

В. С. Конышев, А. Д. Лифанов, К. Ю. Никитина

**ПРИМЕНЕНИЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ
ДЛЯ ПРОГНОЗИРОВАНИЯ ТЕМПЕРАТУРЫ ВСПЫШКИ ОРГАНИЧЕСКИХ СОЕДИНЕНИЙ***Ключевые слова: большие данные, индустрия 4.0, температура вспышки, машинное обучение, искусственный интеллект.*

Повышение требований по безопасности субъектов производственной деятельности привело к ряду технологических изменений в направлении усиления мер по промышленной безопасности. В настоящее время обязательным является разработка паспорта безопасности производственных объектов, особенно если химическое производство связано с использованием горючих жидкостей. При этом, приоритетным в создании безопасных химико-технологических систем является использование цифровых технологий. Данное направление является частью Индустрии 4.0. Во многих странах мерой разделения жидкостей на горючие и легковоспламеняемые является температура вспышки. Данные о температуре вспышки для органических соединений были взяты из базы данных PubChem. В базу данных для данной работы была включена информация о температуре вспышки для 1741 органических веществ. Для упрощения анализа представления органических соединений, мы использовали ключи MACCS, 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 были реализованы гребневая регрессия, алгоритм случайного леса, метод ближайших соседей kNN, метод опорных векторов (SVM), искусственные нейронные сети. Для обучающей выборки полученная классификационная модель случайного леса показала безошибочную классификацию, ошибка прогноза для нее равна 0. Статистические характеристики построенной RF-модели для выборки имеют следующие значения: точность ACC = 0.83, специфичность SPC = 0.81, чувствительность SEN = 0.86, сбалансированная точность BA = 0.83, коэффициент корреляции Мэтью MCC = 0.72. На основе разработанной модели был проведен прогноз принадлежности соединений к определенному классу («1» класс горючий или «0» легковоспламеняемый) для органических соединений, для которых отсутствует экспериментальная информация относительно принадлежности к классу «1» или «0».

V. S. Konyshov, A. D. Lifanov, K. Yu. Nikitin

**APPLICATION OF MACHINE LEARNING METHODS TO PREDICT THE FLASH POINT
OF ORGANIC COMPOUNDS***Keywords: big data, industry 4.0, flashpoint temperature, machine learning, artificial intelligence.*

The increase in safety requirements for industrial entities has led to a number of technological changes in the direction of strengthening industrial safety measures. Currently, it is mandatory to develop a safety data sheet for production facilities, especially if chemical production is associated with the use of flammable liquids. At the same time, the use of digital technologies is a priority in creating safe chemical technology systems. This area is part of Industry 4.0. In many countries, the flash point is a measure of the separation of liquids into flammable and flammable. Flash point data for organic compounds were taken from the PubChem database. Information on the flash point for 1,741 organic substances was included in the database for this work. To simplify the analysis of the representation of organic compounds, we used MACC keys, 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 molecular prints, also known as circular prints with a radius of 2. Within the framework of this work, ridge regression, the random forest algorithm, the kNN nearest neighbor method, the support vector machine (SVM) method, and artificial neural networks were implemented. For the training sample, the resulting classification model of a random forest showed an error-free classification, the prediction error for it is 0. The statistical characteristics of the constructed RF model for the sample have the following values: accuracy ACC = 0.83, sensitivity SPC = 0.81, specificity SEN = 0.86, balanced accuracy BA = 0.83, Matthews correlation coefficient MCC = 0.72. On the basis of the developed model, a forecast was made for the belonging of compounds to a certain class ("1" class combustible or "0" flammable) for organic compounds for which there is no experimental information regarding belonging to class "1" or "0".

Введение

Повышение требований по безопасности субъектов производственной деятельности привело к ряду технологических изменений в направлении усиления мер по промышленной безопасности. В настоящее время обязательным является разработка паспорта безопасности производственных объектов, особенно если химическое производство связано с использованием горючих и легковоспламеняемых жидкостей.

Во многих странах мерой разделения жидкостей на горючие и легковоспламеняемые является температура вспышки. Температура вспышки одна из важнейших характеристик горючести органических соединений [1]. Она определяется как нижняя граница температуры, при которой смесь паров данного вещества с воздухом может быть подожжена при инициировании.

Развитие экспериментальных методов определения температуры вспышки привело к накоплению большого числа экспериментальных

данных [2], которые можно обобщить и проанализировать при необходимости. Допускается использовать как расчетные, так и экспериментальные значения температура вспышки (T_f) [3]. Величины T_f экспериментально определены для многих химических соединений. Однако большинство из них не публикуются в открытых источниках. Имеется ограниченное количество публикаций по применению моделей типа «структура-свойство» для отдельных классов органических соединений, в частности алканов [1] и спиртов [2]. Кроме того, для летучих и токсичных веществ экспериментальное определение этой величины весьма сложно. Поэтому теоретические методы оценки T_f имеют ограничения.

Приоритетным в создании безопасных химико-технологических систем является использование цифровых технологий. Данное направление является частью Индустрии 4.0.

В связи с этим, целью настоящей статьи является разработка модели «структура-свойство» для прогнозирования температуры вспышки органических соединений с применением методов машинного обучения.

Методы исследования

Традиционно, построение модели QSPR (англ. Quantitative Structure-Property Relationship-количественные соотношения «структура-свойство») проходит несколько этапов: подготовка данных, выбор модели, отбор дескрипторов, обучение модели и анализ результатов [4].

Данные о T_f для органических соединений были взяты из базы данных PubChem [5]. В базу данных для данной работы была включена информация о температуре вспышки для 1741 органических веществ.

Собранный набор данных о свойствах органических соединений включал 1741 с их T_f в диапазоне от 85.15 до 698.15 К. Распределение значений температур вспышки органических соединений представлен на рисунке 1.

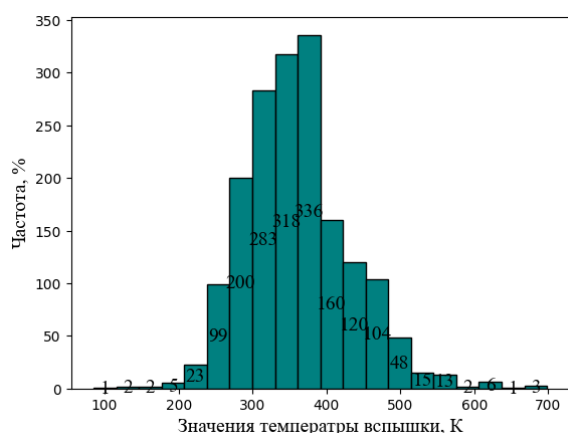


Рис. 1 – Распределение значений температур вспышки органических соединений

Fig. 1 – The distribution of flask point values of organic compounds

В нашей работе в качестве линейных представлений выбрана система SMILES [6, 7]. Далее на основе SMILES генерируется дескриптор, отражающий особенности структуры органических соединений [8]. Данные дескрипторы рассчитаны с использованием библиотеки RDKit 2. RDKit представляет собой инструментальный для хемоинформатики с открытым исходным кодом [9].

В качестве дескрипторов, отражающих структуры органических соединений, использованы молекулярные отпечатки (Morgan's fingerprints) [10, 11] с параметрами: радиус 2, длина битовой строки 1024 бит (чем больше радиус, тем большее окружение атомов учитывается при кодировании фрагментов; чем больше длина битовой строки, тем больше информации о структурных фрагментах можно закодировать и избежать потери информации о связях «структура-свойство»). Иногда в научной литературе можно встретить другое обозначение – ECPF4 (Extended Connectivity Fingerprint) или циркулярный молекулярный отпечаток. Также использовали структурные ключи (MACCS keys) [12, 13]. Отличие структурных ключей от молекулярных отпечатков заключается в том, что при использовании структурных ключей используется список предопределенных фрагментов, а не рассчитывается для конкретной базы данных химических соединений.

Полученные молекулярные отпечатки и структурные ключи используются в качестве входных данных для обучения модели, которая может быть реализована на основе любого метода машинного обучения [13].

В рамках данной работы были реализованы гребневая регрессия, алгоритм случайного леса (Random Forest) [14], метод ближайших соседей (kNN) [15, 16], метод опорных векторов (SVR) [17], искусственные нейронные сети (MLP) [12], градиентный бустинг (XGBoost) [18], AdaBoostRegressor (AdaBoost) [19]. Для построения классификационных моделей были выбраны методы логистическую регрессию (Logistic Regression) [20], алгоритм случайного леса (RandomForest-Classifer) [14], ансамблевые методы машинного обучения (AdaBoostClassifier) [19], многослойный перцептрон (MLPClassifier) [12].

В нашей модели с помощью встроенной функции Scikit-learn - Train test split исходный набор данных разделили на обучающий и тестовый в соотношении 80/20 [21].

Для повышения надежности моделей «структура-свойство» осуществляли процедуру перекрестного скользящего контроля (кросс-валидация) k-fold, которая используется для оценки эффективности и возможности обобщения модели. Иногда данную процедуру обозначают как CV для краткости (от англ. cross-validation – кросс-валидация). Она заключается в разделении набора данных на k подмножеств (блоков) и итеративном использовании k-1 блоков для тренировки и одного для перекрестной проверки модели. Процесс повторяется k раз, где каждый блок используется для перекрестной проверки ровно один раз.

Процедура обучения модели может включать определение гиперпараметров модели. Гиперпараметры модели – характеристики модели, фиксируемые перед началом обучения модели. В алгоритме случайного леса в качестве гиперпараметров выбирают число деревьев, глубина построения дерева и т.д. Данная процедура осуществлялась с помощью инструмента GridSearchCV [22].

Коэффициент детерминации (R^2) рассчитывается следующим образом:

$$R^2 = 1 - \frac{\sum_i (y_i^{\text{pred}} - y_i^{\text{exp}})^2}{\sum_i (y_i^{\text{pred}} - \bar{y}^{\text{exp}})^2} \quad (1)$$

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

Как и коэффициент детерминации R^2 (1), статистический показатель Q^2 (тот же R^2 , после процедуры кросс-валидации) не может принимать значения, превышающие 1, причем, чем ближе его значение к 1, тем выше средняя прогнозирующая способность моделей [22].

Следующим важным статистическим критерием является среднеквадратичная ошибка (RMSE) (2). Среднеквадратичная ошибка RMSE может принимать любое положительное значение, включая ноль, причем, чем ниже, тем выше средняя прогнозирующая способность моделей, оцененная посредством процедуры 5-кратного скользящего контроля. RMSE рассчитывается следующим образом:

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

Таким образом, показатели RMSE и Q^2 являются универсальными показателями прогнозирующей способности моделей [21].

Классификационные модели также характеризуются своими количественными показателями [21].

Для оценки качества классификационных моделей используют следующие параметры [21]:

TN (true negative) - количество правильно классифицированных отрицательных примеров,

TP (true positive) - количество достоверно классифицированных положительных примеров,

FP (false positive) - количество неверно классифицированных положительных примеров,

FN (false negative) - количество неверно классифицированных отрицательных примеров.

Из этих четырех величин формируется Матрица конфликтов для двух классов классификации (Таблица 1). Следствием этого является наличие

разнообразных показателей качества классификационных моделей [21].

Таблица 1 – Матрица конфликтов для бинарной классификации

Table 1 – Conflict matrix for binary classification

		Экспериментальные данные	
		Класс 1	Класс 2
Предсказанные значения	Класс 1	TP	FP
	Класс 2	FN	TN

Как и в случае регрессии, качество моделей оценивается с помощью показателей описательной и прогнозирующей способности. Расчет этих показателей ведется по одним и тем же формулам, но для объектов из разных выборок: обучающей и тестовой, соответственно [23].

Самым простым показателем для оценки качества классификационных моделей есть точность (ACC, accuracy), представляющая собой долю верно классифицированных примеров [23]:

$$ACC = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (3)$$

Также используют такие показатели, как чувствительность (SEN, sensitivity), что характеризует долю корректно классифицированных объектов первого класса («активных») и специфичность (SPC, specificity), характеризующий частицу корректно классифицированных объектов второго класса («неактивных») [23]:

$$SEN = \frac{TP}{(TP + FN)} \quad (4)$$

$$SPC = \frac{TN}{(TN + FP)} \quad (5)$$

Используется также среднее этих двух величин, что есть - сбалансированной точностью (BA) [23]:

$$BA = \frac{(SEN + SPC)}{2} \quad (6)$$

Для оценки качества классификационных задач также используется коэффициент корреляции Мэтью (MCC), который рассчитывается по формуле [23]:

$$MCC = \frac{TP \cdot TN}{\sqrt{(TP + FP) \cdot (TP + FN)}} \times \frac{FP \cdot FN}{\sqrt{(TN + FP) \cdot (TN + FN)}} \quad (7)$$

Результаты и их обсуждение

В результате исследования мы получили адекватные модели «структура-свойство» с высокими статистическими показателями (Табл. 2). Наибольшей предсказательной способностью обладает XGBoost-модель с использованием структурных ключей MACCS в качестве дескрипторов с коэффициентом детерминации для тестовой выборки $R^2 = 0.60$ и ошибкой предсказания $RMSE = 32.25$ К (Табл. 2). После процедуры 5-кратного скользящего контроля коэффициент детерминации $R^2 = 0.63$ немного улучшился, но ошибка предсказания стала больше $RMSE = 43.52$.

Несмотря на то, что при использовании структурных ключей используется список predetermined фрагментов, а не рассчитывается для конкретной базы данных химических соединений, прогностическая способность моделей с использованием этих дескрипторов выше.

Таблица 2 – Результаты прогнозирования температуры вспышки органических соединений с использованием регрессионных моделей

Table 2 – The results of forecasting the flash point of organic compounds using regression models

Модель	Дескрипторы	Без кросс-валидации		С кросс-валидацией	
		R^2	RMSE	R^2	RMSE
Ridge	MACCS	0,53	34,32	0,54	48,42
	ECPF4	0,23	44,70	0,51	49,91
AdaBoost	MACCS	0,44	41,33	0,44	53,12
	ECPF4	0,26	49,82	0,29	60,06
RF	ECPF4	0,42	37,25	0,45	52,67
	MACCS	0,52	49,30	0,53	48,82
XGBoost	MACCS	0,60	32,27	0,63	43,52
	ECPF4	0,49	37,46	0,53	48,94
kNN	MACCS	0,57	34,12	0,56	47,15
	ECPF4	-	51,37	0,16	65,40
MLP	MACCS	0,59	30,91	0,59	45,54
	ECPF4	-	51,41	-	73,97
SVR	MACCS	0,24	46,47	0,49	50,66
	ECPF4	0,18	48,72	0,50	50,19

Для прогноза и интерпретации использовалась модель как результат 5-кратного скользящего контроля (Рис. 2) [21].

Выборка хорошо сбалансирована, количество горючих (класс «1») и легковоспламеняемых («0») органических соединений составляет 1087 и 654 соответственно. Используя набор методов машинного обучения, мы получили классификационные QSAR модели, которые характеризуется удовлетворительной прогностической способностью. Оценка модели осуществлялась в соответствии с внутренней процедурой скользящего контроля моделей: ошибка прогноза для представленной выборки равна 17%.

При визуальной оценке расположение ROC-кривых относительно друг друга показывает прогностическую способность модели. Чем больше

площадь под ROC-кривой, тем точнее предсказывается принадлежность объекта определенному классу. Поскольку качество отдельных моделей находилось на приемлемом уровне, нами построены классификационные модели (Рис. 3).

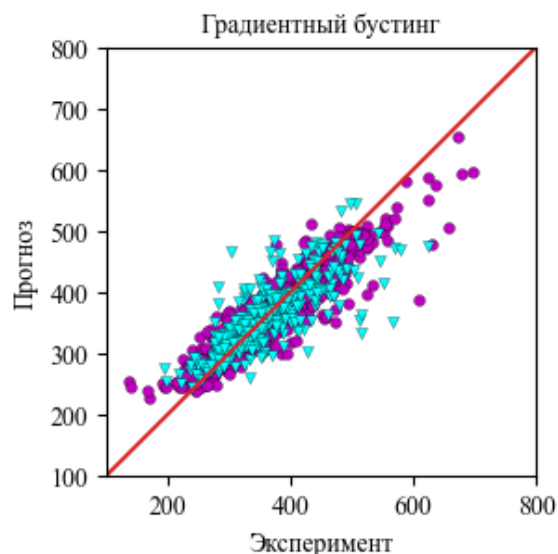


Рис. 2 – Регрессионная модель градиентного бустинга для прогнозирования пределов вспышки органических соединений

Fig. 2 – Regression model of gradient boosting for predicting the flash limits of organic compounds

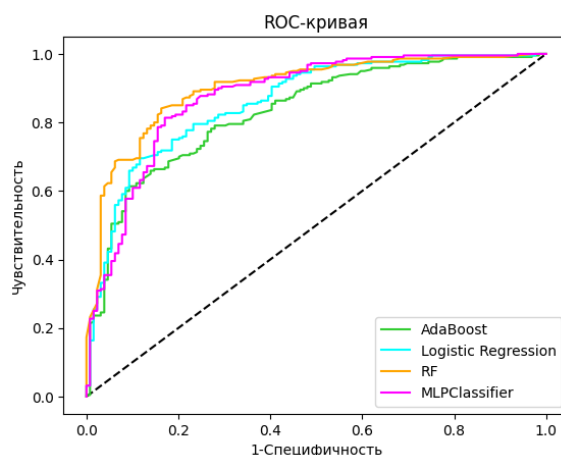


Рис. 3 – Классификационные модели прогнозирования пределов вспышки органических соединений

Fig. 3 – Classification models for predicting the flash limits of organic compounds

Из рис. 3 можно видеть, что ROC-кривая для RF-модели находится выше и левее остальных. Следовательно, предсказательная способность данной модели выше.

На основе результатов прогнозируемых классов для выборки по RF-модели на Рис. 4 представлена матрица конфликтов, где TN и TP – количество верно

установленных классов «0» (легковоспламеняющаяся жидкость) и классов «1» (горючая жидкость), соответственно, FP и FN – количество неверно установленных классов «1» и «0», соответственно (обозначения представлены в Таблице 1, где true label соответствует истинной метке класса, predicted label – метки класса, предсказанной моделью).

Для обучающей выборки полученная классификационная модель случайного леса показала безошибочную классификацию, ошибка прогноза для нее равна 0.

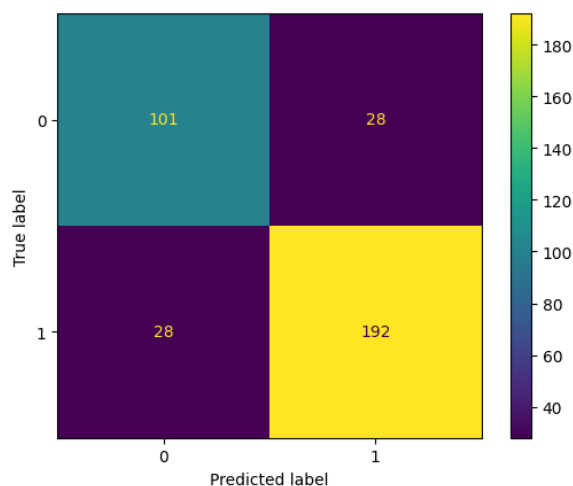


Рис. 4 – Матрица конфликтов для выборки (RF-модель)

Fig. 4 – Conflict matrix for the sample (RF- model)

Статистические характеристики построенной RF-модели для выборки имеют следующие значения: ACC = 0.83, SPC = 0.81, SEN = 0.86, BA = 0.83, MCC = 0.72. На основе разработанной модели был проведен прогноз принадлежности соединений к определенному классу («1» класс горючий или «0» легковоспламеняемый) для органических соединений, для которых отсутствует экспериментальная информация относительно принадлежности к классу «1» или «0».

Анализ прогнозируемых классов для 1471 исследуемых органических соединений показал, что потенциально только 654 среди них являются легковоспламеняемыми. Анализ применимости модели показал, что все исследуемые молекулы входят в область применимости модели [23].

В результате анализа фрагментов горючих и легковоспламеняемых жидкостей не выявили четкой тенденции между вкладами фрагментов. Таким образом, мы получили адекватные классификационные модели, на основе которых проведен прогноз определенных классов соединений для соответствующих свойств, а также На провели сравнительный анализ фрагментов, способствующих или препятствующих горению органических веществ.

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

Литература

1. А.Я. Корольченко, Д.А. Корольченко, *Пожаровзрывоопасность веществ и материалов и средства их тушения*. Справочник: ч.1, Асс. «Пожнаука», Москва, 2004, 713 с.
2. М.Ю. Долوماتов, О.С. Коледин, К.Р. Ахтямова, *Известия высших учебных заведений. Сер. Химия и химическая технология*, **64**, 7, 96-103 (2021).
3. М.Ю. Долوماتов, О.С. Коледин, Э.А. Ковалева, Р.А. Федина, Р.В. Гарипов, М.Р. Валеев, *Башкирский химический журнал*, **29**, 2, 65-70 (2022).
4. T.I. Madzhidov, A. Rakhimbekova, V.A. Afonina, *Mendeleev Communications*, **31**, 6, 769-780 (2021).
5. <https://pubchem.ncbi.nlm.nih.gov/>.
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. C. Kuenneth, W. Schertzer, R. Ramprasad, *Macromolecules*, **54**, 13, 5957-5961 (2021).
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).

References

1. A.Ya. Korolchenko, D.A. Korolchenko Fire and explosion hazard of substances and materials and their extinguishing agents. Handbook: 2 part. Ass. "Pozhnauka", Moskow, 2004, 713 p.
2. M.Y. Dolomatov, O.S. Koledin, K.R. Akhtyamova, *Izvestiya vysshikh uchebnykh zavedeniy. Ser. Chemistry and Chemical technology*, **64**, 7, 96-103 (2021).
3. 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).
4. T.I. Madzhidov, A. Rakhimbekova, V.A. Afonina, *Mendeleev Communications*, **31**, 6, 769-780 (2021).
5. <https://pubchem.ncbi.nlm.nih.gov/>.

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. C. Kuenneth, W. Schertzer, R. Ramprasad, Macromolecules, 54, 13, 5957–5961 (2021).
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).

© **В. С. Коньшев** – студент кафедры общей химической технологии (ОХТ), Казанский национальный исследовательский технологический университет (КНИТУ), Казань, Россия, 678769@mail.ru; **А. Д. Лифанов** – к.х.н., доцент кафедры ОХТ, КНИТУ, lifanov84@mail.ru; **К. Ю. Никитина** – студент кафедры Химии и технологии переработки эластомеров, КНИТУ, karina2004nikitina@mail.ru

© **V. S. Konyshev** – Student of the Department of General Chemical Technology (GCT), Kazan National Research Technological University (KNRTU), Kazan, Russia, 678769@mail.ru; **A. D. Lifanov** – PhD (Chemical Sci.), Associate Professor of the GCT Department, KNRTU, lifanov84@mail.ru; **K. Yu. Nikitina** – Student of the Department of Chemistry and Processing Technology of Elastomers, KNRTU, karina2004nikitina@mail.ru;

Дата поступления рукописи в редакцию – 09.12.24.

Дата принятия рукописи в печать – 14.01.25.