Final Project

Lihui Chen Xuechen Yu Richard Li (Xuanyou)

1. Background and Motivation

Metastatic colorectal cancer (mCRC), which is a significant challenge in oncology, remains something that has a substantial impact on patient morbidity and mortality. Even though there have been a lot of advancements in therapeutic challenges, the prognosis for patients with mCRC is often poor. The selection of first-line therapy is thus a critical determinant of clinical outcomes.

The management of mCRC has evolved with the introduction of targeted therapies, which aim to improve survival rates and quality of life. Among these, the epidermal growth factor receptor (EGFR) inhibitor panitumumab has shown promise when used in combination with the conventional chemotherapy regimen FOLFOX (a combination of folinic acid, fluorouracil, and oxaliplatin).

The trial, NCT00364013, is a randomized, multicenter, phase 3 study. It was designed to evaluate the efficacy of panitumumab in combination with FOLFOX versus FOLFOX alone as first-line therapy for patients with previously untreated mCRC. This study holds a repository of rich clinical data, including patient demographics, treatment details, response criteria, survival metrics, and adverse events. Such comprehensive data presents an invaluable opportunity for the application of advanced statistical and machine learning models to predict patient outcomes.

The motivation behind utilizing these computational models lies in their ability to investigate complex patterns within the data. It potentially leads to the identification of prognostic variables and the development of predictive models for patient survival. By applying these methodologies, the research can shed light on the multifaceted nature of mCRC progression and response to treatment. Moreover, it strives to enhance the decision-making process in clinical settings, enabling personalized medicine approaches and optimizing therapeutic efficacy.

2. Research question

1. How can robust statistical and machine learning models be developed and validated using the dataset from NCT00364013 to accurately predict mortality in patients with metastatic colorectal cancer (mCRC)?
2. What are the key clinical and treatment-related factors that significantly impact the survival of patients with mCRC, and how can their influence be quantified and incorporated into predictive models for patient survival?

3. Data cleaning and exploration

```
library(bis620.2022)
library(readr)
library(dplyr)
```

```
#>
#> Attaching package: 'dplyr'
#> The following objects are masked from 'package:stats':
#>
#>     filter, lag
#> The following objects are masked from 'package:base':
#>
#>     intersect, setdiff, setequal, union
library(tidyr)

data <- adsl %>%
  full_join(biomark, by = 'SUBJID')

data1 = data

my_func <- function() {

  # Function to convert a categorical column to numeric
  convert_to_numeric <- function(column) {
    # Treat NA or blank values as 'unknown'
    column[is.na(column) | column == ""] <- "unknown"

    # Convert the categorical column to a factor and then to numeric
    as.numeric(factor(column)) - 1  # Subtract 1 to start encoding at 0
  }

  # Create dummy variables for 'sex' and 'race', and bind them to the dataset
  if("sex" %in% names(data)) {
    sex_dummies <- model.matrix(~ sex - 1, data)
    colnames(sex_dummies) <- paste("sex", colnames(sex_dummies), sep = "_")
    data <- bind_cols(data, as.data.frame(sex_dummies))
    data <- data %>% select(-sex)
  }

  if("race" %in% names(data)) {
    race_dummies <- model.matrix(~ race - 1, data)
    colnames(race_dummies) <- paste("race", colnames(race_dummies), sep = "_")
    data <- bind_cols(data, as.data.frame(race_dummies))
    data <- data %>% select(-race)
  }

  # Apply the conversion to all other categorical columns except the specified ones
  categorical_columns <- sapply(data, is.character)

  categorical_columns[1] <- FALSE # Assuming the first column is the Subject ID
  #categorical_columns[20:37] <- FALSE # Exclude columns 20 to 37
  categorical_columns[20] <- FALSE
  categorical_columns[22] <- FALSE
  categorical_columns[24] <- FALSE
  categorical_columns[26] <- FALSE
  categorical_columns[28] <- FALSE
  categorical_columns[30] <- FALSE
  categorical_columns[32] <- FALSE
```

```
  categorical_columns[34] <- FALSE
  categorical_columns[36] <- FALSE

  data[categorical_columns] <- lapply(data[categorical_columns], convert_to_numeric)

  print("lala")
}
```

```
# Check Package Installation
library(randomForest)
#> randomForest 4.7-1.1
#> Type rfNews() to see new features/changes/bug fixes.
#>
#> Attaching package: 'randomForest'
#> The following object is masked from 'package:dplyr':
#>
#>     combine
library(rpart.plot)
#> Loading required package: rpart
library(caTools)
library(caret)
#> Loading required package: ggplot2
#>
#> Attaching package: 'ggplot2'
#> The following object is masked from 'package:randomForest':
#>
#>     margin
#> Loading required package: lattice
library(pROC)
#> Type 'citation("pROC")' for a citation.
#>
#> Attaching package: 'pROC'
#> The following objects are masked from 'package:stats':
#>
#>     cov, smooth, var
```

4. Analysis

5. Interpretation and conclusions