Projeto de Graduação

DEPARTAMENTO DE ENGENHARIA ELÉTRICA

19 de dezembro de 2022

O uso de redes neurais para a melhora de desempenho de times de basket na NBB

Lucas Vital Alves da Silva

PUC RIO

Projeto de Graduação

DEPARTAMENTO DE ENGENHARIA ELÉTRICA

O uso de redes neurais para a melhora de desempenho de times de basket na NBB

Aluno: Lucas Vital Alves da Silva

Orientador: Marley Maria B Rebuzzi Vellasco

Trabalho apresentado com requisito parcial à conclusão do curso de Engenharia Elétrica na Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, Brasil.

Agradecimentos

Agradeço sobretudo à Deus por ter me permitido chegar até aqui. Não obstante, agradeço aos meus pais Rosana Vital da Silva e Júlio César Alves da Silva por terem me dado todo apoio e condições para poder dar continuidade aos meus estudos. Por fim, dedico aos meus agradecimentos a professora Marley Vellasco e ao aluno mestrando Eduardo Veras por terem me dado a honra de poder participar desse projeto e por toda orientação e apoio que me foi concedido ao longo dele.

Resumo

Nos últimos anos, houveram diversos avanços tecnológicos que trouxeram inúmeras inovações que provocaram diversas transformações nos mais diversos aspectos da humanidade. Na atualidade, vive-se uma nova fase de revolução tecnológica provocada pelo ganho tanto de capacidade de armazenar dados como também de transmiti-los e processá-los, o que faz possível obter informações valiosas que serão cruciais no processo de tomada de decisões nos mais diversos contextos e aspectos. Não obstante, vivemos num mundo cada vez mais ”data driven”, ou seja, orientado aos dados, e quanto mais uma empresa, instituição ou indivíduo for orientado aos dados em seu processo de tomada de decisão, maior serão as suas probabilidades de sucesso. Portanto, a procura por sistemas que sejam capazes de extrair ”insights” dos dados cada vez mais tem se tornado maior.

O mundo dos esportes também não poderia ficar de fora de toda essa revolução que atualmente está acontecendo, cada vez mais cresce a importância tanto da coleta como processamento de dados que podem gerar ”insights” que permitam a tomadas de decisões que podem ser essenciais nos mais diversos aspectos do esporte. Auxílio na tomada de decisão num momento crucial de um jogo, auxílio na elaboração de treinamentos que maximizem o máximo possível o desempenho de atletas e até mesmo prevenir lesões, dentre outros.

Dentre as muitas possibilidades, uma é o uso de algoritmos de inteligência artificial para determinação do melhor time numa partida de Basquete da NBB. A ideia principal é conseguir avaliar o desempenho dos times que estão se enfrentando numa partida através de uma determinada janela de tempo avaliando a composição dos jogadores em quadra, as variáveis relacionadas a partida e a cada jogador individualmente além de extrair características individuais de cada jogador. Assim, consegue-se saber qual seria a composição ideal de jogadores dentre os disponíveis para que um time maximize as suas chances de vitória numa eventual partida da NBB. Tal análise foi feita mais precisamente para o time de basquete do Flamengo. Deste modo, o objetivo deste projeto foi a criação de um sistema que pudesse prever qual seria a composição de jogadores disponíveis do Flamengo que mais provavelmente se sairia contra um time adversário e assim maximizasse as chances do Flamengo vencer uma partida.

Nesse ínterim, realizou-se um procedimento de análise de dados para selecionar variáveis, para fazer a clusterização de jogadores, criação de variáveis e métricas que pudessem ser úteis na análise, e finalmente o desenvolvimento e treinamento de um sistema de aprendizado de máquina que irá trabalhar para determinar qual é a melhor composição de jogadores que maximizará as chances de vitória. Ao longo do processo diversas linhas de análise foram testadas e os procedimento serão expostos de forma mais detalhadas a seguir.

Palavras-chave: Flamengo, NBB, Esportes, Basquete, redes neurais

The use of neural networks to improve the performance of basketball teams in the Brazilian national basketball league

Abstract

In recent years, there have been several technological advances that have brought numerous innovations that have caused several transformations in the most diverse aspects of humanity. Currently, we are experiencing a new phase of technological revolution caused by the gain in the ability to store data as well as to transmit and process them, which makes it possible to draw valuable conclusions that helped in the decision-making process in the most diverse contexts and aspects. The world of sports could not be left out of this revolution that is currently taking place, the importance of both collecting and processing data that can generate insights that allow decision-making that can be essential in the most diverse sports aspects. Assistance in decision-making at a crucial moment in a game, assistance in the preparation of training that maximizes athletes's performance as much as possible and even injuries prevetion, among others.

Among the many possibilities, one is the use of artificial intelligence algorithms to determine the best team in an NBB (Brasilian League of Basketball) Basketball match. The main idea is to be able to evaluate the performance of the teams that are facing each other in a match through a time window evaluating the composition of the players on the court, analyze variables related to the match and to each player individually and at the same time extract individual characteristics of each player. Thus, it is possible to know what would be the ideal composition of players among those available so that the team as a whole does better than the opponent in a match as a whole. Such an analysis was made more precisely for the Flamengo basketball team, so the objective was to create a system that could predict which would be the composition of available Flamengo players that would be better against an opposing team.

In this manner, a data analysis procedure was carried out to select variables, to cluster players, create variables and metrics that could be useful in the analysis, and finally the development and training of two neural networks that will work together to determine what is the best composition of players. Throughout the process, several lines of analysis were tested and the procedures will be explained in more detail below.

Keywords: Flamengo, NBB, Basketball, Neural Networks, Sports

Sumário

Lista de figuras

Lista de tabelas

1 Introdução

a Motivação

A cada nova partida de Basquete, o técnico de um time tem como algumas de suas atribuições determinar qual o melhor time para enfrentar um determinado adversário, a posição dos jogadores, assim como trabalhar possíveis jogadas considerando as características de cada jogador e realizar as devidas substituições considerando o contexto de um determinado momento do jogo. Obviamente considerando a experiência do técnico, ele possui a habilidade de analisar o time adversário em partidas anteriores e determinar quais jogadores irão compor um bom quinteto contra aquele time adversário. Durante a partida ele também é capaz de determinar se há a necessidade de uma mudança para melhorar o time, caso seu desempenho não seja satisfatório. Entretanto, por melhor que seja um técnico, ele não será capaz de avaliar todas as variáveis de um jogo, pois como qualquer ser humano, ele possui uma capacidade limitada para a absorver o processar os dados, sendo assim, ele será capaz de determinar um bom time, mas nada garante que será capaz de determinar o melhor time para todo e qualquer adversário e toda e qualquer situação de jogo, o que abre margem para erros a serem cometidos e que podem comprometer a vitória de sua equipe.

É justamente nesse contexto que surge a necessidade de algo que auxilie o técnico, algo que seja capaz de processar uma massiva quantidade de dados, avaliar uma grande quantidade de variáveis e assim determinar qual o melhor time contra um determinado adversário numa determinada situação de jogo, é justamente aí que um modelo de aprendizado de máquina pode ser aplicado para ajudar a determinar a melhor formação de quinteto dado um adversário e contexto.

b Objetivo

Esse trabalho possui os seguintes capítulos descritos abaixo:

No capítulo 1, foi apresentado a introdução a este projeto, suas motivações, objetivo e a organização de todo o trabalho aqui desenvolvido.

No capítulo 2 descreve-se os conceitos chave para entender o funcionamento de todo o sistema e faz-se uma apresentação do banco de dados utilizado para o desenvolvimento do projeto. Neste capítulo aborda-se o conceito do que é uma rede neural, dos elementos associados a ela e dinâmica de funcionamento, conceitos relacionados a clusterização, apresentação sobre o que é análise exploratória de variáveis, seleção de variáveis e engenharia de recursos ("Explanatory Data Analysis", "Variable Selection and feature engineering") e faz-se uma breve apresentação do banco de dados utilizado no desenvolvimento deste projeto.

No capítulo 3 se apresenta todo o processo de desenvolvimento até se chegar no modelo atual, isso inclui desde a seleção das variáveis do banco de dados a serem utilizadas no treinamento do sistema, os tratamentos nos dados utilizados, a determinação da estrutura do sistema, a determinação da melhor topologia de rede e possíveis e finalmente as abordagens utilizadas na hora de analisar e utiliza os dados para o treinamento do sistema.

| No capítulo 4 serão apresentados os resultados finais para cada uma das linhas de análise utilizadas. Para cada uma das abordagens será feito uma análise de desempenho para que assim por fim se determina aquela que melhor atendeu ao critério buscado, em especial na última abordagem, apresentou-se a tentativa de utilização do KNN como estágio 1, e a comparação dele com a topologia de rede neural utilizada.

Finalmente no capítulo 5 apresenta-se as conclusões dos resultados obtidos, as dificuldades que foram encontradas no processo, assim como a perspectiva de continuidade para trabalhos futuros e por fim as considerações finais

2 Apresentação de conceitos chave e metodologias utilizadas

a Redes Neurais

Redes neurais constitui-se num algoritmo computacional que tenta emular o funcionamento do cérebro humano. O cérebro humano é constituído de neurônios(células) que se conectam e interagem entre si. A depender de interações anteriores cada neurônio tem uma espécie de memória o que gera como resultado o aprendizado de todo o conjunto a partir das informações recebidas e processadas recebidas por ele.

Analogicamente, é assim que funciona uma rede neural artificial:

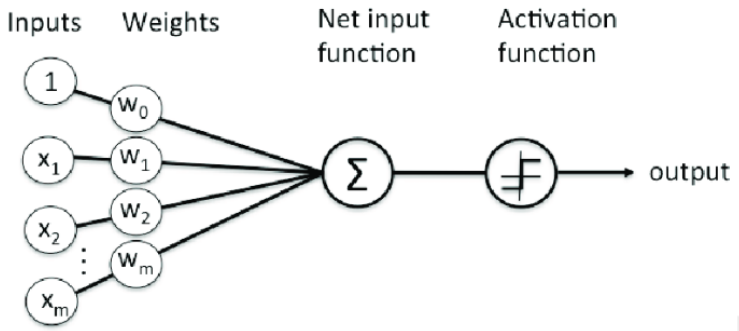

Figura 1: Funcionamento de um neurônio artificial

Acima tem-se ilustrado uma célula assim como também o modelo matemático de como essa célula gera a saída a partir das entradas recebidas. As entradas são combinadas linearmente por pesos, o resultado da soma é somado a um termo denominado bias, ao qual é responsável por determinar o limiar de ativação da célula, isto é, se ela vai gerar uma saída não nula ou não desprezível. Por fim, o resultado dessa combinação é usado como entrada da função de ativação que finalmente gera o resultado final.

Matematicamente falando temos:

$$x = \Sigma_{i=1}^{n} x_i.w_i + bias$$

Fórmula 1: Geração da entrada da função de ativação numa célula da rede

$$output = y = F(x)$$

Fórmula 2: Geração da saída numa célula da rede a partir da função de ativação

Desse modo uma rede neural nada mais é do que o conjunto dessas células trabalhando em conjunto. Tais células são agrupadas em camadas subsequentes, e cada célula de uma camada interage com as células da camada posterior até se gerar um vetor de saída desejada. Abaixo tem-se o exemplo de uma rede neural com 4 entradas, 2 camadas escondidas e 2 saídas:

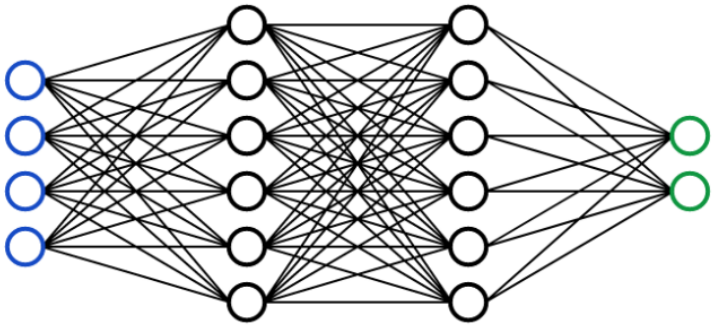

Figura 2: Exemplo de uma rede neural

b O treinamento de um rede neural

Para que uma rede neural seja capaz de prever algo é necessário que ela seja previamente treinada. Neste tópico abordaremos como funciona o treinamento de uma rede neural.
Como qualquer algoritmo de aprendizado de máquina supervisionado, uma rede neural é treinada a partir de um "dataset". Um conjunto de instâncias desse "dataset" é utilizada para o treinamento, parte dos atributos dessa instância são as entradas da rede e outra parte são as saídas dessa rede. Com isso a primeira etapa de treinamento consiste na propagação das entradas ao longo da rede até gerar a saída, a essa etapa chama-se de "foward propagation":

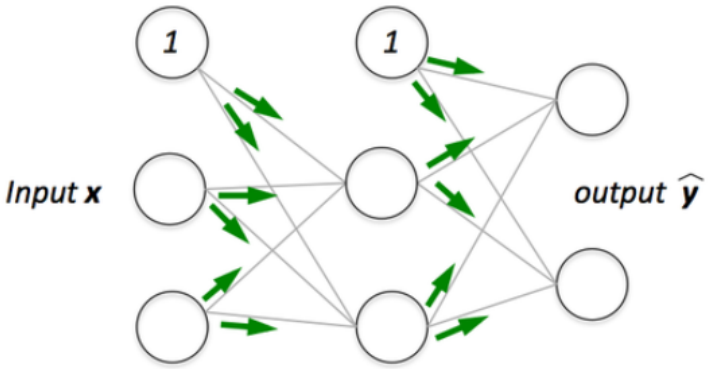

Figura 3: Ilustração forward propagation

Comparando-se a saída gerada($\vec{y}$) com a saídas contidas no atributo da instância, calcula-se uma função erro, a mais comum utilizada nesse tipo de treinamento é a "Mean Squared erro":

$$MSE = \frac{1}{n}\Sigma_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2$$

Fórmula 3: Cálculo Mean-Squared-Error

A partir do erro calculado na função de erro acima, prossegue-se ao procedimento de ajuste dos pesos da rede (parâmetros da rede), pelo procedimento de backpropagation. Tal procedimento é baseado no algoritmo ”Gradient Descend”, que por sua vez calcula o gradiente da função MSE e a partir desse gradiente multiplicado pelo parâmetro de taxa de aprendizado ()ajusta-se cada um dos pesos:

$$\nabla_{\omega(t)} = \begin{bmatrix} \frac{\partial MSE(W^{(t)})}{\partial \omega_1^{(t)}} \\ \frac{\partial MSE(W^{(t)})}{\partial \omega_2^{(t)}} \\ . \\ . \\ . \\ \frac{\partial MSE(W_j^{(t)})}{\partial \omega_n^{(t)}} \end{bmatrix}$$

Fórmula 4: Cálculo gradiente dos pesos

Fórmula 5: Cálculo gradiente do vetor de pesos da camada j de uma rede neural

$$W_{i,j}^{(t+1)} = W_{i,j}^{(t)} - \eta . \nabla MSE(W_j^{(t)})$$

Fórmula 6: Cálculo atualização de peso na camada j da célula i

Logo em seguida todo o procedimento é repetido novamente por vários ciclos denominado época, a quantidade de época é determinada previamente como um parâmetro do sistema. Com isso o treinamento de uma rede neural consiste num problema de otimização da função MSE(W), ou seja, de minimizar seu valor ao máximo possível dentro de uma quantidade de época pré-definida. Vale a pena ressaltar que o algoritmo “Gradient Decent” tem algumas variações, existe uma variação onde se toma todo o dataset de treino para realizar os procedimentos descritos acima, a esse método chama-se “ Batch Gradient Decent”. Em outra variação o processo interativo de treinamento toma randomicamente registros no dataset, a esse método chama-se de “Stochastisc Gradient Decent”. Devido ao fato do primeiro método ser fortemente custoso computacionalmente em “datasets” grandes, tomou-se o segundo método para otimizar a função erro durante o treinamento das redes utilizadas no projeto. Desse modo ao término das interações de treinamento tem-se uma rede neural treinada pronta para ser testada.

c K-nearest Neighborhood

Também conhecido pela abreviação KNN, esse algoritmo tem por principal característica assumir que a semelhança entre instâncias num dataset é medida pela proximidade entre elas, ou seja, quanto menor a distância entre elas mais semelhantes elas são entre si.

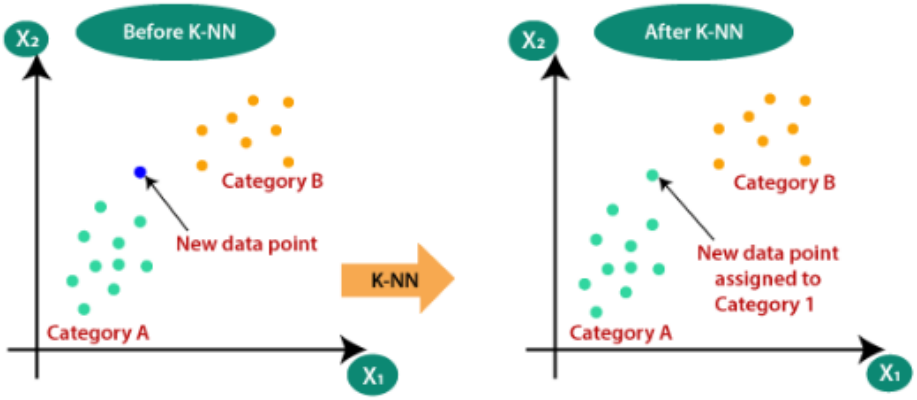

Figura 4: Ilustração algoritmo KNN

Tal algoritmo trata-se de um algoritmo de aprendizado supervisionado classificador. Portanto cada instância no dataset já tem sua classificação previamente informada. Sendo assim pode-se medir o índice de acerto dos algoritmos durantes o treinamento e teste. O KNN possui um funcionamento simples, primeiramente define-se o parâmetro K, esse parâmetro define as K instância mais próximas da instância a ser classifica a partir de uma determinada distância. Existem diversas possibilidades para o cálculo dessas distâncias: elas podem ser uma distância euclidiana, Manhattan, distância de Hamming, etc.

Para fins desse projeto, utilizou-se da distância euclidiana no teste feito com o KNN:

$$d(x,y) = \sqrt{(\Sigma_{i=1}^{n}(y_i - x_i)^2}$$

Fórmula 7: Cálculo da distância euclidiana

Sendo assim, uma vez tendo o parâmetro K definido e um dos métodos de cálculo de distância acima escolhido, o algoritmo consegue achar as K mais próximas, cada um pertencente a sua correspondente classe a que mais aparece dentro do círculo das K mais próximas vai determinar qual a classe do dado a ser classificado:

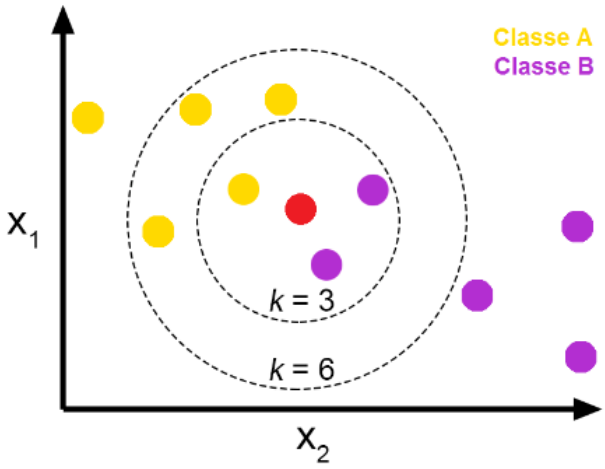

Figura 5: Funcionamento algoritmo KNN

d Apresentação do Dataset

Para realização dos experimentos nesse projeto utilizou-se de um banco de dados fornecido pelo Clube de regatas Flamengo ao qual contém todas as posses de bolas ocorridas entre todas as partidas ocorridas do ano de 2013 ao ano de 2019. O banco de dados totaliza cerca de 273 mil registros e um total de 152 variáveis, das quais algumas estão relacionadas aos parâmetros da partida tal como bloqueios, assistências, roubos de bola, tempo de jogo etc. Algumas outras são parâmetro relacionados a quais jogadores estão em quadra, suas performances individuais de uma classificação dos jogadores em quadra tanto do time de casa quanto do time de fora em clusters previamente determinados por quem gerou o dataset. Ao conjunto de todos os jogadores em quadra, dá-se o nome de ”lineup”.

3 Desenvolvimento do sistema e apresentação de seu funcionamento em sua versão final

a Método de seleção de variáveis

Devido a questões de custo computacionais e performance do modelo, nem todas as variáveis foram utilizadas caso contrário ou poder-se-ia ter um modelo inviável computacionalmente de ser executado e até mesmo com performance comprometida por um ”overfitting”, uma consequência do excesso de variáveis no treinamento de um modelo que faz com que o modelo se desempenho muito bem nos treinos e um desempenho péssimo no teste, ou seja, uma péssima generalização de seu aprendizado.

Desta forma, das 152 variáveis, 61 foram selecionadas de forma empírica considerando os conceitos envolvidos num jogo de basquete. As variáveis selecionadas juntamente com suas respectivas descrições podem ser vista na imagem abaixo:

Variável de Entrada	Descrição
period	Um jogo de basquete é dividido em 4 períodos, essa variável específica qual o período atual.
match_elapsed_minutes	Tempo decorrido de jogo no período
home_score	Pontuação time de casa na partida.
away_score	Pontuação time de Fora na partida.
home_score_difference	Diferença entre a pontuação do time de casa e a pontuação do time de fora numa partida
home_period_fouls	Faltas cometidas pelo time de casa no período atual
away_period_fouls	Faltas cometidas pelo time de fora no período atual
home_quintet_points	Pontos feitos enquanto os 5 jogadores do quinteto de casa estavam em campo, até o momento do início da possession.
home_quintet_assists	Idem para assistências
home_quintet_rebounds	Idem para rebotes
home_quintet_steals	Idem para roubadas
home_quintet_blocks	Idem para bloqueio/toco home_quintet_missed_throws - Idem para número de arremessos perdidos excluindo lance livre
home_quintet_missed_throws	Idem para lances não convertidos
home_quintet_missed_free_throws	Idem para lances livres não convertidos
home_quintet_fouls	Idem para faltas
home_quintet_turnovers	Idem para turnovers
home_quintet_eff_per_min	Mesma métrica de eficiência por minuto, mas calculada considerando os dados do quinteto de casa como um todo
home_quintet_minutes	Número de minutos que o quinteto de casa esteve junto em quadra até o início da possession até o momento do início da possession.
away_quintet_points	Pontos feitos enquanto os 5 jogadores do quinteto de fora estavam em campo, até o momento do início da possession.
away_quintet_assists	Idem para assistências
away_quintet_rebounds	Idem para rebotes
away_quintet_steals	Idem para roubadas
away_quintet_blocks	Idem para bloqueio/toco home_quintet_missed_throws - Idem para número de arremessos perdidos excluindo lance livre
away_quintet_missed_throws	Idem para lances não convertidos
away_quintet_missed_free_throws	Idem para lances livres não convertidos
away_quintet_fouls	Idem para faltas
away_quintet_turnovers	Idem para turnovers
away_quintet_eff_per_min	Mesma métrica de eficiência por minuto, mas calculada considerando os dados do quinteto de casa como um todo
away_quintet_minutes	Número de minutos que o quinteto de casa esteve junto em quadra até o início da possession até o momento do início da possession.
home_cluster_1	Quantidade de jogadores do Cluster 1 que compõe o time de casa em quadra
home_cluster_2	Quantidade de jogadores do Cluster 2 que compõe o time de casa em quadra
home_cluster_3	Quantidade de jogadores do Cluster 3 que compõe o time de casa em quadra
home_cluster_4	Quantidade de jogadores do Cluster 4 que compõe o time de casa em quadra
home_cluster_5	Quantidade de jogadores do Cluster 5 que compõe o time de casa em quadra
home_cluster_6	Quantidade de jogadores do Cluster 6 que compõe o time de casa em quadra
home_cluster_7	Quantidade de jogadores do Cluster 7 que compõe o time de casa em quadra
home_cluster_8	Quantidade de jogadores do Cluster 8 que compõe o time de casa em quadra
home_cluster_9	Quantidade de jogadores do Cluster 9 que compõe o time de casa em quadra
home_cluster_10	Quantidade de jogadores do Cluster 10 que compõe o time de casa em quadra
home_cluster_11	Quantidade de jogadores do Cluster 11 que compõe o time de casa em quadra
home_cluster_12	Quantidade de jogadores do Cluster 12 que compõe o time de casa em quadra
home_cluster_13	Quantidade de jogadores do Cluster 13 que compõe o time de casa em quadra
home_cluster_14	Quantidade de jogadores do Cluster 14 que compõe o time de casa em quadra
home_cluster_15	Quantidade de jogadores do Cluster 15 que compõe o time de casa em quadra
home_cluster_16	Quantidade de jogadores do Cluster 16 que compõe o time de casa em quadra
away_cluster_1	Quantidade de jogadores do Cluster 1 que compõe o time de fora em quadra
away_cluster_2	Quantidade de jogadores do Cluster 2 que compõe o time de fora em quadra
away_cluster_3	Quantidade de jogadores do Cluster 3 que compõe o time de fora em quadra
away_cluster_4	Quantidade de jogadores do Cluster 4 que compõe o time de fora em quadra
away_cluster_5	Quantidade de jogadores do Cluster 5 que compõe o time de fora em quadra
away_cluster_6	Quantidade de jogadores do Cluster 6 que compõe o time de fora em quadra
away_cluster_7	Quantidade de jogadores do Cluster 7 que compõe o time de fora em quadra
away_cluster_8	Quantidade de jogadores do Cluster 8 que compõe o time de fora em quadra
away_cluster_9	Quantidade de jogadores do Cluster 9 que compõe o time de fora em quadra
away_cluster_10	Quantidade de jogadores do Cluster 10 que compõe o time de fora em quadra
away_cluster_11	Quantidade de jogadores do Cluster 11 que compõe o time de fora em quadra
away_cluster_12	Quantidade de jogadores do Cluster 12 que compõe o time de fora em quadra
away_cluster_13	Quantidade de jogadores do Cluster 13 que compõe o time de fora em quadra
away_cluster_14	Quantidade de jogadores do Cluster 14 que compõe o time de fora em quadra

Tabela 1: Variáveis de entrada do sistema

b Desenvolimento do sistema

Como mencionado anteriormente, o principal objetivo deste projeto é o desenvolvimento de um modelo de "Machine Learning" capaz de prever qual seria o melhor time para enfrentar um determinado time adversário numa determinada partida em um dado momento. Para alcançar tal objetivo obteve-se um sistema dividido em 2 estágios, um estágio que diz se o time é melhor, igual ou inferior ao time adversário e outro que diz o quão melhor é o time de casa em relação ao adversário.

Evidentemente que essa foi a versão do sistema que até então fora desenvolvida, e até ter-se conjecturado essa versão houve uma série de procedimentos e testes a serem realizados até chegar esse ponto. Em sua primeira versão o sistema tentava somente prever quem tinha a posse de bola, o time de casa ou o time de fora. Posteriormente cogitou-se tentar prever o saldo de ponto de pontos do time de casa em relação ao time de fora, ou seja, prever a variável `home_result` e viu-se que o desempenho desse modelo de foi melhor que o do primeiro. Uma outra abordagem também se saio muito bem foi tentar prever se o time de casa empatava, ganhava ou perdia para o time adversário, através da classificação categórica do time em campo numa dada posse de bola por uma rede neural. Desse modo chegou-se ao modelo de 2 estágios, onde o estágio 1 classifica o time de casa em relação ao adversário e o estágio 2 mede o quão bom é esse time em relação ao seu adversário.

c Tratamento dos dados

O processo de melhora do modelo não foi linear, concomitantemente ao seu desenvolvimento outros procedimentos foram realizados no que diz respeito a análise e tratamento de dados. Primeiro fez-se uma análise exploratória de dados pra determinar as características estatísticas de uma determinada variável, e qual tratamento deverá ser feito de acordo a suas características
Pode-se citar como exemplo o tratamento da principal variável que se está tentando prever,home_result. Foi-se feita uma análise de distribuição de frequências de seus valores:

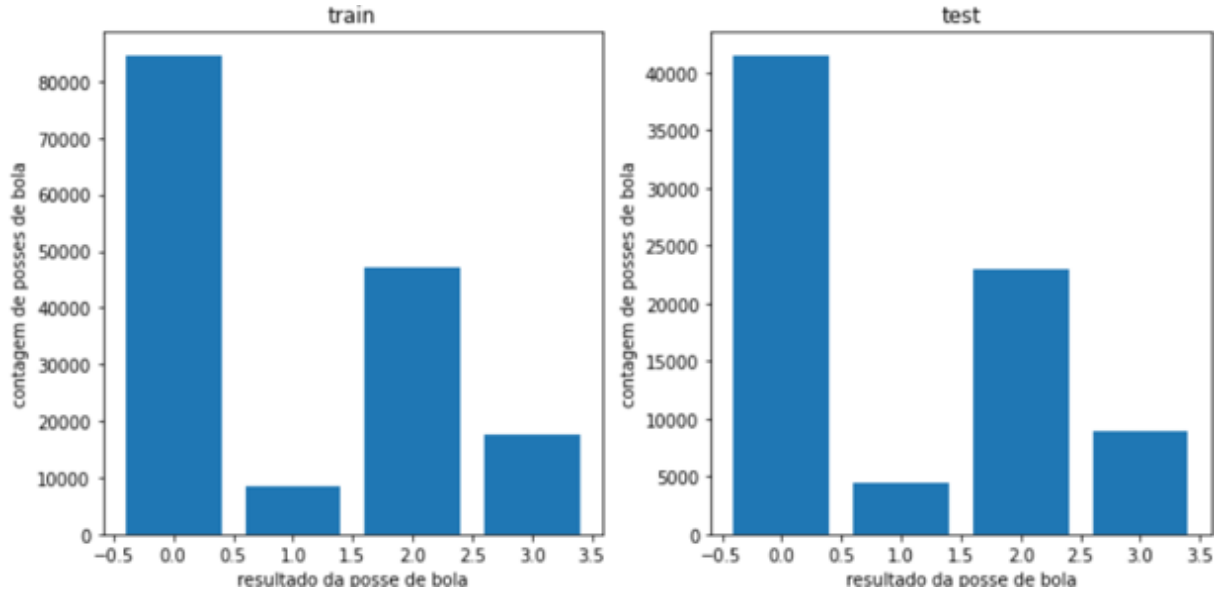

Figura 6: Analise de distribuição `home_result`

Com isso percebeu-se que era necessário fazer um rebalanceamento dos valores dessa variável com a finalidade de evitar vícios de predição na rede, dado que "0" é o valor mais frequente, a rede tenderia a prever 0 já que esse seria o provável de ocorrer.

O mesmo foi feito nas demais variáveis envolvidas, e dentre o procedimento que podemos citar são: análises comparativas de variáveis, análises de distribuição de uma variável, a eliminação de outliers, balanceamento, eliminação de variáveis e a aplicação de codificação para as variáveis categóricas.

d Determinação da topologia de rede neural utilizada

No que diz respeito a determinação da melhor topologia de rede, ou seja, a topologia em termos de quantas camadas escondidas e quantas células por camada, primeiro estabeleceu-se um total de 2 camadas escondidas para esta rede, a quantidade de célula em cada camada foi determinada por um algoritmo que instancia vária topologias de rede, analisa o desempenho de cada uma delas e com isso escolhe a topologia de melhor desempenho. Até então a melhor topologia foi com 60 célula na primeira camada escondida e 30 na segunda.

Por fim vale ressaltar que todas as mudanças aqui citadas ao longo do desenvolvimento do sistema foram feitas de forma incremental e posteriormente testadas, caso houvesse uma melhora significativa a alteração era mantida, caso não descartada. E assim foi implementada todas as mudanças no sistema até se chegar na versão atual.

e Treinamento do sistema

Como citado anteriormente ,em princípio, utilizou-se de duas redes neurais para os 2 estágios. Como todo modelo de "Machine Learning" e redes neurais precisam ser treinadas, ou seja, interativamente alimentadas até consumir toda parte do dataset destinado ao treino, isto é, o "Training Set".

Contudo vale ressaltar que antes de realizar tal procedimento é necessário realizar um tratamento nos dados de entrada antes de realizar o treinamento da rede. Para campos numéricos é necessário realizar uma normalização de valores devido ao fato de que esses dados originalmente estão em diferentes escalas o que distorce o aprendizado do modelo. Para o caso de dados categóricos algum tipo de codificação deve ser implementado. Muitos dados categóricos são simplesmente "strings". Devido ao fato de que o funcionamento das células de uma rede neural estar baseado em combinações lineares e funções matemática não faz sentido alimentar a rede com "strings", então a codificação se encarrega se transformar as categorias em dados numéricos. Dado ao fato dos campos escolhidos para o treinamento do modelo, as 43 variáveis de entrada, A normalização de valores. Para utilizada para os valores numéricos foi a função MinMax:

$$X_{normalizado} = \frac{X - X_{min}}{X_{max} - X_{min}} \times (a - b)$$

Fórmula 8: Cálculo da normalização Min-Máx

Já para os valores categóricos, considerou-se o conjunto de números interiores de 1 a 16 para representar cada um dos clusters de jogadores que compunham o "lineup".

1 Métodos de utilização do dataset no treinamento da rede

Outra abordagem a ser considerada é como as posses de jogo, os registros no dataset, serão considerados ao se treinar a rede. A princípio se conjecturou aquilo que seria o imediatamente mais óbvio, usar cada posse individualmente no treinamento da rede para que fosse possível prever a saída relacionada a cada um dos registros de posse de bola. A saída seria inicialmente a previsão se um time um bem sucedido numa posse de bola e quantos pontos obteve naquela mesma posse, entretanto por questões complexidade, considerou-se como saída somente se naquela posse o time obteve pontos ou não. Além disso, para se pudesse melhorar a capacidade de previsão da rede no treinamento, baseou-se os treinamentos apenas no time de casa. A forma de utilização dos dados pode ser vista na figura abaixo:

	A	D	F	H	I	J	L	AM	CS	DX	EP	EW
1		match_id	home_tean	away_team_name	poss	period	match_elapsed_minutes	home_quintet_id	away_quintet_id	away_quintet_points	home_result	
8	38604	3939	Bauru	Basq. Cearense	7	1	13.833.333.333.333.300	100:146:187:192:478	273:381:11333:11334:11335	3	0	
9	38605	3939	Bauru	Basq. Cearense	8	1	14.833.333.333.333.300	100:146:187:192:478	273:381:11333:11334:11335	3	0	
10	38606	3939	Bauru	Basq. Cearense	9	1	17.666.666.666.666.600	100:146:187:192:478	273:381:11333:11334:11335	3	2	
11	38607	3939	Bauru	Basq. Cearense	10	1	19.833.333.333.333.300	100:146:187:192:478	273:381:11333:11334:11335	3	-1	
12	38608	3939	Bauru	Basq. Cearense	11	1	19.833.333.333.333.300	100:146:187:192:478	273:381:11333:11334:11335	3	0	
13	38609	3939	Bauru	Basq. Cearense	12	1	2.333.333.333.333.330	100:146:187:192:478	273:381:11333:11334:11335	3	3	
14	38610	3939	Bauru	Basq. Cearense	13	1	2.533.333.333.333.330	100:146:187:192:478	273:381:11333:11334:11335	3	0	
15	38611	3939	Bauru	Basq. Cearense	14	1	2.85	100:146:187:192:478	273:381:11333:11334:11335	3	0	
16	38612	3939	Bauru	Basq. Cearense	15	1	29.333.333.333.333.300	100:146:187:192:478	273:381:11333:11334:11335	3	-2	
17	38613	3939	Bauru	Basq. Cearense	16	1	30.666.666.666.666.600	100:146:187:192:478	105:273:381:11334:11335	0	0	
18	38614	3939	Bauru	Basq. Cearense	17	1	30.666.666.666.666.600	100:146:187:192:478	105:273:381:11334:11335	0	-1	
19	38615	3939	Bauru	Basq. Cearense	18	1	30.666.666.666.666.600	100:146:187:192:478	105:273:381:11334:11335	1	0	
20	38616	3939	Bauru	Basq. Cearense	19	1	34.333.333.333.333.300	100:146:187:192:478	105:273:381:11334:11335	1	-2	
21	38617	3939	Bauru	Basq. Cearense	20	1	34.833.333.333.333.300	100:146:187:192:478	105:273:381:11334:11335	3	0	
22	38618	3939	Bauru	Basq. Cearense	21	1	3.95	100:146:187:192:478	105:273:381:11334:11335	3	0	

Tabela 2: Consideração das entradas como posse de bola

Outra forma de análise foi a utilização dos dados em blocos, ou seja, utiliza-se uma certa quantidade de linhas onde o time de casa mantém posses de bola seguidas. Com isso calcula-se a média do conjunto, o que gera um banco de dados para aquela janela, em seguida o banco de dados gerado é utilizado para o treinamento da rede. Assim mede-se se um time obteve vantagem considerando-se um conjunto de posses, o que faz com o ruído nas previsões seja menor e a previsão seja mais provável, ou seja, mais previsível para rede.

	A	D	F	H	I	J	L	AM	CS	DX	EP	EW
1		match_id	home_tean	away_team_name	poss	period	match_elapsed_minutes	home_quintet_id	away_quintet_id	away_quintet_points	home_result	
44	38640	3939	Bauru	Basq. Cearense	43	1	9.7	85:149:187:192:11290	105:273:381:11333:11335	0	0	
45	38641	3939	Bauru	Basq. Cearense	44	1	9.866.666.666.666.660	85:149:187:192:11290	105:273:381:11333:11335	0	0	
46	38642	3939	Bauru	Basq. Cearense	45	2	0.0	85:149:187:192:11290	105:159:273:381:11333	0	0	
47	38643	3939	Bauru	Basq. Cearense	46	2	0.4333333333333335	85:149:187:192:11290	105:159:273:381:11333	0	0	
48	38644	3939	Bauru	Basq. Cearense	47	2	0.7833333333333332	85:149:187:192:11290	105:159:273:381:11333	0	-1	
49	38645	3939	Bauru	Basq. Cearense	48	2	10.333.333.333.333.300	85:149:187:192:11290	105:159:273:381:11333	1	0	
50	38646	3939	Bauru	Basq. Cearense	49	2	14.000.000.000.000.000	85:149:187:192:11290	105:159:273:381:11333	1	0	
51	38647	3939	Bauru	Basq. Cearense	50	2	14.499.999.999.999.900	85:149:175:478:11290	105:159:273:381:11333	1	0	
52	38648	3939	Bauru	Basq. Cearense	51	2	15.999.999.999.999.900	85:149:175:478:11290	105:159:273:381:11333	1	0	
53	38649	3939	Bauru	Basq. Cearense	52	2	17.166.666.666.666.600	85:149:175:478:11290	105:159:273:381:11333	1	0	
54	38650	3939	Bauru	Basq. Cearense	53	2	2.333.333.333.333.330	85:149:175:478:11290	105:159:273:381:11333	1	-2	
55	38651	3939	Bauru	Basq. Cearense	54	2	25.166.666.666.666.600	85:149:175:478:11290	105:159:273:381:11333	3	0	
56	38652	3939	Bauru	Basq. Cearense	55	2	2.666.666.666.666.660	85:149:175:478:11290	105:159:273:381:11333	3	-2	
57	38653	3939	Bauru	Basq. Cearense	56	2	2.75	85:149:175:478:11290	105:159:273:381:11333	5	0	
58	38654	3939	Bauru	Basq. Cearense	57	2	2.75	85:146:175:478:11291	105:159:273:381:11333	5	-1	
59	38655	3939	Bauru	Basq. Cearense	58	2	2.75	85:146:175:478:11291	105:159:273:381:11333	6	2	
60	38656	3939	Bauru	Basq. Cearense	59	2	2.95	85:146:175:478:11291	105:159:273:381:11333	6	0	

Tabela 3: Consideração das entradas como blocos de posses de bola

Uma terceira abordagem foi a utilização dos dados em segmentos, ou seja, em sequencias de

posses de bola para um mesmo time, ou seja, uma janela que não é mais fixa e sim variável. Posteriormente os segmentos são agrupados em um único registro, assim gera-se um novo banco de dados com segmentos e por fim alimenta-se a rede durante o treinamento. Esta abordagem pode ser vista na figura abaixo:

	A	D	F	H	I	J	L	AM	CS	DX	EP
1		match_id	home_team_name	away_team_name	possession_id	period	match_elapsed_minutes	home_quintet_id	away_quintet_id	away_quintet_points	home_result
33	38629	3939	Bauru	Basq. Cearense	32	1	64.833.333.333.333.300	100:146:187:192:478	105:273:381:11334:11335	6	2
34	38630	3939	Bauru	Basq. Cearense	33	1	6.716.666.666.666.660	100:146:187:192:478	105:273:381:11334:11335	6	0
35	38631	3939	Bauru	Basq. Cearense	34	1	7.066.666.666.666.660	149:187:192:478:11290	105:273:381:11333:11335	0	0
36	38632	3939	Bauru	Basq. Cearense	35	1	7.333.333.333.333.330	149:187:192:478:11290	105:273:381:11333:11335	0	0
37	38633	3939	Bauru	Basq. Cearense	36	1	74.833.333.333.333.300	149:187:192:478:11290	30:105:273:11333:11335	0	0
38	38634	3939	Bauru	Basq. Cearense	37	1	7.683.333.333.333.330	149:187:192:478:11290	30:105:273:11333:11335	0	0
39	38635	3939	Bauru	Basq. Cearense	38	1	79.833.333.333.333.300	149:187:192:478:11290	30:105:273:11333:11335	0	3
40	38636	3939	Bauru	Basq. Cearense	39	1	8.166.666.666.666.660	149:187:192:478:11290	30:105:273:11333:11335	0	-2
41	38637	3939	Bauru	Basq. Cearense	40	1	8.583.333.333.333.330	85:149:187:192:11290	30:105:273:11333:11335	2	2
42	38638	3939	Bauru	Basq. Cearense	41	1	8.966.666.666.666.660	85:149:187:192:11290	105:273:381:11333:11335	0	0
43	38639	3939	Bauru	Basq. Cearense	42	1	9.433.333.333.333.330	85:149:187:192:11290	105:273:381:11333:11335	0	2
44	38640	3939	Bauru	Basq. Cearense	43	1	9.7	85:149:187:192:11290	105:273:381:11333:11335	0	0
45	38641	3939	Bauru	Basq. Cearense	44	1	9.866.666.666.666.660	85:149:187:192:11290	105:273:381:11333:11335	0	0
46	38642	3939	Bauru	Basq. Cearense	45	2	0.0	85:149:187:192:11290	105:159:273:381:11333	0	0
47	38643	3939	Bauru	Basq. Cearense	46	2	0.4333333333333335	85:149:187:192:11290	105:159:273:381:11333	0	0
48	38644	3939	Bauru	Basq. Cearense	47	2	0.7833333333333332	85:149:187:192:11290	105:159:273:381:11333	0	-1
49	38645	3939	Bauru	Basq. Cearense	48	2	10.333.333.333.333.300	85:149:187:192:11290	105:159:273:381:11333	1	0
50	38646	3939	Bauru	Basq. Cearense	49	2	14.000.000.000.000.000	85:149:187:192:11290	105:159:273:381:11333	1	0
51	38647	3939	Bauru	Basq. Cearense	50	2	14.499.999.999.999.900	85:149:175:478:11290	105:159:273:381:11333	1	0
52	38648	3939	Bauru	Basq. Cearense	51	2	15.999.999.999.999.900	85:149:175:478:11290	105:159:273:381:11333	1	0
53	38649	3939	Bauru	Basq. Cearense	52	2	17.166.666.666.666.600	85:149:175:478:11290	105:159:273:381:11333	1	0

Tabela 4: Consideração das entradas como segmentos de posses de bola

4 Apresentação de resultados

Considerando todos os procedimentos e abordagem citadas anteriormente, segue-se nesta seção a apresentação dos resultados obtidos a partir dos experimentos realizados a partir de cada uma das abordagens.

Considerando-se cada uma das linhas de análise abordadas, tinha-se como objetivo obter um nível de acurácia de pelo menos 80%, assim testou-se cada uma das linhas de análise em busca do resultado desejado. Sendo assim, segue-se para a apresentação dos resultados de treinamento da rede para cada uma das linhas de análise:

a Treinamento utilizando posses de bola

Nesse treinamento treinou-se o sistema utilizando cada uma das posses de bola como entrada, ou seja, o dataset como um todo e assim obteve-se os seguintes resultados

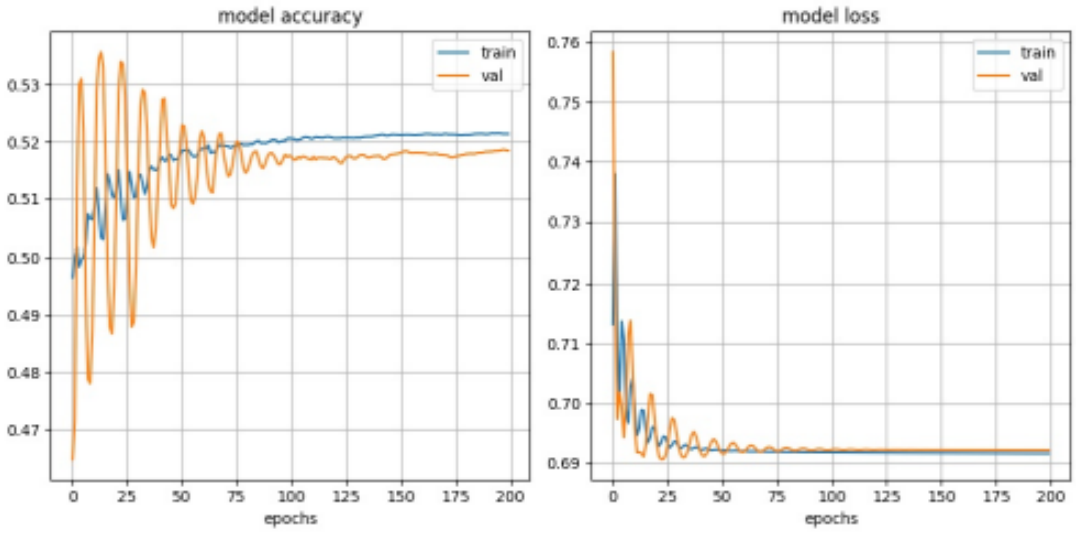

Figura 7: Resultado treino por posse de bola

Nessa abordagem percebe-se que o resultado não atinge um nível de previsão aceitável, tendo um nível de acurácia muito baixa e um valor de erro muito alto.

b Treinamento utilizando segmentos de posses de bola

Nesse tipo de treinamento utilizou-se segmentos de posse de bola conforme anteriormente explicado. Observou-se que o sistema é capaz de acerta mais do que errar:

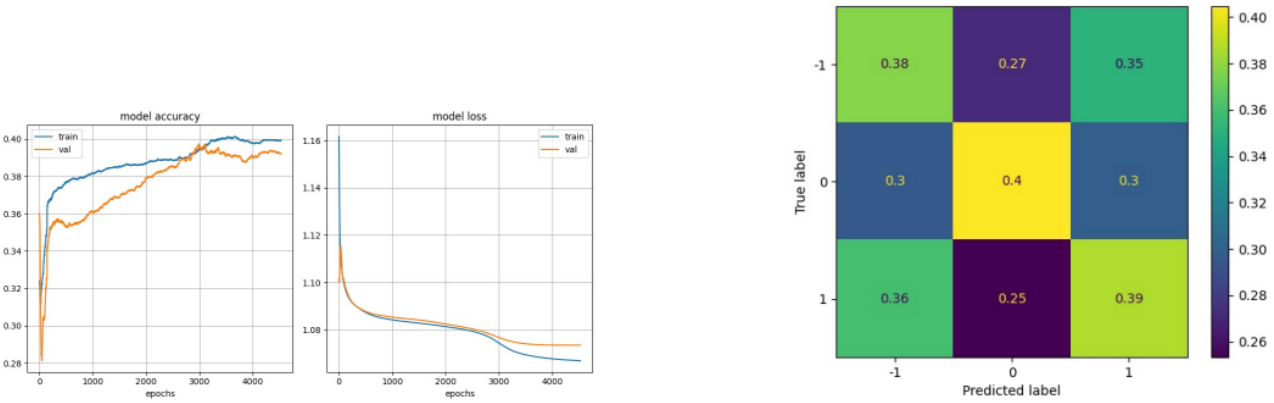

Figura 8: Resultado treino por segmento de posse de bola estagio 1

No entanto sua capacidade de previsão ainda é muito baixa, dada a baixa acurácia nas previsões.

c Treinamento utilizando blocos de posses de bola

Nesse tipo de treinamento houveram 2 abordagens utilizadas, a primeira com um único tamanho fixo de janela e a segunda com vários tamanhos de janela unidos num único dataset. Como explicado anteriormente as janelas são geradas por agrupamento de posses de bola seguidas dentro de um tamanho fixo, na segunda abordagem desta análise o procedimento foi feito para vários tamanhos de janelas gerando um dataset de agrupamento de várias janelas, vários bancos de dados foram gerados e agrupados num único que fora utilizado como treinamento da rede. A segunda abordagem revelou-se mais bem sucedida que simplesmente utilizar um único tamanho de janela como podemos ver no resultado a seguir:

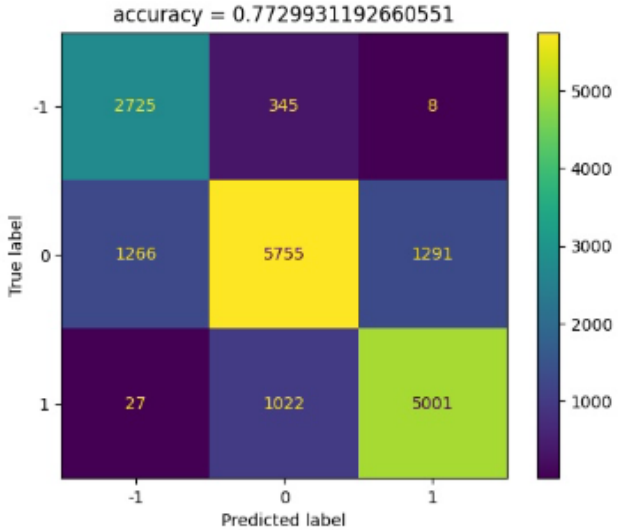

Figura 9: Resultado treino com janelas utilizando rede neural

Como visto no resultado, a abordagem por janelas de forma cumulativa se demonstrou promissora, tendo obtido uma acurácia de 77.7% , algo muito próximo do desejado.
Como tentativa de tentar-se obter um melhor modelo para o estágio 1, também se foi realizado teste utilizando o KNN sob a mesma a abordagem de treino com janelas variadas e de forma

cumulativa, obteve-se o seguinte resultado:

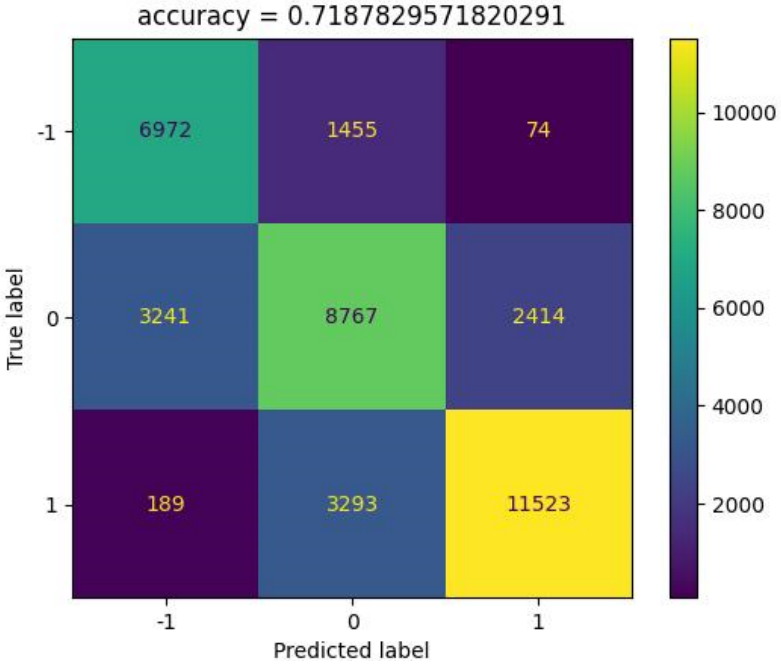

Figura 10: Resultado treino com janelas utilizando KNN

Observa-se que o resultado é até um bom resultado, mas não tão bom quanto o obtido utilizando uma rede neural, sendo assim, acabou-se ficando com o modelo de rede neural para o 1 estágio.

Com a classificação realizada no estágio 1, selecionou-se os as formações de time de casa classificadas como vencedoras e treinou-se o estágio 2:

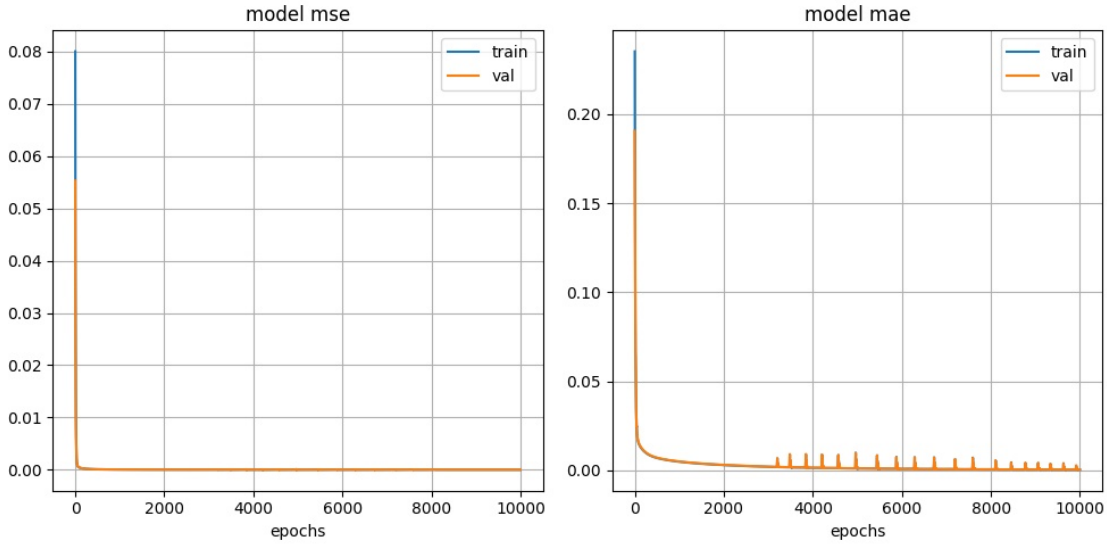

Figura 11: Resultado treino com janelas do estágio 2

Em seguida, foram realizados testes para validar os resultados das predições na variável de saída. O resultado dos testes realizados pode ser visto abaixo:

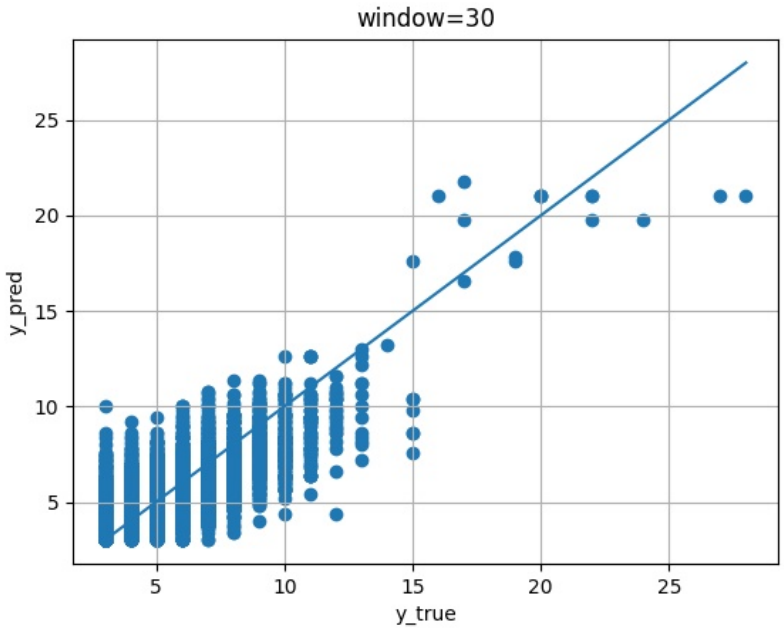

Figura 12: Resultado avaliação das formações de time de casa que venceriam a partida

O que mostra que o sistema desenvolvido obteve um desempenho razoável, o que indica que possíveis melhorias devão ser implementadas para melhorar ainda mais seu desempenho.
Por fim avaliou-se a distribuição de frequência de valores da variável de saída, isto é, saldo de pontos do time de casa em relação ao time adversário, previsto tanto para treino:

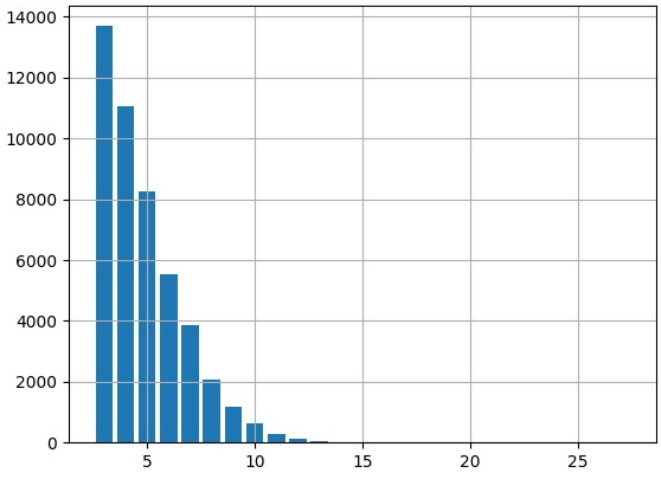

Figura 13: Distribuição de frequência de valores gerados pelo estágio 2 no treino

Quanto para teste:

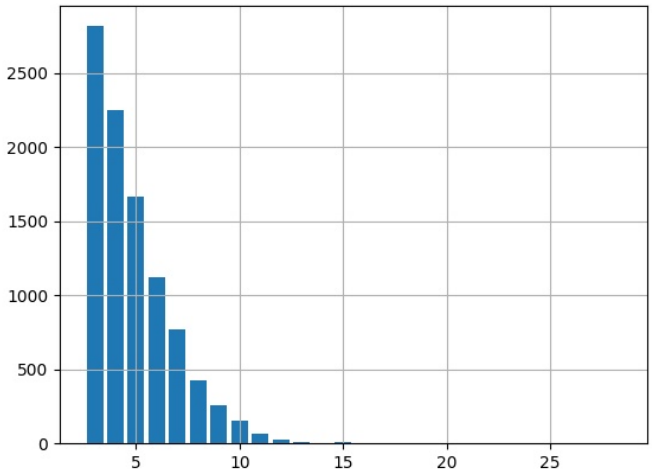

Figura 14: Distribuição de frequência de valores gerados pelo estágio 2 no teste

Percebeu-se que o sistema tem grande probabilidade de prever uma formação de jogadores para o time de casa um pouco melhor do que o time de fora. Formações que exercem forte vantagem são pouco prováveis. Isso não se deve somente a limitações do modelo em si, mas também em limitações do próprio time de casa em termos de jogadores disponíveis e também em relação aos jogadores do time adversário, nem sempre é possível construir uma formação para o time de casa com os jogadores disponíveis muito superior dado aos jogadores adversários que estarão em campo, por isso é pouco frequente as previsões de valores altos para a variável de saída pelo sistema que fora desenvolvido neste projeto.

5 Apresentação de conclusoes e pespectiva de trabalhos futuros

A partir dos resultados apresentados pôde-se constatar que se obteve resultados muito bons dentro das circunstâncias de equilíbrio dos jogos de basquete. Durante o processo de desenvolvimento deste projeto a principal dificuldade encontrada foi o tempo de execução demasiadamente longo devido a capacidade computacional insuficiente para lidar com a complexidade dos algoritmos utilizados e a massiva quantidade de processamento de dados, o que resultou na lentidão da análise de dados, testes e implementação de melhorias no sistema. Sendo assim a análise até então feita não teve tanta consistência. Portanto, considerou-se a análise feita neste projeto uma abordagem inicial que necessitará de melhorias futuramente.

Assim sendo, algumas melhorias poderiam ser consideradas tal como a implementação do desenvolvimento num ambiente cloud onde a capacidade computacional seria menos limitada e algumas automatizações de processos já estariam prontas e haveria ferramenta específicas para determinadas demandas do projeto. Outra melhoria seria uma abordagem considerando que o desempenho dos jogadores pode ser fortemente influenciado pelo contexto de um jogo decisivo ou regular, na análise até então feita o fator motivacional dos jogadores não foi considerado.
Evidentemente que as mudanças propostas não seriam as únicas, o sistema necessitaria de teste em ambientes reais de jogo e de uma implementação de melhoria contínua. Contudo o sistema que fora desenvolvido demonstra-se extremamente promissor para evolução do basquete brasileiro. O técnico de um time de basquete poderia contar com um sistema em tempo real que processa os dados para auxiliar em suas decisões o que potencialmente aumentaria os acertos nas decisões do técnico. Sabe-se que tomar a decisão correta no momento certo é crucial para se ganhar uma partida e que recíproco também é válido, a utilização de um sistema capaz de prever quais seriam as melhores mudanças feitas seria crucial para aumentarem as chances de vitória de um time numa partida de NBB.

Dessa forma, tem-se uma nova perspectiva de evolução para o esporte brasileiro. A tecnologia move no mundo grandes evoluções na medida em que o apresenta novas perspectivas e patamares a serem alcançados. Isso não poderia ser diferente no mundo do esporte, incluindo o basquete brasileiro, o que ficou muito nítido ao longo do desenvolvimento deste projeto.

6 Referências

[1] A. Géron, Hands-On Machine Learning with Scikit-Learn, Keras and TensorFlow., 2019, vol. 2.
[2] Khushijain. K-nearest neighbors. Disponível em: https://medium.com/nerd-for-tech/k-nearest-neighbors-aac72032aaea. Acesso em: 17 SET. 2022
[3] Khushijain. A function approximation of perspective on sensory representations. Disponível em: : https://www.youtube.com/watch?v=qsYx-6rZmoU. Acesso em: 16 SET. 2022
[4] U. Karn. A quick introduction to neural networks. Disponível em: https://www.kd-nuggets.com/2016/11/quick-introduction-neural-networks.html/3. Acesso em: 16 SET. 2022.
[5] M. M. M. M. A. N. A. M. T. T. R. Z. Shamsa Khalid, Muhammad Anees Khan and M. Jehangir. Predicting Risk through Artificial Intelligence Based on Machine Learning Algorithms. Disponível em: https://www.kdnuggets.com/2016/11/quick-introduction-neural-networks.html/3. Acesso em: 18 SET. 2022.
[6] L. Shukla. Natural language processing. Disponível em: https://wandb.ai/ lavanyashukla/vega-plots/reports/Natural-Language-Processing–Vmlldzo2Nzk2Ng. Acesso em: 16 SET. 2022.