ИНФОРМАТИКА И ИНФОРМАЦИОННЫЕ ПРОЦЕССЫ

УДК 004.89 Научная статья
DOI: 10.35330/1991-6639-2023-2-112-18-29
EDN: GXEHAC

Интеллектуальный анализ образовательных данных для прогноза успеваемости студентов вуза

Н. А. Попова[1], Е. С. Егорова[2]

[1] Пензенский государственный университет
440026, Россия, г. Пенза, ул. Красная, 40
[2] Пензенский государственный технологический университет
440039, Россия, г. Пенза, проезд Байдукова/ул. Гагарина, 1а/11

Аннотация. Прогресс в области интеллектуального анализа данных делает возможным использование образовательных данных для повышения качества образовательных процессов. В статье рассмотрены различные методы анализа данных об успеваемости студентов. Основное внимание уделено двум аспектам: во-первых, прогнозирование академических достижений студентов в конце четырехлетней учебной программы по программам бакалавриата; во-вторых, изучение типичных прогрессий учащихся и объединение их с результатами прогнозирования. При прогнозировании было использовано порядка 10 алгоритмов классификации. Предложен подход к улучшению производительности методов классификации, когда атрибуты классификаторов выбираются в процессе их обучения. Определены две важные группы учащихся – с низкими и высокими достижениями. Результаты показывают, что, сосредоточив внимание на небольшом количестве курсов, которые являются показателями особенно хорошей или плохой успеваемости, можно своевременно предупреждать и поддерживать студентов с низкой успеваемостью, а также давать советы и возможности студентам с высокой успеваемостью.

Ключевые слова: анализ образовательных данных, дерево решений, кластеризация, прогнозирование, успеваемость, дистилляция

Поступила 29.03.2023, одобрена после рецензирования 07.04.2023, принята к публикации 10.04.2023

Для цитирования. Попова Н. А., Егорова Е. С. Интеллектуальный анализ образовательных данных для прогноза успеваемости студентов вуза // Известия Кабардино-Балкарского научного центра РАН. 2023. № 2(112). С. 18–29. DOI: 10.35330/1991-6639-2023-2-112-18-29

MSC: 68T09 Original article

Educational data mining for predicting the academic performance of university students

N.A. Popova[1], E.S. Egorova[2]

[1] Penza State University
440026, Russia, Penza, 40 Krasnaya street
[2] Penza State Technological University
440039, Russia, Penza, 1a/11 Baidukova passage/Gagarina street

Abstract. Progress in the field of data mining makes it possible to use educational data to improve the quality of educational processes. This article examines various methods of analyzing student achievement data. The focus is on two aspects: first, predicting students' academic achievements at the

end of a four-year undergraduate curriculum; second, examining typical student progressions and combining them with the prediction results. Approximately 10 classification algorithms were used in the prediction process. An approach to improving the performance of classification methods is proposed where classifier attributes are selected during their training. Two important groups of students were identified: low-achieving and high-achieving students. The results show that by focusing on a small number of courses that are indicators of particularly good or poor performance, it is possible to prevent and support low-achieving students in a timely manner, and to provide advice and opportunities to high-achieving students.

Keywords: analysis of educational data, decision tree, clustering, forecasting, academic performance, dissociation

Submitted 29.03.2023, approved after reviewing 07.04.2023, accepted for publication 10.04.2023

For citation. Popova N.A., Egorova E.S. Educational data mining for predicting the academic performance of university students. *News of the Kabardino-Balkarian Scientific Center of RAS.* 2023. No. 2(112). Pp. 18–29. DOI: 10.35330/1991-6639-2023-2-112-18-29

Введение

Благодаря оцифровке академических процессов университеты генерируют огромное количество данных, относящихся к студентам, в электронном виде. Для них крайне важно эффективно преобразовать этот массив данных в знания, которые помогут преподавателям, сотрудникам деканатов, учебных управлений и руководству университетов анализировать их для улучшения процесса принятия решений. Кроме того, это может также повысить качество образовательных процессов за счет своевременного предоставления информации различным заинтересованным сторонам. Целью методов интеллектуального анализа данных является извлечение значимых знаний из данных. Направление исследований, где применяется интеллектуальный анализ данных, машинное обучение и статистика для изучения информации, которую создают образовательные учреждения, называется анализом образовательных данных (АОД) (Educational Data Mining).

Выделяют пять основных категорий или подходов к анализу образовательных данных [1]: прогнозирование, кластеризация, выявление взаимосвязей, открытие с помощью моделей и дистилляция данных для последующей оценки и принятия решения. Для анализа успеваемости студентов бакалавриата достаточно сочетать три подхода: прогнозирование, кластеризацию и дистилляцию данных.

Цель прогнозирования заключается в том, чтобы предсказать класс или метку объекта данных. Основной ключевой областью применения прогнозирования в АОД является прогнозирование результатов обучения учащихся. Исследования в этой области проводились на разных уровнях детализации: на уровне системы обучения, на уровне курса и т.д. Например, на уровне интеллектуальной системы обучения [2] АОД прогнозирует результаты тестов учащихся, интегрируя информацию о времени и объеме помощи, необходимой студенту для решения проблем. На уровне курса [3] прогнозируют успеваемость по курсу на основе успеваемости студентов на контрольных точках текущей успеваемости и промежуточных экзаменах.

При кластеризации цель состоит в том, чтобы сгруппировать объекты в классы похожих объектов. Поскольку кластеризация используется в интеллектуальном анализе образовательных данных для решения широкого спектра задач, интересной областью является группирование учащихся для изучения моделей типичного поведения. В работе [4] идентифицируют группы учащихся с аналогичной успеваемостью от средней школы до окончания вуза.

Дистилляция данных для человеческого суждения соответствует тому, что другие называют обзорной статистикой и визуализацией, с целью помочь в понимании результатов анализов. Например, в [5] используют интуитивную визуализацию результатов анализа, которая дает представление о процессах обучения учителям, поставщикам электронного обучения и исследователям. В работе [6] дендрограммы сочетаются с тепловыми картами, чтобы обеспечить интуитивную визуализацию отличительных групп студентов.

Целью данного исследования является на основе анализа успеваемости студентов, обучающихся по 4-летней программе бакалавриата по направлению подготовки «Информационные системы и технологии», предсказать успеваемость студентов на ранней стадии освоения учебной программы, определить курсы, которые могут служить индикаторами хорошей или низкой успеваемости в конце обучения, а также установить типичные показатели успеваемости студентов во время учебы и соотнести их с курсами-индикаторами. Для достижения этой цели был разработан подход, состоящий из трех частей.

Во-первых, генерируются несколько классификаторов для прогнозирования успеваемости студентов до окончания образования. Для построения этих классификаторов используются только вступительные баллы из аттестата о среднем образовании и результаты промежуточной аттестации за первый и второй годы обучения в университете, социально-экономические или демографические особенности не учитываются.

Во-вторых, используя эти классификаторы, определяются курсы, которые могут служить эффективными показателями успеваемости студентов в рамках учебной программы. Это позволит поддержать студентов из группы риска или еще больше стимулировать студентов, подающих надежды. Для получения таких курсов-индикаторов был выбран метод построения дерева решений, который для анализа образовательных данных обладает отличными показателями интерпретируемости модели и удовлетворительными показателями точности.

В-третьих, изучается, как академическая успеваемость студентов меняется в течение 4-летней программы обучения, и представляется в виде триангуляций. Используя методы кластеризации, мы делим студентов на группы таким образом, чтобы учащиеся одной и той же группы имели одинаковую типовую успеваемость. Следствием такой группировки является прогноз групп студентов, которые имеют низкие оценки, и студентов с высокими оценками на протяжении всей учебы. Полученные результаты могут использоваться для разработки программы корректировки образовательного процесса студентов: для первой группы предлагать меры по повышению успеваемости, для второй группы предлагать различные мероприятия для дополнительного стимулирования и развития успешных студентов.

1. Данные для исследования

Для проведения исследования были использованы оценки студентов за четыре года обучения. В этом исследовании используются данные академических групп из трех вузов, обучающихся по одинаковым направлениям подготовки с использованием выборки из 210 студентов бакалавриата, которые были зачислены в 2017 и 2018 годах и разделены на две тестовые группы. Набор данных включает два типа переменных:

1. Переменные, связанные с оценками студентов перед поступлением (используются при отборе студентов для поступления в университет): средний балл оценок из аттестата о среднем образовании; балл, набранный студентом на едином государственном экзамене (ЕГЭ)

по математике; общий балл, полученный учащимся при сдаче ЕГЭ, – сумма баллов по русскому языку, математике и информатике/физике.

2. Переменные, связанные с оценками по всем курсам, которые преподаются в течение четырех лет образовательной программы. Полный список курсов, использованных в исследовании, состоит из 50 дисциплин учебного плана. Например, Б1.О.34 Прикладное программное обеспечение, Б1.О.03 Современные информационные технологии, Б1.О.05 Математика, Б1.О.06 Иностранный язык, Б1.О.16 Информационные технологии в профессиональной деятельности.

Оценки в конце обучения рассчитываются как сумма среднего экзаменационного балла за каждый год обучения, причем применяются весовые коэффициенты 0,1 для первого года обучения, 0,2 для второго года, 0,3 для третьего года и 0,4 для четвертого года. Интервал оценки делится на пять возможных значений /категорий: A (90-100), B (80-89), C (70-79), D (60-69) и E (50-59), поскольку в целях прогнозирования требуется ввести более уточненную градацию оценок, нежели принятая в вузах 5-балльная. Такие данные можно получить из внедренной в практику рейтинговой системы. Статистические данные о группах и оценках в выбранных интервалах в конце обучения приведены в таблице 1.

Таблица 1. Статистические данные по оценкам студентов

Table 1. Statistical data on student grades

Группа	Кол-во студ.	Кол-во студ. в интервале A	Кол-во студ. в интервале B	Кол-во студ. в интервале C	Кол-во студ. в интервале D	Кол-во студ. в интервале E
Группа 1	106	1	41	46	14	4
Группа 2	104	-	31	54	18	1

В таблице 2 представлен обзор распределения оценок учащихся за четыре года. Средние оценки для каждого студента за каждый год были рассчитаны как число от 0 (худшая оценка) до 100 (лучшая оценка).

Таблица 2. Распределение годовых оценок по группам

Table 2. Distribution of annual grades by groups

Курс обучения	A (90-100)	B (80-89)	C (70-79)	D (60-69)	E (50-59)
Первая группа					
1-й курс	0	9	56	31	9
2-й курс	0	13	55	25	12
3-й курс	1	47	37	16	4
4-й курс	6	62	26	11	0
Вторая группа					
1-й курс	0	14	55	29	6
2-й курс	0	13	46	34	11
3-й курс	0	30	48	22	4
4-й курс	0	31	54	18	1

Все методы интеллектуального анализа данных в этом исследовании были выполнены с помощью программного обеспечения RapidMiner [7].

2. Методика и результаты исследований

При реализации предложенного подхода к анализу успеваемости студентов на этапе прогнозирования было использовано несколько алгоритмов классификации. Обзор литературы показывает, что в целом не существует единого классификатора, который лучше всего работает во всех контекстах для обеспечения хорошего прогнозирования. Поэтому существует необходимость выяснить, какие классификаторы больше подходят для анализируемых данных. Согласно данным таблицы 1, распределение студентов по интервалам является несбалансированным. Интервал класса «C» содержит наибольшее количество учащихся для обеих групп. Точность прогнозирования студента класса «C» составляет 51,92 %, и это формирует базовую линию, которую мы хотим улучшить. В таблице 3 представлены результаты точности и коэффициента Каппа Коэна (Каппа) 10 классификаторов, которые достигли точности выше базового уровня. Эти результаты показывают, что данные группы 1 могут с достаточной точностью предсказать успеваемость студентов группы 2 в конце обучения, используя оценки до поступления в университет и оценки за 1-й и 2-й курсы. В качестве инструментального средства для проведения исследования была выбрана платформа RapidMiner [8], имеющая в своем составе множество встроенных механизмов интеллектуального анализа данных.

Таблица 3. Точность предсказания классификатора

Table 3. Prediction accuracy of the classifier

Классификатор	**Точность / Каппа**
Дерево решений с индексом Джини	68,27 % / 0,493
Дерево решений с приростом информации	69,23 % / 0,498
Дерево решений с точностью	60,58 % / 0,325
Индукция правила с приростом информации	55,77 % / 0,352
Метод ближайших соседей	74,04 % / 0,583
Наивный байесовский алгоритм	83,65 % / 0,727
Нейронные сети	62,50 % / 0,447
Случайный лес с индексом Джини	71,15 % / 0,543
Случайный лес с приростом информации	69,23 % / 0,426
Случайный лес с точностью	62,50 % / 0,269

Для каждого метода были построены матрицы ошибок. Результирующая матрица ошибок для классификатора «дерево решений с индексом Джини» показана в таблице 4.

Таблица 4. Матрица ошибок

Table 4. Error matrix

Дерево решений с индексом Джини		Фактический				Точность
		B	C	D	E	
Прогнозируемый	B	18	6	0	0	75.00 %
	C	13	38	2	0	71.70 %
	D	0	10	14	0	58.33 %
	E	0	0	2	1	33.33 %
Отзыв класса		58.06 %	70.37 %	77.78 %	100.00 %	

Согласно матрице в первой колонке из 31 (т.е. 18+13) фактического студента класса «B» классификатор правильно предсказал 18 как «B», отзыв для класса «B» составляет 58,06 (т.е. 18/31). В первой строке 24 (18+6) студента были спрогнозированы как класс «B», точность в этом случае для класса «B» составляет 18/24 = 75 %. Оставшиеся столбцы и строки интерпретируются аналогично для других классов. Все правильные прогнозы находятся в диагонали матрицы. Можно заметить, что класс «A» отсутствует. Это связано с тем, что в группе 1 есть только один учащийся, принадлежащий к интервалу «A», а в группе 2 нет ни одного. Получается, что в матрицах ошибок есть только нули для класса «A», поэтому столбец для класса «A» не учитывается. Анализ матриц ошибок показал, что классификаторы испытывают трудности с классами, которые недостаточно представлены данными, такие как «A» и «E». Лучшая точность достигается для хорошо представленных классов, таких как «B» и «C».

На следующем этапе подхода для выделения курсов-индикаторов проанализированы полученные классификаторы. Согласно таблице 3 классификаторы с лучшими показателями производительности – наивный байесовский алгоритм, метод ближайших соседей и случайный лес с индексом Джини. Недостатком этих трех классификаторов является то, что они практически не интерпретируемы для человека: невозможно понять, какие переменные или атрибуты (в нашем контексте какие курсы) влияют на прогнозирование. В отличие от них деревья решений показывают, какие атрибуты приводят к прогнозированию конкретной метки, и поэтому полезны для понимания результатов. Была использована оригинальная техника выбора атрибутов, и она значительно улучшила производительность деревьев решений. Суть техники заключается в том, чтобы выбирать атрибуты деревьев решений в процессе их обучения. Расширив набор данных за счет рассмотрения еще двух последовательных групп студентов, построили деревья решений с учетом четырех критериев [9]: индекс Джини (Gini index), прирост информации (information gain), доля правильных ответов (accuracy), соотношение прироста информации и информации, необходимой для разбиения (gain ratio), используя одну группу для построения модели и последующую группу для ее тестирования. Далее отобрали курсы, встречающиеся как минимум в половине всех деревьев решений, это нам дало набор дисциплин, больше всего влияющих на успеваемость. После этого рассмотрели части путей, встречающихся по крайней мере в половине деревьев решений и ведущих к узлам, помеченным как «B» чистые узлы или, если они нечистые, содержащие элементы с классом «A», и пути, ведущие к узлам, помеченным «D» или «E», в последнем случае нечистые узлы могут включать элементы с классом «C». Курсы, происходящие на тех путях, которые ведут к узлам, помеченным «B», представляют собой показатели хорошей успеваемости, и курсы на путях, ведущих к узлам, помеченным «D» или «E», представляют собой показатели низкой успеваемости. В результате в качестве атрибутов было выбрано 5 курсов: Б1.О.34, Б1.О.03, Б1.О.05, Б1.О.06, Б1.О.16, и на их основе построены деревья решений с точностью выше базового уровня на группе 1 и протестированы на группе 2 (рис. 1).

Полученные результаты можно кратко интерпретировать следующим образом:

– в первый год учащиеся, чьи оценки по дисциплине Б1.О.05 составляют около 63 баллов или меньше, вероятно, будут иметь отметку в интервале «D» или «E» в конце обучения;

– на втором курсе учащиеся, получившие оценки ниже 60 баллов по курсу Б1.О.03 или 53 по Б1.О.16, вероятно, будут иметь отметку в интервале «D» или «E» в конце обучения;

– на втором курсе учащиеся, получившие оценку 80 баллов или выше по курсу Б1.О.03, или учащиеся, набравшие более 86 баллов по Б1.О.34, скорее всего, получат оценку в интервале «A» или «B» в конце обучения.

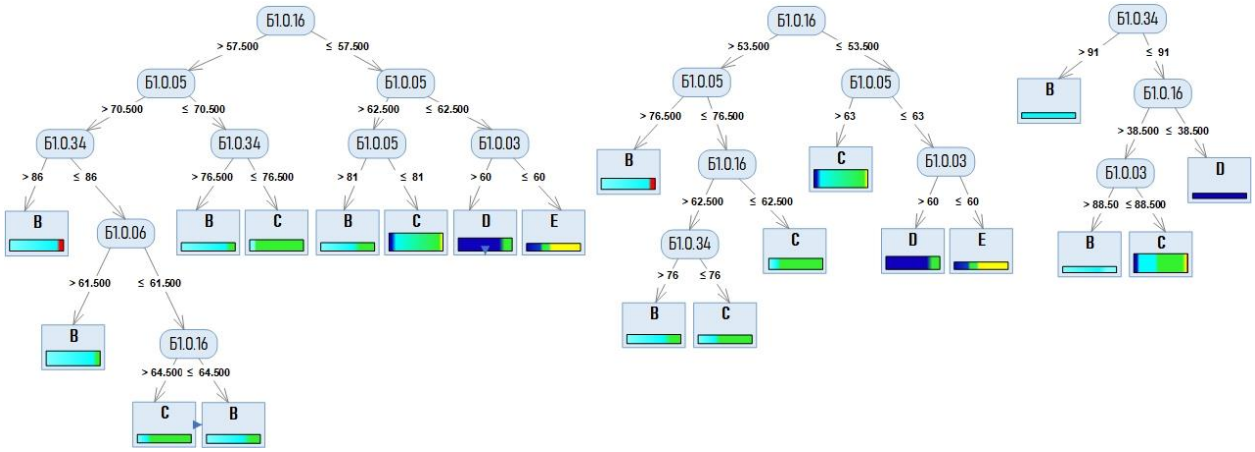

Рис. 1. *Деревья решений с индексом Джини, приростом информации и долей правильных ответов*

Fig. 1. *Decision trees with Gini index, information gain and percentage of correct answers*

Последним этапом подхода является изучение того, как улучшается успеваемость студентов во время учебы, и связи этого прогресса с курсами-индикаторами. С этой целью анализируются типичные модели прогрессирования студентов на всем протяжении их учебы. Прогрессии оценок рассматриваются не по абсолютной шкале, а в сравнении с другими студентами. Каждый студент представлен вектором, состоящим из оценок, полученных за каждый год обучения. Чтобы выявить типичные закономерности прогрессирования на протяжении нескольких лет, студенты ежегодно объединяются в группы на основе своих оценок по промежуточной аттестации по каждой дисциплине. Применяется кластеризация X-средних с евклидовым расстоянием. Таким образом, прогресс каждого студента представлен 4 кортежами, указывающими на кластер, к которому принадлежит студент в каждом году, с целью найти прогрессии, представляющие учащихся с высокой успеваемостью, и прогрессии, представляющие учащихся с низкой успеваемостью.

Для каждой дисциплины на основе оценок группы 1 определены три кластера: «Низкий», «Промежуточный», «Высокий». Центроиды дают среднюю оценку кластера по каждому курсу. Например, для дисциплины первого курса обучения Б1.О.27 определены 3 кластера с центроидами: «Низкий» со средней оценкой 61.714, «Промежуточный» – 75.636, «Высокий» – 87.259.

В таблице 5 представлено общее количество учащихся в каждом кластере за все четыре года для группы 1 и группы 2 соответственно.

Таблица 5 также показывает, что для группы 1 большинство учащихся относится к среднему кластеру во все годы, кроме второго года, когда кластер «Высокий» особенно велик. На третий год промежуточный кластер распадается на два. Как и в группе 1, большинство студентов группы 2 относятся к промежуточному кластеру на всех курсах, кроме первого года, когда кластера «Промежуточный» не существует. Получившиеся

данные хорошо согласуются с данными из таблицы 2, в которой самое большое количество студентов в интервале 70-80 за все четыре года для обеих групп, за исключением четвертого года группы 1.

Таблица 5. Количество студентов по кластерам за четыре года

Table 5. Number of students by clusters for four years

Кластер	Группа 1				Группа 2			
	1-й год	2-й год	3-й год	4-й год	1-й год	2-й год	3-й год	4-й год
Низкий	32	15	14	23	49	20	18	34
Промежуточный	47	25	27 (Промежуточный-Низкий) 32 (Промежуточный-Высокий)	49	-	41	43	44
Высокий	27	66	33	34	55	43	43	26

Сгруппировав каждую группу четыре раза, типичный прогресс учащегося определяется следующим образом. Сначала вычисляется и округляется среднее значение всех центров каждого кластера за каждый год. Например, для группы 1 за первый год обучения среднее значение кластера «Низкий» равно 60, среднее значение кластера «Промежуточный» – 71, а среднее значение кластера «Высокий» – 78; для группы 2 в первый год есть всего 2 кластера: среднее значение кластера «Низкий» равно 60, а среднее значение кластера «Высокий» равно 73. Далее, чтобы получить интуитивно понятное описание того, как успеваемость учащихся в глобальном масштабе изменяется в течение четырех лет обучения, каждый студент описывается четырьмя кортежами, элементами которых являются средние значения центров кластеров, в которых находится студент. Например, для группы 1 студент, принадлежащий кластеру с самой низкой оценкой на первом и втором курсе, с промежуточно-низкой оценкой на третьем курсе и промежуточной оценкой на четвертом курсе будет представлен кортежем (60, 52, 66, 77), а студент, принадлежащий к кластеру с высокими оценками за все четыре года, будет описан кортежем (78, 73, 83, 85).

На рисунке 2 представлены кортежи всех студентов для группы 1 в виде иерархической гистограммы. Высота столбца представляет количество учащихся, характеризуемых одним и тем же кортежем. Диаграмма упорядочена справа налево: низкие значения справа, высокие значения слева. Цвет указывает на кластеры первого года. Второй год изображен внизу диаграммы и разделен на разные части, соответствующие среднему значению кластеров. Каждая из этих частей делится на кластеры 3-го года. Наконец, высший уровень иерархии делит каждую часть кластерами 4-го года, которые нарисованы прямо под столбиками.

Сравнение гистограмм для обеих групп показало, что больше студентов с более высокими оценками на втором курсе и высокими или средними оценками на третьем и четвертом курсах в группе 1, чем в группе 2, что визуализирует тенденцию, представленную в таблице 2. Выделены два критических класса студентов: один класс студентов – наиболее успевающие – обозначен высокой зеленой полосой в крайнем левом углу, второй класс состоит из учеников с низкой успеваемостью, обозначен красной полосой в крайнем правом углу. Учащиеся с высокими показателями успеваемости имеют высокие оценки в каждом из четырех лет, и эта зависимость сохраняется для обеих групп студентов. Учащиеся с низкой успеваемостью имеют низкие оценки за все четыре года, и эта зависимость проявляется

несколько больше в группе 2, чем в группе 1. Другой важной зависимостью для обеих групп является образование классов, которые состоят из учащихся, имеющих промежуточные оценки за все годы или имеющих промежуточные оценки за все годы, кроме одного. Второй и четвертый высшие столбцы рисунка 2 показывают такие классы для группы 1: второй высший столбец – студенты с промежуточными оценками на всех курсах, кроме второго года, где они имеют высокие оценки с кортежем (71, 73, 75, 77), четвертый высший столбец – студенты с промежуточными оценками на всех годах, кроме 1-го курса, где у них низкие оценки с кортежем (60, 62, 66, 77).

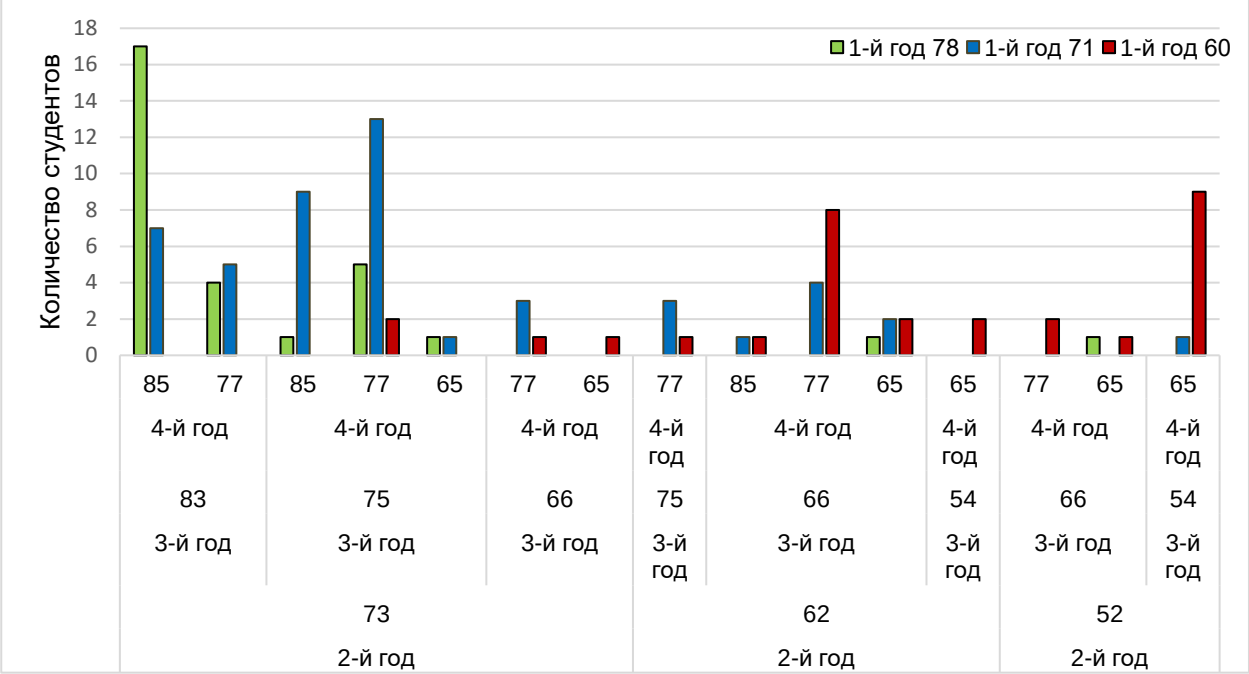

Рис. 2. *Краткое описание кортежей группы 1*

Fig. 2. *Brief description of group 1 tuples*

Визуализация полученных данных показывает, что очень мало нетипичных студентов, таких, которые имеют низкие оценки на 1-м курсе, но затем продвигаются и заканчивают с высокими оценками на 4-м курсе. В группе 2 есть один такой студент, заданный кортежем (60, 69, 70, 84). Еще один интересный и тоже небольшой кластер составляют студенты, имеющие высокие оценки на 1-м курсе, но низкие оценки на всех последующих курсах. Маленькие зеленые полосы справа от диаграммы изображают их. Их двое во 2-й группе (73, 58, 60, 65) и один в 1-й группе (78, 62, 66, 65).

Данный подход к кластеризации студентов по годам позволяет обнаружить небольшие, но интересные кластеры с нетипичными учениками, которые начинают с низких отметок и заканчивают с высокими отметками или наоборот. Эти учащиеся не обнаруживаются при кластеризации с учетом оценок всех лет обучения вместе.

Заключение

В настоящем исследовании был рассмотрен подход к анализу успеваемости студентов бакалавриата с целью предоставить преподавателям и руководителям структурных подразделений университета информацию, которая могла бы помочь им улучшить образовательные программы.

На первом этапе провели прогнозирование успеваемости учащихся, используя только оценки, без каких-либо социально-экономических данных. Результаты показывают, что предсказать успеваемость четырехлетнего обучения, используя оценки довузовского образования и оценки за первый и второй годы обучения, можно с достаточной точностью.

На втором этапе были определены дисциплины, которые могут служить эффективными индикаторами хорошей или плохой успеваемости в программе бакалавриата. Данное исследование проводилось с помощью деревьев решений с учетом четырех критериев: индекс Джини, прирост информации, доля правильных ответов, соотношение прироста информации и информации, необходимой для разбиения.

На заключительном этапе проводилось изучение процесса динамического изменения академической успеваемости студентов в течение четырехлетнего обучения. Подтверждено, что на протяжении каждого года студенты, как правило, получают одни и те же оценки: низкие, промежуточные или высокие по всем предметам. Причем эта закономерность повторяется на протяжении всего срока обучения. Таким образом, выявлены две основные группы: группа высокоуспевающих студентов, которые получали высокие оценки в течение четырех лет, и группа низкоуспевающих студентов, которые получали низкие оценки. Такой подход позволяет уже на ранних курсах выявить студентов, которые будут испытывать трудности на протяжении всей программы обучения и, скорее всего, получат низкий средний балл по диплому. С такими студентами можно проводить различные мероприятия, позволяющие повысить успеваемость. С другой стороны, раннее определение высокоуспевающих студентов позволяет ориентировать их на подготовку и поступление в магистратуру.

СПИСОК ЛИТЕРАТУРЫ

1. *Белоножко П. П., Карпенко А. П., Храмов Д. А.* Анализ образовательных данных: направления и перспективы применения // Интернет-журнал «Науковедение» Том 9. № 4 (2017). URL: http://naukovedenie.ru/PDF/15TVN417.pdf

2. *Мамонтова М. Ю.* Качество учебных достижений: оценка и прогноз на основе результатов критериально-ориентированного тестирования // Образование и наука. Известия УрО РАО. 2009. № 3(60). С. 18–26.

3. *Русаков С. В., Накарякова Н. Н.* Прогнозирование успеваемости студентов первого курса с помощью дерева решений на основе их результатов сдачи ЕГЭ // Наука. Информатизация. Технологии. Образование: Материалы XI международной научно-практической конференции. Екатеринбург: Российский государственный профессионально-педагогический университет, 2018. С. 589–594.

4. *Фирстов В. Е.* Социометрические и информационные аспекты кластеризации обучаемого контингента при организации и оптимизации группового сотрудничества в учебном процессе в школе и вузе // Известия Саратовского университета. Новая серия. Серия: Философия. Психология. Педагогика. 2014. Т. 14. № 1. С. 110–118.

5. *Medvedev D., D'yakonov A.* New Properties of the Data Distillation Method When Working with Tabular Data // Conference proceedings "Analysis of Images, Social Networks and Texts". *Lecture Notes in Computer Science*. Vol. 12602. Springer, Cham. 2021. https://doi.org/10.1007/978-3-030-72610-2_29

6. *Sucholutsky I., Schonlau M.* Soft-Label Dataset Distillation and Text Dataset Distillation // International Joint Conference on Neural Networks, Shenzhen, China, 2021. Pp. 1–8. DOI: 10.1109/IJCNN52387.2021.9533769.

7. *Никонорова М. Л.* Компьютерная модель решения задач классификации в программной среде Rapid Miner // Медицинское образование и профессиональное развитие. 2017. № 2–3(28–29). С. 24–33.

8. *Филяк П. Ю., Виноградов М. А.* Применение Rapid miner и открытых сред как инструментов интеллектуального анализа данных для обеспечения безопасности // Информация и безопасность. 2017. Т. 20. № 4. С. 552–555.

9. *Maimon O., Rokach L.* Data Mining and Knowledge Discovery Handbook. Springer Science, Business Media, 2010. 1285 p. ISBN: 978-0-387-09822-7.

Информация об авторах

Попова Наталия Александровна, канд. техн. наук, доцент кафедры «Математическое обеспечение и применение ЭВМ», Пензенский государственный университет;

440026, Россия, г. Пенза, ул. Красная, 40;

popov.tasha@yandex.ru, ORCID: https://orcid.org/0000-0001-9713-4897

Егорова Екатерина Сергеевна, канд. экон. наук, доцент кафедры «Прикладная информатика», Пензенский государственный технологический университет;

440039, Россия, г. Пенза, проезд Байдукова/ул. Гагарина, 1а/11;

katepost@yandex.ru, ORCID: https://orcid.org/0000-0002-0816-0944

REFERENCES

1. Belonozhko P.P., Karpenko A.P., Khramov D.A. Analysis of educational data: directions and prospects for application. *Internet-zhurnal "Naukovedeniye"* [SCIENCE online journal]. Vol. 9. No. 4 (2017). URL: http://naukovedenie.ru/PDF/15TVN417.pdf (In Russian)

2. Mamontova M.Yu. Quality of educational achievements: assessment and forecast based on the results of criteria-oriented testing. *Education and science. News of the Ural Branch of the Russian Academy of Education*. 2009. No. 3(60). Pp. 18–26. (In Russian)

3. Rusakov S.V., Nakaryakova N.N. Forecasting the progress of first-year students using a decision tree based on their results of passing the exam. *Nauka. Informatizatsiya. Tekhnologii. Obrazovaniye* [Science. Informatization. Technologies. Education]: *Materialy XI mezhdunarodnoy nauchno-prakticheskoy konferentsii.* Yekaterinburg: Rossiyskiy gosudarstvennyy professional'no-pedagogicheskiy universitet, 2018. Pp. 589–594. (In Russian)

4. Firstov V. E. Sociometric and informational aspects of clustering the student contingent in the organization and optimization of group cooperation in the educational process at school and university. *Izvestiya of Saratov University. New series. Series: Philosophy. Psychology. Pedagogy*. 2014. Vol. 14. No. 1. Pp. 110–118. (In Russian)

5. Medvedev D., D'yakonov A. New Properties of the Data Distillation Method When Working with Tabular Data. Conference proceedings "Analysis of Images, Social Networks and Texts". *Lecture Notes in Computer Science*. Vol. 12602. Springer, Cham. 2021. https://doi.org/10.1007/978-3-030-72610-2_29

6. Sucholutsky I., Schonlau M. Soft-Label Dataset Distillation and Text Dataset Distillation. *International Joint Conference on Neural Networks*, Shenzhen, China, 2021. Pp. 1–8. DOI: 10.1109/IJCNN52387.2021.9533769.

7. Nikonorova M.L. Computer model for solving classification problems in the Rapid Miner software environment. *Medical education and professional development*. 2017. No. 2–3(28–29). Pp. 24–33. (In Russian)

8. Filyak P.Yu., Vinogradov M.A. Application of Rapid miner and open environments as data mining tools for security. *Informatsiya i bezopasnost'* [Information and security]. 2017. Vol. 20. No. 4. Pp. 552–555. (In Russian)

9. Maimon O., Rokach L. Data Mining and Knowledge Discovery Handbook. Springer Science, Business Media, 2010. 1285 p. ISBN: 978-0-387-09822-7.

Information about the authors

Popova Nataliya Aleksandrovna, Candidate of Technical Sciences, Associate Professor of the Department of Mathematical Support and Computer Use, Penza State University;

440026, Russia, Penza, 40 Krasnaya street;

popov.tasha@yandex.ru, ORCID: https://orcid.org/0000-0001-9713-4897

Egorova Ekaterina Sergeevna, Candidate of Economic Sciences, Associate Professor of the Department of Applied Informatics, Penza State Technological University;

440039, Russia, Penza, 1a/11 Baidukova passage/Gagarina street;

katepost@yandex.ru, ORCID: https://orcid.org/0000-0002-0816-0944