Фамилия:.....

Имя:.....

Группа:.....

Задача №1

У вас имеются следующие данные (обучающая выборка):

Наблюдение	1	2	3	4	5	6	7	8	9	10
Образование	1	1	0	0	0	1	0	1	1	1
Количество детей	1	2	1	0	1	0	0	1	1	2
Брак	0	1	1	0	0	0	1	1	1	1

1. Для прогнозирования наличия образования с помощью числа детей и брака вы применяете решающее дерево, в котором разбиения осуществляются исходя из энтропии. Используя обучающую выборку постройте решающее дерево глубины 2 и нарисуйте его, указав в каждом листе значение энтропии, а также количество 1 и 0. **(15 баллов)**
   Подсказка: Изначальное разбиение переменной "Количество детей" может производиться всего двумя способами: > 0 и > 1.5.

2. Вы хотите оценить качество вашей модели на тестовой выборке из пяти наблюдений:

Наблюдение	1	2	3	4	5
Образование	0	1	0	1	0
Количество детей	2	1	0	0	0
Брак	1	0	0	0	1

   Используя построенные на обучающей выборке деревья на основании F1-метрики, посчитанной на тестовой выборке, сделайте выбор между деревьями глубины 1 и 2 (считайте, что дерево глубины 1 также строится на основании критерия энтропии). **(10 баллов)**
   Примечание: При определении прогнозных значений переменных используйте порог 0.5:

$$\hat{Y}_i = I\left(\hat{P}\left(Y_i = 1 | X_i\right) \geq 0.5\right).$$

3. Сравните точность прогноза (accuracy) выбранного в предыдущем пункте дерева на обучающей и на тестовой выборке. Сделайте вывод относительно возможного переобучения модели. **(5 баллов)**

4. Используя S-learner, на обучающей выборке оцените средний эффект воздействия наличия брака на образование. В качестве метода оценивания условных математических ожиданий воспользуйтесь решающим деревом, выбранным в предыдущем пункте. **(10 баллов)**

5. Дисперсия прогноза решающего дерева равна 2. При использовании бэггинга из 5 таких решающих деревьев дисперсия прогноза всего ансамбля равняется 1. Определите, чему равняется дисперсия прогноза всего ансамбля при использоваии 10 таких решающих деревьев. **(20 баллов)**
   Примечание: для решения этой задачи не используются данные из условия.

Решение:

1. Представим расчеты энтропии и средней энтропии в форме следующей таблицы:

Переменная	Значение	Число наблюдений	Число единиц	Доля единиц	Энтропия
Количество детей	0	3	1	0.33	0.918
	1 и 2	7	5	0.714	0.863
Средняя энтропия: $0.3 \cdot 0.918 + 0.7 \cdot 0.863 = 0.8795$					
Количество детей	0 и 1	8	4	0.5	1
	2	2	2	1	0
Средняя энтропия: $0.8 \cdot 1 + 0.2 \cdot 0 = 0.8$					
Брак	0	4	2	0.5	1
	1	6	4	0.67	0.918
Средняя энтропия: $0.4 \cdot 1 + 0.6 \cdot 0.918 = 0.9508$					

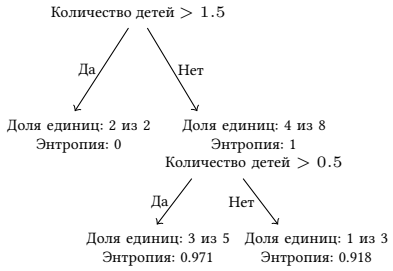
Таким образом, вначале мы разделяем данные в соответствии со значением переменной "Количество детей": те, у кого переменная принимает значение 0 или 1 и все остальные. Обратим внимание, что для всех индивидов, имеющих ровно двоих детей, переменная "Образование" принимает значение 1. Таким образом, дальнейшее разбиение в данной ветви не имеет смысла.

Рассмотрим возможные разбиения во второй ветви:

Переменная	Значение	Число наблюдений	Число единиц	Доля единиц	Энтропия
Количество детей	0	3	1	0.33	0.918
	1	5	3	0.6	0.971
Средняя энтропия: $0.375 \cdot 0.918 + 0.625 \cdot 0.971 = 0.9511$					
Брак	0	4	2	0.5	1
	1	4	2	0.5	1
Средняя энтропия: $0.5 \cdot 1 + 0.5 \cdot 1 = 1$					

Таким образом, на данном шаге мы проводим разбиение также по переменной "Количество детей".

В итоге, наше дерево будет иметь вид:



2. Отметим, что дерево глубины 1 будет отличаться от дерева глубины 2 отсутсвием последних листов. Внесем в таблицу прогнозы, полученные при помощи каждого из деревьев, а также остальные промежуточные показатели, необходимые для расчета метрик.

Наблюдение	1	2	3	4	5
Образование	0	1	0	1	0
Количество детей	2	1	0	0	0
Брак	1	0	0	0	1
Дерево глубины 1					
$\hat{P}(Y_i = 1 \mid X_i)$	1	0.5	0.5	0.5	0.5
$\hat{Y}_i$	1	1	1	1	1
Тип прогноза	FP	TP	FP	TP	FP
$\text{precision} = \frac{2}{3+2} = 0.4$					
$\text{recall} = \frac{2}{2} = 1$					
$\text{F1} = \frac{2 \cdot 0.4 \cdot 1}{0.4+1} \approx 0.57$					
Дерево глубины 2					
$\hat{P}(Y_i = 1 \mid X_i)$	1	$\frac{3}{5}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$
$\hat{Y}_i$	1	1	0	0	0
Тип прогноза	FP	TP	TN	FN	TN
$\text{precision} = \frac{1}{1+1} = 0.5$					
$\text{recall} = \frac{1}{1+1} = 0.5$					
$\text{F1} = \frac{2 \cdot 0.5 \cdot .0.5}{0.5+0.5} = 0.5$					

Таким образом, согласно данному критерию, дерево глубины 1 предпочтительнее дерева глубины 2.

3. На тестовой выборке $\text{ACC}_{\text{test}} = \frac{2}{5} = 0.4$, в то время как на обучающей $\text{ACC}_{\text{train}} = \frac{6}{10} = 0.6$. Точность на обучающей выборке превышает точность на тестовой, что может говорить о возможном переобучении модели.

4. Построенное решающее дерево не зависит от переменной, отвечающей за брак. Используя данное дерево, мы получим одинаковые оценки условных вероятностей получения образования по каждому наблюдению для обоих значений переменной брака.

   Таким образом, оценка среднего эффекта воздействия брака на образование будет равняться $\widehat{ATE} = 0$.

5. Дисперсия прогноза ансамбля из k деревьев равняется $\rho\sigma^2 + \frac{(1-\rho)\sigma^2}{k}$, где ρ обозначает корреляцию между прогнозами, а σ^2 – дисперсию прогноза одного дерева. Тогда $\rho \cdot 2 + \frac{(1-\rho)\cdot 2}{5} = 1$, откуда $\rho = \frac{3}{8}$. Соответственно, дисперсия прогноза ансамбля из десяти деревьев будет равняться:

$$\frac{3}{8} \cdot 2 + \frac{\left(1 - \frac{3}{8}\right) \cdot 2}{10} = 0.875.$$

Задача №2

У вас имеются следующие данные:

Наблюдение i	1	2	3	4	5	6
X_i	$\ln 1$	$\ln 2$	$\ln 3$	$\ln 4$	$\ln 5$	$\ln 6$
T_i	1	0	0	1	0	1
Y_i	10	50	30	20	70	60

1. Вы использовали два метода прогнозирования переменной воздействия T_i: метод ближайших соседей с 3 соседями и логистическую регрессию **без константы**. Метод ближайших соседей был обучен на имеющихся у вас данных, а логистическая регрессия - на некоторых других. Оба метода дают одинаковые оценки условной вероятности для второго наблюдения $P\left(T_2 = 1 | X_2\right)$. Определите оценку параметра логистической регрессии. **(15 баллов)**
   Подсказка: При внутривыборочном прогнозировании методом ближайших соседей само наблюдение рассматривается в качестве ближайшего соседа самого себя.

2. Используя логистическую регрессию, определенную в предыдущем пункте, получите прогнозные значения условных вероятностей $P\left(T_i = 1 | X_i\right)$ для всех наблюдений. **(5 баллов)**
   Подсказка: Считать значения логарифмов нет необходимости.

3. С помощью метода взвешивания на обратные вероятности (inverse probability weighting), определите оценку среднего эффекта воздействия переменной воздействия T_i на целевую переменную Y_i. **(20 баллов)**

Решение:

1. Ближайшими соседями для второго наблюдения являются первое, второе и третье наблюдения. Таким образом, $\hat{P}\left(T_2 = 1|X_2\right) = \frac{1}{3}$. Оценка соответствующей условной вероятности, полученная методом логистической регрессии, будет иметь вид:

$$\hat{P}\left(T_2 = 1|X_2\right) = \frac{1}{1 + e^{-\hat{\beta}X_2}} = \frac{1}{1 + e^{-\hat{\beta}ln2}} = \frac{1}{1 + 2^{-\hat{\beta}}} = \frac{1}{3}.$$

   Отсюда $2^{-\hat{\beta}} = 2$, а значит $\hat{\beta} = -1$.

2. Обратим внимание, что для i-го наблюдения

$$\hat{P}\left(T_i = 1|X_i\right) = \frac{1}{1 + e^{-(-1)\ln i}} = \frac{1}{1 + i}.$$

   Посчитаем и внесем в таблицу соответствующие прогнозы:

№	1	2	3	4	5	6
X_i	$\ln 1$	$\ln 2$	$\ln 3$	$\ln 4$	$\ln 5$	$\ln 6$
T_i	1	0	0	1	0	1
Y_i	10	50	30	20	70	60
$\hat{P}\left(T_i = 1|X_i\right)$	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{1}{4}$	$\frac{1}{5}$	$\frac{1}{6}$	$\frac{1}{7}$

3. Для оценивания условных вероятностей воспользуемся логистической регрессией[1]. Внесем в таблицу некоторые промежуточные расчеты, необходимые для получения оценки эффекта воздействия.

№	1	2	3	4	5	6
X_i	$\ln 1$	$\ln 2$	$\ln 3$	$\ln 4$	$\ln 5$	$\ln 6$
T_i	1	0	0	1	0	1
Y_i	10	50	30	20	70	60
$\hat{P}\left(T_i = 1|X_i\right)$	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{1}{4}$	$\frac{1}{5}$	$\frac{1}{6}$	$\frac{1}{7}$
$1 - \hat{P}\left(T_i = 1|X_i\right)$	$\frac{1}{2}$	$\frac{2}{3}$	$\frac{3}{4}$	$\frac{4}{5}$	$\frac{5}{6}$	$\frac{6}{7}$
$T_i Y_i$	10	0	0	20	0	60
$\left(1 - T_i\right) Y_i$	0	50	30	0	70	0
$\frac{T_i Y_i}{\hat{P}(T_i=1\|X_i)}$	20	0	0	100	0	420
$\frac{(1-T_i)Y_i}{1-\hat{P}(T_i=1\|X_i)}$	0	75	40	0	84	0

   Таким образом, оценка среднего эффекта воздействия переменной воздействия T_i на целевую переменную Y_i равняется:

$$\widehat{ATE} = \frac{(20 + 100 + 420) - (75 + 40 + 84)}{6} \approx 56.83$$

   .

[1]Поскольку в условии нет требований по поводу использования конкретного метода оценивания условных вероятностей, то в качестве альтернативы можно было бы, применить, например, метод ближайших соседей.