Métodos de Machine Learning na Produtividade de Soja

Arthur Hintz, Lucas Sartor

25 nov 2025

Resumo

O trabalho tem como objetivo realizar uma análise de regressão linear múltipla para estimar a produtividade de soja em (Kg/ha) com base nos principais fatores que a influenciam. A análise será dividida em quatro etapas principais: análise descritiva, ajuste do modelo, diagnóstico de influência e teste das suposições do modelo.

Introdução

Segundo a Confederação da Agricultura e Pecuária do Brasil (CNA), o Produto Interno Bruto (PIB) do Agronegócio corresponde a 23,8% em 2023, sendo a soja a commodity de maior valor de produção no Brasil, de acordo com dados divulgados pelo IBGE, link de acesso. Na produção de soja, muitos fatores influenciam o resultado, sendo vários deles incontroláveis, como fatores climáticos. Este ano, por exemplo, houve um aumento de 70,83% na produtividade em relação ao ano anterior, provavelmente relacionado ao volume de precipitação.

Dessa forma, torna-se necessário estimar a produtividade da soja e verificar quais são as principais variáveis que a influenciam, permitindo realizar predições da safra e estimar o potencial de produção a nível nacional.

Inicialmente, será realizada uma análise descritiva dos dados para compreender as variáveis envolvidas e suas inter-relações. Em seguida, o modelo de regressão linear múltipla será ajustado para identificar os fatores mais significativos que afetam a produtividade da soja. O diagnóstico de influência ajudará a identificar pontos de dados que têm um impacto desproporcional no ajuste do modelo, possibilitando a correção ou análise adicional desses pontos. Finalmente, as suposições do modelo de regressão linear serão testadas para garantir a validade das conclusões obtidas.

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

Este estudo é importante porque permite identificar os principais determinantes da produtividade da soja, fornecendo insights valiosos para a tomada de decisões agrícolas e a otimização dos rendimentos das cultivares. Ao compreender melhor os fatores que influenciam a produtividade, produtores e pesquisadores podem implementar práticas agrícolas mais eficazes.

Dados

O banco de dados, inicialmente, continha informações dos 4 blocos de ensaios para cada cultivar. Posteriormente, foi calculada a média dos blocos por cultivar, resultando em 1513 observações e 33 variáveis.

Entretanto, após o precesso de filtragens e tratamento de valores faltantes, esses números foram reduzidos. As variáveis do banco de dados incluem informações sobre as cultivares, características do local, dados climáticos durante o período dos experimentos e componentes químicos do solo.

- **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.

Dentre essas principais características, as variáveis mais significativas utilizadas no modelo foram:

1. `Terras`: divididas em (Altas ou Baixas)
2. `Ambiente`: dividido em (Sequeiro ou Irrigado)
3. `Cultura_Ant`: ("arroz e pousio","aveia", "aveia branca", "aveia e centeio", "aveia e ervilhaca", "azevem", "cevada", nabo")
4. `P_base`: Quantidade de adubação de Fósforo
5. `N_base`: Quantidade de adubação de Nitrogênio
6. `Produtividade`: Produtividade de soja (Kg/ha)
7. `GMR`: Grupo de maturação relativo
8. `Espacamento`: Espaçamento entre linhas do plantio de soja
9. `Temperatura_Max`: média da temperatura máxima durante o período
10. `PH`: PH do solo
11. `M.O.(%)`: Matéria orgânica (%)
12. `Epoca_de_semeadura`: Data de plantio

Sendo as primeiras colunas e observações dadas por:

Safra	COD_PROD	Local	Cod_Estacao_Met	Altitude	Terras	Ambiente	Cultivar
2021/2022	AG117	ALEGRETE	A826	117	BAIXAS	SEQUEIRO	AS 3595 I2X
2021/2022	AG117	ALEGRETE	A826	117	BAIXAS	SEQUEIRO	AS 3615 I2X
2021/2022	AG117	ALEGRETE	A826	117	BAIXAS	SEQUEIRO	BMX COMPACT
2021/2022	AG117	ALEGRETE	A826	117	BAIXAS	SEQUEIRO	BMX CROMO TI
2021/2022	AG117	ALEGRETE	A826	117	BAIXAS	SEQUEIRO	BMX LOTUS IPI

Análise Descritiva

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

Table 2: Data summary

Name	dados
Number of rows	1513
Number of columns	33

Column type frequency:	
character	9
numeric	24
Group variables	None

Variable type: character

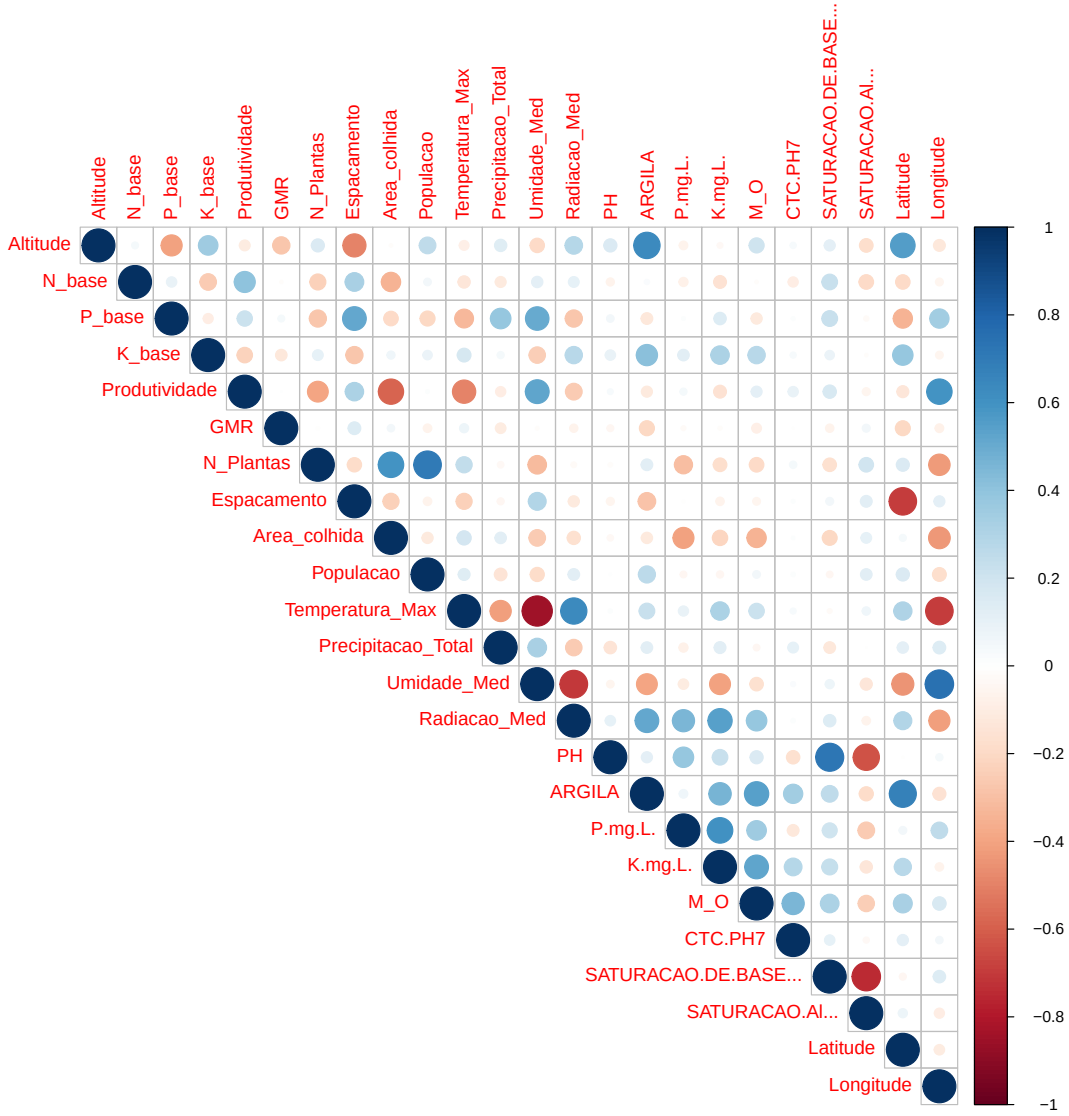
skim_variable	n_missing	min	max	empty	n_unique	whitespace
Safra	0	9	9	0	2	0
COD_PROD	0	5	5	0	34	0
Local	0	5	23	0	27	0
Cod_Estacao_Met	0	4	4	0	18	0
Terras	0	5	6	0	2	0
Ambiente	0	8	8	0	3	0
Cultivar	0	5	25	0	184	0
Cultura_Ant	119	3	17	0	12	0
Epoca_de_semeadura	0	10	10	0	40	0

Variable type: numeric

skim_variable	n_missing	mean	sd	p0	p25	p50	p75	p100
Altitude	0	304.69	217.97	3.00	105.00	288.00	489.00	688.00
N_base	191	12.46	8.56	0.00	6.00	9.20	17.20	40.00
P_base	191	74.45	28.20	40.00	56.00	64.40	92.00	135.00
K_base	191	41.47	32.79	0.00	0.00	45.50	60.00	112.50
Produtividade	0	2630.00	1417.06	185.00	1512.70	2416.70	3622.57	6898.68
GMR	32	5.83	0.43	4.90	5.50	5.80	6.10	8.10
N_Plantas	474	77.28	25.36	12.00	59.75	74.50	92.25	172.00
Espacamento	53	0.47	0.04	0.40	0.45	0.45	0.45	0.58
Area_colhida	247	3.37	0.96	0.90	2.70	3.60	4.05	5.40
Populacao	247	18.12	9.21	0.00	15.25	20.30	24.07	43.98
Temperatura_Max	0	24.42	1.23	22.08	23.59	24.48	25.08	28.10
Precipitacao_Total	0	389.34	88.58	158.70	315.90	372.80	451.20	735.20
Umidade_Med	0	68.61	6.20	52.93	64.45	68.68	72.00	80.19
Radiacao_Med	0	24285.63	2940.83	20411.79	22770.84	23768.85	24520.27	38592.57
PH	0	5.18	0.35	4.50	5.00	5.10	5.30	6.30
ARGILA	18	43.19	20.31	4.00	27.00	44.00	59.00	85.00
P.mg.L.	0	26.14	41.91	2.50	9.00	19.00	27.00	359.20
K.mg.L.	0	130.42	109.59	28.00	72.00	88.00	180.00	636.00
M_O	0	2.51	0.83	1.00	1.90	2.50	2.90	5.90
CTC.PH7	0	14.09	4.21	6.90	11.40	14.00	15.40	35.50
SATURACAO.DE.BASE...	0	62.70	14.10	27.00	54.00	65.00	76.00	86.00
SATURACAO.Al...	0	3.38	5.74	0.00	0.00	2.10	4.00	34.00
Latitude	0	-29.52	1.57	-33.52	-29.70	-29.08	-28.53	-27.55
Longitude	0	-53.65	1.29	-56.68	-53.99	-53.78	-53.21	-49.74

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

Com o gráfico a seguir, podemos verificar as principais correlações entre as variáveis numéricas e já procurar possiveis casos de multicolinearidade e variáveis que podem ser mais significativas para estimar a produtividade de soja

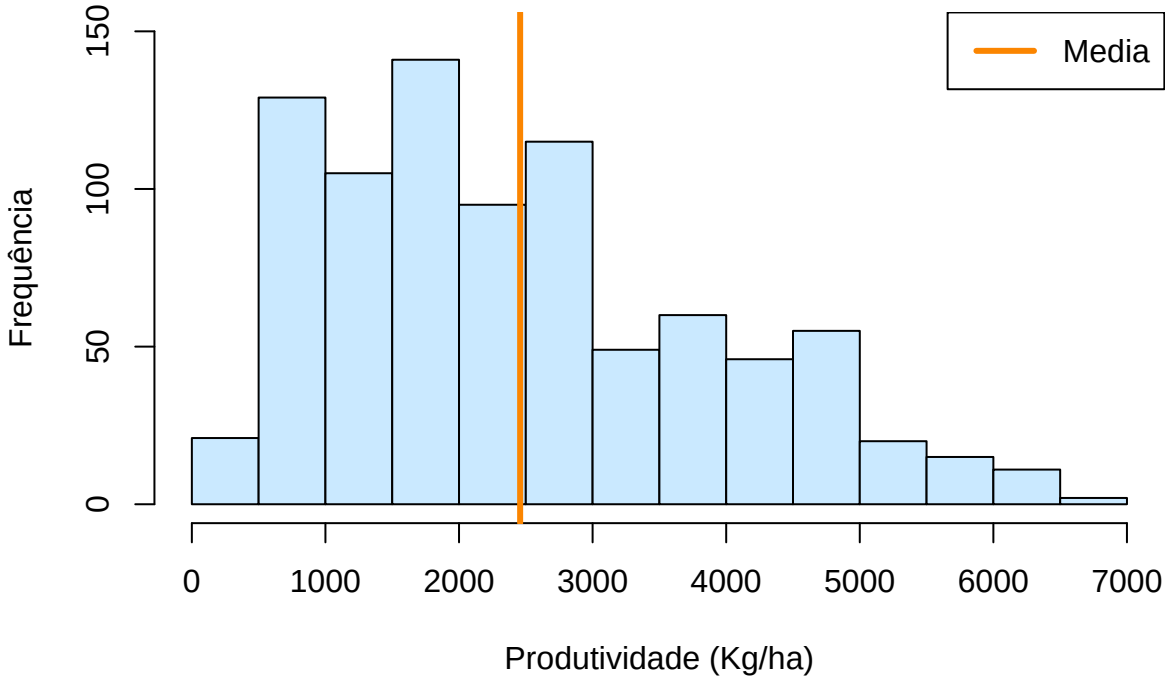


O gráfico de correlação entre as variáveis pode ser interpretado a partir do tamanho dos círculos e da cor. Quanto maior o círculo e mais escura é a cor mais forte é a correlação, também, se puxar mais para o azul é positiva, se for vermelha é negativa.

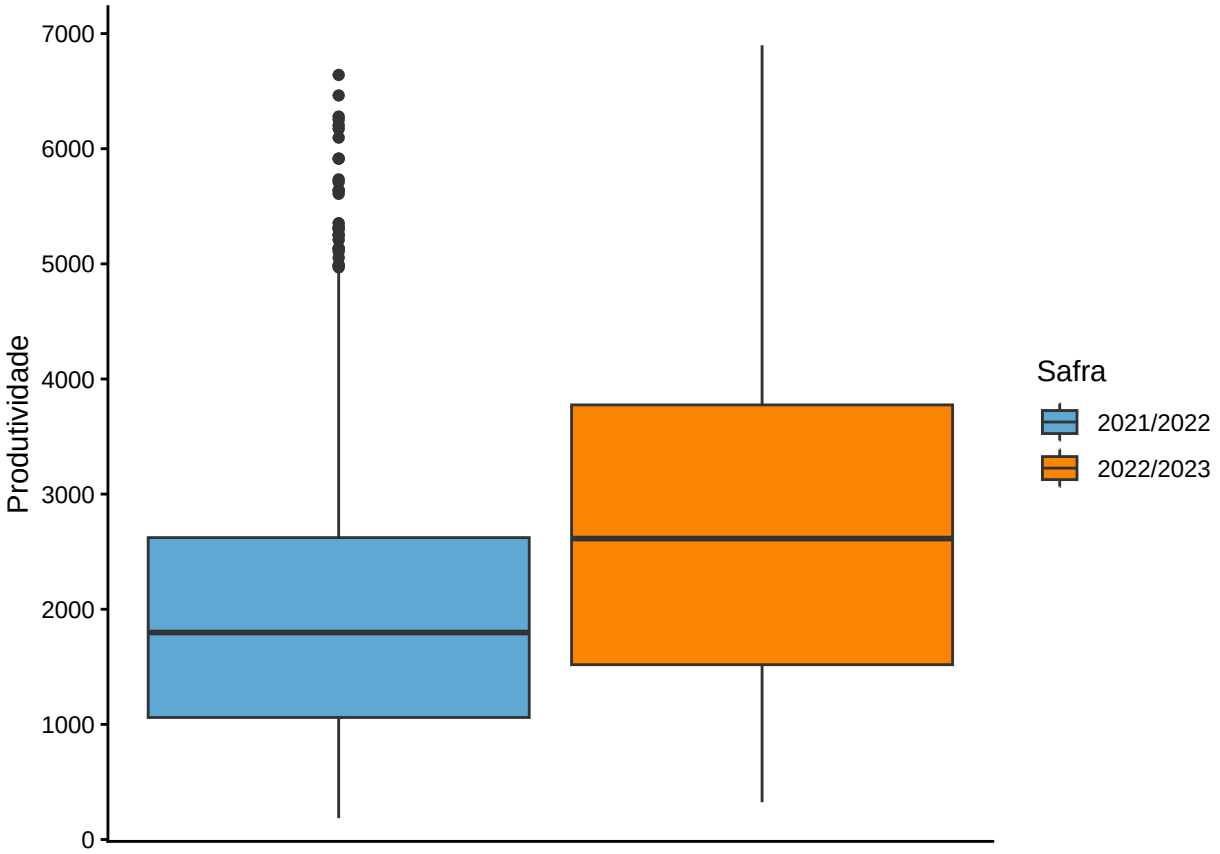
Principais medidas da variável de interesse e um gráfico para mostrar frequência de produtividade

Média da produtividade: 2458

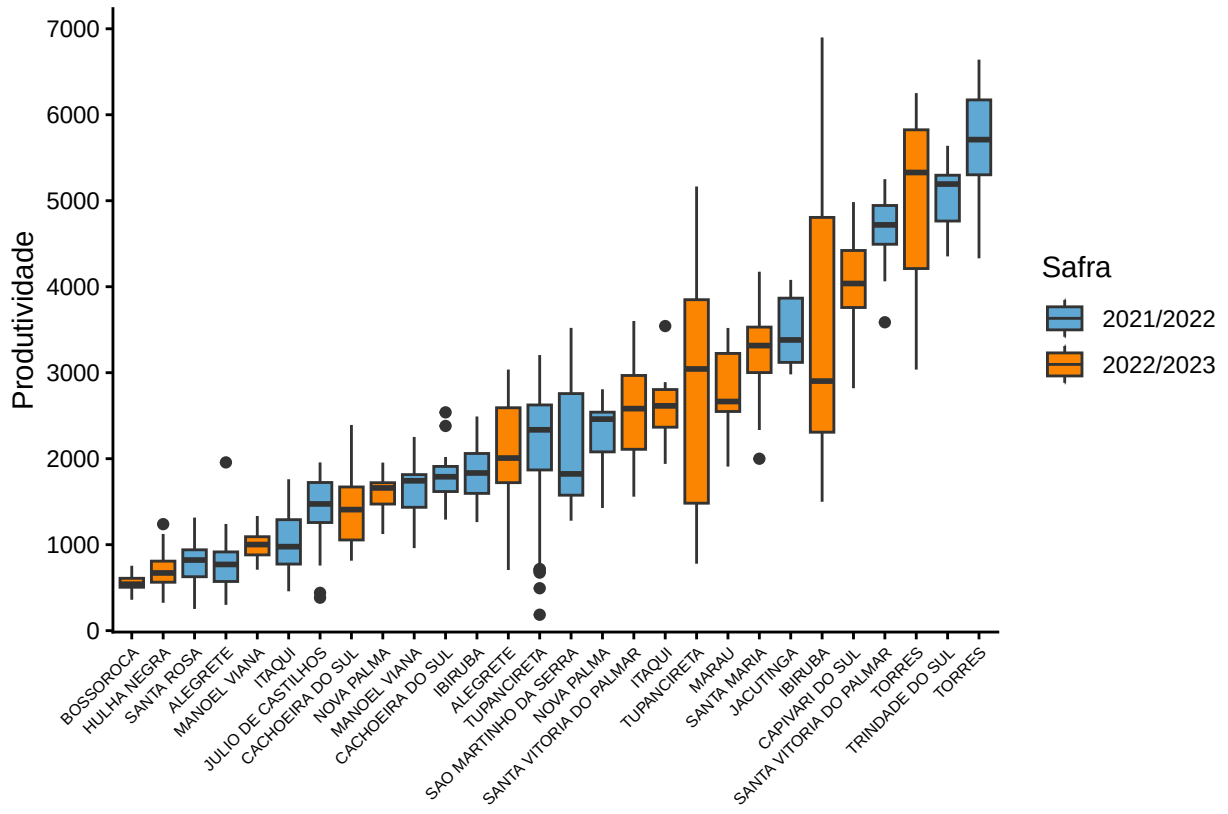
Desvio padrão da produtividade: 1434



O gráfico de boxplot mostra a relação entre os anos das safras com a produtividade



Nota-se um diferença de produtividade entre as duas safras, de certa forma a variável `Safra` deveria ser significativa para a explicação da produtividade de soja, no entanto ja tem relação com outras variáveis do modelo. Além disso, deve explicar a alta variabilidade dos erros, conforme será apresentado nas suposições do modelo.



A partir do gráfico percebe-se a alta variabilidade de produtividade entre os locais de ensaio nas duas safras.

Ajustes dos Dados

1 - Inicialmente foi removida uma cultivar experimental e cultivares com grupos de maturação relativos maiores que 7

2 - Os locais Santa Rosa e Jacutinga apresentaram ser pontos influêntes para o modelo. Logo para o ajuste foi melhor remover essas observações

3 - Locais os quais não tiveram uma cultura antes do plantio de soja ou tiveram mix também foram pontos de alavancagem para o modelo.

4 - Criação da variável `mês`, relacionada a data de plantio, ou seja, invés de ter dia e mês, tem se apenas o mês

5 - Foi alterado a variável `Terras`, em que Baixas recebe 0 e Altas recebe 1. A variável `Ambiente`, "irrigado" = 1 e "sequeiro" = 0

Os dados depois de filtrados e selecionados as variáveis importantes para o modelo, é dado por:

	Terras	Ambiente	Cultura_Ant	P_base	N_base	Produtividade	GMR	Espacamento	Temperatura_
308	0	1	azevem	135.0	9.0	4330.1250	6.1	0.50	23.
183	1	0	aveia	69.0	6.0	2585.2788	5.8	0.45	24.
591	1	1	aveia	43.0	17.2	4275.1000	5.4	0.45	24.

12	0	0	azevem	69.0	13.5	405.2235	5.7	0.45	25.0
631	1	1	trigo	40.0	16.0	3592.3250	5.8	0.45	25.3
682	1	0	trigo	40.0	16.0	906.2000	5.4	0.45	25.3
231	1	1	aveia e ervilhaca	57.5	5.0	2897.3250	6.4	0.45	24.4
254	1	1	aveia e ervilhaca	57.5	5.0	2735.3250	5.9	0.45	24.4
257	1	1	aveia e ervilhaca	57.5	5.0	1058.9250	4.9	0.45	23.0
390	0	0	azevem	56.0	6.0	3135.2500	5.2	0.45	23.0