Breast Cancer Malignancy Predictions Using Machine Learning

Tim Swarts

19/11/2021

Abstract

Contents

Introduction

Breast cancer is the second leading cause of cancer death in women. It is therefore crucial to determine wheter any given found lump in the breasts is a malignant form of breast cancer, or rather benign. Time is key in finding the right diagnosis, afast diagnosis of breast cancer can drastically improve the five year survival rate, but a rather time consuming part of this is gathering data. Because of this, a machine learning algorithm that can accurately predict breast cancer malignancy with just a few data points could prove vital. Studies show that in stage 1 and stage 2 breast cancer the 5 year survival rate is 89.6-97.9%. But in stage 4 the 5 year survival rate drops to 26.2%. This proves that an early diagnosis is life-crucial. This researchs hopes to help with creating a machine learning solution to diagnosing breast cancer malignancy.

Objective

The goal of this research can be encompassed in two main research questions: Which attributes best predict breast cancer malignancy? Is it possible to create a machine learning model that can predict breast cancer malignancy with an accuracy of 80% or higher? In answering these questions it is import to note that in this case, we would want a very sensitive model, that can pick up as many malginant cases as possible. This model would thus have very little false negatives for malignancy, which might introduce a few extra false possitives. This is a fair trade given that the consequences of leaving a malignant lump untreated are far greater than the consequences of treating someone with a benign form.

Theory

Breast cancer is a type of cancer that starts when cells in the breast begin to grow out of control. These cells usually start a tumor which can be seen on an x-ray or felt as a lump. These lumps are mostly benign, but malignant cases exists which need medical attention.

Materials & Methods

Materials

The breast cancer database used in this research was obtained from the University of Wisconsin Hospitals, Madison from Dr. William H. Wolberg. This data is normalized to fit a grading systems that grades the severity of each attribute. A detailed description of what each grade value means for each attribute can be found in the Breast Cancer Diagnosis Web User Interface. Furthermore, exploratory data analysis was carried out in R version 4.0.3 using the following packages:

- ggplot2 2.2.5

- ggpubr 0.4.0

- farff 1.1.1

- tibble 3.1.5

- reshape2 1.4.4

- ggrepel 0.9.1

- dplyr 1.0.7

Training and testing machine learning algorithms was done by using the Explorer application in Weka Version 3.8.4 from The University of Waikato. Lastly, a java program to run the final model to classify new data was coded in Java version 1.14.

Any cost sensitive classifiers in this report were made with the following cost matrix, unless stated otherwise:

Table 1: Cost Matrix

classified as ->	B	M
B	0	1
M	4	0

The code for the Java application is available at the following github repository:
2021thema9java (link)
The exploratory data analysis and the log, can be found in the github repository below:
thema9 (link)

Methods

The data was read into R. Models were tested with 10 fold cross validation.

Results

Correlation of the attributes

An import research question in this analysis was which attributes would best predict breast cancer malignancy. In order to get insight into this, a correlation heat map was made, as to visualize the respective correlations between each of the attributes present in the data set. The resulting plot is shown in figure 1 below:
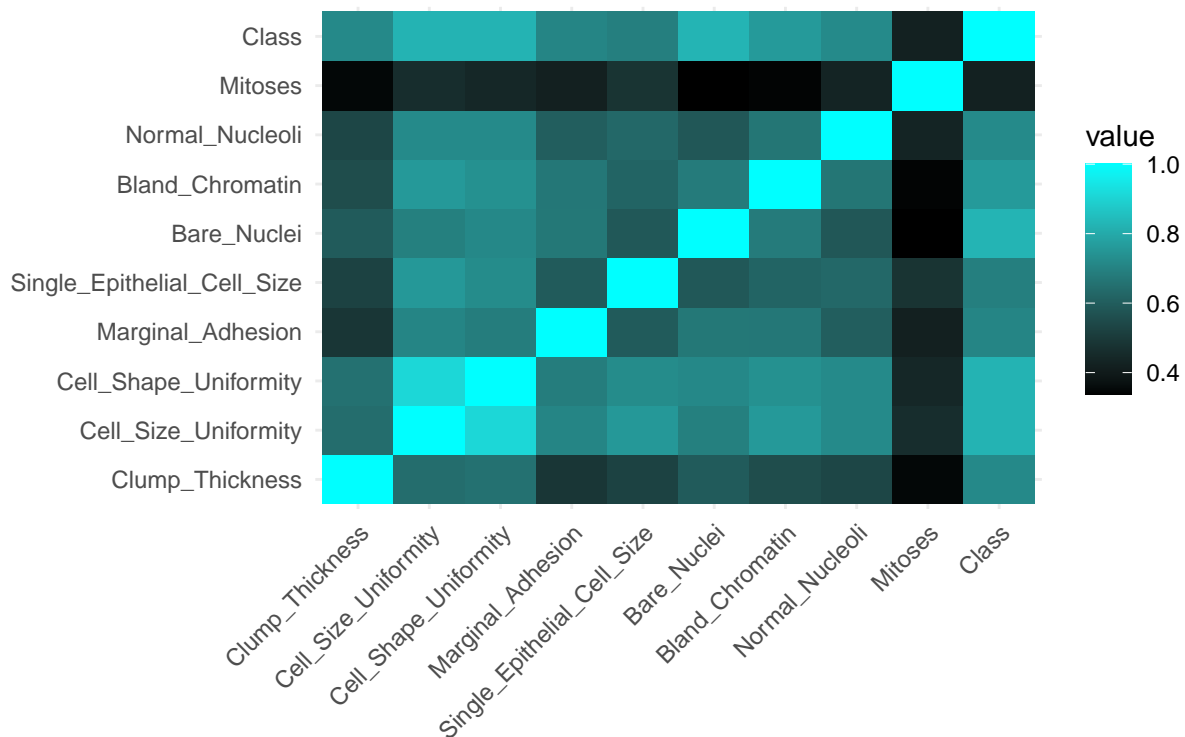

Figure 1 ~ Correlation Heatmap of breast cancer data set.
The lighter the color, the better the correlation between the two columns.

The lightest colors visible in figure 1, besides of course, the diagonal, are those between Cell Shape Uniformity and Cell Size Uniformity. The lightest colors when looking at the class attribute specifically, which is the attribute concerning malignancy, are those between Class & Cell Shape Uniformity, Class & Cell Size Uniformity, and Class and Bare nuclei.

Class Distribution

To get a sense of how the data set was subdivided in the two class attributes, malignant and benign, a pie chart was made that shows the proportions these values respectively. This chart is shown below in figure 2:
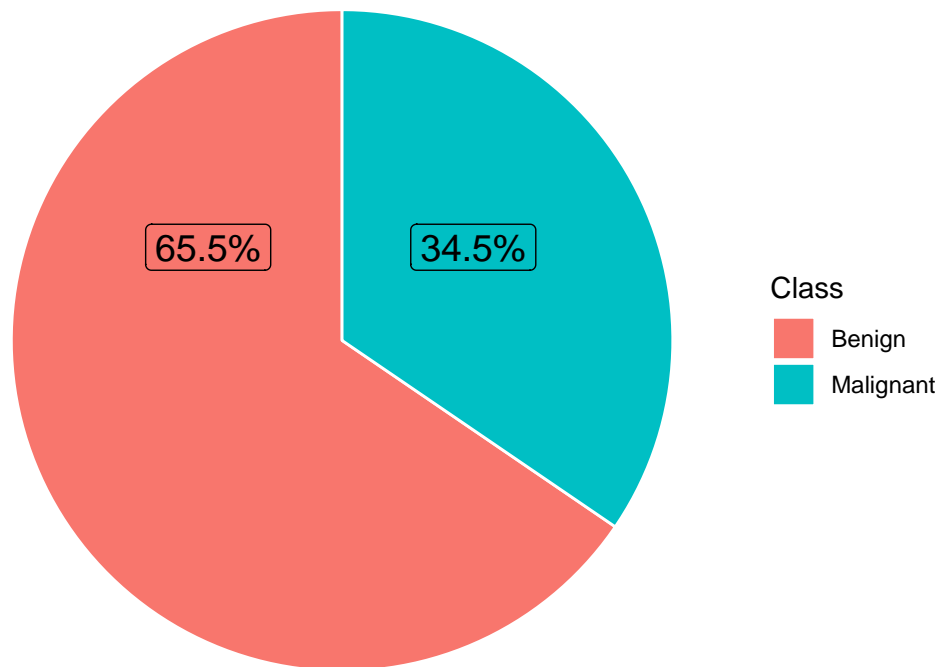
Figure 2 ~ A pie chart of the distrubtion of the classes in the dataset.

Figure 2 shows that roughly two thirds of the data consists of Benign instances. When calculated, we find that only 34.5% of the instances are malignant. This needed to be kept in mind when testing the model, because ZeroR algorithm would in this case already classify 65.5% of the instances correctly by pure chance. Since ZeroR ignores all predictors and simply picks the majority category, a predicting model with an accuracy lower than 65.5% could not be seen as an accurate model.

Relation between Cell Shape Uniformity and Cell Size Uniformity

After seeing the correlations in figure 1, the decision was made to further investigate the relationship between Cell Shape and Size Uniformity and whether or not they would be able to divide the data into the two class labels. A plot that visualizes this relation is depicted in figure 3, shown here:
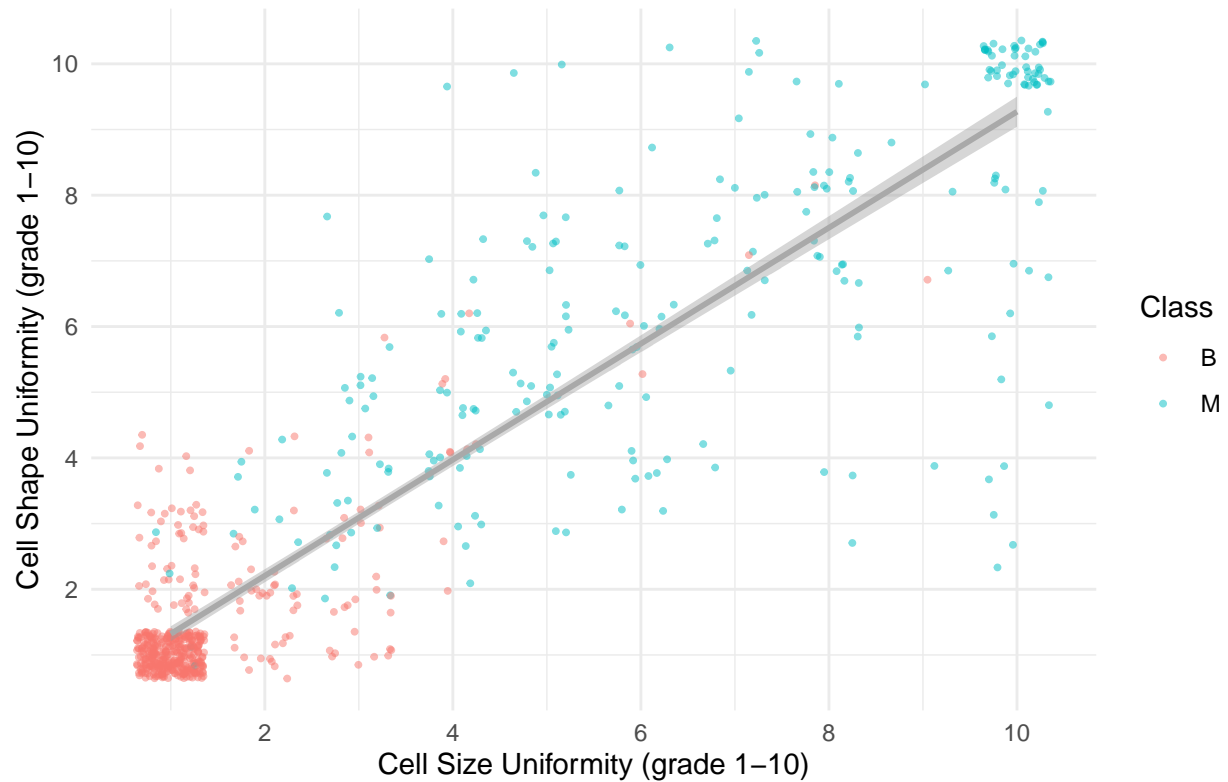
Figure 3 ~ Cell Shape as a function of Cell Size.
The color of the points show their class, red being benign, blue being malignant.

Figure 3 shows the correlation between the attributes Cell_Shape_Uniformity and Cell_Size_Uniformity. When looking at the spread of the two colors in this graph, the benign instances seem to form a group in the bottom left, while the malignant cases are spread across the middle and the top right. The spread noticeably widens near the top right of the plot, showing a bigger spread in values in malignant cases than in benign cases.

Another visualization was made to further illustrate de relation between Cell Shape Uniformity and breast cancer malignancy.
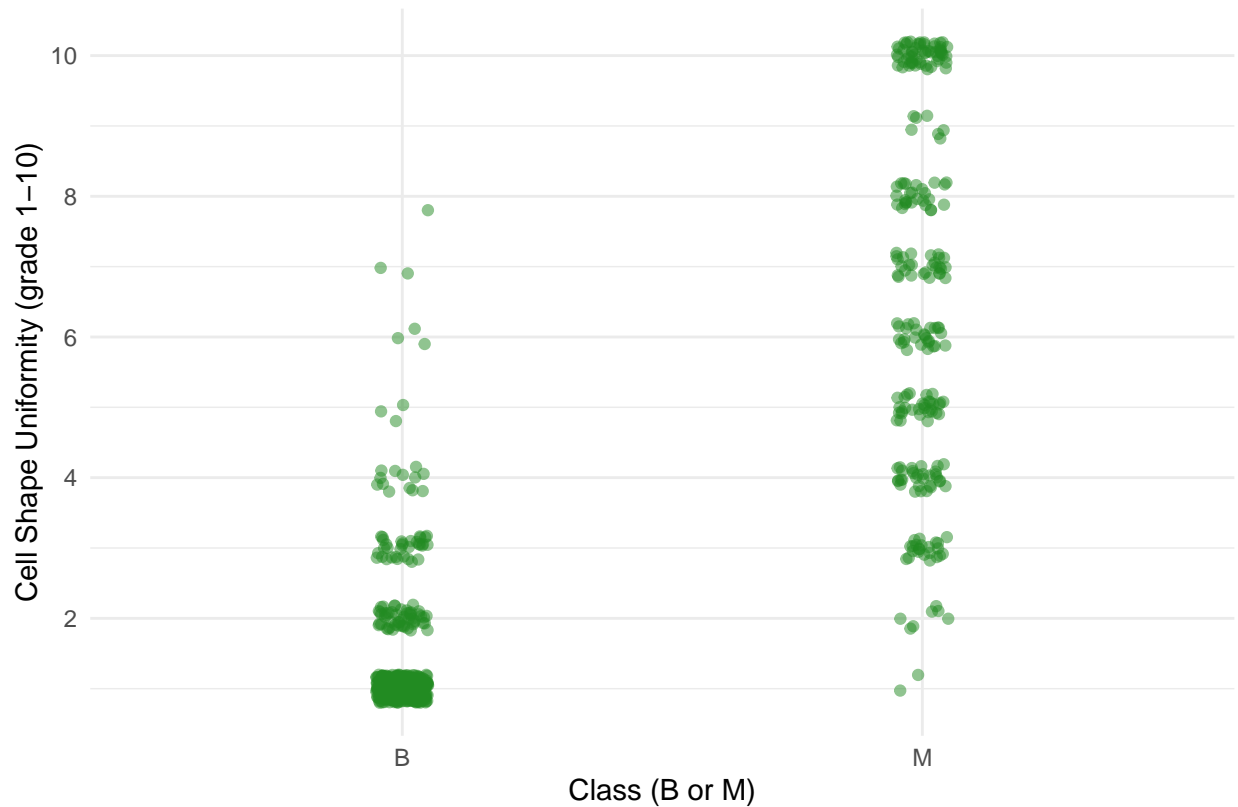
4 ~ Cell Shape as a function of Class. Jitter and transparancy is added to show the denisity of points at each grade.

Figure 4 shows a lot of benign cases at Cell Shape Uniformity grades 1 and 2 and none higher than 8. Malignant cases are spread out, but only 2 points are at grade 1.

Testing Machine Learning Algorithms

Over the span of this reasearch multiple, for the sake of time efficiency, only one is shown here, along with its ROC curve. The model in case is a CostSenstiveClassifier that uses Logistic Regression. This model was trained using only the following attributes as predictors:

- Cell_Size_Uniformity

- Cell_Shape_Uniformity

- Bare_Nuclei

- Bland_Chromatin

This yielded the following results:
Accuracy: 96.77% correctly classified instances
Recall for M: 0.975
Precision for M: 0.936

Table 2: Confusion Matrix Logistic Cost Sensitive Classifier

classified as ->	B	M
B	427	16
M	6	223

As can be seen in the table above, this model only produces 6 false negatives for malignant breast cancer. This is visible in the cell in the column classified as B (Benign) and the row M (actually malignant).

Discussion & Conclusion

The goal of this project was to train a machine learning algorithm to classify breast cancer cases as either benign or malignant as accuratly as possible and to minimize the amount of data needed for said prediction by searching which attributes in the given data set were suited best for the task. This algorithm had to be very sensitive and yiield very little flase negative. Completing this goal was achieved by looking at the correlation between the different predictors and the class attribute (figure 1), and by testing various available algorithms with varying cost matrices. The eventual model had an accuracy of 96.77%, as shown in the results, and yielded only 6 false negatives. This model was then implemented in an executable java application.

Discussion

Looking back at figure 4, something noteworthy is the very clear devide between the two groups (benign and malignant) when spread over Cell Size- and Cell Shape Uniformity. If given the time, we would have liked to explore how accurate one could classify the data with these attributes only, because the plot gives the impression that a single perceptron could fairly adequatly split the data in two. Something to keep in mind however, is that the spread of points in figure 4 noticeably widens near the top right of the plot, showing a bigger spread in values in malignant cases than in benign cases. This makes sense keeping in mind malignant cancers are cells rapidly and uncontrollably deviding, and would thus create alot of diversity in cell uniformity, but this unpredictableness can cause inaccuracy when classifying new data sets using Cell Shape Uniformity and Cell Size Uniformity only.

Conclusion

At the end the most effictive predictors, as can be concluded out of figure 1, were Cell_Size_Uniformity, Cell_Shape_Uniformity, Bare_Nuclei, and Bland_Chromatin. It is possible to classify any given breast lump as either benign or malignant with these attributes alone, with a 96.77% accuracy. This classifcation process has a recall of 0.975. This classifier is however not initended to be used as a sole diagnosing factor, but merely as screening tool that aids doctors in making their diagnosis.

References