

Métodos de Machine Learning na Produtividade de Soja

Arthur Hintz, Lucas Sartor

30 nov 2025

Contents

| | |
|--|----------|
| 1 Resumo | 1 |
| 2 Introdução | 1 |
| 3 Conjunto de Dados | 2 |
| 4 Análise Descritiva | 3 |
| 5 Metodologia | 6 |
| 5.1 Aprendizado não supervisionado | 6 |
| 5.2 Aprendizado supervisionado | 8 |
| 6 Resultados | 9 |

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 reproduzibilidade. 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 incontroláveis, 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 1 que apresenta uma amostra aleatória de 10 linhas do conjunto de dados final.

Table 1: Exemplo das variáveis qualitativa

| Safra | COD_PROD | Local | Terras | Ambiente | Cultivar | Cultura_ |
|-----------|----------|-------------------------|--------|----------|---------------|----------|
| 2021/2022 | SS444 | SAO MARTINHO DA SERRA | ALTAS | SEQUEIRO | BMX ZEUS IPRO | aveia |
| 2022/2023 | DA542 | DOIS IRMAOS DAS MISSOES | ALTAS | IRRIGADO | DM 54IX57 12X | trigo |
| 2021/2022 | SR227 | SANTA ROSA | ALTAS | SEQUEIRO | DM 5958 IPRO | trigo |
| 2022/2023 | SL009 | SANTA VITORIA DO PALMAR | BAIXAS | SEQUEIRO | CZ 15B20 I2X | pousio |
| 2021/2022 | CA070 | CACHOEIRA DO SUL | BAIXAS | SEQUEIRO | DM 5958 IPRO | azevem |
| 2022/2023 | DA542 | DOIS IRMAOS DAS MISSOES | ALTAS | IRRIGADO | AS 3551 XTD | trigo |
| 2022/2023 | BA255 | BOSSOROCA | ALTAS | SEQUEIRO | ST 622 IPRO | trigo |
| 2021/2022 | TA485 | TUPANCIRETA | ALTAS | SEQUEIRO | M 5710 I2X | NA |
| 2021/2022 | IP051 | ITAQUI | BAIXAS | IRRIGADO | TMG 7362 IPRO | pousio |
| 2022/2023 | TA485 | TUPANCIRETA | ALTAS | IRRIGADO | P 95R21 E | aveia |

Table 2: Tabela das variáveis qualitativas

| variaveis | n_faltantes | n_categorias |
|--------------------|-------------|--------------|
| Safra | 0 | 2 |
| COD_PROD | 0 | 34 |
| Local | 0 | 27 |
| Cod_Estacao_Met | 0 | 18 |
| Terras | 0 | 2 |
| Ambiente | 0 | 3 |
| Cultivar | 0 | 184 |
| Cultura_Ant | 119 | 12 |
| Epoca_de_semeadura | 0 | 40 |

4 Análise Descritiva

A análise descritiva dos dados constitui uma etapa fundamental para compreender o comportamento da produtividade de soja no Rio Grande do Sul. Nessa fase, foram avaliadas medidas de tendência central e de dispersão, além das relações entre as variáveis e a forma das distribuições.

A Tabela 2 apresenta as variáveis qualitativas do conjunto de dados. A coluna *n_faltantes* indica a presença de valores ausentes, observando-se que apenas a variável *Cultura_Ant* possui registros faltantes. Já a coluna *n_categorias* evidencia o número de categorias existentes em cada variável qualitativa.

Em complemento, a Tabela 3 exibe as variáveis quantitativas juntamente com suas estatísticas descritivas. São apresentados valores de média, desvio padrão, mínimo, máximo e quartis, permitindo identificar a distribuição dos dados e potenciais assimetrias ou dispersões relevantes para a modelagem.

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.

Table 3: Tabela das variáveis quantitativas

| variaveis | media | dp | min | quartil_25 | mediana | quartil_75 | max |
|----------------------|----------|---------|----------|------------|----------|------------|----------|
| Altitude | 304.69 | 217.97 | 3.00 | 105.00 | 288.00 | 489.00 | 688.00 |
| N_base | 12.46 | 8.56 | 0.00 | 6.00 | 9.20 | 17.20 | 40.00 |
| P_base | 74.45 | 28.20 | 40.00 | 56.00 | 64.40 | 92.00 | 135.00 |
| K_base | 41.47 | 32.79 | 0.00 | 0.00 | 45.50 | 60.00 | 112.50 |
| Produtividade | 2630.00 | 1417.06 | 185.00 | 1512.70 | 2416.70 | 3622.57 | 6898.68 |
| GMR | 5.83 | 0.43 | 4.90 | 5.50 | 5.80 | 6.10 | 8.10 |
| N_Plantas | 77.28 | 25.36 | 12.00 | 59.75 | 74.50 | 92.25 | 172.00 |
| Espacamento | 0.47 | 0.04 | 0.40 | 0.45 | 0.45 | 0.45 | 0.58 |
| Area_colhida | 3.37 | 0.96 | 0.90 | 2.70 | 3.60 | 4.05 | 5.40 |
| Populacao | 18.12 | 9.21 | 0.00 | 15.25 | 20.30 | 24.07 | 43.98 |
| Temperatura_Max | 24.42 | 1.23 | 22.08 | 23.59 | 24.48 | 25.08 | 28.10 |
| Precipitacao_Total | 389.34 | 88.58 | 158.70 | 315.90 | 372.80 | 451.20 | 735.20 |
| Umidade_Med | 68.61 | 6.20 | 52.93 | 64.45 | 68.68 | 72.00 | 80.19 |
| Radiacao_Med | 24285.63 | 2940.83 | 20411.79 | 22770.84 | 23768.85 | 24520.27 | 38592.57 |
| PH | 5.18 | 0.35 | 4.50 | 5.00 | 5.10 | 5.30 | 6.30 |
| ARGILA | 43.19 | 20.31 | 4.00 | 27.00 | 44.00 | 59.00 | 85.00 |
| P.mg.L. | 26.14 | 41.91 | 2.50 | 9.00 | 19.00 | 27.00 | 359.20 |
| K.mg.L. | 130.42 | 109.59 | 28.00 | 72.00 | 88.00 | 180.00 | 636.00 |
| M_O | 2.51 | 0.83 | 1.00 | 1.90 | 2.50 | 2.90 | 5.90 |
| CTC.PH7 | 14.09 | 4.21 | 6.90 | 11.40 | 14.00 | 15.40 | 35.50 |
| SATURACAO.DE.BASE... | 62.70 | 14.10 | 27.00 | 54.00 | 65.00 | 76.00 | 86.00 |
| SATURACAO.Al... | 3.38 | 5.74 | 0.00 | 0.00 | 2.10 | 4.00 | 34.00 |
| Latitude | -29.52 | 1.57 | -33.52 | -29.70 | -29.08 | -28.53 | -27.55 |
| Longitude | -53.65 | 1.29 | -56.68 | -53.99 | -53.78 | -53.21 | -49.74 |

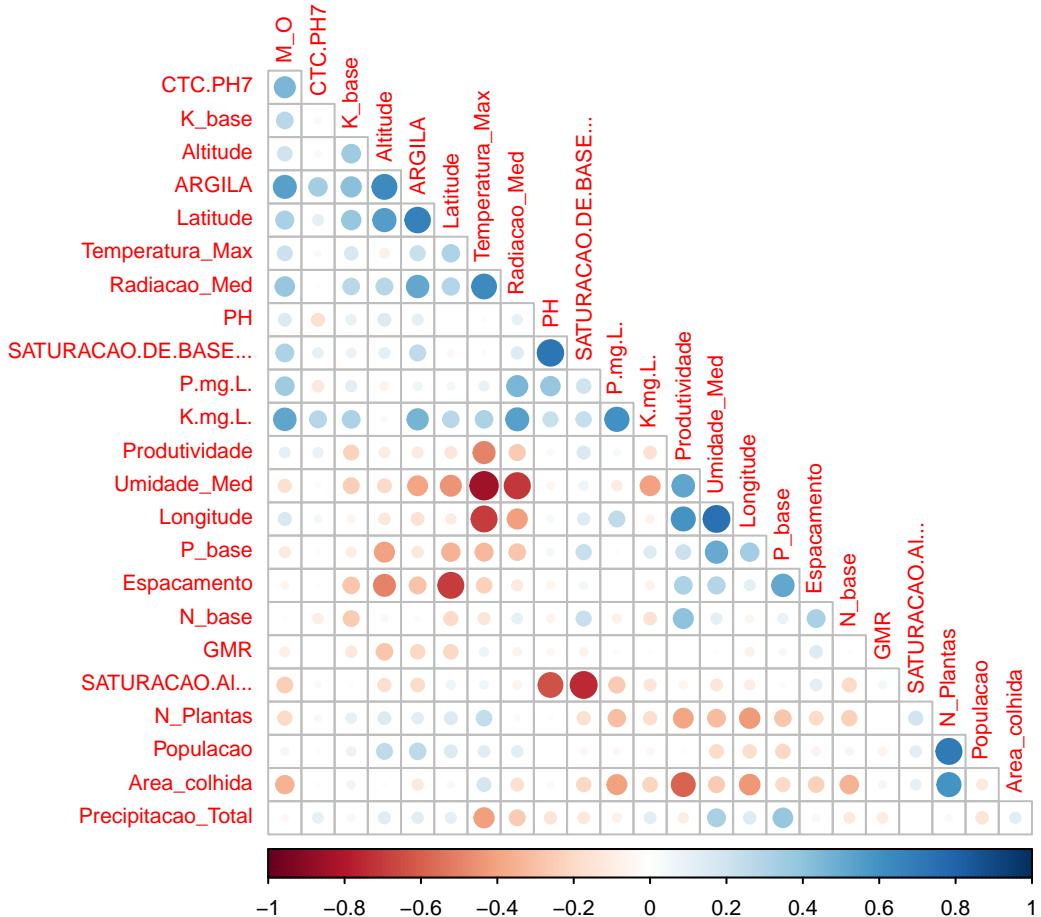


Figure 1: Corplot das Variáveis Numéricas

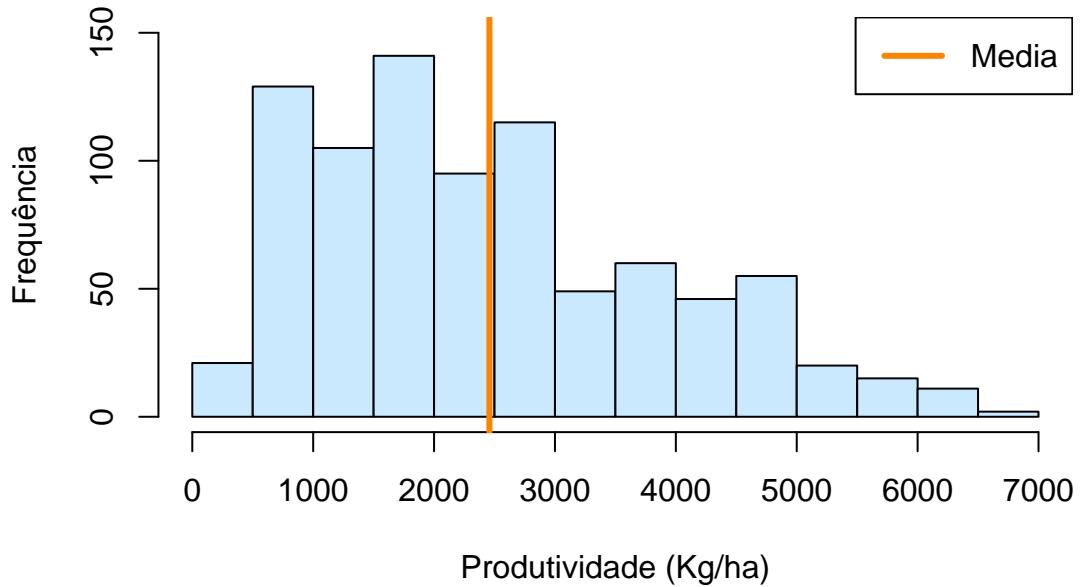


Figure 2: Histograma da Produtividade de Soja

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.

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 partitionar as observações em K grupos distintos, de modo que exista alta homogeneidade

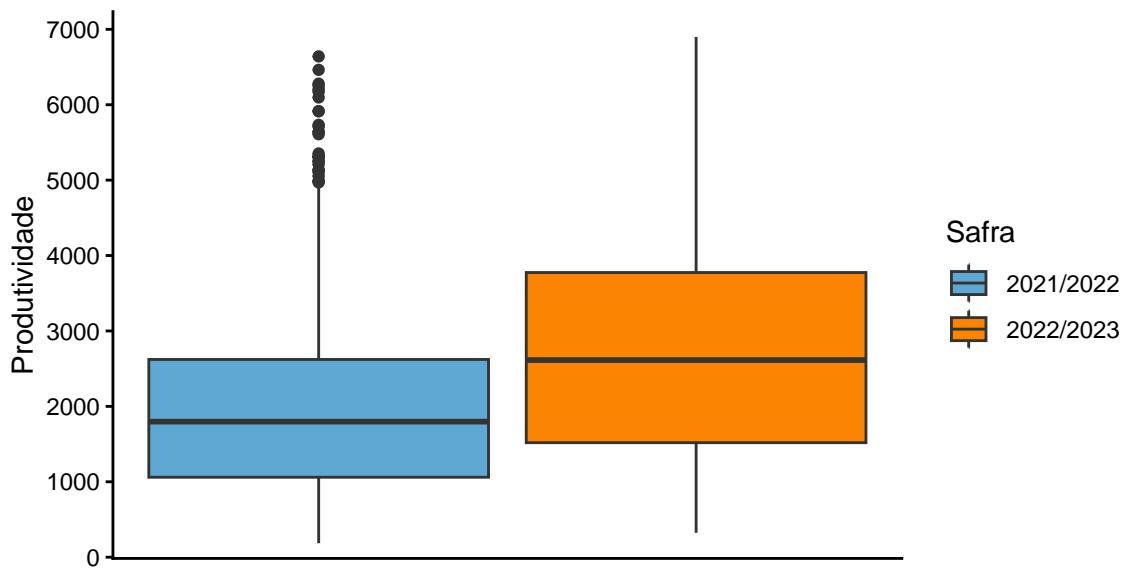


Figure 3: Boxplot da produtividade de soja

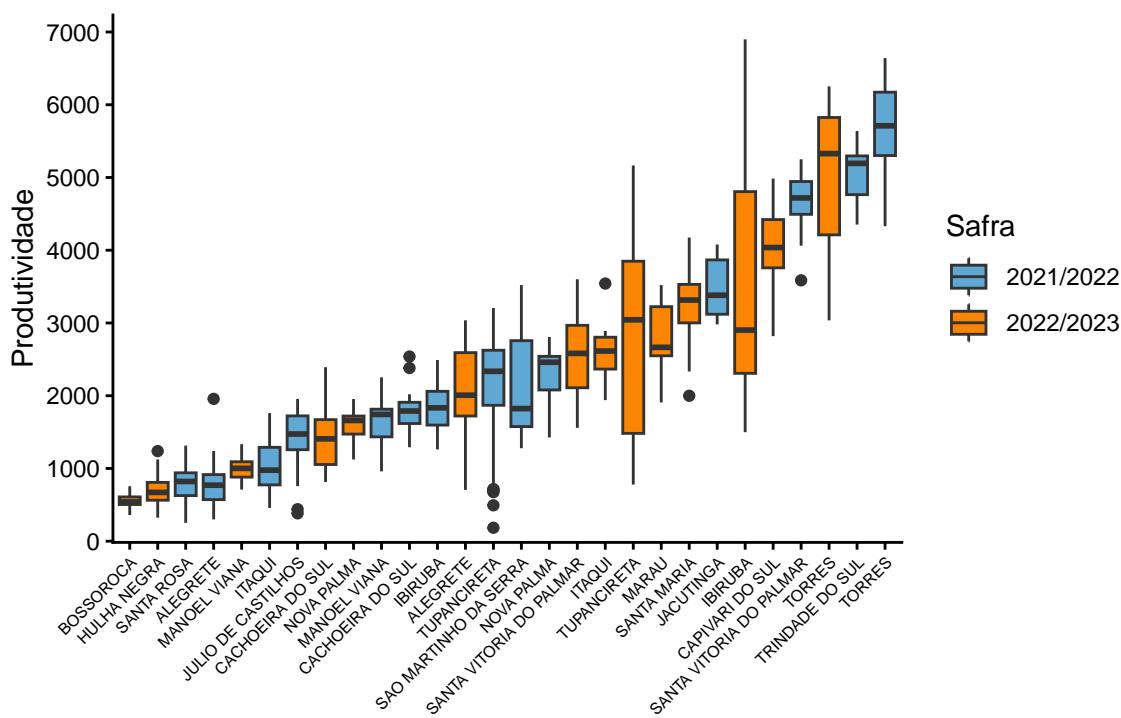


Figure 4: Boxplot da Produtividade separada por Locais e por Safras

dentro dos clusters e heterogeneidade entre os 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:

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

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.

5.2 Aprendizado supervisionado

Com os dados previamente particionados em 4 clusters com base na similaridade das variáveis analisadas, torna-se necessário verificar se essa segmentação também se reflete na variável produtividade, dado que essa variável não foi utilizada para atribuir os grupos. Caso os clusters apresentem diferenças significativas nos níveis de produtividade, será possível inferir que essa variável é influenciada pelas demais variáveis consideradas e que a estrutura de clusters obtida é adequada para avaliar a produção.

Para verificar se a produtividade pode ser explicada pelos clusters, é preciso empregar um método de aprendizado supervisionado, no qual há uma variável resposta a ser prevista e um conjunto de variáveis explicativas utilizadas para modelar essa previsão. O método supervisionado escolhido foi o de árvore de regressão com apenas os clusters determinando a produtividade.

Uma árvore de regressão é um modelo que realiza a predição de uma variável resposta por meio da construção de uma estrutura hierárquica de decisões. Trata-se de um particionamento do espaço das variáveis explicativas, essas divisões podem ser chamadas de nó. Em cada etapa procura-se a variável que minimiza a soma de quadrados de suas respectivas regiões.

Cada particionamento é definido por:

$$\min_{j,s} \left[\sum_{x_i \in R_1(j,s)} (y_i - \bar{y}_{R_1})^2 + \sum_{x_i \in R_2(j,s)} (y_i - \bar{y}_{R_2})^2 \right].$$

em que j representa as variáveis explicativas, s corresponde ao ponto de corte, X são cada linha do banco de dados, $R_1(j,s)$ e $R_2(j,s)$ são as regiões em que X_j são menores ou iguais ao ponto de corte ou maior que o ponto de corte respectivamente. \bar{y}_R é a média da variável resposta em cada região.

Esse processo acontece até que o número mínimo de observações por nó ou o máximo de profundidade da árvore sejam atingidos. Além disso há um custo de complexidade que define se vale a pena uma divisão extra. Esses são hiperparâmetros que foram tunados antes do ajuste para encontrar o melhor modelo.

O modelo por fim estima valores de produtividade para minimizar essa soma para cada um dos clusters.

6 Resultados

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:

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

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.

A partir do método supervisionado é possível estimar a produtividade associada para cada cluster:

Os resultados mostram diferenças expressivas nas médias de cada grupo, colaborando com a tese de que características semelhantes influenciam em sua produtividade final.

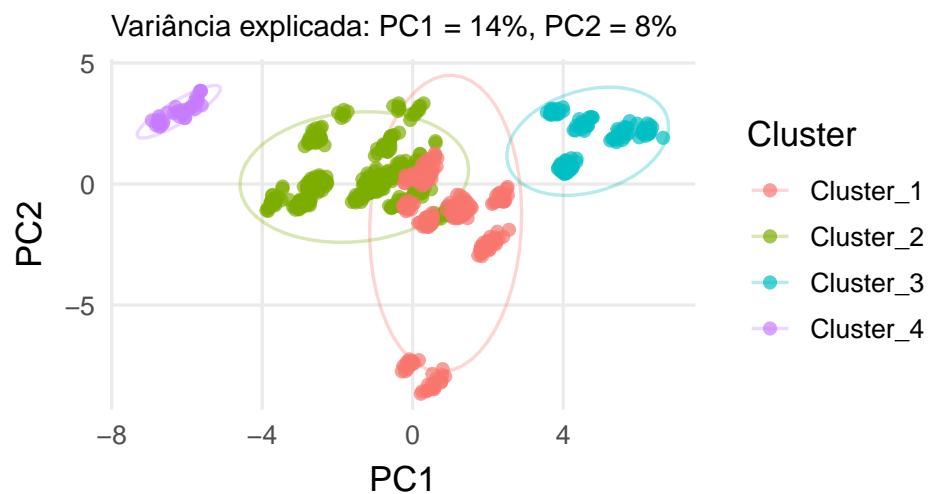


Figure 5: Análise de Componentes Principais

Além disso é interessante notar o padrão geográfico de cada cluster a partir de um mapa do Rio Grande do Sul. Nele será apresentado as médias de cada grupo e sua localização:

Esse mapa mostra divisões do estado e sua influência na divisão dos clusters e suas produtividades. Sendo assim é possível notar o quanto relevante é os fatores naturais de cada região.