

Métodos de Machine Learning na Produtividade de Soja

Arthur Hintz, Lucas Sartor

26 nov 2025

Contents

1	Resumo	1
2	Introdução	1
3	Conjunto de Dados	2
4	Análise Descritiva	3
5	Metodologia	5

1 Resumo

Este relatório descreve a execução de um projeto de Machine Learning no âmbito da disciplina, fazendo o uso inicialmente de um modelo não supervisionado e posteriormente supervisionado. São apresentados objetivos, dados, metodologia, resultados e análises de sensibilidade, com ênfase em validação, regularização e reprodutibilidade. Para validação, empregamos cross-validation (CV)

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.

2 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, obtidos no [link](#). Na produção de soja, muitos fatores influenciam o resultado, sendo vários deles incontrolláveis, 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 previsões 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

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

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.

3 Conjunto de 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 processo 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.

A Tabela ?? apresenta uma amostra aleatória de 10 linhas do conjunto de dados final. Inicialmente foram criadas e ajustadas algumas variáveis:

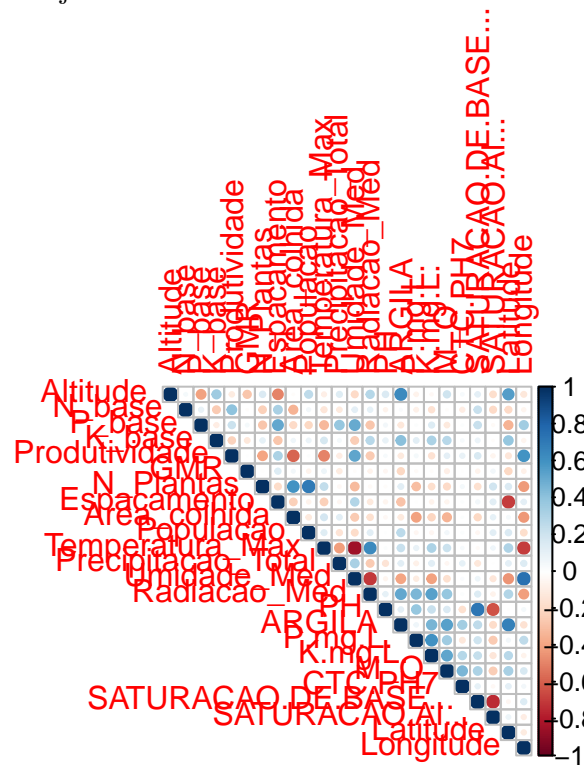
1. Foi criada a variável *dias* a partir da data de semeadura, dessa forma ela entra no modelo como o numero do dia do ano que foi plantado a soja
2. Todas as variáveis que não eram do tipo numéricas foram transformadas para do tipo fator

4 Análise Descritiva

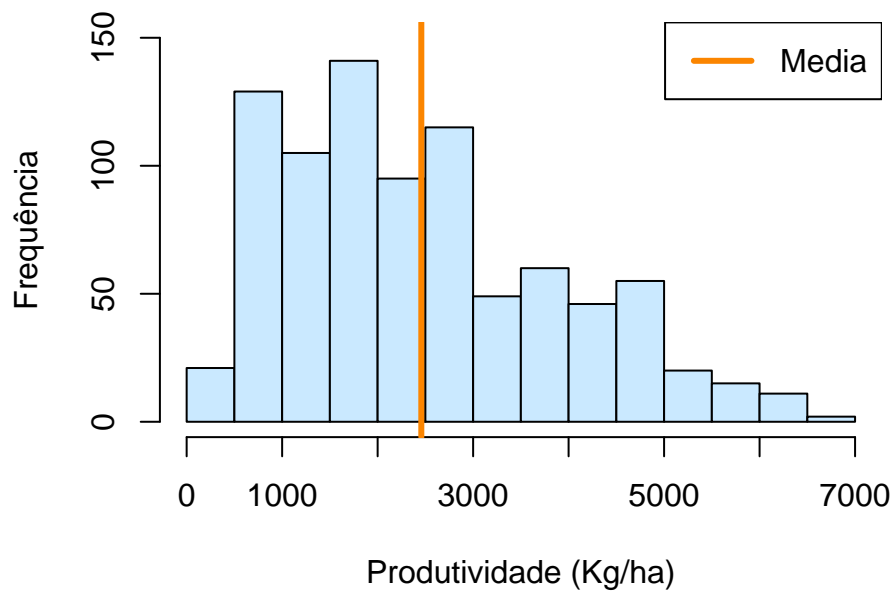
As análises dos dados referem-se a um processo crítico em relação produtividade de soja no RS. Dessa forma, foi verificadas medidas de tendência central, medidas de dispersão, as relações entre as variáveis e suas distribuições.

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

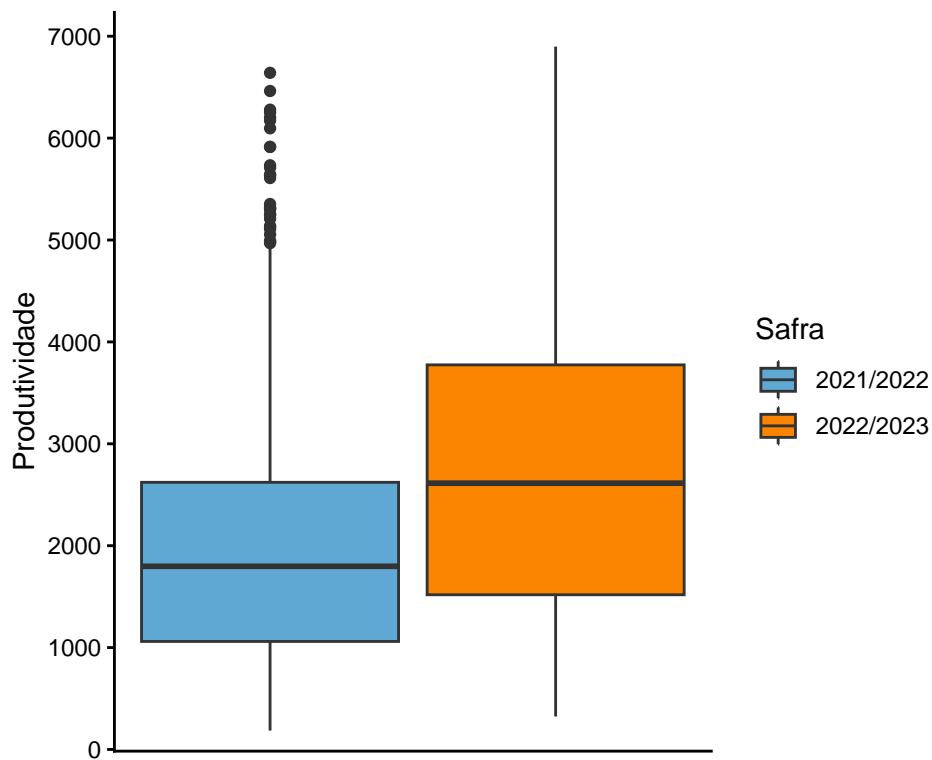
Com o gráfico a seguir, podemos verificar as principais correlações entre as variáveis numéricas e já procurar possíveis casos de multicolinearidade e variáveis que podem ser mais significativas para estimar a produtividade de soja



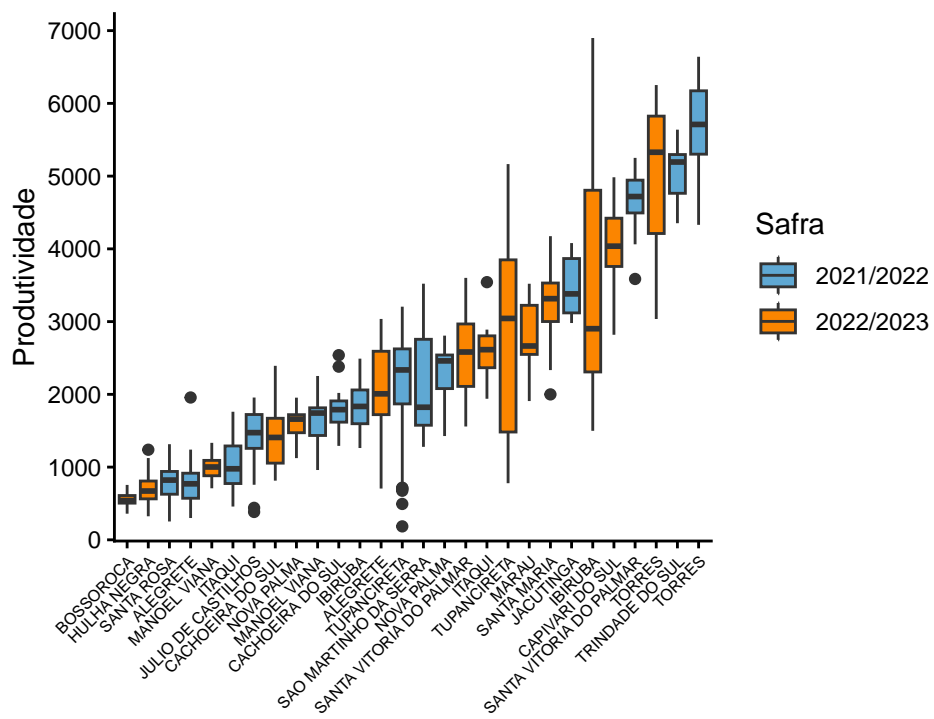
O gráfico de correlação entre as variáveis pode ser interpretado a partir do tamanho dos círculos e da cor. Quanto maior o círculo e mais escura é a cor mais forte é a correlação, também, se puxar mais para o azul é positiva, se for vermelha é negativa.



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



Nota-se uma diferença de produtividade entre as duas safras, de certa forma a variável **Safrã** deveria ser significativa para a explicação da produtividade de soja, no entanto já 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.

5 Metodologia