Breast Cancer Risk Factors: A Logistic Regression Approach

Debargha Chatterjee

Semester: VI

Paper Code: DSE-B2

Roll No: 213544-21-0077

Reg. No: 544-1111-0431-21

Vivekananda College
(Affiliated to the University of Calcutta)

Date: 29/07/2024

Table of Contents

VIVEKANANDA COLLEGE

AFFILIATED TO THE UNIVERSITY OF CALCUTTA

CERTIFICATE

This is certified that the project paper entitled "***Breast Cancer Risk Factors: A Logistic Regression Approach***" is submitted by **DEBARGHA CHATTERJEE** in partial fulfilment of the requirement for the Bachelor degree of Statistics (Honours) is based upon the result of benefited research work carried out by the investigator under guidance and supervision of the professors of Department of Statistics, Vivekananda College, Thakurpukur.

The results of the investigator reported in this project paper have not so far been submitted for any degree or diploma.

DECLARATION

I, **DEBARGHA CHATTERJEE**, a student of B.Sc. Semester-6, Statistics Honours, of University of Calcutta, Registration no: 544-1111-0431-21, Roll no: 213544-21-0077, hereby declare that I have done this piece of project work entitled as "Breast Cancer Risk Factors: A Logistic Regression Approach" under the supervision of Professors of Department of Statistics, Vivekananda College) as a part of B.Sc. Sem-6 examination according to the syllabus paper DSE-B2. I further declare that the piece of project work has not been published elsewhere for any degree or diploma or taken from any published project.

_____

Signature of the Student

ACKNOWLEDGEMENTS

I am indebted to number of person for helping me in the preparation of this project.

Firstly, Dr. Tapan Kumar Poddar, Former Principal, Vivekananda College, University of Calcutta, without whose help I couldn't have been a part of this prestigious college and Prof. Nabakishore Chanda, Teacher-in Charge, Vivekananda College.

I owe a deep debt of gratitude to my supervisors Prof. Nilkanta Mukherjee, Head of the Department of Statistics, Vivekananda College, Prof. Sutapa Biswas, Prof. Riddhi Das Majumder and Dr. Junaid Khan (Faculty Members) for necessary guidance, for this presentation of this dissertation, valuable comments and suggestions. I am extremely grateful to them for the necessary stimulus, support and valuable time, often took pains and stood by me in adverse circumstances. Without their encouragement and inspiration, it was not possible for me to complete this project.

Finally, my earnest thanks go to my friends who were always beside me when I needed them without any excuses and made these three years worthwhile.

This project is not only a mere project. It is the memories spend with the whole department which has created a mutual understanding among us. There are many emotions related to this piece of work, especially respect and duty towards teachers and vice versa; educational attachment with my friends; social attachment with my college.

ABSTRACT OF THE PROJECT

Breast Cancer is a significant public health concern globally, with its incidence steadily increasing over the years. Each year, about 30% of all newly diagnosed cancers in women are breast cancer[1]. In response to this issue, the objective of this project is to implement *Logistic Regression Analysis* to examine the factors associated with the occurrence of breast cancer and develop a predictive model using a comprehensive dataset of breast cancer cases.

The *relevance* of this project lies in its potential to enhance our understanding of the complex interrelations between various risk factors and the occurrence of breast cancer. By identifying significant predictors, healthcare professionals can better assess an individual's risk factors and take preventive measures accordingly. Moreover, the developed model could aid in *early detection of Malignant tumors* (tumors that causes cancer) from the physical features of the tumors. This can potentially lead to improved survival rate among affected individuals.

This project represents a significant contribution to the field of breast cancer research by implementing advanced statistical techniques to unravel the intricate relationships between risk factors and disease occurrence. The aim of this project is to generate actionable insights that can inform clinical practice, public health initiatives and future research endeavors in the fight against breast cancer.

[Keywords- Breast Cancer, Logistic Regression, risk factors, healthcare]

1. Introduction

Breast cancer is the most common type of cancer and the primary cause of mortality among women aged 45-55 years [2] . From being fourth in the list of most common cancers in India during the 1990s, it has now become the first. About 1 in 28 Indian women is likely to develop breast cancer during her lifetime. The chances of developing breast cancer are more (1 in 22) for urban women than their rural counterpart (1 in 60)[3]. In India, the incidence has increased significantly, almost by 50%, between 1965 and 1985. The estimated number of incident cases in India in 2016 was 118000, 98.1% of which were females. In 2022, an estimated 287,850 new cases of invasive breast cancer are expected to be diagnosed in women in the U.S., along with 51,400 new cases of non-invasive (in situ) breast cancer[4]. A man's lifetime risk of breast cancer is about 1 in 833. About 43,250 women in the U.S. are expected to die in 2022 from breast cancer[5]. Women are seriously threatened by breast cancer with high morbidity and mortality. Every year, death rate increases drastically due to breast cancer. Breast cancer cells usually form a tumor that can often be seen on an x-ray or felt as a lump.

Here we predict the malignancy of the tumor based on the 30 characteristics of individual cells of breast cancer obtained from a minimally invasive fine needle aspirate (FNA), to discriminate benign from malignant lumps of a breast mass using Logistic Regression.

As prerequisites we shall be introducing the concepts of a Classification problem, Logistic Regression. And how one can use Logistic Regression to solve a Classification Problem.

1.1. Classification Problems

Classification problems are the set of problems which deal with identifying the categories or the *classes* to which data points or observations belong to. In this problem, we will be having observations with certain features which we are interested in and we will know which classes some of them belong to. For the rest we will be predicting the classes based on the information we already have.

Classification can be of two types: *Binary Classification* when we have just two target classes, and *Multiclass Classification* when we have more than two

target classes. One of the most common algorithms to approach Binary classification problem is **Logistic Regression**.

2. Problem Statement

This section will state our objective and the dataset that we will be using for this purpose.

2.1. Objective

Breast Cancer occurs as a result of abnormal growth of cells in the breast tissue, commonly referred to as a *Tumor.* A tumor can be benign (not cancerous) or malignant (cancerous). This analysis aims to observe which features are most helpful in predicting malignant or benign tumor and to see general trends that may aid us in understanding Breast Cancer better. Based on the 30 characteristics mentioned in the following dataset the goal is to classify whether the breast tumor is benign or malignant.

2.2. About the dataset

We will use the "Breast Cancer Wisconsin (Diagnostic)" (WBCD) dataset, provided by the University of Wisconsin, and hosted by the UCI Machine Learning Repository. In the dataset, 30 characteristics of individual cells of breast cancer are studied. The dataset can be found in this Kaggle.

QR code:

The first column of the dataset corresponds to the patient ID, while the last column represents the diagnosis (the outcome can be "Benign" or "Malignant" based on the type of diagnosis reported). The resulting dataset consists of 569 patients: 212 (37.2%) have an outcome of Malignancy and 357 (62.7%) are Benign.

In detail, the dataset consists of ten real-valued features computed for each tumor cell:

1. Radius (mean of distances from centre to points on the perimeter)
2. Texture (standard deviation of Gray-scale values)
3. Perimeter
4. Area
5. Smoothness (local variation in radius lengths)
6. Compactness (the ratio of the volume and the surface area)
7. Concavity (severity of concave portions of the contour)
8. Concave points (number of concave portions of the contour)
9. Symmetry
10. Fractal dimension (The higher the number, the more abnormal the tissue is.)

The ten real-valued features correspond to the *Mean*, *Standard Error* and the *Worst* or largest (mean of the three largest values of samples obtained from each individual). Column 2 contains the Benign or Malignant outcome.

3.Methodology

Our research methodology primarily consists of the following steps:

1. Cleaning the data with respect to variable names and dealing with missing values, abnormal data and so on.
2. Computing summary statistics and performing exploratory analysis of the data.
3. Discovering relationships between data in terms of graphical visualizations and correlations between variables.
4. Performing logistic regression and computing the accuracy of the model.

3.1. Importing and cleaning the data

We start with loading the required libraries

```
library(tidyverse)
library(ggplot2)
library(dplyr)
library(tidyr)
library(caret)
library(caTools)
library(pROC)
library(reshape2)
library(corrplot)
library(ggcorrplot)
```

Then we load the dataset

```
setwd("C:\\Users\\Dell\\Desktop")
data <- read.csv("data.csv")
```

After that we print the dimensions of the data

```
nrow(data)
[1] 569
ncol(data)
[1] 32
```

The structure of the data is as follows

```
str(data)


'data.frame':     569 obs. of  32 variables:
 $ id                      : int  842302 842517 84300903 84348301 84358402
843786 844359 84458202 844981 84501001 ...
 $ diagnosis               : chr  "M" "M" "M" "M" ...
 $ radius_mean             : num  18 20.6 19.7 11.4 20.3 ...
 $ texture_mean            : num  10.4 17.8 21.2 20.4 14.3 ...
 $ perimeter_mean          : num  122.8 132.9 130 77.6 135.1 ...
$ area_mean               : num  1001 1326 1203 386 1297 ...
 $ smoothness_mean         : num  0.1184 0.0847 0.1096 0.1425 0.1003 ...
 $ compactness_mean        : num  0.2776 0.0786 0.1599 0.2839 0.1328 ...
 $ concavity_mean          : num  0.3001 0.0869 0.1974 0.2414 0.198 ....
$ concave.points_mean     : num  0.1471 0.0702 0.1279 0.1052 0.1043 ...
 $ symmetry_mean           : num  0.242 0.181 0.207 0.26 0.181 ...
$ fractal_dimension_mean : num  0.0787 0.0567 0.06 0.0974 0.0588 ...
 $ radius_se               : num  1.095 0.543 0.746 0.496 0.757 ...
$ texture_se              : num  0.905 0.734 0.787 1.156 0.781 ...
 $ perimeter_se            : num  8.59 3.4 4.58 3.44 5.44 ...
```

```
$ area_se                : num  153.4 74.1 94 27.2 94.4 ...
 $ smoothness_se          : num  0.0064 0.00522 0.00615 0.00911 0.01149 ...
 $ compactness_se         : num  0.049 0.0131 0.0401 0.0746 0.0246 ...
 $ concavity_se           : num  0.0537 0.0186 0.0383 0.0566 0.0569 ...
$ concave.points_se      : num  0.0159 0.0134 0.0206 0.0187 0.0188 ...
 $ symmetry_se            : num  0.03 0.0139 0.0225 0.0596 0.0176 ...
 $ fractal_dimension_se   : num  0.00619 0.00353 0.00457 0.00921 0.00511 ...
 $ radius_worst           : num  25.4 25 23.6 14.9 22.5 ...
 $ texture_worst          : num  17.3 23.4 25.5 26.5 16.7 ...
 $ perimeter_worst        : num  184.6 158.8 152.5 98.9 152.2 ...
 $ area_worst             : num  2019 1956 1709 568 1575 ...
 $ smoothness_worst       : num  0.162 0.124 0.144 0.21 0.137 ...
 $ compactness_worst      : num  0.666 0.187 0.424 0.866 0.205 ...
 $ concavity_worst        : num  0.712 0.242 0.45 0.687 0.4 ...
 $ concave.points_worst   : num  0.265 0.186 0.243 0.258 0.163 ...
 $ symmetry_worst         : num  0.46 0.275 0.361 0.664 0.236 ...
 $ fractal_dimension_worst: num  0.1189 0.089 0.0876 0.173 0.0768 ...
```

The id column is unnecessary to our purpose so we eliminate it

```
data$id <- NULL
```

We preview the dataset

```
head(data)
  diagnosis radius_mean texture_mean perimeter_mean area_mean smoothness_mean
compactness_mean
1         M       17.99        10.38         122.80    1001.0         0.11840
0.27760
2         M       20.57        17.77         132.90    1326.0         0.08474
0.07864
3         M       19.69        21.25         130.00    1203.0         0.10960
0.15990
4         M       11.42        20.38          77.58     386.1         0.14250
0.28390
5         M       20.29        14.34         135.10    1297.0         0.10030
0.13280
6         M       12.45        15.70          82.57     477.1         0.12780
0.17000
concavity_mean concave.points_mean symmetry_mean fractal_dimension_mean
radius_se texture_se
1         0.3001             0.14710        0.2419                0.07871
1.0950     0.9053
2         0.0869             0.07017        0.1812                0.05667
0.5435     0.7339
3         0.1974             0.12790        0.2069                0.05999
0.7456     0.7869
4         0.2414             0.10520        0.2597                0.09744
0.4956     1.1560
5         0.1980             0.10430        0.1809                0.05883
0.7572     0.7813
6         0.1578             0.08089        0.2087                0.07613
0.3345     0.8902
```

```
perimeter_se area_se smoothness_se compactness_se concavity_se
concave.points_se symmetry_se
1        8.589  153.40      0.006399        0.04904      0.05373
0.01587     0.03003
2        3.398   74.08      0.005225        0.01308      0.01860
0.01340     0.01389
3        4.585   94.03      0.006150        0.04006      0.03832
0.02058     0.02250
4        3.445   27.23      0.009110        0.07458      0.05661
0.01867     0.05963
5        5.438   94.44      0.011490        0.02461      0.05688
0.01885     0.01756
6        2.217   27.19      0.007510        0.03345      0.03672
0.01137     0.02165
  fractal_dimension_se radius_worst texture_worst perimeter_worst area_worst
smoothness_worst
1              0.006193        25.38         17.33          184.60     2019.0
0.1622
2              0.003532        24.99         23.41          158.80     1956.0
0.1238
3              0.004571        23.57         25.53          152.50     1709.0
0.1444
4              0.009208        14.91         26.50           98.87      567.7
0.2098
5              0.005115        22.54         16.67          152.20     1575.0
0.1374
6              0.005082        15.47         23.75          103.40      741.6
0.1791
compactness_worst concavity_worst concave.points_worst symmetry_worst
fractal_dimension_worst
1            0.6656          0.7119               0.2654         0.4601
0.11890
2            0.1866          0.2416               0.1860         0.2750
0.08902
3            0.4245          0.4504               0.2430         0.3613
0.08758
4            0.8663          0.6869               0.2575         0.6638
0.17300
5            0.2050          0.4000               0.1625         0.2364
```

3.2. Descriptive Analysis

The summary statistics of the dataset is as follows

```
summary(data)


 diagnosis  radius_mean      texture_mean   perimeter_mean     area_mean
smoothness_mean
 B:357     Min.   : 6.981   Min.   : 9.71   Min.   : 43.79   Min.   : 143.5
Min.   :0.05263
 M:212     1st Qu.:11.700   1st Qu.:16.17   1st Qu.: 75.17   1st Qu.: 420.3
1st Qu.:0.08637
           Median :13.370   Median :18.84   Median : 86.24   Median : 551.1
Median :0.09587
           Mean   :14.127   Mean   :19.29   Mean   : 91.97   Mean   : 654.9
Mean   :0.09636
           3rd Qu.:15.780   3rd Qu.:21.80   3rd Qu.:104.10   3rd Qu.: 782.7
3rd Qu.:0.10530
           Max.   :28.110   Max.   :39.28   Max.   :188.50   Max.   :2501.0
Max.   :0.16340
compactness_mean  concavity_mean    concave.points_mean symmetry_mean
fractal_dimension_mean
 Min.   :0.01938   Min.   :0.00000   Min.   :0.00000     Min.   :0.1060   Min.
:0.04996
 1st Qu.:0.06492   1st Qu.:0.02956   1st Qu.:0.02031     1st Qu.:0.1619   1st
Qu.:0.05770
 Median :0.09263   Median :0.06154   Median :0.03350     Median :0.1792   Median
:0.06154
 Mean   :0.10434   Mean   :0.08880   Mean   :0.04892     Mean   :0.1812   Mean
:0.06280
 3rd Qu.:0.13040   3rd Qu.:0.13070   3rd Qu.:0.07400     3rd Qu.:0.1957   3rd
Qu.:0.06612
 Max.   :0.34540   Max.   :0.42680   Max.   :0.20120     Max.   :0.3040   Max.
:0.09744
radius_se        texture_se      perimeter_se       area_se         smoothness_se
 Min.   :0.1115   Min.   :0.3602   Min.   : 0.757   Min.   :  6.802   Min.
:0.001713
 1st Qu.:0.2324   1st Qu.:0.8339   1st Qu.: 1.606   1st Qu.: 17.850   1st
Qu.:0.005169
 Median :0.3242   Median :1.1080   Median : 2.287   Median : 24.530   Median
:0.006380
 Mean   :0.4052   Mean   :1.2169   Mean   : 2.866   Mean   : 40.337   Mean
:0.007041
 3rd Qu.:0.4789   3rd Qu.:1.4740   3rd Qu.: 3.357   3rd Qu.: 45.190   3rd
Qu.:0.008146
 Max.   :2.8730   Max.   :4.8850   Max.   :21.980   Max.   :542.200   Max.
:0.031130
```

```
compactness_se      concavity_se     concave.points_se   symmetry_se
fractal_dimension_se
 Min.   :0.002252   Min.   :0.00000   Min.   :0.000000   Min.
:0.007882   Min.   :0.0008948
 1st Qu.:0.013080   1st Qu.:0.01509   1st Qu.:0.007638   1st
Qu.:0.015160   1st Qu.:0.0022480
 Median :0.020450   Median :0.02589   Median :0.010930   Median
:0.018730   Median :0.0031870
 Mean   :0.025478   Mean   :0.03189   Mean   :0.011796   Mean
:0.020542   Mean   :0.0037949
 3rd Qu.:0.032450   3rd Qu.:0.04205   3rd Qu.:0.014710   3rd
Qu.:0.023480   3rd Qu.:0.0045580
 Max.   :0.135400   Max.   :0.39600   Max.   :0.052790   Max.
:0.078950   Max.   :0.0298400
radius_worst   texture_worst   perimeter_worst    area_worst
smoothness_worst
 Min.   : 7.93   Min.   :12.02   Min.   : 50.41   Min.   : 185.2   Min.
:0.07117
 1st Qu.:13.01   1st Qu.:21.08   1st Qu.: 84.11   1st Qu.: 515.3   1st
Qu.:0.11660
 Median :14.97   Median :25.41   Median : 97.66   Median : 686.5
Median :0.13130
 Mean   :16.27   Mean   :25.68   Mean   :107.26   Mean   : 880.6   Mean
:0.13237
 3rd Qu.:18.79   3rd Qu.:29.72   3rd Qu.:125.40   3rd Qu.:1084.0   3rd
Qu.:0.14600
 Max.   :36.04   Max.   :49.54   Max.   :251.20   Max.   :4254.0   Max.
:0.22260
 compactness_worst concavity_worst  concave.points_worst symmetry_worst
fractal_dimension_worst
 Min.   :0.02729   Min.   :0.0000   Min.   :0.00000      Min.   :0.1565
Min.   :0.05504
 1st Qu.:0.14720   1st Qu.:0.1145   1st Qu.:0.06493      1st Qu.:0.2504
1st Qu.:0.07146
 Median :0.21190   Median :0.2267   Median :0.09993      Median :0.2822
Median :0.08004
 Mean   :0.25427   Mean   :0.2722   Mean   :0.11461      Mean   :0.2901
Mean   :0.08395
 3rd Qu.:0.33910   3rd Qu.:0.3829   3rd Qu.:0.16140      3rd Qu.:0.3179
3rd Qu.:0.09208
 Max.   :1.05800   Max.   :1.2520   Max.   :0.29100      Max.   :0.6638
Max.   :0.20750
```

In the results displayed, there are 569 records each with 31 columns. Diagnosis is a Categorical variable. Class distribution of the categorical variable: 357 Benign, 212 Malignant.

3.3. Univariate Plots

One of the main goals of visualizing the data here is to observe which features are most helpful in predicting malignancy of a tumor in breast cancer. Here we will analyze the features and try to understand which features have larger predictive value and which does not bring considerable predictive value if we want to create a model that allows us to guess if a tumor is benign or malignant.

From the data, we will visualize how the malignancy of the tumors is distributed with the help of **Pie Chart**. A Pie chart is a circular graph divided into slices to illustrate numerical proportions. Each slice represents a category's contribution to the whole, with the size of the slice corresponding to its percentage.

```
#Create a frequency table
diagnosistab <- table(data$diagnosis)

#Create a pie chart
diagnosis.p.tab <- prop.table(diagnosistab)*100
diagnosis.p.df <- as.data.frame(diagnosis.p.tab)
pielabels <- sprintf("%s - %3.1f%s", diagnosis.p.df[,1],
             diagnosis.p.tab, "%")

pie(diagnosis.p.tab,
    labels = pielabels,
    col = c( "salmon", "lightblue"),
    border = "gainsboro",
    main= "Distribution of Benign and Malignant tumours",
    clockwise = TRUE,
    radius = 1)

legend("topright", c("Benign", "Malignant"),
       fill = c( "salmon", "lightblue"),
       title = "Diagnosis",
       cex = 0.75)
```

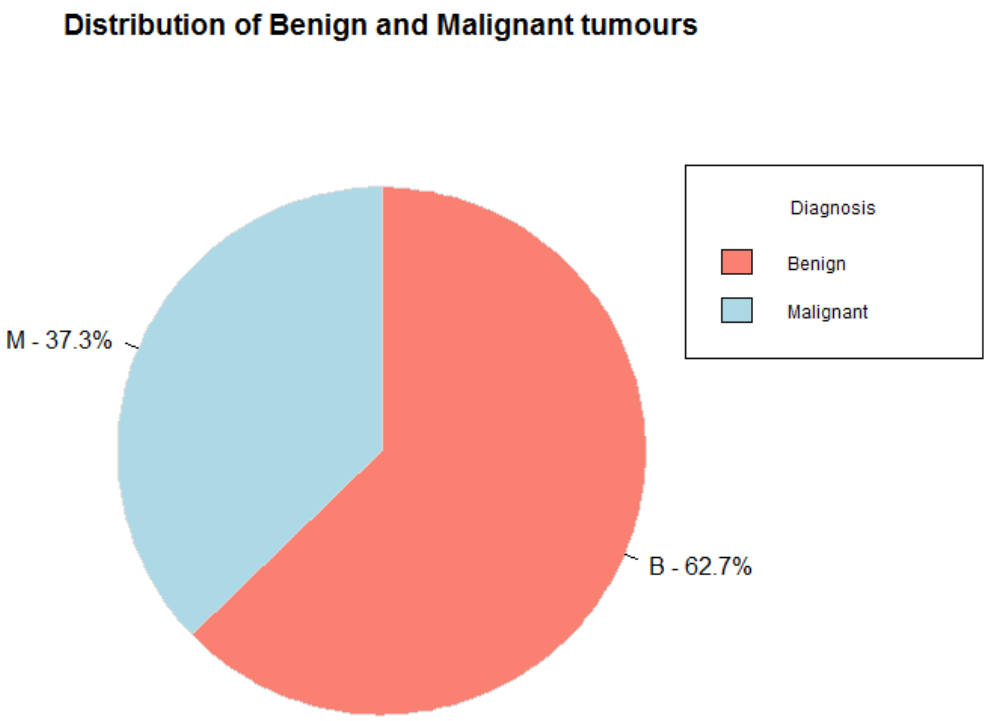

Figure 1: Distribution of Malignancy of tumors

M= Malignant (indicates presence of cancer cells)

B= Benign (indicates absence of cancer cells)

Here we can see that, 357 observations which account for 62.7% of all observations, indicate the absence of cancer cells; 212 which account for 37.3% of all observations, shows the presence of cancerous cells.

The percent for malignant tumors unusually large. The dataset does not represent a typical medical analysis distribution in this case. Typically, we get a considerably large number of benign tumors vs. a small number of cases that represents presence of malignant tumors.

Following this, we observe that we have data corresponding to the 10 parameters that we listed previously. Hence it is in our interest to visualize the values with the help of **Histograms** across the two categories of benign and malignant.

```
data_mean <- data[ , c("diagnosis", "radius_mean",    "texture_mean",
                       "perimeter_mean",  "area_mean","smoothness_mean",
                       "compactness_mean","concavity_mean",
                       "concave.points_mean", "symmetry_mean",
                       "fractal_dimension_mean")]

#plot histograms of "_mean" variables group by diagnosis
ggplot(data= melt(data_mean, id.var= "diagnosis"), mapping = aes(x=value))+
  geom_histogram(bins=10, aes(fill= diagnosis),colour="black",
alpha=0.5)+
  facet_wrap(~variable, scales = "free_x")+
  labs(title = "Histogram of the Following Parameters")+
  theme(plot.title = element_text(hjust = 0.5))
```

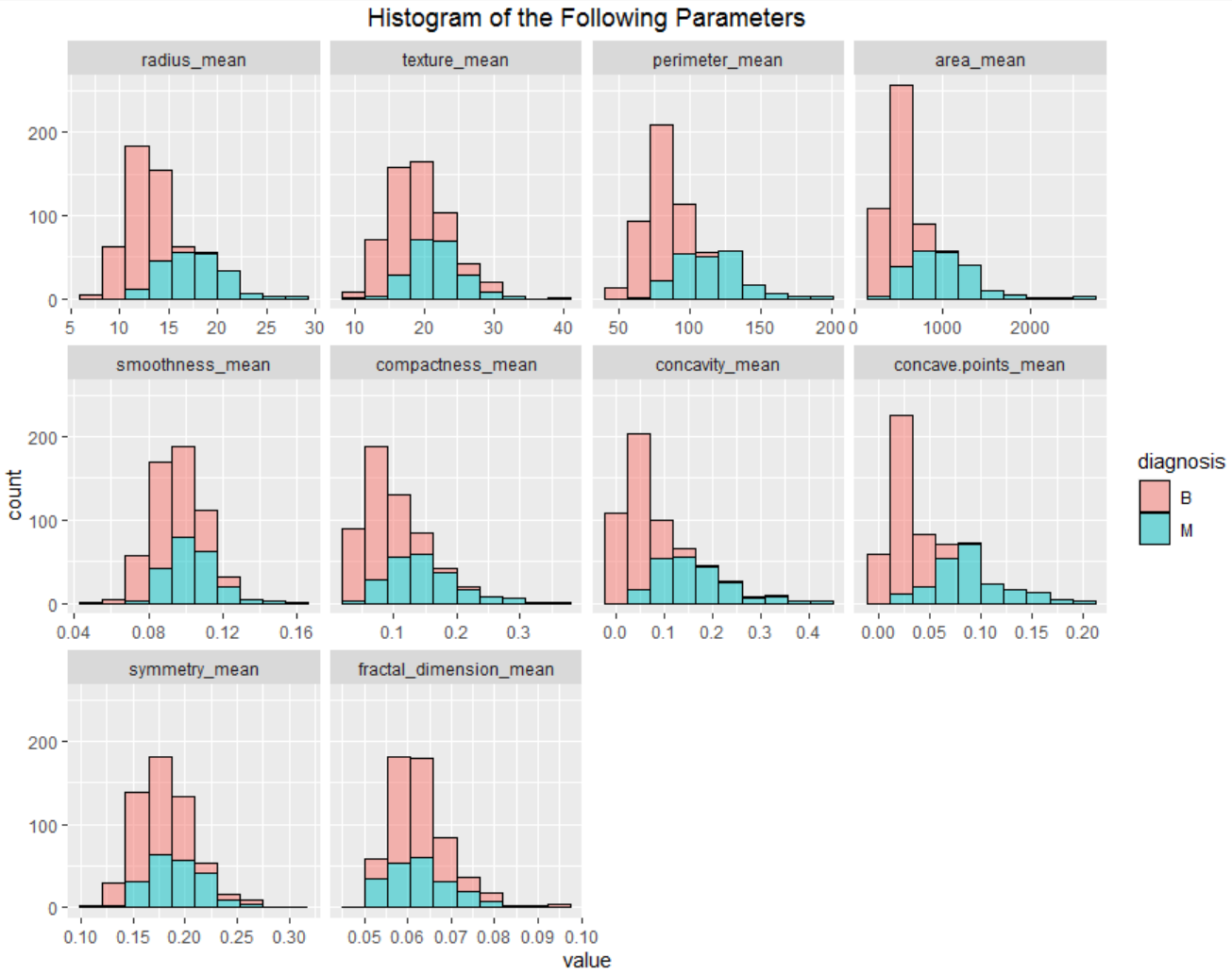

Figure 2: Histogram of the mean of the 10 concerned parameters separated by diagnosis

From the histogram of radius mean, it is observed that higher radius mean of sample cell nuclei corresponds to higher probability of malignancy of the tumor. As for the malignant cell samples, we can see that the radius mean follows normal distribution with mean 17.35 units. The total data ispositively skewed which shows lower number of people have high tumor radius mean in general.

The histogram of the texture mean shows that benign and malignant tumor cells superimpose in terms of texture mean. The entire data on texture mean is almost symmetric in nature, with a slight positive skewness.

The distribution of perimeter mean of tumor cells is similar to that of radius mean, where high perimeter mean corresponds to higher probability of malignancy. The distribution of malignant tumor cells is positively skewed which means lower number of people have very high perimeter mean. The entire distribution of the perimeter mean data is also highly skewed in nature.

Area mean of tumor cells also follows a positively skewed distribution. Higher area mean of tumor cells corresponds to higher probability of malignancy of the tumor. However, the malignant cells follow a normal distribution with mean at 1062.5 units.

Smoothness mean of both benign and malignant tumor cells super- impose on each other. They follow a normal distribution with mean at 0.1 units.

From the histogram, we can see benign and malignant tumor cells superimpose on each other in the range 0.00 to 0.27 units. Very high compactness mean corresponds to high probability of malignancy.

Concavity mean follows a positively skewed distribution, where lower number of cell nuclei have higher concavity mean. Moreover, higher values of concavity mean also corresponds to higher probability of malignancy of tumor cells.

Concave points mean also follow positively skewed distribution. Higher concave points mean corresponds to higher probability of malignancy of tumor. The malignant cells themselves follow normal distribution with mean at 0.1 units.

Symmetry mean of both benign and malignant tumor cells superimpose. They follow a normal distribution with mean at 0.18 units.

Fractal dimension also approximately follows a symmetric distribution. The malignant cells however follow positively skewed distribution. However in general, the benign and malignant cell observations superimpose.

```
data_se <- data[ , c("diagnosis", "radius_se",  "texture_se",
                     "perimeter_se",      "area_se",  "smoothness_se",
                     "compactness_se",    "concavity_se",
      "concave.points_se",
                     "symmetry_se", "fractal_dimension_se")]

#plot histograms of "_se" variables group by diagnosis
ggplot(data= melt(data_se, id.var= "diagnosis"), mapping =
aes(x=value))+
  geom_histogram(bins=10, aes(fill= diagnosis),colour="black", alpha=
0.5)+
  facet_wrap(~variable, scales = "free_x")+
  labs(title = "Histogram of the Following Parameters")+
  theme(plot.title = element_text(hjust = 0.5))
```

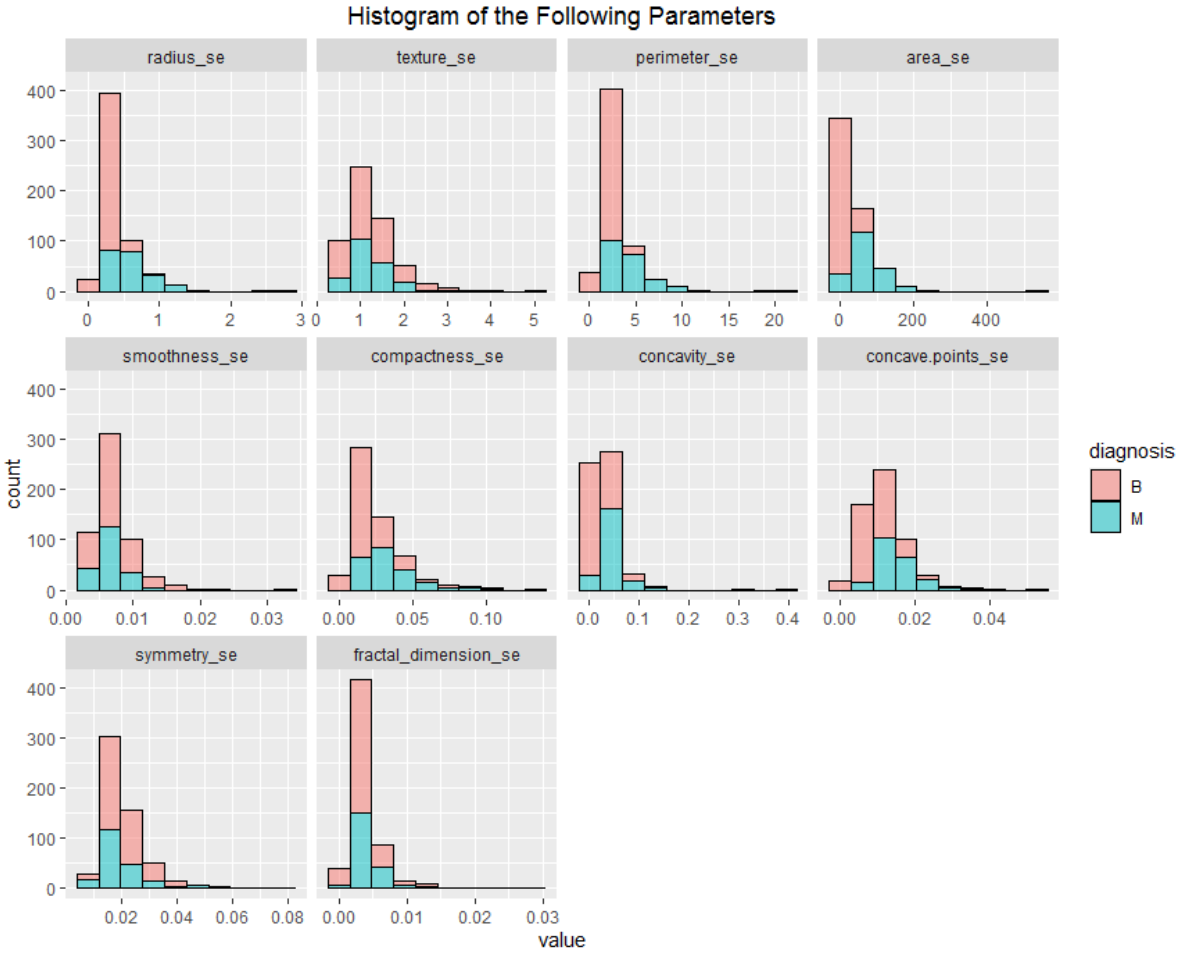

Figure 3: Histogram of the standard error of the 10 concerned parameters separated by diagnosis

From the histogram of radius se, it is observed that higher radius se of sample cell nuclei corresponds to higher probability of malignancy of the tumor. Highly right-skewed distribution for both benign and malignant cases. There's some overlap, but a clear separation between benign and malignant distributions.

The histogram of the texture se shows that there is considerable overlap between benign and malignant cases. Histogram is less skewed compared to radius se, but still right-skewed. Malignant tumors show a slightly wider spread of values.

The distribution of perimeter se of tumor cells is similar to that of radius se, where high perimeter se corresponds to higher probability of malignancy. Benign cases cluster tightly at lower values and malignant cases show more variability.

The histogram of the texture se shows that there is extremely right-skewed distribution for the benign cases. Benign cases are tightly clustered at the lower end of the scale. There is a notable separation between benign and malignant cases.

The distribution of smoothness se of tumor cells shows that there is significant overlap between benign and malignant distributions. Also, malignant cases show a slightly wider spread of values.

The histogram of the compactness se shows right-skewed distribution for both classes. There is some separation between benign and malignant cases, but with considerable overlap. Malignant tumors tend towards higher compactness se values. Benign cases peak at lower values compared to malignant cases.

For concavity se, the histogram shows highly right-skewed for both classes. Many benign cases cluster near zero, suggesting low variability in concavity.

The histogram of the concave points se shows clear separation between benign and malignant, with some overlap. Malignant cases show higher values and more spread and benign cases cluster more tightly at lower values.

For symmetry se, we see a significant overlap between benign and malignant distributions i.e. the classes are less pronounced.

The histogram of fractal dimensions se shows very less separation between the classes.

```
data_worst <- data[ , c("diagnosis", "radius_worst",  "texture_worst",
                        "perimeter_worst",      "area_worst",
      "smoothness_worst",
                        "compactness_worst",    "concavity_worst",
                        "concave.points_worst", "symmetry_worst",
                        "fractal_dimension_worst")]

#plot histograms of "_worst" variables group by diagnosis
ggplot(data= melt(data_worst, id.var= "diagnosis"), mapping =
aes(x=value))+
  geom_histogram(bins=10, aes(fill= diagnosis),colour="black", alpha=
0.5)+
  facet_wrap(~variable, scales = "free_x")+
  labs(title = "Histogram of the Following Parameters")+
  theme(plot.title = element_text(hjust = 0.5))
```

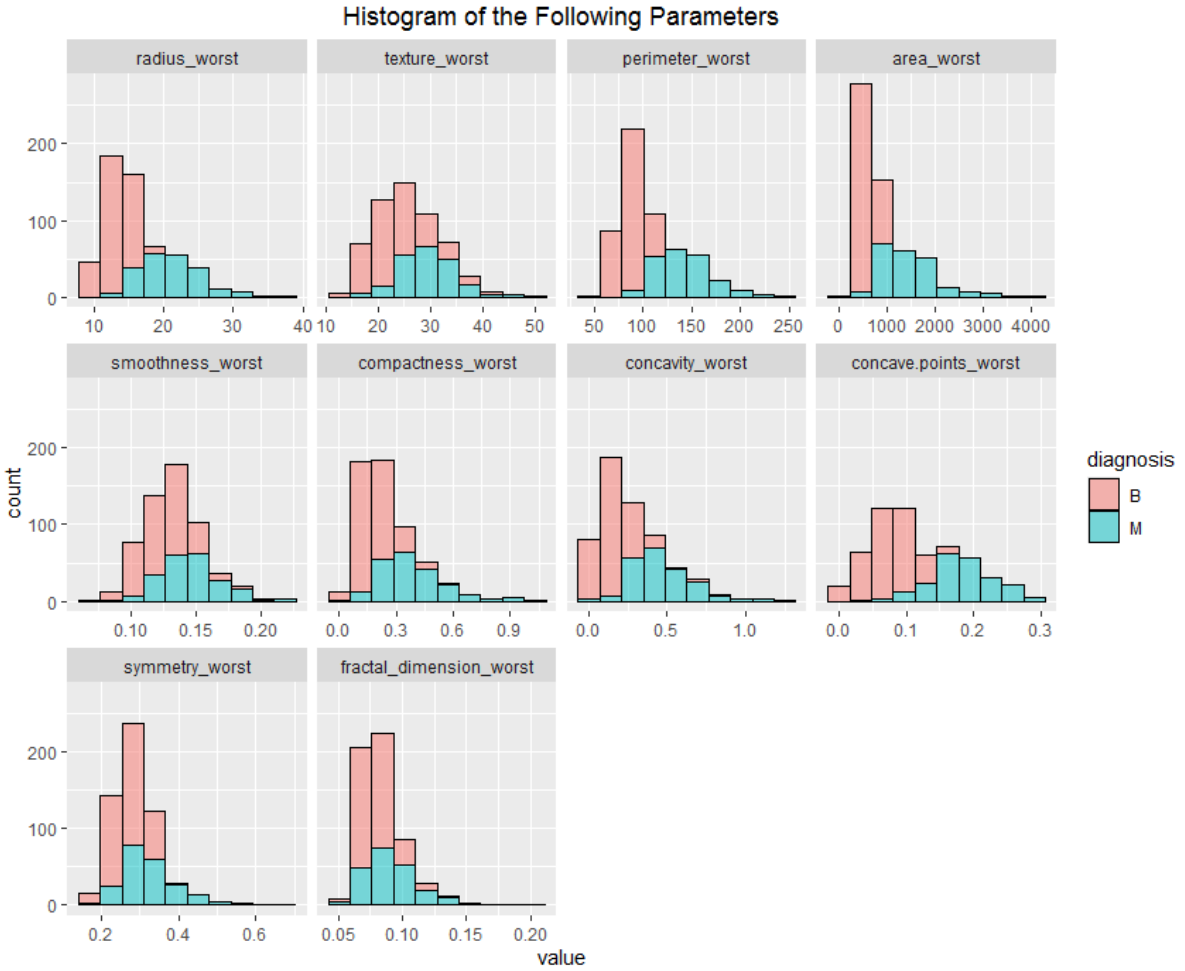

Figure 4: Histogram of the worst values of the 10 concerned parameters separated by diagnosis

From the histogram of radius_worst, we see clear separation between benign (B) and malignant (M) tumors. Malignant tumors tend to have larger worst radius values. Benign tumors are more clustered at lower values, while malignant tumors show a wider spread.

For texture_worst, there is some overlap between benign and malignant tumors. Malignant tumors generally have higher texture_worst values. The distribution is more symmetric compared to radius_worst.

For perimeter_worst, malignant tumors have significantly larger worst perimeter values. Also the histogram has very similar pattern to radius_worst.

For area_worst, malignant tumors show much larger worst area values. Highly right-skewed distribution for both classes, more extreme for malignant. Benign tumors cluster tightly at lower values.

For smoothness_worst, there is considerable overlap between benign and malignant tumors. Slight tendency for malignant tumors to have higher smoothness_worst values. More symmetric distribution compared to size-related features.

For compactness_worst, there is some separation between benign and malignant, but with overlap. It is more distinctive than smoothness_worst, but less than size-related features.

For concavity_worst, we see that many benign tumors cluster near zero concavity_worst. Histogram is highly right-skewed, especially for benign tumors.

For concave.points_worst, we see distinct separation between benign and malignant tumors which makes it potentially strong predictor for malignancy.

For symmetry_worst, we see considerable overlap between benign and malignant tumors which makes it less discriminative compared to many other "worst" features.

For fractal_dimension_worst, we see significant overlap between benign and malignant tumors making it the least distinctive separation among all "worst" features shown.

Comparison of distribution by malignancy shows that there is no perfect separation between any of the features. Although we do have fairly good separations for concave.points worst, concavity_worst, perimeter_worst, area_mean, perimeter_mean. We do have as well tight superposition for some of the values, like symmetry se, smoothness se.

3.4. Graphical visualization of relationships between multiple variables

We are also interested in how the 30 predictor variables relate to each other. To see bivariate relationships among these 30 predictor variables, we will look at their correlation coefficients.

```
#correlation matrix
cormat <- cor(data[ , 2:31])
corrplot(cormat,method = "square", order= "hclust",
         tl.cex = 0.7, tl.col = "black")
```

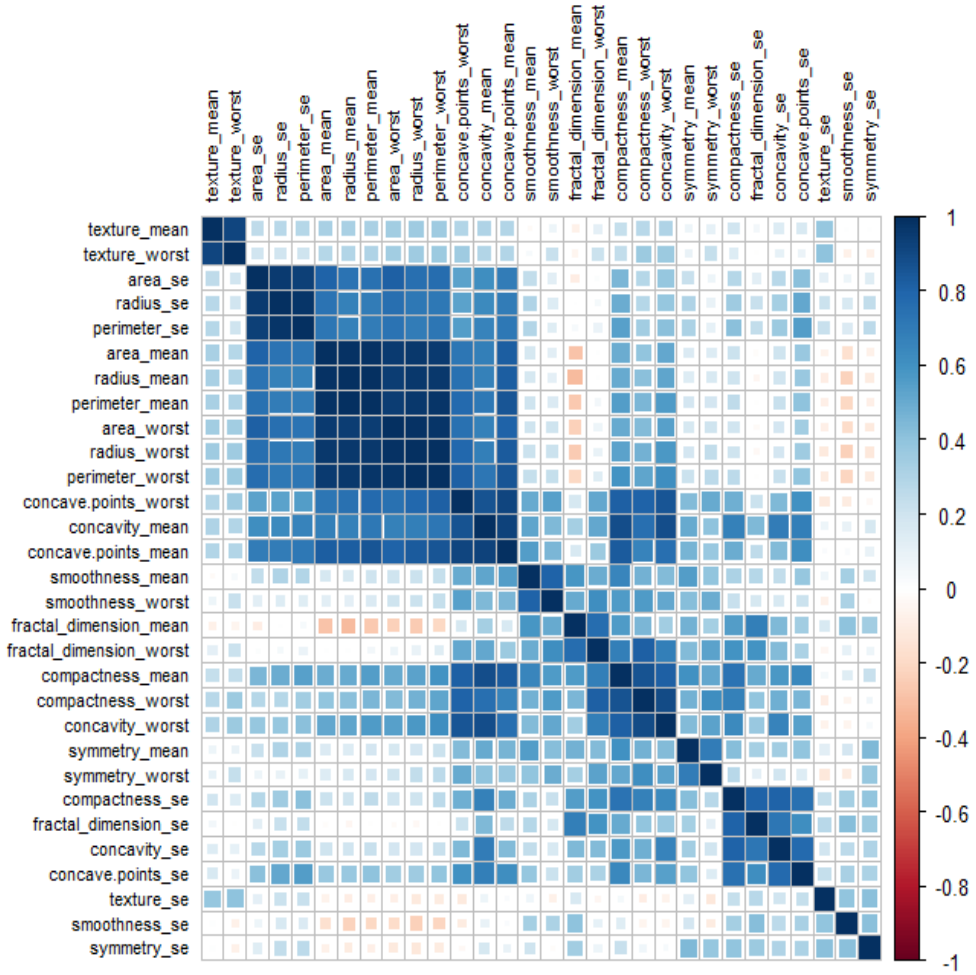

Figure 5: Correlation matrix of the variables

There are quite a few variables that are highly correlated to each other. We setup a threshold of 0.9 and look for the variables which have the correlation coefficient greater than 0.9.

```
highlyCor <- findCorrelation(cormat, cutoff = 0.9, verbose = T, names =
T)
Compare row 7  and column  8 with corr  0.921
  Means:  0.571 vs 0.389 so flagging column 7
Compare row 8  and column  28 with corr  0.91
  Means:  0.542 vs 0.377 so flagging column 8
Compare row 23  and column  21 with corr  0.994
  Means:  0.48 vs 0.367 so flagging column 23
Compare row 21  and column  3 with corr  0.969
  Means:  0.446 vs 0.359 so flagging column 21
Compare row 3  and column  24 with corr  0.942
  Means:  0.414 vs 0.353 so flagging column 3
Compare row 24  and column  1 with corr  0.941
  Means:  0.39 vs 0.349 so flagging column 24
Compare row 1  and column  4 with corr  0.987
  Means:  0.35 vs 0.347 so flagging column 1
Compare row 13  and column  11 with corr  0.973
  Means:  0.372 vs 0.346 so flagging column 13
Compare row 11  and column  14 with corr  0.952
  Means:  0.323 vs 0.347 so flagging column 14
Compare row 22  and column  2 with corr  0.912
  Means:  0.224 vs 0.357 so flagging column 2
All correlations <= 0.9
```

So the following variables are very highly correlated to each other.

```
highlyCor

 [1] "concavity_mean"      "concave.points_mean" "perimeter_worst"
"radius_worst"
 [5] "perimeter_mean"      "area_worst"          "radius_mean"
"perimeter_se"
 [9] "area_se"             "texture_mean"
```

The above mentioned predictor variables are very highly correlated to each other. But there are other predictor variables in the dataset that are correlated to each other. This leads to the infamous Multicollinearity problem.

3.5. Logistic Regression

A logistic regression is a special case of regression analysis which is used whenever the dependent variable is categorical.

Logistic regression estimates the probability of an event occurring, based on a given dataset of independent variables, and not the actual value of the variable. This type of statistical model is often used for classification and predictive analytics. Since the outcome is a probability, the dependent variable is bounded between 0 and 1. In order to do this we use the sigmoid function (a special case of the logistic function),

$$sig(t) = \frac{1}{1 + e^{t}}$$

This is a positive valued function bounded between 0 and 1, as seen in Figure 6 below.

In our case, when the linear regression of y on x is given by $\hat{y} = b_0 + b_1x_1 + b_2x_2 + \cdots + b_kx_k$, we enter this y in place of t in the sigmoid function and we determine the coefficients $b_1, b_2, \ldots b_k$.

$$f(x) = \frac{1}{1 + e^{-y}} = \frac{1}{1 + e^{-(b_0 + b_1x_1 + b_2x_2 + \cdots + b_kx_k)}}$$

In logistic regression, a logit transformation is applied on the odds which is the probability of success divided by the probability of failure. This is also known as the log odds and this logistic function is represented by the formula

$$Logit\big(f(x)\big) = \frac{1}{1 + e^{-f(x)}}$$

$$\ln\left(\frac{f(x)}{1 - f(x)}\right) = b_0 + b_1x_1 + b_2x_2 + \cdots + b_kx_k$$

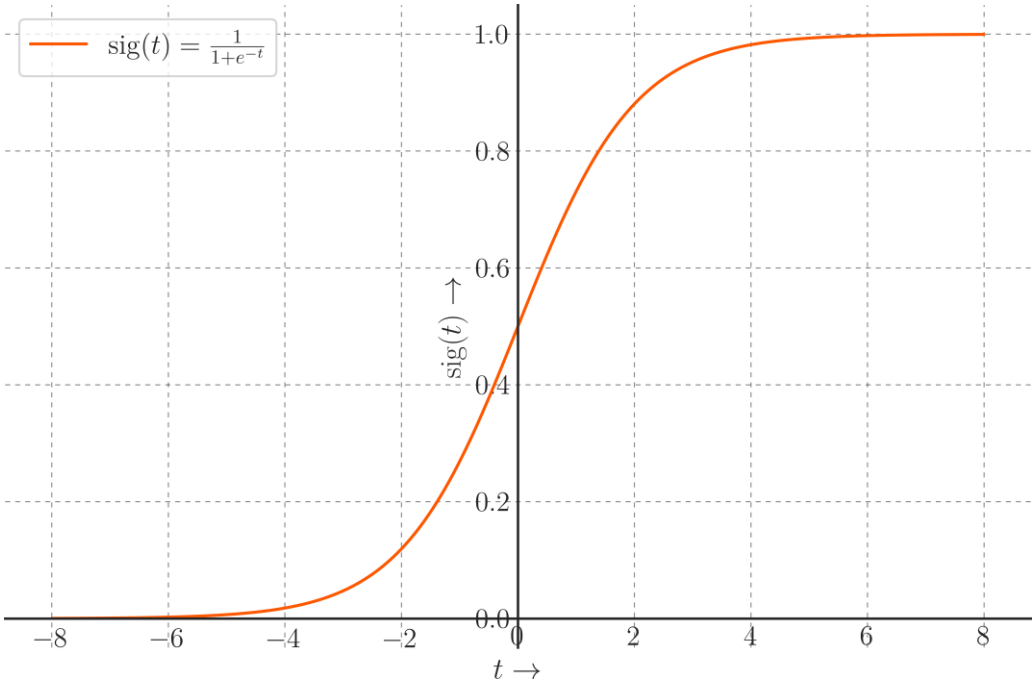

Figure 6: Sigmoid function

In this logistic regression equation, $Logit(f(x))$ is the dependent or response variable and x is the independent variable. The coefficients in this model are commonly estimated via maximum likelihood estimation (MLE). This method tests different values of b_i through multiple iterations to optimize for the best fit of log odds. All of these iterations produce the log likelihood function, and logistic regression seeks to maximize this function to find the best parameter estimate. Once the optimal coefficient (or coefficients if there is more than one independent variable) is found, the conditional probabilities for each observation can be calculated, logged, and summed together to yield a predicted probability.

For binary classification, a probability less than 0.5 will predict 0 while a probability greater than 0.5 will predict 1. After the model has been computed we use the confusion matrix to check for the accuracy of the model.

3.6. Fitting of Logistic Regression Model

In order to predict the diagnosis given the data, we need to fit a logistic regression model to the data. In order to do that we need to convert the “diagnosis” column into a factor as it is a categorical variable.

Also we split the dataset into two data sets: a training set and a test set. The purpose of the training set is to train the model on this data. Then we test the model on the “unknown data” which is in the test set.

```
#converting diagnosis to a factor
data$diagnosis <- as.factor(data$diagnosis)

#split the data into training and test set
split <- sample.split(data, SplitRatio = 0.8)
split
train_data <- subset(data, split)
test_data <- subset(data, !split)
```

Next we fit a generalized linear model, where we specify that we want to predict “diagnosis” with the help of the other variables. The parameter “binomial” signifies that it is a logistic regression problem.

```
#Fit logistic regression model
logistic_model <- glm(diagnosis~., data = train_data, family = binomial)

#predict probabilities on test set
test_data$predictions <- predict(logistic_model, newdata = test_data,
type = "response")

#Convert probabilities to class labels
predicted_classes <- ifelse(test_data$predictions> 0.5, "M", "B")
```

Next we generate the **Confusion Matrix**. It is a simple table which is used to describe the performance of a classifier by providing the numbers of correctly and incorrectly classified observations across the categories. For a binary classification problem, the table looks as below:

	Predicted: NO	Predicted: YES
Actual: NO	TN	FP
Actual: NO	FN	TP

Table 1

Here,

- TN: True Negative
- FP: False Positive
- FN: False Negative
- TP: True Positive

Following this, accuracy is computed as:

$$accuracy = \frac{TP + TN}{TN + FP + FN + TP}$$

In code:

```
#Confusion matrix
confusion_matrix <- table(predicted_classes, test_data$diagnosis)
print(confusion_matrix)

predicted_classes  B  M
                B 74  3
                M  3 49

#calculate accuracy
accuracy <- sum(diag(confusion_matrix))/ sum(confusion_matrix)
print(paste("Accuracy:", accuracy))

[1] "Accuracy: 0.953488372093023"
```

Thus we get ≈ 95% accuracy.

Remark

We get a satisfactory high accuracy of **95%**. But to get this accuracy, we had to consider 30 predictor variables. So we can state that this process is not very efficient as we need 30 independent characteristics of the tumor cells to fit this model. Also from Figure 5, we get that many predictor variables are highly correlated to each other. This leads to Multicollinearity.

Multicollinearity prevents predictive models from producing accurate predictions by increasing model complexity and overfitting. So we need to reduce the dimensions of the dataset.

We can reduce the dimension by using only the principal components that retain maximum variance. Here we shall use the famous dimension reduction algorithm **Principal Component Analysis (PCA)** on the dataset with 30 predictor variables and again perform the logistic regression model on it.

The algorithm of PCA centers and rescale the individual variables. It constructs a set of orthogonal (non-linear, uncorrelated, independent) variables. Principal Component Analysis ensures dimension reduction after more or less retaining the equal amount of variability.

```
x<- data %>% select(-diagnosis)
y<- data$diagnosis

#we standardize the data
x_scaled <- scale(x)

#apply pca
pca<- prcomp(x_scaled, center =T, scale. = T)

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

```
#standard deviation of the components
pca$sdev
[1] 3.64439401 2.38565601 1.67867477 1.40735229 1.28402903 1.09879780 0.82171778
0.69037464
 [9] 0.64567392 0.59219377 0.54213992 0.51103950 0.49128148 0.39624453 0.30681422
0.28260007
[17] 0.24371918 0.22938785 0.22243559 0.17652026 0.17312681 0.16564843 0.15601550
0.13436892
[25] 0.12442376 0.09043030 0.08306903 0.03986650 0.02736427 0.01153451

#getting the proportion of variance
prop_var <- pca$sdev^2 / sum(pca$sdev^2)

#proportion of variance
plot(prop_var, type = "b", xlab = "Components",
     ylab = "Proportion of Variance",
     main = "Variance Explained by Principal Components")+
  grid(nx = 20, ny = 25)
```

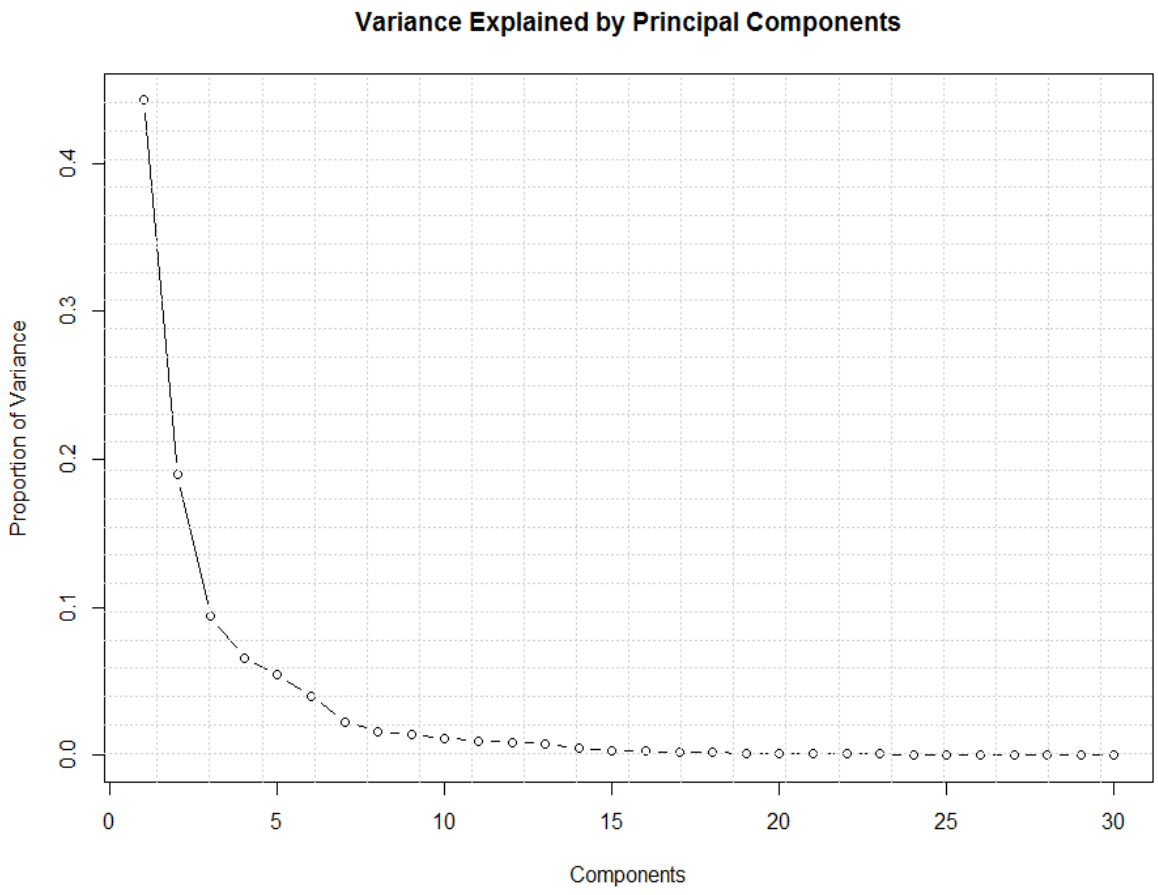

Figure 7: Principal Components Weight

```
#cumulative proportion of variance

cum_prop_var <- cumsum(prop_var)
plot(cum_prop_var, type = "b", xlab = "Components",
     ylab = "Cumulative Proportion of Variance",
     main = "Cumulative Variance Explained by Principal Components")+
  grid(nx = 20, ny = 20)+
  abline(h = 0.95, col = "red", lty=2)+
  abline(h = 0.99, col = "blue", lty=2)
```

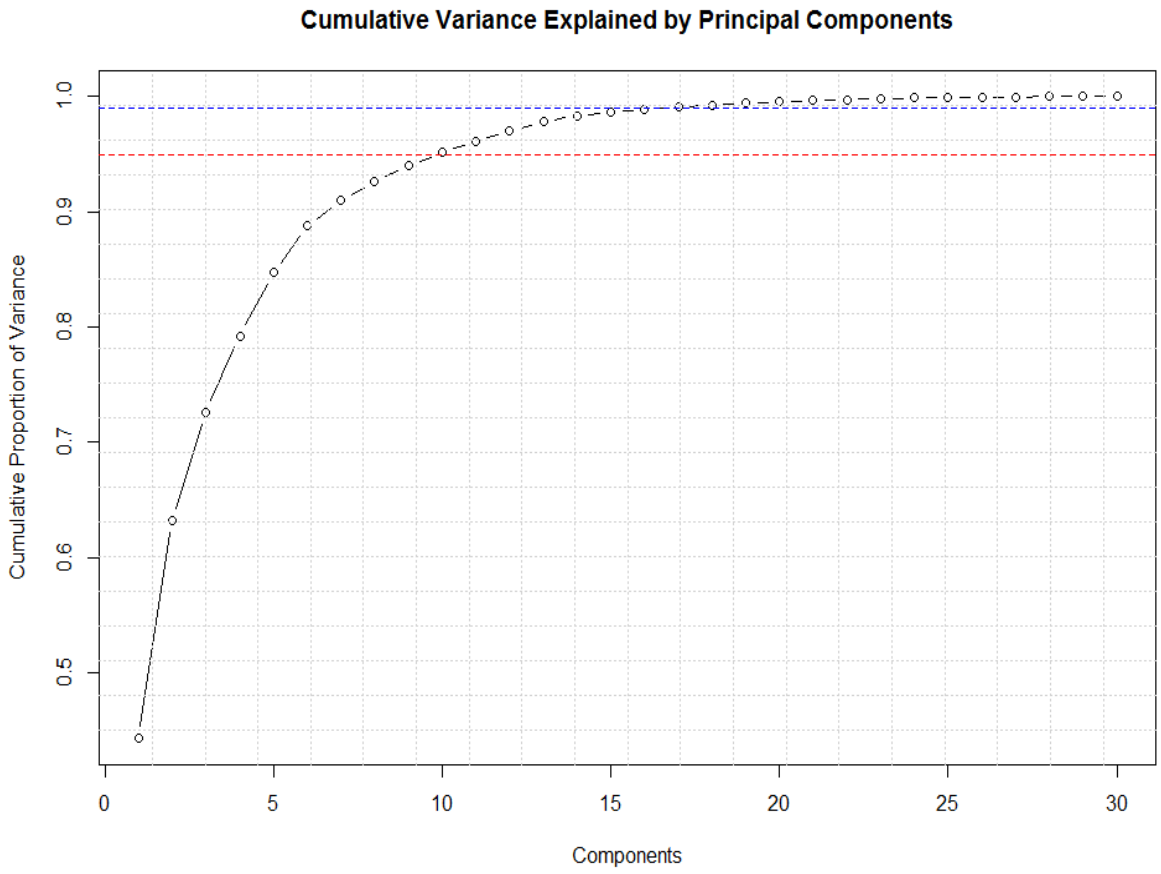

Figure 8: Cumulative Proportion of Variance

From Figure 8, we can see that there are 10 principal components that cover almost 95% (the red dashed line represents the 95% threshold) of the total variability and 17 principal components to cover almost 99% (the blue dashed line represents the 99% threshold) of the total variability.

Now we will find out how much of the variance is explained by the principal components.

```
#percentage of variance explained
pc_var <- pca$sdev^2
pr_var <- pc_var/sum(pc_var)
cat("Principal components with maximum variance:\n")
for (i in 1: length(pr_var)) {
  cat("PC", i, ":", pr_var[i], "\n")
  }
```

```
PC 1 : 0.4427203
PC 2 : 0.1897118
PC 3 : 0.09393163
PC 4 : 0.06602135
PC 5 : 0.05495768
PC 6 : 0.04024522
PC 7 : 0.02250734
PC 8 : 0.01588724
PC 9 : 0.01389649
PC 10 : 0.01168978
PC 11 : 0.00979719
PC 12 : 0.008705379
PC 13 : 0.00804525
PC 14 : 0.005233657
PC 15 : 0.003137832
PC 16 : 0.002662093
PC 17 : 0.001979968
PC 18 : 0.001753959
PC 19 : 0.001649253
PC 20 : 0.001038647
PC 21 : 0.0009990965
PC 22 : 0.0009146468
PC 23 : 0.0008113613
PC 24 : 0.0006018336
PC 25 : 0.0005160424
PC 26 : 0.000272588
PC 27 : 0.0002300155
PC 28 : 5.297793e-05
PC 29 : 2.49601e-05
PC 30 : 4.434827e-06
```

Now we visualize the percentage of variances explained by the first 10 principal components.

```
#scree plot of variance

fviz_eig(pca, addlabels = T, xlab ="Principal Components",
         ylim= c(0,50))+
  theme(plot.title = element_text(hjust = 0.5))
```

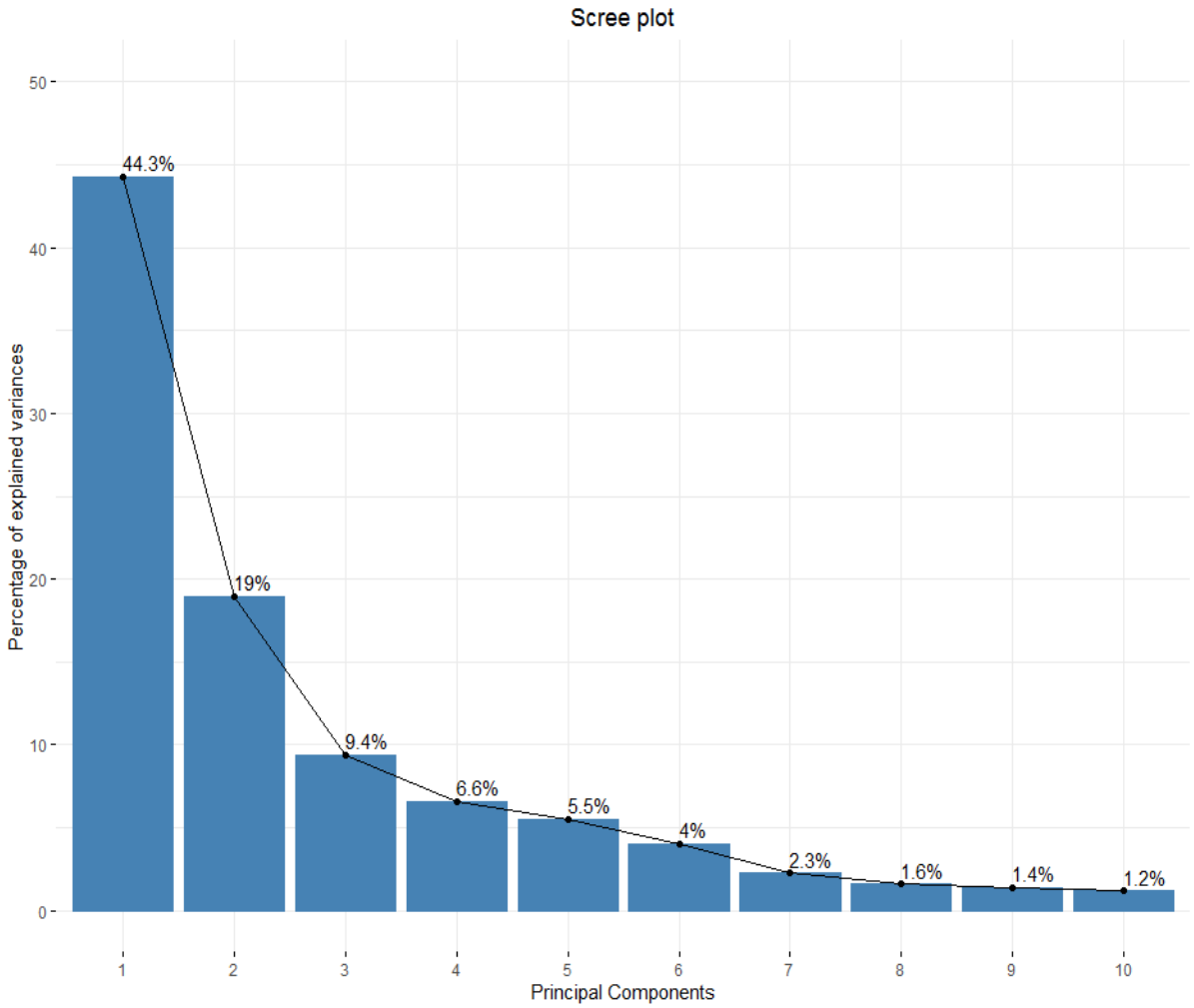

Figure 9: Percentage of Variance Explained by the PC's

Here, we notice that the first two Principal Components cover (44.3+19)=63.3% of the total variability.

After this, we calculate the percentage of variability covered by the 10 Principal Components.

```
#check the explained variance
ex_var <- sum(summary(pca)$importance[2,1:10])
print(paste("Explained variance by the first 10 principal components: ",
            round(ex_var*100, 2),"%"))

[1] "Explained variance by the first 10 principal components:  95.16 %"
```

We can see that, the first 10 Principal Components cover 95.16% of the total variability. In Figure 8, we already mentioned and graphically represented this finding. Here, our goal was to obtain the exact numerical value.

Now, we construct a new data frame with 10 PC's (Principal Components) which explain the most of the variance and we fit the logistic regression on the new data frame

```
#creating a data frame with the first 10 principal components
pca_data<- pca$x[,1:10]
pca_data <- as.data.frame(pca_data)
#adding the diagnosis column to the data frame
pca_data$diagnosis <- y

#now we fit the Logistic Regression model to the data frame
#split the data into training and testing set
split2<- sample.split(pca_data, SplitRatio = 0.8)
split2
train_data2<- subset(pca_data, split2)
test_data2<- subset(pca_data, !split2)

#Fit logistic regression model
logistic_model2 <- glm(diagnosis~., data = train_data2, family =
binomial)

#predict probabilities on test set
test_data2$predictions2 <- predict(logistic_model2, newdata =
test_data2, type = "response")

#Convert probabilities to class labels
predicted_classes2 <- ifelse(test_data2$predictions2> 0.5, "M", "B")
```

```
#Confusion matrix
confusion_matrix2 <- table(predicted_classes2, test_data2$diagnosis)
print(confusion_matrix2)

#calculate accuracy
accuracy2 <- sum(diag(confusion_matrix2))/ sum(confusion_matrix2)
print(paste("Accuracy:", accuracy2))

[1] "Accuracy: 0.961290322580645"
```

Thus we get ≈ **96.1%** accuracy.

Remark

We can see that using 10 Principal Components derived from the dataset, the accuracy of the model has changed. The accuracy has increased by almost 1% in comparison with the first model.

But the most important part is that this model is more efficient and reliable as it is faster than the previous model and also we are working with only 10 principal components rather than working with 30 predictor variables.

4. Conclusion

This study aimed to develop an efficient and accurate model for predicting breast cancer diagnosis using **logistic regression** and **principal component analysis (PCA)**. Our analysis of the Wisconsin Breast Cancer Dataset yielded several significant findings:

1. Initial logistic regression model: Using all 30 predictor variables, we achieved an accuracy of approximately 95%. While this result was promising, it highlighted potential issues of multicollinearity and model complexity.

2. Dimension reduction through PCA: Our application of PCA revealed that the first 10 principal components accounted for 95.16% of the total variance in the dataset. This finding allowed us to significantly reduce the dimensionality of our data while retaining most of its informational content.

3. Improved model efficiency: By applying logistic regression to the 10 principal components, we not only maintained but slightly improved our model's accuracy to 96.1%. This demonstrates that our dimension reduction approach effectively addressed multicollinearity concerns while enhancing model performance.

4. Feature importance: The PCA results indicated that the first two principal components alone explained 63.3% of the total variability, suggesting that a relatively small number of composite features capture much of the relevant information for diagnosis. These results underscore the effectiveness of combining logistic regression with PCA for breast cancer prediction. Our approach not only maintains high accuracy but also improves model interpretability and computational efficiency by reducing the number of input variables from 30 to 10.

5. Significance: The enhanced model efficiency and maintained high accuracy suggest that this approach could be valuable in clinical settings, potentially aiding in faster and more reliable breast cancer diagnoses. However, it's important to note that while our model shows promise, it should be considered as a supportive tool rather than a replacement for expert medical judgement. Future work could focus on external validation of this model with diverse datasets, exploration of other machine learning techniques, and investigation of the biological significance of the principal components identified in this study. Additionally, collaborating with medical professionals to translate these findings into practical diagnostic tools could be a valuable next

step.

In conclusion, this project demonstrates the potential of combining statistical techniques like logistic regression and PCA to create efficient and accurate predictive models for breast cancer diagnosis, contributing to the ongoing efforts to improve early detection and treatment of this critical health issue.

5. References

1. Breast Cancer Facts and Statistics

2. Jemal A, Siegel R, Ward E, Hao Y, Xu J, Thun MJ. Cancer Statistics, 2009. CA Cancer J Clin. 2009; 59:225-49

3. Statistical Analysis of Breast Cancer cases in Indian Women by Omega Hospitals

4. Saxena S, Szabo CI, Chopin S, Brajhoux L, Sinilnikova O, Lenoir G, Goldgar DE, Bhatanager D, BRCA1 and BRCA2 in breast cancer patients. Hum Mutat. 2002;20:473-474

5. The Epidemiology of Breast Cancer- National Center for Biotechnology Information

APPENDIX

1. The Breast cancer Wisconsin dataset:

id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean
842302	M	17.99	10.38	122.8	1001	0.1184	0.2776	0.3001
842517	M	20.57	17.77	132.9	1326	0.08474	0.07864	0.0869
84300903	M	19.69	21.25	130	1203	0.1096	0.1599	0.1974
84348301	M	11.42	20.38	77.58	386.1	0.1425	0.2839	0.2414
84358402	M	20.29	14.34	135.1	1297	0.1003	0.1328	0.198
843786	M	12.45	15.7	82.57	477.1	0.1278	0.17	0.1578
844359	M	18.25	19.98	119.6	1040	0.09463	0.109	0.1127
84458202	M	13.71	20.83	90.2	577.9	0.1189	0.1645	0.09366
844981	M	13	21.82	87.5	519.8	0.1273	0.1932	0.1859

…continued

[Link to display the whole dataset is attached here]

2. The dataset after applying the PCA algorithm:

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
1	-9.18475521	-1.94687003	-1.12217866	3.630536408	1.194059478	1.410183639	2.15747152	0.398056982	-0.156980233
2	-2.385702629	3.764859063	-0.528827374	1.117280773	-0.621228365	0.028631162	0.01334635	-0.2407766	-0.711278966
3	-5.728855491	1.074228589	-0.55126254	0.91128084	0.176930218	0.540976145	-0.667579085	-0.097288135	0.024044486
4	-7.11669126	-10.26655564	-3.229947535	0.152412923	2.958275431	3.050737497	1.428653635	-1.058633755	-1.404204115
5	-3.931842467	1.946358977	1.388544953	2.938054169	-0.546266745	-1.225416405	-0.935389502	-0.635816609	-0.263573547
6	-2.378154625	-3.94645643	-2.932296681	0.940209586	1.055113543	-0.450642134	0.490013955	0.165298432	-0.13335576
7	-2.236915059	2.687666414	-1.638471247	0.1492086	-0.040324037	-0.128835073	-0.301302331	-0.083624609	-0.07995422
8	-2.141414282	-2.338186649	-0.871180715	-0.126931171	1.426181783	-1.255934103	0.973243497	0.652763899	0.247965464
9	-3.172133315	-3.388831138	-3.117243065	-0.60076844	1.520952114	0.55905282	-0.214914597	0.686736938	0.511473855

…continued

[Link to display the PCA dataset is attached here]