MATH 9102 - Probability and Statistical Inference Assignment
Final Assignment

Antonio Silva (D23129331@mytudublin.ie)

2024-05-26

Contents

Abstract

Breast cancer is the most common cancer in women in developed countries, and 12% of breast cancer occurs in women 20-34 yearscite(Hickey et al. 2009).

Advancements in prediction and diagnosis are crucial for maintaining a healthy life. Accurate cancer prediction mechanisms are vital for improving patient treatment and survival rates. Predictive techniques play a significant role in the early diagnosis of breast cancer, allowing for timely intervention and better management of the disease. In this study, we focus on enhancing the accuracy of breast cancer predictions by employing advanced data analysis techniques. Specifically, we utilized Principal Component Analysis (PCA) to reduce the dataset's dimensionality. This step is essential as it helps to simplify the dataset, making the predictive model more efficient and effective. By reducing the number of variables, we can focus on the most significant features that contribute to accurate predictions, thus improving the model's overall performance.

Additionally, we applied Logistic Regression to perform the prediction of the binary outcome (presence or absence of breast cancer). Logistic Regression is well-suited for this task as it provides a clear probabilistic framework for binary classification problems. The dataset used for this analysis is the Breast Cancer Wisconsin (Diagnostic) datasetparencite(Repository 1995), a well-known dataset in the field of medical diagnostics. We conducted our analysis and reporting using *R Studio, specifically version 4.3.2, released on October 31, 2023*.

Introduction

Advances in predictive analytics and diagnostic technologies are crucial for improving public health. Early detection and accurate diagnosis of diseases like cancer significantly improve treatment outcomes and survival rates. Since breast cancer is one of the most common cancers affecting women worldwide, developing reliable prediction methods is essential. This study aims to explore advanced data analysis techniques to improve the accuracy of breast cancer diagnosis.

Predicting breast cancer involves analyzing various physiological and pathological features that indicate malignant tumors. Traditional methods often rely on clinical exams and imaging techniques, which, while effective, can be enhanced by computational methods.

Principal Component Analysis (PCA) is a useful tool for reducing the number of variables in a dataset. By simplifying the data without losing important information, PCA improves the performance of predictive models. In this study, we use PCA on the Breast Cancer Wisconsin (Diagnostic) dataset to streamline the features and improve the prediction model's efficiency. This step is essential for handling high-dimensional data in medical diagnostics and ensures that the model is robust and easy to interpret.

Logistic Regression, a common method for binary classification, is used to predict whether a tumor is malignant or benign. This technique is well-suited for medical diagnostics because it provides probabilities and can handle various predictor variables. By applying Logistic Regression to the PCA-transformed dataset, we aim to achieve high accuracy in predicting breast cancer. Combining PCA and Logistic Regression offers a comprehensive approach to addressing the complexity of breast cancer prediction.

The dataset used in this study is the Breast Cancer Wisconsin (Diagnostic) dataset, a well-known resource in medical research. This dataset includes various features extracted from digitized images of fine needle aspirate (FNA) of breast masses(Repository 1995), making it ideal for predictive analysis. We conducted the analysis and reporting using *R Studio (version 4.3.2, released on October 31, 2023)*, which provides a robust environment for statistical computing and graphics. Through this study, we aim to enhance breast cancer detection and improve patient outcomes, supporting the broader goal of advancing healthcare through data-driven methods.

Method

Participants

The Breast Cancer Wisconsin (Diagnostic) dataset includes data from patients who underwent fine needle aspirate (FNA) of breast masses.

The dataset consists of 569 instances with data collected from real women.

The dataset does not provide personal demographic information such as age, ethnicity, or geographic location, focusing instead on the clinical and pathological features of the breast masses.

For each participant, 30 features were extracted from the FNA samples. These features describe the characteristics of the cell nuclei present in the samples. The features include measurements such as radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension. These features were recorded for both the mean, standard error, and "worst" or largest values across the samples.

Each sample is labeled with a diagnosis indicating whether the breast mass is benign (B) or malignant (M). This binary outcome is used to train and test predictive models aimed at diagnosing breast cancer.

Procedure

Dataset Exploration & Analysis

The dataset includes the following features (variables): We do not have any variables related to the individual, except for the ID. There is a categorical variable indicating whether the tumor is benign or malignant. All other variables are numerical, and for each, the mean, standard error, and worst value were measured during the exam.

Variable	Type	Description
id	Ordinal	Number of the patient
diagnosis	Categorical	M = malignant, B = benign
radius	Continuous	Mean of distances from center to points on the perimeter
texture	Continuous	Standard deviation of gray-scale values
perimeter	Continuous	

Table 2: Descriptive Analysis of the dataset

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
diagnosis*	1	569	1.37	0.48	1.00	1.34	0.00	1.00	2.00	1.00	0.53	-1.73	0.02
radius_mean	2	569	14.13	3.52	13.37	13.82	2.82	6.98	28.11	21.13	0.94	0.81	0.15
texture_mean	3	569	19.29	4.30	18.84	19.04	4.17	9.71	39.28	29.57	0.65	0.73	0.18
perimeter_mean	4	569	91.97	24.30	86.24	89.74	18.84	43.79	188.50	144.71	0.99	0.94	1.02
area_mean	5	569	654.89	351.91	551.10	606.13	227.28	143.50	2501.00	2357.50	1.64	3.59	14.75
smoothness_mean	6	569	0.10	0.01	0.10	0.10	0.01	0.05	0.16	0.11	0.45	0.82	0.00
compactness_mean	7	569	0.10	0.05	0.09	0.10	0.05	0.02	0.35	0.33	1.18	1.61	0.00
concavity_mean	8	569	0.09	0.08	0.06	0.08	0.06	0.00	0.43	0.43	1.39	1.95	0.00
concave.points_mean	9	569	0.05	0.04	0.03	0.04	0.03	0.00	0.20	0.20	1.17	1.03	0.00
symmetry_mean	10	569	0.18	0.03	0.18	0.18	0.03	0.11	0.30	0.20	0.72	1.25	0.00
fractal_dimension_mean	11	569	0.06	0.01	0.06	0.06	0.01	0.05	0.10	0.05	1.30	2.95	0.00
radius_se	12	569	0.41	0.28	0.32	0.36	0.16	0.11	2.87	2.76	3.07	17.45	0.01
texture_se	13	569	1.22	0.55	1.11	1.16	0.47	0.36	4.88	4.52	1.64	5.26	0.02
perimeter_se	14	569	2.87	2.02	2.29	2.51	1.14	0.76	21.98	21.22	3.43	21.12	0.08
area_se	15	569	40.34	45.49	24.53	31.69	13.63	6.80	542.20	535.40	5.42	48.59	1.91
smoothness_se	16	569	0.01	0.00	0.01	0.01	0.00	0.00	0.03	0.03	2.30	10.32	0.00
compactness_se	17	569	0.03	0.02	0.02	0.02	0.01	0.00	0.14	0.13	1.89	5.02	0.00
concavity_se	18	569	0.03	0.03	0.03	0.03	0.02	0.00	0.40	0.40	5.08	48.24	0.00
concave.points_se	19	569	0.01	0.01	0.01	0.01	0.01	0.00	0.05	0.05	1.44	5.04	0.00
symmetry_se	20	569	0.02	0.01	0.02	0.02	0.01	0.01	0.08	0.07	2.18	7.78	0.00
fractal_dimension_se	21	569	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.03	3.90	25.94	0.00
radius_worst	22	569	16.27	4.83	14.97	15.73	3.65	7.93	36.04	28.11	1.10	0.91	0.20
texture_worst	23	569	25.68	6.15	25.41	25.39	6.42	12.02	49.54	37.52	0.50	0.20	0.26
perimeter_worst	24	569	107.26	33.60	97.66	103.42	25.01	50.41	251.20	200.79	1.12	1.04	1.41
area_worst	25	569	880.58	569.36	686.50	788.02	319.65	185.20	4254.00	4068.80	1.85	4.32	23.87
smoothness_worst	26	569	0.13	0.02	0.13	0.13	0.02	0.07	0.22	0.15	0.41	0.49	0.00
compactness_worst	27	569	0.25	0.16	0.21	0.23	0.13	0.03	1.06	1.03	1.47	2.98	0.01
concavity_worst	28	569	0.27	0.21	0.23	0.25	0.20	0.00	1.25	1.25	1.14	1.57	0.01
concave.points_worst	29	569	0.11	0.07	0.10	0.11	0.07	0.00	0.29	0.29	0.49	-0.55	0.00
symmetry_worst	30	569	0.29	0.06	0.28	0.28	0.05	0.16	0.66	0.51	1.43	4.37	0.00
fractal_dimension_worst	31	569	0.08	0.02	0.08	0.08	0.01	0.06	0.21	0.15	1.65	5.16	0.00

Variable	Type	Description
area	Continuous	
smoothness	Continuous	Local variation in radius lenghts
compactness	Continuous	(perimeter ˆ 2 / area - 1)
concavity	Continuous	Severity of concave portions of the contour
symmetry	Continuous	
fractal_dimension	Continuous	“Coastline approximation” - 1

```r
data <- read.csv("data.csv")
desc_data <- data %>% select(-X, -id)
desc_data <- suppressWarnings(describe(desc_data))
desc_data <- desc_data %>% mutate(across(where(is.numeric), round, 2))
kable(desc_data, format = "latex",
      caption = "Descriptive Analysis of the dataset") %>%
  kable_styling(latex_options = c("striped", "scale_down"))
```

The descriptive analysis shown in Table 2 reveals the following observations:

- Most features exhibit some degree of right skewness, indicating that extreme values on the higher side are common;
- The standard errors are relatively low compared to the mean values, suggesting that the measurements are fairly consistent;
- `area` and `radius` show high variability, which may be significant in distinguishing between benign and malignant tumors.

The next tables summarizes the counts of benign and malignant cases in the dataset:

Table 3: Tumors dataset classification

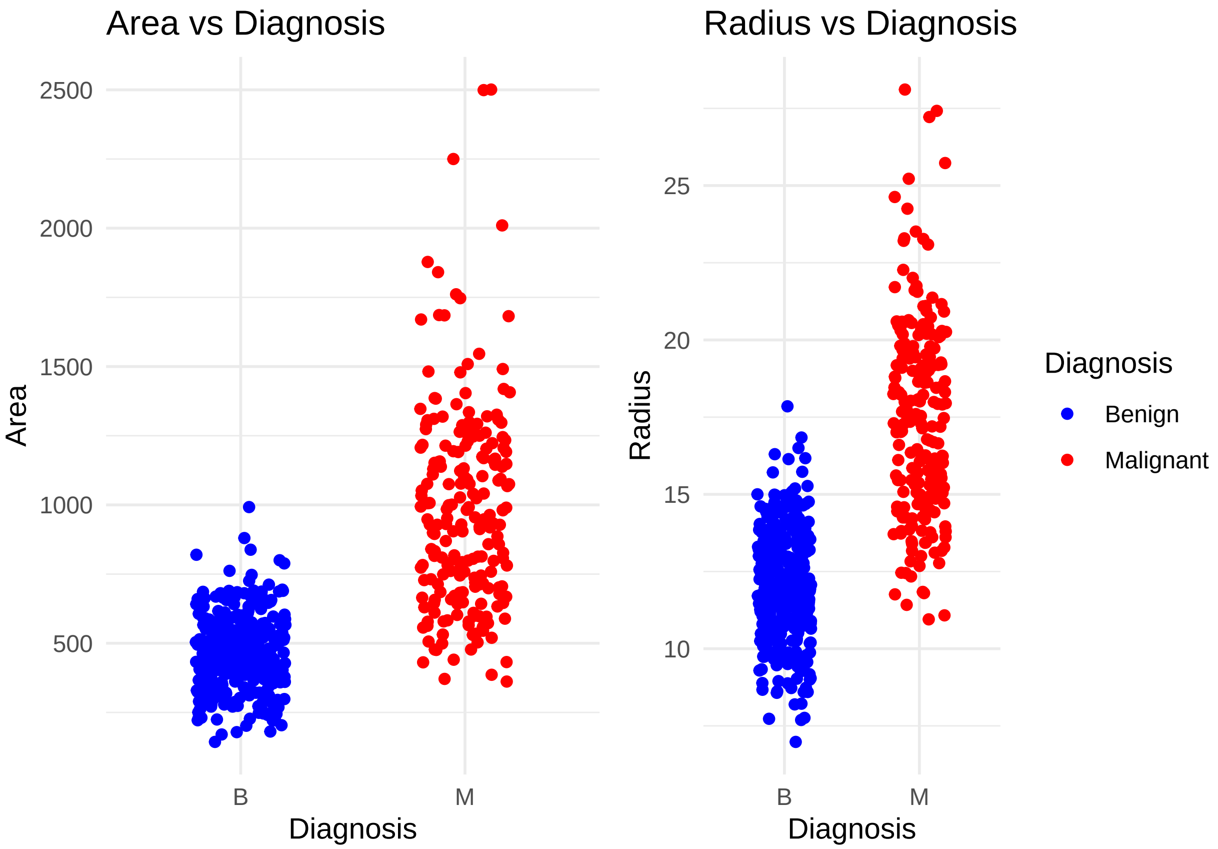
Classification	Frequency
Benign	357
Malignant	212

- Benign: There are 357 cases where the tumor is classified as benign;
- Malignant: There are 212 cases where the tumor is classified as malignant.

This imbalance could affect the performance of predicting model, making it biased towards the more frequent class.

```r
diagnosis_table <- table(data$diagnosis)
names(diagnosis_table) <- c("Benign", "Malignant")
kable(diagnosis_table, format = "latex",
      caption = "Tumors dataset classification",
      col.names=c("Classification", "Frequency"))
```

```r
data$diagnosis <- as.factor(data$diagnosis)

plot_area <- ggplot(data, aes(x = diagnosis, y = area_mean, color = diagnosis)) +
  geom_jitter(width = 0.2) +
  theme_minimal() +
  labs(title = "Area vs Diagnosis",
       x = "Diagnosis",
       y = "Area") +
  scale_color_manual(values = c("B" = "blue", "M" = "red"), name = "Diagnosis") +
  theme(legend.position="none")

plot_radius <- ggplot(data, aes(x = diagnosis, y = radius_mean, color = diagnosis)) +
  geom_jitter(width = 0.2) +
  theme_minimal() +
  labs(title = "Radius vs Diagnosis",
       x = "Diagnosis",
       y = "Radius") +
  scale_color_manual(values = c("B" = "blue", "M" = "red"),
                     name = "Diagnosis",
                     labels = c("Benign", "Malignant"))

grid.arrange(plot_area, plot_radius, ncol = 2)
```



Both area and radius are useful features for distinguishing between benign and malignant tumors. The scatterplots show that malignant tumors tend to have higher values for these features compared to benign tumors.

In terms of data cleaning, there is no need to impute data as there are no missing values. Additionally, we do not need to address outliers because they are related to the diagnosis, and excluding them could negatively impact our analysis.

Feature Selection

Based on dataset we will select only the following variables.

We will exclude all variables ending with `_se` and `_worst`. These variables are highly correlated with their corresponding `_mean` values, reducing their significance.

For example:

- `area_worst` represents the *worst* value for the area;
- `area_se` represents the *standard error* for the area;
- `area_mean` represents the *mean* for the area.

So we believe that the `area_mean` is the most representative feature for the area.

`diagnosis` will be our *dependent variable*.

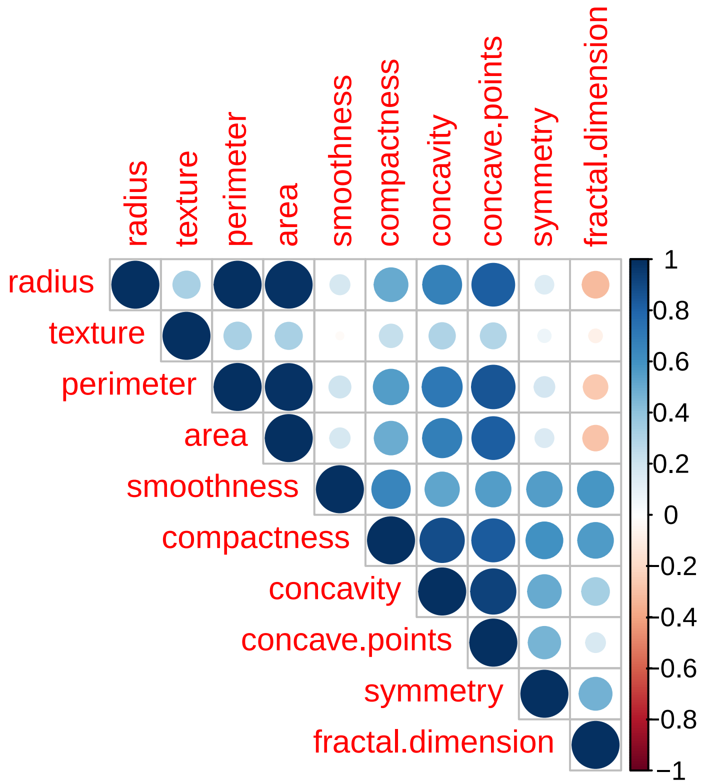
```r
selected_features <- data %>%
  select(contains("_mean"))  %>%
  rename_with(~ sub("_mean$", "", .))  %>%
  rename_with(~ gsub("_", ".", .))

corr_mt <- cor(selected_features)
```

Table 4: Correlation Matrix for Selected Features

	radius	texture	perimeter	area	smoothness	compactness	concavity	concave.points	symmetry	fractal.dimension
radius	**1**	0.32	**1**	**0.99**	0.17	0.51	0.68	**0.82**	0.15	-0.31
texture	0.32	**1**	0.33	0.32	-0.02	0.24	0.3	0.29	0.07	-0.08
perimeter	**1**	0.33	**1**	**0.99**	0.21	0.56	0.72	**0.85**	0.18	-0.26
area	**0.99**	0.32	**0.99**	**1**	0.18	0.5	0.69	**0.82**	0.15	-0.28
smoothness	0.17	-0.02	0.21	0.18	**1**	0.66	0.52	0.55	0.56	0.58
compactness	0.51	0.24	0.56	0.5	0.66	**1**	**0.88**	**0.83**	0.6	0.57
concavity	0.68	0.3	0.72	0.69	0.52	**0.88**	**1**	**0.92**	0.5	0.34
concave.points	**0.82**	0.29	**0.85**	**0.82**	0.55	**0.83**	**0.92**	**1**	0.46	0.17
symmetry	0.15	0.07	0.18	0.15	0.56	0.6	0.5	0.46	**1**	0.48
fractal.dimension	-0.31	-0.08	-0.26	-0.28	0.58	0.57	0.34	0.17	0.48	**1**

```r
corr_mt_formatted <- as.data.frame(corr_mt) %>%
  mutate(across(where(is.numeric), function(x){
    n <- round(x, 2)
    highly_correlated <- abs(x) > 0.9
    high_correlated <- abs(x) > 0.8
    cell_spec(n,
                 format = "latex",
                 color = ifelse(highly_correlated, "red",
                                ifelse(high_correlated, "blue", "black")),
                 bold = ifelse(highly_correlated | high_correlated, T, F))
  }))

kable(corr_mt_formatted, format = "latex", escape = FALSE,
      caption = "Correlation Matrix for Selected Features") %>%
  kable_styling(latex_options = c("striped", "scale_down"))
```

```r
corrplot(corr_mt, type = "upper",
         title = "Correlation Matrix of Selected Features",
         mar = c(0, 0, 2, 0))
```



Correlation Matrix of Selected Features

By the correlation matrix and plot we can conclude:

- There is a *Very High Correlation* (0.9+) between `area`, `perimeter` and `radius`;
- There is also a *Very High Correlation* (0.9+) between `concavity` and `concave.points`;
- There is *High Correlation* (0.8+) between `concavity`, `concave.points`, `compactness`, `area`, `perimeter` and `radius`.

Based on the previous correlation matrix we will discard all the features with correlation bigger than 0.9 between them. So we will:

- Keep the feature `area` and discard `perimeter` and `radius`;
- Keep the feature `concavity` and discard `concave.points`.

Our final selected features are:

- diagnosis (target/dependent variable);
- area;
- texture;
- smoothness;
- compactness;
- concavity;
- symmetry;
- fractal.dimension.

PCA

Before applying PCA, it is essential to select the predictor variables and scale them. This step is important because PCA assumes that the data is normally distributed and is sensitive to the variance of the variables. Standardizing the data ensures that each variable contributes equally to the analysis and that the results

Table 5: PCA Summary

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Standard deviation	1.870985	1.292742	0.8870152	0.7041895	0.6096074	0.3120227	0.2767252
Proportion of Variance	0.500080	0.238740	0.1124000	0.0708400	0.0530900	0.0139100	0.0109400
Cumulative Proportion	0.500080	0.738820	0.8512200	0.9220600	0.9751500	0.9890600	1.0000000

Table 6: Eigenvalues and variance

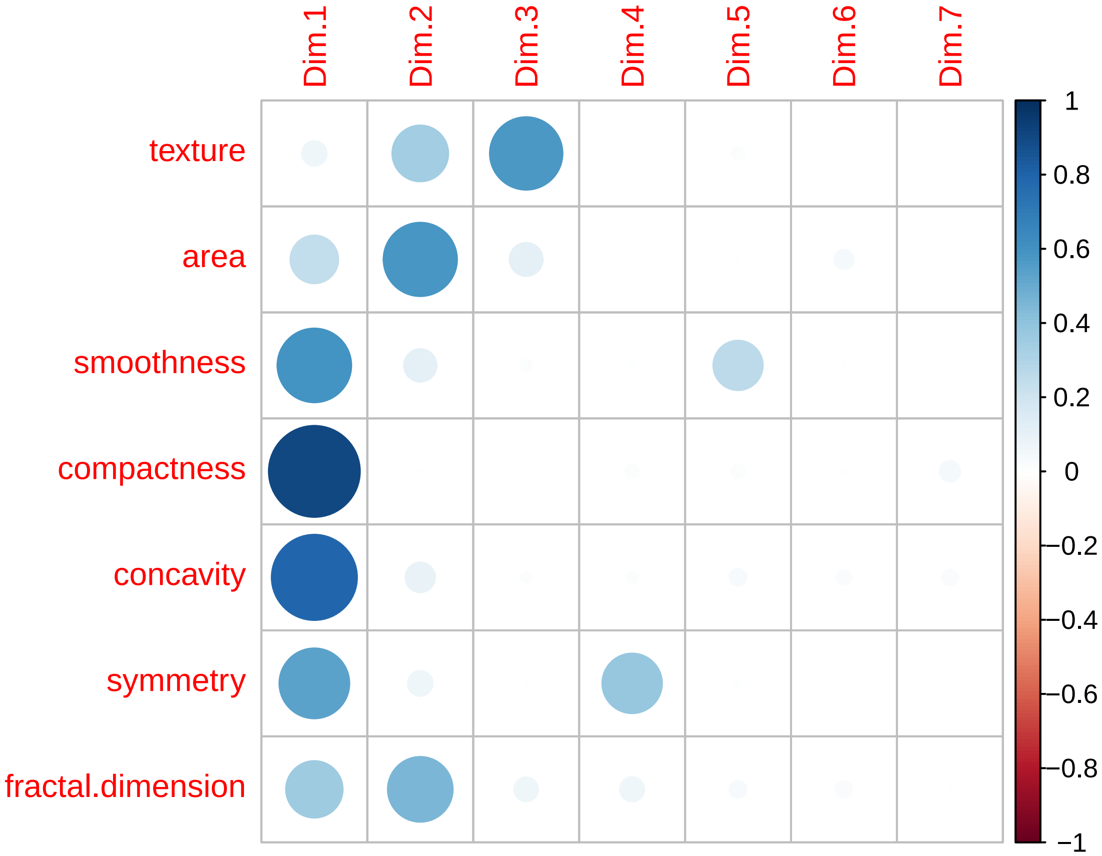
	eigenvalue	variance.percent	cumulative.variance.percent
Dim.1	3.5005842	50.008346	50.00835
Dim.2	1.6711809	23.874013	73.88236
Dim.3	0.7867959	11.239941	85.12230
Dim.4	0.4958828	7.084040	92.20634
Dim.5	0.3716212	5.308874	97.51521
Dim.6	0.0973581	1.390831	98.90605
Dim.7	0.0765768	1.093954	100.00000

are not dominated by variables with higher variance. Luckily we can do it using the `center` and `scale.` parameter from the `prcomp` (Principal Components Analysis) function.

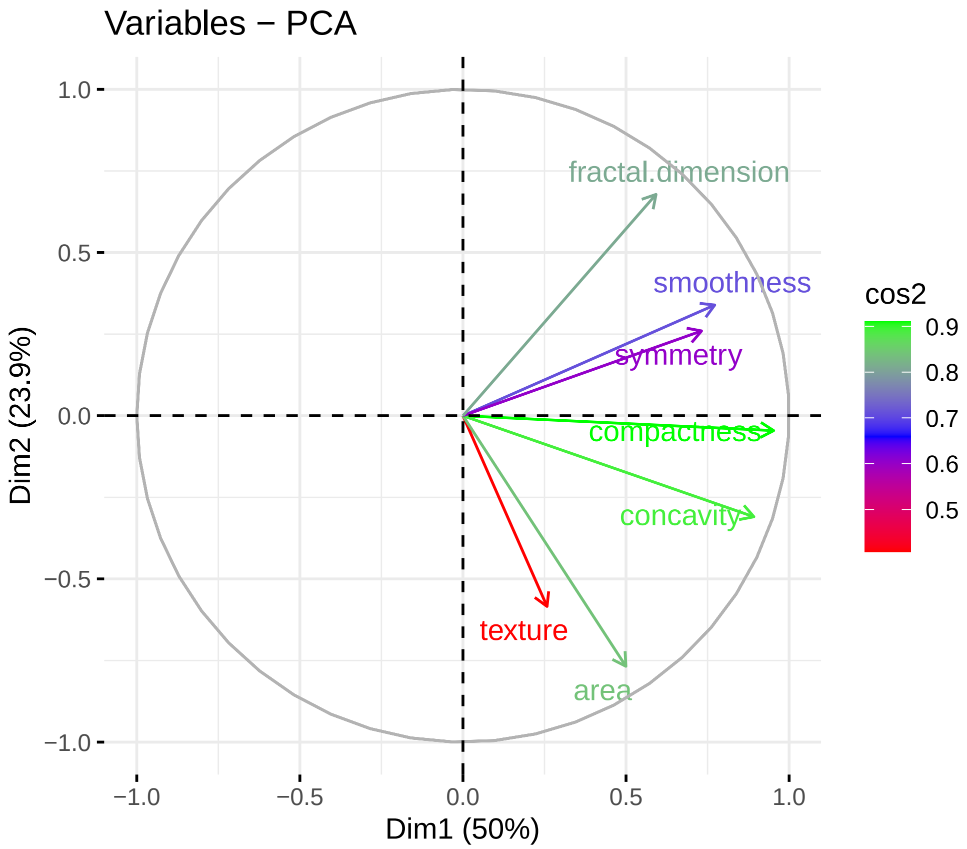
```r
pca_selected <-  data %>%
  select(contains("_mean")) %>%
  rename_with(~ sub("_mean$", "", .))  %>%
  rename_with(~ gsub("_", ".", .)) %>%
  select(-c("perimeter", "radius", "concave.points"))

pca <- prcomp(pca_selected, center = TRUE, scale. = TRUE)

pca_summary <- summary(pca)

kable(pca_summary$importance, format = "latex", escape = FALSE,
      caption = "PCA Summary") %>%
  kable_styling(latex_options = c("striped", "scale_down"))
```

```r
eig.val <- get_eigenvalue(pca)
kable(eig.val, format = "latex", escape = FALSE,
      caption = "Eigenvalues and variance") %>%
  kable_styling(latex_options = c("striped", "scale_down"))
```

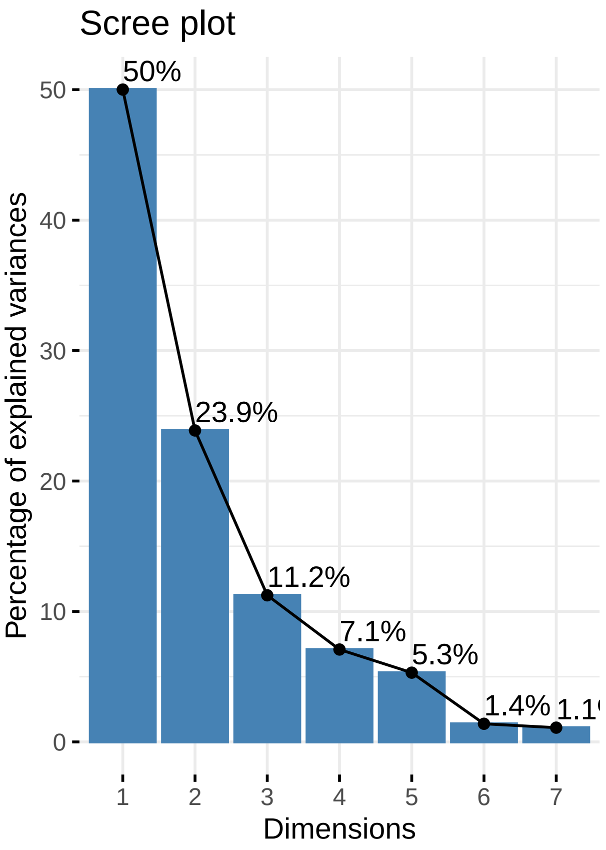
```r
screeplot <- fviz_eig(pca, addlabels = TRUE)

vars = get_pca_var(pca)
corrplot(vars$cos2)
```

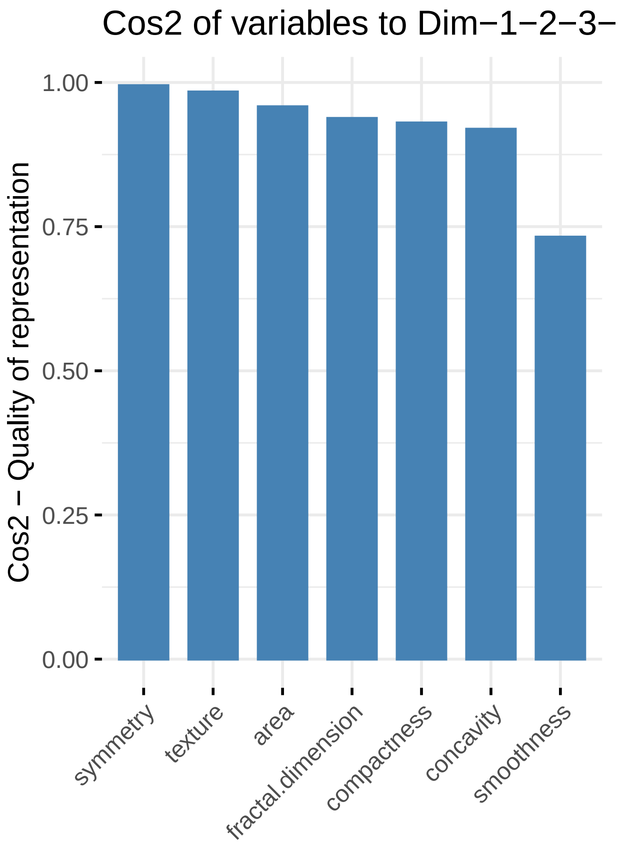


```
top_feature_contributors <- fviz_cos2(pca, choice = "var", axes=1:4)

eigenvectors <- fviz_pca_var(pca, col.var = "cos2",
                gradient.cols = c("red", "blue", "green"),
                repel = TRUE)

eigenvectors
```



```
plot_grid(screeplot, top_feature_contributors)
```





As we can see by the PCA Summary Table we saw that the first four components represents $\sim$ 92.2% of the variance. is It a quite high value and we reduce from 7 to 4 variables to use in our model later it's a $\sim$ 43% reduction of our initial features.

The first four principal components together explain about 92.2% of the total variance. This high percentage indicates that these components effectively summarize the majority of the variability in the dataset.

Initially, we have 7 features. By using PCA, we reduce this number to 4 principal components representing about 43% decrease in the number of dimensions (features)

$$Reduction\ Percentage = \frac{7-4}{7} = 0.4285714286 \approx 43.86\%$$

Using fewer features (4 instead of 7) simplifies the model, making it easier to interpret and faster to compute retaining an high explanatory power.

The eigenvectors and feature contributions plots reveal the following insights:

- All features are positively correlated;
- The most significant features for the first principal component are compactness, concavity, symmetry, and smoothness.
- The second principal component is primarily influenced by area and fractal.dimension;
- The third principal component is dominated by texture;
- Beyond the third dimension, the contributions of the features diminish. This reduction is expected since the first three principal components account for 85.12% of the total variance.

Prepare the selected data for the model

As mentioned earlier, our goal is to predict whether a patient has cancer based on data obtained from a fine needle aspirate (FNA) of a breast mass exam.

Table 7: Model data first rows

PC1	PC2	PC3	PC4	diagnosis
5.1370976	1.641103	-2.0036533	0.1801931	M
-0.4135265	-1.621847	-1.1658057	-0.5143533	M
2.2934161	-1.266826	-0.5936102	-0.4227546	M
6.5007434	3.772279	1.3135824	0.6362549	M
1.1901024	-1.133655	-2.0739767	0.1966861	M
2.7082503	2.205598	-0.3591657	0.4687445	M

We will utilize the first four principal components from the PCA, given their high explanatory power and the reduced set of features they represent.

Given that our *dependent variable is binary* (indicating whether a patient has cancer or not) and we have multiple predictors (the first four principal components from PCA), the most appropriate method for analysis is *multivariate logistic regression*.

This model will predict the probability of having cancer (the probability of diagnosis is malignant or value M).

So our model will be the following one (Assumming the probability of a malignant cancer is $Y = 1$):

$$logit(P(Y = 1)) = \beta_0 + \beta_1 PC1 + \beta_2 PC2 + \beta_3 PC3 + \beta_4 PC4$$

The hypotheses that we want to test with this model is if our model significant predict the cancer diagnosis:

$$\begin{cases} H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0 \\ H_1 : \exists i \in \{1, 2, 3, 4\} : \beta_i \neq 0 \end{cases}$$

- The null hypothesis states that the first four principal components do not significantly predict cancer diagnosis.
- The alternative hypothesis states that the first four principal components do significantly predict cancer diagnosis.

Data preparation

```
model_data <- as.data.frame(pca$x[, 1:4])
model_data$diagnosis <- data$diagnosis

kable(head(model_data), format = "latex",
      caption = "Model data first rows") %>%
  kable_styling(latex_options = c("striped", "scale_down"))
```

```
str(model_data)
```

```
## 'data.frame':    569 obs. of  5 variables:
##  $ PC1      : num  5.137 -0.414 2.293 6.501 1.19 ...
##  $ PC2      : num  1.64 -1.62 -1.27 3.77 -1.13 ...
##  $ PC3      : num  -2.004 -1.166 -0.594 1.314 -2.074 ...
##  $ PC4      : num  0.18 -0.514 -0.423 0.636 0.197 ...
##  $ diagnosis: Factor w/ 2 levels "B","M": 2 2 2 2 2 2 2 2 2 2 ...
```

```
levels(model_data$diagnosis)
```

```
## [1] "B" "M"
```

Model

Model Resulta

Hypotheses definition

Test logistic regression

Classification

NEXT CHAPTER

evaluation results

conclusion

Hickey, M, M Peat, C Saunders, and M Friedlander. 2009. "Breast Cancer in Young Women and Its Impact on Reproductive Function." *Human Reproduction Update* 15 (3): 323–39.

Repository, UCI Machine Learning. 1995. "Breast Cancer Wisconsin (Diagnostic)." 1995. https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic.