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 №2
По курсу
«Оптимизация баз данных систем машинного обучения»
«Обнаружение функциональных зависимостей»
Вариант 4

Выполнил: Журавлев Н.В.

Группа: ИУ5-14М

Дата: 30.10.2023

Проверил:

Плужникова О. Ю.

2023 г.

Задание

Для каждого набора данных:

1) определите ключ с помощью алгоритма HyUCC;
2) сгенерируйте схему базы данных с таблицами в нормальной форме Бойса-Кодда с помощью алгоритма Normalize (автоматически сгенерированную схему БД можно увидеть в папке Metanome\results и в окне "Adding Context for backend");
3) представьте эту схему базы данных в нотации пакета ERwin;
4) преобразуйте схему базы данных в схему с синтетическими ключами;
5) предложите алгоритм заполнения таблиц базы данных (с синтетическими ключами) данными из вашего набора.

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)

Описание наборов данных

adolt.csv – набор данных из базы данных переписи 1994 года, который был выполнен Барри Беккером, чтобы спрогнозировать, превысит ли доход 50 тысяч долларов в год на основе данных переписи населения. Набор содержит:

1. age – возраст; значения - любые численные
2. workclass – доход; значения - Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked
3. fnlwgt - примерная оценка количества людей, которое представляет каждая строка данных; значения - любые численные
4. education – уровень образования; значения - Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool
5. education-num – длительность обучения; значения - любые численные

6. marital-status – семейное положение; значения - Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse

7. occupation – род деятельности; значения - Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces

8. relationship - отношения; значения - Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried

9. race - раса; значения - White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black

10. sex - пол; значения - Female, Male

11. capital-gain – прирост капитала; значения - любые численные

12. capital-loss – потеря капитала; значения - любые численные

13. hours-per-week – количество рабочих часов в неделю; значения - любые численные

14. native-country – родная страна; значения - United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands

15. income – доход; значения - >50K, <=50K.

Estate_valuation.csv – набор рыночных данных по оценке недвижимости, собранный в Синдианском округе, город Нью-Тайбэй, Тайвань. Набор содержит:

1. X1 transaction date – дата заключения сделки; значения - дата в формате год, прошедшая часть (например, 2013,250 = март 2013 г., 2013,500 = июнь 2013)

2. X2 house age – возраст дома; значения – любые численные

3. X3 distance to the nearest MRT – метров до ближайшей станции метро; значения - любые численные

4. X4 number of convenience stores - количество магазинов повседневного спроса в шаговой доступности; значения - любые численные
5. X5 latitude – широта, на которой расположен дом; значения - любые численные
6. X6 longitude – долгота, на которой расположен дом; значения - любые численные
7. Y house price of unit area - стоимость дома за единицу площади; значения - любые численные

Ход работы

Для adult.csv

Данные изначально находились в архиве с расширением data и с файлом описанием колонок в формате name.

Определение ключа с помощью алгоритма HyUCC

С помощью алгоритма HyUCC после выполнения получился результат на рисунке 1.

```
# TABLES
adult.csv       1
# COLUMN
1.column11      11
1.column10      10
1.column13      13
1.column12      12
1.column15      15
1.column14      14
1.column5       5
1.column6       6
1.column3       3
1.column4       4
1.column9       9
1.column7       7
1.column8       8
1.column1       1
1.column2       2
# RESULTS
1,11,13,14,15,2,3,4,7,8
1,11,13,14,15,2,3,5,7,8
```

Рисунок 1. Результат выполнения HyUCC

Генерация схему базы данных

С помощью алгоритма Normalize после выполнения получился результат на рисунке 2.

{"type":"BasicStatistic","columnCombination":{"columnIdentifiers":[{"tableIdentifier":"adult.csv","columnIdentifier":"column1"},{"tableIdent
mn1, adult.csv.column2, adult.csv.column3, adult.csv.column4, adult.csv.column7, adult.csv.column13"}}}
{"type":"BasicStatistic","columnCombination":{"columnIdentifiers":[{"tableIdentifier":"adult.csv","columnIdentifier":"column1"},{"tableIdent
{"type":"BasicStatistic","columnCombination":{"columnIdentifiers":[{"tableIdentifier":"adult.csv","columnIdentifier":"column1"},{"tableIdent
{"type":"BasicStatistic","columnCombination":{"columnIdentifiers":[{"tableIdentifier":"adult.csv","columnIdentifier":"column1"},{"tableIdent
{"type":"BasicStatistic","columnCombination":{"columnIdentifiers":[{"tableIdentifier":"adult.csv","columnIdentifier":"column4"},{"tableIdent

Рисунок 2. Результат выполнения Normalize

Схема базы данных в нотации пакета ERwin

Составим диаграмму полученной ранее схемы базы данных. На рисунке 3 показана получившаяся диаграмма.

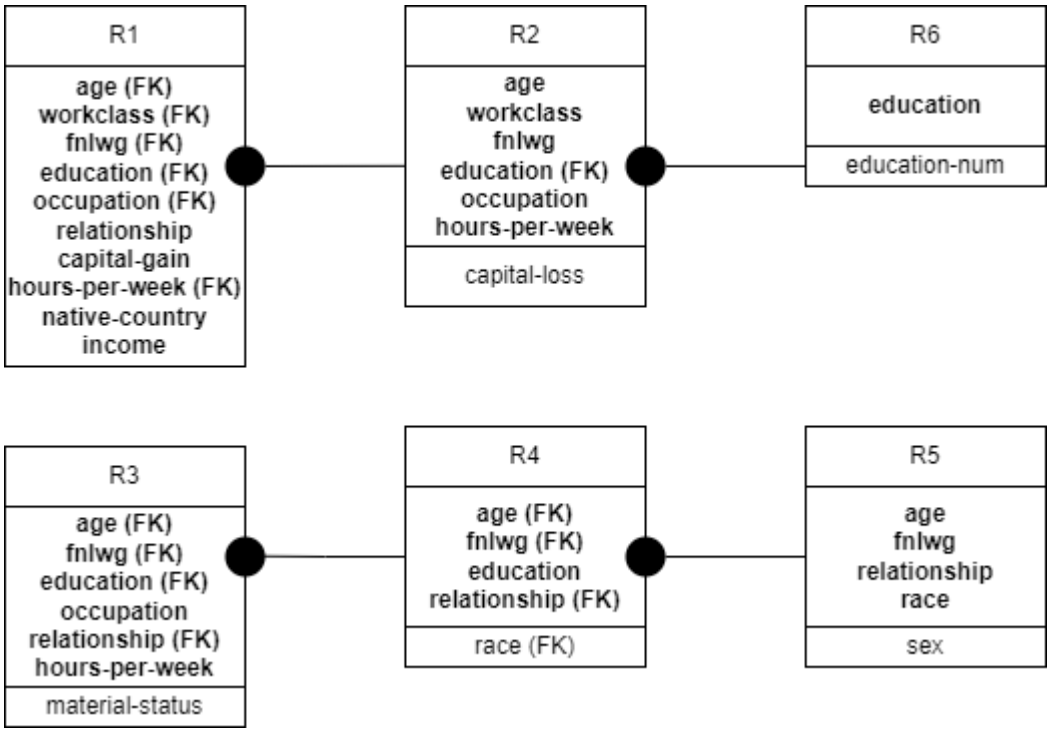

Рисунок 3. Схема базы данных в нотации пакета ERwin

Схема базы данных с синтетическими ключами

Добавим синтетические ключи в полученную схему базы данных. На рисунке 4 показана получившаяся диаграмма.

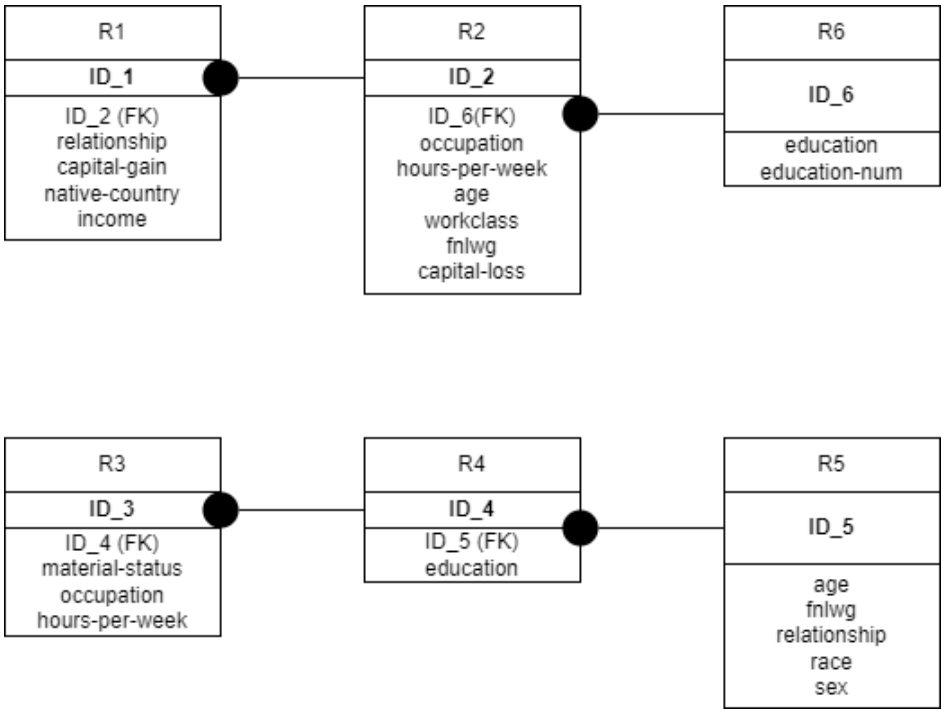

Рисунок 4. Схема базы данных с синтетическими ключами

Алгоритм заполнения таблиц базы данных

Пусть кортеж считывается в переменную s. А ID_1 = 0, ID_2 = 0, ID_3 = 0, ID_4 = 0, ID_5 = 0, ID_6 = 0 и сохраняют своё значение после выполнения для одного кортежа.

1. Если не R6∈(S[education]), то
2. R6.append(S[education], S[education-num], ID_6)
3. Fk_ID_6 = ID_6
4. ID_6 = ID_6 + 1
5. Иначе
6. Fk_ID_6 = R6[S[education], S[education-num]].ID_6
7.
8. Если не R2∈(S[age], S[workclass], S[fnlwg], S[occupation], S[hours-per-week], Fk_ID_6), то
9. R2.append(S[age], S[workclass], S[fnlwg], S[occupation], S[hours-per-week], S[capital-loss], ID_2, Fk_ID_6)
10. Fk_ID_2 = ID_2
11. ID_2 = ID_2 + 1
12. Иначе
13. Fk_ID_2 = R2[S[age], S[workclass], S[fnlwg], S[occupation], S[hours-per-week], S[capital-loss], Fk_ID_6].ID_2
14.
15. Если не R1∈(S[relationship], S[capital-gain], S[native-country], S[income], Fk_ID_2), то
16. R1.append(S[relationship], S[capital-gain], S[native-country], S[income], ID_1, Fk_ID_2)
17. ID_1 = ID_1 + 1
18.

19.Если не R5∈(S[age], S[fnlwg], S[relationship], S[race]), то
20. R5.append(S[age], S[fnlwg], S[relationship], S[race], S[sex], ID_5)
21. Fk_ID_5 = ID_5
22. ID_5 = ID_5 + 1
23.Иначе
24. Fk_ID_5 = R5[S[age], S[fnlwg], S[relationship], S[race], S[sex]].ID_5
25.
26.Если не R4∈(S[education], Fk_ID_5), то
27. R4.append(S[education], ID_4, Fk_ID_5)
28. Fk_ID_4 = ID_4
29. ID_4 = ID_4 + 1
30.Иначе
31. Fk_ID_4 = R4[S[education], Fk_ID_5].ID_4
32.
33.Если не R3∈(S[occupation], S[hours-per-week], Fk_ID_4), то
34. R3.append(S[material-status], S[occupation], S[hours-per-week], ID_3, Fk_ID_4)
35. ID_3 = ID_3 + 1

Для estate_valuation.csv

Данные изначально находились в архиве с расширением xlsx.

Определение ключа с помощью алгоритма HyUCC

С помощью алгоритма HyUCC после выполнения получился результат на рисунке 5.

```
# TABLES
estate_valuation.csv    1
# COLUMN
1.X3 distance to the nearest MRT station        3
1.X6 longitude  6
1.X4 number of convenience stores       4
1.Y house price of unit area    7
1.X1 transaction date   1
1.X5 latitude   5
1.X2 house age  2
# RESULTS
2,3,7
1,3,7
2,5,7
2,6,7
1,2,7
2,4,7
1,5,7
1,6,7
```

Рисунок 5. Результат выполнения HyUCC

Генерация схему базы данных

С помощью алгоритма Normalize после выполнения получился результат на рисунке 6.

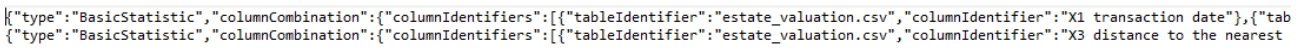

Рисунок 6. Результат выполнения Normalize

Схема базы данных в нотации пакета ERwin

Составим диаграмму полученной ранее схемы базы данных. На рисунке 7 показана получившаяся диаграмма.

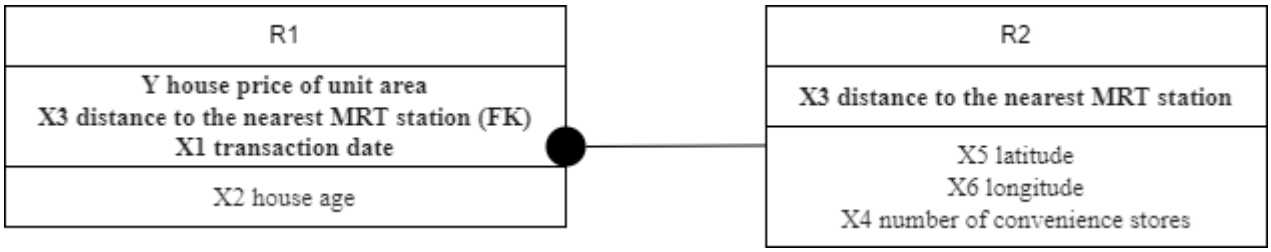

Рисунок 7. Схема базы данных в нотации пакета ERwin

Схема базы данных с синтетическими ключами

Добавим синтетические ключи в полученную схему базы данных. На рисунке 8 показана получившаяся диаграмма.

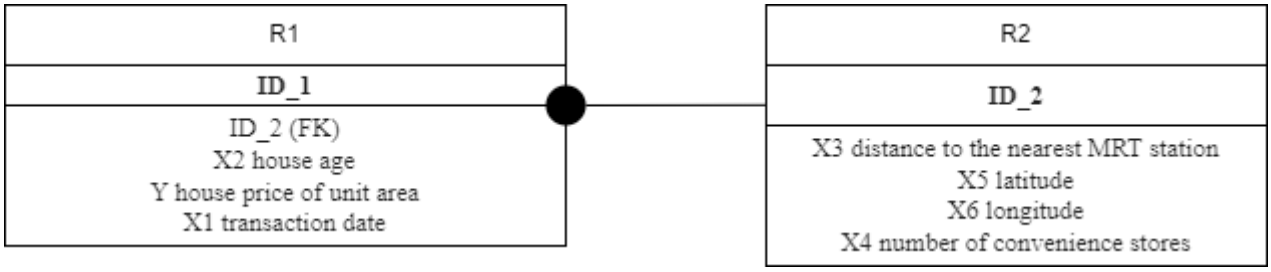

Рисунок 8. Схема базы данных с синтетическими ключами

Алгоритм заполнения таблиц базы данных

Пусть кортеж считывается в переменную s. А ID_1 = 0, ID_2 = 0 и сохраняют своё значение после выполнения для одного кортежа.

1. Если не R2∈(S[X3 distance to the nearest MRT station]), то
2. R2.append(S[X3 distance to the nearest MRT station], S[X5 latitude], S[X6 longitude], S[X4 number of convenience stores], ID_2)
3. Fk_ID_2 = ID_2
4. ID_2 = ID_2 + 1
5. Иначе
6. Fk_ID_2 = R2[S[X3 distance to the nearest MRT station], S[X5 latitude], S[X6 longitude], S[X4 number of convenience stores]].ID_2
7.

8. Если не R1∈(S[Y house price of unit area], S[X3 distance to the nearest MRT station], S[X1 transaction date], Fk_ID_2), то
9. R1.append(S[Y house price of unit area], S[X3 distance to the nearest MRT station], S[X1 transaction date], S[X2 house age], ID_1, Fk_ID_2)
10. ID_1 = ID_1 + 1