mini:2

Capítulo 2

Sports Analytics: Mudando o Jogo

Ígor Barbosa da Costa, Carlos Eduardo Santos Pires e Leandro Balby Marinho

Abstract

In the last decades, researchers have been developing different techniques to understand which factors influence the sporting results and, consequently, what is the role of predictability and randomness in sports. With the evolution of techniques to acquire, store, and process large volumes of information, Sports Analytics has become even more important for discovering new knowledge and transforming the behavior of those involved with the sport. In this chapter we present an introduction to the topic, making a historical contextualization, detailing the types of data used and discussing the process of knowledge discovery for applied research in the domain. In addition, we highlight the relationship between sports data analysis and betting markets. Finally, we present some emerging challenges for the beginning researcher and make our final considerations on this field of research.

Resumo

Nas últimas décadas, pesquisadores vêm desenvolvendo diferentes técnicas para entender quais fatores influenciam os resultados esportivos e, consequentemente, qual o papel da preditibilidade e da aleatoriedade no esporte. Com a evolução das técnicas de aquisição, armazenamento e processamento de grandes volumes de informações, as análises de dados esportivos (Sports Analytics) se tornaram ainda mais importantes para a descoberta de novos conhecimentos e vêm transformando o comportamento daqueles envolvidos com o esporte. Neste capítulo apresentamos uma introdução ao tema, fazendo uma contextualização histórica, detalhando os tipos de dados utilizados e discutindo o processo de descoberta de conhecimento para pesquisas aplicadas no domínio. Além disso, destacamos a relação entre a análise de dados esportivas e os mercados de aposta. Por fim, apresentamos uma série de desafios emergentes para o pesquisador iniciante e fazemos nossas considerações finais sobre esse campo de pesquisa.

2.1. Introdução

No século XX, os esportes passaram por diversas transformações, com destaque para o surgimento das confederações (que cumprem papel regulador) e a profissionalização dos atletas. Além disso, a globalização e a extensão do alcance das mídias sociais permitiram aos fãs acompanhar as façanhas de seus times e atletas diariamente, impulsionando a popularidade de muitos esportes.

Essa crescente popularidade dos esportes permitiu o aumento de investimentos financeiros para clubes e atletas. O esporte profissional se tornou um produto e a indústria esportiva passou a ser um dos maiores mercados de entretenimento no mundo (atualmente, as marcas ligadas ao esporte movimentam cerca 1% de todo o PIB mundial [25]). Os salários dos atletas de destaque dispararam, assim como a cobrança por melhores resultados.

Na constante busca por melhores resultados, a análise de dados esportivos, também conhecida como *Sports Analytics*, ganhou um papel de protagonismo no século XXI. Pode-se definir *Sports Analytics* como a análise de uma coleção de dados históricos que, se realizada adequadamente, pode trazer vantagem competitiva para um time ou atleta.

Em [34], Cokins et al. publicaram um artigo intitulado *"Sports Analytics taxonomy, V.1.0"* que busca classificar as diversas áreas de aplicação da análise de dados no esporte de forma mais abrangente. O trabalho sugere que os ramos da análise esportiva podem ser organizados em três grandes grupos: *Grupo 1*: refere-se ao esporte como competição, ou seja, engloba a análise de equipes, atletas e ligas; *Grupo 2*: refere-se ao esporte como recreação, ou seja, a análise é focada no desempenho e saúde do indivíduo; e *Grupo 3*: tenta superar a incerteza do esporte, ou seja, a análise é voltada para apostas esportivas, jogos de azar, ligas fantasias[1] e jogos eletrônicos.

Na maior parte deste trabalho, trataremos o esporte como competição (*Grupo 1*) e focaremos nas modalidades com interação entre os adversários, como esportes de combate (boxe, judô, MMA, etc.), esportes de campo e taco (beisebol, críquete, etc.), esportes com rede divisória ou parede de rebote (voleibol, tênis, squash, etc.) e esportes de invasão (basquete, futebol, handebol, etc.). A natureza desse tipo de esporte torna a análise de dados uma tarefa mais desafiadora, visto que existe uma diversidade de comportamentos nas estratégias dos times e atletas. Além disso, devido à popularidade dessas modalidades, a quantidade de dados disponíveis para análise está em contínuo crescimento, abrindo novas oportunidades para pesquisa.

A seguir, contextualizaremos *Sports Analytics* na história, destacando o momento em que a análise de dados começou a revolucionar os esportes. Na Seção 2.2, destacaremos um outro domínio impactado pela análise de dados esportivos: o mercado de apostas esportivas (*Grupo 3*). Mostraremos os fundamentos que relacionam *Sports Analytics* e as pesquisas aplicadas em apostas esportivas (*Sports Betting Analytics*). Na Seção 2.3, identificaremos as características dos dados utilizados em *Sports Analytics*, destacando as principais formas de coleta. Além disso, discutiremos como um pesquisador pode ter acesso a essas coleções de dados. Na Seção 2.4, detalharemos os passos da pesquisa apli-

[1] Um tipo de jogo online no qual os participantes escalam equipes imaginárias ou virtuais de jogadores reais de um esporte profissional

cada, discutindo o caminho que o dado percorre até virar uma informação útil, através de exemplos. Na Seção 2.5, apresentaremos os desafios emergentes em *Sports Analytics*. Finalmente, na Seção 2.6, apresentaremos nossas considerações finais a respeito de todo o conteúdo do capítulo.

2.1.1. Contexto Histórico

Apesar de ter recebido maior relevância neste século, a análise de dados esportivos é uma atividade relativamente antiga. Em meados do século XX, alguns pesquisadores já manipulavam dados estatísticos para tentar entender melhor sobre as características dos esportes.

No futebol, Charles Reep, reconhecido como o primeiro analista de dados desta modalidade, criou um sistema notacional para anotar cada lance que ocorria em uma partida. Cada evento do jogo recebia uma categorização detalhada. Por exemplo, para cada passe efetuado, era registrada a posição no campo onde o passe foi originado e finalizado, assim como a distância, a direção e a altura do passe. Os dados coletados durante 15 anos foram base para o artigo científico intitulado *"Skill and Chance in Association Football"* publicado em 1968 [47].

No referido artigo, Reep, juntamente com o estatístico Bernard Benjamin, buscavam entender se os dados coletados podiam revelar padrões previsíveis do esporte. Dentre as descobertas, podemos destacar a definição de que o futebol era um processo estocástico: um chute em cada oito termina em gol, mas era difícil determinar qual deles. Descobriram também que o futebol é um jogo de alternância, pois a maioria das jogadas termina após zero ou um passe completo, ou seja, em uma partida, a posse da bola é trocada, em média, quatrocentas vezes. Além disso, os autores demonstraram que 30% de todas as bolas recuperadas na grande área do adversário se transformam em finalizações ao gol e que metade dos gols eram resultado dessas mesmas bolas recuperadas.

No beisebol, Bill James desafiou a análise tradicional do jogo ao demonstrar que as estatísticas utilizadas pelas equipes para avaliar o desempenho dos atletas estavam equivocadas. Por exemplo, na análise tradicional, a média de rebatidas de um jogador era considerada um indicador importante para definir qual rebatedor deveria iniciar uma partida. Bill, por sua vez, definia que a importância de um jogador estava no modo como ele contribui para as vitórias e não nas suas estatísticas brutas. Ele afirmava que as médias de ataque não significariam nada se o rebatedor não pontuasse. Ideias como esta fizeram com que Bill publicasse uma coleção de "novas estatísticas", denominada *Sabermetrics* [4], para avaliação de atletas de beisebol.

No basquete, Dean Oliver foi um dos pioneiros. Inspirado no *Sabermetrics*, Dean começou a realizar análises semelhantes durante a década de 1980. Buscando quantificar melhor a contribuição dos jogadores para a equipe, criou estatísticas para avaliar o desempenho do time em relação a quantos pontos marcava ou sofria a cada cem posses de bola. Os estudos deram origem ao *APBRmetrics - Association for Professional Basketball Research Metrics* [1].

Entretanto, apesar de pesquisas como as de Reep, James e Oliver, que analisavam o esporte com um certo rigor científico, a análise de dados estatísticos não parecia ter

grande relevância para as tomadas de decisão dos gestores, as quais eram guiadas, principalmente, pelo conhecimento empírico de especialistas do esporte. Foi apenas no início do século XXI, com a história de sucesso do *Oakland Athletics* [21] na MLB[2], que a análise de dados começou a causar uma verdadeira revolução e passou a receber a devida atenção de todos os envolvidos no esporte.

2.1.2. A Revolução dos Dados

Em 2002, o investimento financeiro cada vez maior em algumas franquias[3] estava tornando o beisebol relativamente previsível. As franquias de maior orçamento venciam os campeonatos, enquanto as franquias de baixo orçamento naturalmente ficavam nas últimas colocações e fora dos *play-offs*[4]. Tradicionalmente, as franquias de beisebol trabalhavam da mesma forma. Todas tinham uma equipe de olheiros que decidia quais atletas deveriam ser contratados ou recrutados para a temporada seguinte. Até que, Billy Beane, gerente geral e ex-jogador do *Oakland Athletics*, decidiu que, se ele não conseguia competir com as demais franquias no aspecto financeiro (sua franquia tinha o segundo menor orçamento da MLB), precisava encontrar uma outra vantagem competitiva. Dessa forma, auxiliado pelo economista Paul Depodesta, Beane foi buscar no *Sabermetrics* de James, a solução para nivelar a competição [43].

Beane e Depodeseta montaram o time para temporada 2002 fazendo uma análise baseada apenas nas estatísticas de desempenho do *Sabermetrics*, ignorando aspectos relevantes para os olheiros, tais como idade avançada, porte físico e até mesmo o modo desajeitado de jogar. O resultado foi um time com jogadores de baixo custo, cujas estatísticas denotavam pontos fortes a serem explorados.

A campanha foi surpreendente. O desacreditado time bateu o recorde de 20 vitórias consecutivas da conferência e se classificou para os *play-offs*. Não alcançou o título, mas, no ano seguinte, a inovação foi tema do *best-seller "Moneyball - A arte de ganhar um jogo injusto"* [43]. O sucesso do clube e do livro foram refletidos imediatamente na mudança de ideologia dos grandes clubes que viam seus jogadores com altíssimos salários renderem menos de que aqueles que ganhavam menos.

Desde então, o *Sabermetrics* passou a ser fundamental na seleção dos elencos de todos os clubes da MLB. O tradicional e rico, *Boston Red Sox*, foi além e contratou o próprio Bill James para trabalhar para o clube (onde permanece até hoje). O clube conseguiu em 2004 sagrar-se campeão após 86 anos de jejum, além de ter repetido o feito em 2007 e 2013.

Moneyball mudou não só o Beisebol, como também todos os esportes. Analistas e pesquisadores de outras modalidades passaram a adotar abordagens científicas similares para avaliar o conhecimento empírico relacionado aos esportes. No futebol, por exemplo, Chris Anderson e David Sally publicaram o livro *"Os números do jogo"* [30], compilando uma série de pesquisas que buscavam trazer evidências contrárias às crenças tradicionais do jogo. Dentre as discussões apresentadas nesse trabalho, podemos destacar: a ideia de que o escanteio não deveria ser cobrado para a grande área, a análise de que a demissão de

[2]MLB - Major League Baseball

[3]Nos Estados Unidos, as equipes pertencem a empresas habitualmente chamadas de franquias

[4]Sistema de disputa eliminatório que decide os finalistas de um campeonato

treinadores não melhora o desempenho dos times e a discussão sobre porque um jogador mais fraco (elo fraco) tem mais relevância (negativa) para o resultado do que os melhores jogadores do time (elo forte), ou seja, porque nenhum time é melhor que seu pior jogador.

Se no campo científico os cientistas seguem buscando encontrar padrões no esporte ou fazer predições, no campo esportivo os clubes estão cada vez mais interessados na descoberta de conhecimento através da análise de dados, visando obter vantagem competitiva. Entretanto, além da perspectiva esportiva, o interesse pela criação de modelos preditivos a partir de dados esportivos também tem sido impulsionado pelo crescimento dos mercados de apostas. A seguir, discutiremos a respeito desse domínio.

2.2. Apostas Esportivas: Motivação e Fundamentos

A democratização da Internet impulsionou mundialmente um segmento de mercado que está avaliado atualmente em mais de 3 trilhões de dólares [25]. Estamos falando do mercado de apostas esportivas. Este segmento já representa 37% do mercado de jogos de azar. Embora não exista regulamentação para essa prática em algumas partes do mundo (incluindo o Brasil), as casas de apostas online, hospedadas em lugares onde o jogo é regulamentado, permitem que pessoas de qualquer lugar do mundo possam realizar apostas pela Internet. Dessa forma, existe um grande número de informações disponíveis para análise na própria Web.

Nesta seção, apresentaremos uma fundamentação sobre apostas esportivas, visando dar ao pesquisador o entendimento sobre o significado dos dados desse domínio. Essa fundamentação é importante, pois como veremos na Seção 2.3.1.2, dados do mercado de apostas podem servir como dados de contexto em *Sports Analytics*, da mesma forma que dados esportivos de outras fontes podem servir para resolver desafios das pesquisas aplicadas em apostas esportivas (detalhados na Seção 2.5), também conhecida como *Sports Betting Analytics*.

2.2.1. Conceitos Fundamentais

O mercado de apostas possui dois ramos diferentes: as casas de apostas (*bookmakers market*) e as bolsas de apostas (em inglês, *betting exchange market*). As casas de apostas são as formas tradicionais de realizar apostas. Elas funcionam como reguladoras do mercado, oferecendo diversas oportunidades de apostas para o público. Nesse caso, podemos dizer que o apostador realiza a aposta "contra" uma casa de apostas. Já no outro ramo, o das bolsas de apostas, os apostadores apostam contra outros apostadores através de uma operadora que intermedia a negociação.

O conceito de aposta está estritamente ligado ao conceito de *odd*. De uma forma geral, as *odds* são usadas para definir quanto um apostador receberá se fizer uma aposta bem sucedida. Por essa razão, elas são frequentemente definidas como o "preço" pago por uma aposta premiada. As *odds* são representadas de diferentes formas pelas casas de apostas. Nesta seção, usaremos as *odds* decimais (*european odds*), o formato mais conhecido e utilizado. Outros formatos incluem as *odds* fracionárias (*fractional odds*) e as *odds* americanas (*american odds*).

Pare exemplificar, analisaremos as *odds* decimais para a luta de boxe entre Floyd

Mayweather e Conor McGregor (ver Figura 2.1), realizada em agosto de 2017. Cada resultado possível da luta possui uma *odd* associada e as mesmas nos permitem fazer as seguintes interpretações:

Figura 2.1. *Odds* oferecidas pelo *Sportingbet.com* para a luta entre Floyd Mayweather e Conor McGregor

- Se o apostador fizer uma aposta de $1 em Floyd e ele for o vencedor, o apostador receberá de volta $1.20;
- Se o apostador fizer uma aposta de $1 em Conor e ele for o vencedor, o apostador receberá de volta $4.33;
- Se o apostador fizer uma aposta de $1 no empate e a luta terminar empatada, o apostador receberá de volta $34.00;

Formalmente, considerando o valor apostado como *va* e a *odd* decimal oferecida pela casa de apostas para um determinado resultado como *odd*, podemos definir o potencial valor de retorno *vr* para uma aposta, como mostrado na Equação 1:

$$vr = va * odd \tag{1}$$

Nessa equação, podemos observar que o valor apostado está incluído no potencial valor de retorno, ou seja, para definir o lucro de uma aposta precisamos subtrair uma unidade do valor da *odd* apostada. Formalmente, podemos definir esse *lucro_potencial* como na Equação 2:

$$lucro\_potencial = va * (odd - 1) \tag{2}$$

Uma vez entendido que o valor da *odd* está associado ao retorno que o apostador poderá receber, podemos compreender como as casas de apostas definem o valor da *odd* a ser ofertada. Esse cálculo está diretamente ligado as chances de um determinado evento ocorrer, ou seja, as probabilidades.

2.2.2. *Odds* vs. Probabilidades

Em diversas áreas, é comum que a representação da probabilidade de um determinado evento ocorrer seja dada em termos percentuais. Por exemplo, se jogarmos uma moeda honesta para cima, ela tem 50% de chance de virar cara e 50% de virar coroa. Se jogarmos um dado de seis lados, cada lado tem aproximadamente 16,6% de chance de ocorrer. Conceitualmente, as somas das probabilidades de todos os eventos possíveis deve totalizar 100%.

No mercado de apostas, as *odds* são calculadas a partir dessas probabilidades. Entretanto, não podemos afirmar que essas *odds* refletem exatamente as chances de um determinado evento acontecer. Para perceber essa sutil diferença, precisamos entender o conceito de probabilidade implícita (*implied probability*).

Podemos definir que a probabilidade implícita é inversamente proporcional a *odd*, ou seja, dado um possível resultado x, podemos calcular sua probabilidade implícita *imp*, como na Equação 3:

$$imp_x = \frac{1}{odd_x} \tag{3}$$

Por exemplo, calculando as probabilidades implícitas para os resultados possíveis da luta entre Mayweather e McGregor (ver Figura 2.1), teremos:

- Vitória de Mayweather: $imp_{(Mayweather)} = \frac{1}{1{,}20} = 83,33\%$
- Vitória de McGregor: $imp_{(McGregor)} = \frac{1}{4{,}33} = 23,09\%$
- Empate: $imp_{(empate)} = \frac{1}{34{,}00} = 2,94\%$

Realizando a soma total dessas probabilidades, encontramos o valor 109,36%, que é diferente do valor esperado numa distribuição de probabilidade convencional, que seria 100%. Na prática, as casas de apostas nunca ofertarão *odds* nas quais as somas de suas probabilidade implícitas seja exatamente 100% (na Seção 2.2.3 explicaremos os motivos). Mesmo assim, essa probabilidade é fundamental para que o apostador avalie se o valor esperado (*expected value*) da aposta é positivo. Uma aposta tem "valor esperado positivo"(+*EV*) quando as probabilidades implícitas das *odds* oferecidas pelas casas de apostas são inferiores às probabilidades estimadas pelo apostador. Em outras palavras, se no exemplo da Figura 2.1, o apostador acreditar que McGregor possui 30% de chances de vencer, o apostador parece ter uma boa oportunidade para apostar, dado que a casa de apostas acredita que essa chance é de apenas 23,09%.

2.2.3. O Lucro das Casas de Apostas e o Balanceamento das *Odds*

Na Seção 2.2.2, entendemos que as somas das probabilidades implícitas das *odds* de um evento não totaliza 100% como esperado em um evento justo. Entretanto, é a partir dessa diferença, denominada *overround*, que vem o lucro das casas de apostas. O *overround* é um dos motivos que fazem as casas de apostas serem lucrativas em longo prazo. Elas têm, implicitamente, uma vantagem sobre os apostadores, que varia geralmente de 5% a 12% (no exemplo da Figura 2.1, o valor é de 9,36%). Para melhor entendimento dessa vantagem, vamos voltar ao exemplo da luta de boxe.

Se a casa de apostas tiver feito uma análise correta sobre as chances de cada lutador no evento, isso deverá se refletir no comportamento dos apostadores. Em outras palavras, a distribuição dos valores recebidos pela casa de apostas (em apostas) será semelhante à distribuição das probabilidades oferecidas por ela. Dessa forma, a cada \$109.36 recebidos pela casa de apostas, ela espera que \$83.33 estejam em apostas para Floyd,

$23.09 para Conor e $2.94 para o empate. Nesse cenário ideal, ao final do evento, a casa de apostas poderia ter os seguintes resultados:

- Em caso de vitória de Mayweather, a casa de apostas precisaria pagar $1.2 para cada $1 apostado. Sendo assim, como a casa recebeu $83.33 para esse resultado, pagaria $99.99 aos apostadores, obtendo um lucro de $9.37 sobre o total recebido ($109,36);
- Em caso de vitória de McGregor, a casa de apostas precisaria pagar $4.33 para cada $1 apostado. Nesse caso, como recebeu $23.09, pagaria $99.98 e lucraria $9.38;
- Em caso de empate, a casa precisaria pagar $34.00 para cada $1 recebido. Como recebeu apenas $2.94, pagaria $99.96 e lucraria $9.40.

Observando o cenário apresentado, podemos afirmar que a casa de apostas lucraria independente do resultado do evento. Mas, e se a distribuição dos valores apostados não estiver dentro do esperado? Isso significa que os apostadores, de certa forma, estão discordando das chances oferecidas pela casa. Esse é um dos motivos que pode fazer a casa de apostas recalcular suas probabilidades no decorrer do tempo, numa ação chamada de "balanceamento de *odds*".

Sabendo que as *odds* são publicadas muitos dias (ou meses) antes do início do evento, o balanceamento de *odds* pode ser realizado a qualquer momento, seja durante o evento ou até antes de ele iniciar. Apesar de ser menos comum, o balanceamento antes do evento é realizado para refletir uma nova informação (por exemplo, a contusão de um atleta importante antes de uma partida) ou para se ajustar à percepção do mercado. Entretanto, é durante o evento que existe uma grande variação, visto que as chances podem mudar a cada novo acontecimento (detalharemos isso na Seção 2.2.4). No exemplo da Figura 2.1, é natural que, se no início da luta de boxe, McGregor começar acertando bons golpes, as expectativas mudem e consequentemente as *odds* também.

Por fim, podemos concluir que as casas de apostas não estão necessariamente preocupadas em acertar o resultado final de um evento, mas, sim, em oferecer *odds* que reflitam a expectativa do mercado em relação às chances para cada resultado possível. Para isso, as casas vão estar sempre ajustando suas *odds* (conservando o *overround*) para manter o volume de apostas recebido balanceado e continuar garantindo lucro, independente do resultado do evento.

2.2.4. Casas de Apostas vs. Bolsas de Apostas

Uma vez entendidos os fundamentos que definem uma *odd*, podemos discutir como funcionam os dois ramos do mercado de apostas: casas de apostas e bolsas de apostas.

No ramo de casas de apostas, o apostador realiza a aposta em um determinado resultado, antes ou durante o evento, e aguarda o fim deste para saber se acertou. Em caso de acerto, a casa paga ao apostador de acordo com a *odd* negociada. Em caso de erro, o dinheiro do apostador é lucrado pela casa de apostas. Dentre as principais casas de apostas do mundo estão a Bet365 [13], William Hill [29], Bet-at-Home [12], Bwin [16], Rivalo [24] e Pinnacle [22].

Na bolsa de apostas, a dinâmica é um pouco diferente. Um apostador só consegue realizar uma aposta, caso exista outra pessoa que esteja disposta a apostar contra. Por exemplo, se na luta entre Mayweather e McGregor, um apostador decidir apostar a favor de McGregor, então precisa existir alguém que concorde em apostar contra McGregor (ou seja, em Mayweather ou empate). Existem diversas empresas que intermediam esse tipo de negociação, entre elas se destacam a Betfair.com [15] e a Betdaq.com [14]

A primeira diferença entre esses dois mercados está no valor das *odds* oferecidas. Como as apostas na bolsa são feitas entre pessoas, não existe a necessidade do *overround*, ou seja, as *odds* negociadas são geralmente maiores. Entretanto, as operadoras da bolsa cobram uma taxa de corretagem que varia de 2,5% a 7% sobre a aposta vencedora. Dessa forma, uma *odd* maior na bolsa (do que nas casas de apostas) não significa obrigatoriamente um lucro potencial maior. Para observarmos essas diferenças, vamos comparar as *odds* oferecidas na Betfair para a mesma luta do exemplo anterior, entre Mayweather e McGregor (ver Figura 2.2).

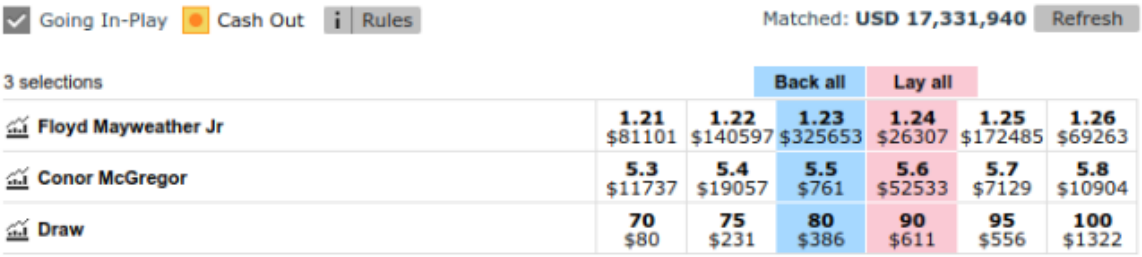

Figura 2.2. *Odds* oferecidas pela *Betfair.com* para a luta entre Floyd Mayweather e Conor McGregor

Ao calcular as probabilidades implícitas de cada *odd*, temos:

- Vitória de Mayweather: $imp_{(Mayweather)} = \frac{1}{1{,}23} = 81,30\%$
- Vitória de McGregor: $imp_{(McGregor)} = \frac{1}{5{,}50} = 18,18\%$
- Empate: $imp_{(empate)} = \frac{1}{80{,}00} = 1,25\%$

Somando as probabilidades da distribuição, encontramos um total de 100,73%, demonstrando um *overround* próximo a zero. Dessa forma, uma aposta de $1 em Mayweather que fosse vencedora, traria um retorno de $1.23. Todavia, se aplicada a taxa máxima de corretagem (7%), o lucro final seria próximo de $0.14 e inferior aos $0.20 que seriam lucrados, caso a aposta fosse feita na casa de apostas. Isso confirma que uma *odd* de 1.20 na casa de apostas pode ser mais lucrativa que uma *odd* de 1.23 encontrada na bolsa de apostas, caso a taxa de corretagem não seja baixa. No caso das *odds* oferecidas para McGregor e para o empate, podemos afirmar que, mesmo com uma taxa alta de corretagem (7%), elas oferecem lucro maior do que nas casas de apostas.

Uma outra diferença entre os ramos, ainda mais significativa, está na dinâmica para encerramento das apostas. Como mencionado anteriormente, é comum que, nas casas de apostas, as pessoas esperem até o final do evento para saber se a aposta teve sucesso,

ou não. Esse tipo de apostador é conhecido como *punter* e no Brasil acabou se tornando também adjetivo para uma forma de aposta. Dessa forma, uma "aposta punter" é aquela em que o apostador aguarda até o final do evento apostado. A bolsa de apostas, apesar de receber diversas apostas *punters*, permite um outro tipo de negociação, conhecida como *trading*. O *trading* esportivo é semelhante ao *trading* das bolsas de valores. O apostador trata a *odd* como o preço de um "ativo", o qual pode comprar (fazer a aposta) e vender segundos ou minutos depois de ter comprado (encerrar a aposta). Nesse caso, a variação positiva ou negativa da *odd* é quem define o lucro ou o prejuízo do apostador.

Assim como na bolsa de valores, existem diversas estratégias para fazer *trading*: *Scalping*, *Swing Trader*, *Dutching*, *Bookmaking*, outre outras. Nesta seção, apenas fizemos uma breve introdução sobre a natureza desse mercado. Em [18], [20], [19] é possível encontrar um extenso material sobre como operar nesse segmento e as possibilidades de apostas.

Do ponto de vista da pesquisa aplicada, alguns desafios serão apresentados na Seção 2.5. Entretanto, antes de discuti-los, precisamos entender as relações entre os dados coletados das casas de apostas e os dados do domínio esportivo (que serão detalhados na Seção a seguir).

2.3. Coleta de Dados

Nos últimos anos, com a crescente relevância de *Sports Analytics* para a indústria esportiva, as tecnologias para coleta automática de dados passaram a ter um grande potencial científico e comercial. Impulsionados pela evolução das tecnologias de armazenamento e processamento de grandes volumes de informações, os novos dispositivos de coleta estão trazendo um desafio ainda maior para os analistas de dados.

Na prática, as equipes estão conseguindo obter coleções heterogêneas de dados (em inglês, *datasets*) relacionadas ao desempenho dos atletas e adversários. Esses *datasets* estão cada vez maiores, com informações que antes eram relativamente impossíveis de serem catalogadas, como, por exemplo, o detalhamento da movimentação dos atletas em campo. A seguir, detalharemos os dados utilizados em *Sports Analytics*, destacando as principais formas de aquisição, de acordo com a classificação de Stein et al. [49]. Em seguida, descreveremos como um pesquisador pode ter acesso a esses dados para iniciar uma pesquisa.

2.3.1. A Origem dos Dados

A aquisição de dados para análise no esporte pode ser realizada de diversas formas. Nos esportes com interação entre os adversários, é comum que os dados sejam extraídos a partir de gravações de vídeos e sensores. Essas informações coletadas podem ser enriquecidas por dados relacionados ao "contexto", como a localização de uma disputa, informações meteorológicas, horário do evento, etc. Informações de contexto também podem ser obtidas a partir de publicações de redes sociais (textos, imagens, *feeds* de vídeo) ou casas de apostas, que refletem a sensação do público em relação aos acontecimentos do evento esportivo. Dessa forma, iremos dividir os tipos de dados sob dois aspectos: os dados extraídos a partir de vídeos e sensores e os dados de contexto.

2.3.1.1. Dados de Vídeos e Sensores

As gravações de vídeos são comuns nos esportes modernos, seja pela perspectiva da mídia (com diversas câmeras espalhadas por estádios e arenas), como também pela perspectiva das equipes que realizam as gravações por conta própria. De uma forma geral, as gravações podem ser consideradas uma das principais fontes de extração de dados, permitindo a indexação de informações referentes a movimentação dos atletas, eventos do jogo e dados descritivos (estatísticos). Empresas como a *STATS* [26] e *Opta Sports* [7] oferecem serviços para catalogar essas informações a partir de suas próprias gravações.

Outra possibilidade para extração desses dados são os sensores colocados diretamente nos jogadores (*wearables*) ou objetos de jogo (bolas, linhas, alvos, etc.). A maioria dos esportes já permitem que as equipes utilizem esses dispositivos, inclusive durante os eventos. Empresas como *Adidas* [10], *Catapult* [17], *Whoop* [28] e a brasileira *One Sports* [6], desenvolveram dispostivos que além de capturar a movimentação dos atletas, rastreiam suas características biométricas.

Dados de Movimentação

Os dados de movimentação, também conhecidos como dados espaço-temporais, descrevem onde um jogador ou objeto está localizado em um momento específico. Essas localizações são geralmente registradas por pares de coordenadas (x,y) e, às vezes, também pela coordenada z.

Os principais clubes do mundo já estão adotando em seus campos de treinamento e arenas sistemas como o *SportVU* da *Stats*. O *SportVU* é um sistema que integra múltiplas câmeras para rastrear as coordenadas de todos os jogadores e da bola (25 vezes por segundo), gerando mais de 3 milhões de dados por treino ou jogo.

Dados de Eventos

Podemos definir evento como um acontecimento relevante de uma disputa esportiva. Uma disputa pode ser descrita como uma sequência ordenada de eventos. A análise automática de vídeo já consegue realizar a detecção de diversos eventos relevantes. Entretanto, também é possível realizar essa catalogação através de notação manual. Na prática, a notação manual pode ter problema de precisão, enquanto que o reconhecimento automático pode ter problemas com a produção de resultados falso-positivos. De forma geral, podemos distinguir os eventos em duas categorias:

- *Eventos baseados em regras*: ocorrem de acordo com as regras de cada modalidade. No futebol, por exemplo, temos os laterais, os escanteios, os gols, etc.;
- *Eventos baseados na interação de atletas*: acontecem a partir da interação do atleta com a bola ou com o adversário. No futebol, por exemplo, temos os passes, os chutes a gol, os cruzamentos, etc.

Dados Descritivos

O desempenho de jogadores e times pode ser caracterizado a partir de dados descritivos. Os dados descritivos incluem tudo o que pode ser contado ou medido durante as

competições. Existe uma grande disponibilidade desse tipo de dado, que inclui resultados históricos, classificações de ligas, registros de carreiras, sumarização de dados de eventos, entre outros.

Historicamente, as estatísticas descritivas são as fontes mais utilizadas para fins analíticos, uma vez que os dados gerados automaticamente por gravações de vídeos são tecnologias relativamente novas.

2.3.1.2. Dados de Contexto

Dados relacionados ao contexto podem ser importantes para algumas análises. Tais dados podem incluir fatores ambientais, tais como a localização do jogo, temperatura ambiente e características do gramado. Fatores dessa natureza, podem, por exemplo, ser utilizados para estimar a qualidade do campo de jogo ou possíveis efeitos sobre o desempenho de um jogador [49].

Além disso, com a popularização das redes sociais, as percepções do público sobre os jogos também podem ser consideradas dados relevantes para as análises. Muitos desses serviços fornecem APIs para coleta seletiva dessas informações, que podem ser realizadas antes, durante e após as competições. Nesse mesmo sentido, esse sentimento coletivo também pode ser observado através do comportamento dos mercados de apostas esportivas (como vimos na Seção 2.2). As casas de apostas disponibilizam uma grande massa de dados que pode ser analisada.

Por fim, outros fatores externos, como notícias de jornais sobre as competições ou dados de documentos Web também podem ser consideradas importantes fontes de dados.

2.3.2. Acesso aos Dados

No tópico anterior, discutimos as características dos dados coletados para *Sports Analytics*. Uma vez compreendida a natureza desses dados, precisamos entender que, na prática, uma parte dessas coleções de dados são confidenciais e acessadas apenas pelos analistas das próprias empresas (que coletam) ou dos clubes (que contratam o serviço). Dessa forma, ainda não existem grandes *datasets* públicos com *dados de movimentação* para a comunidade científica em geral, pois as empresas só disponibilizam esses *datasets* de forma comercial e com um custo bastante elevado.

Por outro lado, existem *datasets* com *dados de eventos* e *dados descritivos* que podem ser acessados a partir de diversas fontes na Internet. Uma lista de bases públicas com dados dessa natureza pode ser encontrada em [42].

Os *dados de contexto*, dependendo da natureza, também podem ser acessados na Internet. Dados de redes sociais ou de casas de apostas, por exemplo, podem ser acessados através de APIs públicas. Uma lista de APIs para redes sociais pode ser encontrada em [9].

Em geral, os dados públicos estão disponíveis de forma estruturada em bases de dados (que podem ser acessadas através de APIs) ou planilhas (que podem ser baixadas por requisição HTTP). Entretanto, em alguns casos, os dados podem estar disponíveis apenas em formato semi-estruturado (por exemplo, embutidos em páginas HTML). Nesse

caso, é comum a realização de duas tarefas em sequência, visando a estruturação do *dataset*: rastreamento (Web crawling) e a raspagem (Web scraping).

O rastreamento é a tarefa de navegar pela Internet de forma automatizada para indexar páginas nas quais os dados desejados estão embutidos. Para realização desta tarefa são desenvolvidos ou utilizados algoritmos conhecidos como *web crawlers* ou *bots*. Existe uma grande número de bibliotecas para esse fim como, por exemplo, a *requests* [23], escrita em *Python*.

A raspagem é a tarefa de extrair informações específicas de documentos Web. Após o rastreamento do documento, analisa-se o mesmo visando compreender sua estrutura. Em seguida, escreve-se um algoritmo (*scraper*) para extrair apenas os dados desejados. Existem diversas bibliotecas para esse fim como, por exemplo, a *BeautifulSoup* [11], escrita em *Python*.

2.4. *Sports Analytics*: Pesquisa Aplicada e Desafios Emergentes

Na Seção 2.3, apresentamos as características dos dados utilizados em *Sports Analytics* e observamos que, com o contínuo crescimento dos dados disponíveis, o uso de métodos computacionais se tornou indispensável para alcançar novos e melhores resultados.

As pesquisas de análise de dados esportivos usando abordagens computacionais são comumente associadas a termos como Inteligência Artificial, Mineração de Dados e Aprendizagem de Máquina. Embora exista uma correlação entre todos esses termos (em alguns casos, podem ser considerados até sinônimos), cada um representa uma perspectiva diferente na modelagem dos dados.

Numa perspectiva geral, *Sports Analytics* é todo o processo de descoberta, interpretação e comunicação de novos conhecimentos obtidos a partir da modelagem de dados esportivos. Numa perspectiva tradicional, modelos podem ser obtidos diretamente a partir de análises estatísticas ou formulações matemáticas. Entretanto, na perspectiva da Ciência da Computação, essas abordagens estatísticas e matemáticas são aplicadas através de técnicas de mineração de dados. Além da estatística, as técnicas de mineração de dados podem utilizar métodos da inteligência artificial. Atualmente, a maioria dos métodos de sucesso de inteligência artificial utilizam aprendizagem de máquina (*machine learning*).

Nesta seção, detalharemos os passos para realização de uma pesquisa aplicada em *Sports Analytics*, destacando técnicas e exemplos que permitam ao pesquisador iniciante ter um ponto de partida para o desenvolvimento de pesquisas na área.

2.4.1. Processo de Descoberta de Conhecimento

Na Ciência da Computação, o processo que compreende todo o ciclo que o dado percorre até virar uma informação útil é conhecido como Processo de Descoberta de Conhecimento em Bases de Dados (*em inglês, Knowledge Discovery in Databases - KDD*). O KDD pode ser definido como um processo não trivial, de extração de informações, previamente desconhecidas e potencialmente úteis a partir de um conjunto de dados [37]. Dessa forma, qualquer pesquisa aplicada em *Sports Analytics* segue implicitamente esse processo.

O KDD contém uma série de passos (ver Figura 2.3) e, por ser um processo interativo, permite que o pesquisador possa intervir no fluxo das atividades, retornando a

passos anteriores quando necessário. A seguir, discutiremos como esse processo deve ser adequado à natureza do objetivo de pesquisa.

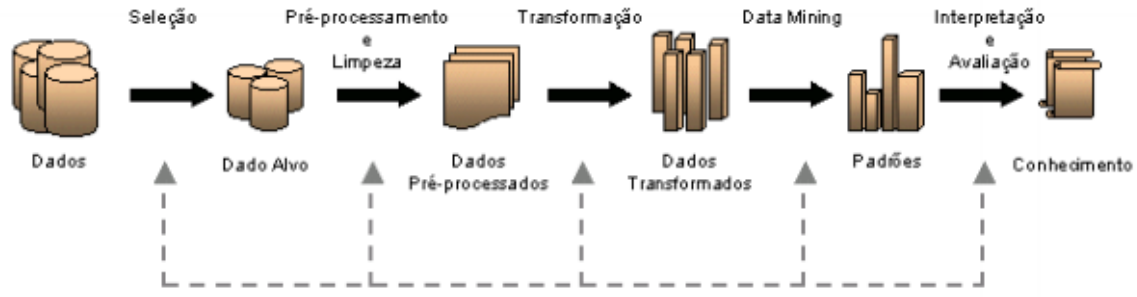

Figura 2.3. Processo de KDD (adaptado de [37])

2.4.1.1. Natureza do Objetivo de Pesquisa

De uma forma superficial, podemos dizer que existem dois objetivos gerais na pesquisa aplicada em *Sports Analytics*: detectar padrões ou realizar predições. Na prática, o processo acabará sendo o mesmo, pois para realizar predições é necessário o reconhecimento dos padrões.

Por exemplo, a detecção de padrões é um objetivo comum na análise de desempenho dos atletas. Os analistas tentam identificar os pontos fortes e fracos de suas equipes ou das equipes adversárias. Em outras palavras, eles buscam entender quais dados podem explicar a derrota ou vitória do seu time. Uma vez entendidos esses dados, os treinadores podem adaptar os treinos para tentar melhorar o desempenho dos atletas. Analogamente, podemos perceber que, se descobrirmos quais dados foram importantes no passado para determinar os resultados das disputas, podemos naturalmente fazer predições para os próximos eventos.

Entretanto, mesmo a modelagem podendo ser semelhante para ambos os objetivos, é preciso ter uma ideia clara sobre a finalidade da pesquisa, antes de iniciar a modelagem. Atualmente, os métodos de mineração de dados podem ser classificados em métodos de "caixa preta"ou métodos de "caixa branca". De uma forma geral, os métodos de caixa preta utilizam abordagens computacionais complexas e tendem a ser mais acurados. Por outro lado, esses métodos geralmente dão pouca informação sobre quais dados foram relevantes para se chegar a determinado resultado. Já os resultados dos métodos caixa-brancas são fáceis de serem interpretados, ao mesmo tempo que podem não ser tão acurados quanto os de caixa-preta.

Dessa forma, reanalisando nosso exemplo anterior, para um analista que deseja entender quais dados influenciaram um determinado resultado, a escolha de métodos caixa-preta pode não trazer informações tão significantes. Por outro lado, para um analista que deseja investir no mercado de apostas, talvez seja mais importante ter um modelo mais acurado, do que entender quais dados podem estar influenciando os resultados das disputas.

2.4.1.2. Seleção de Dados

Uma vez entendida a natureza do objetivo da pesquisa, precisamos decidir quais conjuntos de dados (*datasets*) estão disponíveis e podem ser relevantes para a modelagem do problema. Os *datasets* são compostos por variáveis e registros. As variáveis são conhecidas também como características (*features*), fatores ou atributos, enquanto os registros podem ser denotados como casos, objetos ou observações.

Por exemplo, se desejamos fazer um modelo para prever os resultados de um determinado campeonato, é natural precisarmos de dados históricos para que possamos encontrar variáveis que ajudem nosso modelo a prever corretamente. Considerando que temos disponíveis os resultados desse campeonato nos últimos cinco anos, então cada jogo será um *registro* e cada informação referente ao jogo será uma *variável* (nome do time, gols marcados, data, local, etc.).

Huang & Chang [41], por exemplo, utilizaram diversas variáveis (gols marcados, gols sofridos, chutes pra fora, chutes no alvo, pênaltis, faltas sofridas, faltas cometidas, cartões amarelos, cartões vermelhos, posse de bola, etc.) para fazer predições na Copa do Mundo de Futebol realizada na Alemanha, em 2006.

2.4.2. Pré-processamento e Limpeza de Dados

A seleção de um *dataset* é, geralmente, sucedida por uma análise exploratória, na qual podemos analisar os dados disponíveis e identificar alguns comportamentos iniciais. Para essa análise exploratória, podemos usar estatística descritiva como gráficos descritivos ou descrições tabulares e paramétricas. Este é um passo muito importante a ser realizado junto com um especialista no domínio do esporte em questão, pois através dessas análises podem surgir *insights* importantes para o pré-processamento dos dados.

Após essa análise exploratória, podemos ter mais segurança para realizar a limpeza e o pré-processamento dos dados. O passo de **limpeza de dados** é realizado através de uma série de tarefas, que incluem os tratamentos de valores ausentes, *outliers*, dados inconsistentes e dados duplicados. Já o passo de **pré-processamento** envolve tarefas que serão muito importantes para alcançar o objetivo de pesquisa, visto que, em pesquisas aplicadas, o sucesso dos métodos de mineração de dados dependem principalmente do trabalho de engenharia dos dados (*feature engineering*). Este passo envolve uma série de tarefas, a citar: agregação, amostragem, redução de dimensionalidade, seleção de subconjunto de características, criação de novos recursos, discretização e binarização de variáveis (detalharemos cada uma destas tarefas nos exemplos a seguir).

Vamos supor que desejamos prever os resultados do Campeonato Brasileiro de Futebol de 2017. Para isso, selecionamos um *dataset* que contém todos os resultados da história do Campeonato Brasileiro (1971-2016), pois acreditamos que essa seja uma boa fonte para tentar criar um modelo preditivo. As variáveis disponíveis no *dataset* estão listadas na Tabela 2.1.

Seguindo o processo do KDD, o primeiro passo é avaliar a qualidade dos nossos dados para fazer a limpeza adequada. Uma lista de atividades possíveis seria:

1. *Dados ausentes*: verificar se todos os placares foram devidamente registrados, ou

Tabela 2.1. Variáveis do *dataset* de resultados do Campeonato Brasileiro

CAMP - Ano do campeonato
NROD - Número da rodada
TIMC - Time da casa
TIMV - Time visitante
GOLC - Gols marcados pelo time da casa
GOLV - Gols marcados pelo time visitante
DATA - Data do jogo
LOCJ - Local do jogo (cidade)

seja, se todos os registros tem valores para os atributos *GOLC* e *GOLV*;

2. *Dados inconsistentes*: verificar se todos os valores para *GOLC* e *GOLV* são números inteiros não-negativos;
3. *Dados duplicados*: verificar se não existem dois registros para o mesmo jogo.

Para cada uma dessas atividades, precisamos tomar uma decisão que poderia resultar na eliminação de alguns registros ou na imputação de valores adequados baseados em algum critério.

2.4.2.1. Amostragem

Após a limpeza dos dados, podemos partir para o pré-processamento dos dados. Começaremos fazendo uma *amostragem*. A amostragem é uma abordagem comumente usada para selecionar um subconjunto de registros a serem analisados. No nosso exemplo, podemos selecionar apenas os registos dos campeonatos a partir de 2006, baseado na informação de que, a partir desse ano, o regulamento do campeonato foi alterado. Essa seria uma amostragem simples. Entretanto, para diferentes propósitos, outras formas de amostragem poderiam ser analisadas. Aoki, Assunção & Melo [31], por exemplo, buscaram medir a influência da sorte em alguns esportes e, para isso, selecionaram uma amostra com dados de 198 campeonatos, incluindo 1.503 temporadas de 84 países diferentes para 4 esportes. É natural que para pesquisas dessa magnitude fosse impossível usar todos os dados da população. Entretanto, foi realizada uma boa amostragem estratificada para avaliação.

2.4.2.2. Criação de Novos Atributos

Em seguida, podemos realizar a *criação de novos atributos*. Esta tarefa depende bastante dos *insights* do pesquisador para resolver o problema proposto. No nosso exemplo, podemos observar que não há um atributo único para revelar qual foi o resultado da partida (vitória do time da casa, empate ou vitória do visitante). Para fazer isso, precisaríamos comparar os gols marcados por cada uma das equipes. Dessa forma, poderíamos começar criando esse atributo que sumariza o resultado (RES). Formalmente, dado um jogo x, podemos detonar RES pela Equação 4.

$$RES_x = \begin{cases} C, & \text{se GOLC}_x > \text{ GOLV}_x \\ V, & \text{se GOLC}_x < \text{ GOLV}_x \\ E, & \text{caso contrário} \end{cases} \quad (4)$$

Agora que representamos a variável do resultado (*RES*), precisamos analisar quais outras variáveis podem ser significativas para fazermos predições para *RES*. No momento, temos apenas as identificações dos times, a data do jogo, o local do jogo e o número da rodada. Esses dados trazem pouca ou nenhuma informação a respeito do desempenho prévio dos times. Dessa forma, a partir dos resultados coletados, podemos criar uma diversidade de variáveis relacionadas ao desempenho dos clubes que podem ser mais significativas para um modelo de predição. Algumas das variáveis que podemos derivar dos registros estão listadas na Tabela 2.2.

Tabela 2.2. Variáveis derivadas do *dataset* de resultados do Campeonato Brasileiro

RES - Resultado do Jogo
GOLSP - Gols Marcados no Campeonato (Gols pró)
GOLSC - Gols Sofridos no Campeonato (Gols contra)
NVIT - Números de Vitórias
NEMP - Número de Empates
NDER - Número de Derrotas
NJOG - Número de Jogos Disputados

2.4.2.3. Redução de Dimensionalidade

Uma outra tarefa importante para a modelagem (seja por questão de tempo de processamento, memória utilizada ou até mesmo eficácia) é a *redução de dimensionalidade* do *dataset*. No nosso exemplo, poderíamos fazer reduções simples como transformar as variáveis GOLSP e GOLSC em uma única variável que represente o saldo de gols, ou ainda transformar as variáveis NVIT, NEMP e NDER em uma única variável que represente o número de pontos marcados. Em cenários mais complexos, algumas técnicas robustas podem ser aplicadas para essa redução. Zhao and Cen [52], por exemplo, demonstraram o uso da análise de componentes principais ou PCA (do inglês, *Principal Component Analysis*) para unir 13 variáveis que influenciam o resultado de uma partida de futebol, em uma única variável. PCA é técnica de álgebra linear para atributos contínuos que deriva novos atributos (componentes principais) que sejam combinações lineares dos atributos principais, sejam ortogonais (perpendiculares entre si) e capturem a quantidade máxima de variações nos dados. Uma outra técnica da álgebra linear também utilizada para redução de dimensionalidade é a decomposição de valor único ou SVD (do inglês, *Singular Value Decomposition*). Detalhes comparativos dessas técnicas podem ser encontrados em [38].

2.4.2.4. Seleção de Variáveis

Uma outra forma de reduzir dimensionalidade é através da *seleção de um subconjunto de variáveis*, também conhecida como *feature selection*. No nosso exemplo, intuitivamente, podemos perceber que o número da rodada ou a data do jogo provavelmente não tem qualquer relevância para as predições dos resultados e, portanto, podem ser retiradas da modelagem. Utilizar *features* que não tem relevância para predição, além de aumentar o custo computacional, pode fazer com que a eficácia dos métodos seja comprometida. Existem três formas de realizar essa seleção. A primeira, é ignorar essa tarefa nesse momento, e deixar que essa seleção ocorra naturalmente dentro do algoritmo de mineração de dados (abordagens internas). A outra é usando alguma abordagem independente nesse momento e apenas as variáveis selecionadas para o algoritmo (abordagem de filtro). Por fim, podemos usar alguns métodos de mineração de dados como uma caixa-preta para que esses métodos indiquem qual o melhor subconjunto de atributos (abordagem de envoltório). Tüfekci [51], por exemplo, comparou técnicas de abordagens de filtro e técnicas de envoltório para reduzir um *dataset* com 70 variáveis, visando realizar predições de resultado para o Campeonato Turco de Futebol. Mais detalhes sobre essas técnicas podem ser vistos em [32].

2.4.2.5. Discretização e Binarização

Alguns algoritmos de mineração de dados requerem que os dados das variáveis estejam em formato categórico. Assim, muitas vezes é necessário transformar uma variável contínua em uma variável discreta (discretização) e tanto as variáveis discretas quanto as contínuas podem ser transformadas em atributos binários (binarização). Em resumo, a binarização pode ser considerada a discretização para duas categorias. Existem diversos métodos para realizar essas tarefas. Constatinuou [36], por exemplo, usou um sistema dinâmico [45] para discretizar variáveis como "força do time", "cansaço"e "motivação". A variável "cansaço", por exemplo, era determinada através da quantidade de dias entre um jogo e outro e, depois de aplicada a discretização, denotava valores como "baixa", "média"e "alta". As técnicas mais utilizadas são discutidas em [40].

2.4.2.6. Agregação

Agregação é a combinação de vários registros em um único. Dessa forma a agregação é importante para mudar a escala (de vários registros para um único) e o escopo (o novo registro representa uma nova informação). Do ponto de vista da técnica de mineração, menos dados significa menos tempo de processamento. No nosso exemplo, se o nosso objetivo de pesquisa fosse prever a quantidade de gols do Campeonato Brasileiro de 2017, poderíamos agregar os registros de todos jogos de um ano de campeonato em um único registro com a soma da quantidade de gols de todos os jogos daquele ano. Dessa forma, limitaríamos a quantidade de registros ao número de temporadas do campeonato. Uma desvantagem da agregação é que podemos perder algum detalhe importante dos dados. No nosso exemplo, perderíamos a informação de qual rodada do campeonato possui a maior quantidade de gols.

2.4.3. Transformação de Dados

O processo de transformação é o último antes da mineração de dados e normalmente é realizado de acordo com a técnica que será utilizada. A transformação é aplicada nos valores de uma determinado variável para que eles sejam reduzidos a uma faixa menor de valores. No nosso problema das predições do Campeonato Brasileiro, poderíamos, por exemplo, transformar a quantidade de gols marcados (GOLP) em média de gols marcados, aplicando uma função simples f (Equação 5) que divide o número de gols feitos pelo número de jogos disputados, para todo registro x.

$$f(x) = \frac{GOLP_x}{NJOG_x} \tag{5}$$

Apesar de serem uma representação da mesma característica, esse tipo de transformação deve ser feita com cuidado, pois altera a natureza dos dados. Entretanto, a redução da variação entre eles torna os dados mais apropriados para algumas técnicas de mineração. É comum a normalização por *min-max*, *z-score* ou escala decimal.

2.4.4. Mineração de Dados

Uma vez obtidos os dados transformados, podemos escolher o método de mineração de dados adequado para atingir o objetivo de pesquisa. Na abordagem computacional os principais métodos de mineração utilizam aprendizagem de máquina. Como detalhamos anteriormente, em *Sports Analytics*, dois objetivos gerais são os mais comuns: a realização de predições e o reconhecimento de padrões e

Para modelagem preditiva destacaremos as técnicas de Regressão e de Classificação, enquanto que para reconhecimento de padrões, destacaremos as técnicas de Agrupamento (*clustering*).

2.4.4.1. Regressão

A regressão é uma técnica de modelagem preditiva na qual a variável dependente (variável alvo) é contínua. Em *Sports Analytics*, podemos usar regressão para problemas como: Quantos pontos um time marcará num jogo de basquete? Quantas horas um tenista vai jogar para vencer um campeonato? Por qual preço Cristiano Ronaldo deve ser negociado?

Formalmente, suponha que X representa um conjunto de dados que contem N observações:

$$X = \{(x_i, y_i) | i = 1, 2, 3, ..., N\} \tag{6}$$

Cada x_i corresponde ao conjunto de atributos da observação de índice i (variáveis explicativas) e y_i corresponde à variável alvo (ou resposta). Em outras palavras, y é o que desejamos prever e x é o conjunto de atributos que pode nos ajudar nessa tarefa.

Sendo assim, podemos dizer que regressão é a tarefa de aprender uma função alvo $\boldsymbol{f}$ que mapeie cada conjunto de atributos x em um valor contínuo y. Se a relação entre as variáveis puder ser descrita por uma função linear, podemos dizer que estamos tratando de uma *regressão linear*; caso não seja descrita por uma função linear (uma

função exponencial ou logarítmica, por exemplo) estaremos tratando de uma *regressão não-linear*.

Quando o conjunto de atributos contém apenas um atributo, estamos tratando de uma *regressão simples*. Um regressão linear simples pode ser observada na Figura 2.4 que trata a previsão da quantidade de finalizações de uma partida a partir da quantidade de escanteios [30].

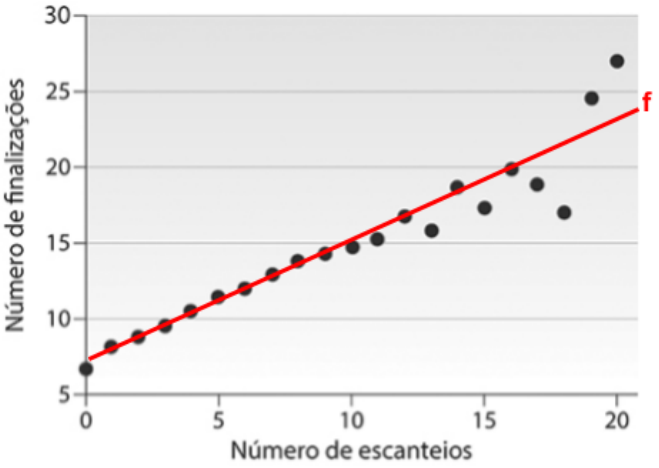

Figura 2.4. Relação entre escanteios e finalizações na *Premier League*, entre 2001-2011 (adaptado de Anderson & Sally [30])

Entretanto, em *Sports Analytics*, é comum que diversos atributos influenciem no valor da variável alvo. No exemplo anterior, poderíamos acrescentar outros atributos (como tempo de posse de bola, número de faltas, etc.) para tentar melhorar a predição da quantidade de finalizações. Nesse caso, quando o conjunto de atributos tem tamanho maior que um, estamos tratando de uma *regressão múltipla*.

Formalmente, a função alvo $\boldsymbol{f}$ de uma regressão linear múltipla pode ser dada pela Equação 7, na qual os valores de $\alpha_1, \alpha_2, ..., \alpha_n$ são os parâmetros a serem estimados na tarefa de regressão e ε é uma variável aleatória desconhecida (erro amostral).

$$f(X) = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + ... + \alpha_n x_n + \varepsilon \tag{7}$$

Em resumo, para se obter a equação estimada, podemos utilizar o método dos mínimos quadrados (MMQ) para obter o menor erro possível. O erro é a representação da soma das diferenças entre o valor real observado e o valor estimado pela regressão. Sendo assim, essa função de erro pode ser denotada, por exemplo, pela soma dos erros quadrados (Equação 9) ou absolutos (Equação 8).

$$Erro\ Absoluto = \sum |y_i - f(x_i)| \tag{8}$$

$$Erro\ Quadrado = \sum (y_i - f(x_i))^2 \tag{9}$$

Em mineração de dados, existem diversas abordagens para estimar os parâmetros da regressão, como Regressão Linear Bayesiana, Regressão por Redes Neurais, Regressão por Árvore de decisão, etc.

2.4.4.2. Agrupamento (*Clustering*)

É uma técnica para agrupar observações segundo algum grau de semelhança, de forma automática. O critério de semelhança é dado por alguma função estatística. O agrupamento é realizado sem intervenção do usuário, sem considerar previamente as características dos grupos e sem o uso de grupos de teste previamente conhecidos para direcionar a classificação.

Os problemas de agrupamento apresentam uma complexidade de ordem exponencial, ou seja, métodos de força bruta, como criar todos os possíveis grupos e escolher a melhor configuração, não são viáveis. Por exemplo, se quisermos agrupar os 100 melhores jogadores de futebol do mundo em 5 grupos, vão existir $5^{100} \approx 10^{70}$ possíveis agrupamentos. Dessa forma, mesmo um computador capaz de testar 109 configurações diferentes por segundo, precisaria de 1.053 anos para terminar a tarefa. Logo, é necessário encontrar uma heurística eficiente que permita resolver o problema [44].

Dois tipos de algoritmos são amplamente utilizados em *Sports Analytics*: os baseados em partição e os baseados em hierarquia. Os primeiros são métodos que buscam agrupar as observações em um número *k* de grupos previamente escolhido, minimizando uma função de custo. Em outras palavras, cada observação será agrupada na classe em que a função de custo é minimizada (ver Figura 2.5). Já os métodos hierárquicos não necessitam que seja definido um número de grupos previamente. Os dados são particionados sucessivamente, produzindo uma representação hierárquica dos agrupamentos (ver Figura 2.6).

Um dos algoritmos de particionamento mais utilizados é o *K-means*, que tem baixa complexidade e grande eficiência computacional em geral. O *K-means* pode ser abstraído em quatro passos:

1. Distribua aleatoriamente *k* pontos (número de clusters) como centros de cluster;
2. Atribua cada observação a um cluster, de forma que a distância da observação ao centro do cluster atribuído seja a menor;
3. Recalcula o centro do cluster usando as observações atribuídas a cada cluster;
4. Repita as etapas 2 e 3 até que as atribuições do cluster não mudem.

Cheng em [27], por exemplo, utilizou *K-means* para classificar os jogadores da NBA em 8 grupos. A rotulagem de grupos foi feita após a tarefa de agrupamento, de acordo com as semelhanças dos jogadores agrupados (ver Figura 2.5). Já Lesmeister [2], utilizou um agrupamento hierárquico, para agrupar 40 *wide receivers*[5] da NFL (ver Figura 2.6).

2.4.4.3. Classificação

É uma técnica de modelagem que mapeia objetos em uma das várias categorias pré-definidas. Na prática, em *Sports Analytics*, podemos usar classificação para fazer pre-

[5]Jogador de posição ofensiva no futebol americano

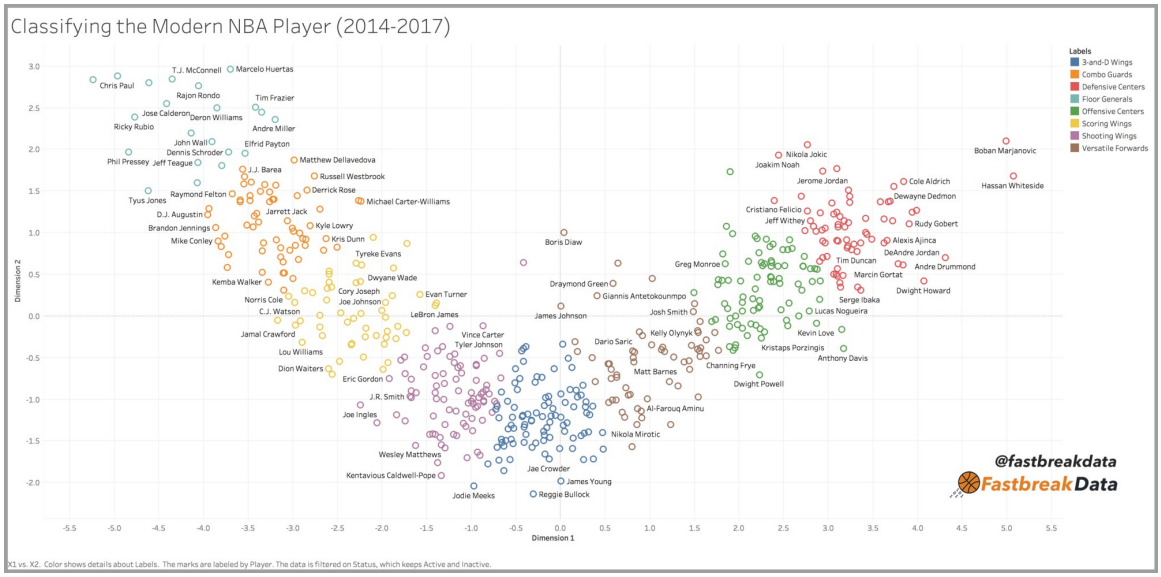

Figura 2.5. Agrupamento particional de jogadores da NBA, entre 2014-2017 (adaptado de Cheng [27])

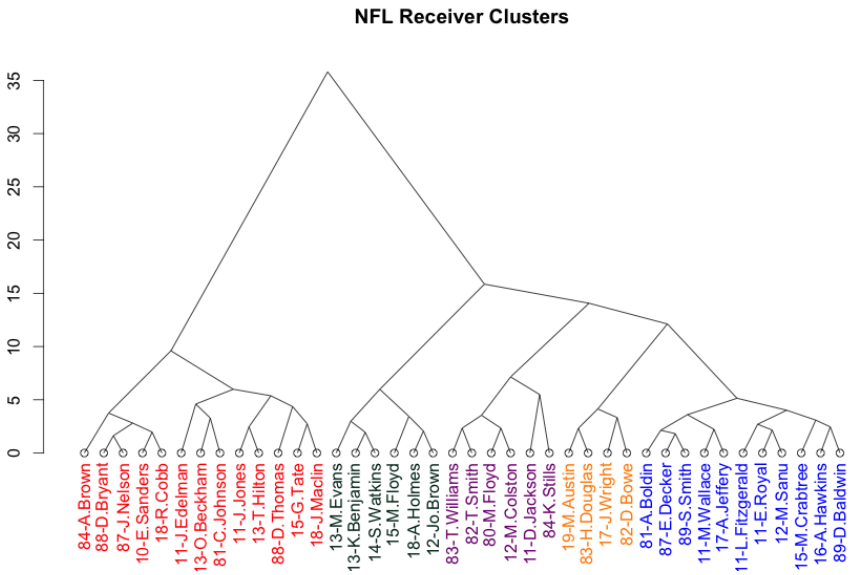

Figura 2.6. Agrupamento hierárquico de 40 *wide receivers* da NFL (adaptado de Lesmeister [27])

dições como, por exemplo: *Quem vencerá um jogo de volêi*? *Em uma partida de futebol, ambos os times marcarão gols*? *Como terminará uma luta de Judô (por pontos*?

A seguir, destacaremos algumas abordagens utilizadas por métodos de classificação em mineração de dados, no escopo de *Sports Analytics*.

Árvores de Decisão

É uma estrutura semelhante a um fluxograma em que cada nó interno representa a avaliação de uma variável, cada ramo representa o resultado do teste e cada nó de folha

representa um rótulo de classe (decisão tomada após o cálculo de todos os atributos). Os caminhos da raiz para a folha representam regras de classificação.

Em [8], por exemplo, o autor utilizou árvores de decisão para prever a posição de um jogador da NBA, a partir de um conjunto de 15 características. A árvore criada para essa classificação pode ser visualizada na Figura 2.7.

Em aprendizagem de máquina, algumas abordagens do estado da arte utilizam múltiplas árvores de decisão. Nesse caso, dois algoritmos bastante utilizados são: Floresta Randômica (*Random Forest*) e *AdaBoost*.

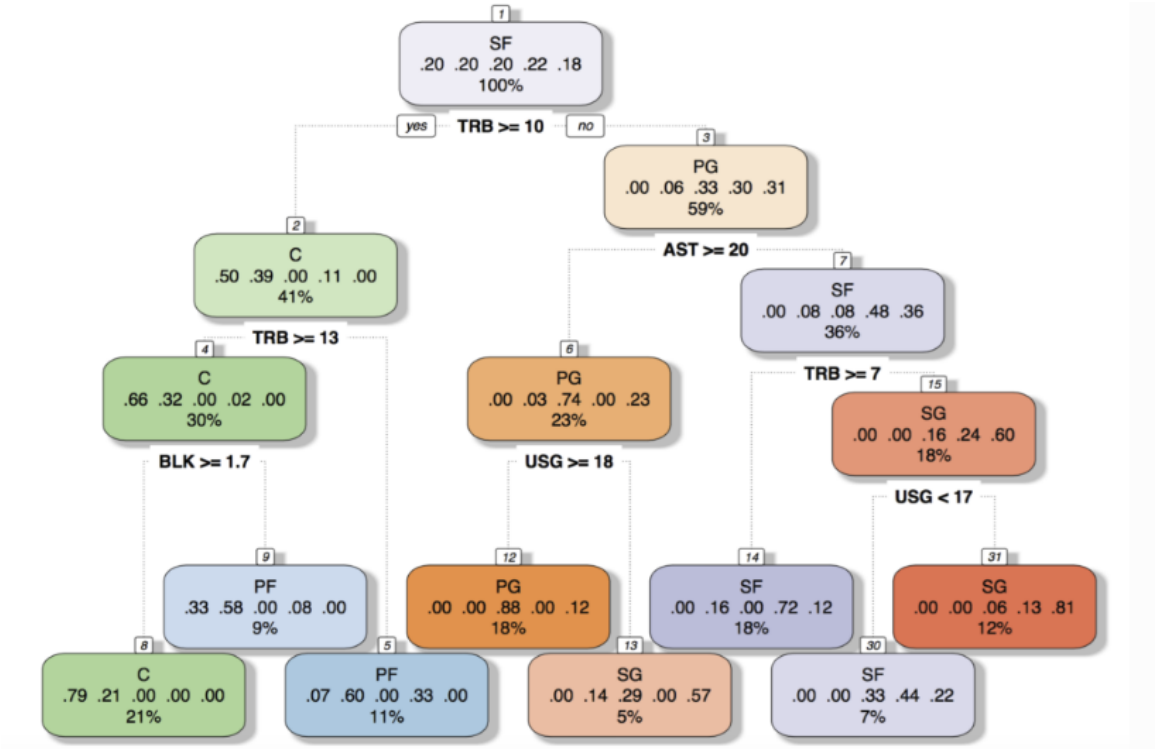

Figura 2.7. Árvore de decisão para classificar jogadores de basquete da NBA por posição (adaptado de Jager [35])

Redes Bayesianas

É um modelo acíclico e probabilístico que representa as variáveis e suas dependências através de um gráfico. Os nós representam as variáveis de um domínio, enquanto os arcos representam as dependências condicionais entre as variáveis. As informações sobre cada nó são dadas através da função de probabilidade que requer um determinado conjunto de valores como entrada e fornece uma distribuição de probabilidade de variáveis como saída.

Constantinou [35] usou Redes Bayesianas para fazer predições de resultados no Campeonato Inglês de Futebol (*Premier League*), através de variáveis objetivas e subjetivas. Uma parte da rede proposta pode ser vista na Figura 2.8.

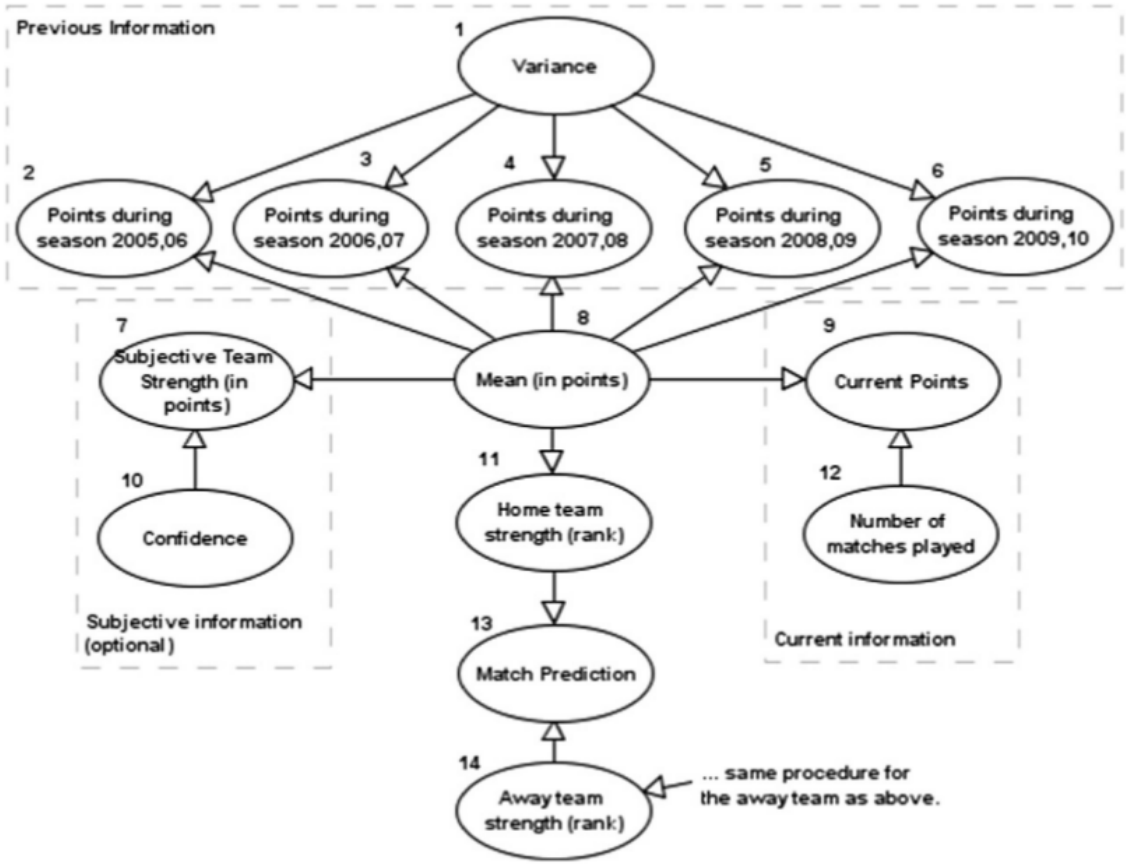

Figura 2.8. Rede Bayesiana para predição de resultados de futebol (adaptado de Constantinou [35])

Redes Neurais

Uma rede neural artificial (ANN) é um modelo computacional baseado em redes neurais biológicas. É uma abordagem de "caixa preta"e, apesar de ser acurada em muitos cenários, não fornece qualquer informação sobre a significância das variáveis independentes.

A técnica consiste em um grupo interconectado de neurônios artificiais (análogos aos do cérebro humano) e processos de informação usando uma abordagem conexionista para computação. Os neurônios dentro da rede trabalham em conjunto (e em paralelo) para produzir uma função de saída. Isso distingue as redes neurais dos programas de computação tradicionais que simplesmente seguem instruções em ordem sequencial. Na maioria dos casos, a ANN é um sistema adaptativo que altera sua estrutura com base em informações que fluem através da rede durante a fase de aprendizagem.

Huang [41], por exemplo, utilizou *multi-layer perceptron* (MLP), um tipo de rede neural, para fazer predições de resultados na Copa do Mundo de Futebol de 2006. A arquitetura utilizada pode ser observada na Figura 2.9.

Máquinas de Vetor de Suporte

A Máquina de Vetor de Suporte (MVS) analisa, para cada observação de conjunto de dados, qual de duas possíveis classes a observação faz parte. Isso faz da MVS um classi-

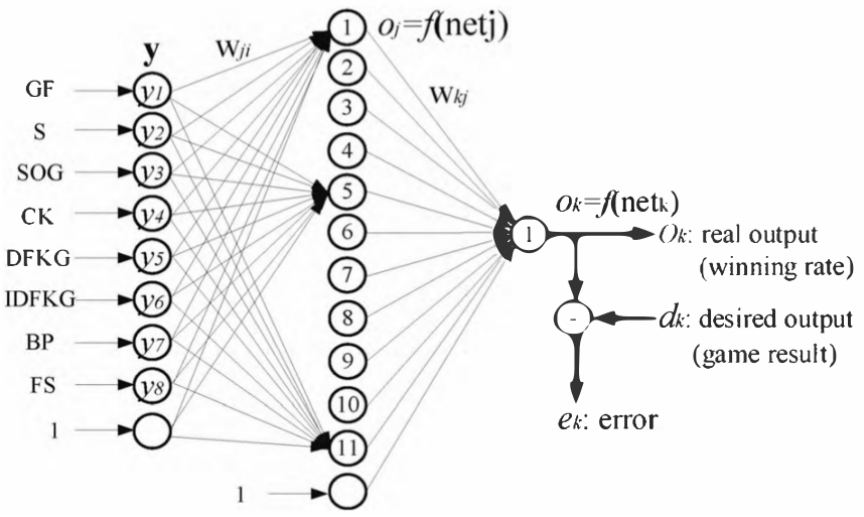

Figura 2.9. Rede Neural para predição de resultados de futebol (adaptado de Huang [35])

ficador linear binário não probabilístico. Em outras palavras, a tarefa da MVS é encontrar uma linha de separação, comumente chamada de hiperplano, entre dados de duas classes. Essa linha busca maximizar a distância entre os pontos mais próximos em relação a cada uma das classes.

Tolbert & Trafalis [50], por exemplo, utilizaram MVS para prever vencedores na MLB. O classificador estimado para esse objetivo pode ser visto na Figura 2.10

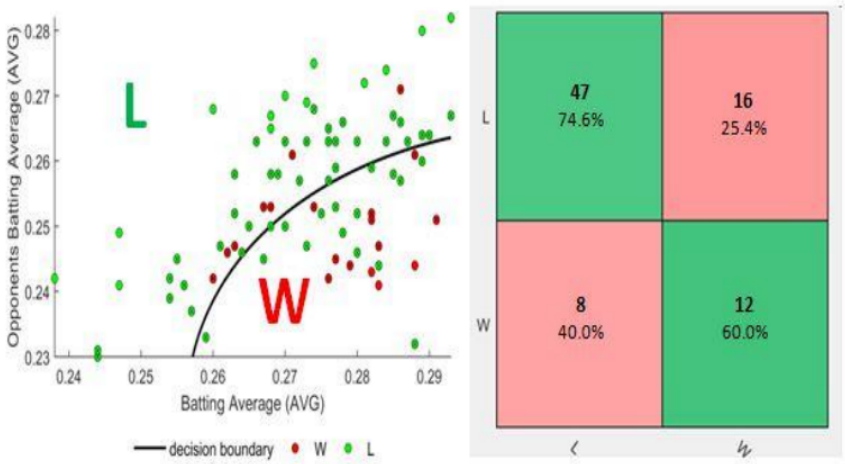

Figura 2.10. Classificador MVS para predição de vencedores no Beisebol (adaptado de Tolbert & Trafalis [50])

2.4.5. Avaliação e Interpretação dos Modelos

A última etapa do KDD é avaliar se o modelo está adequado para o objetivo proposto. Geralmente, a avaliação é parte integrante do processo de desenvolvimento do modelo. Entretanto, avaliar o desempenho do modelo com os dados utilizados para o treinamento não é aceitável na mineração de dados. Dessa forma, existem dois métodos comuns de

avaliação de modelos na mineração de dados: *Hold-Out* e Validação Cruzada (*Cross-Validation*). Para evitar a sobreposição, ambos os métodos utilizam um conjunto de testes (não visto pelo modelo) para avaliar o desempenho do modelo.

No *Hold-Out*, o conjunto de dados é dividido em três subconjuntos:

- **Treinamento**: subconjunto de dados usado para construir modelos preditivos;
- **Validação**: subconjunto de dados utilizado para avaliar o desempenho do modelo construído na fase de treinamento;
- **Teste**: subconjunto (não usado no treinamento) para avaliar o desempenho futuro do modelo. Se um modelo se encaixa no conjunto de treinamento muito melhor do que o conjunto de testes, é provável que este esteja com sobreajuste (*overfitting*).

O conjunto de dados pode ser dividido em quantidades iguais ou não. Geralmente, são selecionados 2/3 dos dados para treinamento e 1/3 para testes.

Quando apenas uma quantidade limitada de dados está disponível para obter uma estimativa do desempenho do modelo, é comum usamos a Validação Cruzada k-fold. Neste tipo de validação, dividimos os dados em k subconjuntos de igual tamanho. Em seguida, um subconjunto é utilizado para testes e os demais são utilizados para estimativa dos parâmetros. Este processo é realizado k vezes, alternando de forma circular o subconjunto de testes. A cada ciclo, calcula-se a acurácia do modelo. Ao término das k interações, calcula-se a acurácia final do modelo sobre os erros encontrados.

A comparação entre modelos pode ser realizada por diversas métricas. Em modelos de regressão, as métricas mais comuns são: Coeficientes de Determinação, Medidas Estatísticas Padrão (Erro Médio, Erro Absoluto Médio e Erro Quadrático Médio) e Medidas Relativas (Erro Percentual, Erro Percentual Médio e Erro Percentual Absoluto Médio).

A comparação entre modelos de classificação também pode ser feita por diferentes métricas ou abordagens como: Taxa de Classificação Incorreta (*misclassification rate*), Matriz de Confusão, F-Measure, Gráficos ROC e Área sob a Curva ROC.

2.5. Desafios Emergentes

Uma vez entendido o detalhamento das pesquisas aplicadas em *Sports Analytics*, podemos listar alguns dos desafios emergentes nesse campo de pesquisa. Do ponto de vista prático, os analistas de dados esportivos têm uma diversidade de desafios para enfrentar, seja no âmbito corporativo (nos clubes ou empresas) ou no âmbito científico.

O grande desafio da pesquisa aplicada está na arte de manipular as variáveis disponíveis para que as técnicas de mineração consigam obter sucesso. Como vimos no processo de KDD, os passos de Pré-Processamento e Transformação são decisivos para que o objetivo seja alcançado. Dessa forma, os desafios emergentes estão relacionados com a natureza dos dados disponíveis. Vejamos:

- **Transformar dados de movimentos em dados de eventos**: fazer essa rotulagem automaticamente tem um alto grau de dificuldade, pois deve ser feita de acordo com

as regras e características de cada esporte. No futebol, por exemplo, a rotulagem de chutes a gol pode ser relativamente trivial, enquanto a rotulagem de uma falta é uma tarefa de alta complexidade. Essas rotulagens podem ser importantes, por exemplo, para gerar os melhores momentos de um jogo, automaticamente.

- **Prever onde os eventos vão acontecer**: os dados também podem ser observados numa perspectiva espacial. Nesse sentido, podemos abstrair quando as coisas aconteceram e nos preocupar apenas com onde aconteceram. Para esse tipo de abordagem, a geometria computacional pode ser o ponto de partida para analisar trajetórias em dados de esportes de interação (ver Figura 2.11). Encontrar padrões nessa sequência também é um tarefa desafiadora. Uma introdução a mineração em dados espaciais pode ser encontrada em [48].

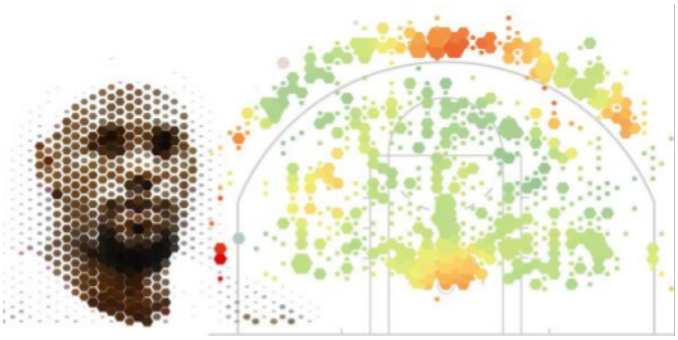

Figura 2.11. Mapa de calor de um jogador de basquete (adaptado de Goldsberry [3])

- **Prever eventos futuros**: os dados podem ser tratados como observações ao longo do tempo, ou seja, como uma série temporal (ver Figura 2.12). Do ponto de vista temporal, podemos abstrair onde as coisas aconteceram e nos preocupar apenas com a sequência de eventos. Dessa forma, encontrar padrões nessas sequências é um tarefa desafiadora. A análise de séries temporais é comum em diversos domínios. Uma lista de técnicas utilizadas pode ser vista em [39].

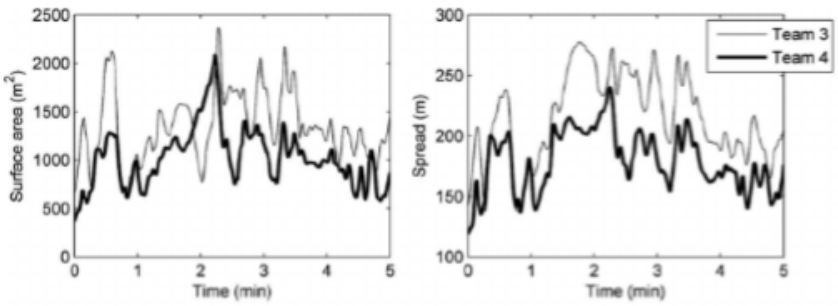

Figura 2.12. Séries temporais do espalhamento dos jogadores e áreas de superfícies cobertas por dois times de futebol em uma disputa (adaptado de Moura et al. [46])

- **Analisar desempenho dos atletas**: a análise de comportamento dos atletas pode ser vista sob a perspectiva individual ou coletiva. Em ambos os casos, a análise dos

dados espaço-temporais (como dados de movimentação) é um dos grandes desafios emergentes na área. Uma introdução sobre o assunto pode ser vista em [33]. Na perspectiva do esporte coletivo, uma abstração para os dados espaço-temporais é interpretá-los como dados de "movimento de grupo". O movimento em grupo também é estudado na biologia e na ciência comportamental. Esses ramos de pesquisa também usam dados de movimentação de pessoas e animais. Atualmente, existe uma diversidade de *datasets* disponíveis para análise [5] com informações dessa natureza. Os pesquisadores acreditam que um dos grandes desafios atuais é encontrar semelhanças e diferenças entre as estratégias das equipes esportivas e o comportamento dos animais em grupo, acreditando que as técnicas de análise desenvolvidas para um domínio possam ser utilizadas no outro [49]. Na perspectiva individual, quantificar o custo-benefício de cada atleta ou encontrar atletas com características semelhantes também são desafios a serem explorados.

- **Prever resultados de jogos, campeonatos ou quantidade de eventos**: a predição de resultados continua sendo um grande desafio, devido à natureza estocástica dos esportes. Nesse contexto, a modelagem de "novas estatísticas"para medição de desempenho das equipes é um desafio permanente. Quantificar a influência da sorte e identificar quais fatores são relevantes para determinar um resultado são sub-tarefas igualmente desafiadoras. Com o crescimento dos mercados de apostas esportivas, diversos outros tipos de predição também passaram a ter relevância, como, por exemplo, quantos escanteios terá um partida de futebol ou quantos gols serão marcados em cada parte do jogo.
- **Resolver desafios específicos de *Sports Betting Analytics***: a análise de dados na perspectiva do mercado financeiro tem uma série de desafios em aberto como, por exemplo: a avaliação de eficiência de mercado (modelos que verifiquem a hipótese de "mercado eficiente", na qual um apostador não consegue obter lucros superiores à média do mercado); a avaliação de estratégias de gestão de banca, difundidas entre os apostadores (*all-in*, Fibonacci, Martingale, Critério de Kelly, valor fixo, etc.); a modelagem de estratégias de *trading* automáticas ou a avaliação da eficiência de outras já utilizadas no mercado financeiro (*Scalping*, *Swing trader*, *Dutching*, etc.); ou ainda, a detecção de fraudes, a partir de apostas suspeitas que possam ter sido originadas a partir de jogos manipulados.

2.6. Considerações Finais

Uma nova era está chegando para o esporte e definitivamente "os dados"são os protagonistas desta revolução. Tudo está sendo mapeado. Jogadores praticam suas atividades com dispositivos que monitoram seus aspectos biométricos. Sistemas de visão computacional estão rastreando todos os passos dos atletas numa granularidade sem precedentes. Na parte administrativa, os clubes identificam o perfil dos seus fãs, descobrindo seu hábitos e analisando seus comentários em redes sociais.

No momento atual, a análise humana ainda é parte da solução. Entretanto, em um futuro próximo, o esporte estará centrado em uma quantidade de dados tão grande que o trabalho de análise não será mais adequado para um humano. O avanço final (que já começou) será a criação de sistemas para o processamento de todas essas informações em

tempo real para que todos os dados sejam transformados em conhecimento, permitindo tomadas de decisão de forma imediata.

O papel do treinador não será extinto, mas definitivamente receberá uma nova conotação. As máquinas serão mais capacitadas que os humanos para avaliar os dados e sugerir mudanças, mas essas mudanças ainda precisarão ser comandadas por pessoas (ao menos, em médio prazo).

Nesse futuro, as máquinas indicarão quais jogadores devem ser contratados e os treinamentos mais eficazes; alertarão quais jogadores estão próximos de se lesionar e como devem ser tratados; avaliarão as táticas da equipe, apontando as falhas e sugerindo as mudanças, em tempo real. Tudo estará, de alguma forma, sendo monitorado minuciosamente.

Os erros humanos de arbitragem também estão com prazo de validade. As regras dos jogos estarão "sob o olhar" de juízes eletrônicos, tornando o esporte mais justo. O mercado de apostas, assim como a bolsa de valores, deverá ser dominada por algoritmos preditivos. Enquanto os jogos de videogame terão times imbatíveis comandados pelas máquinas.

Ao mesmo tempo que esse futuro parece ser inevitável, as oportunidades de pesquisas em análise de dados esportivos se multiplicam. Os atletas continuarão protagonistas, mas os engenheiros e cientistas de dados assumirão papel de destaque nos bastidores. Uma competição paralela e silenciosa que já começou.

Referências

[1] Association for professional basketball research. http://www.apbr.org/. (Acessado em 21/08/2017).

[2] Cluster analysis of the nfl's top wide receivers | r-bloggers. https://www.r-bloggers.com/cluster-analysis-of-the-nfls-top-wide-receivers/. (Acessado em 21/08/2017).

[3] Exploding nba basketball shot heat map analysis - information aesthetics. http://infosthetics.com/archives/2012/06/exploding_nba_basketball_shot_heat_map_analysis.html. (Acessado em 21/08/2017).

[4] A guide to sabermetric research | society for american baseball research. http://sabr.org/sabermetrics. (Acessado em 21/08/2017).

[5] Movebank. https://www.movebank.org/. (Acessado em 21/08/2017).

[6] One sports - a vantagem que gera resultados e campeões. http://www.onesports.com.br/. Acessado em 21/08/2017).

[7] Opta home. http://www.optasports.com/. (Acessado em 21/08/2017).

[8] Predicting nba player positions - nyc data science academy blognyc data science academy blog. http://blog.nycdatascience.com/student-works/predicting-nba-player-positions/. (Accessed on 09/08/2017).

[9] Top 10 social apis: Facebook, twitter and google plus | programmableweb. https://www.programmableweb.com/news/top-10-social-apis-facebook-twitter-and-google-plus/analysis/2015/02/17. (Acessado em 21/08/2017).

[10] Adidas. http://www.adidas.com.br/, 2017. (Acessado em 21/08/2017).

[11] Beautiful soup documentation — beautiful soup 4.4.0 documentation. https://www.crummy.com/software/BeautifulSoup/bs4/doc/, 2017. (Acessado em 21/08/2017).

[12] Bet-at-home.com. https://www.bet-at-home.com/, 08 2017. (Acessado em 21/08/2017).

[13] Bet365.com. https://www.bet365.com/, 08 2017. (Acessado em 21/08/2017).

[14] Betdaq.com. https://www.betdaq.com/, 08 2017. (Acessado em 21/08/2017).

[15] Betfair.com. https://www.betfair.com/exchange/plus/, 08 2017. (Acessado em 21/08/2017).

[16] Bwin.com. https://sports.bwin.com/en/sports, 08 2017. (Acessado em 21/08/2017).

[17] Catapult. http://www.catapultsports.com/, 2017. (Acessado em 21/08/2017).

[18] Clube da aposta • ganhe dinheiro e aprenda como apostar. https://clubedaposta.com/, 2017. (Acessado em 21/08/2017).

[19] Investimento futebol | aprenda a investir e seja um trader esportivo. https://investimentofutebol.com/, 2017. (Acessado em 21/08/2017).

[20] Mercado das apostas - trading esportivo, trader esportivo. http://www.mercadodasapostas.com/, 2017. (Acessado em 21/08/2017).

[21] Oakland athletics. https://www.mlb.com/athletics, 2017. (Acessado em 21/08/2017).

[22] Pinnacle.com. https://www.pinnacle.com/, 08 2017. (Acessado em 21/08/2017).

[23] Requests: Http for humans — requests 2.18.4 documentation. http://docs.python-requests.org/en/master/, 2017. (Acessado em 21/08/2017).

[24] Rivalo.com. https://www.rivalo.com/pt/apostas/, 08 2017. (Acessado em 21/08/2017).

[25] Statista. https://www.statista.com/topics/1740/sports-betting/, 2017. (Acessado em 21/08/2017).

[26] Stats sport vu. https://www.stats.com/sportvu-football/, 2017. (Acessado em 21/08/2017).

[27] Using machine learning to find the 8 types of players in the nba. https://goo.gl/iiMHGp, 2017. (Acessado em 21/08/2017).

[28] Whoop. http://www.whoop.com/, 2017. (Acessado em 21/08/2017).

[29] Williamhill.com. http://sports.williamhill.com/bet/pt, 08 2017. (Acessado em 21/08/2017).

[30] ANDERSON, C., AND SALLY, D. Os números do jogo: Por que tudo o que você sabe sobre futebol está errado. *São Paulo: Paralela* (2013).

[31] AOKI, R., ASSUNCAO, R. M., AND DE MELO, P. O. Luck is hard to beat: The difficulty of sports prediction. *arXiv preprint arXiv:1706.02447* (2017).

[32] CHANDRASHEKAR, G., AND SAHIN, F. A survey on feature selection methods. *Computers & Electrical Engineering 40*, 1 (2014), 16–28.

[33] CHENG, T., HAWORTH, J., ANBAROGLU, B., TANAKSARANOND, G., AND WANG, J. Spatiotemporal data mining. In *Handbook of Regional Science*. Springer, 2014, pp. 1173–1193.

[34] COKINS, G., DEGRANGE, W., CHAMBAL, S., AND WALKER, R. Sports analytics taxonomy, v1.0 - informs. https://www.informs.org/ORMS-Today/Public-Articles/June-Volume-43-Number-3/Sports-analytics-taxonomy-V1.0, 2017. (Acessado em 21/08/2017).

[35] CONSTANTINOU, A. C., FENTON, N. E., AND NEIL, M. pi-football: A bayesian network model for forecasting association football match outcomes. *Knowledge-Based Systems 36* (2012), 322–339.

[36] CONSTANTINOU, A. C., FENTON, N. E., AND NEIL, M. Profiting from an inefficient association football gambling market: Prediction, risk and uncertainty using bayesian networks. *Knowledge-Based Systems 50* (2013), 60–86.

[37] FAYYAD, U. M., PIATETSKY-SHAPIRO, G., SMYTH, P., AND UTHURUSAMY, R. *Advances in knowledge discovery and data mining*, vol. 21. AAAI press Menlo Park, 1996.

[38] FODOR, I. K. A survey of dimension reduction techniques. Tech. rep., Lawrence Livermore National Lab., CA (US), 2002.

[39] FU, T.-C. A review on time series data mining. *Engineering Applications of Artificial Intelligence 24*, 1 (2011), 164–181.

[40] GARCIA, S., LUENGO, J., SÁEZ, J. A., LOPEZ, V., AND HERRERA, F. A survey of discretization techniques: Taxonomy and empirical analysis in supervised learning. *IEEE Transactions on Knowledge and Data Engineering 25*, 4 (2013), 734–750.

[41] HUANG, K.-Y., AND CHANG, W.-L. A neural network method for prediction of 2006 world cup football game. In *Neural Networks (IJCNN), The 2010 International Joint Conference on* (2010), IEEE, pp. 1–8.

[42] JONES, B. Where to find sports data | tableau public. `https://public.tableau.com/s/blog/2014/03/where-find-sports-data?elq=4d0891dfcfd3415e9b057dd998bf8cc3`, 2014. (Acessado em 21/08/2017).

[43] LEWIS, M. *Moneyball: The art of winning an unfair game*. WW Norton & Company, 2004.

[44] LINDEN, R. Técnicas de agrupamento. *Revista de Sistemas de Informação da FSMA 4* (2009), 18–36.

[45] MARQUEZ, D., NEIL, M., AND FENTON, N. Improved reliability modeling using bayesian networks and dynamic discretization. *Reliability Engineering & System Safety 95*, 4 (2010), 412–425.

[46] MOURA, F. A., MARTINS, L. E. B., ANIDO, R. O., RUFFINO, P. R. C., BARROS, R. M., AND CUNHA, S. A. A spectral analysis of team dynamics and tactics in brazilian football. *Journal of Sports Sciences 31*, 14 (2013), 1568–1577.

[47] REEP, C., AND BENJAMIN, B. Skill and chance in association football. *Journal of the Royal Statistical Society. Series A (General) 131*, 4 (1968), 581–585.

[48] SHEKHAR, S., ZHANG, P., HUANG, Y., AND VATSAVAI, R. R. Spatial data mining.

[49] STEIN, M., JANETZKO, H., SEEBACHER, D., JÄGER, A., NAGEL, M., HÖLSCH, J., KOSUB, S., SCHRECK, T., KEIM, D. A., AND GROSSNIKLAUS, M. How to make sense of team sport data: From acquisition to data modeling and research aspects. *Data 2*, 1 (2017), 2.

[50] TOLBERT, B., AND TRAFALIS, T. Predicting major league baseball championship winners through data mining.

[51] TÜFEKCI, P. Prediction of football match results in turkish super league games. In *Proceedings of the Second International Afro-European Conference for Industrial Advancement AECIA 2015* (2016), Springer, pp. 515–526.

[52] ZHAO, Y., AND CEN, Y. *Data mining applications with R*. Academic Press, 2013.

Sobre os Autores

Ígor Barbosa da Costa é professor de computação do Instituto Federal de Educação, Ciência e Tecnologia da Paraíba (IFPB), campus Campina Grande. Graduado em Ciência da Computação pela Universidade Federal de Campina Grande – UFCG (2006) e Mestre pela Universidade Federal de Pernambuco – UFPE (2010). Tem experiência na área de Ciência da Computação, com ênfase em Bancos de Dados e Desenvolvimento, atuando principalmente nos seguintes temas: Mineração de Dados e Descoberta de Conhecimento (com foco atual em dados esportivos).

Carlos Eduardo Santos Pires concluiu a Graduação em Ciência da Computação, em 1997, pela Universidade Federal de Campina Grande (UFCG) e Mestrado em Informática, em 2000, pela mesma instituição. Em 2009, concluiu o Doutorado na Universidade Federal de Pernambuco (UFPE), tendo realizado Doutorado-Sanduíche na Université de Versailles, na França. Atualmente é Professor Adjunto do Departamento de Sistemas e Computação (DSC), da Universidade Federal de Campina Grande (UFCG). Tem experiência na área de Ciência da Computação, com ênfase em Bancos de Dados, atuando principalmente nos seguintes temas: Qualidade de Dados, Integração de Dados, Descoberta de Conhecimento e Big Data.

Leandro Balby Marinho é Doutor em Ciência da Computação, pela Universidade de Hildesheim, Alemanha, 2010. Mestre em Engenharia Elétrica, UFMA, Brasil, 2005. Bacharel em Ciência da Computação, UFMA, Brasil, 2002. Professor do Departamento de Sistemas e Computação da Universidade Federal de Campina Grande (UFCG). Atua como docente, pesquisador e orientador nos cursos de graduação e pós-graduação em ciência da computação. Suas áreas de especialização são: Inteligência Artificial, Mineração de Dados e Recuperação da Informação.