TREINAMENTO EM ANÁLISE DE DADOS USANDO FERRAMENTAS DE FÍSICA DE ALTAS ENERGIAS

Palavras-Chave: Análise de Dados, Machine Learning, LHC

Autores(as):

MATHEUS GUILHERME MIOTTO, IFGW – UNICAMP

Prof. Dr. JUN TAKAHASHI (orientador), IFGW - UNICAMP

INTRODUÇÃO:

A Física de Partículas é uma área de escopo da Física que compreende o estudo das partículas elementares e suas interações fundamentais [1], a qual, ao longo das últimas décadas, desenvolveu uma teoria matemática capaz de descrever a interação eletromagnética, fraca e forte entre as partículas, chamada de "Modelo Padrão" [2]. Com base nessa teoria, as partículas fundamentais do universo são classificadas entre dois grupos: férmions e bósons; que são as partículas que constituem a matéria e as responsáveis pelas interações entre elas, respectivamente [3]. Assim sendo, desde a descoberta do elétron em 1897 até, mais recentemente, a descoberta do Bóson de Higgs em 2012, a Física de Partículas sempre manteve-se em constante movimento para promover avanços científicos, revolucionando o conhecimento até então estabelecido.

É dentro desse contexto, portanto, que aceleradores de partículas, como o "Relativistic Heavy Ion Collider" (RHIC) e o "Large Hadron Collider" (LHC), capazes de promoverem colisões de íons pesados a altas energias foram desenvolvidos. Os aceleradores, em busca de novas descobertas, possuem o objetivo de observar sinais de física além daqueles previstos pelo Modelo Padrão. Destacando o "Large Hadron Collider" (LHC), encontra-se o experimento "A Large Ion Collider Experiment" (ALICE). O experimento, que é uma colaboração do complexo "European Organization for Nuclear Research" (CERN), foi projetado para operar investigações tanto em colisões de núcleos atômicos pesados (e.g., Pb-Pb) como em colisões de sistemas próton-próton (p-p), abordando a física da matéria fortemente interativa e o plasma de quark e glúons em valores extremos de densidade de energia e temperatura [4].

Dessa maneira, uma vez que colisões de partículas são constantemente promovidas pelo experimento ALICE, uma quantidade extraordinária de informação é direcionada para a aquisição de dados. Imensas coleções de dados tabulados que possuem informações sobre as características topológicas e cinemáticas das colisões necessitam de uma análise de dados meticulosa. O presente projeto de iniciação científica, portanto, aborda uma análise de dados em cima de dados experimentais disponibilizados pelo grupo: "Grupo de Física Hadrônica Experimental" (HadrEx); em sua direta contribuição com o experimento ALICE. Particularmente, foca-se o interesse em partículas denominadas bárions multi-estranhos (Ξ^-; Ξ^+; Ω^-; Ω^+) e seus subsequentes decaimentos,

processo chamado de “Decaimento em Cascata”. Assim, o conjunto de dados disponibilizado consta com, aproximadamente, 60 milhões de dados simulados de colisões Pb-Pb, abrangendo as diversas variáveis possíveis.

Assim sendo, surge a demanda por métodos eficazes para a análise proposta. A técnica de Machine Learning toma protagonismo por sua aplicabilidade em análises multivariacionais, como os dados experimentais disponibilizados requerem. Logo, uma seleção criteriosa em cima de diversos modelos disponíveis de Machine Learning deve ser feita para obter o modelo adequado. Modelos de Redes Neurais Profundas tomam a iniciativa por apresentarem maior capacidade e eficiência em análises complexas. Especialmente, “Variational Autoencoders” (VAE) [5] e “Conditional Tabular Generative Adversarial Network” (CTGAN) [6] foram utilizados como modelos regenerativos, com o objetivo de gerar dados sintéticos. O projeto, portanto, tem como principal objetivo desenvolver a análise e compreensão completa dos dados simulados de colisões de íons pesados propostos, assim como a aplicação de ferramentas de Machine Learning para a reprodução de dados sintéticos, processo conhecido como “Data Augmentation” [7].

METODOLOGIA:

Com relação a metodologia proposta, pode-se dividi-la em duas partes: teórica e prática. Em um primeiro momento, é necessário que ocorra uma familiarização com os diversos assuntos abordados previamente que compõem o universo de uma análise complexa de dados simulados de colisões de íons pesados do experimento ALICE. Dessa maneira, três áreas foram estabelecidas para estudo, sendo elas: Física de Partículas; Linguagens de Programação e Análise de Dados; Conceitos e Aplicações de Métodos de Machine Learning.

A partir de artigos sobre colisões de íons pesados, previamente selecionados pelo orientador, o aprendizado em cima da Física de Partículas foi promovido. Citam-se os seguintes artigos escolhidos: “QCP Signatures Revisited” [8]; “Strangeness Production in Quark-Gluon Plasma” [9]; “Heavy Ion Collisions at RHIC and the LHC: Physics Challenges” [10]; “The Little Bang in the laboratory: Heavy Ions LHC with ALICE” [11]; “Signatures of quark-gluon plasma formation in high heavy-ion collisions: a critical review” [12].

Em respeito ao aprendizado de linguagens de programação e análise de dados, foi proposto o estudo em cima de duas linguagens de programação: C++; Python. As linguagens de programação foram estudadas a partir de livros didáticos como: “C++ Language Tutorial” [13]; e cursos especializantes como: “CS50's Introduction to Programming with Python” [14]; para as linguagens C++ e Python, respectivamente. Em conjunto, uma análise de dados em cima de dados de pacientes com câncer de cólon foi desenvolvida, em parceria com a Faculdade de Ciências Médicas (FCM) da Unicamp [15], de maneira a sensibilizar o senso crítico de um cientista de dados a presença de possíveis correlações e *insights* existentes entre variáveis disponibilizadas pelos conjunto de dados.

Por fim, aprimorou-se as habilidades de Machine Learning através dos cursos: “Supervised Machine Learning: Regression and Classification”; “Advanced Learning Algorithms”; “Unsupervised Learning, Recommenders, Reinforcement Learning” [16]; oferecidos pela Universidade de Stanford em colaboração com a empresa “DeepLearning.AI”, disponibilizados pela plataforma “Coursera”.

Logo, em um segundo momento, iniciou-se a abordagem em cima do conjunto de dados simulados do experimento ALICE, associado a reconstrução de partículas via medida de decaimentos secundários. Para isso, dividiu-se a análise em duas principais partes: Análise de Dados; Treinamento do Modelo.

Para a análise dos dados utilizou-se métodos estatísticos para a visualização de possíveis correlações entre as variáveis do conjunto de dados, assim como, uma devida limpeza em cima dos dados, evitando dados faltantes ou incondizentes com a física do problema. Assim, realizou-se um devido estudo em cima do conjunto de dados para posteriores decisões serem tomadas, que beneficiaram o treinamento do modelo.

Com respeito ao treinamento do modelo, utilizou-se duas redes neurais profundas para o treinamento: "Variational Autoencoder" (VAE) e "Conditional Tabular Generative Adversarial Network" (CTGAN); para a tentativa de reprodução de dados sintéticos. O autoencoder variacional teve a seguinte arquitetura designada, o *encoder* recebe os dados de entrada que são processados por três camadas convolucionais 2D, com 32, 64, 128 neurônios e (2 x 7), (1 x 5), (1 x 3) *kernels*, respectivamente. Após as primeiras duas camadas convolucionais, uma camada *MaxPooling2D* com (1 x 2) *pool siz*e e *strides* igual a 1 é aplicada. Então, utiliza-se uma camada *Flatten* para achatar a saída, seguida de duas camadas *Dense* com 896 e 14 neurônios, respectivamente. Atribui-se uma camada *Dropout* entre elas, com uma taxa igual a 0,2. Dois vetores bidimensionais são adquiridos representando a média e o desvio padrão dos valores das variáveis do espaço latente. O *decoder* espelha a arquitetura utilizada pelo *encoder*, onde as camadas convolucionais 2D são substituídas por camadas *ConvTrans2D* e as camadas *MaxPooling2D* são substituídas por *UpSamplig2D* com interpolação escolinha como *bilinear*. Parâmetros de *padding* são utilizados como *same* e as funções de ativação *ReLU* são utilizadas por toda rede neural, com exceção para as saídas do *encoder* e *decoder*, onde utiliza-se uma função linear. A função de custo (Loss) trata-se de uma combinação da função de custo de reconstrução (Reconstruction Loss), dada pelo erro quadrático médio, com a função de custo Divergência de Kullback-Leibler (KL-Divergence). O treinamento do modelo utilizou cerca de cinco milhões de dados em dez épocas.

Ademais, a rede CTGAN utiliza-se de duas redes neurais que competem entre si para a criação de dados sintéticos. Utilizando, portanto, a biblioteca "Synthetic Data Vault" (SDV) [17], disponível na linguagem Python, cria-se a rede neural com os seguintes parâmetros: *enforce_min_max_values*: Verdade; *enforce_rounding*: Verdade; *embedding_dim*: 128; *generator_dim*: (256, 256); *discriminator_dim*: (256, 256); *generator_lr*: 0.0002; *generator_decay*: 1e-06; *discriminator_lr*: 0.0002; *discriminator_decay*: 1e-06; *batch_size*: 500; *discriminator_steps*: 1; *log_frequency*: Verdade; *verbose*: Verdade; *pac*: 10; *cuda*: Verdade. O treinamento do modelo utilizou cerca de cinco milhões de dados em 300 épocas.

RESULTADOS E DISCUSSÃO:

Após a aplicação da metodologia, torna-se a atenção para a visualização e discussão dos resultados. O modelo "Conditional Tabular Generative Adversarial Network" (CTGAN) apresentou uma performance excepcionalmente melhor do que o modelo "Variational Autoencoder" (VAE). Ambos os modelos possuíam a proposta de aprender as distribuições, assim como as correlações dos dados simulados de colisões Pb-Pb, mantendo relações físicas das partículas, porém, apenas a CTGAN demonstrou conseguir reproduzir resultados. O autoencoder variacional não conseguiu reproduzir as distribuições, muito menos as correlações entre os dados simulados, retornando dados sintéticos centrados na média das distribuições para cada variável dos dados simulados.

As figuras a seguir, representam os resultados obtidos através do modelo CTGAN:

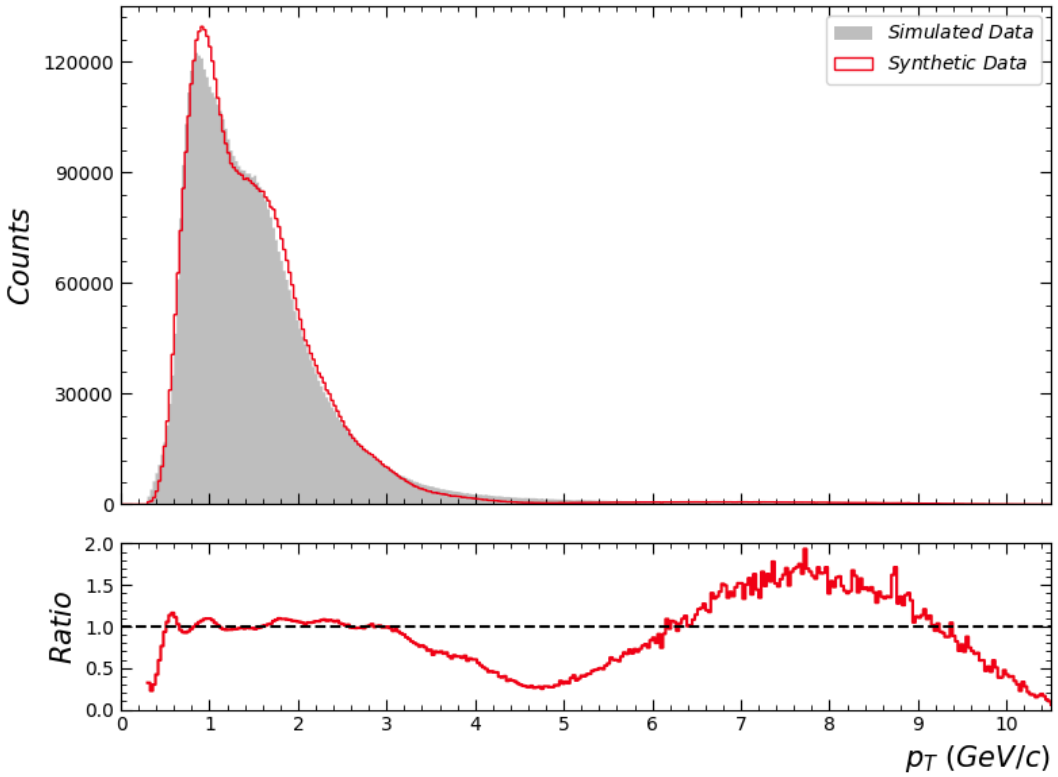

Figura 1: Espectro de momento transverso das partículas (candidatos). Pode-se observar o espectro representado pelo gráfico principal e, abaixo, o gráfico secundário representando a razão entre as distribuições.

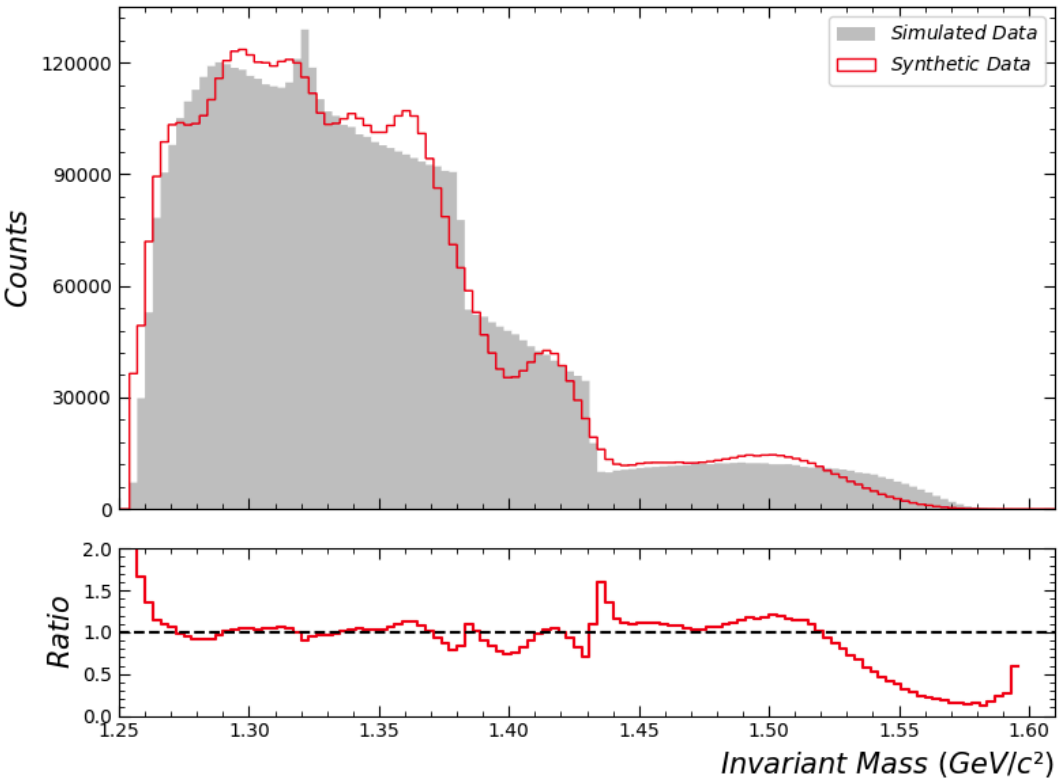

Figura 2: Espectro de massa invariante das partículas (candidatos). Pode-se observar o espectro representado pelo gráfico principal e, abaixo, o gráfico secundário representando a razão entre as distribuições.

CONCLUSÕES:

Com o desenvolvimento desse projeto de iniciação científica, promoveu-se um treinamento qualificado ao aluno em cima de temas como análise de dados e métodos de Machine Learning, envolvendo ferramentas de física de altas energias. Desse modo, tanto conceitos de física de partículas, como meios de análise de dados ou métodos de Machine learning foram estudados e compreendidos. Além disso, o aluno teve acesso a conjuntos de dados simulados de colisões de íons pesados, referentes ao experimento ALICE, disponibilizados pelo grupo "HadrEx". Assim, pode analisar e desenvolver algoritmos de Machine Learning dos mais variados tipos, focando o interesse em modelos regenerativos, com o objetivo de produzir dados sintéticos. Os algoritmos criados tiveram falhas e sucessos, em particular, o modelo CTGAN apresentou resultados satisfatórios até o presente momento, com relação a produção de dados sintéticos. Ao passo que, o modelo VAE fracassou com seu objetivo. Novas investigações devem ser tomadas para entender o motivo da falha do autoencoder variacional em reproduzir os dados sintéticos.

BIBLIOGRAFIA

[1] MARTIN, Brian R.; SHAW, Graham. *Particle Physics*. 4. ed. John Wiley & Sons, 2016.

[2] COTTINGHAM, Noel; GREENWOOD, Derek. *An Introduction to the Standard Model of Particle Physics*. 2. ed. *Cambridge University Press*, 2007.

[3] GRIFFITHS, David. *Introduction to Elementary Physics*. 2. ed. Wiley-VCH, 2008.

[4] The ALICE Collaboration; et al. 2008; JINST 3 S08002.

[5] KINGMA, Diederik P.; WELLING, Max. *An Introduction to Variational Autoencoders*. 2019.

[6] XU, Lei; et al. *Modeling Tabular Data using Conditional GAN*. 2019.

[7] PEREZ, Luis; WANG, Jason. *The effectiveness of data augmentation in image classification using deep learning*. 2017.

[8] HARRIS, John W.; MULLER, Berndt. "QCP Signatures" Revisited.

[9] RAFELSKI; MULLER. Strangeness Production in Quark-Gluon Plasma.

[10] WIEDEMANN, Urs Achim. PANIC, proceedings. Heavy Ion Collisions at RHIC and the LHC: Physics Challenges.

[11] GIUBELLINO, Paolo. The Little Bang in the laboratory: Heavy Ions LHC with ALICE.

[12] BASS, S. A. *et al*. Signatures of quark-gluon plasma formation in high heavy-ion collisions: a critical review.

[13] SOULIÉ, Juan; *C++ Language Tutorial*.

[14] CS50's Introduction to Programming with Python. Disponível em: https://cs50.harvard.edu/python/2022/.

[15] Daniela M.H. Padilha et. al., "Construction of a nomogram for predicting COVID-19 in-hospital mortality: A machine learning analysis" Informatics in Medicine Unlocked 36 (2023) 101138.

[16] Machine Learning Specialization. Disponível em: https://www.coursera.org/specializations/machine-learning-introduction.

[17] SDV - Synthetic Data Vault. Disponível em: https://sdv.dev/.