**Московский государственный технический**

**университет им. Н.Э. Баумана**

Факультет «Информатика с системы управления»

Кафедра ИУ5 «Системы обработки информации и управления»

Курс «Технологии машинного обучения»

Отчёт по лабораторной работе №2

Выполнил: Проверил:

студент группы РТ5-61Б преподаватель каф. ИУ5

Кузнецов А.В. Гапанюк Ю.Е.

Подпись и дата: Подпись и дата:

2023 г.

from google.colab import drive  
drive.mount('/content/drive')

Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.mount("/content/drive", force\_remount=True).

# Описание датасета

Набор данных представляет собой список 200 самых популярных Instagram аккаунтов. Набор включает в себя никнейм в социальной сети, количество подписчиков, постов, "рейтинг влияния", страну и некоторые другие метрики.

Вывод датасета. В наборе присутствуют пропущенные данные и один категориальный признак.

import pandas as pd  
import numpy as np  
  
filename = '/content/drive/My Drive/МГТУ/6 семестр/ТМО/lab02/lab02tmo.csv'  
ds = pd.read\_csv(filename)  
# pd.set\_option('max\_rows', 6)  
  
pd.set\_option('display.max\_colwidth', None)  
pd.set\_option('display.float\_format', '{:.2f}'.format)  
ds.columns = ['Место', 'Никнейм', 'Рейтинг влияния' , 'Публикации', 'Подписчики(млн)', 'Среднее кол-во лайков', 'Вовлеченность', 'Среднее кол-во лайков на новой публикации', 'Всего лайков', 'Страна']  
ds = pd.DataFrame(ds)  
display(ds)

Место Никнейм Рейтинг влияния Публикации Подписчики(млн) \  
0 1 cristiano 92 3.3k 475.80   
1 2 kyliejenner 91 6.9k 366.20   
2 3 leomessi 90 0.89k 357.30   
3 4 selenagomez 93 1.8k 342.70   
4 5 therock 91 6.8k 334.10   
.. ... ... ... ... ...   
195 196 iambeckyg 71 2.3k 33.20   
196 197 nancyajram 81 3.8k 33.20   
197 198 luansantana 79 0.77k 33.20   
198 199 nickjonas 78 2.3k 33.00   
199 200 raisa6690 80 4.2k 32.80   
  
 Среднее кол-во лайков Вовлеченность \  
0 8.7m 1.39%   
1 8.3m 1.62%   
2 6.8m 1.24%   
3 6.2m 0.97%   
4 1.9m 0.20%   
.. ... ...   
195 623.8k 1.40%   
196 390.4k 0.64%   
197 193.3k 0.26%   
198 719.6k 1.42%   
199 232.2k 0.30%   
  
 Среднее кол-во лайков на новой публикации Всего лайков Страна   
0 6.5m 29.0b Spain   
1 5.9m 57.4b United States   
2 4.4m 6.0b NaN   
3 3.3m 11.5b United States   
4 665.3k 12.5b United States   
.. ... ... ...   
195 464.7k 1.4b United States   
196 208.0k 1.5b France   
197 82.6k 149.2m Brazil   
198 467.7k 1.7b United States   
199 97.4k 969.1m Indonesia   
  
[200 rows x 10 columns]

Для удаления пропусков в данных используется функция dropna. После удаления в датасете осталось 153 строки.

ds = ds.dropna()  
display(ds)

Место Никнейм Рейтинг влияния Публикации Подписчики(млн) \  
0 1 cristiano 92 3.3k 475.80   
1 2 kyliejenner 91 6.9k 366.20   
3 4 selenagomez 93 1.8k 342.70   
4 5 therock 91 6.8k 334.10   
5 6 kimkardashian 91 5.6k 329.20   
.. ... ... ... ... ...   
195 196 iambeckyg 71 2.3k 33.20   
196 197 nancyajram 81 3.8k 33.20   
197 198 luansantana 79 0.77k 33.20   
198 199 nickjonas 78 2.3k 33.00   
199 200 raisa6690 80 4.2k 32.80   
  
 Среднее кол-во лайков Вовлеченность \  
0 8.7m 1.39%   
1 8.3m 1.62%   
3 6.2m 0.97%   
4 1.9m 0.20%   
5 3.5m 0.88%   
.. ... ...   
195 623.8k 1.40%   
196 390.4k 0.64%   
197 193.3k 0.26%   
198 719.6k 1.42%   
199 232.2k 0.30%   
  
 Среднее кол-во лайков на новой публикации Всего лайков Страна   
0 6.5m 29.0b Spain   
1 5.9m 57.4b United States   
3 3.3m 11.5b United States   
4 665.3k 12.5b United States   
5 2.9m 19.9b United States   
.. ... ... ...   
195 464.7k 1.4b United States   
196 208.0k 1.5b France   
197 82.6k 149.2m Brazil   
198 467.7k 1.7b United States   
199 97.4k 969.1m Indonesia   
  
[138 rows x 10 columns]

Кодирование категориальных признаков было сделано с помощью Label Encoder. Данный тип кодирования является наиболее часто используемым, преобразование представляет собой однозначное соответствие число <-> уникальное значение категориального признака.

from sklearn.preprocessing import LabelEncoder  
labelencoder = LabelEncoder()  
ds['Страна'] = labelencoder.fit\_transform(ds['Страна'].values)  
display(ds)

Место Никнейм Рейтинг влияния Публикации Подписчики(млн) \  
0 1 cristiano 92 3.3k 475.80   
1 2 kyliejenner 91 6.9k 366.20   
3 4 selenagomez 93 1.8k 342.70   
4 5 therock 91 6.8k 334.10   
5 6 kimkardashian 91 5.6k 329.20   
.. ... ... ... ... ...   
195 196 iambeckyg 71 2.3k 33.20   
196 197 nancyajram 81 3.8k 33.20   
197 198 luansantana 79 0.77k 33.20   
198 199 nickjonas 78 2.3k 33.00   
199 200 raisa6690 80 4.2k 32.80   
  
 Среднее кол-во лайков Вовлеченность \  
0 8.7m 1.39%   
1 8.3m 1.62%   
3 6.2m 0.97%   
4 1.9m 0.20%   
5 3.5m 0.88%   
.. ... ...   
195 623.8k 1.40%   
196 390.4k 0.64%   
197 193.3k 0.26%   
198 719.6k 1.42%   
199 232.2k 0.30%   
  
 Среднее кол-во лайков на новой публикации Всего лайков Страна   
0 6.5m 29.0b 17   
1 5.9m 57.4b 23   
3 3.3m 11.5b 23   
4 665.3k 12.5b 23   
5 2.9m 19.9b 23   
.. ... ... ...   
195 464.7k 1.4b 23   
196 208.0k 1.5b 8   
197 82.6k 149.2m 2   
198 467.7k 1.7b 23   
199 97.4k 969.1m 11   
  
[138 rows x 10 columns]

Для просмотра списка закодированных значений используется функция labelencoder.classes\_ . Реализация Label Encoder в sklearn прежде всего сортирует по алфавиту уникальные значения, потом присваивает им порядковый номер. Таким образом уникальный номер категориального признака -- это его индекс в данном массиве.

labelencoder.classes\_

array(['Anguilla', 'Australia', 'Brazil', 'British Virgin Islands',  
 'Canada', 'Colombia', 'Czech Republic', "CÃ´te d'Ivoire", 'France',  
 'Germany', 'India', 'Indonesia', 'Italy', 'Mexico', 'Netherlands',  
 'Puerto Rico', 'Russia', 'Spain', 'Sweden', 'Switzerland',  
 'Turkey', 'United Arab Emirates', 'United Kingdom',  
 'United States', 'Uruguay'], dtype=object)

При нормализации меняем масштаб данных. Чаще всего масштабирование данных изменяется в диапазоне от 0 до 1.

from sklearn.datasets import load\_iris  
from sklearn import preprocessing  
x\_array = np.array(ds['Рейтинг влияния'])  
normalized\_arr = preprocessing.normalize([x\_array])  
ds['Рейтинг влияния'] = normalized\_arr[0]  
  
  
x\_array = np.array(ds['Подписчики(млн)'])  
normalized\_arr = preprocessing.normalize([x\_array])  
ds['Подписчики(млн)'] = normalized\_arr[0]  
display(ds)

Место Никнейм Рейтинг влияния Публикации Подписчики(млн) \  
0 1 cristiano 0.10 3.3k 0.35   
1 2 kyliejenner 0.09 6.9k 0.27   
3 4 selenagomez 0.10 1.8k 0.25   
4 5 therock 0.09 6.8k 0.24   
5 6 kimkardashian 0.09 5.6k 0.24   
.. ... ... ... ... ...   
195 196 iambeckyg 0.07 2.3k 0.02   
196 197 nancyajram 0.08 3.8k 0.02   
197 198 luansantana 0.08 0.77k 0.02   
198 199 nickjonas 0.08 2.3k 0.02   
199 200 raisa6690 0.08 4.2k 0.02   
  
 Среднее кол-во лайков Вовлеченность \  
0 8.7m 1.39%   
1 8.3m 1.62%   
3 6.2m 0.97%   
4 1.9m 0.20%   
5 3.5m 0.88%   
.. ... ...   
195 623.8k 1.40%   
196 390.4k 0.64%   
197 193.3k 0.26%   
198 719.6k 1.42%   
199 232.2k 0.30%   
  
 Среднее кол-во лайков на новой публикации Всего лайков Страна   
0 6.5m 29.0b 17   
1 5.9m 57.4b 23   
3 3.3m 11.5b 23   
4 665.3k 12.5b 23   
5 2.9m 19.9b 23   
.. ... ... ...   
195 464.7k 1.4b 23   
196 208.0k 1.5b 8   
197 82.6k 149.2m 2   
198 467.7k 1.7b 23   
199 97.4k 969.1m 11   
  
[138 rows x 10 columns]