A Comparative Analysis of Naïve Bayes and Random Forest on the Wisconsin Breast Cancer Dataset

Ashir Valjee and Axil Sudra

Data Science MSc | School of Mathematics, Computer Science and Engineering | City, University of London

Description and motivation of the problem

- Application and comparison of Naïve Bayes and Random Forest algorithms to a medical binary classification problem; assess whether a lump in a patient's breast is benign (non-cancerous) or malignant (cancerous) based on cell characteristics data.
- Contrast preliminary computational results to those obtained by a previous study conducted by Elgedawy, M.N. [1].
- Attempt to improve the performance of both machine learning algorithms.

Initial analysis of the dataset including basic statistics

Analysis and preparation of the dataset

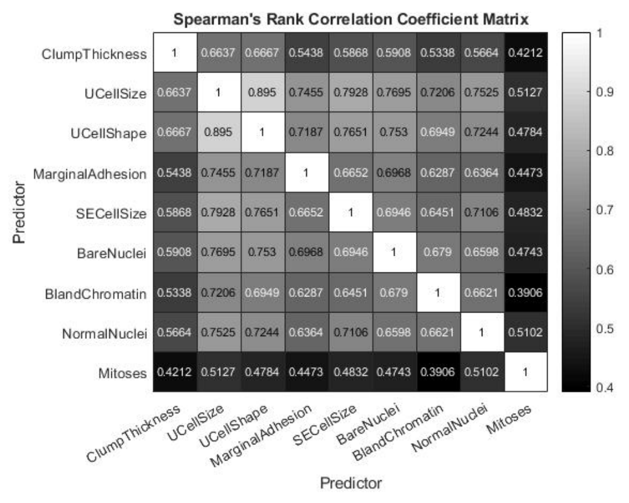
- Dataset: Wisconsin Breast Cancer (Original) from UCI Machine Learning Repository [2].
- The original dataset contained 699 instances, 10 integer-valued predictors and a class label indicating whether a (tumour) lump in a patient's breast is benign (denoted by 2) or malignant (denoted by 4).
- There were 16 instances with missing values which were removed from the dataset. Furthermore, a redundant predictor (*sample identification number*) was removed from the dataset.
- After preprocessing, the dataset had 683 instances and 9 predictors. Each predictor has been converted into numerical values from 1 (normal state) to 10 (abnormal state).
- The frequency table outlines the class imbalance in the dataset; 65.0% of instances were classified as benign and the remaining 35.0% as malignant.

Descriptive statistics of dataset

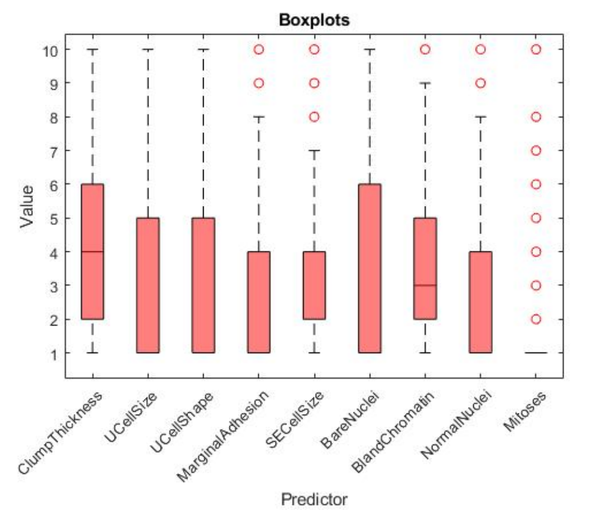
	Benign	Malignant	Total
Frequency	444	239	683
%	65.0%	35.0%	100.0%

- Spearman's Rank Correlation Coefficient (SRCC) was used to assess the relationship between the 9 predictors in the dataset; the results have been depicted on a SMRCC (heatmap) matrix. *Uniformity of Cell Size* and *Uniformity of Cell Shape* had the highest positive SRCC.
- A series of boxplots revealed the right-skewed distribution of the predictors; 60% of the values were either 1 or 2. Note that outliers were marked by red circles of which *Mitoses* had the most.

Statistical visualizations



Predictor	ClumpThickness	UCellSize	UCellShape	MarginalAdhesion	SECellSize	BareNuclei	BlandChromatin	NormalNuclei	Mitoses
ClumpThickness	1	0.6637	0.6667	0.5438	0.5868	0.5908	0.5338	0.5664	0.4212
UCellSize	0.6637	1	0.895	0.7455	0.7928	0.7695	0.7206	0.7525	0.5127
UCellShape	0.6667	0.895	1	0.7187	0.7651	0.753	0.6949	0.7244	0.4784
MarginalAdhesion	0.5438	0.7455	0.7187	1	0.6652	0.6968	0.6287	0.6364	0.4473
SECellSize	0.5868	0.7928	0.7651	0.6652	1	0.6946	0.6451	0.7106	0.4832
BareNuclei	0.5908	0.7695	0.753	0.6968	0.6946	1	0.679	0.6598	0.4743
BlandChromatin	0.5338	0.7206	0.6949	0.6287	0.6451	0.679	1	0.6621	0.3906
NormalNuclei	0.5664	0.7525	0.7244	0.6364	0.7106	0.6598	0.6621	1	0.5102
Mitoses	0.4212	0.5127	0.4784	0.4473	0.4832	0.4743	0.3906	0.5102	1



Summary and general comparison of machine learning models

Naïve Bayes (NB)

- A probabilistic classification model based on Bayes' theorem with the assumption that each predictor is conditionally independent of every other predictor given the value of a class. Hence, the 'naivety' of the model.
- Naïve Bayes Decision Rule: given a vector $X = (x_1,\ x_2, \ldots,\ x_n)$ of n predictors and a vector $C = (c_1,\ c_2,\ \ldots,\ c_k)$ of k classes, Naïve Bayes classifier is defined as:

$$C_{NB_{MAP}} = \underset{K=\{1,\ ...,\ k\}}{\operatorname{argmax}}\ p(c_K) \prod_{i=1}^{n} p(x_i|c_K),$$

  where $C_{NB_{MAP}}$ is the 'Maximum a Posteriori' posterior probability, $p(c_K)$ is the class prior probability, and $\prod_{i=1}^{n} p(x_i|c_K)$ is the likelihood [3].
- The class prior probabilities can be assumed from the data and the likelihood is approximated from the data. It is also possible to use real-world knowledge when choosing the class prior probabilities.

Advantages

- Simple to interpret and computationally efficient (i.e. probabilities required for the model can be assumed from the available data).
- Despite the conditional independence assumption, performance of the model has been reported by many researchers as surprisingly accurate [4].
- Robust to missing values as these values are simply ignored when computing probabilities.

Disadvantages

- Generally outperformed by other more sophisticated machine learning models.
- The key assumption of conditional independence is extremely unrealistic; its almost impossible that a set of predictors are independent in real industry datasets.
- Prone to 'zero probability'; when a particular predictor value in the test set never occurs together with a class in the training set, thus resulting in a zero posterior probability. Laplace Smoothing or additive smoothing can get around this problem [4].

Random Forest (RF)

- A model consisting of an ensemble of decision trees based on bootstrap aggregation (known as bagging); devised by Leo Breiman in 2001 [5].
- Each decision tree in the model is generated from a bootstrapped training dataset using random predictor selection; this is a method that selects a random subset of predictors at each node in the tree to reduce correlation between trees in the overall forest.
- Information gain can help improve the node splitting choice towards smaller trees.
- For classification tasks the model uses a majority vote system for class prediction - an instance is sent down every tree in the model and the most common class across all trees is assigned to that instance. For regression tasks the model calculates the average of the classes predicted by all the trees in the model.

Advantages

- The model can be used for classification and regression tasks.
- In comparison to other general-purpose machine learning models, it is one of the best performers with highly accurate predictions (due to it's inclusion of randomness) and resistant to overfitting when handling a large number of features (predictors) [7].
- Performs surprisingly well with little to no tuning of it's hyper-parameters.
- Can deal with missing values.
- Reduces the variance from individually noisy decision trees by increasing bias.

Disadvantages

- Time consuming in generating predictions as all of the decision trees contained in the model have to function simultaneously. For this reason the model is considered as computationally expensive.
- Considered as a 'black-box' model, thus making it difficult to interpret.

Hypothesis statement

- Elgedawy [1] reports that the Random Forest model slightly outperformed the Naïve Bayes model; they achieved an accuracy of 99.42% and 98.24% respectively. Alternative machine learning comparison papers, such as the one written by Caruana and Niculescu-Mizil [8], also report similar findings.
- Based on these series of analyses, we expect the Random Forest model to marginally exceed the performance of the Naïve Bayes model.
- Through the diligent process of experimentation and tuning model parameters, we anticipate an improvement in the performance of both models.

Description of choice of training and evaluation methodology

- The dataset has 683 points. We are holding out 25% of the data exclusively for testing and the rest will be used for training.
- For Naïve Bayes we will use 10 fold cross validation on the training set. In contrast, we don't use any form of cross validation for Random Forest; this is due to the model using random predictor selection during each tree split [5].
- We will run a grid search to optimize the hyper parameters in our Random Forest model.
- We will use the same performance metrics as Elgedawy [1].
- Since we are trying to classify whether or not a patient has breast cancer, false negatives are very problematic. We will therefore compare the confusion matrix for each model.

Choice of parameters and experimental results

Naïve Bayes (NB)

Parameters

We ran several experiments, first trying to reproduce the results by Elgedawy [1] and then trying to expand on his work. In all of our experiments we implemented 10 fold cross validation:

- Experiment0: Elgedawy [1] uses a 75% \ 25% training test split with no cross validation. He doesn't state it in the paper but we assume he uses a kernel distribution.
- Experiment1: We eliminate the *Uniformity of Cell Shape* predictor based on its high correlation with *Uniformity of Cell Size*. We then compare the results of a normal distribution and kernel distribution for the predictors.
- Experiment2: We then look at changing the Prior distribution to 80% benign and 20% malignant based on information from Stony Brook Care Centre [9]. The Prior was previously calculated from the sample data; it was 65% benign and 35% malignant. We keep a Kernel distribution and the same predictors as experiment one.

Experimental results

- Experiment 0 yielded an accuracy of 96.9% in the training set and 97.1% in the test set. This was somewhat less then the 98.24% that Elgedawy [1] achieved.
- We got a 96.5% test set accuracy with a normal distribution, compared to 96.2% accuracy when using a kernel distribution in experiment 1.
- Using a Prior based on real world knowledge (experiment 2) did not perform as well as calculating the Prior from the training data (95.7% test set accuracy vs 96.2%).
- The run time for our models' was extremely quick.

Performance measures

Training set

Naïve Bayes	Measure	Random Forest
96.10%	Accuracy	99.81%
0.9154	Precision	1
0.9794	Recall	0.9944
0.9463	F-measure	0.9972

Test set

Naïve Bayes	Measure	Random Forest
96.47%	Accuracy	98.82%
0.9344	Precision	0.9831
0.9661	Recall	0.9831
0.9500	F-measure	0.9831

Note: these results are based on experiment 1 for both models (NB uses a normal distribution).

Random Forest (RF)

Parameters

We first tried to reproduce the results by Elgedawy [1] and then tried to expand on his work.

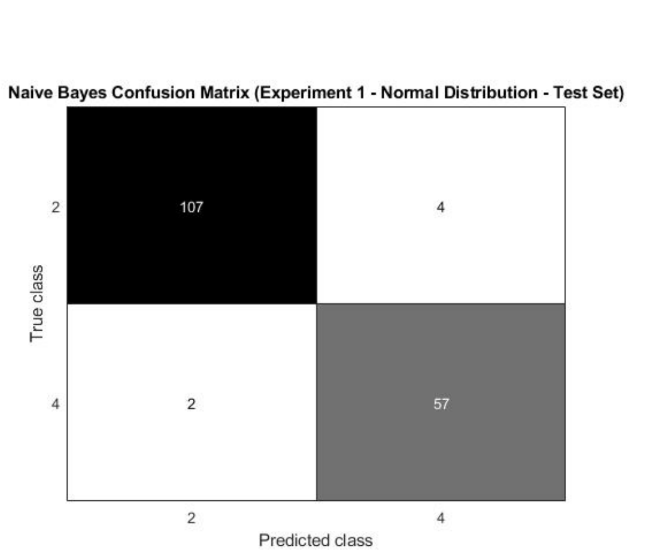
- Experiment0: Elgedawy [1] used 2000 decision trees in his Random Forest model. He doesn't state how many predictors were sampled at random at each node; we assume that it is the square root of the number of predictors, in this case 3.
- Experiment1: Similarly to experiment 1 for Naïve Bayes, We eliminate the *Uniformity of Cell Shape* predictor based on its high correlation with *Uniformity of Cell Size*. We then optimize some hyperparameters via a grid search; varying the number of decision trees and the number of predictors to sample at each node.

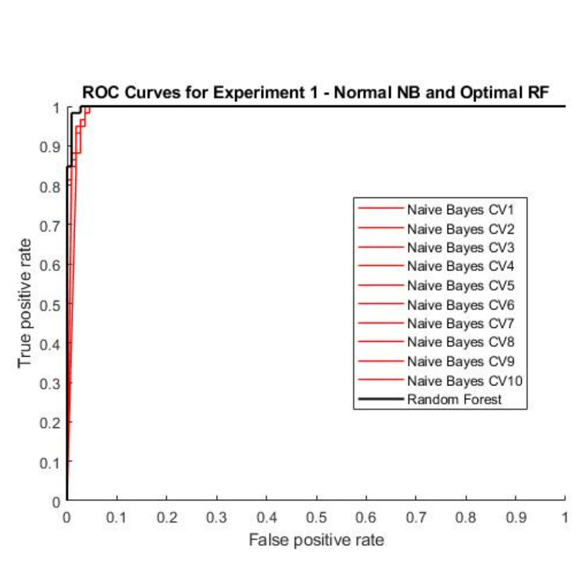
Experimental results

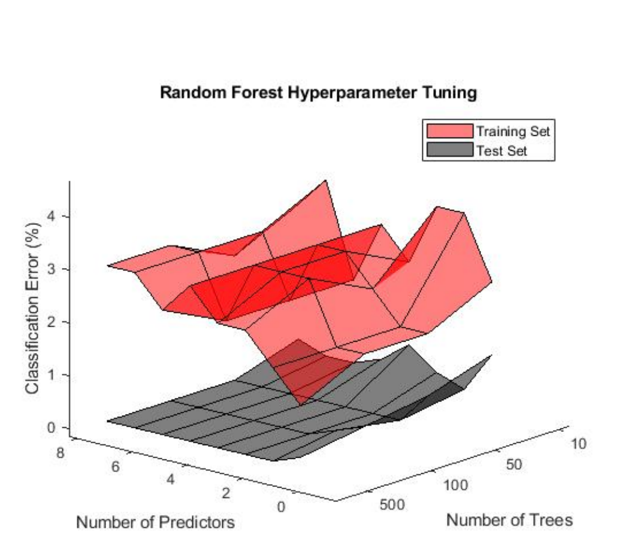
- Experiment 0 yielded an accuracy of 100% in the training set and 97.65% in the test set. This was less then the 99.42% that Elgedawy [1] achieved.
- After eliminating the *Uniformity of Cell Shape* predictor and conducting hyperparameter optimization, we decided to use 500 decision trees and sample one predictor at each node. We achieved a test set accuracy of 98.82% and F-measure of 0.9831 on the test set.
- The most important features in our Random Forest model were *Bare Nuclei*, *Uniformity of Cell Shape* and *Clump Thickness*.
- The run time was considerably longer then Naïve Bayes.

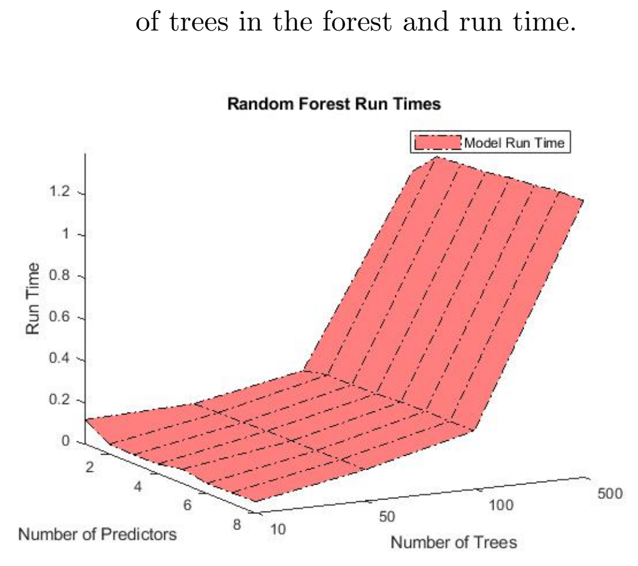
Analysis and critical evaluation of results

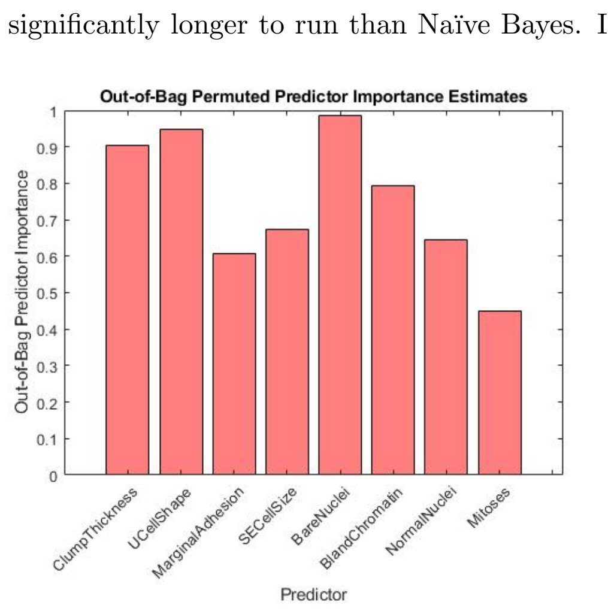
- As expected in our hypothesis, the Random Forest model slightly outperformed our Naïve Bayes model. In this sense our results matched those achieved by Elgedawy [1]. However, we were not able to achieve performance levels quite at the same level. One possible explanation is that Elgedawys [1] implementation was done in R, whereas we used MATLAB and the underlying code libraries are different. Another possibility is simply randomness; we noticed some variability in results when we changed our random seed.
- The results we achieved with Naïve Bayes showed reasonably similar training and test set accuracy. In contrast, our Random Forest models all showed significantly lower accuracy in the test set, suggesting over fitting of the data.
- The fact that using a kernel density distribution for Naïve Bayes did not give better results then a normal distribution was surprising. The box plots analysed during the investigation of the dataset were clearly not normally distributed and were heavily skewed towards lower values.
- Changing the Prior in Naïve Bayes based on real world knowledge slightly reduced accuracy. This is possibly due to the Wisconsin Breast Cancer dataset not necessarily being reflective of the wider population.
- Our cross validation results for Naïve Bayes were similar across validation sets, indicating model stability.
- Breiman [5] states that as the number of trees increases the generalization error for the Random Forest should converge. He also mentions that the error should decrease as the number of predictors sampled at random at each node decreases. During hyperparameter optimization our results reflected these claims.
- As shown in the confusion matrices, there were only two False positive negative predictions in Random Forest, and six in our Naïve Bayes model. These are reflected in the respective recall and precision scores. In a critical field like Breast cancer prediction, this is an important metric.
- The ROC curve plots the false positive rate against the true positive rate. As depicted by the graph, the AUC for both models is very close to 1. Our Naïve Bayes models had an AUC close to 0.9927 vs an AUC of 0.9983 for Random Forest. The AUC is a measure of how well a classifier distinguishes between groups, with 1 being a perfect score. In our case both of our models perform well with Random Forest slightly outperforming. Additionally, our Naïve Bayes cross validation models all had very consistent ROC curves, once again illustrating the stability of our model.
- As expected the Random Forest models took significantly longer to run than Naïve Bayes. Indeed we saw a linear relationship between the number of trees in the forest and run time.

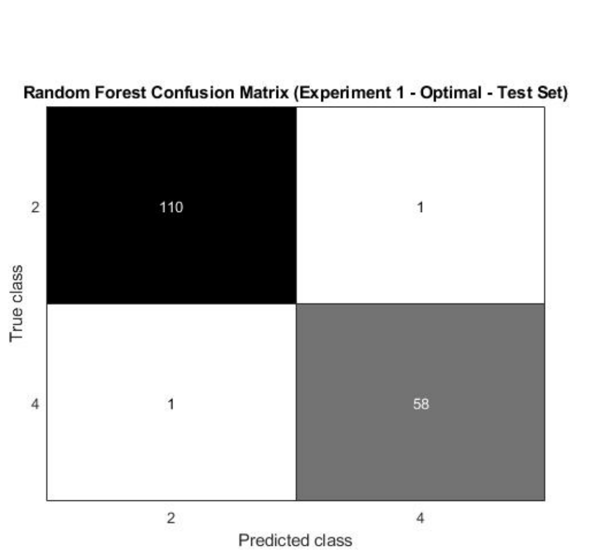












Lessons learned and future work

- Both models performed reasonably well without any tuning. Despite not really having conditionally independent predictors, the Naïve Bayes had good performance.
- In our random forest model consider adding tree depth to our grid search. Furthermore, we could perform a random search before running our grid search.
- Investigate ways to reduce over fitting in the Random Forest; for example, perform cross validation on existing models.
- Future work on Naïve Bayes could try predictor transformation; a log normal transformation could work well given how skewed out predictors are.

References

[1] Elgedawy, M.N. (2015) 'Prediction of Breast Cancer using Random Forest, Support Vector Machines and Naïve Bayes', *International Journal of Engineering and Computer Science*, volume 6 (issue 1), pp. 19884-19889.

[2] Wolberg, Dr. W.H. University of Wisconsin Hospitals (1992), 'Breast Cancer Wisconsin (Original) Data Set', Available at: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original%29, (Accessed 2018).

[3] Rogers, S. and Girolami, M. (2012), 'The Bayes classifier', in *A First Course in Machine Learning.* United States of America: Taylor & Francis Group LLC, pp. 170-173.

[4] Al-Aidaroos K. M., Bakar A. A. and Othman Z. (2010) 'Naïve Bayes Variants in Classification', *International Conference on Information Retrieval Knowledge Management (CAMP)*, pp. 276-281.

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

[6] Hastie T., Tibshirani R. and Friedman J. (2008) 'Definition of Random Forests', in *The Elements of Statistical Learning: Data Mining, Inference, and Prediction (Second Edition).* United States of America: Springer, pp. 587-588.

[7] Biau G. (2012) 'Analysis of a Random Forests Model', *Journal of Machine Learning Research*, volume 13 (issue 1), pp. 1063-1095.

[8] Caruana R. and Niculescu-Mizil A. (2006) 'An Empirical Comparison of Supervised Learning Algorithms', *ICML 2006 Proceedings of the 23rd international conference on Machine learning*, pp. 161-168.

[9] Carol M. Baldwin Breast Care Center (no date) *Different Kinds of Breast Lumps.* Available at: https://cancer.stonybrookmedicine.edu/breast-cancer-team/patients/bse/breastlumps (Accessed: 23 Novemeber 2018).