Instituto de MATEMÁTICA E ESTATÍSTICA UFRGS

DEPARTAMENTO DE ESTATÍSTICA
E-mail: dest@mat.ufrgs.br

Trabalho de Conclusão de Curso

Previsão de resultados no vôlei de praia utilizando modelagem estatística

Lucas Santarossa Alvim

19 de fevereiro de 2024

Lucas Santarossa Alvim

Previsão de resultados no vôlei de praia utilizando modelagem estatística

Trabalho de Conclusão apresentado à comissão de Graduação do Departamento de Estatística da Universidade Federal do Rio Grande do Sul, como parte dos requisitos para obtenção do título de Bacharel em Estatística.

Orientador(a): Prof. Dr. Márcio Valk

Porto Alegre
Fevereiro de 2024

Lucas Santarossa Alvim

Previsão de resultados no vôlei de praia utilizando modelagem estatística

Este Trabalho foi julgado adequado para obtenção dos créditos da disciplina Trabalho de Conclusão de Curso em Estatística e aprovado em sua forma final pela Orientador(a) e pela Banca Examinadora.

Orientador(a):_____
Prof. Dr. Márcio Valk, UFRGS
Doutor(a) pela Universidade Federal do Rio Grande do Sul, Porto Alegre, RS

Banca Examinadora:

Prof. Dr. Danilo Marcondes Filho, UFRGS
Doutor pela Universidade Federal do Rio Grande do Sul – Porto Alegre, RS

Porto Alegre
Fevereiro de 2024

Resumo

Este trabalho visou à previsão de resultados em partidas de vôlei de praia. Foram utilizadas três metodologias: regressão logística, árvores de decisão e *K-Nearest Neighbors* (KNN). A base de dados utilizada contém dados de jogos masculinos e femininos, tanto do circuito norte-americano (AVP) quanto do circuito internacional (FIVB). Foram utilizadas variáveis referentes às estatísticas do jogo, informações de cada dupla, e a aspectos pessoais de cada jogador, como idade e altura. As variáveis referentes às estatísticas de cada jogo foram utilizadas no modelo por meio do cálculo de suas médias móveis simples das quatro partidas imediatamente anteriores à que está sendo prevista. Definiu-se que apenas os jogos da AVP seriam utilizados nas análises. A seleção de variáveis foi feita aplicando-se, para cada gênero, o método *backward stepwise* em duzentos conjuntos diferentes de dados de treino e teste, sendo que as variáveis que fossem eliminadas em mais de 50% destas repetições seriam removidas das análises. Os modelos com apenas as variáveis selecionadas se apresentaram vantajosos para ambos os gêneros devido à maior simplicidade de interpretação e precisão, medidas de ajuste e diagnóstico de resíduos semelhantes ao modelo completo. As médias de acurácia, sensibilidade e especificidade das previsões no modelo de regressão logística com variáveis selecionadas dos jogos masculinos foram de 0.721, 0.719 e 0.706, respectivamente. Para as partidas femininas, estes valores foram de 0.751, 0.742 e 0.755. O método KNN apresentou médias de acurácia, sensibilidade e especificidade de 0.705, 0.711 e 0.700 para os jogos masculinos e de 0.723, 0.701 e 0.745, respectivamente, para os femininos. Utilizando árvores de decisão, estes valores foram de 0.696, 0.697 e 0.695 para as partidas masculinas e de 0.731, 0.751 e 0.711 para as femininas. Ao avaliar os coeficientes da regressão logística para o último conjunto de dados de treino e de teste analisado, bem como a árvore de decisão gerada, ficou claro que as variáveis mais relevantes para a previsão dos resultados foram os rankings das duplas de referência e de oposição. Conclui-se que todos os métodos testados apresentam maior acurácia para as partidas femininas e que, com a metodologia utilizada, o modelo de regressão logística tende a apresentar desempenho preditivo levemente superior à dos outros métodos analisados.

Palavras-Chave: Vôlei de Praia, Regressão Logística, KNN, Árvore de Decisão.

Abstract

This work aimed to predict results in beach volleyball matches. Three methodologies were used: logistic regression, decision trees and K-Nearest Neighbors (KNN). The database used contains data from men's and women's games, both from the North American circuit (AVP) as well as from the international circuit (FIVB). Game statistics, information about each team and personal aspects of each player, such as age and height, were used as variables. Those regarding each game's statistics were used in the model by calculating their simple moving averages from the four matches preceding the one being predicted. It was decided that only AVP games would be used in the analyses. The variable selection was made by applying, for each gender, the backward stepwise method on two hundred different sets of training and testing data, and the variables that were eliminated in more than 50% of these repetitions would be removed from the analyses. Models with only selected variables proved to be advantageous for both genders due to greater simplicity of interpretation and precision, goodness of fit and residuals diagnostics similar to the complete model. The accuracy, sensitivity and specificity of predictions in the model with selected variables for the men's games were 0.721, 0.719 and 0.706, respectively. For the women's matches, these values were 0.751, 0.742 and 0.755, respectively. For both genders, such values were similar or slightly higher when compared to models with all variables, thus proving to be advantageous to use variable selection. The KNN method presented average accuracy, sensitivity and specificity of 0.705, 0.711 and 0.700 for men's games and 0.723, 0.701 and 0.745, respectively, for women's. Using decision trees, these values were 0.696, 0.697 and 0.695 for men's matches and 0.731, 0.751 and 0.711 for women's. When evaluating logistic regression coefficients for the last analyzed training and testing data set, as well as the decision tree generated, it became clear that the most relevant variables for predicting the results were the rankings of the reference and opposition teams. In conclusion, all tested methods showed greater accuracy for women's matches, and the logistic regression model tends to present slightly better predictive performance than the other methods analyzed.

Keywords: Beach Volleyball, Logistic Regression, KNN, Decision Tree.

Sumário

1 Introdução **12**

2 Revisão Bibliográfica **14**
2.1 Estatística no Esporte . . . 14
2.1.1 Estatística no Vôlei de Praia . . . 14
2.2 Aprendizagem Estatística . . . 15
2.2.1 Regressão Logística . . . 16
2.2.2 Árvores de Decisão . . . 17
2.2.3 KNN . . . 18
2.2.4 Medidas de Acurácia . . . 19
2.3 Métodos de Seleção de Variáveis . . . 19
2.3.1 Forward Stepwise . . . 20
2.3.2 Backward Stepwise . . . 20

3 Materiais e Métodos **22**
3.1 Base de Dados . . . 22
3.2 Limpeza e Manipulação da Base de Dados . . . 22
3.2.1 Variáveis de Interesse . . . 22
3.2.2 Dados Faltantes Iniciais e Ajuste de Variáveis Existentes . . . 24
3.2.3 Análise Exploratória . . . 24
3.2.4 Construção de Variáveis . . . 24
3.3 Seleção de Variáveis . . . 25
3.4 Ajuste e Diagnósticos dos Modelos . . . 26
3.5 KNN . . . 27
3.6 Árvores de Decisão . . . 27

4 Resultados e Discussão **28**
4.1 Análise Exploratória . . . 28
4.2 Seleção de Variáveis . . . 30
4.2.1 Masculino . . . 30
4.2.2 Feminino . . . 31
4.3 Correlação entre Variáveis . . . 33
4.3.1 Masculino . . . 33
4.3.2 Feminino . . . 33
4.4 Comparação de Modelos com e sem Seleção de Variáveis . . . 35
4.4.1 Medidas de Ajuste e Diagnósticos dos Modelos . . . 35
4.4.2 Precisão das Previsões . . . 36

4.4.3 Avaliação dos Coeficientes 38
4.5 KNN 39
4.6 Árvores de Decisão 40
4.7 Resumo dos resultados obtidos 42

5 Conclusões e Sugestões para Trabalhos Futuros **44**

Referências Bibliográficas **46**

Lista de Figuras

Figura 2.1: Exemplo do método KNN (James et al., 2013) 18
Figura 2.2: Algoritmo do método *forward stepwise* 20
Figura 2.3: Algoritmo do método *backward stepwise* 21

Figura 3.1: Algoritmo utilizado para a seleção de variáveis em cada gênero . . 26

Figura 4.1: Quantidade de jogos por ano, circuito e gênero 29
Figura 4.2: Percentual de jogadores por nacionalidade, dividido por gênero e circuito 29
Figura 4.3: Distribuição das variáveis de interesse 30
Figura 4.4: Correlações entre variáveis do modelo completo dos jogos masculinos 33
Figura 4.5: Correlações entre variáveis do modelo completo dos jogos femininos 34
Figura 4.6: Gráficos dos envelopes simulados para cada um dos modelos . . . 35
Figura 4.7: Boxplots das acurácias das previsões para cada uma das vinte repetições de cada modelo via regressão logsítica 37
Figura 4.8: Boxplots das acurácias das previsões para cada uma das vinte repetições de cada modelo KNN 40
Figura 4.9: Árvore de decisão para previsão de resultados das partidas das duplas masculinas 41
Figura 4.10: Árvore de decisão para previsão de resultados das partidas das duplas femininas 42

Lista de Tabelas

Tabela 3.1: Variáveis presentes no banco de dados 23

Tabela 4.1: Descrição dos dados analisados 28
Tabela 4.2: Variáveis removidas pelo método *backward stepwise* para os jogos do gênero masculino 31
Tabela 4.3: Variáveis removidas pelo método *backward stepwise* para os jogos do gênero feminino 32
Tabela 4.4: Média das médias do percentual de resíduos dentro dos limites do envelope simulado para cada modelo 35
Tabela 4.5: Médias obtidas para os valores de AIC e pseudo-R^2 de McFadden para cada um dos modelos testados 36
Tabela 4.6: Média da acurácia, sensibilidade e especificidade das previsões dos modelos de regressão logística estudados 37
Tabela 4.7: Coeficientes do modelo com seleção de variáveis das partidas masculinas 38
Tabela 4.8: Coeficientes do modelo com seleção de variáveis das partidas femininas 39
Tabela 4.9: Valor ótimo de K, média da acurácia, sensibilidade e especificidade utilizando o modelo KNN 39
Tabela 4.10: Média da acurácia, sensibilidade e especificidade utilizando o modelo de árvore de classificação 40
Tabela 4.11: Acurácia, sensibilidade e especificidade obtidas pelos diferentes métodos analisados 43

Lista de Variáveis

dupla_opp_avg_age	Média de idade da dupla de oposição
dupla_opp_avg_hgt	Média de altura da dupla de oposição
dupla_opp_home	Variável indicadora de se ao menos um jogador da dupla de oposição está jogando em seu país de origem
dupla_opp_nr_jogos	Número de partidas que a dupla de oposição já disputou junta
dupla_opp_rank	Ranking da dupla de oposição
dupla_opp_sma_avg_aces_points	Soma da média móvel simples do número de *aces* por ponto disputado de cada jogador da dupla de oposição
dupla_opp_sma_avg_blocks_points	Soma da média móvel simples do número de bloqueios por ponto disputado de cada jogador da dupla de oposição
dupla_opp_sma_avg_digs_points	Soma da média móvel simples do número de defesas por ponto disputado de cada jogador da dupla de oposição
dupla_opp_sma_avg_serve_errors_points	Soma da média móvel simples do número de erros de saque por ponto disputado de cada jogador da dupla de oposição
dupla_opp_sma_hitpct	Média móvel simples da eficiência de ataque da dupla de oposição
dupla_opp_streak	Sequência recente de resultados da dupla de oposição
dupla_opp_titulos	Número de títulos ganhos pela dupla de oposição

dupla_ref_avg_age	Média de idade da dupla de referência
dupla_ref_avg_hgt	Média de altura da dupla de referência
dupla_ref_home	Variável indicadora de se ao menos um jogador da dupla de referência está jogando em seu país de origem
dupla_ref_nr_jogos	Número de partidas que a dupla de referência já disputou junta
dupla_ref_rank	Ranking da dupla de referência
dupla_ref_sma_avg_aces_points	Soma da média móvel simples do número de *aces* por ponto disputado de cada jogador da dupla de referência
dupla_ref_sma_avg_blocks_points	Soma da média móvel simples do número de bloqueios por ponto disputado de cada jogador da dupla de referência
dupla_ref_sma_avg_digs_points	Soma da média móvel simples do número de defesas por ponto disputado de cada jogador da dupla de referência
dupla_ref_sma_avg_serve_errors_points	Soma da média móvel simples do número de erros de saque por ponto disputado de cada jogador da dupla de referência
dupla_ref_sma_hitpct	Média móvel simples da eficiência de ataque da dupla de referência
dupla_ref_streak	Sequência recente de resultados da dupla de referência
dupla_ref_titulos	Número de títulos ganhos pela dupla de referência

1 Introdução

A aplicação de métodos estatísticos para previsão de resultados de partidas esportivas é um objeto de estudo cada vez mais relevante no meio acadêmico. Diversos trabalhos relativos a este assunto já foram publicados sobre esportes como futebol, futebol americano, baseball e basquete, por exemplo, utilizando desde modelos mais simples, como a regressão logística, até modelos mais complexos, que utilizam, por exemplo, redes neurais (Horvat e Job, 2020). Além disso, boa parte dos clubes profissionais destes esportes possuem um departamento de análise de desempenho, que se utiliza destes modelos para auxiliar treinadores e gestores na tomada de decisões (Lewis, 2003; Eurosport, 2021). Há, no entanto, esportes em que a aplicação destes métodos foi pouco explorada em publicações científicas, destacando-se o vôlei de praia, uma modalidade que faz parte do programa das olimpíadas de verão a quase 30 anos e que vem se tornando cada vez mais popular em nível internacional nas últimas décadas, atraindo um público acumulado de 425000 pessoas nas olimpíadas de Londres em 2012 (Ioc, 2021). Para exemplificar o quanto o esporte está se tornando global, basta notar que das primeiras 20 edições do Circuito Mundial de Vôlei de Praia masculino, torneio disputado anualmente desde 1989, 14 foram vencidas por duplas brasileiras e 4 por duplas norte-americanas, enquanto das últimas 12 edições apenas 5 foram vencidas por duplas destes países, com destaque de duplas do Catar, Letônia e Noruega neste período. Com isso, visando um maior conhecimento sobre o esporte e a melhora do processo de decisões de técnicos e jogadores, se faz importante a identificação das variáveis relevantes para aumentar a probabilidade de vitória de cada dupla, bem como a construção de um modelo preditivo de resultados que atinja boa precisão em suas previsões.

A utilização de modelos estatísticos mais complexos inúmeras vezes é algo necessário para a previsão de resultados de partidas esportivas com maior acurácia se comparado a modelos mais simples. É importante, porém, que antes de se partir para a aplicação de métodos mais elaborados e computacionalmente intensivos, se conheça quais resultados os métodos mais simples apresentam, de forma que se tenha uma linha de base para comparações futuras e que se evite um maior custo computacional desnecessário nos casos em que os modelos mais básicos apresentam acurácia semelhante aos mais complexos. Assim, dada a escassez de análises preditivas sobre o vôlei de praia, é essencial o estudo da previsão de resultados utilizando um modelo mais comum e tido como simples, como é o caso da regressão logística, o qual é um modelo de classificação que pode ser utilizado para prever resultados das partidas e apresentar informações importantes sobre a relevância das variáveis por meio dos coeficientes de regressão. Desta forma, este trabalho buscará a iden-

tificação das variáveis mais relevantes para se prever a vitória em uma partida e a criação de um modelo de regressão logística visando à correta predição de resultados no vôlei de praia. Além disso, já visando a comparação com o modelo logístico, também serão feitas previsões pelo métodos de árvore de classificação e *K-Nearest Neighbors* (KNN).

2 Revisão Bibliográfica

2.1 Estatística no Esporte

A utilização da estatística no esporte vem aumentando nos últimos anos, tendo em vista o fato de que a quantidade de dados disponíveis relativos a esportes é cada vez maior. O baseball, por exemplo, é um esporte pioneiro na utilização de análises estatísticas, sendo que o uso dos dados é, atualmente, algo crítico e essencial nesta modalidade, o que pode ser demonstrado pelo fato de que todos os times na principal liga profissional de baseball dos Estados Unidos possuem seus próprios departamentos voltados para a produção destas análises e milhões de dólares são investidos neles (Costa et al., 2007).

Além do baseball, existe uma vasta quantidade de publicações referentes aos mais variados esportes. Nesse sentido, Eetvelde et al. (2021), por exemplo, concluiu que a utilização de métodos de machine learning é útil para a previsão de lesões e pode ser importante para a prevenção de futuras contusões, enquanto que, focando nos resultados dos jogos, Chen et al. (2021) e Huang e Lin (2020) utilizaram métodos de machine learning para prever os placares de partidas de basquete e Boshnakov et al. (2017) utilizaram um modelo de Weibull bivariado para prever placares de futebol. No que diz respeito à previsão de resultados de partidas, no entanto, na maior parte das vezes o interesse é definir quem será o vencedor e, sendo assim, destaca-se a utilização de métodos de classificação, nos quais é predita uma classe entre vitória, derrota, ou, no caso do futebol, por exemplo, empate (Horvat e Job, 2020).

2.1.1 Estatística no Vôlei de Praia

Na modalidade de vôlei de quadra, alguns trabalhos foram realizados visando à previsão dos resultados das partidas. Gabrio (2021), por exemplo, utilizou métodos bayesianos hierárquicos com esta finalidade, enquanto Tümer e Koçer (2017) utilizaram redes neurais artificiais para a previsão da colocação das equipes ao final de uma temporada da liga masculina da Turquia. No que diz respeito ao vôlei de praia, no entanto, a produção de trabalhos relacionados à análises estatísticas ainda é escassa se comparada a outros esportes, destacando-se os estudos de Kautz et al. (2017), que utilizou técnicas de *Deep Learning* para analisar padrões de movimentos de atletas durante uma partida visando à prevenção de contusões, e de Wenninger et al. (2020), que utilizou e comparou a acurácia de diferentes modelos de *Machine Learning* na previsão de comportamentos técnicos e táticos dos jogadores em uma partida. Quanto à identificação de fatores mais importantes para se determinar

quem vence uma partida, destacam-se os trabalhos de Peng e Cheng (2023), que, analisando 366 partidas masculinas e 367 femininas entre os anos de 2015 e 2022, usaram árvores de decisao e regressão logística para definir pontos no próprio serviço, erros de ataque e erros do oponente como as principais variáveis para determinar a vitória em uma partida e de Kumar et al. (2021), que utilizou regressão logística e analisou um total de 212 partidas masculinas e 214 partidas femininas entre 2017 e 2019 para identificar pontos em seu próprio serviço, eficiência de ataque, bloqueios, quantidade de erros do oponente, erros de recepção, erros de ataque e quantidade de ataque bloqueados como fatores mais significantes para a determinação da vitória. Não foi encontrada nenhuma bibliografia referente à predição de resultados a partir de variáveis existentes antes da partida ocorrer.

2.2 Aprendizagem Estatística

Aprendizagem estatística se refere a uma vasta gama de ferramentas para modelar e entender bases de dados complexas, sendo uma área da estatística recentemente desenvolvida e em constante crescimento, operando em conjunto com áreas da ciência da computação e, em particular, com aprendizado de máquina (*Machine Learning*). Tais ferramentas podem ser classificadas como supervisionadas ou não-supervisionadas, onde o primeiro caso se refere a modelos estatísticos construídos para estimar ou predizer resultados a partir dos dados de entrada, enquanto o segundo diz respeito a modelos onde não há resultados a serem analisados e deseja-se apenas, por exemplo, encontrar padrões, estruturas e relacionamentos existentes nos dados de entrada (James et al., 2013).

No que diz respeito às variáveis envolvidas nas análises que utilizam aprendizagem estatística, elas podem ser *quantitativas*, apresentando valores numéricos, ou *qualitativas*, também chamadas de *categóricas*, apresentando como valores diferentes categorias ou classes, como por exemplo, gênero e cor do cabelo das pessoas. Alguns métodos, como KNN e *boosting* podem ser utilizados para respostas com ambos os tipos de variáveis, enquanto regressão linear é utilizada exclusivamente para respostas quantitativas e regressão logística apenas para respostas qualitativas. Problemas envolvendo respostas quantitativas são chamados de problema de regressão, enquanto os demais são referidos como problemas de classificação. É importante notar, no entanto, que a regressão logística, apesar de lidar com respostas qualitativas, também pode ser pensada como um caso de regressão, pois estima as probabilidades de ocorrência das classes envolvidas (James et al., 2013).

Uma característica importante dos métodos de aprendizagem estatística é o fato de que, muitas vezes, é preferível utilizar um modelo mais restritivo ao invés de um mais flexível. Isso ocorre porque, quanto mais flexível e adaptável a casos mais complexos é um modelo, mais difícil de interpretar ele tende a ser. Por este motivo, muitas vezes se escolhe um modelo mais simples para se predizer ou inferir o que se deseja, pois, apesar de possivelmente um modelo mais complexo se adaptar melhor aos dados, é importante que se entenda o mais precisamente possível o que de fato este modelo está informando (James et al., 2013).

2.2.1 Regressão Logística

Em um modelo onde a resposta buscada se divide em duas categorias diferentes, o método de regressão logística modela a probabilidade desta resposta pertencer em uma destas duas categorias. Por exemplo, nos casos em que se busca prever qual time ganhará um jogo de vôlei, a regressão logística apresentará a probabilidade de vitória de tal time, dadas outras informações que se tenha sobre a partida.

Sendo Y igual a 1 se o time vencer e 0 caso contrário, X representando a matriz com as informações existentes sobre a partida, β_0 o intercepto e β o vetor de coeficientes lineares, tem-se que o modelo de regressão logística pode ser expresso da seguinte forma:

$$\log\left(\frac{P(Y=1|X)}{1-P(Y=1|X)}\right) = \beta_0 + \beta X, \tag{2.1}$$

onde $P(Y=1|X)$ deve apresentar resultados entre 0 e 1 e é representado pela função logística:

$$P(Y=1|X) = \frac{e^{\beta_0+\beta X}}{1+e^{\beta_0+\beta X}}. \tag{2.2}$$

A expressão $\log\left(\frac{P(Y=1|X)}{1-P(Y=1|X)}\right)$ é chamada de *log-odds* ou *logit*. Desta forma, para cada elemento β_i em β, mantendo-se os demais coeficientes constantes, o aumento de uma unidade de X altera a log-odds por um fator β_i e sabe-se que se β_i for positivo, o aumento de X implica em aumento de $P(Y=1|X)$ (James et al., 2013).

As estimativas de β_0 e β são dadas pelo método de máxima verossimilhança, e são tais que minimizam a função:

$$L = -log\left(\prod_{i:y_i=1}(p(x_i|y_i))\prod_{i':y'_i=0}(1-p(x'_i|y'_i))\right). \tag{2.3}$$

O modelo de regressão logística apresenta alguns pressupostos para que ele possa ser utilizado, sendo eles: a linearidade entre a log-odds e as variáveis independentes; as observações são independentes uma da outra; ausência de multicolinearidade; ausência de *outliers*.

Diagnóstico dos Resíduos

Envelopes Simulados

O método de envelopes simulados é um algoritmo utilizado para verificar se os resíduos do modelo proposto atendem ao pressuposto de distribuição normal. Tal método consiste das seguintes etapas: primeiramente, o modelo é simulado N vezes; em seguida, o modelo é ajustado N vezes utilizando cada uma das N simulações realizadas, e são coletados os resíduos para cada um destes ajustes em cada um dos n pontos analisados; por fim, cada um dos N resíduos de cada um dos n pontos da amostra são organizados em ordem crescente e são calculados os limites superior e inferior do envelope simulado pelos quantis $\frac{(1+conf)}{2}$ e $\frac{(1-conf)}{2}$, respectivamente, sendo $conf$ o nível de confiança estabelecido pelo pesquisador (Atkinson, 1987; Everitt, 1994).

Medidas de Ajuste e Seleção de Modelos

AIC

O critério de informação de Akaike (AIC) é um método utilizado para seleção de modelos, o qual é representado pela equação 2.4, onde $\hat{L}$ é o valor máximo da função de verossimilhança e k é o número de parâmetros estimados no modelo. Logo, o AIC apresenta um *trade-off* entre melhor ajuste e maior complexidade do modelo, sendo que a penalização por um modelo ter mais variáveis evita o *overfitting*. O modelo a ser selecionado é o que apresentar menor valor de AIC.

$$AIC = 2k - 2ln(\hat{L}). \tag{2.4}$$

Pseudo-R² de McFadden

O pseudo R^2 de McFadden é uma medida de ajuste de modelos representada pela equação 2.5, onde L_M é a verossimilhança do modelo sendo ajustado e L_0 a verossimilhança do modelo nulo. Este método foi proposto por McFadden (1972) e, segundo Hensher e Stopher (1979), valores entre 0.2 e 0.4 representam um excelente ajuste.

$$R^2_{McF} = 1 - \frac{ln(L_M)}{ln(L_0)}. \tag{2.5}$$

2.2.2 Árvores de Decisão

Árvores de decisão são algoritmos que visam segmentar as variáveis preditoras em diversas estratificações binárias, de forma a se alcançar o resultado previsto a partir destas subdivisões. Cada ponto onde uma destas subdivisões é realizada é chamado de *nodo*, sendo os nodos finais da árvore chamados de *"nodos terminais"* ou *"folhas"*. Estes métodos se destacam por sua fácil interpretação, mas são inferiores a outros algoritmos mais avançados, como *florestas aleatórias* e *bagging*, em termos de acurácia das previsões. As árvores de decisão podem ser utilizadas tanto para respostas quantitativas quanto qualitativas, sendo chamadas de "árvores de regressão" no primeiro caso e de "árvores de classificação" no segundo.

Árvores de Classificação

Para se definir as regras das segmentações das árvores de decisão, deve-se utilizar um critério que minimize o erro de previsão. No caso das árvores de classificação, um critério comumente utilizado é o índice de Gini, o qual é calculado de acordo com a equação 2.6:

$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}), \tag{2.6}$$

onde $\hat{p}_{mk}$ representa a proporção de observações na m-ésima região que são pertencentes à classe k, e sendo que K é igual a 2 quando a resposta é binária.

As regiões são definidas da seguinte maneira: primeiramente, tendo todas as observações aglomeradas em uma mesma região, o algoritmo busca, para todas as combinações de variáveis preditoras e seus respectivos possiveis valores de corte, a combinação cujo índice de Gini for menor. Tal combinação é, então, utilizada

como critério de segmentação da árvore e, posteriormente, os mesmos passos são realizados para cada um dos novos nodos criados, até que determinado critério de parada seja alcançado. Os valores de cada nodo da árvore, no caso de respostas binárias, representam as classes cuja proporção de ocorrência na respectiva região analisada é maior do que 50%.

A árvore formada pelo processo descrito, apesar de apresentar maior qualidade de previsão dos dados de treinos, pode ser prejudicada por *overfitting* e por uma maior complexidade. Assim, visando melhorar o desempenho do algoritmo, podem ser utiizados métodos que, partindo da árvore inicial completa, a *"podam"* para obter uma *subárvore* menor, de mais fácil interpretabilidade e sem problemas de *overfitting* (James et al., 2013).

2.2.3 KNN

O KNN (*K-nearest neighbors*), é um método que, dado um valor inteiro K e uma observação de teste x_0, identifica os K pontos nos dados de treino que são mais próximos de x_0 e então estima a probabilidade condicional de Y pertencer à classe j de acordo com a seguinte equação:

$$P(Y = j|X = x_0) = \frac{1}{K} \sum_{i \in \mathcal{N}_0} (I(y_i = j)) \tag{2.7}$$

onde $\mathcal{N}_0$ se refere aos K pontos mais próximos de x_0. O método pode ser utilizado tanto para o caso de respostas quantitativas quanto qualitativas e o valor ótimo de K pode ser obtido por validação cruzada. Depois da estimação pela equação 2.7, o modelo classifica o vetor x_0 como pertencendo à classe com maior probabilidade (James et al., 2013). A figura 2.1 mostra um exemplo do funcionamento do algoritmo com $K = 3$ e duas classes distintas, azul e laranja, sendo que a imagem à direita representa a observação de teste x à qual se deseja atribuir uma das classes e o círculo inclui as três observações mais próximas dela. A imagem à esquerda, por sua vez, representa os limites de decisão do algoritmo, e caso a observação de teste estiver na zona laranja ela será classificada como tal, caso contrário será classificada como *azul*.

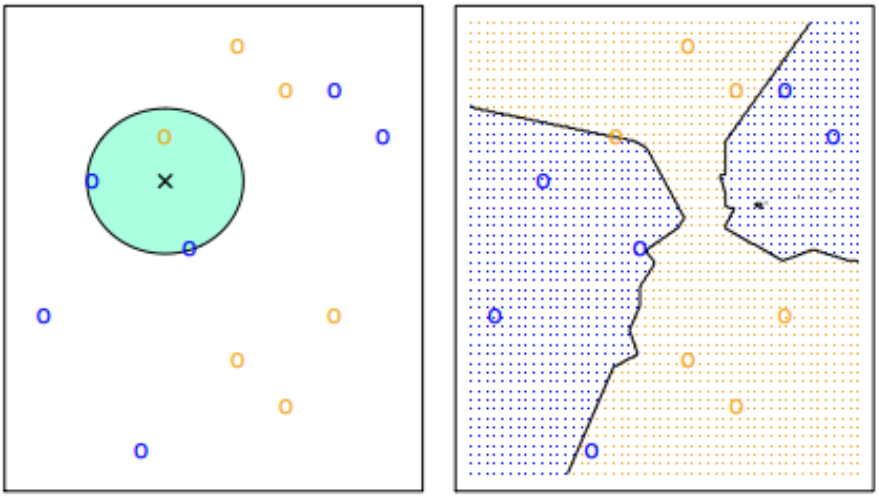

Figura 2.1: Exemplo do método KNN (James et al., 2013)

2.2.4 Medidas de Acurácia

Para os casos em que a resposta desejada é quantitativa, a medida mais comumente utilizada é o erro quadrático médio, ou EQM, sendo que o modelo com menor EQM deve ser o escolhido. Tendo um par de dados de teste (X, Y), formado por n observações e sendo $\hat{f}$ a funçao de predição calculada, o EQM é calculado da seguinte forma:

$$EQM = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2. \tag{2.8}$$

Para casos de classificação, em que a resposta é qualitativa, a medida utilizada para se medir a acurácia do modelo é a taxa de erro de teste (TET). Dado um par de dados de teste (X, Y), formado por n observações e sendo $\hat{Y}$ a classe prevista pelo modelo, a TET é calculada da seguinte forma:

$$\frac{1}{n}\sum_{i=1}^{n}(I(y_i \neq \hat{y}_i)) \tag{2.9}$$

onde I é uma função indicadora que assume valor 1 caso $y_i \neq \hat{y}_i$ e 0 caso contrário. A equação 2.9 é minimizada, em média, pelo classificador de Bayes, o qual verifica a probabilidade de se obter uma classe j dado um vetor de variáveis independentes de teste x_0 e, se a seguinte condição for satisfeita

$$P(Y = j|X = x_0) > 0.5 \tag{2.10}$$

o modelo seleciona a classe j como a previsão correta para o vetor x_0 e outra classe, caso contrário. Tal classificador, no entanto, é impossível de ser obtido para dados reais, pois não se sabe a distribuição condicional de Y dado X. Assim, ele é tido como um padrão ótimo que outros métodos, como o KNN, tentam alcançar.

Além da acurácia geral dos modelos de classificação, pode-se mensurar também a probabilidade deste modelos preverem corretamente cada classe estudada. No caso binário, chama-se uma classe de "negativa" e a outra de "positiva" e calcula-se a especificidade (Equação 2.11) e a sensibilidade (Equação 2.12), que representam, respectivamente, o percentual de "negativos" e o percentual de "positivos" que foram classificados corretamente.

$$Especificidade = \frac{\text{Verdadeiros Negativos}}{\text{Verdadeiros Negativos} + \text{Falsos Positivos}} \tag{2.11}$$

$$Sensibilidade = \frac{\text{Verdadeiros Positivos}}{\text{Verdadeiros Positivos} + \text{Falsos Negativos}} \tag{2.12}$$

2.3 Métodos de Seleção de Variáveis

Muitas vezes, boa parte das variáveis utilizadas nos modelos são irrelevantes ou então redundantes e não estão de fato associadas à resposta estudada. A inclusão de tais variáveis pode levar a uma maior complexidade desnecessária do modelo e a um custo computacional mais elevado. Assim, é importante que se utilize métodos que auxiliem na seleção das variáveis que serão utilizadas no modelo, de forma que os problemas citados sejam evitados. Existem diversos métodos utilizados para este fim, valendo destacar os métodos *forward stepwise* e *backward stepwise*.

2.3.1 Forward Stepwise

O método *forward stepwise* consiste na implementação do algoritmo descrito na figura 2.2. Primeiramente, deve-se estimar o modelo nulo, e, em seguida, deve-se estimar o modelo com uma variável, testando todas as variáveis disponíveis. Após, o modelo com maior ganho de ajuste, de acordo com algum método de escolha, como o AIC, por exemplo, deve ser selecionado. No terceiro passo, o modelo selecionado anteriormente deve ser estimado novamente, adicionando-se uma variável e testando o modelo com todas as variáveis ainda disponíveis. Por fim, deve-se repetir o passo anterior sucessivamente até que nenhuma variável disponível fora do modelo apresente ganho relevante de ajuste.

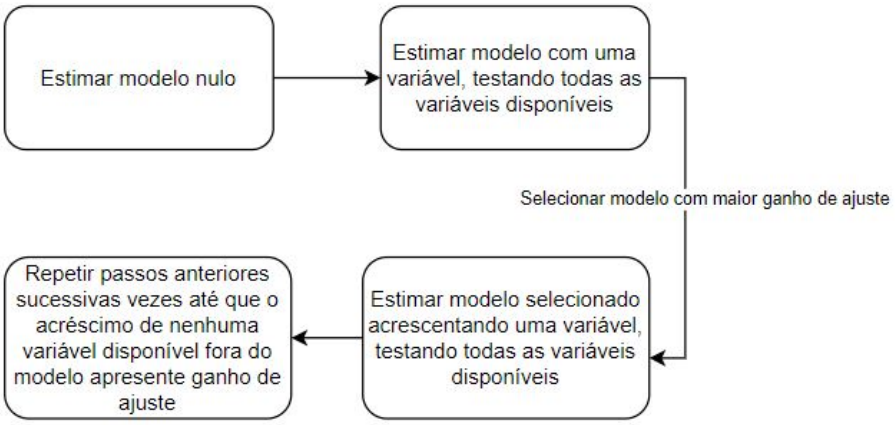

Figura 2.2: Algoritmo do método *forward stepwise*

2.3.2 Backward Stepwise

O método *backward stepwise* é semelhante ao *forward stepwise*, mas ao invés de se adicionar variáveis sucessivamente, elas devem ser removidas. O algoritmo está descrito na figura 2.3. Primeiramente, deve-se estimar o modelo completo, com todas as variáveis. Posteriormente, deve-se estimar o modelo removendo uma variável por vez, selecionando-se o modelo sem a variável que representa menor ganho de ajuste, de acordo com algum criério de escolha, como o AIC, por exemplo. No terceiro passo, o modelo selecionado anteriormente deve ser estimado novamente, removendo-se uma variável por vez e testando o modelo com cada uma destas remoções. Por fim, deve-se repetir os passos anteriores sucessivamente até que a exclusão de qualquer variável resulte em piora de ajuste.

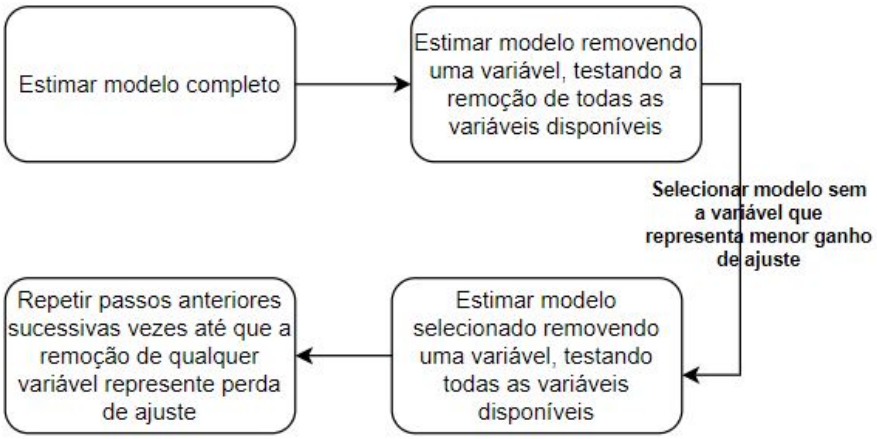

Figura 2.3: Algoritmo do método *backward stepwise*

3 Materiais e Métodos

3.1 Base de Dados

A base de dados utilizada vem do site https://github.com/bigtimestats/beach-volleyball, tendo sido publicada pelo usuário BigTimeStats no github. A base contém dados de mais de 65000 jogos, masculinos e femininos, tanto do circuito norte-americano (Association of Volleyball Professionals - AVP) quanto do circuito internacional (Fédération Internationale de Voleyball - FIVB) de vôlei de praia. As partidas presentes na base compreendem os anos de 2000 a 2022 e as variáveis apresentadas são citadas na Tabela 3.1, a qual também apresenta observações referentes a algumas destas variáveis.

3.2 Limpeza e Manipulação da Base de Dados

Todos os ajustes e manipulações realizados foram feitos utilizando a linguagem de programação R no software RStudio. O principal objetivo desta etapa foi lidar com dados faltantes, ajustar variáveis existentes e construir as variáveis necessárias para a aplicação do modelo de regressão logística visando à previsão de resultados.

3.2.1 Variáveis de Interesse

O processo de limpeza e manipulação de dados consistiu de, primeiramente, escolher as variáveis de interesse que seriam utilizadas nas análises e depois realizar os ajustes necessários. Tais variáveis, as quais são contabilizadas para cada jogador da partida, são as seguintes: *total de aces, total de ataques, total de pontos de ataque, total de bloqueios, total de erros, total de defesas, eficiência de ataque, erros de saque, idade, país e altura.* Além disso, também foi considerada a variável referente ao ranking das duplas no início do torneio.

As variáveis referentes às estatísticas de cada jogo foram utilizadas nos modelos por meio do cálculo de suas médias móveis simples, sendo que foram testadas diferentes janelas de jogos. Por fim, foi definido que, devido ao maior número de jogos disponíveis para análise após a limpeza da base de dados e às precisões semelhantes ao final das análises, as partidas a serem utilizadas no cálculo seriam as quatro imediatamente anteriores à que está sendo prevista.

Tabela 3.1: Variáveis presentes no banco de dados

Dado	**Observação**
Circuito	Circuito referente à partida disputada - FIVB ou AVP
Torneio	-
País de disputa	-
Data da partida	Data da primeira partida do respectivo torneio
Número da partida	Número identificador da partida dentro do respectivo torneio
Gênero dos jogadores	-
Nome dos jogadores	-
Data de nascimento e idade dos jogadores	-
Altura dos jogadores	-
Nacionalidade dos jogadores	-
Ranking das duplas no início do torneio	-
Placar dos sets da partida	-
Duração da partida	-
Grupo e rodada da competição	-
Total de ataques de cada jogador	-
Total de erros de ataque de cada jogador	-
Total de pontos de ataque de cada jogador	-
Eficiência de ataque de cada jogador	Divisão entre a diferença de pontos e erros de ataque sobre o total de ataques
Total de pontos de saque de cada jogador	-
Total de erros de saque de cada jogador	-
Total de pontos de bloqueio de cada jogador	-
Total de defesas de cada jogador	-

3.2.2 Dados Faltantes Iniciais e Ajuste de Variáveis Existentes

Foram removidas partidas que incluíam dados faltantes em ao menos uma das variáveis de interesse. Além disso, como foram utilizadas as médias móveis dos quatro jogos imediatamente anteriores de determinadas variáveis de interesse, qualquer partida em que ao menos um destes quatro jogos apresentasse algum dado faltante também foi posteriormente removida das análises.

Além do cuidado com os dados faltantes, também foi feita a remoção de algumas partidas cujas variáveis de interesse não se apresentaram da maneira desejada para a análise. Por exemplo, nos torneios da AVP há, muitas vezes, a *chave dos vencedores* e a *chave qualificatória*, sendo que nesta última o ranking na base de dados se refere somente ao número do jogo e não à posição da dupla de fato. Logo, os jogos de chave qualificatória foram removidos da análise. Além disso, esta mesma variável não representa a posição da dupla também nos casos em que as partidas eram em torneios do tipo *"Rei da Praia"*, onde era apresentado o ranking individual do jogador da dupla cujo nome era o segundo em ordem alfabética e, sendo assim, partidas deste tipo de torneio também foram removidas. Por fim, jogos em que a dupla vencedora foi definida por desistência ou qualquer outro motivo que não seja de fato a vitória em quadra também foram retirados da análise. Após a realização dos procedimentos citados, a quantidade de jogos restantes para análises foi de 8245 partidas da AVP e 168 da FIVB, totalizando 8413 jogos.

3.2.3 Análise Exploratória

Após o ajuste dos dados faltantes e das variáveis existentes na base, foi feita uma breve análise exploratória da mesma. Para isto, foi descrita a quantidade de partidas, altura e idade média dos jogadores, quantidade de torneios disputados e período abrangido pela base. Além disto, foi verificada a distribuição da nacionalidade dos jogadores e da quantidade de partidas por ano, sendo que todas estas informações foram subdivididas por circuito e gênero.

3.2.4 Construção de Variáveis

As variáveis de interesse que envolviam dados das partidas analisadas, como aces e bloqueios, por exemplo, foram todas ponderadas pelo total de pontos disputados na partida, de forma que não ocorresse distorção para partidas mais equilibradas, em que o número de pontos disputados é relativamente grande. Desta forma, foi primeiramente necessária a construção de uma variável informando o número de pontos jogados nas partidas. Posteriormente, visando à previsão de resultados, foi determinada a dupla de referência para a qual seriam feitas as previsões. Para isto, foi criada uma variável a qual, após o ordenamento da base de dados já sem dados faltantes por circuito, torneio, país, data, gênero e número da partida, foi definida como sendo a dupla vencedora caso o número da linha da base de dados fosse par, e como a dupla perdedora caso contrário.

Depois disto, foi feita a construção das variáveis para se trabalhar com médias móveis simples por jogador, sendo que, como a janela de jogos utilizadas foram as quatro partidas imediatamente anteriores àquela para a qual foi realizada a previsão, os jogos de jogadores com menos de 5 partidas na base de dados foram descartados da

análise. Após esta exclusão, primeiramente foram construídas as variáveis referentes aos aces, número de bloqueios, defesas e erros de saque por total de pontos disputados nas partidas. Depois foi calculada a média móvel simples dessas variáveis na janela descrita anteriormente, além da média móvel simples do total, sem dividir pelo número de pontos disputados, de ataques, erros de ataque e pontos de ataque na mesma janela.

Além dos dados por jogador, também foram elaboradas variáveis relativas às duplas, as quais são: a sequência recente de resultados da dupla, o número de títulos já ganhos e o número de jogos disputados pela dupla até o respectivo jogo analisado. A variável referente à sequência recente de resultados foi montada da seguinte forma: para o primeiro jogo da dupla, a variável é igual a zero; se a dupla vem de uma sequência de 1 vitória, ela é igual a 1; se a sequência é de 2 vitórias, ela é igual a 2, e assim por diante; se, por outro lado, a dupla vem de uma sequência de uma derrota, a variável retorna -1; se a sequência é de duas derrotas, é igual a -2, e assim por diante. Além disso, foi construída uma variável análoga ao "mando de camp" em outros esportes: se ao menos um dos jogadores da dupla forem do país onde o jogo está ocorrendo, considera-se que a dupla joga "em casa", caso contrário, considera-se que ela joga "fora de casa".

Finalmente, as variáveis relativas a cada jogador foram aglomeradas para se obter um valor único para a dupla. Para idade e altura, foi feita a média aritmética simples dos valores de cada jogador, enquanto que para aces, bloqueios, defesas e erros de saque foram somados os valores de médias móveis simples calculados por jogador. A eficiência de ataque, por sua vez, foi calculada da seguinte forma:

$$hitpct_{sma_i} = \frac{\sum_{j=1}^{2}(kills_{sma_{ij}}) - \sum_{j=1}^{2}(errors_{sma_{ij}})}{\sum_{j=1}^{2}(attacks_{sma_{ij}})}, \tag{3.1}$$

onde $hitpct_{sma_i}$ representa a variável de média móvel da eficiência de ataque da dupla i, $kills_{sma_{ij}}$ representa média móvel do total de pontos de ataque do jogador j da dupla i, $errors_{sma_{ij}}$ a média móvel do total de erros e $attacks_{sma_{ij}}$ é a média móvel do total de ataques tentados por este mesmo jogador. Por fim, as variáveis montadas foram atribuídas à dupla de referência ou á dupla de oposição, de acordo com a variável elaborada para este fim.

3.3 Seleção de Variáveis

Para a seleção de variáveis, primeiramente os dados foram divididos em jogos masculinos e femininos. Depois, cada uma dessas partições foi separada em dados de treino e teste, com 80% das partidas indo para o primeiro grupo. Após tal separação, foi montado o modelo de regressão logística completo, com todas as variáveis, através da função *glm* do pacote base do *RStudio*, utilizando o argumento *family* igual a *"binomial"*, sendo que foi aplicado postariormente a tal modelo o método *backward stepwise* para verificação das variáveis mais relevantes, utilizando a função *stepAIC* do pacote *MASS* com argumento *direction* igual a *"backward"*. Foi feita, então, a compilação das variáveis que constavam no modelo completo mas não no modelo com a aplicação do método *backward stepwise*.

O algoritmo descrito no parágrafo anterior foi repetido duzentas vezes tanto para as partidas masculinas quanto para as femininas, de modo que o método de seleção de variáveis fosse aplicado a diversas combinações diferentes de dados de treino. As variáveis que não foram descartadas em mais de 50% das repetições de cada gênero foram utilizadas em um modelo separado no próximo passo, sendo que os procedimentos citados são resumidos pela figura 3.1. Além disso, foi calculado também, por meio da função *cor* do pacote *stats*, o coeficiente de correlação de Pearson para todas as variáveis e também apenas para as variáveis selecionadas.

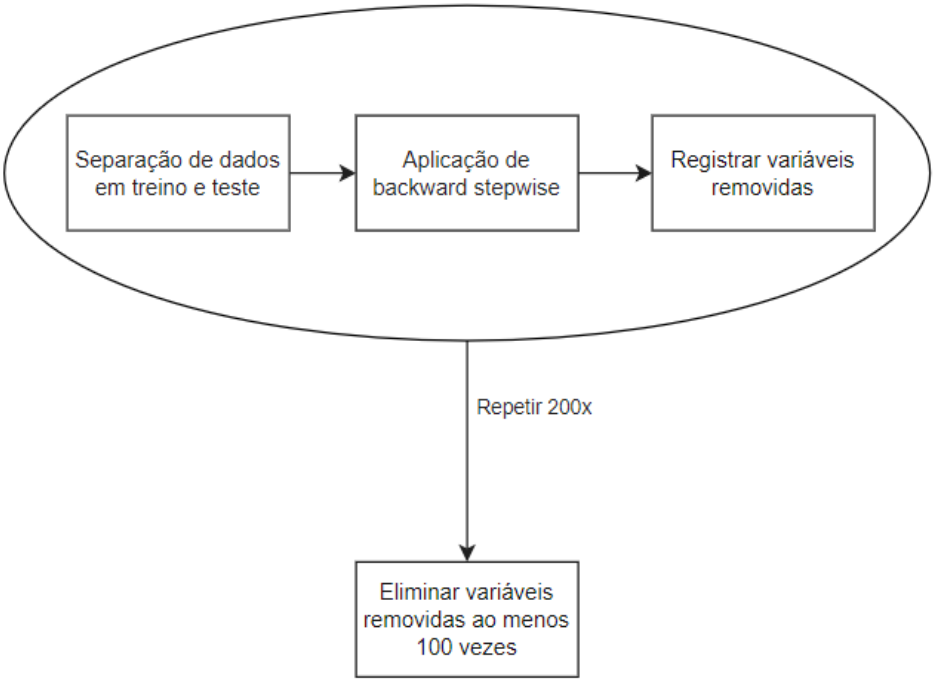

Figura 3.1: Algoritmo utilizado para a seleção de variáveis em cada gênero

3.4 Ajuste e Diagnósticos dos Modelos

Os dados dos jogos masculinos e femininos foram separados, da mesma forma que no item anterior, em dados de treino e de teste. Posteriormente, foi aplicado novamente o modelo de regressão logística utilizando-se todas as variáveis e o mesmo modelo utilizando apenas as variáveis que não foram descartadas em mais de 50% das repetições do passo anterior. Este procedimento foi repetido vinte vezes para os dados de cada gênero, e em cada uma destas repetições foram coletados o Pseudo R^2 de McFadden, o AIC e o percentual de dados dentro dos limites superior e inferior dos envelopes simulados, sendo que tais simulações foram realizadas dez vezes em cada uma das vinte repetições. As funções utilizadas para obter tais dados foram, respectivamente: função *pR2* do pacote *pscl*, função *glm* do pacote base do *RStudio* e função *envelope*, com argumento *rep* igual a vinte e cinco, do pacote *glmtoolbox*. Além destas informações, também foram coletadas as precisões das previsões de cada uma das vinte repetições utilizando a equação 2.9. Por fim, foram avaliados, para a última das vinte reptições realizadas para cada gênero, os coeficientes dos modelos com a aplicação da seleção de variáveis.

3.5 KNN

Depois de coletadas as medidas descritas anteriormente, foi feita a comparação entre o modelo completo e os modelos com seleção de variáveis, para verificar, a partir destas medidas, qual a melhor opção de modelo para utilizar, tanto por motivos de ajuste quanto de acurácia de previsões. Por fim, já utilizando as variáveis do modelo escolhido, foram feitas previsões por meio do método KNN. Para tal, foi utilizado o pacote *caret* e primeiramente foi feito o ajuste da escala das variáveis para valores semelhantes, por meio da função *preProcess*, utilizando os valores *center* e *scale* no argumento *method*. Após, foi encontrado o valor ótimo de K pela função *train*, utilizando o valor *knn* no argumento *method* e validação cruzada como método de controle de treino, permitindo o K variar entre os valores 3, 5, 7, 10, 13, 16, 18 e 20. Por fim, empregando o valor ótimo de K, foi utlizada a função *knn3* para realizar a classificação, para posteriormente se obter a acurácia, sensibilidade e especificidade do modelo. O algoritmo descrito foi repetido por vinte vezes para os dados de cada gênero, assim como foi feito para a regressão logística.

3.6 Árvores de Decisão

As previsões realizadas por meio do método de árvores de decisão foram feitas tanto para obter uma estimativa diferente da precisão quanto para se obter melhor visualização de quais as variáveis mais relevantes para se prever corretamente o resultado de uma partida de vôlei de praia. Tal método foi empregado por meio das funções *decision_tree* e *metrics* do pacote *tidymodels*, bem como a função *rpart.plot* do pacote homônimo. O número mínimo de observações em cada nodo para que o algoritmo continue tentando segmentar a árvore foi definido como sendo igual a vinte, o número mínimo de observações em cada nodo terminal foi definido como sendo igual a seis e o parâmetro de complexidade *cp* utilizado foi 0.01. Este algoritmo, assim como no modelo de regressão logística e no KNN, foi repetido por vinte vezes.

4 Resultados e Discussão

4.1 Análise Exploratória

A quantidade de partidas, média de idade e de altura por partida, bem como a quantidade de torneios disputados e o período abrangido, subdividida por gênero e circuito, após a manipulação e limpeza da base de dados, são mostrados na tabela 4.1. Nota-se que o número de partidas femininas e masculinas, a quantidade de torneios disputados e o período abrangido são semelhantes para cada circuito, enquanto a idade é maior nos jogos da AVP, sendo que os homens são, em média, mais velhos e mais altos que as mulheres. Além disso, é notável a quantidade superior de partidas da AVP em relação à FIVB.

Tabela 4.1: Descrição dos dados analisados

	Gênero	Qtd	Idade Media	Altura Média (cm)	Torneios	Período
AVP	Masculino	4079	31.30	194.65	153	2004-2022
	Feminino	4166	30.78	180.89	150	2004-2022
FIVB	Masculino	80	29.73	195.23	10	2004-2021
	Feminino	88	28.54	181.32	10	2004-2021

A distribuição de jogos por ano, gênero e circuito é apresentada na figura 4.1. No caso do circuito da AVP, percebe-se um pico no número de jogos considerados para análise no ano de 2007, para ambos os gêneros, sendo que a maioria dos jogos concentra-se no período de 2004 a 2010. O circuito da FIVB, por sua vez, apresenta uma quantidade muito menor de jogos, os quais estão dispersos mais igualmente do que as partidas da AVP, e o pico ocorre no ano de 2011 para ambos os gêneros.

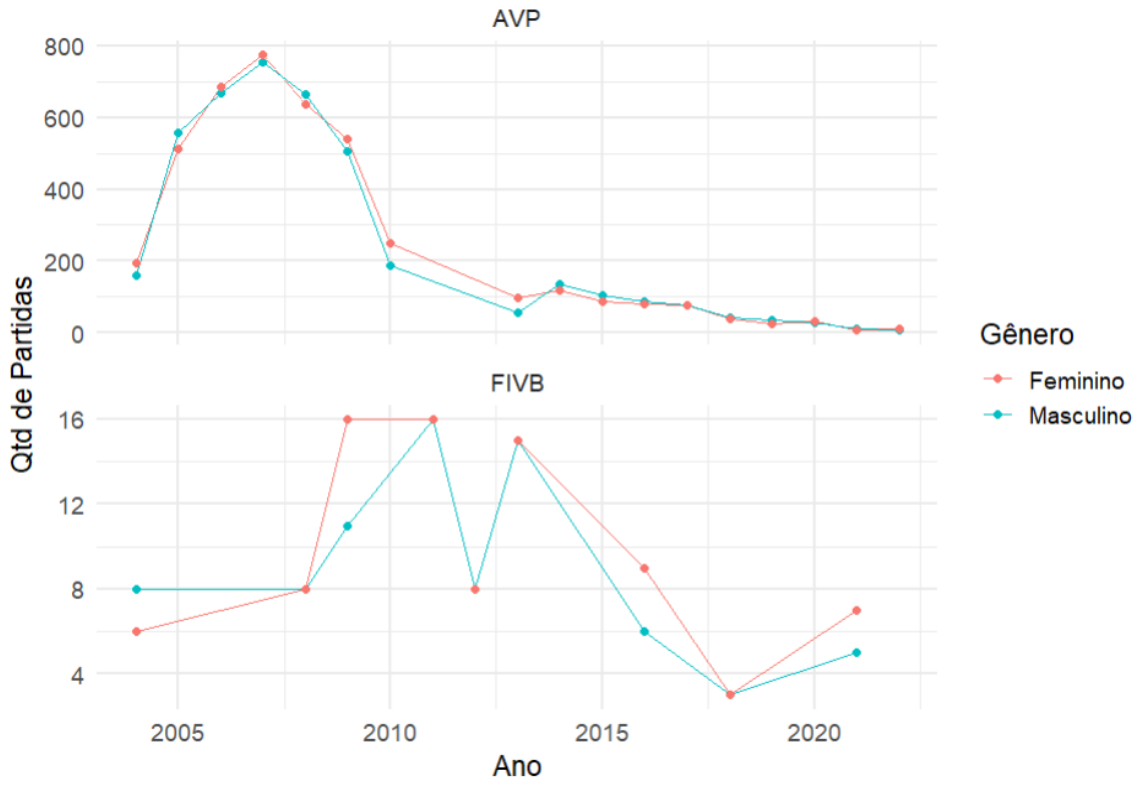

Figura 4.1: Quantidade de jogos por ano, circuito e gênero

A figura 4.2 mostra a distribuição da nacionalidade dos jogadores para cada circuito e gênero. Pode-se observar que a grande maioria dos jogadores da AVP são dos Estados Unidos, o que é esperado, dado que tal circuito é deste país. Na FIVB, por sua vez, a distribuiçao é mais equilibrada entre vários países, com destaque para a Alemanha em ambos os gêneros.

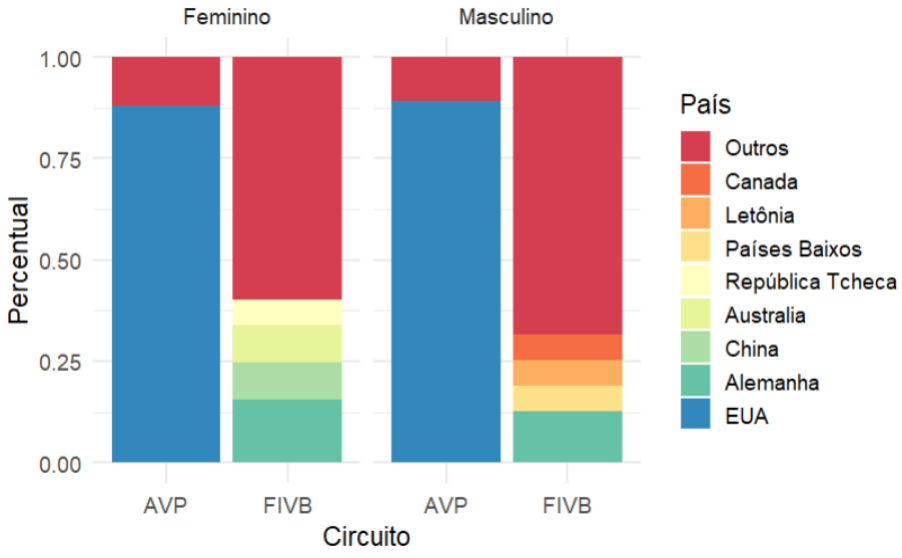

Figura 4.2: Percentual de jogadores por nacionalidade, dividido por gênero e circuito

Perebe-se pela figura 4.1 e pela tabela 4.1 que a quantidade de partidas da FIVB

é muito inferior à da AVP. Além disso, o número de jogos do circuito internacional é escasso para a subdivisão dos dados em dados de treino e de teste e a posterior análise preditiva. Sendo assim, as análises realizadas foram feitas somente para os jogos da AVP.

A distribuição das variáveis de interesse, já separadas por dupla de referência e dupla de oposição, além de discernidas por gêneros, são apresentadas na figura 4.3. Nota-se que as variáveis numéricas contínuas apresentam distribuição aproximadamente normal.

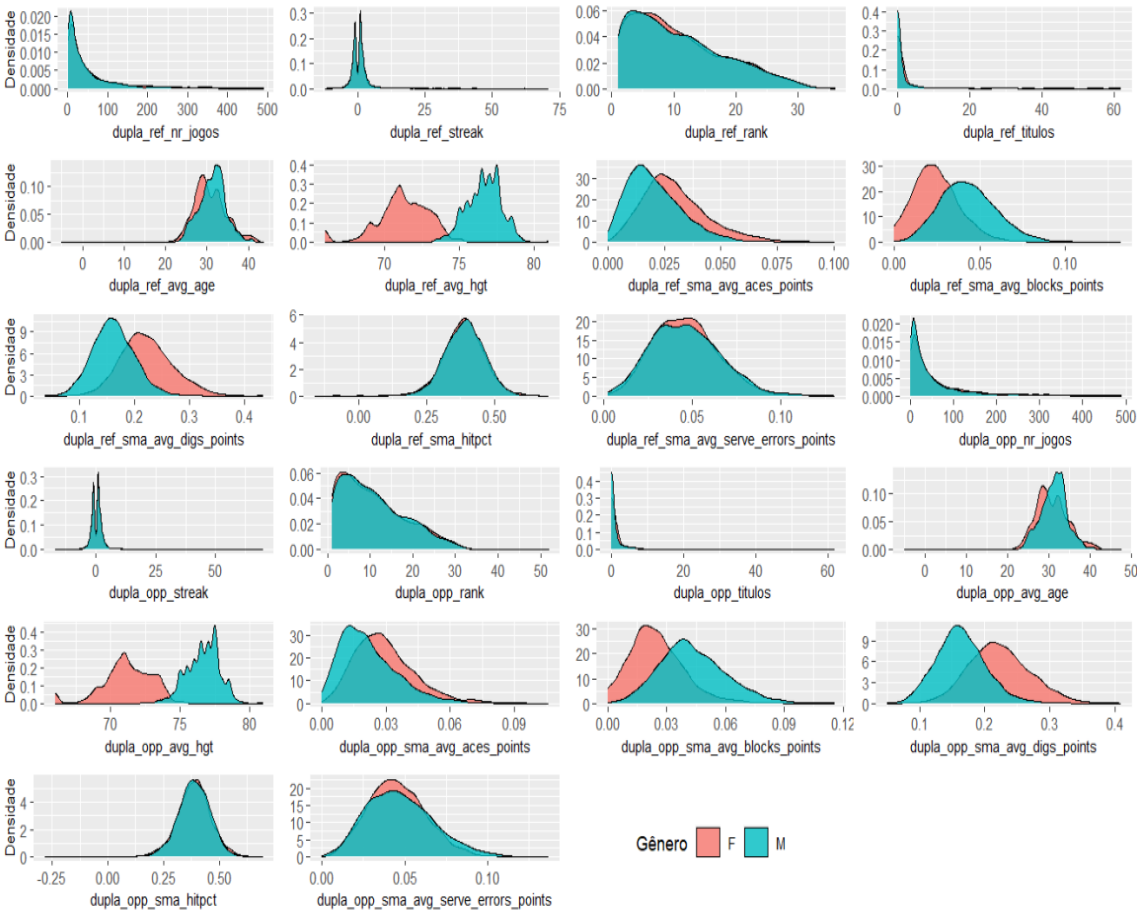

Figura 4.3: Distribuição das variáveis de interesse

4.2 Seleção de Variáveis

4.2.1 Masculino

Para as partidas do gênero masculino, após as duzentas repetições da separação dos dados em treino e teste e aplicação do método *backward stepwise*, a distribuição das variáveis removidas do modelo foi a apresentada na tabela 4.2. Os valores coloridos em vermelho representam os casos em que a respectiva variável foi removida do modelo em mais de 50% das repetições. Nota-se então que tais variáveis são: o mando de quadra, o número de jogos já disputados juntos, a quantidade de erros de saque, a idade média, a sequência recente de resultados e a altura média tanto da dupla de oposição quanto da dupla de referência, bem como a média móvel do

número de aces, defesas e a eficiência de ataque da dupla de referência. As variáveis que restaram foram os rankings e o número de títulos de ambas as duplas, a média móvel de bloqueios por ponto disputado da dupla de referência e as médias móveis de aces, bloqueios e defesas por ponto disuputado, além da média móvel da eficiência de ataque da dupla de oposição.

Tabela 4.2: Variáveis removidas pelo método *backward stepwise* para os jogos do gênero masculino

Variável	Qtd Remoções
dupla_opp_home	200
dupla_ref_nr_jogos	200
dupla_opp_sma_avg_serve_errors_points	198
dupla_ref_avg_age	197
dupla_ref_sma_avg_aces_points	197
dupla_opp_avg_age	194
dupla_opp_streak	191
dupla_ref_home	190
dupla_opp_nr_jogos	176
dupla_opp_avg_hgt	157
dupla_ref_avg_hgt	153
dupla_ref_streak	139
dupla_ref_sma_avg_serve_errors_points	138
dupla_ref_sma_avg_digs_points	120
dupla_ref_sma_hitpct	101
dupla_opp_sma_avg_digs_points	80
dupla_opp_sma_hitpct	16
dupla_opp_titulos	1

4.2.2 Feminino

No caso das partidas do gênero feminino, por sua vez, tal distribuição é apresentada na tabela 4.3. Neste caso, nota-se que as variáveis removidas em mais de 50% das repetições são: a idade média, a sequência recente de resultados e o número de jogos já disputados juntos de ambas as duplas, bem como o mando de quadra, o número de aces e de bloqueios da dupla de oposição, além da altura média, número de erros de saque e quantidade de defesas da dupla de referência. As variáveis que restaram foram os rankings, o número de títulos e a eficiência de ataque de ambas

as duplas, além das médias móveis do número de aces e de bloqueios por ponto disputado da dupla de referência, bem como a altura média e as médias móveis do número de defesas e de erros de saque por ponto disputado da dupla de oposição.

Tabela 4.3: Variáveis removidas pelo método *backward stepwise* para os jogos do gênero feminino

Variável	Qtd Remoções
dupla_ref_avg_age	197
dupla_ref_streak	197
dupla_opp_streak	196
dupla_opp_home	187
dupla_ref_avg_hgt	185
dupla_opp_nr_jogos	183
dupla_ref_sma_avg_serve_errors_points	173
dupla_ref_nr_jogos	161
dupla_opp_avg_age	155
dupla_ref_sma_avg_digs_points	155
dupla_opp_sma_avg_aces_points	134
dupla_opp_sma_avg_blocks_points	116
dupla_opp_avg_hgt	89
dupla_opp_sma_avg_digs_points	81
dupla_ref_home	61
dupla_opp_sma_hitpct	29
dupla_opp_sma_avg_serve_errors_points	13
dupla_ref_sma_avg_aces_points	13
dupla_ref_sma_hitpct	5
dupla_ref_sma_avg_blocks_points	4

Pode-se observar pelos resultados obtidos que as variáveis referentes ao ranking da dupla e ao número de títulos já ganhos pela dupla aparentam ser de grande relevância para o modelo, assim como a eficiência de ataque e a média móvel do número de defesas da dupla de oposição, bem como a média móvel do número de bloqueios da dupla de referência.

4.3 Correlação entre Variáveis

4.3.1 Masculino

As correlações entre as variáveis do modelo masculino completo são apresentadas na figura 4.4. Nota-se que há coeficiente de correlação de Pearson relativamente elevado entre o número de títulos ganhos por uma dupla e a quantidade jogos que ela disputou junta, bem como o número de títulos e a sequência recente de resultados. Outros coeficientes de correlação altos são apresentados entre o ranking da dupla e a quantidade de jogos que ela disputou junta e entre rank e número de títulos ganhos.

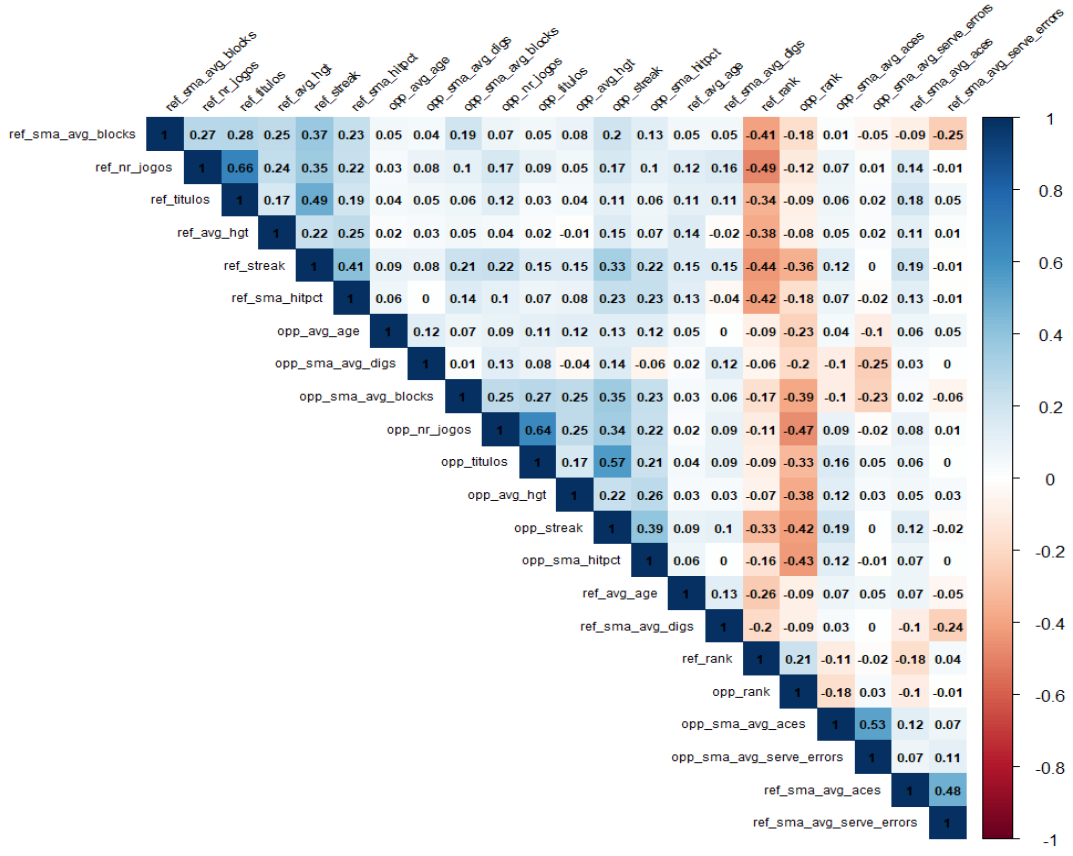

Figura 4.4: Correlações entre variáveis do modelo completo dos jogos masculinos

Para o modelo com seleção de variáveis, são removidas as variáveis citadas na seção 4.2.1. Após tal remoção, observa-se que o maior valor absoluto do coeficiente de correlação é entre as variáveis do ranking da dupla adversária e sua média móvel da eficiência de ataque, com valor de 0.43. Comparando, portanto, com as correlações de todas as variáveis, nota-se uma menor correlação entre as variáveis restantes.

4.3.2 Feminino

As correlações entre as variáveis do modelo feminino completo são apresentadas na figura 4.5. Nota-se que há coeficiente de correlação de Pearson relativamente elevado para as mesmas variáveis do modelo masculino, com destaque para a corre-

lação entre o número de títulos e a sequência recente de resultados da dupla, bem como entre o ranking e a eficiência de ataque da dupla adversária.

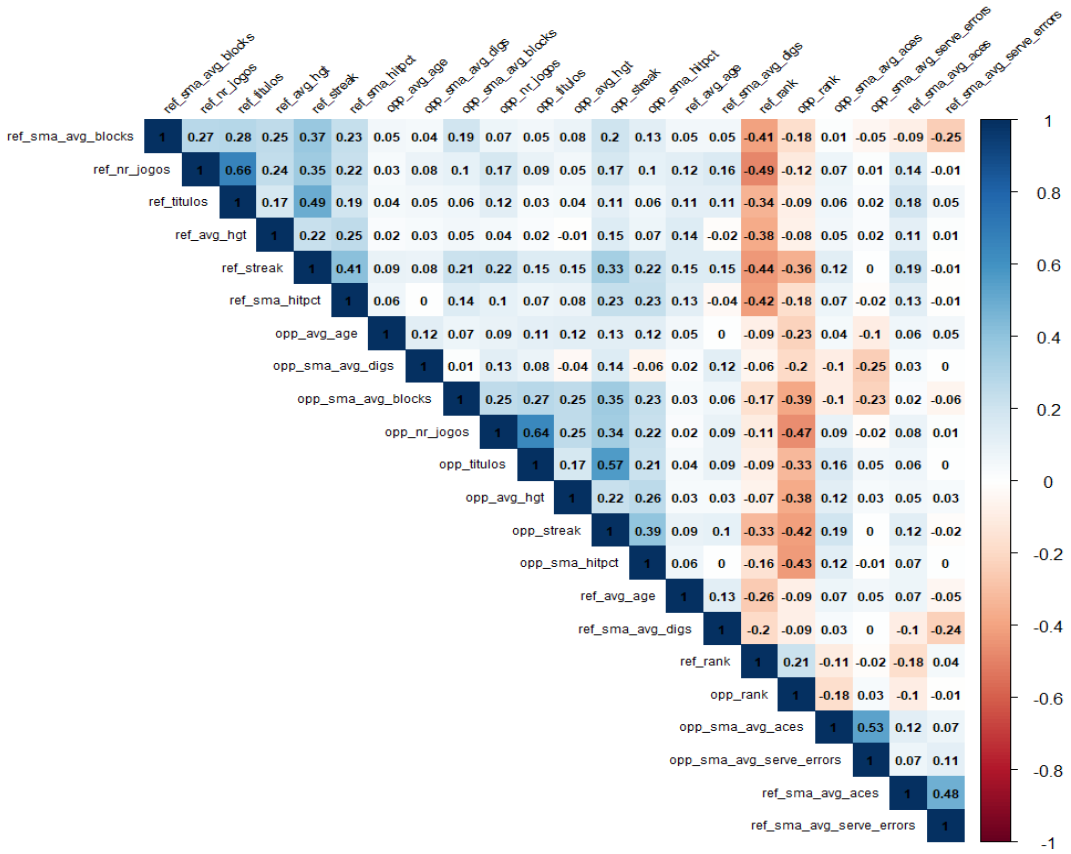

Figura 4.5: Correlações entre variáveis do modelo completo dos jogos femininos

Para o modelo com seleção de variáveis, são removidas as variáveis citadas na seção 4.2.2. Considerando apenas as variáveis restantes, observa-se que o maior valor absoluto do coeficiente de correlação é, assim como no caso das partidas masculinas, entre as variáveis do rank da dupla adversária e sua média móvel da eficiência de ataque, com valor de 0.55. Assim como no caso dos jogos masculinos, portanto, nota-se uma menor correlação entre as variáveis que permanecem após a aplicação da seleção de variáveis.

Em ambos os casos, nota-se que a correlação entre o número de títulos e o ranking das duplas é relativamente elevada, o que pode ocorrer devido ao fato de que o ranking da dupla leva em conta o número de títulos que a mesma ganhou no período de um ano.

4.4 Comparação de Modelos com e sem Seleção de Variáveis

4.4.1 Medidas de Ajuste e Diagnósticos dos Modelos

Diagnóstico dos Resíduos

A simulação dos envelopes foi repetida por dez vezes dentro de cada um dos vinte diferentes dados de treino para cada gênero, sendo que em cada uma destas vinte repetições foi obtida a média da quantidade de dados que estavam dentro dos limites superior e inferior do envelope. Depois disso, foi calculada a média destes vinte valores obtidos, as quais são apresentadas na tabela 4.4. Os gráficos obtidos para os últimos modelos calculados são mostrados na figura 4.6.

Tabela 4.4: Média das médias do percentual de resíduos dentro dos limites do envelope simulado para cada modelo

Modelo	% dentro dos limites
Completo - Masculino	0.888
Seleção de Variáveis - Masculino	0.882
Completo - Feminino	0.879
Seleção de Variáveis - Feminino	0.897

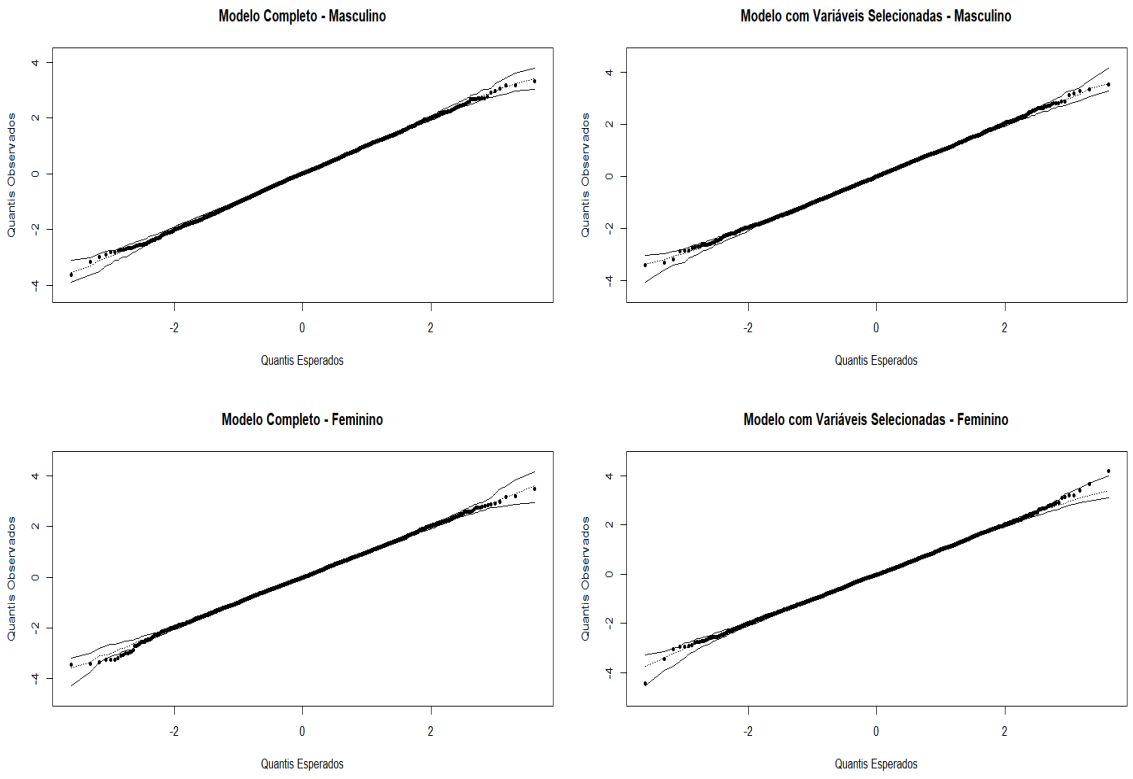

Figura 4.6: Gráficos dos envelopes simulados para cada um dos modelos

Observa-se que o percentual de resíduos dentro dos limites superiores e inferiores vai de 87,9% a 89,7% nos modelos estudados, o que, dada a subjetividade para a rejeição ou não de um modelo, podem ser tomados como valores suficientes para a aceitação dos modelos estudados.

Medidas de Ajuste

As médias dos vinte valores de AIC e pseudo-R^2 de McFadden encontradas para cada modelo são mostradas na table 4.5. Em todas as vinte repetições, o AIC foi levemente inferior para os modelos com seleção de variáveis em relação a seus respectivos modelos completos, mostrando vantagem na utilização dos modelos com menos variáveis. No caso do pseudo-R^2 de McFadden, este valor foi levemente superior para os modelos completos, sendo que em todos os casos o valor ficou entre 0.2 e 0.4, o que, segundo Hensher e Stopher (1979), representa um excelente ajuste.

Tabela 4.5: Médias obtidas para os valores de AIC e pseudo-R^2 de McFadden para cada um dos modelos testados

Modelo	AIC	R^2
Completo - Masculino	3547.284	0.227
Seleção de Variáveis - Masculino	3530.773	0.224
Completo - Feminino	3341.467	0.288
Seleção de Variáveis - Feminino	3328.246	0.285

4.4.2 Precisão das Previsões

As médias de precisão, especificidade e sensibilidade obtidas a partir das vinte repetições para cada modelo testado, bem como a quantidade de vezes em que determinado modelo apresentou maior precisão dentro de seu respectivo gênero, são apresentadas na tabela 4.6. Os resultados, tanto acurácia, sensibilidade e especificidade, são semelhantes entre os modelos completo e com seleção de variáveis, sendo que as previsões dos modelos envolvendo partidas femininas apresentam acurácia superior aos modelos envolvendo partidas masculinas. Nota-se que os valores de espeificidade e sensibilidade são seelhantes dentro de cada modelo, indicando que a probabilidade do algoritmo classificar corretamente as vitórias e aproximadamente igual à de classificar corretamente as derrotas. Os boxplots dos valores de acurácia obtidos são mostrados na figura 4.7, onde é possível notar novamente a semelhança entre as previsões entre os modelos completo e com seleção de variáveis dentro de um mesmo gênero, bem como a maior precisão das previsões dos jogos femininos.

Tabela 4.6: Média da acurácia, sensibilidade e especificidade das previsões dos modelos de regressão logística estudados

Modelo	Acurácia	Qtd com precisão superior (%)	Sensibilidade	Especificidade
Completo - Masculino	0.712	1 (5%)	0.719	0.706
Seleção de Variáveis - Masculino	0.721	19 (95%)	0.731	0.711
Completo - Feminino	0.749	9 (45%)	0.742	0.755
Seleção de Variáveis - Feminino	0.751	11 (55%)	0.744	0.757

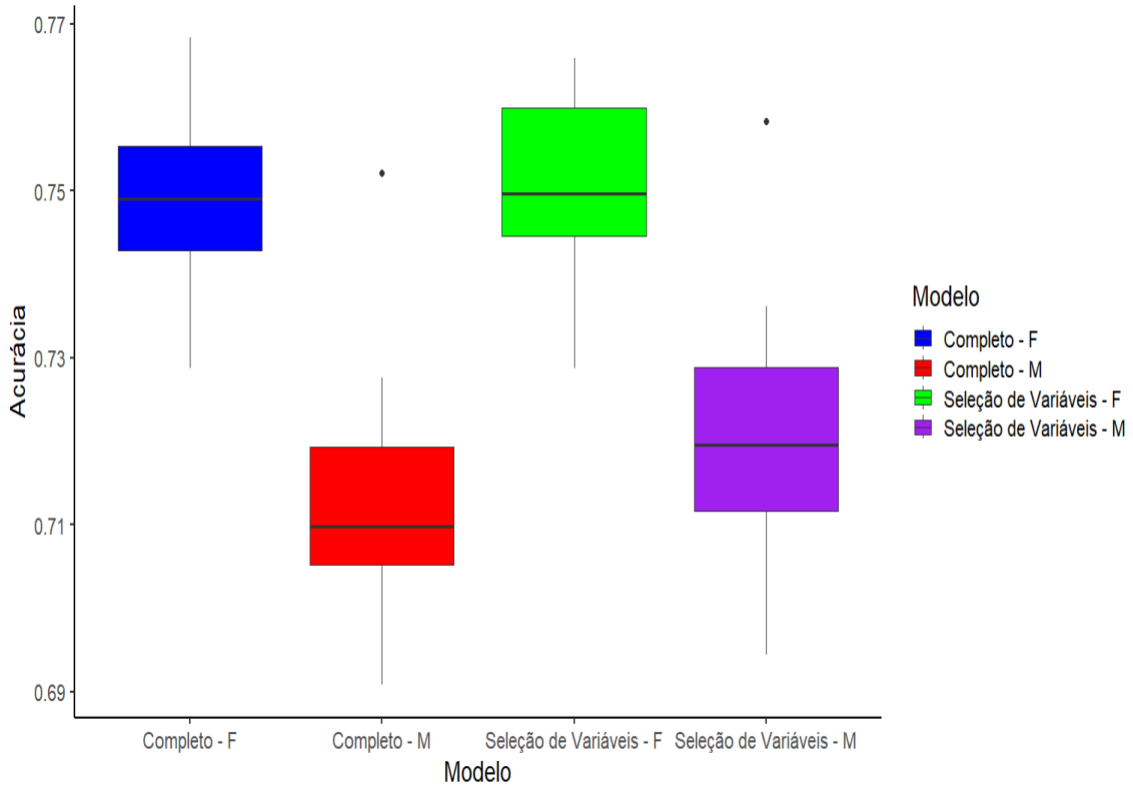

Figura 4.7: Boxplots das acurácias das previsões para cada uma das vinte repetições de cada modelo via regressão logsítica

Observa-se que, tanto para as partidas masculinas quanto para as femininas, os modelos que contam apenas com variáveis selecionadas apresentaram AIC e pseudo-R^2 de McFadden levemente inferiores em relação aos modelos completos. Além disso, nota-se que as precisões das previsões são semelhantes, em cada gênero, para os modelos completo e com seleção de variáveis. Com isso, além do fato de que se reduz o problema da correlação entre variáveis, os melhores modelos para previsão de resultados foram os que utilizaram apenas as variáveis selecionadas. Assim, nas próximas seções serão considerados somente os modelos que contém apenas as variáveis selecionadas.

4.4.3 Avaliação dos Coeficientes

As estimativas dos coeficientes, seus desvios padrão e p-valores apresentados nesta seção se referem ao modelo gerado na última das vinte repetições realizadas para cada gênero. Os valores encontrados para o modelo com seleção de variáveis dos jogos masculinos são mostrados na tabela 4.7. Pode-se observar que, enquanto variáveis como o número de bloqueios e aces apresentam coeficientes cujas estimativas são relativamente elevadas, seus desvio padrão também são altos, e o oposto ocorre para a variável de ranking e número de títulos das duplas, indicando maior estabilidade nos valores destes coeficientes. Nota-se que as variáveis que mais influenciam, com a alteração de uma unidade nos seus valores, a probabilidade de vitória de uma dupla são as referentes às médias móveis da quantidade de bloqueios e pontos de saque da dupla oponente, bem como a média móvel de bloqueios da dupla de referência .

Tabela 4.7: Coeficientes do modelo com seleção de variáveis das partidas masculinas

	Modelo com Seleção de Variáveis		
Variável	**Estimativa**	**DP**	**P-valor**
Intercepto	1.015	0.441	0.021
ref_rank	-0.119	0.007	0.000
ref_titulos	0.089	0.017	0.000
ref_sma_avg_blocks	9.519	2.760	0.001
opp_rank	0.112	0.008	0.000
opp_titulos	-0.054	0.016	0.001
opp_sma_avg_aces	-8.962	3.329	0.007
opp_sma_avg_blocks	-11.188	2.779	0.000
opp_sma_avg_digs	-1.337	1.126	0.235
opp_sma_hitpct	-1.262	0.641	0.049

Os mesmos dados referentes às partidas femininas são mostrados na tabela 4.8. Assim como para os jogos masculinos, nota-se que os desvios padrão das variáveis referentes ao ranking e número de títulos da dupla são relativamente baixos em relação aos das demais variáveis, indicando novamente maior estabilidade nos valores destes coeficientes. De acordo com o modelo, as variáveis que mais influenciam, com a alteração de uma unidade nos seus valores, a probabilidade de vitória de uma dupla são as referentes às médias móveis da quantidade de erro de saques da dupla oponente, de bloqueios e de pontos de saque da dupla de referência.

Tabela 4.8: Coeficientes do modelo com seleção de variáveis das partidas femininas

	Modelo com Seleção de Variáveis		
Variável	**Estimativa**	**DP**	**P-valor**
Intercepto	3.502	1.973	0.076
ref_rank	-0.143	0.008	0.000
ref_titulos	0.098	0.021	0.000
ref_home	-0.382	0.254	0.132
ref_sma_avg_aces	6.278	2.991	0.036
ref_sma_avg_blocks	7.847	3.483	0.024
ref_sma_hitpct	1.645	0.707	0.020
opp_rank	0.161	0.008	0.000
opp_titulos	-0.078	0.017	0.000
opp_avg_hgt	-0.043	0.026	0.094
opp_sma_avg_digs	-1.621	1.028	0.115
opp_sma_hitpct	-1.374	0.691	0.047
opp_sma_avg_serve_errors	-8.021	2.742	0.003

4.5 KNN

O valor ótimo de K mais recorrente e sua frequência nas vinte repetições, bem como a média da acurácia, especificidade e sensibilidade para os vinte conjuntos de dados de treino e testa para os modelos de cada gênero são apresentados na Tabela 4.9. Nota-se que os valores de acurácia, sensibilidade e especificidade do modelo masculino são levemente inferiores aos resultados obtidos via regressão logística. O modelo feminino, por sua vez, apresentou menor acurácia se comparado à regressao logística e, além disso, mostrou menor equilíbrio entre sensibilidade e especificidade, sendo a última maior, indicando maior probabilidade deste algoritmo prever corretamente as derrotas . A figura 4.8 mostra os boxplots com a distribuição das acurácias para as vinte repetições feitas para cada gênero. Nota-se que o modelo feminino tende a apresentar maior acurácia do que o masculino, assim como no modelo de regressão logística.

Tabela 4.9: Valor ótimo de K, média da acurácia, sensibilidade e especificidade utilizando o modelo KNN

Gênero	K ótimo	Acurácia	Sensibilidade	Especificidade
Masculino	20 (70%)	0.705	0.711	0.700
Feminino	20 (45%)	0.723	0.701	0.745

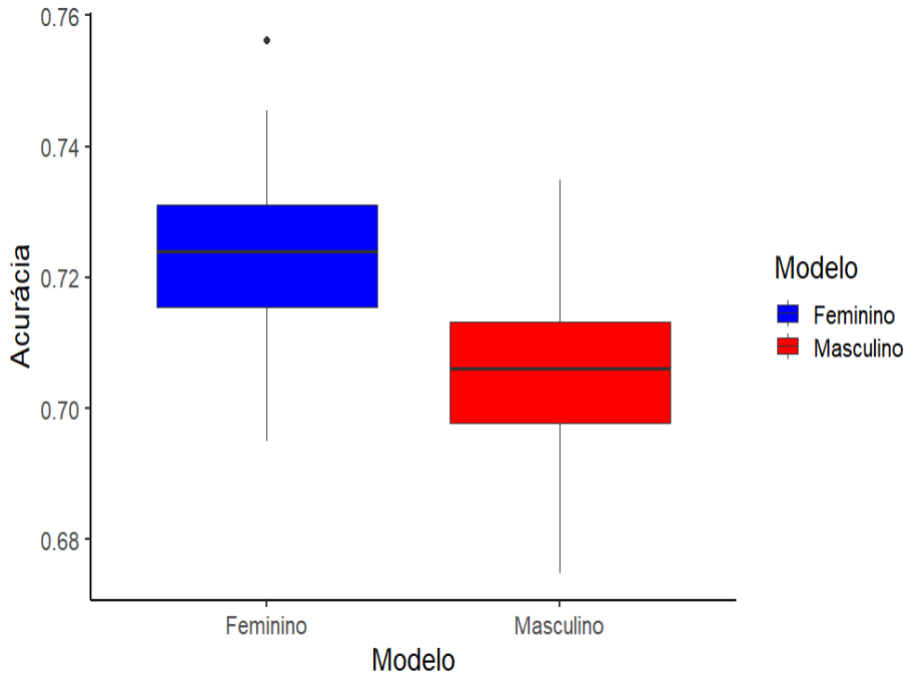

Figura 4.8: Boxplots das acurácias das previsões para cada uma das vinte repetições de cada modelo KNN

4.6 Árvores de Decisão

A média das acurácias, sensibilidades e especificidade obtidas nas vinte repetições feitas do modelo de árvore de decisão são apresentadas na tabela 4.10. Nota-se que a acurácia do modelo feminino é maior do que no masculino e que ambos apresentam acurácias menores do que no modelo de regressão logística, porém semelhantes ao KNN. Além disso, ao contrário do que ocorre no modelo KNN, é possível observar maior sensibilidade no modelo dos jogos femininos, indicando maior probabilidade deste algoritmo prever corretamente as vitórias neste caso.

Tabela 4.10: Média da acurácia, sensibilidade e especificidade utilizando o modelo de árvore de classificação

Gênero	Acurácia	Sensibilidade	Especificidade
Masculino	0.696	0.697	0.695
Feminino	0.731	0.751	0.711

A árvore gerada no último dos vinte modelos calculados para as partidas do respectivo gênero é mostrada na figura 4.9 e na figura 4.10 para os modelos dos jogos masculinos e femininos, respectivamente. Pode-se ver que, em ambos os casos, a variável do ranking da dupla é preponderante sobre todas as outras nas decisões do modelo, com destaque também para o número de títulos das duplas.

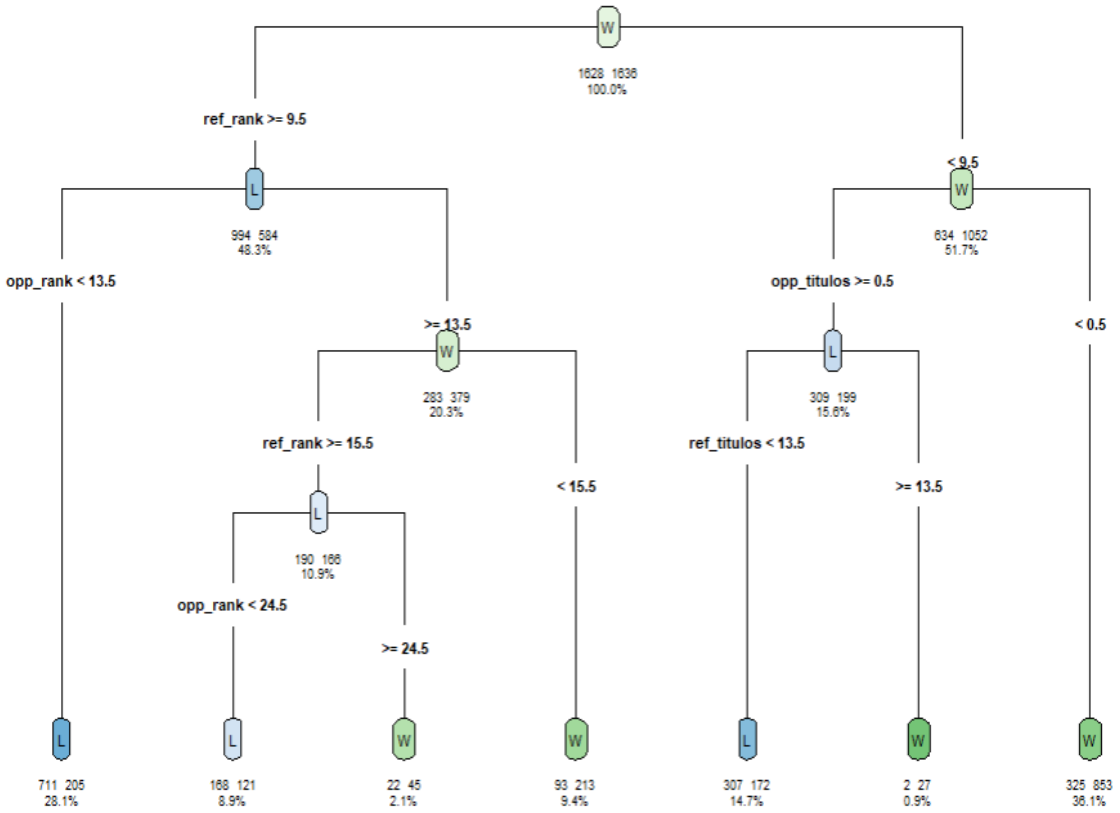

Figura 4.9: Árvore de decisão para previsão de resultados das partidas das duplas masculinas

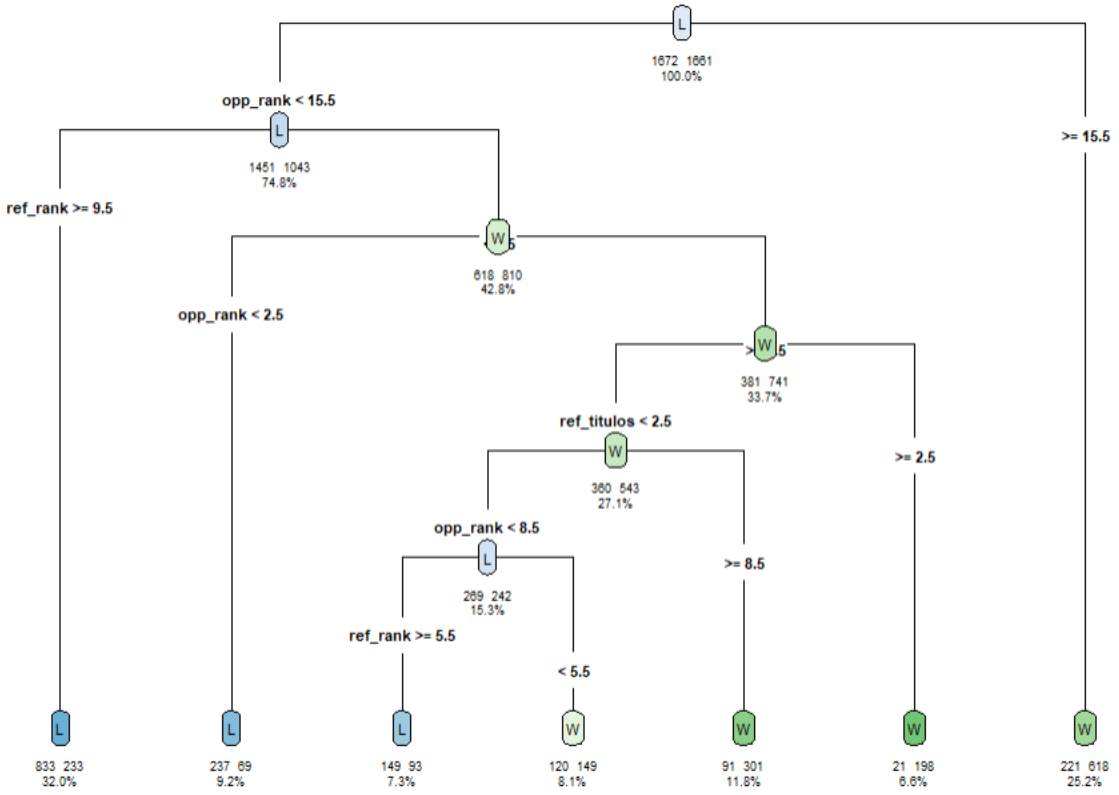

Figura 4.10: Árvore de decisão para previsão de resultados das partidas das duplas femininas

4.7 Resumo dos resultados obtidos

A Tabela 4.11 mostra os resultados de acurácia, sensibilidade e especificidade obtidos para cada um dos três métodos analisados e para cada gênero, sendo que os valores máximos de cada medida estão destacados em negrito. Os resultados de cada método se referem às médias obtidas das vinte vezes que o mesmo foi aplicado ao banco de dados já considerando somente as variáveis selecionadas. Nota-se que, como mencionado nas seções 4.5 e 4.6, as acurácias, sensibilidades e especificidades obtidas, tanto para os jogos do gênero masculino quanto feminino, apresentam valores próximos em cada um dos três métodos, tendendo a ser inferiores na árvore de decisão e KNN em comparação ao modelo de regressão logística. As flutuações encontradas entre os resultados dos modelos podem se dar pelas particularidades de cada método, devendo-se levar em conta, por exemplo, o fato de que árvores de classificação apresentam como característica importante a tendência de priorizar a interpretação e visualização dos resultados em detrimento da acurácia de previsões. Ainda assim, os resultados obtidos por cada um dos métodos variaram numa amplitude máxima de apenas 0.044 um do outro, mostrando estabilidade nos valores.

Tabela 4.11: Acurácia, sensibilidade e especificidade obtidas pelos diferentes métodos analisados

	Medida	Regressão Logística	Árvore de Decisão	KNN
Masculino	Acurácia	**0.721**	0.696	0.705
	Sensibilidade	**0.719**	0.697	0.711
	Especificidade	**0.706**	0.695	0.700
Feminino	Acurácia	**0.751**	0.731	0.723
	Sensibilidade	0.742	**0.751**	0.701
	Especificidade	**0.755**	0.711	0.745

5 Conclusões e Sugestões para Trabalhos Futuros

O presente trabalho avaliou a performance de modelos de regressão logística, KNN e árvore de classificação na previsão de resultados de vôlei de praia, tanto para partidas masculinas quanto femininas. As variáveis utilizadas foram idade e altura média dos jogadores de cada dupla, bem como variáveis referentes ao desempenho passado das duplas e médias móveis de determinadas ações de jogo de cada jogador, ponderadas pela quantidade de pontos disputados na partida. Devido à pouca quantidade de jogos do circuito da FIVB, as análises foram realizadas apenas para o circuito americano.

O método de seleção de variáveis utilizado mostrou, para os jogos masculinos, que o ranking e o número de títulos de ambas as duplas, a média móvel de bloqueios por ponto disputado da dupla de referência, as médias móveis de aces, bloqueios e defesas por ponto disuputado, além da média móvel da eficiência de ataque da dupla de oposição são as variáveis que apresentam maior importância para se definir o resultado de uma partida. Para os jogos femininos, tais variáveis são o ranking, o número de títulos e a eficiência de ataque de ambas as duplas, as médias móveis do número de aces e de bloqueios por ponto disputado da dupla de referência, bem como a altura média e as médias móveis do número de defesas e de erros de saque por ponto disputado da dupla de oposição.

As comparações entre os modelos com a utilização de todas as variáveis e os modelos que utilizaram apenas as variáveis selecionadas mostraram que o problema de correlação entre variáveis é reduzido no segundo caso e que o percentual de pontos dentro dos limites dos envelopes simulados é semelhante entre os dois modelos. Além disso, o AIC apresenta leve vantagem, enquanto o pseudo R^2 de McFadden apresenta leve desvantagem nos modelos com seleção de variáveis.

Para as previsões dos resultados das partidas via regressão logística, a precisão obtida foi semelhante, dentro de cada gênero, entre os modelos estudados, sendo que tais valores foram levemente superiores nos modelos com variáveis selecionadas, apresentando acurácia de 73% nos jogos masculinos e de 75% nos jogos femininos. Considerando esta precisão maior e as medidas de ajuste com valores semelhantes aos modelos com todas as variáveis, os modelos com variáveis selecionadas se apresentam como as melhores opções para se buscar prever os resultados das partidas de vôlei de praia. O algoritmo de árvore de decisão, por sua vez, apresentou valores de precisão levemente inferiores aos obtidos via regressão logística e demonstrou a preponderância da variável de ranking nos modelos analisados, dado que a grande maioria das decisões tomadas no algoritmo tem por base esta variável. Uma possível

explicação para esta maior importância da variável referente ao ranking das duplas é que este valor leva em conta o número de títulos e de vitórias da dupla no último ano. O método KNN apresentou K ótimo igual a vinte na maioria das repetições realizadas tanto para os jogos masculinos quanto os femininos, apresentando acurácia das previsões, sensibilidade e especificidade semelhantes às do modelo de árvore de classificação. A diferença apresentada entre os modelos pode ser explicada pelo fato de que a seleção de variáveis foi feita tendo como objetivo a aplicação da regressão logística, privilegiando maior acurácia para este método e, além disso, pelo fato de que cada modelo apresenta suas particularidades, sendo que árvores de classficação tendem a priorizar a interpretação dos resultados em detrimento de maior acurácia nas previsões, por exemplo.

Para trabalhos futuros, uma sugestão é se obter mais dados das partidas do circuito internacional, para que se possa fazer previsões e inferências sobre estes jogos, bem como compará-los às partidas da AVP. Além disso, em bases de dados com poucas informações faltantes, pode-se trabalhar com médias móveis utilizando janelas de partidas maiores do que os quatro jogos utilizados neste trabalho.

Materiais Suplementares

O código utilizado para a confecção deste trabalho pode ser visualizado pelo site https://github.com/lucas-santarossa/tcc_lucas_alvim_volei_de_praia.

Referências Bibliográficas

Atkinson, A. (1987). *Plots, Transformations, and Regression: An Introduction to Graphical Methods of Diagnostic Regression Analysis*. Oxford science publications. Clarendon Press.

Boshnakov, G., Kharrat, T., e McHale, I. G. (2017). A bivariate weibull count model for forecasting association football scores. *International Journal of Forecasting*, 33(2):458–466.

Chen, W.-J., Jhou, M.-J., Lee, T.-S., e Lu, C.-J. (2021). Hybrid basketball game outcome prediction model by integrating data mining methods for the national basketball association. *Entropy*, 23(4).

Costa, G. B., Huber, M. R., e Saccoman, J. T. (2007). *Understanding sabermetrics: An introduction to the science of baseball statistics*. McFarland e Company.

Eetvelde, H., De Michelis Mendonça, L., Ley, C., Seil, R., e Tischer, T. (2021). Machine learning methods in sport injury prediction and prevention: a systematic review. *Journal of Experimental Orthopaedics*, 8:27.

Eurosport (2021). Super cup - "it was not spontaneous" - chelsea boss thomas tuchel reveals kepa switch was planned months before victory.

Everitt, B. (1994). *A Handbook of Statistical Analyses Using S-PLUS*. Chapman and Hall/CRC.

Gabrio, A. (2021). Bayesian hierarchical models for the prediction of volleyball results. *Journal of Applied Statistics*, 48(2):301–321.

Hensher, D. e Stopher, P. (1979). *Behavioural Travel Modelling*, chapter 13. Routledge.

Horvat, T. e Job, J. (2020). The use of machine learning in sport outcome prediction: A review. *WIREs Data Mining and Knowledge Discovery*, 10(5):e1380.

Huang, M.-L. e Lin, Y.-J. (2020). Regression tree model for predicting game scores for the golden state warriors in the national basketball association. *Symmetry*, 12(5).

Ioc (2021). Net gains - the evolution of beach volleyball.

James, G., Witten, D., Hastie, T., e Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R*. Springer.

Kautz, T., Groh, B. H., Hannink, J., Jensen, U., Strubberg, H., e Eskofier, B. M. (2017). Activity recognition in beach volleyball using a deep convolutional neural network. *Data Mining and Knowledge Discovery*, 31:1678 – 1705.

Kumar, G., Shukla, A., Chhoker, A., e Thapa, R. (2021). Identification of factors determining winning in men's and women's beach volleyball: a logistical regression approach. *Teorìâ ta Metodika Fìzičnogo Vihovannâ*, 21:26–35.

Lewis, M. (2003). *Moneyball*. WW Norton.

McFadden, D. (1972). Conditional logit analysis of qualitative choice behavior.

Peng, Y.-K. e Cheng, S.-C. (2023). Analysis of critical determinant factors for beach volleyball winning in elite men and women teams. *LASE Journal of Sport Science*, 2.

Tümer, A. E. e Koçer, S. (2017). Prediction of team league's rankings in volleyball by artificial neural network method. *International Journal of Performance Analysis in Sport*, 17(3):202–211.

Wenninger, S., Link, D., e Lames, M. (2020). Performance of machine learning models in application to beach volleyball data. *International Journal of Computer Science in Sport*, 19:24–36.