Revista de Estudos Multidisciplinares, São Luís, v. 3, n. 1 (Número Especial XV Encontro Científico da UNDB), jan./mar. 2023. ISSN 2238-7897.

CIÊNCIA DE DADOS NO CONTEXTO PANDÊMICO DA COVID-19: a extração e tratamento de dados como suporte aos profissionais da saúde

DATA SCIENCE IN THE CONTEXT OF THE COVID-19 PANDEMIC: the extraction and processing of data as a support for health professionals

LA CIENCIA DE DATOS EM EL CONTETO DE LA PANDEMIA DE COVID-19: la extracción y procesamiento de datos como apoyo a los profesionales de la salud

Ana Fatima Assunção Leite [1]
Centro Universitário UNDB, São Luís, MA, Brasil.

Daniel Herrera de Oliveira Lemos[2]
Centro Universitário UNDB, São Luís, MA, Brasil.

RESUMO

O objetivo desse estudo é explicitar como a tecnologia por meio da Ciência de Dados pôde agregar à área da saúde durante o contexto da pandemia global de Covid-19. A metodologia empregada durante o desenvolver do algoritmo teve sua raiz no ciclo de vida de Ciência de Dados e todas suas etapas de funcionamento, a linguagem Python fora utilizada para construção do programa juntamente de bibliotecas de aprendizado de Máquina. Ao final da construção do algoritmo, nos resultados, foi constituído também uma *GUI (graphical user interface)* com base no modelo preditivo que objetiva identificar a suscetibilidade do usuário a um risco de morte alto ou baixo por Covid-19 por meio de informações atribuídas por ele mesmo, a precisão marca 92% de assertividade dos dados. Concluindo que o algoritmo é uma ferramenta com capacidade de auxiliar e orientar profissionais de saúde na tomada de decisões assertivas a

[1] Aluna de Engenharia de Software do Centro Universitário UNDB. E-mail: anafatima02@gmail.com.
[2] Prof. Especialista do Centro Universitário UNDB. E-mail: Daniel.lemos@undb.edu.br.

Revista de Estudos Multidisciplinares, São Luís, v. 3, n. 1 (Número Especial XV Encontro Científico da UNDB), jan./mar. 2023. ISSN 2238-7897.

cerca dos grupos de risco, além de auxiliar os próprios usuários do algoritmo a tomar mais medidas de cuidado com esse cenário global.

Palavras-chave: Covid-19; Ciência de Dados; Modelo preditivo; Algoritmos.

ABSTRACT

The objective of this study is to explain how technology through Data Science could add to the health area during the context of the global pandemic of Covid-19. The methodology used during the development of the algorithm had its roots in the Data Science life cycle and all its operating stages, the Python language was used to build the program and Machine Learning libraries. At the end of the construction on the algorithm, in the results a GUI (Graphical User Interface) was also constituted based on the predictive modeling that aims to identify the user-s susceptibility to a high or low risk of death by Covid-19 through information attributed by himself, the precision marks 92% of assertiveness of the data. Concluding that the algorithm is a tool capable of assisting and guiding health professionals in making assertive decisions about risk groups, in addition to helping algorithm users themselves taking more care in this global scenario.

Keywords: Covid-19; Data Science; Predictive modeling; Algorithms.

RESUMEN

El objetivo de este estudio es explicar cómo la tecnología a través de Data Science podría sumar al área de la salud durante el contexto de la pandemia mundial de Covid-19. La metodología utilizada durante el desarrollo del algoritmo tuvo sus raíces en el ciclo de vida de Data Science y todas sus etapas operativas, se utilizó el lenguaje Python para construir el programa y las bibliotecas de Machine Learning. Al final de la construcción del algoritmo, en los resultados también se creó una GUI (interfaz gráfica de usuario) basada en el modelo predictivo que tiene como objetivo identificar la susceptibilidad del usuario a un alto o bajo riesgo de muerte por Covid-19 a través de información atribuida por sí misma, la precisión marca el 92% de la asertividad de los datos. Concluyendo que el algoritmo es una herramienta con capacidad de asistir y orientar a los

Revista de Estudos Multidisciplinares, São Luís, v. 3, n. 1 (Número Especial XV Encontro Científico da UNDB), jan./mar. 2023. ISSN 2238-7897.

profesionales de la salud en la toma de decisiones asertivas sobre los grupos de riesgo, además de ayudar a los propios usuarios del algoritmo a tener más cuidado con este escenario global.

Palabras clave: Covid-19; Ciencia de datos; Modelo predictivo; Algoritmo.

1 INTRODUÇÃO

A transição do ano de 1818 para 1819 foi marcada com uma pandemia global de gripe espanhola e que deixou uma marca de, no mínimo, 50 milhões de vítimas em todos os continentes. O que configurou o poder mortal desse vírus foi a mutação do vírus influenza e um cenário sociopolítico caótico com a Primeira Guerra Mundial em acontecimento, a proliferação do vírus reivindicou a vida de milhões de indivíduos. A pandemia da gripe espanhola contou com três ondas de contágio, onde a segunda foi a que marcou os grandiosos números de mortos em consequência da grande taxa de contaminação da doença em relação às outras ondas (FIOCRUZ, 2020).

Em 2020, o mundo passaria por uma nova pandemia, o vírus SARS-CoV-2 causador da síndrome respiratória aguda grave 2 foi o responsável da vez. O mundo inteiro voltou os olhos para Wuhan, na China, onde os primeiros casos começaram a ser relacionados ao mercado úmido de Huanan (Zhu et al., 2020), já que os infectados compunham um raio de 800m de distância do mercado, posteriormente associou-se o início da infecção com o consumo de animais selvagens que eram comercializados no local, uma vez que o hospedeiro primário até então era o morcego.

Todavia, a transmissão do vírus de humano para humano começou a ocorrer de forma exponencial com o mundo globalizado do século XXI, os profissionais de saúde buscavam artifícios de descobrir características singulares da doença, que a divergisse de uma pneumonia e auxiliasse esses indivíduos nas tomadas de decisões a cerca de tratamentos para os numerosos casos e pacientes. Com isso, a tecnologia reafirmou sua utilidade na comunidade científica, o uso de Aprendizado de Máquina, Algoritmos e Métodos Computacionais serviram como suporte do diagnóstico, estudo e tratamento da doença.

Revista de Estudos Multidisciplinares, São Luís, v. 3, n. 1 (Número Especial XV Encontro Científico da UNDB), jan./mar. 2023. ISSN 2238-7897.

A corrida pela vacina foi marcada pela presença das áreas tecnológica para apoio de todos os profissionais do ramo de biotecnologia e saúde e é nesse cenário que esse artigo irá focar, na aplicabilidade da ciência de dados no cenário pandêmico da COVID-19, os dados que serão analisados tiveram sua origem e coleta durante o ciclo de testagem de vacinas, ciclo esse composto de três requisitos exigidos para aprovação: Imunogenicidade, eficácia e segurança. Tendo em vista sua origem, seu processamento irá resultar em informações que agregarão para identificar possíveis ligações entre infectados, como: idade, comorbidades, localidade etc. Como consequência, os profissionais da saúde poderão agir com base nesse padrão, estudar os grupos de risco e como diminuir a mortalidade.

A análise dos dados é feita em algoritmos baseados no ciclo de vida de Ciência de dados composta de quatro etapas: Coleta, organização, criação de soluções de apoio e acompanhamento dos resultados. E então, por fim, a predição dos dados será a contribuição para auxiliar o cenário pandêmico.

2 REVISÃO DE LITERATURA

O termo central desse artigo é: dado. Citando o Harrod's Librarians' Glossary (PRYTHERCH, 1995, p. 191, tradução nossa) para a definição de dado: "O menor elemento de informação". Tendo isso em mente, o dado é o menor fragmento de informação porque ele isoladamente não possui um significado, necessitando de um contexto de aplicação para que esse passe a significar algo. Com isso, a ciência de dados é o estudo desse fragmento de informação que vai buscar transformá-los em informação para um contexto específico, como: o de tomada de decisões.

A ciência de dados ajuda a responder e realizar questionamentos em diversos contextos sociais e econômicos, pensando nisso, a aplicabilidade dela seria um diferencial na área da saúde para suporte dos profissionais na área em tomadas de decisões. Tendo essa ideia em mente, o artigo buscou trazer como seria o uso da análise preditiva desenvolvida por uma ciência de dados no contexto da pandemia de Covid-19.

A infecção do coronavírus ocorreu de forma exponencial no final de 2019 e em dezembro do mesmo ano a Organização Mundial da Saúde (OMS)

Revista de Estudos Multidisciplinares, São Luís, v. 3, n. 1 (Número Especial XV Encontro Científico da UNDB), jan./mar. 2023. ISSN 2238-7897.

declarou que se tratava de uma crise de saúde global (Zhu et al., 2020), começava então a busca por uma vacina antes que o número de vítimas da pandemia aumentasse. Seres humanos de todas as faixas etárias tiveram reações diferentes e com intensidades diferentes à infecção do SARS-Cov-2, o que fez os estudiosos começarem a busca por um padrão que servisse de apoio aos profissionais da área da saúde na decisão de quem ficaria em isolamento em casa ou deveria ser internado e inclusive conduzido à UTI.

Com esse cenário em mente, a análise preditiva foi originada de técnicas de aprendizado de máquina e ciência de dados, sendo alimentada por dados obtidos na base da SEADE (Fundação Sistema Estadual de Análise de Dados) de São Paulo, que foi o primeiro epicentro dos casos da doença. Essa base era diariamente alimentada com os últimos casos de óbito por Covid-19, com o percentual do número de vítimas que era do sexo feminino ou masculino, faixa etária e se havia comorbidades. Com isso, foi capaz de desenvolver a busca por um padrão de risco, que buscava informar, além dos profissionais citados, os cidadãos de sua susceptibilidade de ser um paciente de risco ou não.

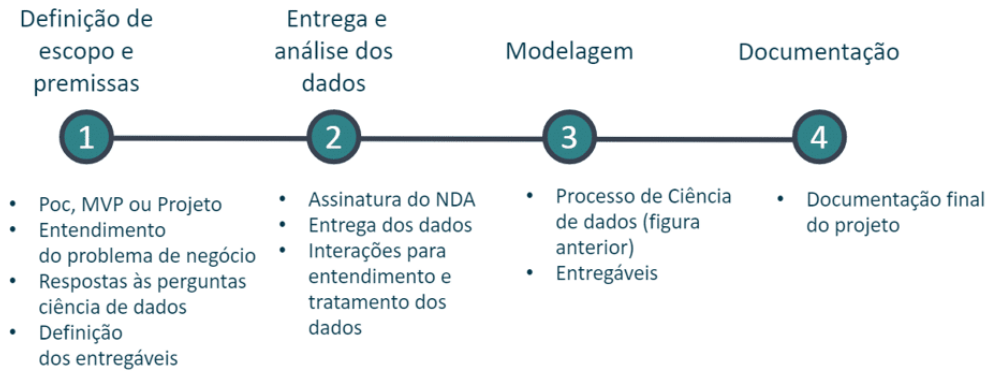
2.1 Fontes dos dados extraídos e trabalhados

Retomando a introdução prévia, a pandemia de Covid-19 é o contexto da fonte de onde os dados foram gerados e pegues para serem trabalhados e analisados, a análise essa que permite uma métrica de mensuração para observar padrões de risco entre os pacientes hospitalizados e cidadãos ainda não infectados e que se encaixam como grupo de risco, com alta susceptibilidade à doença.

Dando segmento, o ciclo de vida dos dados em projetos de Ciência de Dados é o referencial teórico que inicia os trabalhos dos dados, constituído de quatro etapas, sendo elas: Coleta de dados, tratamento dos dados coletados, modelagem e avaliações respectivamente. O algoritmo fundamentado no ciclo em questão serviu para que houvesse uma conexão entre sociedade e profissionais da linha de frente, uma vez que toda informação que fosse desenvolvida com base nisso, ela seria o mais próximo possível do panorama contextual. A figura 1 esquematiza o ciclo descrito acima.

Revista de Estudos Multidisciplinares, São Luís, v. 3, n. 1 (Número Especial XV Encontro Científico da UNDB), jan./mar. 2023. ISSN 2238-7897.

Figura 1 – Ciclo de vida dos Dados na Ciência de dados e suas fases de desenvolvimento de projetos.



Fonte: Tenbu (2021)

O algoritmo que informa o risco de morte por Covid-19 foi construído na linguagem de programação Python, usando uma biblioteca de manipulação e análise de dados pandas, a visualização foi desempenhada pela *matplotlib*, funções matemáticas realizadas pelo *numpy*, *seaborn* para Data Visualization, scikit-*lear* para o aprendizado de máquina e por fim, o *streamlit* que fez a *GUI* do projeto. Sendo esses os componentes do programa, o que os alimenta, os dados, por sua vez foram obtidos da base de casos confirmados e de óbitos por Covid-19 no Estado de São Paulo, devido ser a capital mais populosa do país, com 46,6 milhões de habitantes (IBGE, 2021)

2.2 Tratamento e limpeza dos dados

Os dados nem sempre são advindos de bases já tratadas e organizadas, o que acaba trazendo as bases de dados com campos em branco, colunas vazias e outros detalhes que requerem tratamento adequado antes de serem colocados em uso.

A correção ocorreu com a colocação de números zeros em colunas com campos em branco e transformando a coluna no tipo inteiro, para que não tenha problemas nas operações matemáticas ao algoritmo se deparar com uma célula de números zeros. Isso também foi feito nas colunas “óbito” e “sexo”.

2.3 Desenvolvimento do Algoritmo e a sua avaliação

Revista de Estudos Multidisciplinares, São Luís, v. 3, n. 1 (Número Especial XV Encontro Científico da UNDB), jan./mar. 2023. ISSN 2238-7897.

A questão central do desenvolvimento é a análise preditiva ou melhor, a modelagem preditiva dos dados que foram tratados, essa análise consiste em um procedimento feito com base algo: informação, dados, históricos ou padrões que empregados em um contexto, apresentam uma correlação lógica dos fatos. Com isso, é possível utilizar essa análise para tomada de decisões e estudo de padrões que podem se repetir.

Dando segmento, a predição desses dados foi realizada com o *Random Forest* que é um método de aprendizado de máquina que se baseia em árvores de decisão, empregado em casos de classificação, regressão e outros métodos construtores que vão originar casos de múltipla escolha até um ponto final. Com a previsão já realizada, é preciso ter a avaliação da precisão do algoritmo, buscando comprovar se o resultado que ele entregou está coerente com o cenário real ou não. Essa comparação é feita com a diferença do valor previsto e o campo de informação real e com isso, gera-se a acurácia do algoritmo com uma função já pronta do *Random Forest* que é a *score*.

A acurácia do algoritmo abordado foi de 93.2%, logo, 93% dos valores que foram previstos, estão em consenso com a realidade, são verdadeiros. No próximo tópico será abordado a *GUI* do algoritmo, ou seja, a interface do usuário onde ele terá interação e preencherá os campos para buscar padrões de risco de morte por Covid-19.

O potencial de precisão da *Random Forest* fora colocado em prova ao comparar com outros dois modelos de precisão, sendo eles: *KNN* (*K-Neareast Neighbors*) e *Decision Tree*. Os resultados comparativos estão explícitos na tabela abaixo.

Tabela 1 – Comparativo entre os modelos de predição de dados

Score	Model
93.20	Random Forest
93.20	Decision Tree
79.18	KNN

Revista de Estudos Multidisciplinares, São Luís, v. 3, n. 1 (Número Especial XV Encontro Científico da UNDB), jan./mar. 2023. ISSN 2238-7897.

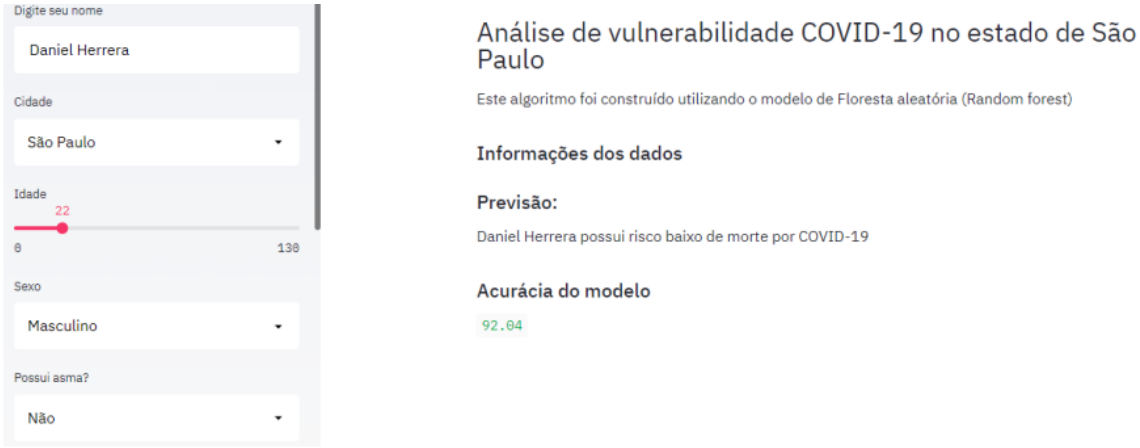
Fonte: Própria (2021).

2.3 *GUI (Graphical User Interface)*

Essa etapa vai focar na interface gráfica que o usuário do algoritmo irá se deparar para uso. Nessa interface, ele deverá submeter informações como: nome, idade, cidade, sexo, problemas respiratórios (como asma, por exemplo), problemas cardíacos, diabetes, obesidade, doenças renais, imunodeficiência, doenças hematológicas, doenças renais, problemas genéticos, se está em pós-parto ou gravidez e problemas neurológicos.

A assertividade do modelo preditivo em questão é de cerca de 92% no exemplo da figura 2, o algoritmo vai realizar uma busca pelos dados que retornem dos valores das pessoas que vieram a óbito seguindo o modelo de informações que o usuário forneceu ao sistema e vai comparar esses casos com os dados reais do banco de dados.

Figura 2 – Interface gráfica de usuário.



Fonte: Própria (2021).

Com isso, o sistema vai classificar qual nível de risco de morte aquele usuário com suas devidas características está enquadrado. Com isso, o alerta para os indivíduos que se encaixam no risco alto para Covid-19 busca pedir um cuidado redobrado daquele cidadão durante a pandemia do vírus SARS-CoV-2, seguindo instruções da Organização Mundial da Saúde (OMS) e respeitando o isolamento como forma de freio na exponencial contaminação do vírus.

Revista de Estudos Multidisciplinares, São Luís, v. 3, n. 1 (Número Especial XV Encontro Científico da UNDB), jan./mar. 2023. ISSN 2238-7897.

3 METODOLOGIA

A natureza desse artigo é de origem aplicada, buscando gerar conhecimento que tenha uso na sociedade empírica, que é especificamente o auxílio na tomada de decisões dos profissionais de saúde com base no modelo preditivo originado por uma ciência de dados durante a pandemia de coronavírus, além de também redobrar o cuidado dos cidadãos com base em seu nível de risco.

Seu objetivo é de base exploratória, investigando e aplicando informações do conteúdo abordado através de pesquisas bibliográficas de artigos e estudos já publicados para maior embasamento do artigo descrito acima sem interferir no cenário real de acontecimento dos fatos até então.

A abordagem desse é quantitativa, mensurando a tendência de um comportamento através dos dados obtidos e possibilitando criar uma mensuração de um padrão que pode repetir-se, dados esses que irão se comportar de forma clara e conclusiva quando empregadas no contexto de tratamento e empregabilidade deles no algoritmo.

Por fim, os procedimentos técnicos são de origem bibliográfica, como comentado no objetivo do estudo, a elaboração dele foi possível através de embasamento teórico prévio e documental com a extração dos dados através do banco de dados da secretaria de saúde de São Paulo, todo tratamento dessa fonte fora realizado devidamente antes de seu uso no algoritmo. Os critérios utilizados para encontrar as fontes bibliográficas mais adequadas foram:

- Recorte temporal do período de 2020 até 2022.
- SciELO e Google Acadêmico como bases de consultas de periódicos.
- AND (e) e NOT (não) como formas de refinar as buscas com critérios de adição e eliminação respectivamente, objetivando mais especificidade dos resultados que retornariam da busca.

4 RESULTADOS E DISCUSSÃO

Os dados que serão resultantes da pesquisa são variantes, mesmo que todos partindo do mesmo pretexto que é o de estudar o risco de morte de

Revista de Estudos Multidisciplinares, São Luís, v. 3, n. 1 (Número Especial XV Encontro Científico da UNDB), jan./mar. 2023. ISSN 2238-7897.

uma pessoa que tenha ou não comorbidades, idade, histórico clínico etc. A assertividade máxima de 93% é o coeficiente de confiabilidade deste algoritmo com os dados informados corretamente, a sua porcentagem pode ser desenvolvida futuramente para uma escala de acerto ainda maior, buscando chegar a precisão máxima. A análise preditiva tem sua origem de, principalmente, dados históricos, casos que já ocorreram. Logo, uma vez que a base de dados que alimenta aquele algoritmo seja maior, a sua gama de possibilidades de diferentes combinações de idades, comorbidade e históricos médicos será paralelamente maior e como consequência teremos uma maior margem comparativa do usuário com suas informações fornecidas para o algoritmo.

A modelagem primitiva que foi empenhada nesse estudo partiu da coleta de dados de treinamento, com isso, a base ainda que seja suficiente para o estudo ser realizado ela é um modelo mais simplório para desenvolvermos as primeiras métricas e identificação de relações entre variáveis, conclui-se, finalmente, que o algoritmo teve seu desempenho satisfatório, alcançando além dos 90% de assertividade em dados empíricos, ou seja, que tocam a realidade. Levando 100% como o máximo da métrica em exatidão, os 93% que foram marcados pelo algoritmo é um indício satisfatório que prova o uso da ciência de dados com aprendizado de máquina e modelagem primitiva ou análise primitiva possui um uso conclusivo na observação de padrão e tomada de decisões originadas de dados.

5 CONSIDERAÇÕES FINAIS

Esse trabalho teve como objetivo primário a contribuição da ciência de dados na ajuda da identificação se um indivíduo possui um risco de nível alto ou baixo de morte por Covid-19. No entanto, ainda há mais linhas com possibilidades de desenvolvimento e aprimoramento, como a hospedagem do algoritmo em um domínio para que usuários acessem e façam seu teste fornecendo todas as informações necessárias para que o programa tenha uma margem de acerto o mais próximo da realidade possível. Uma dessas possibilidades de aprimoramento seria a alimentação do algoritmo com bases de dados de todos os Estados do Brasil e possibilidade de empregar um *Data*

Revista de Estudos Multidisciplinares, São Luís, v. 3, n. 1 (Número Especial XV Encontro Científico da UNDB), jan./mar. 2023. ISSN 2238-7897.

*Visualizatio*n com gráficos que estimem cidades e estados que estão com novas ondas de contaminação e também infográficos com números de brasileiros que tomaram a vacina e quantas doses foram distribuídas no país. Trazendo com isso a geração de valor na sociedade brasileira através de tecnologias e suas aplicabilidades em diferentes áreas de atuação, como a abordada nesse estudo: a saúde.

REFERÊNCIAS

PRYTHERCH, Ray. **Harrod's Librarians' Glossary**. 10. ed. England, UK: Routledge 2005. Pg. 197. Disponível em: Harrod's Librarians' Glossary and Reference Book (unp.ac.id). Acesso em: 06 Nov 2022.

FUNDAÇÃO OSWALDO CRUZ (Fiocruz), Ministério da Saúde, Brasil. **Especial Covid-19 I A Fiocruz em dois tempos: nas pandemias da gripe espanhola e da Covid-19**. 2020. p.2 Disponível em: Especial Covid-19 | A Fiocruz em dois tempos: nas pandemias da gripe espanhola e da Covid-19. Acesso em: 05 Nov 2022.

SEADE. Fundação Sistema Estadual de Análise de Dados. **SP CONTRA O NOVO CORONAVÍRUS**. 2021 p.2. Disponível em: Coronavírus - Dados Completos (seade.gov.br). Acesso em: 05 Nov 2022.

GIL, A. C. **Métodos e técnicas de pesquisa social**. 6. Ed. São Paulo: Atlas, 2008. Disponível em: gil-a-c-mc3a9todos-e-tc3a9cnicas-de-pesquisa-social.pdf (wordpress.com). Acesso em: 04 Nov 2022.

Zhu, N., Zhang, et al. (2020). **China novel coronavirus investigating and research team. a novel coronavirus from patients with pneumonia in china**, 2019. N Engl J Med, 382(8):727– 733.

O CICLO de Vida dos Dados em Projetos de Ciência de Dados. **Tenbu**, 2021. Disponível em: O Ciclo de Vida dos Dados em Projetos de Ciência de Dados - Tenbu. Acesso em: 04 Nov 2022.