

Nomes: Samuel Matias Finkler e Thiago Sordi

Disciplina: Mineração de dados (ELC1098)

Trabalho prático 1 - Etapas de pré processamento do dataset em Python

Para o desenvolvimento deste trabalho foi decidido utilizar a linguagem Python, com a popular biblioteca do Pandas (para manipular dataframes). A primeira coisa feita foi a importação do dataset para o ambiente de execução. Em seguida, foram analisados que tipos de dados presentes, e como estavam representados no dataset.

Com os dados de entrada já importados, o próximo passo foi iniciar o tratamento dos dados presentes no dataframe importado, pois o mesmo apresenta algumas inconsistências e pontos a serem melhorados para uma boa análise sobre os dados.

O primeiro tratamento foi remover do dataframe a coluna onde todos os valores eram nulos, além de remover colunas com informações que não eram cruciais para a análise, como as colunas “Jogadore(a)s.1” e “Partida”.

Com as colunas desnecessárias removidas, a atenção foi voltada para a coluna de jogadores, para fazer um tratamento pois havia divergência entre dados (nomes) que queriam dizer a mesma coisa (representar a mesma pessoa). A coluna inteira foi transformada para lowercase, foram removidos os espaços entre os nomes e alguns caracteres especiais. Também foram retirados acentos de alguns nomes, corrigidos nomes escritos de forma errada, e por fim, eliminado o nome de um jogador que não fazia parte dos 7 jogadores listados na coluna de jogadores (ex: alonso).

Por fim, foram transformadas as colunas “Jogadores” e “Resultado” para OneHot Encode, ou seja, cada categoria (nome ou resultado) virou uma coluna, e em cada linha é representado um valor 0 ou 1 para cada uma dessas colunas. Com essa transformação final, foi possível passar o dataframe através da função *apriori* da biblioteca mlxtend.

Essa função *apriori* em conjunto com a função *association_rules* (o dataframe retornado pela *apriori* é parâmetro para o *association_rules*) retornam um dataframe com tuplas de jogadores, seus suportes e confiança.

Melhor combinação vencedora:

1º Paula e Ariana

2º Paula e Alejandro

3º Alejandro e Juan

| antecedents | consequents | antecedent support | consequent support | support | confidence | lift | leverage | conviction |
|--------------------|-------------|--------------------|--------------------|----------|------------|----------|----------|------------|
| (alejandro, juan) | (GANHOU) | 0.178808 | 0.549669 | 0.119205 | 0.666667 | 1.212851 | 0.020920 | 1.350993 |
| (paula, alejandro) | (GANHOU) | 0.165563 | 0.549669 | 0.119205 | 0.720000 | 1.309880 | 0.028201 | 1.608325 |
| (paula, ariana) | (GANHOU) | 0.231788 | 0.549669 | 0.132450 | 0.571429 | 1.039587 | 0.005044 | 1.050773 |

Melhor jogador:

Alejandro (suporte de 28,48%)

| antecedents | consequents | antecedent support | consequent support | support | confidence | lift | leverage | conviction |
|-------------|-------------|--------------------|--------------------|----------|------------|----------|-----------|------------|
| (alejandro) | (GANHOU) | 0.430464 | 0.549669 | 0.284768 | 0.661538 | 1.203522 | 0.048156 | 1.330524 |
| (ariana) | (GANHOU) | 0.443709 | 0.549669 | 0.251656 | 0.567164 | 1.031829 | 0.007763 | 1.040420 |
| (carlos) | (GANHOU) | 0.278146 | 0.549669 | 0.145695 | 0.523810 | 0.952955 | -0.007193 | 0.945695 |
| (juan) | (GANHOU) | 0.417219 | 0.549669 | 0.245033 | 0.587302 | 1.068464 | 0.015701 | 1.091187 |
| (lucia) | (GANHOU) | 0.264901 | 0.549669 | 0.125828 | 0.475000 | 0.864157 | -0.019780 | 0.857774 |
| (paula) | (GANHOU) | 0.417219 | 0.549669 | 0.238411 | 0.571429 | 1.039587 | 0.009079 | 1.050773 |

Pior jogador:

Ariana (suporte de 19,21%)

| antecedents | consequents | antecedent support | consequent support | support | confidence | lift | leverage | conviction |
|-------------|-------------|--------------------|--------------------|----------|------------|----------|-----------|------------|
| (alejandro) | (Perdeu) | 0.430464 | 0.450331 | 0.145695 | 0.338462 | 0.751584 | -0.048156 | 0.830895 |
| (ariana) | (Perdeu) | 0.443709 | 0.450331 | 0.192053 | 0.432836 | 0.961150 | -0.007763 | 0.969153 |
| (carlos) | (Perdeu) | 0.278146 | 0.450331 | 0.132450 | 0.476190 | 1.057423 | 0.007193 | 1.049368 |
| (fernando) | (Perdeu) | 0.205298 | 0.450331 | 0.132450 | 0.645161 | 1.432638 | 0.039998 | 1.549067 |
| (juan) | (Perdeu) | 0.417219 | 0.450331 | 0.172185 | 0.412698 | 0.916433 | -0.015701 | 0.935923 |
| (lucia) | (Perdeu) | 0.264901 | 0.450331 | 0.139073 | 0.525000 | 1.165809 | 0.019780 | 1.157198 |
| (paula) | (Perdeu) | 0.417219 | 0.450331 | 0.178808 | 0.428571 | 0.951681 | -0.009079 | 0.961921 |