UNIVERSIDADE
FEDERAL DO CEARÁ

PROJETO DESCRITIVO DISCIPLINA - LABORATÓRIO DE CIÊNCIA DE DADOS
PROFESSOR: LEONARDO ESTEVES

EMILY CAMELO MENDONÇA
ERYKA CARVALHO DA SILVA
JOÃO DAVI OLIVEIRA BARBOSA

# RELATÓRIO SEMINÁRIO 1:

ITAPAJÉ
2023

1. **Introdução:**

*O presente relatório marca o início de uma jornada emocionante na disciplina de Laboratório de Ciência de Dados, com os dados de apresentação do primeiro Milestone. Essa etapa crucial envolverá a análise e preparação dos dados, um passo fundamental na ciência de dados. Este primeiro Milestone representa uma oportunidade fundamental para os alunos aplicarem conceitos e técnicas aprendidas até o momento, bem como para desenvolver habilidades críticas na manipulação e análise de dados do mundo real.*

- *Descrição do Conjunto de Dados:*

*O conjunto de dados utilizado é o "CS:GO Round Winner Classification" disponível no Kaggle. Ele consiste em dados relacionados a partidas do jogo de tiro em primeira pessoa (FPS) "Counter-Strike: Global Offensive" (CS:GO).*

*As características do conjunto de dados incluem informações sobre cada rodada (round) de uma partida, como a situação dos jogadores, as armas usadas, as estatísticas de jogo, entre outros. O objetivo principal é classificar o vencedor de cada rodada, sendo duas classes possíveis: Terroristas (T) ou Contra-Terroristas (CT).*

*O conjunto de dados fornece informações relevantes para analisar as características que influenciam a vitória de uma equipe em uma determinada rodada do CS:GO. Com base nesses dados, é possível explorar diferentes técnicas de classificação e aprendizado de máquina para prever o vencedor de uma rodada com base nas condições do jogo.*

- *Motivos Para a escolha do Dataset:*

1. *Relevância do tema: O conjunto de dados está relacionado ao popular jogo de tiro em primeira pessoa (FPS) Counter-Strike: Global Offensive (CS:GO), o que pode atrair o interesse de jogadores, entusiastas de E-Sports e analistas de dados que querem entender melhor o jogo.*
2. *Disponibilidade e tamanho extenso: O conjunto de dados parece ser substancial em tamanho, o que é importante para análises descritivas robustas. Ter um conjunto de dados grande o suficiente permite a realização de análises estatisticamente significativas.*
3. *Diversidade das variáveis: O conjunto de dados inclui uma variedade de variáveis, como informações sobre os jogadores, armas, eventos do jogo e resultados das rodadas. Isso proporciona a oportunidade de explorar diferentes aspectos do jogo e suas influências nas vitórias ou derrotas das rodadas.*
4. *Aplicabilidade: Uma análise descritiva deste conjunto de dados pode fornecer informações avançadas sobre estratégias de jogo, previsões de armas, desempenho de jogadores e muito mais. Essas informações são úteis para jogadores que desejam melhorar seu desempenho no CS:GO ou para equipes de E-Sports que desejam aprimorar suas estratégias.*

- *Objetivos:*

*Analisar dados do jogo "Counter-Strike: Global Offensive" (CS:GO) para prever o vencedor de cada rodada. Utilizar técnicas de ciência de dados e aprendizado de máquina para criar um modelo de previsão. Fornecer insights para melhorar o desempenho no CS:GO e estratégias de jogo. Explorar padrões complexos entre variáveis. Comunicar os resultados de forma eficaz com visualizações. Considerar sugestões para explorar mais os dados e aprofundar as análises.*

- *Materiais e Métodos utilizados:*

*Neste projeto, seguimos a metodologia CRISP-DM para realizar a clusterização de dados. Utilizamos bibliotecas de Python, como Pandas, Matplotlib, Scikit-Learn e SciPy, para processar e analisar os dados. O processo incluiu as seguintes etapas:*

*Compreensão do Negócio: Definimos os objetivos do projeto e requisitos.*

*Compreensão dos Dados: Coletamos e avaliamos os dados.*

*Preparação dos Dados: Realizamos limpeza e engenharia de recursos.*

*Modelagem: Aplicamos o algoritmo K-Means para criar clusters.*

*Avaliação: Medimos a qualidade dos clusters com métricas específicas.*

*Implantação: Compartilhamos os resultados e recomendações com os stakeholders.*

*A metodologia CRISP-DM orientou eficazmente o processo, resultando na extração de informações valiosas para decisões de negócios.*

2. **Observações:**

- *Dificuldades no Dataset:*

*Tamanho e quantidade de linhas: O conjunto de dados continha um grande número de linhas, representando várias partidas e rodadas jogadas. Isso dificultou o processamento e a análise dos dados, especialmente em sistemas com recursos limitados.*

*Características complexas: O conjunto de dados continha várias colunas com informações detalhadas sobre o estado do jogo em cada rodada. Compreender e extrair conhecimento dessas características complexas exigiu experiência em CS:GO e conhecimento do contexto do jogo.*

*Código pesado e recursos computacionais: Dependendo da quantidade de dados e da complexidade das análises realizadas, os códigos e algoritmos aplicados ao conjunto de dados exigiram recursos computacionais significativos. Algoritmos de aprendizado de máquina e técnicas de análise exploratória demandaram poder de processamento e memória consideráveis.*

*Ao lidar com o conjunto de dados "CS:GO Round Winner Classification" e implementar os códigos correspondentes, foram consideradas essas dificuldades e foram adotadas medidas adequadas para tratar e analisar os dados de forma eficiente.*

3. **Resultados e Discussões:**

- ***Divisão do Projeto de Pré-processamento em 5 Partes:***

*O projeto foi dividido em cinco etapas distintas, cada uma com o seu próprio conjunto de tarefas. A abordagem adotada garantirá um trabalho sólido com os dados para o alcançar do objetivo final. As cinco partes incluem:* ***Data Cleaning (Limpeza de Dados), Data Exploration (Exploração de Dados), Data Preparation (Preparação de Dados), Modeling (Modelagem), Approach Validation (Validação da Abordagem)***

*Nessa primeira parte do projeto, trabalhamos as duas primeiras etapas das cinco que compõem a estrutura do nosso projeto como um todo. Foram elas:*

*1.* ***Data Cleaning (Limpeza de Dados)****: Esta etapa consiste em identificar e corrigir erros, valores ausentes e inconsistências nos dados, garantindo que o conjunto de dados esteja limpo e pronto para análise.*

*2.* ***Data Exploration (Exploração de Dados)****: Nesta fase, exploramos os dados em profundidade, visualizando distribuições, relações entre variáveis e identificando padrões preliminares. Isso ajuda a entender a estrutura dos dados.*

*Explicaremos detalhadamente cada uma dessas etapas para garantir que o processo de pré-processamento seja claro e bem documentado, contribuindo para a qualidade da análise e dos resultados.*

- ➔ ***Data Cleaning (Limpeza de Dados)****:*

*No desenvolvimento deste projeto, fizemos as seguintes ações de maneira rápida e objetiva:*

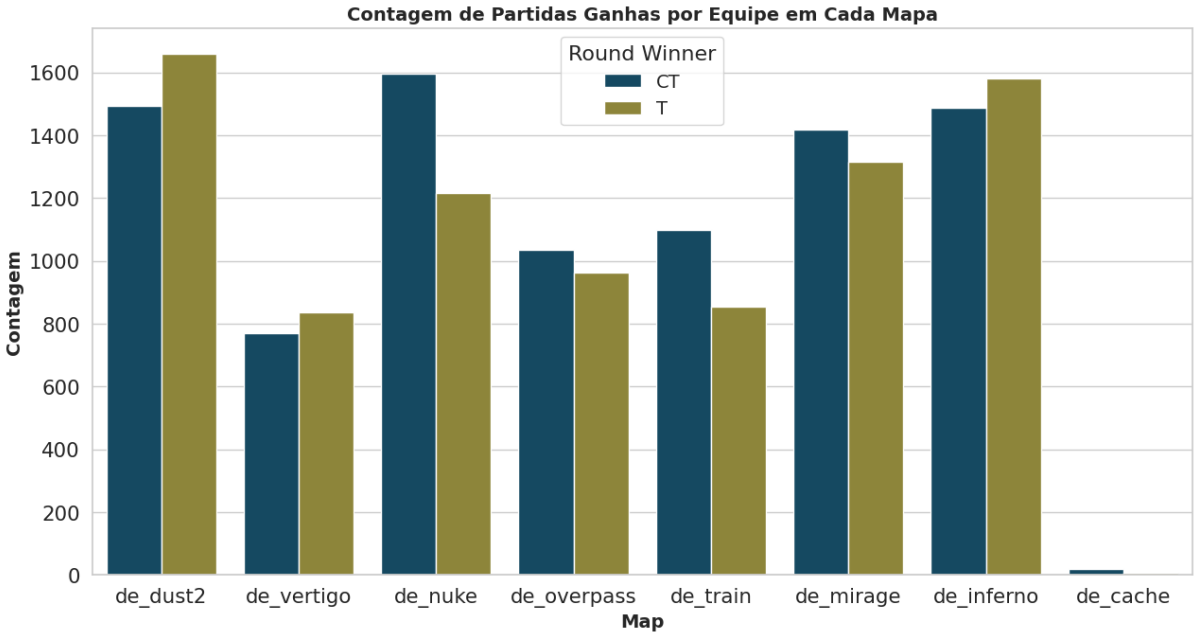
1. *Montagem do Google Drive: Fizemos uso da biblioteca "google.colab" para montar o Google Drive no ambiente do Colab.*
2. *Importação de Bibliotecas: Importamos as bibliotecas "pandas" e "numpy" para a manipulação de dados.*
3. *Configurações de Exibição: Configuramos para exibir um número máximo de colunas e linhas.*
4. *Leitura dos Dados: Lemos um arquivo de dados a partir de um caminho específico no Google Drive.*
5. *Checagem das Primeiras Linhas: Exibimos um trecho do DataFrame com as primeiras linhas dos dados.*
6. *Verificação de Variáveis: Imprimimos as colunas presentes nos dados.*
7. *Verificação de Valores Nulos: Contabilizamos a quantidade de valores nulos em cada coluna e exibimos as colunas que continham valores nulos.*
8. *Informações sobre Tipos de Dados: Obtivemos informações sobre os tipos de dados presentes no DataFrame e a contagem de valores não nulos.*
9. *Verificação de Inconsistências em Colunas Numéricas e de Texto: Exibimos estatísticas descritivas para as colunas numéricas e de texto, permitindo identificar inconsistências.*
10. *Verificação de Duplicatas: Verificamos a presença de linhas duplicadas no DataFrame e exibimos as linhas duplicadas, bem como contamos o total de linhas duplicadas.*
11. *Verificação de Dados Nulos: Realizamos uma verificação final da quantidade total de dados nulos em todo o DataFrame.*
12. *Salvamento dos Dados Limpos: Salvamos os dados limpos em um novo arquivo CSV no Google Drive.*

*Essas etapas garantiram que os dados estivessem prontos para análise subsequente, após a identificação e tratamento de problemas como valores nulos, inconsistências e duplicatas.*

➔ ***Data Exploration (Exploração de Dados)***:

*No código referente à etapa de "Data Exploration (Exploração de Dados)", executamos uma série de ações essenciais. Começamos importando bibliotecas e lendo dados previamente limpos. Em seguida, criamos a coluna 'match_id' para agrupar os dados com base nas colunas 'map', 't_score' e 'ct_score', o que facilitou a análise. Selecionamos um conjunto específico de colunas para visualização e exibimos os valores únicos na coluna 'time_left'. Realizamos a conversão da coluna 'time_left' para um tipo numérico e a arredondamos para intervalos específicos. Para garantir a consistência, removemos duplicatas na coluna 'time_left' para cada 'match_id' e 'round'. Em seguida, plotamos gráficos de barras que mostram a contagem de partidas ganhas por equipe em cada mapa. Além disso, criamos gráficos que representam a porcentagem de uso de armas pelos times T e CT em partidas. Para preparar os dados para a modelagem, identificamos colunas não numéricas, aplicamos o LabelEncoder a essas colunas e normalizamos as colunas restantes usando o RobustScaler. Em seguida, treinamos um modelo RandomForest Classifier e avaliamos a importância das features no modelo. Finalmente, selecionamos as top features com base em sua importância, contribuindo para a preparação dos dados para análises posteriores e modelagem de machine learning.*

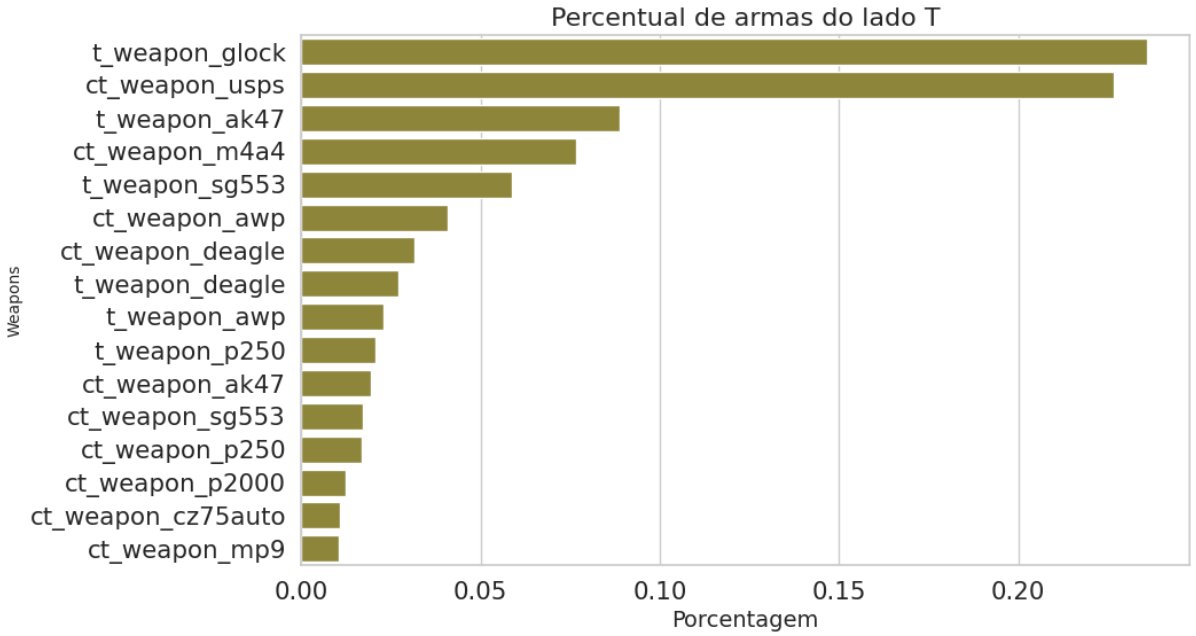
I. ***Gráfico contagem das vitórias de cada equipe em diferentes mapas do jogo***. *Objetivo: proporcionar uma visão visual das tendências de vitórias em contextos específicos.*

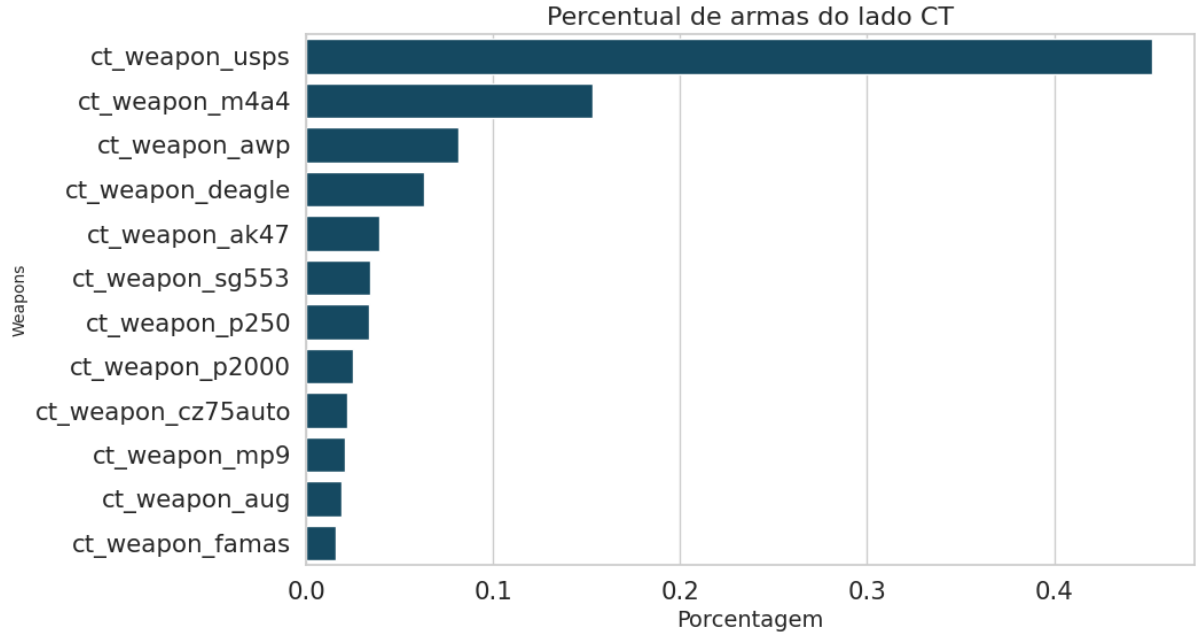



II. ***Distribuição das porcentagens de uso de armas e granadas em partidas.***

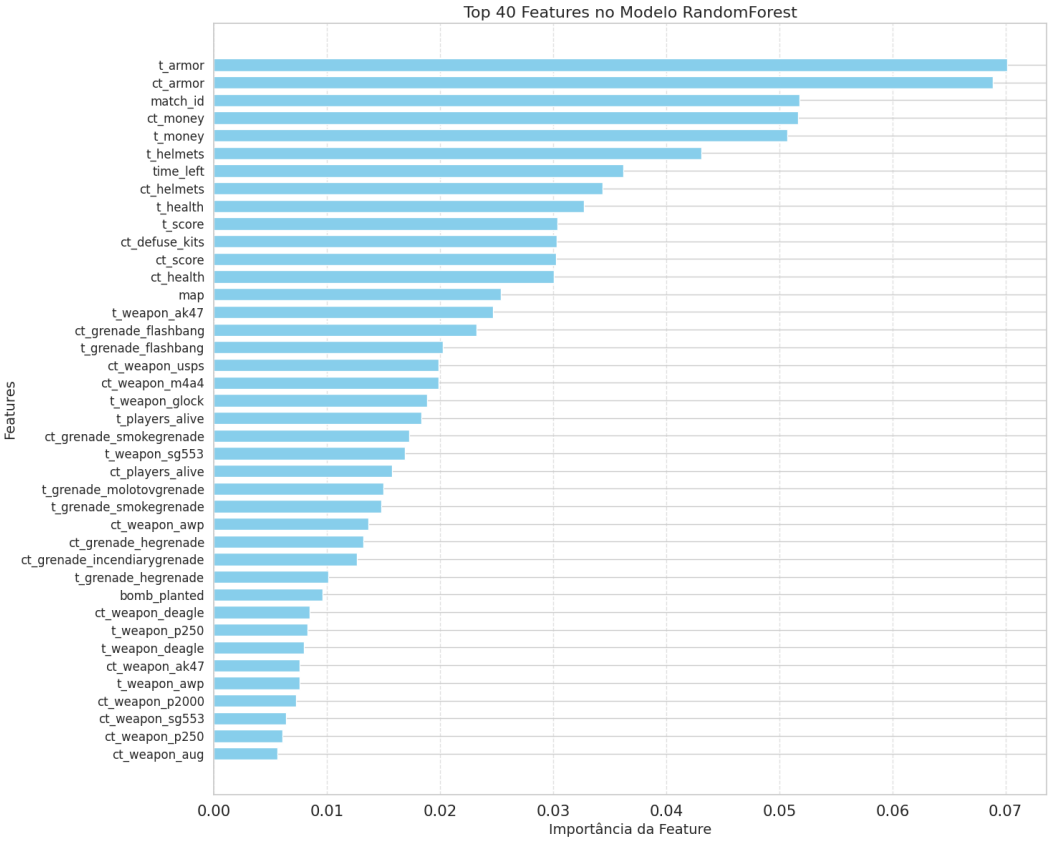
*Os gráficos representam a distribuição das porcentagens de uso de armas em partidas de CS:GO para cada lado (T e CT). Construído a partir do conjunto de dados que inclui informações sobre partidas, jogadores vivos, pontuações e armas utilizadas.*

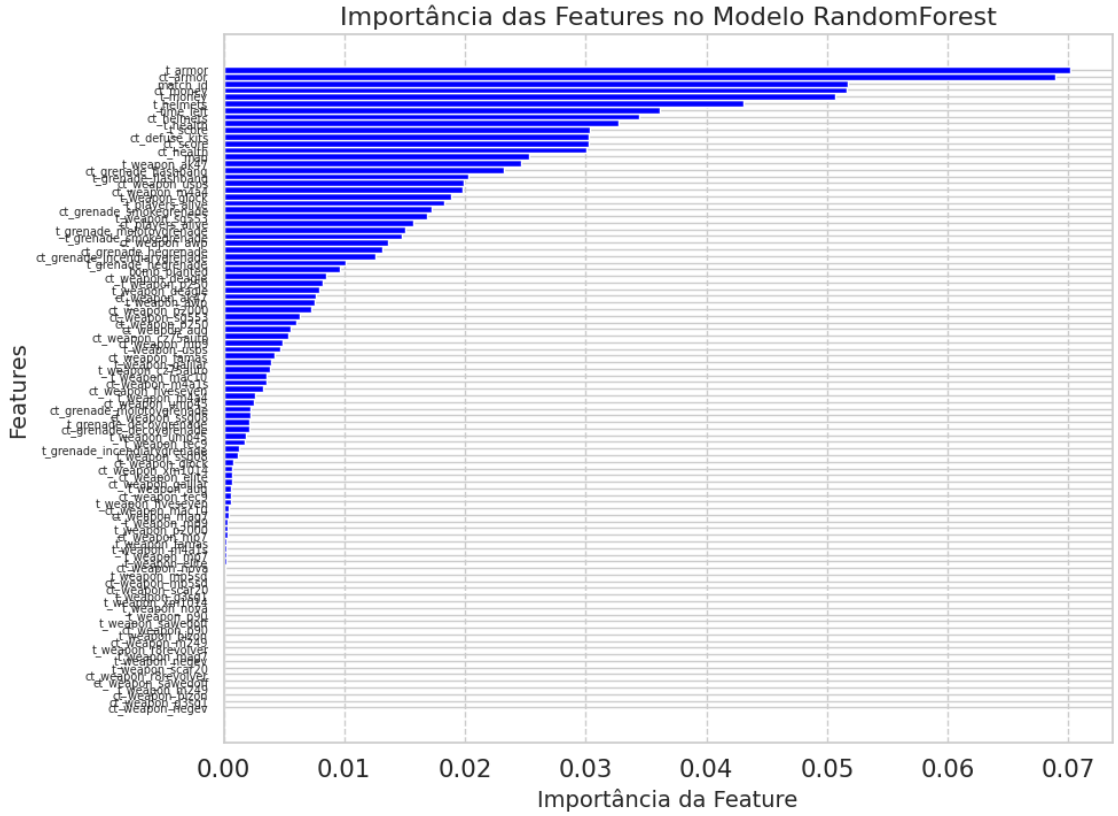
*Em cada gráfico, o eixo horizontal mostra a porcentagem de uso de cada arma, enquanto o eixo vertical exibe as diferentes armas. Cada barra representa uma arma específica, com a altura indicando a porcentagem de vezes que a arma foi usada em relação ao total de armas do lado (T ou CT). O propósito do gráfico é oferecer uma visão visual das preferências de armas em contextos de jogo distintos (T e CT).*




❖ *Realizamos a avaliação da importância das features utilizando o algoritmo Random Forest. O Random Forest é um método de aprendizado de máquina que combina várias árvores de decisão independentes, proporcionando vantagens como redução de overfitting e identificação das variáveis mais importantes. Para preparar os dados, convertemos colunas não numéricas, escalonamos os dados e treinamos o modelo com as features e rótulos. A análise da importância das features auxilia na compreensão das variáveis mais influentes no modelo, apoiando a tomada de decisões.*




*A ordem das importâncias das features é calculada pelo algoritmo do modelo Random Forest. O Random Forest é um ensemble de árvores de decisão que atribui importâncias às features com base em sua contribuição para a redução da impureza nos nós durante o treinamento. Isso é feito calculando a diminuição da impureza média ponderada em cada árvore, considerando como uma feature afeta a mistura de classes no conjunto de dados. As features são ordenadas de acordo com suas importâncias calculadas em ordem decrescente, destacando aquelas que mais contribuem para a redução da impureza nos nós durante a construção das árvores.*

4. **Conclusões:**
    - **Conclusões parciais após o pré-processamento**

*Até o momento, concluímos com êxito as etapas iniciais do projeto, que incluíram a compreensão do negócio, a coleta e avaliação de dados, bem como a limpeza e exploração desses dados. Essas etapas são fundamentais para garantir que os dados estejam prontos para análises subsequentes e modelagem de aprendizado de máquina.*
*Para os próximos passos do projeto, planejamos aprofundar nossa análise, explorar padrões mais complexos e buscar insights adicionais. Estamos empolgados com a perspectiva de aplicar técnicas avançadas de ciência de dados e continuar a desenvolver nosso modelo de previsão, com o objetivo de fornecer informações úteis para jogadores e equipes de eSports no cenário do CS:GO.*
*Este projeto representa um compromisso contínuo com a exploração e compreensão dos dados do CS:GO, e estamos ansiosos para compartilhar nossos resultados à medida que avançamos para as próximas etapas da análise de ciência de dados.*

5. **Próximos Passos:**

*Algumas sugestões e caminhos possíveis que podemos considerar para avançar em nosso projeto de Ciência de Dados:*

- *Exploração Avançada dos Dados: Podemos explorar ainda mais os dados para identificar tendências e padrões complexos entre as variáveis. Isso nos ajudaria a entender melhor como os fatores individuais afetam o resultado das rodadas.*
- *Interpretação de Resultados: Seria interessante analisar cuidadosamente os resultados para entender quais características estão realmente influenciando as vitórias das equipes. Isso poderia revelar insights valiosos para jogadores e equipes de eSports.*
- *Visualizações Avançadas: A criação de visualizações mais complexas e interativas nos ajudaria a comunicar resultados de forma mais eficaz, tornando a análise mais acessível para um público mais amplo.*

*Lembrando que essas são sugestões e possibilidades que podemos considerar à medida que avançamos no projeto. A decisão final sobre qual caminho seguir dependerá dos resultados e das descobertas que faremos ao longo da jornada.*

6. **Referências:**

*Documentação do Scikit-Learn :*

*SCIKIT-APRENDA. Florestas aleatórias no Scikit-Learn. Disponível em: https://r.search.yahoo.com/_ylt=AwrEsaz6czJlmI0JVBHz6Qt.;_ylu=Y29sbwNiZjEEcG9zAzEEdnRpZAMEc2VjA3Ny/RV=2/RE=1697834107/RO=10/RU=https%3a%2f%2fisnca.org%2fpt%2falgoritmo-florestal-aleat%25C3%25B3rio-com-python-e-scikit-learn%2f/RK=2/RS=sQICEFIdxl9MveCk0jIyo_CCjxA-*

*SCIKIT-APRENDA. Métodos de conjunto no Scikit-Learn. Disponível em: https://r.search.yahoo.com/_ylt=AwrNZB4IdDJln_4JlW3z6Qt.;_ylu=Y29sbwNiZjEEcG9zAzEEdnRpZAMEc2VjA3Ny/RV=2/RE=1697834120/RO=10/RU=https%3a%2f%2fawari.com.br%2fscikit-learn%2f/RK=2/RS=bKIyVq.FIVg.tqrK9VF.TQCPEwY-*

*Artigos Acadêmicos :*

*BREIMAN, L. Florestas aleatórias. Aprendizado de máquina, v. 1, pág. 5-32, 2001. Disponível em: https://r.search.yahoo.com/_ylt=AwrEoeU.dDJlZR4K.RTz6Qt.;_ylu=Y29sbwNiZjEEcG9zAzEEdnRpZAMEc2VjA3Ny/RV=2/RE=1697834175/RO=10/RU=https%3a%2f%2fwww.dca.fee.unicamp.br%2f~lboccato%2ftopico_10.1_random_forest.pdf/RK=2/RS=xLcFEELShbaPnVJZdDDsPT4C088-*

*Livros :*

*GÉRON, Aurélien. Aprendizado de máquina prático com Scikit-Learn, Keras e TensorFlow. 2019.* ***Alta Books****.. ISBN:* ***8550803812.***

*MÜLLER, Andreas C.; GUIDO, Sara. Introdução ao aprendizado de máquina com Python. 2016.* ***O'Reilly Media.*** *ISBN:* ***1449369413.***