Mortalidade no Brasil no ano de 2020

**Instituto Federal de Educação,
Ciência e Tecnologia de São Paulo**

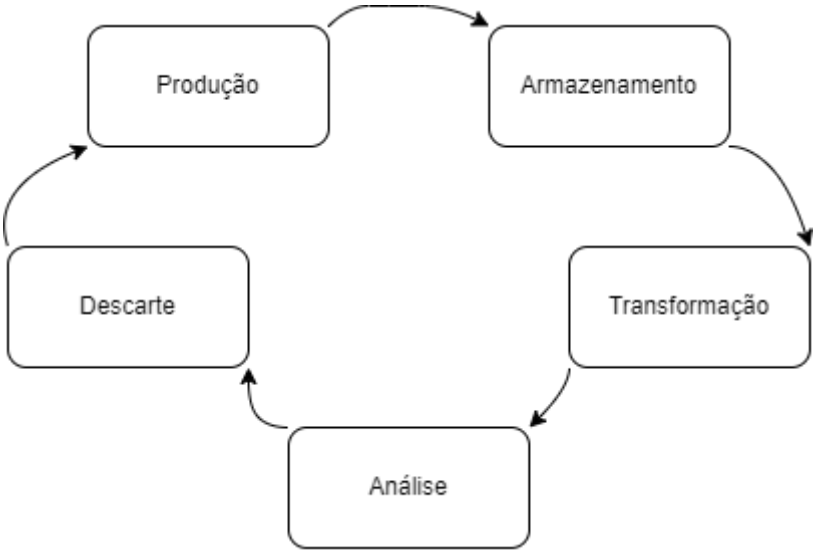
Rian Santos Macedo

Marcos Querino dos Santos e Santos Junior

Intrudução

A proposta deste projeto é aplicar o ciclo de vida do dado em uma base de dados do Ministério da Saúde do Brasil. A base escolhida para análise contém os registros de mortalidade no Brasil no ano de 2020.

O ciclo de vida do dado, citado anteriormente, diz respeito às fases que um conjunto de dados percorre dentro da ciência de dados. Cada fase deste ciclo é abordada em um capítulo deste relatório.



Produção

A base de mortalidade geral foi coletada do sistema OpenDataSus, por meio do pacote microdatasus, feito para liguagem R. A outra base, com os municípios brasileiros, foi baixada manualmente do Moodle Câmpus, ambiente virtual de apoio ao ensino presencial e a distância do IFSP. Os arquvios vieram em formato CSV.

Armazenamento

Ambas as bases foram armazenadas em datasets da Posit Cloud.

Transformação

A transformação necessária na base do Ministério da Saúde, como a codificação utilizada, ficou por parte do pacote microdatasus.

Análise de dados

A fase de análise foi baseada no modelo CRISP-DM, padrão internacional de mineração de

dados. CRISP-DM é o acrônimo para CRoss Industry Standard Process for Data Mining, que em tradução direta pode ser entendido como um padrão de processos de mineração de dados entre indústrias.

Ele especifica os passos necessários para o aproveito de dados a fim de se obter informações e conhecimento sobre eles.

Este padrão se constitui em seis fases:

- Entendimento do negócio
- Entendimento dos dados
- Preparação de dados
- Modelagem - Análise Exploratória
- Modelagem - Analise Implicita
- Avaliação

Entendimento do negócio

O Ministério da Saúde brasileiro desenvolveu o SIM, Sistema de Informação sobre Mortalidade, que unifica declarações de óbito emitidas no país desde 1979. Seu conjunto de informações serve de apoio para o desenvolvimento de políticas públicas com respeito a saúde da população.

Entendimento dos dados

Extraímos os dados de mortalidade geral em 2020, que conta com informações sobre a causa do óbito, local de ocorrência, e características fisícas e socioeconômicas dos indivíduos de forma anonimizada.

Todos os óbitos registrados na base são do estado de São Paulo, não fetais.

Preparação de dados

Identificando os municípios

Durante a preparação dos dados, Foi feita junção da base de mortalidade com uma base de dados sobre os municípios brasileiros. A ação foi necessária pois na base de mortalidade os municípios são referenciados pelo seu código, e a partir da junção foi possível identificá-los.

Gerenciamento de dados ausentes

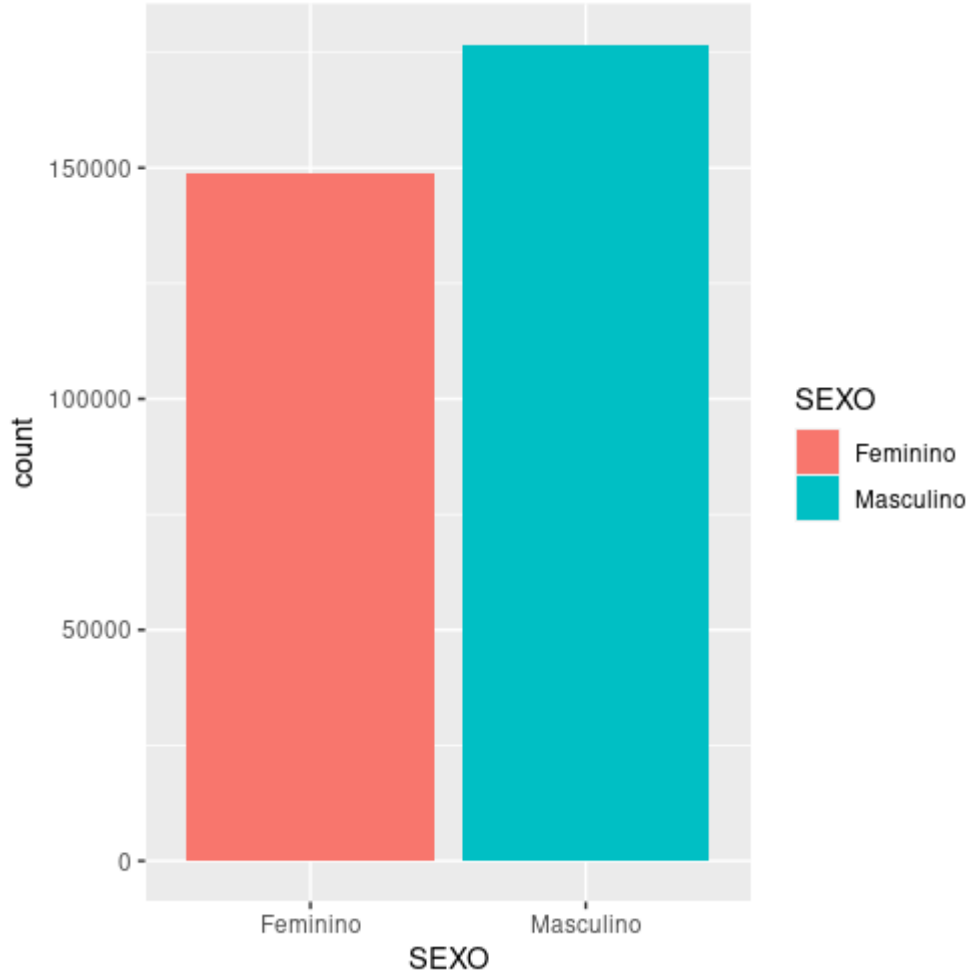
- apagar a linha?
- ou substituir valor?

Padronização dos dados

- IDADE: Todas as idades em dias, meses e de até um ano foram agrupadas na catgoria "até um ano". Demais idades seguem em anos, até agrupar idades maiores ou iguais a 100 em "cem ou mais".

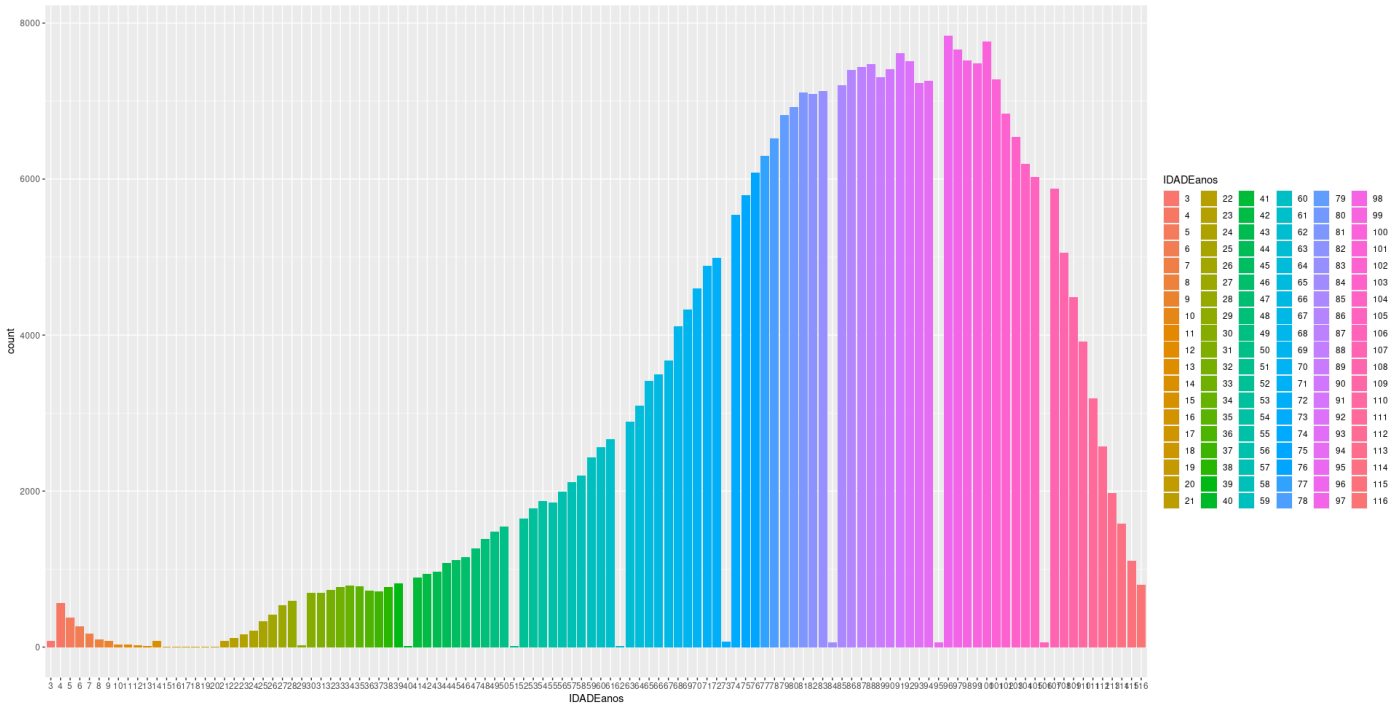
Modelagem - Análise Exploratória

A taxa de mortalidade masculina é maior em relação à feminina

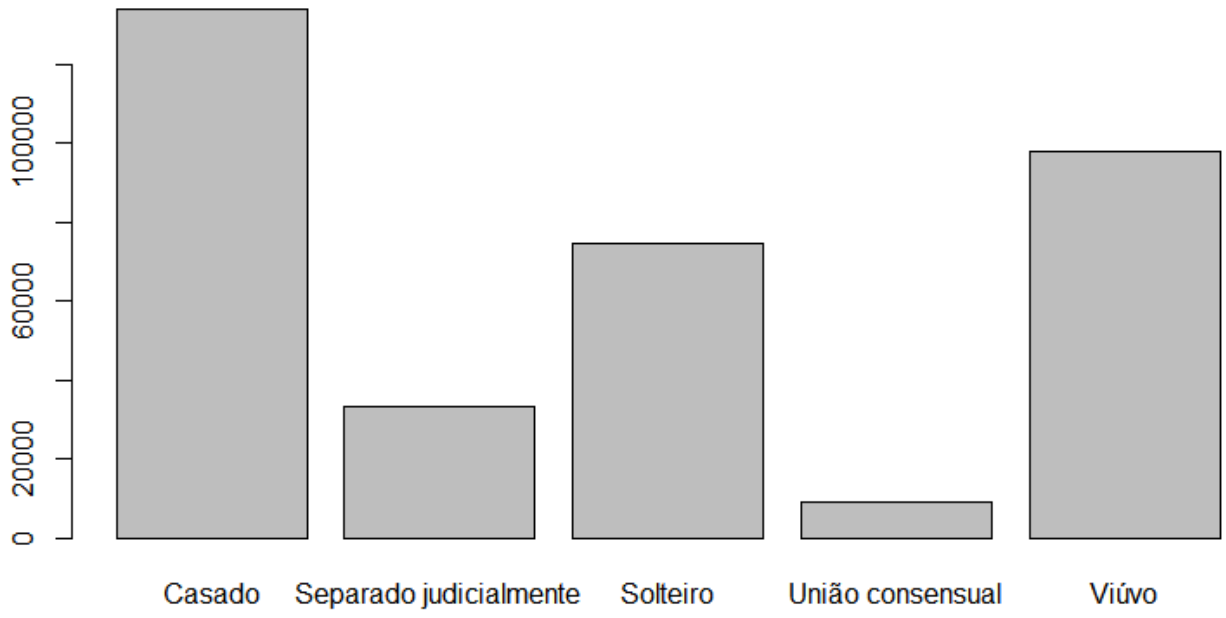


Se uma pessoa viveu até os 15 anos, provavelmente irá viver no mínimo até aos 20.

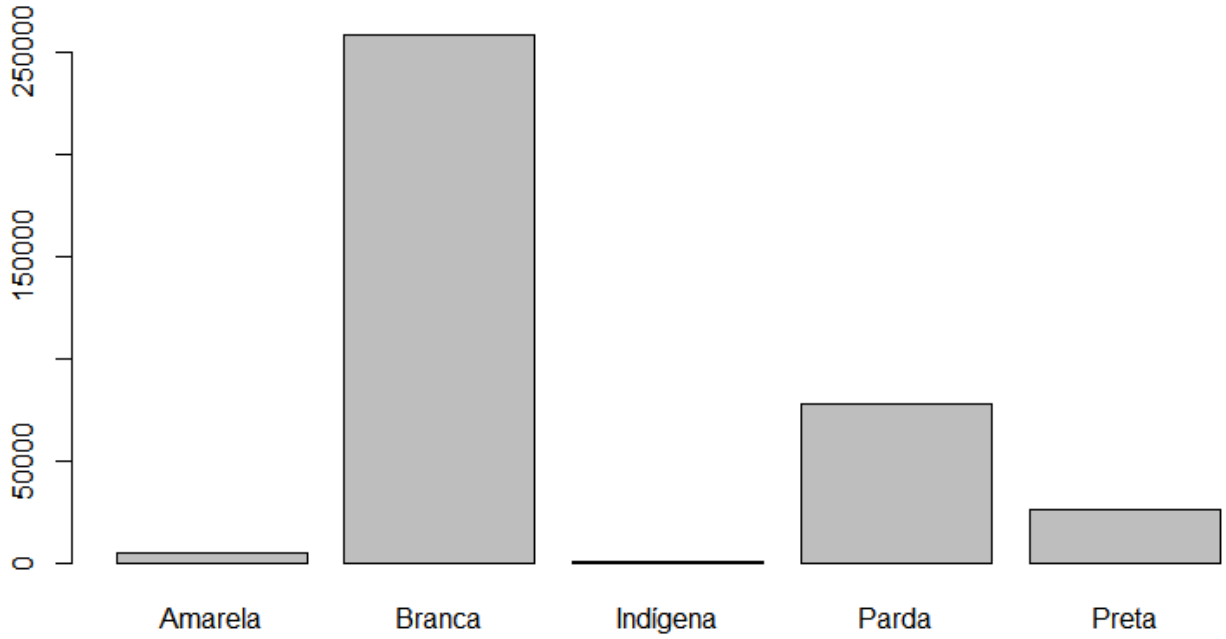
Essa afirmação se baseia no número baixo de óbitos registrados nesta faixa etária.



A Maioria dos óbitos são de pessoas casadas



Numeros de óbitos em relação a raça

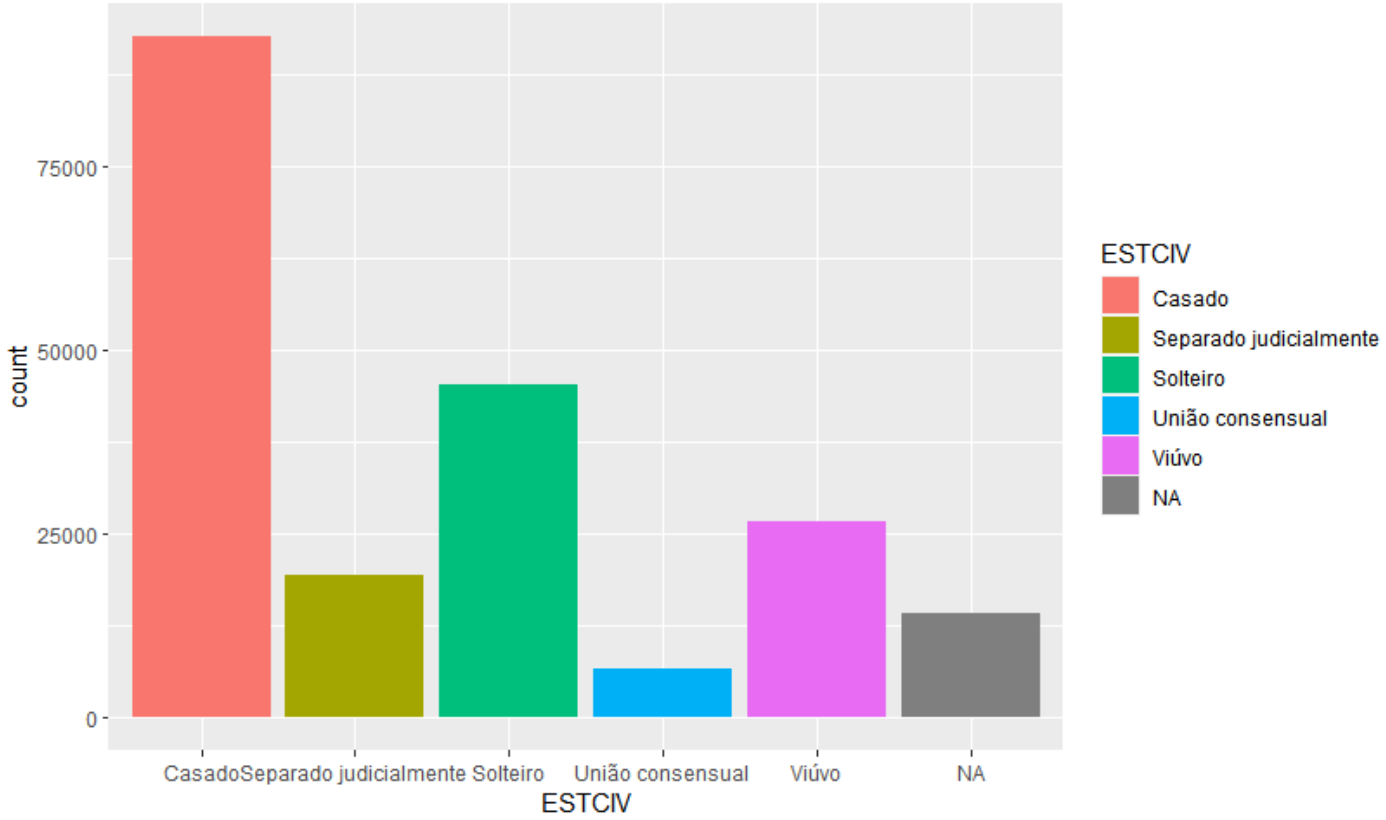


As principais causas de morte foram

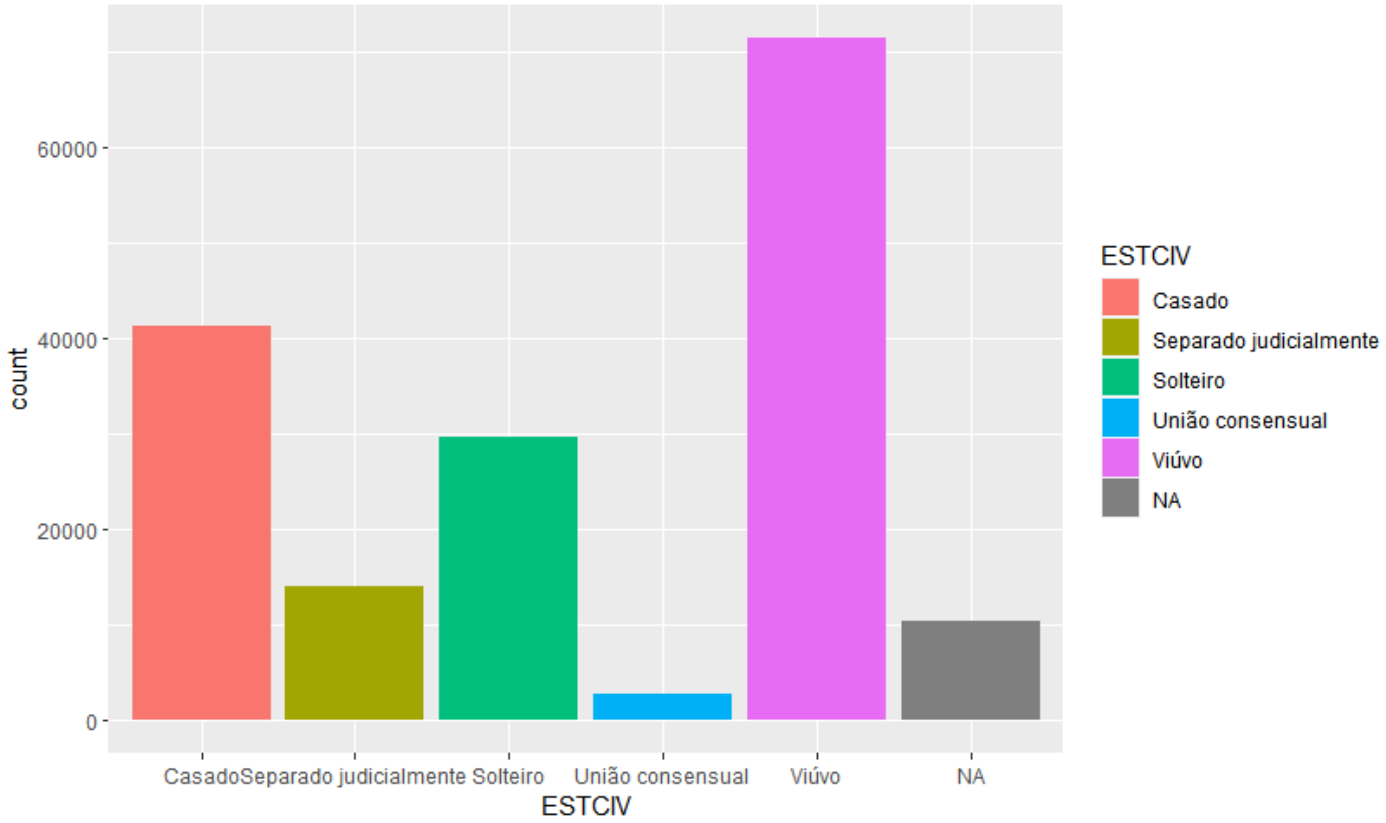
- Coronavírus
- Infarto
- Causas não especificadas,
- Demais transtornos respiratórios, diabetes,
- Neoplasia dos brônquios, infecção urinária
- AVC e alzhieimer

```
# Groups:   CAUSABAS [3,627]
   CAUSABAS     n
   <fct>    <int>
 1 B342     50722
 2 I219     23354
 3 R99      11305
 4 J988      9085
 5 J189      7920
 6 E149      7050
 7 C349      7005
 8 N390      6934
 9 I64       6767
10 G309      6574
```

A maioria dos homens morrem após o matrimônio, estando o cônjuje ainda vivo



Por outro lado, o sexo feminno tende a falecer após seu cônjuje



As ocupações com mais óbitos são: Aposentados, donas de casa e pedreiros.

```
# Groups:   OCUP [1,631]
   OCUP                                                  n
   <fct>                                             <int>
 1 Aposentado/Pensionista                           116126
 2 Dona de Casa                                      59391
 3 Nao informada                                     41393
 4 Ignorada                                          17652
 5 Pedreiro                                           8237
 6 Desempregado cronico ou cuja habitação habitual    7750
 7 Comerciante varejista                              5787
 8 Empregado doméstico nos serviços gerais            5118
 9 Representante comercial autônomo                   5084
10 Motorista de carro de passeio                      4318
```

Modelagem - Analise Implicita

Para a analise implicita foi elaborado um modelo de regressão logistica binomial utilizando como variavel dependente o estado civil(ESTCIV), dividindo em duas classes "viúvo e não viúvo", e utilizando as variaveis CAUSABAS, faixa_idade, SEXO, RACACOR, OCUP como variaveis independentes.

Foi ajustado a idade que era um valores muito extensos para uma variavel categorica de 4 niveis.

Foi utilizado 100000 dados extraidos da base do SIM e removidos os 'NA' desses dados.

Foi alterado a variavel ESTCIV para dicotomica.

Posteriormente foi feito um modelo.

Avaliação

Em relação ao modelo da analise explicita foi feito duas avaliações.

A primeira avaliação é baseada na matrix de confusão gerada pelos dados:

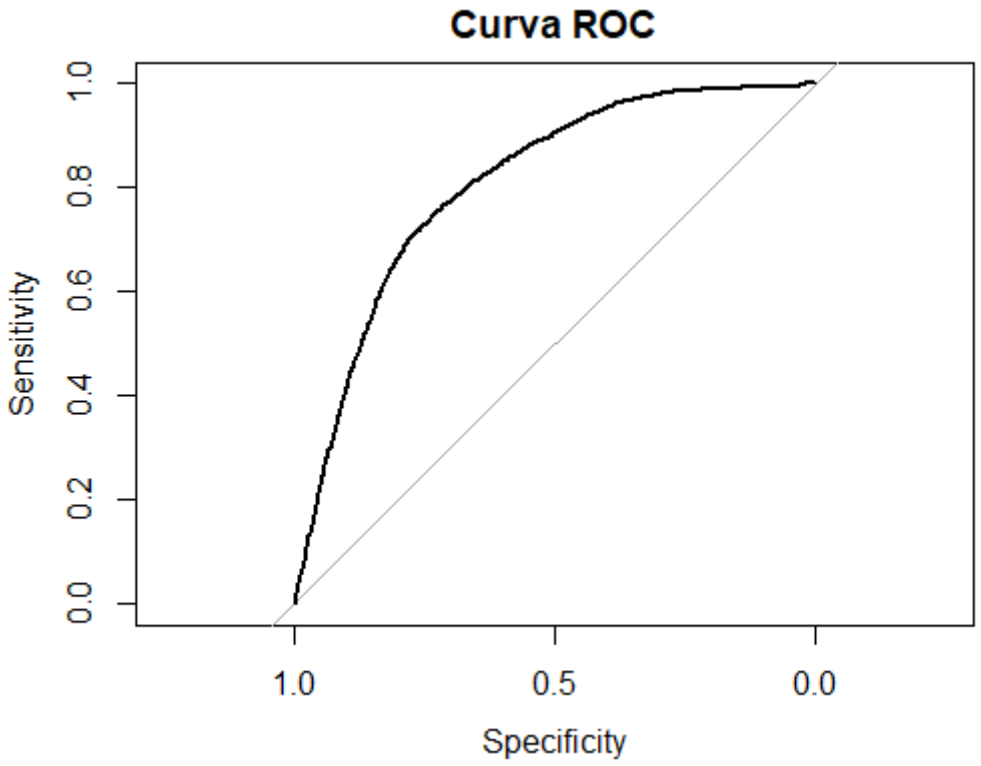
```
> matriz_confusao <- table(dados_teste$resultado, previsoes > 0.5)
> matriz_confusao

       FALSE  TRUE
  FALSE 17931  3402
  TRUE   3410  4789
```

Em seguida extraido informações de acuracia, sensibilidade e especificidade

```
> acuracia <- sum(diag(matriz_confusao)) / sum(matriz_confusao)
> print(acuracia)
[1] 0.769335
```

```
> sensibilidade <- matriz_confusao[2, 2] / sum(matriz_confusao[2, ])
> print(sensibilidade)
[1] 0.5840956
> especificidade <- matriz_confusao[1, 1] / sum(matriz_confusao[1, ])
> print(especificidade)
[1] 0.8405288
```

Por ultimo foi feito a curva ROC com o resultado com base nas predições.



Referências

Bases de dados utilizadas

- Mortalidade Geral 2020
- Municícios brasileiros
- Estrutura do SIM

Trabalhos semelhantes

- Capítulos da Classificação Estatística Internacional de Doenças e Problemas Relacionados à Saúde / CID-10

Ferramentas utilizadas

- Posit Cloud: Ambiente de desenvolvimento de modelos e armazenamento de datasets.

- SALDANHA, Raphael de Freitas; BASTOS, Ronaldo Rocha; BARCELLOS, Christovam. Microdatasus: pacote para download e pré-processamento de microdados do Departamento de Informática do SUS (DATASUS). Cad. Saúde Pública, Rio de Janeiro , v. 35, n. 9, e00032419, 2019 . microdatasus.
- read.dbc: Pacote necessário para usar a biblioteca microdatasus.
- md2pdf: Conversão do REDME do depositório em um arquivo PDF.