|  |  |
| --- | --- |
| 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 №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.

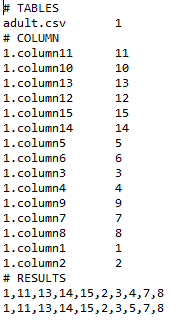


Рисунок 1. Результат выполнения HyUCC

### Генерация схему базы данных

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

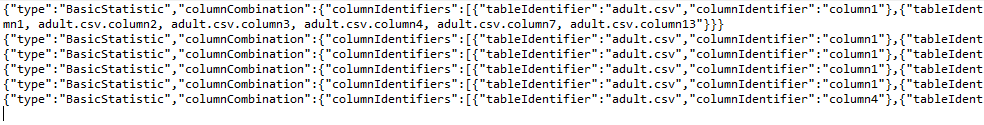


Рисунок 2. Результат выполнения Normalize

### Схема базы данных в нотации пакета ERwin

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

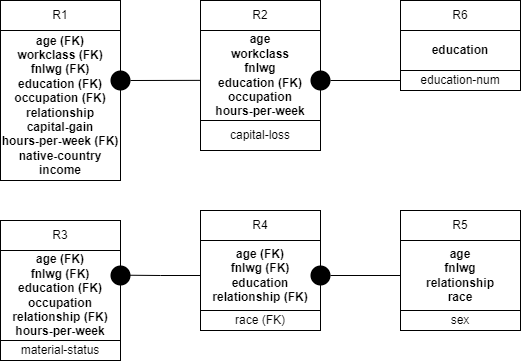


Рисунок 3. Схема базы данных в нотации пакета ERwin

### Схема базы данных с синтетическими ключами

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

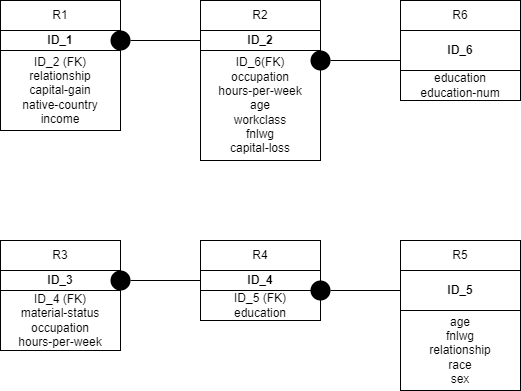


Рисунок 4. Схема базы данных с синтетическими ключами

### 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. Если не R2(S[age], S[workclass], S[fnlwg], S[occupation], S[hours-per-week], Fk\_ID\_6), то
8. R2.append(S[age], S[workclass], S[fnlwg], S[occupation], S[hours-per-week], S[capital-loss], ID\_2, Fk\_ID\_6)
9. Fk\_ID\_2 = ID\_2
10. ID\_2 = ID\_2 + 1
11. Иначе
12. Fk\_ID\_2 = R2[S[age], S[workclass], S[fnlwg], S[occupation], S[hours-per-week], S[capital-loss], Fk\_ID\_6].ID\_2
13. Если не R1(S[relationship], S[capital-gain], S[native-country], S[income], Fk\_ID\_2), то
14. R1.append(S[relationship], S[capital-gain], S[native-country], S[income], ID\_1, Fk\_ID\_2)
15. ID\_1 = ID\_1 + 1
16. Если не R5(S[age], S[fnlwg], S[relationship], S[race]), то
17. R5.append(S[age], S[fnlwg], S[relationship], S[race], S[sex], ID\_5)
18. Fk\_ID\_5 = ID\_5
19. ID\_5 = ID\_5 + 1
20. Иначе
21. Fk\_ID\_5 = R5[S[age], S[fnlwg], S[relationship], S[race], S[sex]].ID\_5
22. Если не R4(S[education], Fk\_ID\_5), то
23. R4.append(S[education], ID\_4, Fk\_ID\_5)
24. Fk\_ID\_4 = ID\_4
25. ID\_4 = ID\_4 + 1
26. Иначе
27. Fk\_ID\_4 = R4[S[education], Fk\_ID\_5].ID\_4
28. Если не R3(S[occupation], S[hours-per-week], Fk\_ID\_4), то
29. R3.append(S[material-status], S[occupation], S[hours-per-week], ID\_3, Fk\_ID\_4)
30. 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.

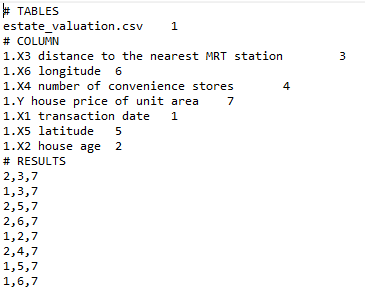


Рисунок 5. Результат выполнения HyUCC

### 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.

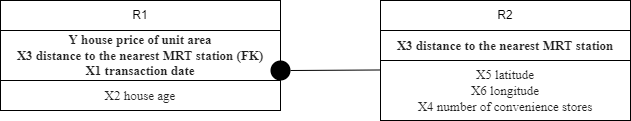


Рисунок 7. Схема базы данных в нотации пакета ERwin

### Схема базы данных с синтетическими ключами

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

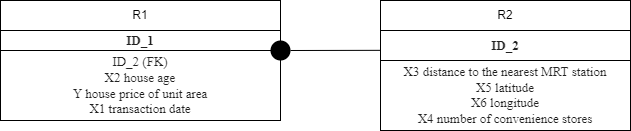


Рисунок 8. Схема базы данных с синтетическими ключами

### 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. 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
8. 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)
9. ID\_1 = ID\_1 + 1