ANÁLISE EXPLORATÓRIA DO CARHOODS10K: UM CONJUNTO DE DADOS DE NÍVEL INDUSTRIAL

Victor Leão da Silva Dias[1]; Lilian Lefol Nani Guarieiro[2]; Erick Giovani Sperandio Nascimento[3]
[1] Vínculo institucional Mestrando em Modelagem Computacional e Tecnologia Industrial; Mestrado - CNPQ; victor.leao@hotmail.com
[2] Centro Universitário SENAI CIMATEC; Salvador - BA; lilian.guarieiro@fieb.org.br
[3] Centro Universitário SENAI CIMATEC; Salvador - BA; erick.sperandio@fieb.org.br

RESUMO

A análise exploratória dos dados (AED) representa uma fase crítica na manipulação de dados de qualquer natureza, principalmente com a finalidade de utiliza-los para a elaboração de modelos preditivos. Este estudo analisou os dados paramétricos e de desempenho mecânico do conjunto CarHoods10k, através da AED utilizando visualizações e estatísticas descritivas, para entender a distribuição dos dados, identificar ausências e correlações entre os parâmetros e determinar o pré-tratamento necessário para aplicar técnicas de aprendizado de máquina. Os resultados apontaram para a falta de correlação consistente entre parâmetros e resultados mecânicos, e variabilidade nos tipos de parâmetros entre as geometrias, indicando a necessidade de um tratamento cuidadoso dos dados para a construção eficaz de modelos preditivos através de aprendizagem de máquina. Conclui-se que a AED é crucial para preparar dados complexos para o desenvolvimento bem-sucedido de modelos de inteligência artificial.

PALAVRAS-CHAVE: Análise exploratória dos dados, CarHoods10k, Aprendizado de máquina.

1. INTRODUÇÃO

O avanço no design de novas peças e estruturas na engenharia tem sido significativo, impulsionado pelo crescimento do poder computacional. No entanto, ainda enfrentamos desafios devido à complexa interação entre os parâmetros de design e as métricas de desempenho. Nesse contexto, a Inteligência Artificial (IA) se apresenta como uma solução promissora para ultrapassar tais obstáculos, oferecendo a capacidade de lidar com grandes quantidades de dados e agilizando os processos de cálculo numérico, reduzindo drasticamente o alto custo computacional [1,2].

A existência de bases de dados benchmark autênticas e abrangentes é fundamental no avanço de técnicas de aprendizagem de máquina destinadas ao uso industrial. Estes são essenciais para testar e avaliar a eficácia de novas técnicas em cenários complexos de design veicular. Contudo, a dificuldade em obter esses bancos de dados, em grande parte devido a preocupações com a confidencialidade e os elevados custos de produção, apresenta um obstáculo notável. Portanto, se faz necessário produzir esses conjuntos de dados, que devem ser precisos e detalhados, que espelhem fielmente as condições e demandas reais da engenharia aplicada [3].

A Análise Exploratória de Dados (AED) representa uma fase crítica em qualquer pesquisa analítica, servindo para o entendimento dos dados, inspecionar distribuições, identificar outliers e anomalias, e orientar a validação de hipóteses. Atuando imediatamente após a coleta e o pré-processamento dos dados, a AED permite uma manipulação e visualização dos dados, facilitando a avaliação de sua qualidade e a formulação de modelos. A AED prepara o caminho para a elaboração de modelos preditivos robustos, especialmente em aprendizado de máquinas, onde a precisão e a estrutura dos dados são fundamentais para o sucesso do modelo [4].

Assim, o objetivo deste trabalho foi executar uma análise exploratória detalhada no conjunto de dados CarHoods10k, com o intuito de compreender a estruturação dos dados, verificar a presença de quaisquer inconsistências ou falhas e explorar as relações entre os diversos parâmetros. Este processo visa planejar qual seria o tratamento de dados mais adequado para prepará-los para a aplicação de técnicas avançadas de aprendizado de máquina.

2. METODOLOGIA

O conjunto de dados analisado foram apenas os dados paramétricos presente no banco de dados CarHoods10k. O mesmo é composto por 10.070 modelos de capôs de automóveis, divididos em 109 geometrias base diferentes com 100 variações dos parâmetros de design. Os modelos se encontram em 3D (em .STL), juntamente com dados de desempenho mecânico (em .CSV), incluindo tensões equivalentes máximas, deformações direcionais máximas e massa. Esse conjunto de dados foi meticulosamente gerado e validado através de um processo CAD automatizado e análise de elementos finitos (FEA). A metodologia empregada na sua criação envolve a idealização de modelos de capôs, removendo detalhes desnecessários

enquanto mantém características essenciais, seguida pela geração automatizada de variantes de capôs e uma rigorosa validação para assegurar a fabricabilidade dos designs [3].

Para tratar de um conjunto de dados tão grande, uma abordagem sistemática e abrangente é necessária para entender as características subjacentes dos dados, identificar padrões e preparar o terreno para aplicação de algoritmos de aprendizado de máquina. Portanto, uma AED foi aplicada, sendo feita uma visualização e entendimento dos dados, tratamento dos dados duplicados ou ausentes, o cálculo de estatísticas descritivas dos dados numéricos e, por último, a construção de gráficos e análise de correlações multivariável [5].

3. RESULTADOS E DISCUSSÃO

A primeira etapa foi realizar a contagem da quantidade total de amostras, de parâmetros considerados e comparar se os parâmetros variados em cada tabela estavam coincidindo. Foram encontrados 41 dados faltantes em uma planilha, não havendo possibilidade de substitui-los, portanto, essa geometria foi retirada da base. Considerando todas as geometrias, a base de dados possui aproximadamente 50 parâmetros diferentes, e todos eles coincidiam entre as tabelas.

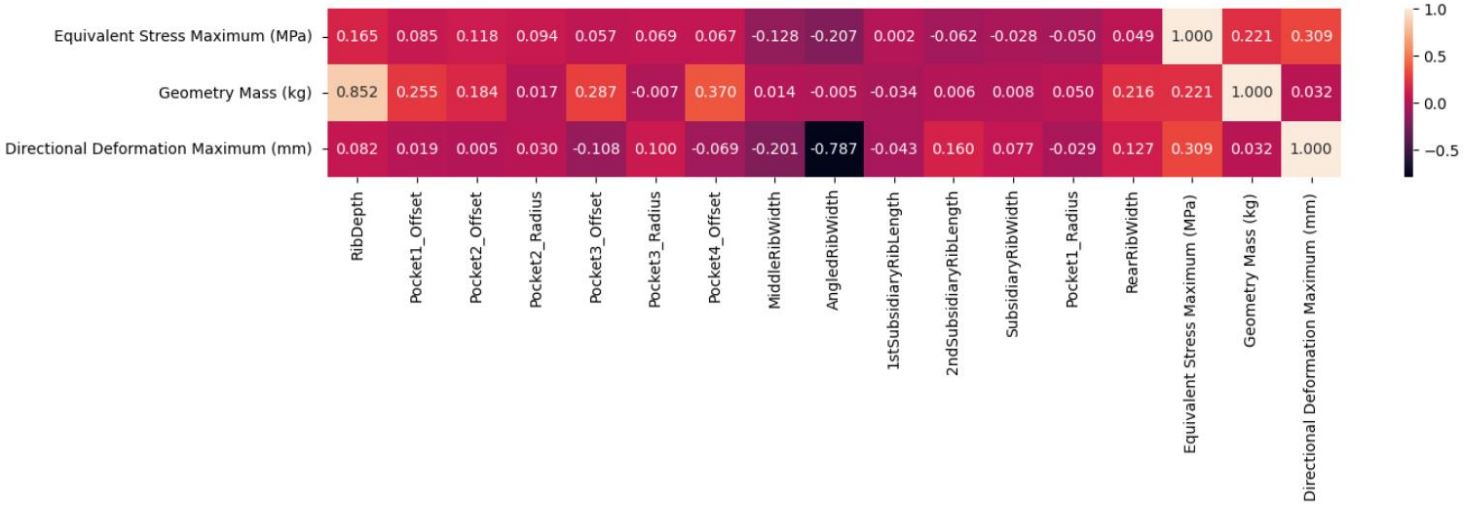
Para a análise exploratória, os dados não foram unificados em uma única base de dados, por haver discrepâncias e variações muito grandes entre eles, tanto no valor como na existência ou não de alguns parâmetros, isso invalidaria a análise de correção e dificultaria a visualização dos dados. Com isso, a análise feita será apenas de uma das geometrias, afim de entender com que tipo de base estamos trabalhando.

Ao avaliar individualmente diferentes tabelas, observa-se que os parâmetros alterados variam de acordo com a geometria do capô. Além disso, há a existência de dados negativos, representando que os dados não necessariamente representam apenas dimensões, mas também posição em relação a algum referencial.

Os gráficos de histograma e dispersão entre os parâmetros de entrada, demonstraram uma distribuição semelhante entre elas, com maior concentração de dados nas extremidades, demonstrando que eles foram criados através de métodos de Design of Experiment (DOE). Já os resultados mecânicos, demonstraram um comportamento contrário, com pouca variação e desvio dos valores de massa e deformação, enquanto a tensão máxima tendeu a ter resultados mais concentrados nas extremidades, ou seja, possuindo valores mais baixos ou mais altos.

A análise de correlação de Pearson, Figura 1, também não demonstrou nenhum padrão de correlação entre os dados, havendo alguns poucos parâmetros demonstrando alta correlação positiva com a massa do capô, provavelmente por se tratar de um resultado diretamente proporcional ao tamanho da peça. As demais correlações não valem a pena serem discutidas por não representar nenhum padrão entre as demais geometrias.

Figura 1 – Mapa de calor da correlação de Pearson entre os parâmetros.



Após feita essa análise exploratória, observa-se uma complexidade em relação a este banco de dados, devido aos diferentes tipos de parâmetros utilizados para cada geometria, as diferentes distribuições entre os parâmetros de entrada e os resultados mecânicos, e a não existência de correlação entre estes. Portanto, se faz necessário avaliar diferentes abordagens de tratamento deste banco de dados, afim de torna-lo factível a sua aplicação em algoritmos de aprendizagem profunda de maneira eficiente. A utilização de, por exemplo, redes neurais nesse conjunto de dados, permitiria uma aceleração na otimização paramétrica no desenvolvimento de novas peças veiculares e permitiria novas abordagens com conjuntos de dados já existentes [2].

4. CONSIDERAÇÕES FINAIS

A análise exploratória realizada no conjunto de dados CarHoods10k revelou insights cruciais sobre a estrutura do banco de dados, incluindo a identificação de dados ausentes e a análise do comportamento dos dados. A complexidade da base de dados se dá pela ausência de correlações consistentes entre os parâmetros e os resultados de simulações mecânicas, além da variação nos tipos de parâmetros em diferentes geometrias. A seleção e o tratamento dos dados serão etapas críticas para o desenvolvimento de um modelo de inteligência artificial capaz de realizar previsões com alta precisão. Portanto, a estratégia de organização e modelagem desses dados se torna fundamental para o sucesso na implementação de um modelo de IA eficaz.

5. REFERÊNCIAS

[1]KOEPPE, A. **Deep Learning in the Finite Element Method**. Dissertation – Institute of General Mechanics, RWTH Aachen University. Alemanha. p. 197. 2021.
[2]VON WYSOCKI, T., RIEGER, F., TSOKAKTSIDIS, D.E., GAUTERIN, F. **Generating Component Designs for an Improved NVH Performance by Using an Artificial Neural Network as an Optimization Metamodel**. Designs, vol. 5, n. 36, 2021.
[3]WOLLSTADT, P., BUJNY, M., RAMNATH, S., SHAH, J. J., DETWILER, D., & MENZEL, S. **CarHoods10k**: An Industry-grade Data Set for Representation Learning and Design Optimization in Engineering Applications. IEEE Transactions on Evolutionary Computation, vol. 26, n. 6, pp. 1221-1235, 2022.
[4]KOMOROWSKI, M., MARSHALL, D., SALCICCIOLI, J., CRUTAIN, Y. **Exploratory Data Analysis**. Secondary Analysis of Electronic Health Records, pp.185-203, 2016.
[5]TUKEY, J. W. **Exploratory Data Analysis**. New York: Addison-Wesley, 1977.