UNIVERSIDADE FEDERAL DE SANTA MARIA

ELC1098 - MINERAÇÃO DE DADOS

Lorenzo Moreira Donatti

Ramon Panazollo

TRABALHO 1

SANTA MARIA, 21 DE NOVEMBRO DE 2022

INTRODUÇÃO

Para a resolução deste trabalho, foi-se utilizada a plataforma Google Colab e a linguagem Python para a extração das regras de associação.

As bibliotecas utilizadas foram: Pandas, Numpy e Mlxtend.

PRÉ PROCESSAMENTO

Na etapa do pré-processamento dos dados os dois Datasets disponibilizados foram baixados. O primeiro problema a ser resolvido foi a escolha da melhor codificação dos caracteres vindos do arquivo CSV, pois haviam caracteres como "ç" no nome dos integrantes. Experimentalmente foi descoberto que o arquivo 1 abria normalmente com a codificação UTF-8 e o arquivo 2 com a codificação "Latin1".

Resolvido este primeiro problema, foi necessário concatenar os dois DataFrames, além de excluir colunas irrelevantes, como "Partida", "Oponentes" e "Amigos".

Após isso, uma checagem de valores únicos foi feita, a fim de averiguar quais duplas e nomes eram mais frequentes, porém foi constatado que existiam caracteres que estavam interferindo na contagem, como por exemplo, espaços antes e depois da virgula ou letras maiúsculas e minúsculas estavam fazendo duplas formadas pelos mesmos integrantes serem classificadas como diferentes. Para resolver este problema algumas medidas foram tomadas, como o uso de funções para a remoção de espaços, a troca do caractere "ç" por "c", e a transformação de todas as letras para maiúsculo.

Depois dessa preparação houve uma diminuição considerável de valores únicos, significando sucesso na identificação de duplas/trios mais frequentes.

Além disso, outra etapa precisou ser feita para que o algoritmo apriori funcionasse perfeitamente, a coluna "Jogadores" precisou ser realocada para três diferentes colunas contendo o nome em separado de cada jogador, pois assim seria possível traçar individualmente os melhores jogadores e duplas.

Por fim, foi necessário o uso da função Transaction_encoder para a transformação do Dataset em valores booleanos, pois o algoritmo à priori disponibilizado pela biblioteca MLxtend aceita entradas apenas nesse modelo.

Depois de todo pré-processamento, foi finalmente possível a aplicação do algoritmo à priori, sendo estes os resultados:

MELHOR DUPLA: BARBARA E FRANÇOIS

PIOR DUPLA: FRANÇOIS E JIMMY

MELHOR JOGADOR: FRANÇOIS

Link para Google Colab: https://colab.research.google.com/drive/1H-kXhcGh-wkxjnxJlDc7eGMKkpPysXj7?usp=sharing