Signature Project

#Required packages

```
#Imported required packages to perform the project
library(corrplot)
library(magrittr)
library(dplyr)
library(ggcorrplot)
library(psych)
library(RVAideMemoire)
library(moments)
library(tidyverse)
library(CatEncoders)
library(DMwR)
library(kernlab)
library(C50)
library(gmodels)
library(caret)
library(Metrics)
library(irr)
library(plotly)
library(cvms)
library(ipred)
library(caretEnsemble)
```

#Data Acquisition

```
#Data imported from the local folder and read it using read.csv function and
#set parameter "stringAsFactors" to "TRUE" to convert the character features into
#factor levels

#Data set links
#https://ieee-dataport.s3.amazonaws.com/open/7249/SEER%20Breast%20Cancer%20Dataset%20.csv?response-cont
#https://www.kaggle.com/datasets/reihanenamdari/breast-cancer?select=Breast_Cancer.csv

ID <- "17XYkkiYNGdp5sY15qNIIsFk04YO_n9IK"

seer_data <- read.csv(sprintf("https://docs.google.com/uc?id=%s&export=download", ID))
head(seer_data)
```

```
##   Age                                                      Race
## 1  43 Other (American Indian/AK Native, Asian/Pacific Islander)
## 2  47 Other (American Indian/AK Native, Asian/Pacific Islander)
## 3  67                                                     White
## 4  46                                                     White
## 5  63                                                     White
## 6  49                                                     White
##                    Marital.Status  X T.Stage N.Stage X6th.Stage
## 1 Married (including common law) NA      T2      N3       IIIC
## 2 Married (including common law) NA      T2      N2       IIIA
## 3 Married (including common law) NA      T2      N1        IIB
## 4                       Divorced NA      T1      N1        IIA
## 5 Married (including common law) NA      T2      N2       IIIA
## 6 Married (including common law) NA      T2      N3       IIIC
##                                 Grade  A.Stage Tumor.Size Estrogen.Status
## 1 Moderately differentiated; Grade II Regional         40        Positive
## 2 Moderately differentiated; Grade II Regional         45        Positive
## 3    Poorly differentiated; Grade III Regional         25        Positive
## 4 Moderately differentiated; Grade II Regional         19        Positive
## 5 Moderately differentiated; Grade II Regional         35        Positive
## 6 Moderately differentiated; Grade II Regional         32        Positive
##   Progesterone.Status Regional.Node.Examined Reginol.Node.Positive
## 1            Positive                     19                    11
## 2            Positive                     25                     9
## 3            Positive                      4                     1
## 4            Positive                     26                     1
## 5            Positive                     21                     5
## 6            Positive                     20                    11
##   Survival.Months Status
## 1               1  Alive
## 2               2  Alive
## 3               2   Dead
## 4               2   Dead
## 5               3   Dead
## 6               3  Alive
```

```
#Removing extra column containing NA values(duplicate column)
SEER_data <- seer_data[,-4]


#Seer cancer data containing 4024 rows and 14 independent columns and 1
#dependent column(target variable)
dim(SEER_data)
```

```
## [1] 4024   15
```

```
#string output showing the factor levels and integer columns
str(SEER_data)
```

```
## ’data.frame':    4024 obs. of  15 variables:
##  $ Age                    : int  43 47 67 46 63 49 64 55 59 67 ...
##  $ Race                   : chr  "Other (American Indian/AK Native, Asian/Pacific Islander)" "Other (A
##  $ Marital.Status         : chr  "Married (including common law)" "Married (including common law)" "Ma
##  $ T.Stage                : chr  "T2" "T2" "T2" "T1" ...
##  $ N.Stage                : chr  "N3" "N2" "N1" "N1" ...
##  $ X6th.Stage             : chr  "IIIC" "IIIA" "IIB" "IIA" ...
##  $ Grade                  : chr  "Moderately differentiated; Grade II" "Moderately differentiated; Gra
##  $ A.Stage                : chr  "Regional" "Regional" "Regional" "Regional" ...
##  $ Tumor.Size             : int  40 45 25 19 35 32 22 15 70 55 ...
##  $ Estrogen.Status       : chr  "Positive" "Positive" "Positive" "Positive" ...
##  $ Progesterone.Status   : chr  "Positive" "Positive" "Positive" "Positive" ...
##  $ Regional.Node.Examined: int  19 25 4 26 21 20 1 9 9 9 ...
##  $ Reginol.Node.Positive : int  11 9 1 1 5 11 1 1 1 9 ...
##  $ Survival.Months       : int  1 2 2 2 3 3 3 3 4 4 ...
##  $ Status                : chr  "Alive" "Alive" "Dead" "Dead" ...
```

```
summary(SEER_data)
```

```
##       Age            Race           Marital.Status       T.Stage
##  Min.   :30.00   Length:4024        Length:4024        Length:4024
##  1st Qu.:47.00   Class :character   Class :character   Class :character
##  Median :54.00   Mode  :character   Mode  :character   Mode  :character
##  Mean   :53.97
##  3rd Qu.:61.00
##  Max.   :69.00
##    N.Stage            X6th.Stage           Grade             A.Stage
##  Length:4024        Length:4024        Length:4024        Length:4024
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##    Tumor.Size     Estrogen.Status    Progesterone.Status Regional.Node.Examined
##  Min.   :  1.00   Length:4024        Length:4024         Min.   : 1.00
##  1st Qu.: 16.00   Class :character   Class :character    1st Qu.: 9.00
##  Median : 25.00   Mode  :character   Mode  :character    Median :14.00
##  Mean   : 30.47                                          Mean   :14.36
##  3rd Qu.: 38.00                                          3rd Qu.:19.00
##  Max.   :140.00                                          Max.   :61.00
##  Reginol.Node.Positive Survival.Months    Status
##  Min.   : 1.000        Min.   :  1.0   Length:4024
##  1st Qu.: 1.000        1st Qu.: 56.0   Class :character
##  Median : 2.000        Median : 73.0   Mode  :character
##  Mean   : 4.158        Mean   : 71.3
##  3rd Qu.: 5.000        3rd Qu.: 90.0
##  Max.   :46.000        Max.   :107.0
```

```
#summary output showing that data set does not containing any NA values and I
#think that there is not much difference between the min-max values of the
#integer columns

#changing column names for convenience and easy to understand
colnames(SEER_data)[1:15] <- c("Age","Race","Marital_Status","T_stage","N_stage","sixth_stage","Grade",
  "A_stage","Tumor_size","Estrogen_status","Progesterone_status",
  "Regional_nodes_examined","Regional_nodes_positive","Survival_months","Status")
```

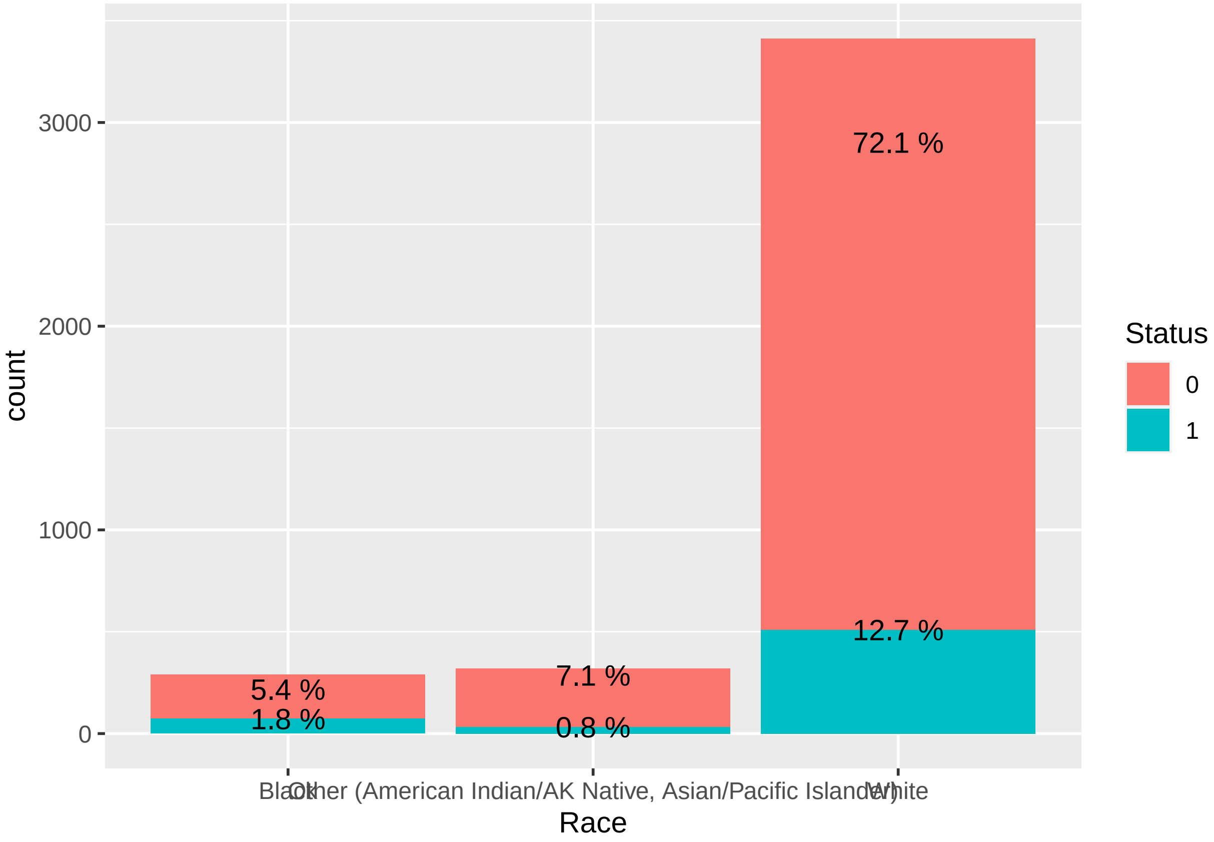
#Data Exploration(EDA)

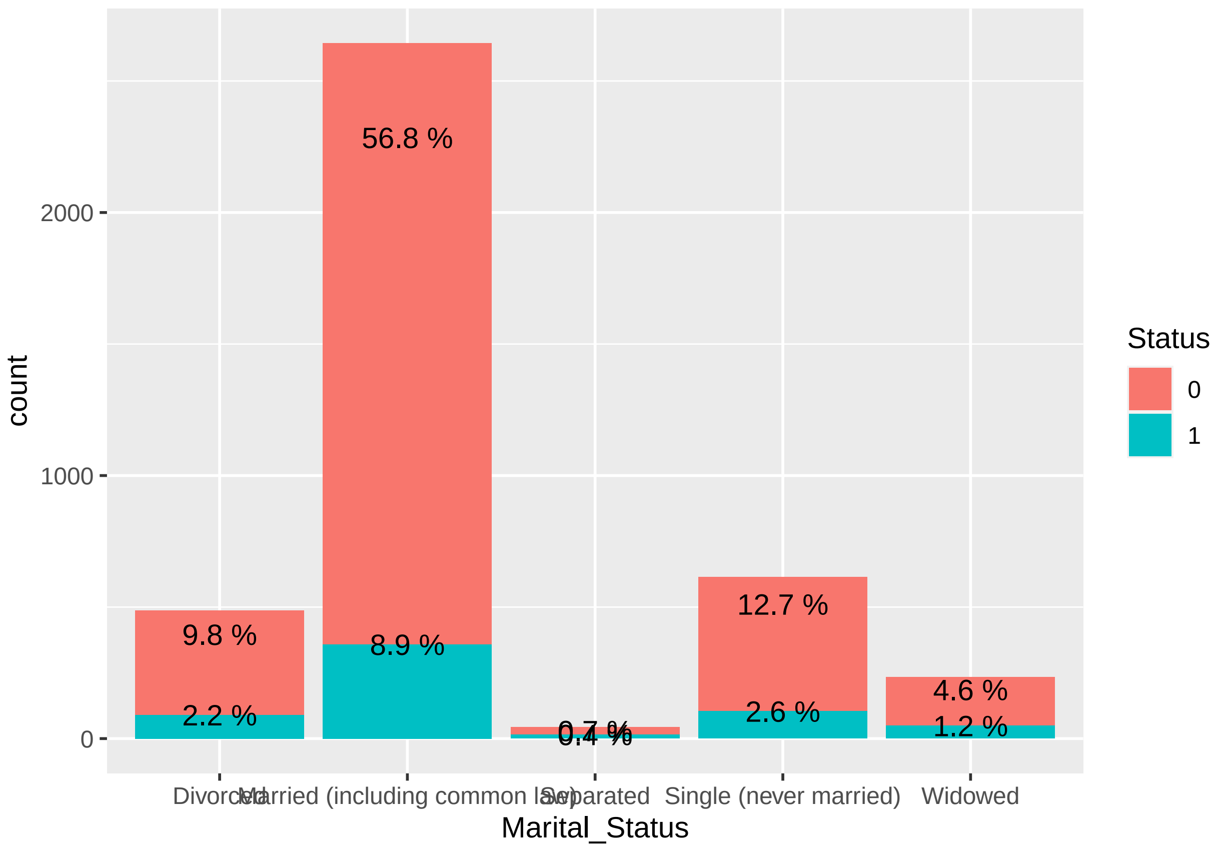
```
#Exploratory data analysis using histograms
#Encoded the target feature Status(Alive=0, Dead=1)

SEER_data$Status <- as.character(SEER_data$Status)
class(SEER_data$Status)
```

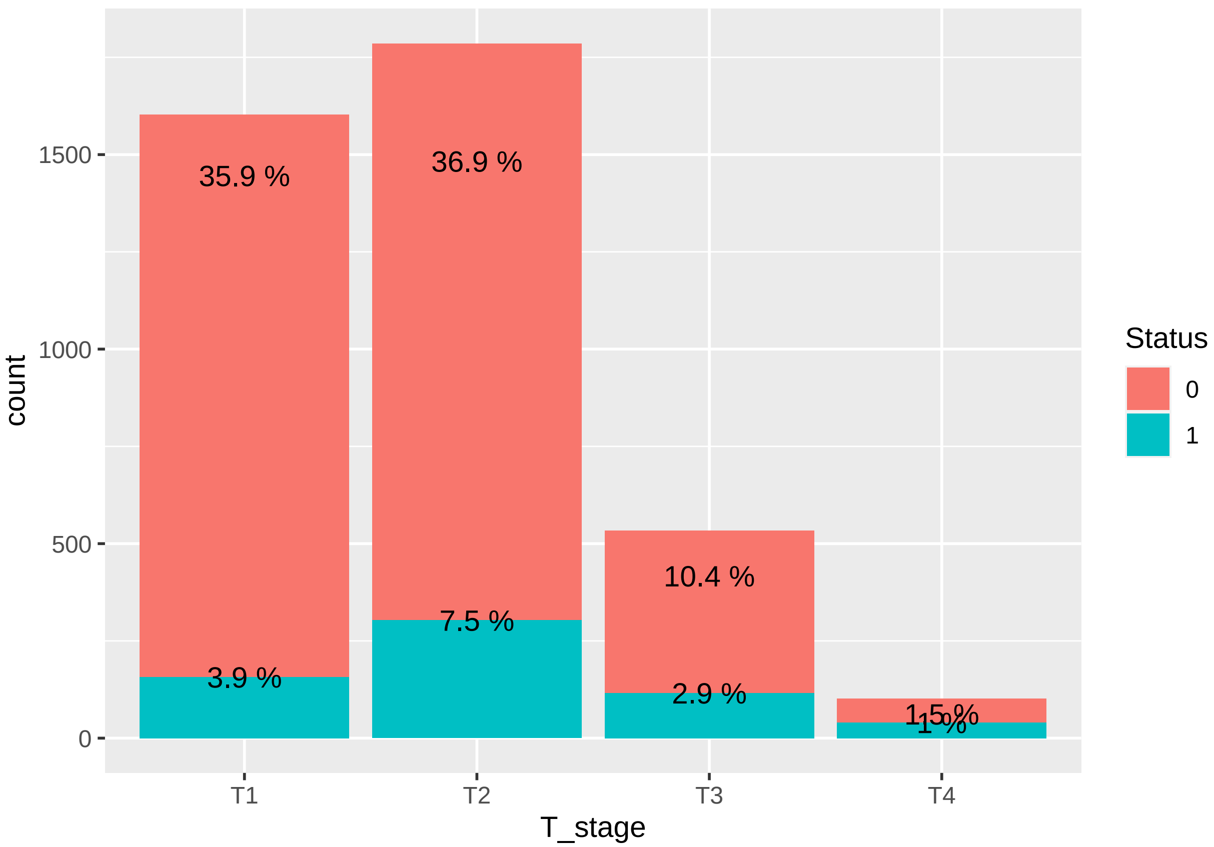
```
## [1] "character"
```

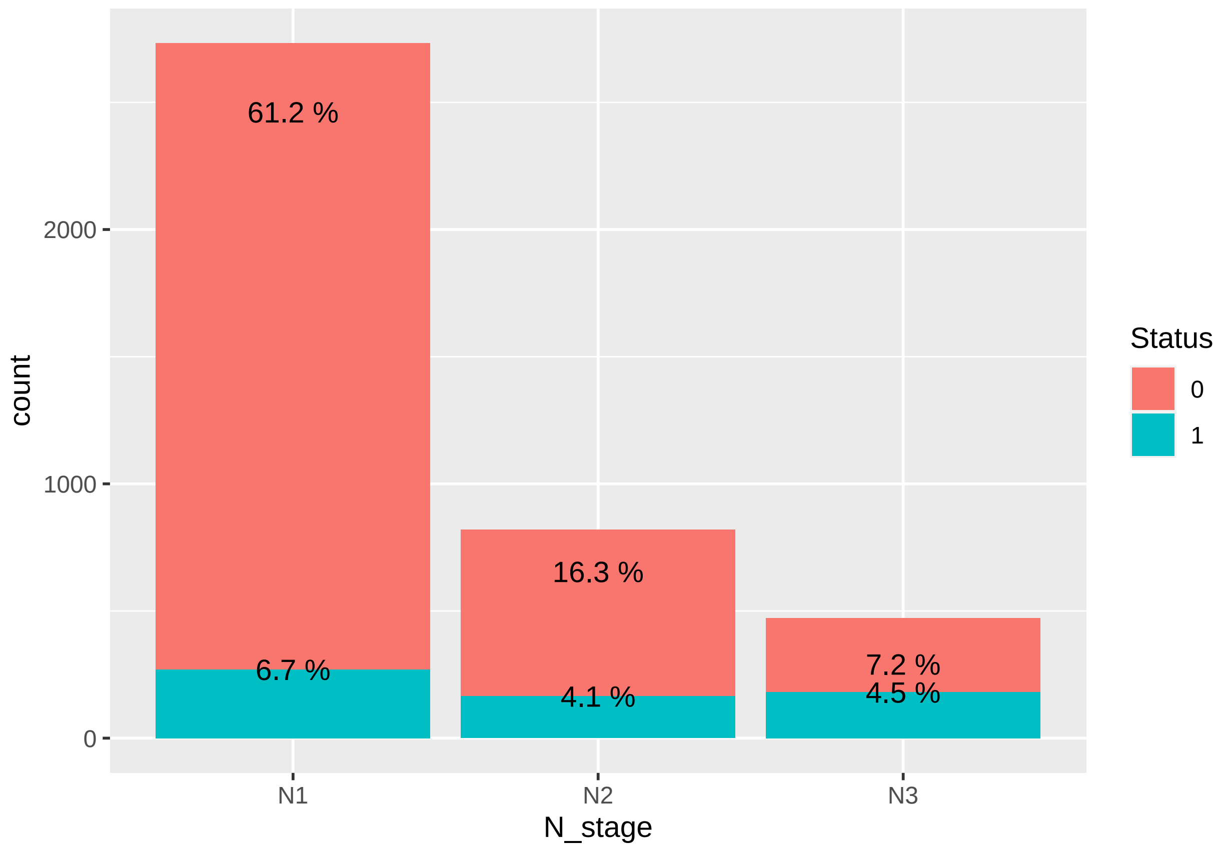
```
#Dummy coding the target variable and assigned Alive to "0" and Dead to "1"
SEER_data$Status[SEER_data$Status=="Alive"] <- 0
SEER_data$Status[SEER_data$Status=="Dead"] <- 1

#Changing the class from character to binary(factor)
SEER_data$Status <- as.factor(SEER_data$Status)


#Factor analysis was done by made bar plots of categorical variables to target
#variable to know how different levels in the categorical variables response to
#the target variable
#Created a list to present the in simple code

cat_variables <- list("Race","Marital_Status","T_stage","N_stage","sixth_stage",
                      "Grade","A_stage","Estrogen_status","Progesterone_status")

#created a for loop to display all the bar plots at once with relation to
#target variable

par(mfrow=c(3,3))
for (i in cat_variables){
 gg_plot <-  ggplot(SEER_data, aes_string(x = i, fill = SEER_data$Status))+
    geom_bar( stat = "count")+ scale_fill_discrete(name = "Status")+geom_text(aes(label=paste(after_sta
                                                                                      ,"%")),
     stat='count',
     nudge_y=0.125)
 print(gg_plot)
}
```

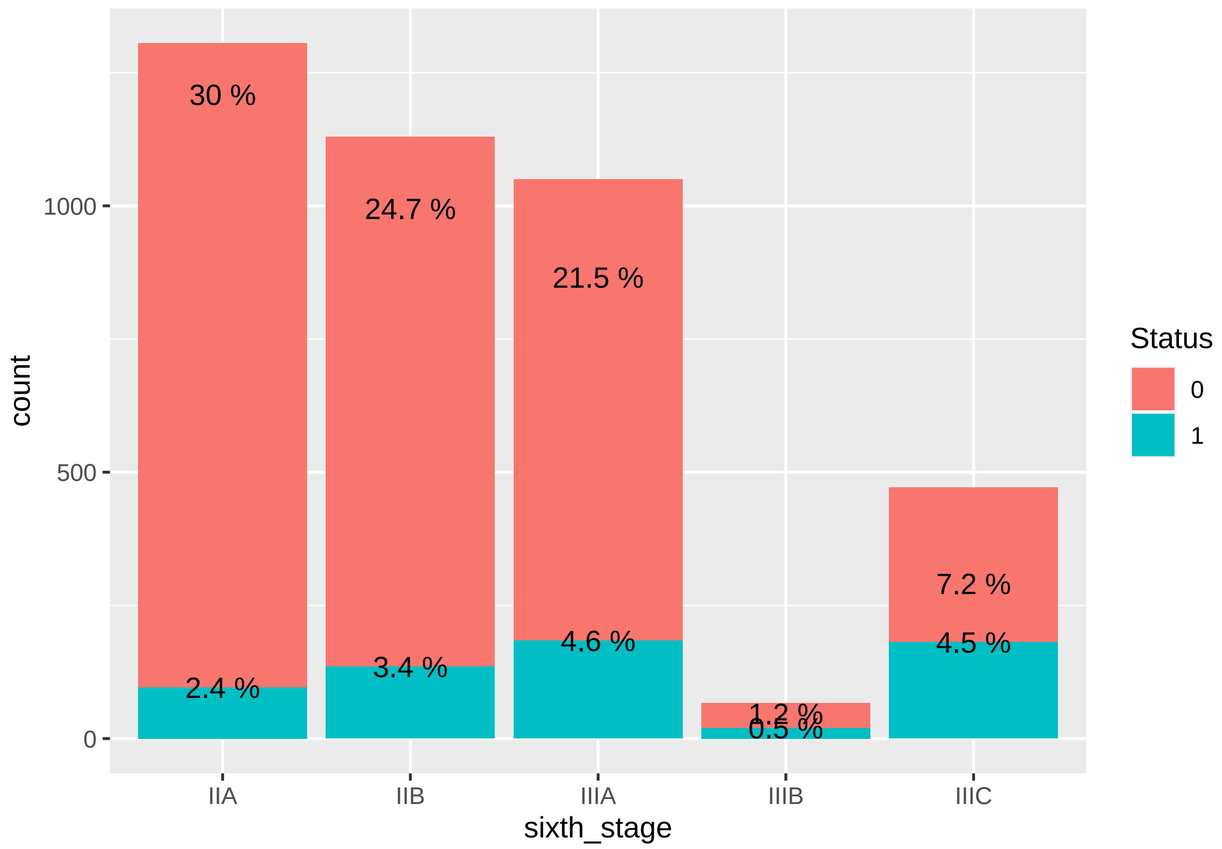
```
## Warning: ‘aes_string()‘ was deprecated in ggplot2 3.0.0.
## i Please use tidy evaluation ideoms with ‘aes()‘
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```

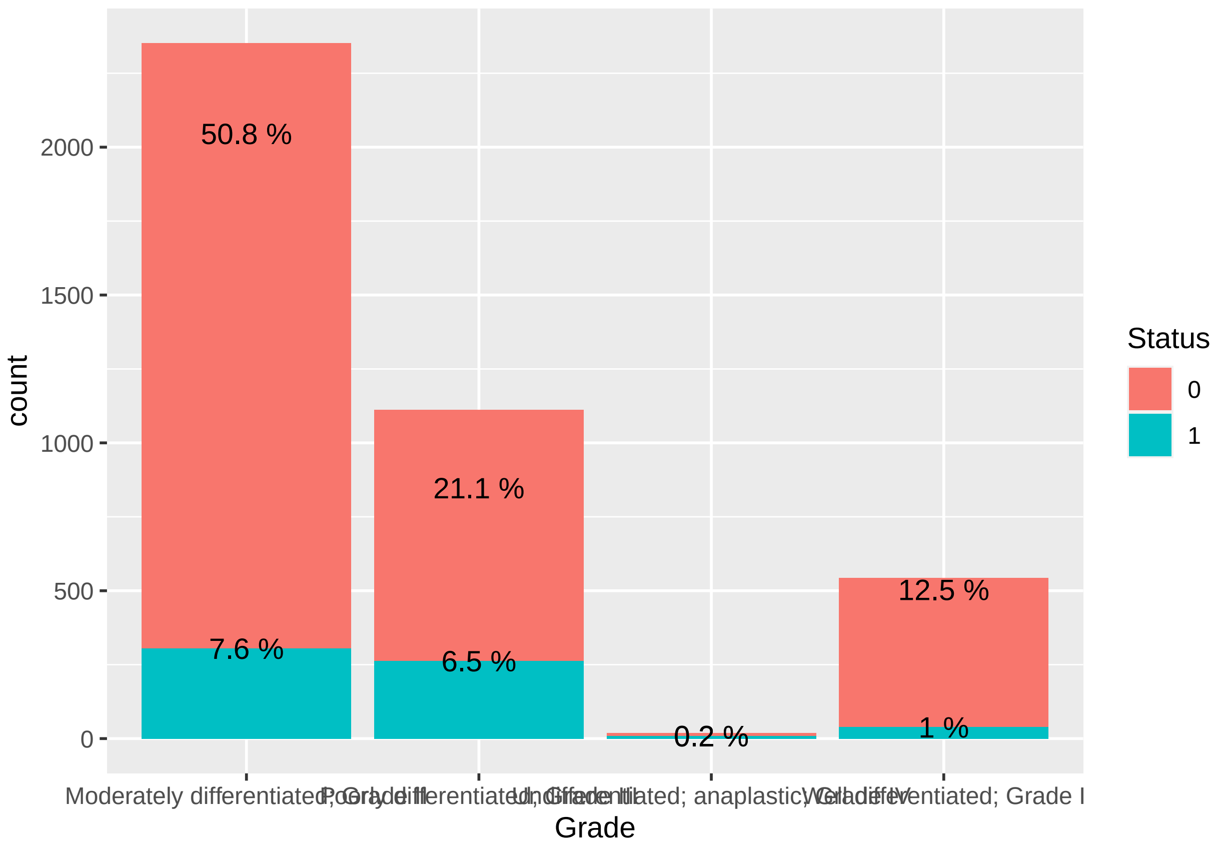

count
3000
2000
1000
0
72.1 %
12.7 %
5.4 %
1.8 %
7.1 %
0.8 %
Black
Other (American Indian/AK Native, Asian/Pacific Islander)
White
Race
Status
0
1

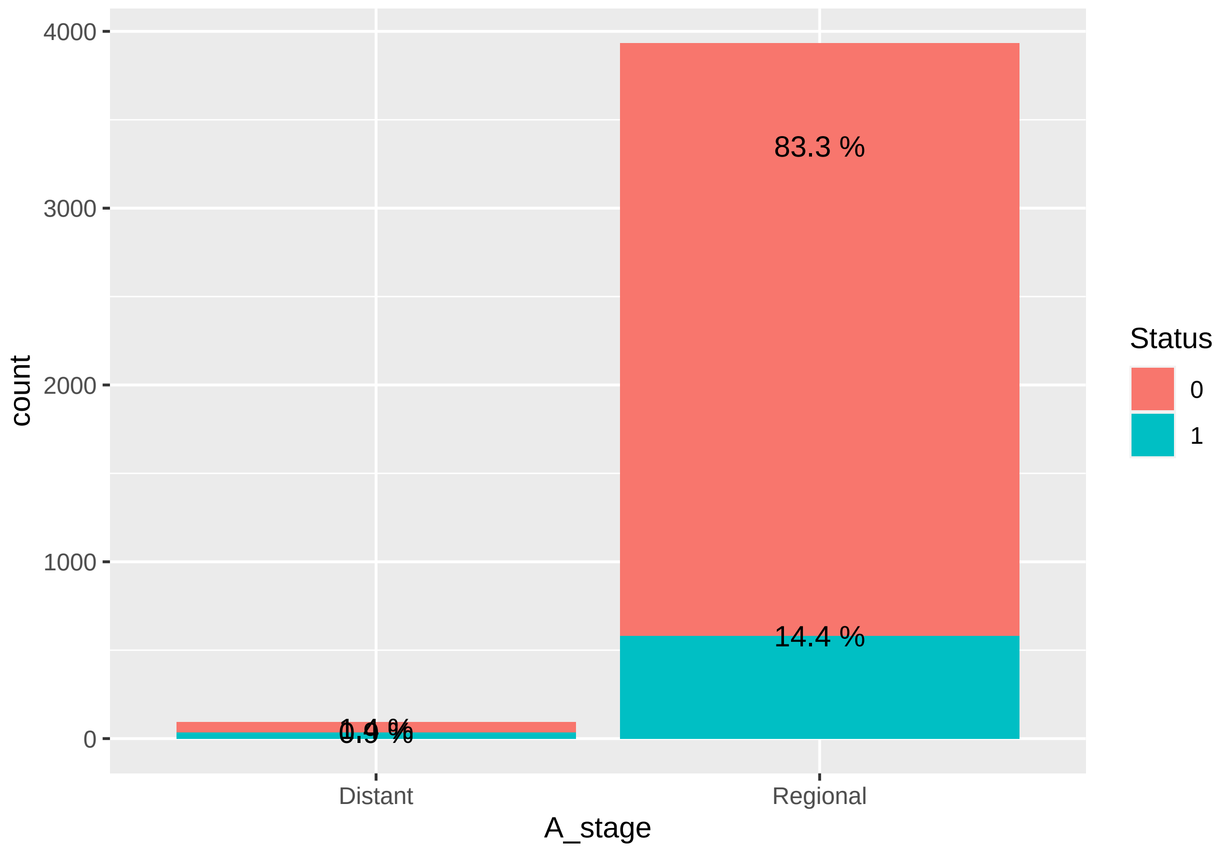

count
2000
1000
0
56.8 %
9.8 %
2.2 %
8.9 %
0.7 %
0.4 %
12.7 %
2.6 %
4.6 %
1.2 %
Divorced
Married (including common law)
Separated
Single (never married)
Widowed
Marital_Status
Status
0
1

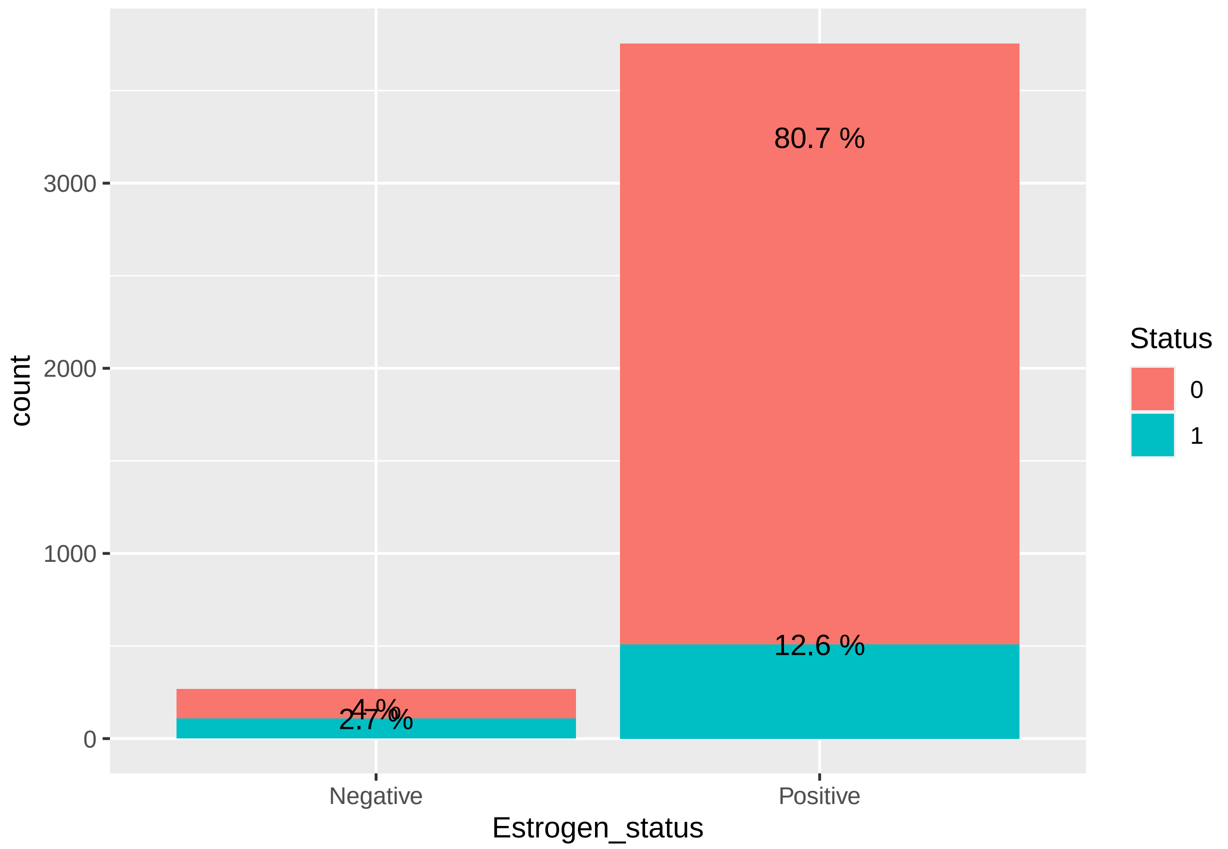

count
1500
1000
500
0
35.9 %
3.9 %
36.9 %
7.5 %
10.4 %
2.9 %
1.5 %
1 %
T1
T2
T3
T4
T_stage
Status
0
1

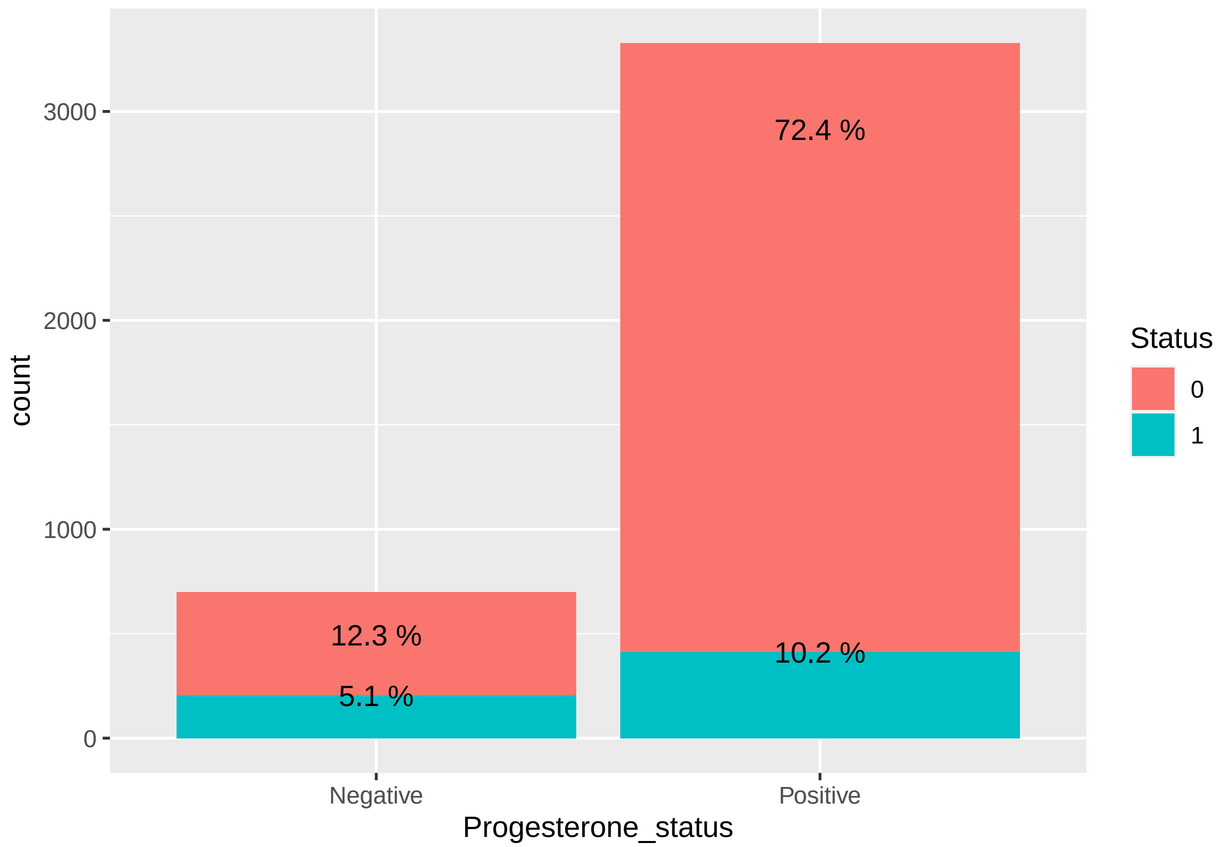

61.2 %
6.7 %
16.3 %
4.1 %
7.2 %
4.5 %
count
2000
1000
0
N1
N2
N3
N_stage
Status
0
1


count
1000
500
0
30 %
2.4 %
24.7 %
3.4 %
21.5 %
4.6 %
1.2 %
0.5 %
7.2 %
4.5 %
IIA
IIB
IIIA
IIIB
IIIC
sixth_stage
Status
0
1


50.8 %
7.6 %
21.1 %
6.5 %
0.2 %
12.5 %
1 %
2000
1500
1000
500
0
count
Moderately differentiated; Grade II
Poorly differentiated; Grade III
Undifferentiated; anaplastic; Grade IV
Well differentiated; Grade I
Grade
Status
0
1


4000
3000
2000
1000
0
count
83.3 %
14.4 %
1.4 %
0.9 %
Distant
Regional
A_stage
Status
0
1

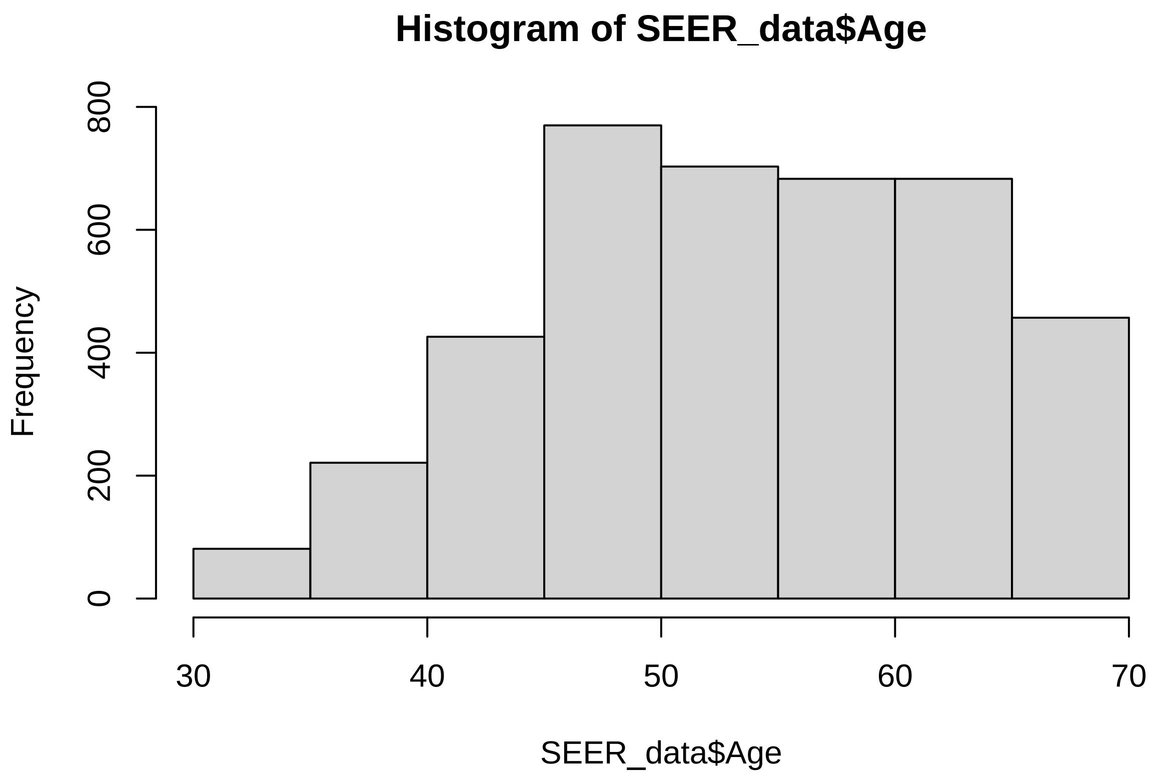

80.7 %
12.6 %
4 %
2.7 %
3000
2000
1000
0
count
Negative
Positive
Estrogen_status
Status
0
1



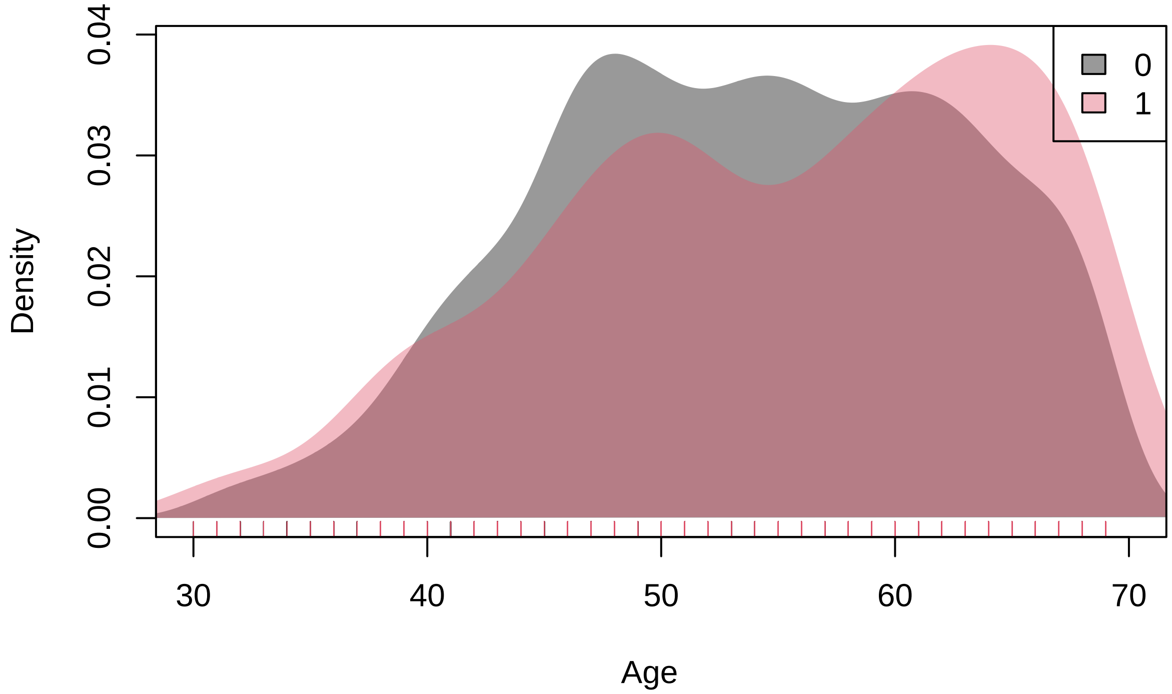
```
paste0("Analysis:From the bar plots, it is very clear that most of the data containing
white married woman(71% and 57%). Where black and other races contribute to very
low percentage. Similarly, separated marital status women are very less compared
to other marital status people.Most of the patients having T1,T2 and N1 stages.
It is indicating that women developing cancer tumor having low size and it's not
wide spread to other areas near the chest.Remaining 1.5% women having T4 stage
and i think those poeple having large tumor with size of > 5cm. Those are most
probably fail the regional_nodes_positive test, because it spread to other organs
in T4 stage. Here one more intersting fact I observed that T4 stage poeple Alive
to dead ratio was 50-50.N1 stage indicated that cancer has spread to near by 1-3
auxillary lymph nodes indicating micrometastasis.As like T4 stage, N3 stage has
also similar ratio to alive and dead patients.Sixth stage groups IIA,IIB and
IIIA constitute majority of the data and same fact repeated here that IIIC stage
people having nearly 60-40% survival rates.Regarding Grade, it might depends on
the above factors as well as estrogen,progesterone levels for each patient. But,
those patients having tumor well differentiated having low survival rates and
majority of the patients were in GradeII(moderately differentiated).Remaining
data features A-stage(Regional), Estrogen(Positive), Progesterone(Positive)
containing most of the patients.")
```

```
## [1] "Analysis:From the bar plots, it is very clear that most of the data containing \nwhite married
```

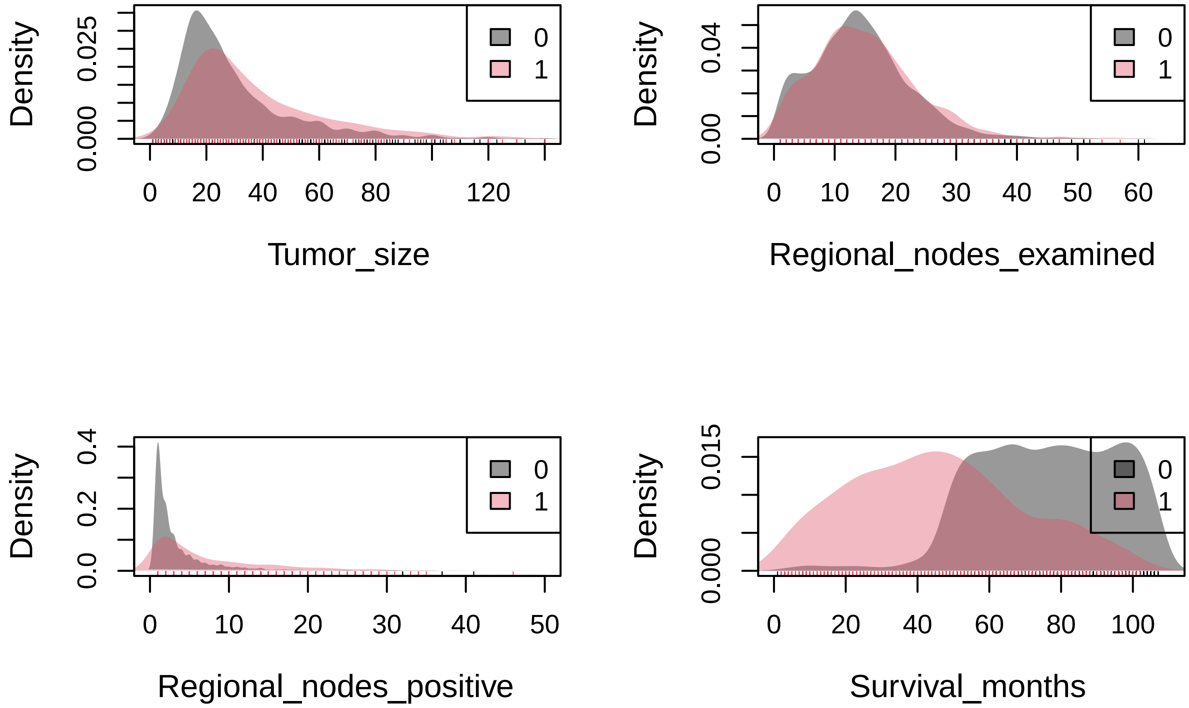
```
#To know the relationship between the continuous variables to target variable
#using byf.hist function to produce dense plots
par(mfrow=c(1,1))
hist(SEER_data$Age)
```



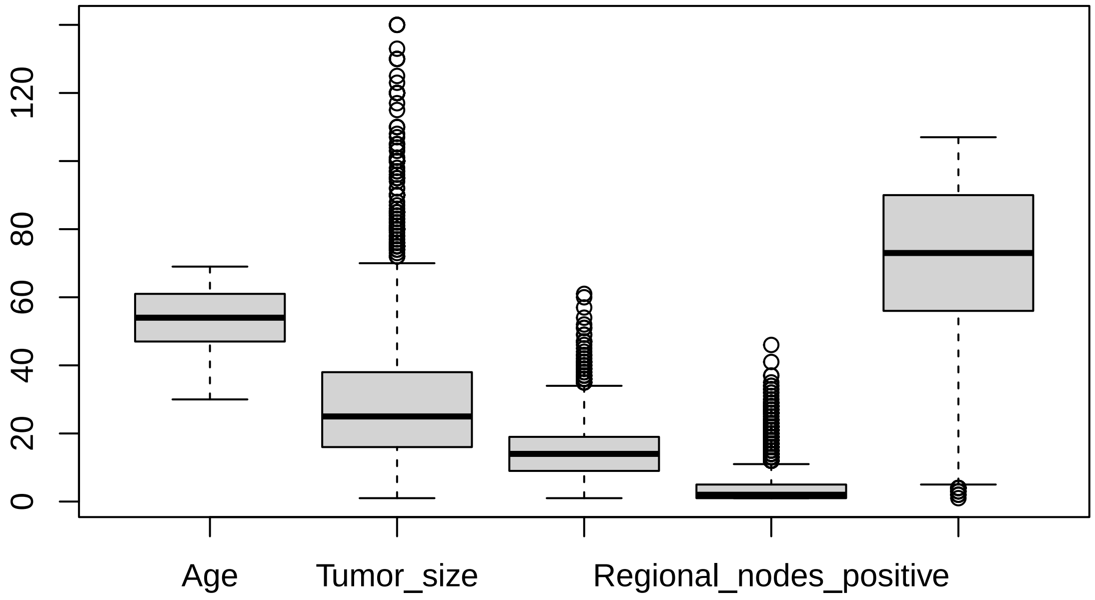
```
byf.hist(Age~Status, data = SEER_data)
```



```
par(mfrow=c(2,2))
byf.hist(Tumor_size~Status, data=SEER_data)
byf.hist(Regional_nodes_examined~Status, data=SEER_data)
byf.hist(Regional_nodes_positive~Status, data=SEER_data)
byf.hist(Survival_months~Status, data=SEER_data)
```



```
#Detection of outliers which are present in the continuous columns. Here I
#used graphical representation box plot to find the outliers in the continuous
#columns
set.seed(123)
par(mfrow=c(1,1))
boxplot(SEER_data[,c("Age","Tumor_size",
  "Regional_nodes_examined","Regional_nodes_positive","Survival_months")])
```



```
#Output"From the plot, it is obvious that columns tumor size,
#regional nodes examined,regional nodes positive and survival months containing
#outliers.It is clear that column regional nodes positive containing large number of #outliers. I had r

for (i in c("Age","Tumor_size",
  "Regional_nodes_examined","Regional_nodes_positive","Survival_months"))
{
  value = SEER_data[,i][SEER_data[,i] %in% boxplot.stats(SEER_data[,i])$out]
  SEER_data[,i][SEER_data[,i] %in% value] = NA
}

#checking the outlier values are replaced with NA values using sum and is.na
#functions. Both tumor size and regional_nodes_positive containing high
#volume of NA values

sum(is.na(SEER_data$Tumor_size))
```

```
## [1] 222
```

```
sum(is.na(SEER_data$Regional_nodes_examined))
```

```
## [1] 72
```

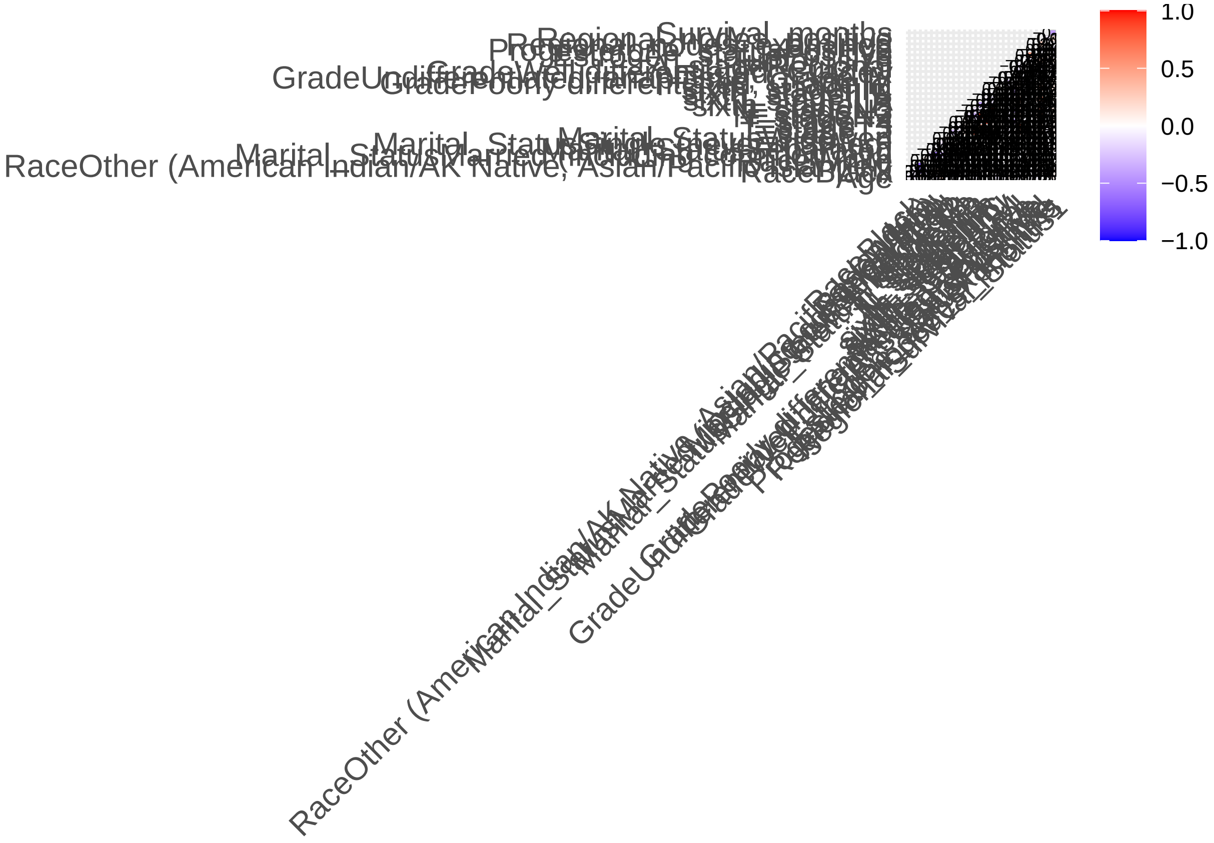
```
sum(is.na(SEER_data$Regional_nodes_positive))
```

```
## [1] 344
```

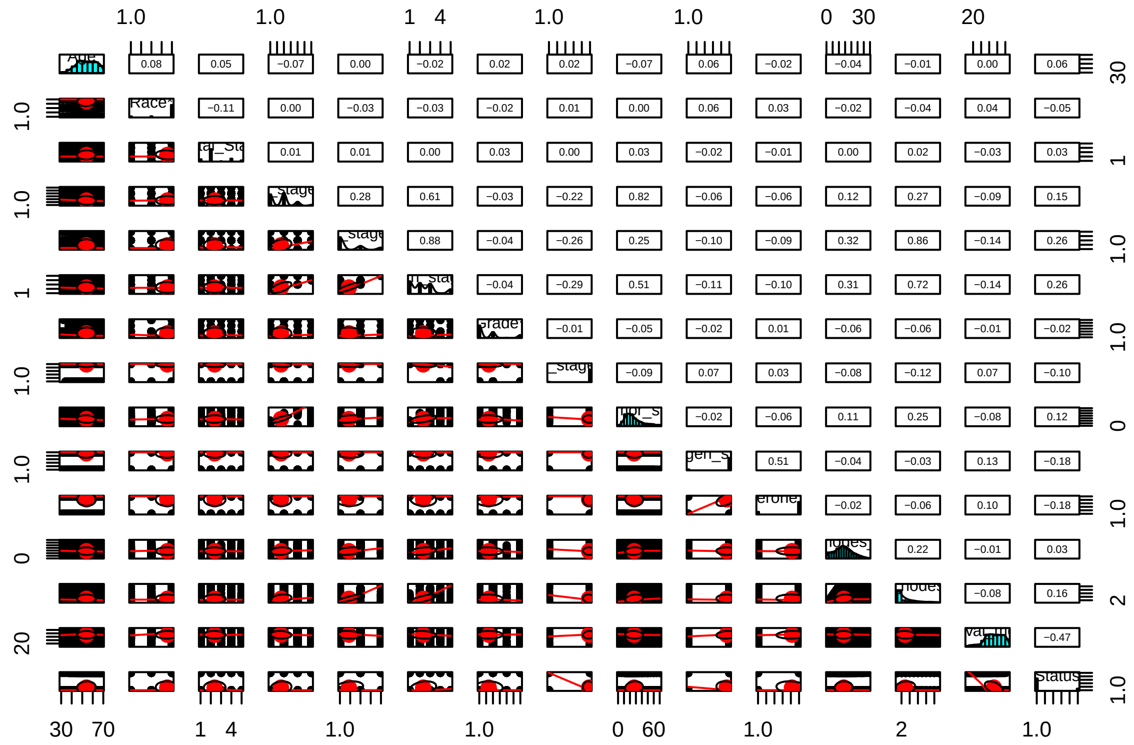
```
sum(is.na(SEER_data$Surival_months))
```

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

```
#correlation
#Here I find correlation using one-hot encoding using model.matrix and plot
#the relations between each variables
model.matrix(~0+., data=SEER_data) %>%
  cor(use="pairwise.complete.obs") %>%
  ggcorrplot(show.diag=FALSE, type="lower", lab=TRUE, lab_size=2)
```



```
#Using another, most commonly used method to find the correlation as well as
#distribution of the data is pairs.panels from the "psych" package

pairs.panels(SEER_data[,c("Age","Race","Marital_Status","T_stage","N_stage",
"sixth_stage","Grade","A_stage","Tumor_size","Estrogen_status","Progesterone_status",
"Regional_nodes_examined","Regional_nodes_positive","Survival_months","Status")])
```



output:It is obvious from these pairs.panels that there is multicollinearity between the independent variables. 6th_stage and N-stage have a very strong positive correlation.N-stage, 6-stage, tumor size,positive regional nodes, nodes examined are all interrelated with each other.

Evaluation of Distribution From the above pairs.panels plot, it is concluded that the continuous columns are distributed differently(skew in the distribution) Near Normal-distribution: Regional_nodes_examined Right-skew-distribution: Regional_nodes_positive, Tumor_size Left_skew_distribution: Survival_months, Age

From the above conclusions, it is mandatory to perform transformation(log, inverse) or standardization of data

#Data cleaning & shaping

##Identify missing values

```
#The above data set containing large volume of NA values and we will identify
#them by using is.na() function

sum(is.na(SEER_data))
```

```
## [1] 656
```

```
#We can observe that the entire data set has significant missing values.
#Therefore, we must replace the missing data in the columns with their
#respective means.
```

#Imputation of data

```
#Replacing large number of missing values in the columns
#with their respective means and survival months containing low volume of
#missing values. So, I decided to remove them instead of keeping them

SEER_data$Regional_nodes_positive[is.na(SEER_data$Regional_nodes_positive)] <- mean(SEER_data$Regional_
SEER_data$Tumor_size[is.na(SEER_data$Tumor_size)] <- mean(SEER_data$Tumor_size, na.rm = TRUE)
SEER_data$Regional_nodes_examined[is.na(SEER_data$Regional_nodes_examined)] <- mean(SEER_data$Regional_

#Removed small volume of missing values in the data column
SEER_breast_cancer_df <- na.omit(SEER_data)


#checking if there any missing values in the data
sum(is.na(SEER_breast_cancer_df))
```

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

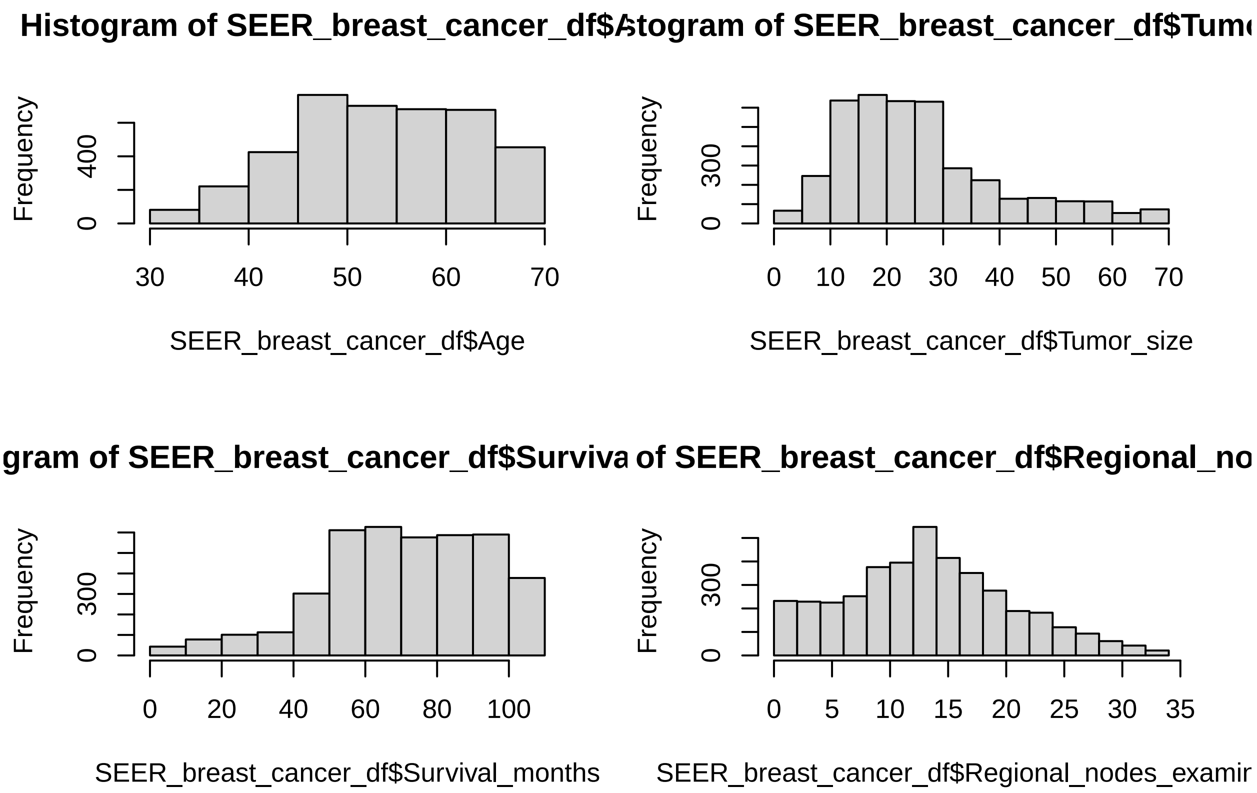
```
#So, now we have zero missing values in the data set. We move forward with
#standardization techniques
```

#Distribution checking

```
#summary stats showing that there is min-max value difference in both tumor_size
#and survival months column
summary(SEER_breast_cancer_df)
```

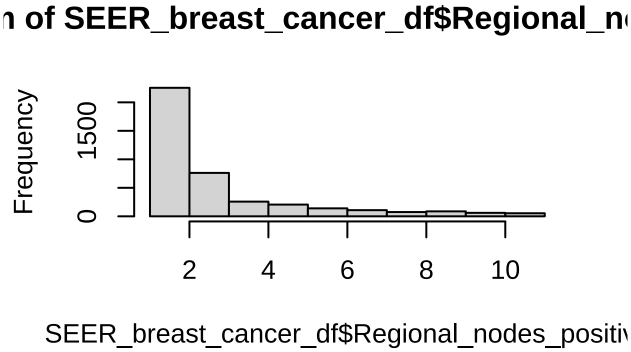
```
##       Age            Race           Marital_Status       T_stage
##  Min.   :30.00   Length:4006        Length:4006        Length:4006
##  1st Qu.:47.00   Class :character   Class :character   Class :character
##  Median :54.00   Mode  :character   Mode  :character   Mode  :character
##  Mean   :53.96
##  3rd Qu.:61.00
##  Max.   :69.00
##    N_stage          sixth_stage          Grade            A_stage
##  Length:4006        Length:4006        Length:4006        Length:4006
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##    Tumor_size   Estrogen_status    Progesterone_status Regional_nodes_examined
##  Min.   : 1.0   Length:4006        Length:4006         Min.   : 1.00
##  1st Qu.:16.0   Class :character   Class :character    1st Qu.: 9.00
##  Median :25.0   Mode  :character   Mode  :character    Median :13.87
##  Mean   :26.9                                          Mean   :13.87
##  3rd Qu.:34.0                                          3rd Qu.:18.00
##  Max.   :70.0                                          Max.   :34.00
##  Regional_nodes_positive Survival_months Status
##  Min.   : 1.000          Min.   :  5.0   0:3403
##  1st Qu.: 1.000          1st Qu.: 56.0   1: 603
##  Median : 2.000          Median : 73.0
##  Mean   : 2.866          Mean   : 71.6
##  3rd Qu.: 3.000          3rd Qu.: 90.0
##  Max.   :11.000          Max.   :107.0
```

```
par(mfrow=c(2,2))
hist(SEER_breast_cancer_df$Age)
hist(SEER_breast_cancer_df$Tumor_size)
hist(SEER_breast_cancer_df$Survival_months)
hist(SEER_breast_cancer_df$Regional_nodes_examined)
```



```
hist(SEER_breast_cancer_df$Regional_nodes_positive)
#From this hist plot, regional_nodes_positive data mostly exist in 1
table(SEER_breast_cancer_df$Regional_nodes_positive)
```

```
##
##                1                2 2.87309782608696                3
##             1515              740              343              419
##                4                5                6                7
##              258              206              140              108
##                8                9               10               11
##               75               87               61               54
```



#Transformation

```
set.seed(123)
#From the above histograms, we have seen that data does not distributed
#normally. We can visually depict that from the above hist plots.So, we need to
#perform transformation(log, sqrt, inverse) for continuous variables to remove
#skewness in the data to perform the model training.
#So standardization does not remove the skewness the data.That's why I chose
#different transformation parameters for different skew's in the data. I took
#these from USCS

#Standardization
#performed min-max normalization on the continuous columns
normalize <- function(x, na.rm = TRUE) {
    return((x- min(x)) /(max(x)-min(x)))
}
SEER_breast_cancer_df[,c(9,12:14)] <- lapply(SEER_breast_cancer_df[,c(9,12:14)],
                                             normalize)
head(SEER_breast_cancer_df)
```

```
##    Age  Race                 Marital_Status T_stage N_stage sixth_stage
## 19  64 White Married (including common law)      T3      N1        IIIA
## 20  31 White                       Divorced      T2      N1         IIB
## 21  31 Black          Single (never married)      T2      N2        IIIA
## 22  41 Black          Single (never married)      T1      N1         IIA
## 23  57 White Married (including common law)      T1      N1         IIA
## 24  40 White         Single (never married)      T1      N1         IIA
##                                   Grade  A_stage Tumor_size Estrogen_status
## 19 Moderately differentiated; Grade II Regional  0.7246377        Positive
## 20 Moderately differentiated; Grade II Regional  0.5942029        Positive
## 21    Poorly differentiated; Grade III Regional  0.4202899        Negative
## 22 Moderately differentiated; Grade II Regional  0.2753623        Negative
## 23 Moderately differentiated; Grade II Regional  0.1739130        Positive
## 24        Well differentiated; Grade I Regional  0.1304348        Positive
##    Progesterone_status Regional_nodes_examined Regional_nodes_positive
## 19            Positive              0.45454545                     0.0
## 20            Positive              0.24242424                     0.2
## 21            Negative              0.36363636                     0.3
## 22            Negative              0.03030303                     0.1
## 23            Positive              0.30303030                     0.0
## 24            Positive              0.06060606                     0.1
##    Survival_months Status
## 19               0      1
## 20               0      0
## 21               0      1
## 22               0      1
## 23               0      1
## 24               0      0
```

```
#Transformation
#From the above plots, columns regional_nodes_positive showing heavy right skew
#and survival column showing moderate right skew
SEER_breast_cancer_df$Regional_nodes_positive <- sqrt(SEER_breast_cancer_df$Regional_nodes_positive)
SEER_breast_cancer_df$Tumor_size <- sqrt(SEER_breast_cancer_df$Tumor_size)


#Comparatively, the normality violation decreased from original data. The
#coefficient values for both Age and regional_nodes_examined features increased
#with transformation. That's the reason I didn't transform them
skewness(SEER_breast_cancer_df$Tumor_size, na.rm = TRUE)
```

```
## [1] 0.2895557
```

```
skewness(SEER_breast_cancer_df$Regional_nodes_examined, na.rm = TRUE)
```

```
## [1] 0.2927328
```

```
skewness(SEER_breast_cancer_df$Regional_nodes_positive, na.rm = TRUE)
```

```
## [1] 0.3988236
```

```
skewness(SEER_breast_cancer_df$Survival_months, na.rm = TRUE)
```

```
## [1] -0.5419102
```

```
skewness(SEER_breast_cancer_df$Age, na.rm = TRUE)
```

```
## [1] -0.2190115
```

```
#Skewness results before transformation
#1.016867
#0.2927328
#1.628163
#-0.5419102
#-0.2190115

#Skewness results after transformation
#0.2895557
#0.2927328
#0.3988236
#-0.5419102
#-0.2190115
```

We got different skewness coefficients and perform transformations accordingly Here, For For Right-skew: log,sqrt,inverse For Left-skew: squares, cubes For Normal-distribution: No parameter required(sqrt)-moderate

For variables with high normality violation value even positive or negative, we should perform inverse transformation. For large violation, use log transformation and for moderate violation we should use sqrt transformation Here Tumor_size(0.91)-sqrt transformation Regional_nodes_examined(0.29)- sqaure root transformation Regional_nodes_positive(1.27)- Inverse transformation Survival_months(-0.54)-log transformation

Note:Here we also check for the linearity and heteroscedasticity, when dependent and independent variables are directly proportional or exhibiting positive correlation, we will first consider "log" transformation.And when they exhibiting negative correlation, we should consider "sqrt" transformation first

#dummy coding

```
#Assigning the transformed data to Encoded_breast_cancer_df
Encoded_breast_cancer_df <- SEER_breast_cancer_df

#using dummyvars from the caret package to perform the dummy coding
Encoded_breast_cancer_df <- as_tibble(predict(
  dummyVars( ~ ., data = Encoded_breast_cancer_df, fullRank = TRUE), newdata = Encoded_breast_cancer_df
#Encoded_breast_cancer_df[,c(2:8,10:11)] <- lapply(Encoded_breast_cancer_df[,c(2:8,10:11)], factor)
head(Encoded_breast_cancer_df)
```

```
## # A tibble: 6 x 27
##     Age RaceOther (American Indian/AK Native,~1 RaceWhite Marital_StatusMarrie~2
##   <dbl>                                   <dbl>     <dbl>                  <dbl>
## 1    64                                       0         1                      1
## 2    31                                       0         1                      0
## 3    31                                       0         0                      0
## 4    41                                       0         0                      0
## 5    57                                       0         1                      1
## 6    40                                       0         1                      0
## # i abbreviated names:
## #   1: `RaceOther (American Indian/AK Native, Asian/Pacific Islander)`,
## #   2: `Marital_StatusMarried (including common law)`
## # i 23 more variables: Marital_StatusSeparated <dbl>,
## #   ‘Marital_StatusSingle (never married)‘ <dbl>, Marital_StatusWidowed <dbl>,
## #   T_stageT2 <dbl>, T_stageT3 <dbl>, T_stageT4 <dbl>, N_stageN2 <dbl>,
## #   N_stageN3 <dbl>, sixth_stageIIB <dbl>, sixth_stageIIIA <dbl>, ...
```

```
factors <- names(which(sapply(Encoded_breast_cancer_df[,-27], is.factor)))

# Label Encoder
for (i in factors){
  encode <- LabelEncoder.fit(Encoded_breast_cancer_df[, i])
  Encoded_breast_cancer_df[, i] <- transform(encode, Encoded_breast_cancer_df[, i])
}
colnames(Encoded_breast_cancer_df)[27] <- "Status"
Encoded_breast_cancer_df$Status <- as.factor(Encoded_breast_cancer_df$Status)
```

#Principal Component Analysis(PCA)

```
#Assigning data set to pca_data
pca_data <- Encoded_breast_cancer_df

#Applying principal component analysis using prcomp from stats on the continuous
#columns
pca_comp <- prcomp(pca_data[,c(1,21,24:26)], center = TRUE)

#summary of the principal components
summary(pca_comp)
```

```
## Importance of components:
##                           PC1     PC2    PC3    PC4     PC5
## Standard deviation     8.9642 0.30873 0.2201 0.2014 0.15981
## Proportion of Variance 0.9974 0.00118 0.0006 0.0005 0.00032
## Cumulative Proportion  0.9974 0.99858 0.9992 0.9997 1.00000
```

```
#First component explains 91% variability and remaining showing similar variability
pca_comp$sdev ^ 2
```

```
## [1] 80.35708423  0.09531487  0.04842750  0.04056164  0.02553991
```

```
print(pca_comp$rotation)
```

```
##                                  PC1          PC2           PC3           PC4
## Age                     -9.999983e-01  0.001389467 -0.0002942179 -0.0003607966
## Tumor_size               1.383738e-03  0.207837059  0.0554471501  0.0467313855
## Regional_nodes_examined  8.590162e-04  0.339523296 -0.2648284779 -0.9024364338
## Regional_nodes_positive  9.069699e-04  0.907676614 -0.0543127427  0.3606673780
## Survival_months          9.799518e-05 -0.132849248 -0.9611667547  0.2309622757
##                                 PC5
## Age                     -0.001156494
## Tumor_size              -0.975470980
## Regional_nodes_examined  0.014055387
## Regional_nodes_positive  0.207584947
## Survival_months         -0.071874646
```

```
#From this pca data, all the variables contribute similarly in different principal
#components.So, I decided to choose all the variables in training the model.
```

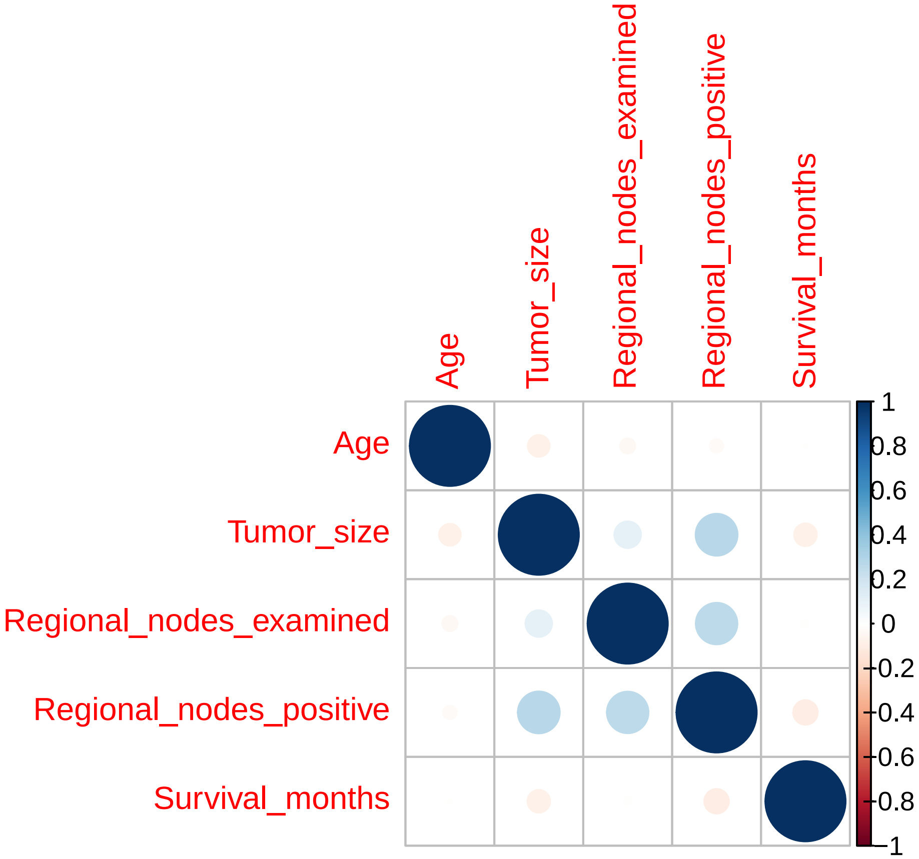
output: From this pca data, all the variables contribute similarly in different principal components.So, I decided to choose all the variables in training the model.

#Feature Engineering

```
cor_Seer <- cor(Encoded_breast_cancer_df[,c(1,21,24:26)])
cor_Seer
```

```
##                                  Age  Tumor_size Regional_nodes_examined
## Age                      1.000000000 -0.07304207             -0.03530731
## Tumor_size              -0.073042072  1.00000000              0.10948978
## Regional_nodes_examined -0.035307313  0.10948978              1.00000000
## Regional_nodes_positive -0.027826093  0.27156960              0.26750438
## Survival_months         -0.003982507 -0.07936258             -0.00926325
##                         Regional_nodes_positive Survival_months
## Age                                 -0.02782609    -0.003982507
## Tumor_size                           0.27156960    -0.079362584
## Regional_nodes_examined              0.26750438    -0.009263250
## Regional_nodes_positive              1.00000000    -0.092563028
## Survival_months                     -0.09256303     1.000000000
```

```
corrplot(cor_Seer)
```



```
#From the corr plot, we have seen that variables tumor_size and
#regional_nodes_positive are positively correlated and regional_nodes_examined
#and regional_nodes_positive are correlated with each other. Feature
#survival_months and age both have no relationship with other variables.

#Need to find correlation between race & marital status to tumor size and regional
#nodes positive
```

#Splitting data

```
#For the given data set, I choose 80% training and 20% validation data set
#splitting
#Imbalanced data splitting
set.seed(123)
without_smote <- createDataPartition(Encoded_breast_cancer_df$Status, p=0.8,
                                     list=FALSE)

train_breast_cancer <- Encoded_breast_cancer_df[without_smote,]
prop.table(table(train_breast_cancer$Status))
```

```
##
##        0        1
## 0.849345 0.150655
```

```
test_breast_cancer <- Encoded_breast_cancer_df[-without_smote,]


#splitting the data and balancing the train data set which containing
#difference in dependent variable
train_breast_cancer_smote <- DMwR::SMOTE(Status~., data=as.data.frame(Encoded_breast_cancer_df[without_s
prop.table(table(train_breast_cancer_smote$Status))
```

```
##
##         0         1
## 0.5714286 0.4285714
```

```
#Here, we see that train data before and after balancing with smote function.
#Now,we perform the models using both imbalanced and balanced training data
```

#selecting models for data set

```
#It's actually one of the difficult task to select the appropriate algorithm for
#specific data and depends on circumstances we need. Here, I decided to go
#through the some of classification algorithms such as Logistic Regression,
#Decision Trees, Support Vector Machine (SVM), Random Forest (RF). It's all
#about #trial-and-error process and finally compare the all the models by specific
#parameters and concluded to one model. Particularly, I chose above because my
#data set containing both categorical and continuous variables and my target
#variable is categorical(binary).
```

#Model1(support vector machines)#Imbalanced data(Model Training)

```
set.seed(123)
#Performing SVM(Support Vector Machines) algorithm on imbalanced training data
svm_model_imb <- ksvm(Status~., data=train_breast_cancer, kernel="vanilladot")
```

```
##  Setting default kernel parameters
```

```
#Evaluating model on test data(unseen data)
fit_svm_imb <- predict(svm_model_imb, test_breast_cancer)

#Table to calculate the Accuracy, precision, recall and F-scores
tab_imb <- table(fit_svm_imb, test_breast_cancer$Status)
svm_cm_imb <- confusionMatrix(tab_imb, positive = "0")
svm_cm_imb
```

```
## Confusion Matrix and Statistics
##
##
## fit_svm_imb   0   1
##           0 670  76
##           1  10  44
##
##                Accuracy : 0.8925
##                  95% CI : (0.8689, 0.9131)
##     No Information Rate : 0.85
##     P-Value [Acc > NIR] : 0.0002797
##
##                   Kappa : 0.455
##
##  Mcnemar's Test P-Value : 2.398e-12
##
##             Sensitivity : 0.9853
##             Specificity : 0.3667
##          Pos Pred Value : 0.8981
##          Neg Pred Value : 0.8148
##              Prevalence : 0.8500
##          Detection Rate : 0.8375
##    Detection Prevalence : 0.9325
##       Balanced Accuracy : 0.6760
##
##        'Positive' Class : 0
##
```

```
table(test_breast_cancer$Status)
```

```
##
##   0   1
## 680 120
```

```
#Results
print(paste("For Imbalanced data:", "Precision is:",caret::precision(tab_imb),
"Recall is:",sensitivity(tab_imb),"F-score is:",caret::F_meas(tab_imb)))
```

```
## [1] "For Imbalanced data: Precision is: 0.898123324396783 Recall is: 0.985294117647059 F-score is: 0
```

#Model1(Support Vector Machines)#Balanced data(Model Training)

```
set.seed(123)
#Performing SVM(Support Vector Machines) algorithm on balanced training data
svm_model_bal <- ksvm(Status~., data=train_breast_cancer_smote, kernel="vanilladot")
```

```
##  Setting default kernel parameters
```

```
#Evaluating model on test data(unseen data)
fit_svm_bal <- predict(svm_model_bal, test_breast_cancer)

#Table to calculate the Accuracy, precision, recall and F-scores
tab_bal <- table(fit_svm_bal, test_breast_cancer$Status)

svm_cm_bal <- confusionMatrix(tab_bal, positive = "0")
svm_cm_bal
```

```
## Confusion Matrix and Statistics
##
##
## fit_svm_bal   0   1
##           0 592  40
##           1  88  80
##
##                Accuracy : 0.84
##                  95% CI : (0.8127, 0.8647)
##     No Information Rate : 0.85
##     P-Value [Acc > NIR] : 0.801
##
##                   Kappa : 0.4613
##
##  Mcnemar's Test P-Value : 3.264e-05
##
##             Sensitivity : 0.8706
##             Specificity : 0.6667
##          Pos Pred Value : 0.9367
##          Neg Pred Value : 0.4762
##              Prevalence : 0.8500
##          Detection Rate : 0.7400
##    Detection Prevalence : 0.7900
##       Balanced Accuracy : 0.7686
##
##        'Positive' Class : 0
##
```

```
#Results
print(paste("For Imbalanced data:", "Precision is:",caret::precision(tab_bal),
"Recall is:",sensitivity(tab_bal),"F-score is:",caret::F_meas(tab_bal)))
```

```
## [1] "For Imbalanced data: Precision is: 0.936708860759494 Recall is: 0.870588235294118 F-score is: 0
```

#Model2(Decision trees)#Imbalanced data(Model Training)

```
set.seed(123)
#imbalanced data
dt_imb <- C5.0(train_breast_cancer[,-27], train_breast_cancer$Status)

fit_dt_imb <- predict(dt_imb, test_breast_cancer)

dt_imb_tab <- table(fit_dt_imb, test_breast_cancer$Status)

dt_cm_imb <- confusionMatrix(dt_imb_tab)

dt_cm_imb
```

```
## Confusion Matrix and Statistics
##
##
## fit_dt_imb   0   1
##          0 673  66
##          1   7  54
##
##                Accuracy : 0.9087
##                  95% CI : (0.8866, 0.9278)
##     No Information Rate : 0.85
##     P-Value [Acc > NIR] : 5.013e-07
##
##                   Kappa : 0.5513
##
##  Mcnemar's Test P-Value : 1.134e-11
##
##             Sensitivity : 0.9897
##             Specificity : 0.4500
##          Pos Pred Value : 0.9107
##          Neg Pred Value : 0.8852
##              Prevalence : 0.8500
##          Detection Rate : 0.8413
##    Detection Prevalence : 0.9237
##       Balanced Accuracy : 0.7199
##
##        'Positive' Class : 0
##
```

```
CrossTable(test_breast_cancer$Status, fit_dt_imb,
 prop.chisq = FALSE, prop.c = FALSE, prop.r = FALSE,
 dnn = c('actual default', 'predicted default'))
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |         N / Table Total |
## |-------------------------|
## 
## 
## Total Observations in Table:  800 
## 
## 
##                | predicted default 
## actual default |         0 |         1 | Row Total | 
## ---------------|-----------|-----------|-----------|
##              0 |       673 |         7 |       680 | 
##                |     0.841 |     0.009 |           | 
## ---------------|-----------|-----------|-----------|
##              1 |        66 |        54 |       120 | 
##                |     0.082 |     0.068 |           | 
## ---------------|-----------|-----------|-----------|
##   Column Total |       739 |        61 |       800 | 
## ---------------|-----------|-----------|-----------|
## 
## 
```

```
#Results
print(paste("For Imbalanced data:", "Precision is:",caret::precision(dt_imb_tab),
"Recall is:",sensitivity(dt_imb_tab),"F-score is:",caret::F_meas(dt_imb_tab)))
```

```
## [1] "For Imbalanced data: Precision is: 0.910690121786198 Recall is: 0.989705882352941 F-score is: 0
```

#Model2(Decision trees)#Balanced data(Model Training)

```
set.seed(123)
#imbalanced data
dt_bal <- C5.0(train_breast_cancer_smote[,-27], train_breast_cancer_smote$Status)

fit_dt_bal <- predict(dt_bal, test_breast_cancer)

dt_bal_tab <- table(fit_dt_bal, test_breast_cancer$Status)

dt_cm_bal <- confusionMatrix(dt_bal_tab, positive = "0")

dt_cm_bal
```

```
## Confusion Matrix and Statistics
## 
##           
## fit_dt_bal   0   1
##          0 627  46
##          1  53  74
## 
##                Accuracy : 0.8762          
##                  95% CI : (0.8514, 0.8983)
##     No Information Rate : 0.85            
##     P-Value [Acc > NIR] : 0.01926         
## 
##                   Kappa : 0.5261          
## 
```

```
##  Mcnemar’s Test P-Value : 0.54649
##
##             Sensitivity : 0.9221
##             Specificity : 0.6167
##          Pos Pred Value : 0.9316
##          Neg Pred Value : 0.5827
##              Prevalence : 0.8500
##          Detection Rate : 0.7837
##    Detection Prevalence : 0.8413
##       Balanced Accuracy : 0.7694
##
##        ’Positive’ Class : 0
##
```

```
CrossTable(test_breast_cancer$Status, fit_dt_bal,
 prop.chisq = FALSE, prop.c = FALSE, prop.r = FALSE,
 dnn = c('actual default', 'predicted default'))
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  800
##
##
##                | predicted default
## actual default |         0 |         1 | Row Total |
## ---------------|-----------|-----------|-----------|
##              0 |       627 |        53 |       680 |
##                |     0.784 |     0.066 |           |
## ---------------|-----------|-----------|-----------|
##              1 |        46 |        74 |       120 |
##                |     0.058 |     0.092 |           |
## ---------------|-----------|-----------|-----------|
##   Column Total |       673 |       127 |       800 |
## ---------------|-----------|-----------|-----------|
##
##
```

```
#Results
print(paste("For Imbalanced data:", "Precision is:",caret::precision(dt_bal_tab),
"Recall is:",sensitivity(dt_bal_tab),"F-score is:",caret::F_meas(dt_bal_tab)))
```

```
## [1] "For Imbalanced data: Precision is: 0.931649331352155 Recall is: 0.922058823529412 F-score is: 0
```

#Model3(Logistic Regression)#Imbalanced data(model training)

```r
#Performing logistic regression model on imbalanced data
log_imb <- glm(formula = Status ~ ., family = binomial(link = "logit"),
    data = train_breast_cancer)

fitted.results <- predict(log_imb,newdata=test_breast_cancer,type='response')
```

```
## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (type == :
## prediction from a rank-deficient fit may be misleading
```

```r
fitted.results <- ifelse(fitted.results > 0.5,1,0)
tab_fitted <- table(fitted.results, test_breast_cancer$Status)

misClasificError <- mean(fitted.results != test_breast_cancer$Status)
print(paste('Accuracy',1-misClasificError))
```

```
## [1] "Accuracy 0.9"
```

```r
#Backward Elimination
Backward_log_imb <- step(log_imb, direction = "backward", trace = TRUE)
```

```
## Start:  AIC=1827.05
## Status ~ Age + ‘RaceOther (American Indian/AK Native, Asian/Pacific Islander)‘ +
##     RaceWhite + ‘Marital_StatusMarried (including common law)‘ +
##     Marital_StatusSeparated + ‘Marital_StatusSingle (never married)‘ +
##     Marital_StatusWidowed + T_stageT2 + T_stageT3 + T_stageT4 +
##     N_stageN2 + N_stageN3 + sixth_stageIIB + sixth_stageIIIA +
##     sixth_stageIIIB + sixth_stageIIIC + ‘GradePoorly differentiated; Grade III‘ +
##     ‘GradeUndifferentiated; anaplastic; Grade IV‘ + ‘GradeWell differentiated; Grade I‘ +
##     A_stageRegional + Tumor_size + Estrogen_statusPositive +
##     Progesterone_statusPositive + Regional_nodes_examined + Regional_nodes_positive +
##     Survival_months
##
##
## Step:  AIC=1827.05
## Status ~ Age + ‘RaceOther (American Indian/AK Native, Asian/Pacific Islander)‘ +
##     RaceWhite + ‘Marital_StatusMarried (including common law)‘ +
##     Marital_StatusSeparated + ‘Marital_StatusSingle (never married)‘ +
##     Marital_StatusWidowed + T_stageT2 + T_stageT3 + T_stageT4 +
##     N_stageN2 + N_stageN3 + sixth_stageIIB + sixth_stageIIIA +
##     sixth_stageIIIB + ‘GradePoorly differentiated; Grade III‘ +
##     ‘GradeUndifferentiated; anaplastic; Grade IV‘ + ‘GradeWell differentiated; Grade I‘ +
##     A_stageRegional + Tumor_size + Estrogen_statusPositive +
##     Progesterone_statusPositive + Regional_nodes_examined + Regional_nodes_positive +
##     Survival_months
##
##                                                                   Df Deviance
## - ‘Marital_StatusMarried (including common law)‘                   1   1775.0
## - sixth_stageIIIA                                                  1   1775.1
## - sixth_stageIIIB                                                  1   1775.4
## - Regional_nodes_positive                                          1   1775.4
## - T_stageT2                                                        1   1775.6
## - ‘Marital_StatusSingle (never married)‘                           1   1775.7
## - Marital_StatusWidowed                                            1   1775.9
## - A_stageRegional                                                  1   1776.6
## - RaceWhite                                                        1   1776.7
## <none>                                                                 1775.0
## - Tumor_size                                                       1   1777.4
## - sixth_stageIIB                                                   1   1777.5
## - Marital_StatusSeparated                                          1   1779.6
## - T_stageT3                                                        1   1779.8
## - Estrogen_statusPositive                                          1   1779.8
## - ‘RaceOther (American Indian/AK Native, Asian/Pacific Islander)‘  1   1780.0
## - Regional_nodes_examined                                          1   1781.8
## - ‘GradeUndifferentiated; anaplastic; Grade IV‘                    1   1782.2
## - T_stageT4                                                        1   1782.4
## - N_stageN2                                                        1   1782.9
## - ‘GradeWell differentiated; Grade I‘                              1   1783.9
## - ‘GradePoorly differentiated; Grade III‘                          1   1784.0
## - Progesterone_statusPositive                                      1   1784.2
## - Age                                                              1   1789.3
## - N_stageN3                                                        1   1804.6
## - Survival_months                                                  1   2327.6
##                                                                      AIC
## - ‘Marital_StatusMarried (including common law)‘                  1825.0
## - sixth_stageIIIA                                                 1825.1
## - sixth_stageIIIB                                                 1825.4
## - Regional_nodes_positive                                         1825.4
## - T_stageT2                                                       1825.6
## - ‘Marital_StatusSingle (never married)‘                          1825.7
## - Marital_StatusWidowed                                           1825.9
## - A_stageRegional                                                 1826.6
## - RaceWhite                                                       1826.7
## <none>                                                            1827.0
## - Tumor_size                                                      1827.4
## - sixth_stageIIB                                                  1827.5
## - Marital_StatusSeparated                                         1829.6
## - T_stageT3                                                       1829.8
## - Estrogen_statusPositive                                         1829.8
## - ‘RaceOther (American Indian/AK Native, Asian/Pacific Islander)‘ 1830.0
## - Regional_nodes_examined                                         1831.8
## - ‘GradeUndifferentiated; anaplastic; Grade IV‘                   1832.2
## - T_stageT4                                                       1832.4
## - N_stageN2                                                       1832.9
## - ‘GradeWell differentiated; Grade I‘                             1833.9
## - ‘GradePoorly differentiated; Grade III‘                         1834.0
## - Progesterone_statusPositive                                     1834.2
## - Age                                                             1839.3
## - N_stageN3                                                       1854.6
## - Survival_months                                                 2377.6
##
## Step:  AIC=1825.05
## Status ~ Age + ‘RaceOther (American Indian/AK Native, Asian/Pacific Islander)‘ +
##     RaceWhite + Marital_StatusSeparated + ‘Marital_StatusSingle (never married)‘ +
##     Marital_StatusWidowed + T_stageT2 + T_stageT3 + T_stageT4 +
##     N_stageN2 + N_stageN3 + sixth_stageIIB + sixth_stageIIIA +
##     sixth_stageIIIB + ‘GradePoorly differentiated; Grade III‘ +
##     `GradeUndifferentiated; anaplastic; Grade IV` + `GradeWell differentiated; Grade I` +
##     A_stageRegional + Tumor_size + Estrogen_statusPositive +
##     Progesterone_statusPositive + Regional_nodes_examined + Regional_nodes_positive +
##     Survival_months
##
##                                                                   Df Deviance
## - sixth_stageIIIA                                                 1   1775.1
## - sixth_stageIIIB                                                 1   1775.4
## - Regional_nodes_positive                                         1   1775.4
## - T_stageT2                                                       1   1775.6
## - `Marital_StatusSingle (never married)`                          1   1776.2
## - Marital_StatusWidowed                                           1   1776.2
## - A_stageRegional                                                 1   1776.6
## - RaceWhite                                                       1   1776.7
## <none>                                                                1775.0
## - Tumor_size                                                      1   1777.4
## - sixth_stageIIB                                                  1   1777.5
## - T_stageT3                                                       1   1779.8
## - Estrogen_statusPositive                                         1   1779.8
## - Marital_StatusSeparated                                         1   1779.9
## - `RaceOther (American Indian/AK Native, Asian/Pacific Islander)` 1   1780.0
## - Regional_nodes_examined                                         1   1781.8
## - `GradeUndifferentiated; anaplastic; Grade IV`                   1   1782.2
## - T_stageT4                                                       1   1782.4
## - N_stageN2                                                       1   1782.9
## - `GradeWell differentiated; Grade I`                             1   1783.9
## - `GradePoorly differentiated; Grade III`                         1   1784.0
## - Progesterone_statusPositive                                     1   1784.2
## - Age                                                             1   1789.3
## - N_stageN3                                                       1   1804.7
## - Survival_months                                                 1   2327.9
##                                                                      AIC
## - sixth_stageIIIA                                                 1823.1
## - sixth_stageIIIB                                                 1823.4
## - Regional_nodes_positive                                         1823.4
## - T_stageT2                                                       1823.6
## - `Marital_StatusSingle (never married)`                          1824.2
## - Marital_StatusWidowed                                           1824.2
## - A_stageRegional                                                 1824.6
## - RaceWhite                                                       1824.7
## <none>                                                            1825.0
## - Tumor_size                                                      1825.4
## - sixth_stageIIB                                                  1825.5
## - T_stageT3                                                       1827.8
## - Estrogen_statusPositive                                         1827.8
## - Marital_StatusSeparated                                         1827.9
## - `RaceOther (American Indian/AK Native, Asian/Pacific Islander)` 1828.0
## - Regional_nodes_examined                                         1829.8
## - `GradeUndifferentiated; anaplastic; Grade IV`                   1830.2
## - T_stageT4                                                       1830.4
## - N_stageN2                                                       1830.9
## - `GradeWell differentiated; Grade I`                             1831.9
## - `GradePoorly differentiated; Grade III`                         1832.0
## - Progesterone_statusPositive                                     1832.2
## - Age                                                             1837.3
## - N_stageN3                                                       1852.7
## - Survival_months                                                 2375.9
##
## Step:  AIC=1823.06
## Status ~ Age + `RaceOther (American Indian/AK Native, Asian/Pacific Islander)` +
##     RaceWhite + Marital_StatusSeparated + `Marital_StatusSingle (never married)` +
##     Marital_StatusWidowed + T_stageT2 + T_stageT3 + T_stageT4 +
##     N_stageN2 + N_stageN3 + sixth_stageIIB + sixth_stageIIIB +
##     `GradePoorly differentiated; Grade III` + `GradeUndifferentiated; anaplastic; Grade IV` +
##     `GradeWell differentiated; Grade I` + A_stageRegional + Tumor_size +
##     Estrogen_statusPositive + Progesterone_statusPositive + Regional_nodes_examined +
##     Regional_nodes_positive + Survival_months
##
##                                                                   Df Deviance
## - sixth_stageIIIB                                                  1   1775.4
## - Regional_nodes_positive                                          1   1775.4
## - T_stageT2                                                        1   1775.6
## - `Marital_StatusSingle (never married)`                           1   1776.2
## - Marital_StatusWidowed                                            1   1776.2
## - A_stageRegional                                                  1   1776.7
## - RaceWhite                                                        1   1776.7
## <none>                                                                 1775.1
## - Tumor_size                                                       1   1777.4
## - sixth_stageIIB                                                   1   1778.6
## - Estrogen_statusPositive                                          1   1779.8
## - Marital_StatusSeparated                                          1   1779.9
## - `RaceOther (American Indian/AK Native, Asian/Pacific Islander)`  1   1780.0
## - T_stageT3                                                        1   1780.9
## - Regional_nodes_examined                                          1   1781.8
## - `GradeUndifferentiated; anaplastic; Grade IV`                    1   1782.2
## - T_stageT4                                                        1   1782.5
## - `GradeWell differentiated; Grade I`                              1   1783.9
## - `GradePoorly differentiated; Grade III`                          1   1784.0
## - Progesterone_statusPositive                                      1   1784.2
## - N_stageN2                                                        1   1787.6
## - Age                                                              1   1789.4
## - N_stageN3                                                        1   1818.0
## - Survival_months                                                  1   2327.9
##                                                                      AIC
## - sixth_stageIIIB                                                 1821.4
## - Regional_nodes_positive                                         1821.4
## - T_stageT2                                                       1821.6
## - `Marital_StatusSingle (never married)`                          1822.2
## - Marital_StatusWidowed                                           1822.2
## - A_stageRegional                                                 1822.7
## - RaceWhite                                                       1822.7
## <none>                                                            1823.1
## - Tumor_size                                                      1823.4
## - sixth_stageIIB                                                  1824.6
## - Estrogen_statusPositive                                         1825.8
## - Marital_StatusSeparated                                         1825.9
## - `RaceOther (American Indian/AK Native, Asian/Pacific Islander)` 1826.0
## - T_stageT3                                                       1826.9
## - Regional_nodes_examined                                          1827.8
## - `GradeUndifferentiated; anaplastic; Grade IV`                    1828.2
## - T_stageT4                                                        1828.5
## - `GradeWell differentiated; Grade I`                              1829.9
## - `GradePoorly differentiated; Grade III`                          1830.0
## - Progesterone_statusPositive                                      1830.2
## - N_stageN2                                                        1833.6
## - Age                                                              1835.4
## - N_stageN3                                                        1864.0
## - Survival_months                                                  2373.9
##
## Step:  AIC=1821.38
## Status ~ Age + `RaceOther (American Indian/AK Native, Asian/Pacific Islander)` +
##     RaceWhite + Marital_StatusSeparated + `Marital_StatusSingle (never married)` +
##     Marital_StatusWidowed + T_stageT2 + T_stageT3 + T_stageT4 +
##     N_stageN2 + N_stageN3 + sixth_stageIIB + `GradePoorly differentiated; Grade III` +
##     `GradeUndifferentiated; anaplastic; Grade IV` + `GradeWell differentiated; Grade I` +
##     A_stageRegional + Tumor_size + Estrogen_statusPositive +
##     Progesterone_statusPositive + Regional_nodes_examined + Regional_nodes_positive +
##     Survival_months
##
##                                                                   Df Deviance
## - Regional_nodes_positive                                          1   1775.7
## - T_stageT2                                                        1   1775.8
## - `Marital_StatusSingle (never married)`                           1   1776.6
## - Marital_StatusWidowed                                            1   1776.6
## - RaceWhite                                                        1   1777.0
## - A_stageRegional                                                  1   1777.0
## <none>                                                                 1775.4
## - Tumor_size                                                       1   1777.6
## - sixth_stageIIB                                                   1   1779.2
## - Estrogen_statusPositive                                          1   1780.0
## - Marital_StatusSeparated                                          1   1780.3
## - `RaceOther (American Indian/AK Native, Asian/Pacific Islander)`  1   1780.3
## - T_stageT3                                                        1   1781.1
## - Regional_nodes_examined                                          1   1782.1
## - `GradeUndifferentiated; anaplastic; Grade IV`                    1   1782.5
## - `GradeWell differentiated; Grade I`                              1   1784.3
## - `GradePoorly differentiated; Grade III`                          1   1784.3
## - Progesterone_statusPositive                                      1   1784.7
## - T_stageT4                                                        1   1787.6
## - N_stageN2                                                        1   1788.5
## - Age                                                              1   1789.7
## - N_stageN3                                                        1   1823.2
## - Survival_months                                                  1   2328.0
##                                                                      AIC
## - Regional_nodes_positive                                         1819.7
## - T_stageT2                                                       1819.8
## - `Marital_StatusSingle (never married)`                          1820.6
## - Marital_StatusWidowed                                           1820.6
## - RaceWhite                                                       1821.0
## - A_stageRegional                                                 1821.0
## <none>                                                            1821.4
## - Tumor_size                                                      1821.6
## - sixth_stageIIB                                                  1823.2
## - Estrogen_statusPositive                                         1824.0
## - Marital_StatusSeparated                                         1824.3
## - ‘RaceOther (American Indian/AK Native, Asian/Pacific Islander)‘ 1824.3
## - T_stageT3                                                       1825.1
## - Regional_nodes_examined                                         1826.1
## - ‘GradeUndifferentiated; anaplastic; Grade IV‘                   1826.5
## - ‘GradeWell differentiated; Grade I‘                             1828.3
## - ‘GradePoorly differentiated; Grade III‘                         1828.3
## - Progesterone_statusPositive                                     1828.7
## - T_stageT4                                                       1831.6
## - N_stageN2                                                       1832.5
## - Age                                                             1833.7
## - N_stageN3                                                       1867.2
## - Survival_months                                                 2372.0
##
## Step:  AIC=1819.65
## Status ~ Age + ‘RaceOther (American Indian/AK Native, Asian/Pacific Islander)‘ +
##     RaceWhite + Marital_StatusSeparated + ‘Marital_StatusSingle (never married)‘ +
##     Marital_StatusWidowed + T_stageT2 + T_stageT3 + T_stageT4 +
##     N_stageN2 + N_stageN3 + sixth_stageIIB + ‘GradePoorly differentiated; Grade III‘ +
##     ‘GradeUndifferentiated; anaplastic; Grade IV‘ + ‘GradeWell differentiated; Grade I‘ +
##     A_stageRegional + Tumor_size + Estrogen_statusPositive +
##     Progesterone_statusPositive + Regional_nodes_examined + Survival_months
##
##                                                                   Df Deviance
## - T_stageT2                                                        1   1776.1
## - Marital_StatusWidowed                                            1   1776.8
## - ‘Marital_StatusSingle (never married)‘                           1   1776.8
## - RaceWhite                                                        1   1777.2
## - A_stageRegional                                                  1   1777.3
## <none>                                                                 1775.7
## - Tumor_size                                                       1   1777.9
## - sixth_stageIIB                                                   1   1779.5
## - Estrogen_statusPositive                                          1   1780.2
## - Marital_StatusSeparated                                          1   1780.5
## - ‘RaceOther (American Indian/AK Native, Asian/Pacific Islander)‘  1   1780.6
## - T_stageT3                                                        1   1781.5
## - Regional_nodes_examined                                          1   1782.2
## - ‘GradeUndifferentiated; anaplastic; Grade IV‘                    1   1782.8
## - ‘GradePoorly differentiated; Grade III‘                          1   1784.5
## - ‘GradeWell differentiated; Grade I‘                              1   1784.5
## - Progesterone_statusPositive                                      1   1785.2
## - T_stageT4                                                        1   1787.9
## - Age                                                              1   1789.9
## - N_stageN2                                                        1   1800.8
## - N_stageN3                                                        1   1841.5
## - Survival_months                                                  1   2328.3
##                                                                      AIC
## - T_stageT2                                                       1818.1
## - Marital_StatusWidowed                                           1818.8
## - ‘Marital_StatusSingle (never married)‘                          1818.8
## - RaceWhite                                                       1819.2
## - A_stageRegional                                                 1819.3
## <none>                                                            1819.7
## - Tumor_size                                                      1819.9
## - sixth_stageIIB                                                  1821.5
## - Estrogen_statusPositive                                         1822.2
## - Marital_StatusSeparated                                         1822.5
## - ‘RaceOther (American Indian/AK Native, Asian/Pacific Islander)‘ 1822.6
## - T_stageT3                                                       1823.5
## - Regional_nodes_examined                                         1824.2
## - ‘GradeUndifferentiated; anaplastic; Grade IV‘                   1824.8
## - ‘GradePoorly differentiated; Grade III‘                         1826.5
## - ‘GradeWell differentiated; Grade I‘                             1826.5
## - Progesterone_statusPositive                                     1827.2
## - T_stageT4                                                       1829.9
## - Age                                                             1831.9
## - N_stageN2                                                       1842.8
## - N_stageN3                                                       1883.5
## - Survival_months                                                 2370.3
## 
## Step:  AIC=1818.11
## Status ~ Age + ‘RaceOther (American Indian/AK Native, Asian/Pacific Islander)‘ + 
##     RaceWhite + Marital_StatusSeparated + ‘Marital_StatusSingle (never married)‘ + 
##     Marital_StatusWidowed + T_stageT3 + T_stageT4 + N_stageN2 + 
##     N_stageN3 + sixth_stageIIB + ‘GradePoorly differentiated; Grade III‘ + 
##     ‘GradeUndifferentiated; anaplastic; Grade IV‘ + ‘GradeWell differentiated; Grade I‘ + 
##     A_stageRegional + Tumor_size + Estrogen_statusPositive + 
##     Progesterone_statusPositive + Regional_nodes_examined + Survival_months
## 
##                                                                     Df Deviance
## - ‘Marital_StatusSingle (never married)‘                             1   1777.3
## - Marital_StatusWidowed                                              1   1777.3
## - RaceWhite                                                          1   1777.7
## - A_stageRegional                                                    1   1777.8
## - Tumor_size                                                         1   1777.9
## <none>                                                                   1776.1
## - Estrogen_statusPositive                                            1   1780.7
## - ‘RaceOther (American Indian/AK Native, Asian/Pacific Islander)‘  1   1781.0
## - Marital_StatusSeparated                                            1   1781.0
## - Regional_nodes_examined                                            1   1782.6
## - T_stageT3                                                          1   1783.2
## - ‘GradeUndifferentiated; anaplastic; Grade IV‘                      1   1783.3
## - ‘GradeWell differentiated; Grade I‘                                1   1784.9
## - ‘GradePoorly differentiated; Grade III‘                            1   1784.9
## - Progesterone_statusPositive                                        1   1785.6
## - sixth_stageIIB                                                     1   1785.9
## - T_stageT4                                                          1   1788.4
## - Age                                                                1   1790.5
## - N_stageN2                                                          1   1812.8
## - N_stageN3                                                          1   1857.8
## - Survival_months                                                    1   2328.9
##                                                                        AIC
## - ‘Marital_StatusSingle (never married)‘                            1817.3
## - Marital_StatusWidowed                                             1817.3
## - RaceWhite                                                         1817.7
## - A_stageRegional                                                   1817.8
## - Tumor_size                                                      1817.9
## <none>                                                            1818.1
## - Estrogen_statusPositive                                         1820.7
## - ‘RaceOther (American Indian/AK Native, Asian/Pacific Islander)‘ 1821.0
## - Marital_StatusSeparated                                         1821.0
## - Regional_nodes_examined                                         1822.6
## - T_stageT3                                                       1823.2
## - ‘GradeUndifferentiated; anaplastic; Grade IV‘                   1823.3
## - ‘GradeWell differentiated; Grade I‘                             1824.9
## - ‘GradePoorly differentiated; Grade III‘                         1824.9
## - Progesterone_statusPositive                                     1825.6
## - sixth_stageIIB                                                  1825.9
## - T_stageT4                                                       1828.4
## - Age                                                             1830.5
## - N_stageN2                                                       1852.8
## - N_stageN3                                                       1897.8
## - Survival_months                                                 2368.9
##
## Step:  AIC=1817.27
## Status ~ Age + ‘RaceOther (American Indian/AK Native, Asian/Pacific Islander)‘ +
##     RaceWhite + Marital_StatusSeparated + Marital_StatusWidowed +
##     T_stageT3 + T_stageT4 + N_stageN2 + N_stageN3 + sixth_stageIIB +
##     ‘GradePoorly differentiated; Grade III‘ + ‘GradeUndifferentiated; anaplastic; Grade IV‘ +
##     ‘GradeWell differentiated; Grade I‘ + A_stageRegional + Tumor_size +
##     Estrogen_statusPositive + Progesterone_statusPositive + Regional_nodes_examined +
##     Survival_months
##
##                                                                   Df Deviance
## - Marital_StatusWidowed                                            1   1778.2
## - A_stageRegional                                                  1   1778.9
## - Tumor_size                                                       1   1778.9
## <none>                                                                 1777.3
## - RaceWhite                                                        1   1779.6
## - Estrogen_statusPositive                                          1   1781.8
## - Marital_StatusSeparated                                          1   1781.9
## - ‘RaceOther (American Indian/AK Native, Asian/Pacific Islander)‘  1   1783.4
## - Regional_nodes_examined                                          1   1783.8
## - T_stageT3                                                        1   1784.3
## - ‘GradeUndifferentiated; anaplastic; Grade IV‘                    1   1784.6
## - ‘GradeWell differentiated; Grade I‘                              1   1786.0
## - ‘GradePoorly differentiated; Grade III‘                          1   1786.3
## - Progesterone_statusPositive                                      1   1786.7
## - sixth_stageIIB                                                   1   1787.1
## - T_stageT4                                                        1   1789.4
## - Age                                                              1   1790.9
## - N_stageN2                                                        1   1814.4
## - N_stageN3                                                        1   1858.5
## - Survival_months                                                  1   2331.3
##                                                                      AIC
## - Marital_StatusWidowed                                           1816.2
## - A_stageRegional                                                 1816.9
## - Tumor_size                                                      1816.9
## <none>                                                            1817.3
## - RaceWhite                                                       1817.6
## - Estrogen_statusPositive                                         1819.8
## - Marital_StatusSeparated                                         1819.9
## - ‘RaceOther (American Indian/AK Native, Asian/Pacific Islander)‘ 1821.4
## - Regional_nodes_examined                                         1821.8
## - T_stageT3                                                       1822.3
## - ‘GradeUndifferentiated; anaplastic; Grade IV‘                   1822.6
## - ‘GradeWell differentiated; Grade I‘                             1824.0
## - ‘GradePoorly differentiated; Grade III‘                         1824.3
## - Progesterone_statusPositive                                     1824.7
## - sixth_stageIIB                                                  1825.1
## - T_stageT4                                                       1827.4
## - Age                                                             1828.9
## - N_stageN2                                                       1852.4
## - N_stageN3                                                       1896.5
## - Survival_months                                                 2369.3
##
## Step:  AIC=1816.25
## Status ~ Age + ‘RaceOther (American Indian/AK Native, Asian/Pacific Islander)‘ +
##     RaceWhite + Marital_StatusSeparated + T_stageT3 + T_stageT4 +
##     N_stageN2 + N_stageN3 + sixth_stageIIB + ‘GradePoorly differentiated; Grade III‘ +
##     ‘GradeUndifferentiated; anaplastic; Grade IV‘ + ‘GradeWell differentiated; Grade I‘ +
##     A_stageRegional + Tumor_size + Estrogen_statusPositive +
##     Progesterone_statusPositive + Regional_nodes_examined + Survival_months
##
##                                                                   Df Deviance
## - Tumor_size                                                       1   1779.8
## - A_stageRegional                                                  1   1779.9
## <none>                                                                 1778.2
## - RaceWhite                                                        1   1780.8
## - Estrogen_statusPositive                                          1   1782.7
## - Marital_StatusSeparated                                          1   1782.7
## - ‘RaceOther (American Indian/AK Native, Asian/Pacific Islander)‘  1   1784.6
## - Regional_nodes_examined                                          1   1784.7
## - T_stageT3                                                        1   1785.4
## - ‘GradeUndifferentiated; anaplastic; Grade IV‘                    1   1785.5
## - ‘GradeWell differentiated; Grade I‘                              1   1786.8
## - ‘GradePoorly differentiated; Grade III‘                          1   1787.4
## - Progesterone_statusPositive                                      1   1787.8
## - sixth_stageIIB                                                   1   1788.2
## - T_stageT4                                                        1   1790.3
## - Age                                                              1   1794.3
## - N_stageN2                                                        1   1815.3
## - N_stageN3                                                        1   1859.7
## - Survival_months                                                  1   2332.5
##                                                                      AIC
## - Tumor_size                                                      1815.8
## - A_stageRegional                                                 1815.9
## <none>                                                            1816.2
## - RaceWhite                                                       1816.8
## - Estrogen_statusPositive                                         1818.7
## - Marital_StatusSeparated                                         1818.7
## - ‘RaceOther (American Indian/AK Native, Asian/Pacific Islander)‘ 1820.6
## - Regional_nodes_examined                                         1820.7
## - T_stageT3                                                       1821.4
## - ‘GradeUndifferentiated; anaplastic; Grade IV‘                   1821.5
## - ‘GradeWell differentiated; Grade I‘                             1822.8
## - ‘GradePoorly differentiated; Grade III‘                         1823.4
## - Progesterone_statusPositive                                     1823.8
## - sixth_stageIIB                                                  1824.2
## - T_stageT4                                                       1826.3
## - Age                                                             1830.3
## - N_stageN2                                                       1851.3
## - N_stageN3                                                       1895.7
## - Survival_months                                                 2368.5
##
## Step:  AIC=1815.83
## Status ~ Age + ‘RaceOther (American Indian/AK Native, Asian/Pacific Islander)‘ +
##     RaceWhite + Marital_StatusSeparated + T_stageT3 + T_stageT4 +
##     N_stageN2 + N_stageN3 + sixth_stageIIB + ‘GradePoorly differentiated; Grade III‘ +
##     ‘GradeUndifferentiated; anaplastic; Grade IV‘ + ‘GradeWell differentiated; Grade I‘ +
##     A_stageRegional + Estrogen_statusPositive + Progesterone_statusPositive +
##     Regional_nodes_examined + Survival_months
##
##                                                                   Df Deviance
## - A_stageRegional                                                  1   1781.6
## <none>                                                                 1779.8
## - RaceWhite                                                        1   1782.5
## - Marital_StatusSeparated                                          1   1784.2
## - Estrogen_statusPositive                                          1   1784.6
## - T_stageT3                                                        1   1785.5
## - Regional_nodes_examined                                          1   1786.2
## - ‘RaceOther (American Indian/AK Native, Asian/Pacific Islander)‘  1   1786.4
## - ‘GradeUndifferentiated; anaplastic; Grade IV‘                    1   1787.1
## - ‘GradeWell differentiated; Grade I‘                              1   1788.0
## - sixth_stageIIB                                                   1   1788.3
## - ‘GradePoorly differentiated; Grade III‘                          1   1788.6
## - Progesterone_statusPositive                                      1   1789.1
## - T_stageT4                                                        1   1790.9
## - Age                                                              1   1795.9
## - N_stageN2                                                        1   1815.8
## - N_stageN3                                                        1   1860.9
## - Survival_months                                                  1   2332.5
##                                                                      AIC
## - A_stageRegional                                                 1815.6
## <none>                                                            1815.8
## - RaceWhite                                                       1816.5
## - Marital_StatusSeparated                                         1818.2
## - Estrogen_statusPositive                                         1818.6
## - T_stageT3                                                       1819.5
## - Regional_nodes_examined                                         1820.2
## - ‘RaceOther (American Indian/AK Native, Asian/Pacific Islander)‘ 1820.4
## - ‘GradeUndifferentiated; anaplastic; Grade IV‘                   1821.1
## - ‘GradeWell differentiated; Grade I‘                             1822.0
## - sixth_stageIIB                                                  1822.3
## - ‘GradePoorly differentiated; Grade III‘                         1822.6
## - Progesterone_statusPositive                                     1823.1
## - T_stageT4                                                       1824.9
## - Age                                                             1829.9
## - N_stageN2                                                       1849.8
## - N_stageN3                                                       1894.9
## - Survival_months                                                 2366.5
##
## Step:  AIC=1815.57
## Status ~ Age + ‘RaceOther (American Indian/AK Native, Asian/Pacific Islander)‘ +
##     RaceWhite + Marital_StatusSeparated + T_stageT3 + T_stageT4 +
##     N_stageN2 + N_stageN3 + sixth_stageIIB + ‘GradePoorly differentiated; Grade III‘ +
##     ‘GradeUndifferentiated; anaplastic; Grade IV‘ + ‘GradeWell differentiated; Grade I‘ +
##     Estrogen_statusPositive + Progesterone_statusPositive + Regional_nodes_examined +
##     Survival_months
##
##                                                                   Df Deviance
## <none>                                                                 1781.6
## - RaceWhite                                                         1   1784.2
## - Marital_StatusSeparated                                           1   1785.8
## - Estrogen_statusPositive                                           1   1786.2
## - T_stageT3                                                         1   1787.2
## - Regional_nodes_examined                                           1   1787.9
## - ‘RaceOther (American Indian/AK Native, Asian/Pacific Islander)‘  1   1788.3
## - ‘GradeUndifferentiated; anaplastic; Grade IV‘                     1   1788.9
## - ‘GradeWell differentiated; Grade I‘                               1   1790.0
## - sixth_stageIIB                                                    1   1790.0
## - ‘GradePoorly differentiated; Grade III‘                           1   1790.8
## - Progesterone_statusPositive                                       1   1791.0
## - T_stageT4                                                         1   1791.1
## - Age                                                               1   1798.3
## - N_stageN2                                                         1   1817.4
## - N_stageN3                                                         1   1861.3
## - Survival_months                                                   1   2332.8
##                                                                         AIC
## <none>                                                            1815.6
## - RaceWhite                                                       1816.2
## - Marital_StatusSeparated                                         1817.8
## - Estrogen_statusPositive                                         1818.2
## - T_stageT3                                                       1819.2
## - Regional_nodes_examined                                         1819.9
## - ‘RaceOther (American Indian/AK Native, Asian/Pacific Islander)‘ 1820.3
## - ‘GradeUndifferentiated; anaplastic; Grade IV‘                   1820.9
## - ‘GradeWell differentiated; Grade I‘                             1822.0
## - sixth_stageIIB                                                  1822.0
## - ‘GradePoorly differentiated; Grade III‘                         1822.8
## - Progesterone_statusPositive                                     1823.0
## - T_stageT4                                                       1823.1
## - Age                                                             1830.3
## - N_stageN2                                                       1849.4
## - N_stageN3                                                       1893.3
## - Survival_months                                                 2364.8
```

```
summary(Backward_log_imb)
```

```
##
## Call:
## glm(formula = Status ~ Age + ‘RaceOther (American Indian/AK Native, Asian/Pacific Islander)‘ +
##     RaceWhite + Marital_StatusSeparated + T_stageT3 + T_stageT4 +
##     N_stageN2 + N_stageN3 + sixth_stageIIB + ‘GradePoorly differentiated; Grade III‘ +
##     ‘GradeUndifferentiated; anaplastic; Grade IV‘ + ‘GradeWell differentiated; Grade I‘ +
##     Estrogen_statusPositive + Progesterone_statusPositive + Regional_nodes_examined +
##     Survival_months, family = binomial(link = "logit"), data = train_breast_cancer)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1963  -0.4647  -0.2683  -0.1358   3.2701
##
## Coefficients:
##                                                                   Estimate
## (Intercept)                                                       0.928275
## Age                                                               0.028155
## ‘RaceOther (American Indian/AK Native, Asian/Pacific Islander)‘ -0.828080
## RaceWhite                                                        -0.347443
## Marital_StatusSeparated                                           1.083461
## T_stageT3                                                         0.412256
## T_stageT4                                                         1.049530
## N_stageN2                                                         1.012029
## N_stageN3                                                         1.748791
## sixth_stageIIB                                                    0.507987
## ‘GradePoorly differentiated; Grade III‘                           0.415482
## ‘GradeUndifferentiated; anaplastic; Grade IV‘                     2.351711
## ‘GradeWell differentiated; Grade I‘                              -0.653054
## Estrogen_statusPositive                                          -0.546921
## Progesterone_statusPositive                                      -0.522161
## Regional_nodes_examined                                          -0.769301
## Survival_months                                                  -6.261683
##                                                                  Std. Error
## (Intercept)                                                        0.514609
## Age                                                                0.006954
## ‘RaceOther (American Indian/AK Native, Asian/Pacific Islander)‘   0.324329
## RaceWhite                                                          0.210194
## Marital_StatusSeparated                                            0.506564
## T_stageT3                                                          0.170862
## T_stageT4                                                          0.327009
## N_stageN2                                                          0.169818
## N_stageN3                                                          0.197894
## sixth_stageIIB                                                     0.175072
## ‘GradePoorly differentiated; Grade III‘                            0.136101
## ‘GradeUndifferentiated; anaplastic; Grade IV‘                      0.803893
## ‘GradeWell differentiated; Grade I‘                                0.236892
## Estrogen_statusPositive                                            0.252820
## Progesterone_statusPositive                                        0.167215
## Regional_nodes_examined                                            0.308087
## Survival_months                                                    0.318587
##                                                                  z value
## (Intercept)                                                        1.804
## Age                                                                4.049
## ‘RaceOther (American Indian/AK Native, Asian/Pacific Islander)‘  -2.553
## RaceWhite                                                         -1.653
## Marital_StatusSeparated                                            2.139
## T_stageT3                                                          2.413
## T_stageT4                                                         3.209
## N_stageN2                                                         5.959
## N_stageN3                                                         8.837
## sixth_stageIIB                                                    2.902
## ‘GradePoorly differentiated; Grade III‘                           3.053
## ‘GradeUndifferentiated; anaplastic; Grade IV‘                     2.925
## ‘GradeWell differentiated; Grade I‘                              -2.757
## Estrogen_statusPositive                                          -2.163
## Progesterone_statusPositive                                      -3.123
## Regional_nodes_examined                                          -2.497
## Survival_months                                                 -19.655
##                                                                 Pr(>|z|)
## (Intercept)                                                      0.07126 .
## Age                                                             5.15e-05 ***
## ‘RaceOther (American Indian/AK Native, Asian/Pacific Islander)‘  0.01067 *
## RaceWhite                                                        0.09834 .
## Marital_StatusSeparated                                          0.03245 *
## T_stageT3                                                        0.01583 *
## T_stageT4                                                        0.00133 **
## N_stageN2                                                       2.53e-09 ***
## N_stageN3                                                        < 2e-16 ***
## sixth_stageIIB                                                   0.00371 **
## ‘GradePoorly differentiated; Grade III‘                          0.00227 **
## ‘GradeUndifferentiated; anaplastic; Grade IV‘                    0.00344 **
## ‘GradeWell differentiated; Grade I‘                              0.00584 **
## Estrogen_statusPositive                                          0.03052 *
## Progesterone_statusPositive                                      0.00179 **
## Regional_nodes_examined                                          0.01252 *
## Survival_months                                                  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2717.7  on 3205  degrees of freedom
## Residual deviance: 1781.6  on 3189  degrees of freedom
## AIC: 1815.6
##
## Number of Fisher Scoring iterations: 6
```

```
fitted.results1 <- predict(Backward_log_imb,newdata=test_breast_cancer,type='response')
fitted.results1 <- ifelse(fitted.results1 > 0.5,1,0)
tab_fitted1 <- table(fitted.results1, test_breast_cancer$Status)

misClasificError <- mean(fitted.results != test_breast_cancer$Status)
print(paste('Accuracy',1-misClasificError))
```

```
## [1] "Accuracy 0.9"
```

```
misClasificError1 <- mean(fitted.results1 != test_breast_cancer$Status)
print(paste('Accuracy',1-misClasificError1))
```

```
## [1] "Accuracy 0.9"
```

```
confusionMatrix(table(fitted.results, test_breast_cancer$Status))
```

```
## Confusion Matrix and Statistics
##
##
## fitted.results   0   1
##              0 668  68
##              1  12  52
##
##                Accuracy : 0.9
##                  95% CI : (0.8771, 0.9199)
##     No Information Rate : 0.85
##     P-Value [Acc > NIR] : 2.001e-05
##
##                   Kappa : 0.5146
##
##  Mcnemar's Test P-Value : 7.788e-10
##
##             Sensitivity : 0.9824
##             Specificity : 0.4333
##          Pos Pred Value : 0.9076
##          Neg Pred Value : 0.8125
##              Prevalence : 0.8500
##          Detection Rate : 0.8350
##    Detection Prevalence : 0.9200
##       Balanced Accuracy : 0.7078
##
##        'Positive' Class : 0
##
```

```
log_cm_imb <- confusionMatrix(table(fitted.results1, test_breast_cancer$Status))


#Results
#Before backward elimination
print(paste("For Imbalanced data:", "Precision is:",caret::precision(tab_fitted),
"Recall is:",sensitivity(tab_fitted),"F-score is:",caret::F_meas(tab_fitted)))
```

```
## [1] "For Imbalanced data: Precision is: 0.907608695652174 Recall is: 0.982352941176471 F-score is: 0
```

```
#After Backward elimination
print(paste("For Imbalanced data", "after backward elimination:", "Precision is:",caret::precision(tab_
"Recall is:",sensitivity(tab_fitted1),"F-score is:",caret::F_meas(tab_fitted1)))
```

```
## [1] "For Imbalanced data after backward elimination: Precision is: 0.90650406504065 Recall is: 0.983
```

```
#We see that, a little improvement in the values after backward elimination
#step() could not remove all the non-signficant variables in the model. we
#can manually drop the non-significant variables having p-value of above 0.05
```

##Model3(Logistic Regression)#Balanced data(model training)

```
#Performing logistic regression model on Balanced data
log_bal <- glm(formula = Status ~ ., family = binomial(link = "logit"),
    data = train_breast_cancer_smote)


fitted.results_bal <- predict(log_bal,newdata=test_breast_cancer,type='response')
```

```
## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (type == :
## prediction from a rank-deficient fit may be misleading
```

```
fitted.results_bal <- ifelse(fitted.results_bal > 0.5,1,0)
tab_fitted_bal <- table(fitted.results, test_breast_cancer$Status)

misClasificError <- mean(fitted.results_bal != test_breast_cancer$Status)
print(paste('Accuracy',1-misClasificError))
```

```
## [1] "Accuracy 0.82875"
```

```
#Backward Elimination
Backward_log_bal <- step(log_bal, direction = "backward", trace = TRUE)
```

```
## Start:  AIC=2975.32
## Status ~ Age + ‘RaceOther (American Indian/AK Native, Asian/Pacific Islander)‘ +
##     RaceWhite + ‘Marital_StatusMarried (including common law)‘ +
##     Marital_StatusSeparated + ‘Marital_StatusSingle (never married)‘ +
##     Marital_StatusWidowed + T_stageT2 + T_stageT3 + T_stageT4 +
##     N_stageN2 + N_stageN3 + sixth_stageIIB + sixth_stageIIIA +
##     sixth_stageIIIB + sixth_stageIIIC + ‘GradePoorly differentiated; Grade III‘ +
##     ‘GradeUndifferentiated; anaplastic; Grade IV‘ + ‘GradeWell differentiated; Grade I‘ +
##     A_stageRegional + Tumor_size + Estrogen_statusPositive +
##     Progesterone_statusPositive + Regional_nodes_examined + Regional_nodes_positive +
##     Survival_months
##
##
## Step:  AIC=2975.32
## Status ~ Age + ‘RaceOther (American Indian/AK Native, Asian/Pacific Islander)‘ +
##     RaceWhite + ‘Marital_StatusMarried (including common law)‘ +
##     Marital_StatusSeparated + ‘Marital_StatusSingle (never married)‘ +
##     Marital_StatusWidowed + T_stageT2 + T_stageT3 + T_stageT4 +
##     N_stageN2 + N_stageN3 + sixth_stageIIB + sixth_stageIIIA +
##     sixth_stageIIIB + ‘GradePoorly differentiated; Grade III‘ +
##     ‘GradeUndifferentiated; anaplastic; Grade IV‘ + ‘GradeWell differentiated; Grade I‘ +
##     A_stageRegional + Tumor_size + Estrogen_statusPositive +
##     Progesterone_statusPositive + Regional_nodes_examined + Regional_nodes_positive +
##     Survival_months
##
##                                                                   Df Deviance
## - T_stageT2                                                        1   2923.3
## - sixth_stageIIIB                                                  1   2923.3
## - ‘Marital_StatusSingle (never married)‘                           1   2923.4
## - Marital_StatusWidowed                                            1   2923.4
## - ‘Marital_StatusMarried (including common law)‘                   1   2924.4
## - sixth_stageIIIA                                                  1   2924.6
## - A_stageRegional                                                  1   2924.8
## - Marital_StatusSeparated                                          1   2925.0
## - Regional_nodes_positive                                          1   2925.1
## - RaceWhite                                                        1   2925.2
## <none>                                                                 2923.3
## - T_stageT3                                                        1   2927.9
## - Tumor_size                                                       1   2928.9
## - Progesterone_statusPositive                                      1   2929.3
## - T_stageT4                                                        1   2929.3
## - Estrogen_statusPositive                                          1   2929.9
## - ‘GradeUndifferentiated; anaplastic; Grade IV‘                    1   2930.0
## - N_stageN2                                                        1   2934.9
## - ‘GradePoorly differentiated; Grade III‘                          1   2937.4
## - ‘RaceOther (American Indian/AK Native, Asian/Pacific Islander)‘  1   2939.4
## - Regional_nodes_examined                                          1   2942.4
## - sixth_stageIIB                                                   1   2945.0
## - Age                                                              1   2950.8
## - ‘GradeWell differentiated; Grade I‘                              1   2962.4
## - N_stageN3                                                        1   2984.3
## - Survival_months                                                  1   3941.0
##                                                                      AIC
## - T_stageT2                                                       2973.3
## - sixth_stageIIIB                                                 2973.3
## - ‘Marital_StatusSingle (never married)‘                          2973.4
## - Marital_StatusWidowed                                           2973.4
## - ‘Marital_StatusMarried (including common law)‘                  2974.4
## - sixth_stageIIIA                                                 2974.6
## - A_stageRegional                                                 2974.8
## - Marital_StatusSeparated                                         2975.0
## - Regional_nodes_positive                                         2975.1
## - RaceWhite                                                       2975.2
## <none>                                                            2975.3
## - T_stageT3                                                       2977.9
## - Tumor_size                                                      2978.9
## - Progesterone_statusPositive                                     2979.3
## - T_stageT4                                                       2979.3
## - Estrogen_statusPositive                                         2979.9
## - ‘GradeUndifferentiated; anaplastic; Grade IV‘                   2980.0
## - N_stageN2                                                       2984.9
## - ‘GradePoorly differentiated; Grade III‘                         2987.4
## - ‘RaceOther (American Indian/AK Native, Asian/Pacific Islander)‘ 2989.4
## - Regional_nodes_examined                                         2992.4
## - sixth_stageIIB                                                  2995.0
## - Age                                                             3000.8
## - ‘GradeWell differentiated; Grade I‘                             3012.4
## - N_stageN3                                                       3034.3
## - Survival_months                                                 3991.0
##
## Step:  AIC=2973.32
## Status ~ Age + ‘RaceOther (American Indian/AK Native, Asian/Pacific Islander)‘ +
##     RaceWhite + ‘Marital_StatusMarried (including common law)‘ +
##     Marital_StatusSeparated + ‘Marital_StatusSingle (never married)‘ +
##     Marital_StatusWidowed + T_stageT3 + T_stageT4 + N_stageN2 +
##     N_stageN3 + sixth_stageIIB + sixth_stageIIIA + sixth_stageIIIB +
##     `GradePoorly differentiated; Grade III` + `GradeUndifferentiated; anaplastic; Grade IV` +
##     `GradeWell differentiated; Grade I` + A_stageRegional + Tumor_size +
##     Estrogen_statusPositive + Progesterone_statusPositive + Regional_nodes_examined +
##     Regional_nodes_positive + Survival_months
##
##                                                                    Df Deviance
## - sixth_stageIIIB                                                   1   2923.3
## - Marital_StatusWidowed                                             1   2923.4
## - `Marital_StatusSingle (never married)`                            1   2923.4
## - `Marital_StatusMarried (including common law)`                    1   2924.4
## - sixth_stageIIIA                                                   1   2924.8
## - A_stageRegional                                                   1   2924.8
## - Marital_StatusSeparated                                           1   2925.0
## - Regional_nodes_positive                                           1   2925.1
## - RaceWhite                                                         1   2925.2
## <none>                                                                  2923.3
## - Progesterone_statusPositive                                       1   2929.3
## - Estrogen_statusPositive                                           1   2929.9
## - `GradeUndifferentiated; anaplastic; Grade IV`                     1   2930.0
## - T_stageT4                                                         1   2930.1
## - Tumor_size                                                        1   2931.2
## - T_stageT3                                                         1   2933.5
## - N_stageN2                                                         1   2934.9
## - `GradePoorly differentiated; Grade III`                           1   2937.5
## - `RaceOther (American Indian/AK Native, Asian/Pacific Islander)`  1   2939.4
## - Regional_nodes_examined                                           1   2942.4
## - Age                                                               1   2950.9
## - `GradeWell differentiated; Grade I`                               1   2962.5
## - sixth_stageIIB                                                    1   2965.6
## - N_stageN3                                                         1   2999.5
## - Survival_months                                                   1   3941.2
##                                                                       AIC
## - sixth_stageIIIB                                                 2971.3
## - Marital_StatusWidowed                                           2971.4
## - `Marital_StatusSingle (never married)`                          2971.4
## - `Marital_StatusMarried (including common law)`                  2972.4
## - sixth_stageIIIA                                                 2972.8
## - A_stageRegional                                                 2972.8
## - Marital_StatusSeparated                                         2973.0
## - Regional_nodes_positive                                         2973.1
## - RaceWhite                                                       2973.2
## <none>                                                            2973.3
## - Progesterone_statusPositive                                     2977.3
## - Estrogen_statusPositive                                         2977.9
## - `GradeUndifferentiated; anaplastic; Grade IV`                   2978.0
## - T_stageT4                                                       2978.1
## - Tumor_size                                                      2979.2
## - T_stageT3                                                       2981.5
## - N_stageN2                                                       2982.9
## - `GradePoorly differentiated; Grade III`                         2985.5
## - `RaceOther (American Indian/AK Native, Asian/Pacific Islander)` 2987.4
## - Regional_nodes_examined                                         2990.4
## - Age                                                             2998.9
## - ‘GradeWell differentiated; Grade I‘                             3010.5
## - sixth_stageIIB                                                  3013.6
## - N_stageN3                                                       3047.5
## - Survival_months                                                 3989.2
##
## Step:  AIC=2971.34
## Status ~ Age + ‘RaceOther (American Indian/AK Native, Asian/Pacific Islander)‘ +
##     RaceWhite + ‘Marital_StatusMarried (including common law)‘ +
##     Marital_StatusSeparated + ‘Marital_StatusSingle (never married)‘ +
##     Marital_StatusWidowed + T_stageT3 + T_stageT4 + N_stageN2 +
##     N_stageN3 + sixth_stageIIB + sixth_stageIIIA + ‘GradePoorly differentiated; Grade III‘ +
##     ‘GradeUndifferentiated; anaplastic; Grade IV‘ + ‘GradeWell differentiated; Grade I‘ +
##     A_stageRegional + Tumor_size + Estrogen_statusPositive +
##     Progesterone_statusPositive + Regional_nodes_examined + Regional_nodes_positive +
##     Survival_months
##
##                                                                   Df Deviance
## - ‘Marital_StatusSingle (never married)‘                           1   2923.4
## - Marital_StatusWidowed                                            1   2923.4
## - ‘Marital_StatusMarried (including common law)‘                   1   2924.4
## - A_stageRegional                                                  1   2924.8
## - sixth_stageIIIA                                                  1   2925.0
## - Marital_StatusSeparated                                          1   2925.0
## - Regional_nodes_positive                                          1   2925.1
## - RaceWhite                                                        1   2925.2
## <none>                                                                2923.3
## - Progesterone_statusPositive                                      1   2929.4
## - Estrogen_statusPositive                                          1   2929.9
## - ‘GradeUndifferentiated; anaplastic; Grade IV‘                    1   2930.0
## - Tumor_size                                                       1   2931.4
## - T_stageT3                                                        1   2933.5
## - N_stageN2                                                        1   2935.0
## - ‘GradePoorly differentiated; Grade III‘                          1   2937.5
## - T_stageT4                                                        1   2939.4
## - ‘RaceOther (American Indian/AK Native, Asian/Pacific Islander)‘  1   2939.5
## - Regional_nodes_examined                                          1   2942.4
## - Age                                                              1   2950.9
## - ‘GradeWell differentiated; Grade I‘                              1   2962.6
## - sixth_stageIIB                                                   1   2966.7
## - N_stageN3                                                        1   3007.0
## - Survival_months                                                  1   3941.5
##                                                                      AIC
## - ‘Marital_StatusSingle (never married)‘                          2969.4
## - Marital_StatusWidowed                                           2969.4
## - ‘Marital_StatusMarried (including common law)‘                  2970.4
## - A_stageRegional                                                 2970.8
## - sixth_stageIIIA                                                 2971.0
## - Marital_StatusSeparated                                         2971.0
## - Regional_nodes_positive                                         2971.1
## - RaceWhite                                                       2971.2
## <none>                                                            2971.3
## - Progesterone_statusPositive                                     2975.4
## - Estrogen_statusPositive                                         2975.9
## - ‘GradeUndifferentiated; anaplastic; Grade IV‘                   2976.0
## - Tumor_size                                                      2977.4
## - T_stageT3                                                       2979.5
## - N_stageN2                                                       2981.0
## - ‘GradePoorly differentiated; Grade III‘                         2983.5
## - T_stageT4                                                       2985.4
## - ‘RaceOther (American Indian/AK Native, Asian/Pacific Islander)‘ 2985.5
## - Regional_nodes_examined                                         2988.4
## - Age                                                             2996.9
## - ‘GradeWell differentiated; Grade I‘                             3008.6
## - sixth_stageIIB                                                  3012.7
## - N_stageN3                                                       3053.0
## - Survival_months                                                 3987.5
##
## Step:  AIC=2969.38
## Status ~ Age + ‘RaceOther (American Indian/AK Native, Asian/Pacific Islander)‘ +
##     RaceWhite + ‘Marital_StatusMarried (including common law)‘ +
##     Marital_StatusSeparated + Marital_StatusWidowed + T_stageT3 +
##     T_stageT4 + N_stageN2 + N_stageN3 + sixth_stageIIB + sixth_stageIIIA +
##     ‘GradePoorly differentiated; Grade III‘ + ‘GradeUndifferentiated; anaplastic; Grade IV‘ +
##     ‘GradeWell differentiated; Grade I‘ + A_stageRegional + Tumor_size +
##     Estrogen_statusPositive + Progesterone_statusPositive + Regional_nodes_examined +
##     Regional_nodes_positive + Survival_months
##
##                                                                   Df Deviance
## - Marital_StatusWidowed                                            1   2923.5
## - A_stageRegional                                                  1   2924.8
## - ‘Marital_StatusMarried (including common law)‘                   1   2924.9
## - sixth_stageIIIA                                                  1   2925.0
## - Regional_nodes_positive                                          1   2925.2
## - RaceWhite                                                        1   2925.2
## - Marital_StatusSeparated                                          1   2925.3
## <none>                                                                 2923.4
## - Progesterone_statusPositive                                      1   2929.4
## - Estrogen_statusPositive                                          1   2930.0
## - ‘GradeUndifferentiated; anaplastic; Grade IV‘                    1   2930.0
## - Tumor_size                                                       1   2931.4
## - T_stageT3                                                        1   2933.6
## - N_stageN2                                                        1   2935.0
## - ‘GradePoorly differentiated; Grade III‘                          1   2937.5
## - T_stageT4                                                        1   2939.4
## - ‘RaceOther (American Indian/AK Native, Asian/Pacific Islander)‘  1   2939.5
## - Regional_nodes_examined                                          1   2942.5
## - Age                                                              1   2951.4
## - ‘GradeWell differentiated; Grade I‘                              1   2962.8
## - sixth_stageIIB                                                   1   2966.7
## - N_stageN3                                                        1   3007.1
## - Survival_months                                                  1   3941.6
##                                                                      AIC
## - Marital_StatusWidowed                                           2967.5
## - A_stageRegional                                                 2968.8
## - ‘Marital_StatusMarried (including common law)‘                  2968.9
## - sixth_stageIIIA                                                 2969.0
## - Regional_nodes_positive                                         2969.2
## - RaceWhite                                                       2969.2
## - Marital_StatusSeparated                                         2969.3
## <none>                                                            2969.4
## - Progesterone_statusPositive                                     2973.4
## - Estrogen_statusPositive                                         2974.0
## - ‘GradeUndifferentiated; anaplastic; Grade IV‘                   2974.0
## - Tumor_size                                                      2975.4
## - T_stageT3                                                       2977.6
## - N_stageN2                                                       2979.0
## - ‘GradePoorly differentiated; Grade III‘                         2981.5
## - T_stageT4                                                       2983.4
## - ‘RaceOther (American Indian/AK Native, Asian/Pacific Islander)‘ 2983.5
## - Regional_nodes_examined                                         2986.5
## - Age                                                             2995.4
## - ‘GradeWell differentiated; Grade I‘                             3006.8
## - sixth_stageIIB                                                  3010.7
## - N_stageN3                                                       3051.1
## - Survival_months                                                 3985.6
##
## Step:  AIC=2967.46
## Status ~ Age + ‘RaceOther (American Indian/AK Native, Asian/Pacific Islander)‘ +
##     RaceWhite + ‘Marital_StatusMarried (including common law)‘ +
##     Marital_StatusSeparated + T_stageT3 + T_stageT4 + N_stageN2 +
##     N_stageN3 + sixth_stageIIB + sixth_stageIIIA + ‘GradePoorly differentiated; Grade III‘ +
##     ‘GradeUndifferentiated; anaplastic; Grade IV‘ + ‘GradeWell differentiated; Grade I‘ +
##     A_stageRegional + Tumor_size + Estrogen_statusPositive +
##     Progesterone_statusPositive + Regional_nodes_examined + Regional_nodes_positive +
##     Survival_months
##
##                                                                   Df Deviance
## - A_stageRegional                                                  1   2924.9
## - sixth_stageIIIA                                                  1   2925.1
## - Regional_nodes_positive                                          1   2925.3
## - RaceWhite                                                        1   2925.3
## - Marital_StatusSeparated                                          1   2925.3
## <none>                                                                 2923.5
## - ‘Marital_StatusMarried (including common law)‘                   1   2925.5
## - Progesterone_statusPositive                                      1   2929.5
## - Estrogen_statusPositive                                          1   2930.1
## - ‘GradeUndifferentiated; anaplastic; Grade IV‘                    1   2930.1
## - Tumor_size                                                       1   2931.4
## - T_stageT3                                                        1   2933.7
## - N_stageN2                                                        1   2935.2
## - ‘GradePoorly differentiated; Grade III‘                          1   2937.6
## - T_stageT4                                                        1   2939.5
## - ‘RaceOther (American Indian/AK Native, Asian/Pacific Islander)‘  1   2939.5
## - Regional_nodes_examined                                          1   2942.5
## - Age                                                              1   2953.3
## - ‘GradeWell differentiated; Grade I‘                              1   2962.9
## - sixth_stageIIB                                                   1   2966.7
## - N_stageN3                                                        1   3007.2
## - Survival_months                                                  1   3941.6
##                                                                      AIC
## - A_stageRegional                                                 2966.9
## - sixth_stageIIIA                                                 2967.1
## - Regional_nodes_positive                                         2967.3
## - RaceWhite                                                       2967.3
## - Marital_StatusSeparated                                         2967.3
## <none>                                                            2967.5
## - ‘Marital_StatusMarried (including common law)‘                  2967.5
## - Progesterone_statusPositive                                     2971.5
## - Estrogen_statusPositive                                         2972.1
## - ‘GradeUndifferentiated; anaplastic; Grade IV‘                   2972.1
## - Tumor_size                                                      2973.4
## - T_stageT3                                                       2975.7
## - N_stageN2                                                       2977.2
## - ‘GradePoorly differentiated; Grade III‘                         2979.6
## - T_stageT4                                                       2981.5
## - ‘RaceOther (American Indian/AK Native, Asian/Pacific Islander)‘ 2981.5
## - Regional_nodes_examined                                         2984.5
## - Age                                                             2995.3
## - ‘GradeWell differentiated; Grade I‘                             3004.9
## - sixth_stageIIB                                                  3008.7
## - N_stageN3                                                       3049.2
## - Survival_months                                                 3983.6
##
## Step:  AIC=2966.9
## Status ~ Age + ‘RaceOther (American Indian/AK Native, Asian/Pacific Islander)‘ +
##     RaceWhite + ‘Marital_StatusMarried (including common law)‘ +
##     Marital_StatusSeparated + T_stageT3 + T_stageT4 + N_stageN2 +
##     N_stageN3 + sixth_stageIIB + sixth_stageIIIA + ‘GradePoorly differentiated; Grade III‘ +
##     ‘GradeUndifferentiated; anaplastic; Grade IV‘ + ‘GradeWell differentiated; Grade I‘ +
##     Tumor_size + Estrogen_statusPositive + Progesterone_statusPositive +
##     Regional_nodes_examined + Regional_nodes_positive + Survival_months
##
##                                                                   Df Deviance
## - sixth_stageIIIA                                                  1   2926.5
## - Marital_StatusSeparated                                          1   2926.6
## - RaceWhite                                                        1   2926.7
## - Regional_nodes_positive                                          1   2926.8
## <none>                                                                 2924.9
## - ‘Marital_StatusMarried (including common law)‘                   1   2927.1
## - Progesterone_statusPositive                                      1   2931.1
## - Estrogen_statusPositive                                          1   2931.5
## - ‘GradeUndifferentiated; anaplastic; Grade IV‘                    1   2931.6
## - Tumor_size                                                       1   2933.3
## - T_stageT3                                                        1   2935.5
## - N_stageN2                                                        1   2936.6
## - T_stageT4                                                        1   2939.5
## - ‘GradePoorly differentiated; Grade III‘                          1   2940.2
## - ‘RaceOther (American Indian/AK Native, Asian/Pacific Islander)‘  1   2940.9
## - Regional_nodes_examined                                          1   2943.9
## - Age                                                              1   2955.5
## - ‘GradeWell differentiated; Grade I‘                              1   2964.1
## - sixth_stageIIB                                                   1   2968.5
## - N_stageN3                                                        1   3007.3
## - Survival_months                                                  1   3941.6
##                                                                      AIC
## - sixth_stageIIIA                                                 2966.5
## - Marital_StatusSeparated                                         2966.6
## - RaceWhite                                                       2966.7
## - Regional_nodes_positive                                         2966.8
## <none>                                                            2966.9
## - ‘Marital_StatusMarried (including common law)‘                  2967.1
## - Progesterone_statusPositive                                     2971.1
## - Estrogen_statusPositive                                         2971.5
## - ‘GradeUndifferentiated; anaplastic; Grade IV‘                   2971.6
## - Tumor_size                                                      2973.3
## - T_stageT3                                                       2975.5
## - N_stageN2                                                       2976.6
## - T_stageT4                                                       2979.5
## - ‘GradePoorly differentiated; Grade III‘                         2980.2
## - ‘RaceOther (American Indian/AK Native, Asian/Pacific Islander)‘ 2980.9
## - Regional_nodes_examined                                         2983.9
## - Age                                                             2995.5
## - ‘GradeWell differentiated; Grade I‘                             3004.1
## - sixth_stageIIB                                                  3008.5
## - N_stageN3                                                       3047.3
## - Survival_months                                                 3981.6
## 
## Step:  AIC=2966.52
## Status ~ Age + ‘RaceOther (American Indian/AK Native, Asian/Pacific Islander)‘ + 
##     RaceWhite + ‘Marital_StatusMarried (including common law)‘ + 
##     Marital_StatusSeparated + T_stageT3 + T_stageT4 + N_stageN2 + 
##     N_stageN3 + sixth_stageIIB + ‘GradePoorly differentiated; Grade III‘ + 
##     ‘GradeUndifferentiated; anaplastic; Grade IV‘ + ‘GradeWell differentiated; Grade I‘ + 
##     Tumor_size + Estrogen_statusPositive + Progesterone_statusPositive + 
##     Regional_nodes_examined + Regional_nodes_positive + Survival_months
## 
##                                                                   Df Deviance
## - RaceWhite                                                        1   2928.3
## - Marital_StatusSeparated                                          1   2928.3
## <none>                                                                 2926.5
## - ‘Marital_StatusMarried (including common law)‘                   1   2928.7
## - Regional_nodes_positive                                          1   2928.7
## - Progesterone_statusPositive                                      1   2932.7
## - ‘GradeUndifferentiated; anaplastic; Grade IV‘                    1   2933.2
## - Estrogen_statusPositive                                          1   2933.2
## - Tumor_size                                                       1   2933.4
## - T_stageT4                                                        1   2939.5
## - ‘GradePoorly differentiated; Grade III‘                          1   2941.9
## - ‘RaceOther (American Indian/AK Native, Asian/Pacific Islander)‘  1   2942.1
## - T_stageT3                                                        1   2942.6
## - Regional_nodes_examined                                          1   2945.3
## - Age                                                              1   2958.8
## - N_stageN2                                                        1   2961.7
## - ‘GradeWell differentiated; Grade I‘                              1   2966.0
## - sixth_stageIIB                                                   1   2970.1
## - N_stageN3                                                        1   3015.2
## - Survival_months                                                  1   3942.2
##                                                                      AIC
## - RaceWhite                                                       2966.3
## - Marital_StatusSeparated                                         2966.3
## <none>                                                            2966.5
## - ‘Marital_StatusMarried (including common law)‘                  2966.7
## - Regional_nodes_positive                                         2966.7
## - Progesterone_statusPositive                                     2970.7
## - ‘GradeUndifferentiated; anaplastic; Grade IV‘                   2971.2
## - Estrogen_statusPositive                                         2971.2
## - Tumor_size                                                      2971.4
## - T_stageT4                                                       2977.5
## - ‘GradePoorly differentiated; Grade III‘                         2979.9
## - ‘RaceOther (American Indian/AK Native, Asian/Pacific Islander)‘ 2980.1
## - T_stageT3                                                       2980.6
## - Regional_nodes_examined                                         2983.3
## - Age                                                             2996.8
## - N_stageN2                                                       2999.7
## - ‘GradeWell differentiated; Grade I‘                             3004.0
## - sixth_stageIIB                                                  3008.1
## - N_stageN3                                                       3053.2
## - Survival_months                                                 3980.2
##
## Step:  AIC=2966.26
## Status ~ Age + ‘RaceOther (American Indian/AK Native, Asian/Pacific Islander)‘ +
##     ‘Marital_StatusMarried (including common law)‘ + Marital_StatusSeparated +
##     T_stageT3 + T_stageT4 + N_stageN2 + N_stageN3 + sixth_stageIIB +
##     ‘GradePoorly differentiated; Grade III‘ + ‘GradeUndifferentiated; anaplastic; Grade IV‘ +
##     ‘GradeWell differentiated; Grade I‘ + Tumor_size + Estrogen_statusPositive +
##     Progesterone_statusPositive + Regional_nodes_examined + Regional_nodes_positive +
##     Survival_months
##
##                                                                   Df Deviance
## - Marital_StatusSeparated                                          1   2929.9
## <none>                                                                 2928.3
## - Regional_nodes_positive                                          1   2930.3
## - ‘Marital_StatusMarried (including common law)‘                   1   2931.4
## - Progesterone_statusPositive                                      1   2934.4
## - Estrogen_statusPositive                                          1   2934.9
## - ‘GradeUndifferentiated; anaplastic; Grade IV‘                    1   2935.0
## - Tumor_size                                                       1   2935.5
## - T_stageT4                                                        1   2941.1
## - T_stageT3                                                        1   2944.2
## - ‘GradePoorly differentiated; Grade III‘                          1   2944.2
## - ‘RaceOther (American Indian/AK Native, Asian/Pacific Islander)‘  1   2944.8
## - Regional_nodes_examined                                          1   2947.1
## - Age                                                              1   2960.0
## - N_stageN2                                                        1   2964.1
## - ‘GradeWell differentiated; Grade I‘                              1   2968.4
## - sixth_stageIIB                                                   1   2973.1
## - N_stageN3                                                        1   3017.4
## - Survival_months                                                  1   3951.3
##                                                                      AIC
## - Marital_StatusSeparated                                         2965.9
## <none>                                                            2966.3
## - Regional_nodes_positive                                         2966.3
## - ‘Marital_StatusMarried (including common law)‘                  2967.4
## - Progesterone_statusPositive                                     2970.4
## - Estrogen_statusPositive                                         2970.9
## - ‘GradeUndifferentiated; anaplastic; Grade IV‘                    2971.0
## - Tumor_size                                                      2971.5
## - T_stageT4                                                       2977.1
## - T_stageT3                                                       2980.2
## - ‘GradePoorly differentiated; Grade III‘                          2980.2
## - ‘RaceOther (American Indian/AK Native, Asian/Pacific Islander)‘ 2980.8
## - Regional_nodes_examined                                         2983.1
## - Age                                                             2996.0
## - N_stageN2                                                       3000.1
## - ‘GradeWell differentiated; Grade I‘                              3004.4
## - sixth_stageIIB                                                  3009.1
## - N_stageN3                                                       3053.4
## - Survival_months                                                 3987.3
##
## Step:  AIC=2965.95
## Status ~ Age + ‘RaceOther (American Indian/AK Native, Asian/Pacific Islander)‘ +
##     ‘Marital_StatusMarried (including common law)‘ + T_stageT3 +
##     T_stageT4 + N_stageN2 + N_stageN3 + sixth_stageIIB + ‘GradePoorly differentiated; Grade III‘ +
##     ‘GradeUndifferentiated; anaplastic; Grade IV‘ + ‘GradeWell differentiated; Grade I‘ +
##     Tumor_size + Estrogen_statusPositive + Progesterone_statusPositive +
##     Regional_nodes_examined + Regional_nodes_positive + Survival_months
##
##                                                                   Df Deviance
## - Regional_nodes_positive                                          1   2931.9
## <none>                                                                 2929.9
## - ‘Marital_StatusMarried (including common law)‘                   1   2933.9
## - Progesterone_statusPositive                                      1   2936.4
## - ‘GradeUndifferentiated; anaplastic; Grade IV‘                    1   2936.6
## - Estrogen_statusPositive                                          1   2936.8
## - Tumor_size                                                       1   2937.2
## - T_stageT4                                                        1   2942.7
## - T_stageT3                                                        1   2945.5
## - ‘GradePoorly differentiated; Grade III‘                          1   2945.7
## - ‘RaceOther (American Indian/AK Native, Asian/Pacific Islander)‘  1   2945.9
## - Regional_nodes_examined                                          1   2949.2
## - Age                                                              1   2961.0
## - N_stageN2                                                        1   2965.8
## - ‘GradeWell differentiated; Grade I‘                              1   2970.1
## - sixth_stageIIB                                                   1   2974.8
## - N_stageN3                                                        1   3019.7
## - Survival_months                                                  1   3954.4
##                                                                      AIC
## - Regional_nodes_positive                                         2965.9
## <none>                                                            2965.9
## - ‘Marital_StatusMarried (including common law)‘                  2967.9
## - Progesterone_statusPositive                                     2970.4
## - ‘GradeUndifferentiated; anaplastic; Grade IV‘                   2970.6
## - Estrogen_statusPositive                                         2970.8
## - Tumor_size                                                      2971.2
## - T_stageT4                                                       2976.7
## - T_stageT3                                                       2979.5
## - ‘GradePoorly differentiated; Grade III‘                         2979.7
## - ‘RaceOther (American Indian/AK Native, Asian/Pacific Islander)‘ 2979.9
## - Regional_nodes_examined                                         2983.2
## - Age                                                             2995.0
## - N_stageN2                                                       2999.8
## - ‘GradeWell differentiated; Grade I‘                             3004.1
## - sixth_stageIIB                                                  3008.8
## - N_stageN3                                                       3053.7
## - Survival_months                                                 3988.4
##
## Step:  AIC=2965.94
## Status ~ Age + ‘RaceOther (American Indian/AK Native, Asian/Pacific Islander)‘ +
##     ‘Marital_StatusMarried (including common law)‘ + T_stageT3 +
##     T_stageT4 + N_stageN2 + N_stageN3 + sixth_stageIIB + ‘GradePoorly differentiated; Grade III‘ +
##     ‘GradeUndifferentiated; anaplastic; Grade IV‘ + ‘GradeWell differentiated; Grade I‘ +
##     Tumor_size + Estrogen_statusPositive + Progesterone_statusPositive +
##     Regional_nodes_examined + Survival_months
##
##                                                                   Df Deviance
## <none>                                                                 2931.9
## - ‘Marital_StatusMarried (including common law)‘                   1   2935.7
## - Progesterone_statusPositive                                      1   2938.6
## - ‘GradeUndifferentiated; anaplastic; Grade IV‘                    1   2938.6
## - Estrogen_statusPositive                                          1   2938.6
## - Tumor_size                                                       1   2938.9
## - T_stageT4                                                        1   2945.8
## - ‘GradePoorly differentiated; Grade III‘                          1   2947.0
## - ‘RaceOther (American Indian/AK Native, Asian/Pacific Islander)‘  1   2948.1
## - T_stageT3                                                        1   2948.4
## - Regional_nodes_examined                                          1   2950.3
## - Age                                                              1   2962.9
## - ‘GradeWell differentiated; Grade I‘                              1   2972.4
## - sixth_stageIIB                                                   1   2978.7
## - N_stageN2                                                        1   3025.7
## - N_stageN3                                                        1   3072.7
## - Survival_months                                                  1   3956.6
##                                                                      AIC
## <none>                                                            2965.9
## - ‘Marital_StatusMarried (including common law)‘                  2967.7
## - Progesterone_statusPositive                                     2970.6
## - ‘GradeUndifferentiated; anaplastic; Grade IV‘                   2970.6
## - Estrogen_statusPositive                                         2970.6
## - Tumor_size                                                      2970.9
## - T_stageT4                                                       2977.8
## - ‘GradePoorly differentiated; Grade III‘                         2979.0
## - ‘RaceOther (American Indian/AK Native, Asian/Pacific Islander)‘ 2980.1
## - T_stageT3                                                       2980.4
## - Regional_nodes_examined                                         2982.3
## - Age                                                             2994.9
## - ‘GradeWell differentiated; Grade I‘                             3004.4
## - sixth_stageIIB                                                  3010.7
## - N_stageN2                                                       3057.7
## - N_stageN3                                                       3104.7
## - Survival_months                                                 3988.6
```

```r
summary(Backward_log_bal)
```

```
##
## Call:
## glm(formula = Status ~ Age + ‘RaceOther (American Indian/AK Native, Asian/Pacific Islander)‘ +
##     ‘Marital_StatusMarried (including common law)‘ + T_stageT3 +
##     T_stageT4 + N_stageN2 + N_stageN3 + sixth_stageIIB + ‘GradePoorly differentiated; Grade III‘ +
##     ‘GradeUndifferentiated; anaplastic; Grade IV‘ + ‘GradeWell differentiated; Grade I‘ +
##     Tumor_size + Estrogen_statusPositive + Progesterone_statusPositive +
##     Regional_nodes_examined + Survival_months, family = binomial(link = "logit"),
##     data = train_breast_cancer_smote)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6677  -0.6796  -0.2620   0.6617   2.8806
##
## Coefficients:
##                                                                   Estimate
## (Intercept)                                                       2.548356
## Age                                                               0.029967
## ‘RaceOther (American Indian/AK Native, Asian/Pacific Islander)‘ -0.796714
## ‘Marital_StatusMarried (including common law)‘                  -0.189223
## T_stageT3                                                         0.680716
## T_stageT4                                                         1.078184
## N_stageN2                                                         1.314130
## N_stageN3                                                         1.823867
## sixth_stageIIB                                                    1.003517
## ‘GradePoorly differentiated; Grade III‘                           0.399166
## ‘GradeUndifferentiated; anaplastic; Grade IV‘                     1.805068
## ‘GradeWell differentiated; Grade I‘                             -1.144990
## Tumor_size                                                      -1.016266
## Estrogen_statusPositive                                         -0.566371
## Progesterone_statusPositive                                     -0.342123
## Regional_nodes_examined                                         -1.010212
## Survival_months                                                 -6.467651
##                                                                 Std. Error
## (Intercept)                                                       0.440546
## Age                                                               0.005435
## ‘RaceOther (American Indian/AK Native, Asian/Pacific Islander)‘   0.204605
## ‘Marital_StatusMarried (including common law)‘                    0.097664
## T_stageT3                                                         0.168449
## T_stageT4                                                         0.293439
## N_stageN2                                                         0.138491
## N_stageN3                                                         0.158871
## sixth_stageIIB                                                    0.148248
## ‘GradePoorly differentiated; Grade III‘                           0.102890
## ‘GradeUndifferentiated; anaplastic; Grade IV‘                     0.670165
## ‘GradeWell differentiated; Grade I‘                               0.190742
## Tumor_size                                                        0.387013
## Estrogen_statusPositive                                           0.220640
## Progesterone_statusPositive                                       0.133029
## Regional_nodes_examined                                           0.237236
## Survival_months                                                   0.251247
##                                                                     z value
## (Intercept)                                                          5.785
## Age                                                                  5.513
## ‘RaceOther (American Indian/AK Native, Asian/Pacific Islander)‘  -3.894
## ‘Marital_StatusMarried (including common law)‘                     -1.937
## T_stageT3                                                            4.041
## T_stageT4                                                            3.674
## N_stageN2                                                            9.489
## N_stageN3                                                           11.480
## sixth_stageIIB                                                       6.769
## ‘GradePoorly differentiated; Grade III‘                              3.880
## ‘GradeUndifferentiated; anaplastic; Grade IV‘                        2.693
## ‘GradeWell differentiated; Grade I‘                                 -6.003
## Tumor_size                                                          -2.626
## Estrogen_statusPositive                                             -2.567
## Progesterone_statusPositive                                         -2.572
## Regional_nodes_examined                                             -4.258
## Survival_months                                                    -25.742
##                                                                     Pr(>|z|)
## (Intercept)                                                        7.27e-09 ***
## Age                                                                3.52e-08 ***
## ‘RaceOther (American Indian/AK Native, Asian/Pacific Islander)‘ 9.86e-05 ***
## ‘Marital_StatusMarried (including common law)‘                     0.052686 .
## T_stageT3                                                          5.32e-05 ***
## T_stageT4                                                          0.000239 ***
## N_stageN2                                                           < 2e-16 ***
## N_stageN3                                                           < 2e-16 ***
## sixth_stageIIB                                                     1.30e-11 ***
## ‘GradePoorly differentiated; Grade III‘                            0.000105 ***
## ‘GradeUndifferentiated; anaplastic; Grade IV‘                      0.007071 **
## ‘GradeWell differentiated; Grade I‘                                1.94e-09 ***
## Tumor_size                                                         0.008641 **
## Estrogen_statusPositive                                            0.010260 *
## Progesterone_statusPositive                                        0.010117 *
## Regional_nodes_examined                                            2.06e-05 ***
## Survival_months                                                     < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4617.8  on 3380  degrees of freedom
## Residual deviance: 2931.9  on 3364  degrees of freedom
## AIC: 2965.9
##
## Number of Fisher Scoring iterations: 5
```

```
fitted.results1_bal <- predict(Backward_log_bal,newdata=test_breast_cancer,type='response')
fitted.results1_bal <- ifelse(fitted.results1_bal > 0.5,1,0)
tab_fitted1_bal <- table(fitted.results1_bal, test_breast_cancer$Status)

misClasificError <- mean(fitted.results1_bal != test_breast_cancer$Status)
print(paste('Accuracy',1-misClasificError))
```

```
## [1] "Accuracy 0.82625"
```

```
misClasificError1 <- mean(fitted.results1 != test_breast_cancer$Status)
print(paste('Accuracy',1-misClasificError1))
```

```
## [1] "Accuracy 0.9"
```

```
confusionMatrix(table(fitted.results_bal, test_breast_cancer$Status))
```

```
## Confusion Matrix and Statistics
##
##
## fitted.results_bal   0   1
##                  0 583  40
##                  1  97  80
##
##                Accuracy : 0.8288
##                  95% CI : (0.8008, 0.8542)
##     No Information Rate : 0.85
##     P-Value [Acc > NIR] : 0.9565
##
##                   Kappa : 0.4383
##
##  Mcnemar's Test P-Value : 1.715e-06
##
##             Sensitivity : 0.8574
##             Specificity : 0.6667
##          Pos Pred Value : 0.9358
##          Neg Pred Value : 0.4520
##              Prevalence : 0.8500
##          Detection Rate : 0.7288
##    Detection Prevalence : 0.7788
##       Balanced Accuracy : 0.7620
##
##        'Positive' Class : 0
##
```

```
log_cm_bal <- confusionMatrix(table(fitted.results1_bal, test_breast_cancer$Status))


#Results
#Before backward elimination
print(paste("For Balanced data:", "Precision is:",caret::precision(tab_fitted_bal),
"Recall is:",sensitivity(tab_fitted_bal),"F-score is:",caret::F_meas(tab_fitted_bal)))
```

```
## [1] "For Balanced data: Precision is: 0.907608695652174 Recall is: 0.982352941176471 F-score is: 0.9
```

```
#After Backward elimination
print(paste("For Balanced data","after backward elimination:", "Precision is:",caret::precision(tab_fit
"Recall is:",sensitivity(tab_fitted1_bal),"F-score is:",caret::F_meas(tab_fitted1_bal)))
```

```
## [1] "For Balanced data after backward elimination: Precision is: 0.934189406099519 Recall is: 0.8558
```

```
#We see that, a little improvement in the values after backward elimination
#step() could not remove all the non-signficant variables in the model. we
#can manually drop the non-significant variables having p-value of above 0.05
#We can evaluate the model performance using the both holdout method and
#k-fold cross validation method
```

#Model Evaluation(Holdout method)(support vector machines)

```
#setting seed to produce same results each time
set.seed(123)
#Performing model evaluation using holdout method the imbalanced training data
#on support vector machines
#model
#The trainControl() function is used to create a set of configuration options
#known as a control object. This object guides the train() function and allows
#for the selection of model evaluation criteria, such as the resampling strategy
#and the measure used for choosing the best model
Grid_svm <- expand.grid(C=c(1:10))
ctrl <- trainControl(method = "LGOCV", p=0.75)
model_svm <- train(Status ~ ., data = train_breast_cancer, method = "svmLinear",
                   trControl = ctrl, tuneGrid=Grid_svm)
model_svm
```

```
## Support Vector Machines with Linear Kernel
##
## 3206 samples
##   26 predictor
##    2 classes: '0', '1'
##
## No pre-processing
## Resampling: Repeated Train/Test Splits Estimated (25 reps, 75%)
## Summary of sample sizes: 2406, 2406, 2406, 2406, 2406, 2406, ...
## Resampling results across tuning parameters:
##
##   C   Accuracy  Kappa
##    1  0.88950   0.4495417
##    2  0.88940   0.4495918
##    3  0.88945   0.4495566
##    4  0.88930   0.4488983
##    5  0.88950   0.4500768
##    6  0.88935   0.4492554
##    7  0.88945   0.4499296
##    8  0.88940   0.4495896
##    9  0.88935   0.4496580
##   10  0.88935   0.4494598
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was C = 1.
```

```
fit_svm <- predict(model_svm, test_breast_cancer)

table(fit_svm, test_breast_cancer$Status)
```

```
##
## fit_svm   0   1
##       0 670  76
##       1  10  44
```

```
##Performing model evaluation using holdout method of the balanced training data
#on support vector machines
#model
ctrl1 <- trainControl(method = "LGOCV", p=0.75)
model_svm_bal <- train(Status ~ ., data = train_breast_cancer_smote,
                       method = "svmLinear", trControl = ctrl1, tuneGrid=Grid_svm)
model_svm_bal
```

```
## Support Vector Machines with Linear Kernel
##
## 3381 samples
##   26 predictor
##    2 classes: '0', '1'
##
## No pre-processing
## Resampling: Repeated Train/Test Splits Estimated (25 reps, 75%)
## Summary of sample sizes: 2536, 2536, 2536, 2536, 2536, 2536, ...
## Resampling results across tuning parameters:
##
##    C   Accuracy   Kappa
##    1  0.7973018  0.5832288
##    2  0.7976331  0.5839280
##    3  0.7977278  0.5841451
##    4  0.7976331  0.5839496
##    5  0.7977278  0.5841774
##    6  0.7977751  0.5842691
##    7  0.7977278  0.5841664
##    8  0.7976805  0.5840747
##    9  0.7975385  0.5837767
##   10  0.7976805  0.5840758
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was C = 6.
```

```
fit_svm1 <- predict(model_svm_bal, test_breast_cancer)

table(fit_svm1, test_breast_cancer$Status)
```

```
##
## fit_svm1   0   1
##        0 592  39
##        1  88  81
```

Results: For support vector machines, Holdout method output showing that, kappa statistic of the balanced data would be 0.606 at tuning parameter C value of 3 and for imbalanced data it would be 0.48 at C parameter value 6, In the next steps, we will tune the hyper parameter to check the any difference in the outputs and will consider the optimal model.

From the output, Tuning C hyper parameter of SVM for holdout method produces fairly similar results for both imbalanced and balanced data. C=6 is best parameter for imbalanced data and C=3 best for balanced data.

#Model Evaluation(Holdout Method)(Decision tree)

```r
#setting seed to get reproducible results
set.seed(123)
#Performing model evaluation using holdout method by taking
#partition of 0.75 of the imbalanced training data on decision tree algorithm
#using C5.0
Grid <- expand.grid(model="tree", trials=c(1,5,10,15,20,25,30),winnow=FALSE)

ctrl <- trainControl(method = "LGOCV", p=0.8, selectionFunction = "oneSE")
model_DT <- train(Status ~ ., data = train_breast_cancer, method = "C5.0",
                   trControl = ctrl, tuneGrid=Grid)
model_DT
```

```
## C5.0
##
## 3206 samples
##   26 predictor
##    2 classes: '0', '1'
##
## No pre-processing
## Resampling: Repeated Train/Test Splits Estimated (25 reps, 80%)
## Summary of sample sizes: 2566, 2566, 2566, 2566, 2566, 2566, ...
## Resampling results across tuning parameters:
##
##   trials  Accuracy   Kappa
##    1      0.8996250  0.5305062
##    5      0.8985625  0.5495790
##   10      0.9001875  0.5481603
##   15      0.9003125  0.5513967
##   20      0.9010625  0.5525106
##   25      0.9009375  0.5520555
##   30      0.9008125  0.5504538
##
## Tuning parameter 'model' was held constant at a value of tree
## Tuning
##  parameter 'winnow' was held constant at a value of FALSE
## Accuracy was used to select the optimal model using  the one SE rule.
## The final values used for the model were trials = 1, model = tree and winnow
##  = FALSE.
```

```r
fit_DT <- predict(model_DT, test_breast_cancer)

table(fit_DT, test_breast_cancer$Status)
```

```
##
## fit_DT   0   1
##      0 673  66
##      1   7  54
```

```r
##Performing model evaluation using holdout method of the balanced training data
#on decision tree algorithm
#model
ctrl <- trainControl(method = "LGOCV", p=0.8, selectionFunction = "oneSE")
model_DT_bal <- train(Status ~ ., data = train_breast_cancer_smote,
                       method = "C5.0", trControl = ctrl, tuneGrid=Grid)
model_DT_bal
```

```
## C5.0
##
## 3381 samples
##   26 predictor
##    2 classes: '0', '1'
##
## No pre-processing
## Resampling: Repeated Train/Test Splits Estimated (25 reps, 80%)
## Summary of sample sizes: 2706, 2706, 2706, 2706, 2706, 2706, ...
## Resampling results across tuning parameters:
##
##   trials  Accuracy   Kappa
##    1      0.8315852  0.6524579
##    5      0.8634667  0.7179799
##   10      0.8807704  0.7534895
##   15      0.8868741  0.7663095
##   20      0.8914370  0.7756012
##   25      0.8938074  0.7806148
##   30      0.8949926  0.7831317
##
## Tuning parameter 'model' was held constant at a value of tree
## Tuning
##  parameter 'winnow' was held constant at a value of FALSE
## Accuracy was used to select the optimal model using  the one SE rule.
## The final values used for the model were trials = 25, model = tree and winnow
##  = FALSE.
```

```r
fit_DT1 <- predict(model_DT_bal, test_breast_cancer)

table(fit_DT1, test_breast_cancer$Status)
```

```
##
## fit_DT1   0   1
##       0 626  50
##       1  54  70
```

Results: Holdout method output showing that, kappa statistic of the balanced data would be optimal at trials=30 with value of 0.78 and for imbalanced data optimal model was at trials=1 wit the value of 0.52.In the next steps, we will tune the hyperparameter to check the any difference in the outputs and will consider the optimal model. In the next steps, no need to tune the hyperparameters.we already did it here with different trials values.

#Model Evaluation(K-fold cross validation)(support vector machines)

```
#setting seed to produce same results each time
set.seed(123)
#Performing model evaluation using k-fold cross validation by taking 10
#repeated folds of the imbalanced training data on support vector machines
#model
#The trainControl() function is used to create a set of configuration options
#known as a control object. This object guides the train() function and allows
#for the selection of model evaluation criteria, such as the resampling strategy
#and the measure used for choosing the best model
Grid_svm <- expand.grid(C=c(1,2,3,4,5,6,7,8,9,10))
ctrl <- trainControl(method = "cv", number = 10)
model_svm <- train(Status ~ ., data = train_breast_cancer, method = "svmLinear",
                   trControl = ctrl, tuneGrid=Grid_svm)
model_svm
```

```
## Support Vector Machines with Linear Kernel
##
## 3206 samples
##   26 predictor
##    2 classes: '0', '1'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 2885, 2886, 2885, 2886, 2885, 2886, ...
## Resampling results across tuning parameters:
##
##   C   Accuracy   Kappa
##    1  0.8923888  0.4649769
##    2  0.8923888  0.4649769
##    3  0.8923888  0.4649769
##    4  0.8923888  0.4649769
##    5  0.8917657  0.4618497
##    6  0.8923888  0.4649769
##    7  0.8923888  0.4649769
##    8  0.8920782  0.4628487
##    9  0.8920763  0.4639780
##   10  0.8917657  0.4618497
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was C = 1.
```

```
fit_svm <- predict(model_svm, test_breast_cancer)

table(fit_svm, test_breast_cancer$Status)
```

```
##
## fit_svm   0   1
##       0 670  76
##       1  10  44
```

```
##Performing model evaluation using k-fold cross validation by taking 10
#repeated folds of the balanced training data on support vector machines
```

```
#model
ctrl1 <- trainControl(method = "cv", number = 10)
model_svm_bal <- train(Status ~ ., data = train_breast_cancer_smote,
                       method = "svmLinear", trControl = ctrl1, tuneGrid=Grid_svm)
model_svm_bal
```

```
## Support Vector Machines with Linear Kernel
##
## 3381 samples
##   26 predictor
##    2 classes: '0', '1'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 3043, 3043, 3043, 3043, 3044, 3042, ...
## Resampling results across tuning parameters:
##
##   C   Accuracy   Kappa
##    1  0.7932604  0.5745130
##    2  0.7941480  0.5763547
##    3  0.7944438  0.5770022
##    4  0.7944438  0.5770022
##    5  0.7947388  0.5776494
##    6  0.7944438  0.5770022
##    7  0.7947388  0.5776494
##    8  0.7944438  0.5770022
##    9  0.7944438  0.5770022
##   10  0.7944438  0.5770022
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was C = 5.
```

```
fit_svm1 <- predict(model_svm_bal, test_breast_cancer)

table(fit_svm1, test_breast_cancer$Status)
```

```
##
## fit_svm1   0   1
##        0 593  40
##        1  87  80
```

Results: The kappa is about “0.59” for balanced data, which agrees with the previous confusion matrix() from caret (the small difference is due to rounding). Using the suggested interpretation, we note that there is good agreement between the classifier's predictions and the actual values on different validation sets.

The final kappa was “0.48” for imbalanced data which is really moderate. It indicates that the model is no better at predicting then chance alone.

Imbalanced data got more accuracy because this is especially important for data sets with severe class imbalance because a classifier can obtain high accuracy simply by always guessing the most frequent class. The kappa statistic will only reward the classifier if it is correct more often than this simplistic strategy.

Tuning C hyper parameter in this model was of no use. Every “C” parameter produces exactly same results. So, default value of C is better to consider.

#Model Evaluation(K-fold cross validation)(Decision Trees)

```r
#setting seed to get reproducible results
set.seed(123)
#Performing model evaluation using k-fold cross validation by taking 10
#repeated folds of the imbalanced training data on decision tree algorithm
#using C5.0
Grid <- expand.grid(model="tree", trials=c(1,5,10,15,20,25,30),winnow=FALSE)

ctrl <- trainControl(method = "cv", number = 10, selectionFunction = "oneSE")
model_DT <- train(Status ~ ., data = train_breast_cancer, method = "C5.0",
                   trControl = ctrl,tuneGrid=Grid)
model_DT
```

```
## C5.0
##
## 3206 samples
##   26 predictor
##    2 classes: '0', '1'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 2885, 2886, 2885, 2886, 2885, 2886, ...
## Resampling results across tuning parameters:
##
##   trials  Accuracy   Kappa
##    1      0.9008078  0.5300372
##    5      0.8961388  0.5372690
##   10      0.9004992  0.5538582
##   15      0.9014328  0.5531395
##   20      0.9005002  0.5481757
##   25      0.9014348  0.5520820
##   30      0.9004982  0.5430428
##
## Tuning parameter 'model' was held constant at a value of tree
## Tuning
##  parameter 'winnow' was held constant at a value of FALSE
## Accuracy was used to select the optimal model using  the one SE rule.
## The final values used for the model were trials = 1, model = tree and winnow
##  = FALSE.
```

```r
fit_DT <- predict(model_DT, test_breast_cancer)

confusionMatrix(table(fit_DT, test_breast_cancer$Status))
```

```
## Confusion Matrix and Statistics
##
##
## fit_DT   0   1
##      0 673  66
##      1   7  54
##
##                Accuracy : 0.9087
##                  95% CI : (0.8866, 0.9278)
##     No Information Rate : 0.85
##     P-Value [Acc > NIR] : 5.013e-07
##
##                   Kappa : 0.5513
##
##  Mcnemar’s Test P-Value : 1.134e-11
##
##             Sensitivity : 0.9897
##             Specificity : 0.4500
##          Pos Pred Value : 0.9107
##          Neg Pred Value : 0.8852
##              Prevalence : 0.8500
##          Detection Rate : 0.8413
##    Detection Prevalence : 0.9237
##       Balanced Accuracy : 0.7199
##
##        ’Positive’ Class : 0
##
```

```
##Performing model evaluation using k-fold cross validation by taking 10
#repeated folds of the balanced training data on decision trees
#model
set.seed(123)
ctrl <- trainControl(method = "cv", number = 10, selectionFunction = "oneSE")
model_DT_bal <- train(Status ~ ., data = train_breast_cancer_smote,
                       method = "C5.0", trControl = ctrl,tuneGrid=Grid)
model_DT_bal
```

```
## C5.0
##
## 3381 samples
##   26 predictor
##    2 classes: ’0’, ’1’
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 3042, 3043, 3044, 3043, 3043, 3043, ...
## Resampling results across tuning parameters:
##
##   trials  Accuracy   Kappa
##    1      0.8281577  0.6446770
##    5      0.8627670  0.7168156
##   10      0.8775538  0.7469327
##   15      0.8881977  0.7687146
##   20      0.8964739  0.7862536
##   25      0.8991410  0.7917940
##   30      0.8997275  0.7929644
##
## Tuning parameter ’model’ was held constant at a value of tree
## Tuning
##  parameter ’winnow’ was held constant at a value of FALSE
## Accuracy was used to select the optimal model using  the one SE rule.
## The final values used for the model were trials = 20, model = tree and winnow
##  = FALSE.
```

```
fit_DT1 <- predict(model_DT_bal, test_breast_cancer)

confusionMatrix(table(fit_DT1, test_breast_cancer$Status))
```

```
## Confusion Matrix and Statistics
##
##
## fit_DT1   0   1
##       0 628  48
##       1  52  72
##
##                Accuracy : 0.875
##                  95% CI : (0.8501, 0.8971)
##     No Information Rate : 0.85
##     P-Value [Acc > NIR] : 0.02469
##
##                   Kappa : 0.5164
##
##  Mcnemar's Test P-Value : 0.76418
##
##             Sensitivity : 0.9235
##             Specificity : 0.6000
##          Pos Pred Value : 0.9290
##          Neg Pred Value : 0.5806
##              Prevalence : 0.8500
##          Detection Rate : 0.7850
##    Detection Prevalence : 0.8450
##       Balanced Accuracy : 0.7618
##
##        'Positive' Class : 0
##
```

Results:The best model here is with balanced data having kappa value of 0.805 at trials = 30 and it is comparatively greater than the kappa value for imbalanced data of trials =1, because here we used selectionfunction "oneSE" instead of base function to get the optimal model.

For balanced data, it produces optimal model with trials=30 For imbalanced data, it produces optimal model with trials=1

Finally, by comparing two decision tree k-fold cross validation, I would probably choose balanced data with trials=30. In the next steps, no need to tune the hyper parameters.we already did it here with different trials values.

#Model Evaluation(Holdout method)(Logistic regression)

```
#setting seed to produce same results each time
set.seed(123)
#Performing model evaluation using logistic regression of the imbalanced training
#data

#The trainControl() function is used to create a set of configuration options
#known as a control object. This object guides the train() function and allows
```

```
#for the selection of model evaluation criteria, such as the resampling strategy
#and the measure used for choosing the best model

ctrl <- trainControl(method = "LGOCV", p=0.8)
model_log <- train(Status ~ ., data = train_breast_cancer, method = "glm",
                   family=binomial(link="logit"),trControl = ctrl)
model_log
```

```
## Generalized Linear Model
##
## 3206 samples
##   26 predictor
##    2 classes: '0', '1'
##
## No pre-processing
## Resampling: Repeated Train/Test Splits Estimated (25 reps, 80%)
## Summary of sample sizes: 2566, 2566, 2566, 2566, 2566, 2566, ...
## Resampling results:
##
##   Accuracy  Kappa
##   0.89075   0.4843035
```

```
fit_log <- predict(model_log, test_breast_cancer)

table(fit_log, test_breast_cancer$Status)
```

```
##
## fit_log   0   1
##       0 668  68
##       1  12  52
```

```
set.seed(123)
##Performing model evaluation using holdout method of partition 0.75
#of the balanced training data on logistic regression model
ctrl1 <- trainControl(method = "LGOCV", p=0.8)
model_log_bal <- train(Status ~ ., data = train_breast_cancer_smote,
                       method = "glm",family=binomial(link="logit"),
                       trControl = ctrl1)
model_log_bal
```

```
## Generalized Linear Model
##
## 3381 samples
##   26 predictor
##    2 classes: '0', '1'
##
## No pre-processing
## Resampling: Repeated Train/Test Splits Estimated (25 reps, 80%)
## Summary of sample sizes: 2706, 2706, 2706, 2706, 2706, 2706, ...
## Resampling results:
##
##   Accuracy   Kappa
##   0.7969778  0.5823727
```

```
fit_log1 <- predict(model_log_bal, test_breast_cancer)

table(fit_log1, test_breast_cancer$Status)
```

```
##
## fit_log1   0   1
##        0 583  40
##        1  97  80
```

#Model Evaluation(k-fold cross validation method)(Logistic regression)

```
#setting seed to get reproducible results
set.seed(123)
#Performing model evaluation using k-fold cross validation by taking 10
#repeated folds of the imbalanced training data on logistic regression algorithm
#using glm with parameter binomial(link="logit")

ctrl <- trainControl(method = "cv", number = 10, selectionFunction = "oneSE")
model_log_k <- train(Status ~ ., data = train_breast_cancer, method = "glm",
                   family=binomial(link="logit"),trControl = ctrl)
model_log_k
```

```
## Generalized Linear Model
##
## 3206 samples
##   26 predictor
##    2 classes: '0', '1'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 2885, 2886, 2885, 2886, 2885, 2886, ...
## Resampling results:
##
##   Accuracy   Kappa
##   0.8920743  0.4845453
```

```
fit_log_k <- predict(model_log_k, test_breast_cancer)

table(fit_log_k, test_breast_cancer$Status)
```

```
##
## fit_log_k   0   1
##         0 668  68
##         1  12  52
```

```
set.seed(123)
##Performing model evaluation using k-fold cross validation by taking 10
#repeated folds of the balanced training data on logistic regression
#model
ctrl <- trainControl(method = "cv", number = 10, selectionFunction = "oneSE")
model_log_bal_k <- train(Status ~ ., data = train_breast_cancer_smote,
                       method = "glm", family=binomial(link="logit"),trControl = ctrl)
model_log_bal_k
```

```
## Generalized Linear Model
##
## 3381 samples
##   26 predictor
##    2 classes: '0', '1'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 3042, 3043, 3044, 3043, 3043, 3043, ...
## Resampling results:
##
##   Accuracy   Kappa
##   0.7965104  0.5814829
```

```
fit_log_bal_k <- predict(model_log_bal_k, test_breast_cancer)

table(fit_log_bal_k, test_breast_cancer$Status)
```

```
##
## fit_log_bal_k   0   1
##             0 583  40
##             1  97  80
```

Results: For support vector machines, Holdout method output showing that, kappa statistic of the balanced data would be 0.606 at tuning parameter C value of 3 and for imbalanced data it would be 0.48 at C parameter value 6, In the next steps, we will tune the hyper parameter to check the any difference in the outputs and will consider the optimal model.

From the output, Tuning C hyper parameter of SVM for holdout method produces fairly similar results for both imbalanced and balanced data. C=6 is best parameter for imbalanced data and C=3 best for balanced data.

Results: The kappa is about "0.59" for balanced data, which agrees with the previous confusion matirx() from caret (the small difference is due to rounding). Using the suggested interpretation, we note that there is good agreement between the classifier's predictions and the actual values on different validation sets.

The final kappa was "0.48" for imbalanced data which is really moderate. It indicates that the model is no better at predicting then chance alone.

Imbalanced data got more accuracy because this is especially important for data sets with severe class imbalance because a classifier can obtain high accuracy simply by always guessing the most frequent class. The kappa statistic will only reward the classifier if it is correct more often than this simplistic strategy.

Tuning C hyper parameter in this model was of no use. Every "C" parameter produces exactly same results. So, default value of C is better to consider.

Results:The best model here is with balanced data having kappa value of 0.805 at trials = 20 and it is comparatively greater than the kappa value for imbalanced data of trials =1, because here we used selectionfunction "oneSE" instead of base function to get the optimal model.

For balanced data, it produces optimal model with trials=20 For imbalanced data, it produces optimal model with trials=1

Finally, by comparing two decision tree k-fold cross validation, I would probably choose balanced data with trials=20. In the next steps, no need to tune the hyper parameters.we already did it here with different trials values.

Results:The best model here is with balanced data having kappa value of 0.805 at trials = 20 and it is comparatively greater than the kappa value for imbalanced data of trials =1, because here we used selectionfunction “oneSE” instead of base function to get the optimal model.

For balanced data, it produces optimal model with trials=20 For imbalanced data, it produces optimal model with trials=1

Finally, by comparing two decision tree k-fold cross validation, I would probably choose balanced data with trials=20. In the next steps, no need to tune the hyper parameters.we already did it here with different trials values.

Logistic regression of k-fold cross validation showing high number of false negative and moderate agreement of actual and predicted values.

Comparing all the models: By comparing all the models above, I would probably choose decision tree on balanced data with kappa value of 0.8 and having low false negatives and best predicting in negative class.

#Model4(Random Forests k-fold)#Imbalanced data

```
library(randomForest)
#setting seed to random number to get reproducible results
set.seed(123)

ctrl <- trainControl(method = "cv", number = 10, selectionFunction = "oneSE")

rf_model <- train(Status ~., data=train_breast_cancer, method="rf",
                  metric=c("Accuracy"), trControl=ctrl)

predict_rf_imb = predict(rf_model, newdata = test_breast_cancer)

# Confusion matrix on test set
confusionMatrix(table(predict_rf_imb, test_breast_cancer$Status))
```

```
## Confusion Matrix and Statistics
##
##
## predict_rf_imb   0   1
##              0 668  59
##              1  12  61
##
##                 Accuracy : 0.9112
##                   95% CI : (0.8894, 0.93)
##      No Information Rate : 0.85
##      P-Value [Acc > NIR] : 1.537e-07
##
##                    Kappa : 0.585
##
##   Mcnemar's Test P-Value : 4.783e-08
##
##              Sensitivity : 0.9824
##              Specificity : 0.5083
##           Pos Pred Value : 0.9188
##           Neg Pred Value : 0.8356
##               Prevalence : 0.8500
##           Detection Rate : 0.8350
##     Detection Prevalence : 0.9087
##       Balanced Accuracy : 0.7453
##
##        'Positive' Class : 0
##
```

#Model5(Random Forests)#Balanced data

```
set.seed(100)

ctrl <- trainControl(method = "cv", number = 10, selectionFunction = "oneSE")

rf_model_bal <- train(Status ~., data=train_breast_cancer_smote, method="rf",
                      metric=c("Accuracy"), trControl=ctrl)

predict_rf_bal = predict(rf_model_bal, newdata = test_breast_cancer)

# Confusion matrix on test set
confusionMatrix(table(predict_rf_bal, test_breast_cancer$Status))
```

```
## Confusion Matrix and Statistics
##
##
## predict_rf_bal   0   1
##              0 622  49
##              1  58  71
##
##                Accuracy : 0.8662
##                  95% CI : (0.8407, 0.8891)
##     No Information Rate : 0.85
##     P-Value [Acc > NIR] : 0.1067
##
##                   Kappa : 0.4912
##
##  Mcnemar's Test P-Value : 0.4393
##
##             Sensitivity : 0.9147
##             Specificity : 0.5917
##          Pos Pred Value : 0.9270
##          Neg Pred Value : 0.5504
##              Prevalence : 0.8500
##          Detection Rate : 0.7775
##    Detection Prevalence : 0.8387
##       Balanced Accuracy : 0.7532
##
##        'Positive' Class : 0
##
```

#Model Tuning & performance improvement(Meta Learning) #Bagging(With homogeneous learners)(same algorithms)

```
RNGversion("3.5.2")
set.seed(300)
```

```r
#Using bagging function to ensemble the model on imbalanced train data by
#taking nbags parameter as 25
bag_imb <- bagging(Status~., data=train_breast_cancer, nbag=25)

bag_pred_imb <- predict(bag_imb, test_breast_cancer)

tab_bag_imb <- table(bag_pred_imb, test_breast_cancer$Status)

confusionMatrix(tab_bag_imb)
```

```
## Confusion Matrix and Statistics
##
##
## bag_pred_imb   0   1
##            0 665  60
##            1  15  60
##
##                Accuracy : 0.9062
##                  95% CI : (0.8839, 0.9255)
##     No Information Rate : 0.85
##     P-Value [Acc > NIR] : 1.542e-06
##
##                   Kappa : 0.5652
##
##  Mcnemar's Test P-Value : 3.761e-07
##
##             Sensitivity : 0.9779
##             Specificity : 0.5000
##          Pos Pred Value : 0.9172
##          Neg Pred Value : 0.8000
##              Prevalence : 0.8500
##          Detection Rate : 0.8313
##    Detection Prevalence : 0.9062
##       Balanced Accuracy : 0.7390
##
##        'Positive' Class : 0
##
```

```r
RNGversion("3.5.2")
set.seed(300)
ctrl <- trainControl(method = "repeatedcv", number = 10)


train_bag <- train_breast_cancer
train_bag[,c(2:20,22,23)] <- lapply(train_bag[,c(2:20,22,23)], factor)

train(Status ~ ., data = train_bag[,c(1,21,24:27)], method="treebag",
 trControl = ctrl) #kappa : 0.49
```

```
## Bagged CART
##
## 3206 samples
##    5 predictor
##    2 classes: '0', '1'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold, repeated 1 times)
## Summary of sample sizes: 2885, 2886, 2886, 2884, 2886, 2885, ...
## Resampling results:
##
##   Accuracy   Kappa
##   0.8895821  0.5006647
```

```
#Using bagging function to ensemble the model on Balanced train data by
#taking nbags parameter as 25
bag_bal <- bagging(Status~., data=train_breast_cancer_smote, nbag=25)

bag_pred_bal <- predict(bag_bal, test_breast_cancer)

tab_bag_bal <- table(bag_pred_bal, test_breast_cancer$Status)

confusionMatrix(tab_bag_bal)
```

```
## Confusion Matrix and Statistics
##
##
## bag_pred_bal   0   1
##            0 608  48
##            1  72  72
##
##                Accuracy : 0.85
##                  95% CI : (0.8233, 0.874)
##     No Information Rate : 0.85
##     P-Value [Acc > NIR] : 0.52433
##
##                   Kappa : 0.4565
##
##  Mcnemar's Test P-Value : 0.03576
##
##             Sensitivity : 0.8941
##             Specificity : 0.6000
##          Pos Pred Value : 0.9268
##          Neg Pred Value : 0.5000
##              Prevalence : 0.8500
##          Detection Rate : 0.7600
##    Detection Prevalence : 0.8200
##       Balanced Accuracy : 0.7471
##
##        'Positive' Class : 0
##
```

```
RNGversion("3.5.2")
set.seed(123)
ctrl <- trainControl(method = "cv", number = 10)
bag_cv_bal <- train(Status ~ ., data = train_breast_cancer_smote[,c(1,21,24:27)],
                    method = "treebag",trControl = ctrl)

bag_cv_bal   #kappa score: 0.769
```

```
## Bagged CART
##
## 3381 samples
##    5 predictor
##    2 classes: '0', '1'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 3043, 3043, 3043, 3043, 3043, 3044, ...
## Resampling results:
##
##   Accuracy  Kappa
##   0.883766  0.7602641
```

Output: For Imbalanced data(Bagging)

The best C5.0 decision tree we adapted previously in this chapter had a 0.43 kappa statistic, and the 0.76 kappa value for this model shows that the bagged tree model works very well. This demonstrates the effectiveness of ensemble methods: when working together, a group of simple learners can outperform extremely complex models.

For Imbalanced data(Bagging) There is little change in the both accuracy and kappa statistic values.

Here, I found interesting that for the bagging process, the number of true negatives(TN) increased for both imbalanced and balanced data sets from the previous decision tree algorithm

#Construction of ensemble model as function

```
Ensemble_model <- function(data, algorithms) {

  # Set the seed for reproducibility
  set.seed(100)

  # Set up the train control object for repeated cross-validation
  control_stacking <- trainControl(
    method = "cv",
    number=10,
    selectionFunction = "oneSE"
  )

  # Train the stacked models using caretList
  stacked_models <- caretList(
    Status ~ .,
    data = data,
    trControl = control_stacking,
    methodList = algorithms,
    family=binomial(link="logit")
  )

  # Generate resampling results for the stacked models
  stacking_results <- resamples(stacked_models)
  # Return the summary of the resampling results
  return(stacking_results)

}
```

#Application of ensemble to make prediction

```
algorithms <- c("svmLinear","C5.0","glm","treebag","rf")
#Imbalanced data
Ensemble_imb <- Ensemble_model(train_breast_cancer[,c(1,21,24:27)], algorithms)
```

```
## Warning in trControlCheck(x = trControl, y = target): trControl$savePredictions
## not ’all’ or ’final’.  Setting to ’final’ so we can ensemble the models.
```

```
## Warning in trControlCheck(x = trControl, y = target): indexes not defined in
## trControl.  Attempting to set them ourselves, so each model in the ensemble
## will have the same resampling indexes.
```

```
## Warning: ’trials’ should be <= 1 for this object. Predictions generated using 1
## trials
```

```
## Warning: ’trials’ should be <= 1 for this object. Predictions generated using 1
## trials
```

```
## Warning: ’trials’ should be <= 1 for this object. Predictions generated using 1
## trials
```

```
## Warning: ’trials’ should be <= 1 for this object. Predictions generated using 1
## trials
```

```
## Warning: ’trials’ should be <= 1 for this object. Predictions generated using 1
## trials
```

```
## Warning: ’trials’ should be <= 1 for this object. Predictions generated using 1
## trials
```

```
## Warning: ’trials’ should be <= 1 for this object. Predictions generated using 1
## trials
```

```
## Warning: ’trials’ should be <= 1 for this object. Predictions generated using 1
## trials
```

```
## Warning: ’trials’ should be <= 1 for this object. Predictions generated using 1
## trials
```

```
## Warning: ’trials’ should be <= 1 for this object. Predictions generated using 1
## trials
```

```
## Warning: ’trials’ should be <= 1 for this object. Predictions generated using 1
## trials
```

```
## Warning: ’trials’ should be <= 1 for this object. Predictions generated using 1
## trials

## Warning: 'trials' should be <= 1 for this object. Predictions generated using 1
## trials

## Warning: 'trials' should be <= 1 for this object. Predictions generated using 1
## trials

## Warning: 'trials' should be <= 1 for this object. Predictions generated using 1
## trials

## Warning: 'trials' should be <= 1 for this object. Predictions generated using 1
## trials
```

```
#Balanced data
Ensemble_bal <- Ensemble_model(train_breast_cancer_smote[,c(1,21,24:27)], algorithms)
```

```
## Warning in trControlCheck(x = trControl, y = target): trControl$savePredictions
## not 'all' or 'final'.  Setting to 'final' so we can ensemble the models.

## Warning in trControlCheck(x = trControl, y = target): indexes not defined in
## trControl.  Attempting to set them ourselves, so each model in the ensemble
## will have the same resampling indexes.
```

```
summary(Ensemble_imb)
```

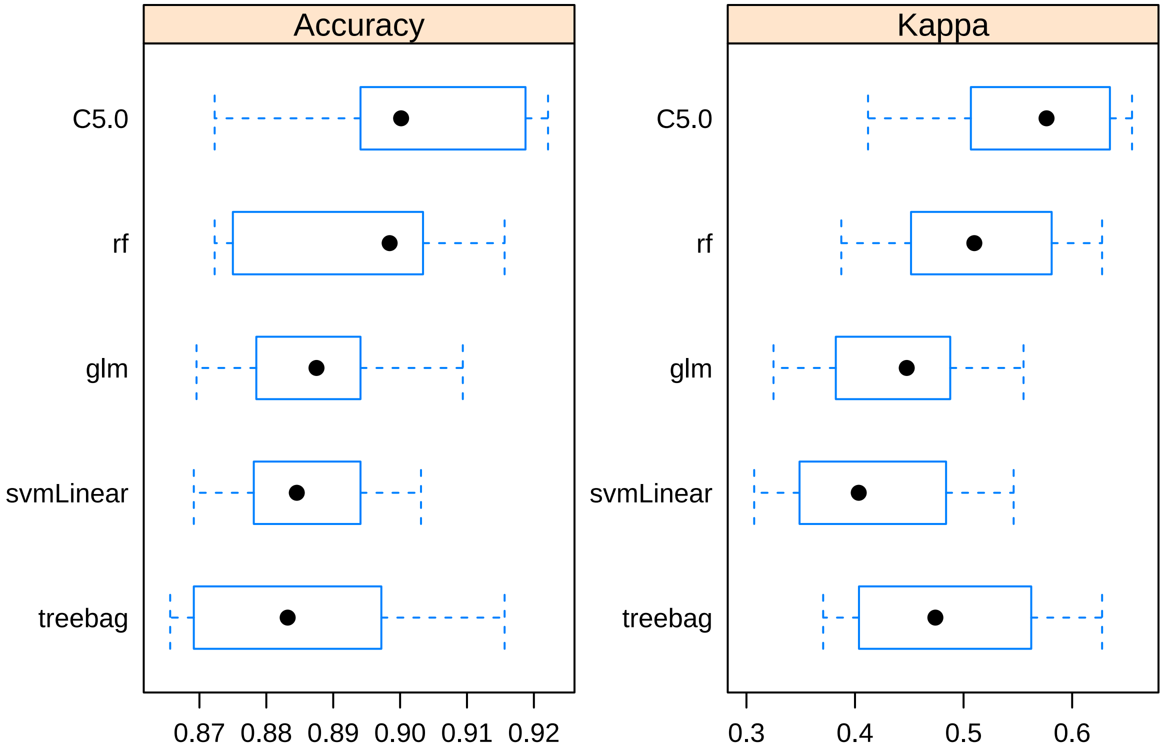
```
##
## Call:
## summary.resamples(object = Ensemble_imb)
##
## Models: svmLinear, C5.0, glm, treebag, rf
## Number of resamples: 10
##
## Accuracy
##                Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
## svmLinear 0.8691589 0.8781250 0.8845551 0.8846025 0.8917445 0.9031250    0
## C5.0      0.8722741 0.8947795 0.9001509 0.9020743 0.9164062 0.9221184    0
## glm       0.8695652 0.8799723 0.8875000 0.8877266 0.8925234 0.9093750    0
## treebag   0.8656250 0.8714004 0.8831910 0.8855342 0.8971159 0.9156250    0
## rf        0.8722741 0.8774338 0.8984375 0.8933409 0.9034268 0.9156250    0
##
## Kappa
##                Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
## svmLinear 0.3071230 0.3572621 0.4034461 0.4137452 0.4701507 0.5461201    0
## C5.0      0.4120091 0.5139780 0.5764337 0.5626342 0.6286821 0.6551724    0
## glm       0.3249476 0.3915666 0.4476591 0.4430823 0.4853416 0.5552367    0
## treebag   0.3706743 0.4132011 0.4740155 0.4845064 0.5557115 0.6275862    0
## rf        0.3873761 0.4560532 0.5099021 0.5115324 0.5791305 0.6275862    0
```

```
summary(Ensemble_bal)
```

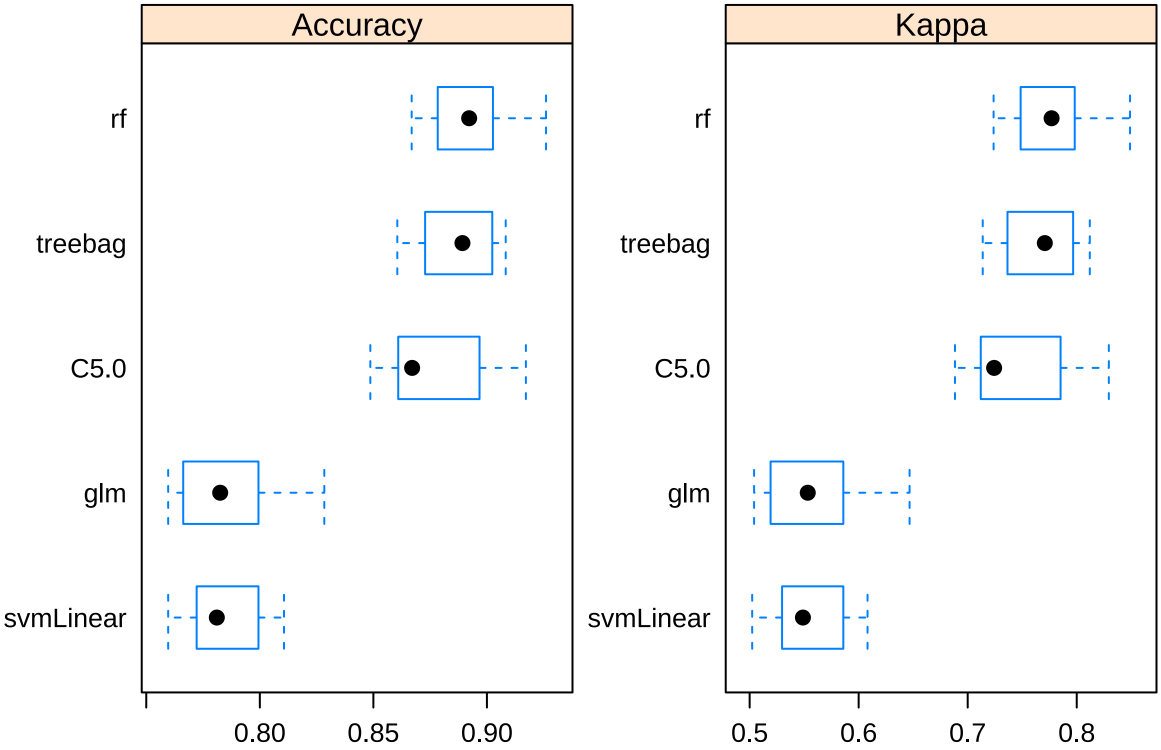
```
##
```

```
## Call:
## summary.resamples(object = Ensemble_bal)
##
## Models: svmLinear, C5.0, glm, treebag, rf
## Number of resamples: 10
##
## Accuracy
##                Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
## svmLinear 0.7596439 0.7721893 0.7810651 0.7837736 0.7963031 0.8106509    0
## C5.0      0.8486647 0.8616864 0.8670603 0.8763564 0.8915013 0.9171598    0
## glm       0.7596439 0.7662722 0.7825444 0.7843654 0.7963031 0.8284024    0
## treebag   0.8605341 0.8765622 0.8892191 0.8875990 0.9001479 0.9082840    0
## rf        0.8668639 0.8799076 0.8921820 0.8938138 0.9018432 0.9260355    0
##
## Kappa
##                Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
## svmLinear 0.5023246 0.5303668 0.5488619 0.5528421 0.5792355 0.6081301    0
## C5.0      0.6883195 0.7135635 0.7242444 0.7435292 0.7736830 0.8294467    0
## glm       0.5040966 0.5194195 0.5532925 0.5555201 0.5799969 0.6467111    0
## treebag   0.7137823 0.7441045 0.7707193 0.7682001 0.7930846 0.8119864    0
## rf        0.7237459 0.7517721 0.7769333 0.7806527 0.7964893 0.8488967    0
```

```r
#Draw box plots to compare the models
scales <- list(x=list(relation="free"), y=list(relation="free"))
#Imbalanced data plots
bwplot(Ensemble_imb, scales=scales)
```



```
#Balanced data plots
bwplot(Ensemble_bal, scales=scales)
```



Results: From the results, None of the above methods performs well. But, when compared to other decision tree and random forests were the best.The Random Forest, bagging are scored well in terms of precision and recall scores when compared to other models and the oversampling strategy for the SMOTE approach after training and testing the models. The Support Vector Classifier and Logistic Regression models, which had the highest recall scores, did not do as well in accurately predicting the 'Alive' class, though. This is due to the fact that for these models, the 'False Positive' value—which indicates the number of times the model erroneously predicted that a patient was dead—was exceptionally high. Using distinct method SMOTE, we also attempted testing the models on datasets that were both imbalanced and balanced. The outcomes, however, fell short of expectations. One possible explanation for this is the lack of a significant correlation between the features in the data set and the target label. We need more variables to get accurate predictions. The performance metrics also demonstrated that these oversampling strategies did not always result in better forecasts.

#Majority voting(for binary classification)

```
predict_status <- function(testdata){
  svm_pred <- predict(svm_model_bal, test_breast_cancer[,-27])
  log_reg_pred <- predict(log_bal, test_breast_cancer[,-27])
  dec_tree_pred <- predict(dt_bal, test_breast_cancer[,-27], type = "class")
  bag_pred <- predict(bag_bal, test_breast_cancer[,-27])
 pred_majority <-  ifelse(sum(svm_pred == "1") + sum(log_reg_pred== "1") + sum(dec_tree_pred == "1") + s

 return(pred_majority)
}
```

```

predict_status(test_breast_cancer)
```

```
## [1] "1"
```

Ensemble method was working well on the test data by majority voting

#Feature engineering task(Additional support)

```
#cancer_basic <- SEER_breast_cancer_df



#cancer_basic[,c(9,12:14)] <- lapply(cancer_basic[,c(9,12:14)], normalize)


#hist(cancer_basic$Age)
#hist(cancer_basic$Tumor_size)
#hist(cancer_basic$Regional_nodes_examined)
#hist(cancer_basic$Regional_nodes_positive)
#hist(cancer_basic$Survival_months)


#skewness(cancer_basic$Tumor_size)
#skewness(cancer_basic$Regional_nodes_examined)
#skewness(cancer_basic$Regional_nodes_positive)
#skewness(cancer_basic$Survival_months)

#skewness((cancer_basic$Tumor_size))
#skewness(sqrt(cancer_basic$Regional_nodes_examined))
#hist((sqrt(cancer_basic$Regional_nodes_positive)))
#skewness((log10(cancer_basic$Survival_months)))
```

#CRISP-DM APPROACH:

For my machine learning project, I have chosen the SEER Breast cancer Dataset. This data set contains information on over 4024 patients and includes features such as age, marital status, race, tumor_size, survival months and whether or not the patient alive or dead. The data set can be found on ieee website at the following URL: https://ieee-dataport.org/open-access/seer-breast-cancer-data Links to an external site.

My goal is to develop a predictive model to identify individuals alive or dead based on their Age, marital status, race and other characteristics. The target variable is binary variable(alive or dead). So, this is classification task

The target variable is one of 15 features (variables) in the data set, which includes 4024 rows overall.Both category and numerical features are present, and some of them have missing values that call for imputation or elimination. The use of machine learning and statistical methods to derive insights from the data and anticipate future outcomes makes this task suitable for classification as a data mining task.

I intend to test a variety of algorithms in order to construct the predictive model and determine the one that works best in this situation. I'll be using random forests, decision trees, support vector machines, and logistic regression among other approaches.These algorithms were chosen because they can handle both numerical and categorical data and are effective for binary classification problems.

BUSINESS UNDERSTANDING: Breast cancer prediction using machine learning entails analyzing data related to breast cancer, such as patient demographics, medical history, genetic factors, and imaging results, using algorithms and statistical models. The goal is to create accurate models that can predict a patient's risk of getting breast cancer or the chance of recurrence in individuals who have already been diagnosed. Health care providers can use these models to identify patients who are at high risk for breast cancer and suggest screening and preventative procedures. They can also be utilized to create personalized treatment strategies for patients based on their unique risk factors.

Companies that sell breast cancer preventative measures and treatment products and services can utilize machine learning algorithms to identify potential clients and customize their advertising messages. A company that sells bras for breast cancer survivors, for example, can use machine learning to identify people who have had mastectomies and target them with personalized advertisements. Machine learning algorithms can be used to examine a massive amount of data and identify patterns and trends that human researchers may not notice. This can assist companies develop new breast cancer treatments and therapies, as well as find new risk factors and prevention strategies.

Furthermore, the use of machine learning to predict breast cancer has the potential to increase the efficiency and accuracy of breast cancer diagnosis. Machine learning algorithms, for example, can be used to scan mammograms and indicate concerning areas that may require further evaluation. This can help reduce the number of false positives and false negatives, boosting overall breast cancer screening accuracy.

References: Rabiei, R., Ayyoubzadeh, S. M., Sohrabei, S., Esmaeili, M., & Atashi, A. (2022, June 1). Prediction of breast cancer using machine learning approaches. Journal of biomedical physics & engineering. Retrieved April 26, 2023, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9175124/#:~:text=The%20proposed%20machine%2Dlearning%20approaches,interventions%20at%20the%20right%20time.

Nasser, M., & Yusof, U. K. (2023, January 3). Deep learning based methods for breast cancer diagnosis: A systematic review and future direction. Diagnostics (Basel, Switzerland). Retrieved April 26, 2023, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9818155/ Alzu'bi, A., Najadat, H., Doulat, W., Al-Shari, O., & Zhou, L. (2021, January 18).

Predicting the recurrence of breast cancer using machine learning algorithms - multimedia tools and applications. SpringerLink. Retrieved April 26, 2023, from https://link.springer.com/article/10.1007/s11042-020-10448-w#:~:text=Machine%20learning%20algorithms%20help%20physicians,and%20molecular%20subtype%20%5B35%5D%