UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Câmpus de Presidente Prudente

GUILHERME DE CONDE REIS

ANÁLISE E MODELAGEM ESTATÍSTICA PARA PREVISÃO DE PARTIDAS COMPETITIVAS DE LEAGUE OF LEGENDS

PRESIDENTE PRUDENTE
2025

GUILHERME DE CONDE REIS

ANÁLISE E MODELAGEM ESTATÍSTICA PARA PREVISÃO DE PARTIDAS COMPETITIVAS DE LEAGUE OF LEGENDS

Relatório Final para Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Estatística da FCT/Unesp para aproveitamento na disciplina Trabalho de Conclusão de Curso.
Orientador: Prof. Dr. Guilherme Aparecido Santos Aguilar.

PRESIDENTE PRUDENTE
2025

R375a

Reis, Guilherme de Conde

Análise e modelagem estatística para previsão de partidas Competitivas de League of Legends / Guilherme de Conde Reis. -- Presidente Prudente, 2025

76 p.

Trabalho de conclusão de curso (Bacharelado - Estatística) - Universidade Estadual Paulista (UNESP), Faculdade de Ciências e Tecnologia, Presidente Prudente

Orientador: Guilherme Aparecido Santos Aguilar Aguilar

1. Regressão logística. 2. Análise de Agrupamento Hierárquico. 3. League Of Legends. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Dados fornecidos pelo autor(a).

TERMO DE APROVAÇÃO

GUILHERME DE CONDE REIS

ANÁLISE E MODELAGEM ESTATÍSTICA PARA PREVISÃO DE PARTIDAS COMPETITIVAS DE LEAGUE OF LEGENDS

Relatório de Final de Trabalho de Conclusão de Curso aprovado como requisito para obtenção de créditos na disciplina Trabalho de Conclusão do curso de graduação em Estatística da Faculdade de Ciências e Tecnologia da Unesp, pela seguinte banca examinadora:

gov.br Documento assinado digitalmente
GUILHERME APARECIDO SANTOS AGUILAR
Data: 02/12/2025 09:16:37-0300
Verifique em https://validar.iti.gov.br

Orientador: _____ **Prof. Dr. Guilherme Aparecido Santos Aguilar**
Departamento de Estatística

gov.br Documento assinado digitalmente
MANOEL IVANILDO SILVESTRE BEZERRA
Data: 10/12/2025 11:32:09-0300
Verifique em https://validar.iti.gov.br

_____
Prof. Dr. Manoel Ivanildo Silvestre Bezerra
Departamento de Estatística

Presidente Prudente, 02 de dezembro de 2025.

Resumo

As metodologias e modelos estatísticos para análise de dados se destacam pela sua forte capacidade de não só previsão futura ou estimativas, mas também explicar e interpretar os dados. O presente trabalho tem como finalidade cumprir 3 objetivos com um propósito, neste veremos a previsão do resultado de partidas do campeonato MSI de League of Legends em 3 momentos diferentes, a partir dos primeiros 10 minutos iniciais de jogo, dos 15 minutos iniciais de jogo e também a partir da seleção de campeões, o propósito dessa análise é, não somente gerar modelos que consigam prever o futuro, mas também interpretar os modelos e os dados a fim de entender quais comportamentos em jogo levam uma equipe a vitória ou a derrota. Para atingir esses objetivos, a metodologia de previsão escolhida ao momento é a Regressão Logística, pela sua boa interpretação, não só isso como também a Análise de Agrupamento Hierárquico, a fim de classificar os campeões que participam das partidas em grupos de membros semelhantes, com o intuito de entender qual grupo pode ou não ter vantagens sobre outro.Os dados utilizados para realização da análise são referentes ao MSI (Mid-Season Invitational), que é o é o mundialito de meio de temporada, em específico dos jogos desde 01/02/2021 até 31/03/2021.

Palavras-chave: League Of Legends, MSI (Mid-Season Invitational), Regressão logística, Análise de Agrupamento Hierárquico.

ABSTRACT

Statistical methodologies and models for data analysis are distinguished by their strong capabilities in not only future prediction and estimation but also in explaining and interpreting data. This study aims to fulfill three objectives with a singular purpose: to predict the outcomes of League of Legends MSI championship matches at three different points in time—after the first 10 minutes of gameplay, after the first 15 minutes of gameplay, and based on the champion selection. The purpose of this analysis is not only to develop models that can forecast future outcomes but also to interpret these models and the data to understand the in-game behaviors that lead a team to victory or defeat. To achieve these objectives, Logistic Regression is employed due to its interpretability, along with Hierarchical Clustering Analysis to classify the champions participating in the matches into groups of similar members, aiming to identify which group may have advantages over another. The data used for this analysis pertains to the MSI (Mid-Season Invitational), specifically from matches played between 01/02/2021 and 31/03/2021.

Keywords: League of Legends, MSI (Mid-Season Invitational), Logistic Regression, Hierarchical Clustering Analysis.

SUMÁRIO

1 Introdução

Nas últimas décadas podemos notar um avanço muito grande na tecnologia, proporcionando uma acessibilidade maior a ela pela população. Com um mundo conectado, novas formas de passar tempo surgiram, como as redes sociais, streamings, conteúdos de live e junto a tudo isso os video games ganharam muito espaço e destaque. Graças a isso, hoje, vivemos em um momento que os games deixaram de ser apenas uma forma de diversão e interação para um âmbito competitivo e profissional. Segundo informações do site Itforum, muitos empregos têm sido movimentados a partir dessa nova modalidade de competitiva como narradores, apresentadores, organizadores de eventos, administradores, marketing. (Itforum, 2019)

Dito isso, um jogo que ganhou destaque mundial é o League of Legends, conhecido popularmente como LOL, lançado em 2009 e hoje oferece, não só, uma forma de entretenimento e diversão, mas também um formato profissional e global de competição. Devido a essa crescente competitividade nos cenários de eSports (esportes eletrônicos), vemos um investimento muito grande das empresas nessa modalidade esportiva, prova disso é o grande aumento na divulgação de torneios, premiações oferecidas e número de times participantes, assim como o investimento em grandes premiações. Segundo a matéria publicada no site GE, a RIOT GAMES, distribuiu aproximadamente R$1.200.000,00 em premiações do MSI (Mid-Season Invitational) no ano de 2022, mostrando o grande poder de crescimento do esporte eletrônico. (GE, 2023)

De uma forma geral, podemos explicar o League o Legends como um jogo competitivo, cujo objetivo é destruir a base inimiga até sua estrutura principal, o “Nexus”, para isso, cada equipe é composta por 5 integrantes, totalizando 10 jogadores na partida. Cada jogador controla um único personagem, sendo que não há repetição de personagens. Para destruir o Nexus e vencer a partida, a equipe se organiza em 3 rotas e possui objetivos secundários, como derrotar monstros da selva, que lhes dão recursos para vencer a partida.

Desenvolvido pela empresa RIOT GAMES, que, por sua vez, disponibiliza os históricos de suas partidas e estatísticas publicamente, de forma que estudos nessa área possam crescer.

O propósito desse projeto é usar a estatística como uma ferramenta de análise das partidas como forma de entender melhor como uma vitória de um time se dá no cenário competitivo, assim como também prever qual time vai ganhar. Através desse estudo espero ajudar a mostrar que a estatística é tão importante nos esportes eletrônicos quanto ela já é hoje nos esportes não eletrônicos. Junto a esse propósito, vale ressaltar a importância dos jogos eletrônicos nas interações sociais das pessoas. Assis (2016, p.25) afirma que “podemos argumentar que os jogos são espaços de comunhão, de encontro para interagir uns com os outros e, portanto, que jogamos para nos comunicar e não apenas para executar comandos e respostas.”. Ao refletir sobre isso, entendemos que estudos nessa área e novas tecnologias podem vir a agregar nesse espaço das interações sociais.

O presente trabalho tem como objetivo geral, através do método de aprendizado de máquina e análise multivariada, entender melhor qual é a dinâmica dos dados das partidas entre si e prever se uma equipe vai ganhar ou perder. E de forma específica atingir os seguintes pontos:

- Entender melhor como as variáveis das partidas se relacionam.
- Prever o resultado de uma partida de quem vence e quem perde.
- Usar dois modelos diferentes.
- Fazer a previsão do resultado da partida em 3 momentos diferentes, a partir da seleção de personagens (início da partida), a partir dos 10 minutos de jogo e a partir dos 15 minutos de jogo.

2 Referencial Teórico

Segundo Bolfarine (2010), na construção de uma análise estatística e até a chegada de um modelo preditivo precisamos entender o comportamento dos dados e quais são os parâmetros que os modelam. Já existem distribuições de probabilidade muito recorrentes em diferentes problemas práticos, através dessas distribuições podemos dar confiança e garantir consistência nas análises e modelos. Desde a regressão linear simples, a distribuição dos dados tem seu papel fundamental, para isso é crucial saber identificar se um determinado objeto de estudo segue uma distribuição e para isso existem artifícios estatísticos como testes e gráficos feito na análise exploratória dos dados que nos ajudam a identificar distribuição nos dados.

Nesse trabalho em questão, sua primeira etapa é estruturada por um estudo inicial dos dados e pela exploração do comportamento dos dados, dessa forma a afirmação sobre os possíveis comportamentos dos dados podem vir através da validação de testes, assumindo 2 grupos, um que ganha partidas e outro que perde partidas testes para dados 2 a 2 são de grande valia. Bussab (1987, p. 330) diz que “Feita determinada afirmação sobre uma população, usualmente sobre um parâmetro dessa, desejamos saber se os resultados experimentais provenientes de uma amostra contrariam ou não tal afirmação”, ou seja, podemos, através dos testes, validar se comportamentos que acreditamos ser influentes na definição de um time vencedor é estatisticamente válido, entrando no conceito de P-valor para valida nossas hipóteses.

Caminhando para o objetivo principal desse trabalho, nos voltamos para os conceitos da Regressão Logística, essa técnica busca estabelecer uma relação entre as variáveis independentes, comumente chamadas de X, com a variável dependente conhecida comumente de Y. Segundo Gonzalez (2018, p. 6) “A regressão logística, que é uma técnica para mineração de dados de resposta categórica, nas suas formas binárias e múltiplas.”, dessa definição, a resposta categórica que buscamos é vitória ou derrota, que já dentro do modelo pode ser interpretado como 1 para caso ocorra a vitória e 0 para ocorrência de derrota. Para contextualizar outros usos da regressão logística, essa é muito usada para análise de crédito, já que a saída do modelo pode ser interpretada como uma probabilidade e ser usada como forma de pontuação para os clientes.

Após a montagem dos modelos devemos escolher o que melhor se desempenha. Capita (2023) mostra em seu trabalho sobre regressão logística em League of Legends, com 3 variáveis independentes, desde a montagem até a escolha do modelo que melhor performa. Para comparação dos modelos ele usa o R^2, AIC e BIC, não só isso, como também Teste Hosmer e Lemeshow que indica qual modelo mais acerta os dados. Apesar do trabalho em questão também tentar usar dá modelagem para prever partidas de League of Legends, o presente estudo comparado ao de Capita (2023) se diferencia por 2 fatores principais, a utilização de 2 modelos para predição e a utilização de mais que 3 variáveis independentes.

Como o interesse desse trabalho em questão é de usar 2 modelos com muitas variáveis, devemos escolher com cuidado que tipo de modelo usar. Calvo (2017, p. 23) afirma que “Há a necessidade de verificar o quão complexo o modelo deve ser, dado a quantidade de parâmetros a ser estimada e o poder computacional alocado para tal tarefa. Muitas vezes métricas de associação entre os atributos de entrada e o atributo alvo auxiliam no descarte de alguns atributos de entrada, diminuindo por consequência a quantidade de dependências a serem criadas.”, a partir disso, entende-se que a escolha das variáveis implica diretamente na escolha do modelo a ser implementado.

3 Metodologia

Como dito antes, o objetivo desse trabalho é tentar prever o resultado das partidas de League of Legends. Para isso, será usado uma base de dados que contêm partidas dos campeonatos competitivos do jogo, sendo que cada linha da base de dados é uma partida no jogo que contém informações dos personagens que compõem cada equipe e as variáveis da partida até os 10 e 15 minutos de jogo, assim como qual foi a equipe vencedora.

Para atingir o objetivo do trabalho, será primeiramente feito uma análise exploratória e multivariada dos dados buscando encontrar as correlações e se possíveis combinações, essa primeira etapa servira como forma de esclarecer o comportamento das variáveis entre si e com o resultado da partida a fim de selecionar apenas variáveis relevantes a previsão que o modelo busca oferecer.

Com o conhecimento das variáveis e os seus comportamentos será escolhido dois modelos para previsão do resultado que melhor se encaixem com as variáveis da base de dados. No primeiro modelo será utilizado a escolha dos personagens de cada equipe para prever, dessa forma teremos uma estimativa de quem vence a partida já no seu início. O segundo modelo levará em consideração os dados da partida e poderá ser treinado de duas formas diferentes, com os dados até 10 minutos de jogo ou 15 minutos de jogo.

Vale ressaltar e esclarecer, que o intuito do primeiro modelo (modelo 1) e do segundo (modelo 2) requerem diferentes dados, já que o modelo 1 leva em consideração apenas a seleção dos campeões e o modelo 2 leva em consideração dados do desempenho da partida até determinado minuto de jogo. De forma geral, o que será feito é uma seleção de quais variáveis da base de dados vão treinar o modelo 1 e quais vão treinar o modelo 2, lembrando que ambos os modelos vão prever os resultados das mesmas partidas.

Caminhando para o objetivo desse trabalho, após os dois modelos treinados, é pretendido comparar os seus desempenhos entre si. Além disso, sabendo que o modelo 1 prevê o resultado da partida a partir de seu início, ou seja, a partir do minuto zero, podemos usar o “chute” do modelo 1 como uma variável do modelo 2 e ver se o seu desempenho melhora, tal interação entre os modelos faz sentido, pois em uma

partida em tempo real, o modelo 1 consegue estimar o resultado da partida assim que os jogadores escolhem seus personagens e o modelo 2 só entra em ação no mínimo aos 10 minutos de jogo.

O Software R será utilizado para as análises estatísticas.

4.Método de Regressão Logística.

A regressão logística é uma técnica estatística utilizada para modelar a relação entre uma variável dependente binária (isto é, que assume dois valores, como "sim" e "não", "1" e "0") e uma ou mais variáveis independentes contínuas ou categóricas independentes.

A grande vantagem e o motivo de escolha dessa técnica para esse trabalho é a interpretação que ela nos proporciona sobre a relação das variáveis explicativas e a variável de interesse.

Em Lemeshow (1989, p.25) se nos lembrarmos de Regressão linear, onde os dados a serem modelados variam de $-\infty$ até $+\infty$ e sua função definida por:

$$Y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p. \quad (1)$$

Podemos ver que a forma como se chega à regressão logística é partindo do que sabemos da linear, agora com uma variável de interesse binária (0 ou 1).

Como na Regressão Linear, nosso Y varia entre -∞ até +∞, assim como a função, precisava – se pensar em uma forma de fazer essa função variar entre 0 e 1, para chegarmos em uma Regressão Logística. Partindo dessa premissa, segue o passo a passo dessa transformação:

1°: Sendo $x = (x_1, x_2, \dots, x_{\mathrm{p}})$ uma coleção de variáveis independentes e assumindo que $p = \pi(x) = P(Y = 1|x)$.

2°: Podem-se começar igualando o $\ln\left(\frac{\hat{p}}{1-\hat{p}}\right)$ com a fórmula da Regressão Linear, dessa maneira é como se estivesse tentando fazer um ajuste para que usasse a regressão linear para modelar o logaritmo natural da chance de o evento ocorrer, pois esse sim varia de -∞ até +∞, mas não é aqui que essa ideia para, ela deve ser desenvolvida para chegar no modelo de Regressão Logística.

$$\ln\left(\frac{\hat{p}}{1-\hat{p}}\right) = b_0 + b_1 x_1 + \cdots + b_p x_p. \quad (2)$$

3°: O próximo passo é isolar o $\hat{p}$, já que ele é a estimativa que queremos dentro de uma regressão logística. Para isso, colocamos a e dos dois lados, da seguinte forma:
Obs: $b_0 + b_1x_1 + \cdots + b_px_p = z$

$$e^{\ln\left(\frac{\hat{p}}{1-\hat{p}}\right)} = e^z. \quad (3)$$

Deixando a fórmula

$$\frac{\hat{p}}{1-\hat{p}} = e^z. \quad (4)$$

4°: Nosso objetivo continua sendo isolar o $\hat{p}$, logo

$$\hat{p} = e^z * (1-\hat{p}). \quad (5)$$

5°: Distribuindo e^z temos:

$$\hat{p} = e^z - \hat{p} * e^z. \quad (6)$$

6°: Isolando $\hat{p}$:

$$\hat{p} + \hat{p} * e^z = e^z. \quad (7)$$

7°: Colocando $\hat{p}$ em evidência:

$$\hat{p} * (1 - e^z) = e^z. \quad (8)$$

8°: Isolando $\hat{p}$, chegamos na seguinte fórmula:

$$\hat{p} = \frac{e^z}{1 + e^z}. \quad (9)$$

Substituindo o Z pela fórmula da Regressão Linear, temos a fórmula da Regressão Logística:

$$\hat{p} = \frac{e^{b_0+b_1x_1+\cdots+b_px_p}}{1 + e^{b_0+b_1x_1+\cdots+b_px_p}}. \quad (10)$$

Ou

$$\pi(\mathrm{x}) = \frac{e^{b_0+b_1x_1+\cdots+b_px_p}}{1 + e^{b_0+b_1x_1+\cdots+b_px_p}}. \quad (11)$$

Dada a equação, agora para estimar seus parâmetros é utilizado o método de Máxima Verossimilhança.

4.1 Máxima verossimilhança

O método da máxima verossimilhança estima parâmetros desconhecidos maximizando a probabilidade dos dados observados. Primeiro, cria-se uma função de verossimilhança que expressa a probabilidade dos dados em relação aos parâmetros. Os valores dos parâmetros que maximizam essa função são escolhidos como estimadores de máxima verossimilhança.

Segundo Lemeshow (1989, p.9) a função:

$$\zeta(\mathrm{x_i}) = \pi(\mathrm{x_i})^{\mathrm{y_i}}[1 - \pi(\mathrm{x_i})]^{1-\mathrm{y_i}}. \quad (12)$$

É uma boa forma de expressar uma contribuição para a função de verossimilhança, pois quando $y_i = 1$ a contribuição da função acontece por parte do $\pi(x_i)$, já quando$y_i = 0$, a contribuição da função acontece pela parte $[1 - \pi(x_i)]$.

A partir disso, o produtório dessa função é feito, assumindo independência dos dados observados, dando a expressão:

$$l(B) = \prod_{i=1}^{n} \zeta(x_i). \quad (13)$$

A ideia agora é derivar a função e igualar a zero para assim encontrar os $\beta's$ que maximizam a expressão, para facilitar essa etapa é feito o logaritmo natural da equação:

$$L(B) = ln[l(B)] = \sum_{i=1}^{n} [y_i \cdot \ln(\pi(x_i) + (1 - y_i) \cdot \ln(1 - \pi(x_i)]. \quad (14)$$

Agora trabalhar com essa expressão é mais simples, fazendo sua derivação igual a zero, chegando nas expressões:

$$\sum_{i=1}^{n} [y_i - \pi(x_i)] = 0. \quad (15)$$

e

$$\sum_{i=1}^{n} x_{ij}[y_i - \pi(x_i)] = 0. \quad (16)$$

Diferente da regressão linear, na regressão logística a solução das equações (15) e (16) não são lineares em relação aos b_0 e b_j com j = 1, 2, ..., p, e assim faz-se necessário métodos especiais, esses são iterativos e se encontram disponíveis em softwares programados de análise de dados com suporte para regressão logística.

4.2 Método de seleção de variáveis

Como Montgomery (2001, p.310) explica, quando o número de variáveis aumenta, fazer todas as possibilidades de modelos acaba sendo inviável, daí que surgiram diferentes métodos que visam apenas alguns modelos possíveis ao invés de todos.

4.2.1 Método Backward

O método backward começa com um modelo que inclui todas as variáveis candidatas e, em seguida, remove variáveis uma a uma, com base em critérios estatísticos, até que a remoção de variáveis não melhore o modelo.

Passos:

1. Início: Começa com um modelo que inclui todas as variáveis preditoras possíveis (modelo completo).
2. Remover Variáveis: Em cada passo, remove-se a variável que menos contribui para o modelo, tipicamente a variável com o maior p-valor.

3. Critério de Parada: O processo continua até que a remoção de variáveis adicionais não aumente significativamente o critério estatístico (por exemplo, quando todas as variáveis restantes têm p-valores menores que um limiar predefinido).

4.2.2 Método Forward

O método forward começa com um modelo vazio e adiciona variáveis uma a uma, com base em critérios estatísticos, até que a adição de novas variáveis não melhore significativamente o modelo.

Passos:

1. Início: Começa com um modelo sem variáveis (modelo nulo).
2. Adicionar Variáveis: Em cada passo, adiciona-se ao modelo a variável que mais melhora o critério estatístico (como AIC, BIC ou p-valor).
3. Critério de Parada: O processo continua até que a adição de novas variáveis não reduza significativamente o critério escolhido (por exemplo, quando o p-valor das variáveis candidatas a entrar é maior que um limiar predefinido).

4.2.3 Método Stepwise

O método stepwise combina aspectos dos métodos forward e backward. Ele começa com um modelo inicial (que pode ser vazio ou conter algumas variáveis) e, em cada etapa, adiciona ou remove variáveis com base em critérios estatísticos, como o valor do AIC (Akaike Information Criterion), BIC (Bayesian Information Criterion) ou testes de significância (p-valores).

Passos:

1. Início: Pode começar com um modelo vazio (sem preditores) ou com um modelo contendo algumas variáveis selecionadas.
2. Adicionar Variáveis: Adiciona-se a variável que melhora mais o critério estatístico escolhido (por exemplo, a que reduz mais o AIC).
3. Remover Variáveis: Após adicionar uma variável, o método verifica se a remoção de alguma das variáveis existentes melhora o critério. Remove-se a variável que menos contribui para o modelo (aquela com o maior p-valor, por exemplo).
4. Repetir: Repete os passos de adição e remoção até que não haja mais melhorias significativas.

4.3 Chance

Chance é a probabilidade do evento acontecer dividida pela probabilidade do evento não acontecer, Ela varia de 0 a infinito.

$$\frac{\hat{p}}{1-\hat{p}}. \tag{17}$$

A partir dela podemos dizer quantas vezes mais ou menos um evento tem probabilidade de acontecer.

Dentro do contexto da regressão logística, podemos definir a chance do evento ocorrer como:

$$\frac{\pi(x)}{1-\pi(x)}. \tag{18}$$

O log da equação (18) é definido como o logit e esse é definido como:

$$\ln\left(\frac{\pi(x)}{1-\pi(x)}\right). \tag{19}$$

4.4 Odds Ratio

Como dito anteriormente, a escolha da regressão logística nesse trabalho foi feita pela interpretação de seus parâmetros, como Montgomery (2001 , p.452) explica, a Odds Ratio pode informar o aumento da probabilidade de sucesso de um evento dado o incremento da variável independente explicativa.

Definição

A Odds Ratio é uma medida que relativiza a chance (odds) de um determinado evento ocorrer em um grupo A com a chance de ocorrer em um grupo B. Trazendo essa definição para dentro do conceito de regressão, podemos dizer comparar o grupo A como x_{i+1} e o grupo B como o x_i. Dessa forma, podemos escrever a odds ratio como

$$\widehat{O_R} = \frac{\text{odds}_{x_i+1}}{\text{odds}_{x_i}} = e^{\widehat{\beta_1}} \tag{20}$$

Ou seja, ela serve para medir a associação entre as variáveis, fazendo isso de dois a dois. Trazendo essa medida para o contexto de análise desse trabalho, o interesse em interpretar essa medida se dá pelo entendimento que ela pode

proporcionar a respeito de quais variáveis e comportamentos em jogo pode fazer com que o time conquiste a vitória.

4.5 Pressupostos para fazer uma regressão logística

Advindo da regressão linear, a logística também tem pressupostos, sendo esses:

- A variável dependente tem que ter um relacionamento linear com as variáveis independentes;
- As observações devem ser independentes um das outras;
- As variáveis independentes não devem apresentar multicolineariedade entre si;
- É importante avaliar a capacidade discriminativa do modelo usando métricas como a curva ROC.

5 Análise de agrupamento.

No presente trabalho, especificamente em objetivos específicos, foi proposto uma previsão de qual time ganha a partir apenas dos personagens que jogam em cada time, tendo isso em mente o agrupamento de indivíduos semelhantes pela metodologia de cluster será utilizado de forma que seja possível certas análises. Para contextualizar, é necessário ressaltar que cada equipe competidora escolhe 5 personagens diferentes por partida, totalizando 10 com as 2 equipes, sendo assim, sabendo que existem mais de 140 campeões diferentes, teríamos inúmeras possibilidades de combinações diferentes de equipes (em específico são 2.081.693.722.421.538.086.400 possibilidades). Nesse caso uma estratégia de estudo é a de identificar indivíduos semelhantes e estudar o desempenho desse grupo seleto ao invés de analisar cada personagem individualmente.

Essa ideia de unir indivíduos semelhantes é conhecido como agrupamento ou Clustering, esse é descrito pela Mingoti (2005, p.21) como parte do grupo de técnicas Análise Multivariadas que simplificam a estrutura dos dados, trazendo uma metodologia prática, que geralmente independe da distribuição dos dados.

Além do que já foi dito, para explicar melhor a ideia de agrupamento no presente trabalho cabe muito bem citar Johnson (2007, p.671) que descreve o agrupamento como uma forma de poder levantar hipóteses a respeito de relacionamentos dos dados, que pode ser traduzido a o presente trabalho como a relação dos grupos na hora de montar uma equipe em busca da vitória, hipóteses como “O grupo A tem vantagem sobre o grupo B?” podem ser relevantes para entender qual equipe tem mais probabilidade de vencer uma partida.

5.1 Distância Euclidiana

A ideia por trás da técnica é criar medidas que possam representar quanto um indivíduo é semelhante ou diferente ao outro para saber quem deve ser agrupado. Essas medidas são denominadas como medidas de dissimilaridade.

Para saber que indivíduos agrupar faz-se necessário criar um critério e medidas para avaliar a quais indivíduos são semelhantes e ou diferentes entre si. Essas

medidas de semelhança são denominadas de Similaridade e as de diferença são de Dissimilaridade, logo é intuitivo que caso a similaridade de 2 indivíduos for alta, tendemos a colocá-los no mesmo grupo, já no caso da dissimilaridade é o contrário, esses indivíduos só são agrupados quanto menor sua medida de distância. Como passo inicial desse trabalho vamos começar com uma medida mais simples de calcular e comumente usada, a Distância Euclidiana, calculada, segundo Mingot (2005, p.165) da seguinte maneira.

$$d(X_l, X_k) = \sum_{i=1}^{p} \sqrt{(X_{il} - X_{ik})^2} \ , \ \ l \neq k \tag{21}$$

Após feita a matriz de distâncias de cada personagem entre si, sem considerar a distância do personagem para si mesmo, pois essa é igual a 0. Vem então a decisão de quais personagens devem ser agrupados

5.2 Agrupamento Hierárquico

O método hierárquico recebe esse nome pois, como descrito em Mingot (2005, p.165) e “As técnicas hierárquicas aglomerativas partem do princípio de que no início do processo de agrupamento tem-se n conglomerados, ou seja, cada elemento do conjunto de dados é considerado como um conglomerado isolado. Em cada passo do algoritmo, elementos amostrais vão se agrupando e formando novos conglomerados até o momento no qual todos os elementos considerados estão num único grupo”, essa definição já nos diz muito sobre a técnica, mas o que podemos enfatizar dela é o fato de que cada indivíduo é inserido a um grupo a cada passo do algoritmo, nos dados a informação de que cada cluster é formado em diferentes níveis de granularidade, nos permitindo posteriormente, observar o histórico de agrupamento para construir o gráfico de Dendrogama.

5.3 Método Ward

Definida a média de distância entre personagens, em função do algoritmo hierárquico seja executado é requerido assumir uma forma de definir a posição do grupo na base para realizar o cálculo, uma medida comumente usada e o Ward, esse definido por Johnson (2007) como um método que minimiza a perda de informação é

feito da seguinte forma, primeiro é calculado o denominado ESS de cada grupo formado, sendo $\bar{x}_{\mathrm{A}}$e $\bar{x}_{\mathrm{B}}$ elementos dos grupo A e B, respectivamente, e na e nb o número de observações de cada grupo, então:

$$ESS_A = \sum_{j=1}^{na} (x_{Aj} - \bar{\mathrm{x}}_A)'(x_{Aj} - \bar{\mathrm{x}}_A) \tag{22}$$

$$ESS_B = \sum_{j=1}^{nb} (x_{Bj} - \bar{x}_B)'(x_{Aj} - \bar{x}_B) \tag{23}$$

Podemos ver que o ESS é a soma de quadrado do desvio da média de cada variável do grupo, a partir do momento que existem K grupos formados, o ESS é calculado pela soma do ESS de cada grupo individual, sendo.

$$ESS = ESS_1 + ESS_2 + \ldots + ESS_k$$

A partir disso, a decisão de qual grupo deve ser unido é feita através da união que produzir o menor aumento do ESS calculado.

5.4 Método Ligação simples (Single Linkage)

Esse método, que é uma alternativa ao método Ward já descrito, propõe uma lógica de agregação diferente, a Ligação simples, segundo Mingot (2005, p.166), define a distância entre grupos a partir da menor distância entre a observações desses. De forma que, sendo G1 um grupo e G2 outro, então a distância desses grupos na matriz de distâncias será dada por:

$$d(G_1, G_2) = \min\{\, d(X_l, X_k, l \neq k,)\} \tag{24}$$

A distância entre dois grupos é definida pela menor distância entre qualquer par de observações, onde uma observação pertence ao primeiro grupo 1 (G1) e a outra pertence ao segundo grupo 2 (G2).

5.5 Método Ligação completa (Complete Linkage)

Já o método de Ligação completa, como o próprio Johnson (2007, p.685) diz, esse se assemelha ao anterior, em que a determinação da distância entre os grupos na matriz de distância agora é dada pela maior distância entre as observações.

$$d(G_1, G_2) = \max\{\, d(X_l, X_k, l \neq k,)\} \tag{25}$$

A distância entre dois grupos é definida pela maior distância entre qualquer par de observações, onde uma observação pertence ao primeiro grupo 1 (G1) e a outra pertence ao segundo grupo 2 (G2).

5.6 Método Centroid (Centróide)

Esse método, utiliza vetores de médias, chamados de centroides para calcular a distância entre os grupos, esses vetores são calculados como sendo: $\overline{X_1}$ o vetor de médias do grupo 1, n_1o número de observações do grupo 1 e X_i o vetor das variáveis do grupo.

$$G_1 = \overline{X_1} = \frac{1}{n_1}\sum_{i \in C_1} X_i \tag{26}$$

O mesmo se repete para o grupo 2

$$G_2 = \overline{X_2} = \frac{1}{n_2}\sum_{i \in C_2} X_i \tag{27}$$

Possuindo os centroides, a distância entre os grupos é definida pela seguinte formula:

$$d(G_1, G_2) = (\overline{X_1} - \overline{X_2})'(\overline{X_1} - \overline{X_2}) \tag{28}$$

Uma característica importante a se destacar deste método, que é citado e exemplificado por Mingot (2005, p.175), é o fato de que, diferentemente dos métodos de ligação simples e completa nos quais a distância entre os agrupamentos tende a aumentar a cada nova fusão, o método do centroide pode apresentar o comportamento oposto. Nesse caso, é possível observar uma redução na distância em que determinado agrupamento é realizado em relação a uma fusão anterior. Essa característica decorre da forma como o método calcula as distâncias com base na posição dos centroides dos grupos, o que pode gerar inversões no dendrograma, fenômeno típico desse tipo de ligação.

5.7 Método Average linkage (Média das distâncias)

Esse determina as distâncias entre grupos a partir da média aritmética de todas as distâncias entre os elementos dos dois grupos, ou seja, após o cálculo da

distância euclidiana entre todos os pares de elementos dos grupos é usada a média dessas distâncias como o valor a representar a distância desses grupos que estão sendo comparados. Dessa forma, a fórmula que define essas distâncias, é dada por:

$$d(G_1, G_2) = \sum_{l \in C_1} \sum_{k \in C_2} \left(\frac{1}{n_1 n_2}\right) d(X_l, X_k) \tag{29}$$

5.8 Método Elbow (Método do cotovelo)

Esse nos ajuda a encontrar um número ideal de grupos através da lógica de variância, a ideia do método é calcular a variabilidade dentro do grupo para que possamos dar uma sugestão do número de grupos através de uma análise gráfica. O total da soma dos quadrados dentro do cluster WSS (Within-Cluster Sum of Squares) mede a homogeneidade dos grupos criados pelo cluster.

Segundo Shi et al. (2021), o número ótimo de K clusters é determinado pelo fato de que, antes de atingir K grupos, o valor de WSS diminui rapidamente até o chamado valor de pico. Após exceder K, o WSS continua a aumentar, permanecendo o valor de pico praticamente inalterado, com um ponto de cotovelo explícito.

Sendo $X = \{x_1, x_2, \dots, x_N\}$ uma base de dados com N observações e K grupos e $G = \{G_1, G_2, \dots, G_K\}$ em que G_{i} representando um grupo individual.

$$\mathrm{WSS} = \sum_{k=1}^{K} \sum_{x_i \in C_k} (x_i - \mu_k)^2 \tag{30}$$

Para montagem do gráfico é observado o valor de WSS para cada número de K grupos.

5.9 Método da Silhueta

Ao contrário do método Elbow, este propõe utilizar a medida de dissimilaridade como ponto de vista para sugerir o número de agrupamentos e também avaliar a qualidade da classificação de cada observação da base de dados individualmente.

Proposto por ROUSSEEUW (1987, apud MACIEL,VINHAS e CÂMARA, 2015). Vale (2005), apresenta a seguinte lógica:

Para cada objeto i, pertencente ao grupo A, calcula-se inicialmente a dissimilaridade média dentro do próprio grupo:

$$a(i) = \frac{1}{|A| - 1} \sum_{j \in A,\, j \neq i} d(i,j) \tag{31}$$

Em seguida, para cada grupo $G \neq A$, calcula-se a dissimilaridade média de i em relação aos objetos desse outro grupo, expressa por:

$$d(i,G) = \frac{1}{|G|} \sum_{j \in G} d(i,j) \tag{32}$$

Após obter essas médias para todos os grupos diferentes de A, define-se a menor dessas distâncias médias, que corresponde ao grupo vizinho mais próximo de i:

$$b(i) = \min_{G \neq A} d\,(i,G) \tag{33}$$

Por fim, o coeficiente de Silhueta do objeto i é obtido pela relação entre a diferença das dissimilaridades e seu valor máximo, conforme:

$$s(i) = \frac{b(i) - a(i)}{\max\{\, a(i),\, b(i)\}} \tag{34}$$

O valor de $s(i)$ varia entre –1 e 1. Valores próximos de 1 indicam que o objeto está bem classificado no seu grupo, já valores próximos de -1 indicam o oposto, uma má classificação.

A partir desse coeficiente podemos interpretar 2 tipos de gráficos, o primeiro é o gráfico da silhueta média por número de grupos, no qual se calcula o valor médio de $s(i)$para cada possível valor de k. Ao observar graficamente esses valores, identifica-se o número de grupos que apresenta a maior média de silhueta, indicando a configuração com melhor separação entre grupos.

O segundo é o gráfico de silhueta individual, que permite avaliar de forma separada por observação o quão bem cada objeto foi alocado ao seu respectivo grupo.

6 Base de dados

A base de dados para fazer o modelo de regressão logística é composta por jogos do MSI (Mid-Season Invitational), que é um campeonato internacional de meio de temporada do jogo, disponibilizada por um usuário que coletou os dados da API do League of Legends, se encontra no site Kaggle, localizado no link :

https://www.kaggle.com/datasets/barthetur/lol-2021-competitive-game-from-012021-to-msi

De todos os dados que a base oferece, apenas as variáveis de interesse para o modelo foram selecionadas.

Quadro 1: variáveis da base de dados e sua descrição até os 15 minutos de jogo

Variável	Descrição
t1_result	Resultado da partida em vitória ou derrota para o time 1 j
t1_goldat15	Ouro total adquirido pelo time 1
t1_xpat15	Experiencia total adquirida
t1_csat15	Total de monstros abatidos pelo time 1
t1_golddiffat15	Diferença dotal de ouro entre os times
t1_xpdiffat15	Diferença total de experiência entre os times
t1_csdiffat15	Diferença do total de monstros abatidos
t1_killsat15	Total de abates realizado pelo time 1
t1_assistsat15	Total de assistências realizado pelo time 1
t1_deathsat15	Total de mortes sofridas pelo time 1
t1p1_goldat15	Ouro adquirido pelo jogador 1 do time 1
t1p1_xpat15	Experiencia adquirida pelo jogador 1 do time 1
t1p1_golddiffat15	Diferença de ouro ente o jogador 1 do time 1 e o jogador 1 do time 2
t1p1_xpdiffat15	Diferença de experiência ente o jogador 1 do time 1 e o jogador 1 do time 2
t1p1_killsat15	Abates realizados pelo jogador 1 do time 1
t1p1_assistsat15	Assistências realizadas pelo jogar 1 do time 1
t1p1_deathsat15	Número de mortes sofridas do jogador 1 do time 1
t1p1_csat15	Número de monstros abatidos do jogador 1 do time 1
t1p1_csdiffat15	Diferença do número de monstros abatidos do jogador 1 do time 1 e o jogador 1 do time 2
t1p2_goldat15	Ouro adquirido pelo jogador 2 do time 1
t1p2_xpat15	Experiencia adquirida pelo jogador 2 do time 1
t1p2_golddiffat15	Diferença de ouro ente o jogador 2 do time 1 e o jogador 2 do time 2
t1p2_xpdiffat15	Diferença de experiência ente o jogador 2 do time 1 e o jogador 2 do time 2
t1p2_killsat15	Abates realizados pelo jogador 2 do time 1
t1p2_assistsat15	Assistências realizadas pelo jogar 2 do time 1

t1p2_deathsat15	Número de mortes sofridas do jogador 2 do time 1
t1p2_csat15	Número de monstros abatidos do jogador 2 do time 1
t1p2_csdiffat15	Diferença do número de monstros abatidos do jogador 2 do time 1 e o jogador 2 do time 2
t1p3_goldat15	Ouro adquirido pelo jogador 3 do time 1
t1p3_xpat15	Experiencia adquirida pelo jogador 3 do time 1
t1p3_golddiffat15	Diferença de ouro ente o jogador 3 do time 1 e o jogador 3 do time 2
t1p3_xpdiffat15	Diferença de experiência ente o jogador 3 do time 1 e o jogador 3 do time 2
t1p3_killsat15	Abates realizados pelo jogador 3 do time 1
t1p3_assistsat15	Assistências realizadas pelo jogar 3 do time 1
t1p3_deathsat15	Número de mortes sofridas do jogador 3 do time 1
t1p3_csat15	Número de monstros abatidos do jogador 3 do time 1
t1p3_csdiffat15	Diferença do número de monstros abatidos do jogador 3 do time 1 e o jogador 3 do time 2
t1p4_goldat15	Ouro adquirido pelo jogador 4 do time 1
t1p4_xpat15	Experiencia adquirida pelo jogador 4 do time 1
t1p4_golddiffat15	Diferença de ouro ente o jogador 4 do time 1 e o jogador 4 do time 2
t1p4_xpdiffat15	Diferença de experiência ente o jogador 1 do time 1 e o jogador 4 do time 2
t1p4_killsat15	Abates realizados pelo jogador 4 do time 1
t1p4_assistsat15	Assistências realizadas pelo jogar 4 do time 1
t1p4_deathsat15	Número de mortes sofridas do jogador 4 do time 1
t1p4_csat15	Número de monstros abatidos do jogador 4 do time 1
t1p4_csdiffat15	Diferença do número de monstros abatidos do jogador 4 do time 1 e o jogador 4 do time 2
t1p5_goldat15	Ouro adquirido pelo jogador 5 do time 1
t1p5_xpat15	Experiencia adquirida pelo jogador 5 do time 1
t1p5_golddiffat15	Diferença de ouro ente o jogador 5 do time 1 e o jogador 5 do time 2
t1p5_xpdiffat15	Diferença de experiência ente o jogador 5 do time 1 e o jogador 5 do time 2
t1p5_killsat15	Abates realizados pelo jogador 5 do time 1
t1p5_assistsat15	Assistências realizadas pelo jogar 5 do time 1
t1p5_deathsat15	Número de mortes sofridas do jogador 5 do time 1
t1p5_csat15	Número de monstros abatidos do jogador 5 do time 1
t1p5_csdiffat15	Diferença do número de monstros abatidos do jogador 5 do time 1 e o jogador 5 do time 2

As mesmas descrições se repetem para as variáveis com terminação “10”, alterando apenas que essas se referem aos dados observados até os 10 minutos de jogo.

Para o presente trabalho é de extrema importância destacar que os termos p1, p2, p3, p4 e p5 se referem a posições específicas dos jogadores no mapa do jogo e são referenciadas da seguinte forma na base de dados:

- p1: se refere ao jogador posicionado na rota superior do mapa, chamada de Top lane.
- p2: se refere ao jogador posicionado na rota flexível do mapa, chamada de Jungle.
- p3: se refere ao jogador posicionado na rota do meio do mapa, chamada de Mid lane.
- p4: se refere ao jogador posicionado na rota inferior do mapa, chamada de Bot lane. Em específico, o jogador p4 recebe o nome de ADC que divide a sua rota com o Suporte.
- P5: se refere ao jogador posicionado na rota inferior do mapa, chamada de Bot lane. Em específico, o jogador p5 recebe o nome de Suporte que divide a sua rota com o ADC.

Ainda com relação às posições dos jogadores é preciso destacar que na mecânica do jogo as rotas compostas por p1, p2, p4 e p5 são comumente referenciadas como “Side lane”, esse termo se dá pela simetria do mapa, no qual a Top lane e Bot lane se encontra nas laterais.

Uma segunda base de dados também será usada para complementar as análises, essa com o propósito de adicionar mais informações e detalhes aos personagens que participam de cada jogo também será usado no método de agrupamento discutido anteriormente, disponibilizada no link:

https://leagueoflegends.fandom.com/wiki/List_of_champions/Base_statistics
Essa apresenta os 18 status básicos de cada personagem do jogo que são usados para definir tanto sua força, quando resistência e até mobilidade.

Sendo esses: HP, HP+, HP5 ,HP5+ ,MP, MP+, MP5, MP5+, AD, AD+, AS, AS+, AR, AR+, MR, MR+, MS, Range.

Uma forma de entender um pouco dessas características funcionam é que elas definem personagens com papeis diferentes entro do jogo, um exemplo disso seria

dizer que se um personagem A tem muito de um status X, logo ele não terá tanto de outro status Y.

7 Resultados

Com cunho de conhecimento e entendimento da base para cumprimento dos objetivos desse trabalho, as primeiras análises são feitas para entender como a análise deve seguir.

7.1 Análise exploratória de dados

Como a intenção desta análise é chegar ao modelo de regressão logística, vamos começar a entender cada variável e suas correlações entre elas para que lá na frente seja possível interpretar quais variáveis possam ser significativas no modelo e evitar a correlação entre elas, o que causaria um possível VIF alto.

Como primeiro passo, vamos entender como cada variável está na base de dados e como ela pode ser tratada dentro da análise. Com exceção das variáveis “t1_result” que é binária (0 ou 1), todas as outras variáveis entram como números inteiros, dessas, a grande diferença são suas escalas e consequentemente suas variâncias, sendo assim vamos dividir as variáveis de baixa escala da de grande escala.

Como forma de facilitar as explicações, nesse tópico a variável “t1_result”, por ser a variável de interesse, será chamada de “y”.

Quadro 2: Tabela de variáveis com escala em milhar

Variável	Média	Desvio padrão	Variável	Média	Desvio padrão
t1_goldat15	25068	1879,9045	t1p3_goldat15	5469	590,2842
t1_xpat15	29647	1547,0083	t1p3_xpat15	7374	567,8686
t1_golddiffat15	278	3105,8363	t1p3_golddiffat15	29	933,8216
t1_xpdiffat15	31,21	2251,1246	t1p3_xpdiffat15	-11	767,2729
t1p1_goldat15	5204	716,3967	t1p4_goldat15	5508	788,9488
t1p1_xpat15	7094	633,6657	t1p4_xpat15	5440	569,1946
t1p1_golddiffat15	49	1174,1738	t1p4_golddiffat15	42	1245,9951
t1p1_xpdiffat15	-1,379	866,8755	t1p4_xpdiffat15	-18	790,6985
t1p2_goldat15	5354	641,7751	t1p5_goldat15	3533	482,6960
t1p2_xpat15	5918	695,4615	t1p5_xpat15	3821	510,5981
t1p2_golddiffat15	101	973,2336	t1p5_golddiffat15	56	651,9139
t1p2_xpdiffat15	38.7	1049,3284	t1p5_xpdiffat15	23	633,5277
t1_goldat10	15820	1011,7437	t1p3_goldat10	3415	332,7590
t1_xpat10	18486	922,7936	t1p3_xpat10	4616	339,9323
t1_golddiffat10	115	1561,7318	t1p3_golddiffat10	11	495,0849
t1_xpdiffat10	38	1208,0637	t1p3_xpdiffat10	-5	438,4490
t1p1_goldat10	3251	383,8977	t1p4_goldat10	3393	407,6353
t1p1_xpat10	4492	390,0480	t1p4_xpat10	3208	329,4740
t1p1_golddiffat10	25	621,1334	t1p4_golddiffat10	13	629,2800
t1p1_xpdiffat10	-3	528,1918	t1p4_xpdiffat10	-5	458,6075
t1p2_goldat10	3468	401,0864	t1p5_goldat10	2293	302,0575
t1p2_xpat10	3691	450,8763	t1p5_xpat10	2480	329,6890
t1p2_golddiffat10	43	584,5004	t1p5_golddiffat10	23	404,6122
t1p2_xpdiffat10	42	653,4018	t1p5_xpdiffat10	9	416,5830

Quadro 3: Tabela de Tabela de variáveis com escala em centenas

Variável	Média	Desvio padrão	Variável	Média	Desvio padrão
t1_csat15	509,6	39,1389	t1_csdiffat15	-0,03	42,4759
t1p1_csat15	118,9	17,2635	t1p3_csdiffat15	-0,3	18,3484
t1p1_csdiffat15	0,2	23,3023	t1p4_csat15	125,8	27,7141
t1p2_csat15	104,2	12,8724	t1p4_csdiffat15	-1,2	35,1695
t1p2_csdiffat15	-0,1	18,9027	t1p5_csat15	25,6	18,5009
t1p3_csat15	134,9	15,9350	t1p5_csdiffat15	1,5	24,7432
t1_csat10	320.9	25.1602	t1_csdiffat10	0,7	27.5087
t1p1_csat10	74.4	11.4728	t1p3_csdiffat10	-0,2	11.8221
t1p1_csdiffat10	0,3	16.0444	t1p4_csat10	77,9	17.9067
t1p2_csat10	68,6	8.9475	t1p4_csdiffat10	-0,6	22.9208
t1p2_csdiffat10	0,4	12.4819	t1p5_csat10	16,1	11.9652
t1p3_csat10	83,7	10.0941	t1p5_csdiffat10	0,9	16.1622

Quadro 4: Tabela de variáveis com escala em unidades

Variável	Média	Desvio padrão	Variável	Média	Desvio padrão
t1_killsat15	4.34	3.0902	t1p3_killsat15	0.91	1.1511
t1_assistsat15	7.26	5.6011	t1p3_assistsat15	1.25	1.3679
t1_deathsat15	4.16	2.9529	t1p3_deathsat15	0.72	0.9059
t1p1_killsat15	0.77	1.0383	t1p4_killsat15	0.99	1.2437
t1p1_assistsat15	1.02	1.2266	t1p4_assistsat15	1.19	1.3921
t1p1_deathsat15	0.90	1.0103	t1p4_deathsat15	0.71	0.9233
t1p2_killsat15	1.26	1.3721	t1p5_killsat15	0.38	0.6869
t1p2_assistsat15	1.69	1.5745	t1p5_assistsat15	2.11	1.9192
t1p2_deathsat15	0.81	0.9352	t1p5_deathsat15	1.00	1.0422
t1_killsat10	2.35	2.0143	t1p3_killsat10	0.47	0.7840
t1_assistsat10	3.63	3.6087	t1p3_assistsat10	0.59	0.8918
t1_deathsat10	2.24	1.9310	t1p3_deathsat10	0.36	0.6304
t1p1_killsat10	0.40	0.6780	t1p4_killsat10	0.54	0.8490
t1p1_assistsat10	0.50	0.8319	t1p4_assistsat10	0.60	0.9546
t1p1_deathsat10	0.50	0.7236	t1p4_deathsat10	0.37	0.6354
t1p2_killsat10	0.72	0.9785	t1p5_killsat10	0.23	0.5002
t1p2_assistsat10	0.87	1.0766	t1p5_assistsat10	1.07	1.2912
t1p2_deathsat10	0.44	0.6641	t1p5_deathsat10	0.57	0.7626

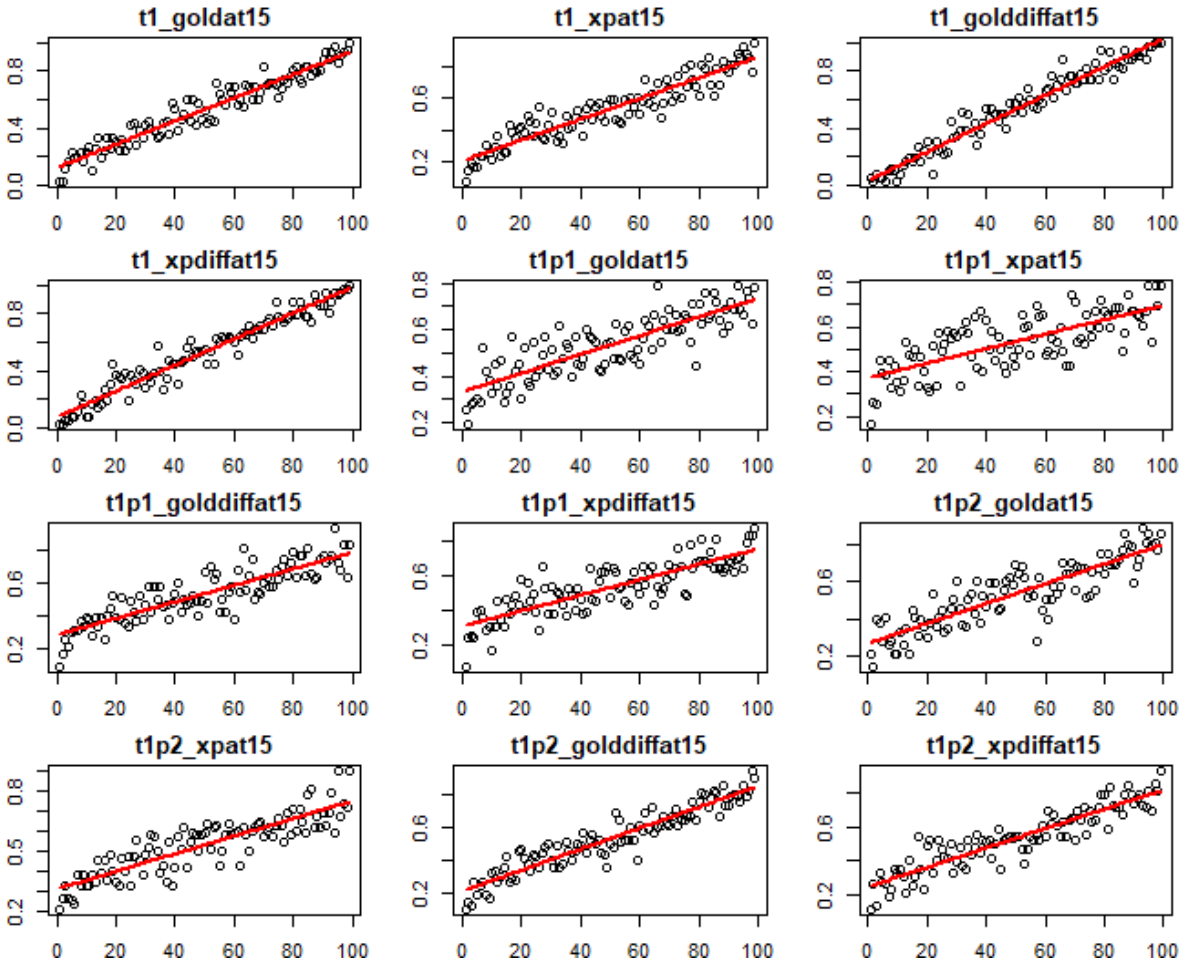
Visto as diferenças entre as variáveis, fica mais fácil para saber interpretar e analisar.

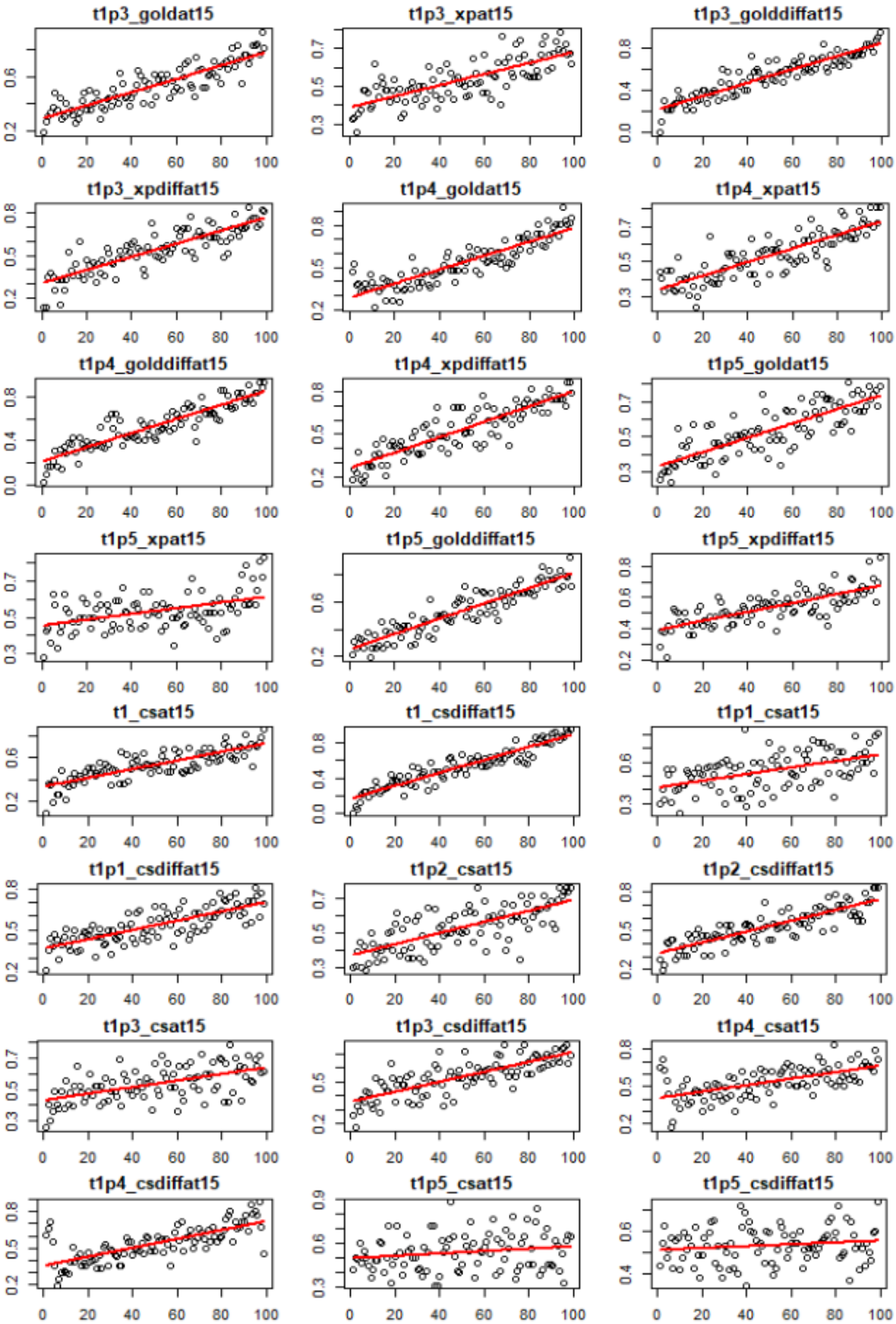
Em função de montar o modelo logístico, vamos verificar se existe relação linear entre as variáveis explicativas e a variável dependente, assim como também verificar se há correlação entre as variáveis explicativas, para que os pressupostos do modelo sejam satisfeitos.

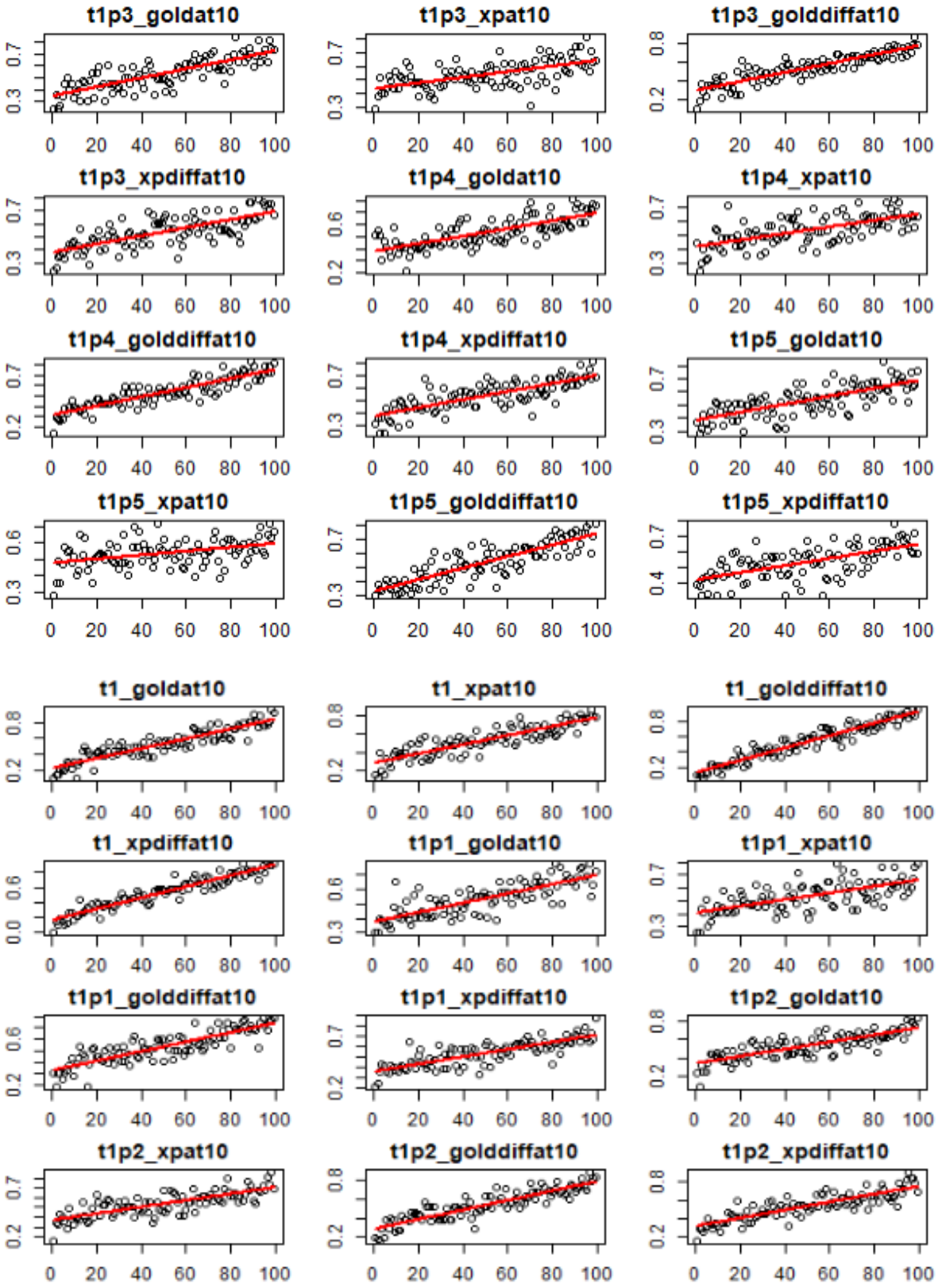
Primeiramente, uma forma de verificar visualmente se há relação entre linear é fazendo o gráfico da concentração de eventos de sucesso por percentil de cada variável explicativa, como esse tipo de gráfico envolve o percentil para visualização, esse é mais adequado para as variáveis que possuem escala em milhar e centena.

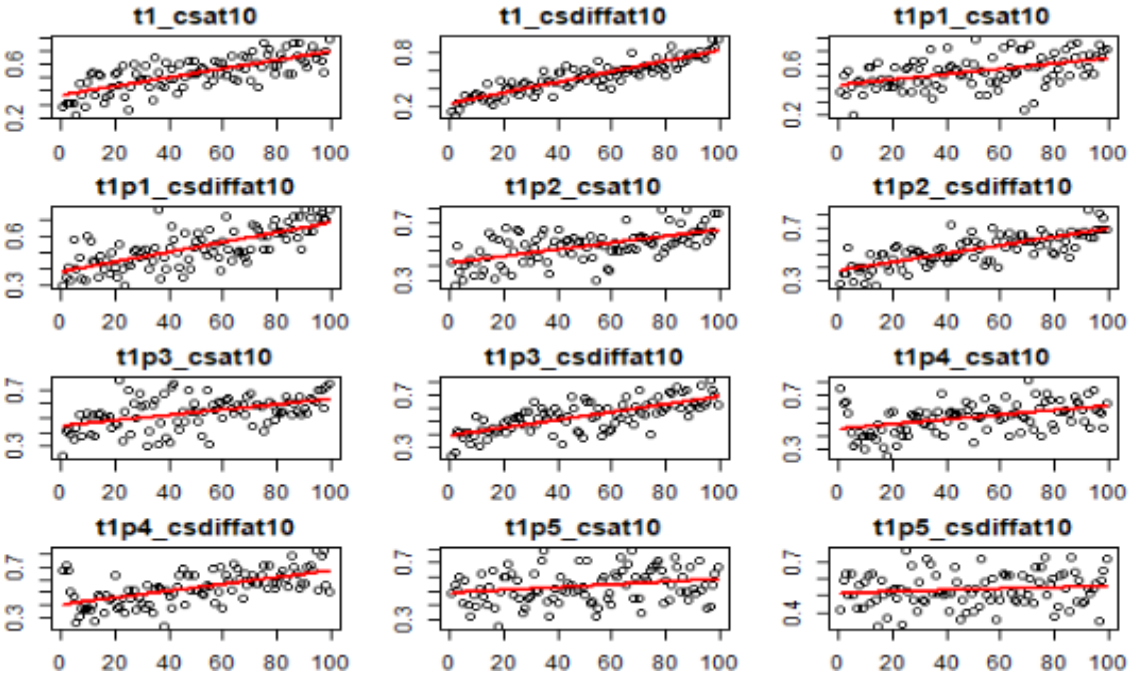
Lembrando que y se refere a “t1_result”.

Figura 1: Gráfico do percentil das variáveis independentes com escala de milhar e centena vs a média do y




t1p3_goldat15
t1p3_xpat15
t1p3_golddiffat15
t1p3_xpdiffat15
t1p4_goldat15
t1p4_xpat15
t1p4_golddiffat15
t1p4_xpdiffat15
t1p5_goldat15
t1p5_xpat15
t1p5_golddiffat15
t1p5_xpdiffat15
t1_csat15
t1_csdiffat15
t1p1_csat15
t1p1_csdiffat15
t1p2_csat15
t1p2_csdiffat15
t1p3_csat15
t1p3_csdiffat15
t1p4_csat15
t1p4_csdiffat15
t1p5_csat15
t1p5_csdiffat15

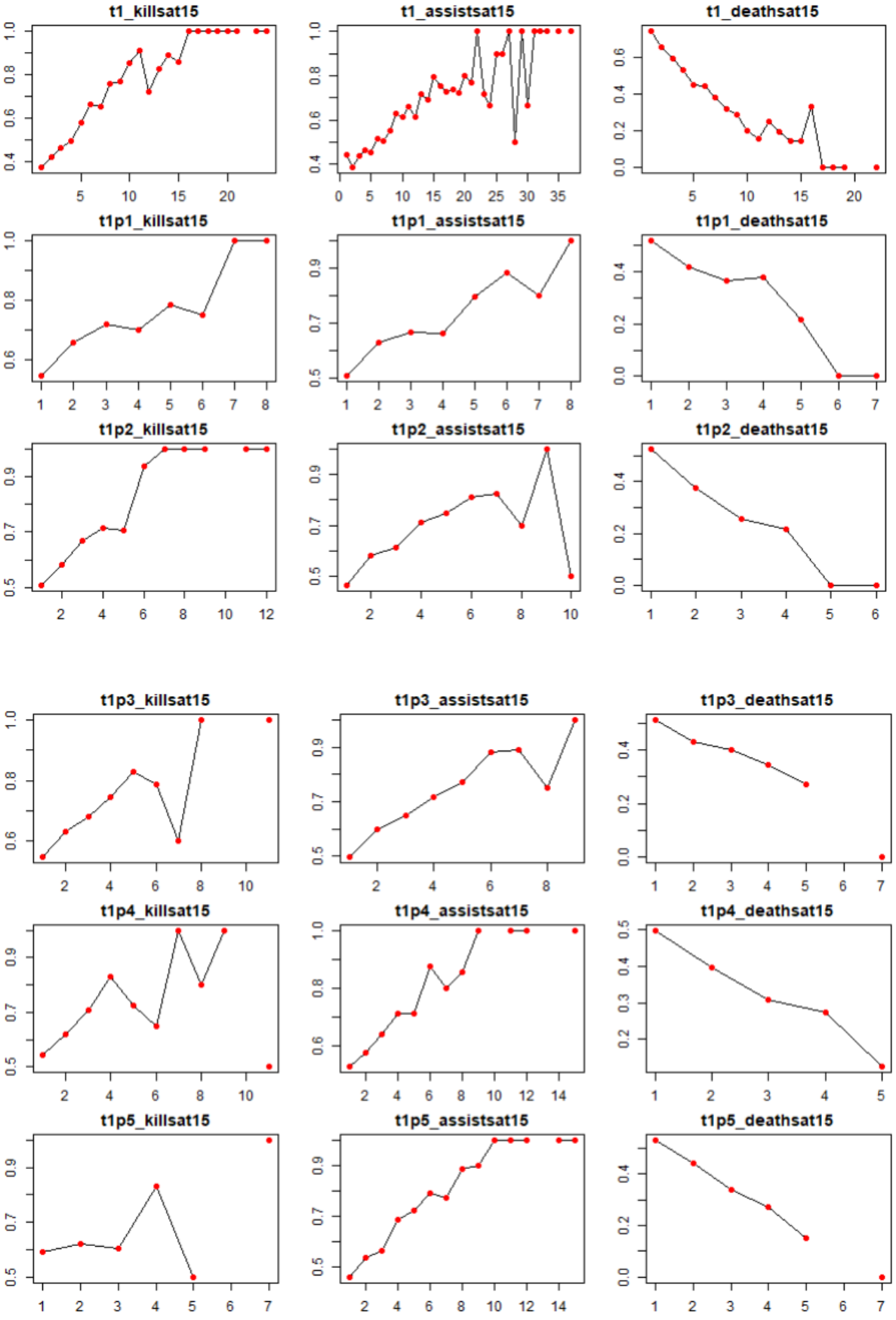

t1p3_goldat10
t1p3_xpat10
t1p3_golddiffat10
t1p3_xpdiffat10
t1p4_goldat10
t1p4_xpat10
t1p4_golddiffat10
t1p4_xpdiffat10
t1p5_goldat10
t1p5_xpat10
t1p5_golddiffat10
t1p5_xpdiffat10
t1_goldat10
t1_xpat10
t1_golddiffat10
t1_xpdiffat10
t1p1_goldat10
t1p1_xpat10
t1p1_golddiffat10
t1p1_xpdiffat10
t1p2_goldat10
t1p2_xpat10
t1p2_golddiffat10
t1p2_xpdiffat10

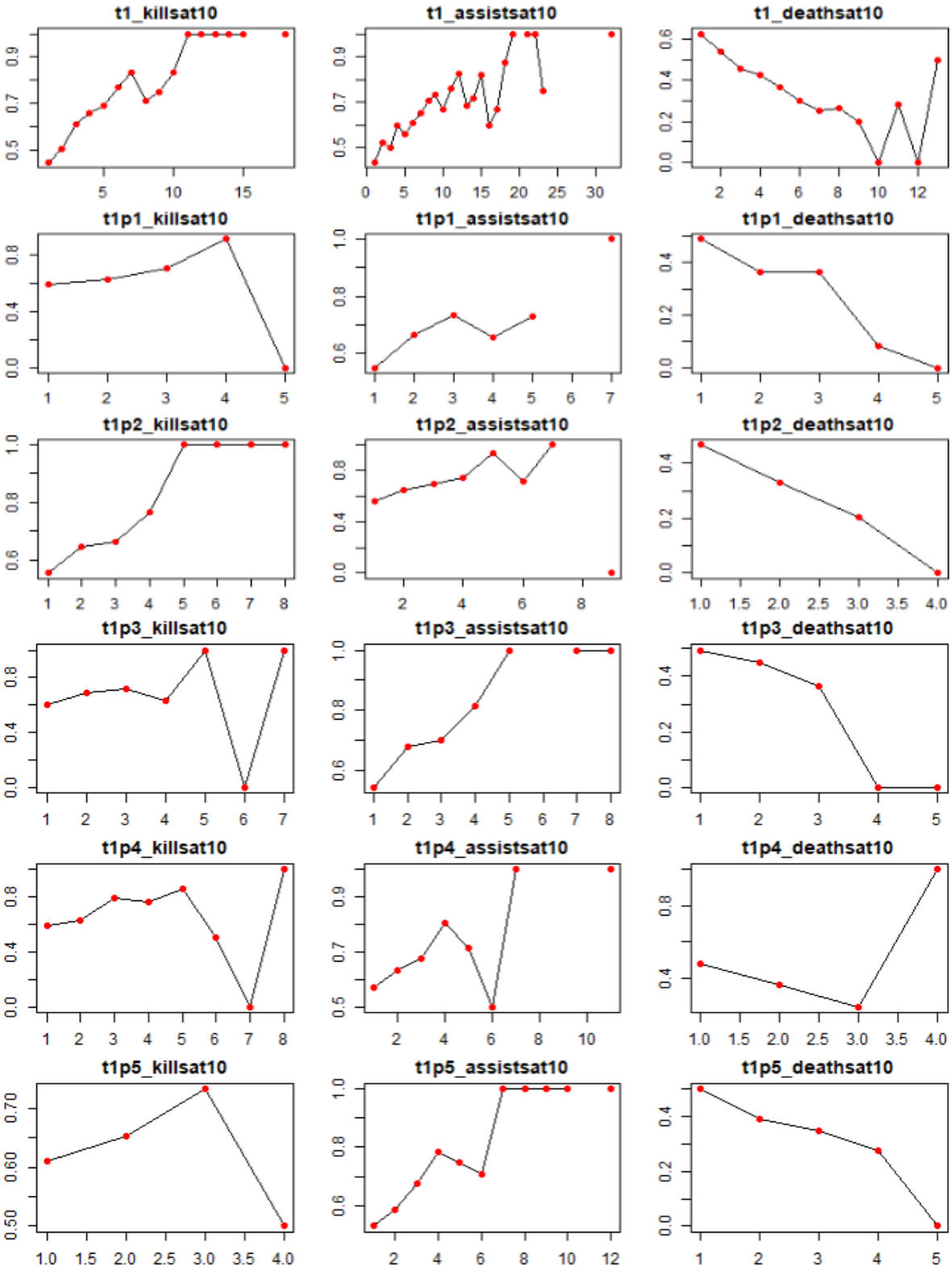


Fica evidente que a maioria dos dados possuem uma relação linear bem visual com o a proporção da variável y, mas variáveis como “t1p5_csat15”, “t1p5_csat10”, “t1p5_csdiffat15” e “t1p5_csdiffat10” não ficam conclusivas.

Passando para as variáveis com escala em unidades ou centenas vamos fazer o gráfico da média de y para cada ponto da escala das variáveis.

Figura 2: Média de y vs variáveis de escala em unidade ou centena





Assim como visto na figura 1, pode-se perceber que a maioria das variáveis com escala em unidades ou dezenas também visualmente apresentam uma relação linear com o y, com exceção das variáveis “t1p5_killsat15”, “t1p3_killsat10”, “t1p4_assistat10”, “t1p4_deathsat10” aparentam não possuir uma relação. Mais

adiante na análise, é interessante revisitar esses gráficos para entender quais variáveis podem ou não ajudar na previsão do modelo.

Agora que destacado a existência de correlação entre a maioria das variáveis explicativas com a variável de interesse, vamos verificar se há multicolinearidade entre as variáveis explicativas.

Para isso é feita a matriz de correlação entre as variáveis e ver quantas possuem correlação baixa, média e forte, como são muitas variáveis, vamos nos ater somente ao número de correlações fortes para ter uma ideia de o quanto isso pode impactar no modelo.

Quadro 5: Número de correlações e sua intensidade entre as variáveis independentes para os dados de até 10 minutos de jogo.

Intensidade da correlação	Correlação	Frequência
Baixa	Correlação < 0.5	2620
Média	0.5 < Correlação < 0.7	168
Alta	Correlação > 0.7	106

Quadro 6: Número de correlações e sua intensidade entre as variáveis independentes para os dados de até 15 minutos de jogo.

Intensidade da correlação	Correlação	Frequência
Baixa	Correlação < 0.5	2568
Média	0.5 < Correlação < 0.7	220
Alta	Correlação > 0.7	128

Dado os quadros 5 e 6, vemos que há um número de correlações altas significativo, dito isso, durante a análise, deve-se atentar uma possível multicolinearidade das variáveis.

7.2 Análise de agrupamento

Para realizar o agrupamento, visto que existe uma diferença na escala entre as variáveis, a padronização dos dados foi feita antes dos cálculos de clustering, definida por: Sendo x o vetor das variáveis aleatórias com $i = 1,2,\dots,18$ e s_i o desvio padrão da variável xi.

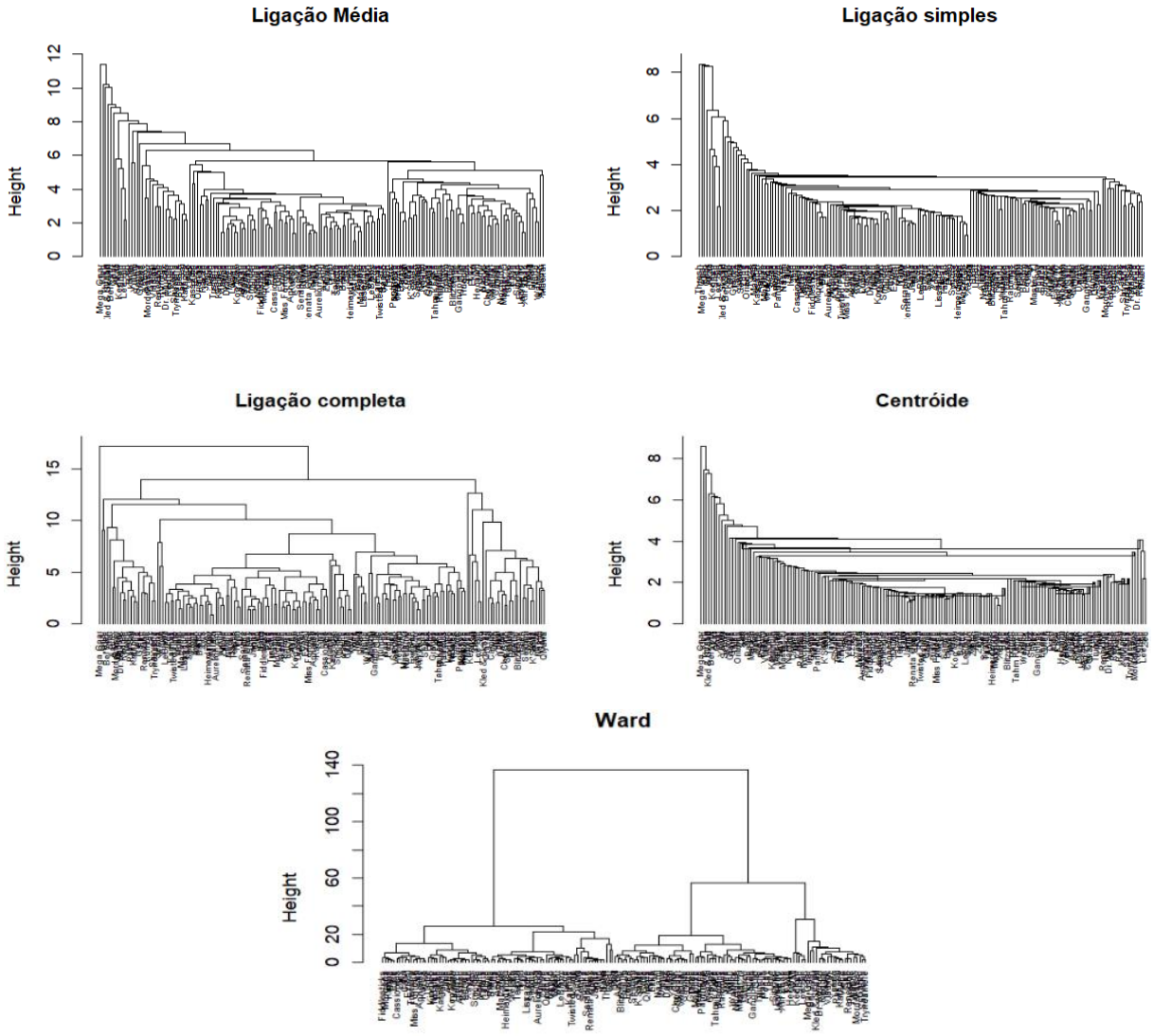
$$\frac{x_i - \overline{x_i}}{s_i} \tag{33}$$

Dado que existem 169 personagens no jogo, a matriz de distância calculada é de ordem 169x169.

Para encontrar o melhor agrupamento esses dados em questão, serão testadas e comparadas as metodologias de ligação média, ligação simples, ligação completa, centroide e ward.

Ao aplicar os diferentes métodos e realizar o Dendrograma temos:

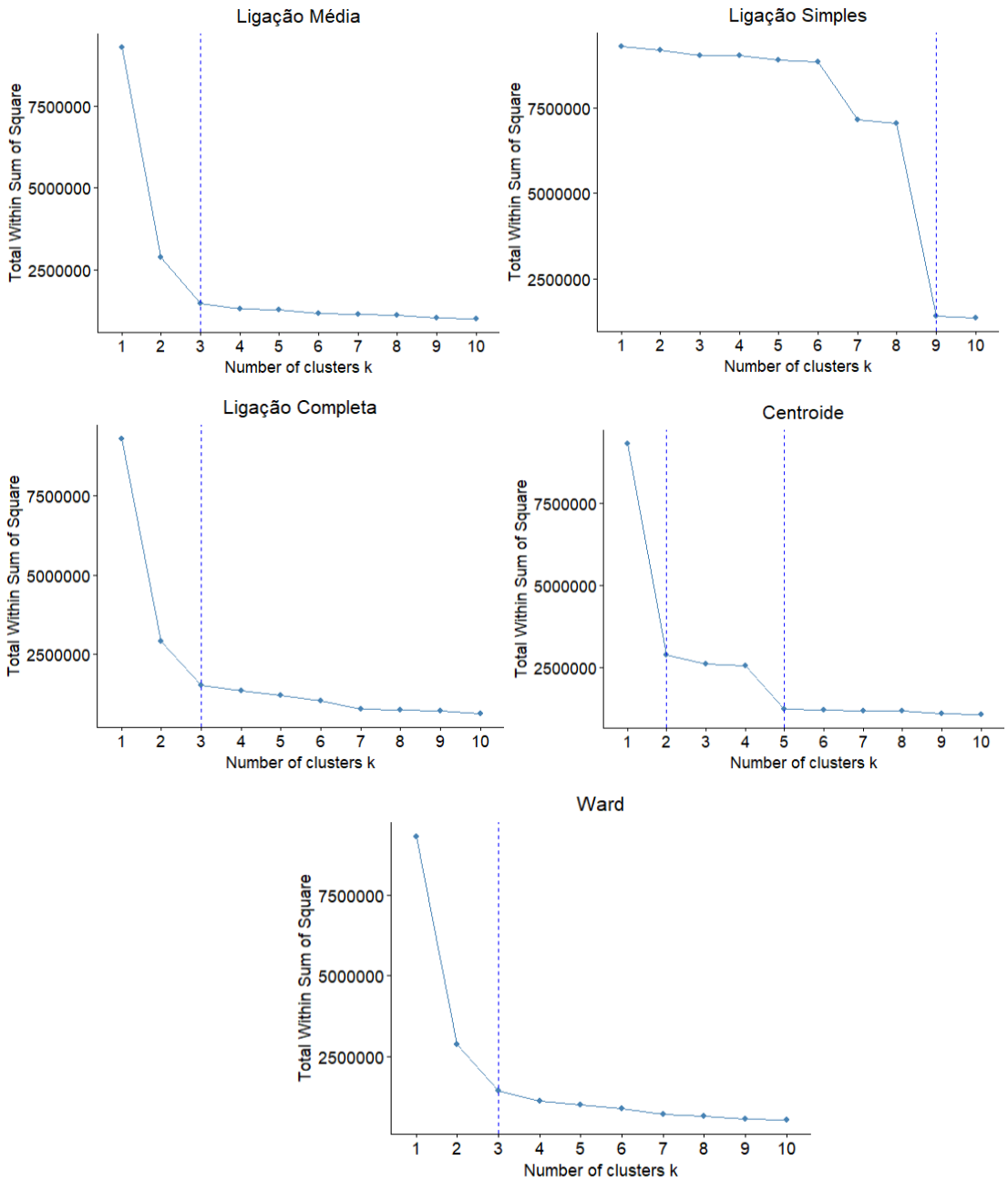
Figura 3: Dendrograma



Ao observar os dendrogramas percebe-se que a maioria dos métodos, com exceção do ward, não apresentou um aumento significativo na distância de agrupamento (eixo height dos gráficos) ao passo que foram realizadas as fusões, destacando o dendrograma do método ward por apresentar linhas compridas (em relação ao eixo height) ao passo que as fusões foram feitas. Para melhor

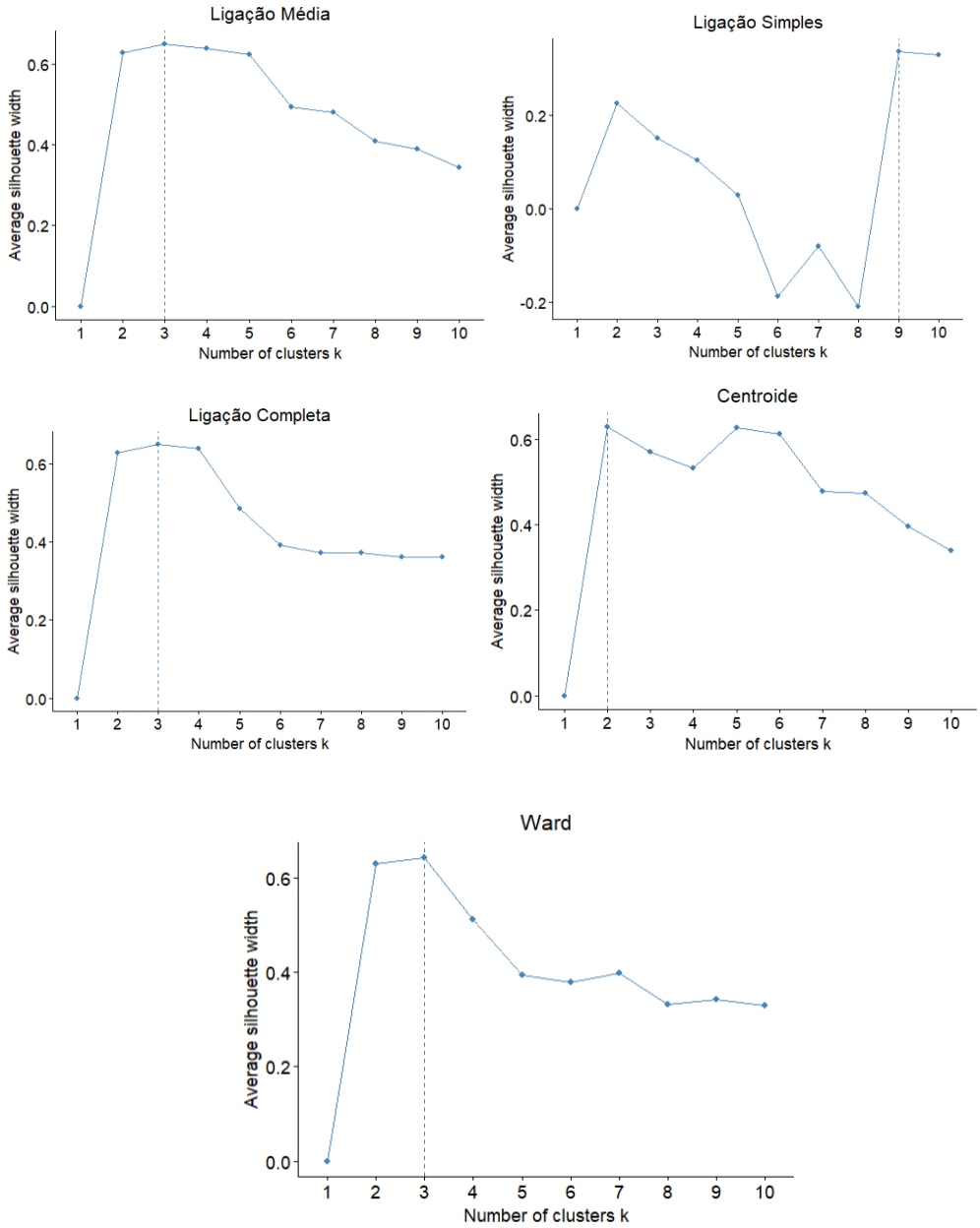
entendimento do desempenho de cada método, será aplicado os gráficos Elbow e Silhouette para verificar a sugestão de número de grupos que cada ligação sugere.

Figura 4: Gráfico Elbow



Pelo método, as ligações média, completa e Ward sugeriram 3 grupos, já a ligação simples sugeriu 9 grupos e o centroide sugeriu 2 ou 5 grupos (valores que podem ser entendidos como o ponto de cotovelo)..

Figura 5: Gráfico da silhueta média por número de grupos

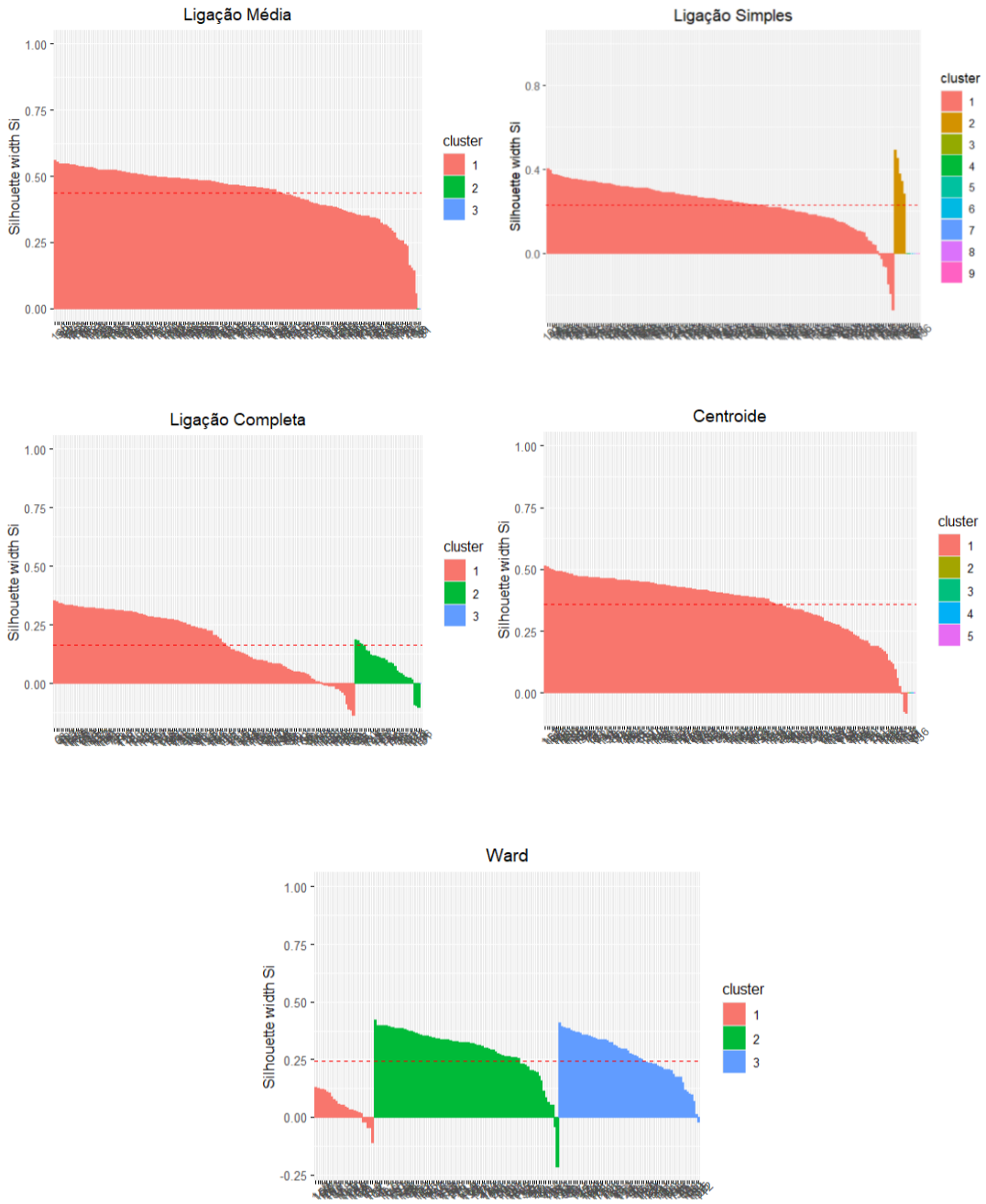


De maneira semelhante ao gráfico Elbow, os métodos de ligação Média, Completa e Ward indicaram a formação de três grupos com base na Silhueta. Já a Ligação Simples sugeriu nove grupos, enquanto o método do Centroide apontou para dois grupos. Vale ressaltar que a Ligação Simples apresentou uma silhueta média máxima relativamente baixa, próxima de 0,2, enquanto as outras se mantiveram mais altas, próximas de 0,6, mostrando melhor desempenho.

Para avaliar a qualidade da classificação proposta por cada técnica, é possível examinar os valores individuais do coeficiente de Silhueta de forma gráfica. Essa

visualização permite verificar o quão bem cada observação foi alocada ao seu respectivo grupo.

Figura 6: gráfico de silhueta individual



Fica evidente que, com exceção do Ward, todos os métodos tendem a agrupar a maioria das observações em um único grupo, além disso o Ward apresentou poucas observações com coeficiente silhueta menor ou igual a zero, o que indica que a grande maioria das observações foi bem classificada.

Desta forma, o agrupamento realizado pelo método Ward com 3 grupos se mostra mais adequado e coerente para o presente trabalho.

Assumindo 3 grupos é possível reduzir o número de combinações diferentes de equipes possíveis de 2.081.693.722.421.538.086.400 para 59.049

7.3 Modelo 10

Será chamado de modelo 10 o modelo de Regressão Logística com o objetivo de prever o resultado da partida utilizando os dados coletados até os 10 iniciais da partida.

Para o segundo modelo proposto, foram adotadas as seguintes estratégias para a aplicação do modelo::

- Como o objetivo do modelo é ser interpretável, a variável t1_goldat10 foi dividida por 1000, pois por apresentar valores altos, na interpretação de parâmetros a leitura dele pode ser feita para cada 1000 de ouro a mais ou a menos da equipe, visto que dividir uma variável por uma constante não afetará na sua seleção ou não pelos p-valores
- Remoção das variáveis com o nome “diff” ao meio: Essas variáveis como por exemplo “t1_golddiffat10”, “t1_xpdiffat10”, “t1_csdiffat10” entre outras não são possíveis de serem coletadas imediatamente durante uma partida, pois essas se referem a diferença de valores observados entre o time 1 e o time 2, como o objetivo do modelo é ser capaz de aplicá-lo em jogos futuros, fora da base de teste, então essas variáveis não entrarão no modelo por sua indisponibilidade na aplicação prática.
- Criação de novas variáveis a fim de simplificar o modelo e evitar multicolinearidade, visto que ainda existem variáveis que se referem a uma mesma medida do jogo, mas para referenciais diferentes. Por exemplo, “t1_goldat10” se refere ao Gold adquirido pelo Time 1 todo até os 10 minutos. De forma semelhante, o “t1p1_goldat10”, “t1p2_goldat10”, “t1p3_goldat10”, “t1p4_goldat10” e “t1p5_goldat10” também se referem ao Gold adquirido, mas individualmente para cada jogador (Player 1, Player 2, Player 3, Player 4 e Player 5, respectivamente). Dessa forma, essas variáveis trazem informações muito semelhantes e correlacionadas que podem gerar problemas ao analisar o VIF posteriormente. Com o objetivo de solucionar isso**, foram criadas** as seguintes variáveis, conforme tabela abaixo:

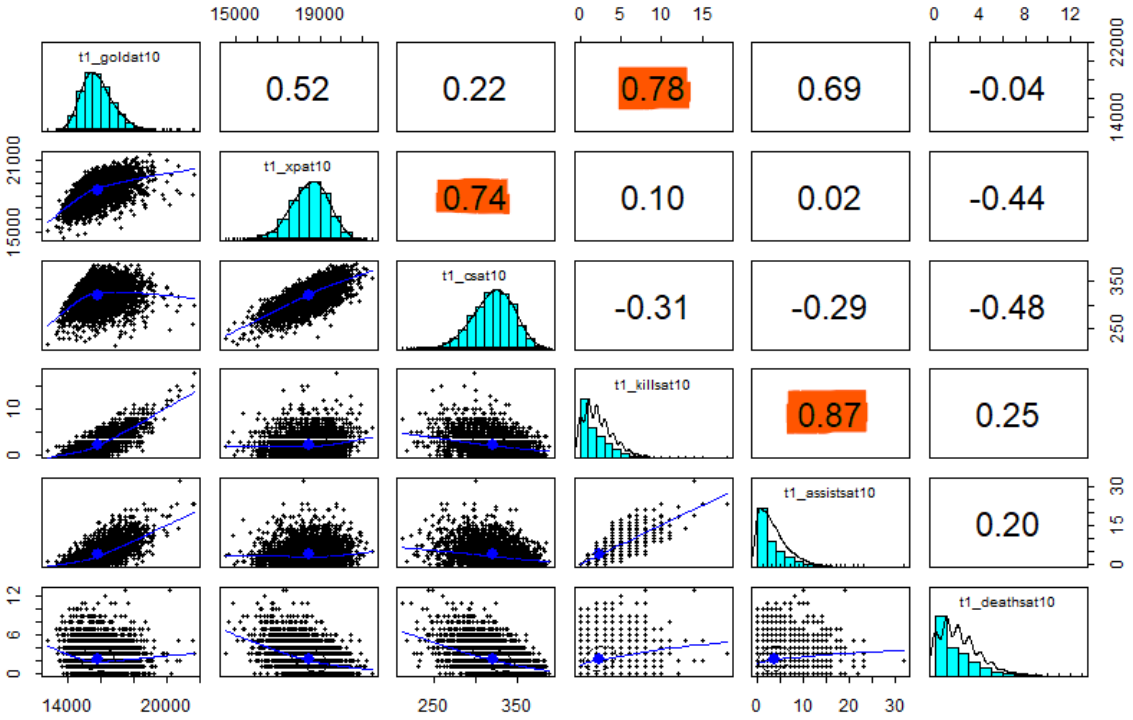
Quadro 8: Variáveis criadas para aplicação no modelo 10

Variável criada	Tipo de variável	Descrição	Níveis
T1_kills_os_Player	Categórica	T1 kills out sized player: Se refere ao jogador com maior número de kills do time 1.	p1, p2, p3, p4, p5 tie_kills (quanto há empate), z_kills (quando ninguém apresenta valor em kills)
T1_gold_os_Player	Categórica	T1 gold out sized player: Se refere ao jogador com maior número de gold do time 1.	p1, p2, p3, p4, p5
T1_xp_os_Player	Categórica	T1 xp out sized player: Se refere ao jogador com maior número de xp do time 1.	p1, p2, p3, p4, p5
T1_cs_os_Player	Categórica	T1 cs out sized player: Se refere ao jogador com maior número de cs do time 1.	p1, p2, p3, p4, p5
T1_assists_os_Player	Categórica	T1 assist out sized player: Se refere ao jogador com maior número de assistências do time 1.	p1, p2, p3, p4, p5 tie_assist (quanto há empate), z_death (quando ninguém apresenta valor em assist)
T1_deaths_os_Player	Categórica	T1 deaths out sized player: Se refere ao jogador com maior número de deaths do time 1.	p1, p2, p3, p4, p5, tie_death (quanto há empate), z_death (quando ninguém apresenta valor em death)

Com essas novas variáveis criadas a partir da avaliação individual de cada jogador em cada métrica, pode-se utilizá-las no lugar das variáveis numéricas originais que causariam alta multicolinearidade.

Ainda a fim de tratar a alta correlação entre variáveis independentes, observa-se** a correlação entre as variáveis t1_killsat10, t1_goldat10, t1_xpat10, t1_csat10, t1_assistat10 e t1_deathsat10.

Figura 7: Gráfico da correlação entre as variáveis



A princípio fica evidente 3 correlações fortes descritas entre as variáveis

- “t1_goldat10” e “t1_killsat10” com correlação forte de 0,78
- “t1_xpat10” e “t1csat10” com correlação forte de 0,74
- “t1_killsat10” e “t1_assistat10” com correlação forte de 0,87

Para evitar multicolinearidade é removido as variáveis “t1_assistat10” e “t1_xpat10” do treinamento do modelo, mantendo a “t1_killsat10” e monitorando os valores de VIF ao final da modelagem.

Feito essa etapa inicial de tratamento de dados, começa a estimativa dos parâmetros na base de treino.

Quadro 9: Proporção de vitórias nas bases de treino e teste

Base	Proporção
Treino	0.5381955
Teste	0.5301887

Com proporções semelhantes podemos seguir com o treinamento do modelo.

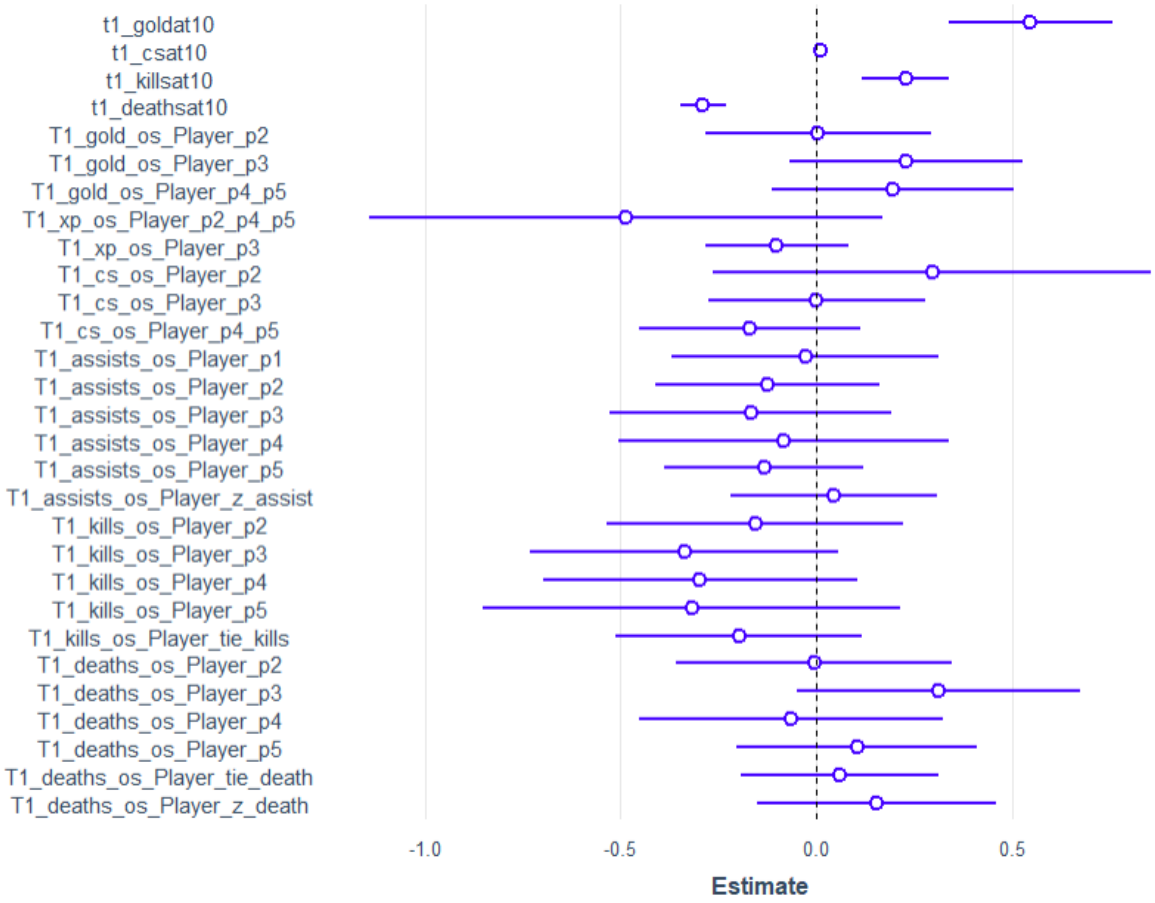
Os valores marcados com (***) são os mais significativos ao modelo, os marcados com (**) são muito significativos ao modelo, os com (*) são apenas

significativos, por fim os marcados com (°) tem o menor nível de significância para o modelo.

Quadro 10: Primeira estimação dos parâmetros do modelo 10

Variável	Estimate	Std. Error	z value	Pr(>	
(Intercept)	-11.668414	1.173805	-9.941	< 2e-16	***
t1_goldat10	0.546911	0.106500	5.135	0.000000282	***
t1_csat10	0.010854	0.002813	3.858	0.000114	***
t1_killsat10	0.228071	0.056781	4.017	0.000059029	***
t1_deathsat10	-0.289176	0.030465	-9.492	< 2e-16	***
T1_gold_os_Player_p2	0.004912	0.147869	0.033	0.973500	
T1_gold_os_Player_p3	0.228572	0.151180	1.512	0.130555	
T1_gold_os_Player_p4_p5	0.195001	0.157383	1.239	0.215337	
T1_xp_os_Player_p2_p4_p5	-0.485416	0.333936	-1.454	0.146052	
T1_xp_os_Player_p3	0.110253	0.093115	1.087	0.276953	
T1_cs_os_Player_p2	0.295801	0.284943	1.038	0.299220	
T1_cs_os_Player_p3	0.001027	0.140827	0.007	0.994183	
T1_cs_os_Player_p4_p5	-0.171288	0.143796	-1.191	0.233580	
T1_assists_os_Player_p1	-0.028663	0.174638	-0.164	0.869629	
T1_assists_os_Player_p2	-0.125034	0.146551	-0.853	0.393652	
T1_assists_os_Player_p3	-0.167983	0.183857	-0.914	0.360895	
T1_assists_os_Player_p4	-0.083560	0.215837	-0.387	0.698650	
T1_assists_os_Player_p5	-0.133731	0.129797	-1.030	0.302865	
T1_assists_os_Player_z_assist	0.044774	0.134937	0.332	0.740027	
T1_kills_os_Player_p2	-0.156729	0.192811	-0.813	0.416296	
T1_kills_os_Player_p3	-0.337721	0.201026	-1.680	0.092959	°
T1_kills_os_Player_p4	-0.297291	0.205061	-1.450	0.147124	
T1_kills_os_Player_p5	-0.318866	0.271970	-1.172	0.241024	
T1_kills_os_Player_tie_kills	-0.197826	0.160646	-1.231	0.218157	
T1_deaths_os_Player_p2	-0.005724	0.179091	-0.032	0.974504	
T1_deaths_os_Player_p3	0.311613	0.183329	1.692	0.090692	°
T1_deaths_os_Player_p4	0.065747	0.197544	0.333	0.739269	
T1_deaths_os_Player_p5	-0.103268	0.157131	-0.657	0.511047	

Figura 8: Gráfico dos intervalos de confiança dos parâmetros para primeira estimação do modelo 10.



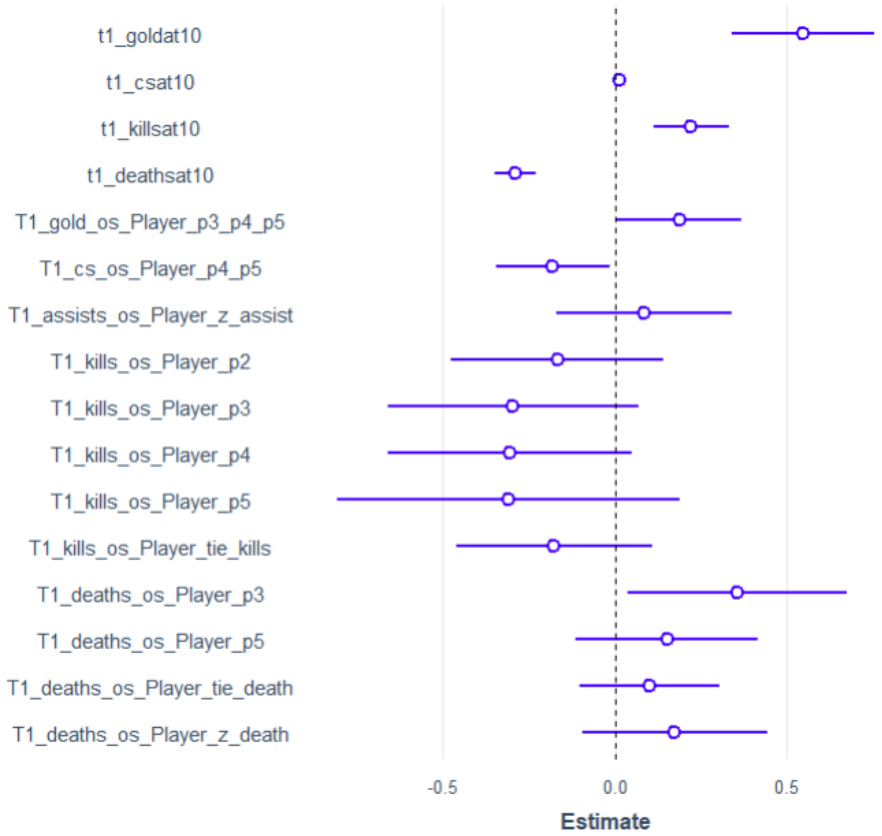
Para simplificar o modelo e aumentar sua parcimônia, realizou-se o agrupamento (pooling) de categorias de variáveis que se comportam de maneira estatisticamente semelhante e que fazem sentido com o contexto do jogo. Este agrupamento foi guiado pela sobreposição dos Intervalos de Confiança (ICs) dos respectivos parâmetros, indicando a falta de diferença significativa em relação à categoria de referência.

- T1_cs_os_Player: Agrupamento dos níveis p1, p2, p3 e em outro grupo o p4 com p5.
- T1_assists_os_Player: Agrupamento dos níveis p1, p2, p3, p4, p5 e tie_assist.
- T1_deaths_os_Player: Agrupamento dos níveis p1, p2, p4.
- T1_gold_os_Player: Agrupamento dos níveis p1 e p2 em um grupo e p3, p4 e p5 outro grupo.
- Remoção da variável xp

Quadro 11: Segunda estimação dos parâmetros do modelo 10

Variável	Estimate	Std. Error	z value	Pr(>	
(Intercept)	-1.17578	114.124	-10.301	< 2e-16	***
t1_goldat10	0.54658	0.10506	5.203	1.96e-07	***
t1_csat10	0.01076	0.00279	3.857	0.000115	***
t1_killsat10	0.22094	0.05584	3.957	0.0000760	***
T1_deathsat10	-0.28977	0.03036	-9.545	< 2e-16	***
T1_gold_os_Player_p3_p4_p5	0.18589	0.09167	2.028	0.042581	*
T1_cs_os_Player_p4_p5	-0.18055	0.08503	-2.123	0.033716	*
T1_assists_os_Player_z_assist	0.08242	0.12996	-0.634	0.525945	
T1_kills_os_Player_p2	-0.16800	0.15779	-1.065	0.286794	
T1_kills_os_Player_p3	-0.29513	0.18640	-1.589	0.112128	
T1_kills_os_Player_p4	-0.30641	0.18073	-1.695	0.090000	°
T1_kills_os_Player_p5	-0.30964	0.25472	-1.216	0.224125	
T1_kills_os_Player_tie_kills	-0.17707	0.14513	-1.218	0.223801	
T1_deaths_os_Player_p3	0.35534	0.15532	2.287	0.022979	*
T1_deaths_os_Player_p5	0.21007	0.13579	1.100	0.269108	
T1_deaths_os_Player_tie_death	0.09911	0.10314	0.961	0.336614	
T1_deaths_os_Player_z_death	0.17325	0.13711	1.264	0.206398	

Figura 9: Gráfico dos intervalos de confiança dos parâmetros para segunda estimação do modelo 10.



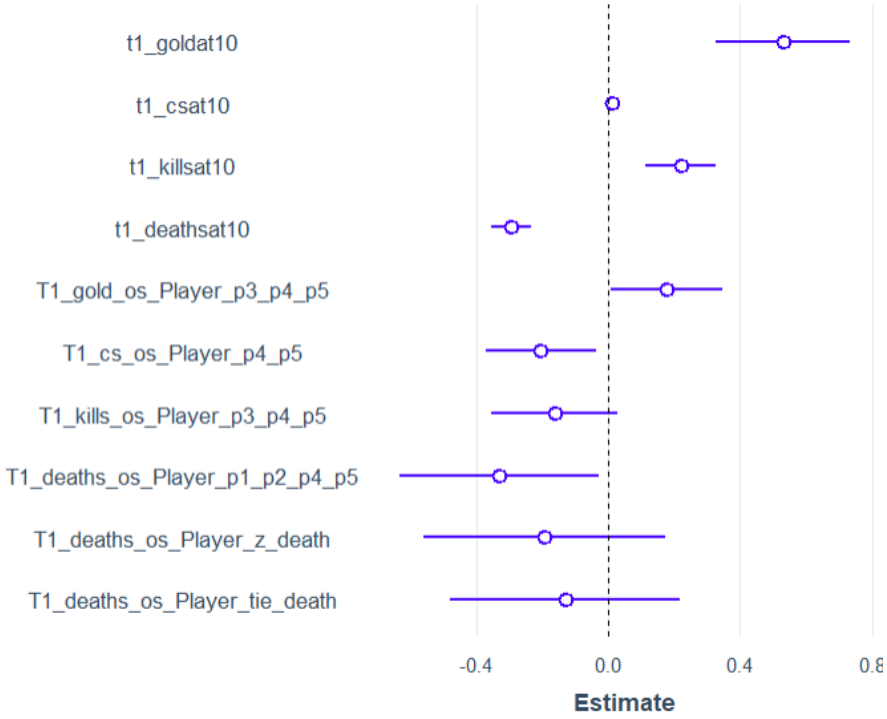
Ainda com diversas dummys não significativas é repetido o processo de agrupamento.

- T1_kills_os_Player: Agrupamento dos níveis p1, p2 e tie_kills em umgrupo e p3, p4 e p5 em outro.
- T1_deaths_os_Player: Agrupamento dos níveis p1, p2, p4, p5.
- Remoção da variável T1_assists_os_Player.

Quadro 12: Terceira estimação dos parâmetros do modelo 10

Variável	Estimate	Std. Error	z value	Pr(>	
(Intercept)	-11.363.409	1.134.525	-10.016	< 2e-16	***
t1_goldat10	0.530491	0.104255	5.088	0.000000361	***
t1_csat10	0.011101	0.002773	4.004	0.000062400	***
t1_killsat10	0.220986	0.054791	4.033	0.000055003	***
t1_deathsat10	-0.293008	0.030256	-9.684	< 2e-16	***
T1_gold_os_Player_p3_p4_p5	0.176990	0.086159	2.054	0.0400	*
T1_cs_os_Player_p4_p5	-0.203334	0.085079	-2.390	0.0169	*
T1_kills_os_Player_p3_p4_p5	-0.162118	0.097850	-1.657	0.0976	°
T1_deaths_os_Player_p1_p2_p4_p5	-0.330446	0.154295	-2.142	0.0322	*
T1_deaths_os_Player_z_death	-0.191844	0.186977	-1.026	0.3049	
T1_deaths_os_Player_tie_death	-0.130463	0.177637	-0.734	0.4627	

Figura 10: Gráfico dos intervalos de confiança dos parâmetros para terceira estimação do modelo 10.



Percebe-se que a variável T1_deaths_os_Player apresenta apenas uma dummy significativa, de forma que se torna interessante realizar o agrupamento dos níveis p1, p2, p4, p5 em um grupo e p3, tie_death e z_death em outro grupo.

Quadro 12: Quarta estimação dos parâmetros do modelo 10

Variável	Estimate	Std. Error	z value	Pr(>	
(Intercept)	-11.678417	1.126513	-10.367	< 2e-16	***
t1_goldat10	0.527021	0.104160	5.060	0.000000420	***
t1_csat10	0.011149	0.002772	4.022	0.0000578	***
t1_killsat10	0.222248	0.054771	4.058	0.0000495	***
t1_deathsat10	-0.284812	0.026045	-10.935	< 2e-16	***
T1_gold_os_Player_p3_p4_p5	0.173814	0.086082	2.019	0.0435	*
T1_cs_os_Player_p4_p5	-0.195539	0.084555	-2.313	0.0207	*
T1_kills_os_Player_p3_p4_p5	-0.160873	0.097818	-1.645	0.1000	
T1_deaths_os_Player_p3_tie_z	0.204218	0.083562	2.444	0.0145	*

Com apenas a variável dummy T1_kills_os_Player_p3_p4_p5 não significativa e sem possibilidade de agrupá-la em outro nível, essa é removida do modelo.

Quadro 13: Quinta estimação dos parâmetros do modelo 10

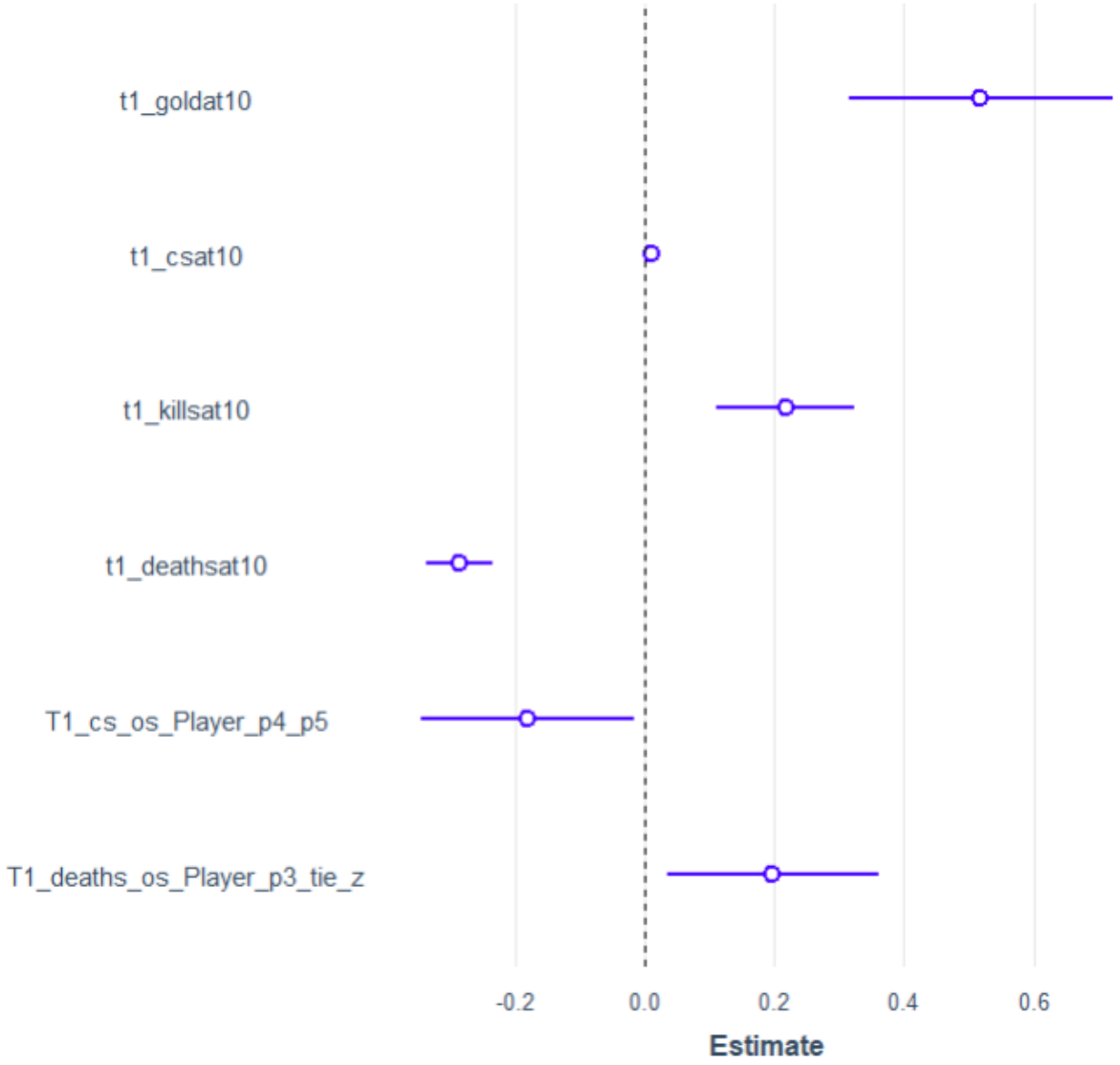
Variável	Estimate	Std. Error	z value	Pr(>	
(Intercept)	-11.631552	1.125548	-10.334	< 2e-16	***
t1_goldat10	0.523587	0.104095	5.030	0.000000491	***
t1_csat10	0.011165	0.002771	4.029	0.0000560	***
t1_killsat10	0.217018	0.054670	3.970	0.0000720	***
t1_deathsat10	-0.285690	0.026026	-10.977	< 2e-16	***
T1_gold_os_Player_p3_p4_p5	0.124347	0.080557	1.544	0.1227	
T1_cs_os_Player_p4_p5	-0.193151	0.084506	-2.286	0.0223	*
T1_deaths_os_Player_p3_tie_z	0.201876	0.083528	2.417	0.0157	*

Nessa etapa fica evidente a importância desse passo a passo mais lento de remoção de variáveis, sendo que ao excluir T1_kills_os_Player do modelo a T1_gold_os_Player_p3_p4_p5 passa a não ser significativa, de forma que agora também pode ser removida simplificando o modelo final.

Quadro 14: Quinta estimação dos parâmetros do modelo 10

Variável	Estimate	Std. Error	z value	Pr(>	
(Intercept)	-11.482365	1.121126	-10.242	< 2e-16	***
t1_goldat10	0.517864	0.104079	4.976	0.000000650	***
t1_csat10	0.011190	0.002769	4.041	0.0000532	***
t1_killsat10	0.216749	0.054676	3.964	0.0000736	***
t1_deathsat10	-0.286499	0.026001	-11.019	< 2e-16	***
T1_cs_os_Player_p4_p5	-0.181523	0.084143	-2.157	0.0310	*
T1_deaths_os_Player_p3_tie_z	0.196932	0.083426	2.361	0.0182	*

Figura 11: Gráfico dos intervalos de confiança dos parâmetros para quinta estimação do modelo 10.



Agora, com o modelo final com todas as variáveis com significância ao nível de 0,05, verifica-se a presença de multicolinearidade antes de se analisar as métricas de ajuste do modelo.

Quadro 15: Valores de VIF do modelo 10.

Variável	VIF	Intensidade da multicolinearidade
t1_goldat10	4.871421	Moderada
t1_csat10	2.757018	Moderada
t1_killsat10	6.093409	De moderada para alta
t1_deathsat10	1.344707	Moderada
T1_cs_os_Player_p4_p5	1.015817	Moderada
T1_deaths_os_Player_p3_tie_z	1.057527	Moderada

Com apenas uma variável com VIF entre 5 e 10 (t1_killsat10), dá-se seguimento à avaliação das medidas de ajuste do modelo e posteriormente à interpretação das Odds Ratio a fim de ver se o modelo está coerente.

7.4 Avaliação do modelo 10

Quadro 16: Valores reais versus preditos.

	0 observado	1 observado
Previsto 0	331	182
Previsto 1	165	382

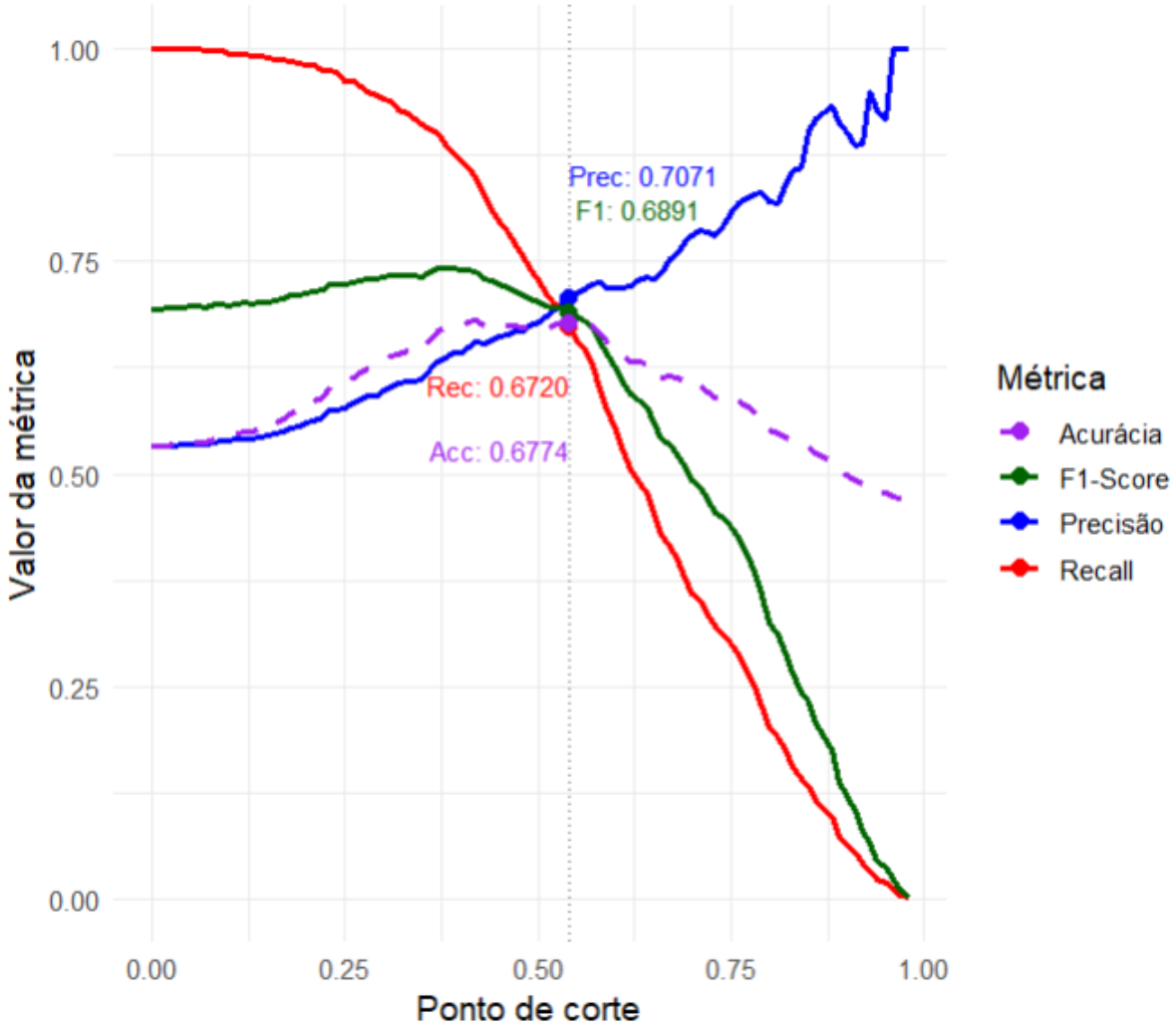
Quadro 17: Medidas de qualidade de ajuste na base de teste.

Acurácia	0,6726415
Precisão	0,6983547
Sensibilidade (Recall)	0,6773050
F1 Score	0,6876688

- Acurácia de 67,26%: O modelo acerta a classificação geral cerca de 67% dos casos na base de teste.
- F1-Score de 0,6877: Esse diz respeito ao equilíbrio entre a Precisão e a Sensibilidade (Recall), ou seja, ele não está dando forte preferência a uma única medida, mas sim buscando um meio termo entre as duas.
- Precisão de 69,83%: Quando o modelo diz que o time 1 vai ganhar (previsto 1), ele está certo aproximadamente de 70% das vezes.
- Sensibilidade (Recall) de 67,73%: Diz que o modelo foi capaz de capturar aproximadamente 68% de todos os casos positivos que realmente ocorreram.

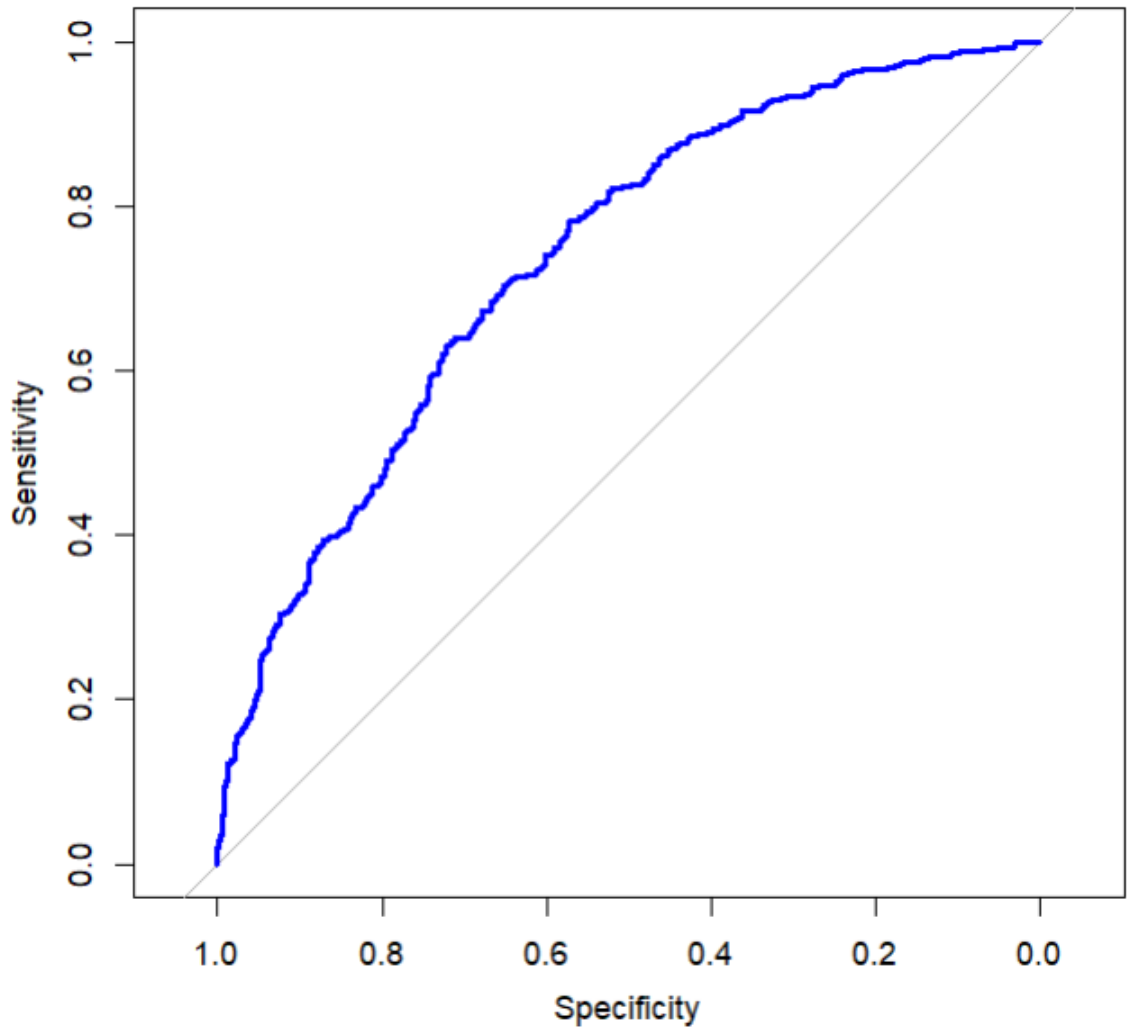
Apesar de assumir 0,54 como corte para previsão do modelo, é possível observar o comportamento de cada medida para diferentes pontos de corte utilizados.

Figura 12: Gráfico da Precisão, Recall, F1 Score, Acurácia por ponto de corte do modelo 10



A taxa de resposta da base se torna um ponto de corte favorável por trazer um equilíbrio entre as medidas de avaliação.

Figura 13: Curva ROC do modelo 10



O modelo de Regressão Logística obteve um AUC (Área Sob a Curva ROC) de 0,7357. Sendo 0,5 o indicativo de um classificador aleatório e 1,0 a discriminação perfeita, o valor alcançado situa o modelo na faixa de discriminação razoável ou aceitável.

A Curva ROC encontra-se moderadamente acima da linha diagonal, confirmando uma qualidade moderada do ajuste, visto que quanto mais abaulada e afastada a curva estiver dessa linha de referência, melhor é o desempenho do modelo.

7.5 Modelo 15

Será chamado de Modelo 15 o modelo de Regressão Logística com o objetivo de prever o resultado da partida utilizando os dados coletados até os 15 minutos iniciais da partida. Para esse modelo, foram tomadas as mesmas precauções e manipulações de dados que o Modelo 10, para haver também o efeito comparativo entre os modelos.

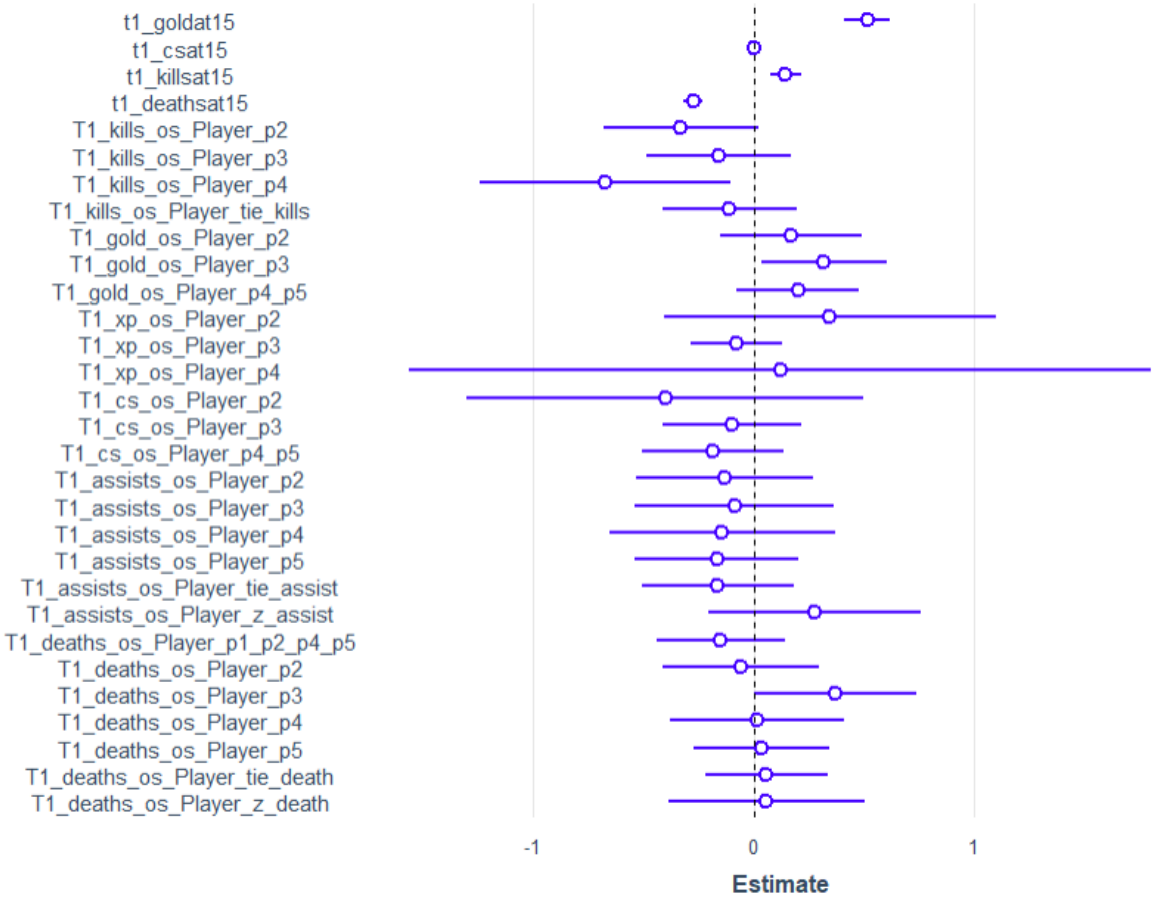
Para o modelo que busca prever o resultado do jogo aos 15 minutos, será seguido os mesmos passos do modelo anterior (que prevê o resultado do jogo aos 10 minutos), de forma que será avaliado quais são os possíveis agrupamentos dos níveis das variáveis categóricas mais lógicos para o contexto do jogo.

Primeira estimação dos parâmetros do modelo 15

Variável	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-13.317501	1.077979	-12.354	< 2e-16	***
t1_goldat15	0.513693	0.053986	9.515	< 2e-16	***
t1_csat15	0.002804	0.001780	1.575	0.1152	
t1_killsat15	0.144638	0.044423	3.199	0.000826	***
t1_deathsat15	-0.275397	0.021317	-12.919	< 2e-16	***
T1_kills_os_Player_p2	-0.331289	0.180193	-1.839	0.0660	°
T1_kills_os_Player_p3	-0.156916	0.167123	-0.939	0.3478	*
T1_kills_os_Player_p4	-0.675955	0.290898	-2.324	0.0201	
T1_kills_os_Player_tie_kills	-0.109629	0.155669	-0.704	0.4813	
T1_gold_os_Player_p2	0.187840	0.163887	1.147	0.2516	*
T1_gold_os_Player_p3	0.316731	0.145200	2.181	0.0292	
T1_gold_os_Player_p4_p5	0.200509	0.141755	1.414	0.1572	
T1_xp_os_Player_p2	0.134877	0.303082	0.896	0.3704	
T1_xp_os_Player_p3	-0.081143	0.106172	-0.764	0.4447	
T1_xp_os_Player_p4	0.119720	0.858628	0.139	0.8891	
T1_cs_os_Player_p2	-0.402388	0.380583	-1.058	0.2898	
T1_cs_os_Player_p3	-0.098866	0.161838	-0.613	0.5401	
T1_cs_os_Player_p4_p5	-0.183907	0.163378	-1.126	0.2603	
T1_assists_os_Player_p2	-0.131922	0.204844	-0.644	0.5196	
T1_assists_os_Player_p3	-0.086921	0.230008	-0.378	0.7056	
T1_assists_os_Player_p4_p5	-0.142970	0.260336	-0.549	0.5829	
T1_assists_os_Player_tie_assist	-0.163920	0.174993	-0.937	0.3489	
T1_assists_os_Player_z_assist	-0.148392	0.245427	-0.604	0.5450	
T1_deaths_os_Player_z_death	-0.150968	0.178434	-1.015	0.3101	
T1_deaths_os_Player_p2	-0.061040	0.181660	-0.336	0.7369	
T1_deaths_os_Player_p3	-0.128877	0.066703	-1.927	0.0486	*
T1_deaths_os_Player_p4	0.012775	0.063000	0.203	0.9495	
T1_deaths_os_Player_p5	0.033755	0.157647	0.214	0.8305	

T1_deaths_os_Player_tie_death	0.056949	0.147149	0.387	0.6989	

Figura 14: Gráfico dos intervalos de confiança dos parâmetros para primeira estimação do modelo 15.



Com a maioria das variáveis não significativas, principalmente as variáveis dummies, é possível planejar agrupamentos entre os níveis das variáveis de forma que ajudem no desempenho e interpretação do modelo.

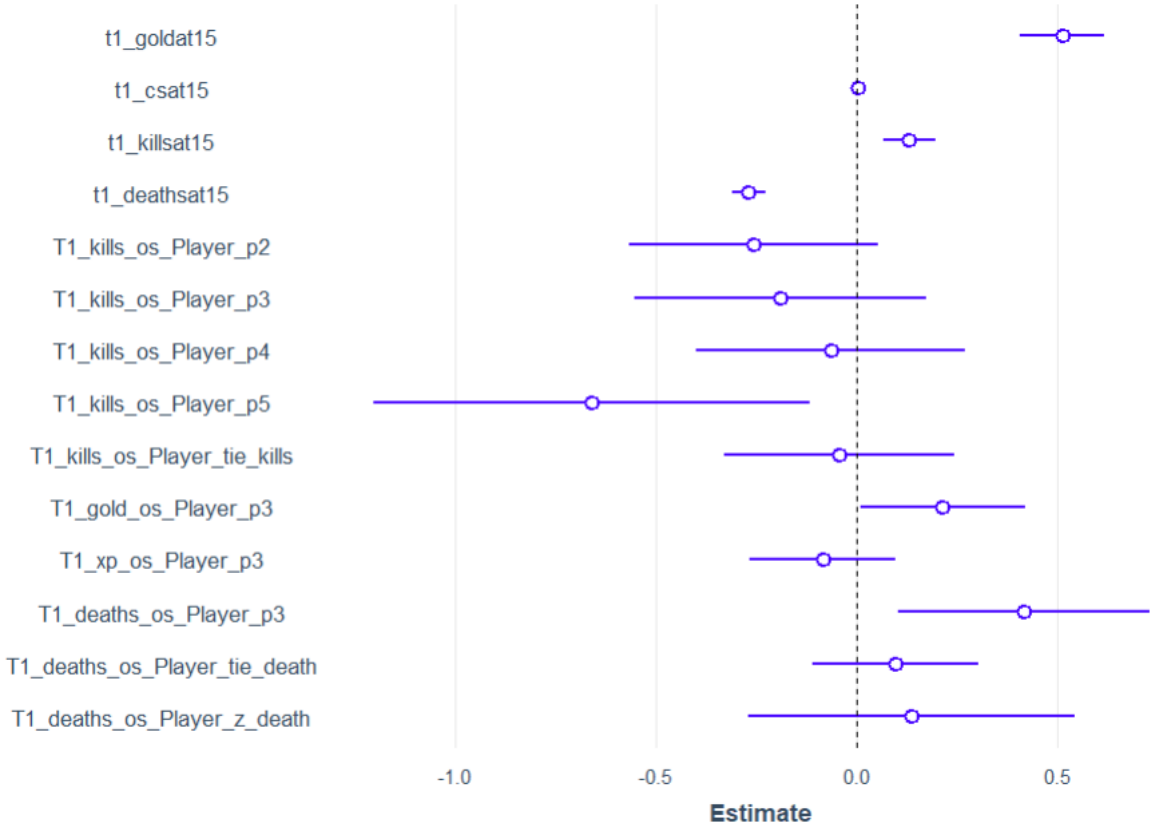
A partir desta análise, são propostos os seguintes agrupamentos para as variáveis categóricas de ***out sized*** de jogadores (identificados como 'os_Player'):

- T1_xp_os_Player: Agrupamento dos níveis p1, p2, p4 e p5.
- T1_deaths_os_Player: Agrupamento dos níveis p1, p2, p4 e p5.
- T1_gold_os_Player: Agrupamento dos níveis p1, p2, p4 e p5.
- Remoção da variável T1_cs_os_Player
- Remoção da variável T1_assist_os_Player

Quadro 19: Segunda estimação dos parâmetros do modelo 15

Variável	Estimate	Std. Error	z value	Pr(>z)	
(Intercept)	-13.527572	1.021998	-13.236	< 2e-16	***
t1_goldat15	0.513016	0.053256	9.633	< 2e-16	***
t1_csat15	0.002917	0.001767	1.650	0.09891	.
t1_killsat15	0.131981	0.033154	3.981	0.0000687	***
t1_deathsat15	-0.268598	0.020878	-12.865	< 2e-16	***
T1_kills_os_Player_p2	-0.256642	0.158021	-1.624	0.10436	
T1_kills_os_Player_p3	-0.189622	0.185425	-1.023	0.30648	
T1_kills_os_Player_p4	-0.063475	0.170793	-0.372	0.71015	
T1_kills_os_Player_p5	-0.659995	0.277778	-2.376	0.01750	*
T1_kills_os_Player_tie_kills	-0.042066	0.146814	-0.287	0.77447	
T1_gold_os_Player_p3	0.215453	0.104570	2.060	0.03936	*
T1_xp_os_Player_p3	-0.083965	0.093194	-0.901	0.36761	
T1_deaths_os_Player_p3	0.419070	0.159954	2.620	0.00879	**
T1_deaths_os_Player_tie_death	0.098995	0.105543	0.938	0.34827	
T1_deaths_os_Player_z_death	0.137494	0.207027	0.664	0.50661	

Figura 15: Gráfico dos intervalos de confiança dos parâmetros para segunda estimação do modelo 15.



Ainda com resultado não satisfatório (presença de variáveis não significativas), repete-se o passo com os seguintes agrupamentos:

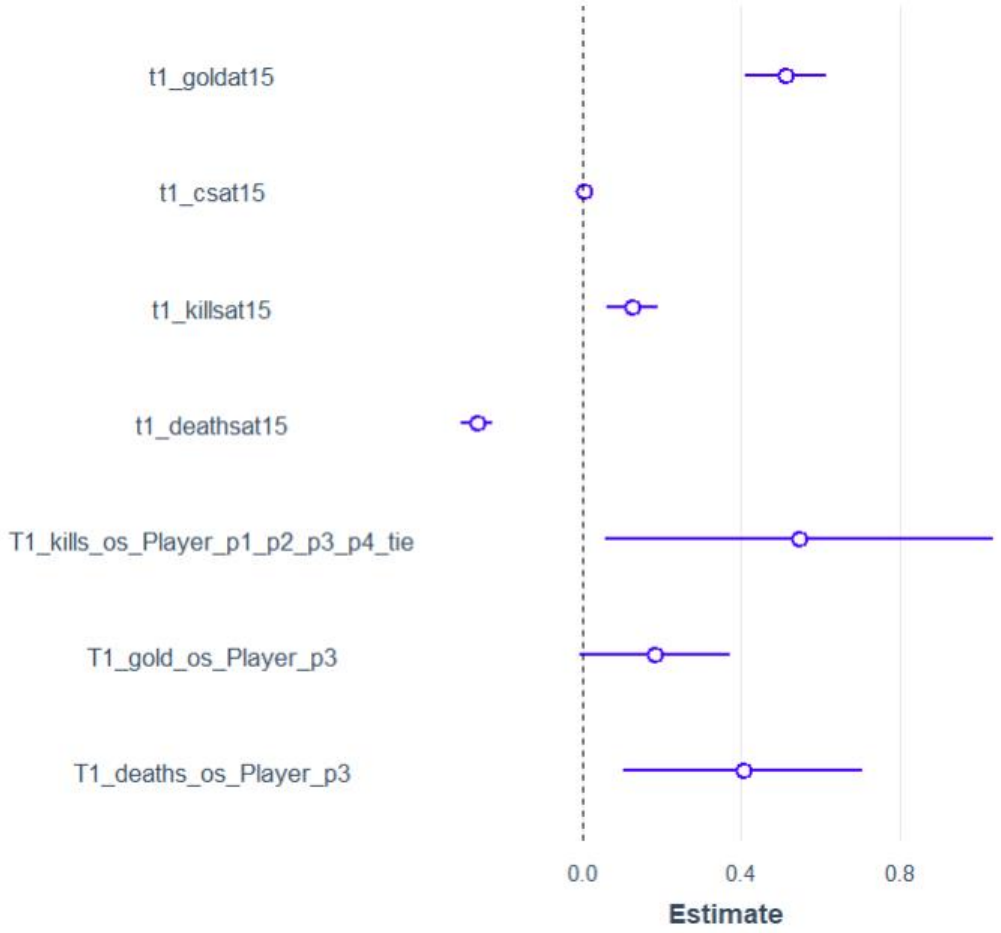
- T1_kills_os_Player: Níveis p1, p2, p3, p4 e tie_kill.

- T1_deaths_os_Player: Níveis p1, p2, p4, p5, tie_death e z_death (Reagrupamento da Categoria de Referência).
- Exclusão: A variável T1_xp_os_Player foi removida do modelo por falta de significância em todos os níveis.

Quadro 20: Terceira estimação dos parâmetros do modelo 15

Variável	Estimate	Std. Error	z value	Pr(>z)	
(Intercept)	-14.222499	1.049407	-13.553	< 2e-16	***
t1_goldat15	0.512759	0.053103	9.656	< 2e-16	***
t1_csat15	0.003008	0.001757	1.712	0.086839	°
t1_killsat15	0.126866	0.032877	3.859	0.000114	***
t1_deathsat15	-0.266920	0.019630	-13.598	< 2e-16	***
T1_kills_os_Player_p1_p2_p3_p4_tie	0.547017	0.249334	2.194	0.028242	*
T1_gold_os_Player_p3	0.183015	0.096433	1.898	0.057716	°
T1_deaths_os_Player_p3	0.405754	0.153851	2.637	0.008356	**

Figura 16: Gráfico dos intervalos de confiança dos parâmetros para terceira estimação do modelo 15.

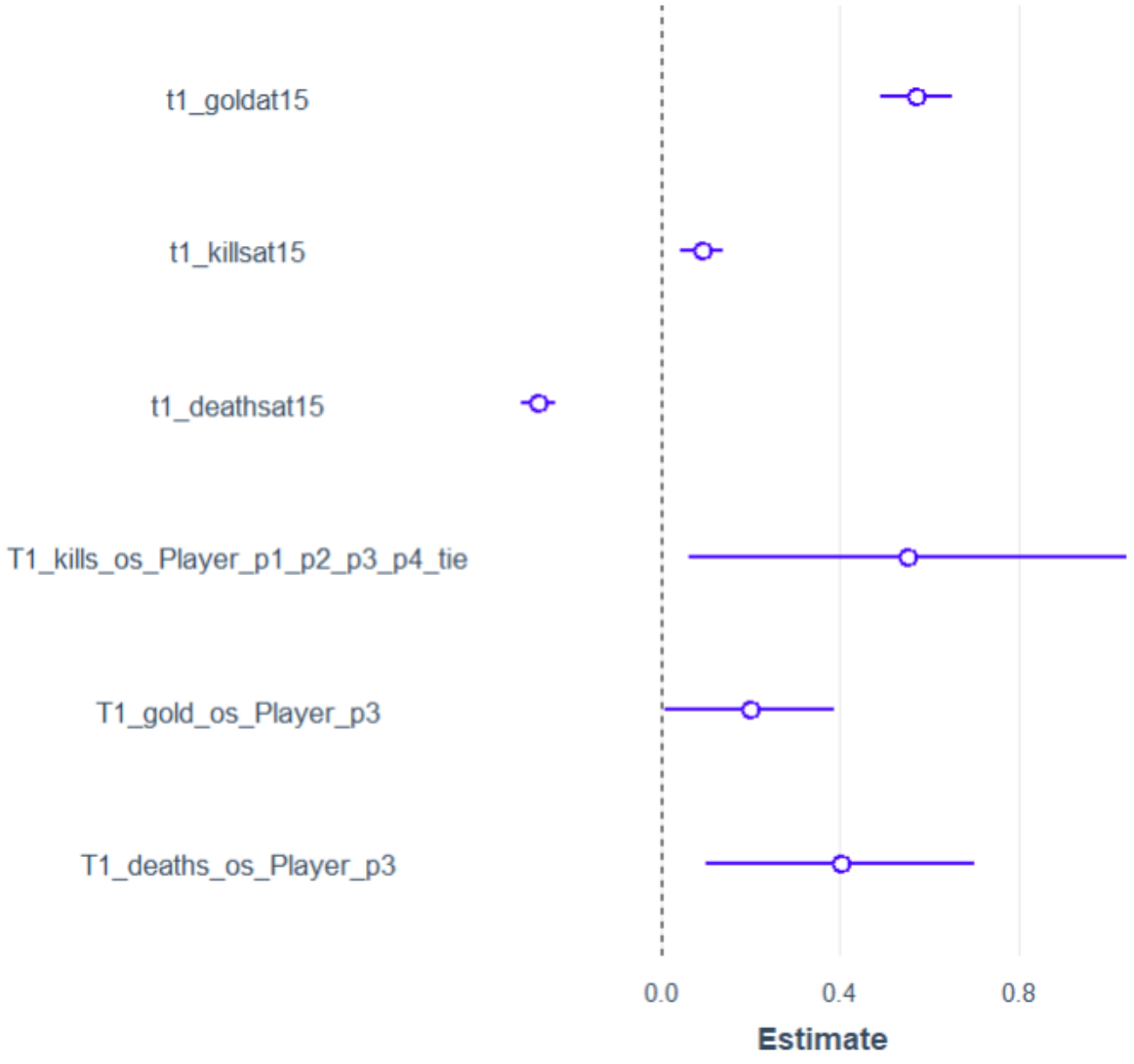


Com quase todas as variáveis significativas ao nível de 0.05 , seguindo a ordem dos p-valores maiores para remoção, é retirada a variável t1_csat15 e recalculado os parâmetros.

Quadro 20: Quarta estimação dos parâmetros do modelo 15

Variável	Estimate	Std. Error	z value	Pr(>z)	
(Intercept)	-13.93768	1.03365	-13.484	< 2e-16	***
t1_goldat15	0.57002	0.04150	13.736	< 2e-16	***
t1_killsat15	0.09046	0.02499	3.620	0.000295	***
t1_deathsat15	-0.27574	0.01896	-14.540	< 2e-16	***
T1_kills_os_Player_p1_p2_p3_p4_tie	0.55160	0.24971	2.209	0.027176	*
T1_gold_os_Player_p3	0.19715	0.09606	2.052	0.040136	*
T1_deaths_os_Player_p3	0.40129	0.15379	2.609	0.009073	**

Figura 17: Gráfico dos intervalos de confiança dos parâmetros para quarta estimação do modelo 15.



Com todos os parâmetros apresentando p-valor significativo, procede-se à avaliação de multicolinearidade antes das métricas de ajuste do modelo.

Quadro 21: Valores de VIF do modelo 15.

Variável	VIF	Intensidade da multicolinearidade
t1_goldat15	1.913229	Moderada
t1_killsat15	2.464159	Moderada
t1_deathsat15	1.339576	Moderada
T1_kills_os_Player_p1_p2_p3_4_tie	1.007663	Moderada
T1_gold_os_Player_p3	1.043434	Moderada

Com nenhuma variável apresentando valor alto de VIF, o modelo se apresenta livre do problema de multicolinearidade, de forma que o próximo passo é a avaliação de desempenho do modelo.

7.6 Avaliação do modelo 15

Utilizando como corte um valor aproximado da taxa de resposta da base (0,54) obtém-se os seguintes resultados:

Quadro 22: Valores reais versus preditos

	0 observado	1 observado
Previsto 0	374	147
Previsto 1	124	415

Quadro 23: Medidas de qualidade de ajuste na base de teste

Acurácia	0,74434
Precisão	0,76994
Sensibilidade (Recall)	0,73843
F1 Score	0,75386

- Acurácia de 74,43%: O modelo acerta a classificação geral em cerca de três quartos das vezes.
- F1-Score de 0,7538: Esse diz a respeito do equilíbrio entre a Precisão e a Sensibilidade (Recall), ou seja, ele não está dando forte preferência a uma única medida, mas sim buscando um meio termo entre as duas.
- Precisão (76.99%): Quando o modelo diz que o time 1 vai ganhar (Previsto 1), ele está certo cerca de 77% das vezes.

- Sensibilidade (Recall) de 73.84%: Diz que o modelo foi capaz de capturar aproximadamente 78% de todos os casos positivos que realmente ocorreram.

Apesar de assumir 0,54 como corte para previsão do modelo, é possível observar o comportamento de cada medida para diferentes pontos de corte utilizados.

Figura 18: Gráfico da Precisão, Recall, F1 Score, Acurácia por ponto de corte do modelo 10

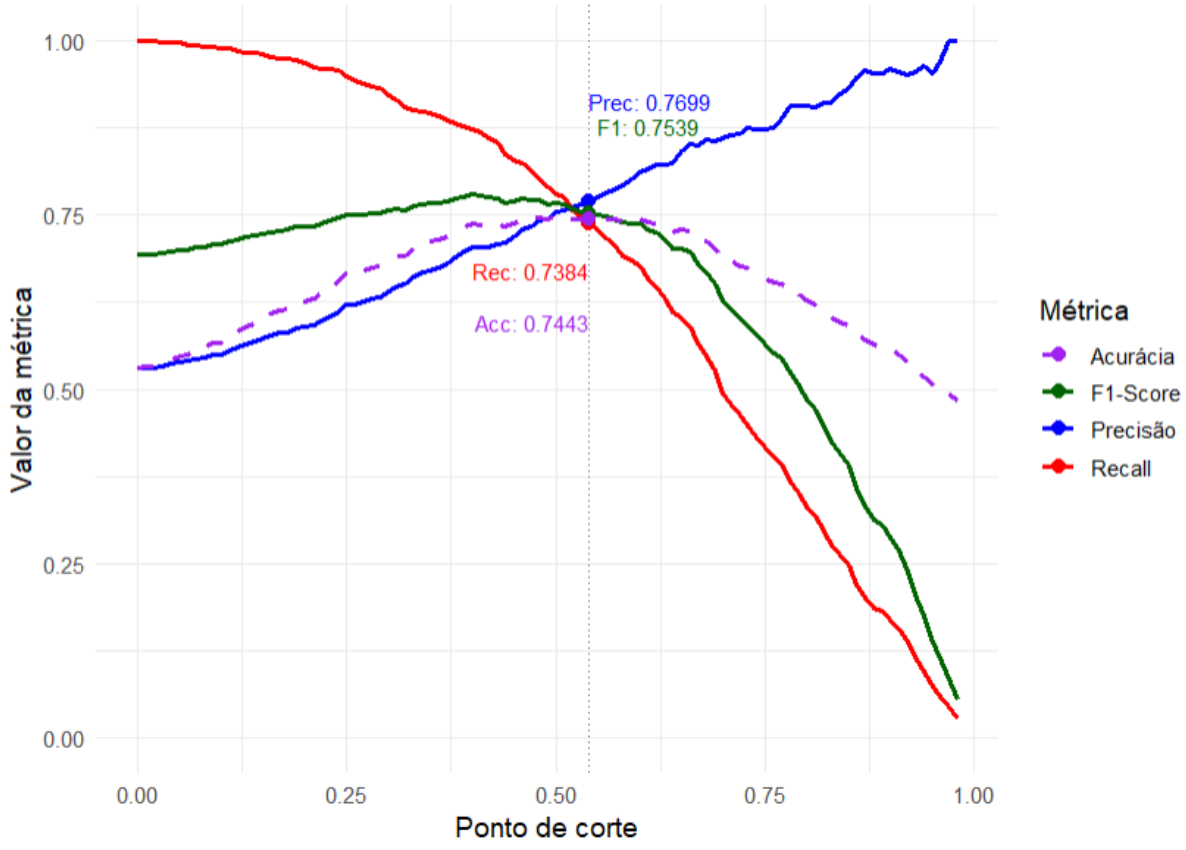
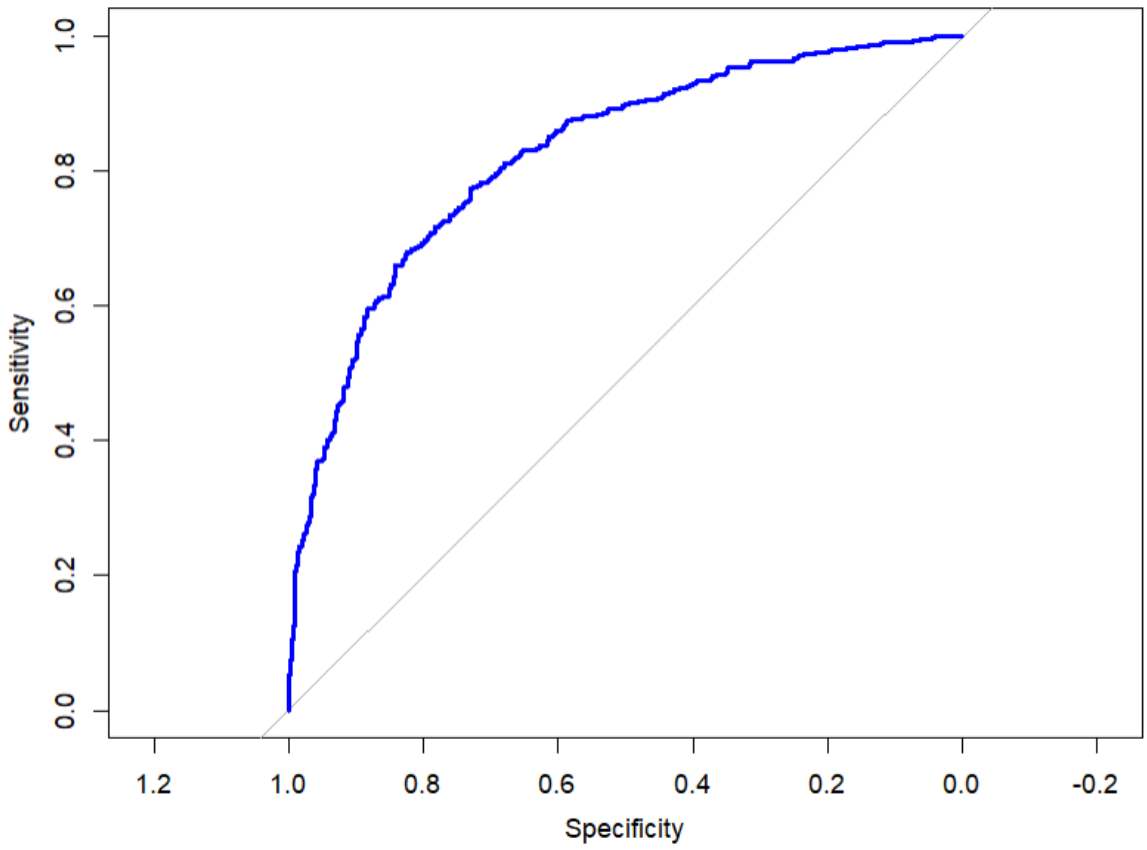


Figura 19: Curva ROC do modelo 15



Apresentando uma AUC (Area Under the Curve) igual a 0,8228, indica uma boa qualidade de ajuste. A Curva ROC encontra-se significativamente acima da linha diagonal, confirmando a validade do ajuste, visto que quanto mais abaulada e afastada a curva estiver dessa linha de referência, melhor é o desempenho do modelo.

8 Interpretação de parâmetros

Como dito anteriormente, a regressão logística possui parâmetros interpretáveis através de seu exponenciação. Dessa forma podemos não só aprender com os dados, mas também com o próprio modelo de regressão.

Dado a OddsRatio, essa pode ser interpretada de 2 formas principais, o valor puro (OR) ou a sua variação (OR – 1) para cunho comparativo.

A interpretação utilizando apenas a OR consiste em avaliar quantas vezes é a chance do evento de interesse ocorrer em um determinado nível de uma variável categórica em comparação ao seu nível de referência. Esse nível de referência é definido no processo de criação das variáveis dummy, no qual uma variável com *n* categorias é convertida em (n − 1) dummies. Assim, a categoria que não é transformada em dummy torna-se automaticamente o nível de referência, servindo como base para interpretar as odds ratios dos demais níveis incluídos no modelo.

Já a utilização de (OR − 1) segue a mesma lógica interpretativa da odds ratio, porém permite expressar o resultado em termos de aumento ou redução proporcional da chance. Enquanto a OR indica a relação entre as chances de dois níveis, o valor de (OR − 1) mostra quanto a chance aumenta ou diminui. Assim, valores positivos indicam aumento da chance de ocorrer o evento de interesse, e valores negativos indicam redução em relação ao nível de referência. Ao adicionar o percentual podemos dizer a relação de aumento ou diminuição em porcentagem. Para esse trabalho será utilizar a interpretação da (OR-1)%.

Para as variáveis numéricas a interpretação da (OR-1)% se dá pelo incremento na unidade da variável independente, ou seja, para cada 1 incremento na variável numérica há um aumento ou redução de (OR-1)% na chance de o evento de interesse ocorrer.

Quadro 24: parâmetros e suas exponenciais modelo 10.

Parâmetro	Estimativa	Exp(estimativa)	Exp(estimativa)-1	(Exp(estimativa)-1)%
(Intercept)	-11.482365	0,000010	-0,999990	-99,99%
t1_goldat10	0.517864	1,678439	0,678439	67,84%
t1_csat10	0.011190	1,011253	0,011253	1,13%
t1_killsat10	0.216749	1,242032	0,242032	24,20%
t1_deathsat10	-0.286499	0,750888	-0,249112	-24,91%
T1_cs_os_Player_p4_p5	-0.181523	0,833999	-0,166001	-16,60%
T1_deaths_os_Player_p3_tie_z	0.196932	1,217661	0,217661	21,77%

Ao observar as (OR-1)% do modelo 10 extrai-se as seguintes interpretações

- t1_goldat10: para cada 1000 de gold (ouro) a mais adquirido pelo time 1 até os 10 minutos de jogo, há um aumento de 67,84% na chance de ganhar o jogo.
- t1_csat10: Para cada 1 unidade de monstro a mais que o time 1 abate há um aumento de 1,13% na chance de vencer o jogo.
- t1_ killsat10: Para cada 1 abate a mais que o time 1 realiza até os 10 minutos de jogo, há um aumento de 24,20% na chance de vencer.
- t1_ deathsat10: Para cada 1 morte a mais que o time 1 sofre até os 10 minutos de jogo, há uma redução de 24,91% na chance de vencer.

Para interpretar o valor de $(\mathrm{OR}-1)\%$ da dummy T1_cs_os_Player_p4_p5, é importante relembrar como essa variável foi agrupada e o significado dos termos p4 e p5. Conforme descrito no índice 6 deste trabalho, p4 e p5 correspondem aos jogadores conhecidos como ADC e Suporte, que atuam juntos na rota inferior (Bot lane).

Também é relevante destacar que, após os agrupamentos realizados na modelagem, a variável T1_cs_os_Player passou a ter apenas dois níveis: o grupo formado por *(p1, p2, p3)* e o grupo formado por *(p4, p5)*. Esses conjuntos podem ser entendidos como “demais rotas” para o primeiro grupo e “Bot lane” para o segundo. Com essa estrutura, a interpretação que pode ser extraída é a seguinte:

- T1_cs_os_Player_p4_p5: A chance de ganhar o jogo reduz em 16,6% se a Bot lane for a rota com maior número de monstros abatidos. A partir desse racional

entende-se que as demais rotas devem desempenhar um papel melhor no abate de monstros do que a rota inferior, de forma que a responsabilidade pelos abates de monstros deve se concentrar sobre os jogadores p1, p2 e p3

De forma semelhante ao procedimento adotado na interpretação anterior, é preciso esclarecer o significado da variável T1_deaths_os_Player e como ela foi agrupada na modelagem. Essa variável indica o jogador em que se concentra o maior número de mortes da equipe e foi organizada em dois níveis: o grupo (p1, p2, p4, p5) e o grupo (p3, tie_deaths, z_deaths).

Retomando a descrição apresentada na Seção 6 deste trabalho, recorda-se que as posições p1, p2, p4 e p5 correspondem às Side lanes. Assim, os agrupamentos podem ser traduzidos como: (p1, p2, p4, p5) representando as Side lanes, e (p3, tie_deaths, z_deaths) representando Mid lane ou ausência de concentração de mortes.

Com essa estrutura, é possível extrair a seguinte interpretação:

- T1_deaths_os_Player_p3_tie_z: Dentre os 10 primeiros minutos de jogo, a chance de ganhar o aumenta em 21,77% caso não haja uma concentração de mortes nas Side lanes , mas sim uma concentração dessas mortes na Mid lane ou sua distribuição pelas rotas. Essa informação diz a respeito da fragilidade das Side lanes em comparação a Mid lane.

Quadro 25: parâmetros e suas exponenciais modelo 15.

Parâmetro	Estimativa	Exp(estimativa)	Exp(estimativa)-1	(Exp(estimativa)-1)%
(Intercept)	-13.93768	0.000001	-0.999998	-99,99%
t1_goldat15	0.57002	1.770014	0,768302	76,83%
t1_killsat15	0.09046	1.113633	0,094678	9,47%
t1_deathsat15	-0.27574	0.757238	-0,240990	-24,09%
T1_kills_os_Player_p1_p2_p3_p4_tie	0.55160	1. 736028	0,736028	73,60%
T1_gold_os_Player_p3	0.19715	1.208917	0,217927	21,79%
T1_deaths_os_Player_p3	0.40129	0.671829	0,493750	49,37%

Agora a interpretação dos parâmetros do modelo 15.

- t1_goldat15: para cada 1000 de gold a mais adquirido pelo time 1 até os 15 minutos de jogo, há um aumento na chance de ganhar o jogo de 76,83%.
- t1_ killsat15: Para cada 1 abate a mais que o time 1 realiza até os 15 minutos de jogo, há um aumento de 9,47% na chance de vencer.
- t1_ deathsat15: Para cada 1 morte a mais que o time 1 sofre até os 15 minutos de jogo, há uma redução de 24,09% na chance de vencer.

A variável T1_kills_os_Player foi reagrupada de modo que, na sua forma final, ela indica apenas uma coisa: se o jogador 5 é ou não o jogador com maior número de abates na equipe. Assim, o grupo p5 representa a situação em que o player 5 concentra os abates, enquanto o agrupamento (p1, p2, p3, p4, tie_kills) indica que o player 5 não é o líder em abates.

Sendo assim pode-se dizer que:

- T1_kills_os_Player_p1_p2_p3_4_tie: A chance de vencer o jogo aumenta em 73,60% caso o player 5 não seja a principal fonte de abates de inimigos até os 15 minutos de jogo.

De forma semelhante ao que ocorreu com a variável T1_kills_os_Player, a variável T1_gold_os_Player também foi reagrupada de forma a cumprir uma função simples: indicar se o player 3 é ou não o jogador com maior concentração de ouro na equipe. Assim, o agrupamento (p1, p2, p4, p5) representa situações em que o player 3 não é quem concentra o ouro, enquanto o nível *p3* indica que ele é o jogador com maior acúmulo de ouro.

- T1_gold_os_Player_p3: A chance de vencer o jogo aumenta em 21,79% caso o player 3 seja o jogador com maior quantidade de gold até os 15 minutos de jogo.

Da mesma maneita que T1_gold_os_Player_p3 narra a concentração de ouro em p3, a variável T1_deaths_os_Player_p3 narra a concentração de mortes em da partida em função de p3.

- T1_deaths_os_Player_p3: A chance de vencer o jogo aumenta em 49,37% caso o player 3 seja o jogador com maior quantidade de gold até os 15 minutos de jogo.

9 Conclusão

Dos modelos propostos no início deste trabalho, que visavam realizar a previsão do resultado da partida em 3 momentos diferentes, a partir da seleção de personagens (início da partida), a partir dos 10 minutos de jogo e a partir dos 15 minutos, apenas 2 modelos apresentaram um desempenho satisfatório, o Modelo 10 e o Modelo 15.

Entre os dois modelos analisados, além da capacidade preditiva para o resultado das partidas, o aspecto mais enriquecedor está na interpretação de seus parâmetros. Esses coeficientes permitem construir uma narrativa sobre o comportamento multivariado dos dados, revelando como cada variável se relaciona com o evento de interesse considerando simultaneamente a presença de todas as demais variáveis incluídas no modelo.

Ao olhar para o modelo 10, este foi capaz de capturar alguns comportamentos importantes para decisão de vitória. Sendo esses:

- Fragilidade das Side lanes dentre os 10 minutos iniciais do jogo: Em casos de que as mortes do time não se concentrem nas rotas laterais a chance de vitória aumenta em 21.77%, destacando a capacidade da Mid lane conseguir absorver mais pressão e ataque do time inimigo em relação as rotas laterais.
- Abate de monstros como uma prioridade maior para Top lane, Mid lane e Jungle, sendo que a chance de ganhar o jogo reduz em 16,6% nos casos em que a Bot lane apresenta a maior quantidade de monstros abatidos.

Ao comparar os 2 modelos e suas interpretações, pontos importantes podem ser levantados.

O abate de monstros total da equipe é mais importante nos 10 minutos iniciais do jogo do que aos 15 minutos, visto que o número de monstros abatidos só teve *p*-valor significativo para o Modelo 10. Além disso, o abate de inimigos (*kills*) se mostrou um objetivo mais importante e impactante nos minutos iniciais, uma vez que a cada abate realizado pela equipe até os 10 primeiros minutos há um aumento de 24,20% na chance de ganhar o jogo, enquanto aos 15 minutos os abates têm um impacto de apenas 9,47% na chance de ganhar.

Ao olhar o modelo 15 individualmente, um dado importante que o modelo traz é o de que o player 5 (Support) apresenta uma baixa capacidade de distribuir vantagem para seu time uma vez que esse não é o jogador com maior quantidade de

abates da equipe, o que aumenta a chance de vitória em 73,6%, indicando que essa posição tem um baixo potencial de pressionar a equipe inimiga e levar o time para a vitória.

A respeito da qualidade de ajuste o modelo 15 apresentou um desempenho significativamente melhor que o modelo 10.

10 Referências

Assis, B. M. de, & Costa, L. M. **Interações sociais e dinâmicas comunicacionais nos games**. Animus. Revista Interamericana De Comunicação Midiática. Disponível em: https://periodicos.ufsm.br/animus/article/view/21855/pdf. Acesso em: 15 nov 2023

BOLFARINE, Heleno; SANDOVAL, Mônica Carneiro. **Introdução à inferência estatística**. 2. ed. Rio de Janeiro: Sociedade Brasileira de Matemática, 2010.

BUSSAB, Wilton de Oliveira; MORETTIN, Pedro Alberto. **Estatística básica.** 4. Ed. São Paulo: Atual, 1987.

CALVO, Teodoro Balbino. **Predição de Partidas de Dota2 via Machine Learning**. Orientador: Prof. Dr. Manoel Ivanildo Silvestre Bezerra. 2017. 48p. TCC (Graduação) – Curso de Estatística, Universidade Estadual Paulista, Presidente Prudente, 2017.

CAPITA, Ítalo Luigi Santa. **Regressão Logística Em League Of Legends**. Orientadora: Profa. Dra. Miriam Rodrigues Silvestre. 2023. 48p. TCC (Graduação) – Curso de Estatística, Universidade Estadual Paulista, Presidente Prudente, 2023.

GE. **MSI 2023 aumentará premiação com venda de pacotes no LoL.** Disponível em: https://ge.globo.com/esports/lol/noticia/2023/04/25/c-msi-2023-aumentara-premiacao-com-venda-de-pacotes-no-lol.ghtml. Acesso em: 15 nov. 2023.

GONZALEZ, Leandro de Azevedo. **Regressão Logística e suas Aplicações.** Orientador: Prof. Dr. Ivo José da Cunha Serra. 2018. 46p. Monografia (Graduação) – Curso de Ciência da Computação, Universidade Federal do Maranhão, São Luís, 2018.

Itforum. **Esports: gerando empregos e se tornando uma indústria gigante.** Disponível em: https://itforum.com.br/noticias/esports-gerando-empregos-e-se-tornando-uma-industria-gigante/#:~:text=Al%C3%A9m%20dos%20jogadores%2C%20hoje%20os,algumas%20carreiras%20citadas%20por%20ele. Acesso em: 15 nov 2023

HOSMER JR., David W, and Stanley Lemeshow. **Applied Logistic Regression**. New York: J. Wiley, 1989. Print.

MINGOTIi, Sueli Aparecida. **Análise de dados através de métodos de estatística multivariada : uma abordagem aplicada**. Belo Horizonte: Ed. UFMG, 2005. Print.

JOHNSON, Richard A. Richard Allen, and Dean W Wichern. **Applied Multivariate Statistical Analysis**. 6. ed.-. Delhi [India: Pearson Education, 2007. Print.

MONTGOMERY, Douglas C, Elizabeth A Peck, and G. Geoffrey Vining. **Introduction to Linear Regression Analysis**. 3. ed.-. New York: John Wiley & Sons, 2001. Print.

SHI, C. et al. **A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm**. *Journal of Wireless Communications and Networking*, v. 2021, n. 31, 2021. Disponível em: https://doi.org/10.1186/s13638-021-01910-w. Acesso em: 17 out 2025

MACIEL, A. M.; VINHAS, L.; CÂMARA, G**. Algoritmos de clustering para separação de culturas agrícolas e tipos de uso e cobertura da Terra utilizando dados de sensoriamento remoto**. In: SIMPÓSIO BRASILEIRO DE SENSORIAMENTO REMOTO – SBSR, 17., 2015. *Anais* [...]. João Pessoa: INPE, 2015. P. 4620-4627. Disponível em: http://www.dsr.inpe.br/sbsr2015/files/p0903.pdf. Acesso em: 17 out. 2025.

11 APÊNDICE – MODELO 0

Será chamado de modelo 0 o modelo de Regressão Logística com o objetivo de prever o resultado da partida a partir da seleção de campeões, ou seja, ainda mesmo no minuto 0 de jogo.

Como visto anteriormente, existem muitos campeões e possibilidades de combinações de times. Dado que cada time possui 5 integrantes, ao colocar no modelo as variáveis que dizem qual campeão cada jogador selecionou, resultaria em 10 variáveis independentes com mais de 100 níveis cada. Dessa forma, será usado o agrupamento pelo método Ward para definir apenas de qual grupo aquele campeão selecionado pertence, resultando em 10 variáveis, cada uma com apenas 3 níveis. A partir desse racional, foram criadas as seguintes variáveis:

Quadro 7: Variáveis criadas a partir da análise de agrupamento para a aplicação no modelo 0

Variável criada	Tipo de variável	Descrição	Níveis
t1p1_champion_grupo	Categórica	Grupo ao qual o campeão selecionado pelo player 1 do time 1 pertence	grupo_1, grupo_2, grupo_3
t1p2_champion_grupo	Categórica	Grupo ao qual o campeão selecionado pelo player 2 do time 1 pertence	grupo_1, grupo_2, grupo_3
t1p3_champion_grupo	Categórica	Grupo ao qual o campeão selecionado pelo player 3 do time 1 pertence	grupo_1, grupo_2, grupo_3
t1p4_champion_grupo	Categórica	Grupo ao qual o campeão selecionado pelo player 4 do time 1 pertence	grupo_1, grupo_2, grupo_3
t1p5_champion_grupo	Categórica	Grupo ao qual o campeão selecionado pelo player 5 do time 1 pertence	grupo_1, grupo_2, grupo_3
t2p1_champion_grupo	Categórica	Grupo ao qual o campeão selecionado pelo player 1 do time 2 pertence	grupo_1, grupo_2, grupo_3
t2p2_champion_grupo	Categórica	Grupo ao qual o campeão selecionado pelo player 2 do time 2 pertence	grupo_1, grupo_2, grupo_3
t2p3_champion_grupo	Categórica	Grupo ao qual o campeão selecionado pelo player 3 do time 2 pertence	grupo_1, grupo_2, grupo_3
t2p4_champion_grupo	Categórica	Grupo ao qual o campeão selecionado pelo player 4 do time 2 pertence	grupo_1, grupo_2, grupo_3
t2p5_champion_grupo	Categórica	Grupo ao qual o campeão selecionado pelo player 5 do time 2 pertence	grupo_1, grupo_2, grupo_3

Ao aplicar a estimação dos parâmetros da regressão logística, utilizando as novas variáveis, obtêm-se os seguintes valores.

Quadro 7: Valores estimados dos parâmetros do modelo 0

Variável	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-1.218.683	1.387.477	-0.878	0.3798	
t1p1_champion_grupo_2	0.145000	0.086721	1.672	0.0945	°
t1p1_champion_grupo_3	-0.075662	0.074110	-1.021	0.3073	
t1p2_champion_grupo_2	0.114102	0.211049	0.541	0.5888	
t1p2_champion_grupo_3	0.167125	0.206424	0.810	0.4182	
t1p3_champion_grupo_2	0.003597	0.129888	0.028	0.9779	
t1p3_champion_grupo_3	-0.009611	0.144264	-0.067	0.9469	
t1p4_champion_grupo_2	1.205.442	1.176.789	1.024	0.3057	
t1p4_champion_grupo_3	0.070525	1.358.566	0.052	0.9586	
t1p5_champion_grupo_2	0.117723	0.173788	0.677	0.4982	
t1p5_champion_grupo_3	0.209529	0.165440	1.266	0.2053	
t2p1_champion_grupo_2	-0.076049	0.088313	-0.861	0.3892	
t2p1_champion_grupo_3	0.076123	0.076501	0.995	0.3197	
t2p2_champion_grupo_2	0.002468	0.192611	0.013	0.9898	
t2p2_champion_grupo_3	0.096773	0.188756	0.513	0.6082	
t2p3_champion_grupo_2	0.048565	0.134684	0.361	0.7184	
t2p3_champion_grupo_3	0.027618	0.146387	0.189	0.8504	
t2p4_champion_grupo_2	-0.039899	0.854264	-0.047	0.9627	
t2p4_champion_grupo_3	0.594249	0.972918	0.611	0.5413	
t2p5_champion_grupo_2	-0.323325	0.180954	-1.787	0.0740	°
t2p5_champion_grupo_3	-0.235297	0.173326	-1.358	0.1746	

A tabela deixa evidente que nenhuma das variáveis é significativamente importante para o modelo. Com o objetivo de simplificar o modelo e avaliar se alguma variável pode ter um poder preditivo significativo, aplica-se o método Stepwise e é retornado apenas o modelo com o intercepto, indicando que, mesmo na presença ou ausência de outras variáveis, nenhuma apresenta significância estatística para o modelo.

Esse resultado é um indicativo de que apenas o conhecimento da seleção dos campeões da partida não é o suficiente para conseguir prever o resultado através da regressão logística.