Научная статья
УДК 338.27
DOI: 10.25683/VOLBI.2025.70.1203

Alexander Vladimirovich Paraskevov
Senior Lecturer of the Department
of Computer Technologies and Systems,
Kuban State Agrarian University
Krasnodar, Russian Federation
paraskevov.alexander@yandex.ru

Александр Владимирович Параскевов
старший преподаватель кафедры
компьютерных технологий и систем,
Кубанский государственный аграрный университет
Краснодар, Российская Федерация
paraskevov.alexander@yandex.ru

Alfira Menligulovna Kumratova
Candidate of Economics,
Associate Professor of the Department of Information Systems,
Kuban State Agrarian University
Krasnodar, Russian Federation
kumratova.a@edu.kubsau.ru

Альфира Менлигуловна Кумратова
канд. экон. наук,
доцент кафедры информационных систем,
Кубанский государственный аграрный университет
Краснодар, Российская Федерация
kumratova.a@edu.kubsau.ru

ЦИФРОВОЙ АНАЛИЗ И ПРОГНОЗИРОВАНИЕ БОЛЬШИХ ДАННЫХ ОБРАЗОВАТЕЛЬНОГО ПРОЦЕССА НА БАЗЕ ПЛАТФОРМЫ LOGINOM

5.2.2 — Математические, статистические и инструментальные методы в экономике

Аннотация. *Большие данные окружают пользователей вне зависимости от их желания. Предприятия любого рода деятельности зачастую генерируют их в огромном объеме, при этом в среднем только каждое третье пытается их обрабатывать, при благоприятном стечении событий цифра достигает 35,5 %, и это лишь за счет сфер информационных технологий, конструкторских бюро и продаж. Авторами была поставлена цель провести исследование приемной кампании высшего учебного заведения на основе сгенерированных данных. Данные максимально приблизить к реальным за счет соответствия показателей конкурса, проходного балла, процентного соотношения абитуриентов из городов и районов и других показателей. Дизайн исследования: результаты и весь ход исследования основывался на сгенерированных данных, но, несмотря на это, результаты представляют подлинный научный и практический интерес — они демонстрируют методику оценки показателей деятельности вне зависимости от направления, форм собственности и сезонности работы. По итогам проведенного исследования сгенерирован репрезентативный набор данных (датасет), проведены операции отбора, кодирования и нормализации данных. Полученные данные максимально приближены к реальным за счет использования реальных показателей проходного балла и количества поданных заявлений. Реальные данные взяты из открытых источников и являются обезличенными. Согласно результатам исследования были выявлены комбинации второстепенных факторов, которые влияют на поступление абитуриента в высшее учебное заведение. Определены ключевые характеристики целевой аудитории учебного заведения. Обучена нейросеть и реализована возможность прогнозирования результатов следующих приемных кампаний. Методы анализа больших данных и машинное обучение играют зачастую решающую роль в определении квазиоптимальных параметров функционирования как отдельных процессов, так и производств в целом. При этом важно соблюдать принцип вариативности подходов. Кластеризация промежуточных результатов или исходных групп помогает добиться значительных успехов. Это определяет новый взгляд на существую систему вне зависимости от отрасли.*

Ключевые слова: *высшее образование, большие данные, методика оценки, анализ данных, датасет, зависимости, машинное обучение, исследование зависимостей, настройка параметров, аналитика, нейросети*

Для цитирования: Параскевов А. В., Кумратова А. М. Цифровой анализ и прогнозирование больших данных образовательного процесса на базе платформы Loginom // Бизнес. Образование. Право. 2025. № 1(70). С. 58—65. DOI: 10.25683/VOLBI.2025.70.1203.

Original article

DIGITAL ANALYSIS AND FORECASTING OF BIG DATA OF THE EDUCATIONAL PROCESS BASED ON THE LOGINOM PLATFORM

5.2.2 — Mathematical, statistical and instrumental methods in economics

Abstract. *Big data surrounds users regardless of their desire. Enterprises of any kind of activity often generate them in a huge volume. At the same time, on average, only one in three tries to process them. With a favorable combination of events, the figure reaches 35.5 %. And this is only due to the areas of information technology, design bureaus and sales. The authors aimed to conduct a study of the admission campaign of a higher educational institution based on the generated data. The data had to be brought as close as possible to the real ones due to the correspondence of the competition indicators, the passing score, the percentage of applicants from cities and districts, and other indicators. The results and the entire course of the study were based on the data*

generated, but despite this, the results are of genuine scientific and practical interest. They demonstrate a methodology for evaluating performance indicators, regardless of the direction, forms of ownership and seasonality of work. Based on the results of the study, a representative data set was generated; data selection, encoding and normalization operations were performed. The data obtained are as close as possible to the real ones due to the use of real indicators of the passing score and the number of applications submitted. The real data was taken from open sources and is depersonalized. According to the results of the study, combinations of secondary factors that affect the applicant's admission to higher education were identified. The key characteristics of the target audience of the educational institution were determined. A neural network was trained and the ability to predict the results of the next admission campaigns was implemented. Big data analysis methods and machine learning often play a crucial role in determining the quasi-optimal parameters of the functioning of both individual processes and industries as a whole. At the same time, it is important to observe the principle of variability of approaches. Clustering intermediate results or initial groups helps to achieve significant success. This defines a new look at the existing system, regardless of the industry.

Keywords: *higher education, big data, assessment methodology, data analysis, data set, dependencies, machine learning, dependency research, parameter setting, analytics, neural networks*

For citation: Paraskevov A. V., Kumratova A. M. Digital analysis and forecasting of big data of the educational process based on the Loginom platform. *Biznes. Obrazovanie. Pravo = Business. Education. Law*. 2025;1(70):58—65. DOI: 10.25683/VOLBI.2025.70.1203.

Введение

Актуальность. Накопление данных происходит вне зависимости от желания человека. Это стало своего рода требованием окружающей действительности. А вот решение обрабатывать их или нет — это уже вопрос современности и актуальности подходов руководящего звена организации. При этом важно учесть правильную организацию хранения данных, при котором они должны соответствовать требованию минимального структурирования. Тогда обработка данных будет происходить максимально быстро и возможное «озеро данных» не превратится в «болото». При этом необходимость обработки данных столь очевидна, сколько и необходима. Она в состоянии помочь выявить неочевидные связи и закономерности, дать объяснение многим, на вид стохастическим, процессам, помочь принять обоснованные решение в задачах оптимизации производственных процессов. Основная особенность состоит в том, что подходы к обработке данных достаточно унифицированы. При этом практически полностью отсутствует привязка к отраслевой принадлежности. Иными словами, при анализе очень малую роль играет фактор происхождения данных. Важно, что они есть, они достаточны, достоверны, полны.

По данным опроса *Tech Pro Research*, только 15 % учреждений сферы образования (при оптимальном раскладе событий) занимаются обработкой больших данных, которые сами генерируют. При этом, если оставаться бинарными, то «затрудняюсь ответить» относится к категории «нет».

Суммируя результаты, можно отметить, что последними, кто обрабатывает большие данные, являются отрасли здравоохранения и образования. При этом существующие в открытом доступе данные (источники *Google Cloud Public Datasets*, *Data.gov*, *Kaggle*, *Global Health Observatory*, *UCI Machine Learning Repository* и др.) у этих направлений являются одними из самых полных (опережает, разве что, большой спорт). Тенденция настораживает.

Кластеризация в области организации дорожного движения позволяет определить группы перекрестков или участков уличной дорожной сети. Затем разработать и применить к ним шаблонные сценарии улучшения плавности потоков, устранения заторов, стимулирования развития системы городского общественного транспорта. В данном случае возможны разнообразные подходы, в т. ч. позволяющие отталкиваться от кластеризации не только самих участков городской транспортной сети, но и районов проживания, категоризация групп населения по транспортным потребностям и классам обеспеченности.

Степень научной разработанности проблемы. Проблемами и вопросами продвижения образовательных услуг посвящены работы Д.А. Шевченко [1], Е. А. Неретиной и А. Б. Макарец, З. С. Жиркова, М. Г. Чардымского и др. В работах В. Ф. Хорошевского рассматриваются особенности организации пространства знаний в Интернет, методы и средства извлечения знаний, а также вопросы использования пространств знаний при создании прикладных интеллектуальных систем.

Роль коммуникаций для высших учебных заведений, в особенности целевых и персонализированных, рассматриваются в работах G. R. Maio, G. S. Linoff, M. J. Berry. Использованию методов интеллектуального анализа данных для оценок работы образовательных сайтов с точки зрения «юзабилити» посвящены работы С. Риза [2].

В работах вышеуказанных ученых рассматриваются вопросы, связанные с повышением имиджа высших учебных заведений, однако не рассматриваются вопросы организации эффективных коммуникаций между вузами и потребителями образовательных услуг, не исследуются проблемы выявления и анализа потребностей будущих потребителей образовательных услуг — абитуриентов. Вопросами повышения конкурентного преимущества вуза в современных условиях занимаются такие ученые, как Е. В. Гугнина [3], И. Ю. Окольнишникова и др. Проблематику продвижения качества предоставления образовательных услуг в своих работах рассматривают М. В. Самсонова, Н. А. Завалько, Н. А. Селезнева и т. д. Вопросы, связанные с разработкой и внедрением образовательных порталов, рассматривают в своих работах В. С. Самсонов, М. П. Лапчик, В. А. Касторнова, С. А. Дочкин, А. Н. Сергеев и т. д. Работы данных ученых посвящены использованию образовательных порталов вузов и существующих методов продвижения образовательной деятельности вуза, но в них не рассматриваются вопросы повышения эффективности процессов продвижения образовательных услуг.

Научная новизна заключается в разработке и применении методики обработки больших данных в избранной отрасли для поиска значимых корреляций как среди отдельных факторов, так и групп переменных, при отбросе основного влияющего фактора.

Целесообразность разработки темы связана с актуальностью темы исследования и различными результатами исследований по обоснованию взаимосвязи между группами переменных.

Цель исследования — провести исследование сгенерированного набора данных для выявления значимых корреляций, а также определения как возможности и целесообразности проведения анализа больших данных в избранной отрасли, так и работоспособности методики в целом.

В качестве **задач** исследования необходимо обозначить следующие:

1. Определение возможности искусственного создания репрезентативного датасета для исследования области.
2. Создание датасета, его нормализация, выполнение операций кодирования и проверка.
3. Определить работоспособность подхода для анализа данных в избранной отрасли.
4. Провести анализ генерированных данных, составить карты Кохонена.
5. Сделать вывод о возможности применения методики.

Теоретическая значимость работы заключена в возможности масштабирования подхода к анализу данных и машинному обучению безотносительно отрасли.

Практическая значимость проводимого исследования применительно к системе вступительных испытаний высшего образования (бакалавриат) заключается в возможности реализации подхода, доказательстве его работоспособности и возможности помочь в определении взаимодействующих переменных. Кластеризация абитуриентов по группам сданных экзаменов, экзаменов по факту поступления, а также нахождение группы параметров помогают увидеть скрытые зависимости. Также определяется привлекательность тех или иных направлений подготовки для абитуриентов.

Основная часть

Методология исследования. В общем случае подход можно унифицировать и свести к определенному алгоритму. Этот алгоритм будет в самых общих чертах схож с парадигмой *MapReduce* (рис. 1).

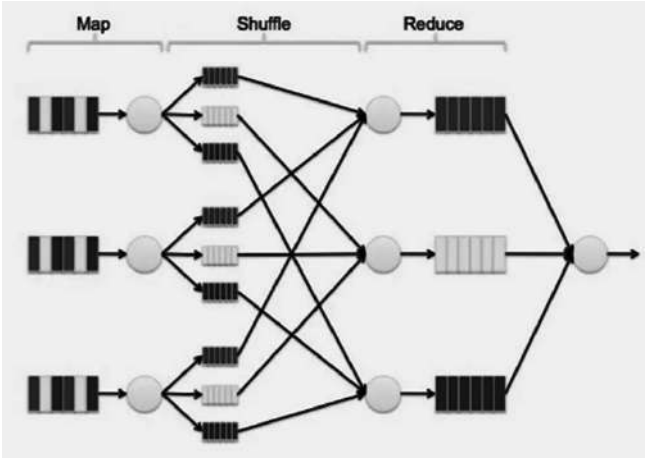

Рис. 1. Парадигма *MapReduce*

В общем случае парадигма состоит из трех стадий: *Map*, *Shuffle* и *Reduce*. На первой стадии, *Map*, производится отбор данных. Рассмотрим процесс работы подхода путем составления алгоритма оценки данных из двух отраслей: высшее образование и производство (косвенно, даже и реализация) продукции. В первую очередь это представляется важным за счет того, что оба направления не связаны между собой, общих черт в них настолько мало, что алгоритм может претендовать на универсальность. На данном этапе его можно называть эвристическим.

Под эвристическим алгоритмом будем понимать последовательность действий, приводящую к верному, с точки зрения практики, результату, но не доказано, что он выдает верные ответы на всем поле входных данных.

Первый этап заключается в сборе и подготовке данных. Сбор данных является одним из ключевых параметров. Недостоверные данные могут приводить к ложным результатам и ошибочным выводам. Данные должны содержать множество полей и записей. Дело в том, что выводы и зависимости могут проявляться самым неожиданным образом. Момент избыточности данных может быть скорее исключением, чем вероятным событием. Нет достоверного представления о возможных парах зависимостей и значений до проведения анализа. Каждое из полей будет являться переменной и в том или ином плане участвовать в поиске решений. Результатом могут быть не только сами переменные, но и их комбинации. Что, в свою очередь, даст возможность посмотреть на комбинации факторов [4].

В сфере высшего образования сбор данных представляется проблематичным, если только учебное заведение само не участвует в разработке тематики. Это связано с полем деятельности Федерального закона от 27 июля 2006 г. 152-ФЗ «О персональных данных». Он говорит нам о том, что абитуриент должен дать разрешение на обработку своих данных университету. При подаче документов для абитуриентов это разрешение является процессом естественным и не выходящим за поле логики событий. В тоже самое время у высших учебных заведений, зачастую, отсутствует исследовательский интерес в данной области. Хотя возможен юридически чистый подход, при котором данные предварительно обезличиваются. Федеральный закон говорит нам, в первом приближении, о том, что нельзя без разрешения обрабатывать и хранить данные, позволяющие однозначно идентифицировать объект.

В этой связи предварительное обезличивание данных представляется замечательным подходом. Необходимо удалить поля с именем, фамилией и отчеством, также с паспортными данными, СНИЛС. Адрес места жительства сокращают до названия населенного пункта, а название улицы и номер дома также удаляют. Для различия записей между собой их просто кодируют, т. е. присваивают персональный идентификационный номер. На выходе мы получаем обезличенные данные об абитуриентах. Если посмотреть на крупные вузы, вспомнить как давно они собирают эти данные в электронном виде, то получится очень приличный, гарантированно репрезентативный датасет. Его особенность будет также в том, что с применением «обучения с учителем» можно правильно настроить не только сами зависимости, но и оптимизировать функции штрафа. А именно они (функции штрафа) помогают спасти модель от недообучения и переобучения [5].

При отсутствии требуемых данных существует единственный способ — генерирование данных. Это связано с множеством трудностей и условностей, но, тем не менее, этот подход видится адекватным для того, чтобы показать работоспособность модели. Доказать, что при несложной перенастройке параметров она сможет работать на реальных данных и выдавать адекватные результаты.

При генерировании данных необходимо учитывать множество параметров: разброс значений возраста, соотношение возрастных групп, соответствие данных по зачислению на I курс (как количественных — контингент студентов, так и качественных, например соотношение поступивших

из городов к остальным населенным пунктам, выбросы в возрастных значениях: имеются ввиду поступившие после армии или средних специальных учебных заведений и другие тонкие моменты), конкурс, проходные баллы и пр. В области производства продукции (далее конкретизировать не представляется необходимым — весь подход легко перестраивается и адаптируется, например, из производства подшипников в производство текстильной продукции) ситуация с данными и датасетами обратная. Здесь производственные предприятия с радостью поделятся ими (безусловно при соблюдении коммерческой тайны) для нахождения зависимостей и возможности тонкой подстройки не только производственных процессов, закупок, штатной численности и прочего, но и областей сбыта конечной продукции.

Технологические процессы на производстве зависят от большего количества факторов, чем в предыдущем примере. Здесь влияют как состояние оборудования, производственной среды, качество материалов от поставщиков, современность оборудования, состояние и настройка цепочек поставки и реализации конечного товара, так и от других факторов, например, уровень социально-экономического развития региона, уровень зарплат, психологическое состояние коллектива, социальные факторы, внешнеполитическая ситуация и пр. [6].

При настройке параметров на производстве риски существенно более велики. Каждая неудачная итерация производственного цикла стоит дорого. Итоговые затраты могут стать слишком велики и приведут к закрытию производства ввиду непосильной величины издержек. Фильтрация данных представляет собой процесс соответствия определенному критерию. Сюда можно отнести также методы удаления пропущенных значений, дубликатов данных и неконсистентных данных. Всё это приводит к тому, что данные становятся грязными. Работа с ними приводит к некоррелирующим результатам. В случае с пропущенными значениями нужно определиться с важностью фактора, его возможной ролью в общем процессе и принять решение: есть ли необходимость восстанавливать пропущенные значения, или стоит полностью удалить переменную. В большинстве случаев при анализе данных недопустимо пренебрегать переменными. Они могут оказывать существенную роль не только сами по себе, но и при группировке нескольких значений. Процесс восстановления данных называется «импутация». Существует несколько способов импутации данных.

1. Не делать ничего и надеяться на алгоритм обработки. При этом совершенно неизвестно как он обойдется с данными и чем заполнит пропуски. Минусов здесь слишком много, чтобы их перечислять.

2. Использование медианы или средней. Метод не точный, не учитывается возможная корреляция между параметрами и отсутствует погрешность, подходит только для количественных значений.

3. Импутация данных самым частым значением или константой. Неточный метод, но возможно применение для качественных значений.

4. Метод *k-NN* (метод ближайших соседей). Взвешенное среднее от количества указанных соседей встраивается в ряд недостающих значений. Большая вычислительная сложность, потому что требует держать в памяти весь датасет, количество соседей нужно подбирать, присутствует чувствительность к выбросам. Но он существенно эффективнее всех предыдущих.

5. Метод *MICE*. Здесь происходит множественная импутация данных. Происходит повышение надежности за счет подтверждения или многократной корректировки предложенного изначально значения, а также возможности работы с разными типами данных. Подразумевает многократное повторение анализа на разных импутированных данных и интеграцию получившихся результатов (рис. 2).

6. Использование глубокого обучения. Точнее всех остальных методов. Вычислительно существенно дороже и в состоянии восстановить значения только по одной переменной. Если нужны другие — необходимо заново проходить процесс обучения. Здесь основным является вопрос целесообразности применения, т. е. существенной ли будет разница между этим методом и остальными. Остальные методы являются разнообразными «вариациями» на тему средних значений и *k-NN* метода [5].

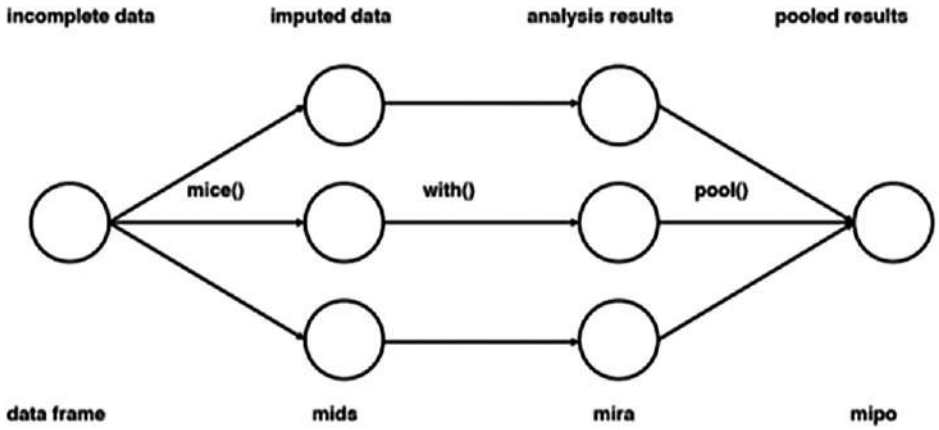

Рис. 2. Алгоритм импутации данных методом *MICE*

Изначально может показаться, что фактор, например, района проживания является несущественным при прогнозировании поступления. Но, если отбросить самый существенный фактор (а это отдельная методика, она играет большую роль, суть в том, что при наличии решающего сильного фактора он отбрасывается и происходит поиск комбинаций из остальных переменных) — сумму баллов, то окажется, что комбинации из нескольких других переменных существенно влияют на значение переменной [7]. В данном контексте необходимо отметить важность как корреляции факторов, так и обратной корреляции. В анализе данных не существует понятия «плохо», обратная

корреляция — это также влияющий фактор, просто наоборот. И им тоже необходимо управлять при настройке параметров. Для обработки данных можно использовать разнообразные инструменты. Одним из удобных является платформа *Loginom*, которая позволяет пользоваться своими решениями в режиме «условного *freeware*» [8]. Рассмотрим процесс обработки сгенерированных данных о системе высшего образования. Конкретно — данные абитуриентов (рис. 3).

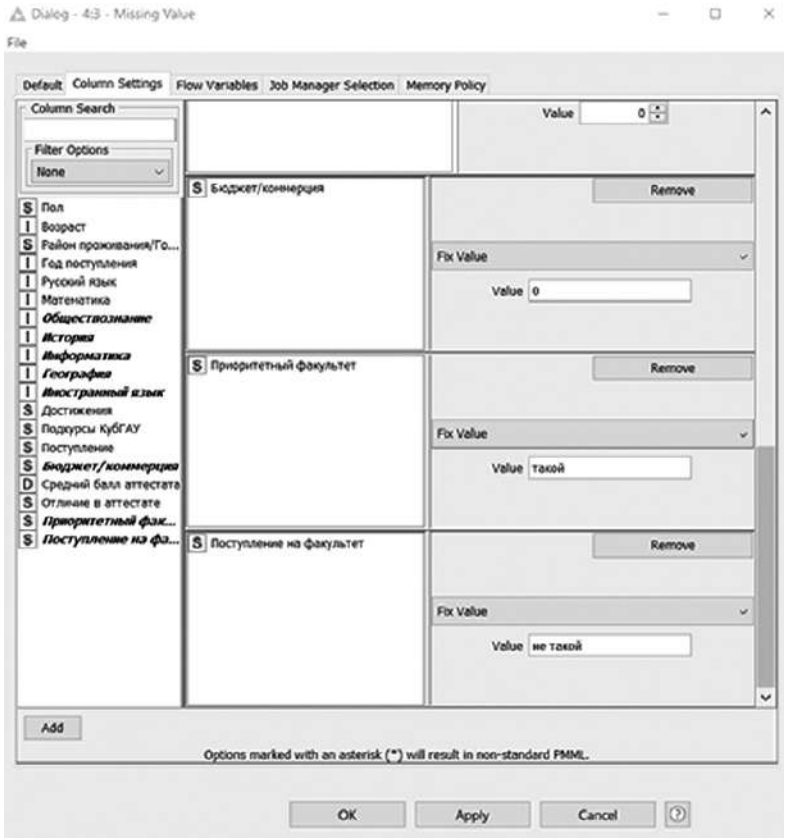

Рис. 3. Настройки *Missing Value*

Необходимо правильно подготовить данные. Обработаем пустые значения, чтобы при обработке строка данных не была пропущена, с помощью *Missing Value*. В настройках, во вкладке *Column Settings* выбрать поля:

– «Обществознание», «История», «Информатика», «География», «Иностранный язык», «Бюджет/коммерция» заменяем на 0;

– «Приоритетный факультет» — «такой»;

– «Поступление на факультет» — «не такой».

Посчитаем сумму баллов сданных экзаменов с помощью *Math Formula*. Это понадобится для дальнейшего анализа. Добавляем новый столбец «Сумма баллов», задавая тип *int* [9].

Затем строковые необходимо закодировать в числовые значения, т. к. текстовые данные мешают получению корректных результатов. Используем *String Manipulation* для каждого столбца отдельно, данные будем замещать:

– «Пол» — toInt(replace(replace($Пол$, «Мужской», «1»), «Женский», «0»));

– «Бюджет/коммерция» — toInt(replace(replace($Бюджет/коммерция$, «Бюджет», «2»), «Коммерция», «1»))

«Достижения», «Подкурсы КубГАУ», «Отличие в аттестате», «Поступление» имеют схожие данные, поэтому к ним с помощью *String Manipulation (Multi Column)* применим — toInt(replace(replace($$CURRENTCOLUMN$$,”Да”, “1”), “Нет”, “0”)). Нужно получить данные о том, поступил ли человек на желаемый факультет. Для этого сравним данные в колонках с помощью того же *String Manipulation* [10].

Результаты добавим в новый столбец «Поступление на желаемый факультет». Если человек поступил на приоритетный факультет, то будет «0», иначе «–1», «1». Теперь правильно закодируем результаты с помощью *Math Formula*. Если значение «0», то поступил, соответственно поставим «1», иначе — «0». Теперь исключим столбцы «Приоритетный факультет» и «Поступление на факультет» из набора, т. к. мы их отработали с помощью *Table Manipulator*.

Результаты исследования. Остается последнее, кодировка «Район проживания/город». Для этого существует *csv*-файл, в котором районам сопоставлено значение «0», а городам — «1». Для чтения используем *CSV Reader*. Зададим путь к файлу, уберем галочку с *Has column header*, во вкладке «Transformation» зададим названия колонкам «Город», «Кодировка». В результате получим скрипт.

Он уже достаточно массивный, поэтому объединим узлы в метаузел. Назовем его *Data prep*. Добавим табличный выход метаузлу. Для этого нажать на «+» справа и выбрать *Table*. Далее нажать правой кнопкой мыши — *Metanode — Open metanode* и подвести выход последнего узла к выходу из метаузла. Поскольку *KNIME* не имеет в себе такой интересный инструмент как самоорганизующиеся карты Кохонена, необходимо использовать *Deductor Studio* [11]. Подготовленные данные необходимо выгрузить. Предварительно отсортируем данные при помощи узла *Sorter* по возрастанию значений в столбце «Год поступления» и разобьем на обучающую

и тестовую выборку по столбцу «Год поступления». Первой станет период с 2015 по 2022, а второй — 2023 г. Для удобства выходы перенесем во входы *Table Manipulator*, чтобы визуально отличать тестовую выборку от обучающей (рис. 4).

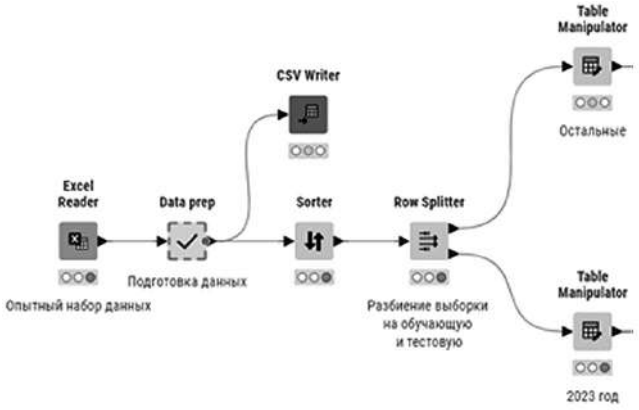

Рис. 4. Скрипт перед анализом данных

Теперь данные можно анализировать. Достаточно простым способ считается статистический анализ. Для этого используем выделенный набор данных за 2023 г. [12]. Соберем статистику всех поступавших за 2023 г. С помощью *Value Counter* посчитаем количество тех, кто поступил и не поступил. Это нужно для определения величины конкурса за место. Теперь следует раскодировать данные обратно, чтобы полученные результаты могли быть поняты другими.

Воспользуемся *String Manipulation*, заменив данные в столбце поступление на «Да» = 1 и «Нет» = 0. Для визуализации применим круговую диаграмму *Pie Chart*. Зададим категорию столбцу «Поступление», а значения по *count*. Демонстрация подписи с категорией, значением внутри диаграммы помогает разделять множества решений. Теперь узнаем, сколько человек выбрали тот или иной экзамен по выбору (обществознание, история, информатика, география, иностранный язык). Это нужно, чтобы определить популярность того или иного предмета среди поступающих. Для этого создадим новый столбец «Выбранный экзамен», в котором закодируем каждый из экзаменов от 1 до 5. Используемый узел — *Math Formula*.

Затем используем *Value Counter*, чтобы посчитать количество записей по значениям в столбце «Выбранный экзамен», и раскодируем данные в новом столбце «Предмет», применив *String Manipulation*. Изучим набор данных поступивших абитуриентов. Также, как и для набора поступающих, соберем надо получить данные о сдаваемых дисциплинах [13]. Дополнительно посчитаем людей, которые родом из городов/районов, используя *Value Counter* по столбцу «Город/Район». Раскодируем данные. Визуализация с помощью круговой диаграммы дает наглядное понимание процесса. Также соберем данные по каждому объекту отдельно. Для этого посчитаем количество записей по значениям «Район/Город проживания». Столбец результирующего набора *RowID* не может быть применен в вычислениях, поэтому перенесем эти значения в новое поле *Row_Id* с помощью *String Manipulation*.

Необходимо раскодировать данные. Для этого нужен узел *Value Lookup*. Подведем на его вход ранее готовые данные из метаузла *Data prep*. Для этого создадим ему 2-й выход, откроем и подведем к выходу результирующую таблицу *Math Formula* с кодировкой от 1 до 63.

Визуализировать результаты можно двумя способами.

Применить узел *Bar Chart*.

Использовать *Table Manipulator*, чтобы переименовать столбец «count» на «Количество поступивших». Отобразить данные с *Table View* [14].

Теперь можно все диаграммы и таблицы объединить в дашборд. Для этого выделить их — правая кнопка мыши — «Create component» с названием *Dashboard*. Дашборд это интерактивная аналитическая панель, графический интерфейс. Смысл в том, что на одном экране расположены все ключевые метрики, показатели цели или процессов. С помощью этих метрик можно выявить и проанализировать тренды и изменения. К каждой диаграмме прописать комментарий, это поможет в дальнейшей навигации. В идеале — каждому узлу задавать описание. Открыть компонент и добавить 3 *Text Output Widget*. В их настройках прописать заголовки «Статистика поступавших/поступивших/не поступивших». Для улучшения понимания структуры данных и нахождения связей в данных построим карты Кохонена. Ранее записанный *csv*-файл нужно открыть в *Deductor Studio*. Для этого применить «Мастер импорта». Указать путь к файлу, формат исходных данных, символ-разделитель.

С помощью «Мастера обработки» к загруженным данным применить инструмент «Карты Кохонена» (рис. 5). Выбрать входные поля (факторы) и выходные (результаты). Первая карта для анализа общей картины поступающих, выбрать поля. Обучающее множество составляет 75 %, а тестовое — 25 %. Верхнее значение ошибки — 0,4, количество эпох — 600. Радиус обучения в начале — 2, в конце — 0,01 со скоростью 1 и 0,5 соответственно. Количество кластеров — 4. Начать обучение, нажать на «Пуск». После выбрать способ отражения — «Карта Кохонена», выбрать выходные столбцы (входные + выходные + кластеры) [15].

Затем исследовать данные с вступительными экзаменами, подготовительными курсами и поступлением. Выберем поля, обучающее множество составляет 75 %, а тестовое — 25%. Верхнее значение ошибки — 0,4, количество эпох — 600. Радиус обучения в начале — 1, в конце — 1 со скоростью 4 и 0,01 соответственно. Количество кластеров — 4.

Также исследуем набор данных, состоящий из города/района, достижений, среднего балла аттестата, отличие в аттестате, поступления, бюджета/коммерции. Обучающее множество — 75 %, а тестовое — 25 %. Верхнее значение ошибки — 0,3, количество эпох — 600. Радиус обучения в начале — 1, в конце — 1 со скоростью 4 и 0,01 соответственно. Количество кластеров — 4.

В результате получим три карты (рис. 6), позволяющие исследовать зависимости между значениями в разных полях (факторах). Кластеры — группа записей набора данных, схожих по значениям в некоторых полях, выбранных алгоритмом. С их помощью можно большое количество абитуриентов можно разделить на отдельные группы.

Затем с помощью полученной функции предскажем результаты для данных 2023 г. Для этого использовать *Regression Predictor*. Выходом будет являться набор с новым полем «Prediction (Поступление)». Оно содержит вероятность поступления. С помощью *Table Manipulator* оставим в наборе полученный столбец и поле «Поступление». Для сравнения следует определить значение вероятности, с которого поступление будет считаться «Да». В данном случае больше или равно 0,66. Использовать *Math Formula* для преобразования. Теперь можно оценить качество регрессии с помощью *Numeric Scorer*. Указать столбец с реальными и предсказанными значениями, дать название новому столбцу «Результат».

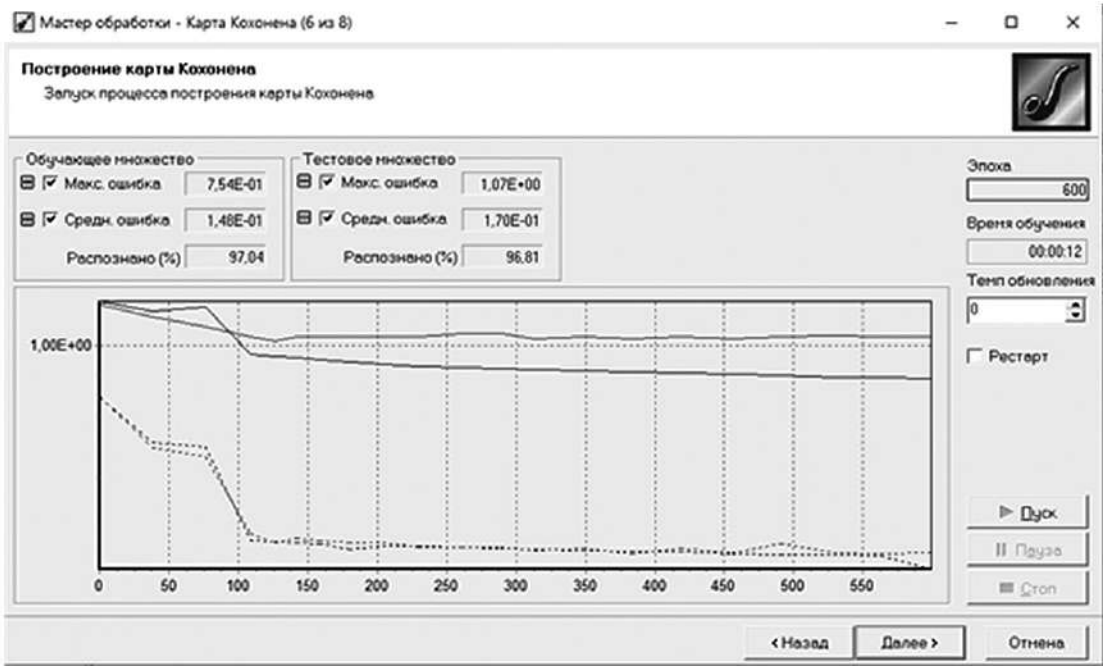

Рис. 5. График обучения

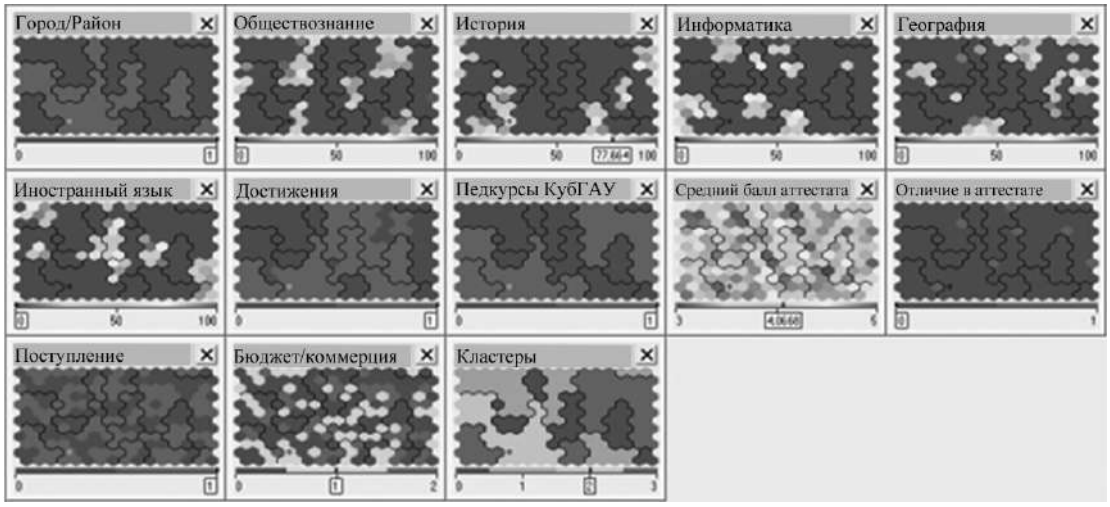

Рис. 6. Результат работы карт Кохонена

Заключение

В результате применения алгоритма получили обработанные результаты анализа приемной кампании в вуз. Они помогут при анализе данных абитуриентов, подведении итогов приемной кампании, формировании стратегических целей и изменении образа учебного заведения. В целом, анализ данных подтверждает способность находить скрытые закономерности, находить точки роста и узкие места в любой отрасли. Это абсолютно необходимый инструмент в продвижении бизнеса и многих аспектах социальной жизни. Правильное применение программных средств позволяет существенно сэкономить временные затраты и повысить точность прогнозирования и обучения. Важнее применения программных средств может быть только методика анализа данных. Она позволяет унифицировать работу с данными, проводить их анализ и корректировку, прогнозирование и импутацию.

СПИСОК ИСТОЧНИКОВ

1. Потребительское поведение / под общ. ред. Д. А. Шевченко. М. : Директ-Медиа, 2023. 112 с.
2. Риза С., Лезерсон У., Оуэн Ш., Уиллс Дж. Spark для профессионалов. Современные паттерны обработки больших данных. М. : Питер, 2023. 737 с.

3. Гугнина Е. В., Резниченко В. Е. Факторный анализ, его методы и виды // России — творческую молодежь : материалы XVI Всерос. науч.-практ. студ. конф. : в 4 т. Волгоград, 2023. Т. 3. С. 51—54.

4. Алексейчик А. Б. Экономический анализ в условиях компьютерной обработки данных: проблемы и направления развития // Актуальные проблемы правовых, экономических и гуманитарных наук : материалы XII Междунар. науч.-практ. конф. профессорско-преподават. состава, аспирантов, магистрантов и студентов. Минск, 2022. С. 60—62.

5. Бекназарова С. С. Анализ данных в распределенных информационных системах // Big Data и анализ высокого уровня : сб. науч. ст. VII Междунар. науч.-практ. конф. Минск, 2021. С. 59—66. (На англ. яз.)

6. Наим Н. А. Анализ и обработка данных // Big Data и анализ высокого уровня : сб. науч. ст. X Междунар. науч.-практ. конф. : в 2 ч. Минск, 2024. Ч. 1. С. 161—163. (На англ. яз.)

7. Галкина Е. В. Анализ потоков данных: применение в экономике и управлении // Вестник Алтайской академии экономики и права. 2021. № 10. Ч. 1. С. 16—20. DOI: 10.17513/vaael.1863.

8. Минин А. С. Сравнительный анализ методов кодирования категориальных данных в задачах линейной регрессии // Тенденции развития науки и образования. 2023. № 94. Ч. 5. С. 98—100. DOI: 10.18411/trnio-02-2023-259.

9. Ефимова Т. Б., Субочева Я. В., Глухова В. А. Интеллектуальный анализ данных в оценке персонала организации // Экономика и предпринимательство. 2023. № 8(157). С. 1395—1397. DOI: 10.34925/EIP.2023.157.8.263.

10. Помазков В. В. Анализ больших массивов данных как инструмент эффективности развития школы // Шамовские чтения : сб. ст. XVI Междунар. науч.-практ. конф. : в 2 т. Москва, 2024. Т. 1. С. 84—89.

11. Шишкина А. А. Сравнительный анализ данных для улучшения условий труда // Известия Тульского государственного университета. Технические науки. 2021. № 9. С. 250—252.

12. Кумратова А. М., Попова М. И. Методы и инструментальные средства визуализации для аналитики в малом бизнесе // Современная экономика: проблемы и решения. 2023. № 2(158). С. 91—98.

13. Параскевов А. В., Шаповалов А. В., Сергеев А. Э., Уварова А. Г. Подходы к анализу больших данных в системе высшего образования // Научный журнал КубГАУ. 2024. № 2(196). С. 193—215.

14. Smith J., Johnson R. Big Data Analytics: Methods and Applications // Proceedings of the International Conference on Big Data. 2021. Pp. 102—115.

15. Garcia A., Martinez E. Data Analysis in Large Volumes: Research Examples // Proceedings of the Big Data Conference. 2019. Pp. 45—58.

REFERENCES

1. Consumer Behavior. D. A. Shevchenko (ed.). Moscow, Direkt-Media, 2023. 112 p. (In Russ.)

2. Ryza S., Laserson U., Owen S., Wills J. Advanced Analytics with Spark. O'Reilly Media, 2015. 260 p.

3. Gugnina E. V., Resnichenko V. E. Factor Analysis, its Methods and Types. *Rossii — tvorcheskuyu molodezh` = Russia – Creative Youth. Materials of the XVI all-Russian scientific and practical student conference*. Volgograd, 2023;3:51—54. (In Russ.)

4. Alekseychik A. B. Economic analysis in conditions of computer data processing: problems and directions of development. *Aktual`nye problemy pravovykh, ekonomicheskikh i gumanitarnykh nauk = Actual problems of legal, economic and humanitarian sciences. Materials of the XII international scientific and practical conference of professors, postgraduates, master students and undergraduates*. Minsk, 2022:60—62. (In Russ.)

5. Beknazarova S. S. Data analysis in distributed information systems. *Big Data i analiz vysokogo urovnya = Big Data advanced analysis. Collection of scientific articles of the VII international scientific and practical conference*. Minsk, 2021:59—66.

6. Naim N. A. Data analysis and processing. *Big Data i analiz vysokogo urovnya = Big Data advanced analysis. Collection of scientific articles of the X international scientific and practical conference*. Minsk, 2024;1:161—163.

7. Galkina E. V. Data flow analysis: application in economics and management. *Vestnik Altaiskoi akademii ekonomiki i prava = Journal of Altai academy of economics and law*. 2021;10-1:16—20. (In Russ.) DOI: 10.17513/vaael.1863.

8. Minin A. S. Comparative analysis of categorical data encoding methods in linear regression problems. *Tendentsii razvitiya nauki i obrazovaniya*. 2023;94-5:98—100. (In Russ.) DOI: 10.18411/trnio-02-2023-259.

9. Efimova T. B., Subocheva Ya. V., Glukhova V. A. Intelligent data analysis in evaluation of the organization's personnel. *Ekonomika i predprinimatel`stvo = Journal of Economy and entrepreneurship*. 2023;8(157):1395—1397. (In Russ.) DOI: 10.34925/EIP.2023.157.8.263.

10. Pomazkov V. V. Analysis of large data sets as a tool for the effectiveness of school development. *Shamov readings. Collection of articles of the XVI international scientific and practical conference*. Moscow, 2024;1:84—89. (In Russ.)

11. Shishkina A. A. Comparative data analysis to improve working conditions. *Izvestiya Tul'skogo gosudarstvennogo universiteta. Tekhnicheskie nauki = News of the Tula state university. Technical sciences*. 2021;9:250—252. (In Russ.)

12. Kumratova A. M., Popova M. I. Methods and visualization tools for analytics in small business. *Sovremennaya ekonomika: problemy i resheniya = Modern economics: problems and solutions*. 2023;2(158):91—98. (In Russ.)

13. Paraskevov A. V., Shapovalov A. V., Sergeev A. E., Uvarova A. G. Approaches to Big Data analysis in the higher education system. *Nauchnyi zhurnal KubGAU = Scientific Journal of KubSAU*. 2024;2(196):193—215. (In Russ.)

14. Smith J., Johnson R. Big Data Analytics: Methods and Applications. *Proceedings of the International Conference on Big Data*. 2021:102—115.

15. Garcia A., Martinez E. Data Analysis in Large Volumes: Research Examples. *Proceedings of the Big Data Conference*. 2019:45—58.

Статья поступила в редакцию 21.12.2024; одобрена после рецензирования 11.01.2025; принята к публикации 13.01.2025.
The article was submitted 21.12.2024; approved after reviewing 11.01.2025; accepted for publication 13.01.2025.