Mineração de Dados
Análises Síndrome Respiratória Aguda Grave com o uso de machine learning com R/tidymodels

Leonardo Pereira Borges

29 de outubro de 2022

?abstractname?

This is a LaTeX simple document skeleton. Use it as a base for your own documents.

Resumo

Este artigo foi realizado usando o conjunto de dados disponível pelo governo "Datasus", referente a Síndrome Respiratória Aguda Grave - incluindo dados da COVID-19. Os dados possui todas as informações de pacientes que realizou o teste para verificação da contração do vírus. "FALTA RESUMO RESULTADOS".

1 Introdução

A mineração de dados envolve o conhecimento de áreas como banco de dados, estatística, aprendizagem de máquina, computação natural, computação de alto desempenho, análise espacial de dados, inteligencia artificial, entre outros. Suas funcionalidades são usadas para especificar os tipo de informações a serem obtidas. Suas tarefas podem ser caracterizadas em duas categorias: as preditivas e a descritivas. Para este trabalho vamos usar as preditivas, que faz a inferência dos dados objetivando predições [1].

Para análises preditivas simples basta ter um grupo de dados onde os dados possam ser comparados. Para encontra um conjunto de dados atualmente temos os sites governamentais, onde disponibilizam vários tipos de dados com fácil acesso e qualquer pessoa pode acessar essas informações. Geralmente eles disponibilizam os dados em pdf e também em formato csv que permite o processamento dos dados por máquinas.

A disponibilização dos dados de forma irrestrita dos órgãos governamentais é muito importante para a transparência das informações e excelente para o programadores para ter vários estilos de dados para teste de aprendizagem de máquina.

Os dados utilizados neste artigo foi extraído do Ministério da Saúde. Dados referente a Banco de Dados de Síndrome Respiratória Aguda Grave - incluindo dados da COVID-19 no ano de 2021.

Vamos utilizar o software R para a criação da aprendizagem de máquina no conjunto de dados.

2 Conjunto de Dados

Será apresentado neste trabalho análises realizadas no conjunto de dados referente a "Síndrome Respiratória Aguda Grave - incluindo dados da COVID-19"do ano de 2021. Os dados podem ser acessados no site `https://opendatasus.saude.gov.br/dataset/srag-2021-e-2022`. Esta página tem como finalidade disponibilizar o legado dos bancos de dados (BD) epidemiológicos de SRAG, da rede de vigilância da Influenza e outros vírus respiratórios, desde o início da sua implantação (2009) até os dias atuais (2022), com a incorporação da vigilância da covid-19. Atualmente, o sistema oficial para o registro dos casos e óbitos por SRAG é o Sistema de Informação da Vigilância Epidemiológica da Gripe (SIVEP-Gripe) [2].

O Ministério da Saúde (MS), por meio da Secretaria de Vigilância em Saúde (SVS), desenvolve a vigilância da Síndrome Respiratória Aguda Grave (SRAG) no Brasil, desde a pandemia de Influenza A(H1N1)pdm09. A partir disso, a vigilância de SRAG foi implantada na rede de vigilância de Influenza e outros vírus respiratórios, que anteriormente atuava exclusivamente com a vigilância sentinela de Síndrome Gripal (SG).Em 2020, a vigilância da COVID-19, a infecção humana causada pelo novo Coronavírus, que vem causando uma pandemia, foi incorporada na rede de vigilância da Influenza e outros vírus respiratórios [2].

A base de dados completa possui 172 colunas, varias informações para vários tipos de trabalhos a serem aplicados. Para este artigo vamos trabalhar com apenas 18 colunas e iremos trabalhar somente

com os dados de Belo Horizonte reduzindo assim a necessidade de grande espaço de memória para processamento.

Os dados serão tratados no R, sendo necessário a instalação de alguns pacotes para a preparação dos dados. Vamos usar o pacote *tydiverse*, pacote principal de nossa analise. O tidyverse é uma coleção opinativa de pacotes R projetados para ciência de dados. Todos os pacotes compartilham uma filosofia de design subjacente, gramática e estruturas de dados.O 'tidyverse' é um conjunto de pacotes que funcionam em harmonia porque compartilham representações de dados comuns e design 'API'. Este pacote foi projetado para facilitar a instalação e o carregamento de vários pacotes 'tidyverse' em uma única etapa. [5]

O tidyverse é uma "coleção opinativa de pacotes R projetados para a ciência de dados", criada com o objetivo de tornar as tarefas de ciência de dados em R mais simples, mais legíveis e mais reproduzíveis. os pacotes são "opinativos"porque são projetados para tornar as tarefas que os autores dos pacotes consideram como boas práticas, fáceis, e fazer as tarefas que são consideradas difíceis.[3]

Para um maior conhecimento e familiarização com a base de dados, abaixo segue a tabela com estatísticas descritivas.

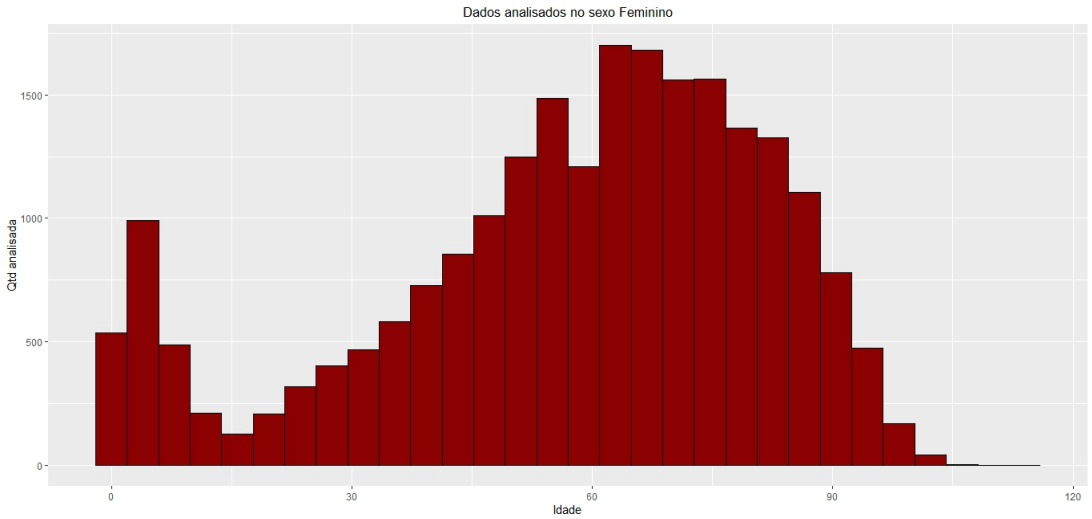
	Sexo	Média	Mediana	Mais novo	Mais velho	Total
1	Feminino	57,2	62	0	114	22.638
2	Masculino	53,3	56	0	106	24.645
3	Ignorado	26,8	5	0	76	11

Tabela 1: Tabela 1: Analise descritiva dos dados.

Abaixo o ggplot para entendermos os comportamentos dos dados. O gráfico representa a quantidade pela idade de cada sexo das pessoas que tiveram a análise coletada na cidade de Belo Horizonte no ano de 2021.

arquitetura

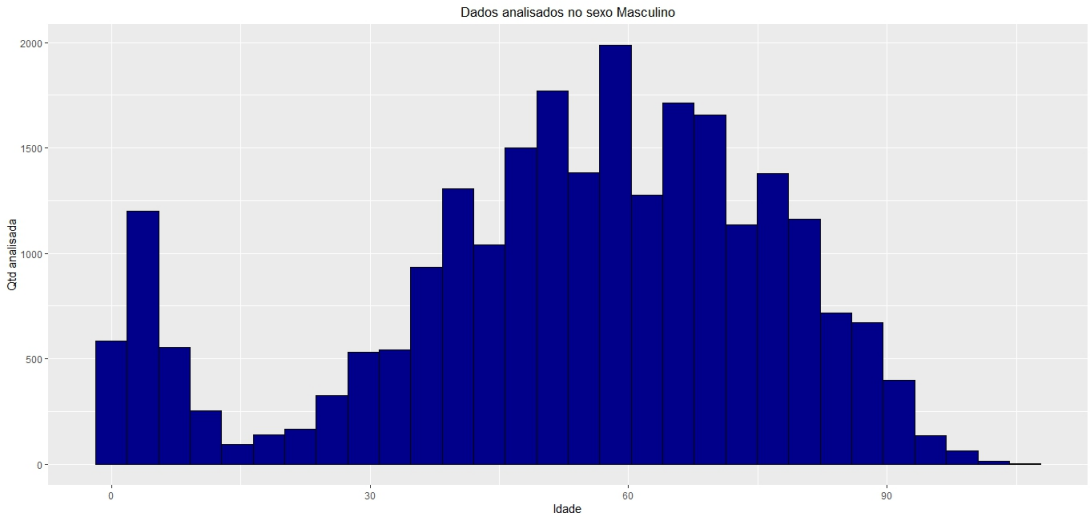
Figura 1: Quantidade de análises realizada no sexo feminino



ggplot da base dados do sexo feminino

arquitetura

Figura 2: Quantidade de análises realizada no sexo masculino



ggplot da base dados do sexo masculino

Observando os dois gráficos, temos que seguem uma distribuição normal com a maior incidencia de analises na faixa etária de 50 a 70 anos de idade, ou seja um grupo de idade independente do sexo, que representa a concentração de analises no ano de 2021.

3 Aprendizagem de Maquina

Em 1959, Arthur Samuel definiu o aprendizado de máquina como o "campo de estudo que dá aos computadores a habilidade de aprender sem serem explicitamente programados". Ou seja, é um método de análise de dados que automatiza a construção de modelos analíticos. É baseado na ideia de que sistemas podem aprender com dados, identificar padrões e tomar decisões com o mínimo de intervenção humana. A importância desse aprendizado se deve principalmente ao fato de que atualmente tem surgido cada vez mais a necessidade de manipulações de grandes volumes e variedades de dados disponíveis[4].

A aprendizagem de maquina é aplicado para prever certos modelos para o conjunto de dados. Possui dois tipos de aprendizado de maquina, o supervisionado e o não supervisionado. Neste conjunto de dados vamos utilizar o aprendizado supervisionado pois sabemos qual o modelo de resposta que queremos.

Com a utilização do software R na aprendizagem de máquina vamos utilizar o método de Tidyverse, um pacote disponível no software R, sendo necessário a sua instalação no inicio do processo de aprendizagem.

O processo de aprendizagem separa o conjunto de dados em dois subconjunto, um conjunto para realizar o treino do modelo e o outro subconjunto para realizar o teste de resultado criado na aprendizagem.

4 Resultado

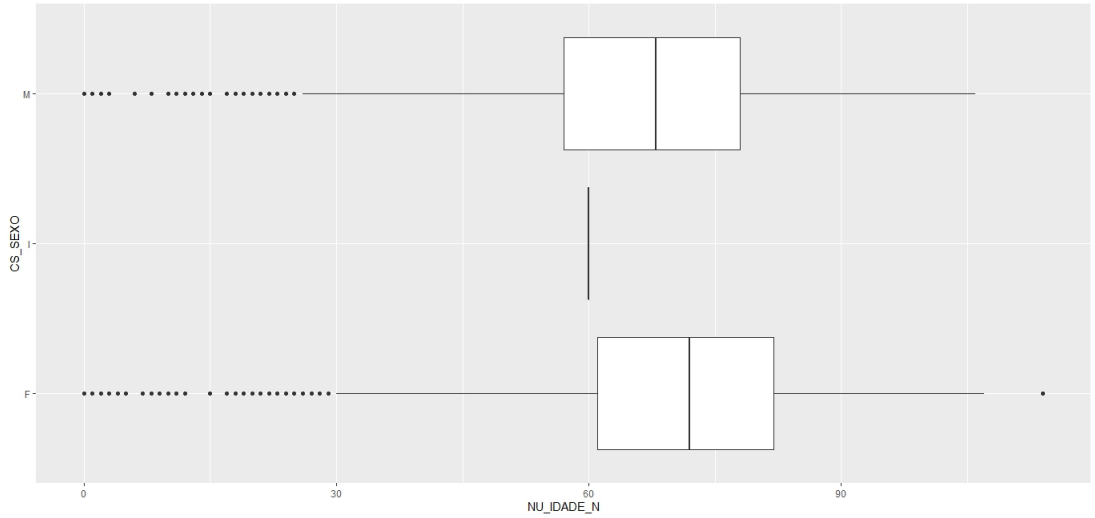
A análise de dados através da aprendizagem de maquina será aplicado ao data set das analises respiratórias agudas de 2021 na cidade de Belo Horizonte no ano de 2021, período em que houve maiores picos de casos de Covid-19 no pais.

Será realizado a classificação de óbitos correlacionado com algumas das características das pessoas que foi coletado para o banco de dados. Será comparado a relação de óbitos com a influencia de vacinados, idade e pessoas acima do peso.

Realizando a comparação entre obitos por sexo e idade não foi identificado uma forte relação, ou seja o sexo e idade não influenciou na quantidade de óbitos durante o período.

arquitetura

Figura 3: òbitos por sexo e idade



Boxplot óbitos comparado o sexo e idade

Referências

[1] L.N. de Castro e D.G. Ferrari. *Introdução a mineração de dados*. Saraiva Educação S.A., 2017. ISBN: 9788547200992. URL: `https://books.google.com.br/books?id=SSlrDwAAQBAJ`.

[2] Datasus. *SRAG 2021 e 2022 - Banco de Dados de Síndrome Respiratória Aguda Grave - incluindo dados da COVID-19*. urlhttps://opendatasus.saude.gov.br/dataset/srag-2021-e-2022. 2022.

[3] Hefin Rhys. *Machine Learning with R, the tidyverse, and mlr*. Simon e Schuster, 2020.

[4] Maqueise Pinheiro e Thaís Machado. *Aprendizado de Máquinas com R*. urlhttps://cienciadedadosuff.github.io/cursos/ 2020.

[5] Hadley Wickham e Maintainer Hadley Wickham. "Package tidyverse". Em: *Easily Install and Load the 'Tidyverse* (2017).

DADOS
https://s3.sa-east-1.amazonaws.com/ckan.saude.gov.br/SRAG/2021/INFLUD21-26-09-2022.csv