УДК 004.89+004.65
DOI: 10.26102/2310-6018/2023.41.2.003

Data Mining в образовании: прогнозирование успеваемости учащихся

Е.С. Егорова[1], Н.А. Попова[2]✉

[1]Пензенский государственный технологический университет, Пенза, Российская Федерация
[2]Пензенский государственный университет, Пенза, Российская Федерация
popov.tasha@yandex.ru✉

Резюме. Способность прогнозировать академические результаты учащихся имеет ценность для любого учебного заведения, стремящегося улучшить успеваемость и мотивацию студентов. Основываясь на сгенерированных прогнозах, учащимся, выявленным как подверженным риску отчисления или неуспеваемости, может быть оказана поддержка более своевременным образом. В статье рассмотрены различные классификационные модели для прогнозирования успеваемости студентов, используя данные, собранные в университетах г. Пензы. Данные включают сведения о зачислении студентов, а также данные о деятельности, полученные из университетской электронной информационно-образовательной среды (ЭИОС). Важным вкладом этого исследования является учет неоднородности учащихся при построении прогностических моделей. Это основано на наблюдении, что учащиеся с различными социально-демографическими особенностями или способами обучения могут проявлять различную мотивацию к обучению. Эксперименты подтвердили гипотезу о том, что модели, обученные с использованием экземпляров в студенческих подгруппах, превосходят модели, построенные с использованием всех экземпляров данных. Кроме того, эксперименты выявили, что учет особенностей как зачисления, так и учебной деятельности помогает более точно идентифицировать уязвимых учащихся. Результаты экспериментов показали, что ни один отдельный метод не обладает превосходной производительностью во всех аспектах. В качестве инструментально средства для создания прогностической модели использовалась отечественная аналитическая платформа Loginom.

Ключевые слова: Data Mining, интеллектуальный анализ образовательных данных, прогнозирование успеваемости учащихся, неоднородность учащихся, электронная информационно-образовательная среда.

Для цитирования: Егорова Е.С., Попова Н.А. Data Mining в образовании: прогнозирование успеваемости учащихся. *Моделирование, оптимизация и информационные технологии.* 2023;11(2). URL: https://moitvivt.ru/ru/journal/pdf?id=1325 DOI: 10.26102/2310-6018/2023.41.2.003

Data Mining in education: predicting student performance

E.S. Egorova[1], N.A. Popova[2]✉

[1]Penza State Technological University, Penza, the Russian Federation
[2]Penza State University, Penza, the Russian Federation
popov.tasha@yandex.ru✉

Abstract. The ability to predict student academic performance is valuable to any institution seeking to improve student achievement and motivation. Based on the predictions generated, students identified as being at risk for expulsion or failure can be supported in a more timely manner. This article discusses various classification models for predicting student performance using data collected from universities in Penza. The data include student enrollment data as well as activity data from the university electronic information and education environment (EIE). An important contribution of this study is the

consideration for student heterogeneity in the construction of predictive models. This is based on the observation that students with different socio-demographic characteristics or modes of learning may exhibit different motivation to learn. Experiments confirmed the hypothesis that models trained using instances in student subgroups outperform models built using all data instances. In addition, the experiments showed that accounting for both enrollment and learning activity patterns helped to identify vulnerable students more accurately. Experimental results have demonstrated that no single method has superior performance in all aspects. The homegrown analytics platform Loginom was employed as a tool to create a predictive model.

Keywords: Data Mining, intellectual analysis of educational data, forecasting of student progress, heterogeneity of students, electronic information and educational environment.

For citation: Egorova E.S., Popova N.A. Data Mining in education: predicting student performance. *Modeling, Optimization and Information Technology*. 2023;11(2). URL: https://moitvivt.ru/ru/journal/pdf?id=1325 DOI: 10.26102/2310-6018/2023.41.2.003 (In Russ.).

Введение

В настоящее время образовательные учреждения сталкиваются с высокой конкуренцией, поэтому необходимо обеспечить эффективное использование ресурсов для улучшения процесса обучения студентов и поощрение мероприятий по сохранности контингента обучающихся и повышению успеваемости. Задача состоит в том, чтобы провести углубленный анализ академической успеваемости учащихся, который может помочь в разработке стратегии поддержки учащихся и улучшить методы преподавания и усвоения материала. В связи с этим учебные заведения могут быть заинтересованы в понимании факторов, предикторов академической успеваемости студентов. Однако это достаточно сложная задача для решения, поскольку огромное количество дополнительных факторов (экономических, социальных, демографических, культурных) могут влиять на академические результаты. Выявление значимых факторов академической успеваемости учащихся требует углубленного анализа данных. Эта задача может быть решена с помощью интеллектуального анализа образовательных данных (АОД) (Educational Data Mining) – процесса обнаружения знаний посредством предоставления ценной информации на основе данных, полученных в образовательной среде [1].

Один из самых популярных методов интеллектуального анализа данных является классификация, которая успешно применяется для прогнозирования динамики различных процессов. Классификация – это контролируемый процесс организации объектов со сходными характеристиками в классы. Подходы к классификации можно в широком смысле разделить на модели «белого ящика», например, дерево решений, и модели «черного ящика», например, искусственные нейронные сети [2]. Исследования в этой области проводились при сочетании различных подходов и уровней детализации. Например, автор работы [3] разработал основанный на нейросети подход для прогнозирования оценок учащихся. Хотя модели «черного ящика» могут обеспечить более высокую точность прогнозирования, интерпретация результатов для этих моделей является сложным процессом, что замедляет осмысление результатов. Некоторые исследователи [4] предложили различные методы для улучшения интерпретации в нескольких методах черного ящика.

Напротив, модели «белого ящика», такие как деревья решений и подходы, основанные на правилах, отображают знания более понятным образом и могут быть непосредственно использованы для дальнейшего принятия решений. Например, в работе [5] был продемонстрирован метод классификации на основе дерева решений для выявления влияющих факторов, разделяющих академически успешных и неуспешных

студентов. В ряде исследований [6, 7] предпринимались попытки использовать методы «черного ящика» и «белого ящика» для прогнозирования успеваемости учащихся, путем рассмотрения активности студентов в различных образовательных средах.

Авторы работы [8] продемонстрировали, что общеуниверситетские прогностические модели часто не учитывают тонкости в разработке курсов, которые влияют на мотивацию студентов, стратегии обучения и успеваемость. Следовательно, прогнозируемые результаты, полученные на основе глобальной модели, могут оказаться бесполезными для преподавателей и руководителей структурных подразделений, поскольку выявленные факторы, которые, как считается, влияют на успеваемость учащихся, могут значительно различаться в разных подгруппах студентов.

Целью данного исследования является построение моделей прогнозирования академической успеваемости учащихся в различных подгруппах, принимая во внимание демографические характеристики студентов, академические особенности и особенности практической учебной деятельности для идентификации среди обучающихся «группы риска». Модель, построенную в подгруппе, будем называть подмоделью, а модель, построенную со всеми экземплярами данных – базовой моделью.

В работе оценивается эффективность предлагаемого подхода с точки зрения прогностической способности, поэтому применяются два метода классификации «черного ящика»: наивный байесовский классификатор (Naive Bayes) и метод опорных векторов (SVM) [9]. Полученные результаты демонстрируют, что в большинстве случаев эти подмодели превосходят базовую модель.

Методология

В исследовании рассматривается неоднородность различных подгрупп студентов и строятся модели классификации этих подгрупп для прогнозирования академических результатов. На первом этапе наборы данных предварительно обрабатываются. На втором этапе подгруппы студентов генерируются из исходных наборов данных с учетом определенных значимых демографических характеристик и академических особенностей студентов. На третьем этапе к поднабору данных применяются различные методы классификации для создания подмоделей студентов. На последнем этапе подмодели оцениваются с использованием различных показателей удобства полученных результатов в принятии решений.

Подход, примененный в данном исследовании, проиллюстрирован на Рисунке 1. Набор данных для исследования включает социально-демографические (возраст, пол, социально-экономический статус) и академические данные, собранные во время набора студентов в вуз, а также данные об активности, полученные с помощью электронной информационно-образовательной среды (ЭИОС) – Moodle [10].

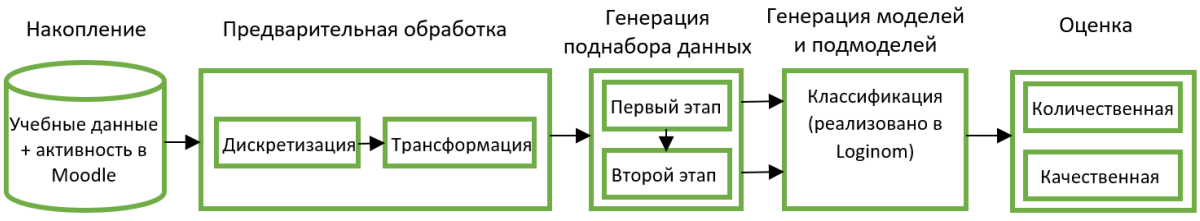

Рисунок 1 – Подход интеллектуального анализа данных для прогнозирования академической успеваемости учащихся

Figure 1 – Data mining approach for predicting student academic performance

Данные, полученные с помощью Moodle, фиксируют участие учащихся в различных мероприятиях (например, заданиях, вики-статьях, форумах и т. д.) и ресурсах (например, книгах и файлах). Каждая запись этого набора данных содержит частоту участия студента в различных мероприятиях определенного курса, поэтому в исходном наборе данных содержится несколько записей для разных курсов, пройденных одним и тем же студентом. В объединенном наборе данных значение параметра активности для студента определяется средним количеством его участия в этом конкретном виде деятельности на всех курсах, которые он посещал в течение учебного года. Значения различных атрибутов академического процесса и деятельности в онлайн-системе поддержки учебного процесса можно найти в Таблице 1.

Таблица 1 – Значение атрибутов, использованных в исследовании
Table 1 – Attribute values used in the study

Тип	Атрибут	Значение
Данные при зачислении	ПОЛ	Пол учащихся
	ОСНОВАНИЕ_ДЛЯ ПОСТУПЛЕНИЯ	Основание для поступления в университет (например, поступление в зрелом возрасте)
	ВОЗРАСТ	Возраст учащихся (если больше 23 лет, то зрелый, в остальном нормальный)
	ФОРМА_ОБУЧЕНИЯ	Форма обучения: очно-заочная или заочная.
	РЕЖИМ_ПОСЕЩАЕМОСТИ	Живущий в городе, в котором находится университет или иногородний
	СОЦИАЛЬНЫЙ-СТАТУС	Социальный статус
	ОБРАЗОВАНИЕ_ПАПЫ	Образовательный статус родителя / опекуна учащегося мужского пола
	ОБРАЗОВАНИЕ_МАМЫ	Образовательный статус женщины-родителя / опекуна учащегося
	ШКОЛА	Школа, которую окончил студент
	БАЛЛ_ЕГЭ	Оценка студента при поступлении в высшее учебное заведение
Активность (деятельность) в ЭИОС	ПРОСМОТР_КНИГ	Просмотр книг и учебных материалов
	ПРОСМОТР_ЗАДАНИЙ	Просмотр учебно-методических материалов
	ПОСЕЩЕНИЕ_ СТРАНИЦЫ_КУРСА	Посещение основной страницы курса
	ОБСУЖДЕНИЕ_ФОРУМ	Добавление обсуждения на форуме курса
	ПОСТ_ФОРУМ	Добавление поста на форум курса
	ПРОСМОТР_ФОРУМ	Просмотр обсуждения на форуме курса
	АКТИВНОСТЬ_ФОРУМ	Просмотр активности на форуме
	ВЫПОЛНЕНИЕ_ЗАДАНИЯ	Количество выполненных заданий
	ТЕСТИРОВАНИЕ	Прохождение тестов по теме
	ФАЙЛЫ_ПРОСМОТР	Просмотр файловых ресурсов

Предварительная обработка данных является важным этапом подготовки данных перед применением методов Data Mining. Предварительная обработка проводится в два этапа, а именно дискретизация и преобразование.

Часть атрибутов, например ПОЛ, ШКОЛА, ОБРАЗОВАНИЕ_МАМЫ представляют категориальные данные и поэтому дискретизированы. Дискретизация выполняется для атрибутов ВОЗРАСТ, СОЦИАЛЬНЫЙ_СТАТУС, БАЛЛ_ЕГЭ, а также для всех атрибутов активности. Все атрибуты активности распределены по четырем квартилям, а именно Q1, Q2, Q3 и Q4, где Q1 представляет наименьшее участие, а Q4 – наибольшее.

При преобразовании данных учитывалось, что классификация выполнялась с помощью платформы Loginom, поэтому данные об академической успеваемости и активности в ЭИОС преобразовывались в формат Excel, поддерживаемый системой.

Наборы данных о зачислении, активности и комбинированные наборы данных разделены на несколько поднаборов для формирования подгрупп учащихся. Разделение набора данных выполняется в два этапа следующим образом:

1. Данные о зачислении, активности и объединенные наборы данных разделяются в соответствии с полом учащегося (мужчины и женщины), возрастом (нормальный и зрелый), форма обучения (очно-заочная или заочная) и режимом посещаемости (местный или иногородний). Следовательно, генерируется восемь поднаборов данных для каждого из наборов данных о зачислении, деятельности и комбинированных наборов данных соответственно.

2. Поднаборы данных для женщин и мужчин дополнительно разделены на еще шесть поднаборов данных в соответствии с возрастом учащихся, формой обучения и режимом посещения. Вложенные наборы данных и их размеры представлены в Таблице 2.

Таблица 2 – Размеры совокупности различных наборов данных
Table 2 – Dimensions of different data set collection

Наборы данных	Вложенные наборы данных	Размер когорты		
		Зачисление	Активность ЭИОС	Комбинированный
Полный тренировочный набор	Всего, из них:	2648	7052	2648
	Мужчины	1986	5211	1986
	Женщины	662	1841	662
	Очно-заочная форма	2160	6123	2160
	Заочная форма	488	929	488
	Местный	2101	5955	2101
	Иногородний	547	1097	547
	Нормальный	1909	5339	1909
	Зрелый	739	1713	739
Мужчины	Очно-заочная форма	1570	4428	1570
	Заочная форма	416	783	416
	Местный	1401	3987	1401
	Иногородний	585	1224	585
	Нормальный	1369	3820	1369
	Зрелый	617	1391	617
Женщины	Очно-заочная форма	590	1695	590
	Заочная форма	72	146	72
	Местный	513	1625	513
	Иногородний	149	216	149
	Нормальный	491	1519	491
	Зрелый	171	322	171

Первоначально для каждого набора данных формируются подгруппы учащихся женского, мужского, нормального возраста, зрелого возраста, обучающихся очно-заочной формы, заочной формы, местный и иногородний, которые будем называть подгруппой первого уровня. После этого мужская и женская подгруппы дополнительно подразделяются на подгруппы нормального возраста, зрелого возраста, очно-заочной формы, заочной формы, подгруппы местных и иногородних студентов, которые относятся к подгруппе второго уровня.

Среди подходов к классификации были выбраны два метода «черного ящика» Байесовский классификатор и метод опорных векторов для генерации подмоделей учащихся.

Наивный байесовский классификатор – это вероятностный метод классификации, основанный на теореме Байеса. Данный метод можно рассматривать как простейший байесовский сетевой классификатор. Этот метод прост в реализации и особенно часто используется с данными большой размерности. Классификатор применяется с дискретными или непрерывными атрибутами.

Метод опорных векторов (SVM) – это метод использует алгоритм оптимизации для обучения машины опорных векторов и относится к типу функциональной классификации, которая работает путем выявления функции. Модели, созданные с помощью этого метода, обычно демонстрируют высокую точность классификации. Этот метод заменяет отсутствующие значения и может обрабатывать многоклассовые задачи, используя попарную классификацию.

Для измерения прогностической способности модели был разработан ряд критериев. В этом исследовании использовались следующие показатели для оценки эффективности различных методов с точки зрения сгенерированных моделей:

1. Точность (T) – доля истинно положительных примеров среди всех примеров, классифицированных классификатором как положительные.

2. Полнота (R) – доля истинно положительных примеров, правильно классифицированных классификатором.

3. F-мера ($F_{\text{мера}}$) – среднее гармоническое значение точности и полноты классификатора:

$$F_{\text{мера}} = \frac{2 \cdot T \cdot R}{T + R}.$$

4. Коэффициент Каппа (k) – сравнивает точность классификатора с точностью, которую, как ожидается, достигнет случайный классификатор. Значение k варьируется в интервале от 0 до 1, где 1 обозначает идеальное предсказание классификатора, а 0 означает не более чем случайное предположение.

5. AUC (Area Under Curve) – площадь под кривой рабочей характеристики приемника (Receiver Operating Characteristic – ROC), указывает на вероятность того, что классификатор оценит случайно выбранный положительный пример более высоко, чем случайно выбранный отрицательный пример. Значение AUC, равное 1, указывает на идеальный классификатор, в то время как 0,5 подразумевает, что классификатор работает как случайные догадки.

Результаты прогнозирования успеваемости учащихся

Для построения и оценки моделей прогнозирования, сгенерированных для подгрупп учащихся, использовали данные о зачислении студентов и активности в ЭИОС как отдельно, так и совместно. Наборы данных были собраны в университетах г. Пензы для студентов, зачисленных на обучение в 2017, 2018 годах и прошедших обучение по программам бакалавриата. В Таблице 3 показаны обучающие и тестирующие наборы данных, использованные в исследовании, для обучения и тестирования характеристик модели соответственно.

Таблица 3 – Сводная база данных
Table 3 – Consolidated database

Набор данных	Количество экземпляров		Атрибуты
	Обучение	Тест	
Данные при зачислении	2648	1362	13
Активность в ЭИОС	7052	3916	14
Комбинированный	2648	1362	27

Далее строим модели прогнозирования успеваемости учащихся в каждой подгруппе, представленные в Таблице 2. Более того, чтобы исследовать эффективность построения моделей в подгруппах, мы также строим базовую модель для всей совокупности и сравниваем показатели подмодели и базовой модели. Эти подмодели создаются для каждого из трех наборов данных: а) данные о зачислении; б) данные о деятельности, полученные из электронной среды Moodle; в) объединенные данные, содержащие как характеристики зачисления, так и активности.

Пример полученных результатов вычисления эффективности студенческих подмоделей с точки зрения выявления учащихся из группы риска с использованием данных о зачислении представлен в Таблице 4.

Таблица 4 – Результаты анализа данных при зачислении
Table 4 – Results of enrolment data mining analysis

Метод	Вложенные наборы данных	Полный набор данных					Женщины					Мужчины				
		T	R	$F_{\text{мера}}$	k	AUC	T	R	$F_{\text{мера}}$	k	AUC	T	R	$F_{\text{мера}}$	k	AUC
Наивный Байес	Исходный набор данных	0,276	0,128	0,18	0,102	0,504	-	-	-	-	-	-	-	-	-	-
	Мужчины	0,411	0,196	0,265	0,138	0,587	-	-	-	-	-	-	-	-	-	-
	Женщины	0,333	0,143	0,2	0,125	0,593	-	-	-	-	-	-	-	-	-	-
	О-З форма	0,495	0,194	0,278	0,167	0,617	0,409	0,153	0,222	0,12	0,605	0,45	0,985	0,435	0,256	0,641
	Заочная форма	0,444	0,353	0,393	0,186	0,681	0,429	0,286	0,343	0,109	0,648	0,318	0,538	0,4	-0,15	0,456
	Местный	0,471	0,25	0,327	0,134	0,349	0,457	0,246	0,319	0,19	0,653	0,39	0,355	0,392	0,21	0,568
	Иногородний	0,425	0,218	0,288	0,126	0,681	0,483	0,259	0,337	0,202	0,678	0,316	0,316	0,316	-0,016	0,547
	Нормальный	0,511	0,263	0,347	0,22	0,65	0,386	0,199	0,262	0,142	0,634	0,465	0,439	0,452	0,213	0,64
	Зрелый	**0,548**	**0,317**	**0,402**	**0,268**	**0,702**	0,458	0,208	0,286	0,116	0,655	**0,49**	**0,495**	**0,492**	**0,27**	**0,706**
Метод опорных векторов	Исходный набор данных	0,322	0,066	0,1	0,075	0,526	-	-	-	-	-	-	-	-	-	-
	Мужчины	0,385	0,122	0,18	0,17	0,576	-	-	-	-	-	-	-	-	-	-
	Женщины	0,385	0,156	0,222	0,146	0,569	-	-	-	-	-	-	-	-	-	-
	О-З форма	0,5	0,136	0,21	0,113	0,537	0,444	0,023	0,043	0,023	0,508	0,6	0,438	0,509	0,38	0,625
	Заочная форма	0,463	0,412	0,436	0,24	0,648	0,429	0,214	0,286	0,086	0,537	0,389	0,538	0,452	0,0137	0,507
	Местный	0,569	0,123	0,2	0,125	0,546	0,395	0,099	0,159	0,075	0,528	0,5	0,568	0,532	0,31	0,661
	Иногородний	0,42	0,13	0,2	0,18	0,566	0,333	0,019	0,035	0,009	0,503	0,267	0,211	0,235	-0,109	0,448
	Нормальный	0,385	0,156	0,22	0,14	0,584	0,333	0,068	0,113	0,047	0,516	0,486	0,486	0,486	0,26	0,626
	Зрелый	**0,508**	**0,513**	**0,51**	**0,28**	**0,751**	0,5	0,19	0,036	0,014	0,505	**0,65**	**0,538**	**0,59**	**0,48**	**0,676**

Установлено, что большинство подмоделей достигают результатов, превосходящих базовую модель, с точки зрения критериев оценки эффективности моделей. Подмодель, представляющая студентов зрелого возраста, лучше всего выявляет неуспевающих студентов. Из двух представленных методов SVM достигает наилучших результатов при построении этой подмодели, при этом значения $F_{\text{мера}}$ и k составляют 51 % и 28 % соответственно. Этот метод также обеспечивает более высокую

AUC в 75,1 % для внешней подмодели. Производительность подмоделей второго уровня можно увидеть в двух последних столбцах, женских и мужских.

Далее проводится интеллектуальный анализ данных о деятельности в ЭИОС. Результаты показывают, что все подмодели достигают результатов, превосходящих базовую модель во всех аспектах. Среди них внешняя подмодель лучше выявляет учащихся из группы риска, при этом значения $F_{\text{мера}}$, k и AUC превышают 66 %, 53 % и 78 % соответственно. Модель SVM достигают наивысшей точности в 79,1 % для подмодели «Женщина-иногородняя», что также приводит к более высокому показателю $F_{\text{мера}}$ на 68,7 %.

Тем не менее, подмодель «Мужчина-зрелый» обладает более высокой производительностью по сравнению со своей базовой моделью («Мужчины»). Подмодель «Мужчина-иногородний» достигает наилучшего результата среди всех подмоделей, созданных на основе студентов мужского и женского пола. Эксперименты также демонстрируют, что SVM работает лучше всего среди всех методов при создании этой подмодели с показателем $F_{\text{мера}}$72 % и значением k 59 %, в то время как значение AUC, достигнутое этой подмоделью, составляет 88 %.

На завершающем этапе осуществлялось прогнозирование успеваемости учащихся с использованием комбинированных данных. Эксперименты, проведенные на объединенном наборе данных, превосходят эксперименты базовой модели. Внешняя подмодель обеспечивает наилучший результат прогнозирования по сравнению с другими подмоделями, на что указывает точность выше 80 % для различных методов. Наивный байесовский классификатор обеспечивает более высокую точность и полноту, составляющие 86 % и 67,8 %, соответственно, для внешней подмодели, что, следовательно, приводит к самому высокому значению $F_{\text{мера}}$ в 76 % среди всех подмоделей. Кроме того, этот метод обеспечивает высочайшую эффективность с точки зрения значений k и AUC – 61 % и 89 % соответственно.

Конкретные женские и мужские подмодели превосходят базовую модель (женщина / мужчина). Эксперименты демонстрируют, что подмодели «Женщина-иногородняя», «Женщина-заочное обучение» и «Женщина-зрелая» достигают лучших результатов по сравнению с базовой женской моделью. Также замечено, что подмодель «Мужчина-иногородний» демонстрирует более высокие показатели в выявлении неуспешных студентов, чем базовая мужская модель. Эксперименты демонстрируют, что эта подмодель набирает наивысший балл как по F-мере, так и по коэффициенту Каппа.

Таким образом, знания, полученные на основе моделей прогнозирования, могут быть полезны учебным заведениям и преподавателям курсов для выявления студентов, подверженных риску академической неуспеваемости на ранних этапах обучения, с тем чтобы можно было своевременно внедрять стратегии упреждающей поддержки.

Это исследование демонстрирует, что результаты, полученные на основе подмоделей, обеспечивают более высокую степень точности, чем базовая модель. Подмодели: «Мужчина-заочная форма», «Мужчина-иногородний» и «Мужчина-зрелый» демонстрируют лучшую результативность в выявлении людей с низкими достижениями как по F-показателю, так и по значениям Каппа. Аналогичные результаты наблюдаются в экспериментах по подгруппам второго уровня. Студентки, которым прогнозируется неуспешность, либо учатся на заочной форме, либо в зрелом возрасте. Результаты показывают, что студенты мужского пола, которые либо учатся на заочной форме, либо иногородние, по прогнозам, имеют более высокий уровень риска академической неуспеваемости.

Сравнение эффективности различных методов

Результаты исследования показали, что ни один отдельный метод не обладает превосходной производительностью с точки зрения анализа различных наборов данных. Для экспериментов с подгруппой первого уровня SVM достигает лучших результатов при интеллектуальном анализе набора данных о зачислении, в то время как метод наивного Байеса лучше всего справляется с интеллектуальным анализом активности и комбинированных данных соответственно.

На Рисунке 2 представлена диаграмма оценки методов с точки зрения правильно классифицированных студентов (как для успешных, так и для неуспешных студентов) внешней подмодели для набора данных, активности и комбинированных наборов данных.

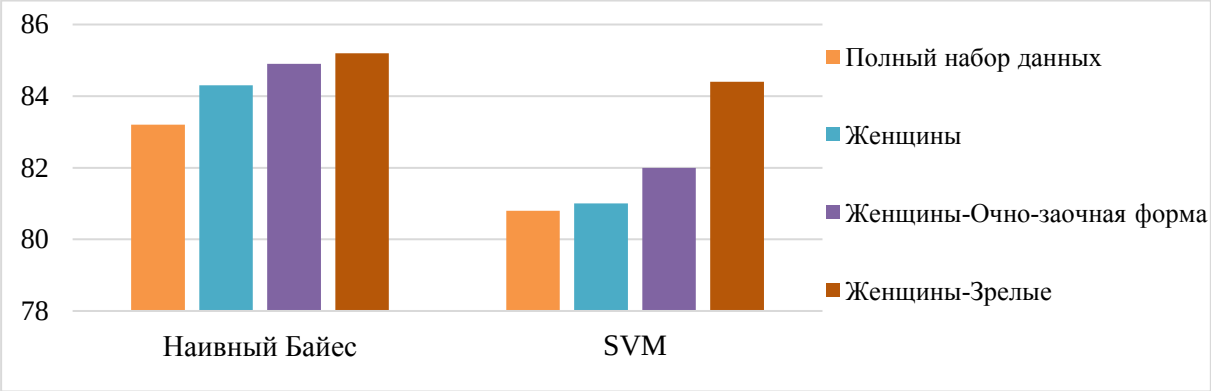

Рисунок 2 – Эффективность различных наборов данных для внешней подмодели
Figure 2 – Efficiency of different datasets for external sub-model

Полученные результаты демонстрируют, что подмодели «Женщина-очно-заочная форма» или «Женщина-зрелая» достигают точности более 60 % для различных методов классификации при рассмотрении только характеристик активности. Эта доля увеличивается до более чем 80 %, если учитывать набор студентов, а также особенности деятельности.

Кроме того, эксперимент с набором данных активности ЭИОС показал, что подмодель «Мужчина-иногородний» достигает точности примерно в 70 %, в то время как при рассмотрении комбинированных данных эта величина увеличивается до 80 %. Модель, сгенерированная на основе комбинированного набора данных, также обеспечивает наилучший результат для правильной классификации как успешных, так и неуспешных студентов. На Рисунке 2 видно, что для каждого метода подмодели, сгенерированные на основе комбинированного набора данных, достигают наилучшего результата прогнозирования.

Заключение

В статье предложена концепция использования неоднородности для получения улучшенных моделей прогнозирования. Результаты эксперимента продемонстрировали эффективность использования студенческих подгрупп для прогнозирования академической успеваемости учащихся. Было показано, что сгенерированные подмодели превосходят базовую модель. Эксперимент выявил, что подмодель, сгенерированная из подгруппы внешних студентов, достигает наилучших результатов.

Кроме того, было продемонстрировано, что полезно исследовать подгруппы второго уровня. Например, эксперименты показали, что подмодели «Женщина-зрелая», «Женщина-иногородняя» и «Женщина-заочная форма» достигают более высоких

результатов прогнозирования, чем женская модель (подмодель первого уровня). Более того, подмодель «Мужчина-иногородний» превосходит мужскую модель. Эти результаты указывают на то, что, хотя не все подмодели второго уровня обеспечивают наиболее точные прогнозы, некоторые из них все же могут дать представление об успеваемости учащихся и, таким образом, помочь в разработке более адресной поддержки учащихся. Кроме того, при рассмотрении комбинированных признаков достигается лучший результат прогнозирования при выявлении неуспешных учащихся по сравнению с рассмотрением признаков по отдельности.

Что касается активности студентов при работе в ЭИОС, было выявлено, что учащиеся с меньшим участием в вики-заданиях с совместным доступом или более низкой частотой просмотра книжных или файловых ресурсов в основном терпят неудачу. Кроме того, обнаружено, что студенты с плохой академической подготовкой принадлежат к более низкому социальному статусу (например, социально-экономический статус или образование родителей) или студенты на заочной форме обучения часто ограничены во времени и не реализуют свой академический потенциал.

Изучая важные социально-демографические и академические факторы, учебное заведение и преподаватели в нем могут выявлять учащихся из группы риска на ранней стадии (до начала их курса) и предпринимать необходимые шаги для поддержки учащихся, проявляющих эти особенности, такие как мониторинг их прогресса путем проведения обычной оценки учебы на протяжении всего учебного года. Более того, учебное заведение может предоставить дополнительную академическую поддержку, например, формируя небольшие группы из таких студентов, чтобы позволить им посещать несколько дополнительных занятий наряду с текущими занятиями по определенной теме. Изучая влияющие факторы, учебное заведение может выявить уязвимых студентов, обладающих специфическими социально-демографическими или академическими особенностями, и посоветовать преподавателям курса следить за их прогрессом на курсе.

СПИСОК ИСТОЧНИКОВ

1. Политов А.Ю., Акжигитов Р.Р., Судариков К.А. Анализ моделей и инструментов предиктивной аналитики для анализа образовательных данных. *Инновации. Наука. Образование.* 2021;28:1055–1065.
2. Salloum S.A., Elnagar A., Shaalan K., Alshurideh M. Mining in Educational Data: Review and Future Directions. *Advances in Intelligent Systems and Computing*. 2020;1153:92–102. DOI: 10.1007/978-3-030-44289-7_9.
3. Пискунов Л.А. Анализ и прогнозирование успеваемости студентов на основе нейронной сети. *Вестник науки*. 2019;3(6):402–405.
4. Васильева Е.Е., Курушин Д.С., Власов С.С. Раннее прогнозирование среднего балла диплома студентов университета: нейросетевой подход. *Международная конференция по мягким вычислениям и измерениям*. 2019;1:366–369.
5. Озерова Г.П., Павленко Г.Ф. Прогнозирование успешности студентов при смешанном обучении с использованием данных учебной аналитики. *Science for Education Today*. 2019;9(6):73–87. DOI: 10.15293/2658-6762.1906.05.
6. Певченко С.С., Блужин В.А. Сравнительный анализ алгоритмов нейронной сети и деревьев принятия решений модели интеллектуального анализа данных. *Молодой ученый*. 2016;132(28):148–154.
7. Шухман А.Е., Парфенов Д.И., Легашев Л.В., Гришина Л.С. Анализ и прогнозирование успеваемости обучающихся при использовании цифровой

образовательной среды. *Высшее образование в России*. 2021;30(8-9):125–133. DOI: 10.31992/0869-3617-2021-30-8-9-125-133.
8. Gasevic D., Dawson S., Rogers T., Gasevic D. Learning analytics should not promote one size fits all: The effects of instructional conditions in predicting academic success. *The Internet and Higher Education*. 2016;28:68–84. DOI: 10.1016/j.iheduc.2015.10.002.
9. Акбархужаев С.А., Абдурахманова Н.Н. Сравнительный анализ методов Наивного Байеса и SVM алгоритмов при классификации текстовых документов. *Молодой ученый*. 2019;267(29):8–10.
10. Булычева П.А. Ошмарина О.Е., Шадрина Е.В. Выявление академически неуспешных студентов на первом году обучения в университете на примере НИУ ВШЭ – Нижний Новгород. *Вестник Нижегородского университета им. Н.И. Лобачевского. Серия: Социальные науки*. 2016;42(2):136–143.

REFERENCES

1. Politov A.Yu., Akzhigitov R.R., Sudarikov K.A. Analiz modelei i instrumentov prediktivnoi analitiki dlya analiza obrazovatel'nykh dannykh. *Innovatsii. Nauka. Obrazovanie.* 2021;28:1055–1065. (In Russ.).
2. Salloum S.A., Elnagar A., Shaalan K., Alshurideh M. Mining in Educational Data: Review and Future Directions. *Advances in Intelligent Systems and Computing*. 2020;1153:92–102. DOI: 10.1007/978-3-030-44289-7_9.
3. Piskunov L.A. Analiz i prognozirovanie uspevaemosti studentov na osnove neironnoi seti. *Vestnik nauki*. 2019;3(6):402–405. (In Russ.).
4. Vasil'eva E.E., Kurushin D.S., Vlasov S.S. Early prediction of the grade point average of university students diploma: neural network approach. *International Conference on Soft Computing and Measurements* . 2019;1:366–369. (In Russ.).
5. Ozerova G.P., Pavlenko G.F. Prediction of student performance in blended learning utilizing learning analytics data. *Science for Education Today.* 2019;9(6):73–87. DOI: 10.15293/2658-6762.1906.05. (In Russ.).
6. Pevchenko S.S., Bluzhin V.A. Sravnitel'nyi analiz algoritmov neironnoi seti i derev'ev prinyatiya reshenii modeli intellektual'nogo analiza dannykh. *Molodoi uchenyi = Young Scientist*. 2016;132(28):148–154. (In Russ.).
7. Shukhman A.E., Parfenov D.I., Legashev L.V., Grishina L.S. Analysis and forecasting students' academic performance using a digital educational environment. *Vysshee obrazovanie v Rossii = Higher Education in Russia.* 2021;30(8-9):125–133. DOI: 10.31992/0869-3617-2021-30-8-9-125-133. (In Russ.).
8. Gasevic D., Dawson S., Rogers T., Gasevic D. Learning analytics should not promote one size fits all: The effects of instructional conditions in predicting academic success. *The Internet and Higher Education*. 2016;28:68–84. DOI: 10.1016/j.iheduc.2015.10.002.
9. Akbarkhuzhaev S.A., Abdurakhmanova N.N. Sravnitel'nyi analiz metodov Naivnogo Baiesa i SVM algoritmov pri klassifikatsii tekstovykh dokumentov. *Molodoi uchenyi = Young Scientist.* 2019;267(29):8–10. (In Russ.).
10. Bulycheva P.A. Oshmarina O.E., Shadrina E.V. Identifying academically "unsuccessful" first-year university students: a case study of Higher School of Economics – Nizhni Novgorod. *Vestnik Nizhegorodskogo universiteta im. N.I. Lobachevskogo. Seriya: Sotsial'nye nauki = Vestnik of Lobachevsky State University of Nizhni Novgorod. Series: Social Sciences.* 2016;42(2):136–143. (In Russ.).

ИНФОРМАЦИЯ ОБ АВТОРАХ / INFORMATION ABOUT THE AUTHORS

Егорова Екатерина Сергеевна, кандидат экономических наук, доцент кафедры «Прикладная информатика», Пензенский государственный технологический университет, Пенза, Российская Федерация.
e-mail: katepost@yandex.ru
ORCID: 0000-0002-0816-0944

Ekaterina Sergeevna Egorova, Candidate of Economic Sciences, Associate Professor at the Department of Applied Informatics, Penza State Technological University, Penza, the Russian Federation.

Попова Наталия Александровна, кандидат технических наук, доцент кафедры «Математическое обеспечение и применение ЭВМ», Пензенский государственный университет, Пенза, Российская Федерация.
e-mail: popov.tasha@yandex.ru
ORCID: 0000-0001-9713-4897

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

Статья поступила в редакцию 02.03.2023; одобрена после рецензирования 03.04.2023; принята к публикации 18.04.2023.

The article was submitted 02.03.2023; approved after reviewing 03.04.2023; accepted for publication 18.04.2023.