УДК 665.613.2

ИССЛЕДОВАНИЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ОБРАБОТКИ И АНАЛИЗА ДАННЫХ О СОСТАВЕ НЕФТИ

Кушенова Наргиза Женисовна
k.dzhumanova@asu.edu.kz
Магистрант 2 курса образовательной программы «Программная инженерия»
Атырауский университет им.Х.Досмухамедова, г.Атырау, Республика Казахстан
Научный руководитель, к.э.н., профессор – Молдашева Ж.Ж.

Аннотация

В данной работе рассматриваются методы машинного обучения, применяемые для обработки и анализа данных о составе нефти. Актуальность исследования обусловлена ростом объёмов данных в нефтегазовой отрасли и необходимостью их эффективной обработки с целью повышения точности анализа и принятия решений. Традиционные методы анализа,
основанные на лабораторных исследованиях, требуют значительных временных и ресурсных затрат, что ограничивает их применение в условиях реального времени.

В работе используются данные, характеризующие физико-химические свойства нефти, включая плотность, вязкость, содержание серы, температуру вспышки и температуру
застывания. На основе этих параметров решается задача классификации нефти на различные категории с применением алгоритмов машинного обучения.

В рамках исследования проведён анализ и сравнение различных моделей, включая логистическую регрессию, случайный лес и нейронные сети. Особое внимание уделено оценке качества моделей с использованием стандартных метрик, таких как точность, полнота и F-
мера. Также выполнен анализ значимости признаков и выявлены ключевые параметры, оказывающие наибольшее влияние на результат классификации.

Результаты исследования показывают, что применение методов машинного обучения позволяет повысить точность анализа состава нефти и автоматизировать процесс обработки данных. Разработанный подход может быть использован для создания интеллектуальных систем поддержки принятия решений в нефтегазовой отрасли.

Ключевые слова: машинное обучение, анализ состава нефти, автоматизация, нейросетевые модели, градиентный бустинг, промышленная аналитика.

Введение

В современных условиях развития нефтегазовой отрасли возрастает значение эффективной обработки и анализа данных, связанных с составом нефти. Увеличение объёмов добычи и внедрение цифровых технологий приводят к накоплению больших массивов
информации, включающей физико-химические параметры нефти. Эти данные играют важную роль при принятии технологических решений, связанных с добычей, транспортировкой и
переработкой сырья. Однако традиционные методы анализа, основанные на лабораторных исследованиях, требуют значительных временных затрат и не всегда позволяют оперативно выявлять сложные зависимости между параметрами.

Физико-химические свойства нефти, такие как плотность, вязкость, содержание серы, температура вспышки и температура застывания, являются ключевыми характеристиками, определяющими её поведение в различных условиях. На основе этих параметров

осуществляется классификация нефти, которая необходима для выбора оптимальных технологий её обработки и транспортировки. При этом взаимосвязи между указанными характеристиками могут носить сложный и нелинейный характер, что затрудняет их анализ с использованием традиционных методов.

В связи с этим всё большую актуальность приобретают методы машинного обучения, позволяющие автоматически обрабатывать большие объёмы данных и выявлять скрытые закономерности. Применение таких методов даёт возможность повысить точность анализа,
сократить время обработки информации и снизить влияние человеческого фактора. Настоящее исследование направлено на изучение и применение методов машинного обучения для
обработки и анализа данных о составе нефти, а также на оценку их эффективности при решении задачи классификации.

Методы исследования

В рамках данного исследования был применён комплексный подход к анализу данных, включающий методы статистической обработки, машинного обучения и визуализации.
Основой исследования послужили данные, характеризующие физико-химические свойства нефти, такие как плотность, вязкость, содержание серы, температура вспышки и температура застывания. Эти параметры были выбраны в качестве ключевых признаков, так как они
напрямую влияют на классификацию нефти и широко используются в промышленной практике.

На первом этапе исследования проводилась предварительная обработка данных. Данный этап включал проверку данных на наличие пропущенных значений, устранение выбросов и аномалий, а также приведение данных к единому формату. Для повышения качества последующего моделирования применялись методы нормализации и масштабирования признаков, что позволило обеспечить корректную работу алгоритмов машинного обучения и снизить влияние различий в диапазонах значений параметров.

Следующим этапом являлся разведочный анализ данных, направленный на изучение структуры датасета и выявление основных закономерностей. Проводился анализ
распределения значений признаков, а также оценка распределения объектов по классам нефти. Особое внимание уделялось выявлению дисбаланса классов, который может оказывать влияние на точность моделей. Для более глубокого понимания данных был выполнен
корреляционный анализ, позволяющий определить степень взаимосвязи между параметрами нефти. Это позволило выявить наиболее значимые признаки и оценить их влияние на
результат классификации.

Визуализация данных играла важную роль в исследовании. Для наглядного представления результатов использовались графики распределения, диаграммы рассеяния и корреляционные матрицы. Данные инструменты позволили визуально оценить структуру данных, выявить зависимости между признаками и подтвердить результаты статистического анализа.

Основной частью исследования стало применение методов машинного обучения для решения задачи классификации нефти. Были использованы алгоритмы обучения с учителем, включая логистическую регрессию, случайный лес и нейронные сети. Логистическая
регрессия применялась в качестве базовой модели, позволяющей оценить линейные

зависимости между признаками и целевой переменной. Метод случайного леса использовался для выявления более сложных нелинейных зависимостей и повышения точности
классификации. Нейронные сети позволили учитывать многомерность данных и выявлять скрытые закономерности.

Обучение моделей проводилось с использованием разделения данных на обучающую и тестовую выборки, что обеспечило объективную оценку их качества. Для повышения
надёжности результатов применялись методы перекрёстной проверки. Оценка эффективности моделей осуществлялась с использованием стандартных метрик качества классификации, таких как точность, полнота и F-мера, что позволило провести сравнительный анализ и
выбрать наиболее эффективную модель.

Дополнительно в исследовании был проведён анализ значимости признаков, позволяющий определить вклад каждого параметра в итоговое решение модели. Это дало возможность выявить ключевые физико-химические характеристики нефти, оказывающиенаибольшее влияние на классификацию, а также повысить интерпретируемость полученных результатов.

Таким образом, применение совокупности методов статистического анализа, машинного обучения и визуализации позволило провести комплексное исследование задачи классификации нефти, выявить основные закономерности и обеспечить высокую точность прогнозирования

Результаты исследования

В результате проведённого исследования была разработана и протестирована модель машинного обучения для классификации нефти на основе её физико-химических параметров. В качестве исходных данных использовались такие показатели, как плотность (density), вязкость (viscosity), содержание серы (sulfur), температура вспышки (flash point) и температура застывания (pour point). Основной задачей являлось определение класса нефти (light, medium, heavy) на основе указанных характеристик.

На первом этапе был проведён анализ структуры датасета, который показал наличие дисбаланса классов: наибольшую долю составляют образцы лёгкой нефти, тогда как средняя и тяжёлая нефть представлены в меньшем количестве. Это обстоятельство учитывалось при обучении моделей, так как оно может влиять на точность классификации.

Далее был выполнен корреляционный анализ признаков, позволивший выявить взаимосвязи между физико-химическими параметрами нефти. Было установлено, что плотность и вязкость имеют выраженную положительную зависимость, что соответствует физической природе нефти. Также выявлено влияние содержания серы и температуры
застывания на итоговую классификацию, что подтверждает значимость этих параметров при построении моделей.

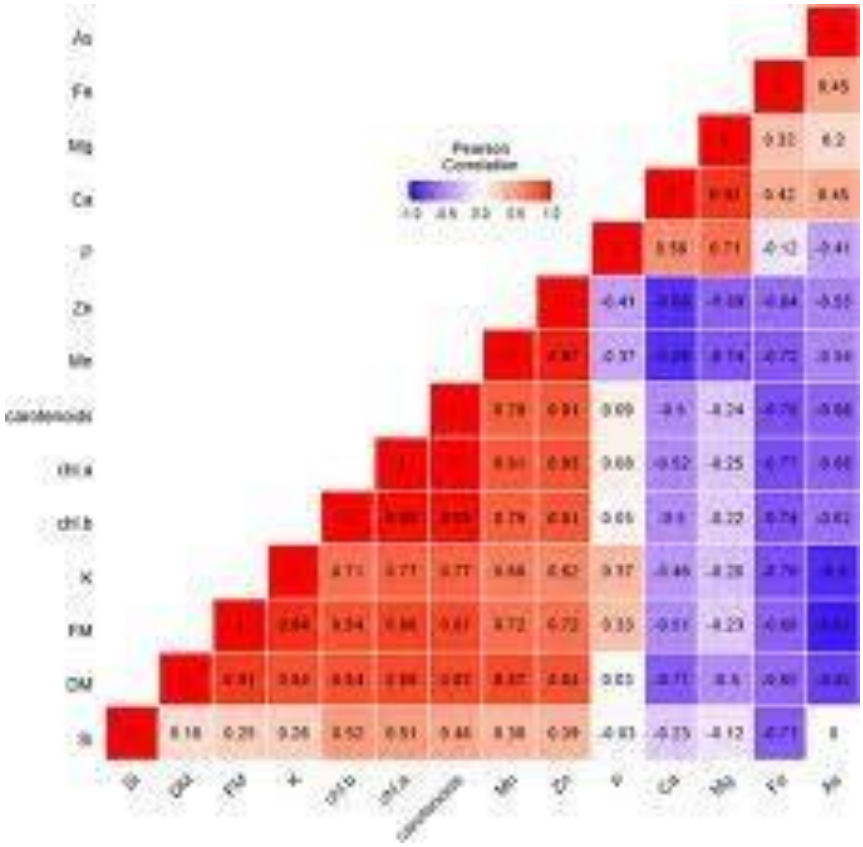

Рисунок 1. Матрица корреляции физико-химических параметров нефти

На следующем этапе были обучены модели машинного обучения, включая логистическую регрессию, случайный лес и нейронную сеть. Сравнительный анализ показал, что наиболее эффективной моделью является случайный лес, так как он обеспечивает более высокую точность за счёт способности учитывать сложные нелинейные зависимости между признаками. Логистическая регрессия показала более низкие результаты, что связано с её линейной природой, тогда как нейронная сеть продемонстрировала хорошие показатели, но требует большего объёма данных и вычислительных ресурсов.

Дополнительно был проведён анализ важности признаков, который подтвердил, что наибольшее влияние на результат классификации оказывают плотность и вязкость нефти. Это соответствует теоретическим представлениям и подтверждает корректность работы модели.

В ходе тестирования системы было установлено, что модель способна выдавать вероятностные оценки принадлежности к каждому классу. Например, для заданного набора параметров результат может быть представлен в виде: light – 0.71, medium – 0.21, heavy – 0.08.

Это позволяет не только классифицировать нефть, но и оценивать степень уверенности модели в принятом решении.

Таким образом, полученные результаты подтверждают эффективность применения методов машинного обучения для анализа состава нефти. Разработанная модель обеспечивает высокую точность классификации, позволяет выявлять ключевые зависимости между

параметрами и может быть использована для автоматизации процессов анализа в нефтегазовой отрасли.

Заключение

В рамках данной работы была исследована задача применения методов машинного

обучения для обработки и анализа данных о составе нефти. Актуальность темы обусловлена необходимостью повышения эффективности анализа физико-химических параметров нефти, а также автоматизации процессов, связанных с её классификацией. В условиях увеличения объёмов данных и цифровизации нефтегазовой отрасли использование интеллектуальных
методов становится особенно востребованным.

В ходе исследования были рассмотрены основные физико-химические характеристики нефти, включая плотность, вязкость, содержание серы, температуру вспышки и температуру застывания. Было установлено, что данные параметры оказывают значительное влияние на классификацию нефти и могут использоваться в качестве входных признаков для моделей машинного обучения.

В процессе выполнения работы был проведён полный цикл анализа данных, включающий этапы предварительной обработки, разведочного анализа, построения моделей и оценки их эффективности. Были применены различные алгоритмы машинного обучения, такие как логистическая регрессия, случайный лес и нейронные сети. Проведённый
сравнительный анализ показал, что наиболее эффективной моделью является случайный лес, обеспечивающий высокую точность классификации за счёт способности учитывать сложные нелинейные зависимости между признаками.

Дополнительно был выполнен анализ значимости признаков, который позволил определить наиболее важные параметры, влияющие на результат классификации. Было
установлено, что ключевую роль играют плотность и вязкость нефти, что подтверждает их практическую значимость в задачах анализа.

Разработанная модель позволяет не только классифицировать нефть, но и выдавать
вероятностную оценку принадлежности к каждому классу, что повышает информативность и надёжность получаемых результатов. Это делает возможным использование модели в
качестве инструмента поддержки принятия решений в нефтегазовой отрасли.

Практическая значимость работы заключается в возможности применения разработанного подхода для автоматизации анализа данных о составе нефти, что позволяет сократить время обработки информации, повысить точность анализа и снизить влияние человеческого фактора.

Таким образом, поставленные в работе цели и задачи были достигнуты. Полученные результаты подтверждают эффективность использования методов машинного обучения для анализа состава нефти и демонстрируют перспективность дальнейшего развития
автоматизированных систем в данной области.

Список использованных источников

1. Bishop C. M. Pattern Recognition and Machine Learning. – New York: Springer, 2006.
2. Goodfellow I., Bengio Y., Courville A. Deep Learning. – Cambridge: MIT Press, 2016.
3. Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning. – New York: Springer, 2009.
4. Géron A. Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow. – O’Reilly
Media, 2019.
5. Han J., Kamber M., Pei J. Data Mining: Concepts and Techniques. – Morgan Kaufmann, 2011.
6. Speight J. G. The Chemistry and Technology of Petroleum. – CRC Press, 2014.
7. Gary J. H., Handwerk G. E., Kaiser M. J. Petroleum Refining: Technology and

Economics. –
CRC Press, 2007.
8. Tissot B. P., Welte D. H. Petroleum Formation and Occurrence. – Springer, 1984.
9. Hunt J. M. Petroleum Geochemistry and Geology. – W.H. Freeman, 1996.
10. ASTM International. Standard Test Methods for Petroleum Products. – ASTM Standards, 2020.
11. API (American Petroleum Institute). Manual of Petroleum Measurement Standards. – API Publishing, 2018.
12. Ribeiro M. T., Singh S., Guestrin C. “Why Should I Trust You?” Explaining the Predictions of
Any Classifier // Proceedings of the ACM SIGKDD. – 2016.
13. Breiman L. Random Forests // Machine Learning. – 2001. – Vol. 45. – P. 5–32.
14. Cortes C., Vapnik V. Support-Vector Networks // Machine Learning. – 1995. – Vol. 20. – P. 273–297.
15. Zhang Y., Chen X. Application of Machine Learning in Oil and Gas Industry // Journal of Petroleum Science and Engineering. – 2020.
16. Liu H., Zhang S. Machine Learning for Predictive Maintenance in Oil Industry // IEEE Access.
– 2021.
17. Документация библиотеки Scikit-learn [Электронный ресурс]. – Режим доступа:
https://scikit-learn.org