Лабораторная работа №1.
Подготовка данных для моделирования

Общая информация по работе

Цель работы –получить навык очистки данных с помощью инструментов библиотеки pandas.

Задачи:

1. Загрузить набор данных.
2. Провести предварительный анализ набора данных.
3. Получить список категориальных и числовых признаков.
4. Вывести количество объектов в наборе для всех категорий
5. Удалить строки из набора, где для каждого числового признака его значения меньше процентиля 2.5 и больше процентиля 97.5.
5. Заменить все отсутствующие значения для признака WaterHeaterSize (размер водонагревателя) значением моды.
6. Устранить отрицательные значения в столбце NumOfRoomsCooled (количество комнат с кондиционированием).
7.Удалить строки, в которых не содержится значения по типу системы отопления (HeatingEquipmentType)
8. Для атрибутов RoofType (тип крыши), StovenFuelType (тип топлива для печи) в случае их отсутствия поставить значение Unknown.
9.Закодировать значения HeatingEquipmentAge (возраст системы отопления)
10. Удалить из набора объекты с редкими категориями (чей процент встречаемости для оставшихся строк меньше 3)
11.Сохранить полученный набор в файл.

Ответ на задание необходимо предоставить в виде файла в формате IPYNB, PY, ODF, Word или PDF. Файл должен содержать следующую информацию: ФИО слушателя, код с комментариями, результаты выполнения программного кода, а также короткое заключение по работе и выводы.

Ход работы

Загрузка набора и подключение библиотек

Подключение к Google Drive

```
from google.colab import drive
drive.mount('/content/gdrive')
```

Подключение библиотек визуализации и работы с наборами данных

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

Отключение предупрежений

```
import warnings
warnings.filterwarnings('ignore')
```

Загрузка набора данных

```
df = pd.read_csv('/content/gdrive/My
Drive/CourseMaterials/homeEnergyConsumption.csv')
df.head()
```

Предварительный анализ набора данных

Получение описательных статистик

```
df.describe()
```

Получение информации об атрибутах

```
df.info()
```

Получение списков категориальных и числовых признаков. Вывод количества объектов в наборе для всех категорий

listOfCategoricalFeatures – список атрибутов категорий, listOfNumericalFeatures – список атрибутов числовых

```
#from pandas.api.types import is_string_dtype
from pandas.api.types import is_numeric_dtype

listOfCategoricalFeatures = []
listOfNumericalFeatures = []

for column in df.columns:
  if not(is_numeric_dtype(df[column])):
    print(f'Признак : {column}.')
    print('Значения и их количество в наборе')
    listOfCategoricalFeatures.append(column)
    print(df[column].value_counts())
  else:
    listOfNumericalFeatures.append(column)

listOfNumericalFeatures.remove('AirConditioning')
```

Вывод списков listOfCategoricalFeatures и listOfNumericalFeatures

```
print('Список категориальных признаков:')
print(listOfCategoricalFeatures)
print('Список числовых признаков:')
print(listOfNumericalFeatures)
```

Удаление строк из набора, где для каждого числового признака его значения меньше процентиля 2.5 и больше процентиля 97.5.

minLimit - процентиль 2.5 maxLimit - процентиль 97.5

```
for column in listOfNumericalFeatures:
    print(column)
    minLimit = df[column].quantile(0.025)
    maxLimit = df[column].quantile(0.975)
    print(minLimit,maxLimit)
    print(df.shape)
```

```
df=df.loc[(df[column]>=minLimit)&(df[column]<=maxLimit)]
```

Замена всех отсутствующих значений для признака WaterHeaterSize (размер водонагревателя) значением моды.

Дополнительно размер переведен в числовое значение, т.к. признак порядковый

```
modeWaterHeaterSize = df['WaterHeaterSize'].mode()[0]
df['WaterHeaterSize'].replace('None',modeWaterHeaterSize,inplace=True)

waterHeaterSizeDict = {'Small' : 1,'Medium' : 2,'Large': 3}
df['WaterHeaterSize'] = df['WaterHeaterSize'].map(waterHeaterSizeDict)
```

Устранение отрицательных значений в столбце NumOfRoomsCooled (количество комнат с кондиционированием)

```
df.loc[df['NumOfRoomsCooled']<0,'NumOfRoomsCooled']=0
```

Удаление строки в наборе, в которых не содержится значения по типу системы отопления (HeatingEquipmentType)

```
df = df.dropna(subset=['HeatingEquipmentType'])
```

Для атрибута RoofType (тип крыши) в случае отсутствия устанавливается значение Unknown.

```
df['RoofType']=df['RoofType'].replace(np.nan,'Unknown')
```

Для атрибута StovenFuelType (тип топлива для печи) в случае отсутствия устанавливается значение Unknown.

Другой вариант обновления значения в столбце.

```
df['StovenFuelType'].replace('-','Unknown',inplace=True)
```

Кодирование категорий HeatingEquipmentAge (возраст системы отопления) числовыми значениями

```
mainHeatingEquipmentAgeDict = {'<2 Years old':1,'2-4 Years old':2,'5-9 Years old':3, '10-14 Years old':4,'15-19 Years old':5,'>20 Years old':6}
df['HeatingEquipmentAge']=df['HeatingEquipmentAge'].map(mainHeatingEquipmentAgeDict)
```

Удалить из набора объекты с редкими категориями (чей процент встречаемости для оставшихся строк меньше 3).

Получить количество объектов по каждой категории

```
for column in listOfCategoricalFeatures:
  print(f'Признак : {column}.')
  print('Значения и их количество в наборе')
  print(df[column].value_counts())
```

Получить пороговое значение (минимальное количество образцов для каждой категории).

```
#3%
threshold = df.shape[0]*0.03
print(threshold)
```

Если количество образцов меньше порогового для категории, то строки с указанной категорией удаляются.

```
for column in listOfCategoricalFeatures:
    counts = df[column].value_counts()
    listOfIndices=counts.index[counts>threshold].tolist()
    df = df.loc[df[column].isin(listOfIndices)]
```

Сохранение полученного набора в файл

```
df.to_csv('/content/gdrive/My Drive/CourseMaterials/homeEnergyConsumption_cleaned.csv',index=False)
```