Predicting breast cancer from digitized images of breast mass

Tiffany A. Timbers, Melissa Lee, Joel Ostblom & Weilin Han

2023-11-09

Table of contents

Summary 1

Introduction 1

Methods 2

Data 2

Analysis 2

Results & Discussion 3

References 6

Summary

Here we attempt to build a classification model using the k-nearest neighbours algorithm which can use breast cancer tumour image measurements to predict whether a newly discovered breast cancer tumour is benign (i.e., is not harmful and does not require treatment) or malignant (i.e., is harmful and requires treatment intervention). Our final classifier performed well on an unseen test data set, with the F2 score, where beta = 2, of 0.87 and an overall accuracy calculated to be 0.87. On the 171 test data cases, it correctly predicted 157. 9 mistakes were predicting a benign tumour as malignant, while 5 mistakes where predicting a malignant tumour as benign. This is somewhat promising for implementing this in the clinic as false positives are less harmful than false negatives. Although they could theoretically cause the patient to undergo unnecessary treatment if the model is used as a decision tool, it is likely that the model is used for initial screening and that there will be a follow up appointment and further testing until treatment commences. However, the observation of even 4 mistakes predicting a

malignant tumour as benign is concerning. As such, we believe further development of this model is needed for it to have clinical utility. Research to improve the model performance and understand the characteristics of incorrectly predicted patients is recommended.

Introduction

Women have a 12.1% lifetime probability of developing breast cancer, and although cancer treatment has improved over the last 30 years, the projected death rate for women's breast cancer is 22.4 deaths per 100,000 in 2019 (Canadian Cancer Statistics Advisory Committee 2019). Early detection has been shown to improve outcomes (Canadian Cancer Statistics Advisory Committee 2019), and thus methods, assays and technologies that help to improve diagnosis may be beneficial for improving outcomes further.

Here we ask if we can use a machine learning algorithm to predict whether a newly discovered tumour is benign or malignant given tumour image measurements. Answering this question is important because traditional methods for tumour diagnosis are quite subjective and can depend on the diagnosing physicians skill as well as experience (Street, Wolberg, and Mangasarian 1993). Furthermore, benign tumours are not normally dangerous; the cells stay in the same place and the tumour stops growing before it gets very large. By contrast, in malignant tumours, the cells invade the surrounding tissue and spread into nearby organs where they can cause serious damage. Thus, if a machine learning algorithm can accurately and effectively predict whether a newly discovered tumour benign or malignant given tumour image measurements this could lead to less subjective, and more scalable breast cancer tumour diagnosis which could contribute to better patient outcomes.

Methods

Data

The data set used in this project is of digitized breast cancer image features created by Dr. William H. Wolberg, W. Nick Street, and Olvi L. Mangasarian at the University of Wisconsin, Madison (Street, Wolberg, and Mangasarian 1993). It was sourced from the UCI Machine Learning Repository (Street, Wolberg, and Mangasarian 1993) and can be found here, specifically this file. Each row in the data set represents summary statistics from measurements of an image of a tumour sample, including the diagnosis (benign or malignant) and several other measurements (e.g., nucleus texture, perimeter, area, etc.). Diagnosis for each image was conducted by physicians.

Analysis

The k-nearest neighbors (k-nn) algorithm was used to build a classification model to predict whether a tumour mass was benign or malignant (found in the class column of the data set). All variables included in the original data set, with the exception of the standard error of fractal dimension, smoothness, symmetry and texture were used to fit the model. Data was split with 70% being partitioned into the training set and 30% being partitioned into the test set. The hyperparameter K was chosen using 30-fold cross validation with the F2 score as the classification metric. Beta was chosen to be set to 2 for the F2 score to increase the weight on recall during fitting because the application is cancer screening and false negatives are very undesirable in such an application. All variables were standardized just prior to model fitting. The Python programming language (Van Rossum and Drake 2009) and the following Python packages were used to perform the analysis: requests (Reitz 2011), zipfile (Van Rossum and Drake 2009), numpy (Harris et al. 2020), Pandas (McKinney 2010), altair (VanderPlas 2018), scikit-learn (Pedregosa et al. 2011). The code used to perform the analysis and create this report can be found here: https://github.com/ttimbers/breast_cancer_predictor_py.

Results & Discussion

To look at whether each of the predictors might be useful to predict the tumour class, we plotted the distributions of each predictor from the training data set and coloured the distribution by class (benign: blue and malignant: orange, Figure 2). In doing this we see that class distributions for all of the mean and max predictors for all the measurements overlap somewhat, but do show quite a difference in their centres and spreads. This is less so for the standard error (se) predictors. In particular, the standard errors of fractal dimension, smoothness, symmetry and texture look very similar in both the distribution centre and spread. Thus, we choose to omit these from our model.

We also looked to see if there was any multicollinearity between any predictors (defined here as correlations between predictors that are greater than 0.9). When we did this, we observed that many predictors suffered from this (Figure 2). As a consequence we identified 13 additional features that should be dropped: the mean radius, perimeter, concavity and concave points, the maximum radius, perimeter, area, texture, concavity, concave points and compactness, and the standard error of the radius and perimeter.

We chose to use a simple classification model using the k-nearest neighbours algorithm. To find the model that best predicted whether a tumour was benign or malignant, we performed 30-fold cross validation using F2 score (beta = 2) as our metric of model prediction performance to select K (number of nearest neighbours). We observed that the optimal K was 7 (Figure 3).

Our prediction model performed well on test data, with a final overall accuracy of 0.87 and F2 (beta = 2) score of 0.87. Other indicators that our model performed well come from the confusion matrix, where it only made 14 mistakes from the 171 test observations. 9 mistakes

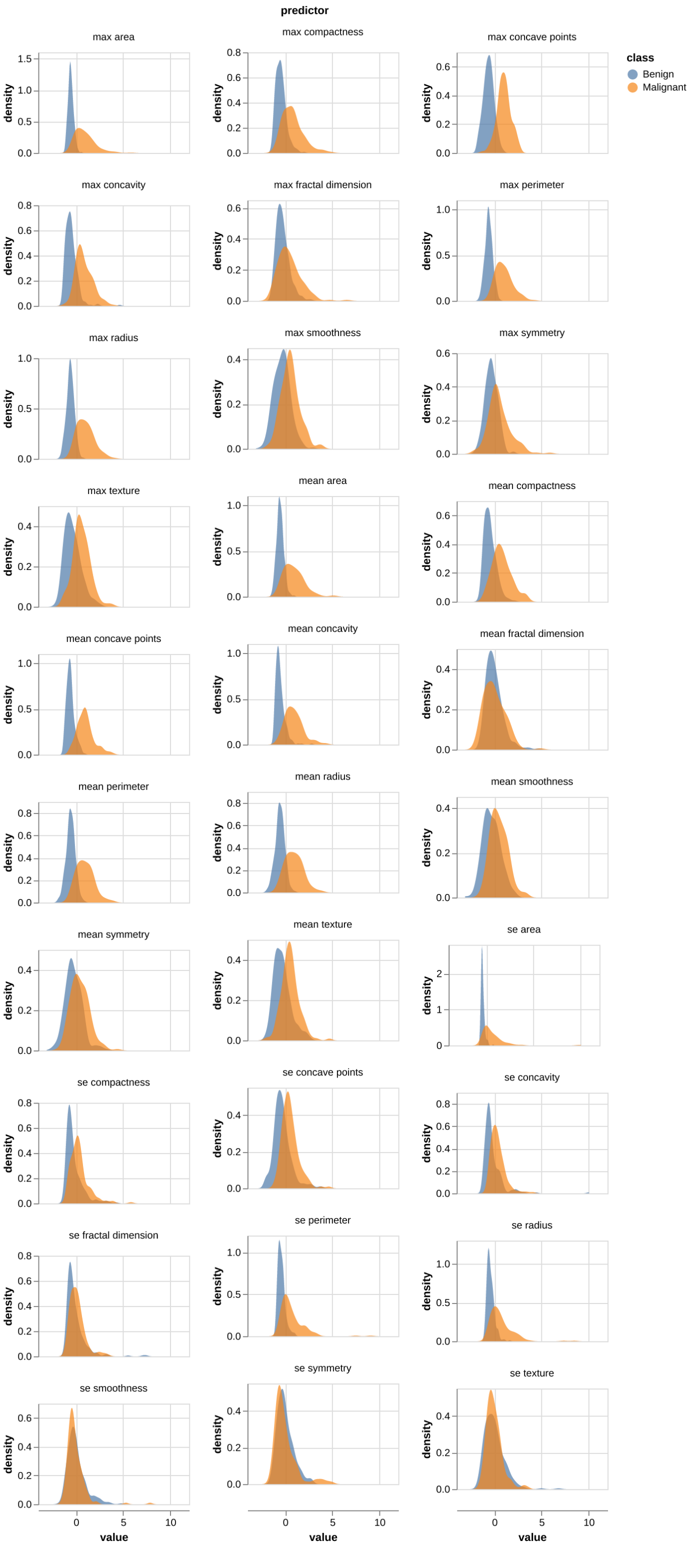

Figure 1: Comparison of the empirical distributions of training data predictors between benign and malignant tumour masses.

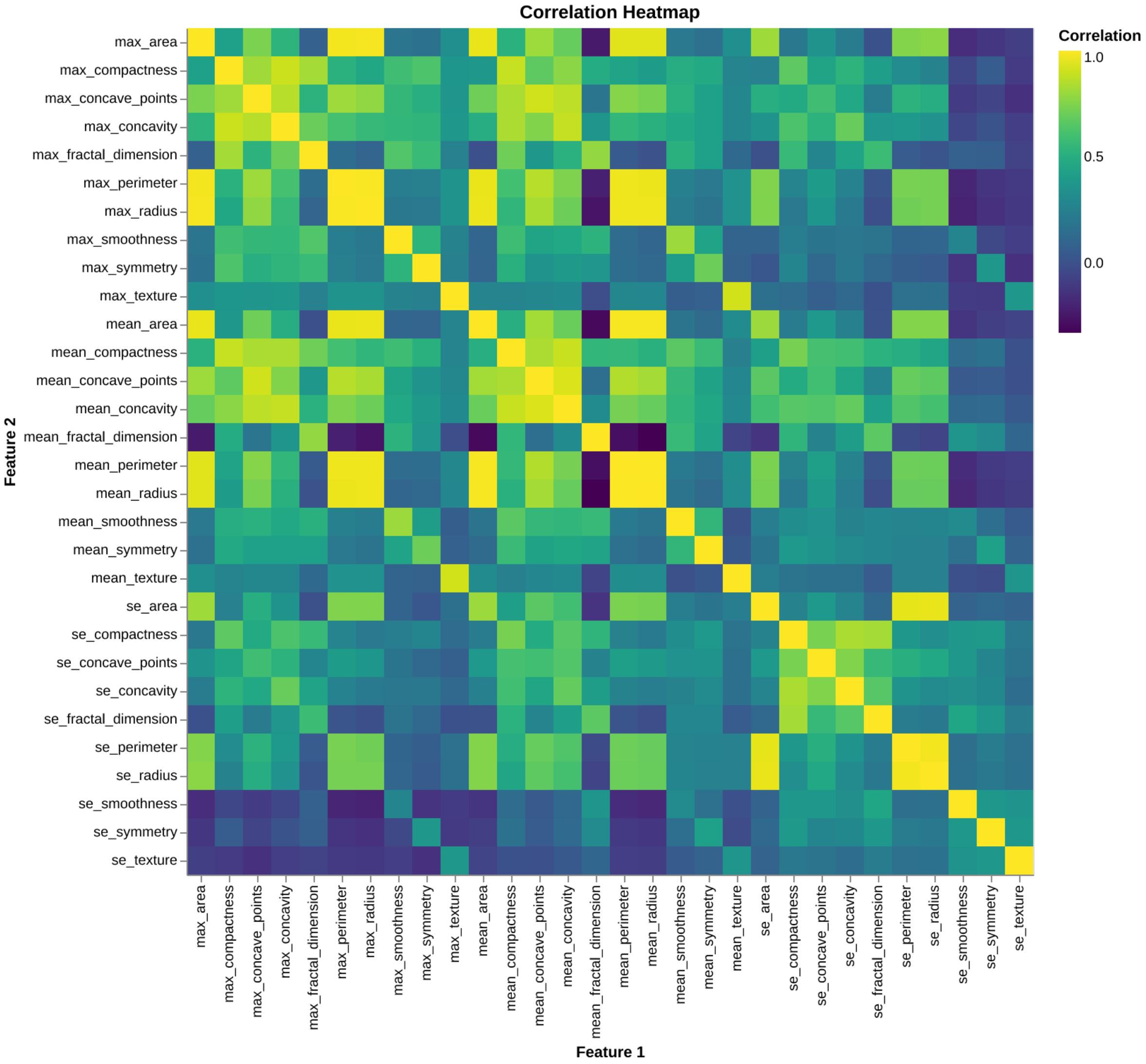

Figure 2: Heatmap of correlations between predictors/features for the breast cancer data set.

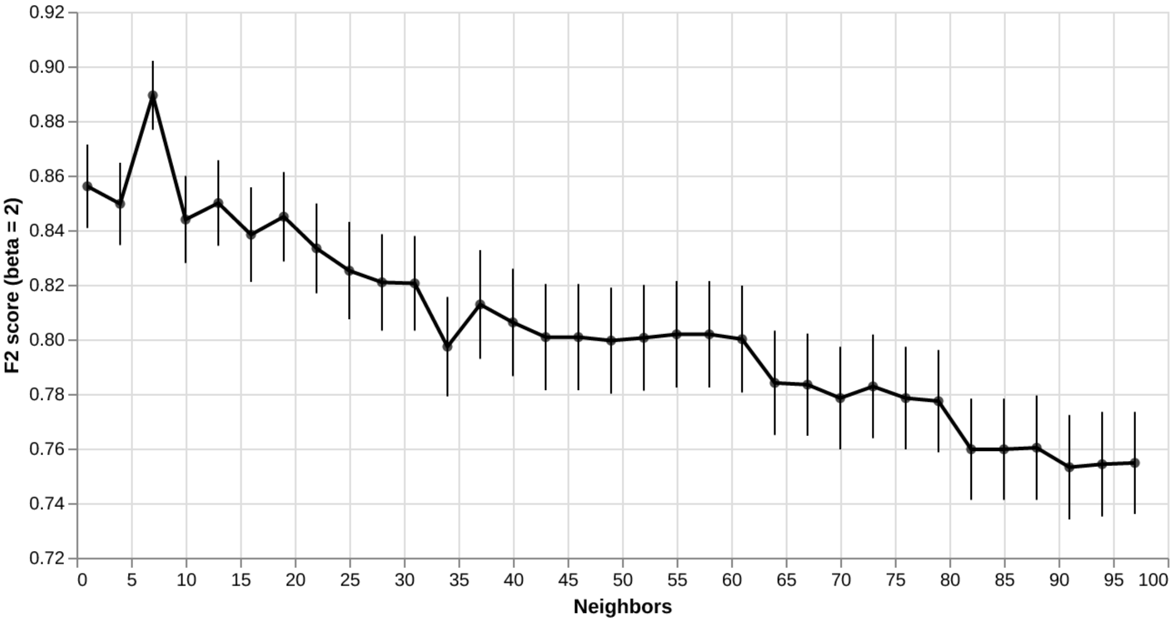

Figure 3: Results from 30-fold cross validation to choose K. F2 score (with beta = 2) was used as the classification metric as K was varied.

were predicting a benign tumour as malignant, while 5 mistakes where predicting a malignant tumour as benign. This is somewhat promising for implementing this in the clinic as false positives are less harmful than false negatives.

Table 1: Confusion matrix of model performance on test data.

Actual label:	Predicted: Benign	Malignant
Benign	102	5
Malignant	9	55

The performance of this model suggests it is not yet ready to be used as a screening tool in a clinical setting, there are several directions that could be explored for to improve it further. First, we could look closely at the 14 misclassified observations and compare them to several observations that were classified correctly (from both classes). The goal of this would be to see which feature(s) may be driving the misclassification and explore whether any feature engineering could be used to help the model better predict on observations that it currently is making mistakes on. Additionally, we would try seeing whether we can get improved predictions using other classifiers. One classifier we might try is random forest forest because it automatically allows for feature interaction, where k-nn does not. Finally, we also might improve the usability of the model in the clinic if we output and report the probability estimates for predictions. If we cannot prevent misclassifications through the approaches suggested above, at least reporting a probability estimates for predictions would allow the clinician to know how confident the model was in its prediction. Thus the clinician may then have the ability to perform additional diagnostic assays if the probability estimates for prediction of a given tumour class is not very high.

References

Canadian Cancer Statistics Advisory Committee. 2019. "Canadian Cancer Statistics." *Canadian Cancer Society*. http://cancer.ca/Canadian-Cancer-Statistics-2019-EN.

Harris, Charles R, K Jarrod Millman, Stéfan J van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, et al. 2020. "Array programming with NumPy." *Nature* 585 (7825): 357–62. https://doi.org/10.1038/s41586-020-2649-2.

McKinney, Wes. 2010. "Data Structures for Statistical Computing in Python." In *Proceedings of the 9th Python in Science Conference*, edited by Stéfan van der Walt and Jarrod Millman, =51–56.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, et al. 2011. "Scikit-learn: Machine Learning in Python." *Journal of Machine Learning Research* 12: 2825–30.

Reitz, Kenneth. 2011. *Requests: HTTP for Humans*. https://requests.readthedocs.io.

Street, W. Nick, W. H. Wolberg, and O. L. Mangasarian. 1993. "Nuclear feature extraction for breast tumor diagnosis." In *Biomedical Image Processing and Biomedical Visualization*, edited by Raj S. Acharya and Dmitry B. Goldgof, 1905:861–70. International Society for Optics; Photonics; SPIE. https://doi.org/10.1117/12.148698.

Van Rossum, Guido, and Fred L. Drake. 2009. *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace.

VanderPlas, Jake. 2018. "Altair: Interactive Statistical Visualizations for Python." *Journal of Open Source Software* 3 (7825, 32): 1057. https://doi.org/10.21105/joss.01057.