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

Отчет
Рубежный контроль №1
По курсу «Технологии машинного обучения»

Вариант 9

ИСПОЛНИТЕЛЬ:

Меркулова Надежда
Группа ИУ5-64

_____

"__"_____2020 г.

ПРЕПОДАВАТЕЛЬ:

Гапанюк. Ю.Е.

_____

"__"_____2020 г.

Москва 2020

1. **Условие**

Задача №2:

Для заданного набора данных проведите обработку пропусков в данных для одного категориального и одного количественного признака. Какие способы обработки пропусков в данных для категориальных и количественных признаков Вы использовали? Какие признаки Вы будете использовать для дальнейшего построения моделей машинного обучения и почему?

Набор данных №1:

https://www.kaggle.com/karangadiya/fifa19

Дополнительное требование:

Для произвольной колонки данных построить график «Скрипичная диаграмма» (violin plot).

2. **Выполнение**

См. на следующей странице

In [1]:

```
import pandas as pd
import numpy as np
```

Извлечение dataset

In [2]:

```
data = pd.read_csv('./data.csv')
data
```

Out[2]:

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

18207 rows × 89 columns

Обработка пропусков в данных

Проверим, есть ли пропущенные значения

In [3]:

```
data.isnull().sum()
```

Out[3]:

```
Unnamed: 0            0
ID                    0
Name                  0
Age                   0
Photo                 0
                   ...
GKHandling           48
GKKicking            48
GKPositioning        48
GKReflexes           48
Release Clause     1564
Length: 89, dtype: int64
```

1. Замена пустых значений на среднее

Выполним замену для количественного признака GKReflexes.

In [4]:

```
1. Количество нулевых значений:
```

```
  File "<ipython-input-4-bb94aec8515a>", line 1
    1. Количество нулевых значений:
                  ^
SyntaxError: invalid syntax
```

In [5]:

```
data['GKReflexes'].isna().sum()
```

Out[5]:

```
48
```

1. Получим среднее:

In [6]:

```
mean = data['GKReflexes'].mean()
mean
```

Out[6]:

```
16.710887163390055
```

1. Выполним замену и проверим количество пустых значений:

In [7]:

```
data['GKReflexes'].fillna(mean, inplace=True)
data['GKReflexes'].isna().sum()
```

Out[7]:

```
0
```

2. Удаление пустых значений

Выполним удаление для категориального признака Club.

1. Количество нулевых значений:

In [8]:

```
data[data['Club'].isna()]['Club'] = ''
data['Club'].isna().sum()
```

Out[8]:

```
241
```

1. Удалим строки, содержащие нулевое значение колонки Club:

In [12]:

```
data = data[~data['Club'].isna()]
data
```

Out[12]:

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

17966 rows × 89 columns

Как можно видеть, количество строк датасета уменьшилось.

1. Проверим количество пустых значений поля Club:

In [13]:

```
data['Club'].isna().sum()
```

Out[13]:

0

Дополнительное задание

Построим график "Скрипичная диаграмма" (Violin plot) для поля Potential

In [10]:

```
import seaborn as sns
sns.violinplot(x=data['Potential'])
```

Out[10]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x11517edd8>
```