APRENDIZADO DE MÁQUINA PARA ANÁLISE DE DADOS PROVENIENTES DE EXPERIMENTOS FÍSICOS

Palavras-Chave: CLASSIFICAÇÃO, APRENDIZADO DE MÁQUINA, FÍSICA DE PARTÍCULAS

Autores(as):
YURI ENDERSON NEVES DE OLIVEIRA, FEEC-UNICAMP
Prof. Dr. ROMIS RIBEIRO DE FAISSOL ATTUX (orientador), FEEC-UNICAMP

INTRODUÇÃO

Historicamente, a física de partículas dependeu de experimentos complexos e teóricos para entender os constituintes fundamentais da matéria e suas interações. Nas décadas de 1950 e 1960, a construção dos primeiros aceleradores de partículas de alta energia, como o Bevatron no Lawrence Berkeley National Laboratory, permitiu a descoberta de novas partículas subatômicas e deu início à era moderna da física de partículas [American Physical Society, 2021].

A análise de dados destas partículas envolve a interpretação de *petabytes* de informações geradas por colisões e experimentos em aceleradores. Essa alta disponibilidade de dados, unida a uma capacidade de processamento avançada, conforme discutido por [Barua et al. 2023], propicia soluções inteligentes exponencialmente mais poderosas. Em contrapartida, as técnicas tradicionais de análise de dados, baseadas em algoritmos estatísticos e computação intensiva, são desafiadas pela escala e complexidade dos dados modernos.

O aprendizado de máquina (ML, do inglês *machine learning*) proporciona soluções poderosas para esses desafios. Redes neurais profundas, algoritmos de clustering e técnicas de aprendizado por reforço são exemplos utilizados para identificar padrões sutis em grandes conjuntos de dados, melhorar a precisão das medições e acelerar a descoberta de novas partículas e fenômenos [CERN, 2021].

Este projeto de iniciação científica tem dois objetivos fundamentais: 1) a realização de um estudo formativo abrangente na área de aprendizado de máquina, o qual permitirá ao aluno a construção de uma base sólida no tema e 2) o uso de técnicas de aprendizado de máquina em bases de dados online de aceleradores de partícula, através pelo conjunto de dados de uma simulação de espalhamento inelástico elétron próton medido por um sistema detector de partículas [Harrison,2018]. Inicialmente, no escopo deste projeto trabalharíamos com a base HIGGS disponível em [MLPhysics, 2023], que inclui a classificação de eventos de colisão a partir de diversos atributos complexos [Baldi et al., 2014]; todavia,

sua elevada dimensão trouxe dificuldades de manipulação, o que fez com que optássemos por seguir os dados de [Harrison,2018].

METODOLOGIA

Na primeira etapa deste projeto, abordaram-se os conceitos fundamentais de aprendizado de máquina, com foco nos modelos básicos de regressão e classificação. O material de estudo teve por base as notas da disciplina de pós-graduação "IA048 - Aprendizado de Máquina" [Boccato e Attux, 2020], em conjunto com os livros [Géron, 2023], [Bishop, 2006] e [Goodfellow et al., 2016]. Foram cobertos modelos de aprendizado supervisionado, não-supervisionado e por reforço, assim como métricas e ferramentas analíticas diversas.

Na segunda etapa, buscou-se trabalhar com os dados disponíveis em [Harrison,2018], uma vez que essa base possui um conjunto de amostras de escopo menor quando comparada à base HIGGS. O Conjunto [Harrisson,2018] consiste em uma simulação simplificada contendo dados de 4 partículas: pósitron (-11), píon (211), kaon (321) e próton (2212), com seis respostas do detector, sendo elas: P (momento, em GeV/c), thetha (ângulo em radianos), beta (ângulo em radianos), nphe (Nº de fotoelétrons), ein (energia interna, em GeV) e eout (energia externa, em GeV).

O processo descrito por essa simulação é utilizado para investigar a estrutura interna dos hádrons, especificamente dos prótons. Nele, uma partícula incidente (fotoelétron) colide com um próton alvo, resultando em uma colisão inelástica em que a energia cinética da partícula incidente não é conservada. Durante o espalhamento inelástico, o próton pode se decompor em seus quarks constituintes, que, subsequentemente, formam um jato hadrônico. Os ângulos de deflexão observados fornecem informações sobre a natureza do processo.

Para o tratamento destes dados utilizou-se a linguagem de programação *Python*, bem como as bibliotecas *pandas*, *numpy*, *seaborn*, *matploitlib* e *sklearn*. Num primeiro momento, gerou-se o histograma de cada um dos parâmetros dos dados (Figura 1).

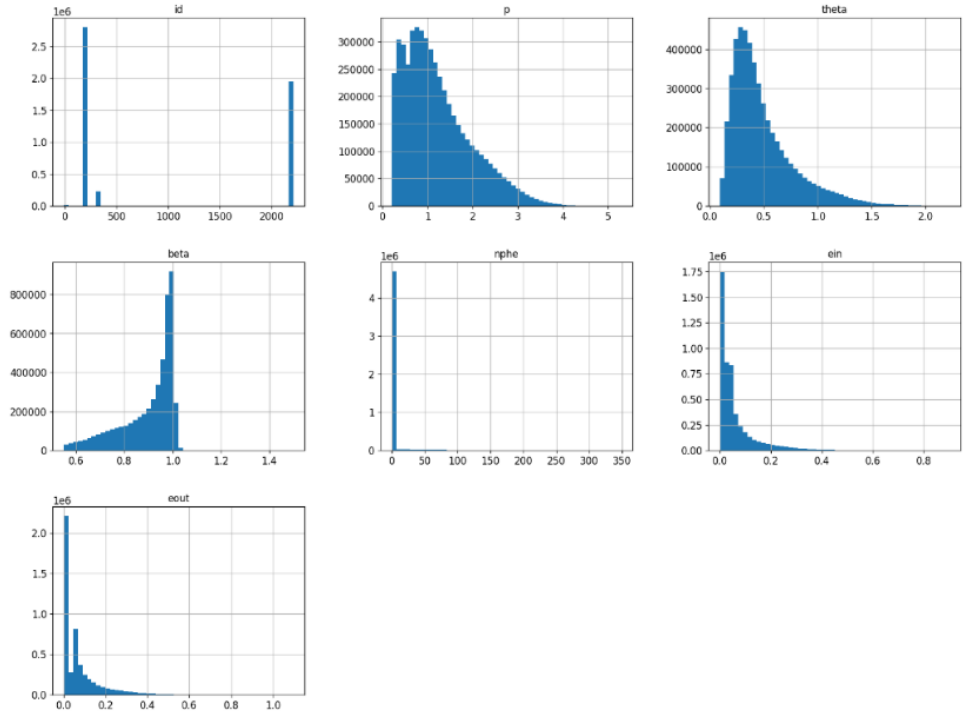


Figura 1: Histograma dos parâmetros.

E em seguida foram criadas as matrizes de dispersão e correlação linear destes parâmetros, ambas indicadas na figura 2 e 3.

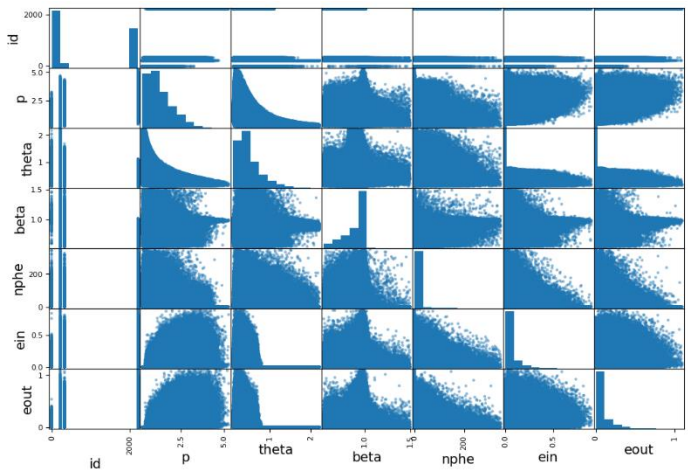


Figura 2: Matriz de Dispersão.

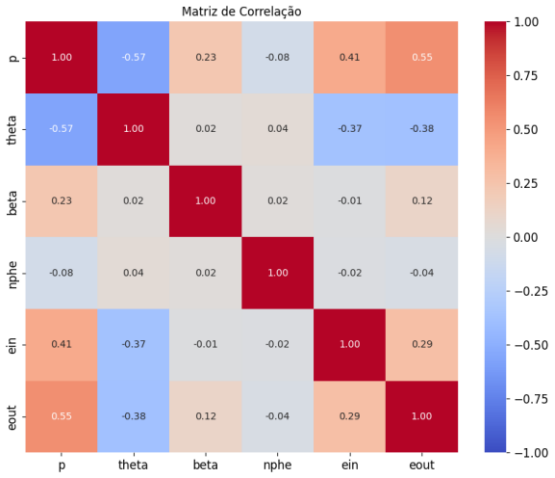
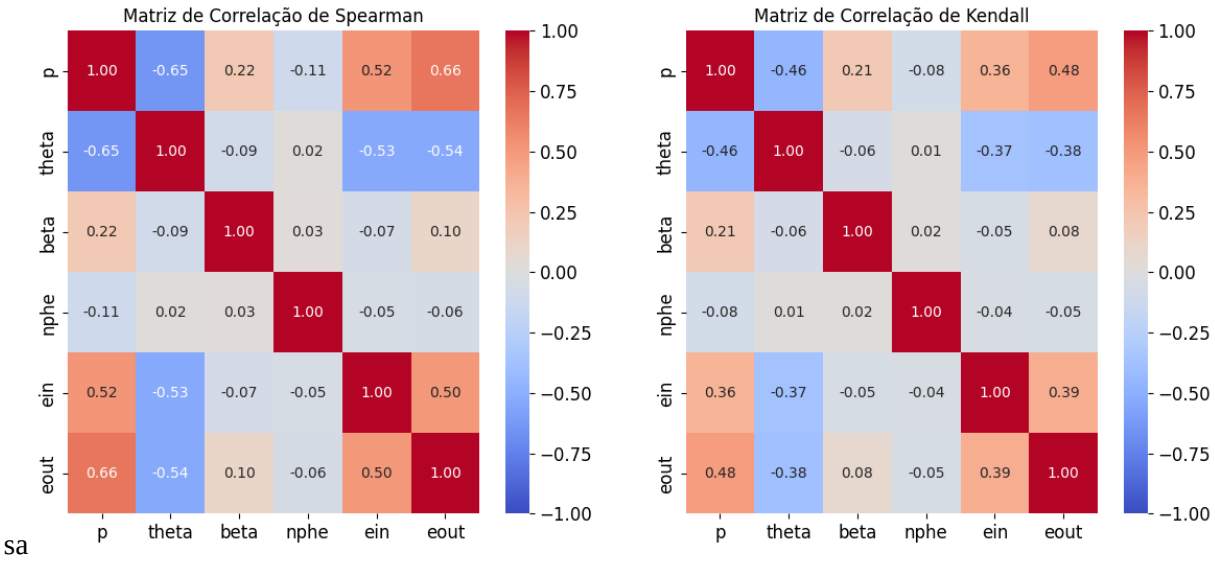


Figura 3: Matriz de correlação linear.

Como não foi observada nenhuma característica que tivesse uma relação de correlação tão evidente, também optou-se por realizar testes como matrizes de correlação de Spearman e Kendall a fim de buscar correlações não-lineares entre os parâmetros. O resultado pode ser visto na figura 4.



sa

Figura 4: Matriz de correlação de spearman e matriz de correlação de Kendall

Posteriormente, iniciou-se a classificação dos dados: o conjunto pid-5M.csv foi carregado e dividido em atributos (X) e classes (Y), tendo como coluno alvo o parâmetro ‘id’. Em seguida, o dataset foi separado em conjuntos de treinamento e teste usando a função "train_test_split" da biblioteca scikit-learn, possibilitando o treinamento do modelo em parte dos dados e a avaliação do desempenho em dados não vistos (com uma proporção de 80% para treino e 20% para teste). Para a classificação foram escolhidos dois modelos específicos: florestas aleatórias e regressão logística.

Para a execução do modelo de floresta aleatória, utilizou-se o a função *RandomForestClassifier com n_estimators* = 5 (valores acima para este conjunto estava causando *overfitting* e um custo de processamento elevado) e para o modelo de regressão logística utilizou-se a função *SGDClassifier,* sendo necessário para esse modelo a normalização para a convergência fosse mais rápida. De forma a tornar ambos modelos robustos, fora utilizada também a validação cruzada através da função *cross_val_score*, da biblioteca Scikit-learn.

Por fim terceira etapa, ainda não finalizada é a realização de outras de técnicas de classificação, como a implementação de uma rede neural profunda e a realização do relatório final, apresentando os resultados definitivos obtidos ao longo da pesquisa.

RESULTADOS E DISCUSSÃO

Ao realizar a comparação entre os modelos de classificação de floresta aleatória e gradiente estocástico, notou-se que o modelo de floresta aleatória se saiu melhor, com acurácia de 96% contra 93% do gradiente. O foco atual agora está em trabalhar essas e demais técnicas de classificação (como uma implementação de uma Rede Neural Profunda) de modo a projetar e implementar uma arquitetura poderosa, capaz de classificar entradas em categorias predefinidas com alta precisão.

CONCLUSÕES

A partir do que foi estudado, foi possível desenvolver um arcabouço de princípios e fundamentos de aprendizado de máquina. Aplicamos técnicas de aprendizado supervisionado e redes neurais para a análise de partículas oriundas de experimentos físicos. Essa jornada proporcionou uma visão abrangente sobre o poder das redes neurais e sua aplicabilidade em tarefas complexas de classificação.

O impacto do aprendizado de máquina na física de partículas é denso, não apenas acelerando descobertas científicas, mas também abrindo novas fronteiras de pesquisa que eram inatingíveis com métodos tradicionais. Esperamos que este trabalho possa, no futuro, estimular novos projetos em redes neurais para aplicações físicas, impulsionando avanços em Inteligência Artificial.

BIBLIOGRAFIA

[American Physical Society, 2021] Berkeley Lab, University of California, San Diego, Sites Recognized for Historical Contributions to Physics, https://www.aps.org/about/news/2021/06/bevatron-mayer-historic-sites, acessado em 19/07/2024.

[Baldi et al., 2014] P. Baldi, P. J. Sadowski, D. Whiteson, “Searching for Exotic Particles in High-Energy Physics with Deep Learning”, Nature Communications, No. 4, 4308, 2014.

[Barua et al., 2023] A. Barua, M. U. Ahmed, S. Begum, “A Systematic Literature Review on Multimodal Machine Learning: Applications, Challenges, Gaps and Future Directions”, IEEE Access, Vol. 11, pp. 14804 – 14831, 2023.

[Bishop, 2006] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.

[Boccato e Attux, 2020] L. Boccato, R. R. F. Attux, *Notas de Aula do Curso IA048 – Aprendizado de Máquina*, FEEC/UNICAMP, 2020.

[CERN, 2021] Key Facts and Figures – CERN Data Centre, https://information-technology.web.cern.ch/sites/default/files/CERNDataCentre_KeyInformation_Nov2021V1.pdf, acessado em 14/05/2023.

[Géron, 2023] A. Géron, *Hands-On Machine Learning with Scikit-Learn, Keras & Tensorflow*, O’Reilly, 2023.

[Goodfellow et al., 2016] I. Goodfellow, Y. Bengio, A. Courville, *Deep Learning*, MIT Press, 2016.

[Harrison,2018].Particle Identification from Detector Responses, https://www.kaggle.com/datasets/naharrison/particle-identification-from-detector-responses/data , acessado em 19/04/2024.

[MLPhysics, 2023] MLPhysics Portal, http://mlphysics.ics.uci.edu/, acessado em 17/06/2024.