Supplementary Material – Oliver Keers

Glossary

Accuracy	The proportion of samples that were classified correctly ((TN + TP)/(TN+TP+FN+FP))
Algorithm	A series of steps or calculations performed in order, to achieve a result
Bag	Bootstrap AGgregate – a random sample that has also been replaced
Bayes Theorem of conditional probability	$p(Y\|X) = \frac{p(X\|Y)p(Y)}{p(X)}$ [3] It is possible to determine the probability of something if you have sufficient information about related events.
Binary classification problem	A machine learning task where the aim is to assign each instance to one of two possible classes X → A or B
Correlation heatmap	A figure showing how one variable changes as another changes, using colour as a scale to show the strength of that relationship. Ranges from -1 to 1, with 0 meaning no relationship and 1 being a linear relationship
Cost	The penalty applied for classifying an instance incorrectly
Decision Tree	A collection of binary decisions (nodes) that are assembled sequentially, to arrive at one of a set of outcomes (leaves)
Gaussian	Normally distributed
Greedy Approach	Maximising gain against the objective at each step
Holdout / Test	Data that the model has not seen at any point during training. Used to assess whether the model works well when presented with new data.
Hyperparameter	Can be adjusted to alter the performance of an algorithm
Instance	A row in the data, in this case corresponding to an individual patient
K-Fold Cross-Validation	Dividing the training set into K smaller datasets. Each of these is predicted by the other K-1, used to assess performance in development.
Naïve	Unworldly – in this case making the assumption that variables are independent, which does not represent the real situation.
Normalized histogram	Graph showing the proportion of the values found within a collection of ranges. Useful for visualising the distribution of variables
Outlier	A data point that is far from the expected values.
Parallel coordinate plot	A graph showing the distribution of each variable, for target classes. Where there is a difference in that variable for the two classes, that may indicate a suitable feature for use in modelling
Posterior Probability	The probability that has been calculated taking account of the information held.
Prior Probability	The probability of something before calculations have been performed to update it.
QQ Plot	A Quantile/Quantile Plot. This illustrates whether data is normally distributed, with a normal distribution following a 45-degree line perfectly.
Sensitivity	The proportion of positive cases that were correctly picked up (TP/(TP+FN))
Specificity	The proportion of negative cases that were correctly classed as negative (TN/(TN+FP))
Standardization	Adjusting the values of variables to make them comparable in scale
Target variable	The attribute we are trying to predict
True/False Negative/Positive	Instances that were correctly/incorrectly classified as non-cancer/cancer respectively
Variable / Feature	A column in the data, an attribute that can be used as part of the model
Z-score	A measure of how far away results are from the mean

Implementation Details & Intermediate Results

Naïve Bayes

10-Fold Accuracy	Description, filename, comments
64.1%	• NBRaw.m • Initial run on all data, non-normalised
64.1%	• NBNormal.m • Normalization of data • **No benefit, not used in subsequent work**
63.2%	• NBNoOutliers.m • Non-normalized, with outliers (|z| > 4) compared to the mean of each classification removed • Reduction in accuracy, and as outliers represent real patients, removal would not be appropriate. • **Not used in subsequent work**
	• NBWeightGrid.m • Grid search to find the optimal weights to apply for outliers (identified by z score) • **Found to be most accurate when instances where z>5 for at least one variable have a weight of 0.125, and 0.3 for 5>z>4**
66.7%	• NBWeight.m • Model using the weightsings for outliers determined in the last step
65.4%	• NBWeightK.m • Applying a kernel to above • Resulted in one additional misclassification, but was preseved for hyperparameter optimisation purposes later on.
	• NBFeat.m • Sequential feature selection. • With no inputs, suggested [1 3 5 8] for both forward and backward selection • With glucose kept, and insulin and HOMA kept out due to colinearity, F: [1 3 8], B: 1 2 3 6 8]
76.9%	• NBWeighFeaK.m • Used to determine accuracy with weightings, using the feature selection • Used the following sets of variables [1 3 8], [1 2 3 8], [1 3 6 8], [1 2 3 6 8] • **[1 3 8] - Age, Glucose & Resistin found to be most accurate and used for subsequent work**
75.6%	• NBCost.m • Applying a costing to penalise for false negatives more. This is to reflect the importance of not missing diagnoses • **Aimed for a Sensitivity of at least 90%, found a cost of 0.7 for false positives to provide the greatest accuracy for this.**
76.9%	• NBOpt.m • Hyperparameter Optimization run • Optimizing kernel and width • Best estimated to be a normal distribtion with width of 6.7532
70.5%	• NBLog8Opt.m • Initial analysis (e.g. qq plots) suggested that log transforming some variables may make them more normally distributed. • Of the remaining variables, this only applied to variable 8 - resistin, so this was attempted • **Resulted in a decrease in accuracy and was not used for subsequent work**
76.9%	• NBBest.m • Model using the best settings found, to save and export for testing

Random Forest

10-Fold Accuracy	Description, filename, comments
70.5%	•RFRaw.m •Initial run on all data, non-normalised
71.7%	•RFExpts.m •Experimental file, rewritten often •MinLeafSize looped for fvalues between 1 and 39. •Optimum found to be 23
71.7%	•RFExpts.m •NumVariablesSampled, looped for 1:9 •optimum found to be 7
73.1%	•RFExpts.m •MinParentSize, looped for 1:40 •Optimum found to be 28
73.1%	•RFExpts.m •MaxNumSplits looped for 1:20 •Optimum found to be 7
74.4%	•RFExpts.m •Looping for both NumVariablesSampled, and MaxNumSplits. •Optimum found to be 5 and 7, respectively
76.9%	•RFRaw.m •Features [1 3 8] selected based on RFFeat.m, work on NB, and reference paper •**Significant improvement on all other work to date, kept going forwards**
78.2%	•RFLoop.m •Gridsearch using For loop to iterate over sample values for MaxNumSplits, MaxNumTrees, and MinLeafSize •**Optimal values found MaxNumSplits: 10, MaxNumTrees: 20, MinLeafSize: 2**
75.6%	•RFOpt.m •Hyperparameter optimisation run looking at MinLeafSize, MaxNumSplits, SplitCriterion & NumVariablesToSample •Optimum festimated: MinLeafSize 2, MaxNumSplits, 77 SplitCriterion deviance & NumVariablesToSample 1 •**Worse than previous model not used for test**

Differences in test classification

Misclassified by NB & RF: 6	Misclassified by NB only: 2
Misclassified by RF only: 3	Misclassified by neither : 27

Additional references:

MATLAB Academy material:

Introduction to Statistical Methods with MATLAB: https://matlabacademy.mathworks.com/R2020a/portal.html?course=stats , last accessed 30/11/20
Machine Learning with MATLAB: https://matlabacademy.mathworks.com/R2020a/portal.html?course=mlml last accessed 30/11/20

MATLAB documentations and example code were used throughout this work to learn the software. These pages were used to inform coding, with example code being adapted as appropriate:

Table: https://uk.mathworks.com/help/matlab/ref/table.html?searchHighlight=variable%20names%20table&s_tid=srchtitle
Histogram: https://uk.mathworks.com/help/matlab/ref/matlab.graphics.chart.primitive.histogram.html
Subplot: https://uk.mathworks.com/help/matlab/ref/subplot.html?searchHighlight=subplot&s_tid=srchtitle
Zscore: https://uk.mathworks.com/help/stats/zscore.html?searchHighlight=zscore&s_tid=srchtitle
Imagesc: https://uk.mathworks.com/help/matlab/ref/imagesc.html?searchHighlight=imagesc&s_tid=srchtitle
Probplot: https://uk.mathworks.com/help/stats/probplot.html?searchHighlight=probplot&s_tid=srchtitle
Corrcoef: https://uk.mathworks.com/help/matlab/ref/corrcoef.html?searchHighlight=corrcoef&s_tid=srchtitle
Normalize: https://uk.mathworks.com/help/matlab/ref/double.normalize.html?searchHighlight=normalize&s_tid=srchtitle
Parallelcoords: https://uk.mathworks.com/help/stats/parallelcoords.html?searchHighlight=parallelcoords&s_tid=srchtitle
Fscmrmr: https://uk.mathworks.com/help/stats/fscmrmr.html?searchHighlight=fscmrmr&s_tid=srchtitle
Cvpartition: https://uk.mathworks.com/help/stats/cvpartition.html?searchHighlight=cvpartition&s_tid=srchtitle
Fitcnb: https://uk.mathworks.com/help/stats/fitcnb.html
Perfcurve: https://uk.mathworks.com/help/stats/perfcurve.html?searchHighlight=perfcurve&s_tid=srchtitle
Crossval: https://uk.mathworks.com/help/stats/crossval.html?searchHighlight=crossval&s_tid=srchtitle
Confusionchart: https://uk.mathworks.com/help/stats/confusionchart.html?searchHighlight=confusionchart&s_tid=srchtitle
Sequentialfs: https://uk.mathworks.com/help/stats/sequentialfs.html?searchHighlight=sequentialfs&s_tid=srchtitle
Fitcensemble: https://uk.mathworks.com/help/stats/fitcensemble.html
templatetree https://uk.mathworks.com/help/stats/templatetree.html
Scatter3: https://uk.mathworks.com/help/matlab/ref/scatter3.html?searchHighlight=scatter3&s_tid=srchtitle
Predict: https://uk.mathworks.com/help/stats/compactclassificationnaivebayes.predict.html
Testcholdout: https://uk.mathworks.com/help/stats/testcholdout.html

Bibliography

Aiping Wang *et al.* (2009) 'An incremental extremely random forest classifier for online learning and tracking', in *2009 16th IEEE International Conference on Image Processing (ICIP)*, pp. 1449–1452. doi: 10.1109/ICIP.2009.5414559.

Austria, Y. D. *et al.* (2019) 'Comparison of Machine Learning Algorithms in Breast Cancer Prediction Using the Coimbra Dataset', *International journal of simulation: systems, science & technology*. doi: 10.5013/IJSSST.a.20.S2.23.

Bishop, C. M. (2006) *Pattern recognition and machine learning*. New York: Springer (Information science and statistics).

Breiman, L. (2001) 'Random Forests', *Machine Learning*, 45(1), pp. 5–32. doi: 10.1023/A:1010933404324. Ch1, Ch3

Crisóstomo, J. *et al.* (2016) 'Hyperresistinemia and metabolic dysregulation: a risky crosstalk in obese breast cancer', *Endocrine*, 53(2), pp. 433–442. doi: 10.1007/s12020-016-0893-x.

F. Y. A'la, A. E. Permanasari and N. A. Setiawan (2019) 'A Comparative Analysis of Tree-based Machine Learning Algorithms for Breast Cancer Detection', in *2019 12th International Conference on Information & Communication Technology and System (ICTS)*. *2019 12th International Conference on Information & Communication Technology and System (ICTS)*, pp. 55–59. doi: 10.1109/ICTS.2019.8850975.

'Full Text' (no date). Available at: https://www.nature.com/articles/s41598-018-25679-z.pdf (Accessed: 28 November 2020).

Dietterich, T. G. (1998) 'Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms', *Neural Computation*, 10(7), pp. 1895–1923. doi: 10.1162/089976698300017197.

M. U. Ghani, T. M. Alam and F. H. Jaskani (2019) 'Comparison of Classification Models for Early Prediction of Breast Cancer', in *2019 International Conference on Innovative Computing (ICIC)*. *2019 International Conference on Innovative Computing (ICIC)*, pp. 1–6. doi: 10.1109/ICIC48496.2019.8966691.

Murphy, K. P. (2012) *Machine learning: a probabilistic perspective*. Cambridge, MA: MIT Press (Adaptive computation and machine learning series).

Naveen, R. K. Sharma and A. Ramachandran Nair (2019) 'Efficient Breast Cancer Prediction Using Ensemble Machine Learning Models', in *2019 4th International Conference on Recent Trends on Electronics, Information, Communication & Technology (RTEICT)*. *2019 4th International Conference on Recent Trends on Electronics, Information, Communication & Technology (RTEICT)*, pp. 100–104. doi: 10.1109/RTEICT46194.2019.9016968.

Patrício, M. *et al.* (2018) 'Using Resistin, glucose, age and BMI to predict the presence of breast cancer', *BMC Cancer*, 18(1), p. 29. doi: 10.1186/s12885-017-3877-1.

Pham, H. and Pham, D. H. (2020) 'A novel generalized logistic dependent model to predict the presence of breast cancer based on biomarkers', *Concurrency and Computation: Practice and Experience*, 32(1), p. e5467. doi: 10.1002/cpe.5467.

Rahman, M. M. *et al.* (2020) 'Machine Learning Based Computer Aided Diagnosis of Breast Cancer Utilizing Anthropometric and Clinical Features', *IRBM*. doi: 10.1016/j.irbm.2020.05.005.

Singh, B. K. (2019) 'Determining relevant biomarkers for prediction of breast cancer using anthropometric and clinical features: A comparative investigation in machine learning paradigm', *Biocybernetics and Biomedical Engineering*, 39(2), pp. 393–409. doi: 10.1016/j.bbe.2019.03.001.

Taniguchi, H., Sato, H. and Shirakawa, T. (2018) 'A machine learning model with human cognitive biases capable of learning from small and biased datasets', *Scientific Reports*, 8(1), p. 7397. doi: 10.1038/s41598-018-25679-z.

Yue, J., Zhao, N. and Liu, L. (2020) 'Prediction and Monitoring Method for Breast Cancer: A Case Study for Data from the University Hospital Centre of Coimbra', *Cancer Management and Research*, Volume 12, pp. 1887–1893. doi: 10.2147/CMAR.S242027.