DOI: 10.15593/2499-9873/2020.4.08
УДК 378.02

Н.Н. Накарякова, С.В. Русаков, О.Л. Русакова

Пермский государственный национальный
исследовательский университет, Пермь, Россия

ПРОГНОЗИРОВАНИЕ ГРУППЫ РИСКА (ПО УСПЕВАЕМОСТИ) СРЕДИ СТУДЕНТОВ ПЕРВОГО КУРСА С ПОМОЩЬЮ ДЕРЕВА РЕШЕНИЙ

Массовое обучение в российских вузах по специальностям (направлениям), связанным с точными и техническими науками, характеризуется высоким уровнем отсева начиная с первого курса обучения. Существующие на сегодняшний день уровень школьного образования, система отбора абитуриентов через процедуру ЕГЭ во многих случаях не гарантируют, что будущие студенты смогут успешно освоить наукоемкие специальности. Упор на личностно-ориентированное, индивидуальное обучение возможен только после того, как студенты проявили себя на первых этапах учебы, поэтому опережающее выявление способности вчерашних абитуриентов эффективно учиться является весьма актуальной задачей.

Рассматриваются методики построения деревьев решений, предназначенных для классификации студентов, выделяя из них множество тех (группу риска), кто с высокой долей вероятности будет отчислен уже по итогам первого учебного цикла (триместра). При этом в качестве входных данных используется минимальная информация о первокурсниках, зафиксированная в их личном деле. Построение модели осуществлялось по данным о студентах направления «Прикладная математика и информатика» Пермского государственного национального исследовательского университета за пятилетний период наборов (2014–2018 гг.). При этом информация 2014–2017 гг. использовалась для обучения, а поток 2018 г. в качестве тестового. На этапе машинного обучения было рассмотрено несколько моделей деревьев решений, которые оптимизировались с помощью балансировки, ограничения на максимальную глубину дерева и минимального количества элементов в листе. Эффективность бинарной классификации оценивалась с помощью матрицы неточностей и целого ряда получаемых на ее основе числовых критериев.

В результате машинного обучения построено дерево решений, которое спрогнозировало попадание в группу риска 16 из 17 человек, отчисленных уже по итогам первого триместра. Иными словами, оказавшихся по ряду причин неспособными к обучению по направлению «Прикладная математика и информатика». Помимо этого, удалось определить уровень значимости различных видов исходных данных, показав, что результаты ЕГЭ в значительной мере определяют успешность студентов на этом этапе обучения. Определение группы риска дает определенные ориентиры для целенаправленной деятельности педагогов и вузовских психологов, что в конечном итоге может послужить основанием для повышения качества обучения и уменьшения отсева. Выполненная работа демонстрирует возможности методов интеллектуального анализа данных при решении плохо формализуемых задач, характерных для этого вида деятельности человека.

Ключевые слова: группа риска, машинное обучение, обучающая и тестовая выборки, дерево решений, балансировка дерева решений, случайный лес, матрица неточностей, ошибка I и II рода, кривая ошибок, значимость параметров модели.

N.N. Nakaryakova, S.V. Rusakov, O.L. Rusakova

Perm State University, Perm, Russian Federation

PREDICTION OF THE RISK GROUP (BY ACADEMIC PERFORMANCE) AMONG FIRST COURSE STUDENTS BY USING THE DECISION TREE METHOD

Mass education in Russian universities in specialties (direction of study) related to the exact and technical sciences is characterized by a high dropout rate, starting from the first year of study. The current level of school education, the system for selecting applicants through the USE procedure, in many cases does not guarantee that future students will be able to successfully master science-intensive specialties. An emphasis on student-centered, individual learning is possible only after students have proven themselves in the early stages of their studies. Therefore, the anticipatory identification of the ability of yesterday's applicants to study effectively is a very urgent task.

In this paper, we consider methods for constructing decision trees designed to classify students, highlighting from them a lot of those (risk group) who, with a high degree of probability, will be expelled after the first academic cycle (trimester). At the same time, the minimum information about the freshmen, recorded in their personal file, is used as input data. The construction of the model was carried out according to the data on students of the applied mathematics and computer science direction of the Perm State National Research University for a five-year period of sets of 2014-2018. At the same time, the information from 2014-2017 was used for training, and the flow of 2018 was used as a test one. At the stage of machine learning, several models of decision trees were considered, which were optimized using balancing, restrictions on the maximum tree depth and the minimum number of elements in a leaf. The effectiveness of the binary classification was assessed using a matrix of inaccuracies and a number of numerical criteria obtained on its basis.

As a result of machine learning, a decision tree was built, which predicted 16 out of 17 people expelled from the first trimester into the risk group. That is, for a number of reasons, they turned out to be incapable of learning in the direction of applied mathematics and computer science. In addition, it was possible to determine the level of significance of various types of initial data, showing that the results of the USE largely determine the success of students at this stage of training. The definition of the risk group provides certain guidelines for the purposeful activity of teachers and university psychologists, which ultimately can serve as a basis for improving the quality of education and reducing dropout rates. The work performed demonstrates the capabilities of data mining methods in solving poorly formalized tasks characteristic of this type of human activity.

Keywords: risk group, machine learning, training and test sample, decision tree, balancing decision tree, random forest, confusion matrix, type I and II errors, error curve, significance of model parameters.

Введение

Практически во всех российских вузах наблюдается существенный отсев студентов первого курса уже по результатам первой сессии, в особенности это касается направлений и специальностей, связанных с точными науками. Отчасти это является результатом слабой школьной подготовки, но также имеет место влияние и других факторов.

В работе [1] показано, что предсказательная способность суммарного балла ЕГЭ является приемлемой для того, чтобы признать этот экзамен валидным инструментом отбора абитуриентов. Авторы

исследования сделали вывод, что предсказательная способность баллов ЕГЭ по отдельным предметам, составляющих суммарный итоговый балл ЕГЭ, примерно одинакова, но все же ЕГЭ по математике и русскому языку являются лучшими предикторами для подавляющего большинства направлений. При этом ЕГЭ по профильным предметам часто оказываются слабее связаны с дальнейшей успеваемостью [1]. Но качество образования определяется не столько баллами ЕГЭ, сколько общим уровнем подготовки, который, в свою очередь, во многом определяется учебным заведением, где учился абитуриент. Так, в работе [2] помимо результатов ЕГЭ рассматривалось влияние характеристики школ на успеваемость студентов. В ходе исследования с помощью эконометрического моделирования авторами было доказано, что студенты из школ с высоким средним баллом ЕГЭ по математике учатся в среднем лучше, чем студенты из школ с низким значением этого балла.

В качестве другой группы факторов, влияющих на успешность обучения, ряд исследователей выделяют тип родительских семей и социально-бытовых условий, в которых проживают студенты. Так, в работе [3] было произведено обобщение исследований взаимосвязи динамики успеваемости студентов в различных типах родительских семей. Авторы пришли к выводу, что успеваемость студентов из малодетных семей значительно выше, чем многодетных: 52 % студентов, находящихся под контролем родителей, обучаются на хорошо и отлично; 30 % – на хорошо и удовлетворительно и 18 % учатся на удовлетворительно и с долгами. Среди студентов из малообеспеченных семей обучается на хорошо и удовлетворительно 85 %, на отлично – 5 % и учатся с долгами – 10 %. В работе [4] было проведено исследование, в котором выборку данных поделили на пять групп по социальному статусу родителей: предприниматель, пенсионер, фермер, рабочий, госслужащий. В ходе исследования авторы установили, что лучшую успеваемость имеют дети госслужащих, рабочих, фермеров, пенсионеров, а наихудшую – дети предпринимателей. Авторы считают, что такие результаты объясняются тем, что студенты из семей предпринимателей не стеснены в материальных благах и, возможно, меньше мотивированы к обучению. В то же время студенты из малообеспеченных семей наиболее мотивированы к учебе с целью

получить престижную профессию и улучшить материальное положение. Успехи студентов из семей госслужащих, возможно, связаны с тем, что родители могут больше уделять внимания детям из-за четкого регламента рабочего дня.

Важную роль в успешности студентов-первокурсников играет их физическое и морально-психологическое состояние. В работе [5] путем психологического тестирования анализировались различные личностные факторы. Была установлена существенная роль способности адаптации к новым условиям и уровня мотивированности на учебную деятельность и результаты обучения.

Попытка учесть большую группу факторов была выполнена в работе [6], где авторы на основе нейросетевой модели «нарисовали портрет» кандидата на попадание в группу риска: студент имеет невысокий балл ЕГЭ по математике, проживает в общежитии, учился не в г. Перми, не изучал в школе английский язык, получает социальную стипендию. Общая точность модели составила около 80 %. Здесь к группе риска относились те студенты, которые с высокой вероятностью могли быть отчислены уже по результатам первой сессии.

Таким образом, можно выделить три основные группы факторов, влияющих на успешность обучения студентов-первокурсников:

– уровень знаний (определяется баллами ЕГЭ и школой, в которой обучался студент);

– социально-бытовые условия (социальный статус семьи и место проживания студента);

– здоровье и личностные качества студента.

При оценке успеваемости студентов в настоящее время широко используются различные методы, характерные для Data Mining. Так, например, в работах [7, 8] для этих целей используется кластерный и дискриминантный анализ, в работе [9] – модели нечеткой логики, в работе [10] – нейросетевые модели.

В настоящей работе по данным о наборах студентов направления «Прикладная математика и информатика» механико-математического факультета ПГНИУ за 2014–2018 гг. проведено исследование на основе данных из первых двух групп факторов. В качестве инструмента, позволяющего выделить группу риска, использовалось дерево решений (decision tree).

1. Деревья решений и методика оценки качества классификации

Метод распознавания, основанный на построении решающего дерева, относится к типу логических методов. В данном алгоритме распознавание объекта осуществляется как процесс прохождения по бинарному дереву из корня в некоторую висячую вершину.

Бинарным корневым деревом называется дерево, имеющее следующие свойства:

– каждая вершина (кроме корневой) имеет одну входящую дугу;

– каждая вершина имеет либо две, либо ни одной выходящей дуги.

Вершины, имеющие две выходящие дуги, называются внутренними, а остальные – терминальными или листьями.

Деревья решений создают иерархическую структуру классифицирующих правил типа if-then («если …, то…»), т.е. в каждой вершине вычисляется определенная логическая функция. Для принятия решения, к какому классу отнести некоторый объект или ситуацию, требуется ответить на вопросы, стоящие в узлах этого дерева, начиная с его корня. В зависимости от полученного значения функции (ответа на вопрос) происходит переход далее по дереву в левую или правую вершину следующего уровня; затем снова следует вопрос, связанный с соответствующим узлом.

Каждая висячая вершина связана с одним из классов, к которому и относится распознаваемый объект, если путь по дереву заканчивается в данной вершине.

Бинарное дерево называется решающим, если выполнены следующие условия:

– каждая внутренняя вершина помечена признаковым предикатом;

– выходящие из вершин дуги помечены значениями, принимаемыми предикатами в вершине;

– концевые вершины помечены метками классов;

– ни в одной ветви дерева нет двух одинаковых вершин [11, 12].

Для решения задачи бинарной классификации с помощью дерева решений необходимо решить, какие вопросы нужно указывать в узлах дерева и в каком порядке. На каждом этапе продвижения по ветвям дерева решений существуют некоторые возможные результаты (отнесение экземпляра к тому или иному классу), которые мы исключили и которые нет. Каждый возможный вопрос разбивает оставшиеся ре-

зультаты в соответствии с их ответами. Для построения наиболее точного классификатора необходимо выбрать вопросы, ответы на которые дают большой прирост информации о том, что должно предсказать дерево. Например, вопрос, положительный ответ на который всегда соответствует результату 1, а отрицательный ответ результату 0 (или наоборот), дает максимальный прирост информации, а вопрос, для которого ни один из ответов не дает много новой информации о том, каким должен быть прогноз, не является хорошим выбором.

Таким образом, каждый этап дерева решений включает в себя задание вопроса, ответ на который разбивает данные на одно или несколько подмножеств [13].

В настоящей работе для оценки качества классификации, выполненной с помощью дерева решений, использовалась матрица неточностей, в которой показаны пары {*предсказанное значение; реальное значение*}. В случае бинарной классификации существует четыре возможных результата:

1) если экземпляр является положительным и классифицируется как положительный, он считается истинно положительным;

2) если экземпляр является положительным и классифицирован как отрицательный, он считается ложноотрицательным;

3) если экземпляр отрицательный и классифицируется как отрицательный, он считается истинно отрицательным;

4) если экземпляр отрицательный и классифицирован как положительный, он считается ложноположительным.

Общий вид такой матрицы представлен на рис. 1.

		Предсказанное значение	
		1 (positive)	0 (negative)
Реальное значение	1	True Positive (TP)	False Positive (FP)
	0	False Negative (FN)	True Negative (TN)

Рис. 1. Общий вид матрицы неточностей

На рис. 1 TP – количество верно классифицированных положительных классов (истинно положительное решение); FN – количество неверно классифицированных отрицательных классов (ложноотрицательное решение); FP – количество неверно классифицированных по-

ложительных классов (ложноположительное решение); TN – количество верно классифицированных отрицательных классов (истинно отрицательное решение). Числа вдоль главной диагонали представляют собой правильные решения, а числа вне этой диагонали – ошибки I и II рода [13].

Для анализа полученных матриц неточности будем использовать такие числовые критерии, как точность и полнота. Точность (precision) – способность классификатора не помечать как положительный образец, который является отрицательным, т.е. не совершать ошибки II рода, определяется формулой precision = TP/(TP + FN). Полнота (recall) – способность классификатора находить все положительные выборки, определяется формулой recall = TP/(TP + FP). Precision и recall могут принимать значения от 0 до 1, включая границы, при этом лучшее значение равняется 1, а худшее 0.

Кроме того, воспользуемся кривой ошибок (ROC). Графики ROC – это двумерные графики, на которых по оси *Y* отображается чувствительность алгоритма классификации TPR = TP/(TP + FN) = Recall, а по оси *X* величина FPR = FP/(FP + TN). Таким образом, кривая ошибок отображает соотношение между долей объектов от общего количества носителей признака, верно классифицированных как несущих признак (истинно положительные результаты), и долей объектов от общего количества объектов, не несущих признака, ошибочно классифицированных как несущих признак (ложноположительные результаты) при варьировании порога решающего правила.

Кривая ошибок – это двумерное изображение производительности классификатора. Для сравнения классификаторов производится «сворачивание» производительности ROC до единого скалярного значения, представляющего ожидаемую производительность. Распространенным методом является вычисление площади под ROC-кривой, сокращенно AUC. Поскольку AUC является частью площади единичного квадрата, его значение всегда будет между 0 и 1. Однако, поскольку случайное угадывание создает диагональную линию, описываемую уравнением $y = x$, между точками с координатами (0; 0) и (1; 1), которая имеет площадь 0,5, ни один реалистичный классификатор не должен иметь AUC меньше 0,5 [14–16].

2. Исходные данные и полученные модели

Настоящее исследование проведено в рамках направления «Прикладная математика и информатика» (ПМиИ) механико-математического факультета Пермского государственного национального исследовательского университета (ПГНИУ). Проверка гипотезы о наличии положительной зависимости между успеваемостью абитуриентов в школе при сдаче ЕГЭ, уровнем среднего образования, местом проживания во время учебы в вузе и их успехами в качестве первокурсников осуществлялась на выборке данных объемом в 461 запись о студентах первого курса направления ПМиИ наборов 2014–2018 гг.

Данное направление характеризуется высокой долей отчисленных студентов по итогам первого триместра обучения, количество и доля которых представлены в табл. 1.

Таблица 1

Количество отчисленных студентов, поступивших в 2014–2018 гг.

Год поступления	Количество поступивших студентов	Количество отчисленных студентов	Доля отчисленных студентов, %
2014	99	23	23,23
2015	86	27	31,39
2016	89	27	30,33
2017	86	17	19,77
2018	101	17	16,83

В течение всего исследуемого периода в учебный план первого триместра продолжительностью 17 недель (сентябрь–декабрь), входили следующие дисциплины: «Математический анализ I», «Алгебра и аналитическая геометрия», «Алгоритмизация и программирование I», «История», «Русский язык и риторика». Именно неуспеваемость по первым трем из них, составляющим основу профессиональной подготовки, являлась основной причиной отчисления студентов уже по итогам первого триместра.

Для построения дерева решений были выбраны и закодированы следующие семь признаков:

– Math – количество баллов ЕГЭ по математике;

– Russian – количество баллов ЕГЭ по русскому языку;

– CS – количество баллов ЕГЭ по информатике;

– City – город учебного довузовского заведения (0 – Пермь, 1 – другой);

– Language – иностранный язык, изучаемый в школе (0 – английский, 1 – другой);

– Home – место проживания (0 – дома с родителями, 1 – общежитие, 2 – съемная квартира или квартира родственников);

– Shcool – учебное заведение (0 – СОШ и другие, 1 – лицей/гимназия/специализированная школа с физико-математическим уклоном).

Целевой переменной был бинарный признак GroupRisk – результаты обучения (0 – студент не попал в группу риска, 1 – студент попал в группу риска). На этапе обучения дерева решений принадлежность к группе риска определялась непосредственно по факту отчисления студента.

При построении дерева решений обучающая выборка (train) включала в себя 360 записей (данные 2014–2017 гг.), тестовая (test) – 101 запись (данные 2018 г.) (см. табл. 1).

Были рассмотрены четыре модели дерева решений:

– с ограничением максимальной глубины выборки (DT.1);

– с ограничением максимальной глубины и балансировкой классов (DT.2);

– с ограничением максимальной глубины, минимального количества экземпляров в листе и балансировкой классов (DT.3);

– модель «случайного леса» (DT.4).

Для выбора оптимальной глубины дерева решений модели DT.1 были обучены и протестированы деревья решений с глубиной от 3 до 20. Критерием расщепления выбиралась энтропия: на каждом этапе построения узла дерева выбирался признак и порог, дающие максимальный прирост информации после расщепления и минимальную энтропию. На основе графиков ROC для каждого из вариантов была выполнена оценка площади под кривыми ошибок AUC. Лучший результат был достигнут при глубине дерева 5.

В связи с описанной выше особенностью набора данных построены классификаторы с автоматическим подбором весов, т.е. балансировкой (модель DT.2). В «сбалансированном» режиме значения целевой переменной используются для автоматической регулировки весов, обратно пропорциональных частотам классов во входных данных, как показано в формуле

$$\text{weight} = \frac{n_{\text{samples}}}{n_{\text{classes}} \cdot \text{bincount}},$$

где n_{samples} – количество экземпляров; n_{classes} – количество классов; bincount – количество вхождений экземпляров каждого класса.

Для модели DT.2 оптимальной глубиной дерева стала 6.

Во избежание переобучения классификатора было использовано ограничение на минимальное количество экземпляров классов в листе дерева (min_samples_leaf). Для того чтобы выбрать оптимальное значение этого параметра, были обучены и протестированы деревья решений со значениями min_samples_leaf от 1 до 20, критерием расщепления «энтропия» и балансировкой классов. На основе сравнения AUC для этих вариантов выбрано min_samples_leaf = 3 (DT.3).

Модель «случайного леса» (DT.4) оптимизировалась аналогичным образом, при этом в ней варьируемым параметром было число деревьев. Для наглядности параметры всех оптимизированных моделей представлены в табл. 2.

Некоторые сравнительные характеристики всех использованных моделей приведены в табл. 3, откуда видно, что наиболее высокую оценку точности имеет модель дерева решений с балансировкой, максимальной глубиной 6, минимальным количеством элементов в листе 3.

Таблица 2

Параметры моделей классификации (*авторские результаты*)

Модель	Количество деревьев	Максимальная глубина дерева	Балансировка	min_samples_leaf
DT.1	1	5	Нет	1
DT.2	1	6	Да	1
DT.3	1	6	Да	3
DT.4	11	4	Да	1

Таблица 3

Сравнение результатов моделирования (*авторские результаты*)

Модель	AUC (train)	AUC (test)	precision	Recall	Матрица неточностей
DT.1	0,82	0,76	0,60	0,18	$\begin{pmatrix} 3 & 14 \\ 2 & 82 \end{pmatrix}$
DT.2	0,87	0,84	0,33	0,94	$\begin{pmatrix} 16 & 1 \\ 32 & 52 \end{pmatrix}$
DT.3	0,88	0,85	0,36	0,94	$\begin{pmatrix} 16 & 1 \\ 28 & 56 \end{pmatrix}$
DT.4	0,85	0,83	0,44	0,71	$\begin{pmatrix} 12 & 5 \\ 15 & 69 \end{pmatrix}$

Из табл. 3 видно, что модели DT.2 и DT.3 правильно предсказывают отчисление 16 студентов из 17 (ошибка I рода только 1), зато у модели DT.3 меньше ошибка II рода, в группу риска попадают 44 человека вместо 48.

Для модели DT.3 получена значимость независимых переменных. Соответствующая гистограмма значимости представлена на рис. 2, откуда видно, что наиболее значимыми признаками являются результаты ЕГЭ, менее значимыми учебное заведение (школа), место проживания и город. Незначимым оказался признак иностранного языка, изучаемого в школе.

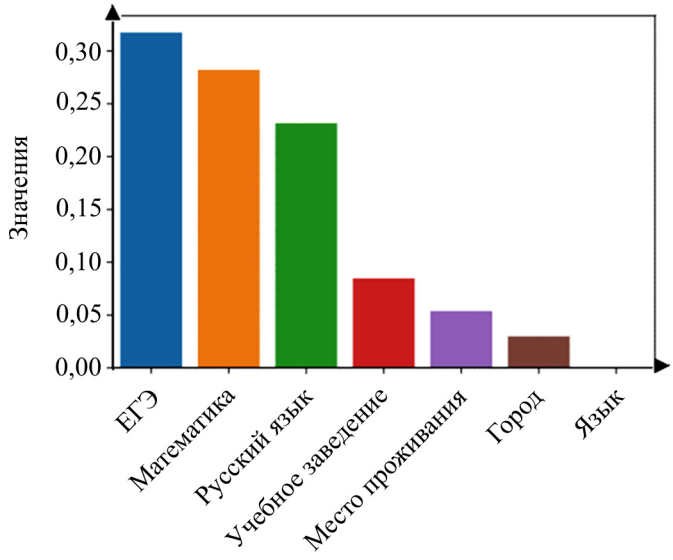

Рис. 2. Значимость признаков модели классификации
(*авторский результат*)

Проходя по цепочке логического вывода модели DT.3, можно отнести очередного студента с некоторой долей вероятности к определенному классу. Поскольку в дереве решений присутствуют листья с нулевой энтропией, можно вывести следующие решающие правила:

– студент с вероятностью 100 % попадает в группу риска, если набранное количество баллов ЕГЭ по информатике менее 51;

– студент, поступивший в университет из другого населенного пункта, с вероятностью 100 % попадает в группу риска, если набран-

ное количество баллов ЕГЭ по информатике менее 75, по математике менее 61, по русскому языку менее 96;

– студент с вероятностью 100 % попадает в группу риска, если набранное количество баллов ЕГЭ по математике менее 70 и по русскому языку менее 68, несмотря на высокие баллы ЕГЭ по информатике.

Эти правила можно интерпретировать как «портрет студента», с вероятностью 100 % попадающего в группу риска. Таким образом, мы можем констатировать, что если в целом по всему потоку набора 2018 г. вероятность быть отчисленным по итогам уже первого триместра составляла около 17 %, то для группы риска эта вероятность уже составляет 36,4 %.

Заключение

Проведенное исследование показало хорошие прогностические свойства «сбалансированных» деревьев решений. Информация о студентах, попавших в группу риска, может помочь тьюторам и преподавателям, ведущим у них занятия с акцентом внимания на потенциально «слабых» студентах. К сожалению, группа риска получается со значительным избытком, что несколько снижает возможности практического применения результатов. Очевидно, что привлечение информации о лично-психологических особенностях студентов (способность к быстрой адаптации, стрессоустойчивость и т.п.), которые можно определить с помощью психологического тестирования, может повысить качество прогнозирования. Причем это дополнительное обследование можно в первую очередь рекомендовать студентам, попавшим в группу риска по итогам классификации с помощью деревьев решений, методика которой описана в настоящей работе.

Список литературы

1. Хавенсон Т.Е., Соловьева А.А. Связь результатов Единого государственного экзамена и успеваемости в вузе // Вопросы образования. – 2014. – № 1. – С. 186–199.

2. Попова Е.А., Шеина М.В. Успеваемость студентов: влияние школы // Современный университет между глобальными вызовами и локальными задачами: сборник материалов, г. Москва, 20–22 октября 2016 г. / под ред. Д.В. Козлова, Н.Г. Малошонок; Национальный исследовательский университет «Высшая школа экономики», Институт образования. – М., 2016. – С. 183–187.

3. Татусь К.Ю., Кузьмина С.В. Влияние родительской семьи на успеваемость студентов // Молодой ученый. – 2016. – № 9.4 (113.4). – С. 69–72.

4. Богданов Е.П., Суханов А.В. О прогнозировании успеваемости студентов по результатам ЕГЭ и атрибутам социального статуса // Актуальные направления научных исследований XXI века: теория и практика. – 2015. – Т. 3, № 7-3 (18-3). – С. 382–386.

5. Кузнецов А.Г., Русаков С.В., Жданова С.Ю. Особенности работы со студентами первого курса (из опыта работы механико-математического факультета Пермского государственного национального исследовательского университета) // Вестник Московского университета. Сер. 20. Педагогическое образование. – 2017. – № 1. – С. 99–110.

6. Русаков С.В., Русакова О.Л., Посохина К.А. Нейросетевая модель прогнозирования группы риска по успеваемости студентов первого курса // Современные информационные технологии и ИТ-образование. – 2018. – Т. 14, № 4. – С. 815–822. DOI: 10.25559/SITITO.14.201804.815-822

7. Шевченко В.А. Прогнозирование успеваемости студентов на основе методов кластерного анализа // Вестник Харьковского национального автомобильно-дорожного университета. – 2015. – Вып. 68. – С. 15–18.

8. Панова Н.Ф., Денисова Н.В. Классификация студентов по уровню успеваемости с помощью аппарата дискриминантного анализа // Вестник Оренбургского государственного университета. – 2014. – № 8 (169). – С. 33–36.

9. Босак С.С. Система оценивания лабораторных работ по дисциплине «Прикладная криптология» на основе моделей нечеткой логики // Донецкие чтения 2016. Образование, наука и вызовы современности: материалы I Междунар. науч. конф., г. Донецк, 16–18 мая 2016 г.: в 6 т. Т. 6. Психологические и педагогические науки / под общ. ред. С.В. Беспаловой; М-во образов. и науки Донец. Народ. Респ., ГОУ ВПО «Донецкий нац. ун-т», Ассоц. юридических вузов России, Междунар. славян. акад. наук, образ., искусств и культуры. – Ростов н/Д: Изд-во Юж. фед. ун-та, 2016 – С. 298–300.

10. Босак С.С. Прогнозирование учебных результатов студентов по курсу «Прикладная криптология» на основе нейронных сетей [Электронный ресурс] // Информатизация образования и методика электронного обучения: I Междунар. науч. конф. в рамках IV Междунар. науч.-образ. форума «Человек, семья и общество: история и перспективы развития», г. Красноярск, 27 сентября–30 сентября 2016 г. / Сибир. фед. ун-т. – Красноярск, 2016. – URL: http://elib.sfu-kras.ru/handle/2311/30621 (дата обращения: 12.10.2020).

11. Classification and regression trees / L. Breiman, J.H. Friedman, R.A. Olshen, C.T. Stone. – Belmont, California: Thomson Wadsworth, 1984. – 368 p.

12. Журавлев Ю.И., Рязанов В.В., Сенько О.В. Распознавание. Математические методы. Программная система. Практические применения. – М.: Фазис, 2006. – 176 с.

13. Michie D., Spiegelhalter D.J., Taylor C.C. Machine learning, neural and statistical classification. – New York: Overseas Press, 2009. – 290 p.

14. Bonnin R. Machine learning for developers. – Packt Publishing, 2017. – 270 p.

15. Fawcett T. ROC Graphs: notes and practical considerations for data mining researchers. – Palo Alto, CA: Hewlett-Packard Company, 2003. – 27 p.

16. Statistical evaluation of diagnostic performance: topics in ROC analysis / K.H. Zou, A. Liu, A.I. Bandos, L. Ohno-Machado, H.E. Rockette. – Chapman and Hall/CRC, 2011. – 245 p.

References

1. Khavenson T.E., Solov'eva A.A. Sviaz' rezul'tatov Edinogo gosudarstvennogo ekzamena i uspevaemosti v vuze [Studying the Relation between the Unified State Exam Points and Higher Education Performance]. *Voprosy obrazovaniia*, 2014, no. 1, pp. 186-199.

2. Popova E.A., Sheina M.V. Uspevaemost' studentov: vliianie shkoly [Student performance: school impact] Sovremennyi universitet mezhdu global'nymi vyzovami i lokal'nymi zadachami: cbornik materialov (Moskva, 20-22 October 2016) Moscow, Higher school of ecoenomics, 2016, pp. 183-187.

3. Tatus' K.Iu., Kuz'mina S.V. Vliianie roditel'skoi sem'i na uspevaemost' studentov [The influence of the parental family on student performance]. *Young scientist*, 2016, no. 9.4 (113.4), pp. 69-72.

4. Bogdanov E.P., Sukhanov A.V. O prognozirovanii uspevaemosti studentov po rezul'tatam EGE i atributam sotsial'nogo statusa [On forecasting the progress of students by the results of the USE and attributes of social status] *Aktual'nye napravleniia nauchnykh issledovanii XXI veka: teoriia i praktika*, 2015, vol. 3, no. 7-3 (18-3), pp. 382-386.

5. Kuznetsov A.G., Rusakov S.V., Zhdanova S.Iu. Osobennosti raboty so studentami pervogo kursa (iz opyta raboty mekhaniko-matematicheskogo fakul'teta Permskogo gosudarstvennogo natsional'nogo issledovatel'skogo universiteta) [Features of work with students of first year]. *Vestnik Moskovskogo universiteta. Ser. 20. Pedagogicheskoe obrazovanie*, 2017, no. 1, pp. 99-110.

6. Rusakov S.V., Rusakova O.L., Posokhina K.A. Neirosetevaia model' prognozirovaniia gruppy riska po uspevaemosti studentov pervogo kursa [Neural network model of predicting the risk group for the accession of students of the first course]. *Modern Information Technologies and IT-Education*, 2018, vol. 14, no. 4, pp. 815-822. DOI: 10.25559/SITITO.14.201804.815-822

7. Shevchenko V.A. Prognozirovanie uspevaemosti studentov na osnove metodov klasternogo analiza [Prognostication of students progress on the basis of

cluster analysis methods]. *Vestnik Khar'kovskogo natsional'nogo avtomobil'no-dorozhnogo universiteta*, 2015, Iss. 68, pp. 15-18.

8. Panova N.F., Denisova N.V. Klassifikatsiia studentov po urovniu uspevaemosti s pomoshch'iu apparata diskriminantnogo analiza [classification of students by academic performance using the discriminant analysis]. *Vestnik of the Orenburg State University*, 2014, no. 8 (169), pp. 33-36.

9. Bosak S.S. Sistema otsenivaniia laboratornykh rabot po distsipline «Prikladnaia kriptologiia» na osnove modelei nechetkoi logiki [Evaluation system for laboratory work in the discipline "Applied Cryptology" based on fuzzy logic models]. Proceedings of the International conference "Donetsk readings 2016. Education, Science and Modern Challenges, Rostov on Don, 16-18 June 2016]. Rostov on Don, South Federal University Publisher, 2016, pp. 298-300.

10. Bosak S.S. Prognozirovanie uchebnykh rezul'tatov studentov po kursu «Prikladnaia kriptologiia» na osnove neironnykh setei [Prediction of student learning results in the course "Applied Cryptology" based on neural networks]. Proceedings of the 1st International conference "Informatization of education and e-learning methodology", 27-30 September 2016. Krasnoyarsk, Siberian Federal university, 2016, available at: http://elib.sfu-kras.ru/handle/2311/30621 (accessed: 12 October 2020).

11. Breiman L., Friedman J.H., Olshen R.A., Stone C.T. Classification and regression trees. Wadsworth, Belmont, California, 1984, 368 p.

12. Zhuravlev Iu.I., Riazanov V.V., Sen'ko O.V. Raspoznavanie. Matematicheskie metody. Programmnaia sistema. Prakticheskie primeneniia [Recognition. Mathematical methods. Software system. Practical applications]. Moscow, Fazis, 2006, 176 p.

13. Michie D., Spiegelhalter D.J., Taylor C.C. Machine learning, neural and statistical classification. New York, Overseas Press, 2009, 290 p.

14. Bonnin R. Machine learning for developers. Packt Publishing, 2017, 270 p.

15. Fawcett T. ROC Graphs: notes and practical considerations for data mining researchers. Palo Alto, CA: Hewlett-Packard Company, 2003, 27 p.

16. Zou K.H., Liu A., Bandos A.I., Ohno-Machado L., Rockette H.E. Statistical evaluation of diagnostic performance: topics in ROC analysis. Chapman and Hall/CRC, 2011, 245 p.

Статья получена: 12.10.2020

Статья принята: 16.11.2020

Сведения об авторах

Накарякова Наталья Николаевна (Пермь, Россия) – магистр кафедры «Прикладная математика и информатика», Пермский государственный

национальный исследовательский университет (614990, Пермь, ул. Букирева 15, e-mail: nata_nakar@mail.ru).

Русаков Сергей Владимирович (Пермь, Россия) – доктор физико-математических наук, профессор, завкафедрой «Прикладная математика и информатика», Пермский государственный национальный исследовательский университет (614990, Пермь, ул. Букирева 15, e-mail: rusakov@psu.ru).

Русакова Ольга Леонидовна (Пермь, Россия) – кандидат физико-математических наук, доцент, доцент кафедры «Прикладная математика и информатика», Пермский государственный национальный исследовательский университет (614990, Пермь, ул. Букирева 15, e-mail: rol58@yandex.ru).

About the authors

Natalia N. Nakaryakova (Perm, Russian Federation) – Master's Student, Department of Applied Mathematics and Informatics, Perm State University (15, Bukireva st., Perm, 614990, e-mail: nata_nakar@mail.ru).

Sergey V. Rusakov (Perm, Russian Federation) – Dr. Habil in Physics and Mathematics, Professor, Head of Department of Applied Mathematics and Informatics, Perm State University (15, Bukireva st., Perm, 614990, e-mail: rusakov@psu.ru).

Olga L. Rusakova (Perm, Russian Federation) – Ph.D. in Physics and Mathematics, Associate Professor, Department of Applied Mathematics and Informatics, Perm State University (15, Bukireva st., Perm, 614990, e-mail: rol58@yandex.ru).

Библиографическое описание статьи согласно ГОСТ Р 7.0.100–2018:

Накарякова, Н.Н. Прогнозирование группы риска (по успеваемости) среди студентов первого курса с помощью дерева решений / Н.Н. Накарякова, С.В. Русаков, О.Л. Русакова. – DOI 10.15593/2499-9873/2020.4.08. – Текст: непосредственный // Прикладная математика и вопросы управления = Applied Mathematics and Control Sciences. – 2020. – № 4. – С. 121–136.

Цитирование статьи в изданиях РИНЦ:

Накарякова Н.Н., Русаков С.В., Русакова О.Л. Прогнозирование группы риска (по успеваемости) среди студентов первого курса с помощью дерева решений // Прикладная математика и вопросы управления. – 2020. – № 4. – С. 121–136. DOI: 10.15593/2499-9873/2020.4.08

Цитирование статьи в references и международных изданиях:

Cite this article as:

Nakaryakova N.N., Rusakov S.V., Rusakova O.L. Prediction of the risk group (by academic performance) among first course students by using decision tree method. *Applied Mathematics and Control Sciences*, 2020, no. 4, pp. 121-136. DOI: 10.15593/2499-9873/2020.4.08 (*in Russian*)