



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

---

---

ФАКУЛЬТЕТ Информатика, системы управления и искусственный интеллект

КАФЕДРА Системы обработки информации и управления

**Домашнее задание №2**

**По курсу**

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

**Вариант №14**

Выполнил:

Студент группы ИУ5-12М

**Пермяков Д.К.**

03.11.2025

2025 г.

## Задание

Для заданного набора данных необходимо решить задачи:

- 1) определить закономерности в наборе данных iris и в наборе данных с опечатками (условные функциональные зависимости CFD);;
- 2) определить правила продукции "ЕСЛИ PR1 ТО PR2" в наборе данных iris и в наборе данных с опечатками (ограничения отказа, т.е. отрицания DC);
- 3) определить зависимости на основе анализа подобных кортежей (сопоставимые зависимости MD);
- 4) определить внешние ключи для соединения таблиц озера данных в одну таблицу перед машинным обучением (зависимости включения IND);

Таблица №1 Вариант студента

ИУ5-12М 2025		Используемые наборы данных	
8	Пермяков Д.К.	набор данных	набора данных с опечатками
Вариант - 14		iris.csv	Iris14.csv

## **Задача №1 – Условные функциональные зависимости (CFD).**

### **Результаты работы алгоритма CFD\_Finder на датасете iris.csv**

Таблица 1 – Результаты CFD\_Finder, iris.csv

<b>Determinant (детерминант)</b>	<b>Dependant (зависимый атрибут)</b>	<b>Pattern Tableau (шаблон)</b>
[iris.csv.column1, iris.csv.column2, iris.csv.column4, iris.csv.column5]	iris.csv.column3	(_,_,_,Iris-virginica)
[iris.csv.column1, iris.csv.column2, iris.csv.column3, iris.csv.column5]	iris.csv.column4	(_,_,_,Iris-versicolor)
[iris.csv.column2, iris.csv.column3, iris.csv.column4, iris.csv.column5]	iris.csv.column1	(_,_,_,Iris-virginica)
[iris.csv.column1, iris.csv.column3, iris.csv.column4, iris.csv.column5]	iris.csv.column2	(_,_,_,Iris-virginica)

Для записей с видом Iris-virginica знание трёх из четырёх числовых столбцов (column1–4) однозначно определяет четвёртый.

Для Iris-versicolor комбинация column1, column2, column3 задаёт column4. Это позволяет восстанавливать пропуски и обнаруживать ошибки в данных.

## Результаты работы алгоритма CFD\_Finder на датасете iris14.csv

Таблица 2 – Результаты алгоритма CFD\_Finder, iris14.csv

Determinant (дeterminант)	Dependant (зависимый атрибут)	Pattern Tableau (шаблон)
[iris14.csv.column1, iris14.csv.column3, iris14.csv.column4, iris14.csv.column5]	iris14.csv.column2	(_,_,_,Iris-virginica)
[iris14.csv.column1, iris14.csv.column2, iris14.csv.column4, iris14.csv.column5]	iris14.csv.column3	(_,_,_,Iris-virginica)
[iris14.csv.column2, iris14.csv.column3, iris14.csv.column4, iris14.csv.column5]	iris14.csv.column1	(_,_,_,Iris-virginica)
[iris14.csv.column1, iris14.csv.column2, iris14.csv.column3, iris14.csv.column5]	iris14.csv.column4	(_,_,_,Iris-versicolor);(_,_,_,Iris-setosa);(_,_,_,Iris-virginica)

Правила, связанные с Iris-virginica, сохранились без изменений.

Правило для столбца 4 нарушилось. Чтобы сохранить зависимость, пришлось включить все значения Species. Возможно, в строках Iris-versicolor появились расхождения (с1–с3, с5 прежние → с4 изменён), поэтому исходная CFD больше не выполняется.

## Задача №2. Ограничения отказа, т.е. отрицания (DC).

### Результаты алгоритма DC на данных iris.csv

представлены в приложении А, в Таблица 3 представлены DC с атрибутом column5 и преобразованные DC в правила продукции вида "ЕСЛИ PR1 ТО предикат для column5".

Таблица 3 – Результаты алгоритма DC, iris.csv

№	DC с атрибутом column5	Преобразованные DC
1	$\forall t_0 \in \text{iris.csv}, t_1 \in \text{iris.csv}:$ $\neg[t_0.\text{column1} = t_1.\text{column1} \wedge$ $t_0.\text{column2} < t_1.\text{column2} \wedge$ $t_0.\text{column5} = t_1.\text{column5}]$	ЕСЛИ column1( $t_0$ )=column1( $t_1$ ) И $\text{column2}(t_0) < \text{column2}(t_1)$ ТО $\text{column5}(t_0) \neq \text{column5}(t_1)$
2	$\forall t_0 \in \text{iris.csv}, t_1 \in \text{iris.csv}:$ $\neg[t_0.\text{column4} < t_1.\text{column4}$ $\wedge$ $t_0.\text{column2} < t_1.\text{column2}$ $\wedge$ $t_0.\text{column5} \neq t_1.\text{column5}]$	ЕСЛИ column4( $t_0$ )<column4( $t_1$ ) И $\text{column2}(t_0) < \text{column2}(t_1)$ ТО $\text{column5}(t_0) = \text{column5}(t_1)$
3	$\forall t_0 \in \text{iris.csv}, t_1 \in \text{iris.csv}:$ $\neg[t_0.\text{column3} \leq t_1.\text{column3}$ $\wedge$ $t_0.\text{column2} < t_1.\text{column2}$ $\wedge$ $t_0.\text{column5} \neq t_1.\text{column5}]$	ЕСЛИ column3( $t_0$ )≤column3( $t_1$ ) И $\text{column2}(t_0) < \text{column2}(t_1)$ ТО $\text{column5}(t_0) = \text{column5}(t_1)$
4	$\forall t_0 \in \text{iris.csv}, t_1 \in \text{iris.csv}:$ $\neg[t_0.\text{column1} = t_1.\text{column1}$ $\wedge$ $t_0.\text{column2} < t_1.\text{column2}$ $\wedge$ $t_0.\text{column5} \neq t_1.\text{column5}]$	ЕСЛИ column1( $t_0$ )=column1( $t_1$ ) И $\text{column2}(t_0) < \text{column2}(t_1)$ ТО $\text{column5}(t_0) = \text{column5}(t_1)$
5	$\forall t_0 \in \text{iris.csv}, t_1 \in \text{iris.csv}:$ $\neg[t_0.\text{column3} = t_1.\text{column3}$ $\wedge$ $t_0.\text{column5} \neq t_1.\text{column5}]$	ЕСЛИ column3( $t_0$ )=column3( $t_1$ ) ТО $\text{column5}(t_0) = \text{column5}(t_1)$
6	$\forall t_0 \in \text{iris.csv}, t_1 \in \text{iris.csv}:$ $\neg[t_0.\text{column1} = t_1.\text{column1}$ $\wedge$ $t_0.\text{column4} < t_1.\text{column4}$ $\wedge$ $t_0.\text{column5} = t_1.\text{column5}]$	ЕСЛИ column1( $t_0$ )=column1( $t_1$ ) И $\text{column4}(t_0) < \text{column4}(t_1)$ ТО $\text{column5}(t_0) \neq \text{column5}(t_1)$
7	$\forall t_0 \in \text{iris.csv}, t_1 \in \text{iris.csv}:$ $\neg[t_0.\text{column3} \leq t_1.\text{column3}$ $\wedge$ $t_0.\text{column2} = t_1.\text{column2}$ $\wedge$ $t_0.\text{column5} \neq t_1.\text{column5}]$	ЕСЛИ column3( $t_0$ )≤column3( $t_1$ ) И $\text{column2}(t_0) = \text{column2}(t_1)$ И $\text{column1}(t_0) \geq \text{column1}(t_1)$ ТО $\text{column5}(t_0) \neq \text{column5}(t_1)$

<b>№</b>	<b>DC с атрибутом column5</b>	<b>Преобразованные DC</b>
	$t_0.column1 \geq t_1.column1 \wedge t_0.column5 = t_1.column5]$	
8	$\forall t_0 \in \text{iris.csv}, t_1 \in \text{iris.csv}: \neg[t_0.column4 = t_1.column4 \wedge t_0.column5 \neq t_1.column5]$	ЕСЛИ $column4(t_0) = column4(t_1)$ ТО $column5(t_0) = column5(t_1)$
9	$\forall t_0 \in \text{iris.csv}, t_1 \in \text{iris.csv}: \neg[t_0.column1 = t_1.column1 \wedge t_0.column3 \leq t_1.column3 \wedge t_0.column5 \neq t_1.column5]$	ЕСЛИ $column1(t_0) = column1(t_1)$ И $column3(t_0) \leq column3(t_1)$ ТО $column5(t_0) = column5(t_1)$
10	$\forall t_0 \in \text{iris.csv}, t_1 \in \text{iris.csv}: \neg[t_0.column3 \leq t_1.column3 \wedge t_0.column4 \geq t_1.column4 \wedge t_0.column5 \neq t_1.column5]$	ЕСЛИ $column3(t_0) \leq column3(t_1)$ И $column4(t_0) \geq column4(t_1)$ ТО $column5(t_0) = column5(t_1)$
11	$\forall t_0 \in \text{iris.csv}, t_1 \in \text{iris.csv}: \neg[t_0.column1 = t_1.column1 \wedge t_0.column4 \leq t_1.column4 \wedge t_0.column5 \neq t_1.column5]$	ЕСЛИ $column1(t_0) = column1(t_1)$ И $column4(t_0) \leq column4(t_1)$ ТО $column5(t_0) = column5(t_1)$

**Результаты алгоритма DC на данных iris14.csv** представлены в приложении Б, в *Таблица 4* представлены DC с атрибутом column5 и преобразованные DC в правила продукции вида "ЕСЛИ PR1 ТО предикат для column5".

Таблица 4 – Результаты алгоритма DC, iris14.csv

<b>№</b>	<b>DC с атрибутом column5</b>	<b>Преобразованные DC</b>
1	$\forall t_0 \in \text{iris14.csv}, t_1 \in \text{iris14.csv}: \neg[t_0.column3 \geq t_1.column3 \wedge t_0.column5 = t_1.column5 \wedge t_0.column1 = t_1.column1]$	ЕСЛИ $column3(t_0) \geq column3(t_1)$ И $column1(t_0) = column1(t_1)$ ТО $column5(t_0) \neq column5(t_1)$
2	$\forall t_0 \in \text{iris14.csv}, t_1 \in \text{iris14.csv}: \neg[t_0.column3 \leq t_1.column3 \wedge t_0.column1 > t_1.column1 \wedge t_0.column5 = t_1.column5 \wedge t_0.column2 < t_1.column2]$	ЕСЛИ $column3(t_0) \leq column3(t_1)$ И $column1(t_0) > column1(t_1)$ И $column2(t_0) < column2(t_1)$ ТО $column5(t_0) \neq column5(t_1)$

<b>№</b>	<b>DC с атрибутом column5</b>	<b>Преобразованные DC</b>
3	$\forall t_0 \in \text{iris14.csv}, t_1 \in \text{iris14.csv}:$ $\neg[t_0.\text{column1} < t_1.\text{column1} \wedge$ $t_0.\text{column2} \geq t_1.\text{column2} \wedge$ $t_0.\text{column5} = t_1.\text{column5} \wedge$ $t_0.\text{column3} > t_1.\text{column3}]$	ЕСЛИ $\text{column1}(t_0) < \text{column1}(t_1)$ И $\text{column2}(t_0) \geq \text{column2}(t_1)$ И $\text{column3}(t_0) >$ $\text{column3}(t_1)$ TO $\text{column5}(t_0) \neq \text{column5}(t_1)$
4	$\forall t_0 \in \text{iris14.csv}, t_1 \in \text{iris14.csv}:$ $\neg[t_0.\text{column5} \neq t_1.\text{column5} \wedge$ $t_0.\text{column3} = t_1.\text{column3} \wedge$ $t_0.\text{column2} \leq t_1.\text{column2}]$	ЕСЛИ $\text{column3}(t_0) = \text{column3}(t_1)$ И $\text{column2}(t_0) \leq \text{column2}(t_1)$ TO $\text{column5}(t_0) = \text{column5}(t_1)$
5	$\forall t_0 \in \text{iris14.csv}, t_1 \in \text{iris14.csv}:$ $\neg[t_0.\text{column2} \geq t_1.\text{column2} \wedge$ $t_0.\text{column5} = t_1.\text{column5} \wedge$ $t_0.\text{column1} = t_1.\text{column1}]$	ЕСЛИ $\text{column2}(t_0) \geq \text{column2}(t_1)$ И $\text{column1}(t_0) = \text{column1}(t_1)$ TO $\text{column5}(t_0) \neq \text{column5}(t_1)$
6	$\forall t_0 \in \text{iris14.csv}, t_1 \in \text{iris14.csv}:$ $\neg[t_0.\text{column2} = t_1.\text{column2} \wedge$ $t_0.\text{column4} < t_1.\text{column4} \wedge$ $t_0.\text{column5} = t_1.\text{column5}]$	ЕСЛИ $\text{column2}(t_0) = \text{column2}(t_1)$ И $\text{column4}(t_0) < \text{column4}(t_1)$ TO $\text{column5}(t_0) \neq \text{column5}(t_1)$
7	$\forall t_0 \in \text{iris14.csv}, t_1 \in \text{iris14.csv}:$ $\neg[t_0.\text{column5} = t_1.\text{column5} \wedge$ $t_0.\text{column3} < t_1.\text{column3} \wedge$ $t_0.\text{column4} > t_1.\text{column4} \wedge$ $t_0.\text{column2} < t_1.\text{column2}]$	ЕСЛИ $\text{column3}(t_0) < \text{column3}(t_1)$ И $\text{column4}(t_0) > \text{column4}(t_1)$ И $\text{column2}(t_0) <$ $\text{column2}(t_1)$ TO $\text{column5}(t_0) \neq \text{column5}(t_1)$
8	$\forall t_0 \in \text{iris14.csv}, t_1 \in \text{iris14.csv}:$ $\neg[t_0.\text{column5} \neq t_1.\text{column5} \wedge$ $t_0.\text{column4} = t_1.\text{column4} \wedge$ $t_0.\text{column3} < t_1.\text{column3} \wedge$ $t_0.\text{column1} \geq t_1.\text{column1}]$	ЕСЛИ $\text{column4}(t_0) = \text{column4}(t_1)$ И $\text{column3}(t_0) < \text{column3}(t_1)$ И $\text{column1}(t_0) \geq$ $\text{column1}(t_1)$ TO $\text{column5}(t_0) = \text{column5}(t_1)$
9	$\forall t_0 \in \text{iris14.csv}, t_1 \in \text{iris14.csv}:$ $\neg[t_0.\text{column4} < t_1.\text{column4} \wedge$ $t_0.\text{column5} = t_1.\text{column5} \wedge$ $t_0.\text{column1} \leq t_1.\text{column1} \wedge$ $t_0.\text{column3} > t_1.\text{column3}]$	ЕСЛИ $\text{column4}(t_0) < \text{column4}(t_1)$ И $\text{column1}(t_0) \leq \text{column1}(t_1)$ И $\text{column3}(t_0) >$ $\text{column3}(t_1)$ TO $\text{column5}(t_0) \neq \text{column5}(t_1)$
10	$\forall t_0 \in \text{iris14.csv}, t_1 \in \text{iris14.csv}:$ $\neg[t_0.\text{column1} < t_1.\text{column1} \wedge$ $t_0.\text{column5} \neq t_1.\text{column5}]$	ЕСЛИ $\text{column1}(t_0) < \text{column1}(t_1)$ И $\text{column3}(t_0) < \text{column3}(t_1)$ И $\text{column2}(t_0) <$

<b>№</b>	<b>DC с атрибутом column5</b>	<b>Преобразованные DC</b>
	$t_0.\text{column3} < t_1.\text{column3} \wedge t_0.\text{column2} < t_1.\text{column2}]$	$\text{column2}(t_1)$ TO $\text{column5}(t_0) = \text{column5}(t_1)$
11	$\forall t_0 \in \text{iris14.csv}, t_1 \in \text{iris14.csv}: \neg[t_0.\text{column5} = t_1.\text{column5} \wedge t_0.\text{column4} > t_1.\text{column4} \wedge t_0.\text{column1} \geq t_1.\text{column1} \wedge t_0.\text{column2} < t_1.\text{column2}]$	ЕСЛИ $\text{column4}(t_0) > \text{column4}(t_1)$ И $\text{column1}(t_0) \geq \text{column1}(t_1)$ И $\text{column2}(t_0) < \text{column2}(t_1)$ TO $\text{column5}(t_0) \neq \text{column5}(t_1)$
12	$\forall t_0 \in \text{iris14.csv}, t_1 \in \text{iris14.csv}: \neg[t_0.\text{column5} \neq t_1.\text{column5} \wedge t_0.\text{column4} = t_1.\text{column4} \wedge t_0.\text{column3} < t_1.\text{column3} \wedge t_0.\text{column2} < t_1.\text{column2}]$	ЕСЛИ $\text{column4}(t_0) = \text{column4}(t_1)$ И $\text{column3}(t_0) < \text{column3}(t_1)$ И $\text{column2}(t_0) < \text{column2}(t_1)$ TO $\text{column5}(t_0) = \text{column5}(t_1)$
13	$\forall t_0 \in \text{iris14.csv}, t_1 \in \text{iris14.csv}: \neg[t_0.\text{column1} > t_1.\text{column1} \wedge t_0.\text{column4} < t_1.\text{column4} \wedge t_0.\text{column5} = t_1.\text{column5} \wedge t_0.\text{column3} < t_1.\text{column3}]$	ЕСЛИ $\text{column1}(t_0) > \text{column1}(t_1)$ И $\text{column4}(t_0) < \text{column4}(t_1)$ И $\text{column3}(t_0) < \text{column3}(t_1)$ TO $\text{column5}(t_0) \neq \text{column5}(t_1)$
14	$\forall t_0 \in \text{iris14.csv}, t_1 \in \text{iris14.csv}: \neg[t_0.\text{column5} \neq t_1.\text{column5} \wedge t_0.\text{column3} = t_1.\text{column3} \wedge t_0.\text{column4} < t_1.\text{column4}]$	ЕСЛИ $\text{column3}(t_0) = \text{column3}(t_1)$ И $\text{column4}(t_0) < \text{column4}(t_1)$ TO $\text{column5}(t_0) = \text{column5}(t_1)$
15	$\forall t_0 \in \text{iris14.csv}, t_1 \in \text{iris14.csv}: \neg[t_0.\text{column5} \neq t_1.\text{column5} \wedge t_0.\text{column4} < t_1.\text{column4} \wedge t_0.\text{column1} \leq t_1.\text{column1} \wedge t_0.\text{column2} < t_1.\text{column2}]$	ЕСЛИ $\text{column4}(t_0) < \text{column4}(t_1)$ И $\text{column1}(t_0) \leq \text{column1}(t_1)$ И $\text{column2}(t_0) < \text{column2}(t_1)$ TO $\text{column5}(t_0) = \text{column5}(t_1)$
16	$\forall t_0 \in \text{iris14.csv}, t_1 \in \text{iris14.csv}: \neg[t_0.\text{column5} \neq t_1.\text{column5} \wedge t_0.\text{column2} \geq t_1.\text{column2} \wedge t_0.\text{column4} = t_1.\text{column4} \wedge t_0.\text{column1} \leq t_1.\text{column1}]$	ЕСЛИ $\text{column2}(t_0) \geq \text{column2}(t_1)$ И $\text{column4}(t_0) = \text{column4}(t_1)$ И $\text{column1}(t_0) \leq \text{column1}(t_1)$ TO $\text{column5}(t_0) = \text{column5}(t_1)$
17	$\forall t_0 \in \text{iris14.csv}, t_1 \in \text{iris14.csv}: \neg[t_0.\text{column5} \neq t_1.\text{column5} \wedge t_0.\text{column3} = t_1.\text{column3} \wedge t_0.\text{column1} \leq t_1.\text{column1}]$	ЕСЛИ $\text{column3}(t_0) = \text{column3}(t_1)$ И $\text{column1}(t_0) \leq \text{column1}(t_1)$ TO $\text{column5}(t_0) = \text{column5}(t_1)$

<b>№</b>	<b>DC с атрибутом column5</b>	<b>Преобразованные DC</b>
18	$\forall t_0 \in \text{iris14.csv}, t_1 \in \text{iris14.csv}:$ $\neg[t_0.\text{column2} \geq t_1.\text{column2} \wedge$ $t_0.\text{column3} = t_1.\text{column3} \wedge$ $t_0.\text{column5} = t_1.\text{column5}]$	ЕСЛИ $\text{column2}(t_0) \geq \text{column2}(t_1)$ И $\text{column3}(t_0) = \text{column3}(t_1)$ TO $\text{column5}(t_0) \neq \text{column5}(t_1)$
19	$\forall t_0 \in \text{iris14.csv}, t_1 \in \text{iris14.csv}:$ $\neg[t_0.\text{column2} = t_1.\text{column2} \wedge$ $t_0.\text{column5} = t_1.\text{column5} \wedge$ $t_0.\text{column3} > t_1.\text{column3}]$	ЕСЛИ $\text{column2}(t_0) = \text{column2}(t_1)$ И $\text{column3}(t_0) > \text{column3}(t_1)$ TO $\text{column5}(t_0) \neq \text{column5}(t_1)$
20	$\forall t_0 \in \text{iris14.csv}, t_1 \in \text{iris14.csv}:$ $\neg[t_0.\text{column3} = t_1.\text{column3} \wedge$ $t_0.\text{column5} = t_1.\text{column5} \wedge$ $t_0.\text{column1} \geq t_1.\text{column1}]$	ЕСЛИ $\text{column3}(t_0) = \text{column3}(t_1)$ И $\text{column1}(t_0) \geq \text{column1}(t_1)$ TO $\text{column5}(t_0) \neq \text{column5}(t_1)$
21	$\forall t_0 \in \text{iris14.csv}, t_1 \in \text{iris14.csv}:$ $\neg[t_0.\text{column2} = t_1.\text{column2} \wedge$ $t_0.\text{column5} = t_1.\text{column5} \wedge$ $t_0.\text{column1} \geq t_1.\text{column1}]$	ЕСЛИ $\text{column2}(t_0) = \text{column2}(t_1)$ И $\text{column1}(t_0) \geq \text{column1}(t_1)$ TO $\text{column5}(t_0) \neq \text{column5}(t_1)$
22	$\forall t_0 \in \text{iris14.csv}, t_1 \in \text{iris14.csv}:$ $\neg[t_0.\text{column5} = t_1.\text{column5} \wedge$ $t_0.\text{column3} < t_1.\text{column3} \wedge$ $t_0.\text{column1} \geq t_1.\text{column1} \wedge$ $t_0.\text{column2} < t_1.\text{column2}]$	ЕСЛИ $\text{column3}(t_0) < \text{column3}(t_1)$ И $\text{column1}(t_0) \geq \text{column1}(t_1)$ И $\text{column2}(t_0) < \text{column2}(t_1)$ TO $\text{column5}(t_0) \neq \text{column5}(t_1)$
23	$\forall t_0 \in \text{iris14.csv}, t_1 \in \text{iris14.csv}:$ $\neg[t_0.\text{column4} < t_1.\text{column4} \wedge$ $t_0.\text{column5} = t_1.\text{column5} \wedge$ $t_0.\text{column1} = t_1.\text{column1}]$	ЕСЛИ $\text{column4}(t_0) < \text{column4}(t_1)$ И $\text{column1}(t_0) = \text{column1}(t_1)$ TO $\text{column5}(t_0) \neq \text{column5}(t_1)$

В «чистом» iris.csv нашлось 11 зависимостей: короткие условия и понятные межвидовые выводы (равенство / неравенство column5).

В iris14.csv из-за опечаток правил стало 23; большинство выводит « $\text{column5} \neq \dots$ », условия длиннее, а межвидовые зависимости почти исчезли. Алгоритм сгенерировал много внутренних запретов внутри одного вида и несколько пар взаимоисключающих правил. Это прямой симптом ошибок:

настоящие закономерности размыты, их место заняли ложные связи, указывающие на загрязнённые записи.

### **Задача №3. Сопоставимые зависимости МД.**

#### **Результаты алгоритма НуМД на данных iris.csv**

Таблица 5 – Результаты алгоритма НуМД, iris.csv

<b>Determinant</b>	<b>Dependant</b>	<b>Зависимости (FD)</b>
[iris.csv.column1(Levenshtein@1), iris.csv.column2(Levenshtein@1), iris.csv.column3(Levenshtein@1)]	iris.csv.column5(Levenshtein@1)	[column1, column2, column3] → column5
[iris.csv.column1(Levenshtein@1), iris.csv.column2(Levenshtein@1), iris.csv.column4(Levenshtein@1)]	iris.csv.column5(Levenshtein@1)	[column1, column2, column4] → column5
[iris.csv.column1(Levenshtein@1), iris.csv.column3(Levenshtein@1), iris.csv.column4(Levenshtein@1)]	iris.csv.column5(Levenshtein@1)	[column1, column3, column4] → column5
[iris.csv.column2(Levenshtein@1), iris.csv.column3(Levenshtein@1), iris.csv.column4(Levenshtein@1)]	iris.csv.column5(Levenshtein@1)	[column2, column3, column4] → column5

Интерпретируя данные результаты, можно сказать, что любая комбинация из трёх числовых признаков однозначно определяет вид ириса, при условии, что значения признаков у двух объектов достаточно близки (в рамках заданной меры сходства).

## Результаты алгоритма HyMD на данных iris14.csv

Таблица 6 – Результаты алгоритма HyMD, iris14.csv

Determinant	Dependant	Зависимость (MD)
[iris14.csv.column1(Levenshtein@1), iris14.csv.column2(Levenshtein@1), iris14.csv.column4(Levenshtein@1)]	iris14.csv.column5(Levenshtein@0.72727272727273)	[column1, column2, column4] → column5
[iris14.csv.column1(Levenshtein@1), iris14.csv.column3(Levenshtein@1), iris14.csv.column4(Levenshtein@1)]	iris14.csv.column5(Levenshtein@0.727272727273)	[column1, column3, column4] → column5
[iris14.csv.column2(Levenshtein@1), iris14.csv.column3(Levenshtein@1), iris14.csv.column4(Levenshtein@1)]	iris14.csv.column5(Levenshtein@0.727272727273)	[column2, column3, column4] → column5

В «чистом» iris.csv расстояние Левенштейна  $\leq 1$  между значениями любых трёх числовых столбцов гарантирует полное совпадение названия вида (column5) при том же пороге 1. Это фактически строгая функциональная зависимость: три числовых признака безошибочно определяют вид.

В iris14.csv опечатки разрушили эту зависимость.

- column4 включен в детерминант
- сходство для column5 = 0,73, то есть считать равными названия, отличающиеся примерно на четверть символов.

Коэффициент подобия рассчитывается как  $1 - (\text{редакционное расстояние} / \text{длина длиннейшей строки})$ . Значение 0,73 означает, что исходная функциональная зависимость нарушена: чтобы сохранить её, пришлось разрешить до 27 % отличий. Исправив искажения в названиях видов, можно вернуть точные ( $r = 1$ ) зависимости, присутствующие в исходном датасете.

#### **Задача №4. Зависимости включения (IND).**

*Таблица 7 – Результаты алгоритма FAIDA с PK Planet*

<b>Dependant (FK)</b>	<b>Referenced (PK)</b>
[WDC_astronomical.csv.Name]	[WDC_age.csv.Planet]
[WDC_game.csv.DivisionName]	[WDC_age.csv.Planet]
[WDC_kepler.csv.Planet]	[WDC_age.csv.Planet]
[WDC_planets.csv.Name]	[WDC_age.csv.Planet]
[WDC_satellites.csv.Planet]	[WDC_age.csv.Planet]
[WDC_age.csv.Planet]	[WDC_appearances.csv.Planet]
[WDC_astronomical.csv.Name]	[WDC_appearances.csv.Planet]
[WDC_game.csv.DivisionName]	[WDC_appearances.csv.Planet]
[WDC_kepler.csv.Planet]	[WDC_appearances.csv.Planet]
[WDC_planets.csv.Name]	[WDC_appearances.csv.Planet]
[WDC_satellites.csv.Planet]	[WDC_appearances.csv.Planet]
[WDC_age.csv.Planet]	[WDC_kepler.csv.Planet]
[WDC_astronomical.csv.Name]	[WDC_kepler.csv.Planet]
[WDC_game.csv.DivisionName]	[WDC_kepler.csv.Planet]
[WDC_planets.csv.Name]	[WDC_kepler.csv.Planet]
[WDC_satellites.csv.Planet]	[WDC_kepler.csv.Planet]
[WDC_age.csv.Planet]	[WDC_planetz.csv.Planet]
[WDC_astronomical.csv.Name]	[WDC_planetz.csv.Planet]
[WDC_game.csv.DivisionName]	[WDC_planetz.csv.Planet]
[WDC_kepler.csv.Planet]	[WDC_planetz.csv.Planet]
[WDC_planets.csv.Name]	[WDC_planetz.csv.Planet]
[WDC_satellites.csv.Planet]	[WDC_planetz.csv.Planet]

Схема в нотации Erwin для первичного ключа представлена на рисунке №1.

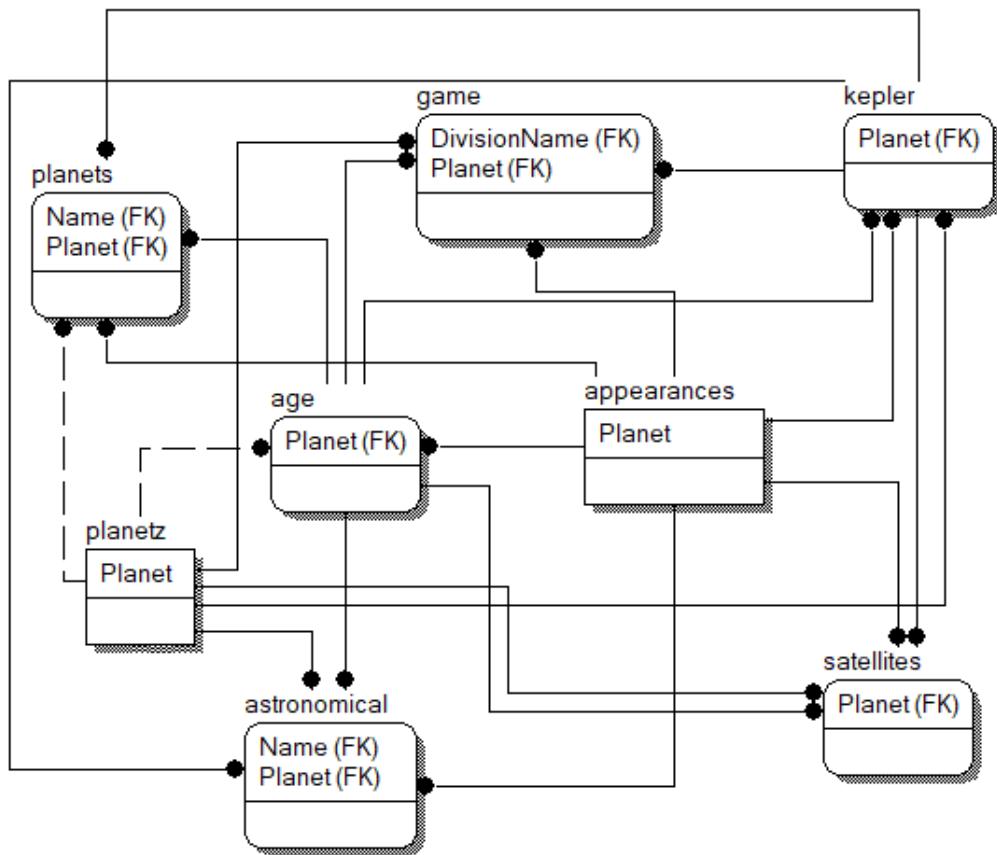


Рисунок №1 – Схема в нотации ERwin для первичного ключа Planet

## Приложение А

Predicates	Size
$\forall t_0 \in \text{iris.csv}, t_1 \in \text{iris.csv}:$ $\neg[t_0.\text{column1} = t_1.\text{column1} \wedge$ $t_0.\text{column4} < t_1.\text{column4} \wedge$ $t_0.\text{column2} < t_1.\text{column2}]$	3
$\forall t_0 \in \text{iris.csv}, t_1 \in \text{iris.csv}:$ $\neg[t_0.\text{column4} = t_1.\text{column4} \wedge$ $t_0.\text{column1} < t_1.\text{column1} \wedge$ $t_0.\text{column3} > t_1.\text{column3}]$	3
$\forall t_0 \in \text{iris.csv}, t_1 \in \text{iris.csv}:$ $\neg[t_0.\text{column3} \leq t_1.\text{column3} \wedge$ $t_0.\text{column2} = t_1.\text{column2} \wedge$ $t_0.\text{column4} \geq t_1.\text{column4}]$	3
$\forall t_0 \in \text{iris.csv}, t_1 \in \text{iris.csv}:$ $\neg[t_0.\text{column3} = t_1.\text{column3} \wedge$ $t_0.\text{column4} = t_1.\text{column4}]$	2
$\forall t_0 \in \text{iris.csv}, t_1 \in \text{iris.csv}:$ $\neg[t_0.\text{column4} < t_1.\text{column4} \wedge$ $t_0.\text{column1} < t_1.\text{column1} \wedge$ $t_0.\text{column3} > t_1.\text{column3}]$	3
$\forall t_0 \in \text{iris.csv}, t_1 \in \text{iris.csv}:$ $\neg[t_0.\text{column1} = t_1.\text{column1} \wedge$ $t_0.\text{column3} \leq t_1.\text{column3} \wedge$ $t_0.\text{column2} < t_1.\text{column2}]$	3
$\forall t_0 \in \text{iris.csv}, t_1 \in \text{iris.csv}:$ $\neg[t_0.\text{column1} = t_1.\text{column1} \wedge$ $t_0.\text{column2} < t_1.\text{column2} \wedge$ $t_0.\text{column5} = t_1.\text{column5}]$	3
$\forall t_0 \in \text{iris.csv}, t_1 \in \text{iris.csv}:$ $\neg[t_0.\text{column5} \neq t_1.\text{column5} \wedge$ $t_0.\text{column4} < t_1.\text{column4}]$	3

t0.column2< t1.column2]	
$\forall t0 \in \text{iris.csv}, t1 \in \text{iris.csv}:$ $\neg[t0.\text{column2} = t1.\text{column2} \wedge t0.\text{column4} = t1.\text{column4}]$	2
$\forall t0 \in \text{iris.csv}, t1 \in \text{iris.csv}:$ $\neg[t0.\text{column5} \neq t1.\text{column5} \wedge t0.\text{column3} \leq t1.\text{column3} \wedge t0.\text{column2} < t1.\text{column2}]$	3
$\forall t0 \in \text{iris.csv}, t1 \in \text{iris.csv}:$ $\neg[t0.\text{column3} = t1.\text{column3} \wedge t0.\text{column2} = t1.\text{column2}]$	2
$\forall t0 \in \text{iris.csv}, t1 \in \text{iris.csv}:$ $\neg[t0.\text{column5} \neq t1.\text{column5} \wedge t0.\text{column1} = t1.\text{column1} \wedge t0.\text{column2} < t1.\text{column2}]$	3
$\forall t0 \in \text{iris.csv}, t1 \in \text{iris.csv}:$ $\neg[t0.\text{column3} \geq t1.\text{column3} \wedge t0.\text{column2} = t1.\text{column2} \wedge t0.\text{column1} < t1.\text{column1}]$	3
$\forall t0 \in \text{iris.csv}, t1 \in \text{iris.csv}:$ $\neg[t0.\text{column1} = t1.\text{column1} \wedge t0.\text{column2} = t1.\text{column2}]$	2
$\forall t0 \in \text{iris.csv}, t1 \in \text{iris.csv}:$ $\neg[t0.\text{column1} > t1.\text{column1} \wedge t0.\text{column4} = t1.\text{column4} \wedge t0.\text{column2} < t1.\text{column2}]$	3
$\forall t0 \in \text{iris.csv}, t1 \in \text{iris.csv}:$ $\neg[t0.\text{column1} = t1.\text{column1} \wedge t0.\text{column4} > t1.\text{column4} \wedge t0.\text{column2} < t1.\text{column2}]$	3
$\forall t0 \in \text{iris.csv}, t1 \in \text{iris.csv}:$ $\neg[t0.\text{column1} = t1.\text{column1}]$	3

$t_0.\text{column2} < t_1.\text{column2} \wedge$ $t_0.\text{column3} > t_1.\text{column3}]$	
$\forall t_0 \in \text{iris.csv}, t_1 \in \text{iris.csv}:$ $\neg [t_0.\text{column5} \neq t_1.\text{column5} \wedge$ $t_0.\text{column3} = t_1.\text{column3}]$	2
$\forall t_0 \in \text{iris.csv}, t_1 \in \text{iris.csv}:$ $\neg [t_0.\text{column1} = t_1.\text{column1} \wedge$ $t_0.\text{column3} \leq t_1.\text{column3} \wedge$ $t_0.\text{column4} \geq t_1.\text{column4}]$	3
$\forall t_0 \in \text{iris.csv}, t_1 \in \text{iris.csv}:$ $\neg [t_0.\text{column1} = t_1.\text{column1} \wedge$ $t_0.\text{column4} < t_1.\text{column4} \wedge$ $t_0.\text{column5} = t_1.\text{column5}]$	3
$\forall t_0 \in \text{iris.csv}, t_1 \in \text{iris.csv}:$ $\neg [t_0.\text{column2} \geq t_1.\text{column2} \wedge$ $t_0.\text{column3} = t_1.\text{column3} \wedge$ $t_0.\text{column4} < t_1.\text{column4}]$	3
$\forall t_0 \in \text{iris.csv}, t_1 \in \text{iris.csv}:$ $\neg [t_0.\text{column3} \leq t_1.\text{column3} \wedge$ $t_0.\text{column2} = t_1.\text{column2} \wedge$ $t_0.\text{column1} \geq t_1.\text{column1} \wedge$ $t_0.\text{column5} = t_1.\text{column5}]$	4
$\forall t_0 \in \text{iris.csv}, t_1 \in \text{iris.csv}:$ $\neg [t_0.\text{column5} \neq t_1.\text{column5} \wedge$ $t_0.\text{column4} = t_1.\text{column4}]$	2
$\forall t_0 \in \text{iris.csv}, t_1 \in \text{iris.csv}:$ $\neg [t_0.\text{column1} = t_1.\text{column1} \wedge$ $t_0.\text{column4} = t_1.\text{column4}]$	2
$\forall t_0 \in \text{iris.csv}, t_1 \in \text{iris.csv}:$ $\neg [t_0.\text{column5} \neq t_1.\text{column5} \wedge$ $t_0.\text{column1} = t_1.\text{column1} \wedge$ $t_0.\text{column3} \leq t_1.\text{column3}]$	3

$\forall t_0 \in \text{iris.csv}, t_1 \in \text{iris.csv}:$ $\neg[t_0.\text{column5} \neq t_1.\text{column5} \wedge$ $t_0.\text{column3} \leq t_1.\text{column3} \wedge$ $t_0.\text{column4} \geq t_1.\text{column4}]$	3
$\forall t_0 \in \text{iris.csv}, t_1 \in \text{iris.csv}:$ $\neg[t_0.\text{column2} \geq t_1.\text{column2} \wedge$ $t_0.\text{column3} = t_1.\text{column3} \wedge$ $t_0.\text{column1} \leq t_1.\text{column1}]$	3
$\forall t_0 \in \text{iris.csv}, t_1 \in \text{iris.csv}:$ $\neg[t_0.\text{column1} = t_1.\text{column1} \wedge$ $t_0.\text{column3} = t_1.\text{column3}]$	2
$\forall t_0 \in \text{iris.csv}, t_1 \in \text{iris.csv}:$ $\neg[t_0.\text{column4} \leq t_1.\text{column4} \wedge$ $t_0.\text{column5} \neq t_1.\text{column5} \wedge$ $t_0.\text{column1} = t_1.\text{column1}]$	3

## Приложение Б

Predicates	Size
$\forall t_0 \in \text{iris14.csv}, t_1 \in \text{iris14.csv}:$ $\neg[t_0.\text{column2} = t_1.\text{column2} \wedge$ $t_0.\text{column1} = t_1.\text{column1}]$	2
$\forall t_0 \in \text{iris14.csv}, t_1 \in \text{iris14.csv}:$ $\neg[t_0.\text{column3} \geq t_1.\text{column3} \wedge$ $t_0.\text{column5} = t_1.\text{column5} \wedge$ $t_0.\text{column1} = t_1.\text{column1}]$	3
$\forall t_0 \in \text{iris14.csv}, t_1 \in \text{iris14.csv}:$ $\neg[t_0.\text{column3} \leq t_1.\text{column3} \wedge$ $t_0.\text{column1} > t_1.\text{column1} \wedge$ $t_0.\text{column5} = t_1.\text{column5} \wedge$ $t_0.\text{column2} < t_1.\text{column2}]$	4
$\forall t_0 \in \text{iris14.csv}, t_1 \in \text{iris14.csv}:$ $\neg[t_0.\text{column1} < t_1.\text{column1} \wedge$ $t_0.\text{column2} \geq t_1.\text{column2} \wedge$ $t_0.\text{column5} = t_1.\text{column5} \wedge$ $t_0.\text{column3} > t_1.\text{column3}]$	4
$\forall t_0 \in \text{iris14.csv}, t_1 \in \text{iris14.csv}:$ $\neg[t_0.\text{column5} \neq t_1.\text{column5} \wedge$ $t_0.\text{column3} = t_1.\text{column3} \wedge$ $t_0.\text{column2} \leq t_1.\text{column2}]$	3
$\forall t_0 \in \text{iris14.csv}, t_1 \in \text{iris14.csv}:$ $\neg[t_0.\text{column2} \geq t_1.\text{column2} \wedge$ $t_0.\text{column5} = t_1.\text{column5} \wedge$ $t_0.\text{column1} = t_1.\text{column1}]$	3
$\forall t_0 \in \text{iris14.csv}, t_1 \in \text{iris14.csv}:$ $\neg[t_0.\text{column2} \geq t_1.\text{column2} \wedge$ $t_0.\text{column3} = t_1.\text{column3} \wedge$ $t_0.\text{column4} < t_1.\text{column4}]$	3
$\forall t_0 \in \text{iris14.csv}, t_1 \in \text{iris14.csv}:$	3

$\neg[t_0.\text{column1} > t_1.\text{column1} \wedge t_0.\text{column4} = t_1.\text{column4} \wedge t_0.\text{column2} < t_1.\text{column2}]$	
$\forall t_0 \in \text{iris14.csv}, t_1 \in \text{iris14.csv}: \neg[t_0.\text{column1} = t_1.\text{column1} \wedge t_0.\text{column3} > t_1.\text{column3} \wedge t_0.\text{column2} < t_1.\text{column2}]$	3
$\forall t_0 \in \text{iris14.csv}, t_1 \in \text{iris14.csv}: \neg[t_0.\text{column2} = t_1.\text{column2} \wedge t_0.\text{column4} < t_1.\text{column4} \wedge t_0.\text{column5} = t_1.\text{column5}]$	3
$\forall t_0 \in \text{iris14.csv}, t_1 \in \text{iris14.csv}: \neg[t_0.\text{column1} > t_1.\text{column1} \wedge t_0.\text{column3} = t_1.\text{column3} \wedge t_0.\text{column4} < t_1.\text{column4}]$	3
$\forall t_0 \in \text{iris14.csv}, t_1 \in \text{iris14.csv}: \neg[t_0.\text{column1} < t_1.\text{column1} \wedge t_0.\text{column4} = t_1.\text{column4} \wedge t_0.\text{column3} > t_1.\text{column3}]$	3
$\forall t_0 \in \text{iris14.csv}, t_1 \in \text{iris14.csv}: \neg[t_0.\text{column3} \leq t_1.\text{column3} \wedge t_0.\text{column1} = t_1.\text{column1} \wedge t_0.\text{column2} < t_1.\text{column2}]$	3
$\forall t_0 \in \text{iris14.csv}, t_1 \in \text{iris14.csv}: \neg[t_0.\text{column5} = t_1.\text{column5} \wedge t_0.\text{column3} < t_1.\text{column3} \wedge t_0.\text{column4} > t_1.\text{column4} \wedge t_0.\text{column2} < t_1.\text{column2}]$	4
$\forall t_0 \in \text{iris14.csv}, t_1 \in \text{iris14.csv}: \neg[t_0.\text{column5} \neq t_1.\text{column5} \wedge t_0.\text{column4} = t_1.\text{column4} \wedge t_0.\text{column3} < t_1.\text{column3}]$	4

$t_0.\text{column1} \geq t_1.\text{column1}]$	
$\forall t_0 \in \text{iris14.csv}, t_1 \in \text{iris14.csv}:$ $\neg [t_0.\text{column4} < t_1.\text{column4} \wedge$ $t_0.\text{column5} = t_1.\text{column5} \wedge$ $t_0.\text{column1} \leq t_1.\text{column1} \wedge$ $t_0.\text{column3} > t_1.\text{column3}]$	4
$\forall t_0 \in \text{iris14.csv}, t_1 \in \text{iris14.csv}:$ $\neg [t_0.\text{column2} = t_1.\text{column2} \wedge$ $t_0.\text{column3} \leq t_1.\text{column3} \wedge$ $t_0.\text{column4} \geq t_1.\text{column4}]$	3
$\forall t_0 \in \text{iris14.csv}, t_1 \in \text{iris14.csv}:$ $\neg [t_0.\text{column1} < t_1.\text{column1} \wedge$ $t_0.\text{column5} \neq t_1.\text{column5} \wedge$ $t_0.\text{column3} < t_1.\text{column3} \wedge$ $t_0.\text{column2} < t_1.\text{column2}]$	4
$\forall t_0 \in \text{iris14.csv}, t_1 \in \text{iris14.csv}:$ $\neg [t_0.\text{column5} = t_1.\text{column5} \wedge$ $t_0.\text{column4} > t_1.\text{column4} \wedge$ $t_0.\text{column1} \geq t_1.\text{column1} \wedge$ $t_0.\text{column2} < t_1.\text{column2}]$	4
$\forall t_0 \in \text{iris14.csv}, t_1 \in \text{iris14.csv}:$ $\neg [t_0.\text{column3} = t_1.\text{column3} \wedge$ $t_0.\text{column4} = t_1.\text{column4}]$	2
$\forall t_0 \in \text{iris14.csv}, t_1 \in \text{iris14.csv}:$ $\neg [t_0.\text{column5} \neq t_1.\text{column5} \wedge$ $t_0.\text{column4} = t_1.\text{column4} \wedge$ $t_0.\text{column3} < t_1.\text{column3} \wedge$ $t_0.\text{column2} < t_1.\text{column2}]$	4
$\forall t_0 \in \text{iris14.csv}, t_1 \in \text{iris14.csv}:$ $\neg [t_0.\text{column1} > t_1.\text{column1} \wedge$ $t_0.\text{column4} < t_1.\text{column4} \wedge$ $t_0.\text{column5} = t_1.\text{column5}]$	4

t0.column3 < t1.column3]	
$\forall t0 \in \text{iris14.csv}, t1 \in \text{iris14.csv}:$ $\neg[t0.\text{column5} \neq t1.\text{column5} \wedge$ t0.column3 = t1.column3 $\wedge$ t0.column4 < t1.column4]	3
$\forall t0 \in \text{iris14.csv}, t1 \in \text{iris14.csv}:$ $\neg[t0.\text{column5} \neq t1.\text{column5} \wedge$ t0.column4 < t1.column4 $\wedge$ t0.column1 $\leq$ t1.column1 $\wedge$ t0.column2 < t1.column2]	4
$\forall t0 \in \text{iris14.csv}, t1 \in \text{iris14.csv}:$ $\neg[t0.\text{column2} = t1.\text{column2} \wedge$ t0.column3 = t1.column3]	2
$\forall t0 \in \text{iris14.csv}, t1 \in \text{iris14.csv}:$ $\neg[t0.\text{column2} = t1.\text{column2} \wedge$ t0.column4 = t1.column4]	2
$\forall t0 \in \text{iris14.csv}, t1 \in \text{iris14.csv}:$ $\neg[t0.\text{column5} \neq t1.\text{column5} \wedge$ t0.column2 $\geq$ t1.column2 $\wedge$ t0.column4 = t1.column4 $\wedge$ t0.column1 $\leq$ t1.column1]	4
$\forall t0 \in \text{iris14.csv}, t1 \in \text{iris14.csv}:$ $\neg[t0.\text{column5} \neq t1.\text{column5} \wedge$ t0.column3 = t1.column3 $\wedge$ t0.column1 $\leq$ t1.column1]	3
$\forall t0 \in \text{iris14.csv}, t1 \in \text{iris14.csv}:$ $\neg[t0.\text{column2} = t1.\text{column2} \wedge$ t0.column3 $\geq$ t1.column3 $\wedge$ t0.column1 < t1.column1]	3
$\forall t0 \in \text{iris14.csv}, t1 \in \text{iris14.csv}:$ $\neg[t0.\text{column2} \geq t1.\text{column2} \wedge$ t0.column3 = t1.column3 $\wedge$	3

t0.column5=t1.column5]	
$\forall t0 \in \text{iris14.csv}, t1 \in \text{iris14.csv}:$ $\neg[t0.\text{column2}=t1.\text{column2} \wedge$ t0.column5=t1.column5 $\wedge$ t0.column3>t1.column3]	3
$\forall t0 \in \text{iris14.csv}, t1 \in \text{iris14.csv}:$ $\neg[t0.\text{column3}=t1.\text{column3} \wedge$ t0.column5=t1.column5 $\wedge$ t0.column1\geq t1.column1]	3
$\forall t0 \in \text{iris14.csv}, t1 \in \text{iris14.csv}:$ $\neg[t0.\text{column3}\leq t1.\text{column3} \wedge$ t0.column4\geq t1.column4 $\wedge$ t0.column1=t1.column1]	3
$\forall t0 \in \text{iris14.csv}, t1 \in \text{iris14.csv}:$ $\neg[t0.\text{column2}=t1.\text{column2} \wedge$ t0.column5=t1.column5 $\wedge$ t0.column1\geq t1.column1]	3
$\forall t0 \in \text{iris14.csv}, t1 \in \text{iris14.csv}:$ $\neg[t0.\text{column5}=t1.\text{column5} \wedge$ t0.column3< t1.column3 $\wedge$ t0.column1\geq t1.column1 $\wedge$ t0.column2< t1.column2]	4
$\forall t0 \in \text{iris14.csv}, t1 \in \text{iris14.csv}:$ $\neg[t0.\text{column1}< t1.\text{column1} \wedge$ t0.column4< t1.column4 $\wedge$ t0.column3> t1.column3]	3
$\forall t0 \in \text{iris14.csv}, t1 \in \text{iris14.csv}:$ $\neg[t0.\text{column4}< t1.\text{column4} \wedge$ t0.column5=t1.column5 $\wedge$ t0.column1=t1.column1]	3
$\forall t0 \in \text{iris14.csv}, t1 \in \text{iris14.csv}:$ $\neg[t0.\text{column3}=t1.\text{column3} \wedge$	2

t0.column1=t1.column1]	
$\forall t0 \in \text{iris14.csv}, t1 \in \text{iris14.csv}:$ $\neg[t0.\text{column4}=t1.\text{column4} \wedge$ t0.column1=t1.column1]	2
$\forall t0 \in \text{iris14.csv}, t1 \in \text{iris14.csv}:$ $\neg[t0.\text{column2} \geq t1.\text{column2} \wedge$ t0.column3=t1.column3 $\wedge$ t0.column1 $\leq$ t1.column1]	3