|  |
| --- |
|  |
| МИНОБРНАУКИ РОССИИ |
| Федеральное государственное бюджетное образовательное учреждение  высшего образования  **"МИРЭА - Российский технологический университет****"**  **РТУ МИРЭА** |
| Институт информационных технологий (ИТ) |
| Кафедра инструментального и прикладного программного обеспечения (ИиППО) |

|  |
| --- |
| **ОТЧЕТ ПО ЛАБОРАТОРНОЙ РАБОТЕ №1** |
| **по дисциплине** |
| **«**Программное обеспечение интеллектуальных систем»  *Изучение технологии анализа данных* |

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |
| Выполнил студент группы ИКБО-02-17 | *Травин М.Б.* |
| Приняла | *Зорина Н.В* |

|  |  |  |
| --- | --- | --- |
| Работа выполнена | «\_\_\_» \_\_\_\_\_\_\_\_ 2020 г. |  |
|  |  |  |
| «Зачтено» | «\_\_\_» \_\_\_\_\_\_\_\_ 2020г. |  |

**Цели работы**

Основной целью работы является познакомиться с мощным графическим инструментом для анализа данных Knime. Работа подразделяется на три цели:

1. Ознакомиться с механизмом поиска шаблонов (правил)
2. Научиться работать с деревом решений
3. Освоить возможности Knime по кластеризации

**Поиск шаблонов**

**Постановка задачи**

Построение потока работ, выполняющего решение задачи анализа рыночной корзины и поиска ассоциативных правил. Данный поток должен выполнять следующую последовательность действий: загрузить данные из текстового файла, преобразовать загруженные данные в специализированный тип данных пакета KNIME, найти частые наборы и ассоциативные правила, вывести результаты.

**Ход работы**

Создадим поток работ, состоящий из следующих узлов: узла чтения файла, узла создания выборки, узла вывода правил и узла вывода результата (интерактивной таблицы) (рис. 1).

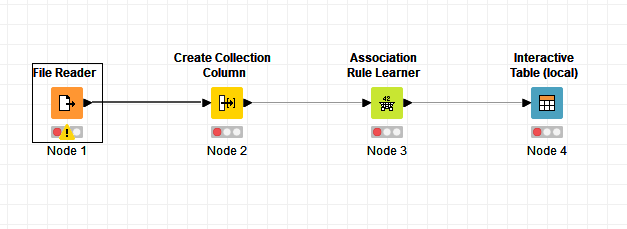


Рисунок 1 – Поток работ.

Настроим узлы следующим образом:

1. Узел чтения файла настроим так, чтобы он считывал файл vgsales.csv, первая строка трактовалась как названия столбцов, поставим разделитель и обработку неполных строк
2. В узле выборки настроим выборку по поля Year и Genre
3. В узле вывода правил, поставим значение minimum support 0.01 для того чтобы увеличить вероятность вхождения элементов в выборку, а в графе Association Rules проставим параметр minimum confidence 0.03, чтобы повысить вероятность создания правила (рис 2).
4. Узел вывода таблица не нуждается в настройке

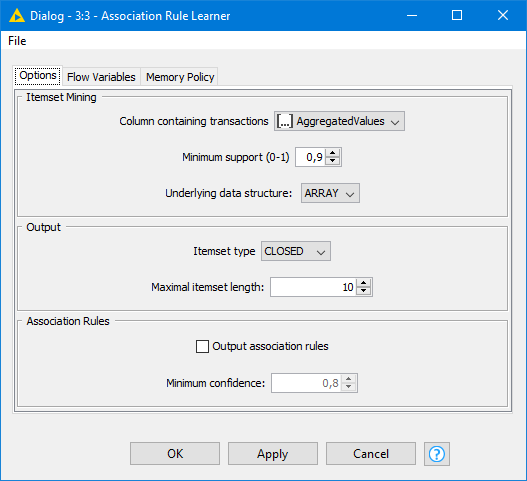


Рисунок 2 – Настройка узла вывода правил.

Выполним (execute) все узлы, после чего посмотрим на отсортированный по параметрам lift и confidence (для наглядного подтверждения верности правил) вывод таблицы (рис. 3).

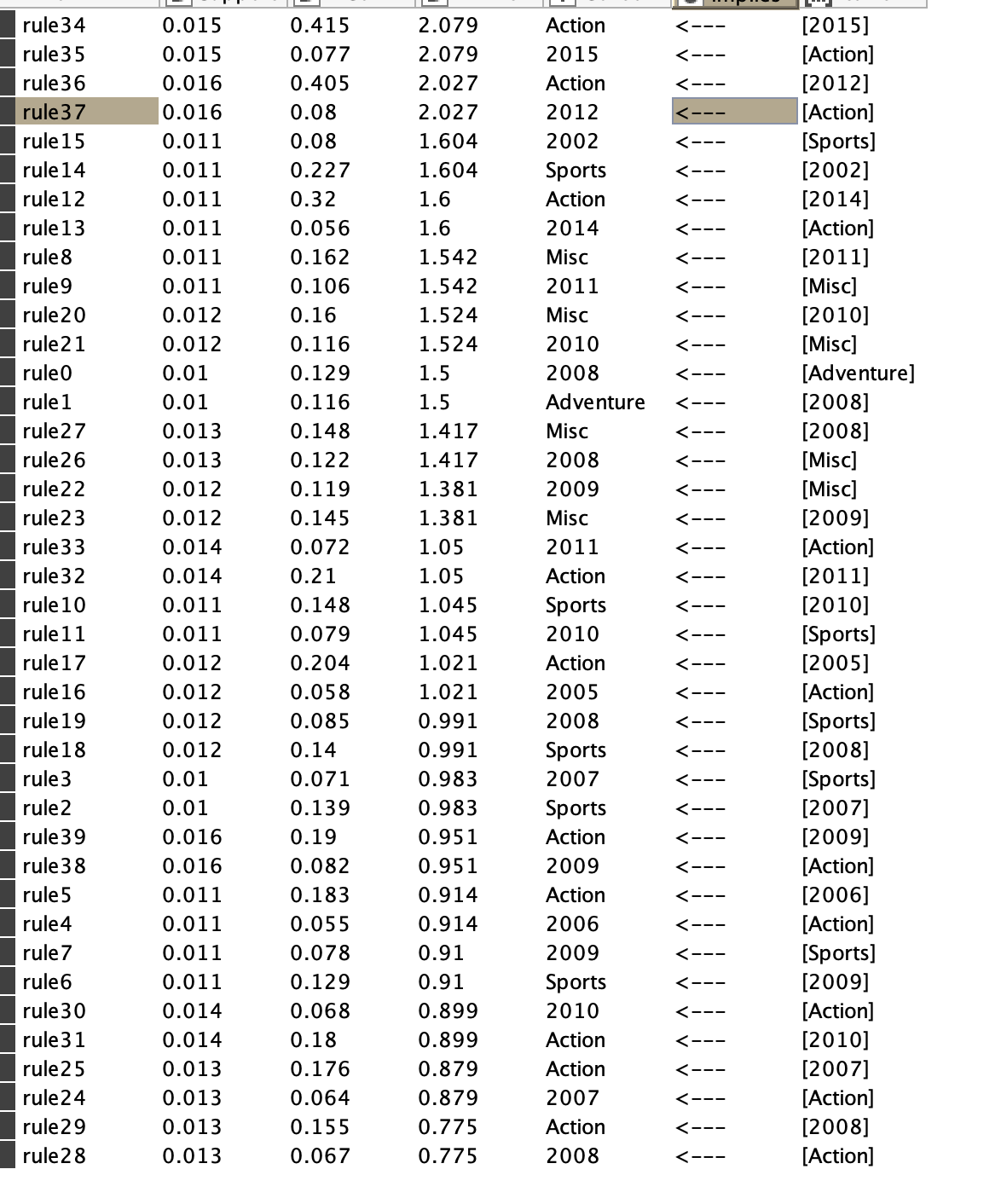


Рисунок 3 – Вывод полученных правил.

Ориентировать по таблице можно следующим образом, чем больше confidence, тем точнее правило, однако вероятность того, что это правило общее, а не частное определяется параметром lift – чем больше lift, тем больше такая вероятность. Следовательно правила 34 и 36, например, являются истинными и показательными: если год 2015 или 2012, то скорее всего будет Action. Обратным же образом работает правило 19 – Sports не гарантирует 2008 год.

**Поиск шаблонов**

**Постановка задачи**

Построение потока работ, выполняющего решение задачи классификации посредством построения дерева решений. Данный поток должен выполнять следующую последовательность действий: загрузить данные из текстового файла, построить дерево решений, вывести результаты.

**Ход работы**

Построим поток работ, состоящий из узла чтения файла, узла фильтрации и узла DTL – создания дерева решения (рис. 4).

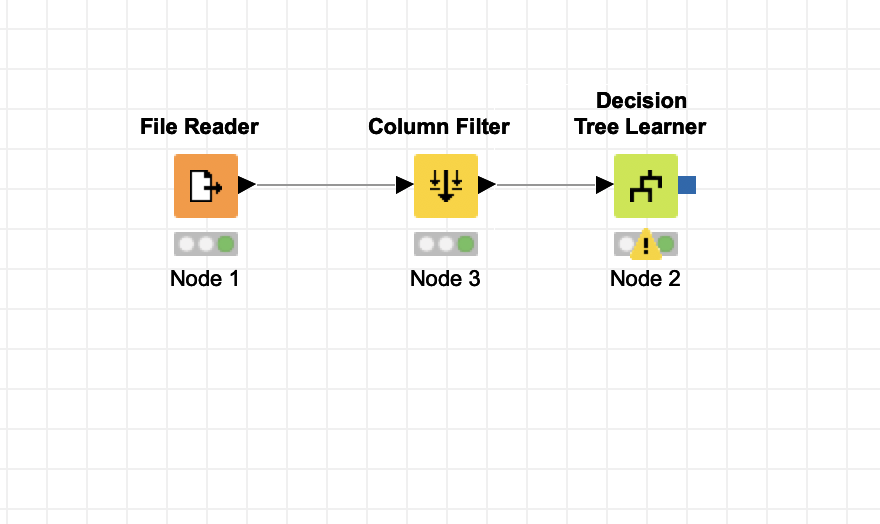


Рисунок 4 – настройка потока работ.

Настроим узел фильтрации так, чтобы не пускать рейтинг фильма и его выручку без учет инфляции дальше, узел DTL настроим так, чтобы параметр genre был основным для разделения. Посмотрим на частичный вывод дерева (рис. 5).

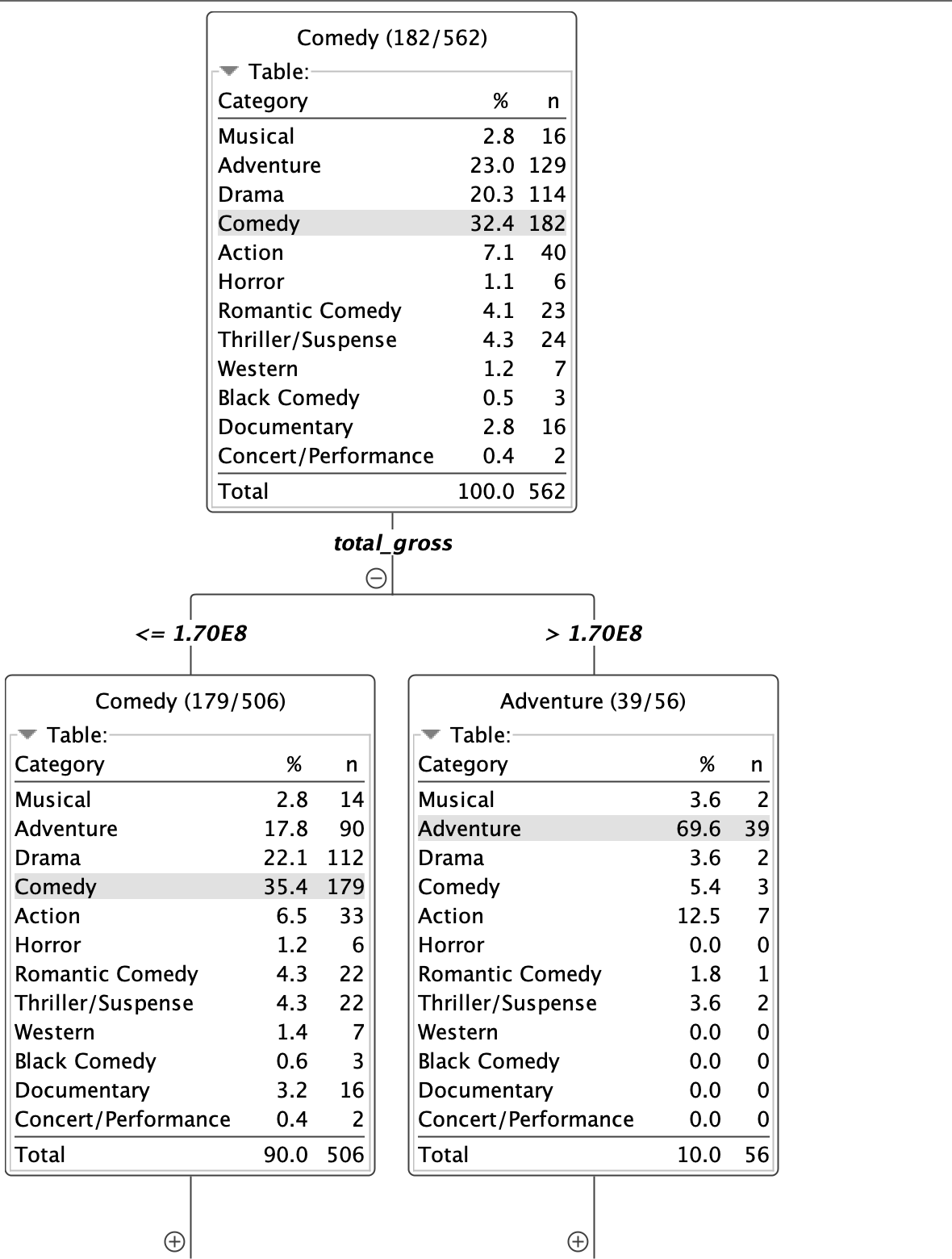


Рисунок 5 – Неполный вывод дерева.

На основании вывода (если раскрывать его дальше), можно сделать выводы о зависимости выручки от жанра. Так, если мы дойдем до финального, так называемого чистого – pure – класса разделения в большую сторону, мы обнаружим, что среди самых «дорогих» фильмов (с выручкой более1 770 000 000) Disney окажутся преимущественно музыкальные фильмы. В целом дерево покажет чистую разделение по количеству выручки между жанрами.

Выведем дерево картинкой, воспользовавшись потоком работ из методического пособия и несколько доработав его, чтобы цвет появился на финальном изображении (рис. 6).

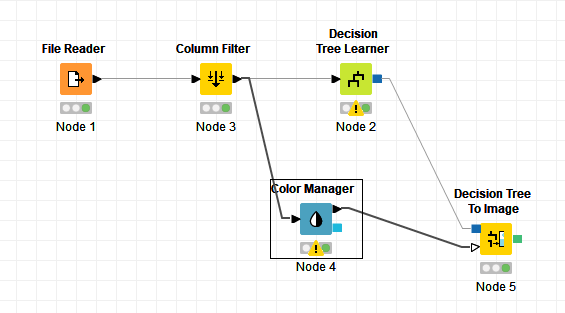


Рисунок 6 – Поток работ для вывод цветного дерева картинкой.

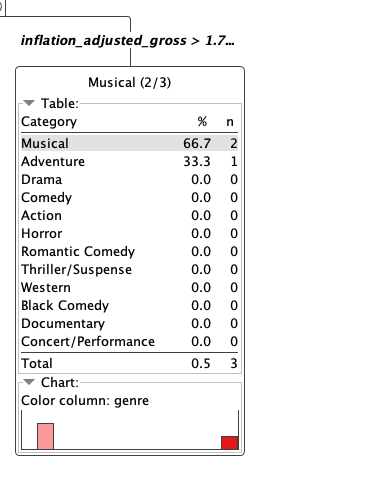


Рисунок 8 – Финальный узел дерева в цвете.

**Кластеризация**

**Постановка задачи**

Построение потока работ, выполняющего решение задачи кластеризации. Данный поток должен выполнять следующую последовательность действий: загрузить данные из текстового файла, построить дерево решений, вывести результаты.

**Ход работы**

Построим поток работ, руководствуясь с тем лишь изменением, что добавим узел фильтрации строк, который сконфигурируем на удаление строк с пустыми значениями Year (рис. 9). Настроим его, посмотрим выводы узлов вывода (рис. 10 и 11)

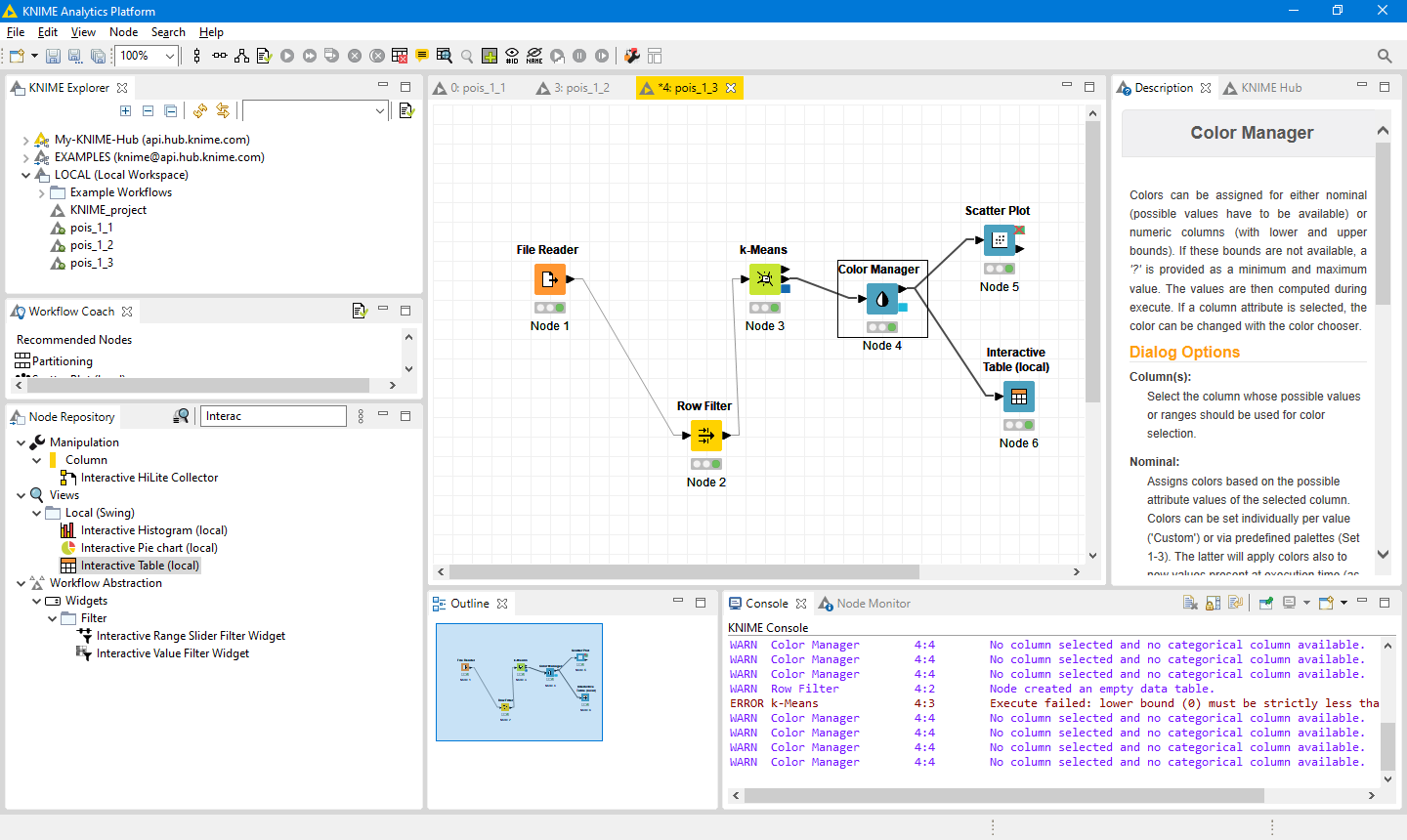


Рисунок 9 – Поток работ для кластеризации.

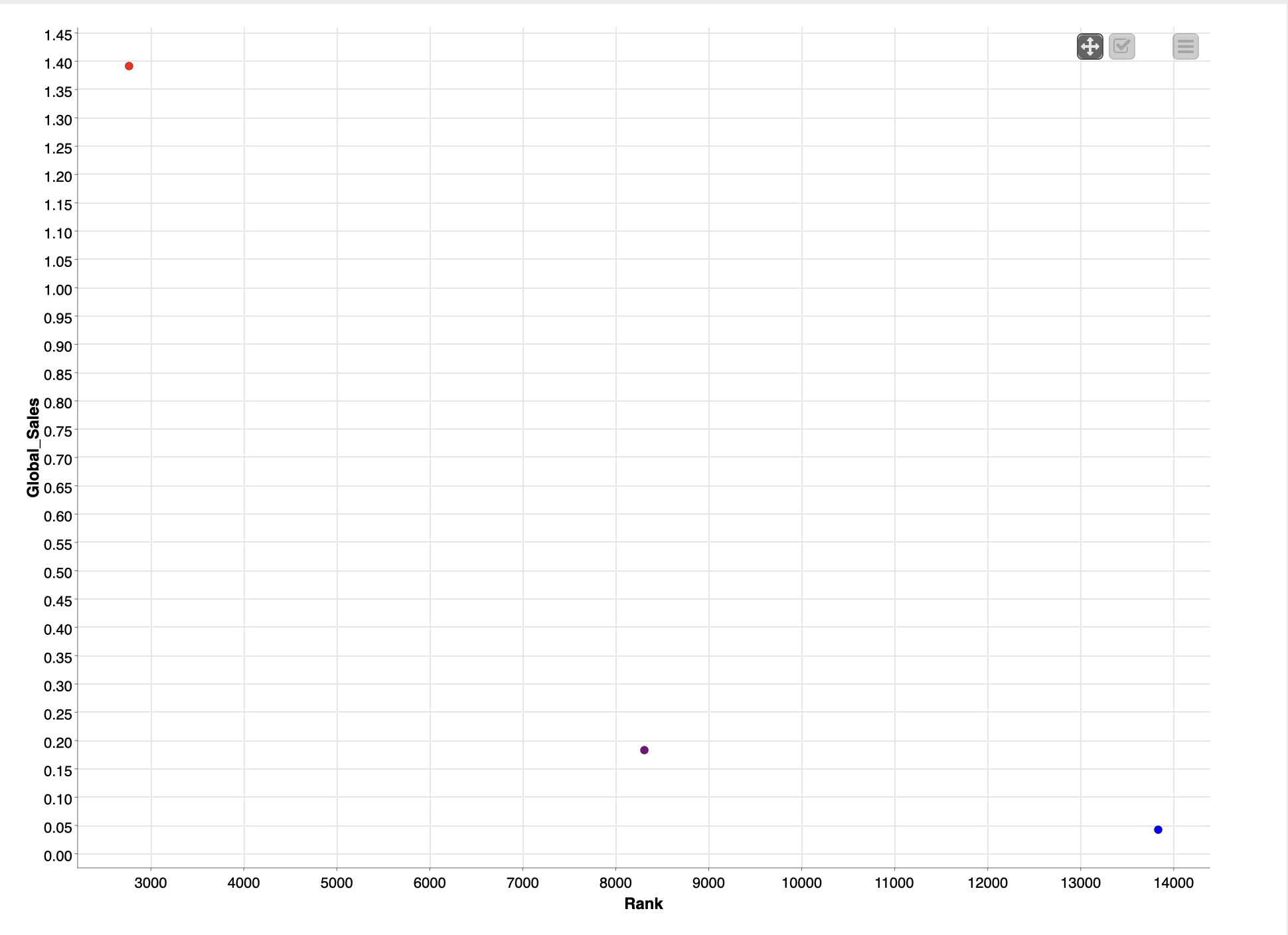


Рисунок 10 – вывод кластеров по цветам.

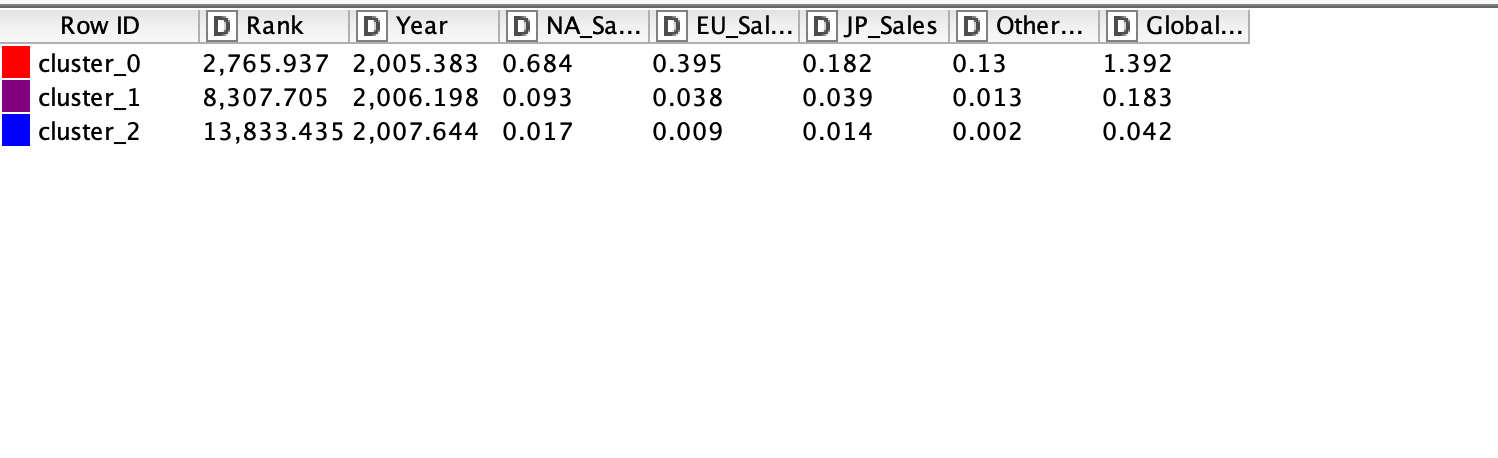


Рисунок 11 – таблица кластеров.

**Вывод**

В ходе выполнения данной работы были освоены навыки обучения дерева решений в Knime по заданному датасету.

Кроме того, были получены навыки кластеризации в Knime по заданному датасету.

Наконец, были усвоены навыки генерации правил в Knime по заданному датасету.

Список используемых источников

1. Docs Knime [Электронный ресурс]. – Режим доступа: https://docs.knime.com/ – (Дата обращения: 28.10.2020).
2. Knime TV [Электронный ресурс]. – Режим доступа: https://www.youtube.com/user/KNIMETV – (Дата обращения: 28.10.2020).
3. Association Rules and the Apriori Algorithm: A Tutorial [Электронный ресурс]. – Режим доступа: https://www.kdnuggets.com/2016/04/association-rules-apriori-algorithm-tutorial.html. – (Дата обращения: 28.10.2020).