Relatório

Limpeza dos dados

Os pacientes 773, 489 e 212 aparentam apresentar erros no registro da data de evolução.

```
data <- readRDS('dados_att.rds') %>%group_by(pac,exame) %>%
mutate(DT_EVOL=replace(DT_EVOL, pac==773, '2017-05-26')) %>%
  mutate(DT_EVOL=replace(DT_EVOL, pac==489, '2017-04-14')) %>%
  mutate(DT_EVOL=replace(DT_EVOL, pac==212, '2017-02-16'))%>%
  mutate(DATA_FINAL =difftime(DT_EVOL,DT_IS,units='days') ) %>%
  mutate(dt_sup=DIAS,dt_inic =lag(DIAS) ) %>% select(-DIAS) %>%
  mutate(dt_ate_inter =difftime(DT_INTERN,DT_IS,units='days') %>% as.numeric )
```

Estruturação dos dados

```
dados <- data %>%
  select(pac,EVOLUCAO,exame,valor,SEXO,DT_NASC,FEBRE:DIARREIA,
         dt_inic,dt_sup,DATA_FINAL,OBITO,dt_ate_inter) %>%
  mutate(Idade = year(as.Date('2019-12-30'))-year(DT_NASC)) %>%
  select(-DT_NASC) %>%
  spread(exame,valor) %>% mutate(TARGET = ifelse(OBITO =='OBITO',1,0)) %>%
  select(-OBITO,-EVOLUCAO)

####### Seleção de variáveis
data_inic <- dados %>% group_by(pac) %>% filter(is.na(dt_inic)) %>% ungroup(pac) %>%
  select(Idade,SEXO,TGO,TGP,HEMORRAGIA,DOR_ABDOM,MIALGIA,Leucocitos,
         Bastoes,Creatinina,Linfocitos,Ureia,DATA_FINAL,TARGET)
```

```
dados <- data %>%
  select(pac,EVOLUCAO,exame,valor,SEXO,DT_NASC,FEBRE:DIARREIA,
         dt_inic,dt_sup,DATA_FINAL,OBITO,dt_ate_inter) %>%
  mutate(Idade = year(as.Date('2019-12-30'))-year(DT_NASC)) %>%
  select(-DT_NASC) %>%
  spread(exame,valor) %>% mutate(TARGET = ifelse(OBITO =='OBITO',1,0)) %>%
  select(-OBITO,-EVOLUCAO)
```

Organização do banco

Nesse cenário todas as observações de um indivíduo que foi a óbito foram consideradas como tendo “Óbito” como variável resposta", essa aplicação é possível e útil pois a utilização do

```
dados2 <- dados %>%
  select(pac,dt_inic,dt_sup,DATA_FINAL,TGO,TGP,TARGET,Idade,Ht,BD,BI,TAP,INR,
         Leucocitos,Bastoes,Creatinina,ICTERICIA,HEMORRAGIA,SEXO,dt_ate_inter
         ,Ureia,GamaGT,Fosfatase_alcal,Plaquetas,SINAL_FAGET) %>%
  mutate(DATA_FINAL = ifelse(DATA_FINAL<0,max(dt_sup),DATA_FINAL))


dados_class <- dados2 %>% group_by(pac) %>% mutate(TARGET=max(TARGET))%>%
  arrange(pac,dt_sup) %>%
```

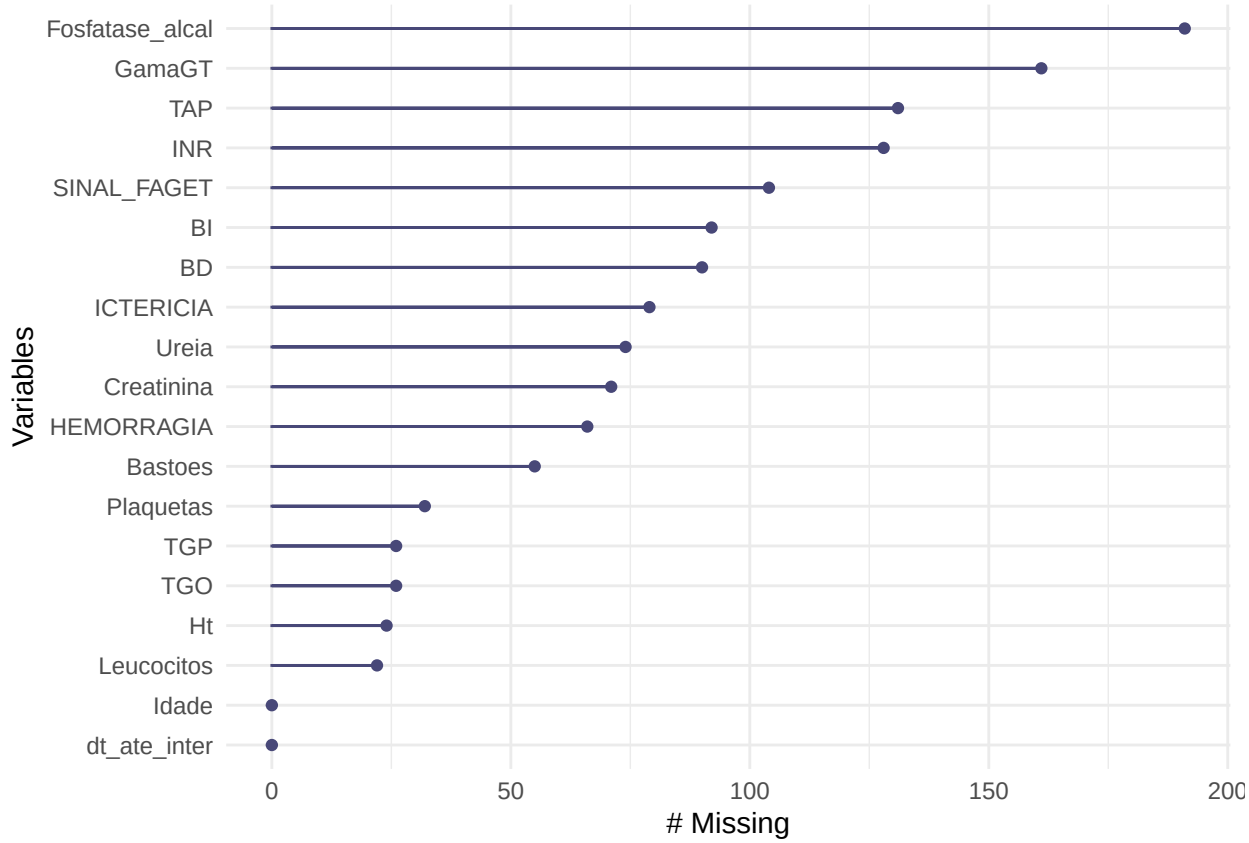
```
  filter(dt_sup <=15)

# dados_completos <- dados2 %>% group_by(pac)%>% filter(any(DATA_FINAL == dt_sup))
#
# dados_faltantes <- dados2 %>% filter(!(pac %in% (dados_completos$pac %>% unique)))
#
# dados_faltantes_2 <- dados_faltantes %>%
#   group_by(pac) %>%
#   summarise(dt_inic = max(dt_sup),dt_sup =max(DATA_FINAL)) %>%
#   bind_rows(dados_faltantes,.) %>% arrange(pac)
#
# dados_final <- bind_rows(dados_completos,dados_faltantes_2)  %>%
#   group_by(pac) %>% mutate(TARGET = ifelse(dt_sup==max(dt_sup),max(TARGET,na.rm = TRUE),0))
#
# dados_input <- dados_final %>% group_by(pac) %>%
#   mutate_at(.vars=vars(DATA_FINAL,TGO,TGP,Ht,Leucocitos,Creatinina,Ht,Idade,SEXO),
#             .funs=list(~na.locf(., na.rm = FALSE))) %>% filter(dt_sup !=0) %>%
#   mutate_at(.vars=vars(DATA_FINAL,TGO,TGP,Ht,Leucocitos,Creatinina,Ht,Idade,SEXO),
#             .funs=list(~na.locf(., na.rm = FALSE,fromLast=TRUE)))
#
# dados_input <- dados_input%>%
#   group_by(pac) %>% mutate(TARGET = ifelse(dt_sup==max(dt_sup),
#       max(TARGET,na.rm = TRUE),0)) %>% ungroup %>%  mutate(dt_inic = ifelse(is.na(dt_inic),0,dt_inic)
```

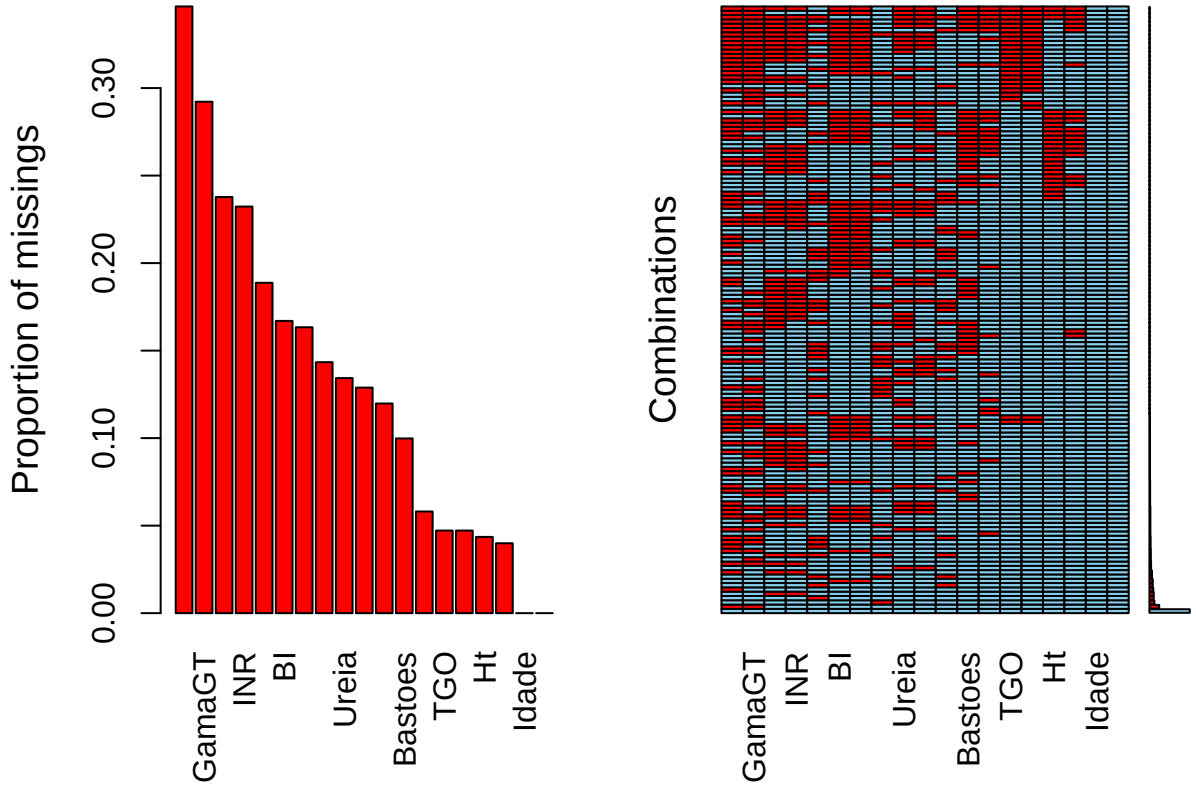
Análise de dados Faltantes

```
require(naniar)
summary_missing <- dados_class %>%ungroup %>%
  miss_var_summary()
not_missing <-summary_missing %>% filter(n_miss==0) %>% select(variable) %>% pull()
```

```
require(naniar)
require(FarctoMineR)
require(missMDA)
require(VIM)
gg_miss_var(dados_class %>% ungroup %>%
    select(-c(pac,dt_inic,dt_sup,DATA_FINAL,TARGET,SEXO)))
```




```
res<-summary(aggr(dados_class %>% ungroup %>%
    select(-c(pac,dt_inic,dt_sup,DATA_FINAL,TARGET,SEXO)), sortVar=TRUE))$combinations
```




```
##
##  Variables sorted by number of missings:
##         Variable      Count
##  Fosfatase_alcal 0.34664247
##           GamaGT 0.29219601
##              TAP 0.23774955
##              INR 0.23230490
##      SINAL_FAGET 0.18874773
##               BI 0.16696915
##               BD 0.16333938
##        ICTERICIA 0.14337568
##            Ureia 0.13430127
##       Creatinina 0.12885662
##       HEMORRAGIA 0.11978221
##          Bastoes 0.09981851
##         Plaquetas 0.05807623
##              TGO 0.04718693
##              TGP 0.04718693
##               Ht 0.04355717
##       Leucocitos 0.03992740
##            Idade 0.00000000
##     dt_ate_inter 0.00000000
```

Devemos evitar utilizar variáveis com muitos valores faltantes

#Inputação de valores utilizando substituindo o valor faltante pelo registro do dia anterior.

Caso o registro do dia anterior também esteja faltante, o registro do dia posterior é utilizado.

```
dados_class_fill <- dados_class %>%group_by(pac) %>%
  fill(c("Plaquetas","Ht",'TGP','Ureia','TGO','Creatinina','ICTERICIA','BD',
         'BI','TAP','INR','Leucocitos'),.direction="updown")
```

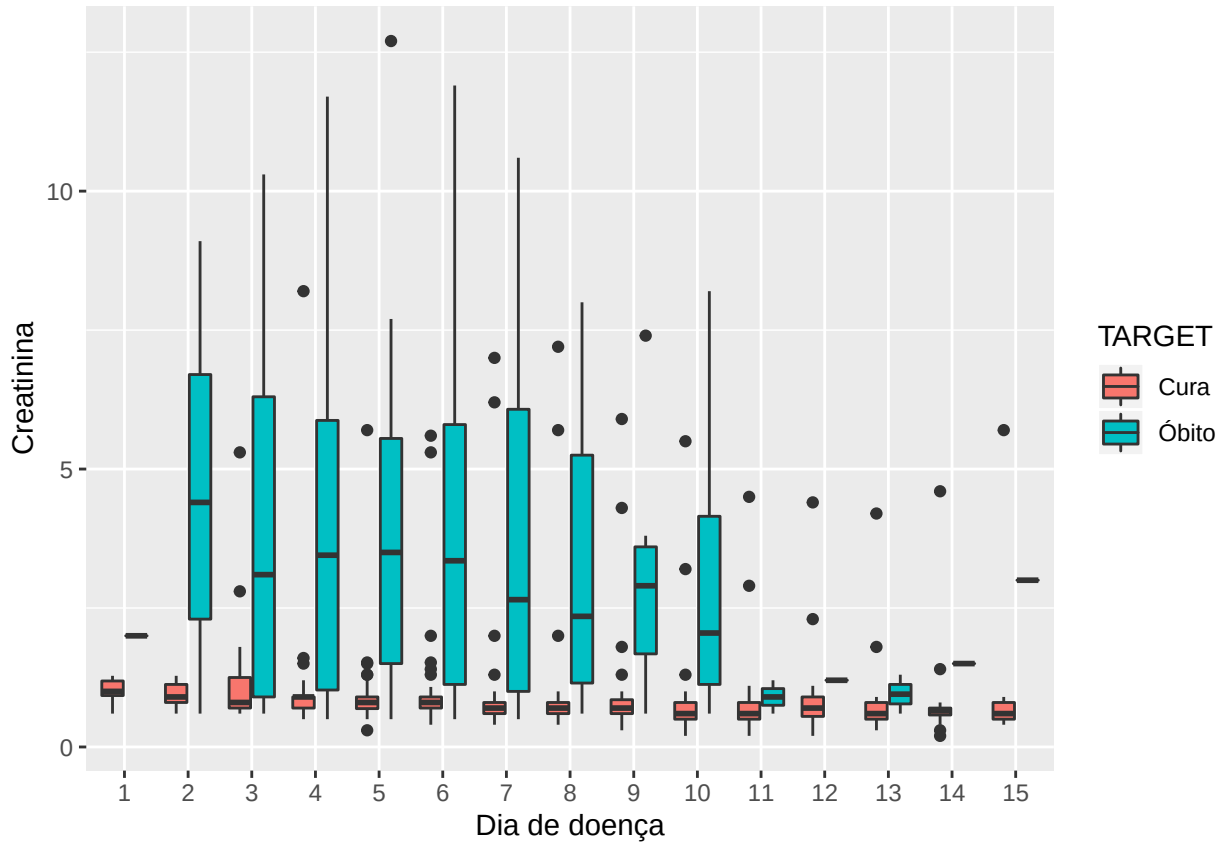
Dados Agrupados

```
media_na <- function(x){mean(x,na.rm=TRUE)}
dados_agrupados <- dados_class_fill %>% group_by(pac) %>% summarise(dt_inic=min(dt_inic),
dt_sup=max(dt_sup))
```

Estatísticas resumo após imputação

Gráfico Creatinina

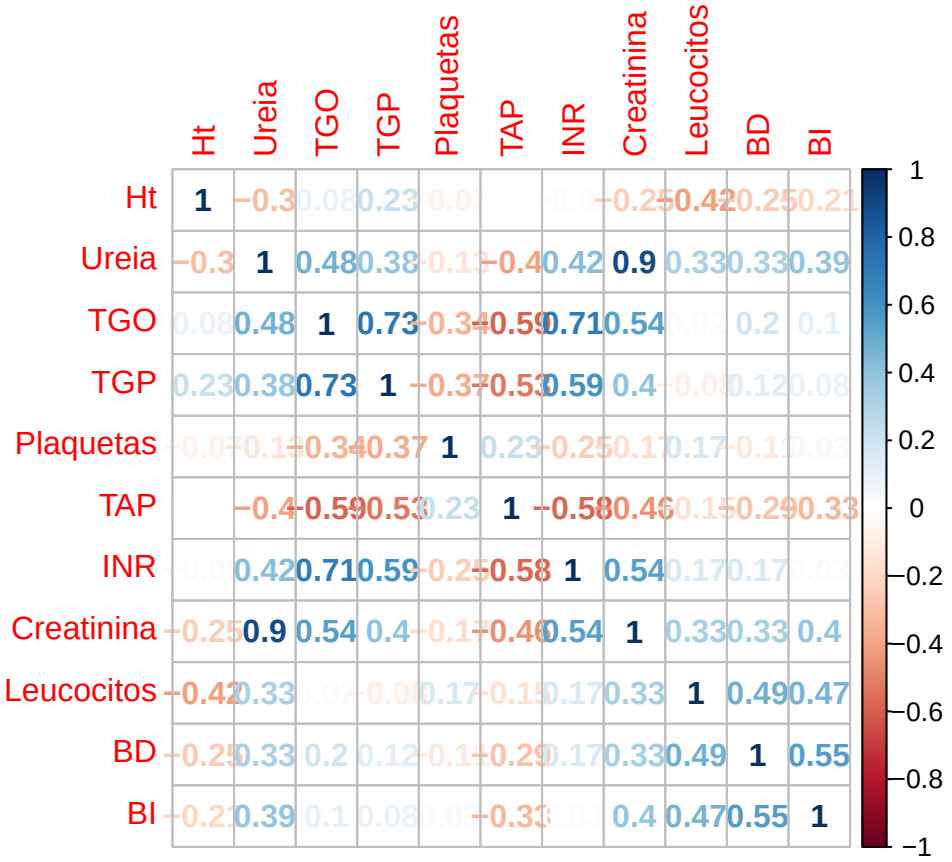
```
ggplot(data=dados_plot %>% filter(dt_sup>0),aes(y=Creatinina,fill=TARGET,x=as.factor(dt_sup)))+
  geom_boxplot()+xlab("Dia de doença")
```



Correlações

```
require(corrplot)
M <- cor(dados_class_fill %>% ungroup%>%select(Ht,Ureia,TGO,TGP,Plaquetas,TAP,INR,
```

```
                                                Creatinina,Leucocitos,BD,BI)%>% drop_na()  )
corrplot(M, method = "number")
```



Ainda existem alguns dados faltantes de ICTERICIA que não podem ser inputados pois não existe o regristro anível de paciente.

Análise longitudinal dos dados (GEE)

O método de equações de estimação generalizado é uma alternativa para a modelagem de dados longitudinais.

Essa técnica de modelagem requer a indicação de uma estrutura de correlação entre o tempo e estima o efeito médio populacional das variáveis explicativas levando em consideração a correlação interna de cada paciente.

O algorítmo obtem a estimativa dos coeficientes β através da maximização da equação abaixo em β:

$$U(\tilde{\beta}) = \sum_{i=1}^{n} \frac{\partial \mu_{ij}}{\partial \beta_p} V_i^{-1} \{Y_i - \mu_i(\tilde{\beta})\}$$

Onde:

- μ_{ij} é o valor da variável explicativa do indivíduo i no tempo j.
- $\beta_p,\ p = 1, \ldots, k$ são os coeficientes das variáveis explicativas
- V_i é a estrutura de correlação do tempo.

```
require(geepack)
dados_scaled <- dados2  %>% mutate(Plaquetas=log2(Plaquetas),Leucocitos=log2(Leucocitos),
                                   TGO=log2(TGO),TGP=log2(TGP)) %>% select(-dt_inic) %>%
  ungroup %>% mutate(pac = as.factor(pac)) %>%
  select(Ht,TGO,pac,dt_sup,Plaquetas,TARGET,ICTERICIA,dt_ate_inter) %>%
  group_by(pac) %>% mutate(number = row_number()) %>%
  drop_na()

ajuste_gee<- geeglm(TARGET~Plaquetas+Ht+TGO+ICTERICIA,id=pac, data=dados_scaled,
family=binomial(link='logit'),waves = dt_sup,corstr='independence', std.err="san.se")


summary(ajuste_gee)
```

```
##
## Call:
## geeglm(formula = TARGET ~ Plaquetas + Ht + TGO + ICTERICIA, family = binomial(link = "logit"),
##     data = dados_scaled, id = pac, waves = dt_sup, corstr = "independence",
##     std.err = "san.se")
##
##  Coefficients:
##               Estimate   Std.err   Wald Pr(>|W|)
## (Intercept)  -18.38082   6.78390  7.341  0.00674 **
## Plaquetas      0.70725   0.39247  3.247  0.07154 .
## Ht            -0.12228   0.05788  4.463  0.03463 *
## TGO            0.88783   0.16658 28.407 9.83e-08 ***
## ICTERICIASIM   1.12288   0.59485  3.563  0.05907 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Estimated Scale Parameters:
##             Estimate Std.err
## (Intercept)   0.9209   1.831
##
## Correlation: Structure = independenceNumber of clusters:   83   Maximum cluster size: 16
```

Interpretação das variávies

Estimativas Negativas indicam diminuição na razão de chances de óbito e estimativas prositivas indicam aumento, mesma interpretação da regressão logística.

- A cada vez que a contangem de Plaquetas dobra a razão de chances de óbito aumenta exp(0.7073)=2.03 vezes
- O aumento de uma unidade de HT aumenta a razão de chances de cura em exp(-0.1223)=0.885 vezes.
- A cada vez que a contangem de TGO dobra a razão de chances de óbito aumenta exp(0.8878)=2.43 vezes
- A presença do sintoma de Icterícia aumenta a razão de chances de óbito em exp(1.1229)=3.07 vezes.

As demais variáveis não se mostraram significativas

Análise da predição

```
prob=predict(ajuste_gee,type=c("response"))

dados_roc <- dados_scaled %>%ungroup %>%
  select(TARGET,Plaquetas,Ht,TGO,ICTERICIA) %>%
mutate(p= prob)
require(pROC)
g <- auc(TARGET ~ p, data = dados_roc)
g
```

```
## Area under the curve: 0.901
```

```
dados_roc <- dados_roc %>% ungroup %>% mutate(predicao = ifelse(p>0.5,1,0))

confusionMatrix(dados_roc$predicao %>% as.factor,dados_roc$TARGET%>% as.factor,positive='1')
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 360  40
##          1  20  71
##
##                Accuracy : 0.878
##                  95% CI : (0.846, 0.905)
##     No Information Rate : 0.774
##     P-Value [Acc > NIR] : 2.96e-09
##
##                   Kappa : 0.627
##
##  Mcnemar's Test P-Value : 0.0142
##
##             Sensitivity : 0.640
##             Specificity : 0.947
##          Pos Pred Value : 0.780
##          Neg Pred Value : 0.900
##              Prevalence : 0.226
##          Detection Rate : 0.145
##    Detection Prevalence : 0.185
```

```
##       Balanced Accuracy : 0.794
##
##        'Positive' Class : 1
##
```

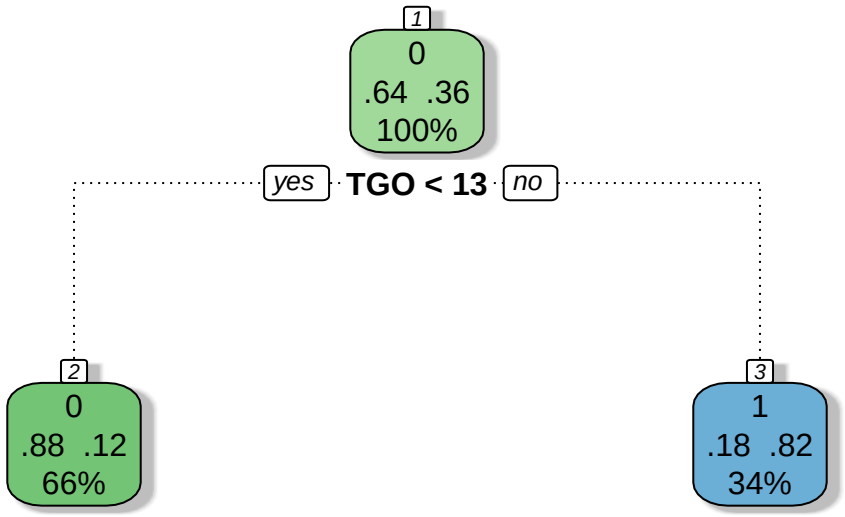
É importante notar que apesar da acurácia não ser extremamente alta, estamos falando da acurária de predição para todos os registros observados. Em alguns períodos de doença a diferenciação de indivíduos pode ser mais complexa.

Árvore de Decisão

Podemos nos basear no primeiro registro observado para construir uma árvore de decisão para tentar prever o óbito.

```
require(rpart)
require(rattle)
dados_primeira_obs=dados2  %>% mutate(Plaquetas=log2(Plaquetas),Leucocitos=log2(Leucocitos),
                                      TGO=log2(TGO),TGP=log2(TGP)) %>% select(-dt_inic) %>% group_by(pa
  filter(dt_sup==min(dt_sup))

Arvore = rpart(TARGET~ Ht + TGO + ICTERICIA  +Plaquetas, data=dados_primeira_obs,maxdepth=6,method='cla
fancyRpartPlot(Arvore)
```



Rattle 2019−dez−11 01:10:52 gusta

```
t_pred <- predict(Arvore,dados_primeira_obs)[,1]


confusionMatrix(ifelse(t_pred<.5,1,0) %>% as.factor,dados_primeira_obs$TARGET %>% as.factor,positive='1
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 58  8
##          1  6 28
##
##                Accuracy : 0.86
##                  95% CI : (0.776, 0.921)
##     No Information Rate : 0.64
##     P-Value [Acc > NIR] : 8.06e-07
##
##                   Kappa : 0.692
##
##  Mcnemar's Test P-Value : 0.789
##
##             Sensitivity : 0.778
##             Specificity : 0.906
##          Pos Pred Value : 0.824
##          Neg Pred Value : 0.879
##              Prevalence : 0.360
##          Detection Rate : 0.280
##    Detection Prevalence : 0.340
##       Balanced Accuracy : 0.842
##
##        'Positive' Class : 1
##
```

O resultado baseando-se apenas em uma variável já tem uma discriminação aceitável, tem em vista que 88% dos casos que não vão a óbito tem o valor do TGO inferior a 2^13 = 8192 e 82% dos casos de óbito te valor do TGO superior a 8192