Breast Cancer Wisconsin

Rose Hazenberg

2021-10-05

Contents

Table 1: Codebook with an overview of the data.

Column.Name	Full.Name	Data.Type	Value.Range
ID	Sample code number	int	NA
clump_thickness	Clump Thickness	int	1 - 10
uniformity_of_cell_size	Uniformity of Cell Size	int	1 - 10
uniformity_of_cell_shape	Uniformity of Cell Shape	int	1 - 10
marginal_adhesion	Marginal Adhesion	int	1 - 10
single_epithelial_cell_size	Single Epithelial Cell Size	int	1 - 10
bare_nuclei	Bare Nuclei	int	1 - 10
bland_chromatin	Bland Chromatin	int	1 - 10
normal_nucleoli	Normal Nucleoli	int	1 - 10
mitoses	Mitoses	int	1 - 10
class	Class	factor	Benign, Malignant

1 Introduction

Despite a great deal of public awareness and scientific research, breast cancer continues to be the most common cancer and the second largest cause of cancer deaths among women. Approximately 12% of U.S. women will be diagnosed with breast cancer, and 3.5% will die of it. The research in clinical practice depends on the analysis of cellular images which is accomplished with a graphical computer program called **Xcyt**, written by one of the authors.[1]

First, a sample of fluid is taken from the patient's breast. The procedure involves using a small-gauge needle to take fluid, known as fine needle aspirate (FNA), directly from a breast lump or mass. An image from the FNA is used for the program Xcyt.[1] This digitized image is used to asses whether a lump in a breast could be malignant (cancerous) or benign (non-cancerous).[2]

1.1 Research question

The research question is: Is it possible to reliably predict the cancer stage based on the uniformity of cell size/shape using machine learning?

1.1.1 Data

The dataset that is used in this paper contains information about breast cancer in Wisconsin and is online available.[3] The data was created by Dr. William H. Wolberg a physician at the University of Wisconsin Hospitals.

The samples are clinical cases of Dr. Wolberg which are periodically classified and are ordered in a chronological grouping. Which is as followed:

Group 1: 367 instances (January 1989)
Group 2: 70 instances (October 1989)
Group 3: 31 instances (February 1990)
Group 4: 17 instances (April 1990)
Group 5: 48 instances (August 1990)
Group 6: 49 instances (Updated January 1991)
Group 7: 31 instances (June 1991)
Group 8: 86 instances (November 1991)

Total: 699 points (as of the donated database on 15 July 1992)[3]

Table 1 shows the codebook of the data which provides an overview of the data frame and the variables.

2 Results

2.0.1 Density plot

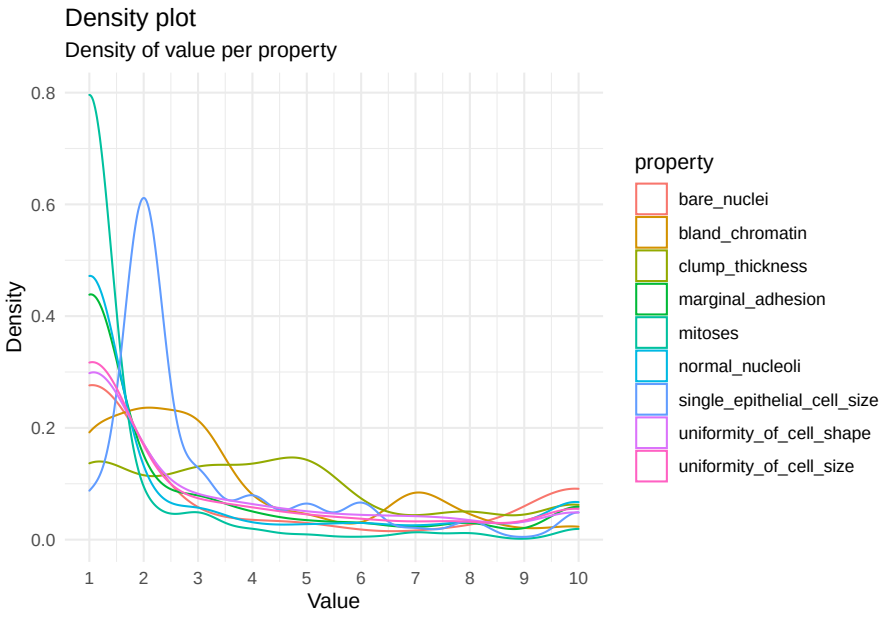

Figure 1: Density plot with the probability density of the value per property.

Figure 1 shows the density plot which shows the density of the values per property. The density plot shows the shape of the distribution of the values per property. The property mitoses shows by the value 1 a density of 0.8 which is the highest of them all. So mitoses mainly consist of the value 1. From value 1 it decreases and has then the lowest density. Single epithelial cell size is the only property with a curve of this distribution. The other properties have a density that runs in waves. The variation in density between the properties lies mainly between the value 1 to 4 because here is the density higher and the density decreases from value 1 until around value 4. After value 4 the lines of the properties run almost around the same density. Thus at higher values, the values approximately occur equal per property.

2.0.2 Boxplot

The distribution of the values per property can be seen in a boxplot 2 which is divided into the two cancer classes. In figure 2 can be seen that for every property of the class malignant a boxplot is created but for the class benign are there only two boxplots created, the non-visible boxplots consist of just a vertical line mainly at value 1 and a lot of outliers. The boxplots of class benign that consist of the line and outliers are all the properties except the properties clump thickness and bland chromatin. This is because benign consists mainly of the value 1 and the other values are thus outliers. And the properties clump thickness and bland chromatin have more variation in their values and consist thus of higher values. The boxplots of the class malignant lie mainly to the right, are longer and don't have any outliers except the property mitosis. Just like in figure 1 shows the boxplot of mitoses in 2 that mitoses mainly consist of the value 1 until 3 so that is why this boxplot is more to the left, shorter, and has outliers for the class malignant. For benign shows, both figure 1 and 2 that there is only a vertical line in the boxplot at value 1 and a high density also at value 1 which thus mainly consist of the value 1.

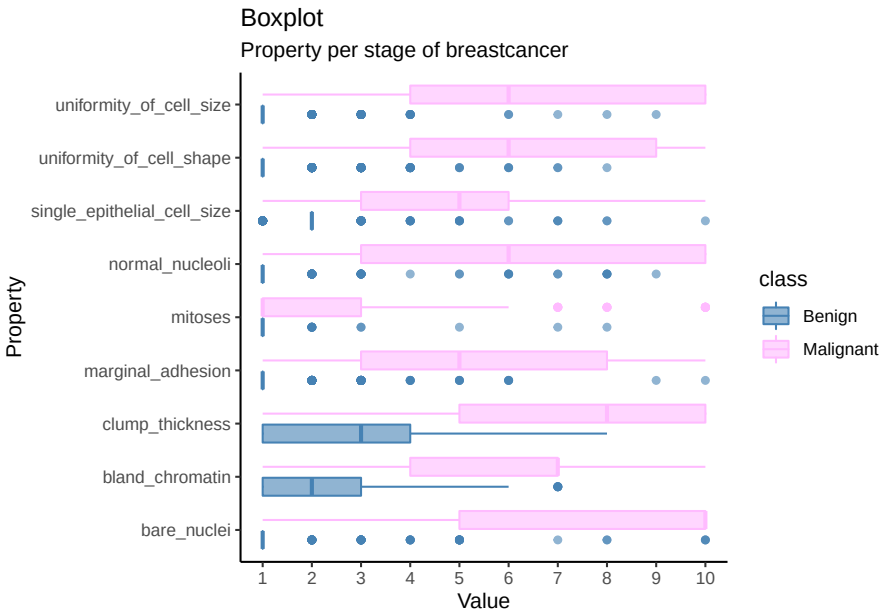

Figure 2: Boxplot showing the distribution of all the properties per class.

Table 2: Total number of the class

Class	Total
Benign	458
Malignant	241

2.0.3 Class table

Table 2 shows the total number of benign and malignant. So here can be seen that there are 458 cases of benign and 241 cases of malignant. But in figure 2 and 3 it seems like there is more malignant, this is because the class malignant consist of higher values, and there is more variation in the values. The class benign consist of lower values mainly the value 1 so there is less variation in between the values.

2.0.4 Relationship

Figure 3: Relationship between the properties and their values for both the classes of breast cancer.

Figure 3 shows the relationship between the properties and their values for both classes. All of the properties have a variation for all values except mitoses here the values are very low. Here is it easy to see that benign consist of mainly 1 values because at value 1 there are a lot of dots in one group for every property except single epithelial cell size there lies the group at value 2. Most of the values for benign lies from 1 until 5. For malignant are the dots scattered over the value range. But there are also groups at value 10 for most of the properties. So here can be seen that malignant mainly consist of value 10. The variation and the distribution of both classes for the properties uniformity of cell size and uniformity of cell shape look alike so these can both be useful to predict the cancer stage.

2.0.5 Heatmap

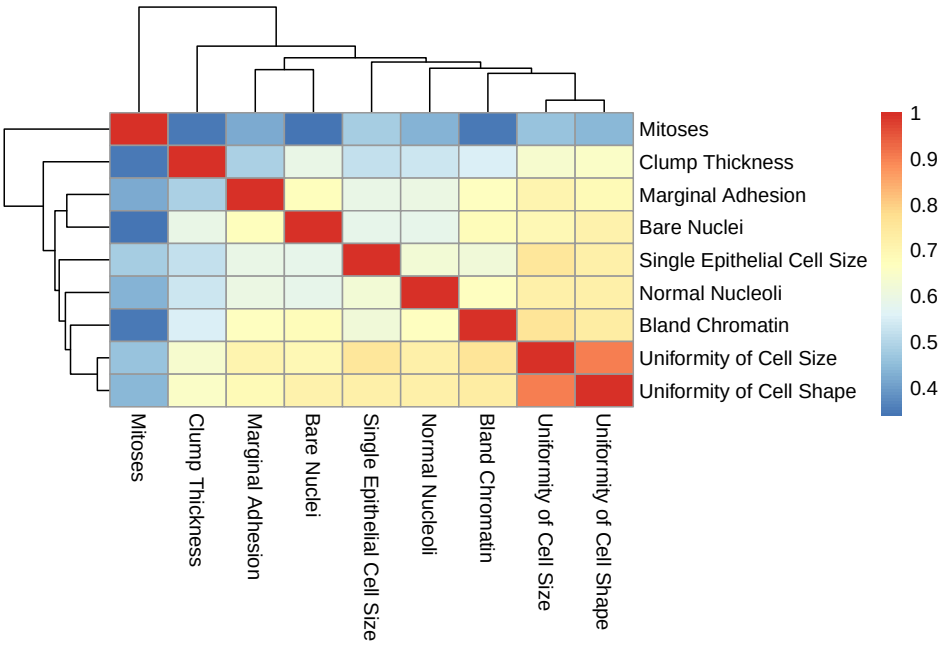

Figure 4: Heatmap with the correlation matrix of the properties.

Figure 4 shows the correlation of the properties. Red shows that the properties are highly correlated and blue shows a low correlation. Here in figure 4 are the properties highly correlated with themselves because they contain the same values. The heatmap shows a patron in the middle because this contains a correlation value between 0.5 and 1 for every property. Only the outside of the heatmap shows a lower correlation for the property mitoses compared with the other properties. Here is the correlation value between 0 and 0.5 and is colored blue. There are two properties with a higher correlation for each other are Uniformity of cell size and Uniformity of cell shape. The blocks of these properties are slightly lighter than the dark ones. So this shows that these properties have a high correlation with each other. This can also be seen at the branches of the dendrogram because here are the two properties in one cluster.

2.0.6 Principal Components Analysis (PCA)

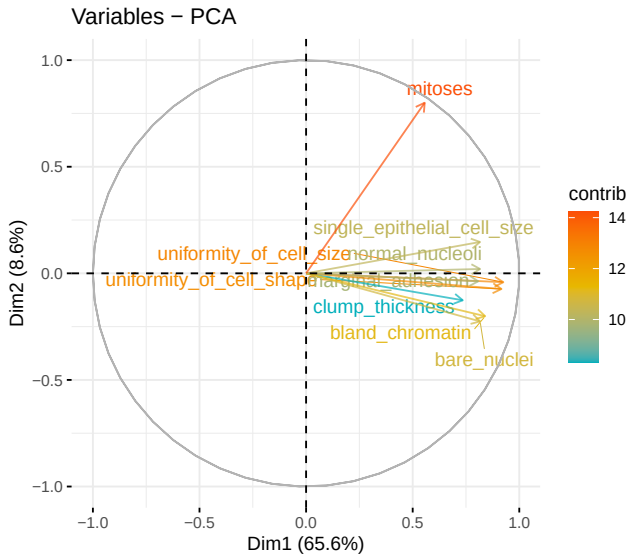

Figure 5: Correlation plot with the contribution of variables.

Figure 5 shows a plot of the contribution per dimension which has two dimensions. The contribution shows a correlation between variables. In the plot, there aren't properties that have an arrow to the left and aren't negative. But some properties point to the bottom right so they are negative for dim2 and positive for dim1. The properties/variables mitoses, uniformity of cell size, and uniformity of cell shape contribute the most to dimensions 1 and 2 because these properties have a high contrib value of 13/14. But mitoses has the highest contrib value and thus contributes the most. The property clump thickness shows the least contribute variable because the variable is colored blue. The other variables lie in between. The properties uniformity of cell size and uniformity of cell shape have the same direction of their arrow and they have the same contrib value.

2.0.7 Heatmap after removal

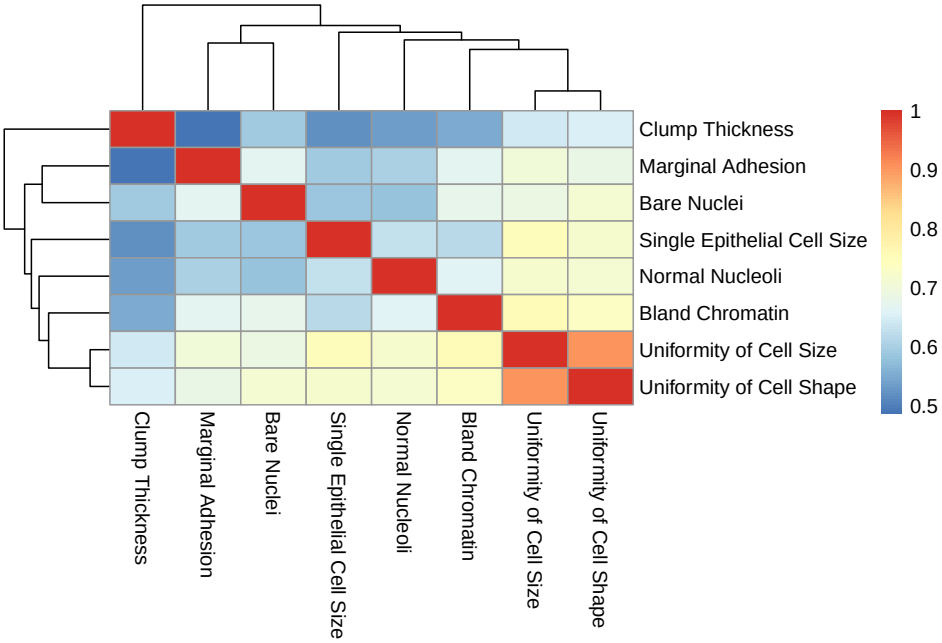

Figure 6: Heatmap with the correlation matrix of the properties. After the removal of the property mitoses.

Figure 6 shows the correlation matrix in a heatmap for all properties. Here can be seen that the property mitoses is removed compared to the heatmap in figure 4. The legend with the colors is also adapted, here is the lowest value 0.5, and in figure 4 0.3. In the heatmap 6 are the properties highly correlated with themselves because they contain the same values. The heatmap shows a better patron than in figure 4 because mitoses isn't shown here otherwise, the figure has remained the same. The two properties Uniformity of cell size and Uniformity of cell shape show a high correlation with each other and are in the same cluster.

3 Discussion and conclusion

3.1 Discussion

In a quick overview can be seen that the properties consist of 1-10 values with a slight variation in their values between the properties. Which can be seen when comparing the results of the figure 1 until figure 5. There is also variation in between the classes in figure 2 and 3. Figure 1 shows the density per value 1-10 of the properties here in the figure is there one with a higher peak than the other properties. All of the properties have a high value for 1 and 1 indicates benign. So indicates that there are more benign which can be shown in table 2. In figure 2 can also be seen that benign consist of low values mainly of value 1 and that malignant consist of high values mainly of value 10. Figure 3 shows a good variation between the value and the classes for the properties. Here you can clearly see groups of the classes between the properties because at the value 1 until 3 are there groups for the class benign and the rest of the values are scattered. For the class malignant are there small groups at value 10 and the rest are scattered across the values. Between the properties, there is a high correlation for one more than the other this is visualized in a heatmap4. In the heatmap they all have a high correlation between 0.5 and 1 except the property mitoses which has a low correlation with other properties. The PCA plot shown in figure 5 shows the contribution of the properties/variables. Here can be seen that the properties point the same way and have a contrib value close to each other. Except for the property mitoses it differs from the other properties and clump thickness but only in contrib value. So mitoses deviates the most in all figures because there are mainly low values in the properties. This can be seen the most in figure 3 because this shows that there are a lot of dots between the value 1 to 4 and it has bigger gaps between the places of the values and there are fewer values above 5 compared to the other properties. Thus mitoses is removed from the dataset as seen in the heatmap in figure 6.

In general, is the data breast cancer of good quality. This is because all of the values associated with a property have almost the same variation and distribution per property like in figure 3. This figure shows a similar distribution for the values of the properties. This is not only for the properties but also for the classes as seen in figure 2. In this boxplot have the boxplots the same length for the class malignant which lies between the value 3 until 10 except for mitoses. And for the class benign is this the same here are the boxplots a line at 1 except for three properties but they aren't bigger than 4. So this shows that the data is good for partitioning in the two classes and that the classes depend on the values whether is it has low or high values.

3.2 Conclusion

As discussed is the property mitoses removed. This was the best column to remove because this property has a lower variation and distribution. Which can be seen in figure 3 because here the values lie low and there is less scattered compared to the other properties. Another reason is because of the low correlation with other properties as shown in figure 4. Here has mitoses the lowest correlation value the lowest of them all. Without mitoses is the dataset suitable for the machine learning process.

The dataset still contains the properties uniformity of cell size and uniformity of cell shape because these are needed for further research and these properties have a high correlation with each other as see in figure 4 and 6.

The dataset could have been ameliorative by providing more information about the data. Because there is little information about the original data and the subject to find but this can be because the data is from 1992. Otherwise, the data is clear in structure and content.

References

[1] Olvi L Mangasarian, W Nick Street, William H Wolberg: *Research article*, Breast Cancer Diagnosis and Prognosis via Linear Programming, December 19 1994, Retrieved from https://www.neuraldesigner.com/learning/examples/breast-cancer-diagnosis on 05-10-2021

[2] Neural Designer: *Machine Learning Examples*, Diagnose breast cancer from fine-needle aspirate images using Neural Designer, Retrieved from the website researchgate.net on 05-10-2021

[3] UCI Machine Learning Repository: *Breast Cancer Wisconsin (Original) Data Set*, Center for Machine Learning and Intelligent Systems

4 Appendix

```
knitr::opts_chunk$set(echo = FALSE)
# Add: chunk caching
knitr::opts_chunk$set(cache = TRUE)
#####################################################
## Codebook
#####################################################
## Load the codebook
codebook <- read.csv(file = "codebook.txt", sep = ";")
kable(codebook, caption = "Codebook with an overview of the data.")
#####################################################
## Libraries
#####################################################
library(kableExtra)
library(dplyr)
library(tidyr)
library(ggplot2)
library(stats)
library(pheatmap)
library(missMDA)
library(FactoMineR)
library(factoextra)
#####################################################
## Load and change the dataset
#####################################################
## Define the data file
datafile <- "data/breast-cancer-wisconsin.data"

## Load the dataset
breastcancer <- read.table(datafile, sep = ",", header = FALSE,
                           na.strings = "?")

## Changed the column names
colnames(breastcancer) <- c("ID", "clump_thickness", "uniformity_of_cell_size",
                            "uniformity_of_cell_shape", "marginal_adhesion",
                            "single_epithelial_cell_size", "bare_nuclei",
                            "bland_chromatin", "normal_nucleoli", "mitoses",
                            "class")

## Create new variables for the column class
breastcancer <- breastcancer %>%
    mutate(class = factor(class, labels = c("Benign", "Malignant"),
                          levels = c(2, 4)))
#####################################################
## Convert the data
#####################################################
## Tidying the data by using pivot_longer
breast_long <- pivot_longer(data = breastcancer,
                            cols = -c("ID", "class"),
                            names_to = "property",
                            values_to = "value")
#####################################################
```

```r
## Density plot
#####################################################
## Plot the density plot using ggplot
ggplot(data = breast_long, mapping = aes(value, color = property)) +
  geom_density() +
  labs(title="Density plot",
       subtitle="Density of value per property",
       x = "Value",
       y = "Density") +
  theme_minimal() +
  scale_x_continuous(breaks = c(1:10))
  #scale_colour_discrete(labels = codebook$Full.Name)
#####################################################
## Boxplot
#####################################################
## Plot the boxplot with ggplot
ggplot(data = breast_long, mapping = aes(property, value, color = class,
                                         fill = class)) +
  geom_boxplot(aes(outlier.colour = class), alpha = 0.6) +
  scale_colour_manual(name = "class", values = c("steelblue", "plum1"),
                      aesthetics = c("color", "fill")) +
  labs(title="Boxplot",
       subtitle="Property per stage of breastcancer",
       x="Property",
       y="Value") +
  coord_flip() +
  theme_classic() +
  scale_y_continuous(breaks = c(1:10))
  #scale_x_discrete(labels = codebook$Full.Name[-1])
#####################################################
## Table of the total per class
#####################################################
## Create table
kable(table(breastcancer$class), col.names = c("Class", "Total"),
caption = "Total number of the class")
#####################################################
## Plot with points
#####################################################
## Plot the figure
ggplot(data = breast_long, mapping = aes(x = property, y = value, color = class)) +
  geom_jitter(alpha = 0.8) +
  scale_colour_manual(name = "class", values = c("steelblue", "plum1"),
                      aesthetics = c("color", "fill")) +
  labs(title = "Relationship between the properties and their values",
       x = "Property", y = "Value", colour = "Class") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1)) +
  scale_y_continuous(breaks = c(1:10))
#####################################################
## Heatmap
#####################################################
## Get the correlation matrix
cormat <- signif(cor(na.omit(breastcancer[2:10])))
```

```r
## Plot the heatmap
pheatmap(cormat, labels_row = codebook$Full.Name[-1],
         labels_col = codebook$Full.Name[-1])
#####################################################
## PCA plot
#####################################################
## Impute dataset with PCA
bc.comp <- imputePCA(breastcancer[,2:10], graph = FALSE)
bc.pca <- PCA(bc.comp$completeObs, scale.unit = TRUE, ncp = 5, graph = FALSE)

## Plot the PCA
fviz_pca_var(bc.pca, col.var = "contrib",
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE, alpha.var = 0.7)
#####################################################
## Remove column and covert data
#####################################################
## Remove the column of mitoses
breast_clean <- breastcancer[-10]

## Tidying the data by using pivot_longer
breast_long <- pivot_longer(data = breast_clean,
                            cols = -c("ID", "class"),
                            names_to = "property",
                            values_to = "value")
#####################################################
## Heatmapp after removal
#####################################################
## Get the correlation matrix
cormat <- signif(cor(na.omit(breast_clean[2:9])))

## Plot the heatmap
pheatmap(cormat, labels_row = codebook$Full.Name[-1],
         labels_col = codebook$Full.Name[-1])
```