Knowledge Discovery in Brazilian Soccer Championship Scout Data

Luís Felipe Corrêa Ortolan[1], **Diego Furtado Silva**[1]

[1]Departamento de Computação – Universidade Federal de São Carlos

luis.ortolan@estudante.ufscar.br,diegofs@ufscar.br

Abstract. *Sports analytics, also known as scout, has gained significant attention thanks to the positive results of its application in a wide variety of sports. Due to the popularity of football, several studies seek to apply Machine Learning on scout data in this sport due to the popularity of soccer. This type of data comprises events countings, such as passes and finishes, and can assist the technical staffs' decision-making process. However, these efforts are focused on European football. In this work, we investigate the knowledge discovery using Machine Learning algorithms in data from scout obtained in Brazilian soccer. With that, we show the current potential and limitations of this approach.*

Resumo. *A análise de dados esportivos, conhecida como scout, tem ganhado grande atenção graças aos resultados positivos de sua aplicação em uma grande variedade de esportes. Devido à popularidade do futebol, há diversos trabalhos que buscam aplicar o Aprendizado de Máquina em dados de scout nesse esporte. Esses dados compreendem contagem de eventos, como passes e finalizações, e podem auxiliar o processo de tomada de decisão por parte de equipes técnicas. No entanto, esses esforços se concentram no futebol europeu. Neste trabalho, investigamos a obtenção de conhecimento por algoritmos de Aprendizado de Máquina em dados de scout obtidos no futebol brasileiro. Com isso, mostramos o potencial e as limitações atuais dessa abordagem.*

1. Introdução

A análise de dados esportivos não é uma área de estudos recente [Alamar 2013]. Em alguns esportes, como o futebol, ela é feita há décadas [Pollard 1986]. Esse tipo de análise, conhecida como *scout*, tem ganho força em diferentes esportes nos últimos anos. Para realizá-la, é necessário contabilizar determinados eventos durante uma partida ou treino, como um chute a gol no futebol ou um arremesso no basquete. Posteriormente, esses dados são utilizados a fim de melhorar o rendimento individual dos atletas ou da equipe.

O *scout* pode ser definido em duas etapas. A primeira consiste na coleta de dados sobre o desempenho de atletas. A aquisição desses dados pode ser feita tanto por um ser humano, que anota todas as ocorrências de determinados eventos, quanto pelo uso de máquinas capazes de detectar cada categoria definida.

Como os dados obtidos não têm em si uma classificação, também faz parte do *scout* uma segunda etapa que consiste na análise desses dados, buscando uma melhor compreensão do desempenho individual ou coletivo dos atletas. O foco da análise é buscar quais fundamentos em que se deve aperfeiçoar o rendimento, tendo impacto no treinamento e planejamento tático da equipe.

Historicamente, a obtenção de conhecimento a partir desse tipo de dados é realizada manualmente. Porém, com a ascensão de ferramentas computacionais para coleta e armazenamento, essas equipes passaram a recorrer a técnicas para tornar a análise esportiva mais prática em grandes volumes de dados, como a mineração de dados [Schumaker et al. 2010, Brefeld and Zimmermann 2017].

Graças à popularidade do esporte, não é incomum encontrar trabalhos que utilizam técnicas de mineração de dados e aprendizado de máquina para extrair conhecimento a partir de dados de futebol [Berrar et al. 2019]. Trabalhos nessa área vão de tarefas como predição de desempenho futuro de atletas [Arndt and Brefeld 2016] à prevenção de lesões [Rossi et al. 2018].

Entretanto, esses esforçam se concentram majoritariamente no futebol europeu. Por exemplo, um trabalho recente chamou a atenção por apresentar um conjunto de dados abertos e detalhados de eventos (passes, finalizações, entre outros) em partidas de futebol, contendo o tempo e a posição de cada um desses eventos [Pappalardo et al. 2019]. No entanto, as partidas contidas nesse conjunto contemplam apenas as ligas da Alemanha, Espanha, França, Itália e Inglaterra.

Neste trabalho, buscamos responder se técnicas de Aprendizado de Máquina bem difundidas são capazes de obter conhecimento a partir de dados do futebol brasileiro. Para isso, utilizamos o melhor e mais completo conjunto de dados público no domínio, extraído do *fantasy game* CartolaFC. A partir desses dados, mostramos como a classificação e a regressão podem ser úteis para enteder a importância de diferentes fundamentos do futebol para o desempenho de atletas ou da equipe. Além disso, mostramos que regras de associação podem auxiliar a compreender a relação entre essas características.

2. Conjunto de Dados

Visto que este trabalho tem interesse no futebol brasileiro, utilizamos um conjunto de dados específico para esse domínio. O conjunto utilizado é composto pelos dados de 1.340 partidas entre os anos de 2014 e 2017 do site CartolaFC[1].

Para cada ano, o conjunto CartolaFC é organizado em quatro arquivos distintos. O primeiro arquivo é uma lista com atributos como nome e *id*, de cada equipe participante da primeira divisão do Campeonato Brasileiro naquele ano. Há também uma lista com todos os atletas que jogaram nesse campeonato naquele ano, além dos seus *ids*, nomes, clube em que jogam e sua posição. Uma lista com todos os jogos realizados naquele ano e seus resultados compõem um terceiro arquivo.

Porém, um quarto arquivo é o mais utilizado nesta pesquisa. Cada linha contempla os dados de *scout* de um determinado jogador em uma determinada partida. Os dados de *scouts* contados neste conjunto de dados são: faltas sofridas (`FS`), passes errados (`PE`), assistências (`A`), finalizações na trave (`FT`), finalizações defendidas (`FD`), finalizações fora do gol (`FF`), gols (`G`), impedimentos (`I`), pênaltis perdidos (`PP`), roubadas de bola (`RB`), faltas cometidas (`FC`), gols contra (`GC`), cartões vermelhos (`CV`), cartões amarelos (`CA`), se o time não tomou nenhum gol (`SG`), defesas difíceis (`DD`), pênaltis defendidos (`DP`) e gols sofridos (`GS`). Alguns desses atributos possuem valor diferente de zero apenas para posições específicas. Esse é o caso do atributo (`SG`) para defensores e os atributos (`DD`),

[1] `https://globoesporte.globo.com/cartola-fc`

(DP) e (GS) para os goleiros. Considerando todos esses atributos, o conjunto CartolaFC contempla cerca de 110.000 anotações de eventos associados a partidas de futebol.

Além desses dados de *scout*, o conjunto ainda apresenta algumas informações relacionadas ao *fantasy game* CartolaFC: o status do jogador antes da partida (machucado, disponível, dúvida), a pontuação do jogador no jogo naquela partida, o seu preço atual no jogo, a variação no preço do jogador naquela partida e a média de variação de preço daquele jogador durante todo o ano.

3. Técnicas Utilizadas

O conjunto de dados do CartolaFC está disponível no repositório e plataforma de competições Kaggle[2]. Por isso, além de artigos utilizando esses dados, há contribuições na forma de *notebooks* técnicos na própria plataforma Kaggle. Em ambos os casos, artigos e *notebooks*, os esforços se encontram nas tarefas de visualização, previsão de resultados e auxílio na escalação do time para otimizar o desempenho no *fantasy game* [Mota et al. 2018, Santos 2019].

Dessa maneira, os esforços deste trabalho se voltam à verificação da utilidade desse tipo de dados de *scout* no futebol brasileiro. Para isso, utilizamos técnicas comuns e bem difundidas de Aprendizado de Máquina para a descoberta de conhecimento no conjunto de dados CartolaFC. Especificamente, utilizamos técnicas de classificação, regressão, descoberta de regras de associação e agrupamento. A seguir, descrevemos como configuramos experimentos para cada uma dessas tarefas. Por motivos de espaço, não descrevemos o funcionamento de cada um dos algoritmos utilizados. No entanto, são algoritmos bem conhecidos e descritos na literatura de Aprendizado de Máquina.

Em todas as tarefas, com exceção da descoberta de regras de associação, foi utilizada a ferramenta *sckit learn* [Pedregosa et al. 2011]. Usualmente, os valores de hiperparâmetros utilizados foram o padrão da ferramenta. Nos casos em que isso não acontece, deixamos claro como escolhemos esses valores.

3.1. Classificação

Para a tarefa de classificação, buscamos compreender como os dados de *scout* influenciaram positiva ou negativamente para o desempenho da equipe nas partidas. Para isso, utilizamos o classificador *Random Forest* (RF), por representar um bom compromisso entre interpretabilidade e acurácia.

Para esse objetivo, criamos um conjunto de dados em que cada exemplo agrega os *scouts* de atletas do "time da casa" e equipe visitante. Como atributo-alvo, utilizamos o vencedor da partida. Ou seja, as classes se referem à vitória do time da casa, do visitante ou ao empate.

Para se chegar ao conjunto de dados agregado, utilizamos duas estratégias distintas. A primeira agrega os dados de todos os atletas pela soma. Por exemplo, para agregar as faltas sofridas, na partida do time A (casa) contra o time B (visitante), há os atributos FS_home e FS_away, respectivamente. Para obter seus valores, são somados os valores de FS de cada jogador da equipe A e da equipe B, respectivamente, observados naquela partida.

[2]https://www.kaggle.com/schiller/cartolafc

Além disso, avaliamos uma segunda estretégia de agregação. Nela, separamos os *scouts* por posição. As posições existentes no conjunto de dados são goleiro, lateral, zagueiro, meia, atacante e técnico. Esse último foi removido do conjunto de dados.

É importante notar que, além da agregação dos dados para formar o conjunto descrito, foi realizada uma limpeza nos dados para garantir o objetivo da tarefa. Uma vez que atributos como número de gols do visitante e da equipe da casa estão diretamente relacionados ao atributo alvo, eles foram removidos do conjunto de dados. Portanto, para cada uma das equipes, foram excluídos os atributos `A`, `G`, `GC`, `GS` e `SG`.

3.2. Regressão

Assim como na tarefa de classificação, utilizamos o RF como método de indução do modelo. Adicionalmente, utilizamos a regressão linear. Por outro lado, a regressão foca no desempenho individual dos atletas. Ou seja, a entrada para o modelo é um exemplo em sua forma primária, contendo o *scout* de um jogador em uma partida.

Como atributo-alvo, utilizamos duas características diferentes. A primeira dela é a pontuação do CartolaFC. Apesar desse valor ser obtido por uma equação definida e aberta ao público, utilizá-la como atributo-alvo nos permite analisar a capacidade de algoritmos de Aprendizado de Máquina em replicar a pontuação sem conhecimento do modelo definido previamente.

O segundo atributo-alvo é a nota dada por comentaristas do Globo Esporte[3]. Essa note é interessante por ser um critério que carrega uma subjetividade que pode, muitas vezes, não estar refletida nos dados. Além disso, ela garante que a técnica empregada tem um significativo poder de generalização, análise que pode ser enviesada utilizando-se apenas a pontuação do atleta no *fantasy game* CartolaFC. Porém, essa informação só está presente nos dados do ano 2014.

Acreditamos que a regressão desses atributos-alvo seja importante para refletir quais ações fazem com que um atleta se destaque positiva ou negativamente. Isso pode ser utilizado para guiar o treinamento individual desses atletas.

3.3. Descoberta de Regras de Associação

A principal dificuldade para se modelar o problema como classificação e regressão é o fato delas serem tarefas de aprendizado supervisionado. Assim, temos que definir atributos-alvo para elas. No entanto, acreditamos que esses atributos não estejam presentes na aplicação prática de Aprendizado de Máquina para dados de *scout*. Por isso, realizamos duas tarefas de aprendizado não supervisionado.

A primeira delas é a tarefa de descoberta de regras de associação. Por meio desse tipo de regras, é possível encontrar padrões que se repetem ao longo de diversas partidas. Por exemplo, podemos associar um número baixo de roubadas de bola a um número maior que zero de gols sofridos. Assim como em qualquer domínio de aplicação, podemos descobrir regras "óbvias", como associar que um jogador que não foi titular não fez nenhum gol, ou pouco interessantes, como relacionar nenhuma falta sofrida com nenhuma finalização na trave. Porém, acreditamos que essa abordagem pode trazer regras interessantes e pouco ou nada conhecidas.

[3] `https://globoesporte.globo.com/`

Para as regras de associação, foi utilizada a biblioteca *mlxtend*[4]. Para obter as regras de associação, utilizamos o algoritmo apriori, implementado nessa ferramenta. Para evitar regras espúrias e que sejam apenas exceções, definimos suporte mínimo de $0,2$ e confiança mínima de $0,8$ para a extração das regras de associação.

As regras de associação foram obtidas a partir de subconjunto de dados compostos pelos *scouts* de atletas de cada posição. Além dos dados de *scout*, foram utilizados a pontuação, nota do GloboEsporte, se o atleta pertencia ao time com mando de campo e se foi titular na partida.

Uma vez que o algoritmo Apriori não lida com atributos numéricos e nominais, mas sim com a ocorrência ou não de um elemento (item) em cada exemplo (transação), os dados foram transformados a fim de se adaptar a essa representação. Para isso, os dados foram discretizados em faixas com semânticas interessantes a cada atributo. Por isso, foram utilizadas diversas faixas de valores para a discretização, cujo valor foi escolhido empiricamente. Além disso, observamos que a discretização por quantis retornou intervalos muito pequenos entre quantis consecutivos. Por isso, foi utilizada a binarização com divisão uniforme de intervalo entre as faixas.

A primeira foi aplicada sobre a pontuação de um atleta e o número de jogos que o atleta participou anteriormente. Esses dados foram divididos em cinco faixas, representando a semântica de muito baixo, baixo, médio, alto e muito alto. Essas faixas são identificadas por 0, 1, 2, 3 e 4, respectivamente. Para os atributos relativos a nota do GloboEsporte e tempo jogado, escolhemos utilizar três valores (0, 1 e 2), representando baixo, médio e alto. Por fim, para os demais atributos, foram utilizadas sete faixas indexadas por valores de 0 a 6, representando baixíssimo, muito baixo, baixo, médio, alto, muito alto, altíssimo. As únicas exceções são os atributos binários, que não sofreram quaisquer alterações. São eles: titularidade e cartões vermelhos e amarelos (`CV` e `CA`).

Após isso, os dados são submetidos à trasnformação *one-hot encoding*. Com isso, os atributos são separados em cada faixa da discretização e binarizados. Por exemplo, em vez do atributo `FF` conter o número de finalizações para fora, ele é transformado em sete atributos binários, representando as diferentes faixas obtidas com a discretização. Ou seja, se o atleta teve um número médio de finalizações para fora, os atributos derivados de `FF` conterão todos o valor zero, com exceção daquele que representa a faixa de número médio de finalizações para fora.

3.4. Agrupamento

Para avaliar a aplicação da tarefa de agrupamento, focamos no agrupamento de partidas. Com isso, poderíamos analisar as semelhanças entre partidas no mesmo grupo e as diferenças nos dados que fazem determinadas partidas estarem em grupos distintos. Por exemplo, poderíamos encontrar que partidas em um mesmo grupo são compostas de vitória de uma determinada equipe e tiveram um número alto de finalizações de seus atacantes. Com isso, poderíamos encontrar algumas relações ocultas nos dados.

Para utilizar algoritmos de agrupamento, realizamos a agregação de dados para contemplar partidas. Assim como fizemos para a tarefa de classificação, a agregação dos dados de *scout* foram pela soma dos valores de todos os atletas para formar um novo

[4] `http://rasbt.github.io/mlxtend`

conjunto de dados e agregados por diferentes posições, para um segundo conjunto. Porém, as duas equipes de uma partida são separadas em objetos diferentes. Ao final, temos um conjunto de dados em que cada linha representa os dados contabilizados de todos os atletas de uma equipe em uma partida. Além dessas transformações, ainda foi criada uma coluna que indicava o resultado da partida para aquela equipe.

Depois da etapa de preparação dos dados, aplicou-se um algoritmo k-means nos dois conjuntos de dados para tentar agrupar os exemplos. Uma vez que esse algoritmo depende da definição de um número de grupos, utilizamos duas técnicas para a escolha do valor para esse hiper-parâmetro. Para isso, avaliamos o grupo com menor valor de silhueta, bem como o ponto de inflexão da curva de erro quadrático médio.

4. Resultados e Discussão

Nesta seção, serão apresentados os resultados obtidos pelas técnicas descritas. Por motivos de espaço, resumimos os resultados obtidos e os discutimos brevemente. No entanto, criamos um *website* complementar a este artigo[5]. Nele, há resultados detalhados para cada uma das tarefas e um link para o repositório contendo todos os códigos utilizados para a obtenção dos resultados.

4.1. Classificação

O resultado da classificação possui dois pontos a serem observados: a eficácia do modelo e a importância que os atributos tiveram na sua construção. A primeira análise é importante para compreender se o algoritmo RF é capaz de aprender relações entre os atributos que sejam relevantes para definir o ganhador de uma partida. O segundo é útil para que a equipe técnica compreenda essas relações, auxiliando, assim, a tomada de decisão.

Os resultados da classificação, no entanto, não chegaram perto do ótimo. Para chegarmos a essa conclusão, avaliamos a acurácia obtida pelo RF. No caso da agregação dos *scouts* para cada posição dos jogadores (vide Seção 3.1), a acurácia foi de $38,95\%$. Além de um desempenho próximo ao classificador aleatório, esse conjunto de dados possui muitos atributos, dificultando a compreensão da importância de cada um deles.

No caso de agregação de atributos para toda a equipe, os resultados são melhores. A acurácia obtida, apesar de ainda longe do ótimo, foi de $52,63\%$. Ainda, o número de atributos é consideravelmente menor, facilitando a interpretação do resultado. Na Figura 1, são apresentadas as importâncias de cada atributo nesse experimento.

Notamos que os atributos mais importantes não são necessariamente os mais intuitivos. Atributos relacionados a finalizações, especificamente finalizações para fora(`FF`), aparecem entre os atributos com maiores importâncias. Porém, os atributos considerados mais relevantes foram passes errados (`PE`) e roubadas de bola (`RB`). Também são importantes atributos de faltas sofridas (`FS`) e cometidas (`FC`). Por outro lado, atributos que raramente são diferentes de zero, como pênaltis perdidos e defendidos (`PP` e `PD`) aparecem como os menos importantes.

Observações com essas podem ser importantes para a tomada de decisão. Por exemplo, com os resultados obtidos, pode-se interpretar que melhorar fundamentos defensivos e passes tendem a ser positivamente decisivos para as equipes. Apesar desse

[5] `https://sites.google.com/view/cartolafc-kdd`

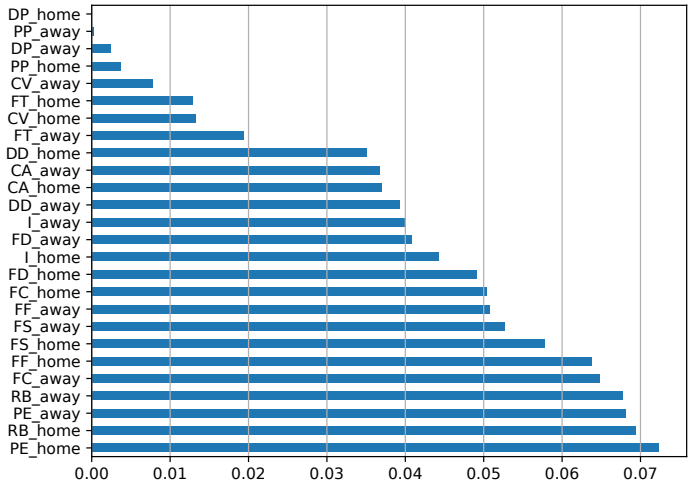

Figura 1. Importância de cada atributo para a classificação de ganhador da partida utilizando RF.

resultado ser em nível de campeonato, ele mostra que é um tipo de análise que pode ser muito relevante para decisões a nível de equipe. Ainda, que o mesmo tipo de análise pode ser replicado em outros níves de abstração para encontrar conhecimento interessante.

4.2. Regressão

Os primeiros resultados observados para a tarefa de regressão são relativos à pontuação do site CartolaFC aos jogadores. Em nossos experimentos, notamos uma grande facilidade de técnicas de regressão em replicar o modelo adotado pelo jogo. Em um primeiro momento, a técnica de Regressão Linear foi usada para checar a viabilidade da tarefa. Os resultados mostraram um coeficiente de determinação (R^2) de $0,99$ e um erro médio quadrático (MSE) de apenas $3,65 \times 10^{-}5$.

Uma forma de observar o quanto o modelo induzido se aproxima daquele pré-definido pela equipe do *fantasy game*, é por meio da comparação entre os coeficientes do modelo induzido e as os valores utilizados pelo CartolaFC para definir a pontuação. Na Figura 2, é apresentado um gráfico em que cada ponto representa uma coordenada (x, y) em que x é a nota somada à pontuação para um determinado dado de *scout* e y é o valor do coeficiente induzido.

Como notado na Seção 3.1, pré-determinamos a utilização do regressor RF para esta tarefa. Nesse caso, os resultados foram um ligeiramente inferiores, com R^2 igual a $0,96$ e MSE na ordem de $1,2x10^{-}4$. Porém, ainda conseguem aproximar muito bem o modelo utilizado pelo CartolaFC. Nota-se que a pontuação dos atletas no conjunto de dados possui alta concentração entre -2 e -1, mas varia de -12 a $21,9$.

Apesar de ser possível replicar o modelo do CartolaFC e dos bons resultados da regressão, percebeu-se que os pontos do jogo não condiziam com um jogador ter jogado bem a partida ou não de acordo com a nota do site GloboEsporte.com. Para avaliar a generalização da regressão para outros cenários, aplicamos os mesmos regressores à nota dada por esses especialistas. Visto que há uma grande subjetividade quando um ser humano atribui notas aos jogadores de acordo com seu entendimento em relação ao desempenho dos atletas, os resultados são significativamente inferiores.

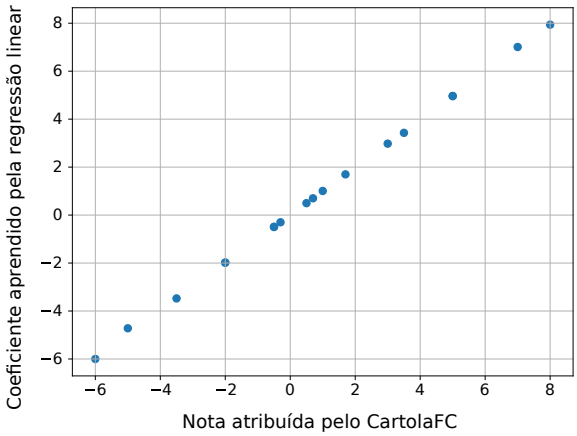

Figura 2. Comparação entre os coeficientes induzidos pela regressão linear e a pontuação de cada *scout* no CartolaFC.

Quando aplicamos o regressor em todos os jogadores, obtivemos R^2 de -0.34 e MSE de $0,59$ com a regressão linear. As mesmas medidas de avaliação resultaram em $-0,58$ e $0,66$ para o RF. Esse resultado mostra que o RF não foi capaz de encontrar quaisquer relações que levem a uma boa regressão da nota.

Para avaliar se esses resultados se devem à dificuldade de refletir o desempenho de determinadas posições, aplicamos os regressores para as posições separadamente. Os coeficientes R^2 e o erro quadrático médio para cada posição são apresentados na Tabela 1. Apesar dos coeficientes de determinação apresentarem valores baixos em alguns casos, o erro médio é relativamente pequeno, visto que as notas variam entre 0 e 10.

Tabela 1. Resultados da regressão da nota atribuída pelo GloboEsporte por posição do atleta.

		Goleiros	Laterais	Zagueiros	Meias	Atacantes
Regressão	R^2	0,38	0,26	0,22	0,28	0,45
linear	MSE	0,43	0,62	0,56	0,61	0,55
Random	R^2	-0,74	-0,96	-0,61	-0,95	-0,13
Forest	MSE	0,55	0,76	0,75	0,69	0,63

Podemos notar que todos os valores de R^2 para o regressor RF foram negativos. Por isso, voltamos a nossa análise apenas para a regressão linear. Mesmo com valores relativamente pequenos, é possível observar fatores que são intuitivos em relação à importância dada pelo regressor a cada atributo – os coeficientes da regressão linear, nesse caso. A Figura 3 mostra os coeficientes aprendidos para as posições zagueiro e atacante, que obtiveram o menor e maior valor de R^2, respectivamente.

Apesar do subajuste do regressor (observado por valores relativamente baixos de R^2), é possível observar pontos interessantes na Figura 3. Por exemplo, gols marcados e cartões vermelhos foram os mais relevantes positiva e negativamente para ambas as posições. Defesas, que são atributos do goleiro, não contam nem positiva e nem negativamente para as posições apresentadas. No entanto, note que o time não sofrer gols (`SG`) é positivamente relevante para os zagueiros e pouco negativo para os atacantes. No caso do atacante, finalizações então entre os atributos com maior valor (positivo). Há atributos

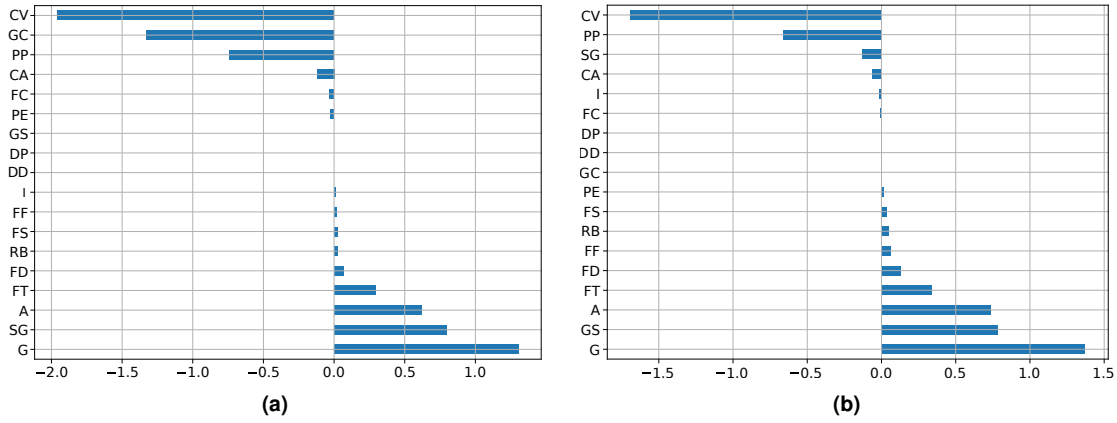

Figura 3. Coeficientes da regressão linear para a nota atribuída pelo GloboEsporte para zagueiros (esquerda) e atacantes (direita).

que possivelmente explicam o subajuste, como o GS com alta relevância para os atacantes. Mas, de modo geral, a importância dada a cada um desses atributos parece explicar bem as notas dadas pelos especialistas.

4.3. Regras de Associação

Como descrito na Seção 3.3, avaliamos a descoberta de regras de associação separando-os em subconjuntos definidos por suas posições, para melhor direcionarmos a descoberta e análise dessas regras.

Para algumas das posições, foram obtidas poucas regras. Goleiros e laterais obtiveram apenas dezenove regras cada. No entanto, um número elevado de regras foi obtido para as demais posições. Os subconjuntos de dados de zagueiros, meias e atacantes obtiveram, respectivamente, 561, 387 e 791 regras. Para evitar um número tão elevado de regras para analisar, aumentamos o valor de suporte mínimo para essa posições para $0,4$. Com isso, o número de regras reduziu para 137, 43 e 30, respectivamente.

Notamos que esses números também contam com um corte superior para o suporte de $0,9$. Isso foi realizado para evitar regras muito óbvias. Um exemplo de regra obtida sem esse filtro é $PD\_0 \Rightarrow PP\_0$, com confiança e suporte próximos a 1. Ou seja, é quase certo que se um goleiro não defendeu um pênalti, também não perdeu um pênalti. O suporte dessa regra é alto pois, com raríssimas exceções na história, um goleiro não costuma cobrar penalidades. Além disso, são poucas as partidas em que um goleiro defende um pênalti.

Ainda assim, algumas regras não interessantes foram descobertas. Por exemplo, a regra $SG \Rightarrow GS\_0$ obteve confiança $1,0$. Essa regra significa que se o time não sofreu nenhum gol, o goleiro não sofreu nenhum gol. Porém, essa é uma regra fácil de filtrar, pois SG é um atributo pensado para jogadores de linha, enquanto GS é exclusivo para goleiros. Portanto, podemos considerar apenas as regras que contenham apenas um desses itens.

Para jogadroes de linha, foi possível encontrar regras mais interessantes. Por exemplo, a regra $nota\_1 \wedge FF\_0 \Rightarrow FD\_0$ obteve suporte de $0,59$ e confiança de $0,89$ para zagueiros. Ou seja, quando sua nota é mediana e ele realizou pouquíssimas finalizações para fora, também teve poucas das suas finalizações defendidas. Uma vez que não analisamos causalidade, acreditamos que essa regra possa ser lida como um zagueiro que fez

pouquíssimas finalizações (para fora ou defendidas), comumente recebe uma nota mediana dos especialistas. Algo similar pode ser feito para atacantes. A regra $nota\_1 \Rightarrow G\_0$ tem $0,99$ de confiança para essa posição. Ou seja, um atacante que não faz gol recebe nota mediana.

Porém, notamos que a grande parte das regras se limita àquelas com átomos na primeira faixa, ou seja, quando o valor do *scout* é baixíssimo, usualmente zero. Isso se deve ao fato de que essas são as regras com maior suporte. Apesar disso, as regras trazem informação aos *scouts* mais relevantes a cada posição. Por exemplo, grande parte das regras para atacantes tem o consequente G_0, ou seja, o atacante não fez gols. A Tabela 2 mostra um exemplo desse fato.

Tabela 2. Exemplos de regras descobertas no conjunto de dados para cada uma das posições de linha: atacante (A), meia (M), lateral (L) e zaguiero (Z).

Regra de associação	Posição	Suporte	*Confiança*
$FD\_0 \Rightarrow G\_0$	A	$0,56$	$0,86$
$FF\_0 \wedge nota\_1 \Rightarrow FD\_0$	M	$0,50$	$0,82$
$SG \Rightarrow nota\_2$	L	$0,27$	$0,87$
$titular \wedge FF\_0 \wedge nota\_1 \Rightarrow tempo\_jogado\_2 \wedge FD\_0$	Z	$0,47$	$0,82$

Esse resultado mostra que a descoberta de regras de associação pode encontrar relações interessantes entre dados de *scout*. Porém, o suporte de algumas regras é altamente afetado pelas características da aplicação. Por exemplo, não marcar um gol é um evento muito mais frequente que marcar um ou mais gols. Por isso, regras que contemplam atributos como G_0 não são descobertas.

Por isso, para alcançar melhores resultados, é preciso lançar mão de algoritmos mais robustos à descoberta de regras com baixo suporte. Isso pode ser visto, por exemplo, para a regra com menor suporte na Tabela 2. Quando um time não sofre gols, um lateral recebe nota alta, com alta confiança. Apesar de que isso pareça mais intuitivo ocorrer com zagueiros, essa regra não é descoberta para tal posição.

4.4. Agrupamento

Ao realizar a tarefa de agrupamento pelo algoritmo k-means, iniciamos pela escolha do número de grupos. Para isso, medimos o erro quadrático médio obtido para todos os valores de k no intervalo de 2 a 59. No entanto, não houve um ponto de inflexão claro, como pode ser visto na Figura 4. Porém, já é possível notar que a agregação feita com detalhamento da posição dos atletas tende a ter menor erro.

A partir dessa indefinição, também observamos o menor valor de silhueta para os dois conjuntos de dados. Novamente, não há um número de grupos claramente melhor que outros. O melhor valor de silhueta, para ambas versões do conjunto de dados, foi para 2 grupos. Ainda assim, esses valores foram muito baixos. Especificamente, $0,18$ sem separação por posição e $0,17$ quando a posição é levada em conta.

Por isso, avaliamos a distribuição dos dados no espaço. Para isso, utilizamos *multidimensional scaling* para embarcar os exemplos em um espaço de duas dimensões preservando a distância relativa entre eles. Esse novo espaço é exibido na Figura 5.

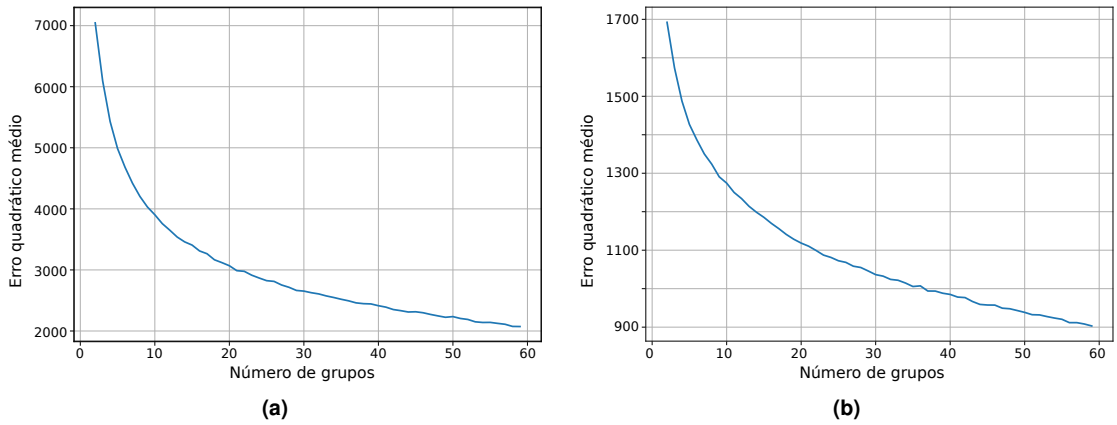

Figura 4. Erro quadrático médio por número de grupos obtidos pelo k-means para os dados agregados sem *(a)* e com *(b)* a separação por posição.

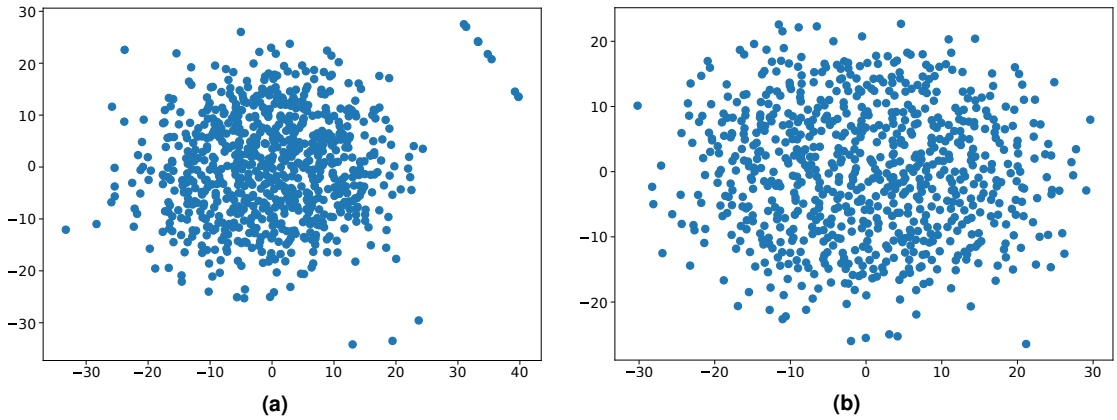

Figura 5. Projeção dos dados agregados sem *(a)* e com *(b)* a separação por posição em duas dimensões pelo algoritmo *multidimensional scaling*.

Notamos, com isso, que não há uma estrutura clara de grupos. Por isso, não é possível utilizar os *scouts* para determinar estruturas de grupos sem manipulações nos dados ou algoritmos mais robustos.

5. Considerações Finais

Os dados de *scout* têm grande relevância na modernização da análise esportiva. Por isso, é importante pesquisar por métodos computacionais que auxiliem na tomada de decisão a partir desses dados. No entanto, não há muitos conjuntos públicos de *scout* para o futebol brasileiro.

Nesse contexto, utilizamos um conjunto de dados derivado do *fantasy game* CartolaFC. Esse é um conjunto de dados relativamente simples, pois tem o objetivo de servir apenas ao jogo. Por esse motivo, algumas das técnicas investigadas não levaram a resultados interessantes.

Mesmo com essa limitação, foi possível demonstrar que o Aprendizado de Máquina pode ser utilizado com sucesso para obter conhecimento sobre dados de *scout*. Especialmente, técnicas de classificação e regressão podem mostrar a importância de fundamentos do esporte para diferentes objetivos, como o desempenho individual ou coletivo.

A descoberta de regras de associação também se mostrou como uma ferramenta com potencial para compreender a relação entre esses fundamentos. Porém, em todos os casos há um grande espaço para melhorias.

Nosso trabalho mostra o potencial de técnicas de Aprendizado de Máquina para a descoberta de conhecimento em dados de *scout*. Porém, com os dados públicos disponíveis, essas técnicas ainda ficam limitadas. Por isso, buscaremos técnicas mais robustas para conseguir melhores resultados com o conjunto de dados utilizado. No entanto, esperamos que trabalhos como este também sejam uma forma de incentivo à criação de conjuntos de dados mais completos para a realização da tarefa.

Agradecimentos

Agredecemos às instituições que, por meio do financiamento à pesquisa, tornaram possível este trabalho: processo 128268/2019-0, Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e processos nº 2017/24340-6 e 2020/07911-2, Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP).

Referências

Alamar, B. C. (2013). *Sports analytics: A guide for coaches, managers, and other decision makers*. Columbia University Press.

Arndt, C. and Brefeld, U. (2016). Predicting the future performance of soccer players. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 9(5):373–382.

Berrar, D., Lopes, P., Davis, J., and Dubitzky, W. (2019). Guest editorial: special issue on machine learning for soccer. *Machine Learning*, 108(1):1–7.

Brefeld, U. and Zimmermann, A. (2017). Guest editorial: Special issue on sports analytics. *Data Mining and Knowledge Discovery*, 31(6):1577–1579.

Mota, E., Coimbra, D., and Peixoto, M. (2018). Cartola fc data analysis: A simulation, analysis, and visualization tool based on cartola fc fantasy game. In *Proceedings of the XIV Brazilian Symposium on Information Systems*, pages 1–8.

Pappalardo, L., Cintia, P., Rossi, A., Massucco, E., Ferragina, P., Pedreschi, D., and Giannotti, F. (2019). A public data set of spatio-temporal match events in soccer competitions. *Scientific data*, 6(1):1–15.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Pollard, R. (1986). Home advantage in soccer: A retrospective analysis. *Journal of sports sciences*, 4(3):237–248.

Rossi, A., Pappalardo, L., Cintia, P., Iaia, F. M., Fernández, J., and Medina, D. (2018). Effective injury forecasting in soccer with gps training data and machine learning. *PloS one*, 13(7):e0201264.

Santos, J. M. A. d. (2019). *Previsões de resultados em partidas do campeonato brasileiro de futebol*. PhD thesis, Fundação Getúlio Vargas.

Schumaker, R. P., Solieman, O. K., and Chen, H. (2010). *Sports data mining*, volume 26. Springer Science & Business Media.