Инженерный вестник Дона, №4 (2024)
ivdon.ru/ru/magazine/archive/n4y2024/9114

Построение и оценка эффективности модели дерева решений для прогнозирования успеваемости обучающихся

О.А. Пырнова[1], А.С. Катасёв[2]

[2]Казанский государственный энергетический университет, Казань
[1]Казанский национальный исследовательский технический университет им. А.Н. Туполева-КАИ, Казань

Аннотация: В данной работе решается задача повышения эффективности образовательной деятельности за счет прогнозирования успеваемости обучающихся на основе внешних и внутренних факторов. Для решения данной задачи построена модель прогнозирования успеваемости обучающихся с использованием языка программирования Python. Исходные данные для построения модели дерева решений взяты с платформы UCI Machine Learning Repository и предварительно обработаны с помощью аналитической платформы Deductor Studio Academic. Приведены результаты работы модели и проведено исследование для оценки эффективности прогнозирования успеваемости обучающихся.
Ключевые слова: прогнозирование, дерево решений, успеваемость обучающихся, влияние факторов, оценка эффективности.

В современных условиях, когда происходит активное развитие цифровых технологий, применение методов автоматического анализа данных в различных областях жизни становится все более популярным. Одной из таких областей является образование, где использование инновационных технологий может помочь улучшить процесс обучения. В частности, использование электронных учебников позволяет получить доступ к информации из любой точки мира и в любое удобное время. Кроме того, благодаря развитию информационных технологий, возможно проведение дистанционных занятий, вебинаров, онлайн-курсов и других форм обучения. Но наиболее эффективным применением информационных технологий в образовании является использование интеллектуальных систем для оптимизации процесса обучения или анализа данных о студентах и их потребностях для дальнейшего подбора необходимого материала [1-4].

Использование интеллектуальных моделей для прогнозирования успеваемости обучающихся позволяет систематически анализировать факторы, влияющие на успеваемость студентов, что, в свою очередь,

помогает улучшить методы преподавания и увеличить эффективность образовательной программы. Такие факторы могут быть связаны как непосредственно с самим процессом обучения, так и с внешней средой.

Для оценки эффективности моделей прогнозирования можно применять различные методы машинного обучения. Один из таких методов - дерево решений, которое строит древовидную структуру, где каждый узел представляет собой условие, а каждый лист – предсказание. Дерево решений позволяет разбивать данные на более мелкие сегменты, основываясь на характеристиках и признаках, и использовать их для принятия решений о классификации или регрессии.

Для решения поставленной задачи выбран набор данных из репозитория машинного обучения UCI (University of California, Irvine) Machine Learning Repository, который содержит информацию об успеваемости обучающихся португальских школ (в возрасте от 15 до 22) по двум предметам – математике и португальскому языку. Данные были собраны в ходе исследования, которое было проведено в течение двух лет с 2008 по 2009 годы. Также список обновлялся в 2012 и 2015 годах. Характеристики набора данных – многовариантные и содержат 33 атрибута, 30 из которых описывают различные факторы, такие, как пол, тип домашнего адреса, образование родителей, их место работы, взаимоотношения в семье, время, которое они затрачивают на различные аспекты их жизни, связанные как с процессом обучения, так и вне его, текущее состояние здоровья и количество пропущенных занятий. Последними тремя атрибутами являются оценки обучающихся [5,6].

Также, при анализе данных было выявлено, что в португальских школах используется 20-балльная шкала оценивания обучающихся. Так, при переводе в пятибалльную систему, она будет соответствовать следующим

значениям: 0-9 баллов – «1»; 10-11 баллов – «2»; 12-13 баллов – «3»; 14-15 баллов – «4»; 16-20 баллов – «5».

Для оценки качества и подготовки данных используется аналитическая платформа Deductor Studio Academic, которая представляет собой программное обеспечение для статистического анализа. В результате работы с аналитической платформой выявлено, что все параметры полностью пригодны для дальнейшего построения модели прогнозирования успеваемости обучающихся. Также, при проведении корреляционного анализа данных, который позволяет оценить силы и направление между переменными, можно увидеть, что 15 показателей превышают определенный порог значимости (0,07), некоторые из них имеют отрицательную величину, что свидетельствует о наличии обратной зависимости (рис. 1).

Матрица корреляции

Входные поля		Корреляция с выходными полями
№	Поле	G3
6	failures	-0,372
2	Medu	0,216
5	studytime	0,170
3	Fedu	0,169
19	Walc	-0,146
21	absences	-0,142
18	Dalc	-0,132
17	goout	-0,119
4	traveltime	-0,118
7	schoolsup	-0,109
13	internet	0,104
1	age	-0,100
9	paid	-0,084
16	freetime	-0,080
20	health	-0,077
14	romantic	-0,064
10	activities	0,063
12	higher	-0,060
15	famrel	0,053
11	nursery	0,041
8	famsup	-0,013

Рис. 1. – Корреляционный анализ

После оценки качества данных, необходимо сформировать обучающую и тестовую выборки для оценки качества модели машинного обучения и её устойчивости [7]. Для этого написан скрипт на языке программирования python с использованием библиотеки scikit-learn, который загружает файл данных, используя библиотеку Pandas, и добавляет новый столбец,

содержащий среднее значений оценок за три периода, а также удаляет столбцы G1, G2, G3 [8, 9]. Затем происходит подготовка данных для обучения модели путем замены категориальных признаков численными, выбора необходимых признаков для обучения и разбиения данных на тренировочную и тестовую выборки [10]. После этого данные делятся на обучающую и тестовую выборки с помощью функции train_test_split из библиотеки scikit-learn. Обучающая выборка составляет 75% от исходных данных, тестовая – 25%. Результаты разбиения сохраняются в переменных X_train, y_train, X_test, y_test.

Далее для оценки эффективности модели на основе дерева решений написан программный код на языке python с использованием библиотек pandas, Scikit-learn и numpy. Также, в данном случае оценка эффективности модели происходит с помощью расчета ошибок прогнозов: Accurancy, Mean Squared Error (MSE) и Root Mean Squared Error (RMSE).

Ниже представлен блок программного кода для выбранной модели:

```
# Обучаем модель
model = DecisionTreeClassifier().fit(x_train, y_train)
# Делаем прогноз
predicted_y = model.predict(x_test)
# Вычисляем среднеквадратичную ошибку
mse = mean_squared_error(y_test, predicted_y)
rmse = np.sqrt(mean_squared_error(y_test, predicted_y))
# Выводим показатели точности
print("Accuracy Score:", accuracy_score(y_test, predicted_y))
print("Decision Tree MSE:", mse)
print("Decision Tree RMSE:", rmse)
```

В коде проводится обучение модели дерева решений и производится прогноз на основе тестовых данных. После этого вычисляются показатели точности и качества модели. Также была построена матрица ошибок, которая позволяет проанализировать и визуализировать правильность спрогнозированных данных. Данная матрица представлена на рисунке 2.

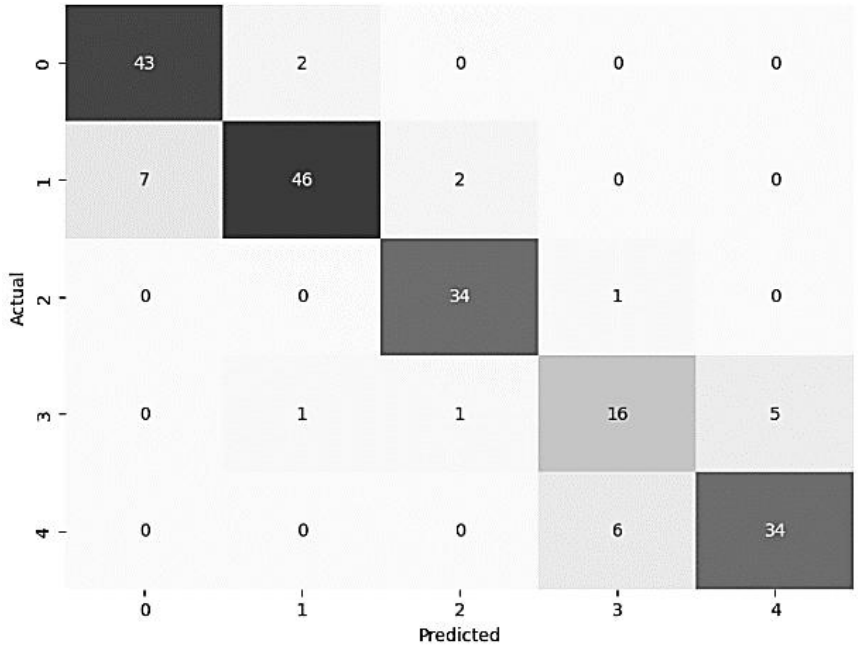

Рис. 2. – Матрица ошибок

В результате можно сделать вывод, что точность составила 87,37 %, что указывает на адекватность построенной модели [10, 11]: она верно распознала 173 примера и 25 – неверно. На основе матрицы ошибок рассчитаны значения метрик precision, recall и F1-score (таб. 1).

Таблица № 1

Результаты оценки производительности модели

Класс объекта	Метрики		
	precision	recall	F1-score
1	0,86	0,96	0,91
2	0,94	0,84	0,88
3	0,92	0,97	0,94
4	0,70	0,70	0,70
5	0,87	0,85	0,86

В ходе проведенных исследований было установлено, что выбранная модель для прогнозирования успеваемости учащихся продемонстрировала высокую эффективность. Это свидетельствует о благоприятном потенциале данной модели, что позволяет использовать ее в целях улучшения образовательного процесса.

Литература

1. Сапрыкина Т.А. О переходе «школа – ВУЗ»: предикторы успеваемости студентов-первокурсников // Высшее образование в России. 2017. № 6. С. 76-87.

2. Татусь К.Ю., Кузьмина С.В. Влияние родительской семьи на успеваемость студентов // Молодой ученый. 2016. № 9-4(113). С. 69-72.

3. Дашкуева П.В., Пырнова О.А., Кузнецов М.Г. Методы подготовки студентов к профессиональному развитию: решения и инновации в университетском образовании // Научное обозрение. Серия 2: Гуманитарные науки. 2024. № 1. С. 87-97.

4. Косулин В. В. Электронные образовательные ресурсы в обучении студентов инженерным дисциплинам // Уральский научный вестник. 2018. Т. 11, № 2. С. 037-042.

5. Cortez P., Silva A.M.G. Using data mining to predict secondary school student performance // Proceedings of 5th Annual Future Business Technology Conference. 2008. P. 5-12.

6. Yagsi M. Educational data mining: prediction of students' academic performance using machine learning algorithms // Smart Learning Environments. 2022. №9. URL: doi.org/10.1186/s40561-022-00192-z.

7. Чувакина К.А. Использование программного комплекса Deductor studio academic для обучения нейросетевому моделированию // Современные инновационные технологии подготовки инженерных кадров для горной промышленности и транспорта. 2016. №1(3). С. 555-559.

8. Майорова Е.С., Зарипова Р.С. Разработка алгоритма переноса стиля изображения с использованием предобученной нейросети // Инженерный вестник Дона. 2024. №1. URL: ivdon.ru/ru/magazine/archive/n2y2024/8997.

9. Овсеенко Г.А. SMART-решения и системы искусственного интеллекта // Информационные технологии в строительных, социальных и экономических системах. 2021. № 2 (24). С. 71-74.

10. Емалетдинова Л.Ю., Кабирова А.Н., Катасев А.С. Методика разработки нейросетевых моделей регуляторов управления техническим объектом // Инженерный вестник Дона. 2023. № 7. URL: ivdon.ru/ru/magazine/archive/n7y2023/8544.

References

1. Saprykina T.A. Vysshee obrazovanie v Rossii. 2017. № 6. pp. 76-87.

2. Tatus K.YU., Kuzmina S.V. Molodoj uchenyj. 2016. №9-4(113). pp.69-72.

3. Dashkueva P.V., Pyrnova O.A., Kuznecov M.G. Nauchnoe obozrenie. Seriya 2: Gumanitarnye nauki. 2024. № 1. pp. 87-97.

4. Kosulin V. V. Ural'skij nauchnyj vestnik. 2018. T. 11, № 2. pp. 037-042.

5. Cortez P., Silva A.M.G. Proceedings of 5th Annual Future Business Technology Conference. 2008. pp. 5-12.

6. Yagsi M. Smart Learning Environments. 2022. №9. URL: doi.org/10.1186/s40561-022-00192-z.

7. CHuvakina K.A. Sovremennye innovacionnye tekhnologii podgotovki inzhenernyh kadrov dlya gornoj promyshlennosti i transporta. 2016. №1 (3). pp. 555-559.

8. Majorova E.S., Zaripova R.S. Inzhenernyj vestnik Dona. 2024. № 1. URL: ivdon.ru/ru/magazine/archive/n2y2024/8997.

9. Ovseenko G.A. Informacionnye tekhnologii v stroitel'nyh, social'nyh i ekonomicheskih sistemah. 2021. № 2 (24). pp. 71-74.

10. Emaletdinova L.YU., Kabirova A.N., Katasev A.S. Inzhenernyj vestnik Dona. 2023. № 7. URL: ivdon.ru/ru/magazine/archive/n7y2023/8544

Дата поступления: 19.02.2024 Дата публикации:28.03.2024