Weka - Dataset Iris

William Hitoshi Sumida

Abril 2020

1 Análise dos Dados

Ao abrir o dataset iris, podemos notar que apresenta 150 amostras, sendo formadas por cinco atributos: comprimento e largura da pétala, comprimento e largura da sépala e a espécie da íris. A tabela 1 apresenta uma visão geral dos dados observados contemplando o menor (Min) e maior (Max) valor de cada atributo, quantidade de amostras faltando (Missing).

	Min	Max	Missing (%)
Comprimento pétala	1,0	6,9	0
Largura pétala	0,1	2,5	0
Comprimento sépala	4,3	7,9	0
Largura sépala	2,0	4,4	0

Table 1: Valores máximos, mínimos e nulos dos atributos.

Podemos também observar na tabela 2 que existem apenas três espécies de íris no dataset e estão igualmente distribuídas, portanto não é necessário fazer balanceamento dos dados.

	Quantidade
Íris setosa	50
Íris versicolor	50
Íris virginica	50

Table 2: Quantidade de amostras de cada espécie

Com base nos dados analisados, podemos deduzir que se trata de um problema de classificação. Como o atributo alvo (espécie de íris) é composto por mais de 2 classes, concluímos que o tipo de classificação é multi-classes.

Utilizando a aba visualize 1 do Weka, são criados gráficos de dispersão com a correlação de todos os atributos entre si.

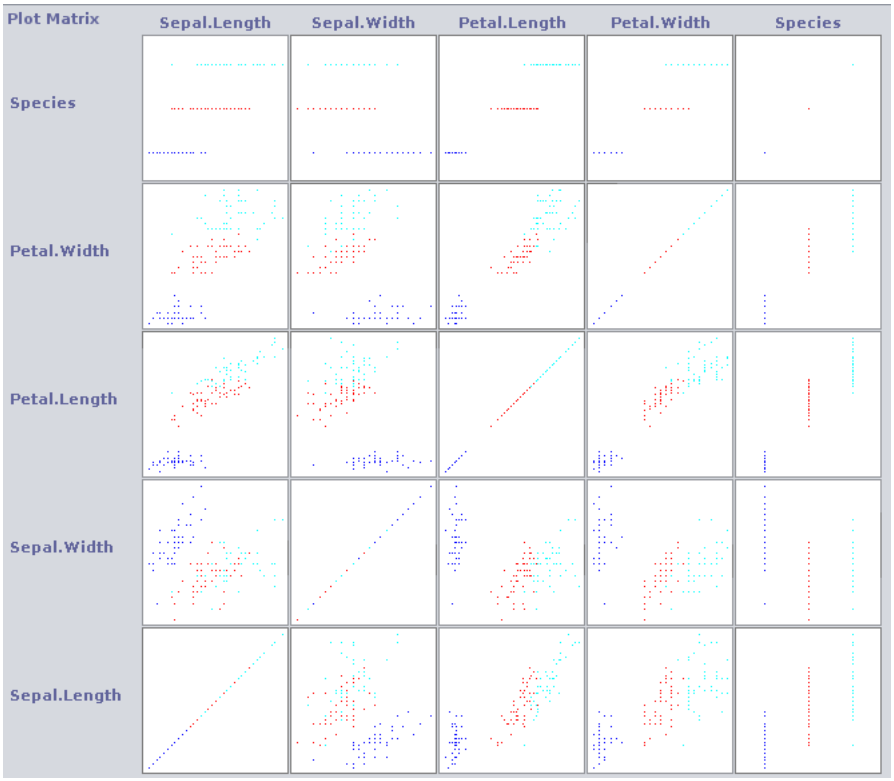

Figure 1: Gráficos de dispersão com a correlação entre os atributos.

Na figura 2 a correlação entre os atributos Comprimento da sépala e Largura da pétala separa as três classes em regiões bem definidas do plano cartesiano, e dentre as outras correlações, é a que melhor define o problema visualmente.

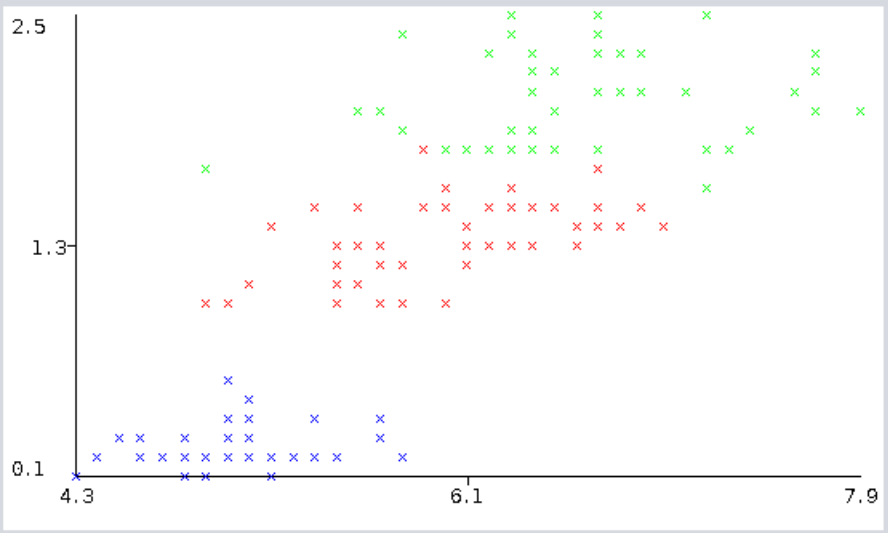

Figure 2: Correlação dos atributos Comprimento da sépala e Largura da pétala.

Também podemos observar que a espécie setosa (identificada pelo azul) possui as coordenadas mais isoladas das outras, assim como pode ser observado nas Figuras 2 e 3.

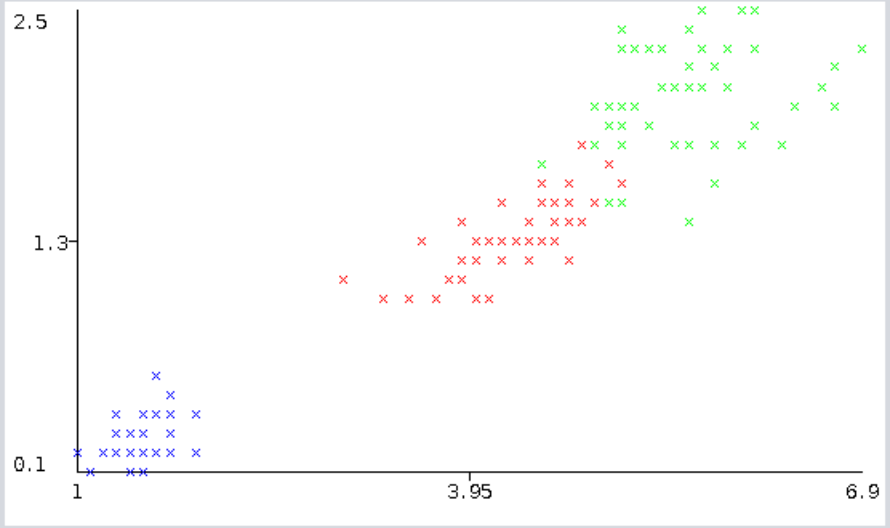

Figure 3: Correlação dos atributos Comprimento da pétala e Largura da pétala.

2 Conclusão

Podemos concluir que o dataset Iris é composto por 5 atributos, quatro atributos numéricos representando comprimento e largura da pétala, comprimento e largura da sépala da íris, e um atributo nominal com a espécie.
Trata-se de um problema de classificação multi-classes, não sendo necessário balanceamento dos dados devido à uniformidade dos dados.
Ao fazer a correlação de todos os atributos, nota-se que visualmente o comprimento da sépala correlacionado com a largura da pétala possuem a melhor distribuição para classificação. Também observando as correlações, é possível identificar que a espécie setosa é a apresenta menor chance do classificador errar.