Utilização de KDD em dados de resultados olímpicos

Samuel Favarin, Vinícius Machado e José Henrique

Com os estudos e a implementação de Data Mining, podemos adquirir conhecimentos importantes e úteis de bases de dados, que muitas vezes parecem só servir para o acúmulo de dados. O objetivo desse projeto é aplicar conhecimentos adquiridos de Data Mining e KDD para validar hipóteses e responder perguntas relacionadas aos jogos olímpicos.

Neste projeto, conseguimos explorar alguns algoritmos clássicos de data mining, como árvores de decisão para classificação com o algoritmo J48, algoritmos de associação como o Apriori e algoritmos de predição com técnicas de regressão linear.

Foi utilizado a ferramenta WEKA para utilizar os algoritmos de data mining, e a utilização de banco de dados com SQL, para facilitar a exploração, tratamento e filtragem de dados.

Base de dados

O dataset escolhido para o desenvolvimento dos estudos de KDD e os estudos de Data Mining foi o *"120 years of Olympic history: athletes and results"* que possui os dados biológicos e dados de medalhas de cada participação dos atletas das olimpíadas de 1896 em Atenas até 2016 no Rio.

O dataset original possuí 271116 linhas e 15 colunas. Cada linha corresponde a um atleta individual que participou das olimpíadas de 1896 a 2016. As colunas tem os seguintes valores:

1. ID - Número único do atleta
2. Name - Nome do atleta
3. Sex - sexo do atleta (M ou F)
4. Age - Idade em número inteiro
5. Height - Altura em centímetros
6. Weight - Peso em kilogramas
7. Team - Nome do Time
8. NOC - Comitê nacional olímpico (sigla do país)
9. Games - Ano e estação
10. Year - Ano da competição
11. Season - Estação da competição (Inverno ou verão)
12. City - Cidade sede
13. Sport - Esporte
14. Event - Evento
15. Medal - Medalha: NA, Bronze, Prata ou Ouro

Questões levantadas:

1. Quais as características básicas dos atletas de ginástica? Ginastas possuem tendência de serem leves?
2. Quantos participantes o Brasil terá nas olimpíadas de 2020?
3. Sabendo que mais da metade dos brasileiros estão acima do peso. Nos esportes olímpicos, qual a relação de peso entre os atletas brasileiros? Este número aumentou com o decorrer das olimpíadas
4. Sabe-se que os Estados Unidos corresponde ao maior número de medalhas acumuladas na história das olimpíadas. Portanto, existe alguma relação entre os esportes onde foram participantes e premiados?

Pré Processamento

Eliminação de dados:

- **Eliminação de dados das olimpíadas de inverno:** Eliminamos todos os dados referentes às olimpíadas de inverno, pelo fato de não serem importantes para nossas questões, evitando o processamento de muito dado inutilizável.
- **Eliminação de nomes e Ids de atletas:** Não houve a necessidade de utilizar dados pontuais como os nomes dos atletas e ids, diminuindo assim também o processamento de dados inúteis.
- **Eliminação de outras Games:** Retiramos a coluna "Games", que basicamente concatenar o ano com o tipo da olimpíada(summer)
- **Eliminação da coluna Event:** Essa coluna especializava o esporte praticado. Por exemplo, o esporte é “Natação”, e o evento é Natação Masculina de 100m.
- **Eliminação da coluna Season:** Essa coluna visava apenas diferenciar se o evento era as olimpíadas de inverno, ou era as olimpíadas de verão.
- **Eliminação da coluna Team:** Eliminamos a coluna Team, que descrevia o time do atleta, como já temos a coluna NOC que é a sigla do país do atleta, se tornou uma coluna inútil. Notamos também a especialização de dados, por exemplo, o Team era Denmark/Sweden e o NOC era DEN referenciando apenas a Dinamarca.

Tratamento de Missing Values:

- Tratamento de valores nulos na coluna “Age”. Notamos a presença de valores nulos em idades, como achamos um dado importante, decidimos manter eles. Para tratar, preenchemos esses dados com a média de todas as idades dos atletas, e

arredondamos para transformar em um valor inteiro. A média obtida foi 25.64, e foi arredondado para cima 26.

- Tratamento de valores nulos em "Height". Para tratar preenchemos com a média também, o valor médio foi 175.4 e arredondamos para 175.
- Tratamento de valores nulos em “Weight”. Para tratar preenchemos para 70.5.

Tratamento de Outliers:

Em Weight notamos uma certa discrepância nos dados, em poucas linhas, cerca de 6 casos, notamos essas coluna ser preenchida com “733.33.33” ou “603.33.33”. Resolvemos tratar esses casos removendo essas linhas da base de dados, pelo fato de terem sido em poucos casos.

Para o cálculo do IMC (Índice de massa corporal), removemos os atletas menores de 20 anos e maiores de 65 que ficam fora da amplitude de idade para o cálculo.

Agrupamento de dados:

Com a exploração dos dados utilizando os algoritmos de classificação, notamos certa dificuldade que os algoritmos levavam para tratar as idades, pesos e altura. Havia uma variação muito grande desses valores. Para resolver esse problema, agrupados os dados em faixas etárias, faixas de peso e faixa de altura. A tabela abaixo demonstra como foi feito esse agrupamento:

Classificação de idades:

Faixa de Idades dos atletas	**Label de saída**
Menor de 20 anos	ADOLECENTE
De 20 anos até 29 anos	JOVEM
De 30 anos até 39	ADULTO
De 40 anos até 49	MEIA IDADE
Acima de 49	SENIOR

Classificação de altura:

Faixa de Pesos dos atletas	**Label de saída**
Abaixo de 149 cm	BAIXO
Acima de 149cm até 169 cm	ALTURA MEDIANA
Acima de 169cm até 189 cm	ALTO
Acima de 189cm	MUITO ALTO

Classificação de peso:

Faixa de Peso dos atletas	**Label de saída**
Menor de 60kg	PESO MUITO LEVE
Maior que 59kg e menor que 80kg	PESO LEVE
Maior que 79kg e menor que 100kg	PESO NORMAL
Maior que 99kg e menor que 120kg	PESADO
Maior que 119kg	MUITO PESADO

Outros tratamentos de dados:

- Transformação dos dados de medalhas (N/A, bronze, silver e gold) para Ganhou e não ganhou.
- Substituição de cidade sede para sigla de país sede.
- Inserção da coluna IMC (Índice de massa corporal).

Etapa de exploração dos dados:

Com os dados limpos, fomos para a etapa de exploração dos dados, fizemos análises e testamos diversos tipos de algoritmos.

1. Fizemos um levantamento do número de ganhadores de medalhas versus não ganhadores de nenhuma medalha, e confirmamos que existe muito mais ganhadores.
2. Alguns questionamentos eram simples de serem resolvidos com algumas linhas de script SQL,
3. Notamos que nas primeiros olimpíadas só participavam homens.
4. Notamos que os EUA possuem o maior número de medalhas acumuladas da história das olimpíadas.

Processo de Data Mining

“Quais as características básicas dos atletas de ginástica? Ginastas possuem tendência de serem leves?”

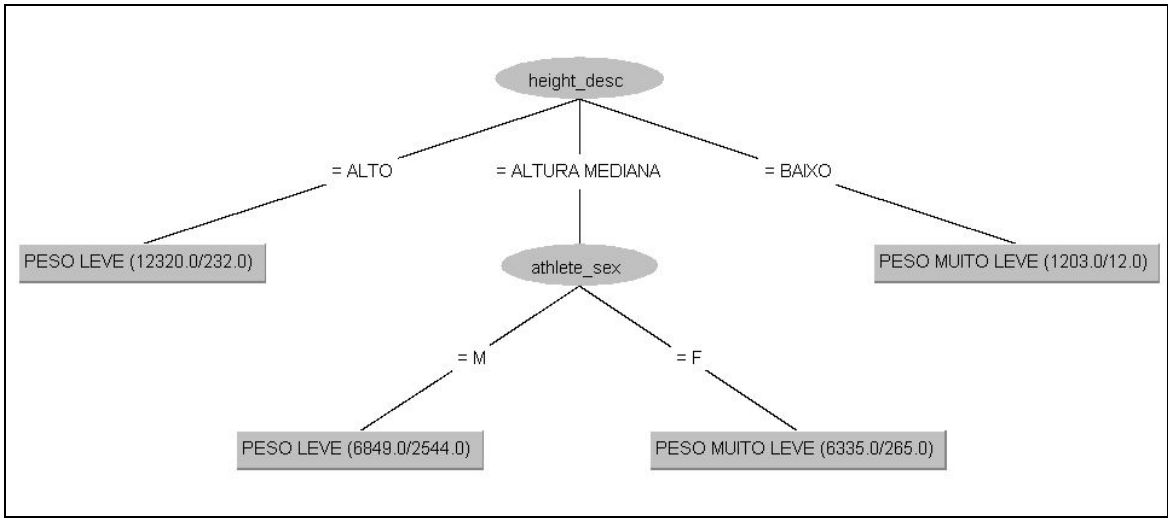
Para responder essas questões, utilizamos de técnicas de classificação para explorar os dados registrados na base de dados. Como queremos identificar as características básicas dos ganhadores de um esporte específico, então filtramos a base para recebermos apenas dados dos atletas que ganharam qualquer tipo de medalha em esportes de ginástica.

Para fazer isso, importamos a base de dados para um banco de dados local, e executamos uma query em SQL que facilitou o processo de filtragem de dados. Após aplicar o filtro, obtivemos 26707 linhas, cada qual representando os dados biológicos do atleta, e se obteve algum resultado ou não. As colunas utilizadas nessa análise foi: “athlete_sex”, “age”, “height”, “weight” e “medal”.

A princípio foi feito uma exploração com todos os atletas ginastas. Alguns conhecimentos prévios foram gerados, por exemplo:

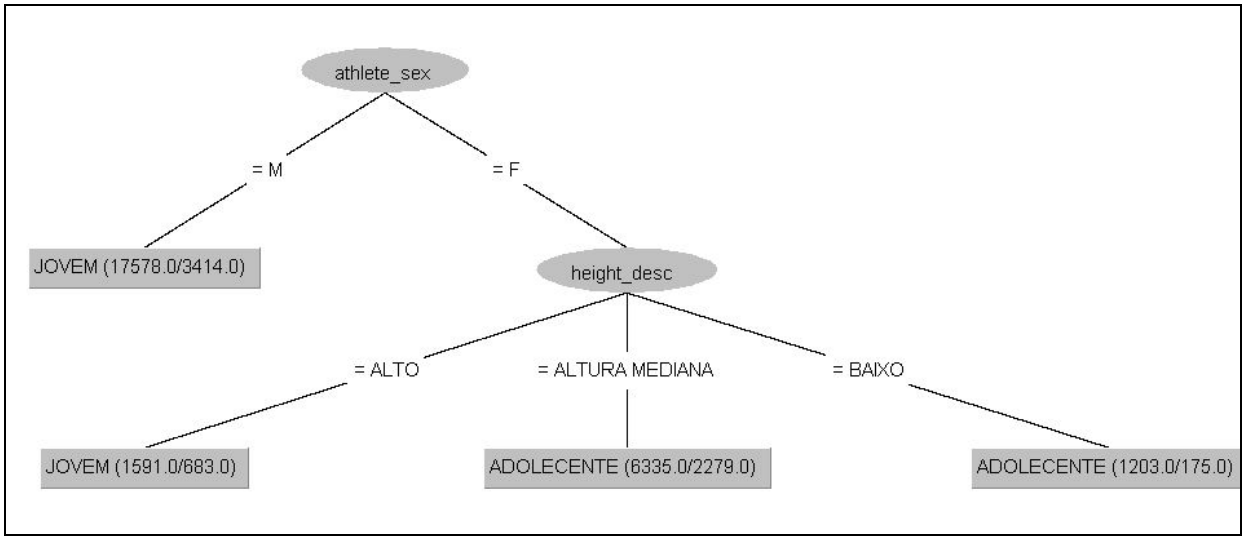
- A maioria dos atletas são jovens
- A maioria dos atletas tem altura de 150cm até 169cm
- A maioria dos atletas tem peso de 60kg até 79 kg

Foi utilizado a técnica de classificação por árvore de decisão com o algoritmo J48. Primeiramente fizemos uma análise classificatória em relação ao peso do atleta, tivemos essa seguinte árvore gerada. Tivemos 88,4% de instâncias classificadas corretamente em um treinamento de 70% e teste de 30%.



Nota-se que pessoas altas tendem a serem mais pesadas que pessoas baixas, e em pessoas de altura medianas, os homens são mais pesados que as mulheres. Com essa árvore conseguimos identificar o padrão de peso entre os atletas.

Após as análises com peso, fizemos uma classificação pela idade, que nos levou a seguinte árvore. Houve 75% de acerto nos resultados. Segundo a matriz de confusão, algumas pessoas adultas foram classificadas como jovem, e algumas pessoas de meia idade foi classificada como jovem. Pensamos que foi devido ao pouco número de casos de pessoas de meia idade (20 atletas).



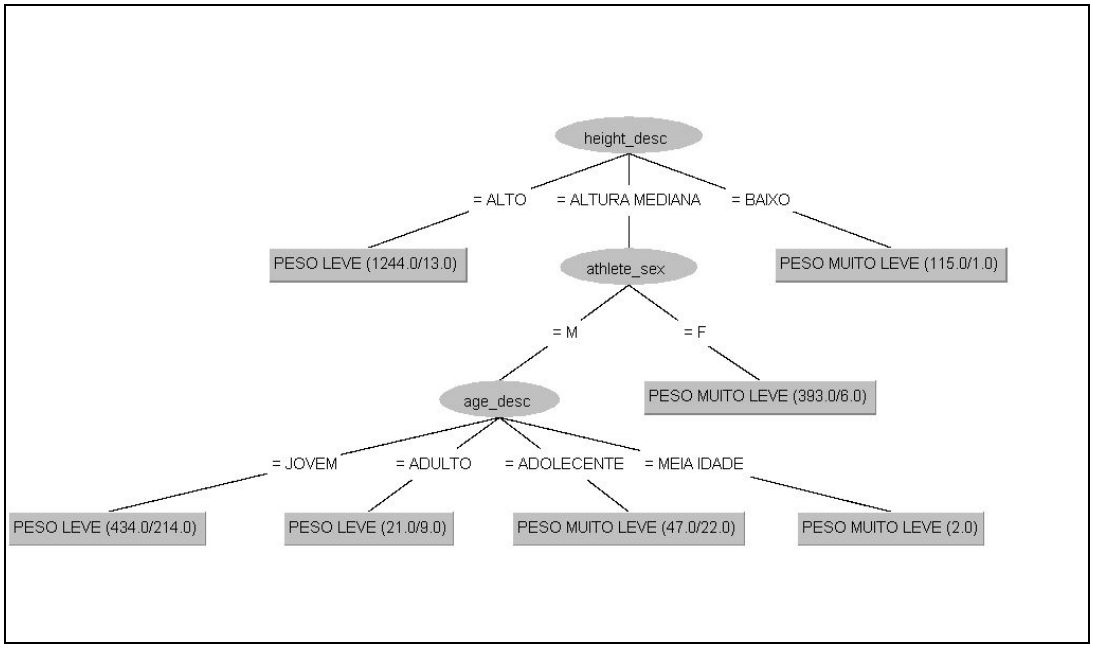
Segundo a árvore, todos os homens que participaram da ginástica são jovens. Em mulheres temos que quando são altas são jovens, e quando não são altas, são adolescentes.

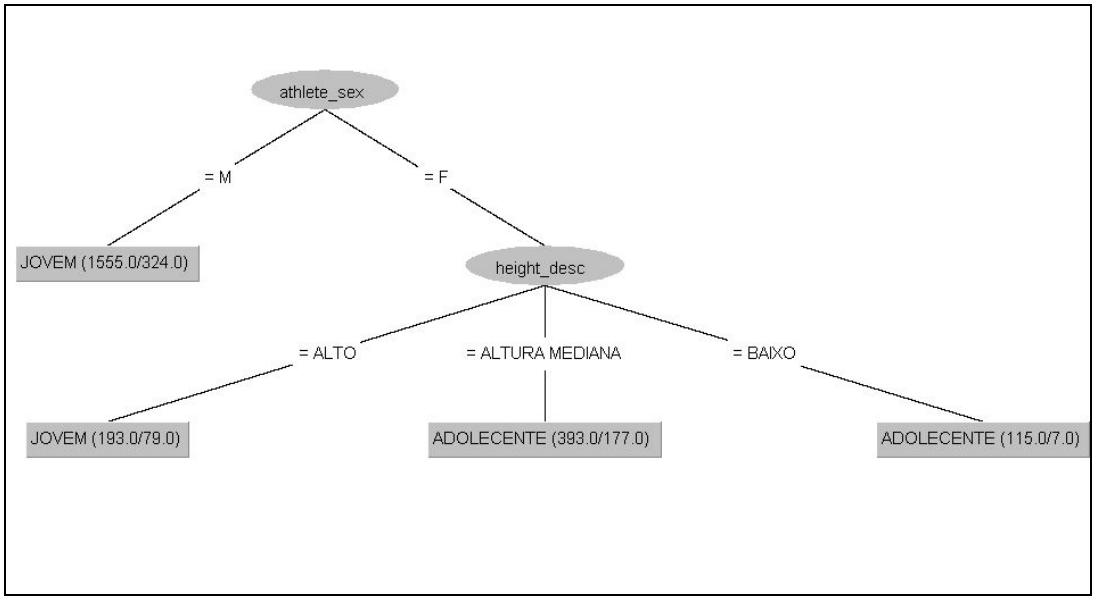
Depois de haver uma visão geral dos atletas em ginástica, foi feito um filtro para estudar apenas os atletas ginastas que ganharam qualquer tipo de medalhas. As colunas utilizadas nessa análise foi: “athlete_sex”, “age”, “heigth” e “weight”.

- A maioria é Jovem
- A maioria é Alto (170cm até 189cm)
- A maioria tem Peso Leve (60kg até 79kg)
- A maioria dos ganhadores foram da União Soviética e depois Estados Unidos da América

As árvores relacionadas a altura e ao peso continuaram no mesmo padrão que as árvores geradas com a população total. Indicando assim que não existe nenhuma característica relevante que leva os atletas a serem campeões, a não ser as características normais de um ginasta.

Segue abaixo em ordem, a primeira árvore é relacionada a altura com precisão de 88% e a segunda a idade com precisão de 75%:





Com as árvores geradas, não conseguimos obter uma precisão maior que 90%, porém acredito que adquirimos níveis satisfatórios para as análises. Com as árvores geradas, conseguimos obter os seguintes informações.

- As características dos ganhadores para as característica da população total não varia.
- Quando o atleta for homem, ele provavelmente será jovem e de peso leve ou muito leve.
- Quando o atleta for mulher, ela será adolescente ou jovem e com peso muito leve.
- Atletas adolescentes serão a maior parte mulheres e com peso muito leve.
- Atletas de ginástica são leves ou muito leves.

Quantos participantes o brasil terá nas olímpiadas de 2020?

Para responder essa questão tivemos que utilizar de técnicas de predição, e para fazer isso utilizamos as regressão linear. O primeiro passo para a solução desse problema foi gerado com a utilização de querys no banco de dados, uma lista de todos os atletas que já representaram o brasil nas olímpiadas.

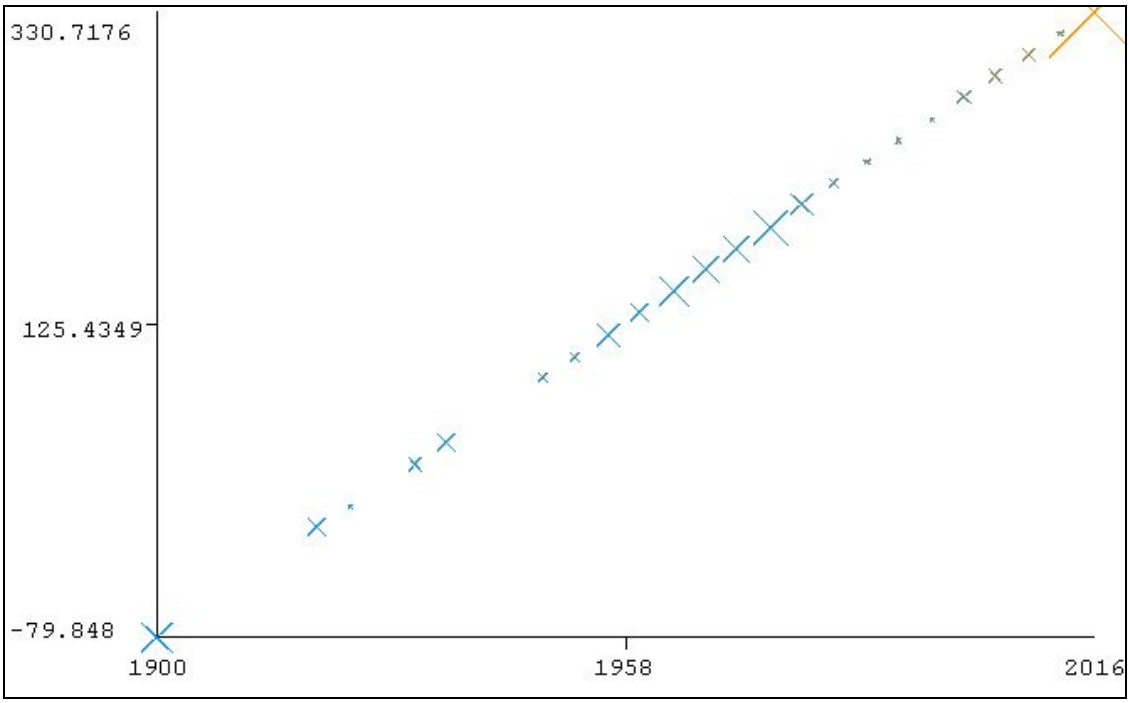
O segundo passo, foi fazer uma filtragem e uma contagem de quantos atletas participaram por cada ano, o resultado foi gerado abaixo:

```
1 select count(*) as num_athlete, event_year from datas where athlete_country = 'BRA' group by event_year
```

datas (2×23)

num_athlete	event_year
3	1.900
38	1.920
18	1.924
67	1.932
94	1.936
105	1.948
119	1.952
63	1.956
86	1.960
65	1.964
89	1.968
104	1.972
89	1.976
149	1.980
197	1.984
233	1.988
233	1.992
251	1.996
236	2.000
318	2.004
338	2.008
306	2.012
583	2.016

Com esses dados obtidos, inserimos os dados no Weka e aplicamos o algoritmo de regressão linear com o parâmetro de número de atletas. Abaixo temos a curva de erro, com o eixo x representando o ano da competição e o eixo y representando a quantidade de atletas:



Nota-se que houve um erro nos anos 1900 e 2016 em relação aos dados reais, isso deve-se muito provavelmente pelo fato que esses anos são os outliers, o máximo e o mínimos dos resultados. A gerada não foi a melhor possível, apesar de haver uma correção de 0.84 houve um erro elevado. Houve tentativa de acrescentar mais colunas no intuito de melhorar o erro, porém sem sucesso. Esses dados foram treinados com a própria base.

```
=== Summary ===

Correlation coefficient                  0.8413
Mean absolute error                     49.7052
Root mean squared error                 71.0388
Relative absolute error                 47.073  %
Root relative squared error             54.0599 %
Total Number of Instances               23
```

Mesmo com resultados que não são os melhores possíveis, decidimos manter esses estudos aqui no relatório, como relato de exploração da técnica de regressão linear. A regressão gerou a seguinte fórmula para calcularmos a quantidade de participantes da próxima olimpíada:

*num_athlete = 3.5394 * event_year -6804.6261*

Com a seguinte fórmula, podemos estimar então que o número de atletas brasileiros da próxima olimpíadas (2020) será de aproximadamente 345 atletas.

Sabendo que mais da metade dos brasileiros estão acima do peso. Nos esportes olímpicos, qual a relação de peso entre os atletas brasileiros? Este número aumentou com o decorrer das olimpíadas?

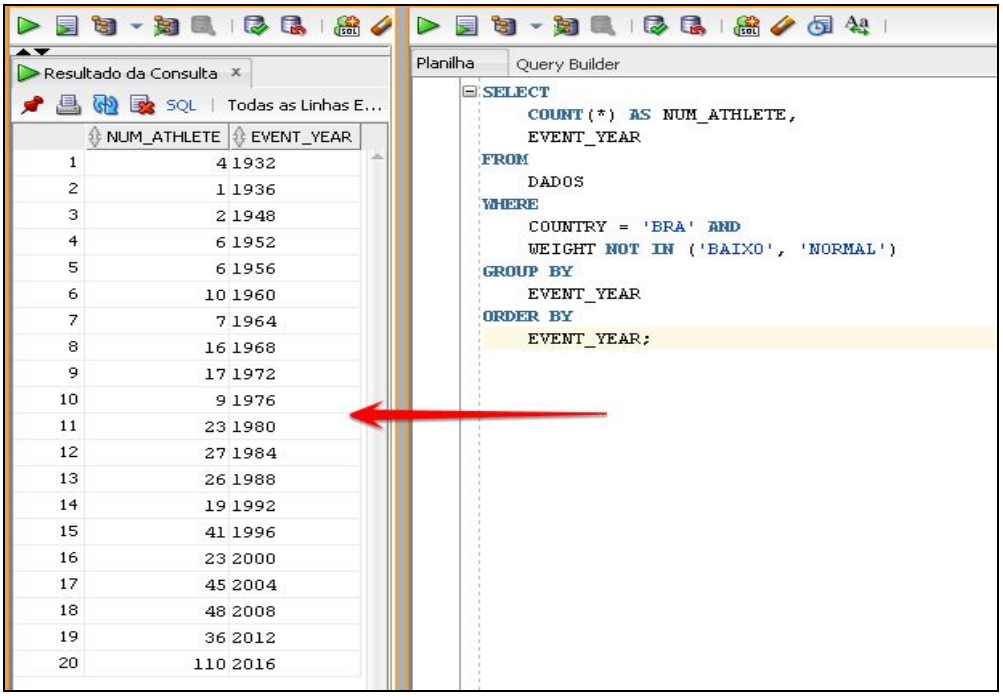
Esta pergunta foi resultado de uma pesquisa exploratória sobre as olimpíadas de 2016 que aconteceu no Brasil no Rio de Janeiro. Segundo o site Estadão, “Há mais de 300 atletas inscritos na Rio 2016 com IMC (índice de massa corporal) superior a 30, que segundo a Organização Mundial de Saúde, é sinal de obesidade”.

A pesquisa posteriormente descreve que índice de massa corporal não deve ser levado em consideração em atletas por não resultar em resultados precisos pelo fato de não travar as informações de massa muscular. Entretanto, resolvemos calcular o IMC independente de não ser 100% preciso para termos uma “base” dos perfis dos atletas de forma “homogênea” e com

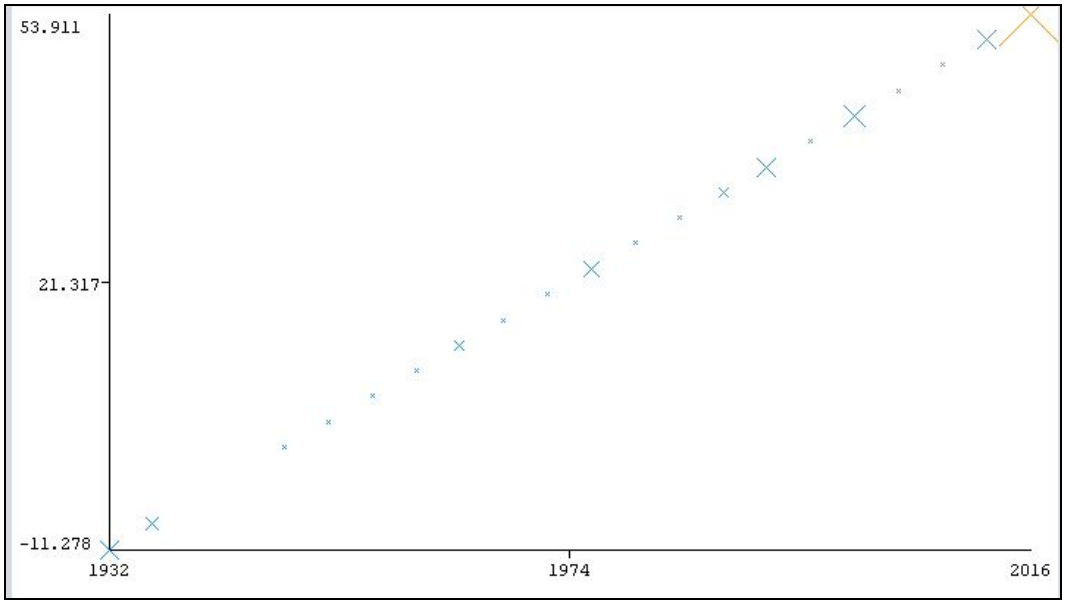
alguma base para o peso, não apenas relacionando com o "senso comum" de muito magro, magro, gordo etc. A tabela utilizada para o cálculo encontra-se a seguir:

Resultado	Situação
Abaixo de 17	Muito abaixo do *peso*
Entre 17 e 18,49	Abaixo do *peso*
Entre 18,5 e 24,99	*Peso* normal
Entre 25 e 29,99	Acima do *peso*
Entre 30 e 34,99	*Obesidade* I
Entre 35 e 39,99	*Obesidade* II (severa)
Acima de 40	*Obesidade* III (mórbida)

Após terem sido feitas todos os cálculos de IMC para os atletas e inserido os dados no banco de dados, utilizando o script SQL a seguir para exportar o conteúdo para ser utilizado no WEKA a fim de encontrar alguma previsão do índice de peso dos atletas brasileiros:



Aplicamos o algoritmo de regressão linear com o parâmetro de número de atletas. Abaixo temos a curva de erro, com o eixo x representando o ano da competição e o eixo y representando a quantidade de atletas:



A regressão gerou a seguinte fórmula para calcularmos a quantidade de atletas acima do peso de acordo com os anos:

*num_athlete = 0.7761 * event_year - 1510.6422*

Com a fórmula acima, pode-se concluir que o número de atletas com peso diferente de “BAIXO” e “NORMAL” aumentou gradativamente com o passar das competições.

Sabe-se que os Estados Unidos corresponde ao maior número de medalhas acumuladas na história das olimpíadas. Portanto, existe alguma relação entre os esportes onde foram participantes e premiados?

Vimos que entre os principais esportes que os Estados Unidos mais ganharam medalhas estavam o Atletismo, natação. O voleibol está na décima quinta posição no ranking de medalhas, mas escolhemos ele para ser um dado diferenciado, “fora do top ranking”. Para estes esportes fizemos um script SQL para “marcar” como “1” se o EUA teve medalha no esporte da coluna correspondente e “0” para vise versa.

Após a criação do script, os dados foram exportados para o Weka a fim de verificar uma possível relação com o algoritmo Apriori. A seguir encontram-se as regras para um grau de confiança de 80% e suporte mínimo de 10%:

1. Swimming=NÃO 13 ==> Volleyball=NÃO 12 <conf:(0.92)>

2. Swimming=NÃO Athletics=SIM 11 ==> Volleyball=NÃO 10 <conf:(0.91)>

3. Swimming=NÃO 13 ==> Athletics=SIM 11 <conf:(0.85)>

4. Volleyball=NÃO Swimming=NÃO 12 ==> Athletics=SIM 10 <conf:(0.83)>

5. Athletics=NÃO 6 ==> Volleyball=NÃO 5 <conf:(0.83)>

6. Athletics=SIM 21 ==> Volleyball=NÃO 17 <conf:(0.81)>

7. Volleyball=SIM 5 ==> Swimming=SIM 4 <conf:(0.8)>

8. Volleyball=SIM 5 ==> Athletics=SIM 4 <conf:(0.8)>

Ao todo oito regras foram encontradas para os três esportes selecionados, e a mais surpreendente para o grupo foi a primeira delas, que se os EUA não obteve medalhas em natação (sejam elas bronze, prata ou ouro), também não obteve medalhas no Voleibol.

Conclusão

Os estudos praticados e explorados nesse trabalho, mostra como podemos adquirir informações interessantes de bases de dados com o processo de Data Mining e KDD. Com o algoritmos, conseguimos tirar análises e conclusões para as perguntas estipuladas.

O processo mais lento e mas trabalhoso foi o pré processamento, onde tivemos que tratar todas as inconsistências e tivemos que adaptar os dados para determinados algoritmos. Todo o processo de mineração de dados se tornou viável, quando fizemos esses procedimentos, um exemplo claro foi a utilização de faixas etárias de idade, peso e altura, que generalizou os dados, porém deu mas facilidade para a análises dos dados nos algoritmos.

A utilização importação da base de dados para um banco tornou o processo de tratamento de dados e filtragem ainda mais fácil. Com a utilização de SQL, conseguimos elaborar filtros, tratar valores nulos e classificar os dados em faixas etárias. Com a utilização de SQL, notamos que com a própria utilização dele conseguimos gerar informações preliminares úteis.

A utilização e exploração dos algoritmos foi bem interessante, mesmo não tendo resultados extremamentes confiáveis, tivemos uma noção do funcionamento de cada algoritmos, e a tentativa de cada vez melhorar seus graus de confiança.

De modo geral, foi obtidos resultados interessantes e úteis com as técnicas de mineração de dados e o processo de KDD, validando a ideia que esses processos são realmente viáveis e úteis para a geração de resultados.

Referências Bibliográficas

- https://www.kaggle.com/heesoo37/120-years-of-olympic-history-athletes-and-results
- https://esportes.estadao.com.br/noticias/jogos-olimpicos,medidas-extremas-confira-os-biotipos-dos-atletas-olimpicos,10000069154
- http://www.tertuliaconscienciologia.org/index.php?option=com_content&task=view&id=2&Itemid=1
- https://www.youtube.com/watch?v=ZvxJoJmqtrM
- https://machinelearningmastery.com/use-regression-machine-learning-algorithms-weka/
- http://www.calculoimc.com.br/tabela-de-imc/