Revista Brasileira de Computação Aplicada, Abril, 2021

DOI: 10.5335/rbca.v13i1.11120
Vol. 13, Nº 1, pp. 42–52
Homepage: seer.upf.br/index.php/rbca/index

ISSN 2176-6649

ARTIGO ORIGINAL

Predição da efetividade da substituição no futebol: caso campeonato Brasileiro da Série A

Prediction of substitution effectiveness in football: Brazilian Tournament First Division case

Nicholas Sangoi Brutti[1], Denio Duarte [1], Guilherme Dal Bianco [1]

[1]Universidade Federal da Fronteira Sul - Campus Chapecó

*nicholassbrutti@gmail.com.; duarte@uffs.edu.br; guilherme.dalbianco@uffs.edu.br

Recebido: 02/06/2020. **Revisado:** 17/09/2020. **Aceito:** 09/11/2020.

Resumo

Substituições são recursos importantes para os técnicos de futebol para melhorar o desempenho de suas equipes durante as partidas. Também podem ser cruciais para o resultado final. A relevância e limitação das substituições em jogos oficiais levaram a vários estudos para identificar a melhor maneira de substituir um jogador, ou seja, o melhor momento durante o jogo e a melhor estratégia. Aprendizado de máquina vem sendo aplicado para construir modelos para melhorar a efetividade das substituições. Neste contexto, este trabalho visa propor a criação de modelos que auxiliem os técnicos das equipes visitantes a escolher o melhor momento e estratégia para realizar a substituição. A escolha do modelo para tratar apenas o time visitante foi baseada nas estatísticas que apontam times visitantes têm menores chances de vencer uma partida e, portanto, a criação de modelos se torna mais desafiadora. Foi utilizado um conjunto de dados extraído de quatro edições do Campeonato Brasileiro da Série A (2015-2018), e, a partir deste conjunto, foram extraídos e propostos atributos para a construção do modelo (*e.g.*, a força do time). Quatro classificadores foram avaliados: *Support Vector Machine*, Árvores de Decisão, *Random Forest* e *K-Nearest Neighbor*. Os resultados mostram que os modelos podem atingir até 90% de F_1-Score, ou seja, podem ser promissores para a análise dos melhores momento e estratégia para substituir um jogador.

Palavras-Chave: Campeonato Brasileiro Série A; efetividade da substituição; Futebol; Aprendizado de Máquina; Análise de dados esportivos.

Abstract

Substitutions are essential for coaches to improve team performance and can be crucial for the match result. The relevance and limitation of substitutions in official matches have conducted several studies to propose an optimal way to substitute a player. That is the best moment during the match or the best strategy. Machine learning approaches are widely applied to build prediction models for improving the effectiveness of substitutions. The odds of the away team to win a match are lower than the home team, so, the prediction task is more challenging. Accordingly, this work proposes to build prediction models to help the away team coach deciding the best strategy and moment to replace a player during the match. We use a dataset extracted from four years of the Brazilian Tournament First Division Championship (2015-2018), and based on this dataset, we extract features from the matches and propose new ones (e.g., the team strength) to build the models. Four classifiers are applied: Support Vector Machine, Decision Tree, Random Forest, and K-Nearest Neighbor. The results show that the models can achieve up to 90% of F_1-Score, *i.e*, they can be promising for analysis of the best moment to replace a player.

Keywords: Brazilian Tournament First Division; Effectiveness of substitution; Football; Machine learning; Sports data analysis

1 Introdução

A análise esportiva é uma área multidisciplinar que busca proporcionar aos técnicos e jogadores informações que contribuam para a melhoria contínua do desempenho esportivo. Entre as diversas áreas estudadas, destaca-se o aprendizado de máquina, que vem ganhando grande notoriedade, impulsionado pelo sucesso no uso em outros esportes, como o basquete e o beisebol (Kumar, 2013). Além disso, com o crescente volume de dados disponíveis, técnicas de aprendizado de máquina são frequentemente adotadas para transformar este extenso volume de dados em informação, o que contribui para o processo de tomada de decisão por parte das comissões técnicas dos clubes (Rein and Memmert, 2016).

As substituições, no futebol, são recursos importantes, dadas sua limitação em partidas oficiais e a capacidade de mudança tática proporcionada, o que pode, muitas vezes, influenciar diretamente no resultado final do jogo (Myers, 2012). Por meio das substituições, o treinador define explicitamente qual é sua intenção em relação ao jogo, ou seja, se sua equipe será mais ofensiva ou defensiva. Nesse cenário, surgiram diversos estudos utilizando modelos estatísticos e de aprendizado de máquina. Alguns deles tratam da influência das substituições no resultado, *e.g.*, (Gomez et al., 2016, Flôres et al., 2019), outros avaliam como as variáveis do jogo influenciam no resultado, como em Rey et al. (2015); alguns, também, aplicam técnicas de aprendizado de máquina para definição de uma regra de decisão que determina os momentos mais favoráveis para que uma substituição ocorra (Myers, 2012, Silva and Swartz, 2016).

Porém, não foram encontrados trabalhos que avaliam a predição da efetividade de substituições, de acordo com o andamento da partida. Neste sentido, este trabalho explora a utilização de algoritmos de aprendizado de máquina baseados em classificadores para criação de dois modelos de predição de substituições do time visitante. O primeiro modelo foca na primeira substituição e o segundo modelo foca na terceira substituição, com o objetivo de classificá-las como efetivas ou não. O estudo concentra-se em uma base de dados controlada, referente ao Campeonato Brasileiro de Futebol Série A (conhecido como Brasileirão), dos anos de 2015 a 2018. Para determinar os modelos mais apropriados para o evento estudado, os algoritmos foram submetidos a testes, e através das métricas de avaliação, elegeram-se os melhores modelos que são os que obtiveram maior capacidade de predição.

Como contribuição deste trabalho destacam-se: a proposição de novos atributos para a construção de modelos de previsão, um conjunto de experimentos para identificar a importância dos atributos e a construção de modelos de previsão utilizando quatro algoritmos classificadores. O foco é no time visitante por duas razões: (*i*) é a predição mais desafiadora pois o time local, historicamente, tem mais vitórias que derrotas (Jamieson, 2010) e (*ii*) um modelo que atenda os dois casos se tornaria bastante complexo.

Para o processo de criação dos modelos, foram considerados dados referentes ao andamento do jogo tais como: o tempo (momento) das substituições, o saldo de gols no momento da substituição e o tipo tático da substituição que é atribuído com base na posição do jogador substituído e do substituto. Além disso, foram utilizados atributos de alguns trabalhos correlatos, como a força do time (vantagem do time da casa, média diferencial de gols – veja Seção 4, atributos HTA e D, respectivamente) (Silva and Swartz, 2016). Também foram propostos novos atributos: a força defensiva e ofensiva, além da diferença entre ambas. Esses atributos foram utilizados para melhorar a segmentação e valorizar os dados disponíveis. Foram selecionados os quatro algoritmos de aprendizado de máquina mais frequentes na literatura para o tipo de problema abordado neste trabalho: *Support Vector Machine* (SVM), *Decision Tree* (DTR), *Random Forest* (RFC) e *K-Nearest Neighbor* (KNN). Os algoritmos foram testados com os melhores hiperparâmetros encontrados aplicando o método `GridSearchCV` da biblioteca scikit-learn[1]. As métricas utilizadas para medir o desempenho são: Precisão, Revocação e F_1-Score.

Os Experimentos realizados mostraram que os modelos de previsão propostos tiveram uma precisão de até 86% na identificação da efetividade das segundas e terceiras substituições do time visitante. Isso demonstra a possível viabilidade da aplicação de tais modelos em cenários reais.

O restante deste artigo está estruturado da seguinte maneira: a próxima seção apresenta alguns conceitos importantes para a compreensão da proposta; na Seção 3 são discutidos alguns trabalhos relacionados; as Seções 4 a 6 apresentam as contribuições deste trabalho; e a Seção 7 conclui e apresenta algumas direções futuras.

2 Referencial Teórico

Esta seção apresenta brevemente alguns conceitos para o apropriado entendimento deste trabalho.

2.1 Aprendizado de Máquina

O aprendizado de máquina é uma subárea da inteligência artificial cujo foco principal é permitir que computadores aprendam a reconhecer padrões para tomarem decisões de forma inteligente com base nos dados de entrada. O processo de aprendizagem recebe um conjunto de entradas e saídas, eventualmente, as saídas podem ser nulas (Duarte and Ståhl, 2019). Com esse conjunto de dados, aplicam-se técnicas de aprendizado para obtenção de um modelo que represente o fenômeno em questão. Esta etapa é conhecida como treinamento. Posteriormente, ocorre a etapa de teste. Nesse caso, o modelo recebe novos dados como entrada, e responde de acordo com o conhecimento obtido na etapa anterior. Ao final, métricas são aplicadas para verificação do nível de representatividade do modelo.

Formalmente, o conjunto de treinamento pode ser representado como $X = \{(x^{(1)}, y^{(1)}), \ldots, (x^{(m)}, y^{(m)})\}$, em que m é o tamanho de X, e $y^{(j)}$ é o rótulo (possivelmente vazio) de $x^{(j)}$ $(1 \leq j \leq m)$. Através do conjunto X, cria-se o modelo (hipótese). Para verificação da hipótese, um novo conjunto de dados X' tal que $\{X' \cap X = \emptyset\}$ é usado como

[1] scikit-learn.org

entrada para o modelo. Caso o rótulo não exista (*i.e.*, $y^{(j)}$ é vazio), o tipo de aprendizado aplicado é o não supervisionado, caso contrário, aplica-se o aprendizado supervisionado. Em aprendizado supervisionado, se os valores de $y^{(j)}$ são discretos, o problema é dito de classificação, se não (*i.e.*, $y^{(j)} \in \mathbb{R}$) é dito de regressão.

Este trabalho se apoia em quatro algoritmos classificadores para criação do modelo de previsão: *Decision Tree* (DTR), *Random Forest* (RFC), *Support Vector Machine* (SVM) e *K-Nearest Neighbor* (KNN).

DTR e RFC são algoritmos que criam modelos baseados em árvores em que os nós são regras e as folhas as classes que o caminho das regras prediz. O RFC difere do DTR pois constrói um conjunto de subárvores de decisão em que cada subárvore prediz classes que são aferidas entre si. A classe mais predominante é a escolhida.

Já os algoritmos SVM e KNN projetam os dados (X) em um hiperplano. O KNN se baseia nas distância entre os pontos para encontrar as classes. Se K pontos são os vizinhos mais próximos de $x^{(j)}$ e a classe predominante é c, $x^{(j)}$ é classificado como c. Por outro lado, o SVM procura encontrar um hiperplano ótimo que separe os dados conforme as classes.

Os algoritmos de aprendizado de máquina possuem várias configurações para otimizar a construção dos modelos, chamadas de hiperparâmetros. Por exemplo, para algoritmos baseados em árvores, a pureza de um nó na árvore e a função de escolha dos atributos mais informativos (*e.g.*, Gini ou Entropia) são alguns dos hiperparâmetros, ou a função de distância entre os pontos (*e.g.*, Manhattan e Euclidiana) ou tipo de núcleo (*e.g.*, Gaussian kernel e Radial basis function (RBF)) para KNN ou SVM, respectivamente.

2.1.1 *Métricas*

A criação de um modelo de predição envolve vários passos e, quando um modelo está pronto, é necessário avaliar o resultado comparando as classes preditas com as classes verdadeiras. Essa avaliação é feita por meio do uso de métricas. As métricas se baseiam nos resultados de predição de uma das classes que podem ser categorizados de quatro formas, conforme a matriz de confusão apresentada na Tabela 1. A categoria *Verdadeiro Positivo* (VP) é a situação em que a classe foi corretamente predita como positiva. A categoria *Verdadeiro Negativo* (VN), por sua vez, é a situação em que a classe foi corretamente predita como negativa. Ao classificar um exemplo que é negativo como positivo, tem-se a categoria de *Falso Positivo* (FP). A categoria de *Falso Negativo* (FN) se dá quando classifica-se um exemplo positivo como negativo.

Tabela 1: Matriz de confusão

	Verdadeiro Positivo	Verdadeiro Negativo
Falso Positivo	VP	FP
Falso Negativo	FN	VN

As principais métricas utilizadas para a validação de modelos de classificação são baseadas na contagem de classes previstas. A acurácia é mais tradicional entre elas e pode ser definida como todos os acertos sobre todos os exemplos e é dada pela Equação 1. Acurácia é a métrica recomendada quando o número de classes do conjunto de dados é balanceado (Duarte and Ståhl, 2019). Se um dado conjunto de dados, por exemplo, possui 96% de exemplos positivos e um modelo retorna apenas exemplos positivos, este modelo terá uma acurácia de 96%. Porém, este mesmo modelo não é capaz de identificar um exemplo negativo.

Quando o conjunto de dados é desbalanceado outras métricas são recomendadas, como a precisão (Eq. (2)), a revocação (Eq. (3)) e a F_1-Score (Eq. (4)), que é a média aritmética ponderada entre a precisão e a revocação.

$$A = \frac{VP + VN}{VP + VN + FP + FN} \quad (1) \qquad P = \frac{TP}{TP + FP} \quad (2)$$

$$R = \frac{TP}{TP + FN} \quad (3) \qquad F_1 = 2 \times \frac{P \times R}{P + R} \quad (4)$$

Intuitivamente, a precisão é usada quando é necessária assertividade para encontrar os exemplos positivos. A revocação, por sua vez, é usada quando é necessária assertividade em não deixar de identificar nenhum exemplo positivo.

2.2 Futebol

O futebol é um esporte amplamente conhecido e é considerado o mais popular do mundo (Morrow, 2003). O futebol tem uma extensa comunidade de espectadores e praticantes e, por meio de seus grandes torneios, é responsável por gerar impactos positivos para economia e promover maior desenvolvimento local (Allmers and Maennig, 2009). Um dos motivos para este sucesso, é o fato de ser um esporte de simples interpretação, o que facilita o entendimento por parte dos espectadores. Outros motivos são explorados em Dunmore and Murray (2013).

Uma partida de futebol oficial é composta por dois times. Cada time apresenta no máximo onze e no mínimo sete jogadores titulares, caso contrário a partida não poderá acontecer. Os times apresentam um banco de reservas com no máximo 12 jogadores suplentes (CBF, 2017). Considerando a situação inicial de um jogo, em cada time há onze jogadores em campo. A comissão técnica, portanto é responsável por determinar estrategicamente o posicionamento de 10 jogadores (obrigatoriamente há um goleiro), de acordo com o esquema de jogo a ser adotado.

Conforme Dunmore and Murray (2013), existem quatro posições básicas:

- Atacante: responsável por propor jogadas ofensivas e finalizações, com o objetivo de marcar gols;
- Defensor: preocupa-se primeiramente em defender seu gol, evitando que o time adversário aproxime-se da área;
- Goleiro: é o jogador mais próximo ao gol de sua equipe. Seu objetivo é evitar que as finalizações dos jogadores adversários se concretizem em gols. É o único capaz de utilizar as mãos e os braços para efetuar a defesa, desde que esteja dentro da área delimitada;
- Meio-campista: são os jogadores mais versáteis, contribuem tanto na defesa quanto no ataque.

Para tornar a definição de Meio-campista mais específica, no meio futebolístico ela é dividida em outras duas

posições. São elas: o Volante que é um Meio-campista que atua predominantemente no setor defensivo, e o Meia-atacante cuja responsabilidade é atuar mais incisivamente no setor ofensivo.

A posição de cada jogador pode variar de acordo com o andamento da partida. Diante das circunstâncias da partida e do plantel de jogadores disponíveis, a comissão técnica pode optar por utilizar mudanças táticas (*i.e*, alteração da formação e substituições). Estas modificações têm o intuito de maximizar o desempenho da equipe, seja para reforçar o setor defensivo e manter o placar, ou para priorizar o ataque e maximizar a probabilidade de vitória (Del Corral et al., 2008).

Uma partida regular de futebol apresenta um total de 90 minutos de jogo divididos em dois tempos de 45 minutos. Toda a partida está suscetível a possuir maior tempo de duração, pois em caso de interrupções no jogo, o tempo parado poderá ser convertido em acréscimos. A responsabilidade de definir o tempo total de acréscimos da partida é do árbitro previamente escalado.

Todas as regulamentações envolvendo o futebol são publicadas pela FIFA (Federação Internacional de Futebol – `www.fifa.com`) que é responsável por propor leis no âmbito futebolístico (Brillinger, 2011).

2.2.1 *Substituições*

Conforme citado anteriormente, durante uma partida os clubes estão sujeitos a mudanças táticas, seja por iniciativa da equipe técnica ao identificar carências, ou devido a problemas físicos dos atletas. Uma alternativa utilizada nessas situações, são as substituições. No futebol oficial as substituições têm regras que diferem em relação a outros esportes. Isso porque há uma limitação no número de substituições. Durante um jogo, são permitidas no máximo três alterações. Além disso, a partir do momento que o jogador foi substituído, ele não poderá reingressar no jogo que encontra-se em andamento.

Essas características reforçam a importância da equipe técnica acertar suas decisões de alterações de jogadores. Em muitos casos, as substituições são fatores que decidem o jogo, seja de maneira positiva ou negativa (Myers, 2012, Flôres et al., 2019), especialmente por se tratar de um recurso limitado (Rey et al., 2015).

Nesse contexto, considerando as dificuldades de se estabelecer o momento correto para uma substituição, diversos trabalhos surgiram com objetivo de analisar dados históricos de partidas e determinar uma regra de decisão, *e.g.*, (Myers, 2012, Silva and Swartz, 2016).

O momento e a estratégia das substituições são os pontos principais para a proposta de um modelo neste trabalho que irá abstrai as mudanças de esquemas táticos, devido a restrição dessa informação do conjunto de dados de entrada. Portanto, a partir do momento em que as substituições anteriores ocorreram, mais as variáveis circunstanciais do jogo em andamento pretende-se classificar a substituição como efetiva ou não.

2.2.2 *Campeonato Brasileiro de Futebol - Série A*

É a principal competição entre clubes de futebol do Brasil e envolve os 20 clubes da elite do futebol nacional. Conhecida popularmente como "Brasileirão", é uma competição organizada pela CBF (Confederação Brasileira de Futebol - `www.cbf.com.br`), que além de possibilitar o título e premiações ao primeiro colocado, permite aos seis primeiros colocados acesso à Taça Libertadores da América. Sendo que os quatro primeiros, têm a vantagem de ingressar diretamente na fase de grupos, enquanto o quinto e sexto colocado participam da fase preliminar (CBF, 2017).

Assim como em outras ligas nacionais, o Brasileirão atualmente segue o modelo de pontos corridos, ou seja, uma competição de longa duração. Uma temporada é composta por 380 jogos incluindo 20 equipes participantes e 38 rodadas disputadas.

3 Trabalhos Relacionados

As substituições são recursos existentes em diversas modalidades esportivas. Porém, cada esporte pode ter uma interpretação diferente sobre sua utilização, conforme definido pela entidade reguladora. O futebol em partidas oficiais, conforme citado anteriormente, apresenta algumas particularidades em comparação aos outros esportes, destaca-se a limitação do número de substituições e a impossibilidade de reingresso de jogadores já substituídos. Os trabalhos, aqui selecionados, tratam exclusivamente do estudo das substituições no contexto do futebol.

Hirotsu and Wright (2002) estudaram o problema de definição do melhor momento para substituição. Eles modelaram a partida de futebol como um processo de decisão de *Markov*, contendo quatro estados. As probabilidades de transição são estimadas através do método estatístico de máxima verossimilhança. Utilizando programação dinâmica busca-se encontrar o tempo ótimo para uma *mudança tática*. Variáveis são incorporadas ao modelo tais como: o local da partida (casa ou fora), tempo restante da partida e a formação tática. Uma característica importante deste trabalho é a consideração das *mudanças táticas*. Para promover uma mudança tática não é necessário acontecer uma substituição de um jogador, basta que o posicionamento seja modificado. Embora os resultados sejam relevantes, o modelo carece de uma maior validação, pois foram utilizados poucos dados que são provenientes de partidas hipotéticas.

Já em (Del Corral et al., 2008), os aspectos determinantes para realização da primeira substituição foram estudados. Para tanto, é utilizado um modelo *hazard* Gaussiano inverso, cujo objetivo é identificar as variáveis estatísticas significativas que explicam as substituições dos jogadores durante uma partida de futebol. Foram utilizados os dados da primeira liga espanhola, nas edições de 2004 e 2005. Foram consideradas algumas variáveis na construção do modelo. Por exemplo, se a substituição foi feita pelo time da casa ou não, a classificação do time no campeonato antes do jogo, a diferença no placar do jogo, tipo da substituição (ofensiva, defensiva ou neutra) e os pontos das equipes nos últimos quatro jogos. Os resultados mostraram que existem elementos estratégicos importantes que determinam o tempo das substituições. Em particular, foi descoberto que o fator mais importante é o placar do jogo no momento da substituição e que as decisões dos treinadores dependem de sua equipe estar jogando em casa ou fora. Além disso, foi evidenciado que as equipes da casa têm uma probabilidade condicional mais alta de fazer sua primeira substituição no intervalo, quando a pressão da

torcida é menor.

O trabalho de (Myers, 2012) foca no fator crítico das substituições: o tempo. Historicamente, os treinadores tendem a fazer as substituições de forma mais reativa do que proativa. Esta abordagem nas substituições pode ser equivocada. Se for preciso esperar sinais para identificar que uma substituição precisa ser feita, pode ser que o momento crítico para aquela determinada alteração já tenha passado. O autor analisou dados históricos de diversas ligas do futebol (*Premier League* inglesa, Série A italiana e *La Liga* espanhola) com o objetivo de propor uma regra de decisão, contendo os melhores intervalos para a ocorrência de uma substituição. As conclusões do trabalho foram: o placar é um fator predominante para realizar a substituição, o momento da substituição difere para o time local e visitante (a maior discrepância é na terceira substituição) e as ligas estudadas têm tempos diferentes no momento da substituição. Finalmente, o estudo propôs as seguintes regras para efetuar as substituições quando uma das equipes estiver perdendo: (*i*) a primeira substituição deve ocorrer antes do minuto 58, (*ii*) a segunda antes do minuto 73 e (*iii*) a terceira antes do minuto 79. Essas regras não são aplicáveis quando as substituições são por lesão ou cartões vermelhos. De acordo com Myers (2012), as equipes que utilizaram esta proposta de substituições, obtiveram uma taxa de sucesso de 38% a 47%, comparada com 17% a 24% de quando não é seguida.

O trabalho de (Rey et al., 2015) examina a influência das variáveis situacionais no tempo e na tática das substituições. O conjunto de dados é composto exclusivamente pelas substituições na Liga dos Campeões da UEFA. Esta competição possui particularidades em relação aos campeonatos abordados nos trabalhos anteriores. Destaca-se o formato do torneio, que neste caso segue o modelo popularmente conhecido como *mata-mata*. Basicamente neste modelo há inicialmente a fase de grupos com um total de 32 times, divididos em 8 grupos. Os grupos são definidos a partir de um sorteio, de forma com que os clubes integrantes não compartilhem o mesmo país de origem. Durante esta etapa cada equipe joga uma partida em casa e outra fora com cada membro do grupo. Após todos os jogos, os dois primeiros colocados de cada grupo se classificam para a próxima fase (UEFA, 2018). Na fase seguinte, a competição adota a prorrogação. Ela é aplicada em casos em que as partidas permanecem empatadas ao término tempo regulamentar, levando em consideração o gol qualificado, ou seja, o gol feito fora de casa. Considerando que o jogo já extrapolou o tempo normal, observa-se que provavelmente haverá substituições nesse período, dado ao alto nível de exigência física de jogos em competições desse molde. Portanto, esses dados são relevantes para construção de um modelo mais próximo ao real. Além disso, duas variáveis dependentes foram consideradas. Primeiramente, o tempo em que a substituição aconteceu junto a um rótulo que representa o número da substituição. A outra variável diz respeito a questão tática da substituição. Dadas as posições do jogador substituído e do jogador substituto, é possível classificar a alteração como defensiva, ofensiva ou neutra. Como resultado foi obtido uma regra de decisão semelhante ao estudo de Myers (2012). Caso o time esteja perdendo: (*i*) a primeira substituição antes do minuto 53, (*ii*) a segunda antes do minuto 71 e (*iii*) a terceira antes do minuto 80. Os resultados mostram que a qualidade do oponente e o placar da partida são fatores chave, que devem ser analisados para determinar a necessidade de substituição ou não. Os autores finalizam elucidando que embora tenham identificado variáveis que implicam diretamente no tempo das substituições, ainda há outras informações abstraídas que poderiam melhorar a qualidade do estudo, como: condições físicas, qualidade do campo, táticas adversárias e aspectos culturais (Rey et al., 2015). Orienta-se que os estudos futuros busquem agregar esse conhecimento na etapa de construção do modelo.

O trabalho de (Silva and Swartz, 2016) apresentou uma análise alternativa sobre a regra encontrada por Myers (2012). Adotou-se uma abordagem utilizando a Regressão Logística Bayesiana, que é caracterizada por aproveitar o conhecimento *a priori*. Também foi utilizada uma maior quantidade de dados, e um atributo que define a *força* do time. O intuito de determinar a força do clube surgiu da hipótese de que um clube com um elenco mais forte provavelmente possui maiores chances de diminuir a diferença de gols. Portanto, as substituições em times mais fortes, em tese, pode apresentar maior ganho em qualidade ao time. Foi encontrado que, com equipes equiparadas, há uma vantagem de gols para o time visitante no segundo tempo. Adicionalmente, observaram que não há tempo discernível durante o segundo tempo, em que houve algum benefício através de substituições.

Os trabalhos citados não abordam a efetividade das substituições sendo que a maioria indica o melhor momento da substituição. Assim, esta proposta tem como objetivo preencher esta lacuna ao promover uma abordagem baseada em aprendizado de máquina para a previsão de efetividade de substituições.

4 Obtenção do Conjunto de Dados e Seleção dos Atributos

A criação de modelos de aprendizado de máquina envolve vários passos, um dos passos mais importantes é a extração dos atributos (Guyon et al., 2008). O processo de extração de atributos inclui a sua construção, redução da dimensionalidade, representação esparsa do conjunto de dados e a seleção dos atributos mais relevantes. Esta seção apresenta a forma que o conjunto de dados foi modelado, bem como, o processo de construção dos atributos a partir deste conjunto.

Os dados utilizados neste trabalho foram extraídos do portal de acompanhamento de jogos em tempo real da UOL Esporte (`www.uol.com.br/esporte`). A extração foi realizada utilizando um *web scraper* através de solicitação dos dados via *XMLHttpRequest*, cuja resposta é um documento *JSON*.

A partir dos dados extraídos, um banco de dados relacional foi gerado baseado no modelo *snowflake* (Chaudhuri and Dayal, 1997). Nele a tabela de fatos é a partida e as dimensões representam os eventos relacionados às partidas. Os eventos são: os gols contra e a favor, cartões vermelhos e amarelos, os pênaltis e as substituições. A Figura 1 apresenta a organização do modelo *snowflake*. Para adicionar informações sobre o tipo tático das substituições, foi necessário extrair o posicionamento do jogador subs-

tituído e o substituto, então foi preciso recuperar estes dados localizados na seção de escalação. Perceba que os dados dos jogadores não são necessários para este trabalho, logo não foram incluídos no modelo. A Tabela `partidas` armazenava, originalmente, 1518 partidas, mas em algumas não existe a informação de posicionamento tanto do jogador substituto quanto do substituído. Embora existam técnicas para imputação de dados, optou-se pela exclusão destas partidas do conjunto de dados. Aplicou-se, também, um filtro e foram selecionadas somente as partidas em que o time visitante realizou as 3 substituições. Dessa forma, após estes procedimentos o conjunto ficou com dados de 1326 partidas, ou seja, foram excluídos dados de 192 partidas. Nesse conjunto, há um total de 3978 substituições para análise e utilização nos algoritmos de aprendizado de máquina. O arquivo *JSON* pode ser acessado em `github.com/nbrutti/uol-export/`.

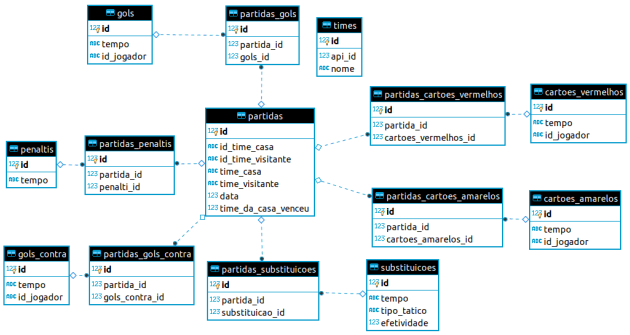

Figura 1: Modelo Entidade Relacionamento.

Após a aquisição dos dados e a construção do modelo do banco de dados, iniciou-se a etapa de criação de novos atributos. O objetivo é aproveitar as informações disponibilizadas no banco de dados para gerar novos atributos que contribuam para a criação do conjunto de treinamento, além disso, espera-se que eles forneçam maior capacidade de classificação e representatividade ao modelo de aprendizado de máquina. Os atributos *HTA (Home Team Advantage)*, *RC* (Ganhador Casa) e *RV* (Ganhador Visitante) foram inspirados em (Silva and Swartz, 2016), enquanto os atributos $F_{[V,C]D}$ (Força Defensiva Visitante ou Casa), $F_{[V,c]O}$ (Força Ofensiva Visitante ou Casa), *DOD* (Diferença Ofensiva para Defensiva) e *DDO* (Diferença Defensiva para Ofensiva) foram criados intuitivamente, de acordo com os dados disponíveis. A descrição desses atributos é apresentada a seguir.

Vantagem do time da casa (`HTA`): é um atributo contínuo, calculado para mensurar a força dos times quando jogam em casa em determinada edição do Brasileirão. Ele não é explicitamente utilizado no conjunto de dados, é um indicador auxiliar para o cálculo dos atributos descritos nas Equações 5 e 6. O cálculo funciona da seguinte maneira: dado o ano de edição, soma-se todas as ocorrências de gols efetuados pelo time da casa. O mesmo é efetuado para os gols do time visitante. Na sequência efetua-se uma subtração entre os gols do time da casa e os gols do time visitante, e depois é realizado a divisão pelo total de jogos da edição:

$$HTA = \frac{\text{total gols casa} - \text{total gols visitante}}{\text{total partidas}}$$

Média diferencial de gols (D): esse atributo é responsável por representar a média diferencial de gols de uma determinada edição do Brasileirão. O cálculo necessita de duas informações prévias: o time T que será submetido à análise e a edição (E) que deseja-se avaliar. Para montagem da equação, primeiramente é calculado a quantidade de gols que o time T efetuou na edição E. Depois, calcula-se a quantidade de gols sofridos pelo time T na edição E. No final, é subtraído o valor de gols marcados e sofridos, e dividido pelo total de jogos:

$$D = \frac{\text{total gols pró} - \text{total gols contra}}{\text{total partidas}}$$

Identificação do provável ganhador (RV e RC): esses atributos são criados com base nos dois anteriores (*i.e.*, `HTA` e D). Utilizando os resultados da média diferencial de gols e do `HTA`, é possível criar uma regra que defina qual o time favorável a vencer, considerando as forças dos dois clubes envolvidos e a qualidade do time da casa em partidas realizadas em seu estádio. Para explicação da regra, suponha que D_j represente o time da casa e D_k o visitante. Se $D_j - D_k + HTA \geq 0$ então o time da casa é favorável a vencer a partida, logo, provavelmente suas substituições serão bem sucedidas. Assim, o atributo RC (ganhador casa) recebe 1, e o atributo RV (ganhador visitante) recebe 0. Caso RC receba 0, significa que o time favorável a conquistar a vitória é o time visitante.

$$RC = \begin{cases} 1, & \text{se } D_j - D_k + HTA \geq 0 \\ 0, & \text{caso contrário} \end{cases} \quad (5)$$

$$RV = \begin{cases} 1, & \text{se } D_j - D_k + HTA < 0 \\ 0, & \text{caso contrário} \end{cases} \quad (6)$$

Força defensiva dos times: como uma medida para determinar a força defensiva dos times participantes de uma partida, foram criados os atributos F_{CD} (representa a força do time da casa) e F_{VD} (representa a do time visitante). A proposta é que clubes com maior força defensiva consigam maior taxa de sucesso em substituições deste tipo. A fórmula a seguir apresenta o cálculo da força defensiva. Perceba que quanto mais próximo o resultado for de 0, maior a força defensiva da equipe:

$$F_{[V,C]D} = \frac{\text{total gols contra}}{\text{total de partidas}}$$

Força ofensiva dos time: semelhante ao raciocínio anterior, tem o objetivo de indicar a força ofensiva dos clubes. A hipótese é de que clubes com maior força ofensiva provavelmente terão sucesso ao realizar substituições ofensivas. O cálculo é feito conforme fórmula a seguir. Perceba que

quanto maior o $F_{[V,C]O}$, maior a força ofensiva da equipe:

$$F_{[V,C]O} = \frac{\text{total gols pró}}{\text{total de partidas}}$$

Diferença entre as forças dos times: foram criados dois atributos para armazenar as diferenças. O primeiro intitulado *DOD* (*i.e.*, diferença ofensiva para defensiva): $DOD = F_{CD} - F_{VO}$. O intuito desse atributo é armazenar a diferença entre F_{CD} e F_{VO}. Uma diferença grande pode indicar que o nível de disparidade entre os dois adversários é significativo. A hipótese é que esta informação também agregue no desenvolvimento e desempenho do trabalho. Já o segundo, com características comuns ao anterior, é chamado de *DDO* (*i.e.*, diferença defensiva para ofensiva): $DDO = F_{VD} - F_{CO}$.
Efetividade da substituição: é um atributo binário que representa o sucesso ou não da substituição. Para o time visitante, a efetividade da substituição está associada a dois fatores: ao tipo tático da substituição e ao modelo de competição adotada. Por padrão, o conjunto de dados não apresenta a informação do tipo tático da substituição, logo foi desenvolvido um método que faz o papel de atribuir um rótulo para a substituição. Os possíveis rótulos da substituição são: ofensivo (*OFF*), defensivo (*DEF*) e sem alteração (*SA*). Essa codificação foi baseada no trabalho de Rey et al. (2015). De forma simplificada, há um método intermediário que recebe a posição de um jogador e atribui um índice ascendente de acordo com o grau de ofensividade. As posições foram subdivididas em seis setores: (*i*) o Goleiro é codificado como 0, (*ii*) posições defensivas (Zagueiro, Lateral-direito e Lateral-esquerdo) como 1, (*iii*) Volantes como 2, (*iv*) Meias armadores como 3, (*v*) Meias-atacantes 4 e (*vi*) Atacantes 5. O valor -1 é o rótulo atribuído quando a informação encontra-se ausente.

O índice implementado tem como objetivo identificar o tipo da substituição. Se o índice atribuído ao jogador substituto for maior que o do substituído, sabe-se que esta substituição foi do tipo ofensiva (*e.g.*, zagueiro vs. atacante e volante vs. meia-atacante). Caso contrário, a substituição é considerada defensiva. Já quando o valor do índice é igual, a substituição é classificada como sem alteração (*SA*).

Para a classificação da substituição como efetiva ou não efetiva é avaliado se, no intervalo entre o momento que a substituição ocorreu e o final do jogo, houveram gols por parte de qualquer uma das equipes. Independentemente do tipo da substituição, se no intervalo houveram gols do time visitante, a substituição é classificada como positiva, afinal houve um incremento favorável no placar. Já se a substituição foi do tipo *SA* ou *DEF* e o time adversário não marcou gols, a substituição também é positiva. Isto significa que o fortalecimento do setor defensivo ou a alteração para recuperação da condição física surtiram efeito. Nos outros casos, a substituição é classificada como negativa.

Também foi conduzida uma análise exploratória do conjunto de dados para entender como um modelo de previsão de efetividade de substituições poderia ser criado. Conforme também identificado por Myers (2012), em aproximadamente 87% dos jogos selecionados, a equipe visitante efetua as três substituições. Em contrapartida, aproximadamente 11% dos jogos realizaram duas substituições, enquanto 1% realizou apenas uma substituição. Um dado significativo que reforça a importância dada pelos clubes para esse recurso.

Outra constatação é em relação ao momento da substituição. Testes empíricos indicaram que a melhor divisão do tempo da partida é de 15 em 15 minutos. Assim, considerando um intervalo de 15 minutos, percebeu-se que o time visitante (objeto de estudo deste trabalho) tende a realizar a terceira e a segunda substituição nos 30 minutos finais da partida. Por outro lado, a primeira substituição se encontra distribuída durante toda a partida, o que mostra quase uma uniformidade do minuto 30 ao 75. A Figura 2 apresenta o histograma com as quantidades de substituições, conforme as fatias de tempo definidas. Esse comportamento é semelhante ao encontrado por Gomez et al. (2016) ao analisar um conjunto de dados distinto. Já o tipo tático das substituições teve a seguinte distribuição quanto ao momento:

- Primeira: 43.7% SA, 33.8% OFF e 22.5% DEF.
- Segunda: 43.0% SA, 33.0% OFF e 24.0% DEF.
- Terceira: 39.0% SA, 32.5% OFF e 28.5% DEF.

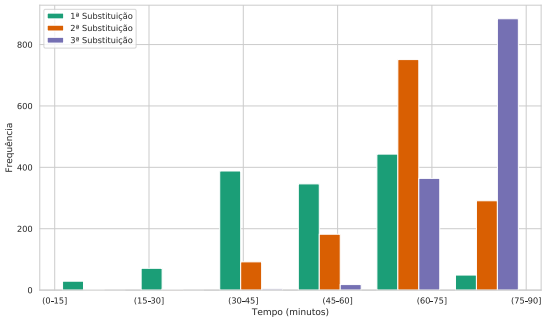

Figura 2: Histograma das substituições.

Pode-se ressaltar que a substituição do tipo tático defensivo (*DEF*) ocorreu com maior probabilidade na terceira substituição. Aliás, é na terceira substituição que há a informação mais relevante para a previsão da efetividade ou não. De acordo com os dados, é nessa substituição que os times priorizaram alterações que visavam tornar o time mais agressivo ou defensivo, abdicando de alterações do tipo *SA*. Este fato, expõe o real interesse do time adversário no minuto T_3 (minuto em que a terceira substituição aconteceu) até o final da partida. A decisão é formulada levando em consideração as circunstâncias da partida (*i.e.*, placar, quantidade de jogadores em campo, entre outros) e até mesmo o placar dos jogos dos clubes adversários na tabela de classificação (*e.g.*, clubes com chances reais de rebaixamento tendem a priorizar a vitória).

5 Experimentos

Esta seção apresenta a metodologia para a execução dos experimentos, bem como os resultados obtidos.

Tabela 2: Extrato do Conjunto de Dados II.

G_1	G_2	G_3	S_1	S_2	S_3	T_1	T_2	T_3	F_{VD}	F_{VO}	F_{CD}	F_{CO}	R_C	R_V	*DOD*	*DDO*	y_3
0	0	0	2	2	2	2	2	3	0.55	0.51	0.55	0.41	0	1	-0.04	0.14	1
1	1	1	2	2	2	1	3	3	0.72	0.43	0.72	0.44	1	0	-0.29	0.28	0
0	0	0	2	1	0	2	2	3	0.70	0.38	0.70	0.43	1	0	-0.32	0.27	1

5.1 Conjunto de Dados

Após a análise do conjunto de dados e a criação de novos atributos, obteve-se o seguinte conjunto de atributos:

- G_n representa o saldo de gols antes da substituição n ser efetuada;
- S_n representa o tipo tático da substituição n, codificado da seguinte maneira: 0 para substituições defensivas, 1 para ofensivas e 2 para as que não promovem alteração tática;
- T_n é composta pela classe relativa ao tempo em que a substituição n aconteceu. Conforme apresentado anteriormente, a duração da partida foi dividida em 6 partes, de 15 minutos cada. Então, T_n pode receber os seguintes valores: 1, 2, 3, 4, 5 e 6. Por exemplo, 3 indica que a substituição ocorreu entre (30, 45] minutos.
- F_{VD} representa a força defensiva do time visitante;
- F_{VO} representa a força ofensiva do time visitante;
- F_{CD} representa a força defensiva do time da casa;
- F_{CO} representa a força ofensiva do time da casa;
- R_C e R_V são atributos binários, responsáveis por informar se o time da casa ou o time visitante é favorável a vencer, respectivamente;
- *DOD* e *DDO* representam a diferença entre as forças ofensiva e defensiva dos dois times.
- y_n apresentam a informação se a substituição n foi efetiva (1) ou não (0).

A partir desses atributos foram construídos dois conjuntos de dados e, então, dois modelos de predição para substituição da equipe visitante: um para prever se a segunda substituição foi efetiva e outro para prever a efetividade da terceira substituição.

No primeiro modelo, o rótulo a ser predito é y_2. A ideia é que, por meio da análise das informações previamente conhecidas referente a primeira substituição e o saldo de gols atual G_2, o modelo classifique a substituição candidata S_2, que ocorrerá no tempo T_2, como efetiva ou não. Além dos atributos propostos, este conjunto possui os atributos G_1, S_1 e T_1 que caracterizam os dados da primeira substituição, o que totaliza 11 atributos.

No segundo modelo, o rótulo a ser predito é o y_3. A ideia é que, por meio do uso das informações previamente conhecidas referentes a primeira e a segunda substituições, o modelo classifique a substituição candidata S_3 que ocorrerá no tempo T_3, como efetiva ou não. Como exemplo, a Tabela 2 apresenta um extrato deste conjunto de dados. Perceba que os atributos retidos para construção do modelo foram praticamente os mesmos do Conjunto I mais os dados da terceira substituição: G_3, S_3 e T_3. Neste extrato, a segunda tupla, por exemplo, apresenta uma terceira substituição não efetiva. Essa tabela mostra que o time visitante estava ganhando por um gol, atributos *G1*, *G2* e *G3*, e, provavelmente, depois da terceira substituição, o time da casa empatou ou venceu. Sendo que as substituições realizadas não foram de ordem tática, todas foram do tipo 2 (veja atributos *S1*, *S2* e *S3*).

Após a definição dos conjuntos de dados, foi realizado um experimento para classificar os atributos mais informativos (*i.e.*, os mais importantes) para construir os modelos. Para tal, foi utilizada o método `selectKBest` da biblioteca *scikit-learn* (Buitinck et al., 2013). A Tabela 3 apresenta a classificação da importância dos atributos em cada conjunto de dados. Perceba que dos cinco primeiros atributos, os dois conjuntos compartilham três: *RV*, *RC* e *DOD*. Os dois primeiros indicam o favorecimento do time da casa ou visitante a vencer, e o último representa a diferença da força ofensiva dos dois times. Nos cinco últimos, os dois conjuntos compartilham os atributos *DDO* e *FCO*, que indicam, respectivamente, a diferença da força defensiva dos times e força ofensiva do time da casa.

Tabela 3: Classificação dos atributos quanto à importância nos dois conjuntos de dados.

Conjunto I		Conjunto II	
Atributo	Score	Atributo	Score
FVO	16.7	T2	19.4
RV	14.6	DOD	17.7
RC	14.6	G3	15.3
DOD	14.5	RC	14.9
G1	10.1	RV	14.9
S2	8.1	G1	14.7
G2	7.5	FVO	14.5
T2	5.3	T1	11.8
FVD	4.8	FVD	9.5
FCD	4.8	FCD	9.5
T1	4.0	G2	9.4
DDO	3.9	S2	6.0
S1	0.5	T3	4.7
FCO	0.2	DDO	2.5
		FCO	1.1
		S3	0.5
		S1	0.2

Com o ordenamento da importância dos atributos, um novo experimento foi realizado para identificar quais combinações de atributos seriam mais determinantes para a construção dos modelos. O experimento iniciou com os dois primeiros atributos, depois os três primeiros, até todas as combinações serem testadas. Durante esse experimento, também foram testados os melhores hiperparâmetros para cada classificador utilizando o método `GridSearchCV` também da biblioteca *scikit-learn*. A Tabela 4 apresenta os melhores valores dos hiperparâmetros considerados para cada classificador depois de executar o processo de ajuste (tuning) do `GridSearchCV`. Perceba que cada classificador possui um conjunto diferente de hiperparâ-

metros (a documentação da *scikit-learn* detalha a função de cada um deles).

A execução dos experimentos para identificar os melhores atributos para os modelos finais levou às seguintes configurações: (*i*) o Conjunto I é composto pelos seis atributos mais informativos: *FVO*, *RV*, *RC*, *DOD*, *G1* e *S2*, e (*ii*) no Conjunto II apenas o atributo *S1* não foi selecionado.

Tabela 4: Melhores valores por hiperparâmetro retornados pelo *GridSearchCV*.

Classificador	*Melhores valores*
RFC	*'criterion': 'entropy', 'max_depth': 5, 'max_features': 'sqrt', 'n_estimators': 1000*
DTR	*'criterion': 'gini', 'max_depth': 3, 'min_samples_leaf': 5*
SVM	*'C': 1, 'decision_function_shape': 'ovo', 'gamma': 'auto', 'kernel': 'rbf', 'shrinking': True*
KNN	*'metric': 'manhattan', 'n_neighbors': 19, 'weights': 'uniform'*

6 Resultados

Para avaliação dos Conjuntos I e II foram utilizadas as métricas de acurácia (A), precisão (P), revocação (R) e F_1-*score* (F_1). A validação ocorreu exclusivamente com 30% do conjunto de dados, o que significa que foram utilizados dados de 398 jogos. Estes dados compõem o conjunto de teste, são independentes e não participaram da fase de treinamento do modelo. Essa medida é importante para que o aprendizado não se torne tendencioso.

Os resultados das métricas para cada um dos modelos gerados pelos classificadores e conjunto de dados são apresentados na Tabela 5. O método de segmentação por estratificação realizou a criação do conjunto de teste automaticamente. O Conjunto I é composto por 153 exemplos da classe 0 (substituições não efetivas) e 245 exemplos da classe 1 (substituições efetivas). Já para o Conjunto II há uma pequena diferença: são 146 exemplos da classe 0 e 252 exemplos da classe 1. Cada algoritmo foi executado dez vezes e as tabelas apresentam as médias dos resultados das métricas. Nas Figuras ?? e ??, esses resultados são apresentados pictoricamente através de gráficos de barra.

Observando novamente a Tabela 5, verifica-se que os melhores desempenhos dos modelos foram aqueles aplicados no Conjunto II. Esse resultado é esperado pois o Conjunto II armazena mais dados do andamento de uma partida, o que permite o classificador fazer uma melhor generalização na criação do modelo. No caso de uma partida de futebol, a melhor métrica a ser utilizada é a precisão, pois ela mede a qualidade com que o modelo cobriu os exemplos positivos e evita gerar falsos positivos, ou seja, substituições que não foram efetivas, mas que o modelo considerou efetiva. Já a acurácia não se caracteriza como uma boa métrica neste contexto, pois nenhum dos conjuntos de dados é balanceado. Por exemplo, no Conjunto I quase 62% dos exemplos são de substituições efetivas (*i.e.*, 245 de 398), assim, um modelo que retornasse apenas a classe efetiva já possuiria esse valor como acurácia.

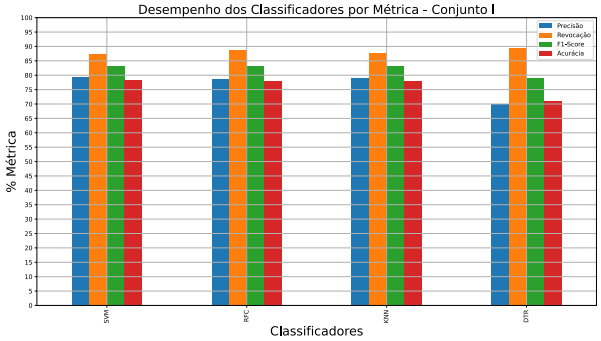

Figura 3: Desempenho Classificadores no Conjunto I.

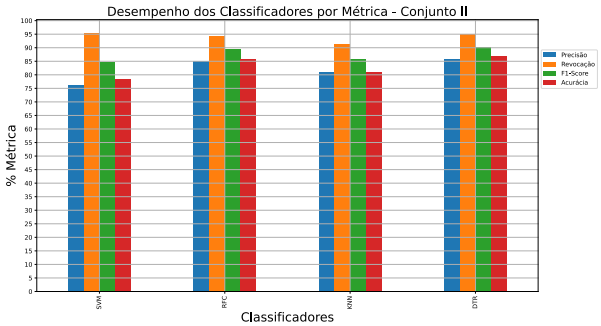

Figura 4: Desempenho Classificadores no Conjunto II.

Dentre os desempenhos do Conjunto I (vide Figura ??), a maioria dos classificadores geraram modelos satisfatórios, sendo que, em termos de precisão, excetuando *DTR*, os desempenhos foram praticamente iguais. Vale ressaltar que os modelos podem ser úteis aos treinadores das equipes visitantes, pois podem ajudá-los na tomada de decisão de quando realizar a segunda substituição. Basta informar os valores para T_2 (tempo da segunda substituição) e S_2 (tipo tático da substituição) e observar o resultado gerado. Nesse caso, se o modelo retornar que a substituição será efetiva, o treinador teria 79.5% na precisão da resposta o que seria um bom indicativo para tomada de decisão.

No caso do Conjunto II (vide Figura ??), os modelos criados com classificadores baseados em árvores tiveram um desempenho superior aos demais, pois chegaram a 86% de precisão. Com base no raciocínio anterior, o treinador poderia simular, através de mudanças dos atributos T_3 e S_3,

Tabela 5: Tabela de Resultados de Execução do Conjuntos I e II.

	Conjunto I				Conjunto II			
Classificador	P (%)	R(%)	F_1(%)	A(%)	P(%)	R(%)	F_1(%)	A(%)
SVM	79.5	87.3	83.3	78.4	76.3	95.6	84.9	78.4
RFC	78.6	88.6	83.3	78.1	85.0	94.4	89.5	86.0
KNN	79.0	87.8	83.2	78.1	81.2	91.3	86.0	81.2
DTR	70.1	89.4	79.1	70.9	86.0	94.9	90.2	87.0

o melhor momento e o melhor tipo de substituição. Nesse caso, a decisão poderia ser tomada com 86% de precisão.

Perceba também que o resultado da revocação foi significativo: em torno de 90% para o Conjunto I e 95% para o Conjunto II. Porém, no contexto da efetividade da substituição, a revocação não é um bom indicador, pois não computa os falsos positivos, apenas o percentual de positivos verdadeiros preditos *versus* os perdidos (falsos negativos). Neste caso, a métrica F_1-Score apresenta uma média ponderada entre a precisão e a revocação, ou seja, considera tantos os falsos positivos quanto os falsos negativos. No Conjunto I, a F_1-Score teve desempenho semelhante para os modelos: 83%. Já para o Conjunto II, o modelo gerado pela DTR teve o melhor desempenho: 90%.

Com base nos experimentos conduzidos é possível afirmar que os modelos propostos para prever a efetividade das substituições apresentam um bom grau de segurança (vide resultado das métricas). Assim, os modelos se tornam um boa ferramenta para auxiliar os treinadores a tomarem decisões durante uma partida de futebol no momento de realizar substituições.

7 Conclusão

Este trabalho teve como objetivo explorar as substituições propostas pelo time visitante nos anos de 2015 a 2018 do Campeonato Brasileiro de Futebol da Série A. Por meio do uso de dados históricos disponíveis no conjunto de dados, avaliou-se a possibilidade da previsão da efetividade da segunda e terceira alterações dos times visitantes, a partir do uso de algoritmos de aprendizagem de máquina.

Dois conjuntos de dados foram construídos, o primeiro armazena dados utilizados na previsão da efetividade da segunda substituição e o segundo na previsão da terceira substituição. Por intermédio das métricas de acurácia, precisão, revocação e F_1-Score, foi possível avaliar os classificadores quanto a sua qualidade. Dentre os algoritmos testados, o primeiro conjunto de dados não apresentou diferença significativa entre os desempenhos e alcançou uma precisão de quase 80%. Já no segundo conjunto, os algoritmos baseados em árvores foram os melhores: chegaram a 86% de precisão. Pode-se concluir, também, que os atributos propostos tiveram um papel importante no desempenho dos modelos.

Portanto, o resultado deste trabalho mostra-se promissor para ser aplicado em situações reais e auxiliar na tomada de decisão sobre as substituições durante uma partida de futebol. Como trabalhos futuros, pode-se citar: adaptar o modelo para prever a eficiência de substituições do time da casa, criar um fator de força para o jogadores envolvidos na substituições e inclusão de dados dos principais campeonatos de outros países no modelo.

Referências

Allmers, S. and Maennig, W. (2009). Economic impacts of the fifa soccer world cups in france 1998, germany 2006, and outlook for south africa 2010, *Eastern economic journal* **35**(4): 500–519. https://doi.org/10.1057/eej.2009.30.

Brillinger, D. R. (2011). Soccer/world football, *in* L. C. Cochran, P. Keskinocak, J. Kharoufeh and J. Smith (eds), *Wiley Encyclopedia of Operations Research and Management Science*, American Cancer Society. https://doi.org/10.1002/9780470400531.eorms0791.

Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., Niculae, V., Prettenhofer, P., Gramfort, A., Grobler, J., Layton, R., VanderPlas, J., Joly, A., Holt, B. and Varoquaux, G. (2013). API design for machine learning software: experiences from the scikit-learn project, *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pp. 108–122.

CBF (2017). Regulamento geral das competições. Disponível em https://bit.ly/2MoUfuy.

Chaudhuri, S. and Dayal, U. (1997). An overview of data warehousing and olap technology, *ACM Sigmod record* **26**(1): 65–74. https://doi.org/10.1145/248603.248616.

Del Corral, J., Barros, C. P. and Prieto-Rodriguez, J. (2008). The determinants of soccer player substitutions: a survival analysis of the spanish soccer league, *Journal of Sports Economics* **9**(2): 160–172. https://doi.org/10.1177/2F1527002507308309.

Duarte, D. and Ståhl, N. (2019). Machine learning: a concise overview, *Data Science in Practice*, Springer, pp. 27–58. https://doi.org/10.1007/978-3-319-97556-6_3.

Dunmore, T. and Murray, S. (2013). *Soccer for dummies*, John Wiley & Sons.

Flôres, F. S., dos Santos, D. L., Carlson, G. R. and Gelain, E. Z. (2019). What can coaches do? the relationship between substituion and results of professional soccer matches, *RBFF-Revista Brasileira de Futsal e Futebol* **11**(43): 215–222. http://www.rbff.com.br/index.php/rbff/article/view/745.

Gomez, M.-A., Lago-Peñas, C. and Owen, L. A. (2016). The influence of substitutions on elite soccer teams' performance, *International Journal of Performance Analysis in Sport* **16**(2): 553–568. https://doi.org/10.1080/24748668.2016.11868908.

Guyon, I., Gunn, S., Nikravesh, M. and Zadeh, L. A. (2008). *Feature extraction: foundations and applications*, Vol. 207, Springer. https://doi.org/10.1007/978-3-540-35488-8.

Hirotsu, N. and Wright, M. (2002). Using a markov process model of an association football match to determine the optimal timing of substitution and tactical decisions, *Journal of the Operational Research Society* **53**(1): 88–96. https://doi.org/10.1057/palgrave.jors.2601254.

Jamieson, J. P. (2010). The home field advantage in athletics: A meta-analysis, *Journal of Applied Social Psychology* **40**(7): 1819–1848.

Kumar, G. (2013). *Machine Learning for Soccer Analytics*, PhD thesis, KU Leuven. https://doi.org/10.13140/RG.2.1.4628.3761.

Morrow, S. (2003). The people's game, *Football, finance and society* . https://doi.org/10.1057/9780230288393.

Myers, B. R. (2012). A proposed decision rule for the timing of soccer substitutions, *Journal of Quantitative Analysis in Sports* **8**(1). https://doi.org/10.1515/1559-0410.1349.

Rein, R. and Memmert, D. (2016). Big data and tactical analysis in elite soccer: future challenges and opportunities for sports science, *SpringerPlus* **5**(1): 1410. https://doi.org/10.1186/s40064-016-3108-2.

Rey, E., Lago-Ballesteros, J. and Padrón-Cabo, A. (2015). Timing and tactical analysis of player substitutions in the uefa champions league, *International Journal of Performance Analysis in Sport* **15**(3): 840–850. https://doi.org/10.1080/24748668.2015.11868835.

Silva, R. M. and Swartz, T. B. (2016). Analysis of substitution times in soccer, *Journal of Quantitative Analysis in Sports* **12**(3): 113–122. https://doi.org/10.1515/jqas-2015-0114.

UEFA (2018). 2018/19 UEFA Champions League regulations. Disponível em https://bit.ly/3aPgeEC.