**RESUMO EXPANDIDO ESTRUTURADO:**

(Tipo: Estudo Original, Revisão, Relato de Caso)

**IAmon: Um Modelo SVM para Classificação de Pokémons Comuns, Lendários, Sub-lendários e Míticos**

IAmon: An SVM Model for Classifying Common, Legendary, Sub-legendary, and Mythical Pokémon

IAmon: Un Modelo SVM para la Clasificación de Pokémon Comunes, Legendarios, Sub-legendarios y Míticos

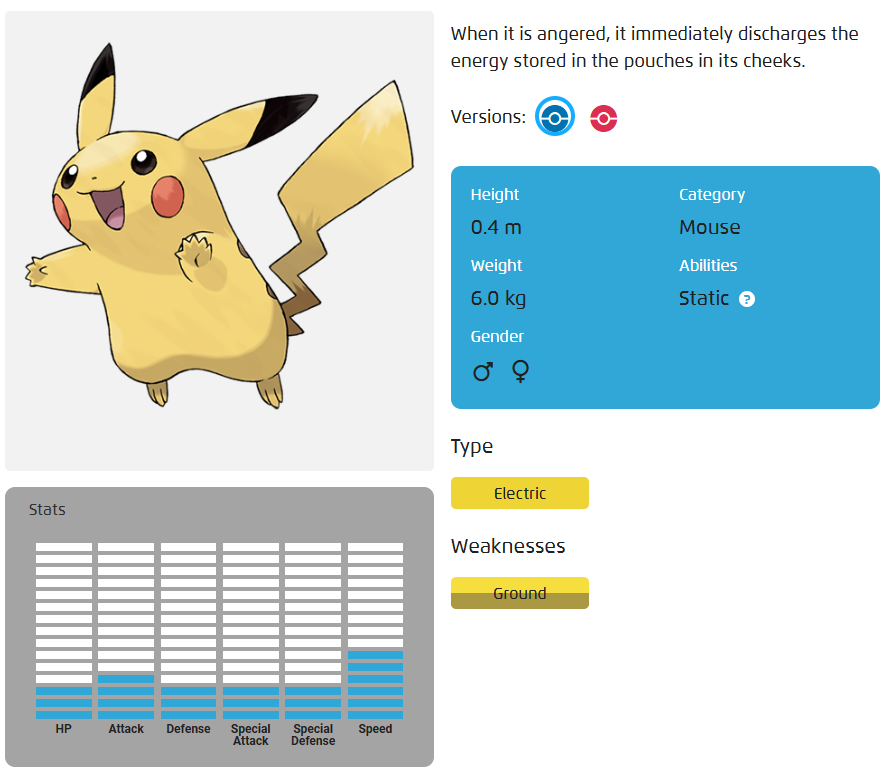
Henry Prado Geraldes¹, Felipe Gomes Amorim¹, Guilherme de Almeida Laurentino¹, Camylla Emy Tahara¹, Gustavo Silva Leão¹, Murilo de Souza Sanches¹, Márcio Pereira Araújo[[1]](#footnote-1), Higor Barreto Campos ¹\*, Fabiano Bezerra Menegidio¹.

**INTRODUÇÃO**

A franquia Pokémon, desenvolvida pela The Pokémon Company, é uma das mais duradouras e populares no mundo dos videogames e da cultura pop, sendo amplamente conhecida por suas criaturas fictícias, cada uma com características únicas. Desde o lançamento da série, a classificação de Pokémons entre comuns, lendários, sub-lendários e míticos tornou-se um aspecto importante tanto para os jogadores quanto para a narrativa do jogo (1). Com o crescimento do aprendizado de máquina, tornou-se possível automatizar essa classificação, oferecendo uma abordagem sistemática e escalável para a identificação de padrões em conjuntos de dados extensos e complexos (2).

O aprendizado de máquina supervisionado, que utiliza algoritmos treinados em dados rotulados, é uma técnica eficiente para tarefas de classificação, incluindo a categorização de personagens com múltiplas características como é o caso dos Pokémons, como ilustrado na Figura 1. Entre os métodos mais aplicados, destaca-se a Máquina de Vetores de Suporte (SVM), devido à sua eficácia em problemas de classificação binária e multiclasse, sua capacidade de lidar com dados de alta dimensionalidade e sua robustez na construção de fronteiras de decisão otimizadas (3-5).

**FIGURA 1.** Pikachu e suas características como altura, peso e suas estatísticas.



**Fonte:** https://www.pokemon.com/br/pokedex/

Este estudo utiliza o modelo SVM para classificar automaticamente Pokémons em diferentes categorias (comuns, lendários, sub-lendários e míticos) com base em dados extraídos de um conjunto de dados disponibilizado pelo Kaggle, que contém atributos detalhados de cada Pokémon, incluindo altura, peso, pontos de saúde, ataque e defesa especiais, e velocidade. O objetivo é verificar a eficácia do SVM na categorização, identificar as características mais relevantes e oferecer uma base técnica para estudos futuros de classificação automatizada em séries de dados similares.

**OBJETIVO**

Este trabalho teve como objetivo desenvolver e avaliar um modelo de aprendizado de máquina, baseado no SVM, para classificar Pokémons em categorias específicas (comuns, lendários, sub-lendários e míticos), analisando as variáveis que influenciam na classificação e medindo a eficácia do modelo em tarefas de categorização multiclasse.

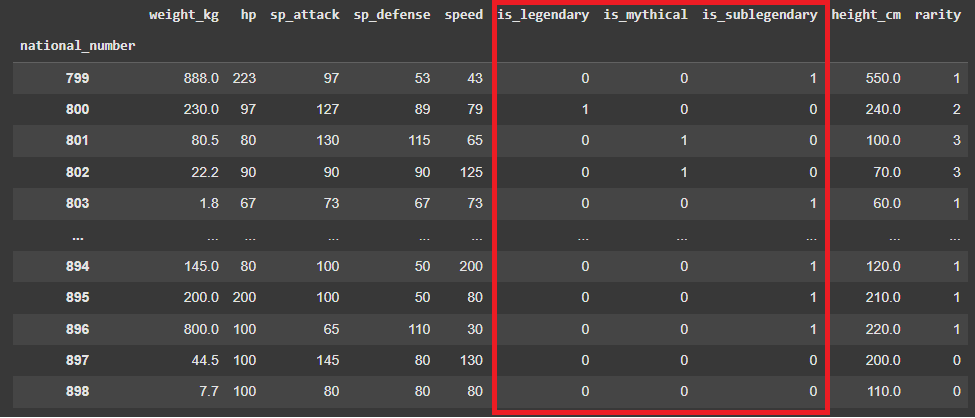
**MATERIAL E MÉTODOS**

**Coleta e Preparação de Dados**

O dataset utilizado foi obtido do Kaggle e contém 801 registros com 10 atributos principais, entre eles: altura, peso, pontos de saúde, ataque especial, defesa especial, velocidade e uma variável indicadora de Pokémon lendário. A plataforma Kaggle tem se destacado pela qualidade dos dados oferecidos para aprendizado de máquina, sendo amplamente utilizada para projetos de ciência de dados e competições de modelagem preditiva (6,7).

Para preparar o dataset, algumas etapas de processamento de dados foram realizadas. Primeiramente, valores inconsistentes ou nulos foram removidos. Em seguida, a altura foi convertida de metros para centímetros para padronizar as unidades, e as variáveis categóricas foram codificadas utilizando a técnica de Label Encoding, Figura 2, transformando-as em variáveis numéricas, o que facilita a análise pelo SVM (8,9). A divisão entre dados de treino (80%) e teste (20%) foi feita usando a função train\_test\_split, com uma semente fixa para garantir a reprodutibilidade dos resultados (10).

**FIGURA 2.** Técnica de Label Encoding para transformação da variável raridade para variável numérica.



**Fonte:** Henry Prado Geraldes et al.,2024

**Configuração do Modelo e Procedimento de Análise**

O modelo SVM foi selecionado devido à sua eficácia em problemas de classificação, especialmente com dados de alta dimensionalidade, como no caso dos atributos dos Pokémons. O modelo foi treinado utilizando o kernel linear e otimizado com a função GridSearchCV, que explora diferentes combinações de hiperparâmetros para maximizar a precisão do modelo (11). Foram ajustados parâmetros como a margem máxima do hiperplano (C) e o gamma (influência de cada ponto de coordenada). Após a otimização, os melhores parâmetros foram definidos como C = 1 e gamma = 0,1 com kernel radial (RBF), o que melhorou o desempenho do modelo para a tarefa de classificação multiclasse (12).

**Avaliação do Modelo**

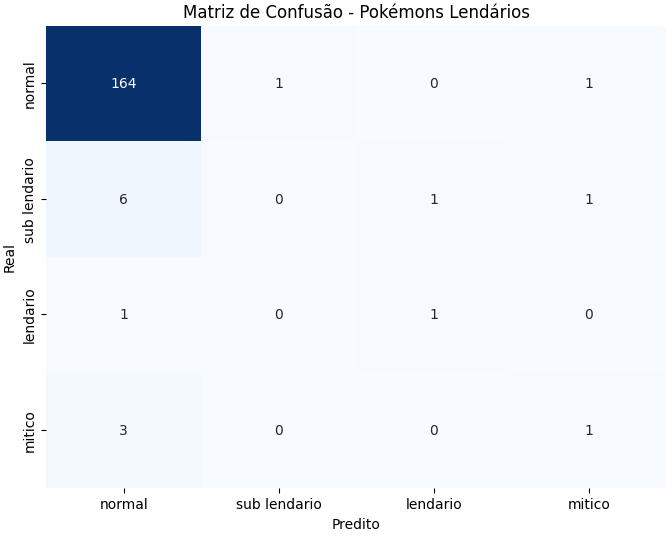
Para medir a precisão do modelo, foram utilizadas métricas de avaliação padrão como acurácia, precisão, recall e F1-score. A matriz de confusão foi utilizada para avaliar a performance do modelo, comparando as previsões com as classes reais e identificando erros de classificação. A análise de importância por permutação foi realizada para determinar a influência de cada variável na acurácia do modelo, auxiliando na identificação das variáveis mais impactantes na classificação (13-15). A validação cruzada foi aplicada com cinco divisões de teste para evitar a dependência de um único conjunto de treino e proporcionar uma avaliação mais robusta (16).

**RESULTADOS E DISCUSSÃO**

**Classificação de Pokémons**

Os resultados obtidos com o modelo SVM demonstram um bom desempenho na classificação de Pokémons, especialmente nas categorias mais comuns, como observado na matriz de confusão na Figura 3. O modelo apresentou uma acurácia geral elevada, com precisão de 0.99 e recall de 1.00 para a classe de Pokémons comuns, indicando que o SVM capturou eficientemente os padrões das variáveis que caracterizam esses Pokémons. Este desempenho é consistente com estudos que sugerem a eficácia do SVM em contextos de classificação com múltiplos atributos e classes bem representadas (3, 5, 10).

**FIGURA 3.** Matriz de confusão de raridade dos Pokémons.



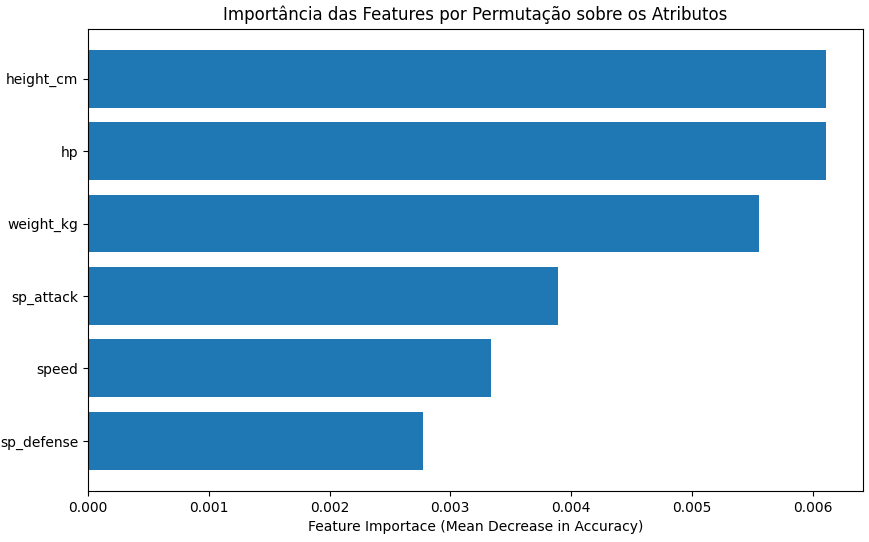
**Fonte:** Henry Prado Geraldes et al.,2024

No entanto, para a categoria de Pokémons lendários, o modelo enfrentou dificuldades, registrando um recall de apenas 0.33. Esse resultado sugere que o modelo teve dificuldade em identificar corretamente os Pokémons lendários, devido à baixa representatividade dessa classe no dataset, caracterizando um problema de desequilíbrio de classes. Em situações de dados desequilibrados, as classes minoritárias tendem a ser menos reconhecidas, levando a um desempenho inferior em termos de recall e precisão para essas categorias (16). A literatura propõe o uso de técnicas como o balanceamento de dados ou métodos de ensemble, como o Random Forest e o Boosting, para minimizar esses efeitos e aprimorar a identificação de classes minoritárias (13, 18).

**Importância das Variáveis**

A análise de importância por permutação revelou que variáveis como pontos de saúde (HP), ataque especial, defesa especial e velocidade foram os principais atributos que influenciaram a classificação correta dos Pokémons, conforme a Figura 4. Esses resultados indicam que características de desempenho em combate têm um papel significativo na diferenciação entre as categorias de Pokémons, especialmente na distinção de classes superiores, como lendários e míticos, que normalmente apresentam estatísticas elevadas nesses atributos (19). Estudos em jogos de estratégia e RPGs indicam que atributos como HP, ataque e defesa são fatores cruciais na definição de status de personagens, o que justifica a relevância dessas variáveis para a classificação de Pokémons (20).

**FIGURA 4.** Importância de cada característica para a classificação da raridade.



**Fonte:** Henry Prado Geraldes et al.,2024

O alto impacto dessas variáveis no modelo também reflete as tendências de design de personagens em jogos de batalha, onde criaturas com características superiores em combate são frequentemente atribuídas a classes raras ou lendárias. Comparado a variáveis como altura e peso, que desempenharam um papel menos significativo, esses atributos de combate demonstram ser determinantes na classificação automática de Pokémons, alinhando-se com outras pesquisas que analisam personagens de jogos eletrônicos e suas categorizações (21, 22).

**Comparação com Outros Modelos**

Embora o SVM tenha se mostrado eficaz, estudos indicam que modelos como Redes Neurais e algoritmos de ensemble, como o Random Forest, poderiam potencialmente oferecer um desempenho superior em tarefas com desequilíbrio de classes, graças à sua maior flexibilidade e capacidade de ajuste dinâmico (11, 17, 23). A literatura reforça que, enquanto o SVM se destaca em problemas de classificação binária e com fronteiras lineares claras, modelos de redes neurais e técnicas de ensemble tendem a se adaptar melhor a fronteiras complexas e a lidar com classes com menor representatividade, como é o caso dos Pokémons lendários (21, 23). A inclusão dessas técnicas em futuras análises poderia melhorar o recall das classes minoritárias, além de oferecer uma perspectiva complementar para o entendimento dos fatores que influenciam a categorização.

**Desafios e Limitações**

O principal desafio enfrentado pelo modelo foi o desequilíbrio de classes, que impactou negativamente o desempenho para a categoria lendária. Esse problema é comumente observado em tarefas de classificação com classes altamente desiguais, onde a classe majoritária domina o aprendizado do modelo (16). O uso de técnicas como o SMOTE (Synthetic Minority Over-sampling Technique) e a aplicação de métricas ajustadas para classes desequilibradas, como a F1-score ponderada, são alternativas recomendadas para melhorar a precisão geral e o recall das classes minoritárias (12, 16). Além disso, o uso de validação cruzada estratificada poderia contribuir para uma avaliação mais precisa do modelo ao considerar a distribuição desigual das classes em cada uma das divisões de validação (14, 15).

Em suma, o modelo SVM mostrou-se adequado para a tarefa de classificação multiclasse de Pokémons, com um desempenho robusto para classes bem representadas. Para otimizar a identificação das classes minoritárias, como a de Pokémons lendários, recomenda-se a incorporação de técnicas de balanceamento de classes e o teste de outros modelos de aprendizado, como redes neurais profundas e métodos de ensemble, que poderiam oferecer melhorias adicionais para a precisão e robustez do modelo.

**CONSIDERAÇÕES FINAIS**

O modelo SVM mostrou-se eficiente para classificar Pokémons comuns, mas apresentou limitações para classes menos representadas, como a de Pokémons lendários, devido ao desequilíbrio de classes. A análise indicou que atributos de combate, como HP e ataque especial, são críticos na diferenciação entre as categorias. Para melhorar o desempenho em classes minoritárias, recomenda-se o uso de técnicas de balanceamento, como o SMOTE, e a experimentação com modelos alternativos, como redes neurais e algoritmos de ensemble. Esses resultados evidenciam o potencial do aprendizado de máquina na classificação automática em contextos de jogos.

**REFERÊNCIAS**

1. Tobin J. Pikachu's global adventure: The rise and fall of Pokémon. Durham: Duke University Press; 2004.
2. Bishop CM. Pattern recognition and machine learning. New York: Springer; 2006.
3. Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995;20(3):273–97.
4. Shalev-Shwartz S, Ben-David S. Understanding machine learning: From theory to algorithms. Cambridge: Cambridge University Press; 2014.
5. Goodfellow I, Bengio Y, Courville A. Deep learning. Cambridge: MIT Press; 2016.
6. Kaggle. The Complete Pokédex Dataset [Internet]. Available from: https://www.kaggle.com/datasets/cristobalmitchell/pokedex.
7. Donoho DL. 50 years of data science. J Comput Graph Stat. 2017;26(4):745–66.
8. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning. New York: Springer; 2009.
9. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine learning in Python. J Mach Learn Res. 2011;12:2825–30.
10. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Proceedings of the 14th international joint conference on Artificial intelligence; 1995.
11. Berrar D. Cross-validation. In: Encyclopedia of bioinformatics and computational biology. Elsevier; 2019.
12. Chang CC, Lin CJ. LIBSVM: A library for support vector machines. ACM Trans Intell Syst Technol. 2011;2(3):27.
13. Breiman L. Random forests. Mach Learn. 2001;45(1):5–32.
14. Stone M. Cross-validation: A review. Math Operations Res. 1974;112(2):111-23.
15. Tibshirani R, Friedman J, Hastie T, et al. The Lasso and generalizations. Boca Raton: CRC Press; 2015.
16. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PloS one. 2015;10(3).
17. Vapnik VN. Statistical learning theory. New York: Wiley; 1998.
18. Freund Y, Schapire RE. Experiments with a new boosting algorithm. In: ICML. 1996.
19. Dietterich TG. Ensemble methods in machine learning. In: International workshop on multiple classifier systems; 2000.
20. Quinlan JR. C4.5: Programs for machine learning. San Francisco: Morgan Kaufmann; 1993.
21. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553):436-44.
22. Liaw A, Wiener M. Classification and regression by randomForest. R news. 2002;2(3):18–22.
23. Cover T, Hart P. Nearest neighbor pattern classification. IEEE Trans Inf Theory. 1967;13(1):21–7.

1. Universidade de Mogi das Cruzes (UMC), Mogi das Cruzes, SP, Brasil.

   **SUBMETIDO EM: XX/2023 | ACEITO EM: XX/2023 | PUBLICADO EM: XX/2023** [↑](#footnote-ref-1)