Projeto 2 - Resultados

Nesse projeto, analisamos um conjunto de 3 diferentes flores com características visuais muito próximas. O conjunto de dados é constituído por 150 objetos (50 de cada classe) com 4 parâmetros que as especificam. Iniciando o processo de caracterização, cruzamos os dados plotando pontos de dois desses parâmetros por vez, alternando quais parâmetros utilizávamos. Tais gráficos podem ser observados na Figura 1

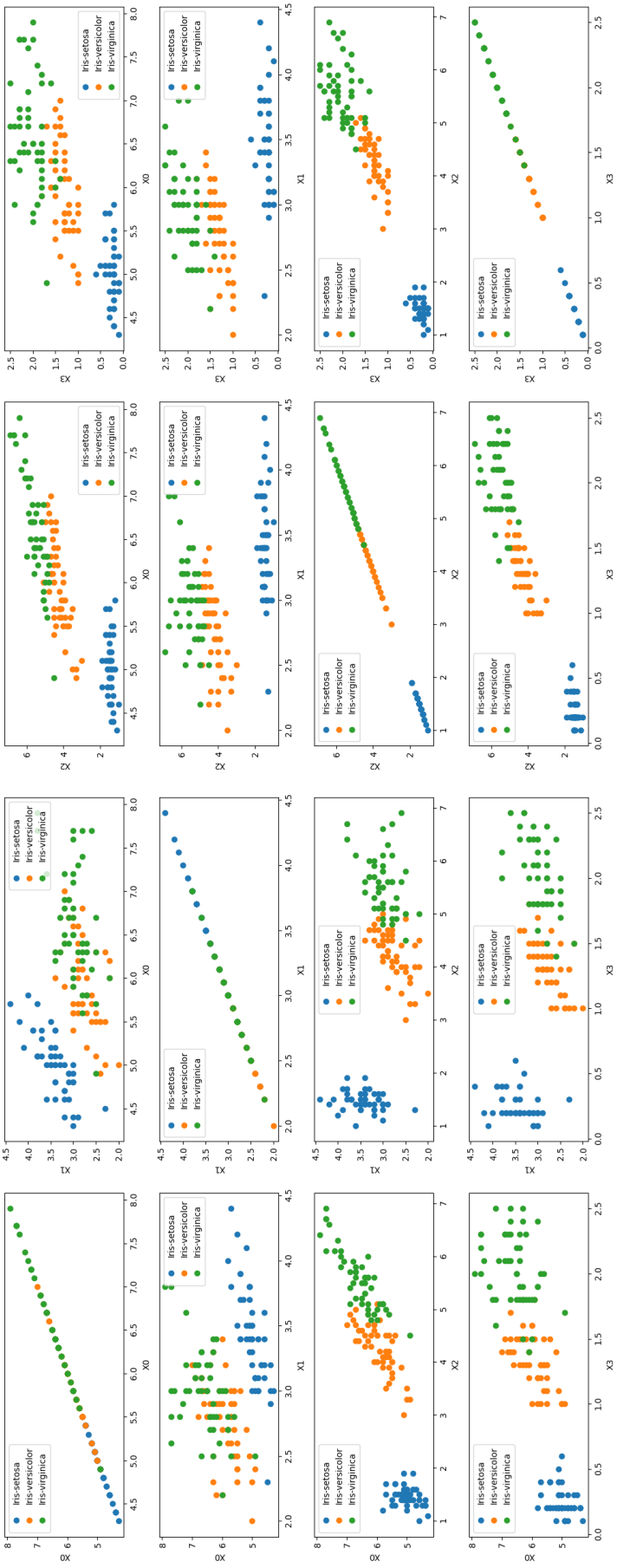

Figura 1: Gráficos de 2 parâmetros alternados.

Dividimos então o conjunto de dados em duas partes iguais (dois conjuntos de 75 flores com 25 de cada tipo) chamadas *train* e *test*. Com o intuito de treinar um algorítimo capaz de diferenciar uma flor a partir de seus atributos, construímos uma função que utiliza o algorítimo de classificação chamado de k-vizinhos. Tal algorítimo classifica um objeto com base na distância euclidiana entre ele e os k objetos conhecidos mais próximos. Em nosso caso, utilizamos $k = 1$ para a classificação. Tentando encontrar a dupla de parâmetros que melhor separa os objetos de cada classe, fizemos o teste de classificação para todas as duplas apresentadas na Figura 1.

A partir da classificação de todos os objetos da classe *test*, implementamos uma função que apresenta a matriz de confusão de cada um dos classificadores treinados. Os resultados dos treinamentos podem ser vistos na Figura 2.

Algumas considerações devem ser feitas sobre o treinamento. O algorítimo implementado se baseia na fixação de um dos objetos da classe *test* e no cálculo das distâncias entre tal objeto e todos os objetos da classe *train* a partir do conjunto de dois dos quatro parâmetros encontrados no dataset. Como são 4 parâmetros, teremos 16 pares de parâmetros. Após os cálculos das distâncias, construímos as matrizes de confusão da Figura 2 passando por cada uma das classificações e somando um ao respectivo local de tal classificação na matriz (de acordo com o rótulo real e o encontrado) seguida da divisão de cada item da matriz por 25, pois esse era o número máximo de flores por categorias.

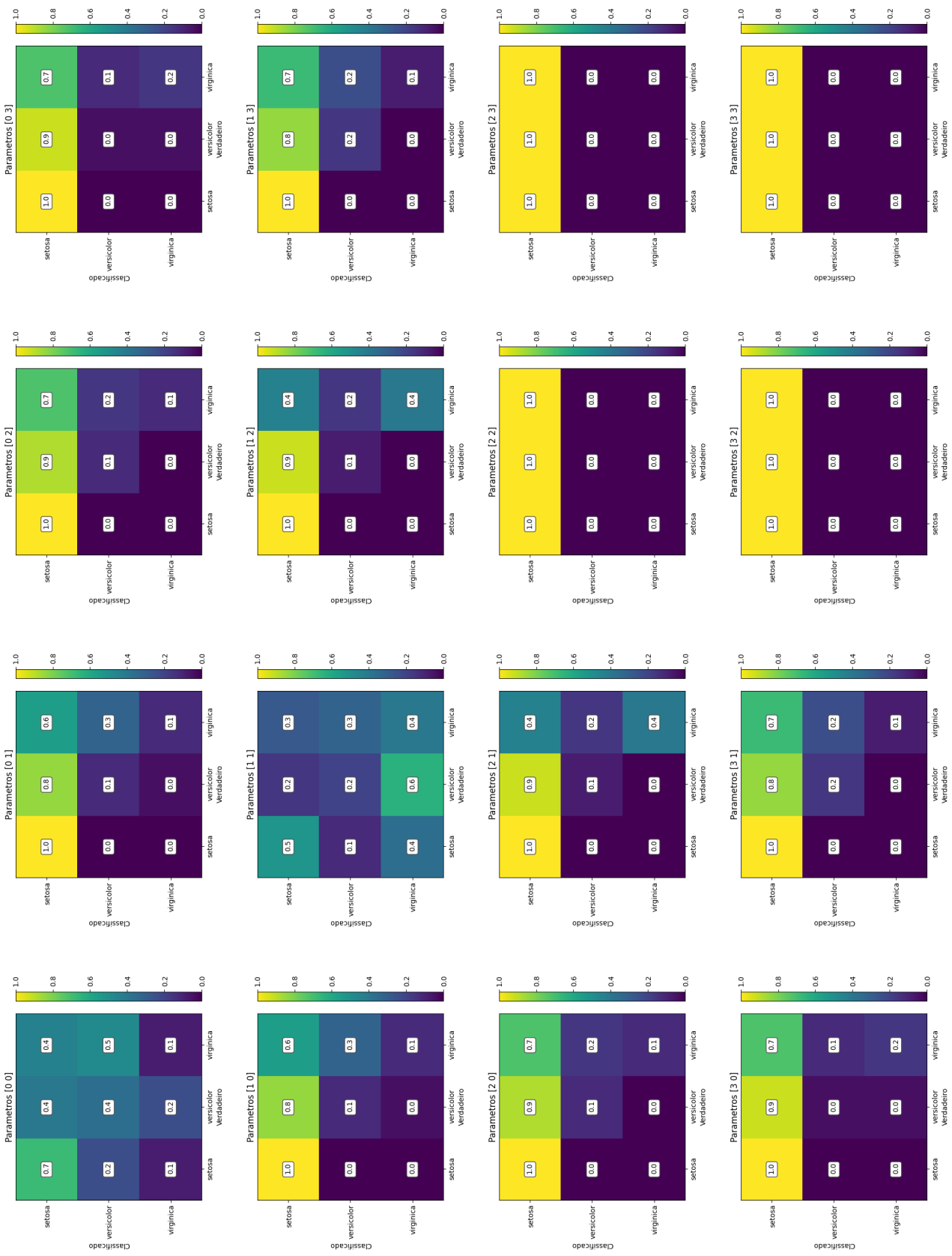

Figura 2: Matrizes de confusão respectivas a cada dupla de parâmetros.

Pode-se notar que a categoria mais fácil de ser diferenciada é a da "Iris-setosa"que apresentou

desempenho melhor que 0.5 em todos os classificadores. Contudo, buscamos encontrar o conjunto de parâmetros que maximiza a diagonal da matriz, isto é, maximiza a classificação de rótulos corretos. Vendo esse a partir desse aspecto, consideramos os parâmetros 1 e 2 como os melhores.

Vale salientar ainda que os conjuntos de parâmetros (2 e 2, 2 e 3, 3 e 2, 3 e 3) nos deram as piores classificações com 100% de acerto no rótulo "Iris-setosa"e 100% de erro nas outras duas categorias. Tal comportamento se mostra inesperado e possivelmente pode ser relacionado a erros na implementação do método.