УДК 004.424.62
МРНТИ 52.47.19

DOI: https://doi.org/10.54859/kjogi108642

Получена: 20.04.2023.
Одобрена: 24.08.2023.
Опубликована: 30.09.2023.

Оригинальное исследование

Прогноз обводнённости новых скважин с помощью машинного обучения

А.Е. Ибраев, Г.С. Камариденова, Б.А. Балуанов, А.С. Елемесов
КМГ Инжиниринг, г. Астана, Казахстан

АННОТАЦИЯ

Обоснование. Бурение новых скважин относится к одним из наиболее эффективных геолого-технических мероприятий. На зрелых месторождениях, характеризующихся высокой выработкой запасов и обводнённостью, подбор проектных точек для бурения является сложной задачей. Прогнозирование параметров новых скважин возможно с помощью применения геолого-гидродинамических моделей либо аналитических методов. В данной работе авторами предложено использование алгоритмов машинного обучения для прогноза пусковых параметров новых скважин на основе обширного набора геологических и промысловых данных.

Цель. В статье приведено описание процесса разработки алгоритмов машинного обучения и продемонстрированы показатели эффективности комплексной модели. В рамках данной работы выполнено апробирование алгоритмов машинного обучения для прогноза пусковой обводнённости потенциальных кандидатов.

Материалы и методы. В рамках данной работы были применены различные методы машинного обучения на геолого-технических промысловых данных.

Результаты. Разработанная комплексная модель показала приемлемые результаты сходимости на основе метрик классификации и регрессии, что говорит о её применимости для прогноза пусковой обводнённости проектных скважин.

Заключение. Данный метод прогнозирования показателей является альтернативным инструментом прогноза пусковой обводнённости новых скважин, позволяющим уточнить и дополнить прогнозные параметры, рассчитанные с помощью геолого-гидродинамической модели или эмпирических зависимостей пусковой обводнённости новых скважин от геологических параметров.

Ключевые слова: *бурение новых скважин, прогноз обводнённости, распределение, датасет, препроцессинг данных, машинное обучение, оценка эффективности.*

Как цитировать:
Ибраев А.Е., Камариденова Г.С., Балуанов Б.А., Елемесов А.С. Прогноз обводнённости новых скважин с помощью машинного обучения // *Вестник нефтегазовой отрасли Казахстана.* 2023. Том 5, №3. С. 20–34. DOI: https://doi.org/10.54859/kjogi108642.

UDC 004.424.62
CSCSTI 52.47.19

DOI: https://doi.org/10.54859/kjogi108642

Received: 20.04.2023.
Accepted: 24.08.2023.
Published: 30.09.2023.

Original article

New well water cut prediction using machine learning

Aktan Ye. Ibrayev, Gaukhar S. Kamaridenova, Bakytzhan A. Baluanov, Azamat S. Yelemessov
KMG Engineering, Astana, Kazakhstan

ANNOTATION

Background: The drilling of new wells is one of the most effective geological and technical activities. In mature fields characterized by high production of reserves and high water availability, the selection of design points for drilling is a difficult task. Forecasting the parameters of new wells is possible by using geological and hydrodynamic models or analytical methods. In this paper, the authors propose the use of machine learning algorithms to predict the initial parameters of new wells based on an extensive set of geological and field data.

Aim: The article describes the process of developing machine learning algorithms and demonstrates the performance indicators of a complex model. As part of this work, testing of machine learning algorithms was performed to predict the start-up water cut of potential candidates.

Materials and methods: Within the framework of this work, various machine learning methods were applied on geological and technical field data.

Results: The developed complex model showed acceptable convergence results based on classification and regression metrics, which indicates its applicability for predicting the start-up water cut of project wells.

Conclusion: This method of predicting indicators is an alternative tool for predicting the start-up water cut of new wells, which makes it possible to clarify and supplement the forecast parameters calculated using a geological and hydrodynamic model or empirical dependencies of the initial water cut of new wells on geological parameters.

Keywords: *drilling of new wells, water cut forecast, distribution, dataset, data preprocessing, machine learning, performance evaluation.*

To cite this article:
Ibrayev AY, Kamaridenova GS, Baluanov BA, Yelemessov AS. New well water cut prediction using machine learning. *Kazakhstan journal for oil & gas industry.* 2023;5(3):20–34. DOI: https://doi.org/10.54859/kjogi108642.

ӘОЖ 004.424.62
ҒТАХР 52.47.19

DOI: https://doi.org/10.54859/kjogi108642

Қабылданды: 20.04.2023.
Мақұлданды: 24.08.2023.
Жарияланды: 30.09.2023.

Түпұнсқа зерттеу

Машиналық оқыту көмегімен жаңа ұңғымалардың сулануын болжау

А.Е. Ібраев, Г.С. Қамариденова, Б.А. Балуанов, А.С. Елемесов
ҚМГ Инжиниринг, Астана қаласы, Қазақстан

АННОТАЦИЯ

Негіздеу. Жаңа ұңғымаларды бұрғылау ең тиімді геологиялық-техникалық шаралардың біріне жатады. Қорлардың жоғары өндірілуімен және сулануымен сипатталатын жетілген кен орындарында бұрғылау үшін жобалау нүктелерін таңдау қиын міндет болып табылады. Жаңа ұңғымалардың көрсеткіштерін болжау геологиялық-гидродинамикалық модельдерді немесе аналитикалық әдістерді қолдану арқылы мүмкін болады. Бұл жұмыста авторлар геологиялық және кәсіптік деректердің кең жиынтығы негізінде жаңа ұңғымалардың іске қосу параметрлерін болжау үшін машиналық оқыту алгоритмдерін қолдануды ұсынды.

Мақсаты. Мақалада машиналық оқыту алгоритмдерін жасау процесі сипатталған және күрделі модельдің өнімділік көрсеткіштері көрсетілген. Осы жұмыстың бір бөлігі ретінде әлеуетті үміткерлердің іске қосу сулануын болжау үшін машиналық оқыту алгоритмдерін тестілеу жүргізілді.

Материалдар мен әдістер. Бұл жұмыс аясында геологиялық-техникалық кәсіптік деректерге әртүрлі машиналық оқыту әдістері қолданылды.

Нәтижелері. Әзірленген интеграцияланған модель топтастыру және регрессия көрсеткіштеріне негізделген конвергенцияның қолайлы нәтижелерін көрсетті, бұл оның жобалық ұңғымаларды іске қосу сулануының болжау үшін қолдану мүмкіндігін көрсетеді.

Қорытынды. Көрсеткіштерді болжаудың бұл әдісі геологиялық-гидродинамикалық модельдің немесе жаңа ұңғымалардың іске қосу сулануының геологиялық көрсеткіштеріне эмпирикалық тәуелділіктерінің көмегімен есептелген болжамды көрсеткіштерді нақтылауға және толықтыруға мүмкіндік беретін жаңа ұңғымалардың іске қосу сулануын болжаудың балама құралы болып табылады.

Негізгі сөздер: *жаңа ұңғымаларды бұрғылау, сулануды болжау, тарату, деректер жинағы, деректерді алдын ала өңдеу, машиналық оқыту, тиімділікті бағалау.*

Дәйексөз келтіру үшін:
Ібраев А.Е., Камариденова Г.С., Балуанов Б.А., Елемесов А.С. Машиналық оқыту көмегімен жаңа ұңғымалардың сулануын болжау // *Қазақстанның мұнай-газ саласының хабаршысы.* 2023. 5 том, №3, 20–34 б. DOI: https://doi.org/10.54859/kjogi108642.

 2957-806X

Введение

Для месторождений, на которых не применяются геолого-гидродинамические модели (далее – ГГДМ), расчёт пусковой обводнённости проектных скважин ведётся эмпирическим путём. Инженеры анализируют обширный набор параметров, чтобы выбрать подходящую зону для бурения. В результате экспертного анализа собранной информации геологи и разработчики оценивают начальную обводнённость проектной точки.

Такой метод несёт в себе риски ошибки из-за влияния человеческого фактора и невозможности полного охвата входных параметров. В этой связи возникла идея создания алгоритма для прогноза пусковой обводнённости новых скважин. Применение алгоритма вместо экспертной оценки позволит использовать единый подход для каждой зоны и минимизировать влияние человеческого фактора.

Для поиска решения авторы собрали массив данных по добыче, геологии и геофизике, который позволил бы рассчитать обводнённость с помощью общепринятых формул либо выявить эмпирические зависимости между обводнённостью и геолого-техническими параметрами.

Обводнённость напрямую зависит от нефтенасыщенности пласта, но при сборе геологической информации не было обнаружено представительных по объёму источников текущей насыщенности. На момент подготовки алгоритма имеющаяся ГГДМ обновлялась более 5 лет назад. Анализ результатов интерпретации геофизических исследований (далее – РИГИС) новых скважин показал, что обводнённость коррелируется с насыщенностью (рис. 1), но собранных материалов РИГИС недостаточно для масштабирования. Экстраполяция значений насыщенности на всю площадь залежи на основе немногочисленных данных РИГИС по скважинам, пробуренным за последние 3–5 лет, неизбежно приведет к значительным ошибкам из-за того, что скважины расположены неравномерно на объектах и площади нефтеносности очень велики.

Из-за отсутствия возможности масштабирования имеющейся информации использование общепринятых формул оказалось невозможным. Далее авторы предприняли попытки поиска эмпирических зависимостей обводнённости от геолого-физических параметров.

Данные на рис. 2–3 свидетельствуют о слабой корреляции обводнённости от одной переменной, а в некоторых случаях указывают на обратную корреляцию обводнённости от входных переменных, что говорит о невозможности применения данных зависимостей для всех объектов без исключения. В связи с этим дальнейшие попытки обнаружения эмпирических зависимостей были прекращены.

Литературный обзор

В работе [1] обсуждаются методы интеграции физических традиционных методов прогнозирования добычи, таких как анализ кривых падения (далее – DCA), и автоматическое применение машинного обучения для подбора и сопоставления данных. Авторы отмечают, что одним из недостатков использования машинного обучения при расчёте показателей является то, что алгоритмы могут давать аномальные результаты. Такие нефизические данные вызывают у специалистов скептицизм. В связи с этим авторами было проведено исследование, в котором за основу был взят принятый и проверенный метод анализа кривых падения. Объектом исследования стали 396 горизонтальных скважин месторождения Баккен. Авторы использовали набор данных из открытых источников. При расчёте показателей применялись модификации кривой Арпса: Stretched Exponential Decline Mode, Duong Model, Pan's Combined Capacitance-Resistance Model, Bayesian Neural Ordinary Differential Equation. Каждая модель характеризуется набором коэффициентов, влияющих на результаты расчёта. Для подбора этих коэффициентов применялись методы автоматического машинного обучения, которые самостоятельно выбирают наилучший алгоритм на основе входных данных и целевых переменных. В статье описаны этапы и применяемые методики проведенного исследования. Итоговые комплексные DCA-модели показали хорошие результаты при сопоставлении фактической и расчётной добычи нефти с историей. Авторы пришли к выводу, что машинное обучение можно использовать для создания прогностических моделей с использованием различных параметров. Одним из направлений совершенствования модели является включение в набор признаков геологической информации.

Статья [2] демонстрирует исследование по интеграции DCA, аналих типовых кривых падения (далее – TCM) и результатов численного моделирования пласта. Авторы отмечают, что для зрелых месторождений наиболее распространенной и доступной информацией являются данные по добыче. Эти данные используются в DCA и TCM. Прогноз по кривым падения (далее – PDA) может предоставить приемлемые характеристики резервуара, но у него есть два недостатка: во-первых, для характеристики резервуара

процесс требует данные о забойном или устьевом давлении в дополнение к данным дебита. Данные о забойном или устьевом давлении обычно недоступны на большинстве зрелых месторождений. Во-вторых, методика, позволяющая интегрировать результаты сотен отдельных скважин в связный анализ всего месторождения или пласта для принятия бизнес-решений, не является частью современного набора инструментов PDA. DCA и TCM были объединены для формирования полного прогноза добычи на основе индивидуальных расчётов для каждой скважины. Общеизвестно, что DCA представляет собой математический метод прогнозирования работы скважины, не имеющий физической основы. Численное моделирование было выполнено с использованием моделирования методом Монте-Карло. Полученные результаты сопоставлены с фактическими данными по 137 скважинам. В этой статье показаны методы интеграции различных методов прогнозирования с численным моделированием.

В работе [3] представлено сравнение результатов прогноза добычи нефти с результатами регрессионного анализа на основе машинного обучения. Авторы выявили сильную корреляцию целевой переменной с технологическими факторами, такими как конструкция скважины и параметры ГРП, на основании которых были построены эмпирические зависимости для расчёта целевой переменной. Также была проведена работа по созданию алгоритма машинного обучения на основе обширной матрицы данных (более 350 параметров) с использованием линейных моделей Лассо из библиотеки scikit-learn. На основании ретроспективного анализа можно сделать вывод, что полученные авторами результаты имеют высокую сходимость с фактическими производственными данными. Основное влияние на продуктивные характеристики скважин оказывают технологические факторы, характеризующие количественные и качественные параметры жидкостей и пропанта, закачиваемых в пласт при ГРП. Выявлена корреляция между типом жидкости ГРП и величиной стимулированного объема пласта, влияющая на показатели добычи нефти анализируемого фонда. Формируются комплексные параметры, характеризующие добычу нефти скважинами.

Работа [4] посвящена проблеме ранжирования потенциальных кандидатов на уплотняющее бурение по полевым данным. Для этого авторами разработан двухэтапный алгоритм классификации заложенных точек на основе показателей работы скважин в окружающей среде, а также геологических показателей без использования ГГДМ. Данная работа демонстрирует возможность точного прогнозирования работы скважины без использования моделей пласта, оперируя обширным набором промысловых данных. Стоит отметить, что авторы использовали метод опорных векторов, т.к. входные данные для их задачи показали хорошую взаимную корреляцию.

В статье [5] авторами проведены исследования по рекуррентному прогнозу пластового давления месторождения на основе изменения дебита скважинной жидкости и предыдущих значений замеров пластового давления. В этой статье обсуждается использование машинного обучения для оценки его эффективности и потенциала для определения и прогнозирования значений пластового давления при разработке нефтяных месторождений в сравнении с традиционными статистическими моделями. Были применены два метода: метод линейной регрессии и модель случайного леса. Результаты работы показали необходимость контроля замеров давления для исключения выбросов и привлечения специалистов для корректировки моделей. Предлагается новый метод прогнозирования пластового давления с помощью машинного обучения, основанный на непараметрической многомерной модели, связывающей работу скважины с течением времени. Предлагаемый метод учитывает динамику показателей, характеризующих работу скважин, а прогнозируемое пластовое давление хорошо коррелирует со значениями, измеренными с помощью гидродинамических исследований скважин.

Авторы статьи [6] разработали прикладной инструмент для экспресс-оценки темпов падения добычи на основе минимального набора исходных данных. В статье описан новый метод прогнозирования кривых падения проектных скважин. Метод основан на интеграции ручной группировки кривых спада и применении алгоритмов машинного обучения. Машинное обучение позволяет находить скрытые связи между функциями и выводом. Статья включает анализ кривых падения для двух типов заканчивания скважин: горизонтальных и наклонных, который показывает, что горизонтальные скважины более эффективны, чем наклонные. Представлен критерий оценки достаточности данных, проведено сравнение с расчётами темпов снижения по Арпсу и оценена производительность двух алгоритмов машинного обучения.

В статье [7] предложены алгоритмы прогнозирования обводнённости новых скважин по данным геофизических исследований.

 DOI: 10.54859/kjogi108642

Анализ корреляции между данными ГИС и обводнённостью не выявил видимых закономерностей. Были изучены четыре метода генерации признаков для набора данных, а также четыре метода классификации (деревья решений, случайный лес, градиентное повышение и нейронные сети). Все методы показали схожие результаты. На наборе данных по результатам интерпретации ГИС были обучены классификаторы, которые показали высокую степень точности прогнозов на контрольных данных, но использование фактических данных не дало результатов.

В статье [8] рассматривается возможность использования аналого-статистических методов при прогнозировании обводнённости добывающих скважин с учетом влияния геолого-технологических показателей. Несмотря на свои широкие возможности, методы моделирования разработки нефтегазовых месторождений, основанные на построении ГГДМ, зачастую требуют неоправданно больших финансовых и временных затрат. Для задач проектирования подходит разработка надёжных статистических оценок, позволяющих контролировать текущую нефтеотдачу. Кроме того, статистические оценки более устойчивы к ошибкам в информации, чем методы моделирования разработки месторождений. Одним из ключевых звеньев является построение для аналогов моделей модифицированных зависимостей обводнённости продукции скважин от степени извлечения извлекаемых запасов нефти.

Таким образом, основываясь на результатах рассмотренных научных работ, была предложена гипотеза о том, что для прогноза обводнённости проектных точек возможно применение алгоритмов машинного обучения. Новизна предложенного в данной статье подхода заключается в том, что в отличие от работы [7] в качестве параметров модели были использованы данные по добыче и соседним скважинам, а не геофизические данные, которые отсутствуют на момент заложения проектных точек. Результаты работы [1] обусловили необходимость проверки корректности входных данных, подаваемых для обучения моделей, и целесообразность применения параметров, которые имеют непосредственное влияние на прогнозную обводнённость скважин. Также рассмотренные работы повлияли на дальнейший выбор применяемых в данном исследовании алгоритмов машинного обучения.

Выбор алгоритмов машинного обучения

Большое количество разнородных и невзаимосвязанных данных навело на идею тестирования алгоритмов машинного обучения для решения поставленной задачи. Машинное обучение – это класс методов искусственного интеллекта, характерной чертой которых является не прямое решение задачи, а обучение за счёт применения решений множества сходных задач. Для построения таких методов используются средства математической статистики, численных методов, математического анализа, методов оптимизации, теории вероятности, теории графов, различные техники работы с данными в цифровой форме. Существует две классические задачи, решаемые с помощью машинного обучения, – классификация и регрессия.

Задача классификации – задача, в которой имеется множество объектов (ситуаций), разделённых некоторым образом на классы. Задано конечное множество объектов, для которых известно, к каким классам они относятся. Это множество называется выборкой. Классовая принадлежность остальных объектов неизвестна. Требуется построить алгоритм, способный классифицировать произвольный объект из исходного множества.

Регрессионный анализ – набор статистических методов исследования влияния одной или нескольких независимых переменных на зависимую переменную. Терминология зависимых и независимых переменных отражает лишь математическую зависимость переменных, а не причинно-следственные отношения.

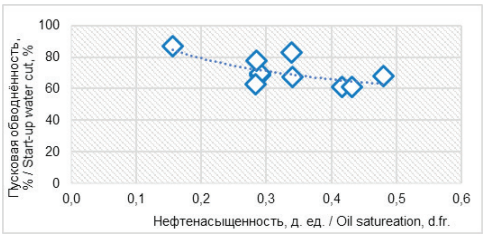


Рисунок 1. Зависимость пусковой обводнённости от нефтенасыщенности
Figure 1. Dependence of start-up water cut on oil saturation

Для решения задачи авторы выбрали алгоритмы Random forest, заключающиеся в использовании комитета (ансамбля) решающих «деревьев» (подмоделей, обученных на небольших частях набора данных). Схема работы Random Forest представлена на рис. 4.

Алгоритм применяется для задач классификации, регрессии и кластеризации. Алгоритм Random forest сочетает в себе две основные идеи: метод бэггинга и метод случайных подпространств. Метод бэггинга подразумевает под собой параллельное

обучение на многочисленных частях данных и выбор лучшего результата путём осреднения полученных результатов. Метод случайных подпространств означает обучение на частях данных, которые выбираются случайным образом. Комбинация методов повышает стабильность и точность алгоритмов обучения.

Основная идея заключается в использовании большого ансамбля решающих деревьев, каждое из которых само по себе даёт невысокое качество классификации, но за счёт большого количества результат получается достоверным. Алгоритмы Random forest обладают следующими преимуществами:

- способность эффективно обрабатывать данные с большим числом признаков и классов;
- нечувствительность к масштабированию значений признаков;
- одинаково хорошо обрабатываются как непрерывные, так и дискретные признаки;
- существуют методы оценивания значимости отдельных признаков в модели.

Для создания модели машинного обучения необходим обширный массив параметров описываемой среды – датасет. В качестве параметров модели использованы такие геолого-физические характеристики, которые описывали бы состояние среды в точках, где отсутствуют скважинные данные. Это означает, что в модель не закладываются значения параметров из РИГИС по одной конкретной скважине, геолого-технические мероприятия, техническое состояние скважины, т.к. такие значения для проектной точки отсутствуют. Поэтому каждая точка характеризуется текущими и накопленными показателями окружения [1] и значениями параметров с геологических карт. Для обучения модели сформирован датасет по ежемесячному действующему фонду скважин, в котором для всех скважин все параметры рассчитывались на каждый календарный месяц в период 2019–2021 гг. Таким образом, итоговый датасет содержит 230 тыс. строк, каждая из которых характеризуется 23 параметрами. В датасете представлены данные по более чем 3 тыс. добывающих скважин по таким параметрам, как накопленная добыча нефти и жидкости, накопленное отработанное время, накопленная и текущая закачка соседних нагнетательных скважин, взаимное расположение скважин и контуров нефтеносности, эксплуатируемый объект, начальные нефтенасыщенные толщины, текущая минерализация пластовой воды и др.

Необходимо отметить, что алгоритмы Random forest автоматически оценивают значимость отдельных признаков и не используют в расчётах нерелевантные параметры.

Обработка данных и разработка моделей

Перед началом расчётов на модели исходные данные прошли предварительную обработку, которая включает в себя отбраковку выбросов по обводнённости и дебитам жидкости и последующее преобразование параметров для нормализации распределения.

На первом этапе обработки из датасета исключены точки с аномальными значениями обводнённости. Рис. 5 отображает распределение обводнённости в датасете в виде диаграммы разброса (или «ящика с усами»). На диаграмме выбросы рисуют точками выше и ниже «усов».

Минимальное значение для исключения выбросов определялось на основе значений интерквантильного размаха согласно формуле (1):

$$lower\_threshold = Q_{25} - 1{,}5 \times IQR \quad (1)$$

где $lower\_threshold$ – минимальное значение обводнённости в выборке; Q_{25} – 25-й перцентиль выборки; IQR – интерквантильный размах выборки, который, в свою очередь, рассчитывается по формуле (2):

$$IQR = Q_{75} - Q_{25} \quad (2)$$

где Q_{75} – 75-й перцентиль выборки.

После отбраковки выбросов по обводнённости из датасета исключались точки, которые лежат за пределами доверительного интервала, рассчитанного на основе средних дебитов жидкости по окружению (рис. 6). Ширина доверительного интервала равна 2 среднеквадратическим отклонениям.

Параметры в датасете характеризуются асимметричными гистограммами распределений. Для повышения точности модели датасет преобразовали с помощью пакета PowerTransformer из библиотеки scikit-learn. На рис. 7 представлен пример преобразования массива по накопленной добыче нефти. Преобразование данных повышает степень различия параметров в датасетах, чтобы алгоритм отчетливо дифференцировал точки при обучении модели.

Для расчётов авторы разработали комплексную модель. Целевая переменная модели – пусковая обводнённость. На первом этапе обработанный датасет проходит через классификатор RandomForestClassifier, который делит скважины на два класса. На втором этапе для каждого класса регрессор RandomForestRegressor предсказывает дискретное значение целевой переменной (рис. 8).

DOI: 10.54859/kjogi108642

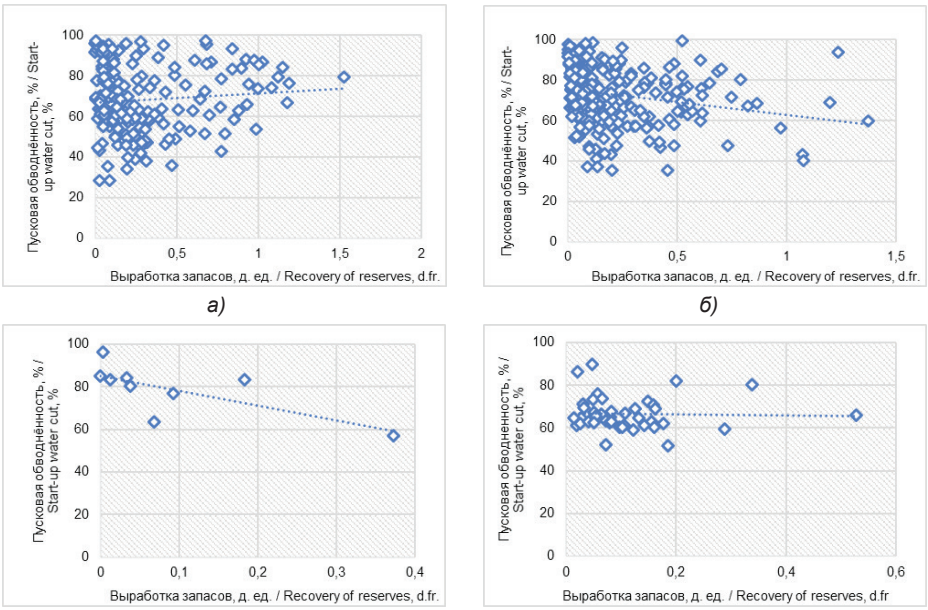


Рисунок 2. Зависимость пусковой обводнённости от выработки запасов
Figure 2. Dependence of initial water cut on oil recovery
а) 13 горизонт / Horizon 13; б) 14 горизонт / Horizon 14; в) 19 горизонт / Horizon 19; г) 20 горизонт / Horizon 20

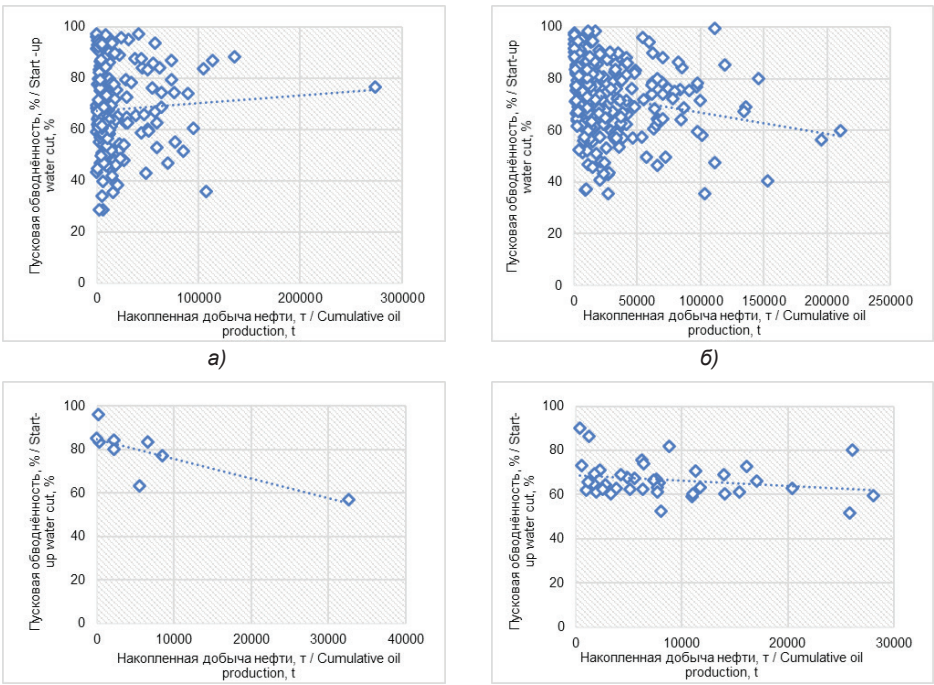


Рисунок 3. Зависимость пусковой обводнённости от накопленной добычи
Figure 3. Dependence of initial water cut on cumulative oil production
а) 13 горизонт / Horizon 13; б) 14 горизонт / Horizon 14; в) 19 горизонт / Horizon 19.; г) 20 горизонт / Horizon 20

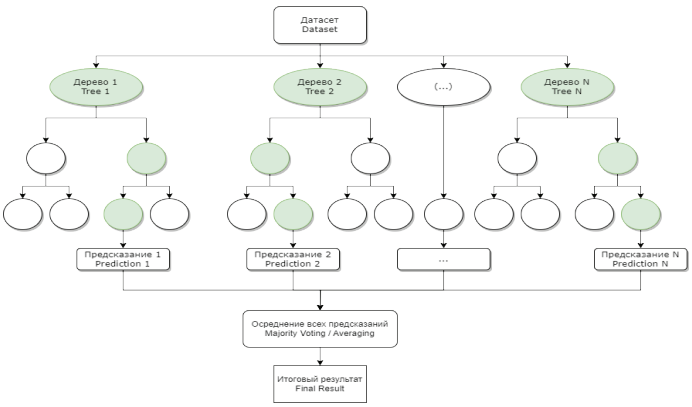


Рисунок 4. Схематическое описание алгоритма Random forest
Figure 4. Random forest algorithm scheme

Для валидации результатов модели проводилась процедура разделения датасета на тренировочную и тестовую выборки с помощью программного пакета train_test_split. Массив со значениями параметров делится случайным образом. Модель обучается на тренировочной выборке, а оценка эффективности проводится на тестовой выборке, которая не применялась для обучения.

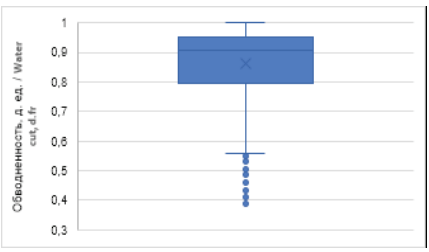


Рисунок 5. Распределение значений обводнённости
Figure 5. Water cut values distribution

При обучении модели возможны случаи, когда качество модели на обучающей выборке существенно превосходит качество модели на тестовой выборке. Построенная модель хорошо объясняет примеры из тренировочной выборки, но относительно плохо работает на примерах из тестовой выборки. Модель запоминает все возможные примеры вместо того, чтобы научиться подмечать особенности. Такое явление называется переобучением.

Для недопущения переобучения модели применяется метод перекрестной проверки Cross-Validation и подбираются гиперпараметры с помощью функции GridSearchCV. При оценке модели датасет разбивается на k частей. Затем модель обучается на k−1 частях, а оставшаяся часть используется для тестирования. Процедура повторяется k раз; в итоге каждая из k частей датасета используется для тестирования. В результате получается оценка эффективности выбранной модели с равномерным использованием данных. При этом модель обучается для каждого из заданных значений гиперпараметров. Для условий данной задачи выполнялся подбор числа деревьев в «лесу» (n_estimators), минимального числа объектов в простейшем элементе модели (min_samples_leaf), максимальной глубины дерева решений (max_depth). Схема обучения модели представлена на рис. 9.

Значения целевой переменной в датасете распределяются асимметрично (рис. 10). Если разделить тренировочный датасет по арифметическому среднему значению целевой переменной, то возникнет дисбаланс количества точек в классах. Это приведет к тому, что классификатор будет склонен предсказывать наиболее представительный класс. По этой причине при обучении классификатора датасет делится на два класса с одинаковым количеством точек в каждом классе, используя медианное значение целевой переменной. Такое деление позволяет сбалансированно разделить датасет на классы и повысить точность предсказания обводнённости.

Оценка полученных результатов

Оценка эффективности моделей машинного обучения производилась с помощью следующих метрик:

- для классификатора – метрики *accuracy, f1_score и matthews_correlation_coefficient* (далее – MCC);
- для регрессора – метрики R^2, *mean_absolute_error и mean_squared_error*.

Оценка классификатора выполняется на основе матрицы ошибок (Confusion Matrix). Матрица ошибок – это метрика производительности классифицирующей модели (рис. 11) [6]. Согласно данной матрице оценка

 DOI: 10.54859/kjogi108642

проводилась по трём метрикам: *accuracy, F1 score и MCC*.

Метрика accuracy отображает отношение общего количества верно предсказанных классов ко всему количеству точек в датасете (3):

$$\text{Accuracy} = \frac{\text{TP} + \text{T N}}{\text{P} + \text{N}} \quad (3)$$

Метрика F1 score позволяет оценить, действительно ли классификатор точен в предсказаниях разных классов либо склонен к предсказанию наиболее представительного класса:

$$\text{F1} = \frac{2\text{TP}}{2\text{TP} + \text{F P} + \text{FN}} \quad (4)$$

Метрика Matthews correlation coefficient, или phi coefficient, – это статистическая метрика, которая дает высокий балл только в том случае, если прогноз даёт хорошие результаты во всех четырёх категориях матрицы ошибок пропорционально как размеру положительных элементов, так и размеру отрицательных элементов в наборе данных [9] (5):

		Предсказанные значения PREDICTED VALUES	
	Общая выборка = P + N	Позитивные (PP)	Негативные (NN)
Фактические значения Actual values	Позитивные (P)	Истинно позитивные (TP)	Ошибочно негативные (FN)
	Негативные (N)	Ошибочно позитивные (FP)	Истинно негативные (TN)

Рисунок 11. Матрица ошибок
Figure 11. Confusion matrix

$$\text{MCC} = \frac{\text{TP} \times \text{T N} - \text{FP} \times \text{F N}}{\sqrt{(\text{TP} + \text{F P})(\text{TP} + \text{FN})(\text{TN} + \text{F P})(\text{TN} + \text{FN})}} \quad (5)$$

Метрика R^2 для регрессора представляет собой коэффициент детерминации – долю дисперсии зависимой переменной y, объясняемой рассматриваемой моделью зависимости, т.е. объясняющими переменными x (6):

$$\text{R}^2 = 1 - \frac{\text{D}[y|x]}{\text{D}[x]} \quad (6)$$

Средняя абсолютная ошибка mean_absolute_error (далее – MAE) рассчитывается как среднее значение абсолютных разностей между целевыми значениями и прогнозами (7):

$$\text{MAE} = \frac{\sum_{i=1}^{n}(y_i - x_i)}{n} \quad (7)$$

Средняя квадратическая ошибка mean_squared_error (далее – MSE) в основном измеряет среднеквадратическую ошибку прогнозов (8):

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i)^2 \quad (8)$$

Для оценки эффективности модели выполнены расчёты обводнённости на основе фактических данных работы скважин за январь 2022 г. Представлены оценки по тренировочной и по тестовой выборке. Метрики модели представлены в табл. 1.

Из табл. 1 видно, что классификатор правильно предсказал категорию скважин для 90% точек в тестовой выборке. Регрессор для класса 1 также достиг приемлемых показателей эффективности на обеих выборках. Средняя ошибка при определении целевой переменной на тестовой выборке составила 3%. При этом регрессор для класса 2 переобучился, потому что значение R^2 на тренировочной выборке значительно выше, чем на тестовой выборке. Регрессор для класса 2 не представляет интереса, т.к. отсутствует практическая необходимость точного прогнозирования для высо-

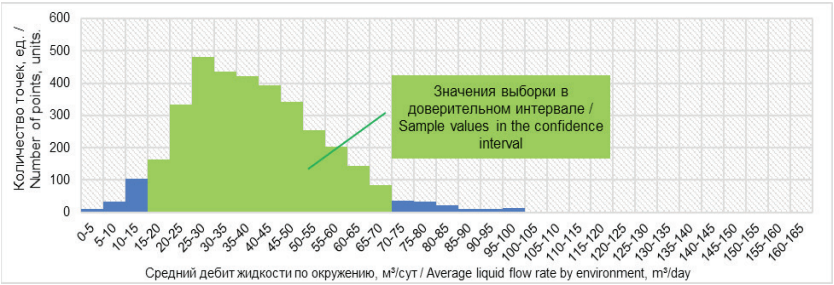
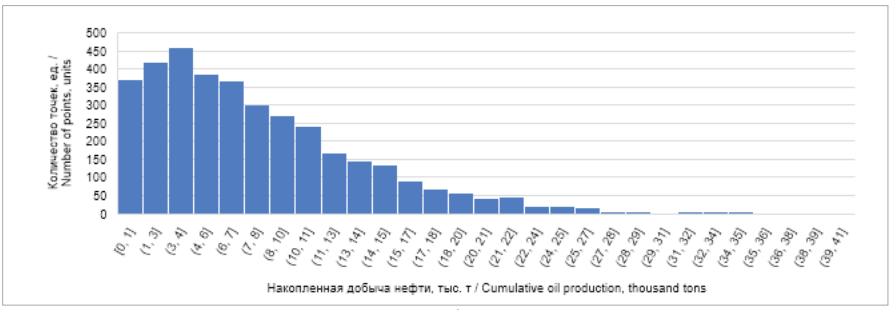
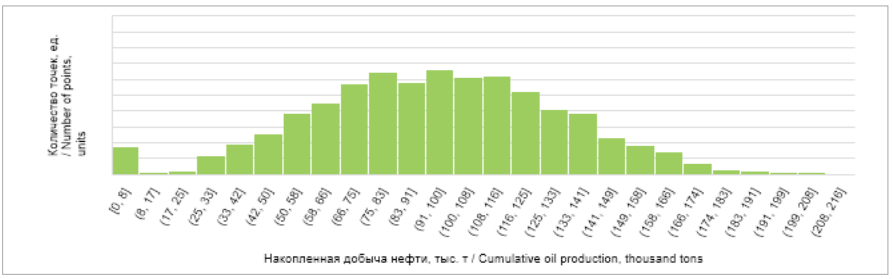


Рисунок 6. Распределение значений средних дебитов жидкости по окружению
Figure 6. Distribution of average liquid flow rate by environment



а)



б)

Рисунок 7. Преобразование данных по накопленной добыче нефти
Figure 7. Cumulative oil production data transformation
а) исходное распределение / initial distribution; б) преобразованные данные / transformed data

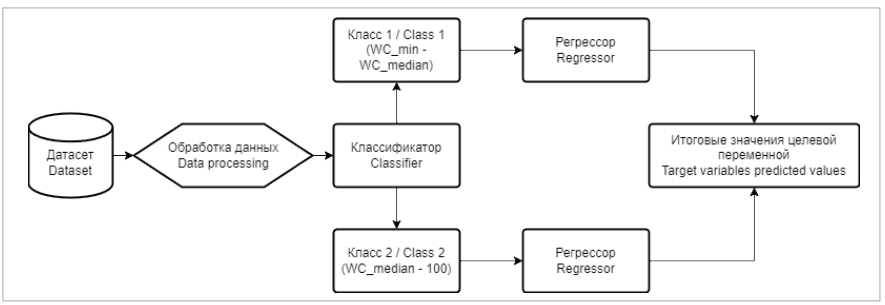


Рисунок 8. Архитектура модели машинного обучения
Figure 8. ML model architecture

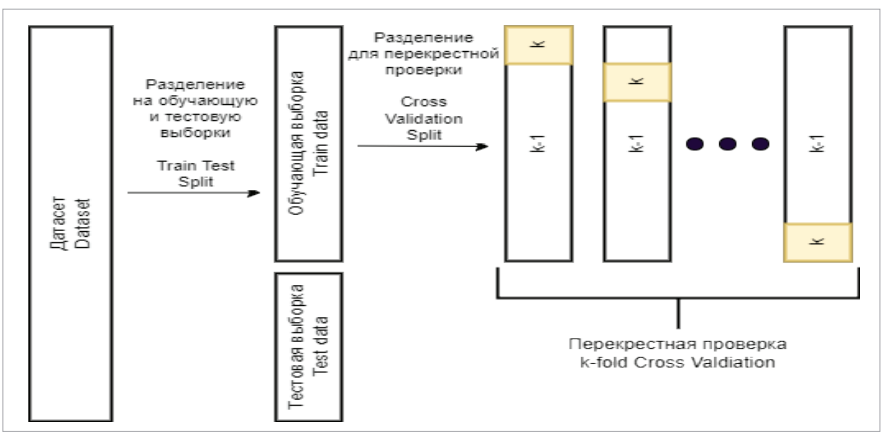


Рисунок 9. Схема обучения модели с использованием функций train_test_split и GridSearchCV
Figure 9. Model training scheme using train_test_split and GridSearchCV functions

DOI: 10.54859/kjogi108642

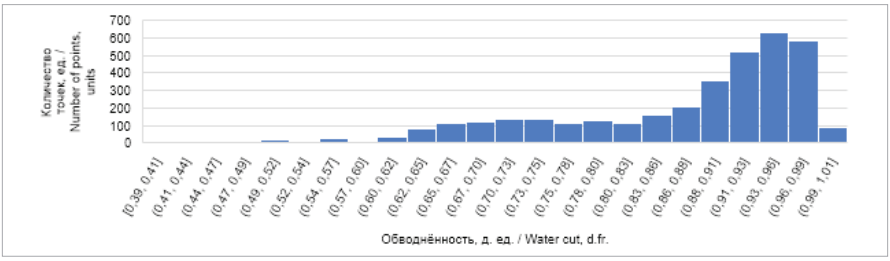


Рисунок 10. Гистограмма распределения значений обводнённости
Figure 10. Water cut values distribution histogram

кообводнённых скважин, потому что бурение таких скважин нерентабельно.

Таблица 1. Оценка эффективности разработанной модели
Table 1. Evaluation of the effectiveness of the developed model

Метрика / Metrics	Тренировочная выборка / Train data	Тестовая выборка / Test data
Классификатор / Classifier		
Accuracy	0,999	0,898
F1 score	0,999	0,897
MCC	0,999	0,797
Границы классов / Class limits	55,9 – 91,8%, 91,8 – 100,0%	
Регрессор для класса 1 / Class 1 regressor (55,9–91,8%)		
R2	0,975	0,818
MAE	0,011	0,030
MSE	0,000254	0,00183
Регрессор для класса 2 / Class 2 regressor (91,8–100,0%)		
R2	0,939	0,554
MAE	0,004	0,011
MSE	0,000026	0,00019

		Предсказанные значения PREDICTED VALUES	
	Общая выборка 41012 скважин / Total sampling of 41012 wells.	Позитивные (PP)	Негативные (NN)
Фактические значения Actual values	Позитивные (P)	18309	2197
	Негативные (N)	1889	18617

Рисунок 12. Матрица ошибок разработанного классификатора
Figure 12. Confusion matrix for the developed classifier

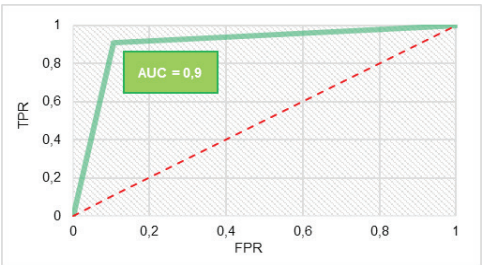


Рисунок 13. ROC-кривая для разработанного классификатора
Figure 13. ROC-curve for the developed classifier

На рис. 12 представлена матрица ошибок по тестовой выборке для разработанного классификатора.

На рис. 13 показана ROC-кривая – график, позволяющий оценить качество бинарной классификации. Этот график отображает соотношение доли объектов от общего количества носителей признака, верно классифицированных как несущие признак (TPR, англ. true positive rate, называемой чувствительностью алгоритма классификации), и доли объектов от общего количества объектов, не несущих признака, ошибочно классифицированных как несущие признак (FPR, англ. false positive rate, величина 1-FPR называется специфичностью алгоритма классификации). Количественная интерпретация ROC даёт показатель AUC (англ. Area Under Curve) — площадь, ограниченная ROC-кривой и осью доли ложных положительных классификаций. Чем выше показатель AUC, тем качественнее классификатор, при этом значение 0,5 демонстрирует непригодность выбранного метода классификации (соответствует случайному гаданию). В нашем случае AUC равен 0,9.

На рис. 14 отображено распределение фактических и предсказанных значений обводнённости для скважин в классе 1 (55,9 – 91,8%). Хотя данный график указывает на наличие корреляции между актуальными и расчётными значениями обводнённости, необходимо отметить, что полученные значения сходимости фактических и предсказанных значений на тестовой выборке указывают на необходимость дальнейшей доработки моделей в части процессинга входных данных и расширения возможных параметров моделей.

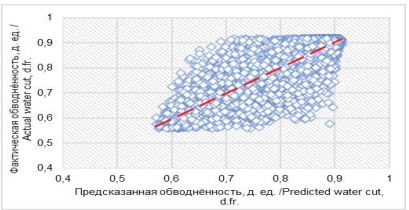


Рисунок 14. Значения предсказанной и фактической обводнённости в тестовой выборке
Figure 14. Predicted and actual values of water cut in test data

Ключевым параметром для оценки эффективности модели для данного исследования была принята метрика MAE, которая для тестовой выборки составила 3% при текущей средней обводнённости на уровне 85–87%. При прогнозе с помощью машинного обучения средняя ошибка получилась ниже значения ошибок при прогнозировании обводнённости на основе экспертного анализа.

Аналогичные расчёты для последующих месяцев показали схожие цифры по эффективности, что говорит о стабильности модели и воспроизводимости расчётов в разные временные периоды.

Выводы

1. Предложенный метод прогнозирования показателей позволяет уточнить и дополнить прогнозные параметры, рассчитанные с помощью ГГДМ или эмпирических зависимостей.

2. Анализ информации по геологии и разработке показал отсутствие статистически надёжных и достаточных для применения эмпирических зависимостей пусковой обводнённости новых скважин от геологических параметров для выбранного месторождения. Некоторые из зависимостей показали отрицательную корреляцию между обводнённостью и величиной остаточных запасов.

3. Анализ результатов интерпретации геофизических исследований новых скважин показал, что обводнённость коррелируется с насыщенностью, но имеющихся материалов недостаточно для масштабирования на все объекты месторождения для решения поставленной задачи.

4. Наличие обширного массива данных по историческим показателям работы скважин позволяет применить методы машинного обучения.

5. Разработанная комплексная модель показала приемлемые уровни эффективности на основе метрик классификации и регрессии, что говорит о её применимости для прогноза пусковой обводнённости проектных скважин.

6. Основываясь на полученных результатах, стоит отметить, что модели требуют доработки в части обработки входных данных для улучшения сходимости фактических и предсказанных значений.

7. Предложенный метод минимизирует вклад человеческого фактора и делает возможным автоматизацию прогноза обводнённости в программных решениях.

ДОПОЛНИТЕЛЬНО

Источник финансирования. Авторы заявляют об отсутствии внешнего финансирования при проведении исследования.

Конфликт интересов. Авторы декларируют отсутствие явных и потенциальных конфликтов интересов, связанных с публикацией настоящей статьи.

Вклад авторов. Все авторы подтверждают соответствие своего авторства международным критериям ICMJE (все авторы внесли существенный вклад в разработку концепции, проведение исследования и подготовку статьи, прочли и одобрили финальную версию перед публикацией). Наибольший вклад распределён следующим образом: Ибраев А.Е. – проведение исследования, написание рукописи, Камариденова Г.С. – проверка результатов, редактирование рукописи, Балуанов Б.А. – сбор, анализ, интерпретация данных исследования, Елемесов А.С. – концепция исследования, проверка результатов.

ADDITIONAL INFORMATION

Funding source. This study was not supported by any external sources of funding.

Competing interests. The authors declare that they have no competing interests.

Authors' contribution. All authors made a substantial contribution to the conception of the work, acquisition, analysis, interpretation of data for the work, drafting and revising the work, final approval of the version to be published and agree to be accountable for all aspects of the work. The greatest contribution is distributed as follows: Aktan Ye. Ibrayev – conduction of the study, writing of the manuscript; Gaukhar S. Kamaridenova – quality check, revision of the manuscript; Bakytzhan A. Baluanov – acquisition, analysis, interpretation of data for the study; Azamat S. Yelemessov – conception of the study, quality check.

СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ

1. *Tadjer A., Hong A., Bratvold R.* Bayesian Deep Decline Curve Analysis: A New Approach for Well Oil Production Modeling and Forecasting // SPE Res Eval & Eng. 2022. Vol. 25. P. 568–582. doi:10.2118/209616-PA.

2. *Gaskari R., Mohaghegh S.D., Jalali J.* An Integrated Technique for Production Data Analysis (PDA) With Application to Mature Fields // SPE Prod & Oper. 2007. Vol. 22. P. 403–416. doi:10.2118/100562-PA.

3. *Шевчук Т.Н., Кашников О.Ю., Мезенцева М.А., и др.* Прогноз показателей добычи из пластов баженовской свиты на основе статистических зависимостей и методов машинного обучения // PROНЕФТЬ. Профессионально о нефти. 2020. №4(18). С. 63–68. doi:10.7868/S2587739920040096.

4. *Колесов В.В., Курганов Д.В.* Расчёт рейтинга скважин-кандидатов при уплотняющем бурении с помощью машинного обучения промысловых данных (метод опорных векторов) // Вестник Самарского Государственного технического университета. Серия «Технические науки». 2019. №1 (61).

5. *Мартюшев Д.А., Пономарева И.Н., Захаров Л.А., Шадров Т.А.* Применение машинного обучения для прогнозирования пластового давления при разработке нефтяных месторождений // Известия Томского политехнического университета. Инжиниринг георесурсов. 2021. Т. 332, № 10. С. 140–149. doi:10.18799/24131830/2021/10/3401.

6. *Габитова С.И., Давлетбакова Л.А., Климов В.Ю., и др.* Методика прогнозирования темпов падения нефти проектных скважин на основе алгоритма машинного обучения // *PROНЕФТЬ. Профессионально о нефти.* 2020. №4(18). С. 69–74. doi:10.7868/S2587739920040102.

7. *Еникеев М.Р., Фазлытдинов М.Ф., Еникеева Л.В., Губайдуллин И.М.* Прогноз обводнённости на проектируемых к бурению скважинах методами машинного обучения // Сборник трудов ИТНТ-2019: V междунар. конф. и молодеж. шк. «Информ. технологии и нанотехнологии» (ITNT-2019). 2019. Т. 4. С. 434–444.

8. *Илюшин П.Ю., Галкин С.В.* Прогноз обводнённости продукции добывающих скважин пермского края с применением аналого-статистических методов // Вестник Пермского национального исследовательского политехнического университета. Геология, Нефтяная и газовая промышленность. 2011. Т. 10, №1. С. 76–84.

9. *Chicco D., Jurman G.* The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation // BMC Genomics. 2020. Vol. 6. doi:10.1186/s12864-019-6413-7.

REFERENCES

1. Tadjer A, Hong A, Bratvold R. Bayesian Deep Decline Curve Analysis: A New Approach for Well Oil Production Modeling and Forecasting. *SPE Res Eval & Eng.* 2022;25:568–582. doi:10.2118/209616-PA.

2. Gaskari R, Mohaghegh SD, Jalali J. An Integrated Technique for Production Data Analysis (PDA) With Application to Mature Fields. *SPE Prod & Oper.* 2007;22:403–416. doi:10.2118/100562-PA.

3. Shevchuk TN, Kashnikov OY, Mezentseva MA, et al. Production Forecast for Bazhen Formation Reservoirs on the Basis of Statistical Analyses and Machine Learning Techniques. *PRONEFT. Professional'no o nefti.* 2020;4(18):63–68. doi:10.7868/S2587739920040096. (In Russ).

4. Kolesov VV, Kurganov DV. Well Ranking for In-Fill Drilling Using Machine Learning with Production and Geological Data. *Vestnik of Samara State Technical University (Technical Sciences Series)*. 2019;1(61). (In Russ).

5. Martyushev DA, Ponomareva IN, Zakharov LA, Shadrov TA. Application of Machine Learning for Forecasting Formation Pressure in Oil Field Development. *Bulletin of the Tomsk Polytechnic University. Geo Assets Engineering.* 2021;332(10):140–149. doi:10.18799/24131830/2021/10/3401. (In Russ).

6. Gabitova SI, Davletbakova SA, Klimov VY, et al. A new method of decline curve forecasting for project wells on the base of machine learning algorithms. *PROneft. Professionally about Oil.* 2020;4:69–74. doi:10.7868/S2587739920040102. (In Russ.).

7. Enikeev MR, Fazlytdinov MF, Enikeeva LV, Gubaidullin IM. The apply of machine learning methods for water cut prediction on the projected wells. *V International Conference on Information Technology and Nanotechnology (ITNT-2019).* 2019;4:434–444. (In Russ.).

8. Ilyushin PY, Galkin SV. Forecast water cut production wells perm with the A-statistical methods. *Vestnik of Perm National Research Polytechnic University. Geology, Oil and Gas Industry.* 2011;10:1. P. 76–84. (In Russ).

9. Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics.* 2020. Vol. 6. doi:10.1186/s12864-019-6413-7.

ИНФОРМАЦИЯ ОБ АВТОРАХ

***Ибраев Актан Ермекович**
e-mail: *a.ibrayev@niikmg.kz.*
Камариденова Гаухар Сериковна
e-mail: *g.kamaridenova@niikmg.kz.*
Балуанов Бакытжан Айтуарович
e-mail: *b.baluanov@niikmg.kz.*
Елемесов Азамат Серикович
e-mail: *ayelemessov@niikmg.kz.*

AUTHORS' INFO

***Aktan Ye. Ibrayev**
e-mail: *a.ibrayev@niikmg.kz.*
Gaukhar S. Kamaridenova
e-mail: *g.kamaridenova@niikmg.kz.*
Bakytzhan A. Baluanov
e-mail: *b.baluanov@niikmg.kz.*
Azamat S. Yelemessov
e-mail: *ayelemessov@niikmg.kz.*

*Автор, ответственный за переписку/Corresponding Author

 DOI: 10.54859/kjogi108642