t3

February 17, 2021

1 Integração e pré-processamento das bases de dados de jogos eletrônicos

Para esta tarefa, utilizaremos somente as biblioteca `pandas`, `numpy` e `scikitlearn`.

[168]:
```
import pandas as pd
import numpy as np
from sklearn.preprocessing import MinMaxScaler

df1 = pd.read_csv('trabalho2_dados_1.csv')
df2 = pd.read_csv('trabalho3_dados_adicionais_1.csv')
df = pd.concat([df1, df2])

print("Total de jogos na base de dados 1: {}".format(len(df1)))
print("Total de jogos na base de dados 2: {}".format(len(df2)))

print("\nBase de dados resultante da operação de concatenação:")

df
```

```
Total de jogos na base de dados 1: 1632
Total de jogos na base de dados 2: 1588

Base de dados resultante da operação de concatenação:
```

[168]:

	nome	plataforma	genero	\
0	Jelly Belly: Ballistic Beans	Wii	Puzzle	
1	Madden NFL 12	PSP	Sports	
2	The Sims 2: Pets	Wii	Simulation	
3	Guilty Gear XX Accent Core Plus	PSP	Fighting	
4	WWE 2K14	PS3	Sports	
…	…	…	…	
1583	Major League Baseball 2K8	PSP	Sports	
1584	Monster Hunter Frontier Online	PS3	Role-Playing	
1585	Battle vs. Chess	PS3	Misc	
1586	AKB1/48: Idol to Guam de Koishitara…	X360	Misc	
1587	PDC World Championship Darts 2008	PSP	Sports	

```
                    editora  vendas lancamento  avaliacao-criticos  \
0     Zoo Digital Publishing    0.02  21-Apr-09                 NaN
1           Electronic Arts    0.20  30-Aug-11                 NaN
2           Electronic Arts    0.46  12-Jun-07                65.0
3                     PQube    0.08   7-Apr-09                 NaN
4       Take-Two Interactive    0.72  29-Oct-13                74.0
…                        …     …        …                    …
1583                Unknown    0.03   3-Mar-08                63.0
1584                    NaN    0.03        NaN                 NaN
1585     TopWare Interactive    0.03        TBA                 NaN
1586                    NaN    0.01        NaN                 NaN
1587      Oxygen Interactive    0.01  16-Jun-09                43.0

      numero-criticos avaliacao-usuarios  numero-usuarios  \
0                 NaN                tbd              NaN
1                 NaN                tbd              NaN
2                15.0                tbd              NaN
3                 NaN                8.3              4.0
4                21.0                7.4             60.0
…                 …                  …                …
1583              5.0                tbd              NaN
1584              NaN                NaN              NaN
1585              NaN                NaN              NaN
1586              NaN                NaN              NaN
1587              7.0                tbd              NaN

                fabricante
0     Zoo Digital Publishing
1                EA Tiburon
2                     Maxis
3           Arc System Works
4                    Yuke's
…                        …
1583             Kush Games
1584                    NaN
1585     TopWare Interactive
1586                    NaN
1587      Oxygen Interactive

[3220 rows x 11 columns]
```

1.1 Tratamento de variáveis categóricas

Há somente dois atributos categóricos: **plataforma** e **gênero**.

[169]:
```
# Atributos categóricos
df['plataforma'] = df['plataforma'].astype('category')
df['genero'] = df['genero'].astype('category')

print("Plataformas: {}".format(df['plataforma'].cat.categories))
print("Gêneros: {}".format(df['genero'].cat.categories))
```

```
Plataformas: Index(['3DS', 'PS3', 'PS4', 'PSP', 'PSV', 'Wii', 'WiiU', 'X360',
'XOne'], dtype='object')
Gêneros: Index(['Action', 'Adventure', 'Fighting', 'Misc', 'Platform', 'Puzzle',
       'Racing', 'Role-Playing', 'Shooter', 'Simulation', 'Sports',
       'Strategy'],
      dtype='object')
```

1.2 Unicidade

O nome do jogo eletrônico atua como chave-primária, portanto é necessário remover os jogos que aparecem repetidos no *dataframe*. Para manter a indexação consistente, utiliza-se a função `reset_index`.

[170]:
```
# Jogos são únicos
df = df.drop_duplicates(subset = ['nome'])
df = df.reset_index(drop = True)

df[['nome']]
```

[170]:

	nome
0	Jelly Belly: Ballistic Beans
1	Madden NFL 12
2	The Sims 2: Pets
3	Guilty Gear XX Accent Core Plus
4	WWE 2K14
…	…
1442	D.C. III: Da Capo III
1443	Street Fighter X Tekken
1444	John Daly's ProStroke Golf
1445	Backyard Sports Football: Rookie Rush
1446	Darkstalkers Resurrection

[1447 rows x 1 columns]

1.3 Tratamento de dados faltantes

Foram identificados os seguintes atributos faltantes em algumas instâncias do *dataframe*:

[171]:
```
df.isnull().sum()
```

[171]:
```
nome                    0
plataforma              0
genero                  0
editora                 4
vendas                  0
lancamento            455
avaliacao-criticos    639
numero-criticos       639
avaliacao-usuarios    470
numero-usuarios       693
fabricante            459
dtype: int64
```

Neste momento, há somente preocupação em tratar dados **numéricos** faltantes, uma vez que estes são os mais sensíveis para a aplicação dos métodos de aprendizado de máquina:

- avaliacao-criticos
- numero-criticos
- avaliacao-usuarios
- numero-usuarios

O preenchimento de dados faltantes será feito através da média geral do gênero do jogo em questão.

[172]:
```
# Converter valores 'tbd' para NaN
df['avaliacao-usuarios'] = pd.to_numeric(df['avaliacao-usuarios'], errors =␣
 ↪'coerce')

# Agrupando por gênero e coletando a média para os valores faltantes
df['avaliacao-criticos'] = df.groupby('genero')['avaliacao-criticos'].
 ↪transform(lambda x: x.fillna(x.mean()))
df['avaliacao-usuarios'] = df.groupby('genero')['avaliacao-usuarios'].
 ↪transform(lambda x: x.fillna(x.mean()))
df['numero-criticos'] = df.groupby('genero')['numero-criticos'].
 ↪transform(lambda x: x.fillna(x.mean())).astype('int')
df['numero-usuarios'] = df.groupby('genero')['numero-usuarios'].
 ↪transform(lambda x: x.fillna(x.mean())).astype('int')

df[['nome', 'genero', 'avaliacao-criticos', 'avaliacao-usuarios',␣
 ↪'numero-criticos', 'numero-usuarios']]
```

[172]:

	nome	genero	avaliacao-criticos	\
0	Jelly Belly: Ballistic Beans	Puzzle	62.095238	
1	Madden NFL 12	Sports	66.848485	
2	The Sims 2: Pets	Simulation	65.000000	
3	Guilty Gear XX Accent Core Plus	Fighting	70.160000	
4	WWE 2K14	Sports	74.000000	
…	…	…	…	
1442	D.C. III: Da Capo III	Adventure	67.439024	

```
1443                Street Fighter X Tekken    Fighting           83.000000
1444             John Daly's ProStroke Golf      Sports           57.000000
1445  Backyard Sports Football: Rookie Rush      Sports           66.848485
1446              Darkstalkers Resurrection    Fighting           80.000000

      avaliacao-usuarios  numero-criticos  numero-usuarios
0               6.133333               17               10
1               6.804167               22               29
2               7.558333               15               31
3               8.300000               30                4
4               7.400000               21               60
…                    …                …                …
1442            7.119048               32              331
1443            4.200000               45              162
1444            6.804167                5               29
1445            6.804167               22               29
1446            7.000000               33               26

[1447 rows x 6 columns]
```

1.4 Normalização dos dados

- Para que métodos de aprendizado de máquina operem melhor sobre atributos categóricos, utiliza-se o método `get_dummies()` para gerar um atributo binário para cada categoria existente.
- Além disso, todos os dados númericos foram normalizados no intervalo [0, 1] utilizando o scaler `MinMaxScaler`.

[173]:
```
# Similar ao LabelBinarizer
df = pd.get_dummies(df, columns = ['plataforma', 'genero'])

df[['vendas', 'avaliacao-criticos', 'numero-criticos', 'avaliacao-usuarios',␣
 ↪'numero-usuarios']] = MinMaxScaler().fit_transform(df[['vendas',␣
 ↪'avaliacao-criticos', 'numero-criticos', 'avaliacao-usuarios',␣
 ↪'numero-usuarios']])

df
```

[173]:
```
                                  nome                 editora    vendas  \
0              Jelly Belly: Ballistic Beans  Zoo Digital Publishing  0.000122
1                           Madden NFL 12        Electronic Arts  0.002322
2                       The Sims 2: Pets        Electronic Arts  0.005499
3           Guilty Gear XX Accent Core Plus                 PQube  0.000855
4                                 WWE 2K14  Take-Two Interactive  0.008677
…                                      …                    …       …
1442                D.C. III: Da Capo III         Kadokawa Games  0.000244
```

```
1443               Street Fighter X Tekken                 Capcom  0.004644
1444            John Daly's ProStroke Golf               O-Games  0.000000
1445  Backyard Sports Football: Rookie Rush                 Atari  0.000855
1446             Darkstalkers Resurrection                 Capcom  0.000244

     lancamento  avaliacao-criticos  numero-criticos  avaliacao-usuarios  \
0     21-Apr-09            0.577591         0.126214            0.627451
1     30-Aug-11            0.633512         0.174757            0.706373
2     12-Jun-07            0.611765         0.106796            0.795098
3      7-Apr-09            0.672471         0.252427            0.882353
4     29-Oct-13            0.717647         0.165049            0.776471
…         …                   …               …                   …
1442        NaN            0.640459         0.271845            0.743417
1443   6-Mar-12            0.823529         0.398058            0.400000
1444   Canceled            0.517647         0.009709            0.706373
1445        NaN            0.633512         0.174757            0.706373
1446  12-Mar-13            0.788235         0.281553            0.729412

      numero-usuarios              fabricante  plataforma_3DS  …  \
0            0.000710  Zoo Digital Publishing               0  …
1            0.002958             EA Tiburon               0  …
2            0.003194                  Maxis               0  …
3            0.000000       Arc System Works               0  …
4            0.006625                Yuke's               0  …
…               …                      …              …  …
1442         0.038684                    NaN              0  …
1443         0.018692                Capcom               0  …
1444         0.002958            Gusto Games              0  …
1445         0.002958                    NaN              0  …
1446         0.002603    Iron Galaxy Studios              0  …

      genero_Fighting  genero_Misc  genero_Platform  genero_Puzzle  \
0                   0            0                0              1
1                   0            0                0              0
2                   0            0                0              0
3                   1            0                0              0
4                   0            0                0              0
…                 …            …                …              …
1442                0            0                0              0
1443                1            0                0              0
1444                0            0                0              0
1445                0            0                0              0
1446                1            0                0              0

      genero_Racing  genero_Role-Playing  genero_Shooter  genero_Simulation  \
0                 0                    0               0                  0
1                 0                    0               0                  0
```

```
2                 0                     0               0                 1
3                 0                     0               0                 0
4                 0                     0               0                 0
…               …                     …               …               …
1442              0                     0               0                 0
1443              0                     0               0                 0
1444              0                     0               0                 0
1445              0                     0               0                 0
1446              0                     0               0                 0

      genero_Sports  genero_Strategy
0                 0                0
1                 1                0
2                 0                0
3                 0                0
4                 1                0
…               …                …
1442              0                0
1443              0                0
1444              1                0
1445              1                0
1446              0                0

[1447 rows x 30 columns]
```

1.5 Detecção de outliers

Para esta base de dados, considerou-se como **outliers** instâncias que apresentam valores extremos (para mais ou para menos) no atributo **vendas**, com intervalo de confiança de 95%. O motivo para esta decisão vem do fato de que, além de ser um atributo importante, todas as instâncias continham um valor inicial daquele, não necessitando de preenchimento artificial.

[174]:
```
df = df[np.abs(df['vendas'] - df['vendas'].mean()) <= 1.96 * df['vendas'].std()]

df
```

[174]:
```
                                  nome                 editora    vendas  \
0              Jelly Belly: Ballistic Beans  Zoo Digital Publishing  0.000122
1                             Madden NFL 12        Electronic Arts  0.002322
2                         The Sims 2: Pets        Electronic Arts  0.005499
3           Guilty Gear XX Accent Core Plus                 PQube  0.000855
4                                  WWE 2K14  Take-Two Interactive  0.008677
…                                        …                     …       …
1442                 D.C. III: Da Capo III         Kadokawa Games  0.000244
1443               Street Fighter X Tekken                Capcom  0.004644
1444           John Daly's ProStroke Golf               O-Games  0.000000
```

```
1445  Backyard Sports Football: Rookie Rush                   Atari  0.000855
1446              Darkstalkers Resurrection                  Capcom  0.000244

     lancamento  avaliacao-criticos  numero-criticos  avaliacao-usuarios  \
0     21-Apr-09            0.577591         0.126214            0.627451
1     30-Aug-11            0.633512         0.174757            0.706373
2     12-Jun-07            0.611765         0.106796            0.795098
3      7-Apr-09            0.672471         0.252427            0.882353
4     29-Oct-13            0.717647         0.165049            0.776471
…         …                   …                …                   …
1442        NaN            0.640459         0.271845            0.743417
1443   6-Mar-12            0.823529         0.398058            0.400000
1444   Canceled            0.517647         0.009709            0.706373
1445        NaN            0.633512         0.174757            0.706373
1446  12-Mar-13            0.788235         0.281553            0.729412

      numero-usuarios              fabricante  plataforma_3DS  …  \
0            0.000710  Zoo Digital Publishing               0  …
1            0.002958              EA Tiburon               0  …
2            0.003194                   Maxis               0  …
3            0.000000       Arc System Works               0  …
4            0.006625                 Yuke's               0  …
…               …                       …              …  …
1442         0.038684                     NaN              0  …
1443         0.018692                  Capcom              0  …
1444         0.002958             Gusto Games              0  …
1445         0.002958                     NaN              0  …
1446         0.002603    Iron Galaxy Studios               0  …

      genero_Fighting  genero_Misc  genero_Platform  genero_Puzzle  \
0                   0            0                0              1
1                   0            0                0              0
2                   0            0                0              0
3                   1            0                0              0
4                   0            0                0              0
…               …          …              …           …
1442                0            0                0              0
1443                1            0                0              0
1444                0            0                0              0
1445                0            0                0              0
1446                1            0                0              0

      genero_Racing  genero_Role-Playing  genero_Shooter  genero_Simulation  \
0                 0                    0               0                  0
1                 0                    0               0                  0
2                 0                    0               0                  1
3                 0                    0               0                  0
```

```
4                 0                    0               0                    0
…               …                    …               …                    …
1442              0                    0               0                    0
1443              0                    0               0                    0
1444              0                    0               0                    0
1445              0                    0               0                    0
1446              0                    0               0                    0

      genero_Sports  genero_Strategy
0                 0                0
1                 1                0
2                 0                0
3                 0                0
4                 1                0
…               …                …
1442              0                0
1443              0                0
1444              1                0
1445              1                0
1446              0                0

[1427 rows x 30 columns]
```

[ ]: