Известия Алтайского государственного университета. 2024. № 1 (135). С. 114–119.
Izvestiya of Altai State University. 2024. No 1 (135). P. 114–119.

Научная статья
УДК 519.6:004.89
DOI: 10.14258/izvasu(2024)1-16

# Исследование оценки качества подготовки обучающихся методами интеллектуального анализа данных

*Татьяна Викторовна Михеева*

Алтайский государственный университет, Барнаул, Россия, mikheeva@math.asu.ru

Original article

# Research on Assessing the Quality of Students' Training Using Data Mining Methods

*Tatyana V. Mikheeva*

Altai State University, Barnaul, Russia, mikheeva@math.asu.ru

Статья посвящена применению методов интеллектуального анализа данных в задачах оценки качества подготовки обучающихся на основе официальных данных результатов выпускных проверочных работ по Алтайскому краю. Использованы методика дескриптивного анализа данных, методы корреляционного анализа данных, кластеризации, классификации, регрессии. В результате проведенного исследования определены лучшие алгоритмы методов интеллектуального анализа данных для решения задачи оценки качества подготовки обучающихся. Проведена оценка методами интеллектуального анализа и выявлены наиболее зависимые закономерности. Исследование выполнено машинными и графическими методами вычислительной математики, а результаты имеют теоретическое и прикладное значение. Разработанная программа может служить основой формирования плана исследования в задачах оценки качества подготовки обучающихся на основе анализа результатов выпускных проверочных работ, а также для обоснования рекомендаций на основе количественных оценок, полученных в результате проведенного анализа.

***Ключевые слова:*** интеллектуальный анализ данных, корреляция, классификация, регрессия, кластеризация, логистическая регрессия, деревья решений, метод k-ближайших соседей, метод k-средних

The article is devoted to the application of data mining methods to assess the quality of general education using the official data of the final examination results in the Altai Region. The methods of descriptive data analysis, correlation data analysis, data clustering, classification, and regression are used in the study. The obtained results identify the best data mining algorithms for the considered problem. The most dependent patterns are revealed during the education quality assessment using methods of intellectual analysis. The conducted study utilizes the computer and graphical methods of computational mathematics and provides the results with theoretical and applied significance. The developed programs can serve as the basis for prospective research plans to assess the quality of education using the analysis of the final examination results, as well as to provide the foundation for the quantitative assessment of such analysis.

***Keywords:*** data mining, correlation, classification, regression, clustering, logistic regression, decision trees, k-nearest neighbors method, k-means method

***Для цитирования:*** Михеева Т.В. Исследование оценки качества подготовки обучающихся методами интеллектуального анализа данных // Известия Алтайского государственного университета. 2024. № 1 (135). С. 114–119. DOI: 10.14258/izvasu(2024)1-16.

***For citation:*** Mikheeva T.V. Research on Assessing the Quality of Students' Training Using Data Mining Methods. *Izvestiya of Altai State University*. 2024. No 1 (135). P. 114–119. (In Russ.). DOI: 10.14258/izvasu(2024)1-16.

**Введение**

В настоящее время интеллектуальный анализ данных широко применяется во многих сферах деятельности человека, в том числе образовательной. Образовательные организации несут ответственность за качество предоставляемых услуг, поэтому одним из показателей качества образования в организациях является успеваемость учащихся [1–4]. Отсюда возникает необходимость достоверной информации о реальных результатах деятельности, а также о росте или снижении уровня образовательного процесса. Для этого ежегодно в России проводятся различные итоговые контрольные работы по отдельным предметам. Целью проведения всероссийских проверочных работ (ВПР) является своевременная диагностика уровня освоения образовательных программ, а также уровня достижения образовательных результатов с учетом требования ФГОС [5].

Данная работа посвящена проведению исследования оценки качества подготовки обучающихся общеобразовательных организаций методами интеллектуального анализа данных на основе результатов ВПР. Исследование выполнено на базе официальных данных результатов ВПР учащихся 4–8 классов образовательных организаций Алтайского края [6].

**Методы**

Для обработки и анализа данных в качестве программного инструмента применялся современный язык программирования Python [7]. Выполнено описание, предварительная обработка исходных данных, исследование, сравнительный анализ. Исходные данные по каждому предмету представляют собой файл в формате .xlsx с двумя листами. Первый лист содержит информацию по программам обучения в школах по каждому классу, второй — данные по каждому обучающемуся, включающие первичные баллы и информацию о месте обучения. Для чтения файла использовалась функция read_excel() из библиотеки pandas, в которую передавался путь к файлу и номер листа. В результате из одного файла выгружались по две таблицы.

Данные имеют некоторые особенности, затрудняющие дальнейшую обработку: пропуски, некорректные значения, неудобный формат данных и др. [8]. Указанные признаки могут привести к затруднениям в процессе их обработки. Поэтому была проведена предварительная обработка данных, что является первым и важным шагом в процессе интеллектуального анализа данных.

Для проведения исследования использовались следующие методы и функции Python [9–11]:

1. Для проведения описательного (дескриптивного) анализа использовалась функция describe() из библиотеки pandas.

2. Для составления матрицы парной корреляции использовалась функция corr() из библиотеки pandas с параметром method='pearson', обозначающим ее тип. Для более наглядного представления коэффициентов матрицы парной корреляции была построена тепловая карта heatmap из библиотеки seaborn, параметрами которой является сама матрица корреляции, linewidth=.5, обозначающей линию, разделяющую ячейки и annot=True — значения корреляции в ячейках.

3. Задача кластеризации была применена для нахождения классов, по которым могут разбиваться данные. Для кластеризации в Python были применены следующие алгоритмы: KMeans, MiniBatchKMeans, AgglomerativeClustering, BisectingKMeans. Наилучшим алгоритмом оказалась агломеративная кластеризация (AgglomerativeClustering). В качестве целевого значения алгоритм выбрал столбец «Оценка за ВПР».

4. Задача классификации была применена для нахождения признаков, по которым были разбиты данные. В качестве класса взята целевая переменная «Отметка за предыдущий период», обозначаемая Y. Данные были разбиты на 4 группы в соответствии с отметкой, полученной исходя из итогового балла. Данный атрибут показывает итоговый результат освоенности школьной программы по предмету. В качестве предикторов были взяты столбцы «Муниципалитет», «Пол», «Код школы», «Процент выполнения», «Класс», «Итого баллов» и «Учебник», «Тип ОО», «Тип поселения», добавленные в переменную X. Таким образом можно узнать, какие из этих признаков сильнее влияют на оценку за работу. Для задач классификации и регрессии для формирования списка предикторов были убраны столбцы, явно зависящие на целевую переменную.

Для многоклассовой классификации наилучшими алгоритмами оказались метод k-ближайшего соседа (KNeighborsClassifier) и метод градиентного спуска (XGBClassifier). Для столбца «Отметка за предыдущий период» наиболее важными признаками стал процент выполнения работы ВПР. Для столбца «Тип ОО» все представленные алгоритмы дали хорошие результаты и наиболее важным признаком стал тип изучаемой программы в школах. Это говорит о том, что для каждого типа образовательной организации (школа, гимназия и др.) представлены разные программы обучения. Для столбца «Учебник» наиболее важными признаками являются школа и тип образовательной организации.

5. Для построения дерева решений использовался класс DecisionTreeClassifier из библиотеки sklearn, который принимает на вход обучающую выборку X и метки классов целевой переменной Y. После чего идет построение прогнозирования целевой переменной на тестовой выборке с помощью созданной модели. Для этого используется функция predict(). Для оценки качества построенной модели использовалась функция classification_report из библиотеки

sklearn, в качестве параметров которой были переданы вектор целевой переменной и его предсказанные значения.

6. Для построения модели по методу «Случайного леса» использовался класс RandomForestClassifier из библиотеки sklearn.

**Результаты**

Проведение описательного анализа обеспечило количественное описание числовых данных. На рисунке 1 приведена таблица с описательным анализом данных по каждому столбцу, имеющему числовой тип данных. Значение count определяет размер выборки (количество записей в каждом столбце, отличных от nan). Min и max определяют минимальное и максимальное значения соответственно по каждому столбцу. Показатель максимального значения столбца «Разница отметок», равный 3, говорит о том, что ребенок за четверть имеет оценку 5, а за работу ВПР — 2, что говорит о сильном завышении оценок в школе. Mean определяет среднее арифметическое значение, рассчитанное как разница суммы всех значений к их общему числу.

| | Вариант | Отметка за предыдущий период | Итого баллов | Оценка за ВПР | Процент выполнения | Разница отметок |
|---|---|---|---|---|---|---|
| count | 110371.000000 | 110224.000000 | 110371.000000 | 110371.000000 | 110371.000000 | 110224.000000 |
| mean | 1.489150 | 3.680260 | 26.612543 | 3.307508 | 57.106957 | 0.372215 |
| std | 0.499885 | 0.672064 | 9.901293 | 0.861797 | 20.767880 | 0.653084 |
| min | 1.000000 | 2.000000 | 0.000000 | 2.000000 | 0.000000 | -3.000000 |
| 25% | 1.000000 | 3.000000 | 20.000000 | 3.000000 | 46.000000 | 0.000000 |
| 50% | 1.000000 | 4.000000 | 27.000000 | 3.000000 | 57.000000 | 0.000000 |
| 75% | 2.000000 | 4.000000 | 33.000000 | 4.000000 | 72.000000 | 1.000000 |
| max | 2.000000 | 5.000000 | 51.000000 | 5.000000 | 100.000000 | 3.000000 |

Рис. 1. Описательный анализ данных по числовым столбцам

По таблице можно сказать, что средний процент выполнения заданий по всему краю равен 57,1 из 100 (по максимальному значению), средняя оценка за ВПР равна 3,3 балла (из 5), а отметка за четверть значительно выше оценки за ВПР — 3,7 балла (из 5), что также может говорить о признаке завышения оценок в школах, а среднее значение варианта, равное 1,5, показывает, что работ с обоими вариантами распределено поровну.

Показатель максимального значения столбца «Разница отметок», равный 3, говорит о том, что ребенок за четверть имеет оценку 5, а за работу ВПР — 2, что говорит о сильном завышении оценок в школе. Mean определяет среднее арифметическое значение, рассчитанное как разница суммы всех значений к их общему числу.

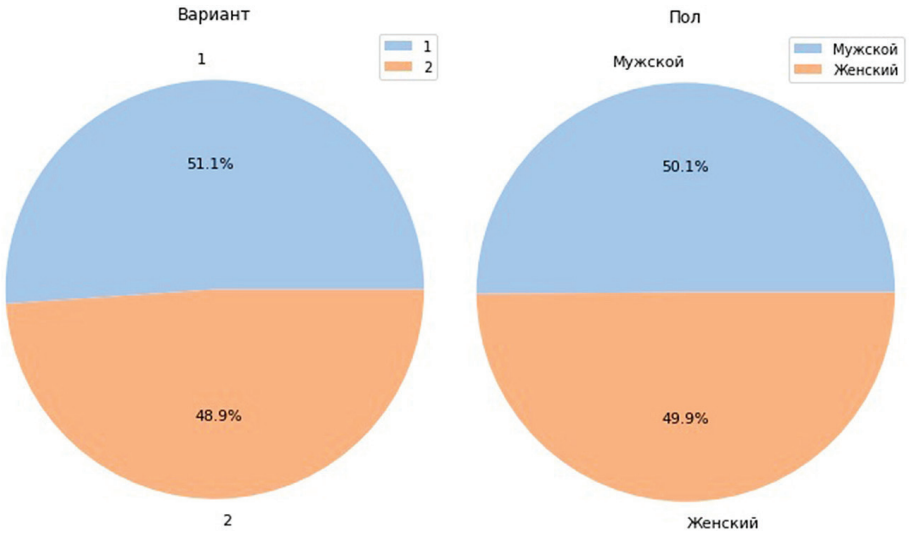


Рис. 2. Диаграммы распределения «Пол», «Вариант»

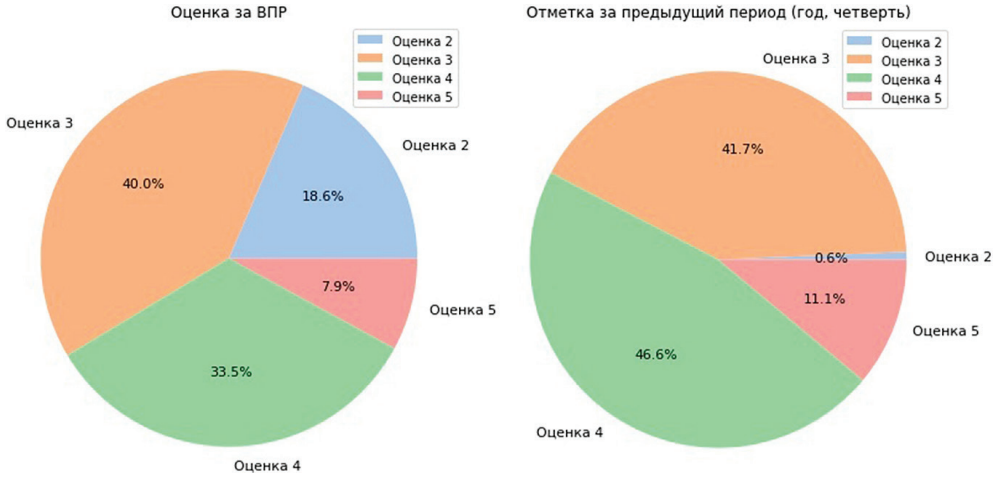


Рис. 3. Диаграммы распределения «Оценка за ВПР», «Отметка за четверть / полугодие»

На основе визуального анализа (рис. 2, 3) были получены следующие результаты:

1. Число мальчиков и девочек, принявших участие в ВПР, одинаково.

2. Двоек, полученных за работу, значительно больше, чем поставленных за предыдущий период, а четверок и пятерок — меньше.

3. Успеваемость в школе, а также результаты за работу по ВПР у девочек лучше, чем у мальчиков.

4. Количество двоек значительно превышает количество пятерок, а за четверть — наоборот, что говорит о том, что в школах оценки завышены.

Для нахождения классов, по которым могут разбиваться данные, была применена кластеризация. Сравнение методов кластеризации по оценкам метрик представлено в таблице 1.

*Таблица 1*

Оценка качества построенных моделей кластеризации различными методами

| Алгоритм | Число кластеров | ARI | AMI | V-measure | Silhouette | Значимый столбец |
|---|---|---|---|---|---|---|
| KMeans | 3 | 0,72 | 0,67 | 0,67 | 0.54 | Оценка за ВПР |
| Agglomerative Clustering | 3 | 0,81 | 0,88 | 0,88 | 0,53 | Оценка за ВПР |
| Bisecting KMeans | 10 | 0.43 | 0.76 | 0.76 | 0,41 | Процент выполнения |

Лучшим алгоритмом оказался алгоритм агломеративной кластеризации, показавший наибольшие оценки по всем метрикам. Важным признаком для предсказания целевой переменной стала оценка за ВПР.

Задача классификации применялась для нахождения признаков, по которым были разбиты данные. В качестве класса была взята целевая переменная «Отметка за предыдущий период». Данные были разбиты на 4 группы в соответствии с отметкой, полученной исходя из итогового балла. Указанный атрибут показывает итоговый результат освоенности школьной программы по предмету. В таблице 2 представлено сравнение методов классификации по оценке F1 и важному признаку по русскому языку, а в таблице 3 — по математике.

*Таблица 2*

Оценки построенных моделей классификации по столбцу «Отметка за предыдущий период» по русскому языку

| Метод | Оценка F1 | Важный признак |
|---|---|---|
| LogisticRegression | 0,65 | – |
| DecisionTreeClassifier | 0,67 | Процент выполнения |
| KNeighborsClassifier | 0,72 | – |
| RandomForestClassifier | 0,69 | Код школы |
| XGBClassifier | 0,74 | Процент выполнения |

*Таблица 3*

Оценки построенных моделей классификации по столбцу «Отметка за предыдущий период» по математике

| Метод | Оценка F1 | Важный признак |
|---|---|---|
| LogisticRegression | 0,63 | – |
| DecisionTreeClassifier | 0,66 | Процент выполнения |
| KNeighborsClassifier | 0,68 | – |
| RandomForestClassifier | 0,68 | Код школы |
| XGBClassifier | 0,71 | Процент выполнения |

Исходя из полученных результатов, можем сделать вывод, что для многоклассовой классификации лучше всего работают ансамблевые методы, а также метод k-ближайших соседей.

Задача регрессии схожа с задачей классификации, только в качестве целевой переменной идет не категориальная, а количественная величина. В качестве целевого значения был использован столбец «Процент выполнения», принимающий значения от 0 до 100.

В качестве предикторов были взяты остальные столбцы таблицы за исключением целевого значения и столбцы «Итого баллов» и «Оценка за ВПР», явно влияющие на целевую переменную. Результаты оценок построенной модели по целевой переменной «Процент выполнения» на основе русского языка представлены в таблице 4.

*Таблица 4*

Оценки построенных моделей регрессии по столбцу «Процент выполнения» по русскому языку

| Метод | MAE | MSE | RMSE | R2 | Важный признак |
|---|---|---|---|---|---|
| LinearRegression | 5,59 | 50,61 | 7,11 | 0,86 | – |
| DecisionTree Regressor | 6,36 | 71,88 | 8,48 | 0,80 | Оценка за четверть |
| KNeighbors Regressor | 14,04 | 314,48 | 17,73 | 0,14 | – |
| RandomForest Regressor | 5,05 | 41,17 | 6,42 | 0,89 | Оценка за четверть |
| XGBRegressor | 4,98 | 38,43 | 6,20 | 0,90 | Оценка за четверть |

Плохую оценку показал метод k-ближайшего соседа (KNeighbors Regressor). Лучшим методом оказался метод градиентного спуска (XGBRegressor), показавший наименьшие ошибки и наибольшую оценку R2. Важным признаком для предсказания целевой переменной стала отметка за четверть, полученная в школе.

**Заключение**

В результате проведенного исследования получены оценки качества подготовки обучающихся методами интеллектуального анализа данных, выявлены наиболее зависимые закономерности. Разработанная программа может служить основой формирования плана исследования в задачах оценки качества подготовки обучающихся на базе анализа результатов ВПР, а также для обоснования рекомендаций на основе количественных оценок, полученных в результате проведенного анализа.

## Библиографический список

1. Методология и критерии оценки качества общего образования в общеобразовательных организациях на основе практики международных исследований качества подготовки обучающихся (утв. приказами Рособрнадзора № 590, Минпросвещения России № 219 от 06.05.2019) (ред. от 11.05.2022) // сайт Консультант плюс. URL: https://www.consultant.ru/document/cons_doc_LAW_325095/ (дата обращения: 18.11.2023).

2. Сергеева С.Ю., Обревко Е.Д. Современные подходы и методы оценки качества образования // Молодой ученый. 2019. № 37 (275). С. 162–165.

3. Соловьев И.В., Филатов С.В. Интегральные оценки качества образования // ИТС. 2014. № 2 (75). С. 14–18.

4. Рямов Р.Ф. Оценка качества образования — инструмент воздействия на развитие системы образования // Вестник Башкирск. ун-та. 2012. № 3. С. 1423–1425.

5. О проведении Федеральной службой по надзору в сфере образования и науки мониторинга качества подготовки обучающихся общеобразовательных организаций в форме всероссийских проверочных работ в 2023 году : Приказ Федеральной службы по надзору в сфере образования и науки от 23.12.2022 г. № 1282. // сайт Федеральной службы по надзору в сфере образования и науки. URL: https://obrnadzor.gov.ru/wp-content/uploads/2023/01/1282.pdf (дата обращения: 18.11.2023).

6. Официальные данные результатов выпускных проверочных работ по Алтайскому краю // сайт Системы аналитики Всероссийских проверочных работ в Алтайском крае. URL: https://stat.22edu.ru/ (дата обращения: 18.11.2023).

7. Маккинни У. Python и анализ данных. М.: ДМК Пресс, 2020. 540 с.

8. Скиена Стивен С. Наука о данных. СПб.: ООО «Диалектика», 2020. 544 с.

9. Абдрахманов М.И. Визуализация данных. Matplotlib. Seaborn. Mayavi, 2020. 412 с.

10. Баймуратов И.Р. Методы автоматизации машинного обучения СПб.: Университет ИТМО, 2020. 40 с.

11. Брантон С.Л., Куц Дж.Н. Анализ данных в науке и технике. М.: ДМК Пресс,, 2021. 574 с.

## References

1. Methodology and Criteria For Assessing The Quality Of General Education In General Education Organizations Based On The Practice Of International Research On The Quality Of Student Training. Edition from 11.05.2022). *Consultant Plus website*. URL: https://www.consultant.ru/document/cons_doc_LAW_325095/ (accessed: 18.11.2023). (In Russ.).

2. Sergeeva S.Yu. Obrevko E.D. Modern Approaches And Methods For Assessing The Quality Of Education. *Molodoj Uchenyj*. 2019. No 37 (275). P. 162–165. (In Russ.).

3. Solov'ev I.V., Filatov S.V. Integral Assessments of The Quality of Education. *ITS*. 2014. No 2 (75). P. 14-18. (In Russ.).

4. Ryamov R.F. Assessing the Quality Of Education Is A Tool For Influencing The Development Of The Education System. *Vestnik Bashkir University*. 2012. No 3. P. 1423–1425. (In Russ.).

5. Order of the Federal Service for Supervision in the Sphere of Education and Science of December 23, 2022 No. 1282 "On the Monitoring by the Federal Service for Supervision in the Sphere Of Education and Science of the Quality of Training of Students in General Education Organizations in the Form of All-Russian Testing in 2023". *Website Of The Federal Service For Supervision Of Education and Science*. URL: https://obrnadzor.gov.ru/wp-content/uploads/2023/01/1282.pdf (accessed: 18.11.2023). (In Russ.).

6. Official Data On The Results Of Final Examinations in the Altai Territory. *Website of the Analytics System of the All-Russian Test Works in the Altai Territory*. URL: https://stat.22edu.ru/ (accessed: 18.11.2023). (In Russ.).

7. Makkinni U. *Python and Data Analysis*. Moscow: DMK Press, 2020. 540 p. (In Russ.).

8. Skiena Stiven S. *Data Science*. Saint Petersburg: OOO «Dialektika», 2020. 544 p. (In Russ.).

9. Abdrahmanov M.I. *Data Visualization. Matplotlib. Seaborn*. Mayavi, 2020. 412 p. (In Russ.).

10. Bajmuratov I.R. *Machine Learning Automation Methods*. Saint Petersburg: Universitet ITMO, 2020. 40 p. (In Russ.).

11. Branton S.L., Kuts J. N. *Data Analysis in Science and Technology*. Moscow: DMK Press, 2021. 574 p. (In Russ.).

***Информация об авторе***

**Т.В. Михеева,** кандидат технических наук, доцент кафедры информатики, Алтайский государственный университет, Барнаул, Россия.

***Information about the author***

**T.V. Mikheeva,** Candidate of Sciences in Technology, Associate Professor of the Department of Computer Science, Altai State University, Barnaul, Russia.