УДК 004.415.25

5.2.2. Математические, статистические и инструментальные методы экономики (физико-математические науки, экономические науки)

ПОДХОДЫ К АНАЛИЗУ БОЛЬШИХ ДАННЫХ В СИСТЕМЕ ВЫСШЕГО ОБРАЗОВАНИЯ

Параскевов Александр Владимирович
РИНЦ SPIN-код: 2792-3483
paraskevov.a@kubsau.ru
Кубанский государственный аграрный университет имени И.Т. Трубилина, Краснодар, Россия

Шаповалов Анатолий Вячеславович
кандидат юридических наук, доцент
РИНЦ SPIN-код: 3424-4583
Кубанский государственный аграрный университет имени И.Т. Трубилина, Краснодар, Россия

Сергеев Александр Эдуардович
кандидат физико-математических наук, доцент

SPIN-код 7837-9566
Кубанский государственный аграрный университет имени И.Т. Трубилина, Краснодар, Россия

Уварова Алина Геннадьевна
бакалавр 3 курса
факультет прикладной информатики
alinauv02@mail.ru
Кубанский государственный аграрный университет имени И.Т. Трубилина, Краснодар, Россия

Методы, основанные на искусственном интеллекте (ИИ) и машинном обучении (ML) достаточно новы, но при этом очень надежны в прогнозировании. Они особенно полезны с данными, имеющими слабую или вовсе отсутствующую структуру. Популярным алгоритмом являются нейронные сети - математические модели, организованные по принципу нервной сети живого организма. К методам обучения таких сетей относятся: обучение с учителем, то есть контролируемое человеком, с предварительной разметкой данных, разбиением выборки на обучающую и тестовую; обучение без учителя возможность выявить закономерности, которые человек не смог бы предположить; обучение с подкреплением, то есть на решения машины дается обратная связь человека, которая

UDC 004.415.25

5.2.2. Mathematical, statistical and instrumental methods of economics (physical and mathematical sciences, economic sciences)

APPROACHES TO BIG DATA ANALYSIS IN THE HIGHER EDUCATION SYSTEM

Paraskevov Alexander Vladimirovich
RSCI SPIN-code 2792-3483
paraskevov.a@kubsau.ru
Kuban State Agrarian University named after I. T. Trubilin, Krasnodar, Russia

Shapovalov Anatoly Vyacheslavovich
Cand.Leg.Sci., associate Professor
RSCI SPIN-code 3424-4583
Kuban State Agrarian University named after I. T. Trubilin, Krasnodar, Russia

Sergeev Alexander Eduardovich
Candidate of Physical and Mathematical Sciences, associate Professor

RSCI SPIN-code 7837-9566
Kuban State Agrarian University named after I. T. Trubilin, Krasnodar, Russia

Uvarova Alina Gennadievna
bachelor of the 3rd year
Department of Applied Informatics
alinauv02@mail.ru
Kuban State Agrarian University named after I. T. Trubilin, Krasnodar, Russia

Methods based on artificial intelligence (AI) and machine learning (ML) are quite new, but at the same time they are very reliable in forecasting. They are especially useful with data that has weak or no structure at all. A popular algorithm is neural networks, mathematical models organized according to the principle of the nervous network of a living organism. The methods of training such networks include: learning with a teacher, that is, supervised by a person, with preliminary data markup, splitting the sample into training and test; learning without a teacher the opportunity to identify patterns that a person would not be able to guess. Reinforcement learning, that is, human feedback is given to the decisions of the machine, which corrects the conclusions. Also included in the category of AI and ML-based methods is a decision tree, a way of

корректирует выводы. Также в категорию методов, основанных на ИИ и ML можно отнести дерево решений - способ представления правил в иерархической последовательной логической структуре, который позволяет соотнести объект или ситуацию на входе с одним или несколькими выходными узлами. Последняя категория основана на применении графических средств анализа сданных. Сюда входят такие методы, как контрольный листок, диаграмма Парето, схема Исикавы, гистограмма, диаграмма разброса, расслоение, контрольная карта, график временного ряда и др. Преимуществом таких методов является простота освоения, использования, возможность применения их в комплексе с другими методами. Они достаточно удобны для восприятия даже не имеющего общего с информационными технологиями человека, поэтому это самый простой, но достаточно надежный способ для принятия решений

representing rules in a hierarchical sequential logical structure that allows you to correlate an object or situation at the input with one or more output nodes. The latter category is based on the use of graphical means of analyzing the submitted data. This includes methods such as checklist, Pareto diagram, Ishikawa diagram, histogram, scatter diagram, layering, control map, time series graph, etc. The advantage of such methods is the ease of mastering, use, and the possibility of using them in combination with other methods. They are quite convenient for the perception of even a person who has nothing to do with information technology, so this is the simplest, but reliable enough way to make decisions

Ключевые слова: АНАЛИЗ ДАННЫХ, БОЛЬШИЕ ДАННЫЕ, ВЫСШЕЕ ОБРАЗОВАНИЕ, МАШИННОЕ ОБУЧЕНИЕ

Keywords: DATA ANALYSIS, BIG DATA, HIGHER EDUCATION, MACHINE LEARNING

http://dx.doi.org/10.21515/1990-4665-196-021

Введение.

Каждый старшеклассник сталкивается с проблемой выбора дальнейшего пути развития, самореализации. Прежде всего необходимо определить направление развития. Некоторые определились с этим еще со средних классов, кто-то же выбирает по остаточному принципу, основываясь на том, что говорят родители, популярности и прибыльности будущей профессии или том, что выбрали их друзья. Этот выбор не основан на личных характеристика и в большом количестве случаев ошибочен. После этого следует вопрос, в каком конкретном месте получать соответствующее образование. Кто-то выберет место для поступления исходя из различных факторов комфорта: местоположение, друзья, то есть выбор основывается на том, куда идут знакомые, другие подойдут к своему выбору ответственно, основываясь на рейтинге ВУЗов, отзывах людей, которые прошли обучение, «силе» направления подготовки. Конечный выбор и тех, и тех будет основываться прежде всего

на сумме необходимых баллов ЕГЭ для поступления, анализируя свои шансы поступить, попасть на бюджетное место или просто желаемый факультет.

Каждый ВУЗ принимает абитуриентов на основании конкурсного отбора в разрезе общего, целевого, льготного и квотного отборов внутри каждого направления обучения. Для того, чтобы привлечь будущих студентов необходимо, чтобы об учебное заведение было «с именем», а также оно имело авторитет. Для этого существует рейтинг ВУЗов. Он строится на территориальном расположении вуза, преподавательском составе, количестве направлений подготовки, их содержание, соотношение бюджетных и коммерческих мест, а также научная, культурная и спортивная активность образовательного обучения, уровня трудоустройства выпускников и многих других факторов. Так что университетам очень важно, чтобы те, кого они принимают для обучения были заинтересованы в своей профессии, развивали науку, были активными и впоследствии работали по профессии. На этом строится рейтинг ВУЗов в глазах потребителя.

В профессии наиболее вероятно останутся те, кто подошли к выбору ответственно, были заинтересованы специальностью, имели хорошую успеваемость во время обучения. Для того, чтобы предугадать это, стоит основываться на среднем балле обучения в школе, наличия отличия в аттестате, высоких баллах абитуриента и других факторах, так как это поможет понять, насколько старателен будет студент.

В процессе реализации деятельности высших учебных заведений порождается множество данных абитуриентов прошлых лет. Они могут быть применены для выявления закономерностей и даже построения прогноза на основе имеющихся данных с помощью технологий Big Data. Ведь для принятия действительно взвешенных решений с целью

повышения рейтинга ВУЗа недостаточно их собирать и хранить, важно правильно проанализировать существующие данные.

Методы анализа данных

Анализ данных – процесс исследования, фильтрации, преобразования и моделирования данных с целью извлечения полезной информации и принятия решений. Происходит это с помощью конкретных математических методов и вычислительных алгоритмов.

Методы анализа данных различны. Использование того или иного зависит от вида задачи, типа данных и их количества. Цель каждого: преобразование данных в параметры для принятия решения с помощью использования новых данных, анализа и детализации информации. Методики анализа массивов данных можно разделить на категории:

– основанные на математике и статистике;

– основанные на искусственном интеллекте и машинном обучении;

– основанные на визуализации и графиках.

Первая категория используется для исследования данных, когда анализируемый параметр изменяется случайным образом. К этой группе относятся методы: регрессионный, дисперсионный и факторный виды анализа, метод сравнения средних, метод сравнения дисперсий, кластерный анализ и др. Данные методы позволяют установить факт зависимости изучаемых явлений от случайных факторов (качественно – дисперсионный, количественно – корреляционный анализ), исследовать связь между этими величинами (регрессионный анализ), выявить роль отдельных факторов в изменении важного параметра (факторный анализ), разбить множество на более мелкие группы, или кластеры, внутри которых находятся схожие на основании какого-то критерия объекты (кластерный анализ).

Big Data – огромные объемы данных, которые постоянно генерируются из множества источников. Для их описания используется правило 5 V:

– volume (объем) указывает на огромные массивы данных, генерируемые каждую секунду;

– velocity (скорость) – скорость создания и передачи данных;

– variety (разнородность) – разнотипность данных, которые могут использоваться;

– veracity (достоверность) – способность достоверно отражать окружающую действительность или предметную область;

– value (ценность) – значимость данных для принятия решений, их полезность.

Данные можно получить абсолютно бесплатно практически отовсюду: из Интернета (социальные сети, веб-страницы, медиа), от различных устройств и их датчиков (аудио-, видеорегистраторы, журналы аудита, IoT-датчики), в процессе реализации различной деятельности (архивы, документация, внутренние сведения организаций, хранилища).

При извлечении данных важно учитывать их тип и формат, доступность (соблюдать правила и ограничения на использование), размеры. Обычно для анализа используют табличное представление, где каждая строка является описанием отдельного объекта, а каждый столбец – его характеристика, признак.

Прежде чем анализировать данные важно их обработать, чтобы уменьшить количество ошибок, ускорить вычисления. Обработка данных включает в себя следующие этапы:

– очистка – удаление лишних, ошибочных или некорректных данных. Устранение дубликатов, пропусков, шумов;

– интеграция, если данные из разных источников, соединение таблиц, приведение их к одному формату;

– трансформация – фильтрация, сортировка, группировка, вычисление новых значений;

– агрегация – обобщение данных путем вычисления итогов по группам. Подсчет в них средних, минимумов, максимумов, сумм.

То, как обрабатывать данные, зависит от их типа: бинарный (два варианта значений), категориальный (вариантов значений больше двух, зачастую строковый тип), целочисленный, непрерывный (содержит числа со знаками после запятой). Например, категориальный – сложен для обработки, поэтому его необходимо трансформировать – каждому уникальному значению присвоить свое числовое обозначение.

Юридические аспекты анализа персональных данных абитуриентов.

Развитие информационных технологий активно влияет на многие сферы жизнедеятельности и является неотъемлемой частью повседневной жизни человека. Однако, данное развитие является не только благом для всего общества, но и помимо положительных сторон следует отметить возникшие проблемы связанные с незаконным сбором, распространением и использованием информации о персональных данных.

Персональные данные (ПД) человека или личностные данные – это нематериальный объект, который представляет высокую ценность. Если по набору информации можно идентифицировать человека, как личность, то это персональные данные. Общим принципом понимания, что относится к персональным данным, а что нет, является следующее: если из приведенных данных можно понять о каком человеке конкретно говорится – это персональные данные.

Население страны каждый день вступают в различные правоотношения, вся информация обрабатывается и хранится на

электронных носителях. Следует отметить, что без должной защиты, информация о персональных данных может попасть к злоумышленникам, что приведет к совершению различных видов правонарушений, за что может наступить административная, уголовная и иные виды юридической ответственности. Не все лица, владеющие информацией о персональных данных граждан, являются злоумышленниками, чаще информация о персональных данных лиц, предоставляющих эти данные, используется в целях, без намерения причинить вред гражданам. Персональная информация может использоваться, в том числе, и в научной деятельности, при проведении различных исследований, в деятельности средств массовой информации, в иных сферах.

Условия сбора, обработки и использования персональных данных человека должны осуществляться с соблюдением определенных принципов и правил.

Для защиты субъектов правовых отношений в информационной сфере, в Российской Федерации, в последние годы активно формируется и развивается законодательство, направленное на защиту персональных данных. Так, 27.07.2006 г. был принят Федеральный закон «О персональных данных», который содержит ряд важных терминов. Указанный нормативный правовой акт является базовым в сфере регулирования отношений, связанных с требованиями по обработке персональных данных граждан и их защиты. В ст. 1 указанного Закона определена сфера действия закона, который направлен на регулирование отношений в области обработки персональных данных органами государственной власти, органами местного самоуправления, физическими и юридическими лицами как с использованием различных средств автоматизации, так и без использования таких средств. Данный Федеральный закон должен способствовать защите частной жизни,

семейной и личной тайны человека и гражданина, а также должен защитить их права при обработке персональных данных операторами.

Согласно ст. 3 Федерального закона от 27.07.2006 г. №152-ФЗ «О персональных данных» персональными данными является любая информация, которая относится к определяемому лицу прямо или косвенно, то есть отнесена к субъекту персональных данных. Обработка средств персональных данных включает в себя любое действие с персональными данными, включая запись, сбор, хранение, систематизацию, накопление, использование, передачу, блокирование, обезличивание, уничтожение, удаление персональных данных о субъекте. Операторами, осуществляющими обработку персональных данных, могут являться органы государственной власти, органы местного самоуправления, а также юридические и физические лица.

В соответствии с Федеральным законом от 27.07.2006 г. №152-ФЗ «О персональных данных» персональными данными будут являться сведения, которые могут позволить идентифицировать человека, то есть сведения, которые относятся к конкретному человеку. Например, в научном исследовании указывается Ф. И. О., дата рождения и другая конфиденциальная информация (место рождения, пол, образование, данные паспорта и т.д.). Любая информация, которая собрана о человеке (общая, специальная) должна быть обработана только с письменного согласия лица.

В рамках настоящей статьи, проведены исследования с использованием некоторых данных об абитуриентах, поступающих в вуз, таких как: пол, возраст, место рождения иные данные. Наличие сведений о лицах, приведенных в статье не создает условий, при которых возможно наступление неблагоприятных последствий, безопасность персональных данных сохранена, неправомерность их использования исключена, так как все данные о лицах, приведенные в исследовательской части статьи, такие

как Ф. И. О., дата рождения, иные личностные данные не применялись, соответственно требования Федерального закона от 27.07.2006 г. №152-ФЗ «О персональных данных» соблюдены, идентифицировать конкретного человека, по имеющимся в статье данным не представляется возможным.

Основная часть.

Для проведения анализа требуется сделать выборку из имеющихся у университета данных за несколько лет и сформировать из них датасет с полями:

– пол;

– возраст;

– район/населенный пункт проживания (без конкретики в виде улицы, дома, квартиры);

– сданные дисциплины и баллы по ним;

– наличие/отсутствие достижений, которые учитываются при поступлении;

– наличие/отсутствие курсов довузовской подготовки;

– средний балл аттестата;

– аттестат с отличием или нет;

– поступление или нет;

– поступление на бюджет или коммерцию;

– на какие факультеты подавал документы;

– на какой факультет поступил.

Для того, чтобы продемонстрировать возможности анализа деятельности ВУЗов и последующей помощи в принятии решений с применением технологий Big Data, был сгенерирован набор опытных данных при помощи языка программирования Python, библиотек random (генерация данных), csv (создание и манипуляция csv файлом), pandas

(считывание html страниц для наполнения справочника по районам/населенным пунктам) в размере 15700 строк со столбцами, отражающими приведенные выше характеристики с годом поступления.

```
pol = ['Мужской', 'Женский']
exams = ['Обществознание', 'История', 'Информатика', 'География', 'Иностранный язык']
facultets = ['Экономический','Учетно-финансовый']
budzh = ['Бюджет', 'Коммерция']
cp = ['Да','Нет']


#формирование листа с районами
url =
'https://ru.wikipedia.org/wiki/%D0%90%D0%B4%D0%BC%D0%B8%D0%BD%D0%B8%D1%81%D1%82%D1%80%D0%B0%D1%82%D0%B8%D0%B2%D0%BD%D0%BE-
%D1%82%D0%B5%D1%80%D1%80%D0%B8%D1%82%D0%BE%D1%80%D0%B8%D0%B0%D0%BB%D1%8C%D0%BD%D0%BE%D0%B5_%D0%B4%D0%B5%D0%BB%D0%B5%D0%BD%D0%B8%D0%B5_%D0%9A%D1%80%D0%B0%D1%81%D0%BD%D0%BE%D0%B4%D0%B0%D1%80%D1%81%D0%BA%D0%BE%D0%B3%D0%BE_%D0%BA%D1%80%D0%B0%D1%8F'
site1 = pandas.read_html(url)[1]
rayons = list(site1['Название'])[1:38]

#формирование листа с городами
url =
'https://ru.wikipedia.org/wiki/%D0%93%D0%BE%D1%80%D0%BE%D0%B4%D1%81%D0%BA%D0%B8%D0%B5_%D0%BD%D0%B0%D1%81%D0%B5%D0%BB%D1%91%D0%BD%D0%BD%D1%8B%D0%B5_%D0%BF%D1%83%D0%BD%D0%BA%D1%82%D1%8B_%D0%9A%D1%80%D0%B0%D1%81%D0%BD%D0%BE%D0%B4%D0%B0%D1%80%D1%81%D0%BA%D0%BE%D0%B3%D0%BE_%D0%BA%D1%80%D0%B0%D1%8F'
site1 = pandas.read_html(url)[0]
towns = list(site1['Название'])

#общий список города/район
ray_town = rayons + towns

#создание справочника городов/районов с кодировкой
with open('Города.csv', mode ='w', newline='') as my_file:
    writer = csv.writer(my_file, dialect='excel')
    for it in ray_town[0:37]:
        writer.writerow([it, 0])
    for it in ray_town[37:63]:
        writer.writerow([it, 1])
```

Приведен фрагмент кода программы генерации опытных данных.

Набор представляет из себя данные поступающих на экономический и учетно-финансовый факультеты за 2015-2023 год.

Пол	Возраст	Район прс	Год посту	Русский я	Математи	Обществс	История	Информа	Географи	Иностран	Достижен	Подкурсы	Поступление	Бюджет/ком	Средний бал	Отличие в аттес	Приоритетный факульт	Поступление на факульте
Женский	18	Усть-Лаби	2022	43	41					44	Да	Нет	Нет		3,1	Нет		
Мужской	17	Приморск	2015	54	99					64	Да	Нет	Да	Коммерция	3,9	Нет	Экономический	Учетно-финансовый
Женский	17	Павловски	2021	38	42			48			Нет	Да	Нет		3	Нет		
Женский	17	Калининс	2023	39	32				42		Нет	Да	Нет		3,3	Нет		
Женский	18	Успенский	2021	68	81					73	Нет	Да	Да	Бюджет	4,6	Нет	Экономический	Учетно-финансовый
Мужской	18	Гулькевич	2017	98	91		61				Нет	Нет	Да	Коммерция	3,8	Нет	Экономический	Экономический
Мужской	18	Кущёвски	2023	61	82		98				Да	Да	Да	Коммерция	3,3	Нет	Экономический	Экономический
Мужской	17	Динской	2019	38	35					48	Да	Нет	Нет		3,9	Нет		
Мужской	17	Славянск	2019	50	50				45		Нет	Нет	Нет		4,7	Нет		
Женский	17	Усть-Лаби	2017	59	66				83		Да	Да	Да	Коммерция	3,9	Нет	Экономический	Экономический
Мужской	17	Апшеронс	2023	38	50				47		Да	Да	Нет		4,1	Нет		
Женский	17	Старомин	2017	50	39			44			Нет	Да	Нет		3,8	Нет		
Мужской	17	Брюховец	2015	61	71	77					Да	Нет	Да	Коммерция	3,7	Нет	Экономический	Экономический
Мужской	18	Мостовск	2019	42	32	50					Нет	Нет	Нет		4,4	Нет		
Мужской	18	Красноар	2022	46	30			42			Да	Нет	Нет		4,3	Нет		
Мужской	18	Славянск	2022	66	77				56		Да	Нет	Да	Бюджет	3,6	Нет	Учетно-финансовый	Учетно-финансовый
Мужской	18	Новоросс	2018	79	77				76		Нет	Да	Да	Бюджет	4,1	Нет	Учетно-финансовый	Учетно-финансовый
Женский	18	Темрюкск	2016	40	29	44					Да	Да	Нет		4,2	Нет		
Женский	18	Ленингра	2022	71	94	65					Нет	Да	Да	Бюджет	3,4	Нет	Учетно-финансовый	Экономический
Мужской	18	Туапсе	2016	36	47	44					Нет	Нет	Нет		4,4	Нет		
Женский	17	Усть-Лаби	2015	78	89		86				Да	Нет	Да	Коммерция	3,1	Нет	Экономический	Экономический
Мужской	18	Лабинск	2019	41	49		42				Нет	Да	Нет		3,8	Нет		
Мужской	17	Апшеронс	2017	36	34		44				Да	Да	Нет		3,8	Нет		
Мужской	18	Красноар	2023	88	100				55		Да	Нет	Да	Бюджет	3,3	Нет	Экономический	Экономический
Мужской	17	Северский	2019	48	30				42		Да	Нет	Нет		3,2	Нет		
Мужской	17	Тбилисск	2020	48	31				50		Нет	Нет	Нет		4,3	Нет		
Женский	17	Новопокр	2019	37	35		45				Нет	Да	Нет		3	Нет		

Рисунок 1 – Набор опытных данных

Славянски	0
Старомин	0
Тбилисски	0
Темрюкск	0
Тимашёвс	0
Тихорецки	0
Туапсинск	0
Успенский	0
Усть-Лаби	0
Щербино	0
Абинск	1
Анапа	1
Апшеронс	1
Армавир	1
Белореченск	1
Геленджи	1
Горячий К	1
Гулькевич	1
Ейск	1
Кореновс	1
Краснода	1
Кропоткин	1
Крымск	1
Курганинс	1
Лабинск	1
Новокуба	1
Новоросс	1

Рисунок 2 – Справочник городов/районов с кодировкой

Пример анализа опытных данных.

Опытный набор данных был проанализирован с помощью свободно распространяемого и открытого программного обеспечения Knime Analytics Platform. Скрипт процесса обработки опытного набора данных.

Сначала считывается подготовленный MS Excel файл с данными. Далее в узле «Data prep» происходит сценарий подготовки данных к обработке. В нем пустые значения в полях сдаваемых дисциплин заменяются на 0, таким же образом устраняются пустые значения в поле бюджет/коммерция. В полях «Приоритетный факультет» и «Поступление на факультет» такие значения заменяются на «а» и «б» соответственно. На основе значений в полях сдаваемых дисциплин высчитывается итоговая сумма баллов и упорядочивается столбец среди других.

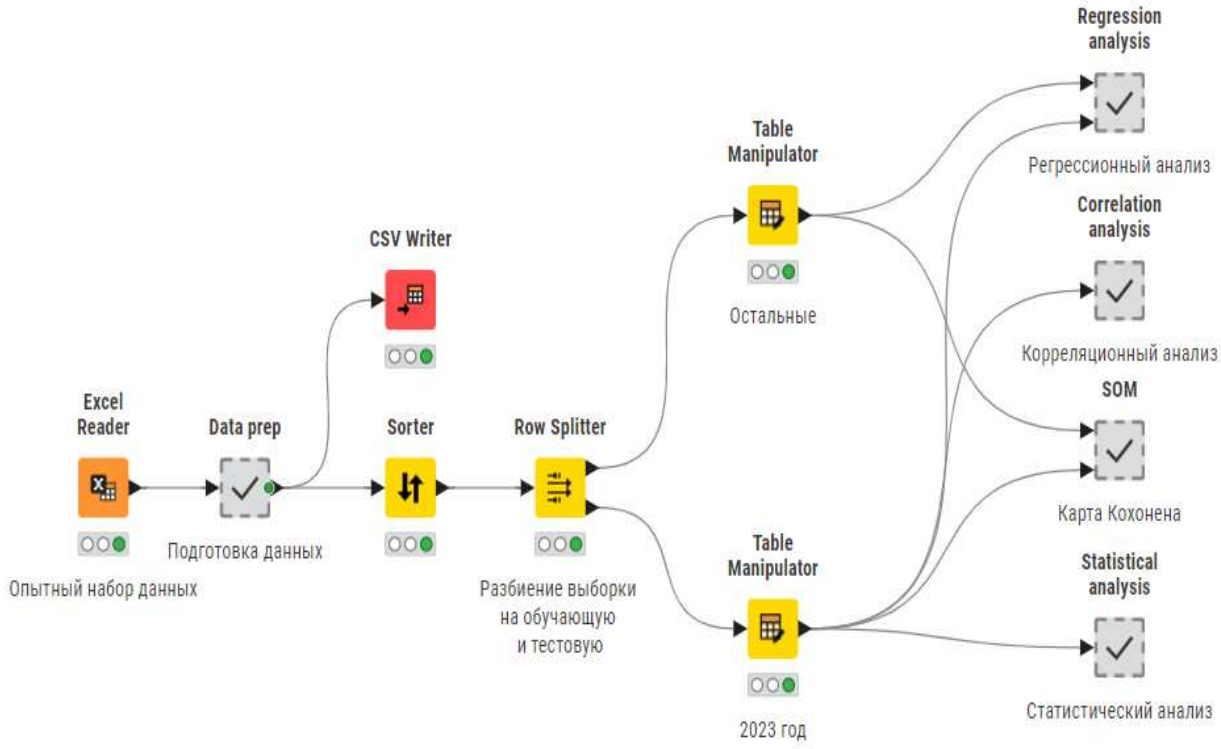


Рисунок 3 – Сценарий обработки набора данных

Затем происходит кодирование данных в полях «Пол», «Достижения», «Подкурсы», «Поступление», «Бюджет/коммерция», на основе «Приоритетный факультет» и «Поступление на факультет» вычисляется новое закодированное поле «Поступление на желаемый факультет», которое отражает зачисление абитуриента на приоритетное направление. Следующий шаг – считывание справочника

«Городов/районов» для кодировки столбца «Город/район» по значениям город/район (1/0), а также создание нового – «Район/город проживания» с кодировкой 63 уникальных значений.

После подготовки набора данных сортируем строки по году поступления и разбиваем данные на до 2023 года и 2023 год и приступаем к анализу. В узле «Statistical analysis» находится статистика по всем поступавшим, поступившим и не поступившим за 2023 год.

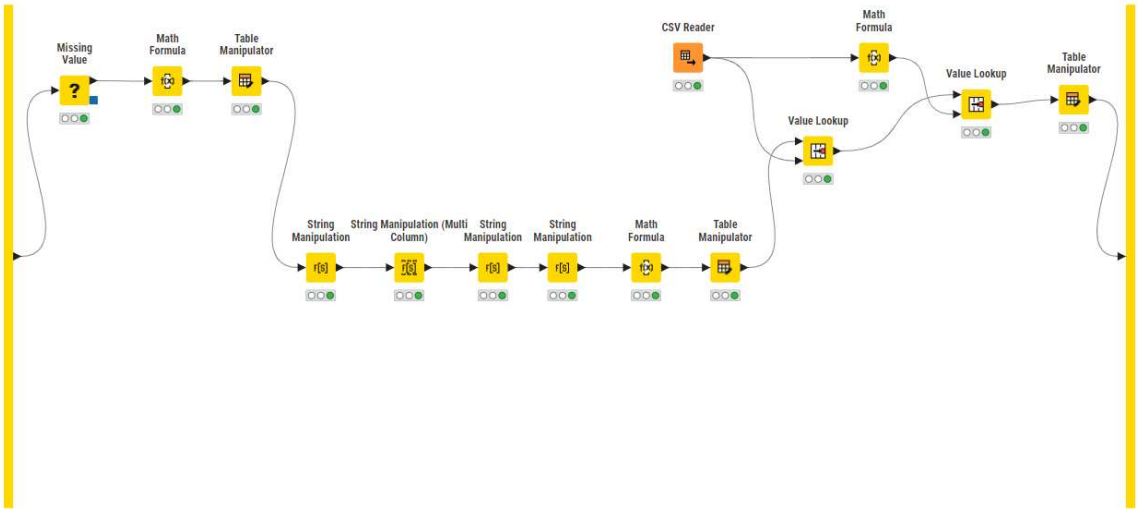


Рисунок 4 – Сценарий подготовки к обработке

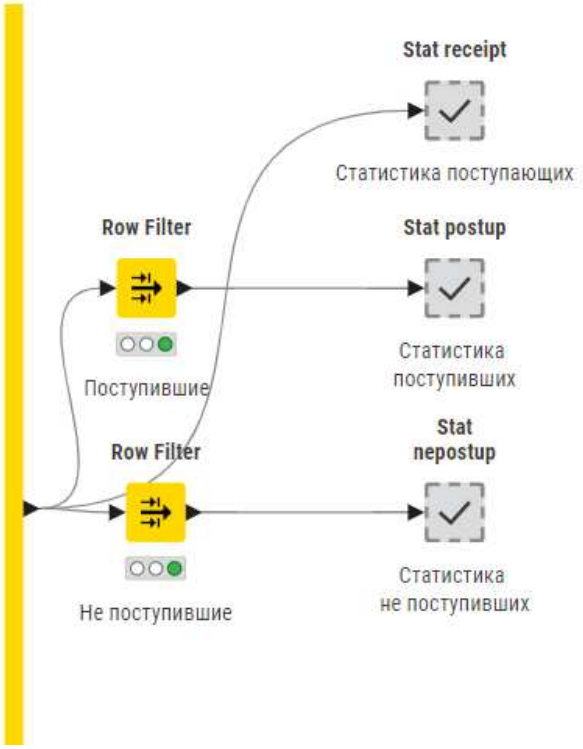


Рисунок 5 – Статистический анализ

В узле «Stat receipt» находятся данные о соотношении поступивших к не поступившим и популярность экзаменов по выбору среди них.

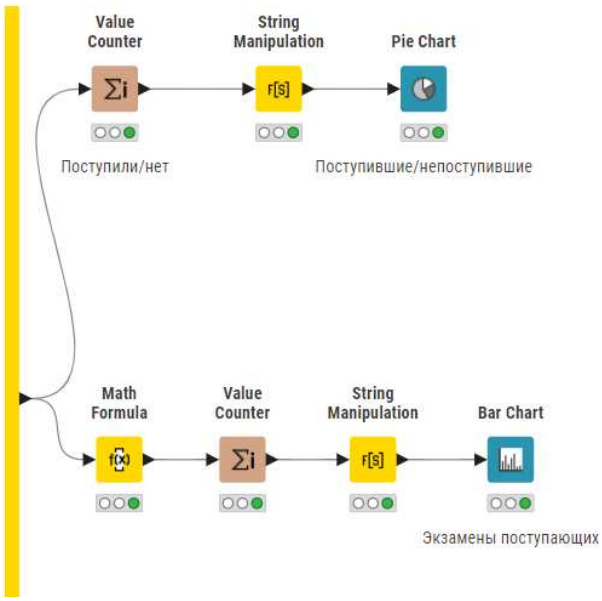


Рисунок 6 – Статистика поступавших

Данные о популярности экзаменов по выбору среди поступивших, соотношение студентов с городов и районов, а также данные о том, откуда они конкретно, проанализированы в узле «Stat postup». Такой же анализ проведен и в узле «Stat nepostup» для не поступивших.

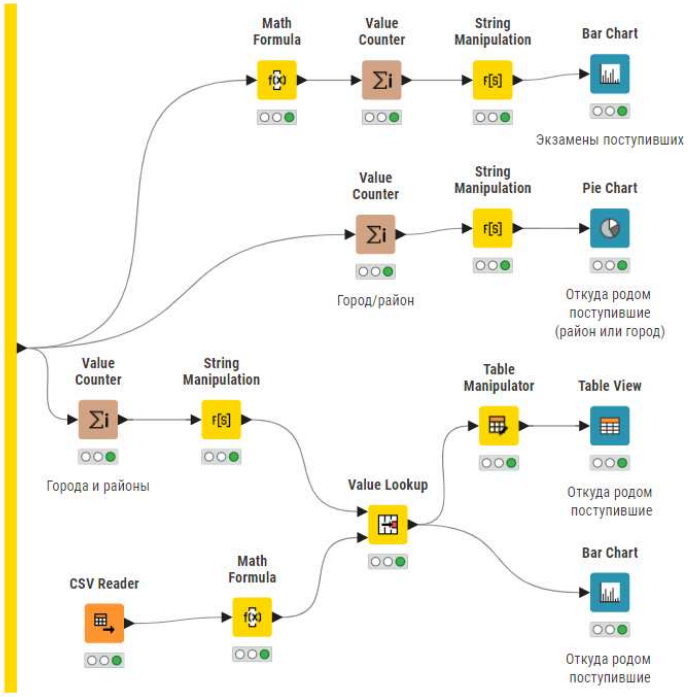


Рисунок 7 – Статистика поступивших

Следующим этапом (узел «SOM») стало построение самоорганизующихся карт Кохонена, который поможет лучше понимать полученные данные, разбить их на отдельные классы, попадание в которые строится на основе признаков «Сумма баллов», «Достижения», «Подкурсы», «Средний балл аттестата», «Отличие в аттестате», «Поступление». Обучив модель на данных до 2023 года, мы разбиваем на группы абитуриентов 2023 года со сходными признаками.

Далее к группам, к которым они принадлежат, соотносим цвет и можем просматривать полученные результаты с помощью графика, изменяя оси на нужные нам соотношения.

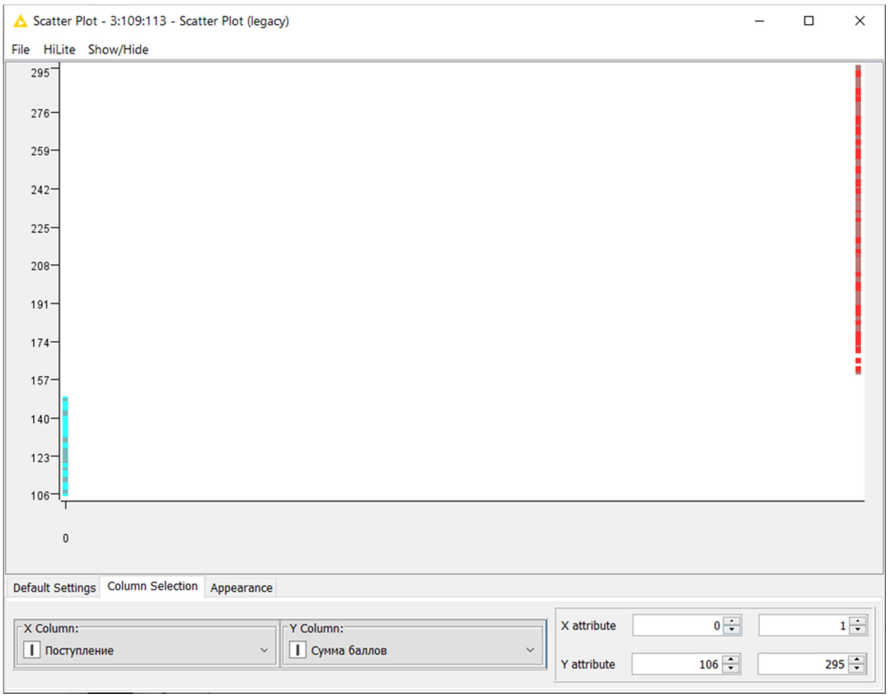


Рисунок 8 – Результаты SOM

В узле «Correlation analysis» находятся 3 узла для анализа зависимостей фактор-переменная, пара факторов-переменная, тройка факторов-переменная. В узле «Corr def» происходит корреляционный анализ поступления, формы обучения, поступления на желаемый факультет от суммы баллов, среднего балла, отличия в аттестате, наличия курсов довузовской подготовки, достижений. В результате получаем таблицу со значением коэффициента корреляции на каждый пункт. Для получения пар факторов в узлах «Corr pairs» и «Corr trio» применяется логическая операция «И», а также кодируется средний балл следующим образом: больше 4,0 значение «1», иначе – «0». Получаем коэффициенты для пар «достижения-подкурсы», «достижения-отличие», «подкурсы-отличие», «достижения-средний балл», «подкурсы-средний балл» и для троек «подкурсы-достижения-отличие», «подкурсы-достижения-средний балл».

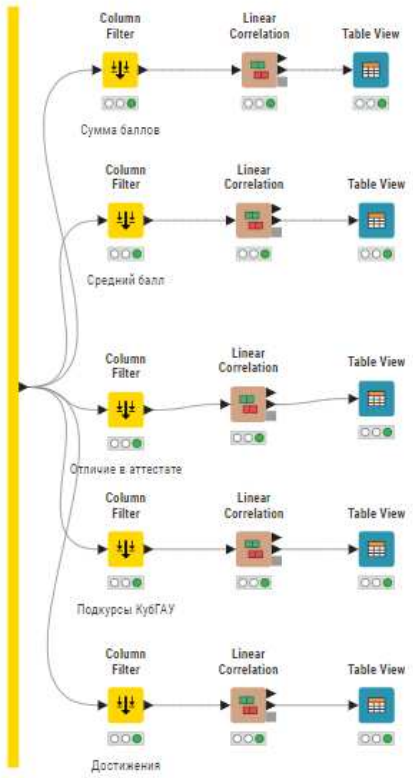


Рисунок 9 – Корреляционный анализ фактор-переменная

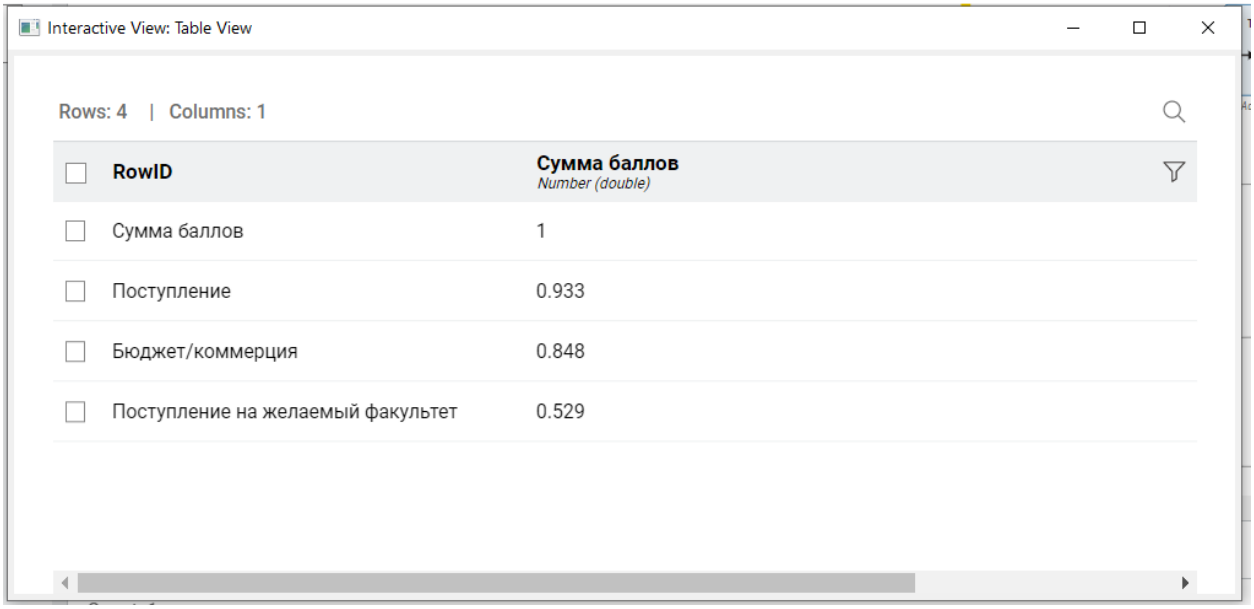


Рисунок 11 – Пример результата корреляционного анализа

Далее с данными до 2023 года был выполнен регрессионный анализ в узел «Regression analysis», на основе которого была обучена модель, и впоследствии был сделан прогноз для 2023 года, предсказанные результаты, округленные до 1, если шанс больше или равен 0,66, иначе – до 0, были сравнены с реальными данными для выяснения правильности работы прогноза (оценка качества классификации) и оценена полученная регрессия.

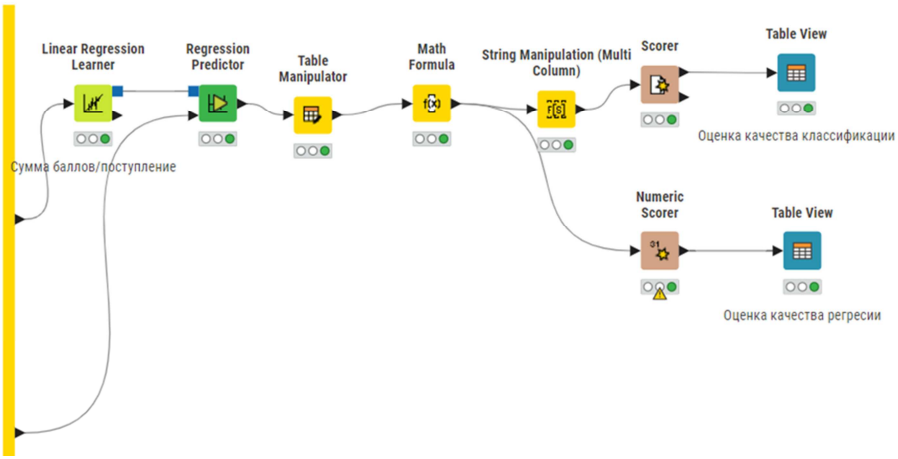


Рисунок 12 – Регрессионный анализ

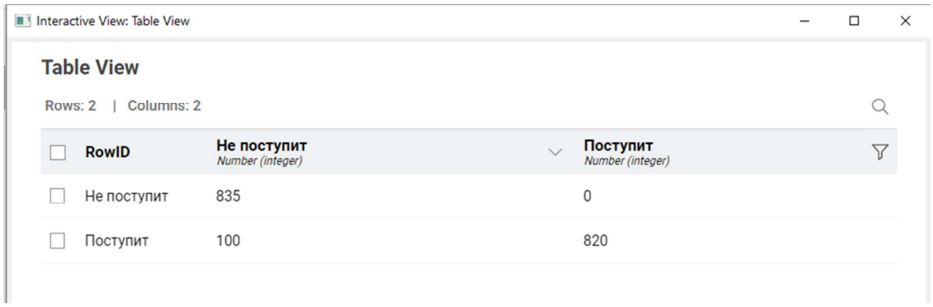


Рисунок 13 – Оценка качества классификации

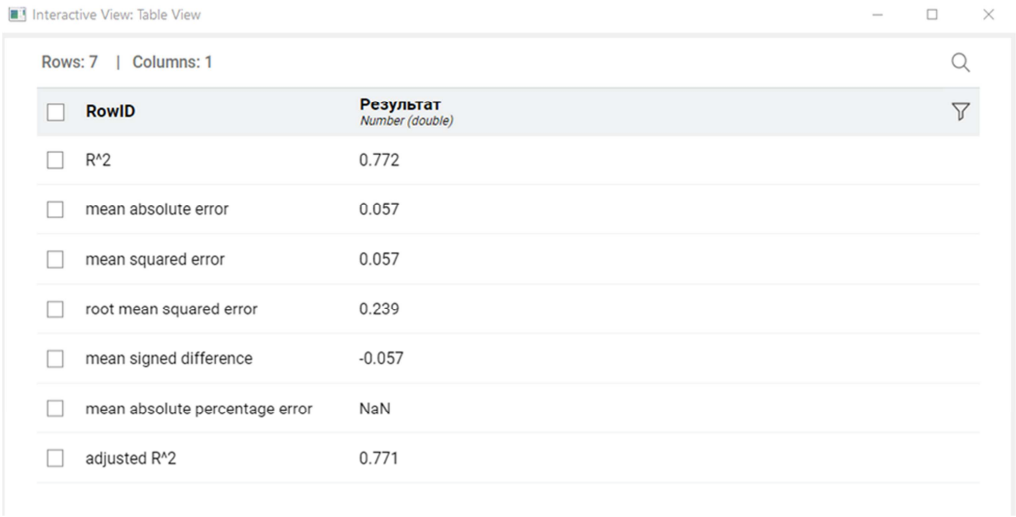


Рисунок 14 – Оценка качества регрессии

Практическое применение и выводы.

Чтобы просмотреть полученные результаты анализа можно заходить в конечные узлы, просматривая таблицы, диаграммы, созданные для упрощения восприятия, или заходить в предшествующий узел для более детального разбора. Обработав опытный набор данных, описанными выше способами, из статистического анализа можно сделать следующие выводы: на одно место в учетно-финансовом и экономическом факультете претендует 2 человека (данные – тестовые, не являются актуальными) – это очень хорошо для абитуриента, так как выше возможность поступить. Самый непопулярный экзамен среди абитуриентов – информатика, самый популярный – обществознание, при этом с ним поступают хуже, чем с историей, географией и иностранным языком, у не поступивших же он встречается чаще всего. Лучше всего поступают с иностранным языком и географией. Возможно, стоит пересмотреть сложность сдачи дисциплины по выбору «Обществознание» на хорошие баллы. Также абитуриент может сделать выбор в пользу географии и иностранного языка, ведь так его шанс поступить будет больше. Больший приток абитуриентов с районов, чем с городов. Лучше всего поступают из г. Гулькевичи. ВУЗу стоит задуматься

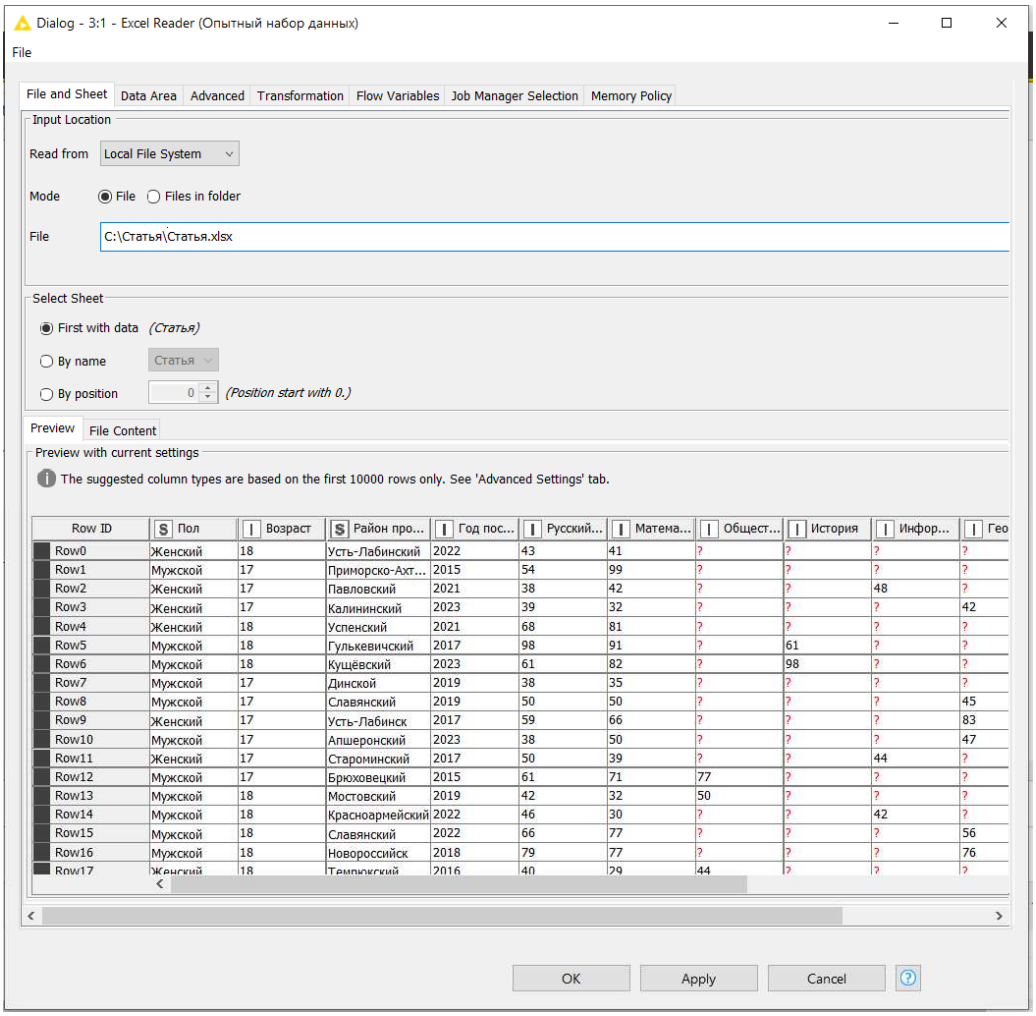
о том, чтобы сотрудничать со школами в данном городе, проводить мероприятия с целью привлечения школьников, ведь он дает больше всего потенциально одаренной молодежи.

Из корреляционного анализа выведена прямая зависимость между суммой баллов и поступлением, бюджетом/коммерцией, поступлением на желаемый факультет, отрицательная зависимость между средним баллом аттестата и поступлением, поступлением на желаемый факультет, исходя из того, что p-значение ≤ 0,07 является статистически значимым. Такая зависимость может говорить о том, что ВУЗ принимает студентов, которые могли просто хорошо подготовиться к сдаваемым дисциплинам, но при этом обучение дается им тяжелее, что впоследствии может привести к отчислению. Зависимости от достижений, подкурсов и отличия в аттестате, от пар и троек факторов не выявлены. Отсутствие зависимости от подготовительных курсов может говорить о их неэффективности, ВУЗу следует улучшить их, а абитуриенту это может помочь в принятии решения о их использовании.

С помощью регрессионного анализа был спрогнозирован результат 2023 года. Доля правильных ответов составила 94,3%, подобранная регрессия имеет коэффициент детерминации 0,772, а средняя абсолютная ошибка – 0,057, что говорит о том, что построенная модель хорошо объясняет набор данных. Это является очень хорошим результатом и позволяет использовать построенную для прогноза модель для предсказания в ближайшем будущем как со стороны ВУЗа (прогнозирование набора), так и для абитуриента.

Построенный сценарий обработки данных может быть использован в собственных целях следующим образом: в самом начальном узле «Excel Reader» необходимо зайти в настройки и в графе file указать путь до соответствующего файла excel. Для корректной работы скрипта следует, чтобы изначальный файл имел те же столбцы, их количество, тип,

название и возможные значения: «Пол» (string), «Возраст» (integer), «Район проживания/Город» (string), «Год поступления» (integer), «Русский язык» (integer), «Математика» (integer), «Обществознание» (integer), «История» (integer), «Информатика» (integer), «География» (integer), «Иностранный язык» (integer), «Достижения» (string), «Подкурсы КубГАУ» (string), «Поступление» (string), «Бюджет/коммерция» (string), «Средний балл аттестата» (double), «Отличие в аттестате» (string), «Приоритетный факультет» (string), «Поступление на факультет» (string). В ином случае придется настраивать узлы под обрабатываемые данные, логика обработки может поменяться.



Row ID	Пол	Возраст	Район про...	Год пос...	Русский...	Матема...	Общест...	История	Инфор...	Гео
Row0	Женский	18	Усть-Лабинский	2022	43	41	?	?	?	?
Row1	Мужской	17	Приморско-Ахт...	2015	54	99	?	?	?	?
Row2	Женский	17	Павловский	2021	38	42	?	?	48	?
Row3	Женский	17	Калининский	2023	39	32	?	?	?	42
Row4	Женский	18	Успенский	2021	68	81	?	?	?	?
Row5	Мужской	18	Гулькевичский	2017	98	91	?	61	?	?
Row6	Мужской	18	Кущёвский	2023	61	82	?	98	?	?
Row7	Мужской	17	Динской	2019	38	35	?	?	?	?
Row8	Мужской	17	Славянский	2019	50	50	?	?	?	45
Row9	Женский	17	Усть-Лабинск	2017	59	66	?	?	?	83
Row10	Мужской	17	Апшеронский	2023	38	50	?	?	?	47
Row11	Женский	17	Староминский	2017	50	39	?	?	44	?
Row12	Мужской	17	Брюховецкий	2015	61	71	77	?	?	?
Row13	Мужской	18	Мостовский	2019	42	32	50	?	?	?
Row14	Мужской	18	Красноармейский	2022	46	30	?	?	42	?
Row15	Мужской	18	Славянский	2022	66	77	?	?	?	56
Row16	Мужской	18	Новороссийск	2018	79	77	?	?	?	76
Row17	Женский	18	Темрюкский	2016	40	29	44	?	?	?

Рисунок 15 – Свойства узла «Excel reader»

Применение методов Big Data в разрезе деятельности высших учебных заведений может позволить проанализировать работу

образовательного учреждения, принять решения по улучшению каких-то процессов: пересмотреть сложность сдаваемых дисциплин, условия принятия решений о зачислении абитуриентов, подготовительные курсы для будущих студентов, начать сотрудничать с определенными районами, городами, проводить мероприятия, повышая интерес и привлекая потенциально одаренную молодежь. Для абитуриентов это позволит вычислять свои шансы на поступление, принимать взвешенное решение о сдачи тех или иных предметов по выбору для увеличения своих шансов, использования подготовительных курсов ВУЗа.

Такие результаты могут быть достигнуты в том числе, если результаты анализа данных из учебного заведения будут опубликованы, допустим, на официальном сайте. Результаты деятельности станут открытыми, что позволит увеличить доверие абитуриентов к ВУЗу, улучшить его рейтинг.

Список литературы

1. Пенкина Ю. Н. Адаптивная система поддержки принятия оперативных решений в управлении IT-проектами / Ю. Н. Пенкина, А. В. Параскевов // Научное обеспечение агропромышленного комплекса, сборник статей по материалам 71-й научно-практической конференции студентов по итогам НИР за 2015 год. Краснодар – Кубанский государственный аграрный университет имени И. Т. Трубилина, 2016.

2. Параскевов А. В. Разработка системы принятия решений при работе с динамическими показателями / А. В. Параскевов, А. Ф. Алексеев // Математические методы и информационно-технические средства. Материалы X Всероссийской научно-практической конференции, Краснодар, 2014.

3. Параскевов А. В. Оценка компетенций бакалавра: возможное решение проблем / А. В. Параскевов, Г. О. Монин // статья в открытом архиве www.researchgate.net, 10.13140/RG.2.2.26177.58721, 2020.

4. Параскевов А. В. Необходимость внедрения информационных технологий / А. В. Параскевов, Д. А. Махлушев, А. А. Ахлёстова // Политематический сетевой электронный научный журнал Кубанского государственного аграрного университета (Научный журнал КубГАУ) [Электронный ресурс]. – Краснодар: КубГАУ, 2022. – №09(183). С. 212 – 222. – IDA [article ID]: 1832209021. – Режим доступа: http://ej.kubagro.ru/2022/09/pdf/21.pdf, 0,688 у.п.л.

5. Параскевов А. В. Анализ опыта внедрения информационных систем в образовательную деятельность / А. В. Параскевов // Точки научного роста: на старте десятилетия науки и технологии. Материалы ежегодной научно-практической конференции преподавателей по итогам НИР за 2022г. Краснодар, 2023.

6. https://www.vsavm.by/knigi/kniga3/780.html

References

1. Penkina Ju. N. Adaptivnaja sistema podderzhki prinjatija operativnyh reshenij v upravlenii IT-proektami / Ju. N. Penkina, A. V. Paraskevov // Nauchnoe obespechenie agropromyshlennogo kompleksa, sbornik statej po materialam 71-j nauchno-prakticheskoj konferencii studentov po itogam NIR za 2015 god. Krasnodar – Kubanskij gosudarstvennyj agrarnyj universitet imeni I. T. Trubilina, 2016.

2. Paraskevov A. V. Razrabotka sistemy prinjatija reshenij pri rabote s dinamicheskimi pokazateljami / A. V. Paraskevov, A. F. Alekseev // Matematicheskie metody i informacionno-tehnicheskie sredstva. Materialy X Vserossijskoj nauchno-prakticheskoj konferencii, Krasnodar, 2014.

3. Paraskevov A. V. Ocenka kompetencij bakalavra: vozmozhnoe reshenie problem / A. V. Paraskevov, G. O. Monin // stat'ja v otkrytom arhive www.researchgate.net, 10.13140/RG.2.2.26177.58721, 2020.

4. Paraskevov A. V. Neobhodimost' vnedrenija informacionnyh tehnologij / A. V. Paraskevov, D. A. Mahlushev, A. A. Ahljostova // Politematicheskij setevoj jelektronnyj nauchnyj zhurnal Kubanskogo gosudarstvennogo agrarnogo universiteta (Nauchnyj zhurnal KubGAU) [Jelektronnyj resurs]. – Krasnodar: KubGAU, 2022. – №09(183). S. 212 – 222. – IDA [article ID]: 1832209021. – Rezhim dostupa: http://ej.kubagro.ru/2022/09/pdf/21.pdf, 0,688 u.p.l.

5. Paraskevov A. V. Analiz opyta vnedrenija informacionnyh sistem v obrazovatel'nuju dejatel'nost' / A. V. Paraskevov // Tochki nauchnogo rosta: na starte desjatiletija nauki i tehnologii. Materialy ezhegodnoj nauchno-prakticheskoj konferencii prepodavatelej po itogam NIR za 2022g. Krasnodar, 2023.

6. https://www.vsavm.by/knigi/kniga3/780.html