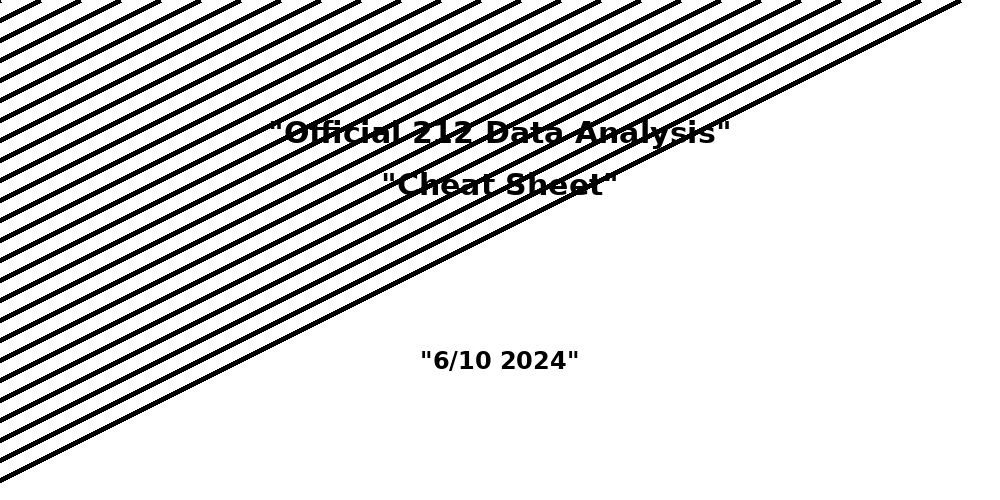
****

[*ГИТХАБ ПМИ МО-1*](https://github.com/esokolov/ml-course-hse)

[*ЯНДЕКС УЧЕБНИК ПО МО*](https://education.yandex.ru/handbook/ml)

[РЕШЕНИЕ ТРЕН ВАР 1](https://drive.google.com/file/d/1Vc6q_5XMR7VoEmb9_PSZRmfKkOe-TUEp/view?usp=drive_link)

[РЕШЕНИЕ ТРЕН ВАР 2](https://drive.google.com/file/d/13W1kaGC9HZo4ThJT36KeMRVS0MY59PnS/view?usp=sharing)

## 

## “ТЕСТОВАЯ ЧАСТЬ”

**Вопрос 1**

Визуализация, приведенная ниже, называется по-английски stacked barplot (многоуровневая диаграмма).

Ниже изображена гистограмма stacked barplot, отображающая распределение населения Нью-Йорка и окрестностей по районам с 1790 по 2010 годы.

На этой гистограмме по оси абсцисс отложены значения одного признака (ниже - год), а по оси ординат - значения другого признака (ниже - процент от общего населения).

Выберите верные утверждения относительно данной гистограммы:

1. Пик количества населения в Нью-Йорке и окрестностях было в 1830 году
2. Доля населения среди всего населения Нью-Йорка и окрестностей в районе Квинс (Queens) монотонно увеличивалась с 1790 по 2010 годы.
3. Среди всех перечисленных районов наименьшая доля населения в 2010 году жила в Staten Island.
4. Больше всего населения в Манхеттене (Manhattan) было в 1830 году

*ANS (график утерян, скорее всего тут бред)*

Без визуализации сложно дать точный ответ. Но логически:

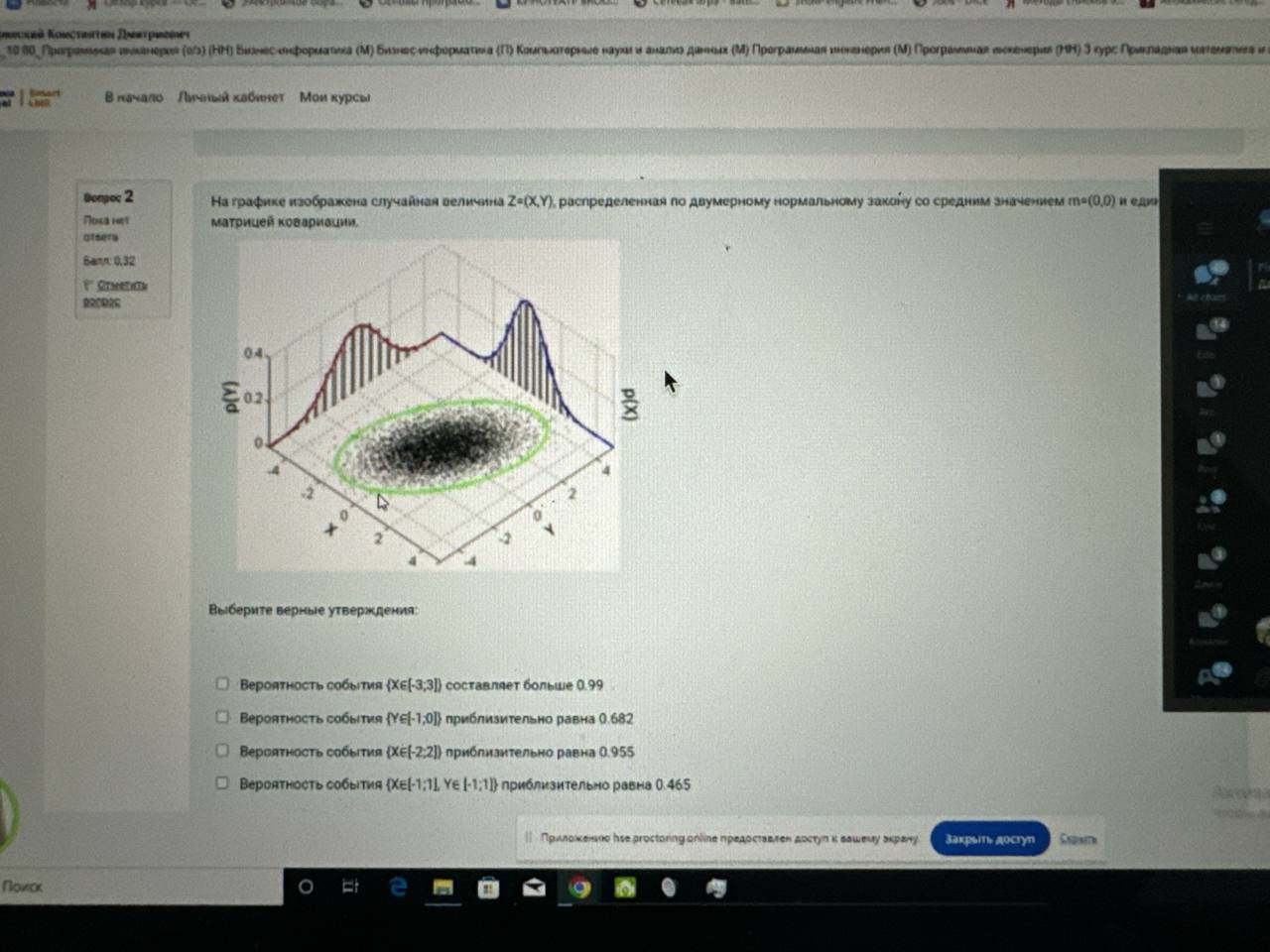
1. **Неверно**, пик населения не обязательно в 1830 году, это слишком специфично без графика.
2. **Может быть верно**, если данные поддерживают это.
3. **Может быть верно**, но требует проверки данных.
4. **Неверно**, опять же слишком специфично без данных.

**Вопрос 2**

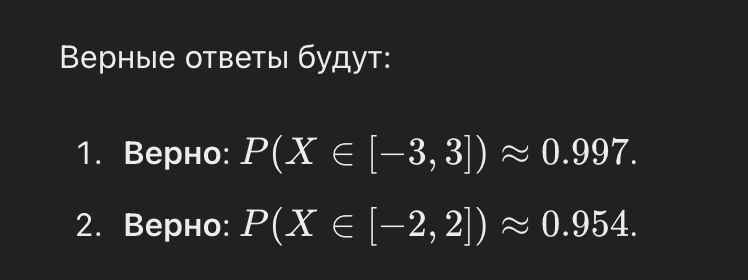
На графике изображена случайная величина Z=(X,Y), распределенная по двумерному нормальному закону со средним значением m=(0,0) и единичной матрицей ковариации.

Выберите верные утверждения:

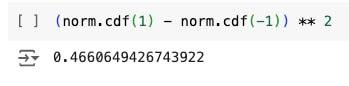
1. Вероятность события {X∈[-3;3]} составляет больше 0.99
2. Вероятность события {Y∈[-1;0]} приблизительно равна 0.682
3. Вероятность события {X∈[-2;2]} приблизительно равна 0.955
4. Вероятность события {X∈[-1;1], Y∈ [-1;1]} приблизительно равна 0.465



*ANS*



*Д. Р. говорит: ответ 4 тоже должен быть, он, видимо, не понял, что распределение двумерное*



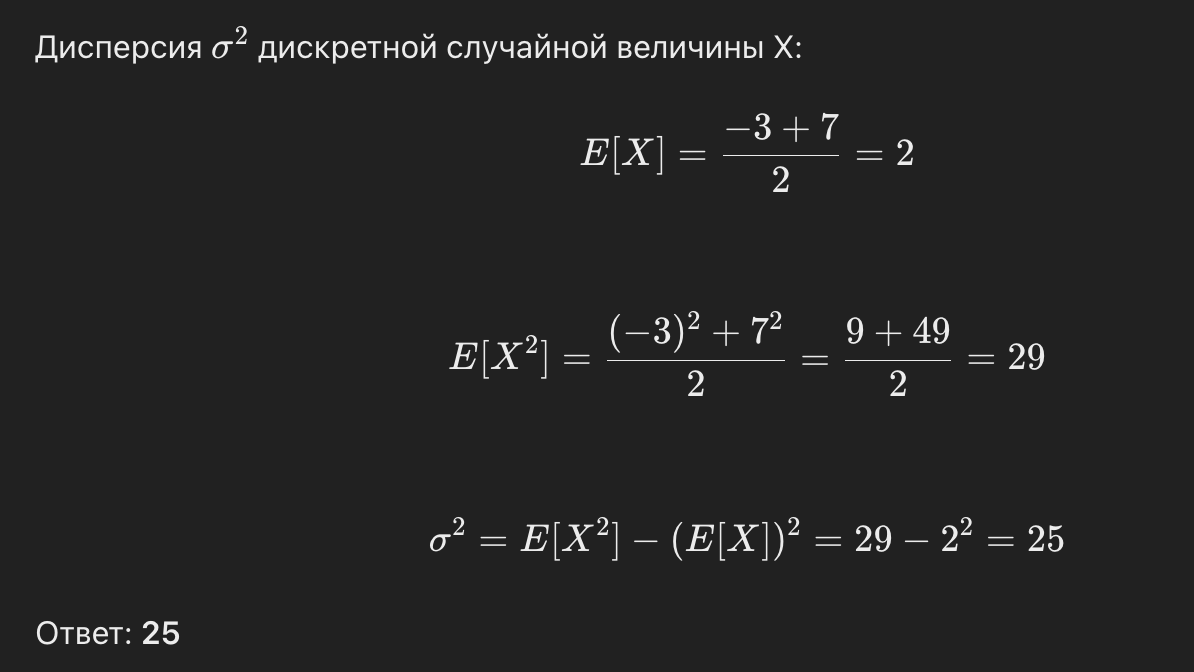
**Вопрос 3**

Известно, что дискретная случайная величина X принимает лишь два значения: -3 и 7, причем с равной вероятностью.

Найдите дисперсию дискретной случайной величины X.

1. 23
2. 25
3. 21
4. 27

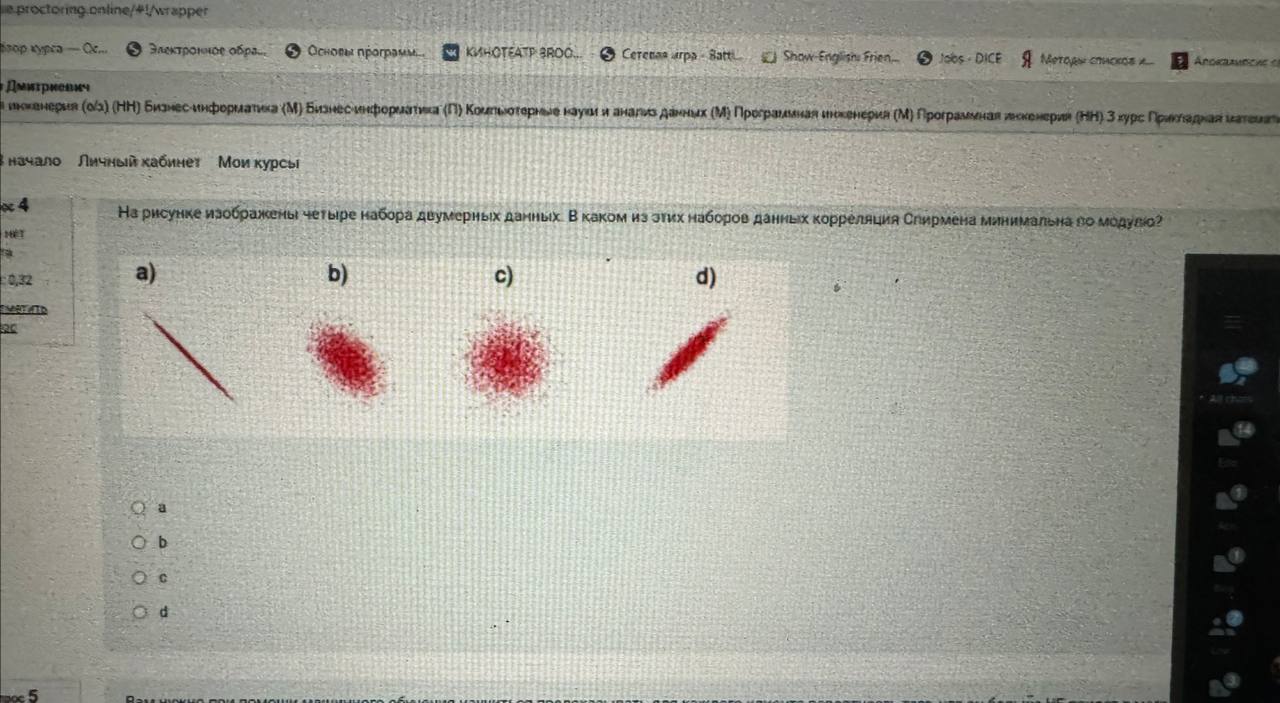
*ANS*



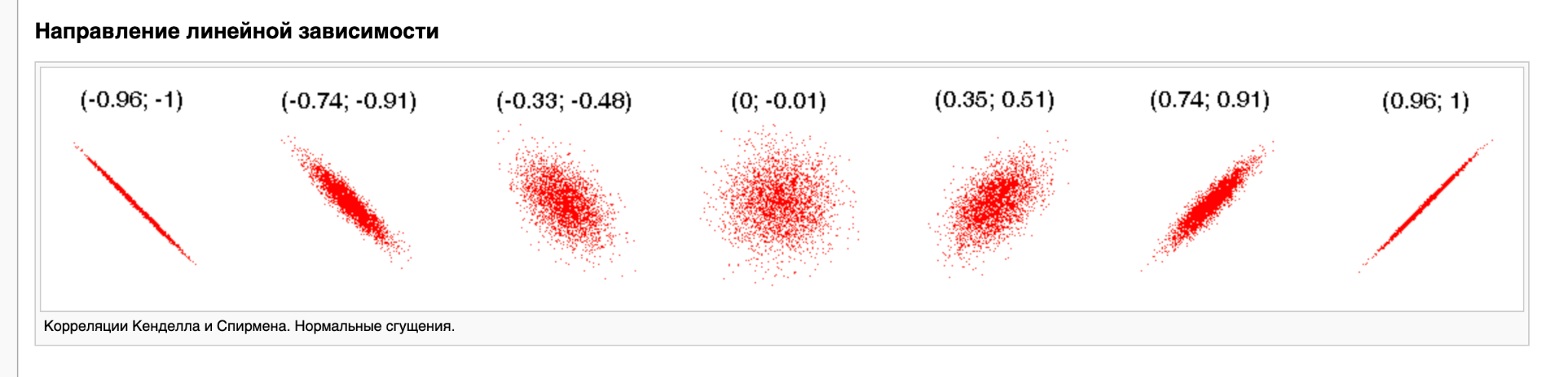
**Вопрос 4**

На рисунке изображены четыре набора двумерных данных. В каком из этих наборов данных корреляция Спирмена минимальна по модулю?

1. a
2. b
3. c
4. d



*ANS* без графика определить невозможно (приступ слепоты), но по факту там с)



**Вопрос 5**

Вам нужно при помощи машинного обучения научиться предсказывать для каждого клиента вероятность того, что он больше НЕ придет в магазин. У вас есть следующие признаки: возраст клиента, сумма последней покупки в магазине, наличие карты лояльности магазина и несколько других характеристик клиента.

Целевая переменная - это бинарная величина, равная 1, если клиент еще придет в магазин, и 0, если не придет. Целевая переменная используется в алгоритме в исходном виде, без каких-либо изменений. Какую или какие из перечисленных метрик можно использовать для оценки качества алгоритма в этой задаче?

1. SMAPE
2. RMSLE (Root Mean Squared Logarithmic Error)
3. PR-AUC
4. WAPE
5. Не подходит ни одна из перечисленных

*ANS* **PR-AUC** подходит для бинарной классификации

**Вопрос 6**

В каких случаях используется L1-регуляризация?

1. При решении задач классификации
2. Для уменьшения недообучения
3. При решении задач регрессии
4. Для снижения размерности признакового пространства
5. Для снижения переобучения
6. Для увеличения сложности модели

*ANS* Ответ: **1, 3, 4, 5**

**Вопрос 7**

Пусть мы пытаемся решить задачу классификации при помощи алгоритмов кластеризации. В качестве ответов (классов) будем брать номера предсказанных моделями кластеров. Какие метрики качества можно использовать в этой задаче для оценки качества классификации при таком подходе?

1. MAE
2. RMSE
3. Silhouette score
4. Гомогенность
5. V-мера

*ANS* Ответ: **Гомогенность, V-мера**

**Вопрос 8**

Для оценки качества модели в задаче бинарной классификации была выбрана метрика precision (точность). Для вычисления метрики сначала получаем уверенности, спрогнозированные моделью (например, методом predict\_proba из sklearn), затем по некоторому порогу переводим вероятности в классы и вычисляем precision.

Какие из преобразований гарантированно не повлияют на метрику?

1. Калибровка вероятностей (при сохранении исходного порога для перевода вероятностей в классы)
2. Изменение порога для перевода вероятностей в классы
3. Все перечисленные преобразования могут повлиять на выбранную метрику
4. Изменение модели

*ANS* Ответ: **Калибровка вероятностей**

**Вопрос 9**

Выберите верные недостатки стохастического градиентного спуска:

1. Метод требует больше памяти для хранения данных, чем классический градиентный спуск
2. Метод может разойтись при неверном выборе градиентного шага
3. Метод гарантирует нахождение минимума с заранее заданной точностью в случае оптимальной стратегии выбора градиентного шага
4. Метод не гарантирует нахождение глобального минимума функции потерь

*ANS* Ответ: **2, 4**

**Вопрос 10**

Для решения задачи регрессии было решено использовать композицию из трех алгоритмов: линейной регрессии, метода опорных векторов с радиальным ядром и решающего дерева.

Какой из алгоритмов мог быть использован?

1. Бутстреп
2. Стекинг
3. Бустинг
4. Бэггинг

*ANS* Ответ: **Стекинг**

**Вопрос 11**

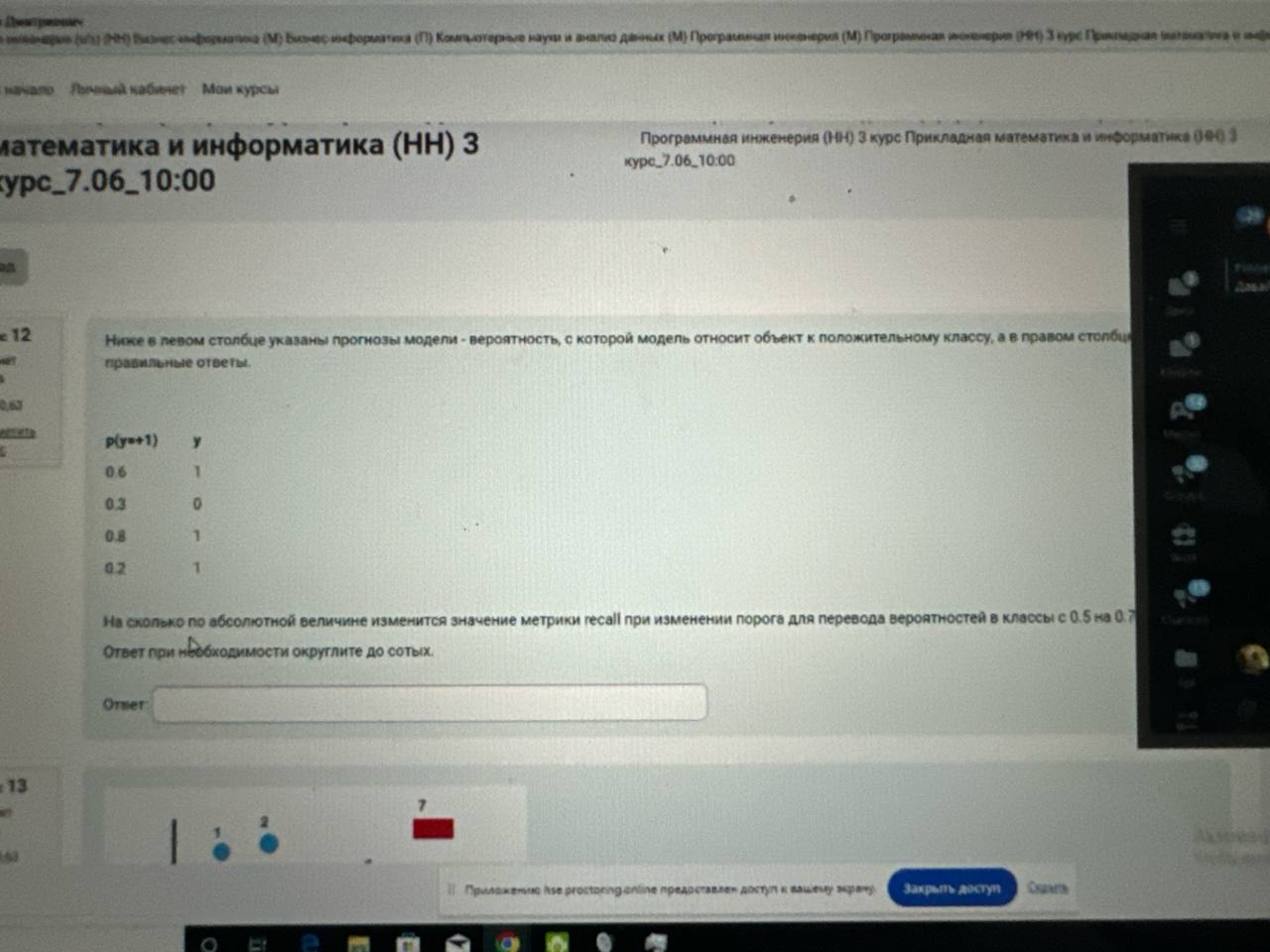
Выберите верные утверждения про иерархическую (агломеративную) кластеризацию:

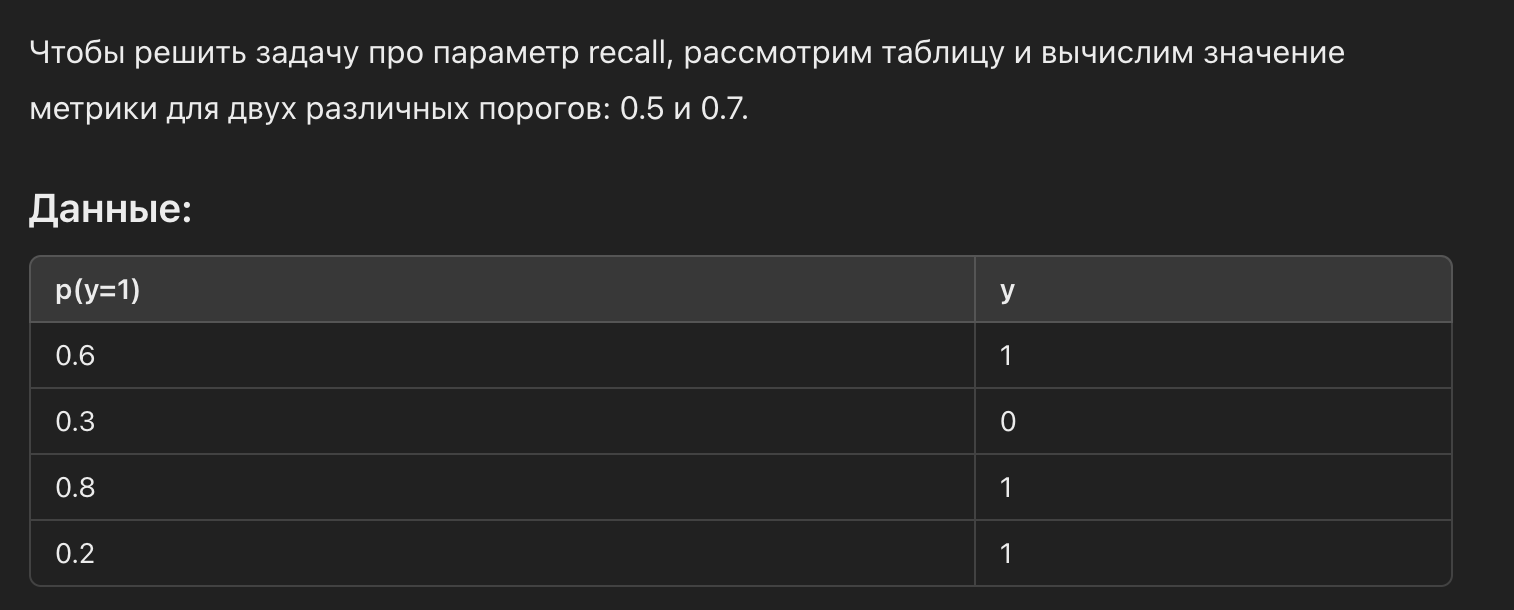
1. Число кластеров в методе можно задать заранее, а можно выбрать, например, при помощи дендрограммы и различных эвристик
2. Результат работы метода можно визуализировать при помощи дендрограммы
3. Метод не умеет находить кластеры сложной формы (умеет выявлять только сплошные однородные кластеры в форме облака)
4. Результат работы метода зависит от способа вычисления расстояния между кластерами на промежуточных этапах

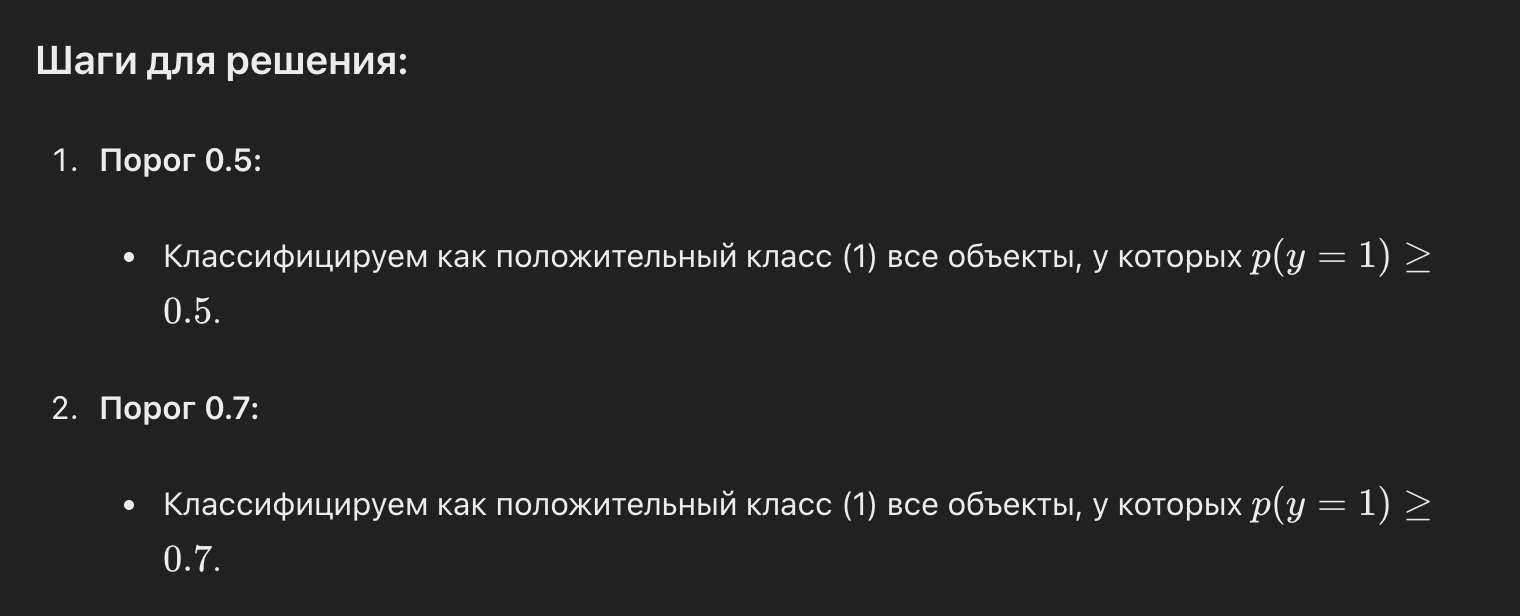
*ANS* Ответ: **1, 2, 4**

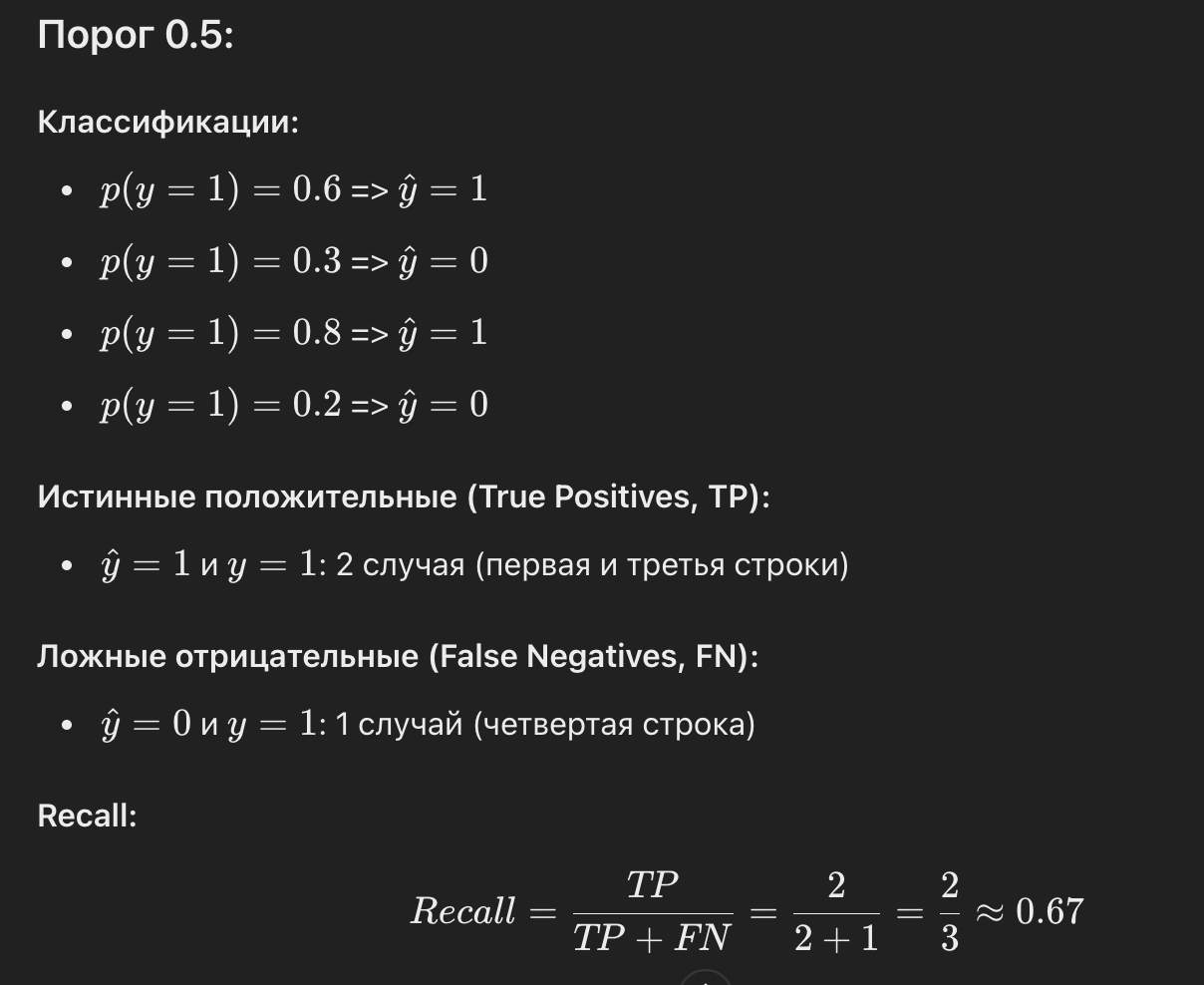
## “ЗАДАЧИ НА ПОДСЧЕТ”

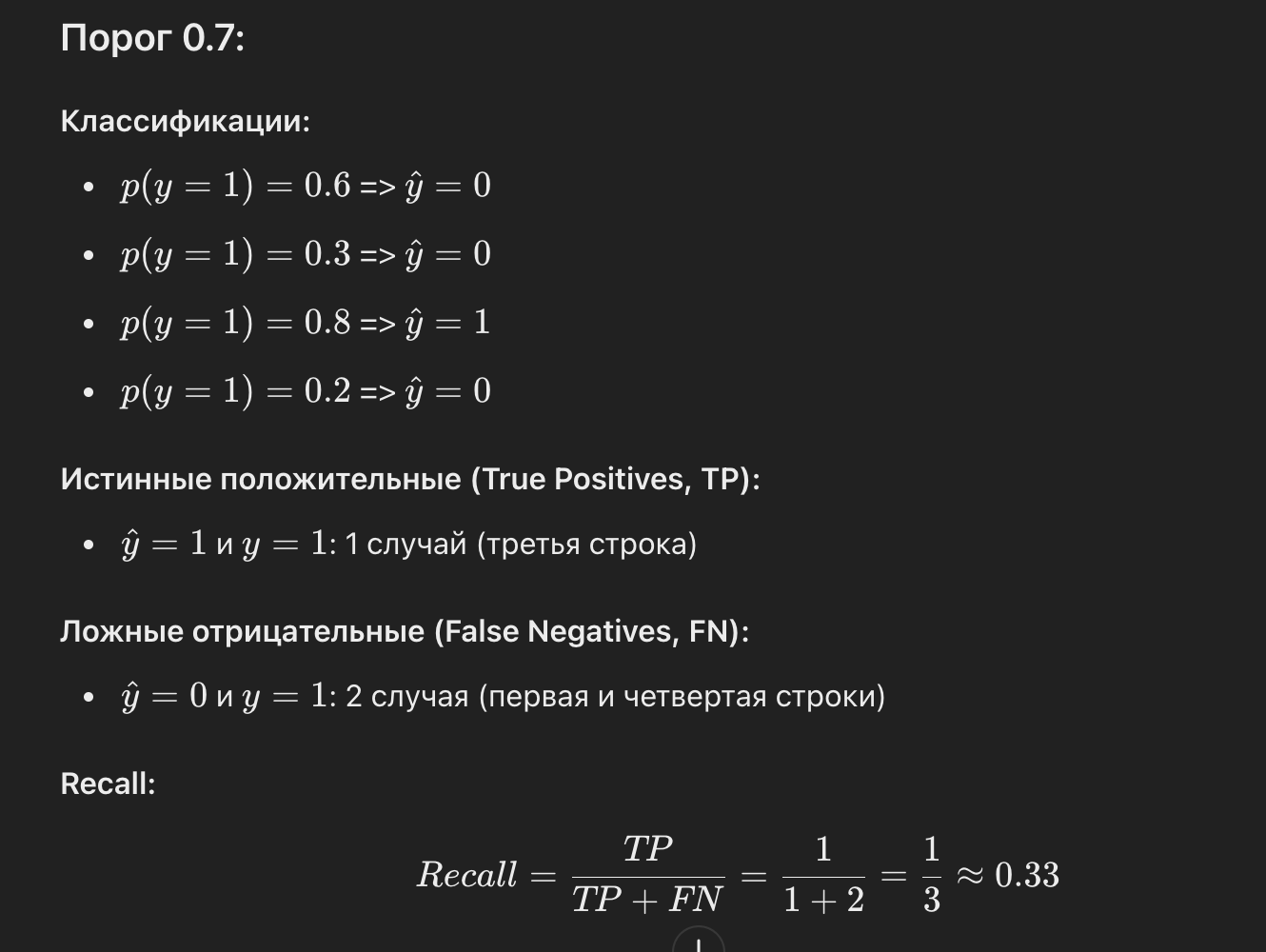
ЗАДАЧА НА RECALL

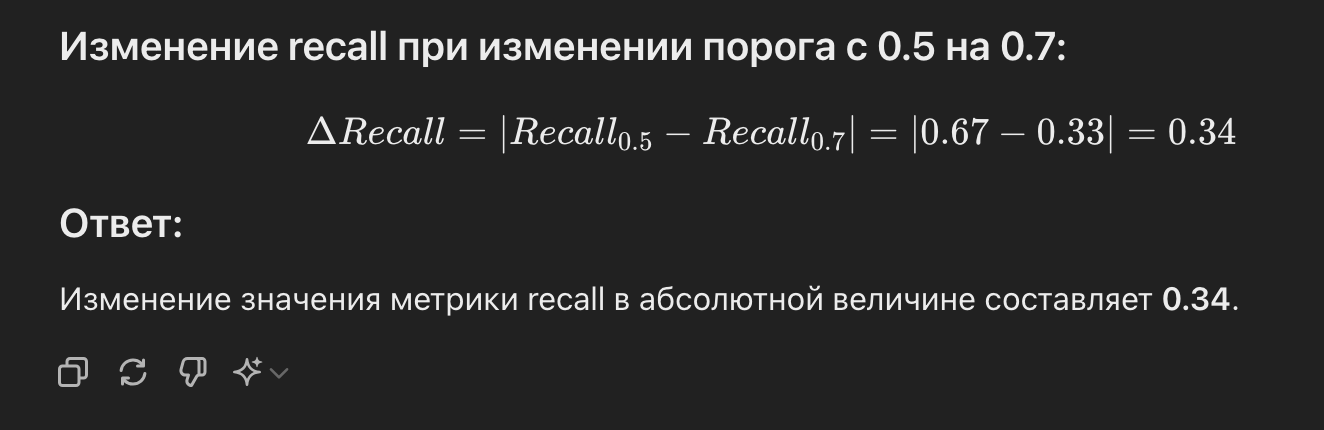






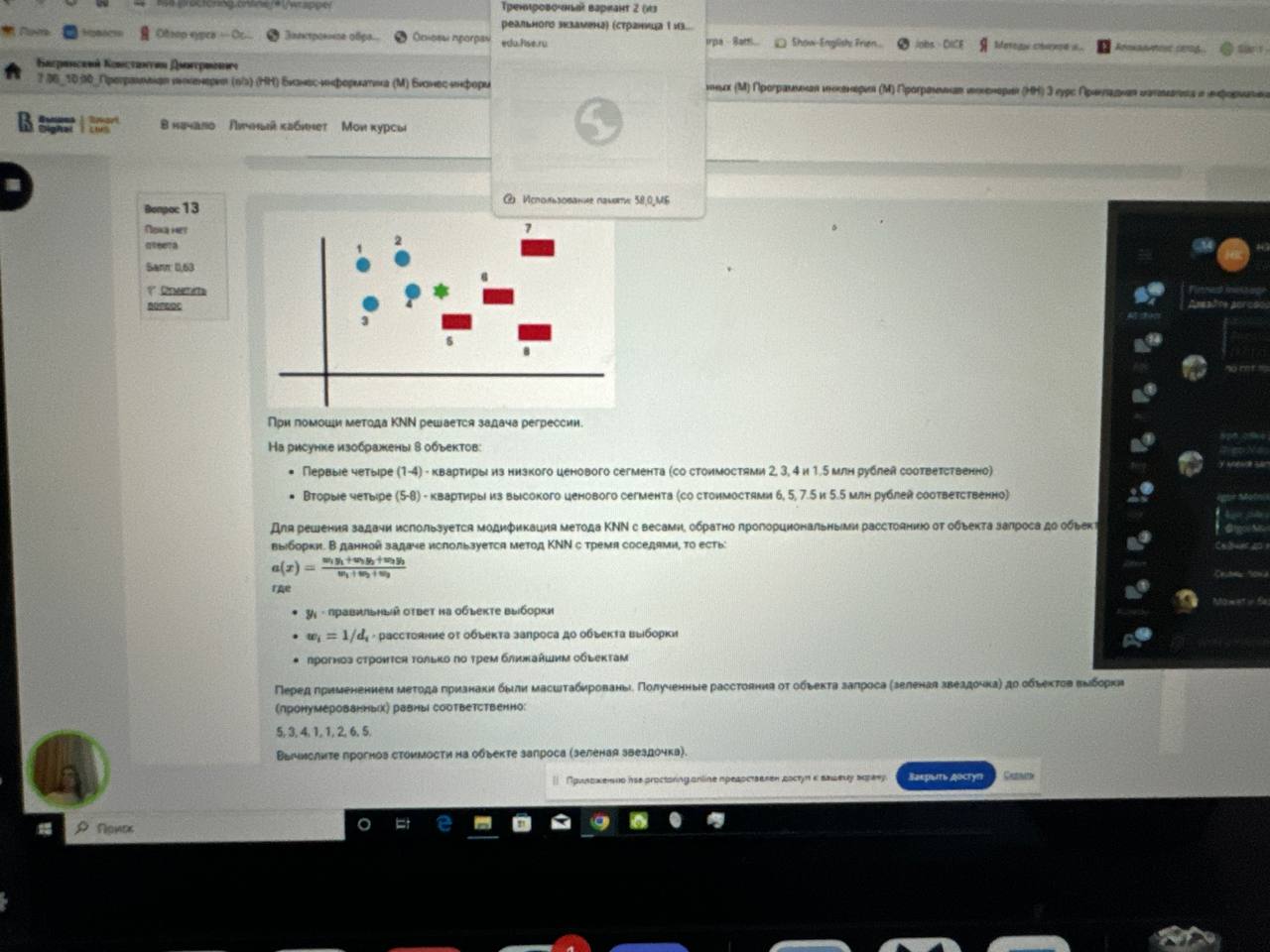


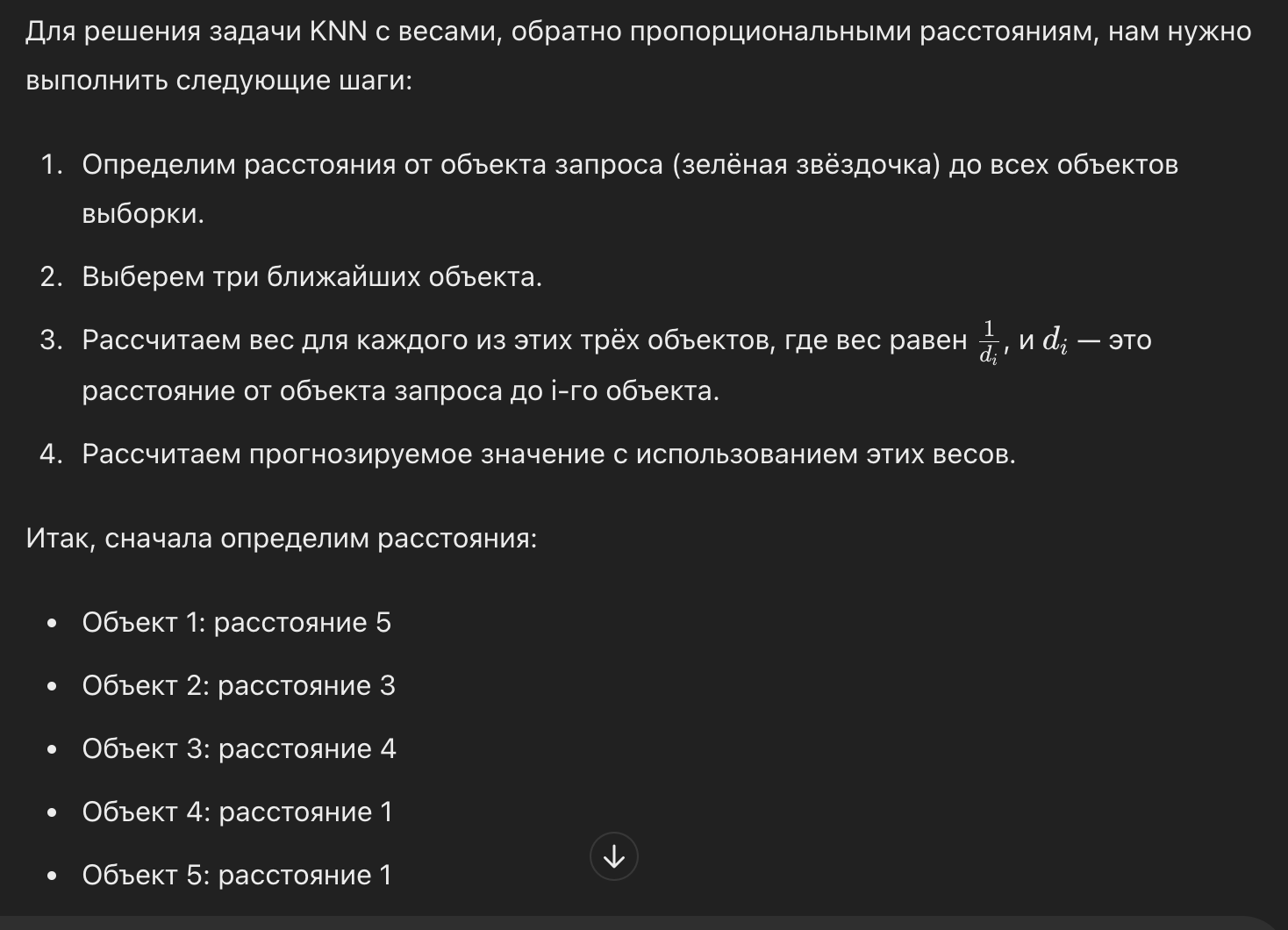


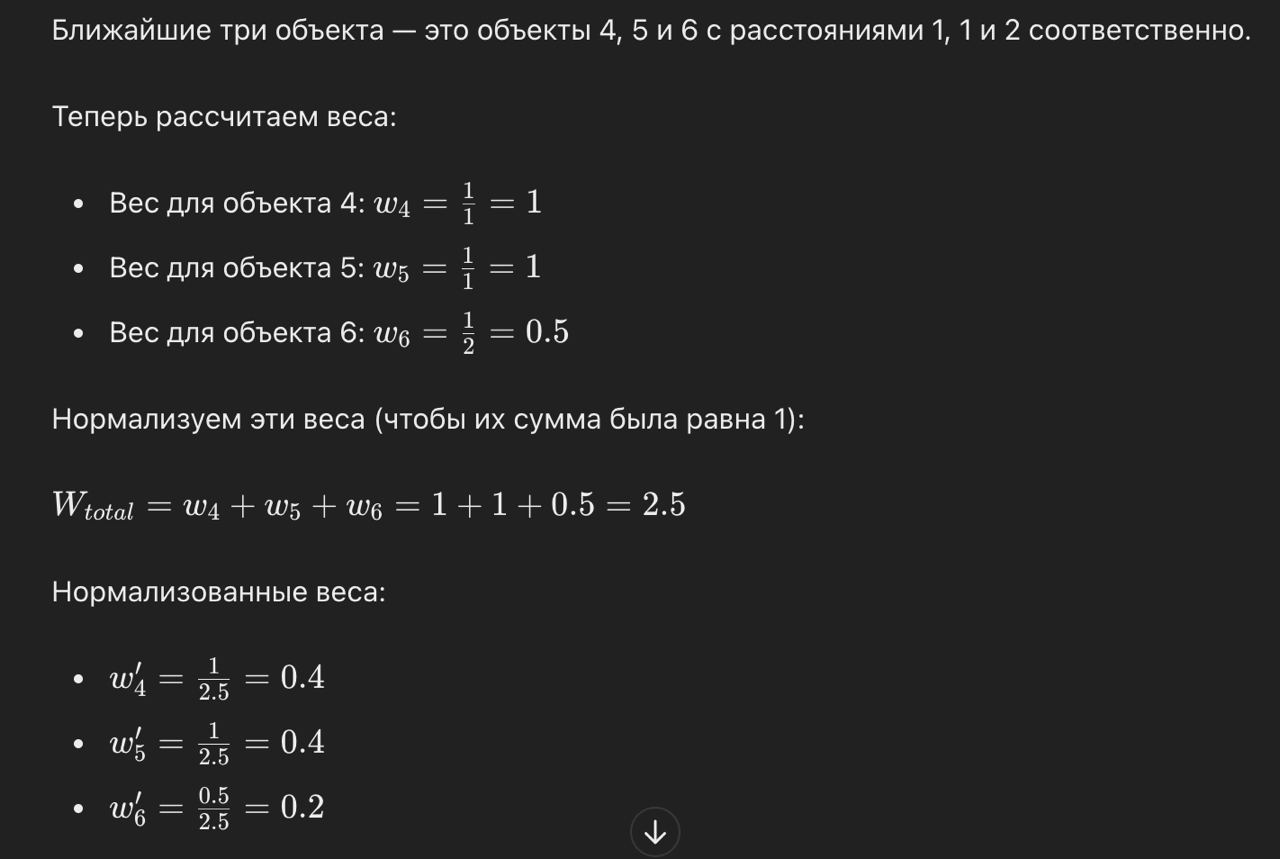


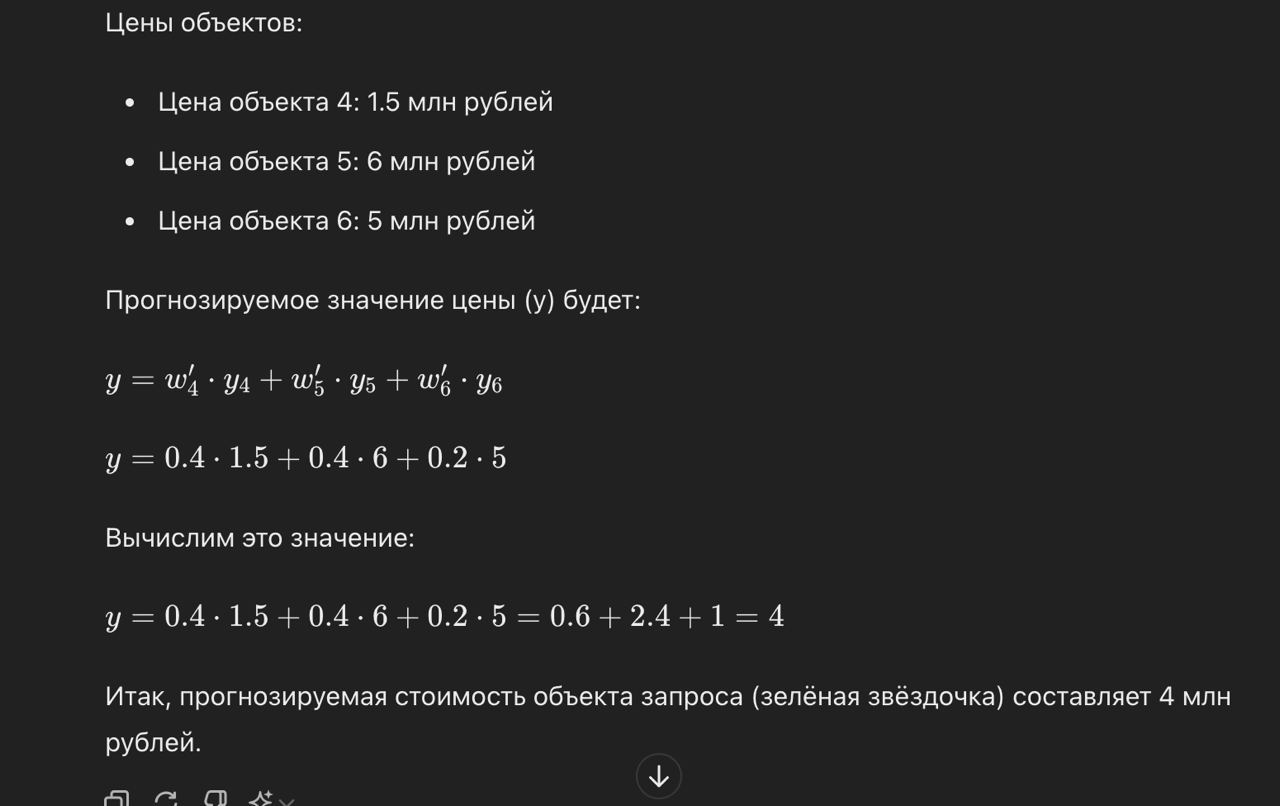
*Д. Р. говорит: округлено на полрешении в задаче на recall (получил 0.34, а должен был 0.33)*

ЗАДАЧА НА МОЩНУЮ РЕГРЕССИЮ ЧЕРЕЗ KNN

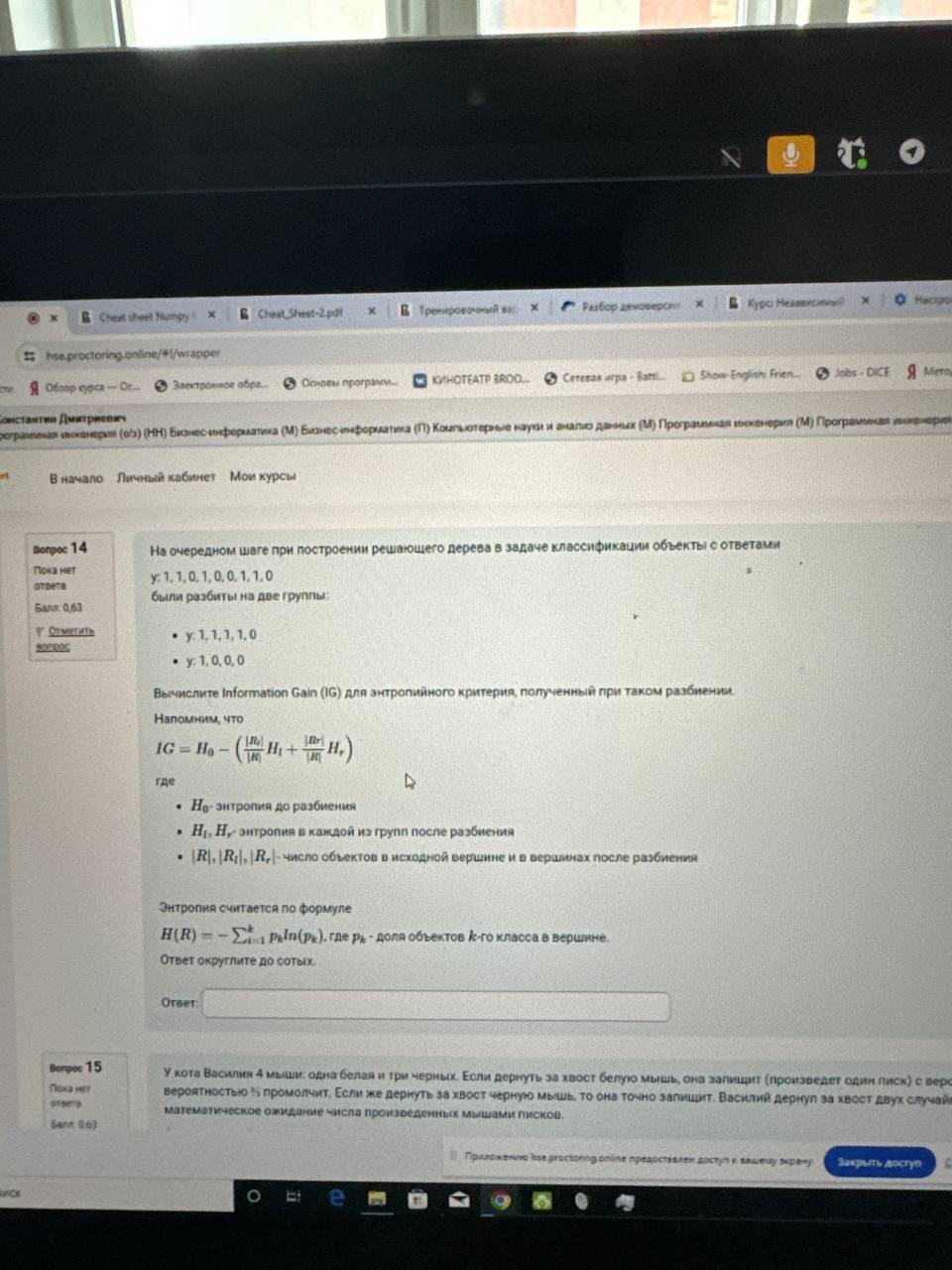


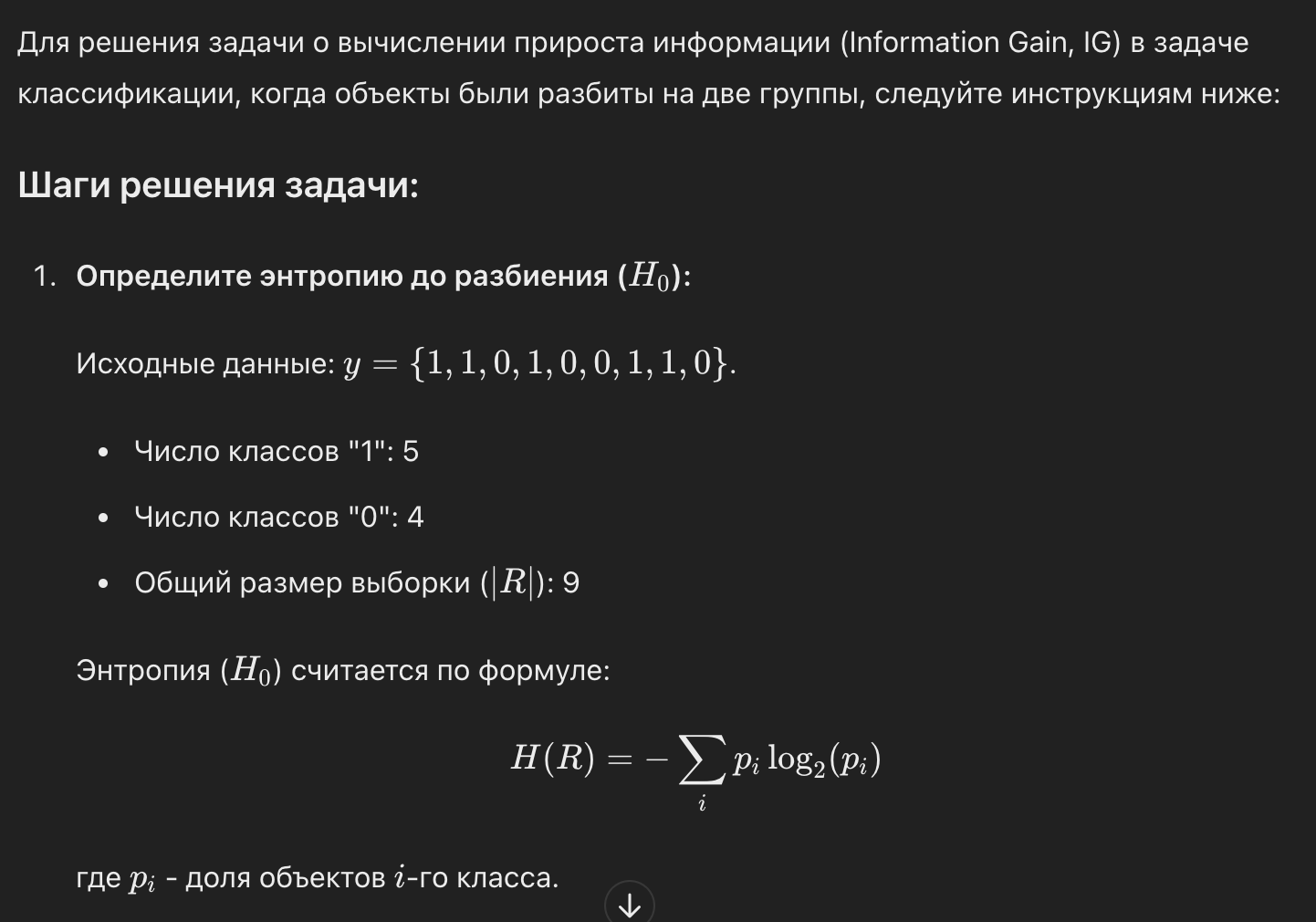


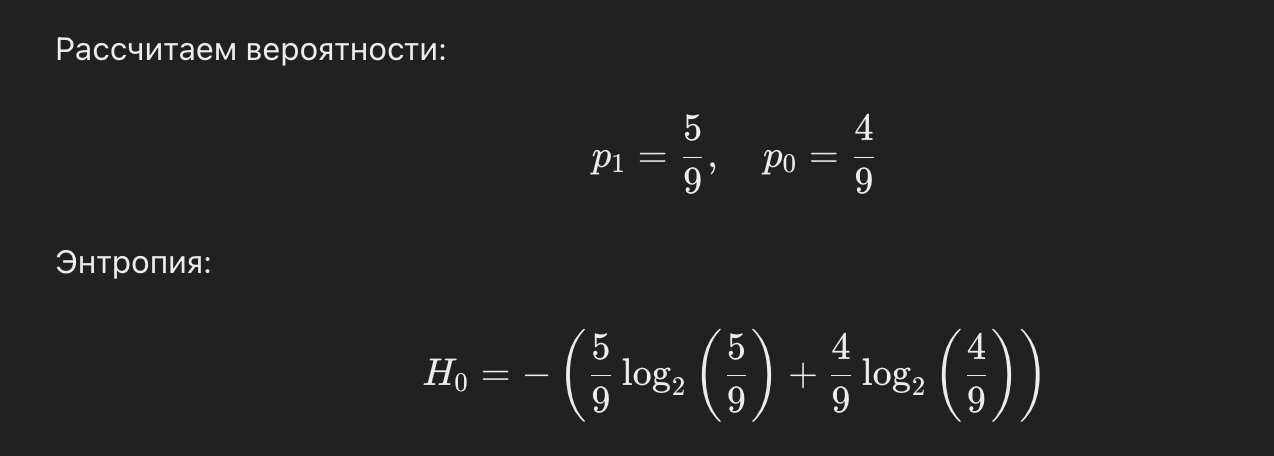


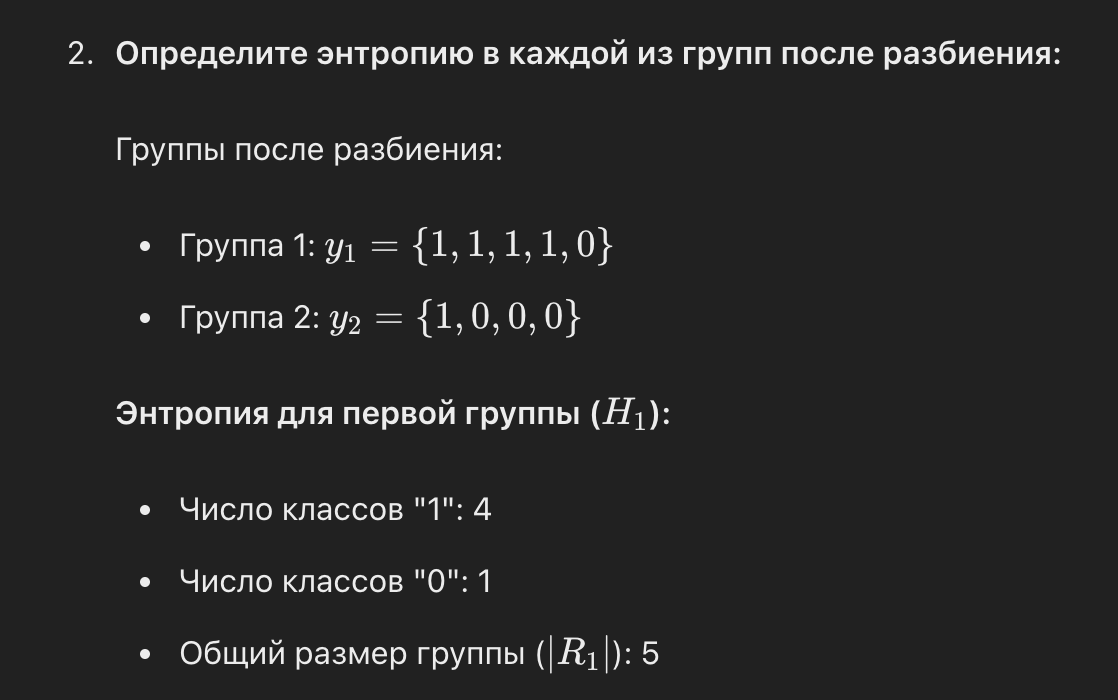


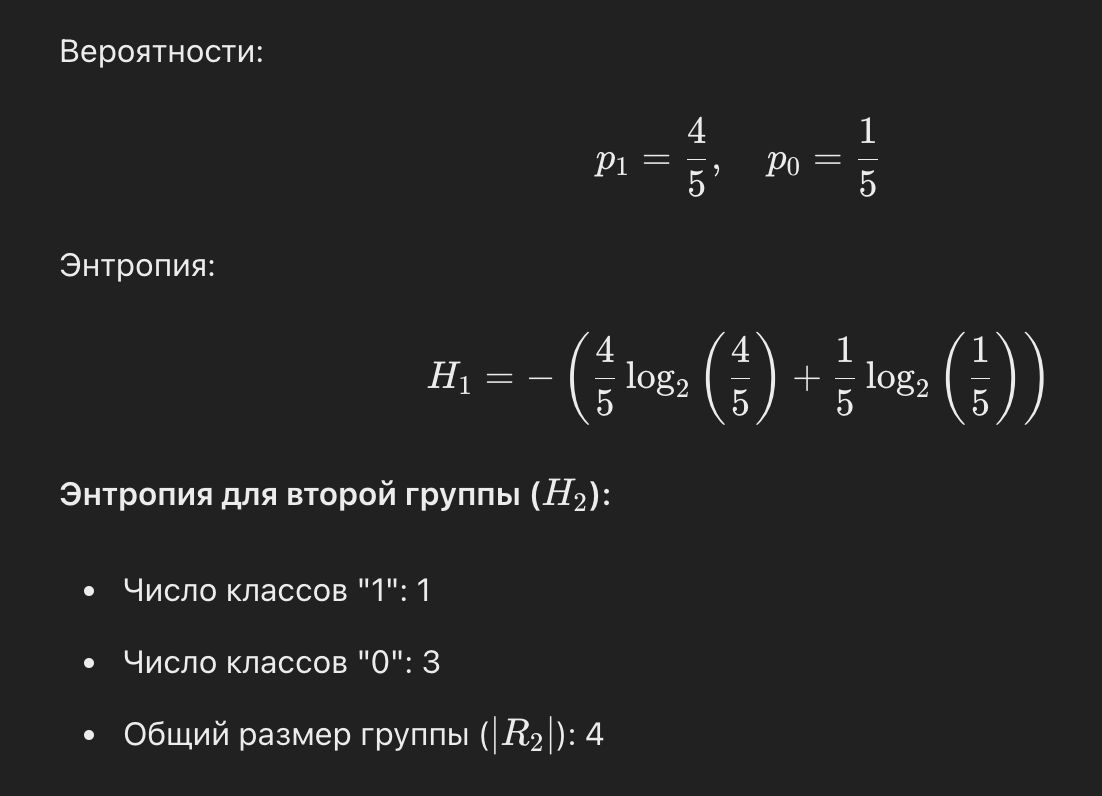
ЗАДАЧА НА РЕШАЮЩЕЕ ДЕРЕВО

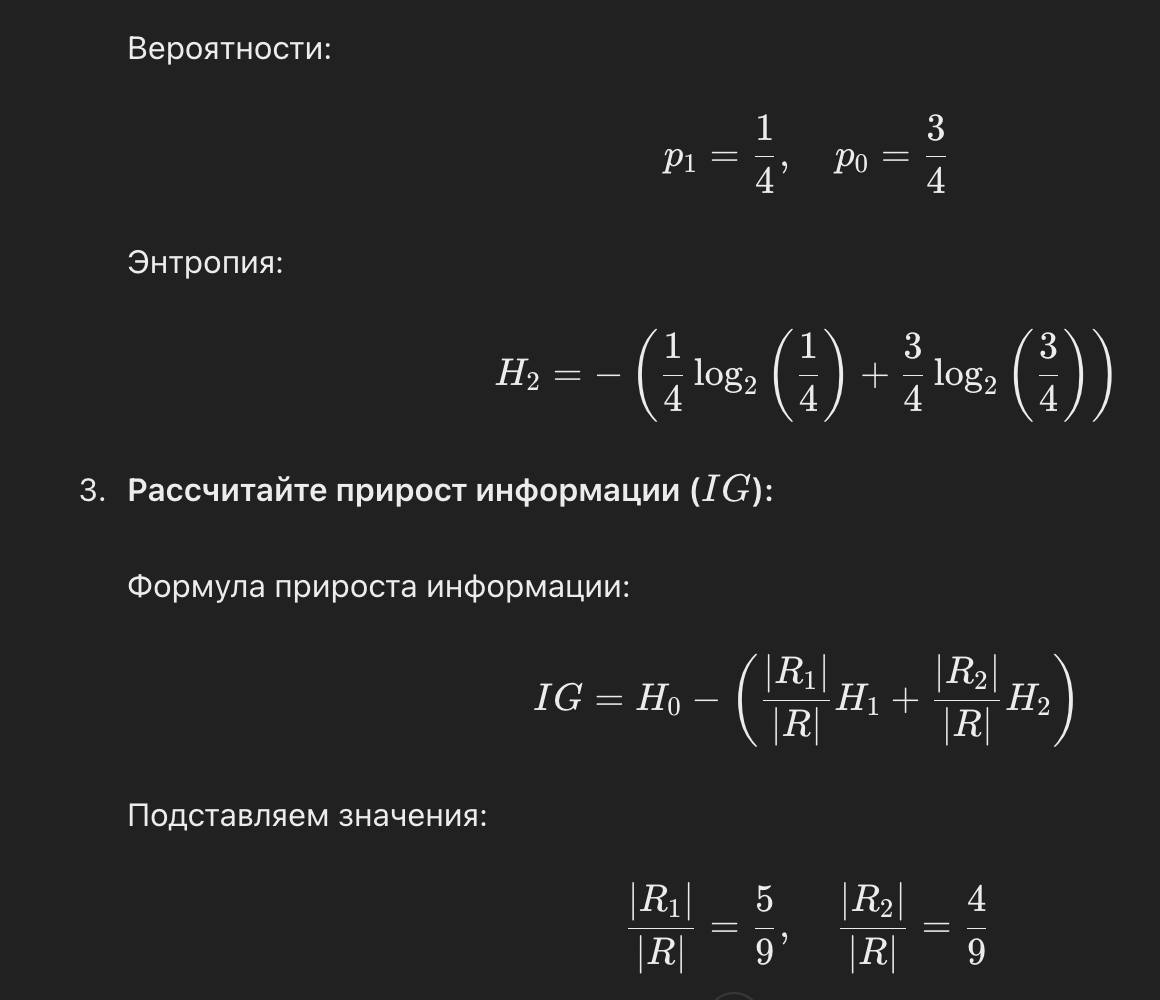


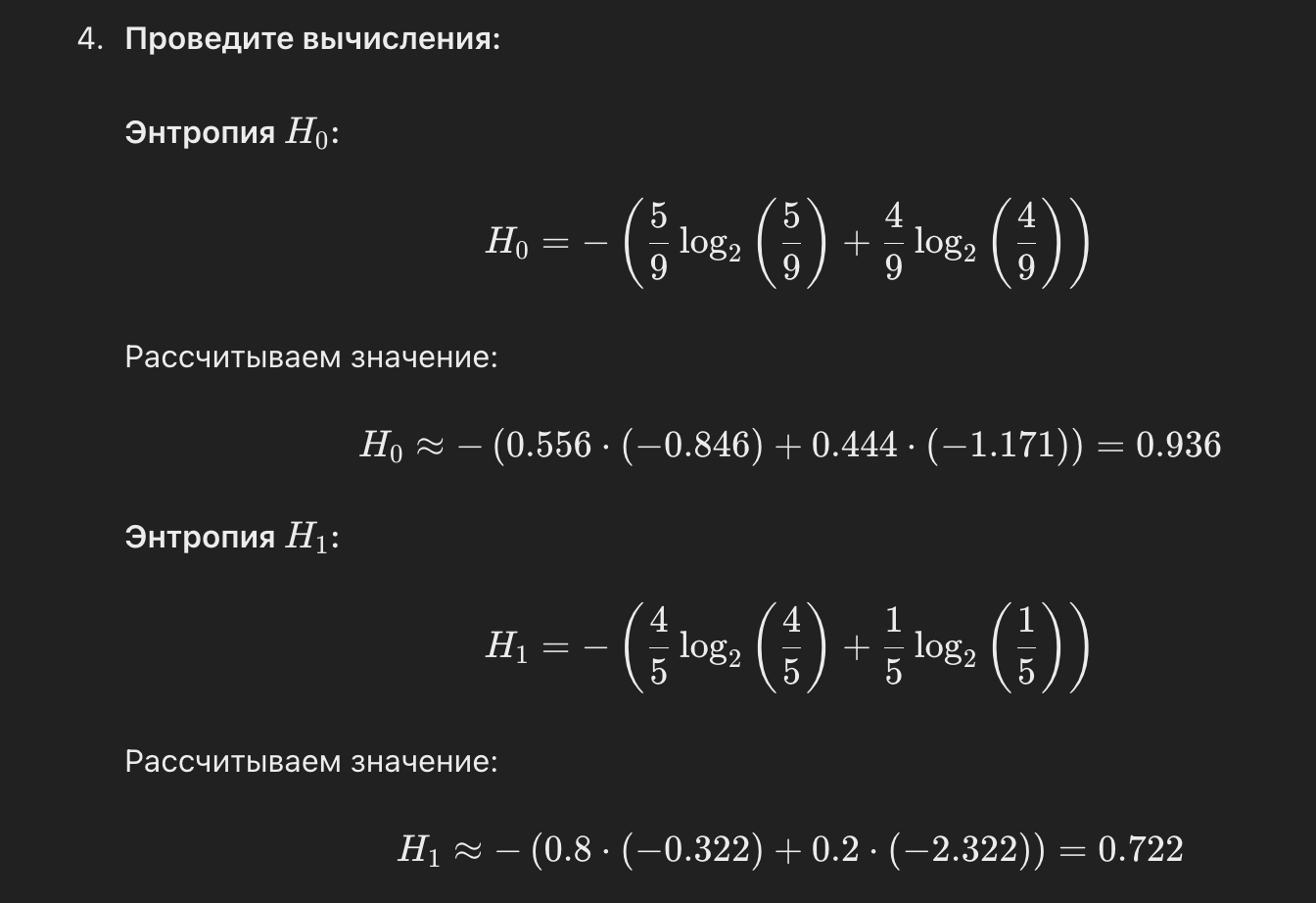


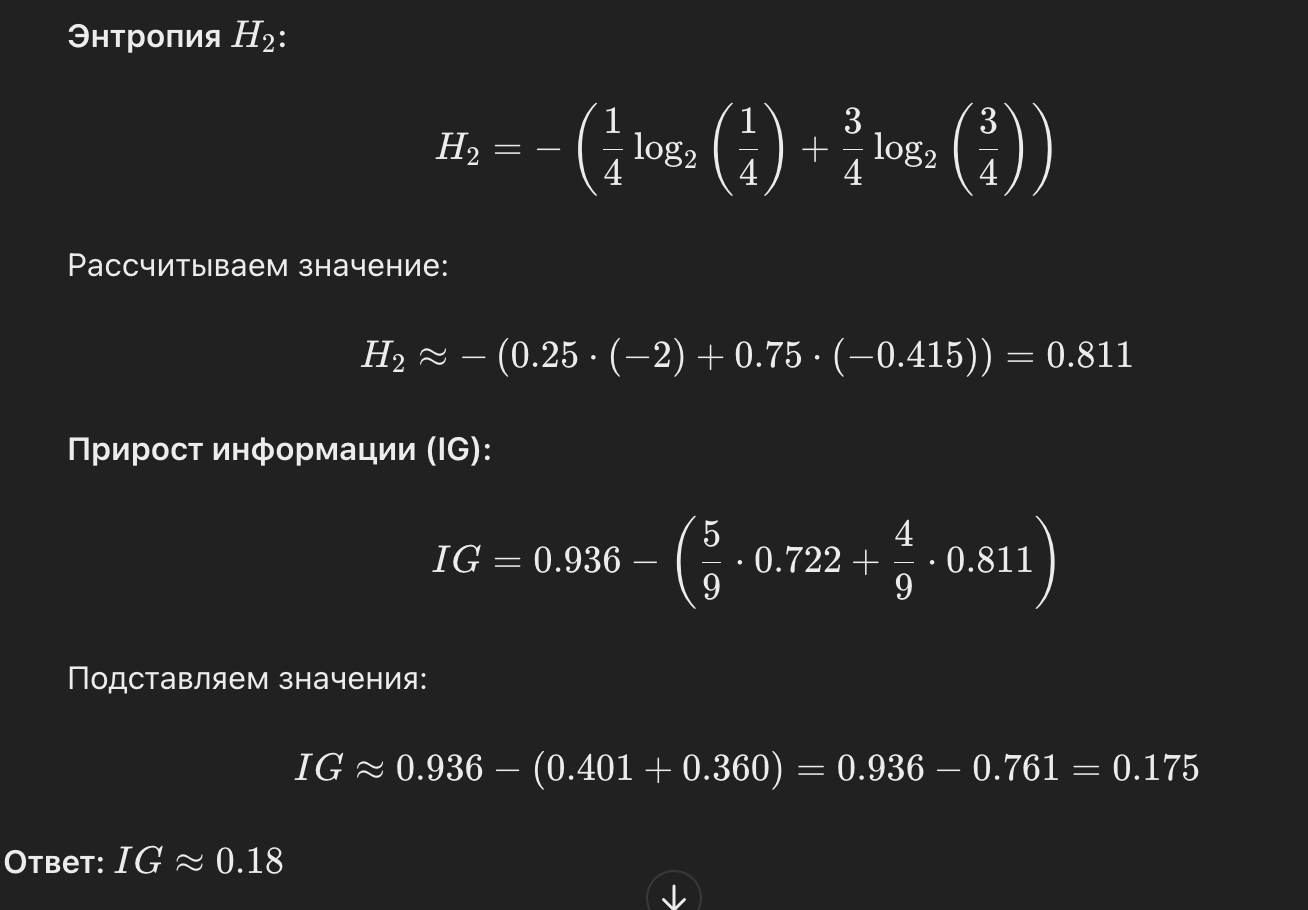


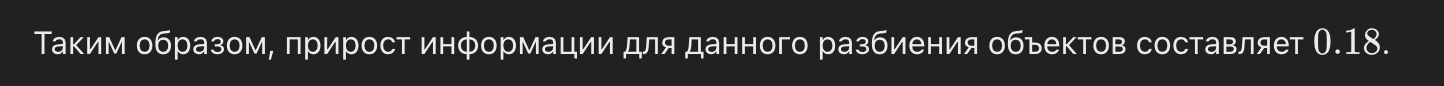












*Д. Р. говорит: в задаче на энтропию должно быть, вроде как, 0.16. Скорее всего, опять округлил несколько раз где не надо*

ЗАДАЧА ПРО МЫШЕЙ НА ТЕРВЕР

