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

Отчет

Рубежный контроль № 1
Вариант № 17

По курсу «Технологии машинного обучения»

ИСПОЛНИТЕЛЬ:
Молева Анастасия
Группа ИУ5-61Б

_____

"__"_____2020 г.

ПРЕПОДАВАТЕЛЬ:
Гапанюк Ю.Е.

_____

"__"_____2020 г.

Москва 2020

Молева А.А. ИУ5-61Б, В-17

Задача № 3

Для заданного набора данных произведите масштабирование данных (для одного признака) и преобразование категориальных признаков в количественные двумя способами (label encoding, one hot encoding) для одного признака. Какие методы Вы использовали для решения задачи и почему?

In [ ]:
```
import pandas as pd
import numpy as np
import zipfile
```

In [121]:
```
import os
import shutil

CUR_DIR = os.path.join(os.path.curdir, 'fifa19.zip')
DATA_PATH = os.path.join("datasets")
def fetch_data(data_path=DATA_PATH, cur_dir=CUR_DIR):
    os.makedirs(data_path, exist_ok=True)
    if os.path.isfile('fifa19.zip'):
        shutil.move(cur_dir, data_path)
    zip_path = os.path.join(data_path, "fifa19.zip")
    zip_file = zipfile.ZipFile(zip_path)
    zip_file.extractall(path=data_path)
    zip_file.close()

fetch_data()
```

```
data = load_data()
data
```

	Unnamed: 0	ID	Name	Age	Photo	Nationality	Flag
0	0	158023	L. Messi	31	https://cdn.sofifa.org/players/4/19/158023.png	Argentina	https://cdn.sofifa.org/
1	1	20801	Cristiano Ronaldo	33	https://cdn.sofifa.org/players/4/19/20801.png	Portugal	https://cdn.sofifa.org/
2	2	190871	Neymar Jr	26	https://cdn.sofifa.org/players/4/19/190871.png	Brazil	https://cdn.sofifa.org/
3	3	193080	De Gea	27	https://cdn.sofifa.org/players/4/19/193080.png	Spain	https://cdn.sofifa.org/
4	4	192985	K. De Bruyne	27	https://cdn.sofifa.org/players/4/19/192985.png	Belgium	https://cdn.sofifa.org/
...	...	...	...	...	...	...	...
18202	18202	238813	J. Lundstram	19	https://cdn.sofifa.org/players/4/19/238813.png	England	https://cdn.sofifa.org/
18203	18203	243165	N. Christoffersson	19	https://cdn.sofifa.org/players/4/19/243165.png	Sweden	https://cdn.sofifa.org/
18204	18204	241638	B. Worman	16	https://cdn.sofifa.org/players/4/19/241638.png	England	https://cdn.sofifa.org/
18205	18205	246268	D. Walker-Rice	17	https://cdn.sofifa.org/players/4/19/246268.png	England	https://cdn.sofifa.org/
18206	18206	246269	G. Nugent	16	https://cdn.sofifa.org/players/4/19/246269.png	England	https://cdn.sofifa.org/

18207 rows × 89 columns

Масштабирование данных (одного признака)

```
: from sklearn.preprocessing import StandardScaler
```

Отмасштабируем 'Release Clause'

Для начала избавимся от пустых строк

```
: data['Release Clause'].isnull().sum()
```

```
: 1564
```

```
: data = data.dropna(axis='index', how='any', subset=['Release Clause'])
```

Преобразуем строки в числа

```
data['Release Clause'] = data['Release Clause'].map(lambda x: str(x)[1:])
data['Release Clause']
```

```
d:\user\desktop\untitled\venv\lib\site-packages\ipykernel_launcher.py:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: http://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html
#returning-a-view-versus-a-copy
  """Entry point for launching an IPython kernel.
```

```
0        226.5M
1        127.1M
2        228.1M
3        138.6M
4        196.4M
          ...
18202      143K
18203      113K
18204      165K
18205      143K
18206      165K
Name: Release Clause, Length: 16643, dtype: object
```

Масштабируем признак

```
scaler = StandardScaler()
scaler.fit(data[['Release Clause']])
data_1 = scaler.transform(data[['Release Clause']])
data['Release Clause'] = data_1
```

```
d:\user\desktop\untitled\venv\lib\site-packages\ipykernel_launcher.py:4: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: http://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html
#returning-a-view-versus-a-copy
  after removing the cwd from sys.path.
```

```
data
```

Преобразование категориальных признаков в количественные

Преобразуем признак 'Club'

Метод LabelEncoder

```
from sklearn.preprocessing import LabelEncoder
```

```
le = LabelEncoder()
cat_le = le.fit_transform(data['Club'])
cat_le
```

```
array([212, 326, 435, ..., 122, 586, 586])
```

```
np.unique(cat_le)
```

```
array([  0,   1,   2,   3,   4,   5,   6,   7,   8,   9,  10,  11,  12,
        13,  14,  15,  16,  17,  18,  19,  20,  21,  22,  23,  24,  25,
        26,  27,  28,  29,  30,  31,  32,  33,  34,  35,  36,  37,  38,
        39,  40,  41,  42,  43,  44,  45,  46,  47,  48,  49,  50,  51,
        52,  53,  54,  55,  56,  57,  58,  59,  60,  61,  62,  63,  64,
        65,  66,  67,  68,  69,  70,  71,  72,  73,  74,  75,  76,  77,
        78,  79,  80,  81,  82,  83,  84,  85,  86,  87,  88,  89,  90,
        91,  92,  93,  94,  95,  96,  97,  98,  99, 100, 101, 102, 103,
       104, 105, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116,
       117, 118, 119, 120, 121, 122, 123, 124, 125, 126, 127, 128, 129,
       130, 131, 132, 133, 134, 135, 136, 137, 138, 139, 140, 141, 142,
       143, 144, 145, 146, 147, 148, 149, 150, 151, 152, 153, 154, 155,
       156, 157, 158, 159, 160, 161, 162, 163, 164, 165, 166, 167, 168,
       169, 170, 171, 172, 173, 174, 175, 176, 177, 178, 179, 180, 181,
       182, 183, 184, 185, 186, 187, 188, 189, 190, 191, 192, 193, 194,
```

OneHotEncoder

```
from sklearn.preprocessing import OneHotEncoder

ohe = OneHotEncoder()
cat_ohe = ohe.fit_transform(data[['Club']])
cat_ohe.A
```

```
array([[0., 0., 0., ..., 0., 0., 0.],
       [0., 0., 0., ..., 0., 0., 0.],
       [0., 0., 0., ..., 0., 0., 0.],
       ...,
       [0., 0., 0., ..., 0., 0., 0.],
       [0., 0., 0., ..., 0., 0., 0.],
       [0., 0., 0., ..., 0., 0., 0.]])
```

```
pd.get_dummies(data['Club'])
```

Можно также сделать с национальностью

```
pd.get_dummies(data['Nationality'])
```

	Afghanistan	Albania	Algeria	Andorra	Angola	Antigua & Barbuda	Argentina	Armenia	Australia	Austria	...	Uganda	Ukrai
0	0	0	0	0	0	0	1	0	0	0	...	0	0
1	0	0	0	0	0	0	0	0	0	0	...	0	0
2	0	0	0	0	0	0	0	0	0	0	...	0	0
3	0	0	0	0	0	0	0	0	0	0	...	0	0
4	0	0	0	0	0	0	0	0	0	0	...	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
18202	0	0	0	0	0	0	0	0	0	0	...	0	0
18203	0	0	0	0	0	0	0	0	0	0	...	0	0
18204	0	0	0	0	0	0	0	0	0	0	...	0	0
18205	0	0	0	0	0	0	0	0	0	0	...	0	0
18206	0	0	0	0	0	0	0	0	0	0	...	0	0