Técnicas de Reamostragem

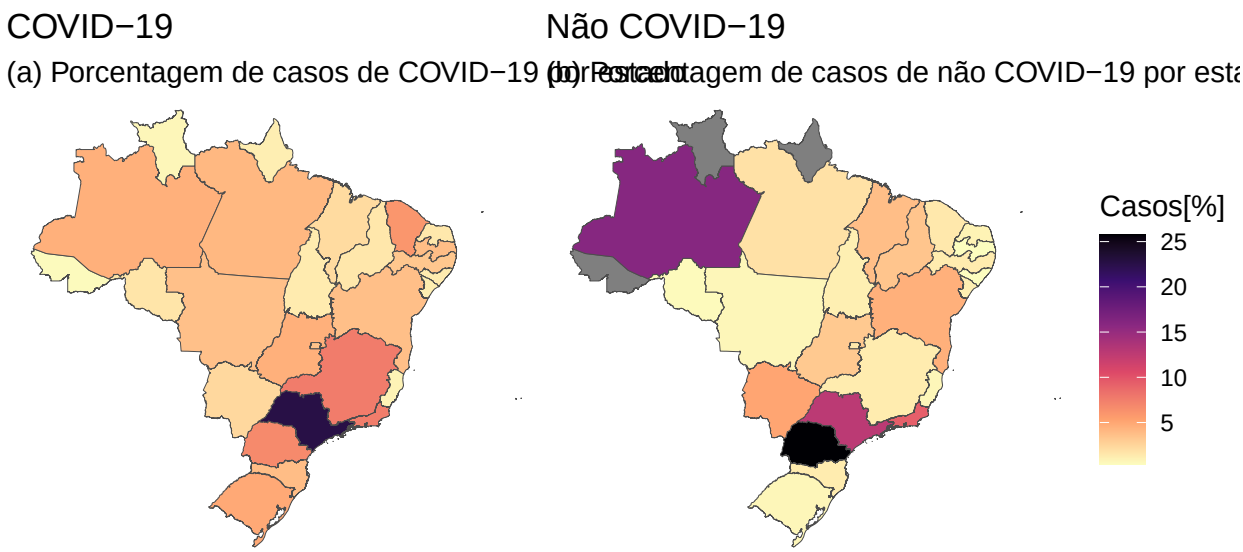
Métodos de Downsample

```
##   |                                                                      |
```

```
## Frequencies
## dados$CLASSI_FIN
## Type: Numeric
##
##                Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
## ----------- ------- --------- -------------- --------- --------------
##           1    1521     4.063          4.063     3.912          3.912
##           2     601     1.606          5.669     1.546          5.458
##           3     108     0.289          5.958     0.278          5.736
##           4   16048    42.874         48.831    41.279         47.015
##           5   19136    51.123         99.955    49.222         96.237
##           9      17     0.045        100.000     0.044         96.281
##        <NA>    1446                             3.719        100.000
##       Total   38877   100.000        100.000   100.000        100.000
```

Casos por estado COVID-19

Casos por estado não COVID-19

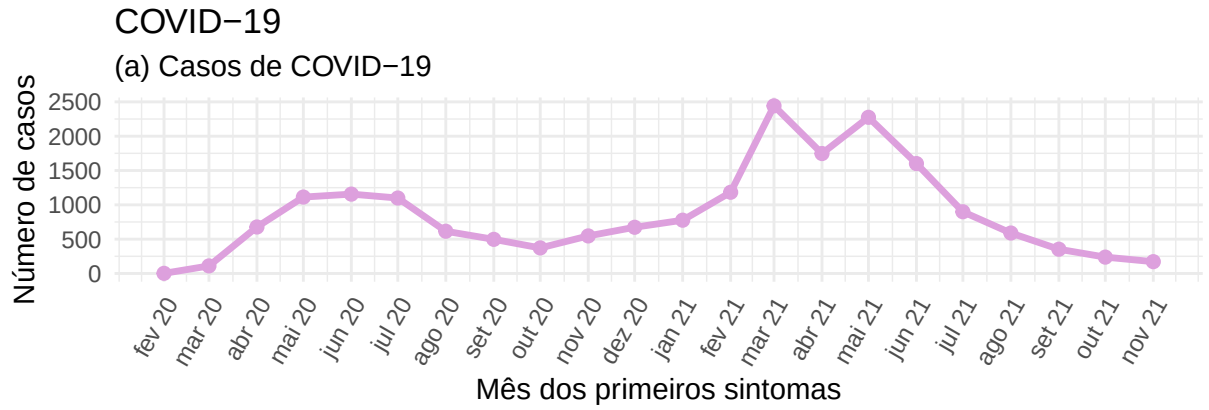


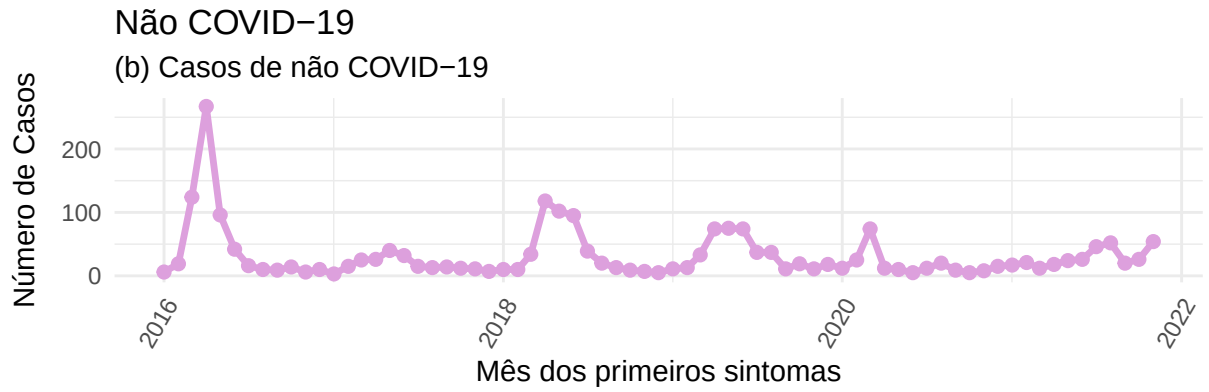
(a) Porcentagem de casos de COVID−19 por estado

(b) Porcentagem de casos de não COVID−19 por esta

Casos Covid-19 por data de primeiros sintomas

Casos não Covid por data de primeiros sintomas






```
##
##  Welch Two Sample t-test
##
## data:  idade by classi_fin
## t = 15.579, df = 2803.3, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group COVID-19 and group Não COVID-19 is no
## 95 percent confidence interval:
##  2.154071 2.774357
## sample estimates:
##     mean in group COVID-19 mean in group Não COVID-19
##                   29.90233                   27.43812
```

```
##
##  Welch Two Sample t-test
##
## data:  tempo_sintomas_notific by classi_fin
## t = 6.4256, df = 2848, p-value = 1.534e-10
## alternative hypothesis: true difference in means between group COVID-19 and group Não COVID-19 is no
## 95 percent confidence interval:
##  2.044827 3.840872
## sample estimates:
##     mean in group COVID-19 mean in group Não COVID-19
##                  10.261235                   7.318386
```

Remonstragem com dados numéricos

DownSample

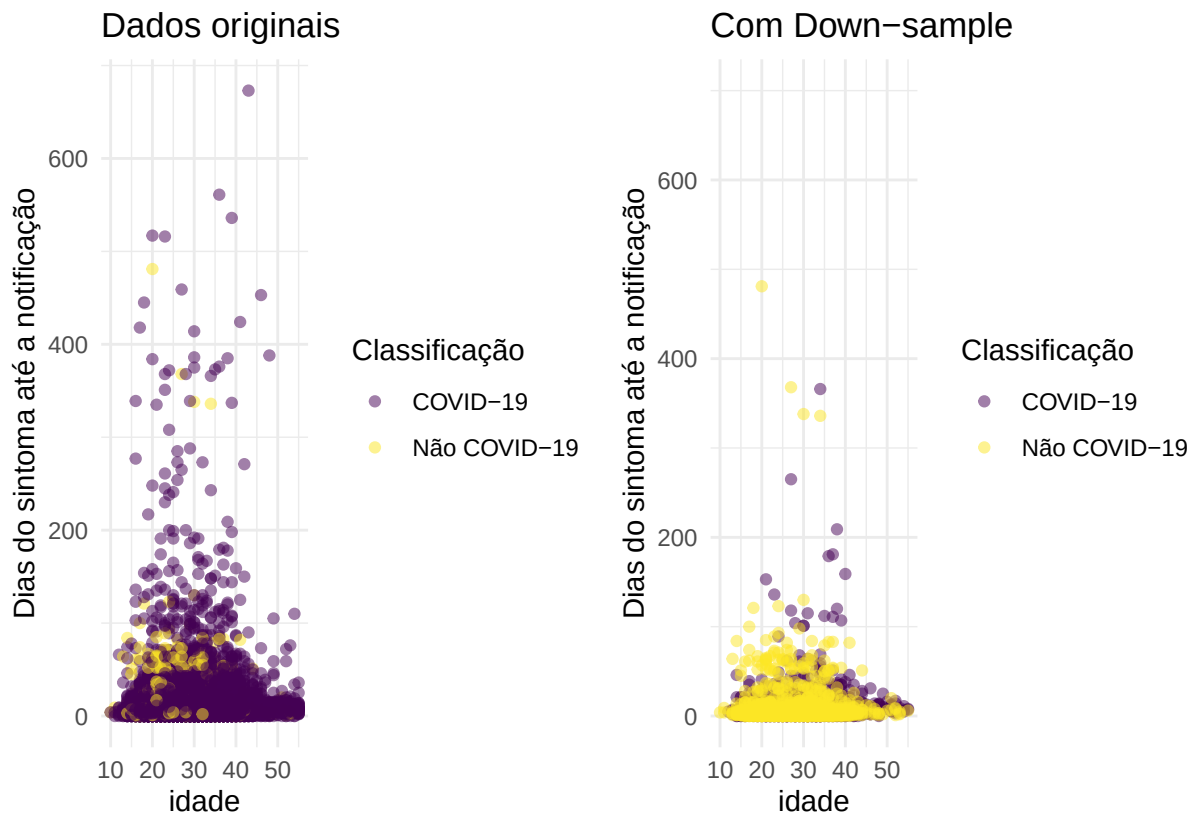
```
data <- data1 %>%
  select(idade, tempo_sintomas_notific,classi_fin) %>%
  drop_na() %>%
  as.data.frame()

g5 <- ggplot(data, aes(idade, tempo_sintomas_notific,color = classi_fin)) +
  geom_point(alpha=0.5) +
  labs(title = "Dados originais") + ylab("Dias do sintoma até a notificação") +
  scale_colour_viridis_d() + labs(color="Classificação")

dados3 <- recipe(classi_fin ~ idade + tempo_sintomas_notific, data = data) %>%
  step_downsample(classi_fin,seed=69) %>%
  prep() %>%
  bake(new_data = NULL)

g6 <-  dados3 %>%
  ggplot(aes(idade, tempo_sintomas_notific, color = classi_fin)) +
  geom_point(alpha=0.5) +
  labs(title = "Com Down-sample") + ylab("Dias do sintoma até a notificação") +
  scale_colour_viridis_d() + labs(color="Classificação") +
  scale_y_continuous(limits = c(0,700))

g5|g6
```




```
datasummary((classi_fin) ~  idade*(n+media+DP+mediana+q25+q75+IQR),
            data = dados3, output = 'markdown')
```

	n	media	DP	mediana	q25	q75	IQR
COVID-19	2230.00	29.98	7.40	30.00	25.00	35.00	10.00
Não COVID-19	2230.00	27.44	7.05	27.00	22.00	32.00	10.00

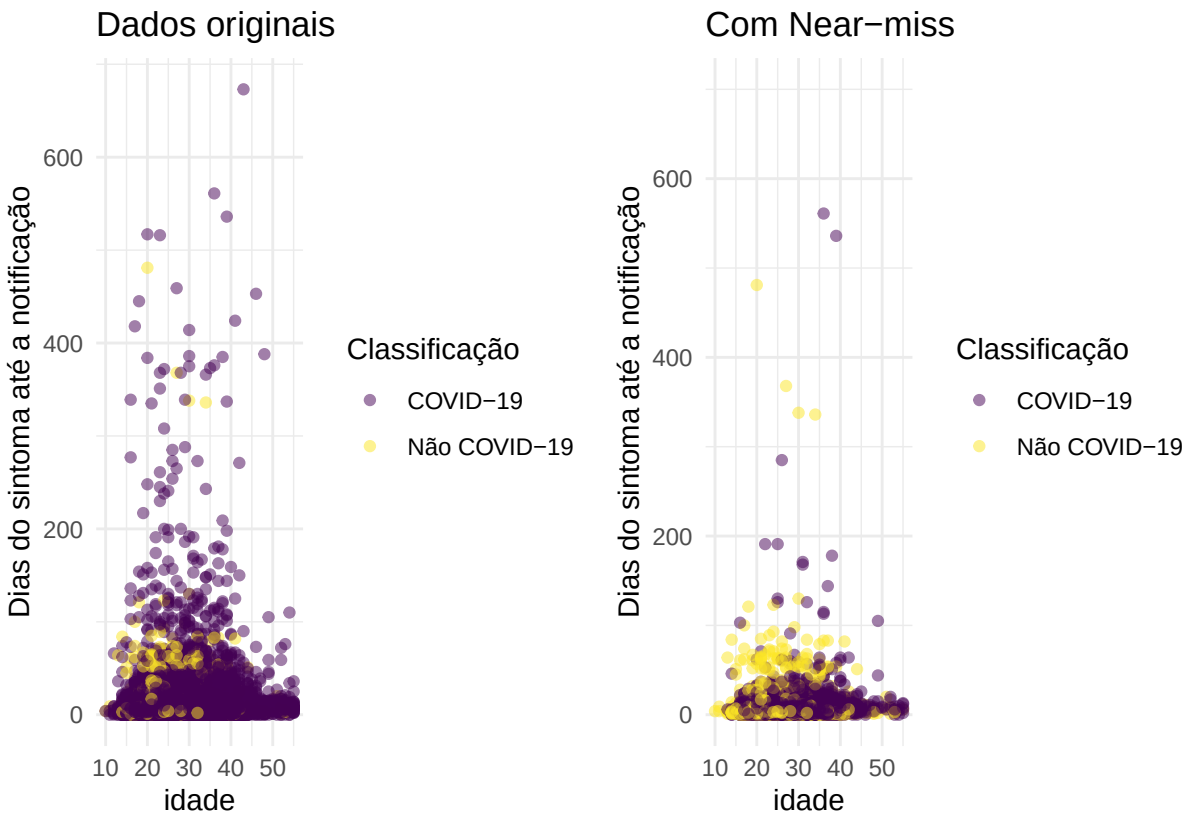
```
datasummary((classi_fin) ~  tempo_sintomas_notific*(n+media+DP+mediana+q25+q75+IQR),
            data = dados3, output = 'markdown')
```

	n	media	DP	mediana	q25	q75	IQR
COVID-19	2230.00	9.66	16.06	7.00	4.00	11.00	7.00
Não COVID-19	2230.00	7.32	20.33	3.00	2.00	6.00	4.00

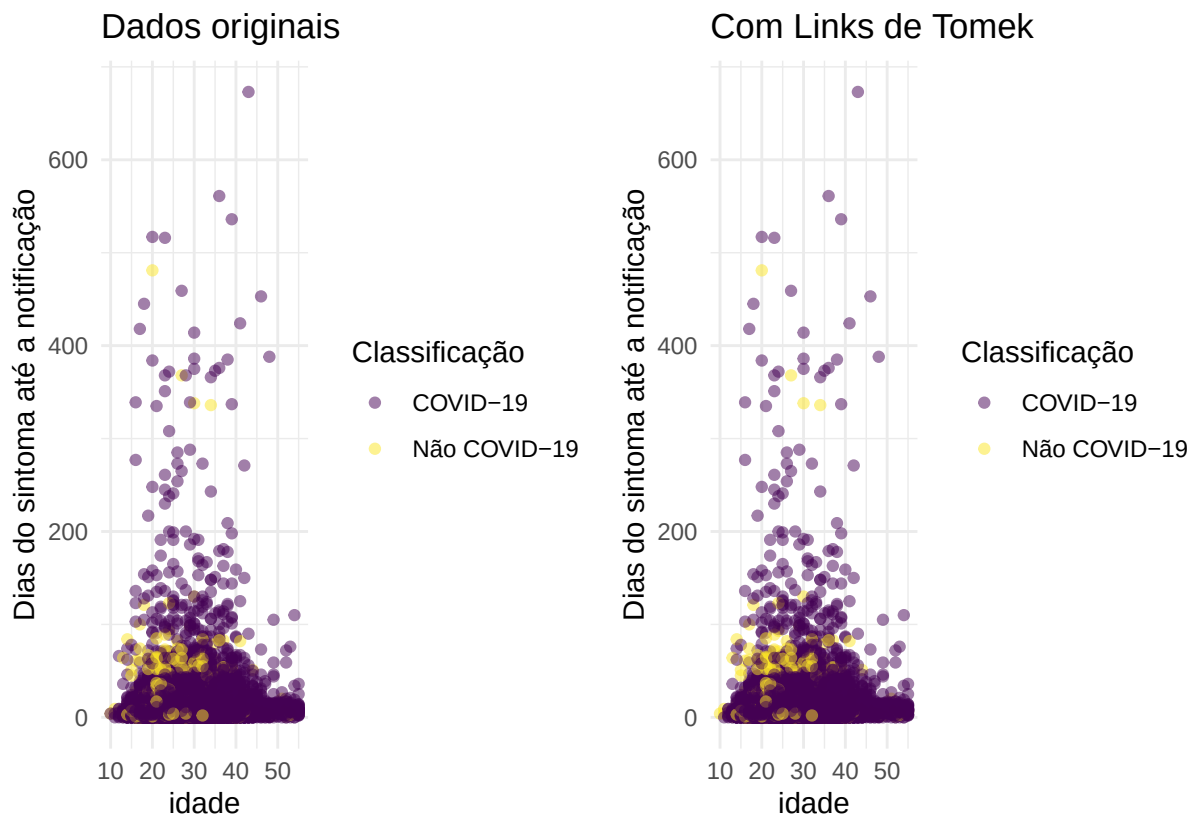
NearMiss

	n	media	DP	mediana	q25	q75	IQR
COVID-19	2230.00	29.99	7.16	30.00	25.00	35.00	10.00
Não COVID-19	2230.00	27.44	7.05	27.00	22.00	32.00	10.00

	n	media	DP	mediana	q25	q75	IQR
COVID-19	2230.00	9.90	21.78	7.00	4.00	11.00	7.00
Não COVID-19	2230.00	7.32	20.33	3.00	2.00	6.00	4.00



Links de Tomek



	n	media	DP	mediana	q25	q75	IQR
COVID-19	18936.00	29.89	7.20	30.00	25.00	35.00	10.00
Não COVID-19	2230.00	27.44	7.05	27.00	22.00	32.00	10.00

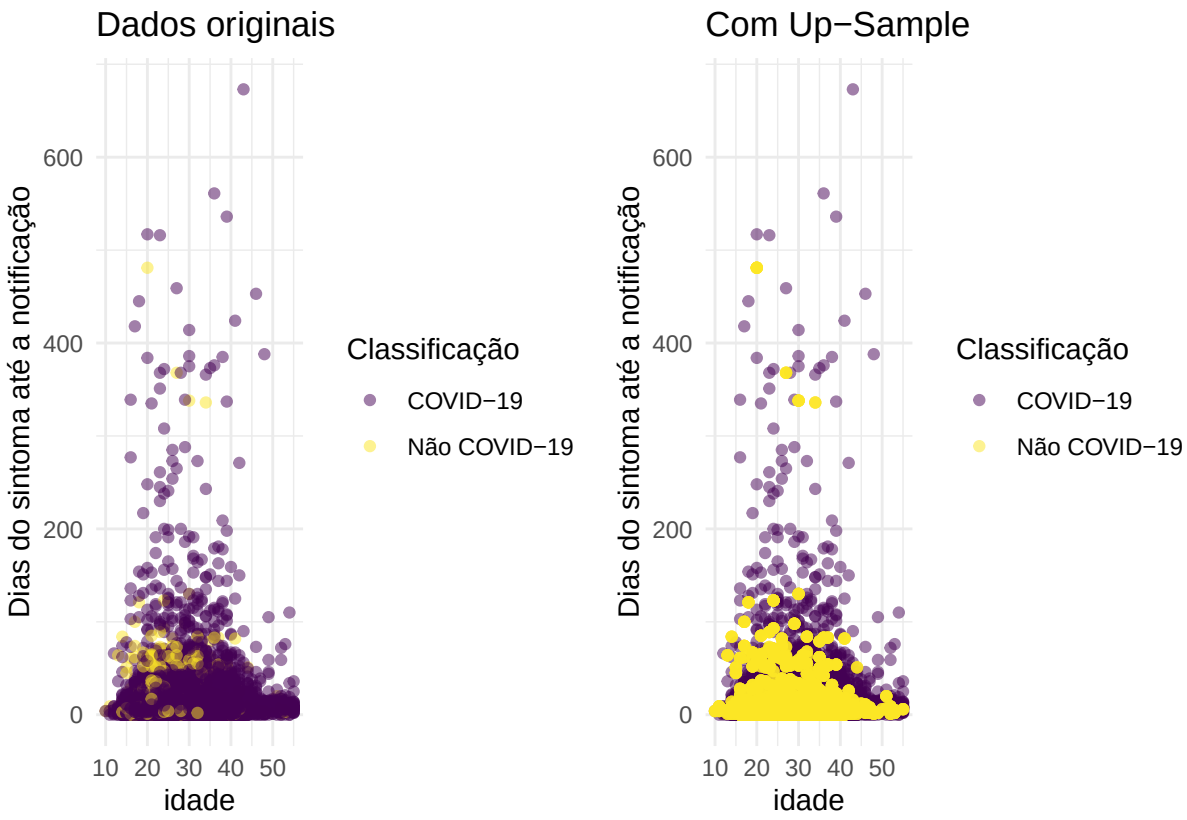
	n	media	DP	mediana	q25	q75	IQR
COVID-19	18936.00	10.11	21.28	7.00	4.00	11.00	7.00
Não COVID-19	2230.00	7.32	20.33	3.00	2.00	6.00	4.00

Métodos de Up-Sample

Upsample

	n	media	DP	mediana	q25	q75	IQR
COVID-19	19136.00	29.90	7.23	30.00	25.00	35.00	10.00
Não COVID-19	19136.00	27.46	7.04	27.00	22.00	32.00	10.00

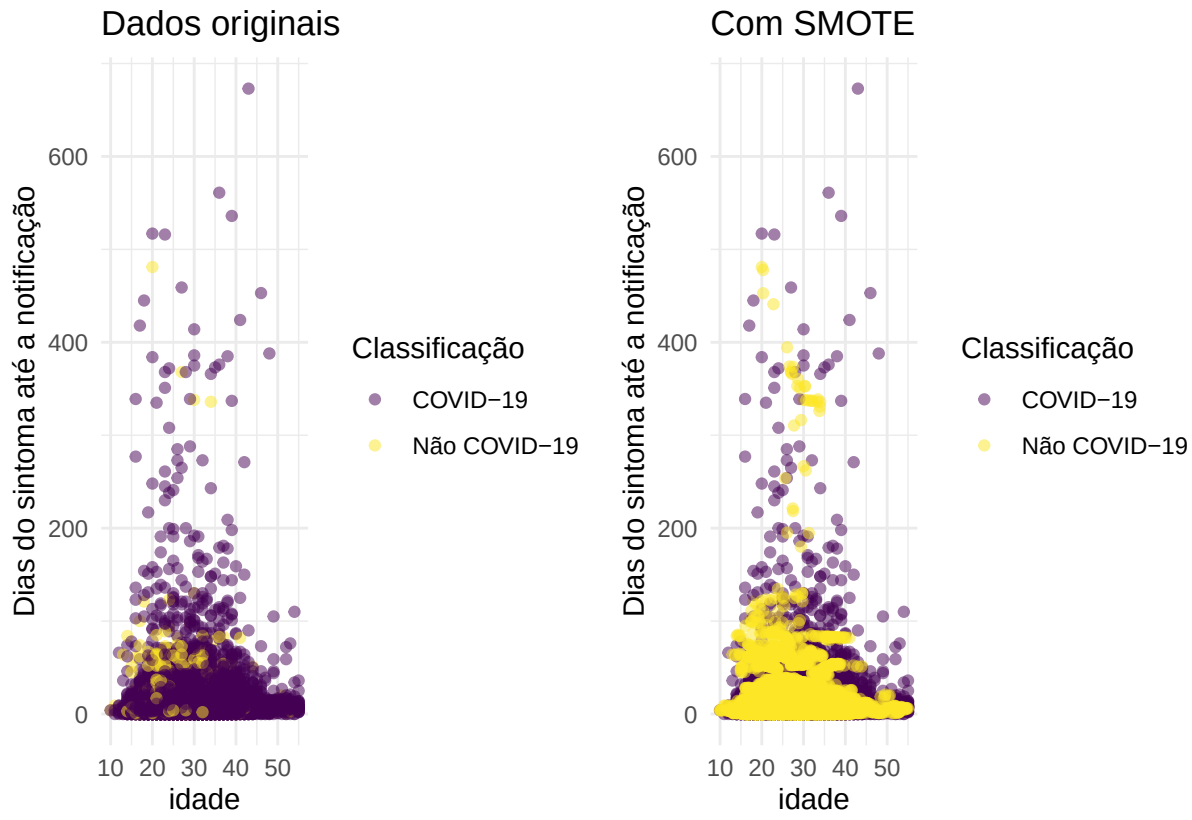
	n	media	DP	mediana	q25	q75	IQR
COVID-19	19136.00	10.26	21.60	7.00	4.00	11.00	7.00
Não COVID-19	19136.00	7.18	19.41	3.00	2.00	6.00	4.00



Smote

	n	media	DP	mediana	q25	q75	IQR
COVID-19	19136.00	29.90	7.23	30.00	25.00	35.00	10.00
Não COVID-19	19136.00	27.43	6.99	27.00	22.00	32.22	10.22

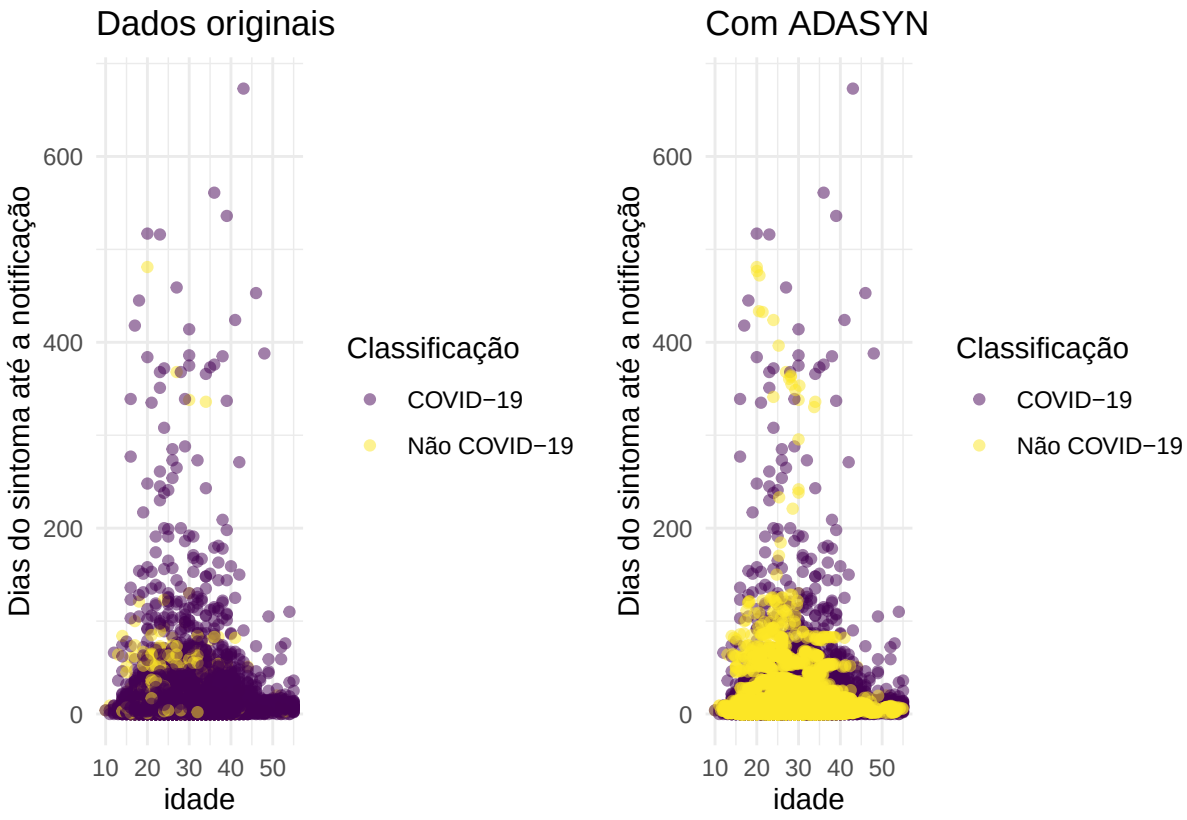
	n	media	DP	mediana	q25	q75	IQR
COVID-19	19136.00	10.26	21.60	7.00	4.00	11.00	7.00
Não COVID-19	19136.00	7.21	18.92	3.00	2.00	6.00	4.00



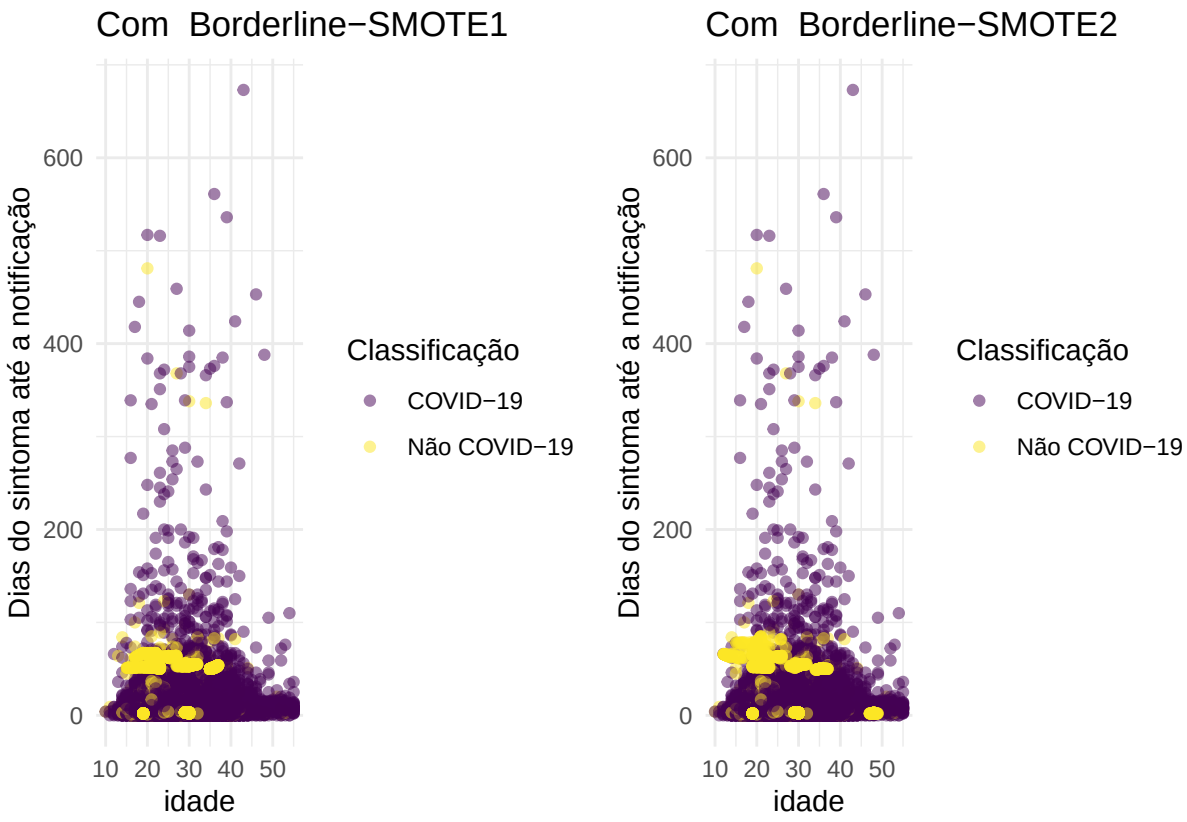
ADASYN

	n	media	DP	mediana	q25	q75	IQR
COVID-19	19136.00	29.90	7.23	30.00	25.00	35.00	10.00
Não COVID-19	19136.00	27.62	6.97	27.00	22.00	33.00	11.00

	n	media	DP	mediana	q25	q75	IQR
COVID-19	19136.00	10.26	21.60	7.00	4.00	11.00	7.00
Não COVID-19	19136.00	6.76	16.90	3.81	2.00	6.00	4.00


Dados originais
Dias do sintoma até a notificação
600
400
200
0
10 20 30 40 50
idade
Classificação
COVID−19
Não COVID−19
Com ADASYN
Dias do sintoma até a notificação
600
400
200
0
10 20 30 40 50
idade
Classificação
COVID−19
Não COVID−19

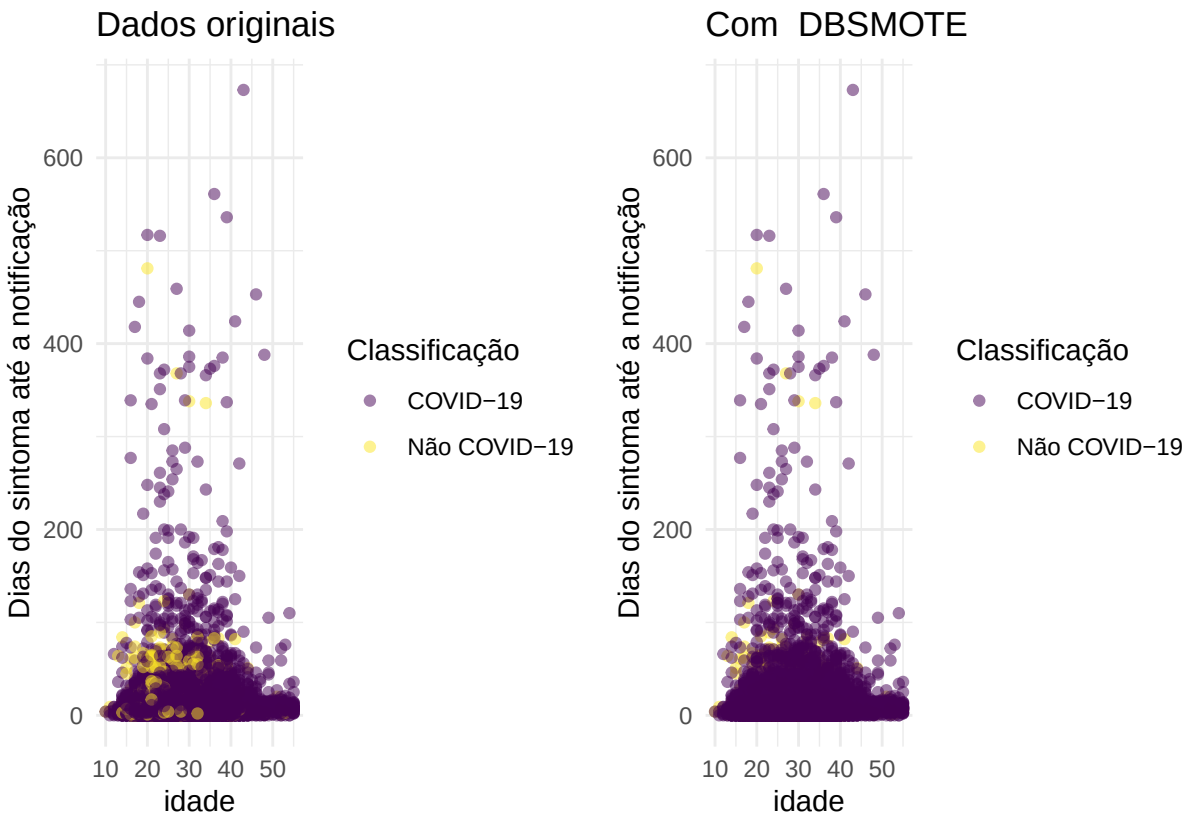
Borderline Smote



	n	media	DP	mediana	q25	q75	IQR
COVID-19	19136.00	29.90	7.23	30.00	25.00	35.00	10.00
Não COVID-19	19136.00	27.82	4.87	30.00	29.00	30.00	1.00

	n	media	DP	mediana	q25	q75	IQR
COVID-19	19136.00	10.26	21.60	7.00	4.00	11.00	7.00
Não COVID-19	19136.00	5.66	13.63	3.00	2.00	3.00	1.00

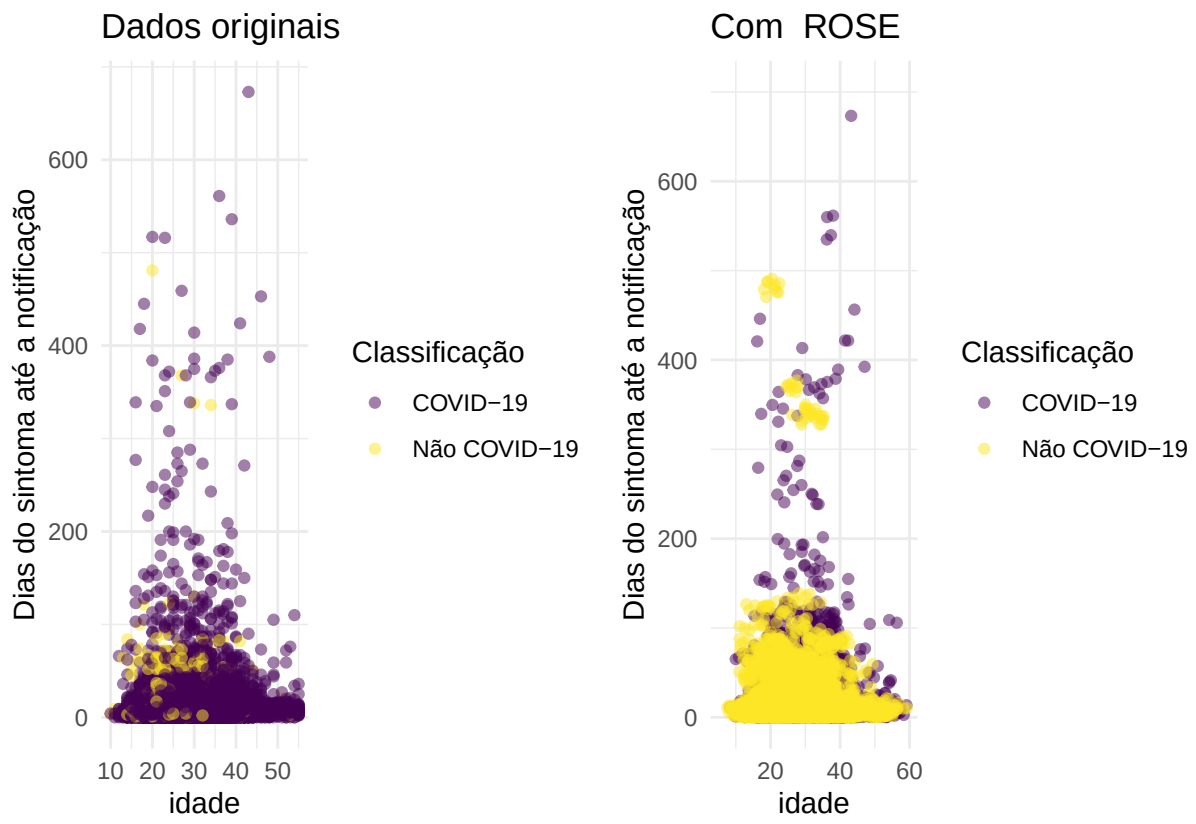
DBSMOTE



	n	media	DP	mediana	q25	q75	IQR
COVID-19	19136.00	29.90	7.23	30.00	25.00	35.00	10.00
Não COVID-19	18254.00	27.13	4.58	27.00	24.78	29.88	5.10

	n	media	DP	mediana	q25	q75	IQR
COVID-19	19136.00	10.26	21.60	7.00	4.00	11.00	7.00
Não COVID-19	18254.00	3.84	7.47	3.00	3.00	3.06	0.06

ROSE



	n	media	DP	mediana	q25	q75	IQR
COVID-19	19317.00	29.86	7.36	29.84	24.51	35.05	10.54
Não COVID-19	18955.00	27.34	7.32	26.97	21.76	32.49	10.72

	n	media	DP	mediana	q25	q75	IQR
COVID-19	19317.00	10.09	21.69	7.39	3.01	12.31	9.30
Não COVID-19	18955.00	7.62	22.56	4.25	-0.20	9.25	9.45

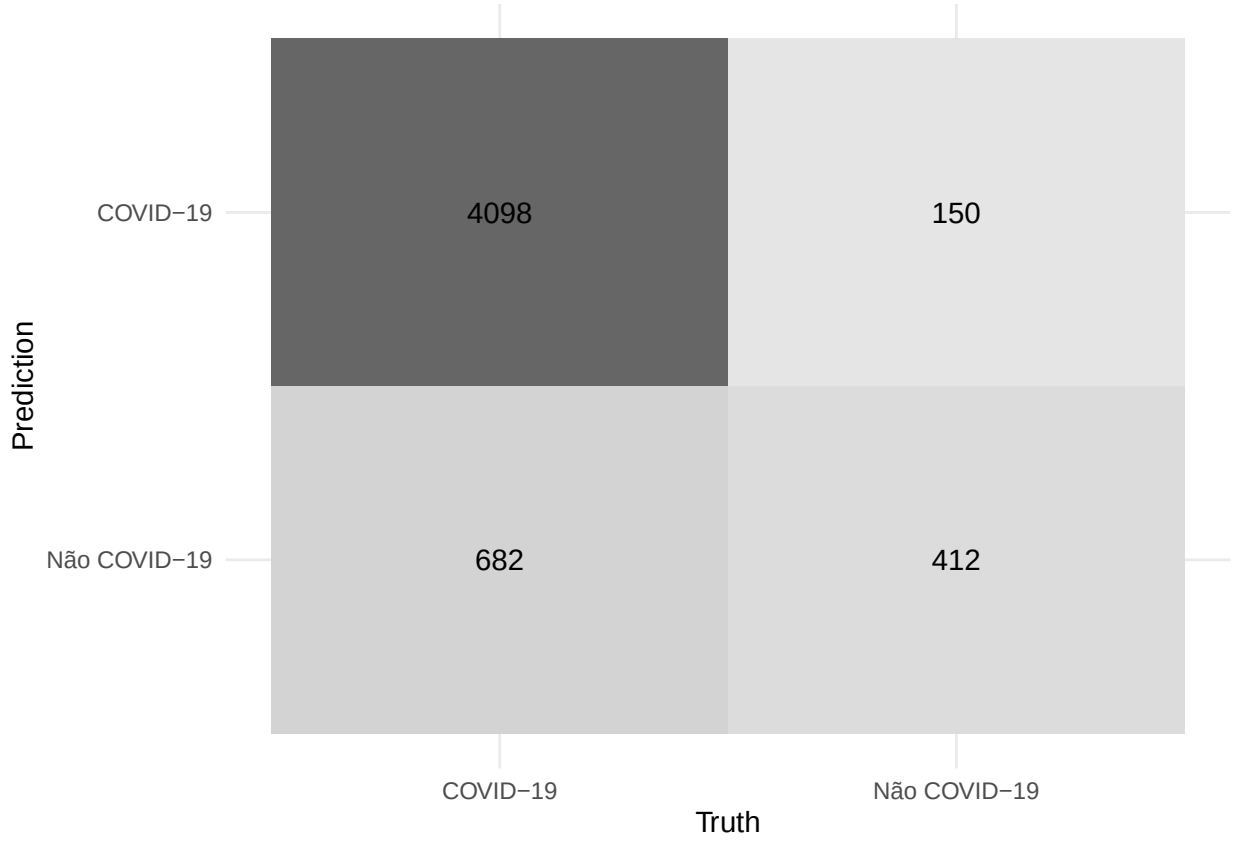
Reamostragem com dados numéricos e categóricos com aplicação do modelo logístico

Dados

Smote NC

```
## # A tibble: 5 x 6
##   .metric     .estimator  mean     n std_err .config
##   <chr>       <chr>      <dbl> <int>   <dbl> <chr>
## 1 npv         binary     0.379    10 0.00756 Preprocessor1_Model1
## 2 ppv         binary     0.967    10 0.00137 Preprocessor1_Model1
```

```
## 3 roc_auc     binary     0.880    10 0.00509 Preprocessor1_Model1
## 4 sensitivity binary     0.857    10 0.00359 Preprocessor1_Model1
## 5 specificity binary     0.750    10 0.00943 Preprocessor1_Model1
```

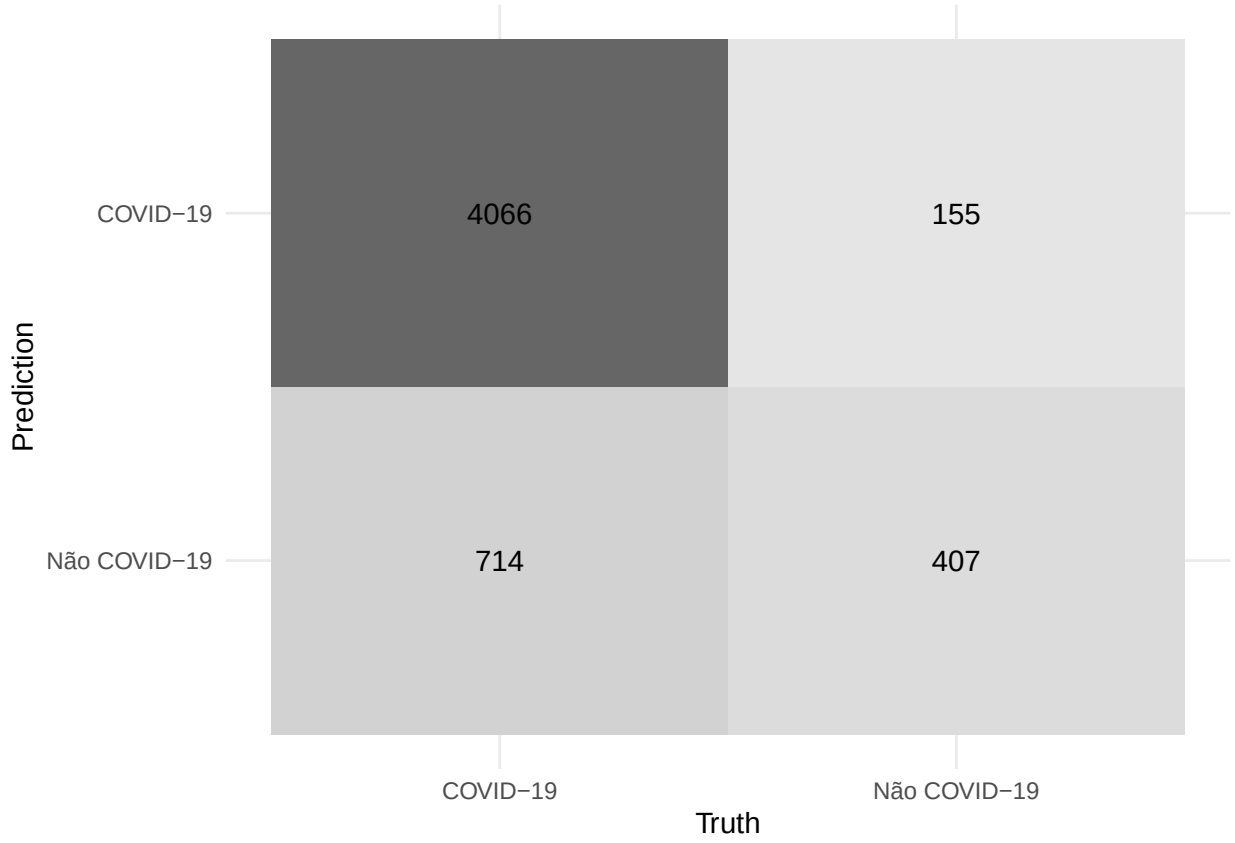


.metric	.estimate
accuracy	0.8443
kap	0.4165
sens	0.8573
spec	0.7331
ppv	0.9647
npv	0.3766
mcc	0.4489
j_index	0.5904
bal_accuracy	0.7952
detection_prevalence	0.7952
precision	0.9647
recall	0.8573
f_meas	0.9078

Adasyn distancia HEOM

```
## # A tibble: 5 x 6
##   .metric     .estimator  mean     n std_err .config
##   <chr>       <chr>      <dbl> <int>   <dbl> <chr>
```

```
## 1 npv         binary     0.856    10 0.00335 Preprocessor1_Model1
## 2 ppv         binary     0.842    10 0.00390 Preprocessor1_Model1
## 3 roc_auc     binary     0.927    10 0.00134 Preprocessor1_Model1
## 4 sensitivity binary     0.858    10 0.00249 Preprocessor1_Model1
## 5 specificity binary     0.840    10 0.00369 Preprocessor1_Model1
```

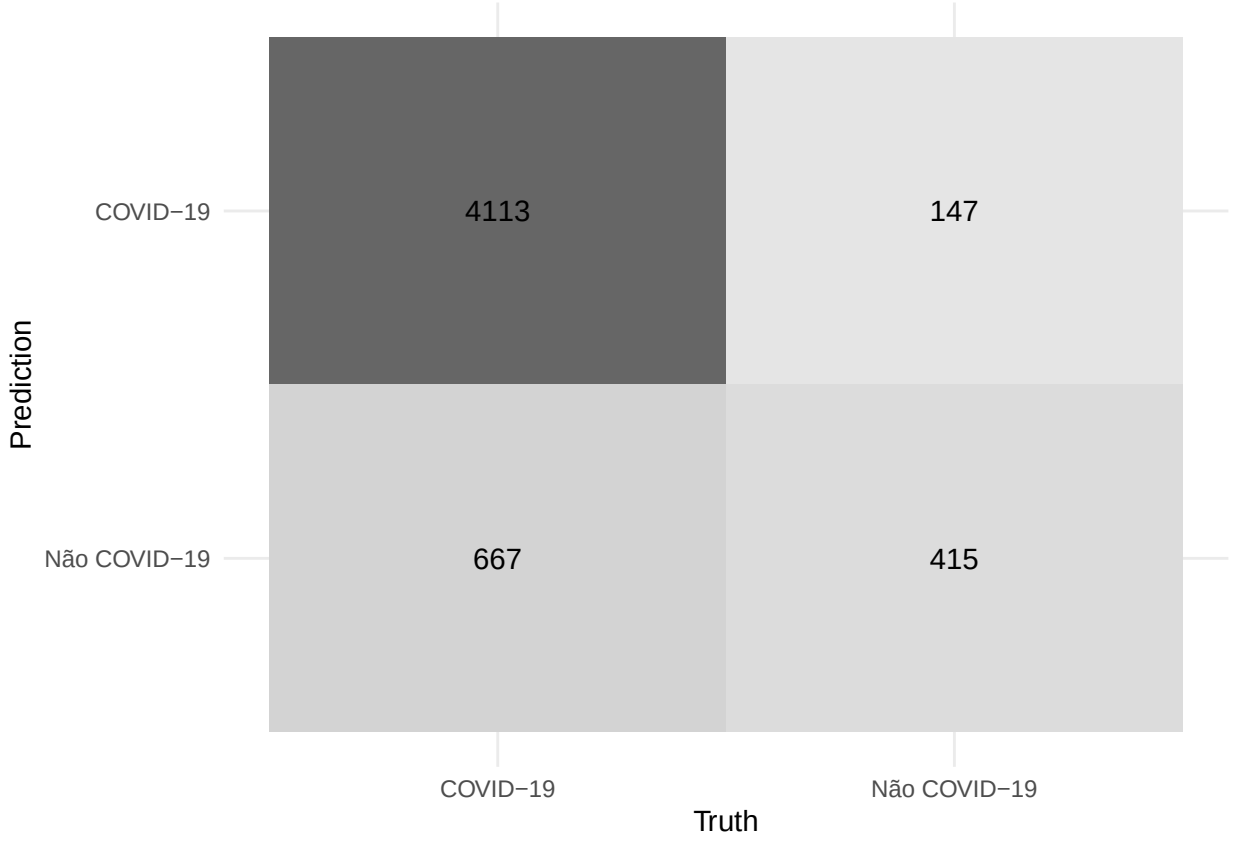


.metric	.estimate
accuracy	0.8373
kap	0.3995
sens	0.8506
spec	0.7242
ppv	0.9633
npv	0.3631
mcc	0.4331
j_index	0.5748
bal_accuracy	0.7874
detection_prevalence	0.7902
precision	0.9633
recall	0.8506
f_meas	0.9035

Smote distencia HEOM

```
## # A tibble: 5 x 6
```

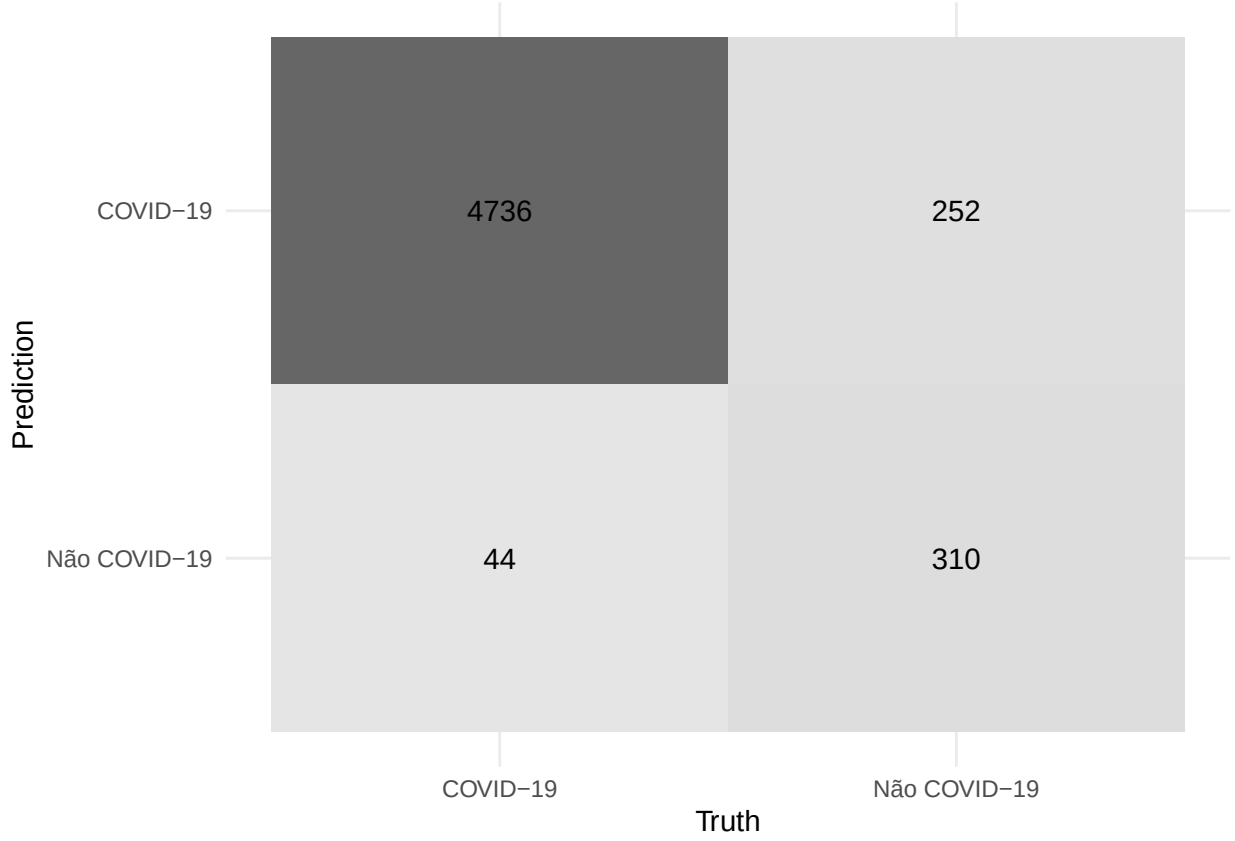
```
##   .metric     .estimator  mean     n std_err .config
##   <chr>       <chr>      <dbl> <int>   <dbl> <chr>
## 1 npv         binary     0.846    10 0.00479 Preprocessor1_Model1
## 2 ppv         binary     0.801    10 0.00287 Preprocessor1_Model1
## 3 roc_auc     binary     0.903    10 0.00222 Preprocessor1_Model1
## 4 sensitivity binary     0.857    10 0.00434 Preprocessor1_Model1
## 5 specificity binary     0.787    10 0.00332 Preprocessor1_Model1
```



.metric	.estimate
accuracy	0.8476
kap	0.4253
sens	0.8605
spec	0.7384
ppv	0.9655
npv	0.3835
mcc	0.4572
j_index	0.5989
bal_accuracy	0.7994
detection_prevalence	0.7975
precision	0.9655
recall	0.8605
f_meas	0.9100

Links de tomek distancia HEOM

```
## # A tibble: 5 x 6
##   .metric     .estimator  mean     n std_err .config
##   <chr>       <chr>      <dbl> <int>   <dbl> <chr>
## 1 npv         binary     0.846    10 0.00479 Preprocessor1_Model1
## 2 ppv         binary     0.801    10 0.00287 Preprocessor1_Model1
## 3 roc_auc     binary     0.903    10 0.00222 Preprocessor1_Model1
## 4 sensitivity binary     0.857    10 0.00434 Preprocessor1_Model1
## 5 specificity binary     0.787    10 0.00332 Preprocessor1_Model1
```

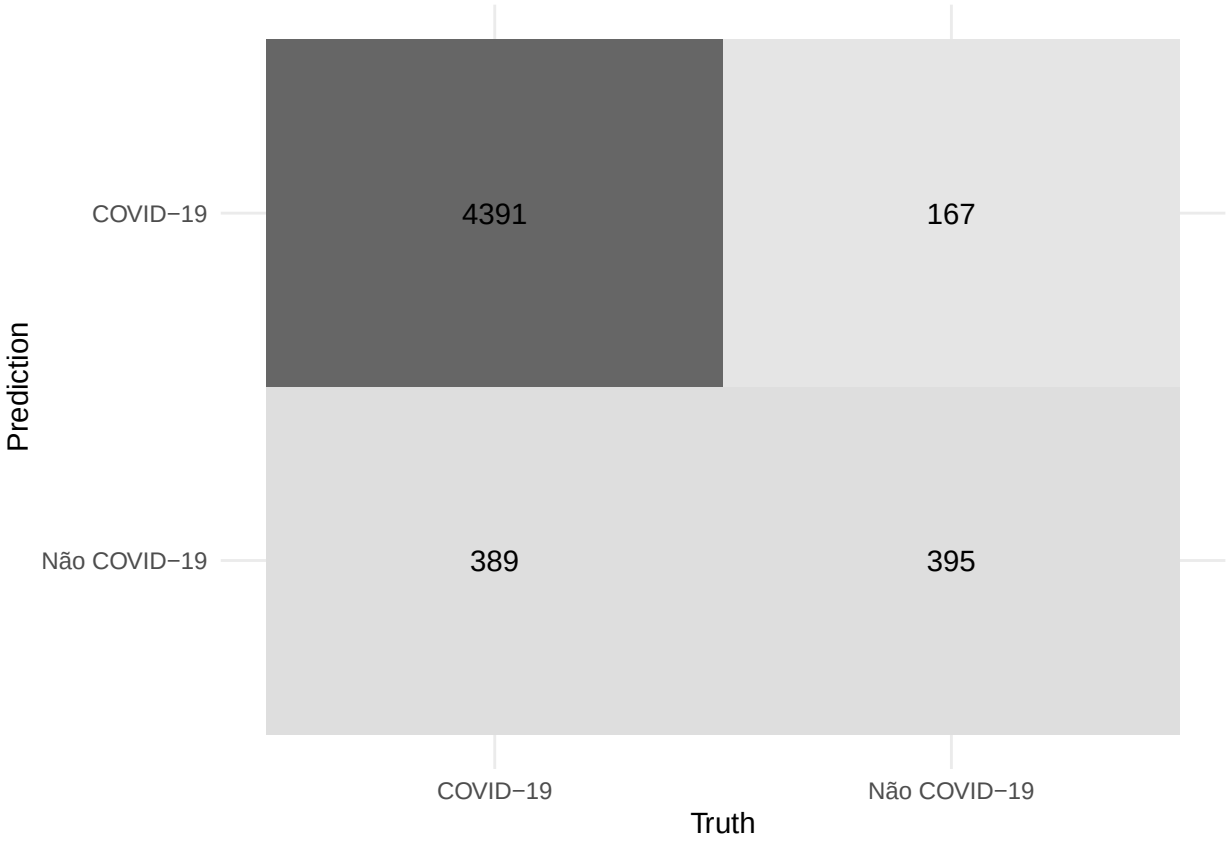


.metric	.estimate
accuracy	0.9446
kap	0.6483
sens	0.9908
spec	0.5516
ppv	0.9495
npv	0.8757
mcc	0.6690
j_index	0.5424
bal_accuracy	0.7712
detection_prevalence	0.9337
precision	0.9495
recall	0.9908

f_meas	0.9697

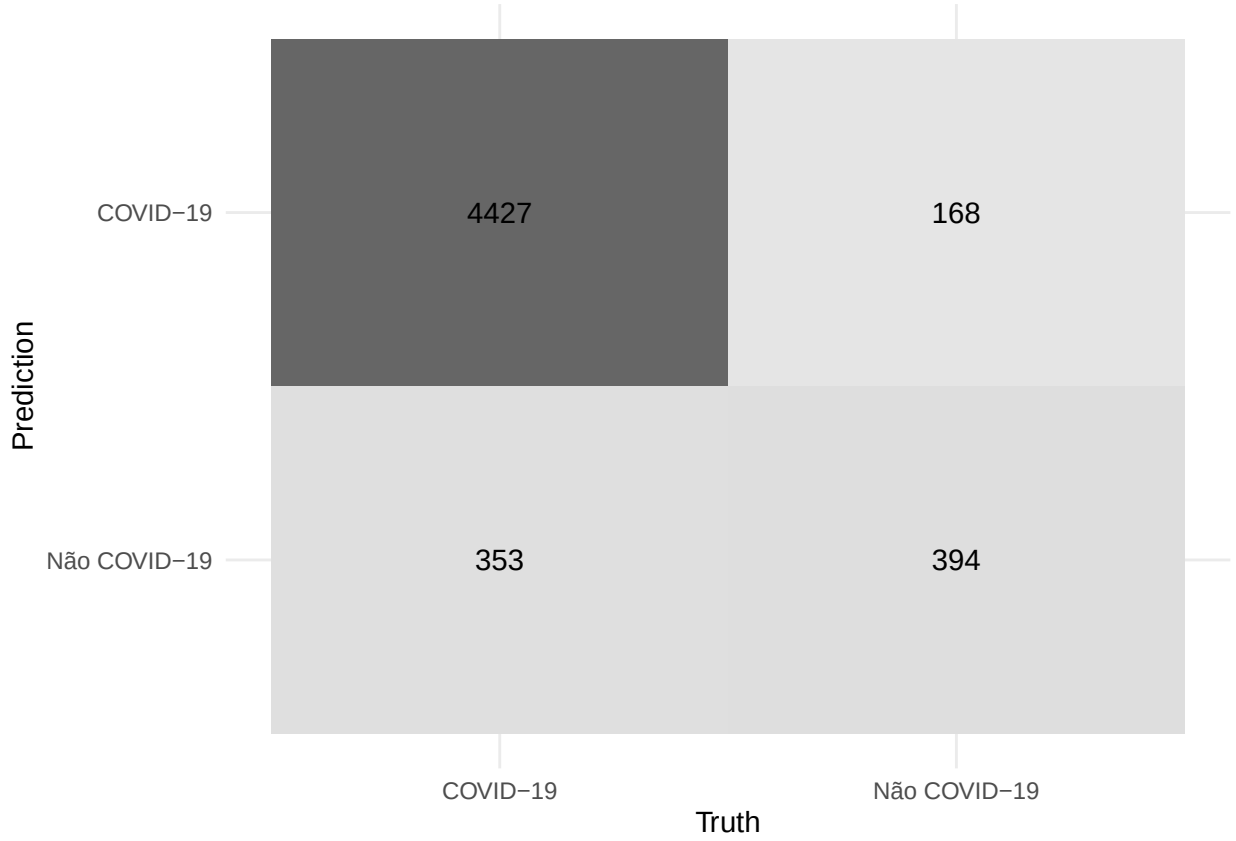
Reamostragem com dados numéricos e categóricos com aplicação do modelo XGBoost

Smote NC



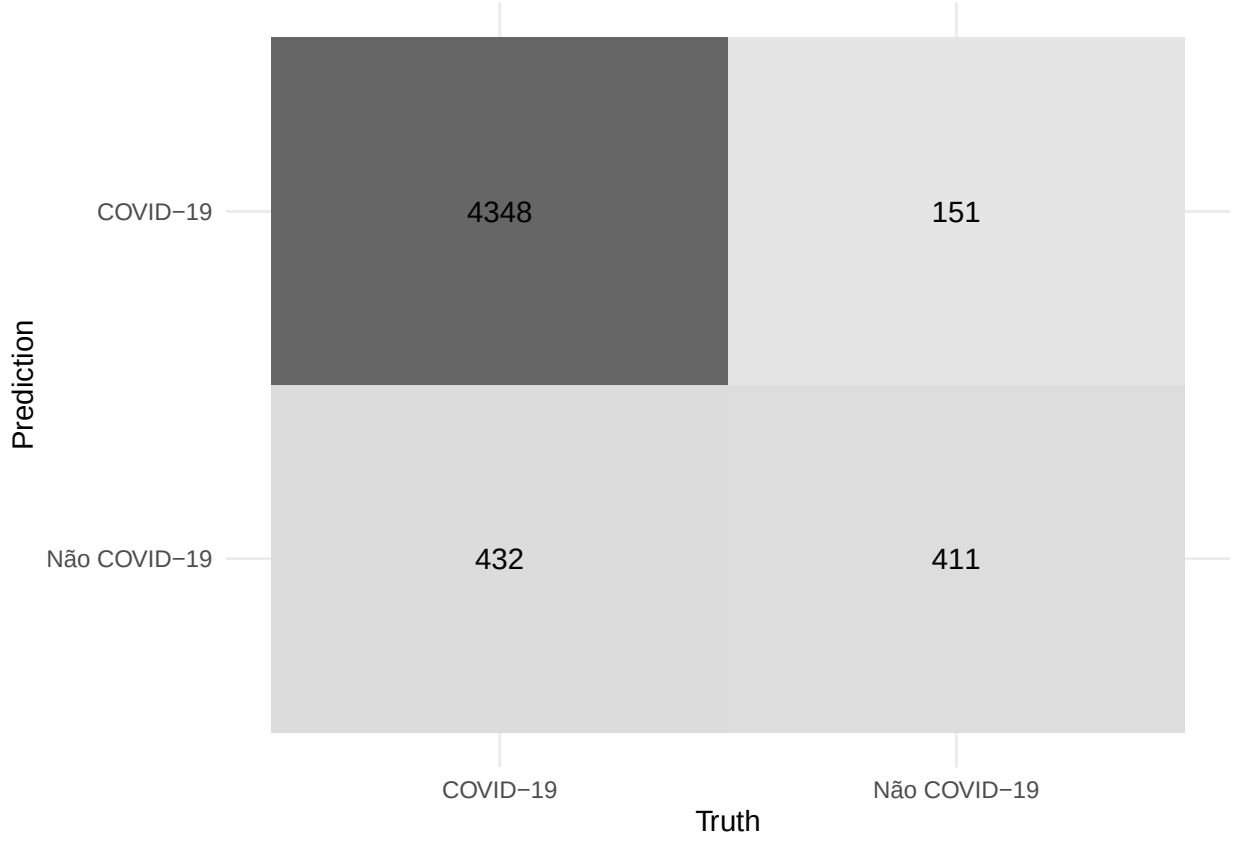
.metric	.estimate
accuracy	0.8959
kap	0.5292
sens	0.9186
spec	0.7028
ppv	0.9634
npv	0.5038
mcc	0.5388
j_index	0.6215
bal_accuracy	0.8107
detection_prevalence	0.8532
precision	0.9634
recall	0.9186
f_meas	0.9405

Adasyn distancia HEOM



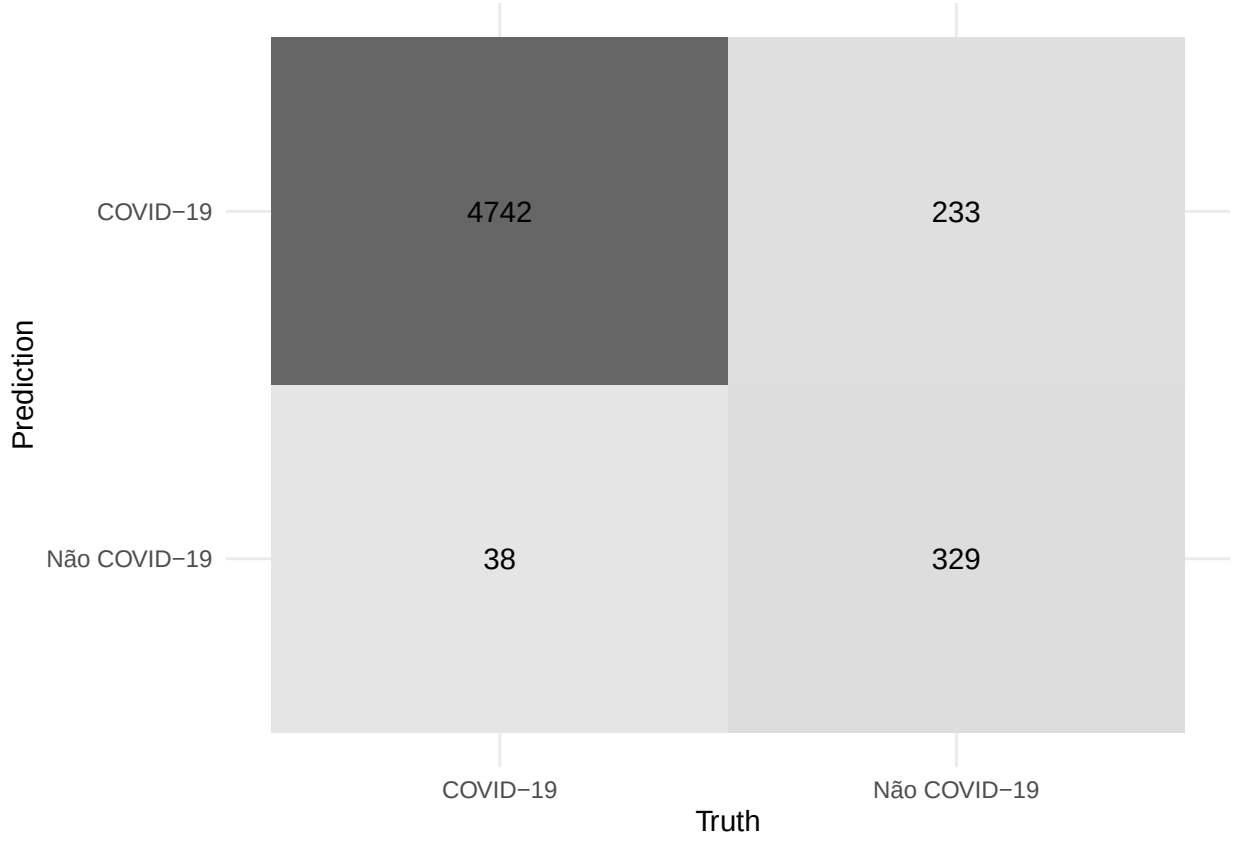
.metric	.estimate
accuracy	0.9025
kap	0.5477
sens	0.9262
spec	0.7011
ppv	0.9634
npv	0.5274
mcc	0.5549
j_index	0.6272
bal_accuracy	0.8136
detection_prevalence	0.8602
precision	0.9634
recall	0.9262
f_meas	0.9444

Smote distancia HEOM



.metric	.estimate
accuracy	0.8909
kap	0.5251
sens	0.9096
spec	0.7313
ppv	0.9664
npv	0.4875
mcc	0.5394
j_index	0.6409
bal_accuracy	0.8205
detection_prevalence	0.8422
precision	0.9664
recall	0.9096
f_meas	0.9372

Links de tomek distancia HEOM



.metric	.estimate
accuracy	0.9493
kap	0.6818
sens	0.9921
spec	0.5854
ppv	0.9532
npv	0.8965
mcc	0.7004
j_index	0.5775
bal_accuracy	0.7887
detection_prevalence	0.9313
precision	0.9532
recall	0.9921
f_meas	0.9722