Breast Cancer Malignancy Predictions Using Machine Learning

Tim Swarts

19/11/2021

Abstract

Breast cancer is the second leading cause of cancer death in women. Time is key in finding the right diagnosis; A fast diagnosis of breast cancer can drastically improve the five year survival rate, but a rather time consuming part of this is gathering the data needed to diagnose the disease. We used data from the University of Wisconsin Hospitals in Madison to train a machine learning algorithm to predict breast cancer malignancy. In the end the created model could predict whether a case of breast cancer was malignant or benign with a 96.77% accuracy, using only four predicting attributes.

Contents

1 Introduction

Breast cancer is the second leading cause of cancer death in women. It is therefore crucial to determine whether 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; A fast 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 research hopes to help with creating a machine learning solution to diagnosing breast cancer malignancy. [1] [2]

1.1 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 important to note that in this case, we would want a very sensitive model, that can pick up as many malignant cases as possible. This model would thus have very little false negatives for malignancy, which might introduce a few extra false positives. 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.

1.2 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. [2]

2 Materials & Methods

2.1 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: https://github.com/TimSwarts/2021thema9Java The exploratory data analysis and the log, can be found in the github repository below:
https://github.com/TimSwarts/thema9

2.2 Methods

The data was read into R, and turned into a tibble, which allowed for better readability. The provided column names did not clearly explain what any given column was about. With the help of the user manual and the provided .names file these were changed to more comprehensible names. A correlation matrix was made using the cor function after which ggplot2 was used to create a correlation heat map. The class values, which used to be integers, either 2 or 4, were changed to nominal labels 'B' and 'M' respectively. After this the amount of instances per class were calculate and the corresponding percentage showing the distribution of the two classes. This was visualized into a pie chart using ggplot2. Further visualization was carried out by ggplot as well. After this, NA's were omitted and the data was written to a CSV file that was then used

in Weka as training and testing data. Models were tested with 10 fold cross validation. The performance of the different algorithms tried were logged. The best model was exported to a .model file and used for the Java application.

3 Results

3.1 Correlation of the attributes

An important 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. In this heat map the strength of the correlation between two attributes in the data set is represented by a colored cell. A lighter color signifies a stronger correlation between the two attributes, and indicates they might be used to predict one another.

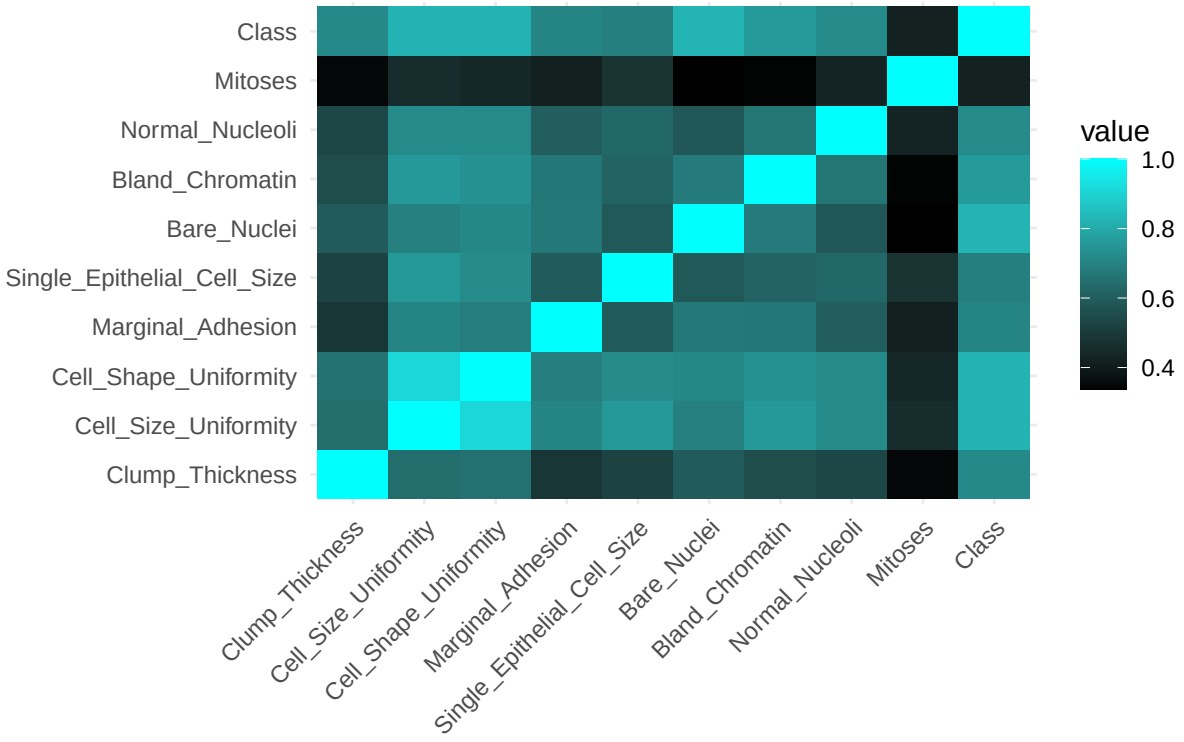


Figure 1 ~ Correlation Heatmap of breast cancer data set.
The lighter the color, the better the correlation between the two columns.

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. As stated before, a lighter color means a higher correlation, these features were thus chosen to be used when training the machine learning model later on.

Another thing to note is that in this plot, the relation between Class and Bare Nuclei stands out as being of use as well. Although its correlation is weaker than that of the aforementioned attributes, it still looks to lie around 0.8, which was a previously set threshold. Upon further analysis, we have found this value to be roughly 0.76. This was deemed high enough and the Bland Chromatin attribute was added to the list attributes used in training the model.

Lastly, the Cell Shape Uniformity and Cell Size Uniformity attributes appear to have a strong correlation among themselves. This too can be useful in predicting the Class, as we will explore later on with figure 3 in section 3.3.

3.2 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. It is important to talk about class distribution when assessing machine learning algorithms, because a skew in the data towards a certain class can influence the accuracy of the model. The pie chart that hopes to give insight into this aspect of the data is shown below in figure 2:

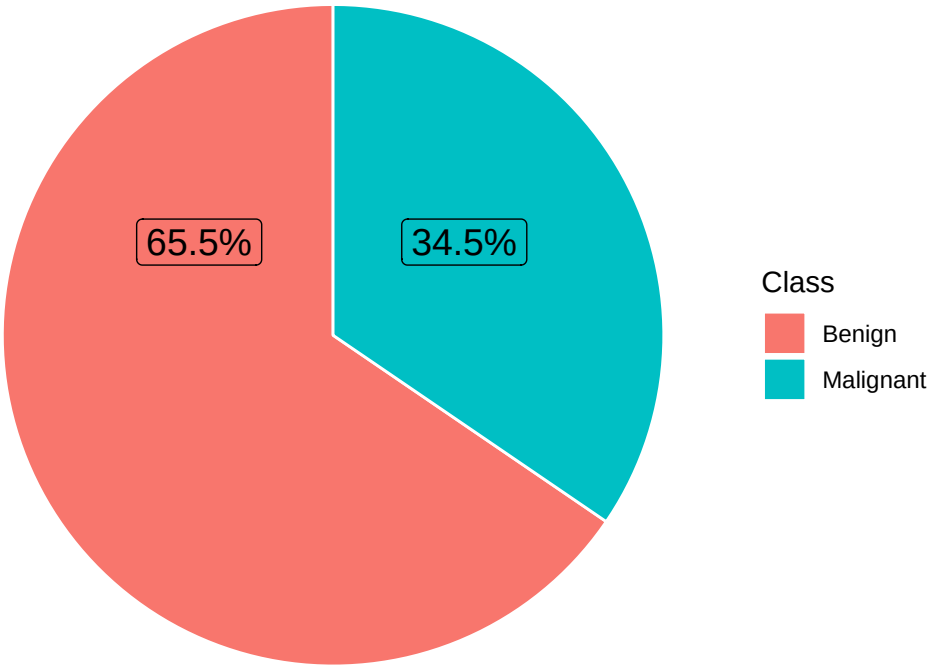


Figure 2 ~ A pie chart of the distrubtion of the classes in the dataset.

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.

3.3 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 below. The points in this plot have been given a *jitter*, meaning they have a small random offset from their actual position. This was necessary to make overlapping points visible.

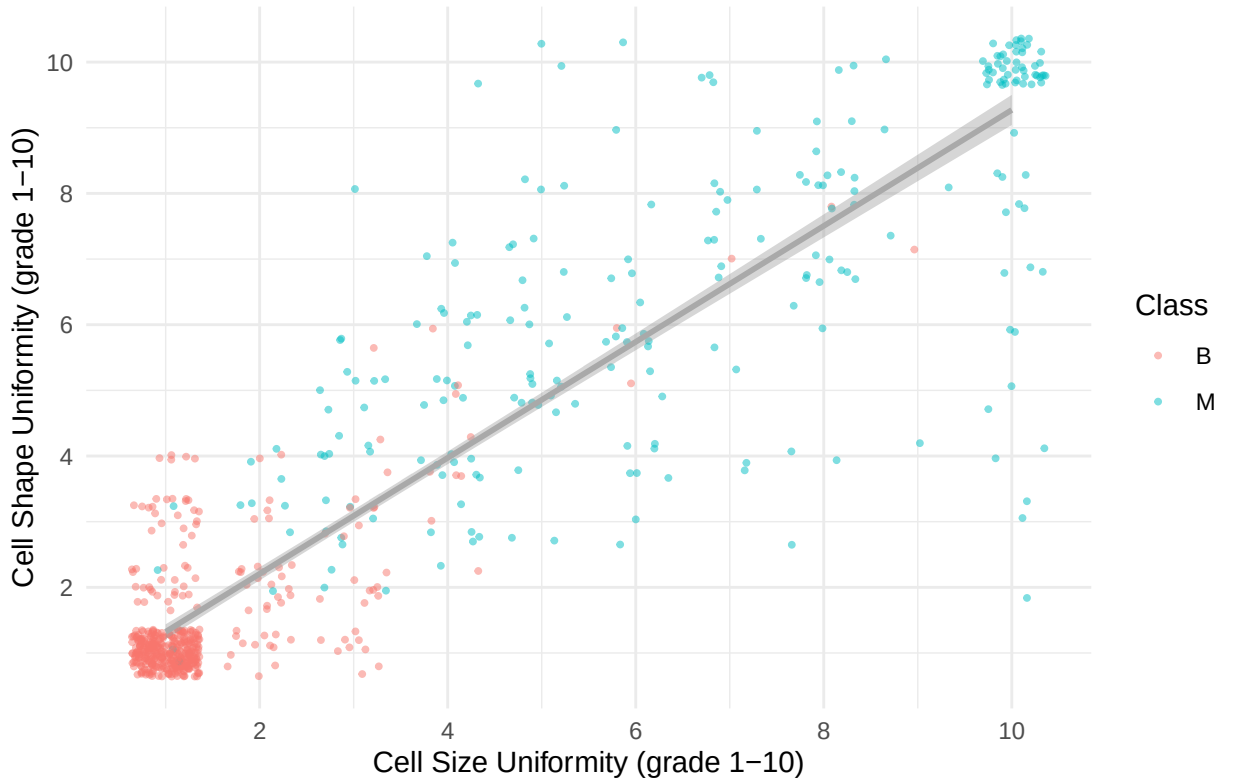
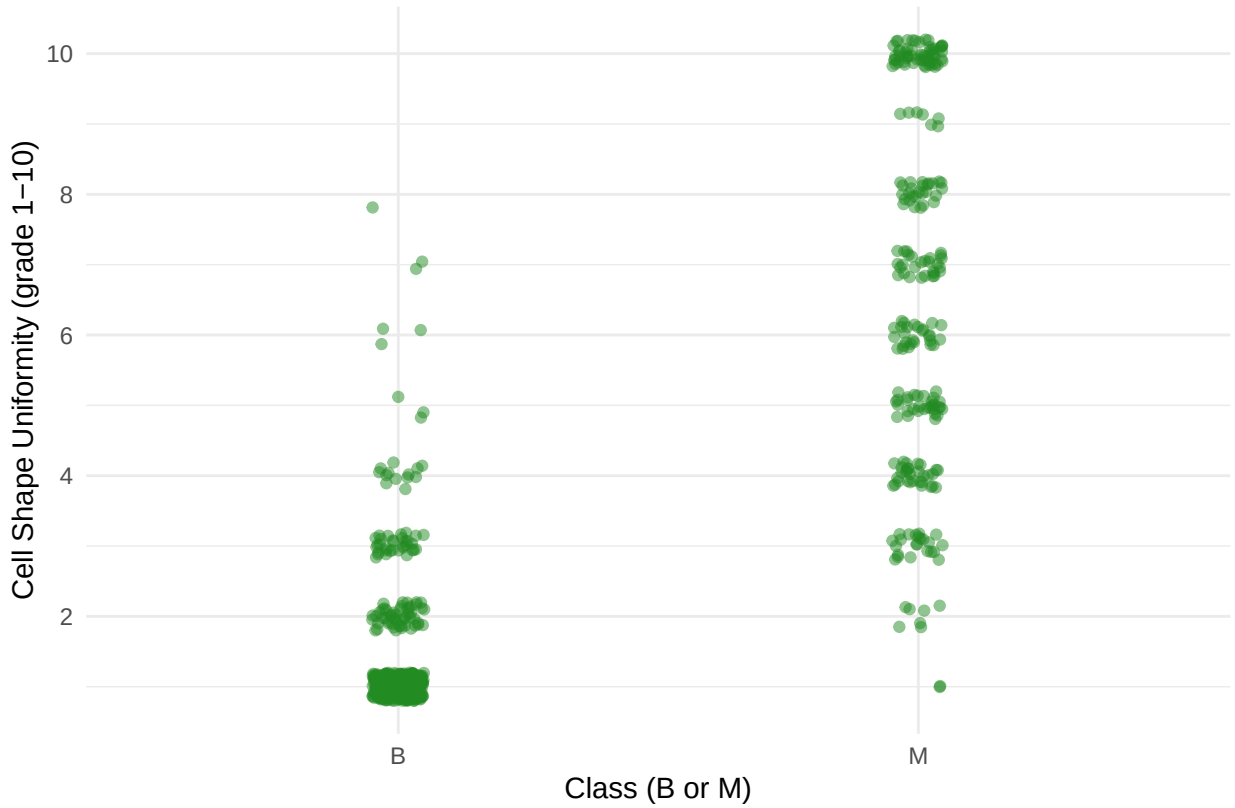


Figure 3 ~ Cell Shape as a function of Cell Size.
The color of the points show their class, red being benign, blue being malignant.

Figure 3 shows the correlation between the attributes Cell_Shape_Uniformity and Cell_Size_Uniformity. Most benign cases seem to rest at uniformity grade of 1, whilst malignant cases are more spread out and tend towards a grade of 10. The linear regression line indicates a strong correlation between the two attributes used in this visualization; Cell Size Uniformity and Cell Shape Uniformity. This is consistent with what we saw in figure 1.

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. The divide between the two classes show multiple outliers, for instance their is a benign case visible at coordinates (9, 7) in the of a field of malignant cases. This suggest these two attributes alone are not sufficient to create an accurate model. For this reason, the other attributes named in section 3.1 and shown in figure 1, were also used.

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.

3.4 Testing Machine Learning Algorithms

Over the span of this research multiple algorithms have been tested to see how they perform on our data set. 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). It also yields 16 false positives (classified as malignant, actually benign), this suggest the model is slightly more sensitive than it is specific. This is expected, because we wanted to minimize false negatives as much as possible in our research. Figure 5, shown below, further illustrates this.

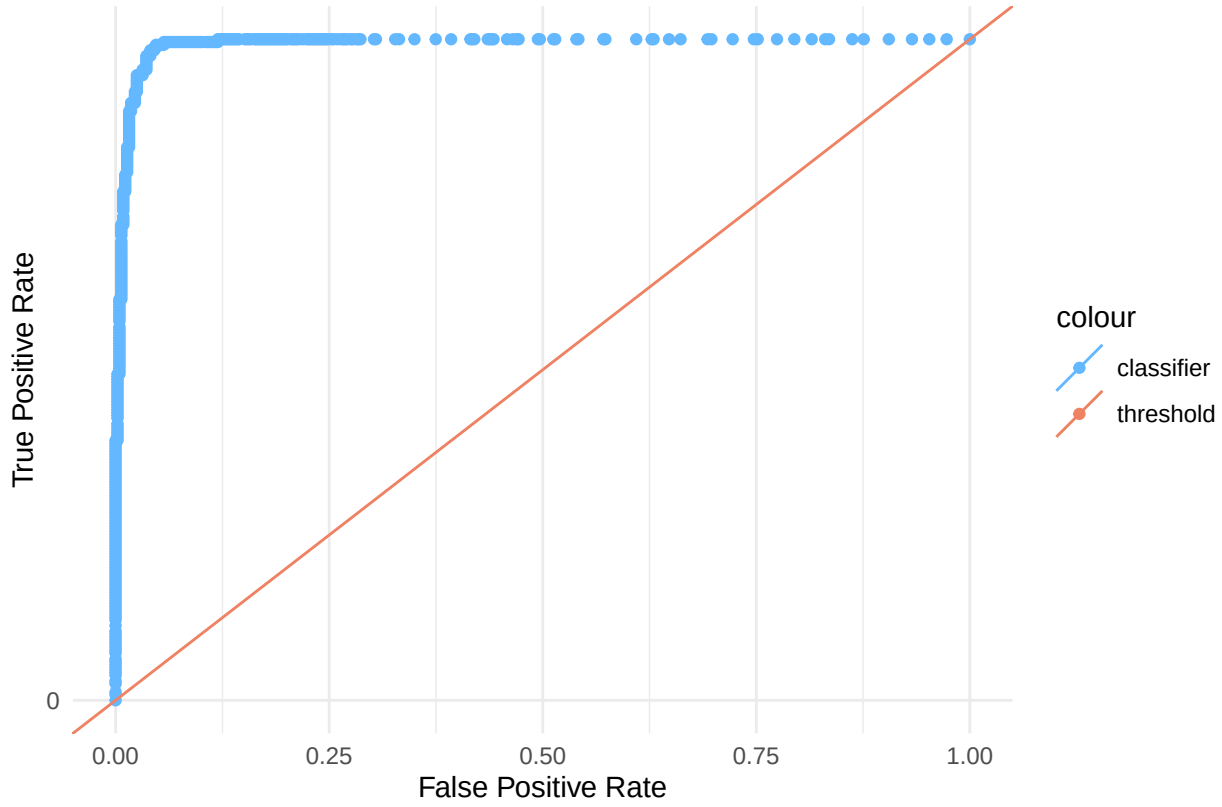


Figure 5 ~ ROC−Curve of model using a cost−sensitive logistic classifier.

The ROC curve in figure 5 shows almost an 90 degree angle in the upper left corner. The AUC of this plot lies at 0.993. The closer this value is to 1, the better the model is at distinguishing between malignant and benign cases. An AUC of 0.993 thus indicates our model is quite capable of making this distinction.

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 accurately 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 yield very little false 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

When we take in account the correlations found in figure 1, along with further data analysis (not shown in this paper, but visible here) it becomes apparent that features related to irregularities in cell growth are correlated most strongly to malignancy. This isn't surprising, seen as abnormal growth is a key characteristic of cancer.

Looking back at figure 4, something noteworthy is the very clear divide 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 adequately 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 dividing, and would thus create a lot 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 effective 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 classification process has a recall of 0.975. This classifier is however not intended to be used as a sole diagnosing factor, but merely as screening tool that aids doctors in making their diagnosis.

References

[1] *Cancer Research UK*. (2021). *Breast Cancer Survival Statistics*. Consulted on November 19th, 2021, from `https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/brea`

[2] *Cancer*. (2021). *About Breast Cancer*. Consulted on November 19th, 2021, from `https://www.cancer.org/cancer/breast-cancer/about/`