Limpeza - Coronavírus RS

Eduardo

2023-03-19

Sumário

Introdução **2**

Seleção de atributos **2**

Transformação dos dados **2**

Sexo . . . 2
Faixa Etária . . . 3
Idade . . . 3
Evolução . . . 3
Hospitalizado . . . 4
UTI . . . 4
Febre . . . 4
Tosse . . . 4
Garganta . . . 5
Dispneia . . . 5
Raça/cor . . . 5
Etnia indígena . . . 6
Profissional da saúde . . . 7
SRAG . . . 7
Pessoa Privada de Liberdade . . . 7

Condições **8**

União dos *datasets* **8**

Filtro por `IDADE` e `EVOLUCAO` . . . 9

Introdução

A Secretaria da Saúde do Rio Grande do Sul, no Painel Coronavírus RS, disponibiliza alguns conjuntos de dados. Desses, escolhemos o conjunto *Casos Confirmados - microdados* e, a partir dele, criamos um *dataset* específico para tarefas de classificação e associação de regras.

Seleção de atributos

Os atributos do conjunto *Casos Confirmados - microdados* são os seguintes.

```
##  [1] "COD_IBGE"              "MUNICIPIO"             "COD_REGIAO_COVID"
##  [4] "REGIAO_COVID"          "SEXO"                  "FAIXAETARIA"
##  [7] "IDADE"                 "CRITERIO"              "DATA_CONFIRMACAO"
## [10] "DATA_SINTOMAS"         "DATA_INCLUSAO"         "DATA_EVOLUCAO"
## [13] "EVOLUCAO"              "HOSPITALIZADO"         "UTI"
## [16] "FEBRE"                 "TOSSE"                 "GARGANTA"
## [19] "DISPNEIA"              "OUTROS"                "CONDICOES"
## [22] "GESTANTE"              "DATA_INCLUSAO_OBITO"   "DATA_EVOLUCAO_ESTIMADA"
## [25] "RACA_COR"              "ETNIA_INDIGENA"        "PROFISSIONAL_SAUDE"
## [28] "BAIRRO"                "SRAG"                  "FONTE_INFORMACAO"
## [31] "PAIS_NASCIMENTO"       "PES_PRIV_LIBERDADE"
```

Queremos criar um *dataset* para a análise da ação do coronavírus nos indivíduos, por isso removemos atributos geográficos e temporais. O atributo `OUTROS` foi ignorado, pois não sabemos que outros sintomas são esses. O atributo `CONDICOES` será utilizado no futuro. Os atributos restantes são os seguintes, e o nome desse subconjunto é `carac`.

```
##  [1] "SEXO"               "FAIXAETARIA"        "IDADE"
##  [4] "EVOLUCAO"           "HOSPITALIZADO"      "UTI"
##  [7] "FEBRE"              "TOSSE"              "GARGANTA"
## [10] "DISPNEIA"           "RACA_COR"           "ETNIA_INDIGENA"
## [13] "PROFISSIONAL_SAUDE" "SRAG"               "PES_PRIV_LIBERDADE"
```

Transformação dos dados

Vamos analisar todas as colunas e transformá-las para diminuir o tamanho do *dataset*, que é muito grande.

Sexo

Os valores que `SEXO` pode assumir são os seguintes.

```
## [1] "Masculino" "Feminino"
```

Aplicamos as seguintes transformações a essa coluna.

```
carac$SEXO[carac$SEXO == 'Masculino'] <- 'M'
carac$SEXO[carac$SEXO == 'Feminino'] <- 'F'
```

Faixa Etária

Os valores que `FAIXAETARIA` pode assumir são os seguintes.

```
##  [1] "<1"        "01 a 04"   "05 a 09"   "10 a 14"   "15 a 19"   "20 a 29"
##  [7] "30 a 39"   "40 a 49"   "50 a 59"   "60 a 69"   "70 a 79"   "80 e mais"
```

Nesse caso, vamos atribuir a cada faixa etária um código numérico.

```
## <1 : 1
## 01 a 04 : 2
## 05 a 09 : 3
## 10 a 14 : 4
## 15 a 19 : 5
## 20 a 29 : 6
## 30 a 39 : 7
## 40 a 49 : 8
## 50 a 59 : 9
## 60 a 69 : 10
## 70 a 79 : 11
## 80 e mais : 12
```

Idade

Não há muito o que fazer com o atributo `IDADE`, contudo percebemos que há um problema.

```
##   [1]   0   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17
##  [19]  18  19  20  21  22  23  24  25  26  27  28  29  30  31  32  33  34  35
##  [37]  36  37  38  39  40  41  42  43  44  45  46  47  48  49  50  51  52  53
##  [55]  54  55  56  57  58  59  60  61  62  63  64  65  66  67  68  69  70  71
##  [73]  72  73  74  75  76  77  78  79  80  81  82  83  84  85  86  87  88  89
##  [91]  90  91  92  93  94  95  96  97  98  99 100 101 102 103 104 105 106 107
## [109] 108 109 110 111 112 113 115 116 118 119 120 121 122 123 124 125 127 128
## [127] 130 131 132 133 134 135 136 137 138 139 140 141 142 143 145 147 150 151
## [145] 153 157 158 163 164 165 166 168 170 171 172 173 174 175 179 180 184 223
## [163] 224 225 226 227 228 229 230 231 243 244 245 250 252 264 281 301 322 323
## [181] 324 325 326 327 329 330 331 332 335 336 337 338 341 342 347 349 354 355
## [199] 356 357 358 360 364 365 366 367 368 369 372 373 375 377 379 390
```

Há exatamente 215 observações com idade maior do que 120 anos. Como esse número não é expressivo, esses casos podem ser removidos do conjunto de dados, mas faremos isso em uma etapa posterior.

Evolução

Esse é um dos atributos mais importantes, pois diz se o paciente veio a óbito ou se recuperou. Os valores que `EVOLUCAO` pode assumir são os seguintes.

```
## [1] "RECUPERADO"          "EM ACOMPANHAMENTO"   "OBITO"
## [4] "OBITO OUTRAS CAUSAS"
```

Por enquanto, deixaremos essa coluna da forma que está.

Hospitalizado

Os valores que `HOSPITALIZADO` pode assumir são os seguintes.

```
unique(carac$HOSPITALIZADO)
```

```
## [1] "NAO" "SIM"
```

Vamos aplicar a seguinte transformação.

```
carac$HOSPITALIZADO[carac$HOSPITALIZADO == 'NAO'] <- 'N'
carac$HOSPITALIZADO[carac$HOSPITALIZADO == 'SIM'] <- 'S'
```

UTI

Os valores que `UTI` pode assumir são os seguintes.

```
unique(carac$UTI)
```

```
## [1] "NAO" "SIM"
```

Vamos aplicar a seguinte transformação.

```
carac$UTI[carac$UTI == 'NAO'] <- 'N'
carac$UTI[carac$UTI == 'SIM'] <- 'S'
```

Febre

Os valores que `FEBRE` pode assumir são os seguintes.

```
unique(carac$FEBRE)
```

```
## [1] "NAO" "SIM" ""
```

Além de abreviar os valores NAO e SIM, vamos assumir que o valor é NAO quando não for informado.

```
carac$FEBRE[carac$FEBRE == ''] <- 'N'
carac$FEBRE[carac$FEBRE == 'NAO'] <- 'N'
carac$FEBRE[carac$FEBRE == 'SIM'] <- 'S'
```

Tosse

Os valores que `TOSSE` pode assumir são os seguintes.

```
unique(carac$TOSSE)
```

```
## [1] "NAO" "SIM" ""
```

Vamos aplicar as mesmas transformações que aplicamos à `FEBRE`.

```
carac$TOSSE[carac$TOSSE == ''] <- 'N'
carac$TOSSE[carac$TOSSE == 'NAO'] <- 'N'
carac$TOSSE[carac$TOSSE == 'SIM'] <- 'S'
```

Garganta

Os valores que `GARGANTA` pode assumir são os seguintes.

```
unique(carac$GARGANTA)
```

```
## [1] "SIM" "NAO" ""
```

Vamos aplicar as mesmas transformações anteriores.

```
carac$GARGANTA[carac$GARGANTA == ''] <- 'N'
carac$GARGANTA[carac$GARGANTA == 'NAO'] <- 'N'
carac$GARGANTA[carac$GARGANTA == 'SIM'] <- 'S'
```

Dispneia

Os valores que `DISPNEIA` pode assumir são os seguintes.

```
unique(carac$DISPNEIA)
```

```
## [1] "NAO" "SIM" ""
```

Vamos aplicar as mesmas transformações anteriores.

```
carac$DISPNEIA[carac$DISPNEIA == ''] <- 'N'
carac$DISPNEIA[carac$DISPNEIA == 'NAO'] <- 'N'
carac$DISPNEIA[carac$DISPNEIA == 'SIM'] <- 'S'
```

Raça/cor

OS valores que `RACA_COR` pode assumir são os seguintes.

```
unique(carac$RACA_COR)
```

```
## [1] "BRANCA"        "PARDA"         "PRETA"         "NAO INFORMADO"
## [5] "AMARELA"       "INDIGENA"
```

Vamos substituir o valor NAO INFORMADO por NA.

```
carac$RACA_COR[carac$RACA_COR == 'NAO INFORMADO'] <- NA
```

Vamos codificar as cores com valores numéricos.

```
## AMARELA : 1
## BRANCA : 2
## INDIGENA : 3
## PARDA : 4
## PRETA : 5
```

Etnia indígena

Os valores que ETNIA_INDIGENA pode assumir são os seguintes.

```
unique(carac$ETNIA_INDIGENA)
```

```
##  [1] "NAO ENCONTRADO"
##  [2] ""
##  [3] "KAINGANG (CAINGANGUE)"
##  [4] "AMANAYE"
##  [5] "AJURU"
##  [6] "GUARANI M BYA"
##  [7] "PANKARU (PANCARU)"
##  [8] "ALANTESU"
##  [9] "GUARANI"
## [10] "GUARANI KAIOWA (PAI TAVYTERA)"
## [11] "TUPI"
## [12] "AIKANA (AIKANA, MAS SAKA,TUBARAO)"
## [13] "NÃO INFORMADO"
## [14] "ACONA (WAKONAS, NACONAS, JAKONA, ACORANES)"
## [15] "KAMBIWA (CAMBIUA)"
## [16] "GUARANI NANDEVA (AVAKATUETE, CHIRIPA, NHANDEWA, AVA GUARANI)"
## [17] "GUAJAJARA (TENETEHARA)"
## [18] "ANAMBE"
## [19] "MURA"
## [20] "ACONA (WAKONAS, NACONAS, JAKONA, ACORAÃ-NES)"
## [21] "XOKLENG (SHOKLENG, XOCLENG)"
## [22] "POTIGUARA"
## [23] "JURUNA"
## [24] "AKUNSU (AKUNT SU)"
## [25] "PANKARARU (PANCARARU)"
## [26] "BANIWA (BANIUA, BANIVA, WALIMANAI, WAKUENAI)"
## [27] "JIRIPANCO (JERIPANCO, GERIPANCO)"
## [28] "TABAJARA"
## [29] "MARUBO"
## [30] "PATAXO"
## [31] "KADIWEU (CADUVEO, CADIUEU)"
## [32] "TERENA"
## [33] "KAIXANA (CAIXANA)"
## [34] "TICUNA (TIKUNA, TUKUNA, MAGUTA)"
## [35] "TRUMAI"
## [36] "AICABA"
## [37] "GUATO"
## [38] "BARE"
## [39] "YANOMAMI YANOMAM (IANOMAMI, IANOAMA, XIRIANA)"
## [40] "CHARRUA"
## [41] "KAYAPO KARARAO (KARARAO)"
## [42] "KAYAPO XICRIM (XIKRIN)"
## [43] "KAIABI (CAIABI, KAYABI)"
## [44] "TARIANO (TARIANA, TALIASERI)"
## [45] "CHIQUITANO (XIQUITANO)"
## [46] "KARIPUNA (CARIPUNA)"
## [47] "CHAMACOCO"
## [48] "KANINDE"
```

```
## [49] "ATIKUM (ATICUM)"
## [50] "ARARA DO ACRE (SHAWANAUA, AMAWAKA)"
## [51] "TAPEBA"
## [52] "KAXIXO"
## [53] "GUARANIÃ  NANDEVAÃ  (AVAKATUETE, CHIRIPA,NHANDEWA, AVA GUARANI)"
## [54] "AHANENAWA"
## [55] "NAMBIKWARA DO CAMPO (HALOTESU, KITHAULU, WAKALITESU, SAWENTES, MANDUKA)"
## [56] "ARARA DO ARIPUANA (ARARA DO BEIRADAO/ARIPUANA)"
## [57] "KAIOWA"
```

Há exatamente 2906849 observações com o valor de NAO ENCONTRADO. Como esse valor é praticamente o número de observações, vamos assumir que ele corresponde a um paciente que não é indígena. Por isso, vamos substuí-lo por NA. Além disso, vamos fazer a mesma coisa para os valores e NÃO INFORMADO, que contam 61752 e 55, respectivamente.

```
carac$ETNIA_INDIGENA[carac$ETNIA_INDIGENA == 'NAO ENCONTRADO'] <- NA
carac$ETNIA_INDIGENA[carac$ETNIA_INDIGENA == ''] <- NA
carac$ETNIA_INDIGENA[carac$ETNIA_INDIGENA == 'NÃO INFORMADO'] <- NA
```

Profissional da saúde

Os valores que `PROFISSIONAL_SAUDE` pode assumir são os seguintes.

```
unique(carac$PROFISSIONAL_SAUDE)
```

```
## [1] "NAO"           "SIM"           "NAO INFORMADO"
```

Há 168528 observações com o valor NAO INFORMADO. Vamos substituí-lo por NA.

```
carac$PROFISSIONAL_SAUDE[carac$PROFISSIONAL_SAUDE == 'NAO INFORMADO'] <- NA
carac$PROFISSIONAL_SAUDE[carac$PROFISSIONAL_SAUDE == 'NAO'] <- 'N'
carac$PROFISSIONAL_SAUDE[carac$PROFISSIONAL_SAUDE == 'SIM'] <- 'S'
```

SRAG

Os valores que `SRAG` pode assumir são os seguintes.

```
unique(carac$SRAG)
```

```
## [1] "NAO" "SIM"
```

Vamos aplicar a transformação para abreviar os valores.

```
carac$SRAG[carac$SRAG == 'NAO'] <- 'N'
carac$SRAG[carac$SRAG == 'SIM'] <- 'S'
```

Pessoa Privada de Liberdade

Os valores que `PES_PRIV_LIBERDADE` podem assumir são os seguintes.

```
unique(carac$PES_PRIV_LIBERDADE)
```

```
## [1] "NAO" "SIM"
```

Vamos aplicar a abreviação.

```
carac$PES_PRIV_LIBERDADE[carac$PES_PRIV_LIBERDADE == 'NAO'] <- 'N'
carac$PES_PRIV_LIBERDADE[carac$PES_PRIV_LIBERDADE == 'SIM'] <- 'S'
```

Condições

Aplicamos um *script* à coluna `CONDICOES` para expandi-la.

```
##  [1] "CARDI"                "IMUNOSSUPRESSAO"      "RESP_DESCOMP"
##  [4] "OBESIDADE"            "DIABETES"             "CARDIOVASC"
##  [7] "MELLITUS"             "GESTANTE"             "CROMOSSOM_FRAG_IMUNO"
## [10] "OUTROS"               "PNEUMATOPATIA"        "RENAL_AVANC"
## [13] "PUERPERA"             "DOWN"                 "HEPATICA"
## [16] "ASMA"                 "HEMATOLOGICA"         "IMUNODEFICIENCIA"
## [19] "RENAL"                "NEUROLOGICA"          "GESTANTE_ALTO_RISCO"
## [22] "CARDI_DESCOMP"
```

Transformamos os valores NA em N e TRUE em S.

```
condi[is.na(condi)] <- 'N'
condi[condi == TRUE] <- 'S'
```

União dos *datasets*

Vamos unir `carac` com `condi`.

```
carac_condi <- cbind(carac, condi)
colnames(carac_condi)
```

```
##  [1] "SEXO"                 "FAIXAETARIA"          "IDADE"
##  [4] "EVOLUCAO"             "HOSPITALIZADO"        "UTI"
##  [7] "FEBRE"                "TOSSE"                "GARGANTA"
## [10] "DISPNEIA"             "RACA_COR"             "ETNIA_INDIGENA"
## [13] "PROFISSIONAL_SAUDE"   "SRAG"                 "PES_PRIV_LIBERDADE"
## [16] "CARDI"                "IMUNOSSUPRESSAO"      "RESP_DESCOMP"
## [19] "OBESIDADE"            "DIABETES"             "CARDIOVASC"
## [22] "MELLITUS"             "GESTANTE"             "CROMOSSOM_FRAG_IMUNO"
## [25] "OUTROS"               "PNEUMATOPATIA"        "RENAL_AVANC"
## [28] "PUERPERA"             "DOWN"                 "HEPATICA"
## [31] "ASMA"                 "HEMATOLOGICA"         "IMUNODEFICIENCIA"
## [34] "RENAL"                "NEUROLOGICA"          "GESTANTE_ALTO_RISCO"
## [37] "CARDI_DESCOMP"
```

Filtro por `IDADE` e `EVOLUCAO`

Por fim, filtramos as observações pela idade e evolução.

```
carac_condi <- filter(carac_condi, IDADE < 120)
carac_condi <- filter(carac_condi, EVOLUCAO == 'OBITO' | EVOLUCAO == 'RECUPERADO')
```