Interpretation of boundaries:

- Linear boundary:

When using a linear kernel, we obtain a classifier that is a straight line in two dimensions. It is clear that this straight line is unable to separate the classes very well. It seems we need a non-linear decision boundary for this case, which can be achieved by using a different kernel.

- Polynomial boundary:

When using a polynomial kernel with the default degree (which is 3). This decision boundary is not a straight line in two dimensions, and is somewhat better able to split the two classes in our dataset than a linear decision boundary. However, it is also clear that we could achieve even better performance on our dataset by using a model that uses even more non-linear transformations of our features. One such a model would be a SVM using the radial basis function.

- Radial basis function:

When using the radial basis kernel, the feature space of our model has an infinite number of dimensions. This helps the model to separate the classes in our dataset a lot better than an SVM using the polynomial (degree 3) and linear kernels can. That is because the model using the radial basis function has access to more higher-order features that can be used to separate these non-linearly separable classes.

Choosing a very low value for C decreases the model's performance (especially when paired with a low value for γ). This indicates underfitting, since a low value for C means that the parameters of the model are heavily regularized. This causes only a small amount of the parameters to have a non-zero value, meaning the model complexity is too low to fit to the data.

Interpretation of heatmap and statement of results:

Choosing a low value for γ similarly diminishes model performance. We know that as γ gets smaller, the model will fit the training data less closely. Thus, the poor performance we get for low values of γ is due to underfitting: our model will be made too simple to capture all patterns present in the data.

Note, however, that this does not mean that increasing both C and γ to even larger values than 10 and 2 (respectively) will yield even better average results upon cross-validation. That is because, if large enough, C and γ will likely cause our model to overfit to the training data, meaning it will show poorer performance when predicting labels for samples it has not seen before than if we had used lower values for γ and C.

We get the best accuracy for $\gamma = 1$ and $C = 2$ (so relatively high values for γ and C).

The best model has an accuracy of 0.983 (rounded to 3 decimals).