MRR Project

Noah KWA MOUTOME - Victor TAN

2023-11-13

Introduction

This project is about the diagnosis of breast cancer. Thanks to the dataset "Breast Cancer Wisconsin (Diagnostic) Data Set" from the computer sciences department at the University of Wisconsin, we will try to predict if a tumor is benign or malignant.

Preprocessing and Data Exploration

Let's take a look at all the missing values of the dataset, if there's any, and clean the dataset:

There's no missing values in this dataset, but we have to drop the last unnamed column and the "id" one which is not usefull. So now, we have 569 observations and 30 covariables to predict 1 target variable, which is the diagnosis.

Now, let's explain the covariables:

There are ten real-valued features computed for each cell nucleus:

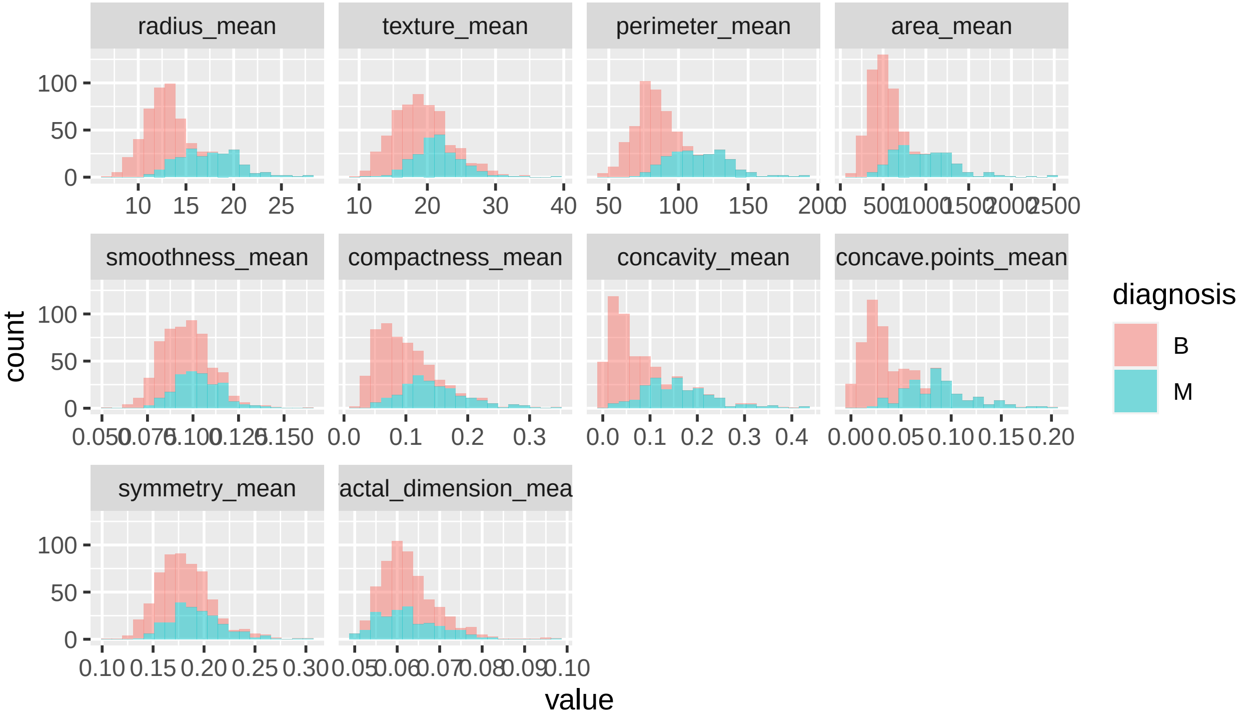
a) radius (mean of distances from center to points on the perimeter)
b) texture (standard deviation of gray-scale values)
c) perimeter
d) area
e) smoothness (local variation in radius lengths)
f) compactness (perimeterˆ2 / area - 1.0)
g) concavity (severity of concave portions of the contour)
h) concave points (number of concave portions of the contour)
i) symmetry
j) fractal dimension ("coastline approximation" - 1)

Then, the mean, the standard error and the "worst" or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 features. For instance, field 3 is Mean Radius, field 13 is Radius SE, field 23 is Worst Radius.

Description of the dataset

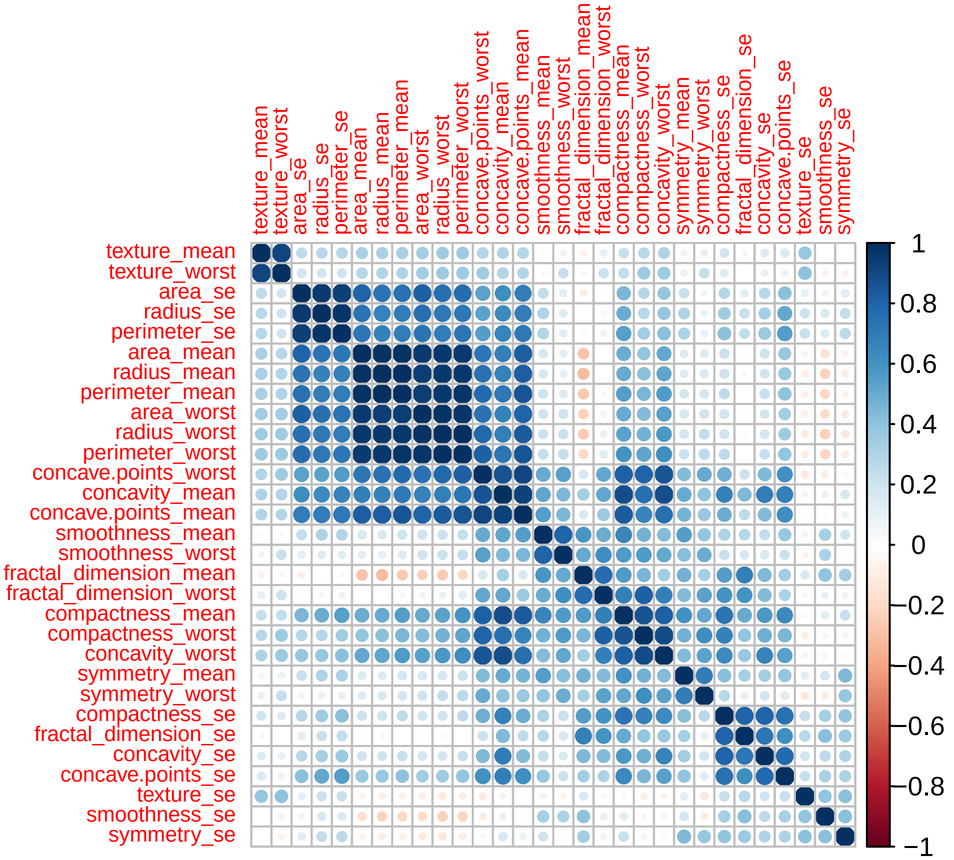
Data visualization

Let's now visualize the data to see if there's any outliers or if the data is normally distributed. These histograms are repetitive, so we will only show the means of the 10 features. The others are still available in the code for the reader to see.



It looks like most of the features are normally distributed, and there seems to be no outliers. However, there's no clear separation between benign and malignant for most of the features except "concave.points_worst", "perimeter_worst", "radius_worst" and "concave.points_mean". So a visualization of the values of these features is not enough.

Let's now look at the coorelation plot :



We can see that there's a strong correlation between some features, which can be a problem for the model.

We can now try 3 diffents approaches, and we will apply the same methods on each of them in the future:

- No covariables deletion
- Covariables deletion by highest correlation
- Covariables deletion by PCA

No covariables deletion

We still use the same dataset.

Covariables deletion by highest correlation

If two covariables are highly correlated, (for instance, if the correlation is higher than 0.8), we will delete the one that has the highest mean of correlations with the others. We can use the function "findCorrelation" from the package "caret" to provide us the said covariables that we will delete.

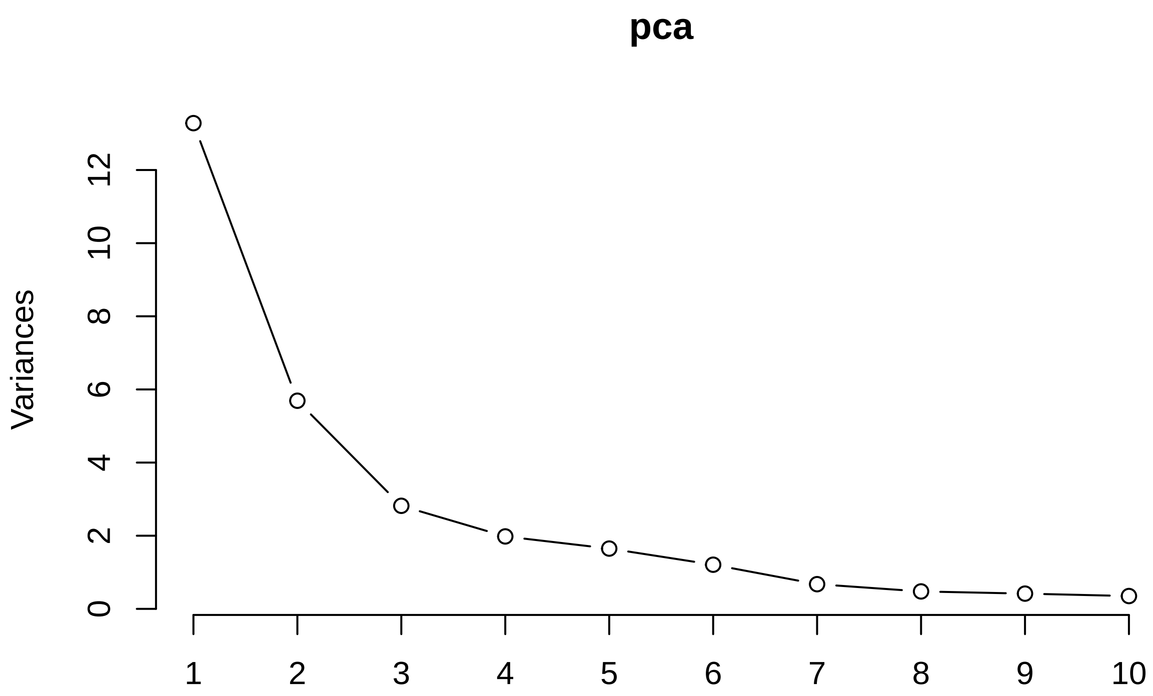
Then, we only keep the covariables that are not in the list "corr_var".

```
##  [1] "diagnosis"              "perimeter_mean"
##  [3] "concave.points_mean"    "symmetry_mean"
##  [5] "fractal_dimension_mean" "radius_se"
##  [7] "area_se"                "compactness_se"
##  [9] "concavity_se"           "concave.points_se"
## [11] "symmetry_se"             "radius_worst"
## [13] "area_worst"              "concave.points_worst"
## [15] "symmetry_worst"          "fractal_dimension_worst"
```

We now have a new dataset with less covariables, only 16. However, this approch is not perfect, because we only scan the correlation between two covariables once: there could still be high correlations between variables if 3 or more covariables were correlated together, and deleting more covariables could reduce to much information.

Covariables deletion by PCA

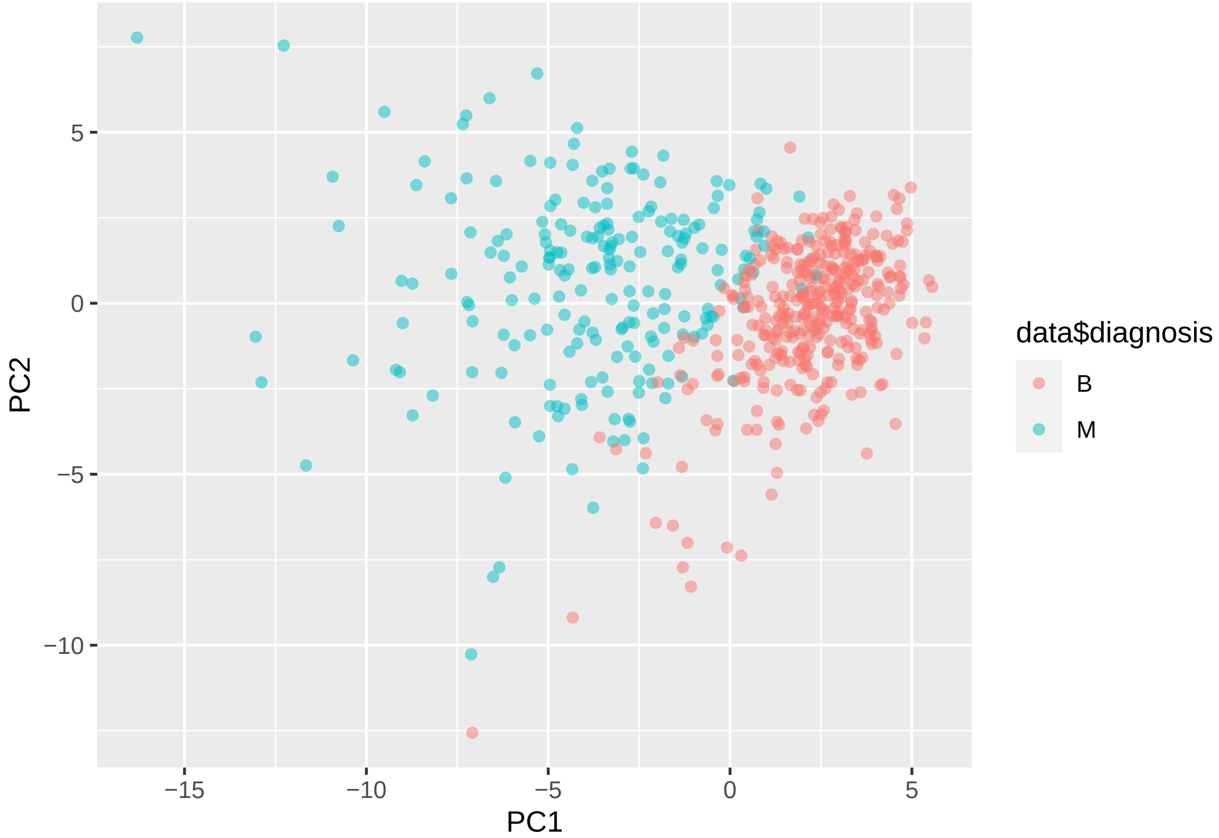
We will now use the PCA method to reduce the number of covariables. We will use the function “prcomp” from the package “stats” to do so. Because each covariable has a different scale, we will use the parameters “center = TRUE” and “scale. = TRUE” to standardize the covariables.



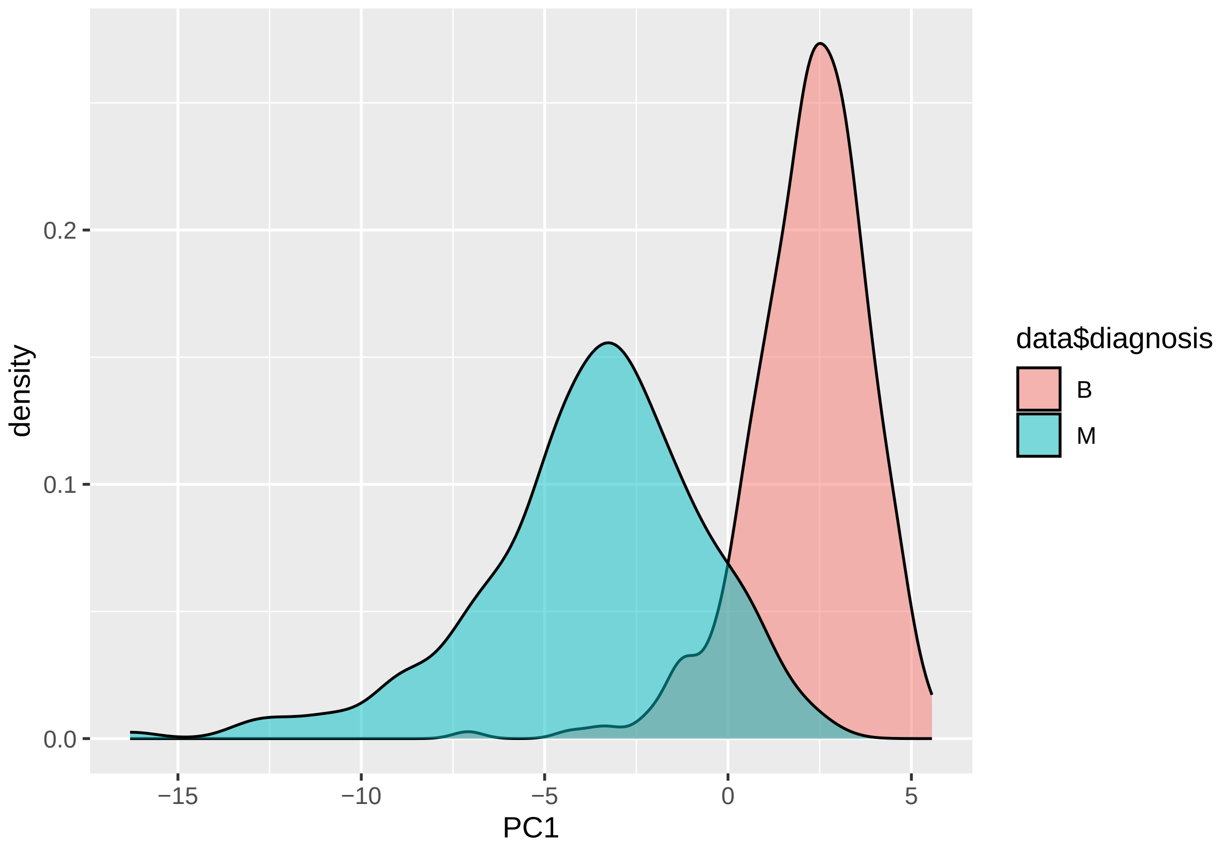
```
## Importance of components:
##                           PC1    PC2     PC3     PC4     PC5     PC6     PC7
## Standard deviation     3.6444 2.3857 1.67867 1.40735 1.28403 1.09880 0.82172
## Proportion of Variance 0.4427 0.1897 0.09393 0.06602 0.05496 0.04025 0.02251
## Cumulative Proportion  0.4427 0.6324 0.72636 0.79239 0.84734 0.88759 0.91010
##                            PC8    PC9    PC10   PC11    PC12    PC13    PC14
## Standard deviation     0.69037 0.6457 0.59219 0.5421 0.51104 0.49128 0.39624
## Proportion of Variance 0.01589 0.0139 0.01169 0.0098 0.00871 0.00805 0.00523
## Cumulative Proportion  0.92598 0.9399 0.95157 0.9614 0.97007 0.97812 0.98335
##                           PC15    PC16    PC17    PC18    PC19    PC20   PC21
## Standard deviation     0.30681 0.28260 0.24372 0.22939 0.22244 0.17652 0.1731
## Proportion of Variance 0.00314 0.00266 0.00198 0.00175 0.00165 0.00104 0.0010
## Cumulative Proportion  0.98649 0.98915 0.99113 0.99288 0.99453 0.99557 0.9966
##                           PC22    PC23   PC24    PC25    PC26    PC27    PC28
## Standard deviation     0.16565 0.15602 0.1344 0.12442 0.09043 0.08307 0.03987
## Proportion of Variance 0.00091 0.00081 0.0006 0.00052 0.00027 0.00023 0.00005
## Cumulative Proportion  0.99749 0.99830 0.9989 0.99942 0.99969 0.99992 0.99997
##                           PC29    PC30
## Standard deviation     0.02736 0.01153
## Proportion of Variance 0.00002 0.00000
## Cumulative Proportion  1.00000 1.00000
```

Let's arbitrarily choose to keep enough covariables to explain at least 99% of the variance. According to the summary, we need to keep 17 covariables to do so (99.11% of the variance explained).

Though, 17 covariables means 17 dimensions, which is too much to visualize. So we can temporarily choose to keep 2 covariables, which will allow us to visualize the data. Together, they explain 63.24% of the variance.



Visually, we can see that there's a clear separation between benign and malignant tumors, but only thanks to the first principal component: If we guess the center of the two clusters of diagnosis, we can see that their PC1 values are far enough from each other to be able to separate them (around -3 for the malignant tumors and 2.5 for the benign ones), but their PC2 values are too close to each other (around 0), so we can't separate them with this principal component. Let's draw only the first distribution on PC1 as the PC2 one is not useful (the reader can still see the PC2 distribution in the code):



With the first PC1, we can see that the means of the distributions are far enough from each other to be able to separate them. But the PC1 only explains 44.27% of the variance, so it doesn't really give us a lot of information. Let's keep the 17 covariables for the next steps.

Model training and selection

We will train models with the 3 approaches.