|  |  |
| --- | --- |
| lu135925on3bu_tmp_3360867a00ce4d37 | **Министерство науки и высшего образования Российской Федерации Федеральное государственное бюджетное образовательное учреждение высшего образования** **«Московский государственный технический университет имени Н.Э. Баумана** **(национальный исследовательский университет)» (МГТУ им. Н.Э. Баумана)** |

ФАКУЛЬТЕТ Информатика и системы управления и искусственный интеллект

КАФЕДРА                  Системы обработки информации и управления

**Домашнее задание №3**

**По курсу**

**«Оптимизация баз данных систем машинного обучения»**

**«Обнаружение условных функциональных зависимостей»**

**Вариант 4**

Выполнил: Журавлев Н.В.

Группа: ИУ5-14М

Дата: 25.11.2023

Проверил:

Плужникова О. Ю.

2023 г.

# **Задание**

Для одного из двух наборов данных:

# Задача 1. Определите наименьшее число признаков набора данных, которые можно использовать для машинного обучения модели по набору данных (например, для обучения нейронной сети).

# Задача 2. Определите функциональные зависимости при условии, что целевой признак равен определённому значению.

# Задача 3. Определите внешние ключи для соединения таблиц озера данных в одну таблицу перед машинным обучением.

# **Набор данных по варианту**

|  |  |  |  |
| --- | --- | --- | --- |
| ИУ5-14М 2023 | | Варианты наборов данных | |
| 4 | Журавлев Николай Вадимович | 4  (07 adult.zip) | 20  (24 real+estate+valuation+data+set.zip) |

# **Ход работы**

### Выбор набора данных

Был выбран набор adult, т.к. при другом наборе после выполнения задачи 2 получается результат "No results found!", а попытка уменьшить параметр MIN\_CONFIDENCE (достоверность) с 1.0 до 0.99 приводит к формированию только шаблонов (\_,\_), что не позволяет выполнить дальнейший пункт задания.

### Задача 1

Выделите в наборе данных множество входных признаков(R) и целевой признак (T). Всего в наборе 32561 записей.

Набор был изначально без неинформативных столбцов, поэтому удалять их не нужно. Были удалены дубликаты записей по признаку R. Таким образом был получен файл RT\_XLSX.

Затем был удалён столбец с выходным признаком T и через алгоритм HyUCC были найдены ключи, представленные на рис.1. После чего был выбран минимальный ключ, а обозначим признаки (атрибуты) ключа через K.

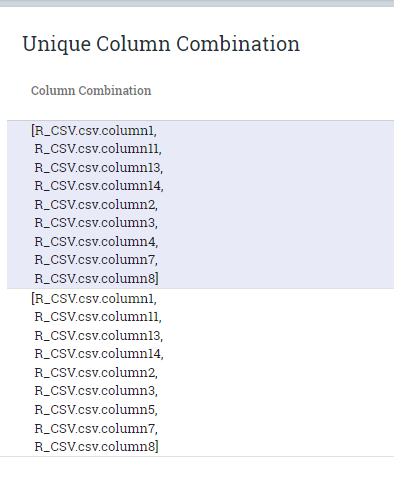


Рисунок 1. Результат выполнения HyUCC

Затем из RT\_XLSX те столбцы в R, которые не входят в ключ K были удалены. Так же удалены записи с дубликатами по признакам K. В результате получился файл KT\_XLSX с 32536 строк.

### Задача 2

Преобразуем xlsx-файл KT\_XLSX в csv-файл KT\_CSV. Для этого файла постоем условные функциональные зависимости (УФЗ) с помощью программы CFDFINDER и получим результат на рис.2.

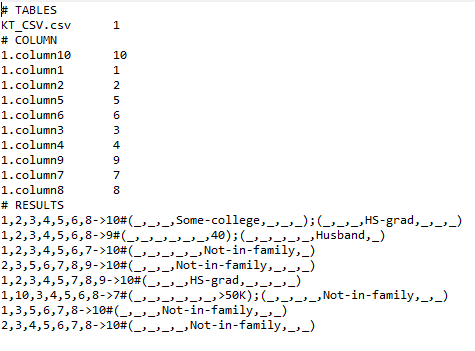


Рисунок 2. Результат выполнения CFDFINDER

У целевого признака имеется всего 2 значения – “>50”, “<=50”. Для каждого из них в файле результата выпишим УФЗ. УФЗ для “>50K” представлена только 1,10,3,4,5,6,8->7#(\_,\_,\_,\_,\_,\_,>50K);(\_,\_,\_,\_,Not-in-family,\_,\_). Для “<=50” УФЗ не найдено.

### Задача 3

Выберете в Metanome алгоритм FAIDA, и следующие наборы данных:

WDC\_symbols.csv,

WDC\_science.csv,

WDC\_satellites.csv,

WDC\_planetz.csv,

WDC\_planets.csv,

WDC\_kepler.csv,

WDC\_game.csv,

WDC\_astronomical.csv,

WDC\_astrology.csv,

WDC\_appearances.csv,

WDC\_age.csv

Построим связи PK-FK для первичного ключа Planet. Результат представлен на рис. 5.

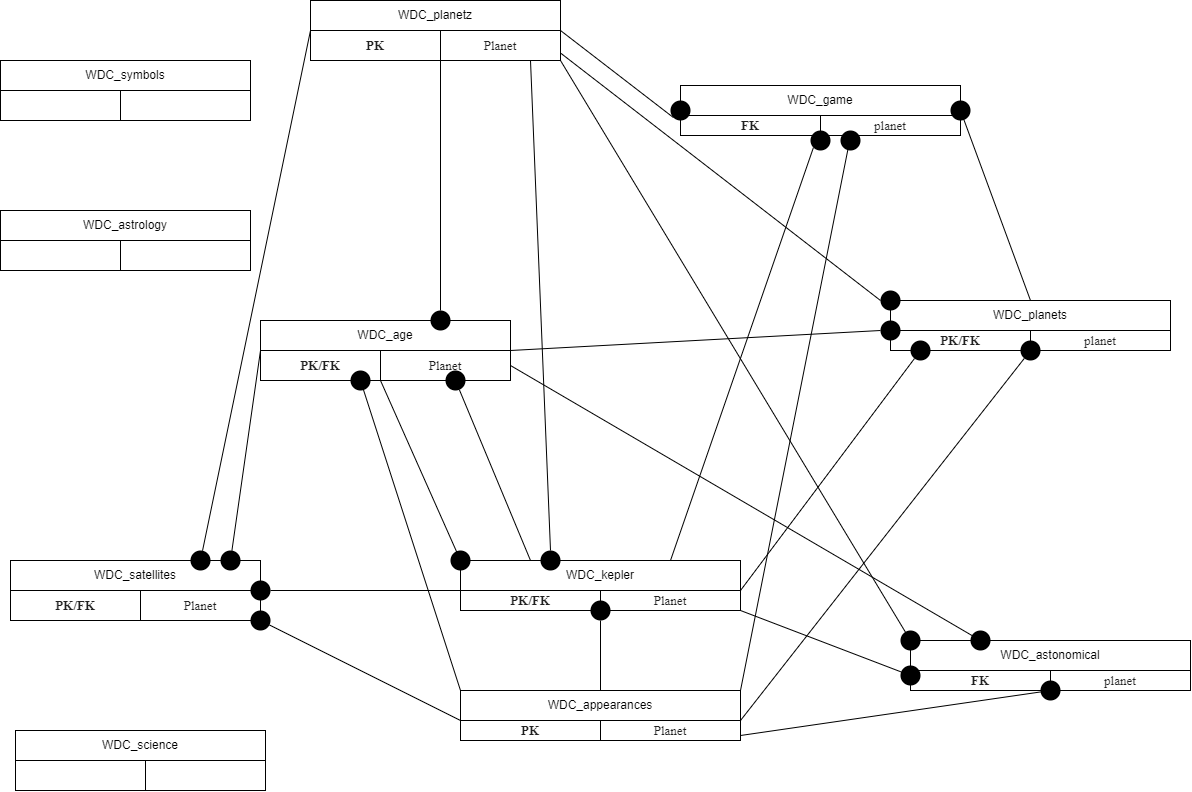


Рисунок 3.Связи PK-FK

Результат алгоритма представлен на рисунках 6, 7, 8, 9, 10.

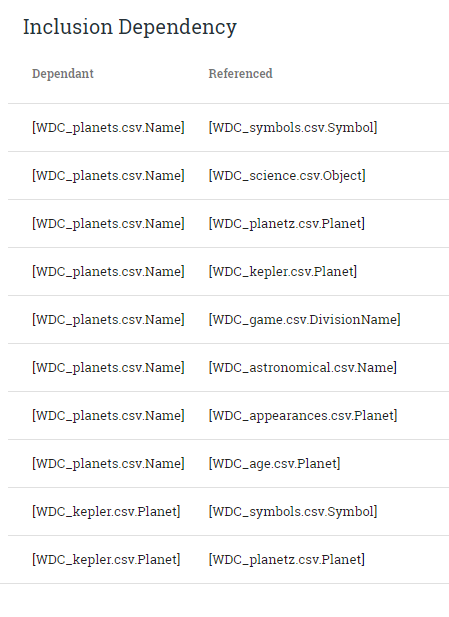


Рисунок 4



Рисунок 5

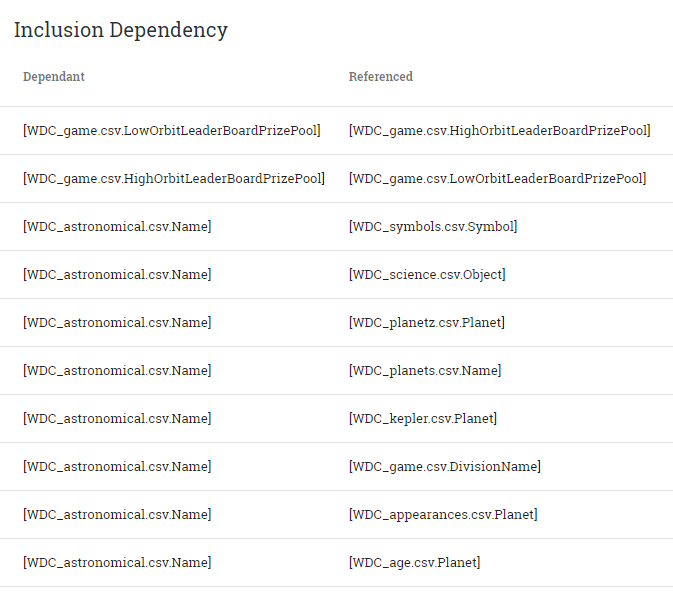


Рисунок 6



Рисунок 7

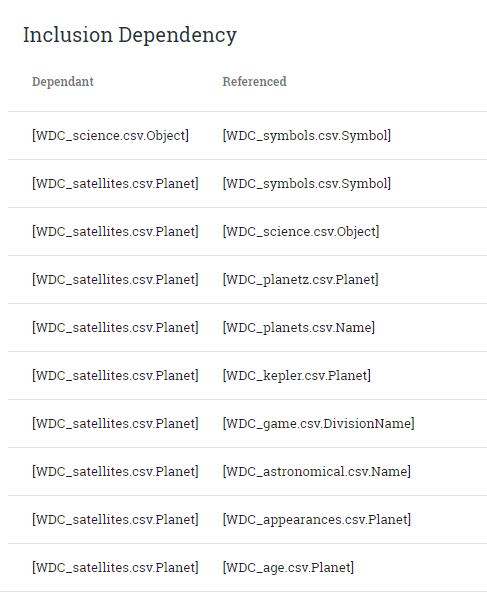


Рисунок 8