ЕГОРОВА Д. К., ГАРИН М. А., САЙФЕТДИНОВ С. Ф.

СОЗДАНИЕ WORKFLOW АНАЛИТИЧЕСКОЙ ПЛАТФОРМЫ KNIME ДЛЯ АНАЛИЗА ДАННЫХ НА ПРИМЕРЕ ВАКАНСИЙ САЙТА HEADHUNTER

Аннотация. В статье приведено создание рабочего процесса аналитической платформы KNIME Analytics Platform. Кластеризация вакансий выполнена методами k-means и density-based.

Ключевые слова: KNIME, workflow, функциональный узел, кластеризация, k-means, density-based алгоритм.

EGOROVA D. K., GARIN M. A., SAIFETDINOV S. F.

CREATION OF WORKFLOW FOR ANALYTICAL PLATFORM KNIME FOR DATA ANALYSIS ON THE EXAMPLE OF VACANCIES ON THE SITE HEADHUNTER

Abstract. The process of creating a KNIME Analytics Platform workflow was explored in the article. Clustering of vacancies was performed by k-means and density-based methods.

Keywords: KNIME, workflow, functional node, clustering, k-means, density-based algorithm.

В настоящее время существует множество инструментальных средств машинного обучения позволяющих осуществлять анализ данных. В работе приведено создание рабочего процесса workflow аналитической платформы KNIME Analytics Platform [1] на примере данных, предоставляемых HeadHunter – одним из самых крупных сайтов по поиску работы и сотрудников в мире [2]. KNIME Analytics Platform – это Java-кроссплатформенное приложение с открытым исходным кодом для анализа данных, объединяющее различные компоненты машинного обучения и интеллектуального анализа посредством модульной конвейерной обработки данных «Lego of Analytics». Выбор HeadHunter в качестве источника данных обуславливается наличием открытого API [3].

Решим задачу кластеризации данных по вакансии «Программист» в населенном пункте «Саранск» по состоянию на 23 мая 2021 года, которая покажет разбиение данных по величине предлагаемой заработной платы. Для этого создадим рабочий процесс (Workflow), считаем данные, воспользовавшись соответствующим узлом (Node) GET Request (Tools & Services→REST Web Services→GET Request), перенесем его из репозитория узлов на рабочее пространство. В конфигурации данного узла размещаем следующий код на языке python, осуществляющий запрос.

Фрагмент кода:

```
https://api.hh.ru/vacancies?text=программист&area=63&per_page=100&only_with_salary=true
```

Для извлечения вакансий воспользуемся узлом JSON Path (Structured Data → JSON→ JSON Path) и соединяем его с GET Request (рис. 1).

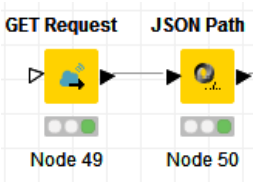


Рис. 1. Узлы, осуществляющие извлечение данных.

Данные были получены в виде таблицы (Рисунок 2), в том числе содержащей поля «from» и «to», означающие границы измерения зарплаты.

Table "default" - Rows: 1 | Spec - Columns: 7 | Properties | Flow Variables

Row ID	Status	Content type	body	from	to
Row0	200	application/json; charset=UTF-8	{ "items": [{ "id": "44770732", "premium": false,	[80000,80000,30000,...	[100000,100000,50000,...

Рис. 2. Таблица извлеченных данных.

Данные в таком виде использовать для кластеризации нельзя. Необходимо выполнить элементы так называемого разведочного анализа данных, а именно осуществить разгруппировку данных по столбцам, фильтрацию этих столбцов, вычисление основных статистик, заполнение пропущенных значений в полях с размером заработной платы (например, медианным значением), так как не все работодатели указали на сайте значения полей «from» и «to» и, наконец, их нормализацию. Для этого были использованы узлы Ungroup (Manipulation→Row→Transform→Ungroup), Column Filter (Manipulation→Colum→ Filter→ Column Filter) и Missing Value (Manipulation→Colum→ Transform→ Missing Value). Узел Box Plot (Views→JavaScript→Box Plot) позволяет построить «ящичковые диаграммы» данных полей «from» и «to». На диаграммах отображаются статистические параметры: минимум, нижний квартиль, медиана, верхний квартиль (соответственно, 25-й, 75-й 50-й процентили) и максимум. Эти параметры называются надежными, поскольку они нечувствительны к экстремальным выбросам. На рисунке 3 представлены данные до нормализации и заполнения пропущенных значений, из которого видно, что данные параметра «to» имеют несколько «выбросов» (точка) и экстремальное значение (крест).

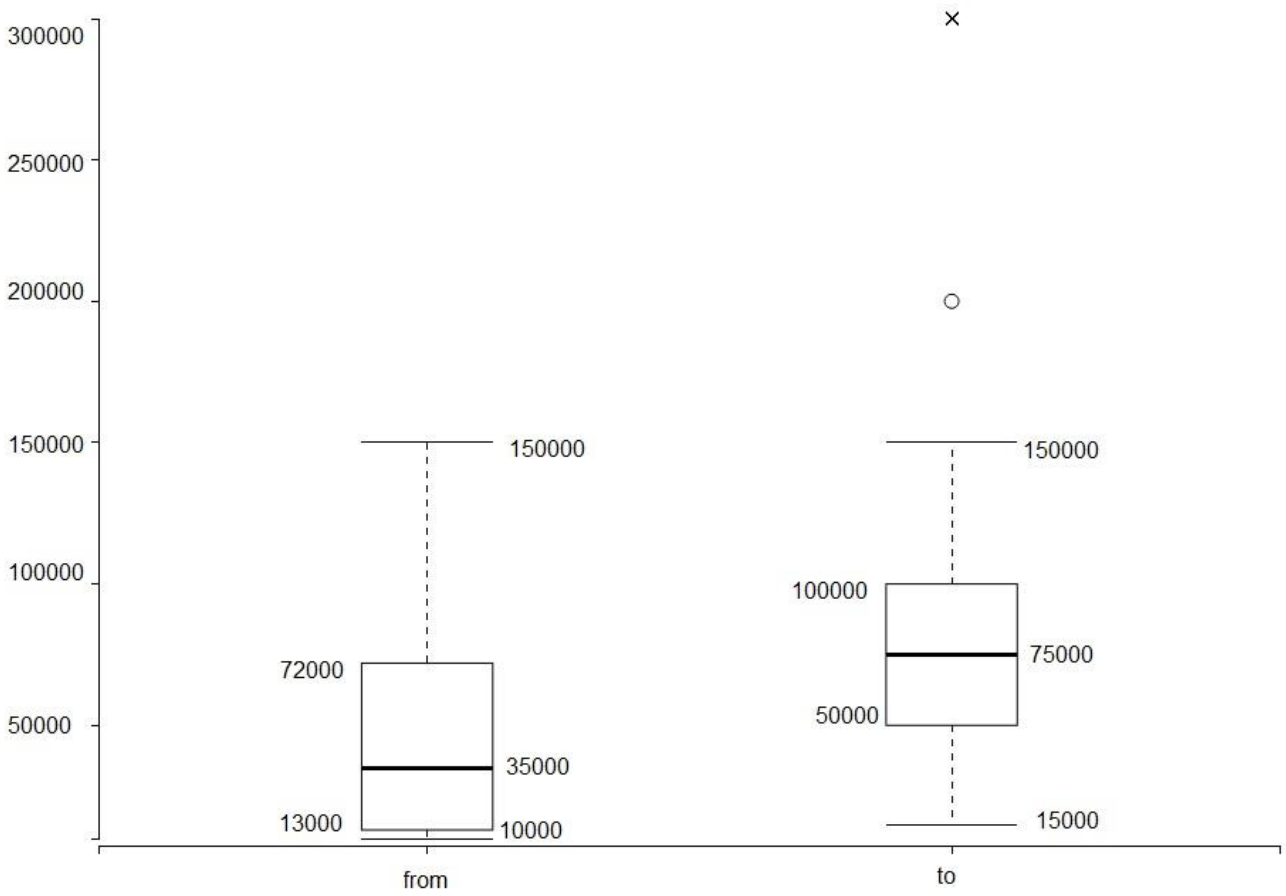


Рис. 3. Ящичковые диаграммы данных полей «from» и «to».

Узел Statistics (Statistics→Hypothesis Testing→Statistics) позволяет найти основные характеристики данных (таблица 1).

Таблица 1

Значения статистик

Статистика	Данные до нормализации		Данные после нормализации и заполнения пропущенных значений	
	«from»	«to»	«from»	«to»
Минимум	10000	15000	0	0
Максимум	150000	300000	1	1
Среднее значение	43288	89453	0,235	0,228
Стандартное отклонение	35421	65296	0,247	0,133
Асимметрия	1,123	1,19	1,18	3,724
Эксцесс	0,472	4,711	1,685	19,591
Общая сумма	2554000	1878520	14,564	14,118
Количество пропущенных значений	3	41	0	0
Количество строк	62	62	62	62

Далее была выполнена кластеризация методом k-means и алгоритмом на основе плотности (density-based алгоритм или DBSCAN). Для этого использовались узлы k-Means (Analytics→Mining→Clustering→k-Means) и DBSCAN(Analytics→Mining→Clustering→ DBSCAN). Параметры кластеризации подбирались путем использования метода силуэтов, который реализован посредством узла Silhouette Coefficient (Analytics→Scoring→ Silhouette Coefficient). Общий коэффициент силуэтов Overal, рассчитанный для каждого метода кластеризации, должны быть не менее 0,5, иначе текущее разбиение на кластеры является нецелесообразным. На рис. 4 представлены результаты работы метода силуэтов для методов k-means (рисунок 4 а) и DBSCAN (рисунок 4 б).

Row ID	D Mean Si...
cluster_2	0.319
cluster_0	0.517
cluster_1	0.757
Overall	0.57

а)

Row ID	D Mean Si...
Cluster_2	0.451
Cluster_4	0.476
Cluster_3	0.975
Noise	-0.272
Cluster_0	0.766
Cluster_1	1
Overall	0.629

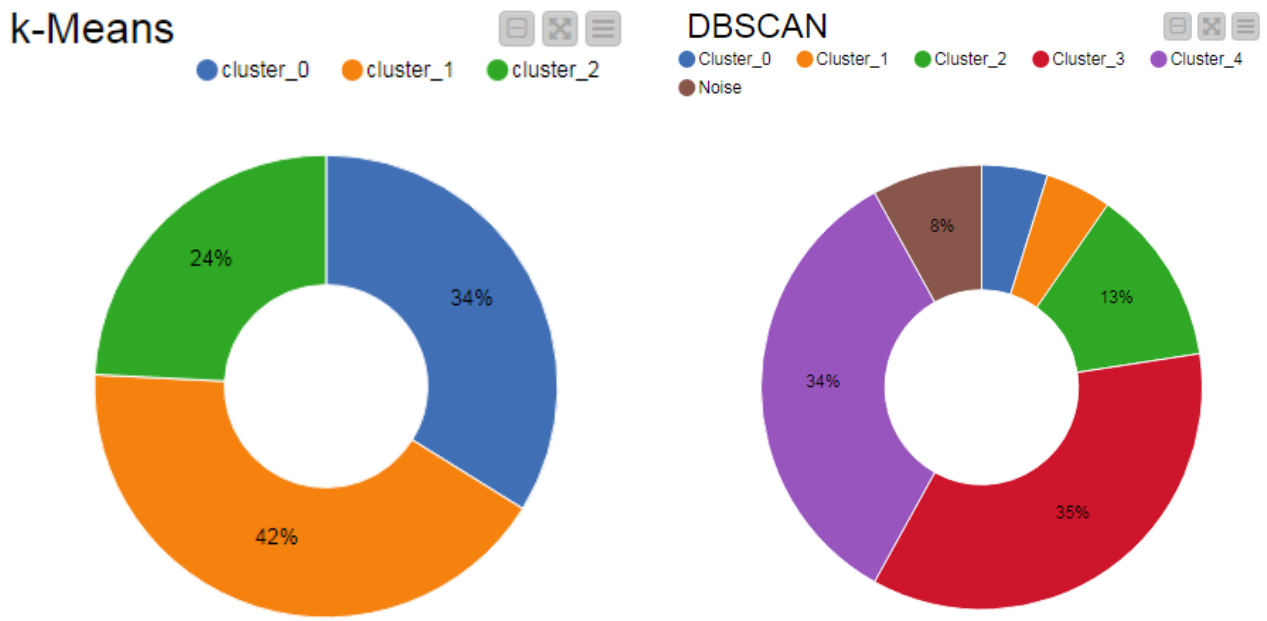
б)

Рис. 4. Коэффициенты силуэтов. а) Метод k-means, б) Метод DBSCAN.

Метод k-means позволяет определить центры кластеров (центроиды) для заранее определенного количества кластеров (его определили с помощью метода силуэтов). K-means выполняет четкую кластеризацию, которая назначает вектор данных ровно одному кластеру и в качестве метрики использует Евклидово расстояние. Алгоритм завершается, когда назначения кластера больше не меняются.

Метод DBSCAN определяет три типа точек в наборе данных. Базовые точки имеют количества соседей большее чем минимальное значение (выбрано MinPts=3) в пределах указанного расстояния (выбрано eps=0,1). Граничные точки находятся в пределах «eps» от основной точки, но имеют меньше соседей «MinPts». Точки шума «Noise» не являются ни основными, ни граничными точками. Кластеры создаются путем соединения основных точек друг с другом. Если основная точка находится в пределах «eps» от другой базовой точки, они называются непосредственно достижимыми по плотности. Все точки, которые находятся в пределах «eps» от основной точки, называются доступными по плотности и считаются частью кластера. Все остальные считаются шумом. Метод DBSCAN требует определения метрики, было выбрано Евклидово расстояние (Analytics→Distance Calculation→Numeric Distances).

Визуализация результатов работы алгоритмов кластеризации приведена на рис. 5 (а – k-means, б – DBSCAN). Кольцевые диаграммы (см. рисунок 5) выполнены на основе библиотеки NVD3 при помощи узла Pie/Donut Chart (Views→JavaScript→Pie/Donut Chart).



а) б)

Рис. 5. Кольцевые диаграммы визуализирующие работу методов кластеризации.

а) Метод k-means, б) Метод DBSCAN.

Так же с помощью узла Scatter Plot (Views→JavaScript→Scatter Plot) может быть выполнена визуализация координат центроидов.

Анализируя результаты кластеризации отметим, что метод k-means разбил вакансии на три кластера, соответствующие средним значениям по заработной плате 13000₽, 40000₽ и 100000₽ соответственно. Метод DBSCAN – на 5 кластеров, выделив при этом порядка 8% зашумленных данных.

На рис. 6 представленных итоговый рабочий процесс анализа вакансий.

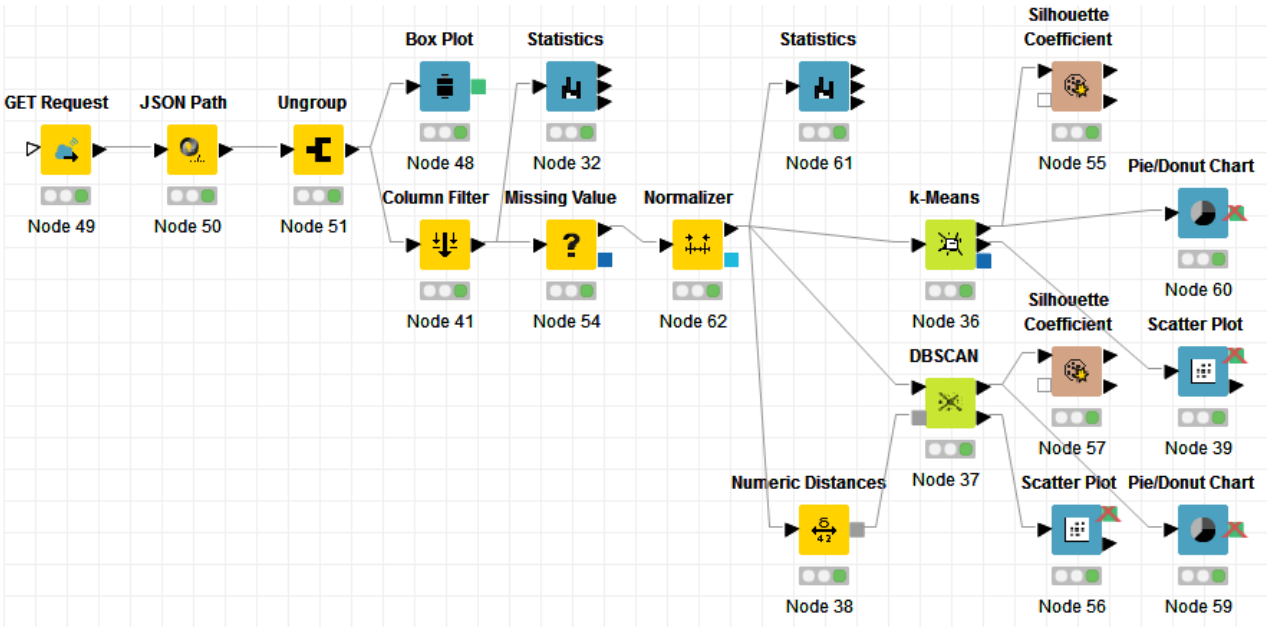


Рис. 6. Рабочий процесс анализа вакансий.

Данный процесс может быть использовать для создания шаблонов отчетов, экспортируемых в такие форматы документов, как doc, ppt, xls, pdf и другие, что может весьма эффективно использоваться для отслеживания информации на рынке вакансий в режиме реального времени.

СПИСОК ЛИТЕРАТУРЫ

1. KNIME Analytics Platform [Электронный ресурс]. – Режим доступа: https://www.knime.com/knime-analytics-platform (дата обращения: 10.09.2021).
2. HeadHunter [Электронный ресурс]. – Режим доступа: https://hh.ru (дата обращения: 10.09.2021).
3. HeadHunter API [Электронный ресурс]. Режим доступа: https://dev.hh.ru (дата обращения 10.09.2021).