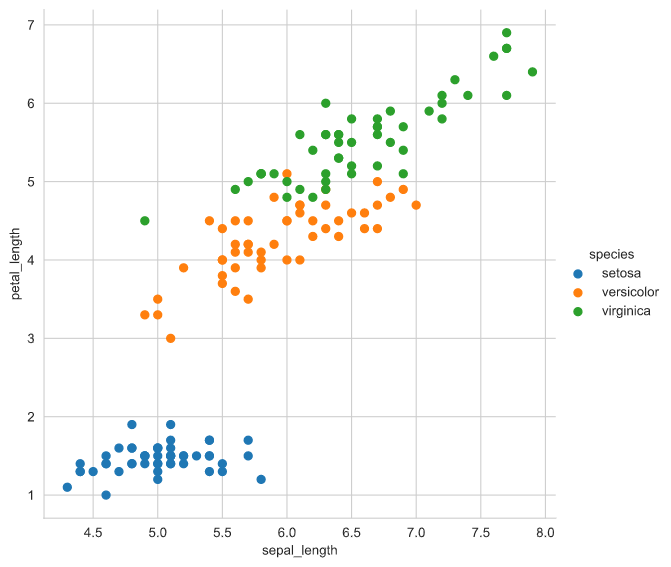
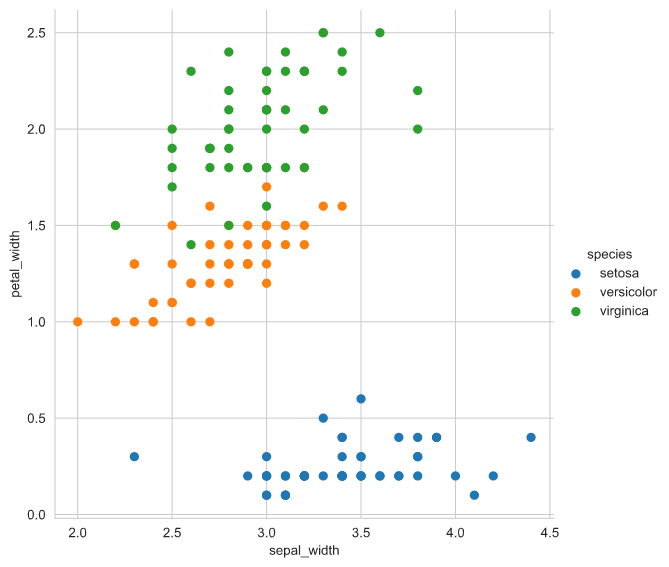
**Trabalho de Grupo**

**K Nearest Neighbors**

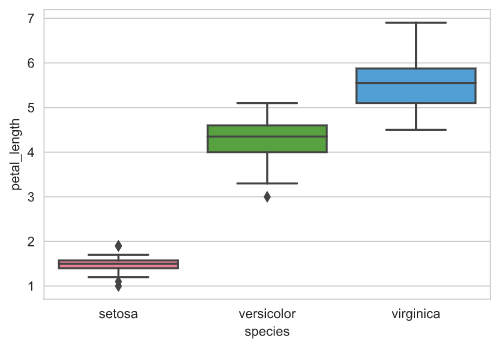
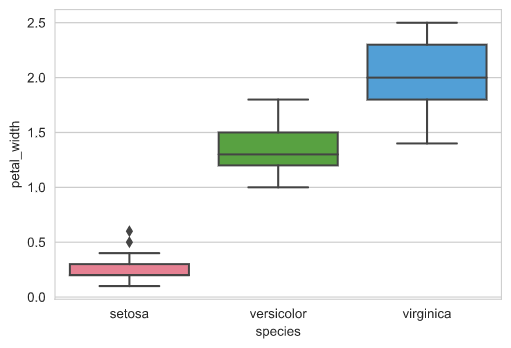
Em Data Mining, K Nearest Neighbors é um método de classificação supervisionada, sem parâmetros, que utiliza os pontos mais próximos de uma nova observação para determinar a sua classificação.

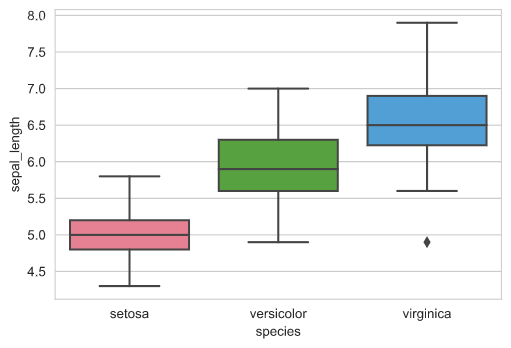
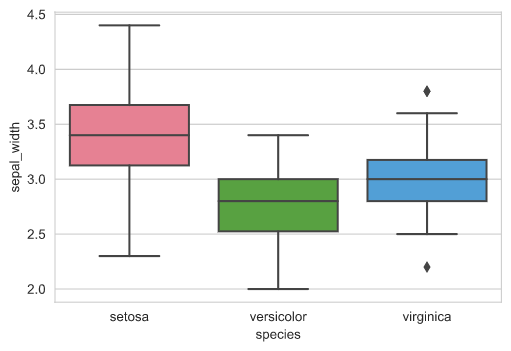
Na realização deste trabalho usamos o dataset da iris, este tem 150 amostras divididas em três classes *Iris Setosa*, *Iris Virginica* e *Iris Versicolor*. Existem quatro variáveis que foram medidas em cada amostra correspondendo ao comprimento e largura das sépalas e pétalas, em centímetros. A partir da combinação destas características Fisher desenvolveu um modelo discriminante linear para distinguir as espécies.

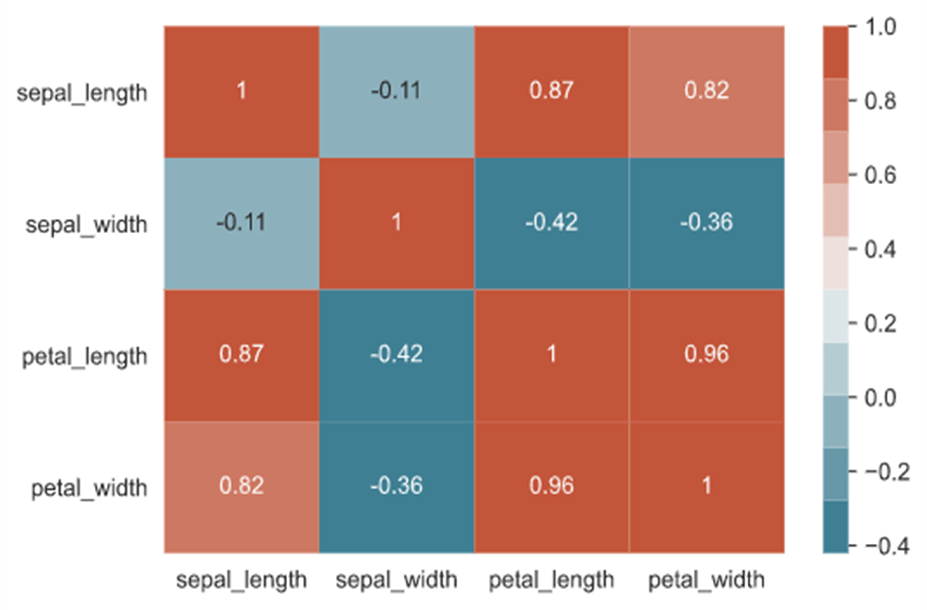
Analisando graficamente as variáveis observamos que a espécie *setosa* corresponde a valores mais baixos relativamente ao comprimento das pétalas e das sépalas, bem como a valores mais baixos da largura da pétala, no entanto de uma forma geral observamos que tende a ter valores mais altos para o comprimento da sépala. Este comportamento da espécie *setosa* faz com que seja mais fácil classificá-la.



Pela análise de boxplot verificamos que existem alguns outliers, no entanto reforça a nossa conclusão relativamente à espécie *setosa*. Tanto pelos gráficos anteriores como pelos boxplot podemos afirmar que a espécie *virginica* terá valores mais altos para o comprimento e a largura das pétalas, bem como para o comprimento das sépalas. Conseguimos observar também que o que define a espécie *versicolor* é ter em geral valores mais baixos em todas as variáveis relativamente à espécie *virginica.*





Pela tabela de correlação verificamos que existe uma alta correlação entre a largura da pétala e o comprimento da sépala. Também se verifica uma grande correlação entre o comprimento da pétala e da sépala. O comprimento e a largura da pétala são aqueles que possuem uma maior correlação o que nos poderá indicar que poderíamos optar apenas por uma das variáveis de forma a não termos informação redundante no nosso dataset.