UNIVERSIDADE FEDERAL DO CEARÁ

CAEN

PEDRO REIS LIMA

PROJETO DE MACHINE-LEARNING I: ANALISANDO A RAZÃO DE TREINO E TESTE EM MACHINE-LEARNING

FORTALEZA

2025

SUMÁRIO

1 INTRODUÇÃO

Em geral, ao se estudar Machine-Learning (ML), os usuários costumam ser orientados a trabalhar com uma razão de treino e teste de 20/80 por exemplo, de forma arbitrária e que é costume na literatura. De fato, esse procedimento é indispensável para evitar viés nos modelos de ML, em especial, para evitar o fenômeno de overfitting (JAMES *et al.*, 2013).

Nesse sentido, parece interessante entender a razão de treino/teste (RTT) como um hiper-parâmetro dos modelos de ML que provavelmente tem impacto nos resultados esperados. Evidentemente, a disponibilidade de dados é uma questão crítica a ser avaliada, os dados de teste não são utilizados para enriquecer o modelo e existe um custo de oportunidade nessa decisão. Especialmente em casos em que a coleta de dados pode ser extremamente custosa, como no desenvolvimento de medicamentos (TAN *et al.*, 2021).

Focando nas aplicações de mercado, a disponibilidade de dados é algo essencial. Por mais que atualmente se trabalhe com bases de dados cada vez maiores, as empresas também modelam seus dados de interesse com uma frequência maior. Por exemplo, se uma industria busca entender a demanda por duas variações de um produto, A e B, é de seu interesse que o melhor produto seja rapidamente identificado. Ou seja, o custo de oportunidade de armazenar dados suficientes para modelagem pode ser elevado a depender do resultado de um produto sobre outro, e é do interesse das firmas minimizar esse custo.

Por isso, o presente projeto busca propor uma análise das implicações de diferentes razões entre teste e treino para modelos de Machine-Learning. Mais especificamente, são avaliados as métricas de Erro Quadrático Médio (EQM) e R^2 (JAMES *et al.*, 2013). O presente projeto buscou trazer resultados preliminares de alguns testes para modelos de Lasso com dados simulados. Além disso, uma breve revisão de literatura foi conduzida.

2 REVISÃO DE LITERATURA

Uma breve revisão de literatura foi conduzida utilizando o repositório do Google Acadêmico. As palavras chaves utilizadas foram "Training and Test Ratio"e "Training and Test Split Ratio". Com isso, 3 artigos representativos da literatura foram selecionados para introduzir a temática.

2.1 Influence of Data Splitting on Performance of Machine Learning Models in Prediction of Shear Strength of Soil (NGUYEN *et al.*, 2021)

Nguyen *et al.* (2021) buscam avaliar e comparar a performance de vários modelos de ML - Artificial Neural Network (ANN), Extreme Learning Machine (ELM), and Boosting Trees (Boosted) - considerando o impacto de diferentes razões de treino e teste. O trabalho focou na análise de solo, avaliando propriedades fundamentais para a construção e design de projetos de engenharia civil. As estatísticas utilizadas foram erro quadrático médio, erro absoluto médio e R^2.

A inovação deste trabalho está em testar a importância da razão treino/teste para trabalhos que avaliam a "solidez"do solo para construções. O uso de ML nessa avaliação não é novo, mas pouco ainda havia se discutido sobre a importância da quebra dos dados. Na verdade, os autores mostram que a literatura de "soil shear strength"já tinham indicativos da importância dessa razão, citando que (PHAM *et al.*, 2020) indicou que, para um pequeno número de opções, aumentar a base de treino tinha implicações positivas para teste e treino até certo ponto.

Em conclusão, os autores ressaltam a importância da qualidade dos dados nos modelos de ML. Em geral, esses modelos tem grande poder preditivo para problemas reais, mas a influência da razão de treino/teste ainda deve ser melhor estudada. Os resultados indicam que a performance dos modelos mudam significativamente dado diferentes RTTs, em que a razão 70/30 indica o nível de separação mais adequado, o que está de acordo com outros artigos em literaturas correlatas a engenharia como predição de deslizamentos.

2.2 Effect of Dataset Size and Train/Test Split Ratios in QSAR/QSPR Multiclass Classification (RÁCZ *et al.*, 2021)

Rácz *et al.* (2021) abordam não só a RTT, mas também o tamanho amostral e seus impactos para modelos de classificação com múltiplas classes. Estes utilizaram 5 algoritmos

diferentes para comparação. Esse trabalho destoa ao trabalhar com classificação de várias classes, fator que introduz variáveis diferentes para análise como por exemplo a necessidade de equilíbrio entre as classes para melhores estimadores. Uma das formas de lidar com casos de "desbalanço"de classes é utilizar modelos de ML com "oversampling"sintético.

Em suma, os autores argumentam que a RTT impactou mais os testes em modelos com maior numero de amostras, o que pode ser uma variável importante para análise, no caso, para modelos de multiplas classes. A sugestão do trabalho indicam a razão de 80/20, especialmente para bases de dados maiores.

2.3 IDEAL DATASET SPLITTING RATIOS IN MACHINE LEARNING ALGORITHMS: GENERAL CONCERNS FOR DATA SCIENTISTS AND DATA ANALYSTS (MURAINA, 2022)

Muraina (2022) ressalta a importância da separação de treino e teste para tratar overfitting. Em sua revisão, ele ressalta que em bases muito grandes, faz sentido utilizar bases de teste menores. O trabalho também menciona os "development sets", utilizado para calibragem de hiper-parâmetros, que também disputariam espaço na amostra completa. Outra quebra possível é entre treino, validação (para selecionar o modelo) e teste. Dessa forma os resultados de teste não enviesariam a análise exploratória do modelo e poderiam testar mais fielmente a reprodutibilidade da análise.

O trabalho buscou analisar as RTTs 50/50, 60/40, 70/30, 80/20 e 90/10 para um modelo de Multilayer perceptron. Aferindo pelas estatísticas acurácia, precisão, recall, erro absoluto médio, raiz do erro quadrático médio, erro absoluto relativo e raiz do erro quadrático relativo. Os resultados não foram lineares na maioria dos indicadores, isto é, existia uma inflexão nos resultados. Por exemplo, para erro quadrático médio os resultados foram, respectivamente, 0.1304, 0.2029, 0.3209, 0.2157 e 0.1301. Esse relacionamento não linear denota como uma especificidade para determinar a razão ótima.

3 METODOLOGIA

A metodologia proposta para o trabalho é quantitativa, rodando modelos de ML para diferentes razões de teste e treino - como 10/90, 20/80, 30/70, 40/60 e 50/50 - e comparando as métricas de MSE e R^2. Para o trabalho final, seria interessante utilizar um algoritmo popular e recorrente na literatura. No momento atual da pesquisa sugere-se os modelos de Random Forest (JAMES *et al.*, 2013), pois são comuns e considerados eficazes em problemas de classificação.

Para o projeto, uma simulação do trabalho foi realizada com um modelo de regressão, o Lasso (JAMES *et al.*, 2013), para dados sintéticos descritos na seção 3.1. Buscou-se estimar o modelo pela biblioteca *glmnet* no R (R-project, 2025a), utilizando validação cruzada com k-folding (JAMES *et al.*, 2013) dado $k = 10$, que é o padrão da biblioteca. Foram testados as razões de treino/teste 10/90 até 50/50 por iteração com "step"1 e calculados respectivos MSE e R^2. Optou-se por Lasso e dados sintéticos pela simplicidade e facilidade de reprodução para validar a ideia do projeto.

Para o trabalho final, as métricas de resultado serão expressadas de forma tanto agregada quanto separada para diversos dados. Essa análise pode permitir explorar hipóteses como a possibilidade de diferentes razões de teste serem viáveis para dados de naturezas diferentes. Esse teste também pode ser replicado para modelos diferentes, indicando que modelos lidam melhor com mais ou menos razão de treino/teste, mas, de forma pragmatica, essa pode ser uma contribuição futura.

3.1 Base de dados

As bases de dados utilizadas para o trabalho final serão as disponíveis no pacote *wooldridge* do software R (R-project, 2025b). O pacote contem diversas bases do livro de introdução a econometria (WOOLDRIDGE, 2020). Serão utilizadas as bases adequadas ao modelo testado (sejam de regressão ou classificação).

Como primeiro passo, uma base com 10.000.000 observações foi simulada no R utilizando variáveis sintéticas, todas descritas por distribuições normais. Como o modelo a ser testado inicialmente é linear, a variável dependente é uma combinação linear das variáveis sintéticas. Estas podem ser sintetizadas da seguinte forma:

- $x_1 \sim \mathcal{N}(0,1)$
- $x_2 \sim \mathcal{N}(0,1)$

- $z_1 \sim \mathcal{N}(0,1)$
- $z_2 \sim \mathcal{N}(0,1)$
- $e_1 \sim \mathcal{N}(0,1)$
- $e_2 \sim \mathcal{N}(0,1)$
- $x_3 = q * z_1 + w * z_2 + e_2$, dado que $q = 1.5, w = 0.5$
- $y = a * x_1 + b * x_2 + c * x_3 + e_1$, dado que $a = 2, b = -1, c = 0.5$

Assim, tem-se uma variável dependente conhecida que se pode estimar por modelos de Lasso. Todas as variáveis são exogenas com excessão de x_3 e y. z_1 foi utilizado como substituição de x_3 para simular a ausência de uma variável que não foi possível acessar. Para reprodução, o código se encontra no apêndice A.

4 RESULTADOS PRELIMINARES

Como mencionado, uma primeira abordagem para o problema foi estimada por Lasso para uma amostra sintética. A forma estrutural utilizada foi:

$$y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 z_1 \tag{4.1}$$

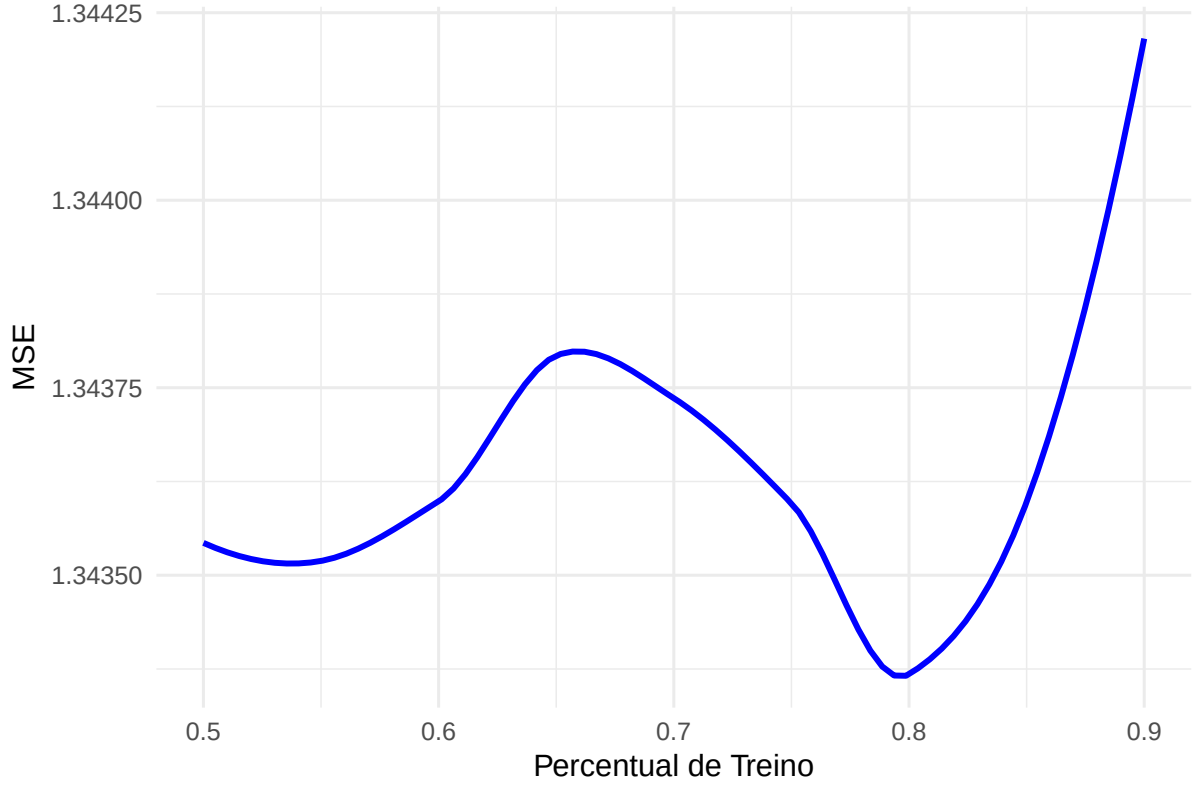
Evidentemente, o ruído não foi utilizado e x_3 foi parcialmente substituída por z_1 de forma a simular os desafios comuns de modelagem de fenômenos estatísticos.

A primeira análise foi a do erro quadrático médio para razões de treino de 0.5 até 0.9, ver Figura 1. Evidenciou-se dois vales na amplitude analisada, com o menor erro em torno de 80% de treino. É de se esperar que ao se aproximar de 100%, a amostra vá apresentar overfitting e o teste tenha maiores disparidades com o modelo. Entretanto, o comportamento de inflexão de que acontece aproximadamente em 65% é inesperado e colabora com o encontrado por (MURAINA, 2022) para modelos de classificação.

Uma crítica necessária a essa análise é que, ao olhar para a magnitude da mudança entres os EQMs, vê-se que a diferença total no gráfico é de 0.00075. Como foi feita uma aplicação simulada, é difícil entender as consequências dessa variação. Evidencia-se, também, a necessidade de introduzir métricas relativas para legibilidade do trabalho. As variáveis relativas também serão importantes para comparação de dados diferentes, que podem ter magnitudes dispares.

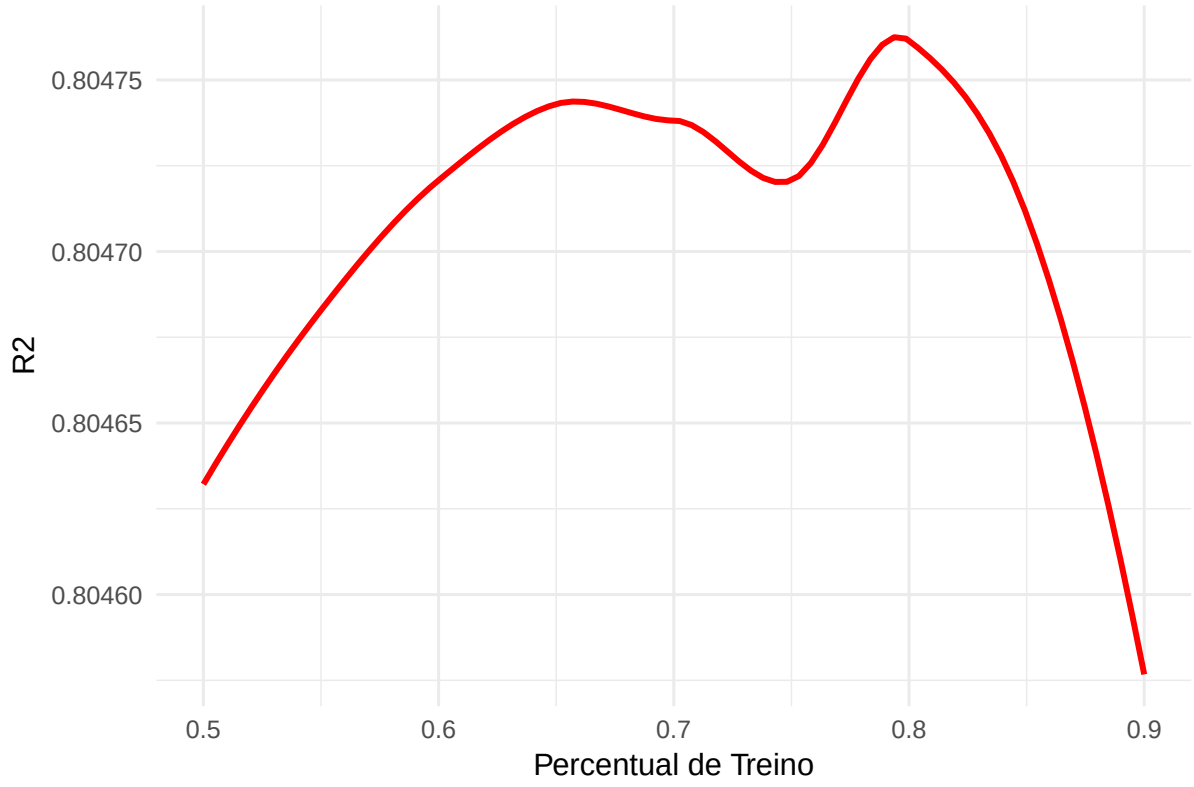
A outra variável analisada, que também teve resultados semelhantes, foi o coeficiente de determinação R^2, veja a Figura 2. O maior poder explicativo foi encontrado com um percentual de treino de 80%, similar ao visto no EQM. Evidentemente, o mesmo processo de overfitting impacta a capacidade do modelo explicar o fenomeno das observações não utilizadas em sua estimação. Também há a existência de uma inflexão em 0.65%, mas parece ter magnitude relativa menor ao que foi observado no EQM.

Figura 1 – EQM dado percentual dedicado ao treino



Fonte: Elaborado pelo autor.

Figura 2 – R^2 dado percentual dedicado ao treino



Fonte: Elaborado pelo autor.

REFERÊNCIAS

JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. *et al.* **An introduction to statistical learning**. [S.l.]: Springer, 2013. v. 112.

MURAINA, I. Ideal dataset splitting ratios in machine learning algorithms: general concerns for data scientists and data analysts. In: **7th international Mardin Artuklu scientific research conference**. [S.l.: s.n.], 2022. p. 496–504.

NGUYEN, Q. H.; LY, H.-B.; HO, L. S.; AL-ANSARI, N.; LE, H. V.; TRAN, V. Q.; PRAKASH, I.; PHAM, B. T. Influence of data splitting on performance of machine learning models in prediction of shear strength of soil. **Mathematical Problems in Engineering**, Wiley Online Library, v. 2021, n. 1, p. 4832864, 2021.

PHAM, B. T.; QI, C.; HO, L. S.; NGUYEN-THOI, T.; AL-ANSARI, N.; NGUYEN, M. D.; NGUYEN, H. D.; LY, H.-B.; LE, H. V.; PRAKASH, I. A novel hybrid soft computing model using random forest and particle swarm optimization for estimation of undrained shear strength of soil. **Sustainability**, MDPI, v. 12, n. 6, p. 2218, 2020.

R-project. **Package 'glmnet'**. 2025. <https://cran.r-project.org/web/packages/glmnet/glmnet.pdf>. Accessed: 2025-03-10.

R-project. **Package 'wooldridge'**. 2025. <https://cran.r-project.org/web/packages/wooldridge/wooldridge.pdf>. Accessed: 2025-03-10.

RÁCZ, A.; BAJUSZ, D.; HÉBERGER, K. Effect of dataset size and train/test split ratios in qsar/qspr multiclass classification. **Molecules**, MDPI, v. 26, n. 4, p. 1111, 2021.

TAN, J.; YANG, J.; WU, S.; CHEN, G.; ZHAO, J. A critical look at the current train/test split in machine learning. **arXiv preprint arXiv:2106.04525**, 2021.

WOOLDRIDGE, J. **Introductory Econometrics: A Modern Approach**. Cengage Learning India, 2020. ISBN 9789355731074. Disponível em: <https://books.google.com.br/books?id=tXqAzwEACAAJ>.

APÊNDICE A – CÓDIGO NO R

```
1
2
3  #Library
4
5  library(glmnet)
6  library(ggplot2)
7
8  #Variables
9
10 #Meta
11
12 n_samples <- 10^7
13
14
15 set.seed(6)
16
17 #Coeficientes populacionais
18
19
20 a <- 2; b <- -1; c <- 0.5
21 q <- 1.5; w <- 0.5
22
23 #Independent
24
25
26 x1 <- rnorm(n_samples, 0, 1)
27 x2 <- rnorm(n_samples, 0, 1)
28 z1 <- rnorm(n_samples, 0, 1)
29 z2 <- rnorm(n_samples, 0, 1)
30 e1 <- rnorm(n_samples, 0, 1)
```

```
e2 <- rnorm(n_samples, 0, 1)

#Dependent


x3 <- q * z1 + w * z2 + e2
y <- a * x1 + b * x2 + c * x3 + e1

#Dataset


df <- data.frame(x1, x2, x3, z1, y)


# dividir treino e teste
train_percentage = 0.8
train_index <- sample(1:n_samples, size = train_percentage
   * n_samples, replace = FALSE)
train <- df[train_index, ]
test <- df[-train_index, ]


X_train <- as.matrix(train[, c("x1", "x2", "z1")])
y_train <- train$y
X_test <- as.matrix(test[, c("x1", "x2", "z1")])
y_test <- test$y

#Basic Modeling


lasso_model <- glmnet(X_train, y_train, alpha = 1, lambda =
    0.1, nfolds = 10)
```

```
61
62
63 y_pred <- predict(lasso_model, X_test)
64
65
66 mse <- mean((y_test - y_pred)^2)
67 r2 <- 1 - sum((y_test - y_pred)^2) / sum((y_test - mean(
       y_test))^2)
68
69
70 cat("MSE:", mse, "\n")
71 cat("R  :", r2, "\n")
72
73 #Ratio testing
74
75
76 df_tests <- data.frame(train_percent = numeric(), MSE =
       numeric(), R2 = numeric())
77
78 for (i in 10:50) {
79   # define o percentual de treino do loop
80   train_percentage = 1 - 0.01*i
81   print(train_percentage)
82
83   # separa a base no treino teste
84   train_index <- sample(1:n_samples, size =
         train_percentage * n_samples, replace = FALSE)
85   train <- df[train_index, ]
86   test <- df[-train_index, ]
87   X_train <- as.matrix(train[, c("x1", "x2", "z1")])
88   y_train <- train$y
89   X_test <- as.matrix(test[, c("x1", "x2", "z1")])
```

```
90    y_test <- test$y
91
92
93    lasso_model <- glmnet(X_train, y_train, alpha = 1, lambda
          = 0.1, nfolds = 10)
94    y_pred <- predict(lasso_model, X_test)
95
96    mse_i <- mean((y_test - y_pred)^2)
97    r2_i <- 1 - sum((y_test - y_pred)^2) / sum((y_test - mean
          (y_test))^2)
98
99    temp <- data.frame(train_percent = train_percentage, MSE
          = mse_i, R2 =  r2_i)
100
101   df_tests <- rbind(df_tests, temp)
102 }
103
104
105 ggplot(df_tests, aes(x = train_percent)) +
106   geom_smooth(aes(y = MSE), color = "blue", se = FALSE) +
          # Linha suave para MSE
107   #geom_smooth(aes(y = R2), color = "red", se = FALSE) +
              # Linha suave para R
108   labs(x = "Percentual de Treino", y = "MSE") +
109   theme_minimal()
110
111
112 ggplot(df_tests, aes(x = train_percent)) +
113   #geom_smooth(aes(y = MSE), color = "blue", se = FALSE) +
          # Linha suave para MSE
114   geom_smooth(aes(y = R2), color = "red", se = FALSE) +
          # Linha suave para R
```

```
115   labs(x = "Percentual de Treino", y = "R2") +
116   theme_minimal()
```