Московский государственный университет
имени М. В. Ломоносова
Факультет вычислительной математики и кибернетики

Задание 2.
Базовые модели

Отчёт
О выполненном задании

Выполнил:
студент 523 группы
Латыпов Ш. И.
latypovshamil2001@gmail.com

Москва
2023

Описание задачи

Выполняется на данных "Термодинамика". Данные разбить на обучение и контроль. Отдельные задачи МО даны как пары (предикторы, цели):

1. PRES TEMPC ==> PHST
2. PRES TEMPC ==> SLIQ
3. PRES TEMPC ZMF$1 ZMF$2 ZMF$3 ZMF$4 ==> PHST
4. PRES TEMPC ZMF$1 ZMF$2 ZMF$3 ZMF$4 ==> SLIQ SGAS
5. PRES TEMPC ZMF$1 ZMF$2 ZMF$3 ZMF$4 SLIQ SGAS ==> DLIQ VISLIQ
6. PRES TEMPC ZMF$1 ZMF$2 ZMF$3 ZMF$4 SLIQ SGAS ==> DGAS VISGAS
7. PRES TEMPC ZMF$1 ZMF$2 ZMF$3 ZMF$4 ==> XMF$1 XMF$2 XMF$3 XMF$4
8. PRES TEMPC ZMF$1 ZMF$2 ZMF$3 ZMF$4 ==> YMF$1 YMF$2 YMF$3 YMF$4

Для каждой задачи:

- указать тип задачи (классификация, регрессия...)
- аргументированно предложить показатели качества (главный, дополнительные)
- из библиотеки scikit-learn аргументированно выбрать модель
- обучить модель (при этом обеспечить соответствие главному показателю качества)
- отчитаться о качестве по всем показателям на обучении и контроле

Данные из файла "ТемпИзменяется.data"

PRES TEMPC ==> PHST

Так как столбец PHST отражает количество фаз и содержит дискретные значения, то задача является задачей классификации.

Нужно проанализировать распределение классов в переменной PHST:

```
import pandas as pd

file_path = 'ТемпИзменяется.data'
data = pd.read_csv(file_path, sep='\s+', engine='python')

class_distribution = data['PHST'].value_counts(normalize=True)

class_distribution
```

Вывод программы:

```
1    0.7884
2    0.2116
Name: PHST, dtype: float64
```

Показатели качества:

- Основным показателем качества будет точность (accuracy).
- Дополнительные показатели включают точность по классам (precision), полноту (recall) и F1-меру для каждого класса, а также общую F1-меру.

Выбор модели из scikit-learn:

- Используется 'GradientBoostingClassifier', так как он хорошо работает для табличных данных.

После разбиение всего датасета на тренировочные и тестовые наборы (8000 строк на тренировочные, 2000 строк на тестовые) с применением стратификации, обучения модели получены результаты оценки:

	accuracy	precision	recall	f1-score
На обучающем наборе данных	79.08%	83.47%	79.08%	70.17%
На тестовом наборе данных	79.15%	62.79%	79.15%	70.03%

После стратифицированного разделения данных, распределение классов в обучающей и тестовой выборках сохранило пропорцию классов исходного набора данных:

В обучающей выборке:

- Класс 1: 78.84%
- Класс 2: 21.16%

В тестовой выборке:

- Класс 1: 78.85%
- Класс 2: 21.15%

PRES TEMPC ==> SLIQ

Сначала посмотрим уникальные значения SLIQ, чтобы определить тип задачи:

```
unique_sliq_values = data['SLIQ'].unique()

sliq_is_discrete = pd.api.types.is_integer_dtype(data['SLIQ'])

print(unique_sliq_values)
print(sliq_is_discrete)
```

Вывод:

```
[1.   0.023083 0.194197 ... 0.033277 0.172826 0.206979]
False
```

Значения непрерывные, значит задача линейной регрессии.

Показатели качества:

- Средняя абсолютная ошибка (MAE) - средняя абсолютная разница между предсказанными и истинными значениями. Показывает, насколько модель ошибается в среднем.
- Среднеквадратичная ошибка (MSE) - среднее квадратов ошибок. Чем выше этот показатель, тем больше в среднем ошибка модели.

- Коэффициент детерминации (R^2) - Какая доля дисперсии зависимой переменной объясняется независмыми переменными в модели. Чем ближе к значению 1, тем идеальнее соответствие

Модель Ridge. После обучения и оценки модели получены результаты:
На обучающем наборе данных:

- MSE = 0.1079
- MAE = 0.2506
- R^2 = 0.0026

На тестовом наборе данных:

- MSE = 0.1092
- MAE = 0.2529
- R^2 = 0.0031

Также была протестирована модель TweedieRegressor, но результаты в ней оказалисб примерно такие же.

PRES TEMPC ZMF$1 ZMF$2 ZMF$3 ZMF$4 ==> PHST

Задача классификации. Используется та же модель GradientBoostingClassifier.
Результаты:

	accuracy	precision	recall	f1-score
На обучающем наборе данных	99.66%	99.66%	99.66%	99.66%
На тестовом наборе данных	97.70%	97.72%	97.70%	97.71%

PRES TEMPC ZMF$1 ZMF$2 ZMF$3 ZMF$4 ==> SLIQ SGAS

Задача многозадачной регрессии. Необходима модель, которая поддерживает многомерный вывод. Для этого подойдет MultiOutputRegressor, которая позволяет обернуть стандартную модель для их применения в многомерной регрессии. Используя GradientBoostingRegressor как базовую модель, получены результаты по показателям MAE, MSE и R^2:
Результаты:
На обучающем наборе данных:

1. Для SLIQ:
    - MSE = 0.0032
    - MAE = 0.0203
    - R^2 = 0.9704
2. Для SGAS:
    - MSE = 0.0032
    - MAE = 0.0203
    - R^2 = 0.9704

На тестовом наборе данных:

1. Для SLIQ:
    - MSE = 0.0052
    - MAE = 0.0242
    - R^2 = 0.9526
2. Для SGAS:
    - MSE = 0.0052
    - MAE = 0.0242
    - R^2 = 0.9526

PRES TEMPC ZMF$1 ZMF$2 ZMF$3 ZMF$4 SLIQ SGAS ==> DLIQ VISLIQ

Задача многозадачной регрессии.

Так как в данных присутствуют строки, в котоорых не записаны никакие данные и имеют значение Nan, необходимо дополнитеьно изменить таблицу с данными и удалить оттуда "пустые"строки.

Показатели MAE, MSE, R^2. Модель MultiOutputRegressor вместе с GradientBoostingRegressor как базовой.

Результаты:

На обучающем наборе данных:

1. Для DLIQ:
    - MSE = 83.7681
    - MAE = 6.1606
    - R^2 = 0.9936
2. Для VISLIQ:
    - MSE = 0.000087
    - MAE = 0.0067
    - R^2 = 0.9950

На тестовом наборе данных:

1. Для DLIQ:
    - MSE = 206.0809
    - MAE = 9.1956
    - R^2 = 0.9846
2. Для VISLIQ:
    - MSE = 0.000170
    - MAE = 0.0091
    - R^2 = 0.9900

PRES TEMPC ZMF$1 ZMF$2 ZMF$3 ZMF$4 SLIQ SGAS ==> DGAS VISGAS

Задача многозадачной регрессии.

Показатели MAE, MSE, R^2. Модель MultiOutputRegressor вместе с GradientBoostingRegressor как базовой.

Результаты:

На обучающем наборе данных:

1. Для DGAS:
   - MSE = 15.1814
   - MAE = 2.5766
   - R^2 = 0.9877
2. Для VISGAS:
   - MSE = 0.000000087
   - MAE = 0.000167
   - R^2 = 0.9814

На тестовом наборе данных:

1. Для DGAS:
   - MSE = 41.0285
   - MAE = .8046
   - R^2 = 0.9697
2. Для VISGAS:
   - MSE = 0.000000282
   - MAE = 0.000261
   - R^2 = 0.9478

PRES TEMPC ZMF$1 ZMF$2 ZMF$3 ZMF$4 ==> XMF$1 XMF$2 XMF$3 XMF$4

Задача многозадачной регрессии.

Показатели MAE, MSE, R^2. Модель MultiOutputRegressor вместе с GradientBoostingRegressor как базовой.

Результаты:

На обучающем наборе данных:

	XMF$1	XMF$2	XMF$3	XMF$4
MSE	0.000027	0.000013	0.000103	0.000224
MAE	0.0037	0.0023	0.0065	0.0107
R^2	0.9915	0.9994	0.9973	0.9957

На тестовом наборе данных:

	XMF$1	XMF$2	XMF$3	XMF$4
MSE	0.000070	0.000036	0.000258	0.000526
MAE	0.0054	0.0037	0.0097	0.0155
R^2	0.9798	0.9984	0.9929	0.9899

PRES TEMPC ZMF$1 ZMF$2 ZMF$3 ZMF$4 ==> YMF$1 YMF$2 YMF$3 YMF$4

Задача многозадачной регрессии.
Показатели MAE, MSE, R^2. Модель RandomForestRegression.
Результаты:

На обучающем наборе данных:

	YMF$1	YMF$2	YMF$3	YMF$4
MSE	0.0000415	0.0000066	0.0000037	0.000000024
MAE	0.0044	0.0016	0.0012	0.000090
R^2	0.9971	0.9995	0.9971	0.9909

На тестовом наборе данных:

	YMF$1	YMF$2	YMF$3	YMF$4
MSE	0.0000991	0.0000322	0.0000096	0.000000165
MAE	0.0067	0.0028	0.0018	0.000154
R^2	0.9931	0.9977	0.9919	0.9552

Данные из файла "ТемпПостоянная.data"

PRES TEMPC ==> PHST

Выбор модели из scikit-learn:

- Используется 'GradientBoostingClassifier', так как он хорошо работает для табличных данных.

После разбиение всего датасета на тренировочные и тестовые наборы (8000 строк на тренировочные, 2000 строк на тестовые), обучения модели получены результаты оценки:

	accuracy	precision	recall	f1-score
На обучающем наборе данных	77.99%	82.85%	77.99%	68.66%
На тестовом наборе данных	78.05%	68.31%	78.05%	68.57%

PRES TEMPC ==> SLIQ

Задача линейной регрессии.
Показатели качества:

- Средняя абсолютная ошибка (MAE)
- Среднеквадратичная ошибка (MSE)
- Коэффициент детерминации (R^2)

Модель Ridge. После обучения и оценки модели получены результаты:
На обучающем наборе данных:

- MSE = 0.1072
- MAE = 0.2508
- R^2 = 0.0038

На тестовом наборе данных:

- MSE = 0.1091
- MAE = 0.2537
- R^2 = 0.00015

Также была протестирована модель TweedieRegressor, но результаты в ней оказались примерно такие же.

PRES TEMPC ZMF$1 ZMF$2 ZMF$3 ZMF$4 ==> PHST

Задача классификации. Используется та же модель GradientBoostingClassifier.
Результаты:

	accuracy	precision	recall	f1-score
На обучающем наборе данных	99.75%	99.75%	99.75%	99.75%
На тестовом наборе данных	98.55%	98.56%	98.55%	98.55%

PRES TEMPC ZMF$1 ZMF$2 ZMF$3 ZMF$4 ==> SLIQ SGAS

Задача многозадачной регрессии. Необходима модель, которая поддерживает многомерный вывод. Для этого подойдет MultiOutputRegressor, которая позволяет обернуть стандартную модель для их применения в многомерной регрессии. Используя GradientBoostingRegressor как базовую модель, получены результаты по показателям MAE, MSE и R^2:

Результаты:

На обучающем наборе данных:

1. Для SLIQ:
   - MSE = 0.00267
   - MAE = 0.0175
   - R^2 = 0.97515
2. Для SGAS:
   - MSE = 0.00267
   - MAE = 0.0175
   - R^2 = 0.97515

На тестовом наборе данных:

1. Для SLIQ:
   - MSE = 0.00415
   - MAE = 0.02134
   - R^2 = 0.96196
2. Для SGAS:
   - MSE = 0.00415
   - MAE = 0.02134
   - R^2 = 0.96196

PRES TEMPC ZMF$1 ZMF$2 ZMF$3 ZMF$4 SLIQ SGAS ==> DLIQ VISLIQ

Задача многозадачной регрессии.

Так как в данных присутствуют строки, в котоорых не записаны никакие данные и имеют значение Nan, необходимо дополнитеьно изменить таблицу с данными и удалить оттуда "пустые"строки.

Показатели MAE, MSE, R^2. Модель MultiOutputRegressor вместе с GradientBoostingRegressor как базовой.

Результаты:

На обучающем наборе данных:

1. Для DLIQ:
   - MSE = 77.2226
   - MAE = 5.6298
   - R^2 = 0.9947

2. Для VISLIQ:

   - MSE = 0.0000664
   - MAE = 0.00558
   - R^2 = 0.9962

На тестовом наборе данных:

1. Для DLIQ:

   - MSE = 138.256
   - MAE = 7.946
   - R^2 = 0.9898

2. Для VISLIQ:

   - MSE = 0.0001557
   - MAE = 0.00829
   - R^2 = 0.9909

PRES TEMPC ZMF$1 ZMF$2 ZMF$3 ZMF$4 SLIQ SGAS ==> DGAS VISGAS

Задача многозадачной регрессии.

Показатели MAE, MSE, R^2. Модель MultiOutputRegressor вместе с GradientBoostingRegressor как базовой.

Результаты:

На обучающем наборе данных:

1. Для DGAS:

   - MSE = 11.1971
   - MAE = 2.1874
   - R^2 = 0.9915

2. Для VISGAS:

   - MSE = 0.00000007
   - MAE = 0.000152
   - R^2 = 0.9860

На тестовом наборе данных:

1. Для DGAS:

   - MSE = 32.049
   - MAE = 3.1195
   - R^2 = 0.9754

2. Для VISGAS:

   - MSE = 0.000000238
   - MAE = 0.000221
   - R^2 = 0.9557

PRES TEMPC ZMF$1 ZMF$2 ZMF$3 ZMF$4 ==> XMF$1 XMF$2 XMF$3 XMF$4

Задача многозадачной регрессии.

Показатели MAE, MSE, R^2. Модель MultiOutputRegressor вместе с GradientBoostingRegressor как базовой.

Результаты:

На обучающем наборе данных:

	XMF$1	XMF$2	XMF$3	XMF$4
MSE	0.0000183	0.0000083	0.0000953	0.000216
MAE	0.00289	0.00174	0.0063	0.01
R^2	0.9939	0.9997	0.9976	0.996

На тестовом наборе данных:

	XMF$1	XMF$2	XMF$3	XMF$4
MSE	0.0000387	0.0000298	0.000218	0.00043
MAE	0.00412	0.0029	0.0094	0.0142
R^2	0.9878	0.9989	0.9947	0.992

PRES TEMPC ZMF$1 ZMF$2 ZMF$3 ZMF$4 ==> YMF$1 YMF$2 YMF$3 YMF$4

Задача многозадачной регрессии.

Показатели MAE, MSE, R^2. Модель RandomForestRegression.

Результаты:

На обучающем наборе данных:

	YMF$1	YMF$2	YMF$3	YMF$4
MSE	0.0000269	0.00000447	0.00000205	0.000000019
MAE	0.00362	0.00123	0.0009	0.000071
R^2	0.9982	0.9997	0.9982	0.9931

На тестовом наборе данных:

	YMF$1	YMF$2	YMF$3	YMF$4
MSE	0.0000532	0.000015	0.00000597	0.00000015
MAE	0.0047	0.002	0.00137	0.000114
R^2	0.9962	0.9989	0.9951	0.9602