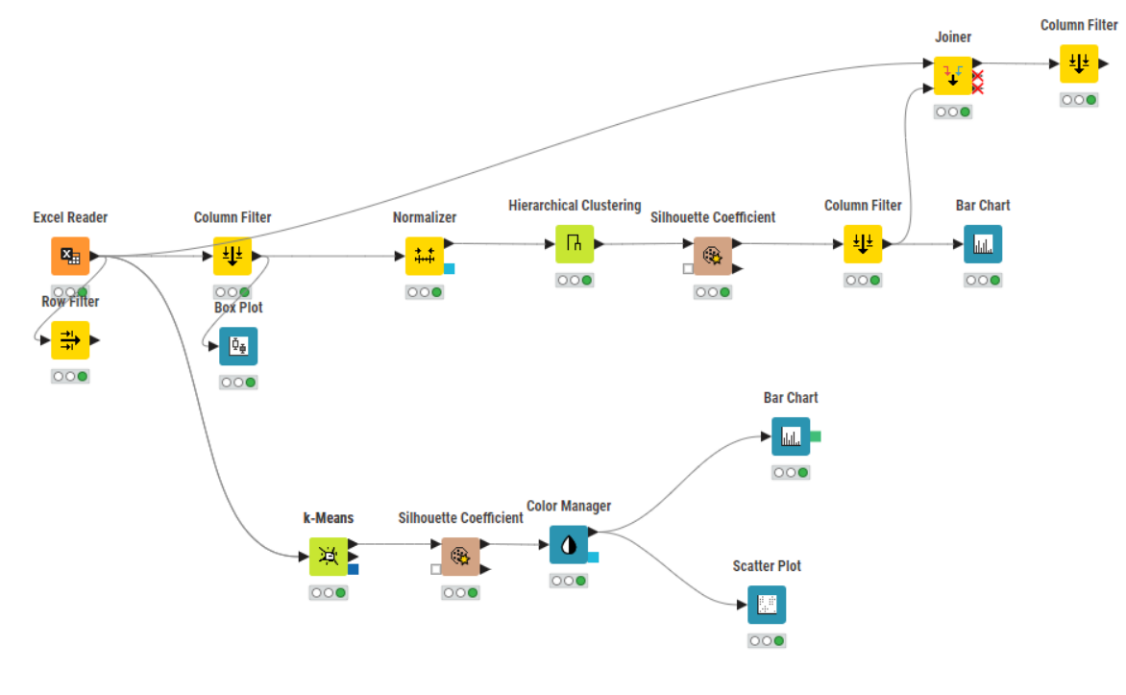
Аналитическая записка по результатам кластерного анализа с использованием программы KNIME

Выполнила: Лёвкина Екатерина гр. Э-2209



**АНАЛИЗ КЛАСТЕРОВ ПО ПОТРЕБЛЕНИЮ ПРОДУКТОВ ПИТАНИЯ**

Изображение выглядит как снимок экрана, текст, диаграмма, График

Контент, сгенерированный ИИ, может содержать ошибки.

Диаграмма показывает средние уровни потребления различных категорий продуктов в каждом из трех кластеров. Можно выделить основные закономерности:

Кластер ‘2’ характеризует наиболее выраженное потребление хлебобулочных— значительно выше, чем в других кластерах (около 45), сравнительно низкое потребление красного и белого мяса, рыбы, молока и крахмалистых продуктов, сбалансированное потребление фруктов и овощей. Можно предположить, что это страны с пищевой моделью, ориентированной на белое мясо и более здоровое питание.

Кластер ‘0’ характеризует умеренное потребление красного мяса (около 20), а также рыбы и молока, относительно высокое потребление белого мяса, яиц и хлебобулочных изделий. В целом, смешанный характер рациона, без ярко выраженных перекосов.

Кластер ‘1’ отличает высокое потребление красного мяса, молока и крахмалистых продуктов, относительно высокое потребление белого мяса и овощей и фруктов. Можно предположить, что это страны с более традиционным европейским или мясо-крахмалистым типом питания.

**РЕГИОНАЛЬНЫЙ АНАЛИЗ**

Изображение выглядит как текст, карта

Контент, сгенерированный ИИ, может содержать ошибки.Изображение выглядит как текст, снимок экрана, число, Параллельный

Контент, сгенерированный ИИ, может содержать ошибки.

* Кластер ‘2’ включает такие страны, как Албания, Болгария, Чехословакия, Греция, Венгрия, Италия, Польша, Португалия, Румыния, Испания, СССР, Югославия

Объединяет страны Средиземноморья (Италия, Испания, Греция, Португалия) и Восточной Европы (СССР, Польша, Румыния, Болгария). Первые ориентированы на белое мясо, овощи и злаки (средиземноморская диета), вторые тяготеют к более углеводному питанию с меньшим потреблением молочных продуктов. В целом, этот кластер характеризуется низким потреблением красного мяса и молочных продуктов и большим акцентом на хлебобулочные.

Страны этого кластера скорее всего характеризуются более здоровым образом жизни и низким уровнем сердечно-сосудистых заболеваний, благодаря высокому потреблению белого мяса, рыбы и овощей.

* Кластер ‘0’ включает страны Центральной Европы - Австрию, Бельгию, Германию, Нидерланды

Эти страны характеризуются высоким потреблением мяса и молока, но сбалансированным рационом в целом. Кластер объединяет страны близкие к традиционной западноевропейской модели питания. Рацион более традиционный для Западной Европы, но не такой мясоориентированный, как у северных стран. Страны данной группы могут быть более склонны к диетам с высоким содержанием насыщенных жиров и углеводов, что потенциально повышает риски ожирения и заболеваний сердца.

Данная группа занимает промежуточное положение, демонстрируя умеренный и сбалансированный рацион.

* Кластер ‘1’ включает страны Западной и Северной Европы: Данию, Финляндию, Францию, Ирландию, Норвегию, Швецию, Швейцарию, Великобританию

Это страны с высоким потреблением красного мяса, молочных продуктов и углеводов (картофель, хлеб), что характерно для северного климата, где калорийное питание исторически было необходимым. Страны характеризуются более калорийным рационом, что может коррелировать с большей распространенностью ожирения и сердечно-сосудистых заболеваний.

**РЕАЛИЗАЦИЯ АЛГОРИТМА КЛАСТЕРИЗАЦИИ**

1. **Инициализация данных**
   1. Импорт

Так как исходный файл представлен в формате CSV, я решила предварительно перевести его в формат xsls, открыв его в Excel. Таким образом, теперь таблицу можно импортировать в программу через Excel Reader.

* 1. Отбор признаков

В кластеризации не участвует столбец country, остальные признаки будем считать значимыми для характеристики элементов набора данных. Таким образом – отбираем все признаки кроме ‘country’, а именно:

* ‘redmeat’ Мясо (Red meat)
* ‘whitemea’ Птица (White meat)
* ‘eggs’ Яйца (Eggs)
* ‘milk’ Молоко (Milk)
* ‘fish’ Рыба (Fish)
* ‘cereals’ Хлебо-булочные (Cereals)
* ‘starch’ Макароны (?) Крахмал (?) (Starchy foods)
* ‘nuts’ Орехи (Pulses, nuts, and oil-seeds)
* ‘fruits\_v’ Фрукты и овощи (Fruits and vegetables)

В данном случае все признаки являются количественными и описывают объем потребеления каждого вида продуктов

1. **Предобработка данных для кластеризации** 
   1. Пропущенных значений в таблице нет
   2. Выбросы

Изображение выглядит как диаграмма, текст, линия, План

Контент, сгенерированный ИИ, может содержать ошибки.

На графике Box Plot видно, что есть экстремальные значения по признакам ‘redmeat’ и ‘fish’. Однако в данном случае убирать данные из рассмотрения нецелесообразно ввиду их смысла – потребление по странам может сильно различаться.

* 1. Масштабирование

Вновь обратимся к графику Box Plot. Видно, что разброс значений по признакам нельзя назвать однородным, а разные диапазоны значений могут иказить результаты работы алгоритма кластеризации.

Использовать будем MinMaxScaler , потому что данные не имеют явных выбросов. Данный метод сохраняет изначальное распределение данных, но сглаживает “выбросы”. К тому же он хорошо работает для модели кластеризации.

* 1. Дубликатов нет

1. **Кластеризация-1.** 
   1. Определение количества кластеров для разбиения

Для кластеризации используем метод Hierarchical Clusteing. С помощью OpenView в нем открываем дендограмму и методом отсечения получаем, что оптимальное количество кластеров – 3:

Изображение выглядит как диаграмма, Прямоугольник, зарисовка, линия

Контент, сгенерированный ИИ, может содержать ошибки.



В данном случае можно видеть, что наиболее значительное разделение происходит примерно на уровне **0.8–1.0**. Если провести горизонтальную линию на этом уровне, она пересечёт около **3-4 линий**. Более точное количество кластеров, можно получить, используя **метод "локтя"** .

Изображение выглядит как текст, График, снимок экрана, линия

Контент, сгенерированный ИИ, может содержать ошибки.



Резкий спад виден на уровне 3. Таким образом, делить будем на 3 кластера.

* 1. Индекс силуэта



График позволяет оценить качество кластеризации. Основной критерий хорошего разбиения – близость к 1 коэффициентов. Для кластеров ‘0’ и ‘1’ наблюдается высокий уровень компактности и разделимости – коэффициент ~0.4 составляет свидетельствует о хорошем разделении этих кластеров, так как их объекты близки к своим центрам и далеки от других кластеров. Кластер ‘2’ наименее однородный и возможно содержит объекты, которые могли бы принадлежать другому кластеру. Однако методом перебора (отклоняясь от значения 3 в меньшую или большую сторону) выявлено, что данное разбиение – самое удачное. В иных случаях индекс силуэта уменьшается для всех групп, следовательно качество кластеризации падает.

