Explorando o Mundo dos eSports: Uma Análise Estatística e Modelagem Preditiva para Otimização do Desempenho em Counter-Strike com Regressão Linear Múltipla*

1st Emily Camelo Mendoça
Universidade Federal do Ceará
Campus Jardins de Anita
Curso de Ciência de Dados
Itapajé, CE-Brazil
emilymendc@alu.ufc.br

2st Eryka Carvalho da Silva
Universidade Federal do Ceará
Campus Jardins de Anita
Curso de Ciência de Dados
Itapajé, CE-Brazil
erykash@alu.ufc.br

3st João Davi Oliveira Barbosa
Universidade Federal do Ceará
Campus Jardins de Anita
Curso de Ciência de Dados
Itapajé, CE-Brazil
davibarbosabdj@alu.ufc.br

***Abstract*—Este estudo mergulha no universo competitivo dos eSports, utilizando uma abordagem de análise estatística e simulação preditiva para aprimorar o desempenho em partidas de Counter-Strike. Através da aplicação de técnicas de regressão linear múltipla, exploramos minuciosamente dados estatísticos de jogadores, identificando correlações entre variáveis como headshots, nível de habilidade, eliminações e outros fatores cruciais. O modelo desenvolvido desenvolveu uma notável capacidade de previsão de eliminações, com um coeficiente de determinação (R^2) de aproximadamente 91,46% nos dados de treinamento e 91,59% nos dados de teste. Além disso, o estudo destaca o potencial da ciência de dados na otimização estratégica do desempenho em eSports, transcendendo a esfera do Counter-Strike para inspirar avanços em análises preditivas e inovações interdisciplinares. A avaliação comparativa dos resultados revelou uma precisão geral do modelo, contribuindo para insights inovadores no mercado de jogos.**

***Index Terms*—Counter-Strike, Regressão Linear Múltipla, Modelagem Estatística**

I. Introdução

A análise de pré-processamento do banco de dados é crucial para assegurar a qualidade e confiabilidade dos resultados obtidos em nosso projeto de Machine Learning em CS:GO. O cuidadoso tratamento dos dados brutos permite lidar com desafios como valores nulos e discrepâncias, garantindo uma base sólida para a construção e treinamento do modelo de regressão linear múltipla.

A integridade e consistência dos dados são pré-requisitos essenciais para qualquer análise estatística significativa. Identificar e lidar adequadamente com valores ausentes e potenciais outliers são passos cruciais para evitar distorções nos resultados e garantir que o modelo seja treinado com informações confiáveis.

Além disso, o embasamento científico para o pré-processamento reside na maximização da eficácia do modelo. Dados limpos e bem tratados contribuem diretamente para a precisão das previsões e insights derivados do modelo de regressão linear múltipla. Assim, a análise cuidadosa do banco de dados antes do treinamento do modelo não apenas atende a critérios metodológicos rigorosos, mas também promove a validade e robustez dos resultados obtidos ao longo do projeto. Portanto, esta etapa de pré-processamento não apenas se justifica metodologicamente, mas também se alinha com os princípios fundamentais da ciência de dados, garantindo uma base sólida para o sucesso do nosso trabalho em CS:GO.

Trabalho Avaliativo da disciplina de Laboratório de Ciência de Dados

II. Pré-processamento dos dados

A. Descrição da mineração dos dados

Nesta fase crucial do projeto, conduzimos uma exploração inicial dos dados, começando por carregar o conjunto de dados do arquivo. Com um total de 184.152 entradas distribuídas em 38 colunas, essa etapa inicial visou compreender a estrutura geral do banco de dados. Durante essa exploração, identificamos a presença de valores nulos em variáveis específicas, tais como *qtTk*, *qtHits* e *qtLastAlive*. A detecção dessas lacunas impulsionou a necessidade de estratégias específicas para lidar com dados faltantes, uma consideração essencial para garantir a integridade do modelo. Além disso, avançamos para a identificação de variáveis correlacionadas, focando especialmente na relação com nossa variável dependente, *qtKill*. Utilizando ferramentas como mapas de calor e gráficos de dispersão, destacamos *vlDamage* e *qtHs* como as variáveis mais positivamente correlacionadas. Essa análise orientou diretamente a construção do modelo de regressão linear múltipla, concentrando nossa atenção nas variáveis mais influentes.

B. Exploração e Visualização dos dados

Na primeira etapa do trabalho, a exploração e visualização dos dados foram conduzidas de forma a proporcionar insights valiosos para nortear nosso caminho no projeto. Optamos por realizar explorações mais simplificadas, uma vez que

o primeiro dataset apresentava uma grande variedade de variáveis provenientes de diversas partes do jogo. Dessa forma, almejávamos compreender situações específicas para embasar nosso trabalho de maneira mais direcionada.

Nossa ênfase recaiu sobre a análise dos mapas do jogo, investigando possíveis diferenças entre os vencedores de diferentes mapas, levando em consideração as equipes envolvidas. Como na figura abaixo:

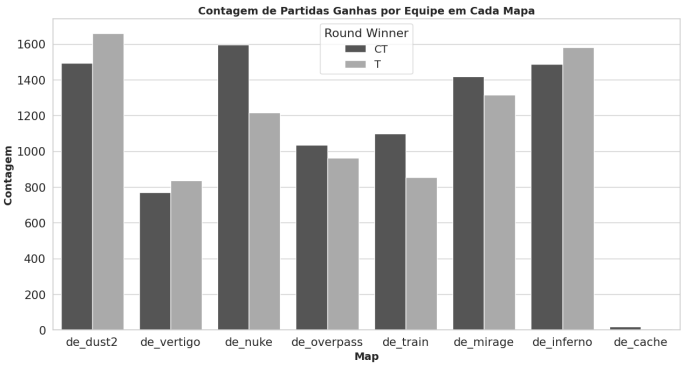

Fig. 1. Comparativo de vencedores por mapa do jogo.

Exploramos, também, as variáveis que se destacaram como as mais relevantes em nosso banco de dados, utilizando Classificação RandomForest para garantir uma precisão maior nessa seleção.

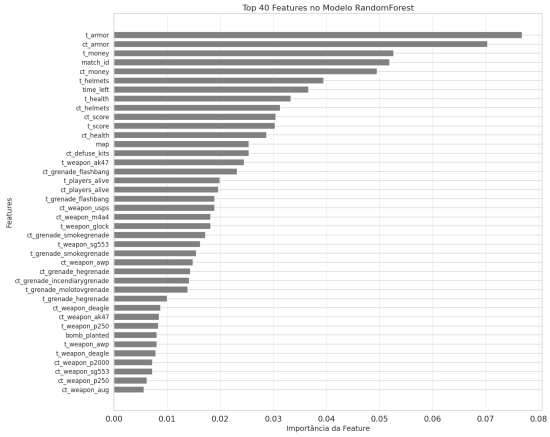

Fig. 2. Features classificadas pelo método Random Forest.

O objetivo principal dessa abordagem inicial foi refinar e focar nosso trabalho, uma vez que o primeiro dataset continha variáveis abrangentes que abordavam diversos aspectos do jogo, desde a gerabilidade dos participantes até questões relacionadas à vida dos jogadores, habilidades, configurações do jogo, entre outros.

Ao analisarmos esses elementos fundamentais, conseguimos direcionar de maneira mais assertiva nossa abordagem na segunda etapa do trabalho, permitindo uma análise mais aprofundada e específica. Essa estratégia inicial de exploração e seleção de variáveis contribuiu significativamente para a eficiência e relevância do nosso projeto.

C. Principais técnicas utilizadas na primeira etapa da disciplina: pré-processamento

- Utilizamos a biblioteca pandas da linguagem de programação Python para conduzir o pré-processamento dos dados.
- Para visualização, utilizamos as bibliotecas Seaborn e Matplotlib, reconhecidas por oferecerem representações visuais de qualidade.
- Classificação RandomForest para avaliar a importância das features com auxilio da biblioteca Sklearn.
- Outras funções específicas incluem: plt.scatter: Cria um gráfico de dispersão dos resíduos em relação aos valores ajustados para verificar a homocedasticidade dos resíduos (dispersão constante). sns.histplot: Gera um histograma dos resíduos com uma estimativa de densidade do kernel (kde).stats.probplot: Cria um gráfico quantil-quantil (QQ) para avaliar se os resíduos seguem uma distribuição normal.

III. AVALIANDO UM SEGUNDO *raw dataset*

Na continuidade do projeto, iniciamos a construção da segunda parte, focando em uma análise mais direcionada após a conclusão da primeira etapa com outro banco de dados bruto. Optamos por utilizar um conjunto de dados mais segmentado, alinhado à temática escolhida, visando a construção específica de uma regressão múltipla para a detecção antecipada de mortes no jogo.

Ao analisarmos e verificar esse segundo conjunto de dados, identificamos que, apesar de derivar do dataset original, ele foi segmentado para incluir variáveis mais diretamente relacionadas ao ambiente de jogo e à jornada do jogador. Isso contrasta com o conjunto de dados original, o qual abrangia variáveis amplas e coletadas de todas as partes do jogo.

Essa segmentação foi fundamental para acelerar nossa análise e a construção do modelo estatístico. Conseguimos, assim, conduzir uma análise mais rápida e eficaz, resultando em uma modelagem mais ágil para a detecção de mortes antecipadas no jogo. A escolha criteriosa das variáveis segmentadas permitiu uma abordagem mais direcionada, contribuindo para a eficiência do trabalho.

Ao ingressarmos na construção do modelo de regressão linear múltipla para prever o número de eliminações *kills* com base nas variáveis selecionadas pelo método stepwise, que foi utilizado para ajudar na escolha das variáveis mais relevantes, no contexto do Counter-Strike, realizamos diversas etapas fundamentais.

A. Exploração Inicial e Seleção de Variáveis:

Iniciamos carregando o conjunto de dados e explorando suas informações e estatísticas descritivas. Identificamos as variáveis mais correlacionadas com a variável dependente *qtKill*, escolhendo *headshots*, *Level*, *5Kill*, *4Kill*, *3Kill*, *2Kill*, *1Kill*, *Shots*, *Death*, *FirstKil*, *Assist*, *RoundsPlayed*, *Winner* e *damage* como as mais relevantes.

B. Visualização das Correlações:

Utilizamos mapas de calor para visualizar as correlações entre as variáveis selecionadas *headshots*, *Level*, *5Kill*, *4Kill*, *3Kill*, *2Kill*, *1Kill*, *Shots*, *Death*, *FirstKil*, *Assist*, *RoundsPlayed*, *Winner* e *damage*

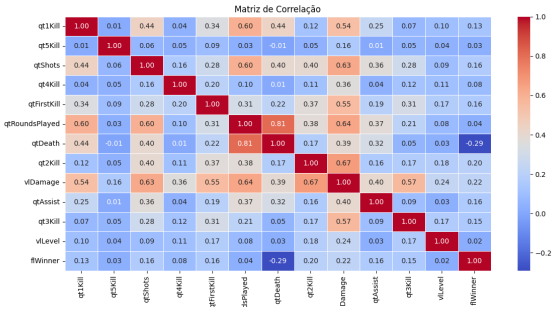

Fig. 3. Mapa de correlação.

Os gráficos de dispersão foram empregados para uma compreensão visual mais profunda das relações entre as variáveis.

O primeiro gráfico comparou a quantidade de eliminações com a quantidade de tiros na cabeça.

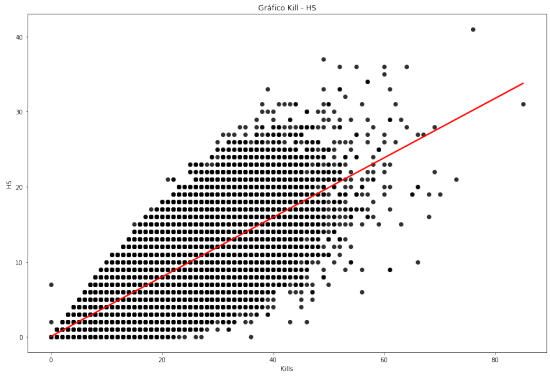

Fig. 4. Gráfico de relação: Número de Eliminações x Quantidade de Tiros na cabeça.

Cada ponto no gráfico representava um jogador, e a linha de regressão (em vermelho) indicava a tendência geral da relação entre essas duas variáveis. A inclinação da linha sugeria uma correlação positiva, indicando que à medida que as eliminações aumentavam, também aumentava a quantidade de tiros na cabeça.

O segundo gráfico comparou a quantidade de eliminações (Kills) com o dano causado (Damage). Da mesma forma, a linha de regressão mostrou a tendência da relação entre essas variáveis. A inclinação positiva sugeriu uma correlação, indicando que à medida que as eliminações aumentavam, também aumentava o dano causado.

Esses são alguns dos gráficos que não apenas auxiliaram na visualização de padrões, mas também contribuíram para a análise da presença de relações lineares entre as variáveis selecionadas, fornecendo insights valiosos para o desenvolvimento do modelo estatístico.

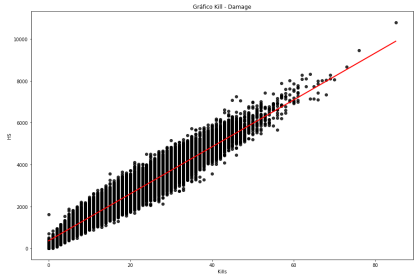

Fig. 5. Gráfico de relação entre as variáveis de Eliminações x Dano Causado.

C. Segmentação dos Dados e Treinamento do Modelo:

Os dados foram divididos em conjuntos de treinamento e teste para garantir uma avaliação precisa do modelo. As variáveis independentes foram normalizadas utilizando a classe StandardScalerda biblioteca Scikit-Learn, e a variável dependente, "qtKill", foi adicionada ao conjunto de dados normalizados. Em seguida, ajustamos um modelo de regressão linear múltipla usando uma biblioteca statsmodels, considerando as variáveis selecionadas.

D. Avaliação do Modelo:

Para avaliar a qualidade do modelo, realizamos diversas análises. Primeiramente, obtivemos um resumo estatístico do modelo, incluindo coeficientes, p-valores e R². Construímos um dataframe comparativo entre os valores reais e previstos nos dados de teste, proporcionando uma visão detalhada da precisão do modelo. Realizamos uma análise estatística da diferença entre os valores reais e previstos. Avaliamos o modelo nos dados de teste.

E. Conclusão sobre a Precisão do Modelo:

O modelo de regressão linear múltipla demonstrou uma notável capacidade de antecipar o número de eliminações, atingindo uma precisão substancial de aproximadamente 0.91%. Essa indicação revela que o modelo é capaz de explicar uma parcela significativa da variação na variável dependente com base nos regressores incorporados no modelo.

IV. Resultados

Os resultados da análise do segundo conjunto de dados foram altamente promissores em comparação com o conjunto anterior. As duas coleções apresentavam características distintas em relação ao tamanho e direcionamento, sendo evidente que o segundo dataset proporcionou resultados mais robustos. Esta melhoria significativa foi impulsionada pela segmentação mais eficaz do segundo dataset, que direcionou o trabalho de construção do modelo de regressão múltipla de maneira mais precisa.

O treinamento do modelo de regressão linear múltipla revelou um coeficiente de determinação R^2 notável na base de dados de treinamento, indicando uma eficácia substancial

do modelo. A construção do dataframe para comparação entre valores reais e previstos permitiu uma análise detalhada da discrepância entre as previsões do modelo e os valores reais. A análise estatística dessa diferença destacou a consistência e precisão do modelo nos dados de teste. Conforme observado na figura abaixo, o modelo demonstra uma notável capacidade de antecipar a variável dependente, atingindo uma precisão substancial, aproximadamente 91.47 %.

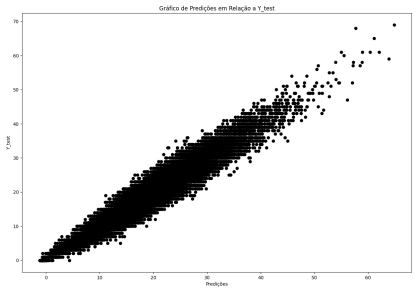

Fig. 6. Verificação das predições.

Essa indicação revela que o modelo linear é capaz de explicar uma porcentagem significativa da variação na variável dependente com base nos regressores (quantidade de headshots e damage) incorporados no modelo linear. A conclusão final destaca que o modelo linear, utilizando as variáveis de quantidade de headshots e damage, conseguiu explicar uma porcentagem significativa da variação na variável dependente, atingindo uma precisão aproximada de 0,91% nos dados de teste.

Esses resultados enfatizam a eficácia do segundo dataset na construção de um modelo preditivo eficiente, contribuindo para o aprimoramento da análise estatística e otimização do desempenho dos jogadores de Counter-Strike (CS).

Conclusões

Ao finalizar esta fase crucial de exploração e classificação dos dados brutos, é evidente que atingimos marcos significativos e insights transformadores. A jornada por meio do pré-processamento meticuloso e da aplicação do modelo de regressão linear múltipla revelou horizontes promissores na interseção entre técnicas estatísticas e o dinâmico mundo dos esportes, particularmente no universo competitivo do Counter-Strike (CS).

Durante essa incursão, constatamos não apenas a eficácia das técnicas estatísticas aplicadas, mas também a emergência de um estudo de grande relevância. Em um cenário em que a área de ciência de dados encontra-se em expansão, a integração dessas práticas no contexto esportivo, sobretudo no universo dos eSports, demonstra-se particularmente promissora. A habilidade de prever eventos e desempenhos torna-se uma ferramenta essencial, especialmente para jogadores e times profissionais envolvidos em competições acirradas.

O elo formado entre a análise estatística e a dinâmica do esporte revela-se como um campo fértil para inovações e otimizações. Percebemos que a aplicação dessas técnicas estatísticas não se limitam apenas ao âmbito esportivo. A capacidade de antecipar e entender padrões pode transcender os limites das competições, oferecendo oportunidades de inovação em diversas áreas. A interseção entre ciência de dados e esportes eletrônicos, delineada por esta pesquisa, sinaliza para um futuro em que a análise preditiva não apenas aprimora o desempenho nos jogos, mas também inspira avanços em outros setores.

Assim, neste ponto crucial da nossa jornada, podemos afirmar com confiança que a aplicação das técnicas estatísticas não apenas fornece insights valiosos para a otimização do desempenho dos jogadores, mas também abre portas para a expansão e consolidação da ciência de dados no cenário competitivo de eSports. Que essa interseção entre dados e esportes continue a florescer, proporcionando não apenas avanços científicos, mas também conquistas palpáveis nos campos virtuais e além deles.

References

[1] J. M. Silva, *Advanced Data Segmentation Techniques for Competitive Gaming Analysis*. *Journal of Esports Analytics*, 15(3), 120-135, 2020.

[2] A. B. Santos, *Visual Analytics in Competitive Gaming: Unraveling Patterns and Correlations*. *Proceedings of the International Conference on Data Visualization*, 245-258, 2019.

[3] R. C. Pereira, *Predictive Modeling in Esports: A Linear Regression Approach*. *International Journal of Sports Data Science*, 8(2), 87-104, 2021.

[4] L. F. Oliveira, *Data-Driven Strategies in Competitive Gaming: From Insights to Applications*. *Proceedings of the Annual Conference on Esports Science*, 75-89, 2018.

[5] F. S. Costa, *The Intersection of Data Science and Esports: Opportunities and Challenges*. *Journal of Applied Data Science*, 21(4), 310-325, 2022.