

## **Анализ данных в аналитической платформе Loginom**

<b>Введение .....</b>	<b>2</b>
<b>1. Исходные данные .....</b>	<b>6</b>
<b>2. Кластеризация .....</b>	<b>7</b>

## Введение

**Loginom** – платформа для решения большого спектра бизнес-задач, требующих обработки больших объемов данных, реализации сложной логики и применения методов машинного обучения.

Используя платформу Loginom, можно решать следующие бизнес-задачи:

- управление рисками: кредитный конвейер, скоринг, антифрод;
- клиентская аналитика: сегментация клиентов, противодействие оттоку, кросс-продажи;
- очистка данных: очистка и удаление дублей, создание золотой записи, стандартизация НСИ (нормативно-справочная информация);
- маркетинг: директ-маркетинг, оптимизация цен, оценка эффективности рекламы;
- логистика: прогнозирование спроса, оптимизация запасов, расчет страховых запасов;
- диагностика: статистический контроль качества, оценка вероятности поломок, цифровые двойники.

Процесс анализа данных в Loginom представлен на рисунке 1.

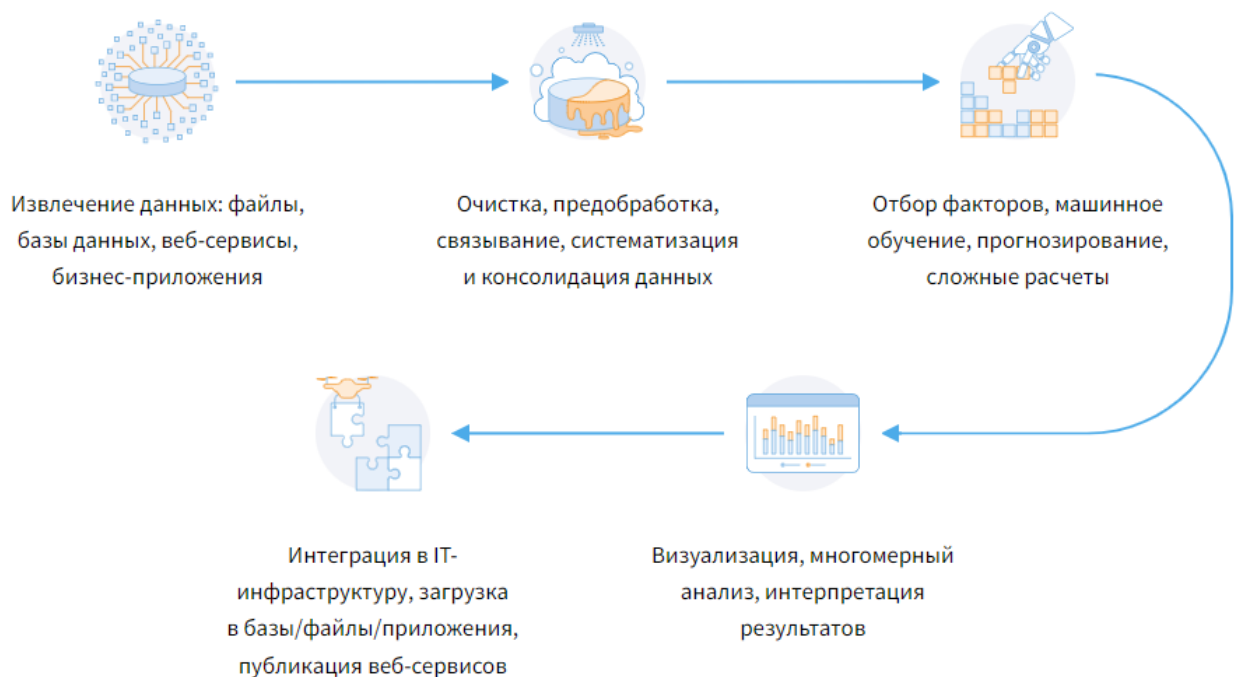


Рисунок 1 – Процесс анализа данных в Loginom

Платформа Loginom позволяет подключиться к множеству источников/приемников данных и настроить ETL-процессы. Интеграция со сторонними веб-сервисами и публикация собственных упрощает интеграцию в IT-инфраструктуру любой компании (рисунок 2).

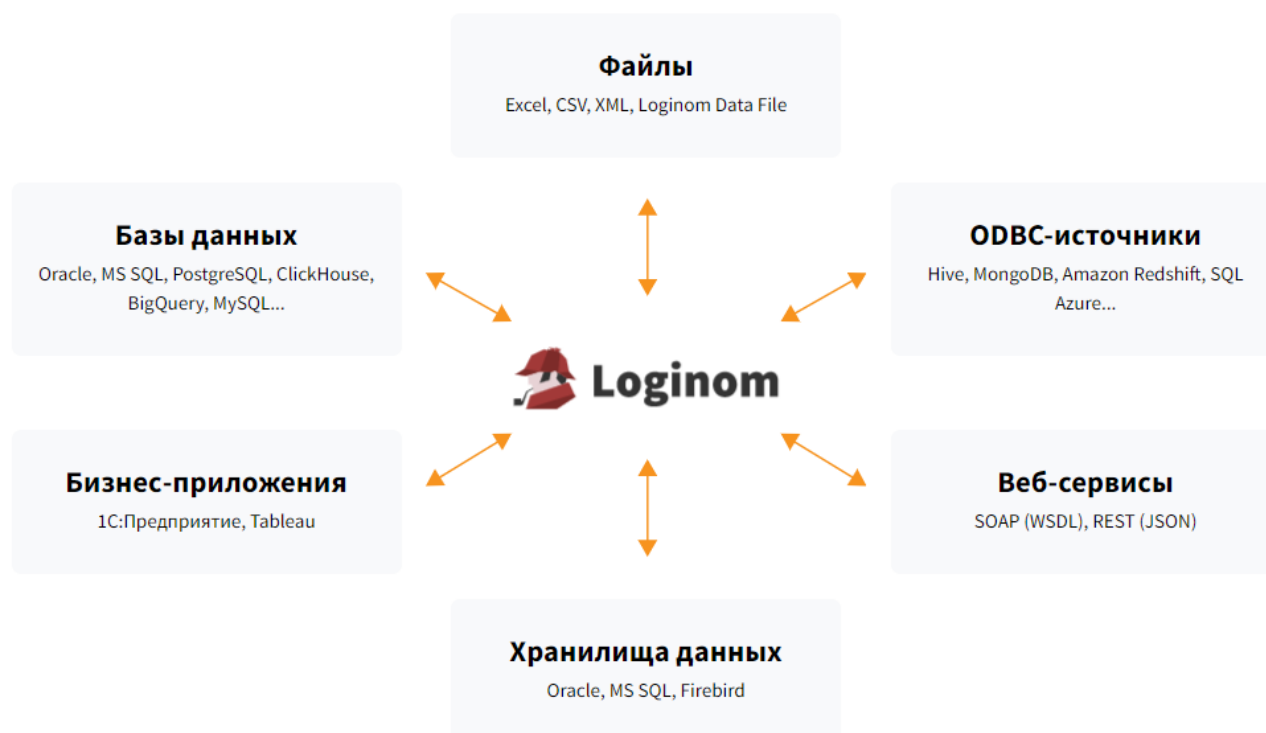


Рисунок 2 – Интеграция данных Loginom

Loginom поддерживает десятки вариантов визуализации больших наборов данных и формирование панелей отчетов для представления результатов обработки конечным пользователям:

- **OLAP–куб.** Визуализация многомерных данных с возможностью сортировки, группировки, фильтрации и агрегации данных и расчета многомерных формул «на лету». Связанные с кросс-таблицей диаграммы. Возможность детализации по любой ячейке.

- **Табличные данные.** Отображение огромных массивов данных в браузере с возможностью фильтрации, сортировки, форматирования. Визуализация любых статистических показателей.

- **Специализированные визуализаторы.** Специализированные визуализаторы позволяют оценить качество модели и интерпретировать результаты. Например, отобразить ROC-кривую и матрицу сопряженности для модели бинарной классификации или сравнить профили кластеров.

- **Панель отчетов.** Возможность вынести любой из настроенных визуализаторов на специальную панель отчетов. Пользователь с правами только на просмотр отчетов будет видеть результаты обработки, но не сценарии, при помощи которых результаты получены.


Для некоммерческого использования есть бесплатная клиентская версия Loginom Community Edition.

Для установки бесплатной версии Loginom Community Edition

(<https://loginom.ru/download>) (рисунок 3) необходимо заполнить анкету и на указанный email придет ссылка на скачивание программного продукта.

# Скачать Loginom Community Edition

Бесплатная версия для некоммерческого использования



**Loginom** — интегрированная Low-code платформа для реализации всех аналитических процессов: от консолидации и подготовки данных до моделирования, развертывания и визуализации

- Визуальная настройка логики обработки и повторное использование компонентов
- Продвинутая аналитика: от простейших формул до машинного обучения
- Интеграция данных: файлы, базы данных, учетные системы, включая Excel и 1С:Предприятие
- Доступ к ClickHouse
- Экспорт в Tableau
- Подключение и вызов REST-сервисов
- Высокая производительность: in-memory, параллельная обработка, быстрые алгоритмы

БЕСПЛАТНО

## Loginom CE 7.0.1

Заполните анкету, и мы вышлем вам ссылку для скачивания на указанный email.

На сайте [loginom.com](https://loginom.com) вы можете скачать версию с английским интерфейсом

Фамилия: \*

Имя: \*

Отчество:

Email: \*

Компания:

Должность:

Телефон: \*

Ваш профессиональный статус: \*

— ▼

☒ Подписаться на новостной дайджест Loginom (рассылка выходит один раз в месяц)

Получить ссылку

Оставляя заявку, вы даете [согласие на обработку персональных данных](#)

Рисунок 3 – Получение дистрибутива Loginom

**Стартовое окно** содержит следующие команды для манипуляции с пакетами (рисунок 4):

- **Создать новый пакет** – позволяет создать и сохранить новый Пакет.
- **Создать черновик** – создает временный Пакет и позволяет работать с ним, не сохраняя его.
- **Открыть пакет** – позволяет открыть существующий Пакет.

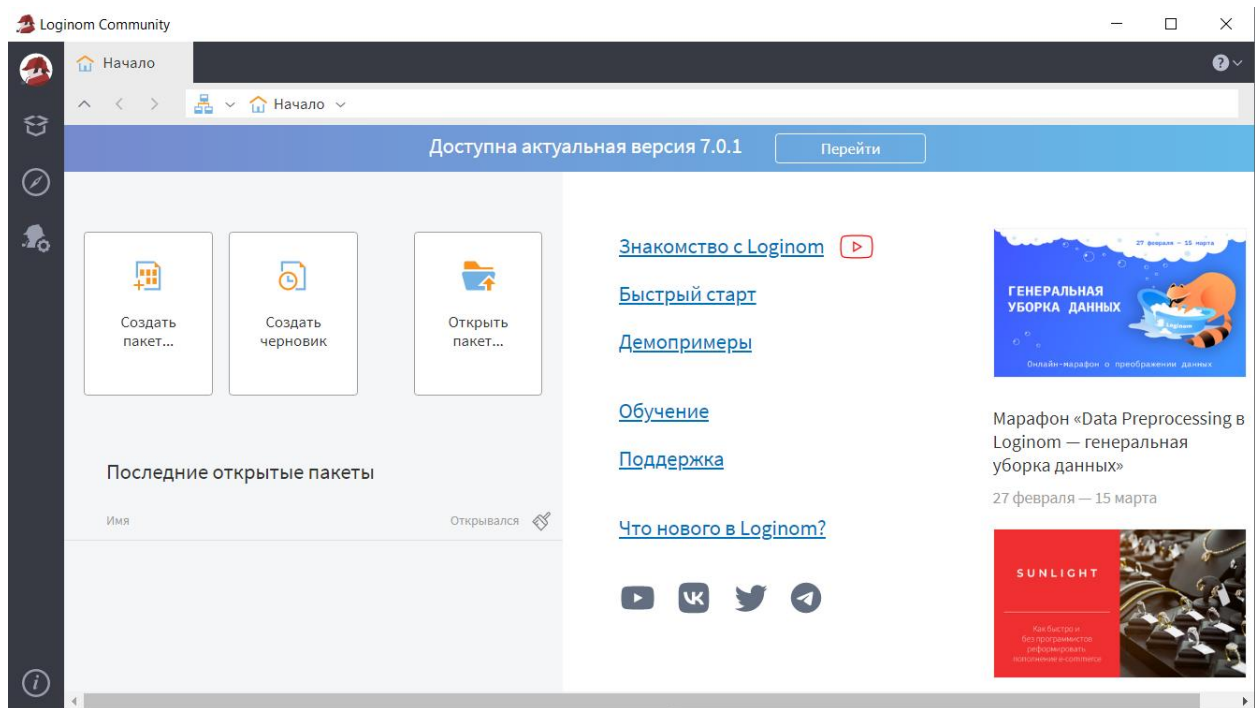


Рисунок 4 – Стартовое окно Loginom

## 1. Исходные данные

Прежде чем приступить к работе, скачайте на локальный диск файл **insurance.csv**. Файл содержит данных о расходах на медицинское обслуживание тех, кто имеет медицинскую страховку. Файл взят с сайта [kaggle.com](https://www.kaggle.com/mirichoi0218/insurance) (<https://www.kaggle.com/mirichoi0218/insurance>).

**Kaggle** – система организации конкурсов по исследованию данных, а также социальная сеть специалистов по обработке данных и машинному обучению. Принадлежит корпорации Google с марта 2017 года.

### Описание переменных набора:

- age – возраст основного бенефициара;
- sex – пол застрахованного;
- bmi – индекс массы тела;
- children – число детей, охваченных медицинским страхованием / число иждивенцев;
- smoker – курит ли застрахованный;
- region – жилой район получателя в США, Северо-Восток, ЮгоВосток, Юго-Запад, Северо-Запад;
- charges – индивидуальные медицинские расходы, оплачиваемые страховкой.

## 2. Кластеризация

Кластеризация (сегментация) – это группировка объектов (наблюдений, событий) на основе данных, описывающих свойства объектов. Объекты внутри кластера должны быть похожими друг на друга и отличаться от других, которые вошли в другие кластеры.

В Loginom обработчик производит кластеризацию объектов на основе алгоритмов k-means и g-means. Если количество кластеров известно, то применяется алгоритм k-means, в противном случае – g-means, который определяет это количество автоматически в рамках заданного интервала.

Создайте **новый пакет** «Медицинское\_страхование». Для этого необходимо выбрать «Создать пакет...» и в окне «Сохранение» указать путь, задать имя файла и нажать кнопку «Сохранить» (рисунок 5).

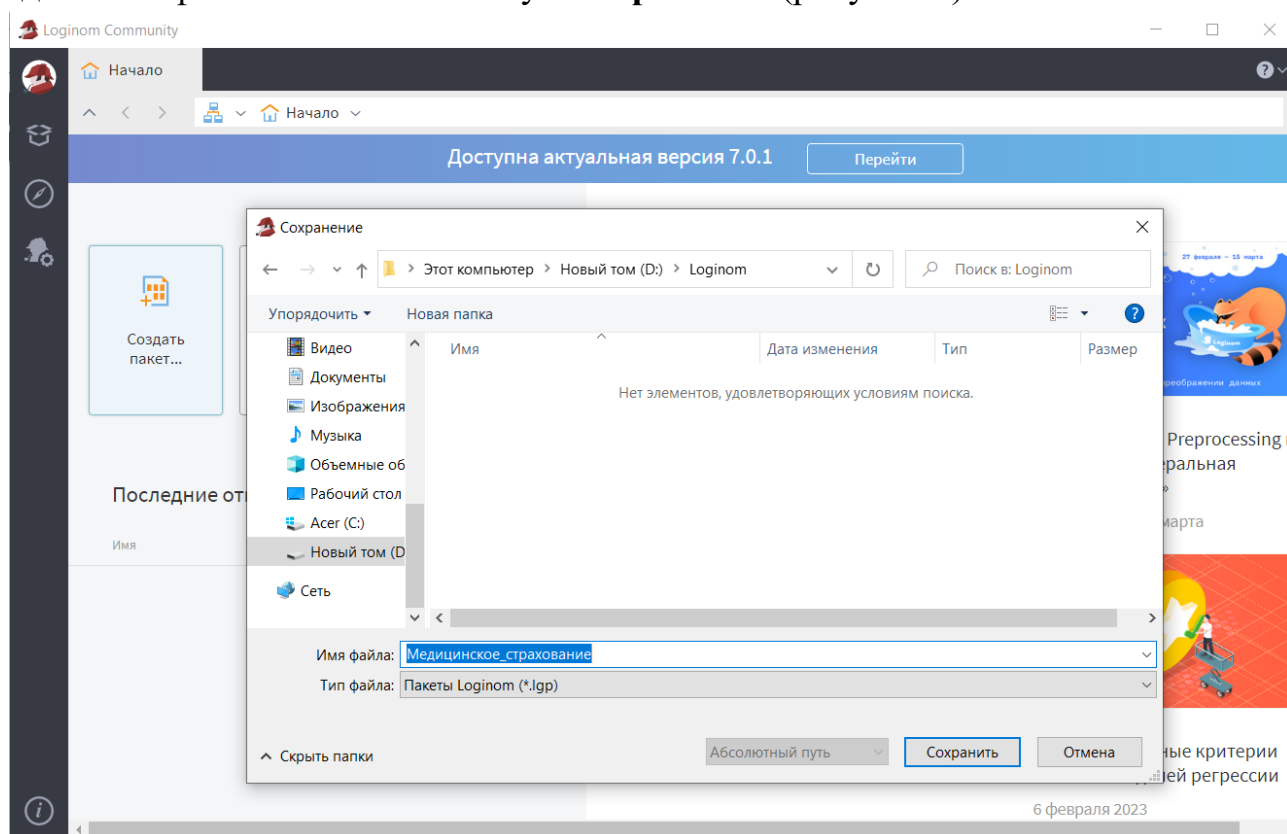


Рисунок 5 – Создание нового пакета

В первом созданном пакете по умолчанию создается «Модуль1», включающий «Сценарий», который пока является пустым (рисунок 6).

На рисунке 6 показаны основные блоки интерфейса:

1. **Главное меню** – позволяет пользователю начать/завершить работу, получить доступ к пакетам и настройкам платформы и др. Состав элементов главного меню может изменяться в зависимости от редакции платформы (Server/Desktop), прав пользователя, а также при работе с визуализаторами.

2. **Адресная строка** – строка, содержащая путь к открытому объекту.

3. **Рабочее пространство** – область, в которой осуществляются основные действия по построению сценария, настройке подключений, отчетов, администрированию платформы и т.п. Состав элементов, доступных команд и визуальное отображение этой области зависят от того, какая страница платформы находится в активном состоянии.

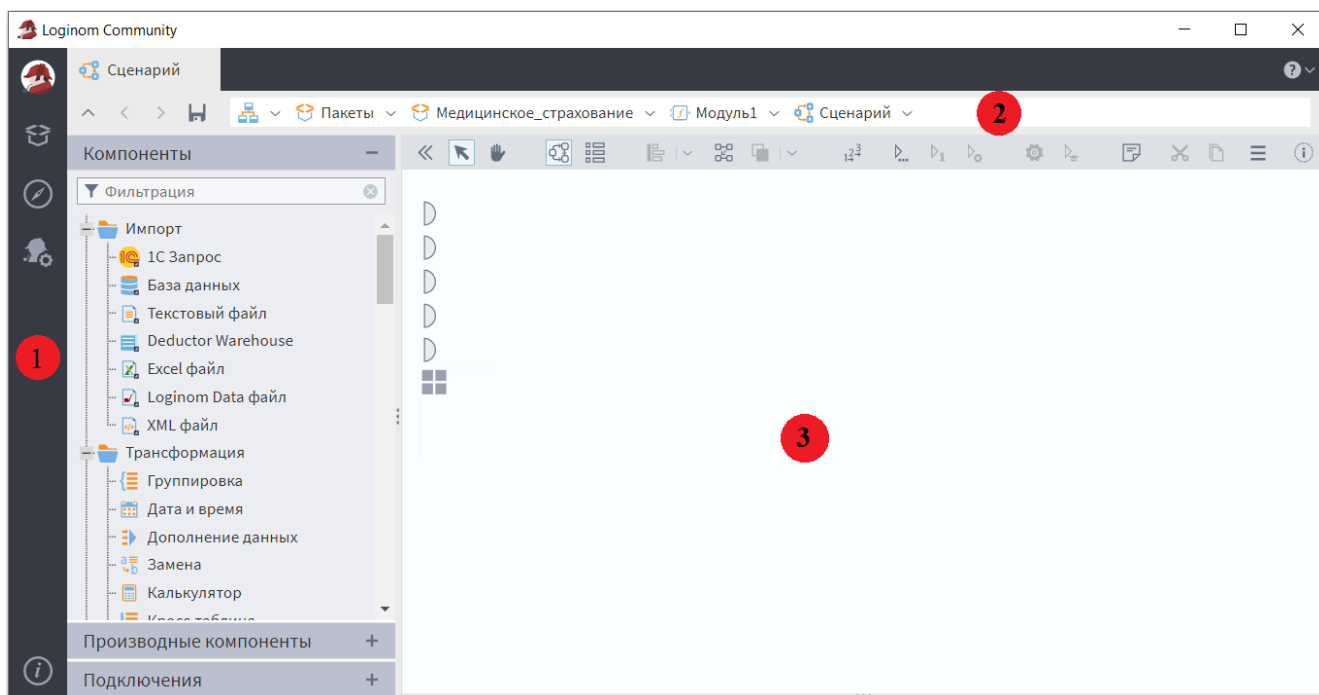


Рисунок 6 – Сценарий пакета «Медицинское\_страхование»

**Сценарий** – главная составная часть модуля и представляет собой последовательность шагов по обработке данных. Шаги задаются узлами из стандартных или производных компонентов.

Сценарий по умолчанию пустой и заполняется необходимыми компонентами в зависимости от решаемой задачи путем их добавления в область сценария.

Компоненты добавляются в сценарий перетаскиванием из панели в рабочую область.

В сценарий из категории «Импорт» добавьте первый узел «Текстовый файл» (рисунок 7).

Для настройки узла «Текстовый файл» необходимо нажать на знак настройки (шестеренка внутри узла) и выполнить следующие шаги:

1. Выполнить импорт из текстового файла. Для этого нужно выбрать нужный файл и нажать на кнопку «Далее» (рисунок 8).



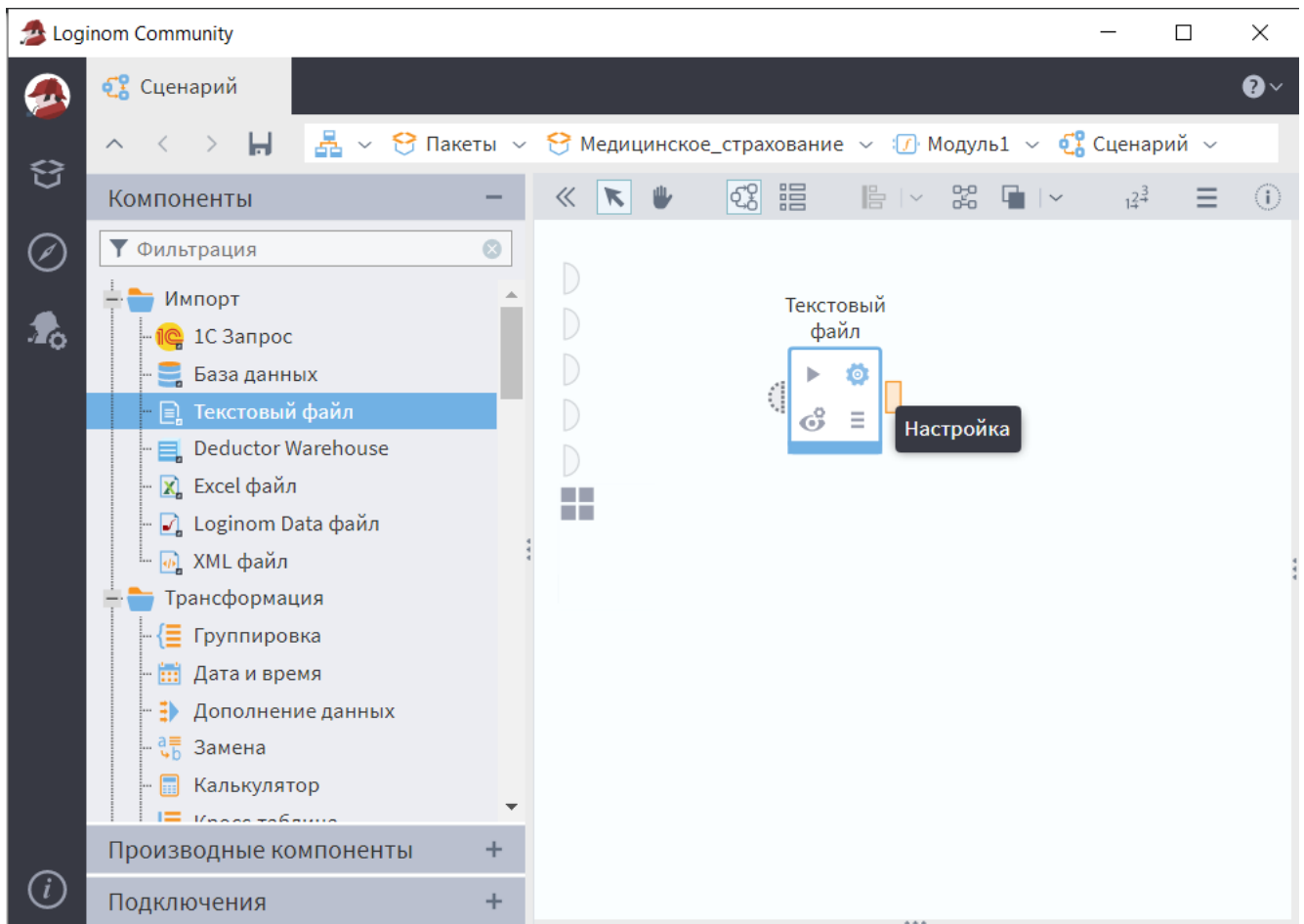


Рисунок 7 – Узел «Текстовый файл»

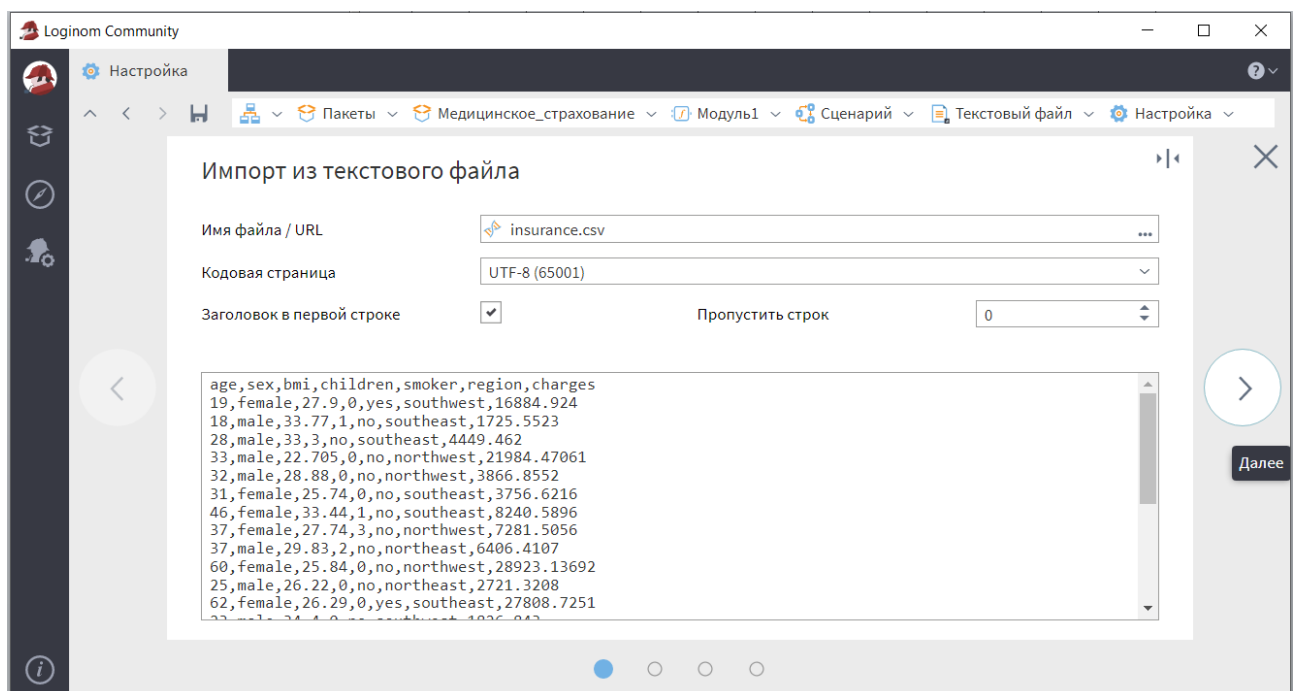


Рисунок 8 – Окно «Импорт из текстового файла»

В разделе **столбцов** выберите «Запятая». При настройке форматов импорта два поля «bmi» и «charges» определены с ошибочным типом данных (рисунок 9). Это связано с десятичным разделителем, который в импортируемых данных точка, а по умолчанию мастер настройки ожидает запятую. Выберите в качестве **десятичного разделителя** «Точка» и нажмите по кнопке «**Определять автоматически**». Теперь тип данных для этих полей будет верно настроен (рисунок 10). Нажмите кнопку «Далее».

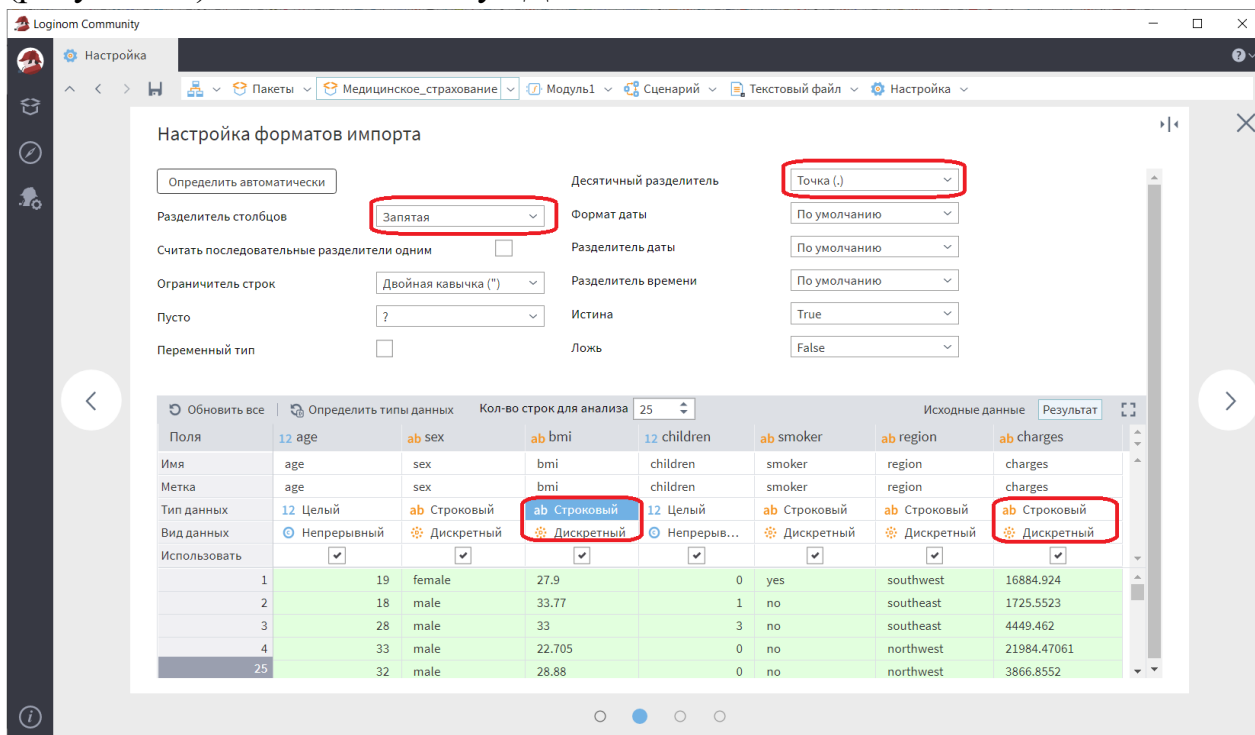


Рисунок 9 – Окно «Настройка формата импорта»

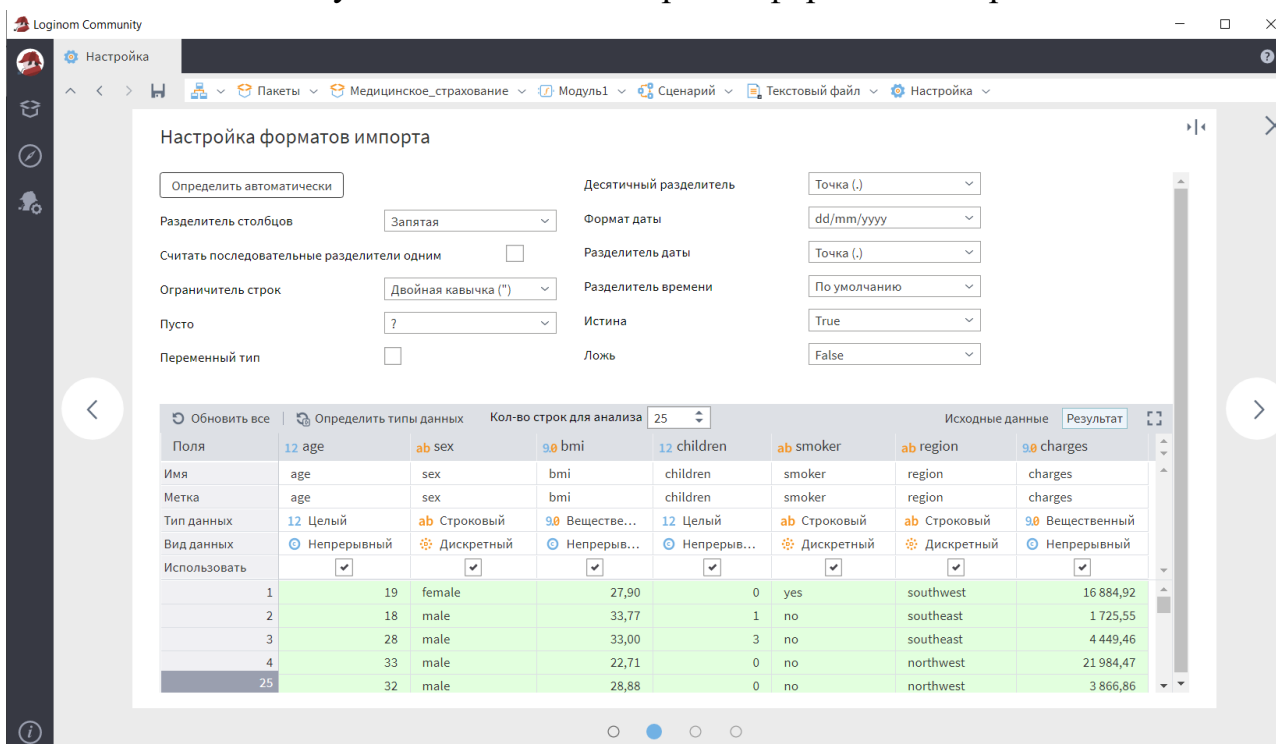


Рисунок 10 – Окно «Настройка формата импорта»

2. Ничего не меняя, нажмите «Далее» (рисунок 11).

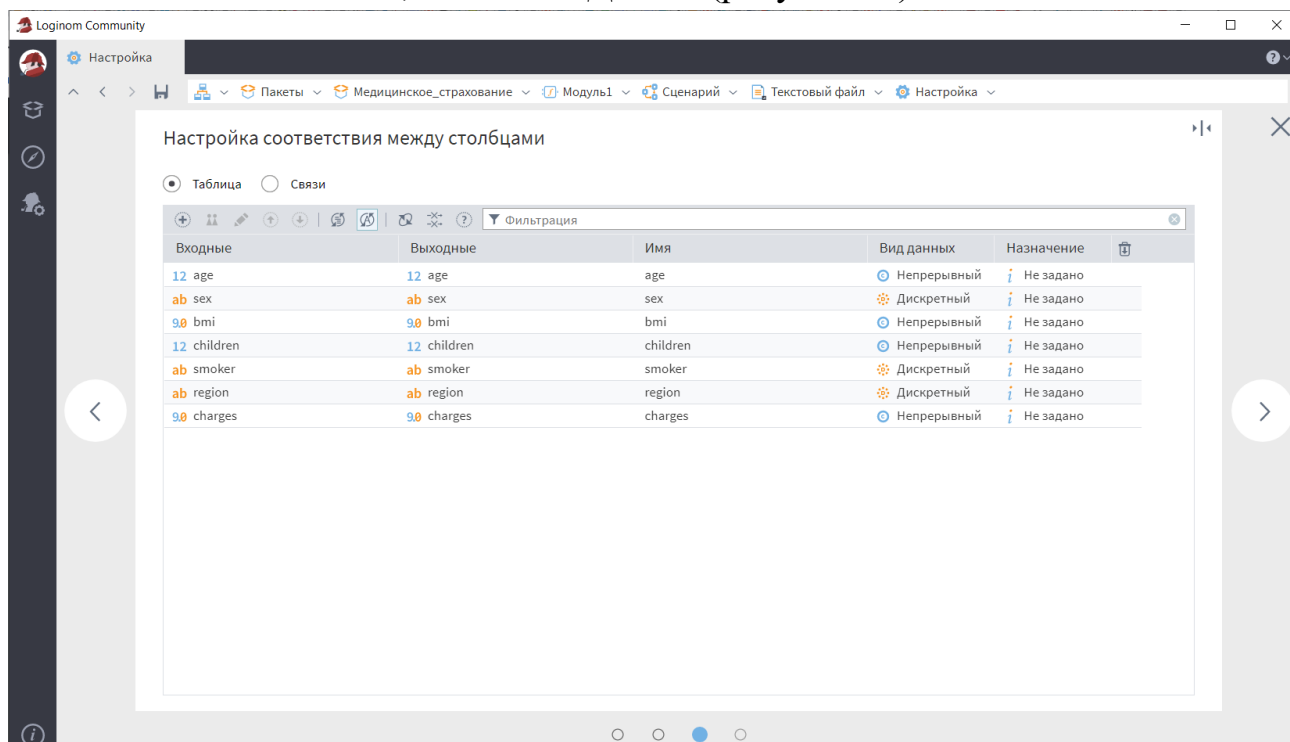


Рисунок 11 – Окно «Настройка соответствия между столбцами»

3. В окне «Описание узла» для завершения настройки нажмите «Сохранить» (рисунок 12).

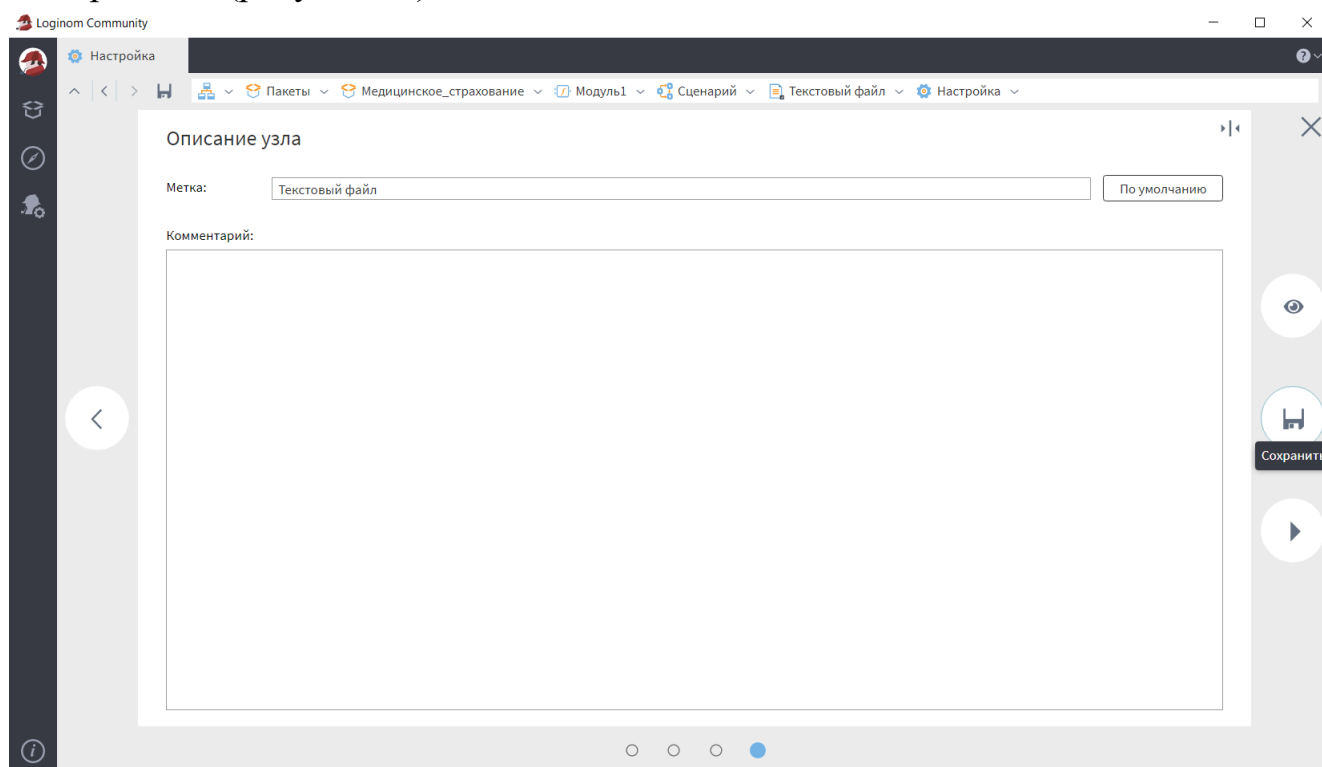


Рисунок 12 – Окно «Описание узла»

Далее добавьте в рабочую область сценария узел «Кластеризация» (рисунок

13).

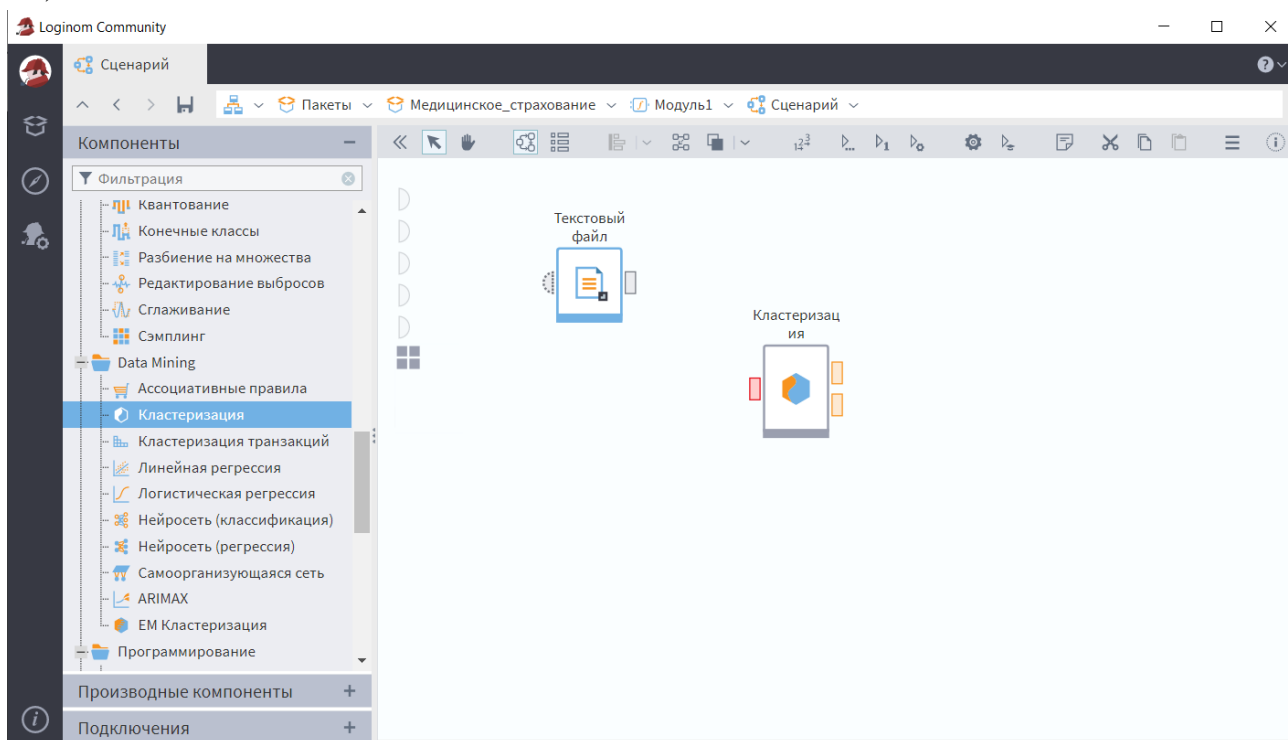


Рисунок 13 – Добавление узла Кластеризация

Далее необходимо установить связь между набором данных из текстового файла с входным источником данных в кластеризации (рисунок 14).

Добавить комментарий (желтый прямоугольный блок) можно с помощью контекстного меню (нажатие правой кнопки мыши).

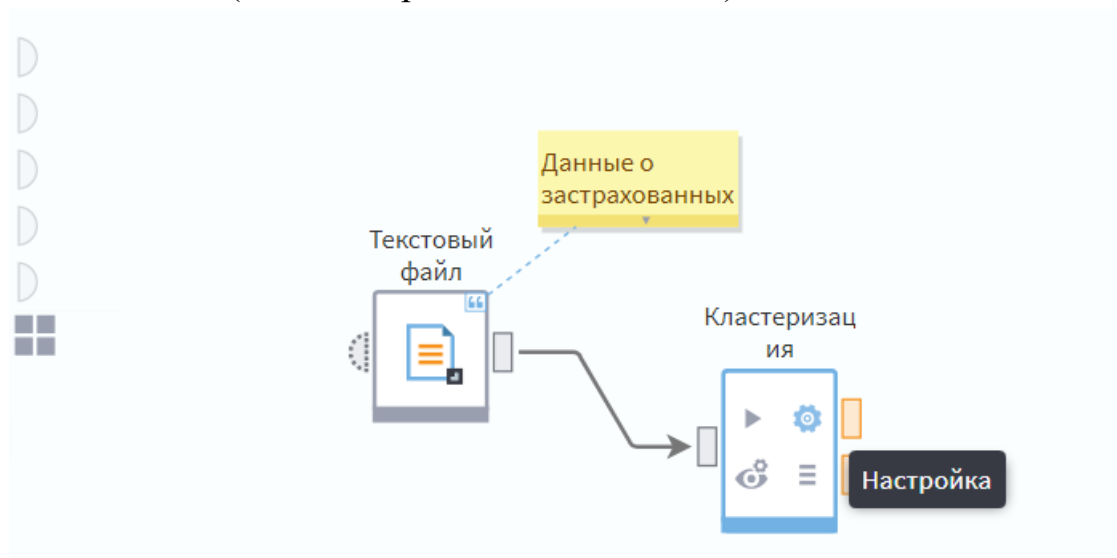


Рисунок 14 – Установка связи

Перейдем к настройке узла «Кластеризация». Для настройки узла необходимо нажать на знак настройки (шестеренка внутри узла) и выполнить следующие шаги:

1. В первом окне необходимо произвести **настройку входных столбцов**

(рисунок 15). Дважды щелкните левой кнопкой мышки на ячейку «Не задано» в столбце «Назначение» для тех полей, по значениям которых должна быть произведена кластеризация. Измените значение ячейки на «Используемое» и нажмите кнопку «Применить» (рисунок 15). Только столбец «charges» не используется для кластеризации (рисунок 16). Нажмите «Далее».

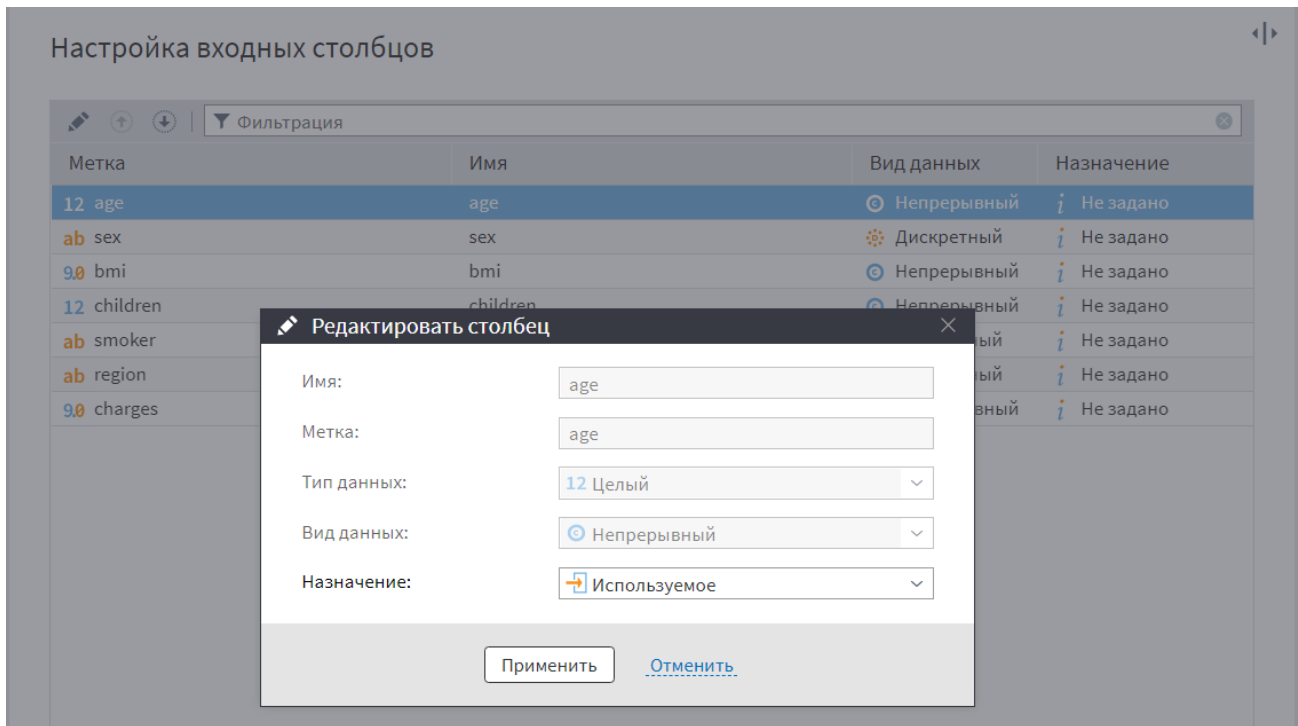


Рисунок 15 – Окно «Настройка входных столбцов»

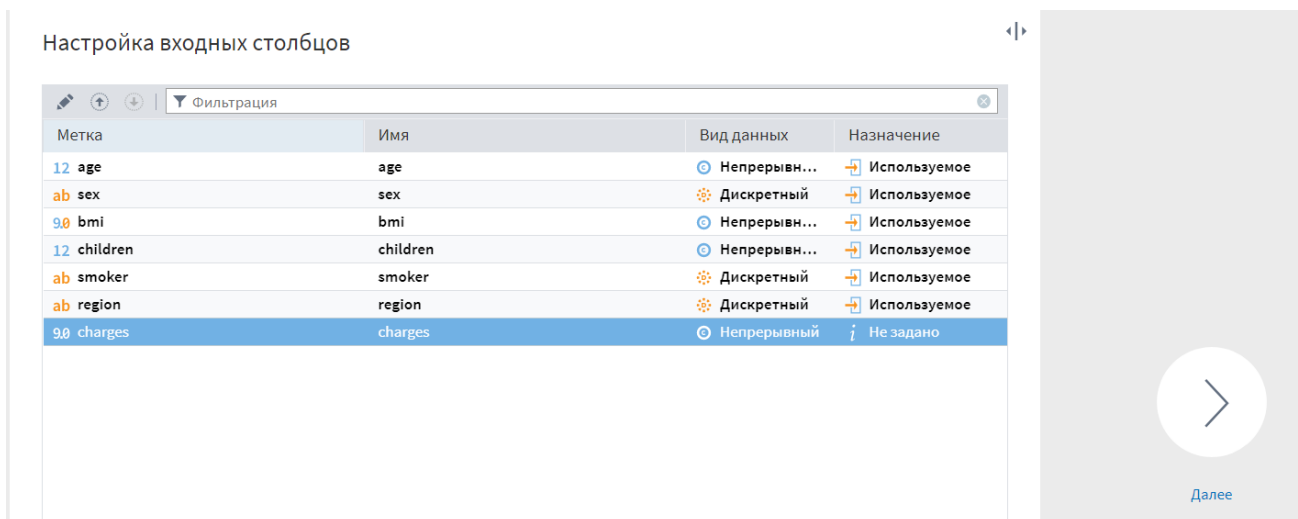


Рисунок 16 – Окно «Настройка входных столбцов»

2. В окне «Настройки нормализации» нажмите на кнопку «Далее» (рисунок 17).

Настройки нормализации

Состояние входа: Не активировано [Активировать](#)

Разрешить пропущенные значения: ☐

Поле	Нормализатор
Входные	
12 age	Стандартизация
ab sex	Индикатор (без опорной категории)
9.0 bmi	Стандартизация
12 children	Стандартизация

Параметры нормализации

Не выбран нормализатор

Далее

Рисунок 17 – Окно «Настройки нормализации»

3. В окне «Кластеризация» примените следующие настройки: уберите галочку с «Автоопределения числа кластеров».

Число кластеров определите равным трем (рисунок 18). В кластеризации будет реализован алгоритм k-means.

Для алгоритма g-means нужно было оставить настройки в этом диалоговом окне по умолчанию, т.е. автоопределение числа кластеров.

Кластеризация

Автоопределение числа кластеров: ☐

Заданное число кластеров

Число кластеров: 3

Автоматическое определение числа кластеров

Минимальное число кластеров: 1

Максимальное число кластеров: 10

Порог разделения кластеров: 1

Далее

Рисунок 18 – Окно «Кластеризация»

4. В окне «**Описания узла**» нажмите кнопку «**Сохранить**» (рисунок 19).

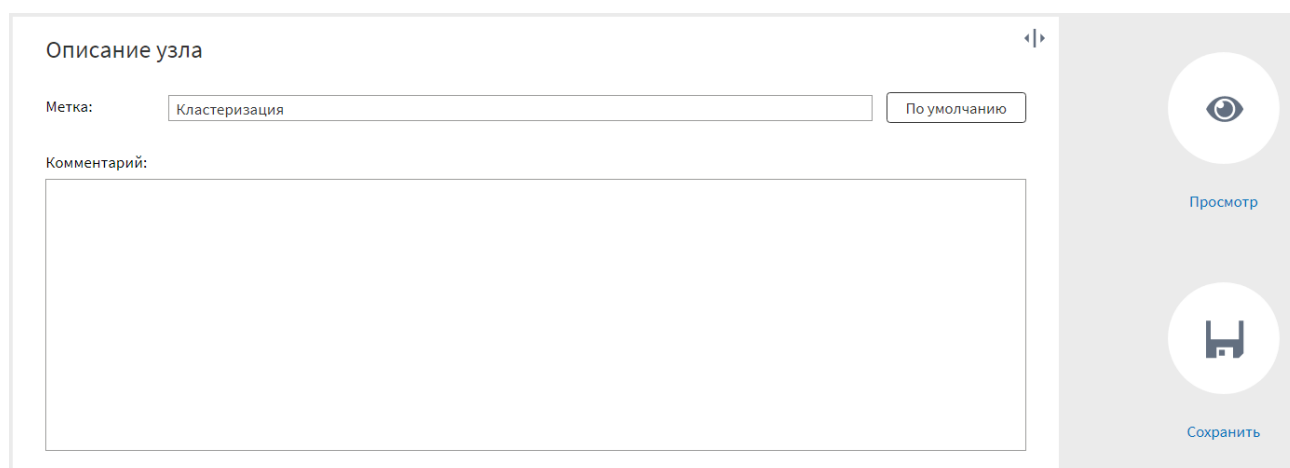


Рисунок 19 – Окно «Описание узла»

В рабочей области вызовите контекстное меню, кликнув правой кнопкой мышки на узле «Кластеризация». В контекстном меню выберем опцию «**Переобучить узел**» (рисунок 20).

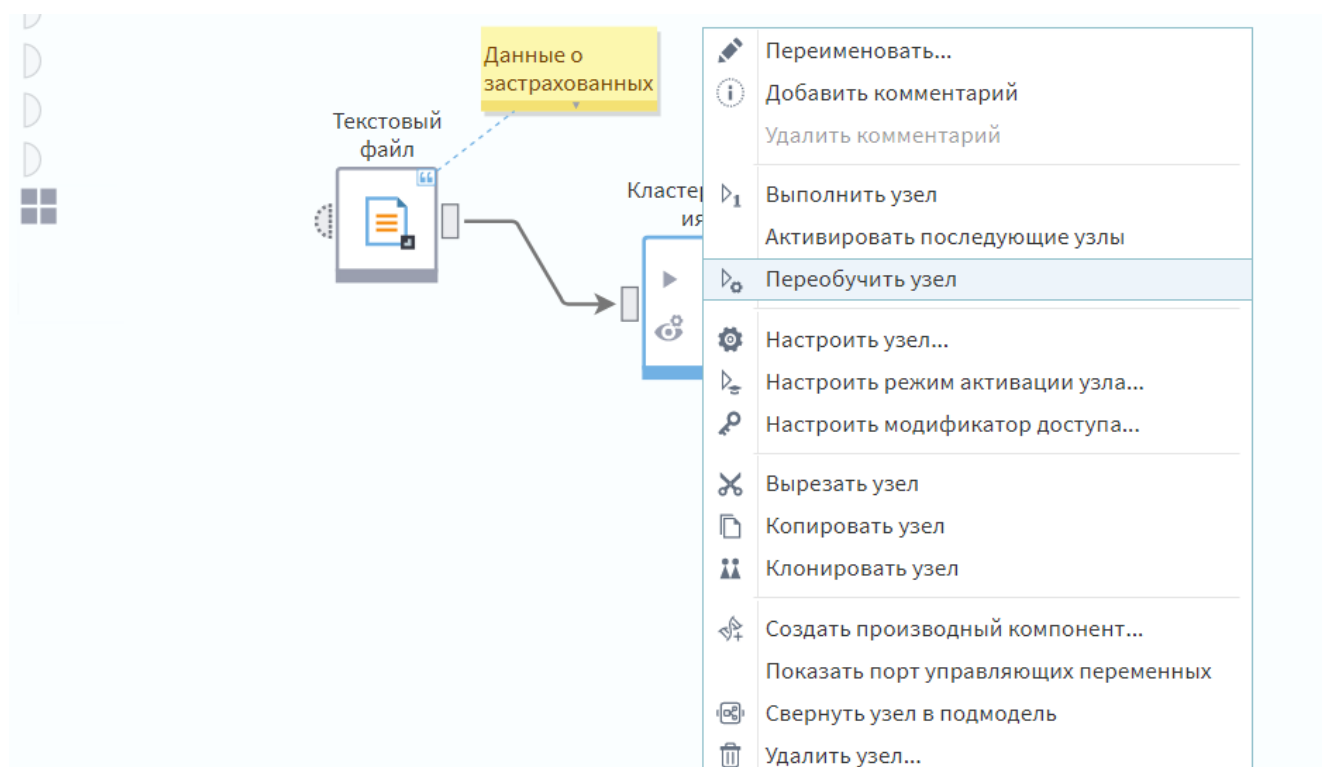


Рисунок 20 – Контекстное меню. Переобучить узел

После выполнения этой операции станут доступны выходные данные узла (рисунок 21, рисунок 22).

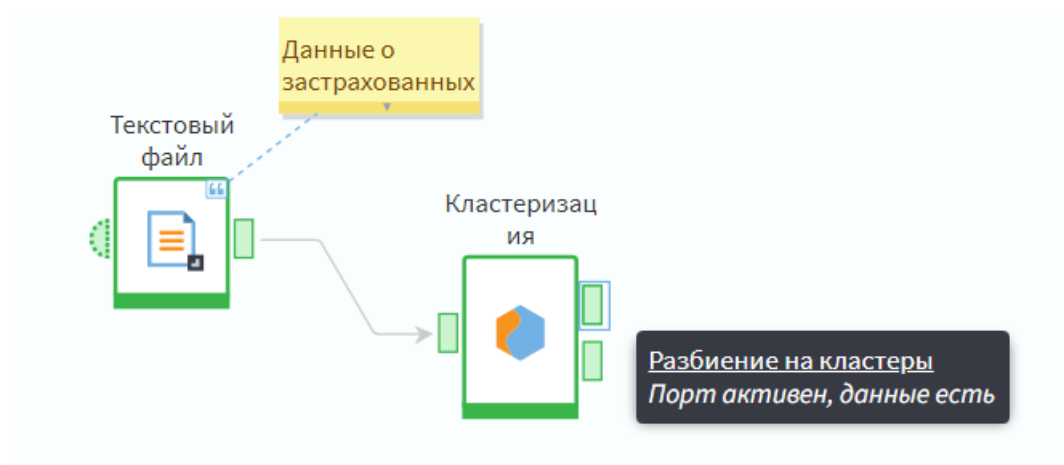


Рисунок 21 – Выходной порт с данными разбиения на кластеры

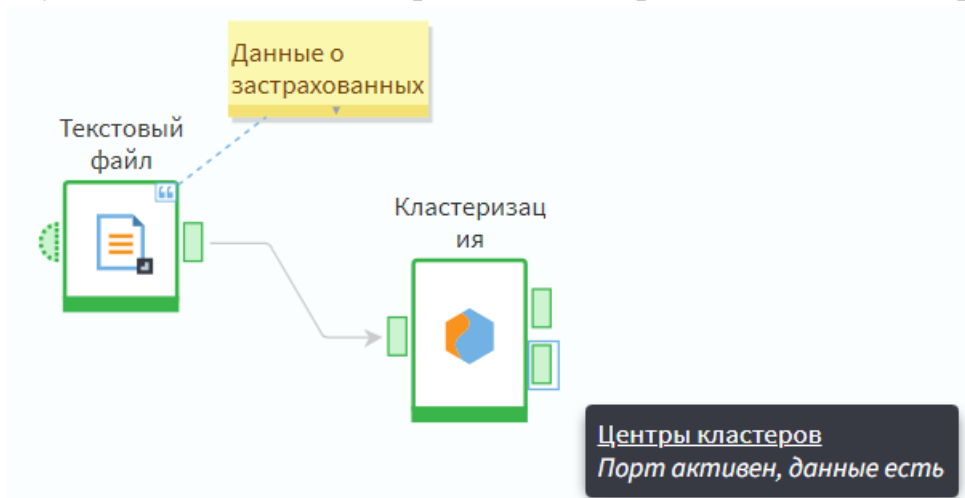


Рисунок 21 – Выходной порт с данными центров кластеров

Выходные данные при нажатии мышки на соответствующем выходном порте отображаются в отдельном окне быстрого просмотра. Но более информативными будут **визуализаторы**, которые можно настроить для узла «Кластеризация». Для этого необходимо на узле «Кластеризация» нажать на значок «**Настройка визуализаторов**» (голубой глазик) (рисунок 22). В результате откроется **окно «Визуализаторы»** (рисунок 23).

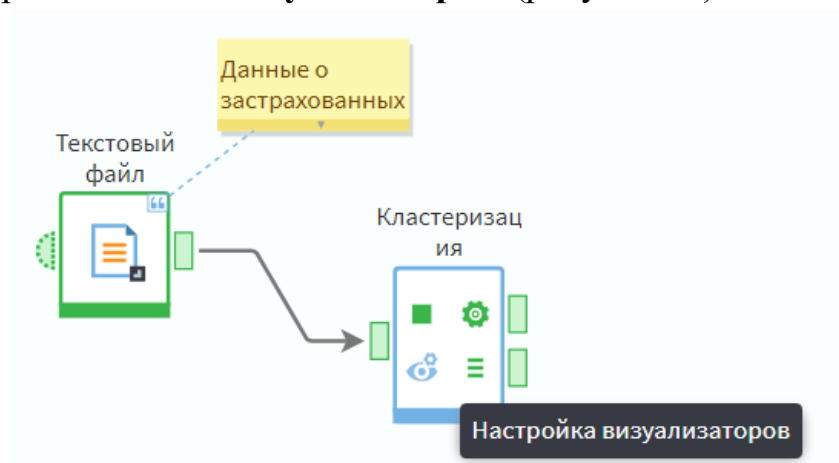


Рисунок 22 – Настройка визуализаторов на узле «Кластеризация»



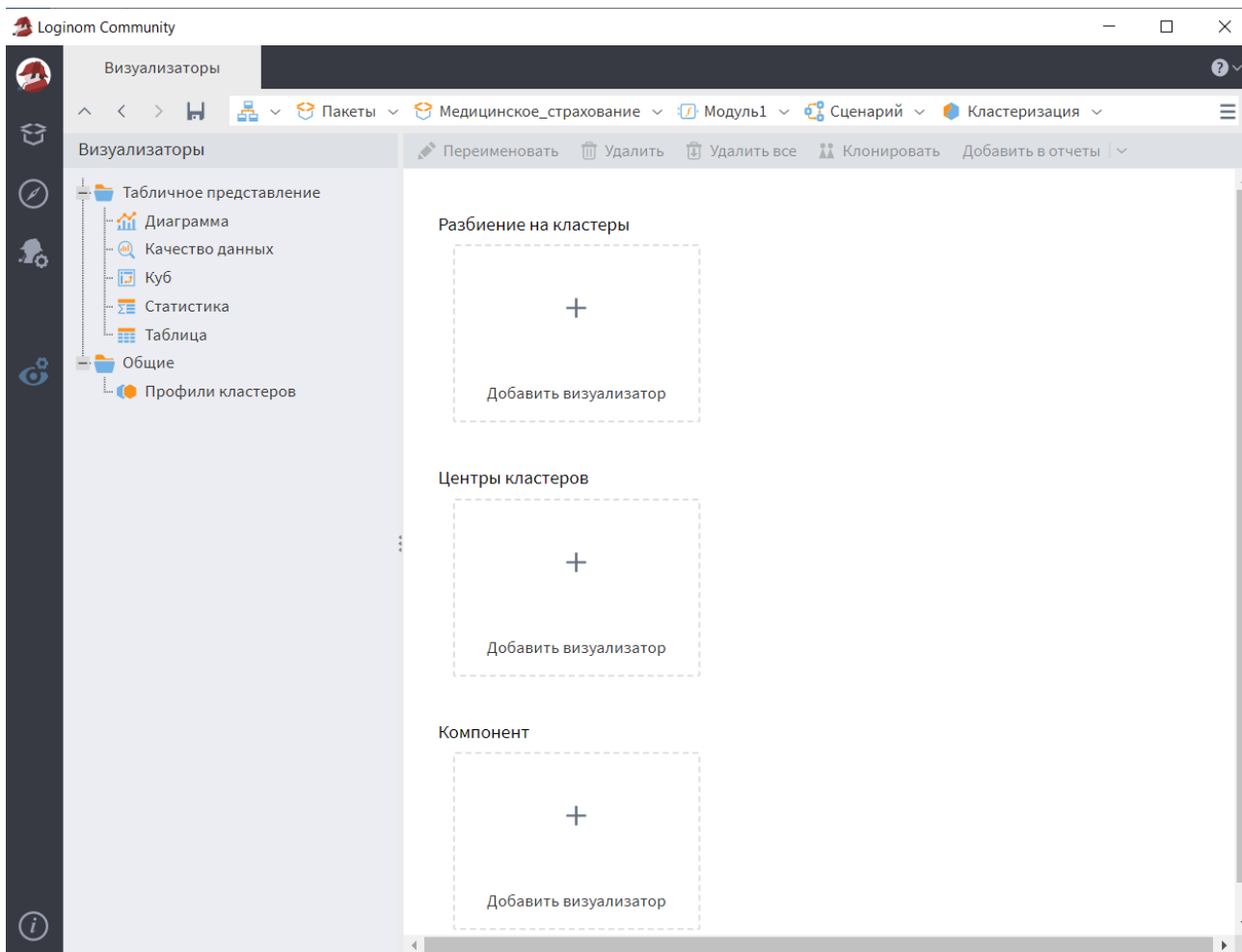


Рисунок 23 – Окно «Визуализаторы» для узла «Кластеризация»

Изначально это окно пусто. Визуализатор можно добавить простым перетаскиванием нужного элемента или выделив элемент в списке и нажав на «+» в рабочей области. После добавления нужного визуализатора для настройки необходимо в него войти (рисунок 24).

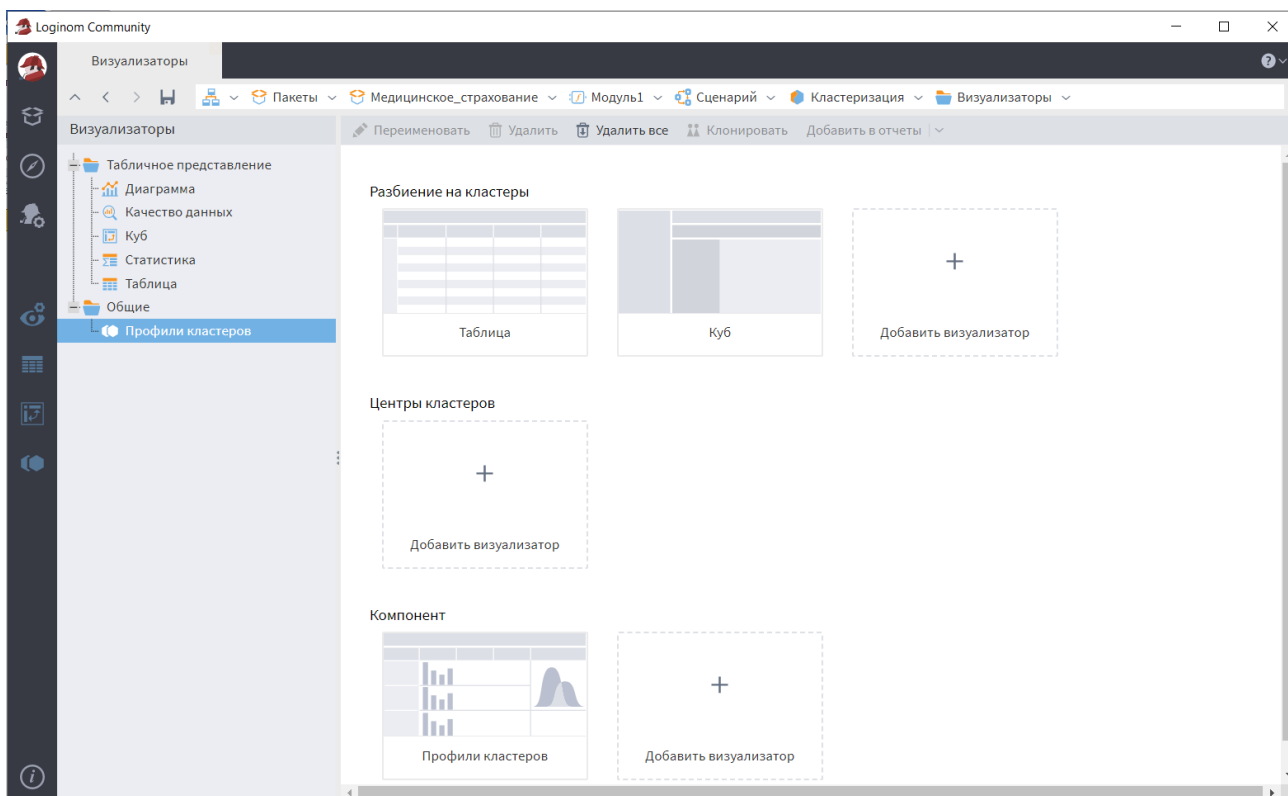



Рисунок 23 – Окно «Визуализаторы» для узла «Кластеризация»

Настройте первый визуализатор «Таблица» (рисунок 24).

#	12 Номер кластера	90 Расстояние до центра кластера	12 age	ab sex	90 bmi	12 children	ab smoker	ab region	90 charges	1 338
1	2	1,729576513	19	female	27,9	0	yes	southwest	16884,924	
2	2	1,49097328	18	male	33,77	1	no	southeast	1725,5523	
3	1	1,508393005	28	male	33	3	no	southeast	4449,462	
4	2	1,745169568	33	male	22,705	0	no	northwest	21984,47061	
5	2	1,307719776	32	male	28,88	0	no	northwest	3866,8552	
6	2	1,402932186	31	female	25,74	0	no	southeast	3756,6216	
7	0	1,339203162	46	female	33,44	1	no	southeast	8240,5896	
8	1	1,338933494	37	female	27,74	3	no	northwest	7281,5056	
9	1	1,328062992	37	male	29,83	2	no	northeast	6406,4107	
10	0	1,613407555	60	female	25,84	0	no	northwest	28923,13692	
11	2	1,33303488	25	male	26,22	0	no	northeast	2721,3208	
12	0	1,960866102	62	female	26,29	0	yes	southeast	27808,7251	
13	2	1,476920757	23	female	34,4	0	no	southwest	1826,843	
14	0	1,81241008	56	female	39,82	0	no	southeast	11090,7178	
15	2	2,599312253	27	male	42,13	0	yes	southeast	39611,7577	
16	2	1,579818902	19	male	24,6	1	no	southwest	1837,237	
17	0	1,230397713	52	female	30,78	1	no	northeast	10797,3362	
18	2	1,546920454	23	male	23,845	0	no	northeast	2395,17155	
19	0	1,893350238	56	male	40,3	0	no	southwest	10602,385	
1 338	2	1,906923827	30	male	35,3	0	yes	southwest	36837,467	

Рисунок 24 – Визуализатор «Таблица»

Проведите сортировку с использованием нескольких полей  Сортировка, установив иерархию: номер кластера, затем упорядочиваем внутри каждого кластера по убыванию возраста, а затем среди застрахованных одного года рождения по убыванию расходов на медицинское обслуживание (рисунок 25).

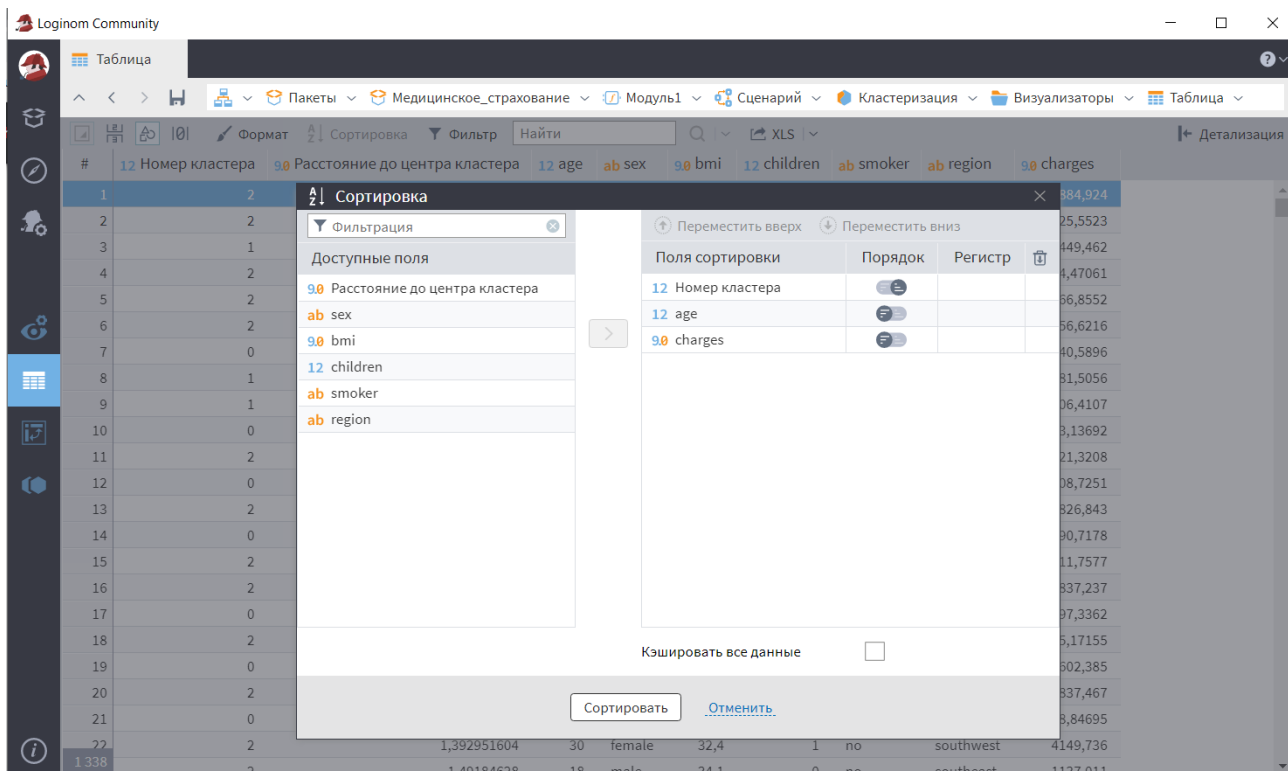


Рисунок 25 – Настройка сортировки по нескольким полям

Если необходимо убрать с экрана какой-либо столбец таблицы, то вызовите контекстное меню по стрелке рядом с именем столбца и уберите галочку в списке столбцов для отображения (рисунок 26).

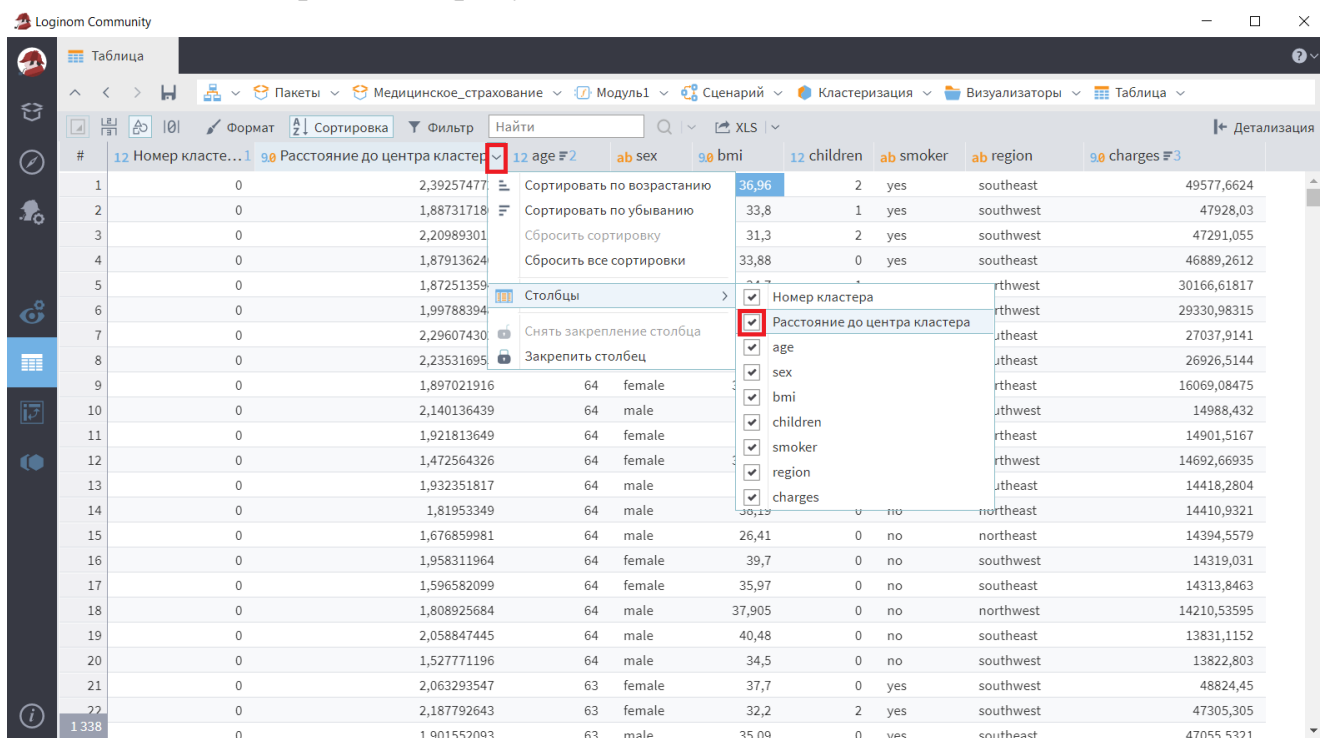


Рисунок 26 – Выбор столбцов для отображения

Итоговой результат настройки визуализатора «Таблица» приведен на рисунке 27.

#	12 Номер класте...	12 age	ab sex	9.0 bmi	12 children	ab smoker	ab region	9.0 charges	Детализация
1	0	46	female	33,44	1	no	southeast	8240,5896	
2	0	60	female	25,84	0	no	northwest	28923,13692	
3	0	62	female	26,29	0	yes	southeast	27808,7251	
4	0	56	female	39,82	0	no	southeast	11090,7178	
5	0	52	female	30,78	1	no	northeast	10797,3362	
6	0	56	male	40,3	0	no	southwest	10602,385	
7	0	60	female	36,005	0	no	northeast	13228,84695	
8	0	63	female	23,085	0	no	northeast	14451,83515	
9	0	63	male	28,31	0	no	northwest	13770,0979	
10	0	60	male	39,9	0	yes	southwest	48173,361	
11	0	38	male	37,05	1	no	northeast	6079,6715	
12	0	55	male	37,3	0	no	southwest	20630,28351	
13	0	60	female	24,53	0	no	southeast	12629,8967	
14	0	48	male	28	1	yes	southwest	23568,272	
15	0	58	female	31,825	2	no	northeast	13607,36875	
16	0	53	female	22,88	1	yes	southeast	23244,7902	
17	0	64	male	24,7	1	no	northwest	30166,61817	
18	0	61	female	39,1	2	no	southwest	14235,072	
19	0	40	female	36,19	0	no	southeast	5920,1041	
20	0	58	male	32,01	1	no	southeast	11946,6259	
21	0	57	male	34,01	0	no	northwest	11356,6609	
22	0	41	female	32,965	0	no	northwest	6571,02435	
1 338	0	45	female	38,285	0	no	northeast	7935,29115	

Рисунок 27 – Визуализатор «Таблица» для узла «Кластеризация»

Сохраните визуализатор «Таблица». Для этого необходимо нажать кнопку



. Чтобы вернуться в окно «Визуализаторы», необходимо на боковой панели



нажать кнопку

Далее перейдите в настройки визуализатора «Куб» (рисунок 28).

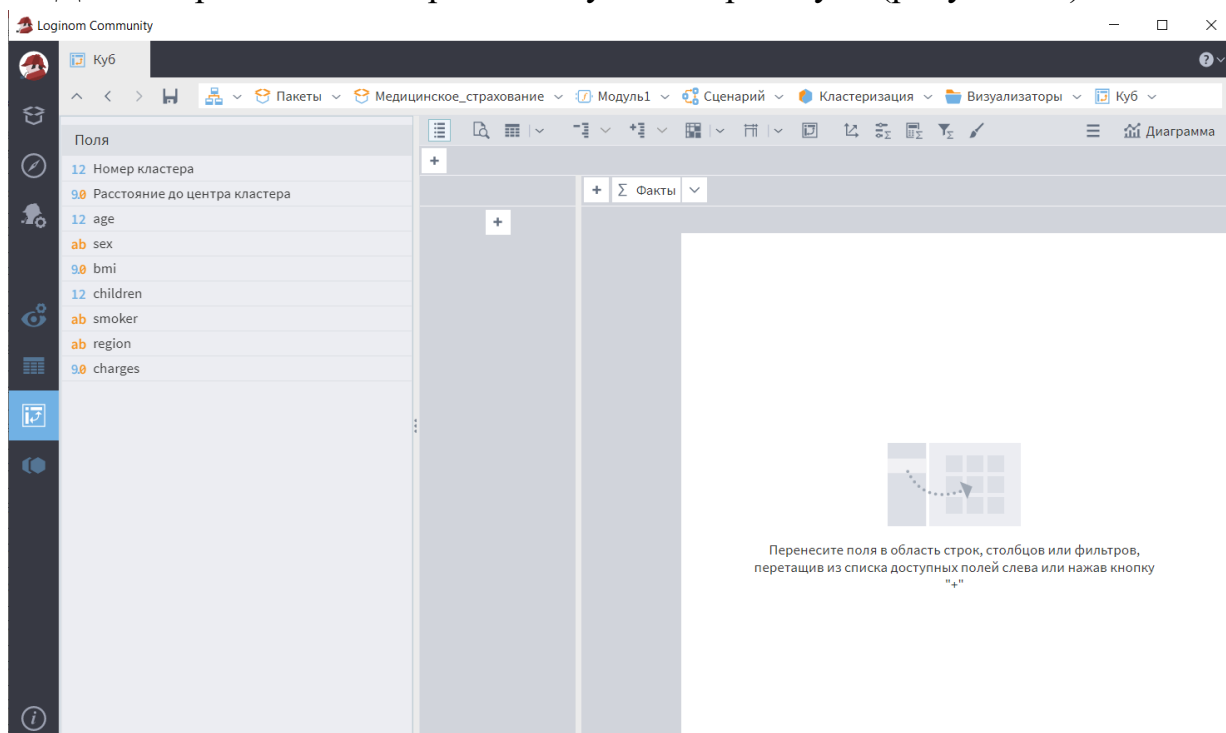


Рисунок 28 – Визуализатор «Куб»

Добавьте измерения и факты для Куба.

Измерения для строк добавляются слева от таблицы (рисунок 29).

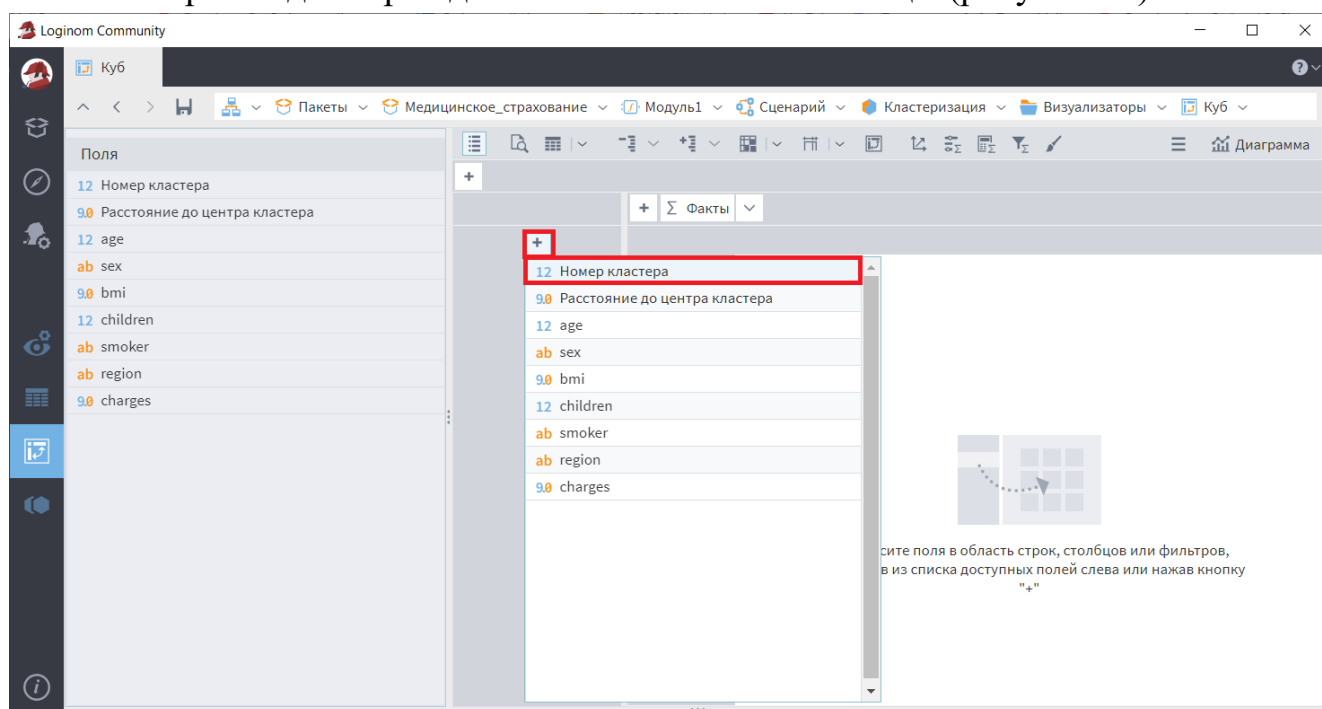


Рисунок 29 – Добавление измерений

Добавьте «Номер кластера», «smoker» (рисунок 30).

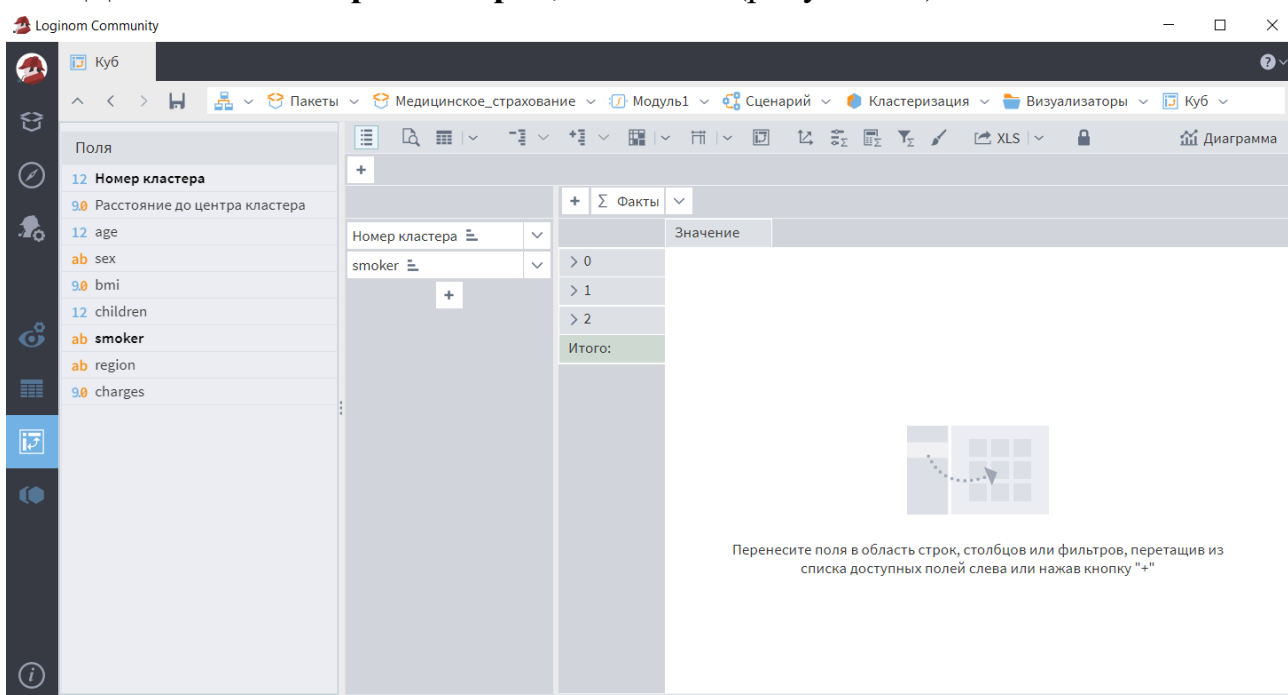


Рисунок 30 – Добавление измерений

С помощью кнопки «**Факты**» добавьте два количественных показателя: «**charges**» и «**bmi**» – и определите для них два способ агрегации – «Сумма» и «Среднее» (рисунок 31).

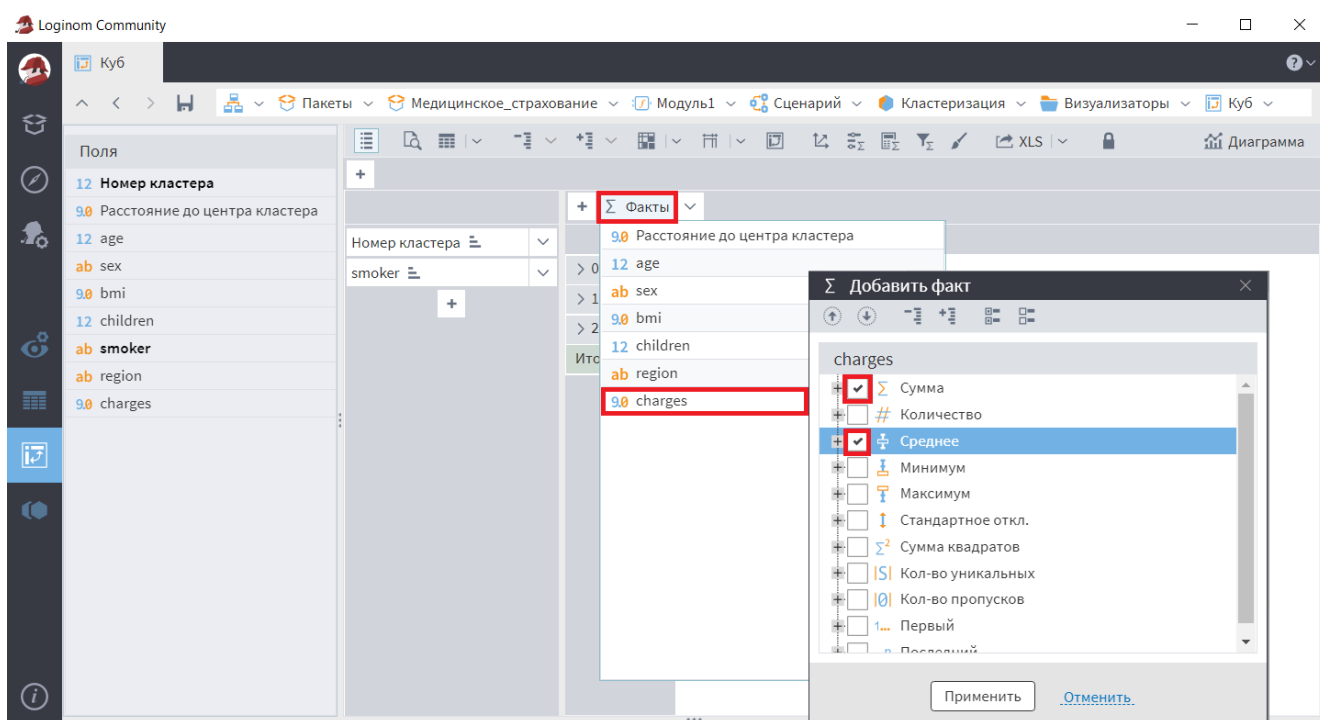


Рисунок 31 – Добавление фактов

Итоговой результат настройки визуализатора «Куб» приведен на рисунке 32.

		charges		bmi	
		Σ Сумма	Σ Средн...	Σ Сумма	Σ Средн...
0	no	4 348 981,69	11 597,28	11 838,54	31,57
	yes	2 922 346,03	36 991,72	2 468,18	31,24
	Итого:	7 271 327,72	16 016,14	14 306,72	31,51
1	no	2 878 873,55	9 469,98	9 419,63	30,99
	yes	2 943 137,30	33 068,96	2 766,20	31,08
	Итого:	5 822 010,85	14 814,28	12 185,82	31,01
> 2		4 662 486,42	9 495,90	14 535,08	29,60
Итого:		17 755 824,99	13 270,42	41 027,62	30,66

Рисунок 32 – Настройки визуализатора Куб

Сохраните визуализатор «Куб», перейдите в настройки визуализатора «Профили кластеров» (рисунок 33).

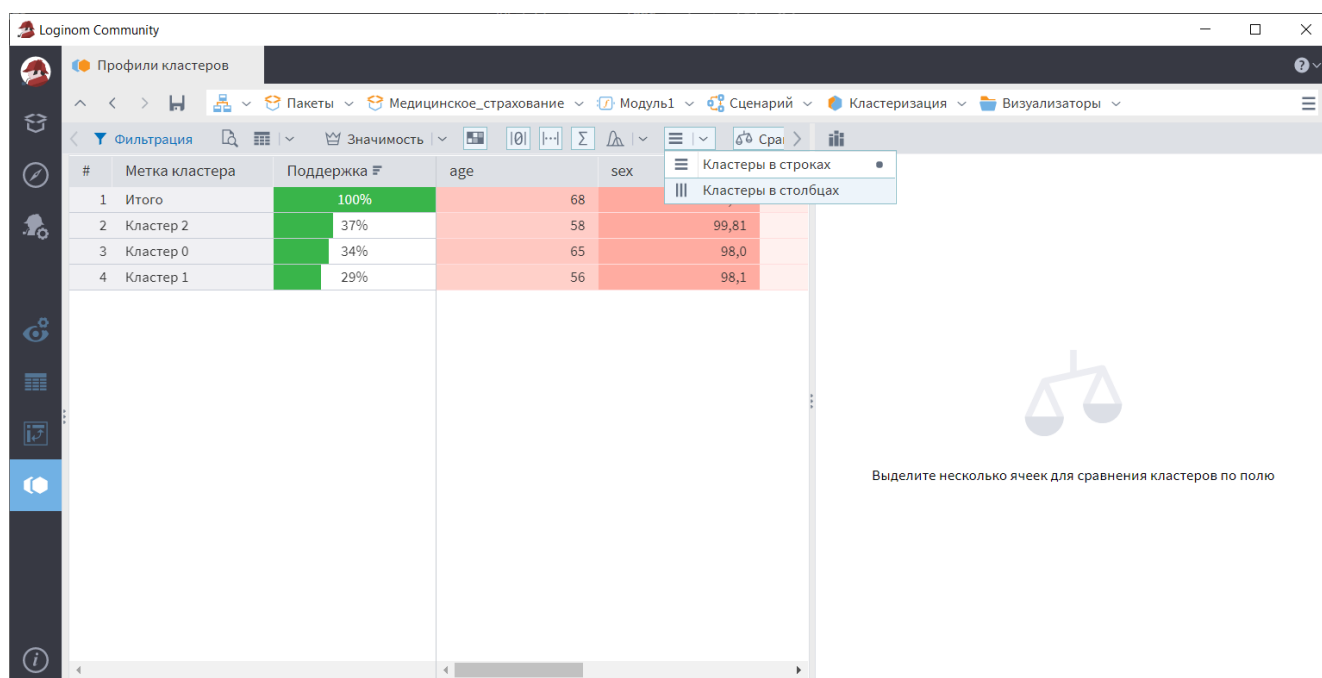


Рисунок 33 – Визуализатор «Профили кластеров»

Протранспонируйте таблицу, расположив кластеры в столбцах (рисунок 33). Если необходимо упорядочить столбцы по номеру кластера, то это можно сделать простым перетаскиванием. Далее выделите три ячейки в строке «age» (возраст) для сравнения (выделение нескольких ячеек одновременно возможно при нажатой клавише Ctrl). Справа от таблицы будет проведено сравнение кластеров по показателю «age» (рисунок 34).

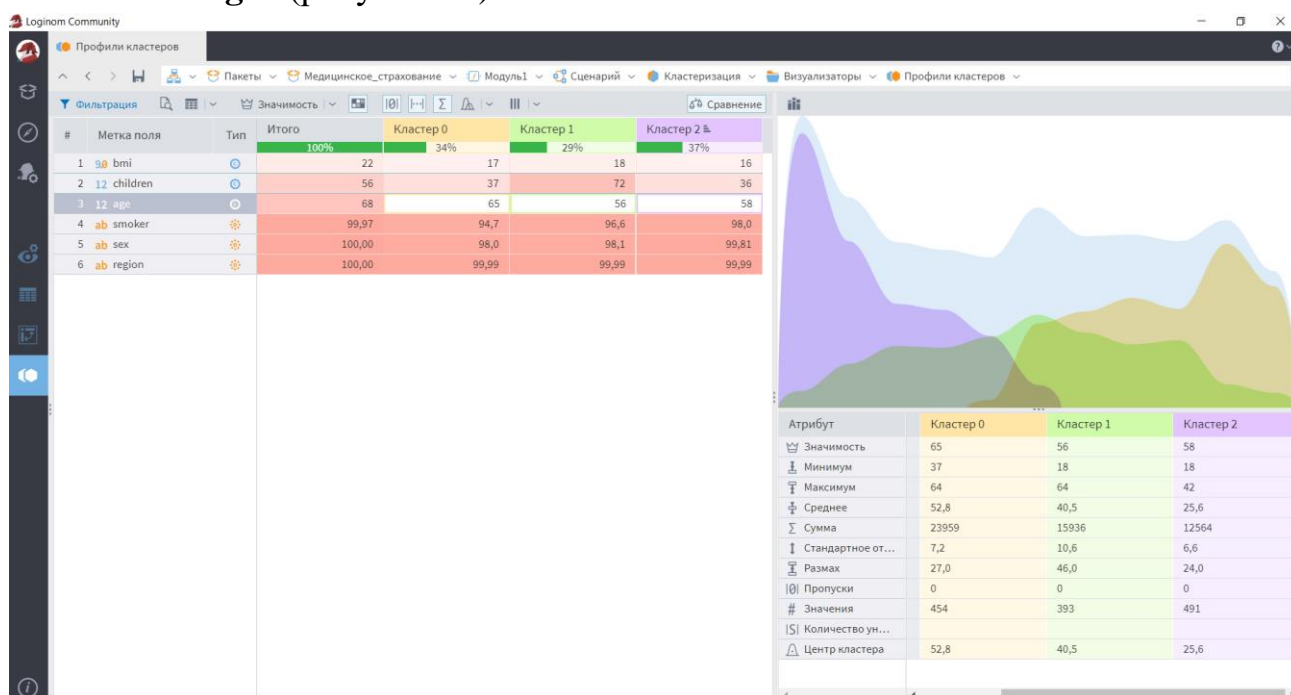



Рисунок 34 – Сравнение кластеров по показателю «age»

**Панель фильтрации** содержит настройки фильтрации основной таблицы.


Эта панель (рисунок 34) вызывается при нажатии кнопки  .

Настройки диапазона фильтруемых значений:

**1. Мощность кластера** – показывает количество строк исходного набора, попавших в кластер. Задаёт диапазон от 0 до числа строк в исходном наборе данных.

**2. Значимость поля** – мера влияния поля на попадание поля в некоторый кластер. Задаёт диапазон от 0 до 100 %.

**3. Значимость ячейки** – мера влияния ячейки на попадание ячейки в некоторый кластер. Задаёт диапазон от 0 до 100 %.

**Детализация** () использует данные исходного набора данных, отфильтрованные по выделенным в основной таблице кластерам. Если выделен кластер «Итого», тогда набор данных детализации совпадает с входным набором. Детализация позволяет узнать, из каких строк исходного набора данных состоит тот или иной кластер.

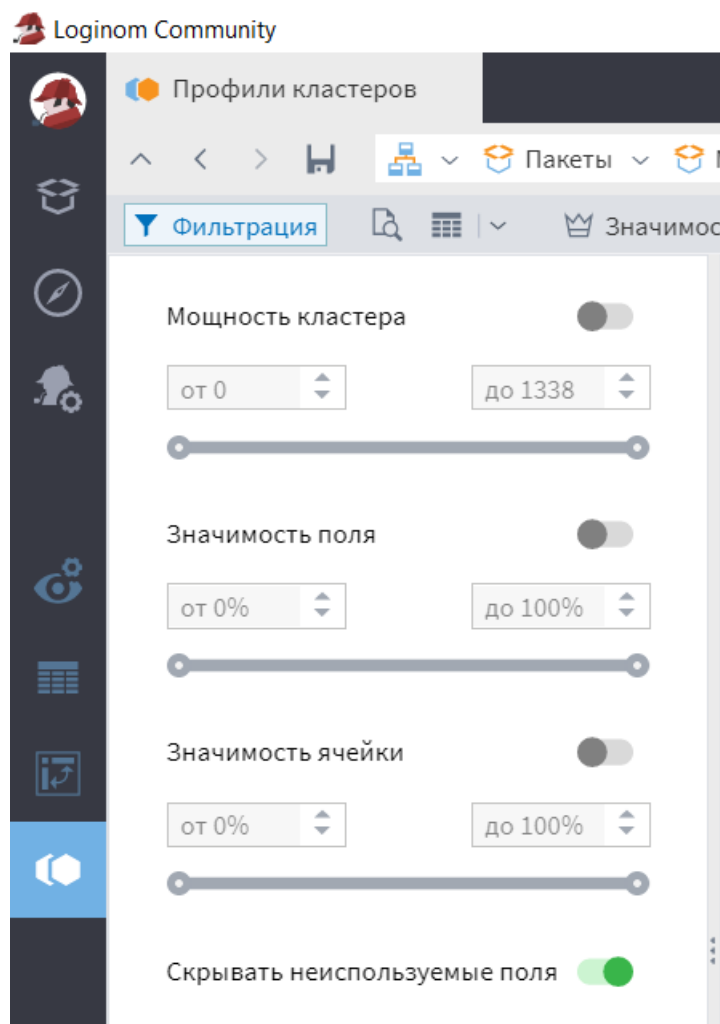


Рисунок 34– Панель фильтрации