energies

MDPI

Article

Unveiling the Feasibility of Coalbed Methane Production Adjustment in Area L through Native Data Reproduction Technology: A Study

Qifan Chang [1], Likun Fan [2], Lihui Zheng [1,*], Xumin Yang [1], Yun Fu [3], Zixuan Kan [4] and Xiaoqing Pan [5]

[1] College of Petroleum Engineering, China University of Petroleum (Beijing), Beijing 102249, China; 2020310178@student.cup.edu.cn (Q.C.); 2021215226@student.cup.edu.cn (X.Y.)
[2] Changqing Oilfield Company, China National Petroleum Corporation, Xi'an 710018, China; flk_cq@petrochina.com.cn
[3] College of Safety and Ocean Engineering, China University of Petroleum (Beijing), Beijing 102249, China; 2020011318@student.cup.edu.cn
[4] College of Information Engineering, Beijing Institute of Petrochemical Technology, Beijing 102617, China; 2021310889@bipt.edu.cn
[5] Beijing LihuiLab Energy Technology Co., Ltd., Beijing 102200, China; asset@lihuilab.cn
* Correspondence: zhenglihui@cup.edu.cn

Abstract: In the L Area, big data techniques are employed to manage the principal controlling factors of coalbed methane (CBM) production, thereby regulating single-well output. Nonetheless, conventional data cleansing and the use of arbitrary thresholds may result in an overemphasis on certain controlling factors, compromising the design and feasibility of optimization schemes. This study introduces a novel approach that leverages raw data without data cleaning and eschews artificial threshold setting for controlling factor identification. The methodology supplements previously overlooked controlling factors, proposing a more pragmatic CBM production adjustment scheme. In addition to the initial five controlling factors, this approach incorporates three additional ones, namely, dynamic fluid level state, drainage velocity, and fracturing displacement. This study presents a practical application case study of the proposed approach, demonstrating its ability to reduce reservoir damage during the coal fracturing process and enhance output through seal adjustments. Utilizing the full spectrum of original data and minimizing human intervention thresholds enriches the information available for model training, thereby facilitating the development of a more efficacious model.

Keywords: coalbed recover; yield optimization scheme; raw data; coalbed methane mining; native data reproduction technology

Citation: Chang, Q.; Fan, L.; Zheng, L.; Yang, X.; Fu, Y.; Kan, Z.; Pan, X. Unveiling the Feasibility of Coalbed Methane Production Adjustment in Area L through Native Data Reproduction Technology: A Study. *Energies* **2023**, *16*, 5709. https://doi.org/10.3390/en16155709

Academic Editor: Shu Tao

Received: 3 June 2023
Revised: 18 July 2023
Accepted: 25 July 2023
Published: 31 July 2023

1. Introduction

In recent years, the utilization of big data has gained prominence as the volume of data generated and documented within the oil and gas industry has experienced a significant upsurge. The versatility of big data has exerted a substantial impact on the oil and gas sector [1]. Applications of big data in this industry include stratum identification [2], drilling process optimization [3], post-development production forecasting [4], and production–sales optimization [5]. Compared to traditional empirical formulas and theoretical models, big data methodologies exhibit enhanced accuracy when leveraging extensive data [6]. Concurrently, big data techniques also demonstrate superiority compared to certain finite element methods. For instance, Xiao et al. [7] employed a global optimization framework facilitated by machine learning to ascertain optimal fracturing parameters for shale gas, which necessitated extensive physical simulations. Consequently, investigating the applications of big data across various scenarios within the petroleum industry is

valuable, particularly in terms of production forecasting and optimization solutions. These applications ultimately underscore the commercial value of big data technology.

The emergence of big data technology has reduced economic costs to some extent; along with this emergence, however, there is an increasing concern over the application of specific oil problems. The lack of unified planning and deployment in the initial stage of oilfield construction leads to data becoming an "information island," which cannot meet the intelligent application of data [8]. Johnston and Guichard analyzed approximately 350 oil and gas Wells in the UK North Sea using drilling data, well logs, and geological formation tops in their study of the use of big data to reduce the risks associated with drilling operations. They reported that two of the most difficult steps in a petroleum engineering project are data collection and processing [9]. Anand explained well how big data reveal a lot of hidden information from the vast amount of data in the oil and gas industry. If a limited amount of data are used, the results will reveal limited patterns, may lack overall insight, and may carry a great deal of uncertainty [10]. However, if a large data set is available and used with more sophisticated tools, more promising patterns can be identified, which may be closer to the true value [1]. Data are the foundation of machine learning or related big data approaches to petroleum engineering problems. How to use data better is more important.

Because raw data cannot be directly used in calculations for the following reasons [11], it is a common practice to carry out two parts of work before data use: data cleaning and screening of main control factors.

- Some algorithms require data to be numerical.
- Some algorithms impose requirements on data. Errors and statistical noise existing in the gathered data should be corrected.
- Complex nonlinear relationships can not be too common.

Data cleaning methods are widely used in petroleum engineering problems for different purposes.

The data cleaning method used by Fabrice Hollender in an earlier study to remove artifacts can significantly improve tomography resolution [12].

Kyungbook Lee used data cleansing to remove zeros from native data during her study of shale gas production [13]. In the continuous production process, due to manual operation, "0" value points will be generated due to the closing of the well, and such values are not naturally generated in the production process. When the computer learns the above data curve, if there is no separate reminder to the computer that this special "0" value point is artificial, it will cause the computer to produce errors in machine learning. This is especially important during long production cycles. Yin et al. found in their study of drill string dynamics that raw measurements may seem noisy, and removing noise/outliers can eliminate drilling/process anomalies. After data cleaning, the optimization from two to six kinds of warning is enabled [14]. Yu et al. found that data cleaning and clustering were critical for improving the performance in all models. However, it was also found that it was necessary to adopt median filtering for input logs to alleviate the aliasing problem caused by data interpolation and eliminate outliers [15].

Cleaning data will homogenize data, which can be easily calculated using a computer [16]. After data cleansing, several sets of data are obtained. A single set of data of the same type is often referred to as a factor or a feature. A feature or a factor can be defined as a single measurable property of the dataset that is being analyzed [17]. In the actual research process, the redundancy of factors is caused by the inability to judge the correlation between factors. Some factors are related to the research goal, and some factors are not related to the research goal or even a kind of interference. Too many unrelated factors participating in the calculation would make it prone to failure.

Therefore, the method of feature selection is proposed to select existing factors rather than directly using all factors [18]. A commonly used approach for some researchers is analyzing the problem and screening the existing factors based on theoretical knowledge. However, such a method is too subjective to discover new horizons, and other permutations

and combinations of features are created by comparing the differences between a single factor and a target. Determine whether the selected combination is consistent with the goal, and if not, try other combinations again until a good match appears [19]. This approach can traverse all permutations when the number is relatively small. Many research studies in the 1990s focused on modeling with less than 40 features, but this has since expanded into the hundreds to tens of thousands of features [20]. It is impossible to screen so many factors by means of manual screening. Then researchers gradually replaced the manual selection process by setting evaluation indicators for comparison.

Based on this approach, a large number of machine learning and statistical analysis metrics are proposed to screen for the right combination of factors [21]. Common statistical indicators, such as the Pearson correlation coefficient and mutual information, are used to evaluate differences between factors and targets. Then, with the application of principal component analysis (PCA), the information contribution of factors is also applied. These methods are easy to scale to different datasets because researchers only need to artificially select an indicator and set a threshold based on their own experience to achieve the screening of factors [22]. Earlier, G. Stewart, in 1989, used artificial intelligence methods for screening of main control factors and extraction of well-test interpretations [23]. Principal component analysis (PCA) has been used for modeling and classification of reservoir properties [24].

The above methods proved that data cleaning and main control factors screening facilitate the use of data in the algorithm model. However, researchers also found that processes such as data cleaning, artificial data removal, and interpolation will lead to the loss of some extremely important information. While cleaning parts of the data can remove noise when studying identified problems, more often, it is impossible to be sure that what is being cleaned has an effect. Shut-in processes such as the one mentioned above can result in zero daily production, so data are thrown out, but the state of the well during shut-in can easily affect subsequent production. The simmering effect of the well closure time may increase the yield after reproduction. There are a large number of "irresolution areas" in the middle depending on whether data are cleansed or not. This situation is more obvious in setting thresholds to filter the controlling factors. Screening the main control factors is accomplished by artificially setting thresholds. In this process, the correlation coefficient and contribution degree can be used to judge the complete correlation when the correlation coefficient is 1 (contribution degree 100%) and the complete non-correlation when the correlation coefficient is 0 (contribution degree 0%). However, when the correlation coefficient is 0.8 and 0.7 (contribution degree 80% and 70%), it is impossible to judge whether the factor is the master factor. This vacillating approach can lead to unconventional results in actual engineering applications.

An example is the single well production in area L which controls the main factors of coalbed methane production. A well-adjusted production using conventional methods considering five main control factor models found that the pre-adjusted single well production increased by 6.6%. The drilling fluid Marsh viscosity and permeability were increased to 150 s and 180 mD, respectively. First, the Marsh viscosity of drilling fluid is too high to flow, making it hard to carry solid phase. Second, the permeability of the coal reservoir is unlikely to be so high. Using the above steps, due to the human screening factors, may lead to the omission of the main control factors, resulting in the weight of factors used for the calculation being increased. If using the main control factors to optimize the construction design reverse calculation, the scope of control often exceeds the actual conditions, thus making the scheme not feasible.

Another reason for the infeasibility of the proposed solution is that the optimization exceeds the applicable range. During the optimization process, the model's performance is influenced by the consistency between data samples and test objectives. A dilemma exists between the model's generalizability and accuracy. In one scenario, when studying a specific region, the model is trained on data from wells within that region. To achieve higher accuracy during the training process, overfitting may occur. While this model

performs well on wells of the same type, the actual wells under consideration might differ from all the wells in the region, necessitating a model with good generalizability that can accommodate these varying well conditions. In another scenario, when using the model to optimize a specific target well, the target may have already exceeded the data range for that region, rendering the optimization results infeasible. Thus, the model presents issues during both its development and application phases. For example, in region L, increasing production from 300 m^3 to 320 m^3 is consistent with the current regional data characteristics, but raising it to 500 m^3 may exceed the model's application range. Directly using the model for calculation would result in unrealistic and unfeasible outcomes.

To address the information loss caused by data cleaning and the manual setting of thresholds, as well as the issue of exceeding the model's application range during implementation, this study introduces a new approach called native data reproduction technology (NDRT), a method that retains all raw data without data cleansing, and the algorithm of big-data cocooning (ABDC), using an elimination method for screening master factors without artificial threshold setting is applied. This method is applied to region L and compared with other methods that fail to form a feasible control scheme after feature removal. Based on this, a feasible coal bed methane (CBM) production optimization scheme is proposed. When the actual application case surpasses the scope of the model's applicability, the optimization scheme is proposed by analyzing the commonalities within the existing schemes provided by the model.

2. Methods

In an effort to preserve a greater extent of information and minimize human interference, the non-destructive raw data transformation (NDRT) method is employed. To ensure seamless computer utilization of all information, this study prioritizes the "retention of native data" as its core principle, enabling the computation of non-numeric data through the implementation of "one-hot encoding" and "tokenization" techniques. To further mitigate human interference, various big data model sets and the "big-data cocooning algorithm" are utilized for principal control factor screening. Moreover, to facilitate the execution of practical construction beyond the model's application scope, a method for selecting superior construction schemes is proposed.

2.1. Enhancing Data Utilization through Native Data Processing

Data in this study exhibit a lack of consistency, with significant amounts of missing, abnormal, and non-numerical data. In order to retain as much information as possible, compared with the traditional method of data cleaning, the traditional method deletes abnormal data and interpolates a small amount of missing information. In this method, these anomalous data are retained in their original form, while missing data are retained as "missing" information instead of being populated with interpolation. Non-numerical data were processed using follow-up methods to make them suitable for inclusion in the analysis.

One-hot encoding, also known as one-bit valid coding, involves the use of N-bit status registers to encode N states, each represented using a separate register bit. This method is particularly useful for handling discontinuous numerical features and can be easily implemented without the need for decoders. Additionally, one-hot encoding expands the feature space to some extent. Non-numerical data may include both correlated and uncorrelated data. For instance, conventional rock types may be represented in text form in these data.

A token is a way to convert a paragraph or several related words into a data structure. For text information, time information, and other information to be understood by human beings, these data cannot be used directly as a number in the computer for calculation. One-hot provides a way to convert uncorrelated text into numerical data. The token complements the conversion of other information into sequence data.

2.2. Model Selection Based on Data Characteristics

Different models have different fit degrees on different data sets due to different data distribution characteristics, although it is possible to optimize the model by adjusting hyperparameters to achieve good results. However, the artificial selection of a certain model will also lead to a mismatch in research objectives. Therefore, various models should be selected as far as possible to compare the accuracy of data in the model so that these data can use their own characteristics to choose a more suitable model. So, this study is a collection of common big data models. The package includes the Random Forest model [25], CatBoost (short for "Category Boosting") [26], Weighted Ensemble model [27], XGBoost (short for " Extreme Gradient Boosting") [28], ExtraTrees (short for "Extra Randomized Trees") [29], K-Nearest Neighbors (KNN) [30], NeuralNetFastAi (Neural network come from FastAi) and NeuralNetTorch (Neural network come from pytorch) [31], KNeighborsUnif is a machine learning model that combines the K-Nearest Neighbors (KNN) algorithm with the Resilient Propagation (RPROP) algorithm.

2.3. Identifying Main Control Factors Using the Algorithm of Big-Data Cocooning

The algorithm of big-data cocooning (ABDC) is a method for identifying the main control factors in a dataset without the need for setting artificial thresholds. This approach begins with an initial dataset containing all relevant metadata and establishes a research target factor and other models. The model is then retrained after removing one factor, and the rankings of the remaining factors are compared between the new and old models to determine the impact of the removed factor. If there is no change in the rankings of the remaining factors, the removed factor is considered unimportant and has a weak effect on model generation. However, if the rankings of other factors change after the removal of a factor, it is considered to be an important influencing factor that cannot be replaced by other factors.

This method begins with an initial dataset containing all metadata, establishes the target factor and other models, and retrains the model after removing one factor at a time. Each square in Figure 1 represents a feature. In the first round of optimization, all features are included in the model training. The yellow squares represent features that, when removed, would affect the ranking of other factors and are therefore retained. The red squares represent features that, when removed, do not affect the ranking of other factors and are therefore removed in the next round. After all the features have been evaluated, several non-important red factors are screened out. The process is then repeated with the remaining features until all the factors are important yellow features. This method was proposed by Professor Zheng and has obtained a good application effect in the evaluation method of coalbed methane permeability [32].

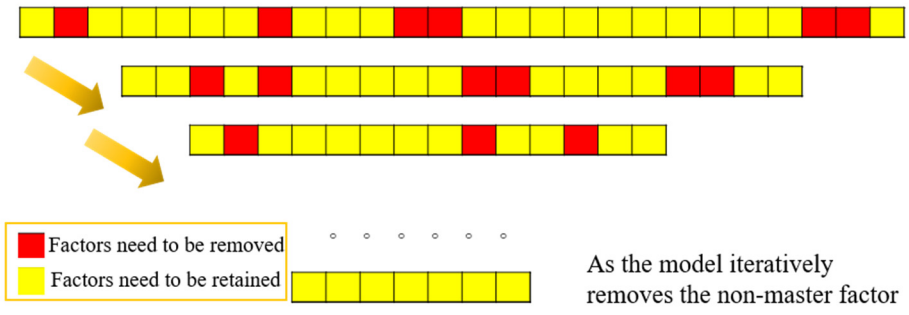

Figure 1. Process diagram of factor selection by ABDC method.

The purpose of this part is to compare the native information method with the ABDC method with the traditional method in terms of handling real problems in the L region. The background and composition of the dataset for this region are provided, and a case study analysis is conducted to examine the differences in handling problems using different methods. The necessary parameter setting details for each method are also discussed.

2.4. Developing Practical Construction Plans within Model's Applicability Scope

In order to achieve a higher degree of production optimization, the direct inversion of optimization solutions using models may lead to exceeding the applicable scope of the model. However, within the appropriate application range, models can provide effective construction solutions. Analyzing the commonalities among these solutions may assist engineers in understanding which aspects the model seeks to adjust. Combining this analysis with engineers' experience enables the proposition of feasible solutions for construction needs that surpass the model's application scope.

Given the extant production wells, one may establish an artificial production improvement standard and utilize this model to investigate feasible schemes. The commonalities of feasibility across these schemes are studied, and the fundamental reasons for the feasibility of each scheme are analyzed. These essential reasons are then employed to design construction schemes:

1. Artificially set a higher yield target than the current plan.
2. Utilize the existing model to design a feasible solution.
3. Analyze the commonality of all proposed solutions.
4. Propose an optimization plan to increase yield based on the identified commonalities.

3. Discussion

The purpose of this section is to discuss the capabilities of information retention and utilization using the ABDC and traditional approaches.

First, it is discussed how much information the method in this study retains from native data and judge its value by comparing the correlation coefficient of missing data.

Secondly, through the comparison of the models, it is judged whether these data and models have improved the accuracy of the model and helped to find the main control factors.

Then the difference between the results of the model in this study and the traditional models are compared, and the infeasible results of the traditional model are explained.

Finally, through a comprehensive analysis of various model schemes, a method for practical application beyond the range of model use is proposed.

3.1. Native Data Can Retain More Useful Information

According to the missing information statistics of the original data, there are 230 missing data points in the total 2208 data points, accounting for about 10% of missing data. In order to show these missing data more intuitively, these missing data in each set of data are listed. To focus more intuitively on the data itself, all features are represented by Arabic numeral numbers. The specific feature number and the name in the actual project are shown in the Appendix A Table A1. The main reasons for missing data include text information that cannot be directly calculated using conventional methods and data that are not measured and recorded in the construction process, as shown in Figure 2 below.

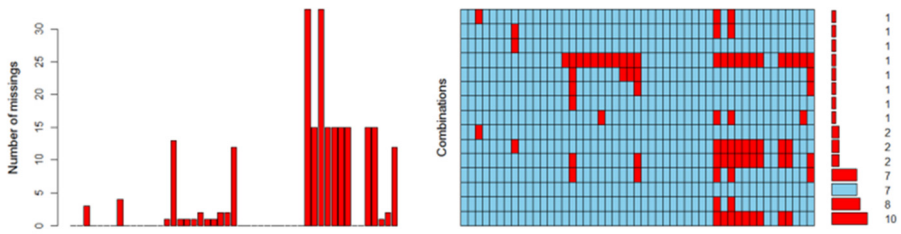

Figure 2. Missing value distribution map.

On the left, the horizontal axis represents factors with missing values, and the vertical axis represents the number of missing values. The red box on the right represents missing data, the horizontal axis represents factors that contain missing values, and the vertical axis represents the position of missing values in the data box.

The right side of Figure 2 shows that each small square represents a data point, with red representing missing data and blue representing numeric data. Data on the left include information about well type, number of perforations, and oil pressure, while data on the right include geological parameters such as rock type, ash content, and location. It can be seen that there is a higher prevalence of missing data among geological parameters, which may be due in part to the fact that these parameters are often expressed in words rather than numeric values.

Through the conversion of text information, the missing statistics of raw data are shown as follows in Figure 3:

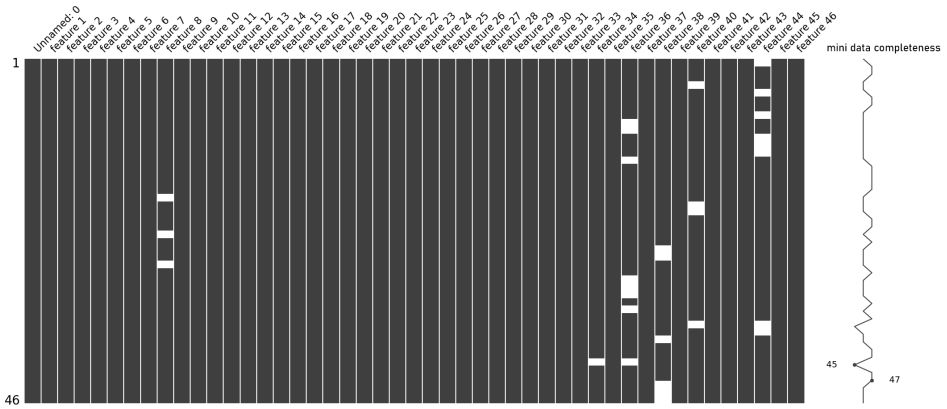

Figure 3. Map of missing value distribution after converting text information.

Figure 3 illustrates that the white squares represent missing data, while the black squares represent calculated values. It can be seen that the amount of missing data has significantly decreased compared with Figure 2, but there are still some unrecorded null values. In this approach, these null values are considered to have information value. To prove that these missing data contain information, the correlation between missing data is calculated, and a thermal map of the missing value correlation is generated.

As shown in Figure 4, there is no digital mark box indicating that there is no correlation between the two missing values, such as Feature 8 and Feature 34 and Feature 34 and Feature 40. At the same time, there is a correlation coefficient of 0.3 between Feature 34 and Feature 36. In these data, the correlation coefficient between drilling pressure and drill bit speed is only 0.28. However, they are usually considered to have an influencing relationship [33]. If such information is easily deleted, a large amount of information will be omitted. At the same time, the traditional method of using correlation to collect and select data will also lead to the exclusion of a large amount of data.

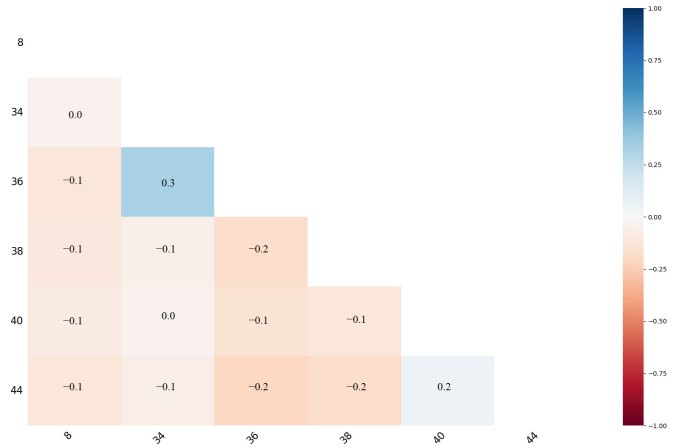

Figure 4. Missing value correlation coefficient heat map.

3.2. Native Data Supplemented the Missing Master Factors

After training using original data in the ensemble model, in order to further evaluate the superiority of the proposed method, an ablation experiment will be conducted, and the results will be compared with other techniques for processing the main controlling factors in region L. Conventional data preprocessing techniques will be applied to these data and used as inputs for the ensemble model, as well as for the existing methods such as support vector machine (SVM), linear regression, and neural network models in region L. These models, numbered 1–9 in this study, correspond to the ensemble models proposed in this method, while the models numbered 10–12 correspond to the existing techniques in region L. The results of each model's training and test data are shown in Table 1:

Table 1. Model error score comparison table.

No.	Model Name	Methods in This Study		Method from Past	
		Training Data Score	Test Data Score	Training Data Score	Test Data Score
1	Random Forest	87.366	139.021	108.614	165.435
2	CatBoost	87.473	164.701	150.630	181.171
3	Weighted Ensemble	90.254	78.353	134.009	93.240
4	XGBoost	99.226	82.740	146.457	90.187
5	ExtraTrees	111.233	197.085	184.390	226.648
6	KNeighbors	113.435	242.088	171.763	271.138
7	NeuralNetFastAI	164.736	173.494	227.435	221.518
8	NeuralNetTorch	184.916	210.433	301.783	252.520
9	KNeighborsUnif	188.183	248.578	246.708	278.407
10	Linear regression			324.699	270.715
11	SVM			252.544	210.556
12	NeuralNet			227.290	189.500

As shown in Table 1, the error scores of nine different models in this study on training and test data are listed. The lower the error score, the higher the accuracy of the model's application effect on these data. When the training data score is low, but the test data error score is high, it indicates that the model is likely overfitting these data. For a more visual comparison, the overfitting case is analyzed using the difference subtracted from training data and test data, as shown in Figure 5.

As Figure 5a shows, the first three models attain about 90 on the training data, among which RandomForest has the best performance. However, this model and CatBoost have a low score on test data and a big difference with the score of training data, so they are not suitable for the scenario where the consistency between test and training data is weak. The overfitting line shows the WeightedEnsemble model, and the NeurlNetFastAI model is low. Although a low overfitting value does not mean that the model is good, a high value indicates that the performance of the model in training and test data is unbalanced. Moreover, the WeightedEnsemble model will perform better on training and test data.

As shown in Figure 5b, these data are cleaned before entering the model training. Among them, the model represented by histograms 1–9 is consistent with the model used in Figure 5a, and the two graphs use the same scale, so it can be seen that after data cleaning, the error scores of training and test data of a large number of models increase. Therefore, it can be considered that using native data is helpful for the model.

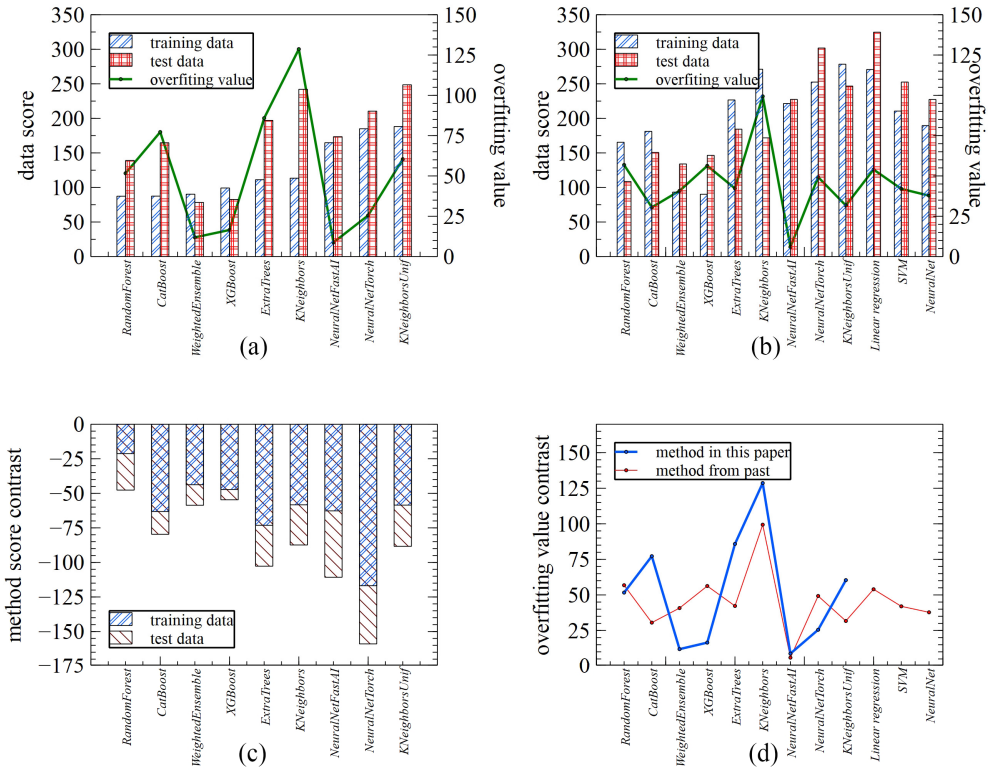

Figure 5. Model score and overfitting comparison chart ((**a**) This study's method error scores the data histogram and overfitting value line chart; (**b**) Other method data error score histograms and overfitting value line charts; (**c**) Stacked histogram of the difference between the scores of this method and other methods; (**d**) This method overfitting values compared to other methods against line charts.).

Figure 5c shows the difference between the error score of this study's method and the score of other methods. The error score after data cleaning is used to obtain worse results, especially in B, the neural network model is more sensitive to data, the score result is the worst, and the analysis is due to the reduction in data information after data cleaning. Compared with other methods already applied in the L region in Figure 5b, such as SVM, linear regression, and neural network, which have been carefully designed and adjusted by researchers, the neural network model is slightly better than the model in the ensemble model, but the score is still only 227.29, which is lower than other ensemble models. In addition, in Figure 5d, the overfitting value fluctuation ranges of the two methods are compared. The previous method has a similar degree of overfitting to the method in this study, but due to the reduction in data information after data cleaning, the range of overfitting fluctuations is reduced, and data cleaning removes unbalanced data.

The results of the analysis encompass a range of algorithms, some of which are designed to process data while others are not. Algorithms such as Cat Boost, Weight Ensemble Method, and Extra Trees possess the capability to effectively handle data quality and demonstrate improved accuracy when applied to original data that are rich in information. In the context of oil field systems, data acquisition is an expensive endeavor, particularly when key decision makers, such as experts in the field, have already applied their knowledge to determine the relevance of data. As such, utilizing raw data in the research process pertaining to oil engineering problems holds significant value.

According to the model calculation, under the existing five main controlling factors, namely permeability, drilling fluid Marsh viscosity, total fracturing fluid volume, reservoir pressure, and coal roof height, three factors, namely the dynamic fluid level state, drainage velocity, and fracturing displacement, which were not mentioned in the existing research literature, were added.

3.3. Native Data Increase the Feasibility of the Scheme

The optimized scheme obtained using this model is compared with that obtained by the conventional model. The relationship between Marsh viscosity, permeability, and production obtained using conventional models is shown in Figure 6.

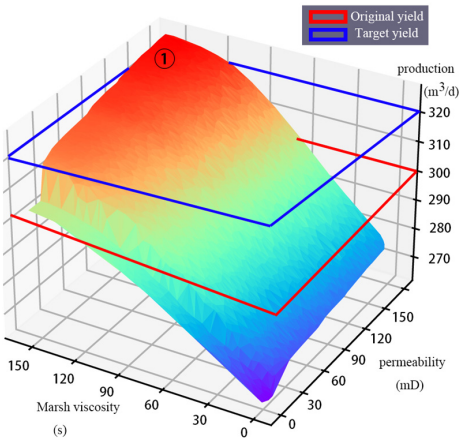

Figure 6. The optimization scheme diagram obtained from the conventional model (① in the figure is the schematic point that meets the target yield. The red box line is when the output is 300, and the blue box line is when the output is 320, with ① position higher than the blue box line).

The original production of this well was 300 m^3 (in the red box line), but only Condition ① was satisfied when the target production was adjusted to 320 m^3 (in the blue box line) in Figure 6. The regulation scheme with conventional methods can only obtain the rule of higher permeability, higher Marsh viscosity of drilling fluid, and higher production, ignoring the influence of other factors, and the Marsh viscosity needs to be adjusted to 150 s, which exceeds the actual construction conditions.

Upon supplementing the missing factors with these original data, the relationship between the main control factors and the yield is shown in Figure 7. Conditions ①, ②, and ③ are all met when the target yield is adjusted. When the permeability is adjusted by 20 mD, the Marsh viscosity is adjusted by 30 s, the drainage rate is adjusted to 12 m^3/d, and the fracturing displacement is 6.3 m^3/h, which meets the engineering construction conditions and makes the scheme feasible.

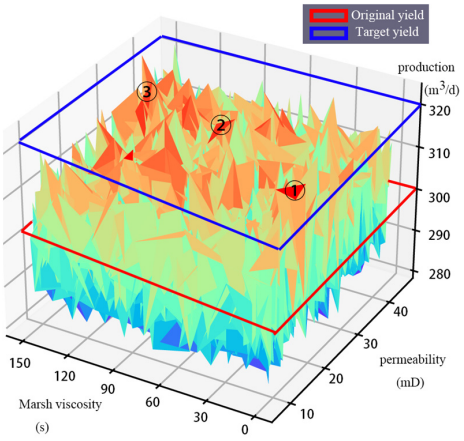

Figure 7. The optimization scheme diagram obtained in this study. (①, ②, ③ in the figure are the schematic points that meet the target yield).

3.4. Plugging Leaks While Reducing Reservoir Damage Is Key to Extraction

The model has been used to calculate the possible scenarios above, and this section analyzes the commonality of these scenarios. The engineering design parameters in the two scenarios are listed in Table 2, and the range of engineering parameters for the L region is provided.

Table 2. Scheme comparison table.

Process	L Region Process Range	Scheme for This Thesis	Feasibility	Scheme for Past	Feasibility
permeability	2–30 mD	20 mD	√	180 mD	×
Marsh viscosity	15–60 s	30 s	√	150 s	×
drainage rate	0–50 m^3/d	12 m^3/d	√	37.3 m^3/d	√
fracturing displacement	4–10 m^3/h	6.3 m^3/h	√	8.1 m^3/h	√

√ It's feasible; × It's not feasible.

The engineering design parameters in the two scenarios are listed in Table 2, and the range of engineering parameters for the L region is provided. As shown in Table 2, the permeability and Marsh viscosity in the original scheme are too large and are not suitable for actual operation. The original production of the well was 300 cubic meters, and an increase in production to 320 cubic meters, or 6%, was desired. In the original protocol, the permeability and Marsh viscosity were increased. This requires huge economic investment and technology research and development for oilfield companies. In this scenario, the drainage rate is adjusted to 32% of the original parameter to drain water at a slower rate while reducing the displacement during fracturing [34].

The adjustment of the above scheme focuses on the fluid characteristics and its corresponding process technology, and it is simpler to regulate by changing the fluid parameters than other methods. Further combined with the literature, analyze how these adjustments to the fluid affect the final yield. One possibility is that reservoir damage is reduced in this way.

Alum et al. developed a mathematical model to show the effect of drilling performance on drilling speed. Their research showed that by maximizing the viscosity of plastics, the permeation rate can be reduced [35].

Wang et al. increased the likelihood of damage to formation using viscous fracturing fluids because the formed filter cake can clog the pore throat of low-permeability rocks [36]. Galindo et al. studied the effect of increasing viscosity on formation damage [37]. The increasing viscosity also increases the risk of high-viscosity fluids staying in the formation, so it also requires the original formation to have a higher porosity, which is conducive to the flow back of the fluid.

Combined with the above scheme, it is proposed that the fluid with the ability to reduce reservoir damage and increase the fluid with deplugging capacity is proposed to achieve the operation, and it is recommended to select a fluid that meets the local operating conditions for this scheme optimization.

4. Field Application

4.1. Basic Information of the Well

The Z-3X well, located in the village of Zhongxiang, Zhengzhuang Town, Qingshui County, Shanxi Province, is a vertical well used for coalbed methane development in the Zhengzhuang block of the Qinnan Jincheng slope belt. The well was drilled to a depth of 710 m, with an artificial well bottom at 702 m, and was completed with casing perforation. The production layer is the #3 coal seam of the Shanxi Formation, which is 4.1 m thick. The roof is 7.2 m thick, containing sandy mudstone, and the floor is 11.8 m thick, also containing sandy mudstone. The perforation section is between 648.7 and 653.2 m, 4.5 m thick. The coal seam has a desorption pressure of 2.2 MPa, a gas content of 23 m^3/t, and a porosity

of 4.3%. The storage conditions of the coalbed methane in the reservoir are good, with strong adsorption/desorption capacity, high gas content, and excellent potential for gas production.

The well has undergone preliminary development, including fracturing and other work, but the production is insufficient. Therefore, it is beneficial for this method to optimize the design of the well.

4.2. Method of Calculation in This Study

To apply this method, the well's data will be incorporated into the model to predict current well production and assess the model's accuracy. We aim to increase current production by 10% and evaluate the feasibility of this plan.

The Z-3X well underwent hydrofracturing in the perforation section and went into production test mining six months later. Gas was observed after 403 days of production, with a peak daily gas output of approximately 310 m^3/d and a peak daily water output of approximately 17 m^3/d. After approximately 690 days of production, the well stopped producing gas, with an average daily water output of about 8 m^3/d and an average daily gas output of about 150 m^3/d. The total gas production was 40,580 m^3, and the total water output was approximately 7750 m^3. The prediction result of the model was an average daily gas production of 161 m^3/d, with an error of 7%, indicating that the model is accurate within the error range.

Feasible schemes recommended by the model are shown in Table 3:

Table 3. The scheme provided by the model of this study.

Plan	Total Volume of Fracturing Fluid (m^3)	Viscosity of Fracturing Fluid	Liquid Level Drop (m)	Drainage Speed (m^3/d)
1	668	87	900	8.6
2	734	80	800	4.3
3	1013	73	900	7.2

Similar to the above results, the purpose of the model's proposed scheme is to minimize reservoir damage during the fracturing process. It aims to enhance the proppant suspension effect of the fracturing fluid, reduce friction resistance, make it easier to flow backward and attempt to reduce water production in the coal seam and lower the resistance to gas production through slow, long-term drainage. Thus, in the application scheme, it is recommended to use a fluid consistent with this scheme, protective of the reservoir, and ideally possesses water-blocking functionality. For this purpose, the literature on materials used in local or similar areas was investigated, and the material deplugging capacity and reservoir protection performance were compared. It is recommended to use a Fuzzy-ball to prevent and plug leaks and reduce formation damage.

4.3. Detailed Construction Process

In practical applications, a Fuzzy-ball is selected for second stimulation fracturing of wells to reduce reservoir damage. The target layer is the #3 coal seam of the Shanxi Formation, with the working well section from 648.70 to 653.20 m. On-site, a cycle is established using a sand mixer and water tank, and materials are added through a cutting funnel. Following the lab formula, in two 50 m^3 mixing tanks, 40 m^3 of clean water, 0.75 t of capsule agent, 0.50 t of fluffy agent, 0.10 t of capsule core agent, and 0.20 t of capsule film agent are sequentially added. The prepared Fuzzy-ball temporary plugging fluid has a density of 0.93 g/cm^3, apparent viscosity of 50 mPa·s, plastic viscosity of 25 mPa·s, shear stress of 28 Pa, and a pH value of 8.

Temporary plugging phase: First, 2.0 to 3.0 m^3/min of active water is discharged to replace 12 m^3 of fluid, testing and opening the original cracks. Then, 2.0 to 3.0 m^3/min of fluffy capsule temporary plugging fluid is discharged to plug 25 m^3 of the water layer. Next,

2.5 to 3.5 m^3/min of Fuzzy-ball temporary plugging fluid is discharged to temporarily plug 52 m^3 of the original cracks. Lastly, 2.0 to 3.0 m^3/min of fluid is discharged to replace 13 m^3 of fluid, pushing the Fuzzy-ball temporary plugging fluid into the water layer and deep into the cracks, enhancing the plugging strength.

After the injection of the Fuzzy-ball temporary plugging fluid is completed, to judge the temporary plugging conditions of the original cracks and water layers, the pump is stopped for 12 min to observe the change in pressure drop. When the pressure is basically stable, or the pressure drop rate is low, it is considered that the Fuzzy-ball temporary plugging fluid has successfully sealed the water layer and the original cracks.

The Fuzzy-ball temporary plugging fluid is compatible with active water fracturing fluid, requiring no isolation fluid, and can proceed directly to fracturing. This scheme uses a total volume of 561 m^3 of fracturing fluid, with the viscosity of the fracturing fluid within the model's calculated range.

4.4. Apply Effects

Compared to the first development of the Z-3X well, gas was observed after 403 days of production, with a peak daily gas output of approximately 310 m^3/d. After approximately 690 days of production, the well stopped producing gas. After this construction, gas production began 245 days after the Z-3X well was fractured, and the gas output gradually increased, indicating that the coal seam was depressurized and the coalbed methane began to desorb. The gas output was 494 m^3/d 355 days after fracturing, and the gas output was still increasing, maintaining gas production. This indicates that the Fuzzy-ball fluid did not damage the gas production ability of the coal reservoir, reduced reservoir damage, and the construction scheme was effective.

5. Conclusions

In this study, we discuss the feasibility of an optimization scheme for coalbed methane (CBM) single-well production, utilizing raw data without data cleaning and applying a main control factor screening method based on the elimination method.

Through our investigation, we discovered that uncleaned raw data retain about 10% more information than data post-cleaning. Correlation analysis indicated that even missing data could maintain a relationship with other information; thus, preserving more data potentially allows for a more accurate representation of the situation. Moreover, upon employing the elimination method, we identified the dynamic fluid level state, drainage velocity, and fracturing displacement, which were previously unfeasible with older schemes. This approach provides a viable optimization plan for CBM wells. Furthermore, in conjunction with literature studies, we adjusted our strategy to mitigate reservoir damage caused by high-viscosity fracturing fluids commonly used in local CBM hydraulic fracturing developments. Our method also introduces a novel approach for practical applications outside the range of typical model use. By focusing on fluid characteristics and their corresponding process technologies, we determined that regulating fluid parameters is more efficient than utilizing other methods. Thus, we propose the use of a specialized fluid designed to reduce reservoir damage and increase deplugging capabilities tailored to local operational conditions.

However, we note that researchers may need to allocate additional time for studying the processing of text information, as training in natural language text can often be more time-consuming. The text information used in this study involves less data and technology than typical natural language processing projects. Future research should aim to integrate a larger array of drilling reports, fracturing reports, production reports, and expert guidance into the big data model to further enhance its effectiveness.

In conclusion, our study demonstrated that utilizing raw data in oil yield analysis can result in better information retention, identification of new controlling factors, and more feasible extraction strategies. We propose that the ABDC method could serve as a valuable

tool for petroleum engineers to optimize their extraction strategies and reduce operational costs.

Author Contributions: Writing—review and editing, project administration, Q.C.; resources, data curation, X.Y.; visualization, L.Z.; resources, L.F.; software, data curation Z.K.; writing—original draft preparation, validation, Y.F. field experiment, X.P. All authors have read and agreed to the published version of the manuscript.

Funding: The authors thank the Ministry of Science and Technology of the People's Republic of China (2016ZX05066) for financial support.

Data Availability Statement: This study did not report any data.

Acknowledgments: The authors are very grateful to China United Coalbed Methane Co., Ltd. Changqing Oilfield Company for providing data and technical support. Thanks to the CBM Exploration and Development Division for providing oilfield application opportunities.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Table A1. The feature number corresponds to the corresponding name in the actual project.

Feature	The Name of the Process	Feature	The Name of the Process	Feature	The Name of the Process
1	Drilling of Well Period (d)	17	Displacement Fluid Volume (m^3)	33	Average Discharge Volume(m^3/d)
2	Total Footage (m)	18	Total Fluid Volume (m^3)	34	Reservoir Pressure (Mpa)
3	Coal Seam Diameter (m)	19	Sand Volume (m^3)	35	Coal Seam Depth (m)
4	Well Type	20	Avg. Sand Ratio (%)	36	Reservoir Temperature (°C)
5	Drilling Fluid Type	21	Oil Pressure (MPa)	37	Coal Seam Thickness (m)
6	Drilling Fluid Viscosity (s)	22	Displacement (m^3/h)	38	Gas Content (m^3/t)
7	Mechanical Drilling Speed (m/h)	23	Pump Depth (m)	39	Effective Porosity (∅)
8	Drilling Pressure (kN)	24	Initial Dynamic Liquid Level (m)	40	Ash Content (%)
9	Rotation Speed (r/min)	25	Initial Bottomhole Flowing Pressure (MPa)	41	Reservoir Layer Group 1
10	Cement Volume (m^3)	26	Gas Appearance Time (d)	42	Reservoir Layer Group 2
11	Avg. Cement Slurry Density (g/cm^3)	27	Dynamic Liquid Level at Gas Appearance (m)	43	Permeability (mD)
12	Volume of Fluid Used to Replace Drilling Fluid (m^3)	28	Casing Pressure at Gas Appearance (MPa)	44	Critical Desorption Pressure (MPa)
13	Perforation Thickness (h)	29	Bottomhole Flowing Pressure at Gas Appearance (MPa)	45	Fracturing Pressure (MPa)
14	Hole Count	30	Cumulative Water Production at Gas Appearance (m^3)	46	Shut-in Pressure (MPa)
15	Pre-fluid Volume (m^3)	31	Drop in Liquid Level (m)	47	Coal Seam Roof Height (m)
16	Carrying Sand Fluid Volume (m^3)	32	Water Discharge Rate (m^3/d)		

References

1. Mohammadpoor, M.; Torabi, F. Big Data analytics in oil and gas industry: An emerging trend. *Petroleum* **2020**, *6*, 321–328. [CrossRef]
2. Popa, A.S.; Grijalva, E.; Cassidy, S.; Medel, J.; Cover, A. Intelligent Use of Big Data for Heavy Oil Reservoir Management. In Proceedings of the SPE Annual Technical Conference and Exhibition, Houston, TX, USA, 28–30 September 2015; SPE: Houston, TX, USA, 2015; p. D021S011R003.
3. Duffy, W.; Rigg, J.; Maidla, E. Efficiency Improvement in the Bakken Realized Through Drilling Data Processing Automation and the Recognition and Standardization of Best Safe Practices. In Proceedings of the SPE/IADC Drilling Conference and Exhibition, The Hague, The Netherlands, 14–16 March 2017; SPE: The Hague, The Netherlands, 2017; p. D012S021R002.

4. Seemann, D.; Williamson, M.; Hasan, S. Improving Resevoir Management through Big Data Technologies. In Proceedings of the SPE Middle East Intelligent Oil and Gas Symposium, Manama, Bahrain, 28–30 October 2013; SPE: Manama, Bahrain, 2013; p. D021S010R001.
5. Rollins, B.T.; Broussard, A.; Cummins, B.; Smiley, A.; Eason, T. Continental Production Allocation and Analysis Through Big Data. In Proceedings of the 5th Unconventional Resources Technology Conference, Austin, TX, USA, 24–26 July 2017; American Association of Petroleum Geologists: Austin, TX, USA, 2017.
6. Mohamadi-Baghmolaei, M.; Azin, R.; Osfouri, S.; Mohamadi-Baghmolaei, R.; Zarei, Z. Prediction of gas compressibility factor using intelligent models. *Nat. Gas Ind. B* **2015**, *2*, 283–294. [CrossRef]
7. Xiao, C.; Zhang, S.; Ma, X.; Zhou, T.; Li, X. Surrogate-assisted hydraulic fracture optimization workflow with applications for shale gas reservoir development: A comparative study of machine learning models. *Nat. Gas Ind. B* **2022**, *9*, 219–231. [CrossRef]
8. Jiang, Y.; Li, J.; Zhang, H.; Wang, Q.; Yu, Y.; He, C.; Liang, M. Construction of data resource sharing center of the Puguang Intelligent Gas Field. *Nat. Gas Ind. B* **2019**, *6*, 215–219. [CrossRef]
9. Johnston, J.; Guichard, A. New Findings in Drilling and Wells using Big Data Analytics. In Proceedings of the Offshore Technology Conference, Houston, TX, USA, 4–7 May 2015; OTC: Houston, TX, USA, 2015; p. OTC-26021-MS.
10. Anand, P. Big Data Is a Big Deal. *J. Pet. Technol.* **2013**, *65*, 18–21. [CrossRef]
11. Salem, A.M.; Yakoot, M.S.; Mahmoud, O. Addressing Diverse Petroleum Industry Problems Using Machine Learning Techniques: Literary Methodology–Spotlight on Predicting Well Integrity Failures. *ACS Omega* **2022**, *7*, 2504–2519. [CrossRef] [PubMed]
12. Hollender, F. Improvement in Borehole Ground-penetrating Radar Tomography: Removal of Artifacts And Use of Wave Separation Algorithm. In Proceedings of the 2000 SEG Annual Meeting, Calgary, AB, Canada, 6–11 August 2000; p. SEG-2000-1923.
13. Lee, K.; Lim, J.; Yoon, D.; Jung, H. Prediction of Shale-Gas Production at Duvernay Formation Using Deep-Learning Algorithm. *SPE J.* **2019**, *24*, 2423–2437. [CrossRef]
14. Yin, Q.; Yang, J.; Tyagi, M.; Zhou, X.; Hou, X.; Wang, N.; Tong, G.; Cao, B. Machine Learning for Deepwater Drilling: Gas-Kick-Alarm Classification Using Pilot-Scale Rig Data with Combined Surface-Riser-Downhole Monitoring. *SPE J.* **2021**, *26*, 1773–1799. [CrossRef]
15. Yu, Y.; Xu, C.; Misra, S.; Li, W.; Ashby, M.; Pan, W.; Deng, T.; Jo, H.; Santos, J.E.; Fu, L.; et al. Synthetic Sonic Log Generation With Machine Learning: A Contest Summary From Five Methods. *Petrophys.—SPWLA J. Form. Eval. Reserv. Descr.* **2021**, *62*, 393–406. [CrossRef]
16. Fernández, A.; García, S.; Galar, M.; Prati, R.C.; Krawczyk, B.; Herrera, F. Introduction to KDD and Data Science. In *Learning from Imbalanced Data Sets*; Springer International Publishing: Cham, Switzerland, 2018; pp. 1–17. ISBN 978-3-319-98073-7.
17. Chandrashekar, G.; Sahin, F. A survey on feature selection methods. *Comput. Electr. Eng.* **2014**, *40*, 16–28. [CrossRef]
18. Li, J.; Cheng, K.; Wang, S.; Morstatter, F.; Trevino, R.P.; Tang, J.; Liu, H. Feature Selection: A Data Perspective. *ACM Comput. Surv.* **2018**, *50*, 1–45. [CrossRef]
19. Arkalgud, R.; McDonald, A.; Brackenridge, R. Automated Selection of Inputs for Log Prediction Models Using a New Feature Selection Method. In Proceedings of the SPWLA 62nd Annual Logging Symposium, Virtual, 17–20 May 2021; p. D041S029R003. [CrossRef]
20. Guyon, I.; Elisseeff, A. An introduction to variable and feature selection. *J. Mach. Learn. Res.* **2003**, *3*, 1157–1182.
21. Sánchez-Maroño, N.; Alonso-Betanzos, A.; Tombilla-Sanromán, M. Filter Methods for Feature Selection—A Comparative Study. In *Intelligent Data Engineering and Automated Learning—IDEAL 2007*; Yin, H., Tino, P., Corchado, E., Byrne, W., Yao, X., Eds.; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2007; Volume 4881, pp. 178–187. ISBN 978-3-540-77225-5.
22. Das, S. Filters, wrappers and a boosting-based hybrid for feature selection. In *Icml*; Citeseer: State College, PA, USA, 2001; Volume 1, pp. 74–81.
23. Stewart, G.; Du, K.F. Feature Selection and Extraction for Well Test Interpretation by an Artificial Intelligence Approach. In Proceedings of the SPE Annual Technical Conference and Exhibition, San Antonio, TX, USA, 8–11 October 1989; p. SPE-19820-MS. [CrossRef]
24. Sarma, P.; Durlofsky, L.J.; Aziz, K. Kernel Principal Component Analysis for Efficient, Differentiable Parameterization of Multipoint Geostatistics. *Math. Geosci.* **2008**, *40*, 3–32. [CrossRef]
25. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
26. Prokhorenkova, L.; Gusev, G.; Vorobev, A.; Dorogush, A.V.; Gulin, A. CatBoost: Unbiased boosting with categorical features. *Adv. Neural Inf. Process. Syst.* **2018**, *31*, 1–11.
27. Sun, Q.; Pfahringer, B. Bagging Ensemble Selection for Regression. In *AI 2012: Advances in Artificial Intelligence*; Thielscher, M., Zhang, D., Eds.; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2012; Volume 7691, pp. 695–706. ISBN 978-3-642-35100-6.
28. Mitchell, R.; Frank, E. Accelerating the XGBoost algorithm using GPU computing. *PeerJ Comput. Sci.* **2017**, *3*, e127. [CrossRef]
29. Saeed, U.; Jan, S.U.; Lee, Y.-D.; Koo, I. Fault diagnosis based on extremely randomized trees in wireless sensor networks. *Reliab. Eng. Syst. Saf.* **2021**, *205*, 107284.
30. Pandya, V.J. Comparing handwritten character recognition by AdaBoostClassifier and KNeighborsClassifier. In Proceedings of the 2016 8th International Conference on Computational Intelligence and Communication Networks (CICN), Tehri, India, 23–25 December 2016; pp. 271–274.

31. Uppari, R.R. Comparison between KERAS Library and FAST. AI Library Using Convolution Neural Network (Image Classification) Model. Doctoral Dissertation, Dublin Business School, Dublin, Ireland, 2020.
32. Lihui, Z.; Xiuyun, L.; Guandong, S.; Wei, Z.; Xuguang, G.; Xiujuan, T. Applicability of working fluid damage assessment methods for coalbed methane reservoirs. *Nat. Gas Ind.* **2018**, *38*, 28–39.
33. Sliwa, T.; Jarosz, K.; Rosen, M.A.; Sojczyńska, A.; Sapińska-Śliwa, A.; Gonet, A.; Fąfera, K.; Kowalski, T.; Ciepielowska, M. Influence of rotation speed and air pressure on the down the hole drilling velocity for borehole heat exchanger installation. *Energies* **2020**, *13*, 2716.
34. Zhai, X.; Chen, H.; Lou, Y.; Wu, H. Prediction and control model of shale induced fracture leakage pressure. *J. Pet. Sci. Eng.* **2021**, *198*, 108186. [CrossRef]
35. Alum, M.A.; Egbon, F. Semi-Analytical Models on the Effect of Drilling Fluid Properties on Rate of Penetration (ROP). In Proceedings of the SPE Nigeria Annual International Conference and Exhibition, Abuja, Nigeria, 30 July–3 August 2011; SPE: Abuja, Nigeria, 2011; p. SPE-150806-MS.
36. Wang, J.Y.; Holditch, S.A.; McVay, D.A. Modeling Fracture-Fluid Cleanup in Tight-Gas Wells. *SPE J.* **2010**, *15*, 783–793. [CrossRef]
37. Galindo, T. Can Proppant Transport be Negatively Affected by Too Much Viscosity? In Proceedings of the SPE Hydraulic Fracturing Technology Conference and Exhibition, The Woodlands, TX, USA, 5–7 February 2019; SPE: The Woodlands, TX, USA, 2019; p. D021S006R002.