Предсказание академической успеваемости студентов на основе индивидуальных когнитивных различий

Е. Е. Котова[1], А.С. Писарев[2]
Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В. И. Ульянова (Ленина)
[1]eekotova@gmail.com, [2]a_pisarev@mail.ru

***Аннотация*. Разработаны модели с применением методов деревьев решений и множественного регрессионного анализа для предсказания итоговой академической успеваемости студентов за весь период обучения (четыре года) в бакалавриате. В моделях учитываются индивидуальные показатели когнитивного потенциала учащихся, идентифицируемые в начале первого семестра с использованием модифицированных компьютерных методик диагностики когнитивных стилей. Апробация моделей проводилась на наборе данных, полученном в процессе обучения (StudCSP-DataSet-23). Полученные результаты подтверждают научную и практическую значимость проведенных исследований в развитии технологий обучения, учитывающих индивидуальные когнитивные различия студентов в обработке информации.**

Ключевые слова: интеллектуальный анализ; образовательный процесс; классификация; академическая успеваемость; когнитивный потенциал; индивидуальные различия

I. Введение

Открытые образовательные ресурсы, курсы обучения, библиотеки образовательного контента, предоставляющие свободный доступ и распространение учебных материалов, привели к масштабному внедрению цифровых технологий в учебный процесс и ускорению развития исследований в области образования.

Одно из направлений относится к адаптивным системам обучения, основанным на моделировании учащегося с применением математических моделей и методов анализа данных, которые позволяют оценить восприятие информации, понимание изучаемого предмета, уровень компетентности, прогресс в обучении, мотивацию, метапознание и другие навыки обучающихся.

Сбор образовательных данных и аналитика обучения становятся все более важными для понимания, как учащиеся учатся, и для прогнозирования академической успеваемости. По мере того, как образовательные учреждения собирают массивы данных об учащихся, от демографических данных до результатов учебной деятельности, возрастает роль автоматизации обработки и интеллектуального анализа, которые помогают дать представление об учебных процессах с тем, чтобы корректировать и адаптировать программы обучения в соответствии с потребностями общества и отдельных учащихся. Это имеет большое значение для оптимизации практики преподавания, разработки систем адаптивного обучения и выявления учащихся, которым может потребоваться дополнительная поддержка [1].

Существует несколько методов моделирования учащихся и прогресса в обучении. Эти модели основаны на данных, собранных посредством явных или неявных взаимодействий учащегося с системой обучения.

Среди широко признанных моделей особенно выделяется направление, основанное на способности отслеживать знания обучающихся в системах адаптивного обучения, которые начинали свое развитие с идей моделирования пользователя в системах адаптивных гипермедиа (Adaptive Hypermedia Systems, AHS) [2 и др.].

II. Постановка задачи прогнозирования результатов обучения

Возможность делать точные прогнозы для отдельных учащихся обеспечивает персонализированный образовательный процесс. К наиболее эффективным для прогнозирования результатов обучения на индивидуальном уровне относятся модели, использующие академические, поведенческие, когнитивные и контекстуальные данные.

Прогнозирование долгосрочной академической успеваемости, например, общего среднего балла студента (Grade Point Average, GPA) и (Cumulative Grade Point Average, CGPA) за весь период обучения в бакалавриате, является важной и сложной задачей интеллектуального анализа данных, которая продемонстрирована в отдельных исследованиях [3, 4]. В отличие от краткосрочных прогнозов успеваемости по одному курсу, долгосрочный прогноз среднего балла за весь период обучения требует моделирования траекторий учащихся на протяжении нескольких лет. Это может дать полезную информацию о факторах, влияющих на общий успех в учебе, и позволит на раннем этапе выявить учащихся группы риска.

Для долгосрочного прогнозирования среднего балла применялся ряд подходов на основе статистики и машинного обучения. В ранних исследованиях использовались модели линейной регрессии с информацией о школьном образовании, такой, как баллы ЕГЭ, и стандартизированные результаты тестов (Scholastic Assessment Test в США, SAT) в качестве предикторов. Эти модели учитывают небольшую дисперсию среднего балла бакалавра, указывая на то, что успеваемость до поступления в университет является информативной, но не достаточной для точных

долгосрочных прогнозов. Более поздние исследовательские работы были сосредоточены на включении данных об успеваемости студентов за первые семестры с использованием таких методов, как моделирование скрытого роста (latent growth modeling) и нейронные сети. Эти модели демонстрируют значительное улучшение прогнозов по сравнению с предыдущими моделями.

Несмотря на то, что прогнозирование среднего балла в долгосрочной перспективе остается сложной задачей, продолжающиеся исследования с применением методов анализа данных предлагают поиск более точных моделей, чтобы превратить эти модели в действенные инструменты поддержки успешного обучения студентов [5].

В докладе предлагается способ предсказания результатов обучения, основанный на моделях классификации и множественном регрессионном анализе, отличающийся применением показателей когнитивного потенциала студентов в качестве дополнительных факторов–предикторов.

III. Сравнение моделей в задачах прогнозирования

Приведем примеры результатов, полученных в некоторых исследованиях.

Достаточная степень точности предсказания выпускного среднего балла (GPA) на четвертом курсе университета с использованием только априорной информации о «доуниверситетской» успеваемости и экзаменационных баллах за первый и второй учебный год, равная 83.5 %, получена без учета социально-экономических, демографических или других факторов [6].

Получен прогноз успешности обучения студентов на основе классификации данных по результатам обучения за первые два семестра с применением Байесовского подхода с 74 %-й точностью [7].

В зарубежных исследованиях в модели достаточно часто включены демографические факторы, в отличие от отечественных исследований, где демографическим различиям придается меньше внимания. Хотя, возможно, в будущих исследованиях, следует уделить внимание и этим факторам.

Например, расширение числа предикторов в моделях с демографическими факторами позволило повысить точность предсказания до 80 %. [8]. В данном исследовании применялся метод деревьев решений. Однако по априорным данным об успеваемости за первый семестр результаты прогноза не превысили 76 % точности. Точность предсказания результатов обучения по математике с использованием демографических факторов достигает 83.9 % с применением Байесовской модели (Naive Bayes) [9].

Следующие ключевые факторы могут входить в модели и методы для прогнозирования успеваемости учащихся.

Объем рабочей памяти (Working memory capacity) – характеризует способность учащихся воспринимать и обрабатывать информацию.

Скорость обработки информации (Processing speed), представленной в мультимодальной форме (визуальной, вербальной, аудиальной) – показывает, насколько быстро учащиеся могут воспринимать информацию и реагировать на нее. Более быстрая и точная обработка связана с лучшими результатами обучения. Тесты на время реакции могут измерить эту способность.

Предварительные знания (Prior knowledge). Наличие определенных знаний обеспечивают основу для интеграции новой информации. Модели могут оценивать предыдущий опыт учащихся и знания основных понятий в качестве предиктора.

Метапознание (Metacognition) – способность отслеживать и контролировать собственное мышление и обучение. Метакогнитивные навыки, такие как самопроверка, позволяют учащимся лучше регулировать обучение.

Мотивация (Motivation). Внутренняя мотивация и самоэффективность в учебе являются сильными предикторами достижений.

Когнитивная нагрузка (Cognitive load). Модели, основанные на теории когнитивной нагрузки, оптимизируют представление учебного материала, чтобы избежать перегрузки.

Саморегуляция (Self-regulation) – навыки постановки целей, использования стратегий и отслеживания прогресса. Студенты с хорошей саморегуляцией достигают больших результатов.

Включение этих факторов в прогностические модели вместе с демографическими и академическими данными позволяет создать более полный профиль способностей и потребностей учащихся в обучении для более точных прогнозов их успеваемости [9].

Однако, следует уточнить, что сбор всех необходимых данных для исследований занимает значительное время. Поэтому важной задачей является получение необходимых данных для построения моделей наиболее быстро, за короткое время.

В этой статье в моделях предсказания общего среднего балла предлагается использовать факторы когнитивного потенциала студентов, которые оценивались с помощью модифицированных компьютерных методик Дж. Кагана (J. Kagan) (модификация Kagan-M), Л. Терстоуна (L.L. Thurstone) (модификация Thurstone-M) и Дж.Р. Струпа (J.R. Stroop) (модификация Stroop-M) в автоматизированной системе экспресс-диагностики [10].

IV. Результаты исследований

При выборе методов и моделей интеллектуального анализа учитывалась не только показатели точности и робастности, но и интерпретируемость. Поэтому, наряду с применением в исследованиях современных моделей в виде нейронных сетей, случайного леса, акцент в данной статье сделан на хорошо интерпретируемых моделях деревьев решений и множественной линейной регрессии.

При прогнозировании на основе моделей классификации, построенных по априорным данным, формируются оценки правдоподобия принадлежности студентов к классам успешности обучения.

Регрессионные модели способны предсказывать численные значения целевого показателя (среднего балла) с различной точностью, которая зависит в том

числе от используемых методов, наборов данных и факторов-предикторов.

A. Предсказание результативности обучения бакалавров на основе моделей классификации

В данном исследовании рассматривались два класса успешности обучения. К группе, условно названной группой «риска», относятся студенты, средний балл которых меньше оценки «хорошо». В наборе данных таким студентам присвоена метка «L». Студентам со средним баллом выше четверки присвоена метка «H».

Задача обучения модели классификации решалась с использованием 50 %-й выборки из набора экспериментальных данных. Сравнительный анализ эффективности различных методов интеллектуального анализа данных (деревьев решений, нейронных сетей, линейной регрессии и др.) проводился в программном комплексе «ОнтоМАСТЕР» [11].

С использованием метода деревьев решений (j48) размерность пространства факторов-предикторов была снижена до двух – среднего балла за первый семестр (*srb*1) и показателя имплицитной обучаемости (*K*1). Фрагмент дерева решений изображен на рис. 1. В данном примере коэффициент имплицитной обучаемости определен на основе выполнения методики Stroop-M

$$K1 = (T1_1 - T1_2) / T1_1,$$

где $T1_1$ – время выполнения первой половины первой серии методики Stroop-M, $T1_2$ – время выполнения второй половины первой серии методики Stroop-M.

```
srb1 <= 3.5: L
srb1 > 3.5
|   K1 <= 0.001488: L
|   K1 > 0.001488: H
```

Рис. 1. Фрагмент дерева решений для классификации студентов на две подгруппы с «удовлетворительным» (*L*) и «хорошим» (*H*) средним баллом

С помощью полученной двухфакторной модели правильно предсказаны классы для 92 % студентов.

B. Предсказание средних баллов на основе модели множественной регрессии

В качестве показателя академической успеваемости использовался средний балл за весь период обучения (четыре года) в бакалавриате по образовательной программе «Управление в технических системах».

Модель линейной регрессии общего среднего балла (*srb*) от среднего балла в первом семестре характеризуется значениями среднеквадратической ошибки (RMSE) 0.23 и коэффициента детерминации R^2=0.57:

$$srb = 0.7243 * srb1 + 1.2667$$

Для повышения точности предсказания среднего балла модель регрессии была дополнена факторами когнитивного потенциала студентов.

Оценка степени влияния скрытых факторов когнитивного потенциала на значение среднего балла осуществлялась с использованием измеряемых показателей результатов решения студентами серий тестовых перцептивно-когнитивных задач по модифицированным компьютерным методикам.

Набор данных «StudCSP-DataSet-23» содержит обезличенную информацию, включая средние баллы и показатели когнитивного потенциала. Выборку составили 40 студентов, обучающиеся в бакалавриате по направлению «Управление в технических системах» (2021–2022 годы).

Средние баллы в подгруппе студентов с «удовлетворительным» (достаточным) результатом обучения (уровнем компетенций) характеризуются нормальным распределением (среднее значение 3.77, стандартное отклонение 0.13, медиана 3.8).

Выбор наиболее значимых показателей для предиктивных моделей проводился на основе корреляционного анализа и метода главных компонент (*Principal component analysis, PCA*). Сильная корреляция (*R=0.84*) обнаружена между средним баллом за весь период обучения (*srb*) и средним баллом за первый семестр (*srb*1). В табл. 1. приведены полученные значения коэффициентов линейной корреляции Пирсона (R) между показателями когнитивного потенциала и средним баллом, весовые показатели их значимости (W) в главном компоненте, в который входят все факторы, перечисленные в табл. 1.

ТАБЛИЦА I. КОЭФФИЦИЕНТЫ КОРРЕЛЯЦИИ И ВЕСА ФАКТОРОВ В ГЛАВНОМ КОМПОНЕНТЕ

№	Фактор	W	R
1	*srb*1	0.31	0.84
2	*K*1	0.24	0.36
3	arff_2	0.12	0.36
4	pars3a_2	0.08	-0.30
5	kerr	0.07	-0.28
6	Hobr1	0.06	-0.28
7	pzmes	0.04	0.25
8	*H*2	0.02	-0.22
9	*K*2	0.02	0.20
10	*T*3minus*T*2	0.015	-0.20
11	*P*3	0.011	-0.19
12	*T*2del*T*1	0.004	0.18
13	*T*2	0.002	0.16
14	arff_1	0.001	-0.14

Факторы для предиктивных моделей, которые объясняют 99.9 % дисперсии, определены методом главных компонент (рис. 2).

Наиболее значимыми факторами по методу главных компонент для построения предиктивной модели являются (табл.1): srb1– средний балл в 1-м семестре; *K*1, *K*2 – соотношение времени решения когнитивных задач в первой и второй половинах двух серий методики Stroop-M; pars3a, arff – параметры моделей динамики решения когнитивных задач; pzmes – показатель, характеризующий когнитивный стиль полезависимость/поленезависимость (методика Thurstone-M); kerr – точность решения задач методики Kagan-M; *T*3minus*T*2, *T*2del*T*1 – различия в обработке неконгруэнтной информации; *T*2, *H*2, *P*3, Hobr1 – характеризуют эффект интерференции при восприятии информации в вербальной и визуальной формах.

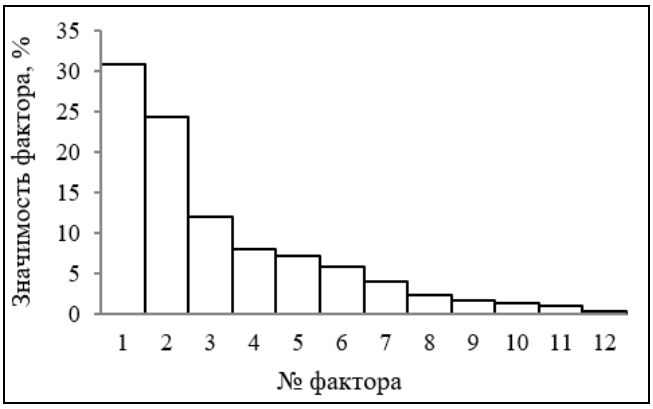

Рис. 2. Выбор значимых факторов для предиктивной модели по методу главных компонент

На основе выбранных факторов была построена модель множественной линейной регрессии для предсказания значений среднего балла. Расчет коэффициентов модели производился на основе 50 %-й выборки из набора данных.

Результаты предсказания средних баллов по регрессионной модели (рис. 3) характеризуются значениями среднеквадратической ошибки (0.15) и коэффициента детерминации (0.9).

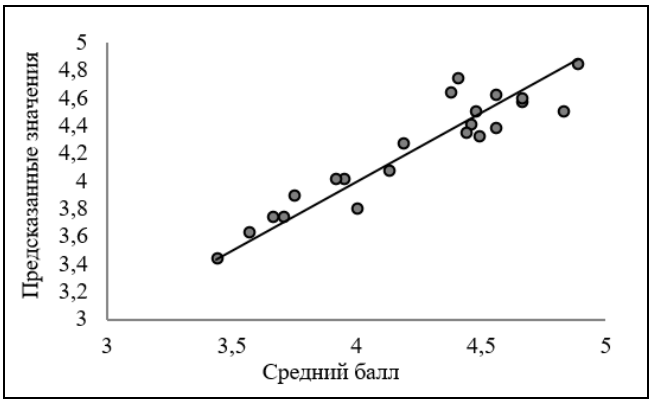

Рис. 3. Предсказанные по модели значения средних баллов

Таким образом, применение дополнительных факторов, характеризующих когнитивный потенциал студента, позволяют повысить точность прогнозирования общего среднего балла.

Прогноз результатов на начальном этапе обучения в вузе позволит заранее планировать распределение дидактических ресурсов/когнитивной нагрузки персонально для студентов с учетом личностных характеристик когнитивно-познавательной сферы.

V. Заключение

В работе представлены результаты исследования предиктивных моделей академической успеваемости студентов, обучавшихся по образовательной программе «Управление в технических системах».

Разработаны модели классификации и множественной регрессии для предсказания успешности обучения, отличающиеся применением показателей когнитивного потенциала студентов в качестве доминирующих факторов-предикторов.

Предварительные результаты апробации модели на наборе данных «StudCSP-DataSet-23» подтверждают возможность применения предлагаемого подхода с использованием среднего балла за первый семестр обучения и факторов когнитивного потенциала студентов для предсказания среднего балла за весь период обучения.

В дальнейшем планируется продолжить исследования для повышения точности предиктивных моделей результативности обучения.

Список литературы

[1] Mandernach J. Innovation in research and teaching. 2023 EDUCAUSE Horizon Report, Teaching and Learning Edition. Boulder, CO: EDUCAUSE. 2023. Pp.45-46.

[2] Brusilovsky P. Methods and techniques of adaptive hypermedia // User Modeling and User Adapted Interaction/ 1996. V. 6. No. 2–3. Pp. 87–129.

[3] Yang S. J., Lu O. H., Huang A. Y., Huang J. C., Ogata H., Lin A. J. Predicting students' academic performance using multiple linear regression and principal component analysis. Journal of Information Processing. 2018. Vol. 26. Pp. 170-176.

[4] Huang S., Fang N. Regression models for predicting student academic performance in an engineering dynamics course. 2010 Annual Conference & Exposition. 2010. Pp. 1-17.

[5] Kotova E.E. Intellectual data analysis in the educational process. Proceedings of 2017 XX IEEE international conference on soft computing and measurements. (SCM). 2017. Pp. 757-759.

[6] Asif R., Merceron A., Pathan M. K. Predicting student academic performance at degree level: a case study. International Journal of Intelligent Systems and Applications. 2015. Vol. 7. No. 1. Pp. 49-61.

[7] Makhtar M., Nawang H., Wan Shamsuddin S.N. Analysis on students performance using naïve bayes classifier. Journal of Theoretical & Applied Information Technology. 2017. Vol. 95. No. 16. Pp. 1-9.

[8] Bilal M., Omar M., Anwar W., Bokhari R. H., Choi G. S. The role of demographic and academic features in a student performance prediction. Scientific Reports. 2022. Vol. 12. No. 1. 9 p.

[9] Roslan M.H.B., Chen C.J. Predicting students' performance in English and Mathematics using data mining techniques. Education and Information Technologies. 2023. Vol. 28. No. 2. Pp. 1427-1453.

[10] Котова Е.Е. Прогнозирование успешности обучения в интегрированной образовательной среде с применением инструментов онлайн аналитики. Компьютерные инструменты в образовании. 2019. № 4. С. 55-80.

[11] Котова Е.Е., Писарев А.С., Писарев И.А. Программный комплекс разработки сценариев анализа данных ОнтоМАСТЕР-Сценарий. Свидетельство о государственной регистрации программы для ЭВМ № 2019613556 от 19.03.2019 г.