Разработка рекомендательного сервиса формирования групп первокурсников

В. О. Дубова

Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина)

1271072@gmail.com

Аннотация. **В работе рассматривается задача обработки данных студентов для составления рекомендаций по формированию академических групп первокурсников с целью повышения их успеваемости. Описано использование моделей классификации RandomForestClassifier (weighted F1-score: 0.33-0.4), KNeighborsClassifier (weighted F1-score: 0.64-0.71) для предсказания оценок студентов. Модель Constrained K-Means применена для кластеризации полученных оценок с целью формирования однородных академических групп. Для проверки предложенного метода были использованы данные студентов Факультета компьютерных технологий и информатики СПбГЭТУ «ЛЭТИ».**

Ключевые слова: рекомендательный сервис; формирование групп; прогнозирование успеваемости

I. Введение

Ежегодно перед работниками ВУЗов возникает задача формирования учебных групп первокурсников. При распределении вручную сложно принять во внимание индивидуальные особенности студентов, поэтому, как правило, группы формируются случайно. Более эффективным может оказаться дифференцированный подход, учитывающий первоначальную подготовку учащихся [1].

Проблему гетерогенности учебных групп можно решить, анализируя информацию о первокурсниках текущего года, собранную во время приёмной кампании, и данные о студентах предыдущих лет. В данной работе рассматривается задача предсказания академических результатов, которые учащиеся продемонстрируют в конце первого семестра обучения. Полученные прогнозы можно использовать для генерации рекомендаций по формированию учебных групп студентов и назначению преподавателей с целью оптимизации среднего балла.

II. Набор данных

Исследование проведено с использованием личных данных и оценок за первый семестр первокурсников, обучавшихся на Факультете компьютерных технологий и информатики СПбГЭТУ «ЛЭТИ» в 2021–2022 и в 2023–2024 годах. Для дальнейшего анализа был отобран ряд параметров, которые могут повлиять на будущие академические успехи студента. Краткое описание параметров приведено в табл. 1.

Все перечисленные параметры указываются абитуриентами при подаче документов в ВУЗ и позволяют произвести первоначальную характеристику студентов. Например, используя один только суммарный балл ЕГЭ, можно на 20 % объяснить академическую успеваемость первокурсника [2].

В анализе также используются оценки, полученные студентами за первый семестр по профильным предметам «Алгебра и геометрия», «Математический анализ» и «Физика». Данные первокурсников за 2021–2022 выступают в качестве обучающей выборки, данные за 2023–2024 – тестовой.

ТАБЛИЦА I. Данные, отобранные для дальнейшего анализа

Параметр	**Расшифровка параметра**
БВИ	Поступил без вступительных испытаний
ИД	Балл за индивидуальные достижения
Направление	Наименование направления подготовки
Общежитие	Потребность в общежитие
ОП	Особые права при поступлении
Оценка_1 DIV	Отношение оценки за ЕГЭ по математике, полученной студентом, к среднему баллу ЕГЭ по стране в год сдачи экзамена
Оценка _2 DIV	Отношение оценки за ЕГЭ по предмету по выбору (информатике или физике), полученной студентом, к среднему баллу ЕГЭ по стране в год сдачи экзамена
Оценка _3 DIV	Отношение оценки за ЕГЭ по русскому языку, полученной студентом, к среднему баллу ЕГЭ по стране в год сдачи экзамена
Предмет_2 Приоритет	Указан предмет вариативного ЕГЭ, результаты которого были учтены при поступлении
ПП	Приоритетное право при поступлении
Тип финансирования	Бюджет/контракт
Целевик	Поступал по целевой квоте
Санкт-Петербург	Студент закончил учебное заведение, расположенное в Санкт-Петербурге
Ленинградская область	Студент закончил учебное заведение, расположенное в Ленинградской области
Другие регионы	Студент закончил учебное заведение, расположенное не в Санкт-Петербурге или Ленинградской области

III. Ход решения

Задача, поставленная в п. 1, решена следующим путём:

- Предобработка данных о студентах.
- Классификация студентов для предсказания их успехов в учёбе по профильным предметам.
- Кластеризация полученных предсказаний для формирования учебных групп.

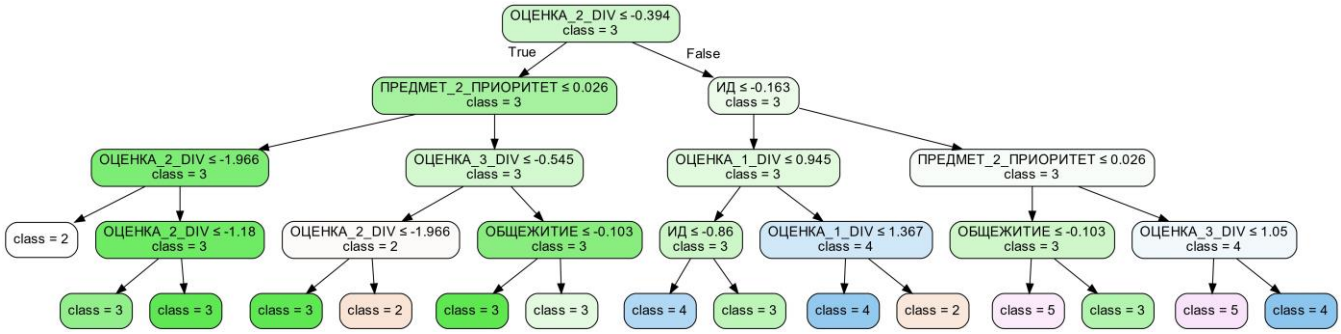


Рис. 1. Пример дерева классификации для предмета «Алгебра и геометрия»

A. Предобработка данных о студентах

Этот этап включает в себя обработку пропущенных и некорректных значений в наборе личных данных студентов и наборе данных с оценками студентов.

В наборе данных с информацией о первокурсниках обрабатываются дубликаты записей, которые возникают, например, в случае ухода студента в академический отпуск или при подаче документов на несколько направлений. В наборе данных с оценками были оставлены только первые попытки сдачи профильных предметов (без учёта допсессий). Оценки учащихся, не соответствующие пятибалльной шкале, например «н/я» – не явился, были обработаны как двойки.

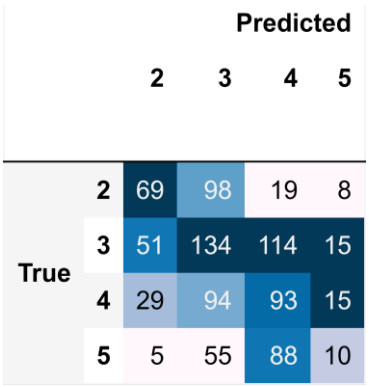
Следующим шагом является формирование общего набора данных, состоящего из объединения набора личных данных студентов и набора данных с оценками. Итоговый набор данных содержит информацию о первокурсниках и их оценках по профильным предметам за первый семестр.

В результате данной обработки был сформирован тренировочный набор данных, включающий в себя информацию о 621 студенте 2021–2022 годов обучения, и тестовый с данными о 897 студентах 2023–2024 годов обучения.

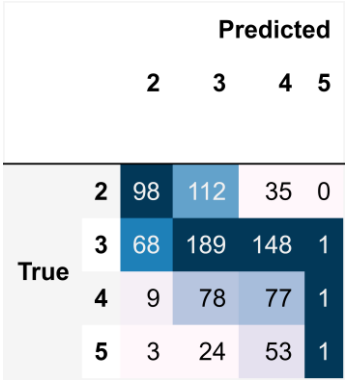
B. Классификация студентов для предсказания их успехов в учёбе по профильным предметам

На данном шаге можно применить алгоритм классификации для предсказания результатов первокурсников в учёбе. В качестве меток класса выступают оценки («2», «3», «4», «5»), полученные студентами за первый семестр обучения. Для классификации был выбран алгоритм RandomForestClassifier [3]. Такую модель легко представить в графическом виде, что открывает возможности для совместного использования мнения эксперта и результатов предсказания классификатора. Пример такого дерева приведён на рис. 1.

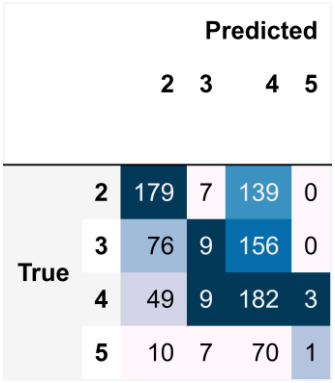
На рис. 2 представлены матрицы ошибок, построенные для результатов классификации. На них можно увидеть, что алгоритм часто путается между соседними оценками, а в случае (в) – «Физика» – предсказывает в основном только «2» и «4».



(а)



(б)



(в)

Рис. 2. Матрицы ошибок при использовании RandomForestClassifier для предметов: (а) – «Алгебра и геометрия», (б) – «Математический анализ», (в) – «Физика»

Для улучшения прогностических возможностей модели был использован альтернативный подход: объединение классов между собой. Анализ возможных комбинаций оценок показал, что наилучший результат достигается при объединении оценок в два класса: «23» – студент получил «2» или «3» и «45» – студент получил «4» или «5». Для классификации был выбран алгоритм KNeighborsClassifier [4], результаты работы которого приведены на рис. 3.

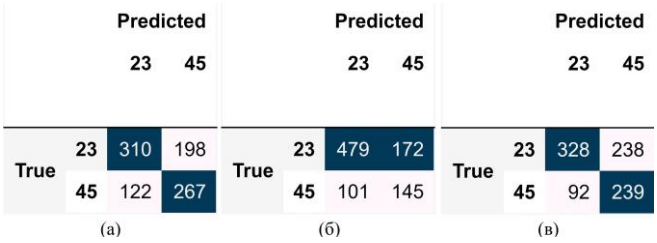


Рис. 3. Матрицы ошибок при использовании KNeighborsClassifier с двумя классами для предметов: (а) – «Алгебра и геометрия», (б) – «Математический анализ», (в) – «Физика»

Для оценки качества работы моделей использовалась метрика weighted F1-score. Weighted F1-score рассчитывается как взвешенное среднее арифметическое множества F1-score, подсчитанных отдельно для каждого класса, в качестве весов выступает количество истинных элементов в классе. Такая метрика позволяет учесть дисбаланс классов и может использоваться в случаях как бинарной, так и мультиклассовой классификации. Результаты работы моделей приведены в табл. 2.

ТАБЛИЦА II. WEIGHTED F1-SCORE ДЛЯ ПРОФИЛЬНЫХ ПРЕДМЕТОВ

Модель	Алгебра и геометрия	Математический анализ	Физика
RandomForest	0.33	0.4	0.35
KNeighbors	0.64	0.71	0.64

При анализе результатов, описанных в табл. 2, было обнаружено, что модель RandomForestClassifier показывает по всем предметам худший результат, чем KNeighborsClassifier. Также преимуществом второй модели является то, что она бинарная – её проще использовать для формирования учебных групп, что и является основной целью данной работы.

C. Кластеризация полученных предсказаний для формирования учебных групп

При помощи алгоритма KNeighborsClassifier можно предсказывать не метки классов, а вероятности принадлежности каждого элемента каждому классу. Эти данные могут быть использованы для создания учебных групп на основе вероятностей получения студентами определенных оценок, а не предсказанных классов («23» или «45»).

Для решения задачи разделения студентов на группы алгоритм кластеризации должен соответствовать ряду условий:

- работать при достаточно большом количестве студентов в выборке;
- иметь возможность задать количество групп (кластеров);
- иметь возможность задать минимальное и максимальное количество студентов в группе.

Большинство классических алгоритмов кластеризации соответствуют только одному или двум пунктам этого списка. Однако метод кластеризации K-Means [5] можно модифицировать так, чтобы он соответствовал описанным условиям.

Стандартный K-Means разбивает множество элементов на заданное количество кластеров, но не регулирует количество элементов внутри кластеров. Идея использования алгоритма поиска потока минимальной стоимости, для задания минимальных размеров кластера, впервые была озвучена в статье [6]. В данной работе использовалась реализация метода Constrained K-Means, модифицированная для регулирования максимального и минимального размеров кластера [7].

Используем полученные на предыдущем этапе вероятности того, что студенты получат оценки «4» или «5» по предметам «Алгебра и геометрия», «Математический анализ» и «Физика». Так как предмета всего три, мы можем визуализировать эти данные без дополнительной обработки. На рис. 4 представлено то, как первокурсники 2023–2024 годов обучения направления «Информационные системы и технологии» были разделены на группы. Каждый студент представлен точкой, цвет которой задаётся его принадлежностью группе. Чем ближе к единице расположена точка по оси одного из предметов, тем выше шанс того, что студент получит по соответствующему предмету «4» или «5».

Видим, что первокурсники распределены по группам хаотично, сформированные вручную группы неоднородны. При этом, с ростом вероятности получения высокой оценки по одному из предметов, растет и вероятность получения высоких оценок по другим предметам.

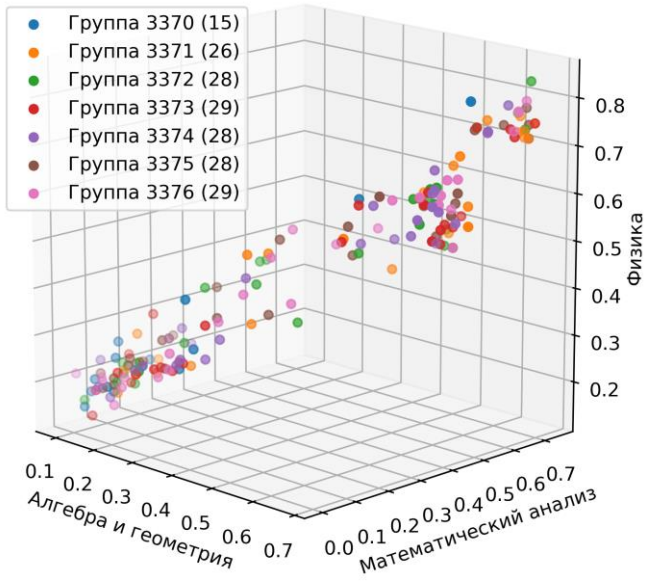


Рис. 4. Распределение студентов направления «Информационные системы и технологии» по группам без применения кластеризации

Теперь применим алгоритм Constrained K-Means к данным студентов каждого направления отдельно, задав ограничения на количество групп (задаётся в зависимости от направления), максимальное и минимальное количество студентов в группе (30 и 15 соответственно). На рис. 5 представлен график полученной кластеризации для направления «Информационные системы и технологии», аналогично распределены студенты других направлений. Видим, что группы сформированы в зависимости от ожидаемой успеваемости первокурсников, учитывающей их индивидуальные особенности и первоначальную подготовку.

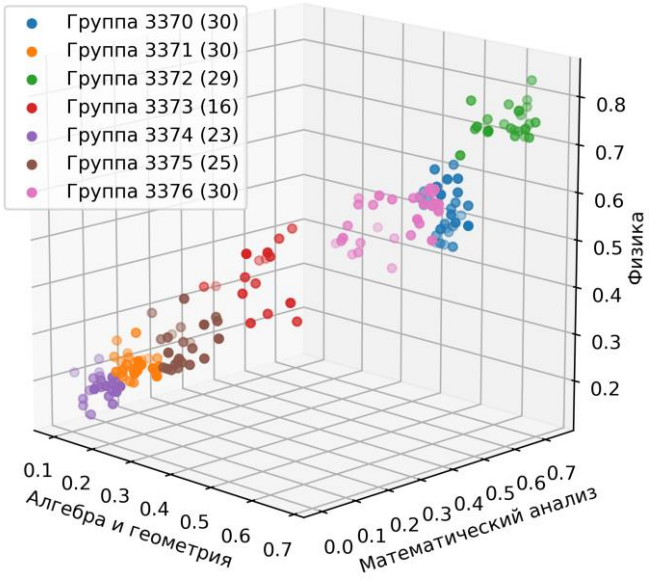


Рис. 5. Распределение студентов направления «Информационные системы и технологии» по группам с применением Constrained K-Means

IV. Заключение

В ходе работы были применены два метода классификации для предсказания ожидаемой успеваемости студентов в первом семестре первого курса. Расчёт показал, что метрика weighted F1-score для алгоритма RandomForestClassifier находится в диапазоне 0.33–0.4, для KNeighborsClassifier – 0.64–0.71. Алгоритм RandomForestClassifier предлагается использовать для предсказания оценок («2», «3», «4», «5») в комбинации с мнением эксперта. Алгоритм KNeighborsClassifier был применён для бинарной классификации оценок («23», «45»).

Был использован метод кластеризации Constrained K-Means для формирования однородных академических групп первокурсников. Предложенный подход позволяет учитывать индивидуальные особенности студентов. Полученные результаты могут быть полезны для ВУЗов при формировании учебных групп с целью улучшения оценок первокурсников и уменьшения числа отчислений.

В 2024–2025 учебном году будет проведена апробация метода для проверки того, что группы первокурсников Факультета компьютерных технологий и информатики, составленные по данной методике, будут учится лучше.

В дальнейшем полученные результаты можно использовать также для назначения преподавателей группам. Опираясь на расписание занятий первого семестра у студентов прошлых лет, планируется дополнить общий набор данных информацией о том, кто вёл лекции, практики и лабораторные («1» – вел, «0» – не вел). Анализируя этот набор данных, можно будет сформулировать рекомендации по назначению преподавателей таким образом, чтобы это повысило успеваемость групп.

Список литературы

[1] Сарычева И.А., Грибкова Ю.В., Голицына Е.В., Запатрина Н.В. «О проведении педагогического эксперимента по дифференцированному подходу к обучению высшей математике в вузе» // Преподаватель XXI век. 2022. №3-1 С. 114-121.

[2] Хавенсон Т.Е., Соловьева А.А. Связь результатов Единого государственного экзамена и успеваемости в вузе // Вопросы образования. 2014. №1 С. 176-199.

[3] Breiman L. Random Forests. // Machine Learning. 2001. №45, С. 5-32

[4] Scikit-learn, KNeighborsClassifier URL: https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html (дата обращения 29.03.2023)

[5] Scikit-learn, KMeans URL: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html (дата обращения 29.03.2023)

[6] Bradley P. S., Bennett K. P., Demiriz. A. Constrained k-means clustering // Microsoft Research, Redmond. 2000. С. 1-8.

[7] GitHub, joshlk/k-means-constrained URL: https://github.com/joshlk/k-means-constrained (дата обращения 29.03.2023)