# Predição de produtividade de sementes

## SCC0230 - Inteligência Artificial (Trabalho 2)

Antonio Carrilho Neto - Aluno UATI
Eduardo Souza Rocha - 11218692
Fábio Verardino de Oliveira - 12674547
Olavo Morais Borges Pereira - 11297792

# Introdução

# Contextualização

- As sementes são um dos insumos agrícolas com maior valor agregado, pois contém a constituição genética da variedade sendo cultivada: um dos fatores determinantes da produtividade.
- Sementes comerciais passam por rigoroso controle de qualidade e supervisão legal, sendo classificadas por sua origem, tecnologia utilizada, entre outros...
- Portanto, é interessante para os produtores de sementes que o processo seja o mais eficiente possível.

# Problema

- O Ministério da Agricultura e Pecuária mantém um registro de dados enviados por produtores de semente a cada safra, que inclui, entre outros dados, as características da área, a classe de semente produzida e, além da produção bruta real, uma estimativa dessa produção.
- Analisando os dados, percebe-se que a estimativa de produção tem, em geral, baixa precisão.
- Logo, temos como objetivo desenvolver um modelo que aumente a precisão dessa predição.

# Dados

- Trata-se de dados inseridos por campos produtores de sementes no SIGEF (Sistema de Gestão Fundiária), contemplando as safras de 2013 a 2023.
- O informe desses dados é requerido pela legislação brasileira de sementes.
- O conjunto de dados utilizado é disponibilizado no Portal de Dados Abertos do Ministério da Agricultura e Pecuária (https://dados.agricultura.gov.br/).
    - “Controle da Produção de Sementes e Mudas – SIGEF”
    - Total de 458.618 registros com 12 colunas.

# Dados

- Exemplo do *dataset*:

| | Safra | Especie | Categoria | Cultivar | Municipio | UF | Status | Data do Plantio | Data de Colheita | Area | Producao bruta | Producao estimada |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 2013/2013 | Zea mays L. | C1 | P4285H | Paraúna | GO | Aprovado | 16/02/2013 | 10/06/2013 | 44.0 | 409.0600 | 140.80 |
| **1** | 2013/2013 | Zea mays L. | C1 | P4285YH | Paraúna | GO | Aprovado | 17/02/2013 | 19/06/2013 | 37.0 | 279.1600 | 118.40 |
| **2** | 2013/2013 | Zea mays L. | C1 | P3646 | Morrinhos | GO | Aprovado | 28/02/2013 | 29/06/2013 | 23.5 | 201.6700 | 105.00 |
| **3** | 2013/2013 | Zea mays L. | C1 | P3161H | Paraúna | GO | Aprovado | 15/02/2013 | 05/06/2013 | 28.0 | 167.7100 | 72.80 |
| **4** | 2013/2013 | Zea mays L. | C1 | BG7061H | Caiapônia | GO | Aprovado | 14/02/2013 | 24/06/2013 | 87.0 | 848.2700 | 279.10 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |

# Pré-processamento

# Dados faltantes

- 317.093 registros não possuem informações de *Data de Colheita* e *Produção Bruta (real)* e, portanto, foram descartados.

- Dessa forma, restam 141.516 registros a serem utilizados na análise (aprox. 30% do *dataset* original).

# Tratamento de outliers

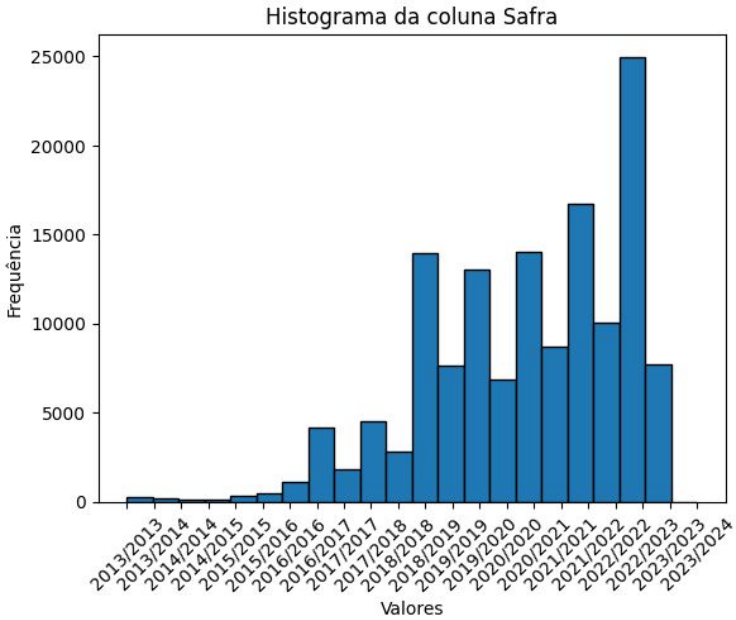
- Calculando o erro médio (RMSE) da produção estimada em relação à produção bruta no conjunto de dados, obtém-se um valor de aprox. 37.498 t, o que indica a presença de *outliers* no conjunto.
- Para tratar esses dados, foi realizado um agrupamento por *k-means* com *k=3*, sendo descartados os dois menores clusters.
- Após isso, restaram 139.646 registros, i.e., cerca de 1,3% foram considerados *outliers*.
- Como resultado, o RMSE passou a ser de 9.193 t, um valor bem mais plausível.

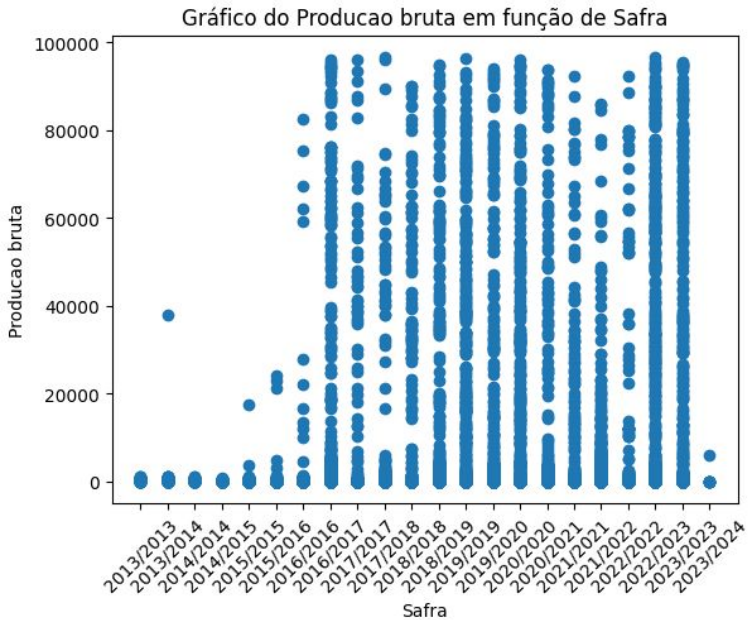
# Distribuição dos dados

- A fim de entender melhor a distribuição dos dados do conjunto e as possíveis correlações (atributos de interesse), foi realizada uma análise estatística do conjunto.

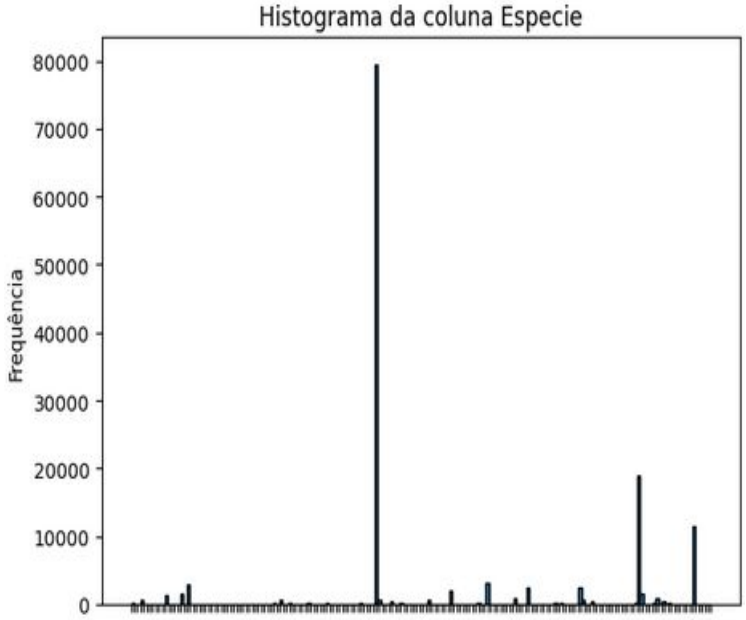
- Foram gerados histogramas e gráficos em função da produção bruta para as seguintes colunas:
    - Safra, Espécie, Categoria, Cultivar, Município, UF, Status e Área

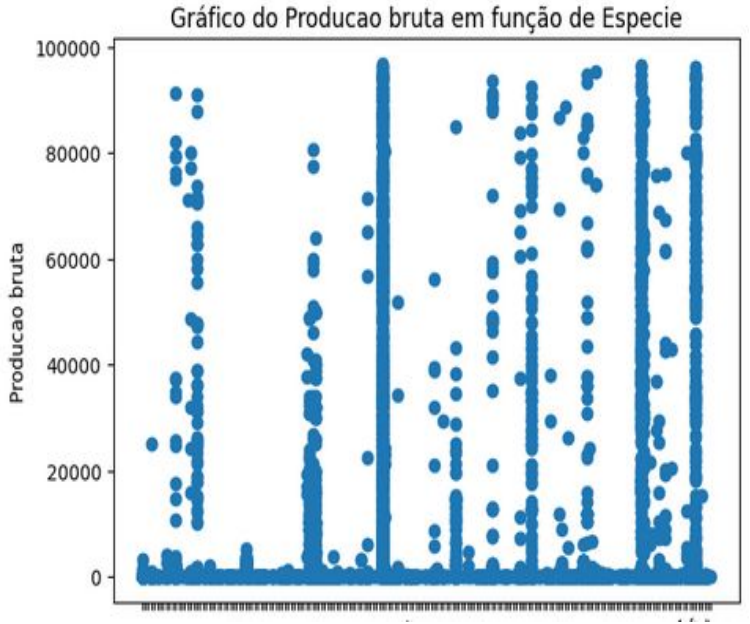
# Distribuição dos dados - Safra



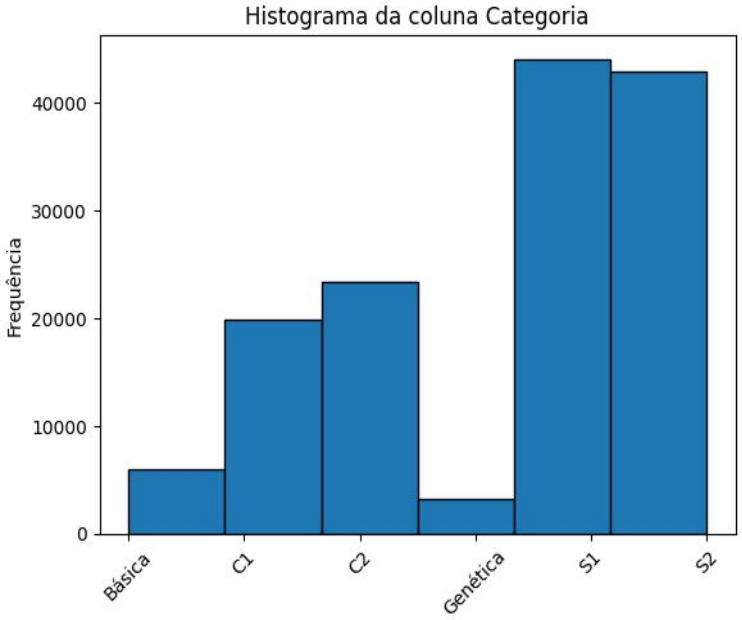



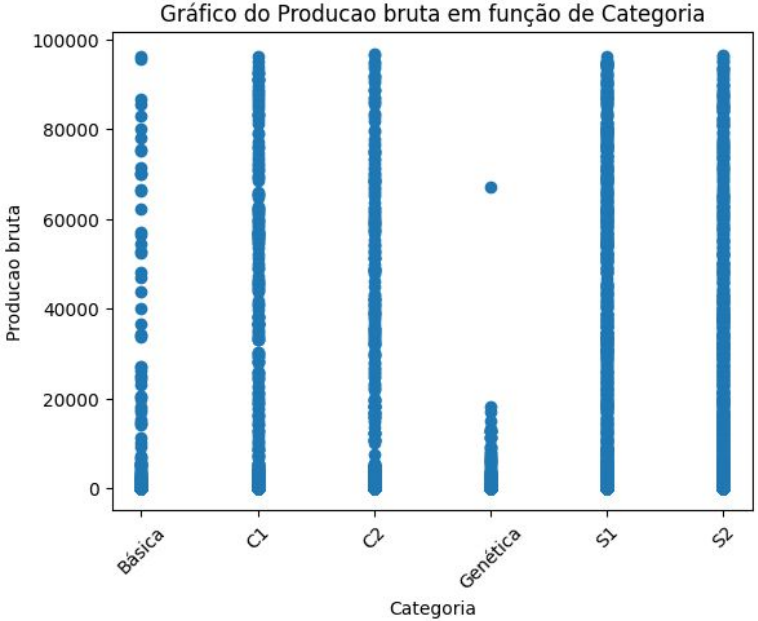
# Distribuição dos dados - Espécie



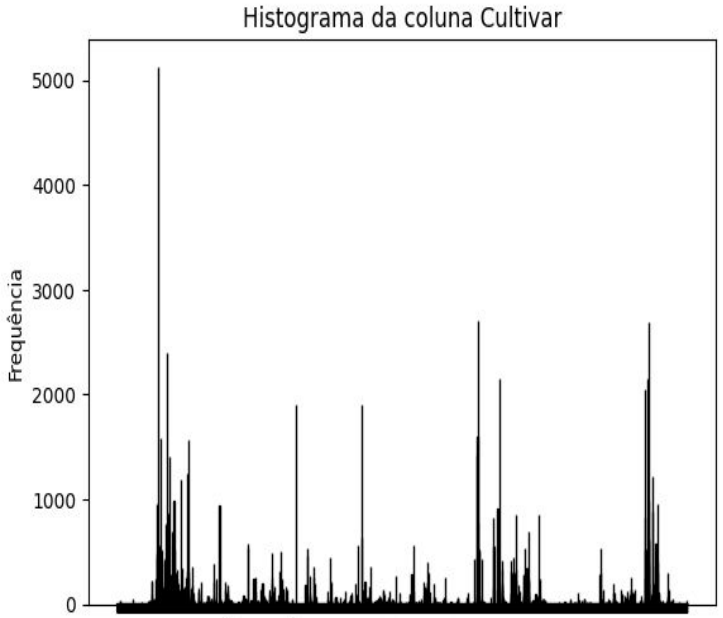



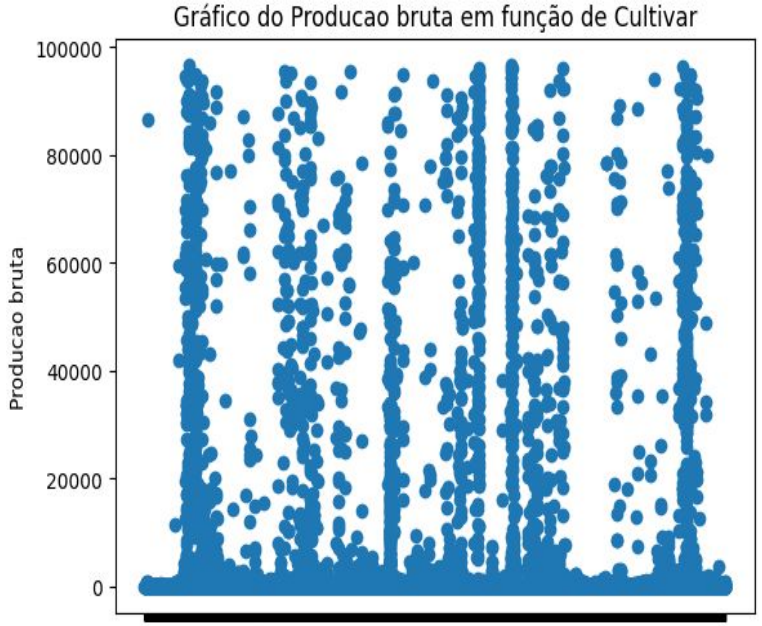
# Distribuição dos dados - Categoria



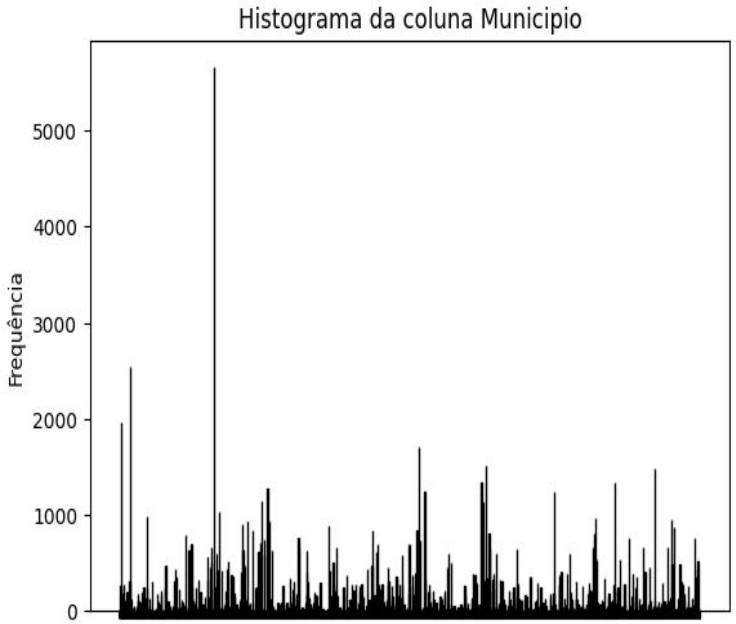



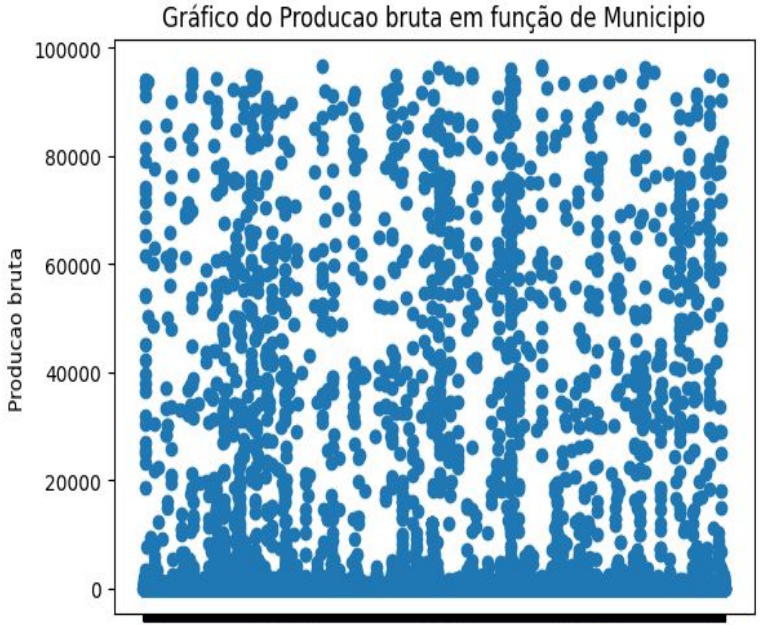
# Distribuição dos dados - Cultivar



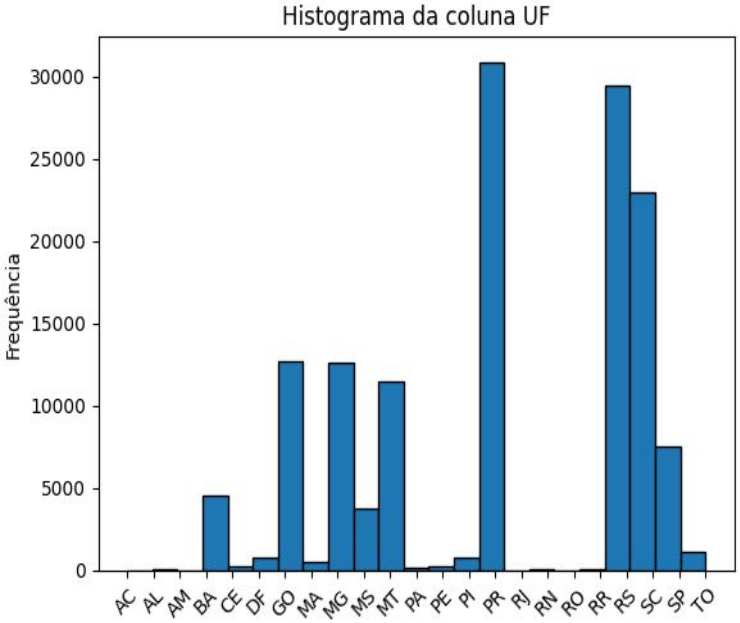



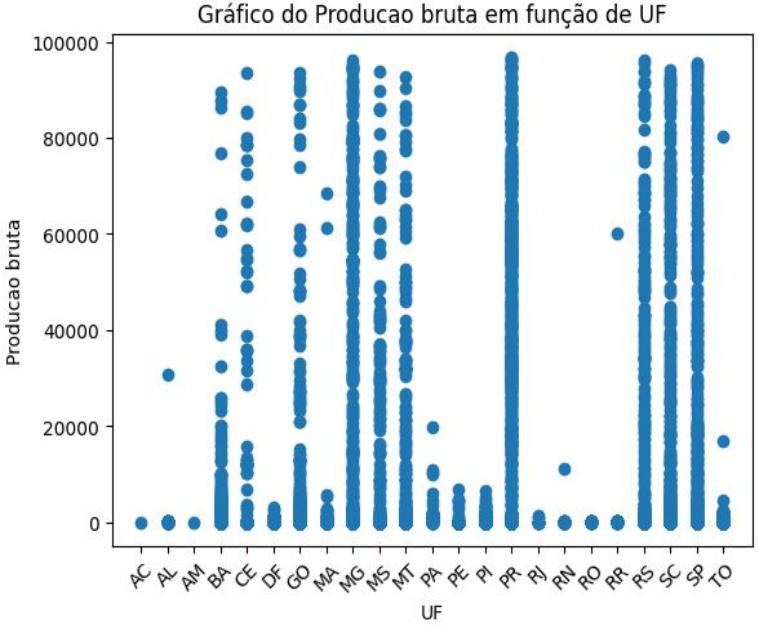
# Distribuição dos dados - Município



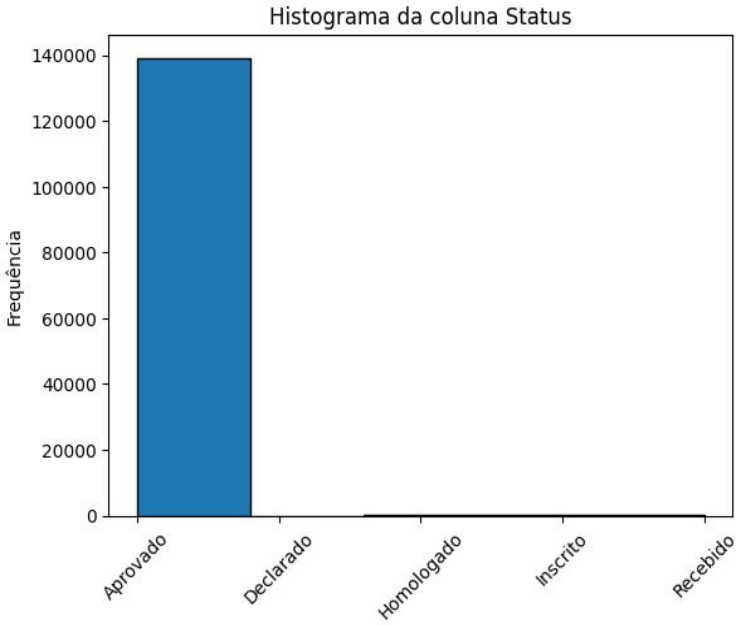



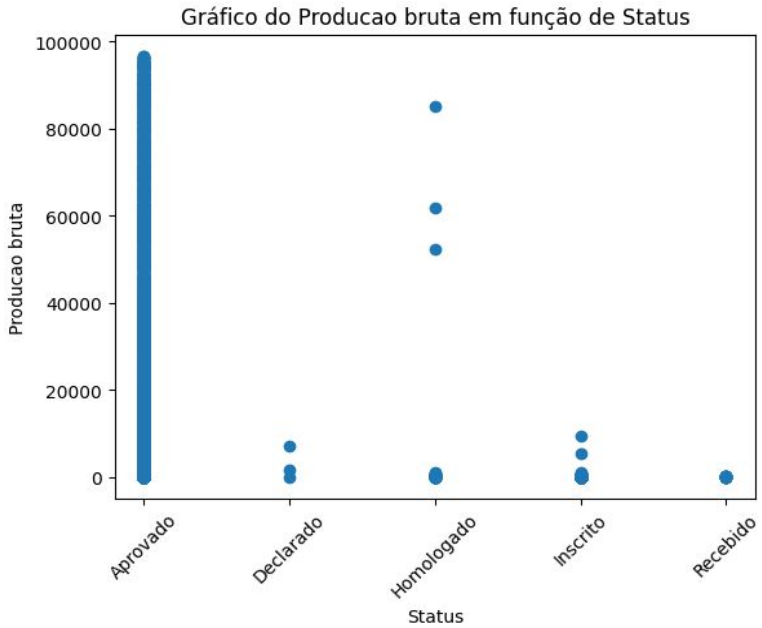
# Distribuição dos dados - UF



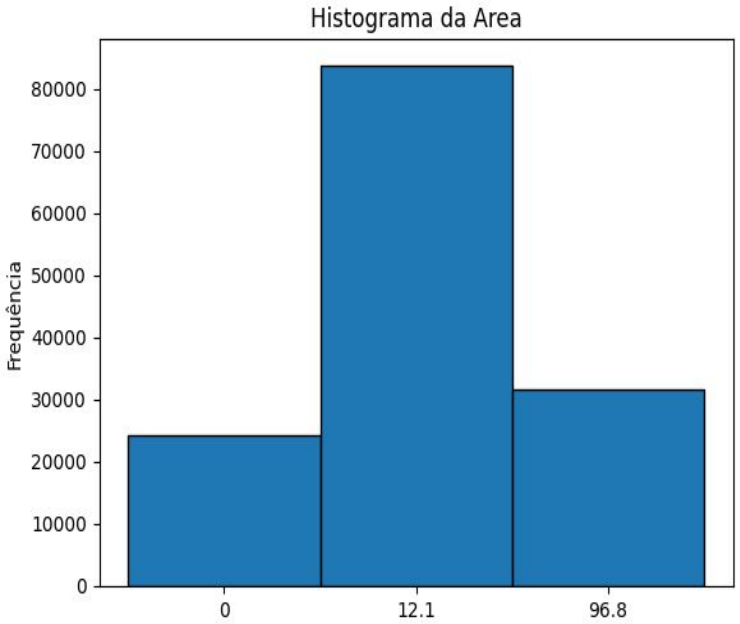



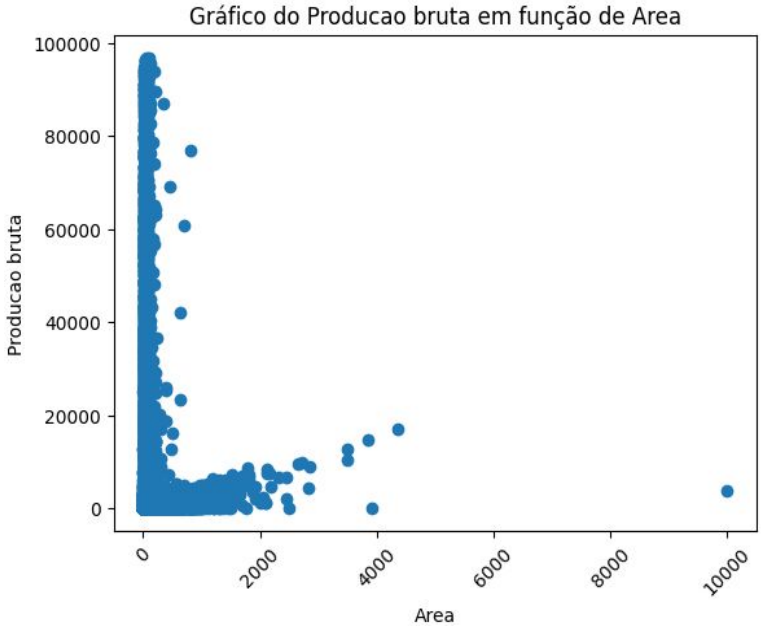
# Distribuição dos dados - Status



# Distribuição dos dados - Área



# Atributos considerados

- A partir dos resultados da análise estatística, conclui-se que devem ser utilizados, para a predição, os atributos: Categoria, UF, Área, Espécie e Safra.
- Como a qualidade das correlações para Espécie e Safra não é tão boa, os modelos serão treinados sobre 4 conjuntos de atributos:
    - Categoria; UF; Área.
    - Categoria; UF; Área; Espécie.
    - Categoria; UF; Área; Safra.
    - Categoria; UF; Área; Espécie; Safra.

# Normalização

- Todos os dados foram normalizados com o *standard scaler*:

$$z = \frac{x - \bar{x}}{\sigma}$$

# Modelos de predição

# Modelos de predição

- Foram testados diversos algoritmos (e parâmetros) de predição, a fim de encontrar o melhor modelo para melhorar a estimativa de produção de sementes, em relação à fornecida pelo *dataset*.
- Os modelos utilizados são fornecidos pela biblioteca *Scikit Learn*.
- Algoritmos testados:
    - Regressão linear; Regressão linear positiva
    - Árvores de decisão Comum, Friedman MSE, Aleatória; Random Forest
    - KNN (K Nearest Neighbors); Gradient Boost
    - MLP (Multilayer Perceptron)

# Análise de resultados

# Métricas

- Após o treinamento de cada um dos 10 modelos para cada um dos 4 *datasets*, foi realizada a predição para cada registro e comparada com a produção bruta reportada (real), sendo utilizada como métrica de erro o RMSE.

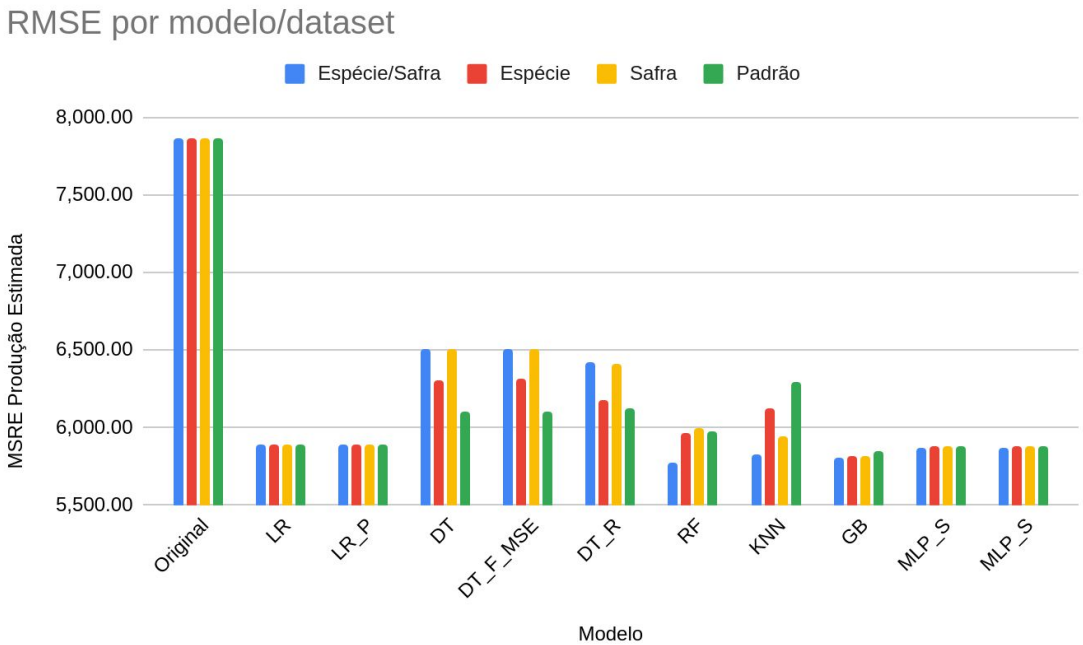
# Métricas

| Modelo\Dataset | Espécie/Safra | Espécie | Safra | Padrão |
|---|---|---|---|---|
| Original | 7,864.79 | 7,864.79 | 7,864.79 | 7,864.79 |
| LR | 5,888.85 | 5,889.51 | 5,890.24 | 5,890.84 |
| LR_P | 5,890.26 | 5,890.26 | 5,892.03 | 5,892.03 |
| DT | 6,506.80 | 6,308.22 | 6,508.80 | 6,105.72 |
| DT_F_MSE | 6,511.80 | 6,318.04 | 6,510.22 | 6,107.31 |
| DT_R | 6,420.10 | 6,174.64 | 6,406.49 | 6,129.58 |
| RF | **5,774.77** | 5,966.45 | 5,993.81 | 5,977.68 |
| KNN | 5,828.03 | 6,129.18 | 5,943.72 | 6,292.39 |
| GB | 5,806.71 | **5,820.14** | **5,822.40** | **5,848.85** |
| MLP_S | 5,873.70 | 5,876.32 | 5,881.44 | 5,884.91 |
| MLP_S | 5,871.11 | 5,875.97 | 5,881.09 | 5,884.51 |

RMSE por modelo/dataset

Em negrito, o melhor resultado para cada dataset. Destacado em verde, o melhor em geral.

# Métricas


RMSE por modelo/dataset
Espécie/Safra
Espécie
Safra
Padrão
8,000.00
7,500.00
7,000.00
6,500.00
6,000.00
5,500.00
MSRE Produção Estimada
Original
LR
LR_P
DT
DT_F_MSE
DT_R
RF
KNN
GB
MLP_S
MLP_S
Modelo

# Conclusões

- Todos os modelos performaram bem melhor que a estimativa presente no conjunto de dados.
- Além disso, o algoritmo de Gradient Boost foi o mais preciso na maioria dos *datasets*.
- Em geral, entretanto, a melhor performance foi exibida pelo Random Forest, no conjunto com Espécie e Safra, apresentando redução de 26,5% no RMSE em relação à original.
- Em contrapartida, a pior precisão foi da Decision Tree (Friedman MSE) no mesmo *dataset*, com redução de 17,2% no RMSE.
- Ou seja, até no pior caso, houve melhora considerável na qualidade da predição.

# Fim