



**MANIPULAÇÃO DE DADOS PARA MODELAGEM EM APRENDIZADO DE MÁQUINA**  
**DATA MANIPULATION FOR MACHINE LEARNING MODELING**LOUREIRO, Arthur Afonso Bitencourt<sup>1</sup>PEREIRA, Gabriella Gomes<sup>2</sup>**RESUMO**

A preparação de dados é vital na aplicação de algoritmos de aprendizado de máquina, garantindo a integridade e coesão dos dados para otimizar o desempenho e obter resultados positivos. Esta etapa é fundamental na criação de modelos de Aprendizado de Máquina, pois os dados são a base de qualquer classificador. Este estudo teve como objetivo o pré-processamento, que inclui a limpeza e remoção de dados irrelevantes, a identificação e correção de erros, o tratamento de valores ausentes e a eliminação de informações desnecessárias, assegurando a qualidade dos dados utilizados nas análises. O processo de manipulação e preparação envolveu a integração de dados de várias fontes, a transformação dos dados para formatos adequados, como normalização ou codificação de variáveis, e a redução da dimensionalidade do conjunto de dados para simplificar a análise e melhorar o desempenho dos algoritmos. A etapa subsequente consistiu na divisão da amostra em conjuntos de treinamento e teste que permite avaliar a capacidade do modelo não apenas com os dados de ajuste, mas também sua habilidade de generalização para novas observações. Uma maior disponibilidade de observações geralmente implica proporções mais elevadas para o conjunto de treinamento, garantindo uma representatividade adequada dos dados no processo de ajuste. Em conclusão, a preparação de dados é fundamental para a qualidade e eficácia das análises em abordagens de aprendizado de máquina, fornecendo uma base sólida para a construção de modelos preditivos.

**PALAVRAS-CHAVE:** Aprendizado de Máquina. Divisão de Amostras. Preparação de Dados.

**ABSTRACT**

Data preparation is vital in the application of machine learning algorithms, ensuring data integrity and cohesion to optimize performance and achieve positive results. This step is fundamental in creating machine learning models, as data is the foundation of any classifier. This study aimed at preprocessing, which includes cleaning and removing irrelevant data, identifying and correcting errors, handling missing values,

---

<sup>1</sup> Professor de Ensino Superior na Universidade Federal de Mato Grosso. Email: arthur.loureiro@sou.ufmt.br

<sup>2</sup> Professora de Ensino Técnico em Enfermagem no Centro de Ensino Unibarra. Email: gabriellapereira728@gmail.com

and eliminating unnecessary information, ensuring the quality of the data used in the analyses. The methodology involved integrating data from various sources, transforming the data into suitable formats, such as normalization or variable encoding, and reducing the dimensionality of the dataset to simplify analysis and improve algorithm performance. The subsequent step consisted of splitting the sample into training and test sets, allowing the model's capacity to be evaluated not only with the fitting data but also its ability to generalize to new observations. A greater availability of observations generally implies higher proportions for the training set, ensuring adequate data representation in the fitting process. In conclusion, data preparation is fundamental to the quality and effectiveness of analyses in machine learning approaches, providing a solid foundation for building predictive models.

**KEYWORDS:** Machine Learning. Sample Splitting. Data Preparation.

## INTRODUÇÃO

O Aprendizado de Máquina está inserido no campo da inteligência artificial, abrangendo o desenvolvimento de algoritmos e modelos com a capacidade de aprender, fazer previsões e tomar decisões fundamentadas em dados (JANIESCH, ZSCHECH, HEINRICH, 2021). Esta metodologia se apoia no pressuposto de que sistemas computacionais podem adquirir conhecimento a partir de exemplos e experiências, ao invés de serem programados de forma explícita.

O principal objetivo do aprendizado de máquina é capacitar computadores a identificar padrões presentes nos dados, realizando previsões ou tomando decisões de maneira autônoma, sem intervenção humana (VERBRAEKEN, 2020). Isto se concretiza através da construção e do treinamento de modelos utilizando algoritmos específicos, os quais têm a habilidade de assimilar informações a partir dos dados de entrada e de generalizar este conhecimento, aplicando-o na formulação de previsões ou na tomada de decisões frente a novos dados.

Existem diversas tipologias de algoritmos de aprendizado de máquina, cada qual com características e aplicabilidades específicas (ESCOVEDO, KOSHIYAMA, 2020). É imperativo ressaltar que o êxito no emprego de aprendizado de máquina é contingente a uma miríade de fatores. Dentre estes, destacam-se a qualidade e representatividade dos dados utilizados no treinamento, a escolha apropriada dos algoritmos e a configuração correta dos modelos (YOUNG et al., 2019).

Para garantir que os dados utilizados em um algoritmo de aprendizado de máquina tenham a qualidade necessária, é fundamental seguir algumas etapas de preparação.

## PREPARAÇÃO DOS DADOS

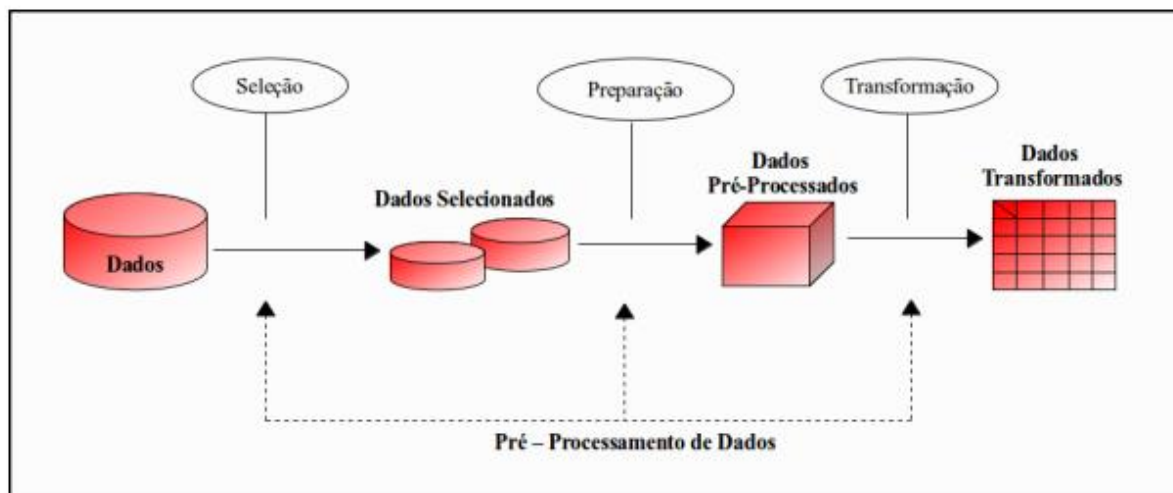
A fase de preparação dos dados é de suma importância na implementação de algoritmos de aprendizado de máquina (*Machine Learning*), conforme destacado por (ESCOVEDO, KOSHIYAMA, 2020), garantindo que os dados apresentados aos algoritmos sejam abrangentes e coesos, com o intuito de otimizar o desempenho e alcançar os melhores resultados possíveis. Conforme ressaltado por Oliveira (2023), o processamento de dados constitui uma etapa imprescindível e recorrente na construção de modelos baseados em abordagens de aprendizado de máquina, uma vez que os dados representam a fundação de qualquer classificador de ML.

A preparação dos dados tem como objetivo a limpeza e eliminação de dados irrelevantes. Para realizá-lo, podemos seguir algumas etapas, conforme descrito por Silva (2021):

- Limpeza de dados: processo que identifica e corrige erros, valores ausentes e dados irrelevantes em conjuntos de dados, assegurando a qualidade necessária para análises de aprendizado de máquina;
- Integração dos dados: processo de combinação de múltiplas fontes de dados em um único conjunto coeso e estruturado, permitindo uma análise mais completa e precisa;
- Transformação dos dados: envolve a conversão de dados brutos para formatos adequados a análises específicas, como normalização e codificação de variáveis;
- Redução dos dados: processo de diminuição da dimensionalidade do conjunto de dados para simplificar a análise e aprimorar o desempenho dos algoritmos de aprendizado de máquina.

A preparação dos dados segue um modelo de pré-processamento, ilustrado na Figura 1, que abrange desde a seleção dos dados até a etapa de processamento, culminando na saída com os dados já transformados.

Figura 1 - Modelo de pré-processamento de dados



Fonte: Juste (2013, p. 20).

Após o pré-processamento dos dados podemos realizar a etapa de divisão de dados.

## DIVISÃO DOS DADOS

Segundo Santos (2018), dividir a amostra em conjuntos de treinamento e teste visa avaliar se o modelo consegue manter um desempenho satisfatório não apenas com os dados usados para seu ajuste (treinamento), mas também quanto à sua capacidade de generalização para novos dados (teste). Normalmente, as proporções utilizadas nessa divisão são de 60:40, 70:30 ou 80:20, sendo que a escolha depende do tamanho do conjunto de dados. De maneira geral, quanto mais observações disponíveis, maior a parte destinada ao treinamento, o que assegura uma representatividade adequada dos dados para ajustar o modelo (RASCHKA, 2017).

A Figura 2 ilustra o roteiro da estrutura de organização para a divisão dos dados em conjuntos de treino, validação e teste.

**Figura 2-** Roteiro para Divisão de dados



Fonte: Adaptado de Raschka (2017).

Essa etapa é fundamental para que o algoritmo consiga validar os dados em relação ao processo de treinamento.

## IDENTIFICAÇÃO DE ATRIBUTOS RELEVANTES

Um método comprovado para a seleção de variáveis e atributos é o *Select K-Best*, uma técnica usada em aprendizado de máquina, disponível na biblioteca *Scikit-Learn* (versão 1.3.2). Seu propósito é selecionar os  $k$  melhores atributos com maior correlação com a variável alvo. Essa correlação é medida por meio de testes estatísticos, como o teste qui-quadrado para variáveis categóricas e o teste F para variáveis contínuas (PRADO, DIGIAMPIETRI, 2020).

As etapas do *Select K-Best* incluem:

- Calcular a pontuação de correlação de cada recurso com a variável alvo;
- Classificar os recursos com base nessas pontuações;
- Selecionar os  $k$  recursos com as pontuações mais altas.

Essa técnica reduz a dimensionalidade dos dados, podendo aumentar a eficiência do modelo e ajudar a evitar o sobreajuste. No entanto, ela não considera

correlações entre os próprios recursos, o que pode ser uma limitação quando essas correlações são relevantes (ROSA DA SILVA, 2022).

## ESCOLHA DO MODELO

Conforme Rauber *et al.* (2023), o aprendizado de máquina foca no desenvolvimento de sistemas capazes de aprender e evoluir com base em suas próprias experiências, sem depender de uma programação explícita, devido à construção de modelos matemáticos e estatísticos baseados nos dados coletados. A escolha do modelo pode ser influenciada por diversos fatores, entre os quais destacam-se (RAUBER *et al.*, 2023):

- **Objetivo da tarefa:** O objetivo da tarefa de aprendizado de máquina é um dos principais fatores que determinam a escolha do modelo. Por exemplo, se a tarefa é classificar e-mails como spam ou não spam, um modelo de classificação seria adequado;
- **Características dos dados:** As características dos dados, como o número de recursos, a presença de valores ausentes, a correlação entre os recursos e a distribuição dos dados, também influenciam a escolha do modelo;
- **Métricas de avaliação:** A métrica de avaliação usada para medir o desempenho do modelo também desempenha um papel crucial na escolha do modelo. A escolha da métrica deve levar em consideração fatores como a proporção de dados de cada classe no *dataset* e o objetivo da previsão (probabilidade, binário, *ranking*, etc);
- **Desempenho do modelo:** O desempenho do modelo em termos de acurácia e velocidade é outro fator importante. Modelos que fornecem alta acurácia e são rápidos para treinar e prever são geralmente preferidos.
- **Facilidade de interpretação:** Modelos que são fáceis de interpretar e explicar são frequentemente preferidos em situações em que a interpretabilidade é importante.

A etapa de preparação de dados influencia diretamente todos esses fatores na aplicação de aprendizado de máquina. Um pré-processamento apropriado garante que os dados sejam adequados ao modelo escolhido e que reflitam de maneira eficaz as variáveis preditivas e a variável alvo.

## CONSIDERAÇÕES FINAIS

Neste estudo, foi demonstrado que a preparação de dados desempenha um papel crucial na eficácia e desempenho de modelos de aprendizado de máquina. Através das etapas de limpeza, transformação, integração e redução da dimensionalidade dos dados, foi possível garantir a qualidade dos dados utilizados, o que resultou em modelos mais eficientes e com maior capacidade de generalização. A divisão adequada dos dados em conjuntos de treinamento e teste foi fundamental para avaliar a capacidade dos modelos em prever novos dados de forma precisa. Além disso, a aplicação da técnica *Select K-Best* contribuiu para a seleção de atributos relevantes, melhorando a eficiência dos modelos ao eliminar variáveis irrelevantes. Dessa forma, os resultados reforçam a importância de um pré-processamento rigoroso e adequado para assegurar a robustez dos modelos e a confiabilidade das previsões realizadas.

Conclui-se que, sem uma preparação cuidadosa e detalhada dos dados, os modelos de aprendizado de máquina não seriam capazes de atingir o desempenho desejado, comprometendo a precisão e a robustez das previsões realizadas. A qualidade dos dados é fundamental para garantir que os algoritmos aprendam de maneira eficiente e generalizem corretamente para novos conjuntos de dados. Caso os dados estejam desorganizados, incompletos ou com ruídos, os modelos podem apresentar resultados distorcidos ou até mesmo falhar em capturar padrões importantes. Assim, a preparação adequada dos dados — que envolve desde a limpeza e normalização até a seleção de atributos relevantes — se mostra essencial não apenas para melhorar o desempenho do modelo, mas também para garantir que as análises sejam consistentes e confiáveis. A importância desta etapa torna-se ainda

mais evidente quando se observa que a maioria dos problemas enfrentados durante o processo de aprendizado de máquina está diretamente relacionada à qualidade dos dados, e não à complexidade dos algoritmos utilizados. Portanto, um pré-processamento eficaz é a chave para o sucesso de qualquer análise de aprendizado de máquina, garantindo resultados que possam ser aplicados de forma prática e com confiança.

## REFERÊNCIAS

ESCOVEDO, Tatiana; KOSHIYAMA, Adriano. Introdução a Data Science: Algoritmos de Machine Learning e métodos de análise. Casa do Código, 2020

JANIESCH, C.; ZSCHECH, P.; HEINRICH, K. Machine learning and deep learning. *Electronic Markets*, v. 31, n. 3, p. 685-695, 2021.

JUSTE, Gleice Ebili. Uma proposta de mineração de dados na base de dados do REDECA utilizando a ferramenta WEKA. 2013. 63 f. Trabalho de Conclusão de Curso (Graduação em Ciência da Computação). Instituto Municipal de Ensino Superior de Assis. Assis, 2013.

OLIVEIRA, Franciana Sokoloski de. Uso de machine learning para a classificação e predição da capacidade de autocura de pasta cimentícia por meio do uso de bactérias [recurso eletrônico]. 63 f., il. color., pdf. Dissertação (Mestrado em Ciência de Materiais) - Universidade Federal de Mato Grosso, Campus Universitário do Araguaia, Barra do Garças, 2023.

PRADO, Fernando F.; DIGIAMPIETRI, Luciano A. A systematic review of automated feature engineering solutions in machine learning problems. In: *Proceedings of the XVI Brazilian Symposium on Information Systems*. 2020. p. 1-7.

RAUBER, Marcelo Fernando et al. Análise do desempenho de aprendizagem de Machine Learning na Educação Básica aplicando a Teoria de Resposta ao Item. In: *Anais do III Simpósio Brasileiro de Educação em Computação*. SBC, 2023. p. 37-48.

RASCHKA, S. *Phyton Machine Learning*. 2. ed. Birmingham: Packet Publishing Ltd, 2017.

ROSA DA SILVA, Maria Eduarda. Seleção de Atributos em aprendizado de máquina para identificação de falha em motores de combustão interna. 2022. 72 f. TCC

(Graduação em Engenharia Mecatrônica.). Universidade Federal de Santa Catarina. Centro Tecnológico de Joinville. 2022.

SANTOS, Hellen Geremias dos. Comparação da performance de algoritmos de machine learning para a análise preditiva em saúde pública e medicina. 206 f. Tese (Doutorado em Ciências). Faculdade de Saúde Pública. Universidade de São Paulo. 2018.

SILVA, Daniel Filipe Baptista Ferreira da. Pré-processamento de Dados e Comparação entre Algoritmos de Machine Learning para a Análise Preditiva de Falhas em Linhas de Produção para o Controlo. 2021. 78 f. Dissertação (Mestrado em Engenharia Informática). Instituto Superior de Engenharia do Porto, Porto, Portugal. 2021.

VERBRAEKEN, Joost et al. A survey on distributed machine learning. Acm computing surveys (csur), v. 53, n. 2, p. 1-33, 2020.

YOUNG, B. A. et al. Can the compressive strength of concrete be estimated from knowledge of the mixture proportions?:New insights from statistical analysis and machine learning methods. Cement and concrete research, v. 115, p. 379-388, 2019.