УДК 547 DOI 10.55421/3034-4689_2025_28_10_102

А. Д. Лифанов, Е. Г. Лифанова

ИСПОЛЬЗОВАНИЕ ГРАДИЕНТНОГО БУСТИНГА ДЛЯ ПРОГНОЗИРОВАНИЯ ТЕМПЕРАТУРЫ ВСПЫШКИ ОРГАНИЧЕСКИХ СОЕДИНЕНИЙ

Ключевые слова: большие данные, индустрия 4.0, температура вспышки, градиентный бустинг, искусственный интеллект.

В настоящее время в химии накоплен большой массив экспериментальных данных. В связи с этим, возникает необходимость совершенствования вычислительных методов хранения, обработки экспериментальных данных. Температура вспышки органических соединений является важным факторов, обеспечивающим безопасность химических производств. Современная химическая промышленность в условиях перехода к Индустрии 4.0 претерпевает глубокие цифровые трансформации из-за повышенных требований к безопасности химических производств. Использование цифровых двойников процессов вызвали значительные изменения в организации химического производства. Так, в настоящее время активно развиваются такие направления Индустрии 4.0 как аддитивные технологии, Интернет вещей и т.д. В таких условиях применение алгоритмов машинного обучения является ключевым инструментом для выявления факторов, влияющих на температуру вспышки органических соединений и повышения эффективности прогнозирования данного параметра. В базу данных для данной работы была включена информация о температурах вспышки для 1741 органических веществ. Данные о температурах вспышки органических соединений были взяты из базы данных PubChem. Для упрощения анализа представления органических соединений, мы использовали 208 дескрипторов RDKit, поскольку они являются одними из лучших дескрипторов для прогнозирования свойств химических соединений. Данные дескрипторы создаются на основе общих ключей подструктуры. Кроме того, модели были рассчитаны с использованием молекулярных отпечатков Моргана, также известных как циркулярные отпечатки с радиусом 2. В рамках данной работы был реализован градиентный бустинг. XGBoost построен на принципах усиления градиента с использованием древовидных алгоритмов обучения для повышения возможностей прогнозного моделирования. Для обучающей выборки полученная классификационная модель градиентного бустинга показала безошибочную классификацию, ошибка прогноза для нее равна 0. Статистические характеристики построенной модели гребневой регрессии для выборки имеют следующие значения: R^2 =0.74 и ошибкой предсказания RMSE=36.36 К.

A. D. Lifanov, E. G. Lifanova

USING GRADIENT BOOSTING FOR PREDICTING THE FLASH POINT OF ORGANIC COMPOUNDS

Keywords: big data, industry 4.0, flash point, gradient boosting, artificial intelligence.

Currently, a large amount of experimental data has been accumulated in chemistry. In this regard, there is a need to improve computational methods for storing and processing experimental data. The flash point of organic compounds is an important factor ensuring the safety of chemical industries. The modern chemical industry, in the context of the transition to Industry 4.0, is undergoing deep digital transformations due to increased safety requirements for chemical industries. The use of digital process twins has caused significant changes in the organization of chemical production. Thus, such areas of Industry 4.0 as additive technologies, the Internet of Things, etc. are currently actively developing. In such conditions, the use of machine learning algorithms is a key tool for identifying factors affecting the flash point of organic compounds and improving the efficiency of predicting this parameter. Information on the flaskpoint temperature for 1741 organic substances was included in the database for this work. The data on flash points of organic compounds were taken from the PubChem database. To simplify the analysis of the representation of organic compounds, we used 208 RDKit descriptors, as they are among the best descriptors for predicting the properties of chemical compounds. These descriptors are created based on the shared keys of the substructure. In addition, the models were calculated using Morgan's molecular fingerprints, also known as circular prints with a radius of 2. As part of this work, gradient boosting was implemented. XGBoost is built on the principles of gradient enhancement using tree-based learning algorithms to enhance predictive modeling capabilities. For the training sample, the obtained gradient boosting model showed an error-free classification, the prediction error for it is 0. The statistical characteristics of the constructed ridge regression model for the sample have the following values: R^2 =0.74 and the prediction error RMSE=36.36 K.

Введение

Температура вспышки органических соединений является важным фактором, обеспечивающим безопасность химических производств. Современная химическая промышленность в условиях перехода к Индустрии 4.0 претерпевает глубокие цифровые трансформации из-за повышенных требований к безопасности химических производств. Индустрия 4.0 предполагает использование цифровых двойников процессов. Это вызвало значительные изменения в организации химического производства. Так, в настоящее время активно развиваются такие направления Индустрии 4.0 как аддитивные технологии, Интернет вещей и т.д. В таких условиях применение алгоритмов машинного обучения является ключевым инструментом для выявления факторов, влияющих органических соединений, в том числе на температуру вспышки и повышения эффективности прогнозирования данного параметра [1].

В работе В.С. Конышева [2] показана целесообразность применения методов машинного обучения для прогнозирования температуры вспышки органических соединений. В данной работе показано, что использование градиентного бустинга позволяет получить высокую точность модели. Однако авторы

рассмотрели только два вида дескрипторов – структурные ключи и молекулярные отпечатки (фингерпринты) Моргана.

Одна из самых эффективных моделей для прогнозирования температуры вспышки органических соединений среди регрессионных методов машинного обучения является модель градиентного бустинга XGBoost. Данный алгоритм приобрел популярность благодаря своей высокой производительности, масштабируемости и эффективности, особенно при работе с большими наборами данных. XGBoost построен на принципах усиления градиента с использованием древовидных алгоритмов обучения для повышения возможностей прогнозного моделирования.
В основе метода лежит ступенчатый поиск оптимальной модели. Пусть произвольно задано *M* шагов для спуска, а $L\left(y,F(x)\right)$ - дифференциальная функция потерь (метрика, которую алгоритм стремится минимизировать) [3]. В качестве дифференциальной функции потерь для регрессии выбирают функцию вида:

$$L\left(y_i, F\left(x\right)\right) = \frac{1}{2}\left(y_i - F\left(x\right)\right)$$

где $F\left(x\right)$ - любая дифференцируемая функция.

Пусть дан набор данных пар дескрипторов x_i и целевых переменных y_i, для которых имеется зависимость вида $y = f\left(x\right)$.

Тогда следующая последовательность шагов позволяет достичь оптимальной модели:

1. инициализация первичной модели:

$$F_o = \arg\min_{\gamma} \sum_{i=1}^{n} L\left(y_i, F\left(x\right)\right) \qquad (1)$$

На данном шаге осуществляется последовательная минимизация функции потерь *L(yi, F(x))*.

2. для следующих *M* шагов рассчитываются псевдо-остатки:

$$r_{im} = -\left[\frac{L\left(y_i, F\left(x\right)\right)}{\partial F\left(x\right)}\right]_{F(x)=F_{m-1}(x)}, i = \overline{1,n} \qquad (2)$$

Данные величины показывают, как именно необходимо подкорректировать прогнозы. Здесь производную следует брать по второму аргументу функции потерь *F(x)*.

3. функциональную зависимость можно описать как множество моделей H, каждый элемент которого $h_m\left(x\right):\left\{\left(x_i, r_{im}\right)\right\}_{i=1}^{n}$ определяется некоторым вектором параметров r_{im}. Решая одномерную оптимизационную задачу происходит оценка мультипликатора:

$$\gamma_m = \arg\min_{\gamma} \sum_{i=1}^{n} L\left(y_i, F_{m-1}\left(x\right) + \gamma h_m\left(x_i\right)\right) \qquad (3)$$

Т.е. предсказание *m*-ой модели есть предсказание *m*-1 модели и ошибки предсказания *m*-ой модели. Мы хотим, чтобы каждая следующая модель компенсировала ошибки предыдущего дерева. Таким образом необходимо найти такие модели $\gamma_m h_m(x_i)$, которые обеспечат минимум функции потерь.

4. проводится градиентный спуск:

$$F_m\left(x\right) = F_{m-1}\left(x\right) + \gamma_m h_m\left(x_i\right) \qquad (4)$$

В результате новые модели должны аппроксимировать отрицательный градиент.

5. после M шагов итерационный спуск останавливается, а сформирована финальная модель $F_m\left(x\right)$ обладает лучшими свойствами.

Одной из отличительных особенностей LightGBM (Light Gradient Boosting Machine) - высокопроизводительный фреймворк с открытым исходным кодом для градиентного бустинга является его способность эффективно обучаться на больших наборах данных и поддерживать высокую скорость выполнения благодаря двум новым методам: односторонней выборки на основе градиента (Gradient-Based one-Side Sampling, GOSS) и метод уменьшения эффективного набора факторов (Exclusive Feature Bundling, EFB) [2]. Способность метода эффективно обучаться на больших наборах данных и поддерживать высокую скорость является особенно важной при разработке метода, ведь предоставляет возможность выстроить подход к тренировке большого количества моделей для каждого объекта сети с учетом индивидуального характера воздействия внешних факторов.

В нашей работе, мы предположили, что учет аномалий (выбросов) в исходном наборе данных позволит существенно повысить прогностическую способность модели. Для достижения поставленной цели был использован ряд методов идентификации аномалий на основе кластеризации [4]:

– K-Nearest Neighbors Detector: для каждого наблюдения расстояние до k-го ближайшего соседа сигнализирует о возможной оценке аномалии;

– Isolation Forest: выполнение разделения данных в форме набора деревьев позволяет проверить насколько изолированным является наблюдение в структуре;

– Angle-Based Outlier Detection (ABOD): связь между каждым наблюдением и его соседом в форме дисперсии его взвешенных баллов косинуса помогает выявить выброс;

– Histogram-based Outlier Detection: получение оценки выбросов путем построения гистограмм;

– Local Correlation Integral (LOCI): идентификация аномалий на основе кластеров, их диаметров и расстояний между кластерами.

Цель работы заключалась в исследовании закономерностей связи химического строения органических соединений с температурами вспышки.

Материалы и методы

Традиционно, построение QSPR (Quantitative structure–property relationship) – модели (установление количественных соотношений структура-свойство) проходит несколько этапов: подготовка данных, выбор модели, отбор дескрипторов, обучение модели и анализ результатов [2].

Данные о температурах вспышки органических соединений были взяты из базы данных PubChem [5].

Собранный набор данных содержал информацию о свойствах 1741 органических соединений с их значениями температуры вспышки в диапазоне от 13.87 до 243.24 кДж/моль.

В нашей работе в качестве линейных представлений выбрана система SMILES [6, 7]. Далее на основе SMILES генерируется дескриптор, отражающий особенности структуры органического соединения [8]. Информация о структуре химических соединений кодируется дескрипторами.

С помощь дескрипторов кодируется информация о структуре химического соединения. Так в качестве дескрипторов можно использовать число атомов углерода, число ароматических циклов, количество двойных связей в органических соединениях. Описание структуры химических соединений осуществляется наборами дескрипторов [9-13]. Данные дескрипторы могут быть легко рассчитаны с использованием библиотеки RDKit 2 [9]. RDKit представляет собой инструментарий для хемоинформатики с открытым исходным кодом [9].

В рамках данной работы был реализован метод градиентного бустинга (XGBoost) [18]. Данный метод машинного обучения имеет преимущества перед дуги методами, например, гребневой регрессией [10], алгоритмом случайного леса [14], методом ближайших соседей kNN [15, 16], методом опорных векторов (SVM) [17]. Данный метод относится к ансамблевым методам и основан на том, что каждая последующая модель исправляет ошибки предыдущей модели, а, например, AdaBoostRegressor (AdaBoost) корректирует веса модели [19].

Для настройки гиперпараметров моделей машинного обучения была использована библиотека для оптимизации Optuna [2]. Данная библиотека эффективно сочетает алгоритм поиска и обрезки (Pruning) для оптимизации затрат при реализации модели. Библиотека Optuna предоставляет возможность распределить вычисления на несколько узлов для ускорения оптимизации на больших вычислительных мощностях.

Выборка была разбита на обучающую и тестовую в отношении 80% и 20% соответственно.

Оценка качества регрессионных моделей осуществлялась с использованием стандартных метрик качества: коэффициент детерминации R^2, средняя абсолютная ошибка (MAE, Mean Absolute Error), средняя абсолютная процентная погрешность (MAPE, Mean Absolute Percentage Error), среднеквадратичная логарифмическая ошибка (RMSLE, root Mean Squared Logarithmic Error), среднеквадратичная ошибка (RMSE).

Коэффициент детерминации R^2, обеспечивающий стандартизированную меру того, однако то, насколько хорошо модель фиксирует отклонения в данных, не всегда полезно при использовании со сложными моделями или при наличии мультиколлинеарности. Коэффициент детерминации (R^2) рассчитывается следующим образом:

$$R^2 = 1 - \frac{\sum_i \left(y_i^{\text{pred}} - y_i^{\text{exp}}\right)^2}{\sum_i \left(y_i^{\text{pred}} - \bar{y}_i^{\text{exp}}\right)^2} \qquad (5)$$

где y_i^{pred} и y_i^{exp} - предсказанные и истинные значения характеристик объектов, соответственно, N - объем выборки.

Для оценки и сравнения моделей также были выбраны другие метрики в качестве индикаторов [8]:

MAE легка в интерпретации и устойчивой к выбросам, однако придает одинаковый вес погрешностям:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i^{pred} - y_i^{\text{exp}}\right| \qquad (6)$$

MAPE удобна для понимания относительного уровня погрешности и сравнения моделей, решающих задачи регрессии для данных разного диапазона распределений:

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{\left|y_i^{pred} - y_i^{\text{exp}}\right|}{\left|y_i^{pred}\right|}, \qquad (7)$$

RMSLE оказывает наибольшее влияние недооценке и прекрасно подходит в комбинации с химической задачей, имеющей соответствующий эффект на температуру вспышки и процесс (например, температура кипения и дескрипторов, характеризующих это свойство):

$$RMSLE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\log\left(1+y_i^{pred}\right) - \log\left(1+y_i^{exp}\right)\right)^2} \qquad (8)$$

По сравнению с тремя предыдущими метриками, метрика RMSE более чувствительна к влиянию больших ошибок. Однако она позволяет выявить наиболее существенные ошибки прогнозирования, требующие особого внимания. RMSE рассчитывается следующим образом:

$$RMSE = \sqrt{\frac{\sum_i \left(y_i^{\text{pred}} - y_i^{\text{exp}}\right)^2}{N-1}} \qquad (9)$$

Нами также проведена процедура 5-кратного скользящего контроля (кросс-валидации, CV). Подробно о процедуре кросс-валидации мы описали в работах [2, 24, 25].

Результаты и их обсуждение

Для поиска оптимального решения для обучения были выбраны следующие методы: ансамбли деревьев решений с использованием алгоритмов случайного леса (Random Forrest), градиентный бустинг (XGBoost) и AdaBoostRegressor [14, 18, 19].

Среди моделей, которые имели наилучшие показатели точности прогнозирования пиковой нагрузки, можно выделить:

1. Random Forest, которые сочетают структуру дерева решений и обучение ансамбля для случайного моделирования деревьев решений и путем смешивания их результатов улучшают показатели точности модели.

Алгоритм Random Forrest начинается с создания нескольких начальных выборок из тренировочного набора данных, которые служат обучающими данными для отдельных деревьев. Для каждой выборки строится дерево решений. Каждое дерево строится путем выбора наилучшего деления на каждом узле (node) из случайного подмножества, что помогает обеспечить разнообразие отдельных деревьев. После построения всех деревьев решений они используются для прогнозирования путем взвешивания [14]. Random Forest широко используются благодаря своей надежности, высокой производительности и простоте использования. Данный алгоритм эффективен в обработке данных большой размерности, предоставляя оценки важности факторов и является менее подверженным переобучению по сравнению с отдельными деревьями решений.

2. Градиентный бустинг (Gradient Boosting), который основан на пошаговом поиске оптимальной модели [18]. Алгоритм был подробно описан во введении. Он позволяет обрабатывать категориальные переменные с помощью бинарного кодирования, не требуя предварительной обработки. Среди других преимуществ алгоритма возможность обрабатывать пропущенные данные и эффективность при работе с наборами данных, содержащими категориальные характеристики с высокой кардинальностью (мощностью и производительностью в работе с множеством атрибутов).

Наивысший результат на тестовой выборке был достигнут при использовании модели XGBoost по RMSE, коэффициенту детерминации и RMSLE (табл. 1). Следует отметить, что градиентный бустинг обладает более высокой прогностической способностью по сравнению с алгоритмом случайного леса. По показателям MAE и MAPE наивысший результат был также зафиксирован при использовании модели градиентного бустинга Extreme Gradient Boosting (XGBoost).

Таблица 1 – Оценка результативности моделей на тестовой выборке

Table 1 – Evaluation of model performance on the test sample

Метод	Обучающая выборка		Тестовая выборка	
	R^2	RMSE	R^2	RMSE
RF	0.52	49.30	0.53	48.82
XGBoost	0.74	36.36	0.66	41.36
AdaBoost	0.58	32.52	0.57	34.12

В результате исследования мы получили адекватные модели «структура-свойство» с высокими статистическими показателями. Для прогноза и интерпретации использовались модели как результат пятикратного скользящего контроля (Рис. 1) [2, 21].

В работах [1, 2] показано, что прогностическая способность модели определяется числом дескрипторов, адекватно описывающих структуру химического соединения.

Лучшими гиперпараметрами модели оказались learning_rate=0.01, n_estimators=500, loss='huber', criterion='squared_error'. По результатам обучения модели RMSE = 36.36 К и R^2=0.74 (на обучающей выборке) и RMSE = 41.36 К и R^2=0.66 (на тестовой выборке). В работе [2] качество модели XGBoost было хуже: RMSE = 32.27 К и R^2=0.60 (с процедурой 5-кратного скользящего контроля).

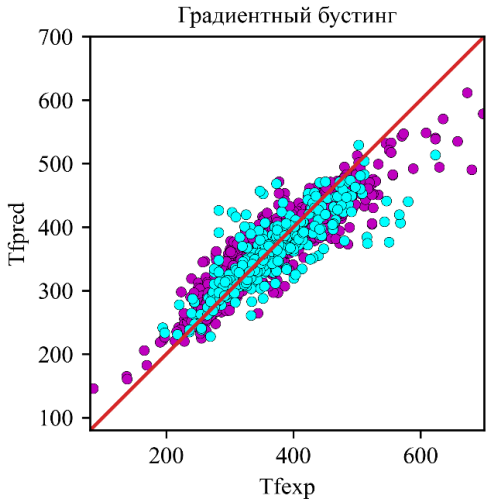

Рис. 1 – Регрессионная модель градиентного бустинга для прогнозирования температуры вспышки органических соединений

Fig. 1 – A gradient boosting model for predicting the flash point of organic compounds

Следующим этапом нашей работы являлось установление вклада отдельных фрагментов на температуру вспышки органических соединений.

Алгоритм градиентного бустинга позволяет установить вклад отдельных дескрипторов в предсказание целевой переменной y.

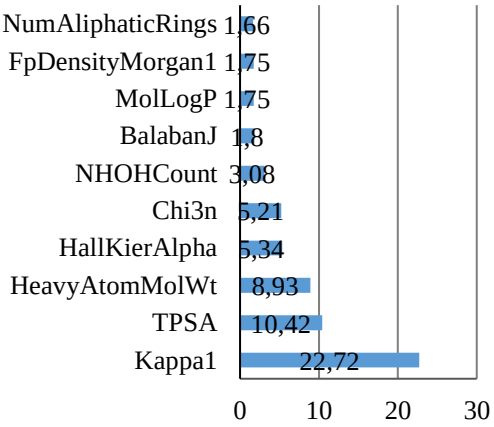

Рис. 2 – Уровень значимости дескрипторов в прогнозирования температуры вспышки органических соединений

Fig. 2 – The significance level of descriptors in predicting the flash point of organic compounds

В результате проведенного анализа нами не выявлено четкой тенденции между вкладами фрагментов на температуру вспышки органических соединений.

Таблица 2 – Обозначение и наименование наиболее значимых дескрипторов в прогнозировании температуры вспышки органических соединений

Table 2 – Designation and name of the most significant descriptors in predicting the flash point of organic compounds

Обозначение	Наименование (на англ.)	Наименование (на русск.)
Kappa1	Kappa alpha index for 1 bonded fragment	Каппа альфа-индекс для 1 связанного фрагмента
TPSA	Topological polarity surface area	Топологическая площадь полярной поверхности
HeavyAtomMolWt	Average molecular weight of a molecule, without hydrogen atoms	Средний молекулярный вес молекулы, без атомов водорода
HallKierAlpha	Hall-Kier alpha value	Число Hall-Kier α
Chi3n	Simple molecular connectivity Chi indices for path order 3	Индекс χ простой молекулярной связи для пути третьего порядка
NHOHCount	NumHDonors	Число доноров водородной связи
BalabanJ	Balaban averaged distance sum connectivity Balaban's J index (J)	Индекс Балабана (J-индекс)
MolLogP	Molar logarithm of the partition coefficient (logP).	Молярный логарифм коэффициента распределения (logP).
FpDensityMorgan1	Fingerprint Density for Morgan Radius 3	Плотность молекулярных отпечатков Моргана радиуса 3
NumAliphaticRings	Number of aliphatic rings	Количество алифатических циклов

Наибольший вклад в температуру вспышки вносят специфические топологические индексы, содержащие некоторую информацию о форме молекулы. Сюда относятся топологический дескриптор Kappa1 (22.72), описывающий форму молекулы на основе распределения длин связей и углов наклона. Индекс Kappa1 рассчитывается относительно соединений с разной степенью разветвлений с тем же числом атомов, что и исследуемая молекула [9]. Этот дескриптор основаны на дескрипторе HallKierAlpha (5.34) и количестве возможных путей в молекуле определенной длины.

Вторым по значимости дескриптором является топологическая площадь полярной поверхности TPSA (10.42), которая представляет собой сумму вкладов в поверхность молекулы полярных атомов, таких как азот, кислород, и связанные с ними атомы водорода [9]. В медицинской химии TPSA используется для оценки способности молекулы проникать через гематоэнцефалический барьер [1]. Это широко используемый показатель в медицинской химии для оптимизации способности лекарственных средств проникать в клетки. Поскольку по определению температура вспышки - наименьшая температура летучего конденсированного вещества, при которой пары над поверхностью вещества способны самовоспламеняться в воздухе под воздействием источника огня, то вклад дескриптора TPSA в температуру вспышки не является удивительным.

Дескриптор HeavyAtomMolWt вычисляет среднюю молекулярную массу соединений, без учета атомов водорода. Отсюда следует что на температуру вспышки влияет количество тяжелых атомов в структуре молекулы.

Дескриптор Chi3n (значимость - 5.21), выбранный в этом исследовании, представляет собой индекс χ третьего порядка. Таким образом, дескриптор Chi3n кодирует информацию о всей молекулярной структуре, определяющую природу молекулы.

Количество алифатических циклов NumAliphaticRings (1.66) характеризует число алифатических (содержащих по крайней мере одну неароматическую связь) циклов в молекуле. Данный дескриптор характеризует количество тех циклов в молекуле, которые имеют неароматические связи.

Исходя из этого, температуру вспышки органических соединений можно рассматривать, как эмпирический функцию, зависящую от электронного строения молекул и взаимного расположения (топологии) атомов [1]. В качестве параметров, отражающих форму молекул, можно использовать параметры, которые характеризуют геометрию молекул: длины связей, валентные углы. Данные параметры определяют возможность конформационных перестроек молекулы.

Несмотря на это, нами построены адекватные регрессионные модели, на основе которых проведен прогноз определенных классов соединений для соответствующих свойств [20-23]. Поэтому наши модели могут быть полезным инструментом, обеспечивающим безопасность химических производств.

Работа выполнена в рамках реализации Программы развития ФГБОУ ВО «КНИТУ» (ПСАЛ Приоритет 2030).

Литература

1.М.Ю. Доломатов, О.С. Коледин, Э.А. Ковалева, Р.А. Федина, Р.В. Гарипов, М.Р. Валеев, *Башкирский химический журнал*, **29**, 2, 65-70 (2022).
2. В.С. Конышев, А.Д. Лифанов, К.Ю. Никитина, Вестник технологического университета, **28**, 4, 112-117 (2025).
3. J. H. Friedman, *Ann. Statist*, **29**, 5, 1189-1232 (2001).
4. S. Fei, Y. Qi, W. Liu, Y. Wang, Z. Wang, H. Zhang, Combustion Science and Technology, **7**, 1-19 (2021).
5. http://www.pubchem.com/.
6. D. Weininger, *Journal of Chemical Information and Computer Sciences*, **28**, 1, 31-36 (1988).
7. D. Weininger, A. Weininger, J.L. Weininger, *Journal of Chemical Information and Computer Sciences*, **29**, 2, 97-101 (1989).
8. R. Todeschini, V. Consonni, New York: Wiley-VCH, 1-680 (2000).
9. https://rdkit.org/
10. J. Gasteiger, T. Engel, Weinheim: Wiley-VCH (2003).
11. J.L. Durant, B.A. Leland, D.R. Henry, J.G. Nourse, *J. Chem. Inf. Comput. Sci*., **42**, 6, 1273-1280 (2002).
12. О.С. Коледин, М.Ю. Доломатов, Р.Ш. Япаев, А.Т. Гильмутдинов, М.Ф. Мухарметов, Р.В. Гарипов, М.Р. Валеев, *Журнал прикладной химии*, 95, 5, 666-671 (2022).
13. H.L. Morgan, *Journal of Chemical Documentation*, **5**, 2, 107-113 (1965).
14. L. Breiman, *Machine Learning*, **45**, 1, 5-32 (2001).
15. Fix, E., J.J. Hodges, Discriminatory analysis: Non-parametric discrimination: Consistency properties. Technical report. USAF School of Aviation Medicine (1951).
16. Fix, E., J.J. Hodges, Discriminatory analysis: Non-parametric discrimination: Small sample performance. Technical report. USAF School of Aviation Medicine (1952).
17. Cortes, C., V. Vapnik, Support-vector networks. *Machine Learning*, **20**, 3, 273-297 (1995).
18. S.B. Gunturi, K. Archana, A. Khandelwal, R. Narayanan, *QSAR & Combinatorial Science*, **27**, 11-12, 1305–1317 (2008).
19. M. Budka, B. Gabrys, *Procedia Computer Science*, **1**, 1, 193–201 (2010).
20. Schein, A.I., L.H. Ungar, *Machine Learning*, **68**, 3, 235-265 (2007).
21. Gramatica, P. *QSAR & Comb Sci*, 694–701 (2007).
22. Meloun M, Militku J, Hill M. *Analyst*, **127**, 433–450 (2002).
23. Tharwat A. *Applied Computing and Informatics*, **17**, 1, 168–192 (2021).
24. А.Д. Лифанов, А.А. Фатыхова, К.Ю. Никитина, *Вестник Технол. ун-та*, **28**, 1, 142-146 (2025).
25. А.Д. Лифанов, К.Ю. Никитина, Е.Г. Лифанова, *Вестник Технологического университета*, **28**, 2, 66-69 (2025).
26. D. Bonchev, N. Trinajstić: Information theory, distance matrix, and molecular branching. J. Chem. Phys. 67 (1977) 4517-4533.

References

1. M.Y. Dolomatov, O.S. Koledin, E.A. Kovaleva, R.A. Fedina, R.V. Garipov, M.R. Valeev, Bashkir Chemical Journal, 29, 2, 65-70 (2022).
2. V.S. Konyshev, A.D. Lifanov, K.Yu. Nikitina, Herald of Technological University, 28, 4, 112-117 (2025).
3. J. H. Friedman, Ann. Statist, 29, 5, 1189-1232 (2001).
4. S. Fei, Y. Qi, W. Liu, Y. Wang, Z. Wang, H. Zhang, Combustion Science and Technology, 7, 1-19 (2021).
5. http://www.chemspider.com/.
6. D. Weininger, Journal of Chemical Information and Computer Sciences, 28, 1, 31-36 (1988).
7. D. Weininger, A. Weininger, J.L. Weininger, Journal of Chemical Information and Computer Sciences, 29, 2, 97-101 (1989).
8. R. Todeschini, V. Consonni, New York: Wiley-VCH, 1-680 (2000).
9. https://rdkit.org/
10. J. Gasteiger, T. Engel, Weinheim: Wiley-VCH (2003).
11. J.L. Durant, B.A. Leland, D.R. Henry, J.G. Nourse, J. Chem. Inf. Comput. Sci., 42, 6, 1273-1280 (2002).
12. O.S. Koledin, M.Yu. Dolomatov, R.Sh. Yapaev, A.T. Gil'mutdinov, M.F. Muharmetov, R.V. Garipov, M.R. Valeev, Journal of Applied Chemistry, **95**, 5, 666-671 (2022).
13. H.L. Morgan, Journal of Chemical Documentation, 5, 2, 107-113 (1965).
14. L. Breiman, Machine Learning, 45, 1, 5-32 (2001).
15. Fix, E., J.J. Hodges, Discriminatory analysis: Non-parametric discrimination: Consistency properties. Technical report. USAF School of Aviation Medicine (1951).
16. Fix, E., J.J. Hodges, Discriminatory analysis: Non-parametric discrimination: Small sample performance. Technical report. USAF School of Aviation Medicine (1952).
17. Cortes, C., V. Vapnik, Support-vector networks. Machine Learning, 20, 3, 273-297 (1995).
18. S.B. Gunturi, K. Archana, A. Khandelwal, R. Narayanan, QSAR & Combinatorial Science, 27, 11-12, 1305–1317 (2008).
19. M. Budka, B. Gabrys, Procedia Computer Science, 1, 1, 193–201 (2010).
20. Schein, A.I., L.H. Ungar, Machine Learning, 68, 3, 235-265 (2007).
21. Gramatica, P. QSAR & Comb Sci, 694–701 (2007).
22. Meloun M, Militku J, Hill M. Analyst, 127, 433–450 (2002).
23. Tharwat A. Applied Computing and Informatics, 17, 1, 168–192 (2021).
24. A.D. Lifanov, A.A. Fatyhova, K.Yu. Nikitina, Herald of Technological University, 28, 1, 142-146 (2025).
25. A.D. Lifanov, K.Yu. Nikitina, E.G. Lifanova, Herald of Technological University 28, 2, 66-69 (2025).

© **А. Д. Лифанов** – к.х.н., доцент кафедры Общей химической технологии, Казанский национальный исследовательский технологический университет, Казань, Россия, lifanov84@mail.ru; **Е. Г. Лифанова** – ассистент кафедры Коммуникативного дизайна, Казанский (Приволжский) федеральный университет, 678769@mail.ru.

© **A. D. Lifanov** – PhD (Chemical Sci.), Associate Professor of the Department of General Chemical Technology, Kazan National Research Technological University, Kazan, Russia, lifanov84@mail.ru; **E. G. Lifanova** – Assistant of the Department of Communicative Design, Kazan Federal University, Kazan, Russia, 678769@mail.ru.

Дата поступления рукописи в редакцию – 04.05.25.
Дата принятия рукописи в печать – 11.09.25.