Вариант 4

Решение отправлять **Балычевой Юлии Евгеньевне** ybalycheva@hse.ru фотографии (сканы) листов с в формате pdf, название файла: 4_Фамилия.pdf

1. Аналитики оценивали зависимость среднегодового изменения температуры воздуха в 87 регионах по 7 различным географическим факторам используя стандартную модель множественной линейной регрессии. Для оценки ковариационной матрицы вектора оценки коэффициентов была вычислена несмещённая дисперсия, которая оказалась равной 3.
   а) Найти коэффициент детерминации получившейся модели, если известно, что дисперсия показателя среднегодового изменения температуры воздуха равна 25 (1 балл).
   б) Определить значимость модели на уровне 0,2 (1 балл).

2. Используя F-критерий однофакторного дисперсионного анализа, определить есть ли зависимость между урожайностью пшеницы и климатом на уровне значимости 20%. (4 балла)

Умеренный климат	Тропический климат	Континентальный климат
5.8	6.7	5.7
3.4	6.2	6.6
3.5	3.3	7.0
4.2	3.5	3.4
	1.8	4.6
	4.8	

3. Построена модель однофакторной линейной регрессии для оценки прибыли предприятия в зависимости от инвестиций в овеществленные технологии. Используя тест Бреуша-Пагана оценить наличие гетероскедастичности на уровне значимости 5% (4 балла).

Предприятие	Инвестиции в овеществленные технологии (млн. руб.)	Прибыль (млн. руб)
ЭкоФерма Технологий	2	10
АгроИнновации	3	6
Зеленый урожай	0	7
Агропроект Сити	1	12
Натуральный век	3	15
Фермерские решения	0	10

4. По данным из предыдущей задачи построить оценку прибыли, используя взвешенный МНК, если известны оценки дисперсии ошибок для каждого наблюдения (4 балла).

$D(\varepsilon_1)$	$D(\varepsilon_2)$	$D(\varepsilon_3)$	$D(\varepsilon_4)$	$D(\varepsilon_5)$	$D(\varepsilon_6)$
5,2	19,5	3,6	0,1	19,5	3,6

5. Перед аналитиками стоит задача построить модель логистической регрессии, предсказывающей выявление заболевших туберкулёзом в общежитии мигрантов. Целевая переменная – заболевание туберкулезом (1- подтвержденный диагноз, 0 – человек здоров). На данном этапе построена модель линейной регрессии по 19 независимым переменным. Результаты приведены в таблице.

y	0	0	0	1	1	1

прогноз у	0,45	0,5	-2	0,3	0,4	4

Найти AUC (3 балла). Выбрать пороговое значение, соответствующее условию задачи, составить матрицу ошибок для выбранного порогового значения, вычислить Recall, Presicion и $F_{0,5}$. (2 балла)

6. Если возможно, привести к линейному виду модель (3 балла):

$$y = \frac{e^{-11x*z-5x-19z-22}}{e^{-2x*z+x-6z+3}}$$

7. Василию необходимо сократить размерность данных от 6 до 2. С этой целью он воспользовался методом главных компонент и вычислил собственные значения и собственные вектора. Привести двумерные данные, которые в результате получил Василий (2 балла). Найти процент дисперсии, который удалось сохранить в полученном наборе (1 балл).
   Оригинальный набор данных:

X_1	X_2	X_3	X_4	X_5	X_6
0	1	9	2	-1	30
-4	-40	9	2	-2	0
2	5	0	0	-5	-3
3	4	-1	-5	8	8
6	5	6	5	5	8

   Найденные собственные значения и соответствующие им собственные вектора:

Собственные значения	Собственные вектора
28,8	(0,11; -0,29; -0,22; -0,46; 0,73; 0,34)
426,5	(-0,15; 0,01; 0,11; -0,40; 0,24, -0,86)
3,9e-15	(-0,09; 0,00; 0,88; -0,37; -0,09; 0,27)
158,8	(-0,94; 0,21; -0,16; -0,02; 0,07; 0,19)
-1,17e-14	(-0,26; -0,93; 0,08; 0,16; -0,13; -0,16)
17,4	(-0,27; -0,08; -0,37; -0,68; -0,62,; 0,08)

8. Отдел маркетинга для анализа действия клиентов в интернет-магазине собрал следующие данные:

	Совершение покупки	Отказ от подписки	Добавление товара в корзину	Переход по ссылке рекламного баннера	Создание аккаунта	Поиск по сайту
Клиент 1	0	0	1	0	0	0
Клиент 2	1	0	0	0	1	0
Клиент 3	1	0	1	1	0	0
Клиент 4	1	1	1	1	0	1
Клиент 5	0	0	1	0	0	0
Клиент 6	0	0	0	0	0	1
Клиент 7	1	0	1	1	1	1
Клиент 8	0	1	1	0	1	1
Клиент 9	1	1	0	1	0	0

а) Вычислить значения support, confidence, lift, conviction для следующего правила: «Добавление товара в корзину без создания аккаунта не приводит к совершению покупки» Приведите интерпретацию. (2 балла)

б) Используя алгоритм Frequent Pattern Growth найти наборы, для которых support>0,3 (4 балла).

9. Используя показатель взаимной информации, оценить взаимосвязь уровней доходов и потребления населения. Привести интерпретацию полученного результата. Показатель уровня дохода разделить на 4 категории в зависимости от квантиля, а показатель уровня потребления на две категории (по медиане). (3 балла)

Доход, тыс. руб.	Потребление тыс. руб.
50	48
200	150
70	310
30	30
25	35
120	150
150	80
110	110
85	95
150	130
130	50
80	60

10. Используя метод MDI определить влияние каждого признака по энтропийному критерию. (4 балла)