МРНТИ 28.23
УДК 519.6

10.51889/2959-5894.2023.83.3.015

Б.К. Асилбеков [1,2], Н.Е. Қалжанов [3], Д.Ә. Болысбек [1,3], К.Ш. Узбекалиев [1]*

[1]Сатбаев Университет, г. Алматы, Казахстан
[2]ТОО «KBTU BIGSoft», г. Алматы, Казахстан
[3]Казахский национальный университет имени аль-Фараби, г. Алматы, Казахстан
[*]e-mail: assilbekov.b@gmail.com

ИССЛЕДОВАНИЕ ЭФФЕКТИВНОСТИ АЛГОРИТМОВ МАШИННОГО ОБУЧЕНИЯ НА ОСНОВЕ ДАННЫХ РАЗЛИЧНЫХ ГОРНЫХ ПОРОД

Аннотация

Проницаемость является важным свойством пористой среды, и ее определение является актуальной задачей. В статье изучается эффективность алгоритмов машинного обучения, такие как RF, GB, SV, Lasso, k-NN и GP, при прогнозировании проницаемости различных пород. В качестве признаков использованы радиус пор, радиус горловины, координационное число, пористость, удельная площадь поверхности, извилистость и проницаемость. Было изучено влияние соотношения обучающего и тестового набора данных (70/30 и 80/20) и количества признаков на производительность алгоритмов. Результаты показали, что алгоритм RF являлся наиболее подходящим для прогнозирования проницаемости с высокой достоверности. Наибольший коэффициент достоверности прогноза составил R^2=0.83, и он был получен при использовании 5 признаков. Алгоритм GB также показал хорошую прогнозирующую способность проницаемости, хотя он выбирал практически одного признака (пористости) как важным. Наибольший коэффициент для него составил R^2=0.73 при 80/20. Результаты также показали, что все алгоритмы, кроме RF, предсказали существенно завышенные минимальные проницаемости. А также, все алгоритмы, кроме SV и k-NN, предсказали среднее значение проницаемости с наименьшими погрешностями.

Ключевые слова: машинное обучение, прогноз проницаемости, микрокомпьютерная томография, мини-образец, поромасштабное моделирование.

Аңдатпа

Б.К. Асилбеков [1,2], Н.Е. Қалжанов [3], Д.Ә. Болысбек [1,3], К.Ш. Узбекалиев [1]
[1]Сәтбаев Университеті, Алматы қ., Қазақстан
[2]«KBTU BIGSoft» ЖШС, Алматы қ., Қазақстан
[2]Әл-Фараби атындағы Қазақ ұлттық университеті, Алматы қ., Қазақстан

ӘРТҮРЛІ ТАУ ЖЫНЫСЫНЫҢ ДЕРЕКТЕРІ НЕГІЗІНДЕ МАШИНАЛЫҚ ОҚУ АЛГОРИТМДЕРІНІҢ ТИІМДІЛІГІН ЗЕРТТЕУ

Өткізгіштік кеуекті орталардың маңызды қасиеті болып табылады, ал оны анықтау актуалды мәселе болып табылады. Бұл жұмыста әртүрлі тау жыныстарының өткізгіштігін болжауда RF, GB, SV, Lasso, k-NN және GP сияқты машиналық оқыту алгоритмдерінің тиімділігі зерттеледі. Белгілер ретінде кеуек радиусы, кеуек мойнының радиусы, координация саны, кеуектілік, беттің меншікті ауданы, бұралу және өткізгіштік қолданылды. Оқу және тестілеу деректер жинағының арақатынасының (70/30 және 80/20) және белгілер санының алгоритмдердің өнімділігіне әсері зерттелді. Нәтижелер RF алгоритмі өткізгіштікті жоғары сенімділікпен болжау үшін ең қолайлы екенін көрсетті. Болжаудың ең жоғары сенімділік коэффициенті R^2=0.83 болды және ол 5 белгі үшін алынды. GB алгоритмі де өткізгіштік үшін жақсы болжау қабілетін көрсетті, дегенмен ол тек бір белгіні ғана (кеуектілікті) маңызды деп тапты. Ол үшін ең жоғары коэффициент 80/20 кезінде R^2=0.73 болды. Нәтижелер сонымен қатар RF алгоритмінен басқа барлық алгоритмдер өткізгіштіктің төмен мәндерін айтарлықтай асыра болжағанын көрсетті. Сондай-ақ, SV және k-NN алгоритмінен басқа барлық алгоритмдер өткізгіштіктің орташа мәнін ең аз қателермен болжады.

Түйін сөздер: машиналық оқыту, өткізгіштікті болжау, микрокомпьютерлік томография, мини-үлгілер, кеуекті масштабта модельдеу.

Abstract

STUDY OF THE EFFICIENCY OF MACHINE LEARNING ALGORITMS BASED ON DATA OF VARIOUS ROCKS

Assilbekov B.K.[1,2], Kalzhanov N.E.[3], Bolysbek D.A.[1,3], Uzbekaliyev K.Sh.[1]

[1]Satbayev University, Almaty, Kazakhstan

[2]JSC «KBTU BIGSoft», Almaty, Kazakhstan

[2]Al-Farabi Kazakh National University, Almaty, Kazakhstan

Permeability is an important property of porous media, and its determination is relevant. This paper studies the effectiveness of machine learning algorithms (RF, GB, SV, Lasso, k-NN and GP) in permeability prediction. Pore and throat radius, coordination number, porosity, specific surface area, tortuosity and permeability were used as features. The influence of the training and testing dataset ratios and the number of features on the algorithm's performance was studied. The results showed that the RF was the most suitable for predicting permeability with high confidence. The highest R^2 was 0.83, and it was obtained for 5 features. GB also showed good predictive ability, although it selected only porosity as important feature. The highest R^2 for it was 0.73 at 80/20. The results also showed that all algorithms except RF significantly overpredicted low permeabilities. And also, all algorithms, except SV and k-NN, predicted the average permeability value with the smallest errors.

Keywords: machine learning, permeability prediction, microcomputed tomography, sub-sample, pore-scale modeling.

Введение

Абсолютная проницаемость является важной макроскопической транспортной характеристикой пористой среды, от которой зависят добыча углеводородов при разработке месторождений, производительность фильтров при очистке воздуха в помещениях, сепарации газожидкостных систем и в каталитических системах и т.д. Проницаемость обычно определяют в лабораторных условиях экспериментальным путем с помощью специальных оборудовании. Лабораторные измерения обычно длятся не мало времени и являются дорогостоящими. Поэтому ее определение альтернативными путями на основе имеющихся аналитических и экспериментальных данных о пористой среде является актуальной задачей.

С развитием методов машинного обучения, они стали применяться для анализа данных и прогнозирования важных характеристик во многих сферах таких как медицина [1], экономика [2, 3], геофизика [4-6] и т.д. При прогнозировании абсолютной проницаемости используют изображения реальных горных пород или синтетических пористых сред, полученные с помощью микрокомпьютерной томографии [5, 7, 8] и цифровых данных, извлеченные тем или иным способом из этих изображений [9, 10]. А также используются каротажные данные скважин для прогнозирования абсолютной проницаемости [6, 11].

Двумерные изображения синтетических пористых сред в совмещении с решеточным методом Больцмана использованы для прогнозирования пористости, извилистости и абсолютной проницаемости в работе [7]. Синтетические пористые среды получены путем случайного распределения квадратных частиц, означающие твердого скелета породы, в квадратной области. Авторы использовали сверточные нейронные сети для определения взаимосвязи между структурой и основными характеристиками синтетической пористой среды. А с помощью решеточного метода Больцмана находили поля течения для дальнейшего вычисления извилистости и проницаемости. Они утверждали, что сверточные нейронные сети показали высокую прогнозирующую способность основных характеристик. Подобное исследование проведено в [8], где было также сказано, что сверточные нейронные сети показали высокую произвоительность в прогнозировании абсолютной проницаемости. Тембли и др. [5] использовали изображения свыше 1000 образцов горной породы, отсканированные с помощью микрокомпьютерной томографии с высоким пространственным разрешением, для построения прогнозной модели проницаемости на основе методов машинного обучения и глубокого обучения. Они показали, что модели проницаемости на основе методов машинного обучения и глубокого обучения предсказали проницаемость с достоверностью 88 и 91%, соответственно. А также показали, что использование методов исскуственного интеллекта позволяет сократит времени расчета проницаемости на три порядка по сравнению с традиционными методами. Прогнозирование абсолютной проницаемости породы на основе каротажных данных скважины с использованием методов машинного обучения приведены в [6, 11].

В настоящей статье изучается прогноз абсолютной проницаемости на основе данных различных пород, истинная проницаемость которых сильно отличается от их среднестатистической

проницаемости. Анализ литератур показал, что при исследованиях эффективности методов машинного обучения, в основном использованы данные синтетических пористых сред и реальных горных породы с несильно изменяющимися характеристиками. При изучении рассматривались 6 методов машинного обучения, эффективность которых была изучена при разных соотношения обучающего и тестового набора данных.

Материалы и методы

Материалы

Как правило, методы машинного обучения используют набор данных с различными признаками в качестве входных данных. С этой целью были подготовлены цифровые модели 266 мини-образцов из готовых моделей различных пород. Каждый мини-образец имеет 7 характеристик в качестве признака (входных) данных. Готовые цифровые модели большего размера 8 различных реальных горных пород и 1 искусственной породы (упакованная из частиц песка) были взяты из библиотеки исследовательской группы профессора М. Бланта из Имперского Колледжа Лондона с открытым доступом, которые показаны на рис. 1. Отметим, что вышесказанные цифровые модели уже были отфильтрованы и отсегментированы. Как видно из рис. 1 поровое пространство песчаных пород образованы пустотами между гранул песка и имеют более однородное распределение по всему образцу по сравнению с строениями карбонатных пород. Эти модели были получены в результате сканирования пород с помощью микрокомпьютерного томографа с пространственным разрешением около 3 мкм.

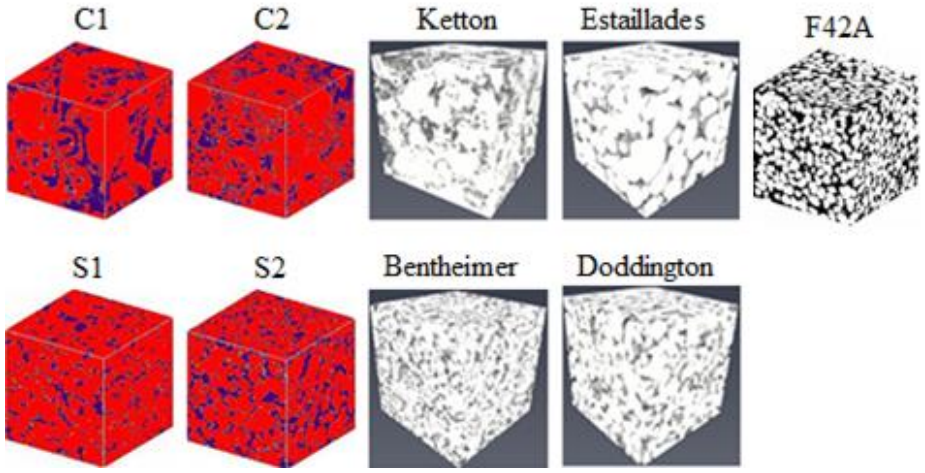

Рисунок 1. Цифровые модели образцов различных пород

С целью увеличения количества входных данных в методы машинного обучения, полученные цифровые модели пород были разделены на разные куски (мини-образцы) меньшего размера (см. рис. 2). Таким образом, всего было подготовлено 266 мини-образцов размеры которых меняется от 0,125 мм до 3 мм.

Проницаемость мини-образцов приведена в табл. 1. Как видно из этой таблицы, рассмотренные образцы имеют существенно отличающаяся проницаемость. Отдельно можно выделить образец упакованного песчаника, который имеет относительно высокую проницаемость. Как видно из табл. 1, карбонатные породы имеют относительно неоднородная и низкая проницаемость по сравнению с песчаными породами. Сравнение минимального, максимального и среднего значения проницаемости говорить, что в карбонатные мини-образцы в основном имеют проницаемость, близкая к минимальной, тогда как песчаные мини-образцы кроме Bentheimer и Doddington имеют более однородные распределения проницаемости. А мини-образцы песчаной упаковки F42A практически имеют одинаковую проницаемость.

Таблица 1. Проницаемость мини-образцов

Образец/ тип породы	*Карбонатный*				*Песчаный*				*Песчаная упаковка*
	C1	*C2*	*Ketton*	*Estaillades*	*S1*	*S2*	*Bentheimer*	*Doddington*	*F42A*
Минимум	*0,097*	*0,012*	*0,0033*	*0,0019*	*0,45*	*1,81*	*0,04*	*0,001*	*7,42*
Максимум	*4,95*	*4,24*	*16,05*	*85,4*	*3,59*	*7,84*	*15,1*	*9,24*	*104,3*
Среднее	*1,65*	*0,87*	*3,08*	*4,09*	*1,36*	*4,44*	*2,53*	*2,16*	*66,0*

Основным и сложным этапом всего процесса был сбор данных из уже подготовленных мини-образцов, так как большинство параметров пористой среды, такие как средний радиус пор, средний радиус горловины пор, средняя извилистость, среднее координационное число и абсолютная проницаемость, будут найдены только при моделировании течения жидкости сквозь эти пористые среды. Набор данных содержит пористость (ϕ), средний радиус пор (r_p, в мкм), средний радиус горловины пор (r_t, в мкм), извилистость (τ), координационное число (N_c), удельную площадь поверхности пор (S_s, в 1/мкм) и абсолютную проницаемость (k, в мкм^2) для каждого мини-образца. В настоящей работе течение жидкости в пористых средах было смоделировано при помощи поросетевого моделирования с использованием специального программного обеспечения Avizo, в процессе которого сначала строится поровая сеть микроструктуры мини-образцов горных пород на основе выделенного порового пространства (рис. 3), затем на ней будет смоделировано течение самой жидкости на основе закона сохранения массы.

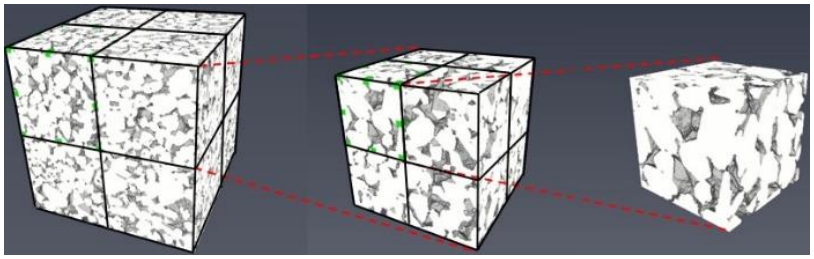

Рисунок 2. Разбиение образцов на более мелкие мини-образцы

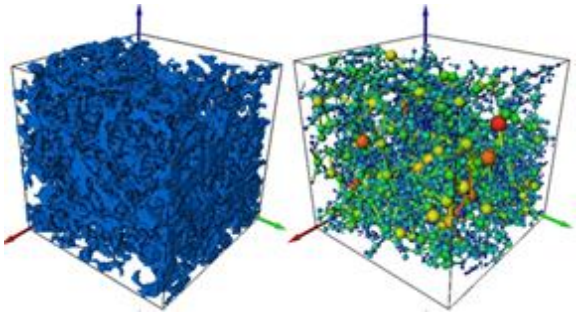

Рисунок 3. Поровое пространство и поровая сеть мини-образца

Поровая сеть – это совокупность отдельных пор и горловин пор, которые в свою очередь представляются в виде сфер и цилиндров. Построение поровой сети основано на алгоритме максимальных шаров, в котором поры заменяются сферами, а горловины пор – цилиндрами. Соответственно, радиусы сфер и цилиндров являются радиусами пор и горловины пор. Поровая сеть дает информацию о распределении пор и горловины пор по их размерам, соответственно мы можем найти средние радиусы пор и горловины пор. Поровая сеть также позволяет определить количество связей конкретной поры с другими соединенными с ней порами, а значит это дает о координационном числе сети. Удельная площадь поверхности пор является одним из важных параметров пористой среды, так как она влияет на абсолютную проницаемость и степени растворения породы разными кислотными составами.

Основным макроскопическим параметром пористой среды является ее проницаемость – она показывает способность пористой среды пропускать через себя жидкости и зависит от многих факторов, такие как пористость, извилистость, удельная площадь поверхности пор и т.д. Но, до сих пор конкретно не установлено какие параметры больше всего влияют на нее. Абсолютная проницаемость среды определяется по закону Дарси. С цифровой модели каждого мини-образца были извлечены 7 его параметров, такие как средний радиус пор, средний радиус горловины пор, извилистость, координационное число и абсолютная проницаемость которые приведены на рис. 4. В таблице на рис. 3 каждая строка означает набор данных из 7 параметров (не включая наименование) для каждого мини-образца. Эти данные являются исходной информацией при их анализе с помощью методов машинного обучения. Этот набор данных имеет целевую переменную: 'k' и входные переменные: 'r_p', 'N_c', 'r_t', 'ϕ', 'S_s', 'τ'. Статистика характеристик использованных мини-образцов приведена в табл. 2.

1 2	Наименование образца и его кусков	Средний радиус пор [мкм]	Среднее координационное число	Средний радиус горловины пор [мкм]	Средняя пористость	Удельная площадь поверхности [1/мкм]	Извилистость	Проницаемость [мкм2]
3	Full Size C1	31,382	5,00	15,064	0,21	0,0503	1,809	1,27736824300
4	C1-1	18,235	4,29	10,160	0,14	0,0234	1,596	0,09680006031
5	C1-2	19,583	5,08	11,921	0,23	0,0265	1,779	4,05577314400
6	C1-3	19,745	2,81	13,590	0,16	0,0250	2,053	0,38747577120
7	C1-4	30,996	4,03	14,737	0,17	0,0234	1,569	0,72855039900
8	C1-5	29,190	3,69	14,091	0,19	0,0248	1,337	1,06907533700
9	C1-6	30,269	4,52	13,718	0,26	0,0301	1,675	0,74851358560
10	C1-7	21,431	4,82	11,523	0,18	0,0251	1,800	1,50289427500
11	C1-8	34,897	5,10	17,082	0,33	0,0334	1,646	4,95092404800
12	Full Size C2	47,434	4,13	20,461	0,14	0,0138	1,888	0,08093792610
13	C2-1	35,080	3,50	17,130	0,24	0,0181	1,693	0,23001029520

Рисунок 4. Набор данных с цифровых моделей мини-образцов

Таблица 2. Статистика характеристик мини-образцов

Наименование параметра	*Минимальное значение*	*Максимальное значение*	*Среднее значение*	*Стандартное отклонение*
r_p *(мкм)*	*11,70*	*188,32*	*77,33*	*26,73*
N_c	*1,71*	*7,48*	*4,13*	*1,18*
r_t *(мкм)*	*10,11*	*73,85*	*27,27*	*11,08*
ϕ	*0,02*	*0,34*	*0,19*	*0,067*
S_s *(1/мкм)*	*0,0014*	*0,050*	*0,013*	*0,005*
τ	*1,05*	*2,66*	*1,51*	*0,19*
k *(мкм^2)*	*0,001*	*104,3*	*9,38*	*21,98*

Методы исследования

При изучении данных применялись такие методы как алгоритм случайного леса, повышения градиента, опорных векторов, ЛАССО, К-ближайших соседей, гауссовского процесса.

Метод случайного леса (Random Forest, RF) – это алгоритм машинного обучения, основанный на ансамбле решающих деревьев. В данном методе каждое дерево решает задачу независимо от других деревьев, в конце ответы всех деревьев усредняются. RF использует множество параметров для управления оптимизацией решением такие как n_estimators, max_depth, max_features и т.д. [11]. Использованные в настоящей работе их значения приведены в табл. 3. Мы подбирали наилучшие параметры для этого метода используя библиотеку GridSearchCV [12], которая поможет упростить перебор параметров.

Первым важным параметром в методе RF является n_estimators – означающее количество деревьев, чем больше деревьев, тем лучше качество, но время настройки и работы RF также пропорционально увеличиваются. Вторым важным параметром является max_features при увеличении которого увеличивается время построения леса, а деревья становятся «более однообразными». Третий параметр – max_depth (глубина дерева) при увеличении которого резко возрастает качество обучении. При использовании неглубоких деревьев (т.е. при малых max_depth) изменение параметров, связанных с ограничением числа объектов в листе и для деления, не приводит к значимому эффекту.

Таблица 3. Методы и их параметры управления

Метод	*Параметры управления*	*Значение*
Случайного леса	*max_depth*	*41*
	max_features	*1*
	n_estimators	*15*
	bootstrap	*True*
	criterion	*MSE*
Повышения градиента	*n_estimators*	*19*
	learning_rate	*0.9*
	max_depth	*1*
	criterion	*friedman_mse*
Опорных векторов	*kernel*	*linear*
	epsilon	*0.5*
	C	*10*
	gamma	*1e-07*
ЛАССО	*alpha*	*0.01*
	max_iter	*11*
К-ближайших соседей	*n_neighbors*	*86*
	p	*1*
	weights	*distance*
Гауссовского процесса	*alpha*	*0.001*
	kernel	*DotProduct (sigma_0=0.1)*

Метод повышения градиента (Gradient Boosting, GB) – это метод преобразования слабообученных моделей в хорошообученные. Этот метод основан на минимизацию функции потери с помощью градиентного спуска [12]. Из-за схожести с методом случайного леса, данный метод имеет практически такие же параметры управления как в методе случайного леса. Параметры управления и их значения приведены в табл. 3.

Метод (регрессор) опорных векторов (Support Vector, SV) является версией метода опорных векторов для использования в задачах регрессии. Данный регрессор основан на нахождение непрерывной (линейной или нелинейной) функции, которая максимально аппроксимирует входные данные внутри заданной трубки с достаточно малым диаметром на основе опорных векторов [4]. Данный метод также имеет свои параметры управления, которые приведены в табл. 3.

ЛАССО (Lasso) – метод, изначально предназначенный для линейной регрессии, который обеспечивает выбора переменной и регуляризацию для повышения точность прогноза. Параметры управления данного метода приведены в табл. 3.

Метод К-ближайших соседей (k-Nearest Neighbors, k-NN) – метод решения задач классификации и регрессии, основанный на поиске ближайших объектов с известными значениями целевой переменной.

Регрессия гауссовского процесса (Gaussian Process, GP) является универсальным непараметрическим методом обучения с учителем, разработанным для решения *регрессии [13]*. Алгоритмы Lasso, k-NN и **GP** рассматриваются в целях сравнения результатов прогноза с помощью других алгоритмов с их результатами.

Результаты и их обсуждения

Собранные данные были анализированы с помощью вышеприведенных методов машинного обучения. Сначала покажем связей между признаками (параметрами пористых сред), которая визуализирована в виде корреляционной матрицы (рис. 5). Данная матрица показывает насколько хорошей или плохой связи имеют каждая пара признаков – чем больше коэффициент в матрице, тем выше корреляция между выбранными признаками. Коэффициент, равный 1 означает идеальную корреляцию. Для отображения корреляционной матрицы использовали библиотеку seaborn. Seaborn – это, по сути, более высокоуровневое API на базе библиотеки matplotlib. Seaborn содержит более адекватные настройки оформления графиков по умолчанию. Также в библиотеке есть достаточно сложные типы визуализации, которые в matplotlib потребовали бы большего количества кода. Как видно из рис. 5, все входные переменные (признаки), кроме удельной площади поверхности и извилистости, имеют высокую корреляцию с целевой переменной.

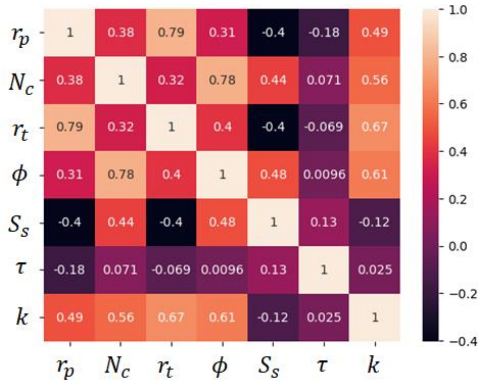

Рисунок 5. Корреляционная матрица для визуализации связей между признаками

Влияние разделения данных на обучение/тест на производительность алгоритмов

После того как выбрали алгоритма машинного обучения, первоочередной задачей является определение списка входных данных (признаков), которые являются наиболее важными при построении прогнозной модели [14]. Для этого у каждого алгоритма есть свойство feature_importances_, с помощью которого можно посмотреть вес (важность) каждого признака в итоговой прогнозной модели. Важность разделения набора входных данных на обучения и тестирования заключается в том, что обучающий набор содержит известные выходные данные, на которых модель учится. В настоящей статье набор данных был разделен в соотношении 70/30 и 80/20.

Результаты прогнозирования проницаемости по рассмотренным выше алгоритмам машинного обучения приведены на рис. 6-8 и табл. 4. Рис. 6 показывает какие из 6 независимых входных данных были выбраны наиболее важными при построении прогнозной модели проницаемости по разным алгоритмам машинного обучения. Отметим, что в целях сравнения здесь и далее важность признака нормализована по максимальному значению важности для каждого алгоритма. Как показывают диаграммы, алгоритмы RF, SV и GP оказались устойчивыми на изменение доли обучающего набора данных – при использовании этих алгоритмов количество важных признаков и их важность практически не менялись. RF и GP выбрали важными 5 из 6 признаков, тогда как SV посчитал наиболее важным всего 2 признаков. Увеличение доли обучающего набора данных существенно повлияло на выбор важных признаков алгоритмами Lasso и k-NN. Если алгоритм Lasso при 70/30 выделил всего 2 признака важным, то для 80/20 важных признаков увеличились до 5, а для k-NN количество важных признаков увеличилось с 3 при 70/30 до 6 для 80/20. Все это показывает чувствительность рассмотренных алгоритмов к количеству обучающих данных.

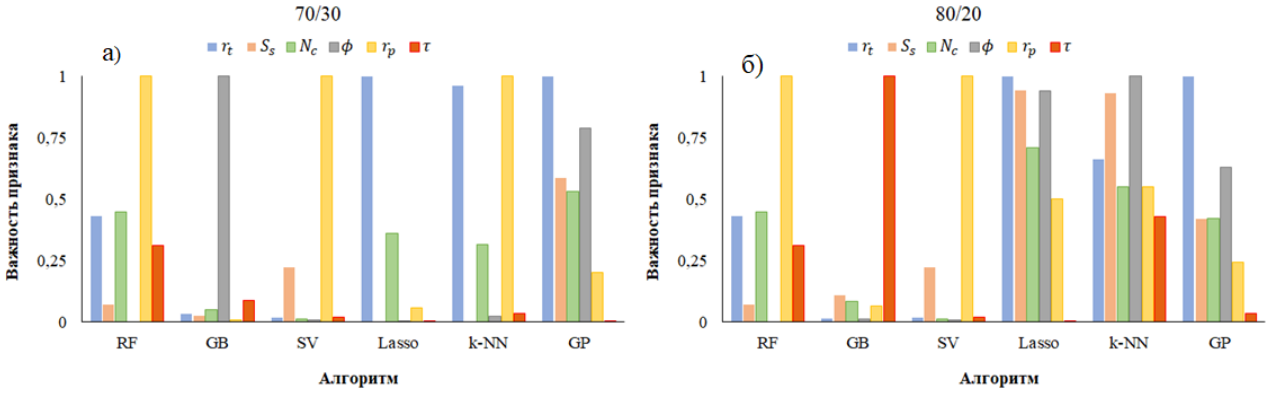

Рисунок 6. Важность признака при разбиении данных в соотношении 70/30 (а) и 80/20 (б)

После того как рассмотренные алгоритмы были обучены проверили их прогнозирующую способность на тестовых данных путем сопоставления их результатов с истинными данными. Результаты сопоставления показаны на рис. 7, где символы разной формы соответствуют разным алгоритмам, а сплошная линия означает идеальную корреляцию между предсказанным и истинным

значениями проницаемости. Здесь и далее, на всех графиках на оси абсцисс и ординат расположены истинная и модельная проницаемости, соответственно.

Как видно из графиков, все алгоритмы машинного обучения предсказали завышенные значения проницаемости, особенно в диапазоне <1 μм^2, а среднее и максимальное значения ложатся близко к линии 1:1, что показывает близкое к истинным значениям прогноза проницаемости с помощью построенных моделей прогноза (рис. 7а-г). Аль-Халифа и др. [9] при изучении карбонатных пород также обнаружили, что качество прогноза методами машинного обучения снижается при низких значениях проницаемости. При увеличении доли обучающего набора данных, практически все алгоритмы предсказали значения проницаемости, которые относительно близко находятся к линии 1:1 при 80/20 (рис. 7в, г) по сравнению с случаем 70/30 (рис. 7а, б). Это подтверждается рис. 8, где приведено количественное сравнение прогнозирующей способности рассмотренных алгоритмов машинного обучения в виде коэффициента корреляции (R^2) между предсказанной и истинной проницаемостями. Как заметно из этого рисунка, корреляция между предсказанной и истинной проницаемостей улучшилась для всех алгоритмов с ростом количества обучающего набора данных.

Относительно высокие коэффициенты корреляции наблюдались у GP (0,75), Lasso (0,77) k-NN (0,784), тогда как для остальных алгоритмов данный коэффициент составил 0,687, 0,723 и 0,732, соответственно для SV, RF и GB. Очевидно, что чем больше коэффициент R^2, тем выше достоверность прогноза, т.е. предсказанная проницаемость ближе к истинной проницаемости. Хотя все алгоритмы предсказали проницаемости, которые высоко коррелируют с истинной проницаемостью, некоторые алгоритмы предсказали отрицательные проницаемости, что неприемлемо. В табл. 4 приведены минимальное, максимальное и среднее значения истинной проницаемости и проницаемости, предсказанные с помощью рассмотренных алгоритмов машинного обучения. Как видно из табл. 4, при прогнозе с помощью алгоритмов SV, Lasso и GP были получены отрицательные минимальные проницаемости. Юн [3] также получил отрицательные значения по алгоритму GB при прогнозе ВВП Японии за 2001-2018 годы. Хотя при помощи GB и k-NN получились положительные значения, они на три порядка выше истинного минимального значения. Самое близкое значение к истинному получилось только у алгоритма RF для обоих случаев разбиения набора входных данных.

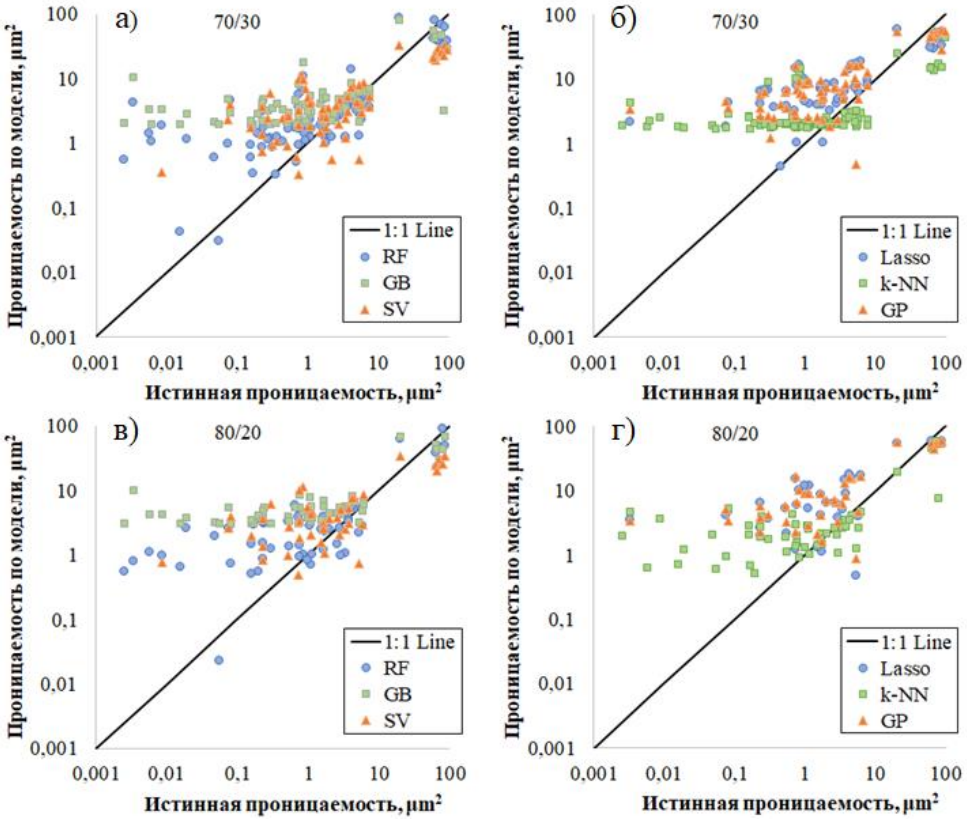

Рисунок 7. Прогнозированная с помощью разных алгоритмов проницаемость в сравнении с истинной проницаемости при разбиении данных в соотношении 70/30 (а) и 80/20 (б)

Таблица 4. Статистика предсказанных проницаемостей при 70/30 и 80/20

Параметр	*Обучение/Тест*	*Истинная*	*RF*	*GB*	*SV*	*Lasso*	*k-NN*	*GP*
Минимум	*70/30*	*0,00249*	*0,03128*	*1,94*	*-10,95*	*-16,37*	*1,66*	*-17,10*
	80/20	*0,00249*	*0,02273*	*3,03*	*-10,79*	*-16,32*	*0,52*	*-18,27*
Максимум	*70/30*	*95,91*	*87,81*	*80,94*	*33,44*	*58,44*	*51,63*	*56,39*
	80/20	*87,07*	*91,15*	*68,40*	*34,18*	*59,12*	*58,35*	*59,36*
Среднее	*70/30*	*9,31*	*8,22*	*7,96*	*4,59*	*7,64*	*5,57*	*8,18*
	80/20	*8,37*	*7,72*	*9,78*	*4,65*	*8,60*	*6,30*	*8,33*

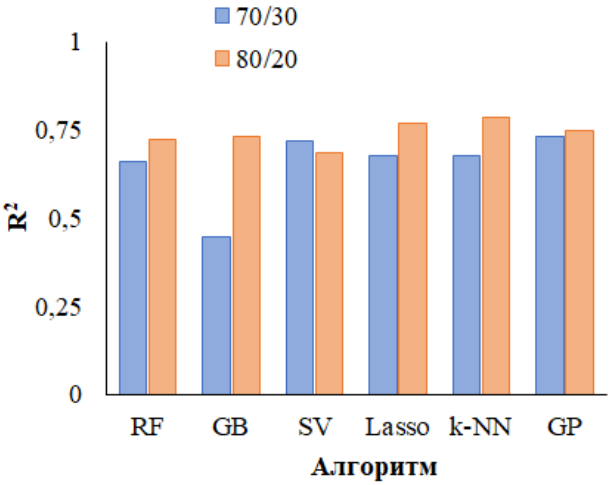

Рисунок 8. Коэффициент достоверности прогноза проницаемости на тестовых данных при разбиении данных в соотношении 70/30 (синие столбцы) и 80/20 (оранжевые столбцы)

Максимальная проницаемость, предсказанная по алгоритмам SV, Lasso, k-NN и GP имеет существенную разницу от истинной проницаемости. Алгоритм GB предсказал проницаемость, которая находится в пределах 20% погрешности от истинной. Максимальная проницаемость, предсказанная по алгоритму RF, достаточно близко находится к истинной проницаемости, относительная погрешность прогноза составляет 8,5 и 4,5%, соответственно при 70/30 и 80/20. Все алгоритмы кроме SV и k-NN предсказали близкие к истинной проницаемости. Это может быть связано с тем, что методы машинного обучения в основном базированы на статистике данных, которые достоверно предсказывают средние значения.

Влияние количества признаков на производительность алгоритмов

А также алгоритмы были обучены на наборе данных, в которых количества признаков менялись от 3 до максимального (т.е. до 6), и составили 'r_p', 'N_c', 'r_t' (3 признака), 'r_p', 'N_c', 'r_t', 'ϕ' (4 признака), 'r_p', 'N_c', 'r_t', 'ϕ', 'S_s' (5 признаков) и 'r_p', 'N_c', 'r_t', 'ϕ', 'S_s', 'τ' (6 признаков), соответственно. При этом соотношение обучающего и тестового набора данных не менялось и составило 70/30 во всех случаях. Целью изменения количества признаков в наборе входных данных является проверка чувствительности модели прогноза проницаемости к количеству независимых переменных, от которых проницаемость могла иметь функциональную зависимость. Очевидно, что было бы лучше если проницаемость будет рассчитана по формуле (модели), которая использует минимальные, но важные параметры (признаки) при этом не теряя точность. Например, эмпирическая формула Козени-Кармана [15], которая позволяет найти абсолютную проницаемость, использует всего лишь три признака (пористость, удельная площадь поверхности и извилистость) и одной параметрической константы.

Какие признаки были выбраны рассмотренными алгоритмами машинного обучения при изменении их количество показаны на рис. 9. Как заметно из этого рисунка, алгоритмы SV, Lasso и k-NN выбрали одни и те же признаки хотя их количество было разным в каждом случае (рис. 9в, г, д). При этом

важность выбранных признаков практически была одинакова. Остальные алгоритмы постарались выбрать практически всех признаков, когда их общее количество менялось (рис. 9а, б, е). Кроме того, важность этих признаков были разными у алгоритмов RF, GB и GP. GP выбрал в качестве самого важного признака среднего радиуса горловины пор r_t не зависимо от количества признаков, а важность остальных признаков была заметна с ростом количества признаков (рис. 9е). А алгоритм GB считал, что наиболее важным признаком является пористость мини-образца ϕ (рис. 9б). В случае с RF наиболее важными признаками считались пористость ϕ и координационное число N_c (рис. 9а).

Распределение предсказанной проницаемости по разным алгоритмам в сравнении с ее истинным значением показано на рис. 10. Можно сказать, что все алгоритмы предсказали визуально близкие к истинному значению проницаемости. Однако, для всех алгоритмов характерно завышенное предсказанные минимальные проницаемости, а максимальное и среднее значения были существенно близко прогнозированы к истинной проницаемости (распределение значении вокруг сплошной линии). Среди всех алгоритмов, RF предсказал наиболее близкие к истинной проницаемости в диапазоне минимальных проницаемостей (0,001-1 μм^2). Как видно из рис. 10, с ростом количества признаков, предсказанная проницаемость по алгоритмам RF (синие круги на рис. 10а, в, д и ж) и GP (оранжевые треугольники на рис. 10б, г, е и з) сблизилась к истинной проницаемости, так как эти методы выбрали наибольшее количество важных признаков (рис. 9а, е).

Параметр, характеризующий насколько хорошо или плохо спрогнозировал проницаемость мини-образцов тот или иной алгоритм показан на рис. 11. На этом рисунке распределены коэффициенты достоверности прогноза, когда количество признаков в наборе входных данных растет. А в табл. 5 приведены минимальное, максимальное и среднее значения проницаемости, полученные с помощью рассмотренных алгоритмов машинного обучения в зависимости от количества важных признаков.

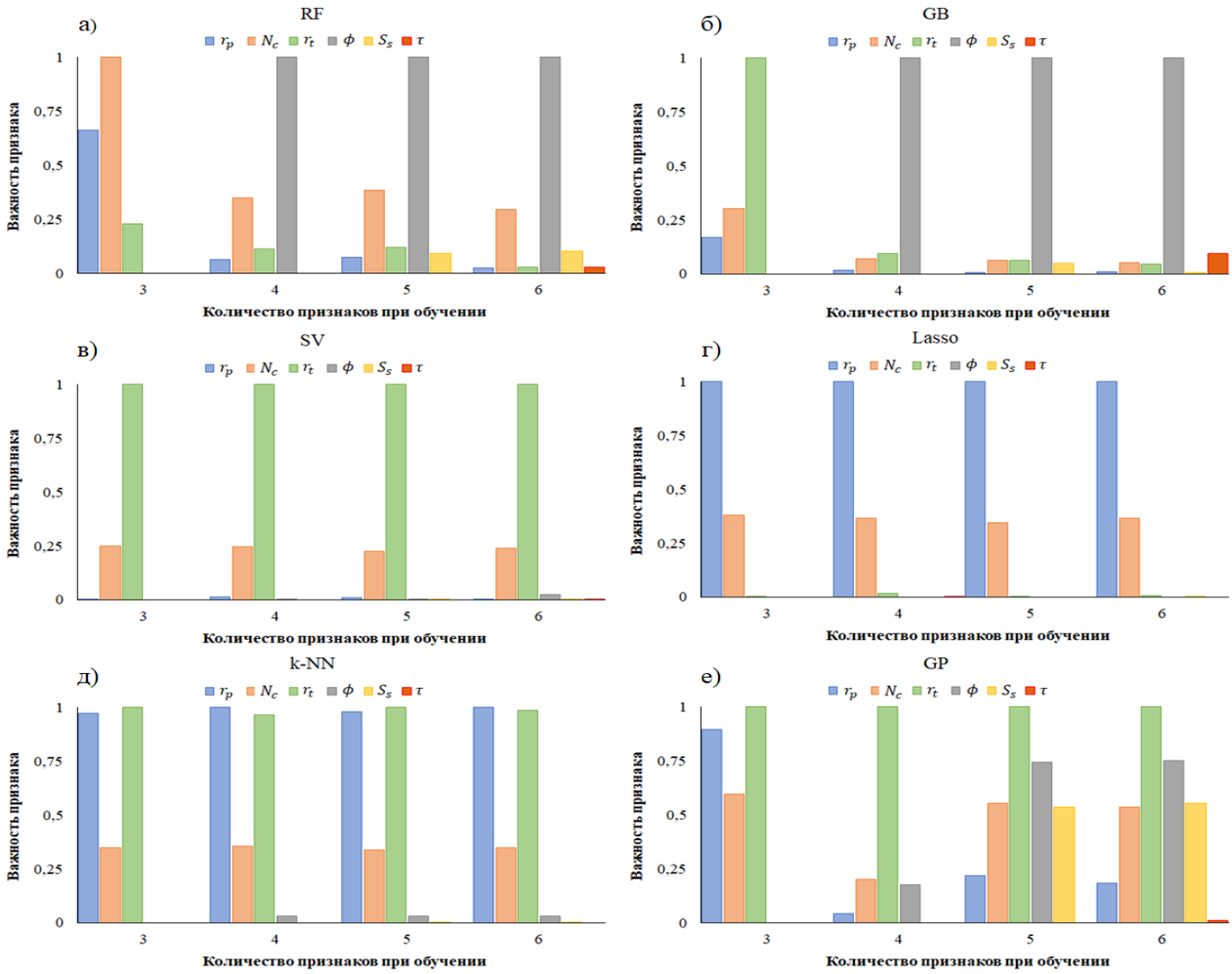

Рисунок 9. Важность признака при их различном количестве в наборе данных для разных алгоритмов машинного обучения

Как заметно на рис. 11, у алгоритмов SV, Lasso и k-NN коэффициент достоверности прогноза практически не изменился с ростом количества признаков в наборе входных данных хотя значение этого коэффициента не низкий. Это связано с выбором практически одинаково количества признаков по их важности (см. рис. 9в, г, д). С другой стороны, R^2 у остальных алгоритмов имеет чувствительность к изменению количества важных признаков. У алгоритмов RF и GP имеется тенденция роста R^2 с ростом количества признаков, т.е. эти алгоритмы точнее предсказывают значения проницаемости при включении большего свойств мини-образцов в набор входных данных. А у алгоритма GB наоборот ухудшается качество прогноза с увеличением количества входных признаков. Это может быть связано с тем, что данный алгоритм считал важным разные признаки при их различном количестве (рис. 9б). Также отметим, что максимальный коэффициент достоверности прогноза R^2=0,83 был достигнут при использовании алгоритма RF, и это было получено при 5 признаках.

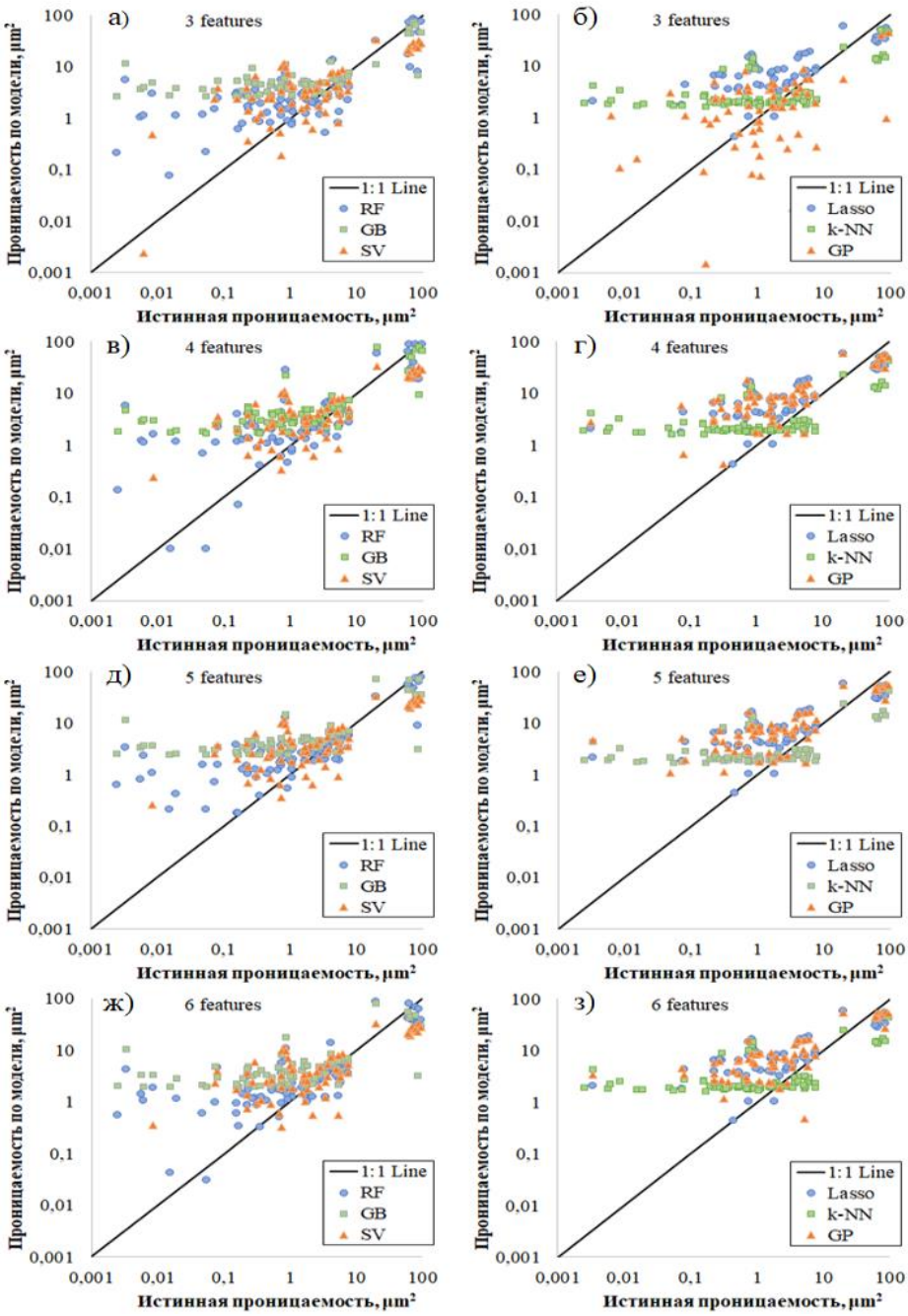

Рисунок 10. Прогнозированная с помощью разных алгоритмов проницаемость в сравнении с истинной проницаемости при разном количестве признаков

В табл. 5 собрана статистика ключевых параметров прогнозного и истинного значения проницаемости. Как видно из этой таблицы, алгоритмы SV, Lasso и GP прогнозировали отрицательные (минимальные) значения проницаемости во всех случаях с количеством признаков, тогда как она должна быть строго положительной. Причем, отрицательные проницаемости имеют большие значения по модули. Хотя, минимальная проницаемость по GB и k-NN положительная, но она все еще имеет большую погрешность от истинной минимальной проницаемости.

Таблица 5. Статистика предсказанных проницаемостей при разных количествах признаков

Параметр	*Признаки*	*Истинная*	*RF*	*GB*	*SV*	*Lasso*	*k-NN*	*GP*
Минимум	*3*	*0,00249*	*0,07540*	*2,66*	*-10,64*	*-16,37*	*1,6557*	*-4,33*
	4		*0,01001*	*1,75*	*-10,94*	*-16,37*	*1,6634*	*-21,50*
	5		*0,17730*	*2,40*	*-10,93*	*-16,37*	*1,6633*	*-17,05*
	6		*0,03128*	*1,94*	*-10,95*	*-16,37*	*1,6563*	*-17,10*
Максимум	*3*	*95,91*	*84,05*	*71,47*	*33,25*	*58,44*	*50,893*	*45,63*
	4		*88,70*	*78,88*	*33,26*	*58,44*	*50,409*	*60,02*
	5		*78,69*	*72,95*	*33,23*	*58,44*	*50,408*	*56,54*
	6		*87,81*	*80,94*	*33,44*	*58,44*	*51,634*	*56,39*
Среднее	*3*	*9,31*	*7,51*	*8,67*	*4,627*	*7,64*	*5,4628*	*2,40*
	4		*9,40*	*9,10*	*4,626*	*7,64*	*5,3918*	*7,63*
	5		*8,04*	*8,63*	*4,640*	*7,64*	*5,3917*	*8,29*
	6		*8,22*	*7,96*	*4,589*	*7,64*	*5,5671*	*8,18*

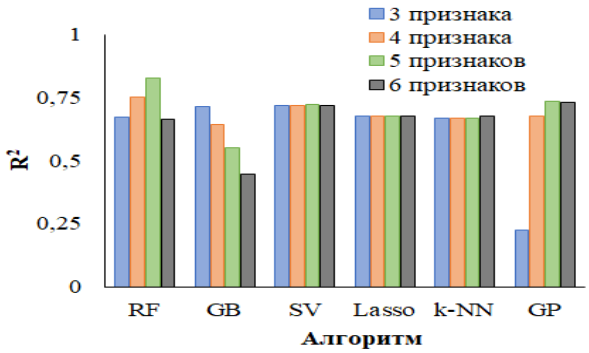

Рисунок 11. Коэффициент достоверности прогноза проницаемости при разных количествах признаков

Наиболее близкое к истинному значению проницаемости (превышающая всего 4 раза) было получено с помощью алгоритма RF при 4 признаках. Наиболее отдаленные от истинного максимального значения проницаемости были получены при помощи алгоритмов SV, Lasso, k-NN и GP, которые имеют погрешность 37-65% от истинной проницаемости. Самое близкое к максимальной истинной проницаемости была полученная с помощью алгоритма RF погрешность которой составляет 7,5%. Проницаемость, предсказанная по GB, также близко находится к истинной. Средние проницаемости, полученные с помощью алгоритмов RF и GB, являются наиболее близкими к истинной проницаемости при 2 признаках, и их погрешности от истинной проницаемости составили 2-3%. Другие средние значения проницаемости при остальных количествах признаков находятся относительно близко по сравнению с результатами других алгоритмов. Наиболее отдаленные средние от истинного значения проницаемости были получены с помощью SV и k-NN, тогда как проницаемости по Lasso и GP находятся в диапазоне погрешности 18-29% от истинной проницаемости.

Заключение

В настоящей статье рассмотрены 6 различных методов машинного обучения для анализа данных и прогноза абсолютной проницаемости 266 образцов различных горных пород размерами от 125 мкм до 3 мм. Анализ полного набора данных показал высокую корреляцию проницаемости с радиусом горловины пор, пористостью, координационным числом и радиусом пор. Разбиение полного набора данных на обучающий и тестовый существенно повлиял на результаты методов GB, Lasso и k-NN: с увеличением доли обучающего набора данных количество важных признаков по выбору этих методов выросло. Все алгоритмы предсказали завышенные минимальные проницаемости при любых разбиениях набора данных. Кроме того, алгоритмы SV, Lasso и GP прогнозировали отрицательные проницаемости. Все алгоритмы постарались близко предсказать максимальное и среднее значения (песчаные породы и песчаная упаковка) к истинному значению проницаемости. Сравнение качеств прогноза всех алгоритмов показал, что алгоритм RF является наиболее подходящим для анализа неоднородных данных, т.е. данных образцов различных породы, свойства которых сильно отличаются от их среднестатистического значения. Наибольший коэффициент достоверности прогноза R^2=0,83 был достигнут при использовании алгоритма RF, и это было получено при 5 признаках. Алгоритмы в основном выбрали в качестве важных признаков (параметров) радиуса пор, радиуса горловины пор и пористости.

Благодарность. *Данное исследование было профинансировано Комитетом Науки Министерства науки и высшего образования Республики Казахстан в рамках программы BR18574136 «Развитие методов глубокого обучения и интеллектуального анализа для решения сложных задач механики и робототехники».*

Список использованной литературы:

1 Rajalingam, B., Priya, R. (2018). Multimodal Medical Image Fusion based on Deep Learning Neural Network for Clinical Treatment Analysis. International Journal of ChemTech Research. https://doi.org/10.20902/IJCTR.2018.110621.

2 Cicceri, G., Inserra, G., & Limosani, M. (2020). A Machine Learning Approach to Forecast Economic Recessions—An Italian Case Study. Mathematics, 8(2), 241. https://doi.org/10.3390/math8020241.

3 Yoon, J. (2021). Forecasting of Real GDP Growth Using Machine Learning Models: Gradient Boosting and Random Forest Approach. Computational Economics, 57(1), 247–265. https://doi.org/10.1007/s10614-020-10054-w.

4 Gholami, R., Shahraki, A. R., & Jamali Paghaleh, M. (2012). Prediction of Hydrocarbon Reservoirs Permeability Using Support Vector Machine. Mathematical Problems in Engineering, 2012, 1–18. https://doi.org/10.1155/2012/670723.

5 Tembely, M., AlSumaiti, A. M., & Alameri, W. (2020). A deep learning perspective on predicting permeability in porous media from network modeling to direct simulation. Computational Geosciences, 24(4), 1541–1556. https://doi.org/10.1007/s10596-020-09963-4.

6 Waszkiewicz, S., Krakowska-Madejska, P., & Puskarczyk, E. (2019). Estimation of absolute permeability using artificial neural networks (multilayer perceptrons) based on well logs and laboratory data from Silurian and Ordovician deposits in SE Poland. Acta Geophysica, 67(6), 1885–1894. https://doi.org/10.1007/s11600-019-00347-6.

7 Болысбек, Д., Асилбеков, Б. и Кульджабеков, А. (2023). Численное изучение влияния растворения породы на поровую структуру карбонатных образцов на основе экспериментальных данных. Вестник «Физико-математические науки». 2(82), 54–63. DOI:https://doi.org/10.51889/2959-5894.2023.82.2.006.

8 Wu, J.-L., Yin, X.-L., & Xiao, H. (2018). Seeing Permeability From Images: Fast Prediction with Convolutional Neural Networks. https://doi.org/10.1016/j.scib.2018.08.006.

9 Al Khalifah, H., Glover, P. W. J., & Lorinczi, P. (2020). Permeability prediction and diagenesis in tight carbonates using machine learning techniques. Marine and Petroleum Geology, 112, 104096. https://doi.org/10.1016/j.marpetgeo.2019.104096.

10 Rabbani, A., & Babaei, M. (2019). Hybrid pore-network and lattice-Boltzmann permeability modelling accelerated by machine learning. Advances in Water Resources, 126, 116–128. https://doi.org/10.1016/j.advwatres.2019.02.012.

11 Rezaee, R., & Ekundayo, J. (2022). Permeability Prediction Using Machine Learning Methods for the CO2 Injectivity of the Precipice Sandstone in Surat Basin, Australia. Energies, 15(6), 2053. https://doi.org/10.3390/en15062053.

12 Erofeev, A., Orlov, D., Ryzhov, A., & Koroteev, D. (2019). Prediction of Porosity and Permeability Alteration Based on Machine Learning Algorithms. Transport in Porous Media, 128(2), 677–700. https://doi.org/10.1007/s11242-019-01265-3.

13 Rodríguez-Rodríguez, I., Rodríguez, J.-V., Woo, W. L., Wei, B., & Pardo-Quiles, D.-J. (2021). A Comparison of Feature Selection and Forecasting Machine Learning Algorithms for Predicting Glycaemia in Type 1 Diabetes Mellitus. Applied Sciences, 11(4), 1742. https://doi.org/10.3390/app11041742.

14 Otchere, D. A., Ganat, T. O. A., Gholami, R., & Lawal, M. (2021). A novel custom ensemble learning model for an improved reservoir permeability and water saturation prediction. Journal of Natural Gas Science and Engineering, 91, 103962. https://doi.org/10.1016/j.jngse.2021.103962.

15 Bolysbek, D. A., Assilbekov, B. K., Akasheva, Z. K., & Soltanbekova, K. A. (2021). ANALYSIS OF THE HETEROGENEITY INFLUENCE ON MAIN PARAMETERS OF POROUS MEDIA AT THE PORE SCALE. Journal of Mathematics, Mechanics and Computer Science, 112(4). https://doi.org/10.26577/JMMCS.2021.v112.i4.06.

References:

1 Rajalingam, B., Priya, R. (2018). Multimodal Medical Image Fusion based on Deep Learning Neural Network for Clinical Treatment Analysis. International Journal of ChemTech Research. https://doi.org/10.20902/IJCTR.2018.110621.

2 Cicceri, G., Inserra, G., & Limosani, M. (2020). A Machine Learning Approach to Forecast Economic Recessions—An Italian Case Study. Mathematics, 8(2), 241. https://doi.org/10.3390/math8020241.

3 Yoon, J. (2021). Forecasting of Real GDP Growth Using Machine Learning Models: Gradient Boosting and Random Forest Approach. Computational Economics, 57(1), 247–265. https://doi.org/10.1007/s10614-020-10054-w.

4 Gholami, R., Shahraki, A. R., & Jamali Paghaleh, M. (2012). Prediction of Hydrocarbon Reservoirs Permeability Using Support Vector Machine. Mathematical Problems in Engineering, 2012, 1–18. https://doi.org/10.1155/2012/670723.

5 Tembely, M., AlSumaiti, A. M., & Alameri, W. (2020). A deep learning perspective on predicting permeability in porous media from network modeling to direct simulation. Computational Geosciences, 24(4), 1541–1556. https://doi.org/10.1007/s10596-020-09963-4.

6 Waszkiewicz, S., Krakowska-Madejska, P., & Puskarczyk, E. (2019). Estimation of absolute permeability using artificial neural networks (multilayer perceptrons) based on well logs and laboratory data from Silurian and Ordovician deposits in SE Poland. Acta Geophysica, 67(6), 1885–1894. https://doi.org/10.1007/s11600-019-00347-6.

7 Bolysbek, D., Asilbekov, B. i Kul'dzhabekov, A. (2023). Chislennoe izuchenie vlijanija rastvorenija porody na porovuju strukturu karbonatnyh obrazcov na osnove jeksperimental''nyh dannyh [Numerical study of the influence of roc dissolution on the pore structure of carbonate samples based on experimental data]. Vestnik «Fiziko-matematicheskie nauki». 2 (82). 54–63. DOI:https://doi.org/10.51889/2959-5894.2023.82.2.006. (In Russian)

8 Wu, J.-L., Yin, X.-L., & Xiao, H. (2018). Seeing Permeability From Images: Fast Prediction with Convolutional Neural Networks. https://doi.org/10.1016/j.scib.2018.08.006.

9 Al Khalifah, H., Glover, P. W. J., & Lorinczi, P. (2020). Permeability prediction and diagenesis in tight carbonates using machine learning techniques. Marine and Petroleum Geology, 112, 104096. https://doi.org/10.1016/j.marpetgeo.2019.104096.

10 Rabbani, A., & Babaei, M. (2019). Hybrid pore-network and lattice-Boltzmann permeability modelling accelerated by machine learning. Advances in Water Resources, 126, 116–128. https://doi.org/10.1016/j.advwatres.2019.02.012.

11 Rezaee, R., & Ekundayo, J. (2022). Permeability Prediction Using Machine Learning Methods for the CO2 Injectivity of the Precipice Sandstone in Surat Basin, Australia. Energies, 15(6), 2053. https://doi.org/10.3390/en15062053.

12 Erofeev, A., Orlov, D., Ryzhov, A., & Koroteev, D. (2019). Prediction of Porosity and Permeability Alteration Based on Machine Learning Algorithms. Transport in Porous Media, 128(2), 677–700. https://doi.org/10.1007/s11242-019-01265-3.

13 Rodríguez-Rodríguez, I., Rodríguez, J.-V., Woo, W. L., Wei, B., & Pardo-Quiles, D.-J. (2021). A Comparison of Feature Selection and Forecasting Machine Learning Algorithms for Predicting Glycaemia in Type 1 Diabetes Mellitus. Applied Sciences, 11(4), 1742. https://doi.org/10.3390/app11041742.

14 Otchere, D. A., Ganat, T. O. A., Gholami, R., & Lawal, M. (2021). A novel custom ensemble learning model for an improved reservoir permeability and water saturation prediction. Journal of Natural Gas Science and Engineering, 91, 103962. https://doi.org/10.1016/j.jngse.2021.103962.

15 Bolysbek, D. A., Assilbekov, B. K., Akasheva, Z. K., & Soltanbekova, K. A. (2021). ANALYSIS OF THE HETEROGENEITY INFLUENCE ON MAIN PARAMETERS OF POROUS MEDIA AT THE PORE SCALE. Journal of Mathematics, Mechanics and Computer Science, 112(4). https://doi.org/10.26577/JMMCS.2021.v112.i4.06.