Московский государственный технический университет им. Н.Э. Баумана
Факультет «Информатика и системы управления»
Кафедра «Автоматизированные системы обработки информации и управления»

Отчет

Лабораторная работа № 2

По курсу «Технологии машинного обучения»

ИСПОЛНИТЕЛЬ:
Группа ИУ5-65Б
Голубев С.Н.

"4" марта 2020 г.

ПРЕПОДАВАТЕЛЬ:
Гапанюк Ю.Е.

_____

"__"_____2021 г.

Москва 2021

1. 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.

2. Скрины jupyter notebook

ЛР №2

Импорт библиотек

B [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
```

B [2]:
```
data = pd.read_csv('android-games.csv')
```

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

Out[3]:

	rank	title	total ratings	installs	average rating	growth (30 days)	growth (60 days)	price	category	5 star ratings	4 star ratings	3 star ratings	2 star ratings	1 star ratings	paid
0	1	Garena Free Fire - The Cobra	80678661	500.0 M	4.33	2.9	7.9	0.0	GAME ACTION	61935712	4478738	2795172	1814999	9654037	False
1	2	PUBG MOBILE: Graffiti Prank	35971961	100.0 M	4.24	2.0	3.1	0.0	GAME ACTION	26670566	2109631	1352610	893674	4945478	False
2	3	Mobile Legends: Bang Bang	25836869	100.0 M	4.08	1.6	3.3	0.0	GAME ACTION	17850942	1796761	1066095	725429	4397640	False
3	4	Brawl Stars	17181659	100.0 M	4.27	4.1	6.6	0.0	GAME ACTION	12493668	1474319	741410	383478	2088781	False
4	5	Sniper 3D: Fun Free Online FPS Shooting Game	14237554	100.0 M	4.33	0.8	1.8	0.0	GAME ACTION	9657878	2124544	1034025	375159	1045945	False

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

Out[4]:
```
rank                int64
title              object
total ratings       int64
installs           object
average rating    float64
growth (30 days)  float64
growth (60 days)  float64
price             float64
category           object
5 star ratings      int64
4 star ratings      int64
3 star ratings      int64
2 star ratings      int64
1 star ratings      int64
paid                 bool
dtype: object
```

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

Out[5]:
```
rank              0
title             0
total ratings     0
installs          0
average rating    0
growth (30 days)  0
growth (60 days)  0
price             0
category          0
5 star ratings    0
4 star ratings    0
3 star ratings    0
2 star ratings    0
1 star ratings    0
paid              0
dtype: int64
```

В [6]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1830 entries, 0 to 1829
Data columns (total 15 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   rank               1830 non-null   int64  
 1   title              1830 non-null   object 
 2   total ratings      1830 non-null   int64  
 3   installs           1830 non-null   object 
 4   average rating     1830 non-null   float64
 5   growth (30 days)   1830 non-null   float64
 6   growth (60 days)   1830 non-null   float64
 7   price              1830 non-null   float64
 8   category           1830 non-null   object 
 9   5 star ratings     1830 non-null   int64  
 10  4 star ratings     1830 non-null   int64  
 11  3 star ratings     1830 non-null   int64  
 12  2 star ratings     1830 non-null   int64  
 13  1 star ratings     1830 non-null   int64  
 14  paid               1830 non-null   bool   
dtypes: bool(1), float64(4), int64(7), object(3)
memory usage: 202.1+ KB
```

Обработка пропусков

В [7]:
```
# Удаляем столбец, которые не несут значимой информации
data.drop(['paid'], axis = 1, inplace = True)
```

В [8]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1830 entries, 0 to 1829
Data columns (total 14 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   rank               1830 non-null   int64  
 1   title              1830 non-null   object 
 2   total ratings      1830 non-null   int64  
 3   installs           1830 non-null   object 
 4   average rating     1830 non-null   float64
 5   growth (30 days)   1830 non-null   float64
 6   growth (60 days)   1830 non-null   float64
 7   price              1830 non-null   float64
 8   category           1830 non-null   object 
 9   5 star ratings     1830 non-null   int64  
 10  4 star ratings     1830 non-null   int64  
 11  3 star ratings     1830 non-null   int64  
 12  2 star ratings     1830 non-null   int64  
 13  1 star ratings     1830 non-null   int64  
dtypes: float64(4), int64(7), object(3)
memory usage: 200.3+ KB
```

В [9]:
```
# Заполняем отсутствующие значения
data['1 star ratings'] = data['1 star ratings'].replace(0,np.nan)
data['1 star ratings'] = data['1 star ratings'].fillna(data['1 star ratings'].mean())
```

В [10]:
```
data.head()
```

Out[10]:

	rank	title	total ratings	installs	average rating	growth (30 days)	growth (60 days)	price	category	5 star ratings	4 star ratings	3 star ratings	2 star ratings	1 star ratings
0	1	Garena Free Fire - The Cobra	80678661	500.0 M	4.33	2.9	7.9	0.0	GAME ACTION	61935712	4478738	2795172	1814999	9654037
1	2	PUBG MOBILE: Graffiti Prank	35971961	100.0 M	4.24	2.0	3.1	0.0	GAME ACTION	26670566	2109631	1352610	893674	4945478
2	3	Mobile Legends: Bang Bang	25836869	100.0 M	4.08	1.6	3.3	0.0	GAME ACTION	17850942	1796761	1066095	725429	4397640
3	4	Brawl Stars	17181659	100.0 M	4.27	4.1	6.6	0.0	GAME ACTION	12493668	1474319	741410	383478	2088781
4	5	Sniper 3D: Fun Free Online FPS Shooting Game	14237554	100.0 M	4.33	0.8	1.8	0.0	GAME ACTION	9657878	2124544	1034025	375159	1045945

В [11]:
```
data.isnull().sum()
# проверим есть ли пропущенные значения
```

Out[11]:
```
rank               0
title              0
total ratings      0
installs           0
average rating     0
growth (30 days)   0
growth (60 days)   0
price              0
category           0
5 star ratings     0
4 star ratings     0
3 star ratings     0
2 star ratings     0
1 star ratings     0
dtype: int64
```