Análise exploratória

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.

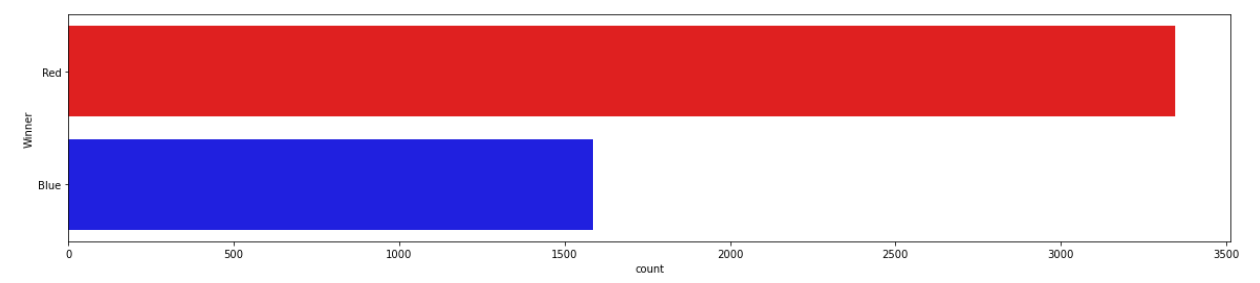
Analisando o dataset de lutas, observamos alguns campos vazios e os substituímos por “Nan”. Também convertemos os campos de data para “datetime” e arredondamos os campos numéricos

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

Sabe-se que 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 nomeclatura 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.

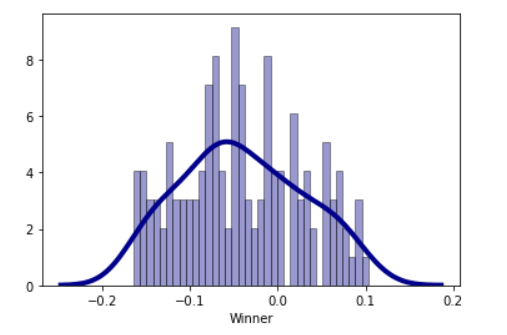
Uma característica marcante do UFC é que o lutador favorito sempre utiliza a cor vermelha e o desafiante fica com a cor azul. Diante deste informação, espera-se que o favoritismo seja do vermelho. A base comprova isso com uma taxa de vitórias do vermelho igual a 67,85%. Esse valor é o nosso baseline.



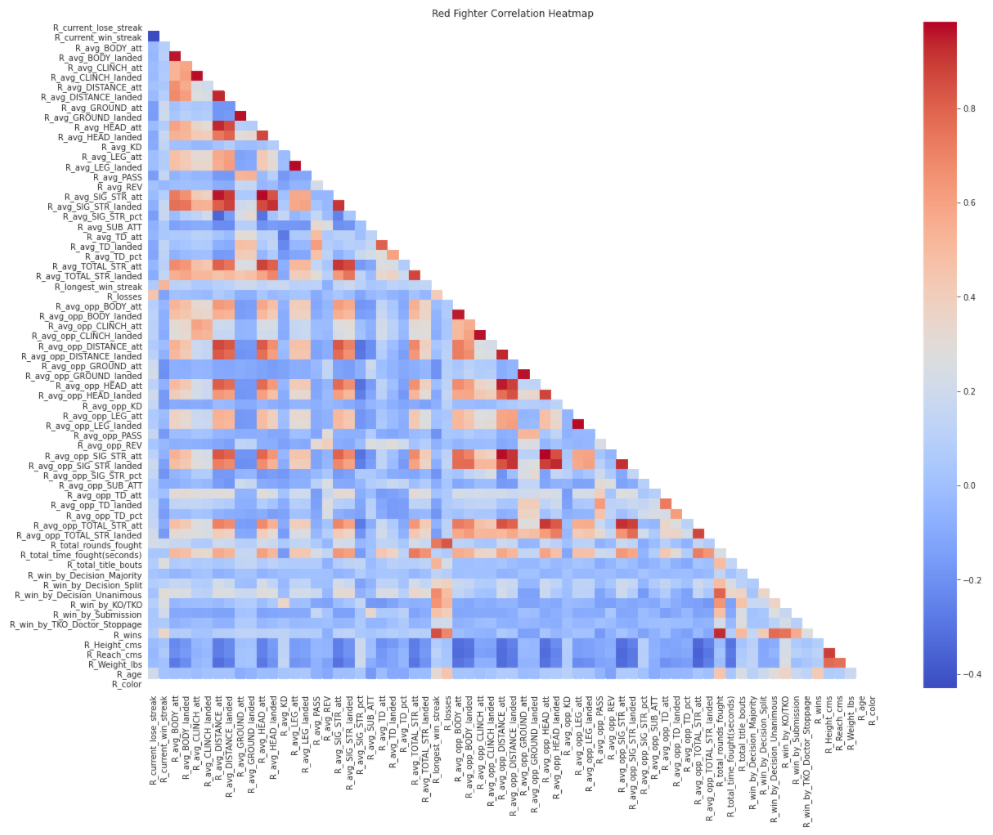
Avançando na análise do dataset, fizemos alguns gráficos de correlação. Cada quadrado mostra a correlação entre as variáveis ​​em cada eixo. Valores mais próximos de zero significam que não há tendência linear entre as duas variáveis. Quanto mais próxima de 1 a correlação é, mais positivamente correlacionados 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.

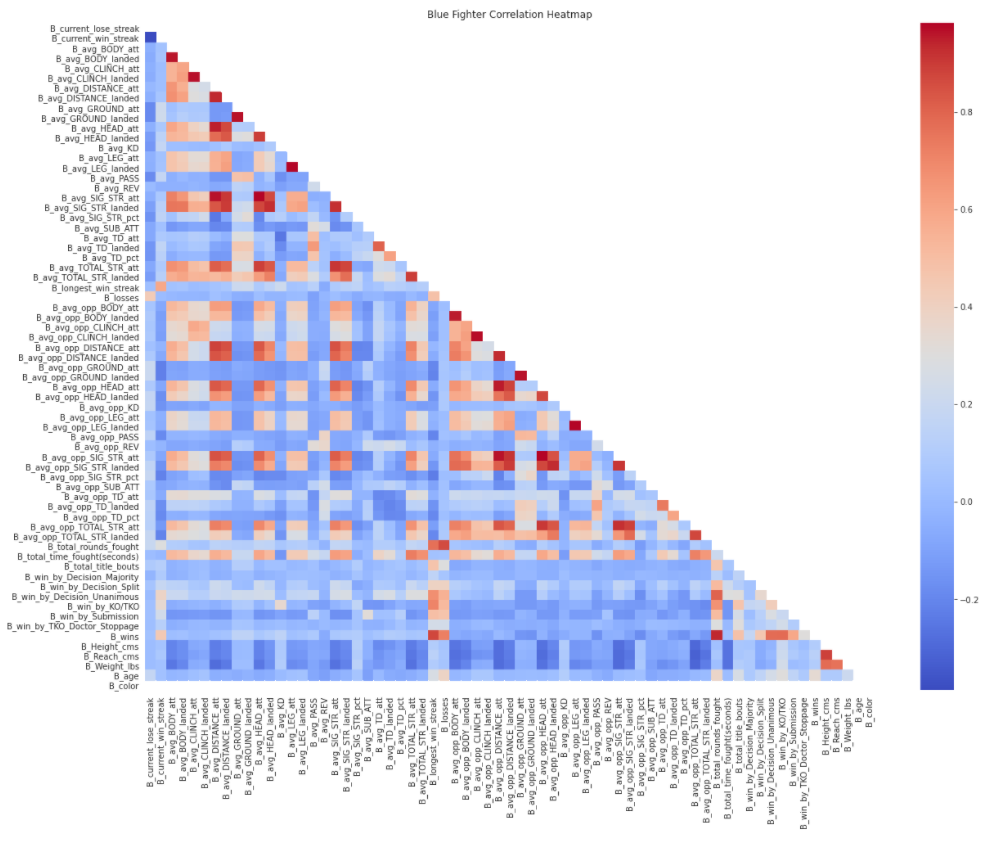
Através dessa análise, podemos identificar pares de variáveis com correlação alta e avaliar a possibilidade de manter só uma delas no modelo, para evitar redundâncias desnecessárias.

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.



Mapas de calor com as variáveis de cada lutador:



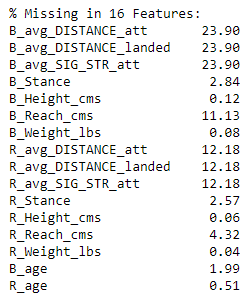


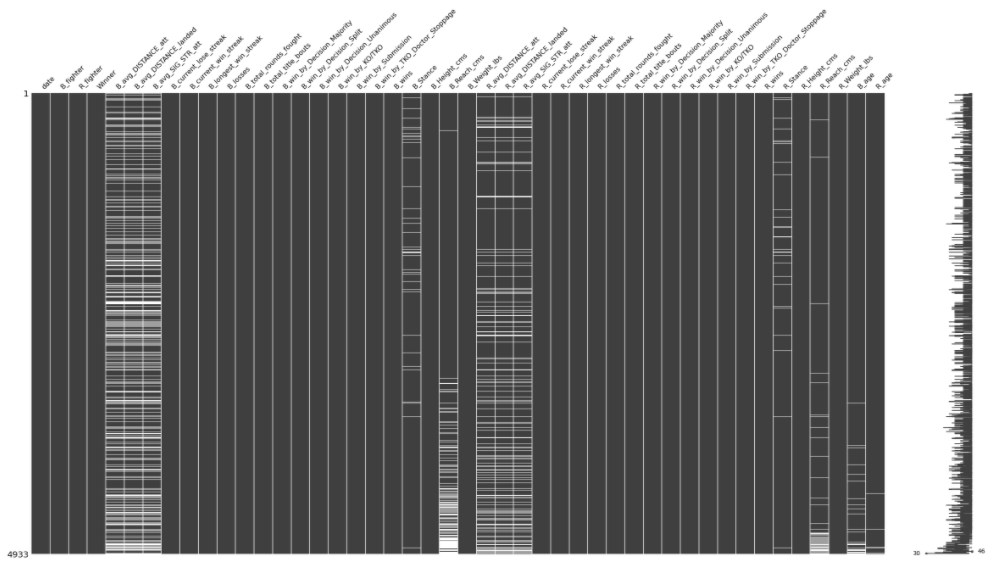
Analisando os mapas de calor e o significado de cada variável, selecionamos as variáveis que julgamos mais relevantes e as que classificamos como importantes, eliminando redundâncias:

['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:





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

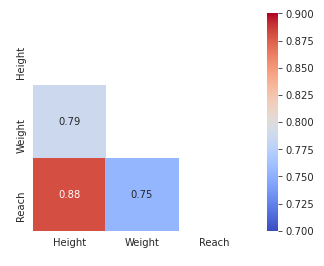
A próxima variável com ausência elevada é a envergadura. Neste caso, não podemos excluir as colunas do dataframe, porque esse é um dos atributos físicos mais importantes de qualquer lutador. Também optamos por não excluir da base os registros sem envergadura, por representarem um percentual importante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

B\_Reach\_cms 11.13

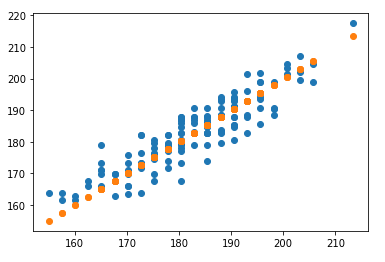
R\_Reach\_cms 4.32

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.



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





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

R\_Stance\_Open Stance

R\_Stance\_Orthodox

R\_Stance\_Sideways

R\_Stance\_Southpaw

R\_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

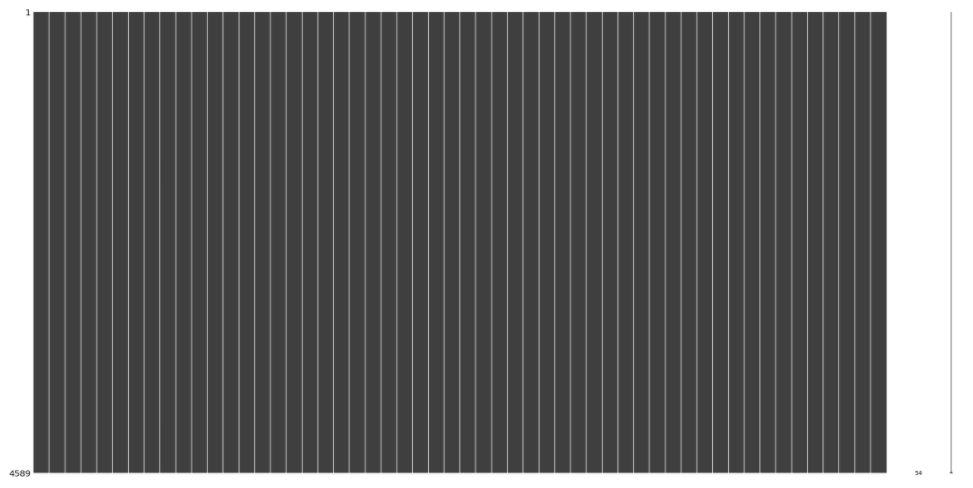
B\_age 1.99

R\_age 0.51

B\_Stance 2.84

R\_Stance 2.57

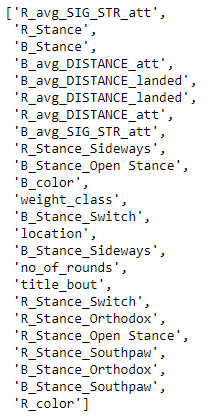
Chegando ao dataset tratado, sem nenhum valor ausente:



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']

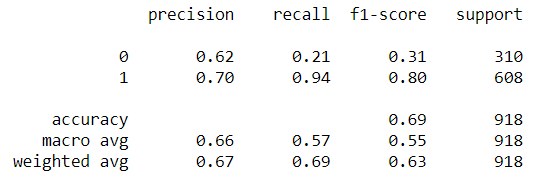
Comparando essa lista final de colunas, com a lista de variáveis que consideramos importantes na análise exploratória inicial e gráficos de correlação, ficamos muito satisfeitos com o resultado, pois as principais diferenças observadas são as colunas que decidimos excluir do dataframe e as colunas categóricas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



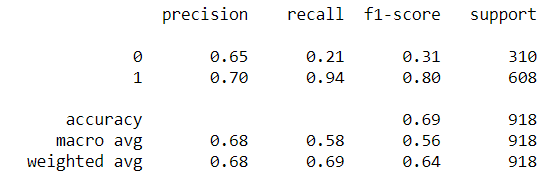
INICIANDO PREVISÃO DE RESULTADOS DE LUTAS

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.

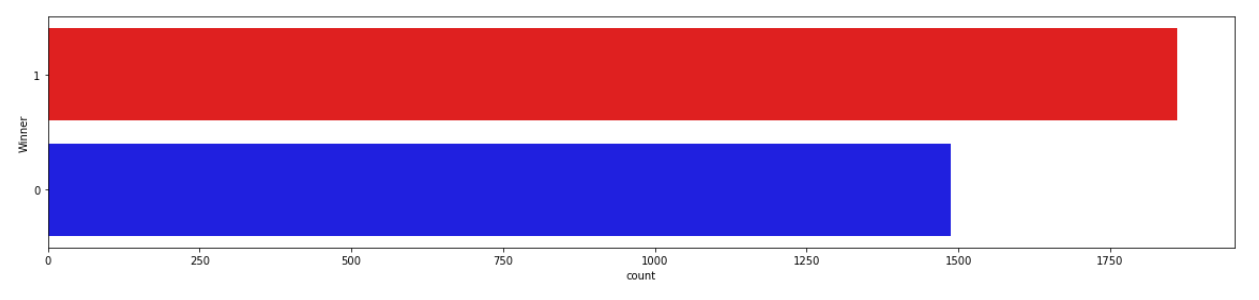
Floresta Aleatória:



Redes Neurais:

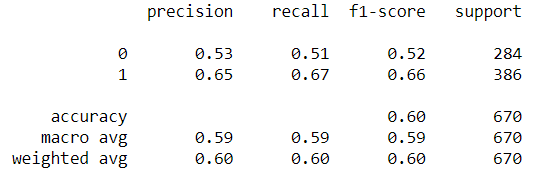


Então partimos para uma abordagem diferente, onde tentamos forçar o balanceamento da base, eliminando 40% dos registros de vitórias do vermelho, de maneira aleatória.

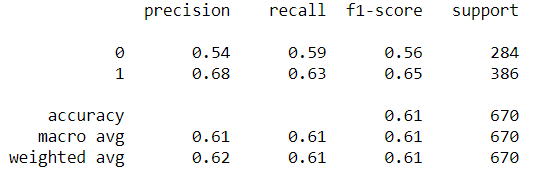


Nessa abordagem, foram eliminados cerca de 1600 registros da base (redução considerável). E mesmo forçando o balanceamento, ainda assim vemos um certo desequilíbrio nos valores de recall e f1-score.

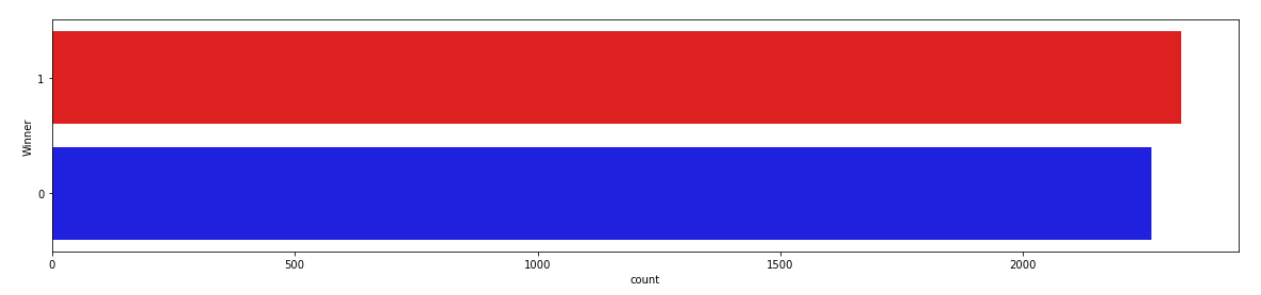
XGBoost:



Redes Neurais:

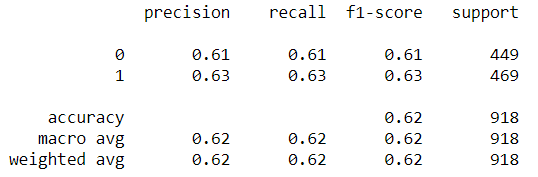


Decidimos então realizar o balanceamento da base de uma maneira diferente, em vez de apagar 40% das vitórias do vermelho, invertemos as cores em 25% das lutas vencidas pelo vermelho. Assim obtivemos um dataset balanceado, sem perda de registros:



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:



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.