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 № 2

Обработка признаков (часть 1)

название лабораторной работы

Группа ИУ5-14М

Студент 28.03.2022 _____ 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 набор данных (датасет), содержащий категориальные и числовые признаки и пропуски в данных. Для выполнения следующих пунктов можно использовать несколько различных наборов данных (один для обработки пропусков, другой для категориальных признаков и т.д.) Просьба не использовать датасет, на котором данная задача решалась в лекции.
2. Для выбранного датасета (датасетов) на основе материалов лекций решить следующие задачи:
   I. устранение пропусков в данных;
   II. кодирование категориальных признаков;
   III. нормализацию числовых признаков.

Текст программы

```
import pandas as pd
df = pd.read_csv('kamyr-digester.csv')
df.info()

#Удаление пустых строк
df1 = df.dropna(axis=0)
df1.isnull().sum()

#Simpleimputer
from numpy import nan
from numpy import isnan
from pandas import read_csv
from sklearn.impute import SimpleImputer
imputer = SimpleImputer(missing_values=nan, strategy='mean')
dfSimpleImputer = pd.DataFrame(imputer.fit_transform(df.iloc[:, 1:]), columns
=df.columns[1:])
dfSimpleImputer['Observation'] = df['Observation']
dfSimpleImputer
dfSimpleImputer.isnull().sum()

#KNNImputer
from sklearn.impute import KNNImputer
knnimputer = KNNImputer(
    n_neighbors=5,
    weights='distance',
    metric='nan_euclidean',
    add_indicator=False,
)

knnimpute_hdata_imputed_temp = knnimputer.fit_transform(df.iloc[:, 1:])
knnimpute_hdata_imputed = pd.DataFrame(knnimpute_hdata_imputed_temp, columns=
df.columns[1:])
knnimpute_hdata_imputed.head()
knnimpute_hdata_imputed
knnimpute_hdata_imputed.isnull().sum()

#Категориальные признаки
! kaggle competitions download -c titanic
dfCat = pd.read_csv('train.csv')
dfCat
dfCat.info()
from sklearn.preprocessing import OneHotEncoder

#Разбиение на две колонки
pd.get_dummies(dfCat[['Sex']]).head()

#Кодирование массива
from sklearn.preprocessing import LabelEncoder
le = LabelEncoder()
```

```
cat_enc_le = le.fit_transform(dfCat['Name'])
dfCat['Name'].unique()
import numpy as np
np.unique(cat_enc_le)
le.inverse_transform([0, 1, 2, 3])

#Нормализация
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import scipy.stats as stats

def diagnostic_plots(df, variable):
    plt.figure(figsize=(15,6))
    # гистограмма
    plt.subplot(1, 2, 1)
    df[variable].hist(bins=30)
    ## Q-Q plot
    plt.subplot(1, 2, 2)
    stats.probplot(df[variable], dist="norm", plot=plt)
    plt.show()

#Логарифмическое преобразование
df['ChipRate_log'] = np.log(df['ChipRate'])
diagnostic_plots(df, 'ChipRate_log')

#Обратное преобразование
df['ChipRate_reciprocal'] = 1 / (df['ChipRate'])
diagnostic_plots(df, 'ChipRate_reciprocal')

#Квадратный корень
df['ChipRate_sqr'] = df['ChipRate']**(1/2)
diagnostic_plots(df, 'ChipRate_sqr')
```

Экранные формы

	Observation	Y-Kappa	ChipRate	BF-CMratio	BlowFlow	ChipLevel4	T-upperExt-2	T-lowerExt-2	UCZAA	WhiteFlow-4	...	SteamFlow-4	Lower-HeatT-3	Upper-HeatT-3	ChipMass-4	WeakLiquorF	BlackFlow-2	WeakWashF	SteamHeatF-3	T-Top-Chips-4	SulphidityL-4
0	31-00:00	23.10	16.520	121.717	1177.607	169.805	358.282	329.545	1.443	599.253	...	67.122	329.432	303.099	175.964	1127.197	1319.039	257.325	54.612	252.077	NaN
1	31-01:00	27.60	16.810	79.022	1328.360	341.327	351.050	329.067	1.549	537.201	...	60.012	330.823	304.879	163.202	665.975	1297.317	241.182	46.603	251.406	29.11
2	31-02:00	23.19	16.709	79.562	1329.407	239.161	350.022	329.260	1.600	549.611	...	61.304	329.140	303.383	164.013	677.534	1327.072	237.272	51.795	251.335	NaN
3	31-03:00	23.60	16.478	81.011	1334.877	213.527	350.938	331.142	1.604	623.362	...	68.496	328.875	302.254	181.487	767.853	1324.461	239.478	54.846	250.312	29.02
4	31-04:00	22.90	15.618	93.244	1334.168	243.131	351.640	332.709	NaN	638.672	...	70.022	328.352	300.954	183.929	888.448	1343.424	215.372	54.186	249.916	29.01
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
296	12-08:00	20.40	14.233	89.790	1278.006	379.458	354.290	315.558	1.515	491.374	...	60.424	331.980	308.078	140.301	975.016	1344.835	388.676	47.803	252.311	NaN
297	12-09:00	20.90	15.167	84.640	1283.706	339.440	354.803	311.041	1.635	532.419	...	65.561	332.924	307.626	145.299	832.906	1344.708	388.911	49.524	251.833	30.29
298	12-10:00	24.98	NaN	85.034	1278.345	368.564	357.723	321.387	NaN	520.365	...	65.729	332.523	307.169	151.544	905.639	1344.469	418.979	48.135	251.614	30.47
299	12-11:00	21.00	NaN	88.013	1307.722	278.842	357.438	323.757	NaN	553.070	...	65.795	331.263	306.400	157.954	908.691	1344.588	462.712	54.373	251.197	NaN
300	12-12:00	21.40	NaN	85.490	1255.986	273.484	361.365	322.689	NaN	590.199	...	71.456	333.032	308.732	174.069	986.206	1348.747	457.313	53.194	251.324	30.46

301 rows × 23 columns

Рисунок 1 – Датасет 1

	Observation	Y-Kappa	ChipRate	BF-CMratio	BlowFlow	ChipLevel4	T-upperExt-2	T-lowerExt-2	UCZAA	WhiteFlow-4	...	SteamFlow-4	Lower-HeatT-3	Upper-HeatT-3	ChipMass-4
1	31-01:00	27.60	16.810	79.022	1328.360	341.327	351.050	329.067	1.549	537.201	...	60.012	330.823	304.879	163.202
3	31-03:00	23.60	16.478	81.011	1334.877	213.527	350.938	331.142	1.604	623.362	...	68.496	328.875	302.254	181.487
6	31-06:00	22.65	14.100	91.887	1307.852	288.989	352.321	331.162	1.468	625.549	...	71.298	329.662	301.539	179.886
8	31-08:00	24.70	13.850	96.208	1334.892	362.511	352.372	327.358	1.515	553.172	...	64.249	332.264	305.419	166.120
10	31-10:00	24.40	14.117	85.998	1330.104	394.234	348.089	319.027	1.429	540.558	...	62.179	329.831	302.652	163.258
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
289	12-01:00	19.90	11.333	87.405	1033.565	369.383	343.515	302.364	1.592	452.718	...	55.963	330.842	308.789	128.701
291	12-03:00	22.00	11.858	93.199	1171.206	366.787	345.261	310.115	1.513	428.202	...	52.494	330.589	309.152	122.011
293	12-05:00	19.00	12.425	92.905	1272.030	316.226	345.811	307.806	1.633	469.045	...	60.307	329.997	308.072	137.719
295	12-07:00	20.50	13.358	97.662	1304.597	377.678	347.672	313.147	1.546	496.460	...	60.119	332.615	308.575	141.076
297	12-09:00	20.90	15.167	84.640	1283.706	339.440	354.803	311.041	1.635	532.419	...	65.561	332.924	307.626	145.299

131 rows × 23 columns

Рисунок 2 – Удаление пустых строк

	Y-Kappa	ChipRate	BF-CMratio	BlowFlow	ChipLevel4	T-upperExt-2	T-lowerExt-2	UCZAA	WhiteFlow-4	AAWhiteSt-4	...	Lower-HeatT-3	Upper-HeatT-3	ChipMass-4	WeakLiquorF
0	23.10	16.52000	121.717	1177.607	169.805	358.282	329.545	1.443000	599.253	6.143012	...	329.432	303.099	175.964	1127.197
1	27.60	16.81000	79.022	1328.360	341.327	351.050	329.067	1.549000	537.201	6.076000	...	330.823	304.879	163.202	665.975
2	23.19	16.70900	79.562	1329.407	239.161	350.022	329.260	1.600000	549.611	6.143012	...	329.140	303.383	164.013	677.534
3	23.60	16.47800	81.011	1334.877	213.527	350.938	331.142	1.604000	623.362	6.054000	...	328.875	302.254	181.487	767.853
4	22.90	15.61800	93.244	1334.168	243.131	351.640	332.709	1.490588	638.672	6.110000	...	328.352	300.954	183.929	888.448
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
296	20.40	14.23300	89.790	1278.006	379.458	354.290	315.558	1.515000	491.374	6.143012	...	331.980	308.078	140.301	975.016
297	20.90	15.16700	84.640	1283.706	339.440	354.803	311.041	1.635000	532.419	6.340000	...	332.924	307.626	145.299	832.906
298	24.98	14.33867	85.034	1278.345	368.564	357.723	321.387	1.490588	520.365	6.220000	...	332.523	307.169	151.544	905.639
299	21.00	14.33867	88.013	1307.722	278.842	357.438	323.757	1.490588	553.070	6.143012	...	331.263	306.400	157.954	908.691
300	21.40	14.33867	85.490	1255.986	273.484	361.365	322.689	1.490588	590.199	6.230000	...	333.032	308.732	174.069	986.206

301 rows × 23 columns

Рисунок 3 – Замена средним значением

	Y-Kappa	ChipRate	BF-CMratio	BlowFlow	ChipLevel4	T-upperExt-2	T-lowerExt-2	UCZAA	WhiteFlow-4	AAWhiteSt-4	...	SteamFlow-4	Lower-HeatT-3	Upper-HeatT-3	ChipMass-4	WeakLiquorF	BlackFlow-2	WeakWashF	SteamHeatF-3	T-Top-Chips-4	SulphidityL-4
0	23.10	16.520000	121.717	1177.607	169.805	358.282	329.545	1.443000	599.253	6.088805	...	67.122	329.432	303.099	175.964	1127.197	1319.039	257.325	54.612	252.077	31.273819
1	27.60	16.810000	79.022	1328.360	341.327	351.050	329.067	1.549000	537.201	6.076000	...	60.012	330.823	304.879	163.202	665.975	1297.317	241.182	46.603	251.406	29.110000
2	23.19	16.709000	79.562	1329.407	239.161	350.022	329.260	1.600000	549.611	6.077649	...	61.304	329.140	303.383	164.013	677.534	1327.072	237.272	51.795	251.335	29.384961
3	23.60	16.478000	81.011	1334.877	213.527	350.938	331.142	1.604000	623.362	6.054000	...	68.496	328.875	302.254	181.487	767.853	1324.461	239.478	54.846	250.312	29.020000
4	22.90	15.618000	93.244	1334.168	243.131	351.640	332.709	1.514654	638.672	6.110000	...	70.022	328.352	300.954	183.929	888.448	1343.424	215.372	54.186	249.916	29.010000
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
296	20.40	14.233000	89.790	1278.006	379.458	354.290	315.558	1.515000	491.374	6.235430	...	60.424	331.980	308.078	140.301	975.016	1344.835	388.676	47.803	252.311	30.776069
297	20.90	15.167000	84.640	1283.706	339.440	354.803	311.041	1.635000	532.419	6.340000	...	65.561	332.924	307.626	145.299	832.906	1344.708	388.911	49.524	251.833	30.290000
298	24.98	14.794154	85.034	1278.345	368.564	357.723	321.387	1.570349	520.365	6.220000	...	65.729	332.523	307.169	151.544	905.639	1344.469	418.979	48.135	251.614	30.470000
299	21.00	14.997380	88.013	1307.722	278.842	357.438	323.757	1.582153	553.070	6.173173	...	65.795	331.263	306.400	157.954	908.691	1344.588	462.712	54.373	251.197	30.224469
300	21.40	14.847598	85.490	1255.986	273.484	361.365	322.689	1.573451	590.199	6.230000	...	71.456	333.032	308.732	174.069	986.206	1348.747	457.313	53.194	251.324	30.460000

301 rows × 22 columns

Рисунок 4 – Заполнение по ближайшим соседям

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S
...	...	...	...	...	...	...	...	...	...	...	...	...
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.0000	NaN	S
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.0000	B42	S
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	23.4500	NaN	S
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.0000	C148	C
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.7500	NaN	Q

891 rows × 12 columns

Рисунок 5 – Датасет 2

```
pd.get_dummies(dfCat[['Sex']]).head()
```

	Sex_female	Sex_male
0	0	1
1	1	0
2	1	0
3	1	0
4	0	1

Рисунок 6 – Разделение на две колонки

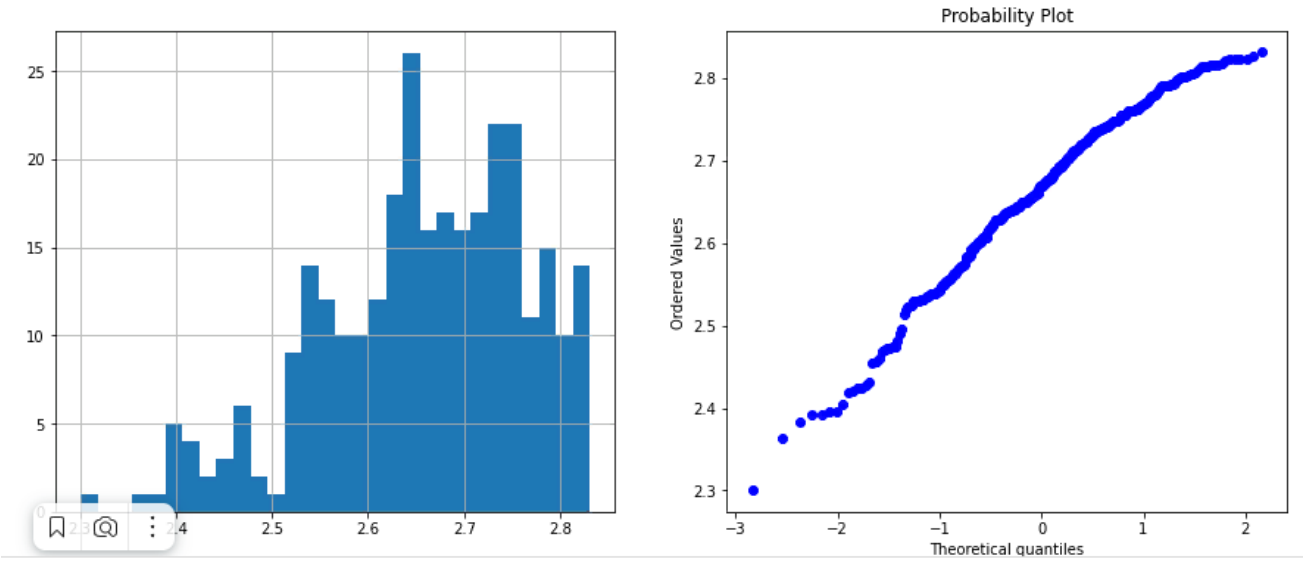

Рисунок 7 – Логарифмическое преобразование

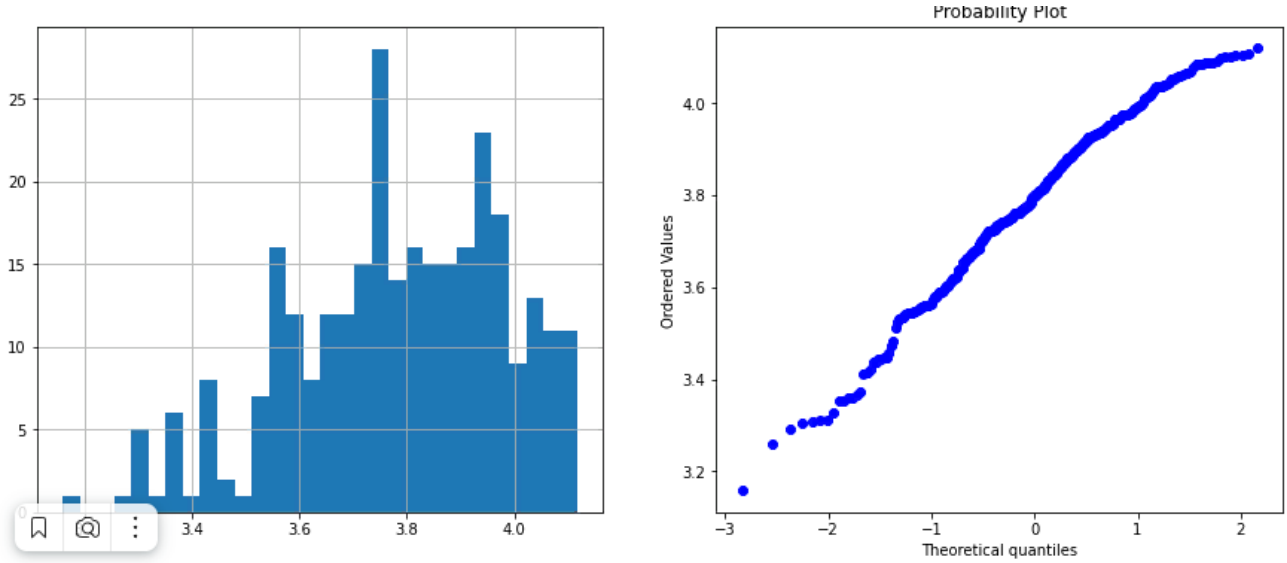

Рисунок 8 – Обратное преобразование

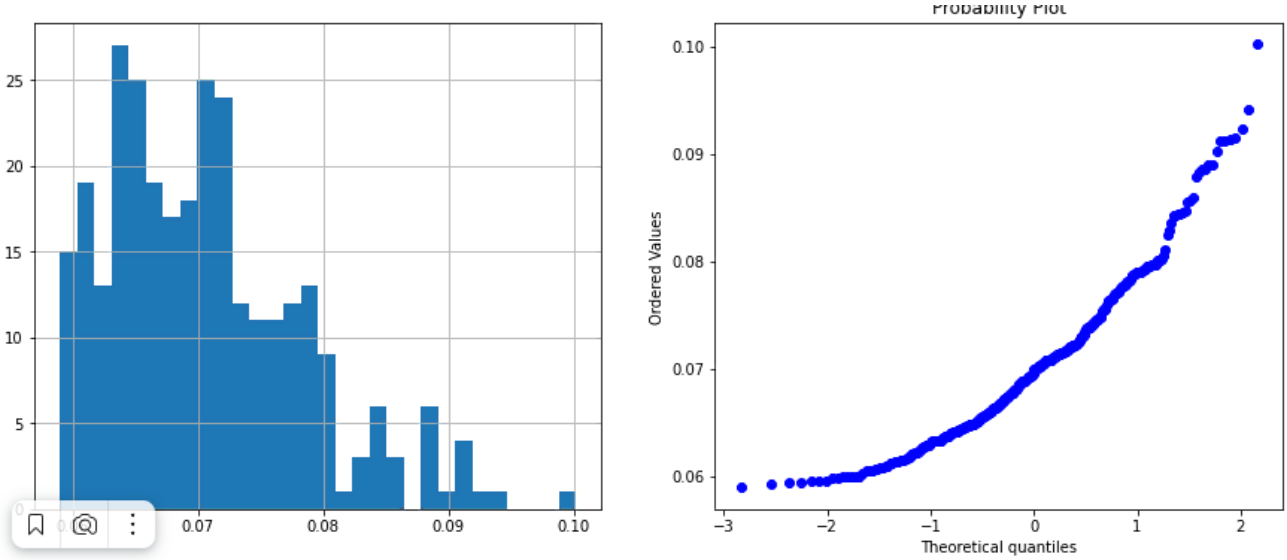

Рисунок 9 – Квадратный корень

Выводы

В результате проделанной работы были решены следующие задачи: устранение пропусков в данных; кодирование категориальных признаков; нормализация числовых признаков.