MINISTRY OF EDUCATION AND SCIENCE OF THE REPUBLIC OF KAZAKHSTAN

СӘТБАЕВ УНИВЕРСИТЕТІ SATBAYEV UNIVERSITY

School of geology, petroleum and mining engineering

Department of Petroleum Engineering

Bukharbayeva A.N.

Identification of the most effective candidate well for hydraulic fracturing using machine learning

DIPLOMA PROJECT

5B070800 – Oil and Gas business

Almaty 2021

MINISTRY OF EDUCATION AND SCIENCE OF THE REPUBLIC OF KAZAKHSTAN

СӘТБАЕВ УНИВЕРСИТЕТІ SATBAYEV UNIVERSITY

School of geology, petroleum and mining engineering

Department of Petroleum Engineering

APPROVED FOR DEFENSE

Head of the Petroleum
Engineering Department
Dairov Zh.K., PG in
Petroleum Engineering

DIPLOMA PROJECT

Topic: « Identification of the most effective candidate well for hydraulic fracturing using machine learning»

5B070800 – Oil and Gas business

Performed by Bukharbayeva A.N.

Academic adviser
PG in Petroleum Engineering
Dairov Zh.K.

Almaty 2021

StrikePlagiarism.com SATBAYEV UNIVERSITY

Дата отчета 5/12/2021
Дата редактирования ---

Метаданные

Название
Identification of the most effective candidate well for hydraulic fracturing using machine learning

Автор	Научный руководитель
Бухарбаева Айдана	**Жасулан Дайров**

Подразделение
ИГНиГД

Список возможных попыток манипуляций с текстом

В этом разделе вы найдете информацию, касающуюся манипуляций в тексте, с целью изменить результаты проверки. Для того, кто оценивает работу на бумажном носителе или в электронном формате, манипуляции могут быть невидимы (может быть также целенаправленное вписывание ошибок). Следует оценить, являются ли изменения преднамеренными или нет.

Замена букв	ß	3
Интервалы	A→	0
Микропробелы	▯	21
Белые знаки	▨	0
Парафразы (SmartMarks)	a	32

Объем найденных подобий

Обратите внимание!Высокие значения коэффициентов не означают плагиат. Отчет должен быть проанализирован экспертом.

2.74% 2.74% КП1	**0.00%** 0.00% КП2	**0.43%** 0.43% КЦ
25 Длина фразы для коэффициента подобия 2	**14942** Количество слов	**96031** Количество символов

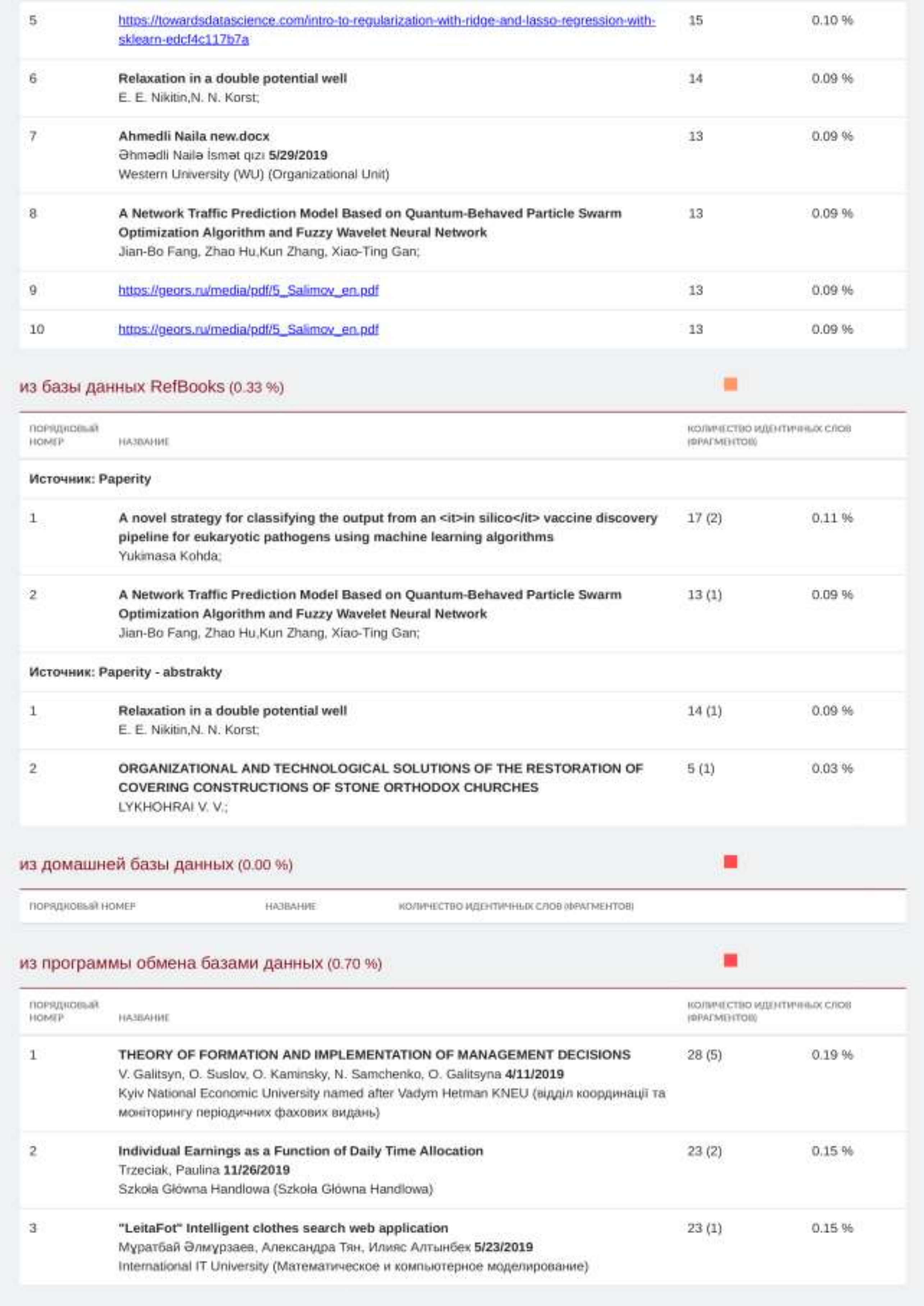
Подобия по списку источников

Просмотрите список и проанализируйте, в особенности, те фрагменты, которые превышают КП №2 (выделенные жирным шрифтом). Используйте ссылку «Обозначить фрагмент» и обратите внимание на то, являются ли выделенные фрагменты повторяющимися короткими фразами, разбросанными в документе (совпадающие сходства), многочисленными короткими фразами расположенные рядом друг с другом (парафразирование) или обширными фрагментами без указания источника ("криптоцитаты").

10 самых длинных фраз

Цвет текста

ПОРЯДКОВЫЙ НОМЕР	НАЗВАНИЕ И АДРЕС ИСТОЧНИКА URL (НАЗВАНИЕ БАЗЫ)	КОЛИЧЕСТВО ИДЕНТИЧНЫХ СЛОВ (ФРАГМЕНТОВ)	
1	http://e-journal.president.ac.id/presunivojs/index.php/JMEM/article/view/322	23	0.15 %
2	**"LeitaFot" Intelligent clothes search web application** Мұратбай Әлмұрзаев, Александра Тян, Илияс Алтынбек **5/23/2019** International IT University (Математическое и компьютерное моделирование)	23	0.15 %
3	https://mindrightdetroit.com/faq/readers-ask-reasons-why-fracking-is-bad.html	21	0.14 %
4	**Individual Earnings as a Function of Daily Time Allocation** Trzeciak, Paulina **11/26/2019** Szkoła Główna Handlowa (Szkoła Główna Handlowa)	16	0.11 %



5	https://towardsdatascience.com/intro-to-regularization-with-ridge-and-lasso-regression-with-sklearn-edcf4c117b7a	15	0.10 %
6	**Relaxation in a double potential well** E. E. Nikitin,N. N. Korst;	14	0.09 %
7	**Ahmedli Naila new.docx** Əhmədli Nailə İsmət qızı **5/29/2019** Western University (WU) (Organizational Unit)	13	0.09 %
8	**A Network Traffic Prediction Model Based on Quantum-Behaved Particle Swarm Optimization Algorithm and Fuzzy Wavelet Neural Network** Jian-Bo Fang, Zhao Hu,Kun Zhang, Xiao-Ting Gan;	13	0.09 %
9	https://geors.ru/media/pdf/5_Salimov_en.pdf	13	0.09 %
10	https://geors.ru/media/pdf/5_Salimov_en.pdf	13	0.09 %

из базы данных RefBooks (0.33 %) 🟧

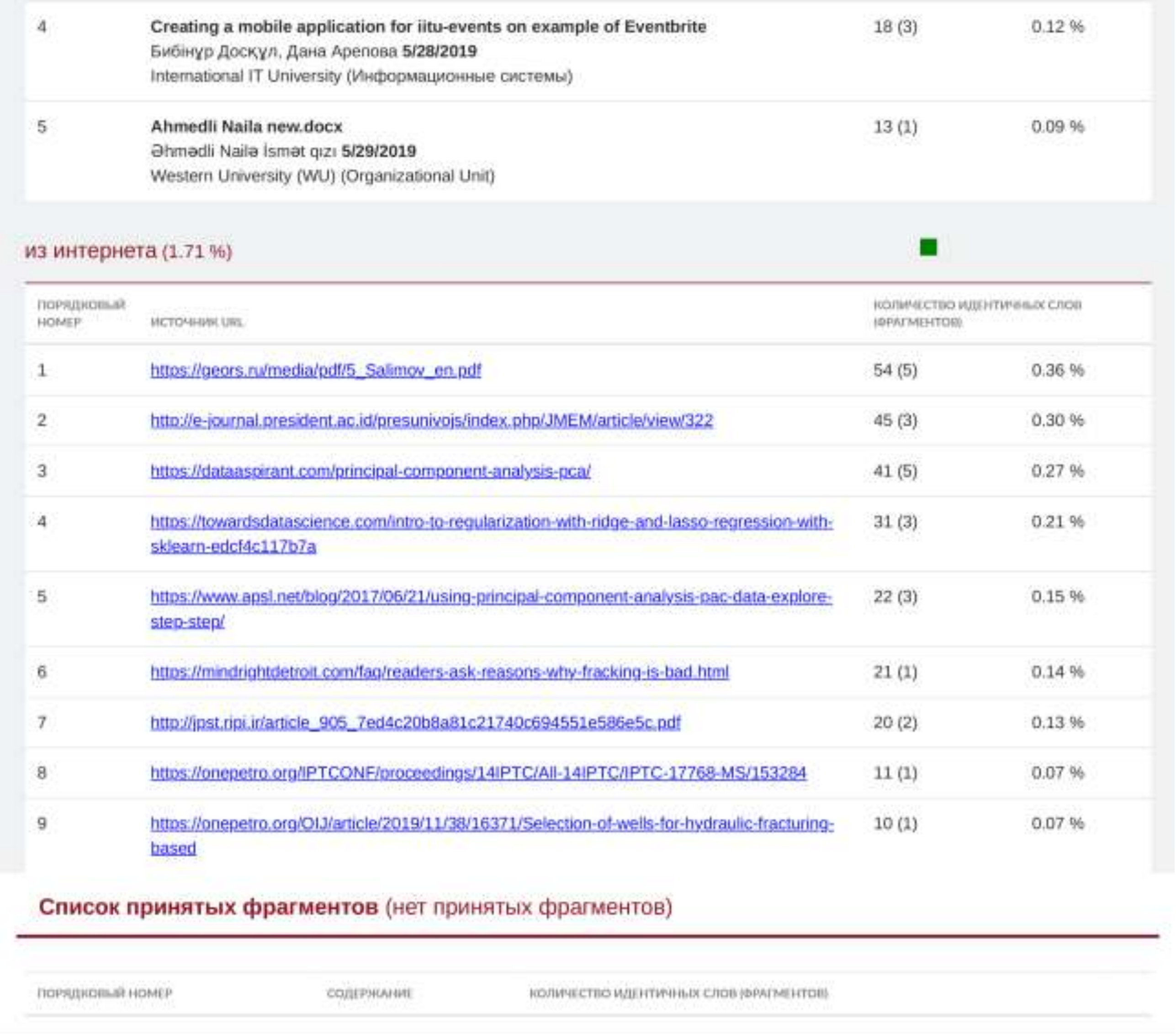
ПОРЯДКОВЫЙ НОМЕР	НАЗВАНИЕ	КОЛИЧЕСТВО ИДЕНТИЧНЫХ СЛОВ (ФРАГМЕНТОВ)	
Источник: Paperity			
1	**A novel strategy for classifying the output from an <it>in silico</it> vaccine discovery pipeline for eukaryotic pathogens using machine learning algorithms** Yukimasa Kohda;	17 (2)	0.11 %
2	**A Network Traffic Prediction Model Based on Quantum-Behaved Particle Swarm Optimization Algorithm and Fuzzy Wavelet Neural Network** Jian-Bo Fang, Zhao Hu,Kun Zhang, Xiao-Ting Gan;	13 (1)	0.09 %
Источник: Paperity - abstrakty			
1	**Relaxation in a double potential well** E. E. Nikitin,N. N. Korst;	14 (1)	0.09 %
2	**ORGANIZATIONAL AND TECHNOLOGICAL SOLUTIONS OF THE RESTORATION OF COVERING CONSTRUCTIONS OF STONE ORTHODOX CHURCHES** LYKHOHRAI V. V.;	5 (1)	0.03 %

из домашней базы данных (0.00 %) 🟥

ПОРЯДКОВЫЙ НОМЕР	НАЗВАНИЕ	КОЛИЧЕСТВО ИДЕНТИЧНЫХ СЛОВ (ФРАГМЕНТОВ)

из программы обмена базами данных (0.70 %) 🟥

ПОРЯДКОВЫЙ НОМЕР	НАЗВАНИЕ	КОЛИЧЕСТВО ИДЕНТИЧНЫХ СЛОВ (ФРАГМЕНТОВ)	
1	**THEORY OF FORMATION AND IMPLEMENTATION OF MANAGEMENT DECISIONS** V. Galitsyn, O. Suslov, O. Kaminsky, N. Samchenko, O. Galitsyna **4/11/2019** Kyiv National Economic University named after Vadym Hetman KNEU (відділ координації та моніторингу періодичних фахових видань)	28 (5)	0.19 %
2	**Individual Earnings as a Function of Daily Time Allocation** Trzeciak, Paulina **11/26/2019** Szkoła Główna Handlowa (Szkoła Główna Handlowa)	23 (2)	0.15 %
3	**"LeitaFot" Intelligent clothes search web application** Мұратбай Әлмұрзаев, Александра Тян, Илияс Алтынбек **5/23/2019** International IT University (Математическое и компьютерное моделирование)	23 (1)	0.15 %



4	**Creating a mobile application for iitu-events on example of Eventbrite** Бибінұр Досқұл, Дана Арепова **5/28/2019** International IT University (Информационные системы)	18 (3)	0.12 %
5	**Ahmedli Naila new.docx** Əhmədli Nailə İsmət qızı **5/29/2019** Western University (WU) (Organizational Unit)	13 (1)	0.09 %

из интернета (1.71 %) 🟩

ПОРЯДКОВЫЙ НОМЕР	ИСТОЧНИК URL	КОЛИЧЕСТВО ИДЕНТИЧНЫХ СЛОВ (ФРАГМЕНТОВ)	
1	https://geors.ru/media/pdf/5_Salimov_en.pdf	54 (5)	0.36 %
2	http://e-journal.president.ac.id/presunivojs/index.php/JMEM/article/view/322	45 (3)	0.30 %
3	https://dataaspirant.com/principal-component-analysis-pca/	41 (5)	0.27 %
4	https://towardsdatascience.com/intro-to-regularization-with-ridge-and-lasso-regression-with-sklearn-edcf4c117b7a	31 (3)	0.21 %
5	https://www.apsl.net/blog/2017/06/21/using-principal-component-analysis-pac-data-explore-step-step/	22 (3)	0.15 %
6	https://mindrightdetroit.com/faq/readers-ask-reasons-why-fracking-is-bad.html	21 (1)	0.14 %
7	http://jpst.ripi.ir/article_905_7ed4c20b8a81c21740c694551e586e5c.pdf	20 (2)	0.13 %
8	https://onepetro.org/IPTCONF/proceedings/14IPTC/All-14IPTC/IPTC-17768-MS/153284	11 (1)	0.07 %
9	https://onepetro.org/OIJ/article/2019/11/38/16371/Selection-of-wells-for-hydraulic-fracturing-based	10 (1)	0.07 %

Список принятых фрагментов (нет принятых фрагментов)

ПОРЯДКОВЫЙ НОМЕР	СОДЕРЖАНИЕ	КОЛИЧЕСТВО ИДЕНТИЧНЫХ СЛОВ (ФРАГМЕНТОВ)

MINISTRY OF EDUCATION AND SCIENCE OF THE REPUBLIC OF KAZAKHSTAN

СӘТБАЕВ УНИВЕРСИТЕТІ SATBAYEV UNIVERSITY

School of geology, petroleum and mining engineering
Department of Petroleum Engineering

CONFIRM

Head of the Petroleum
Engineering Department
Dairov Zh.K., PG in
Petroleum Engineering

TASK

for completing the diploma project

For student: Bukharbayeva A.N.

Topic: «Identification of the most effective candidate well for hydraulic fracturing using machine learning»

Approved by the order of university rector №2131-b from “24” November 2020

Deadline for completion the work “18” May 2021

Initial data for the diploma project: data of the information system "Geographically distributed data bank" (TBD), interpretation results of well logging and well testing of the Uzen field.

Summary of the diploma project:

a) *Identification of a list of factors that affect the selection process of candidates for hydraulic fracturing*
b) *Establishment a database of training wells*
c) *Determining the most effective algorithm for the data sample under consideration*
d) *Forecasting production on potential candidates*
e) *Economic calculations*

List of graphic material: *presented 16 slides of presentation of the diploma project*

Recommended main literature: (Michael Economides, 2004)

THE SCHEDULE

For the diploma project preparation

Name of sections, list of issues being developed	Submission deadlines to the Academic adviser	Notes
Justification of expediency. Setting goals and objectives, an explanation of the relevance.	08.02.2021	Task completed
Evaluation of the geology, history of exploration, production, analysis of HF of the Uzen field.	15.02.2021	Task completed
Detailed review of the research material, comparative analysis.	18.02.2021	Task completed
Determining the list of parameters that affect on the selection of candidate wells for hydraulic fracturing.	01.03.2021	Task completed
Creating a database: review and identify 100 wells where hydraulic fracturing was performed (with positive and negative results).	11.03.2021	Task completed
Technical description of the project: development and use of various models of machine learning algorithms.	22.03.2021	Task completed
Applicability assessment: production forecasting	15.04.2021	Task completed
Economic justification	30.04.2021	Task completed

SIGNATURES

Of consultants and standard controller for the completed diploma work, indicating the relevant sections of the work (project).

The section titles	Consultant name (academic degree, title)	Date	Signature
Theoretical part	PG, Dairov Zh.K.	08.02.2021	
Technical description of the project	PG, Dairov Zh.K.	22.03.2021	
Economic model	PG, Dairov Zh.K.	15.04.2021	
Normcontrol	PG, Dairov Zh.K.	10.05.2021	

Academic Adviser PG, Dairov Zh.K.

The task was completed by student: Bukharbayeva A.N.

Date **"18" May 2021**

ANNOTATION

This project presents a methodology for identifying the most effective candidates for hydraulic fracturing based on well performance forecasting. The methodology is developed with the use of massive data mining tools.

Hydraulic fracturing provides a significant impact on the recovery of production, and therefore requires a huge investment due to operating costs. Applying the correct design to a non-efficient candidate ensures that this method of production intensification is not appropriate. Due to the fact that the selection of candidates is the first stage of the implementation of the hydraulic fracturing process, the use of the developed technology will significantly reduce the economic risks.

The third stage of development of the Uzen field is characterized by the need for hydraulic fracturing to increase oil recovery. Every year, more than 100 hydraulic fracturing operations are carried out at this field. The huge volume of the current fund of wells complicates the process of selecting candidates. This problem requires a comprehensive approach to solving and determines the relevance of this study.

To date, the selection of a candidate well for hydraulic fracturing is a significantly time-consuming process. The time factor is due to the requirement to consider a huge amount of information about the history of production, the geological and technical measures carried out for the specified well, the method of development of the field as a whole, etc. The developed technology not only provides accelerated analysis of significant amounts of information, but also represents the qualitative predictive power of data mining tools for geologically complex reservoirs.

АННОТАЦИЯ

В данном проекте представлена методика выявления наиболее эффективных кандидатов для проведения гидравлического разрыва пласта на основе прогнозирования производительности скважин. Методология разработана с применением инструментов интеллектуального анализа массивных данных.

Гидроразрыв пласта обеспечивает значительное влияние на восстановление производства, и соответственно требует огромных капиталовложений, обусловленных операционными затратами. Применение правильного дизайна к не эффективному кандидату гарантирует нецелесообразность применения данной метода интенсификации добычи. В связи с тем, что подбор кандидатов является первым этапом осуществления процесса ГРП, применение разработанной технологии, значительно снизит экономический риски.

Третья стадия разработки месторождения Узень характеризуется необходимостью проведения ГРП для увеличения нефтеотдачи. Ежегодно на данном месторождении проводится более 100 операции ГРП. Огромный объем текущего фонда скважин усложняет процесс подбора кандидатов. Данная проблема требует комплексного подхода к решению и обуславливает актуальность данного исследования.

На сегодняшний день подбор скважины-кандидата для проведения гидравлического разрыва пласта является значительно время затратным процессом. Фактор времени обусловлен требованием рассмотрения огромного массива информации об истории добычи, проведенных геолого-технических мероприятий указанной скважины, способа разработки месторождения в целом и др. Разработанная технология не только обеспечивает ускоренный анализ значительных объемов информации, но и представляет качественную прогностическую силу инструментов интеллектуального анализа данных для геологически сложных коллекторов.

АҢДАТПА

Бұл жобада ұңғымалардың өнімін болжау барысында, қабат гидро жарылысын жүргізу үшін ең тиімді үміткер ұңғыманы анықтау әдістемесі ұсынылған. Жалпы әдістеме көптеген мәліметтер мен деректерді іздеу, талдау барысында жасалды .

Қабат гидро жарылысы өндірісті қалпына келтіруге айтарлықтай әсер етеді, және сәйкесінше операциялық шығындарға байланысты үлкен көлемдегі инвестицияларды қажет етеді. Тиімді емес үміткерге дұрыс дизайнды қолдану, мұнай өндіруді арттыруға, осы әдісті қолданудың қажет еместігіне кепілдік береді. Үміткер ұңғымаларды іріктеу, қабат гидро жарылысы процесін жүзеге асырудың бірінші кезеңі болғандықтан, дамыған технологияны қолдану экономикалық тәуекелдерді едәуір төмендетеді.

Өзен кен орнын игерудің үшінші кезеңі, мұнай өндіруді арттыру үшін қабат гидро жарылысын жүргізу қажеттілігімен сипатталады. Жыл сайын Өзен кен орнында 100-ден астам қабат гидро жарылысы операциялары жүргізіледі. Қазіргі уақыттағы ұңғымалар қорының көптігі, ұңғыманы таңдау процесін қиындатады. Бұл мәселені шешу жан-жақты көзқарасты қажет етеді, сонымен қатар зерттеу жұмысының өзектілігі болып отыр.

Бүгінгі таңда қабат гидро жарылысын жүргізу үшін, үміткер ұңғыманы таңдау айтарлықтай уақытты қажет ететін процесс болып табылады. Уақыт факторы көрсетілген ұңғыманың геологиялық – технологиялық жұмыстары жүргізілген өндіру тарихы, кен орнын тұтастай игеру әдісі, және тағы да басқа ақпараттың үлкен көлемін қарастыру талабымен анықталады. Дамыған технология - ақпараттың едәуір көлемін жедел талдауды қамтамасыз етіп қана қоймай, сонымен қатар геологиялық күрделі коллекторлар үшін, деректерді іздеу құралдарының жоғары сапалы болжамды түрін ұсынады.

CONTENT

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

A distinctive feature of the oil industry of Kazakhstan is the presence of a significant part of the developed fields at a late stage of development. Under conditions of depletion of reserves, these fields are characterized by a drop in production and high water cut. Therefore, the problem of improving efficiency requires special attention, since the question of economic profitability remains open. Currently, hydraulic fracturing is one of the most effective methods of intensifying production. And the correct selection of an effective candidate well, in turn, determines the success of hydraulic fracturing.

The practical application of this work is of increased importance directly at the Uzen field, because it is not feasible to develop at this field without using methods and technologies to increase the intensification of oil production, namely hydraulic fracturing. In addition to the above, Uzen is the largest, multi-layer field, where more than a hundred hydraulic fracturing operations are carried out annually. Since the fund of existing wells is large, effective options are often overlooked when selecting a candidate for hydraulic fracturing, and the company does not receive the possible profit. The use of machine learning allows to significantly speed up the analysis of information. The scientific novelty of the study lies in the simultaneous use of modern approaches of machine learning and analysis of influencing parameters in the task of optimizing the process of selecting a well for conducting a hydraulic reservoir. As a result of the study, various machine learning algorithms were considered and described to identify the best candidate for hydraulic fracturing. The factors that mostly influence the process of well selection and their limitations for the use of machine learning methods are determined.

The purpose of this research is to develop an algorithm for identifying the best hydraulic fracturing candidate well based on a set list of parameters.

In accordance with the purpose of the study, the following tasks are set in the work:

- Research and establishment of a list of criteria for the selection of candidate wells for the Uzen field;
- Collecting data on wells where the hydraulic fracturing operation was performed;
- Conducting a quality check of the data representing the subject of the study;
- Development and application of various machine learning algorithms to identify the best candidate well;
- Conducting a comparative analysis and selecting the most successful model;
- Predicting production after hydraulic fracturing using the selected algorithm.

The object of the research is a sample of current production wells, where the hydraulic fracturing process was carried out. The sample size of the study is one hundred wells. The sample was carried out among wells 13, 14 horizons.

MAIN PART

1. THEORETICAL PART

1.1Tectonics and lithological-stratigraphic characteristics of the Uzen field

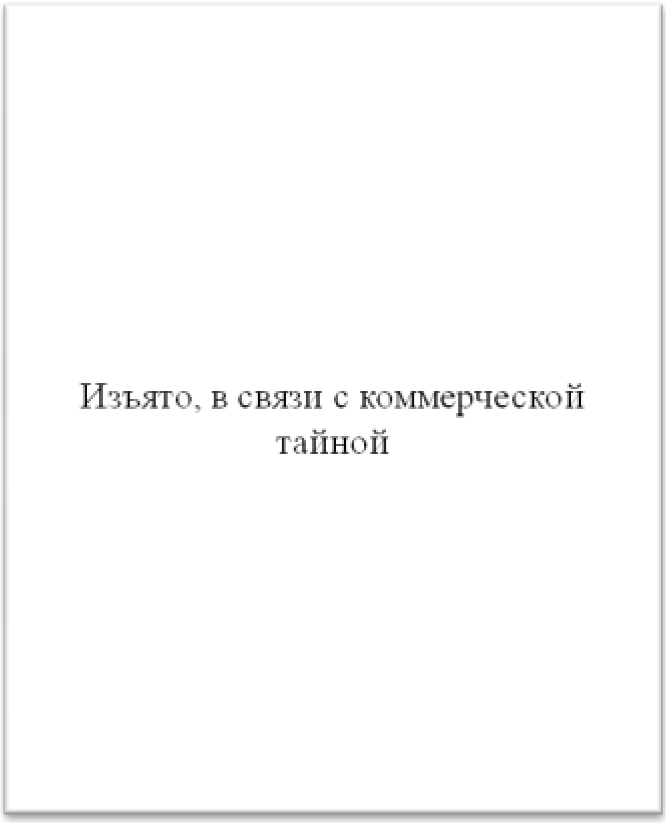
The Uzen deposit is located in the steppe part of the Southern Mangyshlak and is administratively part of the Karakiyansky district of the Mangystau region of the Republic of Kazakhstan.

Tectonically, the Uzen uplift is associated with the Zhetybai-Uzen tectonic stage, which complicates the northern side of the South Mangyshlak trough (Figure1). The region under consideration is part of the Turan Plate, which is part of the Central Eurasian young Epigercin platform. In the section here, three structural floors are distinguished, separated from each other by regional stratigraphic and angular inconsistencies.

The Zhetybai-Uzen tectonic stage, being a second-order structural element, is confined to the northern side of the South Mangyshlak trough and stretches from north-west to south-east for 200 km with a stage width of about 40 km. From the north, the stage is bounded by a regional fault that complicates the southern wing of the Beke-Bashkuduk rampart, in the west it borders on the Segendyk depression and the Karagiinsky saddle, and in the east-with the Kokumbai stage. The stage is separated from the Zhazgurlinsky depression in the south by a deep fault of the sublatitudinal strike, reflected in the platform cover by a flexure-like inflection (Figure 1).

Along the sedimentary cover within the Zhetybai-Uzen stage, three anticline lines are traced, oriented along the strike of the stage. From north to south, the most elevated Uzen-Karamandybas, then Zhetybai, and the most submerged Tenge-Tasbulat anticline lines are distinguished.

The largest local structure of the Zhetybai-Uzen stage is the Uzen uplift, which is a gentle anticlinal fold, the axis of which extends from east-southeast to west-northwest.



Изъято, в связи с коммерческой тайной

Figure 1-Tectonic map of Mangyshlak: I-Mangyshlak dislocation system; II-South Mangyshlak trough; IIA-Zhetybai-Uzen stage; IIB-Kokumbai tectonic Stage;

According to the roof 13 of the productive horizon, which is stratigraphically related to the Kellovian stage of the Upper Jurassic, the size of the Uzen fold is 34.5 x 10.0 km, and the elevation amplitude is about 300 m. The morphology of the fold is characterized by the asymmetry of the wings and periclinals. The northern wing is flat (the angles of incidence of rocks are 30), and the southern wing is steeper with angles of incidence of 5-60. The periclines of the structure are of different sizes: the eastern one is shorter than the western one and, accordingly, the fold axis dips in the eastern direction more sharply than in the western one. Within the more gentle northern wing of the fold and its western pericline, the sinking of rocks occurs unevenly with the formation of protruding areas. The shape of the fold and its spatial position coincide in different stratigraphic horizons of the Jurassic and Cretaceous. With depth, the amplitude of the rise and the angles of fall of the rocks on the wings increase, which is explained by the inherited nature of its development.

The structure is complicated by six domes, most clearly traced in the lower productive horizons: the Main Vault, the Humurun, Northwestern and Aksai, Parsumurun and East Parsumurun domes.

Deep drilling at the Uzen field uncovered a sedimentary complex with a thickness of 4500 m, in the structure of which rocks of Triassic, Jurassic, Cretaceous, Paleogene, Neogene and Quaternary ages take part. Within the Uzen structure, deep drilling uncovered Lower Triassic deposits with a thickness of 698 to 2250 m, represented by the Indian and Olenekian tiers.

Jurassic sediments, which are associated with the industrial oil and gas content of the Uzen field, transgressingly lie on the eroded surface of the Triassic rock complex. The Lower, middle and upper divisions of the Jurassic system are distinguished by the results of the study of fauna, flora and data of spore-pollen analysis. The Jurassic sediments are clearly divided into two complexes according to their lithological composition: the terrigenous complex of rocks of the Lower and Middle Jurassic and the carbonate complex of the Upper Jurassic. The undifferentiated deposits of the Lower Jurassic are represented by the interbedding of sandstones, siltstones, argellite-like black carbonaceous clays with organic plant remains and coal inclusions.

Productive deposits of the 13-18 horizons of the Uzen deposit are represented by an uneven alternation of terrigenous rocks – sandstones, siltstones, clays and transitional lithological differences between them. Among them there are thin layers of limestones, marls, siderite, coals, and accumulations of charred plant detritus. In the calcareous differences of siltstones and clays, cores, fragments and impressions of bivalve shells are often found, sometimes small aggregations of pyrite.

1.2 Structure of oil and gas deposits of the considered horizons

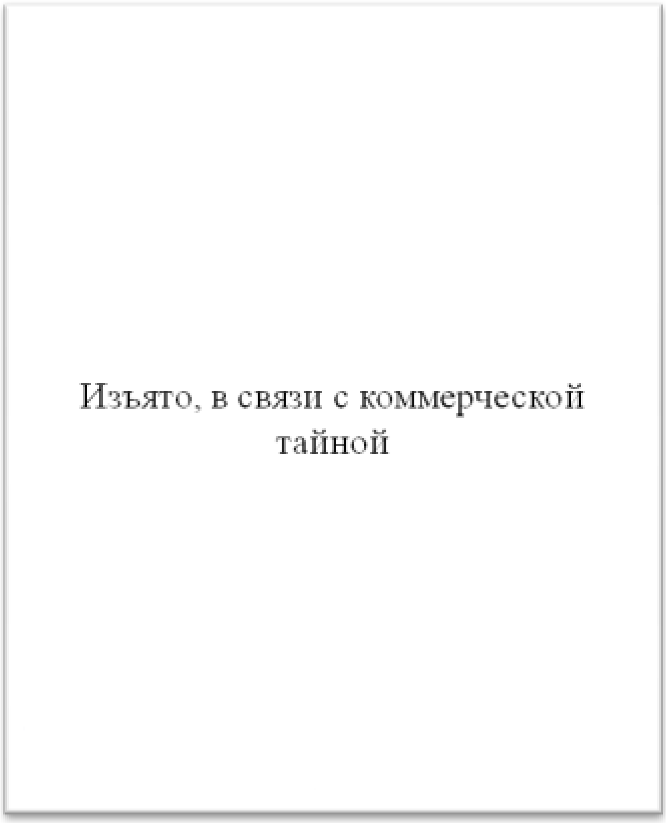
A characteristic feature of the productive strata of the 13-18 horizons is a high heterogeneity, which is expressed in the complex nature of the distribution of reservoir layers over the area and section of the field and the significant variability of their filtration and reservoir properties.

Along with a fairly confident correlation not only of the horizons, but also of individual bundles within the entire Uzen structure, there are significant differences in the number and volume of deposits installed on different domes, which is associated with the complex nature of the distribution of reservoir layers within the bundles and horizons.

The porosity of productive reservoir rocks varies from 14.0% (lower limit) to 41.7% (horizon 13). The tendency to decrease porosity from top to bottom along the section of the productive strata, despite some deviations due to the lithological heterogeneity of layers and bundles, is maintained quite clearly. The permeability of productive reservoirs of all bundles in the composition of 13-18 horizons varies in extremely wide ranges – from 0.001 to 7.301 mD (14 horizon).

Horizon 13

Within the horizon, the thickness of which varies from 40 to 56 m, 12 sand-siltstone layers are traced, united according to the accepted scheme of dividing the productive section into 5 bundles (Figure 2). Large reservoir thicknesses are not characteristic of this horizon and, in addition to the sand lens that can be traced through the listed blocks, they occur over the area of the deposit in small areas, without affecting the general idea of the structure of the horizon as the most heterogeneous in the Jurassic productive section of the Uzen deposit.



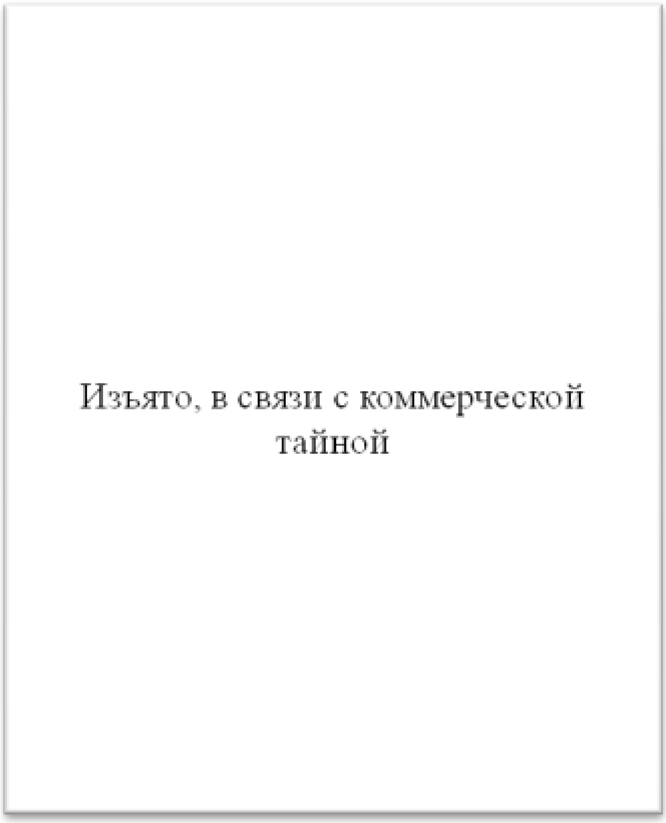
Изъято, в связи с коммерческой тайной

Figure 2-Geological and statistical cross-section of the horizon 13

The water-oil zones on each deposit of the thirteenth horizon, having a small width, are characterized by the development of reservoirs of small thickness. Therefore, a limited number of production wells were drilled in this zone, usually injection wells, and only a few wells were tested before the start of water injection to determine the nature of reservoir saturation.

Horizon 14

The horizon is separated from the overlying horizon 13 by a well-maintained clay section in thickness and area. The thickness of the horizon varies from 60 m in the east to 80 m in the west of the structure. However, the decrease in thickness from west to east does not affect the structure of the horizon as a whole, within which 15 sand-siltstone layers were identified as a result of detailed reservoir correlation (Figure 3).



Изъято, в связи с коммерческой тайной

Figure 3-Geological and statistical cross-section of the horizon 14

1.3 A brief overview of the application of machine learning in the process of selecting candidates for hydraulic fracturing.

Currently, the application areas of machine learning are expanding every day. This branch of artificial intelligence is widespread not only in industry, trade, various sectors of the economy, but also in everyday life. Machine learning is an extensive sub-division of artificial intelligence, the methodology of which is that the computer does not just use a pre-written algorithm, but learns how to solve a set problem. Machine learning quickly automates the process of creating an analytical model, and also allows computers to independently adapt to new scenarios.

The oil and gas industry is the largest source for big data, therefore, the use of machine learning algorithms in this area, first of all, optimizes the economic component of the company. In the oil industry, the range of algorithms used is wide: the choice of the field development option, the calculation of reserves, the calculation of the expected flow rate after the intensification of production, the assessment of filtration and reservoir properties, and more. The hydraulic fracturing process is not an exception to the above list.

Hydraulic fracturing is one of the most common and effective methods for increasing oil production of reservoirs. The increase in hydraulic fracturing operations in recent years has led to the formation of huge amounts of data. As a result, the ability to implement machine learning in the hydraulic fracturing process in order to improve results has also increased. The process of selecting a candidate well is the first and result-determining step in the implementation of the hydraulic fracturing process. The use of artificial intelligence algorithms in selecting the desired target well reveals the potential of research, revealing the relationship between the impact factors. In general, there is no traditional, generally accepted methodology for applying machine learning to the process of selecting wells for hydraulic fracturing. The development of a methodology for the rapid selection of wells for hydraulic fracturing based on machine learning is presented in the paper (Akhmetov A., 2018). The authors developed a machine learning model based on neural networks to estimate the average annual level of oil production after hydraulic fracturing treatment. Based on the selected model, forecasts are made for other wells that are potential for hydraulic fracturing applications. There is a growing body of research around the world focused on applying big data analysis to the problem of hydraulic fracturing optimization. The use of advanced techniques, such as artificial neural networks, significantly reduces the uncertainty in the selection of candidate wells (Aryanto Agus, 2017). Non-linearity is the main advantage of neural networks, as a nonlinear relationship between the predicted and actual process parameters is established. In the above work, the relationship between the input and output data set is determined to reveal the optimal distribution of the membership function, which allows for more efficient prediction of candidate selection and fracture optimization. The paper (Alimkhanov R., 2014) presents a methodology for selecting wells for hydraulic fracturing operations using Data Mining tools. The impact of various geological and field conditions on the efficiency of hydraulic treatment of the reservoir was also assessed. Classification models have been developed to divide potential candidates into groups: effective and ineffective. In addition to the above, the authors have proposed regression models for predicting the flow rate and water cut after hydraulic fracturing. A recent study (Vanina A. S., 2020), containing a sample of 5,000 wells, presents a detailed process for optimizing hydraulic fracturing design. The input parameters are divided into reservoir, well, and design parameters of the hydraulic fracturing design. The estimated parameter is the three-month cumulative oil production. As a machine learning tool, several boosting algorithms are used and compared. Testing of the developed methodology for pilot wells shows the effectiveness of the developed approach. Traditional methods of well selection for hydraulic fracturing do not take into account all the non-linearity of the process. The results of research and the experience of recent years have shown us the possibility of implementing machine learning in the process of analyzing and predicting hydraulic fracturing operations.

1.4 Parameters and their influence on candidate-well selection for hydraulic fracturing.

The selection of a suitable candidate for hydraulic fracturing determines the ultimate success of the entire process. This process is the first stage in hydraulic fracturing operations (Figure 4). The use of machine learning to identify the impact of a number of parameters on the volume of production after a fracking operation is a powerful tool, as it leads to an integrated approach.

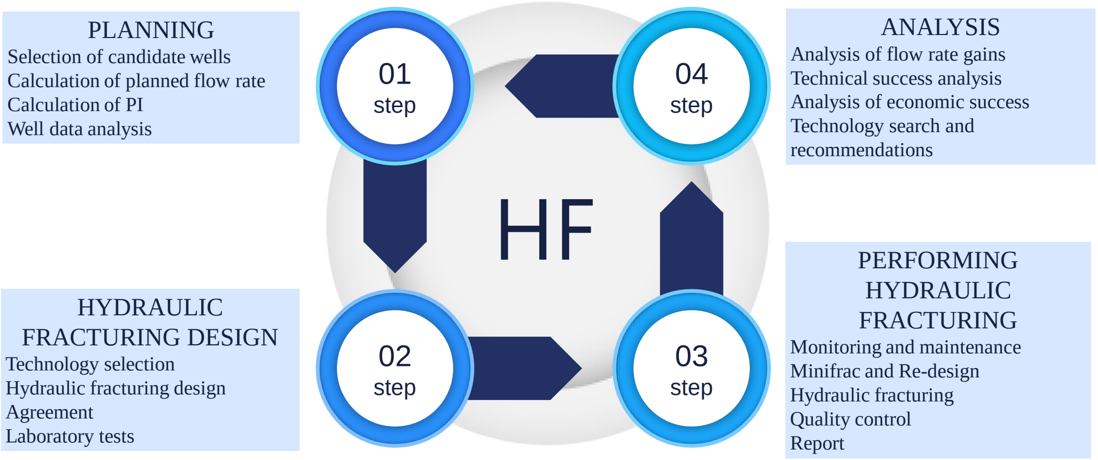

PLANNING
Selection of candidate wells
Calculation of planned flow rate
Calculation of PI
Well data analysis
01
step
04
step
ANALYSIS
Analysis of flow rate gains
Technical success analysis
Analysis of economic success
Technology search and recommendations
HF
HYDRAULIC FRACTURING DESIGN
Technology selection
Hydraulic fracturing design
Agreement
Laboratory tests
02
step
03
step
PERFORMING HYDRAULIC FRACTURING
Monitoring and maintenance
Minifrac and Re-design
Hydraulic fracturing
Quality control
Report

Figure 4- Block diagram showing the stages of hydraulic fracturing operations

In practice, a table with a list of parameters and their boundary values is used to select a candidate. For example, the water content of the candidate well should not exceed 50%, or the effective oil-saturated thickness should not be less than 0.8 m (Салимов О.В., 2017). Using the threshold methodology, a large number of effective candidates are overlooked, due to the fact that they are not suitable only for a certain parameter. The presence of errors in the methodology of threshold values: 1-a well that does not meet the criteria, but the hydraulic fracturing on it will be effective; 2-a well that meets the criteria, but the hydraulic fracturing on it will be ineffective, because the use of other approaches in the selection of candidate wells (Салимов О.В., 2017). The criteria that are taken into account when selecting candidates for hydraulic fracturing in this study will be discussed below. In the future, the relationship between the influences of the selected criteria on the flow rate after hydraulic fracturing is revealed. This procedure is applied to wells 13, 14 horizons of the Uzen field.

Parameters that affect the selection of a candidate well:

Net Pay Thickness. According to Darcy's law, as the effective thickness of the reservoir increases, the oil flow rate increases proportionally. Because the main purpose of production intensification is to increase profits by increasing production volumes, the economy is the driving force in deciding which well is most suitable for carrying out the inflow intensification. Since the hydraulic formation operation is relatively resource-intensive, the best candidate wells must have significant volumes of hydrocarbons. Following this point of view, large values of the oil-saturated thickness are preferred.

Reservoir permeability. This parameter refers to the petrophysical properties of the intervals. In the field under consideration, wells with moderate and relatively high permeability have good production indicators, therefore, do not belong to potential candidates. When conducting hydraulic fracturing, as a rule, wells with low permeability are considered as an effective candidate.

Reservoir pressure. Effective candidates for hydraulic fracturing are wells with a smaller drop in reservoir pressure relative to the initial one. Low reservoir pressure at wells during hydraulic fracturing is the reason for the failure to achieve the planned flow rate. In this study, the ratio of the current reservoir pressure to the initial pressure is taken as a parameter that estimates the reservoir pressure. This is due to the fact that the sample was made from two horizons with different initial reservoir pressures. For non-perforated objects, you must specify the primary reservoir pressure or the results in neighboring wells.

Water Cut. A distinctive feature of the Uzen deposit is the high degree of water cut of the layers. In this regard, during the hydraulic fracturing operation, candidates with critical water cut values are not neglected. However, lower values of the degree of water cut are preferable, since the share of oil in the extracted liquid will be higher.

Oil rate. Wells with a low oil flow rate prior to fracking are also suitable candidates. Because as a result of the intensification under consideration, the oil flow rate increases, therefore, the value of the productivity index increases, showing the potential and ability of the well to produce products.

Distance to the nearest production well. When conducting hydraulic fracturing in a field drilled on a dense grid, working oil wells can interact. The reason is the interference of wells, which leads to the economic inexpediency of the hydraulic fracturing process. Wells with long distances to neighboring wells are preferred. The minimum and maximum distances to the producing well are 94 and 881m, respectively.

Distance to the nearest injection well. The breakthrough of the injected water is also the reason for the failure to achieve the predicted flow rate. The optimal fracture length ensures the efficiency of the hydraulic fracturing process, but it is necessary to take into account the radius of the well drainage zone and the proximity of the injection wells. According to the data sample, the minimum and maximum distance to the nearest injection well is 152 and 1194 m, respectively.

New planned fracturing interval. The considered horizons of the Uzen deposit are characterized by an inhomogeneous structure of productive layers. Oil-saturated layers are an alternation of permeable sandy and impermeable clay layers, which indicates a high degree of dissection of the layers. Hydraulic fracturing under these conditions increases oil recovery, due to the involvement in the development of oil reserves in heterogeneous and dismembered reservoirs. The possibility of additional perforation is highlighted as a parameter taken into account when selecting a candidate for hydraulic fracturing.

The success of hydraulic fracturing is primarily determined by the selection of a suitable candidate well. The most important parameters for selecting a candidate for hydraulic fracturing, considered in this study, are presented in table 1. As a source of information, various sources were used to create a database of potential candidates for the specified parameters. To determine the distance to the nearest production and injection wells, we have mastered the work on the NGT Smart computer program. The information system "Geographically distributed data bank" (TBD) is a resource of information on the remaining criteria.

Table 1- Criteria affecting the selection of candidates for hydraulic fracturing

Parameter	Units	Effect	Source of information
Additional perforation	Qualitative	Positive	TBD-Perforation intervals
Net Pay Thickness	[m]	Positive	TBD-Perforation intervals
Reservoir pressure	[atm/atm]	Positive	TBD-Well test
Permeability	[mD]	Negative	TBD-Well test
Oil rate	[t/day]	Negative	TBD-Operational parameters
Water Cut	[%]	Negative	TBD-Operational parameters
Distance to IW	[m]	Positive	NGT Smart
Distance to PW	[m]	Positive	NGT Smart

The main purpose of the TBD is to create a single field of geological, geophysical and field information for decision-making. The TBD system operates in real time. For example, the indications of water cut and oil flow rate before hydraulic fracturing are determined from operational data on liquid production. The values of permeability and reservoir pressure are contained in hydrodynamic studies (the method of steady-state sampling, buildup, etc.). Data on additional perforation and effective intervals are determined from the reporting documentation.

1.5 Summary of Chapter 1

This chapter discusses the theoretical aspects of this study. Uzen is one of the largest deposits on the territory of our country. The geological structure and heterogeneity of the reservoirs, the high water content of the reservoirs cause a complicated process of oil production. The use of various methods of inflow intensification in this field is not just a frequent practice, but a necessity. The huge fund of existing wells and various reasons for not achieving the planned indicators during hydraulic fracturing determine the relevance of research in this area.

The current economic situation dictates new realities and conditions for oil production. The use of machine learning in the oil industry is one of the priorities, as it guarantees an increase in the productivity and efficiency of the technologies used.

The introduction of machine learning algorithms in the process of selecting wells for hydraulic fracturing operations provides accelerated analysis of a large amount of data and the ability to achieve a high level of economic efficiency of the hydraulic fracturing process.

The selection of candidates requires the integration of various sectors, such as geology, reservoir engineering, petrophysics, manufacturing, geomechanics, and stimulation engineering. According to the practice of hydraulic fracturing, there is a contradiction in the methodology used for selecting candidate wells. Each method requires a careful approach, characterized by its own disadvantages and advantages. The Uzen deposits are multi– layered and heterogeneous, so 13, 14 horizons were chosen as the object of research. Because the qualitative results are the result of a detailed approach.

The first chapter also considers the significance of the parameters that are taken into account in this study as factors affecting the flow rate after hydraulic fracturing. The selection of an effective candidate for hydraulic fracturing ensures that the economic aspects of the operation are met and that it is economically viable.

2. TECHNICAL DESCRIPTION OF THE PROJECT

The presented research is aimed at identifying an effective candidate for conducting the hydraulic fracturing process using machine learning.

The step-by-step scheme of the study is shown in Figure 5. This chapter will cover the stages of data collection, data analysis, and algorithm modeling.

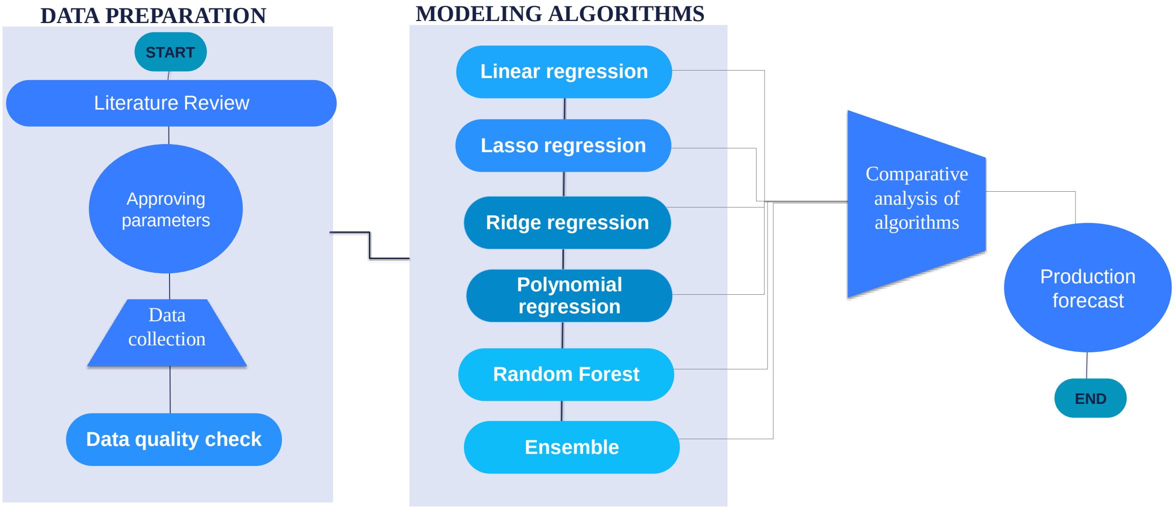


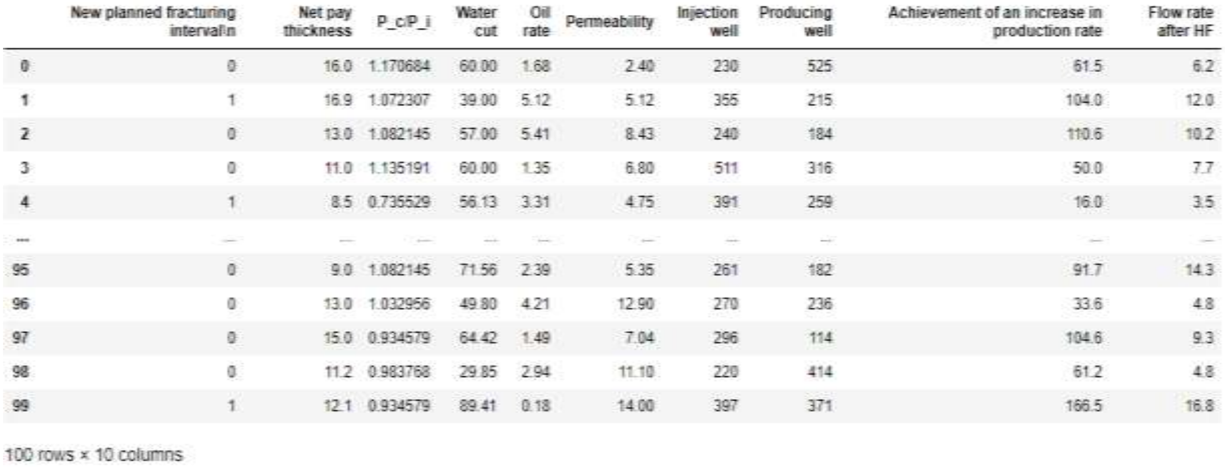
Figure 5- Flowchart of research

The code for the developed models is written in the Jupyter notebook interactive computing environment based on Python. Python is one of the most common high-level object-oriented programming languages. Its popularity is due to its simple syntax, portability, and the most commonly used functions from the standard library. Classical numerical solution algorithms are implemented in packages used in scientific computing-numpy, scipy, matplotlib, and sympy. The possibilities of implementing two-dimensional and three-dimensional data are implemented using the matplotlib package. The basis of the NumPy and SciPy packages is numerical calculations, but also symbolic calculations. In the present study, the Scikit-learn package is the most powerful and widely used, as it provides a variety of algorithms. With the use of this package, preprocessing of data, changing the dimensions of variables, and cluster analysis were carried out.

2.1 Data: collection, preparation, quality check, analysis

Data collection. One of the most important tasks of this research is the formation of a database of qualitative data for machine learning algorithms. The use of low-quality data leads to problems associated with data preprocessing and overestimated performance error.

The generated database is presented in the form of processed and structured information in tabular form. The objects are the rows of this table - the wells on which hydraulic fracturing was carried out. And the columns represent the parameters under consideration. Therefore, the average oil flow rate after the application of geological and technical measures is a dependent variable on predictors (influencing parameters). The number of rows corresponds to the number of wells (100), the number of columns-10.



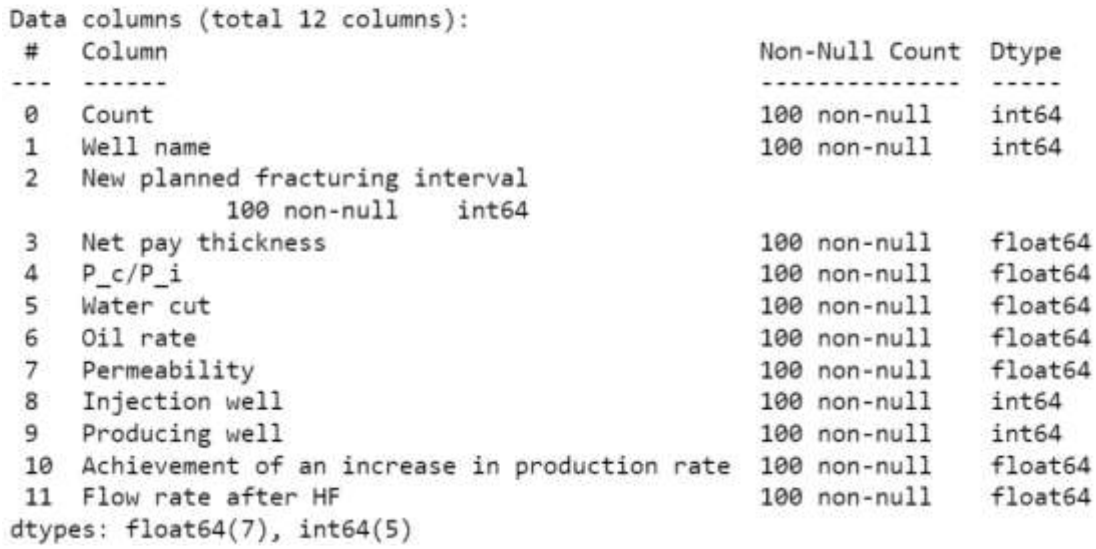
	New planned fracturing interval\n	Net pay thickness	P_c/P_i	Water cut	Oil rate	Permeability	Injection well	Producing well	Achievement of an increase in production rate	Flow rate after HF
0	0	16.0	1.170684	60.00	1.68	2.40	230	525	61.5	6.2
1	1	16.9	1.072307	39.00	5.12	5.12	355	215	104.0	12.0
2	0	13.0	1.082145	57.00	5.41	8.43	240	184	110.6	10.2
3	0	11.0	1.135191	60.00	1.35	6.80	511	316	50.0	7.7
4	1	8.5	0.735529	56.13	3.31	4.75	391	259	16.0	3.5
...	...	...	...	...	...	...	...	...	...	...
95	0	9.0	1.082145	71.56	2.39	5.35	261	182	91.7	14.3
96	0	13.0	1.032956	49.80	4.21	12.90	270	236	33.6	4.8
97	0	15.0	0.934579	64.42	1.49	7.04	296	114	104.6	9.3
98	0	11.2	0.983768	29.85	2.94	11.10	220	414	61.2	4.8
99	1	12.1	0.934579	89.41	0.18	14.00	397	371	166.5	16.8

100 rows × 10 columns

Figure 6-Two-dimensional data table-DataFrame (Jupyter notebook)

Data preparation. The data preparation process is necessary for the subsequent analysis of the information using machine learning algorithms. Data cleaning, data cleansing, or scrubbing is all about data mining - the stage in which the process of identifying and removing inconsistencies in data is carried out. Incorrect conclusions, inadequate statistics are the result of incorrect, duplicate and lost information. In this regard, data cleaning is an integral procedure when using machine learning algorithms.

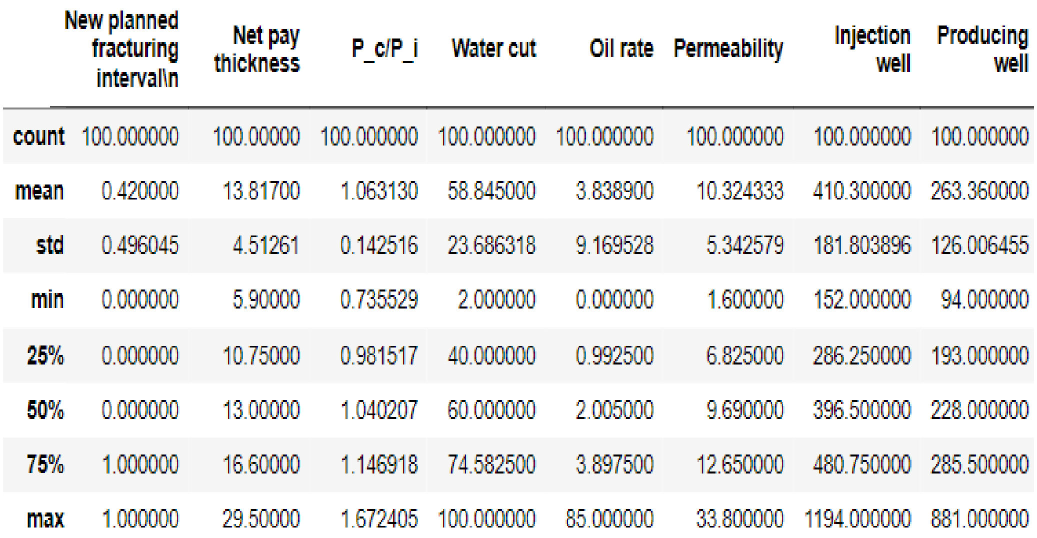
Data preparation is the process of converting the source data into a form suitable for modeling. Machine learning algorithms require the input data to be numbers. Therefore, if the data contains values that are not numbers, you will need to change this data to numbers.



```
Data columns (total 12 columns):
 #   Column                                          Non-Null Count  Dtype  
---  ------                                          --------------  -----  
 0   Count                                           100 non-null    int64  
 1   Well name                                       100 non-null    int64  
 2   New planned fracturing interval
                   100 non-null    int64  
 3   Net pay thickness                               100 non-null    float64
 4   P_c/P_i                                         100 non-null    float64
 5   Water cut                                       100 non-null    float64
 6   Oil rate                                        100 non-null    float64
 7   Permeability                                    100 non-null    float64
 8   Injection well                                  100 non-null    int64  
 9   Producing well                                  100 non-null    int64  
 10  Achievement of an increase in production rate  100 non-null    float64
 11  Flow rate after HF                              100 non-null    float64
dtypes: float64(7), int64(5)
```

Figure 7- Output of the type of values in the data set.

Data quality. Using info () , it is observed that the data type in the sample in question is integer or float. No null values were found in the data set either. Using the describe () method, you can get a statistical analysis of the data set. The statistical summary of the input variables shows that each variable has a very different scale (Figure 8).



	New planned fracturing interval\n	Net pay thickness	P_c/P_i	Water cut	Oil rate	Permeability	Injection well	Producing well
count	100.000000	100.00000	100.000000	100.000000	100.000000	100.000000	100.000000	100.000000
mean	0.420000	13.81700	1.063130	58.845000	3.838900	10.324333	410.300000	263.360000
std	0.496045	4.51261	0.142516	23.686318	9.169528	5.342579	181.803896	126.006455
min	0.000000	5.90000	0.735529	2.000000	0.000000	1.600000	152.000000	94.000000
25%	0.000000	10.75000	0.981517	40.000000	0.992500	6.825000	286.250000	193.000000
50%	0.000000	13.00000	1.040207	60.000000	2.005000	9.690000	396.500000	228.000000
75%	1.000000	16.60000	1.146918	74.582500	3.897500	12.650000	480.750000	285.500000
max	1.000000	29.50000	1.672405	100.000000	85.000000	33.800000	1194.000000	881.000000

Figure 8-Output of the statistical summary of the dataset

A histogram is created for each input variable. The output graphs confirm a different range for each input variable and show that the variables have different scales.

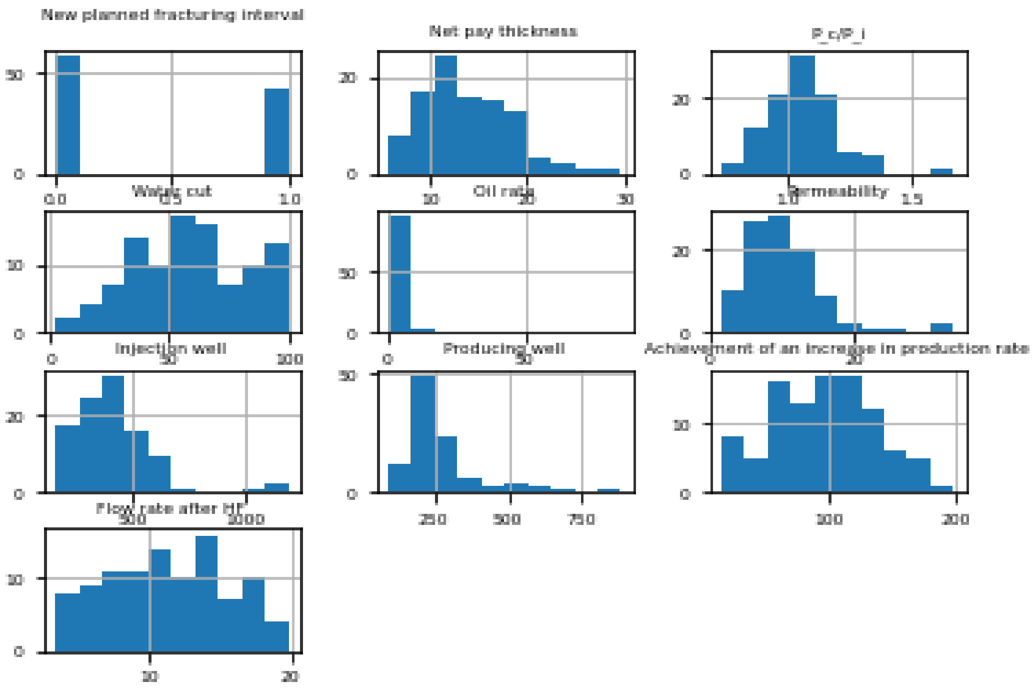


Figure 9- Histogram plots of input variables

Data quality is a multidimensional thing. The data must be simultaneously accessible, accurate, coherent, complete, consistent, defined, and relevant. A larger amount of the time resource of this study is based on the collection and preparation of data, because high – quality data is the key to a high-quality result.

Data analysis. Data analysis is a prior step before applying algorithms. The data analysis carried out in this study is divided into two sections. First of all, we will study the sample of wells from the point of view of bias and objectivity. The purpose of this analysis is to prove that the author was objective when making the selection. This means that the sampling of wells should have both positive and negative results after hydraulic fracturing.

Let's analyze the statistics of oil flow rate after hydraulic fracturing, because this variable is the target. A numerical explanation of the figure below is provided in Table 2.

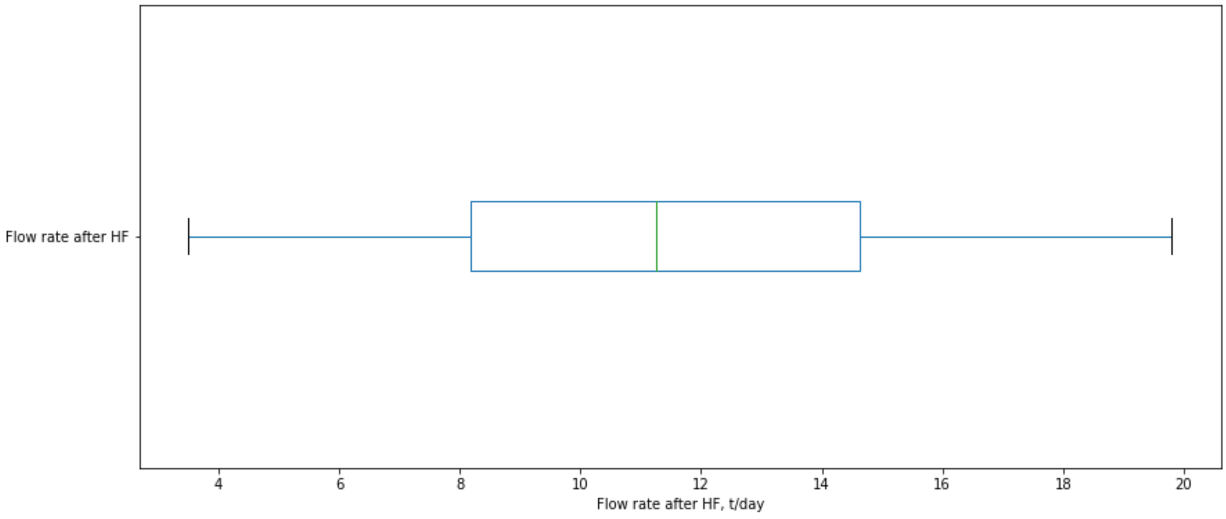


Figure 10- Sample box plot: flow rate after hydraulic fracturing

A span chart is a tool for representing numerical data in terms of quartiles. The "whiskers" represent straight lines coming out of the box, they are necessary to visualize the degree of spread (variance) beyond the upper and lower quartiles. This descriptive statistics tool allows you to quickly and efficiently explore data sets.

Figure 10 shows a diagram of the oil flow rate after the intensification of the sample under consideration. According to the figure, the average oil flow rate is 11.4 t/day, the minimum – 3.5 t/day, the maximum-19.8 t / day.

Table 2- Sample statistics: flow rate after hydraulic fracturing

Statistics	Value, [t/d]
mean	11,403
std	4,148667
min	3,5
25%	8,175
50%	11,25
75%	14,625
max	19,8

Figure 11 shows the sample density graph, which is a graphical tool for distributing data over a certain time interval. This type of visualization is a modification of the histogram, where nuclear smoothing is used to display the values, which allows displaying a smoother distribution. According to the analysis, it becomes clear that in the data sample under consideration, most wells produce an oil inflow after hydraulic fracturing from 9 to 12 tons/day. The green line corresponds to the average oil flow rate, and the red line corresponds to the median (Figure 11).

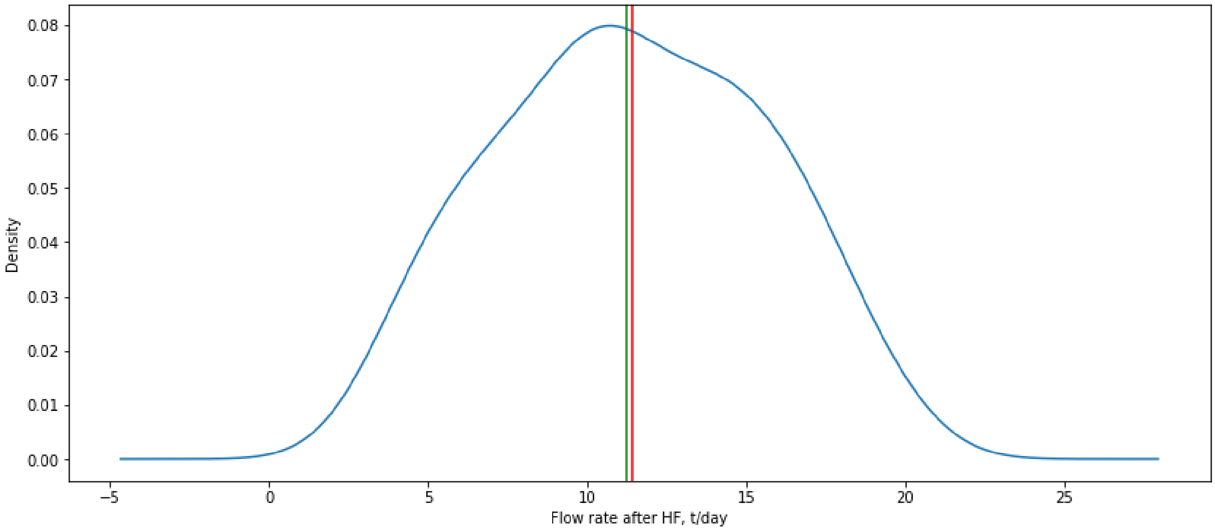


Figure 11- Sample density plot: flow rate after hydraulic fracturing

For the next stage of data sampling analysis, it is necessary to divide the wells into classes based on water cut. The classification is carried out according to the following limit values of water cut classes (Figure 12):

- Low-watered wells (0-40%) -21;
- Medium watered wells (40-80%) -50;
- High-watered wells (80-100%) -29.

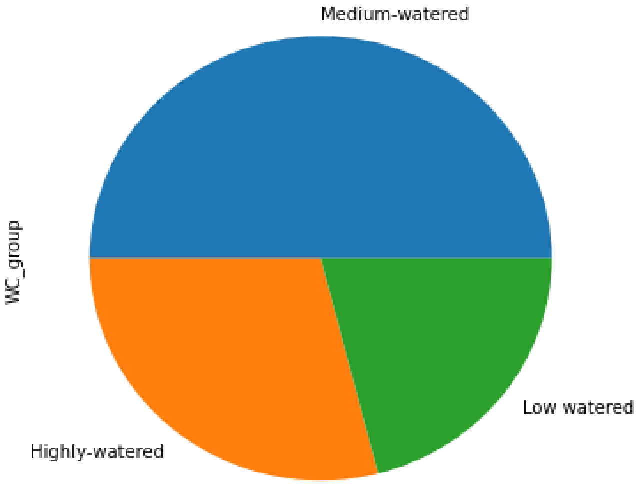


Figure 12- Classification of wells by water cut parameter

As mentioned earlier, at the Uzen field, the problem of water cut is the most common during hydraulic fracturing. The figure below shows the statistics of oil flow rate after intensification relative to the water cut groups.

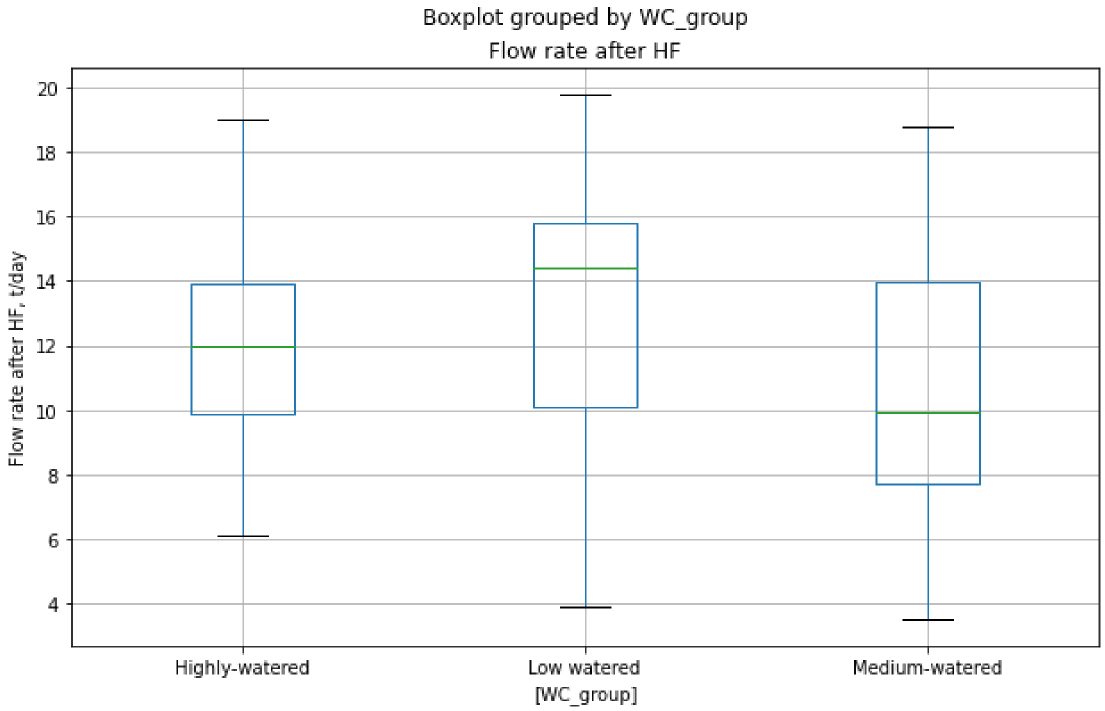


Figure 13- Box sample diagram: oil flow rate by water cut class

It should be noted that low-water wells have the highest oil inflow, which is expected. However, high-watered wells produce an average of 12 t / day compared to 10 t/day of medium-watered wells. This phenomenon is explained by the size of the sample by water cut classes and the possibility of carrying out repair and insulation work before hydraulic fracturing.

In order to argue for the objectivity of the sample of wells, it is necessary to conduct a comparative analysis with wells that have been fractured during the last three years (2020/19/18).

Table 3- Statistics of flow rate after HF: comparative analysis

Statistics	mean	std	min	25%	50%	75%	max	count
Sample, [t/d]	11.403	4.1487	3.5	8.175	11.25	14.625	19.8	100
2020, [t/d]	10.7	4.966	0.7	7.3	10.7	14.85	25.7	110
2019, [t/d]	11.609	5.4797	0.7	7.575	11.35	14.875	29.6	122
2018, [t/d]	11.059	6.454	0.1	6.9	10.7	14.6	39.4	145

Table 3 presents a statistical analysis of the oil flow rate after hydraulic fracturing over the past three years at the Uzen field.

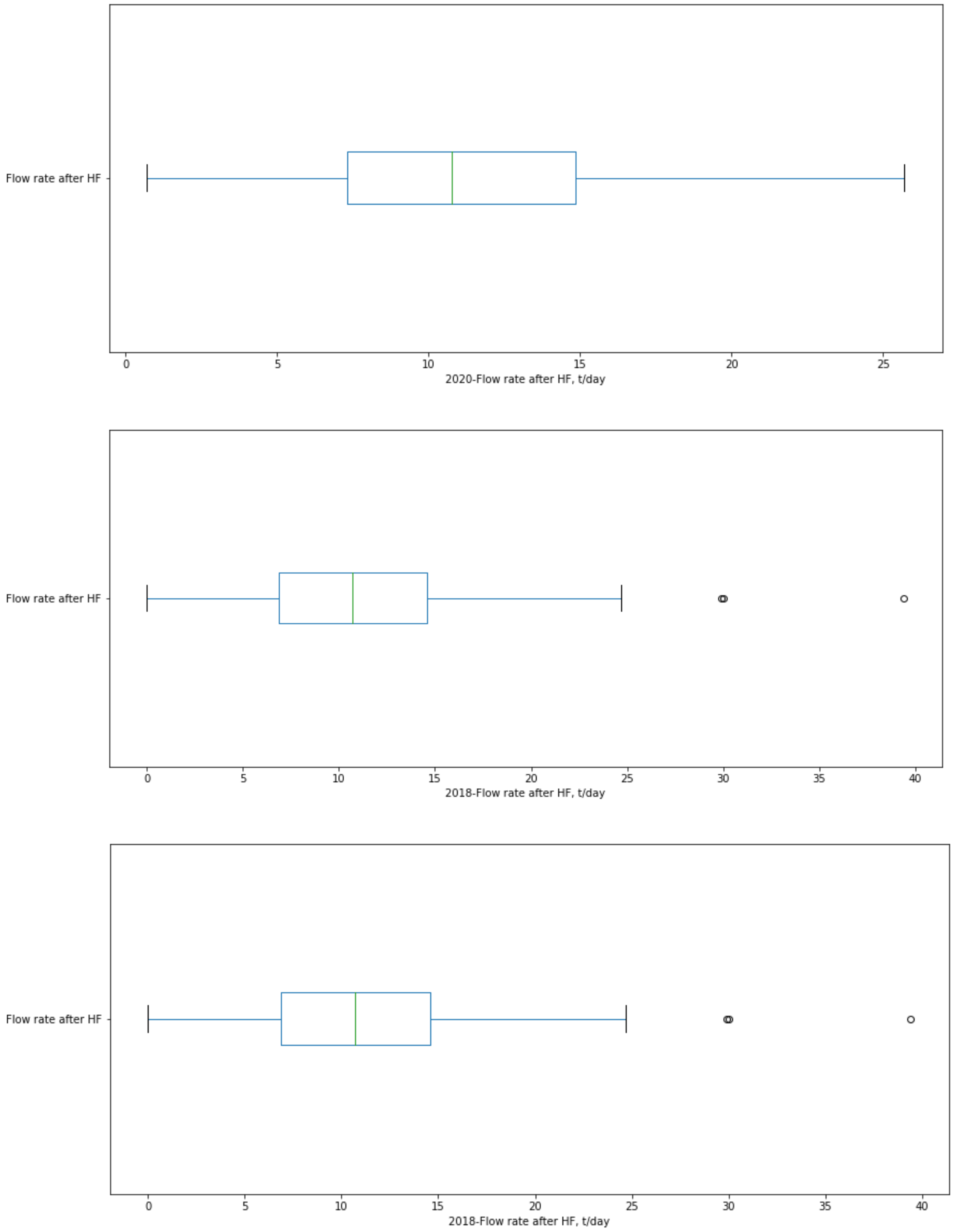


Figure 14- Box plot (2020/19/18): flow rate after hydraulic fracturing

Based on the comparative analysis, it can be argued that the average oil flow rate after hydraulic fracturing is in a small range of changes in the flow rate (10.7-11.6 t / day) for all hydraulic fracturing operations in a time period of three years. It should also be noted that there are significant deviations in the maximum flow rate for 2018. When studying the graph of the well flow rate density for the years under consideration, it becomes clear that the most frequent increase is also as close as possible to the sample produced (Appendix B).

Table 4- Distribution of wells by water content groups: comparative analysis

Group	Low watered	Medium-watered	Highly-watered
Sample	21	50	29
2020	11	67	32
2019	7	78	39
2018	10	85	50

According to the above table, the sample corresponds to the hydraulic fracturing carried out over the past three years, according to the numerical distribution relative to the water cut classes. The conducted data analysis allows us to state that the data sample of 100 wells is absolutely random and objective in relation to the majority of hydraulic fracturing operations conducted in 2020/19/18 at the Uzen field.

Previously, the analysis of the target variable was carried out, then the data analysis will be presented, based on the parameters that affect the oil flow rate after hydraulic fracturing. Data science looks at the relationships between two or more variables in a particular data set. The purpose of correlation analysis is to determine the degree of connection between the random variables under consideration. This type of analysis allows measuring the degree of connectivity of two or more phenomena, to detect unknown causes of connections, etc. A relationship is called a correlation if each value of a factor attribute corresponds to a well-defined non-random value of the resulting attribute. The linear correlation coefficient is a tool that evaluates the degree of closeness of this relationship.

In the area of machine learning, the correlation of independent variables is called multicollinearity. This phenomenon is undesirable for the data set, as it makes it difficult to analyze and evaluate the final result. The retraining of the model is a consequence of the manifestation of the multicollinear system. In addition, redundant coefficients increase the complexity of the machine learning model, hence the training time increases. Also, when using a multiple regression model, the interpretation of the parameters becomes more complicated - the regression parameters lose their meaning. The final result is large standard errors, and the regression model will not be applicable for forecasting.

The correlation matrix is used as a tool for estimating the degree of correlation of parameters (Figure 15).

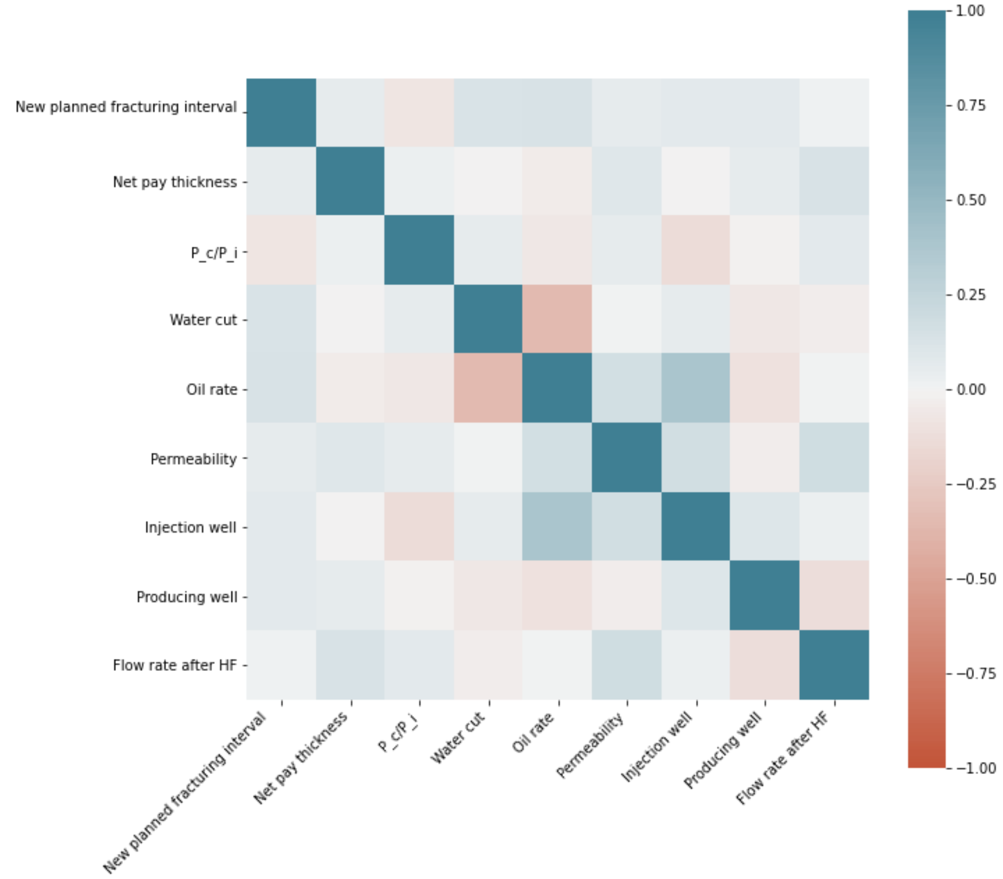
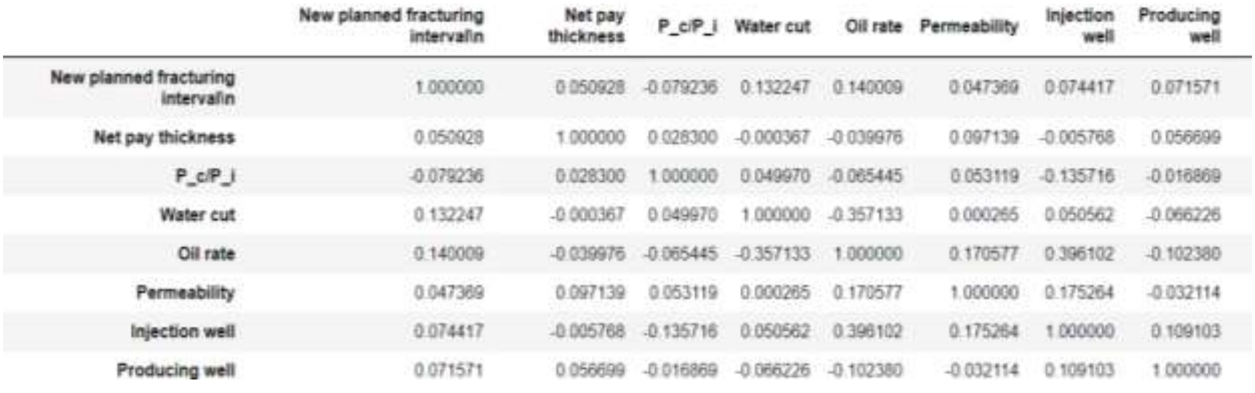


Figure 15- Correlation matrix of parameters

According to the above figure, each cell of the correlation matrix is a "correlation coefficient" between two variables corresponding to the row and column of the cell. The correlation coefficient for each pair of parameters is shown in Figure 16.



	New planned fracturing intervalin	Net pay thickness	P_c/P_i	Water cut	Oil rate	Permeability	Injection well	Producing well
New planned fracturing intervalin	1.000000	0.050928	-0.079236	0.132247	0.140009	0.047369	0.074417	0.071571
Net pay thickness	0.050928	1.000000	0.028300	-0.000367	-0.039976	0.097139	-0.005768	0.056699
P_c/P_i	-0.079236	0.028300	1.000000	0.049970	-0.065445	0.053119	-0.135716	-0.016869
Water cut	0.132247	-0.000367	0.049970	1.000000	-0.357133	0.000265	0.050562	-0.066226
Oil rate	0.140009	-0.039976	-0.065445	-0.357133	1.000000	0.170577	0.396102	-0.102380
Permeability	0.047369	0.097139	0.053119	0.000265	0.170577	1.000000	0.175264	-0.032114
Injection well	0.074417	-0.005768	-0.135716	0.050562	0.396102	0.175264	1.000000	0.109103
Producing well	0.071571	0.056699	-0.016869	-0.066226	-0.102380	-0.032114	0.109103	1.000000

Figure 16- Output of the numerical expression of the correlation matrix

The Pearson coefficient is a coefficient defined as the covariance between two variables divided by the product of the standard deviations of these two variables.

$$\rho(X,Y) = \frac{COV(X,Y)}{\sigma_X * \sigma_Y} \quad \textbf{(1)}$$

The value $\rho(X,Y)$lies in the range [-1 ; 1]. Values close to +1 indicate a strong positive relationship between X and Y, while values close to -1 indicate a strong negative relationship between X and Y. Values close to zero indicate that there is no relationship between X and Y. Each cell in the matrix above is also represented by shades of color. Here, red shades of color indicate a negative correlation, while blue shades correspond to a positive correlation. According to the above analysis, no critically correlated data were found. Therefore, for the further application of machine learning algorithms, the previously considered list of influencing parameters is approved.

2.2 Development of models for predicting the flow rate after hydraulic fracturing

The development of a model for predicting the flow rate after a hydraulic fracturing operation is based on a machine learning algorithm. In this study, the code contains an algorithm for finding solutions independently through the integrated use of statistical data. Further, certain patterns are identified, on the basis of which the flow rate after hydraulic fracturing is predicted for potential candidates.

The main problem of machine learning is that today there is no single flexible algorithm applicable to any data sample, regardless of the optimization industry. In addition to the above, it should be noted that different types of algorithms have a different degree of efficiency of the result. For example, you can't say that decision trees work better than neural networks in all cases, and vice versa. The structure and size of the data set largely determine the success of the type of algorithm used. For this reason, during the research work, a number of algorithms were used to identify the most efficient and suitable for the existing data sample.

Machine learning algorithms can be described as learning the objective function *f* that best matches the input variables *X* and the output variable *Y: Y = f(X).* The most common task in machine learning is to predict *Y* values for new *X* values. This is called predictive modeling, which aims to make the most accurate prediction possible. Next, we will consider the algorithms used in this study to predict the oil flow rate after hydraulic fracturing.

2.2.1 Regression Algorithms: Linear regression

The most basic and fundamental algorithm used to identify the relationship between a dependent variable and one or more independent variables is linear regression. This algorithm is focused on finding the "best match line". The best match line is found by minimizing the squared distances between the points and the best match line.

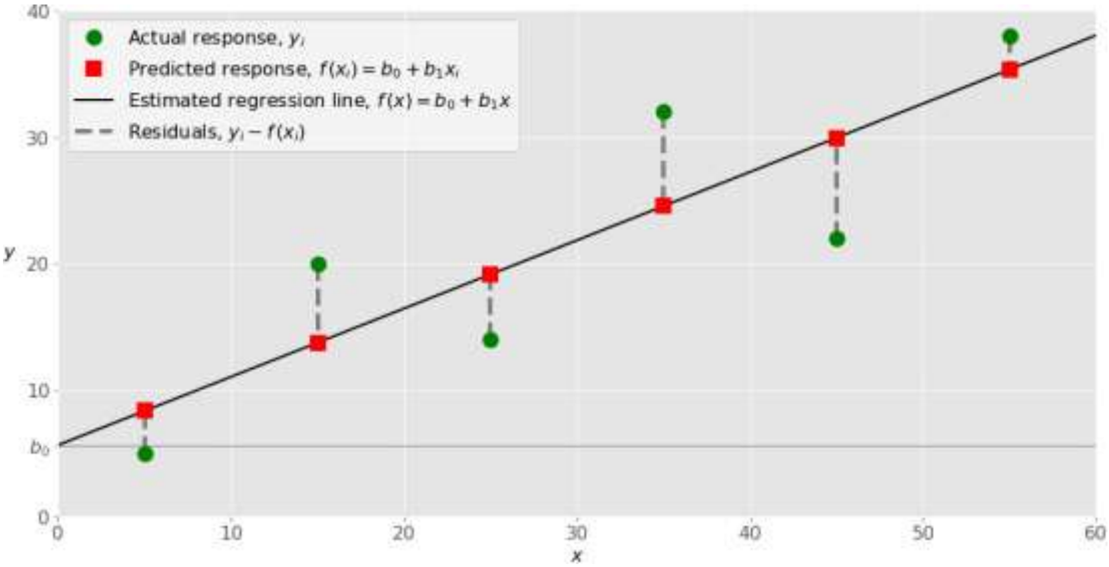


Figure 17- One-dimensional linear regression model [7]

In this study, multiple regression is considered, because the predicted flow rate after hydraulic fracturing depends on several independent variables (factors discussed in the first chapter). The equation below represents a linear regression model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \quad \textbf{(2)}$$

where,

$x_1, x_2, x_n - sets\ of\ input\ values;$

$y - model\ output\ data;$

$\beta_0, \beta_1, \beta_n -\ coefficients\ of\ a\ linear\ equation.$

2.2.2 Regularization Algorithms: Least Absolute Shrinkage and Selection Operator

Linear regression is unstable if it shows an overestimated degree of dependence on training data, which usually leads to the phenomenon of overfitting. The use of the regularization method avoids the consequences of an unstable model. Regularization is based on the imposition of additional constraints on the initial parameters, which prevent excessive complexity of the model. The LASSO regression model uses coefficient compression, meaning that the data approaches the mean value.

The introduction of an additional regularization term in the optimization functional of the model determines the effectiveness of the LASSO regression method. The following formula expresses the condition for minimizing the squared error in parameter estimation:

$$\beta = \sum \left(y_{true} - y_{pred}\right)^2 + \lambda \sum_{i=1}^{n} |\beta_i| \qquad (3)$$

where,

λ- regularization parameter that has the meaning of a penalty for complexity.

In this case, a certain compromise is reached between the regression error and the dimension of the used feature space, expressed by the sum of the absolute values of the coefficients $|\beta_i|$. During minimization, some coefficients become equal to zero, which, in fact, determines the selection of informative features. This compression process allows you to get the most stable and accurate estimates of the true parameters. In LASSO regression, instead of taking the square of each coefficient, their absolute values are taken.

2.2.3 Regularization Algorithms: Ridge regression

Ridge regression, like LASSO regression, is a modification of linear regression. The similarity of Lasso regression to ridge regression is the application of the compression process in both cases. Both algorithms are used with a high degree of efficiency to sample data with a large number of influencing features, the correlation of which can lead to a multicollinear system. However, the main difference in these types of regression is that in Ridge regression, none of the coefficients becomes zero, which can be observed in LASSO regression. Let’s look at the cost function of Ridge:

$$\beta_R = \sum \left(y_{true} - y_{pred}\right)^2 + \lambda * \sum_{i=1}^{n} \beta_i^2 \tag{4}$$

Using Ridge regression avoids the effect of over fitting with lower coefficients. Lambda (λ) is a constant, hence it has the same scaling effect on all coefficients. Regression Ridge has a detailed approach, as it allows you to make important features more pronounced and reduce the influence of factors that have the least effect. This is because when squaring a larger number, the result is an even larger number. On the contrary, if you select a low value (for example, 0.01), the result will decrease significantly. But when squaring numbers less than one, and then multiplying it by 0.01, we get an even smaller result. Thus, this algorithm works perfectly even with a high degree of correlation of influencing factors. Since the influence of all factors is taken into account, but the coefficients are distributed among the factors depending on the correlation.

2.2.4 Polynomial regression

The polynomial regression is a special case of the previously considered linear regression. The polynomial regression algorithm models the relationship between the independent variable (x) and the dependent variable (y) as an n-th degree polynomial. The main equation of the polynomial regression is given below:

$$y = b_0 + b_1 x_1 + b_2 x_1^2 + \cdots + b_n x_1^n \tag{5}$$

Polynomial regression is perfectly applicable to a data set that is characterized by non-linearity. Because using a linear model to the above data set will increase the loss function, increase the error rate, and therefore decrease the accuracy of the result. However, the presence of one or two outliers has a significant impact on the results of the power analysis. This means that the polynomial regression is very sensitive to outliers. This type of regression is often used in mathematical statistics when modeling the trend components of time series.

2.2.5 Random Forest Regression

The Random Forest algorithm is one of the most common algorithms used in machine learning. This is due to its versatility: the algorithm is applicable in solving problems of classification, regression, anomaly search, clustering, etc. A random forest is a collection of a certain number of decision trees. This means that trees constructed "randomly" make up a Random forest. Each tree is formed from a selection of rows, and a different selection of objects is selected at each node for partitioning. Each of the "random" trees models its own individual prediction. Further, the available forecasts are averaged to obtain a single, more accurate result. The figure below shows the structure of the Random Forest algorithm.

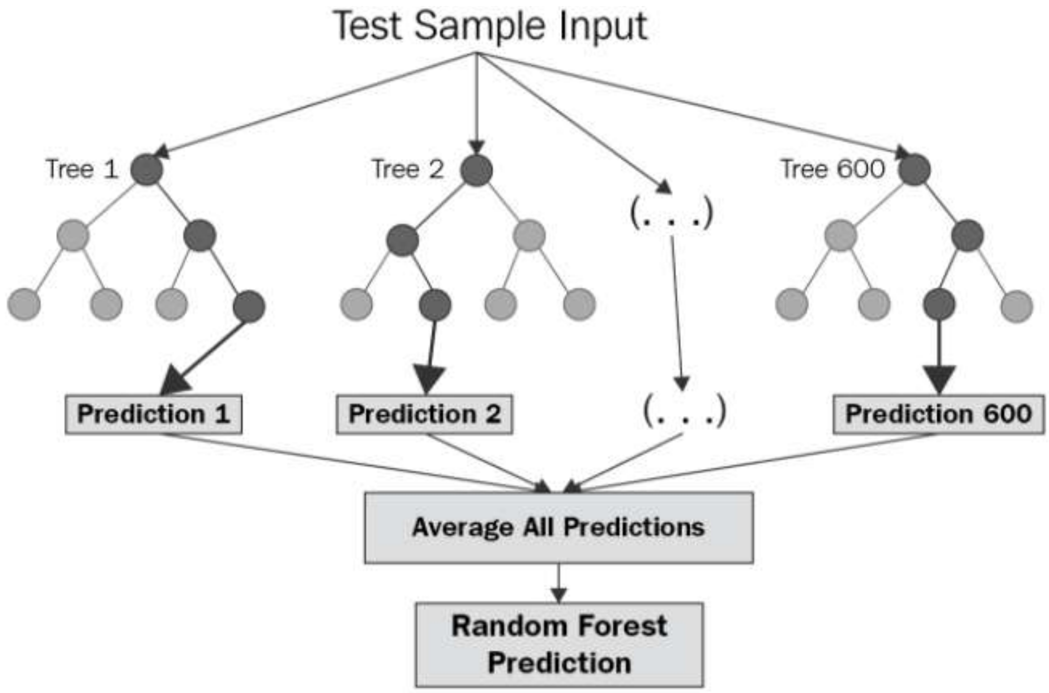


Figure 18- Structure of a Random Forest Regression [8]

2.2.6 Ensemble method

To date, the methods of ensembling are powerful tools that are most often used in machine learning. The popularity and high degree of effectiveness is explained by the assumption that combining several models together leads to the creation of a much more powerful model. Getting the best prediction performance is the main goal of any machine learning algorithm. Thus, an ensemble of methods improves the prediction result by using several training algorithms. The flexibility of ensemble methods is provided by a larger number of parameters than individual models of algorithms. It should also be noted that with the ensemble method, it is important to synthesize distinctive models, because if similar algorithms are combined, the error increases.

It should be noted that in the case of ensembling, the numeric input variables change the scale to the standard range. When using the standardization function, each source variable is scaled individually, subtracting the mean (called centering) and dividing by the standard deviation to shift the distribution to produce a zero mean and a unit standard deviation.

In this study, eight factors influence the forecast flow rate after hydraulic fracturing. Including more features does not always improve the performance of the algorithm. The principal components method solves this problem by reducing the dimension of the feature data space. The data set in question, like any data sample, contains noise. The phenomenon of model over fitting is a consequence of the presence of data noise. Using the principal components method avoids this problem. It is assumed that the variance of the noise is small relative to the variance of the data itself, and after converting the data by the principal component method, the transformed data (components) whose variances are small will be considered noise. They can be safely excluded from subsequent training, assuming that the quality of the training model, at least, will not decrease.

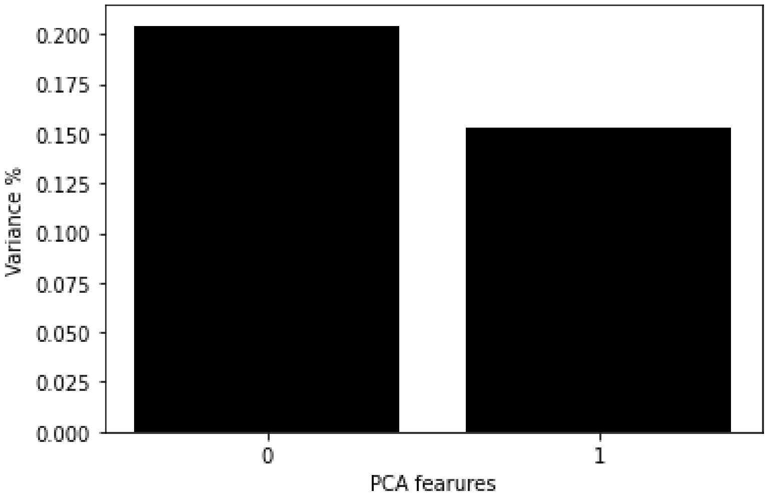


Figure 19- Output principal component analysis-variance reduction

Principal component analysis (PCA) is an exploratory approach to reducing the dimension of a data set, in this case to 2D, used in exploratory data analysis to create predictive models. This method focuses on finding an orthonormal basis for data, sorting measurements in order of importance, and excluding low-significance measurements. The PCA method is described by the eigenvectors and eigenvalues of the covariance matrix. The eigenvectors (principal components) determine the direction of the new attribute space, and the eigenvalues determine its magnitude. Reducing the dimension of the data leads to their projection into a smaller subspace, where the eigenvectors form the axes (Figure 20). The figure below shows the process of changing an eight-dimensional space to a two-dimensional space.

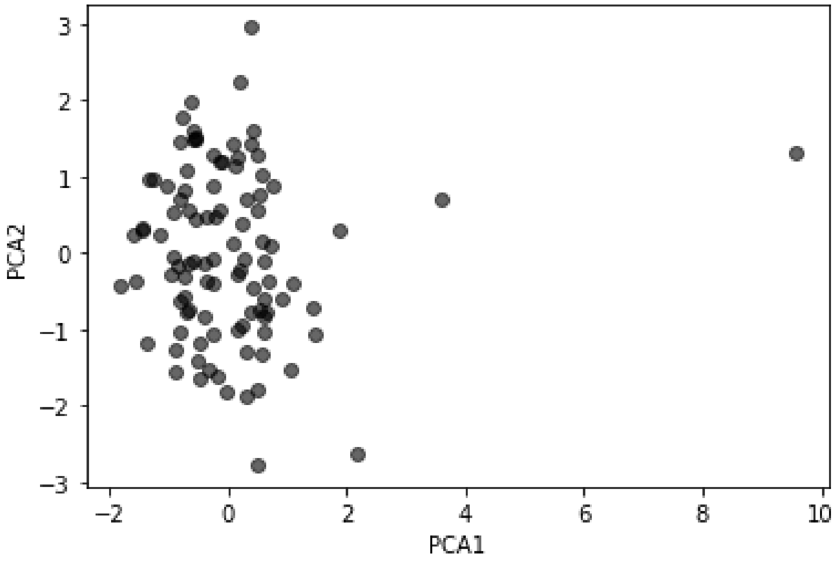


Figure 20- Decorrelation of a new feature space

Next, we use the k-means implementation of clustering, which is a machine learning method that identifies clusters of data objects in a data set. In general, clustering involves dividing data into groups (clusters). Clusters are defined as groups that are more similar to other objects in their cluster than to data objects in other clusters. The main element of the algorithm works based on the process of maximizing expectations. The waiting step assigns each data point its nearest centroid. Then, at the maximization stage, the average value of all points for each cluster is calculated and a new centroid is set. The quality of the defined clusters is based on calculating the sum of squared errors (SSE) after the centroids converge or coincide with the destination of the previous iteration. SSE is defined as the sum of the squared Euclidean distances of each point to its nearest centroid. Since this is a measure of error, the goal of k-means is to try to minimize this value (Figure 21). The appendix provides an example of binding a sample of data, namely each well, to a specific cluster.

After clustering, a Random Forest regression is applied to the selected number of clusters. And then the forecast of oil production after the hydraulic fracturing operation is modeled.

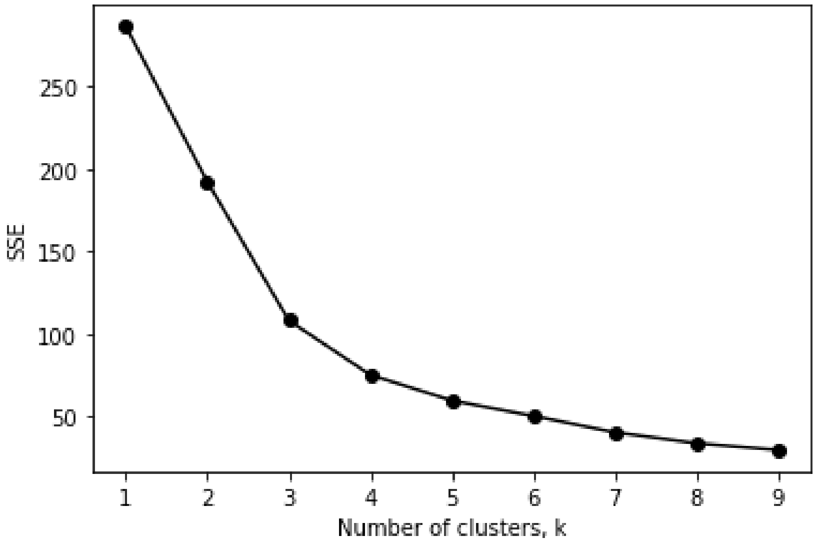


Figure 21- Determination of the elbow point

2.3 Summary of Chapter 2

There's no such thing as a free lunch. The main idea of this expression in the key of this study is that there is no single flexible algorithm for any data set, regardless of the industry of the problem being solved. But this is the whole point of interest: finding the most suitable algorithm model for the problem under consideration.

The linear regression algorithm is the simplest, but it has good performance. However, restrictions on the freedom of maneuver are the reason for the low frequency of applying this algorithm to real data. To a greater extent, the linear regression model is used as a base model for comparison with other machine learning algorithms.

Problems related to the phenomenon of overfitting and bias are solved using Ridge and LASSO regression. LASSO regression allows to exclude features that have little effect on the prediction of production.

When solving a complex forecasting problem, a common case is that the use of any of the algorithms does not provide the desired quality of dependency recovery. In such cases, there is the creation of a composition of algorithms, namely, the use of the method of ensembles of models. Due to the fact that this study considers a relatively massive data set, the inclusion of the principal components and clustering method in the modeling helps to avoid the influence of data noise and the phenomenon of multicollinearity by reducing the data dimension.

DISCUSSION

3.1 Selection and analysis of algorithm quality metrics

In the second chapter, we discussed the various types of algorithms that were used in this study to predict the flow rate of oil after a hydraulic fracturing operation. The comparison of the results is an integral part of the entire modeling process, this process is carried out on the basis of the selected metric. The choice of a suitable evaluation system-metrics - has an impact on how the performance of machine learning algorithms is measured and compared. And the most important point is that the metric influences the final choice of the working algorithm for the problem being solved.

In this study, we consider the problem of predictive modeling, which results in the prediction of a numerical value. This type of task is fundamentally different from classification tasks, which involve predicting the class label. Therefore, using classification accuracy to evaluate forecasts is incompetent. This leads to the fact that it is necessary to use the error indicators developed for regression models. As I mentioned earlier, predictive modeling is a task that is solved using historical data to predict new data. Predictive modeling can be described as a mathematical problem of approximating the function of mapping input variables to output variables. Therefore, in this case, it is impossible to assess the accuracy of the developed models. The performance of the regression model is characterized by the approximation of forecasts to their expected values.

The most common quality measures in regression problems are the following errors:

- Mean absolute error (MAE);
- Root mean square error (RMSE);
- The coefficient of determination (R^2) .

Mean absolute error. The sum of the absolute differences between the simulated and actual values forms the average absolute error. This type of error characterizes how incorrect the forecast values are. However, having an idea of the magnitude of the numeric value, there is no knowledge of the overestimated or insufficient performance of the algorithm. The formal equation corresponding to this type of error is presented below:

$$MAE = \frac{1}{n} * \sum y_{true} - y_{pred} \quad \textbf{(6)}$$

Root mean square error. This type of error is the square root of the average square of the entire error. RMSE is a good measure of accuracy, but subject to comparing the prediction errors of different model configurations for a particular variable, rather than between variables, as the scale effect is apparent. This is a measure of how well the regression line matches the data points. The formula for calculating RMSE is as follows:

$$RMSE = \sqrt[2]{\frac{\sum_{i=1}^{n}\left(y_{true} - y_{pred}\right)^2}{n}} \tag{7}$$

The coefficient of determination. For quality control during training, the root-mean-square error is effectively used, but this error does not give a concept of the degree of correctness of the problem being solved. Therefore, to compare the available machine learning algorithms, it is necessary to enter the coefficient of determination. The coefficient of determination measures the proportion of variance explained by the model in the total variance of the target variable. Provided that this error is as close to one as possible, the developed model explains the data well, but if it is close to zero, then the forecasts are comparable in quality to the constant prediction. The coefficient of determination is calculated according to the following formula:

$$R^2 = \frac{\sum_{i=1}^{n}\left(y'_{true} - y_{pred}\right)^2}{\sum_{i=1}^{n}\left(y_{true} - y_{pred}\right)^2} \tag{8}$$

3.2Target object of forecasting: oil flow rate or the percentage of achieving the planned flow rate

In this section, I will demonstrate one of the problems that I encountered while conducting research. The main task of the algorithm is to simulate the forecast of oil production after the hydraulic fracturing process. At the first stage of the study, the target production forecast represented the percentage of achieving the planned flow rate after hydraulic fracturing. This value depends on the value of the planned flow rate, which in turn is determined by a number of factors. The planned flow rate is a dynamic value, which largely depends on the economic conditions of the oil and gas market. The percentage of achievement of the planned flow rate can take different values, the hydraulic fracturing is considered successful if the achievement value exceeds one hundred percent. Predicting this value, the following error results were obtained: MAE=13.6961, RMSE=16.8249. Table 5 shows the forecast of the percentage of achieving the planned flow rate for ten random wells from the considered data set.

Table 5- Percentage of reaching the planned flow rate: output forecast / actual values

Well #	Achievement of an increase in production rate, [%]	
	Actual values	Forecast values
1	61,5	76,57
2	104	103,75
3	110,6	96,56
4	50	64,57
5	16	49,84
6	91,7	98,66
7	33,6	54,56
8	104,6	100,07
9	61,2	75,39
10	166,5	142,62

The increased amount of errors and the lack of knowledge of the actual flow rate after hydraulic fracturing reveals that this target value does not provide the desired quality of the result. The percentage of achieving the planned flow rate is not a variable proportional to the factors discussed in the first chapter. Therefore, choosing the correct forecast target value is the most important step in the simulation. To improve the results, the oil flow rate (t/day) after hydraulic fracturing was selected as the target value. The results of forecasting and comparative analysis are presented in the next section.

3.3 Comparative analysis of the results of forecasting the oil flow rate from various algorithms

This section is devoted to the comparative analysis of the results of various algorithms for modeling the forecast of oil production after hydraulic fracturing. As mentioned in section 3.1, the identification of the most appropriate algorithm for the data sample under consideration will be based on the error metric.

The results of predicting the linear regression algorithm on random ten wells are presented in Table 6. The visualization of the correlation of actual and predicted values is shown in Figure 22.

Table 6- Predicting oil production after hydraulic fracturing: Linear Regression

Actual values, [t/day]	11.01	5.3	10.2	14.7	16.3	14.3	15.8	13.5	6.6	6.9
Predicted values, [t/day]	13.818	10.069	11.849	10.164	7.654	10.949	6.844	12.009	12.623	11.954

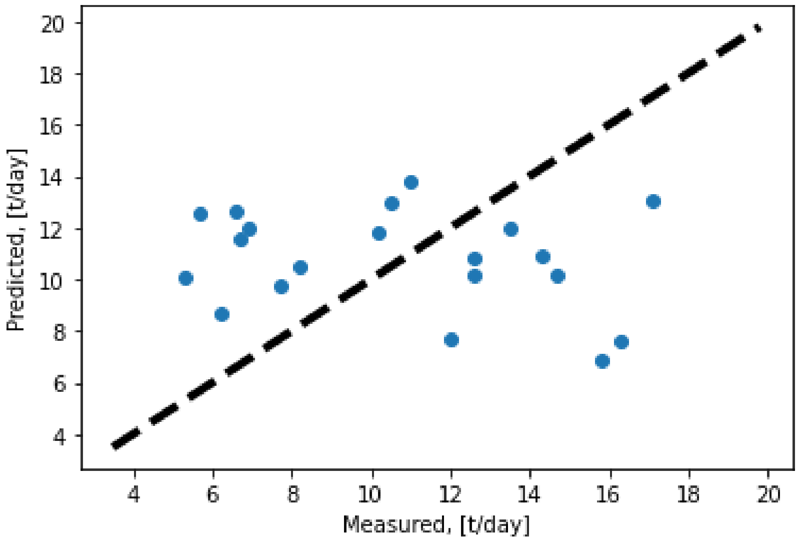


Figure 22- Output: Linear Regression

It should be noted that there is no excellent correlation between the forecast and actual values of oil production, because there are significant deviations in the predictions.

Next, consider the modifications of linear regression. The results of predicting the LASSO regression algorithm on random ten wells are presented in Table 7. A visualization of the correlation of actual and predicted values is shown in Figure 23.

Table 7- Predicting oil production after hydraulic fracturing: LASSO Regression

Actual values, [t/day]	11.01	5.30	10.2	14.7	16.3	14.3	15.8	13.5	6.6	6.9
Predicted values, [t/day]	11.386	6.278	12.485	12.463	13.314	13.183	12.571	14.103	8.638	7.536

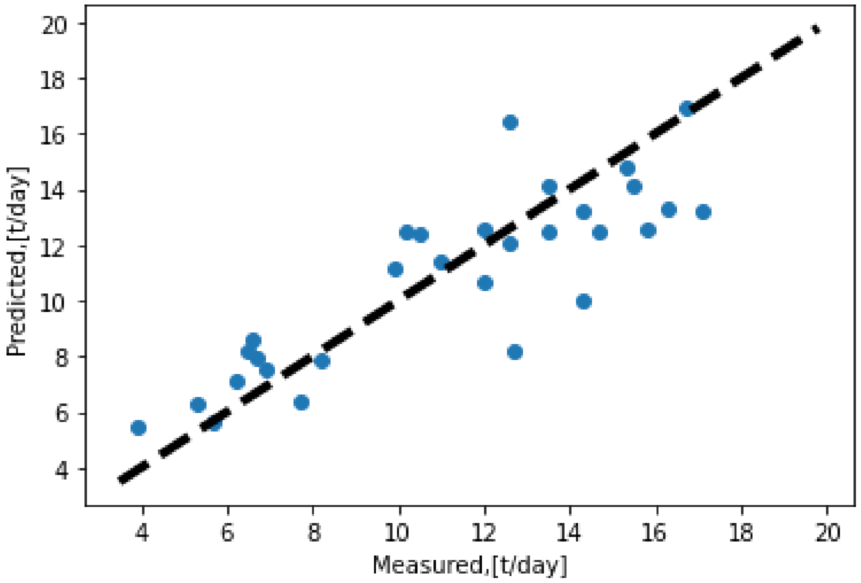


Figure 23- Output: LASSO Regression

When using the LASSO model, the results of predicting the oil flow rate are significantly improved. In Table 7, there are no high drifts of oil values in comparison with the linear regression forecasts. The error is determined when the actual oil values exceed the average value.

Similar results of the LASSO algorithm are observed when using Ridge regression. Differences in the forecast flow rate in regression, the Ridge has reduced values only in hundredths (Table 8).

Table 8- Predicting oil production after hydraulic fracturing: Ridge Regression

Actual values, [t/day]	11.01	5.30	10.2	14.7	16.3	14.3	15.8	13.5	6.6	6.9
Predicted values, [t/day]	11.006	6.091	12.414	12.14	13.068	13.009	12.073	14.025	8.519	7.466

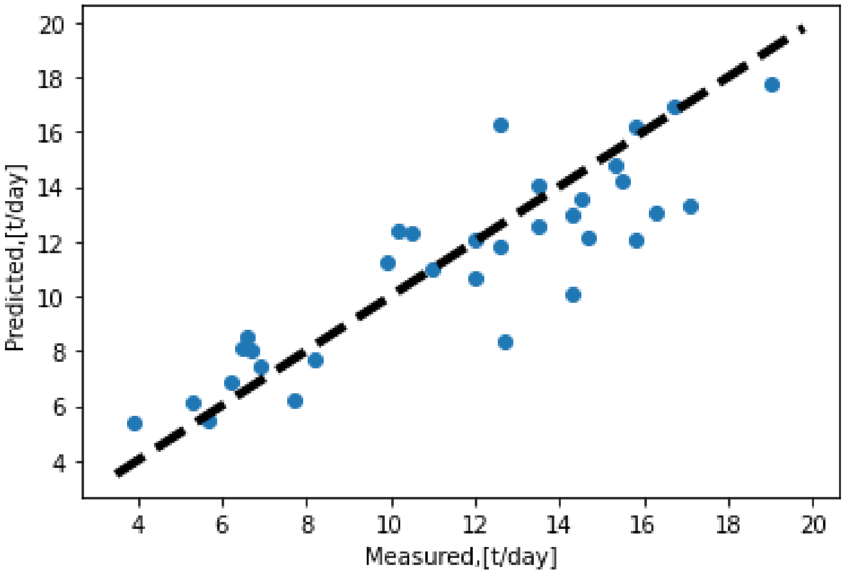


Figure 24- Output: Ridge Regression

Next, we will consider the results of applying the polynomial regression algorithm. In the table below, there is a significant difference between the actual and projected oil flow rates.

Table 9- Predicting oil production after hydraulic fracturing: Polynomial regression

Actual values, [t/day]	6.8	3.90	5.8	9.4	17.4	12.7	13.8	6.7	12.5	6.2
Predicted values, [t/day]	15.829	9.631	11.956	10.962	7.791	11.093	6.045	12.041	12.002	11.998

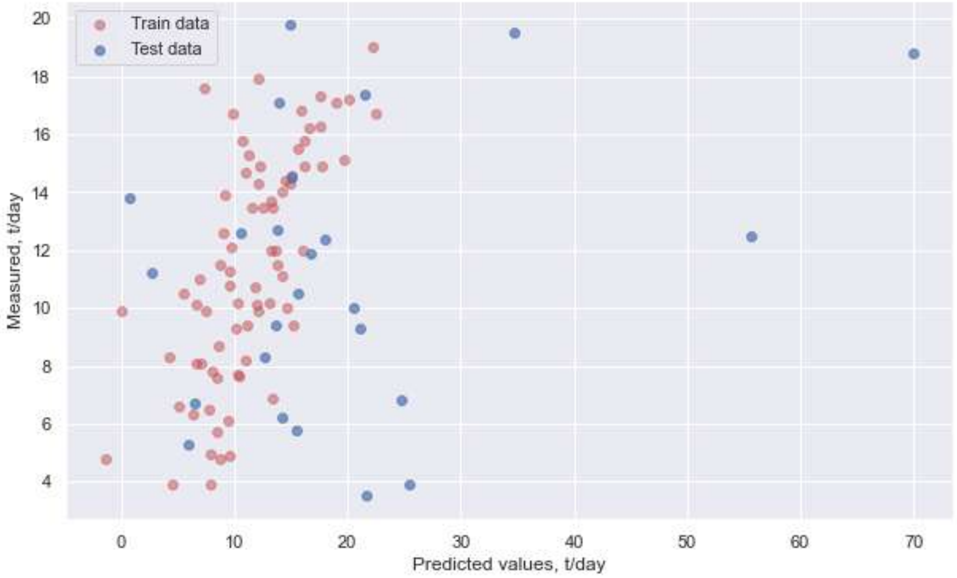


Figure 25- Output: Polynomial regression

According to Figure 24, we observe a less pronounced correlation directly in the test data. Therefore, this forecast behavior is the cause of high final errors. A similar situation is observed when applying the Random Forest algorithm. The difference in the results of the forecast of the flow rate of training and test wells can be observed in Figure 25.

Table 10- Predicting oil production after hydraulic fracturing: Random Forest Regression

Actual values, [t/day]	11.01	5.30	10.2	14.7	16.3	14.3	15.8	13.5	6.6	6.9
Predicted values, [t/day]	15.212	12.32	12.163	12.229	10.079	9.892	9.31	12.198	13.888	10.658

Let's move on to the final algorithm for predicting the oil flow rate after hydraulic fracturing - the method of ensembling. From the visual interpretation of Table 11, we can conclude that this algorithm is the most suitable for the data sample under consideration.

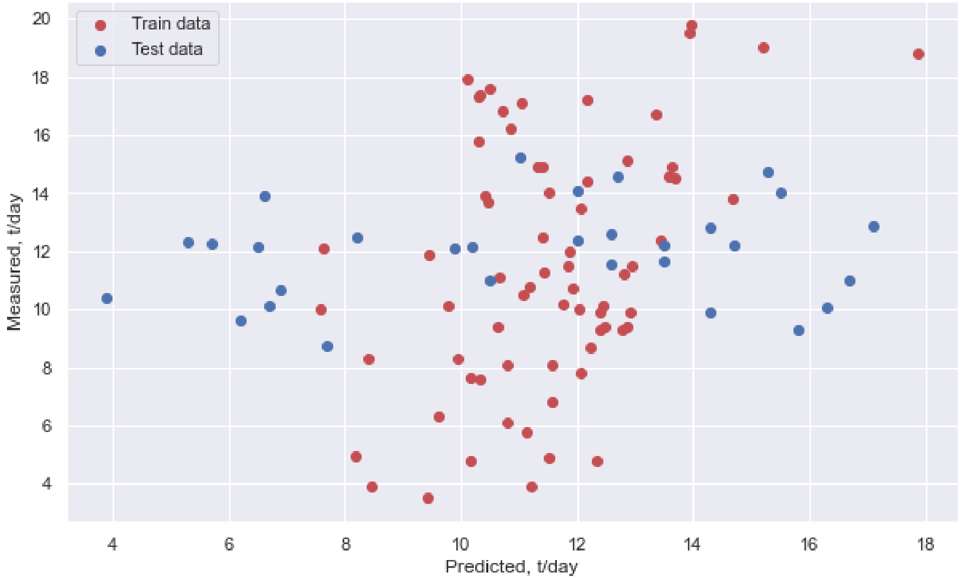


Figure 26- Output: Random Forest Regression

Table 11- Predicting oil production after hydraulic fracturing: Ensemble method

Actual values, [t/day]	6.2	12.00	10.2	7.7	3.5	14.3	4.8	9.3	4.8	16.8
Predicted values, [t/day]	7.362	11.94	10.44	8.208	6.045	12.86	6.985	9.465	7.877	14.912

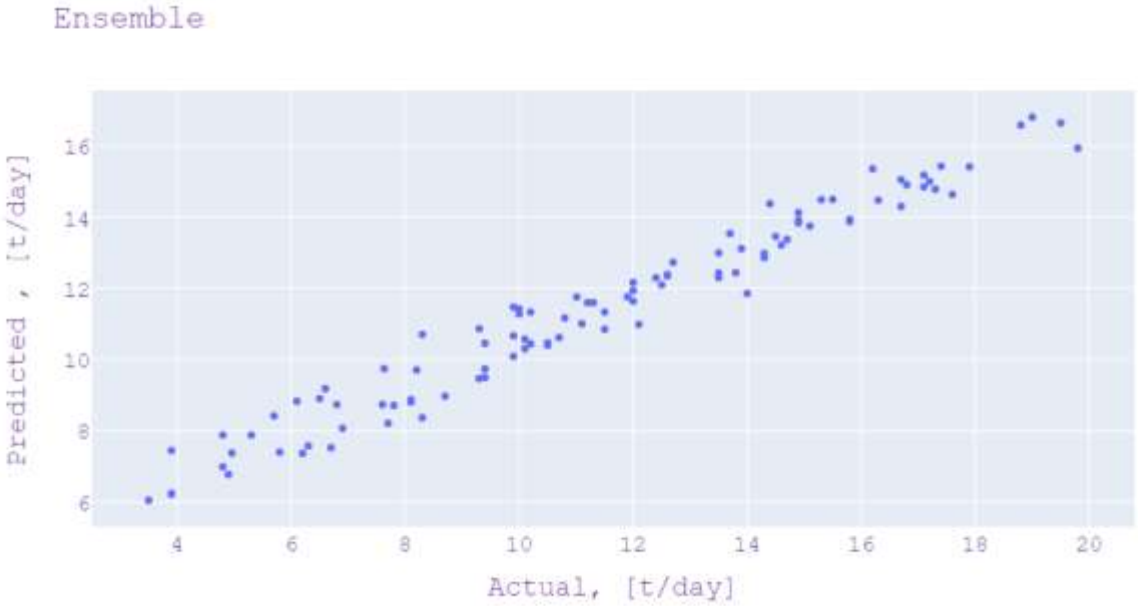


Figure 27- Ensemble method

However, there are increased values of the forecast flow rate at low actual values. This phenomenon is explained by the fact that in fact, low flow rates are the "noise" of the sample under consideration. These "noises" relative to the average flow rate are also significantly underestimated, which is the reason for the presence of forecasting errors in the ensemble method.

Table 12- Comparative analysis of algorithms

Algorithm	MAE	RMSE	R^2
Linear regression	4.039	4.582	0.4928
LASSO Regression	1.623	2.039	0.721
Ridge Regression	1.546	1.98	0.757
Polynomial regression	4.378	5.41	0.678
Random Forest Regression	1.208	1.469	0.664
Ensemble method	0.1475	1.5449	0.8591

The above table also proves that the method of model ensembles has comparatively the best results. The coefficient of determination is as close as possible to one (0.86) with the method of ensembling. Therefore, this algorithm is most suitable for identifying the best candidate for hydraulic fracturing operations.

3.4 Forecasting production on potential candidates for hydraulic fracturing

The practical significance of this research work lies in the application of the developed algorithm to potential wells. After conducting a comparative analysis and selecting the most effective algorithm for solving the current problem, the next step is to implement the process of selecting the best candidate for hydraulic fracturing. To implement this task, a database of potential candidates (Appendix D) was formed from the fund of current wells of the Uzen field. After filling in the database based on previously established criteria, I applied the selected algorithm (an ensemble of models) to this database. The results of predicting the oil flow rate for potential candidates are shown in Figure 28:

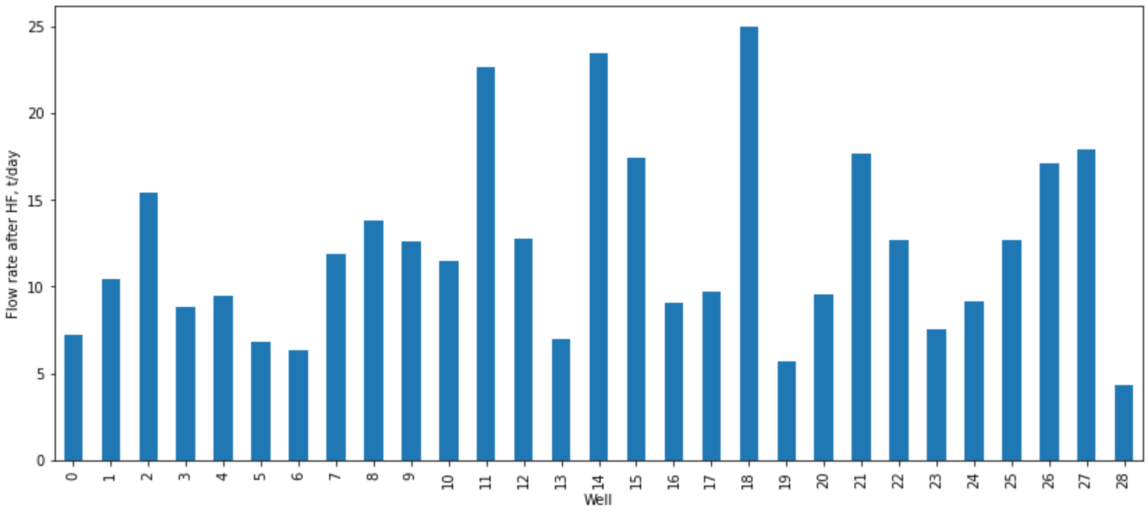


Figure 28- Production forecasting: potential candidates for hydraulic fracturing

The application of the chosen algorithm allows to determine the best candidates for fracking operations among potential wells. For example, wells 11, 14, 18 are the best candidates, because their flow rate exceeds 20t/day. In comparison with wells 28, 19 - the forecast production of which does not exceed 6 t / day. This technology allows the selection of candidates for hydraulic fracturing operations to be carried out in an accelerated format.

Checking the quality of the algorithm performance was the next stage of this study. After the request was made, materials were received on the results of hydraulic fracturing carried out in 2021 (January, February, March). Next, a database was created from a sample of potential candidates, including wells with a recent hydraulic fracturing operation. The next stage is a comparative analysis of the forecast flow rate of the candidates considered earlier and the actual flow rates of the wells where the hydraulic fracturing was carried out in 2021. The forecast results are shown in the figure below: MAE=4.7; RMSE=5.902; $R^2 = 0.5364$. In comparison with the previously studied sample of 100 wells, there is a hanging error value. It should also be noted that the greatest deviation in the forecast values of the flow rate is observed at the actual low oil flow rates (left part, Figure 29). The existing margin of error is due to a number of factors:

- The projected target flow rate is the average flow rate for three months after the hydraulic fracturing. Since hydraulic fracturing at the wells under consideration was carried out relatively recently, the wells did not reach their potential. For a qualitative analysis, the time factor is necessary.

- The sample of candidate wells is a new and unknown data set for the algorithm. Therefore, the presence of a certain percentage of errors, for the specified reason, cannot be avoided.

- The scale of the original sample is relatively small. Increasing the number of wells that present historical data will improve the flow rate prediction result and significantly reduce the error.

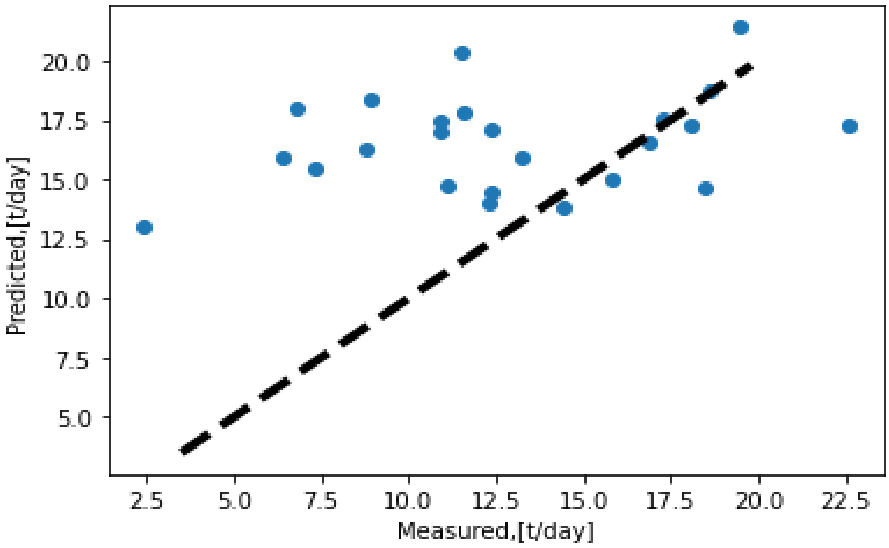


Figure 29- Production forecasting: Hydraulic Fracturing Wells - 2021

3.5 Summary of Chapter 3

This chapter is devoted to a comparative analysis of the algorithms used and explains the practical significance of this research work.

The identification of the most suitable and working algorithm for the data sample under consideration was carried out on the basis of various types of errors. The use of absolute mean error and mean square error does not provide the desired quality of the result. First of all, because this metric shows only the magnitude, but not the direction of the error. In addition, this metric on high-rate wells summarizes the increased error, in comparison with low-rate wells. The use of the coefficient of determination avoids the existing problems.

The ensemble model method is the most working algorithm applicable to the current problem, due to the fact that it has a minimum error value. This algorithm was used to identify the best candidates for hydraulic fracturing. And later, a comparative analysis was carried out with the wells where the hydraulic fracturing was carried out in 2021. This analysis proved the practical applicability of machine learning in the field of hydraulic fracturing.

4 ECONOMIC MODEL

Hydraulic fracturing is one of the most important discoveries in the field of energy in the last fifty years. This technology has significantly increased the volume of produced hydrocarbons. This growth has dramatically lowered energy prices, strengthened energy security, and even reduced air pollution and carbon dioxide emissions by replacing coal in power generation.

The result of the application of various technologies aimed at the extraction of hydrocarbons depends on many factors, both geological, physical, and chemical, as well as technological. As a rule, the more complex the process or technology of oil production, the more qualitative and quantitative parameters and properties should be taken into account when evaluating its effectiveness. One of the most complex and expensive technologies aimed at increasing the degree of oil recovery is hydraulic fracturing. To solve the problems of predicting hydraulic fracturing, to achieve sufficiently high technical and economic indicators of the effectiveness of measures, an objective assessment is necessary.

The economic justification of the proposed measures is necessary, because only on the basis of economic indicators, such as the indicator of the annual economic effect of hydraulic fracturing, the economic efficiency of capital investments can be judged on the economic efficiency of the proposed measures. The profitability index (PI) characterizes the economic return on investment and represents the ratio of the total net income to the total volume of capital investments, its value is interpreted as follows: if PI >1, the project is effective, if PI <1 – the project is not profitable.

$$NPV = \sum_{n=`1}^{n} \frac{R_n - OPEX}{(1+i)^n} - CAPEX \qquad \textbf{(9)}$$

where,

R- additional revenue from hydraulic fracturing;

OPEX- operating costs for additional oil production;

CAPEX – hydraulic fracturing costs;

i - discount rate.

In the framework of this study, the initial stage of the hydraulic fracturing operation was considered - the selection of a candidate well. The cost-effectiveness of hydraulic fracturing in this project is estimated using net discounted income. As mentioned earlier, the cost-effectiveness of the project is affected by a number of factors. Optimization of the fracture geometry based on its size and conductivity belongs to the stage of hydraulic fracturing design (Appendix D). Of course, this stage has a preferred impact on the economy of this technology, since it determines the main operating costs. In this section, the oil gain from the hydraulic fracturing operation is the main factor in the decision to conduct hydraulic fracturing.

Therefore, the economic feasibility of the project is based on the volume of additional production obtained by selecting the right candidate for hydraulic fracturing.

Table 13- Comparative analysis: an economic model

Case#	Additional production, 10^3 tons	Number of hydraulic fracturing operations, units	Average growth, t/day	Percentage of economic success by PI>1
Case#1	137	72	12.2	88
Case#2	235	141	8.5	77

In the framework of this study, a comparative analysis of the economic model of behavior with the use of the developed technology (Case#1, Table 13) and without (Case#2, Table 13) was carried out. The use of this technology not only allows to save time by performing accelerated analysis, but also significantly increase the company's profit. In the first case, the use of machine learning in the process of selecting wells for hydraulic fracturing allows not only to exclude negative candidates, but also not to leave "overboard" effective wells. The second option involves conducting hydraulic fracturing on positive and negative candidates. Thus, the average increase in oil production by 3.7 t/day of the first option exceeds the second one. At the same time, the planned indicators of cumulative production are achieved with a smaller number of hydraulic fracturing operations. Consequently, lower operating costs and higher profits due to additional oil production ensure the economic profitability of the project (Figure 30).

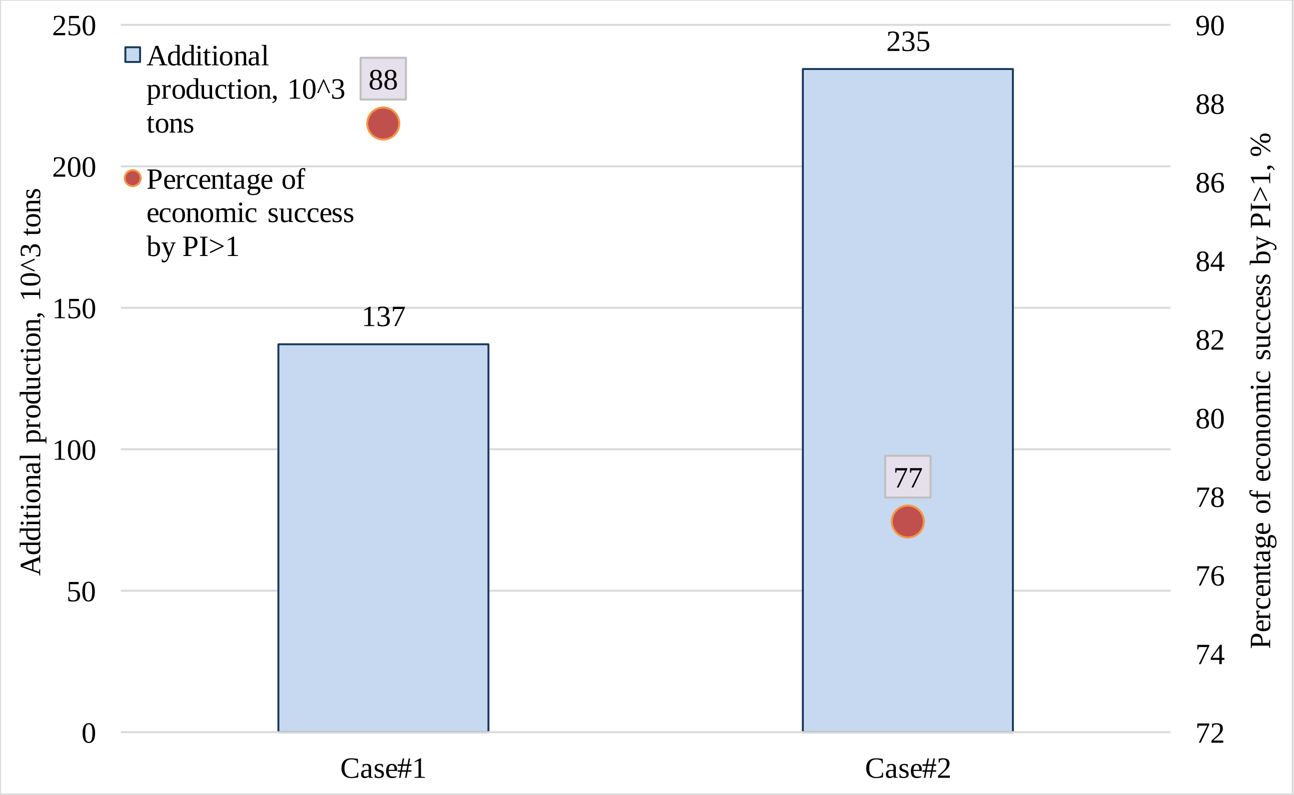


Figure 30- Economic justification of the project

CONCLUSION

In conclusion, it should be noted that the main goal of the project was to create a machine learning algorithm to identify the best candidate well for hydraulic fracturing. Development without applying the hydraulic fracturing methodology at the Uzen field is impossible. The question of improving and optimizing the method under consideration by correctly determining the impact candidate is relevant. Based on the literature review, a list of factors influencing the effectiveness of this methodology for increasing oil recovery was established.

Summing up the results of the implemented project, the selected algorithm model performs automated processing and analysis of a large data stream. Based on the oil production forecast, the best candidates for hydraulic fracturing are determined. The use of the developed technology allows not only to reduce operating costs by screening out negative candidates, but also not to miss out on the benefits of selecting the most efficient wells.

The weaknesses of this project are that in the framework of machine learning, it is impossible to avoid the presence of errors due to the phenomenon of "over fitting" data. However, this work has a deep potential for further research. The developed model is primarily dynamic. Increasing the number of influencing criteria will allow for a more detailed and in-depth research. In addition, changing the scale of the data sample makes it possible to significantly reduce the error value, thereby improving the forecast of oil production. As a result, the main task was completed.

LIST OF NOMENCLATURE

mD	Millidarcy (units of permeability)
m	Meter (unit of length)
atm	Atmosphere (unit of pressure)
P_i	Initial reservoir pressure
P_c	Current reservoir pressure
ρ	Pearson coefficient
σ	Standard deviation
β_n	Coefficient of a linear equation
λ	Regularization parameter
2D	Two-dimensional space
$\boldsymbol{R^2}$	Coefficient of determination
R	Additional revenue from hydraulic fracturing
i	Discount rate
$COV(X,Y)$	Covariance between two variables

LIST OF ABBREVIATIONS

HF	Hydraulic Fracturing
PI	Productivity Index
WC	Water Cut
IW	Injection Well
PW	Production Well
TBD	Information system «Geographically distributed data bank»
std	Standard deviation
min	Minimum sample value
max	Maximum sample value
LASSO	Least Absolute Shrinkage and Selection Operator
SSE	Sum of squared errors
PCA	Principal component analysis
MAE	Mean absolute error
RMSE	Root mean square error
NPV	Net present value

LIST OF REFERENCES

1. Akhmetov A., S. G. (2018). Operational Selection of Wells for Hydraulic Fracturing Treatment Through Machine Learning. (pp. 1-5). Saint Petersburg : European Association of Geoscientists & Engineers.

2. Alimkhanov R., S. I. (2014). Application of Data Mining Tools for Analysis and Prediction of Hydraulic Fracturing Efficiency for the BV8 Reservoir of the Povkh Oil Field . Moscow: SPE.

3. Aryanto Agus, S. K. (2017). Hydraulic fracturing candidate-well selection using artificial intelligence approach. Buku.

4. Behzad Mehrgini, H. M. (2014). Recognising the Effective Parameters and their Influence on Candidate-Well Selection for Hydraulic Fracturing Treatment by Decision Making Method. Kuala Lumpur: International Petroleum Technology Conference .

5. Boney, M. J. (2000). *Reservoir Stimulation in Petroleum Production.* ADVANTEK International.

6. GushchinA.E. (2015). *Methods of ensembling learning algorithms.* Moscow.

7. https://morioh.com/p/2d4243726fdb. Random Forest Regression., (p. Online).

8. https://proglib.io/p/linear-regression/. Linear Regression in Python., (p. Online).

9. Michael Economides, R. O. (2004). *Unified Fracture Design.* Alvin, Texas.

10. Mukku, V. (2019). Increase in Oil Production: Methodology & Best Practices for Hydraulic Fracturing Candidate Selection and Execution in Assam-Arakan Basin. Manama,: SPE.

11. Roshanai Heydarabadi, M. J. (2010). Criteria for Selecting a Candidate Well for Hydraulic Fracturing. Tinapa – Calabar: SPE.

12. S. Ai-Haddad. (1998). Hydraulic Fracturing in High Permeability Wells.

13. Stanton, J. (2013). *Introduction to data science* . Syracuse: Syracuse University.

14. Vanina A. S., P. V. (2020). Applying Machine Learning Methods to Exploit Unprofitable Reserves. Moscow: SPE.

15. Мелкумова Л.Э., Ш. С. (2017). Сравнение методов Ридж-регрессии и LASSO в задачах обработки данных. *Наука о данных* , 1748-1755.

16. (2017). *Общие требования к построению, изложению, оформлению и содержанию текстового и графического материала.* Алматы: СТ КазНИТУ – 09 – 2017.

17. Салимов О.В., Н. А. (2017). О КРИТЕРИЯХ ПОДБОРА СКВАЖИН ДЛЯ ГИДРОРАЗРЫВА ПЛАСТА. *Георесурсы* , 368-373.

APPENDICES

Appendix A. Database of training and test wells

New planned fracturing interval	Net pay thickness, m	P_c/P_i	Water cut, %	Oil rate, t/d	Permeability, mD	Injection well, m	Producing well, m	Achievement of an increase in production rate, %	Flow rate after HF, t/d	WC_group
0	16	1.17068	60	1.68	2.4	230	525	61.5	6.2	Medium-watered
1	16.9	1.07231	39	5.12	5.12	355	215	104	12	Low watered
0	13	1.08215	57	5.41	8.43	240	184	110.6	10.2	Medium-watered
0	11	1.13519	60	1.35	6.8	511	316	50	7.7	Medium-watered
1	8.5	0.73553	56.13	3.31	4.75	391	259	16	3.5	Medium-watered
0	14.5	1.20086	22.5	3.4	1.6	367	201	126	16.2	Low watered
0	15	1.17272	77.5	1.14	6.25	265	214	157	14.9	Medium-watered
0	19.6	1.23839	21.33	5.91	6.25	218	279	102	14.4	Low watered
0	8	1.05076	41.5	1.55	9.72	549	445	66	8.3	Medium-watered
0	12.6	0.96632	63	4.04	8.13	245	200	29	4.9	Medium-watered
1	24	0.86312	76.09	3.73	7.64	1079	246	120	12.5	Medium-watered
0	128	1.2196	40	2.52	18	521	185	99	13.5	Highly-watered

Continuation of Appendix A. Database of training and test wells

New planned fracturing interval	Net pay thickness, m	P_c/P_i	Water cut, %	Oil rate, t/d	Permeability, mD	Injection well, m	Producing well, m	Achievement of an increase in production rate, %	Flow rate after HF, t/d	WC_group
1	13.3	1.0789	60	3.36	9.62	181	177	104	10.5	Medium-watered
1	16	1.18052	32	8.56	12.6	197	222	41.7	14.9	Low watered
1	12.7	0.89127	15	85	17.1	1178	201	95	10.2	Low watered
0	8.4	1.08828	50	2.1	11.3	449	127	77	7.8	Medium-watered
0	9	1.03199	44	0.94	4.65	399	186	160	15.8	Medium-watered
1	14.2	1.37727	90.32	1.06	6.3	305	295	66	6.1	Highly-watered
0	12	0.94343	38	5.2	19.5	307	189	172.7	19.8	Low watered
0	10	1.32283	60	1.01	7	390	264	68	7.6	Medium-watered
0	10	0.98377	89.6	0.44	4.45	389	168	103.3	7.632	Highly-watered
1	10.1	1.13133	98.57	0.23	11.5	314	189	152.6	10.7	Highly-watered
1	25.4	1.01328	53.45	1.56	8.84	152	172	126	13.8	Medium-watered
0	13.9	0.85374	26	1.86	3.99	437	246	199	17.3	Low watered
1	20.6	1.32283	59.2	2.4	6.25	333	276	65	8.7	Medium-watered
0	11	1.0789	20	2.01	9.54	509	238	22	3.9	Low watered
0	17.8	1.03199	66	1.43	14.1	774	193	77	9.3	Medium-watered
0	11	0.8725	80.94	1.28	12	555	228	122	14.7	Highly-watered
0	9.6	1.0789	85.63	1.81	6.833333	1194	315	148	16.3	Highly-watered

Continuation of Appendix A. Database of training and test wells

New planned fracturing interval	Net pay thickness, m	P_c/P_i	Water cut, %	Oil rate, t/d	Permeability, mD	Injection well, m	Producing well, m	Achievement of an increase in production rate, %	Flow rate after HF, t/d	WC_group
0	14	1.03199	94	0.5	5.15	470	168	70	8.2	Highly-watered
1	15.9	1.14363	20	2.01	13.1	428	285	102	11.2	Low watered
0	18	1.03199	40	4.03	6.28	195	94	75	12.4	Highly-watered
1	14.5	0.93817	66.65	0.84	8	189	119	94	11.5	Medium-watered
0	11	0.99447	69.5	1.28	5	221	157	143	17.1	Medium-watered
0	10.6	0.96632	36	2.68	17.2	476	237	21	3.9	Low watered
0	11	1.01328	91.7	0.28	7.64	443	207	173.4	12.6	Highly-watered
1	13	1.08215	91.8	3.72	7.08	396	205	100.9	10.8	Highly-watered
0	9.1	1.03199	54	3.86	11.9	571	279	137.9	17.6	Medium-watered
0	20	0.84436	68.07	2.68	14.7	479	194	120.1	14.6	Medium-watered
0	10.5	1.24777	51.45	7.33	9.95	409	193	60.7	12.7	Medium-watered
0	6	1.06014	2	8.22	9.37	287	191	116.7	14.9	Low watered
0	17.2	1.03762	24	6.38	11.2	399	217	129.1	15.1	Low watered
1	10.3	1.00385	43	5.26	9.92	569	382	41.5	5.3	Medium-watered
1	18.6	1.1305	74.08	3.59	5.84	449	194	107.8	13.5	Medium-watered
0	16.5	0.97476	90.48	0.08	7.1	581	278	94.3	17.4	Highly-watered
0	19.8	1.03199	73	2.27	8.61	456	151	124.5	9.4	Medium-watered

Continuation of Appendix A. Database of training and test wells

New planned fracturing interval	Net pay thickness, m	P_c/P_i	Water cut, %	Oil rate, t/d	Permeability, mD	Injection well, m	Producing well, m	Achievement of an increase in production rate, %	Flow rate after HF, t/d	WC_group
0	11	1.0789	38	3.64	13.3	392	193	87.5	10.1	Low watered
0	10.1	1.00291	80.87	0.32	12.8	622	540	134.3	12	Highly-watered
0	29.5	0.89127	40	0.5	12.5	399	204	69.2	9.9	Highly-watered
0	6	1.10423	45	2.31	10.8	283	243	67.8	8.1	Medium-watered
1	19.5	1.14457	98	0.94	24.6	486	260	166.4	19	Highly-watered
0	12.1	1.17272	71	1.46	6.99	599	514	87.7	10	Medium-watered
1	7.7	1.15395	40	35.7	15.5	639	197	59.6	6.6	Highly-watered
0	12.6	0.84436	30	1.76	3	474	379	30.5	3.9	Low watered
0	20.1	1.18052	85.71	1.08	7.24	243	273	120.8	11.5	Highly-watered
0	14	1.32809	37	5.29	4.4	339	134	92.3	10	Low watered
1	16.5	1.00344	40	5.03	17	536	250	104.4	17.1	Highly-watered
0	17.6	1.2789	21.7	9.85	13.4	333	255	111	16.7	Low watered
1	14.9	0.98377	34	2.21	11.7	501	881	124	15.8	Low watered
0	11	0.9513	85.71	1.44	32	280	210	89	11.01	Highly-watered
0	5.9	0.98377	22.96	6.46	10.3	296	208	135	15.3	Low watered
0	9	1.08215	96.55	0.06	13.3	157	105	67	6.5	Highly-watered
0	20	1.13133	60.33	0.33	22.5	544	193	113	15.5	Medium-watered
1	13.1	1.13133	55	3.78	6.14	208	326	164	16.7	Medium-watered
1	11	1.18052	64.69	1.48	3.3	310	198	113	11.1	Medium-watered

Continuation of Appendix A. Database of training and test wells

New planned fracturing interval	Net pay thickness, m	P_c/P_i	Water cut, %	Oil rate, t/d	Permeability, mD	Injection well, m	Producing well, m	Achievement of an increase in production rate, %	Flow rate after HF, t/d	WC_group
1	11.3	1.25922	66.87	3.61	6.77	428	215	60	10.5	Medium-watered
0	19.4	1.06247	59.5	0.68	9.18	258	685	56	4.96	Medium-watered
0	15.9	0.93458	40	4.03	12.5	443	505	75	10.1	Highly-watered
1	15.7	1.03296	88	0.3	8.61	396	196	133	13.5	Highly-watered
0	11	0.82637	70.29	0.5	4.25	301	271	91	11.9	Medium-watered
0	16.1	1.04279	67.7	0.81	8	536	246	47	5.8	Medium-watered
0	17.5	1.18052	92.6	0.68	5.71	224	339	143	13.9	Highly-watered
0	17.1	1.16085	56.57	1.82	9.98	344	179	84.3	9.9	Medium-watered
1	8	1.03296	100	0	15	607	252	119.5	12.6	Highly-watered
1	16.2	1.08215	86	1.29	10.8	455	250	54.5	6.9	Highly-watered
1	13.7	0.88539	46.27	0.9	13.3	404	381	115.6	14.3	Medium-watered
1	18	0.86572	68.89	3.92	15.6	367	220	114.8	14.5	Medium-watered
1	11	1.18052	100	0	7.4	416	304	74.8	8.3	Highly-watered
1	12.7	1.23955	81	0.32	5.5	418	300	154.6	17.9	Highly-watered
0	20.8	1.33792	60.36	2	10	374	586	45.4	6.3	Medium-watered
1	10.8	1.03296	52	2.42	9.7	570	204	56	6.7	Medium-watered

Continuation of Appendix A. Database of training and test wells

New planned fracturing interval	Net pay thickness, m	P_c/P_i	Water cut, %	Oil rate, t/d	Permeability, mD	Injection well, m	Producing well, m	Achievement of an increase in production rate, %	Flow rate after HF, t/d	WC_group
1	8.3	0.9149	97	0.23	12.5	359	301	127	13.7	Highly-watered
1	7.5	1.09198	44	0.47	8.29	224	275	77.8	9.4	Medium-watered
1	19.5	0.79685	71	1.22	11.4	450	349	25.8	6.8	Medium-watered
1	12	1.1215	100	0	9.7	496	199	97.8	11.3	Highly-watered
0	10.2	1.18052	53.5	7.41	9	265	135	113.1	17.2	Medium-watered
1	9.8	1.16085	42	3.89	16.2	398	274	89	9.9	Medium-watered
1	12.5	1.03296	38	2.6	8.8	440	632	131.6	12.1	Low watered
0	19.9	1.67241	70	4.28	33.8	526	229	151.1	18.8	Medium-watered
1	8	1.03296	44	1.98	9.68	248	211	119.9	14	Medium-watered
1	23	0.90507	10	14.4	11.1	408	228	126.1	19.5	Low watered
1	13.8	0.99361	65.2	1.75	15.1	330	287	97.8	9.4	Medium-watered
0	16.2	0.96409	61	7.2	10	318	191	85.1	12	Medium-watered
1	11.5	1.03296	63.7	1.52	10.6	606	199	71.8	8.1	Medium-watered
1	19.5	0.90507	65	4.4	14.1	284	307	20.4	5.7	Medium-watered

Appendix B. Data: collection, preparation, quality check, analysis

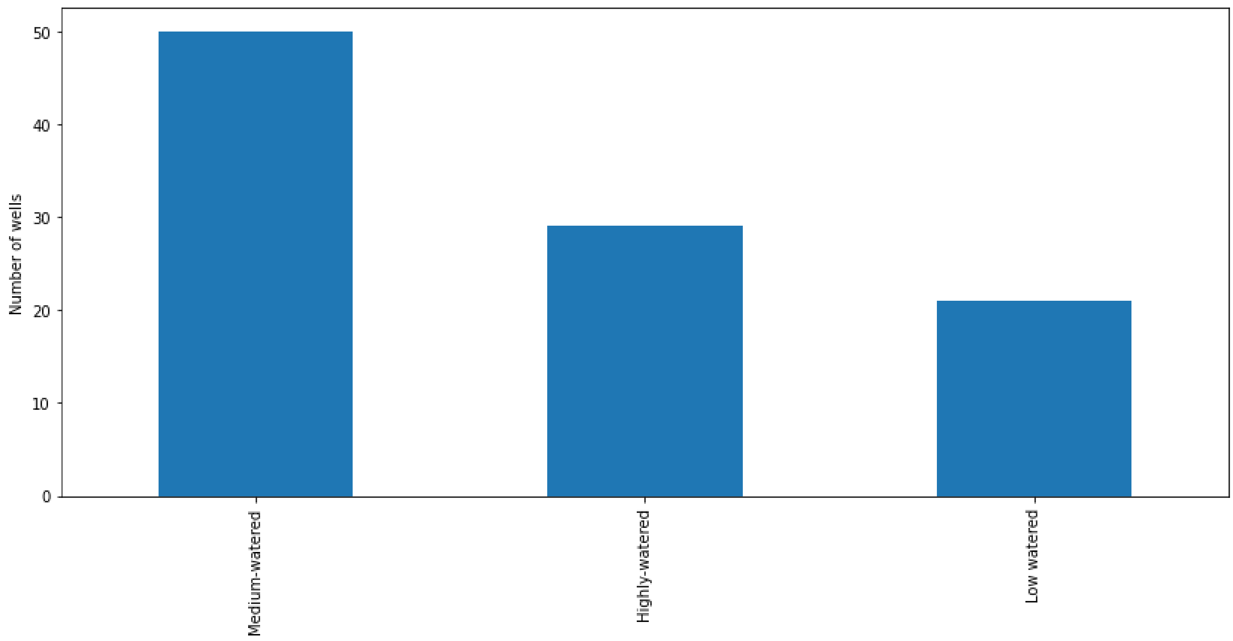


Figure B.1- Data sampling: cclassification of wells by water cut parameter

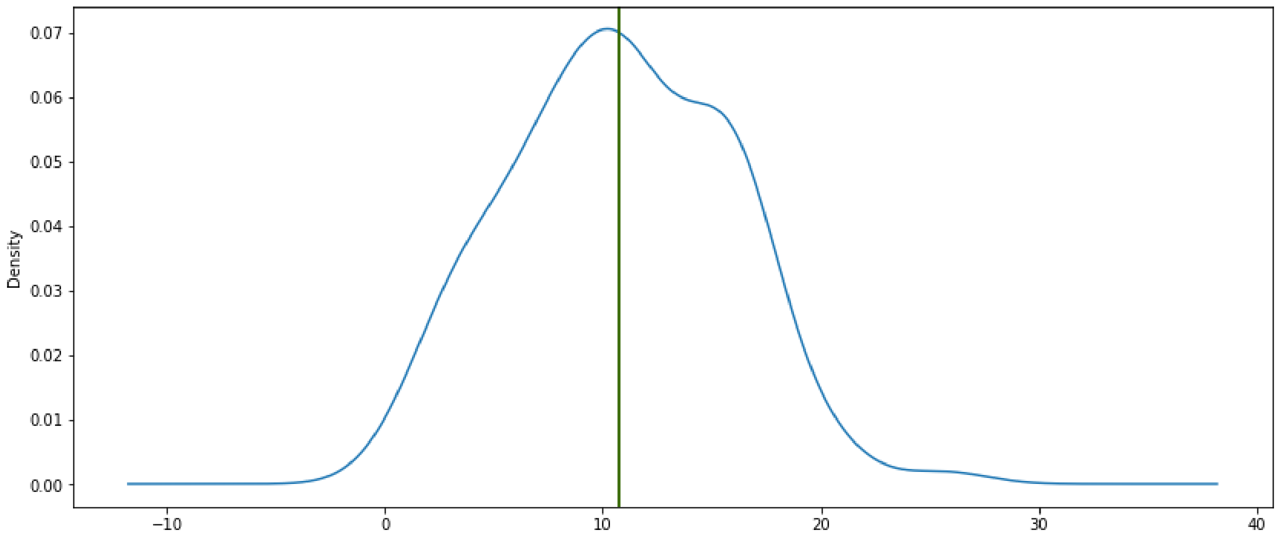


Figure B.2- Data sampling: Average oil flow rate after hydraulic fracturing, [t/d]

Continuation of Appendix B. Data: collection, preparation, quality check, analysis

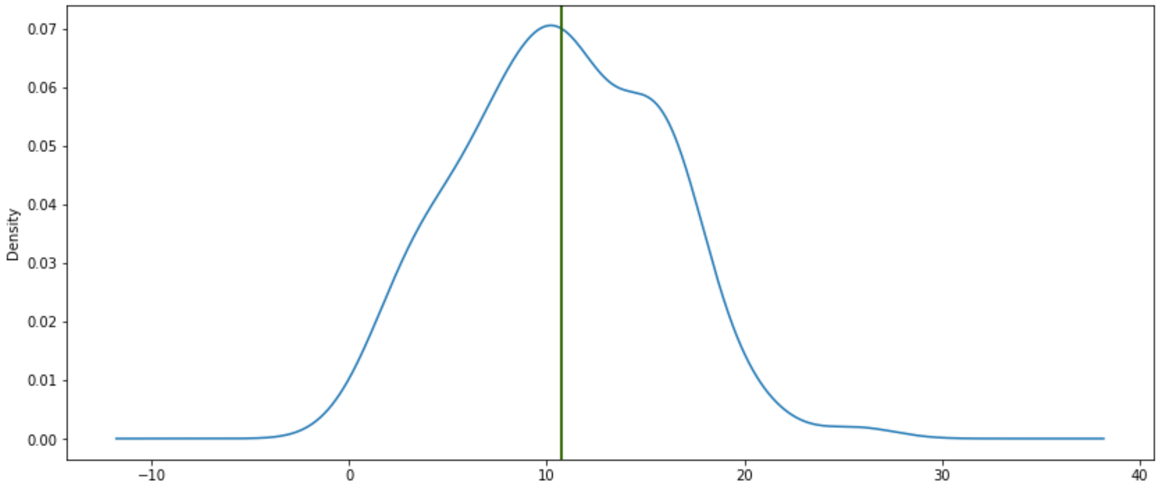


Figure B.3- 2020 HF: Average oil flow rate after hydraulic fracturing, [t/d]

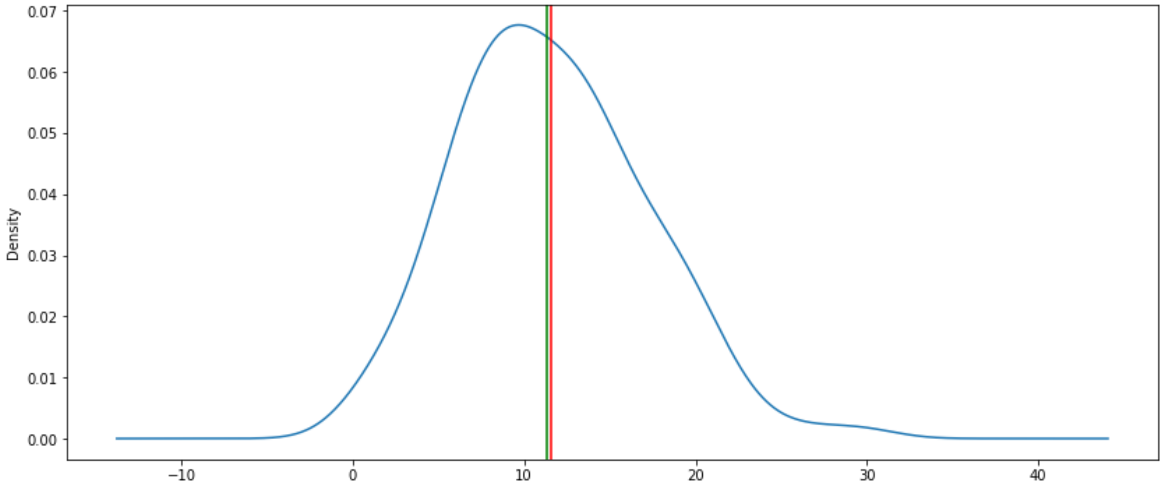


Figure B.4- 2019 HF: Average oil flow rate after hydraulic fracturing, [t/d]

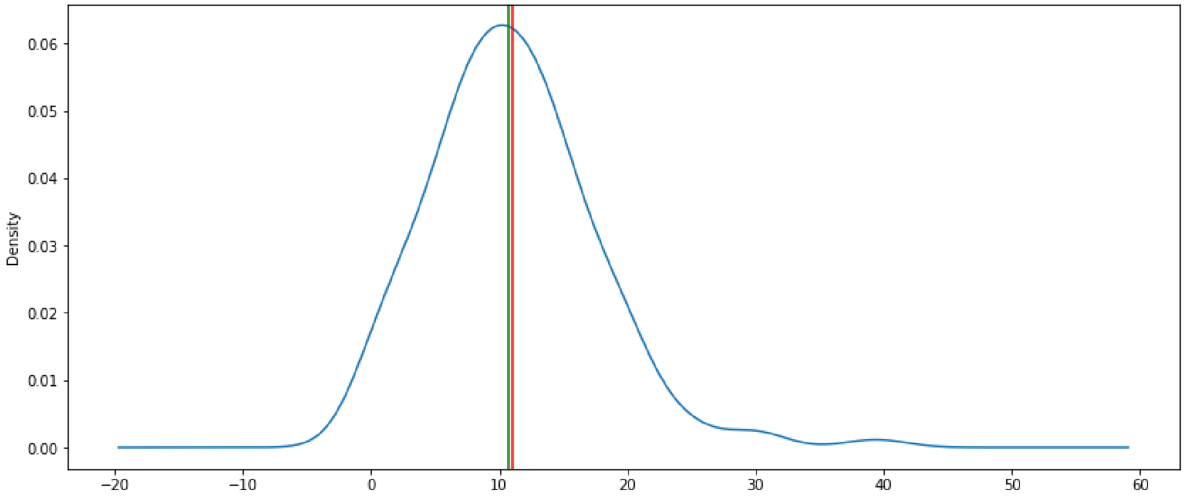


Figure B.5- 2018 HF: Average oil flow rate after hydraulic fracturing, [t/d]

Continuation of Appendix B. Data: collection, preparation, quality check, analysis

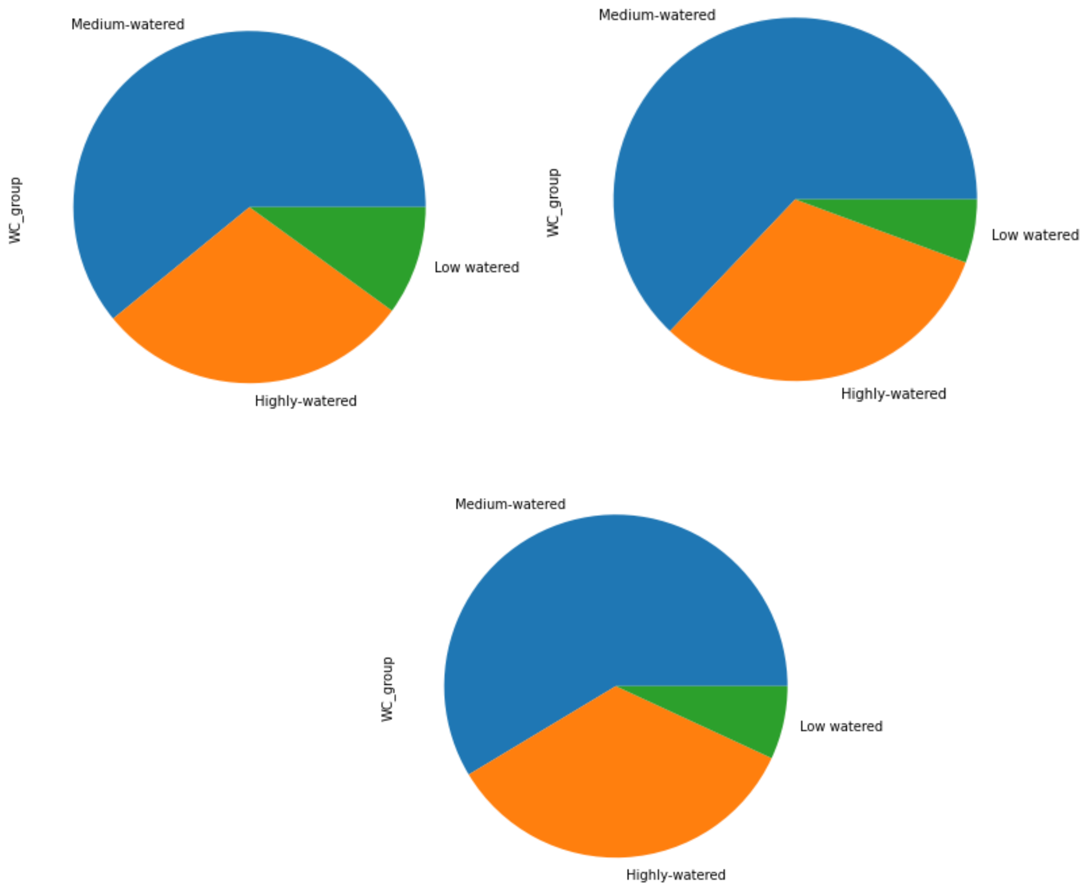


Figure B.6- 2020/19/18 HF: cclassification of wells by water cut parameter

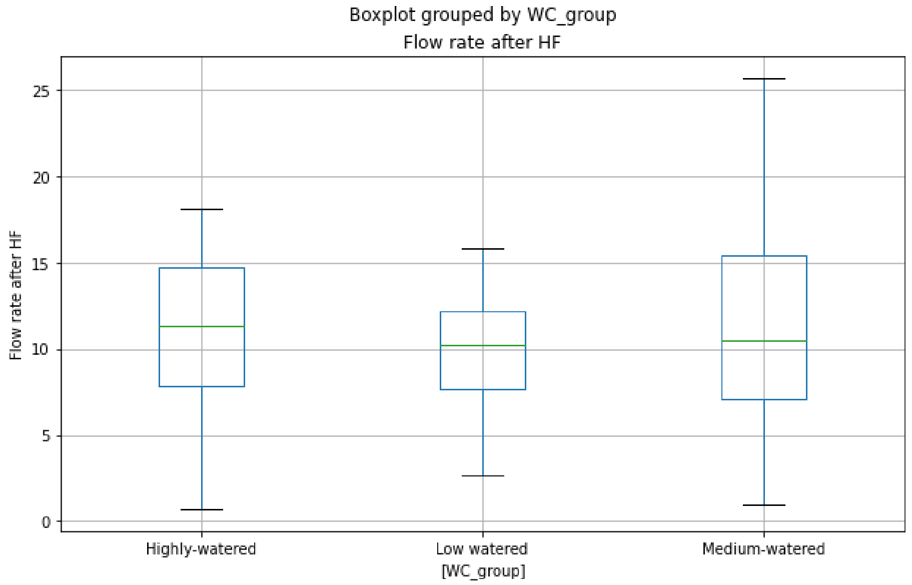


Figure B.7- 2020: Production statistics by WC group

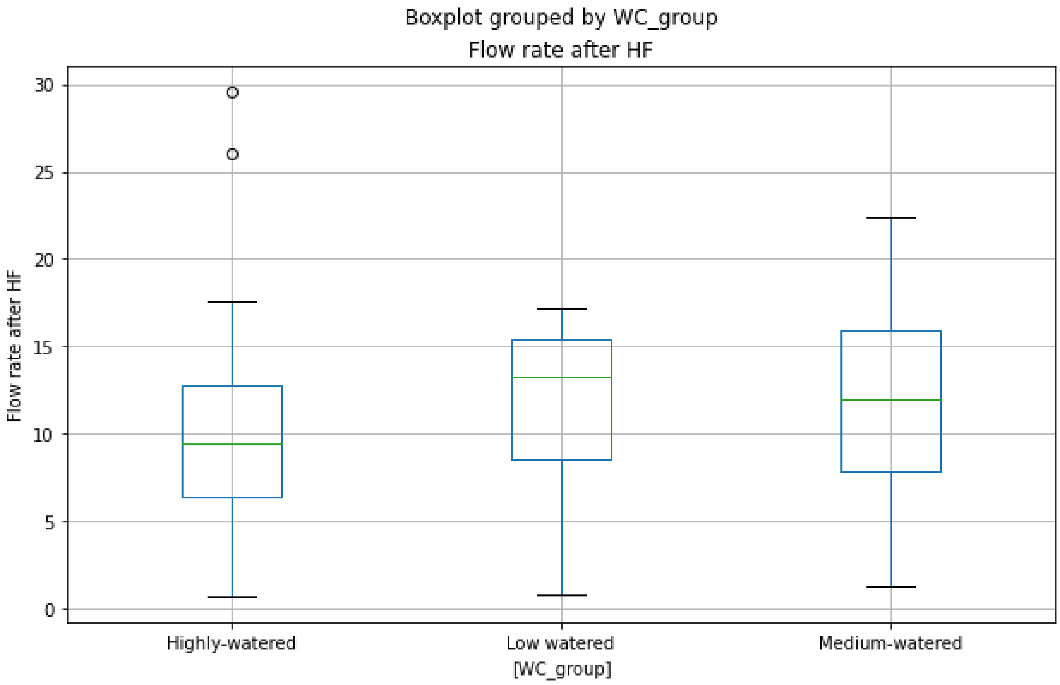


Figure B.8- 2019: Production statistics by WC group

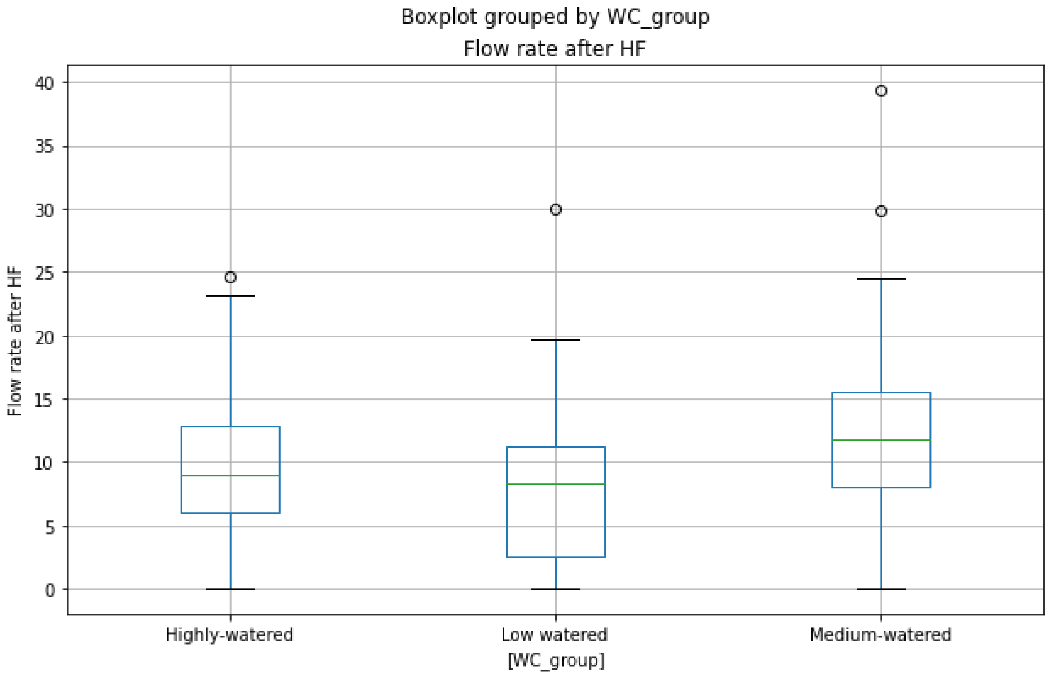


Figure B.9- 2018: Production statistics by WC group

Appendix C. Machine learning: output data

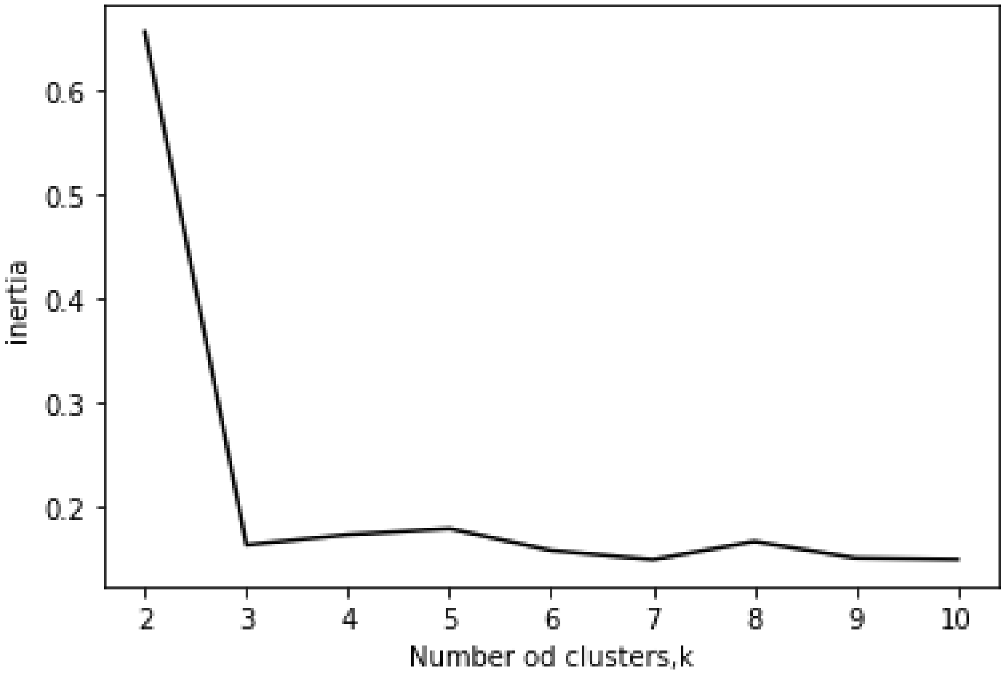


Figure C.1- Determination of the elbow point

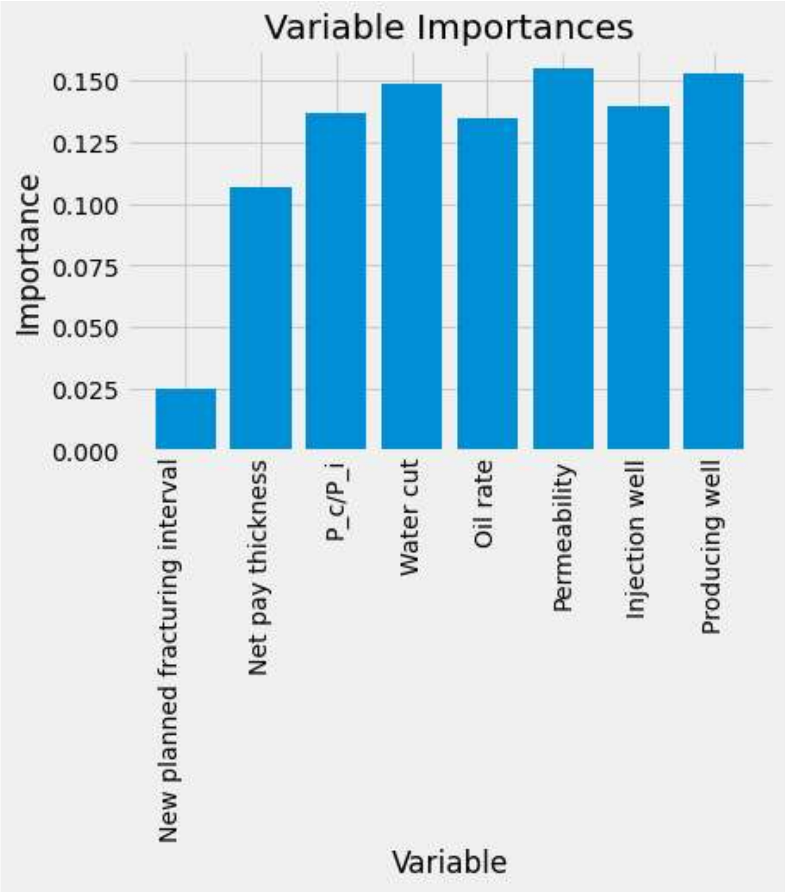
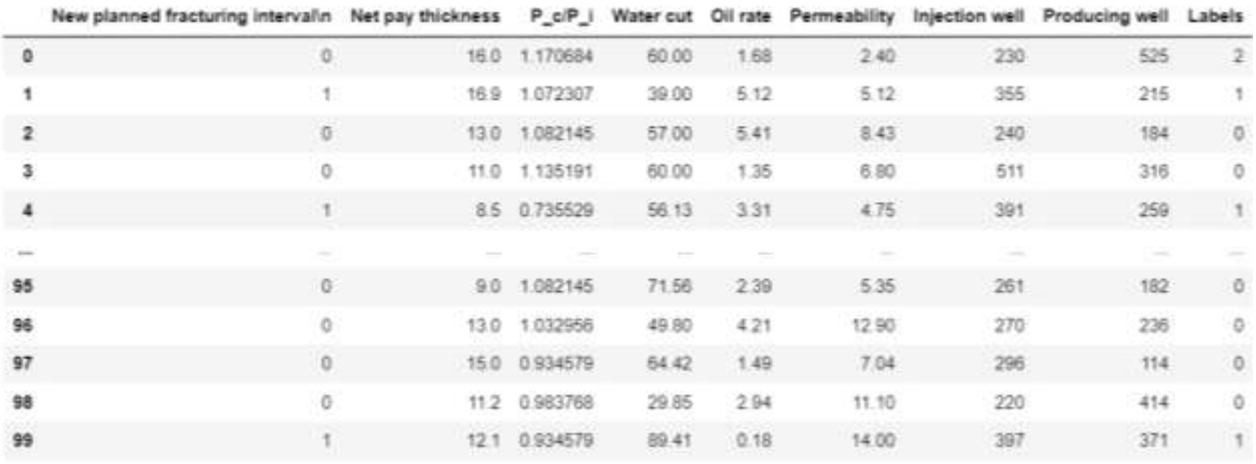


Figure C.2- Influence of parameters on well performance forecasting

Continuation of Appendix C. Machine learning: output data



	New planned fracturing interval\n	Net pay thickness	P_c/P_i	Water cut	Oil rate	Permeability	Injection well	Producing well	Labels
0	0	16.0	1.170684	60.00	1.68	2.40	230	525	2
1	1	16.9	1.072307	39.00	5.12	5.12	355	215	1
2	0	13.0	1.082145	57.00	5.41	8.43	240	184	0
3	0	11.0	1.135191	60.00	1.35	6.80	511	316	0
4	1	8.5	0.735529	56.13	3.31	4.75	391	259	1
...	...	...	...	...	...	...	...	...	...
95	0	9.0	1.082145	71.56	2.39	5.35	261	182	0
96	0	13.0	1.032956	49.80	4.21	12.90	270	236	0
97	0	15.0	0.934579	64.42	1.49	7.04	296	114	0
98	0	11.2	0.983768	29.85	2.94	11.10	220	414	0
99	1	12.1	0.934579	89.41	0.18	14.00	397	371	1

Figure C.3- Example of K-Mean well classification

Appendix D. Forecasting production on potential candidates for hydraulic fracturing

Table D.1- Database of potential hydraulic fracturing candidates

New planned fracturing interval	Net pay thickness	P_c/P_i	Water cut	Oil rate	Permeability	Injection well	Producing well
1	14	0.99	86	0.4	13	674	236
1	12.7	1.05	79	0.3	5.2	374	204
1	9.2	0.99	81	1.3	12.1	283	199
1	16.2	0.85	77	2.1	6	166	222
0	13	1.18	58	2.9	8.9	164	195
0	12.9	1.14	62	3.5	11.3	231	415
0	15.8	1.01	74	1.9	7.2	321	74
0	18.5	1.16	76	2.6	6.6	172	454
1	5.5	1.06	62	2.7	11.5	297	233
0	12.7	1.13	62	4.4	19.7	514	288
0	10	0.88	64	3	8.9	171	472
1	8.7	1.04	95	0.8	12.6	182	224
0	9	1.09	92	1.9	17.6	328	145
1	13.9	0.81	97	0.5	14.5	414	273
0	8.1	1.31	75	1.8	11.5	387	181
1	6.7	1.28	76	2.2	12	251	283
0	10	1.08	83	1.8	4.9	80	222
0	26.3	0.93	75	0.8	3.2	182	373
1	13.1	1.46	61	3	8	353	333
1	20.1	0.94	86	2.3	10.2	252	231
0	10.4	0.81	54	1.8	6.5	263	208
0	8.3	0.93	95	1.2	13.7	374	162
1	32	1.13	78	3.7	11	819	282
0	20.7	1.11	54	3.7	9.6	248	318

Continuation of Appendix D. Forecasting production on potential candidates for hydraulic fracturing

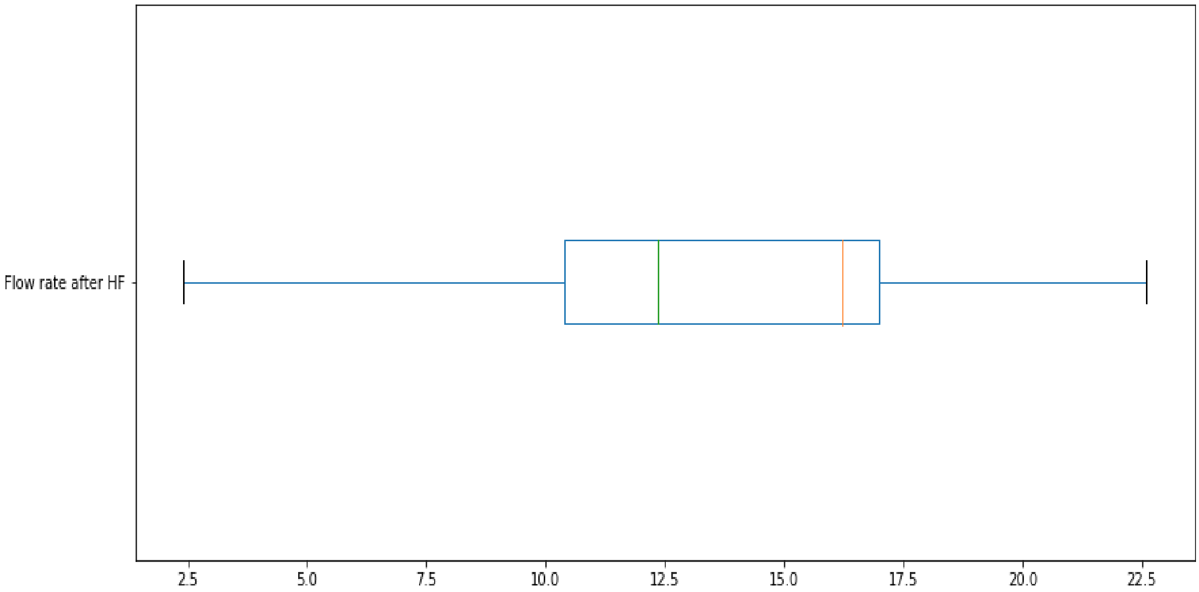
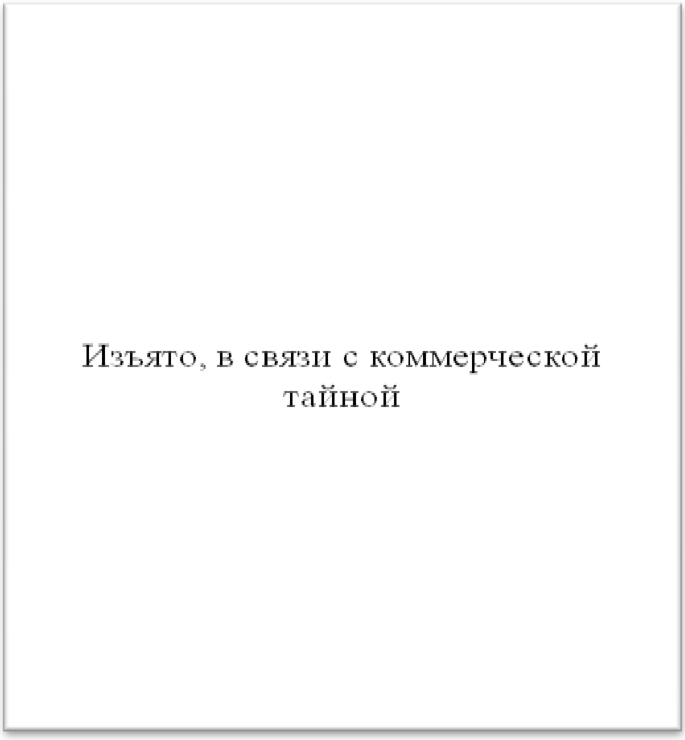


Figure D.1- Comparison of the average forecasted (red line) and actual (green line) oil production rates

Appendix E. Economic justification

Изъято, в связи с коммерческой тайной

Figure E.1- Dependence of the cost of hydraulic fracturing on the fracture length



Изъято, в связи с коммерческой тайной

Figure E.2- Discounted return of investments