NAZARBAYEV UNIVERSITY

Development of Failure Prediction Model for the Oil and Gas Industry

(Capstone Project)

Master of Engineering Management (2021-2023)

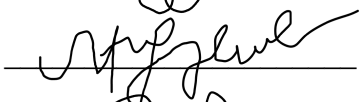
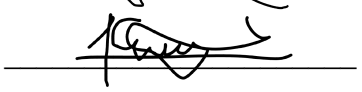
Students name: Altynay Takisheva
Minura Nugumanova
Nursaule Batyrgali
Saiyn Kurmankulov

Supervisor: Professor Essam Shehab

Astana 2023

Declaration Form

We, the undersigned, hereby declare that this report entitled “Development of Failure Prediction Model for the Oil and Gas Equipment” is a result of our own project work except for quotations and citations which have been acknowledged. We also declare that it has not been previously or concurrently submitted for any other degree at Nazarbayev University.

Name:	(Signed)
Altynay Takisheva	
Minura Nugumanova	
Nursaule Batyrgali	
Saiyn Kurmankulov	

Date: 10.05.2023

Abstract

The oil and gas is a crucial sector in the global economy, and the Republic of Kazakhstan has established itself as a significant player in this field. As the industry continues to evolve, companies are actively seeking ways to leverage advanced digital technologies to optimize their operations, enhance safety, and increase profitability. This capstone project represents a cutting-edge effort to develop a framework for failure prediction in the oil and gas sector.

Through close collaboration with Caspi Neft, which is a leading digitalization pioneer in the Kazakhstani market, this study has produced a range of innovative deliverables. These include the failure prediction model for the oil and gas equipment, an example of a well. In addition, the project has developed a comprehensive framework for the implementation of the model, which enables it to be readily integrated into the enterprise resource planning (ERP) system of the company.

To achieve these outcomes, the project team employed a range of advanced machine learning algorithms, leveraging a rich dataset provided by Caspi Neft. This dataset included detailed information on operational conditions, downtime history, and other critical variables, allowing the team to generate a highly accurate failure prediction model with a validation accuracy of over 90%. The model was further validated through expert review, demonstrating its robustness and applicability to real-world conditions.

The oil and gas industry is characterized by a dynamic and evolving dataset, which presents challenges for long-term prediction and modeling. In addition, the industry is subject to strict confidentiality requirements, which can limit the availability of data for research purposes. Nonetheless, the present project represents a significant step forward in the establishment of advanced digital technologies for the oil and gas industry, with the potential to drive cost reductions, enhance safety, and increase efficiency in this vital sector.

Acknowledgements

We would like to take this opportunity to express our sincere gratitude to everyone who has contributed to the success of our capstone project.

First and foremost, we are immensely grateful to our supervisor, Professor Essam Shehab, for his unwavering guidance, support, and encouragement throughout the project. His expert advice and valuable insights have been instrumental in shaping our research and achieving our goals. We would also like to extend our appreciation to Appstream and Caspi Neft company for their extended assistance and support. We are grateful to have had the opportunity to collaborate with such esteemed organizations and for their willingness to provide us with the resources and information we needed to carry out our research.

We would like to thank Almat Imashev, Senior Reservoir Engineer at Caspi Neft, and Shynggys Bimagambetov, Head of the Project Management Office at Appstream, for their technical support and informative feedback from our meetings. We appreciate their assistance and support in providing us with the necessary data and helping us progress through our project. Their input has significantly improved the quality of our results, and we are grateful for their time and effort.

We would like to specifically thank Professor Dimitrios Emeris for his support and validation of our results. We also want to mention the valuable contribution of Serik Imashev, Head of Automation Engineering at KMG Engineering, in the validation of the features for the model and frameworks.

Furthermore, we would like to express our gratitude to our industrial supervisor Gani Kazbek for his constant communication and organizational skills.

Lastly, we want to express our heartfelt thanks to Caspi Neft and the School of Engineering for the provision of such a joint project. We appreciate the opportunity to work on a real industrial project, and the experience gained has been invaluable to our learning and growth.

Once again, we would like to express our deepest gratitude to everyone who has contributed to the success of our capstone project. Thank you for your support, encouragement, and valuable insights.

TABLE OF CONTENT

Abstract **3**
Acknowledgements **4**
LIST OF TABLES **7**
LIST OF FIGURES **8**
LIST OF ABBREVIATIONS **10**
CHAPTER 1: INTRODUCTION 11
1.1 Background 11
1.2 Research Motivation 12
1.3 Aim and Objectives 13
1.4 Scope of the Project 13
1.5 Structure of the Report 14
CHAPTER 2: LITERATURE REVIEW 16
2.1 Introduction 16
2.2 Maintenance Techniques 17
2.2.1 Reactive Maintenance (Corrective Maintenance) 18
2.2.2 Preventive Maintenance (Time-based Maintenance) 18
2.2.3. Predictive Maintenance (Condition-based Maintenance) 19
2.2.3.1 Physical model 20
2.2.3.2 Knowledge-based model 20
2.2.3.3. Data-driven method 21
2.3 Machine learning (ML) models for predictive maintenance of equipment 21
2.3.1 ML algorithms used for Oil and gas equipment 23
2.3.2 ML algorithms used for Oil and gas equipment 26
2.4 Oil production well and its failures 27
2.4.1 Oil well structure 27
2.4.2 Well Orientation 28
2.4.3 Well events 29
2.4.3 Oil well failures 30
2.4.3.1 Cement 30
2.4.3.2 Water cut 31
2.4.3.3 Rate of inflow 33
2.4.3.4 Rapid Productivity Loss 33
2.5 Research Gap Analysis 33
CHAPTER 3: RESEARCH METHODOLOGY 34
3.1 Introduction 34
3.2 Project Development Phases 35
3.3 Workflow of the Project 37
CHAPTER 4. CURRENT PRACTICE AT CASPI NEFT 39
4.1 Introduction 39
4.2 Company Overview 39

4.3 Processes and Operations at Caspi Neft 40
4.4 “Smart Oilfield” Project 43
CHAPTER 5. DATA ANALYSIS & PREDICTIVE MODEL CONSTRUCTION 46
5.1 Introduction 46
5.2 Identification of optimal parameters for the model 46
5.3 Data Collection and Acquisition 49
5.4 Data Pre-processing 50
5.4.1 Data source: First dataset from the oilfield 50
5.4.2 Data source: Second dataset from the oilfield 52
5.5 Development of classification and RNN model 55
5.5.1 Development of classification model 55
5.5.1.1 Validation of classification model 59
5.5.2 Development of Long-Short Term Memory 60
5.5.2.1 Validation of Long-Short Term Memory 63
Chapter 6. FRAMEWORK DEVELOPMENT 66
6.1 Introduction 66
6.2 Decision-based Framework for PdM Implementation 66
6.3 Framework for Failure Prediction in the Oil and gas industry 67
6.3.1 Failure Prediction model process 69
6.4 TO-BE analysis 71
CHAPTER 7. EXPERTS VALIDATION 72
CHAPTER 8: CONCLUSION AND FUTURE WORK 75
8.1 Conclusion 75
8.2 Limitations and Challenges 76
8.3 Further Work 77
Reference list **78**
Appendices **85**
Appendix A: Classification Models **85**
Appendix B: LSTM model **89**

LIST OF TABLES

Table 5.1 Optimal parameters influencing the well downtime…….46

Table 7.1 Experts' validation……61

Table 7.2 List of meetings with Air Astana and academic supervisors.……...62

LIST OF FIGURES

Figure 1.1: Scope of the Project……..……..13

Figure 1.2: Schematic of the Report structure…….……14

Figure 2.1: Literature Review areas…….……16

Figure 2.2: ML process…….21

Figure 2.3: SVM graphical representation……22

Figure 2.4: Random forest graphical representation…….23

Figure 2.5: Neural network graphical representation……...24

Figure 2.6: LSTM graphical representation……..25

Figure 2.7: Oil production well structure…….……...……..26

Figure 2.8: Typical orientation of oil production wells……..……..27

Figure 2.9: Well events…….……28

Figure 2.10: Excessive water production……..29

Figure 3.1: Methodology of the Project……33

Figure 4.1: West and East wings of the Ayrankol oilfield……34

Figure 4.2: Extraction of oil in the Ayrankol……...35

Figure 4.3: Preparation of the oil in the Ayrankol……36

Figure 4.4: AS-IS representation of the current maintenance initiation……...38

Figure 4.5: The architecture of "Smart Oilfield".....39

Figure 5.1: Cause-effect diagram of the oil well's most common downtimes……...…..42

Figure 5.2: Cause-effect diagram of the parameters affecting the high water cut of the well…..43

Figure 5.3: Raw parameters of the production wells before and after the preparation process(light space depicts the absence of the data).....45

Figure 5.4: Categories of production wells' downtime types……...46

Figure 5.5: State of production wells parameters after preparation……..……...48

Figure 5.6: The heatmap of the second dataset features…….……..49

Figure 5.7: Individual and cumulative variance described by PCA…….……55

Figure 5.8: Principal Comonent Analysis…….……56

Figure 5.9: Preliminary normalization of k-means clustering…….…….58

Figure 5.10: Results of the 3 classification models…..…….…….……...60

Figure 5.11: LSTM layers in the model architecture……...…….61

Figure 5.12: Summary of the model with layers……..…….…….…….62

Figure 5.13: Performance graph of LSTM model……...……..…..……..63

Figure 5.14: Actual vs Predicted data for 200 days……...…….……..63

Figure 5.15: K-fold cross-validation result…….……...…….……..64

Figure 6.1: Decision-based Framework for Predictive maintenance model selection…….53

Figure 6.2: Framework for failure prediction in the oil and gas industry……54

Figure 6.3: Flowchart of the failure prediction model…….56

Figure 6.4: TO-BE representation of the maintenance initiation in the company……57

LIST OF ABBREVIATIONS

ANN	Artificial Neural Network
APMS	Automatic Plant Matery Station
BSW	Basic Sediment and Water
CM	Corrective Maintenance
CBM	Condition-based Maintenance
CPS	Cyber-Physical System
ES	Expert System
ESP	Electrical Submersible Pumps
FL	Fuzzy Logic
I/O	Input/Output
IoT	Internet of Things
LSTM	Long Short-Term Memory
ML	Machine Learning
NDA	Non-Disclosure Agreement
NLP	Natural Language Processing
NN	Neural Network
OCP	Oil Collection Point
OEM	Original Equipment Manufacturer
OFM	Oilfield Manager
PCA	Principal Component Analysis
PdM	Predictive Maintenance
PM	Preventive Maintenance
SVM	Support Vector Machine
SEM	Submersible Electrical Motor
TBM	Time-based Maintenance
WPO	Well Optimization Portfolio

CHAPTER 1: INTRODUCTION

1.1 Background

The goal of each nation all over the world is to become a thriving and economically stable country. In order to achieve this target one of the ways is to effectively and efficiently use the resources of the state. The Republic of Kazakhstan is rich in natural resources like coal, oil, and natural gas. The country has proven on-shore oil fields mostly located in the west Kazakhstan region, managing for current time 172 oil and 42 gas fields near the Caspian Sea (Karatayev and Clarke, 2014). Specifically, the oil and gas industry provides a considerable portion of the national GDP. Hence, the oil and gas industry is one of the bases of Kazakhstan's economy (Kazakhstan Energy Sector Review, 2022). However, despite the importance of the industry and global development trends, digital transformation has not been achieved yet. As a result, the country's dynamic industry omits crucial benefits of digital technology, like cost reduction, resource utilization reduction, increased efficiency, reduced downtimes, and lowered risks (Bousdekis et al., 2019).

In order to adequately respond to global digital development and provide value-producing further opportunities, Kazakhstan oil and gas companies need to take action for achieving full digital transformation. The country's prospective industry nowadays is presented by many companies. And one of the companies that aimed to be the pioneer of doing business in a more efficient way is Caspi Neft.

Caspi Neft is a leading oil-producing company operating in the Caspian region, which since 1997 has been engaged in a comprehensive study of the subsoil, development, exploration, and production of hydrocarbons as well as its storage and export. Nowadays, Caspi Neft faces the problem of a decrease in oil production due to the depletion of the oilfield. Consequently, this only strives the company for prompt and resolute digital solutions. The digital transformation started with the cooperation of Kazakhstani IT consulting and software development companies, and now Appstream is responsible for current digital development changes. In addition to this, Caspi Neft undergoes the process of digital transformation to successfully accomplish the project "Smart Oil Field" (Galushko, 2023). The next stage of this project is to introduce a failure prediction framework in the company as this action is proved to be an effective way to reduce costs and increase the efficiency of the oilfields.

1.2 Research Motivation

In the oil and gas sector, breakdowns and equipment failures can have far-reaching economic, safety, and environmental consequences. For instance, the repair of major rotating equipment damage can take months or even years, leading to significant losses in income. Bevilacqua and Braglia (2000) found that up to 70% of the industry's total production costs are associated with maintaining rotating machinery. As such, selecting the best equipment maintenance plan is essential for fault detection and preventing output interruption.

Traditional periodic maintenance is a common approach in the industry, whereby equipment is checked at set intervals to ensure it is in good condition. However, this method is not foolproof, as unexpected equipment failures can occur at any time and cause unnecessary shutdowns (Susto et al., 2014). To address this limitation, the industry has embraced condition-based maintenance (CBM), which uses sensors to monitor equipment health while it is in use (Figueroa Barraza et al., 2022). This predictive maintenance technique enables maintenance to be scheduled in a manner that minimizes disruption and optimizes costs by assessing equipment criticality and risk.

To advance digitalization in the industry, Ngu et al. (2019) identified several key enablers for fault prediction and maintenance. These include the application of smart sensors, the creation of Internet of Things (IoT) systems that can process vast amounts of data, and enhanced connectivity through the use of wireless systems and remote viewing of process operating data and equipment health.

Although numerous approaches to predictive maintenance have been reported in the literature, most studies have focused on fault detection in specific rotary equipment. For example, Abbasi et al. (2018) developed a model for gas compressor failure detection using Multiple Linear Regression, while Orru et al. (2020) employed machine learning techniques for predicting faults in centrifugal pumps. Additionally, Li et al. (2019), Qian et al. (2018), and Zhang et al. (2019) developed machine-learning algorithms for diagnosing roller-bearing faults. However, there is a dearth of detailed information on the parameters required for analyzing failures in oil wells, particularly in the context of specific oil fields such as those in Kazakhstan.

Therefore, this study aims to develop and discuss a fault prediction model for the entire oil well system in the Ayrankol oilfield. The paper proposes a framework for implementing the model, considering all necessary requirements. This research fills a critical gap in the literature, and its findings could significantly enhance fault prediction and maintenance in the oil and gas industry, particularly in Kazakhstan.

1.3 Aim and Objectives

The main aim of this Capstone Project is to develop a framework for the failure prediction of equipment in the oil and gas industry. To achieve the main aim of the project, the overall objectives are to:

- Conduct a comprehensive literature review on predictive maintenance in the oil and gas industry as well as advanced statistical methods utilized;
- Identify the critical parameters for wells failure prediction;
- Capture the current capabilities and activities of Caspi Neft company (AS-IS analysis) to access the ability to adapt the failure prediction model;
- Develop the prototype of the failure prediction model based on real data from the oilfield;
- Develop a framework for the implementation of equipment failure prediction in oil and gas companies;
- Validate the model through cross-validation and experts' judgments.

As a result of the Capstone Project, Caspi Neft as a first stakeholder received the failure prediction model, which was trained on their data and can be applied to the Ayrankol oilfield. At the same time, other Kazakhstani and not only companies may also adopt the outlined framework as a step of predictive maintenance.

1.4 Scope of the Project

The scope of the project consists of achieving a basic understanding of the oil and gas industry and the need for intervention in the performance of one of its main assets. And the main deliverable is to create a machine-learning model for the undesirable well events in the oil field. For achieving this the next core knowledge is needed: vertical orientation of the wells, common types of failures, and finding optimal parameters influencing the undesirable well events. Besides the industries' knowledge, data mining, and processing techniques, a machine learning model and its validation will be created. However, the programming part and machine learning models for the separate parts of the oil field other than the wells will not be delivered. In addition, the economics and implementation of this model by the company can not be controlled during the project performance.

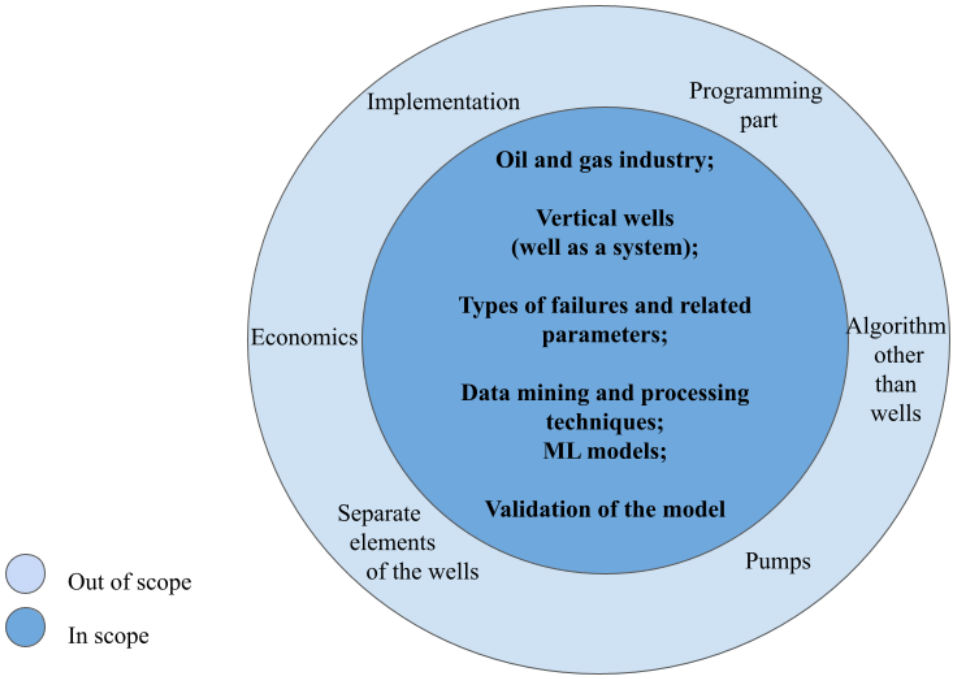


Figure 1.1: Scope of the Project

1.5 Structure of the Report

This report is organized into seven chapters, and its structure is shown in Figure 1.2. The current Introduction Chapter includes a background of the company, a problem statement and relevance, scope, aim, objectives, and outcomes of the project. The comprehensive literature review with areas, information on oil and gas wells, data processing techniques, and machine learning models are presented in Chapter 2 of the current report. Furthermore, Chapter 3 is devoted to the methodology of the project, including stages and the project's flow chart. The next Chapter 4 is dedicated to the current practice of Caspi Neft connected to the operations, maintenance procedures, performed projects, and data gathered and stored in the company. AS-IS analysis of the current practice is also presented in Chapter 4. Accounting for the outputs of the previous sections, Chapter 5 outlines the data acquisition and its processing, feature engineering, and machine learning models applied to the processed data. One of the final sections is Chapter 6, where the framework for the implementation of failure prediction of the equipment is described. In addition, the validation of both model outcomes and the framework is included in Chapter 6. Chapter 7 is dedicated to the validation of the model and framework. Finally, Chapter 8 draws the conclusions of the project with future work prospects.

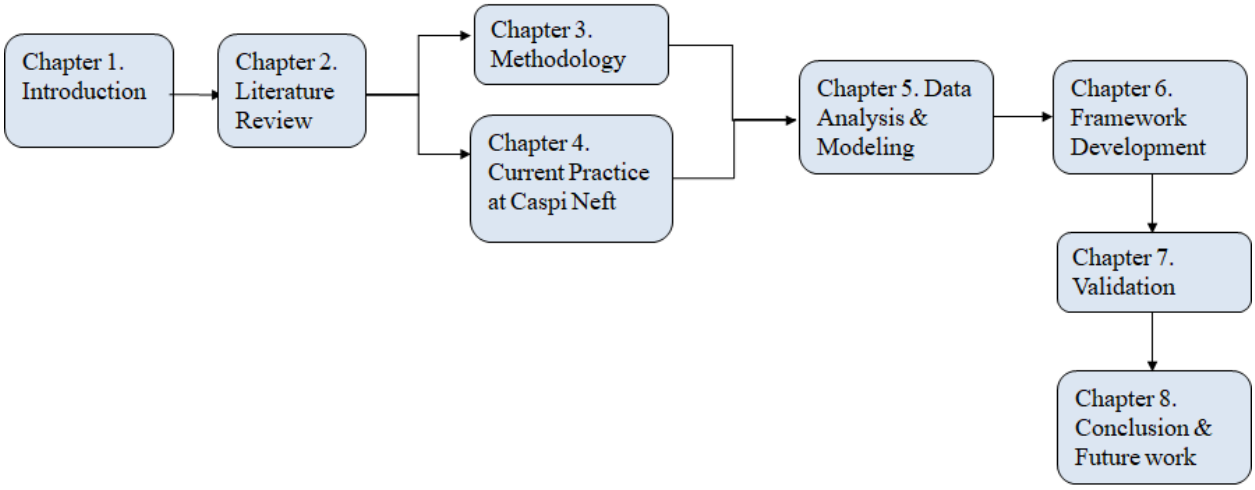


Figure 1.2: Schematic of the Report Structure

CHAPTER 2: LITERATURE REVIEW

2.1 Introduction

The oil and gas sector is highly reliant on the efficient operation of industrial machines, and any equipment failure can result in significant financial losses and safety hazards. Therefore, maintenance activities are crucial to ensure the smooth running of equipment and minimize downtime. Industries are increasingly employing predictive maintenance to optimize maintenance tasks, which involves the application of machine learning algorithms to anticipate the possibility of equipment failure. This section provides an in-depth exploration of maintenance strategies executed in industrial settings, including preventive maintenance, corrective maintenance, and condition-based maintenance. Moreover, the section briefly summarizes the reasons for failure that appear in oil and gas equipment and provides the optimum parameters for failure prognosis. It also examines the application of machine learning algorithms in predictive maintenance and how they affect the procedure for decision-making.

The researched literature considered in the present work was based on a comprehensive search of the academic databases Google Scholar and Scopus, and various keywords such as "failure prediction models," "predictive maintenance" and "preventive maintenance," and "ML algorithms for predictive maintenance" were used for finding the most appropriate research studies. The following subsection aims to provide an informative summary about the current state of predictive maintenance, particularly failure prediction in the oil and gas industry, as well as its potential benefits for improving equipment reliability and lowering operational costs, failure parameters, and recent advances in using ML algorithms for these purposes. The literature review areas are presented in Figure 2.1.

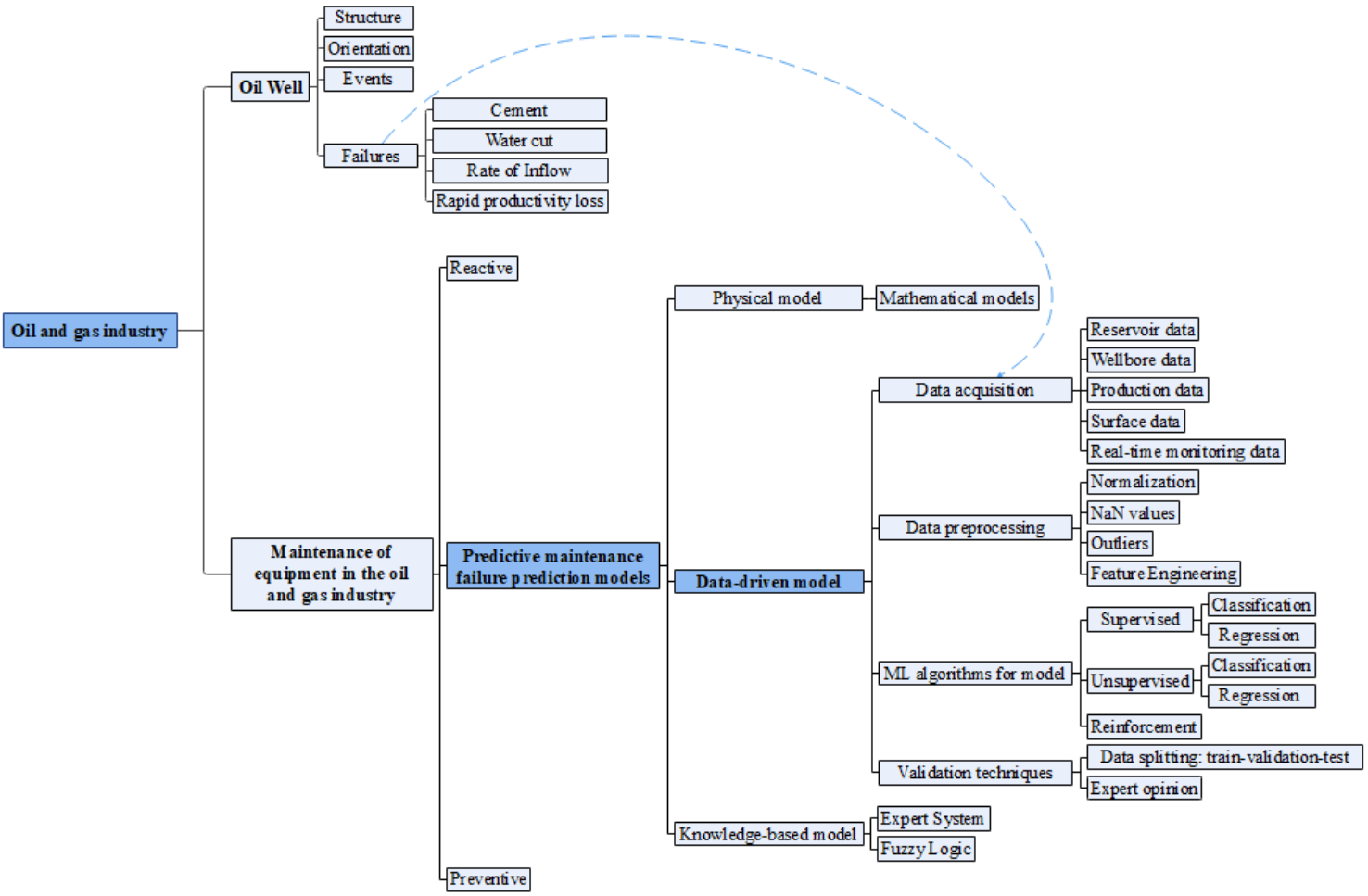


Figure 2.1: Literature Review areas

2.2 Maintenance Techniques

Maintenance plays a critical role in the oil and gas sector, where equipment reliability is important to guarantee an uninterrupted and risk-free operation of facilities,. Unplanned downtimes can cause substantial financial losses for oil and gas companies. A study found that an average of 3.65 days of unplanned downtime per year, which is equivalent to 1% of production facility downtime, can result in a loss of $5.037 million for these companies (Baker, 2016). To minimize facility downtime, it is necessary to conduct regular maintenance. Maintenance can be divided into three types: reactive, preventive, and predictive maintenance. The employment of reactive and proactive maintenance approaches in the oil and gas industry, as well as the impact they have in avoiding downtime and maximizing equipment reliability, is investigated in the following section of the literature study.

.

2.2.1 Reactive Maintenance (Corrective Maintenance)

Repairing machinery after a failure occurs is known as reactive maintenance, also referred to as Corrective Maintenance (CM) (Wanasinghe, 2020). Unplanned downtime is a common feature of this strategy, which can also lead to higher repair costs and output losses. Reactive maintenance is typically used in situations where it is not feasible or cost-effective to perform preventive maintenance or when the equipment failure rate is low (Duffua et al., 2001). However, relying solely on reactive maintenance can lead to increased downtime, production losses, and safety hazards, and it can also negatively impact the overall reliability of equipment (Blanchard, Verma and Peterson, 1995). Therefore, oil and gas companies should aim to implement a balanced maintenance approach that combines reactive and preventive maintenance to ensure equipment reliability and minimize downtime (Elwerfalli and Al-Maqespi, 2021).

2.2.2 Preventive Maintenance (Time-based Maintenance)

The PM (Preventive Maintenance) strategy is an alternative to the CM (Corrective Maintenance) strategy (Ahmad and Kamaruddin, 2012). Whereas the CM strategy involves repairing equipment only when it fails, the PM strategy involves regular inspections and maintenance to prevent potential failures before they occur (Alazemi et al., 2019). The purpose of PM is to reduce downtime and repair costs by addressing possible issues before they become major difficulties.

Preventive maintenance (PM) can be done in the industry through two methods: experience or original equipment manufacturer (OEM) guidelines, both of which are scientific in origin. The experience-based method is a commonly used PM approach that is carried out at periodic times (Sheu et al., 1995). In this method of implementing preventive maintenance, there are no fixed prescribed procedures. As a result, the experience and expertise of maintenance engineers and technicians become a valuable asset to the organization. Their experience in identifying potential issues and performing maintenance activities helps to ensure the optimal performance of the equipment and reduce the risk of failures.

The OEM recommendation-based method of implementing preventive maintenance involves carrying out maintenance at fixed intervals, such as every 1000 hours or every 10 days, according to the manufacturer's manuals. This strategy, however, may not be appropriate when the goal is to minimize operational expenses while maximizing the performance of machine (Adenuga, Diemuodeke and Kuye, 2023). This is because the manufacturer's suggestions may not take into consideration the equipment's individual operating conditions or the company's maintenance

objectives (Labib, 2004; Tam, Chan and Price, 2006). As a result, a more customized approach to PM may be required to optimize maintenance schedules and reduce costs.

2.2.3. Predictive Maintenance (Condition-based Maintenance)

Predictive maintenance (PdM), also known as Condition-based Maintenance (CBM), is a prominent and up-to-date maintenance technique that has received a lot of attention in the literature (Moya, 2004; Jardine, Lin, and Banjevic, 2006; Ahmad and Kamaruddin, 2012). CBM was launched in 1975 as a technique to maximize the effectiveness of PM (Preventive Maintenance) decision-making, according to Ahmad and Kamaruddin (2012). CBM involves continuously monitoring the state of equipment using techniques including vibration analysis, oil analysis, thermography, and other non-destructive testing procedures. CBM can detect possible difficulties and forecast when maintenance is required by regularly tracking the status of the equipment, allowing maintenance activities to be undertaken just when necessary or shortly before the breakdown (Abbasi, Lim and San Yam, 2019). This approach reduces maintenance costs and equipment downtime while improving reliability and extending the equipment's lifespan.

According to Wanasinghe (2020), a predictive maintenance system typically comprises of a number of significant components, including a sensor network for data acquisition, algorithms for data processing, a data storage system, and a human-machine interface for operators to interact with and observe the insights. This framework can be used in an Internet of Things-based CPS (Cyber-Physical System) to collect data from IoT-enabled smart sensors. The acquired data can subsequently be analysed utilizing edge or fog computing systems, enabling quick decision-making based on data to be made in order to avoid possibly fatal failures. By using this approach, maintenance can be performed more efficiently and effectively, reducing downtime and increasing overall equipment reliability.

Peng, Dong and Zuo (2010) identified that in a CBM program, diagnostics and prognostics are two crucial components. Diagnostics focuses on detecting, isolating, and identifying faults when abnormalities occur in the system. This approach contributes to determining the core cause of an issue, allowing maintenance personnel to take corrective action promptly and reduce downtime. Prognostics, on the other hand, is concerned with predicting flaws and damage prior to occurring. This enables maintenance staff to take proactive measures such as repairing or replacing components before they break. Furthermore, the authors divided the prognostic model into three categories: physical, knowledge-based, and data-driven.

2.2.3.1 Physical model

Physical model-based approaches often employ mathematical models that are related to physical processes which have a direct or indirect impact on the status of the components being monitored (Peng, Don, and Zuo, 2010). These physical models are developed by subject matter specialists, and the parameters utilized in the models are validated using large data sets. Physical model-based prognostic approaches necessitate specific mechanistic knowledge and theories relevant to the systems being studied (Luo et al., 2003). By employing physical model-based approaches, it is possible to gain an in-depth knowledge of the monitoring processes and to make more accurate predictions about their condition and performance (Eghbali, Ayatollahi and Boozarjomehry, 2016). Physical model-based techniques have the disadvantage of being more expensive and specialized to specific components, limiting its applicability to a wider range of equipment (Brotherton et al., 2000). In addition, building an accurate physical model can be very challenging.

2.2.3.2 Knowledge-based model

Knowledge-based techniques have emerged as a possible alternative to physical-based models. Expert systems (ES) and fuzzy logic (FL) are two common knowledge-based techniques.

ES, which has been in use since the mid-1960s, is a computer system designed to solve problems that are typically managed by human personnel (Biagetti and Sciubba, 2004). It can be viewed as a system that demonstrates expert knowledge in a specific topic, and its performance may be assessed through the combination of computing power with logical reasoning. Human experts' knowledge is stored in computers as rules, which are then employed by the ES to develop answers by emulating human thinking and inference (Hassannayebi et al., 2021). The process of acquiring expertise in a specific area and translating it into a set of rules that an expert system can use is challenging. Furthermore, ES are constrained in their ability to deal with new scenarios that are not explicitly listed in their knowledge databases.

Fuzzy logic (FL) is known as a problem-solving approach which enables the derivation of definite conclusions from imprecise, noisy, missing, or ambiguous input information (Senouci, El-Abbasy and Zayed, 2014). It is utilized across a wide range of systems, including small microcontrollers and large data acquisition and control systems. Unlike traditional discrete values, FL utilizes fuzzy sets and continuum mathematics to model system behavior, which provides a balance between rigorous analytical modeling and qualitative simulation (Alakbari et al., 2021). FL utilizes linguistic variables to represent and reason with incomplete or inaccurate information, making it intuitive and human-like. Additionally, FL offers extensive simulation capabilities.

2.2.3.3. Data-driven method

Data-driven prognostic methodology is a type of prognostic technique that utilizes data-driven models to predict the failure of equipment (Mazumder, Salman and Li, 2021). The data acquired from the system is utilized to develop a model that can predict the system's future behavior in this manner. These models are generated utilizing statistical and machine learning algorithms based on pattern recognition theory, such as regression modeling, time series evaluation, neural networks, and support vector machines (SVM).

The primary advantage of the data-driven prognostic methodology is its ability to handle complex systems with large amounts of data (Sircar et al., 2021). These models can capture the nonlinear relationships between system variables, which are often difficult to model analytically. Furthermore, data-driven models are adaptive and can update their predictions based on new data, which is essential for real-time prognostics.

However, the accuracy of data-driven models is determined by the quality and quantity of data utilized to train the model, as well as the model used (Wang et al., 2023). Insufficient or biased data can result in inaccurate predictions. Therefore, data preprocessing and feature selection are crucial steps in the data-driven prognostic methodology. The next subsections are aimed to identify the optimum parameters for equipment failure prediction data-driven model and the review of machine-learning algorithms used for these purposes.

2.3 Machine learning (ML) models for predictive maintenance of equipment

One of the most important tools of data-driven models is ML, the general principles of which as well as distinct models are discussed in this section. Also, the application of ML algorithms in the oil and gas industry is considered.

Machine learning is a type of artificial intelligence that allows systems to learn, improve, and predict outcomes automatically from their own experience without explicit programming. Basically, machine learning includes training on the datasets of features and respective outputs, which means the identification of correlations and patterns in the data and making adjustments to the features. As a result of machine learning, the trained model is obtained that in turn is able to predict the outcomes from the new data input. The schematic representation of the ML process is presented in Figure 2.2.

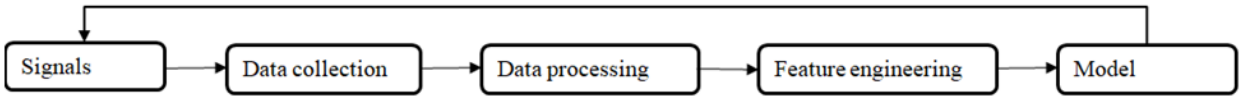


Figure 2.2: ML process

Generally, three types of machine learning exist: supervised, unsupervised, and reinforcement learning. Supervised learning implies that an algorithm is trained on a dataset with labels on how to map input data (features) to an output (Delua, 2021). The algorithm learns to generalize from the I/O (input/output) examples during the training phase in order to predict the output for the new input dataset. Supervised learning is of special importance in natural language processing (NLP), classification of image, recognition of speech, and predicting customer behavior in marketing.

Unsupervised learning is usually utilized in those cases when the labeled data, such as I/O pairs are not available or scarce as the algorithm finds its own structure in the data provided to it. There are various types of unsupervised learning algorithms like clustering (grouping of similar examples), dimensionality reduction (reducing the number of features as much as possible not to lose information), and anomaly detection (detection of unusual examples in the dataset) (Delua, 2021). Although unsupervised learning can be more challenging in implementation and evaluation than supervised learning due to the absence of functions to optimize, it found to be extremely useful in numerous fields, such as market segmentation, image segmentation, and anomaly detection in cyber security.

Reinforcement learning is aimed at maximizing the cumulative reward of the system; an agent interacts with the environment and feedback in the from of penalties or rewards depending on the agent's actions is provided. Generally, there are two types of reinforcement learning algorithms: model-based which predicts the next point, and reward based on the current state, and model-free, which is trained by experience. Reinforcement learning is particularly important to solve complex problems, for example, in the game of Go, game playing, robotics, and control systems (Wuest et al., 2021).

2.3.1 ML algorithms used for Oil and gas equipment

In this section, some of the most common ML algorithms utilized in the scope of predictive maintenance in oil and gas will be discussed.

The first one is a classical supervised learning algorithm utilized for classification and regression purposes – support vector machine (SVM). The concept of SVM is based on finding the hyperplane (maybe a line or a plane), which will separate the data point into different classes. That hyperplane should maximize the margin (difference) between the closes data points from various classes, so the larger the margin, the better SVM works. One of the benefits of SVM is the ability to handle many features and large amounts of data but with a small number of training examples (Shrivastava et al., 2010). The prominent fields of SVM application are text classification, image classification, bioinformatics, and finance.

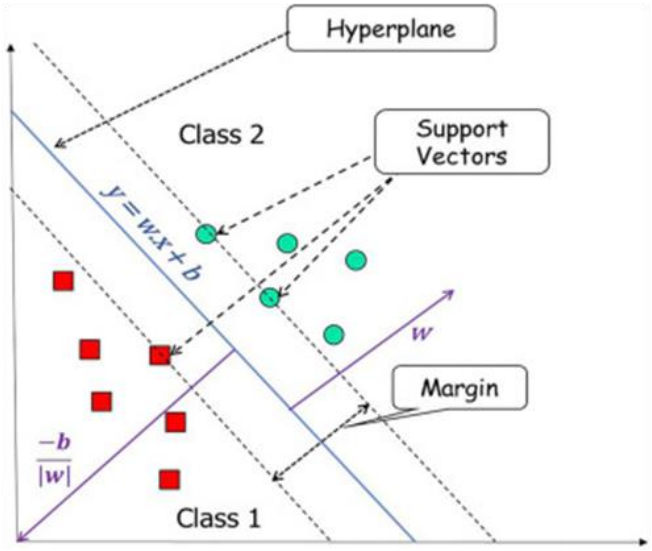


Figure 2.3: SVM graphical representation (Rani et al., 2022)

Another popular ML algorithm widely used for predictive maintenance purposes (both classification and regression) is Random Forest, which is basically the advanced form of the Decision Tree algorithm. Random Forest is a group of decision trees, and each tree is represented by the subsets of training data. This large number of decision trees contributes to the large number of "best split" nodes, which are determined by Gini impurity or information gain. When it comes to the prediction of the output based on the new data point, the major vote from individual trees from an ensemble is taken. Compared to a Decision Tree, Random Forest allows for reducing overfitting due to a large number of trees and increasing the accuracy of the predictions. Due to its nature, Random Forest can handle high-dimensional datasets with a large number of features and diminish the effect of missing data points and outliers. Random Forests found its wide application in various fields such as finance, healthcare, and ecology.

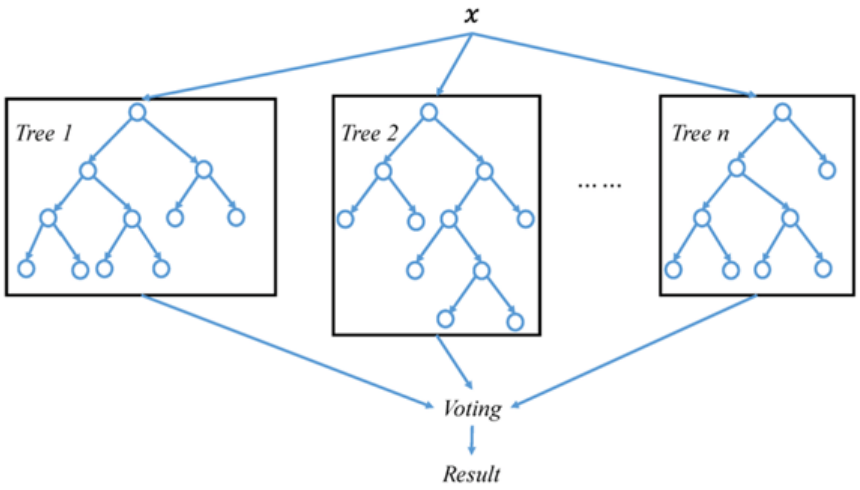


Figure 2.4: Random forest graphical representation (Wang et al., 2019)

A neural network (NN) is a variant of machine learning algorithm, which simulates human brain functions. The component of NN is interconnected nodes forming layers, weights, and biases. The input data enters the first layer of the model, where it is mathematically manipulated. Then this refined input enters the following layers of the model until it reaches the last and can be called an output. The neural network algorithm is designed according to backpropagation, meaning that biases and weights in the layers are adjusted during the training process by calculating the difference between the current state and the desirable output (Carvalho et al., 2019). The particular importance of NN algorithms comes when the relations between input and output data are not clear, complex, and non-linear, however, the datasets for NN should be quite dense. On the other hand, it provides the benefit of being able to process large and complex datasets. NN found their application in solving many difficult tasks, such as image and speech recognition, although they are highly prone to overfitting.

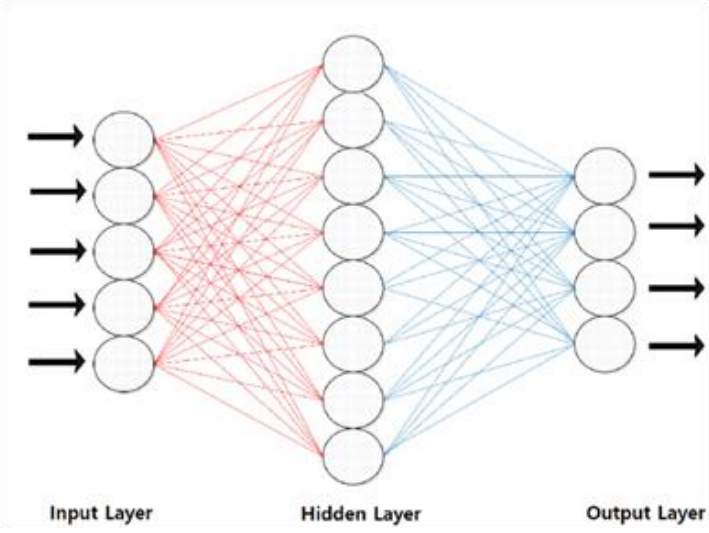


Figure 2.5: Neural network graphical representation (Huh et al., 2020)

Long Short-Term Memory (LSTM) is a type of the recurrent neural network, which possesses improved memory compared to RNN; it also, solves the issue of too small useful gradients propagating from the output of the model to the layers of input, thus eroding the update of the weights, consequently leading to the poor performance of the model (Lindemann et al., 2021). Moreover, as it comes from the name of LSTM, it is capable to capture the long-term dependencies of the sequential data and manipulate this data due to its complex architecture, which consists of input, forget, and output gates. The input gate controls the new data inserted into the algorithm, while forget gate in LSTM can decide whether the data should be kept or deleted, so only essential data is stored and processed further in this model. Finally, the output gate estimates the impact of the data point on the output (Siami-Namini, Tavakoli & Siami-Namin, 2020). Overall, this mechanism allows LSTMs to selectively delete or keep data at each time step, thus modeling long-term dependencies in sequences. Due to its unique ability to model complex patterns, LSTM found its wide application in various fields, such as time-series prediction and state-of-the-art deep learning algorithms.

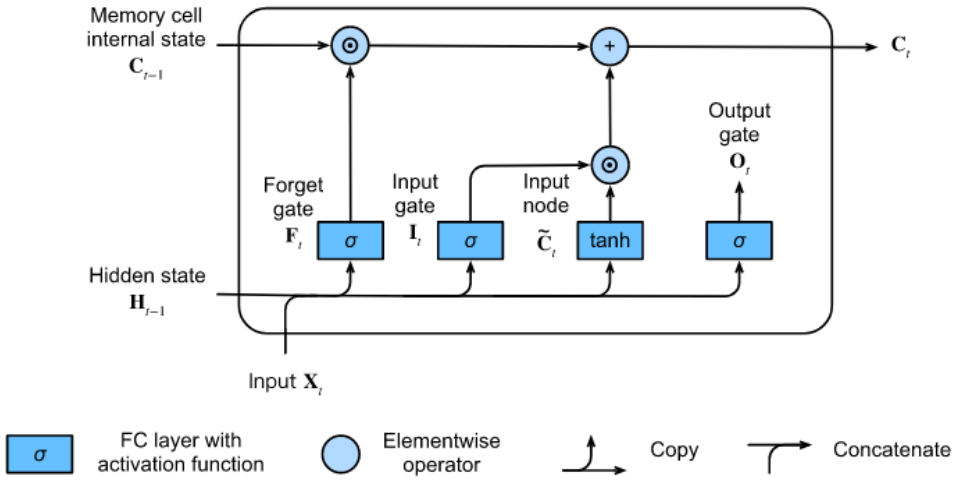


Figure 2.6: LSTM graphical representation (Dive into Deep Learning, 2022)

2.3.2 ML algorithms used for Oil and gas equipment

It is clear that various types of ML with their algorithms have found a wide application in solving difficult problems in different areas due to their strong pattern recognition, robustness, and adaptability. Predictive maintenance in the oil and gas industry is not an exception. Overall, the rotating equipment, especially represented by pumps in the oil and gas industry is a common topic for various research papers. For instance, the paper by Abbasi et al. (2019) summarizes the utilization of ML algorithms under the knowledge-based approach for predictive maintenance of the rotating equipment, namely these methods are the k-nearest neighbor (k-NN) algorithm, Bayesian classifier,

support vector machine (SVM), and artificial neural network (ANN). At the same time, Abdalla et al. (2022) predicted failures in electrical submersible pumps (ESPs) by analyzing real-time sensor data using an unsupervised learning technique called extreme gradient boosting trees (XGBoosting), which gives an accuracy of 0.71 on the test set and the signals 7 days in advance of the failure. Cline et al. (2017) experimented with various ML algorithms to predict the failure of several components of two families of oil and gas equipment from downstream exploiters based on the inspection datasets. The utilized predictive analytics algorithms included linear regression, logistic regression, neural networks, decision trees, random forest, and gradient boosting machines; as a result, 61% of the connector failures were identified by the implementation of ML. However, limited research is conducted on the equipment from the oil and gas industry other than pumps or rotating equipment. One of the research papers is dedicated to the predictive maintenance of offshore wells, which was constructed on the publicly available dataset by means of deep learning feature extraction and the following ML algorithms: Random Forest, Nearest Neighbours, Gaussian Naive Bayes, and Quadratic Discriminant Analysis (Gatta et al., 2021).

2.4 Oil Production Well and its Failures

In order to properly implement the above-mentioned machine learning models and get the highest accuracy with beneficial validation results, a basic understanding of the oilfield is required. To be more specific, the required knowledge area includes the well construction, its orientation, well events, delving into the failures,and briefly introducing the parameters affecting the downtimes that will be implemented into the machine learning algorithms.

2.4.1 Oil well structure

An oil production well is a hole drilled into the Earth. And the purpose of each oil well is to produce oil or gas to the surface. Production wells almost always bring petroleum products, including some natural gas and water. The main elements of the well are casing and wellbore. The casing is a metal pipe embedded into the borehole to avoid the problem of collapsing. As can be seen in Figure 2.7, all wells have a borehole that's entered into the subsurface, a casing of a well that's embedded into the gap, and a special perforated zone inside the store that permits the oil/gas to come to the well casing and be extricated (Pyramid Environmental, 2015; energyeducation.ca, n.d). Oil wells have a more complex pumping framework related to extricating petroleum, and they are by and large penetrated much more profoundly than water wells. Furthermore, they regularly will have numerous layers of casing encompassing the well that expand to diverse profundities in arrange to get the most extreme entering required to reach the arrangement including the petroleum store and also permit

both liquids and gas to move through the well framework (including conductor, surface, middle and production casing).

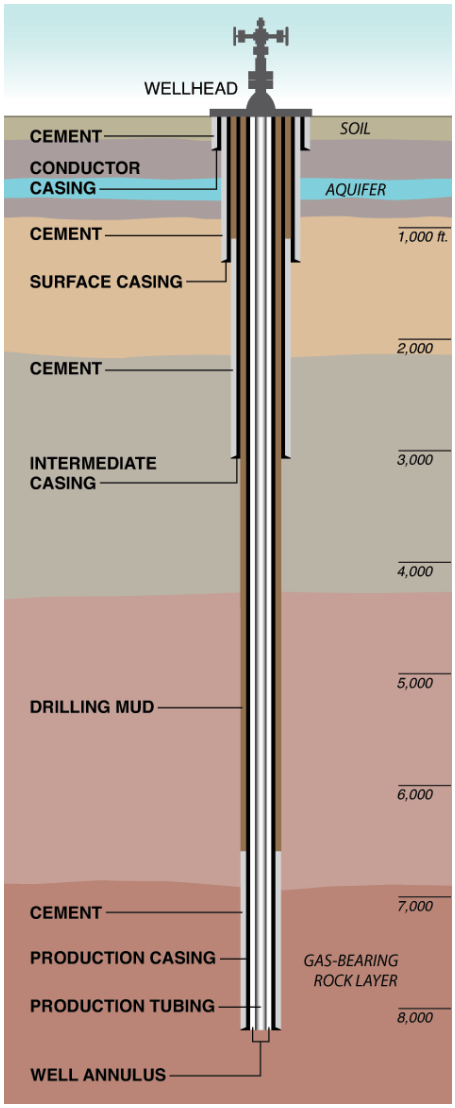


Figure 2.7: Oil production well structure (Pyramid Environmental, 2015)

2.4.2 Well Orientation

Oil production wells are one of the most important assets of the oil industry. The main goal of these wells is oil production or exploitation. The classification of wells primarily depends on the conditions of geology (King, n.d.). Based on the deviation angle of the barrel from its vertical axis, there are 3 main types as shown in Figure 2.2: horizontal (comprising approximately 90 degrees), deviated (in excess of 5 degrees) and vertical (5 degrees maximum).

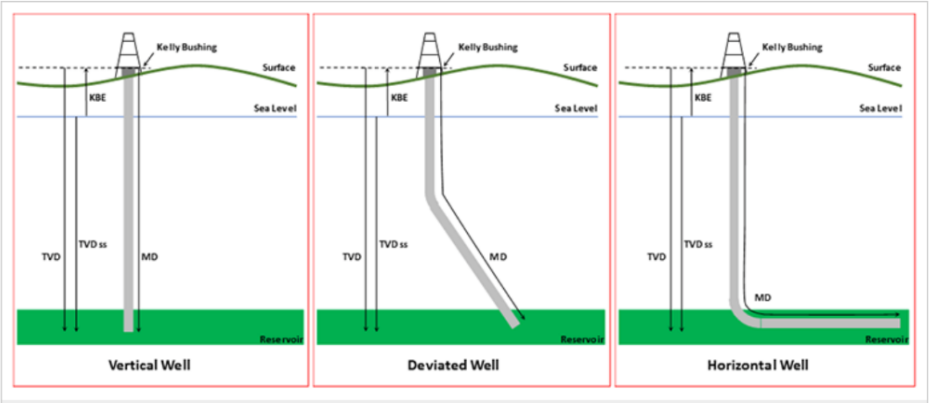


Figure 2.8: Typical orientation of oil production wells (King, n.d.)

From the early beginning of the oil industry, the most common on-shore wells have been vertical wells, despite the current classification. This is due to their simplicity and economics. In addition, due to their simple design, vertical wells are a rather budget option for drilling. However, nowadays wells are designed horizontally (Sciencealpha, 2019).

In spite of well orientation, the productivity of oil production wells depends on the execution of a subsurface equipment system that may fail with the presence of water, sand, corrosion, temperature or pressure variations and other outside factors. In addition to this, one or even several failures may occur during a well's performance and in these cases, losses may be significant and as a consequence expensive, as corrective maintenance processes must be performed only after the production interruption (Aalsalem, et al., 2019; Madrid and Min, 2020).

2.4.3 Well events

Well event is a change of a well parameter over time. This event changes the performance of the well. Different modes of well behavior can be described by a set of behavioral parameters (measured by different sensors). Primarily, well events are divided into 3 groups: expected events, unexpected events (failures) and noise. Expected events include changes in signals based on a known basis, for example: planned technological operations and daily changes (fluctuations) of parameters, like temperature or pressure. On the contrary, unexpected events or anomalies are changes affecting known parameters, these may include equipment failure. And the noise (or symptoms of other possible anomalies) is the change that does not primarily relate to the well or reservoir event (Elichev et al., 2019). Besides the described division, well events also can be divided into 3 additional groups, as shown in Figure 2.9**.** This classification also includes parameters of change of the formation, fluctuations inside the well or the well' failure.

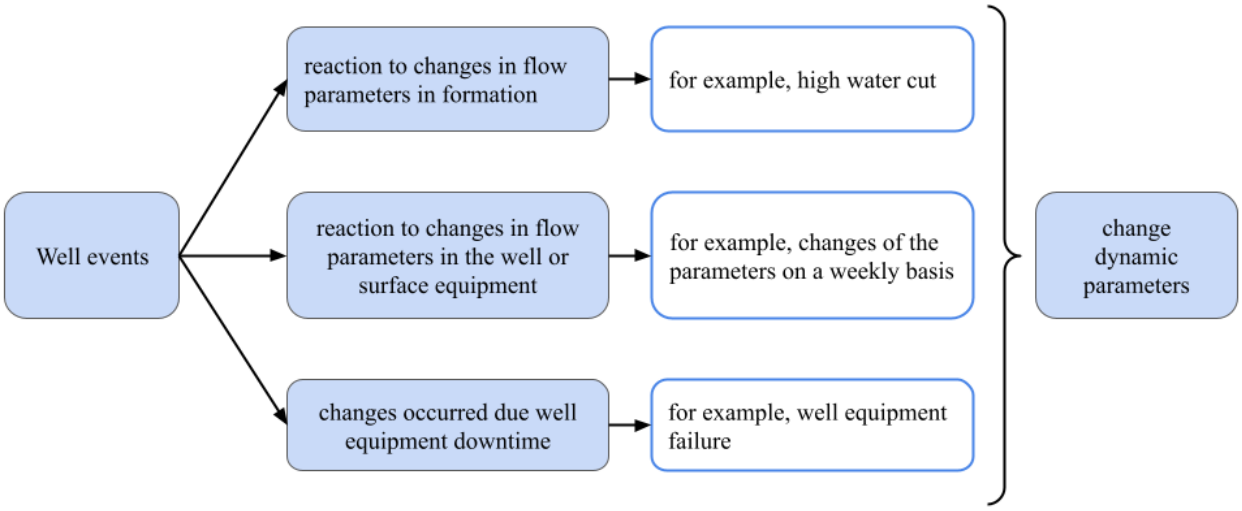


Figure 2.9: Well events

It is worth noting that according to the practice from many fields, not all events are recorded, furthermore only considerable events that can affect the production are noted in the reports. On one hand, recording all normal "healthy" or minor operations do not lead to considerable advantages. However, from another side, the failure identification by above described machine learning algorithms must be constructed on all kinds of data including expected, unexpected events and even noise. Generally, several parameters are monitored and collected from the oil field sensors with high frequency: wellhead pressure, bottom-hole pressure, flow line pressure and, also temperature.

Hence the objective of classification of all kinds of events is also required. In order to create well events prediction models, raw data and expert knowledge data are needed.

2.4.3 Oil well failures

As it was previously mentioned, oil well, as one of the important industry assets, requires frequent monitoring in order to work for a longer time without downtimes to expand the lifetime of wellhead industrial equipment and avoid or decrease considerable financial losses. Since it is challenging to conduct well maintenance activities and control every well in person by arranging engineers, new smart and intelligent oil fields can unravel common failures. Considering the list of these possible problems, water cuts and cement are the most common and widespread.

2.4.3.1 Cement

Besides the water cut, another common failure of wells is cementing, which refers to mechanical failures. Well cement is a significant process in the penetrating and completion of oil wells. Its primary purpose is to maintain the integrity of the well behind the casing and give the opportunity for continuous zonal isolation to prevent security threats and environmental issues (Tahmourpour and Griffith, 2007; Shadravan and Amani, 2012; Wu et al, 2022). However, there are various factors that can not positively influence the integrity, such as mud channeling along the cement sheath (which can act as an enabler for hydrocarbon leaks) and casing centralization. In addition to this list, there are also casing perforations, fracturing, stimulations, and acid treatments that can lead to the failure of cementing including changes in the wellbore state during well production (Shadravan and Amani, 2012).

Changes in temperature between casing-cement-formation can also result in different failure modes that negatively affect the integrity of the set cement sealing. Additionally, changes in temperature and pressure throughout the well's life cycle can create various mechanisms of failure that could potentially compromise the set cement sealing integrity. Specialists highlight that downtimes of the cement sheath generally appear because of the well pressure or temperature fluctuations. (Tahmourpour and Griffith, 2007). However, despite well cementing being a common type of well failure, controversially the cement failure can not be predicted based on the current parameters from the well sensors, except by measuring the pressure and temperature parameters' cumulative influence.

2.4.3.2 Water cut

Another important parameter in reservoir engineering is water cut and its high rate is one the most common downtime on the oilfields. Primarily, water is the main element of every stage of the oilfield, however, no operator wants to produce it on the oilfield (Bailey, et al., 2000). They are interested in how to save on expenses and spend less on water loading. According to statistics, it was calculated that daily approximately 75 million barrels of oil production are produced together with 210 million barrels of water globally, this is about one-third of the daily production of oil (Aalsalem, et al., 2019). Unfortunately, there is no option to control the level of water. When the water level exceeds the acceptable rate (referring to the parameter of the water cut percentage) engineers close the well.

However, nowadays, water produced with crude oil has increased rapidly in the oil fields because of formations of high-conductivity thick sandstone, as well as in some stratified formations,

as hydraulic fracturing, is considered the most common. The process of water cut is illustrated in Figure 2.10.

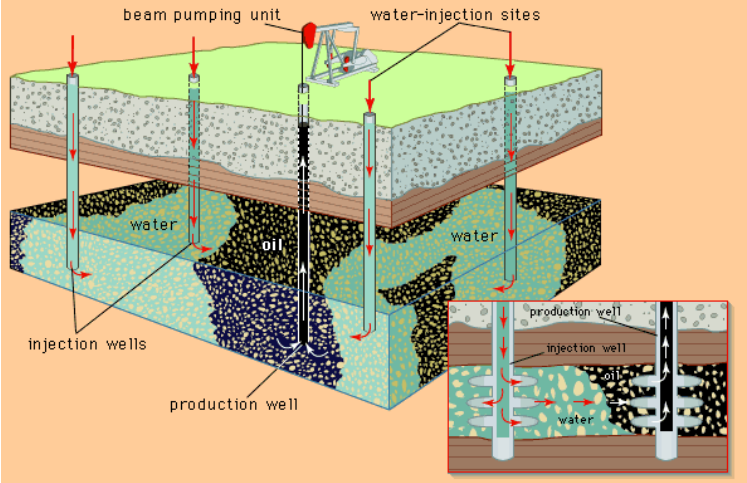


Figure 2.10 Excessive water production. (Alimohammadi, 2018)

Other researchers add that it is worth considering that oil field production is limited by the facilities of the surface. They explain that grown water production (high water cut of the well) has different influences on the well. First, operating costs grow because of the high cost of facilities' water-handling. Second, if this additional facility is not available, production of the oil can be reduced because of the inability to manage the grown liquid fluid (Madrid and Min, 2020). Nevertheless, profits can be negatively impacted by both cases and as well as well closure due to this failure (Al-Fadhli, et al., 2020; Bailey, et al., 2000). Also, some specialists argue (Kewen et al., 2011; Vargas et al., 2019) that older wells are mostly subject to high water cut failure, and as a consequence, much attention should be paid to the age of the well.

However, produced water (with the oil) currently is considered one of the main problems of the oil fields. This issue may cause premature abandonment of the oil fields, and it can lead to the decreasing production rates, decrease recoverable reserves, and constitute not beneficial influence on the environment. The problem is that one barrel of produced water requires the same amount of effort as the volume of oil, and sometimes it happens that one barrel of water shows even less amount of the oil (Aalsalem, et al., 2019).

For indicating the high water cut Basic Sediment and Water (BSW) parameter is used. This parameter is a ratio between the water, sediment and liquid flow rates, the requirement for these parameters is measuring with normal rate of temperature and pressure (Vargas et al., 2019). As it was

mentioned above, during the well age, and because of avoiding oil production decline, this ratio is expected to grow due to grown water production from the natural reservoir aquifer or artificial injection. However, a sudden growth of BSW may cause issues related to flow assurance, decrease of oil production, incrustation and also industrial plant processing. As a consequence, all these issues lead to the need for well closure.

It can be seen that it is rather important from different perspectives to predict the possibility of the high water cut of the well. This action includes the estimation of the reserves, the indicator of the BSW rate, reservoir management, and constant monitoring. Also, the high water cut prediction is a current considerable issue for oil wells with a high density of description (facture) or low permeability (Madrid and Min, 2020). However, less attention has been paid to predicting the high water cut of the well. And as a result due to its cost and ubiquity water cut has been a problem and a significant challenge to engineers.

2.4.3.3 Rate of inflow

The flow instability of the well can be rather easily detected by changes in monitored variables (even one). The hint here is that the exceeding relevant changes of measures should not be with significant amplitudes. However, this characteristic can indicate another failure, unplanned event, or severe slugging (lack of frequency between these fluctuations) (Theyab, 2018; Vargas et al., 2019). As a result, prediction of the flow instability is crucial for avoiding all the points connected with more severe failures, especially severe slugging.

2.4.3.4 Rapid Productivity Loss

The well’s rate of productivity depends on several measured parameters: the pressure of the reservoir, rate of basic sediment and water (BSW), the viscosity of the oil, and the well’s physical parameters for example, the diameter of the production line. When these rates grow or decline, that is a sign that the system's capability is not enough to get over the losses or the flow slows. Timely identification of this problem allows operators not to lose the productivity of the well, because of changing the operation (Hausler et al., 2015; Vargas et al., 2019)

2.5 Research Gap Analysis

This Research Gap Analysis took into account Kazakhstan's current level of research in the field of oil and gas well failure prediction. A preliminary review of academic databases such as

Scopus, Web of Science, and Google Scholar revealed that some research on oil and gas well failure prediction is available, but relatively few works, particularly those focusing on Kazakhstan.

The following research gaps were identified based on the available literature:

- There is a lack of study in Kazakhstan on the specific failure prediction of oil and gas equipment. While limited research has been conducted on the failure prediction of oil and gas equipment in general, few studies have been conducted in Kazakhstan. This is a critical gap since the conditions and limits encountered in Kazakhstan may be unique, necessitating the creation of location-specific failure prediction models.
- Insufficient use of machine learning and data analytics techniques is the second issue that was identified. The analysis of massive volumes of data is required for oil and gas equipment failure prediction, and machine learning and data analytics techniques can be used to uncover patterns and predict equipment breakdowns. However, there has been a shortage of such research efforts in Kazakhstan that have particularly applied these approaches for oil and gas equipment failure prediction.
- Failure prediction models often focus on the technical aspects of equipment failure, ignoring environmental influences. However, environmental conditions such as high water levels on the line might also contribute to equipment failure. There have been only a few studies that have taken these parameters into account while developing failure prediction models. Moreover, the majority of the study focused on the specific components of the well such as pumps and valves. There are only a couple of studies that look at well as a full system.
- Limited data availability is another key difficulty in the research of failure prediction of oil and gas equipment. Companies may be unwilling to give data in some circumstances due to confidentiality issues, making it harder for academics to construct robust failure prediction models.

In conclusion, the research gap analysis suggests a need for additional studies on the failure prediction of oil wells, particularly in Kazakhstan. These studies should take into account the region's particular conditions and difficulties and should apply machine learning and data analytics approaches when possible. In addition, environmental factors should be considered, and the economic impact of equipment failure should be calculated.

CHAPTER 3: RESEARCH METHODOLOGY

3.1 Introduction

The main focus of this research project is to develop the machine learning model using real data from the oilfield and to outline the framework for the implementation of wells' failure prediction in oil and gas companies. In order to achieve this goal both qualitative methods in the form of interviews for validation as well as quantitative methods in the form of advanced analytical methods were applied to this project. The project was carried out in five phases: initiation, planning, execution, closure, and control over the project's flow. In addition, the methodology of the project with decision-making in a form of a chart was developed in this chapter to illustrate the process and furtherly use it in the framework development. Each activity described in the phases of the methodology correlates with the project's aim and objectives to fulfil the requirements of the work.

3.2 Project Development Phases

Step 1: Initiation and Planning

Initiation is one of the most crucial steps of the project due to the identification of the preliminary scope, aim, and objectives as well as initial requirements gathered from the company. At this stage, mutual understanding between two parties may dictate the future successful completion of the project. After the kick-off meeting, the virtual visit to the oilfield and interviews regarding the insights of Caspi Neft took place to understand the operations of the company; a non-disclosure agreement (NDA) was signed and the first bunch of internal documents were transferred. As soon as the team received clear instructions and initial information on the project, the planning stage began with the refinement and documentation of the exact aim, objectives, requirements, and validation techniques in the Client Brief document. This document in turn was then consulted with the industrial supervisors, and the main deliverables with acceptance criteria, key dates, and the description were agreed upon one more time. At the same time, the comprehensive literature review on the pre-defined scope was initiated and conducted at this phase to assess the ability of the team to perform the project and the ability of the company to transfer the required data.

Step 2: Execution

After the comprehensive literature review analysis, the team defined the clear scope of the project and started to work on the data-driven failure prediction model. The execution stage of the project began with the collection of real-time data from the database of company. The historical data

containing all required information regarding well performance, well execution dates, and data on previous failures were merged for further analysis.

The project's execution stage follows the algorithms of failure prediction model development. The initial stage of the data analysis involved the selection of features required for determining the failure probability of the well. The feature selection was done with the assistance of the expert, who also validated the methodology of the project.

After the data acquisition step, data preprocessing, which includes importing libraries, identification of missing values, and feature engineering, was conducted. The feature engineering approach contains several steps such as Data type conversion, Handling outliers, Visualizations and identification of patterns, the correlation between the attributes, Generation of profile reports, and Splitting the dataset. After finding the data preprocessing stage, the data were input into machine learning algorithms.

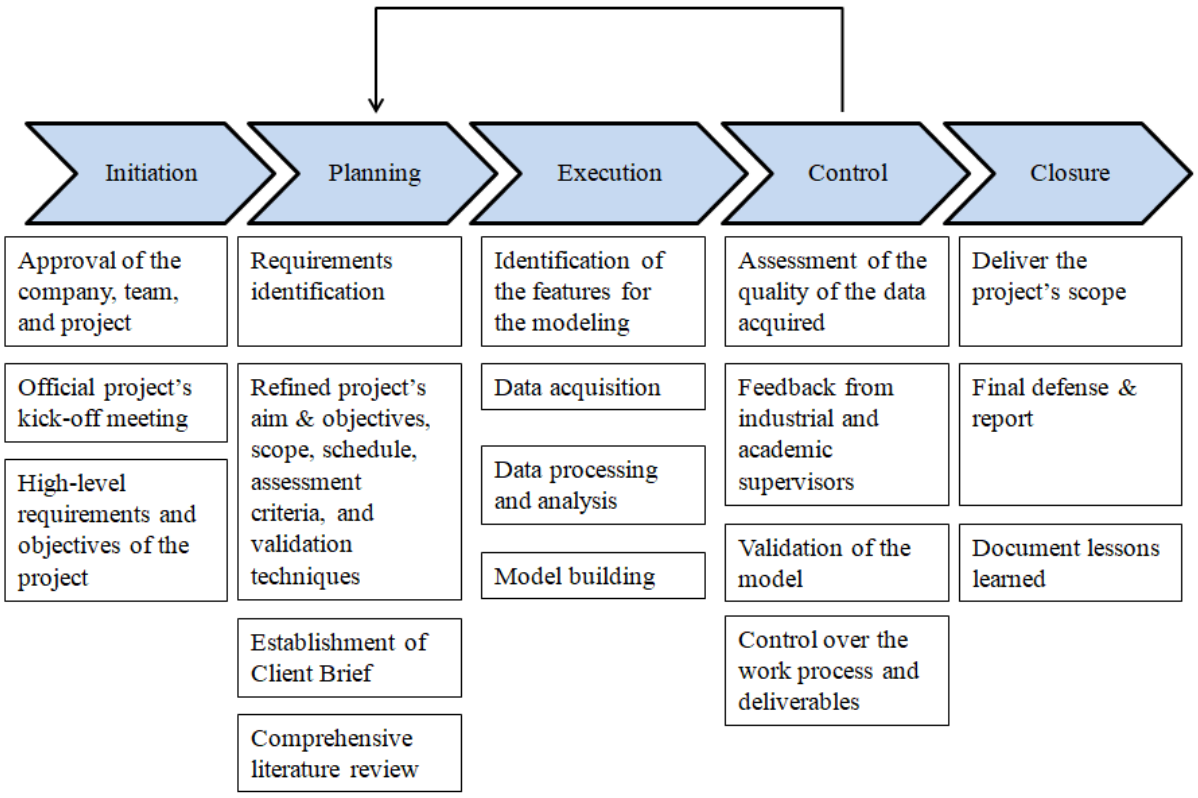


Figure 3.1: Methodology of the Project

Step 3. Control and Closure

Due to the high dependence on real data from the oilfield, its assessment and feedback to the company about it is a crucial part of the project's control. Moreover, weekly meetings, reports, and

presentations were a routine of the project in order to fix the problems in the early beginning. Self-control over the work performed and continuous comparison with the desired outcomes of the project was a crucial part of the project to eliminate the migration from the scope defined. The validation of the machine learning model and framework of the implementation of failure prediction was the last step of the control stage. The closure of the project was aligned with the submission of the official documents, such as the academic report and presentation as well as the presentation for the company. The lessons learned from the project were documented as well.

3.3 Workflow of the Project

Generally, the project's workflow included two main components: literature review and assessment of input data from the collaborating company Caspi Neft. The critical moment of the project was to coincide with the parameters for the failure prediction model from the literature review and data input from the company. Moreover, the assessment of the quality of the data included the presence of the necessary information (operating conditions, production rate, and downtime history) as well as the reasonable quality of the datasets. Several feedback loops were considered in the process of the project due to the recurrent nature of the work on obtaining the datasets and collaboration with the company. As soon as the appropriate datasets were obtained, the data preprocessing was applied to the data followed by the development of a fault prediction model with suitable machine learning algorithms. The next step was to validate the model; in the case of unsatisfactory accuracy, the data preprocessing and model were revised, otherwise, the development of the framework proceeded.

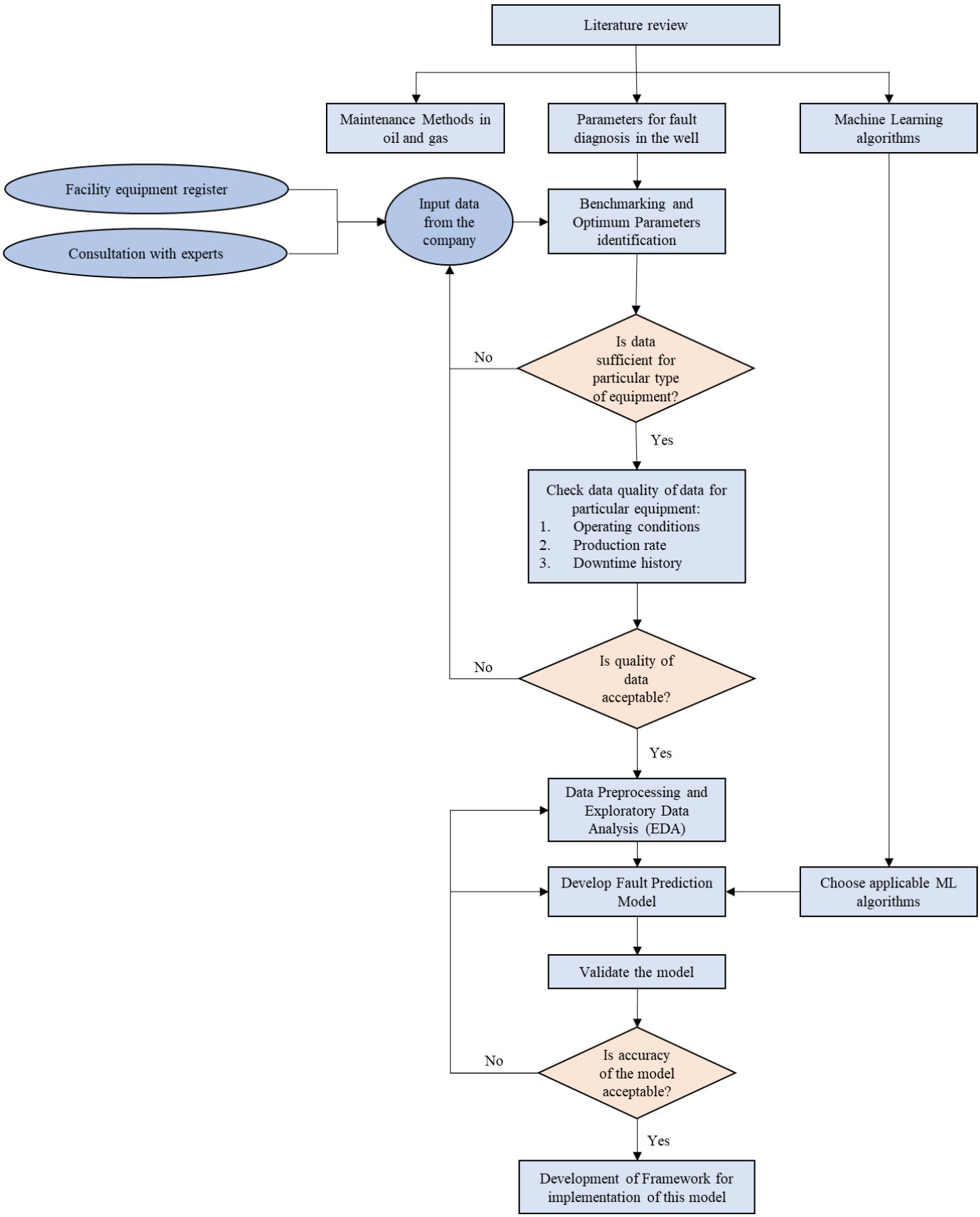


Figure 3.2: Project's process flow chart

CHAPTER 4. CURRENT PRACTICE AT CASPI NEFT

4.1 Introduction

In this chapter, the background, current projects, and practices of the collaborating company, Caspi Neft JSC, are described and analyzed. The first subchapter is aimed at the acquaintance with the company and its current activities. Subsequently, the main operations of the company, precisely extraction, and preparation of the crude oil are presented for the full understanding of processes in the company. AS-IS analysis of the maintenance initiation in the company is performed for further improvement suggestions in the project. One of the most important aspects of Caspi Neft, “Smart Oilfield” project is also taken into account in this Chapter.

4.2 Company Overview

Caspi Neft JSC is one of the main oil and gas companies in the Caspian region, having been involved in comprehensive subsoil exploration, search, exploration, and the extraction of hydrocarbon raw materials in the designated area, storage, and crude oil export since 1997 (АО «Каспий Нефть», n.d.). Since its founding, the company has carried out exploration and appraisal activities in accordance with the license and contract issued by the Government of the Republic of Kazakhstan. Over the years, the business has conducted exploration, including an appraisal of commercial oil reserves and trial exploitation of the multi-layered Ayrankol oil field. The oilfield is located in the Zhylyoi district of the Atyrau region of the Republic of Kazakhstan, while the company's headquarters are in Atyrau city. According to 2021 data, the overall number of personnel is 290 (АО «Каспий Нефть», n.d.).

In 2022, the company was listed as number 18 in the largest private companies in Kazakhstan and ranked number 23 among the country's greatest taxpayers (АО «Каспий Нефть», n.d.). Moreover, from 2018 to November 2022, the company allocated 2.5 billion tenges to personnel training, socio-economic development of the region, and sponsorship support (Chervinskiy, 2022). Since 2019, the shareholder has been reinvesting a significant portion of these funds into the company. “Caspi Neft was one of the first oil-producing companies in the country to implement a "digital field" system at the Ayrankol field. Starting with the automation of production in 2019, the company achieved digitization of production data in 2021” (Chervinskiy, 2022). Furthermore, they launched a strategic analysis center on the field, allowing for analytics and process modeling. As a result, by reprocessing and reinterpreting seismic exploration data, after building a geological and hydrodynamic model of the field and drilling appraisal wells, the resource base of Ayrankol was

increased by 4 million tons of oil (Chervinskiy, 2022). That is, almost 40% of the initial recoverable reserves.

Caspi Neft's dedication to innovation and the use of technology to enhance its operations was honored in 2023 at the "Digital Almaty 2023: Digital Partnership in a New Reality" forum, where the organization received the Tech Garden Award 2022 in the "Best Digital Solution" category (Galushko, 2023).

4.3 Processes and Operations at Caspi Neft

The following description of process flow and production steps were prepared with the assistance of information provided by specialists of Caspi Neft Company. Overall, the processes in the Ayrankol can be divided into two groups, those dedicated to the extraction of crude oil and the preparation of oil for trade.

There are 203 production wells and numerous injection wells in the area of the oilfield, which in turn is divided into two wings: West and East sides. The West wing of the oilfield has a lower number of wells, however, the viscosity of the oil extracted there is considerably higher, making this oil more expensive and profitable compared to the oil from the East wing. At the same time, the East wing accounts for more production wells located there. Actually, the liquid pumped in the wells is not 100% pure oil; this liquid is called process fluid and it consists of oil, water, and gas in the case of the Ayrankol field. Depending on the percentage of each component in the process fluid, the decision on whether the well should operate or not can be drawn. For example, if the water content is higher than the oil, it may be not profitable to pump this process fluid, so the well is closed for some time.

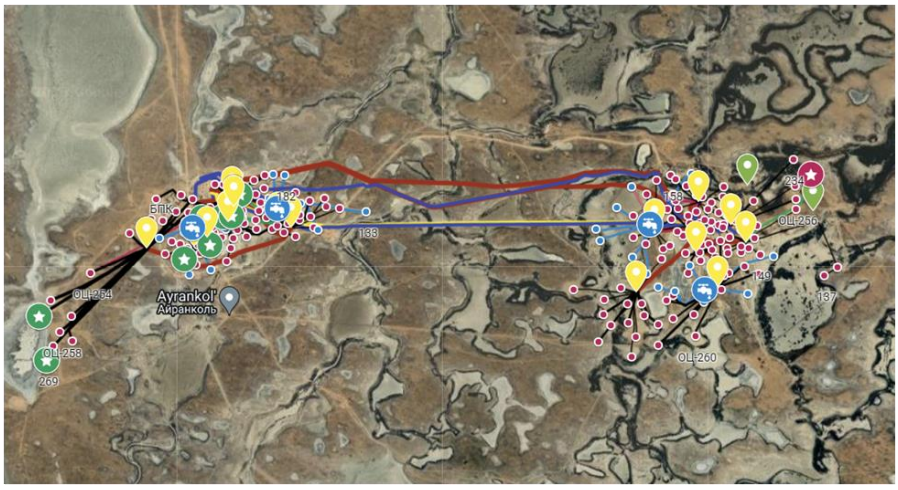


Figure 4.1: West and East wings of the Ayrankol oilfield

After the extraction of the process fluid from all 203 production wells, it is transported to the corresponding oil collection points. Some of the points are designed only for the collection of high-viscous oil and some are equipped with Automatic Pad Metering Stations (APMS), which measure the process fluid technological parameters and its volume. Generally, process fluid from 6 to 8 wells is gathered at one oil collection point. From the oil collection points, all process fluid is transferred to the biggest collection point, and some samples of the fluid are taken to the laboratory to check the properties of the liquid.

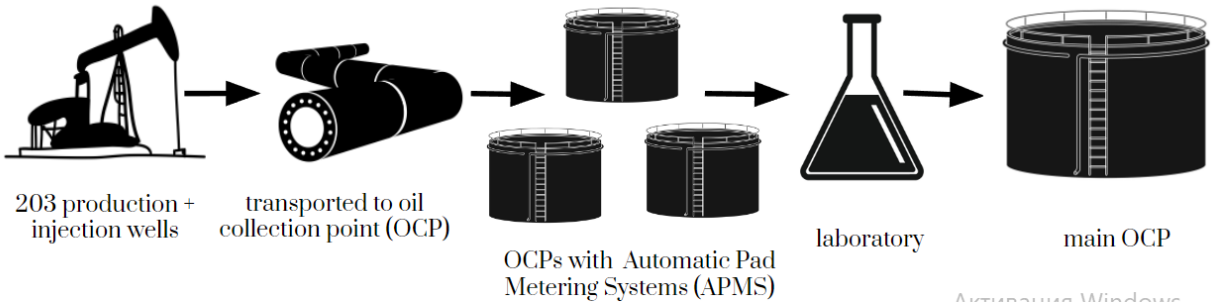


Figure 4.2: Extraction of the oil in the Ayrankol

The following step is to prepare the crude oil for trade and transportation through pipelines of the Caspian Pipeline Consortium (Uzen-Atyrau-Samara pipeline) to the Russian Port of Novorossiysk, where about 80% of all Kazakhstani extracted oil is gathered and shipped to the European market (Wood Mackenzie, 2022). To prepare the process fluid Caspi Neft implements several stages of separation of oil from the impurities, namely gas, and water. First of all, the process fluid goes through the scrubber, where the gas is separated from the liquid. The obtained associated gas is then utilized as fuel for the oilfield's operation, which makes Caspi Neft a self-sustainable entity in terms of energy consumption. In order to separate the oil from water, several fractional distillation columns are used with the bypass streams to achieve the desired purity of the oil. Water separated in this process is sent to the tank, and when a reasonable volume of water is collected there, it is sent to the injection wells to pump the oil from the oil reservoir. The last stage of oil preparation is a flow of the oil through heat exchangers, which increases the temperature of the oil to the standard one, thus decreasing its viscosity and density according to the trade rules. Finally, the prepared oil is transported through the pipelines.

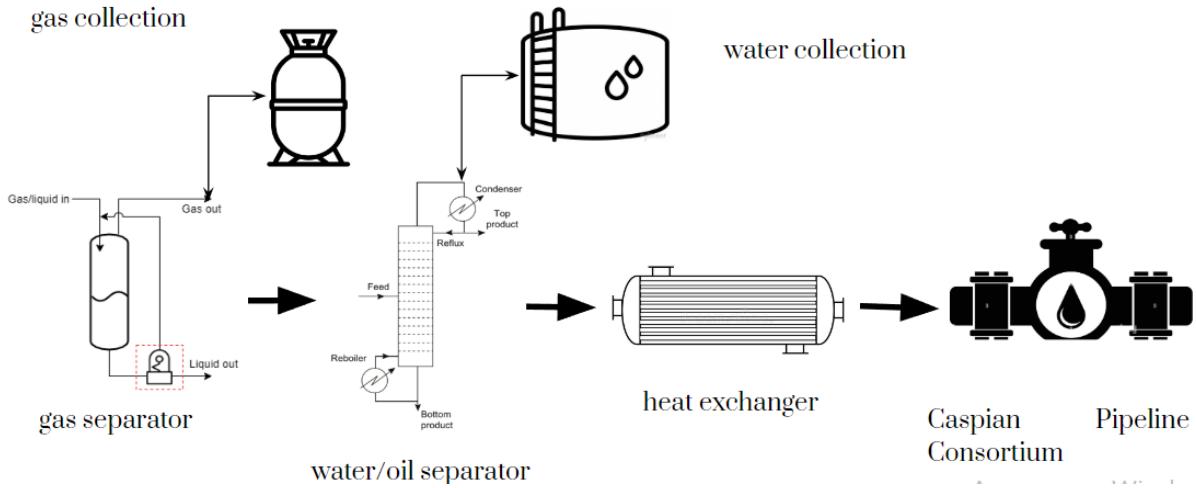


Figure 4.3: Preparation of the oil in the Ayrankol

The process described above is executed in a continuous manner for 24 hours 7 days a week as the extraction of crude oil is profitable, yet comprehensive production. To allow uninterrupted production, technical inspection and maintenance of the equipment are performed in a timely manner described in the official plans and passports of the equipment. Currently, the initiation of the maintenance activities, thus closure of the wells in Caspi Neft is performed in the manner described in Figure 4.4. Three inputs of the maintenance initiation are the repair plan, maintenance services plan,n and equipment manufacturers recommendations outlined in the passports of equipment. At the same time, the assessment of the state of the equipment by engineers can also initiate the closure of the well. From these four inputs, it can be said that the company exploits preventive maintenance (plans and passports) and reactive maintenance (assessment). These maintenance techniques were previously discussed in the Literature Review section of the current Report as costly and ineffective in a long-term perspective compared to predictive maintenance techniques. As a result of such input, sometimes unnecessary closure of the wells occur in Caspi Neft after which wells are restarted. However, closure, inspection, and restart of the wells require time that in turn reflects in production and money losses for the company.

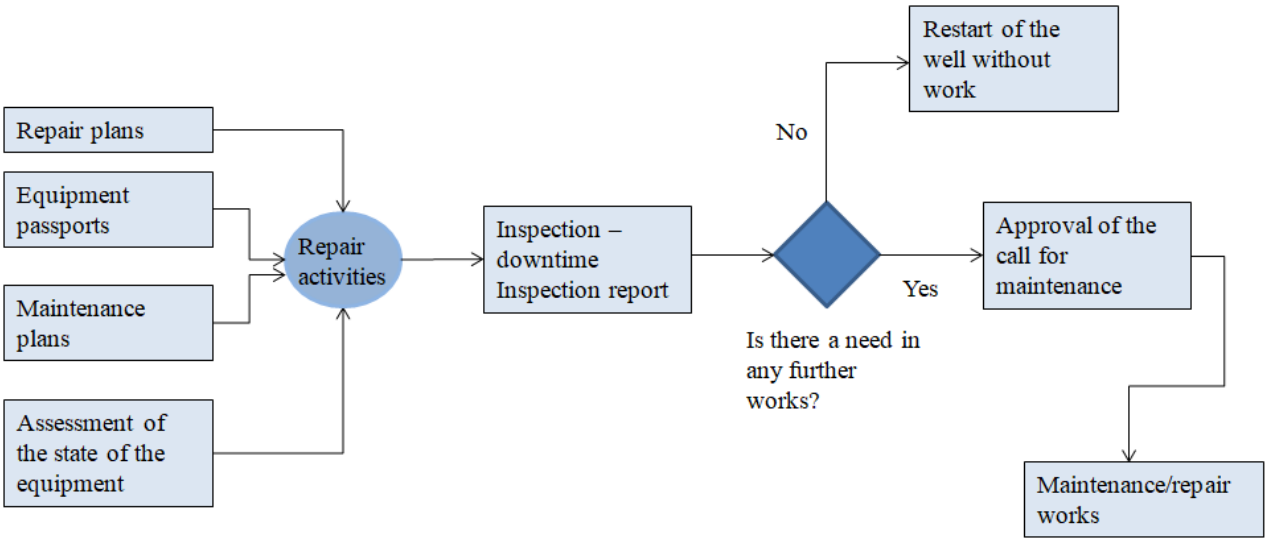


Figure 4.4: AS-IS representation of the current maintenance initiation

4.4 "Smart Oilfield" Project

Subsequently, after analyzing the production process and its attributes, the next part of the paper is subjected to the identification of the company's challenges. Each company in the oil and gas industry wants to maintain its oil supply through the full utilization of the reservoir's capacity or by finding new oil deposits. However, in the example of the considered company, due to the transition to a later stage of field development of Caspi Neft and based on the approved project documents and expert forecasts, production levels of the oilfield have been declining since 2018. And as a consequence, nowadays, Caspi Neft is approaching a problem when it has to face the results of an irreversible problem of oil production decline. Therefore, the company is trying to achieve the task of involving innovative solutions to achieve the improvement of the company's performance.

In order to solve the above-mentioned problem of oil depletion and maintain the level of oil production, the company has already started searching for new technologies. Currently, they are trying to achieve the desired results of modernization and automation of production facilities through a complex project "Smart Oilfield". The "Smart Oilfield" project is created for one important purpose increasing the efficiency of the company and its cost saving.

According to the material from the interviews with the specialists from the company, this project was initiated in 2017 with the construction of geological and hydrodynamic models of oil deposits and improving the company's production processes (Galushko, 2023). Currently, the project is still in the execution phase approaching its final stage. Initially, the intellectual part of this innovative solution consists of four cycles/modes, as it is shown in Figure 4.5.

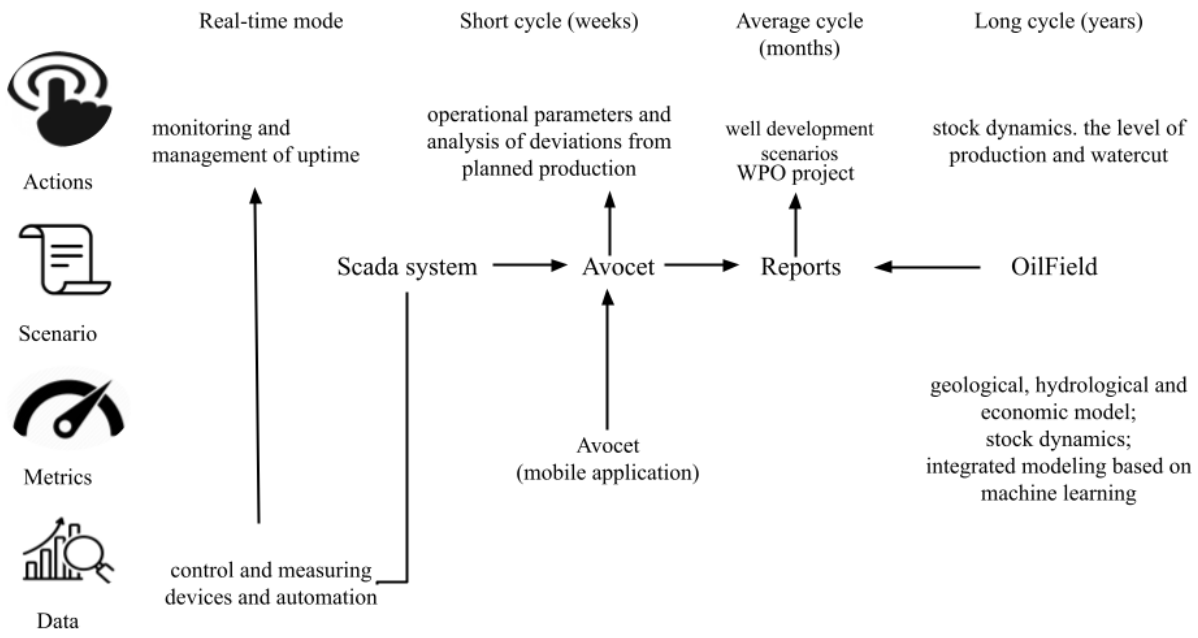


Figure 4.5: The architecture of "Smart Oilfield"

Firstly, some high-yield wells and pipelines are equipped with sensors, which measure the properties of the process fluid, its volume as well as the parameters of the equipment to track the conditions of both. It enabled real-time monitoring of production facilities and their remote control from the analytical center of the field management to ensure failure-free operation. However, the data is mostly presented on a weekly basis and does not collect parameters by the hours which might complicate the process of further maintenance actions.

Secondly, based on the data of the first cycle, plans for days or weeks are calculated, operational parameters are analyzed and technological modes of well operation are derived to maintain planned production. Nevertheless, this data is not used for forecasting the component's life and as a result, it can not help to optimize the timing of the element's unplanned downtime.

The third cycle includes the Oilfield Manager (OFM) software, which allows the specialist to analyze the development of the field and plan scenarios for the operation of oil deposits in the medium term. Another key tool of this cycle is the WPO (Well Portfolio Optimization) platform, which collects data on production and geological and technical measures for wells for the entire development period.

However, up to now, the project is only in its development stage, and achieving the final stage assumes the effective usage of the monitored data. This final stage of the project includes the valuable outcome of the stored data. Based on the 4 cycles project's implementation plan there is a critical

need for a failure prediction framework that will consider the condition of the company's equipment. The failure prediction model will be created on the collected data from the sensors (equipment log). It is expected that the advanced version of this framework will help to extend the part lifetime, reduce unexpected events, avoid expensive oil field shutdowns and equipment damage.

Overall, the failure prediction model will help to optimize the use of the system's elements and increase its productivity. As a result, it will help to minimize the costs of maintenance activities by avoiding early and frequent production suspending.

CHAPTER 5. DATA ANALYSIS & PREDICTIVE MODEL DEVELOPMENT

5.1 Introduction

In this chapter, a summary of the parameters identification pertaining to the list of optimal for the data collection and its further modeling is provided. The result of obtaining the requirements is depicted through the cause-effect diagrams. Taking into account these parameters, further data preprocessing was performed, before the creation of a failure prediction through the construction of classification and recurrent neural network models. This proposed model helps oil and gas companies to implement one element of predictive maintenance, precisely failure prediction for their production wells based on their current capabilities. In order to obtain real datasets and access to experts in the field, the current project is performed in collaboration with Kazakhstani oil producer, Caspi Neft.

5.2 Optimal Parameters

The literature review chapter includes an investigation of wells, their structure, main types, and common failure modes. This section covers the attributing parameters responsible for the wells' downtimes for further model construction. Understanding the current state and the history of these parameters are the keys to the successful prediction of well interventions and other related actions. It is worth mentioning that Chapter 4 on the current practice of Ayrankol "Smart Oilfield" identified that a large amount of data is monitored during the operation of the wells and other oil field equipment. The data that can be used to make decisions about the well downtime includes both interpreted data from experts and raw data directly from sensors, which may be structured or unstructured. Sorting these parameters, and output data from sensors, makes it more useful for the decision-making process during field operations and failure prediction for the wells. Besides parameters of the well from sensors, expert evaluation from various areas of the industry needs to be involved, for example, geology, engineering, and production. The reason for that is the diversity and the strong correlation of these parameters. However, finding the parameter with the highest influence on oil well performance is of current great interest and value for the optimization of the production process.

Table 5.1 illustrates common oil well monitored parameters identified according to the literature review analysis and provided data. In addition to the raw parameters monitored from the oilfield, the table below shows parameters that were further requested from the company.

Table 5.1: Optimal parameters influencing the well downtime

Monitored parameters in the oilfield	Requested parameters
• tumbling head pressure • casing pressure • bottom hole pressure • flow line pressure • frequency • temperature • current • flow rate of a pump • fluid flow rate • basic sediments and water (BSW)	• well orientation; • well age; • physical parameters; • downtime type; • comments to the undesirable well events (downtimes);

Along with the well's commonly monitored parameter identification, the production data delivered by the company was analyzed. Overall, taking into account both received versions of the company's datasets (initial and updated based on our request), the provided dataset includes the operational history of the pressure (tumbling head, flow line, casing, and bottom hole pressures), submersible electric motor characteristics (frequency, efficiency, temperature, current), fluid flow rate and BSW. Along with the initial data parameters received from the company, the well orientation, its exploration date and age, the physical parameters, failures, and their characteristics with any additional comments were also asked from the industrial supervisors of the team. These parameters are considered valuable for failure prediction, especially for the common undesirable events occurring in the considered oilfield. Unfortunately, this information was not transferred due to confidentiality. Taking into account the information from the literature and the 2 versions of the provided data from the field the following cause-effect ("fishbone") diagram was created for further stages of data analysis and modeling (Figure 5.1).

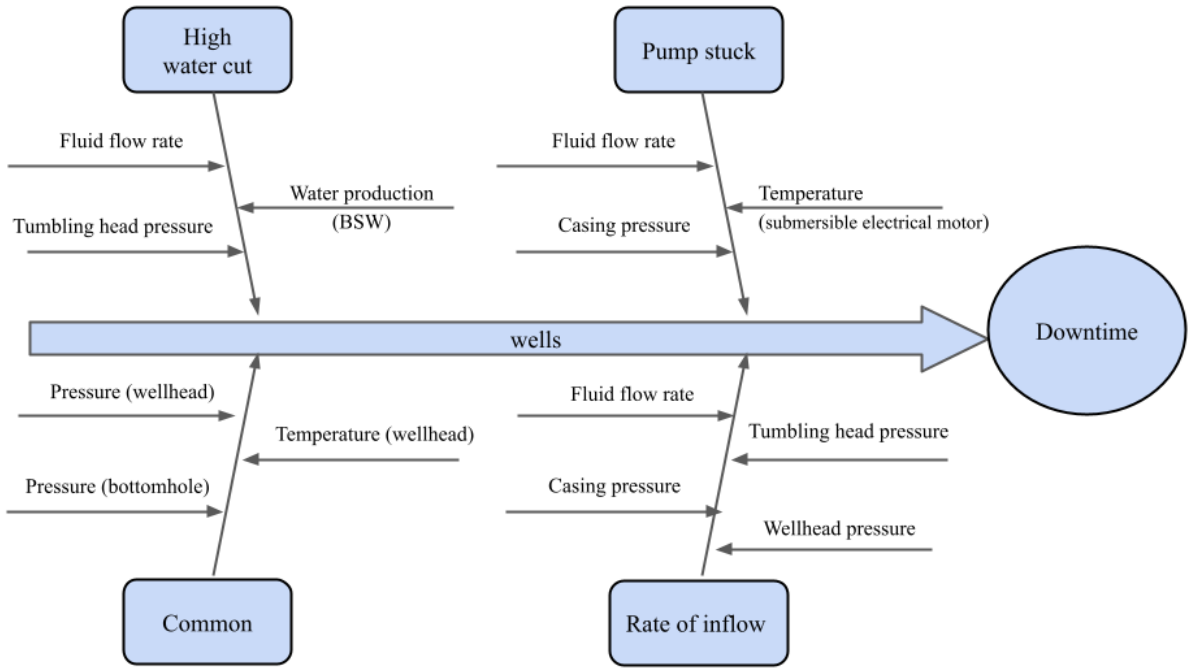


Figure 5.1: Cause-effect diagram of the oil well's most common downtimes

This fishbone diagram represents the most frequent unplanned well downtimes and the possible parameters affecting the failure (its indicators' increasing or decreasing). These parameters were further implemented for the predictive model construction. According to the analysis, it was identified that the key parameter indicating a "not healthy" well's performance is the fluid flow rate.

As was mentioned in the Literature Review and observed from the dataset delivered by the company, the most common failure of the oilfield is a high watercut. And the diagram below (Figure 5.2) presents the parameters that may affect the high watercut of the well.

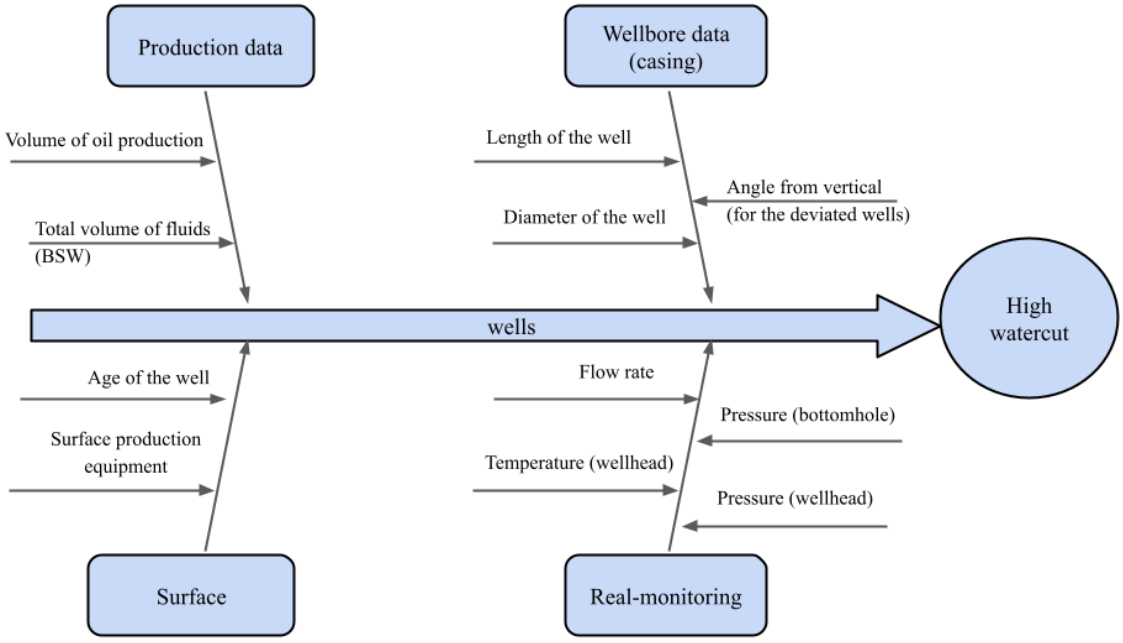


Figure 5.2: Cause-effect diagram of the parameters affecting the high water cut of the well

Overall, based on the table and diagrams it is expected that the parameters described above are reasonable and optimal enough and they can be used for further tasks associated with creating a model for predicting undesirable events in oil wells.

5.3 Data Collection and Acquisition

The Ayrankol oilfield uses an operational management system that allows operators and engineers to receive real-time data on oil production. Using the system of dispatch control and data acquisition SCADA control room every second receives the parameters of 43 high-yielding production wells. However, it should be emphasized that this system is used for proactive response and saves a record only once a day at the closure of operation day. The remaining parameters of more than 200 wells are recorded manually by the field operators, also once a day, into the AVOCET data collection and configuration platform. Data is stored in more than 10 separate system tables.

The company provided data on "Daily measurements of production wells" (production wells), "Daily measurements of production wells at the Automatic Group Metering Unit/Automated Block for Measuring High-viscosity Oil" (flow rate) and "Downtime at the wells" (downtime). The production wells dataset contained 459,277 rows and 27 columns with an 11 years time period ranging from 01.01.2012 to 27.01.2023 for 209 production wells. The flowrate dataset contained 505,409 rows and 11 columns ranging from 08.09.1900 to 01.18.2023 with parameters for 220 wells. It was initially clear that this dataset had a large number of duplicates. The third dataset downtime

consisted of 1260 rows and 9 columns with a time period ranging from 01.01.1990 to 01.30.2023 with downtimes for 190 wells. All three tables were obtained in CSV format.

After an extended negotiation process, the second dataset was obtained from the company which contained various pressure types in a production well, pump and its submersible electric motor (SEM) indicators, and other parameters with more than 241,259 rows and 17 columns. This table contained data for 284 wells for a 2 year period from 01.01.2020 to 31.12.2022. It is similarly based on a daily measure at the closure of operation day.

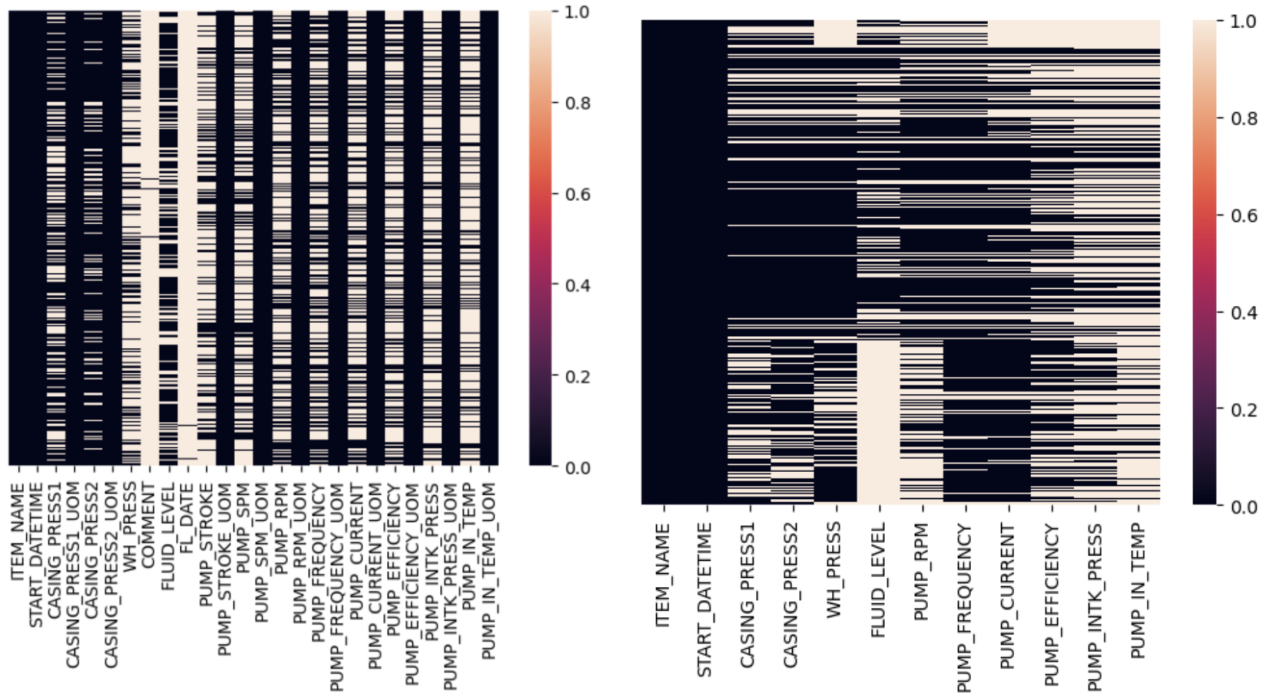
5.4 Data Pre-processing

The data preprocessing stage was divided into two parts. The analysis stage of the first one included a collection of data in one table on the main parameters and downtimes, as it contained such information. While the second dataset did not contain downtime types, it was necessary to identify production wells failures. The next part was focused on preparing raw data to be used in the ML model development. As the dataset did not have downtimes, preprocessing is significantly necessary to ensure the accuracy and reliability of predictions.

5.4.1 Data source: First dataset from the oilfield

To optimize the process of data preparation and speed up further analysis due to a large amount of data, Python programming language was used. The main libraries are Pandas, Numpy, Matplotlib, and Seaborn.

The primary step in data preparation for oil production wells, flow rate and downtime tables provided by Caspi Neft company involved importing the data using the Pandas library. As the main downtimes were logged after 2020, the data was filtered to include only records starting from 2020. From the 27 columns that the production wells dataset contained 13 columns were deleted due to uselessness as a unit of measure of parameters and a substantial number of empty cells as liquid level measurement time, pump stroke length, and number of pump strokes. After cleaning the data, using the Matplotlib library to create visualizations that assist in the identification of trends and patterns in the data, a heatmap chart was used to visualize the consistency and completeness of the table (Figure 5.3).



a)before preprocessing

b) after preprocessing

Figure 5.3 Raw parameters of the production wells before and after the preparation process (light space depicts the absence of the data).

Moving forward in preparation of the flow rate table, it was discovered that it contained 383,351 duplicated rows out of 505,409, which made it useless for the model construction.

As for the downtime dataset, it consisted of well name, downtime start date and time, downtime end date and time, downtime duration in hours, downtime categories, and 4 sub-categories. The data covered 1215 downtimes for 175 production wells divided to 2 main categories: scheduled and unplanned downtimes. Each category also contained 3 sub-categories strating from major parts of the wells, reasons, and some cases including some of the details. Both of the downtime types had 14 main subtypes that stated the main cause of a failure (Figure 5.4).

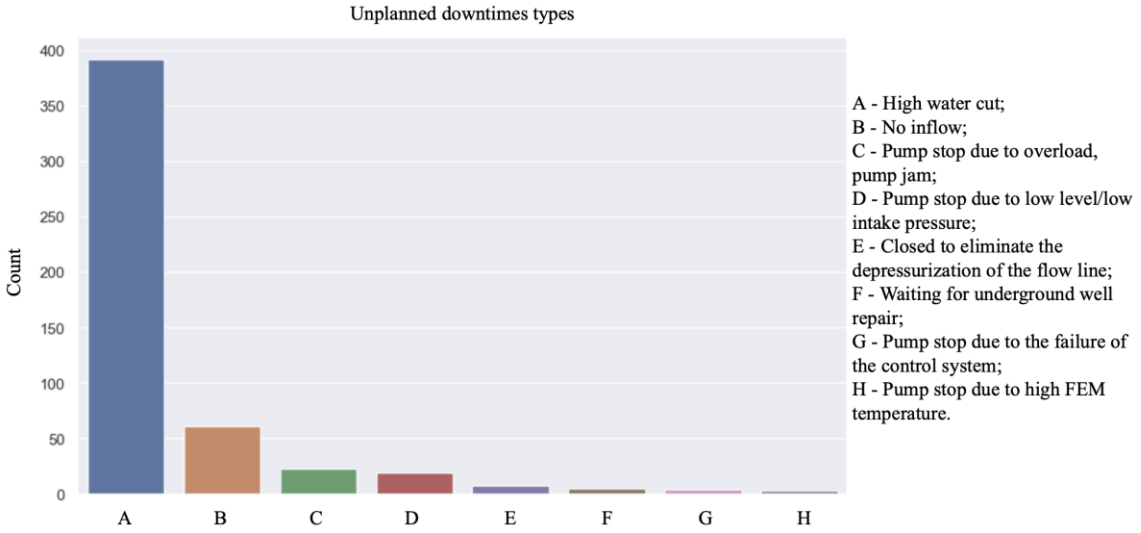


Figure 5.4: Categories of production wells' downtime types

At the stage of data transformation, the inconsistency in value ranges in the production wells data frame and human errors in the data entry process were identified in the analysis phase that complicates the further work with the dataset. It also should be mentioned that most of the downtime types within the production wells dataset are not recorded in the downtime table. Moreover, the class imbalance was also an issue with the downtime table, as a disproportionate number of observations in "High water cut" compared to other failure types would lead to biased results.

5.4.2 Data source: Second dataset from the oilfield

The second dataset form the oilfield contained the following columns for further analysis "well name", "date of measure", "line pressure", "buffer pressure", "annular pressure", "bottomhole pressure", "frequency SEM", "pump_performance", "temperature SEM", "current SEM", "fluid rate SEM", "water cut", "commissioning date", "dynamic fluid level", "Active_hours", "Gas_rate", "Wellhead temperature". Except "commissioning date" the rest of the features are needed for model construction. Since the table had inconsistency and lack of data, 66 production oil wells with constant performance were selected.

The datasets may contain missing values, which can lead to challenges during model construction and training. The practical solution involves removing entire rows that contain missing values, it's most appropriate for large datasets. However, such an approach may result in the loss of important information. The other way involves imputing missing values in numeric columns by using various techniques such as mean, median, or mode. In this case, it is crucial to preserve the integrity

and prevent bias in the data. As the remaining empty values were filled with zeros or approximate values based on historical performance in order not to lose important data. In addition, stable parameters that do not show anomalies and are more suitable for planned shutdowns have also been removed. Likewise, the range of each feature was taken into account and ensured that it was in the correct range of the well's parameters. Since preserving data consistency was a significantly important aspect, some empty cells were filled with average and median values for a certain period, and the influence of various parameters on each other when the filling was also considered. As the state of features for a random production well after the preparation stage can be seen in Figure 5.5.

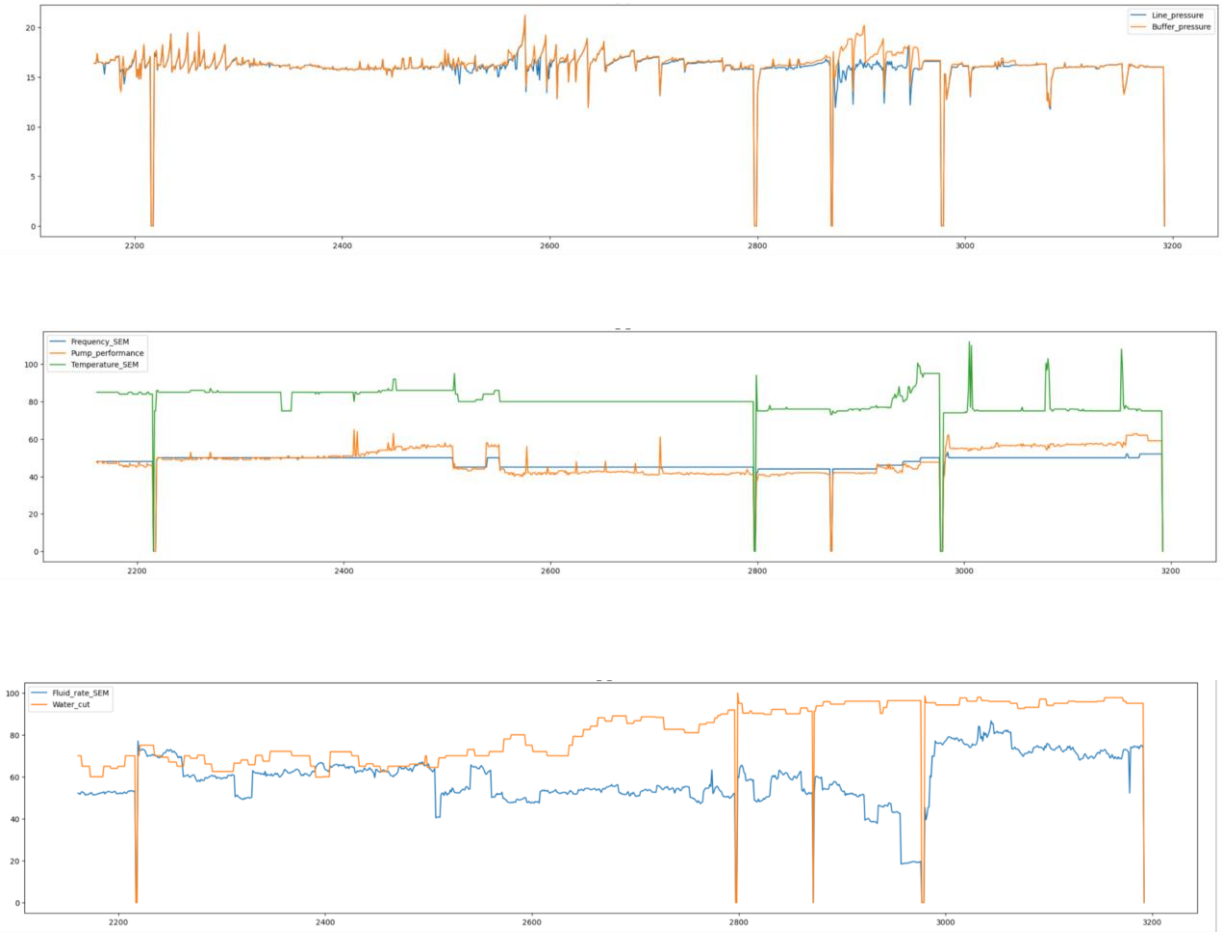


Figure 5.5: State of production wells parameters after preparation

The creation of new columns is often necessary for a machine-learning model construction because the model performance indicators are dependent on the consistency, relevance and quality of the data used to train it. Creating new columns or features, can increase the richness of the data and provide the model with more useful information to train. This may include transforming existing data, such as extracting numeric values from text fields, combining multiple fields to create a new variable,

or applying mathematical operations to existing columns. In addition, creating new columns allows for capturing patterns and relationships that may not be immediately visible in the original data. In this way, we can improve the accuracy and robustness of a machine learning model, allowing it to make more accurate predictions based on new, unseen data. Therefore, the new feature "Duration of downtime" in hours presents the idle time per day based on "Active hours" which in turn was removed, and "Failure" presents cases of failure itself.

The correlation between features is an essential aspect of building accurate and robust machine learning models (Figure 5.6). It helps to identify the relationships of a target variable and input features and taking into account possible issues with underfitting and overfitting the model. When features are highly correlated, they can be redundant and may not add much value to the model. On the other hand, features that are weakly correlated may not provide sufficient information to the model. The correlation matrix was computed using pandas corr(), which provided the correlation coefficients for all pairs of features. Based on the correlation matrix, it can be determined which pairs of features were highly correlated, negatively or uncorrelated, and further selected for the model construction. The final version of preprocessed contained 77,872 rows and 17 columns.

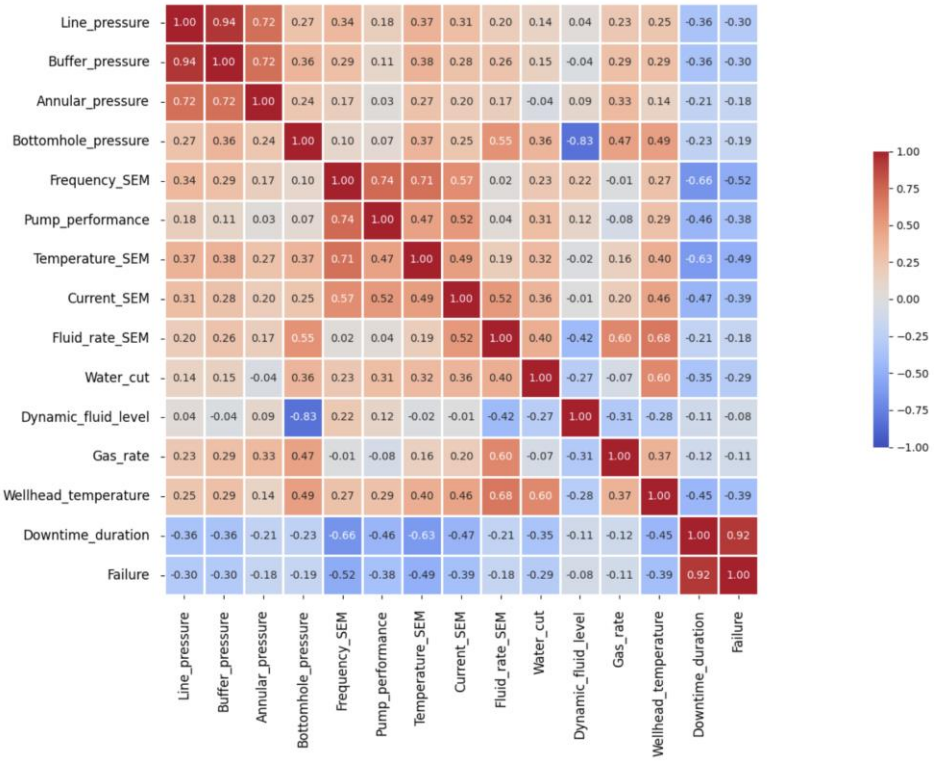


Figure 5.6: The heatmap of the second dataset features

5.5 Development of classification and RNN model

This section aims to develop a classification model and a recurrent neural network (RNN) model for predicting the failure of production wells. Both supervised and unsupervised machine learning algorithms were used. Firstly, unsupervised machine learning algorithms, specifically the clustering technique, were utilized to discover patterns or structures in the data and group failure types accordingly with further usage of classification algorithms. Subsequently, the LSTM algorithm was employed for predictive modeling.

5.5.1 Development of classification model

Once defined with a dataset, clean up, and preparation for a model, the best practice for building machine learning requires additional steps. With the increasing prevalence of large datasets across various fields, it has become necessary to find effective ways of interpreting them. According to Jolliffe & Cadima (2016), principal component analysis (PCA) remains one of the most established and commonly used techniques and its core principle is straightforward - by reducing the dimensionality of a dataset, PCA aims to preserve the majority of the statistical information or variability in the data. In PCA, each principal component explains the magnitude of the variance of the data which is the individual variance. From the given results, it can be seen that the first principal component explains the majority of the variance (97.72%) in the data, followed by the second and third principal components which explain 1.46% of the variance (Figure 5.7). The remaining principal components explain a very small amount of variance, with the last principal component explaining less than 0.1%.

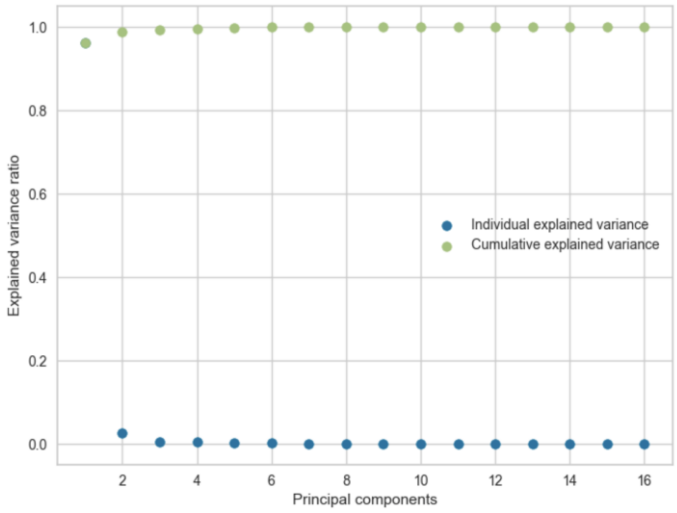


Figure 5.7: Individual and cumulative variance described by PCA

The cumulative variance also represents the total amount of variance in the data explained by a given number of principal components. From the given results, it can be seen that the first principal component explains 97.72% of the variance in the data, while the first three principal components together explain 99.50% of the variance. As more principal components are added, the cumulative variance approaches 100%, indicating that the variance in the data can be well-represented.

The provided data in the data frame (Figure 5.8) has three columns: PC1, PC2, and Downtime_duration. The PC1 and PC2 columns are the first and second principal components obtained through PCA. PCA is mainly used to reduce dimensionality, while preserving most of the statistical information in the original dataset to transform from higher to lower-dimensional dataset. The Downtime_duration column represents the duration of downtime for each observation. It is not clear from this information how this column relates to the PCA analysis, as typically PCA is an unsupervised learning technique used to uncover patterns and reduce dimensionality without using any external information. The analysis of parameter data for a unit's degradation over time and the building of a classification model can be extremely challenging in many cases where the start and end positions of other units are very close to each other. This proximity can make it difficult to analyze the relationships between the data and accurately classify the units.

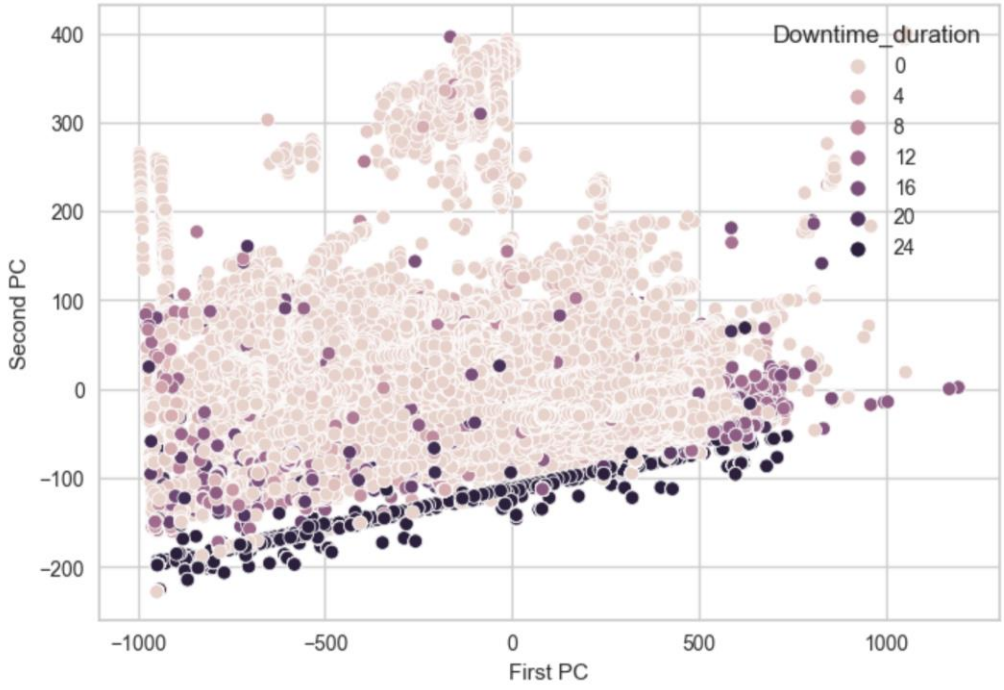


Figure 5.8: Principal Component Analysis

Clustering is a technique of unsupervised ML, which groups similar data based on certain patterns, inherent characteristics and properties that vary from dataset to dataset to clusters. The main

aim of clustering is to outline the patterns and structures in the dataset without any labels previously given. Moreover, clustering is needed in various applications such as customer segmentation, anomaly detection, image segmentation, and many more. It can help identify subgroups or clusters within a large dataset, which can be useful for targeted marketing, personalized recommendations, or understanding complex relationships between different data points.

For clustering purposes, the K-means algorithm was used. K-means algorithm divides a given dataset into k number of clusters interactively with a purpose to minimize the sum of squared distances between each data point of the dataset and the centroid of its assigned cluster identified by the algorithm. In simple terms, k-means tries to find k number of centres that are representative of the points in the dataset and then assigns each point to the nearest centre. K-means is an efficient technique, which is able to handle large datasets, therefore it found its wide application in clustering tasks.

The classification model after clustering was trained and tested using the following types of machine learning models:

1. Decision Tree Classifier;

2. Random Forrest Classifier;

3. Support Vector Machine Classifier.

Normalization is used to scale numerical features in a dataset to a common range. This is done to ensure that all features contribute equally to the analysis and modeling process, regardless of their original scale or units of measurement. ML algorithms are generally sensitive to the range and scale of input data, its features, values and relationships with dominant features that can lead to bias and errors. Moreover, it also assists in the prevention of datasets from overfitting. Normalization helps to avoid this issue by bringing all features to a similar scale, typically between 0 and 1 or -1 and 1. There are different techniques for normalization, including min-max scaling, z-score scaling, and log scaling, among others. The distribution and characteristics of the data and the requirements of the specific machine learning algorithm dictates which normalization technique should be used. So, before deciding which normalization method to use, it can be primarily tested by running 3 different types of normalizations on the k-means clustering algorithm to see the results. Overall, the choice of scaling method influences the clustering results significantly, and it is important to experiment with different scaling methods to stop with the one that shows the best clustering solution for a given dataset and a number of clusters.

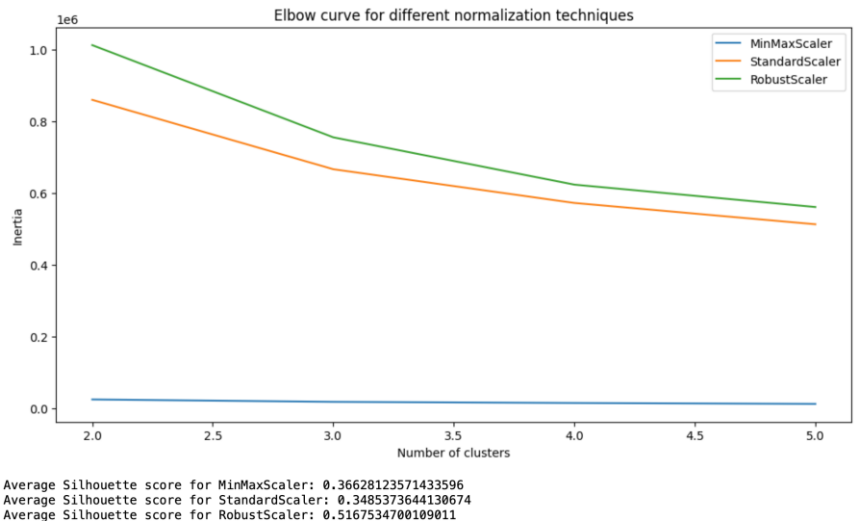


Figure 5.9: Preliminary normalization for k-means clustering

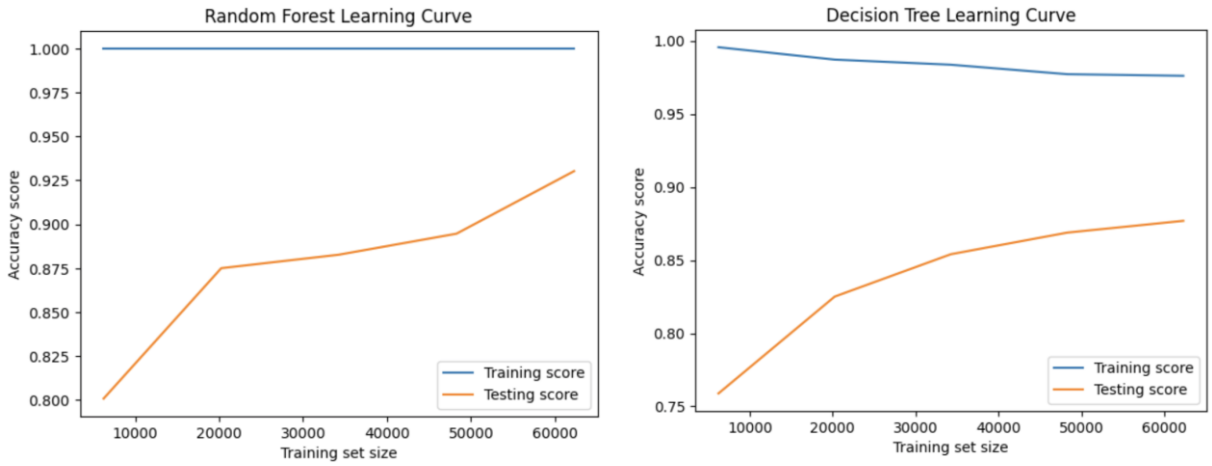
Silhouette score and Elbow score are two commonly used metrics for evaluating the quality of results in K-means clustering. Silhouette score assess the extent to which an object is similar to its own cluster compared to other clusters identified (Kumar, 2020). Silhouette score may take the values from -1 to 1, where a higher score (closer to 1) means that the object is well-matched to its own cluster assigned and poorly matched to other neighbouring clusters, so the clusters are well-separated. At the same time, a score closer to 0 indicates overlapping clusters, while a negative score demonstrates that the object may be easily assigned to the wrong cluster. The elbow score, on the other hand, is a function of the number of clusters that measures the amount of variance. The score is obtained by plotting the sum of squared distances between the data points and their nearest cluster centres. Based on the results, the RobustScaler seems to produce the best clustering solution, with an average Silhouette score of 0.4931 (Figure 5.9). This suggests that the clusters are well separated and the data points are tightly clustered within their respective clusters.

For the model construction several columns that had too high and low correlation (Figure 5.6) were eliminated due to the risk of data unnecessity for specific, bias and overfitting: ‘'Well_name', 'Date', 'Well_age', 'Dynamic_fluid_level', 'Gas_rate' and 'Line_pressure'. After normalization, the Silhouette method to determine the optimal number (k) of clusters was used and the value of 4 that maximizes the average silhouette score was identified with a final Silhouette score of 0.497. Despite all possible combinations of features, correlations between them and the number of clusters, this Silhouette score did not exceed 0.5 and other clustering scores that were attempted to calculate as Elbow method and DBSCAN,. Accordingly, despite the results of subsequent building of classification models, they cannot be calculated as accurate or ready for use in the field. This was the optimal number of clusters for the dataset. Based on this the classification can be started.

The k-means clustering gave understanding of the underlying patterns in the dataset, which allowed to better select and tune developed classification models. The Decision Tree model was the most efficient, which is an important consideration for applications with limited computing resources. Overall, analysis demonstrated the importance of selecting the right model for the specific problem and dataset at hand, and how clustering techniques can be used to inform and improve classification model selection. The detailed code for classification models is presented in Appendix A.

5.5.1.1 Validation of classification model

In this study, the performance of three different classification algorithms were investigated after unpsupervised clustering - Decision Tree, Random Forest, and Support Vector Machine (SVM) - in predicting the target variable. The data was splited into training and validation sets (80/20) and trained each of the three models using the same hyperparameters. The evaluation of the models' performance on the validation set using metrics such as accuracy, precision, recall, and F1 score was performed (Figure 5.10). Based on the comparison of three different classification models: Decision Tree, Random Forest, and SVM. While all three models produced good results, the Decision Tree model achieved the best result due to its speed and low computing power consumption. Overall, the model demonstrated the importance of choosing the right classification algorithm for a given dataset and problem, and highlighted the potential of Decision Tree model in achieving high accuracy and predictive power in less time and computational power consumption.



	Accuracy	Precision	Recall	F1-Score	Training time	Prediction time
Decision Tree	0.994504	0.994509	0.994504	0.994506	0.425731	0.004166
Random Forest	0.997278	0.997278	0.997278	0.997277	3.959229	0.122276
SVM	0.999281	0.999281	0.999281	0.999281	81.444668	0.090647

Figure 5.10: Results of 3 classification models.

5.5.2 Development of Long-Short Term Memory

In recent years, recurrent neural networks (RNNs) have become a popular choice for modeling sequential data, such as time series and natural language. However, LSTM models have shown great potential in predicting various phenomena in the oil and gas industry, from production rates and reservoir performance to equipment health and failure prediction. By leveraging the sequential nature of time-series data, LSTM models can capture complex patterns and relationships that are not easily discernible through traditional statistical approaches.

In order to prepare LSTM modeling, it is common practice to apply normalization techniques to ensure that all features are on a similar scale. In this study, three common normalization techniques: MinMaxScaler, StandardScaler, and RobustScaler were compared, to determine which one would be the most suitable for use with LSTM models in predicting oil field production rates. To evaluate the performance of each scaler, the LSTM model was trained on the normalized data and evaluated their accuracy using mean squared error (MSE) and mean absolute percentage error (MAPE) metrics. We found that the StandardScaler consistently outperformed the other two scalers, achieving lower MSE and MAPE scores across multiple test scenarios. Based on these results, it was concluded that StandardScaler is the most suitable normalization technique for use with LSTM models in predicting oil field production rates. However, it is important to note that the choice of scaler may depend on the specific characteristics of the data and the nature of the prediction task. As such, it is always advisable to experiment with different normalization techniques and evaluate their performance before settling on a final choice.

A Recurrent Neural Network (RNN) model was constructed using Keras' "Sequential" API, layer by layer. The model architecture consists of an embedding layer followed by four LSTM layers, four dropout layers, and a final dense layer. The model was designed to learn patterns in sequential data and make predictions based on them. A visualization of the model architecture is shown in the accompanying figure. The use of multiple LSTM layers allows the model to capture complex relationships between variables over time, while the dropout layers help prevent overfitting. Finally, the dense layer produces a single output value, which represents the predicted outcome of the model. The detailed code for RNN is presented in Appendix B.

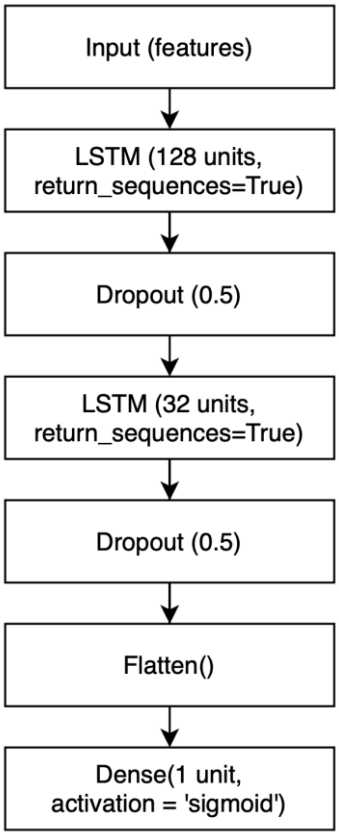


Figure 5.11: LSTM layers in the model architecture

Each layer in the LSTM model serves a specific purpose in learning and predicting patterns in sequential data:

1. Input layer: The input layer defines the shape of the input data, which is a time series with a number of timesteps and features.

2. LSTM layers: The LSTM layers are the main components of the model, and they are responsible for learning the sequential patterns in the input data. In this model, four LSTM layers are used, with decreasing numbers of units. The first LSTM layer has 500 units, which allows it to capture more complex relationships between variables. The subsequent layers have fewer units, which helps the model learn more general patterns in the data.

3. Dropout layers: Dropout layers help prevent overfitting by randomly dropping out some of the nodes in the LSTM layers during training. In this model, a dropout rate of 0.2 is used after each LSTM layer.

4. Dense layer: The final dense layer produces a single output value, which represents the predicted outcome of the model. A sigmoid activation function is used in the dense layer to produce a probability value between 0 and 1.

Overall, the use of multiple LSTM layers with decreasing numbers of units, combined with dropout layers, helps the model capture complex patterns in sequential data while avoiding overfitting. The final dense layer produces a single prediction based on the learned patterns in the input data.

Layer (type)	Output Shape	Param #
lstm_40 (LSTM)	(None, 1, 128)	73216
dropout_40 (Dropout)	(None, 1, 128)	0
lstm_41 (LSTM)	(None, 1, 32)	20608
dropout_41 (Dropout)	(None, 1, 32)	0
flatten_13 (Flatten)	(None, 32)	0
dense_16 (Dense)	(None, 1)	33

Total params: 93,857
Trainable params: 93,857
Non-trainable params: 0

Figure 5.12: Summary of the model with layers

5.5.2.1 Validation of Long-Short Term Memory

This subsection aims to describe the validation of the LSTM model constructed to predict oil production wells' failure. The performance of the model was evaluated using accuracy, which yielded an impressive result of 96.9%. The learning curve of the model represented in Figure 5.13 below with the best validation performance is roughly 0.21 at 11th epoch. The training process required 11 iterations and included 5 validation checks. To avoid overfitting, which is a common problem in training neural network algorithms, the training was stopped when the model reached the maximum level of generalization.

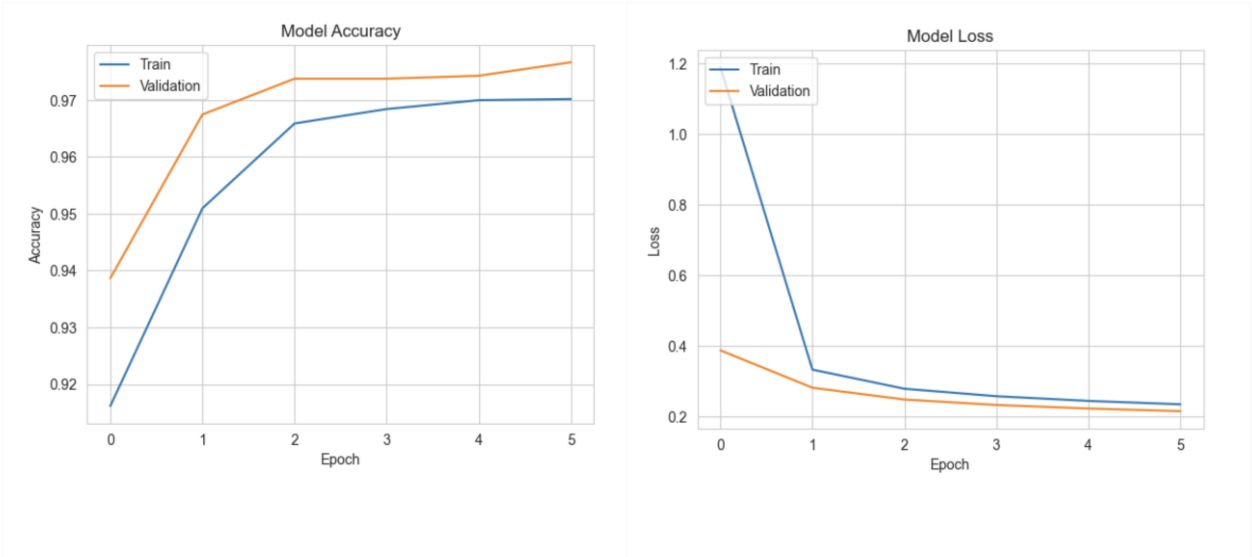
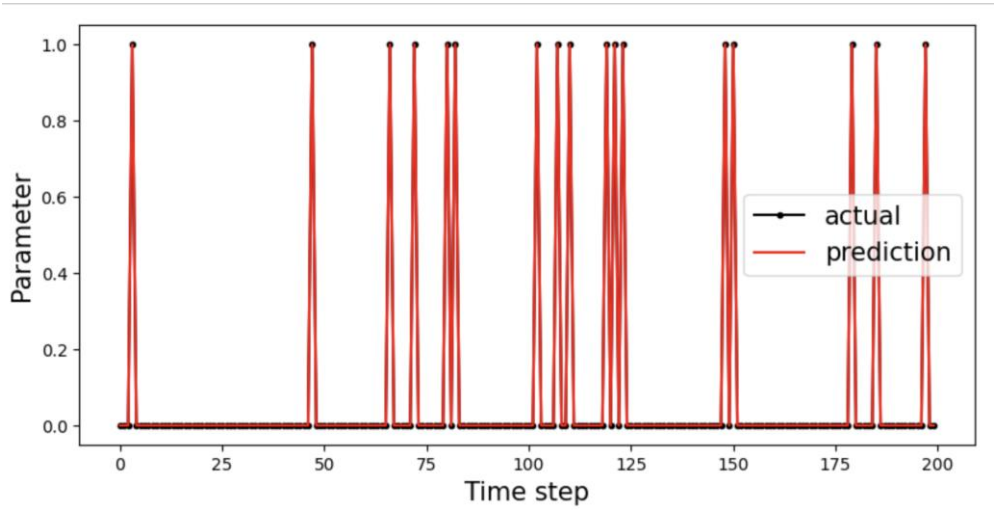


Figure 5.13: ***Performance graph of LSTM model***

The success of the LSTM model can be attributed to several factors. The use of multiple LSTM layers allowed the model to learn and capture complex patterns within the data. The dropout layers were also essential in preventing overfitting and improving the generalization of the model. Lastly, the dense layer with a sigmoid activation function was able to output a probability value, which made it ideal for binary classification tasks.



5.14: Actual vs Predicted data for 200 days.

The Figure 5.15 compared the actual and predicted data using a line graph. The x-axis represents the time step, while the y-axis represents the parameter values. The black line and dots represents the actual values, while the red line represents the predicted ones. The graph depicts that the predicted values were closely following the actual values. The model is performing well in predicting the parameter values, indicating its accuracy in the prediction of production oil wells failures. The good result indicates that the LSTM model is successfully capturing the patterns and trends in the input data and predicting the output values with high accuracy.

A K-fold cross-validation is a widely used technique in ML to assess the performance of a model on a limited data set. In this study, a k-fold cross-validation method to evaluate the performance of an LSTM model (Lyashenko, 2023). This method randomly divides the data set into k equally sized subsets or folds, trains the model on k-1 folds, and tests it on the remaining fold (Lyashenko, 2023). To evaluate the accuracy of the LSTM model 5-fold cross-validation was conducted. The mean accuracy and standard deviation were computed for each fold. The results revealed that the model achieved an average accuracy of 97.5%, with a standard deviation of 0.15 (Figure 5.15), which suggests the model performed consistently well on all folds. This suggests that the model performed consistently well across different folds and did not overfit to the training data. The high accuracy and

low standard deviation further indicate that the model's performance is reliable and can be generalized to new data sets.

```
Score for fold 5: loss of 0.21519042551517487; accuracy of 97.66528606414795%
------------------------------------------------------------------------
Score per fold
------------------------------------------------------------------------
> Fold 1 - Loss: 0.21740877628326416 - Accuracy: 97.68010377883911%
------------------------------------------------------------------------
> Fold 2 - Loss: 0.21763446927070618 - Accuracy: 97.37622141838074%
------------------------------------------------------------------------
> Fold 3 - Loss: 0.21077491343021393 - Accuracy: 97.68010377883911%
------------------------------------------------------------------------
> Fold 4 - Loss: 0.21440406143665314 - Accuracy: 97.37622141838074%
------------------------------------------------------------------------
> Fold 5 - Loss: 0.21519042551517487 - Accuracy: 97.66528606414795%
------------------------------------------------------------------------
Average scores for all folds:
> Accuracy: 97.55558729171753 (+- 0.14655153689994485)
> Loss: 0.21508252918720244
------------------------------------------------------------------------
```

5.15: K-Fold cross-validation result.

Overall, the cross-validation results validate the effectiveness of the LSTM model in predicting the failure of production oil wells. The high accuracy and consistency of the model suggest that it could be a reliable tool for identifying faulty units in the future.

Chapter 6. FRAMEWORK DEVELOPMENT

6.1 Introduction

The establishment of a decision-based framework for perspective maintenance and a framework for failure prediction in the oil and gas industry is covered in this chapter. Both frameworks reflect various processes, such as choosing the best monitoring methods, analyzing the data, and modeling techniques. The chapter is comprised of two parts: dedicated to the framework developed for the decision-based maintenance model selection and the failure prediction model.

6.2 Decision-based Framework for PdM Implementation

As mentioned in the Introduction chapter, the main aim of implementing the model is to decrease equipment maintenance costs caused by downtimes. Therefore, the initial stage of the implementation is to decide which model to adopt for equipment maintenance and control. Figure 6.1 shows a Decision-based Framework for maintenance technique selection.

The Framework is supported by three main steps that help make decisions in the Predictive Maintenance implementation process:

1. To evaluate the need for proactive maintenance techniques;
2. To decide on the type of proactive maintenance;
3. To check the applicability of three different models:
    a. Decision-based model
    b. Knowledge-based model
    c. Physical model.

Any oil and gas company first should evaluate the cost efficiency of using maintenance techniques in the long and short terms by relying on Step 1. If a company has limited resources and a budget for maintenance planning, reactive maintenance would be an optimal choice for them since there are no regular maintenance costs or expenses associated with preventative measures. However, in case of frequent failures of important equipment, proactive maintenance can assist in seeing possible issues early and fixing them before they grow into bigger problems, lowering the risk of equipment failure and saving money on emergency repairs and downtime.

The second important step is to decide the efficiency of two proactive maintenance techniques: preventive and predictive maintenance. The most important difference between them is

that preventive maintenance schedules rely on the original equipment manufacturer's (OEM) recommendation and assign monitoring activities at fixed intervals. If such an approach is efficient, then preventive maintenance (time-based maintenance) is the right choice. However, if the goal is to improve maintenance schedules and cut expenses, Predictive maintenance is the best option to choose.

After the decision on implementing predictive maintenance is made, the next important point is to select the right model for implementation. The knowledge-based approach works only if the company has experienced and competent experts to foresee unpredictable events relying on their knowledge. In case of their absence, the company can consider the possibility of implementing the other two models. To implement a physical model, a thorough understanding of the physics involved in the equipment's operation and the potential failure points are required. As for the data-driven model, high-quality data is the main requirement for implementation.

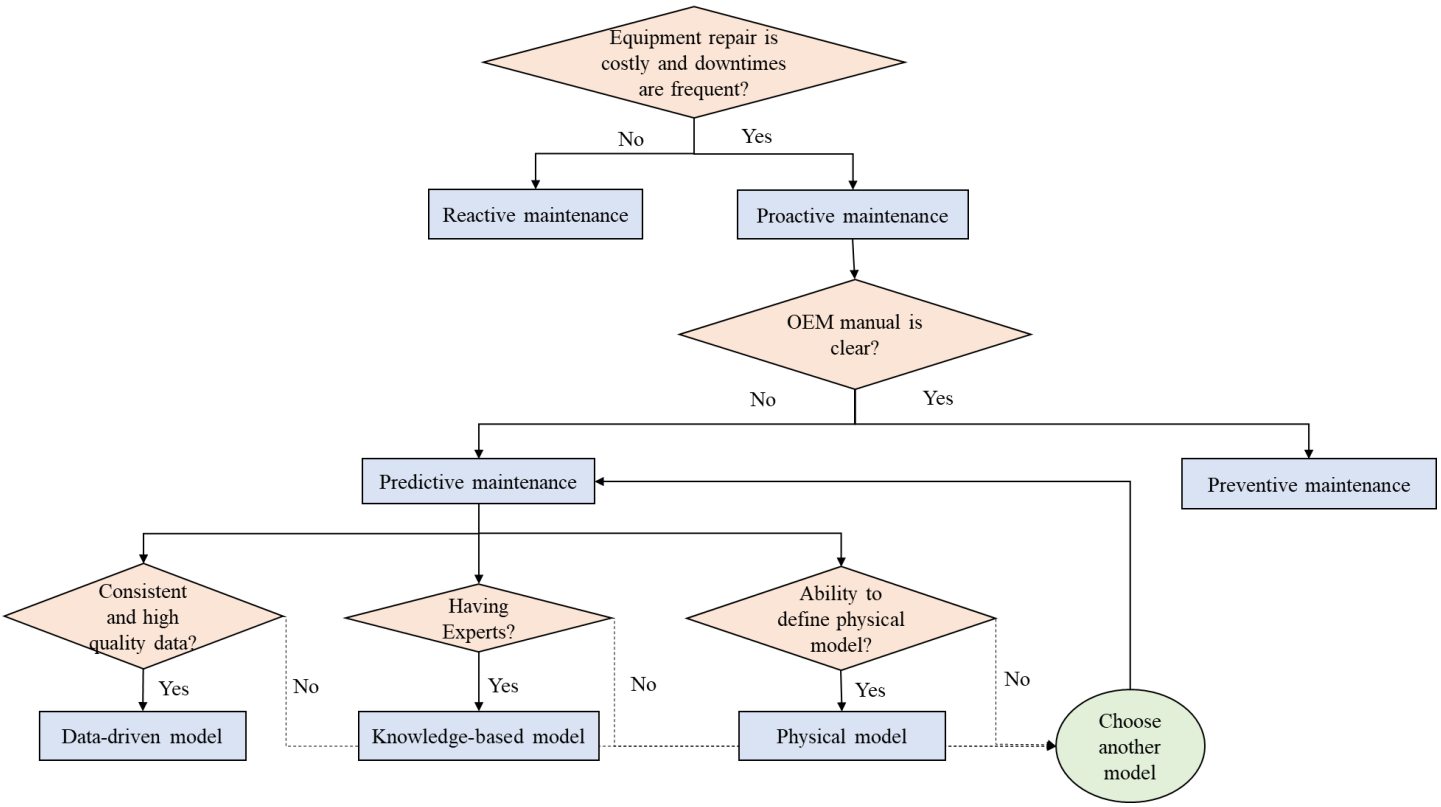


Figure 6.1: Decision-based Framework for Predictive Maintenance model selection

6.3 Framework for Failure Prediction in the Oil and gas industry

This subchapter discusses the detailed components of the framework for failure prediction of oil and gas equipment in the scope of the data-driven PdM. Failure prediction is one of the essential elements of predictive maintenance that can help detect probable equipment failures before they

happen, enabling proactive maintenance and lowering the chance and longevity of unplanned downtime. As shown in Figure 6.2, the main elements of the framework include IoT sensors, a failure prediction model, and an output that informs decision-making.

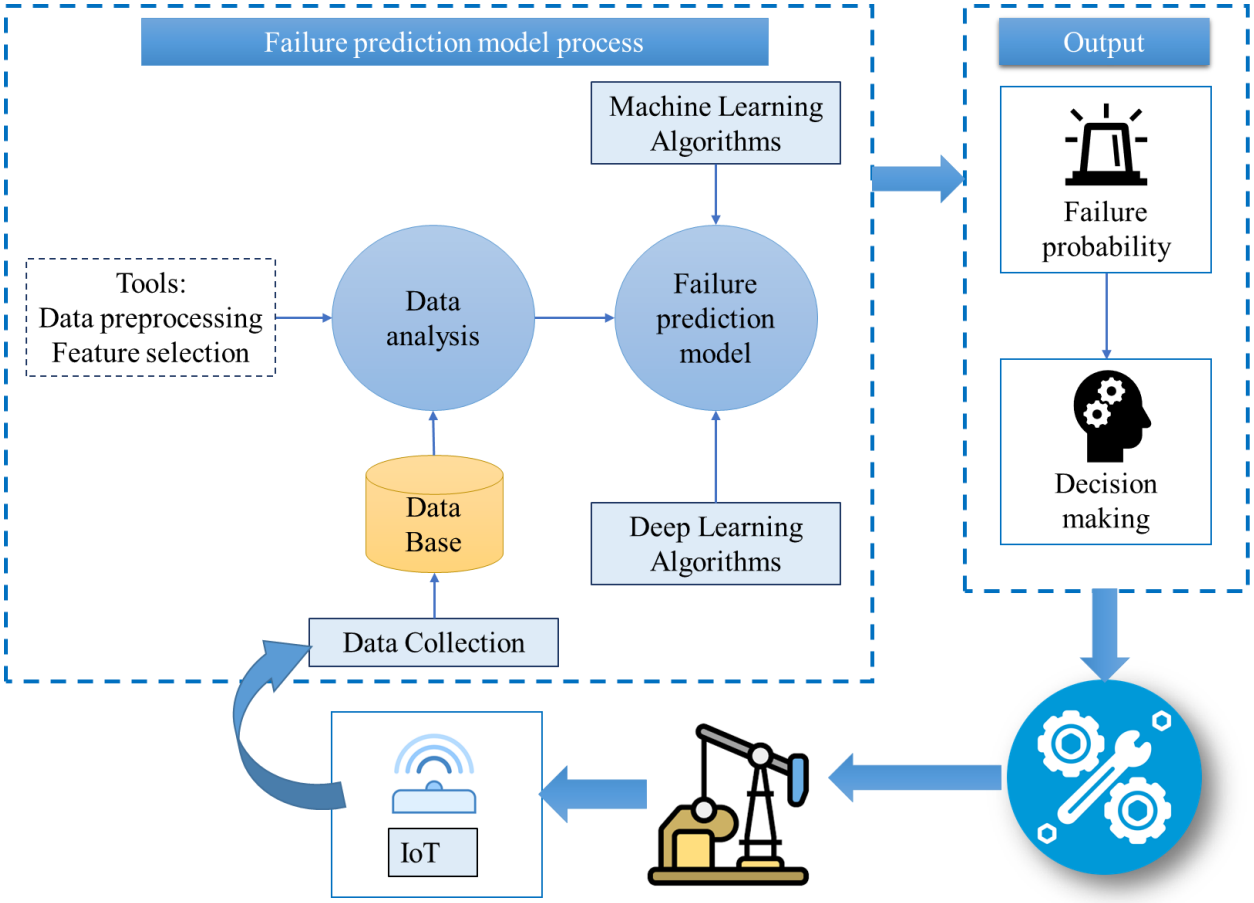


Figure 6.2: Framework for failure prediction in the oil and gas industry

IoT sensors are essential to the framework since they allow for the gathering and transfer of well data. Temperature, pressure, flow rate, vibration, and other variables are among the many factors that these sensors are intended to track. These sensors collect data, which is then sent to a central database for further analysis.

The framework's foundation is the failure prediction model, which analyzes data provided by IoT sensors using powerful algorithms to detect possible problems. The model is intended to detect patterns and anomalies in the data that may suggest potential equipment problems. The model is trained on previous data on equipment failures and other pertinent parameters. A detailed discussion of the process to design a failure prediction model is presented in Chapter 6.2.1.

The model's output is used to inform the decision-making process in the oil and gas industry. When possible faults are discovered, the output can include alarms and notifications. The output can

be linked with current systems and workflows to provide decision-makers with the information they need to act fast and effectively.

6.3.1 Failure Prediction model process

In this subchapter, the steps of the framework for failure prediction of oil and gas equipment are described. Figure 6.3 illustrates the flowchart, which consists of seven main steps: data collection, data preprocessing, feature selection, model training, model evaluation, and deployment. Each step is essential for building an accurate and reliable model for predicting equipment failure. Moreover, the left side of the figure shows the technological enablers required for each step of the model.

1. Data Collection: data is acquired from different sources in this step, including sensors, maintenance records, and equipment logs. Data obtained may include operational circumstances, equipment performance, maintenance activities, and environmental considerations. The purpose of data collection is to acquire as much information as possible about the equipment being monitored. The data collected must be accurate, reliable, and thorough. Before proceeding, any missing or erroneous data must be found and corrected. Since data quality can affect model accuracy, it is critical to ensure that the data is of good quality. Thus, the powerful database that saves high-frequency dynamic data is an essential part of the process.

2. The second stage is data preparation. The collected data is cleaned, processed, and prepared for analysis in this step. Data preprocessing consists of multiple sub-steps, including data cleaning, integration, transformation, and reduction. Data cleaning involves eliminating any redundant or irrelevant data, addressing any errors or inconsistencies, and dealing with missing data. Data integration includes merging data from several sources and guaranteeing consistency and compatibility. Data transformation is the process of turning data into a format that can be analyzed, such as numerical or categorical data. Data reduction is the process of shrinking a data set by deleting duplicated or irrelevant elements.

3. Feature selection. The most relevant and important features from the preprocessed data set are chosen in this step. Identifying the features with the highest predictive power for equipment failure is what feature selection entails. Feature selection strategies can be used to prioritize features and select the top features for the model. The purpose of feature selection is to minimize the number of features in the model while retaining predictive power. Overfitting can occur when there are too many features, while underfitting occurs when there are too few characteristics.

4. Model training is the failure prediction framework's fourth step. Machine learning models are trained using the selected features and the preprocessed data set in this step. Based on the input features, the models are trained to forecast the probability of equipment failure. Several machine learning algorithms, such as linear regression, decision trees, support vector machines (SVM), and neural networks, can be used to forecast failure. The algorithm chosen is determined by the type of data and the complexity of the challenge. The data is separated into training and testing sets during model training. The training set is used to train the model, whereas the testing set is used to assess the model's performance.

5. In the model validation step, the trained model's performance is assessed using measures such as accuracy, precision, cross-validation, recall, and F1-score. The testing set, which was not used during model training, is used to evaluate the model. The purpose of model evaluation is to analyze the model's accuracy and reliability. To measure the model's efficacy, its performance can be compared to other models or industry standards.

6. The final step is deployment and monitoring. Deploying the developed model in a real-time environment for continuous monitoring of equipment health. The model can be integrated with existing monitoring systems or used as a standalone system.

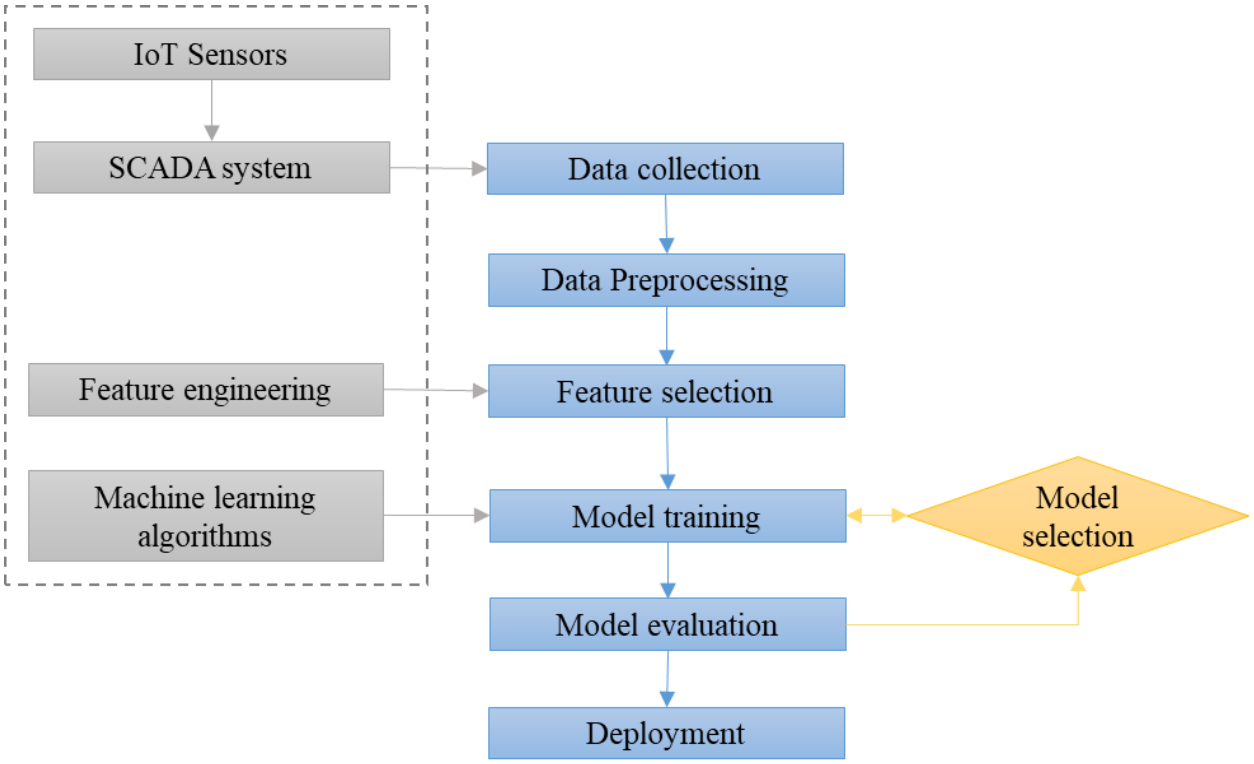


Figure 6.3: Flowchart of the failure prediction model

6.4 TO-BE analysis

In Chapter 4 dedicated to the Current Practice at Caspi Neft, the AS-IS representation of maintenance initiation in the company was presented. From that analysis, it was clear that reactive and preventive maintenance techniques are utilized in the company. The implementation of the failure prediction framework in the company will lead to the modification of the maintenance process in terms of reduced inputs. For instance, all preventive maintenance elements, such as repair plans, maintenance plans, and equipment passports are to be eliminated and replaced by the failure prediction alarm, which is produced by the model trained on the real data from the oilfield. At the same time, due to the accurate and timely prediction, the number of unpredicted failures will be reduced, so there will be no need for a thorough assessment of the state of the equipment. However, it is still not recommended to eliminate this element of reactive maintenance totally due to safety concerns and force-majeure situations. Moreover, due to the failure prediction of the wells, unnecessary closures and downtimes (planned ones) will be excluded, so all the closures will be effective - leading to the maintenance and repair works. Overall, it can be said that the frequency of maintenance initiation will be reduced considerably with the implementation of the Failure Prediction Framework presented in Figure 6.2 of the current chapter.

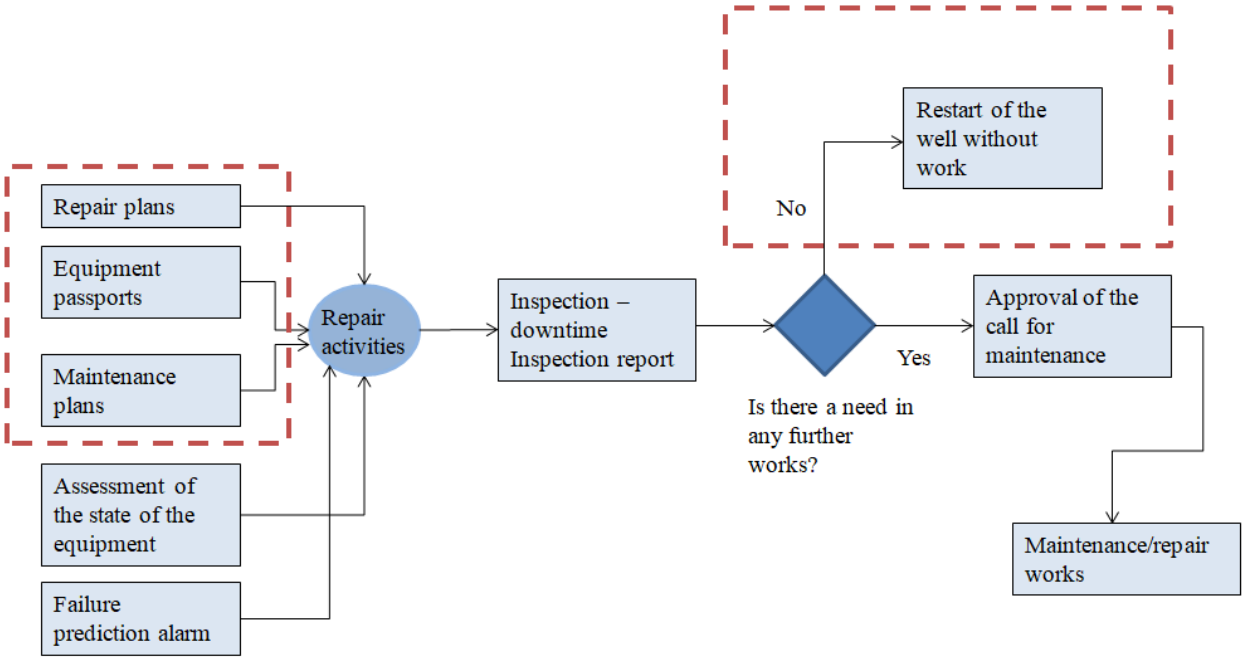


Figure 6.4: TO-BE representation of the maintenance initiation in the company

CHAPTER 7. VALIDATION

Validation is a crucial part of any project because it allows teams to verify the correctness of the results obtained and conclusions drawn. In this project both quantitative validation in a form of cross-validation of the model mentioned in 5.5.2.1 part of the project as well as qualitative validation described in this chapter were performed. Due to the complexity of the current project, assistance from both industrial (oil and gas) and technical (machine learning, data analysis) experts was required. Therefore, two vital moments of the project, such as scope alteration and data acquisition were consulted with both oil and gas oilfield and technical experts. Subsequently, these experts validated two major deliverables of the project: the failure prediction model and framework. At the same time, the benefits of the validation were the chance to obtain it from people, who have expertise in both the oil and gas industry as well as data analysis and ML techniques. Qualitative validation considered both the model and framework developed under the current project, while the control over the model itself was also implemented. The qualitative validation from experts is illustrated in Table 7.1 below.

Table 7.1: Experts' validation

Model/ Framework	**Expert**	**Position/ Area of Expertise**	**Validation**	**Recommendation**	**Project Adjustments**
Model & Framework	Expert 1	Industrial partners of this projects, which provide IT services for big enterprises, including Caspi Neft oil and gas company	Model with high accuracy and comprehensive framework. The best result possible from the given data	To develop more models for other equipment based on this projects	Addition of the recommendation to the future work

Model & Framework	Expert 2	Automation engineer and expert of digitalization of oil and gas entities	Expert expressed an opinion that with the given quality of the data, the model works very good and accuracy is very high. At the same time, the AS-IS and TO-BE states of the companies with the implementation of the failure prediction framework were approved.	To use more dynamic data for the failure prediction model to obtain more reliable results; to take into account advantage of reduced spare parts for the equipment.	The benefit of the framework implementation in terms of inventory management were included.
Model & Framework	Expert 3	An expert from academia with cross-functional expertise in both oil and gas projects and data analysis	Reasonable results of the model and comprehensive framework	Elaboration on the data problems and challenges, focus on failure prediction, not maintenance	The names of the models and framework were adjusted

At the same time, there were four different types of meetings, which were conducted throughout the project with different stakeholders. All of them influenced the results of the project and assisted in the improvement of the work. During the meetings with oil and gas experts, the necessity of the new dataset was revealed and proved. Meetings with cross-functional experts contributed to the change of the scope and consequently the title of the project. And last but not least is the regular meetings with our academic and industrial supervisors with weekly presentations and reports, which guided the workflow and provided feedback for the work performed.

Table 7.2: List of meetings with Air Astana and academic supervisors

Type of meeting	Frequency	Attendants	Discussions
Weekly meetings with academic and industrial supervisors	Once a week	Academic supervisor Industrial supervisors Team	Progress of the project Future work Current challenges
Meeting with industrial supervisors	By necessity	Industrial supervisors Team	Mostly comments regarding the data required for the project
Meeting with oil and gas expert (Expert 2)	By necessity	Industrial supervisors Expert 2 Team	Verification of the parameters included in the model Consultation about the quality of the data
Meeting with a cross-functional expert (Expert 3)	By necessity	Team Expert 3	The changing scope of the project

There were four different types of meetings, which were conducted throughout the project with different stakeholders. All of them influenced the results of the project and assisted in the improvement of the work. During the meetings with oil and gas experts, the necessity of the new dataset was revealed and proved. Meetings with cross-functional expert contributed to the change of the scope and consequently the title of the project. And last but not least is the regular meetings with our academic and industrial supervisors with weekly presentations and reports, which guided the workflow and provided feedback for the work performed.

CHAPTER 8. CONCLUSION AND FUTURE WORK

8.1 Conclusion

The project aimed to develop the well failure prediction model and framework for its implementation in the oil and gas sector. For this purpose, the study was conducted in collaboration with the oil and gas enterprise Caspi Neft, which initiated the project of “Smart Oilfield ". The current maintenance practices of the company consider reactive repair works and scheduled maintenance activities based on the manufacturer's recommendations, which are estimated to be impractical in terms of operational costs and costs related to downtimes. Therefore, there is a need for the implementation of a failure prediction model in the scope of Condition-based Maintenance.

From the comprehensive literature analysis, it was concluded that the most appropriate model in the scope of predictive maintenance is the Data-driven model. Thus, the study completed the failure prediction model using machine learning algorithms such as LSTM. The main finding of the project is the failure prediction model with an accuracy of 90-95%. The proposed model is capable of predicting the possible failure and calculating the probability of its occurrence. Additionally, the model was validated by experts from the oil and gas field and machine learning engineers to investigate its relevance and precision.

Besides the model, the second delivery of the project is the framework for the failure prediction model for oil and gas equipment. The prerequisite to the framework is to investigate the implementation relevance of PdM according to a decision-based framework. The failure prediction framework consists of IoT devices that collect all required data such as sensors, maintenance records, and equipment logs, the failure prediction model itself and the output of the model in the form of the system alarm. The maintenance actions that would be carried out according to the system’s alarm is the next step of the advanced framework, which is not in the scope of this project.

Overall, all set objectives are achieved successfully, starting from the preliminary research on the oil and gas industry and assessing the current capabilities of the company, generating a failure prediction model using identified optimal parameters, followed by its validation and adaptable framework for implementation of this model. The further subchapter describes the limitations of the project that restricted the scale of the project.

8.2 Limitations and Challenges

As was mentioned in the Literature review, the quality and consistency of the data is the key success factor for the failure prediction model in the scope of Data-driven PdM. Unfortunately, due to the sensitive nature of the strategic direction of the industry, data availability and quantity of information is the inhibitor to research in this area, including the assessment of the accuracy of ML algorithms.

The main limitation of the project is related to the datasets provided by the company:

- The confidentiality of the data in the oil industry led to insufficient parameters for the model construction. As a result, there was a lack of information on the following features: downtime history and comments on failure type, physical parameters and types of well, and location of the assets. Moreover, the company representatives altered the actual names of the wells which led to an inaccurate and inconsistent dataset due to human factors. Additionally, the clear classification of the wells according to the pump types such as centrifugal, screw, and beam-balanced pumping units was not provided. The major issue related to confidentiality is the unwillingness of the company to provide well logs with failures type, which was not agreed upon in the initiation stage of the project. Due to this fact, the model couldn't classify the types of downtime based on provided parameters and was limited only to the probability of undesirable events.

- The inadequate collection of data with low-frequency results in 24-hour period prediction, which might not be considered fully cost-efficient. Ideally, for the model, a dynamic dataset with at least 2 hours of frequency is required. However, since the database of the company is recorded only on a daily basis, consequently, the model gives predictions only for this period of time.

- The quality of the dataset transferred first was questionable due to the following reasons: missing data cells, duplication of data and dates, inadequate range of parameters(such as pressure), and inconsistency between various cells and dataset lists. As a result, such data cannot be utilized for the ML algorithm as it is very dependent on the quality of the data. Moreover, feature engineering and other data preprocessing techniques are not useful in coping with such problems due to the absence of data and logic. Therefore, another dataset was requested from the company to proceed further. Luckily, the new dataset obtained did not possess the same problems, however, it had its own limitations as well, for instance, the absence of corresponding downtimes of the wells. This fact totally changed the approach of

the modeling. Overall, these data quality problems lead to significant changes in the Methodology of the project, which is reflected in the project's flow chart in Chapter 3 of the current report.

- The initial scope of the project was altered due to the insufficient data collection in Caspi Neft. As Predictive Maintenance originally focuses on mechanical equipment, the initial system for consideration was oil and gas pumps. However, after a thorough literature review and required parameters identification, it was revealed that the company is not able to provide such data for the model because they did not install suitable IoT sensors. As a result, the scope of the project was changed from pumps to well systems due to the variability and availability of data from sensors.

- The environmental factors such as geological parameters and location information were not considered in the model construction due to the lack of such data. If the company provided all parameters mentioned in the fishbone diagrams in Chapter 5, not mechanical failures such as watercut would be predicted.

8.3 Further Work

If the abovementioned limitations related to data would be reconsidered, the applicability of the proposed model would be extended as follows:

- Firstly, the created model could be improved in terms of versatility by adding an additional function for the identification of possible failure types. For this purpose, downtime history with corresponding collected parameters should be provided.

- Secondly, the model could be further ameliorated in terms of precision by increasing the frequency of the monitored data (recommended at least every 2 hours). As it was mentioned in the framework, high-frequency dynamic datasets are required to obtain the model that can inform about the failure in the optimum timeframe.

- Thirdly, the suggested model could be also enhanced in terms of the scope of the failures, extending it from the mechanical to natural undesirable events.

Reference list

Aalsalem, M. Y., Khan, W. Z., Gharibi, W., & Armi, N. (2017). An intelligent oil and gas well monitoring system based on Internet of Things. In *2017 International Conference on Radar, Antenna, Microwave, Electronics, and Telecommunications (ICRAMET)*, (p. 124-127), 10.1109/ICRAMET.2017.8253159

Abbasi, T., Lim, K. H., Rosli, N. S., Ismail, I., & Ibrahim, R. (2018, August). Development of predictive maintenance interface using multiple linear regression. In *2018 International Conference on Intelligent and Advanced System (ICIAS)* (pp. 1-5). IEEE.

Abbasi, T., Lim, K. H., & San Yam, K. (2019, April). Predictive maintenance of oil and gas equipment using recurrent neural network. In *Iop conference series: Materials science and engineering* (Vol. 495, No. 1, p. 012067). IOP Publishing.

Abdalla, R., Samara, H., Perozo, N., Carvajal, C. P., & Jaeger, P. (2022). Machine Learning Approach for Predictive Maintenance of the Electrical Submersible Pumps (ESPs). ACS omega, 7(21), 17641-17651.

Adenuga, O. D., Diemuodeke, O. E., & Kuye, A. O. (2023). Development of Maintenance Management Strategy Based on Reliability Centered Maintenance for Marginal Oilfield Production Facilities. *Engineering*, *15*(3), 143-162.

Alakbari, F. S., Mohyaldinn, M. E., Ayoub, M. A., Muhsan, A. S., & Hussein, I. A. (2021). A robust fuzzy logic-based model for predicting the critical total drawdown in sand production in oil and gas wells. *PloS one*, *16*(4), e0250466.

Alazemi, A. S. K. R., Ali, M. Y., & Daud, M. R. C. (2019). Preventive maintenance of boiler: A case of Kuwait industry. *International Journal of Engineering Materials and Manufacture*, *4*(2), 48-58.

Al-Fadhli, W., Kurma, R., Bhatia, K., Alboueshi, A., & Abdelbaky, A. (2020). Water Control in High-Water-Cut Well Using Cutting-Edge Polymers: Production Optimization Methodology Applied in Burgan Field, Kuwait. In *SPE Nigeria Annual International Conference and Exhibition*, https://doi.org/10.2118/203709-MS

Ahmad, R., & Kamaruddin, S. (2012). An overview of time-based and condition-based maintenance in industrial application. *Computers & industrial engineering*, *63*(1), 135-149.

Alimohammadi, H. (2018, May 14). Diagnosing and Attacking Excessive Water Production. [Image attached]. LinkedIn. https://www.linkedin.com/pulse/diagnosing-attacking-excessive-water-production-hamzeh-alimohammadi/

Bailey, B., Crabtree, M., Turie, J., Elphick, J., Kuchuk, F., Romano, C., and Roodhart, L. (2000), Water Control, *Spring*, (30-51), https://www.petroleumengineers.ru/sites/default/files/water_production_solving.pdf

Baker, H. (2016). The Impact of Digital on Unplanned Downtime: An Offshore Oil And Gas Perspective.

Bevilacqua, M., & Braglia, M. (2000). The analytic hierarchy process applied to maintenance strategy selection. *Reliability Engineering & System Safety*, *70*(1), 71-83.

Biagetti, T., & Sciubba, E. (2004). Automatic diagnostics and prognostics of energy conversion processes via knowledge-based systems. *Energy*, *29*(12-15), 2553-2572.

Blanchard, B. S., Verma, D. C., & Peterson, E. L. (1995). *Maintainability: a key to effective serviceability and maintenance management* (Vol. 13). John Wiley & Sons.

Bousdekis, A., Apostolou, D., & Mentzas, G. (2019). Predictive maintenance in the 4th industrial revolution: Benefits, business opportunities, and managerial implications. *IEEE Engineering Management Review*, *48*(1), 57-62, doi: 10.1109/EMR.2019.2958037

Brotherton, T., Jahns, G., Jacobs, J., & Wroblewski, D. (2000, March). Prognosis of faults in gas turbine engines. In *2000 IEEE Aerospace Conference. Proceedings (Cat. No. 00TH8484)* (Vol. 6, pp. 163-171). IEEE.

Chervinskiy, O. (2022). Чем отзовется троллинг предпринимателей за получение дивидендов. [What will be the response to trolling entrepreneurs for receiving dividends]. Inbusiness.kz. Retrieved on March 25, 2023, from https://inbusiness.kz/ru/news/chem-otzovetsya-trolling-predprinimatelej-za-poluchenie-dividendov

Cline B., Niculescu R. S., Huffman D. & Deckel B., Predictive maintenance applications for machine learning. (2017). *Annual Reliability and Maintainability Symposium (RAMS)*, Orlando, FL, USA, 2017, pp. 1-7.

Davies, R. J., Almond, S., Ward, R. S., Jackson, R. B., Adams, C., Worrall, F., ... & Whitehead, M. A. (2015). Reply:“Oil and gas wells and their integrity: Implications for shale and unconventional resource exploitation”. *Marine and Petroleum Geology*, *59*, 674-675.

Duffuaa, S. O., Ben-Daya, M., Al-Sultan, K. S., & Andijani, A. A. (2001). A generic conceptual simulation model for maintenance systems. *Journal of Quality in Maintenance Engineering*, *7*(3), 207-219.

Elichev, V., Bilogan, A., Litvinenko, K., Khabibullin, R., Alferov, A., & Vodopyan, A. (2019, October). Understanding events well with machine learning. In *SPE Russian Petroleum Technology Conference*. OnePetro, 1-12, https://doi.org/10.2118/196861-MS.

Elwerfalli, A., & Al-Maqespi, S. (2021). Selection of Appropriate Maintenance Strategy for Oil and Gas Equipment Using Analytical Hierarchy Process (AHP). In *Proceedings of the International Conference on Industrial Engineering and Operations Management Sao Paulo, Brazil*.

Eghbali, S., Ayatollahi, S., & Boozarjomehry, R. B. (2016). New expert system for enhanced oil recovery screening in non-fractured oil reservoirs. *Fuzzy Sets and Systems*, *293*, 80-94.

Energy Education. (n.d.). *Oil well*. Retrieved March 23, 2023, from https://energyeducation.ca/encyclopedia/Oil_well

Figueroa Barraza, J., Guarda Bräuning, L., Benites Perez, R., Morais, C. B., Martins, M. R., & Droguett, E. L. (2022). Deep learning health state prognostics of physical assets in the Oil and Gas industry. *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability*, *236*(4), 598-616.

Galushko, M. (2023). “Каспий нефть” победила в номинации “Лучшее цифровое решение”. [Caspi Neft won the nomination "The best digital solution]. Inbusiness.kz

Gatta F., Giampaolo F., Chiaro D., & Piccialli F. (2022). Predictive maintenance for offshore oil wells by means of deep learning features extraction. *Expert System*s.

Hassannayebi, E., Nourian, R., Mousavi, S. M., Alizadeh, S. M. S., & Memarpour, M. (2022). Predictive analytics for fault reasoning in gas flow control facility: A hybrid fuzzy theory and expert system approach. *Journal of Loss Prevention in the Process Industries*, *77*, 104796.

Hausler, R. H., Krishnamurthy, R. M., & Sherar, B. W. (2015, March). Observation of productivity loss in large oil wells due to scale formation without apparent production of formation brine. In *CORROSION 2015*. OnePetro.

IEA (2022), *Kazakhstan 2022 Energy Sector Review*, OECD Publishing, Paris, https://doi.org/10.1787/73d1d69f-en.

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences*, 374(2065). https://doi.org/10.1098/rsta.2015.0202

Karatayev, M., & Clarke, M. L. (2014). Current energy resources in Kazakhstan and the future potential of renewables: A review. *Energy Procedia*, *59*, 97-104, doi:10.1016/j.egypro.2014.10.354

Kewen, L., Yangtze, U., Xianghai, R., Li, L., & Xiaodong, F. (2011, May). A new model for predicting water cut in oil reservoirs. In *SPE EUROPEC/EAGE Annual Conference and Exhibition*. OnePetro., 1-8, https://doi.org/10.2118/143481-MS

King, G.(n.d.). *Well Orientation.* https://www.e-education.psu.edu/png301/node/648

Kumar, S. (202, October 18). Silhouette Method — Better than Elbow Method to find Optimal Clusters. Medium. https://towardsdatascience.com/silhouette-method-better-than-elbow-method-to-find-optimal-clusters-378d62ff6891

Labib, A. W. (2004). A decision analysis model for maintenance policy selection using a CMMS. *Journal of Quality in Maintenance Engineering*, *10*(3), 191-202.

Li, K., Xiong, M., Li, F., Su, L., & Wu, J. (2019). A novel fault diagnosis algorithm for rotating machinery based on a sparsity and neighborhood preserving deep extreme learning machine. *Neurocomputing*, *350*, 261-270.

Lindemann, B., Müller, T., Vietz, H., Jazdi, N. & Weyrich, M. (2021). A survey on long short-term memory networks for time series prediction. Procedia CIRP. 99. 650-655. 10.1016/j.procir.2021.03.088.

Long Short-Term Memory (LSTM) - Dive into Deep Learning 1.0.0-beta0 documentation. (n.d.). Retrieved April 16, 2023, from https://d2l.ai/chapter_recurrent-modern/lstm.html

Luo, J., Bixby, A., Pattipati, K., Qiao, L., Kawamoto, M., & Chigusa, S. (2003, October). An interacting multiple model approach to model-based prognostics. In *SMC'03 Conference Proceedings. 2003 IEEE International Conference on Systems, Man and Cybernetics. Conference Theme-System Security and Assurance (Cat. No. 03CH37483)* (Vol. 1, pp. 189-194). IEEE.

Lyashenko, V. (2023). Cross-Validation in Machine Learning: How to Do It Right. neptune.ai. https://neptune.ai/blog/cross-validation-in-machine-learning-how-to-do-it-right

Madrid, J., & Min, A. (2020). Reducing Oil Well Downtime with a Machine Learning Recommender System, [Master's thesis, Massachusetts Institute of Technology] https://hdl.handle.net/1721.1/126390

Moya, M. C. C. (2004). The control of the setting up of a predictive maintenance programme using a system of indicators. *Omega*, *32*(1), 57-75.

Mazumder, R. K., Salman, A. M., & Li, Y. (2021). Failure risk analysis of pipelines using data-driven machine learning algorithms. *Structural safety*, *89*, 102047.

Ngu, K. M., Philip, N., & Sahlan, S. (2019). Proactive and predictive maintenance strategies and application for instrumentation & control in the oil & gas industry. *International Journal of Integrated Engineering*, *11*(4).

Orrù, P. F., Zoccheddu, A., Sassu, L., Mattia, C., Cozza, R., & Arena, S. (2020). Machine learning approach using MLP and SVM algorithms for the fault prediction of a centrifugal pump in the oil and gas industry. *Sustainability*, *12*(11), 4776.

Pyramidenvironmental.(n.d.). *Newsletter: Supply wells, Monitor Wells, and Oil Wells – What's the Difference?* Retrieved March 20, 2023, from https://pyramidenvironmental.com/september-2015-newsletter-supply-wells-monitor-wells-and-oil-wells-whats-the-difference/

Peng, Y., Dong, M., & Zuo, M. J. (2010). Current status of machine prognostics in condition-based maintenance: a review. *The International Journal of Advanced Manufacturing Technology*, *50*, 297-313.

Qian, W., Li, S., Wang, J., & Wu, Q. (2018). A novel supervised sparse feature extraction method and its application on rotating machine fault diagnosis. *Neurocomputing*, *320*, 129-140.

Rani A., Kumar N., Kumar J., Sinha N.K. (2022). Machine learning for soil moisture assessment. *In Cognitive Data Science in Sustainable Computing*, 143-168.

Sciencealpha. (2019, October 3). *Oil well types, structure, construction and development phases.* https://sciencealpha.com/oil-well-types-structure-construction-and-development-phases

Siami Namini, S., Tavakoli, N. & Siami Namin, A. (2019). The Performance of LSTM and BiLSTM in Forecasting Time Series. 3285-3292. 10.1109/BigData47090.2019.9005997.

Sircar, A., Yadav, K., Rayavarapu, K., Bist, N., & Oza, H. (2021). Application of machine learning and artificial intelligence in oil and gas industry. *Petroleum Research*, *6*(4), 379-391.

Senouci, A., El-Abbasy, M. S., & Zayed, T. (2014). Fuzzy-based model for predicting failure of oil pipelines. *Journal of Infrastructure Systems*, *20*(4), 04014018.

Sheu, S. H., Griffith, W. S., & Nakagawa, T. (1995). Extended optimal replacement model with random minimal repair costs. *European Journal of Operational Research*, *85*(3), 636-649.

Shadravan, A., & Amani, M. (2012). HPHT 101-what petroleum engineers and geoscientists should know about high pressure high temperature wells environment. *Energy Science and Technology*, *4*(2), 36-60, DOI:10.3968/J.EST.1923847920120402.635

Susto, G. A., Schirru, A., Pampuri, S., McLoone, S., & Beghi, A. (2014). Machine learning for predictive maintenance: A multiple classifier approach. *IEEE transactions on industrial informatics*, *11*(3), 812-820.

Tam, A. S. B., Chan, W. M., & Price, J. W. H. (2006). Optimal maintenance intervals for a multi-component system. *Production Planning and Control*, *17*(8), 769-779

Tahmourpour, F., & Griffith, J. E. (2007). Use of finite element analysis to engineer the cement sheath for production operations. *Journal of Canadian Petroleum Technology*, *46*(05), https://doi.org/10.2118/07-05-TN1

Theyab, M. (2018). Severe Slugging Control: Simulation of Real Case Study. *J Environ Res*, *2*(*1*), 5.

Vargas, R. E. V., Munaro, C. J., Ciarelli, P. M., Medeiros, A. G., do Amaral, B. G., Barrionuevo, D. C., ... & Magalhães, L. P. (2019). A realistic and public dataset with rare undesirable real events in oil wells. *Journal of Petroleum Science and Engineering*, 181 *(106223)*, 1-9, https://doi.org/10.1016/j.petrol.2019.106223

Wanasinghe, T. R., Gosine, R. G., James, L. A., Mann, G. K., De Silva, O., & Warrian, P. J. (2020). The internet of things in the oil and gas industry: a systematic review. *IEEE Internet of Things Journal*, *7*(9), 8654-8673.

Wang, Y., Pan, Z., Zheng, J., Qian, L. & Mingtao, L. (2019). A hybrid ensemble method for pulsar candidate classification. *Astrophysics and Space Science.*

Wang, Q., Song, Y., Zhang, X., Dong, L., Xi, Y., Zeng, D., ... & Luo, H. (2023). Evolution of corrosion prediction models for oil and gas pipelines: From empirical-driven to data-driven. *Engineering Failure Analysis*, 107097.

Wood Mackenzie. (2022). Airankul (Caspi Neft). Retrieved from https://www.woodmac.com/reports/upstream-oil-and-gas-airankul-caspi-neft-73375060/

Wu, Z., Wu, G., Xing, X., Yang, J., Liu, S., Xu, H., & Cheng, X. (2022). Effect of hole on oil well cement and failure mechanism: application for oil and gas wells. *ACS omega*, *7*(7), 5972-5981, https://doi.org/10.1021/acsomega.1c06275

Zhang, Z., Li, S., Wang, J., Xin, Y., & An, Z. (2019). General normalized sparse filtering: A novel unsupervised learning method for rotating machinery fault diagnosis. *Mechanical Systems and Signal Processing*, *124*, 596-612.

АО «Каспий Нефть». (n.d.). [Caspi Neft JSC]. Forbes Kazakhstan. Retrieved on March 16, 2023, from https://forbes.kz/ranking/object/897

Appendices

Appendix A: Classification Models

```
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
import numpy as np
from sklearn.cluster import KMeans
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.svm import SVC
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score, mean_squared_error
```

```
from sklearn.metrics import accuracy_score, confusion_matrix
from sklearn.metrics import silhouette_score

df = pd.read_excel('KMG.xlsx', engine='openpyxl')
data = df.drop(['Date', 'Well_name', 'Dynamic_fluid_level', 'Water_cut', 'Gas_rate', 'Line_pressure'],
axis=1)

kmeans = KMeans(n_clusters=4, init='k-means++', random_state=42)
kmeans.fit(data)

data['Cluster'] = kmeans.labels_

silhouette_score(data, kmeans.labels_)

from sklearn.metrics import f1_score, precision_score, recall_score, accuracy_score,
classification_report
import time

model_performance = pd.DataFrame(columns=['Accuracy', 'Precision',\
                                'Recall', 'F1-Score', 'Training time',\
                                'Prediction time'])

def log_scores(model_name, y_test, y_predictions):
    accuracy = accuracy_score(y_test, y_predictions)
    precision = precision_score(y_test, y_predictions, average='weighted')
    recall = recall_score(y_test, y_predictions, average='weighted')
    precision = precision_score(y_test, y_predictions, average='weighted')
    f1 = f1_score(y_test, y_predictions, average='weighted')

    # save the scores in model_performance dataframe
    model_performance.loc[model_name] = [accuracy, precision, recall, f1,
                            end_train-start, end_predict-end_train]


X = data.drop(["Failure","Cluster"], axis=1)
y = data["Cluster"]

# Split the data into training and test set
X_train, X_test, y_train, y_test = train_test_split(X, y,
                                        test_size = 0.25,
                                        random_state = 42)

start = time.time()
model = DecisionTreeClassifier(max_depth = 8).fit(X_train, y_train)
end_train = time.time()
y_predictions = model.predict(X_test)
end_predict = time.time()

# evaluate the model
log_scores("Decision Tree", y_test, y_predictions)
```

```
from sklearn.model_selection import learning_curve

train_sizes, train_scores, test_scores = learning_curve(model, X, y, cv=5)

# plot the learning curve
plt.plot(train_sizes, train_scores.mean(axis=1), label='Training score')
plt.plot(train_sizes, test_scores .mean(axis=1), label='Testing score')
plt.legend(loc='best')
plt.xlabel('Training set size')
plt.ylabel('Accuracy score')
plt.title('Decision Tree Learning Curve')
plt.show()

start = time.time()
model = RandomForestClassifier(n_estimators=100, n_jobs=-1,
                               random_state=0, bootstrap=True).fit(X_train, y_train)
end_train = time.time()
y_predictions = model.predict(X_test)
end_predict = time.time()

# evaluate the model
log_scores("Random Forest", y_test, y_predictions)
print("Random Forest Model\n" + classification_report(y_test, y_predictions))

# plot the learning curve
plt.plot(train_sizes, train_scores.mean(axis=1), label='Training score')
plt.plot(train_sizes, test_scores .mean(axis=1), label='Testing score')
plt.legend(loc='best')
plt.xlabel('Training set size')
plt.ylabel('Accuracy score')
plt.title('Random Forest Learning Curve')
plt.show()


# train the SVM model
start_train = time.time()
model = SVC(kernel='linear', random_state=42).fit(X_train, y_train)
end_train = time.time()
y_predictions = model.predict(X_test) # predictions from the testset
end_predict = time.time()


# evaluate the model
log_scores("SVM", y_test, y_predictions)
print("SVM Model\n" + classification_report(y_test, y_predictions))

# plot the learning curve
plt.plot(train_sizes, train_scores.mean(axis=1), label='Training score')
plt.plot(train_sizes, test_scores .mean(axis=1), label='Testing score')
plt.legend(loc='best')
plt.xlabel('Training set size')
```

```
plt.ylabel('Accuracy score')
plt.title('SVM Learning Curve')
plt.show()

model_performance
```

Appendix B: LSTM model

```
import pandas as pd
import numpy as np
import tensorflow as tf
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler, OneHotEncoder
from sklearn.compose import ColumnTransformer
from keras.models import Sequential
from keras.layers import Dense, LSTM, Dropout
import keras
from keras.callbacks import EarlyStopping
import matplotlib.pyplot as plt
import seaborn as sns

df = pd.read_excel('KMG.xlsx', engine='openpyxl')
df = df.drop(['Well_name','Date'], axis=1)

X = df.iloc[:, :-1].values
y = df.iloc[:, -1].values

# feature scaling
sc = StandardScaler()
X = sc.fit_transform(X)

# split the data into training and testing sets (80/20)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# reshape input data into 3D tensor (samples, timesteps, features)
X_train = np.reshape(X_train, (X_train.shape[0], 1, X_train.shape[1]))
X_test = np.reshape(X_test, (X_test.shape[0], 1, X_test.shape[1]))

model = Sequential()
model.add(LSTM(units=128, return_sequences=True, input_shape=(X_train.shape[1],
X_train.shape[2])))
model.add(Dropout(0.5))
model.add(LSTM(units=32, return_sequences=True))
model.add(Dropout(0.5))
model.add(Flatten())
model.add(Dense(units=1, activation='sigmoid'))

model.compile(optimizer='adam',loss='binary_crossentropy', metrics=['accuracy'])

early_stop = EarlyStopping(monitor='val_loss', patience=5, min_delta=0.001, mode='max',
verbose=1)
history = model.fit(X_train, y_train, epochs=50, batch_size=64, validation_data=(X_test, y_test),
callbacks=[early_stop], verbose=2)

# plot accuracy
plt.plot(history.history['accuracy'])
```

```
plt.plot(history.history['val_accuracy'])
plt.title('Model Accuracy')
plt.ylabel('Accuracy')
plt.xlabel('Epoch')
plt.legend(['train', 'val'], loc='upper left')
plt.show()

# evaluate the model
_, accuracy = model.evaluate(X_test, y_test)
print('Accuracy: %.2f' % (accuracy*100))

# plot loss
plt.plot(history.history['loss'])
plt.plot(history.history['val_loss'])
plt.title('Model Loss')
plt.ylabel('Loss')
plt.xlabel('Epoch')
plt.legend(['train', 'val'], loc='upper left')
plt.show()

# predict on test set
y_pred = model.predict(X_test)
y_pred = (y_pred > 0.5)

# create new table with predicted and test values
df_results = pd.DataFrame({'Actual': y_test, 'Predicted': y_pred.ravel()})
print(df_results)

nearest_failure_prob = model.predict(X_test)
nearest_failure = np.round(nearest_failure_prob).flatten()

# Create a new DataFrame to store the predicted values
results_df = pd.DataFrame({
    'Actual Failure': y_test,
    'Predicted Nearest Failure': nearest_failure,
    'Probability of Nearest Failure': nearest_failure_prob.flatten()
})

# Compare the predicted values with the actual values
results_df['Correct Prediction'] = results_df['Actual Failure'] == results_df['Predicted Nearest
Failure']
accuracy = results_df['Correct Prediction'].mean()

# Print the accuracy
print("Accuracy: {:.2f}%".format(accuracy * 100))

# Print the results DataFrame
print(results_df.head())

model.summary()
```

```
plt.scatter(x=results_df.index, y=results_df['Probability of Nearest Failure'], c=results_df['Correct Prediction'], cmap='cool')
plt.axhline(y=0.5, color='black', linestyle='--')
plt.xlabel('Sample Index')
plt.ylabel('Probability of Nearest Failure')
plt.title('Predicted Probabilities vs Actual Failure Statuses')
plt.show()

# create x-axis values
x_values = range(200)

# plot actual vs predicted values
plt.figure(figsize=(8,4))
plt.plot(x_values, y_test[:200], marker='.', label="actual", color = 'black')
plt.plot(x_values, y_pred[:,0][:200], 'r', label="prediction")

# adjust plot layout
plt.tight_layout()
plt.ylabel('Parameter', size=15)
plt.xlabel('Time step', size=15)
plt.legend(fontsize=15)
plt.show()
```