Cancer Prediction

authour

2023-12-10

Introduction

Worldwide, breast cancer is the most frequent cancer to affect women. It affects about 2.1 million people in 2015 alone and makes up 25% of all cancer cases. It all begins when breast cells start to proliferate uncontrollably. Usually, these cells develop into tumors that are felt as lumps in the breast area or that are visible on X-rays. The main obstacle to its discovery is determining whether a tumor is benign (not cancerous) or malignant (cancerous). Please finish the analysis of the Breast Cancer Wisconsin (Diagnostic) Dataset and machine learning (using SVMs) to classify these tumors.

#Data description Link to the dataset

https://www.kaggle.com/datasets/yasserh/breast-cancer-dataset

• This dataset contains information on features that help Build classification models to predict whether the cancer type is Malignant or Benign. Machine learning algorithms can be used to create prediction models with this data. Utilize this dataset for visualization, exploration, and data cleaning.

#Objective: • Understand the Dataset & cleanup (if required). • Build classification models to predict whether the cancer type is Malignant or Benign. and find the most important features

The research question

What features contribute the most when building classification models to predict whether the cancer type is Malignant or Benign?
The factors or parameters from the dataset that can be utilized are 'radius_mean', 'texture_mean', 'perimeter_mean' and 'area_mean',

Computational Methods

Data-driven, computational approach may be useful Because a data-driven, computational method makes it possible to analyze a lot of data and find patterns and interactions between variables, it might be helpful in addressing the research topic. When developing classification models to determine whether a cancer type is benign or malignant, the research question asks about the traits or qualities of the disease that are most significant. Stated differently, the goal of the research is to identify the critical variables that are important in differentiating between benign and malignant tumors. A data-driven, computational approach may be useful in addressing the research topic since it enables the analysis of large amounts of data and the discovery of patterns and relationships between variables. In this case, the technique can help determine the most important features.

A computational and data-driven method is proposed to address this question. This implies that in order to extract useful insights from the data, the research would need to analyze already-existing data on cancer cases, possibly with the aid of statistical models and algorithms. This method would entail gathering pertinent information, doing statistical analyses, and developing classification models in order to pinpoint the salient characteristics that are most important in determining the kind of cancer. In general, the research topic implies that it would be beneficial to use a data-driven, computational method to identify the factors that have the greatest influence when developing classification models that predict whether a cancer is benign or malignant.

```r
library(ggplot2)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v lubridate 1.9.2     v tibble    3.2.1
## v purrr     1.0.2     v tidyr     1.3.0
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
df1 = read.csv("C:\\Users\\nakka\\OneDrive\\Documents\\breast-cancer.csv")

head(df1, 3)
```

```
##         id diagnosis radius_mean texture_mean perimeter_mean area_mean
## 1   842302         M       17.99        10.38          122.8      1001
## 2   842517         M       20.57        17.77          132.9      1326
## 3 84300903         M       19.69        21.25          130.0      1203
##   smoothness_mean compactness_mean concavity_mean concave.points_mean
## 1         0.11840          0.27760         0.3001             0.14710
## 2         0.08474          0.07864         0.0869             0.07017
## 3         0.10960          0.15990         0.1974             0.12790
##   symmetry_mean fractal_dimension_mean radius_se texture_se perimeter_se
## 1        0.2419                0.07871    1.0950     0.9053        8.589
## 2        0.1812                0.05667    0.5435     0.7339        3.398
## 3        0.2069                0.05999    0.7456     0.7869        4.585
##   area_se smoothness_se compactness_se concavity_se concave.points_se
## 1  153.40      0.006399        0.04904      0.05373           0.01587
## 2   74.08      0.005225        0.01308      0.01860           0.01340
## 3   94.03      0.006150        0.04006      0.03832           0.02058
##   symmetry_se fractal_dimension_se radius_worst texture_worst perimeter_worst
## 1     0.03003             0.006193        25.38         17.33           184.6
## 2     0.01389             0.003532        24.99         23.41           158.8
## 3     0.02250             0.004571        23.57         25.53           152.5
##   area_worst smoothness_worst compactness_worst concavity_worst
## 1       2019           0.1622            0.6656          0.7119
## 2       1956           0.1238            0.1866          0.2416
## 3       1709           0.1444            0.4245          0.4504
##   concave.points_worst symmetry_worst fractal_dimension_worst
## 1               0.2654         0.4601                 0.11890
## 2               0.1860         0.2750                 0.08902
## 3               0.2430         0.3613                 0.08758
```

Load required libraries

```r
dim(df1)
```

```
## [1] 569  32
```

# • For the choosen dataset, what are the necessary data wrangling steps to make the data ready 1. Remove missing values: Use the “na.omit” function to remove rows with missing values in the dataframe df1. This step is performed using the command “df1 <- na.omit(df1)”.

2. Check for missing values: Use the “sum(is.na())” function to count the number of missing values in the dataframe df1. This step is performed using the command “sum(is.na(df1))”.
3. Convert data type: Convert the “diagnosis” column from string to numeric. In this case, the “M” value is converted to 1 and the “B” value is converted to 0. This step can be performed using the “ifelse” function and the command “df1$diagnosis < -ifelse(df$1diagnosis ==”M”, 1, 0)“.

By performing these steps, the data is prepared for subsequent analyses by removing missing values and converting the necessary columns to the appropriate data types.

```r
df1 = na.omit(df1)

#checking for missing values
sum(is.na(df1))#
```

```
## [1] 0
```

```r
summary(df1)
```

```
##        id              diagnosis          radius_mean      texture_mean
##  Min.   :     8670   Length:569         Min.   : 6.981   Min.   : 9.71
##  1st Qu.:   869218   Class :character   1st Qu.:11.700   1st Qu.:16.17
##  Median :   906024   Mode  :character   Median :13.370   Median :18.84
##  Mean   : 30371831                      Mean   :14.127   Mean   :19.29
##  3rd Qu.:  8813129                      3rd Qu.:15.780   3rd Qu.:21.80
##  Max.   :911320502                      Max.   :28.110   Max.   :39.28
##  perimeter_mean     area_mean      smoothness_mean   compactness_mean
##  Min.   : 43.79   Min.   : 143.5   Min.   :0.05263   Min.   :0.01938
##  1st Qu.: 75.17   1st Qu.: 420.3   1st Qu.:0.08637   1st Qu.:0.06492
##  Median : 86.24   Median : 551.1   Median :0.09587   Median :0.09263
##  Mean   : 91.97   Mean   : 654.9   Mean   :0.09636   Mean   :0.10434
##  3rd Qu.:104.10   3rd Qu.: 782.7   3rd Qu.:0.10530   3rd Qu.:0.13040
##  Max.   :188.50   Max.   :2501.0   Max.   :0.16340   Max.   :0.34540
##  concavity_mean    concave.points_mean symmetry_mean    fractal_dimension_mean
##  Min.   :0.00000   Min.   :0.00000     Min.   :0.1060   Min.   :0.04996
##  1st Qu.:0.02956   1st Qu.:0.02031     1st Qu.:0.1619   1st Qu.:0.05770
##  Median :0.06154   Median :0.03350     Median :0.1792   Median :0.06154
##  Mean   :0.08880   Mean   :0.04892     Mean   :0.1812   Mean   :0.06280
##  3rd Qu.:0.13070   3rd Qu.:0.07400     3rd Qu.:0.1957   3rd Qu.:0.06612
##  Max.   :0.42680   Max.   :0.20120     Max.   :0.3040   Max.   :0.09744
##    radius_se        texture_se      perimeter_se       area_se
##  Min.   :0.1115   Min.   :0.3602   Min.   : 0.757   Min.   :  6.802
##  1st Qu.:0.2324   1st Qu.:0.8339   1st Qu.: 1.606   1st Qu.: 17.850
##  Median :0.3242   Median :1.1080   Median : 2.287   Median : 24.530
##  Mean   :0.4052   Mean   :1.2169   Mean   : 2.866   Mean   : 40.337
##  3rd Qu.:0.4789   3rd Qu.:1.4740   3rd Qu.: 3.357   3rd Qu.: 45.190
##  Max.   :2.8730   Max.   :4.8850   Max.   :21.980   Max.   :542.200
##  smoothness_se      compactness_se      concavity_se     concave.points_se
##  Min.   :0.001713   Min.   :0.002252   Min.   :0.00000   Min.   :0.000000
##  1st Qu.:0.005169   1st Qu.:0.013080   1st Qu.:0.01509   1st Qu.:0.007638
##  Median :0.006380   Median :0.020450   Median :0.02589   Median :0.010930
##  Mean   :0.007041   Mean   :0.025478   Mean   :0.03189   Mean   :0.011796
##  3rd Qu.:0.008146   3rd Qu.:0.032450   3rd Qu.:0.04205   3rd Qu.:0.014710
##  Max.   :0.031130   Max.   :0.135400   Max.   :0.39600   Max.   :0.052790
##   symmetry_se       fractal_dimension_se  radius_worst   texture_worst
##  Min.   :0.007882   Min.   :0.0008948    Min.   : 7.93   Min.   :12.02
##  1st Qu.:0.015160   1st Qu.:0.0022480    1st Qu.:13.01   1st Qu.:21.08
##  Median :0.018730   Median :0.0031870    Median :14.97   Median :25.41
##  Mean   :0.020542   Mean   :0.0037949    Mean   :16.27   Mean   :25.68
##  3rd Qu.:0.023480   3rd Qu.:0.0045580    3rd Qu.:18.79   3rd Qu.:29.72
##  Max.   :0.078950   Max.   :0.0298400    Max.   :36.04   Max.   :49.54
##  perimeter_worst    area_worst     smoothness_worst  compactness_worst
##  Min.   : 50.41   Min.   : 185.2   Min.   :0.07117   Min.   :0.02729
##  1st Qu.: 84.11   1st Qu.: 515.3   1st Qu.:0.11660   1st Qu.:0.14720
##  Median : 97.66   Median : 686.5   Median :0.13130   Median :0.21190
##  Mean   :107.26   Mean   : 880.6   Mean   :0.13237   Mean   :0.25427
##  3rd Qu.:125.40   3rd Qu.:1084.0   3rd Qu.:0.14600   3rd Qu.:0.33910
##  Max.   :251.20   Max.   :4254.0   Max.   :0.22260   Max.   :1.05800
##  concavity_worst  concave.points_worst symmetry_worst   fractal_dimension_worst
##  Min.   :0.0000   Min.   :0.00000      Min.   :0.1565   Min.   :0.05504
##  1st Qu.:0.1145   1st Qu.:0.06493      1st Qu.:0.2504   1st Qu.:0.07146
##  Median :0.2267   Median :0.09993      Median :0.2822   Median :0.08004
##  Mean   :0.2722   Mean   :0.11461      Mean   :0.2901   Mean   :0.08395
##  3rd Qu.:0.3829   3rd Qu.:0.16140      3rd Qu.:0.3179   3rd Qu.:0.09208
##  Max.   :1.2520   Max.   :0.29100      Max.   :0.6638   Max.   :0.20750
```

```
colnames(df1)
```

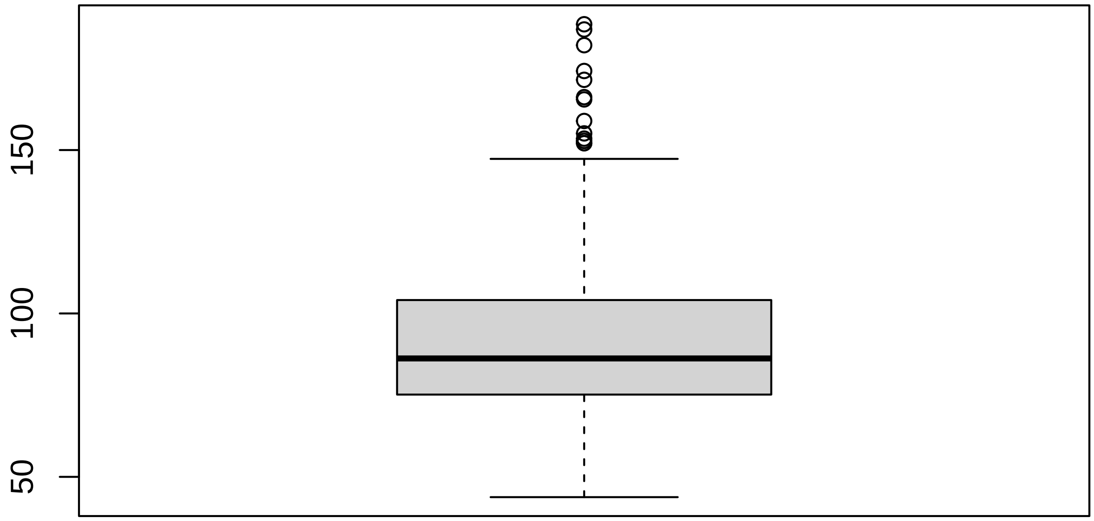
```
##  [1] "id"                      "diagnosis"
##  [3] "radius_mean"             "texture_mean"
##  [5] "perimeter_mean"          "area_mean"
##  [7] "smoothness_mean"         "compactness_mean"
##  [9] "concavity_mean"          "concave.points_mean"
## [11] "symmetry_mean"           "fractal_dimension_mean"
## [13] "radius_se"               "texture_se"
## [15] "perimeter_se"            "area_se"
## [17] "smoothness_se"           "compactness_se"
## [19] "concavity_se"            "concave.points_se"
## [21] "symmetry_se"             "fractal_dimension_se"
## [23] "radius_worst"            "texture_worst"
## [25] "perimeter_worst"         "area_worst"
## [27] "smoothness_worst"        "compactness_worst"
## [29] "concavity_worst"         "concave.points_worst"
## [31] "symmetry_worst"          "fractal_dimension_worst"
```

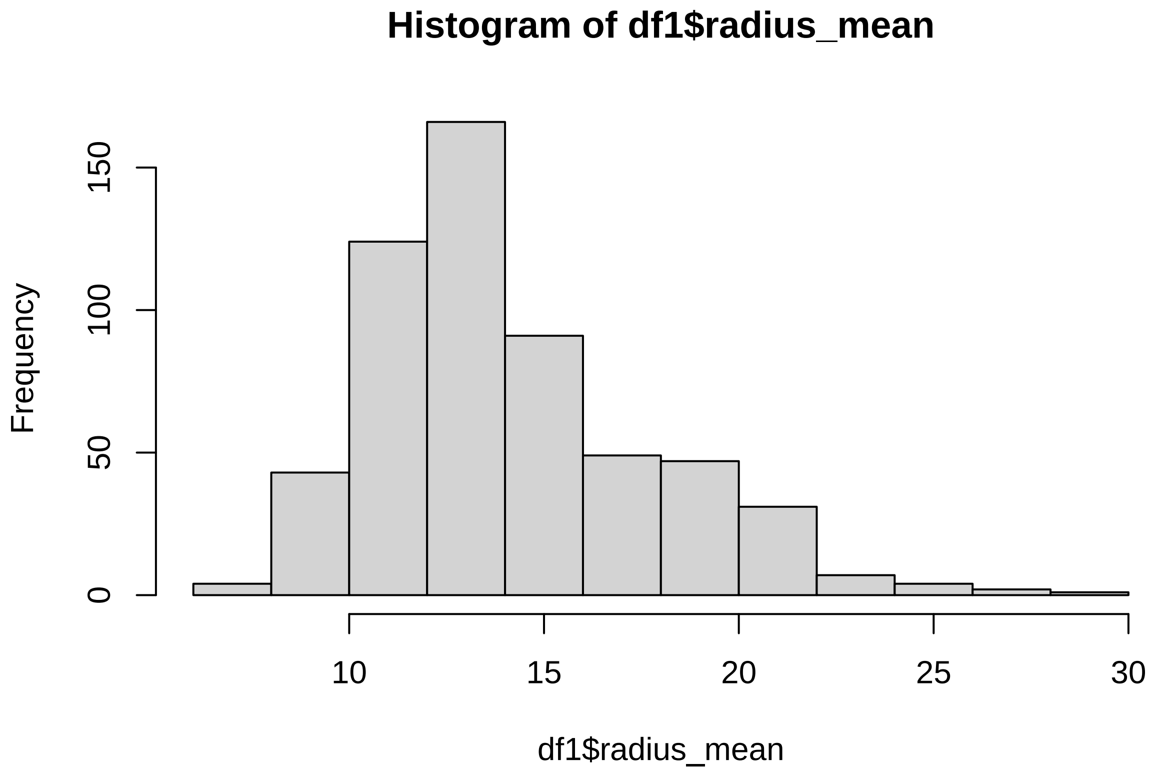
```r
#convert from string to numeric
df1$diagnosis = ifelse(df1$diagnosis == "M", 1, 0)
summary(df1)
```

```
##        id              diagnosis       radius_mean      texture_mean
##  Min.   :     8670   Min.   :0.0000   Min.   : 6.981   Min.   : 9.71
##  1st Qu.:   869218   1st Qu.:0.0000   1st Qu.:11.700   1st Qu.:16.17
##  Median :   906024   Median :0.0000   Median :13.370   Median :18.84
##  Mean   : 30371831   Mean   :0.3726   Mean   :14.127   Mean   :19.29
##  3rd Qu.:  8813129   3rd Qu.:1.0000   3rd Qu.:15.780   3rd Qu.:21.80
##  Max.   :911320502   Max.   :1.0000   Max.   :28.110   Max.   :39.28
##  perimeter_mean     area_mean      smoothness_mean   compactness_mean
##  Min.   : 43.79   Min.   : 143.5   Min.   :0.05263   Min.   :0.01938
##  1st Qu.: 75.17   1st Qu.: 420.3   1st Qu.:0.08637   1st Qu.:0.06492
##  Median : 86.24   Median : 551.1   Median :0.09587   Median :0.09263
##  Mean   : 91.97   Mean   : 654.9   Mean   :0.09636   Mean   :0.10434
##  3rd Qu.:104.10   3rd Qu.: 782.7   3rd Qu.:0.10530   3rd Qu.:0.13040
##  Max.   :188.50   Max.   :2501.0   Max.   :0.16340   Max.   :0.34540
##  concavity_mean    concave.points_mean symmetry_mean    fractal_dimension_mean
##  Min.   :0.00000   Min.   :0.00000     Min.   :0.1060   Min.   :0.04996
##  1st Qu.:0.02956   1st Qu.:0.02031     1st Qu.:0.1619   1st Qu.:0.05770
##  Median :0.06154   Median :0.03350     Median :0.1792   Median :0.06154
##  Mean   :0.08880   Mean   :0.04892     Mean   :0.1812   Mean   :0.06280
##  3rd Qu.:0.13070   3rd Qu.:0.07400     3rd Qu.:0.1957   3rd Qu.:0.06612
##  Max.   :0.42680   Max.   :0.20120     Max.   :0.3040   Max.   :0.09744
##    radius_se         texture_se        perimeter_se        area_se
##  Min.   :0.1115   Min.   :0.3602   Min.   : 0.757   Min.   :  6.802
##  1st Qu.:0.2324   1st Qu.:0.8339   1st Qu.: 1.606   1st Qu.: 17.850
##  Median :0.3242   Median :1.1080   Median : 2.287   Median : 24.530
##  Mean   :0.4052   Mean   :1.2169   Mean   : 2.866   Mean   : 40.337
##  3rd Qu.:0.4789   3rd Qu.:1.4740   3rd Qu.: 3.357   3rd Qu.: 45.190
##  Max.   :2.8730   Max.   :4.8850   Max.   :21.980   Max.   :542.200
##  smoothness_se      compactness_se      concavity_se      concave.points_se
##  Min.   :0.001713   Min.   :0.002252   Min.   :0.00000   Min.   :0.000000
##  1st Qu.:0.005169   1st Qu.:0.013080   1st Qu.:0.01509   1st Qu.:0.007638
##  Median :0.006380   Median :0.020450   Median :0.02589   Median :0.010930
##  Mean   :0.007041   Mean   :0.025478   Mean   :0.03189   Mean   :0.011796
##  3rd Qu.:0.008146   3rd Qu.:0.032450   3rd Qu.:0.04205   3rd Qu.:0.014710
##  Max.   :0.031130   Max.   :0.135400   Max.   :0.39600   Max.   :0.052790
##   symmetry_se       fractal_dimension_se  radius_worst   texture_worst
##  Min.   :0.007882   Min.   :0.0008948    Min.   : 7.93   Min.   :12.02
##  1st Qu.:0.015160   1st Qu.:0.0022480    1st Qu.:13.01   1st Qu.:21.08
##  Median :0.018730   Median :0.0031870    Median :14.97   Median :25.41
##  Mean   :0.020542   Mean   :0.0037949    Mean   :16.27   Mean   :25.68
##  3rd Qu.:0.023480   3rd Qu.:0.0045580    3rd Qu.:18.79   3rd Qu.:29.72
##  Max.   :0.078950   Max.   :0.0298400    Max.   :36.04   Max.   :49.54
##  perimeter_worst    area_worst      smoothness_worst  compactness_worst
##  Min.   : 50.41   Min.   : 185.2   Min.   :0.07117   Min.   :0.02729
##  1st Qu.: 84.11   1st Qu.: 515.3   1st Qu.:0.11660   1st Qu.:0.14720
##  Median : 97.66   Median : 686.5   Median :0.13130   Median :0.21190
##  Mean   :107.26   Mean   : 880.6   Mean   :0.13237   Mean   :0.25427
##  3rd Qu.:125.40   3rd Qu.:1084.0   3rd Qu.:0.14600   3rd Qu.:0.33910
##  Max.   :251.20   Max.   :4254.0   Max.   :0.22260   Max.   :1.05800
##  concavity_worst  concave.points_worst symmetry_worst   fractal_dimension_worst
##  Min.   :0.0000   Min.   :0.00000      Min.   :0.1565   Min.   :0.05504
##  1st Qu.:0.1145   1st Qu.:0.06493      1st Qu.:0.2504   1st Qu.:0.07146
##  Median :0.2267   Median :0.09993      Median :0.2822   Median :0.08004
##  Mean   :0.2722   Mean   :0.11461      Mean   :0.2901   Mean   :0.08395
##  3rd Qu.:0.3829   3rd Qu.:0.16140      3rd Qu.:0.3179   3rd Qu.:0.09208
##  Max.   :1.2520   Max.   :0.29100      Max.   :0.6638   Max.   :0.20750
```

Exploratory analyses - EDA

perimeter_mean had some outliers as shown by the boxplot





Data Analysis and Results

```
#CORRELATION analysis
temp = df1 |>
  dplyr::select('radius_mean', 'texture_mean', 'perimeter_mean', 'area_mean', 'diagnosis' )
head(temp)
```

```
##   radius_mean texture_mean perimeter_mean area_mean diagnosis
## 1       17.99        10.38         122.80    1001.0         1
## 2       20.57        17.77         132.90    1326.0         1
## 3       19.69        21.25         130.00    1203.0         1
## 4       11.42        20.38          77.58     386.1         1
## 5       20.29        14.34         135.10    1297.0         1
## 6       12.45        15.70          82.57     477.1         1
```

```
#install.packages("lattice")
library(lattice)

# rounding to 2 decimal places
corr_m = round(cor(temp),2)
head(corr_m)
```

```
##                  radius_mean texture_mean perimeter_mean area_mean diagnosis
## radius_mean           1.00          0.32            1.00       0.99       0.73
## texture_mean          0.32          1.00            0.33       0.32       0.42
## perimeter_mean        1.00          0.33            1.00       0.99       0.74
## area_mean             0.99          0.32            0.99       1.00       0.71
## diagnosis             0.73          0.42            0.74       0.71       1.00
```

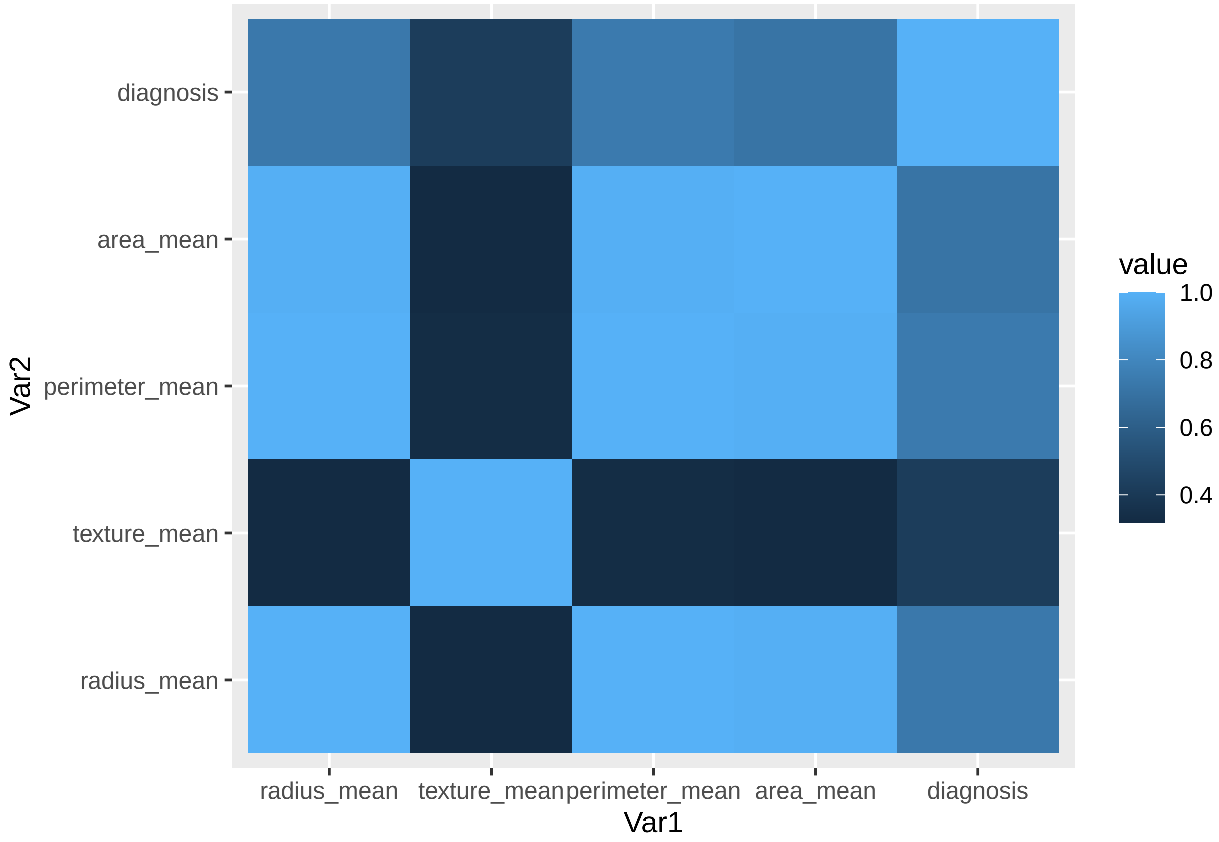
These correlations show the relationship between the variable "diagnosis" (indicating whether a breast tumor is malignant or benign) and different features of the tumors: radius_mean, texture_mean, perimeter_mean, and area_mean.

- The correlation between "diagnosis" and "radius_mean" is positive with a value of 0.73. This indicates that as the average radius of the tumor increases, the likelihood of the tumor being diagnosed as malignant also increases.
- The correlation between "diagnosis" and "texture_mean" is positive but weaker, with a value of 0.42. This suggests that there is a moderate association between the texture of the tumor and the diagnosis, but it is not as strong as the relationship with radius_mean.
- The correlation between "diagnosis" and "perimeter_mean" is strong, with a value of 0.74. This means that as the average perimeter of the tumor increases, the chance of it being diagnosed as malignant also increases.
- The correlation between "diagnosis" and "area_mean" is positive and has a value of 0.71. This indicates that there is a strong positive association between the average area of the tumor and the diagnosis. As the area increases, the likelihood of the tumor being malignant also increases.

```
## Warning: package 'reshape2' was built under R version 4.3.2
```

```
##
## Attaching package: 'reshape2'
```

```
## The following object is masked from 'package:tidyr':
##
##     smiths
```



Modeling Techniques

logistic regression was used as a modeling technique to predict cancer. The code snippet provided demonstrates how logistic regression was implemented using the glm() function in R. The dependent variable, "diagnosis," represents the presence or absence of cancer, and the independent variables, "radius_mean," "texture_mean," "perimeter_mean," and "area_mean," are the predictors used in the model.

The glm() function is applied to the dataset "df1," and the family argument is set to "binomial" to indicate that we are performing binary logistic regression. This means that the outcome variable, diagnosis, is binary (presence or absence of cancer) and follows a binomial distribution.

By running this code, the logistic regression model is estimated, which allows us to predict the probability of cancer based on the values of the predictor variables. The model takes into account the relationship between the predictors and the outcome variable and provides coefficients that quantify the effect of each predictor on the likelihood of having cancer.

```
#df_model1 = subset(df_clean, select = c(Purchased_numeric,Income))
model = glm(  diagnosis ~  radius_mean + texture_mean + perimeter_mean + area_mean, data = df1, family
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
summary(model)$coef
```

```
##                    Estimate Std. Error   z value     Pr(>|z|)
## (Intercept)      1.77290702 6.87010704  0.258061 7.963598e-01
## radius_mean    -9.42873795 1.63958422 -5.750688 8.888080e-09
## texture_mean    0.23760964 0.04602853  5.162226 2.440306e-07
## perimeter_mean  1.15065585 0.16435846  7.000892 2.543377e-12
## area_mean       0.03277012 0.01182456  2.771361 5.582245e-03
```

```
coef(model)
```

```
##    (Intercept)    radius_mean   texture_mean perimeter_mean      area_mean
##     1.77290702    -9.42873795     0.23760964     1.15065585     0.03277012
```

FEATURE SELECTION

```
library(caret)
```

```
## Warning: package 'caret' was built under R version 4.3.2
```

```
##
## Attaching package: 'caret'
```

```
## The following object is masked from 'package:purrr':
##
##     lift
```

```
important_features = varImp(model)
important_features
```

```
##                 Overall
## radius_mean    5.750688
## texture_mean   5.162226
## perimeter_mean 7.000892
## area_mean      2.771361
```

```
#DISPLAYS THE TOP FEATURES
```

The research question aims to determine which features are most important in building classification models to predict whether a cancer type is Malignant or Benign. The factors or parameters considered for analysis are 'radius_mean', 'texture_mean', 'perimeter_mean', and 'area_mean'. The feature selection analysis revealed that 'perimeter_mean' is the most significant feature, followed by 'radius_mean' and 'texture_mean'. 'Area_mean' was found to contribute the least in predicting the cancer type.

```
model2 = glm(  diagnosis ~  radius_mean + perimeter_mean , data = df1, family = binomial)
summary(model2)$coef
```

```
##                  Estimate Std. Error   z value     Pr(>|z|)
## (Intercept)    -13.301275  1.4138803 -9.407639 5.074068e-21
## radius_mean     -5.741509  0.8994975 -6.383019 1.736305e-10
## perimeter_mean   1.020808  0.1397836  7.302777 2.818883e-13
```

```
coef(model2)
```

```
##    (Intercept)    radius_mean perimeter_mean
##     -13.301275      -5.741509       1.020808
```

Interpretation

The estimate for the Intercept is -13.301275, indicating that the log chances of a positive diagnosis are -13.301275 when both “radius_mean” and “perimeter_mean” are 0.

The estimate for “radius_mean” is -5.741509, meaning that the log probabilities of a positive diagnosis drop by -5.741509 for every unit increase in “radius_mean” while keeping all other variables constant.

The estimate for “perimeter_mean” is 1.020808, meaning that the log probabilities of a positive diagnosis rise by 1.020808 for every unit increase in “perimeter_mean” while keeping all other variables constant.

With a statistically significant p-value for each coefficient, it is possible that they all significantly deviate from zero and have an effect on the outcome variable.

Conclusion

The goal of the study is to use a logistic regression model to ascertain the significance of various variables in predicting the kind of cancer (malignant or benign). Four features—“radius_mean,” “texture_mean,” “perimeter_mean,” and “area_mean”—are included in the analysis, and their importance in determining the kind of cancer is assessed.

The research can be deemed more generalizable if the dataset is typical of the entire population and includes a wide variety of cancer cases.

It is necessary to take into account any potential limitations with this analysis. First off, the study only takes into account four features; other significant features may exist that are left out of the model. There’s a chance that leaving out some features could compromise the model’s precision and applicability.

It is necessary to take into account any potential limitations with this analysis. First off, the study only takes into account four features; other significant features may exist that are left out of the model. There’s a chance that leaving out some features could compromise the model’s precision and applicability.

Furthermore, the cautionary note “fitted probabilities numerically 0 or 1 occurred” raises the possibility of a separation problem in the data, which could result in inaccurate parameter estimations. Either greater data collection or the application of regularization strategies like ridge or lasso regression can be used to solve this problem.

Moreover, it’s possible that the analysis’s findings cannot be applied to other cancer kinds or demographics. The analysis is particular to the dataset that was used.

It would be advantageous to take into account a larger dataset with a more varied range of cases in order to enhance the analysis. This could enhance the generalizability of the findings and assist capture the diversity in various cancer types. To find the most pertinent features for predicting cancer type, it would also be beneficial to investigate other feature selection methods like correlation analysis or recursive feature removal.

In conclusion, there are restrictions on the analysis’s scope and generalizability even if it sheds light on the significance of particular characteristics in predicting the kind of cancer. Extensive and varied datasets and additional feature selection methods can be used in future study to enhance the precision and relevance of the results.