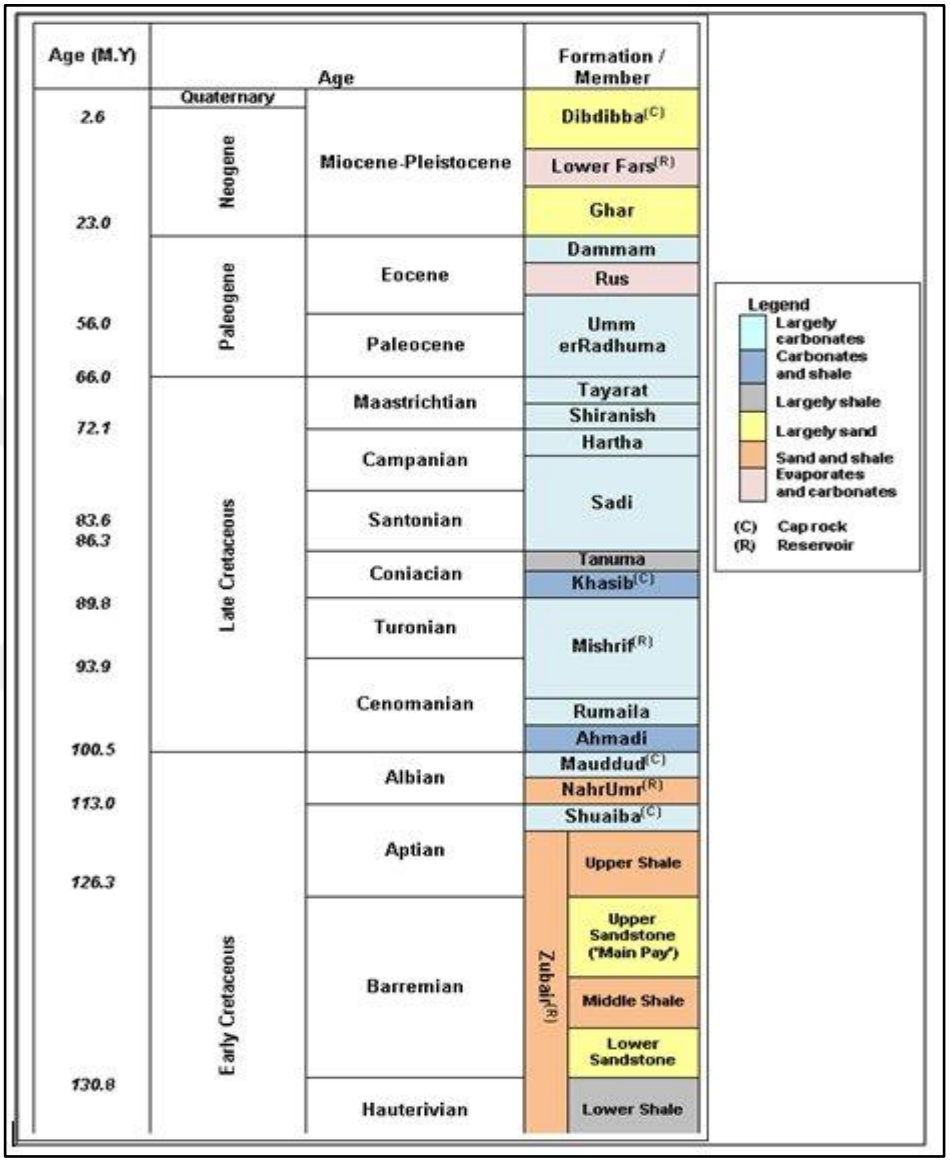
Figure-2: North and South Rumail oilfield



Rumaila oil field is one of the world's greatest oilfields with about a billion barrels of recoverable oil contained in it. İt is located 50Km to the west of Basra in the South of Iraq. The subsurface geology of the region showing similarity throughout the whole

south of the Iraq region. The main reservoirs of the field are Zubair formation, upper shale reservoir and Mishrif reservoir. Formations include in the research with the lithology description and their structures are shown in figure-3 below.

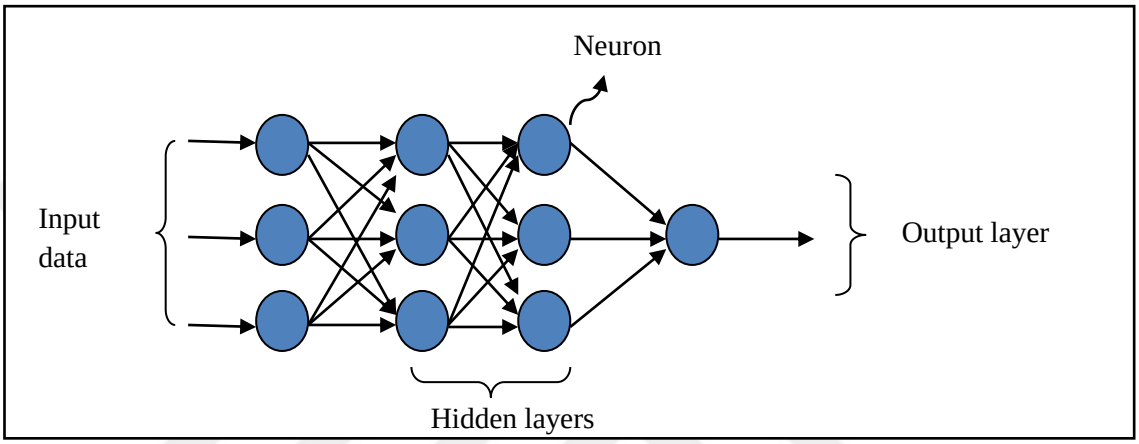
Figure-3: Formation lithology description



APPENDIX A.2 THE NEURAL NETWORK WORKING MECHANISM

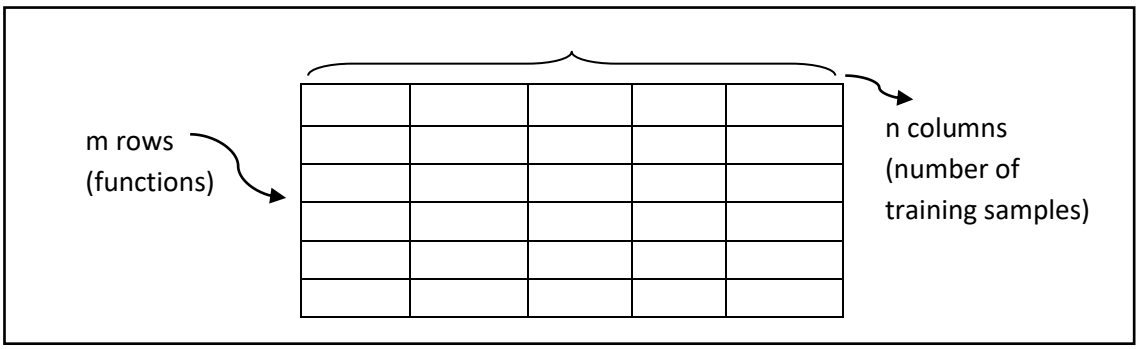
The neural network consists of several logical components; input data, weights, neurons, hidden layers, activation function (learning procedure), and the output layer. Those components form the basic structure of each neural network as illustrated in figure-1 below.

Figure-1: Neural network logical components



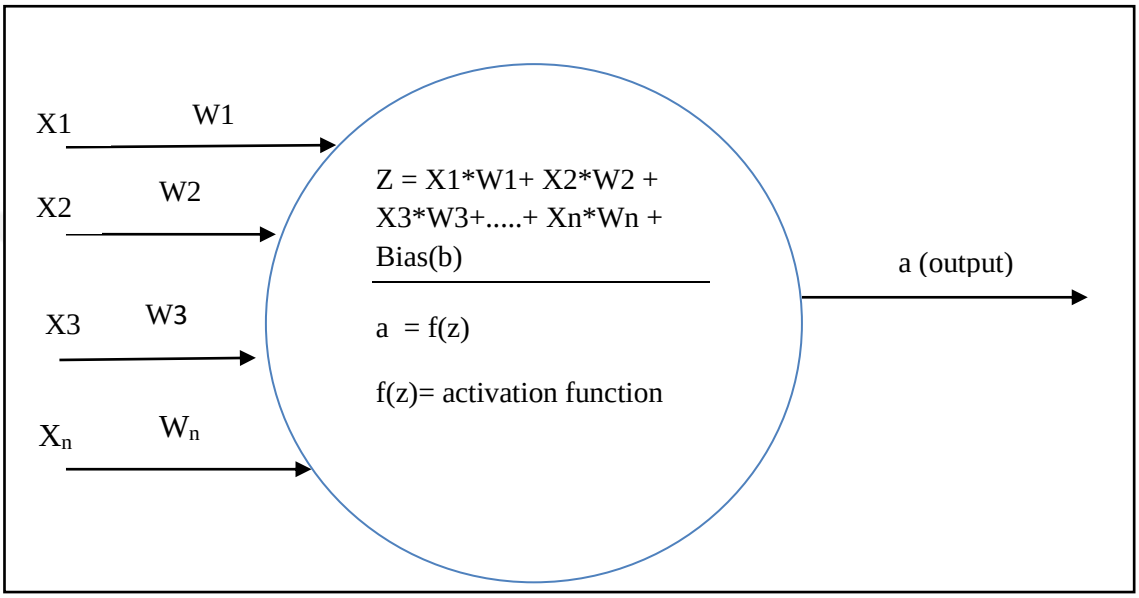
The input data is introduced to the hidden layers to produced the output of each hidden layer which in turn will be the input of the next layer and so on, until reaching the final output layer. All the input data should be numerical data, and categorical data should be transformed into numerical values in order for the NN to be able to interpret it. The input data is in the form of an n-dimensional matrix with the shape of (mxn) where m is the number of rows (functions), and n is the number of columns (number of training samples) (Moolayil 2019), the structure is shown in figure-2.

Figure-2: n-dimensional matrix of input data



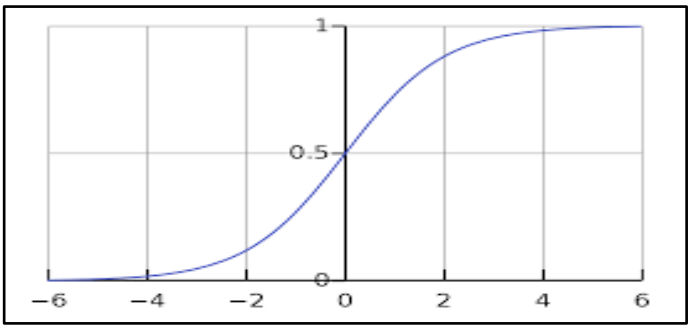
Neurons are the units responsible for the computation of output. A neuron receives the data from the previous neuron and sum all the inputs multiplied by their corresponding weights together which is denoted as Z. The computed input (Z) transferred to output with the use of the activation function. The structure and computation process of a single neuron is shownin figure-3 below.

Figure-3: Structure of a single neuron



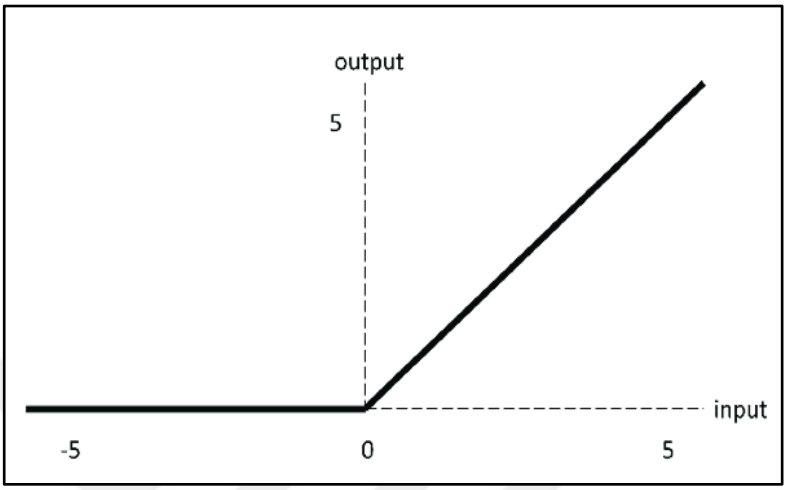
The activation function used to compute the output of a layer depending on the computed input (Z). There are several kinds of activation functions, but the most common are sigmoid and ReLU functions. The sigmoid function used to set the output between 0 and 1, and defined as ($\frac{1}{1+e^{-z}}$). it is known to enhance the learning process very well (Moolayil 2019), figure-4 describes the sigmoid function.

Figure-4: The sigmoid function



ReLU is another function defined as f(Z)= max(0, Z), when the Z is negative value it would output zero, while when Z is positive value it would output the same value (Moolayil 2019). Figure-5 shows the ReLU function concept.

Figure-5: ReLU function



Another term in the NN is the loss function which helps the NN to improve it's learning process in each iteration. İt used to measure the difference between actual and predicted values. Several kinds of loss function used in the neural network depending on data outcome, whether it is regression or classification problem. For regression problems, the most common function is the mean squared error (MSE) which measures the average squared difference of actual and predicted values with the mathematical term:

$\sum_{n=1}^{k} \frac{(Actual-Predicted)^2}{k}$, where k = number of training samples **(**2.1)

Also, the mean absolute error used to measure the average absolute error between actual and predicted values as shown in the equation (2.2) below.

$$\sum_{n=1}^{k} |Actual - Predicted| \quad (2.2)$$

The optimization process of the NN depends on reducing the loss function to a certain limit; therefore, optimization algorithms used to update the weights of neurons to minimize the loss function (Moolayil 2019). The optimization algorithms work on the basis of derivatives, partial derivatives, and chain rule to determine how much change to apply on weights to make a change in the loss function and network.

The most famous optimizers are Stochastic Gradient Descent (SGD) and Adaptive Moment Estimation (Adam). The SGD considered as fast optimizer compared with other optimizers, but produce a very noisy curve. İt updates the weights with the following formula:

$$W: = W - \alpha \frac{dw}{dj} \quad (2.3)$$

Where; α = learning rate

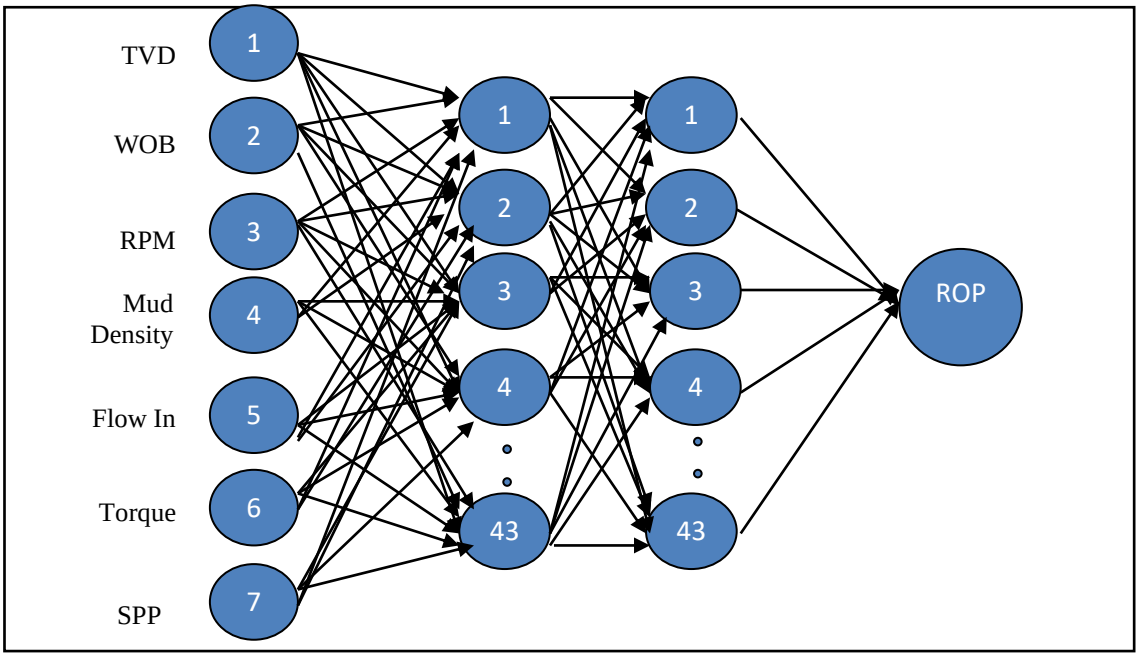
$\frac{dw}{dj}$ = loss

Adam is the most popular optimizer, it depends on computing the momentum and variance of the loss which helps in smoothing the learning curve effectively with the following formula:

$$W: = W - \alpha * (\text{momentum and variance combined}) \quad (2.4)$$

Considering the neural network developed in this project, the NN structure is shown in figure-6 below.

Figure-6: The developed ANN structure



2 hidden layers with 43 neurons in each layer, the activation function is sigmoid for all the layers, MSE used as the loss function, Adam is the optimizer algorithm, and ROP is the output parameter. The two dimensional matrix of input data would be as follow:

Table-1: 2 dimensional matrix of the input parameters

One training sample

$X_{1,1}$	$X_{1,2}$	$X_{1,3}$	…..	$X_{1,95558}$
$X_{2,1}$	$X_{2,2}$	$X_{2,3}$	…..	$X_{2,95558}$
$X_{3,1}$	$X_{3,2}$	$X_{3,3}$	…..	$X_{3,95558}$
$X_{4,1}$	$X_{4,2}$	$X_{4,3}$	…..	$X_{4,95558}$
$X_{5,1}$	$X_{5,2}$	$X_{5,3}$	…..	$X_{5,95558}$
$X_{6,1}$	$X_{6,2}$	$X_{6,3}$	…..	$X_{6,95558}$
$X_{7,1}$	$X_{7,2}$	$X_{7,3}$	…..	$X_{7,95558}$

$X_{1,1}$ refers to X_1 in the training sample one, while $X_{1,95558}$ refers to X_1 in the training sample 95558 and so on for the other parameters. Calculating $Z^{[1]}$ which is the computed value of the first hidden layer would be as follow:

$$Z^{[1]} = W^{[1]} * X + b^{[1]} \tag{2.5}$$

İn matrix forum :

$Z^{[1]} = W^{[1]} *X + b^{[1]}$

(43, 95558) = (43,7) * (7, 95558) + (43, 95558)

$$\begin{bmatrix} W1,1 & W2,1 & W3,1 & W4,1 & W5,1 & W6,1 & W7,1 \\ W1,2 & W2,2 & W3,2 & W4,2 & W5,2 & W6,2 & W7,2 \\ W1,3 & W2,3 & W3,3 & W4,3 & W5,3 & W6,3 & W7,3 \\ ... & ... & ... & ... & ... & ... & ... \\ W1,43 & W2,43 & W3,43 & W4,43 & W5,43 & W6,43 & W7,43 \end{bmatrix} * \begin{bmatrix} X1,1 & X1,2 & X1,3 & ... & X1,95558 \\ X2,1 & X2,2 & X2,3 & ... & X2,95558 \\ X3,1 & X3,2 & X3,3 & ... & X3,9558 \\ X4,1 & X4,2 & X4,3 & ... & X4,9558 \\ X5,1 & X5,2 & X5,3 & ... & X5,9558 \\ X6,1 & X6,2 & X6,3 & ... & X6,9558 \\ X7,1 & X7,2 & X7,3 & ... & X7,9558 \end{bmatrix} + \begin{bmatrix} b1,1 & b2,1 & b3,1 & ... & b95558,1 \\ b1,2 & b2,2 & b3,2 & ... & b95558,2 \\ b1,3 & b2,3 & b3,3 & ... & b95558,3 \\ ... & ... & ... & ... & ... \\ b1,43 & b2,43 & b3,43 & ... & b95558,43 \end{bmatrix}$$

$$= \begin{bmatrix} Z1,1 & Z2,1 & Z3,1 & ... & Z95558,1 \\ Z1,2 & Z2,2 & Z3,2 & ... & Z95558,2 \\ b1,3 & Z2,3 & Z3,3 & ... & Z95558,3 \\ ... & ... & ... & ... & ... \\ Z1,43 & Z2,43 & Z3,43 & ... & Z95558,43 \end{bmatrix}$$

This computation process produces the Z value of the first layer and for each training sample in each neuron. The same process would be done for the second layer. Then, the activation function sigmoid transmits the Z value into the output of each neuron until reaching the final output. The loss function would calculate the error between the actual value and the value produced by the network, all this process is known as the feedforward, the following equations represent the process of feedforward:

$$Z^{[1]} = W^{[1]} * X + b^{[1]} \quad , \quad Z^{[2]} = W^{[2]} * X + b^{[2]}$$

$$a^{[1]} = f(Z^{[1]}) \quad (2.6) \quad , \quad a^{[2]} = f(Z^{[2]}) \quad (2.7)$$

$$L = \text{loss function} = \sum_{n=1}^{k} \frac{(\text{Actual} - \text{Predicted})^2}{k}$$

Feedforward

The next step is to use an algorithm that can send feedback to the system and determining the change that leads to the correct way of decreasing loss function, this algorithm is known as the backpropagation (Moolayil 2019). The updating is done after processing all the samples in a batch. A batch is a determined number of training samples, processing a full batch called iteration. Processing the whole batches in the training sample called epoch. By running a number of iterations and epochs, the network updates the weights to improve its predictions for the used training samples. The backpropagation starts by calculating the derivative of the loss function, output of last layer ($a^{[2]}$), then calculating all the other values gradually as shown in the equations below with updating the weights used in the network. The equations below describe the case used in this research of two hidden layers and should be effective for each neuron as well.

$$dZ^{[2]} = da^{[2]} * f(Z^{[2]}) \quad (2.8)$$

$$dW^{[2]} = dZ^{[2]} * a^{[1]} \quad (2.9)$$

$$db^{[2]} = dZ^{[2]} \quad (2.10)$$

Backpropagation

$$W^{[2]} := W^{[2]} - \alpha * (\text{momentum and variance together}) \quad (2.11)$$

$$b^{[2]} := b^{[2]} - \alpha * (\text{momentum and variance together}) \quad (2.12)$$

Update

$$da^{[1]} = W^{[2]} * dZ^{[2]} \quad (2.13)$$

$$dZ^{[1]} = da^{[1]} * f(Z^{[1]}) \quad (2.14)$$

$$dW^{[1]} = dZ^{[1]} * X \quad (2.15)$$

$$db^{[1]} = dZ^{[1]} \quad (2.16)$$

Backpropagation

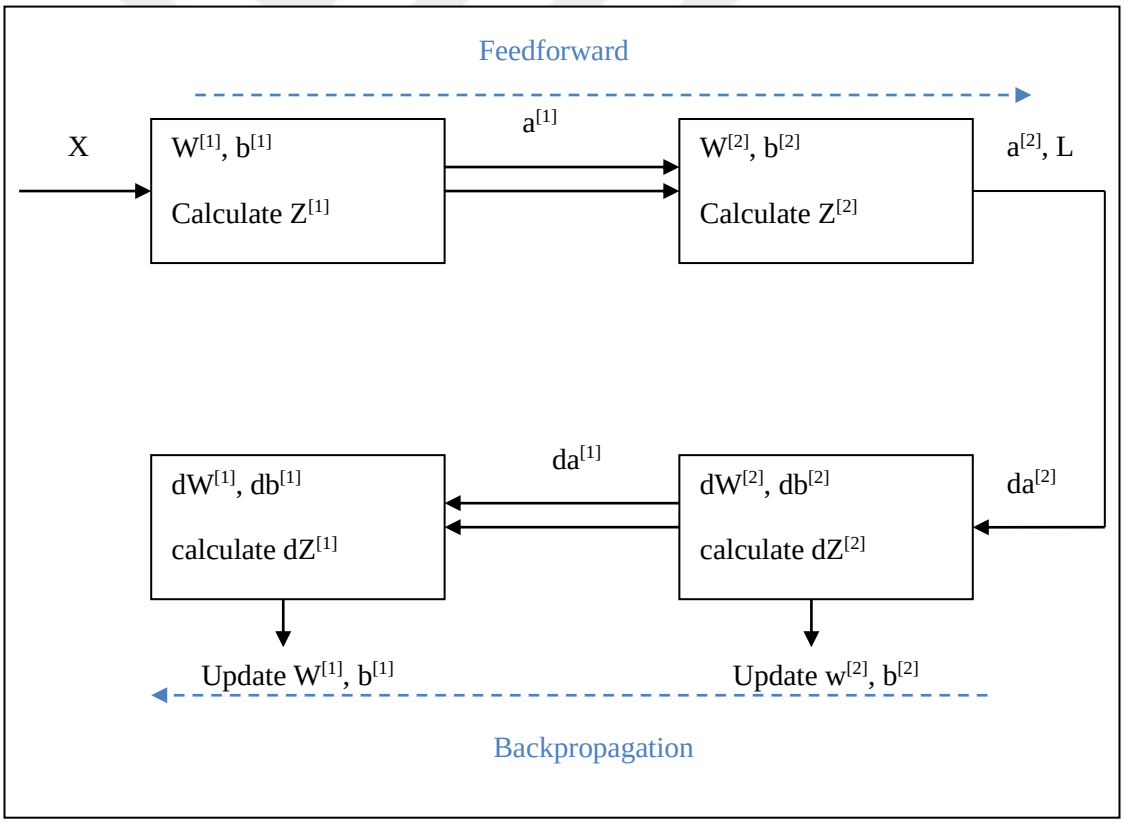
$$W^{[1]} := W^{[1]} - \alpha * (\text{momentum and variance together}) \quad (2.17)$$

$$b^{[1]} := b^{[1]} - \alpha * (\text{momentum and variance together}) \quad (2.18)$$

Update

a simple figure to illustrate the procedure of feedforward and backpropagation is shown in figure-7 below.

Figure-7: Feedforward and backpropagation



This procedure continues until finding the optimal values of weight that could reduce the cost function of the network and produce very well predictions. All the above explanation forms the mathematical procedure of the working mechanism of the neural network, while the code part was explained in chapter five of this thesis.

APPENDIX A.3 BUILDING THE ANN

More details about the data used to build the network will be included in this Appendix, as well as more details of the results obtained during the project. A quick look at the distribution of the features is shown in figure-1 and figure-2.

Figure-1: Distribution of WOB, RPM, TVD, and Torque

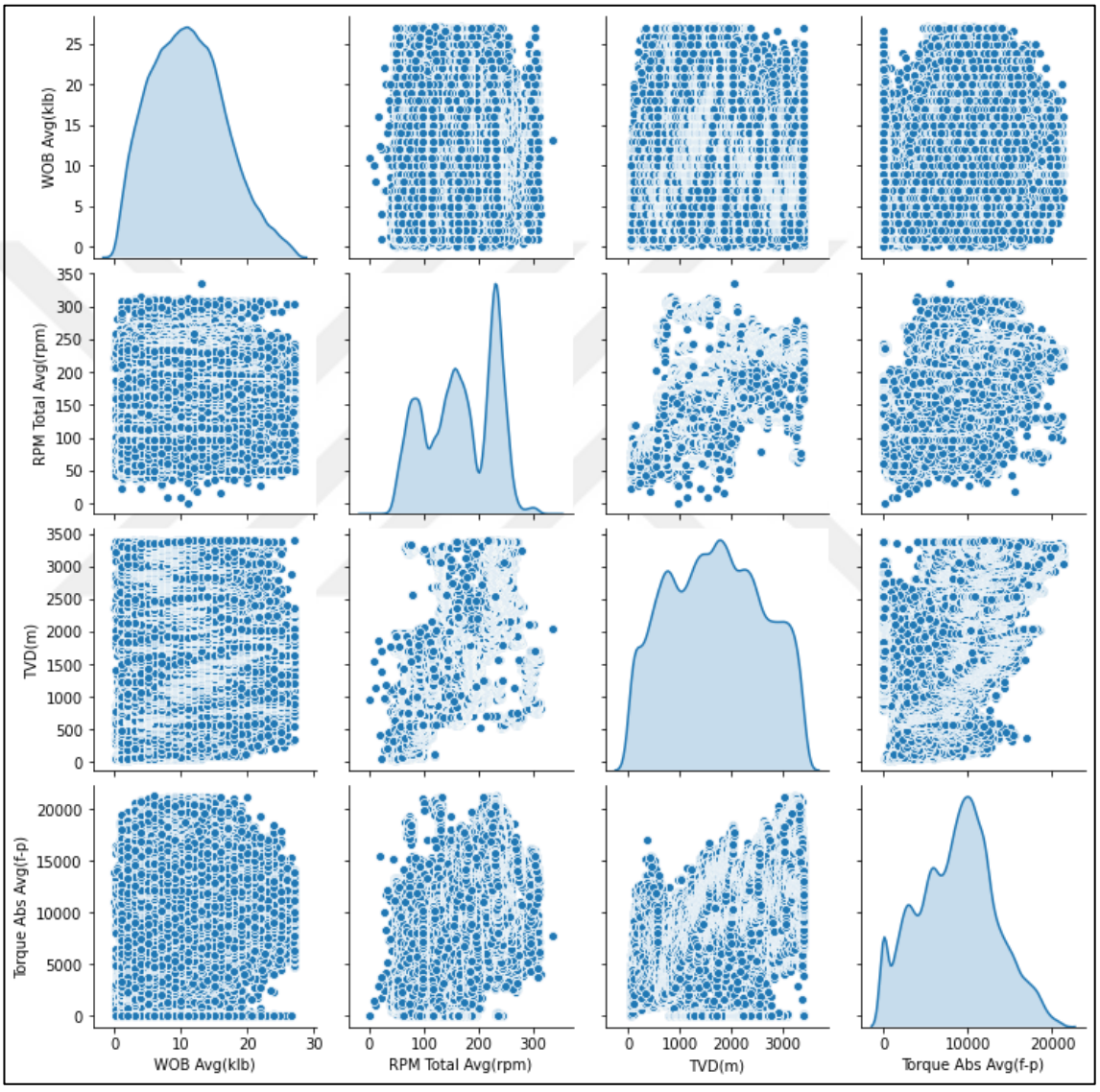
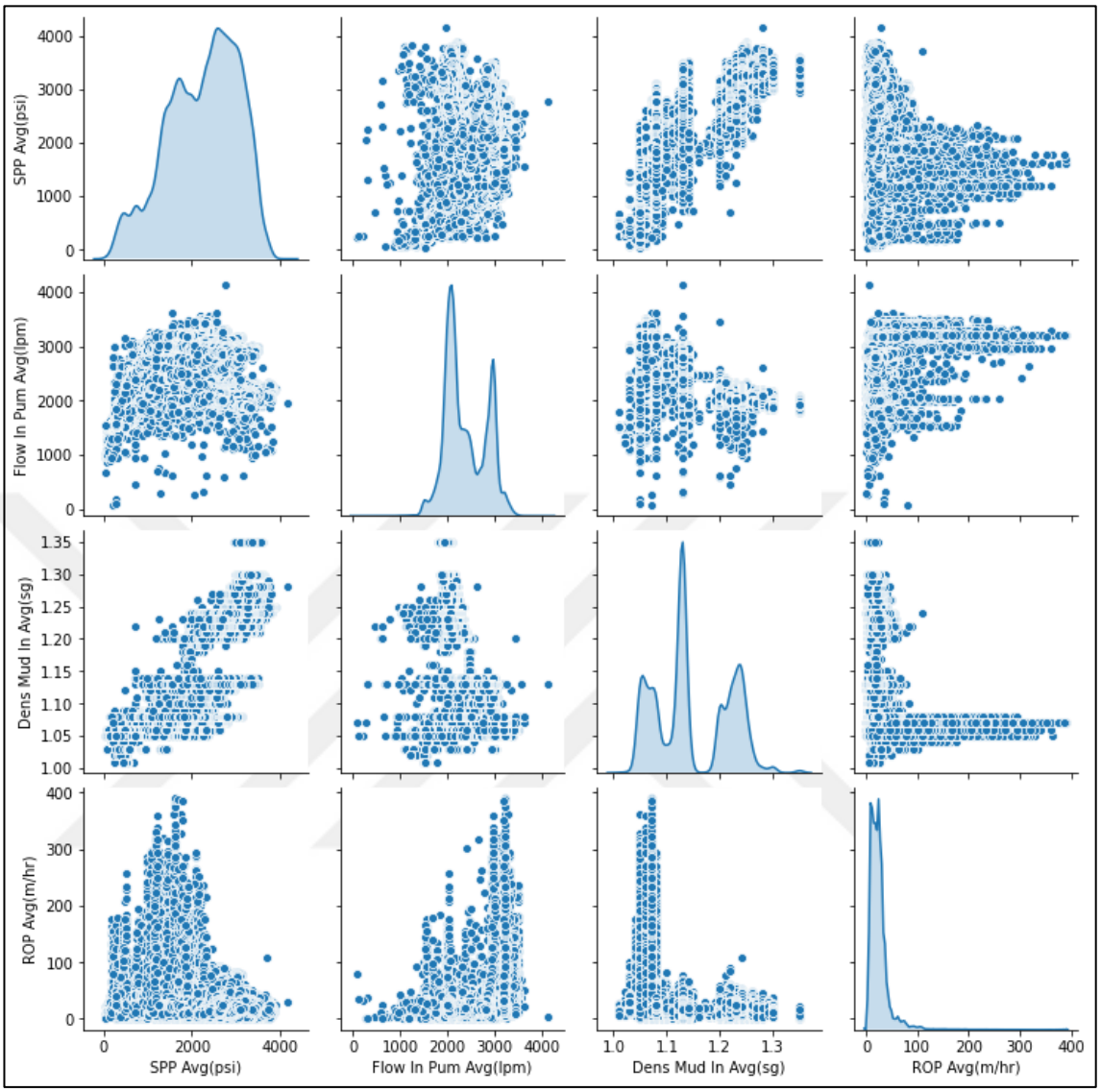


Figure-2: Distribution of SPP, Flow in, Mud density and ROP



Box plot and distribution plot used here to show the outliers founded in some features utilized in building the model. Figure-3 and Figure-4 describes the outliers in WOB, Torque, Flow in, and ROP, while figure-5 and figure-6 show the features after the elimination of the outliers.

Figure-3: Detection of outliers with Box plot

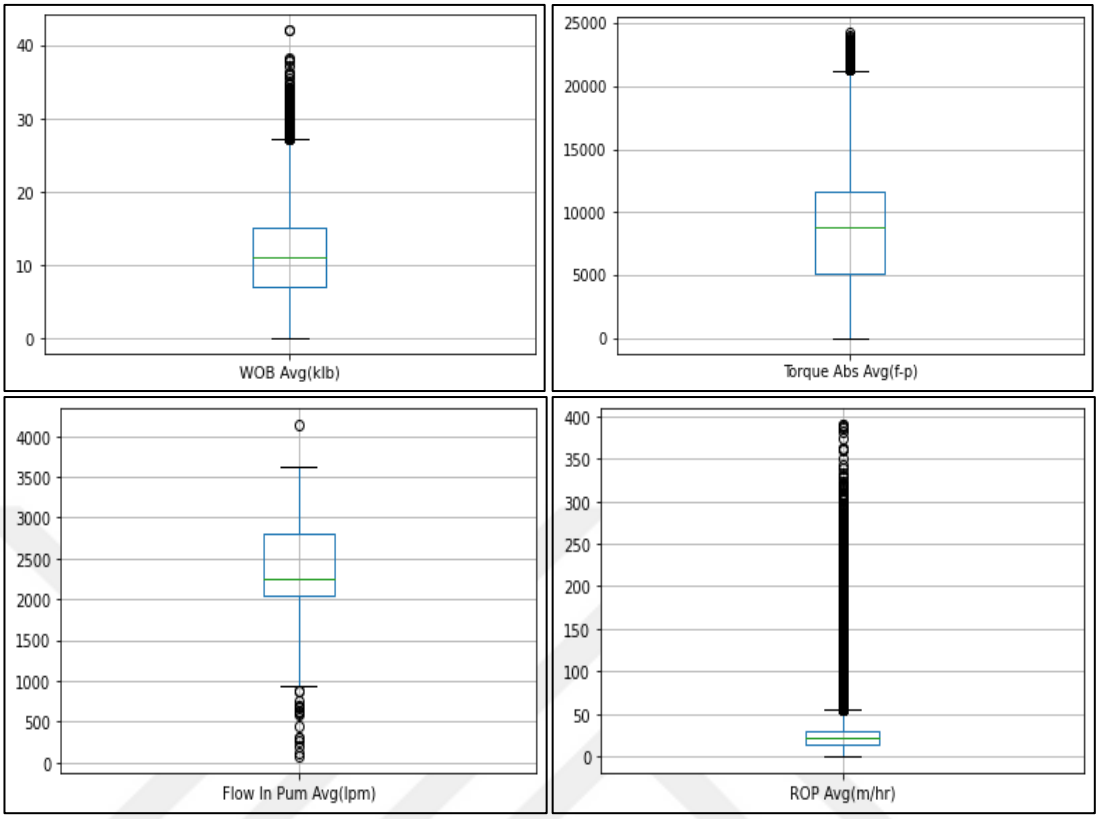


Figure-4: Detection of outliers with distribution plot

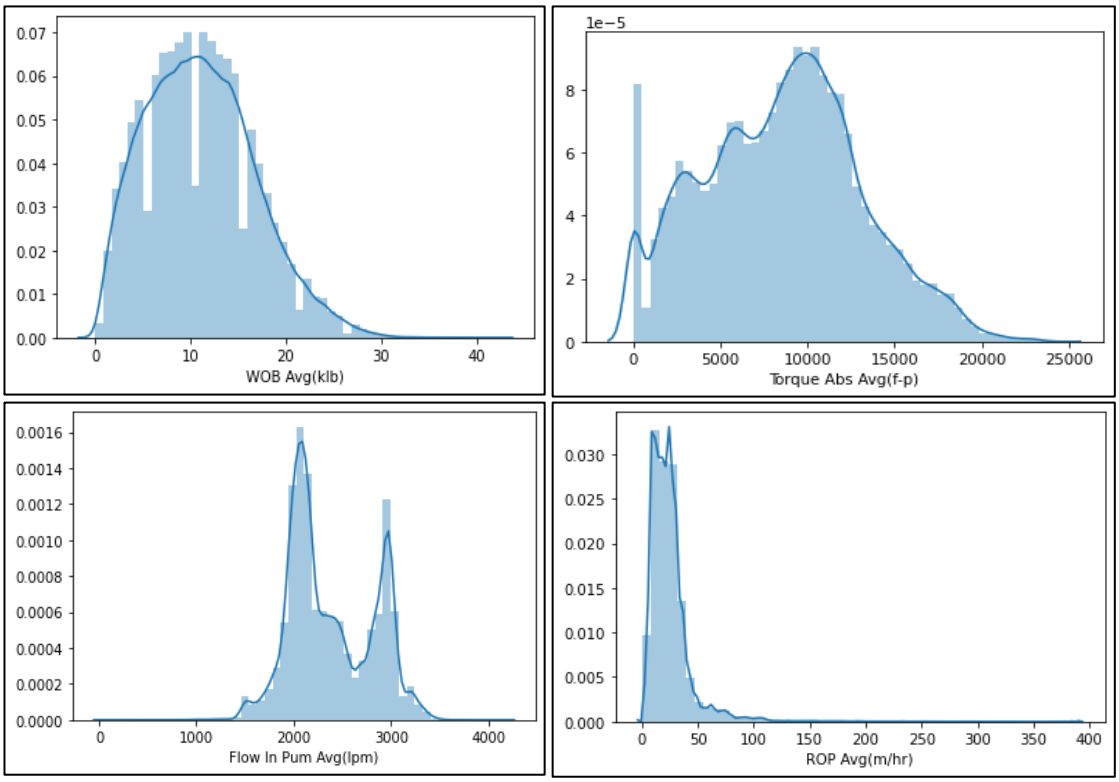


Figure-5: Illustration of the elimination of outliers with Box plot

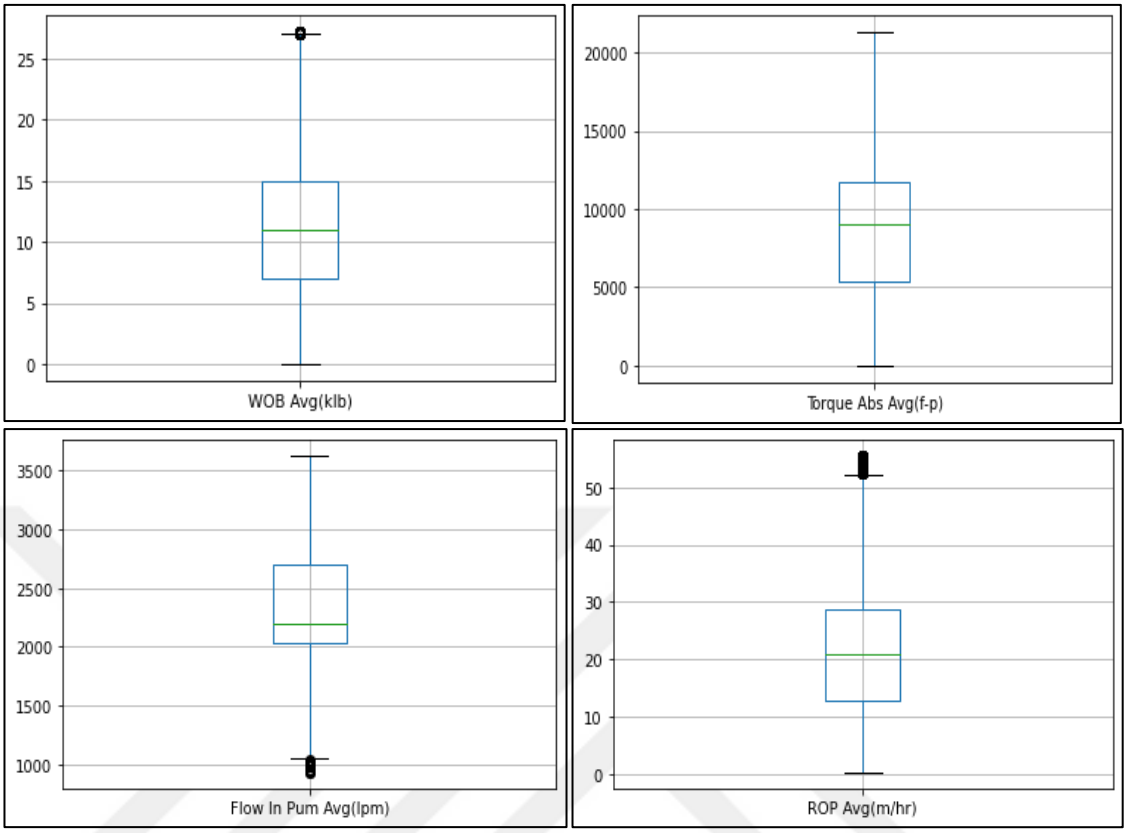
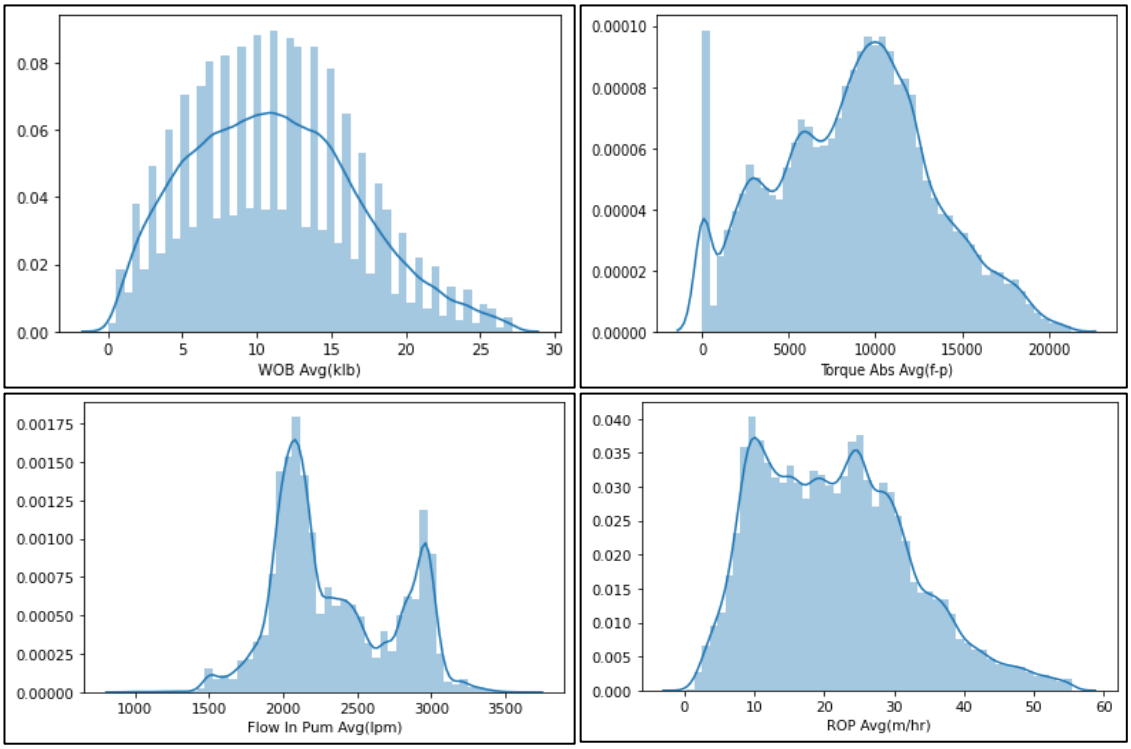


Figure-6: Illustration of the elimination of outliers with distribution plot



Some of the trials conducted to find the optimal number of layers and neurons are shown in table-1, while table-2 shows the use of different activation functions to reach the optimal structure of ANN.

Table-1: Results of three layers and different number of neurons

Layer	Neuron	Loss	Mae	Val Loss	Val Mae	R2 Train	R2 Test
1	8	0.0194	0.1065	0.019	0.105	0.478	0.47
	12	0.0181	0.1021	0.0174	0.1001	0.52	0.513
	16	0.0176	0.0999	0.0174	0.1012	0.521	0.51
	20	0.0174	0.0991	0.0167	0.0973	0.539	0.529
	26	0.0168	0.097	0.0166	0.0963	0.546	0.539
	32	0.017	0.0976	0.0165	0.0956	0.547	0.541
	38	0.0172	0.0987	0.0165	0.0966	0.545	0.537
	40	0.0167	0.0963	0.0163	0.0954	0.555	0.55
	43	0.0166	0.0965	0.0161	0.0946	0.557	0.547
2	8	0.0146	0.089	0.0143	0.088	0.61	0.595
	12	0.0137	0.0856	0.0134	0.0843	0.639	0.618
	16	0.0095	0.0714	0.0097	0.0726	0.645	0.632
	20	0.013	0.0835	0.013	0.0835	0.656	0.639
	26	0.0127	0.082	0.0127	0.0815	0.661	0.641
	32	0.0128	0.0821	0.0128	0.0822	0.66	0.641
	38	0.0123	0.08	0.0122	0.0796	0.676	0.654
	40	0.0129	0.0828	0.013	0.0833	0.656	0.641
	40	0.0115	0.0776	0.012	0.0786	0.693	0.673
	43	0.0121	0.0795	0.0122	0.0796	0.679	0.654
3	8	0.0127	0.0816	0.0129	0.0827	0.661	0.638
	20	0.0108	0.0743	0.0114	0.0757	0.715	0.675
	32	0.0106	0.0743	0.0114	0.0763	0.713	0.68
	40	0.0103	0.0713	0.0112	0.0747	0.727	0.693
	42	0.0101	0.0714	0.0112	0.0757	0.727	0.68
	43	0.0103	0.0724	0.0113	0.0753	0.740	0.700

Table- 2: Results of different activation functions

Layer 1	Layer 2	Layer 3	Output	Loss	Mae	Val Loss	Val Mae	R2 Train	R2 Test
relu	sigmoid	sigmoid	sigmoid	0.011	0.0752	0.0117	0.0769	0.709	0.672
relu	relu	sigmoid	sigmoid	0.0138	0.0859	0.0132	0.084	0.645	0.626
relu	relu	relu	sigmoid	0.0145	0.0886	0.0146	0.0888	0.607	0.593
sigmoid	sigmoid	sigmoid	tanh	0.0106	0.0743	0.0115	0.0765	0.718	0.684
tanh	tanh	tanh	sigmoid	0.0132	0.0843	0.0134	0.0848	0.651	0.625

After running several iterations with different combinations of hyperparameters, the summary of the final structure is shown in figure-7.

Figure-7: Summary of the model

```
[ ] model.summary()

Model: "sequential_1"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense_1 (Dense)              (None, 43)                344
_________________________________________________________________
dense_2 (Dense)              (None, 43)                1892
_________________________________________________________________
dense_3 (Dense)              (None, 43)                1892
_________________________________________________________________
dense_4 (Dense)              (None, 1)                 44
=================================================================
Total params: 4,172
Trainable params: 4,172
Non-trainable params: 0
_________________________________________________________________
```

The results of the predicted rate of penetration against the actual rate of penetration in both training and testing sets are shown in figure-8 and figure-9 respectively.

Figure-8: Actual and predicted ROP comparison in training dataset

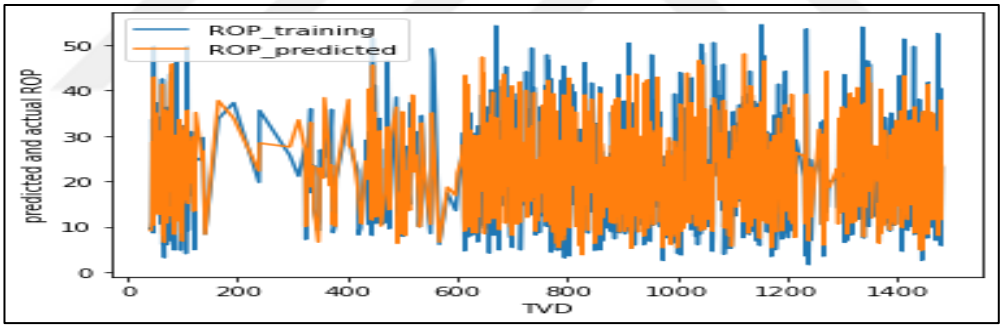
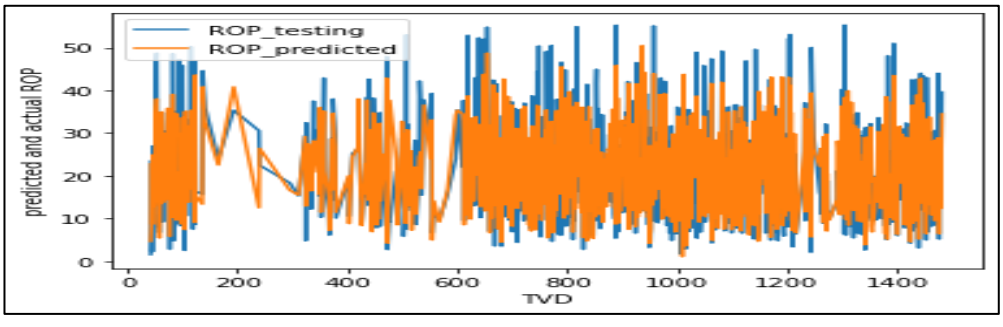


Figure-9: Actual and predicted ROP comparison in testing dataset



The model then implemented with new data to check the possibility of using the model in new applications. The description of new data is shown in figure-10.

Figure-10: Description of the data of the new well

```
[110] data1.describe()
```

	TVD(m)	WOB Avg(klb)	RPM Total Avg(rpm)	Torque Abs Avg(f-p)	SPP Avg(psi)	Flow In Pum Avg(lpm)	Dens Mud In Avg(sg)	ROP Avg(m/hr)
count	3318.000000	3318.000000	3318.000000	3318.000000	3318.000000	3318.000000	3318.000000	3318.000000
mean	1716.821398	9.888186	191.521398	5887.886679	2182.009042	2431.839964	1.172824	31.506450
std	963.865617	4.269823	55.030931	1733.353810	616.472495	474.195848	0.079118	38.246117
min	45.000000	1.000000	34.000000	51.000000	221.000000	1496.000000	1.050000	3.600000
25%	889.050000	7.000000	166.000000	4985.000000	1849.250000	2033.000000	1.130000	15.900000
50%	1718.300000	10.000000	190.000000	6013.000000	2270.000000	2195.000000	1.140000	22.300000
75%	2552.450000	13.000000	239.000000	6892.000000	2537.750000	2960.750000	1.230000	26.200000
max	3381.600000	23.000000	250.000000	16713.000000	3311.000000	3315.000000	1.320000	342.000000

Results of using the model to predict the rate of penetration gave the following result in normalized and matrix form which is shown in figure-11.

Figure-11: Predicted ROP values in matrix form

```
y_pred
```

```
array([[0.56418175],
       [0.4955141 ],
       [0.693058  ],
       ...,
       [0.13325576],
       [0.13380945],
       [0.4204659 ]], dtype=float32)
```

Denormalizing the values into the original values gave the following result which is shown in figure-12.

Figure-12: Predicted values of ROP

```
pd.DataFrame(y_real_pred).head(20)
```

	0
0	24.474726
1	21.934021
2	29.243147
3	20.551821
4	15.941999
5	25.638685
6	15.944246
7	28.967255
8	22.217278
9	16.270262
10	22.006634
11	22.165279
12	26.598997
13	28.466154

The generated model saved using the Pickle package in order to be used directly on new data as shown in figure-13.

Figure-13: Saving and loading the model with Pickle