МГТУ имени Баумана
Факультет «Информатика и Системы управления»
Кафедра «Системы обработки информации и управления»
Дисциплина «Теория машинного обучения»

Отчет по лабораторной работе №2

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

Выполнил:
Студент группы ИУ5-61Б
Гапчук Л.Д.

Преподаватель:
Гапанюк Ю.Е.

Москва, 2020г.

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

Задание:

1. Выбрать набор данных (датасет), содержащий категориальные 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.)
2. Для выбранного датасета (датасетов) на основе материалов лекции решить следующие задачи:
   - обработку пропусков в данных;
   - кодирование категориальных признаков;
   - масштабирование данных.

Импорт библиотек

In [1]:
```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from pandas.plotting import scatter_matrix
import warnings
warnings.filterwarnings('ignore')
sns.set(style="ticks")
%matplotlib inline
```

In [2]:
```
data = pd.read_csv('reddit_vm.csv')
```

In [3]:
```
data.head()
```

Out[3]:

	title	score	id	url	comms_num	created	body	timestamp
0	Health Canada approves AstraZeneca COVID-19 va...	7	lt74vw	https://www.canadaforums.ca/2021/02/health-can...	0	1.614400e+09	NaN	2021-02-27 06:33:45
1	COVID-19 in Canada: 'Vaccination passports' a ...	2	lsh0ij	https://www.canadaforums.ca/2021/02/covid-19-i...	1	1.614316e+09	NaN	2021-02-26 07:11:07
2	Coronavirus variants could fuel Canada's third...	6	lohlle	https://www.canadaforums.ca/2021/02/coronaviru...	0	1.613887e+09	NaN	2021-02-21 07:50:08
3	Canadian government to extend COVID-19 emergen...	1	lnptv8	https://www.canadaforums.ca/2021/02/canadian-g...	0	1.613796e+09	NaN	2021-02-20 06:35:13
4	Canada: Pfizer is 'extremely committed' to mee...	6	lkslm6	https://www.canadaforums.ca/2021/02/canada-pfi...	0	1.613468e+09	NaN	2021-02-16 11:36:28

In [4]:
```
data.dtypes
```

Out[4]:
```
title         object
score          int64
id            object
url           object
comms_num      int64
created      float64
body          object
timestamp     object
dtype: object
```

In [5]:
```
data.isnull().sum()
# проверим есть ли пропущенные значения
```

Out[5]:
```
title          0
score          0
id             0
url          984
comms_num      0
created        0
body         365
timestamp      0
dtype: int64
```

In [6]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1429 entries, 0 to 1428
Data columns (total 8 columns):
 #   Column     Non-Null Count  Dtype  
---  ------     --------------  -----  
 0   title      1429 non-null   object 
 1   score      1429 non-null   int64  
 2   id         1429 non-null   object 
 3   url        445 non-null    object 
 4   comms_num  1429 non-null   int64  
 5   created    1429 non-null   float64
 6   body       1064 non-null   object 
 7   timestamp  1429 non-null   object 
dtypes: float64(1), int64(2), object(5)
memory usage: 89.4+ KB
```

Обработка пропусков

In [7]:
```
# Удаляем столбцы, которые не несут значимой информации
data.drop(['id','score'], axis = 1, inplace = True)
```

In [8]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1429 entries, 0 to 1428
Data columns (total 6 columns):
 #   Column     Non-Null Count  Dtype  
---  ------     --------------  -----  
 0   title      1429 non-null   object 
 1   url        445 non-null    object 
 2   comms_num  1429 non-null   int64  
 3   created    1429 non-null   float64
 4   body       1064 non-null   object 
 5   timestamp  1429 non-null   object 
dtypes: float64(1), int64(1), object(4)
memory usage: 67.1+ KB
```

In [9]:
```
# Заполняем отсутствующие значения
data = data.fillna("Num")
data.head()
```

Out[9]:

	title	url	comms_num	created	body	timestamp
0	Health Canada approves AstraZeneca COVID-19 va...	https://www.canadaforums.ca/2021/02/health-can...	0	1.614400e+09	Num	2021-02-27 06:33:45
1	COVID-19 in Canada: 'Vaccination passports' a ...	https://www.canadaforums.ca/2021/02/covid-19-i...	1	1.614316e+09	Num	2021-02-26 07:11:07
2	Coronavirus variants could fuel Canada's third...	https://www.canadaforums.ca/2021/02/coronaviru...	0	1.613887e+09	Num	2021-02-21 07:50:08
3	Canadian government to extend COVID-19 emergen...	https://www.canadaforums.ca/2021/02/canadian-g...	0	1.613796e+09	Num	2021-02-20 06:35:13
4	Canada: Pfizer is 'extremely committed' to mee...	https://www.canadaforums.ca/2021/02/canada-pfi...	0	1.613468e+09	Num	2021-02-16 11:36:28

In [10]:
```
data.isnull().sum()
# проверим есть ли пропущенные значения
```

Out[10]:
```
title        0
url          0
comms_num    0
created      0
body         0
timestamp    0
dtype: int64
```

Импорт библиотек

In [1]:
```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
from sklearn.impute import SimpleImputer
from sklearn.model_selection import train_test_split
```

In [2]:
```
data = pd.read_csv('train.csv')
```

In [3]:
```
data.head()
```

Out[3]:

	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

In [4]:
```
data['Embarked'].value_counts()
```

Out[4]:
```
S    644
C    168
Q     77
Name: Embarked, dtype: int64
```

In [5]:
```
# Кодируем признаки Pclass и Embarked в отдельные столбцы
data = pd.get_dummies(data, columns=['Pclass','Embarked'])
```

In [6]:
```
# Пол кодируем в 1/0
data['IsMale']=data.Sex.replace({'female':0,'male':1})
data.drop('Sex', axis = 1, inplace = True)
```

In [7]:
```
data.head()
```

Out[7]:

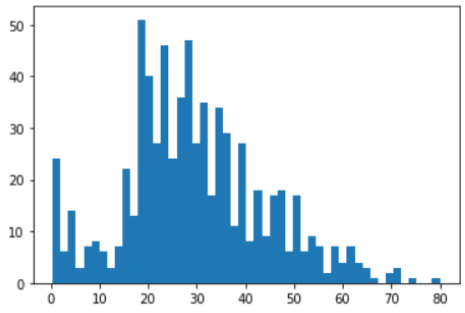
	PassengerId	Survived	Name	Age	SibSp	Parch	Ticket	Fare	Cabin	Pclass_1	Pclass_2	Pclass_3	Embarked_C	Embarked_Q	Embarked_S	I
0	1	0	Braund, Mr. Owen Harris	22.0	1	0	A/5 21171	7.2500	NaN	0	0	1	0	0	1	
1	2	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	38.0	1	0	PC 17599	71.2833	C85	1	0	0	1	0	0	
2	3	1	Heikkinen, Miss. Laina	26.0	0	0	STON/O2. 3101282	7.9250	NaN	0	0	1	0	0	1	
3	4	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	35.0	1	0	113803	53.1000	C123	1	0	0	0	0	1	
4	5	0	Allen, Mr. William Henry	35.0	0	0	373450	8.0500	NaN	0	0	1	0	0	1	

Масштабирование значений

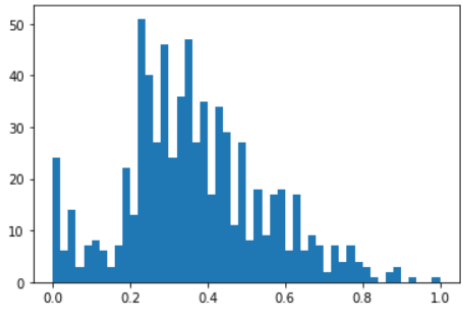
In [8]:
```
from sklearn.preprocessing import StandardScaler, MinMaxScaler, StandardScaler, Normalizer
```

In [9]:
```
sc1 = MinMaxScaler()
sc1_data = sc1.fit_transform(data[['Age']])
```

In [10]:
```
plt.hist(data['Age'], 50)
plt.show()
```



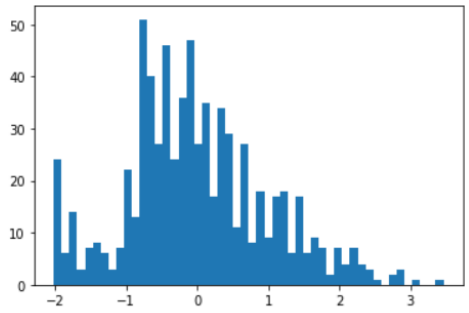
In [11]:
```
plt.hist(sc1_data, 50)
plt.show()
```



In [12]:
```
# Удаляем столбцы, которые не несут значимой информации
data.drop(['Cabin','Name','Ticket'], axis = 1, inplace = True)
```

In [13]:
```
sc2 = StandardScaler()
sc2_data = sc2.fit_transform(data[['Age']])
```

In [14]:
```
plt.hist(sc2_data, 50)
plt.show()
```



In [15]:
```
data.head()
```

Out[15]:

	PassengerId	Survived	Age	SibSp	Parch	Fare	Pclass_1	Pclass_2	Pclass_3	Embarked_C	Embarked_Q	Embarked_S	IsMale
0	1	0	22.0	1	0	7.2500	0	0	1	0	0	1	1
1	2	1	38.0	1	0	71.2833	1	0	0	1	0	0	0
2	3	1	26.0	0	0	7.9250	0	0	1	0	0	1	0
3	4	1	35.0	1	0	53.1000	1	0	0	0	0	1	0
4	5	0	35.0	0	0	8.0500	0	0	1	0	0	1	1