A Comparison of Naïve Bayes and Random Forest Machine Learning Models, Applied to the Breast Cancer Wisconsin (Diagnostic) Dataset

Elisa Chimenton and Qiqi Su

Description and Motivation of the Problem

- Predicting the outcome of a binary classification task: whether the breast cell nucleus is benign or malignant.
- Compare and contrast the performance of two machine learning models for this task: Naïve Bayes and Random Forest.
- We will compare our results to Street etc al., [1] and Shahnaz et al.,[2]

Initial Analysis of the Datasets including Basic Stats with Data Visualisations

Analysis and Preparation of the Dataset

- Dataset: Breast Cancer Wisconsin (Diagnostic) Dataset from UCI.
- The dataset has 569 instances and 32 attributes with no missing values, where 30 out 32 are real-valued input attributes.
- Class distribution: 357 benign (62.74%) and 212 malignant (37.26%).
- The 30 features of the original dataset represent the mean, standard error and worst of 10 different measurements of the cell nucleus.
- A selection of the normplot of each features (fig 1) show that area standard error (AS) is one of the least normalised features in the data, whereas concave point worst (CpW) is one of the most normalised features.
- The Covariance heatmap (fig 2) clearly indicates the correlation between each feature. Such that radius mean (RM) and fractal dimension mean (FDM) has one of the least correlations, whereas area mean (AM) and perimeter worst (PW) has one of the most correlations.
- A simple cross validation method and simple Naïve Bayes classier model were conducted on the entire dataset and 3 of the subsets separately before the grid search, showed that a simple Naïve Bayes on the entire dataset gives the best accuracy result.

Initial Data Visualisation

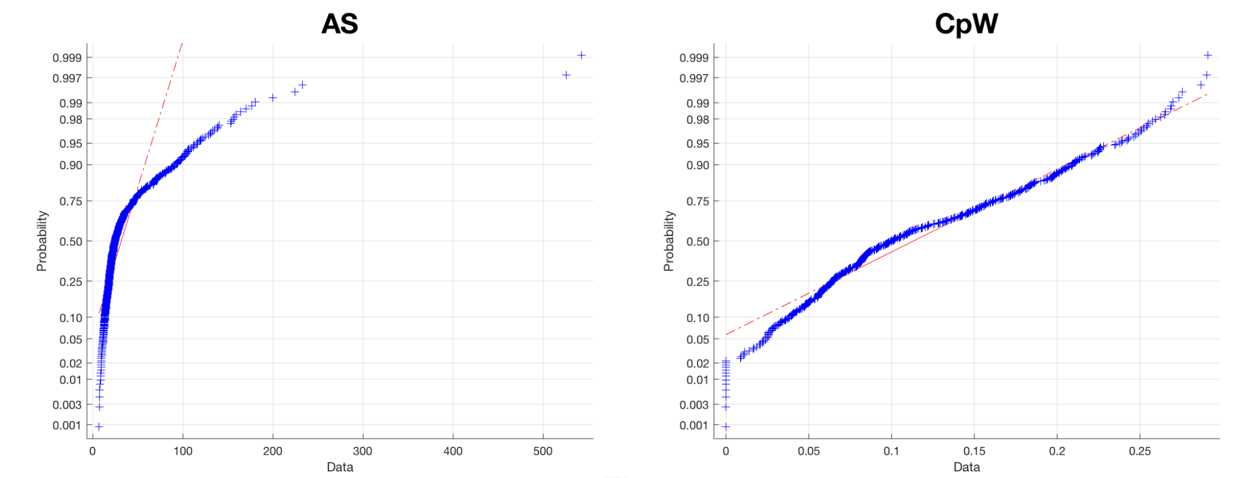


Fig 1

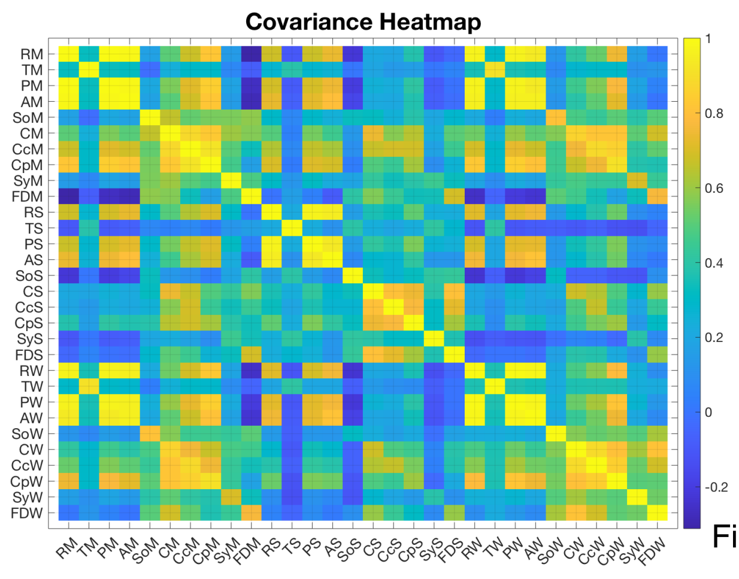


Fig 2

Pros and Cons of the Two Models

Naïve Bayes (NB)

Naïve Bayes classifiers are a family of algorithms using Bayes Theorem to obtain a set of probabilities of which class each predictor belongs to by counting the frequency and combinations of values in a given dataset. The model takes the assumption that each pair of features are independent of each other and make an equal contribution to the outcome; therefore, the model reduces the dimensionality of the dataset from a high-dimensional prediction task to a one-dimensional kernel density problem.

Pros:

- Gives a relatively accurate result despite the simplicity of the model.
- Fast and easy to implement.
- Can handle both continuous and categorical variables.
- Reduces the dimensionality of the data.

Cons:

- Less accurate than a slower method such as RF.
- Assumes the conditional probability of each independent variables are statically independent.
- In the case of binary classification, NB is substandard for non-linear separable concepts as it can only learn linear discriminant functions [3].

Random forest (RF)

Random forests split the data into training and testing sets. Bagging is used to create new training set, with replacement from the original dataset. The reasons to use bagging are to enhance accuracy and estimates generalisation error of the combined ensemble of trees [4]. Random Forests use bootstrapped data, thus not every sample is used to build every tree. The training dataset is the bootstrapped data and testing is the remaining sample. These remaining samples in Random Forest are called Out-of-Bag (OOB) data.

Pros:

- Overcomes problems of overfitting and bias.
- Reliable performance and high accuracy compared with best supervised learning models.
- It is relatively robust to outliers and noise.
- Explainable, simple to understand, it is not considered 'black-box' model.

Cons:

- Ensemble models is less interpretable than an individual decision tree.
- High computational cost and uses a lot of memory.
- Predictions are slower than simple models such as NB.

Hypothesis Statement

- Both models should give relatively accurate results, and RF should be more accurate than NB [2].
- Street et al., [1] stated that the best accuracy obtained using one separating plane, estimated accuracy 97.5% using repeated 10-fold cross validation.
- Shahnaz et al.,[2] has also found that Random Forest is the best classifier for 50% train and 50% test data with 96.83% accuracy.

Description of choice of training and evaluation methodology

- Randomly split the datapoint to 70% Training(1) and 30% Testing sets, and then further split the Training(1) set further to 70% Training(2) and 30% Validation sets for NB only.
- Compare the results with a 50% Training and 50% Testing sets.
- Conduct a 10-fold cross validation method on the Training(2) and Validation sets for NB, varying the hyper-parameters in the grid search for both models to find the best model as well as avoid overfitting.
- Also conduct an Bayesian Optimisation method for NB and compare the results with 10-fold cross validation method.
- Use the best model obtained from 10-fold cross validation method for NB and use Out-of-Bag (OOB) errors to estimate generalisation performance for RF, to train and test on the Training(1) and Testing sets, minimising error in the prediction accuracy.
- Compare the final result using Area under Curve (AUC), ROC chart and Positive and Negative predictive value.

Choice of parameters and experimental results

Naive Bayes

Parameters:

- Different distribution: normal, kernel and multivariate multinomial distribution (mvmn).
- Different width for kernel distribution.
- Different values for prior probabilities.

Main experimental results:

- Both Bayesian Optimisation method and 10-fold cross validation method show that normal distribution gives the best accuracy result, in comparison to 10 different width of the kernel distribution and mvmn distribution. However, 10-fold cross validation method still gives the best prediction accuracy overall.
- Baysain Optimisation method also reached 30 objective evaluations, with total elapsed time of 83.1631 seconds.
- The sample prior, which is calculated from the dataset, showed a clear superior performance compare to different values of prior probabilities in the range of -0.05:0.05.

Random Forest:

Parameters:

- Number of trees in the ensemble.
- Minimum number of observations per tree leaf.
- Number of predictors to sample per node.

Main experimental results:

- Accuracy improves when increase number of trees.
- Increasing the number of trees to 50 improves the accuracy to approximately 96.98%, and running time decreases significantly.
- Increasing the number of tress to 100 gives a better result.
- The most optimal number trees are 80, number of leaves are 6 and number of predicators are 8.

Data Visualisation on the Final Result for NB

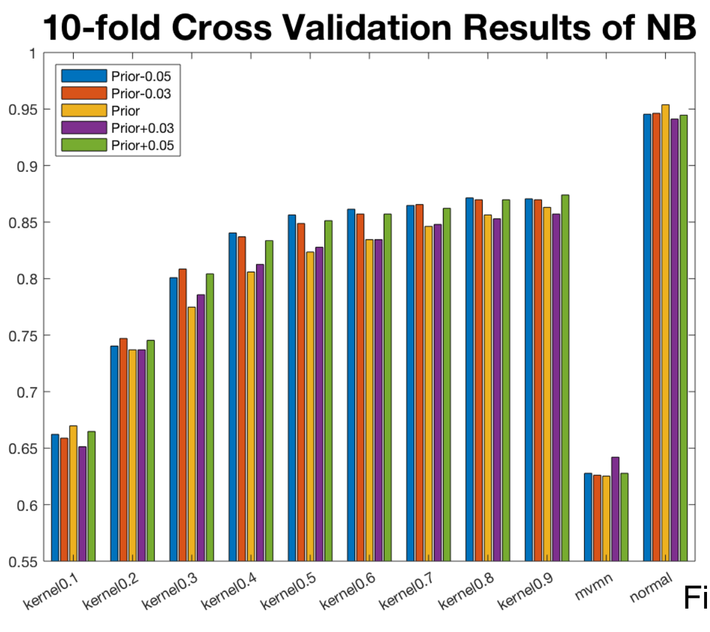
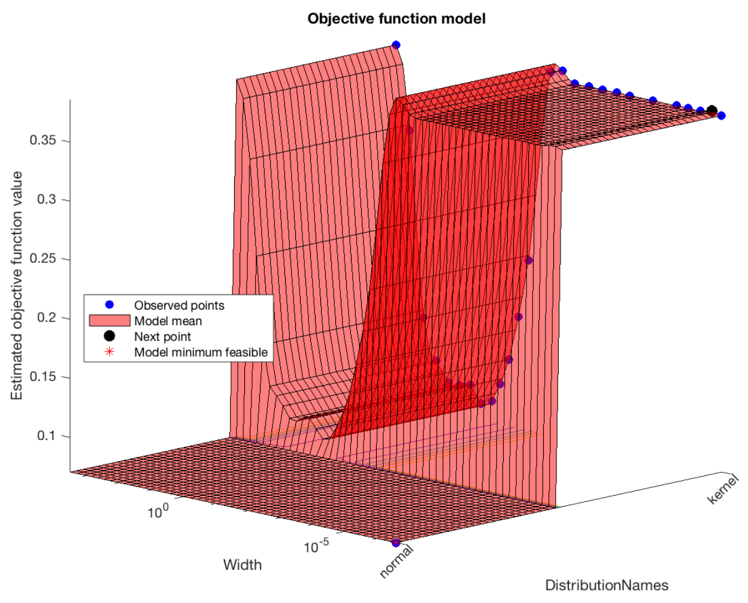
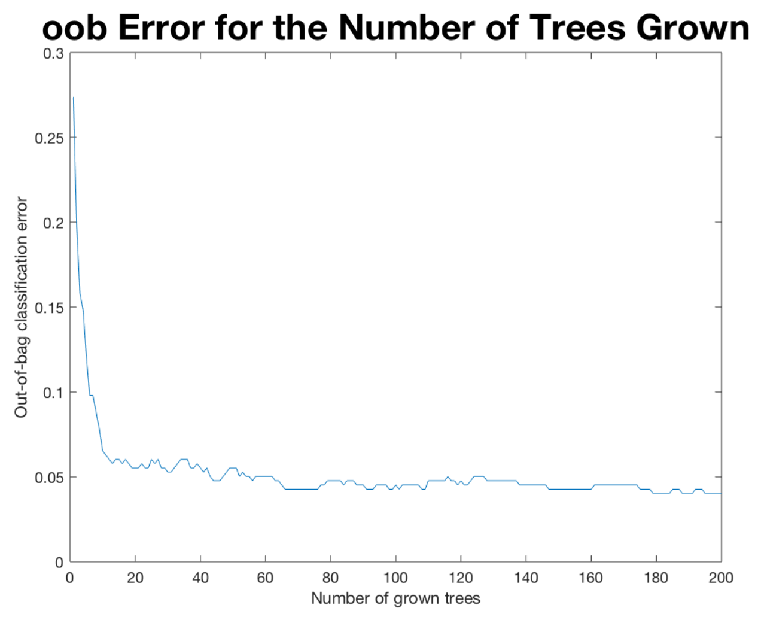
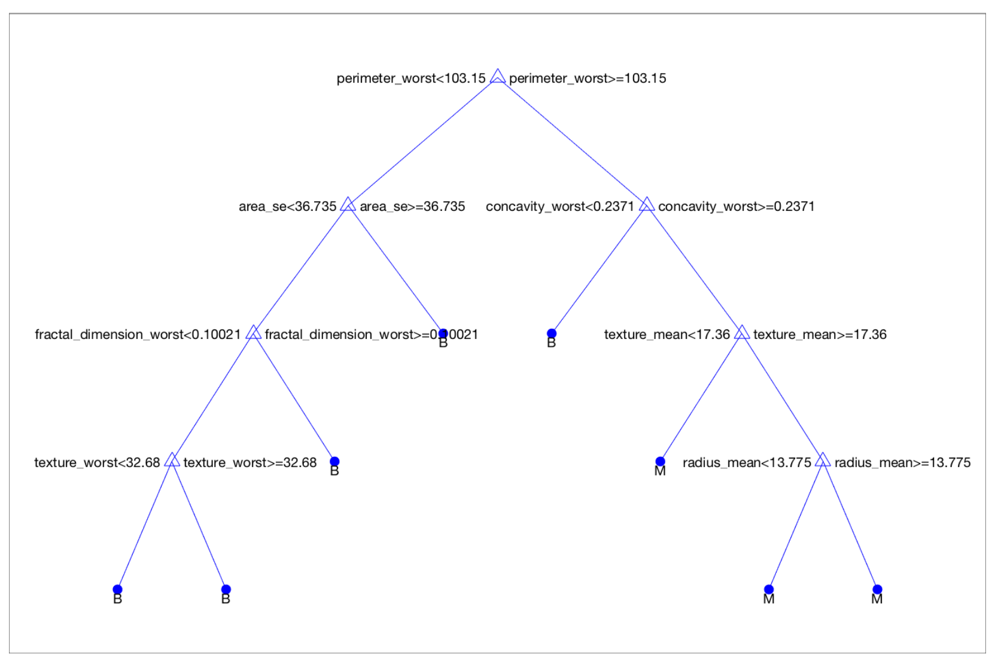


Fig 3



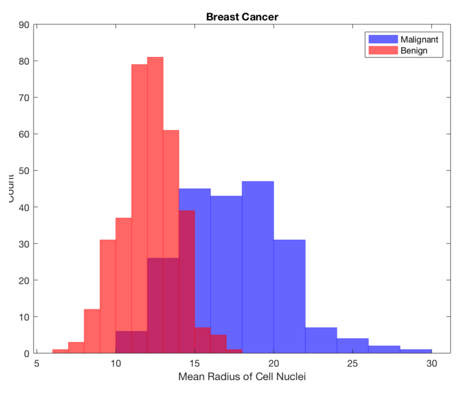
Data Visualisation on the Final Result for RF

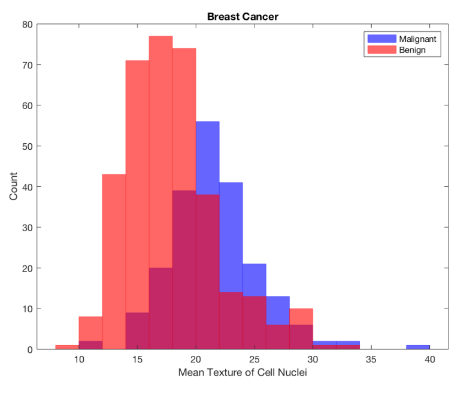


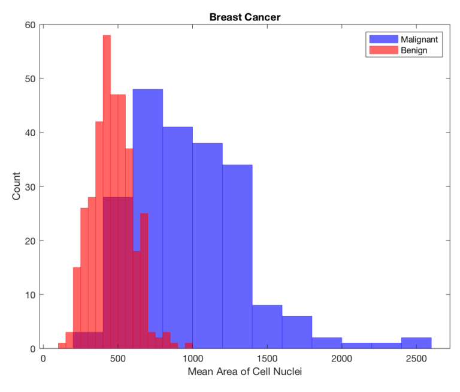


Analysis and critical evaluation of results

- In a medical diagnosis scenario such as Breast cancer diagnosis the prediction of the model should be very accurate and especially never predict false-negatives, for example, incorrectly predicting a malignant cancer is benign. The patient would not obtain correctly health-care and subsequently would cause serious complications.
- In this case, RF performed better than NB as expected. However, both models performed well as the results were above 90%.
- For a 70/30 split of the data, NB achieved a higher accuracy of 92.3977% than Shahnaz et al., [2] who achieved 91.8129%. Whereas RF only achieved 94.7368% compared to [2] who achieved 96.4912%
- For a 50/50 split of the data, the accuracy for NB is 91.55% and RF is 96.49%, compared to [2] who achieved 91.9014% for NB and 96.8310% for RF.
- Therefore, RF is still the better classifier for a 50/50 split of the data in our study, which is a significant contribution in early detection of the breast cancer.
- Analysing the means of all features was perceived that the values of texture, smoothness, symmetry or fractal does not show clear division over diagnosis.
- Each feature has a different dominance over diagnosis criterion, normally larger values of these parameters corresponding to malignant class. The overlapping part would be the "suspicious" diagnosis and it require more attention from specialists in order to decide correct treatment.

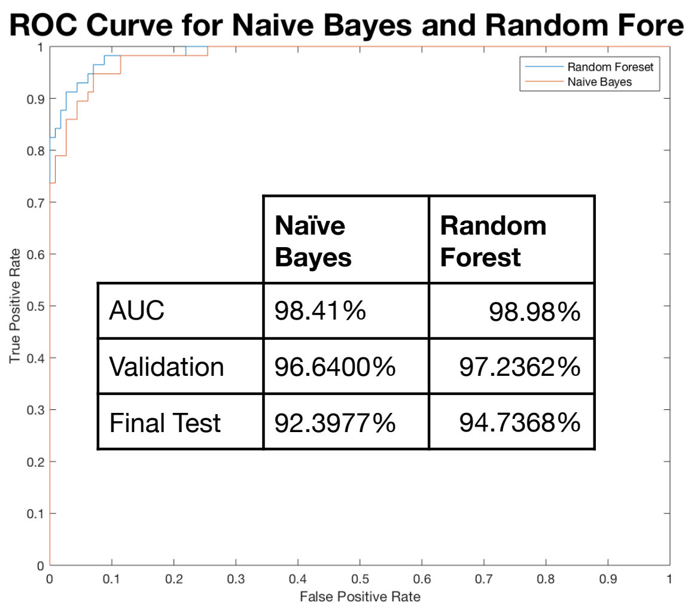






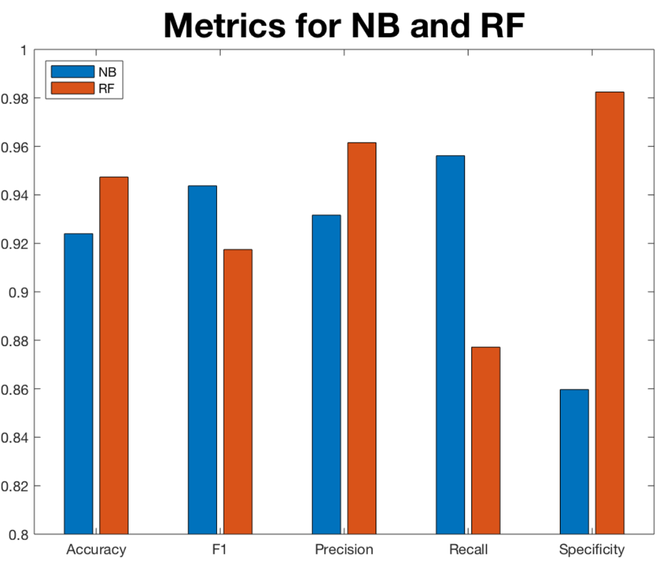
- Normal distribution for NB showed a clear superior performance compare to the other distributions, and the sample prior probabilities also gave the best accuracy result compare to the other values, as shown by fig 3.
- Validation accuracy of RF achieved the best results around 60-80 trees, however, computation time also increases with number of trees. When increased the number of trees to 100-200, no better results were achieved.
- RF used method of bagging that seems to enhance accuracy and reduce variance as shown by Breiman [4].
- Brieman et al., [5] showed that estimation of out-of-bag error is considered accurate when use the test set of the same size of the training set.

ROC Curve for Naive Bayes and Random Forest



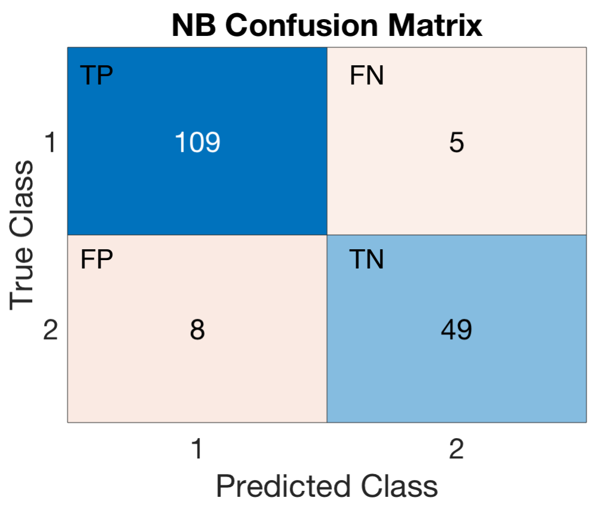
	Naïve Bayes	Random Forest
AUC	98.41%	98.98%
Validation	96.6400%	97.2362%
Final Test	92.3977%	94.7368%

Metrics for NB and RF



- As expected, NB was a lot quicker to implement than RF; the total elapsed time for NB to run is 114 seconds and RF is 2235 seconds (approx. 20 times longer than NB).
- The performance of the classification models can also be evaluated by the confusion matrix below. These results give insight on the errors being made by the classifiers as well as the types of errors being made.

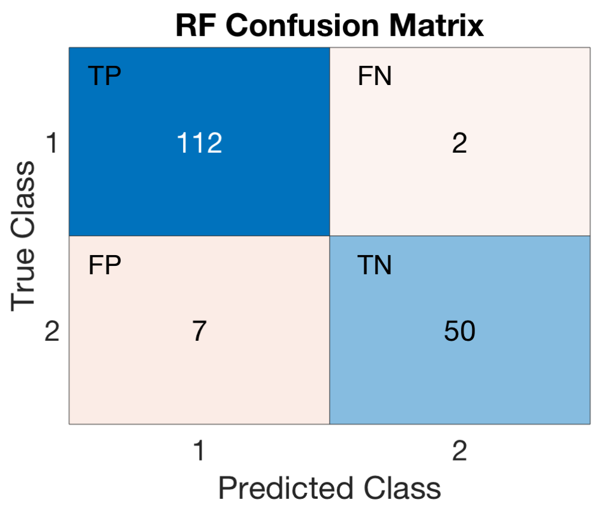
NB Confusion Matrix



	1	2
1	TP 109	FN 5
2	FP 8	TN 49

	NB	RF
Accuracy	92.40%	94.74%
Precision	93.16%	94.12%
Recall/ Sensitivity	95.61%	98.25%
Specificity	85.96%	87.72%
F1 Score	94.37%	96.14%

RF Confusion Matrix



	1	2
1	TP 112	FN 2
2	FP 7	TN 50

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad Precision = \frac{TP}{TP+FP} \quad Recall = \frac{TP}{TP+FN} \quad Specificity = \frac{TN}{TN+FP} \quad F1 = \frac{2*Precision*Recall}{Precision+Recall}$$

- When comparing the metrics of the confusion matrix generated to Street et al., [1] ,which has specificity of 86% and sensitivity of 90%, NB has only outperformed on sensitivity with 95.61% and RF has outperformed on both with 87.72% and 98.25% respectively.
- The precision (positive predictive value) of both models are good.
- The metrics of RF outperformed the metrics of NB in all areas.
- It is possible to observe when the model makes mistakes. Both models tend to predict more false-positives than false-negatives with the misclassification rate of 5.26% for RF and 7.6% for NB , this could be due to imbalanced dataset.

Lessons learned and further work

- There are limited hyperparameters available for a gird search for Naïve Bayes, therefore, to optimise the model a better understanding of the data is required.
- As this dataset is relatively small, the computational increases in time for RF model was not a concern but it should be considered when working with bigger datasets.
- Future work on Naïve Bayes: as shown by Rish [3], the model performs the worst in between two extreme scenarios: completely independent features and functionally dependent features. Therefore, the characteristics of the data plays a vital role in the effects the performance of the model.
- Future work on Random Forest: could consider using entropy based information gain or Gini index to decide the best features to use.
- For future work with imbalanced classes, some resampling techniques could be used to help to improve results. This is because most of the algorithms are designed to maximise accuracy and reduce error.
- A better understanding on the metrics of the confusion matrix is desired, so that the models can be improved based on this information.
- In a medical diagnosis case such as our study, Naïve Bayes and Random Forest might not be the most suitable classifier model for obtaining the best accuracy result, as shown by Shahnaz et al., [2].
- However, a hybrid model with Random Forest's ability to perform feature selection and the speed of Naïve Bayes should be considered, such that the prediction accuracy improves and run time reduces as shown by Chihab etl al., [6].

References:
[1] W.N. Street, W.H. Wolberg and O.L. Mangasarian, "Nuclear feature extraction for breast tumor diagnosis", in *IS&T/SPIE 1993 International Symposium on Electronic Imaging: Science and Technology, volume 1905, pages 861-870, San Jose, CA, 1993*.
[2] C. Shahnaz, J. Hossain, S. A. Fattah, S. Ghosh and A. I. Khan, "Efficient approaches for accuracy improvement of breast cancer classification using wisconsin database," *2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC)*, Dhaka, 2017, pp. 792-797.
[3] I. Rish, "An Empirical Study of the Naïve Bayes Classifier", IJCAI 2001 Work Empir Methods Artif Intell. 3.
[4] L. Breiman, Random Forests. Machine Learning, 45, 5-32, 2001.
[5] L.Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees*. Boca Raton, FL: CRC Press, 1984.
[6] Y. Chihab, A. Ait. Ouhman, M. Erritali, B.E. Ouahidi, "Detection & Classification of Internet Intrusion Based on the Combination of Random Forest and Naïve Bayes", in *International Journal of Engineering and Technology (IJET), Vol 5 No 3 Jun-Jul 2013*.