ST662 Topics in Data Analytics
Analysis of Breast Cancer Data 2022

Introduction

In 2020, breast cancer was the leading cause of cancer deaths among women across the world (ACS, 2022). There are many research projects in progress to improve early detection rates as early detection leads to better outcomes (ACS, 2022). Breast cancer can be detected through complex tests, including blood-based markers. The discovery of a simpler predictive blood test has the potential to make a significant contribution both in terms of improved early detection and the opportunity cost of the current population-based screening system.

Given enough of the right kind of data, a machine learning algorithm can abstract and generalise data into a predictive equation which can evaluate the likelihood of achieving a successful prediction (Lantz, 2019). A data scientist has many options to choose from such as classification models and cluster models, including generalised linear models (GLM), linear discriminant analysis (LDA), quadratic discriminant analysis (QDA) and K-nearest neighbours (KNN).

In the process of developing a prediction model based on blood results for breast cancer, one needs to consider that public health screening test results need to be sensitive to the diseases screened for and specific to the occurrence of the disease. For cancer screening, the standards for sensitivity are higher than specificity, as missing a cancer diagnosis (type II error) has more grave consequences than misdiagnosing cancer (type I error). Algorithms can provide these measures as well as an overall error rate for predictions.

This paper explores published breast cancer data (Patrício et al, 2018) and its relationship to a breast cancer diagnosis and then uses this data to develop predictive models. The predictive models (GLM, LDA, QDA and KNN) were evaluated for accuracy and specificity, and the KNN model (using all predictors, scaled, k=5) was found to be the most accurate (23% error rate) model but the QDA model (using three predictors) was the most effective in minimising Type II errors (13% error rate). This paper recommends the KNN model for overall accuracy but recommends in future analysis it is validated with larger datasets and other models e.g. support vector machine models (Nindrea, 2018). Additionally better age control in control groups is recommended as well as consideration for splitting women into pre and post-menopausal groups (Crisóstomo, 2022).

Methods

Published data (Patrício et al, 2018) included the records of 116 women. The health status of the two groups was the classification type response variable: 52 healthy participants and 64 patients. There were 9 predictors: BMI and age, and 7 blood results MCP.1, leptin, resistin, adiponectin plus a group of glucose-related results: HOMA, glucose, insulin.

To understand the relationship between the predictors and breast cancer, we loaded the data into R studio (RStudio 2021.09.0) to create visualisations and predictive models. Initially, to validate the classification groups, age and BMI were compared for significant differences with Welch T-Tests and all predictors were correlated.

Boxplots for each predictor were then visualised between groups, to compare their median and interquartile ranges and to examine outliers. Scatterplots were produced including regression lines for each blood result against age and they enabled observations of trends that were different from the boxplots. Further boxplots for each predictor against age ranges were also produced to check the variation between groups. To examine the relationship between age and the predictors by classification group, Welch T-Tests were performed.

To investigate the possibility of building a predictive model to help identify breast cancer, the following machine learning models were selected: GLM, LDA, QDA, and KNN. A regression algorithm was used to select the optimum number of predictors that offer a significant explanation of the variability of the data. The validation set approach (regsubset) was used to decide which predictors to include for linear regression. This predictor choice was then replicated in all models for comparison purposes. We evaluated principal component analysis to examine clusters and reduce the variables used to analyse the data.

Additionally, individual predictors that were identified as being of interest during the exploration stage (boxplots/scatterplots/t-tests), were used to generate all the models. Finally, all predictors were also used to generate models.

Cross-Validation was used to evaluate the models as this method is beneficial when the data volume is small. In this method, several folds are used to create test data, with one set of data kept aside for testing in set of training folds. Each fold was applied iteratively to test the predictive power of the model.

For KNN, several nearest neighbours were tested, and the best k was chosen for the minimum overall error rate. Data were scaled for the KNN model.

All model outputs were independently checked and validated within the team. The results of the models were output as a table and a graph of overall error rates by model type including accuracy, specificity, and sensitivity.

Results

The evaluation of the data showed that the groups were comparable in age and BMI, there was no significant difference between the median age (p-value=0.65) or BMI (p-value=0.13) for classification groups. It was noted there was minimal correlation among the predictors (Figure 1). The exceptions were HOMA and glucose and HOMA and insulin which were strong positive correlations and insulin and glucose, and BMI and leptin, were positive correlations.

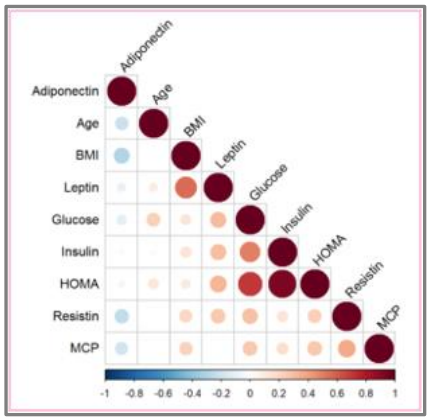

Figure 1 Correlation between predictors

For the boxplots of predictors by group (Figure 2), the age of healthy participants was wider than for patients. Both patients and healthy participants displayed outliers in all blood predictors, with one standout outlier in resistin. The visual inspection of boxplots showed the predictors glucose, insulin, HOMA, and resistin were different between groups.

Subsequently, the scatterplot with linear analysis (Figure 3) showed that leptin and BMI have an opposite relationship across the age range and may also be useful predictors. MCP.1, glucose, and HOMA showed higher values for patients across all ages, with the difference widening in older participants between patients and healthy.

Boxplots for patient group and age groups (Appendix 1, Figure 5) showed greater increases in older age categories for BMI, Glucose, Insulin, HOMA, Leptin and Resistin for patients than healthy participants.

Our findings from PCA were that no clusters emerged for healthy and patient groups. The proportion of variability explained for the first two components was considered low and so was not used in subsequent analysis. (Appendix 1, Figure 6)

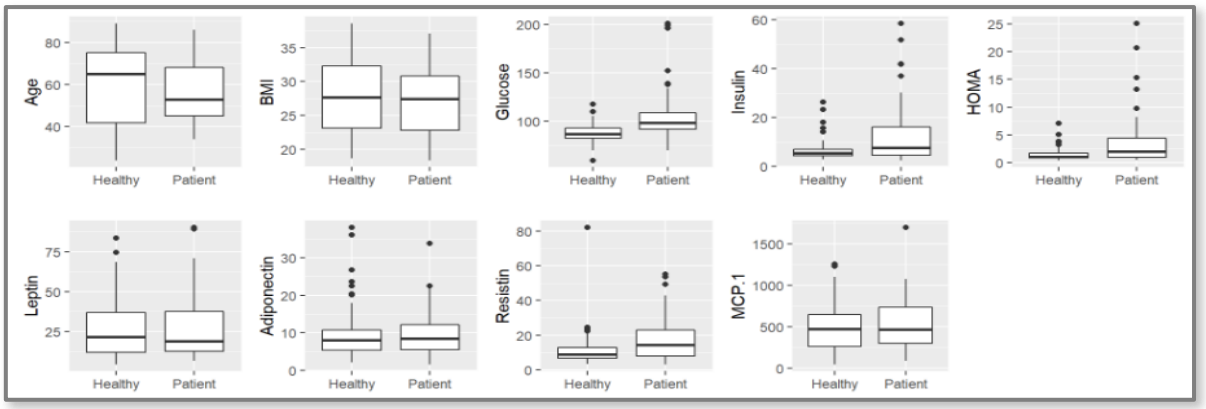

Figure 2:Boxplot comparison of predictors by healthy participants and patient groups.

Four tests were completed across all models with a variety of predictors. From a regression algorithm plot (Appendix 1, Figures 7 and 8) 5 predictors were chosen (option 1) and were BMI, glucose, insulin, HOMA, and resistin. From exploratory analysis, 3 predictors were chosen (option 2) which were BMI, resistin and HOMA (HOMA to represent the three correlated blood results: HOMA, glucose, insulin). From the exploratory analysis of the data, 7 predictors were chosen and were age, BMI, HOMA, glucose, insulin, leptin, resistin (option 3). Lastly, all predictors (option 4) were selected. The analysis of nearest neighbours options showed that k=5 gave the minimum overall error.

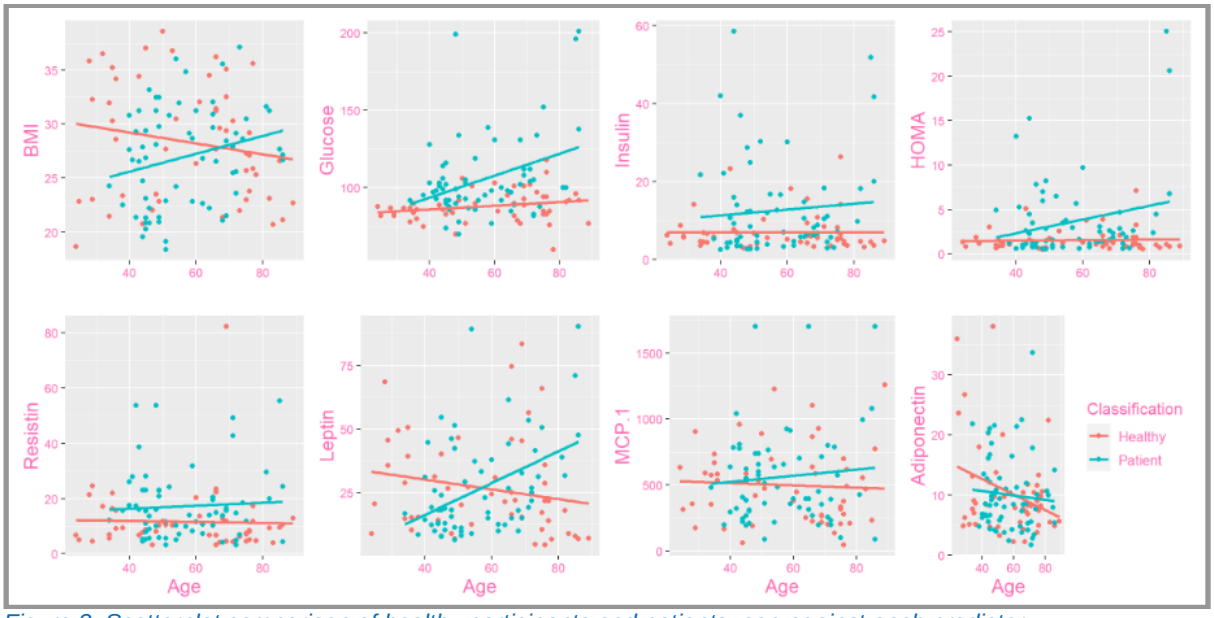

Figure 3: Scatterplot comparison of healthy participants and patients, age against each predictor.

Figure 4 shows which predictor group delivered the best result for each machine learning model with the overall error score, a score for type I, and type II errors (see Appendix 1, Figure 9 for all results). The best model for accuracy is KNN (all predictors), with an overall error between 0.23-0.32. The model with the lowest type I error detection is GLM (with 3 predictors), discounting the models where results are classified completely in one error category. The best model for type II error detection is QDA (3 predictors).

Figure 4 Graph showing overall, Type I and Type II errors by model and predictor choice.

Conclusions

The groups were comparable for age and BMI, but the age variation between groups from the visual analysis was a concern particularly because age could be better controlled in the healthy group. The youngest patient was 34, but there were 6 younger healthy participants than this. A further observation was that as the groups aged, predictors varied more, a recommendation is to consider women in a subsequent study to control for the potential influence of menopause (Crisóstomo, 2022).

The visual analysis with boxplots and scatterplots was a valuable exercise, prior to modelling, to support familiarisation with the dataset. This analysis also led to the selection of the best predictors (Option 1 and 2) for the sensitivity of results, other results benefitted from other approaches to selecting predictors.

The model that was the best performer in predicting cancer was a KNN, with a k=5, cross-validated on scaled data with all predictors. The model has an overall error rate of 23%, with an 18% likelihood of misdiagnosing cancer. This model could be improved by validation with a significantly larger dataset and investigating the earlier suggestion of standardising age categories or breaking into younger/older groups of women.

The model that was the best performer in determining the accuracy of positive diagnosis was QDA with three predictors (Age, HOMA and resistin). This model has a type II error rate of 13%. This model is however not such a good performer on type I errors (62%), and it could lead to a high proportion of unnecessary additional tests.

The models chosen were supervised classification models, parametric (GLM, LDA, QDA) and non-parametric (KNN). Parametric models support simple decision boundaries between patients and healthy participants and lead to a stable model based on normal distributions of means and variances. However, the boxplots in our investigations showed a noteworthy degree of difference, supporting the assertion that groups more homogenous in age may be a method to improve model outcomes.

This research would benefit from a business context so that it could be focused on health management needs. For instance, prioritising type II errors or large scale public health testing or fast access tests may influence the selection of a model or the ability to either collate more data or validate models as part of the business process for data collection.

This analysis demonstrates it is possible to predict breast cancer (accuracy 68-77%). KNN is the recommended model for breast cancer prediction, using all predictors. Other models, including support vector models, have proved more successful (Nindrea, 2018). Further research could consider different models, and additional model predictors (e.g. menopause). It is recommended groups be better matched on age, separating ages into older and younger cohorts, and controlling better for age would improve models.

References

ACS 2022 https://www.cancer.org/cancer/breast-cancer/screening-tests-and-early-detection/american-cancer-society-recommendations-for-the-early-detection-of-breast-cancer.html accessed 8/4/2022

Crisóstomo, J., Matafome, P., Santos-Silva, D., Gomes, A.L., Gomes, M., Patrício, M., Letra, L., Sarmento-Ribeiro, A.B., Santos, L., Seiça, R., 2022. Hyperresistinemia and metabolic dysregulation: a risky crosstalk in obese breast cancer. Endocrine, 53, pp433-442.

Lantz, B., 2019. Machine Learning with R, 3rd edition. Packt publishers, Birmingham.

Nindrea, R.D., Aryandono, T., Lazuardi, L., Dwiphrahasto, I., 2018. Diagnostic Test Accuracy of Different Machine Learning Algorithms for Breast Cancer Risk Calculation: a Meta-Analysis. Asian Pacific J. Cancer Prevention, Vol 19.

Patrício, M., Pereira, J., Crisóstomo, J., Matafome, P., Gomes, M., Seiça, R., and Caramelo, F., 2018. Using Resistin, glucose, age, and BMI to predict the presence of breast cancer. BMC Cancer, 18(1).

Appendix

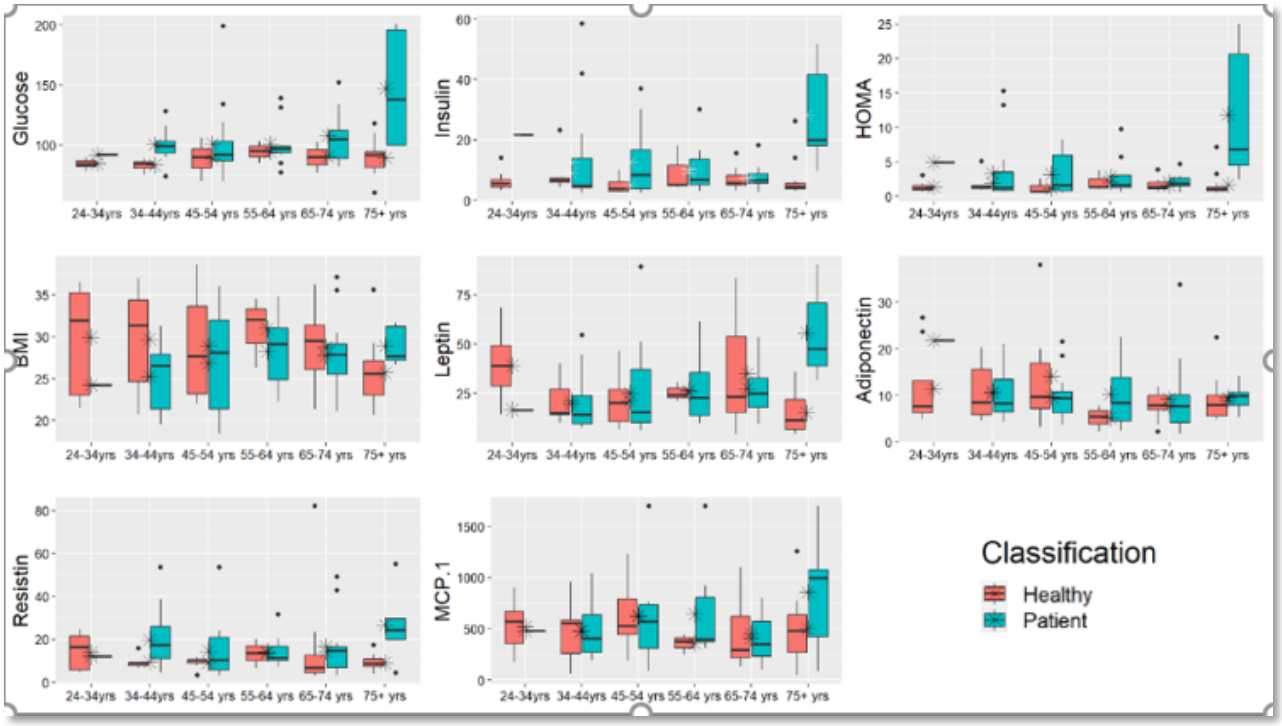

Figure 5: Boxplot comparisons by age groupings for healthy participants and patients against each predictor.

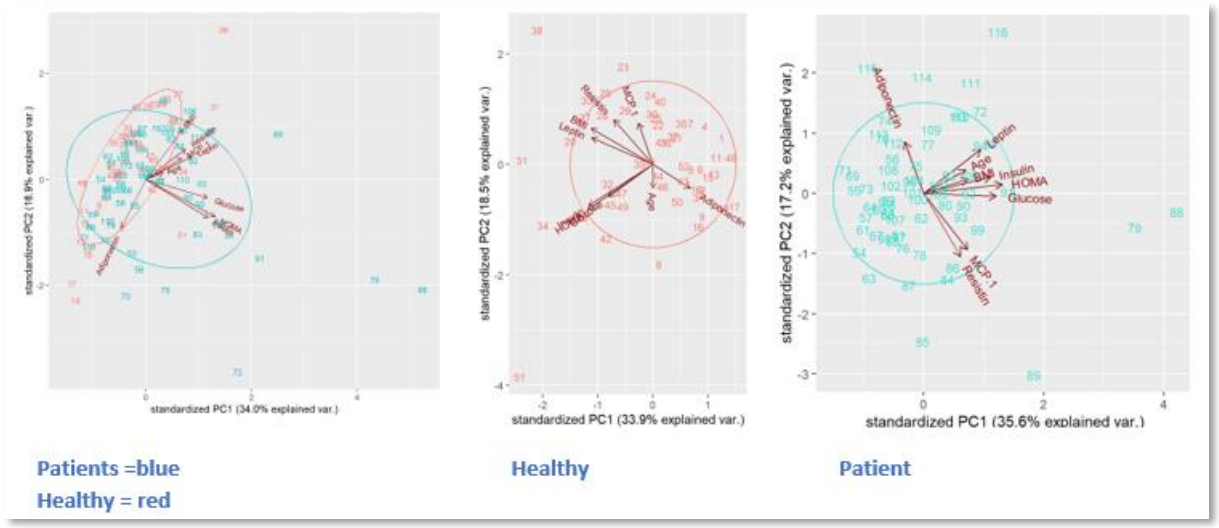

Figure 6: Principal component analysis results for the whole group and separately by patient and healthy participants

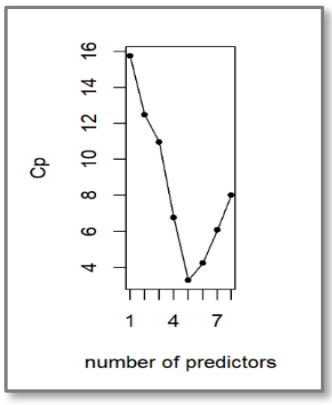

Figure 7: Scree plot to select 5 of predictors

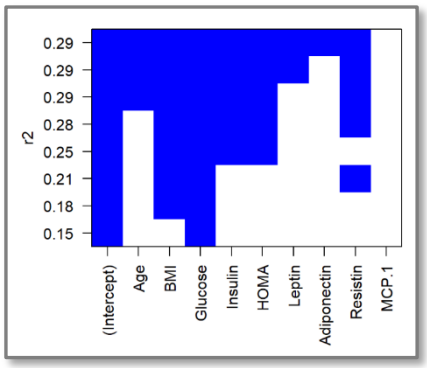

Figure 8: Regsubset for choice of predictors -(Glucose,BMI,Insulin,HOMA and Resistin)

OPTION 1 Predictors: Glucose / BMI / Insulin / HOMA / Resistin

	Overall Error	Sensitivity	Specificity	Type I Error	Type II Error
glm	0.2586207	0.6916667	0.7876493	0.2123507	0.3083330
lda	0.2844828	0.6833333	0.7434763	0.2565237	0.3266667
qda	0.3793103	0.8375000	0.4052410	0.5947590	0.1625000
knn	0.4310345	0.0416667	1.0000000	0.0000000	0.9583333

OPTION 2 Predictors: BMI / HOMA / Resistin

	Overall Error	Sensitivity	Specificity	Type I Error	Type II Error
glm	0.2931034	0.6333333	0.7697921	0.2302079	0.3666667
lda	0.3103448	0.5833333	0.7797988	0.2202012	0.4166667
qda*	0.3879310	0.8666667	0.3804732	0.6195268	0.1333333
knn	0.3965517	0.1083330	1.0000000	0.0000000	0.8916667

OPTION 3 Predictors: Age / Resistin / Leptin / BMI / Glucose / Insulin / HOMA

	Overall Error	Sensitivity	Specificity	Type I Error	Type II Error
glm	0.2931034	0.6750000	0.7387771	0.2612229	0.3250000
lda	0.3189655	0.6291667	0.7437630	0.2565237	0.3708333
qda	0.3275862	0.8458333	0.5076294	0.4923706	0.1541667
knn	0.4827586	0.3041667	0.7267249	0.2732751	0.6958333

OPTION 4 All Predictors

	Overall Error	Sensitivity	Specificity	Type I Error	Type II Error
glm	0.3103448	0.6208333	0.7566342	0.2433658	0.3791667
lda	0.3275862	0.6041667	0.7434763	0.2565237	0.3958333
qda	0.3275862	0.7833333	0.5580495	0.4419505	0.2166667
knn**	0.2327586	0.8208333	0.7240159	0.2759841	0.1791667

Figure 9: Tabulation of all error rates for all models and predictors *: lowest Overall Type II Error, **: lowest Overall Error