Using machine learning to predict the malignancy of breast cancer tumor cells

Rienk Heins

9/14/2021

Goal of the research

The goal of this project is to answer the following question:

Using machine learning, can the malignancy of a breast clump be predicted by the cell and nuclei attributes and which are the best predictors?

To answer this question the data used will be visualized and discussed in the following EDA.

Unpacking the data

The data for this research was obtained from the University of Wisconsin Hospitals, Madison from Dr. William H. Wolberg. The data is a combination from the results of four research studies, namely O. L. Mangasarian and W. H. Wolberg: "Cancer diagnosis via linear programming", William H. Wolberg and O.L. Mangasarian: "Multisurface method of pattern separation for medical diagnosis applied to breast cytology", O. L. Mangasarian, R. Setiono, and W.H. Wolberg: "Pattern recognition via linear programming: Theory and application to medical diagnosis" and K. P. Bennett & O. L. Mangasarian: "Robust linear programming discrimination of two linearly inseparable sets".

First the data will be loaded in and a small portion will be shown for observation.

```r
# Data is loaded from the file
Data <- read.table(file = "breast-cancer-wisconsin.data", sep = ",", na.strings = "?")
# Row names are taken from the names file of the Wisconsin data set and added to the data frame
row_names <- c("ID", "Clump_Thickness", "Uniformity_of_Cell_Size", "Uniformity_of_Cell_Shape", "Marginal
names(Data) <- row_names
# Labels are changed to benign and malignant instead of the 2 and 4 used in the data set for clarity an
Class_numeric <- Data$Class
Data$Class <- factor(Data$Class, labels = c("Benign", "Malignant"))
# Head function used to show a small sample of the data
head(Data, 10)
```

```
##         ID Clump_Thickness Uniformity_of_Cell_Size Uniformity_of_Cell_Shape
## 1  1000025               5                       1                        1
## 2  1002945               5                       4                        4
## 3  1015425               3                       1                        1
## 4  1016277               6                       8                        8
## 5  1017023               4                       1                        1
## 6  1017122              8                      10                      10
## 7  1018099              1                       1                       1
## 8  1018561              2                       1                       2
## 9  1033078              2                       1                       1
## 10 1033078              4                       2                       1
##    Marginal_Adhesion Single_Epithelial_Cell_Size Bare_Nuclei Bland_Chromatin
## 1                  1                           2           1               3
## 2                  5                           7          10               3
## 3                  1                           2           2               3
## 4                  1                           3           4               3
## 5                  3                           2           1               3
## 6                  8                           7          10               9
## 7                  1                           2          10               3
## 8                  1                           2           1               3
## 9                  1                           2           1               1
## 10                 1                           2           1               2
##    Normal_Nucleoli Mitoses     Class
## 1                1       1    Benign
## 2                2       1    Benign
## 3                1       1    Benign
## 4                7       1    Benign
## 5                1       1    Benign
## 6                7       1 Malignant
## 7                1       1    Benign
## 8                1       1    Benign
## 9                1       5    Benign
## 10               1       1    Benign
```

As seen above the data contains the patient id, thickness of the clump, cell information, nuclei data and last the class. The class contains if the cell is benign or malignant. Id gives the id numbers of the patients. Further as seen above the rest of the attributes range from 1-10. This is a score given to the attributes based on percentages, with a low score standing for a more normal situation and a high score meaning the attribute in question is severely different than a normal cell. Clump thickness assesses if cells are mono or multilayered, and if so how much. Uniformity of cell size and shape evaluates the consistency of these features of the cells in the sample. Marginal adhesion quantifies the amount of cells that stick together. Single epithelial cell size measures the enlargement of epithelial cells. Bare nuclei measures the amount of nuclei that are surrounded by cytoplasm and which are not. Bland chromatin rates the texture of the nucleus. Normal nucleoli determines the visibility of the nucleoli and mitosis describes the level of mitotic activity.

Next up the data will be analysed. The first step will be a simple summary to look if statistic measures as the mean and quartiles make sense.

```r
# Summary function used on all attributes appart from the ID
summary(Data[,c(2:10)])
```

```
##  Clump_Thickness  Uniformity_of_Cell_Size Uniformity_of_Cell_Shape
##  Min.   : 1.000   Min.   : 1.000          Min.   : 1.000
##  1st Qu.: 2.000   1st Qu.: 1.000          1st Qu.: 1.000
##  Median : 4.000   Median : 1.000          Median : 1.000
##  Mean   : 4.418   Mean   : 3.134          Mean   : 3.207
##  3rd Qu.: 6.000   3rd Qu.: 5.000          3rd Qu.: 5.000
##  Max.   :10.000   Max.   :10.000          Max.   :10.000
##
```

```
##  Marginal_Adhesion Single_Epithelial_Cell_Size  Bare_Nuclei
##  Min.   : 1.000    Min.   : 1.000              Min.   : 1.000
##  1st Qu.: 1.000    1st Qu.: 2.000              1st Qu.: 1.000
##  Median : 1.000    Median : 2.000              Median : 1.000
##  Mean   : 2.807    Mean   : 3.216              Mean   : 3.545
##  3rd Qu.: 4.000    3rd Qu.: 4.000              3rd Qu.: 6.000
##  Max.   :10.000    Max.   :10.000              Max.   :10.000
##                                                NA's   :16
##  Bland_Chromatin  Normal_Nucleoli     Mitoses
##  Min.   : 1.000   Min.   : 1.000   Min.   : 1.000
##  1st Qu.: 2.000   1st Qu.: 1.000   1st Qu.: 1.000
##  Median : 3.000   Median : 1.000   Median : 1.000
##  Mean   : 3.438   Mean   : 2.867   Mean   : 1.589
##  3rd Qu.: 5.000   3rd Qu.: 4.000   3rd Qu.: 1.000
##  Max.   :10.000   Max.   :10.000   Max.   :10.000
##
```

As seen from the summary most data seems to make sense, they are all as expected having a min of 1 and a max of 10 as the data is graded from 1 to 10 with most data having a mean between the 1st and 3rd quartile with most data points sitting on the lower end of the grading scale, indicating that the attribute has mutated, but not by a lot. The only attribute which brakes this trend is mitosis, which has a 1st quartile, median and 3rd quartile of 1, with a mean of above 1, suggesting that a very large amount of the data has a grade of 1. This might indicate that it wouldn't be the best predictor for the cancer cell outcome. Furthermore Bare Nuclei is the only attribute with missing values, which will have to be removed for the machine learning steps, but with only 16 NA's in a dataset of 700 patients this still leaves a lot of data to use.

Next up a the correlation function of R will be used to see which attributes correlate the most with the class attribute. This will give a first impression of which attributes might be good predictors of the malignancy of a tumor cell.

```
# Correlation function is used
correlation <- cor(x = Class_numeric, y = Data[,c(2:10)], use = "complete.obs")
# Correlation data is set into a data frame
cor_data <- as.data.frame(correlation)
# Data is sorted for better visualization
sort(cor_data, decreasing = T)
```

```
##   Bare_Nuclei Uniformity_of_Cell_Shape Uniformity_of_Cell_Size Bland_Chromatin
## 1   0.8226959                0.8218909               0.8208014       0.7582276
##   Normal_Nucleoli Clump_Thickness Marginal_Adhesion Single_Epithelial_Cell_Size
## 1       0.7186772       0.7147899         0.7062941                   0.6909582
##     Mitoses
## 1 0.4234479
```

As seen above Bare nuclei and uniformity of the cell seem to be the best predictors, all scoring higher than a 0.8 in terms of correlation. Furthermore almost all the attributes seem to correlate decently well with the exception of mitoses, as the other attributes score higher or are around a 0.7.

Data visualization

In the next part of this EDA the data will be visualized using plots made by using the ggplot library of Rstudio.

To start the visualization process a multiple box plot will be made containing all the attributes to get a first view of their ability to classify a tumor cell.

```r
# Data is changed to long data to be used in the multiple box plot
long_data <- pivot_longer(data = Data[,c(2:11)], !Class, names_to = "Attribute", values_to = "Count")
# Multiple boxplot is created using ggplot
ggplot(long_data, aes(y = Count, x = Attribute, fill = Class)) +
  geom_boxplot(width = 0.5, outlier.size = 0.5) +
  ggtitle("Multiple boxplot of all attributes") +
  xlab("Attributes") +
  ylab("Grades(1-10)") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

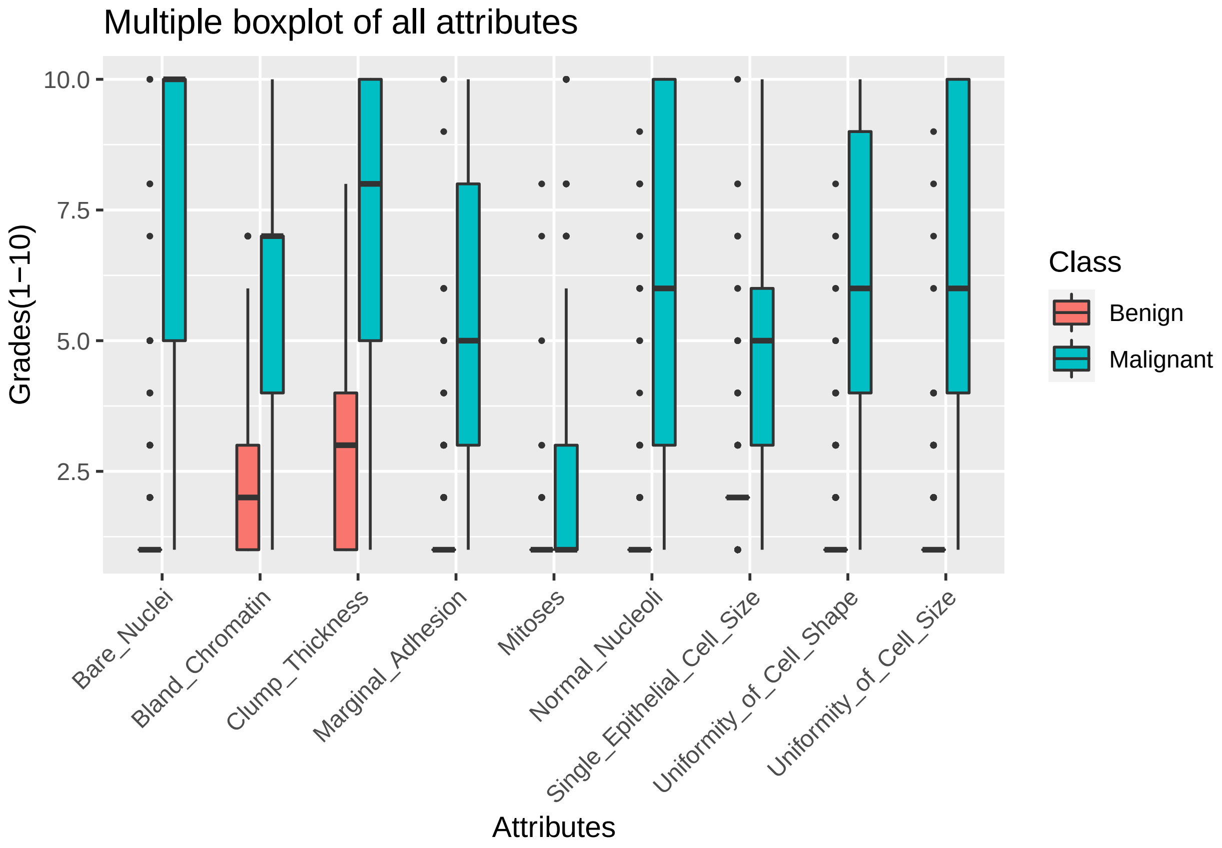

Figure 1: Box plot containg all the attributes from the Wisconsin breast cancer dataset and their spread in benign and malignant cancer cells

As seen in the figure above most of the boxes don't match at any values with the exception of mitoses. This is positive as this means that there are not a lot of scores which could be either benign or malignant so each score should give a decent prediction of what cancer type the cell is. Mostly bare nuclei, uniformity of cell size and shape, marginal adhesion and normal nucleoli look promising as all these attributes have most their benign cases on a value of 1 with some outliers and all malignant cases on higher scores. Suggesting that cell changes in these attributes may cause a cancer cell to become malignant.

With figure 1 showing that mitoses is the only attribute that has boxes on the same value of 1, combined with the fact that it has a weird summary and low correlation it has to be concluded that it most likely isn't a good predictor and thus will not be used in the machine learning steps to come.

Next up some of the more interesting attributes will be plotted separately to get a better view of their relationship with the cancer cell class attribute.

The first attributes that will be plotted separately are the uniformity of cell size and shape as both show how consistent the cells are with their uniformity and also both show high correlation with the malignancy of the cell.

```r
# Jitter plot if made using ggplot
ggplot(data = Data, aes(x = Uniformity_of_Cell_Size, y = Uniformity_of_Cell_Shape, colour = Class)) +
  geom_jitter(width = 0.2, height = 0.2, alpha = 0.3, size = 0.8) +
  geom_smooth(method = "lm", SE = F) +
  xlab("Uniformity of the cell size(graded)") +
  ylab("Uniformity of the cell shape(graded)") +
  ggtitle("Effects of uniformity of cell size and shape on state of breast cancer cells")
```

```
## `geom_smooth()` using formula 'y ~ x'
```

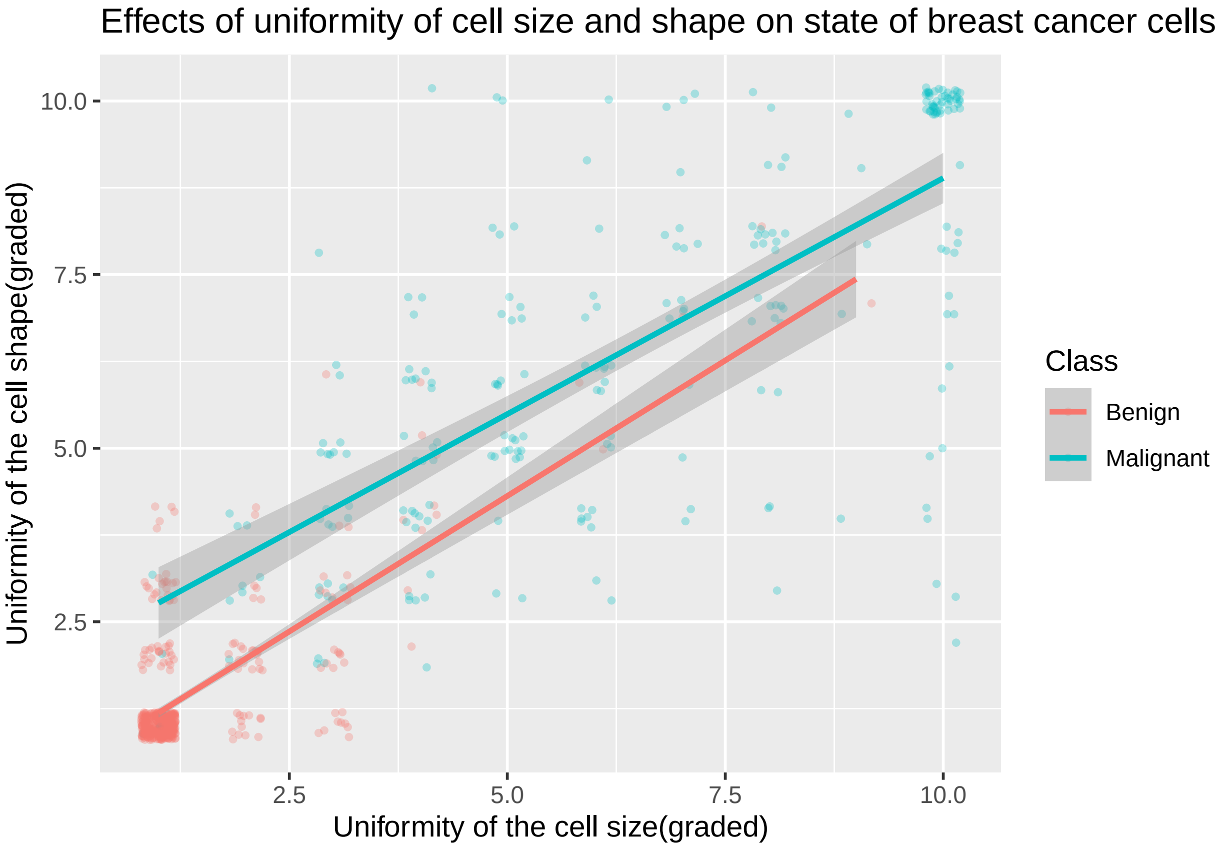

Figure 2: Shows the uniformity of the cell size vs the uniformity of the cell shape and their effects on the malignancy of the cell

As seen in figure 2 there is an overall trend in the rising of both the uniformity ratings causing the cells to be more malignant. This might indicate that the higher scores on these attributes, the higher the chance that the cell might be malignant. On the other hand a lower score has more benign cases, but still contains some malignant cases. This suggests that a higher score can be used to say a cell might be malignant, yet a lower score might not directly mean that a cell is benign, unless the score is as low as 1 or 2 where most benign cases reside.

For the next plot bare nuclei will be plotted against normal nucleoli as bare nuclei is the highest scoring attribute in the correlation matrix. While the normal nucleoli also scored well here, it is also part of the nucleus and thus might have some relation with the bare nuclei score, thus helping for the prediction.

```r
# Jitter plot is created using ggplot
ggplot(data = Data, aes(x = Bare_Nuclei, y = Normal_Nucleoli, colour = Class)) +
  geom_jitter(width = 0.2, height = 0.2, alpha = 0.5, size = 0.8) +
  geom_smooth(method = "lm", SE = F) +
  ggtitle("Effects of nuclei transfomrations on breast tumor cell malignancy") +
  xlab("Nuclei surrounder by cytoplasm(graded)") +
  ylab("Visibility of nucleoli(graded)")
```

```
## `geom_smooth()` using formula 'y ~ x'
```

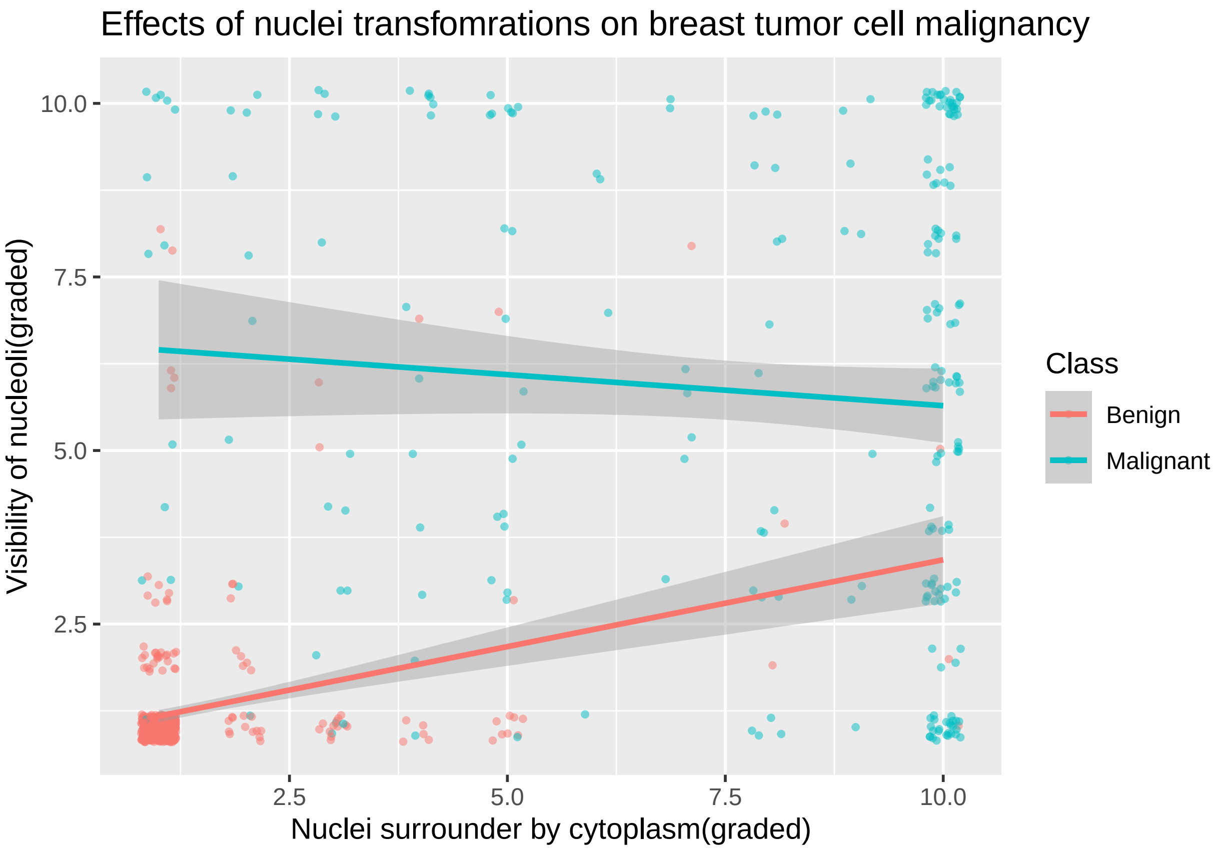

Figure 3: The effects of nuclei changes on the malignancy of a breast cancer cell

A thing to note about figure 3 is that most of the malignant cases are at a 10 for bare nuclei, while there are still quite some benign cases to as high as 5. Meaning that even with a higher score on the attribute a cancer cell doesn't have to be malignant, yet as expected a very high score does. Yet both attributes have the problem that even on a lower score the cancer cell can be malignant. Given this mostly happens when the other attribute is on a higher score but this does mean that both possibly can't be used as a good indicator on their own.

The last attribute that will be plotted is the bland chromatin, as this is the last attribute concerning the cell nucleus. The attribute will be plotted in a box plot.

```r
# Box plot is created by ggplot
ggplot(data = Data, aes(x = Class, y = Bland_Chromatin, fill = Class)) +
  geom_boxplot() +
  ggtitle("Effects of nucleus texture of breast cancer cells") +
  xlab("Class of breast cancer cell") +
  ylab("Nucleus texture(graded)")
```

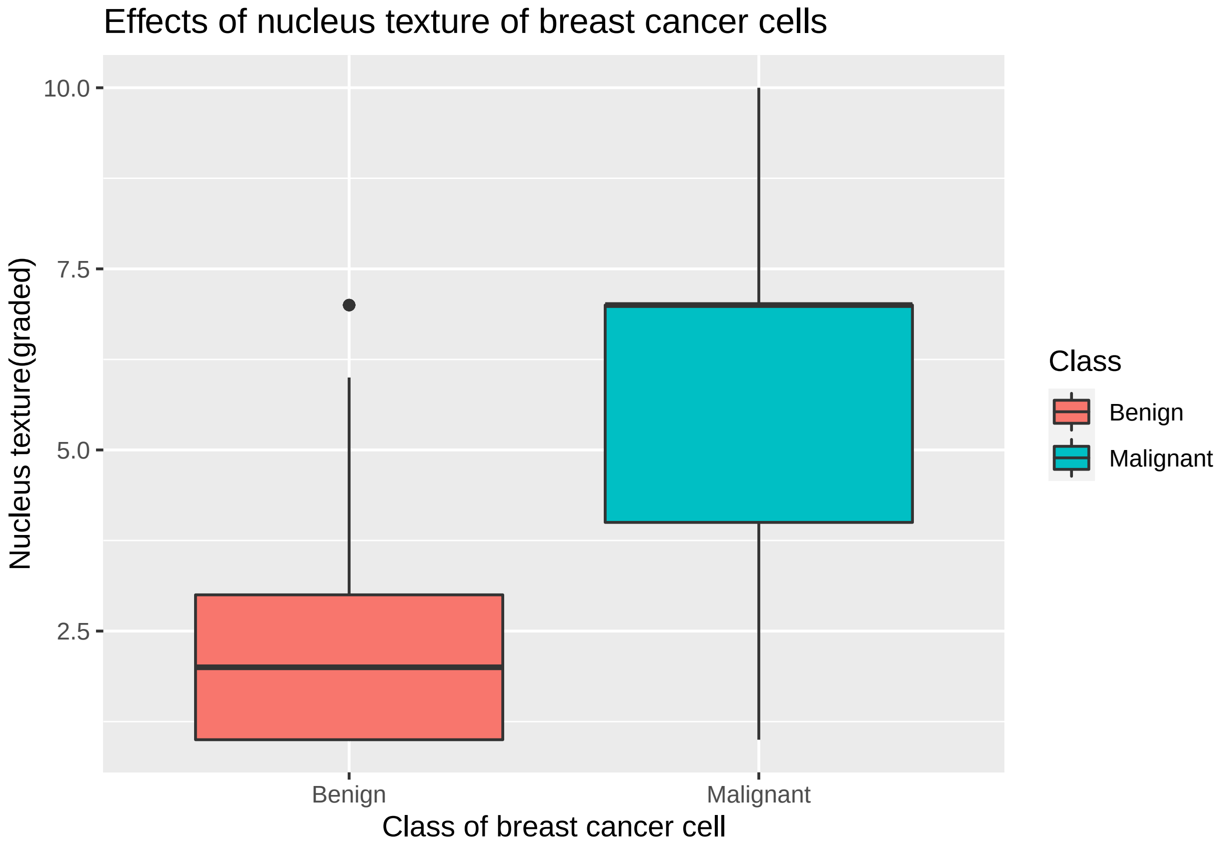

Figure 4: Effects of the bland chromatin attribute describing the nucleus texture on breast cancer cell malignancy

Figure 4 shows that just as seen in figure 1, the boxes don't match with is a positive. Yet the distance in the high end of the benign box and the start of the malignant box isn't very large. Which might lead to scores around the 3 and 4 which are the boxes higher and lower ends having to need more data from other attributes to give an accurate prediction. Also malignant cases have outliers to as low as 1, also suggesting more data is needed. A good thing to note is that benign cases have outliers to only a score of 7, meaning that any case above this score has a very high chance of being malignant.