Министерство науки и высшего образования Российской Федерации

Федеральное государственное бюджетное образовательное учреждение

высшего образования «Кубанский государственный университет»

Кафедра информационных технологий

**ОТЧЕТ**

о выполнении лабораторной работы №1

по дисциплине «Машинное обучение»

Выполнили: ст. гр. 4ИТ

Хижний Е.Г.

Проверил: ст. препод.

Харченко А.В.

Краснодар

2022

Лабораторная работа №1.

**КЛАСТЕРИЗАЦИЯ. ПОСТРОЕНИЕ МОДЕЛИ КЛАСТЕРИЗАЦИИ В ORANGE**

**Цель и задача работы:**изучить основные методы кластеризации с использованием приложения «Orange Data Mining».

**Ход работы**

В качества набора данных выбран стандартный файл zoo.bsk, содержащий информацию о 101 животном и их характеристиках.

Для добавления файла нажмем на виджет File, показанный на рис.1.

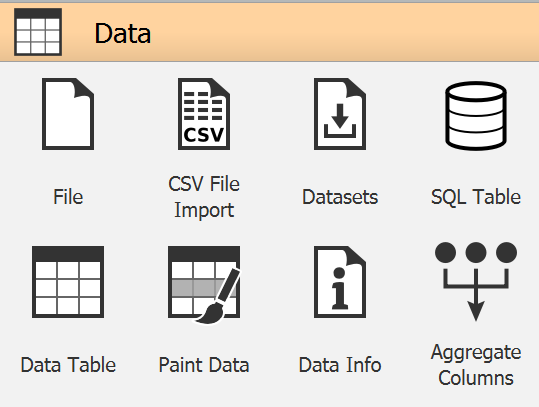


Рисунок 1 – Поле виджетов

На рис. 2 показано окно данного виджета, выберем необходимый файл и нажмем “Apply”.

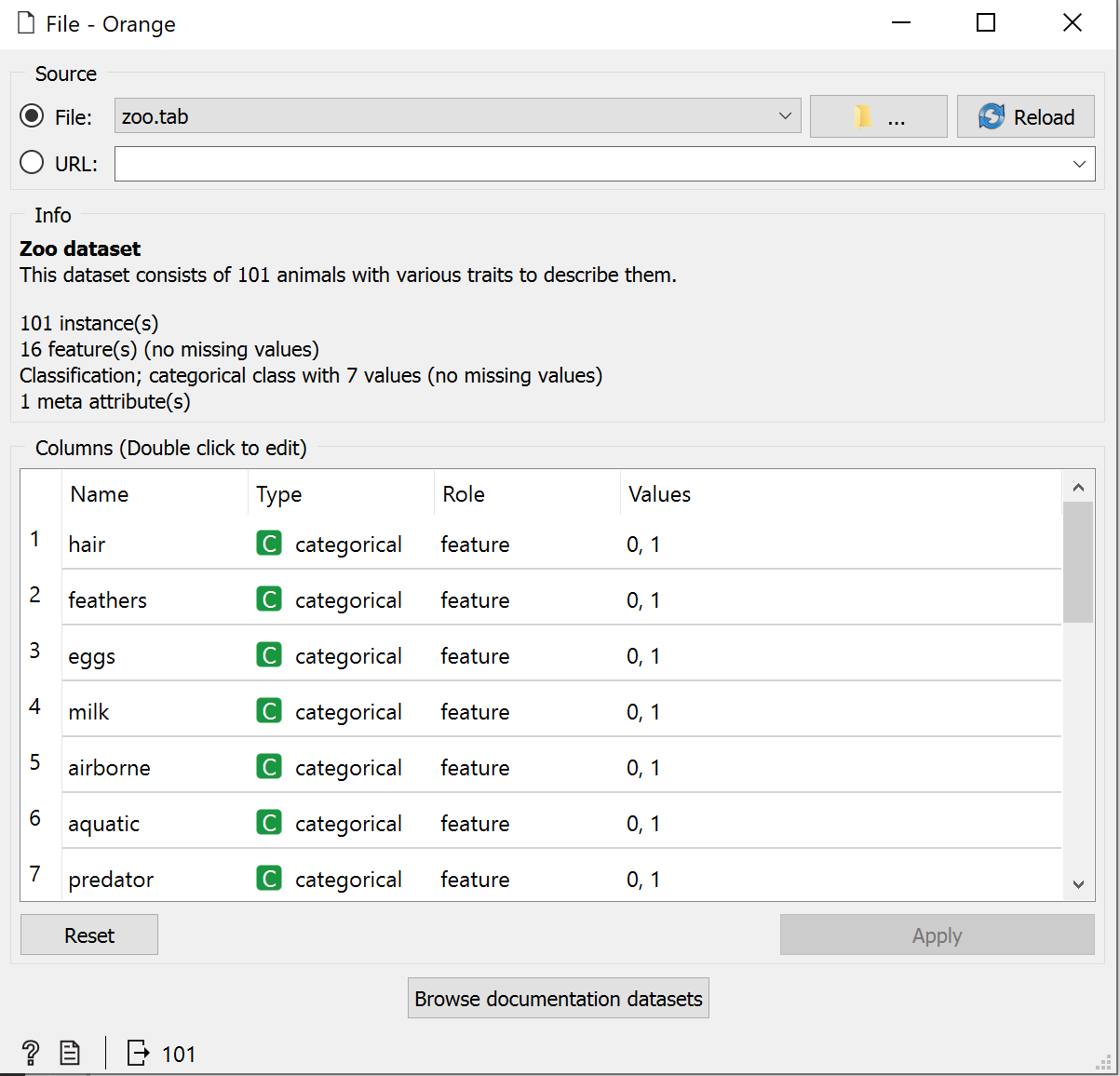


Рисунок 2 – Настройка виджета File

Осуществим необходимую фильтрацию/очистку данных с помощью виджетов Select Rows и Select Columns в соответствии с заданием. От виджета файл передадим данные Select Columns, а затем Select Rows, как показано на рис. 3

Изображение выглядит как текст, часы

Автоматически созданное описание

Рисунок 3 – Фильтрация данных

На рис. 4 и 5 показаны фильтрации переменных и наблюдений, которые игнорируем.

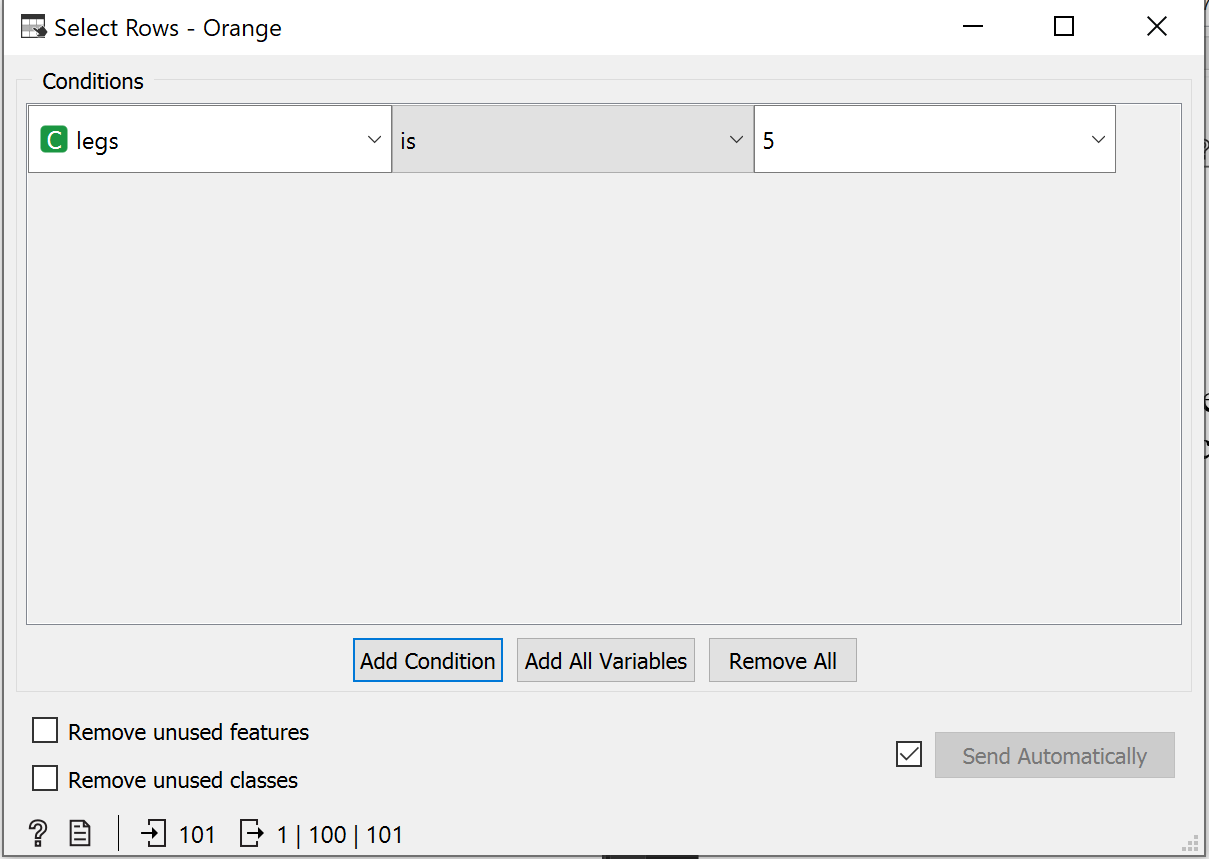


Рисунок 4 – Фильтрация данных по строкам

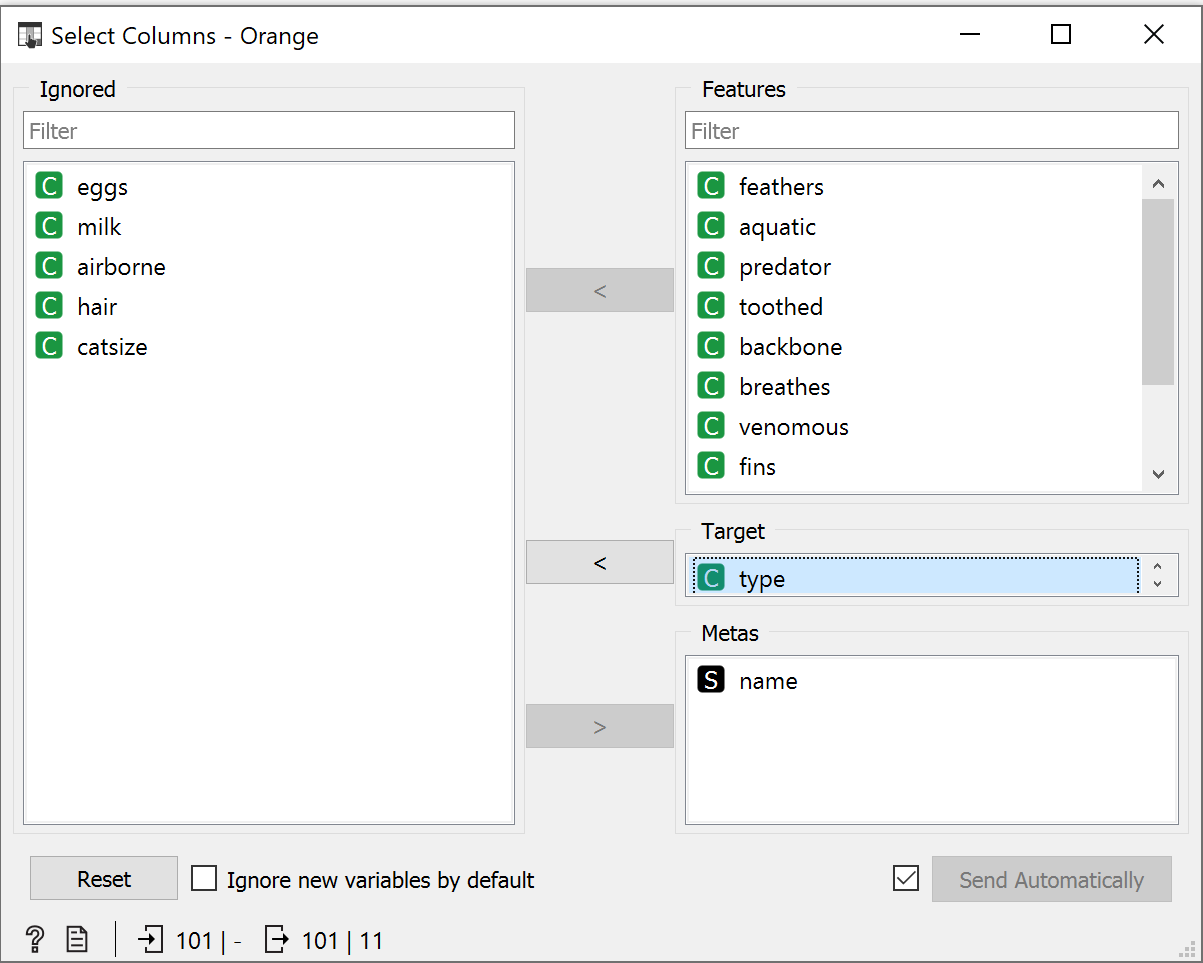


Рисунок 5 – Фильтрация данных по столбцам

Осуществим кластеризацию при помощи 2-х виджетов k-means: с фиксированным и подобранным количеством кластеров. Для этого из Select Columns передадим данные в разные k-means, как показанное на рис.6.

Изображение выглядит как текст, электроника

Автоматически созданное описание

Рисунок 6 – Передача данные в разные k-means

Разобьём данные на 3 (фиксированное) и от 2 до 8 (подобранное) кластеров, алгоритмом начального выбора центров кластеров k-Means++(заключается в том, что первые центры кластеров выбираются случайно, последующие центры выбираются из оставшихся объектов с вероятностью, пропорциональной квадрату расстояния до ближайшего центра).

Изображение выглядит как текст

Автоматически созданное описание

Рисунок 7 – Настройка фиксированной кластеризации

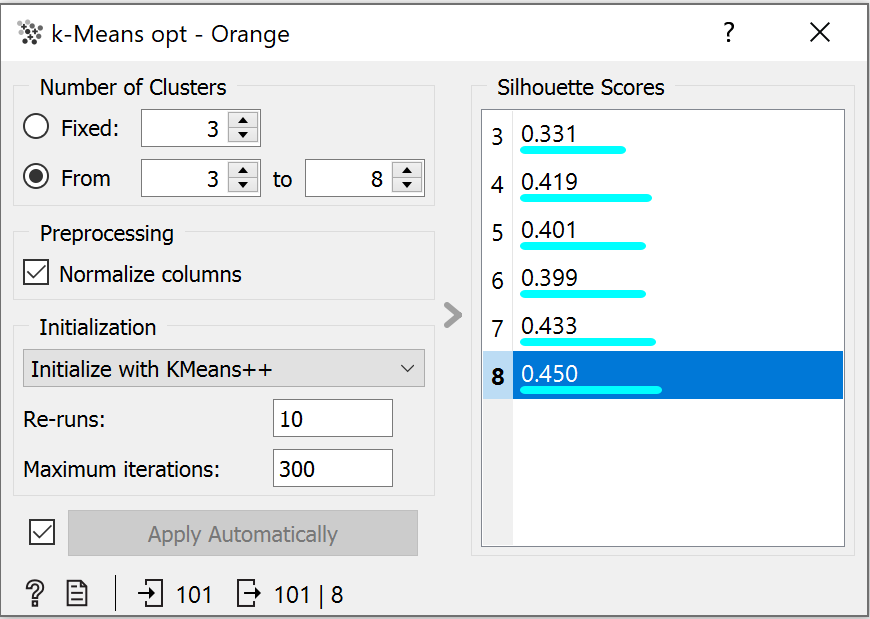


Рисунок 8 – Настройка подобранной кластеризации

Построим диаграмму рассеивания при помощи виджета Scatter Plot для обоих видов кластеризации (рис.9, 10).

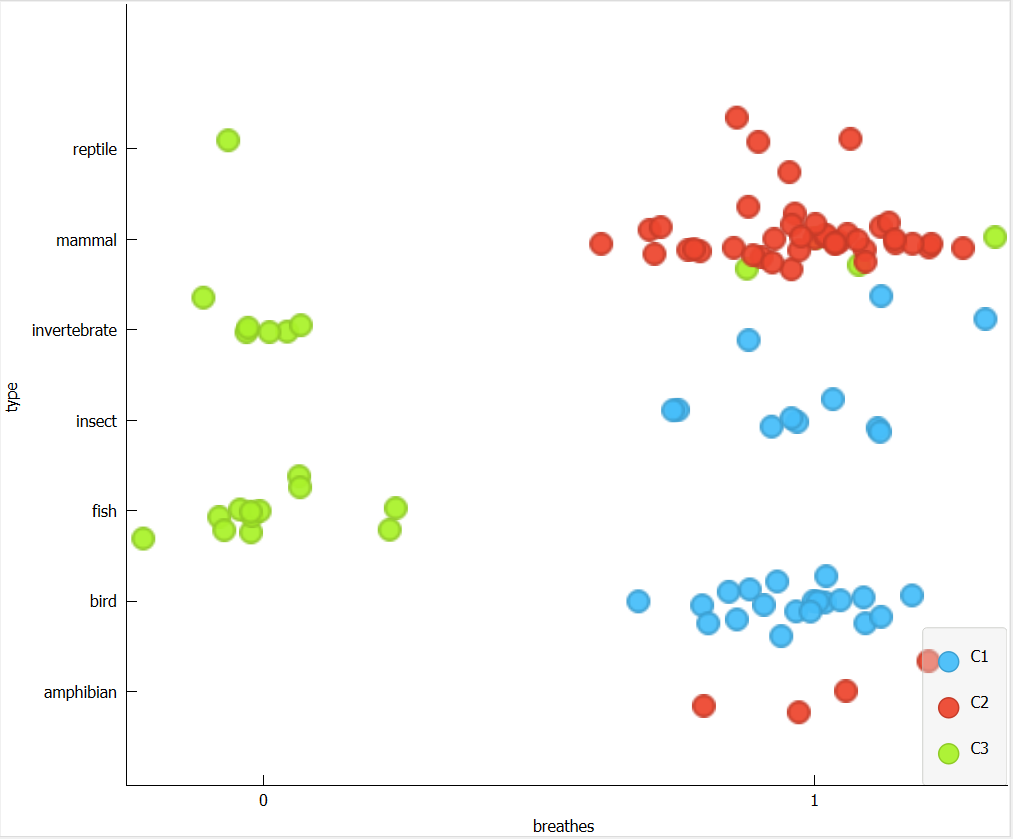


Рисунок 9 – Фиксированная кластеризация

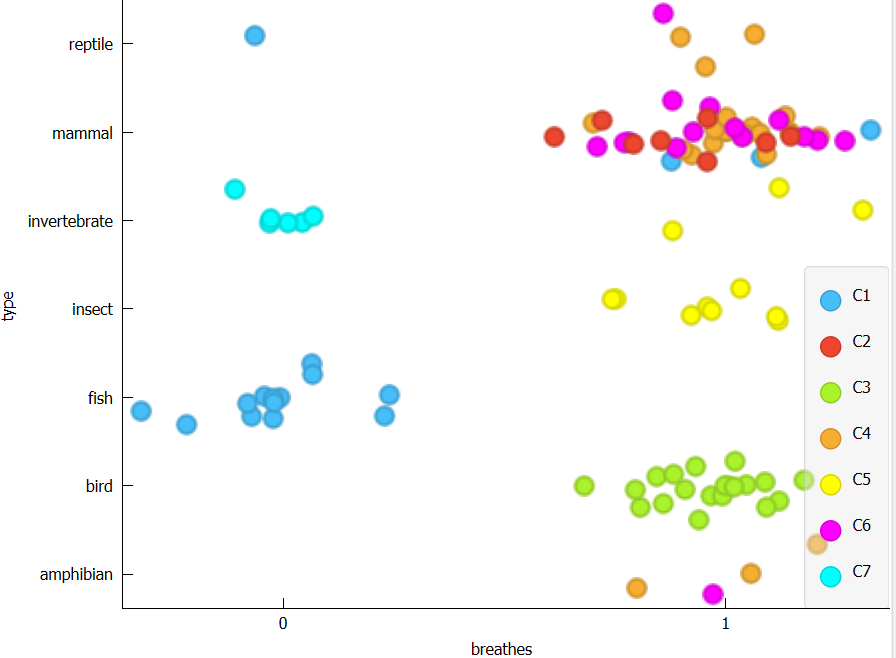


Рисунок 10 – Оптимизированная кластеризация

Можно сделать вывод, что кластеризация на рис.9 разбила на более четкие кластеры по видам, чем на рис.10.

Построим иерархическую модель данных на основе меры расстояния евклидового, манхэттенского и Хэмминга. Произведем иерархическую кластеризацию и для каждой кластеризации построим диаграмму рассеивания (рис. 12–14).

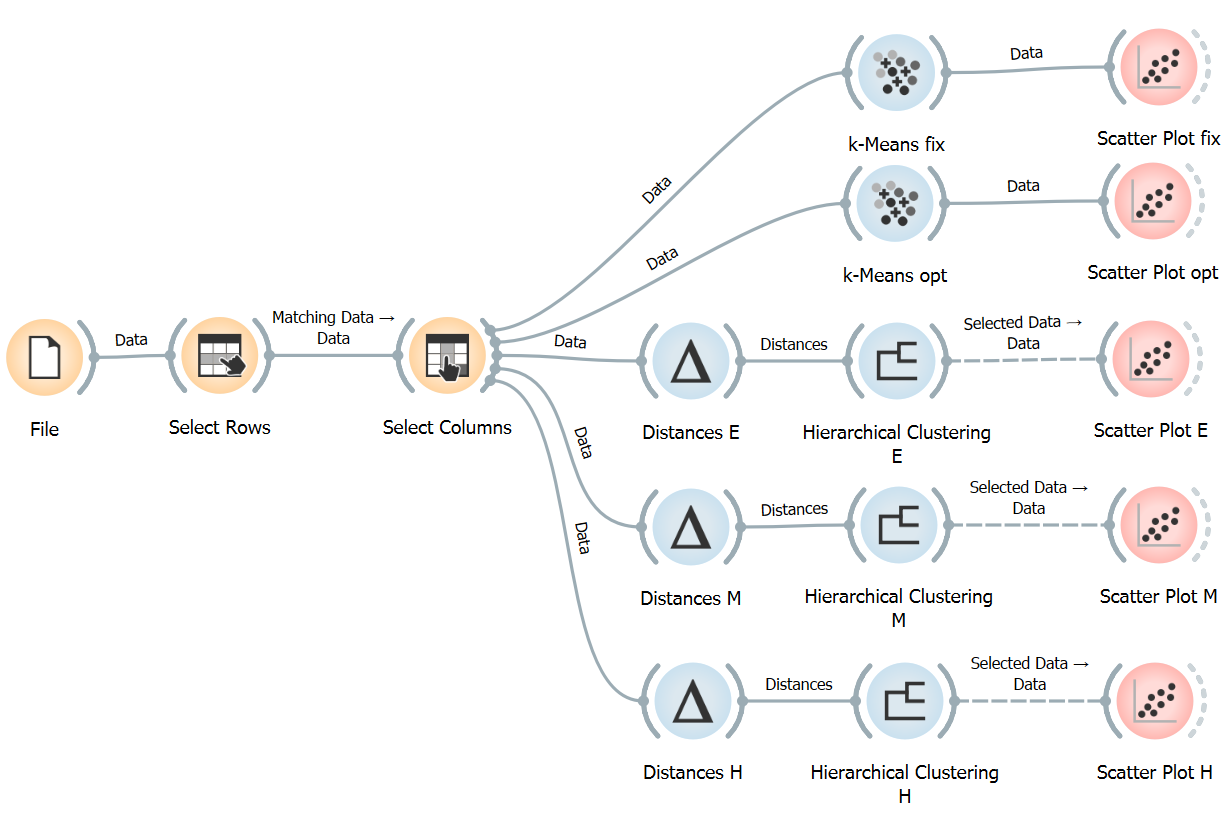


Рисунок 11 – Иерархическая кластеризация

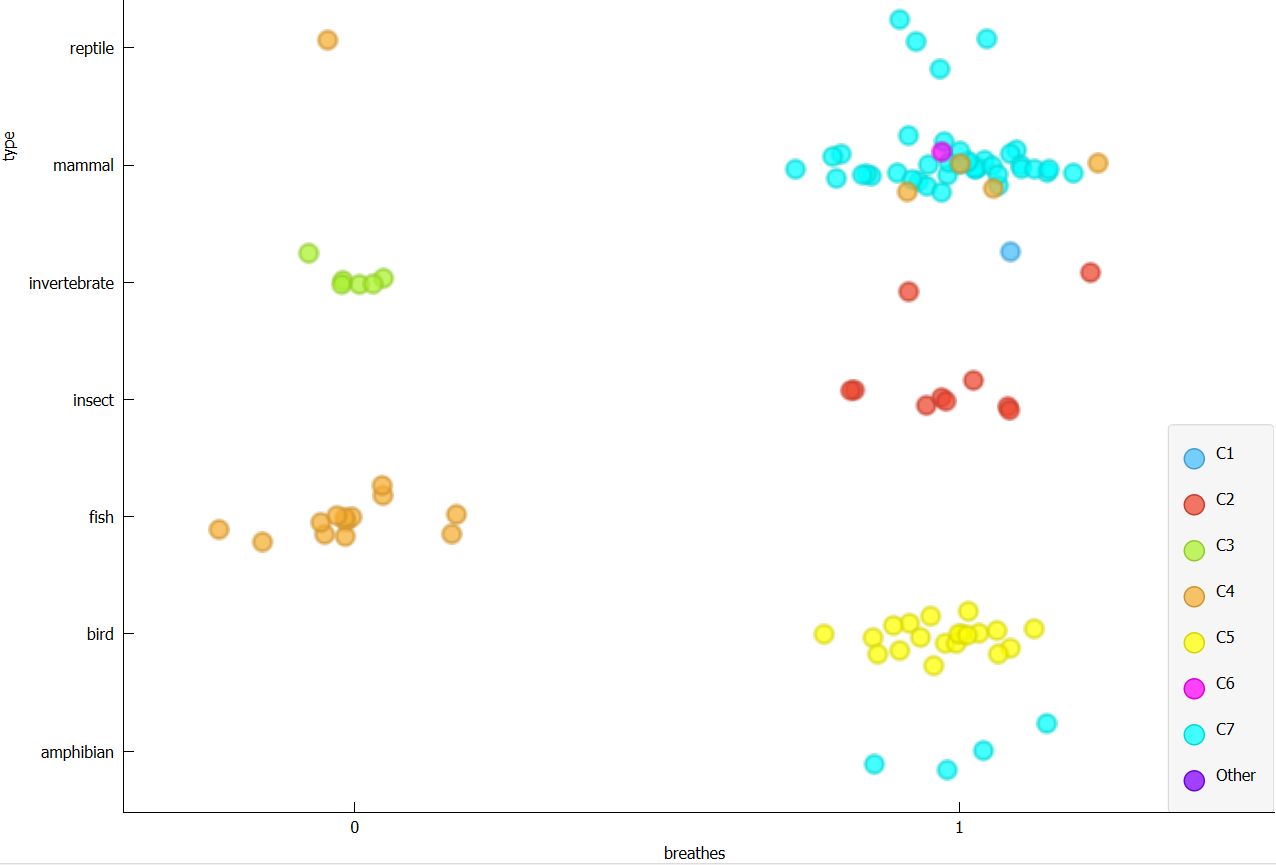


Рисунок 12 – Иерархическая кластеризация для евклидового расстояния

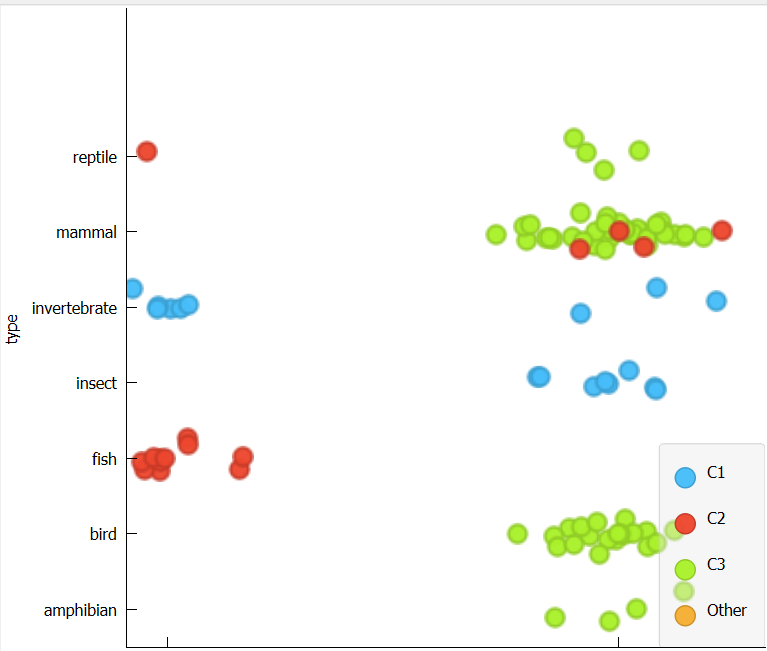


Рисунок 13 – Иерархическая кластеризация для манхэттенского расстояния

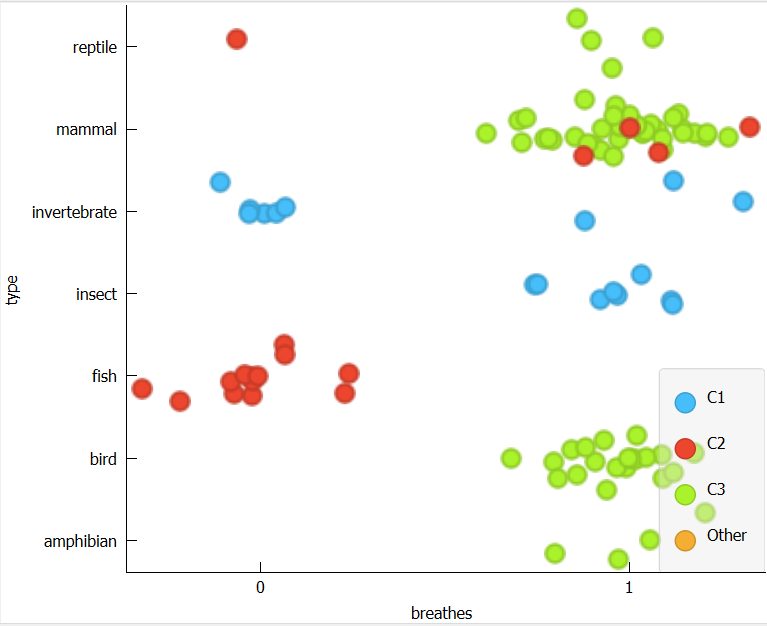


Рисунок 14 – Иерархическая кластеризация для расстояния Хэмминга

Отключим фильтрацию данных, те будем передавать данные напрямую в методы кластеризации и повторить эксперимент.

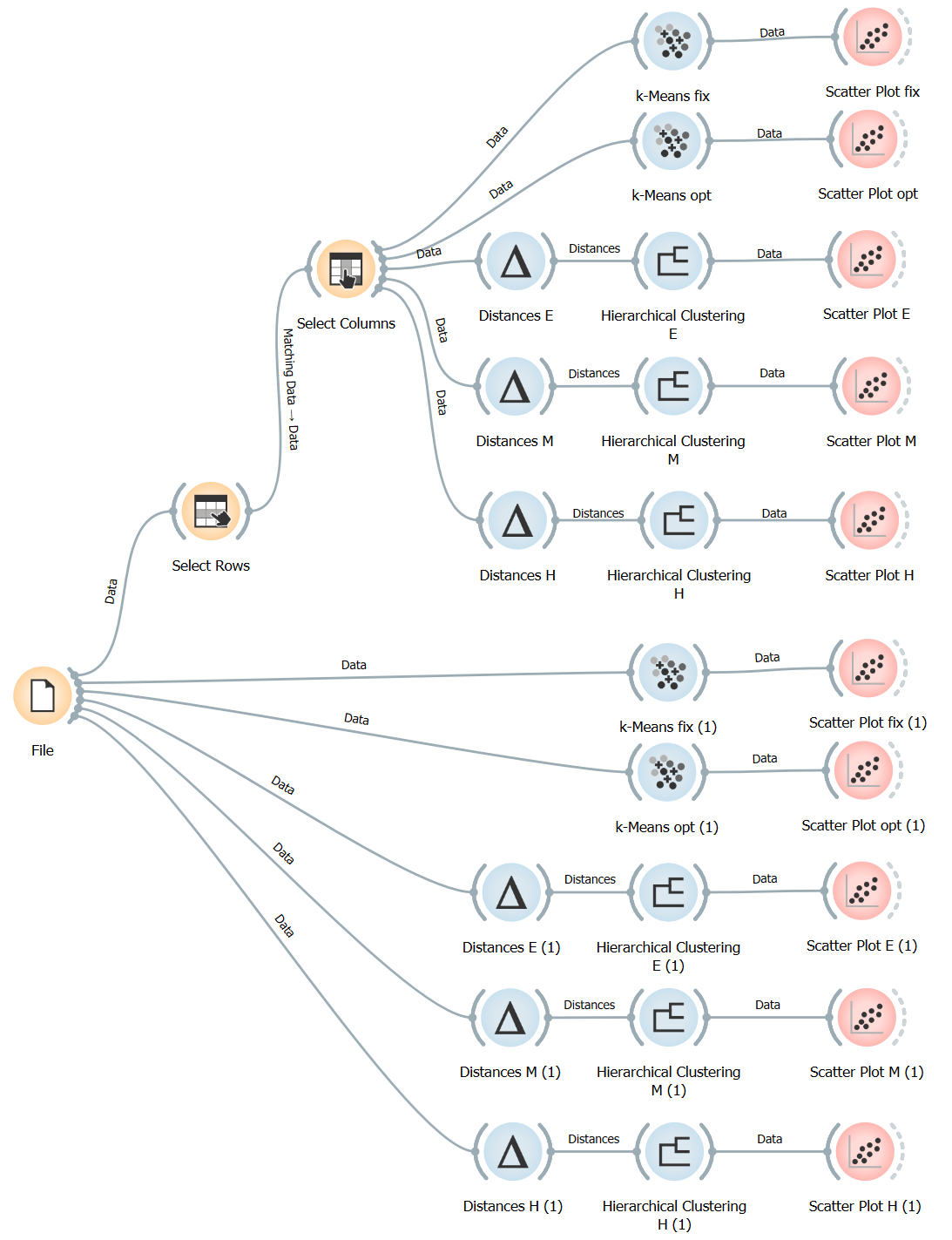


Рисунок 15 – Кластеризация без фильтрации

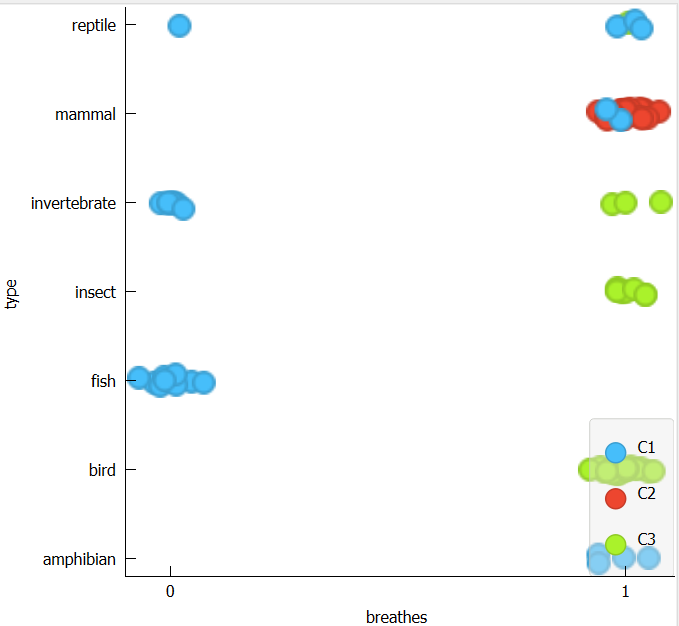


Рисунок 16 – Фиксированная кластеризация

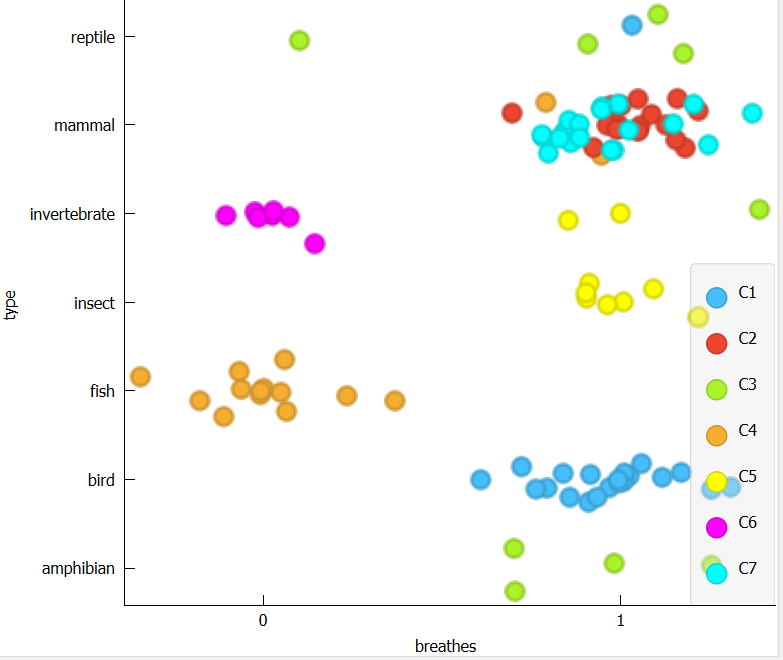


Рисунок 17 – Оптимизированная кластеризация

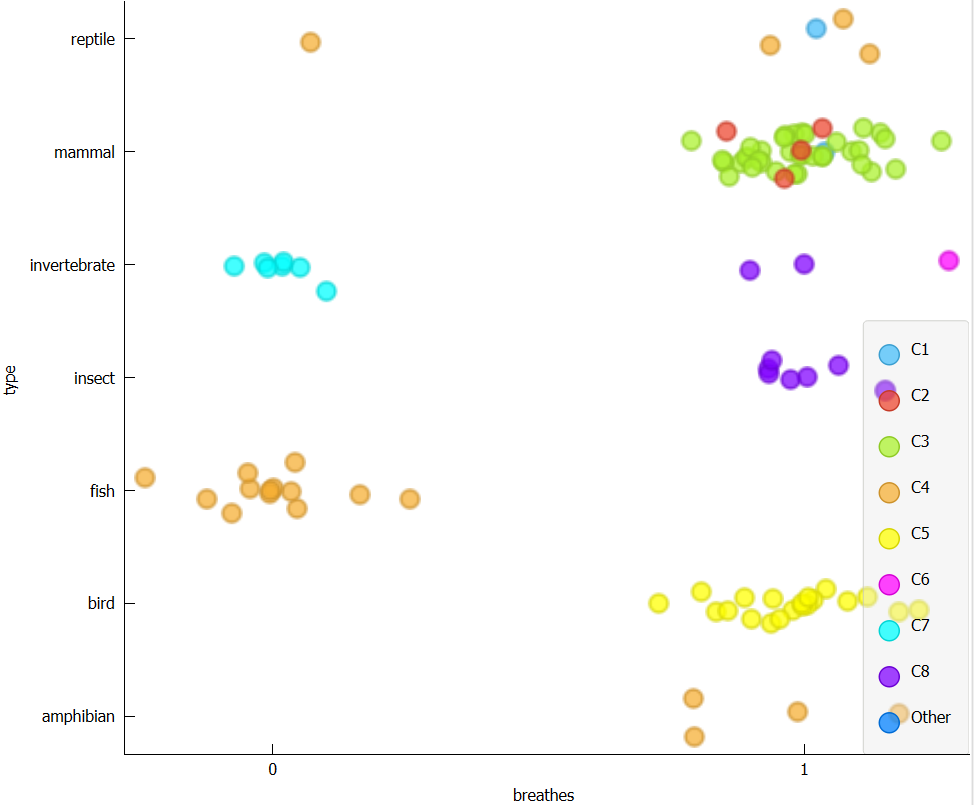


Рисунок 18 – Иерархическая кластеризация для евклидового расстояния

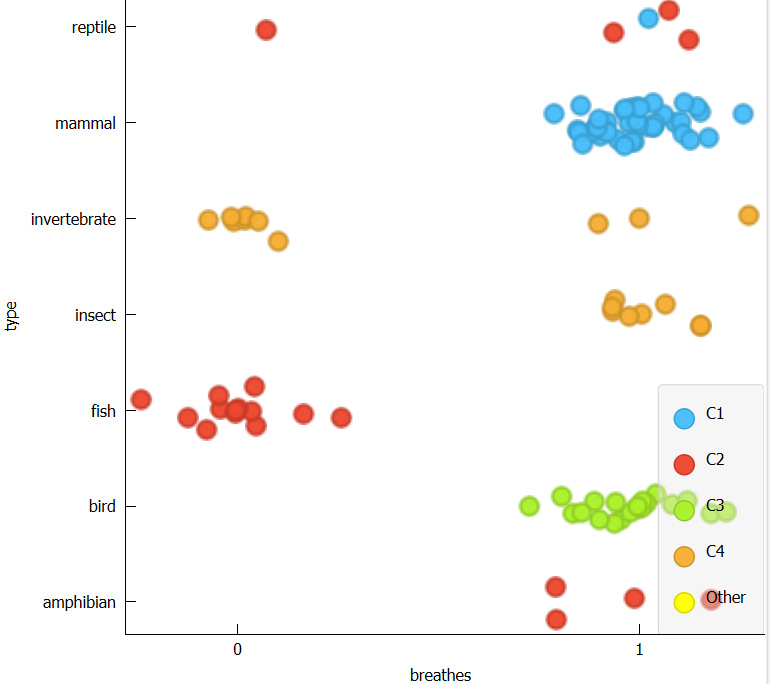


Рисунок 19 – Иерархическая кластеризация для манхэттенского расстояния

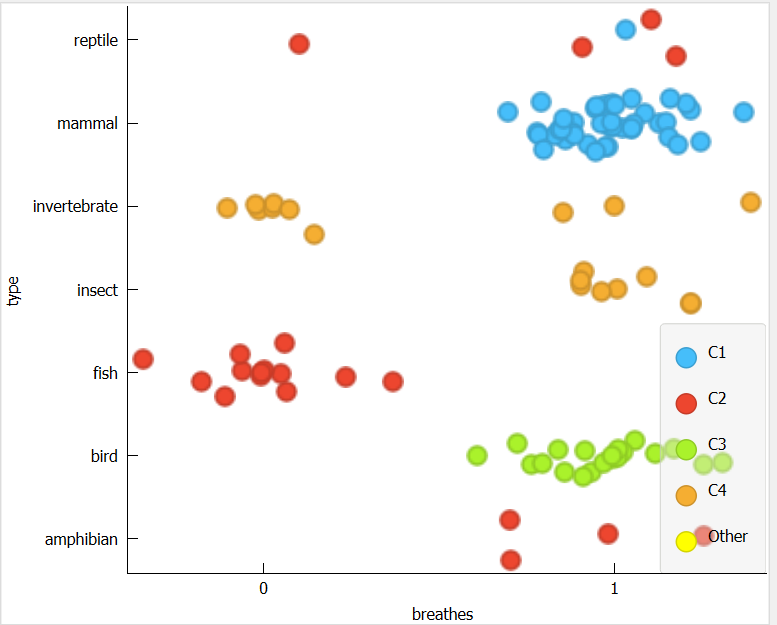


Рисунок 20 – Иерархическая кластеризация для расстояния Хэмминга

По приведенным данным можно сделать вывод, что кластеризация на исходных данных дает позволяет более точно и наглядно разбивать группы по кластерам.