Московский государственный технический университет им. Н.Э. Баумана.

Факультет «Информатика и управление»

Кафедра «Системы обработки информации и управления»

Курс «Теория машинного обучения»

Отчет по лабораторной работе №2

Выполнила:
студентка группы ИУ5-64
Бредня Елизавета

Подпись и дата:

Проверил:
преподаватель каф. ИУ5
Гапанюк Ю.Е.

Подпись и дата:

Москва, 2022 г.

Описание задания

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.

Текст программы и её результаты

```
[6]  import matplotlib.pyplot as plt
     from matplotlib import pyplot
     import missingno
     import seaborn as sns
     import numpy as np # linear algebra
     import pandas as pd
```

```
data = pd.read_csv('/content/train.csv')
data.head()
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450

```
[8]  data.isna().sum()
```

```
PassengerId      0
Survived         0
Pclass           0
Name             0
Sex              0
Age            177
SibSp            0
Parch            0
Ticket           0
Fare             0
Cabin          687
Embarked         2
dtype: int64
```

```
[9]  m = data['Age'].mean()
     m
```

```
29.69911764705882
```

```
[10] data['Age'] = data['Age'].replace(np.nan, 29)
```

```
[11] data = data.drop(['Cabin'], axis=1)
```

```
[12] data['Embarked'].unique()
```

```
array(['S', 'C', 'Q', nan], dtype=object)
```

```
[13] data['Embarked'] = data['Embarked'].replace(np.nan, 'Q')
```

```
[14] data.isna().sum()
```

```
PassengerId    0
Survived       0
Pclass         0
Name           0
Sex            0
Age            0
SibSp          0
Parch          0
Ticket         0
Fare           0
Embarked       0
dtype: int64
```

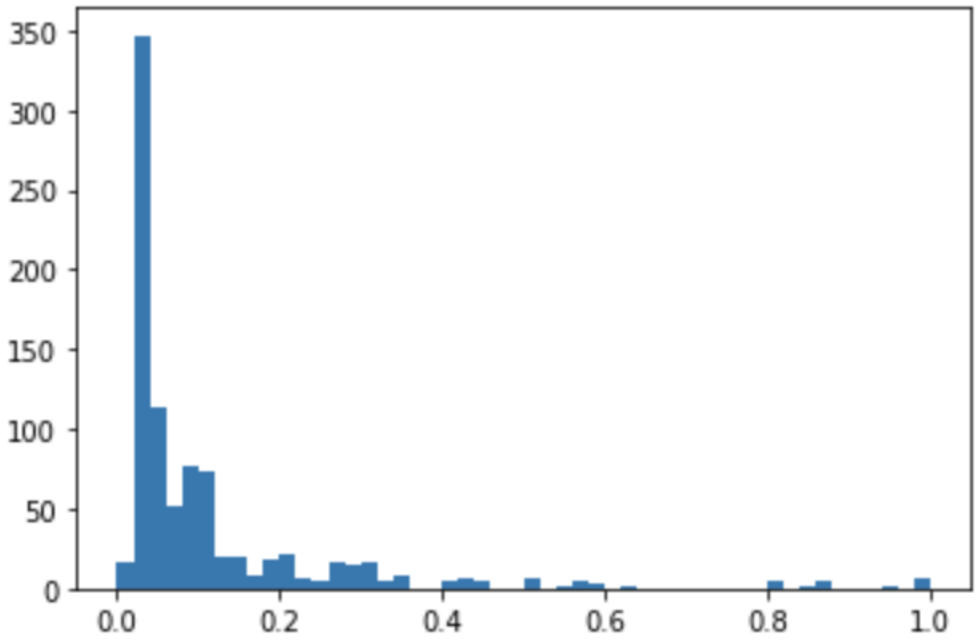
Кодирование категориальных признаков

```
[15] data = pd.get_dummies(data, columns=['Sex', 'Embarked'])
     del data['Name']
     del data['Ticket']
     data.head()
```

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare	Sex_female	Sex_male	Em
0	1	0	3	22.0	1	0	7.2500	0	1	
1	2	1	1	38.0	1	0	71.2833	1	0	
2	3	1	3	26.0	0	0	7.9250	1	0	
3	4	1	1	35.0	1	0	53.1000	1	0	
4	5	0	3	35.0	0	0	8.0500	0	1	

```
[16] data = data.loc[data['Fare'] < 400]
```

```
[17] from sklearn.preprocessing import MinMaxScaler, StandardScaler, Normalizer
     sc1 = MinMaxScaler()
     sc1_data = sc1.fit_transform(data[['Fare']])
     plt.hist(sc1_data, 50)
     plt.show()
```


350
300
250
200
150
100
50
0
0.0
0.2
0.4
0.6
0.8
1.0