

УДК 579.63

DOI: 10.46548/21vek-2022-1157-0022

**ПРОГНОЗИРОВАНИЕ ПАТОГЕННОСТИ МИКРОБА КАК НЕГАТИВНОГО ФАКТОРА  
ПРОИЗВОДСТВЕННОЙ СРЕДЫ НА ОСНОВЕ КОМПЛЕКСА  
КОСВЕННЫХ НАТИВНЫХ ПРИЗНАКОВ**

© 2022

**Земскова Анастасия Романовна**, магистрант**Липина Татьяна Олеговна**, бакалавр**Кузьмин Антон Алексеевич**, кандидат биологических наук,

доцент кафедры «Биотехнологии и техносферная безопасность»

*Пензенский государственный технологический университет*

(440039, г. Пенза, проезд Байдукова/ул. Гагарина, д. 1а/11, kuzmin-puh@yandex.ru)

**Аннотация.** Работа посвящена оценке эффективности определения патогенности микроба на основе комплекса нативных признаков, не связанных напрямую с опасностью для здоровья человека (форма клеток, окраска по Грамму, оптимальные  $pH$  и температура, тип местообитания, подвижность), определяемых непосредственно в среде обитания микроорганизма (*in vivo*), с целью разработки методов анализа и прогноза профессиональной заболеваемости. Кластеризацию образцов проводили методом *Random Forest* без обучения. Качество модели оценивали с помощью *SVM*-алгоритма по показателям точности (общему проценту правильно классифицированных образцов) и чувствительности (способности модели правильно классифицировать патогенные образцы), рассчитываемым по матрице ошибок. Для определения перечня значимых для модели признаков построили пять моделей, последовательно исключая предикты с минимальным значением индекса чистоты. Наиболее точной оказалась классификационная модель №3 (точность – 78,21%, чувствительность – 69,23%), построенная на четырех предиктах – форма клеток, тип местообитания, оптимальные для жизни микроба  $pH$  и температуры. Наиболее чувствительной (и при этом наименее точной), а, следовательно, наиболее эффективной в определении патогенности микроба стала модель №4 (точность – 66,67%, чувствительность – 87,18%), включающая признаки типа местообитания и оптимумов  $pH$  и температуры.

**Ключевые слова:** патогенность, производственная среда, опасные и вредные факторы, профессиональная заболеваемость, микробы, бактерии, тип местообитания, оптимум, подвижность, форма клеток, окраска по Грамму, случайный лес, *SVM*-алгоритм, матрица ошибок, машинное обучение.

**MICROBE'S PATHOGENICITY PREDICTION ON COMPLEX OF INDERECTIVE NATIVE FEATURES**

© 2022

**Zemskova Anastasiya Romanovna**, master's student of Biotechnology and Technosphere Safety Department**Lipina Tatyana Olegovna**, student of Biotechnology and Technosphere Safety Department**Kuzmin Anton Alekseevich**, candidate of biological sciences,

associate professor of Biotechnology and Technosphere Safety Department

*Penza state technological university*

(Russia, 440039, Penza, Pr. Baidukova/Gagarina Street, 1a / 11, kuzmin-puh@yandex.ru)

**Abstract.** Article is devoted to quality evaluation of a microbe's pathogenicity detection on complex of native features directly not related with human health hazard (cell shape, Gram stain,  $pH$  and temperature optimums, habitat, motility), obtained *in vivo* for developing methods of occupational morbidity analysis and prediction. Samples clusterization was conducted using unsupervised *Random Forest* algorithm. Model quality was estimated with *SVM*-algorithm on parameters of Accuracy and Sensitivity (model's ability to classify pathogenic samples correctly) from confusion matrix. To reveal significant predicts five models were built consequentially excluding predicts with lowest Gini impurity. Most accurate model (#3, Accuracy – 78,21%, Sensitivity – 69,23%) were built on four predicts – cell shape, habitat,  $pH$  and temperature optimums. Most sensitive (wherein less accurate) and consequently most effective model (#4, Accuracy – 66,67%, Sensitivity – 87,18%) in a microbe's pathogenicity detection. This model built on three predicts – habitat,  $pH$  and temperature optimums.

**Keywords:** pathogenicity, work environment, dangerous and harmful factors, occupational morbidity, microbes, bacteria, habitat, optimum, motility, cell shape, Gram stain, Random Forest, *SVM*-algorithm, confusion matrix, machine learning.

**Введение.** Микрофлора, являясь неотъемлемой частью производственной среды, относится к биологическим ее факторам и влияет на состояние здоровья работника, обуславливая определенный уровень профессиональной заболеваемости. Поэтому разработка методов учета, анализа и прогноза патогенности представителей микрофлоры является актуальным

направлением развития охраны труда и безопасности деятельности человека.

Патогенность (от др.-греч. *πάθος* – страдание, болезнь и *γένεσις* – возникновение, первоисточник) – способность быть причиной (порождать) патологии (болезни, отклонения от нормы) – полидетерминантная, генотипическая характеристика определённого

микроорганизма или вируса, ответственная за создание специфических структур (например, капсула, экзотоксины) или отвечающая за поведение, нарушающее целостность тканей организма животных или человека. Патогенность характеризуется специфичностью, то есть способностью вызывать типичные для определённого возбудителя патофизиологические и морфологические изменения в определённых тканях и органах, при условии естественного для него способа заражения [1].

Для разработки адекватных мер обеспечения инфекционной безопасности микроорганизмы классифицируются в зависимости от степени их патогенности. Микробы, способные вызывать патологии у человека, классифицируются от условно-патогенных микроорганизмов до возбудителей особо опасных заболеваний [2].

Разделение на группы патогенности позволяет, кроме оценки рисков, создавать отдельные требования для каждой из групп по выделению чистой культуры, хранению и транспортировке материала, допуску и условиям работы с ним, а также по проведению других профилактических и противоэпидемических мероприятий, в том числе режимно-ограничительных, с целью недопущения заражения и распространения инфекций [3].

В соответствии с этим в России приняты санитарные правила, устанавливающие требования к организационным, санитарно-противоэпидемическим (профилактическим) мероприятиям, направленным на обеспечение личной и общественной безопасности, защиту окружающей среды при работе с патогенными биологическими агентами [4].

Патогенность микроба является не только важным параметром биологической безопасности среды обитания человека, но и лимитирует использование микроорганизма в качестве продуцента в промышленном производстве биологически активных веществ, а также в качестве объекта научных биотехнологических разработок.

По причине трудоемкости методики [5] прямых исследований опасных для человека свойств микроба всегда сопряжены с риском утечки биоматериалов и потенциального инфицирования больших групп населения в краткие сроки [6]. Дополнительными негативными факторами заражения могут стать высокие темпы мутации [7], определяющие резистентность к имеющимся антидотам [8], а также сочетание факторов новой среды обитания, превращающих условно-патогенные микробы в патогенные [9, 10].

Патогенность микроорганизма нельзя рассматривать в отрыве от других его нативных биологических признаков и особенностей экологии [11]. Логично предположить, что способность вызывать заболевания человека может зависеть от биотопических предпочтений микроба, его подвижности, кислотности среды, температурного оптимума, особенностей цитоморфологии (форма клеток, окраска по Граму), т.е. признаков, не связанных с патогенностью напрямую.

Современные методы машинного обучения позволяют строить и обучать модели, способные прогнозировать те или иные свойства объектов на основе анализа имеющихся данных о выборке [12]. Развитие этого направления исследований позволит не только снизить риски прямых исследований опасных микробов, но и предсказывать их патогенность на основе комплекса нативных, доступных для сбора и объективных характеристик.

**Цель работы** – построить эффективную модель, прогнозирующую патогенность микроба по комплексу биоэкологических признаков.

**Материал и результаты исследования.** Материал работы получен при анализе особенностей биологии представителей микробиоты ( $n = 78$ ) [13]. Патогенность микроба прогнозировалась на основе комплекса нативных признаков, напрямую не связанных с патогенностью, определяемых непосредственно в среде обитания организма (*in vivo*): морфологических (форма клеток и окраска по Граму) и экологических (оптимальные  $pH$  и температура, тип местообитания, подвижность). Кластеризацию образцов проводили методом *Random Forest* без обучения. Т.е. информация о патогенности образцов не входила в исходные данные модели [14]. Преимуществами алгоритма являются возможность анализировать как категориальные (качественные), так и числовые (количественные) признаки, а также «встроенная» в алгоритм кроссвалидация, при которой часть выборки, полученная случайным образом, используется для тренировки модели, а другая часть (*out of bootstrap*) – для проверки модели [15]. Значимость признака для классификации определяли с помощью индекса чистоты (*Gini impurity*) [16]. Качество полученной модели оценивали с помощью *SVM*-алгоритма [17], по общему проценту правильно классифицированных образцов [18], показателям чувствительности и специфичности [19], доле ложно-отрицательных образцов (реально патогенных образцов, ошибочно классифицированных моделью как непатогенные) и другим индексам, рассчитываемым по матрице ошибок [20–27]. Для определения перечня значимых для модели признаков, последовательно исключали предикты с минимальным значением индекса чистоты при контроле потери качества классификации. Структурирование данных проводили в программе *Microsoft Excel* [28], анализ и графическое представление результатов – с помощью языка программирования *R* [29], в среде программирования *RStudio* [30].

**Модель №1** получена при классификации выборки по патогенности на основе шести предиктов. Выбор метода без обучения обусловлен отсутствием прямой связи между признаками микроорганизмов и их патогенностью. Точность классификации составила 74,36%, чувствительность (способность модели правильно классифицировать патогенные образцы) – 64,1%. По нашему мнению, последний показатель более уместен при выявлении опасности (или безопасности) микроба для здоровья человека

и информативен с точки зрения поставленной задачи исследования – выявления патогенности по нативным признакам, напрямую с ней не связанным.

Значения индексов чистоты предиктов представлены в таблице 1.

Таблица 1 – Значения индекса чистоты (Gini impurity index) предиктов классификационных моделей

Признак	Модель №1	Модель №2	Модель №3	Модель №4	Модель №5
Окраска по Граму	3.16	-	-	-	-
Тип местообитания	9.83	10.37	11.28	7.84	-
Подвижность	3.59	4.09	-	-	-
Форма клеток	6.57	7.05	7.09	-	-
Оптимум pH	19.32	21.75	24.96	17.82	28.68
Оптимальная температура	20.01	23.06	26.54	18.99	28.93

По значениям индекса чистоты предикты располагаются в следующей последовательности (по уменьшению значимости для качества классификации): оптимальная температура, оптимум pH (оба – числовые, количественные), тип местообитания, форма клеток, подвижность, окраска по Граму (все – категориальные, качественные).

Диаграммы рассеяния наблюдаемых и ожидаемых значений патогенности при многомерном шкалировании (*multi-dimensional scaling*, [31]) дистанций между образцами в матрице близости (*proximity matrix*) представлены на рисунке 1.

Модель №2 получена при классификации выборки по патогенности на основе пяти предиктов (за исключением окраски по Граму). Точность классификации составила 76,92%, чувствительность – 69,23%, что превышает аналогичные показатели модели №1. Следовательно, окраска по Граму не является существен-

но значимым предиктом классификации.

Диаграммы рассеяния наблюдаемых и ожидаемых значений патогенности представлены на рисунке 2.

Модель №3 получена при классификации выборки по патогенности на основе четырех предиктов (за исключением окраски по Граму и подвижности). Точность классификации составила 78,21 %, чувствительность – 69,23 %. Следовательно, подвижность не является существенно значимым предиктом при выявлении патогенного микроорганизма.

Диаграммы рассеяния наблюдаемых и ожидаемых значений патогенности представлены на рисунке 3.

Модель №4 получена при классификации выборки по патогенности на основе трех предиктов (за исключением окраски по Граму, подвижности и формы клеток). Точность классификации составила 66,67%, чувствительность – 87,18%. Следовательно, форма клеток является существенно значимым предиктом при классификации микроба и не существенна при определении патогенности последнего, т.к. снижает чувствительность модели.

Диаграммы рассеяния наблюдаемых и ожидаемых значений патогенности представлены на рисунке 4.

Модель №5 получена при классификации выборки по патогенности на основе двух предиктов (оптимальных для роста значений pH и температуры). Точность классификации составила 69,23%, чувствительность – 53,85%. Следовательно, тип местообитания является существенно значимым предиктом при классификации микроба и определении патогенности последнего.

Диаграммы рассеяния наблюдаемых и ожидаемых значений патогенности представлены на рисунке 5.

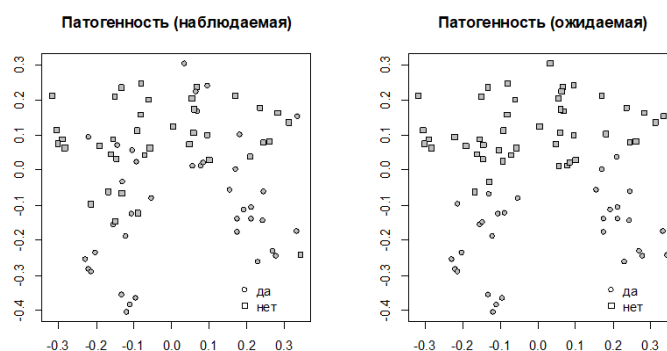


Рисунок 1 – Диаграммы рассеяния наблюдаемых и ожидаемых значений патогенности образцов микроорганизмов (модель №1 – 6 предиктов)

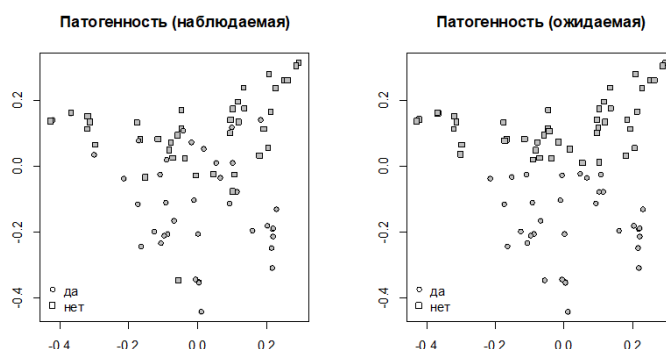


Рисунок 2 – Диаграммы рассеяния наблюдаемых и ожидаемых значений патогенности образцов микроорганизмов (модель №2 – 5 предиктов за исключением окраски по Граму)

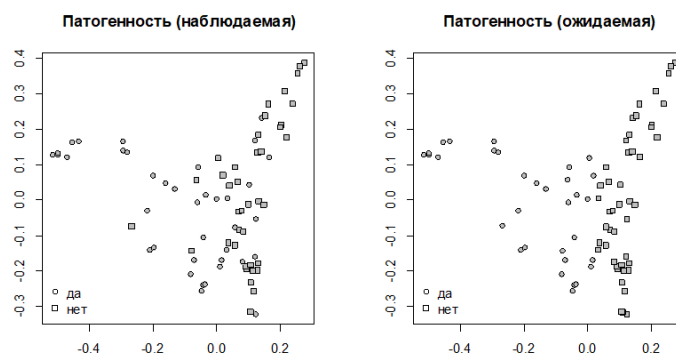


Рисунок 3 – Диаграммы рассеяния наблюдаемых и ожидаемых значений патогенности образцов микроорганизмов (модель №3 – 4 предиктов за исключением окраски по Граму и подвижности клеток)

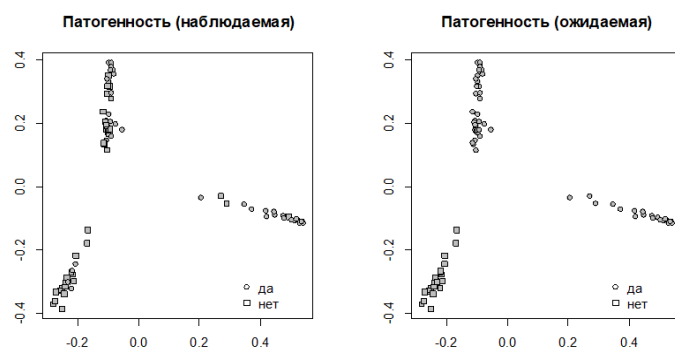


Рисунок 4 – Диаграммы рассеяния наблюдаемых и ожидаемых значений патогенности образцов микроорганизмов (модель №4 – 3 предикта за исключением окраски по Граму, подвижности и формы клеток)

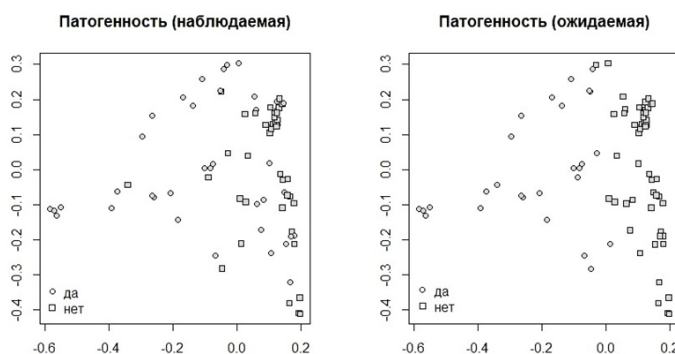


Рисунок 5 – Диаграммы рассеяния наблюдаемых и ожидаемых значений патогенности образцов микроорганизмов (модель №5 – 2 предикта – оптимальные для роста значений pH и температуры)

Совокупность индексов, рассчитанных с помощью матриц ошибок, полученных при сопоставлении наблюдаемых (исходно данных) и ожидаемых (полученных в результате работы SVM-алгоритма) значений патогенности (табл. 2), позволяет провести дополнительную оценку качества той или иной модели в классификации микробов на опасные (патогенные) и безопасные (непатогенные) для здоровья человека. В частности, модель № 3 (четыре предикта) наиболее точна в определении непатогенных (безопасных) микробов.

Таблица 2 – Совокупность индексов, рассчитанных с помощью матриц ошибок, полученных при сопоставлении наблюдаемых и ожидаемых значений патогенности

Index	Модель №1	Модель №2	Модель №3	Модель №4	Модель №5
Prevalence	50	50	50	50	50
Accuracy	74.36	76.92	78.21	66.67	69.23
PPV	80.65	81.82	84.38	61.82	77.78
FDR	19.35	18.18	15.62	38.18	22.22
FOR	29.79	26.67	26.09	21.74	35.29
NPV	70.21	73.33	73.91	78.26	64.71
F1score	0.7114	0.75	0.76	0.72	0.64
Sensitivity	64.1	69.23	69.23	87.18	53.85
FNR	35.9	30.77	30.77	12.82	46.15
Fall-out	15.38	15.38	12.82	53.85	15.38
Specificity	84.62	84.62	87.18	46.15	84.62
LR+	4.17	4.50	5.40	1.62	3.50
LR-	0.42	0.36	0.35	0.28	0.55
DOR	9.82	12.38	15.30	5.83	6.42



**Закключение.** Таким образом, в результате проведенных исследований получены модели, позволяющие прогнозировать опасность того или иного микроорганизма для здоровья работника, предсказывать патогенность микроба как потенциально вредного биологического фактора производственной среды. Наиболее точной оказалась классификационная модель №3, построенная на четырех предиктах – форма клеток, тип местообитания, оптимальные для жизни микроба pH и температура. В то время как наиболее чувствительной (при этом, наименее точной), а, следовательно, наиболее эффективной в определении патогенности (опасности) микроба стала модель №4, включающая признаки типа местообитания и оптимумов pH и температуры, достаточные для выполнения поставленной задачи исследования.

#### СПИСОК ЛИТЕРАТУРЫ:

1. Л.Б. Борисов. Медицинская микробиология, вирусология и иммунология. — МИА, 2005. — С. 191. — ISBN 5-89481-278-X.
2. Ширококов В. П. Медицинская микробиология, вирусология и иммунология / Перевод: Андрианова Т. В. и др. // Винница: Нова Книга. — 2015. — 856 с. ISBN 978-966-382-200-6. (С. 296).
3. Донецкая Э. Г.-А. Клиническая микробиология: руководство // М.: ГЭОТАР-Медиа. — 2011. — 480 с. ISBN 978-5-9704-1830-7. (С. 21-27).
4. Г. Г. Онищенко. Санитарно-эпидемиологические правила СП 1.3.2322-08. — Бюллетень нормативных актов федеральных органов исполнительной власти от 12.05.2008 г. № 19, 2008. — С. 20-34.
5. Санитарно-эпидемиологические правила СП 1.3.3118-13 «Безопасность работы с микроорганизмами I — II групп патогенности (опасности)» / Приложение 3 // М.: Федеральный центр гигиены и эпидемиологии Роспотребнадзора. — 2014. — 195 с. ISBN 978-5-7508-1342-1. (С. 111-125).
6. Безопасность жизнедеятельности: Учеб. для вузов / С.В.Белов, А.В.Ильницкая, А.Ф.Козьяков и др.; Под общ. ред. С.В.Белова. 2-е изд., испр. и доп. М.: Высш. шк., 1999. — 448 с.
7. BR Levin, V Perrot, Nina Walker. Compensatory Mutations, Antibiotic Resistance and the Population Genetics of Adaptive Evolution in Bacteria. Genetics March 1, 2000 vol. 154 no. 3 985—997.
8. Cassir, N; Rolain, JM; Brouqui, P. A new strategy to fight antimicrobial resistance: the revival of old antibiotics (англ.) // Frontiers in microbiology journal. — 2014. — Vol. 5. — P. 551. — doi:10.3389/fmicb.2014.00551. — PMID 25368610.
9. Levy, Stuart B. Factors impacting on the problem of antibiotic resistance (англ.) // Journal of Antimicrobial Chemotherapy (англ.)рус.: journal. — 2002. — 1 January (vol. 49, no. 1). — P. 25—30. — ISSN 0305-7453. — doi:10.1093/jac/49.1.25. — PMID 11751763.
10. Martinez, J. L., & Olivares, J. (2012). Environmental Pollution By Antibiotic Resistance Genes. In P. L. Keen, & M. H. Montforts, Antimicrobial Resistance in the Environment (pp. 151—171). Hoboken, N.J.: John Wiley & Sons.
11. Teale, F.H. (1933), Factors influencing the pathogenicity of bacteria. J. Pathol., 37: 185-232. <https://doi.org/10.1002/path.1700370204>.
12. Mitchell, Tom (1997). Machine Learning. New York: McGraw Hill. ISBN 0-07-042807-7. OCLC 36417892.
13. Kim, Seung & Goodfellow, Michael. (2015). Bergey's Manual of Systematics of Archaea and Bacteria. 10.1002/9781118960608.gbm00186.
14. Breiman, L. (2001), Random Forests, Machine Learning 45(1), 5-32.
15. Breiman, L (2002), "Manual On Setting Up, Using, And Understanding Random Forests V3.1"
16. Breiman L. et al. (1984). Classification and Regression Trees. CRC Press, Boca Raton.
17. Chih-Chung Chang and Chih-Jen Lin, LIBSVM: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2:27:1–27:27, 2011. Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.
18. Stehman, Stephen V. (1997). "Selecting and interpreting measures of thematic classification accuracy". Remote Sensing of Environment. 62 (1): 77–89. Bibcode:1997RSEnv..62...77S. doi:10.1016/S0034-4257(97)00083-7.
19. Yerushalmy J (1947). "Statistical problems in assessing methods of medical diagnosis with special reference to x-ray techniques". Public Health Reports. 62 (2): 1432–39. doi:10.2307/4586294. JSTOR 4586294. PMID 20340527.
20. Fawcett, Tom (2006). "An Introduction to ROC Analysis" (PDF). Pattern Recognition Letters. 27 (8): 861–874. doi:10.1016/j.patrec.2005.10.010.
21. Pirayonesi S. Madeh; El-Diraby Tamer E. (2020-03-01). "Data Analytics in Asset Management: Cost-Effective Prediction of the Pavement Condition Index". Journal of Infrastructure Systems. 26 (1): 04019036. doi:10.1061/(ASCE)IS.1943-555X.0000512.
22. Powers, David M. W. (2011). "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation". Journal of Machine Learning Technologies. 2 (1): 37–63.
23. Ting, Kai Ming (2011). Sammut, Claude; Webb, Geoffrey I. (eds.). Encyclopedia of machine learning. Springer. doi:10.1007/978-0-387-30164-8. ISBN 978-0-387-30164-8.
24. Brooks, Harold; Brown, Barb; Ebert, Beth; Ferro, Chris; Jolliffe, Ian; Koh, Tieh-Yong; Roebber, Paul; Stephenson, David (2015-01-26). "WWRP/WGNE Joint Working Group on Forecast Verification Research". Collaboration for Australian Weather and Climate Research. World Meteorological Organisation. Retrieved 2019-07-17.
25. Chicco D, Jurman G (January 2020). "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation". BMC Genomics. 21 (1): 6-1–6-13. doi:10.1186/s12864-019-6413-7. PMC 6941312. PMID 31898477.
26. Chicco D, Toetsch N, Jurman G (February 2021). "The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation". BioData Mining. 14 (13): 1-22. doi:10.1186/s13040-021-00244-z. PMC 7863449. PMID 33541410.
27. Tharwat A. (August 2018). "Classification assessment methods". Applied Computing and Informatics. doi:10.1016/j.aci.2018.08.003
28. Microsoft Corporation, 2018. Microsoft Excel, Available at: <https://office.microsoft.com/excel>.
29. R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.
30. RStudio Team (2019). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA URL <http://www.rstudio.com/>.
31. Gower, J. C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika, 53, 325-328. 10.2307/2333639.

Статья поступила в редакцию 14.01.2022

Статья принята к публикации 10.03.2022