BREAST CANCER SURVIVAL PREDICTION FOR DECISION SUPPORT SYSTEM

```
#install.packages(c("survival", "survminer", "caret", "randomForest", "e1071")

library(tidyverse)
library(dplyr)
library(survival)
library(survminer)
library(caret)
library(randomForest)
library(e1071)
library(survival)
library(survminer)
```

DATA CLEANING

```
data <- read.csv("Breast Cancer METABRIC.csv")
head(data)
```

```
##   Patient.ID Age.at.Diagnosis Type.of.Breast.Surgery   Cancer.Type
## 1    MB-0000            75.65             Mastectomy Breast Cancer
## 2    MB-0002            43.19      Breast Conserving Breast Cancer
## 3    MB-0005            48.87             Mastectomy Breast Cancer
## 4    MB-0006            47.68             Mastectomy Breast Cancer
## 5    MB-0008            76.97             Mastectomy Breast Cancer
## 6    MB-0010            78.77             Mastectomy Breast Cancer
##                        Cancer.Type.Detailed Cellularity Chemotherapy
## 1          Breast Invasive Ductal Carcinoma                       No
## 2          Breast Invasive Ductal Carcinoma        High           No
## 3          Breast Invasive Ductal Carcinoma        High          Yes
## 4 Breast Mixed Ductal and Lobular Carcinoma    Moderate          Yes
## 5 Breast Mixed Ductal and Lobular Carcinoma        High          Yes
## 6          Breast Invasive Ductal Carcinoma    Moderate           No
##   Pam50...Claudin.low.subtype Cohort ER.status.measured.by.IHC ER.Status
```

```
## 1                claudin-low      1                  Positve  Positive
## 2                       LumA      1                  Positve  Positive
## 3                       LumB      1                  Positve  Positive
## 4                       LumB      1                  Positve  Positive
## 5                       LumB      1                  Positve  Positive
## 6                       LumB      1                  Positve  Positive
##   Neoplasm.Histologic.Grade HER2.status.measured.by.SNP6 HER2.Status
## 1                         3                    Neutral    Negative
## 2                         3                    Neutral    Negative
## 3                         2                    Neutral    Negative
## 4                         2                    Neutral    Negative
## 5                         3                    Neutral    Negative
## 6                         3                    Neutral    Negative
##   Tumor.Other.Histologic.Subtype Hormone.Therapy Inferred.Menopausal.State
## 1                     Ductal/NST             Yes                      Post
## 2                     Ductal/NST             Yes                       Pre
## 3                     Ductal/NST             Yes                       Pre
## 4                          Mixed             Yes                       Pre
## 5                          Mixed             Yes                      Post
## 6                     Ductal/NST             Yes                      Post
##   Integrative.Cluster Primary.Tumor.Laterality Lymph.nodes.examined.positiv
## 1               4ER+                    Right                             1
## 2               4ER+                    Right
## 3                  3                    Right
## 4                  9                    Right
## 5                  9                    Right
## 6                  7                     Left
##   Mutation.Count Nottingham.prognostic.index Oncotree.Code
## 1             NA                       6.044           IDC
## 2              2                       4.020           IDC
## 3              2                       4.030           IDC
## 4              1                       4.050          MDLC
## 5              2                       6.080          MDLC
## 6              4                       4.062           IDC
##   Overall.Survival..Months. Overall.Survival.Status PR.Status Radio.Therapy
## 1                 140.50000                  Living  Negative           Yes
## 2                  84.63333                  Living  Positive           Yes
## 3                 163.70000                Deceased  Positive            No
## 4                 164.93333                  Living  Positive           Yes
## 5                  41.36667                Deceased  Positive           Yes
## 6                   7.80000                Deceased  Positive           Yes
##   Relapse.Free.Status..Months. Relapse.Free.Status    Sex
## 1                       138.65        Not Recurred Female
## 2                        83.52        Not Recurred Female
## 3                       151.28            Recurred Female
## 4                       162.76        Not Recurred Female
## 5                        18.55            Recurred Female
## 6                         2.89            Recurred Female
##   X3.Gene.classifier.subtype Tumor.Size Tumor.Stage Patient.s.Vital.Status
```

```
## 1                  ER-/HER2-         22           2                Living
## 2      ER+/HER2- High Prolif         10           1                Living
## 3                                    15           2        Died of Disease
## 4                                    25           2                Living
## 5      ER+/HER2- High Prolif         40           2        Died of Disease
## 6      ER+/HER2- High Prolif         31           4        Died of Disease
```

```
colSums(is.na(data))
```

```
##                      Patient.ID              Age.at.Diagnosis
##                               0                            11
##         Type.of.Breast.Surgery                   Cancer.Type
##                               0                             0
##           Cancer.Type.Detailed                   Cellularity
##                               0                             0
##                   Chemotherapy    Pam50...Claudin.low.subtype
##                               0                             0
##                         Cohort      ER.status.measured.by.IHC
##                             11                              0
##                      ER.Status      Neoplasm.Histologic.Grade
##                               0                           121
##   HER2.status.measured.by.SNP6                   HER2.Status
##                               0                             0
## Tumor.Other.Histologic.Subtype               Hormone.Therapy
##                               0                             0
##      Inferred.Menopausal.State           Integrative.Cluster
##                               0                             0
##       Primary.Tumor.Laterality  Lymph.nodes.examined.positive
##                               0                           266
##                 Mutation.Count    Nottingham.prognostic.index
##                             152                           222
##                  Oncotree.Code      Overall.Survival..Months.
##                               0                           528
##        Overall.Survival.Status                      PR.Status
##                               0                             0
##                  Radio.Therapy   Relapse.Free.Status..Months.
##                               0                           121
##            Relapse.Free.Status                           Sex
##                               0                             0
##     X3.Gene.classifier.subtype                    Tumor.Size
##                               0                           149
##                    Tumor.Stage         Patient.s.Vital.Status
##                             721                             0
```

```r
colnames(data)
```

```
##  [1] "Patient.ID"                     "Age.at.Diagnosis"
##  [3] "Type.of.Breast.Surgery"         "Cancer.Type"
##  [5] "Cancer.Type.Detailed"           "Cellularity"
##  [7] "Chemotherapy"                   "Pam50...Claudin.low.subtype"
##  [9] "Cohort"                         "ER.status.measured.by.IHC"
## [11] "ER.Status"                      "Neoplasm.Histologic.Grade"
## [13] "HER2.status.measured.by.SNP6"   "HER2.Status"
## [15] "Tumor.Other.Histologic.Subtype" "Hormone.Therapy"
## [17] "Inferred.Menopausal.State"      "Integrative.Cluster"
## [19] "Primary.Tumor.Laterality"       "Lymph.nodes.examined.positive"
## [21] "Mutation.Count"                 "Nottingham.prognostic.index"
## [23] "Oncotree.Code"                  "Overall.Survival..Months."
## [25] "Overall.Survival.Status"        "PR.Status"
## [27] "Radio.Therapy"                  "Relapse.Free.Status..Months."
## [29] "Relapse.Free.Status"            "Sex"
## [31] "X3.Gene.classifier.subtype"     "Tumor.Size"
## [33] "Tumor.Stage"                    "Patient.s.Vital.Status"
```

```r
selected_cols <- c(
  "Patient.ID",
  "Age.at.Diagnosis", "Inferred.Menopausal.State",
  "Tumor.Size", "Tumor.Stage", "Neoplasm.Histologic.Grade",
  "Lymph.nodes.examined.positive", "Cancer.Type.Detailed",
  "Nottingham.prognostic.index",
  "Pam50...Claudin.low.subtype", "ER.Status", "PR.Status",
  "HER2.Status", "X3.Gene.classifier.subtype",
  "Chemotherapy", "Hormone.Therapy", "Radio.Therapy",
  "Overall.Survival..Months.", "Overall.Survival.Status",
  "Relapse.Free.Status", "Relapse.Free.Status..Months."
)
# Subset the data
metabric_sub <- data %>%
  select(all_of(selected_cols))

colSums(is.na(metabric_sub))
```

```
##                    Patient.ID              Age.at.Diagnosis
##                             0                            11
##     Inferred.Menopausal.State                    Tumor.Size
##                             0                           149
##                   Tumor.Stage     Neoplasm.Histologic.Grade
##                           721                           121
```

```
## Lymph.nodes.examined.positive          Cancer.Type.Detailed
##                            266                             0
##   Nottingham.prognostic.index   Pam50...Claudin.low.subtype
##                            222                             0
##                     ER.Status                    PR.Status
##                             0                             0
##                   HER2.Status    X3.Gene.classifier.subtype
##                             0                             0
##                  Chemotherapy              Hormone.Therapy
##                             0                             0
##                 Radio.Therapy     Overall.Survival..Months.
##                             0                           528
##       Overall.Survival.Status           Relapse.Free.Status
##                             0                             0
##  Relapse.Free.Status..Months.
##                           121
```

```
head(data)
```

```
##   Patient.ID Age.at.Diagnosis Type.of.Breast.Surgery   Cancer.Type
## 1    MB-0000            75.65             Mastectomy Breast Cancer
## 2    MB-0002            43.19      Breast Conserving Breast Cancer
## 3    MB-0005            48.87             Mastectomy Breast Cancer
## 4    MB-0006            47.68             Mastectomy Breast Cancer
## 5    MB-0008            76.97             Mastectomy Breast Cancer
## 6    MB-0010            78.77             Mastectomy Breast Cancer
##                        Cancer.Type.Detailed Cellularity Chemotherapy
## 1          Breast Invasive Ductal Carcinoma                       No
## 2          Breast Invasive Ductal Carcinoma        High           No
## 3          Breast Invasive Ductal Carcinoma        High          Yes
## 4 Breast Mixed Ductal and Lobular Carcinoma    Moderate          Yes
## 5 Breast Mixed Ductal and Lobular Carcinoma        High          Yes
## 6          Breast Invasive Ductal Carcinoma    Moderate           No
##   Pam50...Claudin.low.subtype Cohort ER.status.measured.by.IHC ER.Status
## 1                claudin-low      1                   Positve  Positive
## 2                       LumA      1                   Positve  Positive
## 3                       LumB      1                   Positve  Positive
## 4                       LumB      1                   Positve  Positive
## 5                       LumB      1                   Positve  Positive
## 6                       LumB      1                   Positve  Positive
##   Neoplasm.Histologic.Grade HER2.status.measured.by.SNP6 HER2.Status
## 1                         3                      Neutral    Negative
## 2                         3                      Neutral    Negative
## 3                         2                      Neutral    Negative
## 4                         2                      Neutral    Negative
## 5                         3                      Neutral    Negative
```

```
## 6                        3                     Neutral    Negative
##   Tumor.Other.Histologic.Subtype Hormone.Therapy Inferred.Menopausal.State
## 1                     Ductal/NST             Yes                      Post
## 2                     Ductal/NST             Yes                       Pre
## 3                     Ductal/NST             Yes                       Pre
## 4                          Mixed             Yes                       Pre
## 5                          Mixed             Yes                      Post
## 6                     Ductal/NST             Yes                      Post
##   Integrative.Cluster Primary.Tumor.Laterality Lymph.nodes.examined.positiv
## 1               4ER+                    Right                             1
## 2               4ER+                    Right
## 3                  3                    Right
## 4                  9                    Right
## 5                  9                    Right
## 6                  7                     Left
##   Mutation.Count Nottingham.prognostic.index Oncotree.Code
## 1             NA                       6.044           IDC
## 2              2                       4.020           IDC
## 3              2                       4.030           IDC
## 4              1                       4.050          MDLC
## 5              2                       6.080          MDLC
## 6              4                       4.062           IDC
##   Overall.Survival..Months. Overall.Survival.Status PR.Status Radio.Therapy
## 1                 140.50000                  Living  Negative           Yes
## 2                  84.63333                  Living  Positive           Yes
## 3                 163.70000                Deceased  Positive            No
## 4                 164.93333                  Living  Positive           Yes
## 5                  41.36667                Deceased  Positive           Yes
## 6                   7.80000                Deceased  Positive           Yes
##   Relapse.Free.Status..Months. Relapse.Free.Status    Sex
## 1                       138.65       Not Recurred Female
## 2                        83.52       Not Recurred Female
## 3                       151.28           Recurred Female
## 4                       162.76       Not Recurred Female
## 5                        18.55           Recurred Female
## 6                         2.89           Recurred Female
##   X3.Gene.classifier.subtype Tumor.Size Tumor.Stage Patient.s.Vital.Status
## 1                ER-/HER2-         22           2                Living
## 2     ER+/HER2- High Prolif         10           1                Living
## 3                                   15           2       Died of Disease
## 4                                   25           2                Living
## 5     ER+/HER2- High Prolif         40           2       Died of Disease
## 6     ER+/HER2- High Prolif         31           4       Died of Disease
```

Handling Missing Values

```r
# Drop only the rows where Overall.Survival.Months_ is NA
# metabric_sub$Age.at.Diagnosis[is.na(metabric_sub$Age.at.Diagnosis)] <- media
# metabric_sub$Tumor.Size[is.na(metabric_sub$Tumor.Size)] <- median(metabric_s
# metabric_sub$Tumor.Stage[is.na(metabric_sub$Tumor.Stage)] <- "Unknown"
# mode_grade <- names(sort(table(metabric_sub$Neoplasm.Histologic.Grade), decr
# metabric_sub$Neoplasm.Histologic.Grade[is.na(metabric_sub$Neoplasm.Histologi
# metabric_sub$Lymph.nodes.examined.positive[is.na(metabric_sub$Lymph.nodes.ex
# metabric_sub$Nottingham.prognostic.index[is.na(metabric_sub$Nottingham.progr


mode_value <- "ER+/HER2- Low Prolif"

# Replace blank spaces or NAs in 'X3.Gene.classifier.subtype' with the mode va
metabric_sub$X3.Gene.classifier.subtype[metabric_sub$X3.Gene.classifier.subtyp

# Check if replacement was successful
table(metabric_sub$X3.Gene.classifier.subtype)
```

```
##
##             ER-/HER2- ER+/HER2- High Prolif  ER+/HER2- Low Prolif
##                   309                   617                  1385
##                 HER2+
##                   198
```

```r
metabric_sub <- na.omit(metabric_sub)

colSums(is.na(metabric_sub))
```

```
##                    Patient.ID              Age.at.Diagnosis
##                             0                             0
##     Inferred.Menopausal.State                    Tumor.Size
##                             0                             0
##                   Tumor.Stage     Neoplasm.Histologic.Grade
##                             0                             0
## Lymph.nodes.examined.positive          Cancer.Type.Detailed
##                             0                             0
##   Nottingham.prognostic.index   Pam50...Claudin.low.subtype
##                             0                             0
##                     ER.Status                     PR.Status
##                             0                             0
##                   HER2.Status    X3.Gene.classifier.subtype
##                             0                             0
##                  Chemotherapy               Hormone.Therapy
```

```
##                             0                             0
##                Radio.Therapy     Overall.Survival..Months.
##                             0                             0
##       Overall.Survival.Status           Relapse.Free.Status
##                             0                             0
##  Relapse.Free.Status..Months.
##                             0
```

```
dim(metabric_sub)
```

```
## [1] 1354   21
```

```
# Clean column names for easier reference
colnames(metabric_sub) <- gsub("[\\.]+", "_", colnames(metabric_sub))
# View(metabric_sub)
head(metabric_sub)
```

```
##   Patient_ID Age_at_Diagnosis Inferred_Menopausal_State Tumor_Size Tumor_St
## 1    MB-0000            75.65                      Post         22
## 2    MB-0002            43.19                       Pre         10
## 3    MB-0005            48.87                       Pre         15
## 4    MB-0006            47.68                       Pre         25
## 5    MB-0008            76.97                      Post         40
## 6    MB-0010            78.77                      Post         31
##   Neoplasm_Histologic_Grade Lymph_nodes_examined_positive
## 1                         3                            10
## 2                         3                             0
## 3                         2                             1
## 4                         2                             3
## 5                         3                             8
## 6                         3                             0
##                       Cancer_Type_Detailed Nottingham_prognostic_index
## 1          Breast Invasive Ductal Carcinoma                       6.044
## 2          Breast Invasive Ductal Carcinoma                       4.020
## 3          Breast Invasive Ductal Carcinoma                       4.030
## 4 Breast Mixed Ductal and Lobular Carcinoma                       4.050
## 5 Breast Mixed Ductal and Lobular Carcinoma                       6.080
## 6          Breast Invasive Ductal Carcinoma                       4.062
##   Pam50_Claudin_low_subtype ER_Status PR_Status HER2_Status
## 1               claudin-low  Positive  Negative    Negative
## 2                      LumA  Positive  Positive    Negative
## 3                      LumB  Positive  Positive    Negative
## 4                      LumB  Positive  Positive    Negative
## 5                      LumB  Positive  Positive    Negative
```

```
## 6                     LumB  Positive  Positive    Negative
##   X3_Gene_classifier_subtype Chemotherapy Hormone_Therapy Radio_Therapy
## 1                  ER-/HER2-           No             Yes           Yes
## 2      ER+/HER2- High Prolif           No             Yes           Yes
## 3       ER+/HER2- Low Prolif          Yes             Yes            No
## 4       ER+/HER2- Low Prolif          Yes             Yes           Yes
## 5      ER+/HER2- High Prolif          Yes             Yes           Yes
## 6      ER+/HER2- High Prolif           No             Yes           Yes
##   Overall_Survival_Months_ Overall_Survival_Status Relapse_Free_Status
## 1                140.50000                  Living        Not Recurred
## 2                 84.63333                  Living        Not Recurred
## 3                163.70000                Deceased            Recurred
## 4                164.93333                  Living        Not Recurred
## 5                 41.36667                Deceased            Recurred
## 6                  7.80000                Deceased            Recurred
##   Relapse_Free_Status_Months_
## 1                      138.65
## 2                       83.52
## 3                      151.28
## 4                      162.76
## 5                       18.55
## 6                        2.89
```

Encode Categorical Variables with Clinical Coding Where Applicable

We'll encode key variables based on clinical standards, e.g.,: Tumor Grade (Histologic): Convert to ordinal (1, 2, 3) Tumor Stage: Standard TNM categories (Stage I, II, III, IV) ER/PR/HER2 Status: Binary (Positive = 1, Negative = 0) Treatment flags: Binary (Yes = 1, No = 0)

```r
library(dplyr)

# Remove rows with missing values
metabric_clean <- na.omit(metabric_sub)

# Binary encoding for Overall Survival
metabric_clean$Surv_Status <- ifelse(metabric_clean$Overall_Survival_Status ==

# Binary encoding for Relapse Status (Recurred = 1, Not Recurred = 0)
metabric_clean$Relapse_Status <- ifelse(metabric_clean$Relapse_Free_Status ==

# Rename Relapse-Free Survival Months
metabric_clean$Relapse_Months <- metabric_clean$Relapse_Free_Status_Months_
```

```r
# Encode Tumor Stage as ordered factor
metabric_clean$Tumor_Stage <- factor(metabric_clean$Tumor_Stage,
                                     levels = c(1, 2, 3, 4),
                                     labels = c("Stage I", "Stage II", "Stage

# Treat 0 stage as missing if applicable
metabric_clean$Tumor_Stage[metabric_clean$Tumor_Stage == "0"] <- NA

# Encode ER/PR/HER2 status
metabric_clean$ER <- ifelse(metabric_clean$ER_Status == "Positive", 1, 0)
metabric_clean$PR <- ifelse(metabric_clean$PR_Status == "Positive", 1, 0)
metabric_clean$HER2 <- ifelse(metabric_clean$HER2_Status == "Positive", 1, 0)

# Encode Menopausal State
metabric_clean$Menopause <- ifelse(metabric_clean$Inferred_Menopausal_State ==

# Encode therapies
metabric_clean$Chemo <- ifelse(metabric_clean$Chemotherapy == "Yes", 1, 0)
metabric_clean$Hormone <- ifelse(metabric_clean$Hormone_Therapy == "Yes", 1, 0
metabric_clean$Radio <- ifelse(metabric_clean$Radio_Therapy == "Yes", 1, 0)

# Factorize Cancer Type
metabric_clean$Cancer_Type_Detailed <- factor(metabric_clean$Cancer_Type_Detai
                                              "Breast Invasive Lobular C
                                              "Breast Mixed Ductal and L

# Optional: Check the mapping
table(metabric_clean$Cancer_Type_Detailed)
```

```
##
##                                  Breast
##                                       7
##          Breast Invasive Ductal Carcinoma
##                                    1064
##         Breast Invasive Lobular Carcinoma
##                                      93
##  Breast Invasive Mixed Mucinous Carcinoma
##                                      17
## Breast Mixed Ductal and Lobular Carcinoma
##                                     165
##                 Invasive Breast Carcinoma
##                                       8
##                                  Others
##                                       0
```

```r
# Drop NA rows (after transformations)
metabric_clean <- na.omit(metabric_clean)
```

```r
# Final model-ready dataset including relapse variables
model_data <- metabric_clean %>%
  select(Surv_Status, Overall_Survival_Months_, Relapse_Status, Relapse_Months
         Age_at_Diagnosis, Tumor_Size, Cancer_Type_Detailed,
         Lymph_nodes_examined_positive, Nottingham_prognostic_index, Tumor_Sta
         Neoplasm_Histologic_Grade, ER, PR, HER2, Chemo, Hormone, Radio, Menop

# Preview data
head(model_data)
```

```
##   Surv_Status Overall_Survival_Months_ Relapse_Status Relapse_Months
## 1           0                140.50000              0         138.65
## 2           0                 84.63333              0          83.52
## 3           1                163.70000              1         151.28
## 4           0                164.93333              0         162.76
## 5           1                 41.36667              1          18.55
## 6           1                  7.80000              1           2.89
##   Age_at_Diagnosis Tumor_Size                      Cancer_Type_Detailed
## 1            75.65         22          Breast Invasive Ductal Carcinoma
## 2            43.19         10          Breast Invasive Ductal Carcinoma
## 3            48.87         15          Breast Invasive Ductal Carcinoma
## 4            47.68         25 Breast Mixed Ductal and Lobular Carcinoma
## 5            76.97         40 Breast Mixed Ductal and Lobular Carcinoma
## 6            78.77         31          Breast Invasive Ductal Carcinoma
##   Lymph_nodes_examined_positive Nottingham_prognostic_index Tumor_Stage
## 1                            10                       6.044    Stage II
## 2                             0                       4.020     Stage I
## 3                             1                       4.030    Stage II
## 4                             3                       4.050    Stage II
## 5                             8                       6.080    Stage II
## 6                             0                       4.062    Stage IV
##   Neoplasm_Histologic_Grade ER PR HER2 Chemo Hormone Radio Menopause
## 1                         3  1  0    0     0       1     1         1
## 2                         3  1  1    0     0       1     1         0
## 3                         2  1  1    0     1       1     0         0
## 4                         2  1  1    0     1       1     1         0
## 5                         3  1  1    0     1       1     1         1
## 6                         3  1  1    0     0       1     1         1
```

```r
colSums(is.na(model_data))
```

```
##                   Surv_Status      Overall_Survival_Months_
##                             0                             0
##               Relapse_Status                Relapse_Months
##                             0                             0
```

```
##              Age_at_Diagnosis                     Tumor_Size
##                             0                              0
##          Cancer_Type_Detailed Lymph_nodes_examined_positive
##                             0                              0
##   Nottingham_prognostic_index                   Tumor_Stage
##                             0                              0
##     Neoplasm_Histologic_Grade                             ER
##                             0                              0
##                            PR                           HER2
##                             0                              0
##                         Chemo                        Hormone
##                             0                              0
##                         Radio                     Menopause
##                             0                              0
```

```
write.csv(model_data, 'model_data.csv')
```

DATA MODELLING FOR SURVIVAL ANALYSIS

```
model_data$Tumor_Stage <- factor(model_data$Tumor_Stage,
                                 levels = c("Stage I", "Stage II", "Stage III"

model_data$ER <- factor(model_data$ER, levels = c(0, 1))  # Estrogen Receptor
model_data$PR <- factor(model_data$PR, levels = c(0, 1))  # Progesterone Recep
model_data$HER2 <- factor(model_data$HER2, levels = c(0, 1))  # HER2 status
model_data$Chemo <- factor(model_data$Chemo, levels = c(0, 1))  # Chemotherapy
model_data$Hormone <- factor(model_data$Hormone, levels = c(0, 1))  # Hormone
model_data$Radio <- factor(model_data$Radio, levels = c(0, 1))  # Radiotherapy
model_data$Menopause <- factor(model_data$Menopause, levels = c(0, 1))  # Menc
# Convert Cancer Type to a factor
model_data$Cancer_Type_Detailed <- factor(model_data$Cancer_Type_Detailed)
model_data$Neoplasm_Histologic_Grade <- factor(model_data$Neoplasm_Histologic_
model_data$Cancer_Type_Detailed <- factor(model_data$Cancer_Type_Detailed,
                                          levels = c("Breast", "Breast Invasive
                                                     "Breast Invasive Lobular C
                                                     "Breast Mixed Ductal and L

# Ensure Overall_Survival_Months_ is numeric
model_data$Overall_Survival_Months_ <- as.numeric(model_data$Overall_Survival_
```

```
head(model_data)
```

```
##   Surv_Status Overall_Survival_Months_ Relapse_Status Relapse_Months
## 1           0                140.50000              0         138.65
## 2           0                 84.63333              0          83.52
## 3           1                163.70000              1         151.28
## 4           0                164.93333              0         162.76
## 5           1                 41.36667              1          18.55
## 6           1                  7.80000              1           2.89
##   Age_at_Diagnosis Tumor_Size                      Cancer_Type_Detailed
## 1            75.65         22          Breast Invasive Ductal Carcinoma
## 2            43.19         10          Breast Invasive Ductal Carcinoma
## 3            48.87         15          Breast Invasive Ductal Carcinoma
## 4            47.68         25 Breast Mixed Ductal and Lobular Carcinoma
## 5            76.97         40 Breast Mixed Ductal and Lobular Carcinoma
## 6            78.77         31          Breast Invasive Ductal Carcinoma
##   Lymph_nodes_examined_positive Nottingham_prognostic_index Tumor_Stage
## 1                            10                       6.044    Stage II
## 2                             0                       4.020     Stage I
## 3                             1                       4.030    Stage II
## 4                             3                       4.050    Stage II
## 5                             8                       6.080    Stage II
## 6                             0                       4.062    Stage IV
##   Neoplasm_Histologic_Grade ER PR HER2 Chemo Hormone Radio Menopause
## 1                         3  1  0    0     0       1     1         1
## 2                         3  1  1    0     0       1     1         0
## 3                         2  1  1    0     1       1     0         0
## 4                         2  1  1    0     1       1     1         0
## 5                         3  1  1    0     1       1     1         1
## 6                         3  1  1    0     0       1     1         1
```

DATA PREPROCESSING AND SPLITTING

```
# Prepare the data for binary classification (Survival Status)
model_data$Surv_Status <- factor(model_data$Surv_Status, levels = c(0, 1))  #
model_data$Relapse_Status <- factor(model_data$Relapse_Status, levels = c(0, 1


# Separate predictors and target variable
sm_train <- model_data$Overall_Survival_Months_
rm_train <- model_data$Relapse_Months
x_train_m <- model_data[, setdiff(names(model_data), c("Surv_Status", "Overall

sm_test <- model_data$Overall_Survival_Months_
```

```
rm_test <- model_data$Relapse_Months
x_test_m <- model_data[, setdiff(names(model_data), c("Surv_Status", "Overall_



# Split data into training and testing sets (80% training, 20% testing)
set.seed(123)
train_index <- createDataPartition(model_data$Surv_Status, p = 0.8, list = FAL
train_data <- model_data[train_index, ]
test_data <- model_data[-train_index, ]

# Convert Surv_Status to factor for classification
train_data$Surv_Status <- as.factor(train_data$Surv_Status)
test_data$Surv_Status <- as.factor(test_data$Surv_Status)

# Separate predictors and target variable
ss_train <- train_data$Surv_Status
#sm_train <- train_data$Overall_Survival_Months_
rs_train <- train_data$Relapse_Status
#rm_train <- train_data$Relapse_Months
x_train <- train_data[, setdiff(names(train_data), c("Surv_Status", "Overall_S

ss_test <- test_data$Surv_Status
#sm_test <- test_data$Overall_Survival_Months_
rs_test <- test_data$Relapse_Status
#rm_test <- test_data$Relapse_Months
x_test <- test_data[, setdiff(names(test_data), c("Surv_Status", "Overall_Surv
```

Scaling for models like XGboost and LR

Had to exclude this, it does not impove any of the models.

```
# library(caret)
#
# # 1. Identify predictors
# predictors <- x_train
#
# # 2. Create preprocessing recipe (centering and scaling numeric variables)
# preprocess_model <- preProcess(predictors, method = c("center", "scale"))
#
# # 3. Apply preprocessing to train and test sets
# x_train_scaled <- predict(preprocess_model, newdata = x_train)
# x_test_scaled <- predict(preprocess_model, newdata = x_test)
#
```

```
# # 4. Save the preprocessor for future use
# saveRDS(preprocess_model, file = "preprocess_model.rds")
```

TRAINING MODELS

OVERALL SURVIVAL STATUS

```
ss_log_reg_model <- glm(ss_train ~ .,
                        data = x_train, family = "binomial")

# Summary of the model
summary(ss_log_reg_model)
```

```
##
## Call:
## glm(formula = ss_train ~ ., family = "binomial", data = x_train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4809  -0.9615   0.4235   0.9266   1.8945
##
## Coefficients:
##                                                                 Estimate
## (Intercept)                                                  -20.408611
## Age_at_Diagnosis                                               0.079169
## Tumor_Size                                                     0.024230
## Cancer_Type_DetailedBreast Invasive Ductal Carcinoma          14.932043
## Cancer_Type_DetailedBreast Invasive Lobular Carcinoma         14.592510
## Cancer_Type_DetailedBreast Invasive Mixed Mucinous Carcinoma  14.589481
## Cancer_Type_DetailedBreast Mixed Ductal and Lobular Carcinoma 15.239360
## Cancer_Type_DetailedInvasive Breast Carcinoma                 15.373967
## Lymph_nodes_examined_positive                                  0.110490
## Nottingham_prognostic_index                                    0.332174
## Tumor_StageStage II                                            0.215559
## Tumor_StageStage III                                          -0.220816
## Tumor_StageStage IV                                           14.470412
## Neoplasm_Histologic_Grade2                                     0.141109
## Neoplasm_Histologic_Grade3                                     0.019854
## ER1                                                            0.066105
## PR1                                                            0.015285
## HER21                                                          0.683128
## Chemo1                                                        -0.469606
```

```
## Hormone1                                                          -0.771407
## Radio1                                                            -0.567175
## Menopause1                                                        -0.681276
##                                                                  Std. Error
## (Intercept)                                                      705.779969
## Age_at_Diagnosis                                                   0.009304
## Tumor_Size                                                         0.006813
## Cancer_Type_DetailedBreast Invasive Ductal Carcinoma             705.779667
## Cancer_Type_DetailedBreast Invasive Lobular Carcinoma            705.779716
## Cancer_Type_DetailedBreast Invasive Mixed Mucinous Carcinoma     705.779929
## Cancer_Type_DetailedBreast Mixed Ductal and Lobular Carcinoma    705.779690
## Cancer_Type_DetailedInvasive Breast Carcinoma                    705.780124
## Lymph_nodes_examined_positive                                      0.046081
## Nottingham_prognostic_index                                        0.207030
## Tumor_StageStage II                                                0.205616
## Tumor_StageStage III                                               0.415745
## Tumor_StageStage IV                                              452.449506
## Neoplasm_Histologic_Grade2                                         0.339355
## Neoplasm_Histologic_Grade3                                         0.491100
## ER1                                                                0.230984
## PR1                                                                0.168019
## HER21                                                              0.237365
## Chemo1                                                             0.236604
## Hormone1                                                           0.174045
## Radio1                                                             0.156087
## Menopause1                                                         0.248355
##                                                                  z value Pr(>|
## (Intercept)                                                       -0.029 0.976
## Age_at_Diagnosis                                                   8.509  < 2e
## Tumor_Size                                                         3.557 0.000
## Cancer_Type_DetailedBreast Invasive Ductal Carcinoma               0.021 0.983
## Cancer_Type_DetailedBreast Invasive Lobular Carcinoma              0.021 0.983
## Cancer_Type_DetailedBreast Invasive Mixed Mucinous Carcinoma       0.021 0.983
## Cancer_Type_DetailedBreast Mixed Ductal and Lobular Carcinoma      0.022 0.982
## Cancer_Type_DetailedInvasive Breast Carcinoma                      0.022 0.982
## Lymph_nodes_examined_positive                                      2.398 0.016
## Nottingham_prognostic_index                                        1.604 0.108
## Tumor_StageStage II                                                1.048 0.294
## Tumor_StageStage III                                              -0.531 0.595
## Tumor_StageStage IV                                                0.032 0.974
## Neoplasm_Histologic_Grade2                                         0.416 0.677
## Neoplasm_Histologic_Grade3                                         0.040 0.967
## ER1                                                                0.286 0.774
## PR1                                                                0.091 0.927
## HER21                                                              2.878 0.004
## Chemo1                                                            -1.985 0.047
## Hormone1                                                          -4.432 9.33e
## Radio1                                                            -3.634 0.000
## Menopause1                                                        -2.743 0.006
```

```
##
## (Intercept)
## Age_at_Diagnosis                                                  ***
## Tumor_Size                                                        ***
## Cancer_Type_DetailedBreast Invasive Ductal Carcinoma
## Cancer_Type_DetailedBreast Invasive Lobular Carcinoma
## Cancer_Type_DetailedBreast Invasive Mixed Mucinous Carcinoma
## Cancer_Type_DetailedBreast Mixed Ductal and Lobular Carcinoma
## Cancer_Type_DetailedInvasive Breast Carcinoma
## Lymph_nodes_examined_positive                                     *
## Nottingham_prognostic_index
## Tumor_StageStage II
## Tumor_StageStage III
## Tumor_StageStage IV
## Neoplasm_Histologic_Grade2
## Neoplasm_Histologic_Grade3
## ER1
## PR1
## HER21                                                             **
## Chemo1                                                            *
## Hormone1                                                          ***
## Radio1                                                            ***
## Menopause1                                                        **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1486.0  on 1082  degrees of freedom
## Residual deviance: 1219.7  on 1061  degrees of freedom
## AIC: 1263.7
##
## Number of Fisher Scoring iterations: 14
```

```r
# Predict on test data
ss_predictions_status <- predict(ss_log_reg_model, newdata = x_test, type = "r
ss_pred_class_status <- ifelse(ss_predictions_status > 0.5, 1, 0)  # 1 = Dead,

# Confusion Matrix to evaluate the performance
ss_conf_matrix_status <- confusionMatrix(factor(ss_pred_class_status), factor(
print(ss_conf_matrix_status)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
```

```
##          0  66  49
##          1  53 102
##
##                Accuracy : 0.6222
##                  95% CI : (0.5615, 0.6803)
##     No Information Rate : 0.5593
##     P-Value [Acc > NIR] : 0.02107
##
##                   Kappa : 0.2309
##
##  Mcnemar's Test P-Value : 0.76643
##
##             Sensitivity : 0.5546
##             Specificity : 0.6755
##          Pos Pred Value : 0.5739
##          Neg Pred Value : 0.6581
##              Prevalence : 0.4407
##          Detection Rate : 0.2444
##    Detection Prevalence : 0.4259
##       Balanced Accuracy : 0.6151
##
##        'Positive' Class : 0
##
```

Random Forest

```r
library(randomForest)

set.seed(123)
ss_rf_model <- randomForest(ss_train ~ Age_at_Diagnosis + Tumor_Size + Cancer_
                        Lymph_nodes_examined_positive + Nottingham_prognostic
                        Tumor_Stage + Neoplasm_Histologic_Grade + ER + PR + H
                        Chemo + Hormone + Radio + Menopause,
                        data = x_train, ntree = 500, importance = TRUE)


# Predict and evaluate
ss_rf_prop <- predict(ss_rf_model, newdata = x_test, type = "prob")
# Pick the class (column name) with the highest probability for each observati
ss_rf_pred <- colnames(ss_rf_prop)[max.col(ss_rf_prop, ties.method = "first")]
ss_rf_pred <- factor(ss_rf_pred, levels = c("0", "1"))



confusionMatrix(ss_rf_pred, factor(ss_test))
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0  68  46
##          1  51 105
##
##                Accuracy : 0.6407
##                  95% CI : (0.5804, 0.698)
##     No Information Rate : 0.5593
##     P-Value [Acc > NIR] : 0.003963
##
##                   Kappa : 0.268
##
##  Mcnemar's Test P-Value : 0.684641
##
##             Sensitivity : 0.5714
##             Specificity : 0.6954
##          Pos Pred Value : 0.5965
##          Neg Pred Value : 0.6731
##              Prevalence : 0.4407
##          Detection Rate : 0.2519
##    Detection Prevalence : 0.4222
##       Balanced Accuracy : 0.6334
##
##        'Positive' Class : 0
##
```

```
library(glmnet)
```

```
## Loading required package: Matrix

##
## Attaching package: 'Matrix'

## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack

## Loaded glmnet 4.1-8
```

```
# Create matrices
x_train_matrix <- model.matrix(~ . -1, data = x_train)
x_test_matrix <- model.matrix(~ . -1, data = x_test)
```

```r
# Fit model (alpha=0.5 for Elastic Net)
ss_cv_model <- cv.glmnet(x_train_matrix, ss_train, alpha = 0.5, family = "binc

# Predict
ss_enet_pred <- predict(ss_cv_model, newx = x_test_matrix, s = "lambda.min", t
ss_enet_pred_class <- ifelse(ss_enet_pred > 0.5, 1, 0)
confusionMatrix(as.factor(ss_enet_pred_class), as.factor(ss_test))
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0  65  46
##          1  54 105
##
##                Accuracy : 0.6296
##                  95% CI : (0.569, 0.6874)
##     No Information Rate : 0.5593
##     P-Value [Acc > NIR] : 0.01128
##
##                   Kappa : 0.2433
##
##  Mcnemar's Test P-Value : 0.48393
##
##             Sensitivity : 0.5462
##             Specificity : 0.6954
##          Pos Pred Value : 0.5856
##          Neg Pred Value : 0.6604
##              Prevalence : 0.4407
##          Detection Rate : 0.2407
##    Detection Prevalence : 0.4111
##       Balanced Accuracy : 0.6208
##
##        'Positive' Class : 0
##
```

```r
library(xgboost)
```

```
##
## Attaching package: 'xgboost'

## The following object is masked from 'package:dplyr':
##
##     slice
```

```
ss_xgb_model <- xgboost(data = x_train_matrix, label = as.numeric(ss_train) -
                    objective = "binary:logistic", nrounds = 100, max.depth =
```

```
## [1]  train-logloss:0.664455
## [2]  train-logloss:0.640821
## [3]  train-logloss:0.619684
## [4]  train-logloss:0.600481
## [5]  train-logloss:0.583174
## [6]  train-logloss:0.567996
## [7]  train-logloss:0.555717
## [8]  train-logloss:0.544071
## [9]  train-logloss:0.533561
## [10] train-logloss:0.523305
## [11] train-logloss:0.511615
## [12] train-logloss:0.503242
## [13] train-logloss:0.495808
## [14] train-logloss:0.487474
## [15] train-logloss:0.480598
## [16] train-logloss:0.473563
## [17] train-logloss:0.468656
## [18] train-logloss:0.463318
## [19] train-logloss:0.458227
## [20] train-logloss:0.453075
## [21] train-logloss:0.447734
## [22] train-logloss:0.443354
## [23] train-logloss:0.439020
## [24] train-logloss:0.436200
## [25] train-logloss:0.432216
## [26] train-logloss:0.429768
## [27] train-logloss:0.426955
## [28] train-logloss:0.425001
## [29] train-logloss:0.423215
## [30] train-logloss:0.421045
## [31] train-logloss:0.418977
## [32] train-logloss:0.413817
## [33] train-logloss:0.411056
## [34] train-logloss:0.408501
## [35] train-logloss:0.406366
## [36] train-logloss:0.403823
## [37] train-logloss:0.402026
## [38] train-logloss:0.399203
## [39] train-logloss:0.394680
## [40] train-logloss:0.392309
## [41] train-logloss:0.390500
## [42] train-logloss:0.388749
## [43] train-logloss:0.386386
```

```
## [44] train-logloss:0.385245
## [45] train-logloss:0.384234
## [46] train-logloss:0.380597
## [47] train-logloss:0.379226
## [48] train-logloss:0.376557
## [49] train-logloss:0.375986
## [50] train-logloss:0.372361
## [51] train-logloss:0.370947
## [52] train-logloss:0.368468
## [53] train-logloss:0.365958
## [54] train-logloss:0.364999
## [55] train-logloss:0.364151
## [56] train-logloss:0.363678
## [57] train-logloss:0.362360
## [58] train-logloss:0.361500
## [59] train-logloss:0.359207
## [60] train-logloss:0.357795
## [61] train-logloss:0.354989
## [62] train-logloss:0.353136
## [63] train-logloss:0.352345
## [64] train-logloss:0.350728
## [65] train-logloss:0.350231
## [66] train-logloss:0.347474
## [67] train-logloss:0.345888
## [68] train-logloss:0.344768
## [69] train-logloss:0.343065
## [70] train-logloss:0.342461
## [71] train-logloss:0.342151
## [72] train-logloss:0.339947
## [73] train-logloss:0.337364
## [74] train-logloss:0.336486
## [75] train-logloss:0.333977
## [76] train-logloss:0.331237
## [77] train-logloss:0.330783
## [78] train-logloss:0.329187
## [79] train-logloss:0.328309
## [80] train-logloss:0.328047
## [81] train-logloss:0.327546
## [82] train-logloss:0.324814
## [83] train-logloss:0.322472
## [84] train-logloss:0.320116
## [85] train-logloss:0.318995
## [86] train-logloss:0.318617
## [87] train-logloss:0.316268
## [88] train-logloss:0.314195
## [89] train-logloss:0.313983
## [90] train-logloss:0.313296
## [91] train-logloss:0.311388
## [92] train-logloss:0.309420
```

```
## [93] train-logloss:0.307886
## [94] train-logloss:0.306693
## [95] train-logloss:0.304707
## [96] train-logloss:0.302642
## [97] train-logloss:0.302180
## [98] train-logloss:0.301986
## [99] train-logloss:0.300879
## [100]    train-logloss:0.299357
```

```
# Prediction
ss_xgb_pred <- predict(ss_xgb_model, newdata = x_test_matrix)
ss_xgb_pred_class <- ifelse(ss_xgb_pred > 0.5, 1, 0)
confusionMatrix(as.factor(ss_xgb_pred_class), ss_test)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0  74  44
##          1  45 107
##
##                Accuracy : 0.6704
##                  95% CI : (0.6108, 0.7261)
##     No Information Rate : 0.5593
##     P-Value [Acc > NIR] : 0.0001258
##
##                   Kappa : 0.3308
##
##  Mcnemar's Test P-Value : 1.0000000
##
##             Sensitivity : 0.6218
##             Specificity : 0.7086
##          Pos Pred Value : 0.6271
##          Neg Pred Value : 0.7039
##              Prevalence : 0.4407
##          Detection Rate : 0.2741
##    Detection Prevalence : 0.4370
##       Balanced Accuracy : 0.6652
##
##        'Positive' Class : 0
##
```

OVERALL SURVIVAL MONTHS

```r
# Train a Linear Regression model for Overall Survival Months
sm_linear_reg_model <- lm(sm_train ~ .,
                      data = x_train_m)

# Summary of the model
summary(sm_linear_reg_model)
```

```
##
## Call:
## lm(formula = sm_train ~ ., data = x_train_m)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -163.649  -55.611   -6.093   53.334  203.023
##
## Coefficients:
##                                                                  Estimate
## (Intercept)                                                    260.410319
## Age_at_Diagnosis                                                -1.651061
## Tumor_Size                                                      -0.435050
## Cancer_Type_DetailedBreast Invasive Ductal Carcinoma           -26.943848
## Cancer_Type_DetailedBreast Invasive Lobular Carcinoma          -18.983588
## Cancer_Type_DetailedBreast Invasive Mixed Mucinous Carcinoma   -31.020236
## Cancer_Type_DetailedBreast Mixed Ductal and Lobular Carcinoma -22.738829
## Cancer_Type_DetailedInvasive Breast Carcinoma                  -46.246566
## Lymph_nodes_examined_positive                                   -2.538651
## Nottingham_prognostic_index                                      0.003554
## Tumor_StageStage II                                            -11.371732
## Tumor_StageStage III                                           -17.913580
## Tumor_StageStage IV                                            -50.934796
## Neoplasm_Histologic_Grade2                                       2.109243
## Neoplasm_Histologic_Grade3                                       1.904268
## ER1                                                             10.704004
## PR1                                                              6.518525
## HER21                                                          -19.015201
## Chemo1                                                         -23.572374
## Hormone1                                                       -12.459468
## Radio1                                                           4.876476
## Menopause1                                                      19.725234
##                                                                Std. Error
## (Intercept)                                                     31.783408
## Age_at_Diagnosis                                                 0.243179
## Tumor_Size                                                       0.161036
## Cancer_Type_DetailedBreast Invasive Ductal Carcinoma            27.363239
## Cancer_Type_DetailedBreast Invasive Lobular Carcinoma           28.317864
## Cancer_Type_DetailedBreast Invasive Mixed Mucinous Carcinoma    32.443583
## Cancer_Type_DetailedBreast Mixed Ductal and Lobular Carcinoma   27.823932
```

```
## Cancer_Type_DetailedInvasive Breast Carcinoma                  37.313772
## Lymph_nodes_examined_positive                                   0.765118
## Nottingham_prognostic_index                                     4.902119
## Tumor_StageStage II                                             5.684033
## Tumor_StageStage III                                           10.902455
## Tumor_StageStage IV                                            24.899308
## Neoplasm_Histologic_Grade2                                      8.948385
## Neoplasm_Histologic_Grade3                                     12.390692
## ER1                                                             6.551385
## PR1                                                             4.734642
## HER21                                                           6.322885
## Chemo1                                                          6.509047
## Hormone1                                                        4.851625
## Radio1                                                          4.356467
## Menopause1                                                      7.098695
##                                                               t value Pr(>|
## (Intercept)                                                     8.193 5.91e
## Age_at_Diagnosis                                               -6.789 1.69e
## Tumor_Size                                                     -2.702 0.006
## Cancer_Type_DetailedBreast Invasive Ductal Carcinoma           -0.985 0.324
## Cancer_Type_DetailedBreast Invasive Lobular Carcinoma          -0.670 0.502
## Cancer_Type_DetailedBreast Invasive Mixed Mucinous Carcinoma   -0.956 0.339
## Cancer_Type_DetailedBreast Mixed Ductal and Lobular Carcinoma  -0.817 0.413
## Cancer_Type_DetailedInvasive Breast Carcinoma                  -1.239 0.215
## Lymph_nodes_examined_positive                                  -3.318 0.000
## Nottingham_prognostic_index                                     0.001 0.999
## Tumor_StageStage II                                            -2.001 0.045
## Tumor_StageStage III                                           -1.643 0.100
## Tumor_StageStage IV                                            -2.046 0.040
## Neoplasm_Histologic_Grade2                                      0.236 0.813
## Neoplasm_Histologic_Grade3                                      0.154 0.877
## ER1                                                             1.634 0.102
## PR1                                                             1.377 0.168
## HER21                                                          -3.007 0.002
## Chemo1                                                         -3.621 0.000
## Hormone1                                                       -2.568 0.010
## Radio1                                                          1.119 0.263
## Menopause1                                                      2.779 0.005
##
## (Intercept)                                                   ***
## Age_at_Diagnosis                                              ***
## Tumor_Size                                                    **
## Cancer_Type_DetailedBreast Invasive Ductal Carcinoma
## Cancer_Type_DetailedBreast Invasive Lobular Carcinoma
## Cancer_Type_DetailedBreast Invasive Mixed Mucinous Carcinoma
## Cancer_Type_DetailedBreast Mixed Ductal and Lobular Carcinoma
## Cancer_Type_DetailedInvasive Breast Carcinoma
## Lymph_nodes_examined_positive                                 ***
## Nottingham_prognostic_index
```

```
## Tumor_StageStage II                                              *
## Tumor_StageStage III
## Tumor_StageStage IV                                              *
## Neoplasm_Histologic_Grade2
## Neoplasm_Histologic_Grade3
## ER1
## PR1
## HER21                                                            **
## Chemo1                                                           ***
## Hormone1                                                         *
## Radio1
## Menopause1                                                       **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 71.91 on 1331 degrees of freedom
## Multiple R-squared:  0.1643, Adjusted R-squared:  0.1512
## F-statistic: 12.46 on 21 and 1331 DF,  p-value: < 2.2e-16
```

```r
# Predict on test data
sm_predictions_osm <- predict(sm_linear_reg_model, newdata = x_test_m)

# Evaluate performance with RMSE (Root Mean Squared Error)
sm_rmse_osm <- sqrt(mean((sm_predictions_osm - sm_test)^2))
print(paste("RMSE for Linear Regression:", sm_rmse_osm))
```

```
## [1] "RMSE for Linear Regression: 71.3202598917447"
```

```r
library(randomForest)

# Train a Random Forest Regression model
sm_rf_reg_model <- randomForest(sm_train ~ .,
                                data = x_train_m, ntree = 500)

# Predict on test data
sm_rf_reg_pred <- predict(sm_rf_reg_model, newdata = x_test_m)

# Evaluate performance with RMSE (Root Mean Squared Error)
sm_rf_reg_rmse <- sqrt(mean((sm_rf_reg_pred - sm_test)^2))
print(paste("RMSE for Random Forest Regression:", sm_rf_reg_rmse))
```

```
## [1] "RMSE for Random Forest Regression: 39.5820120878021"
```

```r
library(glmnet)

# Create matrices
x_train_matrix_m <- model.matrix(~ . -1, data = x_train_m)
x_test_matrix_m <- model.matrix(~ . -1, data = x_test_m)

# Create matrices
# Fit model (alpha=0.5 for Elastic Net)
sm_cv_model_osm <- cv.glmnet(x_train_matrix_m, sm_train, alpha = 0.5)

# Predict
sm_enet_pred_osm <- predict(sm_cv_model_osm, newx = x_test_matrix_m, s = "lamb

# Evaluate performance with RMSE (Root Mean Squared Error)
sm_enet_rmse_osm <- sqrt(mean((sm_enet_pred_osm - sm_test)^2))
print(paste("RMSE for Elastic Net Regression:", sm_enet_rmse_osm))
```

```
## [1] "RMSE for Elastic Net Regression: 71.3578520831614"
```

```r
library(xgboost)

# Train an XGBoost model
sm_xgb_model_osm <- xgboost(data = x_train_matrix_m, label = sm_train,
                            objective = "reg:squarederror", nrounds = 100, max.de
```

```
## [1]  train-rmse:137.177275
## [2]  train-rmse:126.956768
## [3]  train-rmse:117.879695
## [4]  train-rmse:109.964111
## [5]  train-rmse:103.040434
## [6]  train-rmse:96.992666
## [7]  train-rmse:91.720173
## [8]  train-rmse:87.082584
## [9]  train-rmse:83.136605
## [10] train-rmse:79.573643
## [11] train-rmse:76.546675
## [12] train-rmse:73.966494
## [13] train-rmse:71.676178
## [14] train-rmse:69.687130
## [15] train-rmse:67.827123
## [16] train-rmse:66.338631
## [17] train-rmse:65.128015
## [18] train-rmse:63.875869
```

```
## [19] train-rmse:62.788494
## [20] train-rmse:61.869823
## [21] train-rmse:60.938759
## [22] train-rmse:60.126314
## [23] train-rmse:59.433286
## [24] train-rmse:58.931066
## [25] train-rmse:58.496346
## [26] train-rmse:57.988552
## [27] train-rmse:57.410080
## [28] train-rmse:57.109919
## [29] train-rmse:56.899601
## [30] train-rmse:56.587587
## [31] train-rmse:55.992156
## [32] train-rmse:55.857183
## [33] train-rmse:55.762664
## [34] train-rmse:55.364882
## [35] train-rmse:55.253307
## [36] train-rmse:54.999324
## [37] train-rmse:54.770655
## [38] train-rmse:54.559911
## [39] train-rmse:54.207711
## [40] train-rmse:54.124541
## [41] train-rmse:54.074528
## [42] train-rmse:54.009309
## [43] train-rmse:53.972700
## [44] train-rmse:53.813297
## [45] train-rmse:53.526834
## [46] train-rmse:53.332611
## [47] train-rmse:53.184800
## [48] train-rmse:53.138078
## [49] train-rmse:52.762053
## [50] train-rmse:52.728303
## [51] train-rmse:52.528783
## [52] train-rmse:52.218222
## [53] train-rmse:52.129354
## [54] train-rmse:51.859881
## [55] train-rmse:51.747723
## [56] train-rmse:51.514871
## [57] train-rmse:51.390714
## [58] train-rmse:51.317331
## [59] train-rmse:51.274231
## [60] train-rmse:51.203500
## [61] train-rmse:51.106740
## [62] train-rmse:51.070418
## [63] train-rmse:50.960612
## [64] train-rmse:50.939083
## [65] train-rmse:50.653186
## [66] train-rmse:50.380899
## [67] train-rmse:50.281096
```

```
## [68] train-rmse:50.175224
## [69] train-rmse:50.077963
## [70] train-rmse:49.744254
## [71] train-rmse:49.569645
## [72] train-rmse:49.329140
## [73] train-rmse:49.260004
## [74] train-rmse:49.102543
## [75] train-rmse:48.940396
## [76] train-rmse:48.778764
## [77] train-rmse:48.652026
## [78] train-rmse:48.609682
## [79] train-rmse:48.156571
## [80] train-rmse:47.957327
## [81] train-rmse:47.920835
## [82] train-rmse:47.618279
## [83] train-rmse:47.437497
## [84] train-rmse:47.355706
## [85] train-rmse:47.196855
## [86] train-rmse:46.836336
## [87] train-rmse:46.809511
## [88] train-rmse:46.495159
## [89] train-rmse:46.340071
## [90] train-rmse:46.184213
## [91] train-rmse:46.152751
## [92] train-rmse:46.017012
## [93] train-rmse:45.715802
## [94] train-rmse:45.656873
## [95] train-rmse:45.499337
## [96] train-rmse:45.255152
## [97] train-rmse:45.173125
## [98] train-rmse:45.006028
## [99] train-rmse:44.856898
## [100]    train-rmse:44.574881
```

```
# Predict
sm_xgb_pred_osm <- predict(sm_xgb_model_osm, newdata = x_test_matrix_m)

# Evaluate performance with RMSE (Root Mean Squared Error)
sm_xgb_rmse_osm <- sqrt(mean((sm_xgb_pred_osm - sm_test)^2))
print(paste("RMSE for XGBoost Regression:", sm_xgb_rmse_osm))
```

```
## [1] "RMSE for XGBoost Regression: 44.5748811370455"
```

RELAPSE STATUS MODEL TRAINING

```r
rs_log_reg_model <- glm(rs_train ~ .,
                    data = x_train, family = "binomial")

# Summary of the model
summary(rs_log_reg_model)
```

```
##
## Call:
## glm(formula = rs_train ~ ., family = "binomial", data = x_train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5181  -0.9687  -0.7765   1.2191   1.9951
##
## Coefficients:
##                                                                 Estimate
## (Intercept)                                                   -2.597224
## Age_at_Diagnosis                                              -0.005932
## Tumor_Size                                                     0.015542
## Cancer_Type_DetailedBreast Invasive Ductal Carcinoma           0.497091
## Cancer_Type_DetailedBreast Invasive Lobular Carcinoma          0.598398
## Cancer_Type_DetailedBreast Invasive Mixed Mucinous Carcinoma  -0.435222
## Cancer_Type_DetailedBreast Mixed Ductal and Lobular Carcinoma  0.595358
## Cancer_Type_DetailedInvasive Breast Carcinoma                  0.262641
## Lymph_nodes_examined_positive                                  0.049746
## Nottingham_prognostic_index                                    0.423722
## Tumor_StageStage II                                            0.007001
## Tumor_StageStage III                                           0.039525
## Tumor_StageStage IV                                           15.609078
## Neoplasm_Histologic_Grade2                                     0.014027
## Neoplasm_Histologic_Grade3                                    -0.162012
## ER1                                                            0.433749
## PR1                                                            0.036446
## HER21                                                          0.647733
## Chemo1                                                        -0.222071
## Hormone1                                                      -0.469076
## Radio1                                                        -0.149293
## Menopause1                                                    -0.099048
##                                                              Std. Error
## (Intercept)                                                    1.297250
## Age_at_Diagnosis                                               0.008034
## Tumor_Size                                                     0.006016
## Cancer_Type_DetailedBreast Invasive Ductal Carcinoma           1.170257
## Cancer_Type_DetailedBreast Invasive Lobular Carcinoma          1.196551
## Cancer_Type_DetailedBreast Invasive Mixed Mucinous Carcinoma   1.399595
## Cancer_Type_DetailedBreast Mixed Ductal and Lobular Carcinoma  1.182663
## Cancer_Type_DetailedInvasive Breast Carcinoma                  1.385545
```

```
## Lymph_nodes_examined_positive                                  0.032280
## Nottingham_prognostic_index                                    0.175074
## Tumor_StageStage II                                            0.196146
## Tumor_StageStage III                                           0.368831
## Tumor_StageStage IV                                          466.774632
## Neoplasm_Histologic_Grade2                                     0.322172
## Neoplasm_Histologic_Grade3                                     0.440592
## ER1                                                            0.221183
## PR1                                                            0.156622
## HER21                                                          0.214845
## Chemo1                                                         0.216869
## Hormone1                                                       0.164082
## Radio1                                                         0.145136
## Menopause1                                                     0.232546
##                                                              z value Pr(>|
## (Intercept)                                                   -2.002  0.04
## Age_at_Diagnosis                                              -0.738  0.46
## Tumor_Size                                                     2.583  0.00
## Cancer_Type_DetailedBreast Invasive Ductal Carcinoma           0.425  0.67
## Cancer_Type_DetailedBreast Invasive Lobular Carcinoma          0.500  0.61
## Cancer_Type_DetailedBreast Invasive Mixed Mucinous Carcinoma  -0.311  0.75
## Cancer_Type_DetailedBreast Mixed Ductal and Lobular Carcinoma  0.503  0.61
## Cancer_Type_DetailedInvasive Breast Carcinoma                  0.190  0.84
## Lymph_nodes_examined_positive                                  1.541  0.12
## Nottingham_prognostic_index                                    2.420  0.01
## Tumor_StageStage II                                            0.036  0.97
## Tumor_StageStage III                                           0.107  0.91
## Tumor_StageStage IV                                            0.033  0.97
## Neoplasm_Histologic_Grade2                                     0.044  0.96
## Neoplasm_Histologic_Grade3                                    -0.368  0.71
## ER1                                                            1.961  0.04
## PR1                                                            0.233  0.81
## HER21                                                          3.015  0.00
## Chemo1                                                        -1.024  0.30
## Hormone1                                                      -2.859  0.00
## Radio1                                                        -1.029  0.30
## Menopause1                                                    -0.426  0.67
##
## (Intercept)                                                  *
## Age_at_Diagnosis
## Tumor_Size                                                   **
## Cancer_Type_DetailedBreast Invasive Ductal Carcinoma
## Cancer_Type_DetailedBreast Invasive Lobular Carcinoma
## Cancer_Type_DetailedBreast Invasive Mixed Mucinous Carcinoma
## Cancer_Type_DetailedBreast Mixed Ductal and Lobular Carcinoma
## Cancer_Type_DetailedInvasive Breast Carcinoma
## Lymph_nodes_examined_positive
## Nottingham_prognostic_index                                  *
## Tumor_StageStage II
```

```
## Tumor_StageStage III
## Tumor_StageStage IV
## Neoplasm_Histologic_Grade2
## Neoplasm_Histologic_Grade3
## ER1                                                          *
## PR1
## HER21                                                        **
## Chemo1
## Hormone1                                                     **
## Radio1
## Menopause1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1470.3  on 1082  degrees of freedom
## Residual deviance: 1357.0  on 1061  degrees of freedom
## AIC: 1401
##
## Number of Fisher Scoring iterations: 14
```

```
# Predict on test data
predictions_status <- predict(rs_log_reg_model, newdata = x_test, type = "resp
pred_class_status <- ifelse(predictions_status > 0.5, 1, 0)  # 1 = Dead, 0 = A

# Confusion Matrix to evaluate the performance
rs_conf_matrix_status <- confusionMatrix(factor(pred_class_status), factor(rs_
print(rs_conf_matrix_status)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 130  66
##          1  28  46
##
##                Accuracy : 0.6519
##                  95% CI : (0.5917, 0.7086)
##     No Information Rate : 0.5852
##     P-Value [Acc > NIR] : 0.0147253
##
##                   Kappa : 0.2456
##
##  Mcnemar's Test P-Value : 0.0001355
##
```

```
##             Sensitivity : 0.8228
##             Specificity : 0.4107
##          Pos Pred Value : 0.6633
##          Neg Pred Value : 0.6216
##              Prevalence : 0.5852
##          Detection Rate : 0.4815
##    Detection Prevalence : 0.7259
##       Balanced Accuracy : 0.6167
##
##        'Positive' Class : 0
##
```

Random Forest

```
library(randomForest)

set.seed(123)
rs_rf_model <- randomForest(rs_train ~ .,
                        data = x_train, ntree = 500, importance = TRUE)

# Predict and evaluate
rs_rf_pred <- predict(rs_rf_model, newdata = x_test)

confusionMatrix(rs_rf_pred, factor(rs_test))
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 125  61
##          1  33  51
##
##                Accuracy : 0.6519
##                  95% CI : (0.5917, 0.7086)
##     No Information Rate : 0.5852
##     P-Value [Acc > NIR] : 0.014725
##
##                   Kappa : 0.2558
##
##  Mcnemar's Test P-Value : 0.005355
##
##             Sensitivity : 0.7911
##             Specificity : 0.4554
##          Pos Pred Value : 0.6720
##          Neg Pred Value : 0.6071
##              Prevalence : 0.5852
```

```
##          Detection Rate : 0.4630
##    Detection Prevalence : 0.6889
##       Balanced Accuracy : 0.6232
##
##        'Positive' Class : 0
##
```

```
library(glmnet)

# Fit model (alpha=0.5 for Elastic Net)
rs_cv_model <- cv.glmnet(x_train_matrix, rs_train, alpha = 0.5, family = "binc

# Predict
enet_pred <- predict(rs_cv_model, newx = x_test_matrix, s = "lambda.min", type
enet_pred_class <- ifelse(enet_pred > 0.5, 1, 0)
confusionMatrix(as.factor(enet_pred_class), as.factor(rs_test))
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 137  71
##          1  21  41
##
##                Accuracy : 0.6593
##                  95% CI : (0.5994, 0.7156)
##     No Information Rate : 0.5852
##     P-Value [Acc > NIR] : 0.007555
##
##                   Kappa : 0.2494
##
##  Mcnemar's Test P-Value : 3.245e-07
##
##             Sensitivity : 0.8671
##             Specificity : 0.3661
##          Pos Pred Value : 0.6587
##          Neg Pred Value : 0.6613
##              Prevalence : 0.5852
##          Detection Rate : 0.5074
##    Detection Prevalence : 0.7704
##       Balanced Accuracy : 0.6166
##
##        'Positive' Class : 0
##
```

```r
library(xgboost)


rs_xgb_model <- xgboost(data = x_train_matrix, label = as.numeric(rs_train) -
                        objective = "binary:logistic", nrounds = 100, max.depth =
```

```
## [1]  train-logloss:0.676005
## [2]  train-logloss:0.657922
## [3]  train-logloss:0.643503
## [4]  train-logloss:0.630730
## [5]  train-logloss:0.620522
## [6]  train-logloss:0.610882
## [7]  train-logloss:0.602138
## [8]  train-logloss:0.592953
## [9]  train-logloss:0.583342
## [10] train-logloss:0.573740
## [11] train-logloss:0.566739
## [12] train-logloss:0.559794
## [13] train-logloss:0.554281
## [14] train-logloss:0.547588
## [15] train-logloss:0.541701
## [16] train-logloss:0.536439
## [17] train-logloss:0.533251
## [18] train-logloss:0.530702
## [19] train-logloss:0.526895
## [20] train-logloss:0.524057
## [21] train-logloss:0.521670
## [22] train-logloss:0.518490
## [23] train-logloss:0.516284
## [24] train-logloss:0.513161
## [25] train-logloss:0.511464
## [26] train-logloss:0.509485
## [27] train-logloss:0.508306
## [28] train-logloss:0.505584
## [29] train-logloss:0.504536
## [30] train-logloss:0.503209
## [31] train-logloss:0.501233
## [32] train-logloss:0.500661
## [33] train-logloss:0.497295
## [34] train-logloss:0.496254
## [35] train-logloss:0.493966
## [36] train-logloss:0.492812
## [37] train-logloss:0.490410
## [38] train-logloss:0.489602
## [39] train-logloss:0.487966
## [40] train-logloss:0.486634
```

```
## [41] train-logloss:0.485623
## [42] train-logloss:0.484954
## [43] train-logloss:0.482581
## [44] train-logloss:0.481396
## [45] train-logloss:0.480862
## [46] train-logloss:0.479177
## [47] train-logloss:0.477340
## [48] train-logloss:0.476874
## [49] train-logloss:0.473061
## [50] train-logloss:0.471311
## [51] train-logloss:0.467008
## [52] train-logloss:0.466418
## [53] train-logloss:0.465089
## [54] train-logloss:0.462433
## [55] train-logloss:0.461054
## [56] train-logloss:0.460116
## [57] train-logloss:0.453618
## [58] train-logloss:0.452099
## [59] train-logloss:0.451657
## [60] train-logloss:0.450354
## [61] train-logloss:0.449716
## [62] train-logloss:0.447656
## [63] train-logloss:0.446277
## [64] train-logloss:0.442243
## [65] train-logloss:0.437837
## [66] train-logloss:0.437566
## [67] train-logloss:0.437090
## [68] train-logloss:0.435150
## [69] train-logloss:0.434314
## [70] train-logloss:0.430566
## [71] train-logloss:0.429652
## [72] train-logloss:0.426643
## [73] train-logloss:0.425655
## [74] train-logloss:0.422264
## [75] train-logloss:0.420015
## [76] train-logloss:0.416601
## [77] train-logloss:0.415703
## [78] train-logloss:0.415474
## [79] train-logloss:0.414348
## [80] train-logloss:0.412571
## [81] train-logloss:0.411883
## [82] train-logloss:0.411011
## [83] train-logloss:0.409990
## [84] train-logloss:0.409607
## [85] train-logloss:0.409148
## [86] train-logloss:0.405997
## [87] train-logloss:0.404499
## [88] train-logloss:0.403029
## [89] train-logloss:0.401924
```

```
## [90] train-logloss:0.398837
## [91] train-logloss:0.397637
## [92] train-logloss:0.396691
## [93] train-logloss:0.396054
## [94] train-logloss:0.395557
## [95] train-logloss:0.394583
## [96] train-logloss:0.393729
## [97] train-logloss:0.393577
## [98] train-logloss:0.393042
## [99] train-logloss:0.392623
## [100]    train-logloss:0.392095
```

```
# Prediction
xgb_pred <- predict(rs_xgb_model, newdata = x_test_matrix)
xgb_pred_class <- ifelse(xgb_pred > 0.5, 1, 0)
confusionMatrix(as.factor(xgb_pred_class), factor(rs_test))
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 119  53
##          1  39  59
##
##                Accuracy : 0.6593
##                  95% CI : (0.5994, 0.7156)
##     No Information Rate : 0.5852
##     P-Value [Acc > NIR] : 0.007555
##
##                   Kappa : 0.2851
##
##  Mcnemar's Test P-Value : 0.175308
##
##             Sensitivity : 0.7532
##             Specificity : 0.5268
##          Pos Pred Value : 0.6919
##          Neg Pred Value : 0.6020
##              Prevalence : 0.5852
##          Detection Rate : 0.4407
##    Detection Prevalence : 0.6370
##       Balanced Accuracy : 0.6400
##
##        'Positive' Class : 0
##
```

OVERALL RELAPSE MONTHS

```
rm_linear_reg_model <- lm(rm_train ~ Age_at_Diagnosis + Tumor_Size + Cancer_Ty
                          Lymph_nodes_examined_positive +
                          Nottingham_prognostic_index + Tumor_Stage + Neoplasm_
                          PR + HER2 + Chemo + Hormone + Radio + Menopause,
                        data = x_train_m)

# Summary of the model
summary(rm_linear_reg_model)
```

```
##
## Call:
## lm(formula = rm_train ~ Age_at_Diagnosis + Tumor_Size + Cancer_Type_Detaile
##     Lymph_nodes_examined_positive + Nottingham_prognostic_index +
##     Tumor_Stage + Neoplasm_Histologic_Grade + ER + PR + HER2 +
##     Chemo + Hormone + Radio + Menopause, data = x_train_m)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -148.086  -58.804   -9.628   55.273  215.993
##
## Coefficients:
##                                                                  Estimate
## (Intercept)                                                      225.4679
## Age_at_Diagnosis                                                  -1.1796
## Tumor_Size                                                        -0.5365
## Cancer_Type_DetailedBreast Invasive Ductal Carcinoma            -31.1670
## Cancer_Type_DetailedBreast Invasive Lobular Carcinoma           -29.2029
## Cancer_Type_DetailedBreast Invasive Mixed Mucinous Carcinoma    -21.6378
## Cancer_Type_DetailedBreast Mixed Ductal and Lobular Carcinoma   -27.2817
## Cancer_Type_DetailedInvasive Breast Carcinoma                   -36.8975
## Lymph_nodes_examined_positive                                    -2.8765
## Nottingham_prognostic_index                                       0.8580
## Tumor_StageStage II                                              -9.6177
## Tumor_StageStage III                                            -16.5862
## Tumor_StageStage IV                                             -55.3280
## Neoplasm_Histologic_Grade2                                       -2.7395
## Neoplasm_Histologic_Grade3                                       -1.9405
## ER1                                                               0.9561
## PR1                                                               4.1172
## HER21                                                           -18.8594
## Chemo1                                                          -16.2923
## Hormone1                                                         -5.1531
## Radio1                                                            5.7151
```

```
## Menopause1                                                    19.7036
##                                                              Std. Error
## (Intercept)                                                     32.8092
## Age_at_Diagnosis                                                 0.2510
## Tumor_Size                                                       0.1662
## Cancer_Type_DetailedBreast Invasive Ductal Carcinoma            28.2464
## Cancer_Type_DetailedBreast Invasive Lobular Carcinoma           29.2318
## Cancer_Type_DetailedBreast Invasive Mixed Mucinous Carcinoma    33.4907
## Cancer_Type_DetailedBreast Mixed Ductal and Lobular Carcinoma   28.7220
## Cancer_Type_DetailedInvasive Breast Carcinoma                   38.5181
## Lymph_nodes_examined_positive                                    0.7898
## Nottingham_prognostic_index                                      5.0603
## Tumor_StageStage II                                              5.8675
## Tumor_StageStage III                                            11.2543
## Tumor_StageStage IV                                             25.7029
## Neoplasm_Histologic_Grade2                                       9.2372
## Neoplasm_Histologic_Grade3                                      12.7906
## ER1                                                              6.7628
## PR1                                                              4.8875
## HER21                                                            6.5270
## Chemo1                                                           6.7191
## Hormone1                                                         5.0082
## Radio1                                                           4.4971
## Menopause1                                                       7.3278
##                                                             t value Pr(>|
## (Intercept)                                                   6.872 9.69e
## Age_at_Diagnosis                                             -4.699 2.89e
## Tumor_Size                                                   -3.227 0.001
## Cancer_Type_DetailedBreast Invasive Ductal Carcinoma         -1.103 0.270
## Cancer_Type_DetailedBreast Invasive Lobular Carcinoma        -0.999 0.317
## Cancer_Type_DetailedBreast Invasive Mixed Mucinous Carcinoma -0.646 0.518
## Cancer_Type_DetailedBreast Mixed Ductal and Lobular Carcinoma -0.950 0.342
## Cancer_Type_DetailedInvasive Breast Carcinoma                -0.958 0.338
## Lymph_nodes_examined_positive                                -3.642 0.000
## Nottingham_prognostic_index                                   0.170 0.865
## Tumor_StageStage II                                          -1.639 0.101
## Tumor_StageStage III                                         -1.474 0.140
## Tumor_StageStage IV                                          -2.153 0.031
## Neoplasm_Histologic_Grade2                                   -0.297 0.766
## Neoplasm_Histologic_Grade3                                   -0.152 0.879
## ER1                                                           0.141 0.887
## PR1                                                           0.842 0.399
## HER21                                                        -2.889 0.003
## Chemo1                                                       -2.425 0.015
## Hormone1                                                     -1.029 0.303
## Radio1                                                        1.271 0.204
## Menopause1                                                    2.689 0.007
##
## (Intercept)                                                 ***
```

```
## Age_at_Diagnosis                                                  ***
## Tumor_Size                                                        **
## Cancer_Type_DetailedBreast Invasive Ductal Carcinoma
## Cancer_Type_DetailedBreast Invasive Lobular Carcinoma
## Cancer_Type_DetailedBreast Invasive Mixed Mucinous Carcinoma
## Cancer_Type_DetailedBreast Mixed Ductal and Lobular Carcinoma
## Cancer_Type_DetailedInvasive Breast Carcinoma
## Lymph_nodes_examined_positive                                     ***
## Nottingham_prognostic_index
## Tumor_StageStage II
## Tumor_StageStage III
## Tumor_StageStage IV                                               *
## Neoplasm_Histologic_Grade2
## Neoplasm_Histologic_Grade3
## ER1
## PR1
## HER21                                                             **
## Chemo1                                                            *
## Hormone1
## Radio1
## Menopause1                                                        **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 74.23 on 1331 degrees of freedom
## Multiple R-squared:  0.1197, Adjusted R-squared:  0.1058
## F-statistic: 8.617 on 21 and 1331 DF,  p-value: < 2.2e-16
```

```
# Predict on test data
rm_predictions_osm <- predict(rm_linear_reg_model, newdata = x_train_m)

# Evaluate performance with RMSE (Root Mean Squared Error)
rm_rmse_osm <- sqrt(mean((rm_predictions_osm - rm_test)^2))
print(paste("RMSE for Linear Regression:", rm_rmse_osm))
```

```
## [1] "RMSE for Linear Regression: 73.6221256869938"
```

```
library(randomForest)

# Train a Random Forest Regression model
rm_rf_reg_model <- randomForest(rm_train ~ Age_at_Diagnosis + Tumor_Size + Car
                                Lymph_nodes_examined_positive + Nottingham_prognc
                                Neoplasm_Histologic_Grade + ER + PR + HER2 + Chem
                                data = x_train_m, ntree = 500)
```

```r
# Predict on test data
rm_rf_reg_pred <- predict(rm_rf_reg_model, newdata = x_train_m)

# Evaluate performance with RMSE (Root Mean Squared Error)
rm_rf_reg_rmse <- sqrt(mean((rm_rf_reg_pred - rm_test)^2))
print(paste("RMSE for Random Forest Regression:", rm_rf_reg_rmse))
```

```
## [1] "RMSE for Random Forest Regression: 41.171194465206"
```

```r
library(glmnet)

# Create matrices
# Fit model (alpha=0.5 for Elastic Net)
rm_cv_model_osm <- cv.glmnet(x_train_matrix_m, rm_train, alpha = 0.5)

# Predict
rm_enet_pred_osm <- predict(rm_cv_model_osm, newx = x_train_matrix_m, s = "lam

# Evaluate performance with RMSE (Root Mean Squared Error)
rm_enet_rmse_osm <- sqrt(mean((rm_enet_pred_osm - rm_test)^2))
print(paste("RMSE for Elastic Net Regression:", rm_enet_rmse_osm))
```

```
## [1] "RMSE for Elastic Net Regression: 73.7088928473147"
```

```r
library(xgboost)

# Train an XGBoost model
rm_xgb_model_osm <- xgboost(data = x_train_matrix_m, label = rm_train,
                            objective = "reg:squarederror", nrounds = 100, max.de
```

```
## [1]  train-rmse:126.300596
## [2]  train-rmse:117.829612
## [3]  train-rmse:110.415597
## [4]  train-rmse:103.913536
## [5]  train-rmse:98.217631
## [6]  train-rmse:93.328480
## [7]  train-rmse:89.011855
## [8]  train-rmse:85.313855
## [9]  train-rmse:81.982055
## [10] train-rmse:79.195468
## [11] train-rmse:76.807601
```

```
## [12] train-rmse:74.594402
## [13] train-rmse:72.889168
## [14] train-rmse:71.197075
## [15] train-rmse:69.635382
## [16] train-rmse:68.585154
## [17] train-rmse:67.359246
## [18] train-rmse:66.640804
## [19] train-rmse:66.020047
## [20] train-rmse:65.157438
## [21] train-rmse:64.287514
## [22] train-rmse:63.663597
## [23] train-rmse:63.296035
## [24] train-rmse:62.826605
## [25] train-rmse:62.517395
## [26] train-rmse:62.320173
## [27] train-rmse:61.500348
## [28] train-rmse:61.189667
## [29] train-rmse:60.792547
## [30] train-rmse:60.609998
## [31] train-rmse:60.475578
## [32] train-rmse:60.345974
## [33] train-rmse:60.261655
## [34] train-rmse:60.134742
## [35] train-rmse:59.930572
## [36] train-rmse:59.768069
## [37] train-rmse:59.590252
## [38] train-rmse:59.416691
## [39] train-rmse:59.272476
## [40] train-rmse:58.942284
## [41] train-rmse:58.793292
## [42] train-rmse:58.677808
## [43] train-rmse:58.636556
## [44] train-rmse:58.481721
## [45] train-rmse:58.354241
## [46] train-rmse:58.223180
## [47] train-rmse:57.920687
## [48] train-rmse:57.815919
## [49] train-rmse:57.499041
## [50] train-rmse:56.800525
## [51] train-rmse:56.734877
## [52] train-rmse:56.532416
## [53] train-rmse:56.369891
## [54] train-rmse:56.342593
## [55] train-rmse:56.223318
## [56] train-rmse:55.902741
## [57] train-rmse:55.798326
## [58] train-rmse:55.706411
## [59] train-rmse:55.573600
## [60] train-rmse:55.327907
```

```
## [61] train-rmse:55.087262
## [62] train-rmse:54.847539
## [63] train-rmse:54.463337
## [64] train-rmse:54.226224
## [65] train-rmse:53.730362
## [66] train-rmse:53.489414
## [67] train-rmse:53.222826
## [68] train-rmse:53.043247
## [69] train-rmse:52.704688
## [70] train-rmse:52.521917
## [71] train-rmse:52.328714
## [72] train-rmse:52.151608
## [73] train-rmse:51.934513
## [74] train-rmse:51.561755
## [75] train-rmse:51.214531
## [76] train-rmse:51.062125
## [77] train-rmse:50.906834
## [78] train-rmse:50.567461
## [79] train-rmse:50.220886
## [80] train-rmse:50.037418
## [81] train-rmse:49.684625
## [82] train-rmse:49.619417
## [83] train-rmse:49.260274
## [84] train-rmse:49.123117
## [85] train-rmse:48.899909
## [86] train-rmse:48.712869
## [87] train-rmse:48.664231
## [88] train-rmse:48.279982
## [89] train-rmse:48.173340
## [90] train-rmse:48.033510
## [91] train-rmse:47.631827
## [92] train-rmse:47.298556
## [93] train-rmse:47.144161
## [94] train-rmse:46.983937
## [95] train-rmse:46.634108
## [96] train-rmse:46.380341
## [97] train-rmse:46.024123
## [98] train-rmse:45.920198
## [99] train-rmse:45.711891
## [100]    train-rmse:45.601603
```

```r
# Predict
rm_xgb_pred_osm <- predict(rm_xgb_model_osm, newdata = x_train_matrix_m)

# Evaluate performance with RMSE (Root Mean Squared Error)
rm_xgb_rmse_osm <- sqrt(mean((rm_xgb_pred_osm - rm_test)^2))
print(paste("RMSE for XGBoost Regression:", rm_xgb_rmse_osm))
```

```
## [1] "RMSE for XGBoost Regression: 45.6016025023595"
```

```r
# Collect classification results
classification_results <- data.frame(
  Model = c(
    "Overall Survival - Logistic Regression",
    "Overall Survival - Random Forest",
    "Overall Survival - Elastic Net",
    "Overall Survival - XGBoost",
    "Relapse Status - Logistic Regression",
    "Relapse Status - Random Forest",
    "Relapse Status - Elastic Net",
    "Relapse Status - XGBoost"
  ),
  Accuracy = c(
    ss_conf_matrix_status$overall["Accuracy"],
    confusionMatrix(ss_rf_pred, factor(ss_test))$overall["Accuracy"],
    confusionMatrix(as.factor(ss_enet_pred_class), as.factor(ss_test))$overall
    confusionMatrix(as.factor(ss_xgb_pred_class), ss_test)$overall["Accuracy"]
    rs_conf_matrix_status$overall["Accuracy"],
    confusionMatrix(rs_rf_pred, factor(rs_test))$overall["Accuracy"],
    confusionMatrix(as.factor(enet_pred_class), as.factor(rs_test))$overall["A
    confusionMatrix(as.factor(xgb_pred_class), factor(rs_test))$overall["Accur
  ),
  Kappa = c(
    ss_conf_matrix_status$overall["Kappa"],
    confusionMatrix(ss_rf_pred, factor(ss_test))$overall["Kappa"],
    confusionMatrix(as.factor(enet_pred_class), as.factor(ss_test))$overall["K
    confusionMatrix(as.factor(xgb_pred_class), ss_test)$overall["Kappa"],
    rs_conf_matrix_status$overall["Kappa"],
    confusionMatrix(rs_rf_pred, factor(rs_test))$overall["Kappa"],
    confusionMatrix(as.factor(enet_pred_class), as.factor(rs_test))$overall["K
    confusionMatrix(as.factor(xgb_pred_class), factor(rs_test))$overall["Kappa
  )
)

# Print the table
print(classification_results)
```

```
##                                 Model  Accuracy     Kappa
## 1 Overall Survival - Logistic Regression 0.6222222 0.2309411
## 2       Overall Survival - Random Forest 0.6407407 0.2679859
## 3         Overall Survival - Elastic Net 0.6296296 0.1298407
## 4             Overall Survival - XGBoost 0.6703704 0.1462507
## 5   Relapse Status - Logistic Regression 0.6518519 0.2456307
```

```
## 6         Relapse Status - Random Forest 0.6518519 0.2558058
## 7           Relapse Status - Elastic Net 0.6592593 0.2493654
## 8               Relapse Status - XGBoost 0.6592593 0.2851387
```

```r
# Collect regression results
regression_results <- data.frame(
  Model = c(
    "Overall Survival Months - Linear Regression",
    "Overall Survival Months - Random Forest",
    "Overall Survival Months - Elastic Net",
    "Overall Survival Months - XGBoost",
    "Relapse Months - Linear Regression",
    "Relapse Months - Random Forest",
    "Relapse Months - Elastic Net",
    "Relapse Months - XGBoost"
  ),
  RMSE = c(
    sm_rmse_osm,
    sm_rf_reg_rmse,
    sm_enet_rmse_osm,
    sm_xgb_rmse_osm,
    rm_rmse_osm,
    rm_rf_reg_rmse,
    rm_enet_rmse_osm,
    rm_xgb_rmse_osm
  )
)

# Print the table
print(regression_results)
```

```
##                                       Model     RMSE
## 1 Overall Survival Months - Linear Regression 71.32026
## 2     Overall Survival Months - Random Forest 39.58201
## 3       Overall Survival Months - Elastic Net 71.35785
## 4           Overall Survival Months - XGBoost 44.57488
## 5          Relapse Months - Linear Regression 73.62213
## 6              Relapse Months - Random Forest 41.17119
## 7                Relapse Months - Elastic Net 73.70889
## 8                    Relapse Months - XGBoost 45.60160
```

SAVING THE BEST MODEL

```r
saveRDS(ss_rf_model, "ss_rf_model.rds")
saveRDS(sm_rf_reg_model, "sm_rf_reg_model.rds")
saveRDS(rs_cv_model, "rs_cv_model.rds")
saveRDS(rs_xgb_model, "rs_xgb_model.rds")
saveRDS(rm_rf_reg_model, "rm_rf_reg_model.rds")
```

SIMULATION

```r
# Simulate a new data point for prediction
new_data <- data.frame(
  Age_at_Diagnosis = 50,
  Tumor_Size = 3.5,  # Example size in cm
  Cancer_Type_Detailed = 3,  # E.g., "Breast Invasive Ductal Carcinoma"
  Lymph_nodes_examined_positive = 2,
  Nottingham_prognostic_index = 4.5,
  Tumor_Stage = 2,  # Tumor stage as a factor
  Neoplasm_Histologic_Grade = 2,  # Example grade
  ER = 1,  # ER positive
  PR = 1,  # PR positive
  HER2 = 0,  # HER2 negative
  Chemo = 1,  # Chemotherapy received
  Hormone = 0,  # Hormone therapy not received
  Radio = 1,  # Radiotherapy received
  Menopause = 1  # Post-menopausal
)



# View the new data
new_data
```

```
##   Age_at_Diagnosis Tumor_Size Cancer_Type_Detailed
## 1               50        3.5                     3
##   Lymph_nodes_examined_positive Nottingham_prognostic_index Tumor_Stage
## 1                             2                         4.5           2
##   Neoplasm_Histologic_Grade ER PR HER2 Chemo Hormone Radio Menopause
## 1                         2  1  1    0     1       0     1         1
```

```r
# Convert Tumor Stage to a factor with appropriate labels
new_data$Tumor_Stage <- factor(new_data$Tumor_Stage,
                               levels = c(1, 2, 3, 4),
                               labels = c("Stage I", "Stage II", "Stage III",
```

```r
# Encode ER/PR/HER2 status as binary (Positive = 1, Negative = 0)
new_data$ER <- factor(new_data$ER, levels = c(0, 1))
new_data$PR <- factor(new_data$PR,levels = c(0, 1))
new_data$HER2 <- factor(new_data$HER2, levels = c(0, 1))

new_data$Neoplasm_Histologic_Grade <- factor(new_data$Neoplasm_Histologic_Grad

# Encode Menopausal state as binary
new_data$Menopause <- factor(new_data$Menopause, levels = c(0, 1))

# Encode therapies (Chemotherapy, Hormone Therapy, Radiotherapy) as binary
new_data$Chemo <- factor(new_data$Chemo, levels = c(0, 1))
new_data$Hormone <- factor(new_data$Hormone, levels = c(0, 1))
new_data$Radio <- factor(new_data$Radio, levels = c(0, 1))

# Convert Cancer Type Detailed to factor and numeric
new_data$Cancer_Type_Detailed <- factor(new_data$Cancer_Type_Detailed, levels
                                        labels = c("Breast", "Breast Invasive
                                                   "Breast Invasive Lobular C
                                                   "Breast Mixed Ductal and L




# Ensure necessary columns are present
new_data <- new_data %>%
  select(Age_at_Diagnosis, Tumor_Size, Cancer_Type_Detailed,
         Lymph_nodes_examined_positive, Nottingham_prognostic_index, Tumor_Sta
         Neoplasm_Histologic_Grade, ER, PR, HER2, Chemo, Hormone, Radio, Menop
head(new_data)
```

```
##   Age_at_Diagnosis Tumor_Size              Cancer_Type_Detailed
## 1               50        3.5 Breast Invasive Lobular Carcinoma
##   Lymph_nodes_examined_positive Nottingham_prognostic_index Tumor_Stage
## 1                             2                         4.5    Stage II
##   Neoplasm_Histologic_Grade ER PR HER2 Chemo Hormone Radio Menopause
## 1                         2  1  1    0     1       0     1         1
```

Relapse Free Status

```r
# Convert new_data to a numeric matrix
new_x <- model.matrix(~ . - 1, data = new_data)

# Now predict
```

```
rs_prediction <- predict(rs_cv_model, newx = new_x, s = "lambda.min", type = "

# Output the predicted survival months
rs_prediction
```

```
##   lambda.min
## 1   0.397719
```

Relapse Free Months

```
# Predict survival status using the pre-trained Elastic Net model (cv_model)
rm_prediction <- predict(rm_rf_reg_model, newdata = new_data)


# Output the predicted survival status
rm_prediction
```

```
##        1
## 149.5592
```

Survival Months

```
#setequal(colnames(x_new_data), colnames(x_train_matrix))  # Should return TRU

# Predict survival months using the pre-trained XGBoost model (xgb_model_osm)
sm_prediction <- predict(sm_rf_reg_model, newdata = new_data)

# Output the predicted survival months
sm_prediction
```

```
##        1
## 181.9691
```

Survival Status

```
# Create matrix for Elastic Net prediction of Survival Status
x_new_data <- model.matrix(~ . -1, data = new_data)
```

```r
# Rename new data columns to match training data
colnames(x_new_data)[colnames(x_new_data) == "ERYes"] <- "ER1"
colnames(x_new_data)[colnames(x_new_data) == "PRYes"] <- "PR1"
colnames(x_new_data)[colnames(x_new_data) == "HER2Yes"] <- "HER21"
colnames(x_new_data)[colnames(x_new_data) == "ChemoYes"] <- "Chemo1"
colnames(x_new_data)[colnames(x_new_data) == "HormoneYes"] <- "Hormone1"
colnames(x_new_data)[colnames(x_new_data) == "RadioYes"] <- "Radio1"
colnames(x_new_data)[colnames(x_new_data) == "MenopausePost"] <- "Menopause1"


# Predict survival status using the pre-trained Elastic Net model (cv_model)
ss_prediction <- predict(ss_rf_model, newdata = new_data)


# Output the predicted survival status
ss_prediction
```

```
## 1
## 0
## Levels: 0 1
```