Instituto INFINET

Escola Superior da Tecnologia da Informação

Pós-Graduação MIT em BIG DATA

Cesar Mansur

Leonardo Arariba

Romário Gomes

Estimativa da Probabilidade de Vitória em Lutas no UFC a partir de um modelo preditivo utilizando ....

RIO DE JANEIRO

2021

Cesar Mansur

Leonardo Arariba

Romário Gomes

Estimativa da Probabilidade de Vitória em Lutas no UFC a partir de um modelo preditivo utilizando ....

Trabalho de Conclusão de Bloco apresentado ao programa de Pós-graduação MIT em BIG DATA do Instituto INFINET, como requisito final para avaliação da disciplina XXXXXX

Orientador: Kleyton Cotta

RIO DE JANEIRO

2021

RESUMO

O presente trabalho propõe analisar lutas do UFC onde permite observar diversos lutadores competindo mais de uma vez. Dessa forma, os estudos envolvendo a estimação da probabilidade de vitória e do comportamento das variáveis de interesse necessita que se utilize alguma metodologia que considere a correlação existente entre as unidades amostrais. Em acréscimo, cada luta é formada por dois lutadores, de modo que também não exista independência entre ambos. Uma das formas de se considerar a existência de dependência entre as unidades amostrais é com as Equações de Estimação Generalizadas (GEE). Essa metodologia, proposta por Zeger e Liang (1986) se diferencia dos Modelos Lineares Generalizados por incorporar uma matriz de correlação de trabalho que modela a dependência entre as unidades amostrais. Este trabalho busca, também, comparar os resultados produzidos por diferentes estruturas dematrizes de correlação ao considerar a estimação da probabilidade de vitória das lutas, no contexto explicativo e preditivo. Foram calculados escores para representar diversas situações de combate. Com a metodologia GEE, verificou-se que nas 1111 lutas analisadaso escore correspondente à efetividade de ataque de golpes se mostrou mais relacionado com a vitória, sendo uma prática a ser considerada nos treinamentos. Neste caso, a consideração da estrutura de dependência que existe quando os lutadores são avaliados mais de uma vez se mostrou mais eficiente do que considerar a luta como um agrupamento, mesmo que a soma da probabilidade de vitória não seja iguala 1. Neste caso, as matrizes apresentaram comportamentos similares. Ao considerar a estimação de um modelo preditivo, a variável ODDS, um escore dabanca de apostas, apresentou um maior percentual de acerto do que os modelos só com os escores. Os modelos com os escoresea ODDS apresentaram percentual de acerto maior que 60%. Ao se calcular um ganho fictício em relação às apostas, os modelos com os escores apresentaram ganhos, enquanto que os modelos com a ODDS resultaram em prejuízo, mostrando que a modelagem com os escores foi mais eficiente para este fim.

**Palavras chaves:** Python

ABSTRACT

O presente trabalho propõe analisar lutas do UFC onde permite observar diversos lutadores competindo mais de uma vez. Dessa forma, os estudos envolvendo a estimação da probabilidade de vitória e do comportamento das variáveis de interesse necessita que se utilize alguma metodologia que considere a correlação existente entre as unidades amostrais. Em acréscimo, cada luta é formada por dois lutadores, de modo que também não exista independência entre ambos. Uma das formas de se considerar a existência de dependência entre as unidades amostrais é com as Equações de Estimação Generalizadas (GEE). Essa metodologia, proposta por Zeger e Liang (1986) se diferencia dos Modelos Lineares Generalizados por incorporar uma matriz de correlação de trabalho que modela a dependência entre as unidades amostrais. Este trabalho busca, também, comparar os resultados produzidos por diferentes estruturas de matrizes de correlação ao considerar a estimação da probabilidade de vitória das lutas, no contexto preditivo. Foram calculados escores para representar diversas situações de combate. Com a metodologia GEE, verificou-se que nas 1111 lutas analisadaso escore correspondente à efetividade de ataque de golpes se mostrou mais relacionado com a vitória, sendo uma prática a ser considerada nos treinamentos. Neste caso, a consideração da estrutura de dependência que existe quando os lutadores são avaliados mais de uma vez se mostrou mais eficiente do que considerar a luta como um agrupamento, mesmo que a soma da probabilidade de vitória não seja iguala 1. Neste caso, as matrizes apresentaram comportamentos similares. Ao considerar a estimação de um modelo preditivo, a variável ODDS, um escore dabanca de apostas, apresentou um maior percentual de acerto do que os modelos só com os escores. Os modelos com os escoresea ODDS apresentaram percentual de acerto maior que 60%. Ao se calcular um ganho fictício em relação às apostas, os modelos com os escores apresentaram ganhos, enquanto que os modelos com a ODDS resultaram em prejuízo, mostrando que a modelagem com os escores foi mais eficiente para este fim

**Keyswords:** Python

**LISTA DE ILUSTRAÇÕES**

[Figura 1: gráfico favoritismo 15](#_Toc67599555)

[Figura 2: Gráfico de correlação 16](#_Toc67599556)

[Figura 3: Mapa de calor vermelho 17](#_Toc67599557)

[Figura 4: Mapa de calor azul 18](#_Toc67599558)

[Figura 5: avaliar nome 19](#_Toc67599559)

[Figura 6: Avaliar nome 2 20](#_Toc67599560)

[Figura 7: Modelo predição envergadura 21](#_Toc67599561)

[Figura 8: Modelo 1 21](#_Toc67599562)

[Figura 9: Avaliar nome 4 23](#_Toc67599563)

[Figura 10: Avaliar nome 25](#_Toc67599564)

[Figura 11: avaliar nome 25](#_Toc67599565)

[Figura 12: Avaliar nome 5 26](#_Toc67599566)

[Figura 13: avaliar nome 26](#_Toc67599567)

[Figura 14: Avaliar nome 26](#_Toc67599568)

[Figura 15: Avaliar nome 6 27](#_Toc67599569)

[Figura 16: avaliar nome 27](#_Toc67599570)

**SUMÁRIO**

[**1** **INTRODUÇÃO** 7](#_Toc67604750)

[1.1 Apresentação 7](#_Toc67604751)

[1.2 Justificativa 7](#_Toc67604752)

[1.3 Problema 7](#_Toc67604753)

[1.4 Objetivo 8](#_Toc67604754)

[**2** **REFERENCIAL TEÓRICO** 9](#_Toc67604755)

[2.1 Ultimate Fight Championship: A Competição 9](#_Toc67604756)

[**3** **METODOLOGIA** 11](#_Toc67604757)

[3.1 Instrumento de coleta de dados 11](#_Toc67604758)

[3.2 Análise de dados 13](#_Toc67604759)

[**4** **PREVISÃO DE RESULTADOS DE LUTAS** 23](#_Toc67604760)

[**5** **CONCLUSÃO** 26](#_Toc67604761)

[**6** **REFERÊNCIAS** 27](#_Toc67604762)

# **INTRODUÇÃO**

## Apresentação

O *Ultimate Fighting Championship* (UFC) foi de uma entidade pouco conhecida fora do círculo esportivo das lutas para ser a liga dominante no MMA em meados da década de 2000, consolidando uma marca que transpassou o mundo esportivo e passou a estar presente como um ícone cultural dos tempos atuais.

Segundo a Revista Rolling Stone, o UFC chegou a ser considerado o segundo esporte preferido dos brasileiros, atrás apenas do futebol. O evento UFC Rio de 2011, por exemplo, esgotou todos os 14 mil ingressos em menos de 90 minutos de vendas, demonstrando a grande expectativa no país. Tal rapidez apenas reflete essa magnitude em termos globais.

Atualmente, segundo a organização do UFC, o evento é transmitido para mais de 145 países, com uma média de 800 milhões de televisores ligados por evento. E com a popularidade do esporte e da marca, as casas de apostas oferecem uma ampla gama de probabilidades, assim como mercados para apostar no UFC, em todas as lutas.

## Justificativa

Este trabalho procurou analisar todas as lutas registradas no site ufcstats.com de 1993 até 2019, e prever os vencedores de cada luta com base na série histórica. Para isso, utilizamos algoritmos de aprendizado de máquina na linguagem de programação Python junto com a base de dados “*UFC-Fight historical data from 1993 to 2019*”, baixada no Kaggle em 08/12/2020.

## Problema

A base contém dados de lutas antigas onde não se tem registro de muitas variáveis e métricas. Além disso, nas primeiras edições da competição não havia categorias por peso, e a luta ocorria em um round único e sem limite de tempo. Essas características dificultam a criação de um modelo.

## Objetivo

Este trabalho tem como objetivo criar um modelo de previsão dos resultados das lutas do UFC, utilizando algoritmos de *Machine Learning*. Serão utilizadas as características e estatísticas de competições anteriores dos lutadores.

# **REFERENCIAL TEÓRICO**

## Ultimate Fight Championship: A Competição

Criado em 1993 como uma organização profissional de Artes Marciais Mistas (MMA - *Mixed Martial Arts*), o UFC revolucionou a indústria da luta e hoje se destaca tanto como uma marca global premium de esporte quanto como uma empresa de produção de conteúdo e o maior provedor de eventos Pay-Per-View (PPV) do mundo.

O UFC segue uma história e uma tradição de MMA competitivo que remonta ao Pancrácio, uma luta introduzida nos Jogos Olímpicos gregos no ano de 648 a.C. Nos anos 80, uma forma brasileira de MMA conhecida como Vale-Tudo despertou o interesse local pelo esporte. O UFC então introduziu o MMA organizado e sancionado nos Estados Unidos.

O objetivo era encontrar o "Campeão Supremo de Luta" (Ultimate Fighting Champion) organizando um torneio de uma noite com os melhores atletas das diversas modalidades de artes marciais, incluindo karatê, jiu-jítsu, boxe, kickboxing, grappling, wrestling, sumô e outros esportes de combate. O vencedor do torneio seria coroado o campeão.

O primeiro evento foi realizado em 1993 na McNichols Sports Arena em Denver, Colorado. As primeiras competições do *Ultimate Fighting Championship* buscavam identificar a arte marcial mais eficaz em uma competição com regras mínimas e em um round único sem limite de tempo. Também não havia categorias de peso entre os competidores. Em eventos subsequentes, os lutadores começaram a adotar técnicas eficazes de mais de uma disciplina, o que indiretamente ajudou a criar o MMA.

No UFC 5, foi introduzido a primeira luta única, uma revanche do UFC 1 com o tricampeão Royce Gracie e Ken Shamrock, chamada de "*The Superfight*". Isso se mostrou um desenvolvimento importante, porque as lutas únicas contariam com lutadores que não sofreram nenhum dano anterior no mesmo evento, ao contrário das lutas de torneios. Mais tarde, o "*Superfight*" acabaria eliminando completamente as partidas do torneio.

No final dos anos 90, o UFC passou a receber fortes críticas nos Estados Unidos, e muitos Estados passaram a proibir o evento. Em resposta, o UFC aumentou a cooperação com as comissões atléticas estaduais e redesenhou suas regras para remover os elementos menos palatáveis das lutas, ao mesmo tempo em que manteve os elementos centrais da trocação e de grappling. No UFC 12, foi introduzida categorias de peso e o banimento da “pesca de anzol”. No UFC 14, as luvas passaram a ser obrigatórias, enquanto os chutes na cabeça do oponente no chão foram proibidos. O UFC 15, viu limitações como puxar os cabelos, e proibiu golpes na nuca e na cabeça, cabeçadas, manipulações de pequenas articulações e golpes na virilha. Com rounds de cinco minutos introduzidos no UFC 21, o UFC gradualmente se renomeou como um esporte, ao invés de um espetáculo.

A popularidade do esporte também foi notada pela comunidade de apostas esportivas quando o BodogLife.com, um site de apostas online, declarou em 2007 que naquele ano o UFC, pela primeira vez, ultrapassou o boxe em termos de receita de apostas. Na verdade, o UFC já havia quebrado os recordes de todos os tempos da indústria do pay-per-view em um único ano de negócios, gerando mais de 220 milhões de dólares em receitas em 2006, superando a WWE e o boxe.

Atualmente, o UFC se divide em nove categorias de peso:

* Peso Palha (*Strawweight*) - até 52,2 kg /115 lb (Feminino)
* Peso Mosca (*Flyweight*) - até 56,7kg / 125 lb (Masculino e Feminino)
* Peso Galo (*Bantamweight*) - até 61,2 kg / 135 lb (Masculino e Feminino)
* Peso Pena (*Featherweight*) - até 65,8kg / 145 lb (Masculino e Feminino)
* Peso Leve (*Lightweight*) - até 70,3 kg / 155 lb
* Peso Meio-Médio (*Welterweight*) - até 77,1 kg / 170 lb
* Peso Médio (*Middleweight*) - até 83,9 kg / 185 lb
* Peso Meio-Pesado (*Light Heavyweig*ht) - até 92,9 kg / 205 lb
* Peso Pesado (*Heavyweight*) - até 120,2 kg / 265 lb

# **METODOLOGIA**

## Instrumento de coleta de dados

O processo de coleta de dados se deu através da base de dados com todas as lutas do UFC na história da organização. Cada linha contém informações sobre os lutadores, detalhes da luta e o vencedor. Os dados foram extraídos do site do ufcstats.com pelo Rajeev Warrier. Este site contém muitas informações sobre cada luta e cada evento. Para extrair os dados, ele utilizou o Beautifulsoup, e o Pandas para processá-los. Além do arquivo processado (data.csv), também foram disponibilizados os arquivos brutos (raw\_fighter\_details.csv, raw\_total\_fight\_data.csv, preprocessed\_data.csv).

No dataset data.csv, os lutadores são representados por “*Red*” e “*Blue’*', de acordo com a cor da luva. Foram adicionados os prefixos “R” e “B” para representar as características dos competidores. Historicamente, o lutador *Red* é o favorito.

Abaixo podemos observar as siglas de cada coluna da base:

* \_opp\_ - média de dano feito pelo oponente no lutador;
* KD - número de *knockdowns*;
* SIG\_STR - número de ataques certeiros;
* SIG\_STR\_pct - percentual de ataques certeiros;
* TOTAL\_STR - total de ataques certeiros;
* TD - número de quedas;
* TD\_pct -percentual de quedas;
* SUB\_ATT - número de tentativas de *submission*;
* PASS - número de passadas de guarda;
* REV - número de reversões;
* HEAD - número de ataques certeiros na cabeça;
* BODY - número de ataques certeiros no corpo;
* CLINCH - número de *clinch*;
* GROUND - número de ataques certeiros no chão;
* win\_by - método de vitória;
* last\_round - última rodada da luta (por exemplo, se foi um nocaute no primeiro lugar, então será 1);
* last\_round\_time - tempo de luta no último round;
* Format - formato da luta (3 rodadas, 5 rodadas etc.);
* Referee - nome do arbitro;
* date - data da luta;
* location - local em que o evento ocorreu;
* Fight\_type - categoria e se é uma luta pelo título ou não;
* Winner - vencedor da luta;
* Stance - postura do lutador (ortodoxo, canhoto, etc.);
* Height\_cms - altura em centímetros;
* Reach\_cms - envergadura do lutador em centímetros;
* Weight\_lbs - peso do lutador em libras (lbs)
* age - idade do lutador
* title\_bout - valor booleano de se é luta pelo título ou não
* weight\_class - classe de peso da luta (peso galo, peso pesado, peso mosca feminino, etc.)
* no\_of\_rounds - número de rounds agendados;
* current\_lose\_streak - quantidade atual de perdas simultâneas do lutador;
* current\_win\_streak - quantidade atual de vitórias simultâneas do lutador;
* draw - número de empates na carreira do lutador no UFC;
* wins - número de vitórias na carreira do lutador no UFC;
* losses - número de derrotas na carreira do lutador no UFC;
* total\_rounds\_fought - média do total de rounds lutados pelo lutador;
* total\_time\_fought - contagem do tempo total gasto lutando em segundos;
* total\_title\_bouts - número total de disputas de título pelo lutador;
* win\_by\_Decision\_Majority - número de vitórias por decisão da maioria dos juízes no UFC;
* win\_by\_Decision\_Split - número de vitórias por decisão dividida dos juízes no UFC;
* win\_by\_Decision\_Unanimous - número de vitórias por decisão unânime dos juízes no UFC;
* win\_by\_KO/TKO - número de vitórias por nocaute do lutador no UFC;
* win\_by\_Submission - número de vitórias por finalização do lutador no UFC;
* win\_by\_TKO\_Doctor\_Stoppage - número de vitórias por paralisação médica do lutador no UFC;

Cada linha é uma compilação das estatísticas e características dos lutadores na competição. Assim, por exemplo, o lutador vermelho tem as estatísticas médias compiladas de todas as lutas, exceto a atual. As estatísticas incluem o dano feito pelo lutador vermelho no oponente e o dano feito pelo oponente no lutador (representado por 'opp' nas colunas) em todas as lutas que esse lutador vermelho teve, exceto esta porque não ocorreu ainda (nos dados). A mesma informação existe para o lutador azul. A variável de destino é “*Winner*”, que é a única coluna que informa o que aconteceu.

Na base de dados, algumas lutas receberam a nomenclatura *Open Weight* e *Catch Weight*, referente ao início do UFC quando não havia divisões de categoria por peso e as lutas ocorriam em um round único. Para melhorar o modelo, essas categorias foram removidas.

## Análise de dados

A análise dos dados obtidos foi realizada utilizando a ferramenta Python, como solicitado. Inicialmente, carregamos as bibliotecas necessárias e os dados necessários.

O arquivo “raw\_fighter\_details.csv” (dataset de lutadores) contém os atributos físicos de cada um dos atletas, dentre os quais destacamos: altura, peso e envergadura.

O arquivo “data.csv” (dataset de lutas) contém dados da luta, atributos físicos dos dois lutadores, atributos relacionados às suas carreiras e seus desempenhos nas lutas mais recentes.

Analisando o dataset de lutas, observamos alguns campos vazios e os substituímos por “Nan”. Também convertemos os campos de data para “*datetime*” e arredondamos os campos numéricos

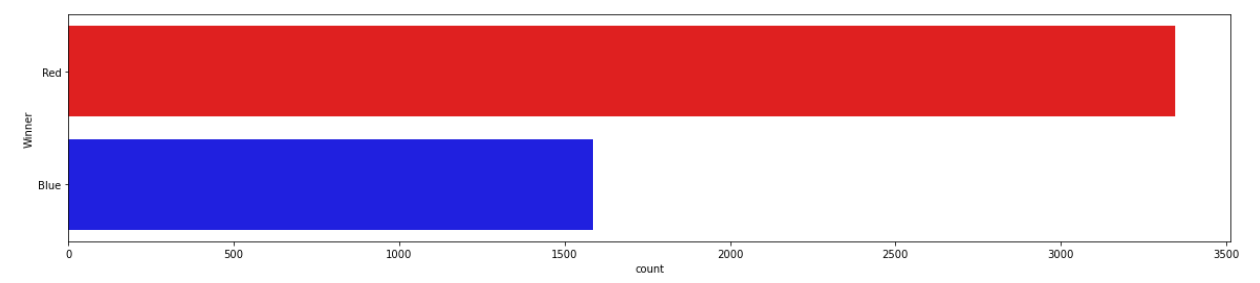
Vimos que os atributos físicos faltantes na base de lutas também não constam no dataset de lutadores.

Como exposto anteriormente, no começo do UFC não havia divisão de categorias por peso e o combate acontece em um round único sem tempo. Na base de dados, essas lutas receberam a nomenclatura de “*Open Weight*” e “*Catch Weight*”. Para a criação de um melhor modelo, essas categorias foram removidas da base.

Como nosso objetivo é prever o vencedor da luta, excluímos do dataset de lutas os registros de empate.

Uma característica marcante do UFC é que o lutador favorito sempre utiliza a cor vermelha e o desafiante fica com a cor azul. Diante desta informação, espera-se que o favoritismo seja do vermelho. A base comprova isso com uma taxa de vitórias do vermelho igual a 67,85%. Esse valor é o nosso baseline

Figura 1: gráfico favoritismo



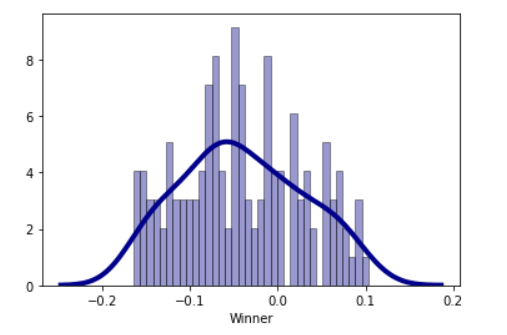
Fonte:

Avançando na análise do dataset, fizemos alguns gráficos de correlação. Cada quadrado mostra a correlação entre as variáveis em cada eixo. Valores mais próximos de zero significam que não há tendência linear entre as duas variáveis. Quanto mais próxima de 1 a correlação é, mais positivamente correlacionados eles são; isto é, à medida que um aumenta, o outro aumenta e quanto mais próximo de 1, mais forte é essa relação. Uma correlação mais próxima de -1 é semelhante, mas em vez de aumentar, uma variável diminuirá à medida que a outra aumenta. Quanto maior o número e mais escura a cor, maior é a correlação entre as duas variáveis.

Através dessa análise, podemos identificar pares de variáveis com correlação alta e avaliar a possibilidade de manter só uma delas no modelo, para evitar redundâncias desnecessárias.

Analisando a correlação de cada variável isoladamente com a variável “*Winner*”, vencedor da luta, vemos que não há uma variável que se destaque, de onde concluímos que a combinação das várias variáveis é que pode contribuir para o resultado da luta.

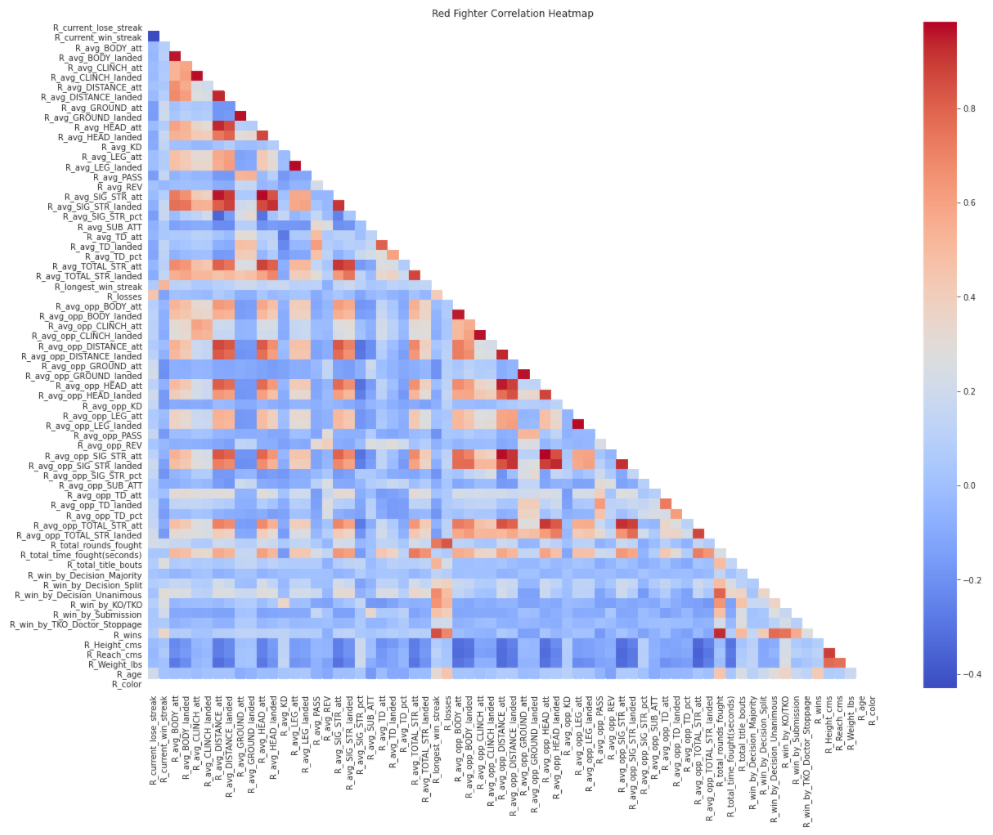
Figura 2: Gráfico de correlação



Fonte:

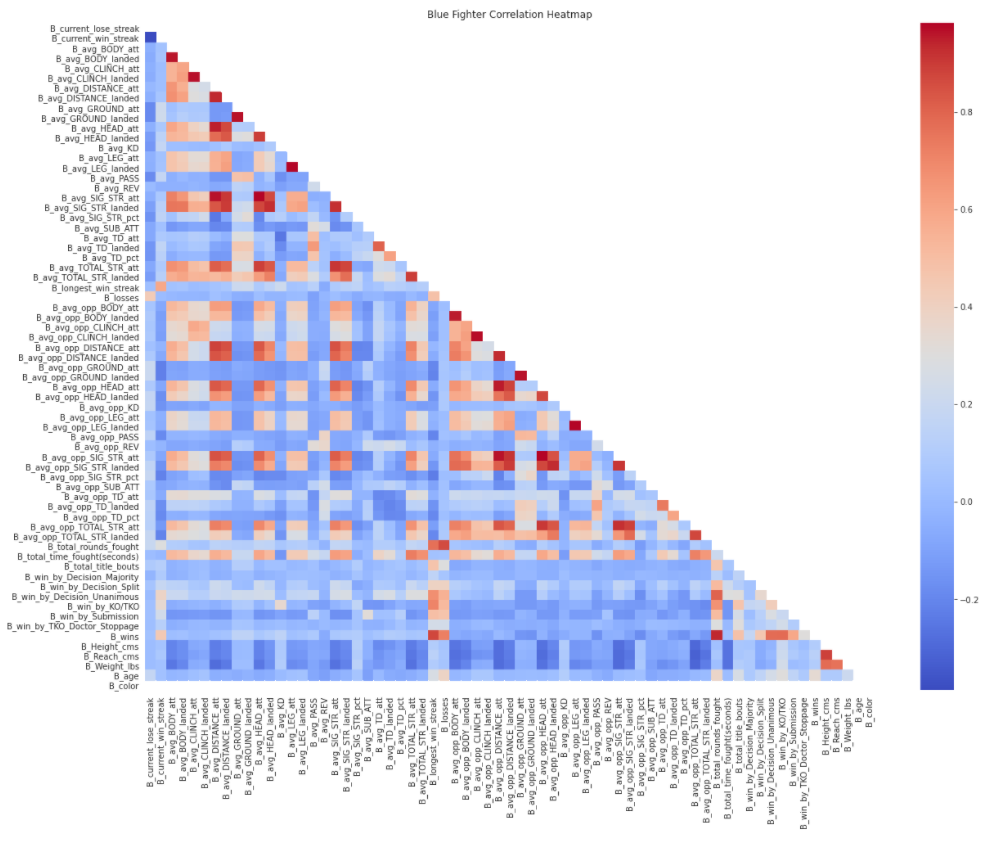
Mapas de calor com as variáveis de cada lutador:

Figura 3: Mapa de calor vermelho



Fonte:

Figura 4: Mapa de calor azul



Fonte:

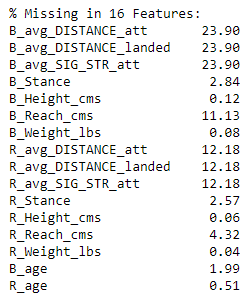
Analisando os mapas de calor e o significado de cada variável, selecionamos as variáveis que julgamos mais relevantes e as que classificamos como importantes, eliminando redundâncias:

['date', 'B\_fighter', 'R\_fighter', 'Winner', 'B\_avg\_DISTANCE\_att', 'B\_avg\_DISTANCE\_landed', 'B\_avg\_SIG\_STR\_att', 'B\_current\_lose\_streak', 'B\_current\_win\_streak', 'B\_longest\_win\_streak', 'B\_losses', 'B\_total\_rounds\_fought', 'B\_total\_title\_bouts', 'B\_win\_by\_Decision\_Majority', 'B\_win\_by\_Decision\_Split', 'B\_win\_by\_Decision\_Unanimous', 'B\_win\_by\_KO/TKO', 'B\_win\_by\_Submission', 'B\_win\_by\_TKO\_Doctor\_Stoppage', 'B\_wins', 'B\_Stance', 'B\_Height\_cms', 'B\_Reach\_cms', 'B\_Weight\_lbs', 'R\_avg\_DISTANCE\_att', 'R\_avg\_DISTANCE\_landed', 'R\_avg\_SIG\_STR\_att', 'R\_current\_lose\_streak', 'R\_current\_win\_streak', 'R\_longest\_win\_streak', 'R\_losses', 'R\_total\_rounds\_fought', 'R\_total\_title\_bouts', 'R\_win\_by\_Decision\_Majority', 'R\_win\_by\_Decision\_Split', 'R\_win\_by\_Decision\_Unanimous', 'R\_win\_by\_KO/TKO', 'R\_win\_by\_Submission', 'R\_win\_by\_TKO\_Doctor\_Stoppage', 'R\_wins', 'R\_Stance', 'R\_Height\_cms', 'R\_Reach\_cms', 'R\_Weight\_lbs', 'B\_age', 'R\_age']

Excluindo registros sem altura e/ou sem peso

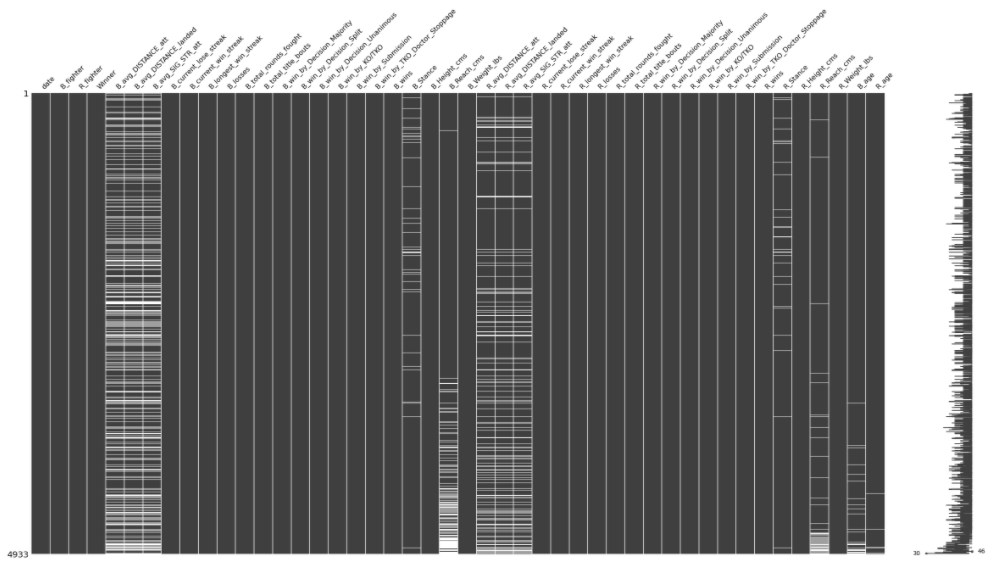
Observando o percentual de dados faltantes no dataset original, vemos uma grande quantidade de variáveis com taxa de ausência de 23,9%. Após filtrar as variáveis consideradas importantes, chegamos a um dataset bem mais homogêneo:

Figura 5: avaliar nome



Fonte:

Figura 6: Avaliar nome 2



Fonte:

Decidimos então eliminar 3 variáveis com maior taxa de ausência, removendo essas colunas do dataframe:

B\_avg\_DISTANCE\_att 23.90

B\_avg\_DISTANCE\_landed 23.90

B\_avg\_SIG\_STR\_att 23.90

R\_avg\_DISTANCE\_att 12.18

R\_avg\_DISTANCE\_landed 12.18

R\_avg\_SIG\_STR\_att 12.18

A próxima variável com ausência elevada é a envergadura. Neste caso, não podemos excluir as colunas do dataframe, porque esse é um dos atributos físicos mais importantes de qualquer lutador. Também optamos por não excluir da base os registros sem envergadura, por representarem um percentual importante no total de registros.

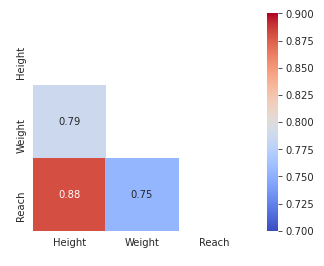
B\_Reach\_cms 11.13

R\_Reach\_cms 4.32

Pensamos inicialmente em fazer um cálculo médio da envergadura por categoria, mas vimos que essa abordagem poderia introduzir um ruído significativo no modelo, pois dentro de uma mesma categoria temos valores bem distantes.

Então, visto que a altura, principalmente, possui altíssima correlação com a envergadura, optamos por fazer um modelo de predição para obter a envergadura dos lutadores que estão sem esse dado.

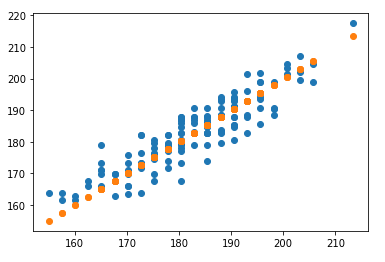
Figura 7: Modelo predição envergadura



Fonte:

Rodamos alguns modelos de regressão para prever a envergadura e o melhor resultado foi obtido com Redes Neurais.

Figura 8: Modelo 1



Fonte:



Utilizamos esse modelo para prever a envergadura e preenchemos o dataset com os valores obtidos. Após completar os valores de envergadura para os lutadores azul e vermelho, convertemos as colunas “B\_Reach\_cms” e “R\_Reach\_cms” para float.

Através desse modelo, conseguimos manter o dataset com a quantidade de registros original.

Depois convertemos as variáveis categóricas “B\_Stance” e “R\_Stance” em colunas numéricas:

B\_Stance\_Open Stance

B\_Stance\_Orthodox

B\_Stance\_Sideways

B\_Stance\_Southpaw

B\_Stance\_Switch

Substituímos True e False por 1 e 0 na coluna 'title\_bout'

E finalmente, eliminamos os registros em que as seguintes variáveis estavam vazias (quantidade irrisória):

B\_Height\_cms 0.12

R\_Height\_cms 0.06

B\_Weight\_lbs 0.08

R\_Weight\_lbs 0.04

B\_age 1.99

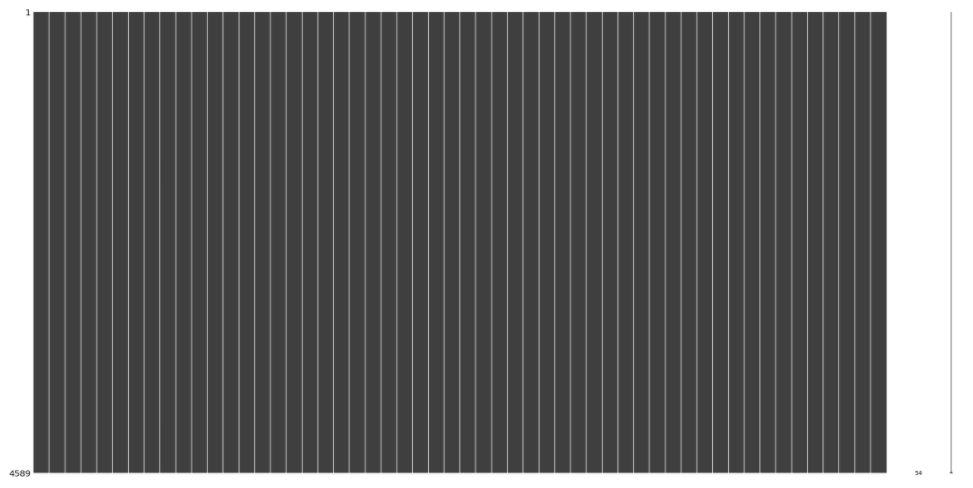
R\_age 0.51

B\_Stance 2.84

R\_Stance 2.57

Chegando ao dataset tratado, sem nenhum valor ausente:

Figura 9: Avaliar nome 4

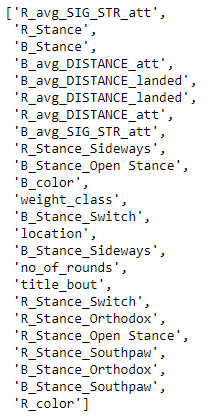


Fonte: criado pelo autor

Lista final de colunas, após tratamento completo da base:

['R\_fighter', 'B\_fighter', 'date', 'location', 'Winner', 'title\_bout', 'weight\_class', 'no\_of\_rounds', 'B\_current\_lose\_streak', 'B\_current\_win\_streak', 'B\_longest\_win\_streak', 'B\_losses', 'B\_total\_rounds\_fought', 'B\_total\_title\_bouts', 'B\_win\_by\_Decision\_Majority', 'B\_win\_by\_Decision\_Split', 'B\_win\_by\_Decision\_Unanimous', 'B\_win\_by\_KO/TKO', 'B\_win\_by\_Submission', 'B\_win\_by\_TKO\_Doctor\_Stoppage', 'B\_wins', 'B\_Height\_cms', 'B\_Reach\_cms', 'B\_Weight\_lbs', 'R\_current\_lose\_streak', 'R\_current\_win\_streak', 'R\_longest\_win\_streak', 'R\_losses', 'R\_total\_rounds\_fought', 'R\_total\_title\_bouts', 'R\_win\_by\_Decision\_Majority', 'R\_win\_by\_Decision\_Split', 'R\_win\_by\_Decision\_Unanimous', 'R\_win\_by\_KO/TKO', 'R\_win\_by\_Submission', 'R\_win\_by\_TKO\_Doctor\_Stoppage', 'R\_wins', 'R\_Height\_cms', 'R\_Reach\_cms', 'R\_Weight\_lbs', 'B\_age', 'R\_age', 'B\_color', 'R\_color', 'B\_Stance\_Open Stance', 'B\_Stance\_Orthodox', 'B\_Stance\_Sideways', 'B\_Stance\_Southpaw', 'B\_Stance\_Switch', 'R\_Stance\_Open Stance', 'R\_Stance\_Orthodox', 'R\_Stance\_Southpaw', 'R\_Stance\_Switch', 'R\_Stance\_Sideways']

Comparando essa lista final de colunas, com a lista de variáveis que consideramos importantes na análise exploratória inicial e gráficos de correlação, ficamos muito satisfeitos com o resultado, pois as principais diferenças observadas são as colunas que decidimos excluir do dataframe e as colunas categóricas que foram convertidas:

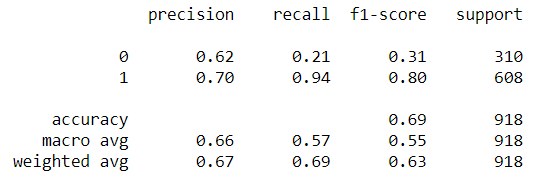


# **PREVISÃO DE RESULTADOS DE LUTAS**

Partindo de uma base desbalanceada, os melhores resultados obtidos foram com os modelos Floresta Aleatória e Rede Neural, que obtiveram acurácia de 69%, que não é um valor expressivo, visto que apenas a cor da luva representa um favoritismo de 67,85% para o vermelho (baseline). Além disso os modelos também apresentaram forte desequilíbrio no recall.

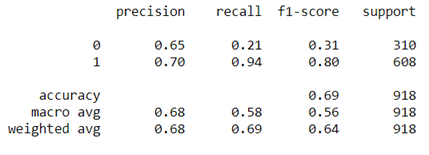
Floresta Aleatória:

Figura 10: Avaliar nome



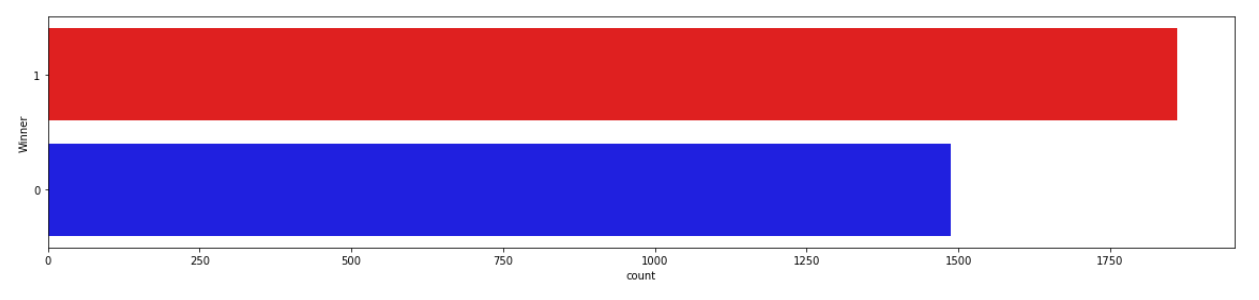
Redes Neurais:

Figura 11: avaliar nome



Então partimos para uma abordagem diferente, onde tentamos forçar o balanceamento da base, eliminando 40% dos registros de vitórias do vermelho, de maneira aleatória.

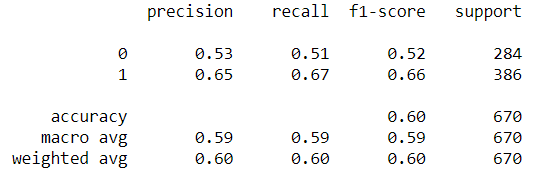
Figura 12: Avaliar nome 5



Nessa abordagem, foram eliminados cerca de 1600 registros da base (redução considerável). E mesmo forçando o balanceamento, ainda assim vemos um certo desequilíbrio nos valores de recall e f1-score.

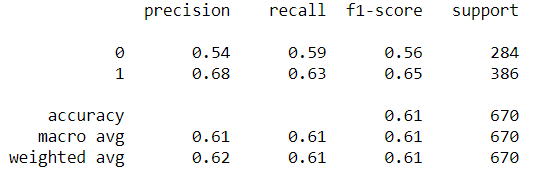
XGBoost:

Figura 13: avaliar nome



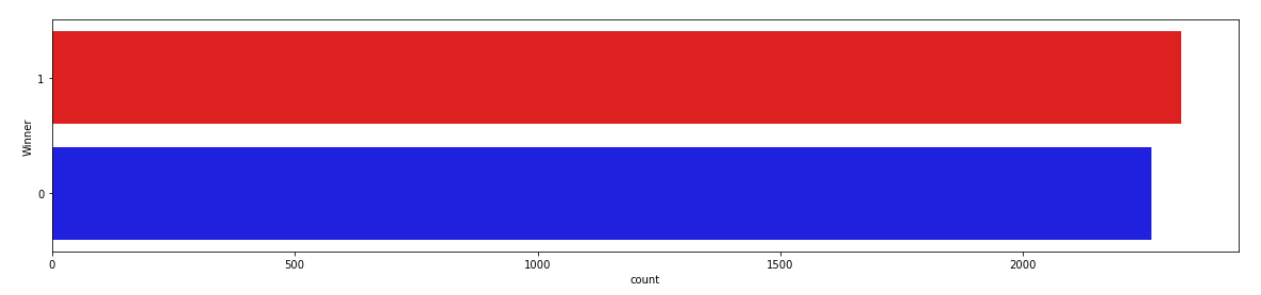
Redes Neurais:

Figura 14: Avaliar nome



Decidimos então realizar o balanceamento da base de uma maneira diferente, em vez de apagar 40% das vitórias do vermelho, invertemos as cores em 25% das lutas vencidas pelo vermelho. Assim obtivemos um dataset balanceado, sem perda de registros:

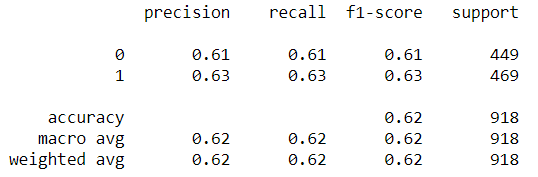
Figura 15: Avaliar nome 6



Neste modelo, a Rede Neural apresentou o melhor resultado, com valores equilibrados de precision e recall, além de uma acurácia de 62% sobre uma base balanceada.

Redes Neurais:

Figura 16: avaliar nome



Filtrando probabilidades superiores a 85%, conseguimos obter taxas de acerto de até 88,89%, o que representa um resultado extremamente satisfatório.

Lutas de 2020:

Modelagem por categoria de peso

Os resultados por categoria de peso não são expressivos, provavelmente por termos amostras bem reduzidas em cada uma das categorias.

# **CONCLUSÃO**

Algumas situações de combate não foram consideradas neste trabalho e podem ser abordadas, como o tempo que o lutador fica em determinadas ações da luta. Além disso, outras variáveis como “Dias sem lutar”, se o lutador estava se recuperando de lesão ou não e as especialidades de cada lutador podem colaborar para aumentar o acerto do modelo preditivo

# **REFERÊNCIAS**

TOLENTINO, Volney. **Novas parcerias compõem estratégia para crescimento do UFC**, 2019 Disponível em: <<https://cebolaverde.com.br/esportes/novas-parcerias-compoem-estrategia-para-crescimento-do-ufc/>> Acessado em XX de XX, XXXX.

<https://rollingstone.uol.com.br/edicao/59/o-esporte-numero-2-do-brasil/>

<https://www.ufc.com.br/news/ufc-networkr-disponivel-em-20-paises-da-america-latina-partir-de-1o-de-setembro#:~:text=Al%C3%A9m%20do%20seu%20alcance%20no,mundo%2C%20em%2028%20idiomas%20diferentes>.

<https://www.kaggle.com/rajeevw/ufcdata>

<http://www.ufcstats.com/>