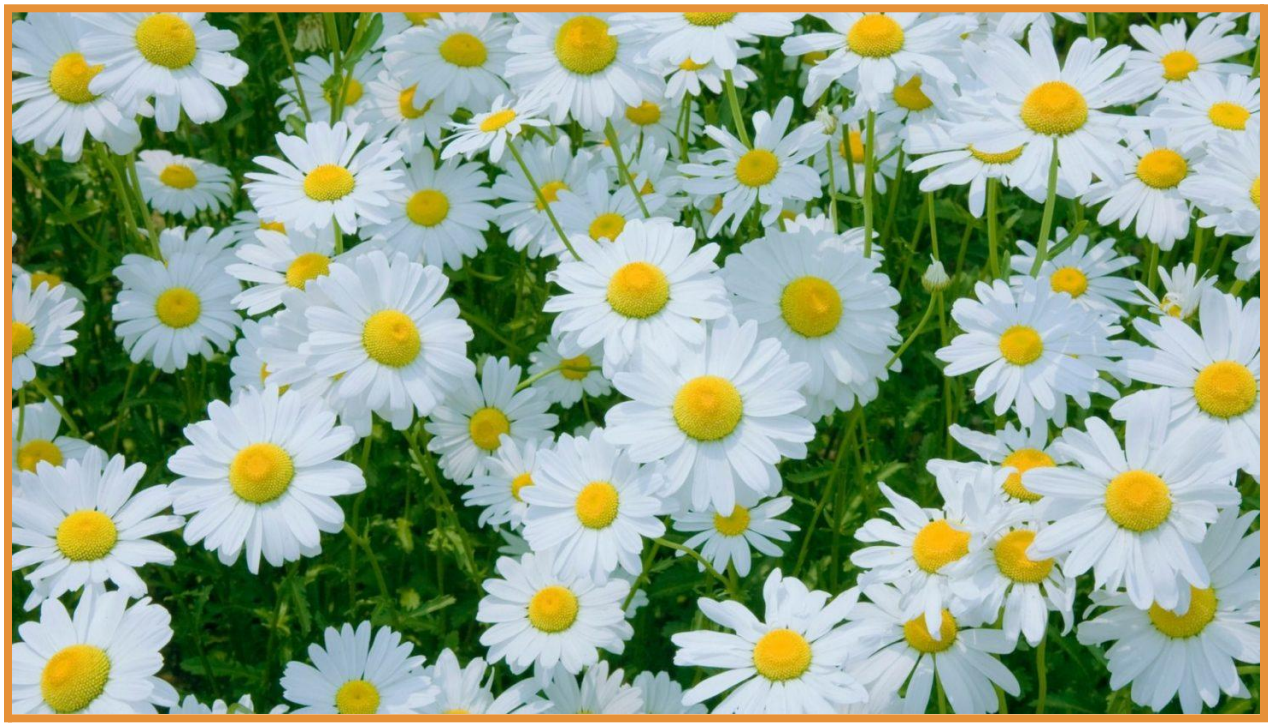
# Teste ANOVA, teste de Tukey e Regressão Logística

*Um exemplo aplicado à base "iris"*



**Guilherme Ceacero**

30/12/2023

## INTRODUÇÃO

A análise da base de dados “iris” tem como objetivo investigar possíveis diferenças no comprimento das pétalas entre diferentes espécies de flores. Utilizando técnicas de visualização de dados e testes estatísticos, pretende-se explorar e entender as características distintivas das espécies presentes no conjunto de dados, e tomar decisões precisas de classificação de espécie de flores dado que só observamos uma determinada característica, tal como o comprimento da pétala da flor.

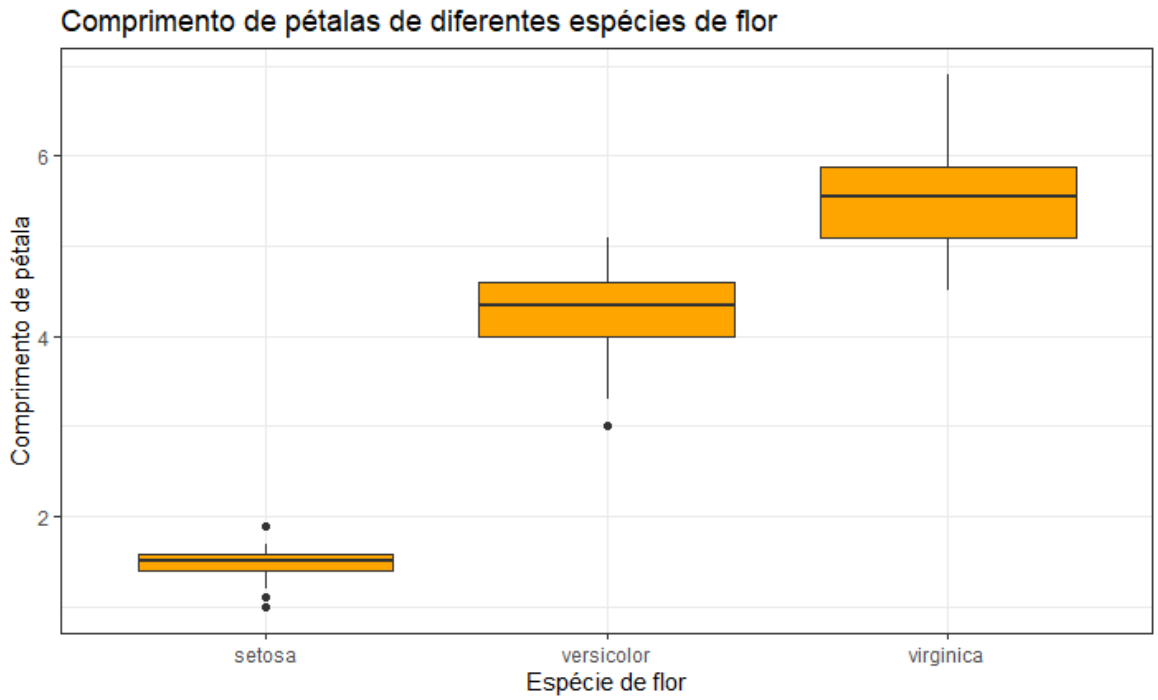
## MATERIAIS

- Pacotes R Utilizados:
    - ggplot2: Para visualização gráfica.
    - dplyr: Para manipulação de dados.
    - broom: Para predição a partir de modelos.
- Conjunto de Dados:
    - Foi utilizada a base de dados “iris” nativa do R, que contém dados sobre flores, tais como espécie, comprimento de pétala e comprimento de sépala.
- Testes de Hipóteses:
    - Teste de Shapiro-Wilk para avaliar a normalidade da distribuição do comprimento de pétalas dentro de cada espécie de flor
    - Teste ANOVA para avaliar se existe diferença estatisticamente significativa no comprimento médio de pétalas entre cada uma das espécies estudadas
    - Teste de Tukey para encontrar quais pares de espécies possuem diferença estatisticamente significante no comprimento médio das pétalas
- Ferramentas Gráficas:
    - Gráfico de Boxplot para visualizar o comprimento médio de pétalas em cada espécie de flor
    - Gráfico de histograma para visualizar se a curva da distribuição Normal ajusta bem os dados da base
    - Gráfico de Intervalos de Confiança de Tukey para avaliar quais pares de

espécies de flor possuem diferença estatisticamente significante em comprimento de pétala
- Gráfico de dispersão com ajuste de modelo de regressão logística

## RESULTADOS

Visualização Inicial - para uma primeira abordagem, criou-se um gráfico de boxplot para visualizar o comportamento do comprimento das pétalas em cada espécie de flor. O gráfico indicou diferenças aparentes.

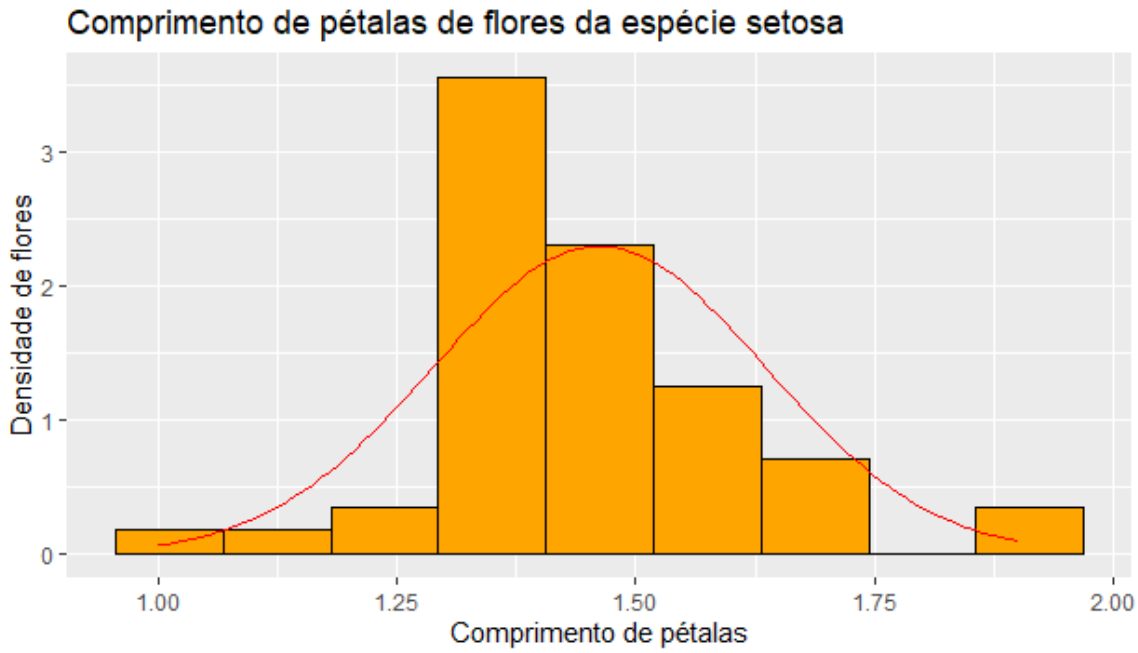


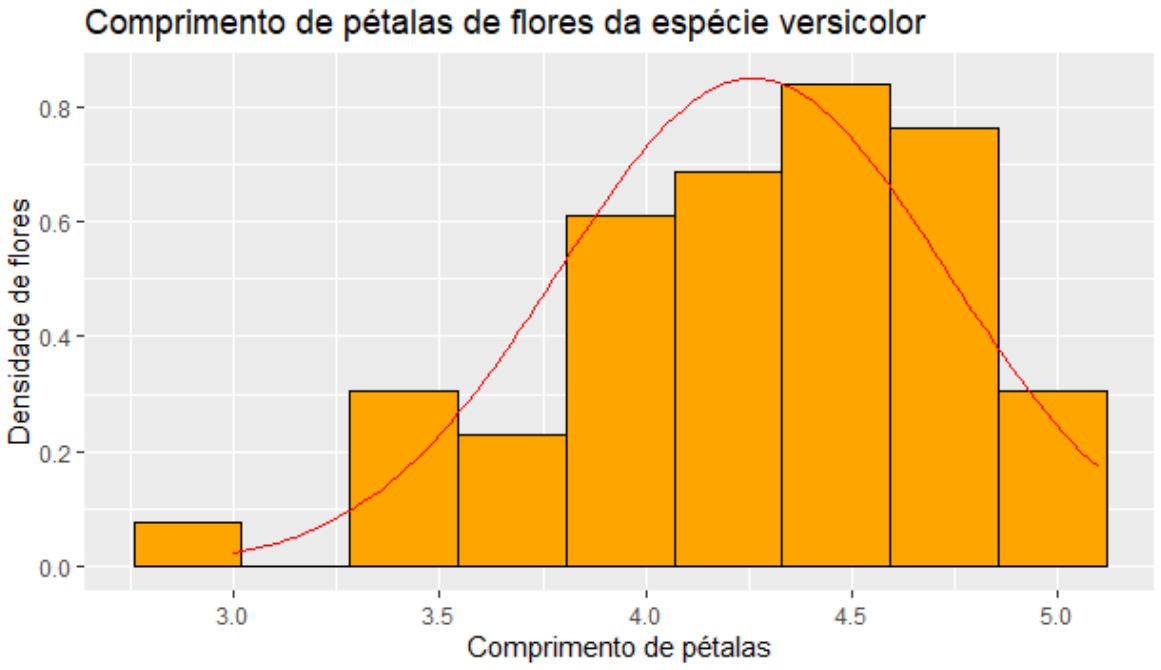

E então, já se pensando em realizar o teste ANOVA e o teste de Tukey para avaliar essas diferenças vistas graficamente, foram realizados testes de Shapiro-Wilk para avaliar a normalidade da distribuição de comprimento de pétalas dentro de cada espécie de flor.

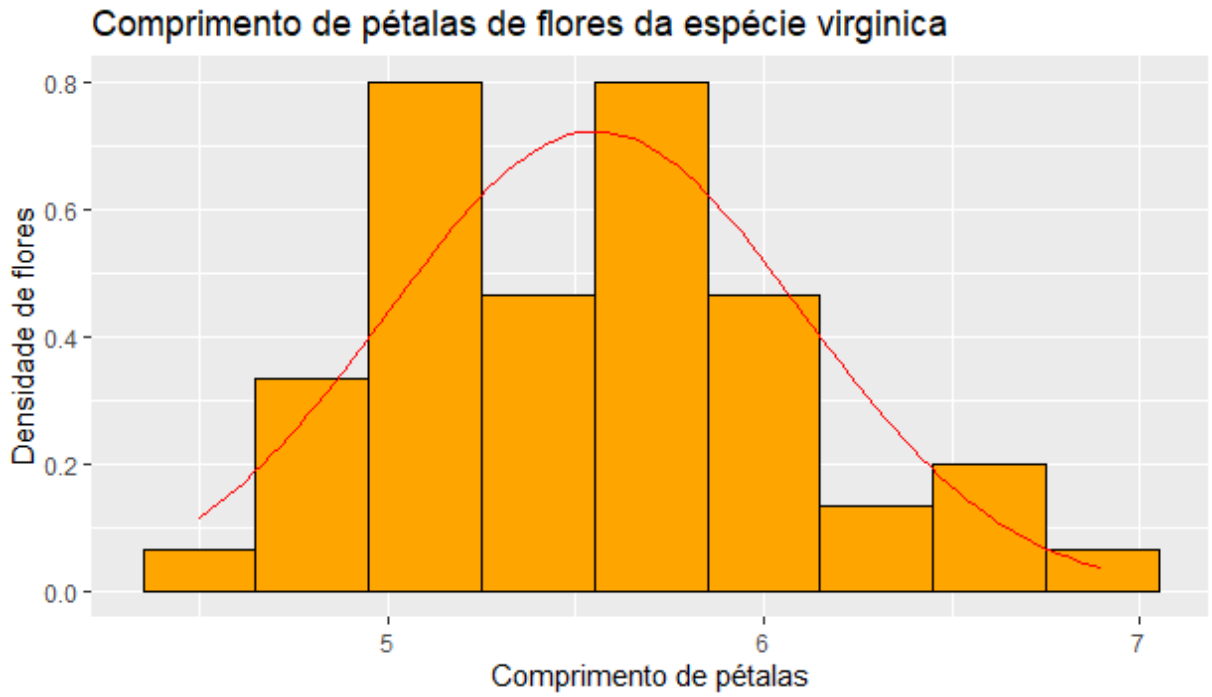
Inicialmente, para a espécie *setosa*, o p-valor encontrando no teste de Shapiro-Wilk foi 0.05481, valor perto do limite de aceitação da hipótese nula de normalidade, dado que o índice de significância utilizado nesse estudo será de 5%.

Para a espécie *versicolor*, o p-valor encontrado foi de 0.1585; e para a espécie *virginiana*, o p-valor encontrado foi de 0.1098. Adotando o índice de significância de 5%, não rejeita-se a hipótese nula em nenhum dos testes, isto é, de que a distribuição do comprimento das pétalas das flores em cada espécie de flor segue distribuição Normal.

Para melhor visualizar este fato, foram construídos gráficos de histograma dos dados de comprimento de pétala juntamente com o ajuste da curva da distribuição Normal teórica, com média e variância igual à do conjunto de dados. Veja:






E, graficamente, os ajustes foram suficientemente bons para dar prosseguimento aos testes de ANOVA e Tukey. Veja o resultado do teste de ANOVA:

```
             Df Sum Sq Mean Sq F value Pr(>F)

iris$Species   2  437.1  218.55    1180 <2e-16 ***
Residuals    147   27.2    0.19
```

O p-valor resultante do teste ANOVA construído para avaliar se diferentes espécies possuem diferença estatisticamente significante no tamanho de suas pétalas é muito baixo. Dessa forma, tem-se evidências para inferir que a diferença no tamanho das pétalas em cada espécie de flor estudada é estatisticamente significante. Agora, afim de descobrir exatamente quais pares de flores diferem, foi escolhido o teste de Tukey.

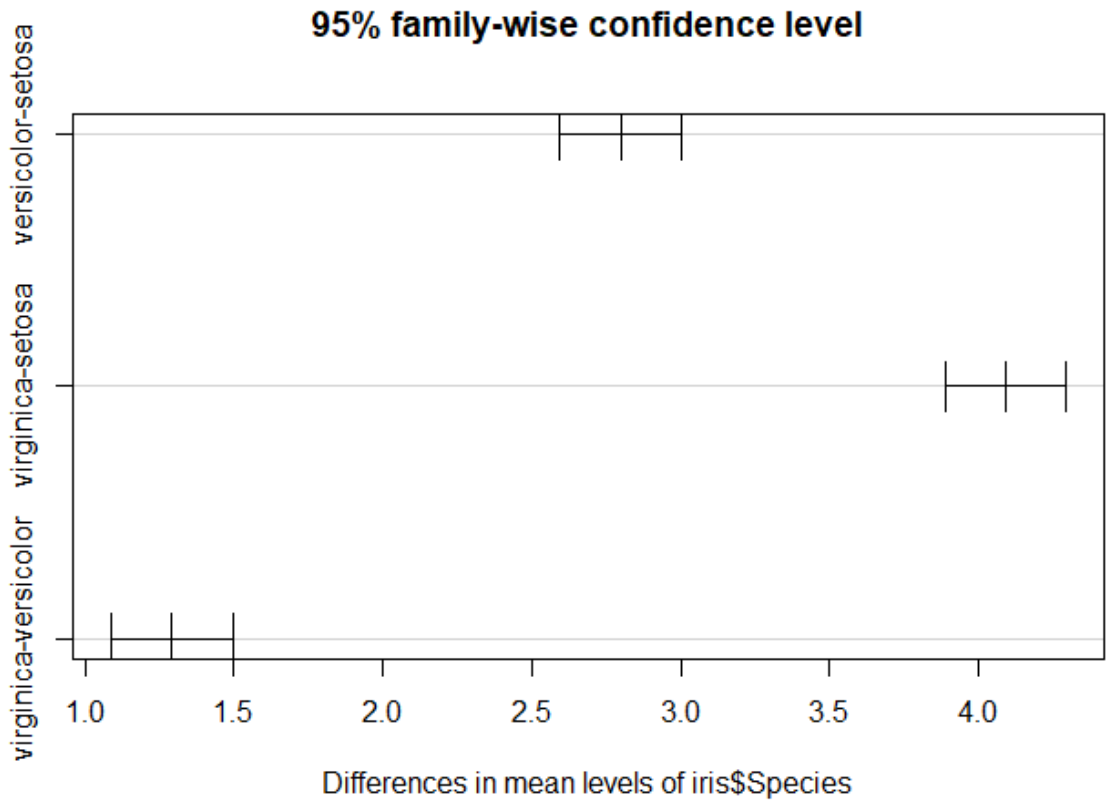
A princípio, como visto no primeiro gráfico, era de se esperar que a espécie *setosa* seria um valor discrepante o suficiente para o teste de ANOVA apontar diferença, e naturalmente, o teste de Tukey apontar que o par *setosa-virginica* e *setosa-versicolor* são diferentes. Mas, quanto ao par *versicolor-virginica,* não se podia ter certeza apenas pelo gráfico. Veja os resultados do teste de Tukey:

```
Tukey multiple comparisons of means
    95% family-wise confidence level
    factor levels have been ordered

Fit: aov(formula = iris$Petal.Length ~ iris$Species)

$`iris$Species`
                      diff     lwr     upr p adj
versicolor-setosa    2.798 2.59422 3.00178     0
virginica-setosa     4.090 3.88622 4.29378     0
virginica-versicolor 1.292 1.08822 1.49578     0
```

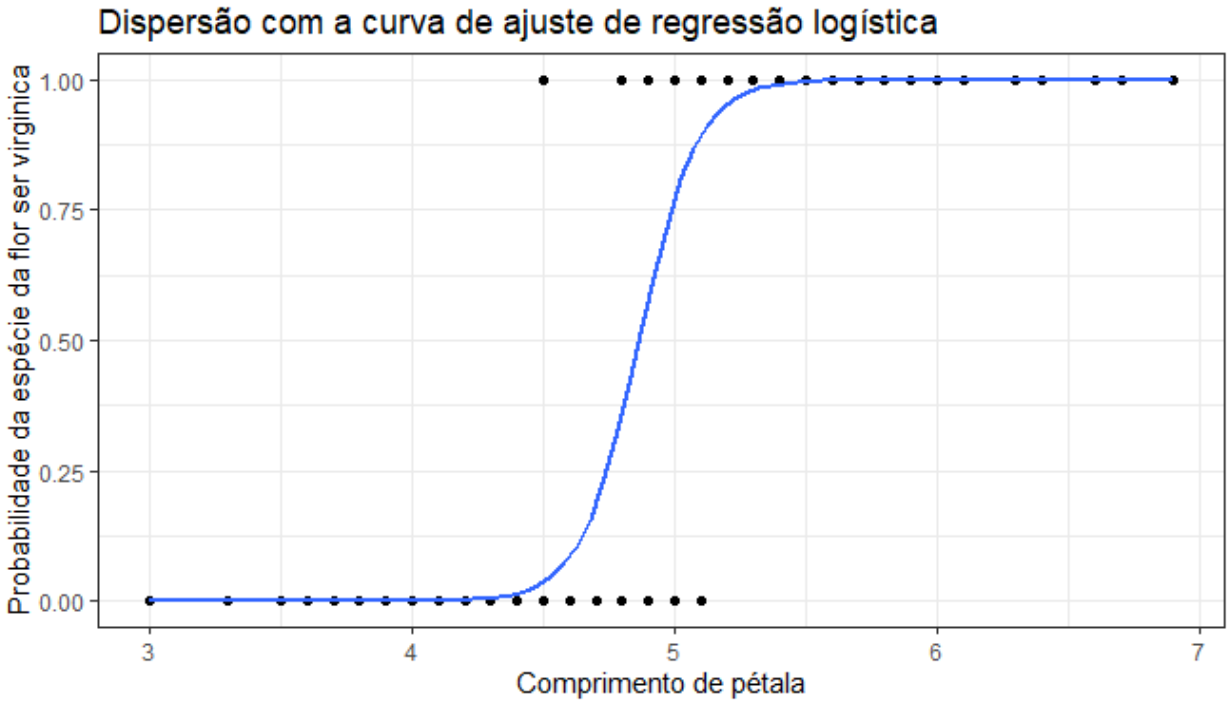
O teste evidencia que, em fato, todos os pares possuem diferença estatisticamente significante; inclusive o par *versicolor-virginica.* Graficamente, fica mais fácil visualizar o resultado do teste:




Os intervalos de confiança no gráfico representam os intervalos estimados para as diferenças no comprimento de pétalas entre cada par de espécies. Veja que nenhum dos intervalos de confiança contém o valor 0; e por isso, pode-se afirmar que as espécies todas possuem diferença estatisticamente significante entre si, no que se refere ao comprimento de pétalas.

Quanto à modelagem estatística, o foco será ajudar na tomada de decisão referente à classificação da espécie quando se observa o comprimento de pétala da flor. Nesse caso, como a espécie *setosa* possui pétalas consideravelmente menores, o modelo será ajustado levando em consideração somente as espécies *versicolor-virginica*.

Para a construção do gráfico, foi criada na base de dados uma variável *dummy* que assume valores 0 ou 1; zero quando a linha for referente à uma flor da espécie *versicolor*, e 1 quando a linha for referente à uma flor da espécie *virginica*. Com essa variável *dummy*, então, pode-se ajustar um modelo de regressão logística indicando a probabilidade de uma flor ser de uma espécie específica dada a informação do comprimento de pétala.




Com o modelo, então, pode-se calcular a probabilidade de uma nova flor observada pertencer à uma espécie apenas pelo comprimento de pétala. E com essa probabilidade, seguindo uma regra de decisão, estimar a espécie da flor.

## CONCLUSÃO

Os resultados obtidos sugerem que há diferenças significativas nos comprimentos das pétalas entre as espécies de flores estudadas. Além disso, o modelo de regressão logística demonstrou ser uma ferramenta útil para prever a espécie da flor com base no comprimento da pétala.