Breast Cancer

Taha Anizan

10/04/2021

Contents

1 Introduction

The Wisconsin breast cancer (diagnostic) database is a set of labelled multivariate data that has been used for classification based machine learning since it was first used in the 1990s and subsequently donated to the UCI machine learning repository. The objective of this project was to use this data set to train different algorithms in order to accurately diagnosis breast cancer based on a prediction as to whether a given sample of cells was malignant or benign tumour mass.

This report sets out the exploratory analysis of the data including the relationship between samples, and features, followed by an overview of the different methods deployed to develop predictive algorithms. The results are presented before a discussion on the relative performance of the models, the limitations of the models as well as the data itself, and opportunities for future work.

2 Exploratory Analysis

2.1 Data set overview

The Breast Cancer Wisconsin (Diagnostic) data set is a data.table, data.frame consisting of 32 columns and 569 rows.

There are 569 unique patient ID numbers, confirming that each row represents a sample from a unique patient. The diagnosis column includes character strings that classify whether the samples were diagnosed as benign (B) or malignant (M). The data set consists of 357 (100, 1) benign samples and 212 (100, 1) malignant samples.

Table 1: Description of nuclear features

Feature	Description
Radius	Mean of distances from center to points on the perimeter of individual nuclei
Texture	Variance (standard deviation) of grey-scale intensitities in the component pixels
Perimeter	Perimeter length of each nucleus
Area	Area as measured by counting pixels within each nucleus
Smoothness	Local variation in radius lengths
Compactness	Combination of perimeter and area using the formula: (perimeter^2 / area - 1.0)
Concavity	Number and severity of concavities (indentations) in the nuclear contour
Concave Points	Number of concavities in the nuclear contour
Symmetry	Symmetry of the nuclei as measured by length differences between lines perpendicular to the major axis and the cell boundary
Fractal Dim	Fractal dimension based on the 'coastline approximation' - 1.0

Each of the features described are such that larger values will typically indicate a higher likelihood of malignancy given that they reflect larger cells and/or more irregular shapes.

The ID column is not required for this project and was removed. The diagnosis column was reclassified as categorical data using the base 'as.factor()' function, with 2 levels, one for benign masses (B) and the other for malignant masses (M).

Prior to normalising the data and exploring distance and clustering of samples and features, the data-set was split into train and test sets in an 80:20 split.

Table 2: Mean scores

radius_m	texture_m	perimeter_m	area_m	smoothness_m	compactness_m	concavity_m	concave_points_m	symmetry_m	fractal_dim_m
Min. : 6.98	Min. : 9.7	Min. : 43.8	Min. : 144	Min. :0.0526	Min. :0.019	Min. :0.000	Min. :0.0000	Min. :0.106	Min. :0.0500
1st Qu.:11.70	1st Qu.:16.2	1st Qu.: 75.2	1st Qu.: 420	1st Qu.:0.0864	1st Qu.:0.065	1st Qu.:0.030	1st Qu.:0.0203	1st Qu.:0.162	1st Qu.:0.0577
Median :13.37	Median :18.8	Median : 86.2	Median : 551	Median :0.0959	Median :0.093	Median :0.062	Median :0.0335	Median :0.179	Median :0.0615
Mean :14.13	Mean :19.3	Mean : 92.0	Mean : 655	Mean :0.0964	Mean :0.104	Mean :0.089	Mean :0.0489	Mean :0.181	Mean :0.0628
3rd Qu.:15.78	3rd Qu.:21.8	3rd Qu.:104.1	3rd Qu.: 783	3rd Qu.:0.1053	3rd Qu.:0.130	3rd Qu.:0.131	3rd Qu.:0.0740	3rd Qu.:0.196	3rd Qu.:0.0661
Max. :28.11	Max. :39.3	Max. :188.5	Max. :2501	Max. :0.1634	Max. :0.345	Max. :0.427	Max. :0.2012	Max. :0.304	Max. :0.0974

Table 3: Worst scores

radius_w	texture_w	perimeter_w	area_w	smoothness_w	compactness_w	concavity_w	concave_points_w	symmetry_w	fractal_dim_w
Min. : 7.9	Min. :12.0	Min. : 50.4	Min. : 185	Min. :0.0712	Min. :0.027	Min. :0.000	Min. :0.0000	Min. :0.156	Min. :0.0550
1st Qu.:13.0	1st Qu.:21.1	1st Qu.: 84.1	1st Qu.: 515	1st Qu.:0.1166	1st Qu.:0.147	1st Qu.:0.114	1st Qu.:0.0649	1st Qu.:0.250	1st Qu.:0.0715
Median :15.0	Median :25.4	Median : 97.7	Median : 686	Median :0.1313	Median :0.212	Median :0.227	Median :0.0999	Median :0.282	Median :0.0800
Mean :16.3	Mean :25.7	Mean :107.3	Mean : 881	Mean :0.1324	Mean :0.254	Mean :0.272	Mean :0.1146	Mean :0.290	Mean :0.0839
3rd Qu.:18.8	3rd Qu.:29.7	3rd Qu.:125.4	3rd Qu.:1084	3rd Qu.:0.1460	3rd Qu.:0.339	3rd Qu.:0.383	3rd Qu.:0.1614	3rd Qu.:0.318	3rd Qu.:0.0921
Max. :36.0	Max. :49.5	Max. :251.2	Max. :4254	Max. :0.2226	Max. :1.058	Max. :1.252	Max. :0.2910	Max. :0.664	Max. :0.2075

Table 4: Standard error scores

radius_se	texture_se	perimeter_se	area_se	smoothness_se	compactness_se	concavity_se	concave_points_se	symmetry_se	fractal_dim_se
Min. :0.112	Min. :0.36	Min. : 0.76	Min. : 7	Min. :0.00171	Min. :0.0023	Min. :0.000	Min. :0.0000	Min. :0.0079	Min. :0.00089
1st Qu.:0.232	1st Qu.:0.83	1st Qu.: 1.61	1st Qu.: 18	1st Qu.:0.00517	1st Qu.:0.0131	1st Qu.:0.015	1st Qu.:0.0076	1st Qu.:0.0152	1st Qu.:0.00225
Median :0.324	Median :1.11	Median : 2.29	Median : 25	Median :0.00638	Median :0.0204	Median :0.026	Median :0.0109	Median :0.0187	Median :0.00319
Mean :0.405	Mean :1.22	Mean : 2.87	Mean : 40	Mean :0.00704	Mean :0.0255	Mean :0.032	Mean :0.0118	Mean :0.0205	Mean :0.00379
3rd Qu.:0.479	3rd Qu.:1.47	3rd Qu.: 3.36	3rd Qu.: 45	3rd Qu.:0.00815	3rd Qu.:0.0324	3rd Qu.:0.042	3rd Qu.:0.0147	3rd Qu.:0.0235	3rd Qu.:0.00456
Max. :2.873	Max. :4.88	Max. :21.98	Max. :542	Max. :0.03113	Max. :0.1354	Max. :0.396	Max. :0.0528	Max. :0.0790	Max. :0.02984

The balance of classes was consistent between the train set (malignant = 100, 1) and test set (malignant = 100, 1). Further data exploration was conducted with the train set only.

The train set was normalised using the sweep function firstly to centre each data point (x) around zero by subtracting the sample mean ($\overline{x}$) by column and then to scale each data point by dividing by the sample standard deviation (S) by column, yielding z-scores (1).

$$z = \frac{(x - \overline{x})}{S} \tag{1}$$

One way of measuring the variance between samples is to calculate the Euclidean distance as

$$d(x_1, x_2) = \sqrt{\sum_{i=1}^{569} (x_{1,i} - x_{2,i})^2} \tag{2}$$

The average distance between all samples included in the train set was 6.99. Benign samples were closer to each other (6.39) than to malignant samples (8.53). Of note, benign samples were also closer to each other than malignant samples (8.02) were from each other, indicating greater variance in the measured features in malignant cells.

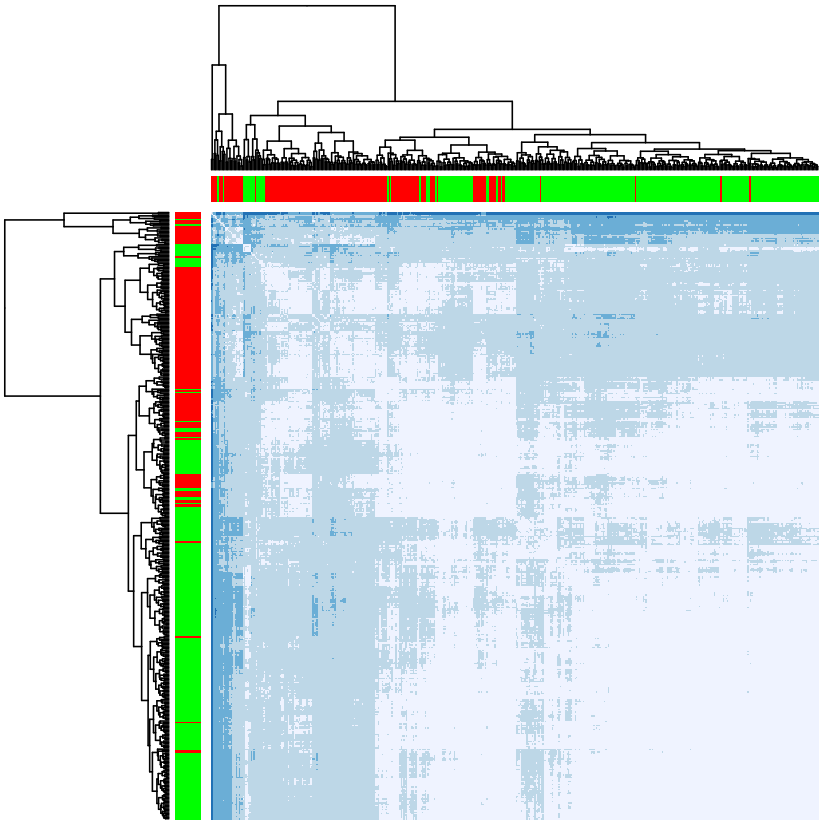


Figure 1: Heatmap of distance between benign (green) and malignant (red) samples

2.1.1 Variance within features

The nearZeroVar function within the caret package demonstrated that none of the features in the train set had either zero or near zero variance. The lowest percentage of unique values for any feature was 77.533 and the mean of all features was 91.601. Moreover, the mean frequency ratio was 1.81 with 100, 1 of all features having a frequency ratio score of 2 or less. This analysis supports the inclusion of all features in the algorithm based on their within feature variance.

2.1.2 Hierarchical clustering

The train set was transposed so that the features were moved from the columns of the matrix to the rows and the dist function operated to calculate the Euclidean distance between each feature.

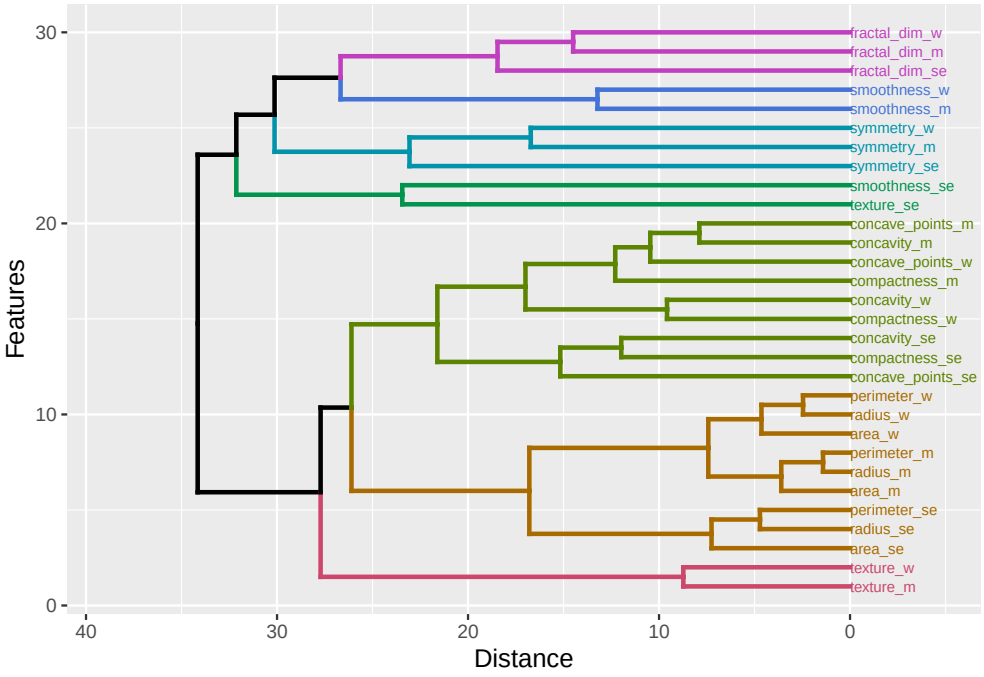


Figure 2: Dendrogram of hierarchical clusters of features

The features that are closest together are those that measure nuclear size, including radius, perimeter and area. Those features that measure shape rather than size are further apart from each other, although concavity, compactness and the number of concave points are relatively near to each other.

2.1.3 Correlation between features

$$r_{x_1,x_2} = \frac{\sum_{i=1}^{n}\left(x_{1,i}-\overline{x}_1\right)\left(x_{2,i}-\overline{x}_2\right)}{\sqrt{\sum_{i=1}^{n}\left(x_{1,i}-\overline{x}_1\right)^2}\sqrt{\sum_{i=1}^{n}\left(x_{2,i}-\overline{x}_2\right)^2}} \tag{3}$$

Overall, there is a low level of correlation between features; indeed, the mean correlation coefficient of features in the train set is 0.43. The greatest correlation is between the various measures of cell size, namely radius, perimeter and area and, to a lesser extent between some of the measures of cell shape, namely fractal dimension, symmetry and smoothness.

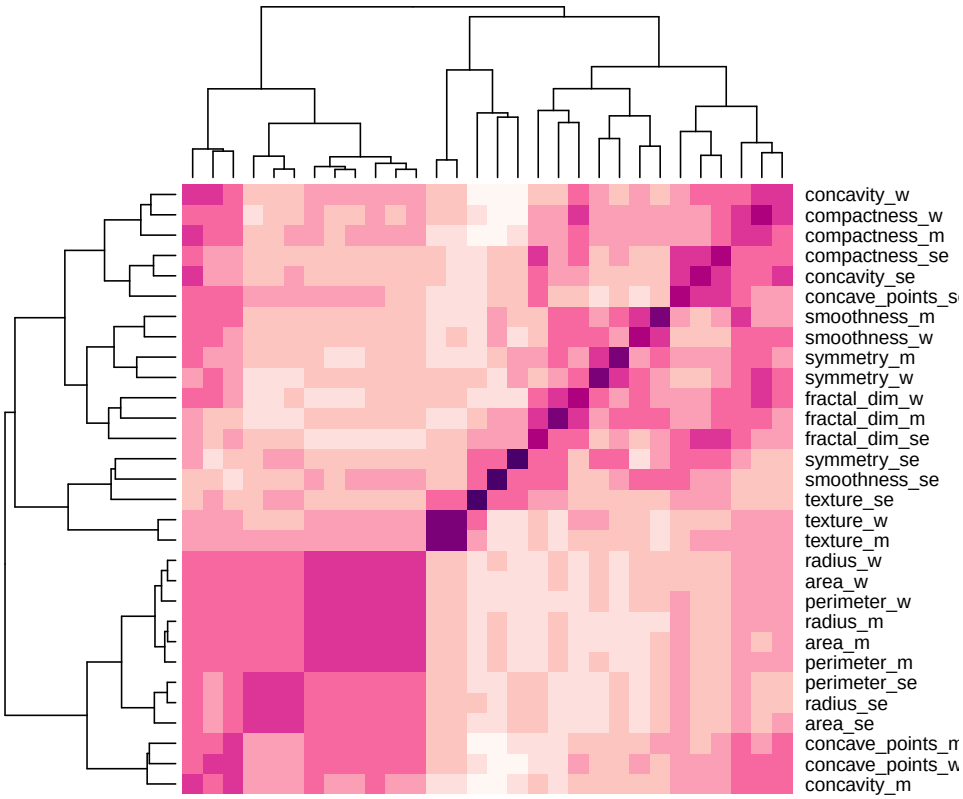


Figure 3: Heatmap of correlation between features

Ten features have a correlation of 0.9 or more, namely concavity_m, concave_points_m, perimeter_w, radius_w, perimeter_m, area_w, radius_m, perimeter_se, area_se, and texture_m. Excluding these features from unsupervised methods of developing the predictive algorithm may be beneficial.

2.1.4 Principal Component Analysis

Principal component analysis (PCA) is a technique for transforming data-sets in order to reduce dimensionality without reducing the number of features by identifying the principal components which explain as much of the data variance as possible. PCA can be used to improve visualisation of multidimensional data and, potentially, to improve the predictive accuracy of classification models.

Table 5: First 10 Principal Components

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
Standard deviation	3.692	2.354	1.654	1.406	1.243	1.103	0.844	0.706	0.616	0.612
Proportion of Variance	0.454	0.185	0.091	0.066	0.052	0.041	0.024	0.017	0.013	0.012
Cumulative Proportion	0.454	0.639	0.730	0.796	0.848	0.888	0.912	0.929	0.941	0.954

In most cases the spread is greater for malignant masses than for benign masses. PC1 is the only component for which the interquartile ranges do not overlap. Principal component analysis does not take into account the classification of data, in this case the diagnosis assigned to each sample.

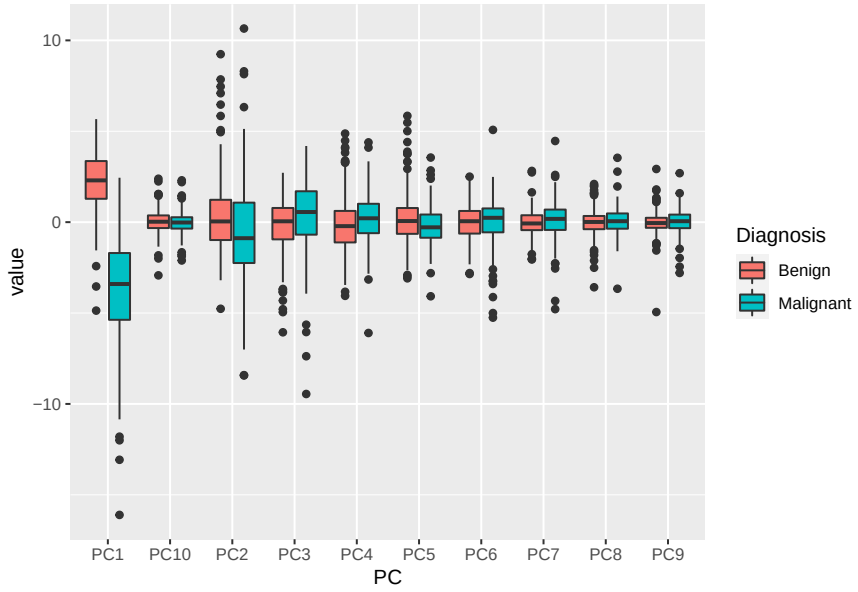


Figure 4: Box plots of top 10 PCs by diagnosis

The graph shows that the malignant data-points are more spread out than the benign data-points and that more of the variance can be accounted for on the x-axis (PC1) than on the y-axis (PC2). Ellipses help to visualise this even better, firstly with a larger ellipse for malignant data-points than for benign data-points and considerable separation of data by classification, despite some overlap. This analysis support the use of PCA in algorithm development to predict diagnosis from this data-set.

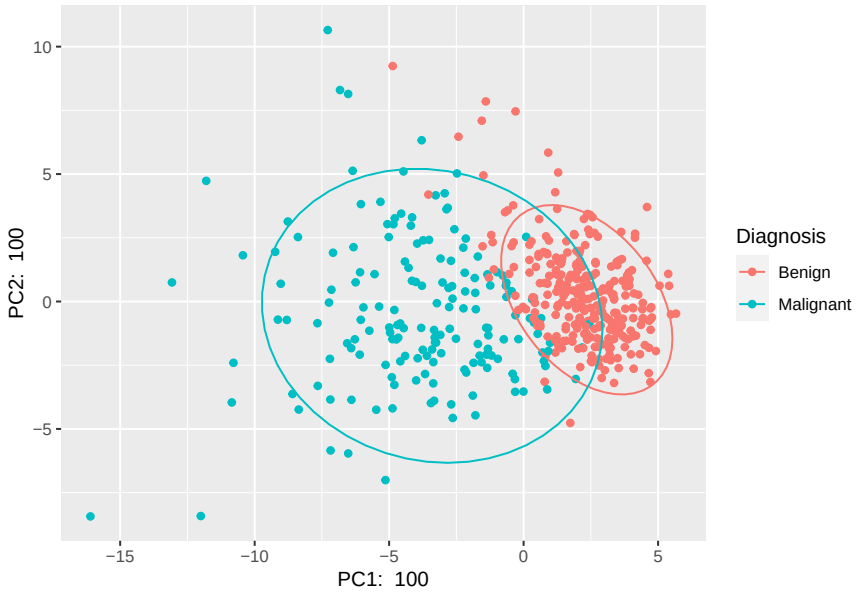


Figure 5: Scatter plot of PC1 and PC2 by diagnosis

3 Methods

3.1 Pre-processing

The exploratory analysis of the Wisconsin breast cancer (diagnostic) data-set revealed patterns across both samples and features that support the use of machine learning techniques to develop predictive algorithms.

An empty data frame was generated in which to store key performance metrics for each model developed, namely the overall accuracy of the model (4), the sensitivity, or true positive rate (5), the specificity, or true negative rate (6).

$$\text{Accuracy} = \frac{\text{True Positive} + \text{True Negative}}{\text{True Positive} + \text{True Negative} + \text{False Positive} + \text{False Negative}} \tag{4}$$

$$\text{Sensitivity} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \tag{5}$$

$$\text{Specificity} = \frac{\text{True Negative}}{\text{True Negative} + \text{False Positive}} \tag{6}$$

The F1 score, a harmonic mean of precision, or positive predictive value, and sensitivity (7) is another measure of a model's accuracy and was also included. To aid analysis, the false negative rates (8) and false positive rates (9) were also computed.

$$\text{F1 score} = \frac{2(\text{True Positive})}{2(\text{True Positive}) + \text{False Positive} + \text{False Negative}} \tag{7}$$

$$\text{False Negative Rate} = \frac{\text{False Negative}}{\text{False Negative} + \text{True Positive}} = 1 \text{ - Sensitivity} \tag{8}$$

$$\text{False Positive Rate} = \frac{\text{False Positive}}{\text{False Positive} + \text{True Negative}} = 1 \text{ - Specificity} \tag{9}$$

Cross-validation is an important technique used to measure performance of a model without recourse to the test data-set, allowing the test set to be reserved for the final hold-out test of each model and minimising the risk of over-fitting. It is also a useful technique for tuning parameters for those models that require it (e.g., to tune the number of neighbours, k, to include in a k-nearest neighbours model). For the purposes of measuring performance within the resamples only the final predictions and summary performance metrics (accuracy and kappa scores) were saved based on the tuned parameters where applicable. Kappa scores are another measure of the agreement between observed (p_o) and expected values (p_e) and, unlike overall accuracy, take account of the chance that a prediction (or observed value) will match the true (or expected) value.

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{10}$$

3.2 Random sampling

The first model to be tested involved random sampling with equal probability for each outcome, i.e. equivalent to a coin toss. Given the imbalance in prevalence of benign and malignant samples in the data-set, a second model was built with weighted random sampling, where the prevalence of each class of samples was used to define the probability of each outcome within the random sample.

3.2.1 k-means clustering

k-means clustering is another form of unsupervised modelling where the number of clusters is defined in advance. k-means clustering is an attractive option for large data-sets because it is a relatively simple approach to clustering and as such is computationally faster than hierarchical clustering.

Two version of the k-means model were developed. The first used the normalised data from the full train data-set. The second selected out those features which were highly correlated, i.e. the ten features where the correlation coefficient exceeded the 0.9 cutoff defined in the exploratory analysis. k-means clustering places greater weight on larger clusters and variables that are highly correlated (that form a large cluster) may therefore carry greater weight in the prediction algorithm.

3.2.2 Generative modelling

$$p(\mathbf{x}) = \Pr(Y = 1|\mathbf{X} = \mathbf{x}) = \frac{f_{\mathbf{X}|Y=1}(\mathbf{x}) \Pr(Y = 1)}{f_{\mathbf{X}|Y=0}(\mathbf{x}) \Pr(Y = 0) + f_{\mathbf{X}|Y=1}(\mathbf{x}) \Pr(Y = 1)} \tag{11}$$

The Naive Bayes model assumes that all features within the data-set are equally important and independent.

Other generative models include linear discriminative analysis (LDA) and quadratic discriminative analysis (QDA).

3.2.2.1 Logistic Regression Logistic regression is the most commonly used form of generalised linear model (GLM). Linear regression assumes that the predictor, X, and the outcome Y, follow a bivariate normal distribution such that the conditional expectation, i.e. the expected outcome Y for a given predictor X, fits the regression line.

$$p(x) = \Pr(Y = 1|X = x) = \beta_0 + \beta_1 x \tag{12}$$

$$g\{\Pr(Y = 1|X = x)\} = \beta_0 + \beta_1 x \tag{13}$$

3.2.2.2 Nearest neighbour model The k-Nearest neighbour model (kNN) is a simple approach to supervised machine learning that assumes proximity equates to similarity, once again measuring the Euclidean distance between two points in multidimensional data. Unlike hierarchical and k-means clustering, the KNN model is a form of supervised learning, i.e. it relies on and makes use of the diagnosis labels in the training set in order to predict diagnosis in an unlabelled test set.

3.2.2.3 Random forest model Many algorithms, including some of those described above, suffer from diminished performance due to multidimensionality of data. As has been described, PCA can be useful to reduce the number of dimensions required as part of pre-processing of data prior to training one of these algorithms. Decisions trees are another way to address this issue, effectively partitioning the data such that final predictions can be made on a smaller subset of predictors.

3.2.3 Ensemble model

Ensembles are combinations of individual model predictions that seek to improve both stability and accuracy of the final result, just as the random forest algorithm uses combinations of individual decision trees. There is no established convention for selecting which models to include in the ensemble. One approach is to establish a performance cutoff within the training sets, via cross-validation, in order to avoid selection based on performance in the test set.

4 Results

4.1 Overall performance

Table 6: Key performance metrics for each model

Method	Accuracy	Sensitivity	Specificity	F1	FNR	FPR
Random sampling						
Random sample	0.55	0.60	0.51	0.50	100	1
Weighted random sample	0.46	0.23	0.60	0.24	100	1
Unsupervised models						
K-means clustering	0.89	0.86	0.90	0.85	100	1
K-means (without highly correlated features)	0.76	0.63	0.83	0.66	100	1
Generative models						
NA	NA	NA	NA	NA	NA	NA
Linear Discriminant Analysis	0.96	0.88	1.00	0.94	100	1
Quadratic Discriminant Analysis	0.97	0.98	0.96	0.95	100	1
Quadratic Discriminant Analysis (with PCA)	0.97	0.95	0.99	0.96	100	1
Discriminative models						
Logistic regression	0.97	0.91	1.00	0.95	100	1
Logistic regression (with PCA)	0.99	0.98	1.00	0.99	100	1
K Nearest Neighbour	0.97	0.93	1.00	0.96	100	1
Random Forest	0.97	0.95	0.99	0.96	100	1
NA	NA	NA	NA	NA	NA	NA
Ensemble						
Ensemble	0.99	0.98	1.00	0.99	100	1

Note:
FNR = false negative rate; FPR = false positive rate; PCA = principal component analysis

k-means clustering improved accuracy, with an F1 score of 0.85 but still not to an acceptable level, with a false negative rate of 100 and a false positive rate of 1. Of note, removing the highly correlated features (i.e. those with a correlation coefficient above 0.9) from the data-set reduced the accuracy of the k-means clustering model, yielding a reduced F1 score of 0.66 and reducing both the specificity and, in particular, the sensitivity of the model.

Linear discriminant analysis and quadratic discriminant analysis improved on the performance of the Naive Bayes model, with F1 scores of 0.94 and 0.95 respectively. The LDA model achieved a specificity of 1.00 (i.e. FPR of 1) but this was offset by reduced sensitivity (0.88), i.e. an unacceptable FNR of 100.

The QDA model delivered a better balance between FNR (100) and FPR (1). Of note, dimension reduction through pre-processing the training data with PCA improved the specificity of the QDA model, reducing the FPR to 1 but it reduced the sensitivity, i.e. increased the FNR to 100.

The discriminative models of supervised learning were the best performing models with this data-set. Logistic regression achieved an overall accuracy of 0.97. This was improved further to 0.99 with dimension reduction using PCA by improving the sensitivity of the model, achieving FNR and FPR of only 100 and 1 respectively.

The nearest neighbours model performed best when the number of neighbours, k, was defined as three. On this basis, the overall accuracy of 0.97 but with lower sensitivity (0.93).

4.2 Variable importance

The worst performing discriminative model, logistic regression appears to be most reliant on mean and standard error scores for features related to nuclear shape in the top five variables of importance, namely,

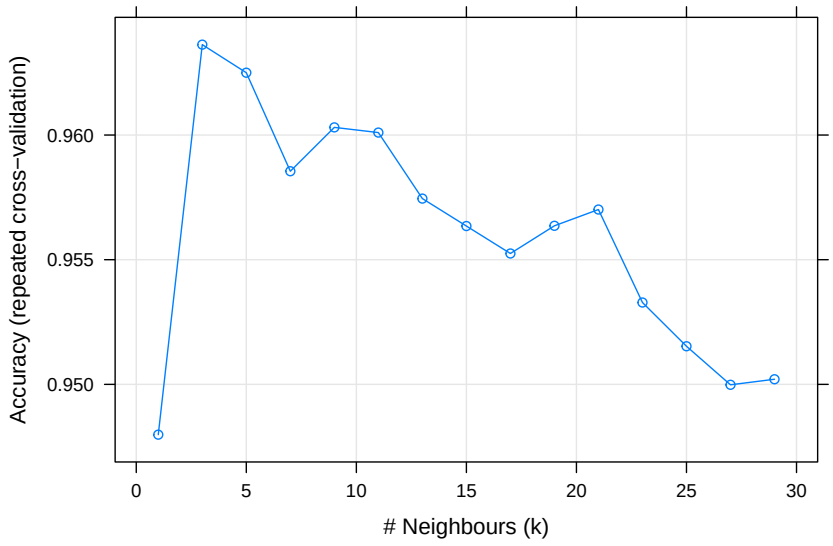


Figure 6: Tuning results for nearest neighbour model during cross-validation

compactness_m, concavity_m, symmetry_m, compactness_se, and fractal_dim_se.

Looking at the best performing models, by comparison, the worst scores feature heavily in the top five variables of importance for the nearest neighbour model (4 out of 5), the random forest model (4 out of 5)

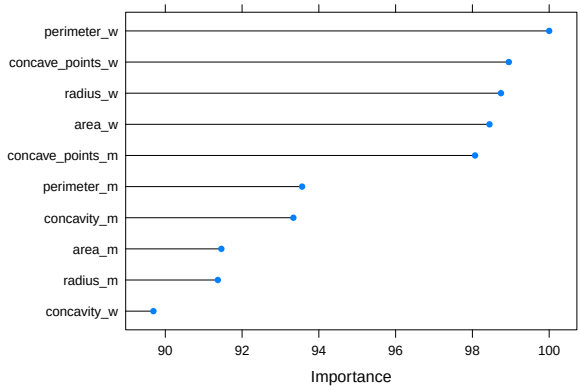
5 Discussion

Despite the success of the neural network, there is merit in selecting the ensemble of supervised models as the preferred algorithm given that it also correctly predicted diagnosis and has the benefit of mitigating the risk of over-training with an individual model, making it more likely to provide reproducible results in different data-sets that include the same feature information.

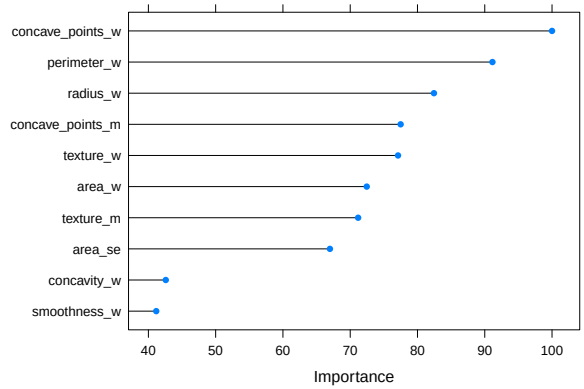
The current dataset suffers from a number of inherent biases that represent possible limitations to the reproducibilty of the performance achieved here. The samples were collected from a single site and from a consecutive series of patients. Operator biases will have included those responsible for conducting the biopsies, digitising the images to measure each of the features and even the clinical diagnoses made to classify each sample as benign or malignant. The methods for capturing nuclear size and shape information in 1995 were relatively rudimentary and more advanced image processing techniques available today would complement the complex machine learning algorithms (such as convolutional neural networking) now available to capture differences between benign and malignant samples.

6 Conclusions

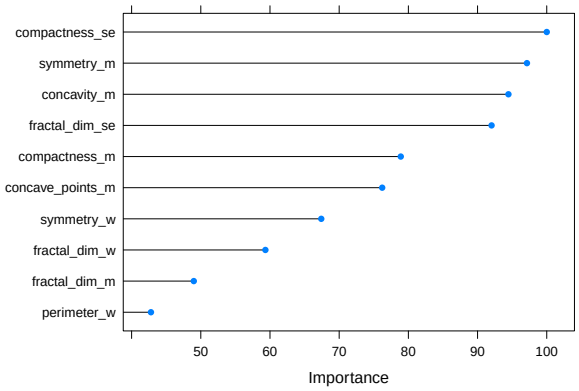
The objective of this project was to use the Wisconsin breast cancer data set to train different algorithms to accurately diagnosis breast cancer. An exploratory analysis of the data revealed that measures of both distance and correlation of nuclear features could be useful in both clustering and classifying individual samples. All of the models developed performed better than random sampling but supervised learning was more accurate than unsupervised learning, and discriminative models were more effective than generative models. The neural network was the most successful individual model, with perfect accuracy within the test data and this performance was matched with an ensemble of supervised models.



(a) K Nearest Neighbours (metric: ROC curve)



(b) Random Forest (metric: model specific)



(c) Neural Network (metric: model specific)

Figure 7: Variable importance

These results confirm the potential of machine learning to accurately predict diagnosis of breast cancer using samples obtained via fine needle aspiration biopsy, with high levels of both sensitivity and specificity.