Análise entre algoritmos de aprendizado de máquina para suportar a predição do posicionamento do jogador de futebol

Randal Gasparini [1]
Alexandre Álvaro [2]

Resumo: O esporte de alto desempenho está em busca, cada vez mais, do suporte de tecnologias que visam o auxílio aos atletas e aos profissionais que os acompanham. Atualmente existem diversas tecnologias que são utilizadas no segmento, como análise de deslocamento de jogadores através de GPS (*Global Position System*), mapeamento através da captura de vídeo, monitoramento de frequência cardíaca, acelerômetro, giroscópio, dentre outros. Entretanto, a comissão técnica possui pouco respaldo tecnológico de ferramentas que consideram as características de movimentação do jogador, as quais poderiam apoiá-las na tomada de decisão em relação ao posicionamento ideal do atleta em campo. Através da coleta de dados de GPS é possível determinar velocidade, distância percorrida, faixas de aceleração, posicionamento, dentre outras informações. Nesse sentido, esse trabalho visa analisar o posicionamento ideal de um jogador de futebol através de algoritmos de aprendizado de máquina, sendo utilizados os modelos de Regressão Logística com Regularização, Redes Neurais e Máquinas de Vetores de Suporte. Os resultados obtidos indicaram que o modelo SVM se sobressaiu aos demais, entretanto são necessários trabalhos futuros para a busca de uma taxa de acerto mais elevada.

***Palavras-chave*:** aprendizado de máquina, análise futebolística, gps, máquinas de vetores de suporte, redes neurais artificiais, regressão logística

Abstract: *The high sport performance is always searching the best technology to support athletes and all professional that helps him. Today there are many technologies in this area, like GPS (Global Position System) motion analysis, video motion analysis, heart rate monitoring, accelerometer, gyroscope and others. The coach has low support from technologies tools that analisys soccer player's characteristics in a game. This informations could help him to decide the ideal position of the athletics. Getting GPS data is possible to define parameters as speed, running distance, acceleration ranges, field positions and others informations. The objective of this study is to find the ideal position for a soccer player using these three machine learning algoritms: Logistic regression with regularization, artificial neural networks and support vector machines. The results show that support vector machines is better than three, but are need future works to find a higher hit rate.*

***Keywords*:** *artificial neural networks, gps, learning machine, logistic regression, soccer analysis, support vector machines*

1 Introdução

O posicionamento dos jogadores de futebol é, em suma, definido prioritariamente pelo treinador do time. Essa análise e decisão são baseadas inteiramente na percepção do técnico e, em alguns casos, conforme as características anatômicas e antropométricas do atleta [2]. A antropometria é o ramo das ciências humanas que estuda as medidas do corpo, particularmente o tamanho e a forma [1]. Por se tratar de um esporte coletivo é de suma

[1]Programa de Pós-Graduação em Ciência da Computação (PPGCC) - Universidade Federal de São Carlos (UFSCar) - Campus Sorocaba - Sorocaba (SP) - Brasil
{gasparini.randal@gmail.com}
[2]Departamento de Computação (DComp) - Universidade Federal de São Carlos (UFSCar) - Campus Sorocaba - Sorocaba (SP) - Brasil
{alvaro@ufscar.br}

http://dx.doi.org/10.5335/rbca.v9i2.6454

importância que os jogadores de futebol tenham as suas posições bem definidas dentro do time, uma vez que isso contribui para a formação tática e impacta diretamente na atuação da equipe.

Sob o olhar da tecnologia e seus avanços, diversos recursos são utilizados para avaliar o desempenho do atleta profissional em seu treinamento e também durante os jogos. Desse modo, é possível identificar algumas tecnologias que objetivam avaliar o desempenho do atleta profissional, dentre elas: dispositivo com GPS e acelerômetro para monitorar o deslocamento do jogador, plataforma de salto para avaliar o desgaste muscular, plataforma de foto-célula para monitorar a explosão muscular, frequencímetro visando avaliar a performance do batimento cardíaco, dentre outros dispositivos de monitoramento [3].

Por outro lado, esses equipamentos coletam informações individualizadas e não as processam para obter dados relevantes e estratégicos para a comissão técnica. Nesse sentido, a área de inteligência artificial pode auxiliar fornecendo apoio de algoritmos específicos que são capazes de criar sistemas de auxílio à tomada de decisão. Mais especificamente, os estudos em aprendizado de máquina permitem que determinadas técnicas computacionais sejam treinadas através de entradas selecionadas de dados, para posteriormente aprender qual a saída (rótulo) mais indicada para determinado conjunto de parâmetros.

Esse estudo tem como objetivo analisar, de forma anônima, dados de times da Séria A do Campeonato Brasileiro sob o aspecto quantitativo, pois envolvem números e classes, sendo adotados métodos compatíveis. Os algoritmos utilizados foram: Regressão Logística, Redes Neurais e SVM. A seção 2 apresenta os trabalhos relacionados; a Seção 3 salienta um breve *background* da área esportiva; a Seção 4 apresenta o estudo, a aplicação dos algoritmos e seus resultados; e a Seção 5 conclui esse trabalho, bem como apresenta os principais direcionamentos para Trabalhos Futuros.

2 Trabalhos Relacionados

O futebol profissional, em âmbito mundial e igualmente no Brasil, recebe altos investimentos em tecnologia, processos, equipamentos e profissionais qualificados. Analisar os dados individuais da atuação de cada jogador em campo não é, definitivamente, uma tarefa fácil. Métodos tradicionais podem ser morosos e dispendiosos para o clube. A informação rápida e atual é um poderoso diferencial para o time, seus jogadores e a comissão técnica [4].

A análise por vídeo, a fim de obter a movimentação e distribuição dos jogadores, é um dos métodos passíveis de utilização. Nos anos noventa, a coleta dessas informações se dava através da repetição de jogos gravados em fitas de vídeo. Observadores anotavam todas as informações possíveis, como chutes a gol, faltas, passes longos, dentre outros. Em paralelo, análises físicas eram realizadas, como testes de corridas, velocidade e resistência anaeróbica dos atletas[3]. Todos esses dados eram confrontados, obtendo informações preciosas sobre o time e seus jogadores [6].

Mais recentemente a mesma técnica de utilização de imagens continua a ser empregada, mas de forma muito mais automática e autônoma. A captura ocorre através de modernas câmeras instaladas em pontos estratégicos e de boa amplitude visual no estádio. Esse método é denominado de Sistema de Rastreamento Automático (SRA). A captura dos frames é de excelente qualidade visual, a fim de que sejam analisadas posteriormente por um software específico, o qual baseia seus cálculos na análise de variância, tomando por base as posições obtidas a partir das imagens [7, 8].

Outra técnica possível e muito viável é a obtenção dos dados dos jogadores através de equipamentos baseados no sistema Global Position System (GPS). Esses são carregados junto às roupas dos jogadores durante as suas atividades físicas, principalmente durante os jogos oficiais. Posteriormente, as informações coletadas são transferidas para um software específico, o qual também é capaz de gerar dados quantitativos sobre os atletas [9].

Ambos os modelos são capazes de determinar o trajeto executado pelo atleta, permitindo assim a obtenção de mais informações, como distância percorrida, organização posicional do time, características individuais, dentre outras. Em relação à acurácia das tecnologias de coleta aqui apresentadas, ambas possuem um nível de precisão confiável e similar entre si. Desse modo é seguro optar por qualquer um dos métodos [10].

Comparando o processo de instalação dos equipamentos para a captura dos dados para ambos os métodos,

[3]capacidade de repetir por diversas vezes a corrida na faixa máxima de aceleração, sem perda considerável de velocidade [5]

é preciso considerar que o SRA exige um considerável investimento e preparação, uma vez que um conjunto de câmeras específicas é necessário. Somado a isso, a instalação é muitas vezes complexa, pois os pontos de fixação dos equipamentos são estratégicos e precisam ter grande amplitude visual [11, 7]. Por sua vez, o GPS demanda a necessidade de equipamentos acoplados à roupa de cada atleta, além da tecnologia e software compatíveis para a interpretação dos dados [10].

Considerando a disponibilização de dados GPS de jogadores profissionais da primeira divisão do futebol brasileiro pela empresa OneSports[4], esse estudo tem como proposta validar ou refutar a teoria de que é possível inferir, através de algoritmos de aprendizado de máquina, o posicionamento ideal dos jogadores de futebol através de coordenadas GPS. Os resultados obtidos serão analisados, determinando a viabilidade da aplicação das técnicas propostas.

3 Contextualização do Futebol de Campo

Um time de futebol é composto por 11 jogadores, sendo que todos possuem posições bem definidas e propostas específicas, considerando o contexto coletivo e individual. Com exceção do goleiro, o qual possui pouca movimentação em campo, os demais possuem deslocamentos, acelerações e desacelerações específicas dentro de um time, conforme a sua posição e a organização tática da equipe.

O objetivo geral de qualquer time de futebol é chegar à meta adversária, uma vez que vence o jogo aquele que efetuar o maior número de gols. Entretanto, é importante que cada jogador conheça, respeite e exerça a sua posição tática dentro de campo, uma vez que se trata de um esporte coletivo. A Tabela 1 demonstra os nomes usualmente adotados para as posições dos jogadores de futebol - exceto goleiro. A mesma nomenclatura será utilizada nesse estudo.

Tabela 1: Posicionamento dos jogadores de futebol que serão considerados nesse estudo [12]

Posição	Descrição
Atacante	Jogador que tem por objetivo receber a bola no campo de ataque do seu time e avançar à grande área adversária
Atacante de Área	Recebe a mesma função de um atacante, entretanto o seu raio de atuação está mais focado dentro da área do time oponente
Lateral Direito	Jogador que atua pelo lado direito do seu time. A sua principal função é permitir que a bola saia do setor defensivo e chegue até os jogadores do meio campo. Por possuírem características de velocistas, muitas vezes fazem a ligação direta com os jogadores de ataque
Lateral Esquerdo	Mesma função que o lateral direito, entretanto atua pelo lado esquerdo
Meio Campista	Jogadores que podem ajudar o ataque ou a defesa. Ficam posicionados no centro de campo e têm papel importante no trabalho da jogada ofensiva
Volante	Jogadores que atuam alternadamente entre o meio de campo e a defesa. Tem por objetivo principal anular o ataque adversário, gerando, sempre que possível, contra-ataques a favor do seu time
Zagueiro	Jogadores designados especificamente para defender o seu time e eliminar o risco de gol do adversário

A relevância desse estudo está ligada à qualidade da seleção de jogadores para determinada função dentro de um time. Uma vez que a decisão final cabe ao treinador, usualmente pela sua observação e conhecimento, há riscos de uma classificação incorreta do posicionamento do atleta, podendo comprometer o desempenho técnico e tático do time. Outro fator importante é que o próprio futuro do jogador pode sofrer prejuízos, uma vez que ele pode não ser o esportista mais indicado para a posição que vem jogando.

[4]http://www.onesports.com.br

4 Análise entre Algoritmos de Aprendizado de Máquina

4.1 Estudo

A base de dados, a qual respalda esse estudo, contém informações de posicionamento, aceleração, distância percorrida pelos jogadores de futebol, dentre outras, obtidas a partir de posicionamento GPS. A coleta foi realizada através de chips instalados nas roupas dos atletas.

A cardinalidade da tabela, com dados brutos, ou seja, sem qualquer tratamento, é de 1.518 tuplas.

Os dados foram fornecidos por uma empresa especializada na coleta de dados GPS de atletas profissionais de futebol, a OneSports. As informações que servirão como objeto desse estudo são de três dos principais times brasileiros de futebol, na categoria masculina. Os dados foram coletados entre os anos de 2013 e 2016. A base original detalha as ocorrências de cada jogador, contendo as seguintes colunas:

- Descrição e resultado do jogo;
- Posição do jogador;
- Data do jogo;
- Hora de início;
- Hora de término;
- Duração em minutos;
- Velocidade máxima do jogador em quilômetros;
- Velocidade média do jogador em quilômetros;
- Aceleração máxima em metros por segundo;
- Distância total percorrida em quilômetros;
- Metros percorridos com velocidade de 0.01 a 5.99 quilômetros por hora;
- Metros percorridos com velocidade de 6 a 7.99 quilômetros por hora;
- Metros percorridos com velocidade de 8 a 9.99 quilômetros por hora;
- Metros percorridos com velocidade de 10 a 14.99 quilômetros por hora;
- Metros percorridos com velocidade de 15 a 22.99 quilômetros por hora;
- Metros percorridos com velocidade acima de 23 quilômetros por hora;
- Percentual dos metros percorridos com velocidade de 0.01 a 5.99 quilômetros por hora;
- Percentual dos metros percorridos com velocidade de 6 a 7.99 quilômetros por hora;
- Percentual dos metros percorridos com velocidade de 8 a 9.99 quilômetros por hora;
- Percentual dos metros percorridos com velocidade de 10 a 14.99 quilômetros por hora;
- Percentual dos metros percorridos com velocidade de 15 a 22.99 quilômetros por hora;
- Percentual dos metros percorridos com velocidade acima de 23 quilômetros por hora;
- Tempo de atividade com velocidade de 0.01 a 5.99 quilômetros por hora;
- Tempo de atividade com velocidade de 6 a 7.99 quilômetros por hora;
- Tempo de atividade com velocidade de 8 a 9.99 quilômetros por hora;

- Tempo de atividade com velocidade de 10 a 14.99 quilômetros por hora;
- Tempo de atividade com velocidade de 15 a 22.99 quilômetros por hora;
- Tempo de atividade com velocidade acima de 23 quilômetros por hora;
- Número de ações com velocidade entre 15 a 22.99 quilômetros por hora;
- Número de ações com velocidade acima de 23 quilômetros por hora.

Ainda sobre a base de dados, é importante destacar que cada tupla contém os dados de apenas um tempo do jogo. As últimas duas colunas referem-se ao número total de ações nas referidas velocidades, ou seja, por quantas vezes o jogador entrou/atingiu as devidas faixas de aceleração.

4.2 Objetivos

Considerando a importância de auxiliar, dar suporte e apoio na decisão do treinador quanto à posição ideal de um atleta, esse trabalho objetiva:

- Aplicar algoritmos específicos de aprendizado de máquina a fim de inferir rótulos para os atletas;
- Com os resultados gerados, realizar uma análise a fim de detectar o algoritmo mais indicado para o estudo proposto;

4.3 Hipóteses de Pesquisa

O objetivo central desse estudo é determinar qual técnica de aprendizado de máquina melhor define a posição ideal de um jogador de futebol através do seu posicionamento GPS.

A fim de responder a questão de pesquisa, as seguintes hipóteses foram definidas:

- H_{a0} - Nenhum dos algoritmos propostos obteve uma taxa de acerto confiável;
- H_{a1} - O algoritmo de Regressão Logística com Regularização demonstrou ser mais confiável quanto à taxa de acerto, quando comparado com os demais;
- H_{a2} - O algoritmo de Redes Neurais demonstrou ser mais confiável quanto à taxa de acerto, quando comparado com os demais;
- H_{a3} - O algoritmo de Máquinas de Vetores de Suporte demonstrou ser mais confiável quanto à taxa de acerto, quando comparado com os demais;
- H_{a4} - Todos os algoritmos se demonstraram confiáveis, retornando taxas de acerto com um grau de confiabilidade aceitável e seguro.

4.4 Técnicas de Aprendizado de Máquina utilizadas

Como técnicas de comparação na predição dos dados, serão utilizados os seguintes algoritmos de aprendizado de máquina:

- Regressão Logística com Regularização

  Método supervisionado para predição de rótulos através de parâmetros binários. Para bases onde os dados não são desse tipo, como o estudado nesse caso, é realizada uma regularização a fim de manter todos os parâmetros num intervalo escalar reduzido [13];

- Redes Neurais Artificiais

  Baseado no funcionamento e comunicação dos neurônios do ser humano, uma rede neural artificial tem características não-lineares. Para ser funcional, primeiramente é necessário um treinamento com uma base conhecida, onde a rede aprende o relacionamento dos pares entrada-saída, sendo capaz de, posteriormente, inferir rótulos às entradas ainda desconhecidas. Para problemas com mais de uma classe de saída, como o encontrado nesse trabalho, é necessária uma transformação no tipo do rótulo, passando de inteiro para vetor categórico de *bits*, o qual foi realizado, a fim de atender à especificidade do algoritmo [14]. O método será supervisionado nesse estudo, uma vez que é apresentado o conjunto de treino com as suas respectivas saídas;

- Máquinas de Vetores de Suporte

  Método supervisionado para predição de rótulos utilizando técnicas de regressão e classificação. Baseia-se na construção de hiperplanos ou espaços dimensionais infinitos. Trata-se de um classificador com técnicas avançadas e de considerável robustez nas predições [15];

Para todos os algoritmos mencionados acima, existe a necessidade de treinamento prévio com amostras e rótulos já conhecidos. Essa etapa é fundamental, uma vez que permite o ajustamento do classificador e possibilita gerar predições de modo mais consistente para entradas desconhecidas [16].

A maneira mais usual de validação dos rótulos inferidos pelo algoritmo de predição está na separação da base de dados em partes: treinamento, validação e teste. Essa divisão pode ocorrer através de dinâmicas diferentes, mas consiste basicamente em treinar o modelo com uma porcentagem da base, validar[5] com outra fração de dados e testar com o restante. Desse modo, como o rótulo real já é conhecido - pois está na base, o mesmo é comparado com a predição do algoritmo. Ao término desse processo é possível medir o desempenho do método, através do percentual de acerto, o qual representa as amostras devidamente classificadas.

4.5 Operação

4.5.1 Remoção de colunas irrelevantes

Uma vez que a base de dados foi fornecida na sua totalidade, e desse modo seria impossível a extração de dados paramétricos relevantes, colunas sem significância foram removidas, conforme descrito na Tabela 2.

4.5.2 Remoção de tuplas inconsistentes

A segunda análise sobre a base consistiu em identificar tuplas inconsistentes. Essa tarefa é importante, pois elimina dados irrelevantes, podendo confundir o algoritmo de aprendizado de máquina. As tuplas eliminadas seguiram o critério:

- Aquelas com informações GPS em branco, onde provavelmente ocorreu alguma falha na leitura;
- Tuplas com informações sobre a calibragem[6] do aparelho
- Tuplas com informações sobre treinos técnicos e táticos
- Seis tuplas foram removidas, pois representavam atividades que duraram dez minutos ou menos. Foram consideradas apenas registros com tempo igual ou superior a dez minutos, o que representa uma taxa maior que 20% de um período de 45 minutos de uma partida oficial de futebol

[5] Etapa que calcula os melhores parâmetros a fim de ajustar melhor o classificador

[6] A fim de manter a acurácia e precisão do GPS, o aparelho precisa muitas vezes ser calibrado, a fim de realizar a leitura correta das posições do jogador em campo [17]

Tabela 2: Colunas da base original e os tratamentos sofridos a fim de que seja possível a parametrização dos dados

Coluna(s)	Justificativa
Descrição e Resultado do Jogo	O resultado do jogo foi desprezado, uma vez que não é o foco desse estudo analisar o desempenho coletivo do time. A análise ocorre de maneira individual, ou seja, analisando o comportamento isolado de cada jogador
Data	Condições climáticas podem mudar o comportamento do jogador, entretanto a data do jogo passou a ser irrelevante uma vez que não existem informações adicionais, como temperatura, umidade, velocidade do vento, dentre outras
Hora de Início e Final	Os horários de início e término do jogo não são relevantes para a análise GPS
Percentual dos metros percorridos em determinada velocidade	A informação é idêntica aos metros percorridos, entretanto utilizando outra notação. Isso gera redundância de informações, sendo assim removida
Tempo de atividade com determinada velocidade	As colunas tempo e metros são dispostas de dados diferentes, entretanto a essência da informação é igual, sendo que ambas representam os metros percorridos em determinada velocidade, podendo ser desconsiderada nesse caso, pois promoverá redundância de informações

4.5.3 Conversões

Com o objetivo de manter os dados sob uma mesma ótica, algumas conversões de tipo foram realizadas, sendo essas:

- Substituição do separador da casa decimal de vírgula para ponto, mantendo assim o padrão requerido pela tipagem flutuante;
- Conversão dos campos com tipo tempo (hora) para ponto flutuante;
- As posições dos jogadores foram alteradas de texto para números, permitindo a sua categorização, conforme pode ser observado na Tabela 3.

Tabela 3: Posição dos jogadores convertida de texto para número, permitindo assim a sua categorização.

Posição	Conversão	Tuplas
Atacante	1	158
Lateral	2	352
Meio Campista	3	125
Volante	4	455
Zagueiro	5	334
		1424 (total)

4.5.4 Balanceamento dos dados

Conforme pode ser observado na Tabela 3, não existe um equilíbrio na quantidade de tuplas para cada posição. Essa característica pode tornar o algoritmo de predição tendencioso, uma vez que existe um desbalanceamento

natural dos dados. Segundo [18] existem duas formas de contornar esse cenário, seja replicando tuplas das classes minoritárias ou eliminando informações das majoritárias. Nesse estudo, será adotada a segunda abordagem, diminuindo o volume de dados, conforme demonstrado na Tabela 4.

Tabela 4: Dados após o balanceamento da base.

Posição	Conversão	Tuplas
Atacante	1	125
Lateral	2	125
Meio Campista	3	125
Volante	4	125
Zagueiro	5	125
		625 (total)

Ainda com o objetivo de evitar qualquer tipo de tendência sobre os dados, os mesmos foram reorganizados de modo aleatório, evitando o agrupamento de tuplas com rótulos iguais.

4.6 Visão do Arquivo de Saída

Após realizadas todas as etapas do pré-processamento da base, a mesma está disponível para ser utilizada pelos algoritmos propostos nesse estudo. A tabela 5 traz um pequeno trecho da base após todo o tratamento, permitindo assim a visualização final dos dados.

Tabela 5: Visão parcial da base após pré-processamento

Vel Máx*	Vel Méd*	Acel Máx (m/s^2)	Dist Total (m)	% Tempo 15 a 23*	% Tempo > 23*	Ações 15 a 23*	Ações > 23*	Posição (rótulo)
21.13	2.94	3.17	1460	30	0	3	0	1
25.06	3.86	3.93	1479	77	14	5	1	4
26.31	4.01	3.29	1485	86	7	11	1	5
29.08	4.85	3.95	1538	184	42	17	3	2
29.08	4.85	3.95	1538	184	42	17	3	2
29.85	4.63	4.43	1557	64	15	7	1	4
27.6	4.5	2.78	1562	121	29	8	3	3
27.6	4.5	2.78	1562	121	29	8	3	3
27.6	4.5	2.78	1562	121	29	8	3	3
27.6	4.5	2.78	1562	121	29	8	3	3
26.43	5.53	4.06	1597	178	66	9	5	2
26.43	5.53	4.06	1597	178	66	9	5	2

*km/h

4.6.1 Aplicação dos Algoritmos

Uma vez tratada a base de dados, conforme descrito na seção 4.5, passa-se à implementação e execução dos algoritmos detalhados na seção 4.4.

Para a construção e execução dos classificadores foi utilizada a ferramenta MatLab®R2014b, valendo-se das *toolboxes* oferecidas pela ferramenta. Os algoritmos foram divididos em arquivos separados. A estrutura de código para os três modelos a serem testados é a que segue:

- Carregamento dos dados;

- Separação das amostras e dos rótulos em matrizes separadas;
- Tratamento dos rótulos, quando necessário;
- Normalização das amostras, quando necessária;
- Divisão da base em treinamento, validação e teste;
- Dez execuções do algoritmo englobando treinamento, ajuste através da validação e teste nas frações dos dados, realizando assim o *10-fold*;
- Por fim serão calculados os falsos positivos, verdadeiros positivos, falsos negativos e verdadeiros negativos, obtendo assim a acurácia do modelo.

A fim de aumentar a confiança dos resultados obtidos, cada algoritmo foi executado dez vezes, além do *10-fold*. Os resultados foram consistentes em todas as execuções, demonstrando assim consistência no modelo e nos resultados gerados. Abaixo é possível visualizar parte do trecho de código utilizado para a regressão logística.

```
1. load  dados.csv;
2. amostras = dados(1:end,1:end-1);
3. rotulos = dados(1:end,end);
4. amostras = normaliza(amostras);
5. rotulos = categorical(rotulos);
6. [B,dev,stats] = mnrfit(amostras, rotulos);
```

4.7 Análise

A análise e validação apenas são possíveis pelo fato de toda a base de dados ser rotulada, ou seja, é conhecida a posição ideal do jogador, segundo a classificação do treinador do time. É possível, desse modo o cálculo da acurácia, conforme os verdadeiros positivos (vp), verdadeiros negativos (vn), falsos positivos (fp) e falsos negativos (fn), através da seguinte fórmula:

$$Ac = \frac{vp + vn}{vp + vn + fp + fn} \tag{1}$$

A definição em detalhes dos parâmetros é demonstrada na Tabela 6.

Tabela 6: Detalhamento dos parâmetros utilizados para calcular a acurácia de cada modelo

Verdadeiros Positivos	Quando o algoritmo e treinador fazem uma afirmação acertada sobre a posição de um determinado jogador e ambos estão corretos
Verdadeiros Negativos	Quando o algoritmo e treinador fazem uma negação acertada sobre a posição de um determinado jogador e ambos estão corretos
Falsos Positivos	Quando o algoritmo faz uma predição diferente do rótulo, entretanto a definição do treinador está correta
Falsos Negativos	Quando o algoritmo faz uma predição diferente do rótulo, entretanto a definição do treinador está incorreta, ficando confirmada a informação inferida pelo algoritmo

A princípio, o cenário encontrado nesse estudo não permite a checagem dos Falsos Positivos e Falsos Negativos, pois não é possível validar a informação com o treinador e com o atleta.

A fim de obter uma alternativa à limitação explicada acima, foi calculado o desvio padrão da posição. Desse modo, para os casos onde houve divergência da definição do posicionamento ideal entre a predição do algoritmo e do treinador, é possível a realização de uma nova checagem, mas com base no desvio padrão do jogador *versus* o desvio padrão da posição. Desse modo é possível obter um resultado a partir de outra métrica, gerando novas informações que permitissem aprofundar o entendimento da acurácia obtida, conforme demonstrado na tabela 6.

Segundo [4] é possível determinar a média do desvio padrão de um time, possibilitando assim que as características posicionais sejam analisadas sob essa perspectiva. Nesse estudo, a média do desvio padrão também foi calculada da maneira apropriada, entretanto os resultados obtidos demonstraram-se ineficientes para definir a posição do jogador, uma vez que a variação para algumas posições é muito pequena, gerando um diferencial comparativo muito tênue, conforme pode ser verificado na tabela 7.

Tabela 7: Média do desvio padrão muito próxima para algumas posições

Atacante	1140
Lateral	1188
Meio Campista	1177
Volante	1007
Zagueiro	1127

Desse modo, esse estudo considerará a rotulagem da base como certa e indiscutível, permitindo o cálculo da acurácia e posterior escolha do melhor algoritmo para o problema em questão. Os resultados obtidos são demonstrados na Tabela 8.

Tabela 8: Base balanceada: Resultados das taxas de acerto de cada algoritmo após dez execuções em *10-fold*

Algoritmo	Taxa de Acerto	Taxa de Erro
SVM	57.44%	42.56%
Redes Neurais	54.25%	45.75%
Regressão Logística	45.33%	54.67%

A Figura 1 evidencia a superioridade do modelo SVM perante os outros modelos, destacando a taxa de acerto descendente e o erro ascendente quando comparados com os outros algoritmos.

4.8 Aplicação dos Algoritmos com a Base Desbalanceada

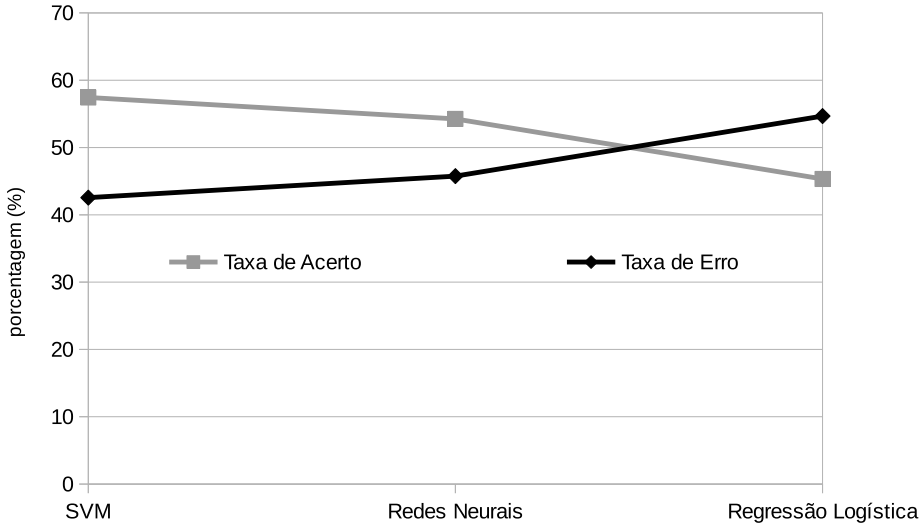
A fim de melhor entender o desempenho dos algoritmos com a base e o problema proposto e, evitar qualquer indução da base original favoreça um determinado algoritmo, optou-se por uma nova execução dos algoritmos utilizando a base pré-processada, entretanto sem a execução do balanceamento das classes. Ao término desse, o resultado divergiu quando comparado com a aplicação utilizando a base balanceada, demonstrando superioridade da taxa de acerto para o algoritmo de regressão logística, conforme pode ser conferido na tabela 9.

Tabela 9: Base desbalanceada: Resultados das taxas de acerto de cada algoritmo após dez execuções em *10-fold*

Algoritmo	Taxa de Acerto	Taxa de Erro
SVM	51.47%	48.53%
Redes Neurais	48.13%	51.87%
Regressão Logística	58.07%	41.93%

Apesar do resultado ser divergente nesse teste, é prudente não considerá-lo. Segundo [18] o desbalanceamento de uma base não tende a prejudicar o desempenho e a acurácia do algoritmo de aprendizado de máquina. Entretanto é prejudicial que a fase de treinamento ocorra com uma base desbalanceada, uma vez que o modelo

Figura 1: Gráfico que demonstra a superioridade do modelo SVM quando comparado com os outros algoritmos



ganha características tendenciosas. Nesse estudo a mesma base para serviu para treinamento, validação e teste, utilizando para tal o *10-fold*. Desse modo, não é confiável aceitar um resultado originado a partir do treinamento com classes majoritárias e minoritárias, permanecendo como válido o resultado obtido a partir da aplicação dos algoritmos com a base balanceada.

4.9 Ameaças à Validade

As principais ameaças à validade foram tratadas previamente nesse trabalho, principalmente àquelas referentes à regularização da base de dados, repetidas execuções dos algoritmos e validação por *10-fold*.

Entretanto é necessário destacar a existência de duas possíveis ameaças à validade. A primeira refere-se ao fato do jogador atuar em uma posição definida pelo técnico, podendo assim assumir características próprias da função, ou seja, os seus parâmetros de posicionamento em campo podem se sobressair às suas características individuais, distorcendo os dados coletados.

A segunda ameaça refere-se à baixa cardinalidade da tabela após o balanceamento dos dados. Sempre que possível é recomendável a utilização de bases com algumas milhares de tuplas em aprendizado de máquina, uma vez que esse fator permite um melhor ajustamento do algoritmo. Entretanto convém reforçar a importância do balanceamento que foi executado no pré-processamento, mesmo resultando em uma cardinalidade menor. Durante a realização desse trabalho, por motivos técnicos, não foi possível a obtenção de bases com um maior volume de tuplas.

5 Conclusões e Trabalhos Futuros

Esse trabalho buscou analisar os algoritmos regressão logística com regularização, redes neurais e SVM, identificando assim aquele com melhor taxa de acerto na predição do posicionamento de jogadores de futebol através de dados GPS.

A regressão logística com regularização obteve a menor taxa de acerto (45.33%) entre os três métodos estudados. Desse modo pode-se refutar a hipótese H_{a1}.

A rede neural artificial obteve uma taxa de acerto superior em relação à regressão logística com regularização (54.25%), entretanto ainda abaixo do modelo SVM, refutando assim a hipótese H_{a2}.

O modelo SVM obteve a melhor taxa de acerto dos três métodos analisados (57.44%), demonstrando ser o algoritmo mais indicado para o problema em questão. Entretanto, é necessário considerar que a sua taxa de acerto foi abaixo de 60%. Desse modo, para esse estudo é possível confirmar a hipótese nula H_{a0} e também refutar a hipótese H_{a3}.

Como evidenciado, apesar da hipótese H_{a3} se confirmar para essa análise, não é possível confiar integralmente na predição do modelo SVM para o cenário aplicado, pois a taxa de erro ainda é alta, podendo induzir a um posicionamento equivocado do jogador.

Considerando a aceitação da hipótese H_{a0}, fica também refutada a hipótese H_{a4}.

Limitando-se exclusivamente aos procedimentos aplicados, os resultados obtidos e considerando as deficiências encontradas na base de dados original, como baixa cardinalidade e disparidade na quantificação das classes, há motivos claros para a superioridade do modelo SVM. Trata-se de um algoritmo baseado em otimização, criando assim um hiperplano, o qual permite um melhor ajustamento e separação das classes. A sua robustez, em comparação aos demais métodos, permite um melhor desempenho às predições de problemas reais, vindo ao encontro desse trabalho. Por fim, é um modelo que aceita bem dados complexos não-lineares. Não existe uma complexidade considerável na base estudada, entretanto, a mesma demonstra uma não-linearidade muito expressiva, uma vez que o desvio padrão para cada posição se apresenta muito alto, conforme pode ser visualizado na tabela 7 [19].

O resultado obtido nesse estudo pode ter sido influenciado por alguns motivos, englobando a base de dados, informações tendenciosas ou o emprego de algoritmos que não foram objetos de análise nesse estudo. Desse modo, fica aberta a possibilidade para os seguintes trabalhos futuros:

- Repetição do estudo utilizando os mesmos algoritmos, mas com uma base de dados com maior cardinalidade;
- Aplicação de outros métodos supervisionados de aprendizado de máquina a fim de analisar a taxa de acerto;
- Aplicação de outros métodos não-supervisionados de aprendizado de máquina a fim de analisar a taxa de acerto em relação aos métodos supervisionados;
- Obtenção de dados com mais métricas a fim de eliminar qualquer ameaça à validade.

Agradecimentos

Os autores agradecem à OneSports pela disponibilização dos dados de GPS dos times de futebol.

Referências

[1] RODRIGUEZ-AÑEZ, C. R. A antropometria e sua aplicação na ergonomia. *Revista Brasileira de Cineantropometria & Desenvolvimento Humano*, v. 3, n. 1, p. 102–108, 2001. ISSN 1415-8426. Disponível em: <http://portalbiocursos.com.br/ohs/data/docs/51/20-_A_ANTROPOMETRIA_E_SUA_APLICAYYO_NA_ERGONOMIA.pdf>. Acesso em: 30 jun. 2017.

[2] GIL, S. M. et al. Physiological and anthropometric characteristics of young soccer players according to their playing position: relevance for the selection process. *The Journal of Strength & Conditioning Research*, LWW, v. 21, n. 2, p. 438–445, 2007. ISSN 1064-8011. Disponível em: <http://journals.lww.com/nsca-jscr/Fulltext/2007/05000/PHYSIOLOGICAL_AND_ANTHROPOMETRIC_CHARACTERISTICS.26.aspx>. Acesso em: 30 jun. 2017.

[3] OKAZAKI, V. H. A. et al. Ciência e tecnologia aplicada à melhoria do desempenho esportivo. *Revista Mackenzie de Educação Física e Esporte*, v. 11, n. 1, 2012. ISSN 1980-6892. Disponível em: <http://cev.org.br/biblioteca/ciencia-tecnologia-aplicada-melhoria-desempenho-esportivo>. Acesso em: 30 jun. 2017.

[4] CARLING, C. et al. The role of motion analysis in elite soccer. *Sports Medicine*, v. 38, n. 10, p. 839–862, 2008. ISSN 1179-2035. Disponível em: <http://dx.doi.org/10.2165/00007256-200838100-00004>. Acesso em: 30 jun. 2017.

[5] SIENKIEWICZ-DIANZENZA, E.; RUSIN, M.; STUPNICKI, R. Resistência anaeróbica de jogadores de futebol. *Fitness & performance journal*, Colégio Brasileiro de Atividade Física, Saúde e Esporte, n. 3, p. 199–203, 2009. ISSN 1676-5133. Disponível em: <https://dialnet.unirioja.es/descarga/articulo/2977271.pdf>. Acesso em: 30 jun. 2017.

[6] MEYER, T.; OHLENDORF, K.; KINDERMANN, W. Longitudinal analysis of endurance and sprint abilities in elite german soccer players. *Deutsche Zeitschrift für Sportmedizin*, v. 7, n. 8, p. 271–277, 2000. ISSN 2510-5264. Disponível em: <https://www.researchgate.net/publication/282684850_Longitudinal_analysis_of_endurance_and_sprint_abilities_in_elite_German_soccer_players>. Acesso em: 30 jun. 2017.

[7] BARROS, R. M. et al. Analysis of the distances covered by first division brazilian soccer players obtained with an automatic tracking method. *Journal of Sports Science and Medicine*, p. 233–242, 2007. ISSN 1303-2968. Disponível em: <http://hdl.handle.net/11449/69706>. Acesso em: 30 jun. 2017.

[8] SALVO, V. D. et al. Performance characteristics according to playing position in elite soccer. *International journal of sports medicine*, Georg Thieme Verlag KG Stuttgart, New York, NY, USA, v. 28, n. 3, p. 222–227, March 2007. ISSN 0172-4622. Disponível em: <https://doi.org/10.1055/s-2006-924294>. Acesso em: 30 jun. 2017.

[9] BARBERO-ÁLVAREZ, J. C. et al. The validity and reliability of a global positioning satellite system device to assess speed and repeated sprint ability (rsa) in athletes. *Journal of Science and Medicine in Sport*, Elsevier, v. 13, n. 2, p. 232–235, 2010. ISSN 1440-2440. Disponível em: <http://dx.doi.org/10.1016/j.jsams.2009.02.005>. Acesso em: 30 jun. 2017.

[10] EDGECOMB, S.; NORTON, K. Comparison of global positioning and computer-based tracking systems for measuring player movement distance during australian football. *Journal of Science and Medicine in Sport*, v. 9, n. 1, p. 25 – 32, 2006. ISSN 1440-2440. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1440244006000053>. Acesso em: 30 jun. 2017.

[11] PROZONE. *Find your Sports Data*. 2016. Disponível em: <http://prozonesports.stats.com>. Acesso em: 11 set. 2016.

[12] SCAGLIA, A. J. et al. Escolinha de futebol: uma questão pedagógica. *Motriz*, v. 2, n. 1, p. 36–43, 1996. ISSN 1980-6574. Disponível em: <http://www.periodicos.rc.biblioteca.unesp.br/index.php/motriz/article/view/6513>. Acesso em: 03 abr. 2017.

[13] CESSIE, S. L.; HOUWELINGEN, J. C. V. Ridge estimators in logistic regression. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, [Wiley, Royal Statistical Society], v. 41, n. 1, p. 191–201, 1992. ISSN 00359254, 14679876. Disponível em: <http://www.jstor.org/stable/2347628>. Acesso em: 30 jun. 2017.

[14] RAUBER, T. W. Redes neurais artificiais. 1997. Disponível em: <https://inf.ufes.br/~thomas/pubs/eri98.pdf>. Acesso em: 03 abr. 2017.

[15] BURGES, C. J. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, v. 2, n. 2, p. 121–167, 1998. ISSN 1573-756X. Disponível em: <http://dx.doi.org/10.1023/A:1009715923555>. Acesso em: 30 jun. 2017.

[16] DREISEITL, S.; OHNO-MACHADO, L. Logistic regression and artificial neural network classification models: a methodology review. *Journal of Biomedical Informatics*, v. 35, n. 5-6, p. 352 – 359, 2002. ISSN 1532-0464. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1532046403000340>. Acesso em: 30 jun. 2017.

[17] YEH, T. K. et al. Construction and uncertainty evaluation of a calibration system for gps receivers. *Metrologia*, v. 43, n. 5, p. 451, 2006. ISSN 1681-7575. Disponível em: <http://stacks.iop.org/0026-1394/43/i=5/a=017>. Acesso em: 30 jun. 2017.

[18] BATISTA, G. E. A. P. A.; PRATI, R. C.; MONARD, M. C. A study of the behavior of several methods for balancing machine learning training data. *SIGKDD Explor. Newsl.*, ACM, New York, NY, USA, v. 6, n. 1, p. 20–29, jun. 2004. ISSN 1931-0145. Disponível em: <http://doi.acm.org/10.1145/1007730.1007735>. Acesso em: 30 jun. 2017.

[19] HEARST, M. A. et al. Support vector machines. *IEEE Intelligent Systems and their Applications*, v. 13, n. 4, p. 18–28, July 1998. ISSN 1094-7167. Disponível em: <http://ieeexplore.ieee.org/abstract/document/708428/>. Acesso em: 30 jun. 2017.