**Задание 1 – 2 балла**

Опишите своими словами лемму Джонсона-Линдештрауса, чем она важна в машинном обучении?

**Ответ:**

Лемма Джонсона-Линдстрауса говорит о том, что можно спроецировать точки из пространства высокой размерности в пространство гораздо меньшей размерности так, чтобы расстояния между ними почти не изменились. Это важно в работе с большими данными, потому что помогает уменьшить размерность данных, ускорить вычисления и снизить требования к памяти, при этом сохраняя структуру данных

**Задание 2 - 2 балла**

Опишите алгоритм градиентного спуска, зачем он нужен в машинном обучении?

**Ответ:**

Градиентный спуск – это метод оптимизации, который используется для минимизации функции потерь. Он нужен в машинном обучении, чтобы находить такие параметры модели, при которых ошибка предсказания минимальна. Алгоритм работает так: вычисляется градиент функции потерь по параметрам, и параметры обновляются в направлении, противоположном градиенту, с шагом, определяемым скоростью обучения

**Задание 3 – 2 балла**

Напишите формулу преобразования Бокса-кокса, зачем оно нужно?

Преобразование Бокса-Кокса применяется к положительным числам и нужно для того, чтобы сделать распределение данных ближе к нормальному

Формула выглядит так:

**для λ ≠ 0:** (y^λ - 1) / λ

**для λ = 0:** ln(y)

Это помогает улучшить результаты статистического анализа и моделей, чувствительных к распределению данных, например, линейной регрессии

Однако не всегда в данных присутствуют положительные значения, поэтому в таких случаях применяют обобщение преобразования Бокса-Кокса, которое называется Йео-Джонсона, который позволяет работать как с положительными, так и с отрицательными значениями (в отличие от Бокса-Кокса, логарифмирования и прочих методов преобразования данных к нормальному распределению). Особенность преобразования Йео-Джонсона в том, что формулы у Йео-Джонсона зависят не только от параметра **λ** (равенство и неравенство двум)**,** но и от знака **y** (строго больше или меньше 0 и (больше или равно), (меньше или равно) 0)

**Задание 4 – 1 балл**

Найдите сотрудников, которые получают больше, чем их менеджеры.

+-------------+---------+

| Column Name | Type |

+-------------+---------+

| id | int |

| name | varchar |

| salary | int |

| managerId | int |

+-------------+---------+

**Ответ (SQL):**

SELECT e.name

FROM employee e

JOIN employee m ON e.managerId = m.id

WHERE e.salary > m.salary

**Ответ (pandas):**

Условно, df – наш датасет

merged = df.merge(df, left\_on=‘managerId’, right\_on=‘id’, suffixes=(‘‘, ‘\_manager’))

result = merged[merged[‘salary’] > merged[‘salary\_manager’]][[‘name’]]

Или если **через query** (что быстрее и удобнее):

merged = df.merge(df, left\_on=‘managerId’, right\_on=‘id’, suffixes=(‘‘, ‘\_manager’))

result = merged.query(‘salary > salary\_manager’)[[‘name’]]

**Задание 6 -2 балла**

Две колонии бактерий A и B были запущены в чашках Петри, каждая бактерия покрывала 50% своей чашки. Через 20 дней бактерия A стала покрывать 70% чашки, а бактерия B – 40% чашки.

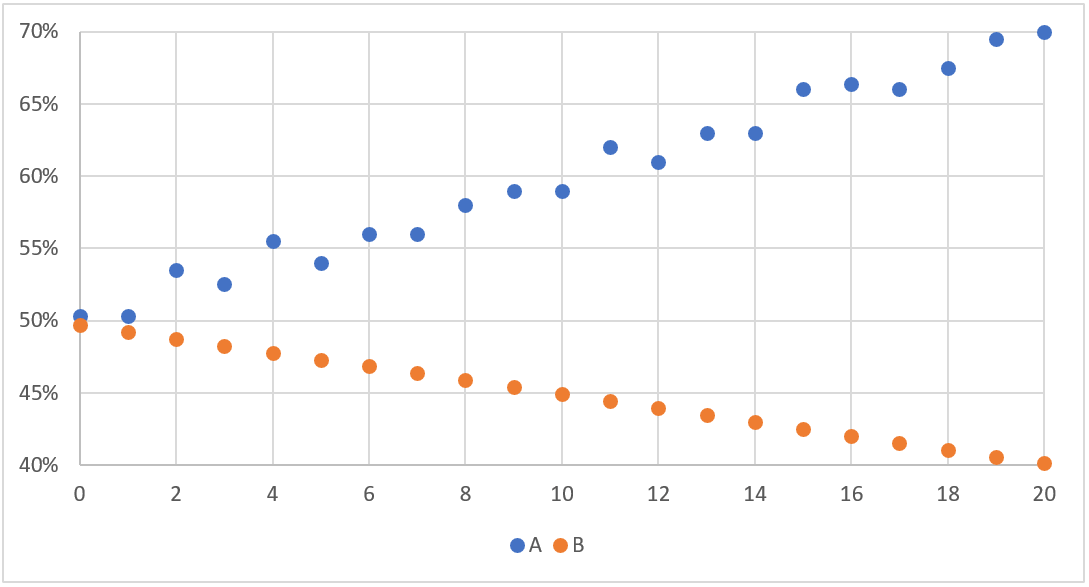
Для какой бактерии рост более линейный?

1. A
2. **B**

Судя по графику, рост более линейный у **бактерии B,** так какоранжевые точки идут более ровно вниз по прямой**,** а у бактерии А сначала чуть более медленно, затем чуть более быстро, что больше похоже на полулогарифмический или экспоненциальный рост, но не линейный

Приблизительно, какой коэффициент корреляции Пирсона между ростом бактерии B и количеством прошедших дней?

1. 1
2. [0.2, 0.6]
3. 0
4. **[-0.6, -0.2]**
5. -1



Коэффициент Пирсона точно будет отрицательным, потому что график, близкий к линейному, имеет отрицательный наклон. Я попытался всмотреться близко в график – вроде, линия не полностью ровная и есть небольшие шумы, что говорит о том, что корреляция точно не -1, значит, **ответ d ([-0.6, -0.2])**