РК 1. Харитонов А.А. ИУ5-64Б вариант 16

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

```
import pandas as pd
import numpy as np
pd.options.mode.chained_assignment = None
df = pd.read_csv('HousingData.csv')
df.info()
df.head()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 506 entries, 0 to 505
Data columns (total 14 columns):
 #   Column   Non-Null Count  Dtype  
---  ------   --------------  -----  
 0   CRIM     486 non-null    float64
 1   ZN       486 non-null    float64
 2   INDUS    486 non-null    float64
 3   CHAS     486 non-null    float64
 4   NOX      506 non-null    float64
 5   RM       506 non-null    float64
 6   AGE      486 non-null    float64
 7   DIS      506 non-null    float64
 8   RAD      506 non-null    int64  
 9   TAX      506 non-null    int64  
 10  PTRATIO  506 non-null    float64
 11  B        506 non-null    float64
 12  LSTAT    486 non-null    float64
 13  MEDV     506 non-null    float64
dtypes: float64(12), int64(2)
memory usage: 55.5 KB
```

```
      CRIM    ZN  INDUS  CHAS    NOX     RM   AGE     DIS  RAD  TAX
PTRATIO  \
0  0.00632  18.0   2.31   0.0  0.538  6.575  65.2  4.0900    1  296
15.3
1  0.02731   0.0   7.07   0.0  0.469  6.421  78.9  4.9671    2  242
17.8
2  0.02729   0.0   7.07   0.0  0.469  7.185  61.1  4.9671    2  242
17.8
```

```
3  0.03237   0.0   2.18   0.0  0.458  6.998  45.8  6.0622    3  222
18.7
4  0.06905   0.0   2.18   0.0  0.458  7.147  54.2  6.0622    3  222
18.7

        B  LSTAT  MEDV
0  396.90   4.98  24.0
1  396.90   9.14  21.6
2  392.83   4.03  34.7
3  394.63   2.94  33.4
4  396.90    NaN  36.2
```

Заметим, что столбец RAD содержит закодированные категориальные данные индекса доступности к радиальным магистралям. Заметим, что данные категориальный признак не имеет пропусков.

```
print('Значения признака: ', df['RAD'].unique())
print('Количество пропусков: ', df['RAD'].isna().sum())
```

```
Значения признака:  [ 1  2  3  5  4  8  6  7 24]
Количество пропусков:  0
```

Искусственно добавим пропуски в признак RAD

```
omission_count = 20
df_len = df['RAD'].count()
for i in range(0, omission_count):
    df['RAD'][i * df_len // omission_count] = None
print('Количество пропусков: ', df['RAD'].isna().sum())
```

```
Количество пропусков:  20
```

Заполнение пропусков в признаках CRIM и RAD

Рассмотрим количество пропусков в данных признака CRIM

```
print('Количество пропусков: ', df['CRIM'].isna().sum())
```

```
Количество пропусков:  20
```

Заменим пропуски данных в столбце CRIM на средние значения с помощью метода mean()

```
df['CRIM'] = df['CRIM'].fillna(df['CRIM'].mean())
print('Количество пропусков: ', df['CRIM'].isna().sum())
```

```
Количество пропусков:  0
```

Рассмотрим категориальный признак RAD. Пропуски в нём мы создали искусственно. Заметим, что этот признак имеет выброс данных в категории 24. Устраним его. Также введём категорию 0 в качестве идентификатора отсутствия данных признака RAD.

```
df['RAD'] = df['RAD'].replace(24, 0)
df['RAD'] = df['RAD'].replace(np.NaN, 0)
print('Значения признака: ', df['RAD'].unique())
```

```
Значения признака:  [0. 2. 3. 5. 4. 8. 6. 1. 7.]
```

Таким образом мы сохраняем данные. В случае если признак RAD не будет иметь значения при построении модели, мы сможем его откинуть, сохранить данные записей, где данный признак был пропущен