Министерство образования Республики Беларусь

Учреждение образования

«Брестский государственный технический университет»

Кафедра ИИТ

**Лабораторная работа №2**

По дисциплине «Методы обработки данных»

Выполнил:

студент 4 курса

группы ПО-7

Комиссаров А.Е.

Проверил:

Савицкий Ю.В.

Брест, 2023

**Цель:** изучить основы методов Machine Learning в контексте задачи кластерного анализа (cluster analysis), приобрести навыки работы с методами Machine Learning в системе STATISTICA StatSoft, осуществить обработку методами Machine Learning индивидуального набора данных и интерпретацию результатов.

**Вариант:** 7

**Ход работы:**

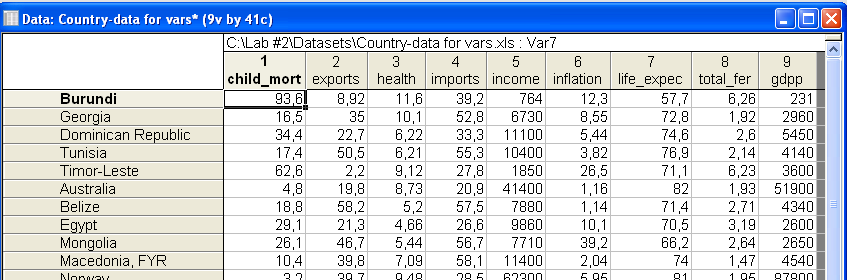
1) изучить теоретические сведения

2) приобрести навыки работы с методами Machine Learning в контексте задачи кластерного анализа в системе STATISTICA StatSoft, реализуя приведенный ниже пример

3) на основе приобретенных практических навыков осуществить все этапы обработки методами Machine Learning в контексте задачи кластерного анализа и интерпретацию результатов согласно варианту индивидуального задания

4) оформить отчет и подготовиться к защите лабораторной работы по полученным результатам и контрольным вопросам

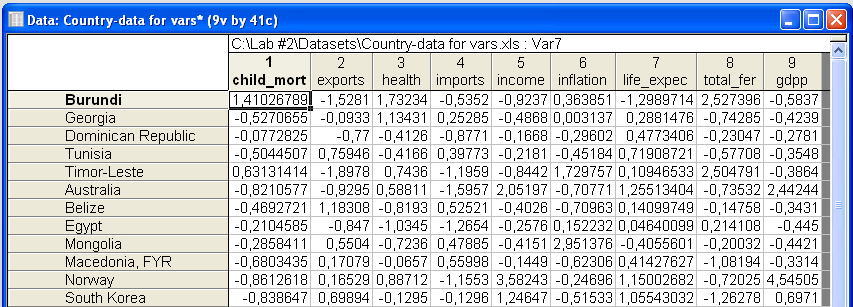
Импорт данных: загружаем 7 лист из вариантов заданий



Исходный файл содержит информацию о различных странах мира – экспорт, импорт, инфляция, средняя продолжительность жизни и тд.

Поскольку различные измерения используют абсолютно различные типы шкал, данные необходимо стандартизовать. Для стандартизации в меню Данные выберем пункт Standartize, после чего каждая переменная будет иметь среднее 0 и стандартное отклонение 1.

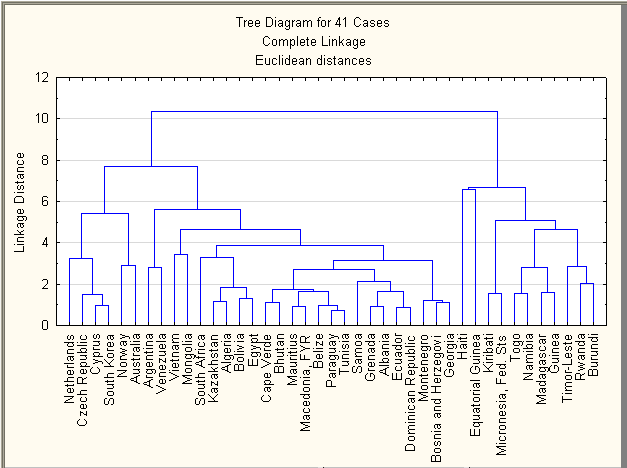
Таблица со стандартизованными переменными приведена дальше.



Шаг 1. Иерархическая классификация

На первом этапе выясним, формируют ли автомобили "естественные" кластеры.

Выберем Кластерный анализ в меню Анализ (Statistics) и проведём древовидную кластеризацию, результатом которой будет построено дендограммное дерево:

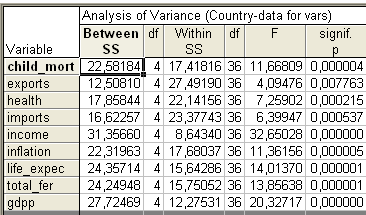


На данном графике отображено разделение на кластеры с каждым шагом, где в самом конце исходные элементы разбиты попарно. Здесь можно выделить, смотря на график на дистанции 2-6 (по оси Linkage Distance), около 3-5 кластеров.

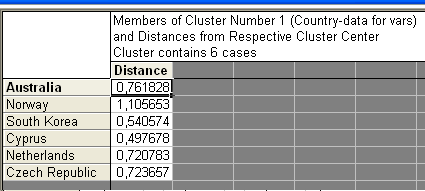
Шаг 2. Кластеризация методом К средних.

Исходя из визуального представления результатов, можно сделать предположение, что страны образуют 3-5 естественных кластера. Проверим данное предположение, разбив исходные данные методом К средних на 5 кластеров, и проверим значимость различия между полученными группами.

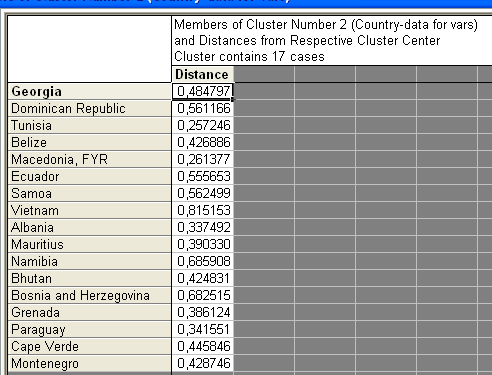
В Стартовой панели модуля Кластерный анализ выберем Кластеризация методом К средних



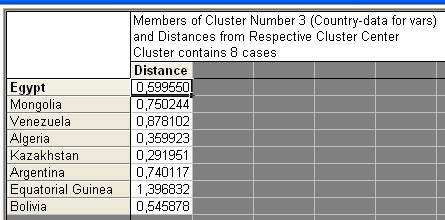
Итак, значение р<0.05, что говорит о значимом различии.

Нажмем кнопку Элементы кластеров и расстояния для просмотра наблюдений, входящих в каждый из кластеров.

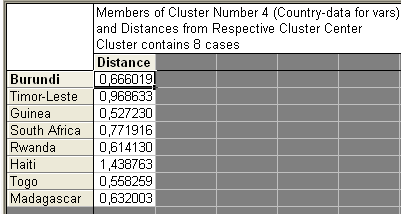
Первый кластер -



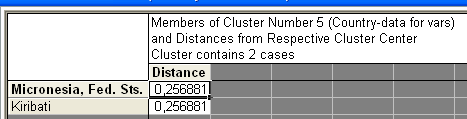
Второй кластер -



Третий кластер -



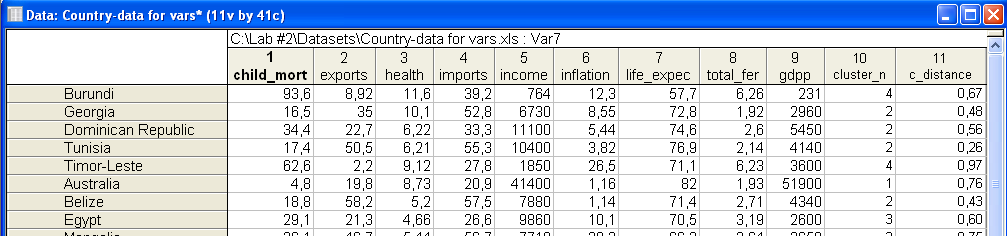
Четвёртый кластер -



Пятый кластер -

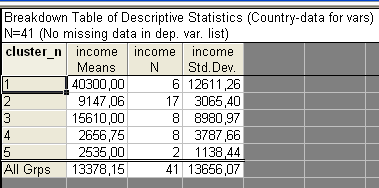
Шаг 3. Описательные статистики

Нажмем кнопку Сохранить классификацию и расстояния. Таблица стандартизованных данных дополнится информацией о кластере, к которому принадлежит наблюдение, евклидовом расстоянии и номере наблюдения.

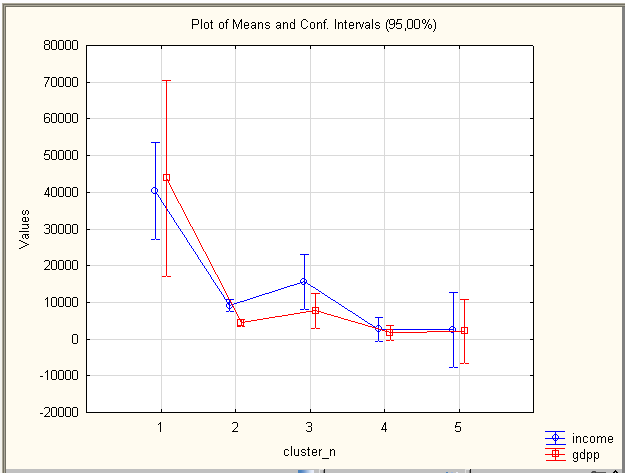
****

Cluster\_n – номер кластера, c\_distance – расстояние от центра кластера.

Теперь для каждого кластера можно вычислить основные описательные статистики. В меню Анализ – Основные статистики и таблицы выберем опцию Группировка и Однофакторный Дисперсионный анализ (Breakdown & one-way ANOVA). В опции Variables определяем исследуемые переменные.

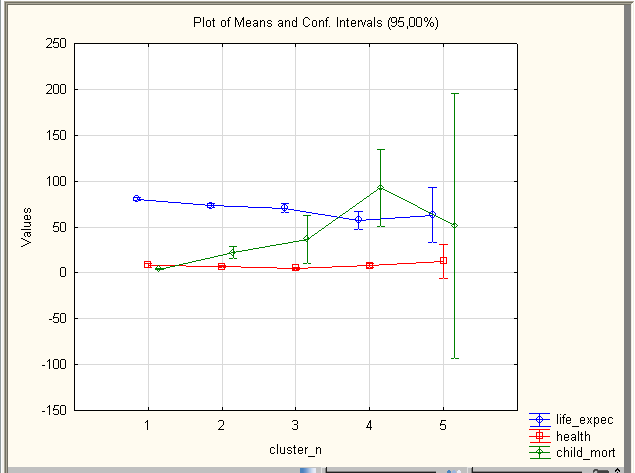


Повторяя процесс для каждой переменной, можно получить таблицы описательных статистик для каждого из показателей. Данная информация помогает увидеть различия между элементами разных кластерамов. Однако можно отразить всё на графике, где будет показано более наглядно.

Например: 

На данном графике отражены кластеры 1,2,3,4,5 (слева направо) и разным цветом показаны переменные – средний доход по стране, синим цветом и ВВП страны, красным. По графику видно, что у кластера стран с высоким ВВП также высокий доход на душу населения.

Следующий пример:



На данном графике отражены средняя продолжительность жизни, затраты на здоровье\медицину по стране и смертность среди детей. Видно что в странах с более высокими затратами ниже смертность и выше средняя продолжительность жизни. Эти данные могут использовать в мед. учреждениях либо при распределении средств по странам для отправки материальной помощи в виде медикаментов.

**Выводы:** я изучил основы методов Machine Learning в контексте задачи кластерного анализа (cluster analysis), приобрёл навыки работы с методами Machine Learning в системе STATISTICA StatSoft, осуществил обработку методами Machine Learning индивидуального набора данных и интерпретировал результат.