Московский Государственный Технический Университет им. Н. Э. Баумана

Факультет «Информатика и Системы управления»

Кафедра «Автоматизированные системы обработки информации и управления»

Дисциплина «Технологии машинного обучения»

Отчёт по лабораторной работе №2

«Обработка пропусков в данных, кодирование категориальных признаков, масштабирование данных.»

Выполнил:

Студент группы ИУ5ц-83Б

Костников И.А.

Преподаватель:

Гапанюк Ю.Е.

Москва, 2020 г.

1 Цель работы

Изучение способов предварительной обработки данных для дальнейшего формирования моделей.

2 Краткое описание

Подготовка данных

3 Текст программы

Текст программы представлена во втором файле

4 Экранные формы с примерами выполнения программы.

```
[4]
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

[5]
data = pd.read_csv('winemag-data-130k-v2.csv', sep=",")

[6]
print("Размер таблицы: ", data.shape)

Размер таблицы:  (129971, 14)

[7]
data.dtypes

Unnamed: 0                int64
country                  object
description              object
designation              object
points                    int64
price                   float64
province                 object
region_1                 object
region_2                 object
taster_name              object
taster_twitter_handle    object
title                    object
variety                  object
winery                   object
dtype: object
```

```
[8]
# Сумма пропущенных значений
isnull = data.isnull().sum()
print(isnull)

Unnamed: 0                  0
country                    63
description                 0
designation             37465
points                      0
price                    8996
province                   63
region_1                21247
region_2                79460
taster_name             26244
taster_twitter_handle   31213
title                       0
variety                     1

[9]
data.head()
```

	Unnamed: 0	country	description	designation	points	price	province	region_1	region_2	taster_name	taster_twitter_handle	title	variety	winery
0	0	Italy	Aromas include tropical fruit, broom, brimston...	Vulkà Bianco	87	NaN	Sicily & Sardinia	Etna	NaN	Kerin O'Keefe	@kerinokeefe	Nicosia 2013 Vulkà Bianco (Etna)	White Blend	Nicosia
1	1	Portugal	This is ripe and fruity, a wine that is smooth...	Avidagos	87	15.0	Douro	NaN	NaN	Roger Voss	@vossroger	Quinta dos Avidagos 2011 Avidagos Red (Douro)	Portuguese Red	Quinta dos Avidagos
2	2	US	Tart and snappy, the flavors of lime flesh and...	NaN	87	14.0	Oregon	Willamette Valley	Willamette Valley	Paul Gregutt	@paulgwine	Rainstorm 2013 Pinot Gris (Willamette Valley)	Pinot Gris	Rainstorm
3	3	US	Pineapple rind, lemon pith and orange blossom ...	Reserve Late Harvest	87	13.0	Michigan	Lake Michigan Shore	NaN	Alexander Peartree	NaN	St. Julian 2013 Reserve Late Harvest Riesling ...	Riesling	St. Julian
4	4	US	Much like the regular bottling from 2012, this...	Vintner's Reserve Wild Child Block	87	65.0	Oregon	Willamette Valley	Willamette Valley	Paul Gregutt	@paulgwine	Sweet Cheeks 2012 Vintner's Reserve Wild Child...	Pinot Noir	Sweet Cheeks

2. Обработка данных

2.1. Удаление значений

[10]
```
# Удаление столбцов
newdata1 = data.dropna(axis=1)
newdata1.shape
```

```
(129971, 5)
```

[11]
```
newdata1.dtypes
```

```
Unnamed: 0      int64
description    object
points          int64
title          object
winery         object
dtype: object
```

[12]
```
# Удаление строк
newdata2 = data.dropna(axis=0)
newdata2.shape
```

```
(22387, 14)
```

SimpleImputer

[16]
```
from sklearn.impute import SimpleImputer
from sklearn.impute import MissingIndicator
```

[17]
```
sort_null_data = data[data_num]
data_price = sort_null_data[['price']]
data_price.head()
```

	price
0	NaN
1	15.0
2	14.0
3	13.0
4	65.0

[18]
```
implicator = MissingIndicator()
values = implicator.fit_transform(data_price)
values
```

```
array([[ True],
       [False],
       [False],
       ...,
       [False],
       [False],
       [False]])
```

Обработка категориальных признаков

[24]
```
for key in mass3:
    print("{} - {} - ({}) {}%".format(key[0], key[1], key[2], key[3]))
```

```
country - object - (63) 0.04847%
designation - object - (37465) 28.82566%
province - object - (63) 0.04847%
region_1 - object - (21247) 16.34749%
region_2 - object - (79460) 61.13672%
taster_name - object - (26244) 20.1922%
taster_twitter_handle - object - (31213) 24.01536%
variety - object - (1) 0.00077%
```

[25]
```
sort_null_data_obj = data[data_num_obj]
data_region = sort_null_data_obj[['region_1']]
data_region.head()
```

	region_1
0	Etna
1	NaN
2	Willamette Valley
3	Lake Michigan Shore
4	Willamette Valley

Кодирование категориальных признаков

```
[69] data_frame = pd.DataFrame({'region': reion_values.T[0]})
```

```
[64] from sklearn.preprocessing import LabelEncoder, OneHotEncoder
```

```
[74] le = LabelEncoder()
     data_label_en = le.fit_transform(data_frame)
```

```
[75] data_frame['region'].unique()
```

```
array(['Etna', 'NA', 'Willamette Valley', ..., 'Del Veneto',
       'Bardolino Superiore', 'Paestum'], dtype=object)
```

```
[76] data_label_en
```

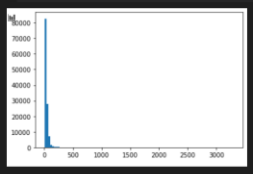
```
array([ 424,  738, 1218, ...,   21,   21,   21])
```

Маштабирование данных

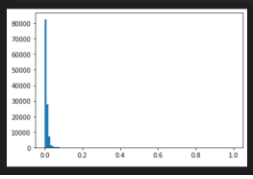
```
[88] from sklearn.preprocessing import MinMaxScaler, StandardScaler, Normalizer
```

```
[89] sc1 = MinMaxScaler()
     sc1_data = sc1.fit_transform(data[['price']])
```

```
[92] plt.hist(data[['price']], 100)
     plt.show()
```



```
[93] plt.hist(sc1_data, 100)
     plt.show()
```



5 Вывод

В данной лабораторной работе я научился обрабатывать пропуски в таблице, кодированию категориальных признаков и масштабированию данных