Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Московский государственный технический университет
имени Н.Э. Баумана
(национальный исследовательский университет)»
(МГТУ им. Н.Э. Баумана)

ФАКУЛЬТЕТ Информатика и системы управления (ИУ)

КАФЕДРА Система обработки информации и управления

ДИСЦИПЛИНА Методы машинного обучения

ОТЧЕТ ПО ЛАБОРАТОРНОЙ РАБОТЕ № 3

Обработка признаков (часть 2)

название лабораторной работы

Группа ИУ5-24М

Студент 13.04.2022 _____ Молева А. А.
дата выполнения работы *подпись* *фамилия, и.о.*

Преподаватель _____ Гапанюк Ю. Е.
подпись *фамилия, и.о.*

Москва, 2022 г.

Цель работы

Цель лабораторной работы: изучение продвинутых способов предварительной обработки данных для дальнейшего формирования моделей.

Задание

1. Выбрать один или несколько наборов данных (датасетов) для решения следующих задач. Каждая задача может быть решена на отдельном датасете, или несколько задач могут быть решены на одном датасете. Просьба не использовать датасет, на котором данная задача решалась в лекции.
2. Для выбранного датасета (датасетов) на основе материалов лекций решить следующие задачи:
   I. масштабирование признаков (не менее чем тремя способами);
   II. обработку выбросов для числовых признаков (по одному способу для удаления выбросов и для замены выбросов);
   III. обработку по крайней мере одного нестандартного признака (который не является числовым или категориальным);
   IV. отбор признаков:
      - один метод из группы методов фильтрации (filter methods);
      - один метод из группы методов обертывания (wrapper methods);
      - один метод из группы методов вложений (embedded methods).

Текст программы

```
import pandas as pd
df = pd.read_csv('data.csv')
df.info()
df[df.author.isna()]

#Масштабирование признаков
#StandartScaler
df_digits = df.select_dtypes(include=[int, float], exclude=None)
df_digits_columns = df_digits.columns
# Обучаем StandardScaler на всей выборке и масштабируем
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
df_scaled_part = pd.DataFrame(scaler.fit_transform(df_digits), columns=df_dig
its_columns)
df_scaled = pd.concat([df_scaled_part, df.select_dtypes(include=None, exclude
=[int, float])], axis=1)
df_scaled

#Mean Normalisation
class MeanNormalisation:

    def fit(self, df):
        self.means = df.mean(axis=0)
        maxs = df.max(axis=0)
        mins = df.min(axis=0)
        self.ranges = maxs - mins

    def transform(self, df):
        param_df_scaled = (df - self.means) / self.ranges
        return param_df_scaled

    def fit_transform(self, df):
        self.fit(df)
        return self.transform(df)
sc21 = MeanNormalisation()
data_mean_scaled = sc21.fit_transform(df_digits)
data_mean_scaled.describe()
df_mean_scaled2 = pd.concat([data_mean_scaled, df.select_dtypes(include=None,
 exclude=[int, float])], axis=1)
df_mean_scaled2.head()

#MinMax
# Обучаем StandardScaler на всей выборке и масштабируем
from sklearn.preprocessing import MinMaxScaler

minMaxScaler = MinMaxScaler()
df_min_max_scaled_part = pd.DataFrame(minMaxScaler.fit_transform(df_digits),
columns=df_digits_columns)
```

```
df_min_max_scaled = pd.concat([df_min_max_scaled_part, df.select_dtypes(inclu
de=None, exclude=[int, float])], axis=1)
df_min_max_scaled.head()

#Обработка выбросов для числовых признаков
#Удаление выбросов
import numpy as np

df_without_blowout = df[(df['views'] < np.quantile(df['views'], 0.95)) \
                        & (df['views'] > np.quantile(df['views'], 0.05))]
df_without_blowout.head()
df_without_blowout.shape
print(round(100 - 100 * df_without_blowout.shape[0] / df.shape[0], 2), '% был
о удалено данных')

#Замена выбросов
df_chg = df.copy()
df_chg['views'] = np.where(df_chg['views'] > np.quantile(df_chg['views'], 0.9
5), \
                           np.quantile(df_chg['views'], 0.95), df_chg['vi
ews'])
df_chg['views'] = np.where(df_chg['views'] < np.quantile(df_chg['views'], 0.0
5), \
                           np.quantile(df_chg['views'], 0.05), df_chg['vi
ews'])
df_chg.head()
df_chg.shape

#Обработка нестандартных признаков
df['date'] =  pd.to_datetime(df['date'])
df.dtypes

#Отбор признаков
#Метод фильтрации
import seaborn as sns
sns.heatmap(df.corr(), annot=True, fmt='.3f')
# Формирование DataFrame с сильными корреляциями
def make_corr_df(df):
    cr = df.corr()
    cr = cr.abs().unstack()
    cr = cr.sort_values(ascending=False)
    cr = cr[cr >= 0.3]
    cr = cr[cr < 1]
    cr = pd.DataFrame(cr).reset_index()
    cr.columns = ['f1', 'f2', 'corr']
    return cr
# Обнаружение групп коррелирующих признаков
def corr_groups(cr):
    grouped_feature_list = []
    correlated_groups = []
```

```
    for feature in cr['f1'].unique():
        if feature not in grouped_feature_list:
            # находим коррелирующие признаки
            correlated_block = cr[cr['f1'] == feature]
            cur_dups = list(correlated_block['f2'].unique()) + [feature]
            grouped_feature_list = grouped_feature_list + cur_dups
            correlated_groups.append(cur_dups)
    return correlated_groups
# Группы коррелирующих признаков
corr_groups(make_corr_df(df))

#Метод из группы методов вложений
df2 = pd.read_spss('1ResearchProjectData.sav')
from sklearn.preprocessing import LabelEncoder

labelEnc = LabelEncoder()
col_dict = {}
for col in df2.select_dtypes(include=None, exclude=float):
    k = labelEnc.fit_transform(df2[col])
    col_dict[col] = k
df2_labeled = pd.concat([pd.DataFrame(col_dict), df2.select_dtypes(include=fl
oat)], axis=1)
df2_labeled
df2_labeled.info()
df2_labeled_notna = df2_labeled.dropna(axis=0)
df2_labeled_notna
from sklearn.linear_model import LogisticRegression
from sklearn.svm import LinearSVC

# Используем L1-регуляризацию
e_lr1 = LogisticRegression(C=1000, solver='liblinear', penalty='l1', max_iter
=500, random_state=1)
X, y = df2_labeled_notna.drop('Score', axis=1), df2_labeled_notna['Score']
e_lr1.fit(X, y)
# Коэффициенты регрессии
# e_lr1.coef_
# Все 4 признака являются "хорошими"
sel_e_lr1 = SelectFromModel(e_lr1)
sel_e_lr1.fit(X, y)
sel_e_lr1.get_support()

#Метод обертывания
import joblib
import sys

sys.modules['sklearn.externals.joblib'] = joblib
from mlxtend.feature_selection import ExhaustiveFeatureSelector as EFS
from sklearn.neighbors import KNeighborsClassifier

knn = KNeighborsClassifier(n_neighbors=3)
efs1 = EFS(knn,
```

```
            min_features=2,
            max_features=4,
            scoring='neg_mean_absolute_error',
            print_progress=True,
            cv=5)

efs1 = efs1.fit(X, y, custom_feature_names=X.columns)

print('Best accuracy score: %.2f' % efs1.best_score_)
print('Best subset (indices):', efs1.best_idx_)
```

Экранные формы

	title	author	date	views	likes	link
0	Climate action needs new frontline leadership	Ozawa Bineshi Albert	December 2021	404000	12000	https://ted.com/talks/ozawa_bineshi_albert_cli...
1	The dark history of the overthrow of Hawaii	Sydney Iaukea	February 2022	214000	6400	https://ted.com/talks/sydney_iaukea_the_dark_h...
2	How play can spark new ideas for your business	Martin Reeves	September 2021	412000	12000	https://ted.com/talks/martin_reeves_how_play_c...
3	Why is China appointing judges to combat clima...	James K. Thornton	October 2021	427000	12000	https://ted.com/talks/james_k_thornton_why_is_...
4	Cement's carbon problem — and 2 ways to fix it	Mahendra Singhi	October 2021	2400	72	https://ted.com/talks/mahendra_singhi_cement_s...
...	...	...	...	...	...	...
5435	The best stats you've ever seen	Hans Rosling	February 2006	15000000	458000	https://ted.com/talks/hans_rosling_the_best_st...
5436	Do schools kill creativity?	Sir Ken Robinson	February 2006	72000000	2100000	https://ted.com/talks/sir_ken_robinson_do_scho...
5437	Greening the ghetto	Majora Carter	February 2006	2900000	88000	https://ted.com/talks/majora_carter_greening_t...
5438	Simplicity sells	David Pogue	February 2006	2000000	60000	https://ted.com/talks/david_pogue_simplicity_s...
5439	Averting the climate crisis	Al Gore	February 2006	3600000	109000	https://ted.com/talks/al_gore_averting_the_cli...

5440 rows × 6 columns

Рисунок 1 – Датасет 1

	views	likes	title	author	date	link
0	-0.464727	-0.470170	Climate action needs new frontline leadership	Ozawa Bineshi Albert	2021-12-01	https://ted.com/talks/ozawa_bineshi_albert_cli...
1	-0.517997	-0.522197	The dark history of the overthrow of Hawaii	Sydney Iaukea	2022-02-01	https://ted.com/talks/sydney_iaukea_the_dark_h...
2	-0.462485	-0.470170	How play can spark new ideas for your business	Martin Reeves	2021-09-01	https://ted.com/talks/martin_reeves_how_play_c...
3	-0.458279	-0.470170	Why is China appointing judges to combat clima...	James K. Thornton	2021-10-01	https://ted.com/talks/james_k_thornton_why_is_...
4	-0.577322	-0.580987	Cement's carbon problem — and 2 ways to fix it	Mahendra Singhi	2021-10-01	https://ted.com/talks/mahendra_singhi_cement_s...
...	...	...	...	...	...	...
5435	3.627491	3.673392	The best stats you've ever seen	Hans Rosling	2006-02-01	https://ted.com/talks/hans_rosling_the_best_st...
5436	19.608336	18.928387	Do schools kill creativity?	Sir Ken Robinson	2006-02-01	https://ted.com/talks/sir_ken_robinson_do_scho...
5437	0.235065	0.235908	Greening the ghetto	Majora Carter	2006-02-01	https://ted.com/talks/majora_carter_greening_t...
5438	-0.017264	-0.024226	Simplicity sells	David Pogue	2006-02-01	https://ted.com/talks/david_pogue_simplicity_s...
5439	0.431321	0.431008	Averting the climate crisis	Al Gore	2006-02-01	https://ted.com/talks/al_gore_averting_the_cli...

5440 rows × 6 columns

Рисунок 2 – StandardScaler

	views	likes	title	author	date	link
0	-0.023022	-0.024099	Climate action needs new frontline leadership	Ozawa Bineshi Albert	2021-12-01	https://ted.com/talks/ozawa_bineshi_albert_cli...
1	-0.025661	-0.026766	The dark history of the overthrow of Hawaii	Sydney Iaukea	2022-02-01	https://ted.com/talks/sydney_iaukea_the_dark_h...
2	-0.022911	-0.024099	How play can spark new ideas for your business	Martin Reeves	2021-09-01	https://ted.com/talks/martin_reeves_how_play_c...
3	-0.022703	-0.024099	Why is China appointing judges to combat clima...	James K. Thornton	2021-10-01	https://ted.com/talks/james_k_thornton_why_is_...
4	-0.028600	-0.029779	Cement's carbon problem — and 2 ways to fix it	Mahendra Singhi	2021-10-01	https://ted.com/talks/mahendra_singhi_cement_s...

Рисунок 3 – Mean Normalisation

	views	likes	title	author	date	link
0	0.005604	0.005707	Climate action needs new frontline leadership	Ozawa Bineshi Albert	2021-12-01	https://ted.com/talks/ozawa_bineshi_albert_cli...
1	0.002965	0.003040	The dark history of the overthrow of Hawaii	Sydney Iaukea	2022-02-01	https://ted.com/talks/sydney_iaukea_the_dark_h...
2	0.005715	0.005707	How play can spark new ideas for your business	Martin Reeves	2021-09-01	https://ted.com/talks/martin_reeves_how_play_c...
3	0.005923	0.005707	Why is China appointing judges to combat clima...	James K. Thornton	2021-10-01	https://ted.com/talks/james_k_thornton_why_is_...
4	0.000026	0.000027	Cement's carbon problem — and 2 ways to fix it	Mahendra Singhi	2021-10-01	https://ted.com/talks/mahendra_singhi_cement_s...

Рисунок 4 – MinMax-масштабирование

	title	author	date	views	likes	link
0	Climate action needs new frontline leadership	Ozawa Bineshi Albert	2021-12-01	404000	12000	https://ted.com/talks/ozawa_bineshi_albert_cli...
1	The dark history of the overthrow of Hawaii	Sydney Iaukea	2022-02-01	214000	6400	https://ted.com/talks/sydney_iaukea_the_dark_h...
2	How play can spark new ideas for your business	Martin Reeves	2021-09-01	412000	12000	https://ted.com/talks/martin_reeves_how_play_c...
3	Why is China appointing judges to combat clima...	James K. Thornton	2021-10-01	427000	12000	https://ted.com/talks/james_k_thornton_why_is_...
5	The tragedy of air pollution — and an urgent d...	Rosamund Adoo-Kissi-Debrah	2021-10-01	422000	12000	https://ted.com/talks/rosamund_adoo_kissi_debr...

```
[ ] df_without_blowout.shape
```

```
(4892, 6)
```

```
[ ] print(round(100 - 100 * df_without_blowout.shape[0] / df.shape[0], 2), '% было удалено данных')
```

```
10.07 % было удалено данных
```

Рисунок 5 – Удаление выбросов

```
[ ] df_chg = df.copy()
    df_chg['views'] = np.where(df_chg['views'] > np.quantile(df_chg['views'], 0.95), \
                               np.quantile(df_chg['views'], 0.95), df_chg['views'])
    df_chg['views'] = np.where(df_chg['views'] < np.quantile(df_chg['views'], 0.05), \
                               np.quantile(df_chg['views'], 0.05), df_chg['views'])
```

```
df_chg.head()
```

	title	author	date	views	likes	link
0	Climate action needs new frontline leadership	Ozawa Bineshi Albert	2021-12-01	404000.0	12000	https://ted.com/talks/ozawa_bineshi_albert_cli...
1	The dark history of the overthrow of Hawaii	Sydney Iaukea	2022-02-01	214000.0	6400	https://ted.com/talks/sydney_iaukea_the_dark_h...
2	How play can spark new ideas for your business	Martin Reeves	2021-09-01	412000.0	12000	https://ted.com/talks/martin_reeves_how_play_c...
3	Why is China appointing judges to combat clima...	James K. Thornton	2021-10-01	427000.0	12000	https://ted.com/talks/james_k_thornton_why_is_...
4	Cement's carbon problem — and 2 ways to fix it	Mahendra Singhi	2021-10-01	18000.0	72	https://ted.com/talks/mahendra_singhi_cement_s...

```
[ ] df_chg.shape
```

```
(5440, 6)
```

Рисунок 6 – Замена выбросов

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5440 entries, 0 to 5439
Data columns (total 7 columns):
 #   Column      Non-Null Count  Dtype         
---  ------      --------------  -----         
 0   title       5440 non-null   object        
 1   author      5439 non-null   object        
 2   date        5440 non-null   datetime64[ns]
 3   views       5440 non-null   int64         
 4   likes       5440 non-null   int64         
 5   link        5440 non-null   object        
 6   split_date  5440 non-null   datetime64[ns]
dtypes: datetime64[ns](2), int64(2), object(3)
memory usage: 297.6+ KB
```

```
[ ] df['date'] =  pd.to_datetime(df['date'])
```

```
[ ] df.dtypes
```

```
title                 object
author                object
date          datetime64[ns]
views                  int64
likes                  int64
link                  object
split_date    datetime64[ns]
dtype: object
```

Рисунок 7 – Обработка нестандартного признака

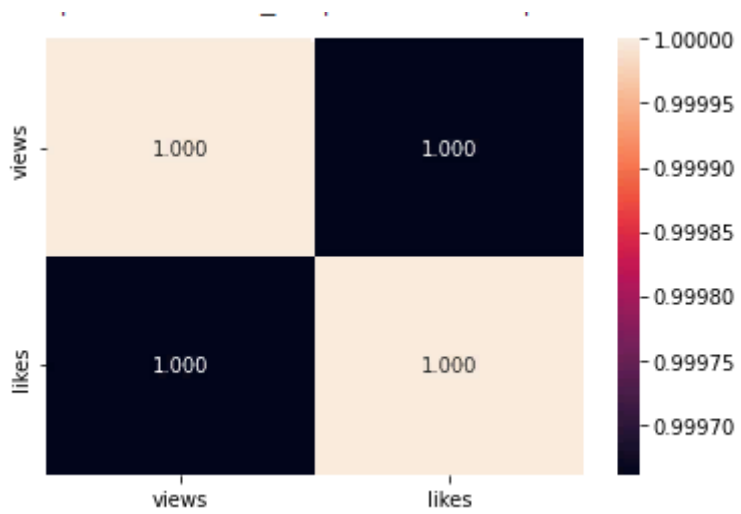

Рисунок 8 – Метод фильтрации (Корреляция признаков)

```
from sklearn.feature_selection import SelectFromModel

# Все 6 признаков являются "хорошими"
sel_e_lr1 = SelectFromModel(e_lr1)
sel_e_lr1.fit(X, y)
sel_e_lr1.get_support()
```

```
/usr/local/lib/python3.7/dist-packages/sklearn/svm/_base.py:1208: ConvergenceWarning: Liblinear failed to converge, increase the number of iterations.
  ConvergenceWarning,
array([ True,  True,  True,  True,  True,  True])
```

Рисунок 9 – Метод из группы методов вложений

```
  UserWarning,
Features: 50/50Best accuracy score: -18.08
Best subset (indices): (0, 1, 2)
```

Рисунок 10 – Метод обертывания (wrapper methods)

Выводы

В результате проделанной работы были решены следующие задачи: масштабирование признаков; обработка выбросов; обработка нестандартного признака; отбор признаков.