

Métodos de Machine Learning na Produtividade de Soja

Arthur Hintz, Lucas Sartor

27 nov 2025

Contents

1	Resumo	1
2	Introdução	1
3	Conjunto de Dados	2
4	Análise Descritiva	3
5	Metodologia	7
5.1	Aprendizado não supervisionado	7

1 Resumo

Este relatório apresenta o desenvolvimento de um projeto de Machine Learning realizado no âmbito da disciplina, empregando inicialmente métodos não supervisionados e, posteriormente, técnicas supervisionadas. São descritos os objetivos, o conjunto de dados, a metodologia empregada e os resultados obtidos, incluindo análises de sensibilidade com ênfase em validação, regularização e reprodutibilidade. Para a avaliação do desempenho preditivo, adotou-se a abordagem de validação cruzada (cross-validation).

2 Introdução

Segundo a Confederação da Agricultura e Pecuária do Brasil (CNA), o agronegócio representou 23,8% do Produto Interno Bruto (PIB) brasileiro em 2023. Entre as commodities agrícolas, a soja destaca-se como aquela de maior valor de produção no país, conforme dados do IBGE. A produtividade das lavouras sofre influência de diversos fatores, muitos deles incontrolláveis, como condições climáticas. No último ciclo, por exemplo, observou-se um aumento de 70,83% na produtividade em relação ao ano anterior, possivelmente associado ao maior volume de precipitação.

Diante desse contexto, torna-se essencial analisar as regiões do Rio Grande do Sul que apresentam maiores desafios produtivos, agrupando áreas com características semelhantes.

Tal abordagem auxilia na compreensão dos fatores que diferenciam os ambientes de produção e pode apoiar estratégias de manejo mais eficientes.

Inicialmente, realiza-se uma análise descritiva do conjunto de dados para compreender as variáveis envolvidas e suas inter-relações. Em seguida, aplica-se o método k-means para identificar agrupamentos homogêneos (clusters). Posteriormente, estima-se o potencial produtivo de cada cluster por meio de um modelo de regressão.

Os dados utilizados neste estudo foram disponibilizados pela empresa Crops Team e foram coletados a partir de experimentos com cultivares de soja conduzidos em diversas localidades do estado do Rio Grande do Sul durante as safras de 2021/2022 e 2022/2023.

Este estudo é relevante pois permite identificar os principais determinantes da produtividade da soja, fornecendo subsídios para a tomada de decisões agrícolas e para a otimização dos rendimentos das cultivares. Ao compreender melhor os fatores que influenciam a produtividade, produtores e pesquisadores podem implementar práticas agrícolas mais eficientes, maximizando o potencial produtivo das regiões analisadas.

3 Conjunto de Dados

O conjunto de dados original continha informações provenientes dos quatro blocos de cada ensaio. As médias dos blocos por cultivar foram então calculadas, resultando em 1513 observações e 33 variáveis. Após o processo de filtragem e tratamento de valores faltantes, obteve-se o banco de dados final utilizado nas análises.

- **Cultivares:** Informações sobre as diferentes variedades de soja utilizadas nos experimentos.
- **Localização:** Características geográficas dos locais onde os experimentos foram conduzidos.
- **Dados Climáticos:** Informações sobre precipitação, temperatura e outras condições climáticas durante o período dos experimentos.
- **Componentes Químicos do Solo:** Dados sobre a composição química do solo, incluindo níveis de nutrientes e pH.

Além disso:

1. Criou-se a variável dias, representando o número do dia do ano correspondente à data de semeadura.
2. Todas as variáveis não numéricas foram convertidas para fatores (factors) antes da modelagem.

Dessa forma, resultou na seguinte Tabela ?? que apresenta uma amostra aleatória de 10 linhas do conjunto de dados final.

Table 1: (#tab:tab:dados)Banco de Dados.

	Safra	COD_PROD	Local	Cod_Estacao_Met	Altitude	Terras	Ambiente
1332	2022/2023	TA485	TUPANCIRETA	A886	485	ALTAS	IRRIGADO
1207	2022/2023	SM100	SANTA MARIA	A803	105	ALTAS	SEQUEIRO
591	2022/2023	AG117	ALEGRETE	A826	117	BAIXAS	SEQUEIRO
12	2021/2022	AG117	ALEGRETE	A826	117	BAIXAS	SEQUEIRO
631	2022/2023	BA255	BOSSOROCA	A852	255	ALTAS	SEQUEIRO
1255	2022/2023	SV124	SAO VICENTE DO SUL	A889	124	ALTAS	SEQUEIRO
1278	2022/2023	SV124	SAO VICENTE DO SUL	A889	124	ALTAS	SEQUEIRO
1281	2022/2023	SV124	SAO VICENTE DO SUL	A889	124	ALTAS	SEQUEIRO
899	2022/2023	JV579	JACUTINGA	A828	579	ALTAS	SEQUEIRO
867	2022/2023	IR393	IBIRUBA	A883	457	ALTAS	IRRIGADO

Table 2: Data summary

Name	dados
Number of rows	1513
Number of columns	33
Column type frequency:	
character	9
numeric	24
Group variables	None

4 Análise Descritiva

As análises dos dados referem-se a um processo crítico em relação produtividade de soja no RS. Dessa forma, foi verificadas medidas de tendência central, medidas de dispersão, as relações entre as variáveis e suas distribuições.

Variable type: character

skim_variable	n_missing	min	max	empty	n_unique	whitespace
Safra	0	9	9	0	2	0
COD_PROD	0	5	5	0	34	0
Local	0	5	23	0	27	0
Cod_Estacao_Met	0	4	4	0	18	0
Terras	0	5	6	0	2	0
Ambiente	0	8	8	0	3	0
Cultivar	0	5	25	0	184	0
Cultura_Ant	119	3	17	0	12	0
Epoca_de_semeadura	0	10	10	0	40	0

Variable type: numeric

skim_variable	n_missing	mean	sd	p0	p25	p50	
Altitude	0	304.69	217.97	3.00	105.00	288.00	
N_base	191	12.46	8.56	0.00	6.00	9.20	
P_base	191	74.45	28.20	40.00	56.00	64.40	
K_base	191	41.47	32.79	0.00	0.00	45.50	
Produtividade	0	2630.00	1417.06	185.00	1512.70	2416.70	
GMR	32	5.83	0.43	4.90	5.50	5.80	
N_Plantas	474	77.28	25.36	12.00	59.75	74.50	
Espacamento	53	0.47	0.04	0.40	0.45	0.45	
Area_colhida	247	3.37	0.96	0.90	2.70	3.60	
Populacao	247	18.12	9.21	0.00	15.25	20.30	
Temperatura_Max	0	24.42	1.23	22.08	23.59	24.48	
Precipitacao_Total	0	389.34	88.58	158.70	315.90	372.80	
Umidade_Med	0	68.61	6.20	52.93	64.45	68.68	
Radiacao_Med	0	24285.63	2940.83	20411.79	22770.84	23768.85	2
PH	0	5.18	0.35	4.50	5.00	5.10	
ARGILA	18	43.19	20.31	4.00	27.00	44.00	
P.mg.L.	0	26.14	41.91	2.50	9.00	19.00	
K.mg.L.	0	130.42	109.59	28.00	72.00	88.00	
M_O	0	2.51	0.83	1.00	1.90	2.50	
CTC.PH7	0	14.09	4.21	6.90	11.40	14.00	
SATURACAO.DE.BASE...	0	62.70	14.10	27.00	54.00	65.00	
SATURACAO.AL...	0	3.38	5.74	0.00	0.00	2.10	
Latitude	0	-29.52	1.57	-33.52	-29.70	-29.08	
Longitude	0	-53.65	1.29	-56.68	-53.99	-53.78	

Podemos verificar valores faltantes em algumas variáveis no banco de dados, dessa forma, foram retiradas essas observações devido a falta de informação para o preenchimento correto desses NA's. Além disso, podemos ter uma ideia da média e da distribuição dos dados

A Figura 1 apresenta a matriz de correlação das variáveis numéricas. A intensidade e o tamanho dos círculos permitem identificar rapidamente associações fortes — positivas quando em azul, negativas quando em vermelho. Essa etapa auxilia na identificação de possíveis casos de multicolinearidade e variáveis potencialmente relevantes para explicar a produtividade.

A Figura 2 mostra a distribuição da produtividade (Kg/ha) ao longo das duas safras analisadas. A linha vertical laranja indica a média geral, igual a 2457.65.

A Figura 3 compara a produtividade entre as safras 2021/2022 e 2022/2023. Verifica-se um aumento expressivo no segundo ciclo, provavelmente associado a condições climáticas mais favoráveis.

A Figura 4 ilustra a variabilidade entre locais. Observa-se grande amplitude produtiva, com valores inferiores a 1000 kg/ha em Bossoroca e superiores a 6000 kg/ha na cidade de Torres, evidenciando forte heterogeneidade espacial.

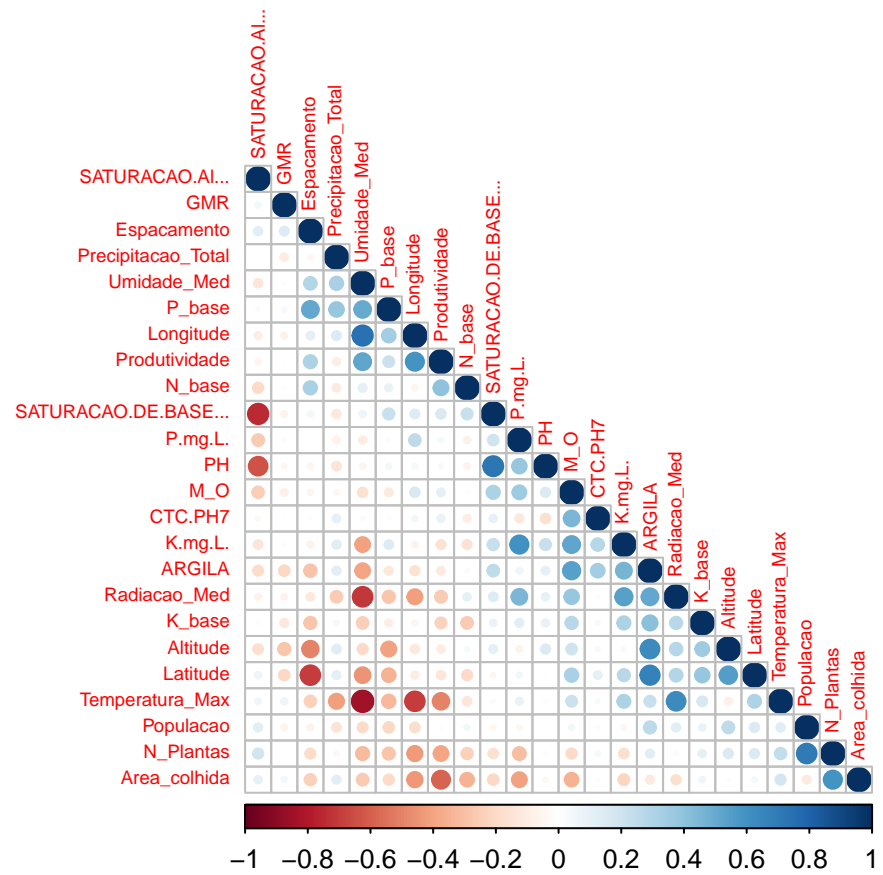


Figure 1: Corplot das Variáveis Numéricas

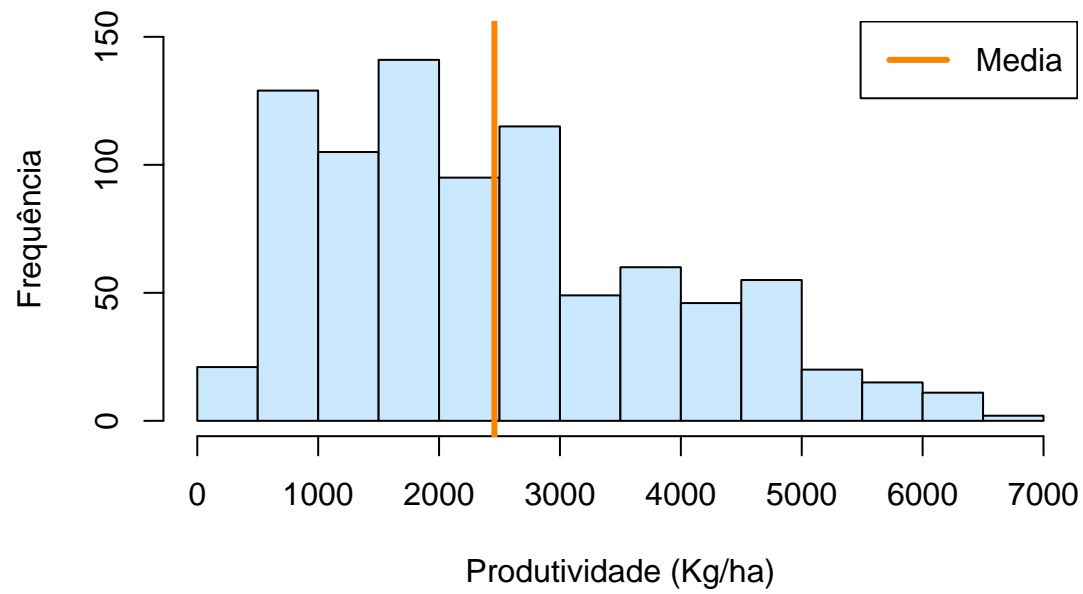


Figure 2: Histograma da Produtividade de Soja

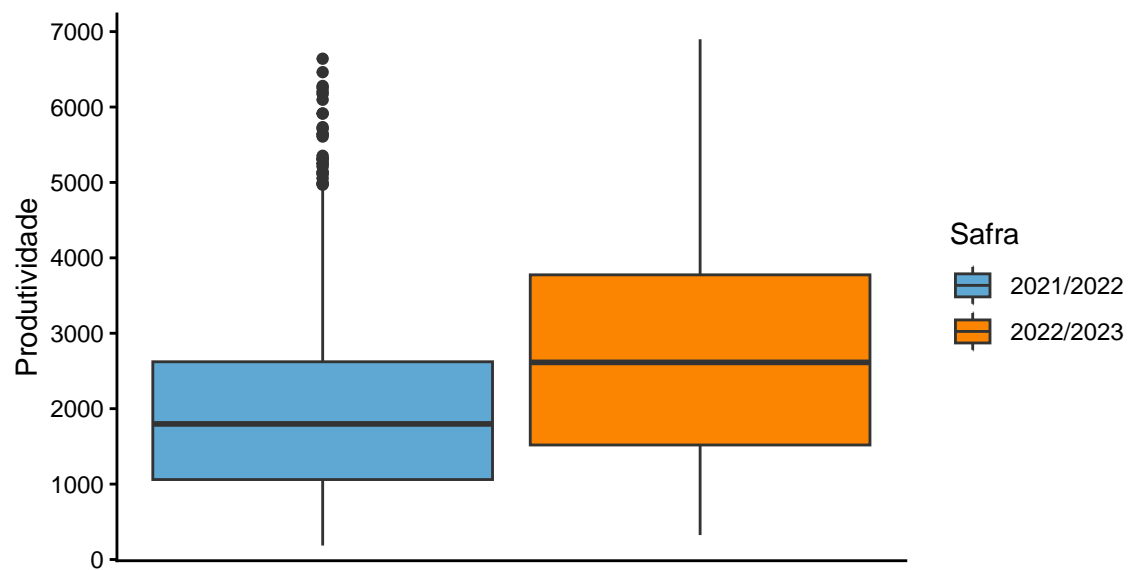


Figure 3: Boxplot da produtividade de soja

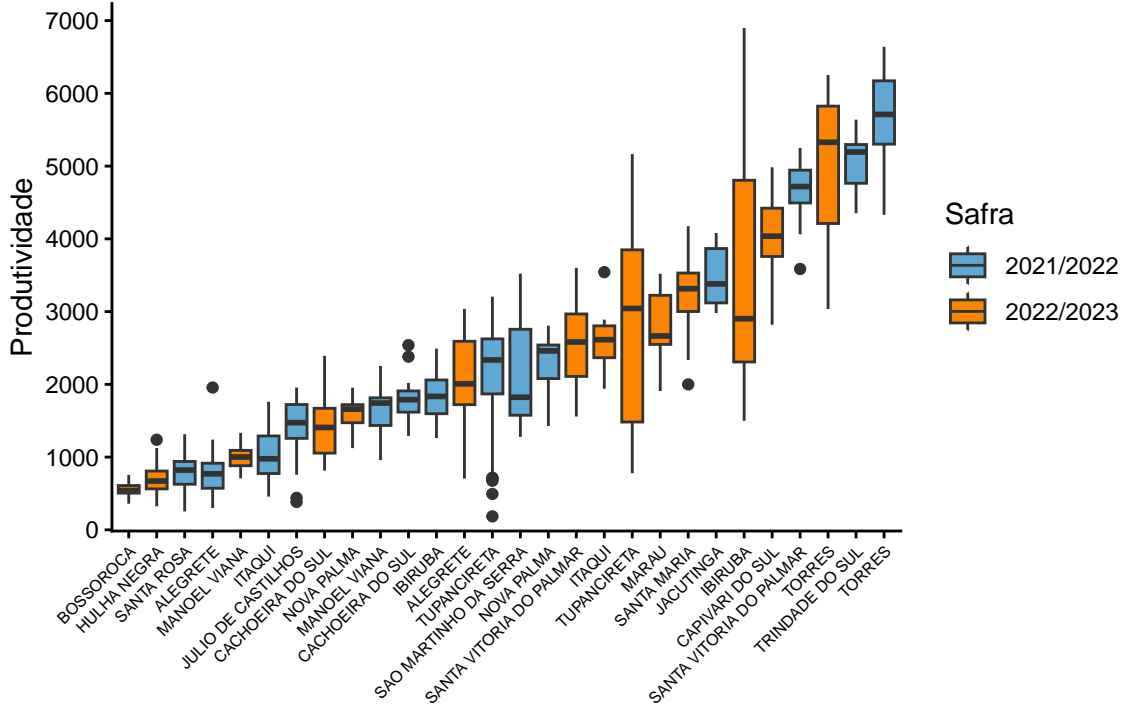


Figure 4: Boxplot da Produtividade separada por Locais e por Safra

5 Metodologia

A metodologia adotada neste estudo foi dividida em duas etapas principais:

- (i) um procedimento de aprendizado não supervisionado, voltado à identificação de padrões estruturais nos ambientes experimentais; e
- (ii) um processo de aprendizado supervisionado, cujo objetivo é a predição da produtividade de soja utilizando os agrupamentos previamente identificados.

5.1 Aprendizado não supervisionado

Para a etapa não supervisionada, empregou-se o algoritmo K-Means, cuja finalidade é particionar as observações em K grupos distintos, de modo que exista alta homogeneidade intra-cluster e heterogeneidade inter-cluster. Cada observação pertence exclusivamente a um único grupo, e a definição dos clusters é realizada minimizando a variabilidade interna de cada conjunto. Sejam C_1, \dots, C_K os conjuntos de índices das observações pertencentes a cada cluster e seja $X \in R^{n \times p}$ a matriz de dados padronizados. O critério otimizado pelo K-Means pode ser escrito como:

$$\text{minimize}_{C_1, \dots, C_K} \left\{ \sum_{k=1}^K \frac{1}{|C_k|} \sum_{i, i' \in C_k} \sum_{j=1}^p (x_{ij} - x_{i'j})^2 \right\}.$$

Em outras palavras, o método busca minimizar a soma das distâncias quadráticas entre cada observação e o centróide de seu respectivo grupo. O algoritmo segue as seguintes etapas: 1. Inicialização: cada observação recebe aleatoriamente um rótulo entre 1 e K. These serve as initial cluster assignments for the observations. 1. Iteração (até convergência): a. recalcular os centróides de cada cluster; b. reatribuir cada observação ao cluster cujo centróide apresente menor distância Euclidiana.

A seleção do número adequado de clusters constitui uma etapa essencial do processo de agrupamento. Neste trabalho, foram avaliados valores de $K \in \{2, 3, 4, 5, 6, 7, 8\}$. Para determinar o particionamento mais apropriado, utilizou-se a métrica Within-Cluster Sum of Squares (WSS), definida como a soma das distâncias quadráticas entre cada observação e o centróide de seu respectivo cluster. Essa medida reflete a coerência interna dos grupos: valores menores indicam clusters mais compactos e homogêneos.

Formalmente, o WSS é expresso por:

$$\text{WSS}(K) = \sum_{k=1}^K \sum_{i \in C_k} \|x_i - \mu_k\|^2.$$

em que C_k representa o conjunto de observações atribuídas ao cluster k e μ_k é seu centróide.

É importante destacar que o WSS sempre diminui à medida que o número de clusters aumenta, mesmo quando essa divisão não corresponde a uma estrutura real dos dados. Dessa forma, torna-se necessário identificar um ponto após o qual o decréscimo no WSS deixa de ser significativo — o chamado método do cotovelo.

No presente estudo, ainda que o WSS apresentasse redução contínua para valores crescentes de K , essa redução ocorreu de maneira aproximadamente linear, sem indicar um cotovelo claramente definido. Assim, considerando também o conhecimento prévio sobre a estrutura dos ambientes experimentais e a necessidade de manter interpretabilidade, optou-se por adotar $K = 4$ clusters.

Para inspeção visual dos clusters, realizou-se uma Análise de Componentes Principais (PCA), um método que projeta os dados para um espaço de menor dimensão preservando a maior variabilidade possível. O primeiro componente principal é definido por:

$$Z_1 = \phi_{11}X_1 + \phi_{21}X_2 + \dots + \phi_{p1}X_p,$$

em que o vetor $\phi_1 = (\phi_{11}, \phi_{21}, \dots, \phi_{p1})^\top$ é obtido pela maximização da variância projetada sob a restrição de normalização.

A PCA permite representar os clusters em duas dimensões, fornecendo uma visão clara de sua dispersão e separabilidade. A Figura 5 ilustra esta projeção, mostrando que os quatro grupos apresentam estrutura coerente, ainda que com alguma sobreposição esperada dada a complexidade dos ambientes agrícolas.

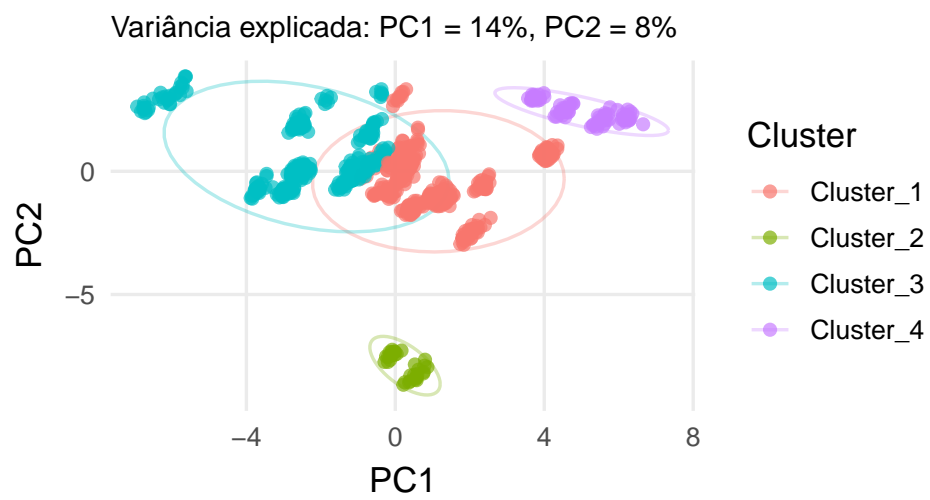


Figure 5: Análise de Componentes Principais