title

Alisson Rosa e Vítor Pereira

Resumo

Pokémon são criaturas que vivem em todos os lugares, livres na natureza ou com os humanos, cada Pokémom tem seu tipo, pontos fortes e fracos. Com isso o objetivo desse trabalho é analisar suas estatísticas, desenvolvendo gráficos e tabelas e também construindo um modelo que dados as características do Pokémon ele irá nos fornecer uma predição se o Pokémon é lendário ou não.

Sumário

1 Introdução

Há mais de 20 anos, crianças do mundo inteiro vêm descobrindo o mundo encantado de Pokémon e muitas delas se tornam fãs para a vida toda. Hoje, a família de produtos Pokémon inclui videogames, o jogo de cartas Pokémon Estampas Ilustradas, uma série de animação, filmes, brinquedos, livros e muito mais, mas afinal que são Pokémons? Pokémons são criaturas fictícias que pertencem ao universo da série de mesmo nome - Pokémon, são semelhantes a animais do mundo real, podendo viver em bandos ou individualmente, mas também podem ser inspirados em objetos inanimados como velas, sorvetes, chaveiro e outros instrumentos. Originalmente, a série foi criada como um jogo de videogame e, com a sua popularização, se espalhou para diversos outros formatos, como séries de TV, filmes e livros.

A palavra pokémon é a contração de duas palavras em inglês: pocket, que significa bolso; e monster, que significa monstro. Assim, um pokémon é um "monstro de bolso", na tradução literal, além de ser uma contração esse seria o nome original da série, devido ao lugar onde os Pokémons são armazenados: as pokébolas, uma espécie de bola pequena para pode-los transportar com mais facilidade, sendo essas basicamente suas casas. Assim as criaturas poderiam descansar após suas batalhas, sendo essa sua principal função explorada no universo Pokémon, em que os monstrinhos lutam de acordo com habilidades da sua tipagem (Fogo, Água, Planta, Pedra, Elétrico, Voador, Lutador, Psíquico, Fantasma, entre outros.).

2 Análise Descritiva

Cada Pokémon tem seus próprios atributos, como HP (Vida), Attack (Ataque), Defense (Defesa), Speed (Velocidade) e outros mais especificos como:

- **Generation** (geração): Uma Geração em Pokémon é um grupo de jogos separados de acordo com os Pokémon que estão incluídos nela. Cada geração possui novos Pokémon, ataques e habilidades que não existem nas gerações anteriores. Aqui portanto cada Pokémon tera sua respectiva geração, sendo tratata como uma variável de fator.
- **Type** (Tipo): São classificações a que estão submetidos todos os Pokémon e técnicas (movimentos). A partir dos tipos, além de ser possível conhecer um pouco mais a natureza de cada Pokémon, é possível também elaborar estratégias de batalha. Isso porque cada tipo tem vantagens e desvantagens sobre outros tipos. Cada Pokémon pode pertencer a até dois tipos, sendo o primeiro deles o primário (Type 1) e o outro, o secundário (Type 2). Por outro lado, cada movimento tem só um tipo. Um Pokémon pode ter até quatro movimentos, mas elas não precisam ser do mesmo tipo que a criatura.
- **Special Attacks** (Sp. $< >$) : Ataques Especiais são movimentos que dão mais dano do que os anteriores, porém possuem um limitador de uso em forma de barra que deve ser carregada. Assim essa variável é dividida em **Sp. Atk** que é a força do ataque especial e **Sp. Def** que é a defesa do ataque especial.
- **Legendary** (Lendário): Pokémon Lendário (Inglês: Legendary Pokémon) é a denominação dada a uma espécie de Pokémon altamente poderosa, raríssima ou, em alguns casos, até mesmo de um único indivíduo, da qual muito se fala em lendas e mitos no mundo Pokémon, e cuja aparição é extremamente rara. Na seção de modelagem utilizaremos como variável a ser predita o Pokémon ser lendário ou não.

2.0.1 Constraste de Atributos

Nessa subseção vamos vislumbrar os atributos dos Pokémon constrastando entre os lendários e não lendários. Primeiro vejamos a média dos atributos dentre as classificações

Tabela 1: Média dos atributos entre as classificações

Legendary	HP	Attack	Defense	Sp. Atk	Sp. Def	Speed	n
Não	67.2	75.7	71.6	68.5	68.9	65.5	735
Sim	92.7	116.7	99.7	122.2	105.9	100.2	65

Assim, como esperado os pokémom lendários possuem atributos superiores (na média) do que os não lendários, note que a força do ataque especial dos pokémon lendários é 1.78 vezes maior que os não lendários.

2.0.2 Tipos e classificação

Vamos aqui estudar a quantidade de tipos por classificação dos Pokémons. O gráfico **??** fornece um vislumbre

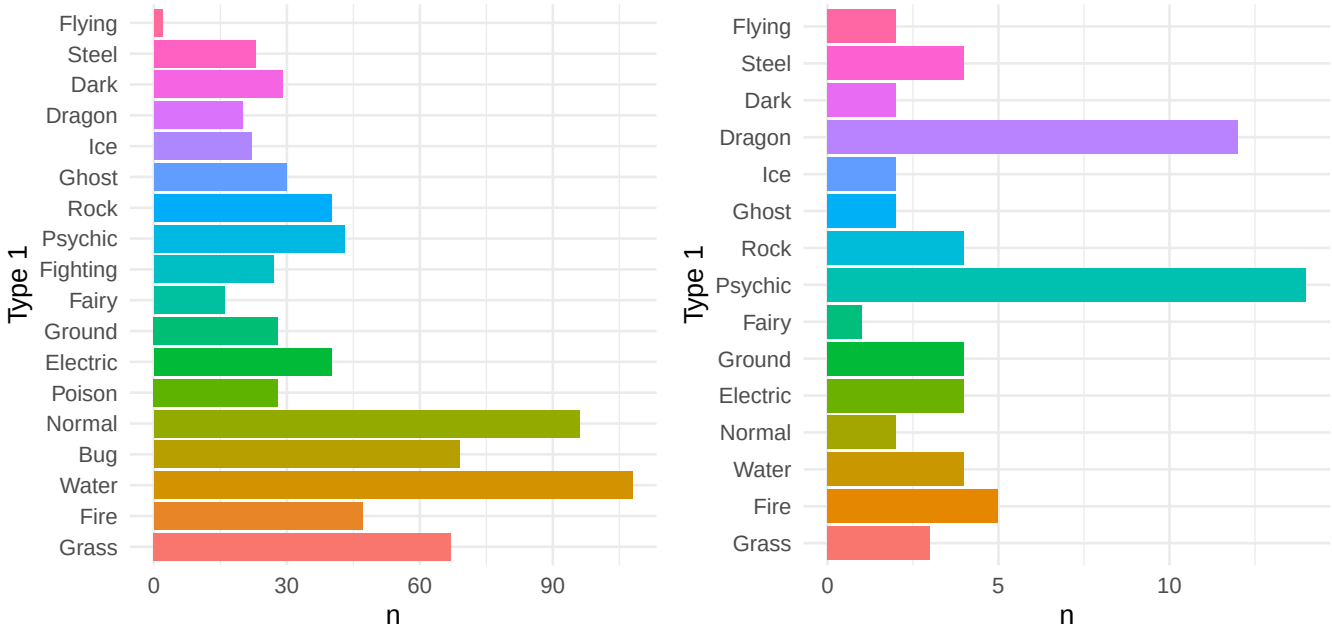


Figura 1: Frequência de Pokémons Lendários e Não Lendários por Tipo

(#fig:ref, E)

3 Análise Preditiva

Vamos aqui utilizar como variável de desfecho a classificação do pokémom, sendo portanto Lendário ou não, vamos testar a saber 3 modelos para predição: Random Forest, Regressão logística e Xgboost. Como métrica de avaliação vamos utilizar a área sobre a curva roc e não a acurácia como erroneamente muitos fazem, pois como vimos, a proporção de pokémons não lendários é 0.08, portanto se os modelos predizerem não lendário para todas observações teremos 8% de acurácia.

3.1 Regressão Logística

Regressão logistíca é um dos principais modelo estatístico atuais, pode ser descrito [1] como modelo linear generalizado (MLG). Vamos considerar p a probabilidade de sucesso de uma certa variável binária, ou seja uma variável que tem distribuição Bernoulli.

O MLG usando como função de ligação logit pode ser escrito da seguinte maneira:

$$\log\left(\frac{p}{1-p}\right) = \sum_{i=1}^{n} \beta_i X_i \quad \text{onde} \quad X_0 = 1$$

Definindo $\sum_{i=1}^{n} \beta_i X_i$ como η fica fácil ver que p pode ser escrito como:

$$p = \frac{e^{\eta}}{1+e^{\eta}}$$

Apesar do modelo de regressão logística ser mais utilizado em análise inferencial, podemos também fazer predições de classes binárias se colocarmos um limiar para a saída (p) do modelo ser classificado como de certa classe, em outras palavras se $p \geq T$, onde T é um certo limite pré estabelecido, como não temos em mãos o modelo populacional trabalhamos com a predição $\hat{p}$ para a classificação do Pokémon ser lendário, aqui utilizamos $T = 0.5$.

[1] Ou também um caso simples de uma neural network.

3.2 Random Forest

Árvores de decisão são modelos que já existem a um certo tempo, apesar de terem uma grande vantagem em interpretabilidade são fracas em termos preditivos, assim a idéia de Random Forest é combinar diversas árvores alterando o conjunto de treinamento de cada uma elas para gerar diversidade na predição, as árvores podem individualmente não serem fortes preditoras mas queremos no geral a predição combinada delas seja. Uma peculiaridade da Random Forest é que podemos ver a importância[2] das variáveis, o que é ilustrado pelo gráfico 2

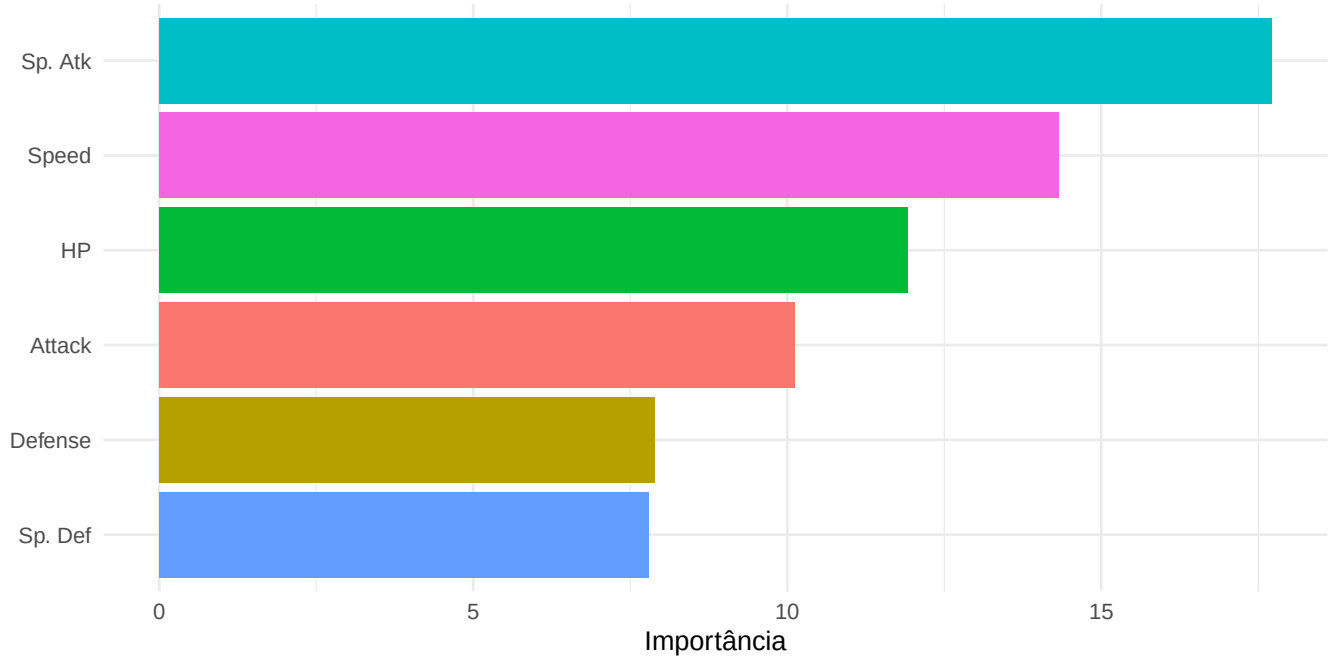


Figura 2: Importância das variáveis na Random Forest nos dados de treino

3.3 Xgboost

É notável que algoritmos de boosting atualmente são o estado da arte para dados estruturados, nele as árvores vão crescendo usando informações das árvores anteriores, isso quer dizer que não fazemos bootstrap dos dados, mas cada árvore trabalha com uma versão diferente dos dados originais, vamos aqui ajustar xgboost para comparar com os modelos anteriores.

Uma medida interessante é a matriz de confusão que pode ser vista como uma tabela que possue os valores reais cruzados com os valores preditos, vejamos para os 3 modelos ajustados como a matriz de confusão fica para os dados de teste:

Tabela 2: Matriz de confusão para os dados de teste no modelo de Regressão logística

	Predição		
Legendary	1	0	Total
1	9	9	18
0	1	181	182
Total	10	190	200

comentários

[2]Importância aqui: Decréscimo médio na impureza.

Tabela 3: Matriz de confusão para os dados de teste no modelo de Random Forest

	Predição		
Legendary	1	0	Total
1	9	9	18
0	2	180	182
Total	11	189	200

comentários

Tabela 4: Matriz de confusão para os dados de teste no modelo de XgBoost

	Predição		
Legendary	1	0	Total
1	10	8	18
0	2	180	182
Total	12	188	200

Da matriz de confusão podemos derivar as seguintes métricas:

- Valor predito positivo (`ppv`): Que é definido como sendo a proporção de predições positivas que foram corretamentes previstas
- Valor predito negativo (`npv`): Por definição é a proporção de predições negativas que foram corretamentes previstas
- Sensibilidade (`sens`): É a proporção de previsões corretas dos casos positivos
- Especificidade (`spec`): É a proporção de previsões corretas dos casos negativos

O gráfico ref fornece o vislumbre de como as métricas se comportam para os 3 modelos ajustados na validação cruzada

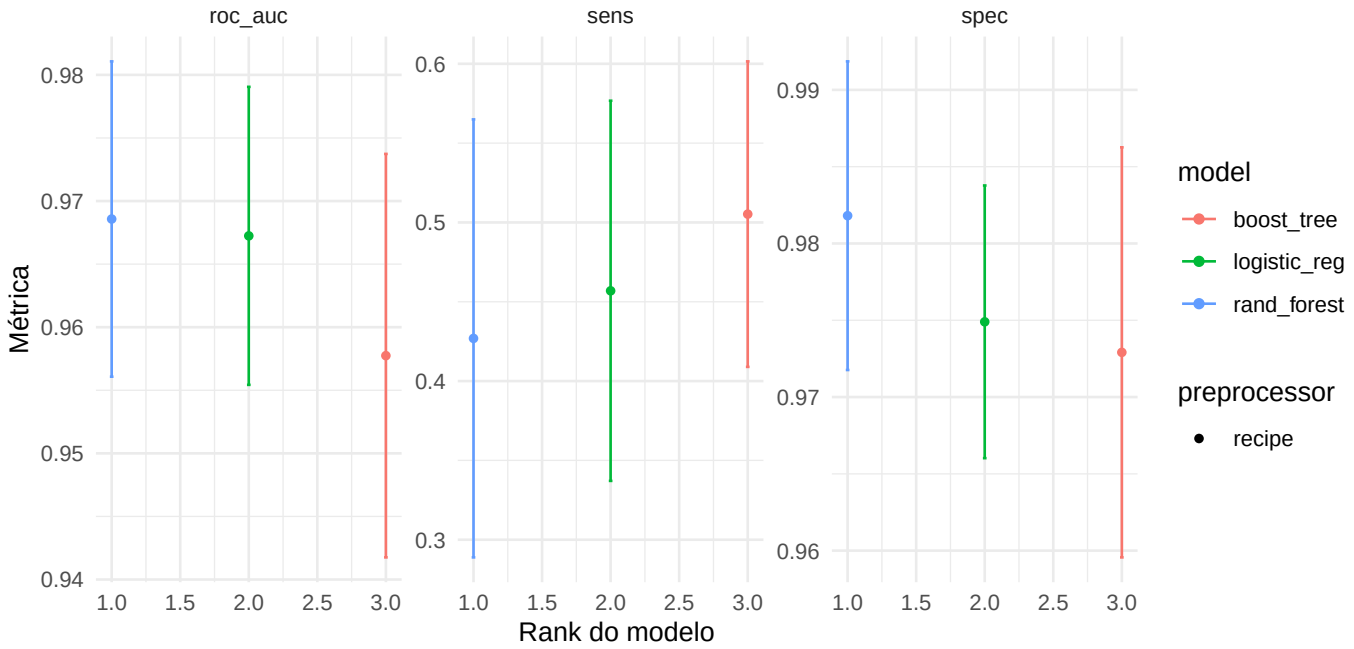


Figura 3: Métricas dos modelos na validação cruzada

E para os dados de teste temos:

Tabela 5: Métricas nos dados de teste

Modelo	sens	spec	ppv	npv	roc_auc
Reg_log	0.500	0.995	0.900	0.953	0.984
Rand_Forest	0.500	0.989	0.818	0.952	0.988
xgboost	0.556	0.989	0.833	0.957	0.981

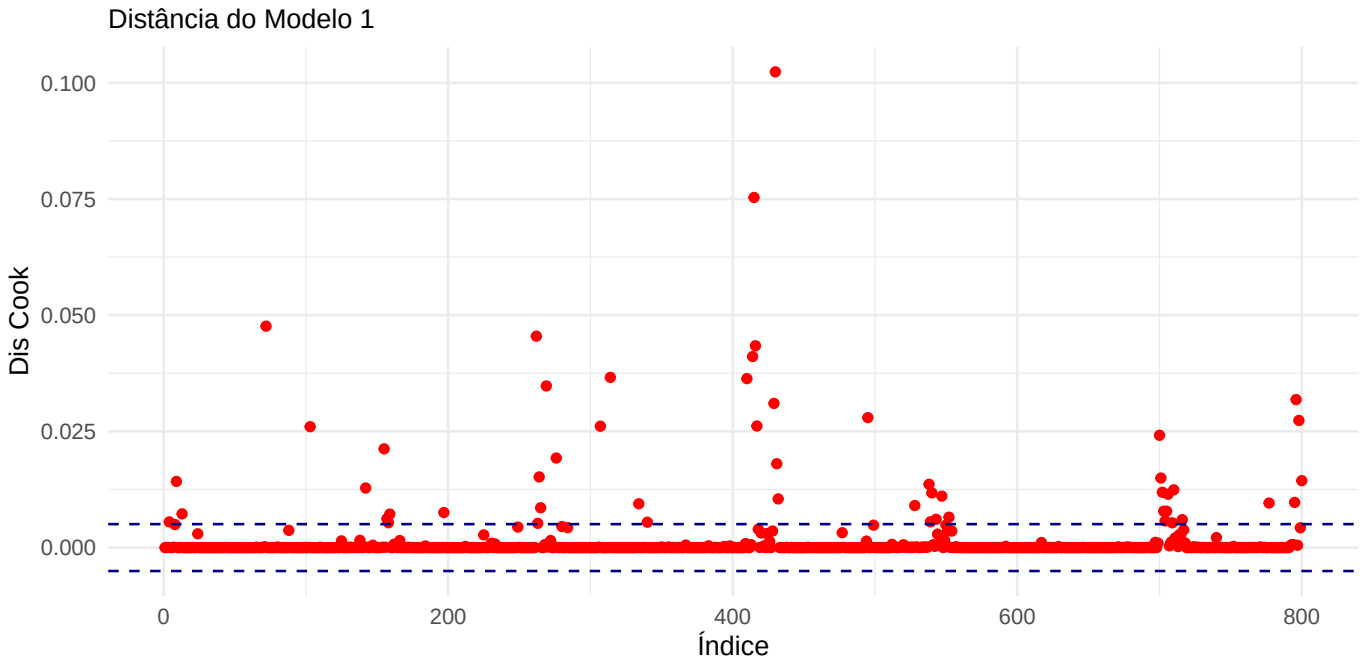
4 Análise Inferencial

4.1 Análise de Dignóstico

Vamos nessa seção avaliar a existência de pontos influentes no modelo de Regressão logística

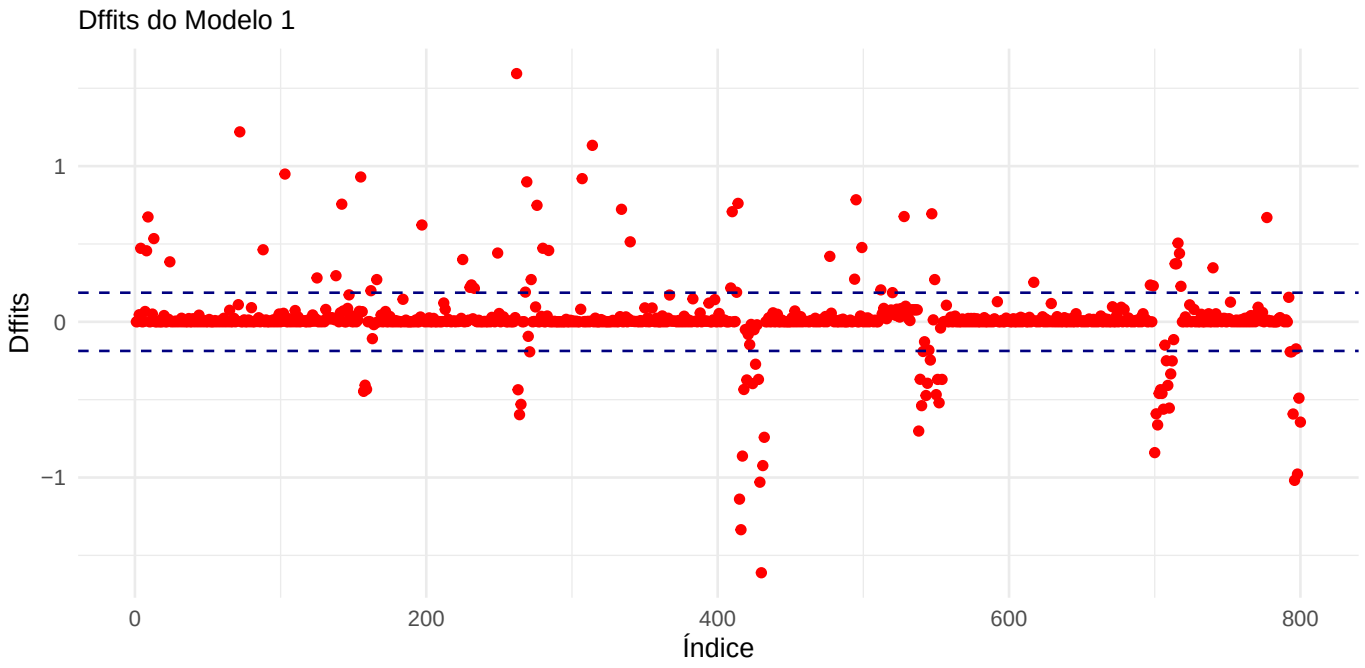
4.1.1 Distância de cook

Tem-se também a distância de cook, que fornece a influência da observação i sobre todos os n valores ajustados,

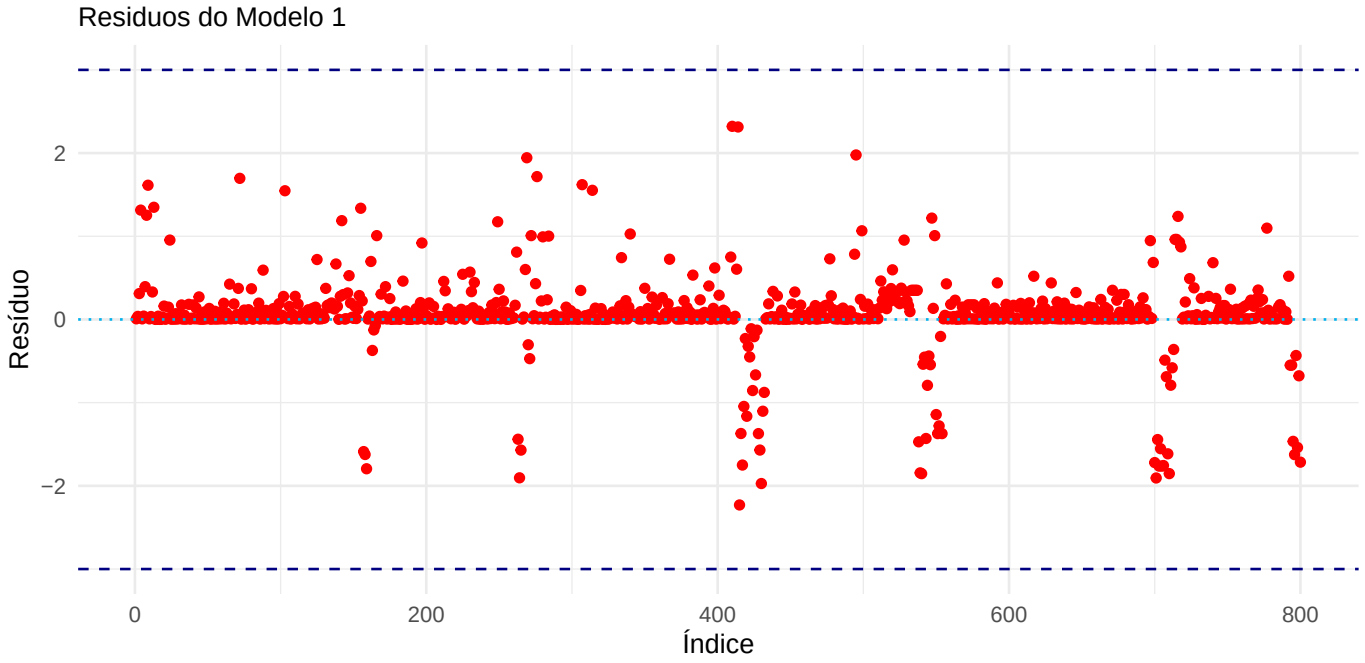


4.1.2 Dffits

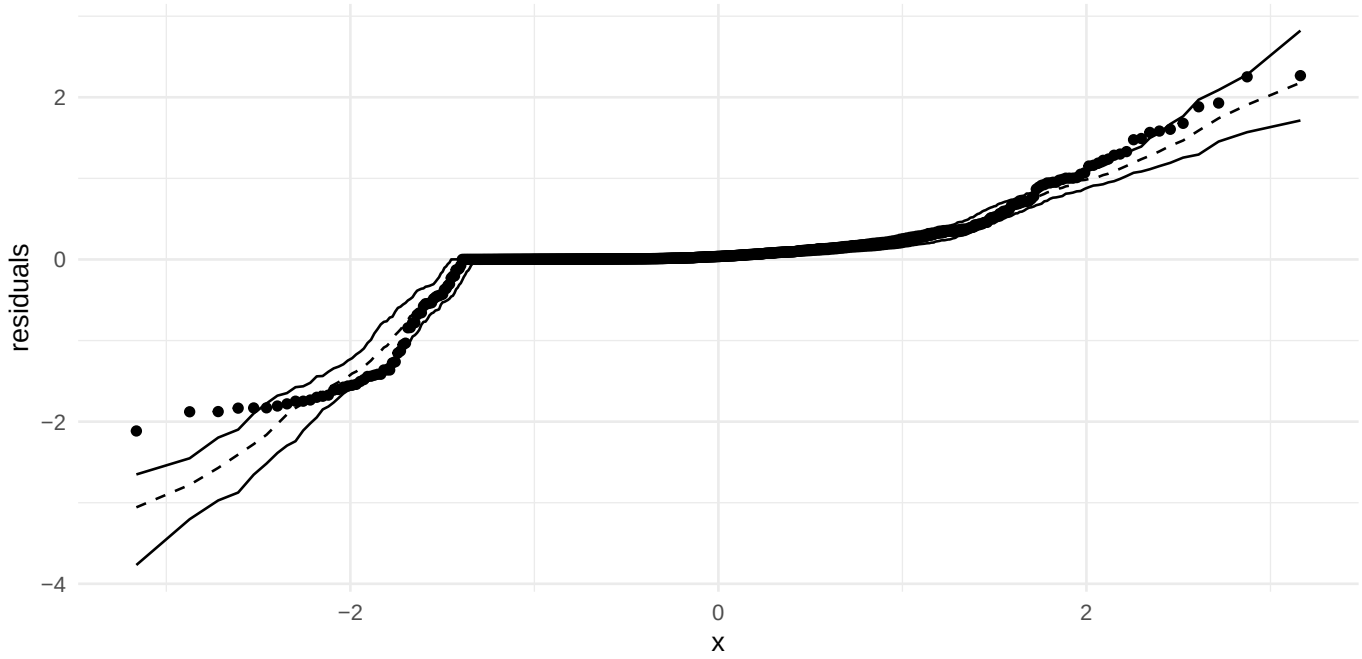
No diagnóstico dffits, que informam o grau de influência que a observação i tem sobre o valor seu próprio valor ajustado $\hat{y}_i$, percebe-se que:



O gráfico de resíduos também é importante para verificarmos visualmente a média dos resíduos e se existe algum valor fora do limite de 3 desvios padrões, pois esses possui baixíssima probabilidade de serem observados, no gráfico abaixo verificamos que todos os estados estão dentro dos limites:



E por último o envelope simulado, que fornce um vislumbre se a distribuição é adequada para o ajuste



Você faz referência cruzada de figuras assim: Figura **??**

Referências

[1] Casella G, Berger RL. Inferência estatística. Cengage Learning; 2021.

[2] R Core Team. R: A language and environment for statistical computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2021. Available from: https://www.R-project.org/.

[3] Pokémon wiki [Internet]. Fandom. Available from: https://pokemon.fandom.com/wiki/Pok%C3%A9mon_Wiki.

[4] Pokémon [Internet]. Wikipedia. Wikimedia Foundation; 2022. Available from: https://en.wikipedia.org/wiki/Pok%C3%A9mon.