Preparation

Download the synthetic classification dataset from
HW_classification_dataset.csv

Download the Wisconsin Breast Cancer ***diagnostic*** dataset from
https://archive.ics.uci.edu/ml/datasets.php
Make sure to get the diagnostic dataset (file named "wdbc.data"), not one of the others (prognostic, original) in the same directory, and the corresponding description "wdbc.names".

Create a new project in Orange.

Note: whenever you add an item to the report, in the text box label that item with the corresponding Part and Step number and a short title.

Part 1: reading and examining synthetic dataset

1. Read in the synthetic classification data file into Orange and make sure it is interpreted so that the row # (index) column is ignored (since you don't want it as a feature), and the class status column is set to be a target. You can use the "Data Table" widget to examine the data and confirm that you got the right number of lines, and a target is set. Check the "File" and "CSV File Import" widgets, which have slightly different options for handling headers and column identities. Be sure to click on the various columns to set (to the extent possible) how that column is handled. You may need to use the "Select Column" widget to turn the class status column from a feature into a target.
2. Attach a "Distributions" widget to the data and visualize the distributions of the two features (f0 and f1). Select the class status in the "Split By" item to see distributions separated by the two classes. Add the Distributions widget to the report while it is showing one of the features, and then again while it is showing the other. Expand the report window if needed to see the actual report (to the right of the list of widgets) and text box. In each text box enter a label (Part 1.2) and a short phrase, e.g. "feature 0 distribution", "feature 1 distribution", etc. Be sure to do this for each text box from here on. In the text box after the 2nd distribution say whether it looks like it will be straightforward to create a model that cleanly separates the two classes, and explain what aspects of the distributions lead you to this conclusion.

Part 2: classifying the synthetic dataset

3. Add a "Logistic Regression" widget and a "Test and Score" widget. Connect the regression widget to the testing widget as Learner -> Learner. Connect the data from the file to the testing widget as Data -> Data. Make sure there are no dashed connection lines (which indicates no data). Open testing widget and make sure that target class is set to "in", and it's using 10-fold cross validation with stratification. Make sure all the metrics we discussed (that are available) are included in the metrics chart (ctrl-click on the chart header). Add the test & score widget to the report, and in the text box write a brief explanation of what the value of each metric tells you about how good the model is.
4. Connect a "Confusion Matrix" widget to the test widget. Open the confusion matrix results window and add the confusion matrix to the report. In the text box write the numbers of true positives and negatives and false positives and negatives, and say whether this model is doing a good job at classification based on the confusion matrix results.
5. Connect a Scatterplot widget to the processed file data. Use the two features for the axes and color by class. Add the scatterplot widget to the report, and in the text box explain whether you think a linear model (like Orange's "Logistic Regression" widget) should be able to classify the data effectively, and why.
6. Save the project (the workflow itself) to a file with your name in it for submission to canvas.
7. Make sure the report has two distribution widget reports (labeled features 1 and 2), a test & score report, a confusion matrix report, and a scatterplot report, and that each text box says what part & step that item corresponds to. Save the report as html.
8. Upload the saved workflow and html report to the canvas assignment submission.

Part 3: wisconsin breast cancer database

1. Look at the raw data file, and see if it has a header. If it does not, read the description to figure out sensible column names, and create a header line. Note that the observed features (after the 2st two columns) are in 3 groups of 10, so first the 10 mean "radius, texture ….", then 10 standard error "radius, texture, …", then 10 worst "radius, texture, …".
2. Duplicate the workflow from the previous parts, most easily by copying the saved workflow ".ows" file (not the report) to a new name. Modify it to read from the breast cancer data file. You may need to rename the file so Orange can see it. You may also need to modify the file reading and processing depending on whether this file has the same format details, such as a column of sample indices and headers. Make sure no lines are lost and the benign/malignant column is used as the target.
3. Open the report window (View menu) and delete any existing items.
4. View the distributions of each feature and determine if there are some that look promising for classification with this model. Add the Distribution widget to the report , making sure that if there is a promising feature, it is selected. In the text box in the

report, explain why you think it will be promising, or if not, explain why not. Do all the distributions have the same typical magnitude? Say what this implies about preprocessing the data.

5. View the confusion matrix, and add it to the report. In the text box explain whether the model is doing a good job, and why, based on the confusion matrix.
6. Add a Preprocessing widget before the logistic regression as Preprocessor -> Preprocessor. Make sure the only thing it does is normalization (shift mean to 0 and std. dev. to 1)
7. Look at the confusion matrix results again and add the confusion matrix to the report again**.** Compare the results to the results without preprocessing (should be able to see both in the report). In the report text box, discuss the changes, and whether they are expected based on what you saw in earlier steps.
8. Save the workflow
9. Open the report window, confirm that there's only the Distributions and two copies of the Confusion matrix entries (all text boxes labeled with part.step and a short title), and save it to html.
10. Upload both workflow and html report to canvas.