УДК 004. 004.9. 004.89

ПРОГНОЗУВАННЯ УСПІШНОСТІ СТУДЕНТІВ НА ОСНОВІ МЕТОДІВ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАНИХ НАВЧАЛЬНИХ ПРОГРАМ

В. Р. Вергун

Національний університет «Львівська політехніка», вул. С. Бандери, 12, Львів, 79013, Україна

Розглянуто задачу прогнозування успішності студентів навчальних програм. Зібрано та сформовано вибірку даних на основі аналізу резюме 101 кандидата на навчання за програмою «Інженерія програмного забезпечення». Досліджено низку методів машинного навчання для розв'язання поставленої задачі. Оцінка ефективності роботи обраних методів машинного навчання під час розв'язання поставленої задачі відбувалася на основі різноманітних показників. Встановлено, що найвищу точність роботи забезпечують логістична регресія та алгоритм випадкового лісу. Виявлено, що ефективність застосування цих методів не є високою та має велику залежність від вибірки даних, конкретного контексту та поставленої задачі.

Ключові слова: *інтелектуальний аналіз даних, прогнозування, random forest, задача регресії, успішність, навчальні програми.*

Постановка проблеми. Ефективний розв'язок задачі прогнозування успішності студентів навчальних програм відкриває багато можливостей у створенні більш якісних навчальних програм, процесів та результатів. Дослідження, що аналізують можливості прогнозування, здебільшого зосереджені на виявленні факторів, що мають вплив на продуктивність студентів, та пошуку нових методів, що дають змогу точніше вирішувати цю задачу. Здебільшого використовується процес інтелектуального аналізу даних (ІАД), що використовує дані, які згенеровано різноманітними системами управління навчальним процесом для пошуку шаблонів та прихованих закономірностей. Найбільш застосовуваними підходами інтелектуального аналізу даних у навчанні є кластеризація, класифікація та асоціація. Ці методики широко використовуються для створення рекомендацій та вдосконалення навчальних програм та процесів. Окрім цього, їх застосовують для визначення обґрунтованого розуміння критеріїв, для подальшого прогнозування та створення систем прийняття рішень. Навчальний прогрес будь-якого суб'єкта навчання може бути розумно та якісно оцінений, з огляду на взаємодію різних факторів та виявлення асоціації.

Аналіз останніх досліджень та публікацій. Незважаючи на велику кількість даних, згенерованих різноманітними системами управління навчанням, у більшості досліджень розглядаються однотипні фактори під час прогнозування успішності.

Одним із таких факторів є оцінка успішності обраного студента. Вибір такого фактора є очевидним, оскільки його легко порівнювати та він має чітке значення [1]. Поточна середня оцінка є істотним фактором, що впливає на прогноз успішності [2]. Також може бути розглянута комбінація оцінок із попередніх періодів навчання. Проте якщо розглядати нашу країну, особливо IT-спеціальності, ми не можемо бути впевнені, що середня оцінка є важливим фактором, який має вплив на подальшу кар'єру та можливість отримати роботу за спеціальністю [3]. Тут потрібно зауважити, що саме конкретний список компетенцій, які має опанувати студент, є важливішим фактором для подальшого працевлаштування. Окрім цього, визначення таких компетенцій є окремим великим завданням під час побудови будь-якої навчальної програми [5].

Водночас існує інший виклик, пов'язаний із прогнозуванням успішності та фінальних оцінок. Прогнозування сьогодні загалом здійснюється як одноразове завдання і не передбачає постійного відслідковування прогресу в навчанні та перегляду результатів прогнозування, базуючись на результатах останніх етапів навчання та перевірки знань. Отже, алгоритми прогнозування мають враховувати не тільки поточну середню оцінку відповідного студента, а й загальний темп та прогрес у засвоєнні матеріалу. Адже кожен студент має свій стиль, підхід та метод навчання. З огляду на це, можна досягти більш точного прогнозованого значення. Хоча при цьому загальна складність сильно зростає [4].

Мета статті — прогнозування успішності проходження вступного тесту потенційними кандидатами на навчання засобами машинного навчання.

Методи дослідження. Дані для цього дослідження зібрано на основі аналізу резюме 101 кандидата на навчання за програмою «Інженерія програмного забезпечення» на комерційній основі.

У табл. 1 подано структуру записів даних студента та опис усіх атрибутів. Серед заявників набору даних за гендерною ознакою — 73 % чоловіків і 28 % жінок. Ця статистика відображає поточний розподіл за статевою ознакою у сфері IT [6].

Потенційні кандидати класифікуються на 2 категорії залежно від результатів початкового тесту. У вибірці даних 23 % кандидатів правильно відповіли на більше ніж 66 % запитань тестування. Для аналізу та візуалізації використовувалося програмне забезпечення Weka. Для дослідження обрано такі алгоритми машинного навчання: Logistic Regression (LR), Support Vector Machines (SVM), Multilayer Perceptron (MP), Random Forest (RF). Їх вибір зумовлено точними результатами роботи в наявних працях [7].

Таблиця 1

Опис атрибутів вибірки дослідження

Атрибут	Опис
1	2
Test Score	Результат початкового тесту: Так: результат тесту> = 66 % Ні: результат тесту <66 %

Продовження табл. 1

1	2
Age	3 категорії віку студентів: <= 22: поточний студент університету > 22 і <29: випускник > = 29: заявник, який хоче змінити професію
Gender	Чоловіча чи жіноча
Degree	Y: має або здобуває освіту, пов'язану з галуззю програмного забезпечення N: має або здобуває освіту, не пов'язану з галуззю програмного забезпечення
Experience	Y: має будь-який досвід на будь-якій позиції N: жодного досвіду
Training	Y: пройшли додаткове навчання, пов'язане з інженерією програмного забезпечення. Онлайн або офлайн N: немає інформації про участь у будь-якому навчанні, пов'язаному з розробкою програмного забезпечення

Random Forest (RF). Це один із поширених методів машинного навчання, що полягає у використанні ансамблю дерев рішень [8, 9]. Застосовується для задач класифікації, регресії і кластеризації. Дерево рішень будується з використанням навчальної вибірки та поняття ентропії. На кожному вузлі обирається один атрибут з даних, який найефективніше ділить навчальну множину на підмножини, що найбільш розрізняються. Головним критерієм для вибору є нормований приріст інформації. Атрибут із найбільшим нормалізованим приростом інформації вибирається для прийняття рішення щодо поділу даних у вузлі дерева.

Logistic Regression (LR). Логістична регресія — статистичний регресійний метод моделювання залежності між векторною змінною та скаляром (вихідним значенням). Цей метод є узагальненням методу лінійної регресії і застосовується у випадку, коли залежна змінна може набувати лише скінченної множини значень. Параметри оцінюються на основі валідаційної вибірки зазвичай за допомогою методу максимальної правдоподібності. Основна відмінність та перевага такого підходу від інших моделей і алгоритмів є оцінка результату, яку можна було б розглядати як значення ймовірності для певного класу.

Support Vector Machines (SVM). Метод опорних векторів — категорія універсальних мереж прямого поширення, яку запропонував в 1963 р. Вапнік [10]. Метод SVM набув поширення в останнє десятиліття для задач класифікації, регресії та ідентифікації. Важливою властивістю SVM є те, що визначення параметрів моделі відповідає задачі випуклої оптимізації (convex optimization), і тому будь-який локальний розв'язок також є глобальним. Такий підхід до класифікації передбачає розгляд поняття поділу (margin), яке визначається як мінімальна відстань від гіперплощини до зразків з вибірки. Розділяюча гіперплощина будується в такий спосіб, щоб максимізувати значення поділу. Розміщення гіперплощини

визначається підмножиною точок даних, відомих як опорні вектори. Для того щоб зробити класифікатор більш потужним, у 1992 р. запропоновано спосіб створення нелінійного класифікатора, в основу якого покладено перехід від скалярних добутків до довільних ядер.

Multilayer Perceptron (MLP). Багатошаровими персептрони називають нейронні мережі прямого поширення. Вхідний сигнал у таких мережах поширюється в прямому напрямку, від шару до шару. Багатошаровий персептрон загалом складається з таких елементів: безлічі вхідних вузлів, які утворюють вхідний шар; одного або декількох прихованих шарів обчислювальних нейронів; одного вихідного шару нейронів. Багатошаровий персептрон — узагальнення одношарового персептрона Розенблатта.

Для оцінки ефективності роботи обраних методів машинного навчання використано такі показники: Kappa statistics; Mean absolute error (MAE); root-mean-square error (RMSE); TP rate; FP rate; precision, and MCC; F-measure; ROC Area; PRC Area. Процес перехресної перевірки повторювався 10 разів для кожного виконання алгоритму.

Основні результати. Результати роботи кожного досліджуваного методу для розв'язання поставленої задачі подано у табл. 2.

Відповідно до отриманих результатів, Multilayer Perceptron та Support Vector Machines мають подібні значення RMSE. Аналогічно подібні значення мають алгоритми LR та RF. SVM має найнижче значення Kappa Statistic. Якщо брати до уваги значення K statistic та RMSE, то можна вважати, що алгоритмами з найкращою швидкодією є Random Forest та LR.

Таблиця 2

Результати роботи досліджуваних методів

	Random Forest	Multilayer Perceptron	Logistic Regression	Support Vector Machines
K statistic	0.0342	0.0026	0.067	0
MAE	0.3153	0.3295	0.3445	0.2277
RMSE	0.4469	0.4735	0.4344	0.4772

Результати для метрик TP rate, FP rate, precision, recall, f-measure, MCC, ROC та PRC подані у табл. 3.

Таблиця 3

Результати показників ефективності TP rate, FP rate, precision, recall, f-measure, MCC, ROC та PRC

	Class	TP Rate	FP Rate	Precision	Recall	f-Measure	MCC	ROC	PRC
Multilayer Perceptron	0	0.872	0.870	0.773	0.872	0.819	0.003	0.589	0.280
	1	0.130	0.128	0.231	0.130	0.167	0.003	0.589	0.835
	Weighted Avg.	0.703	0.701	0.649	0.703	0.671	0.003	0.589	0.708

Продовження табл. 3

Random Forest	0	0.897	0.870	0.778	0.897	0.833	0.038	0.650	0.877
	1	0.130	0.103	0.273	0.130	0.176	0.038	0.650	0.316
	Weighted Avg.	0.723	0.695	0.663	0.723	0.684	0.038	0.650	0.749
Logistic Regression	0	0.962	0.913	0.781	0.962	0.863	0.094	0.570	0.846
	1	0.087	0.038	0.4	0.087	0.143	0.094	0.570	0.287
	Weighted Avg.	0.762	0.714	0.694	0.762	0.698	0.094	0.570	0.719
Support Vector Machines	0	1.000	1.000	0.772	1.000	0.872	-	0.500	0.772
	1	0.000	0.000	-	-	-	-	0.500	0.228
	Weighted Avg.	-	0.772	-	0.772	-	-	0.500	0.648

Як видно з результатів, алгоритм Logistic Regression демонструє найвищий середній показник TP Rate — 0.762. Тобто приблизно 76.2 % даних із вибірки були коректно прокласифіковані. Найнижчий TP Rate отримано під час використання Multilayer Perceptron, а для алгоритму SVM його взагалі не вдалось отримати.

Якщо переглянути інші значення метрик точності, то можна простежити, що Logistic Regression практично за усіма показниками показує найвище значення. Наступним за ним за показниками є Random Forest. Проте ці результати не можна назвати високими. Для ілюстрації отриманих результатів на рис. 1, 2 і 3 наведено ROC-криві для методів LR, RF та MP відповідно. Найвище значення ROC отримано для алгоритму Random Forest.

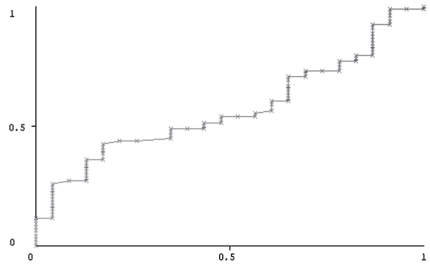

Рис 1. ROC-крива для методу *Logistic Regression*

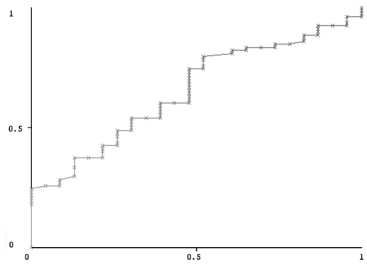

Рис 2. *ROC*-крива для методу *Random Forest*

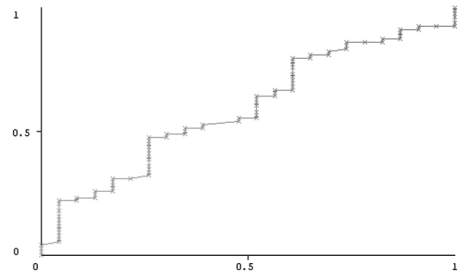

Рис 3. ROC-крива для методу *Multilayer Perceptron*

Додатково у табл. 4 подано матрицю помилок.

Таблиця 4

Матриця помилок для обраних методів

	Logistic Regression		Random Forest		Multilayer Perceptron		Support Vector Machines	
	(a)	(b)	(a)	(b)	(a)	(b)	(a)	(b)
a=1	2	21	3	20	3	20	0	23
b=0	3	75	8	70	10	68	0	78

Висновки. На основі даних 101 кандидата на навчання за програмою «Інженерія програмного забезпечення» проаналізовано ефективність чотирьох обраних алгоритмів машинного навчання для розв'язання задачі класифікації кандидатів відповідно до результатів вступного тесту. Встановлено, що найвищі показники ефективності отримано під час використання алгоритмів Logistic Regression та Random Forest, а, своєю чергою, під час використання Support Vector Machines отримано найгірші результати серед досліджуваних методів.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Hellas A. et. al. Predicting academic performance: a systematic literature review. In Proceedings Companion of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education (ITiCSE 2018 Companion). ACM, New York, NY, USA, 2018. Pp. 175–199. doi: https://doi.org/10.1145/3293881.3295783.
2. World Economic Forum Global Competitiveness Report. 2011. URL: http://www.feg.org.ua/docs/sistema_en.pdf (дата звернення 30.05.2019).
3. Sarker F et al. Student's performance prediction by using institutional internal and external open data sources. CSEDU13 : 5th International Conference on Computer Supported Education. Aachen, Germany, 2013.
4. Xu J., Han Y., Marcu D. Progressive Prediction of Student Performance in College Programs. Proceedings of the ThirtyFirst AAAI Conference on Artificial Intelligence (AAAI-17). 2017. Pp. 1604–1610.
5. Oleksiv I., Izonin I., Kharchuk V., Tkachenko R., Doroshenko A. Identification of IT Sector Stakeholder's Requirements to Masters Program in Information System in Lviv Region. In:

Ermolayev, V., Suárez-Figueroa, M. C., Ławrynowicz, A., Palma, R., Yakovyna, V., Mayr, H. C., Nikitchenko, M., and Spivakovsky, A. (Eds.): ICT in Education, Research and Industrial Applications. Proc. 14-th Int. Conf. ICTERI 2018. Volume I: Main Conference. Kyiv, Ukraine, 2018. May 14–17. Pp.112–120. CEUR-WS.org.

6. How the IT industry of Ukraine and Eastern Europe works: a report. 2019. URL: https://ain.ua/en/2019/02/15/it-industry-of-ukraine-and-eastern-europe/ (дата звернення 30.05.2019).
7. Вергун В. Огляд методів розв'язання задачі класифікації в інтелектуальному аналізі даних навчальних програм. Науковий вісник НЛТУ України. 2019. Т. 29. № 5 (в друці).
8. Pirotti F., Sunar F., Piragnolo M. Benchmark of machine learning methods for classification of a Sentine l-2 image. *International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences.* 2016. Vol. 41. Pp. 335–340.
9. Breiman L. Random forests. *Machine learning.* 2001. Vol. 45. № 1. Pp. 5–32.
10. Haykin S. Neural networks and learning machines. *Upper Saddle River.* NJ, USA : Pearson, 2009. Vol. 3. 938 p.

REFERENCES

1. Hellas, A. et. al. (2018). Predicting academic performance: a systematic literature review. In Proceedings Companion of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education (ITiCSE 2018 Companion). ACM, New York, NY, USA, 175–199. doi: https://doi.org/10.1145/3293881.3295783 (in English).
2. World Economic Forum Global Competitiveness Report. (2011). Retrieved from http://www.feg.org.ua/docs/sistema_en.pdf (дата звернення 30.05.2019) (in English).
3. Sarker, F et al. (2013). Student's performance prediction by using institutional internal and external open data sources. CSEDU13 : 5th International Conference on Computer Supported Education. Aachen, Germany (in English).
4. Xu, J., Han, Y., Marcu, D. (2017). Progressive Prediction of Student Performance in College Programs. Proceedings of the ThirtyFirst AAAI Conference on Artificial Intelligence (AAAI-17), 1604–1610 (in English).
5. Oleksiv, I., Izonin, I., Kharchuk, V., Tkachenko, R., Doroshenko, A. (2018). Identification of IT Sector Stakeholder's Requirements to Masters Program in Information System in Lviv Region. In: Ermolayev, V., Suárez-Figueroa, M. C., Ławrynowicz, A., Palma, R., Yakovyna, V., Mayr, H. C., Nikitchenko, M., and Spivakovsky, A. (Eds.): ICT in Education, Research and Industrial Applications. Proc. 14-th Int. Conf. ICTERI 2018. Volume I: Main Conference. Kyiv, Ukraine, May 14–17, 112–120. CEUR-WS.org (in English).
6. How the IT industry of Ukraine and Eastern Europe works: a report. (2019). Retrieved from https://ain.ua/en/2019/02/15/it-industry-of-ukraine-and-eastern-europe/ (дата звернення 30.05.2019) (in English).
7. Verhun, V. (2019). Ohliad metodiv rozviazannia zadachi klasyfikatsii v intelektual-nomu analizi danykh navchalnykh prohram: Naukovyi visnyk NLTU Ukrainy, 29, 5 (v drutsi) (in Ukrainian).
8. Pirotti, F., Sunar, F., Piragnolo, M. (2016). Benchmark of machine learning methods for classification of a Sentine l-2 image. *International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences,* 41, 335–340 (in English).

9. Breiman, L. (2001). Random forests. *Machine learning*, 45, 1, 5–32 (in English).
10. Haykin, S. (2009). Neural networks and learning machines. *Upper Saddle River.* NJ, USA : Pearson, 3 (in English).

doi: 10.32403/1998-6912-2019-1-58-71-78

EDUCATION DATA MINING METHODS FOR PREDICTING STUDENTS PERFORMANCE

V. R. Verhun

Lviv Polytechnic National University,
12, S. Bandera St., Lviv, 79013, Ukraine
vverhun@gmail.com

The ability to predict student performance may open many opportunities in creation of advanced and personalized educational programs as well as affect educational process and quality of education. Most of the scientific publications are aimed to research feasibility of predicting final students' grades and finding the new methods and approaches that allow to make predictions in more clear and precise way. In most cases, the data mining method is used. All the data generated by any learning management systems is considered for predictive modelling. Nowadays many research publications in area of educational data mining is published but still there are many potential fields of research since all the scientific results are highly depended to context and data structure. Prediction of students' performance also depends on learning styles and teaching styles.

In this research, the problem of predicting the success of students of educational programs has been considered. The methods of machine learning for solving the classification tasks have been selected for predicting student success. The data set, on the basis of which the research on the productivity of the selected methods has been conducted, has been described. Selected performance metrics have been used to evaluate the accuracy of the selected methods. The methods of solving the classification problem with the best performance indicators have been determined by experimental way. It has been established that the highest accuracy of work is provided by the methods of logistic regression and random forest. All the performance indicators of the algorithms have been given, as well as the matrix of errors. It has been found that the efficiency of applying these methods of methods is not high, and it has a large dependency on the data set, the context and the defined problem.

Keywords: *data mining, forecasting, random forest, regression, success, curriculum, education, education data mining.*

Стаття надійшла до редакції 07.02.2019
Received 07.02.2019