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.

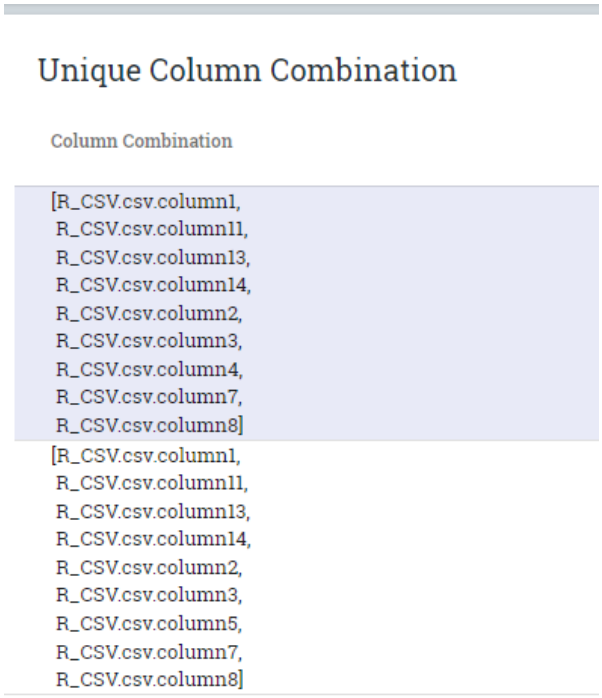

Рисунок 1. Результат выполнения HyUCC

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 построем условные функциональные зависимости (УФЗ) с помощью программы CFDFINDER и получим результат на рис.2.

```
# TABLES
KT_CSV.csv      1
# COLUMN
1.column10      10
1.column1       1
1.column2       2
1.column5       5
1.column6       6
1.column3       3
1.column4       4
1.column9       9
1.column7       7
1.column8       8
# RESULTS
1,2,3,4,5,6,8->10#(_,_,_,Some-college,_,_,_);(_,_,_,HS-grad,_,_,_)
1,2,3,4,5,6,8->9#(_,_,_,_,_,_,40);(_,_,_,_,_,Husband,_)
1,2,3,4,5,6,7->10#(_,_,_,_,_,Not-in-family,_)
2,3,5,6,7,8,9->10#(_,_,_,Not-in-family,_,_,_)
1,2,3,4,5,7,8,9->10#(_,_,_,HS-grad,_,_,_,_)
1,10,3,4,5,6,8->7#(_,_,_,_,_,_,_,>50K);(_,_,_,_,Not-in-family,_,_)
1,3,5,6,7,8->10#(_,_,_,Not-in-family,_,_)
2,3,4,5,6,7,8->10#(_,_,_,_,Not-in-family,_,_)
```

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.

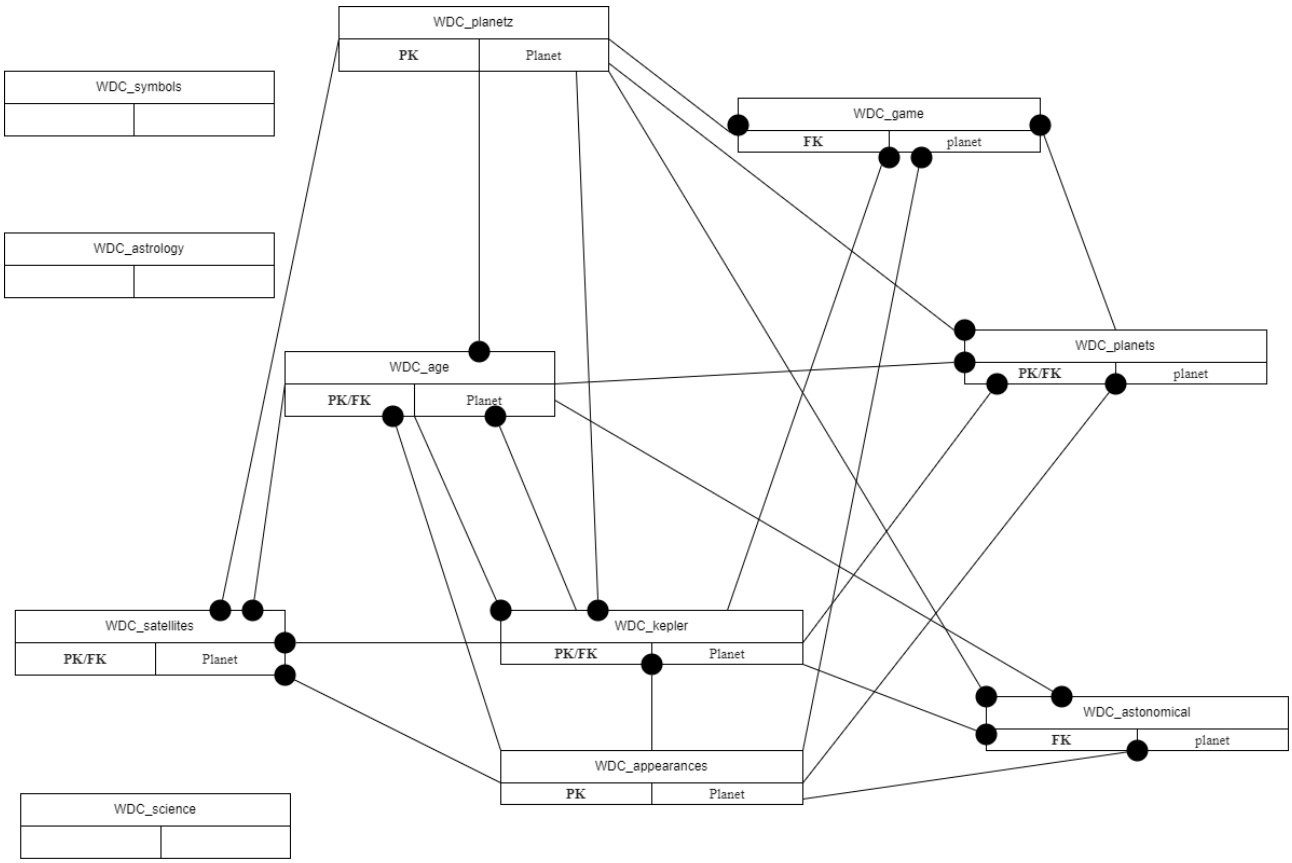

Рисунок 3.Связи PK-FK

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

Inclusion Dependency

Dependant	Referenced
[WDC_planets.csv.Name]	[WDC_symbols.csv.Symbol]
[WDC_planets.csv.Name]	[WDC_science.csv.Object]
[WDC_planets.csv.Name]	[WDC_planetz.csv.Planet]
[WDC_planets.csv.Name]	[WDC_kepler.csv.Planet]
[WDC_planets.csv.Name]	[WDC_game.csv.DivisionName]
[WDC_planets.csv.Name]	[WDC_astronomical.csv.Name]
[WDC_planets.csv.Name]	[WDC_appearances.csv.Planet]
[WDC_planets.csv.Name]	[WDC_age.csv.Planet]
[WDC_kepler.csv.Planet]	[WDC_symbols.csv.Symbol]
[WDC_kepler.csv.Planet]	[WDC_planetz.csv.Planet]

Рисунок 4

Inclusion Dependency

Dependant	Referenced
[WDC_kepler.csv.Planet]	[WDC_appearances.csv.Planet]
[WDC_kepler.csv.Planet]	[WDC_age.csv.Planet]
[WDC_game.csv.DivisionName]	[WDC_symbols.csv.Symbol]
[WDC_game.csv.DivisionName]	[WDC_science.csv.Object]
[WDC_game.csv.DivisionName]	[WDC_planetz.csv.Planet]
[WDC_game.csv.DivisionName]	[WDC_planets.csv.Name]
[WDC_game.csv.DivisionName]	[WDC_kepler.csv.Planet]
[WDC_game.csv.DivisionName]	[WDC_astronomical.csv.Name]
[WDC_game.csv.DivisionName]	[WDC_appearances.csv.Planet]
[WDC_game.csv.DivisionName]	[WDC_age.csv.Planet]

Рисунок 5

Inclusion Dependency

Dependant	Referenced
[WDC_game.csv.LowOrbitLeaderBoardPrizePool]	[WDC_game.csv.HighOrbitLeaderBoardPrizePool]
[WDC_game.csv.HighOrbitLeaderBoardPrizePool]	[WDC_game.csv.LowOrbitLeaderBoardPrizePool]
[WDC_astronomical.csv.Name]	[WDC_symbols.csv.Symbol]
[WDC_astronomical.csv.Name]	[WDC_science.csv.Object]
[WDC_astronomical.csv.Name]	[WDC_planetz.csv.Planet]
[WDC_astronomical.csv.Name]	[WDC_planets.csv.Name]
[WDC_astronomical.csv.Name]	[WDC_kepler.csv.Planet]
[WDC_astronomical.csv.Name]	[WDC_game.csv.DivisionName]
[WDC_astronomical.csv.Name]	[WDC_appearances.csv.Planet]
[WDC_astronomical.csv.Name]	[WDC_age.csv.Planet]

Рисунок 6

Inclusion Dependency

Dependant	Referenced
[WDC_astrology.csv.Planetary Joy]	[WDC_astrology.csv.Fall]
[WDC_age.csv.Planet]	[WDC_symbols.csv.Symbol]
[WDC_age.csv.Planet]	[WDC_planetz.csv.Planet]
[WDC_age.csv.Planet]	[WDC_kepler.csv.Planet]
[WDC_age.csv.Planet]	[WDC_appearances.csv.Planet]
[WDC_astrology.csv.Domicile, WDC_astrology.csv.Exaltation]	[WDC_astrology.csv.Detriment, WDC_astrology.csv.Fall]
[WDC_astrology.csv.Domicile, WDC_astrology.csv.Fall]	[WDC_astrology.csv.Detriment, WDC_astrology.csv.Exaltation]
[WDC_astrology.csv.Detriment, WDC_astrology.csv.Exaltation]	[WDC_astrology.csv.Domicile, WDC_astrology.csv.Fall]
[WDC_astrology.csv.Detriment, WDC_astrology.csv.Fall]	[WDC_astrology.csv.Domicile, WDC_astrology.csv.Exaltation]

Рисунок 7

Inclusion Dependency

Dependant	Referenced
[WDC_science.csv.Object]	[WDC_symbols.csv.Symbol]
[WDC_satellites.csv.Planet]	[WDC_symbols.csv.Symbol]
[WDC_satellites.csv.Planet]	[WDC_science.csv.Object]
[WDC_satellites.csv.Planet]	[WDC_planetz.csv.Planet]
[WDC_satellites.csv.Planet]	[WDC_planets.csv.Name]
[WDC_satellites.csv.Planet]	[WDC_kepler.csv.Planet]
[WDC_satellites.csv.Planet]	[WDC_game.csv.DivisionName]
[WDC_satellites.csv.Planet]	[WDC_astronomical.csv.Name]
[WDC_satellites.csv.Planet]	[WDC_appearances.csv.Planet]
[WDC_satellites.csv.Planet]	[WDC_age.csv.Planet]

Рисунок 8