Utilização de técnicas de Data Mining e KDD

Samuel Favarin, Vinícius Machado e José Henrique

Base de dados

- “120 anos da história das olimpíadas: atletas e resultados”
- Dados biológicos e dados de medalhas de cada participação dos atletas das olimpíadas de 1896 em Atenas até 2016 no Rio.
- 271116 linhas e 15 colunas.

Base de dados

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

Eliminação dos dados

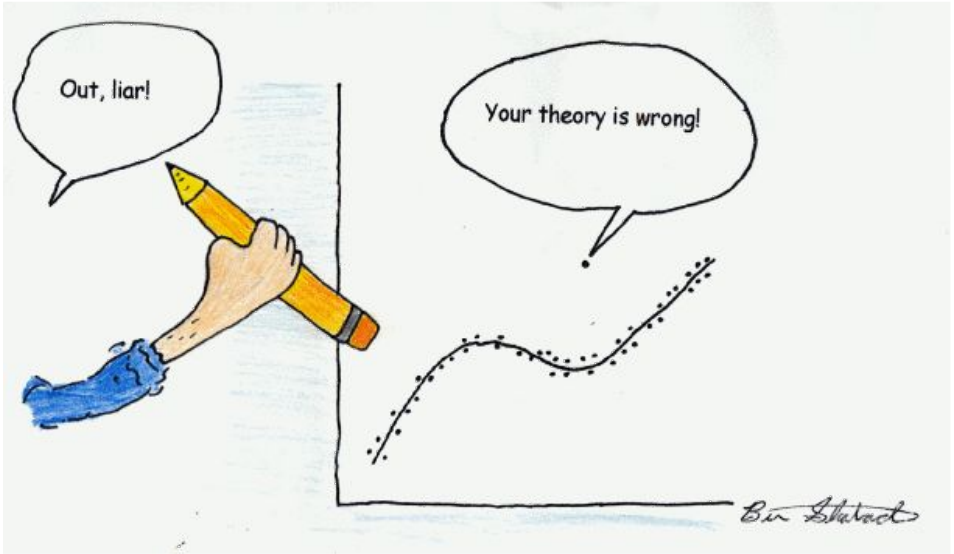
- Olimpíadas de inverno
- Nomes e Ids de atletas
- Coluna Games
- Coluna Event
- Coluna Season
- Coluna Team

Tratamento de Missing Values

- Coluna “Age”: A média obtida foi 25.64, e foi arredondado para cima 26.
- Coluna "Height". Valor médio foi 175.4 e foi arredondado para 175.
- Coluna “Weight”. Valor médio para 70.5.

Tratamento Outliers

- Na coluna “Weight” cerca de 6 casos estavam preenchidos com “733.33.33” ou “603.33.33”.
    - Solução: Remoção dessas linhas
- 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

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 peso:

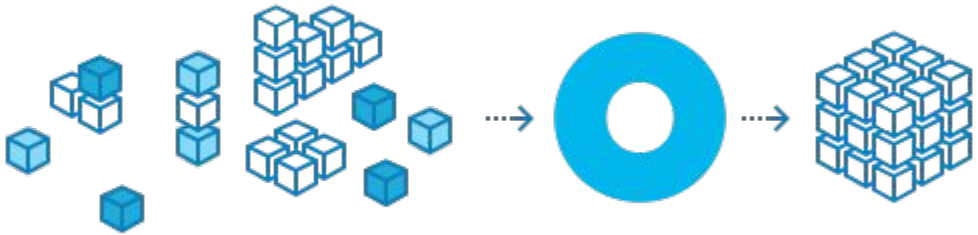
Faixa de Peso dos atletas	Label de saída
Menor de 60kg	PESO MUITO LEVE
Maior que 59kg e menor que 80kg	PESO LEVE
Maior que 79kg e menor que 100kg	PESO NORMAL
Maior que 99kg e menor que 120kg	PESADO
Maior que 119kg	MUITO PESADO

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

Substituição de dados

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



Perguntas e hipóteses

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 os Estados Unidos foram participantes e premiados?

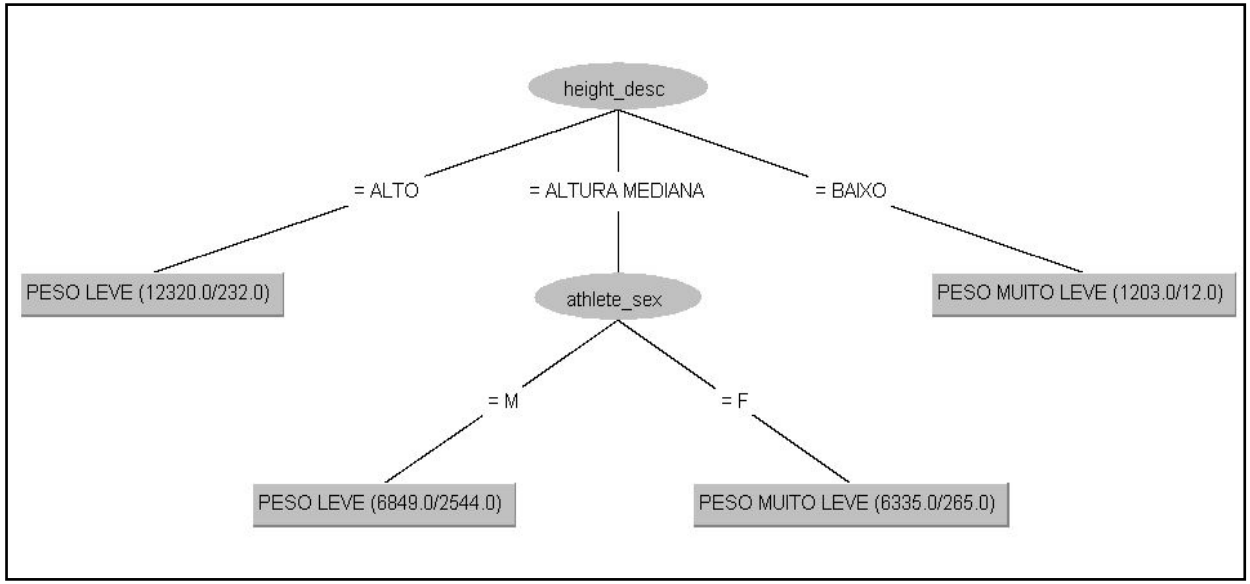
Quais as características básicas dos atletas de ginástica? Ginastas possuem tendência de serem leves?

- Filtragem dos dados para receber apenas atletas de ginástica
- Colunas “athlete_sex”, “age”, “height”, “weight” e “medal”.
- Problema de classificação
    - Algoritmo Árvore de decisão J48

1ª Análise

Análise classificatória em relação ao peso de atletas de ginástica:

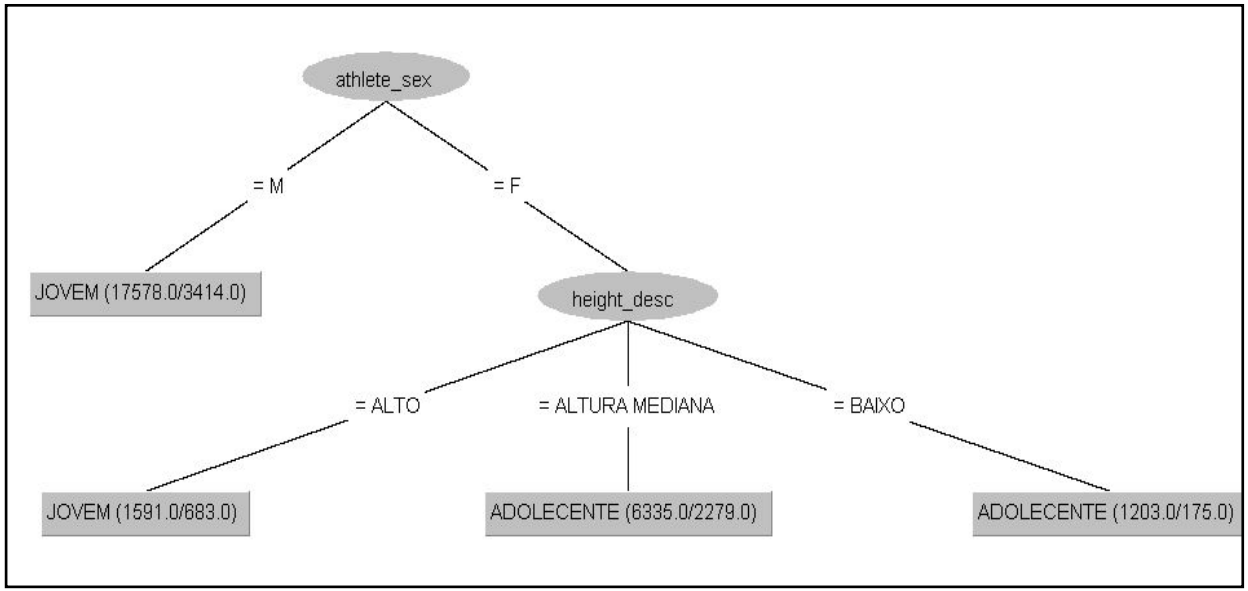
Instâncias analisadas corretamente: 88,4%



2ª Análise

Análise classificatória em relação a idade de atletas de ginástica:

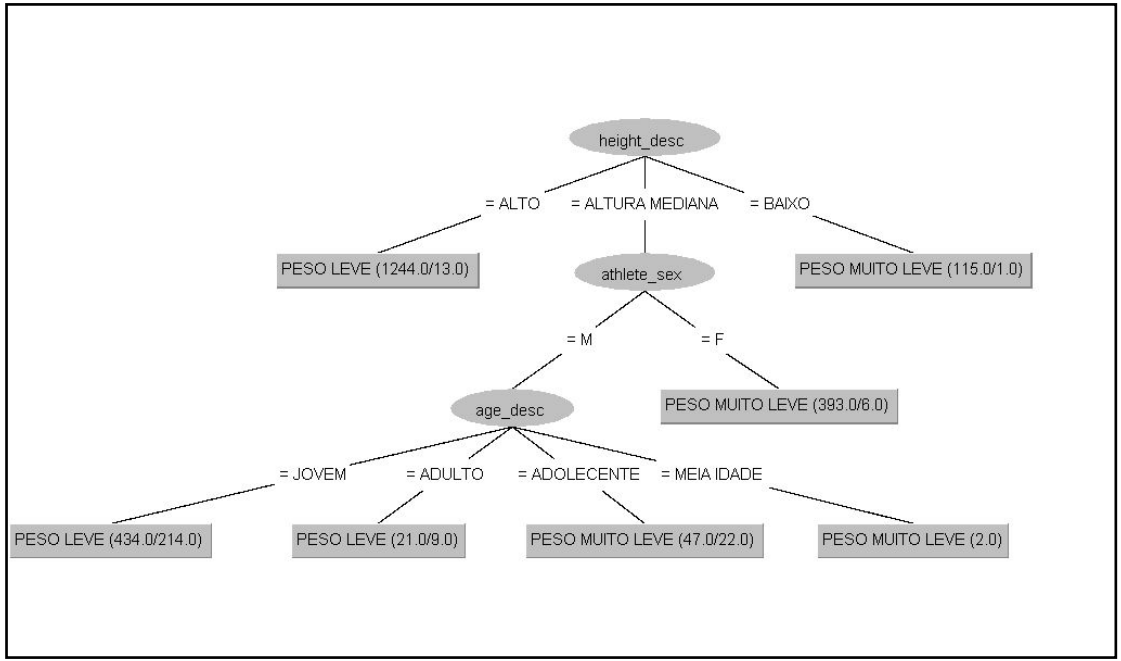
Instâncias analisadas corretamente: 75%

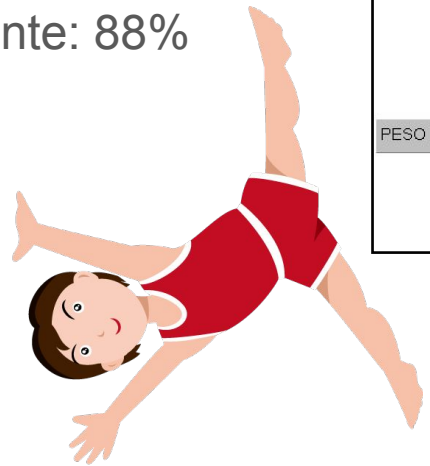


3ª Análise

Análise classificatória em relação ao peso de atletas de ginástica que ganharam alguma medalha:

Instâncias analisadas corretamente: 88%

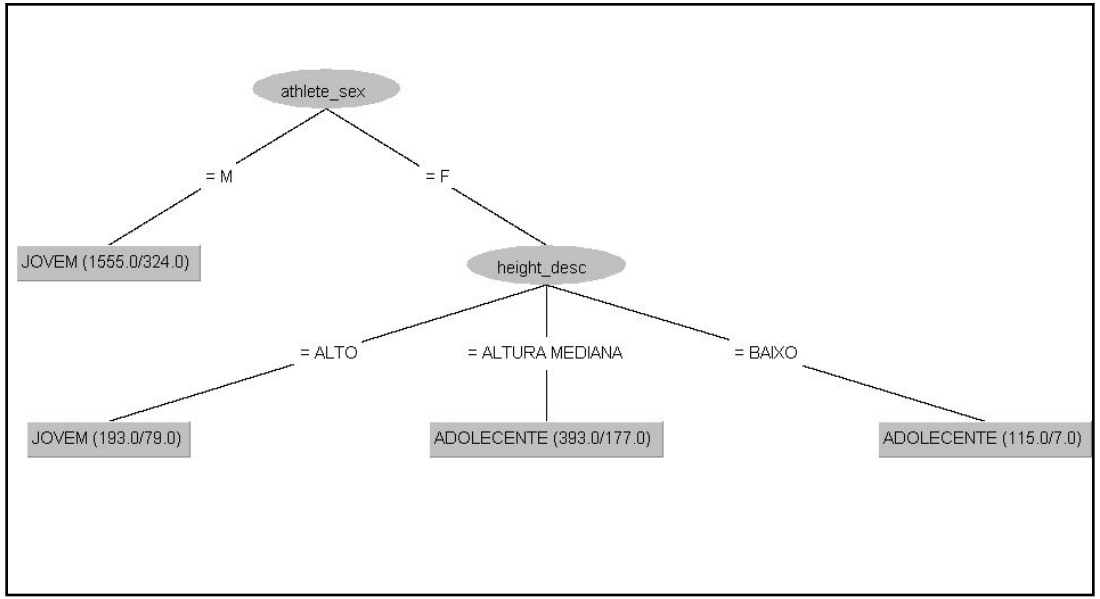




4ª Análise

Análise classificatória em relação a idade dos atletas de ginástica que ganharam alguma medalha:

Instâncias analisadas corretamente: 75%



Portanto...

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

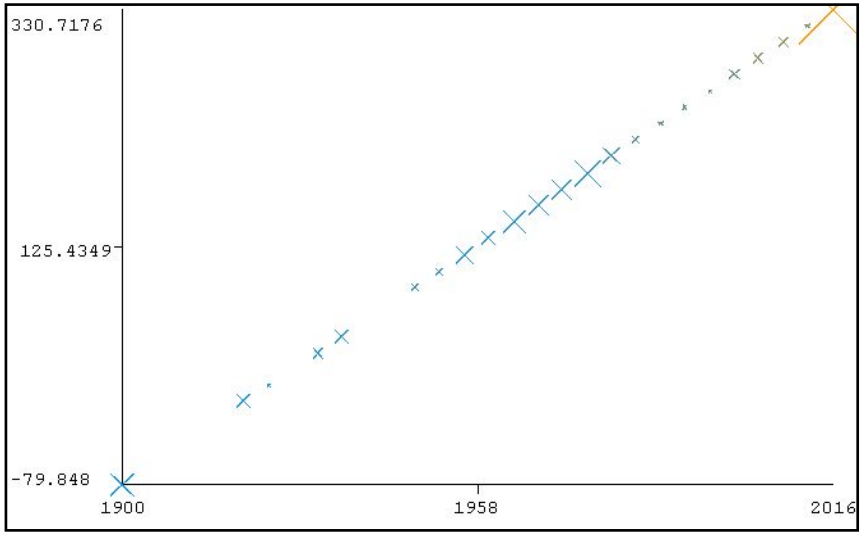
Quantos participantes o Brasil terá nas olimpíadas de 2020?

- Filtro por todos os atletas que já representaram o brasil nas olímpiadas.
- Regreção Linear

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

Quantos participantes o Brasil terá nas olimpíadas de 2020?

- Eixo x representando o ano da competição
- Eixo y representando a quantidade de atletas



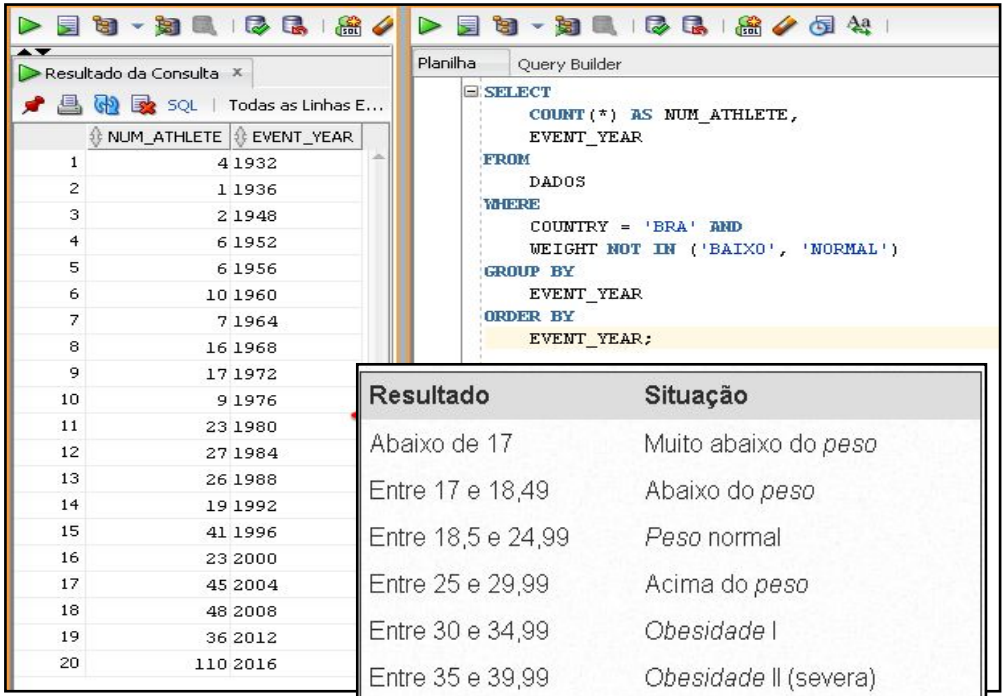

```
=== Summary ===

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
```

Num_athlete = 3.5394 * event_year -6804.6261

Num_athlete = 3.5394 * 2020 -6804.6261

Num_athlete ~= 345

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?

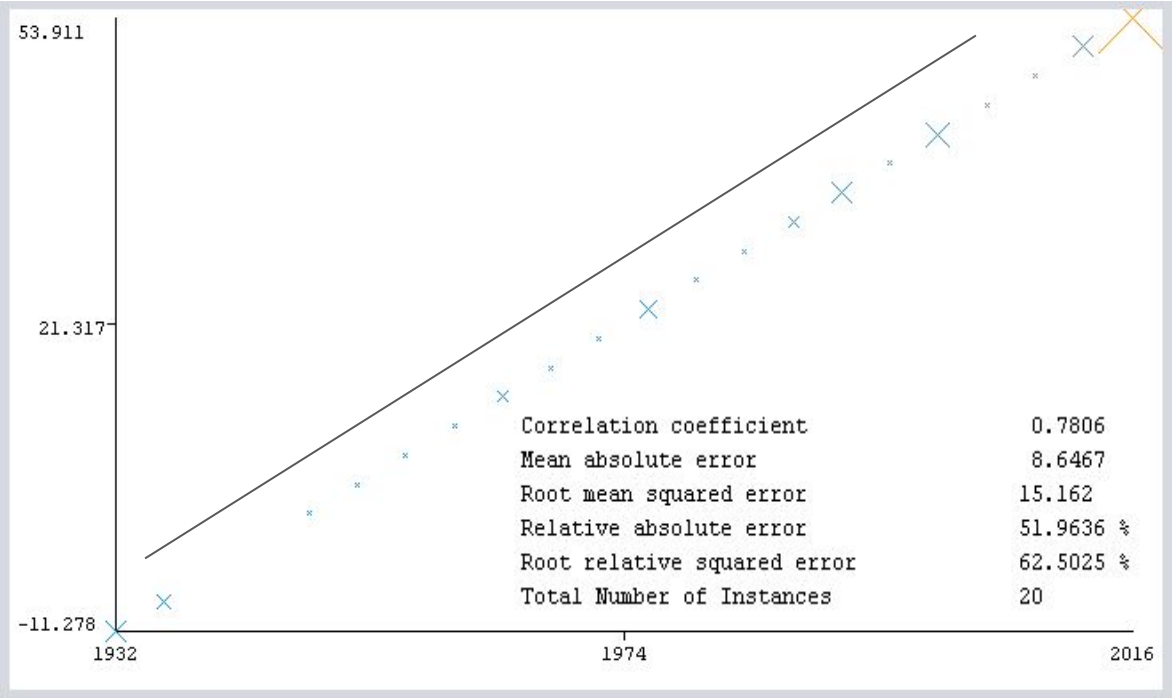


Resultado	Situação
Abaixo de 17	Muito abaixo do *peso*
Entre 17 e 18,49	Abaixo do *peso*
Entre 18,5 e 24,99	*Peso* normal
Entre 25 e 29,99	Acima do *peso*
Entre 30 e 34,99	*Obesidade* I
Entre 35 e 39,99	*Obesidade* II (severa)
Acima de 40	*Obesidade* III (mórbida)

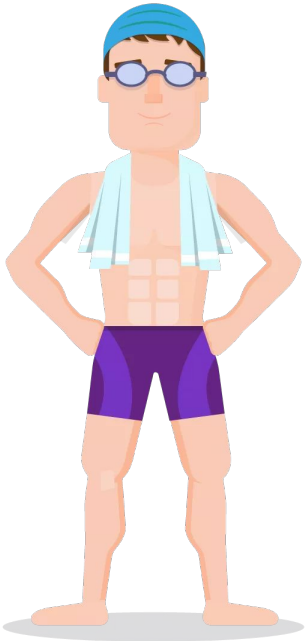
- Selecionado uma contagem de atletas onde o IMC é diferente de BAIXO ou NORMAL e nacionalidade brasileira por olimpíada
- Movimentação do arquivo para ARFF
- Aplicação de regressão linear

Num_Athlete = 0.7761 * EVENT_YEAR -1510.6422

- Durante as olimpíadas o número de atletas brasileiros acima do peso(segundo IMC) aumentou

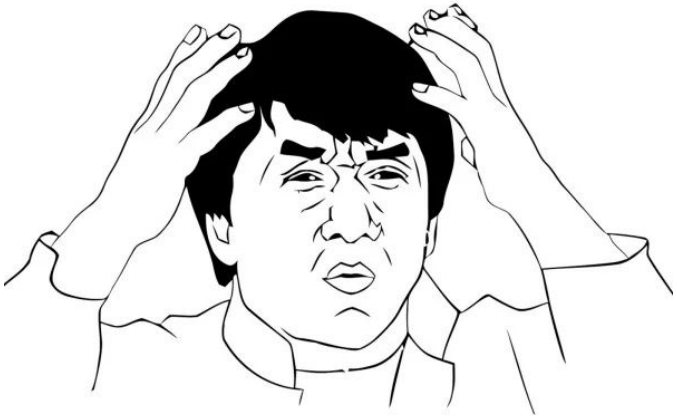


Sabe-se que os Estados Unidos corresponde ao maior número de medalhas acumuladas na história das olimpíadas. Portanto, existe alguma relação entre os esportes onde foram participantes e premiados?



- Atletismo e natação são os 2 principais esportes dos EUA em relação a medalhas.
- Verificar uma possível relação com o algoritmo Apriori entre os 2 principais esportes com 1 outro qualquer.
- A seguir encontram-se as regras para um grau de confiança de 80% e suporte mínimo de 10%:

Regras geradas



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=IM 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)>

Conclusão

- KDD e Data Mining gera informações úteis
- Pré-processamento dos dados é o processo mais lento
- WEKA é uma ferramenta fácil de manipulação
- SQL ajuda muito no pré processamento de dados
- Algumas respostas eram possíveis utilizando-se apenas de scripts SQL

Referência

- 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/