Kim Conger 8.30.22

Assignment #1

a) There are 150 total samples. This K-fold technique requires the splitting of the data in half, so 75 samples were in the model training set.

b) By the reasoning above, 75 samples were in the model test set.

c) The confusion matrix is given as

	setosa	versicolor	virginica
setosa	5	10	11
versicolor	8	7	11
virginica	11	6	6

setosa → ↑ versicolor ← virginica

$$\text{Accuracy} = \frac{\text{diagonal}}{\text{total sample}}$$

$$= \frac{5+7+6}{75}$$

[Accuracy = 0.24 (matches python value)]

d) Note that, as discussed in class, the in-class and sklearn precision & recall equations are "flipped." For instance,

$$P_{class} = \frac{\#\text{ correct}}{\text{sum of row}} \quad (1)$$

$$P_{python} = \frac{\#\text{ correct}}{\text{sum of column}} \quad (2)$$

I will use P_{python} for simpler comparison with notebook printed values.

$$P_{setosa} = \frac{5}{5+8+11} = 0.208, \text{ compared with } 0.21 \checkmark$$

$$P_{vers} = \frac{7}{10+7+6} = 0.304, \text{ compared with } 0.30 \checkmark$$

$$P_{virgin} = \frac{6}{6+11+11} = 0.214, \text{ compared with } 0.21 \checkmark$$

→

(2)

e) Similarly,

$$R_{class} = \frac{\text{\# correct}}{\text{sum of column}} \quad (3)$$

$$R_{python} = \frac{\text{\# correct}}{\text{sum of row}} \quad (4)$$

and I will use (4).

$R_{setosa} = \frac{5}{5+10+11} = 0.192$, compared with 0.19 ✓

$R_{versi} = \frac{7}{8+7+11} = 0.269$, compared with 0.27 ✓

$R_{virgin} = \frac{6}{6+6+11} = 0.261$, compared with 0.26 ✓

f) The general F1 formula is given by $F = 2\frac{(P \cdot R)}{(P+R)}$.

$F_{setosa} = 2\frac{(0.21)(0.19)}{(0.21+0.19)} = 0.199$, compared with 0.20 ✓

$F_{versi} = 2\frac{(0.30)(0.27)}{(0.30+0.27)} = 0.284$, compared with 0.29 ~✓

(Note: using 0.304 and 0.269 yields 0.285, with a rounded value being 0.29.) ✓

$F_{virgin} = 2\frac{(0.21)(0.26)}{(0.21+0.26)} = 0.232$, compared with 0.24 ~✓

(Note: similar to the above case, using the calculated values of 0.214 and 0.261 yields a # closer to 0.24.) ✓