Previsão da evolução de pacientes do estado de Alagoas acometidos com Covid-19 utilizando métodos de Machine Learning

Gean Fernandes da Silva

26 de outubro de 2022

Resumo

O uso de modelos de previsão na evolução do Covid-19 é fundamental para auxiliar nas tomadas de decisões eficazes para o melhor enfrentamento da pandemia. Diante desse contexto, o objetivo desse trabalho é apresentar os resultados obtidos através do uso de técnicas computacionais que possa predizer qual a evolução de pacientes acometidos com o Coronavírus no estado de Alagoas, através da análise de informações como data de atendimento, idade, sexo, município de residência e comorbidades. Foi utilizado uma abordagem baseada na mineração de dados, com extração de informações através de algoritmos de Aprendizado de Máquina. Para preparação dos dados foram usadas algumas ferramentas do tidyverse como: readr, dplyr e ggplopt2. Ao todo foram definidas quatro etapas como: coleta, preparação e caracterização dos dados e elaboração do modelo de predição. Os resultados obtidos indicam que essa abordagem pode auxiliar na melhor compilação de dados referentes ao estágio final de cada paciente acometido com a doença, com uma acurácia de 0,9585, trazendo uma visão precisa da evolução dos pacientes no período da pandemia.

Palavras chave: Pandemia. Coronavírus. Predição. Mineração de dados.

1 Introdução

Com o passar dos anos, a aprimoração de tecnologias em todos os âmbitos, possibilitou melhores informações e disseminação de dados, por meio de organizações e pessoas [DRE19]. A publicação de dados permite o acesso direto a informações, portanto é fundamental ser publicado em formato legível e sem restrição de licenças, patentes ou mecanismos de controle [Bra22]. Esse cenário deve ser disponibilizado na sua forma bruta para que a sociedade possa produzir cruzamentos, interpretação e aplicações úteis [DUT13]. Com essas informações e com a evolução das estruturas computacionais é possível que as organizações consigam armazenar grandes quantidades de dados, ampliando a diversidade de fontes e tipos, que em conjunto possibilita uma análise vasta cada vez mais completa e detalhada [SOU20].

A aprendizagem de máquina no geral, pode ser classificada em aprendizagem supervisionada, onde os algoritmos ajustam parâmetros de um modelo a partir do erro medido entre respostas obtidas e

esperadas, os não supervisionados onde, os parâmetros de um modelo são ajustados com base na maximização de medidas de qualidade das respostas obtidas e os semi- supervisionados que é caracterizado pelo uso de algoritmos híbridos, que fazem uso dos recursos de correção de erro e de maximização de medidas de qualidade, conforme necessário [BRU15].

No final do ano de 2019, o surto do vírus SARS-CoV-2, tornou-se um dos grandes desafios do século XXI [BRI22]. A velocidade e a extensão da disseminação da Covid-19 que levou ao status de pandemia desafiou muitos sistemas de saúde. Incertezas referente as características do vírus, contaminação, grau de periculosidade, faixa etária mais propensa e progressão clínica da doença dificultou a contabilização e precisão dos reais dados da pandemia [CAN21].

Desse modo, o uso de modelos de previsão da evolução dos quadros de Covid-19 tornou-se essencial para auxiliar nas tomadas de decisões eficazes para o melhor enfrentamento da pandemia. Todavia, treinar bons modelos e executar ajustes de parâmetros e hiperparâmetros são tarefas desafiadoras quando se trata de um cenário de pandemia, onde há poucos dados disponíveis [SHI20].

Diante desse contexto, o objetivo desse trabalho é apresentar os resultados obtidos através das técnicas de *Random Forest* na construção de um modelo que possa predizer qual a evolução de pacientes acometidos com a doença no estado de Alagoas, através da análise de informações pessoais como data de atendimento, idade, sexo, município de residência e comorbidades.

2 Conjunto de Dados

Os dados coletados são oriundos da pandemia do Coronavírus, disponíveis no banco de dados do estado de Alagoas no site: https://dados.gov.br/dataset/painel-covid-19-em-alagoas , onde é evidenciado as informações interativas sobre a pandemia do Covid-19 em Alagoas e seus municípios. Com informações e atualizações diárias de saúde, estatísticas e geoespaciais para o acompanhamento dos casos em Alagoas. Os dados foram retirados no período de março de 2020 até junho de 2022.

Para preparação dos dados foram usadas algumas ferramentas do *tidyverse* como: *readr*, *dplyr* e *ggplopt2*. A *readr* é utilizada para fazer a leitura de dados e foi desenvolvida para ser um modo rápido e fácil de importar dados de fontes distintas. Já a *dplyr* é um pacote que realiza transformações de dados, aliando simplicidade e eficiência, com funções de selecionar colunas, ordenar bases, filtrar linhas, criar e modificar colunas, agrupar bases e sumarizar a base. E o *ggplopt2* é utilizado para visualização dos dados com a essência de construir camadas por camadas. Além de uma filosofia elaborada, ainda traz outras vantagens como gráficos bonitos, fácil personalização.

3 Aprendizagem de Máquinas

Aprendizagem de máquinas são conjuntos de elementos computacionais capazes de adquirir conhecimento a partir de dados, se baseando em duas etapas distintas: predição e validação. A escolha de um algoritmo não é algo simples na predição, pois não há um único algoritmo capaz de apresentar boa performance em todas as aplicações, sendo de grande importância a comparação para garantia de um modelo preditivo satisfatório do problema em questão. [SIL20].

O modelo utilizado no trabalho foi o de *Random Forest,* implementado como *rand_forest()* na linguagem R nd *tidymodels*. De acordo com [LIM20], esse modelo consiste em um classificador composto por múltiplas árvores, conhecidos como floresta de decisão. Essas são treinadas usando os diferentes subconjuntos dos dados destinados para o treinamento. Para classificar uma nova amostra, o vetor de entrada deve ser transmitido para cada árvore da floresta fornecendo um resultado de classificação. A floresta, então, escolhe a classificação que obtém mais "votos"(para o resultado da classificação discreta) ou a média de todas as árvores da floresta (para o resultado da classificação numérica).

Em *rand_forest()*, os hiperparâmetros são empregados para aumentar o poder de predição tais como: *tress* que indica o número de árvores construídas pelo algoritmo antes de tomar uma votação ou fazer uma média de predições, o *mtry* que indica o número de preditores que serão amostrados aleatoriamente em cada divisão ao criar os modelos de árvore e o *min_n* que indica o número mínimo de pontos de dados em um nó que são necessários para que o nó seja dividido.

4 Experimentos e Resultados

O trabalho descrito aqui teve como objetivo a construção de um modelo que possa predizer a evolução de pacientes acometidos com a Covid-19 no estado de Alagoas, através da análise de informações dos pacientes como data de atendimento, idade, sexo, município de residência e comorbidades. Foram definidas quatro etapas principais para o presente trabalho: coleta de dados, preparação dos dados, caracterização dos dados e criação do modelo de predição.

4.1 Coleta de Dados

Os dados coletados para a realização deste trabalho apresentam os seguintes atributos:

- Data de atendimento: data em que o paciente deu entrada no sistema;
- Idade: idade do paciente;
- Sexo: sexo do paciente;

- Município de residência: município de residência do paciente;
- Classificação: classificação final do paciente;
- Comorbidades: comorbidades apresentas pelo paciente;
- Situação do paciente: evolução do paciente;
- Data do óbito: data do óbito do paciente, se houver;
- Data de confirmação: data da publicação do exame de Covid-19.

Podemos observar na figura 1, os atributos encontrados no conjunto de dados após a utilização do comando *glimpse()*.

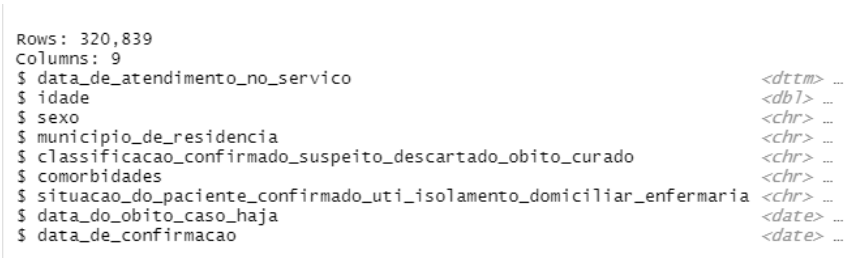

Figura 1: Atributos do Conjunto de Dados.

4.2 Preparação dos Dados

Após a coleta foi realizada a preparação dos dados. Essa etapa é primordial por ser o primeiro passo no processo de mineração. Inicialmente, foram realizadas a leitura dos dados, por meio do pacote *readr* disponível no *tidyverse*. Em seguida, foi feito a formatação dos nomes e colunas do dataframe, utilizando o pacote *janitor* e a eliminação das colunas de índices irrelevantes pra criação do modelo de predição, sendo possível a visualização das primeiras linhas da *dataframe* com a o comando *head()* e a visualização das últimas linhas do *dataframe* com o comando *tail()*.

Foi necessário a renomeação manual das colunas do *dataframe* mesmo com a utilização do pacote *janitor*, devido a presença de nomes poluídos. Em seguida, foi realizada a conversão dos dados de algumas colunas do *dataframe*. A coluna "data_atendimento" foi convertida para o tipo *Date*, enquanto as colunas "sexo", "municipio_residencia", "classificacao_doenca", "comorbidades" e "situacao_paciente" foram convertidos para o tipo *Factor*. Além disso, todas as letras foram convertidas em minúsculas.

Posteriormente, foi realizada a visualização das análises estatísticas dos dados através do comando *skim()*, disponível no pacote *skimr*. Após analisar o resumo estatístico fornecido pelo *skim()*, constatou-se que seria necessário realizar o tratamento de dados faltantes nas colunas "comorbidades" e "idade" e a eliminação da coluna "data_obito", uma vez que ela continha grande quantidade de dados faltantes.

Para a coluna “comorbidades” todos os dados faltantes foram convertidos em “sem comorbidade” enquanto na coluna “idade” a observação com dado faltante foi eliminada. Logo após, foi realizado uma nova visualização dos dados com o comando *skim()*, para confirmar que todos os problemas com dados faltantes foram resolvidos.

4.3 Caracterização dos Dados

Para a caracterização dos dados foi utilizado algumas ferramentas do *tidyverse* como: *dplyr* e *ggplot2* e algumas técnicas de agrupamento junto com a função s*ummarise()* para analisar as principais características dos dados.

Os dados foram agrupados de acordo com o atributo “data_atendimento” e foi utilizada a função *summarise()* para conseguir encontrar a quantidade de casos atendidos por dia e o maior e menor número de casos ocorridos por dia. Sendo assim, foi possível constatar 3.325 casos no dia de maior atendimento e apenas um no dia de menor atendimento.

Já para o agrupamento dos dados com atributo “idade”, a função *summarise()* foi usada para conseguir encontrar a quantidade de casos por cada uma das idades, incluindo a idade que obteve mais casos e a idade que obteve menos casos. Foi observado que pessoas com 39 anos, constituiu um total de 7.747 casos, enquanto idades de 111,114, 115 e 120 anos representou apenas um caso de ocorrência.

Para o agrupamento dos dados com atributo “sexo” a função *summarise()* foi utilizada para encontrar a quantidade de casos por sexo. Foi verificado uma superioridade de casos femininos em relação aos masculinos com um total de 183.707 no sexo feminino e 137.131 do sexo masculino. Posteriormente, utilizando a mesma função, para o agrupamento dos dados de “municipio_residencia” foi possível visualizar uma maior quantidade de casos no município de Maceió com um total de 125.031 casos, e uma menor quantidade no município de Jundiá com apenas 195 casos.

Para o agrupamento dos dados de “comorbidades”, utilizando a função *summarise()*, a maior comorbidade encontrada foi “Diabetes/has”, com um total de 769 casos. No agrupamento dos dados com atributo “situacao_paciente”, foi utilizado a mesma função, com objetivo de encontrar possíveis situações em que cada paciente se encontrava. A maior situação encontrada foi a de “isolamento domiciliar”, com 185.513 casos, enquanto a menor situação foi a de “óbito por outras pessoas” com 32 casos. Logo após as análises utilizando a ferramenta *dplyr*, foi realizada a análise dos dados através da criação e visualização de gráficos com a ferramenta *ggplot2*.

Foi criado um *boxplot* através da variável continua “idade” e da variável categórica “situacao_paciente”. Através dessa análise pode-se constatar que na maioria dos casos em que os pacientes tiveram a situação “recuperado” e “isolamento domiciliar” eles tinham menos de 50 anos. Enquanto na maioria dos casos em que os pacientes tiveram a situação “óbito”, “óbito por outras causas” e

“hospitalizado” os pacientes tinham mais de 38 anos. Podemos observar na figura 2, o gráfico boxplox criado através da idade e da situação do paciente.

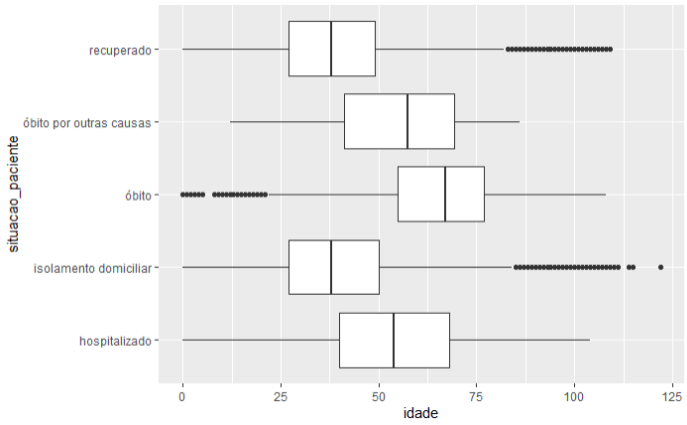

Figura 2: Boxplot criado através da idade e da situação do paciente. Imagem gráfico.

Foi criado um gráfico histograma através dos atributos “data_atendimento” e “situacao_ paciente”. Através dessa análise pode-se constatar que a maior quantidade de casos ocorreu no início do ano de 2022 e que em todo o período analisado as maiores situações encontradas foi a de ”isolamento domiciliar”seguida por ”recuperado”. Podemos observar na figura 3, o histograma criado através da data de atendimento e da situação do paciente.

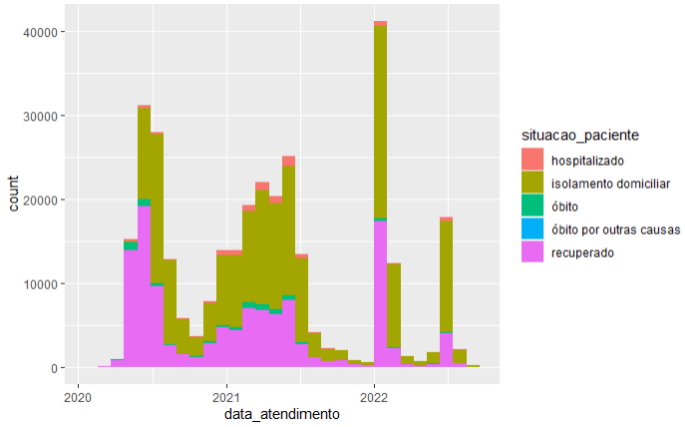

Figura 3: Histograma criado através da data de atendimento e da situação do paciente. Imagem gráfico.

4.4 Criação do Modelo de Predição

Após a etapa de caracterização dos dados foi realizada a construção de um modelo com o objetivo de predizer a evolução de pacientes acometidos com a Covid-19 no estado de Alagoas, utilizando a biblioteca *tidymodels*.

Para isso, foi realizada a separação dos dados de treino e teste utilizando a função *initial_split()* para realizar a divisão dos dados, e as funções *training()* e *testing()* para acessar a base de treino e a base de teste, respectivamente. Logo após, foi realizada a criação da receita utilizando a função *recipe()*, onde foi informado que todos os atributos influenciam na variável resposta. Nessa etapa, também foi criada uma especificação para normalizar os dados numéricos para que tenham um desvio padrão de um e uma média de zero.

Posteriormente, foi realizada a criação do modelo da floresta de decisão utilizando a função *rand_forest()*. Alguns hiperparâmetros utilizados nessa etapa foram: número de árvores do modelo que foi definido como "1000", o modo do modelo definido como "classificação", a *engine* do modelo, chamada de "*ranger*" e os parâmetros *mtry* e *min_n* que foram definidos como *tune()* para que fossem escolhidos os melhores valores no processo de tunagem.

Na criação do *workflow* foi utilizado a função *workflow()*, definindo o modelo e a receita criada, anteriormente. Em seguida foi realizada a reamostragem por validação cruzada utilizando a função *vfold_cv()* dividindo os dados em 10 amostras. A criação do *grid* de hiperparâmetros foi definida com os *mtry* e *min_n*. Para *mtry* foi designado os valores de 1, 2, 3 e 5, enquanto para *min_n* 4, 16 e 64. Na elaboração da tunagem, foram definidos o *workflow*, a reamostragem e o *grid*, todos criados anteriormente. Em seguida, foi realizada a análise dos resultados nesse processo. De acordo com a figura 4, a melhor métrica de acurácia foi encontrada utilizando o valor de 5 para o hiperparâmetro *mtry* e o valor de 64 para o hiperparâmetro *min_n*.

mtry <dbl>	min_n <dbl>	.metric <chr>	.estimator <chr>	mean <dbl>	n <int>	std_err <dbl>	.config <chr>
5	64	roc_auc	hand_till	0.8082635	3	0.007494723	Preprocessor1_Model12
3	64	roc_auc	hand_till	0.7969806	3	0.001678094	Preprocessor1_Model11
3	16	roc_auc	hand_till	0.7791819	3	0.002434658	Preprocessor1_Model07
5	16	roc_auc	hand_till	0.7790905	3	0.000468666	Preprocessor1_Model08
2	4	roc_auc	hand_till	0.7728172	3	0.006870284	Preprocessor1_Model02

Figura 4: Visualização dos melhores valores para os hiperparâmetros mtry e $\min_n$.$Imagem gráfico$

Após esse processo, foi realizada a seleção do melhor *workflow*, por meio da função *select_best()*, que em seguida foi finalizada através da função *finalize_workflow()*. Para melhor ajuste do *workflow*, foi usado a função *last_fit()*. Posteriormente, foi realizado o treinamento do modelo utilizando a base de treino e a função *fit()*. Após o treinamento do modelo foi realizada a predição com a função *predict()*. Por fim, foi efetuado a verificação da métrica de avaliação da acurácia, utilizando a função *metrics()*, e a acurácia encontrada pelo modelo de predição foi igual a 0.9585.

4.5 Considerações Finais

Sabe-se que durante esse período o uso de ferramentas da mineração de dados foram cruciais nas tomadas de decisões e designação de atividades que auxiliaram no enfrentamento da pandemia. Base-

ado nessa perspectiva, o uso de técnicas de aprendizagem de máquinas para predição da evolução de pacientes na região de Alagoas permitiu a mensuração da quantidade de casos e da característica de cada paciente acometido com a doença nessa região, sendo possível predizer a evolução de cada caso de acordo com os dados abertos disponíveis pelo governo. A acurácia encontrada foi de 0.9585, refletindo em uma alta precisão dos dados e algoritmos trabalhados.

Referências

[Bra22] Brasil. Agencia nacional de mineração de dados abertos. *https://www.gov.br/anm/pt-br/acesso-a-informacao/acoes-e-programas/dados-abertos*, 2022.

[BRI22] Sávio Breno Pires BRITO. Pandemia da covid-19: o maior desafio do século xxi. vigilância sanitária em debate. *http://dx.doi.org/10.22239/2317-269x.01531.*, 2022.

[BRU15] Sarajane; FREIRE Valdinei; LIMA Clodoaldo BRUNIALTI, Lucas; PERES. Aprendizado de máquina em sistemas de recomendação baseados em conteúdo textual: Uma revisão sistemática. *https://doi.org/10.5753/sbsi.2015.5818*, 2015.

[CAN21] Maria Grazia; CAPUA Ilaria CANNISTRACI, Carlo Vittorio; VALSECCHI. Age-sex population adjusted analysis of disease severity in epidemics as a tool to devise public health policies for covid-19. *http://dx.doi.org/10.1038/s41598-021-89615-4*, 2021.

[DRE19] Josef DREXL. Technical aspects of artificial intelligence: an understanding from an intellectual property perspective. *http://dx.doi.org/10.2139/ssrn.3465577*, 2019.

[DUT13] Karen Maria Gross DUTRA, Claudio Crossetti; LOPES. Dados abertos: Uma forma inovadora de transparência. *http://www.sgc.goias.gov.br/upload/arquivos/2014-04/dados-abertos—uma-forma-inovadora-de-transparencia1.pdf*, 2013.

[LIM20] Rosalvo Fereira LIMA, Jardel Ribeiro de; OLIVEIRA NETO. . eigenface vs random forest: um estudo comparativo em reconhecimento facial. *http://dx.doi.org/10.5753/sbseg.2017.19517*, 2020.

[SHI20] Gitanjali R. SHINDE. . forecasting models for coronavirus disease (covid-19): a survey of the state-of-the-art. sn computer science. *http://dx.doi.org/10.1007/s42979-020-00209-9*, 2020.

[SIL20] Lina Yara Monteiro Rebouças SILVEIRA, Francisca Raquel de Vasconcelos; MOREIRA. Utilização de algoritmos de aprendizagem de máquina na predição de arboviroses transmitidas pelo aedes aegypti. *http://dx.doi.org/10.21439/conexoes.v14i1.1824*, 2020.

[SOU20] Luca; ALVARENGA Miguel Bastos. SOUZA, Allan Rocha de; SCHIRRU. Direitos autorais e mineração de dados e textos no combate à covid-19 no brasil. *http://dx.doi.org/10.18617/liinc.v16i2.5536*, 2020.