DOI 10.31891/2307-5732-2024-331-41
УДК 004.8

СТАЦЕНКО ВОЛОДИМИР
Київський національний університет технологій та дизайну
https://orcid.org/0000-0002-3932-792X
e-mail: statsenko.v@knutd.edu.ua
ПИЛИПЕНКО ВЛАДИСЛАВ
Київський національний університет технологій та дизайну
https://orcid.org/0000-0002-2761-4817
e-mail: pylypenko.vi@knutd.edu.ua

ОЦІНЮВАННЯ ЕФЕКТИВНОСТІ МОДЕЛІ ПРОГНОЗУВАННЯ УСПІШНОСТІ МЕТОДАМИ МАШИННОГО НАВЧАННЯ

*В роботі проведено оцінювання створеної моделі для прогнозування успішності користувачів навчальної платформи Moodle із метою визначення ефективності та доцільності її використання. Визначення ефективності та якості моделі виконано по наступним показникам: чутливість (Sensitivity), специфічність (Specificity) та збалансована точність (Balanced Accuracy). Також побудовано ROC-криву для оцінки здатності класифікатора правильно розпізнавати позитивні класи і відхиляти негативні класи при зміні порогового значення. Розраховано AUC (*Area Under Curve*). Проаналізовано методи оцінювання ризиків успішності та вимоги до створення моделей на базі методів машинного навчання. На основі користувацьких даних із бази даних навчальної платформи Moodle сформовано показники, що впливають на успішність студентів. Побудовано логістично-регресійну модель для прогнозування успішності, яку випробувано на практиці. Створена модель дозволяє виконати прогнозування успішності студентів із точністю 84%. Загальна ефективність моделі складає 89%. Використання логістично-регресійної моделі для класифікації успішності користувачів платформи Moodle дозволить створити ефективну модель для прогнозування успішності студентів.*

Використання створеної моделі для прогнозування дозволить оперативно аналізувати успішність користувачів і формувати, за необхідності, відповідні рейтинги. Прогноз, отриманий за допомогою моделі буде корисний як для викладачів, так і закладів освіти, оскільки дозволить планувати зміни в навчальних програмах та матеріалах, а також освітньому процесі в цілому.

Ключові слова: логістична регресія; Machine Learning; Python; Scikit-learn; Moodle.

STATSENKO VOLODYMYR, PYLYPENKO VLADYSLAV
Kyiv National University of Technologies and Design

ASSESSMENT OF THE EFFICIENCY OF THE SUCCESS PREDICTION MODEL USING MACHINE LEARNING METHODS

The work evaluated the created model for predicting the success of users of the Moodle educational platform in order to determine the effectiveness and expediency of its use. The determination of the efficiency and quality of the model was performed according to the following indicators: sensitivity, specificity and balanced accuracy. Also, a ROC curve was constructed to reflect the classifier's ability to correctly recognize positive classes and reject negative classes when the threshold value changes, and the AUC (Area Under Curve) was determined. Methods of assessing success risks and requirements for creating models based on machine learning methods are analyzed. On the basis of user data from the database of the Moodle educational platform, indicators affecting the success of students were formed. A logistic regression model was built for predicting success, which was then tested in practice. The created model makes it possible to predict the success of students with an accuracy of 84%. The overall efficiency of the model is 89%. It was established that the Scikit-learn library provides an opportunity to create an effective model for solving classification problems in machine learning. The use of a logistic regression model to classify the success of users of the Moodle platform will allow creating a model that allows you to predict the success of students based on the collected data. Using machine learning methods and Python libraries, the quality and efficiency of the model was determined. A solution to the problem of predicting students' success is proposed by creating a model based on machine learning methods and the Moodle platform. A Python program was developed to analyze the data of users of the Moodle platform.

The presented information shows that choosing the Scikit-Learn library will allow creating an effective model for processing data and predicting results. The use of the created model for forecasting will allow to quickly analyze the success of users and form, if necessary, appropriate ratings. The forecast obtained with the help of the model will be useful both for teachers and educational institutions. Because it will allow planning changes in educational programs and materials, as well as the educational process as a whole.

Keywords: logistic regression; Machine Learning; Python; Scikit-learn; Moodle.

Постановка проблеми

Освіта сьогодні є одним із основних факторів, що впливають на майбутні можливості та кар'єрний розвиток студентів. Високий рівень успішності в навчанні відкриває двері до перспективних посад і можливостей для особистого зростання. В подальшому це створює позитивний вплив на економіку країни. Адже достатня кількість кваліфікованих фахівців збільшує конкурентоспроможність, інноваційність і прискорює розвиток економіки. Можливість прогнозування успішності, як освітнього ризику [1], має велике значення, оскільки швидке реагування може збільшити шанси студента на успішну здачу сесії. Аналіз факторів, що впливають на успішність, грає важливу роль у прогнозуванні цього результату.

Оскільки успішність є одним із ключових освітніх ризиків, заходи із прогнозування дозволяють виявляти проблемні ситуації на ранній стадії, чим забезпечують можливість запобігти негативним наслідкам. У світовій і вітчизняній практиці використовується більше 30 методів загального оцінювання ризику, характеристика яких наведена в ДСТУ IEC/ISO 31010:2013 [2]. Оцінювання ризику, в даному випадку успішності, дає змогу тим, хто приймає рішення, а також відповідальним сторонам краще розуміти, які

фактори можуть на це впливати та яка результативність засобів контролю. Найефективнішим інструментом для визначення ймовірності, як правило, є математичні та статистичні моделі. Тому у даній роботі основна увага буде зосереджена на них. Це допоможе визначити, які саме студенти належать до групи ризику. Для формування таблиці з факторами, що впливають на успішність, використані дані користувачів із навчальної платформи Moodle [3]. Дані являють собою записи з бази даних Moodle, які експортовані в csv формат.

Аналіз останніх джерел

Аналіз літературних джерел показав, що прогнозування успішності є важливим напрямом, який постійно удосконалюються за рахунок використання комп'ютерних засобів та програмних рішень.

Згідно з [4] прогнозування можна виконувати на базі методів машинного навчання, використовуючи метод випадкового лісу "Random Forest" для задач класифікації. Він добре підходить для прогнозування категорії або класу нового зразка на основі його характеристик. Розрахунки точності моделі показують, що вибір бібліотеки Scikit-Learn дозволяє створити ефективну модель обробки даних і прогнозування результатів. Отримана загальна точність розробленої моделі, становила 83%.

Згідно з [5] для оцінки ефективності та якості моделі було запропоновано наступні ключові показники: чутливість (Sensitivity), специфічність (Specificity), збалансована точність (Balanced Accuracy). А також побудовано ROC-криву для відображення здатності класифікатора правильно розпізнавати позитивні класи і відхиляти негативні класи при зміні порогового значення та визначено AUC (Area Under Curve).

Метою роботи є проведення якісної оцінки та ефективності створеної моделі прогнозування успішності студентів на платформі управління навчанням Moodle.

Виклад основного матеріалу

У роботі були проаналізовані фактори які впливають на успішність та запропоновано загальний підхід до створення моделі з прогнозування успішності студентів. З бази даних Moodle були обрані параметри, що тим чи іншим чином пов'язані з успішністю. Їх значення були експортовані та збережені у csv форматі (файл moodle_sdata.csv). Всього експортовано записи 2000 студентів. Після цього дані були оброблені наступним чином. Для кожного студента було взяте середнє значення оцінки по: дисципліні, курсу, тесту та модулю. А також додано загальне значення відсотку відвідуваності. Перелік цих параметрів наведено у табл. 1.

Таблиця 1

Вихідні параметри для створення моделі прогнозування успішності користувачів платформи Moodle

Назва	User ID	First Name	Last Name	Visit Percent	Test Marks	Module Marks	Course Marks	Is Success
Тип	Integer	String	String	Integer	Integer	Integer	Integer	Integer
Знач.	3850911	#####	#####	55	10	8	75	1
Знач.	3850912	#####	#####	75	7	7	87	1
Знач.	3850913	#####	#####	25	3	2	55	0

де UserID – унікальний ідентифікатор користувача на платформі;

FirstName, LastName – ім'я та прізвище користувача;

VisitPercent – загальний відсоток відвідування (від 0 до 100);

TestMarks – середній бал за модуль (від 0 до 10);

ModuleMarks – середній бал за модуль (від 0 до 10);

CourseMarks – середній бал за предмет/курс (від 0 до 100);

IsSuccess – характеристика успішності (1- так, 0- ні).

Створення моделі успішності студентів здійснювалось за допомогою методу машинного навчання – логістичної регресії, та задачі – класифікації [6]. Загальний підхід до створення моделі успішності студентів представлено у табл. 2.

Таблиця 2

Опис алгоритму для створення моделі прогнозування успішності

Назва етапу	Опис
Збір даних	Збирання даних, що включають: загальний відсоток відвідуваності, середній бал з тестів, середній бал з предметів курсу та інші.
Передобробка даних	Виконання обробку пропущених значень, нормалізацію або стандартизацію ознак, а також провести видалення шуму і збалансування класів.
Вибір задачі та методу	Використана здача: класифікації, використаний метод: логістичної регресії.
Розбиття даних	Розділення даних на тренувальний та тестовий набори.
Навчання моделі	Застосування обраної моделі до тренувальних даних та навчання її передбачати успішність студентів на основі вхідних ознак.
Оцінка моделі	Розрахунок наступних метрик: точності (accuracy), збалансованої точності (balanced accuracy), специфічності (specificity), чутливості (sensitivity), побудова ROC-кривої і визначення AUC.
Застосування моделі	Після оцінки та валідації моделі її можна застосовувати для вирішення задач з прогнозування успішності.

Перед виконанням навчання моделі вихідні дані були розділені на тренувальну та тестову вибірки для того, щоб перевірити, наскільки добре модель, навчена на тренувальній вибірці, може передбачати класи нових даних. Обсяг даних взятих для обробки складав 2000 вибірок користувачів із бази даних, які були розподілені у відношенні 1400/600. З яких тренувальна вибірка містила – 1400, а тестова – 600. Ділення даних на тренувальну та тестову вибірки допомагає уникнути перенавчання (overfitting) моделі [7].

Оцінка та перевірка якості моделі здійснювалась на основі тестової вибірки. Основними критеріями ефективності моделі були обрані показники: точність, збалансована точність, чутливість, специфічність, AUC та ROC-крива. Ці показники розраховуються на основі так званої матриці помилок (confusion matrix) [8]. Кращим результатом класифікації є такий, для якого кількість правильно класифікованих випадків максимальна, а кількість неправильно класифікованих випадків мінімізована. Дане відношення значень представлено у табл. 3.

Таблиця 3

Загальний вигляд матриці помилок

		Актуальні значення	
		True	**False**
Предиковані значення	True	True Positives (TP) (істинно позитивне)	False Negatives (FN) (хибно негативне)
	False	False Positives (FP) (хибно позитивне)	True Negatives (TN) (істинно негативне)

Матриця помилок моделі дозволяє нам порахувати, для скількох студентів прогнозування було виконано правильно. Зображення отриманої матриці помилок, для створеної моделі, представлене на рис. 1.

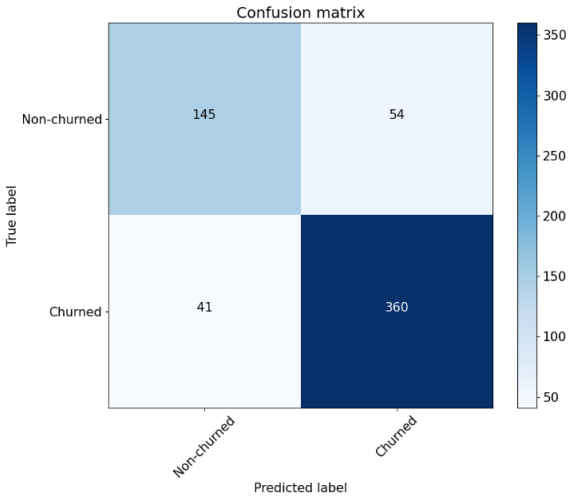

Рис. 1. Матриця помилок моделі із прогнозування успішності

Виходячи з отриманої матриці проводимо розрахунок загальної точності класифікації моделі. Точність (Accuracy) показує, який відсоток прикладів був правильно класифікований [9]. Вираз для визначення точності можна записати у вигляді наступної формули:

$$Accuracy = (TP+TN)/(TP+TN+FP+FN), \quad (1)$$

де TP (true positives) – кількість правильно передбачених позитивних класів;

TN (true negatives) – кількість правильно передбачених негативних класів;

FP (false positives) – кількість неправильно передбачених позитивних класів;

FN (false negatives) – кількість неправильно передбачених негативних класів.

Після проведення підрахунків показник Accuracy становить 0,8416. Це означає, що модель правильно класифікує 84% тестових даних. Це є досить високим показником, але для більш повної оцінки ефективності моделі необхідно визначити інші метрики та фактори.

Наступним важливим фактором у оцінці моделі є чутливість (Sensitivity) [10]. Чутливість дозволяє визначити наскільки добре модель виявляє позитивні випадки. Вираз для визначення чутливості можна записати у вигляді наступної формули:

$$Sensitivity = TP/(TP + FN) \quad (2)$$

де TP (true positives) – кількість правильно передбачених позитивних класів;

FN (false negatives) – кількість неправильно передбачених негативних класів.

Після проведення розрахунків показник Sensitivity становить 0,8977. Такий результат означає, що модель може прогнозувати майже 90% можливого фактору «успішності» правильно.

Наступним фактором оцінки є специфічність (Specificity) [10]. Специфічність дозволяє визначити наскільки добре модель виявляє негативні випадки. Вираз для визначення специфічності можна записати у вигляді наступної формули:

$$Specificity = TN/(TN+FP) \quad (3)$$

де TN (true negatives) – кількість правильно передбачених негативних класів;

FP (false positives) – кількість неправильно передбачених позитивних класів.

Після проведення підрахунків показник Specificity становить 0,7286. Результат означає, що модель може прогнозувати лише 73% можливого фактору «неуспішності» правильно.

Оскільки звичайна точність (Accuracy) може бути використана тільки для вимірювання загальної точності класифікатора, вона може бути непоказовою, коли дані не збалансовані і кількість прикладів одного класу перевищує кількість прикладів іншого класу. Тому для отримання оцінки загальної ефективності моделі бінарного класифікатора, враховуючи баланс між класами даних, було розраховано збалансовану точність (Balanced Accuracy) [11]. Вона враховує як чутливість (True Positive Rate), так і специфічність (True Negative Rate) класифікатора. Вираз для визначення збалансованої точності можна записати у вигляді наступної формули:

$$Balanced\ Accuracy = (Sensitivity\ +\ Specificity)\ /\ 2 \quad (4)$$

Після проведення розрахунків показник Balanced Accuracy склав 0,8131. Результат показує загальну ефективність моделі як 81% у виявленні як позитивних, так і негативних випадків, з урахуванням дисбалансу класів у наборі даних. Результат вказує на середню точність (Accuracy) для кожного класу, з урахуванням його розподілу. І враховує, наскільки добре модель передбачає як позитивні, так і негативні випадки, та надає рівномірну оцінку ефективності для обох класів, незалежно від їх кількості. У підсумку загальна ефективність складає 81%.

Щоб наглядно оцінити здатність моделі до правильної класифікації, враховуючи різні значення порогового значення було побудовано ROC-криву (Receiver Operating Characteristic) [12]. ROC-крива відображає здатність класифікатора правильно розпізнавати позитивні класи та відхиляти негативні класи при зміні порогового значення. Вона дозволяє враховувати компроміс між чутливістю та специфічністю класифікатора та зробити розгляд результатів моделі класифікації більш об'єктивним. Чим більше вигнута вгору і вліво ділянка під ROC-кривою, тим ефективність моделі краща. В бібліотеці scikit-learn, за замовчуванням, значення лінії розподілу cut-off для бінарних класифікаторів встановлено як 0,5. Це означає, що якщо прогнозна ймовірність класу 1 (позитивного класу) більше або дорівнює 0,5, то об'єкт буде класифіковано як позитивний, в іншому випадку – як негативний.

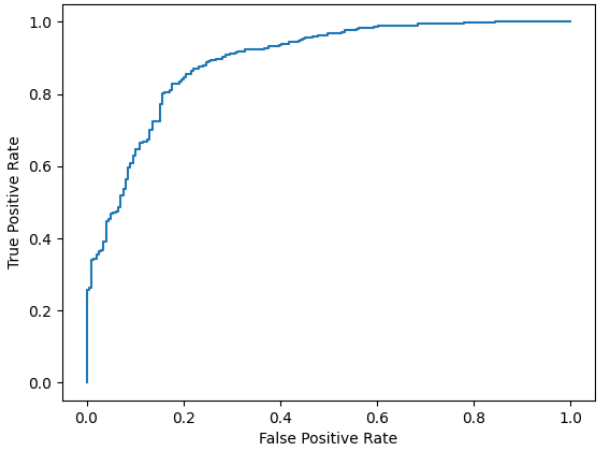

Рис. 2. Графік ROC-кривої

Для оцінки загальної ефективності моделі незалежно від вибору порогового значення було використано параметр AUC (Area Under Curve) [13]. Він обчислюється як площа під ROC-кривою, і може приймати значення в діапазоні від 0 до 1. Чим більше значення AUC, тим краща якість моделі класифікації. Якщо AUC дорівнює 0,5, це вказує на випадковий класифікатор, а значення менше 0,5 вказують на зворотну кореляцію між прогнозованими та дійсними мітками класу. Завдяки обчисленню площі під кривою, можна зрозуміти міру її «хорошості», чим далі крива від діагональної лінії, тим вона краща. Вираз для визначення AUC можна записати у вигляді наступної формули:

$$AUC = \sum_{n=1}^{\infty}(\mathrm{TPR(i+1)} - \mathrm{TPR(i)}) * (\mathrm{FPR(i)} + \mathrm{FPR(i+1)})/2 \quad (5)$$

де TPR(i) - чутливість (True Positive Rate) для і-го порогового значення;

FPR(i) - специфічність (1 - False Positive Rate) для і-го порогового значення.

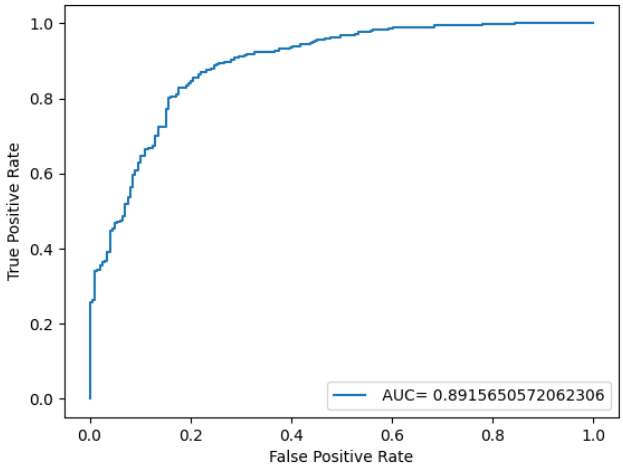

Рис. 3. Графік ROC-кривої з обчисленою площею AUC

Після проведення розрахунків показник AUC склав 0,8915. Отримане значення 89% свідчить про якість моделі класифікації, та гарну дискримінаційну силу моделі. Отримана модель із загальною точністю 84% і ефективністю 89% є досить доброю початковою точкою, але в процесі подальшого дослідження може знадобитися додаткове вдосконалення для підвищення даних показників. Також важливим є збільшення кількості даних для отримання більш точного результату.

Результати показали, що отримана загальна точність 84% є вищою ніж у методу випадкового лісу (Random Forest), яка склала 83% в проведеному раніше дослідженні [4]. Також було проведено визначення та порівняння збалансованої точності (Balanced Accuracy), яка склала 81% в поточному дослідженні та 77% в проведеному раніше. Використання логістичної регресії для бінарної класифікації краще працює для простих моделей з невеликою кількістю функцій і коли залежність між ознаками і вихідними класами лінійна або логістична. Проте випадковий ліс є ансамблевим методом, який зазвичай працює краще в тих випадках, коли взаємозв'язки між ознаками та вихідними класами більш складні, нелінійні або коли є багато ознак. Він може автоматично враховувати важливість ознак і робити кращі передбачення, ніж лінійні моделі на складних даних.

Висновки

1.У роботі створено модель, що дозволяє на основі даних інформації про дії користувачів платформи Moodle виконати прогнозування їх успішності.

2.Для побудови моделі використано логістичну регресію, як статистичний алгоритм машинного навчання. Він має відносно високу точність та низьку тривалість процесу навчання. Завдяки використанню даного рішення отримано ефективну модель із прогнозування, про що свідчить отримана матриця помилок.

3.Розраховано загальну точність розробленої моделі, яка становить 84% та збалансовану точність, яка становить 81%.

4.Визначено оцінку загальної ефективності розробленої моделі, яка становить 89%.

5.Проведено порівняння загальної та збалансованої точності між методами випадкового лісу та логістичної регресії.

6.Підвищення точності моделі можливе за рахунок розширення вихідних даних, що потребує створення відповідних додатків (плагінів) для платформи Moodle, та є перспективним напрямом розвитку таких систем.

Література

1. ISO 31000:2009(en) Risk management - Principles and guidelines. 2009. https://www.iso.org/obp/ui/#iso:std:iso:31000:ed-1:v1:en.

2. Керування ризиком. Методи загального оцінювання ризиків. Мінекономрозвитку України. 2015. https://khoda.gov.ua/image/catalog/files/dstu%2031010.pdf.

3. Overview of the Moodle educational platform. 2022. https://docs.moodle.org/401/en/About_Moodle.

4. Пилипенко В. І., Стаценко В. В. Прогнозування активності користувачів платформи Moodle на базі методів машинного навчання. Вісник Хмельницького національного університету. 2023. № 4. С. 257–261.

5. Стаценко В.В., Пилипенко В.І. Оцінка ефективності моделі прогнозування активності користувачів Moodle методами машинного навчання. VII Міжнародна науково-практична конференція «Мехатронні системи: інновації та інжиніринг» – «MSIE – 2023», 23 листопада 2023. КНУТД. С. 28-29.

6. TAN, Haoyuan. Machine learning algorithm for classification. In: Journal of Physics: Conference Series. IOP Publishing, 2021. p. 012016.

7. Model underfitting vs. Overfitting. 2024. https://scikit-learn.org/stable/auto_examples/model_selection/plot_underfitting_overfitting.html.

8. Compute confusion matrix in Scikit-learn metrics. 2024. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.confusion_matrix.html.

9. LEE, Julian. Analysis of precision and accuracy in a simple model of machine learning. Journal of the Korean Physical Society, 2017, 71: 866-870.

10. Determination of specificity and sensitivity. 2024. https://en.wikipedia.org/wiki/Sensitivity_and_specificity.

11. Balanced accuracy in Scikit-learn metrics. 2024. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.balanced_accuracy_score.html.

12. Receiver Operating Characteristic (ROC-curve). 2022. https://scikit-learn.org/1.0/auto_examples/model_selection/plot_roc.html.

13. How to Calculate AUC. 2021. https://www.statology.org/auc-in-python/.

References

1. ISO 31000:2009(en) Risk management - Principles and guidelines. 2009. https://www.iso.org/obp/ui/#iso:std:iso:31000:ed-1:v1:en.

2. Keruvannia ryzykom. Metody zahalnoho otsiniuvannia ryzykiv. Minekonomrozvytku Ukrainy. 2015. https://khoda.gov.ua/image/catalog/files/dstu%2031010.pdf.

3. Overview of the Moodle educational platform. 2022. https://docs.moodle.org/401/en/About_Moodle.

4. Pylypenko V. I., Statsenko V. V. Prohnozuvannia aktyvnosti korystuvachiv platformy Moodle na bazi metodiv mashynnoho navchannia. Visnyk Khmelnytskoho natsionalnoho universytetu. 2023. № 4. S. 257–261.

5. Statsenko V.V., Pylypenko V.I. Otsinka efektyvnosti modeli prohnozuvannia aktyvnosti korystuvachiv Moodle metodamy mashynnoho navchannia. VII Mizhnarodna naukovo-praktychna konferentsiia «Mekhatronni systemy: innovatsii ta inzhynirynh» – «MSIE – 2023», 23 lystopada 2023. KNUTD. S. 28-29.

6. TAN, Haoyuan. Machine learning algorithm for classification. In: Journal of Physics: Conference Series. IOP Publishing, 2021. p. 012016.

7. Model underfitting vs. Overfitting. 2024. https://scikit-learn.org/stable/auto_examples/model_selection/plot_underfitting_overfitting.html.

8. Compute confusion matrix in Scikit-learn metrics. 2024. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.confusion_matrix.html.

9. LEE, Julian. Analysis of precision and accuracy in a simple model of machine learning. Journal of the Korean Physical Society, 2017, 71: 866-870.

10. Determination of specificity and sensitivity. 2024. https://en.wikipedia.org/wiki/Sensitivity_and_specificity.

11. Balanced accuracy in Scikit-learn metrics. 2024. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.balanced_accuracy_score.html.

12. Receiver Operating Characteristic (ROC-curve). 2022. https://scikit-learn.org/1.0/auto_examples/model_selection/plot_roc.html.

13. How to Calculate AUC. 2021. https://www.statology.org/auc-in-python/.