УДК 004.852

МОДЕЛЬ ПРОГНОЗИРОВАНИЯ УСПЕВАЕМОСТИ СТУДЕНТОВ С ИСПОЛЬЗОВАНИЕМ ДАННЫХ О РЕЗУЛЬТАТАХ ИЗМЕРИТЕЛЬНЫХ МАТЕРИАЛАХ И ПОСЕЩАЕМОСТИ СТУДЕНТАМИ ЗАНЯТИЙ

Потапова К.А.

МИРЭА – Российский технологический университет, 119454, Россия, г. Москва, проспект Вернадского, 78, e-mail: ksurashanti@gmail.com

В статье описывается создание модели прогнозирования успеваемости студентов на основе анализа данных о результатах измерительных материалов и посещаемости занятий. В ходе исследования были использованы методы корреляционного анализа и машинного обучения, в результате чего была выбрана модель линейного дискриминантного анализа (LDA), показавшая хорошие результаты. Разработанная модель может помочь улучшить качество образования, а также может быть адаптирована для применения в других областях прогнозирования на основе множественных переменных.

Ключевые слова: прогнозирование, корреляционный анализ, машинное обучение, линейный дискриминантный анализ, образовательный процесс

CREATION OF A MODEL FOR PREDICTING STUDENT PERFORMANCE USING DATA ABOUT THE RESULTS OF MEASUREMENT MATERIALS AND CLASS ATTENDANCE

Potapova K.A.

MIREA - Russian Technological University, 119454, Russia, Moscow, Vernadsky prospect, 78, e-mail: ksurashanti@gmail.com

The article describes the creation of models for predicting the probability of students based on the analysis of data on the vulnerability of measuring materials and class attendance. During the study, methods of correlation analysis and machine learning were used, as a result of which a linear discriminant analysis (LDA) model was selected, which showed good results. The developed model can help improve the quality of education, and can also be adapted for use in other areas of forecasting based on multiple processes.

Keywords: forecasting, correlation analysis, machine learning, linear discriminant analysis, educational process

Введение

В современном образовательном процессе актуален вопрос об оптимизации и индивидуализации обучения, что требует разработки и внедрения новых методов оценки успеваемости студентов [1].

Одним из перспективных направлений в этой области является создание моделей прогнозирования успеваемости [2] на основе анализа различных факторов, влияющих на результаты обучения. В данной статье предлагается подход к созданию модели прогнозирования успеваемости студентов, основанной на данных об оценках за контрольные работы и посещаемости занятий.

Для разработки модели прогнозирования успеваемости будут использованы методы машинного обучения и анализа данных. Для этого предлагается использовать имеющиеся данные за семестр о результатах выполнения двух контрольных работ и посещаемости студентов. Предлагается предсказать две целевые переменные: оценку за итоговый экзамен и оценку за курсовую. В исследовании рассматривается моделирование бинарного исхода «сдал/не сдал», а также прогнозирование конкретной оценки студента на экзамене и курсовой.

Данные представлены в формате Excel, обработка и моделирование проведены в Jupiter Notebook на языке Python с использованием ряда библиотек: pandas, numpy, seaborn, matplotlib и sklearn [3].

Анализ данных

Набор данных на рисунке 1 содержит 187 строк и 26 столбцов: из них 8 соответствуют посещению лекций, 16 – посещению практических занятий, и 2 – результатам контрольных работ. Добавлен столбец об общем количестве посещенных занятий для удобства визуализации информации.

```
Ввод [9]: stud_train.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 187 entries, 0 to 186
Data columns (total 26 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   lk1        187 non-null    int64
 1   lk2        187 non-null    int64
 2   lk3        187 non-null    int64
 3   lk4        187 non-null    int64
 4   lk5        187 non-null    int64
 5   lk6        187 non-null    int64
 6   lk7        187 non-null    int64
 7   lk8        187 non-null    int64
 8   pr1        187 non-null    int64
 9   pr2        187 non-null    int64
 10  pr3        187 non-null    int64
 11  pr4        187 non-null    int64
 12  pr5        187 non-null    int64
 13  pr6        187 non-null    int64
 14  pr7        187 non-null    int64
 15  pr8        187 non-null    int64
 16  pr9        187 non-null    int64
 17  pr10       187 non-null    int64
 18  pr11       187 non-null    int64
 19  pr12       187 non-null    int64
 20  pr13       187 non-null    int64
 21  pr14       187 non-null    int64
 22  pr15       187 non-null    int64
 23  pr16       187 non-null    int64
 24  kr1        187 non-null    int64
 25  kr2        187 non-null    int64
```

Рисунок 1. Структура данных об успеваемости и посещаемости

У переменной посещаемости два возможных значения: 0 – занятие не посещено студентом, 1 – студент присутствовал на занятии. Строки, у которых было больше чем 5 незаполненных значений о посещении, удалены из таблицы. В остальных случаях пропущенные значения заменены 0.

Переменные оценок за контрольные работы имеют три возможных значения: 0 означает, что студент не пришел на контрольную или не справился с ней, 4 и 5 – положительные оценки.

Исследуем данные. На 2 рисунке представлены два графика: зависимость оценки за курсовую (kurs) от количества посещённых занятий и зависимость оценки за экзамен (ekz) от количества посещенных занятий.

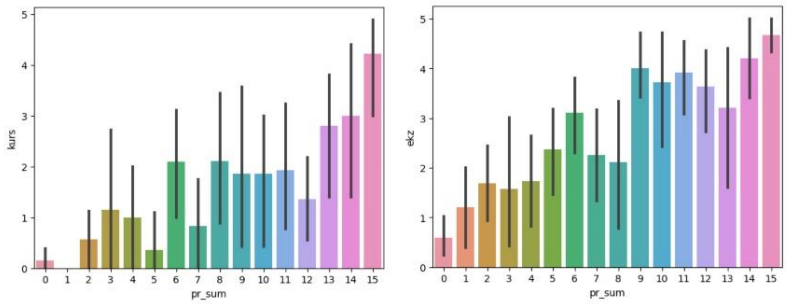

Рисунок 2. Зависимость оценок от количества посещенных занятий

На 3 рисунке представлено среднее количество посещенных практических занятий студентами, получившими за экзамен неудовлетворительно, удовлетворительно, хорошо и отлично.

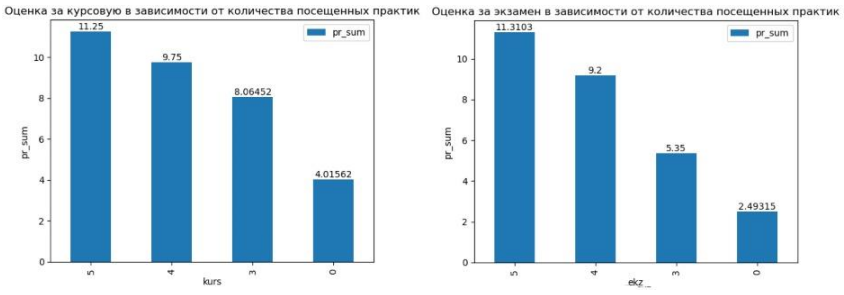

Рисунок 3. Среднее количество посещенных занятий

На обоих рисунках наблюдаем линейную зависимость. Чем больше занятий посетил студент – тем выше оценку в среднем получил на экзамене и сдаче курсовой. Также можно отметить, что студенты в среднем хуже справляются с курсовой работой, чем с экзаменом.

На рисунке 4 изображена матрица корреляций Пирсона. Самая высокая корреляция наблюдается между оценкой за первую и вторую контрольную работу (0.78), что говорит о высокой взаимосвязи этих факторов. Можно предположить, что студенты, успешно сдавшие первую контрольную работу, тратили больше времени на подготовку ко второй контрольной работе.

Рисунок 4. Матрица корреляций Пирсона

Также заметна довольно высокая взаимосвязь между другими факторами, например, между оценкой за экзамен и посещаемостью занятий (0.64). Высокая корреляция позволяет предположить возможность эффективного моделирования итоговых оценок на основе этих факторов.

Анализ методов классификации

Для построения модели предсказания целевых переменных следует выбрать оптимальный алгоритм. В таблице 1 рассмотрены преимущества и недостатки трёх алгоритмов машинного обучения: линейный дискриминантный анализ (LDA), деревья решений [4] и метод k-ближайших соседей [5].

Таблица 1. Сравнение разных методов классификации

Метод	**Преимущества**	**Недостатки**
LDA – работает путем вычисления расстояния между точкой данных и средним значением каждого класса и выбора класса с наименьшим расстоянием.	• Быстрый и простой для вычисления • Не требует большого количества обучающих данных	• Чувствителен к выбросам • Плохо работает на больших наборах данных с большим количеством признаков
Деревья решений – метод заключается в построении дерева, где каждый узел представляет собой условие, а ветви дерева соответствуют результатам выполнения этого условия.	• Возможность проверить созданный набор правил, простота интерпретации • Возможность работы с большим объёмом данных • В случае хорошей структурированности данных показывает высокую точность классификации	• Склонны к переобучению • Невозможность работы с непрерывными признаками
k-ближайших соседей – основан на поиске ближайших соседей точки, которую нужно классифицировать или спрогнозировать, и принятии решения на основе ответов	• Не требует предварительной обработки данных • Работает с разными типами данных, в том числе числовыми, текстовыми и категориальными	• Требует больших расходов памяти • Медленно работает на больших объёмах данных

Для построения модели предсказания был выбран метод LDA, поскольку объём данных для анализа невелик, есть выраженная корреляция между столбцами, а выбросы в данных отсутствуют.

В основе дискриминантного анализа лежит предположение о том, что описания объектов каждого k-го класса представляют собой реализации многомерной случайной величины [6]. Эта случайная величина распределена по нормальному закону $N_m(\mu; \Sigma_k)$ со средними μ_k и ковариационной матрицей (1):

$$c_{\bar{k}} \frac{1}{n_{\bar{k}}1} \sum_{i=1}^{n_k} \left(x_{i_{\bar{k}}} \mu_k\right)^T (x_{iK} - \mu_K) \quad (1)$$

Алгоритм построения модели предсказания

Алгоритм построения модели предсказания с использованием атрибута LinearDiscriminantAnalysis из библиотеки sklearn на примере прогнозирования оценки за курсовую работу состоит в следующем:

- Выделяем целевую переменную на рисунке 5.

Рисунок 5. Выделение целевой переменной: оценки за курсовую работу

- Разделяем данные на выборку для обучения и тестирования в соотношении 20/80 с помощью атрибута train_test_split на рисунке 6. Проверяем размер тестовых данных.

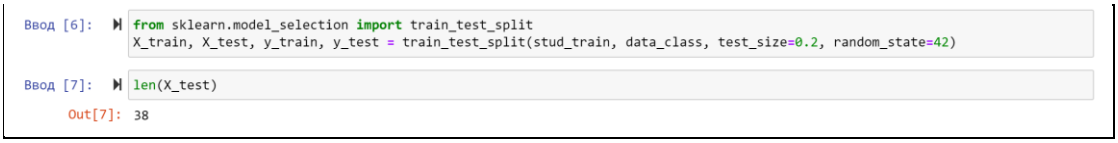

Рисунок 6. Разделение данных для обучения и тестирования

- На рисунке 7 строим модель LDA, обучаем данные.

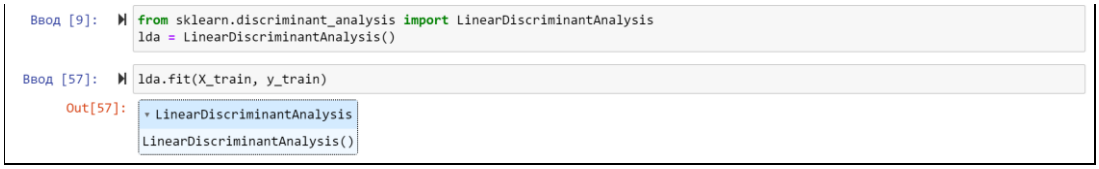

Рисунок 7. Модель LDA

- Делаем прогноз на тестовой выборке на рисунке 8.

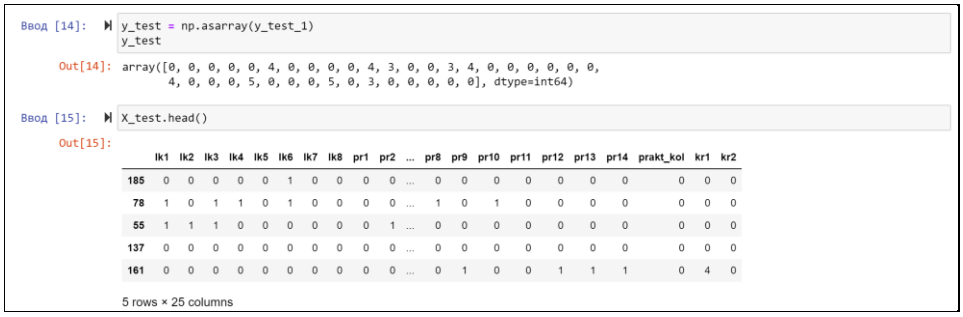

Рисунок 8. Прогнозирование на основе тестовой выборки

- Рассчитываем точность модели на тестовой выборке с помощью модуля accuracy_score [7]. Accuracy_score высчитывает правильно спрогнозированную долю выборки на основе истинно-положительных (TP), истинно-отрицательных (TN), ложно-положительных (FP) и ложно-отрицательных (FN) показателей (2):

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

На рисунке 9 в строке с индексом 0 реальные значения, в строке с индексом 1 – предсказанные.

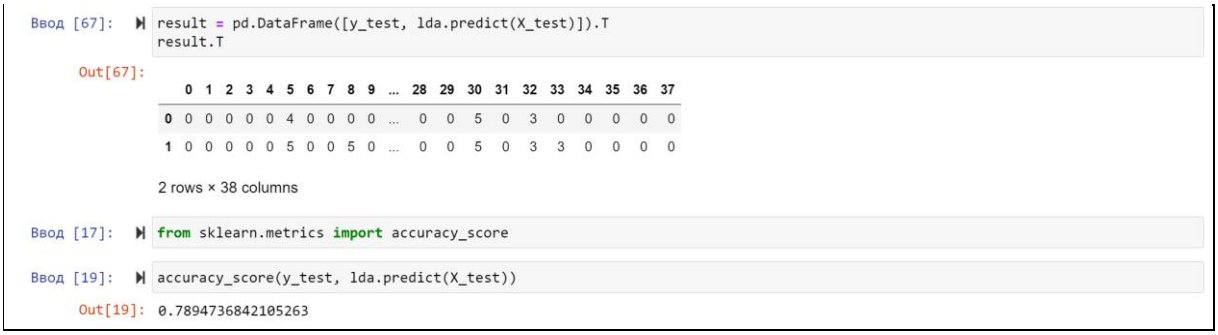

Рисунок 9. Расчёт точности модели

Точность модели составила 78%. С этой вероятностью модель верно предскажет оценку за курсовую работу на основе посещений и оценок в течение семестра.

Попробуем сделать прогноз сдачи экзамена на основе бинарного исхода: 0 – студент не справился с экзаменом, 1 – студент получил любую положительную оценку.

Для этого необходимо провести обработку данных в столбце «ekz», заменив положительные оценки на 1. После чего повторить все шаги алгоритма. Пример показан на рисунке 10.

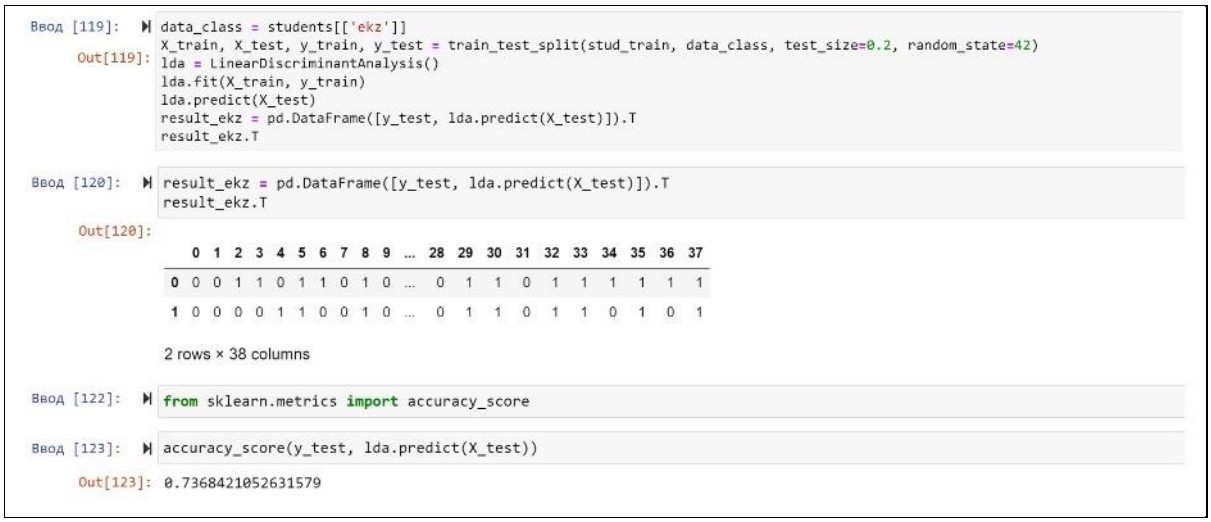

Рисунок 10. Построение модели LDA, прогнозирование и проверка точности для определения успешности сдачи экзамена в бинарном исходе

Точность прогноза составила 73%.

Заключение

Модель на основе линейного дискриминантного анализа показала хорошую точность (73% и 78%) в прогнозировании успеваемости студентов.

Разработанная модель может помочь преподавателям и студентам определить, какие аспекты обучения наиболее важны для достижения успеха, а также позволит преподавателям выявлять в течение семестра студентов, требующих особого внимания. Кроме того, данная работа может стать основой для дальнейших исследований в области прогнозирования успеваемости и анализа образовательных данных.

Список литературы

1. Де С., Басу Р. и Гангули Д., Предиктивная аналитика успеваемости студентов // 5-я Международная конференция IEEE по интеллектуальным вычислениям и системам управления (ICICCS) – 2021, стр. 128-132
2. Моисеев В.Б., Зубков А.Ф., Деркаченко В.Н., Прогнозирование успеваемости студентов по общепрофессиональным и специальным дисциплинам на основе регрессионных моделей. // Информационные и телекоммуникационные технологии в образовании — Научно-технические ведомости СПбГПУ 6’ 2010. 169-173
3. Жерон А., Практическое машинное обучение с помощью Scikit-Learn и TensorFlow: концепции, инструменты и методы для создания интеллектуальных систем // O’Reilly Media, Inc., 2017. 568 с.
4. Кафтанников И.П., Парасич А.В., Особенности применения деревьев решений в задачах классификации // Bulletin of the South Ural University. Ser. Computer Technologies, Automatic Control, Radio Electronics. – 2015. – № 3. – С. 26-32
5. Бабаев А.М., Шемякина М.А., Ляшов М.В., Обзор классических методов машинного обучения в контексте решения задач классификации // Форум молодых учёных – 2018. – № 11(27). – С. 137-143
6. Петухов Д.Е., Ткаченко А.В., Белов Ю.С. Линейный дискриминантный анализ как контролируемый подход в задачах уменьшения размерности данных // Научное обозрение. Технические науки. – 2020. – № 2. – С. 5-9
7. Плас Дж. Вандер, Python для сложных задач: наука о данных и машинное обучение // Серия «Бестселлеры O’Reilly» — СПб.: Питер, 2018. — 576 с.

Reference

1. De, S., Basu, R., & Ganguly, D., Predictive analytics of student academic performance // IEEE 5th International Conference on Intelligent Computing and Control Systems (ICICCS) – 2021, pp. 128-132
2. Moiseev V.B., Zubkov A.F., Derkachenko V.N., Forecasting student performance in general professional and special disciplines based on regression models. // Information and telecommunication technologies in education – Scientific and Technical Journal of St. Petersburg State Polytechnic University 6’ 2010. 169-173
3. Géron A. Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. Sebastopol, O’Reilly Media, Inc., – 2017. 568 p.
4. Kaftannikov I.P., Parasich A.V., Features of the use of decision trees in classification problems // Bulletin of the South Ural University. Ser. Computer Technologies, Automatic Control, Radio Electronics. – 2015. – No. 3. – P. 26-32
5. Babaev A.M., Shemyakina M.A., Lyashov M.V., Review of classical machine learning methods in the context of solving classification problems // Forum of Young Scientists – 2018. – No. 11(27). pp. 137-143
6. Petukhov D.E., Tkachenko A.V., Belov Yu.S. Linear discriminant analysis as a controlled approach to solving data dimension problems // Scientific review. Technical science. – 2020. – No. 2. – P. 5-9
7. Plas J. Vander, Python for complex problems: data science and machine learning // O’Reilly Bestsellers Series – St. Petersburg: Peter, 2018. - 576 pp.