Previsão dos resultados da Copa do Mundo 2022

Aprendizagem de Máquina

CMI104

Mariana Marques Cabral - GRR20205954

Murilo Stellfeld de Oliveira Poloi - GRR20185705

Matemática Industrial

Professores:

Prof. Eduardo Vargas Ferreira

Prof. Lucas Garcia Pedroso

Universidade Federal do Paraná

6 de fevereiro de 2023

1 Introdução

Em 2022 tivemos a vigésima segunda edição da Copa do Mundo FIFA. A Copa do Mundo é sem dúvida um dos eventos mais notáveis no mundo esportivo e possui muita importância para a economia do país que sedia o torneio.

Apesar de ser um grande desafio para os times, é um desafio ainda maior para estatísticos e matemáticos pelo fato de que é díficil prever exatamente o que acontecerá em uma partida de futebol, ou de qualquer esporte. Até mesmo em partidas entre um time considerado muito fraco e um time considerado muito forte, ainda há a possibilidade da equipe favorita não sair vitoriosa, quem dirá prever placar, quem fará o(s) gol(s), entre outros.

Neste trabalho, vamos fazer uma abordagem básica, considerando apenas resultados e placares de todas as edições anteriores da Copa do Mundo, de 1930 à 2018 (as edições de 1942 e 1946 não ocorreram devido à segunda guerra mundial) e utilizaremos conceitos de matemática, estatística e especificamente aprendizagem de máquina para simular cada fase da edição de 2022 e verificar que o esporte é, na maior parte do tempo, imprevisível. Também iremos comentar sobre como tratamos a base de dados afim de fazer uma análise dos mesmos buscando obter alguma conclusão sobre qual ou quais times estão mais aptos a ganhar com base em número de gols, número de vitórias, entre outros.

Vale ressaltar que nem mesmo modelos mais complexos como modelos feitos por casas de apostas ou cientistas com bases de dados maiores conseguem prever os desfechos das partidas com tanta confiança, considerando que a previsão foi feita antes do ínicio do torneio.

2 Material e métodos

As bases de dados utilizadas foram retiradas de:

- `https://www.kaggle.com/datasets/abecklas/fifa-world-cup`
- `https://data.world/rezaghari/fifa-worldcup-2018/workspace/file?filename=2018_worldcup_v3.csv`
- `https://fixturedownload.com/results/fifa-world-cup-2022`

Em essência, foram utilizadas somente duas bases, a primeira, contendo os resultados históricos da competição Copa do Mundo desde 1930 até 2018, com exceção dos anos em que não ocorreu o torneio (1942 e 1946) devido a segunda guerra mundial.

A segunda base de dados contém o chaveamento geral da Copa do Mundo de 2022, isto é, todos as partidas que ocorreram durante fase de grupos e fases eliminatórias. Para a nossa aplicação foi considerado somente o chaveamento da fase de grupos, visto que vamos formar os confrontos das fases eliminatórias utlizando nosso modelo de predição.

Sobre tratamento de base de dados, a primeira continha colunas irrelevantes para a análise de dados e construção do modelo, como por exemplo, data e arbitragem de cada partida. Foram mantidas somente as colunas contendo ano, grupo/fase referente a partida, público presente, seleções jogando e resultado. O *dataset* não continha, por algum motivo, o público da partida entre Alemanha e Argélia pelas oitavas de finais da edição de 2014, e o número em questão foi obtido no artigo da *Wikipédia* referente a Copa do Mundo de 2014, seção das oitavas de finais. Essa base também possuía *strings* dos nomes dos times com algum erro (ver notebook no github). O tamanho da base em questão é de 917x14, antes da manipulação da mesma.

Para o caso da segunda base, foi somente necessário remover a coluna com os resultados (placar final da partida), data e estádio. Nosso modelo não diz a quantidade de gols de cada time em uma partida, somente as posições finais de cada time no grupo. O tamanho da base em questão é de 65x8, antes da manipulação da mesma.

O projeto foi construído inteiramente utilizando a linguagem de programação *Python* com auxílio do *Jupyter Notebook* e/ou *Google Colab*, com exceção da junção da base de dados de 1930 à 2014 com a de 2018, que foi feito via *software* de planilhas eletrônicas. A parte escrita do trabalho foi feita tem LaTeX via *Overleaf*. O notebook criado pode ser acessado em `https://github.com/murlopoloi/CMI104/tree/main/Entrega3`.

O método utilizado para obter os resultados foi construir uma função de probabilidade de Poisson com os parâmetros relacionados a 'força' de cada time em relação a média de gols feitos e a quantidade de gols sofridos por cada equipe. Cada jogo possuí parâmetros λ diferentes, e a métrica utilizada foi de obter um valor que relaciona o poderio defensivo e ofensivo de dois times para ambos os casos de mandante ou visitante. A partir da função de Poisson, simulamos a pontuação de cada grupo (sem restrições lógicas, por exemplo, há a possibilidade de um time ter 8 pontos dependendo dos parâmetros utilizados, mesmo sendo impossível fazer essa quantidade de pontos em três partidas. Após a primeira simulação, selecionamos os primeiros e segundos lugares de cada grupo e os distribuímos de acordo com o chaveamento utilizado na competição, repetindo o processo até o fim do torneio.

Algo importante de se dizer é que, devido ao fato de ser a primeira vez do país Catar na competição, o mesmo não possui nenhum dado histórico na competição, e portanto, partidas dele e/ou contra ele foram desconsideradas.

3 Resultados e discussão

3.1 Análise de Dados

Para a análise de dados das bases escolhidas, decidimos verificar as seguintes relações:

- Gols por edição;
- Gols por partida em cada edição;
- Partidas por edição;
- Gols% por time;
- Número de vitórias, derrotas e empates por time;
- Top 10 times que fizeram mais gols;
- Top 10 times que sofreram mais gols;
- Total de público por edição;
- Médias, mínimos e máximos gerais.

As análises feitas sobre os gráficos estarão no fim desta seção.

Gols por edição:

Gols por partida em cada edição:

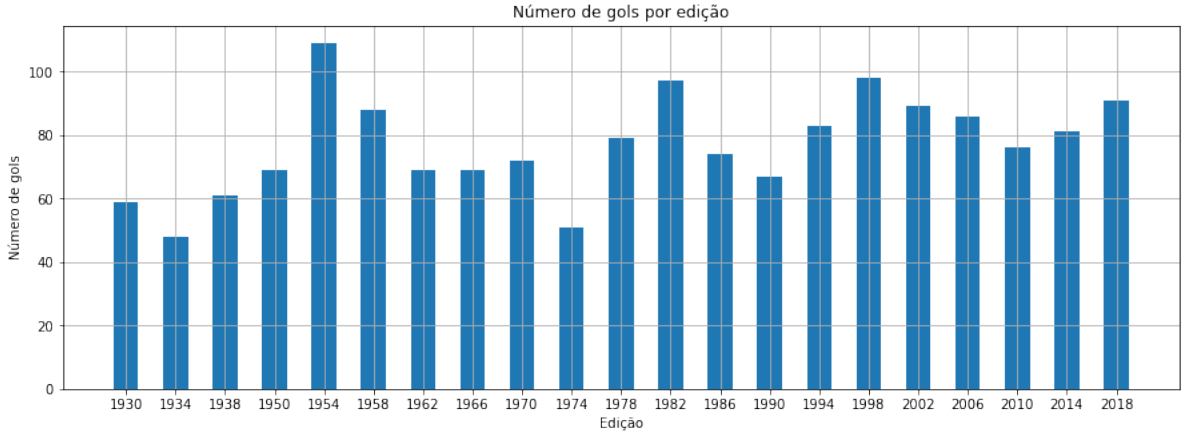

Figura 1: Número de gols por edição.

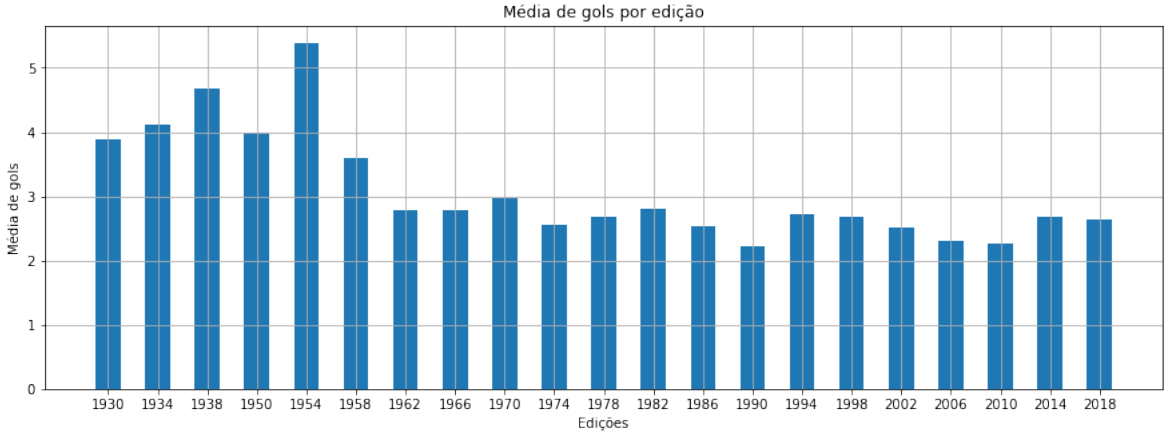

Figura 2: Média de gols por partida de cada edição.

Partidas por edição:

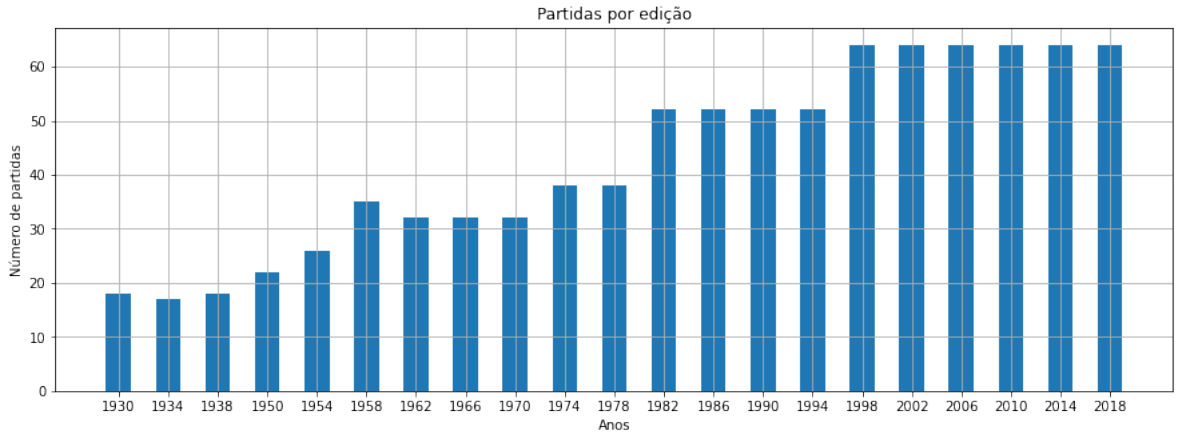

Figura 3: Número de partidas por edição.

Gols% por time:

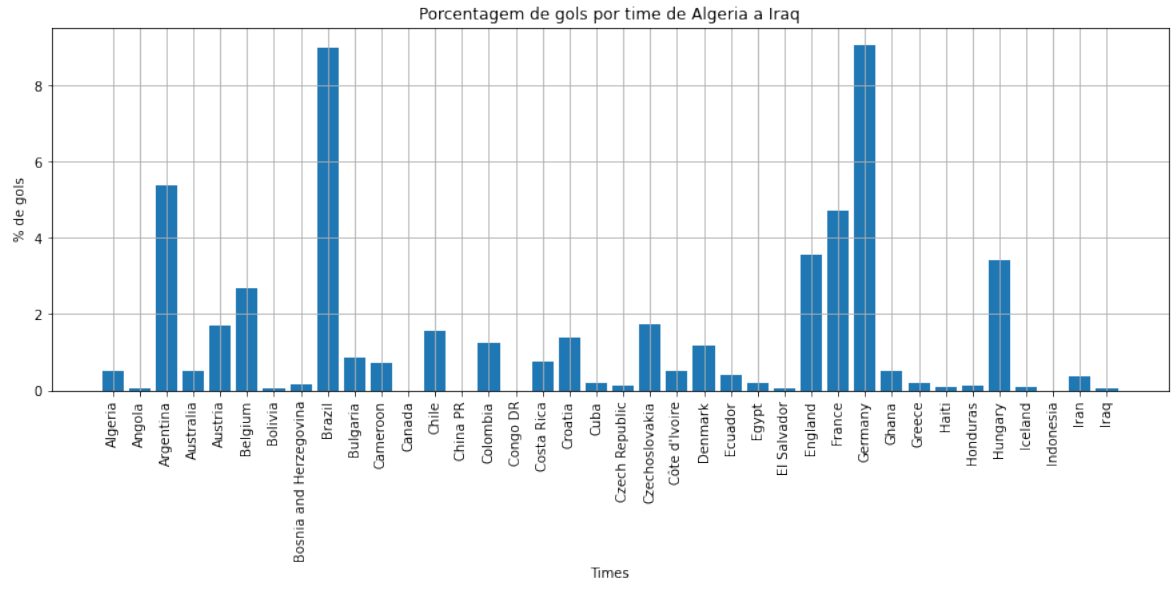

Figura 4: Porcentagem de gols por time de Algeria a Iraq.

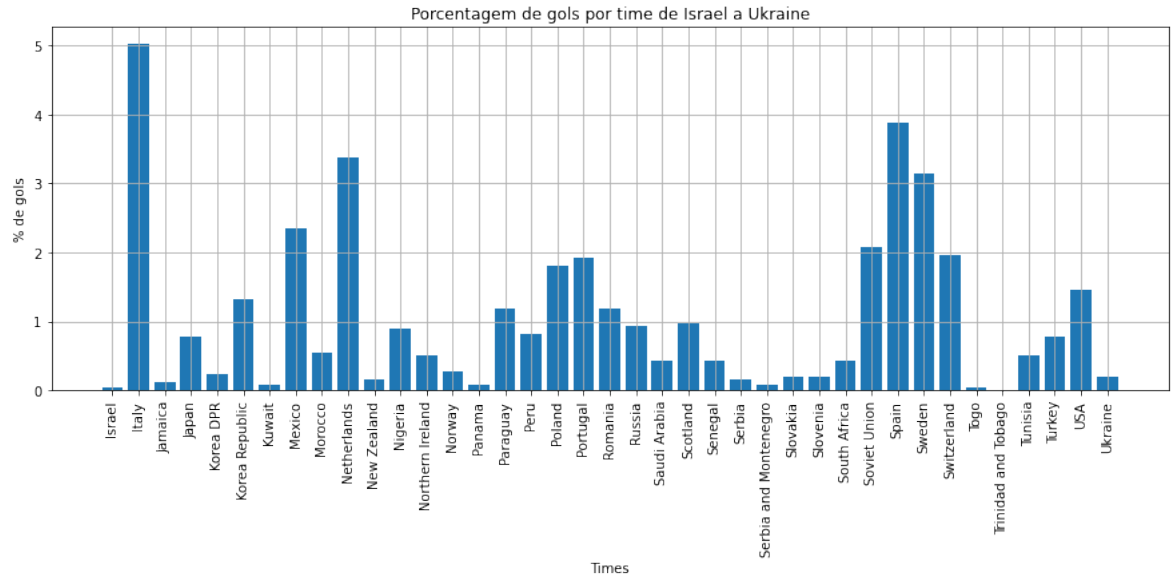

Figura 5: Porcentagem de gols por time de Israel a Ukraine.

Número de vitórias, derrotas e empates por time:

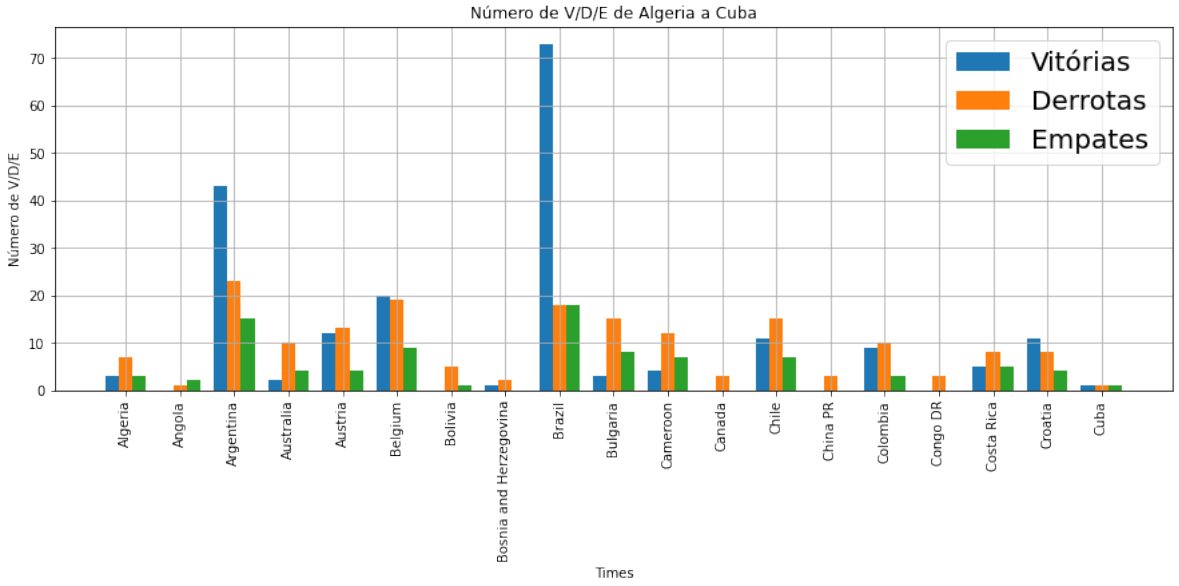

Figura 6: Número de vitórias/empates/derrotas por time de Algeria a Cuba.

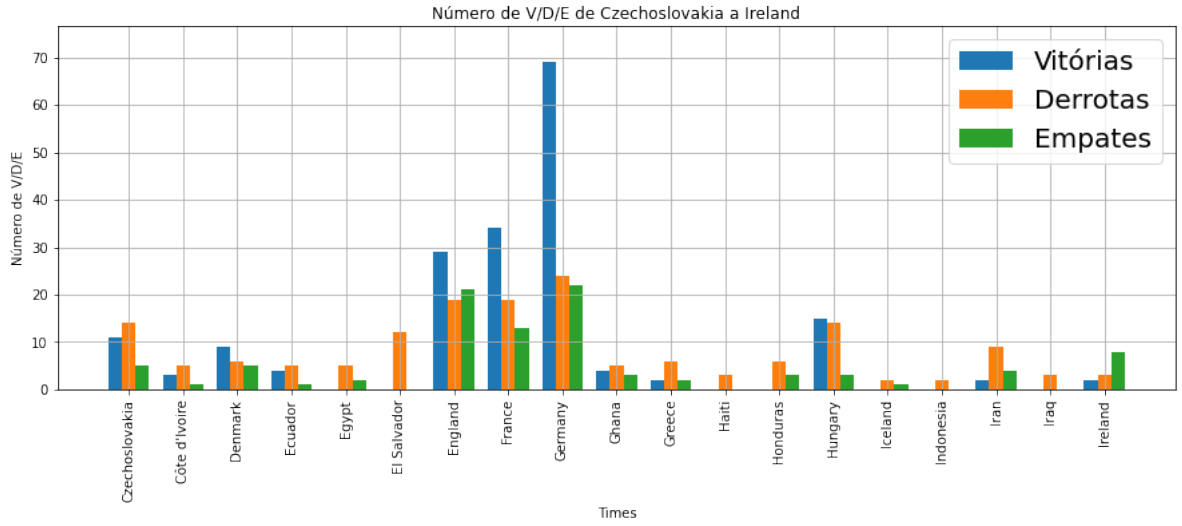

Figura 7: Número de vitórias/empates/derrotas por time de Czechoslovakia a Ireland.

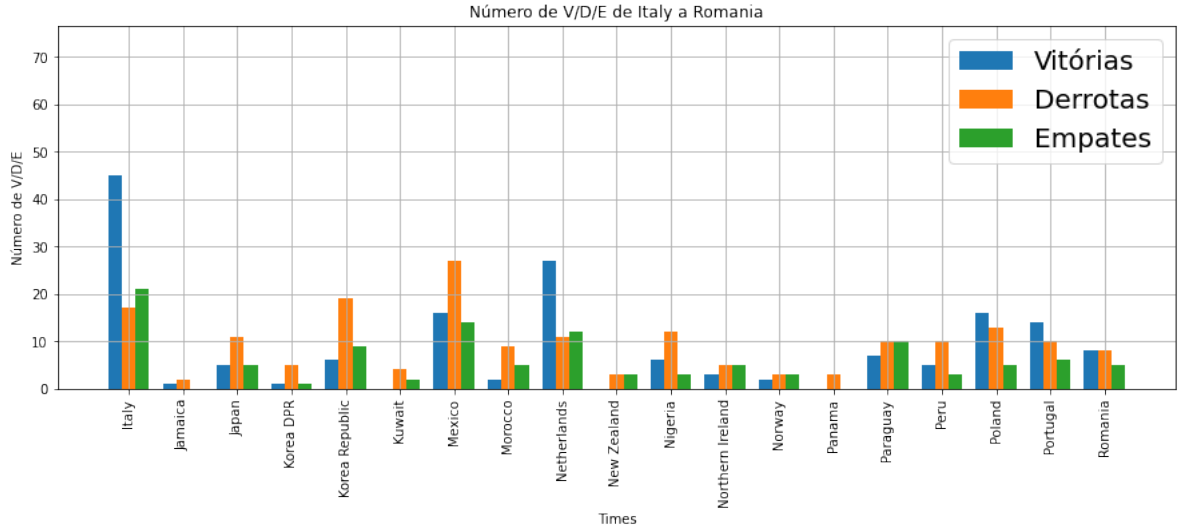

Figura 8: Número de vitórias/empates/derrotas por time de Italy a Romania.

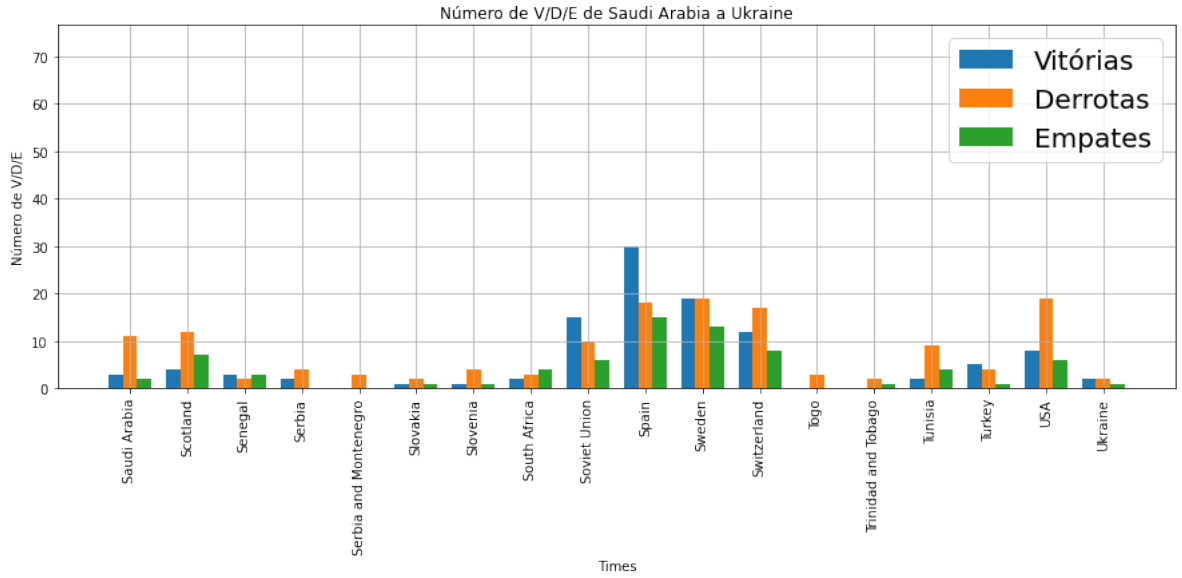

Figura 9: Número de vitórias/empates/derrotas por time de Saudi Arabia a Ukraine.

Top 10 times com mais gols:

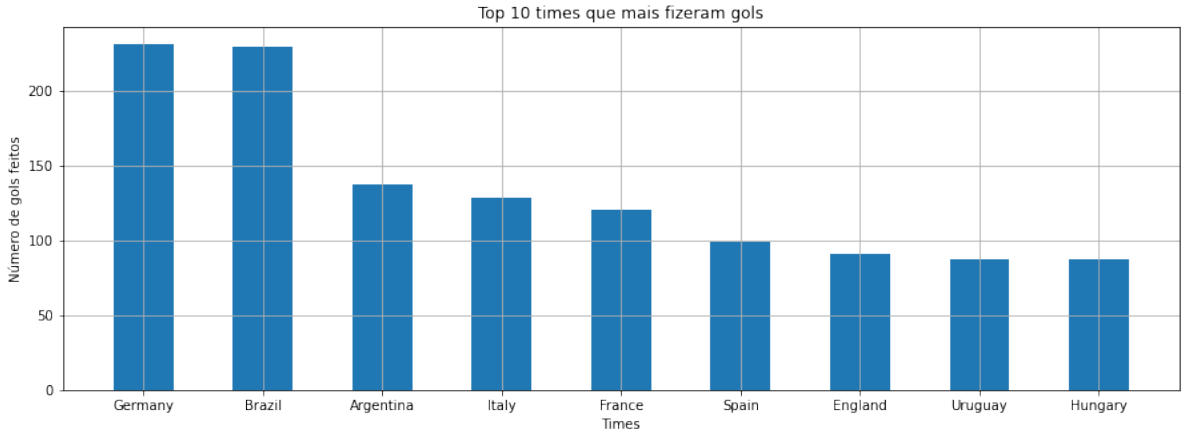

Figura 10: Os dez times que mais fizeram gols.

Top 10 times que mais sofreram mais gols

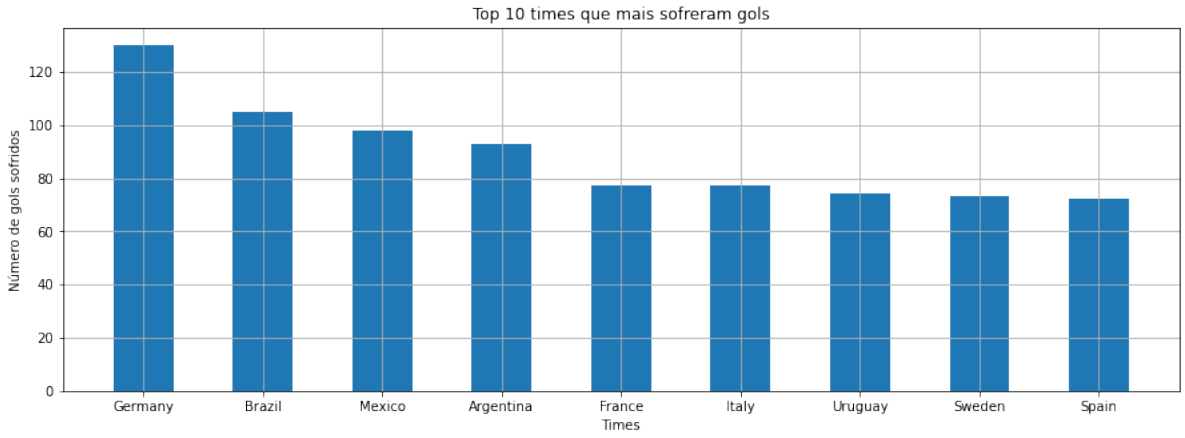

Figura 11: Os dez times que mais sofreram gols.

Total de público por edição:

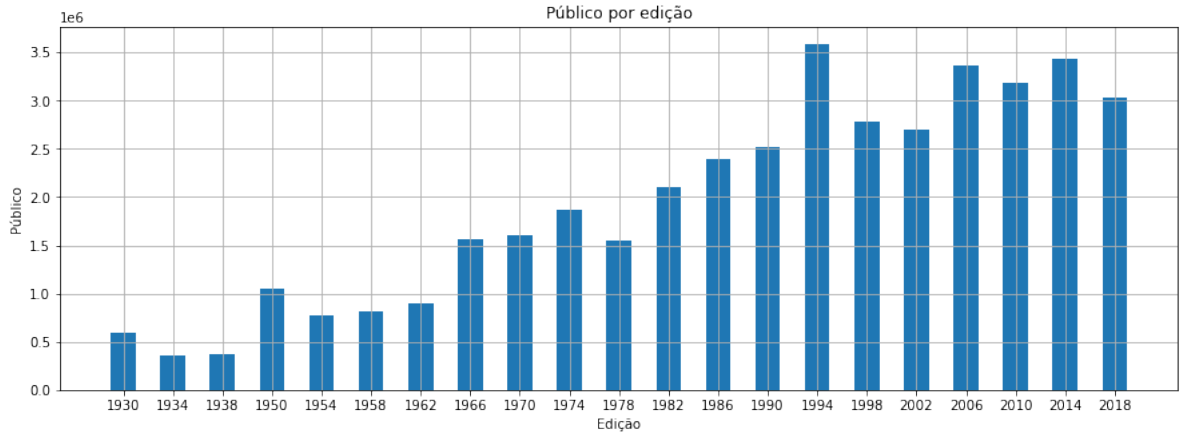

Figura 12: Total de público em cada edição.

Médias, mínimos e máximos gerais:

É visível que o edição de 1954 foi o ano com mais gols e maior média de gols, mas o ano em questão foi apenas o quinto menor em termo de número de partidas por edição. Com isso, é possível dizer que havia um certo desnível entre as equipes: algumas tinham facilidade de fazer gol, enquanto outras tinham facilidade de tomar gol. Nas útlimas edições, viemos tendo seleções que sequer era esperado que passassem da fase de grupos chegando em fases eliminatórias e avançando nelas.

	Year	Home Team Goals	Away Team Goals	Attendance
mean	1986.916	1.796	1.036	45036.087
min	1930.000	0.000	0.000	2000.000
max	2018.000	10.000	7.000	173850.000

Figura 13: Descrição dos dados usando média, minímos e máximos.

Em termos de quantidade ou média percentual de gols, o time da Alemanha ganha do Brasil por pouco. É notável que, por motivos políticos, a Alemanha foi dividida entre ocidental e oriental até o ano de 1990, mas só começou a participar como o país unificado que conhecemos hoje em 1994. Apesar de so ter tido uma participação em 1974, a alemanha oriental fez um total de 5 gols, o que foi suficiente para ficar a frente do Brasil.

O número de partidas tendeu a se normalizar, a partir de 1982 foi criado um sistema mais constante de jogos e o número de partidas somente aumentou a partir de 1998 devido ao aumento de equipes participantes.

Países como Brasil, Alemanha, França, Inglaterra, Itália, Espanha e Argentina tem, em geral mais de trinta vitórias cada, são países que num geral tem uma média de gols por time alta também, ou seja, é plausível dizer que partidas envolvem estes times há boas chances de haver(em) gol(s). É possível fazer a conexão também do número de vitórias baixo, ou de derrotas alto, com times que tiveram poucas participações e/ou gols.

Os países que mais fizeram gols costumam aparecer nos que mais sofreram também. Esse fato, provavelmente se dá também pelo número de participações de cada seleção.

A edição com mais espectadores nos estádios foi a de 1994, que ocorreu nos Estados Unidos. O país sede era o terceiro mais populoso na época.

Sobre as médias, minímos e máximos, é notável que no geral, os times que tem ”mando de campo” tendem a fazer mais gols que os times ”visitantes”. Podemos ver também o menor e maior público até 2018, assim como a média de público geral.

3.2 Pré-processamento e engenharia de características da base

Como mencionado na Entrega 1, a base de dados utilizada possuía somente dados até a edição de 2014 da Copa do Mundo, assim, foi necessário agregar os dados da edição de 2018. Como ambas bases não eram tão grandes, esse ajuste foi feito via planilha eletrônica.

Para tratar e manipular os dados utilizados neste projeto foi necessário remover

colunas com informações julgadas desnecessárias do *dataset*, como por exemplo, data e horário, estádio, cidade, árbitros, entre outros.

Foi necessário arrumar os nomes de alguns dos times, pois estes continham erros de codificação (como foi o caso da Costa do Marfim), ou ainda, fundir ou alterar, por decisão própia, o nome de alguns países, por exemplo, ”Zaire” passou a ser chamado de ”Congo DR”, ”Dutch East Indies” de ”Indonesia” e ”Germany FR” e ”German DR” passaram a se chamar ”Germany”.

A base de dados utilizada e os tratamentos feitos nela, assim como a parte de análise de dados podem ser vistas nos arquivos .csv e .ipynb em `https://github.com/murlopoloi/CMI104/tree/main/Entrega2`.

3.3 Resultados

3.3.1 Previsão do torneio completo

	Group A	Pts A
1st	Netherlands	4
2nd	Senegal	2
3rd	Ecuador	2
4th	Qatar	0

	Group B	Pts B
1st	England	6
2nd	Wales	5
3rd	USA	3
4th	Iran	2

	Group C	Pts C
1st	Argentina	7
2nd	Poland	5
3rd	Mexico	3
4th	Saudi Arabia	1

	Group D	Pts D
1st	France	7
2nd	Denmark	5
3rd	Tunisia	3
4th	Australia	2

	Group E	Pts E
1st	Germany	6
2nd	Spain	5
3rd	Japan	3
4th	Costa Rica	3

	Group F	Pts F
1st	Croatia	7
2nd	Belgium	6
3rd	Morocco	4
4th	Canada	0

	Group G	Pts G
1st	Brazil	7
2nd	Switzerland	4
3rd	Serbia	3
4th	Cameroon	2

	Group H	Pts H
1st	Portugal	6
2nd	Uruguay	5
3rd	Ghana	4
4th	Korea Republic	2

Figura 14: Fase de grupo gerada pelo modelo.

Round Number	Group	Home Team	Winner	Away Team
Round of 16	EF1	Netherlands	**Netherlands**	Wales
Round of 16	EF2	Argentina	**Argentina**	Denmark
Round of 16	EF3	Germany	**Germany**	Belgium
Round of 16	EF4	Brazil	**Brazil**	Uruguay
Round of 16	EF5	England	**England**	Senegal
Round of 16	EF6	France	**France**	Poland
Round of 16	EF7	Croatia	**Spain**	Spain
Round of 16	EF8	Portugal	**Portugal**	Switzerland

Figura 15: Oitavas de final gerada pelo modelo.

Round Number	Group	Home Team	Winner	Away Team
Quarter Finals	QF1	Netherlands	**Netherlands**	Argentina
Quarter Finals	QF2	Germany	**Brazil**	Brazil
Quarter Finals	QF3	England	**France**	France
Quarter Finals	QF4	Spain	**Portugal**	Portugal

Figura 16: Quartas de final gerada pelo modelo.

Round Number	Group	Home Team	Winner	Away Team
Semi Finals	SF1	Netherlands	**Brazil**	Brazil
Semi Finals	SF2	France	**France**	Portugal

Round Number	Group	Home Team	Winner	Away Team
Third Place	3rd	France	**Germany**	Germany
Finals	Finals	Netherlands	**Brazil**	Brazil

Figura 17: Semifinais e finais geradas pelo modelo.

3.3.2 Previsão a partir das oitavas de final

De acordo com o modelo construído, os resultados de todas as fases eliminatórias, seguindo os resultados reais das fases de grupo, se dariam da seguinte forma:

Round Number	Group	Home Team	Winner	Away Team
Round of 16	EF1	Netherlands	**Netherlands**	USA
Round of 16	EF2	Japan	**Croatia**	Croatia
Round of 16	EF3	England	**England**	Senegal
Round of 16	EF4	Morocco	**Morocco**	Spain
Round of 16	EF5	Argentina	**Argentina**	Australia
Round of 16	EF6	Brazil	**Brazil**	Korea Republic
Round of 16	EF7	France	**France**	Poland
Round of 16	EF8	Portugal	**Portugal**	Switzerland

Figura 18: Oitavas de final a partir de resultados reais.

Round Number	Group	Home Team	Winner	Away Team
Quarter Finals	QF1	Netherlands	**Netherlands**	Argentina
Quarter Finals	QF2	Croatia	**Brazil**	Brazil
Quarter Finals	QF3	England	**France**	France
Quarter Finals	QF4	Morocco	**Portugal**	Portugal

Figura 19: Quartas de final a partir de resultados reais.

Round Number	Group	Home Team	Winner	Away Team
Semi Finals	SF1	Netherlands	**Brazil**	Brazil
Semi Finals	SF2	France	**France**	Portugal

Round Number	Group	Home Team	Winner	Away Team
Third Place	3rd	Netherlands	**Netherlands**	Portugal
Finals	Finals	Brazil	**Brazil**	France

Figura 20: Semifinais e finais a partir de resultados reais.

4 Conclusão

Comparando os resultados obtidos pelo modelo considenrando os confrontos predefinidos, obtemos a seguinte tabela:

Fase	**Grupos**	**Oitavas**	**Quartas**	**Semifinais**	**Finais**
Número de acertos	18	8	1	2	1
Número de erros	14	0	3	0	1

Num contexto geral, a partir do ínicio do torneio, o modelo previu corretamente

Fase	**Grupos**	**Oitavas**	**Quartas**	**Semifinais**	**Finais**
Número de acertos	18	3	1	0	0
Número de erros	14	5	3	2	2

É nítido que para prever o chaveamento completo da competição, o modelo não funciona muito bem, com uma acurácia de aproximadamente 45%, enquanto que no quesito de prever jogos a cada fase, ele teve uma acurácia de 62.5%.

Há várias outras maneiras de estimar os parâmetros de λ utilizados pela função de *Poisson*. Infelizmente não foi possível testar para algum destas outras maneiras afim de comparar resultados, porém fica a possibilidade para um estudo futuro.