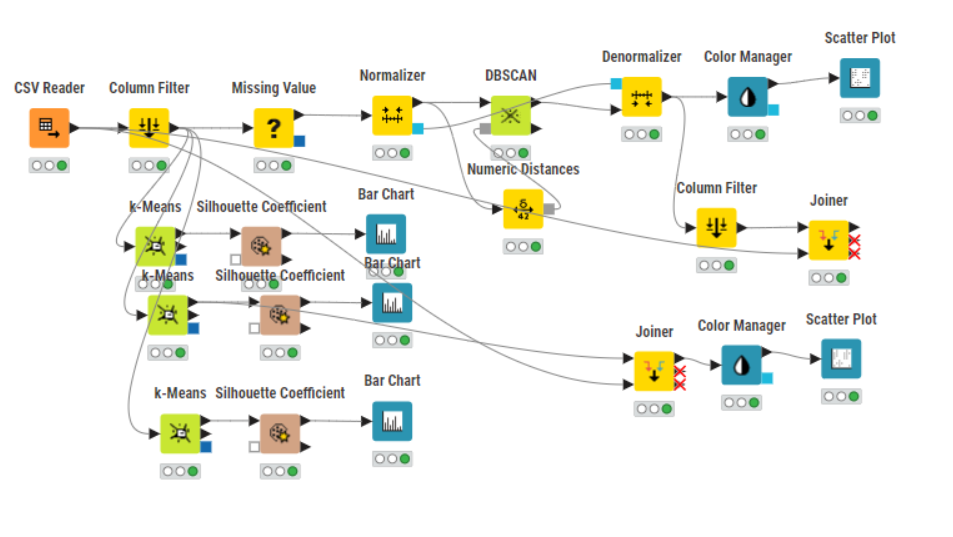
**Аналитический отчет по результатам кластерного анализа с использованием KNIME**

Выполнила: Лёвкина Екатерина гр. Э-2209

****

**Данные:**

SepalLength – длина чашелистика

SepalWidth - ширина чашелистика

PetalLenghth – длина лепестка

PetalWidth - ширина лепестка

Name – название вида ириса

1. **Количество кластеров**

Определим количество кластеров через построение столбчатых диаграмм среднего значения коэффициентов силуэта для кластеров. Рассмотрим три случая, построив коэффициент силуэта через K-means:

1. 2 кластера

Для 2-х кластеров значение коэффициента силуэта варьируется в пределах от 0.6 до 0.8.

1. 3 кластера

Для 3-х кластеров значение коэффициента силуэта варьируется в пределах от 0.4 до 0.8

Для обоих случаев значение для кластера ‘0’ выше, что указывает на явное разграничение элементов данного кластера и оставшихся данных. В кластер ‘0’ попадают iris-setosa, значит параметры для этого вида существенно отличны от других видов.

1. 4 кластера

Изображение выглядит как текст, снимок экрана, Прямоугольник, диаграмма

Контент, сгенерированный ИИ, может содержать ошибки.Изображение выглядит как снимок экрана, текст, диаграмма, Прямоугольник

Контент, сгенерированный ИИ, может содержать ошибки.Для 4-х кластеров значение коэффициента силуэта варьируется в окрестности 0.4

Изображение выглядит как текст, снимок экрана, Красочность, Прямоугольник

Контент, сгенерированный ИИ, может содержать ошибки.

Вывод: необходимо распределить данные на два кластера, так как это наиболее оптимально (значение коэффициента силуэта для кластеров в данном случае близко к 1)

1. **Подготовка к кластеризации**

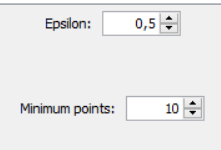
Для кластеризации отбираем только количественные признаки (все кроме Name), проводим нормализацию методом z-score – более предпочтительном при кластеризации, т.к. он позволяет скорректировать влияние влияние атрибутов с большими размерами и множеством выбросов и сравнивать сходства между признаками

1. **DBSCAN**

В отличие от K-means, где нужно заранее знать количество кластеров, DBSCAN находит их автоматически, группируя точки данных, которые расположены близко друг к другу.

Изображение выглядит как текст, снимок экрана, диаграмма, линия

Контент, сгенерированный ИИ, может содержать ошибки.Изображение выглядит как текст, диаграмма, снимок экрана, График

Контент, сгенерированный ИИ, может содержать ошибки.

Изображение выглядит как текст, диаграмма, снимок экрана, График

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

Контент, сгенерированный ИИ, может содержать ошибки.

Изображение выглядит как текст, диаграмма, снимок экрана, линия

Контент, сгенерированный ИИ, может содержать ошибки.

Основные параметры алгоритма: радиус окрестности и минимальное количество точек, которое может быть в одном кластере. Если расстояние между любыми двумя точками меньше или равно заданному радиусу, эти точки считаются соседними.

Видно, что при изменении параметров DBSCAN изменяется число кластеров, на которые разбиваются данные. При увеличении обоих параметров – радиуса окрестности и минимального количества точек (по умолчанию – 1 и 3 соответственно) до 1,5 и 20 данные распределились на два кластера и меньше точек воспринялось за «шум» в данных. При уменьшении параметров до 0,5 10 соответственно количество кластеров увеличилось и вместе с тем относительно большое количество наблюдений было воспринято за шум

1. **Сравнение результатов**

Изображение выглядит как текст, диаграмма, График, линия

Контент, сгенерированный ИИ, может содержать ошибки.Чтобы сравнить результаты кластеризации, проведенной двумя способами, стоит визуализировать результат k-means. Для этого соединим исходную таблицу и k-means через Joiner.

Изображение выглядит как текст, диаграмма, линия, График

Контент, сгенерированный ИИ, может содержать ошибки.

|  |  |
| --- | --- |
| DBSCAN | K-means |
| Cluster 0 iris-setosa  Cluster 1 – iris-versicolor и iris-virginica  Noise - iris-virginica (1 наблюдение) | Cluster 0 iris-setosa и несколько наблюдений iris-versicolor  Cluster 1 – iris-versicolor и iris-virginica |

**ВЫВОДЫ**

1. Оптимальное количество кластеров – два. Анализ коэффициента силуэта показал, что для двух кластеров его значения находятся в диапазоне 0.6–0.8, что говорит о хорошей четкости разделения.
2. DBSCAN четко выделяет один кластер для iris-setosa, а второй – для iris-versicolor и iris-virginica.
3. DBSCAN выявляет те же группы, что и K-means, но также относит часть данных к "шуму". Этот метод оказался полезен для выявления выбросов и учета плотности данных.
4. Несмотря на то, что видов исследуемых ирисов три алгоритмы кластеризации разделили данные на ДВА кластера, так как iris-setosa сильно отличается от других видов - этот вид имеет значительно отличающиеся параметры (например, более короткие лепестки), поэтому легко выделяется в отдельный кластер. Два других вида (iris-versicolor и iris-virginica) имеют пересекающиеся параметры, что мешает их четкому разделению. Коэффициент силуэта подтверждает, что два кластера — лучший вариант. Попытка разделить данные на три кластера приводит к менее выраженному разделению (значения коэффициента силуэта снижаются). Это указывает на то, что граница между iris-versicolor и iris-virginica размыта, и их проще объединить в один кластер.