Московский государственный технический университет им. Н. Э. Баумана

Курс «Технологии машинного обучения»

Отчёт по рубежному контролю №1

«Технологии разведочного анализа и обработки данных.»

Вариант № 16

Выполнил:
Пронченко Т.А.
группа ИУ5-62Б

Проверил:
Гапанюк Ю.Е.

Дата: 14.04.25

Подпись:

Дата:

Подпись:

Москва, 2025 г.

Полученное задание

Номер варианта: **16**

Номер задачи: **2**

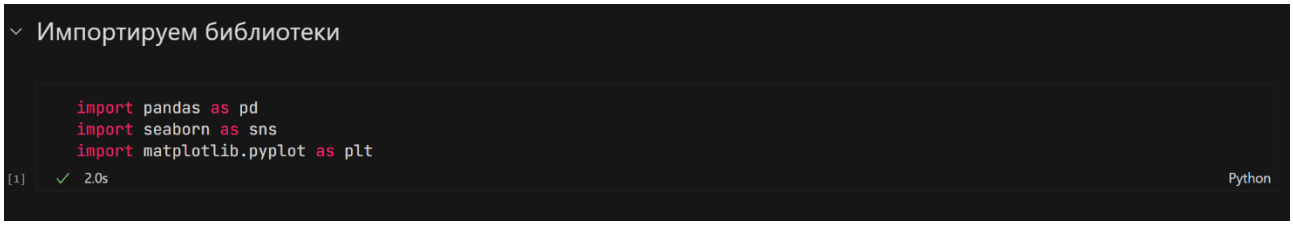
1. Номер набора данных, указанного в задаче: **8**

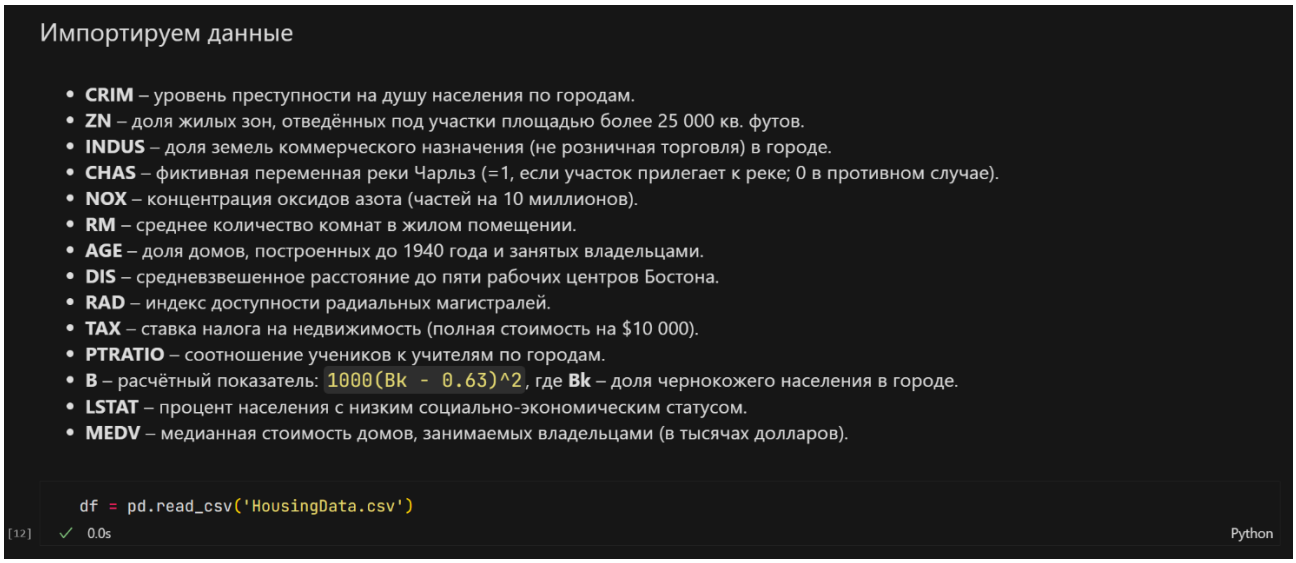
   https://www.kaggle.com/datasets/altavish/boston-housing-dataset

Задача №2.

Для заданного набора данных проведите обработку пропусков в данных для одного категориального и одного количественного признака. Какие способы обработки пропусков в данных для категориальных и количественных признаков Вы использовали? Какие признаки Вы будете использовать для дальнейшего построения моделей машинного обучения и почему?

Ход выполнения





Основные характеристики датасета

```
df.head()
```

[13] ✓ 0.0s Python

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV
0	0.00632	18.0	2.31	0.0	0.538	6.575	65.2	4.0900	1	296	15.3	396.90	4.98	24.0
1	0.02731	0.0	7.07	0.0	0.469	6.421	78.9	4.9671	2	242	17.8	396.90	9.14	21.6
2	0.02729	0.0	7.07	0.0	0.469	7.185	61.1	4.9671	2	242	17.8	392.83	4.03	34.7
3	0.03237	0.0	2.18	0.0	0.458	6.998	45.8	6.0622	3	222	18.7	394.63	2.94	33.4
4	0.06905	0.0	2.18	0.0	0.458	7.147	54.2	6.0622	3	222	18.7	396.90	NaN	36.2

```
# Размер датасета - 506 строк, 14 столбцов
df.shape
```

[14] ✓ 0.0s Python

```
(506, 14)
```

В качестве категориального признака выберем бинарный признак «CHAS» - переменная реки Чарльз, указывающая прилегает ли участок к реке. В качестве количественного признака выберем признак «LSTAT» - процент населения с низким социально-экономическим статусом.

Проверим признаки на пропуски

```
print("Пропуски до обработки:")
print(df[['CHAS', 'LSTAT']].isnull().sum())
```

[21] ✓ 0.0s Python

```
Пропуски до обработки:
CHAS     20
LSTAT    20
dtype: int64
```

Заполним пропуски

```
# Заполнение пропусков в CHAS (категориальный)
if df['CHAS'].isnull().sum() > 0:
    df.fillna({'CHAS': df['CHAS'].mode()[0]}, inplace=True)
    print("Пропуски в CHAS заполнены модой.")
else:
    print("Пропусков в CHAS нет.")

# Заполнение пропусков в LSTAT (количественный)
if df['AGE'].isnull().sum() > 0:
    df.fillna({'LSTAT': df['LSTAT'].median()}, inplace=True)
    print("Пропуски в LSTAT заполнены медианой.")
else:
    print("Пропусков в LSTAT нет.")
```

[23] ✓ 0.0s Python

```
Пропуски в CHAS заполнены модой.
Пропуски в LSTAT заполнены медианой.
```

```
# Проверка после обработки
print("\nПропуски после обработки:")
print(df[['CHAS', 'LSTAT']].isnull().sum())
```

[24] ✓ 0.0s Python

```
Пропуски после обработки:
CHAS     0
LSTAT    0
dtype: int64
```

Какие способы обработки пропусков в данных для категориальных и количественных признаков вы использовали? Какие признаки Вы будете использовать для дальнейшего построения моделей машинного обучения и почему?

1) Признак CHAS является бинарной переменной, принимающей значения 0 или 1, где 1 обозначает расположение недвижимости вблизи реки Чарльз, а 0 - отсутствие такого соседства. Для обработки пропущенных значений в данном признаке могут быть применены следующие методы:

1. **Замена на моду (наиболее частое значение)**
   Данный подход целесообразен при небольшом количестве пропусков и значительном преобладании одного из значений.

2. **Создание новой категории "Unknown"**
   При существенном количестве пропусков рекомендуется введение дополнительной категории.

3. **Предсказание пропущенных значений**
   При наличии значительного числа пропусков возможно применение простых моделей для прогнозирования отсутствующих значений на основе других признаков.

Т.к. в нашем случае пропуски составили менее 5% и значение 0 преобладает замена на моду будет самым оптимальным способом.

2) Признак LSTAT отражает процент населения с низким социально-экономическим статусом. Методы обработки пропусков:

1. **Замена на медиану**
   Медиана является устойчивой к выбросам мерой центральной тенденции

2. **Замена на среднее значение**
   При нормальном распределении данных и отсутствии выбросов допустимо использование среднего

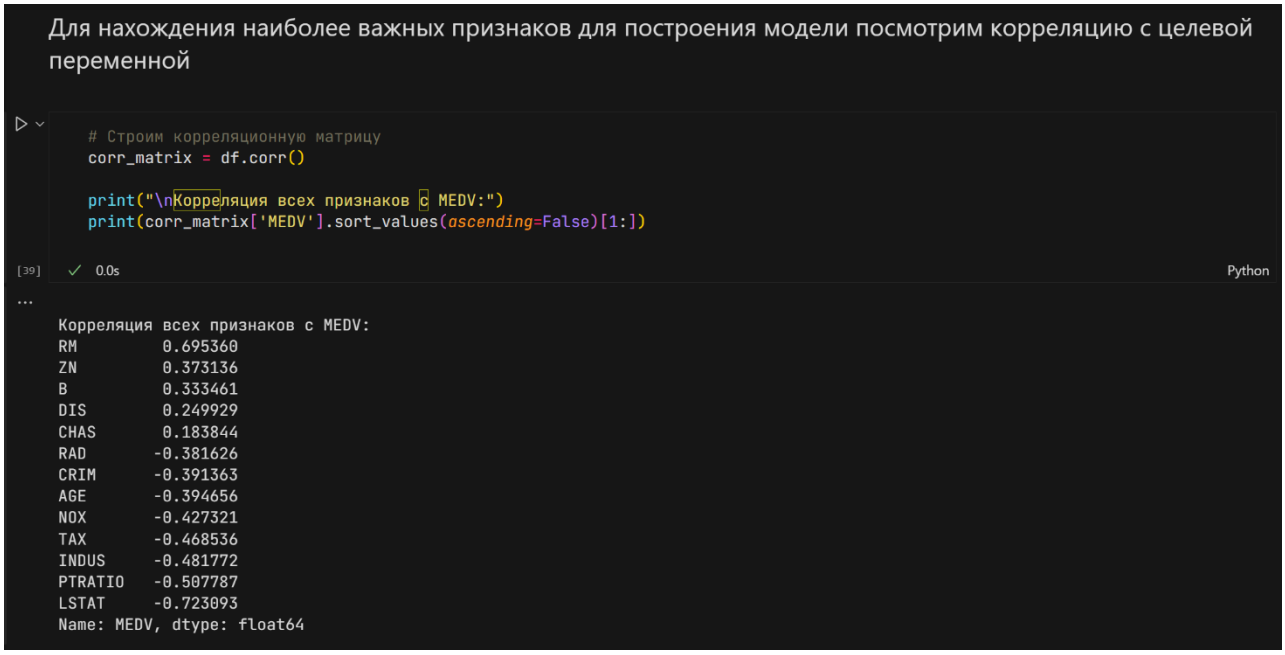
3. **Замена на константу**
   В некоторых случаях применяется заполнение фиксированным значением (0 или -1), однако это может исказить данные.

4. **Использование KNN Imputer**
   Для большого числа пропусков эффективен алгоритм k-ближайших соседей:

Т.к. в нашем случае пропусков немного, и можно заметить выбросы на диаграмме, наиболее оптимальным способом будет замена на медиану.

3) Выбор признаков для построения модели машинного обучения

Для датасета Boston Housing (CRIM, ZN, INDUS, CHAS, NOX, RM, AGE, DIS, RAD, TAX, PTRATIO, B, LSTAT, MEDV) отбор наиболее информативных признаков осуществлялся на основе корреляции с целевой переменной (MEDV - стоимость жилья).



Наиболее значимые признаки:

- **LSTAT** (% населения с низким статусом): обратная зависимость от стоимости
- **RM** (среднее число комнат): сильная положительная корреляция с ценой
- **PTRATIO** (соотношение учеников и учителей): влияет на привлекательность района
- **INDUS** (доля нежилых площадей): обратная зависимость от стоимости - чем больше промышленных зон, тем ниже цены на жилье
- **TAX** (налог на имущество): высокая налоговая нагрузка снижает стоимость недвижимости

Несмотря на то, что некоторые признаки выделяются на фоне других, в нашем случае для построения моделей машинного лучше использовать все признаки, т.к. каждый из них все равно оказывает значительное влияние на стоимость участка.