УДК 004

Кластерный анализ данных о рейтинге лучших вузов России с помощью программного пакета визуального программирования Orange

Голубева Евгения Павловна
Приамурский государственный университет имени Шолом-Алейхема
Студент

Аннотация
Цель данной статьи – выполнить кластерный анализ данных о рейтинге лучших вузов России за 2021 г. Для кластерного анализа был использован программный пакет визуального программирования на основе компонентов для визуализации данных Orange и данные о рейтинге лучших вузов России. С помощью средств визуализации Orange выполнили кластерный анализ данных о рейтинге лучших вузов России и получили итоговую схему.
Ключевые слова: Orange, виджет, визуализация данных, кластерный анализ.

Cluster analysis of data on the ranking of the best universities in Russia using the Orange visual programming software package

Golubeva Evgeniya Pavlovna
Sholom-Aleichem Priamursky State University
Student

Abstract
The purpose of this article is to perform a cluster analysis of data on the ranking of the best universities in Russia for 2021. For cluster analysis, a visual programming software package based on Orange data visualization components and data on the ranking of the best universities in Russia were used. With the help of Orange visualization tools, we performed a cluster analysis of data on the ranking of the best universities in Russia and obtained the final scheme.
Keywords: Orange, widget, data visualization, cluster analysis.

1 Введение

1.1 Актуальность

Современный мир стал невероятно зависим от данных, которые создаются и хранятся в большом количестве в различных сферах деятельности. При этом, для их обработки и анализа требуются соответствующие методы и инструменты. Одним из наиболее популярных методов анализа данных является кластерный анализ.

Преимущества кластерного анализа включают возможность автоматического выделения групп объектов на основе их характеристик, что позволяет сократить время и усилия, затрачиваемые на анализ данных

вручную. Кроме того, кластерный анализ может помочь выявить скрытые связи между объектами, которые могут быть невидимы на первый взгляд.

Программа Orange благодаря своим функциям позволяет провести кластерный анализ данных.

1.2 Обзор исследований

А. В. Леонов в статье рассматривал основные алгоритмы кластеризации категориальных данных применительно к различным типам пользовательских интерфейсов, определяются их достоинства и недостатки. [1]. В данной статье проанализировали основные аспекты кластерного анализа больших объемов данных при помощи различных методов, их сравнения и выделения наиболее эффективного А. В. Клименко, И. С Слащев [2]. Д. В. Гринченков, Ф. Х. Нгуен, Т. Т. Нгуен, Д. А. Горбушин выполнили краткий обзор и сравнительный анализ возможностей алгоритмов, используемых для интеллектуального анализа данных [3]. В статье рассматривали исследование программного обеспечения Data Mining Ю. С. Кривенко, А.Т. Минасян и А.О. Разиньков [4]. С.В. Пальмов и А.А. Жуйкова в статье описали функционал аналитического пакета Orange, предназначенный для поиска часто встречающихся наборов элементов и ассоциативных правил [5]. В статье рассмотрел использование методов кластеризации в программе Orange на основе реальной базы данных. Н. Юсупов[6]. С. С. Мастевой и А.Н. Петрова статью посвятили краткому обзору методов Data Mining в период информационной эры [7].

1.3 Цель исследования

Цель исследования - выполнить кластерный анализ данных с помощью программы Orange.

2 Материалы и методы

Для кластерного анализа используется программа Orange. Работа будет происходить на готовых данных рейтинг вузов.xlsx, скачать которые можно по ссылке:
https://docs.google.com/spreadsheets/d/1GsA6zDjMFrzWnXgcXubjvJ0jdu_CT_MC/edit?usp=sharing&ouid=104272149632818699735&rtpof=true&sd=true

3 Результаты и обсуждения

Перед началом работы требуется установить Orange с официального сайта и установить.

1) Создадим новый файл (рис.1).

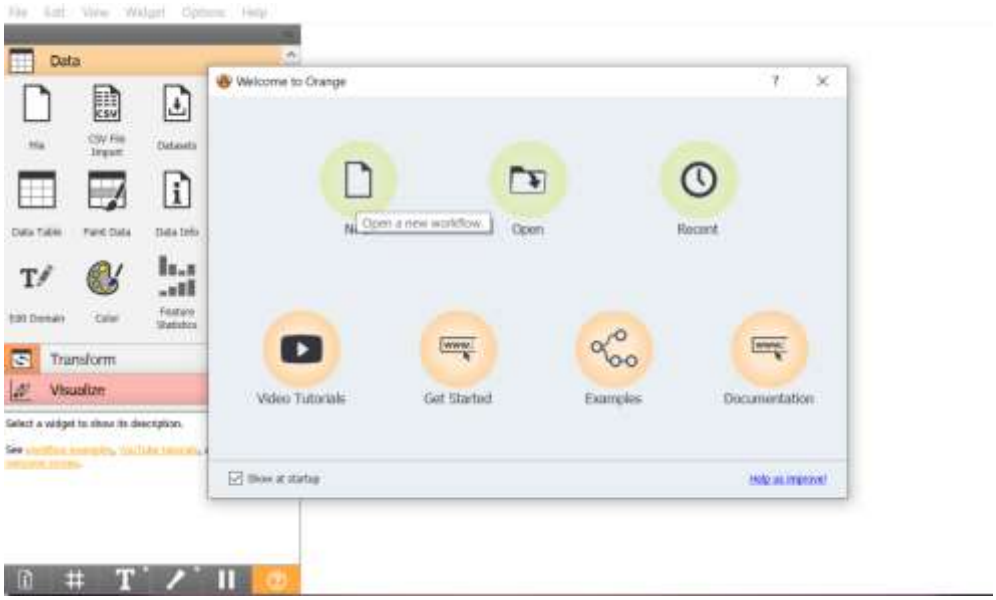

Рис.1. Создание нового файла

2) Добавляем виджет File на холст (Рис.2).

Виджет File считывает файл входных данных (таблицу данных с экземплярами данных) и отправляет набор данных в выходной канал. История последних открытых файлов хранится в виджете. Виджет также включает каталог с образцами наборов данных, которые поставляются с предустановленным с Orange.

Виджет считывает данные из Excel (.xlsx), простых файлов с разделителями табуляции (.txt), файлов, разделенных запятыми (.csv) или URL-адресов.

Рис.2. Добавление виджета File на холст

Есть 3 способа добавления виджета на холст:

1. Дважды щелкните на виджет.
2. Перетащите виджет на холст.
3. Щелкните правой кнопкой мыши на холсте для меню виджета.

3) Чтобы добавить файл необходимо открыть виджет file на холсте (Рис.3).

Рис.3. Открытие виджета File

4) Открылось диалоговое окно File (Рис.4).

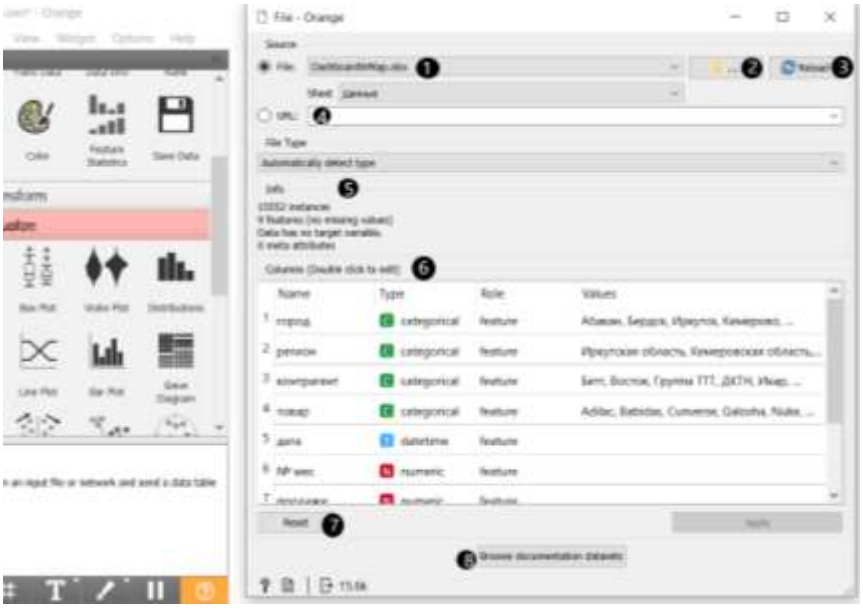

Рис.4. Диалоговое окно File

Описание диалогового окна File (Рис.4).

1. Просмотр ранее открытых файлов данных или загрузка любой из образцов.
2. Найти файл данных.
3. Перезагрузка выбранного в данный момент файл данных.
4. Вставка данных из URL-адресов, включая данные из Google Таблиц.
5. Информация о загруженном наборе данных: размер набора данных, количество и типы объектов данных.
6. Дополнительные сведения о функциях в наборе данных. Объекты можно редактировать, дважды щелкнув по ним. Пользователь может изменить имена атрибутов, выбрать тип переменной для каждого атрибута

(Continuous, Nominal, String, Datetime) и выбрать способ дальнейшего определения атрибутов (как Features, Targets или Meta). Пользователь также может проигнорировать атрибут.

7. Сброс.
8. Просмотр наборов данных документации.

5) Добавляем файл рейтинг вузов.xlsx (Рис.5).

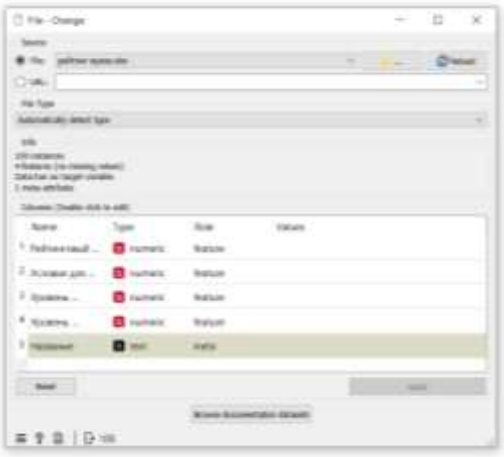

Рис.5. Добавление файла

6) Добавляем виджет Data Table на холст (Рис.6).

Data Table - получает один или несколько наборов данных в виде входных данных и представляет их в виде электронной таблицы. Экземпляры данных могут быть отсортированы по значениям атрибутов. Виджет также поддерживает ручной выбор экземпляров данных.

Рис.6. Добавление виджета Data Table

7) Чтобы посмотреть данные таблицы необходимо соединить два виджета на холсте File и Data Table (Рис.7).

Рис.7. Соединение виджетов

8) Открываем виджет Data Table что бы просмотреть данные загруженной таблицы (Рис.8).

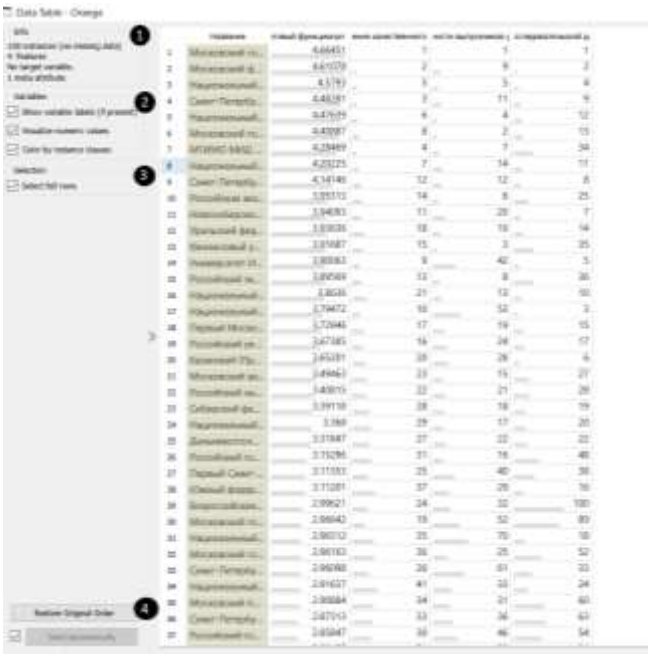

Рис.8. Диалоговое окно Data Table

Описание диалогового окна Data Table (Рис.8).

1. Сведения о текущем размере набора данных, количестве и типах атрибутов
2. Значения непрерывных атрибутов могут быть визуализированы с помощью баров; знамя может быть отнесен к различным классам.
3. Экземпляры данных (строки) могут быть выбраны и отправлены на выход виджета канал.
4. Используется кнопка Restore Original Order для изменения порядка экземпляров данных после сортировки на основе атрибутов.

9) Для того чтобы провести кластерный анализ необходимо использовать виджет K-Means. Кластеризация К-средних (K-Means) – простой метод

разделения множества данных на К различных непересекающихся кластеров. Для выполнения кластеризации сначала нужно определить желаемое число кластеров К, затем алгоритм К средних будет относить каждое наблюдение в точности к одному из К кластеров.

10) Из раздела unsupervised выбираем виджет k-Means, добавляем на холст и соединяем с виджетом Data Table (рис.9).

Рис.9. Добавление виджета k-Means

11) Открываем виджет k-Means, выбираем количество кластеров 4 (рис.10).

Рис.10. Настройка виджета k-Means

12) Для того чтобы просмотреть данные виджета k-Means добавляем на холст виджет Data Table и соединяем с k-Means (рис.11).

Рис.11. Соединение виджетов

13) Открываем виджет Data Table. В таблице появилось 2 новых столбца Cluster и Silhouette. Столбец Cluster содержит информацию к какому кластеру относится вуз (рис.12).

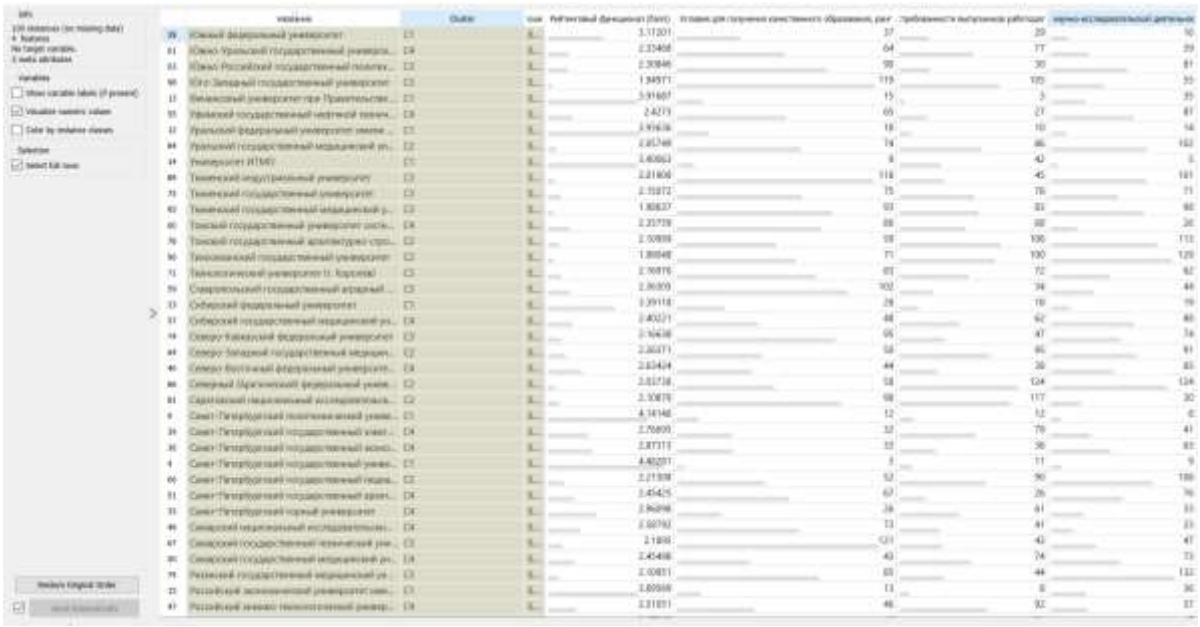

Рис.12. Данные виджета k-Means

14) Для того чтобы визуально посмотреть распределение вузов по схожим признакам (баллов), на холст добавляем виджет scarlett plot и соединяем с виджетом k-Means.

15) Открываем Scatter Plot (Рис.13).

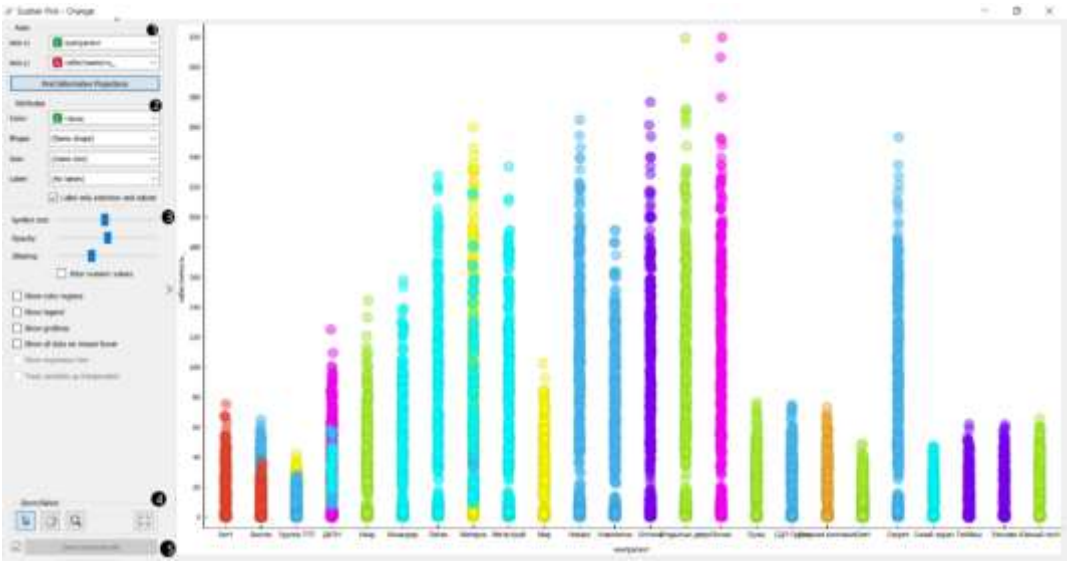

Рис.13. Диалоговое окно Scatter Plot

1. Выбор атрибутов x и y. Оптимизируйте свою проекцию с помощью функции поиска информативных проекций. Эта функция оценивает пары атрибутов по средней точности классификации и возвращает пару с наибольшим баллом с одновременным обновлением визуализации.

2. Атрибуты: Задайте цвет отображаемых точек (вы получите цвета для категориальных значений и сине-зелено-желтые точки для числовых). Задайте метку, форму и размер, чтобы различать точки. Метка только выбранных точек позволяет выбирать отдельные экземпляры данных и помечать только их.

3. Задайте размер и непрозрачность символа для всех точек данных. Установите дрожание, чтобы предотвратить перекрытие точек. Дрожание будет случайным образом рассеивать точки только вокруг категориальных значений. Если установлен флажок Числовые значения Джиттера, точки также рассеиваются вокруг их фактических числовых значений.

• Отображение цветовых областей цветов графика по классам.

• Показать легенду отображает легенду справа. Щелкните и перетащите легенду, чтобы переместить ее.

• Показать линии сетки отображает сетку за графиком.

• Показывать все данные при наведении указателя мыши позволяет создавать информационные пузырьки, если курсор помещен на точку.

• Показать линию регрессии рисует линию регрессии для пары числовых атрибутов. Если для раскраски графика выбрана категориальная переменная, будут отображаться отдельные линии регрессии для каждого значения класса.

• Отношение к переменным как к независимым соответствиям линии регрессии к группе точек (минимизация расстояния от точек), а не как к функции x (минимизация вертикальных расстояний).

1. Выбор, масштабирование, панорамирование и масштабирование по размеру - это варианты изучения графика. Ручной выбор экземпляров данных работает как инструмент углового/квадратного выделения. Дважды

щелкните, чтобы переместить проекцию. Прокрутка вверх или вниз для масштабирования.

2. Если установлен флажок отправить автоматически, изменения сообщаются автоматически. Также можно нажать кнопку Отправить.

16) Используя Scarlett plot проведем кластерный анализ данных Рейтингового функционала вуза, для этого выберем атрибуты в axis x «Cluster», а в axis y «Рейтинговый функционал вуза». С помощью графика можно увидеть, что группа кластера 1 (C1) имеет наивысший балл по рейтинговому функционалу (Рис.14).

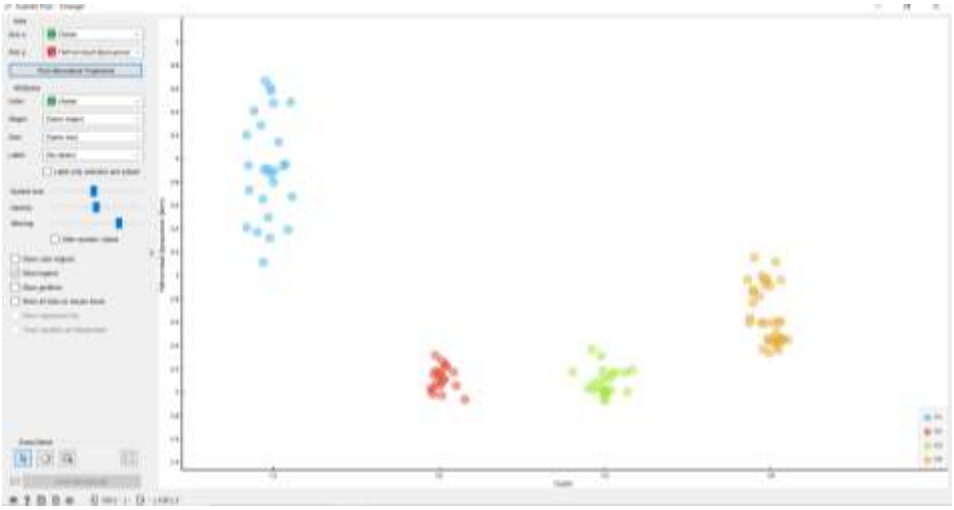

Рис.14. Кластерный анализ данных Рейтингового функционала вуза

17) По такому же принципу проведем кластерный анализ данных условий для получения качественного образования вуза, для этого выберем атрибуты в axis x «Cluster», а в axis y «Условия для получения качественного образования». С помощью графика можно увидеть, что группа кластера 3 (C3) имеет наивысший балл по получению качественного образования вуза (Рис.15).

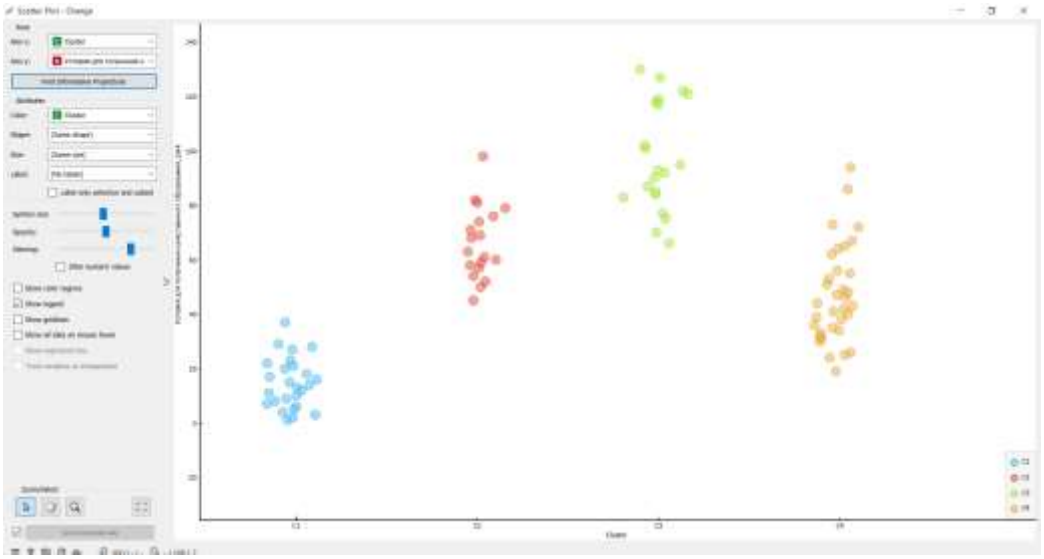

Рис.15. Кластерный анализ данных условий для получения качественного образования вуза

18) Также проведем кластерный анализ уровня востребованности выпускников среди работодателей, выберем атрибуты в axis x «Cluster», а в axis y «Уровень востребованности выпускников работодателями». На графике можно увидеть, что группа кластера 2 (C2) имеет наивысший балл среди других университетов по востребованности среди работодателей (Рис.16).

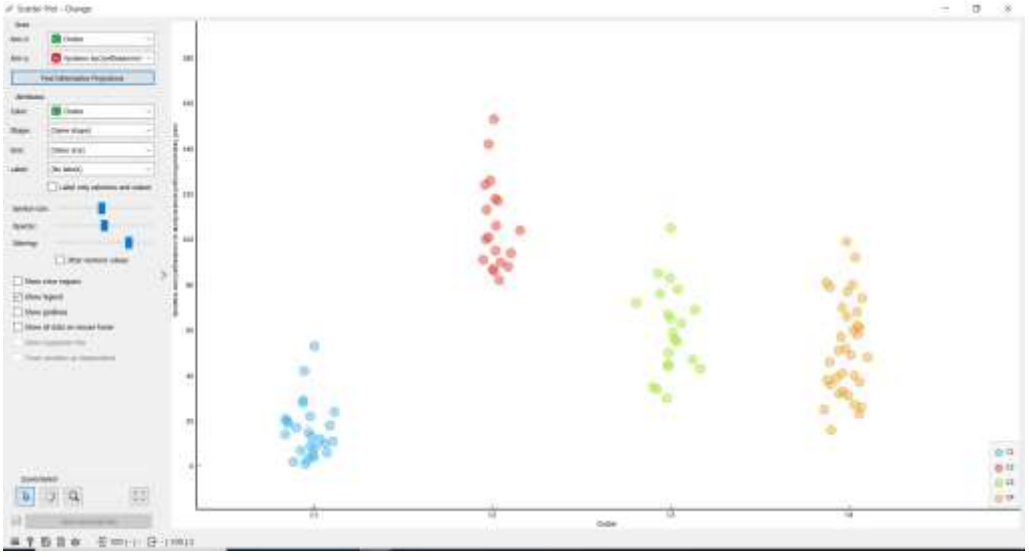

Рис.16. Кластерный анализ уровня востребованности выпускников среди работодателей

19) Далее проведем еще один кластерный анализ уровня научно-исследовательской деятельности, в атрибуте axis y выберем «а в axis y «Уровень востребованности выпускников работодателями». На графике можно заметить, что группа кластера 2 (C2) имеет наивысший балл среди других групп кластера (рис.17).

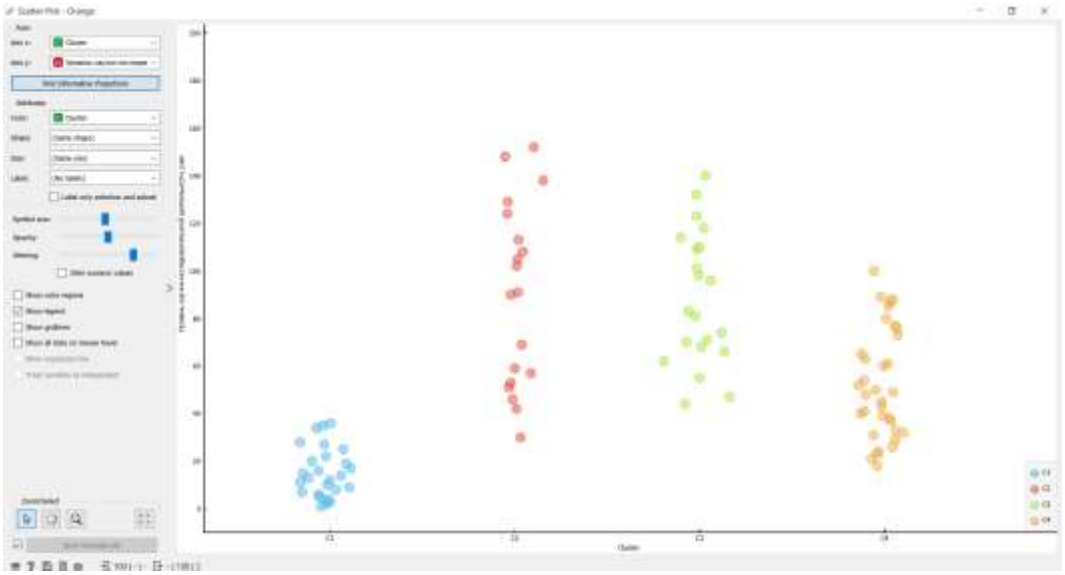

Рис.17. Кластерный анализ уровня научно-исследовательской деятельности

20) В итоге получилась готовая схема разведочного анализа (Рис.18).

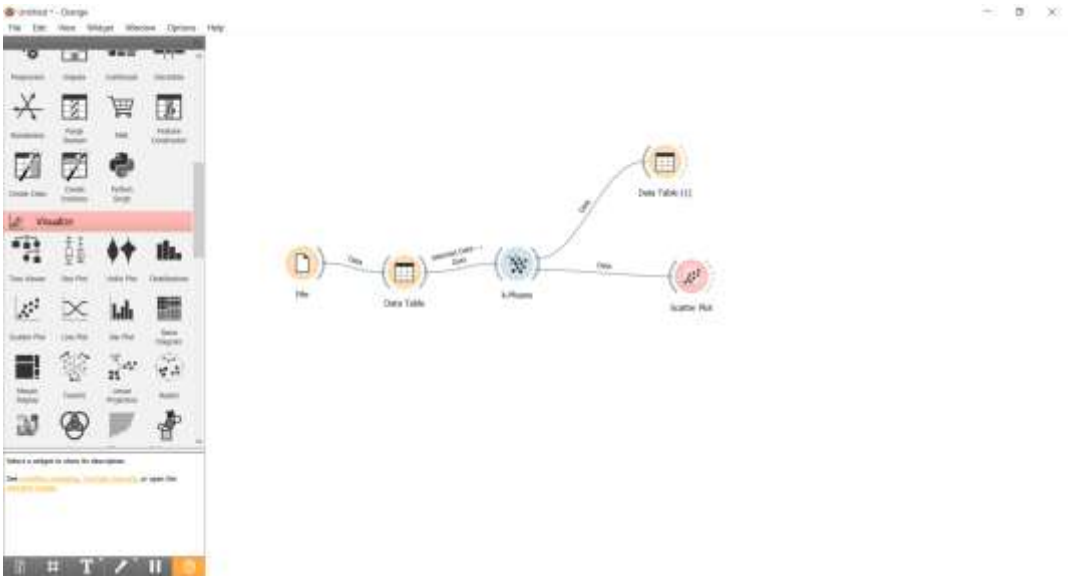

Рис.18. Итоговая схема

Выводы

В данной работе был выполнен кластерный анализ данных с помощью программного пакета визуального программирования на основе компонентов для визуализации данных Orange. С помощью виджетов File, Data Table, k-Means, Scatter Plot и выполнили кластерный анализ данных о рейтинге лучших вузов России по разным критериям.

Библиографический список

1. Мастевной С. С., Петрова А. Н. Data mining: обзор методов и области их применения // Наука, инновации и технологии: от идей к внедрению. 2022. С. 38-40.

2. Клименко А. В., Слащев И. С. Кластерный анализ данных //Вестник науки. 2019. Т. 1. №. 1. С. 159-163.
3. Гринченков Д. В. и др. Сравнительный анализ алгоритмов интеллектуального анализа данных //Моделирование. Теория, методы и средства. 2016. С. 263-266.
4. Маматкасымова А.Т., Кульматова Н.А. Orange: Использование системы визуального программирования при обработке больших данных// Материаловедение. 2022. №1(36). С. 2232.
5. Кривенко Ю. С., Минасян А. Т., Разиньков А. О. Исследование технологий интеллектуального анализа данных (Data Mining) // Актуальные проблемы управления в электронной экономике. 2018. С. 182-184.
6. Юсупов Н. Исследование методов кластеризации в программе Orange //Молодежная школа-семинар по проблемам управления в технических системах имени АА Вавилова. 2020. Т. 1. С. 35-37.
7. Токарев А. И., Брякин А. Н. Разведочный анализ данных и data mining // Перспективные направления развития отечественных информационных технологий. 2017. С. 201-203