Instituto INFNET
Escola Superior da Tecnologia da Informação
Pós-Graduação MIT em BIG DATA

Cesar Mansur
Leonardo Arariba
Romario Gomes

Estimativa da probabilidade de vitória em lutas no UFC a partir de um modelo preditivo utilizando aprendizado de máquina em linguagem *Python*

RIO DE JANEIRO

2021

Cesar Mansur
Leonardo Arariba
Romario Gomes

Estimativa da probabilidade de vitória em lutas no UFC a partir de um modelo preditivo utilizando aprendizado de máquina em linguagem *Python*

Trabalho de Conclusão de Bloco apresentado ao programa de Pós-graduação MIT em BIG DATA do Instituto INFNET, como requisito final para avaliação da disciplina “Indexação e Tratamento de Dados Heterogêneos: Valor”

Orientador: Kleyton Cotta

RIO DE JANEIRO

2021

RESUMO

O *Ultimate Fighting Championship* ou UFC como é comumente conhecido, foi fundado em 1993 e rapidamente se tornou conhecido como maior evento de artes marciais mistas (do inglês: MMA – *Mixed Martial Arts*), superando outras modalidades, como por exemplo, o boxe. Com base no levantamento de dados históricos dos lutadores e resultados de lutas anteriores, juntamente com o banco de dados oficial de atletas do UFC, esse trabalho pretende responder se é possível prever os resultados de futuras lutas no UFC, utilizando apenas dados conhecidos sobre os lutadores antes de entrarem no octógono.

Palavras chave: UFC, *Python*, Ciência de Dados, Aprendizado de Máquina

ABSTRACT

The Ultimate Fighting Championship or UFC as it is commonly known, was founded in 1993 and quickly became known as the biggest mixed martial arts event (in English: MMA – Mixed Martial Arts), surpassing other modalities, such as boxing. Based on the survey of historical fighter data and results of previous fights, along with the official database of UFC athletes, this work aims to answer whether it is possible to predict the results of future fights in the UFC, using only known fighter data before entering the octagon.

***Keywords*:** UFC, *Python*, *Data Science*, *Machine Learning*

LISTA DE ILUSTRAÇÕES

SUMÁRIO

1 INTRODUÇÃO

1.1 Apresentação

O *Ultimate Fighting Championship* evoluiu 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.

1.2 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 (**kaggle.com**) em 08 de dezembro, 2020.

1.3 Problema

A base contém dados de lutas antigas onde não há 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, sem limite de tempo predefinido. Essas características dificultam a criação de um modelo.

1.4 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 físicas, atributos dos lutadores e seus históricos de resultados em lutas anteriores.

2 REFERENCIAL TEÓRICO

2.1 *Ultimate Fighting 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

3 METODOLOGIA

3.1 Instrumento de coleta de dados

O processo se deu através da coleta da base de dados com todas as lutas do UFC na história da organização, até o final do ano de 2019. A base inicial possui 5144 regsitros e 145 colunas. 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 das luvas (vermelha e azul, respectivamente). Foram adicionados os prefixos "R" e "B" para representar as características dos competidores. Historicamente, o lutador *Red* é o favorito para vencer a luta.

Abaixo podemos observar os significados das 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 de características dos lutadores e seus dados históricos na competição. Assim, por exemplo, o lutador vermelho tem os dados obtidos de todas as suas 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, ou seja, quem venceu.

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, foram removidas da base as lutas referentes a essas categorias. Essa remoção reduziu a quantidade de registros na base para 5014.

3.2 Análise de dados

A análise dos dados obtidos foi realizada utilizando a linguagem *Python*, como solicitado. Inicialmente, carregamos as bibliotecas 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.

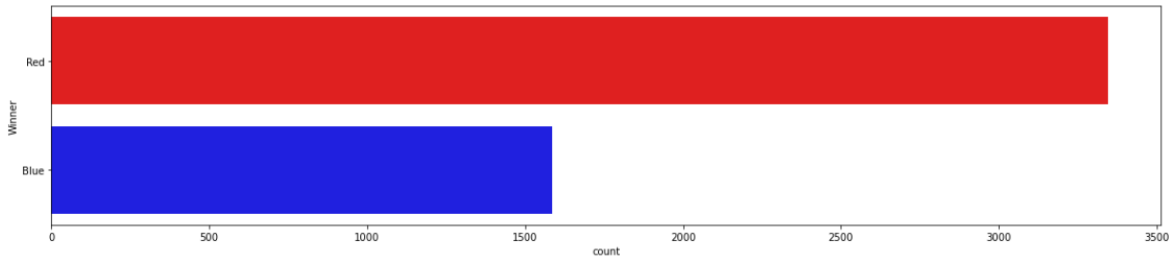
Vimos que os atributos físicos faltantes na base de lutas também não constam no *dataset* de lutadores.

Como exposto anteriormente, no começo do UFC não havia divisão de categorias por peso e o combate acontecia em um *round* único sem tempo predefinido. Por este motivo os registros dessas lutas foram removidos.

Como nosso objetivo é prever o vencedor da luta, excluímos do *dataset* de lutas os registros de empate, reduzindo a quantidade de registros para 4933.

Uma característica marcante do UFC é que o lutador favorito sempre utiliza a luva 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.

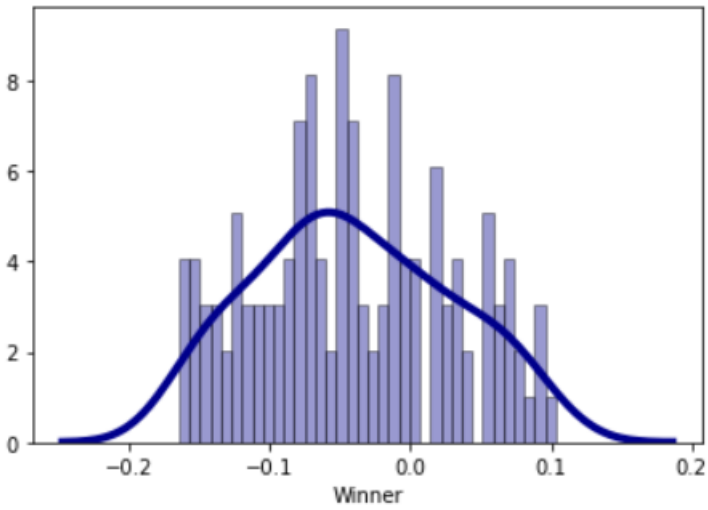
Figura 1: Gráfico Baseline – Base Desbalanceada



Fonte: Criado pelo autor

Avançando na análise do *dataset*, fizemos alguns gráficos de correlação. 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 de diversas variáveis é o que pode contribuir para o resultado da luta.

Figura 2: Gráfico de correlação



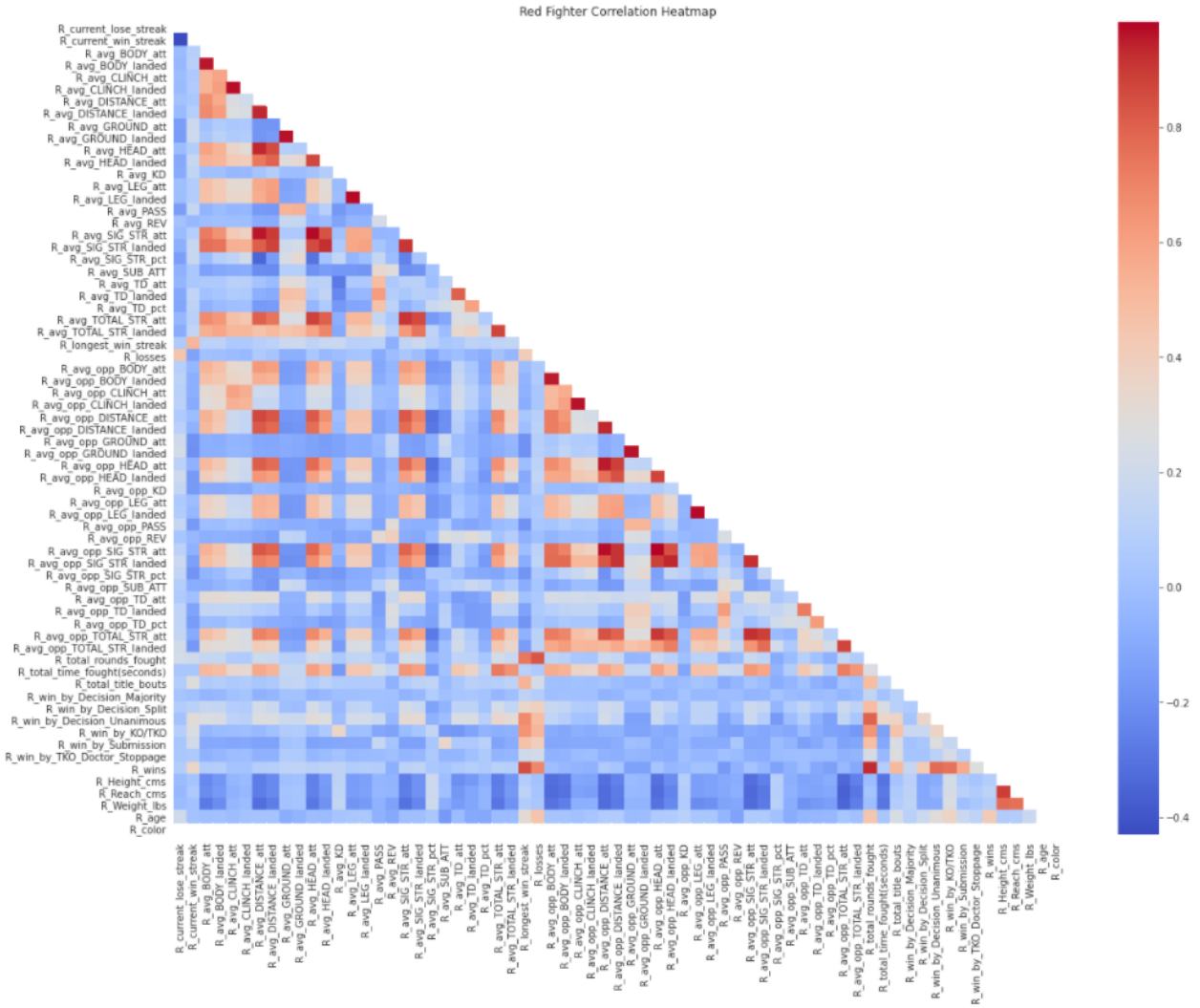
Fonte: Criado pelo autor

Analisando mapas de calor com as variáveis dos lutadores, vemos que 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.

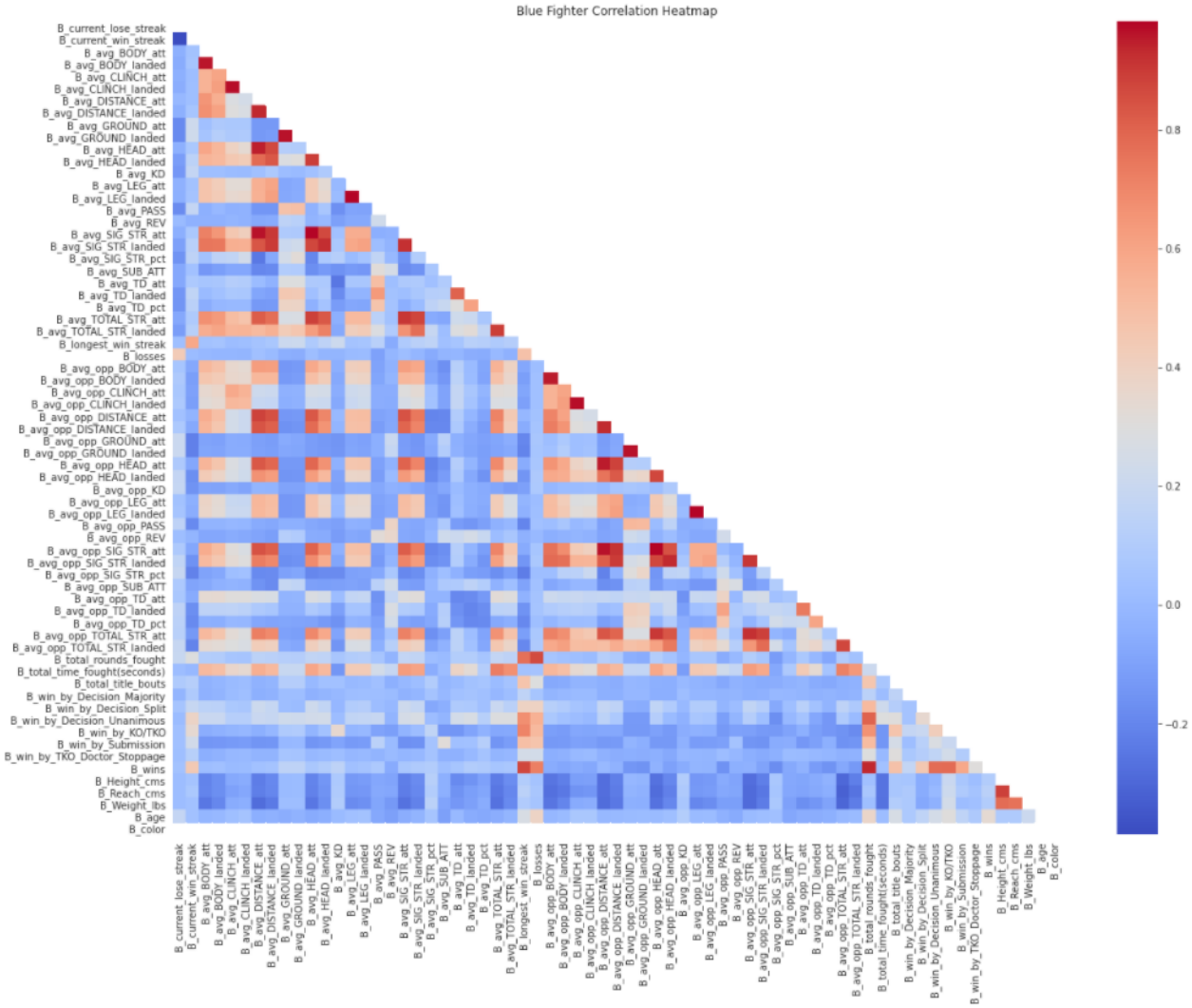
Através dessa análise, podemos identificar pares de variáveis com correlação alta e avaliar a possibilidade de manter só uma delas no modelo, para evitar redundâncias desnecessárias.

Figura 3: Mapa de calor – Lutador vermelho



Fonte: Criado pelo autor

Figura 4: Mapa de calor – Lutador azul



Fonte: Criado pelo autor

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']
```

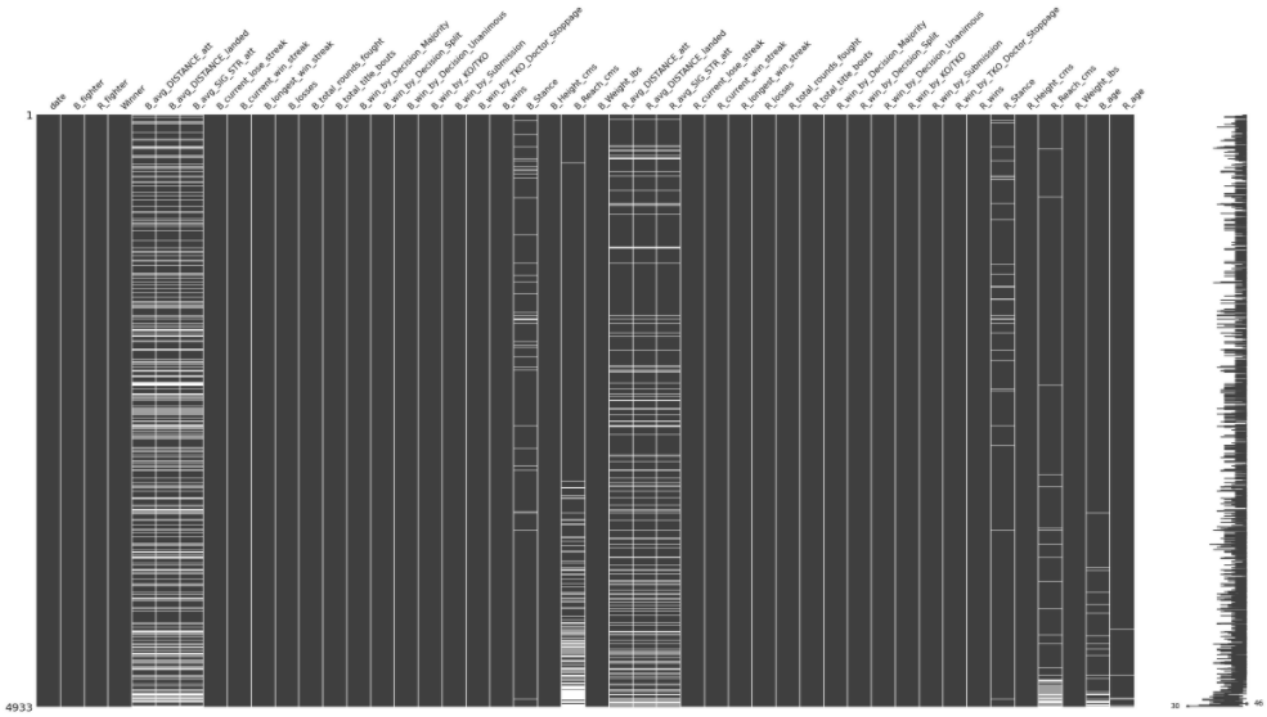
Observando o percentual de dados faltantes no dataset original, vemos uma grande quantidade de variáveis com taxa de ausência de 23,9%. Após filtrar as variáveis consideradas importantes, chegamos a um dataset bem mais homogêneo:

Figura 5: Taxa de ausência – Variáveis importantes

```
% Missing in 16 Features:
B_avg_DISTANCE_att       23.90
B_avg_DISTANCE_landed    23.90
B_avg_SIG_STR_att        23.90
B_Stance                  2.84
B_Height_cms              0.12
B_Reach_cms              11.13
B_Weight_lbs              0.08
R_avg_DISTANCE_att       12.18
R_avg_DISTANCE_landed    12.18
R_avg_SIG_STR_att        12.18
R_Stance                  2.57
R_Height_cms              0.06
R_Reach_cms               4.32
R_Weight_lbs              0.04
B_age                     1.99
R_age                     0.51
```

Fonte: Criado pelo autor

Figura 6: Gráfico Nan – Base original



Fonte: Criado pelo autor

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

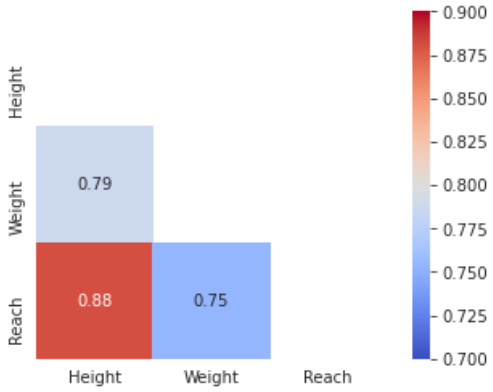
```
B_avg_DISTANCE_att          23.90
B_avg_DISTANCE_landed       23.90
B_avg_SIG_STR_att           23.90
R_avg_DISTANCE_att          12.18
R_avg_DISTANCE_landed       12.18
R_avg_SIG_STR_att           12.18
```

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.

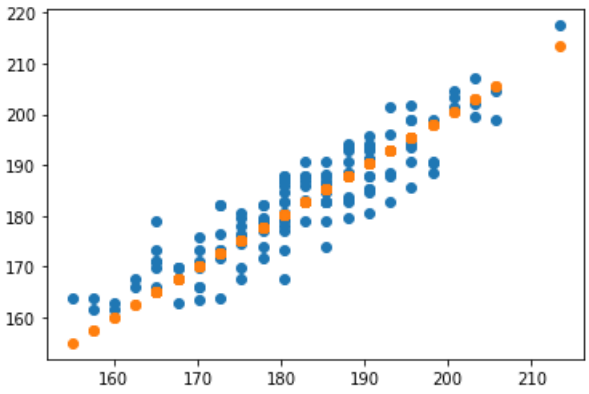
Figura 7: Mapa de calor – Envergadura x Altura x Peso



Fonte: Criado pelo autor

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

Figura 8: Modelo Rede Neural – Envergadura



Métricas:

```
MAE: 3.5154778221330614
MSE: 21.376598997981255
RMSE: 4.623483426809407
```

Fonte: Criado pelo autor

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 a quantidade de registros no *dataset*, ou seja, 4933. Depois convertemos as variáveis categóricas “B_Stance” e “R_Stance” em colunas numéricas:

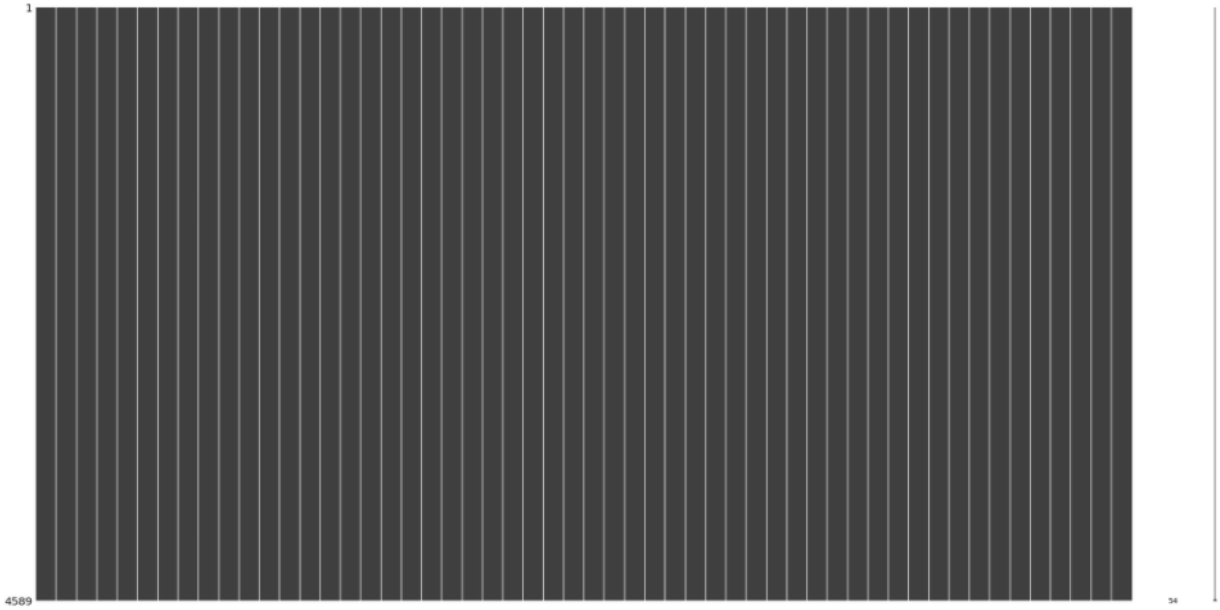
```
B_Stance_Open Stance
B_Stance_Orthodox
B_Stance_Sideways
B_Stance_Southpaw
B_Stance_Switch
```

Substituímos “True” e “False” por 1 e 0, respectivamente, 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
```

Chegamos então ao *dataset* tratado, sem nenhum valor ausente e um total de 4589 registros e 54 colunas:

Figura 9: Gráfico Nan – Base tratada



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.

Figura 10: Comparação de colunas

```
['R_avg_SIG_STR_att',
 'R_Stance',
 'B_Stance',
 'B_avg_DISTANCE_att',
 'B_avg_DISTANCE_landed',
 'R_avg_DISTANCE_landed',
 'R_avg_DISTANCE_att',
 'B_avg_SIG_STR_att',
 'R_Stance_Sideways',
 'B_Stance_Open Stance',
 'B_color',
 'weight_class',
 'B_Stance_Switch',
 'location',
 'B_Stance_Sideways',
 'no_of_rounds',
 'title_bout',
 'R_Stance_Switch',
 'R_Stance_Orthodox',
 'R_Stance_Open Stance',
 'R_Stance_Southpaw',
 'B_Stance_Orthodox',
 'B_Stance_Southpaw',
 'R_color']
```

Fonte: Criado pelo autor

4 PREVISÃO DE RESULTADOS DE LUTAS

4.1 Cenário 1 – Base Desbalanceada

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.

Figura 11: Floresta Aleatória – Base desbalanceada

	precision	recall	f1-score	support
0	0.62	0.21	0.31	310
1	0.70	0.94	0.80	608
accuracy			0.69	918
macro avg	0.66	0.57	0.55	918
weighted avg	0.67	0.69	0.63	918

Fonte: Criado pelo autor

Figura 12: XGBoost – Base desbalanceada

	precision	recall	f1-score	support
0	0.47	0.32	0.38	310
1	0.70	0.82	0.75	608
accuracy			0.65	918
macro avg	0.59	0.57	0.57	918
weighted avg	0.62	0.65	0.63	918

Fonte: Criado pelo autor

Figura 13: Rede Neural – Base desbalanceada

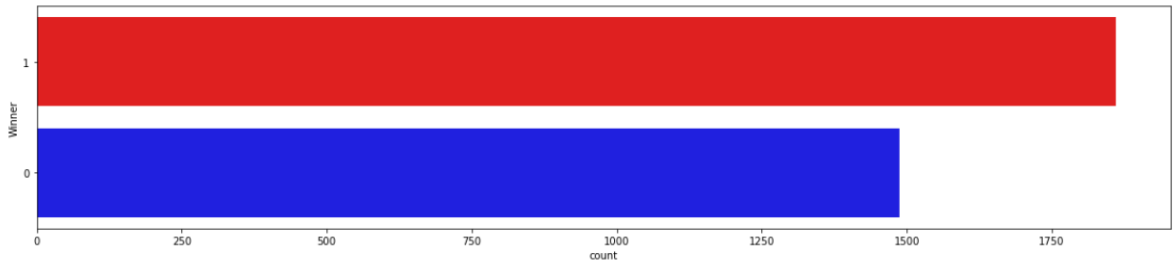
	precision	recall	f1-score	support
0	0.65	0.21	0.31	310
1	0.70	0.94	0.80	608
accuracy			0.69	918
macro avg	0.68	0.58	0.56	918
weighted avg	0.68	0.69	0.64	918

Fonte: Criado pelo autor

4.2 Cenário 2 – Base Reduzida

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

Figura 14: Gráfico Baseline – Base reduzida



Fonte: Criado pelo autor

Nessa abordagem, foram eliminadas 1240 vitórias do lutador vermelho (redução considerável) e base ficou com 3349 registros. Com essa proposta, atingimos uma acurácia de 61% sobre um *baseline* de 56%. Mesmo forçando o balanceamento, ainda assim tivemos um certo desequilíbrio nos valores de *recall* e *f1-score*.

Figura 15: Floresta Aleatória – Base reduzida

	precision	recall	f1-score	support
0	0.55	0.51	0.53	284
1	0.66	0.69	0.67	386
accuracy			0.61	670
macro avg	0.60	0.60	0.60	670
weighted avg	0.61	0.61	0.61	670

Fonte: Criado pelo autor

Figura 16: XGBoost – Base reduzida

	precision	recall	f1-score	support
0	0.53	0.51	0.52	284
1	0.65	0.67	0.66	386
accuracy			0.60	670
macro avg	0.59	0.59	0.59	670
weighted avg	0.60	0.60	0.60	670

Fonte: Criado pelo autor

Figura 17: Rede Neural – Base reduzida

	precision	recall	f1-score	support
0	0.54	0.59	0.56	284
1	0.68	0.63	0.65	386
accuracy			0.61	670
macro avg	0.61	0.61	0.61	670
weighted avg	0.62	0.61	0.61	670

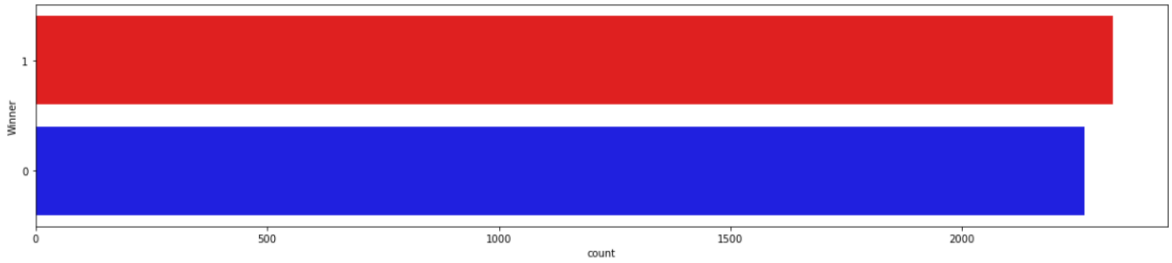
Fonte: Criado pelo autor

Conforme aumentamos a quantidade de vitórias vermelhas excluídas da base, aproximando o balanceamento ao valor de 50%, maior foi o desequilíbrio observado no recall, o que provavelmente se explica pelo fato de que o lutador vermelho é o favorito, então ao forçar o balanceamento da base introduzimos consequentemente um favoritismo para o lutador azul.

4.3 Cenário 3 – Base Balanceada

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, mantendo a base com 4589 registros e partindo para os modelos com um baseline de 51%.

Figura 18: Gráfico Baseline – Base Balanceada



Fonte: Criado pelo autor

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

Figura 19: Floresta Aleatória – Base balanceada

	precision	recall	f1-score	support
0	0.57	0.59	0.58	449
1	0.59	0.57	0.58	469
accuracy			0.58	918
macro avg	0.58	0.58	0.58	918
weighted avg	0.58	0.58	0.58	918

Fonte: Criado pelo autor

Figura 20: XGBoost – Base balanceada

	precision	recall	f1-score	support
0	0.58	0.63	0.60	449
1	0.61	0.57	0.59	469
accuracy			0.60	918
macro avg	0.60	0.60	0.60	918
weighted avg	0.60	0.60	0.60	918

Fonte: Criado pelo autor

Figura 21: Rede Neural – Base balanceada

	precision	recall	f1-score	support
0	0.61	0.61	0.61	449
1	0.63	0.63	0.63	469
accuracy			0.62	918
macro avg	0.62	0.62	0.62	918
weighted avg	0.62	0.62	0.62	918

Fonte: Criado pelo autor

Elegemos então o modelo de Rede Neural do Cenário 3 como o melhor deste trabalho e deste ponto em diante ele será referenciado como **modelo final**.

A figura abaixo mostra os valores das probabilidades de vitória de cada lutador, de acordo com o modelo final, assim como o vencedor real da luta.

Figura 22: Tabela de predição com probabilidades – Base balanceada

	Winner	R_fighter	B_fighter	date	location	weight_class	Resultado_Previsto	Prob_Win_Blue	Prob_Win_Red
0	Blue	Nam Phan	Mike Brown	2011-08-06	Philadelphia, Pennsylvania, USA	Featherweight	Red	0.429753	0.570247
1	Blue	Vik Grujic	Brendan O'Reilly	2015-05-09	Adelaide, South Australia, Australia	Welterweight	Red	0.476783	0.523217
2	Blue	Daniel Roberts	Claude Patrick	2011-04-30	Toronto, Ontario, Canada	Welterweight	Red	0.478677	0.521323
3	Red	Stephan Bonnar	Eric Schafer	2007-10-20	Cincinnati, Ohio, USA	Light Heavyweight	Red	0.312940	0.687060
4	Blue	Mike Pierce	Jon Fitch	2009-12-12	Memphis, Tennessee, USA	Welterweight	Blue	0.827640	0.172360
5	Red	Jake Matthews	Li Jingliang	2018-02-10	Perth, Western Australia, Australia	Welterweight	Blue	0.574999	0.425001
6	Blue	Kuniyoshi Hironaka	Thiago Alves	2007-09-19	Las Vegas, Nevada, USA	Welterweight	Blue	0.636134	0.363866
7	Blue	Darrell Montague	Willie Gates	2015-07-12	Las Vegas, Nevada, USA	Flyweight	Blue	0.604638	0.395362
8	Blue	David Branch	Jack Hermansson	2019-03-30	Philadelphia, Pennsylvania, USA	Middleweight	Blue	0.717351	0.282649
9	Blue	Cub Swanson	Frankie Edgar	2018-04-21	Atlantic City, New Jersey, USA	Featherweight	Blue	0.501446	0.498554
10	Blue	Boston Salmon	Khalid Taha	2019-04-13	Atlanta, Georgia, USA	Bantamweight	Blue	0.712788	0.287212
11	Blue	Dan Henderson	Rashad Evans	2013-06-15	Winnipeg, Manitoba, Canada	Light Heavyweight	Blue	0.745946	0.254054
12	Blue	Jared Hamman	Michael Kuiper	2012-08-11	Denver, Colorado, USA	Middleweight	Red	0.303861	0.696139
13	Red	Cezar Ferreira	Daniel Sarafian	2013-11-09	Goiania, Goias, Brazil	Middleweight	Red	0.240448	0.759552
14	Red	Denis Kang	Xavier Foupa-Pokam	2009-04-18	Montreal, Quebec, Canada	Middleweight	Red	0.330089	0.669911

Fonte: Criado pelo autor

Filtrando probabilidades superiores a 70%, conseguimos obter taxas de acerto de até 75%. Já com o filtro em 85%, as taxas de acerto chegam a 88%. Esses resultados foram extremamente satisfatórios.

4.4 Avaliando lutas recentes, que não constam na base

Pesquisando no site **ufcstats.com**, montamos um *dataframe* novo com 10 lutas ocorridas em 2020 e 2021, para testar as previsões do modelo final. Nessa amostra, obtivemos 70% de acerto nas previsões.

Figura 23: Tabela de predição com probabilidades – Lutas recentes

	Winner	R_fighter	B_fighter	date	location	weight_class	Resultado_Previsto	Prob_Win_Blue	Prob_Win_Red
0	Red	Dan Ige	Gavin Tucker	2021-03-13	Las Vegas, Nevada, USA	Featherweight	Red	0.237969	0.762031
1	Blue	Thiago Santos	Aleksandar Rakic	2021-03-06	Las Vegas, Nevada, USA	LightHeavyweight	Blue	0.708747	0.291253
2	Red	Arman Tsarukyan	Matt Frevola	2021-01-23	Abu Dhabi, Abu Dhabi, United Arab Emirates	Lightweight	Red	0.291771	0.708229
3	Blue	Bobby Green	Thiago Moises	2020-10-31	Las Vegas, Nevada, USA	Lightweight	Blue	0.734127	0.265873
4	Blue	Katlyn Chookagian	Jessica Andrade	2020-10-17	Abu Dhabi, Abu Dhabi, United Arab Emirates	WomenFlyweight	Blue	0.773346	0.226654
5	Red	Loma Lookboonmee	Jinh Yu Frey	2020-10-03	Abu Dhabi, Abu Dhabi, United Arab Emirates	WomenStrawweight	Blue	0.561870	0.438130
6	Blue	Alessio Di Chirico	Zak Cummings	2020-08-29	Las Vegas, Nevada, USA	Middleweight	Blue	0.723541	0.276459
7	Red	Petr Yan	Jose Aldo	2020-07-11	Abu Dhabi, Abu Dhabi, United Arab Emirates	Bantamweight	Blue	0.660775	0.339225
8	Red	Vicente Luque	Niko Price	2020-05-09	Jacksonville, Florida, USA	Welterweight	Red	0.363608	0.636392
9	Red	Amanda Ribas	Randa Markos	2020-03-14	Brasilia, Distrito Federal, Brazil	WomenStrawweight	Blue	0.704259	0.295741

Fonte: Criado pelo autor

4.5 Modelagem por categoria de peso

Executando o modelo final sobre o *dataset* dividido por categorias de peso, os resultados não foram expressivos.

Figura 24: Previsão por categorias de peso

Categoria de Peso	Precisão
Flyweight	62%
Bantamweight	46%
Featherweight	55%
Lightweight	56%
Welterweight	60%
Middleweight	48%
Light Heavyweight	57%
Heavyweight	55%
Women's Strawweight	55%
Women's Flyweight	43%
Women's Bantamweight	42%
Women's Featherweight	50%

Fonte: Criado pelo autor

Acreditamos que o resultado ruim se deve ao fato de termos uma grande quantidade de categorias, então após a divisão do *dataset,* cada uma delas ficou com uma quantidade registros bastante reduzida.

Figura 25: Quantidade de registros por categoria de peso

	weight_class	Qtd
0	Bantamweight	340
1	Featherweight	411
2	Flyweight	182
3	Heavyweight	455
4	Light Heavyweight	474
5	Lightweight	894
6	Middleweight	681
7	Welterweight	914
8	Women's Bantamweight	92
9	Women's Featherweight	8
10	Women's Flyweight	32
11	Women's Strawweight	106

Fonte: Criado pelo autor

5 CONCLUSÃO

Desde a criação do UFC, em 1993, o evento passou por muitas mudanças para conseguir se popularizar e contornar as críticas, se transformando em um dos espetáculos mais atraentes do século 21. A empolgante modalidade mostrou ao mundo um mercado muito promissor. Pessoas lotam arenas, ingressos são esgotados em questão de minutos e os números de pay-per-view batem recordes.

Dado o crescente número de espectadores, também cresceu a quantidade de apostas nos resultados das lutas, passando esportes tradicionais como o boxe. Com o objetivo de explorar esse mercado, utilizamos algoritmos de aprendizado de máquinas em linguagem *Python* e conseguimos criar um modelo de previsão com resultados bastante satisfatórios.

Com o modelo final testado, foi possível estimar os resultados de lutas tendo como entrada atributos físicos e lutas passadas dos competidores. Foi possível prever resultados com uma acurácia de 62%, chegando a superar os 88% de taxa de acerto quando considerados os resultados com probabilidades superiores a 85%. O modelo final também foi validado utilizando dados de lutas recentes que não estavam presentes no *dataset* original, alcançando uma taxa de acertos de 70%.

Durante o desenvolvimento do trabalho, tentamos algumas abordagens não muito convencionais, que trouxeram resultados bastante positivos. Durante a fase de tratamento dos dados, criamos um modelo de regressão para prever a envergadura (em substituição à estimativa por mediana, por exemplo). E no modelo final tivemos a idéia de inverter as cores das luvas em algumas lutas para tornar a base balanceada (em vez de remover registros ou gerar dados artificias, por exemplo), conseguindo eliminar qualquer tendência de vitória do vermelho, mantendo a quantidade de registros inalterada. E a melhor parte é que essas propostas refletiram em resultados melhores.

Referente a realização de apostas, acreditamos que nosso modelo final possa ser utilizado como um bom balizador para tomada de decisão, principalmente quando a probabilidade de vitória do desafiante for bem alta. Ainda assim, cabe ressaltar que sempre haverá um risco envolvido e que a decisão de apostar é individual e responsabilidade única de quem decidir apostar.

6 REFERÊNCIAS

PEREIRA, Bernard. **Análise de características mais influentes em lutas do UFC**. Monografia (Bacharelado em Estatística) – Escola Nacional de Ciências Estatísticas, Instituto Brasileiro de Geografia e Estatística. Rio de Janeiro, 2018.

TAVARES, Henrique. **O SUCESSO DO UFC E SEU PAPEL COMO INFLUENCIADOR DA MODALIDADE MMA EM BRASÍLIA**. Monografia (Bacharelado em Publicidade e Propaganda) – Centro Universitário de Brasília. Brasília, 2012.

TOLENTINO, Volney. **Novas parcerias compõem estratégia para crescimento do UFC**. Cebola Verde, 2019. Disponível em: <https://cebolaverde.com.br/esportes/novas-parcerias-compoem-estrategia-para-crescimento-do-ufc/>. Acesso em 20 de dezembro, 2020.

AGOSTINI, Tiago. **O Esporte Número 2 do Brasil?** Rolling Stone, 2011. Disponível em: <https://rollingstone.uol.com.br/edicao/59/o-esporte-numero-2-do-brasil/>. Acesso em: 20 de dezembro, 2020.

UFC-Fight historical data from 1993 to 2019. Kaggle, 2020. Disponível em: <https://www.kaggle.com/rajeevw/ufcdata >. Acesso em 08 de dezembro de 2020.

Stats | UFC. UFC Stats, 2020. Disponível em: <http://www.ufcstats.com/>. Acesso em 08 de dezembro de 2020.

Como o MMA evoluiu da 'brutalidade' se tornou um negócio mais valioso que o Real Madrid. G1, 2016. Disponível em <http://g1.globo.com/economia/negocios/noticia/2016/07/como-o-mma-evoluiu-da-brutalidade-se-tornou-um-negocio-mais-valioso-que-o-real-madrid.html>. Acesso em 21 de dezembro de 2020.

UFC NETWORK® DISPONÍVEL EM 20 PAÍSES DA AMÉRICA LATINA A PARTIR DE 1º DE SETEMBRO. UFC, 2021. Disponível em: <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>. Acesso em 10 de março de 2021.