A Small Comparison: Naïve Bayes and Random Forest Applied to the Coimbra Breast Cancer Dataset

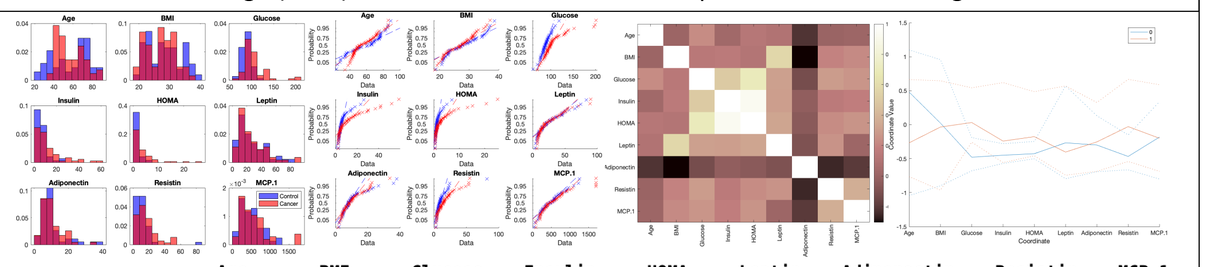
Oliver Keers

Description and Motivation of the Problem

- Breast cancer screening often utilizes mammography, a process that can be expensive, is stigmatized in some societies, and is less effective in younger women. Routine blood tests may help avoid these issues, if they can accurately predict those at risk of breast cancer.
- This is a binary classification problem, with individuals either being assigned to the cancer patient group, or the healthy control group based on 9 numeric variables.
- This is a very small dataset, with only 116 instances, and so approaches that work well with large datasets may be less effective here. Naïve Bayes and Random Forests have been selected to see whether simple or more complicated algorithms are better to use with small datasets. Results will be compared with the work of Singh (2019)[1] to assess the use of the same predictive models using this dataset.

Initial Analysis of the Dataset Including Basic Statistics

- Breast Cancer Coimbra Dataset from UCI ML Database from work by Crisóstomo *et al.*, (2016)[2]
- 116 individuals, with one target variable (cancer status) and nine features (all quantitative).
- A small class imbalance exists with 64 cancer patients and 52 from control group.
- All features bar Age and BMI are obtained from blood samples taken from fasting patients.
- Mean, standard deviation, and median were calculated for each variable, for each patient group.
- Comparative normalized histograms show the distribution of markers for both classes.
- QQ plots for both classes show that variables are not normally distributed.
- Correlation heatmap indicates a relationship between Glucose, Insulin & HOMA, and BMI & Leptin.
- A parallel coordinate plot (solid = median, dotted = quartiles) indicates the greatest differences between the two groups with respect to Age, and Resistin levels, so these are likely to be particularly useful for modelling. Glucose, Insulin and HOMA also display a notable difference between the two patient groups, so selecting from this features is likely to be useful.



	Age	BMI	Glucose	Insulin	HOMA	Leptin	Adiponectin	Resistin	MCP.1
mean(cancer)	58.077	28.317	88.231	6.9338	1.5524	26.638	10.328	11.615	499.73
mean(control)	56.672	26.985	105.56	12.513	3.6233	26.597	10.061	17.254	563.02
std(cancer)	18.958	5.4275	10.192	4.8598	1.2183	19.335	7.6313	11.447	292.24
std(control)	13.493	4.6204	26.557	12.318	4.5889	19.212	6.1895	12.637	384
median(cancer)	65	27.694	87	5.4835	1.1397	21.495	8.1278	8.9292	471.32
median(control)	53	27.408	98.5	7.58	2.0523	18.878	8.4464	14.372	465.37

Naïve Bayes

- Generative model using Bayes' theorem of conditional probability for independent events.
- For each class, the density distribution of the variables is calculated. This is combined with the probability that the class is observed to determine posterior class probability densities.
- New inputs are assigned to whichever class has the largest posterior probability for their features.[3]

Pros

- ✓ Easily interpretable
- ✓ Posteriors indicate classification confidence[3]
- ✓ Simple algorithm: fast and scales well
- ✓ Can perform well with small datasets[4]

Cons

- ✖ Assumes variables completely independent
- ✖ Sensitive to class imbalance in training data[4]
- ✖ Has performed poorly on this dataset with ~74% accuracy on holdout data[1, 5]

Random Forest

- Creates a group of decision trees on randomly selected & replaced subsets of data using a random selection of variables to limit over-correlation of predictors.
- An ensemble is then produced by aggregating the decision trees.[6]
- New inputs are classified by each tree and a majority vote determines classification.[7]

Pros

- ✓ Very good predictive accuracy[6]
- ✓ Random nature reduces chance of overfitting
- ✓ Robust to outliers and noise
- ✓ Can be parallelised[7]

Cons

- ✖ Blackbox method – difficult to interpret
- ✖ Slow to train
- ✖ Random nature can affect reproducibility
- ✖ Performance akin to NB on small datasets [1, 4]

Hypothesis Statement

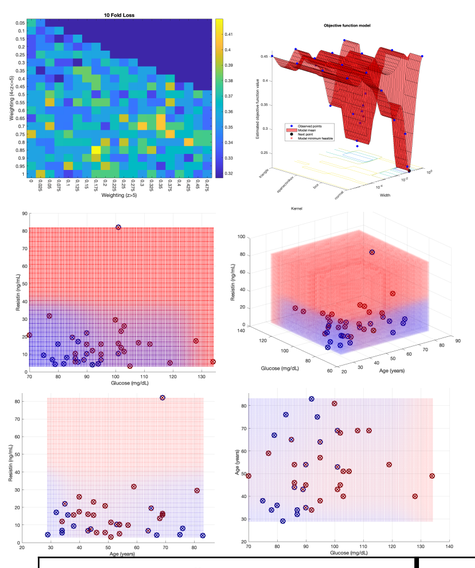
- The accuracy of the two algorithms will be equal when applied to holdout data, RF will be more computationally intensive to achieve these results.
- This is based on Singh (2019)[1] who achieved accuracy of 73.7% & 76.3% with NB & RF, respectively.
- Naïve Bayes was found to be a more specific classifier (NB: 94% , RF: 79%), while Random Forest was found to be more sensitive (NB: 57%, RF 75%)

Description of the choice of training and evaluation methodology

- 33% of the data removed and kept as test data to allow for comparability with reference paper.
- 10-Fold cross-validation performed during training to assess performance of both models
- For each model, hyperparameters were tuned to maximise accuracy of predictions from 10-fold cross-validation at each iteration of model building.
- Performance of both models on test data will be evaluated on accuracy, sensitivity, and specificity.

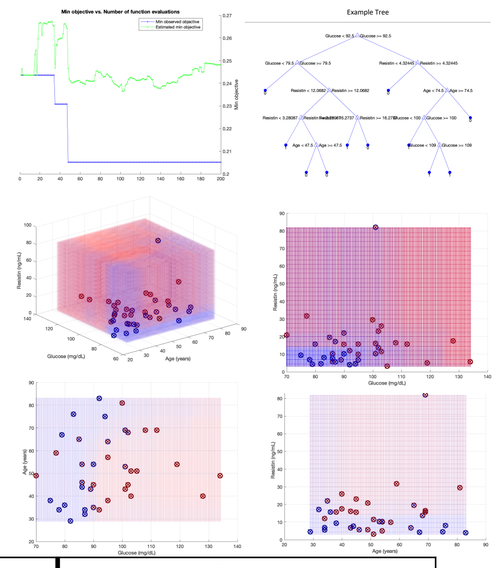
Choice of Parameters & Experimental Results: Naïve Bayes

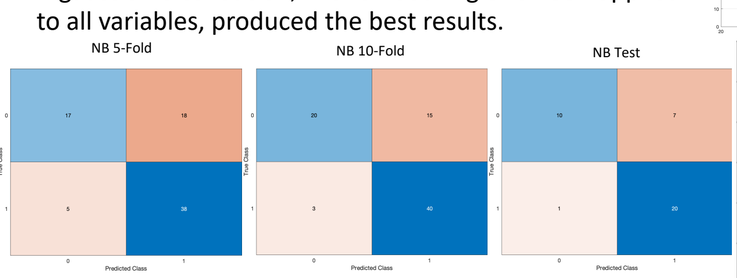
- Parameters chosen manually through a greedy approach, maximizing 10-fold accuracy at each step.
- Hyperparameters optimized using Bayesian Optimisation as the last stage of this process.
- Standardization of variables had no impact on accuracy.
- Outlier removal found to be less beneficial than reducing outlier weights. Optimal weights found by a grid search.
- Sequential feature selection followed by validation attempts indicated Age, Glucose & Resistin were the best predictors of cancer status.
- Sensitivity could be improved by 10% adding a reduced cost of 0.7, at the expense of specificity.
- Log transformation negatively impacted performance.
- A gaussian distribution, with smoothing of 6.7532 applied to all variables, produced the best results.



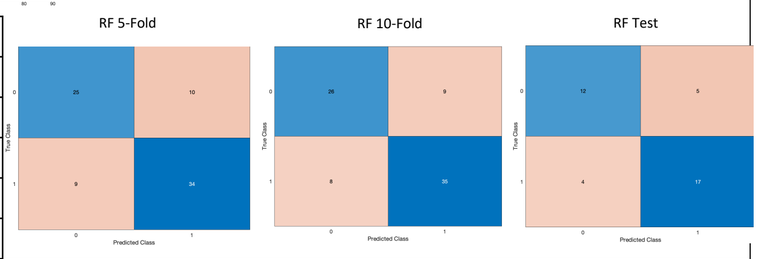
Choice of Parameters & Experimental Results: Random Forest

- Several parameters evaluated iteratively during initial exploratory work
- Feature selection seen to improve results, with Age, Glucose & Resistin chosen. These were found to provide more accurate predictions, and also allows for a fair comparison with Naïve Bayes and the reference paper.
- Manual grid search and in-built Bayesian optimization of hyperparameters were performed.
- The grid search led to the use of 20 trees for hyperparameter optimization.
- Bayesian hyperparameter optimization gave a less accurate model and so was not used as the final model.
- Final model trained had a bag of 20 trees, with a maximum of 10 splits and minimum leaf size of 2.





Naïve Bayes				Random Forest		
5-Fold	10-Fold	Test (33%)		5-Fold	10-Fold	Test (33%)
70.513%	76.923%	78.947%	Accuracy	75.641%	78.205%	76.316%
88.372%	93.023%	95.238%	Sensitivity	79.070%	81.395%	80.952%
48.571%	57.143%	58.824%	Specificity	71.429%	74.286%	70.588%
0.819	0.846	0.804	AUC	0.771	0.797	0.838



Analysis & Critical Evaluation of Results

- Both models appear to have struggled with classifying patients with Glucose levels between 85 and 95 mg/dL and Resistin levels under 25 ng/mL. The main differences in the classification space between the two models are a more negative region for RF at higher Resistin levels, and a more negative region for NB when predicting for younger individuals.
- Naïve Bayes performed very similarly on both cross-validation and the holdout data, unexpectedly performing slightly better on the unseen test data than it did during cross-validation.
- Accuracy of NB was higher than Singh (2019)[1] for 5-fold (71% vs 65%), 10-fold (77% vs 63%) and holdout data (79% vs 74%), and AUC was greater for all division protocols. This model was highly sensitive, while his was highly specific. Given the aim is to identify potential breast cancer cases, a sensitive model is more useful than a specific one due to the consequences of missing cases. The model trained here will provide false positives in c. 41% of healthy cases, which could cause distress if used as a diagnostic test but would make it effective as a tool used to screen patients suitable for mammography.
- Random Forest is usually considered to be a highly accurate algorithm, but its performance on this dataset has not been particularly strong. While more accurate during 5-fold, its accuracy during 10-fold validation is similar to NB, and slightly worse on the holdout data (imbalanced misclassification cost was not imposed for RF which may account for the differences in sensitivity & specificity compared to NB).
- Accuracy of Random Forest was found to be within 0.25% of that found by Singh (2019)[1] during 10-Fold cross-validation, with sensitivity and specificity also being very similar(Δ=1.4%, Δ=1.2% respectively). For 5-Fold, this discrepancy was approximately 2.5% across these metrics, with this model having the inferior performance. When applied to unseen test data, this model had the same accuracy of 76% as Singh (2019)[1] , but as with NB the model here is more sensitive (81% vs 75%) and less specific (71% vs 79%). AUC was similar to Singh (2019)[1] for both validations, but much greater than his on holdout data.
- The moderate performance of RF may be due to the small dataset, and selection of 3 variables to use during training. Feature selection improved accuracy over all 9 features (76.92% vs 67.95%) or a random subset of any size (<=71.79%) during training. Random Forest uses a random selection of variables and instances to train each tree[6], and so the paucity of each may have negatively impacted performance.
- Random Forest usually generalizes best with a large forest of shallow trees[7], but the model tuned in this instance was only 20 trees, which had relatively many splits and small leaves for such a small dataset. The likely impact of this would have been overfitting the training data, which would likely have caused a drop in accuracy from validation to holdout performance, however this was not observed.
- Training Naïve Bayes required much more work on my part, with far more thought needing to be put into deciding things like features to use and misclassification costs. Random Forest was much easier for the user to optimize, as it was possible to set up loops, although hyperparameter optimization was found to be deleterious to performance. NB took approximately double the time for me to optimize than RF.
- The corollary was a much greater computational demand for Random Forest. For hyperparameter optimization, RF took 3x longer than NB per evaluation. RF also took 4x longer to predict the test values.
- There was no significant difference in the predictive performance of the two algorithms on the test data (p=.6875), despite the significant additional computation required for Random Forests. The reduced computational expense, combined with the additional sensitivity, would therefore make the Naïve Bayes model a better choice on this dataset for intended use as a screening tool.

Lessons Learned & Future Work:

- More sophisticated machine learning approaches do not always provide better results, sometimes a simpler approach is best. Algorithms that work well on large datasets may not be best for small ones.
- Automatic hyperparameter optimization may not be as effective as manual tuning by a human, or a well-constructed grid search.
- Taking decisions iteratively in a greedy manner is a very effective approach, but also labour intensive.
- Future work on Random Forests – look at at different ensemble methods using decision trees. Both gradient boosting and using bags of decision trees have been found to provide better results. [8, 9]
- Future work on Naïve Bayes – explore implementation of Bayesian Networks for this dataset, as the naïve assumption of independence between features does not hold for this dataset. As the dataset is small and slightly imbalanced, applying Loosely Symmetric Naïve Bayes[4], would also be interesting.
- This data contained more cancer than control patients, which does not reflect the prevalence of breast cancer in the general population. Naive Bayes is particularly sensitive to the distribution of classes in the training data. Future work could see whether Naïve Bayes would still outperform Random Forest if applied to a dataset that more accurately reflected the rate of breast cancer in the population.

References

[1] 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

[2] Crisóstomo, J. *et al.* (2016) 'Hyperresistinemia and metabolic dysregulation: a risky crosstalk in obese breast cancer', *Endocrine*, 53(2), pp. 433–442

[3] Bishop, C. M. (2006) *Pattern recognition and machine learning*. New York: Springer (Information science and statistics). Ch 1

[4] 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.

[5] 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.

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

[7] Breiman, L. (2001) 'Random Forests', *Machine Learning*, 45(1), pp. 5–32.

[8] 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.

[9] 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.